UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

MODELOS MULTIVARIANTES INTERNOS DE MEDICIÓN DE RIESGOS DE CRÉDITO, ACORDES CON BASILEA II

Tesis Doctoral

Fernando Mallo Fernández

Salamanca, Julio de 2011

MODELOS MULTIVARIANTES INTERNOS DE MEDICIÓN DE RIESGOS DE CRÉDITO, ACORDES CON BASILEA II.

Memoria que para optar al Grado de Doctor,
por la Universidad de Salamanca, presenta:

Fernando Mallo Fernández.

Salamanca

Julio de 2011

Universidad de Salamanca

Departamento de Estadística

Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora Titular del Departamento de Estadística

de la Universidad de Salamanca

CERTIFICA:

Que **D. Fernando Mallo Fernández**, Licenciado con Grado en Ciencias Matemáticas, Estadística e Investigación Operativa, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: *"Modelos Multivariantes Internos de Medición de Riesgos de Crédito, Acordes con Basilea II"* y para que conste, firma el presente certificado en Salamanca, el 17 de febrero de 2011.

Nunca andes por el
camino trazado,
pues él te conduce
únicamente hacia
donde los otros fueron.

(Graham Bell)

"¿Para qué repetir los
errores antiguos
habiendo tantos errores nuevos
que cometer?"

(Bertrand Russel)

AGRADECIMIENTOS

Deseo agradecer a mi Directora de Tesis, Dra. M. P. Galindo Villardón, su dedicación, su tiempo y los conocimientos aportados a lo largo de estos años.

Agradezco a *Caja España-Duero* la importante contribución a esta Tesis Doctoral al proporcionarme los datos sobre riesgo de crédito necesarios para poder construir el *modelo de credit scoring proactivo* que presento, además de su constante apoyo y confianza en mi trabajo.

A mis compañeros y amigos gracias por sus sugerencias y su disponibilidad a escuchar y a decir siempre esa palabra que hacía falta.

Gracias también a SAS® Institute, SAU, por poner a mi disposición, de forma totalmente altruista, sus eficaces herramientas de software estadístico, sin las cuales este trabajo no hubiera sido posible: SAS® Enterprise Guide, SAS® Enterprise Miner for Desktop, Credit Scoring for SAS® Enterprise Miner Client, SAS® Visual Data Discovery for Desktop y SAS®/IML, especialmente a Luis Méndez, Presidente de SAS® España y a Ignacio Poch.

Y un agradecimiento muy especial a mi familia, sobre todo a mi esposa Josefina, con quienes compartí esta experiencia tan apasionante, aunque nada fácil, por estar, tolerar, comprender y alentar y a esa preciosidad, Indira, que a pesar de llevar entre nosotros poco más de año y medio, consiguió que de vez en cuando me olvidase de las dificultades de esta investigación.

Y a todos cuantos me rodearon, gracias porque, sin que yo sepa a veces cómo, me ayudaron a lograrlo.

ÍNDICE

CAPÍTULO 1

INTRODUCCIÓN.

1.1 ESCENARIO ACTUAL DE LOS *MODELOS DE CREDIT SCORING.*

Hasta finales de la primera mitad del siglo XIX los clientes de los bancos constituían un número relativamente pequeño (según los estándares de hoy) de clientes de élite que podían ser "diseccionados" y evaluados de forma individual, y los préstamos, con la finalidad de satisfacer distintas necesidades, constituían un número manejable. Sería en el primer cuarto del siglo XX cuando la situación comienza a cambiar de rumbo según se va ampliando la base de clientes potenciales.

Desde comienzos de la segunda mitad del siglo XIX hasta nuestros días, esta percepción de la banca prácticamente ha desaparecido y, sobre todo en los últimos veinticinco años, ha habido un cambio radical en la forma en que los banqueros afrontan el riesgo en general y el riesgo de crédito en particular. Básicamente, el cambio es una respuesta directa al enorme aumento en la complejidad del negocio de crédito, en la tecnología informática, en las comunicaciones y en los conocimientos financieros.

En la última década del siglo XX y en la primera del XXI el sector de Bancos y Cajas de Ahorros, en adelante Entidades Financieras, ha experimentado una fuerte implantación en España, este hecho, independientemente de la profunda crisis que actualmente atraviesa el sistema financiero a escala mundial, ha venido acompañado de una fuerte vinculación de los clientes con su Entidad. Esta vinculación del cliente ha conllevado que en la actualidad las Entidades Financieras posean un volumen importante de clientes consumidores de productos financieros, tanto de pasivo como de activo. Si a esto se añade que, sobre todo durante el último quinquenio del siglo pasado y hasta finales del primer semestre de 2007, los precios a la baja del dinero amenazaban con reducir drásticamente los beneficios de las Entidades, lo que hizo que estas reaccionaran aumentado el volumen de operaciones concedidas, se entiende la necesidad de automatizar el proceso de concesión. Ya no era posible determinar la solvencia por el limitado número de expertos analistas de crédito basándose en su criterio y experiencia, puesto que era materialmente imposible que pudiesen inspeccionar las particularidades de cada solicitante de crédito, incluyendo los detalles de sus características socio demográficas, las condiciones económicas y sus intenciones de cumplimiento de pago para decidir si aceptar o rechazar la solicitud.

Como respuesta a la masificación del crédito, las importantes cantidades solicitadas por empresas o grupos empresariales y al problema que representaba la quiebra de significativas firmas Financieras, en junio de 2004 el Comité de Supervisión Bancaria del

Banco Internacional de Pagos de Basilea emitió el conocido documento sobre el Nuevo Acuerdo de Capital de Basilea II, (NBA, N*ew Basel* Capital *Accord)***,** referido a sus nuevas recomendaciones sobre requerimientos de capital a las Entidades Financieras, BCBS (2004),.

Por otra parte la crisis financiera que comenzó en 2007 llevó al Comité de Basilea a realizar una reforma de Basilea II con un nuevo acuerdo llamado de Basilea III. Apenas puesto en marcha el nuevo acuerdo de capital de Basilea II, con el objetivo de mejorar la habilidad del sector financiero para absorber el impacto de periodos de estrés financiero y económico, como el desencadenado en 2007, y evitar el contagio de dicho impacto a otros sectores de la economía, el día 12 de setiembre de 2010 el grupo de gobernadores de bancos centrales y los jefes de supervisión bancaria, bajo la presidencia del gobernador del Banco Central Europeo, Jean-Claude Trichet, aprobaron en Basilea una serie de medidas, acuerdos de Basilea III, esbozadas primero en los acuerdos del 26 de julio, para fortalecer la regulación global de capital y liquidez de los bancos. Entre otras importantes medidas, anunciaron que habían decidido aumentar los capitales mínimos de la banca, si bien establecieron un calendario de aplicación gradual, en el que también se incluye la tasa de liquidez que han de mantener las Entidades de Crédito para quedar a cubierto de cualquier riesgo ante una situación de tensiones en los mercados financieros, como las que se produjeron después de la quiebra del banco *Lehman Brothers*, en septiembre 2008.

Estas reformas están dirigidas a dos niveles diferentes, por un lado buscan fortalecer la posición individual de los bancos en periodos de estrés financiero, y por otro, reducir los riesgos a nivel sectorial que pueden surgir y la prociclidad de los mismos. Las reformas se traducen en una serie de obligaciones para las Entidades Financieras:

- Exigencia de capital de mayor calidad, transparencia y consistencia.
  Cuando las cosas van mal al mirar la caja lo único que nos queda es el capital desembolsado y reservas no repartidas. Es decir, que al final, pese a la sofisticación de los productos financieros, en momentos de tensión en el sistema la capacidad de absorber pérdidas radica en el capital de "toda la vida". Aspecto, sin duda, a tener en cuenta con la presente crisis, (TORNABELL, (2010), experto en temas financieros en ESADE).
- Fortalecer y mejorar la cobertura del riesgo.
- Introducción de un ratio de apalancamiento.
- Reducir prociclicidad.

- Establecimiento de un nuevo estándar global de liquidez mínima.

Analizaremos la forma en que influyen estas reformas, que se pretende sean aplicadas a todas las Entidades Financieras a partir de 2012, en la temática que tratamos en nuestra Tesis Doctoral una vez que expongamos como influyen los acuerdos de Basilea II sobre la misma.

De entre las novedades aportadas por el acuerdo Basilea II, con respecto al riesgo de crédito, destaca la Propuesta de *Sistemas Internos de Calificación de Riesgo de Crédito,* (IRB, *Internal Rating Based-approach*): conveniencia de que las Entidades Financieras gestionen el riesgo de crédito mediante *modelos propios de medición de riesgo. Estos* modelos han de ser construidos empleando los datos internos de cada Entidad para la estimación de los diversos componentes del riesgo de crédito: la probabilidad de incumplimiento (PD, *Probability of Default*), la pérdida esperada dado el incumplimiento, (LGD, *Loss Given Default*), y la exposición al riesgo en el momento del incumplimiento, (EAD, *Exposure at Default*), de las distintas carteras.

Para que un modelo sea considerado IRB debe incentivar la mejora de los sistemas y sus resultados, a la vez que valorar el riesgo y su estructura con mayor precisión. Por tanto, la aplicación de la propuesta IRB exige que las Instituciones Financieras proporcionen definiciones consistentes y bien estructuradas de sus sistemas internos, al tiempo que deben demostrar la robustez y eficiencia de los modelos aplicados, en definitiva se requiere "la adopción por parte de la Entidades Financieras de prácticas más sólidas en la gestión y medición de riesgos". A parte de que esto incide directamente en la calidad del riesgo asumido, lo que afecta de forma importante a la cartera de resultados, la adopción de métodos avanzados de gestión más sensibles al riesgo conlleva menores requerimientos de capital económico, cuestión en la que han incidido de forma importante los acuerdos de Basilea III, y aspecto importante en la actual reestructuración del Sistema Financiero Europeo. Ambos aspectos han incentivado, y continuarán haciéndolo en el futuro más inmediato, a las Entidades Financieras a una mejora continua en sus sistemas de calificación de los acreditados y de sus créditos.

El Comité de Basilea II y el Banco de España han querido propiciar la aplicación de modelos integrales en las Entidades, aplicables a todas o a la mayor parte de sus carteras, de tal forma que la aplicación de los modelos en su fase más avanzada sirva para:

1. Aprobar o denegar operaciones. Momento clave en la generación de la calidad de la inversión.

2. Asignar capital a las operaciones en función de su riesgo (RORAC, *Return On Risk-Adjusted Capital*) y fijar el precio mínimo de las operaciones (*Pricing*).

3. Calcular los requerimientos de Capital Regulatorio y Provisiones.

En los últimos años ha habido un aumento significativo en el uso de herramientas de la evaluación del riesgo de crédito en las principales Entidades Financieras de todo el mundo. De hecho, los modelos de riesgo de crédito han adquirido una importancia sin precedentes a partir de que Basilea II ha permitido a estas Entidades, en base de las exigencias de capital, construir por cuenta propia s*istemas internos de calificación de riesgo de crédito,* es decir, construidos con información interna de la Entidad, dada la capacidad explicativa y la confiabilidad de estos sistemas. Este es uno de los factores clave que en los últimos años ha llevado a las Entidades Financieras en todo el mundo a poner mucho énfasis y esfuerzo en el desarrollo de nuevas sistemas de riesgo de crédito o a modificar los existentes.

Con respecto a cómo afecta Basilea III a la temática que trata nuestra Tesis Doctoral, es evidente que la cuestión fundamental que se plantean las Entidades Financieras consiste en cómo abordar el riesgo para obtener un crecimiento rentable de acuerdo con los requerimientos de Basilea III y de las directrices del G20.

En este sentido, hay tres ámbitos principales en los que las Entidades Financieras pueden empezar a mejorar, (Culp y Nadal, 2010):

1.- *Las Entidades Financieras deben acelerar el proceso de integración de las funciones de riesgo y financieras*.

- El director financiero y el director de riesgos deben trabajar a partir de un único conjunto de información, y deben compartir una misma visión de las interrelaciones entre crédito, mercado y riesgo operativo.
- El riesgo debe considerarse como capacidad de crecimiento y aumento de la rentabilidad, no como una función de back-office o centrada en el cumplimiento.

2.- *Las Entidades Financieras deben mejorar la gestión global de datos*. Con Basilea III, de cara al futuro, los bancos deberán elevar significativamente el nivel de focalización y de rigor en relación con la calidad y el mantenimiento de datos en el conjunto de actividades de front office y back office. *Se mantendrán los modelos de riesgo estándar y les corresponderá a las propias Entidades asegurar que sus datos y modelos están a la altura de esta tarea*.

3.- *Las Entidades Financieras deben prepararse para el denominado "crecimiento inteligente".* En Asia, los préstamos a pequeñas y medianas empresas representan una gran oportunidad para las Entidades en economías de sostén como Corea del Sur, Singapur y Australia. Al entrar en estos nuevos mercados, las Entidades Financieras pueden atenuar el riesgo crediticio y operativo mediante la analítica y *automatizando la evaluación de la solvencia*.

La crisis financiera actual está evidenciando que no cabe esperar que los modelos de riesgo de crédito, incluso los modelos internos IRB, que sin duda constituyen un importante avance en la medición del riesgo, contemplen todos los escenarios posibles en que se desenvuelve este riesgo; de hecho estos modelos no han podido predecir sucesos de incumplimiento como los que ocurrieron en la segunda mitad del año 2008 y en 2009 en Estados Unidos y en Europa. Algunas Entidades Financieras disponían de modelos calibrados con promedios de algo más de cinco años, pero los sucesos a que nos hemos referido hacía más de 80 años que no acontecían. Se hizo necesario, por tanto, contar con instrumentos alternativos a los modelos estadísticos de predicción y clasificación, uno de estos instrumentos es la tasa de apalancamiento (*leverage ratio*), que es una medida más gruesa de capital y que consiste en una relación entre el *capital de nivel uno* y todos los *activos sin ponderar*. Entonces, *incluso, aquellas instituciones, que estén en modelos IRB avanzados, tendrán que respetar esta prudente tasa.* Este es *un cambio sustancial en la vinculación de la concesión de una operación a la puntuación otorgada por el modelo, pero no disminuye en absoluto la necesidad de estas potentes herramientas en el cálculo de la probabilidad de default para predecir la exposición al default y la pérdida esperada debida al default en orden al computo del capital requerido por los acuerdo de Basilea III y la concesión de crédito*.

En esta Tesis Doctoral nuestro interés principal está en el *credit scoring* (puntuación del crédito), que habitualmente forma parte del *sistema de calificación del riesgo de crédito* de cualquier Entidad Financiera. Es habitual que en la literatura se haga referencia a los modelos de la *credit scoring* como "*modelos o algoritmos cuyo objetivo consiste en determinar si procede o no conceder créditos a un solicitante, por lo que consiste en un problema de clasificación típica, en particular, un problema de clasificación binaria que tiene por objeto agrupar los solicitantes, ya sea como " no default",* (*solvente o bueno*) *o " default",* (*no solvente o malo*)". El punto de vista adoptado en esta Tesis Doctoral es que el objetivo anterior, siendo muy importante, no es el único, sobre todo desde la perspectiva

IRB de Basilea II, (BCBS (2006), por lo que contemplamos la *credit scoring* dentro del siguiente esquema, más general:

Según (BCBS (2006), § III, 444), en la aproximación IRB de Basilea II, la *probabilidad de default* juega un papel de extraordinaria importancia en los s*istemas de calificación del riesgo de crédito*, puesto que es la base para calificar a los acreditados y a sus créditos, así como para calcular la *pérdida esperada* en orden al cómputo del *capital económico*. Además deberá ser también la herramienta básica para la clasificación de nuevos solicitante de crédito entre "default" o "no default" en un horizonte próximo y todo bajo el condicionante de que es necesario poder establecer la influencia de las características observadas en el comportamiento de los acreditados frente al "default". Como consecuencia, un sistema de calificación de riesgo de crédito deberá contar al menos con tres instrumentos básicos:

1.- Una *Función de Probabilidad de Default* , $PD(\boldsymbol{X})$, que ha de basarse en la información interna de la Entidad, $(\mathbf{X},\mathrm{Y})$, y que ha de permitir conocer la influencia de las variables explicativas de riesgo de crédito, $\mathbf{X}$, en el comportamiento del acreditado frente al default, variable respuesta Y.

Entenderemos que un acreditado ha entrado en default cuando, de acuerdo con las especificaciones de Basilea II, ha incumplido con sus obligaciones de pago respecto al crédito contraído con alguna Entidad Financiera (BCBS, (2006), § III, 452, 453).

2.- Una *Función de Calificación de los Acreditados y de sus Créditos* , $S(\boldsymbol{X})$.
La tendencia es que en el futuro se requiera un criterio unidimensional como medida para el merecimiento o la calidad del crédito, y dado que la realidad que afecta al comportamiento del cliente frente al default conlleva múltiples factores es necesario transformar el conjunto de datos multidimensionales en un conjunto unidimensional con la mínima pérdida de información y preservando algunas propiedades tales como la monotonicidad.

La función $S(\boldsymbol{X})$, deberá contener la máxima información posible disponible sobre el merecimiento de crédito sobre el acreditado para obtener la probabilidad de default condicionada a esa información.

Las calificaciones del acreditado deberán reflejar la evaluación que realice el banco de la capacidad y voluntad del acreditado de atenerse al contenido contractual, incluso en condiciones económicas adversas o ante acontecimientos inesperados, (BCBS (2006), § III, 415).

En el mundo de los sistemas de calificación, en general, y en los modelos de credit scoring, en particular, es usual referirse a $S(\boldsymbol{X})$ como función de calificación de acreditados, puesto que puede interpretarse como la puntuación de merecimiento de crédito que le corresponde al cliente en base a las características observadas sobre él. Nos referiremos en toda esta memoria expresamente con este nombre a la función $S(\boldsymbol{X})$, a sus expansiones por funciones de base y a cualquiera de sus estimadores.

3.- Un Clasificador de Nuevos Solicitantes de Crédito en las clases de "default" y "no default" en un horizonte próximo, $\Gamma(\boldsymbol{X})$.

Además el NBC de Basilea II condiciona el número de clases a considerar en el comportamiento del cliente frente a las obligaciones de pago a dos {default} y {no default}, según se desprende de (BCBS (2006), § III, 465, que dice: "el banco podrá: (i) utilizar una estimación de la P*robabilidad de Default*, PD, adecuada para poder inferir la *Pérdida Media Esperada a largo plazo ponderada por el incumplimiento* o *pérdida dado el default*, LGD, o bien (ii) utilizar la segunda para inferir la primera".

Nos encontramos, de este modo, ante tres problemas binarios multivariantes, uno de *predicción*, otro de *calificación* y un tercero de *clasificación*, que han de resolverse a partir de la información proporcionada por el conjunto de características observadas sobre los acreditados, $\boldsymbol{X} = \left(X_1,..,X_j,..,X_P\right)^T$, vector de p variables aleatorias definidas sobre un espacio de probabilidad generado por la población de acreditados, y una variable respuesta aleatoria binaria, Y , que indica el comportamiento del acreditado frente al cumplimiento de sus obligaciones de pago, con valores 1, no cumple, (default), y 0, cumple, (no default).

Desde nuestra perspectiva, un sistema de credit scoring engloba los tres instrumentos: una *Función de Probabilidad de Default*, $PD(\boldsymbol{X})$, una *Función de*

Calificación de los Acreditados $S(\boldsymbol{X})$ y un *Clasificador de Nuevos Solicitantes de Crédito* en las clases de "default" y "no default", $\Gamma(\boldsymbol{X})$.

Habitualmente en la literatura suele hablarse de *scoring* cuando la puntuación se refiere a clientes particulares y *rating* cuando se refiere a empresas, en esta Tesis Doctoral, con el fin de simplificación, usaremos el término credit scoring para referirnos indistintamente a la puntuación y calificación de acreditados independientemente de que sean particulares o empresas, si bien nuestra aplicación práctica se desarrollara sobre un segmento de acreditados particulares.

Un modelo de *credit scoring* puede definirse, por tanto, como "*un modelo que nos permite usar el conocimiento sobre el cumplimiento en las obligaciones de pago y características de productos de crédito en el pasado, para pronosticar el cumplimiento de las obligaciones de pago de préstamos en el futuro*". De este modo cuando un analista de crédito valora el riesgo comparando mentalmente una solicitud de crédito en el presente con la experiencia que este mismo analista ha acumulado con otros clientes con solicitudes parecidas, está aplicando un modelo de credit scoring, aunque sea un modelo implícito y subjetivo.

Casos particulares del modelo de *credit scoring*, tal como lo hemos definido en el párrafo anterior, son los modelos estadísticos que se caracterizan por que usan el conocimiento cuantitativo acerca del cumplimiento de las obligaciones de pago y características de los productos de crédito pasados registrados en bases de datos electrónicas, para pronosticar el cumplimiento por parte de los acreditados de sus obligaciones de pago. Desde la perspectiva de Basilea II, los modelos de credit scoring proporcionan la probabilidad de default o impago, califican a los acreditados y clasifican nuevas solicitudes de crédito.

Los modelos de *credit scoring se clasifican en dos grandes grupos:*

1.- *Reactivos*, responden a una solicitud de crédito por parte del cliente. Para medir el riesgo se basan -principalmente- en los datos que aporta el cliente en la solicitud de crédito. Son válidos para evaluar tanto a clientes como a no clientes.

2.- *Proactivos*, se anticipan a las necesidades del cliente. Para medir el riesgo se basan en los datos que la Entidad ya dispone sobre el cliente. Están especializados en la evaluación de clientes vinculados. Si se cuenta con suficientes datos, un modelo proactivo de credit scoring hecho a medida puede llegar a reflejar perfectamente el perfil de los clientes de la Entidad.

1.2 OBJETIVOS DE LA TESIS DOCTORAL.

Una tarea esencial en la construcción de los modelos de credit scoring consiste en *la especificación del modelo* que conlleva necesariamente, al menos, fijar el nexo de unión entre la probabilidad de default y las variables explicativas del riesgo de crédito y fijar la expansión lineal de funciones de base, expresión formal de la relación de dependencia entre las componentes anteriores.

La situación en que todas las variables explicativas del riesgo de crédito son, o al menos se pueden suponer, lineales es con mucho la situación más apetecible, puesto que con ello se consiguen modelos paramétricos con lo que se alcanzan las cualidades idóneas para nuestro objetivo, cualidades ligadas a la triple calidad que los acuerdos de Basilea II requiere a un modelo, calidad explicativa, predictiva y discriminante, lo que se traduce en equilibrio entre flexibilidad y dimensión, interpretabilidad y capacidad de generalización.

Desde luego la linealidad es una característica muy deseable aunque no siempre alcanzable, y cuando no se alcance, lo que ocurre con frecuencia en la práctica, serán necesarios métodos alternativos a los modelos lineales, pero para ello debemos de enfrentarnos al *dilema de Occam* con decisión, lo que significa *obtener el modelo de la forma más sencilla que sea posible, salvaguardando la eficacia de los objetivos perseguidos adaptados a la política de riesgos de la Entidad Financiera.*

Nuestro objetivo principal consiste en la formulación de modelos de credit scoring, adecuados a los requerimientos de Basilea II, desde la óptica IRB, cuyas estructuras funcionales contemplen la posible no linealidad de las variables explicativas del riesgo de crédito en su relación con el cumplimiento de las obligaciones de pago de los acreditados.

Este objetivo general está formado por 3 objetivos específicos:

1.- Conseguir una *visión general y unificada de las técnicas más actuales de la estimación de las probabilidades de default, de la calificación de acreditados y de su clasificación*, formalizando sus estructuras funcionales como expansión lineal de funciones de base.

2**.-** *Plantear y formalizar nuevos modelos de estimación de la probabilidad de default, de la calificación de acreditados y de su clasificación, capaces de ir más allá de la linealidad establecida en los modelos clásicos de regresión y clasificación,* sin

necesidad de recurrir a algoritmos adaptativos artificiosos que generalmente no cumplen los requerimientos de Basilea II.

3.- Construir un *modelo proactivo de credit scoring*, bajo la óptica IRB de Basilea II, a partir de datos reales sobre acreditados, proporcionados por una Entidad Financiera española, comparando su rendimiento con otros modelos competidores. La estructura funcional del modelo, desarrollado y expuesto en todas sus fases, vendrá expresada como una expansión lineal de funciones de base, sugeridos por Hastie y Tibshirani (1996).

1.3 ANTECEDENTES.

1.3.1 Estado del arte.

A partir de los acuerdos sobre los requerimientos de capital económico de Basilea II, las Entidades Financieras deben calcular para sus carteras de crédito la *exposición al default*, EAD, y *la pérdida dado el default*, LGD, lo que implica calcular las probabilidades de default de los acreditados y sus créditos, PD, con lo que el tema trasciende ya las populares *tarjetas de puntuación (ScoreCard),* para calificar acreditados y solicitantes de crédito, desarrolladas por FAIR ISAAC, consultora americana de riesgo de crédito, en la década de 1960, y los desarrollos iniciales de Altman (1968) y Merton (1974).

Con la evolución de la tecnología informática, la mejora de las técnicas estadísticas convencionales y la aparición de algoritmos de aprendizaje máquina, se puede automatizar el tratamiento de los datos de los acreditados y solicitante de crédito para calcular la probabilidad de default, la función de calificación de acreditados y el clasificador Bayes optimal de nuevas solicitudes de crédito, todo ello basado en el merecimiento de crédito que para la Entidad Financiera merezcan los clientes vinculados o potenciales. Algunos de estos nuevos modelos, los menos, están demostrando que son una solución viable, desde la perspectiva de Basilea II, para afrontar el aumento en las solicitudes de crédito y la falta de expertos en riesgo de crédito; otros, los más, o no solucionan satisfactoriamente el problema.

Un primer grupo de las técnicas disponibles para el *credit scoring* intentan estimar la probabilidad de default de forma directa, por ejemplo, los k *Vecinos Más Próximos,* (**k-NN**, *k-Nearest Neighbors algorithm*), (Fix y Hodges, 1951, Loftsgaarden y Quesenberry, 1965, Moore y Henrichon, 1969, Wagner, 1973, Devroyé y Wagner, 1977), los métodos basados en los *estimadores de la densidad por funciones núcleo,* (**KDE**, *kernel density estimator*), (Fix y Hodges, 1951, Rossemblat, 1956, Parzen, 1962, Silverman, 1986, Scott,

1992), en particular el *estimador de Nadaraya-Watson*, (**NWE**, *Nadaraya-Watson estimator),* (Nadaraya, 1964, Watson, 1964), entre otros. Pero, ni los estimadores de la densidad por funciones núcleo ni los estimadores en vecindades locales, a través de los vecinos más próximos, son la solución; los primeros por que los estimadores de la densidad multivariante por funciones núcleo están aquejados de la *maldición de la dimensionalidad,* (*curse of dimensionality*), lo que los hace impracticables para más de cinco variables, y los segundos por que, aparte de proporcionar las estructuras menos suaves, suelen sobreajustar los datos lo que implica que sean muy poco generalizables, en el sentido que resuelven mal la clasificación de solicitantes de crédito que no hayan sido incluidos en la muestra de entrenamiento del modelo.

Como alternativa se pensó en los *Modelos Lineales de Probabilidad,* (**PLM**, *Probability Linear Model),* bajo la hipótesis simple de que *la función de calificación de acreditados* $S(\boldsymbol{X})$ *es lineal en* $\boldsymbol{X}=(X_1,...,X_P)^T$, es decir, hay razones suficientes para suponer que todas las variables explicativas son lineales, o que al menos este supuesto puede recoger adecuadamente la relación de dependencia entre la variable estado de default y las variables explicativas. Pero estos modelos, ajustados por regresión lineal, constituyeron un fracaso desde su propia concepción conceptual, por cuanto el estimador pronostica valores fuera del intervalo cerrado de números reales $[0,1]$, por lo que difícilmente pueden pronosticar una probabilidad. A pesar de este inconveniente, que los hace rechazables totalmente, estos modelos presentan una atractiva característica para nuestros objetivos, su estructura lineal es fácilmente interpretable.

Se hacían necesarias técnicas que proporcionaran modelos de probabilidad de default sencillos y fácilmente interpretables, preferiblemente lineales, a partir de los cuales se obtuviese la función de calificación de acreditados y un clasificador binario de default y no default.

Inicialmente se utilizó el Análisis Discriminante Lineal, (**LDA**, *Linear Discriminant Analisys*), (Fisher, 1936, Ladd, 1966, Lachenbruch, 1975), que no es más que la *Regresión Logística Lineal*, (**LLR**, Linear Logistic Regression), en el caso particular de que las distribuciones de X condicionadas al default y no default sean normales de distintas medias e igual matriz de covarianza, en cuyo caso la función de acreditados coincide con la frontera de clasificación del LDA,

clasificador Bayes optimal de nuevas solicitudes de crédito, y la probabilidad de default consiste en la transformación logística de tal función. Desgraciadamente en la práctica de los sistemas de calificación de acreditados las hipótesis de normalidad de las distribuciones de X condicionadas al default y no default no se verifican casi nunca y lo habitual es no conocer ni las probabilidades a priori ni las distribuciones de las verosimilitudes; de estas últimas, con frecuencia, no sólo no se conocen los parámetros sino incluso no se conoce la forma. Esto provocó que el foco de atención se orientase hacia las técnicas que proporcionando modelos igualmente parsimoniosos y fácilmente interpretables no requiriesen hipótesis tan restrictivas e irreales, en el mundo de la economía, como la normalidad de las distribuciones de las funciones de verosimilitud.

La atención se volcó en dos técnicas que resuelven el problema anterior, la *Regresión Probit,* (**LPR** o **PROBIT**, *Linear Probit Regression*), (Fechner, 1860, Gaddum, 1933, Bliss, 1934a y 1934b, 1935, Ashford y Sowden, 1970, Bock y Gibbons, 1996), y la *Regresión Logística Lineal*, LLR, (Berkson, 1944, 1950, Cornfield, 1951, 1956, Farrel, 1954, Aitchison y Brown, 1957, Adam, 1958, Cox, 1970, Ohlson, 1980, Hosmer y Lemeshow, 1989, 2000). Ambas técnicas proporcionan modelos que pertenecen a la familia de los *Modelos Lineales Generalizados*, GLM, (Nelder y Wedderburn, 1972). La única diferencia entre los modelos logísticos probit lineales y los logísticos lineales reside en la función de enlace entre la variable respuesta y las variables explicativas, probit y logística respectivamente, que, por otra parte, son muy similares.

La estimación del modelo conlleva la optimización de una función objetivo basada en una función de pérdida. Un método muy popular, conocido como *Principio de Inducción* consiste en minimizar la pérdida media empírica, error empírico, que consiste en la medida de la discrepancia entre los valores pronosticados por el modelo y los valores observados. Dado que el principio de inducción puede presentar dos serios problemas: *la no unicidad de la solución y el sobreajuste o infraajuste*, características estas dos últimas que restan al modelo capacidad de generalización, es necesario cuando esto ocurra acometer la R*egularización del Modelo,* que pretende conseguir modelos de tendencia antes que modelos muy ajustados localmente, es decir pretende "*suavizar el modelo*". El instrumento utilizado para ello es el *Funcional de Riesgo Regularizado o Estructural,* suma del riesgo empírico y un término de regularización o penalización.

La regularización de un modelo funciona como sigue: si se estima, por ejemplo, un modelo LLR y por alguna razón se observa infraajuste o sobreajuste, será necesario reajustar el modelo utilizando como función objetivo un funcional de riesgo regularizado. Si se utiliza como término de regularización la cantidad $J(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2$, siendo $\boldsymbol{\beta}$ el vector de parámetros del modelo LOGIT, a excepción del término intercepto, y $\|\bullet\|$ la norma Euclídea en $\mathbb{R}^p$, estaremos ante la *Regresión Logística Lineal L_2-Penalizada*, (**LLR_L_2**, *L_2-Penalized Logistic Regression*).

Por su estructura lineal los modelos PROBIT y LOGIT son los más parsimoniosos, amigables y de más fácil interpretación de los modelos paramétricos de estimación de la probabilidad de default, y, de hecho, **LLR**, por su eficacia ampliamente probada, es la técnica líder en *credit scoring* a nivel mundial, sobre todo en crédito a particulares, y, además, ambas técnicas cumplen adecuadamente los requerimientos de Basilea II, a pesar de lo cual no están exentas de críticas dentro de la industria de los sistemas de calificación del crédito; la más importante se corresponde con el hecho de que *no incorporan la posible no linealidad de las variables explicativas del riesgo de crédito*.

En la segunda mitad de los años 90, basados en la *Teoría Estadística de Aprendizaje,* se introdujeron como alternativa a los métodos logísticos los *Métodos de Separación*, que consisten en clasificadores que son extensiones del modelo *Perceptron*, (Rosenblat, 1958, 1962), que se ajustan por optimización de la función de pérdida "*margen suave" o vector soporte*. El modelo conocido en la literatura como *perceptron* desde hace más de 50 años, es un clasificador que computa una combinación lineal de las variables explicativas, el *Hiperplano Separador,* SH, y devuelve el signo. Estos algoritmos de aprendizaje, que intentan encontrar un hiperplano separador que minimice la distancia de los puntos mal clasificados a la frontera de decisión, establecieron los fundamentos de los modelos de redes neuronales de las décadas 1980 y 1990.

De entre los muchos problemas asociados con perceptron destaca sobre todo el hecho de que cuando los datos son separables, existen varias soluciones y cuál de ellas se encuentre depende de los valores de inicialización, (Ripley, 1996).

Con el objetivo de resolver el problema de la no unicidad de la solución que presenta SH, Vapnik (1996) introdujo el *Hiperplano Separador Optimal,* (**OSH**,

O*ptimal* ***S****eparating Hyperplane*), que separa las dos clases y maximiza la distancia para los puntos más próximos de una clase a la otra, es decir maximiza una noción geométrica de margen M, lo que se consigue a base de añadir al hiperplano separador restricciones adicionales.

Cuando los datos no son separables no existe el OHS, por lo que es necesario recurrir a técnicas alternativas. Una alternativa conocida como *Clasificador Vector Soporte,* (**SV**, *Support Vector*), (Vapnik, 1996), extensión del OSH, permite solapamientos de clases en el espacio de las variables explicativas, pero minimiza una medida del alcance de este solapamiento. Una forma de tratar con los solapamientos es relajar la maximización del margen M permitiendo que algunos puntos estén en el lado equivocado del margen.

Los métodos OHS y SV son muy populares por su buena ejecución en la clasificación binaria y según Härdle et al. (2007) las primeras aplicaciones prácticas apuntan a que estos métodos tienen una eficacia superior en la clasificación de nuevos solicitantes de crédito en comparación con el análisis discriminante, LDA y modelo LOGIT, sobre todo para muestras pequeñas. Sin embargo *tanto OSH como el clasificador SV tienen una debilidad importante desde el punto de vista de los sistemas de calificación de acreditados, no proporcionan directamente estimadores de la probabilidad de default subyacente y, además, presentan la misma debilidad que LOGIT y PROBIT, no incorporan la posible no linealidad de las variables explicativas del riesgo de crédito*.

Es poco realista pensar que la función de calificación sea lineal para todas las variables explicativas, de hecho en la práctica ocurre con frecuencia que la relación del estado de default con algunas de las variables importantes para explicar este estado no se manifiesta de forma lineal, por lo que los analistas de riesgos necesitan conocer a menudo de forma más precisa el modo en el que las variables explicativas X se asocian con el estado de default. Es necesario, por tanto, incorporar métodos alternativos a los modelos lineales. Esta necesidad provocó que en la década de 1980 se iniciara una febril actividad en la investigación para caracterizar la no linealidad de las variables explicativas en los modelos de credit scoring, ya fuera a través de modificaciones importantes en las técnicas existentes o desarrollando nuevos modelos estadísticos y algoritmos de aprendizaje máquina, actividad que continúa en la actualidad.

Se han abierto varias vías para *caracterizar la no linealidad de las variables explicativas en los modelos estadísticos de probabilidad de default:*

1.- *Una primera vía que intenta incorporar la no linealidad eliminando la maldición de la dimensionalidad la constituyen los Modelos Aditivos Generalizados,* (**GAM**, *Generalized Additive Model*), (Leontief, 1947, Friedman y Stuetzle, 1981, Stone, 1985, Hastie y Tibshirani, 1990).

Estos modelos son extensiones de los Modelos Lineales Generalizados, (**GLM**, *Generalized Linear Model*), *que combinan la flexibilidad de los modelos semiparamétricos de variables explicativas multidimensionales con la precisión estadística típica de una variable explicativa unidimensional, por lo que la maldición de la dimensionalidad no está presente en estos modelos. Los GAM, aparte de proporcionar eficaces reglas de predicción y clasificación, proporcionan directamente la probabilidad de default, la función de calificación correspondiente y el clasificador Bayes optimal de acreditados y solicitantes de crédito, así como herramientas para encontrar la importancia subyacente de las diferentes variables explicativas, lo que confiere a los modelos aditivos la habilidad de descubrir los patrones no lineales sin sacrificar la interpretabilidad, además, contemplan métodos estadísticos automáticos más flexibles que los métodos logísticos ordinarios.*

Dos técnicas logísticas pertenecientes a la familia de los GAM son la *Regresión Logística Aditiva*, (**ALR**, *Additive Logistic Regression*), y la *Regresión Logística Aditiva Regularizada*, (**RALR**, *Regularized Additive Logistic Regression*). Un caso particular de RALR es la *Regresión Logística Aditiva Regularizada por Splines,* (**SRALR**, *Splines Regularized Additive Logistic Regression*), que consiste en especificar la no linealidad a través de splines. Estas técnicas tienen en su contra que *la aditividad de los efectos de las variables explicativas es una exigencia bastante comprometida por ingenua*.

2.- Una segunda vía, menos comprometida que la exigencia de aditividad de los efectos de las variables explicativas y mucho más simple en su desarrollo teórico, consiste en dividir el espacio de características en un conjunto de híper-rectángulos, que constituyen una Partición Recursiva Binaria, (**BRP**, *Binary Recursive Partition*), que puede representarse en un grafo conexo sin ciclos, y entonces ajustar un simple modelo (igual a una constante) en cada uno, obteniendo de este modo una regla

de predicción de la probabilidad de default, el *Árbol de Clasificación*, (**TREE**, C*lassification Tree*), (Sonquist y Morgan, 1964, Morgan y Messenger, 1973, Breiman et al., 1984, Quinlan, 1986, 1987).

Los árboles de clasificación, aunque son conceptualmente simples, son muy atractivos por cuanto están dotados de una gran facilidad de interpretación, lo que constituye su mejor baza de cara a los requerimientos de Basilea II. Una de sus mayores fortalezas es que detectan de forma automática estructuras complejas, incluida la no linealidad, entre variables. Pero en la otra cara de la moneda las debilidades no son poca pues, aparte de su alta varianza, la probabilidad estimada es constante a trozos, lo que no es la forma que usualmente pensamos para la función subyacente real. La aproximación por saltos a la función de respuesta también implica que la estimación de la probabilidad no se encuentra entre las mejores, (el error de predicción puede ser mayor que en otros modelos más flexibles).

Por tanto, a pesar de que la facilidad de interpretación de los árboles de clasificación los convierte en métodos muy usuales de estimación de la probabilidad de default y de la clasificación de acreditados y solicitantes de crédito, y que están relativamente bien vistos por las Entidades Reguladoras de los Sistemas Financieros Internacionales, en particular del Banco de España, nos parece conveniente que su uso se acompañe de una gran dosis de cautela.

3.- Una tercera vía viene de la mano de una técnica a caballo entre las técnica de particiones recursivas, algoritmo CART, y los modelos aditivos Generalizados, GAM, *Splines de Regresión Adaptativos Multivariantes,* (**MARS***, Multivariate Adaptive Regression Splines*), (Friedman, 1991).

Las técnicas de construcción y estimación de los modelos de regresión MARS son técnicas semiparamétricas adaptativas, generalización de los modelos de regresión lineal paso a paso, que modelizan de forma automática relaciones no lineales e interacciones entre la variable respuesta y las variables explicativas. MARS utiliza algoritmos paso a paso hacia adelante para seleccionar los términos del modelo, *splines bisagra*, seguido por un procedimiento para podarlos.

El modelo MARS trabaja notablemente bien, pero no usa la información de que la variable respuesta estado de default es binaria, por lo que no restringe los valores ajustados entre 0 y 1. Para estimar las probabilidades en el caso de una *variable*

respuesta múltiple, (Stone et al., 1997) desarrollaron la técnica (**POLYMARS***, Multi-*Logistic *Multivariate Adaptive Regression Splines*) utilizando como función de enlace entre las probabilidades y la función que describe la relación de dependencia entre la variable respuesta y las variables independientes la transformación logística, de la que el caso de respuesta binaria es un caso particular, (**BIMARS**, *Binary*-Logistic *Multivariate Adaptive Regression Splines*).

Tampoco el método BIMARS resuelve satisfactoriamente la estimación de la probabilidad de default, por cuanto se configura la estructura formal del modelo suponiendo que todas las variables pueden ser especificadas por splines bisagra. Es evidente que una situación donde todas las variables pudiesen especificarse en ese modo sería absolutamente excepcional. Sin embargo es posible que la no linealidad de alguna de las variables pueda especificarse por los splines bisagra, hecho que tendremos en cuenta a la hora de construir el modelo HLLM proactivo que proponemos en esta Tesis Doctoral.

4.- Una cuarta vía para explorar ciertos aspectos de la estructura del modelo, con antecedentes en el Análisis de Componentes Principales, (**PCA**, *Principal Component Analysis*), consiste en proyectar los puntos sobre ciertas direcciones de interés*, Regresión de Búsqueda de la Proyección,* (**PPR**, *Projection Pursuit Regression*), (Friedman y Stuetzle, 1981, Jones y Sibson, 1987, Hall, 1989), usando la aproximación de funciones de alta dimensión por funciones más simples, (Kruskal, 1969, Friedman y Tukey, 1974).

Una limitación obvia de las particiones recursivas utilizadas en los árboles de decisión es que las fracturas de las variables explicativas se producen solamente en paralelo a las proyecciones de cada coordenada en particular. Las funciones constantes a trozos consideradas para predecir el default en los árboles de decisión en un sistema de coordenadas diferente al anterior, por ejemplo girado y trasladado, no podrían aproximarse bien.

Los modelos PPR son modelos aditivos y, por tanto, son generalizaciones de los modelos lineales que combinan la flexibilidad del modelado no paramétrico de inputs multidimensionales con la precisión estadística típica de una variable explicativa unidimensional, por lo que, la maldición de la dimensionalidad no está presente en estos modelos.

Pero a pesar de que el procedimiento trabaja notablemente bien, (Ripley, 1996), ni en la estructura ni en la estimación del modelo se tiene en cuenta su naturaleza binaria, tal como ocurre en MARS, de este modo no se restringen los valores ajustados a estar entre cero y uno y el ajuste se efectúa minimizando la suma de los cuadrados residuales, es decir el riesgo empírico cuadrático. Por tanto, el método sólo es útil para clasificación y no verifica los requerimientos más básicos de Basilea II. Para resolver este problema Roosen y Hastie (1994) presentaron una formulación para el caso de respuesta binaria, a la que llamaron *Regresión Logística de Búsqueda de la Proyección*, PPLR, dentro de un contexto de Búsqueda de la Proyección Generalizada para modelos de la familia exponencial. Estos autores usaron el armazón de la regresión logística con el modelo PPR con la idea de que un procedimiento hecho a la medida específicamente para respuesta binaria puede aproximar con más éxito la superficie subyacente que relaciona X e Y .

Pero, al igual que BIMARS, PPLR tampoco nos parece una satisfactoria adecuada para estimar la probabilidad de default, puesto que también en este método se configura la estructura formal del modelo suponiendo que todas las variables pueden ser especificadas por funciones caballete o sierra, las direcciones de interés a buscar. Si unimos a ese inconveniente la dificultad de interpretar las funciones sierra, *a causa de que cada variable explicativa entra en el modelo de forma compleja y multifacética, debemos concluir que, a diferencia de MARS cuyas funciones bisagra son de fácil explicación, las funciones sierra no parecen los más adecuados para especificar la no linealidad en modelos de credit scoring.*

5.- Una quinta vía, proveniente de área del *aprendizaje automático,* la configuran las siguiente técnicas: las *Redes Neuronales Artificiales*, (**ANN**, *Artificial Neural Network*), (Mcculloh y Pitts, 1943, Widrow y Hoff, 1960, Rosenblatt, 1962, Werbos, 1974, Parker, 1985, Rumelhart et al., 1986, Bishop, 1995, Ripley, 1996, Haykin, 1998, Hastie et al., 2001, 2009), los *Sistemas Inmunes Artificia*les, (**AIS**, *Artificial Immune Systems*), (Watkins y Bogges, 2002, Watkins et al., 2004), los *Algoritmos Genéticos*, (**GA**, G*enetic **A**lgorithm*)**,** (Holland, 1987, Goldberg, 1989, Davis, 1991, Fogel, 2006). Estas técnicas son clasificadores de eficacia probada en muchos campos de actividad, pero, en general, no cumplen los requisitos de Basilea II; la mayoría no proporcionan la probabilidad de default, aparte de que las posibilidades de que pueda interpretarse la pertenencia de un

acreditado a una clase en función de los valores observados de las variables explicativas sobre él son escasas.

Las R*edes Neuronales Artificiales*, ANNs, no gozan en general de buena fama en el campo de la predicción por cuanto ha existido un gran número de redes neuronales de tipo publicitario que han conseguido que estos modelos sean vistos como *cajas negras mágicas y misteriosas*, y si bien unos casos lo son, otros pueden englobarse dentro de los *modelos lineales de funciones de las variables explicativas,* (Bishop, 1995, Hastie et al., 2009), es decir son modelos del mismo tipo que, por ejemplo, la regresión polinómica en la que existe una única variable y cada función de base es una potencia de esta variable, o como una extensión de esta aproximación, en la que se divide el espacio de entrada en diferentes regiones y se aproxima cada región por un polinomio distinto, *regresión por splines* (Hastie et al., 2009), pero con una gran diferencia, las capas ocultas imposibilitan que el modelo explique el papel de las variables explicativas en el default pronosticado.

Una técnica muy especial de la familia de las ANNs, que se mueve en la misma línea que PPLR y con un modelo muy similar, es el *Modelo Perceptron de Capa Simple de transmisión de información hacia adelante y una sola capa oculta*, (**SLPM**, *Single Layer Perceptron Model*), donde si el algoritmo de aprendizaje es de retropropagación se denota por SLPBP, que conecta a esta familia con los *Modelos Lineales Generalizados, GLM* . De hecho si consideramos la red SLPM con la *transformación logística* y *funciones* de *Unidad Sigmoide (US)* con el modelo ajustado por el *criterio de pérdida de entropía cruzada,* estaremos ante un *modelo de regresión logística lineal en las capas ocultas,* y todos los parámetros son estimados por *máxima verosimilitud.*

A pesar de que algunas de las propiedades y características de las ANNs las han convertido en herramientas usadas con frecuencia en la resolución con éxito de problemas reales de gran complejidad, como por ejemplo, el diagnóstico médico, la decodificación del ADN y también en la calificación de créditos y acreditados en algunas Entidades Financieras, estos modelos están más orientados a la decisión sobre la concesión de un crédito que a estimar un modelo de probabilidad concebido desde la óptica de los acuerdos de Basilea II, en ese sentido, aparte de que no proporcionan en muchos casos la probabilidad de default, la capa oculta, (los modelos neuronales se construyen usando una arquitectura de capas fijas, *capa de entrada, oculta y de salida*), constituye siempre un problema, pues la naturaleza

de esa capa conlleva que no sea fácil explicar la relación entre el default y las variables explicativas del riesgo de crédito que, como venimos insistiendo, es imprescindible desde los requerimientos de Basilea II.

Con argumentos similares a los utilizados en el caso de redes neuronales podemos concluir que los *Sistemas Inmunes Artificia*les, AISs, y los *Algoritmos Genéticos*, GAs, hasta donde se nos alcanza su desarrollo actual, no resuelven de modo satrisfactorio el desarrollo de modelos de credit scoring desde la óptica de Basilea II.

6.- *Una sexta vía consiste en expandir el espacio original de entrada a un espacio de Hilbert, de forma que la carencia de un hiperplano separador en el espacio original sea suplida por una hipersuperficie en el espacio agrandado de Hilbert,* $\mathcal{H}_K$*,* para luego controlar la complejidad del modelo ajustando la función con el criterio de minimización del error empírico asociado a una cierta función de pérdida regularizada. Los métodos que estiman el modelo de acuerdo con la filosofía anterior se llaman *Modelos Regularizados por Núcleos,* (**KRM**, *Kernel Regularized Model*). La expansión del espacio original al espacio de Hilbert agrandado se consigue a través del *teorema representer no paramétrico,* (*Representer Theorem*), (Kimeldorf y Wahba, 1971, Cox y O'Sullivan, 1990, Schölkopf et al., 2001), que permite representar la función de calificación optimal como una combinación lineal de *funciones núcleo* $K(\bullet,\bullet)$*; en este* caso los núcleos computan productos interiores en espacios de características de alta dimensión.

Los modelos obtenidos por técnicas KRM vienen caracterizados por la elección del núcleo y la función de pérdida utilizada para el ajuste. En el entorno de los sistemas de calificación de acreditados, las dos funciones de pérdida más habituales en la regularización por núcleos son la pérdida logística y la pérdida bisagra, que respectivamente darán lugar a las técnicas *Regresión Logística Regularizada por Núcleos*, (**KLR**, *Kernel Regularized Logistic Regression*), (Green y Yandell, 1985, Hastie y Tibshirani, 1990, Wahba et al., 1995, Zhu y Hastie, 2004, , Hastie et al., 2001, 2009, y las *Maquinas de Vector Soporte,* (**SVM**, *Support vector Machines*), (Vapnik, 1996, Evgeniou et al., 1999, Wahba et al., 2000).

El principal inconveniente de la Regresión Logística por Núcleos Regularizada es que todos los vectores del conjunto de entrenamiento están involucrados en la solución final, lo que no es aceptable para grandes conjuntos de datos, debido al elevado número de

parámetros distintos de cero, como a veces ocurre en los sistemas de calificación de acreditados.

Bajo la misma filosofía que el método SV, que permite solapamientos de clases en el espacio de las variables explicativas a diferencia de OSH, si bien desde una óptica distinta que consiste en *suplir la carencia de un hiperplano separador en el espacio original con una hipersuperficie, a través del teorema representer no paramétrico, en el espacio agrandado de Hilbert,* $\mathcal{H}_K$.

A pesar de la fuerte debilidad que caracteriza a los modelos de vector soporte, OSH, SV y SVM, no es posible estimar directamente la probabilidad de default, sus defensores resaltan sus dos grandes fortalezas como métodos de clasificación, maximizan el margen y la frontera de clasificación está definido en términos de combinaciones lineales de los *puntos soporte* x_i, puntos que se definen sobre la frontera de la franja como $\alpha_i > 0$, que generalmente son un número muy reducido respecto del total de acreditados analizados. Apoyados en los dos puntos fuertes mencionados, que convierten a los modelos de vector soporte en magníficos clasificadores y restando importancia tanto al cálculo directo de la probabilidad de default como a la capacidad explicativa de los modelos, no son pocos los artículos aparecidos, hasta bien avanzada la primera década de este siglo, en que estos métodos parecen resultar claros vencedores frente a los modelos logísticos.

Zhu y Hastie (2004) y Rosset et al. (2004) demostraron que KLR es también una técnica maximizadora del margen, por lo que la ventaja de las técnicas vector soporte por ese hecho se desvaneció. Y, por otro, el principal inconveniente de la Regresión Logística Regularizada por Núcleos, *todos los vectores del conjunto de entrenamiento están envueltos en la solución final*, lo que se traduce en un número, con frecuencia inaceptable, de parámetros distintos de cero, fue resuelto un año más tarde también por Zhu y Hastie (2005) con su propuesta de evitar ese inconveniente usando un algoritmo de selección hacia delante que aproxima la expansión completa con un número fijado de parámetros α_i distintos de cero. Esta aproximación puede ser interpretada como añadir al criterio de optimización un término extra penalizando el número de coeficientes no cero, en esto consistió su propuesta de *Maquinas de Vector Importado*, (**IVM**, *Import Vector Machines*), o *Regresión Logística por Núcleos con Vector Importado*, (**KLR**$_{\text{IV}}$, *Import Vector Kernel Logistic Regression*).

Desde el punto de vista de Basilea II, a pesar de que KLR[IV] presenta todas las fortalezas de las técnicas de vector soporte y, además, resuelve el fundamental problema de estimar directamente la probabilidad de default que presentan estas técnicas, no nos parece una técnica conveniente para construir modelos de *puntuación de crédito* por cuanto no se resuelve el problema de la interpretabilidad, pues este modelo, como técnica KLR que es, describe la relación de dependencia de la variable estado de default con los acreditados y no con las variables de riesgo de crédito. No nos cabe ninguna duda q*ue como clasificador tiene tantas fortalezas como SVM, pero estas no son suficientes para convertirlo en un método satisfactorio en los sistemas de calificación del riesgo de crédito desde la óptica IRB de Basilea II.*

7.- La séptima vía supone una situación intermedia entre los modelos totalmente lineales, casi siempre poco realistas, y algunos de los descritos en las vías anteriores, lo que los convierte en métodos que participan de las ventajas de unos y otros e intentan resolver los inconvenientes de ambos, son los *Modelos Parcialmente Lineales Generalizados*, (**GPLM**, *Generalized Linear Partially Models*), que se configuran considerando como hipótesis de partida que se tiene información fundada sobre el hecho de que una o varias de las variables de interés tienen influencia lineal sobre el comportamiento frente al default y el resto influencia no lineal.

Dentro de la familia de modelos GPLM los más desarrollados son los pertenecientes a la subfamilia de los *Modelos Logísticos Parcialmente Lineales*, (**LPLM**, *Linear Partially Logistic Model*), (Severini y Wong, 1992, Severini y Staniswalis, 1994, Chen, 1995, Müller, 2001, Müller y Härdle, 2003, Peng, 2004, Peng y Wang, 2004, Härdle et al., 2004a).

La estructura formal más general del modelo logístico parcialmente lineal es una extensión semiparamétrica del modelo logístico lineal que se expresa como la suma de una componente lineal y una componente no lineal. La componente no lineal se diseña como una expansión no paramétrica infinito dimensional, es decir, a través de una función de suavizado no paramétrica de dimensión infinita que opera sobre un argumento multidimensional de las variables explicativas del estado de default y se calcula de una manera flexible, por ejemplo, cualquier método de suavizado no paramétrico, por lo que permiten un trato más flexible, para un subconjunto de las variables explicativas, que los modelos logísticos lineales.

Müller y Härdle (2003) propusieron un modelo logístico parcialmente lineal donde se exige que las variables de la componente no lineal sean absolutamente continuas, puesto que en su planteamiento la función no paramétrica de la componente no lineal se determina por un método de suavizado no paramétrico que requiere la continuidad absoluta de las variables explicativas. Esta técnica permite estimar simultáneamente las puntuaciones de crédito y las probabilidades de default, lo que provocó un creciente interés en los LPLM para rediseñar los sistemas de calificación del crédito y de los acreditados de acuerdo con los requerimientos del NBCA, BIS II.

Los Modelos Logísticos Parcialmente Lineales, a pesar de que resuelven algunos importantes problemas planteados en la estimación de la probabilidad de default: *se plantean en términos del logit de la probabilidad, permiten captar la linealidad paramétricamente, consideran la componente no lineal y no son excesivamente complejos, siguen sin resolver un problema importante desde el enfoque de Basilea II,* la parte no lineal se estima a través de una función no paramétrica, infinito dimensional, que generalmente no permite explicar la aportación de cada variable explicativa del riesgo de crédito al estado de default. Por esta razón *tampoco los consideramos modelos satisfactorios para nuestro objetivo de estimar la probabilidad de default, calificar a los acreditados y clasificar a nuevos solicitantes de crédito de acuerdo con los requerimientos de Basilea II. Además, como ocurre con casi todos los métodos no paramétricos estas técnicas están aquejadas de la maldición de la dimensionalidad, pero constituyen la semilla que origina los modelos HLLM que proponemos en esta Tesis Doctoral, bajo la sugerencia de Hastie y Tibshirani* (1996).

Con el fin perseguido de resolver los dos problemas que afectan a la técnica anterior, surgió una peculiar familia de modelos logísticos parcialmente lineales, los *Modelos Logísticos Aditivos Parcialmente Lineales*, (**LPALM**, *Linear Partially Additive Logistic Model*), que pueden considerarse desde dos ópticas: como *Modelos Logísticos Aditivos con una Componente Lineal* o bien como una extensión de los *Modelos Logísticos Lineales con una Componente No Lineal Aditiva*.

La componente no lineal en estos modelos se configura como una estructura aditiva de los efectos no lineales, y, por tanto, la formulación general de los LPALM puede verse como una extensión semiparamétrica del modelo LOGIT.

Como de costumbre se recurre a la aditividad para resolver el problema de maldición de la dimensionalidad y para conseguir un modelo más interpretable. Decíamos en la descripción de los modelos GAM, que los modelos aditivos presentan dos muy buenas cualidades, por un lado, dado que cada uno de los términos aditivos se estima usando un suavizador univariante, se esquiva la maldición de la dimensionalidad y, por otro, las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas del riesgo observadas sobre los acreditados. Pues bien estas cualidades son aplicables a la parte no lineal del modelo, la componente lineal las tiene garantizadas por construcción.

A pesar de que esta segunda propiedad hace a los modelos aditivos, sobre todo a los obtenidos a través de la pérdida logística particularmente atractivos en los sistemas de calificación de acreditados en orden al cumplimiento del requerimiento de Basilea II, respecto de la interpretabilidad del modelo, tal como se dijo en la en la primera vía, *la aditividad de los efectos de las variables explicativas es una exigencia bastante comprometida*.

8.- Una octava vía para contemplar en el modelo cualquier estructura compleja, tal como la no linealidad, la constituyen los modelos *Logísticos Basados en Funciones de Proximidad*, (**PFBLM**, *Proximity Function-Based Logistic Model*), (Matusita, 1956, Krzanowski, 1987, Cuadras, 1989, Cuadras y Fortiana, 1995, Cuadras et al., 1997, Boj et al., 2009a). Esta técnica utiliza del análisis discriminante a través de *funciones de proximidad entre individuos y poblaciones*, introducido por (Cuadras et al., 1997), para, sin el conocimiento de la distribución de las poblaciones de default y de no default, obtener *una regla "matusita" para la clasificación de nuevas solicitudes de crédito.*

Las *funciones de proximidad* se definen a partir funciones de distancia con la propiedad *Euclídea* para las población de default y no default y de la *variabilidad geométrica* definida por una disimilaridad, (Cuadras y Fortiana, 1995). Boj et al. (2009a), dentro del esquema anterior y con el fin de aplicar el método al *credit scoring*, definieron el logit de la probabilidad de default como la diferencia de las funciones de proximidad para la población de no default y default, en este orden. Por lo que se tiene que la función de enlace entre la probabilidad de default y la función de calificación obtenida a partir de

las *funciones de proximidad* entre los acreditados de las poblaciones de default y no default es la *transformación logística*.

Tenemos, por tanto, que sobre la base de este enfoque del análisis discriminante a través de funciones de proximidad entre individuos y poblaciones, (Cuadras et al., 1997), es posible obtener, sin el conocimiento de la distribución de las poblaciones de default y de no default, un planteamiento general de Modelos Logísticos basados en funciones de proximidad y, por tanto, en distancias. Utilizar las funciones de proximidad nos permite de modo natural utilizar como variables explicativas del riesgo de crédito variables numéricas y categóricas, así como incorporar la no linealidad al modelo. Además, este modelo tiene la ventaja de que siempre se puede definir una distancia entre las observaciones correspondientes a los acreditados, por lo que es siempre es posible plantearlo.

Una característica interesante del enfoque anterior es que las *funciones de proximidad pueden venir definidas por transformaciones que generan representaciones del espacio métrico de los acreditados en espacios de Hilbert,* $\mathcal{H}$, es decir, *las funciones de distancia que generan las funciones de proximidad son Euclídeas* en el sentido de los Escalogramas Multidimensionales Métricos, (**MDS**, *Metric Multidimensional Scaling*), (Gower (1982).

El modelo actualmente, tal como nosotros lo conocemos, está *orientado a la clasificación* dada la dificultad de interpretación de las funciones de proximidad en términos de las variables explicativas del riesgo. Además para un número grande de acreditados, lo más usual en las aplicaciones reales de *credit scoring*, *está técnica requiere tiempos de computación impracticables*.

El método posee una buena parte de las cualidades para que se sitúe en la vía de las técnicas viables para ser utilizadas en la *puntuación del crédito*, de acuerdo con los requerimientos de Basilea II, pero creemos que todavía no son suficientes. A pesar de que proporciona la probabilidad de default y un potente clasificador, la función de calificación y, sobre todo, su interpretación aún dista bastante de lo que Basilea II recomienda, sin embargo la posibilidad de cuantificar la importancia relativa de los factores potenciales de riesgo (Boj et al., 2009b), parece apuntar en una buena dirección y no cabe duda de que merece la pena, dado el potencial del método, continuar investigando estos aspectos.

9.- Por último, analizamos la novena vía, clave en esta Tesis Doctoral, iniciada por Hastie y Tibshirani (1996), que sugirieron que una forma natural de generalizar los *Modelos Logísticos Lineales* consiste en reemplazar las variables explicativas

por una versión no paramétrica de las mismas, superficies arbitrarias de regresión tales como las funciones sierra, superficies más estructuradas de alta dimensión como las funciones bisagra obtenidas por el procedimiento MARS de Friedman (1991), o modelos paramétricos de mayor complejidad como los modelos aditivos o *los modelos cuya estructura funcional viene dada por expansiones lineales de funciones de base*, (*linear expansión of bases funciones*).

Como veremos en el capítulo 5, la mayor parte de las técnicas a las que nos hemos venido refiriendo responden a la sugerencia de Hastie y Tibshirani (1996), sin embargo los Modelos de Probabilidad Generalizado como representación de una *expansión lineal por funciones de base del vector de variables explicativas*, hasta donde se nos alcanza, no se han llegado a formalizar ni a estudiar de manera unificada, a pesar de que Hastie et al. (2009), dedican el capítulo 5 de su obra a las *expansiones de base y regularización* y textualmente dicen en la página 139: "*En este capítulo y el siguiente se discuten los populares métodos para ir más allá de la linealidad. La idea central en este capítulo es aumentar/reemplazar el vector de entradas X con variables adicionales, las cuales son transformaciones de X, y entonces usar modelos lineales en este nuevo espacio deducido de las características de entrada."*

Aplicar a nuestro caso la sugerencia de Hastie y Tibshirani (1996) supone que **la estructura funcional de un modelo que represente la relación existente entre una conveniente transformación de la probabilidad de default y las variables de riesgo de crédito explicativas del estado de default, incluidas las variables no lineales, viene dada por una expansión lineal de funciones de base de las variables de entrad**a. Estimar la función de calificación es ahora equivalente a encontrar los coeficientes de las funciones de base en la expansión lineal, y algunas veces, los parámetros desconocidos de las funciones de base si es el caso. "L*a belleza de esta aproximación es que una vez que las funciones de base han sido determinadas, los modelos son lineales en este nuevo espacio expansionado de las características de entrada resultantes, y el ajuste se hace como para los modelos lineales",* (Hastie et al. 2009).

La Regresión Logística Lineal es una técnica para estimar modelos de puntuación de crédito cuya conveniencia, como ya hemos dicho, suscita un amplio consenso entre investigadores, expertos en riesgo de crédito y Entidades Financieras, sobre todo para préstamos a particulares, sin embargo, en la necesidad de introducir o no la no

linealidad, y sobre todo, en que técnica es la más adecuada o más conveniente para contemplarla no hay consenso.

1.3.2 Aplicaciones prácticas notables de modelos con componente lineal y componente no lineal en credit scoring.

En esta subsección haremos un breve recorrido sobre algunas de las aplicaciones prácticas, representativas de las que introducen, por las más variadas técnicas, la no linealidad en los modelos de la credit scoring, y que contribuyeron a motivar gran parte del grueso de las contribuciones de esta Tesis Doctoral, (Müller y Härdle, 2003, Härdle et al., 2005, Siddiqi, 2006, Liu y Cela, 2007, Hervás-Martinez et al., 2007, Liu y Cela, 2009).

1.3.2.1 Müller y Härdle (2003). *Exploring Credit Data.*

La primera aplicación práctica sólida de modelos con una componente lineal y una componente no lineal en el dominio del credit scoring se debe a Müller y Härdle (2003) que presentaron el *Modelo Logístico Parcialmente Lineal*, LPLM, como una herramienta exploratoria para una situación práctica usando datos bancarios de pequeños consumidores provenientes de un banco francés, con 6.180 casos y 24 variables, ocho numéricas y quince categóricas.

En primer lugar, especifican el modelo logístico lineal para todas las variables, $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=2}^{24}\beta_r X_r$, que se ajusta al conjunto total de datos disponibles por máxima verosimilitud, centrando su atención en aquellas variables explicativas numéricas cuyos coeficientes no son significativamente distintos de cero, por cuanto la no significación de los coeficientes indica que estas variables o bien no tienen influencia sobre la respuesta o bien su influencia no es lineal, es decir su especificación es insuficiente.

Una vez determinadas, a través de test estadísticos adecuados, las variables que presentan no linealidad en el modelo, y que, por tanto, son las variables a considerar para una modificación no lineal del mismo, analizan el modelo logístico cuya estructura funcional viene dada por expansiones lineales de funciones de base correspondiente, incorporando la no linealidad a través de funciones no paramétricas infinito dimensionales. Por ejemplo, una variable que en su trabajo resultó no lineal en el modelo es la etiquetada como X5, el modelo LPLM adquiere la siguiente estructura funcional

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=2}^{24}\beta_r h_r\left(X\right)=\ \beta_0+\sum_{r=2,r\neq 5}^{24}\beta_r X_r+h_5\left(X_5\right) \qquad (1.1)$$

donde $h_r\left(X\right)=X_r$, para $r=2,...,24$ y $r\neq 5$.

Müller y Härdle (2003) estiman el modelo (1.1) utilizando la *máxima verosimilitud semiparamétrica*, una integración de la técnica de optimización de verosimilitud clásica (paramétrica) para estimar β_0 y los β_j y una técnica local de optimización de la verosimilitud de suavizado para estimar las funciones no paramétricas infinito-dimensionales $h_r(\bullet)$, es decir, el *Algoritmo de Newton-Raphson para el Perfil de Verosimilitud*, (Severini y Staniswalis, 1994).

Es evidente que la interpretabilidad se resiente puesto que no se interpreta con la misma facilidad la contribución, por ejemplo, de la variable X2 al $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)$, que viene dada por $\hat{\beta}_2$, que la de X5 cuya contribución al modelo viene dado por la compleja expresión $\hat{h}_5\left(X_5\right)$, obtenida por iteración, mediante el Algoritmo de Newton-Raphson para el Perfil de Verosimilitud, y donde el valor actual conlleva la estimación asintótica de funciones no paramétricas y pesos locales definidos por funciones núcleo multidimensionales. Creemos que es esta la razón por la que los autores de referencia propusieron su modelo como una herramienta exploratoria más que como un modelo susceptible de ser adoptado como herramienta de pronóstico final, conscientes de su deficiencia en términos explicativos en la práctica de la credit scoring.

1.3.2.2 Bonilla et al. (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring.

En este trabajo Bonilla et al. (2003) realizan un exhaustivo estudio de la capacidad explicativa de dos modelos paramétricos lineales, la *Regresión Logística Lineal*, LLR, y el *Análisis Lineal Discriminante*, LDA, y cinco técnicas no paramétricas de clasificación, dos de ellas son árboles de clasificación obtenidos por el algoritmo CART, el primero, y por el algoritmo C4.5. (Quinlan, 1993) el segundo; las otras tres son la *Regresión Local Ponderada*, WLR, *Weighted Local Regression,* (Cleveland, 1979)*, los Splines de Regresión Adaptativos Multivariantes, MARS, y las Redes Neuronales Artificiales,* ANNs.

Para su estudio, *sin ninguna relación con los requerimientos de Basilea II*, utilizan los datos de un estudio previo de Quinlan (1987) con 690 acreditados demandantes de una tarjeta de crédito, sobre los que se han observado 14 características. Dada la escasez de

datos, dividen su muestra en una de entrenamiento con 600 individuos y otra test con 90. Para evitar el problema del sobre aprendizaje, debido, por un lado, a la escasez de datos, y, por otro, a la sobre parametrización típica de los métodos no paramétricos, pues con frecuencia la estructura es tan compleja que el modelo ha "memorizado" la muestra, lo que se traduce en una débil capacidad de generalización, utilizan el *método de validación cruzada*, (Stone, 1974), para elegir la estructura idónea de los modelos no paramétricos, con el fin de obtener una adecuada generalización del problema.

La principal conclusión de este estudio es que "*los modelos no paramétricos no dominan de forma sistemática a los paramétricos*", lo que según sus autores "*contradice en cierta medida algunos resultados de la literatura*".

Adicionalmente comprueban que el *método de validación cruzada,* aunque les ha permitido obtener la estructura óptima*, no ha resultado adecuado en el problema de la generalización,* puesto que en todos los modelos el error esperado resultó superior al real.

Por último, "*a pesar de que la ANN ha resultado ser la de mayor capacidad explicativa, superando a todos los demás modelos, la escasez de datos dificulta severamente una adecuada comparación entre los modelos, por lo que no es posible asegurar definitivamente si esta aparente mejora es o no estadísticamente significativa".*

Su sugerencia es: *"en lo respecta al proceso de toma de decisiones, es posible que un método que combine las predicciones de los modelos individuales podría resultar más adecuado en el problema que estamos analizando,* (Olmeda y Fernández, 1997, Kumar y Olmeda, 1999).

Este interesante artículo abunda en la comparación de técnicas desde el punto de vista de la clasificación, y, a pesar de no considerar los requerimientos que han de verificar los modelos desde el punto de vista IRB de Basilea II, trata dos cuestiones de máximo interés para nuestra Tesis Doctoral. En primer lugar, *la capacidad de predicción de los modelos* respecto de lo que concluyen que los modelos no paramétricos no dominan de forma sistemática a los paramétricos, en contra de lo que suele ser habitual en la literatura, y, en segundo lugar, *sugieren la utilización de un método que combine las predicciones de los modelos individuales*. Entendemos que su sugerencia de *combinar las predicciones de los modelos individuales* puede venir motivada por la percepción de que ninguna de las técnicas individuales resuelve bien el problema de asignación de los demandantes de tarjetas de crédito en una las dos clase default o no default; de ser así esto encaja con las percepciones que motivan esta Tesis Doctoral respecto no sólo de las técnicas que estos

autores comparan sino también de todas las que nosotros hemos revisado en la subsección 1.3.1.

Su sugerencia de considerar un método que combine las predicciones de los modelos individuales es inaceptable desde el punto de vista de Basilea II, por la dificultad que entraña la interpretación de los resultados.

1.3.2.3 Härdle et al. (2005, 2007). *Predicting Bankruptcy with Support Vector Machines. Statistical Tools for Finance and Insurance.*

Decíamos en la subsección 1.3.1 que la mayor debilidad de las técnicas de vector soporte, SV y SVM, consiste en que no proporcionan la probabilidad de default. A pesar de que las aplicaciones de SVM al análisis de la credit scoring no es muy abundante en la literatura financiera, en los primeros articulo debidos a Härdle et al. (2005) y Härdle et al. (2007) y en uno de reciente aparición Härdle et al. (2011), se defiende que, en comparación con LDA y LOGIT, SVM tiene un rendimiento superior pronosticando la probabilidad de default y puntuando el crédito. Su afirmación referida a la probabilidad de default se debe al hecho de que estos autores han intentado transformar en funciones monótonas decrecientes de la probabilidad de default la función de puntuación proporcionada por SVM,

$$S(X) = \beta_0 + \sum_{r=1}^{N} \beta_r K(X, x_r) \tag{1.2}$$

donde N es el número de acreditados y las funciones de base vienen dadas por $h_r(X) = K(X, x_r) = \left(K(x_1, x_r), ..., K(x_i, x_r), ..., K(x_N, x_r)\right)$, siendo $K(\bullet,\bullet)$ una función núcleo.

Estos autores a partir de la función $S(X)$ resultante de la aplicación de SVM sobre la muestra de entrenamiento establecen un ranking de acreditados según el valor de $S(X)$, z_i igual a la posición del acreditado con observación x_i dentro del total de acreditados ordenados según el valor de $S(X)$. A continuación utilizan una técnica de suavizado estándar para una evaluación preliminar de las probabilidades de default para todos los acreditados de la muestra de entrenamiento:

$$\widetilde{PD}(Z) = \frac{\sum_{i=1}^{N} w(z - z_i) I(y_i = 1)}{\sum_{i=1}^{N} w(z - z_i)} \tag{1.3}$$

dónde $w(z-z_i)=\exp\left\{(z-z_i)^2/2h^2\right\}$. Es decir, utilizando el rango de los acreditados en lugar de las puntuaciones $S(X)$, se obtiene un suavizador k_NN con pesos Gaussianos

$$\frac{w(z-z_i)}{\sum_{i=1}^{N} w(z-z_i)}\ .$$

Las probabilidades de default preliminares obtenidas en (1.3) no son necesariamente funciones monótonas de la función de calificación. La monotonización se consigue en un segundo paso usando el algoritmo Pool Adjacent Violator (PAV), (Barlow et al., 1972).
Como resultado se obtienen las probabilidades de default monotonizadas para cada observación de la muestra de entrenamiento.

El artificioso modo de calcular las probabilidades de default iniciado por Härdle et al. (2005) no ha cuajado en el mundo de la puntuación de crédito y hasta donde se nos alcanza tan sólo en un artículo, debido a Chen et al. (2006), se ha utilizado esta técnica, (el articulo versa sobre la predicción de las probabilidades de default de empresas Alemanas de la base de datos CREDITREFORM entre los años 1996 y 2000.

1.3.2.4 Siddiqi (2006). *Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring.*

Es un modelo divulgado por Siddiqi (2006) basado en un procedimiento muy popular, iniciado por la consultora FAIR ISAAC, a principios de los años 60, que consiste en *el desarrollo del modelo usando una agrupación óptima de atributos de las variables explicativas del riesgo, (tramado automático o supervisado con criterios de riesgos), y utilizando únicamente la regresión logística lineal, LLR, para estimar dicho modelo, para, finalmente, a partir del mismo construir la "herramienta de decisión" conocida como tarjeta de puntuación*. Es lógico que esta metodología se haya popularizado por cuanto, por un lado, es una metodología muy intuitiva y con sólidos fundamentos teóricos y, por otro, se vienen observado adecuados resultados en su aplicación, sobre todo en los sistemas proactivos de calificación de acreditados.

La técnica de Siddiqi (2006) ha sido utilizada por la Confederación Española de Cajas de Ahorros (CECA) en la construcción de Modelos Proactivos de Credit Scoring para algunas Cajas de Ahorros Españolas. Esta metodología, caracterizada por estimar el modelo por la técnica LLR a partir de variables tramadas, tiene al menos dos fortalezas que la hacen "aceptablemente buena", si bien, como contrapunto existen las correspondientes debilidades.

Por un lado, la agrupación óptima de los atributos de una característica X_j en tramos disjuntos, $\{R_{j1},...,R_{jK_j}\}$, que cubren exhaustivamente su rango, junto a la asignación a cada acreditado del *peso de la evidencia del tramo al que pertenece el acreditado en la característica considerada,* $WOE(x_{ij}) = Log\left(\frac{\text{Tasa de Buenos en R}_{jk}}{\text{Tasa de Malos en R}_{jk}}\right)$, conlleva que las nuevas variables adquieran importantes y deseables propiedades, tanto desde la óptica estadística como desde la óptica del riesgo de crédito. Con respecto a las propiedades estadísticas, la agrupación óptima ofrece una vía muy sencilla para resolver problemas asociados a los datos tales como d*atos faltantes,* v*alores extremos y clases raras*, (Siddiqi, 2006). Además, la agrupación óptima puede ser controlada, es decir, en vez de tramos automáticos pueden hacerse tramos diseñados con criterios de riesgos, lo que permite *controlar el desarrollo del proceso.*

Por otro lado, el uso de la regresión logística en esta metodología conlleva que se fije la estructura formal de la relación entre las variables explicativas del riesgo y la probabilidad de default en un modelo de probabilidad generalizado con la función de enlace logística,

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{j=0}^{p}\beta_j h_j\left(X\right)=\beta_0+\sum_{j=0}^{p}\beta_j WOE\left(X_j\right) \qquad (1.4)$$

Por lo que el modelo encaja bien con los requerimientos de Basilea II, en cuanto a conseguir un modelo parsimonioso, no sobreajustado y, por tanto, adecuadamente generalizable, y, además, muy explicativo, por cuanto, el peso de las variables viene dado por los coeficientes de la combinación lineal.

La mayor debilidad que presenta el tramado de variables es que al sustituir la variable original por la correspondiente función de base o variable tramada es necesario aceptar un relativo coste de posible pérdida de información y la renuncia a matices que en algunos casos podrían revelarse como importantes para la eficacia del modelo, lo que no parece razonable en variables para las que el tramado no sea estrictamente necesario.

Otra importante debilidad del método, a pesar de que Siddiqi (2006), destaca entre sus fortalezas que "*el tramado de las variables permite modelar dependencias no lineales a través de modelos lineales*", la constituye el hecho de que no se consideran las distintas formas de no linealidad que pueden presentar

las variables consideradas con información relevante para la construcción del modelo, sino solamente aquella que el tramado es capaz de transformar.

1.3.2.5 Liu y Cela (2007). *Improving Credit Scoring by Generalized Additive Model. Data Mining and Predictive Modeling.*

Liu y Cela (2007) comparan los *resultados de un modelo GAM, concretamente un modelo LPALM, con el modelo LLR, ambos modelos ajustados sobre los mismos datos del banco francés utilizados por* Müller y Härdle (2003). *El modelo, semiparamétrico, consta de una componente paramétrica lineal de 20 variables que se habían revelado significativamente lineales en el modelo LLR*

$$\beta_0 + \beta_2 X_2 + \beta_6 X_6 + \sum_{j=8}^{24} \beta_j X_j \tag{1.5}$$

y una componente aditiva no lineal de 3 variables que se habían revelado como no lineales mediante la *suma acumulada de la residuos* y el test del supremo de Kolmogorov, (Lin, et al., 2002),

$$h_4(X) + h_5(X) + h_7(X) \tag{1.6}$$

Una diferencia con el trabajo de Müller y Härdle (2003) es que Liu y Cela (2007) *para especificar las variables no lineales utilizan como funciones base splines de suavizado cúbicos,* (Wahba, 1990, Green y Silverman, 1994).

El modelo utilizado por Liu y Cela (2007) es también un modelo logístico por expansión lineal de funciones de base, donde las funciones de base de la componente no lineal son splines cúbicos naturales. El modelo se expresa entonces

$$\text{logit}\left(P(Y=1/\ X=x)\right) = \sum_{j=p_1+1}^{p} \beta_{j(N+1)} + \sum_{j=1}^{p_1} \beta_j X_j + \sum_{j=p_1+1}^{p} \beta_{j(N+2)} X_j + \sum_{j=p_1+1}^{p} \left(\sum_{k=1}^{N} \beta_{jk} \left| X_j - \xi_k \right|^3 \right) \tag{1.7}$$

Si bien es cierto que utilizar splines cúbicos naturales resuelve el problema de la elección de la localización de los nodos, puesto que fijan los nodos en los N puntos $\{x_i\}_{i=1,\ldots,N}$, no es menos cierto que esto puede conllevar problemas de sobreajuste, debido al elevado número de nudos, que no siempre se resuelven satisfactoriamente aplicando una penalización "ondulante" a la función objetivo del ajuste.

Por otro lado, el hecho de que el procedimiento GAM de SAS® V9.2, utilizado por los autores para ajustar el modelo, calcule el error empírico con la función de pérdida

cuadrática en lugar de la logística ignora el hecho de que la distribución del estado de default Y es Bernouilli de parámetro p .

El primer problema puede suponer que el modelo generalice pobremente, con el consabido riesgo de admisión de solicitudes de crédito con alto riesgo de default, lo que incide en los resultados de la Entidad Financiera, y el segundo provoca que el modelo no sea adecuado para estimar la probabilidad de default, lo que afectará a los requerimientos de capital, según Basilea II y III. Además el problema de la interpretabilidad no se resuelve de forma satisfactoria, como ya había ocurrido con la aportación de Müller y Härdle (2003), no hay más que observar detenidamente la expresión (1.7) y pensar en explicarle a un cliente como afecta cada valor de las variables observadas sobre la puntuación que se le asigna, a parte de la dificultad computacional cuando el modelo se entrene con miles de de acreditados. Por esta razón, es habitual que los paquetes de software solo proporcionen para los términos ajustados por splines cúbicos naturales los valores pronosticados.

Por otro lado, los propios autores reconocen que si bien todos los indicadores y test estadísticos de ajuste y poder discriminante están a favor de LPALM, el desempeño del modelo puede ser muy sensible a la muestra de datos utilizados para entrenar el modelo. Como ellos mismos señalan en la introducción de su artículo, la principal crítica a los modelos LPALM es que se ajustan más a los datos, como ocurre para cualquier Modelo Aditivo Generalizado, lo que conlleva un mayor fracaso en la generalización de las predicciones.

Como era de esperar, la tasa de error de clasificación de la regresión logística aditiva parcialmente lineal resultó algo menor, en 0,28, que la de la regresión logística lineal para los datos de entrenamiento. Pero, a pesar de que LPALM supera el rendimiento de la regresión logística en la tasa de error de clasificación en 0.60 para la muestra test, tabla 1.1, la alta posibilidad de mala generalización y la mayor dificultad en la interpretación creemos que no compensa, desde el enfoque IRB de Basilea II, la introducción de los Splines Cúbicos Naturales como funciones base para las variables no lineales.

Tabla 1.1.- Comparación entre las tasas de clasificación errónea de LLR y LPALM, (Liu y Cela, 2007).

Tasa de errores de clasificación			
Conjunto de Entrenamiento		Conjunto Test	
LLR	LPALM	LLR	LPALM
8,42%	8,14%	9,80%	9,20%

Además, los propios autores admiten en sus conclusiones que a pesar de que LPALM supera a la LLR en al menos dos aspectos, (relaja el supuesto de linealidad entre las variables explicativas y la respuesta y, mediante la incorporación de los efectos no lineales, ayuda a descubrir el patrón oculto que subyace detrás de las variables explicativas, lo que resulta en un mejor rendimiento en la predicción), "*la no linealidad no es intuitiva para los administradores de empresas*". Como alternativa, aconsejan utilizar el tramado de variables, (Siddiqi, 2006), para aproximar el efecto no lineal una vez descubierto por LPALM, en este sentido se suman al planteamiento de Müller et al. (2003) de considerar esta técnica como un método exploratorio y especificar la no linealidad con un método que permita una interpretación más sencilla de la importancia relativa de cada factor de riesgo en la relación con el estado de default.

Por ejemplo, un efecto no lineal en LPALM se puede categorizar incorporando conocimiento de riesgo de crédito y a continuación una vez sustituidas estas variables por las nuevas categóricas ajustar un modelo LLR. Puntualizan que "al hacerlo así, la interpretabilidad puede ser mejorada, si bien pagando el precio de una menor capacidad explicativa". Es decir, proponen utilizar el procedimiento de FAIR ISAAC y SIDDIQI pero sólo para aquellas variables para las que la técnica LPALM descubra no linealidad subyacente. Y finalizan su artículo sentenciando que "*el equilibrio entre la flexibilidad y la interpretabilidad de los modelos* LPALM *ofrece una alternativa prometedora al modelo lineal generalizado en la mayoría de las situaciones en que este sea aplicable*".

1.3.2.6 Hervás-Martínez et al. (2007). *Aprendizaje mediante la hibridación de técnicas heurísticas y estadísticas de optimización en regresión logística binaria.*

Hervás-Martinez et al. (2007), para salvar el efecto de la no linealidad de las variables explicativas y reducir el error en los bordes del dominio del conjunto de datos, proponen un modelo de regresión logística basado en la hibridación de un modelo estándar de

regresión logística y un modelo de red neuronal de unidades producto, PUNN, introduciendo el término no lineal en el modelo mediante funciones de las variables explicativas obtenidas mediante la multiplicación de potencias de las covariables iniciales, las cuales expresan las posibles fuertes interconexiones existentes entre las covariables. El modelo que proponen viene dado por la expresión

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r\left(X\right)=\beta_0+\sum_{r=1}^{p}\beta_r X_r+\sum_{r=p+1}^{q}\beta_r\prod_{j=1}^{p}X^{w_{jr}} \qquad (1.8)$$

Estas funciones complican a priori el modelo, cuando, en general, no existe ninguna razón para hacerlo más complejo de lo estrictamente necesario, dilema de Occam, como supone el hecho de añadir a la componente no lineal transformaciones de variables que se manifiesten claramente lineales. Del mismo modo que no hay ninguna razón para especificar la no linealidad de forma homogénea para todas las variables no lineales, como es el caso del estudio de Hervás-Martinez et al. (2007), donde la no linealidad de todas las variables se pretenda captar de forma homogénea a través de *unidades producto* UP, independientemente de que para algunas variables concretas esta pudiera ser la especificación adecuada.

Este tipo de modelos y la metodología seguida en su aplicación no se ajusta totalmente a los requerimientos de Basilea II. Si bien es posible obtener los tres instrumentos básicos requeridos a los modelos de *puntuación de acreditados*, probabilidad de default, función de calificación de acreditados y clasificador de nuevas solicitudes de crédito, no es posible explicar de manera satisfactoria, a través de su estructura funcional, la contribución de todas y cada una de las variables explicativas.

Por otra parte, para valorar el rendimiento del modelo LRLPU frente a otros esquemas de aprendizaje en clasificación los autores utilizan los dos conjuntos de datos del proyecto Statlog, Blake y Merz (1998), a los que nos referiremos en el apartado a) de la sub-sección 1.4.1, para clasificar a los acreditados de dos bancos en las poblaciones de default y no default. Para validar el rendimiento de su modelo LRLPU en cada uno de los dos conjuntos de datos utilizan la *proporción correcta de clasificación* (CCR) para el conjunto de generalización, que se define como el *porcentaje de patrones de los datos correctamente clasificados,* pero como explicamos en el capítulo 4 el error total de clasificación no es un criterio apropiado para medir el rendimiento de un modelo de clasificación de acreditados, aunque desgraciadamente se utiliza, por simplicidad y no con poca frecuencia, para comparar diferentes modelos de clasificación.

1.3.2.7 Liu y Cela. (2009). Generalizations of Generalized Additive Model (GAM): A Case of Credit Risk Modeling.

Liu y Cela (2009) presentaron en el SAS Global Forum' 2009, un trabajo, continuación de su artículo del año 2007, al que nos hemos referido en el apartado 1.3.2.5, en el que, reconociendo el papel de líder que juega la técnica de Regresión Logística Lineal en los modelos de credit scoring debido a su simplicidad, ponderan el prometedor futuro de GAM, (entre ellos los LPALM), en la construcción de modelos de riesgos de crédito por su combinación de flexibilidad e interpretabilidad. En la introducción de su artículo resaltan textualmente, "aunque conceptualmente atractivos los GAM no están exentos de críticas, su estructura funcional parcialmente no paramétrica hace difícil la calificación de nuevos solicitantes de crédito directamente desde los datos, lo que dificulta que se utilicen en entorno de negocios y que se implementen en escenarios de producción". A pesar de ello, resaltan la superioridad de los GAM sobre la LLR, *sólo cuando el objetivo es la clasificación.*

En su artículo, Liu y Cela (2009), presentan una combinación de los modelos GAM con los Árboles de Regresión y Clasificación, TREE, y con los Splines de Regresión Adaptativos Multivariantes, MARS, para mejorar la interpretabilidad de los GAM pretendiendo que estas dos nuevas técnicas de construcción de modelos se derrumben dentro del armazón de los Modelos Lineales Generales, GLM, que, según sus propias palabras, "son más familiares a los constructores de modelos" de calificación de acreditados.

La solución que aportan Liu y Cela (2009) al problema de estimar los parámetros de cada término no lineal deducidos del modelo GAM consiste en utilizar aproximaciones constantes y lineales a trozos. Si bien hay muchas maneras de llegar a tales aproximaciones, sea en percentiles o por la experiencia, proponen un modelo basado en la utilización de la técnica TREE, para abordar el problema con aproximaciones constantes a trozos, y en la utilización de la técnica MARS que es capaz de abordar el mismo problema con aproximaciones lineales a trozos.

Utilizan la técnica TREE a través del algoritmo CART, (Breiman et al., 1984), con una sola variable independiente y una variable dependiente, que son el correspondiente término no lineal y estado de default respectivamente. Después del desarrollo de CART se utiliza la *aproximación constante a trozos resultante como una variable categórica*. Una vez obtenidas estas variables LPALM *se desploma para convertirse en* LLR.

Para los datos de entrenamiento, los mismos pertenecientes a un banco francés utilizados por Müller y Härdle (2003), el modelo híbrido considerado por Liu y Cela se comporta de manera similar a LPALM. Sin embargo, este generaliza mejor que LLR y LPALM para los datos de validación con un área bajo la curva ROC AUC=0.77. Los autores entienden que, dado que una ventaja de CART es la resistencia a los valores extremos, esta mejora marginal podría resultar en una reducción del sobreajuste.

Los autores, del mismo modo en que han utilizado TREE, utilizan la técnica MARS. Después del desarrollo de MARS utilizan la *aproximación lineal a trozos resultante como una variable numérica* y a continuación reemplazan con las nuevas variables obtenidas los términos no lineales de LPALM para, finalmente, estimar el modelo con la inclusión de las nuevas variables.

Con respecto a la capacidad explicativa, no existe una notable diferencia en comparación con LLR y el LPALM para los datos de entrenamiento. Este nuevo modelo no parece generalizar bien para la validación de datos con AUC = 0,75, frente a 0,78, para los datos de entrenamiento. En opinión de los autores, este bajo rendimiento es probable que sea debido al sobreajuste, inconveniente bien conocido de MARS.

Liu y Cela (2009) acaban su artículo concluyendo: "*Nuestros resultados muestran que el modelo híbrido que combina la flexibilidad de los* LPALM *y la resistencia al sobreajuste de CART es el que produce el mejor resultado*".

Sin embargo en los artículos de Liu y Cela (2007, 2009) existe a nuestro juicio un serio inconveniente, se especifica la no linealidad para las variables X5 y X7 del mismo modo. En el primero, 2007, por splines cúbicos naturales y en el segundo, 2009, según dos versiones, por funciones constantes a trozos obtenidas de CART unidimensional para la variable en cuestión o por funciones lineales a trozos obtenidas por aplicación de MARS en la misma forma. Pero nada hay que asegure que la no linealidad de todas las variables no lineales deba especificarse de la misma forma, pues la no linealidad puede manifestarse de infinitas formas distintas, excepto que se comprueba que efectivamente así es, por nuestra parte, creemos que la confirmación de tal extremo sería un hecho excepcional.

1.3.3 Síntesis de los apartados 1.3.1 y 1.3.2.

1.- Desde comienzos de la última década del siglo pasado vienen proliferando aplicaciones de la mayoría de las técnicas mostradas en la subsección 1.3.1 en las más diversas aéreas de la actividad humana, Medicina, Microbiología, Genética, Ecología, Agricultura, Sociología, Economía y Empresa y un sinfín de otras, y seguramente en la mayoría con elevado éxito, lo que no contradice el hecho de que muchas presenten importantes debilidades en su aplicación al *credit scoring* con respecto al enfoque IRB de Basilea II.

2.- Las fortalezas y debilidades de un modelo dependen del entorno y condicionantes en los que el modelo debe captar la relación de dependencia entre la variable respuesta y las variables explicativas, así, por ejemplo, en la credit scoring los condicionantes, a parte de los propios de la construcción de modelos estadísticos, vienen determinados por los requerimientos de Basilea II, por lo que es imprescindible que el modelo proporcione la probabilidad de default, requerimiento que puede perfectamente no ser necesario en otra actividad y en este caso la ausencia de la probabilidad condicionada a la pertenencia a una clase no puede considerarse una debilidad del modelo. Es decir, no es la técnica la que presenta las debilidades sino su aplicación indebida, por ejemplo, no es la LLR la que presenta la debilidad de no contemplar la no linealidad, sino en todo caso será la utilización de una técnica diseñada sólo para variables lineales en situaciones donde ese no sea el caso.

1.4 MOTIVACIÓN.

Nuestra motivación se basa en el hecho de que existe la necesidad de desarrollar mejores modelos de *credit scoring*, debido, por un lado, a la necesidad de calcular la probabilidad de default para fijar el capital económico requerido según Basilea III, y, por otro, al gran ahorro que se puede obtener como consecuencia de utilizar modelos que funcionen bien y los altos costos involucrados en la clasificación errónea de solicitantes de créditos. El desarrollo de un modelo de *credit scoring* consta, en general, de 8 fases principales, a saber:

1. Preanálisis de los Datos Poblacionales;
2. Generación de las Muestras;
3. Selección de Variables;
4. Especificación del Modelo;
5. Estimación y ajuste del Modelo;
6. Evaluación y Selección del Modelo

7. Validación del Poder Discriminante del Modelo
8. Calibración del Modelo.

Si bien todos estas fases son fundamentales en el desarrollo de modelos de *credit scoring*, las fases de *selección de variables*, *especificación del modelo* y *desarrollo del mismo* son de particular interés para esta investigación. Creemos que en estas tres fases existe mucho margen para mejorar el rendimiento de modelos de *credit scoring*.

1.4.1 Fase de Selección de Variables.

En lo que se refiere a la fase de la selección de variables, se ha encontrado en la literatura que el uso de las variables adecuadas, (relación con el estado de default y capacidad explicativa y predictiva), podría conducir a mejoras significativas en el rendimiento de los modelos de credit scoring. Desgraciadamente, son pocos los estudios que han examinado el uso de técnicas de selección de variables en *credit scoring*. Creemos que esto es así por las siguientes razones:

a) La mayoría de los estudios sobre *credit scoring* hacen uso de conjuntos de datos de referencia que contienen unas pocas variables ya depuradas. Los dos conjuntos más utilizados son los proporcionados por un banco Australiano y un banco Alemán, ambos están públicamente disponibles en la Universidad Irvine de California (BLAKE y MERZ, 1998). Estos datos han sido usados en el proyecto Statlog, diseñado para contrastar y evaluar la tasa de clasificación y capacidad de predicción de algoritmos de aprendizaje estadístico y lógico en orden a determinar los más eficaces para la industria de Sistemas de Calificación de Riesgo de Crédito, MICHIE et al. (1994). Un tercer conjunto de datos se puede obtener de THOMAS et al. (2002) y un cuarto provienen de un estudio previo realizado por QUINLAN (1987), 600 individuos y 14 características de demandantes de una tarjeta de crédito. Los cuatro conjuntos de datos han sido usados para comparar el rendimiento de distintos modelos de *credit scoring*, pero en los cuatro el número de variables es pequeño, menor de 30, por lo que los autores de los distintos estudios no acometen el proceso de selección de variables.

b) A pesar de que obtener conjuntos de datos reales con muchas variables es usualmente difícil, nosotros afortunadamente, por cortesía de CajaEspaña, contamos con un conjunto de datos reales consistente en 93.761 acreditados de la Caja sobre los que se han observado 63 variables en dos períodos de tiempo distintos, el primero el 30 de noviembre de 2007 y el segundo el 30 de noviembre de 2008. La información de detalle

proporcionada por CajaEspaña, relevante para el desarrollo y explotación de los modelos de credit scoring proactivos, viene estructurada en una serie de "visiones" parciales, cada una de las cuales proporciona una perspectiva concreta de la relación del cliente con la Entidad a través de un determinado tipo de información a una fecha determinada. En concreto estas visiones son: *Visión General del Cliente,* (Información socio – demográfica, Relación de actividad del cliente con la Entidad, Relación activo – pasivo); *Visión de Pasivo,* (Pasivo a la vista, Pasivo no vista, Total pasivo); *Visión de Activo,* (Ahorro en descubierto, Préstamos de finalidad particular con garantía personal, Préstamos de finalidad particular con garantía hipotecaria. Total préstamos de finalidad particular, Cirbe, Total activo); *Visión de Servicios*; *Visión de Incidencias; Visión de Incumplimiento.*

En este caso la fase de selección de variables es un requisito imprescindible para el desarrollo de un modelo adecuado, primero porque seguramente entre algunas de las variables haya colinealidad, segundo, porque es posible que no todas las variables expliquen adecuadamente el comportamiento del acreditado frente al default, por lo que será necesario seleccionar las más significativas, tercero, porque ante dos variables de explicación similar del estado de default y relacionadas, siempre será preferible aquella cuya especificación en la estructura del modelo sea menos compleja, y, cuarto, porque según la opinión de los expertos en riesgo de crédito es deseable que, en la medida de lo posible, estén involucradas en el desarrollo de los modelos variables de todas las visiones parciales del cliente, lo que implica un detallado análisis de la importancia que, en relación con la explicación del default, tienen las variables dentro de la visión a la que pertenecen.

Todo ello motiva en *nuestra investigación, en primer lugar, la preparación de la información mediante la aplicación de varias técnicas estadísticas, - que han demostrado tener éxito en muchos otros campos - en el contexto de la puntuación de acreditados, adaptadas en algunos aspectos concretos a este contexto, y, en segundo lugar, la selección de las variables de riesgo de crédito más relevantes y adecuadas que deberán intervenir en los modelos con el fin de mejorar su rendimiento.*

1.4.2 Fase de Especificación del Modelo.

En la sección 1.3 hemos analizados brevemente los modelos que usualmente se utilizan para representar la relación de dependencia entre la variable estado de default y las variables relevantes de riesgo de crédito. Hemos destacado las características más sobresalientes, sus fortalezas y debilidades, sobre todo en

relación con la estimación de la probabilidad de default desde la óptica de los acuerdos de Basilea II, de un buen número de técnicas que, sin constituir un conjunto exhaustivo, creemos que representan adecuadamente todo casi todo el espectro que con mayor o menor acierto se utilizan para construir modelos de probabilidad de default.

Ninguna de las técnicas analizadas en la referida sección, desde las que producen los modelos más rugosos y sobreajustados, k_NN y NWE, hasta los modelos más parsimoniosos, amigables y fáciles de interpretar, los modelos lineales, LOGIT y, PROBIT, resuelven de forma satisfactoria los tres problemas binarios multivariantes, *predicción, calificación* y *clasificación*, que conlleva el credit scoring desde la óptica de Basilea II. Este panorama motiva la *búsqueda de técnicas y metodologías que resuelvan las debilidades presentadas por las técnicas analizadas en 1.3.1.*

Por otro lado se observa, en primer lugar, que, con excepción de los estimadores no paramétricos de la probabilidad de default, *k-vecinos más próximos*, K_NN, los *estimadores de la densidad por funciones núcleo*, y el *estimador de Nadaraya_Watson*, las estructuras formales de todos los modelos a los que nos hemos referido son susceptibles de expresarse como expansiones lineales de funciones de base, tarea que con fines unificadores hemos realizado y mostramos en el capítulo 5. Esto nos motiva para proponer que la estructura formal del modelo para recoger la no linealidad de aquellas variables necesarias que lo requieran se exprese según la sugerencia de Hastie y Tibshirani (1996), es decir, como *expansión lineal de funciones de base*.

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r\left(X\right) \tag{1.10}$$

Y en segundo lugar, se ha de destacar, que también a excepción de los estimadores no paramétricos de la probabilidad de default, los métodos que hemos analizado en la sección 1.3 atacan el problema de encontrar la estructura formal más adecuada para el modelo desde distintas estrategias respecto de construir la expansión lineal de funciones de base, pero la mayoría de ellas tienen en común que las funciones de base utilizadas poseen todas la misma estructura. Sin embargo, no hay ninguna razón para el supuesto de homogeneidad en las funciones de base. Es natural pensar que algunas variables provocan efectos lineales sobre la expansión lineal y otras efectos no lineales, y debemos tener presente que *la no linealidad puede manifestarse de infinitas formas en los espacios de representación más usuales, como son los espacios de Hilbert*. Esto motiva

buscar para cada variable con linealidad no significativa y con no linealidad confirmada por las técnicas apropiadas para ello las funciones de base específicas más adecuadas, de este modo *la estructura formal de nuestro modelo de mejora deberá consistir en una expansión lineal híbrida de funciones de base*, en el sentido de que tendrán distinta naturaleza.

Todo lo anterior motiva otra parte importante de esta Tesis Doctoral, pues nos sitúa ante la necesidad de formalizar una clase de modelos, más adecuados que los actuales, donde se combinan ideas de Modelos Logísticos Lineales, LLM, (Mejor aproximación a la relación de dependencia teórica entre variables respuesta y variables explicativas y alto grado de interpretabilidad), Modelos Logísticos Parcialmente Lineales, MLPL, (que conservando la interpretabilidad introducen la flexibilidad necesaria para contemplar la no linealidad), y de modelos logísticos expansiones lineales de funciones de base específicas para cada variable X, es decir, expansiones lineales híbridas cuya característica más destacable es que una vez que las funciones de base, nuevas variables, han sido determinadas, los modelos son lineales en estas nuevas variables, resultando modelos logísticos lineales a los que nos referiremos como *Modelos Logísticos Lineales Híbridos*, **HLLM**.

Se motiva de este modo la formalización de la estructura funcional de los modelos HLLM así como la metodología para construirla en la práctica y el análisis de algunas funciones de base apropiadas para describir la no linealidad en nuestra investigación.

1.4.3 Fase de Desarrollo del Modelo.

Una vez que se ha especificado la estructura formal del modelo *desde la perspectiva IRB de Basilea II,* la siguiente tarea consiste en acometer la fase de *desarrollo del modelo, en sus aspectos teórico y práctico.* Esta fase está relacionada con la implementación de una técnica, y su correspondiente metodología, que resuelvan satisfactoriamente los tres problemas binarios multivariantes que conlleva el credit scoring desde dicha perspectiva, a los que nos hemos referido en la sección 1.1, *predicción, calificación* y *clasificación*. El enfoque principal de nuestra investigación consiste precisamente en el desarrollo de un modelo con las características anteriores y con entrenamiento, evaluación, validación y calibración con datos reales; normalmente muestras de la población de acreditados.

1.4.3.1 Aspectos Teóricos del Desarrollo del Modelo.

Tal como hemos expuesto en la sección 1.3.2, estamos motivados a utilizar *Modelos Logísticos Lineales Híbridos*, HLLM, que son una subfamilia de *Modelos Logísticos Lineales de Probabilidad por expansiones lineales de funciones de base.* Pretendemos establecer una metodología común basada en tres hipótesis muy generales para formalizar la estructura funcional de estos modelos generales y su estimación, basada en el principio de inducción, pues, hasta donde se nos alcanza, no conocemos autores que hayan llegado a formularlos en todos sus aspectos de forma integral.

Queremos que el modelo exprese la relación existente entre una conveniente transformación de la probabilidad de default y la función de calificación de acreditados, lo que motiva modelos de probabilidad generalizados cuya estructura funcional quedará fijada a través de la correspondiente expansión por funciones de base $S(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)$, y sólo resta establecer un método para estimarla.

1.4.3.2 Desarrollo de un Modelo de Credit Scoring HLLM Proactivo.

Como es habitual la propuesta de un modelo de puntuación de acreditados junto con una metodología concreta de aplicación ha de conllevar necesariamente la verificación de sus potenciales cualidades y la comparación con otros modelos ya existentes que lo justifiquen como aportación a la caja de herramientas de los sistemas de calificación del riesgo de crédito, hecho que motiva que en nuestra investigación se incluya el desarrollo de un Modelo Logístico Lineal Híbrido por Expansiones Lineales de Funciones de Base, HLLM, Proactivo de Credit scoring.

Sin duda alguna, una de las fases más crítica de cualquier tesis de esta naturaleza la constituye el proceso de extracción de la información, dadas las características que se requiere presente la misma. A pesar, de que esta información está disponible en las Entidades Financieras, bajo el pretexto de su carácter confidencial no suele encontrarse a disposición de los investigadores, para, por un lado, proceder a la selección de las variables que verdaderamente condicionan el comportamiento de los acreditados frente al default o la relación de las intenciones de los solicitantes de crédito respecto del cumplimiento de sus obligaciones de pago, por otro lado, para realizar el necesario contraste de los desarrollos teóricos, sobre todo para medir los rendimientos alcanzados por las distintas herramientas estadísticas aportadas, y, por último, como elementos de

aprendizaje, sobre todo en la construcción de técnicas basadas en el *principio de dejar hablar únicamente a los datos, es decir, en los modelos no paramétricos.*

Las dificultades de acceso a los datos está provocando que la mayoría de las aplicaciones aparecidas en la literatura sobre los modelos de credit scoring, utilicen los conjuntos de datos a los que nos hemos referido en la subsección 1.4.1 apartado a). Si tenemos en cuenta que crece cada día, y sobre todo en la concepción IRB de Basilea II, el número de de expertos en riesgo de crédito que apuestan por modelos de puntuación de crédito hechos a la medida de la realidad de cada Entidad Financiera, es evidente que la situación es claramente preocupante, salvo que cada Entidad o grupo asociado de Entidades Financieras cree costosísimas unidades de investigación. Es evidente que son necesarios adecuados convenios, donde se establezca de forma rigurosa y fiable la protección de la información, al uso de los que vienen fructificando en el desarrollo de la investigación en otras áreas como, por ejemplo, la medicina.

En esta investigación se ha contado, lo que constituye una de sus fortalezas, con la inestimable colaboración de una Entidad Financiera Española que ha puesto a nuestra disposición información estadística de su magnifico *datamar*t de riesgos, con los preceptivos filtros para preservar su debida confidencialidad, y la larga experiencia tanto de sus especialistas informáticos en procesos de extracción de la información como la de sus expertos analistas en detección de la información relevante para el riesgo de crédito. En concreto, se dispone de datos relativos al *segmento de acreditados con préstamos a particulares,* con fechas de visión a 30 de noviembre de 2007 y a 30 de noviembre de 2008.

Se cuenta con 63 variables observadas sobre el segmento de 72.062 acreditados con préstamos a particulares, relativas a la relación global del cliente con la Entidad, a la relación a través de productos de pasivo y de riesgo, a su relación a través del consumo de servicios ofrecidos por la Entidad, a su vinculación a través de la domiciliación de la nómina y/o recibos y al comportamiento del cliente con la Entidad en los productos de riesgo contratados, así como a la calificación, según la definición de incumplimiento del Nuevo Acuerdo de Capital de Basilea II, que la Entidad asigna al cliente de acuerdo con su comportamiento.

Contar con la información que hemos descrito en el párrafo anterior es, sin duda alguna, una importante motivación para *desarrollar de forma detallada en todas sus fases el modelo de credit scoring que proponemos,* desarrollo que no es habitual en la literatura

especializada, y compararlo con otros modelos alternativos utilizados en la industria del credit scoring.

1.5 IMPORTANCIA DE LA INVESTIGACION.

Esta investigación es importante por, al menos, tres razones principales:

1. En primer lugar, la familia de modelos que proponemos, HLLM, extensión natural de los modelos lineales, con la capacidad para recoger la no linealidad y conservando las cualidades más robustas en relación con Basilea II, pueden constituirse en eficaces herramientas para los sistemas de calificación de acreditados de las Entidades Financieras. Estos modelos posibilitan que, con costes computacionales asumibles, estas Entidades puedan disponer de *tarjetas de credit scoring* (*credit scorecard*) desde la óptica IRB de Basilea II sin renunciar a utilizar modelos con estructuras sencillas, (dilema de Occam), pero en las que sea posible especificar la complejidad con que puede manifestarse la no linealidad de algunas de las variables importantes en la explicación del riesgo de crédito. Por tanto, los modelos HLLM representan un complemento necesario a los modelos existentes en la literatura de los sistemas de calificación del crédito.

2.- En segundo lugar, formalizamos los modelos HLLM, desde la formulación general de los *Modelos Logísticos de Probabilidad por expansiones lineales de funciones de base,* desarrollo c*uya importancia* estriba, en primer lugar, en el hecho de que dota de rigor y consistencia la utilización de los modelos HLLM, y, en segundo lugar, en que unifica la formulación de la estimación de los modelos HLLM regularizados y no regularizados.

   Por tanto, el diseño de los modelos HLLM puede ser importante, tanto desde el punto de vista teórico como del práctico, para quienes deseen mejorar los modelos parcialmente lineales de probabilidad generalizados por expansiones lineales de funciones de base utilizando funciones de enlace distintas a la logística o utilizando los resultados de nuestro modelo proactivo como base para futuras mejoras y para la construcción de modelos a la medida en la industria del credit scoring.

3. No es habitual encontrar en la literatura sobre la credit scoring una exposición detalla y exhaustiva, contemplando el *preanálisis de los datos poblacionales, la exploración de los datos de entrenamiento, la generación de las muestras, la selección de variables, la especificación de las funciones de base y la estimación, la selección, la capacidad de generalización, la validación del poder discriminante y la calibración del modelo,* y

todo ello en una situación de riqueza de datos reales proporcionados por una Entidad Financiera, cuyo proceso de obtención ha estado desde su inicio orientado a modelos proactivos, desde la óptica IRB de Basilea II.

1.6 CONTRIBUCIONES A LOS MODELOS DE CREDIT SCORING DESDE LA ÓPTICA IRB DE BASILEA II.

En esta Tesis Doctoral, en la investigación sobre el *desarrollo de mejores modelos proactivos de credit scoring desde la óptica IRB de Basilea II,* se hacen varias contribuciones en tres áreas distintas:

1.- Novedosa visión general unificada de las técnicas más actuales de credit scoring, mostradas en 1.3, formalizando sus estructuras funcionales como expansión de funciones de base.

2.- Formalización de la estructura funcional de los Modelos Logísticos Lineales Híbridos, HLLM, que constituyen una extensión natural de los modelos logísticos lineales a los que se añade la capacidad para recoger la no linealidad, a través de expansiones lineales de funciones de base, sin perder ninguna de las cualidades más importantes en relación con los requerimientos de Basilea II, así como la metodología para su estimación.

3.- Desarrollo de una metodología para el proceso completo de construcción de un modelo proactivo de credit scoring HLLM sobre los datos reales de una Entidad Financiera, desde la óptica IRB de Basilea II.

Con respecto a la primera contribución, **punto 1**, **en la investigación hemos formalizado la estructura funcional de los *Modelos de Probabilidad Generalizados, GPM, por expansiones lineales de funciones de base,***

$$g\left(P\left(Y=1/\,X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(X\right) \tag{1.11}$$

a partir de tres hipótesis muy generales, que nos permiten concebirlos desde una perspectiva unificada. La estructura del modelo quedará perfectamente determinada una vez que se fije la matriz de funciones de base $\boldsymbol{H}(\boldsymbol{X})$.

Como veremos en el capítulo 5, los modelos correspondientes a prácticamente todas las técnicas analizadas en esta Tesis Doctoral, y que representan el amplio espectro de las utilizadas en mayor o menor medida y con mayor o menor éxito en los sistemas de calificación del riesgo de crédito, pueden expresarse según (1.11).

La introducción de un término de regularización en la función objetivo es una forma de solucionar el problema del desconocimiento de la estructura de la función de calificación de acreditados $\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)$, lo que implica otra importante elección, la del *término de regularización* $\lambda \boldsymbol{J}\left(\boldsymbol{\beta}^T \boldsymbol{H}(X)\right)$. Además, con ello resolvemos problemas como el sobreajuste o la no unicidad de la solución y suele dar muy buenos resultados cuando el número de variables es muy elevado en relación con el número de acreditados.

Con respecto al **punto 2**, el foco de atención en esta Tesis Doctoral se centra en modelos con función de enlace logit cuya función de calificación de acreditados es una expansión lineal de funciones de base,

$$\operatorname{logit}\left(P(Y=1/X=x)\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X) \qquad (1.12)$$

Lo que permite contemplar en el modelo la no linealidad sin aumentar significativamente la complejidad de su estructura funcional.

A los modelos con estructura formal (1.12), cuyo método de construcción se enmarca dentro de los *métodos de restricción supervisada,* les llamaremos ***Modelos Logísticos Lineales Híbridos por expansiones lineales de funciones de base,*** **HLLM**. El calificativo "híbrido" se debe a que cada variable no lineal, se puede expresar como una combinación de funciones de base específicas diferentes, como por ejemplo, funciones polinómicas, logarítmicas, pesos de la evidencia obtenidas por un proceso de tramado óptimo de la variable considerada, funciones constantes a trozos resultados de particiones recursivas de Árboles de Clasificación, por ejemplo, obtenidas por el algoritmo CART, funciones lineales a trozos, funciones base de splines, ya sean de regresión, de suavizado o de penalización, funciones sierra obtenidas por el método PPR sobre una sola variable, funciones bisagra obtenidas por el procedimiento MARS, funciones de base radial, RBF, funciones sigmoides, (US) (propias de las redes neuronales), etc..

Respecto de la estimación del modelo HLLM, esta la hemos planteado como el *Problema General de Estimación de los Modelos Logísticos por expansión lineal de funciones de base*, en el que la función objetivo a minimizar se configura como la suma del *termino de ajuste*, media de las discrepancias entre los valores de default observados y los pronosticados por el modelo para los datos de entrenamiento, y del *término de regularización*, *funcional de regularización* que pretende, por un lado, evitar la

complejidad del modelo que conduce al indeseado sobreajuste, y, por otro, solucionar el problema de la posible no unicidad de la solución.

Con respecto al **punto 3**, se realizan tres contribuciones**,** *la primera contribución se refiere a la fase de preanálisis de los datos poblacionales*. Con el fin de eliminar la multicolinealidad, es habitual que en la fase de pre análisis de los datos poblacionales se analice esta a partir de los estadísticos *Inflación de varianza*, VIF, y *Tolerancia*, TOL. **Dado que la inflación de varianza y la tolerancia son muy sensibles a la influencia muestral, en esta Tesis Doctoral utilizamos *técnicas bootstrap* para su estimación, formulando los dos estadísticos como estadísticos de conjunto, *Bagging*, respectivamente**. Ambos estadísticos, VIF_Bag y Tol_Bag, se calculan ajustando el modelo de regresión $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, siendo Y la variables estado de default, por el método de mínimos cuadrados ordinarios, MCO, utilizando B remuestras bootstraping.

La segunda contribución se enmarca dentro de la fase de selección de variables. La estrategia de exploración y selección de variables adoptada en esta Tesis Doctoral se encauzó por derroteros distintos a los habituales.

Cuando se seleccionan automáticamente variables linealmente significativas, entre cuyas interrelaciones existe demasiado "ruido", con técnicas paso a paso basadas en la Regresión Logística Lineal tales como RLL Forward, RLL Backward y RLL Stepwise, las estimaciones de los coeficientes de regresión, sus errores estándar y los intervalos de confianza suelen presentar problema de inestabilidad y sesgo por lo que se hace necesario adoptar estrategias alternativas.

Según refleja la literatura, es habitual que para evitar los problemas anteriores la estrategia de exploración y selección de variables se encamine en la dirección sugerida por Efron y Tibshirani (1993), emprendida por Shtatland et al. (2003) y continuada por Liu y Cela (2007) y Hayden et al. (2009*).* En esta tesis Doctoral adoptamos la estrategia de los autores anteriores pero añadiendo un elemento diferenciador, *fundir en una técnica, la Regresión Logística Lineal Backward con remuestreo Bootstrap, BLLR_Bag, las características especializadas de los métodos de selección automática por regresión logística lineal paso a paso con la capacidad de las técnicas bootstraps para eliminar el sesgo y dotar de estabilidad a los coeficientes de regresión, una vez eliminada la influencia muestral.*

La tercera contribución significativa consiste en la metodología seguida para seleccionar, en la práctica, las funciones de base que configuran la estructura del modelo HLLM. Es decir, la forma en que se construye la componente no lineal del modelo:

1.- En primer lugar se considera un modelo de partida, $M_{inicial}$, que deberá estar formado por todas las variables que muestren una relación de linealidad con el estado de default altamente significativa. Este punto es de necesidad de acuerdo con los requerimientos de Basilea II, pues difícilmente cualquier otra estructura resulta más fácilmente interpretable que la lineal, además de que las variables originales muestran la información en toda su plenitud.

2.- En segundo lugar se hace uso de un *método constructivo* para seleccionar la combinación lineal de funciones de base "más prometedora" para cada variable de un conjunto de candidatas. Para seleccionar las funciones de base más prometedoras se construyen los modelos alternativos respectivos que se ajustan por *Regresión Logística Lineal* y se seleccionan las funciones de base que, con criterios de bondad de ajuste, poder explicativo y discriminante y eficacia clasificadora, en las muestras de entrenamiento y validación, en combinación con los requerimientos de Basilea II y con la política de riesgos de la Entidad Financiera, parezcan más adecuadas.

   Al mismo tiempo se construirán modelos alternativos considerando expansiones lineales por funciones de base que aunque no sean las más prometedoras parezcan en principio adecuadas.

3.- Como resultado del proceso de construcción del apartado 2, se llega a un conjunto de modelos con estructura inicialmente valida pero que, en principio, podrían sobreajustar los datos, no poseer la cualidad de facilidad interpretativa etc. Por lo que es necesario aplicar una técnica de poda para reducir el número de funciones de base.

   Como método de poda hemos utilizado un *Proceso de Selección Limitada de Funciones de Base hacia Atrás*, en concreto un *Ajuste Paso a Paso Hacia Atrás,* hasta conseguir un modelo óptimo desde el punto de vista estadístico y desde el enfoque de los requerimientos de Basilea II.

1.7 ESQUEMA DE LA TESIS DOCTORAL.

El resto de la Tesis Doctoral consta de tres partes. La primera parte consiste en tres capítulos, en el primero, Capítulo 2, *Probabilidad de Default y Modelos de Credit Scoring*, se exponen los conceptos y características básicas de la probabilidad de default y la estimación de esta probabilidad en vecindades locales. En el segundo, Capítulo 3, *Estimación, Evaluación y Selección de Modelos de Probabilidad de Default,* se estudian los aspectos más destacados de la estimación, evaluación y Selección del modelo. El tercero, Capítulo 4, está dedicado a la *Validación y Calibración de los Modelos de Credit Scoring.* En la segunda parte, central en esta Tesis Doctoral, por un lado, en el Capítulo 5, *Funciones de Base en Modelos Notables de Credit Scoring,* se presenta de forma novedosa el estado actual del conocimiento sobre los modelos de *credit scoring*, y por otro, en el Capítulo 6, se desarrolla nuestra propuesta: *Modelos Logísticos Lineales Híbridos, HLLM.* La tercera parte de la Tesis Doctoral, Capitulo 7, se dedica a la *Construcción de un Modelo de Credit Scoring Proactivo Logístico Lineal por Expansión Híbrida de Funciones de Base, HLLM,* a partir de 63 variables de riesgo de crédito observadas sobre 76.607 acreditados en el *segmento de préstamos a particulares* de una Entidad Financiera Española.

1.- Primera parte.

El Capitulo 2 está dedicado a tres partes diferenciadas, por un lado, al *concepto y características básicas de la probabilidad de default*, conceptos generales de la teoría matemática de la probabilidad y de las variables aleatorias y sus distribuciones orientados al estudio de la probabilidad condicionada del estado de default a las variables de riesgo de crédito que caracterizan a la población de acreditados, secciones 2.1, 2.2, 2.3 y 2.4. En esta primera parte el protagonista es el *teorema de Bayes* que nos proporciona un importante instrumento teórico que relaciona la probabilidad de default con la función de calificación de acreditados a través de la transformación logística. En la sección 2.4, se expone la taxonomía de los métodos de estimación, desde la perspectiva del grado de conocimiento de los parámetros del modelo.

Por otro lado, utilizando el instrumento teórico comentado en el párrafo anterior es posible estimar probabilidad de default de forma directa a través de *vecindades locales, (k-vecinos más próximos y estimación de la densidad por funciones núcleo),* aspecto al que dedicamos la sección 2.5.

Por último, por distribuirse el estado de default según una variable aleatoria de Bernoulli, se tiene un instrumento que relaciona la probabilidad de default con la esperanza matemática del estado de default condicionado a la observación de las variables explicativas del riesgo de crédito. Este segundo instrumento teórico nos permite utilizar el enorme potencial de las herramientas de la regresión para construir modelos estadísticos capaces de predecir el default, a la vez que nos proporciona la expresión formal de la relación entre el default y la función de calificación de acreditados, síntesis de las variables explicativas del riesgo de crédito. Este hecho, junto con la constatación de que la estimación directa no resuelve satisfactoriamente, desde la óptica de Basilea II, el problema de la estimación de la probabilidad de default, justifican que una parte importante del resto del capítulo se dedique al *concepto y características de los modelos estadísticos de credit scoring*, sección 2.6, y a *la especificación de la estructura formal de un modelo de credit scoring*, sección 2.7.

En cuanto al Capítulo 3, tras una sección de introducción, 3.1, se abordan los aspectos más significativos de la ingente maquinaria necesaria para la *estimación de la probabilidad de default desde la óptica de la regresión*, a través de modelos estadísticos que relacionan una transformación de la probabilidad de default con la función de calificación de acreditados, es decir de *modelos de credit scoring, e*n las secciones 3.2, 3.3, 3.4 y 3.5.

La sección 3.2 se dedica a la estimación de los modelos de *credit scoring* analizando los principales instrumentos actualmente disponibles para esta tarea, no sólo para los *Modelos Logísticos Generales*, familia a la que pertenecen los *Modelos Logísticos Lineales Híbridos, HLLM,* objetivo principal de esta Tesis Doctoral, sino también otras familias de interés comparativo, los *Modelos Probit* y los *Modelos de Vector Soporte*.

Para evitar los problemas de la no unicidad de la solución así como el sobreajuste o infraajuste, que restan al modelo capacidad de generalización, se dedica la sección 3.3 a la r*egularización del modelo de probabilidad de default, con la que se* pretende conseguir modelos de tendencia antes que modelos muy ajustados localmente, utilizando para ello la *Funcional de Riesgo Regularizado.*

Tras la especificación y estimación de un modelo de *credit scoring* es necesario comprobar si los datos se ajustan adecuadamente al mismo, bondad de ajuste, es decir, se ha de

valorar la discrepancia entre los datos observados y los ajustados por el modelo, de modo que a *menor discrepancia mejor será el ajuste*. A estas cuestiones se dedica la sección 3.4. Otro aspecto fundamental a considerar en los modelos de credit scoring es su *capacidad de predicción*, que en el caso de variable de respuesta binaria coincide con su *poder discriminante*. En la sección 3.5 se describen los métodos más importantes para *validar la capacidad de predicción de un modelo de credit scoring* y se muestra cómo usarlos para *la selección del modelo*.

Por último, existe otra tarea a realizar, a la que Hastie et al. (2009) se refieren como la *fijación del modelo,* que consiste en, una vez elegido el mejor modelo, *estimar su capacidad de generalización o error de predicción esperado sobre nuevos datos*, usualmente la muestra test, a la que se dedica la última sección 3.6.

Por lo que respecta al Capítulo 4, la validación se fundamenta en el enfoque IRB de Basilea II que permite a las Entidades Financieras construir los modelos de riesgo para sus carteras de crédito a partir de información propia, pero requiere la *validación del proceso anterior* (BCBS (2006), § III 500): "*Los bancos deberán contar con un buen sistema para validar la precisión y coherencia de los sistemas de calificación, los procesos y la estimación de todos los componentes de riesgo pertinentes. Asimismo, deberán demostrar a sus supervisores que su proceso de validación interna les permite evaluar, de forma consistente y significativa, el funcionamiento de los sistemas de calificación interna y de estimación de riesgos*".

En la sección 4.1 se hace una síntesis de las recomendaciones del Comité de Supervisión Bancaria del Banco Internacional de Pagos de Basilea II, (BCBS, 2006), sobre la organización y las pautas que deben seguir las Entidades Financieras desde el enfoque IRB para *validar la precisión y coherencia de los sistemas de calificación*.

Entre las recomendaciones anteriores existen tres que analizamos especialmente en este capítulo, una referida a la *Validación Teórica*, sección 4.2, y dos aspectos de validación cuantitativa, el *Poder Discriminante del Modelo*, sección 4.3, y la *Calibración de la Probabilidad de Default,* sección 4.4, problema importante al que el grupo de validación de Basilea II prestó una atención considerable.

2.- Segunda parte.

El Capítulo 5 cuenta con una introducción y otras 7 secciones; después de la introducción, sección 5.1, en las cinco siguientes, 5.2-5.6, se hace una *revisión de los modelos de credit*

scoring más utilizados y estudiados en la actualidad desde la óptica del grado de conocimiento que se posee sobre la relación de dependencia entre el estado de default y las variables explicativas del riesgo de crédito y de la posible representación de la función de calificación de acreditados como expansión lineal de las funciones de base.

En la sección 5.2*, Modelos basados en el Desconocimiento Total de* $\boldsymbol{S}(\boldsymbol{X})$, se analizan los modelos construidos sobre el supuesto de que *no se conoce la distribución poblacional ni ninguna estructura formal que refleje la relación de dependencia entre la variable estado de default y las variables explicativas del riesgo de crédito.*

En la sección 5.3, *Modelos basados en el Desconocimiento Casi Total de* $\boldsymbol{S}(\boldsymbol{X})$, la idea básica en este caso consiste en "aumentar" nuestro conocimiento a priori como resultado de asumir la hipótesis de que *la función de calificación de acreditados es "suave", en el sentido de que para valores similares de las variables explicativas corresponden estadísticamente respuestas similares.*

En la sección 5.4*, Modelos basados en el Conocimiento Casi Total de* $\boldsymbol{S}(\boldsymbol{X})$.Dos populares situaciones representativas de la situación de conocimiento casi total de la distribución de las variables explicativas condicionadas al estado de default y de no default vienen dadas por el *Análisis Discriminante Lineal, LDA*, y por el *Análisis Discriminante Cuadrático, QDA.*

En la sección 5.5*, Modelos basados en el Conocimiento Total de* $\boldsymbol{S}(\boldsymbol{X})$, se analiza una situación que se encuentra en el otro extremo de los modelos analizados en la sección 5.2, se supone que conocemos la distribución conjunta de las variables explicativas y la variable respuesta, $P(X,Y)$, o las funciones de verosimilitud condicionada a las poblaciones de default, $f_{Y=1}(X)$, y de no default $f_{Y=0}(X)$, o bien la razón de verosimilitud, $LR(X)$, así como la probabilidad de default a priori, $P(Y=1)$.

En la sección 5.6*, Modelos basados en el Conocimiento parcial de* $\boldsymbol{S}(\boldsymbol{X})$, se analizan técnicas, *modelos estadísticos*, normalmente de regresión, o *algoritmos*, usualmente basados en la *teoría del aprendizaje*, que permitan estimar modelos generales de la probabilidad de default situados entre el conocimiento total y el total desconocimiento, es decir, en un *conocimiento parcial razonable*. Esta

sección se divide en dos apartados, uno para los métodos lineales y, otro para los métodos no lineales.

Por último, en la sección 5.7, *Taxonomía de los Modelos de Credit Scoring* , se muestra una clasificación de los modelos de credit scoring más relevantes.

El Capítulo 6 está dedicado a nuestra propuesta de *Modelos Logísticos Lineales Híbridos de Credit Scoring*, HLLM, obtenidos como consecuencia de expandir la componente no lineal de los modelos *Logísticos Parcialmente Lineales*, LPLM.

En la sección 6.2 se formaliza la estructura funcional y la estimación d*e los Modelos de Probabilidad Lineales Generalizados por funciones de base*, GLPM, familia de modelos donde se encuadra la subfamilia de nuestro interés, *Modelos Logístico Lineales Híbridos*, HLLM.

La estructura funcional de los modelos HLLM se formaliza en la sección 6.3, donde, además, se hace una *revisión muy somera de las aproximaciones a los modelos de probabilidad de default generalizados por expansiones lineales de las variables originales a través de funciones de base* que en los últimos años han propuesto distintos autores, (Müller, 2000, 2001, Müller y Härdle, 2003, Siddiqi, 2006, Liu y Cela, 2007, Liu y Cela, 2009).

En la sección 6.4 se plantea y resuelve el Problema General de Estimación de los Modelos Logísticos por expansiones lineales de funciones de base, que incluye la regularización del modelo, y consiste en una extensión, motivada por especificar la no linealidad a través de funciones de base, del Problema de Estimación del Modelo LOGIT.

En la sección 6.5 se particulariza la estimación del problema general de los modelos logísticos a los modelos HLLM.

En la sección 6.6 se expone el método supervisado que proponemos para el *proceso de selección de las funciones de base para la componente no lineal del modelo HLLM.*

Por último en la sección 6.7, se analizan con detalle algunas de las funciones de base más destacadas: *Funciones de base polinómicas*, *Funciones Constantes a Trozos obtenidas a partir de Indicadores de Particiones Recursivas*, *Pesos de la Evidencia asignados a Tramados de las Variables o a Particiones Recursivas*, *Splines Cúbicos Restringidos de Stone y Koo*, RCS, *Funciones de Base Bisagra obtenidas por MARS Univariante*, *Funciones*

de Base Radial, RBF, todas ellas serán utilizadas en el desarrollo del modelo proactivo de credit scoring que exponemos en el Capítulo 7.

3.- Tercera parte.

En el capítulo 7 se exponen todas las fases de desarrollo del modelo estructuradas en 8 secciones, la primera de ellas es la introducción, sección 7.1. En la sección 7.2 se realiza el *pre análisis de los datos poblacionales*, el *tramado de variables, el tratamiento de datos faltantes y de los valores extremos*, *la correlación y multicolinealidad entre las variables explicativas y el poder explicativo de las mismas*, y por último, *la selección de las variables explicativas y de la muestras de entrenamiento, validación y test.*

En la sección 7.3 se expone la metodología utilizada y los resultados obtenidos de la *exploración de los datos de entrenamiento*. La tarea fundamental emprendida en esta sección se orienta a explorar la linealidad y no linealidad de las variables de riesgo de crédito en relación con el logit de la probabilidad de default, *método de los residuos acumulados* de Lin et al. (2002), con el *test supremo de Kolmogorov,* (Su y Wei, 1991).

En la Sección 7.4, *Especificación y Ajuste del Modelo.* El modelo se especifica mediante una *expansión lineal de funciones de base de las variables explicativas, HHLM. El ajuste se resuelve a través de la solución del problema de estimación de los modelos lineales generalizados por expansión lineal de funciones de base.*

En la Sección 7.5, *Evaluación, Generalización y Selección del Modelo*, se evalúa si los modelos ajustados en la etapa de estimación son modelos válidos, más allá de que presenten un ajuste adecuado a los datos, evaluación que se materializa *utilizando, el Criterio de Información Bayesiano,* BIC, bondad de ajuste, y el Área Bajo la Curva ROC, AUC, poder discriminante de los modelos, sobre la muestra de validación. A continuación *se analiza si los modelos preseleccionados son suficientemente generalizables.*

En la Sección 7.6, *Reducción de la Complejidad del Modelo,* se realiza un "proceso de poda" del modelo seleccionado con el fin de obtener un parsimonioso modelo final.

En la Sección 7.7, se valida el *poder discriminante del modelo de credit scoring final*. test de Kolmogorov-Smirnov, tasa de precisión AR, área bajo la curva ROC, AUC y el test de Mann_Whitney.

En la Sección 7.8, se valida la *calibración del modelo de credit scoring, es decir, se analiza si es necesaria* la corrección de la *probabilidad de default,* en concreto se comprueba si la

probabilidad de default pronosticada por el modelo de credit scoring finalmente seleccionado a 30 de noviembre de 2007 debe ser revisada dadas las tasas de default realmente observadas un año después. Test Binomial, (Engelmann y Rauhmeier (2006)), *test Chi-cuadrado* (Hosmer y Lemeshow, 2000) y el *test de Spiegelhalter,* (Spiegelhalter, 1986, BCBS, 2005c).

En la sección 7.9 se realiza un *análisis comparativo de distintas técnicas de construcción de modelos de credit scoring*. Se compara el rendimiento del modelo *HLLM* con otros modelos ajustados sobre la muestra de entrenamiento: HLPM, TREE, SLPM, k-NN, *y* SMV. El rendimiento se evalúa sobre: los resultados del *ajuste del modelo a los datos de entrenamiento*, del *poder discriminante del modelo* sobre la población total de acreditados a 30 de noviembre de 2007 y de la *calibración del modelo*, es decir, el *grado de concordancia* entre las probabilidades pronosticadas, para el total de acreditados en esa fecha y las tasas de default realmente observadas un año después.

La Tesis Doctoral acaba con una relación de *Conclusiones,* seguida de la lista de Referencias Bibliográficas.

CAPÍTULO 2

PROBABILIDAD DE DEFAULT Y MODELOS DE CREDIT SCORING.

2.1. INTRODUCCIÓN.

En esta Tesis Doctoral asumimos que los datos observados sobre los acreditados son muestras, $\boldsymbol{\tau} = \left\{\left(x_i, y_i\right) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N} \in \left(X \times Y\right)^N$, de una distribución, usualmente desconocida, de las variables explicativas del riesgo de crédito y de la variable binaria indicador de incumplimiento, $\boldsymbol{F_{X,Y}}(\mathbf{x},\mathbf{y})$. Un principio fundamental es que *toda la información estadística de los datos está almacenada en la distribución de densidad conjunta* $\boldsymbol{f_{X,Y}}(\mathbf{x},\mathbf{y})$**.** Basándonos en este principio, la estimación de la densidad es la tarea fundamental para conocer la relación entre el estado de default y las características observadas sobre el acreditado o el solicitante de crédito, así como para calificarlo, clasificarlo y conocer la influencia que cada variable explicativa juega con respecto a su comportamiento ante el default.

Por tanto, en primera instancia, nuestro interés radica en el comportamiento del "indicador de default", $\mathbf{Y}$, condicionado a los valores de $\boldsymbol{X} = \left(X_1,..,X_j,..,X_P\right)^T$, es decir, *estamos interesados en la variable condicionada* $\boldsymbol{Y / X = x}$, $\boldsymbol{x} \in \mathbb{R}^p$. La probabilidad de la variable $\mathbf{Y}$ condicionada a un valor $\boldsymbol{x}$ de $\mathbf{X}$, $\boldsymbol{P\left(Y / X = x\right)}$, $\boldsymbol{x} \in \mathbb{R}^p$, llamada probabilidad a posteriori, es clave en esta Tesis Doctoral por dos razones principalmente:

- Por un lado, la probabilidad de default se define precisamente como la probabilidad a posteriori del comportamiento del acreditado condicionado a la información de riesgo de crédito que lo caracteriza. Esto será fundamental para poder conocer la pérdida esperada de nuestras carteras de acreditados y de sus créditos. Además, como veremos, *la función de calificación "teórica" de los acreditados no es más que la transformación logit de la probabilidad de default*.

- Por otro lado, la probabilidad a posteriori es una función clave en el proceso de clasificación tanto de los acreditados como de los nuevos solicitantes de crédito, puesto que, como veremos, un clasificador óptimal (Clasificador de Bayes) clasifica a un acreditado de acuerdo con su probabilidad a posteriori. En *la práctica no es posible implementar la regla de clasificación de Bayes a causa precisamente de que las probabilidades a*

posteriori no son conocidas por lo que deberán ser estimadas desde los datos.

De este modo de los tres elementos clave a estudiar, dos de ellos, la calificación y la clasificación de acreditados y solicitantes de crédito se basan en el tercero, la probabilidad de default, Por tanto, *es fundamental encontrar un estimador "tan ajustado como sea posible" a esta probabilidad.*

La estimación de las probabilidades de default para los deudores individuales es el primer paso para evaluar el riesgo de crédito y las pérdidas potenciales a que se enfrenta un inversionista o una Entidad Financiera**.** Las probabilidades de default pueden ser fijadas a priori, y luego cada préstamo debe ser adecuadamente asignado a una categoría de calificación. Estas categorías pueden determinarse por las tasas de morosidad de años anteriores o estimarse a partir de probabilidades de default, calculadas por sistemas estadísticos de credit scoring, por sistemas de agregación del conocimiento de expertos o por la combinación de ambos. Esta estimación será en general un reto debido a la usual limitación en la disponibilidad de datos.

Como veremos en la sección 2.2, existen dos instrumentos teóricos que relacionan la probabilidad de default con la función de calificación de acreditados:

$$(1)\quad \text{logit}\left(P\left(Y=1/\ X=x\right)\right)=S(x)=logit\left(p\right)+\log\left(LR(x)\right),\quad x\in\mathbb{R}^{p}$$

$$(2)\quad P\left(Y=1/\ X=x\right)=\Lambda\left(S(x)\right)=E\left[Y\ /\ X=x\right]$$

donde $P\left(Y=1/\ X=x\right)$ es la probabilidad de default, $S(x)$ es la función de calificación, $LR(x)$ es la razón de las funciones de verosimilitud de las poblaciones de default y no default, $\Lambda(\bullet)$ es la función de distribución acumulada logística $L(0,1)$, $\Lambda\left(z\right)=\dfrac{1}{1+e^{-z}}$, y $E\left[Y\ /\ X=x\right]$ es la esperanza matemática del estado de default condicionado a la observación x del vector aleatorio de variables explicativas del riesgo de crédito.

Por un lado, utilizando la expresión (1), que formaliza la relación entre la probabilidad de default y la función de calificación de acreditados mediante la transformación logística, se obtienen dos estimadores no paramétricos de los vecinos más próximos de la razón de verosimilitud, a partir de los cuales es posible estimar de forma directa la probabilidad de default. Para especificar la vecindad local para cada uno de los estimadores utilizamos las

métricas de los k vecinos más próximos y de las funciones núcleo de Parzen respectivamente.

Si bien es verdad que tanto los estimadores $k-NN$ como las funciones núcleo de Parzen gozan de adecuadas propiedades asintóticas, siendo los segundos más suaves y sofisticados que los primeros, no es menos cierto que ambos adolecen de dos males principales, uno, sobreajustanlos datos, por lo que no son apropiados para la clasificación de nuevos acreditados, requerimiento clave del NACB II, y, dos, sobre ellos se cierne la *maldición de la dimensionalidad* acompañada de tiempos de computación impracticables.

Como alternativa a los métodos del tipo anterior se encuentran los métodos que utilizan la expresión (2) ajustando los datos a modelos estadísticos, asumiendo algunas hipótesis estructurales para la relación que liga el estado de default con las variables explicativas del riesgo de crédito, lo que conduce a modelos estadísticos paramétricos o semiparamétricos, en función del grado de conocimiento real o asumido sobre la estructura del modelo.

2.2. PROBABILIDAD DE DEFAULT. CONCEPTO Y CARACTERÍSTICAS.

En primer lugar haremos una breve descripción de los conceptos y características básicas asociados a la probabilidad de default.

Consideremos la población de acreditados actuales y potenciales de una Entidad Financiera, representada por el espacio poblacional $\Omega=\{\Pi_0,\Pi_1\}$ en relación con el cumplimiento de las obligaciones de pago respecto al crédito contraído con la Entidad, $\Pi_0=\{no\ default\}$, $\Pi_1=\{default\}$. La variable aleatoria objetivo, con respecto a la probabilidad de default, es la variable respuesta binaria Y, "*indicador del incumplimiento*", con valores {0,1}, que está definida sobre el espacio de probabilidad $(\Omega,A(\Omega),\mathrm{P}(\bullet))$ en la forma siguiente:

$$Y(\omega)=I_{\Pi_1}(\omega)=\begin{cases}1\ si\ \omega\in\Pi_1\\0\ si\ \omega\in\Pi 0\end{cases},\quad \forall\omega\in\Omega \tag{2.1}$$

donde $A(\Omega)$ es el álgebra de sucesos generados por Ω, y $\mathrm{P}(\bullet)$ es una probabilidad definida sobre el espacio aleatorio $(\Omega,A(\Omega))$.

La cantidad $P(\Pi_1)=P\{\omega\in\Omega/\omega\in\Pi_1\}$, que notaremos por $P(Y=1)$, es la *probabilidad "a priori"* de que un acreditado o un solicitante de crédito no c mpla

con sus obligaciones de pago frente al crédito y $P(\Pi_0) = P\{\omega \in \Omega / \omega \in \Pi_0\}$, notada $P(Y=0)$, es su probabilidad complementaria.

La variable Y se distribuye según una variable aleatoria de Bernouille de parámetro $p = P(Y=1)$, **Y ~ Be(p)**, con función masa de probabilidad

$$P(Y = y) = p^y (1-p)^{1-y} \tag{2.2}$$

El conjunto de características observadas sobre los acreditados constituye un vector de p variables aleatorias $\boldsymbol{X} = (X_1,..,X_j,..,X_P)^T$ definidas sobre un espacio de probabilidad generado sobre el espacio de acreditados.

Si por $f_X(x)$ notamos la densidad del vector aleatorio p-dimensional $\boldsymbol{X}$ en el punto $x = (x_1,....,x_p)^T \in \mathbb{R}^p$, y por $f_0(x)$ y $f_1(x)$ las *funciones de verosimilitud* de las observaciones $x \in \mathbb{R}^p$ condicionadas a la certeza de la población de procedencia, que vienen dadas por:

$$f_k(x) = f_{X/Y=1}(x) = \frac{P_{X,Y}(X=x, Y=k)}{P(Y=k)}, \quad x \in \mathbb{R}^p, \quad k=0,1 \tag{2.3}$$

entonces la función de probabilidad del indicador de incumplimiento Y, condicionada a la información proporcionada por la observación $x \in \mathbb{R}^p$ *del vector de características* X se define, para $k=0,1$, como

$$P(Y=k / X=x) = \begin{cases} \dfrac{P_{Y,X}(Y=k, X=x)}{f_X(x)} & x \in \mathbb{R}^p / f_X(x) > 0 \\ 0 & \textit{en otro caso} \end{cases} \tag{2.4}$$

donde $P_{Y,X}(y,x)$ *es la distribución conjunta de* X *e* Y *y* $f_X(x)$ *es la densidad marginal de* X *en el punto* $(x,y) \in \mathbb{R}^p \times \{0,1\}$.

Definición 2.1.- a) La cantidad $P(Y=1/X=x)$ *es la "probabilidad de default a posteriori", o "probabilidad de default condicional", una vez conocido el valor* $x \in \mathbb{R}^p$, *a la que nos referiremos en adelante como "probabilidad de default".*

Complementariamente $P(Y=0/X=x)$ *es la "probabilidad de no default" una vez conocido* x, $P(Y=0/X=x) = 1-P(Y=1/X=x)$.

b) *La <u>función de probabilidad de "default"</u> es, por tanto, la función real con valores en* $[0,1]$ *de p variables reales que asigna a cada observación* $x\in\mathbb{R}^p$ *la probabilidad de "default" para esa observación:*

$$\begin{aligned} P(Y=1/X=\bullet):\mathbb{R}^p &\longrightarrow [0,1] \\ x &\longrightarrow P(Y=1/X=x), \quad x\in\mathbb{R}^p \end{aligned} \tag{2.5}$$

c) *Se llama <u>razón de verosimilitud de</u>* X *a la función cociente de las funciones de verosimilitud*

$$\begin{aligned} &LR(\cdot): I\subset\mathbb{R}^p \longrightarrow \mathbb{R} \\ &x \longrightarrow LR(x)=\frac{f_1(x)}{f_0(x)}, \quad \forall x\in\mathbb{R}^p \text{ que verifique: } f_0(x)>0 \end{aligned} \tag{2.6}$$

d) *Supuesto el hiperintervalo* $I = I_1\times\ldots\times I_j\times\ldots\times I_p$, *se llama <u>razón de verosimilitud marginal de</u>* X_j *<u>en</u>* $x_j\in I_j\subset\mathbb{R}$ *al cociente de las verosimilitudes marginales de* X_j *condicionada al default y al no default en* x_j, *respectivamente, y se expresa*

$$LR(x_j)=\frac{f_{1j}(x_j)}{f_{0j}(x_j)}, \quad \forall x_j\in\mathbb{R}, \quad f_{0j}(x_j)>0 \tag{2.7}$$

Por el *<u>Teorema de Bayes</u>* se tiene,

$$P(Y=k/X=x)=\frac{P(Y=k)\ f_k(x)}{f_X(x)}, \quad \forall x\in\mathbb{R}^p, \quad \mathrm{k}=0,1 \tag{2.8}$$

es decir, las probabilidades de default y no default, para una determinada observación $x\in\mathbb{R}^p$, se obtienen a partir de las probabilidades a priori del estado del default y de las verosimilitudes de las variables explicativas $(X_1,\ldots,X_p)$:

$$\begin{aligned} P(Y=1/X=x)&=\frac{P(Y=1)\ f_1(x)}{f_X(x)}, \quad x\in\mathbb{R}^p \\ P(Y=0/X=x)&=\frac{P(Y=0)\ f_0(x)}{f_X(x)}, \quad x\in\mathbb{R}^p. \end{aligned} \tag{2.9}$$

Esta relación es fundamental por cuanto *nos muestra el modelo matemático teórico que relaciona la probabilidad de default con las* variables explicativas $(X_1,\ldots,X_p)$. Por esta razón a lo largo de la tesis enfatizaremos el hecho de que los modelos de estimación de la probabilidad deberán preservar esta relación.

Obsérvese que, cuando las variables $(X_1,\ldots,X_p)$ son absolutamente continuas en $x\in\mathbb{R}^p$, se tiene $P(X=x)=0$, por lo que $P(Y=1/X=x)$ es una probabilidad condicional en el sentido no elemental y ha de ser manejada con cuidado.

2.3. LA PROBABILIDAD DE DEFAULT COMO TRANSFORMACIÓN LOGÍSTICA DE LA RAZÓN DE VEROSIMILITUD.

Siendo $odds(z)=\dfrac{z}{1-z}$, $z\in[0,1]$, y $p=P(Y=1)$, de la expresión (2.9) se sigue que

$$odds\left(P(Y=1/X=x)\right)=odds(p)\ \times\ LR(x),\quad \forall x\in Rango(X)\subset\mathbb{R}^p$$

de donde

$$\log\left(odds\left(P(Y=1/X=x)\right)\right)=\log\left(odds(p)\right)+\log\left(LR(x)\right)$$

por lo que, utilizando la notación $\text{logit}(z)=\log\left(\dfrac{z}{1-z}\right)$, $z\in[0,1]=$, se obtiene

$$logit\left(P(Y=1/X=x)\right)=logit(\mathrm{p})+log\left(LR(x)\right)\quad x\in Rango(X) \tag{2.10}$$

donde $logit(\mathrm{p})$ es un término constante no dependiente de x.

La parte derecha de la igualdad (2.10) *es una función en* x, *que representaremos por* $S(X)$, *que contiene la información proporcionada por la probabilidad de default a priori más toda la proporcionada por la razón de verosimilitud, ambas en escala logarítmica,*

$$S(X)=logit(\mathrm{p})+log\left(LR(X)\right) \tag{2.11}$$

La función $S(X)$ representa, por tanto, toda la información contenida en las variables explicativas $(X_1,\ldots,X_p)$ sobre el comportamiento del cliente frente al cumplimiento de sus obligaciones de crédito, por lo que en el contexto de los sistemas de calificación se asocia $S(X)$ a la calidad crediticia del acreditado (variable latente no observada). *A la variable aleatoria* $S(X)$ *se le llamará en esta*

Tesis Doctoral, indistintamente, función de calificación o función de puntuación crediticia.

Una de las muchas misiones de la función $S(X)$, o bien de una de sus estimaciones, es pronosticar el futuro estado de Y para cada acreditado, a partir de la información que contiene sobre el merecimiento de crédito.

La función de calificación $S(X)$ es clave para la construcción de la calificación de los acreditados, y también juega un papel importante en la construcción de la regla de decisión sobre la pertenencia o no de un cliente a la población de default en un horizonte próximo.

A partir de la expresión (2.10) podemos concluir que *la probabilidad de que un acreditado o un solicitante de crédito, para el que se ha observado un valor* $x=(x_1,....,x_p)^T \in Rango(X)\subset \mathbb{R}^p$ *del vector aleatorio* p*-dimensional* $\boldsymbol{X}=(X_1,...,X_P)^T$, *vaya a incumplir sus obligaciones de pago en el horizonte temporal próximo, viene dada* por

$$P(Y=1/\ X=x)=\Lambda\left(S(x)\right)=\frac{1}{1+e^{-\{logit(p)+\log(LR(x))\}}},\quad \forall x\in Rango(X) \tag{2.12}$$

donde $\Lambda(\bullet)$ es la función de distribución acumulada logística $L(0,1)$, $\Lambda(z)=\frac{1}{1+e^{-z}}$, $\forall z\in\mathbb{R}$, $p=P(Y=1)$ es la probabilidad de default a priori y $LR(x)$ es la razón de verosimilitud de X.

Queremos destacar fundamentalmente el hecho de que, tal como indica (2.12), *la probabilidad de default enlaza con las características de riesgo de crédito* $(X_1,...,X_P)$*, a través de la función de distribución logística, esto nos orienta claramente sobre parte de la estructura básica de los modelos de estimación de la probabilidad de default.*

La probabilidad de default, por tanto, puede expresarse en la forma

$$P(Y=1/\ X=x)=\Lambda\left(S(x)\right) \tag{2.13}$$

lo que es equivalente a

$$S(X)=logit\left(P(Y=1/\ X=x)\right) \tag{2.14}$$

donde $S(X)$ viene dada por (2.11).

Las igualdades reflejadas en las expresiones (2.13) o (2.14), indistintamente, constituyen un primer instrumento teórico básico para calcular la probabilidad de default.

La precisión con que será posible determinar la probabilidad de default, a partir (2.13) o (2.14), dependerá del conocimiento que se posea sobre la función $S(X)=logit(\mathrm{p})+log\left(LR(X)\right)$, que se situará en uno de los innumerables estadios intermedios entre las fronteras irreales del conocimiento total y el total desconocimiento. En el supuesto, muy poco probable, de que se conozca la distribución de las probabilidades a priori de la variable estado del default, Y, y las verosimilitudes del vector de variables explicativas, $\boldsymbol{X}=(X_1,...,X_P)^T$, para las poblaciones de default y no default, la probabilidad de default quedará perfectamente determinada, lo que ilustraremos a continuación, ejemplo 2.1, con una función de calificación de acreditados muy popular, la *función discriminante de Fisher*.

Ejemplo 2.1.- Si las distribuciones de X condicionadas al default y no default sean normales de distintas medias e igual matriz de covarianza, $X/(Y=1)\sim N(\mu_1,\Sigma)$, $X/(Y=0)\sim N(\mu_0,\Sigma)$, se tiene

$$f_{X/Y=k}(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\},\quad \mathrm{k=0,\ 1,}$$

de donde,

$$log\left(LR(x)\right)=-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)$$

por tanto,

$$S(X)=\beta_0+\boldsymbol{\beta}^T X \qquad (2.15)$$

siendo $\beta_0=logit(p)-\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$ y $\boldsymbol{\beta}=\Sigma^{-1}(\mu_1-\mu_0)$

Dado que $logit(p)$ es un término constante no dependiente de X, la relación anterior indica que con las hipótesis establecidas, la función de calificación de acreditados coincide con la frontera de clasificación del *Análisis Lineal Discriminante*, (LDA), (FISHER (1936), LADD (1966), LACHENBRUCH, (1975)).

La probabilidad de default resulta entonces $\mathrm{P}(Y=1/\ X=x)=\Lambda\left(\beta_0+\beta^T x\right)$, o, equivalentemente, $logit\left(\mathrm{P}(Y=1/\ X=x)\right)=\beta_0+\beta^T x$. *Es decir, bajo la hipótesis de*

normalidad de las poblaciones y matrices de covarianza iguales, el logit de la probabilidad de default se relaciona con el vector aleatorio X a través de la función discriminante de Fisher, frontera de clasificación del LDA.

En este caso, el modelo que liga la probabilidad de default con las variables explicativas del riesgo de crédito tiene propiedades clave en los requerimientos de Basilea II, si bien hemos de advertir que en la práctica de los sistemas de calificación sus hipótesis no se verifican casi nunca, como ocurre en la mayoría de los sucesos económicos donde el libre albedrío de los agentes económicos, al contrario de lo que ocurre en los sucesos naturales, deja muy poco espacio a la campana de Gauss.

El modelo se obtiene directamente a través del *teorema de Bayes*, de las probabilidades a priori del default y no default y de las verosimilitudes de X condicionadas a las poblaciones de default y no default, supuestamente conocidas, sin error posible.

Una característica adicional de este modelo es que no es necesario hacer ninguna suposición ni sobre la forma de las distribuciones condicionadas ni sobre sus parámetros, puesto que se suponen conocidos, y, en el peor de los casos, estos últimos podrían ser estimados a través de estimadores con convenientes propiedades como consistencia, eficiencia, convergencia asintótica, etc.

Sin embargo, desgraciadamente en la práctica de los sistemas de calificación lo habitual es no conocer ni las probabilidades a priori ni las distribuciones de las verosimilitudes. De estas últimas con frecuencia no sólo no se conocen los parámetros sino incluso no se conoce la forma. Las probabilidades a priori pueden ser estimadas fácilmente a partir del estimador de máxima verosimilitud de la proporción de "default" en la muestra de acreditados, puesto que se conoce perfectamente la distribución de la variable aleatoria respuesta Y, distribución de Bernouilli de parámetro p. Si no se conoce el parámetro p, puede ser estimado paramétricamente utilizando una muestra aleatoria simple de tamaño N, $(Y_1,...,Y_N)$, a través de la función de verosimilitud de la muestra

$$L(p) = p^{\sum_{i=1}^{n} y_i} (1-p)^{n-\sum_{i=1}^{n} y_i} \qquad (2.16)$$

El estimador de máxima verosimilitud de p, $\hat{p}$, es la proporción de "default" en la muestra que, como todo estimador de máxima verosimilitud, es consistente, asintóticamente eficiente y normal.

De no conocer las verosimilitudes, o la razón de verosimilitud, no tendremos más remedio que estimarlas, lo que analizaremos en la sección 2.5 de este capítulo, pero la estimación directa de las verosimilitudes, sin más información que los datos de riesgo de crédito observados sobre los acreditados, conlleva problemas importantes que, como veremos, imposibilitan una solución satisfactoria. Pero afortunadamente contamos con una segunda herramienta teórica para estimar la probabilidad de default, puesto que al ser la variable aleatoria respuesta Y binaria con distribución de Bernoulli de parámetro $p = P(Y = 1/\ X = x)$, se puede expresar la probabilidad de default, de forma alternativa, a través de la esperanza condicional, $E[Y / X = x]$,

$$P(Y = 1/\ X = x) = E[Y / X = x] \tag{2.17}$$

por tanto,
$$E[Y / X = x] = \Lambda(S(x)) \tag{2.18}$$

es decir, la función $\Lambda(S(X))$ coincide con la función de regresión de Y sobre X, razón por lo que se le llama ***regresión logística***, en sentido amplio.

Dado que la esperanza condicional coincide con la regresión de Y sobre X, en el caso frecuente de no conocer el modelo teórico y todos sus parámetros, se puede trasladar toda la potencia de la regresión a la estimación de las probabilidades de default, lo que nos permitirá explicar y en su caso predecir ya sea por métodos paramétricos, semiparamétricos o no paramétricos, el valor de Y dados los valores de X, todo ello desde la perspectiva de que nuestro objetivo será buscar los efectos de los factores simples y encontrar el mejor modelo.

El objetivo consiste estimar la función $r(x) = E[Y / X = x]$ *para predecir el valor de* $P(Y = 1/\ X = x)$ *dados los valores de* X *a través de un modelo de ajuste de la relación de dependencia entre el default y las variables explicativas del riesgo de crédito*

$$\text{logit}\left(\hat{P}(Y = 1/\ X = x)\right) = \hat{S}(x) + \varepsilon \tag{2.19}$$

donde se asume que el término de error ε tiene distribución logística de media 0 y varianza $\frac{\pi^2}{3}$. Además, se asume que se observa el default si $\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)$ es positivo:

$$\hat{y}=\begin{cases}1 & si\ \ \text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)>0\\ 0 & \text{en otro caso}\end{cases} \tag{2.20}$$

2.4. MÉTODOS DE ESTIMACIÓN DESDE EL PUNTO DE VISTA DEL CONOCIMIENTO DISPONIBLE.

Desde el punto de vista del conocimiento disponible sobre la forma y parámetros de las distribuciones de las variables explicativas, podemos dividir los métodos de estimación de la probabilidad de default, y de la función de calificación de acreditados, en tres categorías: *métodos paramétricos*, no *paramétricos* y *semiparamétricos*.

2.4.1 Métodos Paramétricos.

Los métodos paramétricos son aquellos para cuya utilización se requiere o bien un conocimiento total sobre la forma y parámetros de las distribuciones de las variables explicativas o, si es razonable y está justificado, o bien suposiciones sobre tales elementos. Una vez que se fija alguna función $S(\mathrm{X})$, de una clase conocida de funciones de densidad, se debe utilizar una técnica adecuada para ajustar esta función a los datos de entrenamiento. Obtenemos así un estimador $\hat{P}\left(Y=1/\ X=x\right)$ de la probabilidad de default o bien un estimador $\hat{S}(\mathrm{X})$ de la función de calificación del crédito, que deberán tener convenientes propiedades como consistencia, eficiencia, convergencia asintótica etc. Hacer "*fuertes suposiciones sobre los parámetros que caracterizan a la distribución poblacional*", implica correr el riesgo de especificaciones estructurales erróneas.

2.4.2 Métodos No Paramétricos.

Del conocimiento total, real o supuesto, sobre la forma y parámetros de las distribuciones de las variables, pasamos al extremo contrario, donde ni se conoce, ni se supone, ni las observaciones manifiestan explícitamente ninguna forma concreta de la distribución. Ante el desconocimiento de la relación entre las variables predictivas y la variable respuesta, parece razonable actuar bajo la filosofía de "*dejar hablar a los datos*". En la vía no paramétrica se supone, con una gran dosis de realismo, que no conocemos la

distribución de origen de los datos, lo que en términos estadísticos consiste en suponer que la distribución del vector aleatorio poblacional no está caracterizada por un parámetro finito dimensional. En este enfoque "no se supone un modelo paramétrico para la distribución poblacional".

El problema típico de estimar una distribución conjunta de alta dimensión, $S(\bullet):\mathbb{R}^p \to \mathbb{R}$, sobre la base de una muestra de (X,Y), $\tau = \left\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N} \subset (X \times Y)^N$, es que inmediatamente crece la varianza alrededor de las estimaciones como resultado de que los datos pueden estar muy desparramados, problema observado por BELLMAN (1961) y que los autores anglosajones designan con el nombre de *curse of dimensionality* (HUBER, 1985), *la maldición de la dimensionalidad* en castellano, a lo que hay que añadir el tiempo de computación con grandes conjuntos de datos (Una ilustración muy elaborada de la maldición de la dimensionalidad puede verse en HASTIE et al. (2009)).

La maldición de la dimensionalidad hace virtualmente imposible estimar $S(\mathrm{X})$ directamente en una vía no paramétrica completa sin aceptar algunas hipótesis estructurales. Por tanto, *los métodos que se utilizan en la práctica son modelos flexibles de naturaleza no paramétrica sobre los que se asumen algunas hipótesis estructurales.*

2.4.3 Métodos Semiparamétricos.

Una vía intermedia entre los métodos paramétricos y los no paramétricos la constituyen las técnicas semiparamétricas, que participan de las ventajas de unas y otras, aunque también de sus inconvenientes, pero que sin duda es una vía muy razonable en el universo financiero, por cuanto *si se dispone de un número importante de variables observadas sobre la población de acreditados, si bien es cierto que con frecuencia no se conoce todo sobre su distribución no es menos cierto que siempre se conoce "algo"*. La consideración anterior es generalizable a cualquier campo de actividad humana, razón por la cual están surgiendo con gran auge las técnicas de estimación semiparamétrica. Como veremos, estos modelos híbridos están imponiéndose con gran fuerza en la industria del credit scoring.

2.5. ESTIMACIÓN DIRECTA DE LA PROBABILIDAD DE DEFAULT EN VECINDADES LOCALES.

2.5.1 Introducción.

En esta sección abordamos la estimación directa de la probabilidad de default, y correspondientemente de la función de calificación de acreditados, bajo la hipótesis de que no conocemos la distribución de origen de los datos, lo que en términos estadísticos consiste en suponer que la distribución del vector aleatorio poblacional no está caracterizada por un parámetro finito dimensional. En este enfoque simplemente se actúa con la filosofía de "*dejar hablar a los datos*". Por tanto, d*ejaremos que hablen sólo los datos, escuchándolos estimaremos directamente la probabilidad de default, en vecindades locales, a través de los vecinos más próximos y sólo a partir de ahí entraremos en la tertulia.*

Como se ha reflejado en las secciones 2.2 y 2.3 de este capítulo, para estimar la probabilidad de default, $P(Y=1/\ X=x)$, se cuenta con dos herramientas teóricas básicas, por un lado (2.11) y (2.13) y, por otro, (2.17). En el primer caso se estima la función de calificación de los acreditados $S(\mathrm{X})$, a partir de estimadores adecuados de la razón de verosimilitud $LR(X)$ y, por tanto, las funciones de verosimilitud $f_0(X)$ y $f_1(X)$, y en el segundo, la probabilidad de default en $x \in \mathbb{R}^p$ coincide con la esperanza condicional en ese punto, por lo que se puede intentar estimar esta directamente o aplicar toda la maquinaria de la regresión para estimar la probabilidad de default.

Procederemos en esta sección a la estimación de la probabilidad de default por los vecinos más próximos vía la primera herramienta (2.11) y (2.13), (apartados 2.5.2 y 2.5.3).

De acuerdo con la definición de la razón de verosimilitud de las variables explicativas del riesgo de crédito, se puede obtener un estimador para $LR(X)$ en función de los estimadores de las verosimilitudes conjuntas de dichas variables condicionadas a las poblaciones de default, $f_1(X)$, y no default, $f_0(X)$. Nuestro objetivo será, por tanto, estimar estas verosimilitudes, y, por tanto, la razón de verosimilitud en cada punto $x \in D \in \mathbb{R}^p$, siendo D el dominio de interés. La probabilidad de default, se estima, en cada punto, de acuerdo con la relación teórica (2.12), es decir

$$\hat{P}(Y=1/\ X=x)=\frac{1}{1+e^{-\left\{logit(\hat{p})+\log\left(\widehat{LR}(x)\right)\right\}}},\quad x \in D \subset \mathbb{R}^p \tag{2.21}$$

donde $\widehat{LR}(x) = \dfrac{\hat{f}_1(x)}{\hat{f}_0(x)}$, *con* $\hat{f}_0(x) > 0$.

El estimador de máxima verosimilitud de p, $\hat{p} = \hat{P}(Y=1)$, probabilidad de default a priori, viene dado por

$$\hat{p} = \frac{\sum_{i=1}^{N} I_{\{y_i=1\}}}{N} \tag{2.22}$$

que, como todo estimador de máxima verosimilitud, es consistente, asintóticamente eficiente y normal. Por tanto, $\text{logit}(\hat{p})$ es un estimador de $\text{logit}(p)$ con las mismas propiedades que $\hat{p}$. Además, el intervalo de confianza a nivel $(1-\alpha)$ para $\hat{p}$ es $\left[\hat{p} - z_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{N}}, \hat{p} + z_{1-\alpha/2}\sqrt{\dfrac{p(1-p)}{N}}\right]$.

El problema se concentra principalmente en la estimación de $f_0(X)$ y $f_1(X)$, cuyos estimadores $\hat{f}_0(X)$ y $\hat{f}_1(X)$ se han de obtener bajo el supuesto de que no se conocen las distribuciones, ni siquiera la forma, tanto de X como de $X/(Y=k)$, $k=0,1$. Por tanto, será necesario estimarlas por métodos no paramétricos, es decir deberemos *"dejar hablar a los datos"*. Como consecuencia, nuestro objetivo, consiste en encontrar un estimador de las funciones de verosimilitud, $\hat{f}_0(X)$ y $\hat{f}_1(X)$, a partir de una muestra de N observaciones, $\tau = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1,\ldots,N}$, obtenida de las poblaciones de acreditados de default y no default.

2.5.2 Estimación de la Razón de Verosimilitudes por los k Vecinos Más Próximos, K-NN.

El primer estimador de la densidad fue propuesto, en un trabajo no publicado, por FIX y HODGES (1951) como una vía para liberar el análisis discriminante de las rígidas hipótesis de normalidad multivariante distribucional si bien el primer trabajo publicado explícitamente sobre el tema se debe a ROSENBLATT (1956). A partir de estos trabajos se propusieron muchos métodos, todos ellos fundamentados en el concepto inicial de "vecinos más próximos" de un punto x, es decir, conjunto de los k puntos más próximos a x en distancia Euclídea, $V_k(x)$.

Los métodos de los *vecinos más próximos* tienen ya una larga historia tanto en el campo de la estimación de densidades como en el campo de la clasificación. *Un estimador muy conocido de la funciones de densidad es el estimador* $k\text{-}NN$ debido a LOFTSGAARDEN y QUESENBERRY (1965), que bajo la hipótesis de $X_1,...,X_N$ variables aleatorias independientes e idénticamente distribuidas con valores en $\mathbb{R}^p$ y con densidad de probabilidad común $f(x)$, viene dado por la proporción de acreditados en la vecindad ponderada inversamente por el volumen de la más pequeña esfera centrada en x que contiene los k vecinos más próximos a x, $Vol_k(x)$

$$f_N(x) = \frac{k}{NVol_k(x)} \tag{2.23}$$

Es razonable pensar que el volumen de la hiperesfera que encierra un número fijo de puntos, k, es menor en regiones de vecindad densamente pobladas que en regiones donde los puntos están más dispersos, figura 2.1. *Este sencillo planteamiento es la base de la estimación no paramétrica mediante los* k *vecinos más próximos ya que el volumen de la hiperesfera que encierra a* k *puntos está relacionado con el valor de la función de densidad de probabilidad en el centro de la hiperesfera.*

En el panel izquierdo de la figura 2.1 se representa gráficamente el diagrama de dispersión para la variable bidimensional (X_1, X_2) y en el panel derecho tres hiperesferas, círculos en el caso bidimensional, con los 6 vecinos más próximos a los puntos centrales de los círculos (x_{11}, x_{12}), (x_{21}, x_{22}) y (x_{31}, x_{32}). Es obvio que la mayor densidad corresponde a los puntos para los cuales el área del círculo es menor.

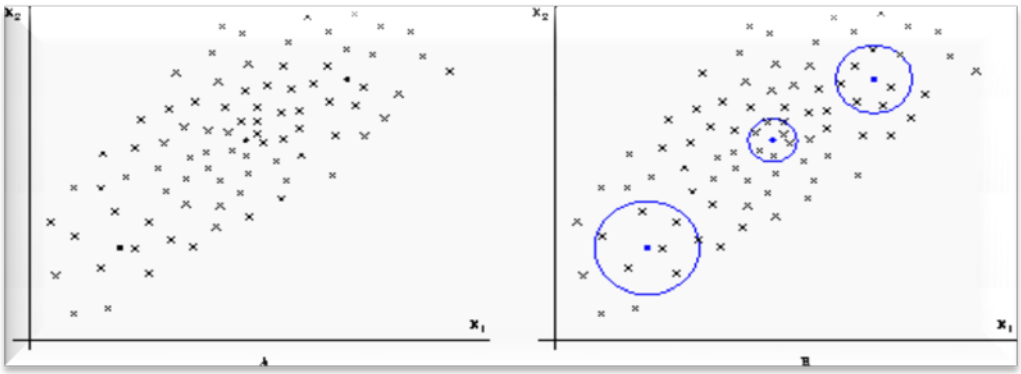


Figura 2.1.- Tres hiperesferas de dimensión 2 con los 6 vecinos más próximos, con densidades inversamente proporcionales a su volumen, área en este caso.

DEVROYÉ y WAGNER (1977), basándose en los trabajos de FIX y HODGES (1951), LOFTSGAARDEN y QUESENBERRY (1965), MOORE y HENRICHON (1969) y WAGNER (1973), demostraron que el *estimador de la densidad de* $f(x)$ *por los k vecinos más próximos,* bajo las condiciones muy razonables y sencillas siguientes

$$\lim_{N\to\infty}\frac{k(N)}{N}\to 0 \quad \text{y} \quad \lim_{N\to\infty}\frac{k(N)}{\log(N)}\to\infty$$

es *robusto y uniformemente consistente*, es decir,

$$\lim_{N\to\infty}\left|f_N(x)-f(x)\right|\to 0$$

Este teorema constituye una base teórica importante, puesto que si para muestras grandes el estimador $k\text{-}NN$ no fuera robusto y consistente es presumible que sería excesivamente inestable para muestras pequeñas, pero esto no es suficiente por cuanto, con frecuencia, en los Sistemas de Calificación no se cuenta normalmente con muestras grandes, sobre todo de la población de default.

De contar con muestras suficientemente grandes, la consistencia y robustez del estimador $k\text{-}NN$ para estas muestras podría convertirlo en un candidato a tener en cuenta para estimar las funciones de verosimilitud en los modelos de credit scoring, si no fuera porque contamos con el siguiente problema añadido: a pesar de que el estimador $k\text{-}NN$ se basa en una hipótesis estructural del modelo muy flexible, por cuanto no requiere apenas hipótesis, es muy poco suave, lo que conduce a un modelo que se ajusta con frecuencia muy estrechamente a los datos, sobreajuste, por lo que resulta ser muy poco eficaz en la generalización, cualidad imprescindible para la correcta clasificación de una nueva solicitud de crédito.

A partir del teorema de DEVROYÉ y WAGNER (1977) *podemos estimar directamente las verosimilitudes en una vecindad* $V_k(x)$, a través de los estimadores

$$\hat{f}_l(x)=\hat{P}_k(X=x/Y=l)=\frac{k_l(x)}{N_l Vol_k(x)},\quad l=0,1 \tag{2.24}$$

y con estos estimadores construir el *estimador de la razón de verosimilitud* $\widehat{LR}(x)$ *en la vecindad* $V_k(x)$, $\widehat{LR}(x)=\dfrac{\hat{f}_1(x)}{\hat{f}_0(x)}$, $x\in\mathbb{R}^p$, con $f_0(x)>0$,

$$\widehat{LR}_k(x)=\frac{\dfrac{k_1(x)}{N_1 Vol_k(x)}}{\dfrac{k_0(x)}{N_0 Vol_k(x)}}=\frac{\dfrac{k_1(x)}{N_1}}{\dfrac{k_0(x)}{N_0}}=\left[\frac{\text{proporción del total de default en } V_k(x)}{\text{proporción del total de no default en } V_k(x)}\right] \tag{2.25}$$

Una vez obtenido el estimador k-NN para la razón de verosimilitud se puede construir un estimador, también por los k vecinos más próximos, para la probabilidad de default sustituyendo en (2.17) $\text{logit}(\hat{p})$ y $\widehat{LR}(x)$ respectivamente por (2.22) y (2.25). El estimador de máxima verosimilitud , EVM, de $logit(p)$ se obtiene de forma sencilla, pues al ser $\hat{p}=\frac{1}{N}\sum_{i=1}^{N} I_{\{y_i=1\}}$ el EVM de p, se tiene $logit(\hat{p})=logit\left(\frac{1}{N}\sum_{i=1}^{N} I_{\{y_i=1\}}\right)$. Por tanto

$$\text{logit}\left(\hat{P}(Y=1/\ X=x)\right)=logit\left(\frac{\sum_{i=1}^{N} I_{\{y_i=1\}}}{N}\right)+\log\left(\frac{\dfrac{k_1(x)}{N_1}}{\dfrac{k_0(x)}{N_0}}\right),\quad \forall x\in D\subset\mathbb{R}^p \tag{2.26}$$

Tras algunas sencillas operaciones algebraicas, (2.26) se puede expresar en la forma

$$\text{logit}\left(\hat{P}(Y=1/\ X=x)\right)=\log\left(\frac{N_1}{N_0}\right)+\log\left(\frac{N_0 k_1(x)}{N_1 k_0(x)}\right)=\log\left(\frac{k_1(x)}{k_0(x)}\right)$$

Por lo que *el estimador por los k vecinos más próximos de la probabilidad de default* viene dado por

$$\hat{P}(Y=1/\ X=x)=\frac{k_1(x)}{k} \tag{2.27}$$

dónde $k_1(x)=\sum_{x_i\in V_k(x)} I_{\{y_i=1\}}=\sum_{x_i\in V_k(x)} y_i$.

De (2.27) se sigue que *el estimador para la función de calificación de acreditados por los k vecinos más próximos en la vecindad* $V_k(x)$ viene dado por

$$\hat{S}_k(x)=\text{logit}\left(\frac{1}{k}\sum_{x_i\in V_k(x)} y_i\right) \tag{2.28}$$

Por otro lado, de (2.28) se sigue que *el estimador de la función de calificación de acreditados por los k vecinos más próximos en* x se puede expresar en función de los pesos de la evidencia en cada vecindad de x:

$$\hat{S}_k(x) = \text{logit}\left(\frac{1}{k}\sum_{x_i \in V_k(x)} y_i\right) = \log\left[\frac{\sum_{x_i \in V_k(x)} y_i}{k - \sum_{x_i \in V_k(x)} y_i}\right]$$

$$= \log\left[\frac{\text{número de default en } V_k(x)}{\text{número de no default en } V_k(x)}\right] = WOE\left(V_k(x)\right), \quad x \in \mathbb{R}^p \tag{2.29}$$

Podemos generalizar el estimador (2.23), expresándolo como un promedio ponderado en la forma siguiente

$$\hat{P}_k(X = x / Y = l) = \frac{1}{N_l}\sum_{i=1}^{N_l} w_{N_l i}(x) I_{\{y_i = 1\}}, \; l = 0,1 \tag{2.30}$$

donde

$$w_{N_l i}(x) = \frac{k_l(x)}{Vol_k(x)}, \; l = 0,1 \tag{2.31}$$

De este modo " la contribución o peso de un elemento de la vecindad $V_k(x)$, $w_{N_l i}(x)$, *a la densidad en* x *es mayor cuanto más cercano se encuentre de* x. Se puede ver que al tomar promedios ponderados de los *k* vecinos más próximos el método puede evitar el impacto de observaciones con ruido aislado.

Las ponderaciones impiden que incurramos en el riesgo de permitir que todas las observaciones de entrenamiento contribuyan a la estimación de las funciones de verosimilitud, ya que las muy distantes no tendrían peso asociado. La desventaja de considerar todas las observaciones seria su lenta respuesta (método global), por lo que es preferible un método local en el que solo se consideren los vecinos más próximos. Esta mejora es muy efectiva en muchos problemas prácticos. El estimador ponderado es robusto ante los ruidos de datos y suficientemente efectivo en conjuntos de datos grandes. Insistimos en que su principal problema es que es muy poco suave, por lo que sobreajusta los datos con el consabido problema de generalización. Por esta razón y por las pistas que proporcionan los pesos (2.31) como factores en (2.30) se generaliza el concepto de "*estimador por los k vecinos más próximos*" de la función de verosimilitud al estimador por "*vecinos más próximos*" según la siguiente definición:

Definición 2.2.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios de valores de $\mathbb{R}^p \times \mathbb{R}$ independientes e idénticamente distribuidos. Consideremos el siguiente estimador de la función de verosimilitud $P(X = x / Y = l), \; x \in \mathbb{R}^p$:

$$\hat{P}_{N_l}\left(X=x/Y=l\right)=\frac{1}{N_l}\sum_{i=1}^{N_l}w_{N_l i}(x)I_{\{y_i=l\}}=\frac{1}{N_l}\boldsymbol{W}_{N_l}^{T}(x)\boldsymbol{I}_{\{\boldsymbol{y=l}\}},\quad \forall x\in\mathbb{R}^p \tag{2.32}$$

donde $\boldsymbol{W}_{N_l}(x)=\left(w_{N_l 1}(x),...,w_{N_l N}(x)\right)^T$ es el vector de pesos y cada peso $w_{N_l i}(x)$ es una función Borel-medible de $X, X_1, X_2,..., X_N$ e $\boldsymbol{I}_{\{\boldsymbol{y=l}\}}=\left(I_{\{y_1=l\}},...,I_{\{y_i=l\}},...I_{\{y_N=l\}}\right)$. El *estimador de los vecinos más próximos de la función* $P(X=x/Y=l),\ x\in\mathbb{R}^p$ viene dado por la expresión (2.32) que depende de pesos convenientes asignados.

La *estimación de la función de verosimilitud por los vecinos más próximos* depende de los pesos asignados al estimador (2.32), que podemos expresar, por ejemplo, en la forma

$$w_{N_l i}(x)=\frac{m_h\left(x_i,x\right)}{Vol_h(x)} \tag{2.33}$$

donde $m_h\left(x_i,x\right)$ es una métrica que especifica la vecindad local con tamaño de ventana h. El estimador (2.32) adopta entonces la forma:

$$\hat{P}_{N_l}\left(X=x/Y=l\right)=\sum_{i=1}^{N_l}\frac{m_h\left(x_i,x\right)}{N_l Vol_h(x)}I_{\{y_i=l\}},\quad x\in\mathbb{R}^p \tag{2.34}$$

A pesar de que nuestro interés en la generalización del estimador dada por la definición 2.2 se orienta a la estimación por funciones núcleo, de la que resultan estimadores mucho más suaves que k-NN, y, por tanto, a la métrica que define vecindades apropiadas para ello, nos parece conveniente a efectos de la compresión de los métodos de estimación por núcleos, exponer aquí como se obtiene el estimador k-NN de la razón de verosimilitud y de la probabilidad de default a partir de la métrica que especifica las vecindades de los k vecinos más próximos.

El estimador k-NN de LOFTSGAARDEN y QUESENBERRY (1965) es un caso particular de (2.34) donde las vecindades vienen especificadas por la métrica

$$m_k\left(x_i,x\right)=I_{\left(\|x_i-x\|\le\|x_i-x_{(k)}\|\right)} \tag{2.35}$$

donde $x_{(k)}$ es el vecino de orden $k-ésimo$ en el ranking de distancias a x, e $I_{(A)}$ es el indicador del conjunto A. Es decir, $V_k(x)$ es el conjunto de los k puntos más próximos

a x en distancia Euclídea, o dicho de otro modo, la hiperesfera de los k vecinos más próximos a x.

Puesto que $I_{\left(\|x_i-x\|\leq\|x_i-x_{(k)}\|\right)}=I_{\{x_i\in N_k(x)\}}$, se tiene que el estimador (2.34) adopta la forma

$$\hat{P}_{N_l}\left(X=x/Y=l\right)=\sum_{i=1}^{N_l}\frac{I_{\{x_i\in V_k(x)\}}}{N_l Vol_k(x)}I_{\{y_i=l\}},\quad x\in\mathbb{R}^p \tag{2.36}$$

y puesto que $\sum_{i=1}^{N}I_{\{x_i\in V_k(x)\}}I_{\{y_i=l\}}=\sum_{x_i\in V_k(x)}y_i=k_l(x),\quad l=0,1$, se tiene

$$\hat{P}\left(X=x/Y=l\right)=\frac{k_l(x)}{N_l Vol_k(x)},\quad x\in\mathbb{R}^p \tag{2.37}$$

$k_l(x)$ indica la proporción de acreditados de la clase $y=l$ que se encuentran en la vecindad de los k vecinos más próximos a x, $V_k(x)$.

2.5.3 Estimación de la Razón de Verosimilitudes por Funciones Núcleo.

En este caso se parte de la equivalencia entre la expresión (2.11) y la siguiente

$$S(x)=\text{logit}(p)+\left\{\log\left(f_1(x)\right)-\log\left(f_0(x)\right)\right\},\quad x\in\mathbb{R}^p \tag{2.38}$$

y, por tanto, un estimador de $S(x)$ en un punto x viene dado por

$$\hat{S}(x)=\text{logit}(\hat{p})+\left\{\log\left(\hat{f}_1(x)\right)-\log\left(\hat{f}_0(x)\right)\right\},\quad x\in\mathbb{R}^p \tag{2.39}$$

Para completar el estimador de $S(X)$ en el punto x solo resta obtener los estimadores locales de las funciones de densidad $f_1(X)$ y $f_0(X)$ en este punto. Una vez obtenidos los valores $\hat{f}_1(x)$ y $\hat{f}_0(x)$ y el estimador de máxima verosimilitud $\hat{p}$, del mismo modo que para $k-NN$ se calcula el valor $\hat{S}(x)=\text{logit}(\hat{p})+\left\{\log\left(\hat{f}_1(x)\right)-\log\left(\hat{f}_0(x)\right)\right\},\ x\in\mathbb{R}^p$ obteniendo finalmente el estimador local por funciones núcleo de la probabilidad de default en el punto x: $\hat{P}\left(Y=1/X=x\right)=\Lambda\left(\hat{S}(x)\right)=\frac{1}{1+e^{-\hat{S}(x)}}$.

La técnica de estimación de la densidad más popular desde el enfoque no paramétrico ha evolucionado desde los vecinos más próximos de FIX y HODGES (1951), pasando por el *histograma,* la primitiva herramienta de síntesis para visualizar la función de densidad, y por el *estimador ingenuo de ROSEMBLATT*

(1956) hasta llegar a los *métodos de estimación por núcleos*, PARZEN (1962), más complejos y sofisticados.

Veremos a continuación como han ido evolucionando los métodos de *estimación de la densidad por vecinos próximos,* considerando las métricas que definen las vecindades locales apropiadas para obtener el estimador de las verosimilitudes.

Un popular y notable ejemplo de métrica que especifica una vecindad local, en el caso en que el vector de variables explicativas sea unidimensional, viene dada por

$$m_h(x, x_i) = I_{|x_i - x| \le h} \text{ , con } \quad x, x_i \in B_m \tag{2.40}$$

donde B_m es un intervalo de $\mathbb{R}$. Dado que $\forall x, x_i \in B_j$ *de amplitud* h se verifica $I_{|x_i - x| \le h} = I_{B_j}(x) = I_{(x_i \in B_j)} I_{(x \in B_j)}$, se tiene

$$m_h(x, x_i) = I_{B_j}(x) = I_{(x_i \in B_j)} I_{(x \in B_j)} \text{ , con } \quad x, x_i \in B_m \tag{2.41}$$

La métrica (2.40) da origen al popular y "humilde" *Histograma,* estimador por vecinos más próximos de las funciones de verosimilitud $P(X = x / Y = l)$ en una vecindad de ventana h, (parámetro de localización y suavización que controla el ancho de los intervalos del histograma), que podemos obtener en la forma siguiente:

$$\begin{aligned} \hat{P}_h(X = x / Y = l) &= \frac{1}{N_l h} \sum_{i=1}^{N} m_h(x_i, x) I_{\{y_i = l\}}, \quad \forall x \in \mathbb{R}, \quad l = 0,1 \\ &= \frac{1}{N_l h} \sum_{i=1}^{N} \sum_{m=1}^{q} I_{(x \in B_m)} I_{(x_i \in B_m)} I_{\{y_i = l\}} \\ &= \sum_{m=1}^{q} \left(\frac{1}{N_l h} \sum_{i=1}^{N} I_{(x_i \in B_m)} I_{\{y_i = l\}} \right) I_{(x \in B_m)} = \sum_{m=1}^{q} \hat{\beta}_m I_{(x \in B_m)} \end{aligned} \tag{2.42}$$

dónde $\hat{\beta}_m = \frac{1}{N_k h} \sum_{i=1}^{N} I_{(x_i \in B_m)} I_{\{y_i = k\}}$. Se pone el subíndice h en la notación del estimador para explicitar la dependencia del mismo de la amplitud del intervalo h, a la vez que también depende del origen elegido.

La vertiente más conocida del histograma es su utilidad como *herramienta de análisis y visualización de datos*. Este aspecto resulta especialmente valioso cuando contamos con gran número de datos y se desea tener una primera idea visual rápida acerca de su estructura.

El segundo aspecto, menos popular quizá, está relacionado con la inferencia: *el histograma es en realidad el estimador no paramétrico de la función de densidad más sencillo que puede utilizarse para reemplazar a los modelos paramétricos usuales (y, en particular, al omnipresente modelo normal) cuando hay razones para dudar de ellos*.

Los resultados del caso univariante se pueden extender al caso multivariante y es relativamente sencillo probar que el *estimador histograma* es un estimador consistente de las funciones de verosimilitud, lo que podría convertirlo, en principio, en una herramienta muy útil para estimar las verosimilitudes y, por tanto, la probabilidad de default.

Consideremos, por ejemplo, la edad de los acreditados de una Entidad Financiera

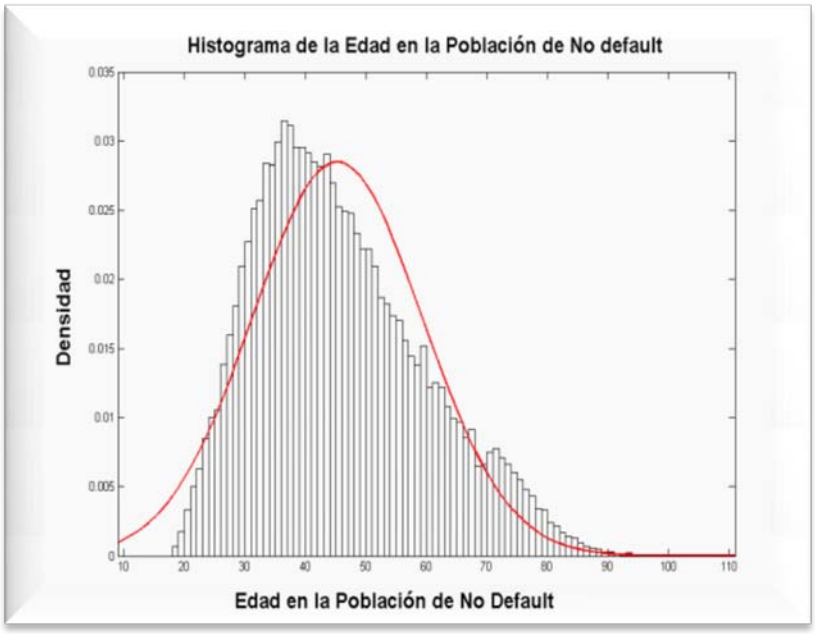


Figura 2.2. Histograma versus Campana de Gauss.

La simple representación de la edad de los acreditados mediante los clásicos histogramas, figura 2.2, resulta muy significativa.

El histograma presenta grandes inconvenientes entre los que destacan:

1) Los resultados dependen del origen.
2) La elección de la amplitud de los subintervalos es discrecional.
3) Un histograma es siempre discontinuo por naturaleza, puesto que se obtiene a través de una función constante sobre cada subintervalo

(función escalera), es decir, continua a trozos cuando muchas veces la función a estimar $f(x)$ es continua.

El efecto del ancho de ventana h y la dependencia del origen se observa en la figura 2.3, donde se representan 4 histogramas de la distribución de la edad de los prestatarios incumplidores confeccionados sobre una población de acreditados clasificados como default para diferentes valores de h, $h=0,2$ (ancho de intervalo óptimo, con la "regla de a dedo"), $h=1,8$, $h=3,2$ y $h=7,2$, todos ellos con el mismo origen $x_0=18$, edad mínima requerida a nuestros prestatarios. Como puede observase en la figura 2.3 los cuatro histogramas tienen una estructura diferente.

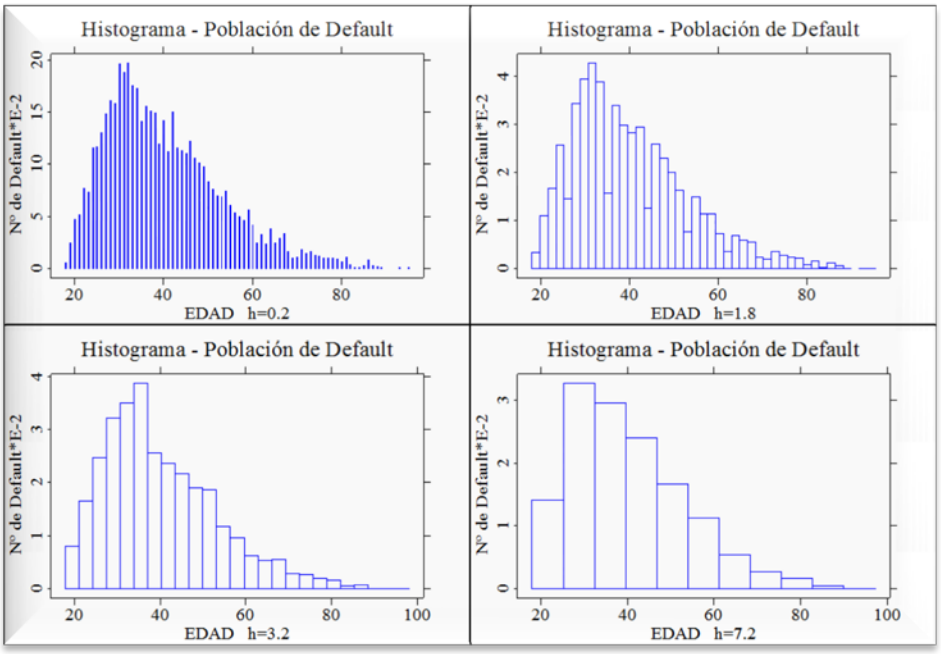


Figura 2.3.- Histogramas de la misma distribución con diferente h.

A pesar de los esfuerzos de investigación realizados por numerosos autores para encontrar un método objetivo que proporcione el ancho de ventana óptimo, es muy clarificador observar que, aún siendo muy importantes los avances registrados desde mediados de la década de 1950, que han amortiguado bastante el problema de la elección de la amplitud, aún hoy en día el método más utilizado se llama "regla de a dedo".

Para remediar el hecho de que orígenes diferentes proporcionen histogramas diferentes, SCOTT (1985), propuso el método "*Average Shifted Histograms*", ASH. La idea básica consiste en promediar varios histogramas desplazados resultando

otro histograma más suavizado. Scott, considerando una colección de m histogramas $\hat{f}_1, \hat{f}_2, ..., \hat{f}_m$ con el mismo ancho de banda h, pero con diferentes orígenes para los intervalos $x_0 = 0, \frac{h}{m}, \frac{2h}{m}, ..., \frac{(m-1)}{m}$, respectivamente, definió el *estimador ASH* como

$$\hat{f}_h(x) = \frac{1}{m}\sum_{k=0}^{m-1} \hat{f}_{h,k}(x) \tag{2.43}$$

expresión, que se desarrolla en la forma siguiente:

$$\hat{f}_h(x) = \frac{1}{m}\sum_{k=0}^{m-1} \hat{f}_{h,k}(x) = \frac{1}{m}\sum_{k=0}^{m-1}\left\{\frac{1}{Nh}\sum_{i=1}^{N}\left(\sum_j I_{(x_i \in B_{jk})} I_{(x \in B_{jk})}\right)\right\} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{mh}\sum_{k=0}^{m-1}\left(\sum_j I_{(x_i \in B_{jk})} I_{(x \in B_{jk})}\right)\right\}$$

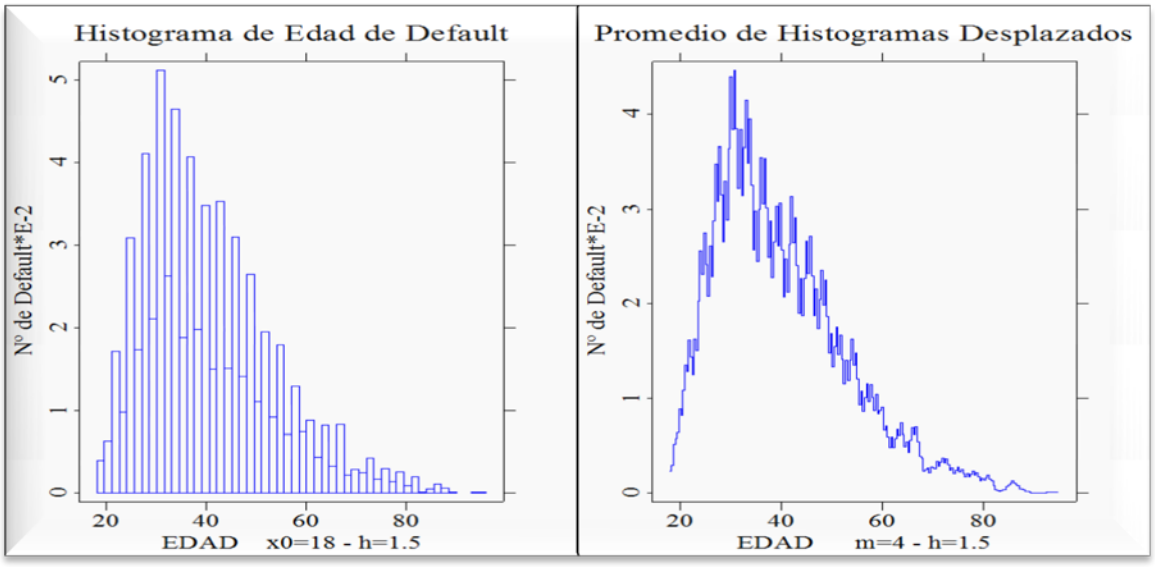


Figura 2.4.- Comparación de un histograma ordinario con un promedio de histogramas desplazados.

m es el número de desplazamientos.

Si nos fijamos en la figura 2.4, los promedios desplazados parecen ser histogramas con un ancho de banda menor que el que corresponde al histograma de partida, pero es importante resaltar que el histograma promedio de histogramas desplazados no se corresponde conceptualmente con un histograma ordinario con menor ancho de banda.

Los histogramas ASH dependen de m y nótese que cuando $m \to \infty$ el histograma no depende del origen, es decir, la función "escalera" discontinua converge a una

función continua. Este comportamiento asintótico puede alcanzarse directamente por una técnica diferente: *la estimación de la densidad por núcleos*.

A pesar de la importante mejora introducida por SCOTT (1985), la continuidad a trozos de los histogramas, incluido el ASH, implica que su primera derivada es cero en casi todo punto, esto los hace completamente inadecuados para estimar la derivada de la función de densidad.

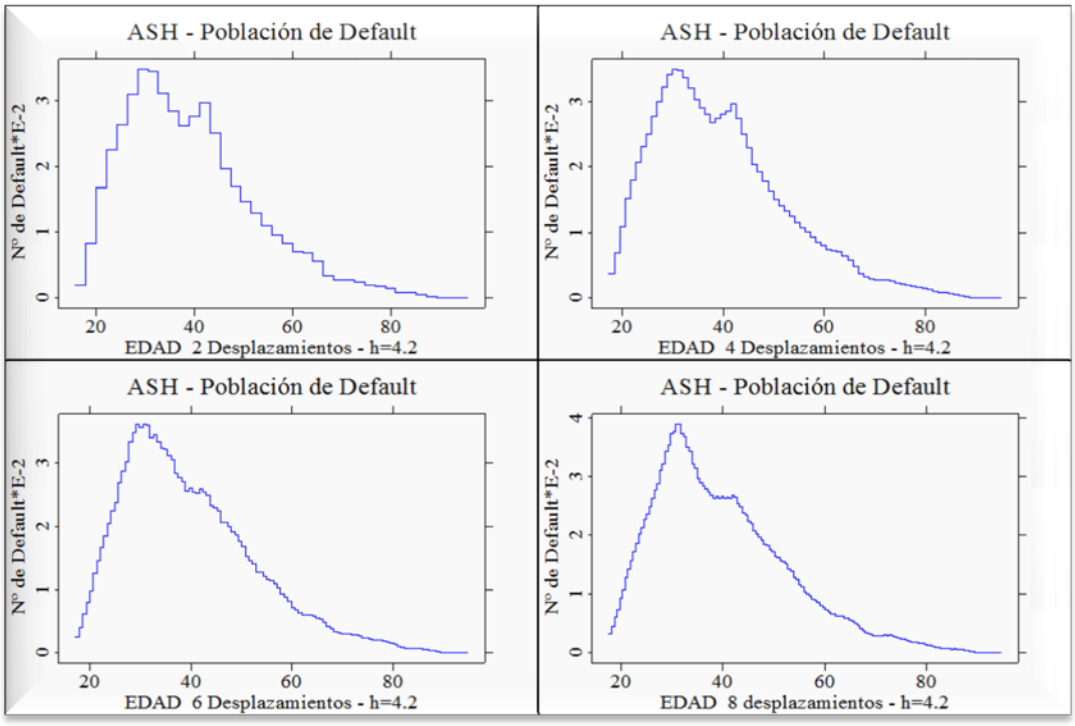


Figura 2.5.- Histogramas ASH con el mismo h y distintos desplazamientos.

Con la intención de resolver el problema de discontinuidad de los *histogramas*, suavizando la estimación obtenida por el histograma ordinario, ROSEMBLATT (1956) introdujo, tres décadas antes que SCOTT (1985), el *histograma móvil* (también conocido como *histograma ingenuo*) como un estimador natural de la relación entre la función de densidad y la función de distribución:

$$f(x) = F_X'(x) = \lim_{h\to 0} \frac{F_X(x+h) - F_X(x-h)}{2h} \tag{2.44}$$

Rosemblatt definió su *estimador ingenuo* $\hat{f}_h$ *de* $f(x)$, como

$$\hat{f}_h(x, x_1, .., x_N) = \lim_{h\to 0} \frac{\hat{F}_N(x+h) - \hat{F}_N(x-h)}{2h} \tag{2.45}$$

donde la distribución empírica $\hat{F}(x)$, viene dada por $\hat{F}(x) = F_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{x_i \le x\}}$

La definición del estimador de Rosemblatt está motivado por el hecho de que $f(x) = \lim_{h\to 0} \frac{1}{2h} P(x-h < X \le x+h)$. Este estimador es discontinuo en $x_i \pm h$, por lo que también da estimaciones continuas a trozos, aunque estas son más suaves que en el histograma ordinario y es el punto de partida usado por PARZEN (1962) para sus populares estimadores tipo núcleo, pues este observó que dado que $\frac{F_n(x+h)-F_N(x-h)}{2h} = \frac{1}{2hN}\sum_{1}^{N} I_{(x-h<x_i\le x+h)} = \frac{1}{2hN}\sum_{1}^{N} I_{(-1\le \frac{x-x_i}{h}<1)}$, el estimador ingenuo de Rosemblatt se puede expresar como

$$\hat{f}_h(x, x_1, .., x_n) = \frac{1}{Nh}\sum_{i=1}^{N} w\left(\frac{x - x_i}{h}\right) \tag{2.46}$$

donde $w(z) = \begin{cases} \frac{1}{2} & si \quad z \in [-1,1) \\ 0 & en \quad otro\ caso \end{cases}$, es decir, w(z) es la densidad de una variable aleatoria con distribución uniforme en el intervalo [-1,1).

Nótese que el estimador de Rosemblatt en vez de partir el intervalo de interés en intervalos fijos, considera una amplitud o ancho de ventana fijo $2h$, centrado en el punto en el que se desea realizar la estimación, de hecho el estimador (2.45) puede ser considerado una modificación del histograma ordinario donde cada punto es el centro del un intervalo de anchura $2h$.

Por último, se demuestra fácilmente que el *estimador de Rosemblatt* $\hat{f}_h(x)$ *es un estimador insesgado, y consistente, es decir, sesgo de* $\left(\hat{f}_h(x)\right) = E\left[\left(\hat{f}_h(x)\right) - f_h(x)\right] = 0$ *y* $\hat{f}_h(x) \xrightarrow{P} f(x)$ *en todo punto de continuidad de* $f(x)$.

Como puede observarse en la figura 2.6, incluso la versión más sofisticada de la clásica estimación por histogramas, el *histograma de Rosemblatt*, a pesar de ser un estimador consistente, en algunos ejemplos reales no es lo suficientemente suave, pues no deja de ser una función continua a trozos, al tratarse de una combinación lineal de funciones lineales a trozos

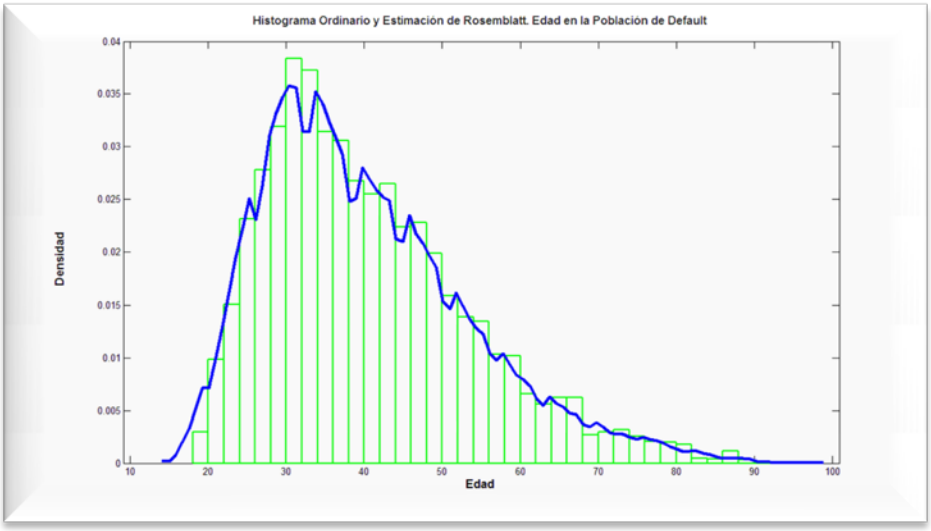


Figura 2.6.- Comparación del histograma ordinario con el histograma de Rosemblatt.

PARZEN (1962), a partir de los trabajos de ROSEMBLATT (1956) y de WHITTLE (1958), desarrolló una aproximación que resuelve las importantes dificultades que presentan los histogramas como estimadores de las funciones de densidad, entre ellas la continuidad a trozos. Parzen resolvió el problema a través de sus populares *estimadores de la densidad por funciones núcleos*. "*De este modo Parzen, sin pretenderlo, resolvía el problema de la rugosa y bacheada apariencia del estimador* $k-NN$ *", HASTIE et al. (2009).*

2.5.3.1 Funciones Núcleo Univariantes.

Definición 2.3.- **(Parzen (1962))**. Sea X una variable aleatoria unidimensional y $\{X_i\}_{i=1,\ldots,N}$ una muestra aleatoria simple de X. Se define el *estimador núcleo de función núcleo* $K(\bullet)$ *de la función de densidad desconocida* $f(x)$ como

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{2.47}$$

donde la función

$$\begin{aligned} K(\bullet) &: \mathbb{R} \to \mathbb{R} \\ t &\to K(t) \end{aligned} \tag{2.48}$$

es una función de densidad prefijada, llamada *función núcleo univariante de ventana* h, que verifica ciertas condiciones de regularidad (2.49), (generalmente es una función de densidad simétrica como, por ejemplo, la distribución normal), y h es un parámetro de suavizado, llamado ventana o amplitud de banda que debe

tender a cero lentamente, ($\lim_{N\to\infty} h = 0$, $\lim_{n\to\infty} Nh = \infty$), para poder asegurar que $\hat{f}_h$ tiende a la verdadera densidad f.

La estimación de la densidad por núcleos no es más que un promedio ponderado por la distancia de las observaciones al punto a ser estimado. El peso lo determinarán la función núcleo elegida y el valor de h. Cuanto mayor sea el valor de h, mayor será el peso de aquellos elementos de las observaciones que se encuentren alejados del punto, es por esto que a h se le llama generalmente "*ancho de banda o ventana*".

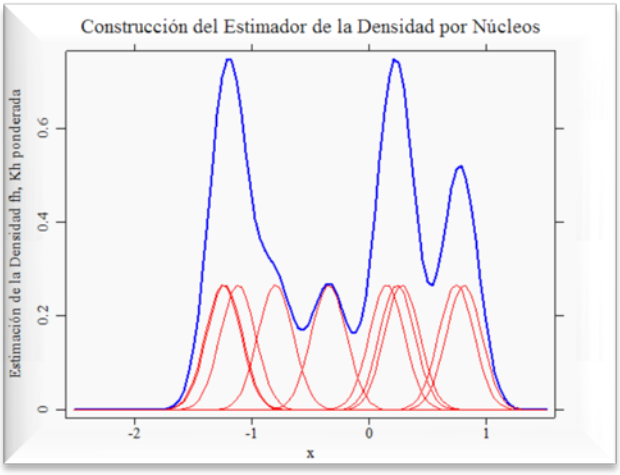


Figura 2.7.- Estimador de la densidad por núcleos.

Tal como se observa en la figura 2.7, el estimador núcleo puede interpretarse como una suma de "protuberancias" situadas en las observaciones.

La función núcleo $K(\bullet)$ determina la forma de las protuberancias mientras que el parámetro h determina su anchura. Cada protuberancia está centrada en una de las observaciones. En cada observación x se calcula $\hat{f}_h(x)$ sumando verticalmente las protuberancias. Obviamente, diferentes valores de h cambian la apariencia de las protuberancias y, como, consecuencia, la apariencia de su suma. La ventana h determina la cantidad de suavización de la estimación, siendo el límite cuando h tiende a cero una suma de funcionales delta de Dirac en los puntos de las observaciones.

El siguiente teorema nos proporciona las condiciones que ha de cumplir la función núcleo para que el estimador definido en (2.47) sea asintóticamente insesgado y consistente:

Teorema.2.1.- S*i la función* $K(t)$ *satisface:*

$$|K(t)| < \infty \qquad \forall t \ (K(t) \text{ acotada}$$

$$K(-t) = K(t) \quad \forall t \ (K(t) \text{ simétrica})$$

$$\int_{-\infty}^{+\infty} |K(t)| \, dt < \infty \ (K(t) \text{ absolutamente integrable}) \tag{2.49}$$

$$|tK(t)| \xrightarrow[|t|\to\infty]{} 0 \Rightarrow \int_{-\infty}^{+\infty} tK(t)dt = 0$$

$$\int_{-\infty}^{+\infty} K(t)dt = 1$$

entonces el estimador definido en (2.47) *es asintóticamente insesgado y consistente en todos los puntos x en los cuales la función de densidad de probabilidad es continua, es decir,* $E\left[\hat{f}_h(x)\right] = f_h(x) \xrightarrow[h\to 0]{} f(x)$ *y* $V\left[\hat{f}_h(x)\right] = \frac{1}{Nh} f_h'(x) \xrightarrow[\substack{h\to 0 \\ Nh\to\infty}]{} 0$.

En la tabla 2.1 se muestra una relación de las funciones núcleo más usuales así como de sus rangos y en la figura 2.8 se muestran las representaciones gráficas correspondientes.

Los problemas más importantes que actualmente se plantean en la estimación no paramétrica de densidades por núcleos son, por un lado, que la elección de la amplitud de ancho de banda es discrecional y, por otro, encontrar la función núcleo adecuada. El primero es el punto más complicado, si el parámetro de suavizado se elige demasiado pequeño, el estimador aparece *infrasuavizado*, e incorpora demasiado *ruido*, reflejado en la presencia de muchas modas (máximos relativos) *espúreas* que, de hecho no aparecen en la densidad que se quiere estimar. Por el contrario, si h se elige demasiado grande, se da el fenómeno contrario, de *sobresuavizació*n y el estimador es casi insensible a los datos.

Si nos fijamos en la figura 2.9, donde se estima la función de densidad para ciertos datos con el mismo ancho de banda $h = 5$, se observa de modo intuitivo la propiedad de herencia de las condiciones de continuidad y diferenciabilidad entre los estimadores de núcleo y las funciones núcleos utilizadas:

Tabla 2.1.- Funciones núcleo más notables.

Funciones Núcleo más notables		
Núcleo	K(t)	Rango
Rectangular	$\frac{1}{2a}$	$(-a < t < a)$
Uniforme	$\frac{1}{2}$	$(-1 < t < 1)$
Gaussiano	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)$	$(-\infty < t < \infty)$
Triangular	$1-\lvert t\rvert$	$(-1 < t < 1)$
Epanechnikov	$\frac{3}{4}(1-t^2)$	$(-1 < t < 1)$
Biweight	$\frac{15}{16}(1-t^2)^2$	$(-1 < t < 1)$
Triweight	$\frac{35}{32}(1-t^2)^3$	$(-1 < t < 1)$

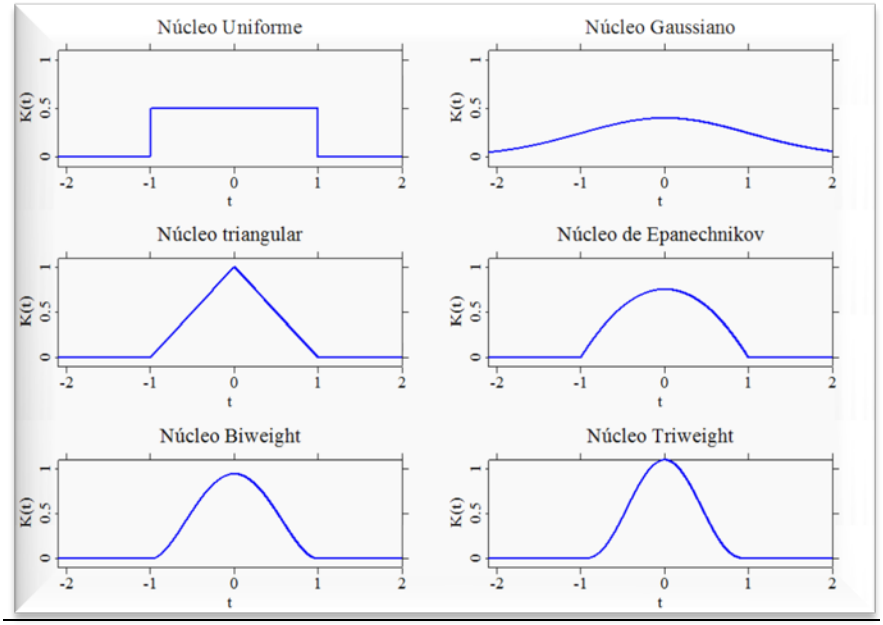


Figura 2.8.- Representación de las funciones núcleo de la tabla 2.1.

Las estimaciones basadas en los núcleos Uniforme y Triangular reflejan la forma de las funciones núcleo subyacentes con sus aspectos mellados, sobre todo el núcleo Uniforme.

Por otro lado, los estimadores que utilizan los núcleos Gaussiano, Biweight y Triweigth tienen representaciones gráficas más suaves y continuas.

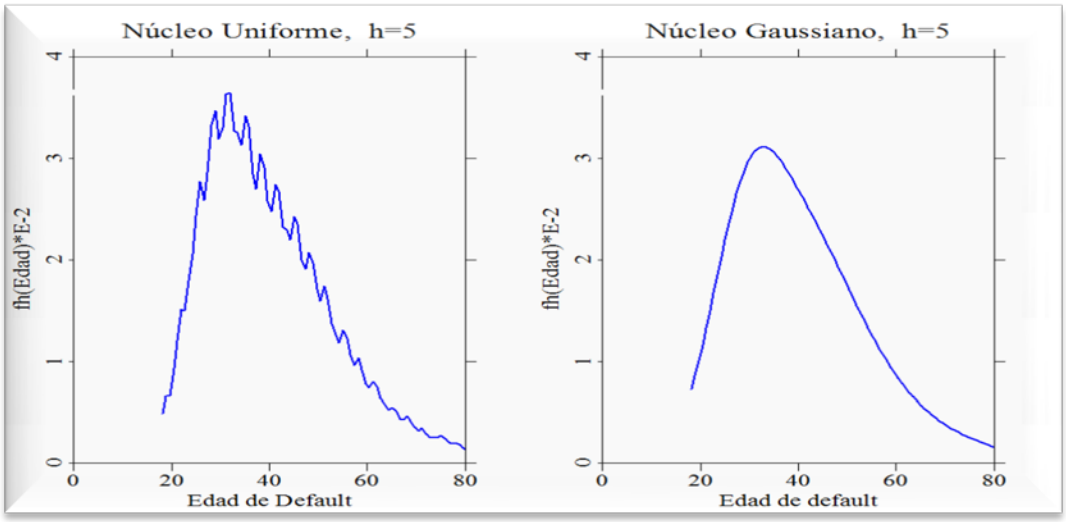


Figura 2.9.- Herencia de las condiciones de continuidad y diferenciabilidad (1).

Las diferencias no se observan solamente cuando las estimaciones están basadas en núcleos continuos o no continuos, puesto que si observamos en la figura 2.10 la estimación con núcleo de Epanechnikov, continuo, difiere más de la basada en el núcleo Biweigth, continuo, que en el triangular que es discontinuo.

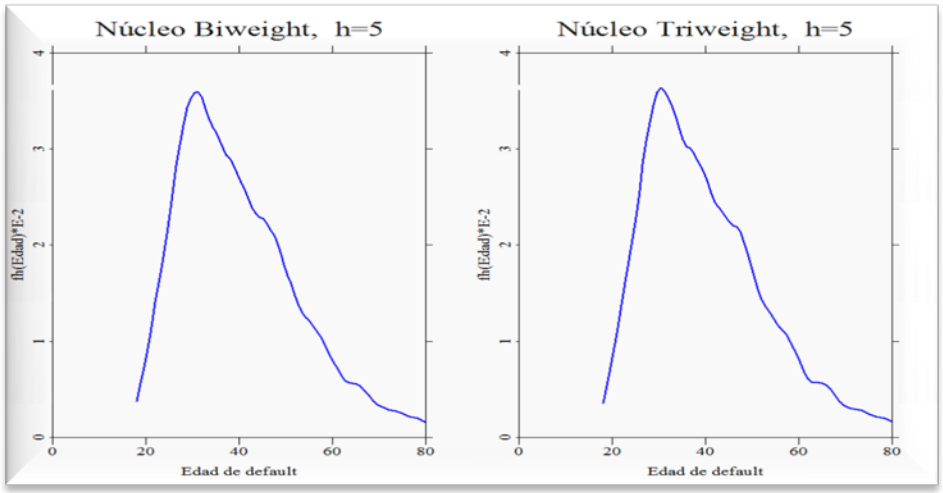


Figura 2.10.- Herencia de las condiciones de continuidad y diferenciabilidad (2).

Nos encontramos, por tanto, ante un dilema, hemos de buscar un ancho de ventana, h, óptimo, pero este óptimo no garantiza el mismo grado de suavizado si se usan diferentes funciones núcleo. Desgraciadamente *no existe el mejor método para seleccionar los anchos de banda de los estimadores de densidad por núcleos;* incluso criterios asintóticamente óptimos suelen mostrar un mal funcionamiento en estudios de simulación. *Como consecuencia* (HÄRDLE et al., 2004a) *recomiendan, determinar anchos de banda con diferentes métodos de selección y comparar las estimaciones de densidad resultantes.*

Sin ninguna duda, la principal característica de los estimadores por funciones núcleos se describe perfectamente en el siguiente párrafo, entresacado de HASTIE et al. (2009), "el estimador $k-NN$ está lleno de baches y se prefiere el estimador "suave" de Parzen, a causa de que este estimador cuenta observaciones estrechamente cercanas a x con pesos que decrecen con la distancia a x". El estimador de la densidad de Parzen es el equivalente al promedio local, y se ha ido mejorando, sobre todo en la línea de la regresión local. Esta técnica se ha ido afianzando para estimar por métodos de vecinos próximos funciones de densidad condicional de variables unidimensionales, puesto que resuelve el problema razonablemente bien, SILVERMAN (1986) y SCOTT (1992), HASTIE et al (2009).

La métrica
$$m_h\left(x, x_i\right) = K\left(\frac{x_i - x}{h}\right) \tag{2.50}$$
define la vecindad de los vecinos próximos que se utiliza en la estimación local de una densidad univariante mediante *funciones núcleo univariantes* con ventana h. Por tanto el estimador de las verosimilitudes por vecinos próximos (2.32) con la métrica (2.50) se formula en los siguientes términos:

$$\hat{P}_{N_l}\left(X = x / Y = l\right) = \frac{1}{N_l h}\sum_{i=1}^{N_l} K\left(\frac{x - x_i}{h}\right) I_{\{y_i = l\}}, \quad \forall x \in \mathbb{R}^p \tag{2.51}$$

La estimación por núcleos de Parzen de densidades univariantes es, por un lado, una estimación suave y eficaz, y, por otro, sencilla de calcular, pero a efectos de cálculo, la estimación multivariante, a la que dedicaremos el último apartado de esta sección, es harina de otro costal.

2.5.3.2 Funciones Núcleo Multivariantes.

Si definimos la vecindad local a través de la métrica

$$m_{h_1,\ldots,h_p}(x,x_i)=\frac{1}{\prod_{j=1}^{p}h_j}K\left(\frac{x_{i1}-x_1}{h_1},\ldots,\frac{x_{ip}-x_p}{h_p}\right) \qquad (2.52)$$

$$\text{con ventana } \boldsymbol{h}=(h_1,\ldots,h_p)$$

donde la función multivariante

$$\begin{aligned} K(\bullet):\mathbb{R}^p &\to \mathbb{R} \\ t &\to K(t)=K(t_1,\ldots,t_p) \end{aligned} \qquad (2.53)$$

es una función de densidad prefijada, llamada *función núcleo multivariante de ventana* $\boldsymbol{h}=(h_1,\ldots,h_p)$, que verifica ciertas condiciones de regularidad, entonces el estimador (2.32) con la métrica (2.52) adopta la forma:

$$\hat{P}_{N_l}(X=x/Y=l)=\frac{1}{N_l}\sum_{i=1}^{N_l}\frac{1}{\prod_{j=1}^{p}h_j}K\left(\frac{x_1-x_{i1}}{h_1},\ldots,\frac{x_p-x_{ip}}{h_p}\right)I_{\{y_i=l\}},\quad x\in\mathbb{R}^p \qquad (2.54)$$

donde la función núcleo es una función de densidad prefijada que verifica ciertas condiciones de regularidad, generalmente es simétrica como, por ejemplo, la distribución normal, y h es un parámetro de suavizado, llamado ventana o amplitud de banda que debe "tender a cero lentamente", $\left(\lim_{N\to\infty}h=0,\ \lim_{n\to\infty}Nh=\infty\right)$, para poder asegurar que $\hat{P}_{N_l}(X=x/Y=l)$ tiende a la verdadera densidad $P(X=x/Y=l)$.

Sobre la forma que ha de tener la función multidimensional $K(t)=K(t_1,\ldots,t_p)$, la opción elegida usualmente, es el núcleo multiplicativo $K(t)=\prod_{j=1}^{p}K(t_j)$, con K función núcleo univariante, en este caso

$$m_{h_1,\ldots,h_p}(x,x_i)=\prod_{j=1}^{p}\frac{1}{h_j}K\left(\frac{x_{ij}-x_j}{h_j}\right),\quad \textit{con ventana } \boldsymbol{h}=(h_1,\ldots,h_p) \qquad (2.55)$$

y entonces la expresión (2.54) se transforma en

$$\hat{f}_l(x) = \hat{P}_{N_l}\left(X = x / Y = l\right) = \frac{1}{N_l}\sum_{i=1}^{N_l}\left\{\prod_{j=1}^{p}\frac{1}{h_j}K\left(\frac{x_j - x_{ij}}{h_j}\right)\right\} \tag{2.56}$$

Para núcleos univariantes con soporte $[-1,1]$, (como el núcleo Epanechnikov $\kappa(t) = 3/4(1-t^2)I(|t| \le 1))$, se usan las observaciones en un cubo alrededor de t para estimar la densidad en el punto t. Este tipo de núcleos se llaman *esféricos o radiales simétricos* puesto que $\kappa(t)$ tiene el mismo valor para todo t sobre una esfera alrededor de cero.

Una aproximación muy general es usar una matriz ventana (no singular) $\boldsymbol{H}$. La forma general para el estimador de densidad multivariante es entonces, SILVERMAN (1986) y SCOTT (1992),

$$\hat{f}_{\boldsymbol{H}}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{det(\boldsymbol{H})}\kappa\left\{\boldsymbol{H}^{-1}(x - x_i)\right\} = \frac{1}{N}\sum_{i=1}^{N}\kappa_{\boldsymbol{H}}(x - x_i)$$

$$donde \quad \kappa_{\boldsymbol{H}}(\bullet) = \frac{1}{det(\boldsymbol{H})}\kappa\left\{\boldsymbol{H}^{-1}\bullet\right\} \tag{2.57}$$

Las propiedades asintóticas y la selección de ventanas son análogas al caso unidimensional, pero más complejas. Además la "maldición de la dimensionalidad" se cierne sobre este estimador multivariante. Un problema especial lo constituye su visualización gráfica; intersecciones de baja dimensionalidad, proyecciones o gráficos de contorno pueden visualizar solo parte de las características de la función de densidad. Por otro lado los tiempos de computación se disparan creciendo rápidamente con el número de variables a considerar en el análisis.

Los estimadores por núcleos de las funciones de densidad conjunta de las poblaciones de default, $f_1(x)$, y no default, $f_0(x)$, tomando como núcleo, por ejemplo, el producto de p núcleos univariantes de Epanechnikov, adoptan la siguiente forma:

$$\hat{f}_{lh_l}(x) = \frac{1}{N_l}\sum_{i=1}^{N_l}\left\{\prod_{j=1}^{p}\frac{1}{h_{lj}}\kappa\left(\frac{x_j - x_{ij}}{h_{lj}}\right)\right\} = \frac{1}{N_k}\sum_{i=1}^{N_l}\left\{\prod_{j=1}^{p}\frac{1}{h_{lj}}\frac{3}{4}\left(1 - \left(\frac{x_j - x_{ij}}{h_{lj}}\right)^2\right)I\left(\left|\frac{x_j - x_{ij}}{h_{lj}}\right| \le 1\right)\right\} \tag{2.58}$$

$$\forall x \in D \subset \mathbb{R}^p, \quad l = 1,0$$

Una vez obtenidos los estimadores $\hat{f}_1(x) = \hat{f}_{1h_1}(x)$ y $\hat{f}_0(x) = \hat{f}_{0h_0}(x)$ por funciones núcleo y el estimador de máxima verosimilitud $\hat{p}$, el estimador de $P(Y=1/X=x)$ se obtiene en la forma $\hat{P}(Y=1/X=x) = \left(1+\exp\left(-\text{logit}(\hat{p}) + \left\{\log\left(\hat{f}_1(x)\right) - \log\left(\hat{f}_0(x)\right)\right\}\right)\right)^{-1}$.

Este sería un excelente método para estimar la probabilidad de default si no fuera por que cuando el número de variables p es elevado la razón de convergencia es muy lenta comparada con el caso univariante, por lo que el esfuerzo computacional de esta técnica crece con el número de dimensiones, "*la maldición de la dimensionalidad*". Es, por tanto, virtualmente imposible estimar la densidad de probabilidad de default a través de funciones núcleo sin "aceptar algunas hipótesis estructurales", de ahí que este método no suela aplicarse cuando $p \geq 5$, lo que lo hace poco eficaz para nuestros propósitos.

Una de las hipótesis estructurales que se ha propuesto y que no es raro ver aplicada en trabajos prácticos presentados en artículos relacionados con los métodos de aprendizaje, es la llamada "*hipótesis ingenua de Bayes*". Esta hipótesis, con el fin de evitar la maldición de la dimensionalidad que se cierne sobre la aplicación de las funciones núcleo multivariantes, asume que *los elementos del vector aleatorio* X *condicionados a las poblaciones de default y no default son independientes entre sí.* De este modo, con la hipótesis ingenua el problema de estimar la función de calificación de acreditados se reduce a un problema de estimación de densidades por funciones núcleos univariantes.

El método más popular basado en la hipótesis ingenua de Bayes se conoce como *Clasificador Ingenuo de Bayes*, BNC. La formalización de la hipótesis ingenua de Bayes viene dada por la expresión

$$f_l(x) = f_{X/Y=l}(x) = \prod_{j=1}^{p} f_{X_j/Y=l}(x_j) = \prod_{j=1}^{p} f_{lj}(x_j), \quad l = 0,1 \tag{2.59}$$

de donde $\log(f_1(x)) - \log(f_0(x)) = \sum_{j=1}^{p}\left\{\log\left(f_{1j}(x_j)\right) - \log\left(f_{0j}(x_j)\right)\right\}$, $j = 1,..,p$, y, por tanto,

$$S(X) = logit(p) + \sum_{j=1}^{p}\left\{\log\left(f_{1j}(X_j)\right) - \log\left(f_{0j}(X_j)\right)\right\} \tag{2.60}$$

de donde el estimador local de $S(X)$ en el punto x_0 se expresa, bajo la hipótesis ingenua, en la forma

$$\hat{S}(x_0) = \text{logit}(\hat{p}) + \sum_{j=1}^{p} \left\{ \log\left(\hat{f}_{1j}(x_{0j})\right) - \log\left(\hat{f}_{0j}(x_{0j})\right) \right\}, \quad x_0 \in \mathbb{R}^p \tag{2.61}$$

Es decir, el estimador local de la función de calificación en el punto x_0 se obtiene, en este caso, sin más que estimar las funciones $\left\{ f_{1j}(x_j) \right\}_{j=1,\ldots,p}$ y $\left\{ f_{0j}(x_j) \right\}_{j=1,\ldots,p}$ en este punto por funciones núcleo univariantes de ventana h, $K\left(\dfrac{x - x_0}{h}\right)$, según la definición 2.3:

$$\hat{f}_{lj}(x_{0j}) = \frac{1}{N_l h} \sum_{i=1}^{N_l} K\left(\frac{x_{ij} - x_{0j}}{h}\right), \quad j = 1,\ldots,p, \; l = 0,1, \; x_{ij} \in V_h(x_{0j}) \tag{2.62}$$

y sustituir en (2.61).

En la construcción de los modelos scoring el nombre de *hipótesis ingenua* encaja perfectamente, por cuanto la independencia condicional de las variables observadas sobre los acreditados en relación con su comportamiento frente al default es "prácticamente inexistente".

Una hipótesis más razonable que la anterior es la debida a LARSEN (2005) que desarrolló el llamado *Clasificador Ingenuo de Bayes Generalizado,* GNBC. La propuesta de Larsen consiste en relajar la hipótesis de independencia en la forma siguiente

$$\log\left(\frac{f_1(x)}{f_0(x)}\right) = \sum_{j=1}^{p} \left(\log\left(f_{1j}(x_j)\right) - \log\left(f_{0j}(x_j)\right) + b_j(x_j) \right) \tag{2.63}$$

donde los $b_j(x_j)$ representan el sesgo marginal atribuido al efecto ingenuo $\log\left(\dfrac{f_{1j}(x_j)}{f_{0j}(x_j)}\right)$, que nos indica cómo cambia el efecto de x_j en presencia del resto de variables explicativas, lo que es una valiosa aportación de la técnica en la línea de los requerimientos de Basilea II.

De este modo se descompone el sesgo total en p términos, donde el j-ésimo término es una función de X_j, corrigiéndose los sesgos causados por la suposición ingenua de independencia condicional. En este caso se tiene

$$S(X)=\text{logit}(p)+\sum_{j=1}^{p}\left\{\log\left(f_{1j}(X_j)\right)-\log\left(f_{0j}(X_j)\right)+b_j(X_j)\right\} \tag{2.64}$$

Por lo que el estimador local de $S(X)$ en el punto x_0 se expresa

$$\hat{S}(x_0)=\text{logit}(\hat{p})+\sum_{j=1}^{p}\left\{\log\left(\hat{f}_{1j}(x_{0j})\right)-\log\left(\hat{f}_{0j}(x_{0j})\right)+\hat{b}_j(x_{0j})\right\},\ x_0=(x_{01},\ldots,x_{0p})\in\mathbb{R}^p \tag{2.65}$$

Los sesgos marginales, $b_j(X_j)$ pueden ser estimados a través del algoritmo iterativo Backfitting, algoritmo con estructura de Gauss-Seidel introducido por FRIEDMAN Y STUETZLE (1981) y perfeccionado por HASTIE Y TIBSHIRANI (1990) para los modelos aditivos, ligeramente modificado para este caso por LARSEN (2005).

Nótese que en los tres casos de estimación local directa de la probabilidad de default

$$\begin{aligned}
&\text{logit}\left(\hat{P}(Y=1/X=x)\right)=\text{logit}(\hat{p})+\left\{\log\left(\hat{f}_1(x)\right)-\log\left(\hat{f}_0(x)\right)\right\}\\
&\text{logit}\left(\hat{P}(Y=1/X=x)\right)=\text{logit}(\hat{p})+\sum_{j=1}^{p}\left\{\log\left(\hat{f}_{1j}(x_j)\right)-\log\left(\hat{f}_{0j}(x_j)\right)\right\}\\
&\text{logit}\left(\hat{P}(Y=1/X=x)\right)=\text{logit}(\hat{p})+\sum_{j=1}^{p}\left\{\log\left(\hat{f}_{1j}(x_j)\right)-\log\left(\hat{f}_{0j}(x_j)\right)+\hat{b}_j(x_j)\right\}
\end{aligned} \tag{2.66}$$

se consideró como hipótesis de partida que *la probabilidad de default se relaciona con las variables explicativas del riesgo de crédito a través de la transformación logística.* También en los tres casos el único conocimiento consiste en *una muestra aleatoria de observaciones de variables explicativas y del estado de default para N miembros de la población de acreditados,* $\tau=\left\{(x_i,y_i)\in\mathbb{R}^p\times\mathbb{R}\right\}_{i=1,\ldots,N}$.

Debido a que para variable respuesta binaria se tiene $P(Y=1/X=x)=E[Y/X=x]$, realmente en los tres casos de (2.66) el estimador del logit de la probabilidad de default en un punto x_0 se obtiene de forma equivalente al ajuste de un modelo logístico a los datos de los acreditados que pertenecen a una vecindad $V_h(x_0)$ definida por una métrica $m_h(x,x_0)$, es decir, a falta de más conocimiento sobre la estructura de la relación de dependencia entre el logit de la probabilidad de default y las variables explicativas se utilizan las peculiaridades locales de los datos, en cada punto x_0.

Por otro lado, las estructuras de la función de calificación de acreditados reflejadas en (2.66) sugieren que *con la "suficiente" información es posible establecer*

un modelo logístico cuya función de calificación sea una combinación lineal de funciones de las variables explicativas que refleje la relación de dependencia entre X *e* Y*, lo que constituye una idea clave en esta Tesis Doctoral.*

En resumen, las técnicas de estimación directa de la probabilidad de default analizadas en este apartado, a excepción del GBNC de Larsen, conllevan importantes problemas. El estimador $k-NN$ tiende a ser usualmente poco suave, lo que conlleva sobreajuste y en el *estimador por funciones núcleos multivariantes,* a pesar de que se utiliza la relación teórica que liga la variable respuesta con las variables explicativas, se enfrenta a la maldición de la dimensionalidad salvo que se acepte, lo que en general para las variables de riesgo de crédito es inaceptable, la hipótesis ingenua de Bayes.

Afortunadamente la relación (2.17) nos permite volcar todo el potencial de las técnicas de regresión en la búsqueda de soluciones aceptables para el problema de estimar la probabilidad de default. Pero será necesario formular hipótesis basadas en indicios razonables que nos permitan plantear modelos de la probabilidad de default realizables estadísticamente y aceptables desde el punto de vista de los requerimientos de Basilea II. Cualquier atisbo de conocimiento sobre la estructura formal deberá ser aprovechado en beneficio del objetivo anterior.

Una revisión de las técnicas de estimación no paramétrica de la densidad puede verse en SILVERMAN (1986), uno de los primeros textos de una larga lista publicados en los últimos 25 años sobre las técnicas de estimación de la densidad. Otros textos de interés son SCOTT (1992), FAN, J. y GIJBELS (1996), HASTIE et al. (2009), EFRON (2004), HÄRDLE et al. (2004a), WASSERMAN (2006).

2.6 MODELOS ESTADÍSTICOS DE CREDIT SCORING. CONCEPTO Y CARACTERÍSTICAS.

Desde una perspectiva muy general y en sentido amplio, podemos afirmar que un modelo estadístico consiste en una formalización de la variabilidad observada sobre un conjunto de datos, en la que se distinguen dos elementos, la variabilidad sistemática y la variabilidad aleatoria (LINDSEY, 1995). Esta formalización responde, por tanto, a la siguiente expresión:

$$Respuesta = Componente\ Sistemática + Componente\ Aleatoria \tag{2.67}$$

La componente sistemática representa la forma en que los valores de ciertas variables explicativas explican la variabilidad en la respuesta y se describe generalmente mediante

un modelo de regresión. Dado que la componente sistemática del modelo describe una respuesta «ideal», para considerar las fluctuaciones en la respuesta debemos incluir una componente probabilística en el modelo, la componente aleatoria o residual, que describe, mediante una distribución de probabilidad, en qué medida la variable respuesta observada se desvía de la respuesta esperada a partir de la parte sistemática del modelo.

Una forma paralela a la anterior de especificar la relación entre las componentes sistemática y aleatoria es mediante la ecuación:

$$Datos = Modelo + Error \tag{2.68}$$

donde el modelo se corresponde con la variabilidad de los datos explicada por la componente sistemática y el error es la variabilidad no explicada o componente aleatoria del modelo y representa la discrepancia entre los datos observados y los pronosticados por la componente sistemática.

Nuestro objetivo consiste en construir un modelo predictivo para estimar la probabilidad de default, $P(Y=1/X=x)$ *a través de* $E[Y/X=x]$*, o, equivalentemente, estimar la función de regresión* $r(x)=E[Y/X=x]$ *a través de un modelo de ajuste ya sea por métodos paramétricos, semiparamétricos o no paramétricos.*

En general, buscaremos que el modelo exprese la relación existente entre una conveniente transformación de la probabilidad de default, $g(\bullet)$, y la función de calificación de acreditados:

$$g\left(P(Y=1/X=x)\right)=S(x) \tag{2.69}$$

Tabla 2.2- Modelos en función de la transformación $g(\bullet)$.

Familia de Modelos	**Transformación** $g(\bullet)$
Probabilidad	$g\left(P(Y=1/X=x)\right)=P(Y=1/X=x)=S(X)$
Logísticos	$g\left(P(Y=1/X=x)\right)=\Lambda^{-1}\left(P(Y=1/X=x)\right)=S(X)$
Probit	$g\left(P(Y=1/X=x)\right)=\Phi^{-1}\left(P(Y=1/X=x)\right)=S(X)$
Vector Soporte	$g\left(P(Y=1/X=x)\right)=Sign\left\{P(Y=1/X=x)-\frac{1}{2}\right\}=S(X)$

donde $\Lambda(z)$ es la función de distribución logística de parámetros 0 y 1, $\Lambda(z)=\frac{1}{1+e^{-z}}$, $\forall z\in\mathbb{R}^p$, $\Phi(z)$ es la función de distribución normal estandarizada, $\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{u^2}{2}}du$, $\forall z\in\mathbb{R}$ y $Sign\{z\}=\begin{cases}- & si\ z<0\\ + & si\ z\geq 0\end{cases}$, $\forall z\in\mathbb{R}^p$.

Las tres transformaciones más utilizadas en los modelos actuales de credit scoring, capítulo 5, son las correspondientes a las tres últimas filas de la tabla 2.2.

Una vez fijado el nexo de unión entre la probabilidad de default y la función de calificación de acreditados, estimaremos la probabilidad de default, $P(Y=1/X=x)=E[Y/X=x]$, a través del correspondiente modelo de ajuste

$$g\left(\hat{P}(Y=1/X=x)\right)=\hat{S}(x)+\varepsilon \tag{2.70}$$

Además, se asume que se observa el default si $g\left(\hat{P}(Y=1/X=x)\right)$ es positivo:

$$\hat{y}=\begin{cases}1 & si\ \ g\left(\hat{P}(Y=1/X=x)\right)>0\\ 0 & si\ \ g\left(\hat{P}(Y=1/X=x)\right)\leq 0\end{cases} \tag{2.71}$$

Debido a que la variable respuesta Y es binaria, (2.10) indica que *el nexo natural que une el "verdadero" modelo de las variables explicativas con la variable respuesta es la transformación logística*, $g(\bullet)=\text{logit}(\bullet)$, *que da origen, junto con la función de pérdida logística, a los modelos a los que, en sentido amplio, nos referiremos como* ***modelos logísticos.*** Por esta razón, *primaremos a lo largo de toda esta Tesis Doctoral los modelos logísticos y utilizaremos los demás como técnicas de apoyo o contraste*. En línea con este planteamiento analizaremos de forma destacada los modelos logísticos de expresión general,

$$\text{logit}\left(P(Y=1/X=x)\right)=S(x) \tag{2.72}$$

y cuyo modelo de ajuste viene dado por

$$\text{logit}\left(\hat{P}(Y=1/X=x)\right)=\hat{S}(x)+\varepsilon \tag{2.73}$$

donde se asume que el término de error ε tiene distribución logística de media 0 y varianza $\frac{\pi^2}{3}$.

En realidad, como ya hemos comentado, es muy difícil conocer la función $S(x)$, por lo que para estimarla contamos con la opción de establecer hipótesis lo más realistas posible sobre el modelo de dependencia que liga la probabilidad de default con las variables explicativas, $X=(X_1,....,X_p)$.

En general, para una transformación (2.69), la relación de dependencia entre la probabilidad de default y las variables explicativas X, que se quiere estimar, es miembro de una *familia de modelos estadísticos*, $\boldsymbol{F}=\left\{S:\ X\ \rightarrow\ g\left(P\left(Y=1/\ X=x\right)\right)\right\}$. Por tanto, de forma habitual el problema consiste en estimar la probabilidad de default, $P\left(Y=1/\ X=x\right)$, o, equivalentemente, la función de calificación, $S(x)$. De este modo nos encontramos ante un problema de ajuste de un *modelo estadístico en sentido amplio,* es decir, se deberá construir o un *modelo estadístico* o bien u*n algoritmo predictivos* para alcanzar conclusiones sobre el comportamiento del acreditado frente a sus obligaciones de pago a través de los valores de ciertas variables explicativas del riesgo de crédito observados sobre él, al que nos referiremos como ***modelo de credit scoring.***

Nótese que al final del párrafo anterior no hemos dicho "estimar la probabilidad de default" o "clasificar un cliente en una clase de riesgo", cuestiones que van implícitas, sino que hemos enfatizado en algo más general, exigido por Basilea II, *en el hecho de* ***alcanzar conclusiones sobre el comportamiento de default del acreditado***.

En términos de la transformación logística, a través de la probabilidad de default podría entonces obtenerse la puntuación crediticia de los acreditados según la ecuación $S(x)=logit\left(P\left(Y=1/\ X=x\right)\right)$ y, recíprocamente, obtenida la función de puntuación se obtiene la probabilidad de default, $P\left(Y=1/\ X=x\right)=\Lambda(S(x))$.

La literatura sobre los modelos estadísticos es muy amplia y no entra dentro de nuestros objetivos su discusión en profundidad, pero existe un aspecto muy importante en la construcción de modelos que creemos conveniente comentar brevemente por su importancia en el enfoque metodológico de esta Tesis Doctoral, *modelos estadístico versus algoritmos de aprendizaje*.

A pesar de su utilización muy anterior, el auge de los modelos estadísticos, como herramientas para el análisis de datos, se inició en los años 80, fundamentalmente debido a los nuevos desarrollos de la Estadística y la evolución en la capacidad de los ordenadores y su popularización, (LUNNEBORG, 1994). La aplicación intensiva de las técnicas de computación a la construcción de modelos en todos los dominios científicos durante las dos últimas décadas del siglo pasado fue tal que se generó una nueva cultura, apodada la cultura de los algoritmos, (BREIMAN, 2001). En su célebre y polémico artículo de 2001, Breiman dice que existen dos culturas en el uso de los modelos estadísticos para alcanzar conclusiones a partir de los datos, la *cultura de los modelos estadísticos* y la *cultura de los algoritmos*.

Por tratar con la variabilidad y la incertidumbre nos encontramos ante modelos probabilísticos o estocásticos, (BARTHOLOMEW, 1995), que se caracterizan por que asumen la existencia de un error resultante de la desviación entre el fenómeno observado y su representación mediante el modelo, por esta razón *sean modelos de datos o algoritmos los englobamos todos dentro del epígrafe "modelos estadísticos",* pues al contrario de lo defendido BREIMAN (2001), en vez de considerar los algoritmos alternativos a los modelos estadísticos los consideramos complementarios.

Cultura de los modelos estadísticos: formalmente el mecanismo de estimación puede representarse, por ejemplo, en la forma

$$g(y) \leftarrow \textit{regresion logística} \leftarrow x \tag{2.74}$$

Un modelo de este tipo se valida a través de herramientas estadísticas tales como el test de bondad de ajuste, el examen de residuos o error empírico, el área bajo la curva Roc, etc.

Cultura de los algoritmos: el mecanismo a través del cual se relacionan la variable respuesta y las variables explicativas es *"una caja negra" con un complejo y desconocido mecanismo interior*. El objetivo en esta aproximación es encontrar un algoritmo que operando sobre x pronostique los valores de $g(y)$. Tal mecanismo se puede representar en la forma siguiente:

$$g(y) \leftarrow \textit{desconocido} \leftarrow x \tag{2.75}$$

Con respecto al enfoque que sea necesario dar en cada caso a la construcción del modelo, insistimos que en los sistemas de calificación de acreditados necesitamos, además de predecir el default y puntuar y clasificar a los acreditados y a los solicitantes de crédito, conocer lo mejor posible la relación del estado de default con la información que sobre créditos, acreditados y solicitantes de créditos poseen los bancos y otras Entidades Financieras; lo demanda Basilea II, lo demanda la calidad del riesgo, y, por tanto, la rentabilidad de la actividad financiera, y lo que no es menos importante, lo demanda el cliente.

Un modelo estadístico es fundamentalmente una herramienta estadística para estudiar la variabilidad de los datos observados, ya sea la variabilidad sistemática o la variabilidad aleatoria, por tanto, el modelo dependerá tanto del objetivo que se persigue con el estudio de la variabilidad como de los datos disponibles para estudiarla.

En primer lugar, el modelo debe responder al triple objetivo perseguido en credit scoring: estimar las probabilidades de default, estimar la función de calificación de los acreditados, y clasificar a los nuevos solicitantes de crédito dentro de una de las poblaciones, default y no default.

En segundo lugar, debe obtenerse en función de los datos disponibles que no siempre son los más adecuados para especificar los mejores modelos.

En tercer y último lugar, el modelo depende del grado de conocimiento tanto sobre la distribución conjunta de la variable respuesta y las variables explicativas como sobre las distribuciones conjuntas de las variables condicionadas al default y al no default.

La combinación de los tres requerimientos del párrafo anterior determina todas las propiedades que ha de tener el modelo más idóneo para cada situación concreta. Sin embargo, existen algunas pautas generales, casi siempre exigibles y en todo caso deseables, que ha de cumplir cualquier modelo con el que se pretendan obtener rendimientos adecuados. Expondremos aquí cuatro aspectos, que consideramos fundamentales, (STONE, 1985), ligados a la triple calidad que Basilea II requiere a un modelo, *calidad explicativa, predictiva y discriminante*:

1º.- *Flexibilidad del modelo*. Capacidad para describir situaciones de naturaleza diferente.

2º.- *Dimensión del modelo.* Ligada a la varianza de las estimaciones, que crece rápidamente para N fijado si *p* aumenta (problema de *la maldición de la dimensionalidad*), lo que conlleva la inestabilidad del modelo estimado. Problema que se resuelve con excesiva frecuencia, no siempre motivada, suponiendo una estructura aditiva para el modelo, con estimación de funciones univariantes.

Los dos aspectos, flexibilidad y dimensión están estrechamente ligados, y un modelo adecuado ha de basarse sobre un compromiso de equilibrio entre flexibilidad y dimensión.

3º.- *Interpretabilidad del modelo.* Es fundamental para la comprensión de la estructura probabilística subyacente, de acuerdo con los requerimientos del NACB II. El modelo ha de revelar la relación entre el estado de default y una característica de riesgo particular, condicionado a la presencia de las demás características. Hay un argumento definitivo para defender la interpretabilidad de los modelos de calificación y clasificación del riesgo, *los clientes insisten realmente en modelos interpretables*.

4º.- Generalización. El modelo no solo ha de describir bien la relación entre el estado de default y las características de riesgo seleccionadas por los expertos y clasificar lo más correctamente posible a los acreditados que han servido de base para la construcción del modelo, sino que además es fundamental que extienda esta relación y clasifique correctamente a nuevos acreditados distintos a los de entrenamiento, característica conocida como *modelo generalizable* o como *no sobre ajustado*.

El modelo ha de obtenerse de la forma más sencilla que sea posible, salvaguardando la eficacia de sus objetivos. *Las descripciones deben mantenerse lo más simples posibles hasta el momento en que se demuestre que resultan inadecuadas*, es el famoso *dilema de Occam*, que, en palabras de Albert Einstein, viene a decir, "*que el modelo sea sencillo, lo más sencillo posible, pero no más*".

Una función compleja puede describir los datos muestrales muy bien y sin embargo no generalizar bien los resultados a la población en estudio por estar excesivamente sobre ajustado, también puede darse lo contrario, tal como se muestra en la figura 2.11.

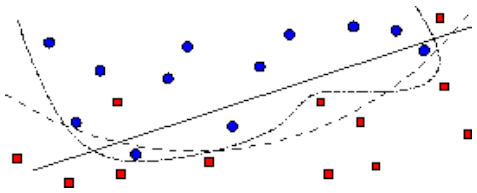


Figura 2.11. Ilustración del infra y sobreajustea una pequeña muestra. La línea recta consigue un bajo ajuste y produce errores muestrales. La línea compleja (*línea discontinua por puntos*) apenas tiene errores muestrales pero puede no generalizar bien sobre datos no muestrales. La función con complejidad intermedia (*línea discontinua*) parece capturar la mejor frontera de decisión.

Con frecuencia existe la tentación de solucionar el dilema de Occam recurriendo a argumentos subjetivos, en este caso estaremos en una situación en que es peor el remedio que la enfermedad. Es fundamental solucionar el dilema de Occam utilizando tan sólo argumentos objetivos, pues, aunque es complejo, conviene recordar que *son muchos los constructores de modelos estadísticos que entienden que tal construcción es un arte*.

En otro orden de cosas, tal como se expuso en la sección 1.4, el desarrollo de un modelo de *credit scoring* consta de 8 fases y decíamos allí que si bien todas ellas son fundamentales en el desarrollo de un modelo, las fases de *selección de variables*, *especificación del modelo* y *desarrollo del mismo* son de particular interés para esta investigación. A continuación veremos algunos de los aspectos más relevantes en la especificación del modelo, aspectos muy importantes por cuanto *la estructura formal elegida para describir la relación entre el estado de default y las variables explicativas condicionará todo el proceso de construcción del modelo en su conjunto*.

2.7. ESPECIFICACIÓN DE LA ESTRUCTURA FUNCIONAL DE UN MODELO DE CREDIT SCORING .

En general, *especificar la estructura funcional de un modelo consiste en seleccionar de entre un conjunto de estructuras posibles aquella más relevante para describir las principales características de la variable estado de default*. Este proceso implica tomar decisiones que conciernen a la formulación de la componente sistemática, los supuestos sobre la componente aleatoria y cómo se combinan las dos componentes en el modelo. La especificación del modelo está sustentada tanto en la teoría sustantiva como en la observación de los datos que nos conducirán a la identificación de los aspectos más relevantes en relación con el estado de default.

A este respecto BOX et al. (1988) señalan que

> *«la identificación es un proceso informal en el que la construcción de gráficos, el análisis preliminar de los datos en sus múltiples facetas y la reflexión sobre las relaciones entre los elementos fundamentales del sistema para el que se quiere construir el modelo se emplean para llegar a una clase de modelos que valga la pena considerar más profundamente».*

El objetivo de la especificación del modelo estadístico consiste en hallar aquella estructura formal que se espera generalice mejor, a través de la ejecución de diferentes modelos en orden a elegir el más idóneo, usando los datos de entrenamiento disponibles, es decir, desde la óptica de Basilea II, *aquella formulación que mejor prediga la probabilidad de default, que mejor explique el comportamiento del acreditado frente al default y que clasifique correctamente a nuevos acreditados distintos a los de entrenamiento,* (*que sea generalizable*).

En este ámbito, no debemos confundir términos como estimación (cálculo de los parámetros del modelo), con la especificación del modelo (identificar la forma funcional más apropiada), con la propia selección del modelo que, desde una óptica estadística, supone la estimación y estudio del término de error del modelo y su capacidad explicativa.

No hay modelos perfectos, ya que cada modelo diseñado es una descripción particular de la realidad, y al menos en teoría, podríamos tener más de un modelo con verosimilitudes, ajuste y capacidad explicativa adecuadas. Aceptar la idea de una estructura formal única para un modelo como dogma suele conducir a evidentes contradicciones que solo pueden resolverse si se rechaza tal idea.

No existe un modelo estadístico ni un algoritmo de aprendizaje que sea inherentemente mejor que otros, (entendemos por mejor que tenga un error de generalización inferior). En esto coinciden la mayoría de los autores y nuestra experiencia. En la misma línea entendemos que *ningún modelo estadístico ni ningún algoritmo son universalmente buenos para todos los problemas*. De hecho, *si promediamos sobre todos los problemas, todos los modelos y todos los algoritmos seguramente se tendría el mismo error*. Esto es independiente tanto de la distribución conjunta del vector de variables aleatorias explicativas, $P(X)$, como del tamaño de los datos. Lo razonable es pensar que *para una tarea y un conjunto de datos, puede existir un modelo estadístico o un algoritmo optimal*.

Por tanto, en realidad no existe un "súper modelo" que sea bueno para todos los propósitos, e incluso en el mismo estudio puede ser necesario, a menudo, contar con al menos dos tipos de modelos: uno para la descripción y la interpretación y otro para la predicción, si bien, en una interpretación estricta de las recomendaciones del Nuevo Acuerdo de Capital de Basilea II, en la construcción de los modelos de credit scoring deberemos contar con un modelo único que integre ambas capacidades. Es necesario, por tanto, dejarnos guiar por el sentido común, siendo el mejor modelo el que, a juicio del analista o experto en riesgos mejor describe o explica el estado de default, previa garantía del estadístico de que el ajuste, la capacidad explicativa y poder discriminante es el adecuado a la política de riesgos marcada por la Entidad Financiera y que deberá ser corroborado con la calibración del modelo al final del horizonte temporal prefijado.

Dicho esto, y teniendo presentes los problemas que pudiera causarnos no controlar personalmente la selección del mejor modelo, sino dejarlo en manos de un algoritmo, hay que decir que la selección automatizada del modelo puede ahorrarnos gran cantidad de tiempo si lo usamos con carácter exploratorio, pero sería un gran error que el modelo real que se implemente en explotación se seleccione de este modo. El modelo se obtendrá de diferentes métodos o bien del mismo método con diferentes parámetros, en caso de ser paramétrico.

La estructura formal de nuestros modelos para la probabilidad de default, $P(Y=1/X=x)$, $x \in \mathbb{R}^p$, ha de basarse necesariamente en (2.69) y (2.70), e, i*ndependientemente de la transformación* $g(\bullet)$ *que se adopte, la estimación de la probabilidad y de la función de calificación dependerá del conocimiento que se posea sobre la forma y parámetros de las distribuciones de las variables explicativas, ya sea de* $P(X,Y)$ *o bien de* $P(Y=1/X=x)$. Dependiendo de este conocimiento podremos utilizar métodos paramétricos, métodos no paramétricos o bien métodos semiparamétricos pero, en todo caso, $S(x)$ es una función sobre la que realizaremos las hipótesis necesarias para obtener el "*mejor modelo posible*" para alcanzar nuestro triple objetivo de pronosticar el default, calificar a los acreditados y clasificar a los nuevos solicitantes de crédito.

Tendremos el modelo especificado una vez que conozcamos en todos sus términos la expresión $g\left(P(Y=1/X=x)\right)=S(x)$. La especificación del modelo habrá de realizarse contando con los datos disponibles y el conocimiento sobre los mismos, así como con el

conocimiento experto en riesgo de crédito y todo ello bajo el enfoque de los requerimientos de Basilea II. Con respecto a la especificación del modelo los requerimientos de Basilea II se enfocan especialmente en dos aspectos:

1.- Calidad de la información a utilizar en la estimación de la probabilidad de default.

Los acuerdos de Basilea II, requieren que en la estimación de la probabilidad de default se cumplan los siguientes requisitos, (relacionados con los datos disponibles, con las variables explicativas y con los juicios expertos), (BCBS (2006), § III, 417):

- ✓ Las variables deben formar un conjunto razonable de variables explicativas.
- ✓ Debe ser posible verificar la exactitud, la exhaustividad y la idoneidad de los datos.
- ✓ Los datos a analizar deben ser representativos de la población.
- ✓ El juicio humano debe tener en cuenta información no considerada por el modelo.

2.- Capacidad del modelo para describir el comportamiento del default.

Se deberá construir un modelo estadístico que permita alcanzar conclusiones sobre el comportamiento de default del acreditado a través de los valores de ciertas variables explicativas del riesgo de crédito observados sobre él.

Insistimos en que los modelos de un adecuado sistema de calificación de acreditados, además de predecir el default y puntuar y clasificar a los acreditados y a los solicitantes de crédito, deben proporcionar el mayor volumen posible de información sobre la relación del estado de default con los datos que sobre créditos, acreditados y solicitantes de crédito poseen los bancos.

CAPÍTULO 3

ESTIMACIÓN Y EVALUACIÓN DE MODELOS DE CREDIT SCORING.

3.1. INTRODUCCIÓN.

Como consecuencia del teorema de Bayes, la expresión matemática exacta del modelo de puntuación crediticia en función de la probabilidad de default a priori y de la razón de verosimilitud de las variables explicativas, si ésta está definida, viene dada por (2.10), donde $g(\bullet)$ es una transformación estrictamente monótona de la probabilidad de default $P(Y=1/X=x)$.

Desgraciadamente, por un lado, la razón de verosimilitud, a diferencia de la esperanza condicional, no está siempre definida y, por otro, las funciones de densidad condicional multivariantes están afectadas por la maldición de la dimensionalidad y exacerbadas por el limitado número de defaults, por lo que hemos de afirmar que, hasta la fecha, la razón de verosimilitud no es en general un valor práctico en el contexto de la modelización del riesgo de crédito. Dos notables ejemplos de modelos logísticos basados en la razón de verosimilitud, donde se conocen las funciones de densidad condicional multivariantes, son los modelos lineal discriminante, LDA, y el discriminante cuadrático, QDA, pero *estos modelos requieren la hipótesis de normalidad multivariante que difícilmente se cumplirá en los datos de riesgos relacionados con el comportamiento de acreditado y solicitantes de crédito frente al default*.

Por ello *es necesario contar con modelos de regresión que permitan estimar la esperanza condicional, que en nuestro caso de respuesta binaria coincide con la probabilidad de default,* $P(Y=1/X=x)=E[Y/X=x]$. Por esta razón, en este capítulo se abordan los aspectos que nos parecen más significativos de la ingente maquinaria necesaria para la estimación de la probabilidad de default desde la óptica de la regresión de la esperanza condicional del estado de default sobre las variables explicativas del riesgo de crédito. La regresión se realiza a través de modelos estadísticos que relacionan una transformación de la probabilidad de default con la función de calificación de acreditados, los *modelos de credit scoring:*

(1) En primer lugar analizaremos los principales instrumentos actualmente disponibles para la *estimación de los modelos de credit scoring,* tales como la *función de pérdida*, que es clave en la construcción de la función objetivo a optimizar, el *riesgo empírico*, que consiste en la medida de la discrepancia entre los valores pronosticados por el modelo y los valores observados, estadístico a optimizar siguiendo el *principio de inducción*, el *riesgo esperado*,

o esperanza matemática del riesgo empírico y la *pérdida y el riesgo locales*. Analizaremos estos instrumentos no sólo para los *Modelos Logísticos*, familia a la que pertenecen los *Modelos Logísticos Lineales por expansión lineal híbrida de funciones de base,* HLLM, objetivo principal de esta Tesis Doctoral, sino también otras familias de interés comparativo, los *Modelos Probit* y los *Modelos de Vector Soporte*. Por otro lado, puesto que los métodos basados en el riesgo empírico global, son con frecuencia relativamente insensibles a las peculiaridades vecinales de los datos, es necesario considerar la versión local de la estimación de los modelos anteriores.

(2) Dado que el principio de inducción puede presentar, sobre todo para muestras pequeñas, los problemas del sobreajuste e infraajuste, características que restan al modelo capacidad de generalización, es necesario en tal caso, contemplar la *regularización del modelo de probabilidad de default*. La regularización pretende conseguir modelos de tendencia antes que modelos muy ajustados localmente, es decir pretende "*suavizar el modelo*", para lo que utiliza la *funcional de riesgo regularizado o estructural,* suma de un término de regularización o penalización más el riesgo empírico, concepto ligado con los splines penalizados clave en la "suavización de funciones".

(3) Tras la especificación y estimación de un modelo es necesario evaluar si los datos se ajustan adecuadamente al mismo, es decir, se ha de valorar la discrepancia entre los datos observados y los ajustados por el modelo. Usualmente la aproximación utilizada para contrastar la *bondad de ajuste* consiste en comparar los datos observados con los datos pronosticados para los datos de entrenamiento. Se trata de medir por tanto las discrepancias entre unos y otros de modo que a menor discrepancia mejor será el ajuste.

(4) Otro aspecto fundamental a considerar en los modelos de credit scoring es su *capacidad de predicción*, que en el caso de variable de respuesta binaria coincide con su *poder discriminante*. A menudo varios modelos estadísticos compiten por explicar los datos, y la comparación de modelos es el procedimiento utilizado para obtener el modelo más parsimonioso que reproduzca mejor el comportamiento de los datos observados. Para la s*elección del modelo se e*stiman diferentes modelos en orden a elegir el mejor.

(5) Por último, existe otra tarea a realizar que consiste en, una vez elegido el mejor modelo, estimar su *capacidad de generalización* o *error de predicción esperado* sobre nuevos datos, usualmente la muestra test.

La eficacia de la generalización de un modelo de ajuste está en relación directa con su capacidad de predicción sobre datos test obtenidos independientemente de la muestra de entrenamiento. Es, por tanto, muy importante fijar este comportamiento en la práctica, puesto que esto guía la elección del modelo de ajuste y da una medida de la calidad del modelo finalmente elegido.

3.2. ESTIMACIÓN DE MODELOS DE CREDIT SCORING.

Como hemos visto en el capítulo 2, la relación de dependencia entre $g\left(P\left(Y=1/\ X=x\right)\right)$ y las variables explicativas X que se quiere estimar adopta la forma de una función $S(X)$, función de calificación de acreditados, que se asume es miembro de una familia de funciones $\boldsymbol{F}=\left\{S:\ x\in\mathbb{R}^p\ \rightarrow\ g\left(P\left(Y=1/\ X=x\right)\right)\in\mathbb{R}\right\}$. Para ajustar el modelo, una vez especificado, es decir, una vez fijadas las funciones g y $S(\bullet)\in\boldsymbol{F}$, es necesaria una muestra de observaciones de la población de acreditados, $\boldsymbol{\tau}=\left\{(x_i,y_i)\in\mathbb{R}^p\times\mathbb{R}\right\}_{i=1,\ldots,N}\in\left(X\times Y\right)^N$, normalmente aleatoria simple, donde $X\times\mathrm{Y}$ es el espacio de las observaciones, y a continuación mediante un criterio de optimización fijado se obtendrá el estimador $\hat{S}(\bullet)$ óptimo.

Para decidir entre varias funciones posibles cual describe mejor la dependencia observada se introducen los conceptos de *función de pérdida, riesgo esperado y riesgo empírico,* que pasamos a tratar en la subsección siguiente.

3.2.1 Función de Pérdida, Riesgo Esperado y Riesgo Empírico.

Definición 3.1.- Sea $X\in\mathbb{R}^p$ un vector aleatorio explicativo e $Y\in\mathbb{R}$ una variable aleatoria perteneciente a una familia de variables respuesta $\boldsymbol{Y}$, con distribución conjunta $P\left(X,Y\right)$. Sea $S\left(X\right)$ perteneciente a una familia $\boldsymbol{F}$ de funciones para predecir Y dados los valores de X . Se dice que la función

$$\begin{array}{rccc}\ell: & \boldsymbol{Y}\times\boldsymbol{F} & \rightarrow & \mathbb{R}\\ & \left(Y,S\left(X\right)\right) & \rightarrow & \ell\left(Y,S\left(X\right)\right)\end{array} \tag{3.1}$$

es una *función de pérdida para penalizar los errores en la predicción* si es acotada y mide el coste de la discrepancia entre la función de pronóstico $S(X)$ y la variable aleatoria respuesta Y.

La función de pérdida $\ell(Y,S(X))$ es una herramienta clave para estimar el modelo, puesto que a partir de ella se obtiene el criterio para ajustarlo a los datos, y, dado que en este caso se usa en problemas de optimización, debe ser convexa.

Definición 3.2.- Se llama *riesgo esperado* asociado a la función de pérdida $\ell(Y,S(X))$ a la cantidad

$$R_\ell(Y,S(X)) = E_P\left[\ell(Y,S(X))\right] \tag{3.2}$$

donde E_P es la esperanza matemática con respecto a la distribución conjunta $P(X,Y)$ del vector aleatorio de variables explicativas X y la variable aleatoria respuesta Y.

Un criterio razonable para elegir $S(X)$, consiste en minimizar el riesgo esperado, (3.2), y, dado que $P(X,Y) = P(Y/X)P(X)$, se verifica

$$E_P\left[\ell(Y,S(X))\right] = E_X E_{Y/X}\left[\ell(Y,S(X))/X\right] \tag{3.3}$$

por lo que es suficiente minimizar punto a punto (3.3)

$$S(x) = \min_S E_{Y/X}\left[\ell(y,S(X))/x\right] \tag{3.4}$$

Definición 3.3.- Dada una muestra aleatoria $\tau = \left\{(x_i,y_i)\in\mathbb{R}^p\times\mathbb{R}\right\}_{i=1,\ldots,N} \in (X\times Y)^N$, si la función $\ell(Y,S(X))$ es una función de pérdida en la predicción de Y por $S(X)$,

a) La cantidad

$$\ell(y_i,S(x_i)) \tag{3.5}$$

se llama *pérdida empírica para la observación* (x_i,y_i).

b) La cantidad

$$\hat{\mathrm{R}}_{emp\ell}(Y,S(X)) = \frac{1}{N}\sum_{i=1}^{N}\ell(y_i,S(x_i)) \tag{3.6}$$

es la pérdida media empírica, estimador del riesgo esperado (3.2), llamado *Riesgo Empírico*. En los problemas de optimización es equivalente optimizar el riesgo

empírico que optimizar la cantidad $L_{emp\ell}\left(Y,S\left(X\right)\right)=\sum_{i=1}^{N}\ell\left(y_i,S(x_i)\right)$ a la que llamaremos *pérdida empírica.*

Un ejemplo de función de pérdida, particularmente importante en nuestra Tesis Doctoral, lo constituye la función de *pérdida logística*, que viene motivada como se expone a continuación:

Dada una muestra, τ, de la ecuación (2.13) se deduce que

$$\begin{aligned} P\left(Y=1/\ X=x_i\right)&=\Lambda\left(S(x_i)\right)\\ P\left(Y=0/\ X=x_i\right)&=1-\Lambda\left(S(x_i)\right)\end{aligned} \tag{3.7}$$

Las dos igualdades de (3.7) se pueden sintetizar en la forma

$$P\left(Y=y_i\ /\ X=x_i\right)=\Lambda\left(S(x_i)\right)^{y_i}\left[1-\Lambda\left(S(x_i)\right)\right]^{1-y_i} \tag{3.8}$$

Por otro lado, dado que la función de verosimilitud asociada con la muestra es

$$L\left(S(x)\right)=\prod_{i=1}^{N}P\left(Y=y_i\ /\ X=x_i\right)=\prod_{i=1}^{N}\Lambda\left(S(x_i)\right)^{y_i}\left[1-\Lambda\left(S(x_i)\right)\right]^{1-y_i} \tag{3.9}$$

se tiene

$$\begin{aligned} -\log\left(L\left(S(x)\right)\right)&=-\log\left(\prod_{i=1}^{N}\Lambda\left(S(x_i)\right)^{y_i}\left[1-\Lambda\left(S(x_i)\right)\right]^{1-y_i}\right)\\ &=\sum_{i=1}^{N}-\left(y_iS(x_i)-\log\left(1+e^{S(x_i)}\right)\right)\end{aligned} \tag{3.10}$$

que matricialmente se puede expresar como

$$-\log\left(L\left(\boldsymbol{S}(x)\right)\right)=-\left[\boldsymbol{S}(x)^T\boldsymbol{Y}-\log\left(1+\exp\left(\boldsymbol{S}(x)\right)\right)^T\mathbf{1}\right] \tag{3.11}$$

donde $\mathbf{Y}=(y_1,...y_N)^T$, $\mathbf{1}=(1,...1)^T$ y $\boldsymbol{S}(x)=\left(S(x_1),...,S(x_N)\right)^T$

La ecuación (3.10) nos sugiere una función de pérdida para medir la discrepancia entre los valores pronosticados, $S(x_i)$, y los verdaderos valores observados para la variable respuesta, y_i, teniendo en cuenta la transformación logística $\text{logit}\left(\text{P}\left(Y=y_i\ /\ X=x_i\right)\right)$

$$\ell\left(y_i,S(x_i)\right)=-y_iS(x_i)+\log\left(1+e^{S(x_i)}\right) \tag{3.12}$$

por lo que se tiene

$$\frac{-\log\left(L\left(S(x)\right)\right)}{N} = \frac{\sum_{i=1}^{N} -\left(y_i S(x_i) - \log\left(1+e^{S(x_i)}\right)\right)}{N} \tag{3.13}$$

(3.13) refleja, por tanto, la *pérdida logística empírica media o riesgo logístico empírico* $\widehat{\mathrm{R}}_{emp\ell}\left(y, S(x)\right)$.

Los argumenos anteriores motivan la siguiente definición:

Definición 3.4.- La función

$$\ell\left(Y, S\left(X\right)\right) = -\left(YS\left(X\right) - \log\left(1+e^{S(X)}\right)\right) \tag{3.14}$$

es una función de pérdida para penalizar los errores en la predicción llamada función de *pérdida log-verosimilitud binomial negativa* o también, como utilizaremos con frecuencia en esta Tesis Doctoral, *pérdida logística*.

A pesar de que la función de pérdida logística es la función de pérdida natural para el ajuste de la función de calificación de acreditados, tal como revelan (3.7) y (3.10), tomamos en consideración otras tres funciones de pérdidas: la d*esviación cuadrática, la log-verosimilitud probit negativa y la pérdida bisagra de las maquinas vector soporte* (SVM). La primera nos conduce a plantear la relación de dependencia entre el estado de default y las variables explicativas del riesgo, no como un modelo de probabilidad sino como un *modelo de probabilidad generalizado, tal como los modelos logísticos o probit*. A través de la segunda se construye una técnica que ha sido ampliamente utilizada en el credit scoring, el modelo Probit, y que creemos es útil contrastar con las técnicas apoyadas en la pérdida logística. Con la misma intención que la anterior, pero con motivos más actuales, pues da origen al método Maquinas de Vector Soporte, SVM, técnica de la teoría del aprendizaje máquina que viene siendo comparado con frecuencia con las técnicas logísticas.

La función de pérdida más popular y conocida en el ajuste de modelos de regresión en general es sin duda alguna *la pérdida cuadrática,* que da lugar a los métodos de estimación más populares, *mínimos cuadrados cuyo* objetivo es minimizar la *suma de los cuadrados residuales,*

$$RRS\left(S(x)\right) = \sum_{i=1}^{N}\left(y_i - S(x_i)\right)^2 \tag{3.15}$$

Definición 3.5.- a) La función

$$\ell(Y,S(X))=\left(Y-S(X)\right)^2 \tag{3.16}$$

es una función de pérdida llamada *pérdida cuadrática.*

b) El riesgo empírico asociado con la muestra, llamado *riesgo empírico cuadrático*, se expresa en este caso en la forma:

$$\widehat{\mathrm{R}}_{emp\ell}\left(y,S(x)\right)=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-S(x_i)\right)^2 \tag{3.17}$$

La construcción de la pérdida log-verosimilitud Probit negativa, o para abreviar, *pérdida Probit*, es muy similar al de pérdida logística, pero suponiendo en este caso que la función de enlace entre la probabilidad de default y la función de calificación de acreditados es la función de distribución normal estandarizada $\Phi(z)$, $Z\sim N(0,1)$,

$$\Phi\left(z\right)=\frac{1}{\left(2\pi\right)^{1/2}}\int_{-\infty}^{z}e^{-\frac{u^2}{2}}du \tag{3.18}$$

En este caso se tiene

$$\begin{aligned}P\left(Y=1/\,X=x_i\right)&=\Phi\left(S(x_i)\right)\\P\left(Y=0/\,X=x_i\right)&=1-\Phi\left(S(x_i)\right)\end{aligned} \tag{3.19}$$

que se puede sintetizar en

$$P\left(Y=y_i\,/\,X=x_i\right)=\Phi\left(S(x_i)\right)^{y_i}\left[1-\Phi\left(S(x_i)\right)\right]^{1-y_i} \tag{3.20}$$

De igual forma que para la pérdida logística, se puede definir una función de pérdida para medir la discrepancia entre los valores pronosticados, $S(x_i)$, y los verdaderos valores observados para la variable respuesta, y_i, teniendo en cuenta la transformación probit, $\text{probit}\left(P\left(Y=y_i\,/\,X=x_i\right)\right)=S\left(x_i\right)$,

$$\begin{aligned}\ell\left(y_i,S(x_i)\right)&=-\log\left(\Phi\left(S(x_i)\right)^{y_i}\left[1-\Phi\left(S(x_i)\right)\right]^{1-y_i}\right)\\&=-y_i\log\left(\frac{1-\Phi\left(S(x_i)\right)}{\Phi\left(S(x_i)\right)}\right)-\log\left(1-\Phi\left(S(x_i)\right)\right)\end{aligned} \tag{3.21}$$

Con lo que la función de log verosimilitud negativa asociada con la muestra es

$$-\log L\left(S(x)\right)=-\sum_{i=1}^{N}\left[y_i\log\left(\frac{1-\Phi\left(S(x_i)\right)}{\Phi\left(S(x_i)\right)}\right)+\log\left(1-\Phi\left(S(x_i)\right)\right)\right] \tag{3.22}$$

De nuevo, en forma análoga a como ocurre para la pérdida logística, (3.22) refleja la pérdida empírica media o riesgo empírico $\widehat{\mathrm{R}}_{emp\ell}\left(y,S(x)\right)$, por lo que se puede caracterizar la pérdida probit de acuerdo con la definición 3.6.

Definición 3.6.- La función

$$\ell\left(Y,S\left(X\right)\right)=-\log\left(\Phi\left(S\left(X\right)\right)^{Y}\left[1-\Phi\left(S\left(X\right)\right)\right]^{1-Y}\right) \tag{3.23}$$

es una función de pérdida llamada *pérdida log-verosimilitud negativa Probit.*

Por lo que respecta a la *pérdida bisagra de Support Vector Machines*, esta viene motivada por el hecho de que los métodos de clasificación SVM provienen de la técnica Perceptron, y del Hiperplano Separador Optimal (OSH), técnicas concebidas como clasificadores a través de un hiperplano separador que hace mínima la distancia de los puntos mal clasificados a la frontera de decisión. Por tanto, la hipótesis de partida, que ha dado lugar a la función de pérdida bisagra de SVM, tomando la variable respuesta valores +1 para representar el default y -1 para el no default, es que si una respuesta $y_i=1$ está mal clasificada, entonces $S\left(x_i\right)<0$ y si una respuesta $y_i=-1$ está mal clasificada, entonces $S\left(x_i\right)\geq 0$, es decir, en modo compacto, $y_iS\left(x_i\right)\geq 1$, $i=1,...,N$, por tanto, la discrepancia entre los valores pronosticados, $S(x_i)$, y los verdaderos valores observados para la variable respuesta, y_i, viene dada por

$$\ell(y_i,S\left(x_i\right))=\left[1-y_iS\left(x_i\right)\right]_+=\text{máximo}\left\{\ 0\ ,\ 1-y_iS\left(x_i\right)\ \right\} \tag{3.24}$$

Por lo que podemos definir una *función de pérdida* para medir el coste de la discrepancia entre la función de pronóstico $S\left(X\right)$ y la variable aleatoria respuesta, Y, según la definición 3.7.

Definición 3.7.- La función

$$\ell(Y,S\left(X\right))=\left[1-YS\left(X\right)\right]_+=\text{máximo}\left\{\ 0\ ,\ 1-YS\left(X\right)\ \right\} \tag{3.25}$$

es una función de pérdida llamada *pérdida bisagra de SVM o margen suave*.

De acuerdo con (3.24), el *riesgo empírico bisagra* o *margen suave* se expresa en la forma:

$$\widehat{\mathrm{R}}_{emp\ell}\left(y,S(x)\right)=\frac{1}{N}\sum_{i=1}^{N}\left[\left[1-y_iS(x_i)\right]_+\right] \tag{3.26}$$

Tabla 3.1- Funciones de pérdida más usuales.

Pérdida	Función de Pérdida
Cuadrática	$(Y - S(X))^2$
Logística	$-\left[YS(X) - \log\left(1+\exp\left(S(X)\right)\right)\right]$
Probit	$-\left[Y\log\left(\Phi\left(S(X)\right)\right)+(1-Y)\log\left(1-\Phi\left(S(X)\right)\right)\right]$
SVM	$\left[1-YS(X)\right]_+$

Fuente: Hastie et al. (2009) y elaboración propia.

En la tabla 3.2 se muestra la formulación del riesgo empírico, en expresión matricial, correspondiente a las funciones de pérdida más notables, para una muestra de entrenamiento τ.

Tabla 3.2- Riesgo empírico para las funciones de pérdida más usuales.

Pérdida	Riesgo Empírico
Cuadrática	$\frac{1}{N}(\boldsymbol{Y}-\boldsymbol{S}(x))^T(\boldsymbol{Y}-\boldsymbol{S}(x))$
Logit	$-\frac{1}{N}\left[\boldsymbol{Y}^T\boldsymbol{S}(x)-\mathbf{1}^T\log\left(1+\exp\left(\boldsymbol{S}(x)\right)\right)\right]$
Probit	$-\frac{1}{N}\left[\boldsymbol{Y}^T\log\left(\boldsymbol{\Phi}\left(\boldsymbol{S}(x)\right)\right)+(\mathbf{1}-\boldsymbol{Y})^T\log\left(1-\boldsymbol{\Phi}\left(\boldsymbol{S}(x)\right)\right)\right]$
SVM	$\frac{1}{N}\left[\mathbf{1}^T\left(1-\boldsymbol{Y}\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)\right]_+$

Fuente: Hastie et al. (2009) y elaboración propia.

donde $\boldsymbol{S}(x)=\left(S(x_1),...,S(x_N)\right)^T$, $x_i=\left(x_{i1},...,x_{ip}\right)^T$, $i=1,...,N$, $\mathbf{Y}=(y_1,...y_N)^T$, $\mathbf{1}=(1,...1)^T$,

$$[Z]_+=\begin{cases} z & si \quad z>0 \\ 0 & si \quad z\le 0 \end{cases}.$$

3.2.2 Función de Pérdida Empírica Local y Riesgo Empírico Local.

Los métodos de estimación basados en la función de pérdida, y, por tanto, en el riesgo esperado global, son con frecuencia relativamente insensibles a peculiaridades vecinales de los datos, tales como agrupaciones locales y agrupaciones de los datos en algunas zonas, particularmente en las colas. Los métodos basados en la pérdida local evitan estos inconvenientes y son imprescindibles cuando poco o nada se conoce sobre la distribución de los datos disponibles para la estimación de modelos. La idea es ajustar localmente $S(X)$ en una "vecindad local" centrada en $x_0 \in \mathbb{R}^p$ de radio $h \in \mathbb{R}$ especificada por una determinada métrica $m_h(x_0, x)$ en $\mathbb{R}^p$.

Definición 3.8.- Se llama *región de vecindad de radio* h *para* $x_0 \in \mathbb{R}^p$ especificada por una métrica $m_h(x_0, x)$, al conjunto de los puntos $x \in \mathbb{R}^p$ para los que $d_{mh}(x_0, x) \le h$, siendo $d_{mh}(x_0, x)$ la distancia inducida por la métrica $m_h(x_0, x) = \|x_0 - x\|$, es decir, a la hiperesfera

$$V_{mh}(x_0) = \left\{ x \in \mathbb{R}^p \,/\, d_{mh}(x_0, x) \le h \right\} \tag{3.27}$$

Definición 3.9.- Dada una muestra τ, una función de pérdida particular $\ell : Y \times \boldsymbol{F} \rightarrow \mathbb{R}$ y una vecindad local $V_{mh}(x_0) = \left\{ x \in \mathbb{R}^p \,/\, m_h(x_i, x_0) \le h \right\}$ especificada por una métrica $m_h(x_i, x_0)$,

a) A la cantidad

$$\ell_{V_{mh}(x_0)}\left(y_i, S(x_i)\right) = m_h(x_i, x_0)\,\ell\left(y_i, S(x_i)\right) \tag{3.28}$$

se le llama *pérdida empírica local para* (x_i, y_i) *en la vecindad local* $V_{mh}(x_0)$.

b) La función

$$\widehat{\mathrm{R}}_{emp\,\ell_{V_{mh}(x_0)}}(y, S(x)) = \frac{1}{N}\sum_{i=1}^{N} \ell_{V_{mh}(x_0)}\left(y_i, S(x_i)\right) = \frac{1}{N}\sum_{i=1}^{N} m_h(x_i, x_0)\,\ell\left(y_i, S(x_i)\right) \tag{3.29}$$

es un estimador, conocido como *riesgo empírico local o pérdida empírica media en la vecindad local* $V_{mh}(x_0)$.

La métrica m_h en (3.29) está orientada a impedir que se tengan en cuenta los sumandos correspondientes a los puntos x_i que no se encuentren dentro de la vecindad $V_{mh}(x_0)$.

El uso de funciones de pérdida local permite obtener modelos más flexibles, puesto que, aunque minimizando el riesgo empírico local se ajusta un modelo entero en una vecindad de un punto $x_0 \in \mathbb{R}^p$, $V_{mh}(x_0)$, el modelo sólo se usa para evaluar el ajuste en el punto x_0.

Tabla 3.3.- Riesgo local empírico para las funciones de pérdida más usuales.

Pérdida	Riesgo Local Empírico en $x_0 \in R^p$ $\left(métrica\ m_h(x_0,x)\right)$
Cuadrática Local	$\frac{1}{N}\left(\boldsymbol{M}_h(x,x_0)\circ(\boldsymbol{Y}-\boldsymbol{S}(x))\right)^T(\boldsymbol{Y}-\boldsymbol{S}(x))$
Logística Local	$-\frac{1}{N}\left[\left(\boldsymbol{M}_h(x,x_0)\circ\boldsymbol{Y}\right)^T\boldsymbol{S}(x)-\boldsymbol{M}_h(x,x_0)^T\log\left(1+\exp\left(\boldsymbol{S}(x)\right)\right)\right]$
Probit Local	$-\frac{1}{N}\left[\left(\boldsymbol{M}_h(x,x_0)\circ\boldsymbol{Y}\right)^T\log\left(\Phi\left(\boldsymbol{S}(x)\right)\right)+\left(\boldsymbol{M}_h(x,x_0)\circ(\boldsymbol{1}-\boldsymbol{Y})\right)^T\log\left(1-\Phi\left(\boldsymbol{S}(x)\right)\right)\right]$
SVM Local	$\frac{1}{N}\left[\boldsymbol{1}^T\left(\boldsymbol{M}_h(x,x_0)-\left(\boldsymbol{M}_h(x,x_0)\circ\boldsymbol{Y}\right)\boldsymbol{S}(x)\right)\right]_+$

donde $\boldsymbol{M}_h(x,x_0)=\left(m_h(x_1,x_0),...,m_h(x_i,x_0),...,m_h(x_N,x_0)\right)^T$, $\boldsymbol{S}(x)=\left(S(x_1),...,S(x_i),...,S(x_N)\right)^T$

$x_i=\left(x_{i1},...,x_{ij},...,x_{ip}\right)^T$, $i=1,...,N$, $\mathbf{Y}=(y_1,...,y_i,...,y_N)^T$, $\mathbf{1}=(1,...1)^T$, $[Z]_+=\begin{cases} z & si\ \ z>0 \\ 0 & si\ \ z\le 0 \end{cases}$ y el

el símbolo $\circ$ indica producto de Hadamard de matrices: $(A\circ B)_{ij}=\left(A_{ij}.B_{ij}\right)$.

En la tabla 3.3 se muestra la formulación matricial del riesgo empírico local correspondiente a las funciones de pérdida locales más notables, para una muestra $\boldsymbol{\tau}$ y una métrica $m_h(x_i,x_0)$ que define la vecindad local $V_{mh}(x_0)$.

3.2.3. Principio de Inducción.

Si se conoce $P(X,Y)$, distribución conjunta de (X,Y), o bien $P(X/Y)$, verosimilitud de (X/Y), el problema de estimar el modelo de dependencia de una transformación de la variable estado de default Y, respecto de las características X, a través de una *función de estimación* $S\in\boldsymbol{F}$ y una función de pérdida $\ell(Y,S(X))$ puede ser fácilmente resuelto sin más que establecer un método que *encuentre la función que minimice el riesgo esperado,* $R_\ell(Y,S(X))=E_P\left[\ell(Y,S(X))\right]$, *de entre todas las funciones* $S\in\boldsymbol{F}$. Pero en la práctica lo frecuente es no conocer tales funciones de probabilidad, lo que implica que no

podamos computar directamente el riesgo esperado asociado a $\ell(Y,S(X))$ (esperanza matemática de la función de pérdida) y, por tanto, para seleccionar la función de calificación de acreditados de acuerdo con la expresión siguiente:

$$S(X) = \min_{S} E_{Y/X}\left[\ell(Y,S(X))/X\right] \tag{3.30}$$

será necesario disponer de una muestra de observaciones y minimizar la esperanza de la función de pérdida con respecto a una distribución empírica $\hat{P}_{empírica}(X,Y)$, es decir, minimizar el riesgo empírico. Este método, conocido como *"principio de inducción"*, posiblemente el método más universal para la estimación de dependencias, se desarrolla en tres pasos:

1).- En primer lugar se define *una clase de funciones* $\boldsymbol{F}$, colección de todos los modelos seleccionables para explicar los datos muestrales.

2).- *A continuación se fija una apropiada función de pérdida* $\ell(Y,S(X))$.

3).- Y finalmente *se establece un método que, de entre todas las funciones* $S \in \boldsymbol{F}$, *encuentre la que minimize el riesgo empírico,* $\hat{R}_{emp\ell}(Y,S(X)) = \frac{1}{N}\sum_{i=1}^{N}\ell(y_i,S(x_i))$, es decir que resuelva el problema

$$\min_{S} \sum_{i=1}^{N}\ell(y_i,S(x_i)) \tag{3.31}$$

Definición 3.10.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p \times \mathbb{R}$. Se dice que $\hat{S}(X)$ *es el Estimador de Mínimo Riesgo de* $S(X)$, $X \in \mathbb{R}^p$, si es la solución del problema de optimización (3.31).

3.2.4 Ajuste de Modelos con Funciones de Pérdida Notables.

Habíamos visto en el capítulo 2 que los modelos de nuestro interés, los que expresan la relación existente entre una conveniente transformación de la probabilidad de default, $g(\bullet)$, y la función de calificación de acreditados de probabilidad, $g(P(Y=1/X=x)) = S(X)$, se clasifican en primer lugar en función de esta transformación: modelos de probabilidad, logísticos, probit y de vector soporte. Pues bien, aparece un nuevo elemento que caracteriza los distintos modelos, la función de pérdida. Pero como puede observarse hemos llegado a las funciones de pérdida precisamente a partir de las correspondientes transformaciones, por lo que los modelos vendrán caracterizados indistintamente por cualquiera de las dos funciones.

3.2.4.1 Modelos de Probabilidad.

Como deciamos, una de las funciones de pérdida utilizadas con más frecuencia en la estimación de dependencias entre variables es la pérdida cuadrática, $\ell\left(Y,S\left(X\right)\right)=\left(Y-S\left(X\right)\right)^2$, a partir de la cual se obtiene la pérdida esperada condicional en la forma siguiente,

$$E_{Y/X}\left[\ell\left(Y,S\left(X\right)\right)\right]=P(Y=1/\,X=x)\left(1-S\left(X\right)\right)^2+P(Y=0/\,X=x)\left(0-S\left(X\right)\right)^2$$

El riesgo esperado $R_\ell\left(Y,S\left(X\right)\right)$ se minimiza, punto a punto, resolviendo el problema de optimización $\min\limits_S E_{Y/X}\left[\ell\left(Y,S\left(X\right)\right)/\,x\right]$. El mínimo se alcanza en

$$S\left(X\right)=E\left[Y/\,X=x\right] \tag{3.32}$$

Por tanto, la función de estimación de $S\left(X\right)$ que minimiza el riesgo asociado a la función de pérdida $\left(Y-S\left(X\right)\right)^2$ es miembro de la familia de esperanzas condicionales de Y sobre X,

$$\boldsymbol{F}_{Modelos\,\mathrm{Pr}obabilidad}=\left\{S\left(X\right)/\;S\left(X\right)=E\left[Y/\,X=x\right]\right\}$$

es decir, $$Y=S\left(X\right)+\varepsilon \tag{3.33}$$

Como puede observarse, en este caso la variable respuesta Y se relaciona directamente con las variables explicativas $X=\left(X_1,\ldots,X_p\right)$.

Definición 3.11.- Sean $\left(X,Y\right)$, $\left(X_1,Y_1\right),\ldots,\left(X_N,Y_N\right)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p\times\mathbb{R}$.

Se dice que $\hat{\mathrm{S}}\,(x)$, $x\in\mathbb{R}^p$, es el *Estimador de Mínimos Cuadrados Ordinarios* de la función $S(X)$, $X\in\mathbb{R}^p$, si es la solución del Problema de Mínimos Cuadrados Ordinarios siguiente

$$\min_S\ \sum_{i=1}^{N}\left(y_i-S\left(x_i\right)\right)^2 \tag{3.34}$$

Esta familia de modelos, de estructura muy simple, presenta un gran problema en la estimación de la probabilidad de default, problema que ilustramos con el siguiente ejemplo:

Supongamos que la función de calificación $S(X)$ es lineal en $X=(X_1,...,X_P)$, es decir, $S(X)=\beta_0+\sum_{j=1}^{p}\beta_j X_j$, se obtiene entonces el modelo de probabilidad lineal

$$P(y=1/\ X=x)=S(X)=\beta_0+\sum_{j=1}^{p}\beta_j x_j \tag{3.35}$$

A pesar de que tanto la estimación como la interpretación del modelo (3.35) son muy simples, presenta un fuerte problema que lo hace inadecuado para estimar una medida de probabilidad. De ser cierta la relación (3.35), $S(X)$ sería una medida de probabilidad, por lo que debería tomar valores en $[0,1]$, sin embargo $S(X)$ es una recta que puede tener imágenes en un intervalo distinto del adecuado, en principio en toda la recta real, pues ninguna restricción se ha hecho sobre el modelo en ese sentido.

Para resolver el problema anterior utilizaremos funciones de pérdida apropiadas, tales como la logística y la probit, lo que es equivalente a plantear modelos de probabilidad generalizados, donde las funciones de enlace entre la esperanza condicional y las variables explicativas son funciones de distribución con lo que nos aseguramos de que la esperanza condicional tome valores en [0,1].

3.2.4.2 Modelos Logísticos.

La estimación de la función de calificación de acreditados, $S(X)\in \boldsymbol{F}$, de los modelos más populares de credit scoring se realiza utilizando la función de pérdida logística, $\ell_\Lambda(Y,S(X))=-\left(Y\ S(X)-\log\left(1+e^{S(X)}\right)\right)$. En este caso,

$$E_{Y/X}\left[\ell\left(Y,S(X)\right)\right]=P(Y=1/\ X=x)\left(-S(X)+\log\left(1+e^{S(X)}\right)\right)$$
$$+P(Y=0/\ X=x)\left(\log\left(1+e^{S(X)}\right)\right)$$

y, por tanto, el riesgo esperado $R_\ell\left(Y,S(X)\right)$ alcanza el mínimo en el punto

$$S(x)=\log\left(\frac{P(Y=1/\ X=x)}{P(Y=0/\ X=x)}\right) \tag{3.36}$$

La relación de dependencia de la variable estado de default, Y, respecto de las características X observadas sobre los acreditados, que minimiza el riesgo asociado a la pérdida logística $-\left(Y\ S(X)-\log\left(1+e^{S(X)}\right)\right)$, se expresa a través de la siguiente relación:

$$P(Y=1/\ X=x)=\Lambda\left(S(X)\right) \tag{3.37}$$

es decir, la *función de estimación de mínimo riesgo, para la función de pérdida logística, es miembro de la familia de las funciones logit de la probabilidad de default,* $\boldsymbol{F}_{Logit}=\left\{S(X)/\ S(x)=\text{logit}\left(P\left(Y=1/\ X=x\right)\right)\right\}$, *por lo que la relación de dependencia se podrá estimar, con mínimo riesgo esperado, a través del modelo de regresión logística:*

$$\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)=\hat{S}(x)+\varepsilon \tag{3.38}$$

De (3.37) y (3.38) se sigue que ajustar a través de la regresión logística, de forma "suave", la función $S(X)$, conduce a la estimación "suave" de la probabilidad de default $P\left(Y=1/\ X=x\right)$.

Definición 3.12.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p\times\mathbb{R}$. Se llama *estimador logístico de la función de calificación* $S(X)$, $X\in\mathbb{R}^p$, a la función $\hat{S}(x)=\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)$, $x\in\mathbb{R}^p$, si es la solución del problema de optimización siguiente:

$$\underset{S}{Mín}\ -\sum_{i=1}^{N}\left[y_i S(x_i)-\log\left(1+e^{S(x_i)}\right)\right] \tag{3.39}$$

conocido como *Problema de Optimización Logístico.*

Una vez obtenida la función de calificación estimada $\hat{S}(x)$, la probabilidad de default se obtendrá a través del estimador

$$\hat{P}(Y=1/\ X=x)=\Lambda\left(\hat{S}(x)\right)=\frac{1}{1+e^{-\hat{S}(x)}} \tag{3.40}$$

3.2.4.3 Modelos Probit.

La estimación de los *modelos Probit* se obtiene en los mismos supuestos que para los modelos logísticos, pero cambiando la función de pérdida lógistica por la pérdida probit, $\ell_\Phi\left(Y,S(X)\right)=Y\log\left(\Phi\left(S(X)\right)\right)+(1-Y)\log\left(1-\Phi\left(S(X)\right)\right)$. En este caso

$$\begin{aligned}E_{Y/X}\left[\ell_\Phi\left(Y,S(X)\right)\right]&=P(Y=1/\ X=x)\log\left(\Phi\left(S(X)\right)\right)\\&+P(Y=0/\ X=x)\log\left(1-\Phi\left(S(X)\right)\right)\end{aligned}$$

El mínimo $\underset{S}{mín}\,E_{Y/X}\left[\ell_\Phi\left(Y,S(X)\right)/\ X\right]$ se alcanza en

$$S(x)=\Phi^{-1}\left(P(Y=1/\,X=x)\right) \tag{3.41}$$

Por tanto, la *función de estimación de mínimo riesgo* de la dependencia del estado de default con respecto a las variables explicativas X, con pérdida probit, *es miembro de la familia de las funciones probit de la probabilidad de default*, $\boldsymbol{F}_{Probit}=\left\{S(X)/\,S(X)=\ probit\left(P(Y=1/\,X=x)\right)\right\}$, *por lo que la relación de dependencia se podrá estimar, con mínimo riesgo esperado, a través del modelo de regresión probit*

$$\text{probit}\left(\hat{P}(Y=1/\,X=x)\right)=\Phi^{-1}(\hat{P}(Y=1/\,X=x))=\hat{S}(x)+\varepsilon \tag{3.42}$$

o, equivalentemente, $$\hat{P}(Y=1/\,X=x)=\Phi(\hat{S}(x))+u \tag{3.43}$$

Definición 3.13.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p\times\mathbb{R}$. Se llama *estimador probit de la función de calificación* $S(X)$, $X\in\mathbb{R}^p$, a la función $\hat{S}(x)=\text{probit}\left(\hat{P}(Y=1/\,X=x)\right)$, $x\in\mathbb{R}^p$, si es la solución del problema de optimización siguiente:

$$\underset{S}{Mín}\ -\sum_{i=1}^{N}\left[y_i\log\left(\Phi(S(x_i))\right)+(1-y_i)\log\left(1-\Phi(S(x_i))\right)\right] \tag{3.44}$$

conocido como *Problema de Optimización Probit.*

Una vez obtenida la función de calificación $\hat{S}(x)$, la probabilidad de default se obtendrá a través del estimador

$$\hat{P}(Y=1/\,X=x)=\Phi\left(\hat{S}(x)\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\hat{S}(x)}e^{-\frac{u^2}{2}}du \tag{3.45}$$

3.2.4.4 Modelos Vector Soporte.

En este caso, como es habitual en la Teoría del Aprendizaje, se considera que la variable respuesta toma valores $y\in\{-1,+1\}$ y la función de pérdida utilizada es la pérdida bisagra de vector soporte, $\ell\left(Y,S(X)\right)=\left[1-YS(X)\right]_+=\text{máximo}\ \left\{\ 0\ ,\ 1-YS(X)\ \right\}$.

La esperanza condicional se expresa entonces,

$$E_{Y/X}\left[\ell_{SVM}\left(Y,S(X)\right)\right]=P(Y=+1/\,X=x)\left[1-S(X)\right]_+ +P(Y=-1/\,X=x)\left[1+S(X)\right]_+$$

El problema de optimización $\min_S E_{Y/X}\left[\ell_{SVM}\left(Y,S(X)\right)/X\right]$ alcanza el mínimo en

$$S(x)=\begin{cases}+1 \quad si\ P(Y=+1/X=x)\geq\dfrac{1}{2}\\ -1 \quad si\ P(Y=-1/X=x)<\dfrac{1}{2}\end{cases}$$

es decir, en

$$S(x)=\ Sign\left\{P\left(Y=1/X=x\right)-\frac{1}{2}\right\} \tag{3.46}$$

La función de estimación de la relación de dependencia del estado de default respecto de las variables explicativas X, que minimiza el riesgo asociado a $\left[1-Yf\left(X\right)\right]_+$, *es miembro de la familia de las funciones*

$$\boldsymbol{F}_{SVM}=\left\{S\left(X\right)/S\left(X\right)=\ Sign\left(P\left(Y=1/X=x\right)-\frac{1}{2}\right)\right\} \tag{3.47}$$

Obsérvese que en este caso la función de estimación es una función binaria de clasificación. No es posible obtener una probabilidad de forma directa y sólo será posible estimar con mínimo riesgo esperado el hecho de si la probabilidad de default es menor o igual a 0.5 o mayor a 0.5,

$$Sign\left(\hat{P}\left(Y=1/X=x\right)-\frac{1}{2}\right)=\hat{S}(x)+\mathrm{u} \tag{3.48}$$

Definición 3.14.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p\times\mathbb{R}$. Se llama *estimador vector soporte de la función* $S(X)$, $X\in\mathbb{R}^p$, a la función $\hat{\mathrm{S}}\ (x)=Sign\left(\hat{P}\left(Y=1/X=x\right)-\frac{1}{2}\right)$, $x\in\mathbb{R}^p$, solución del problema de optimización siguiente:

$$\underset{S}{Mín}\ -\sum_{i=1}^{N}\left[1-y_iS(x_i)\right]_+ \tag{3.49}$$

conocido como *Problema de Optimización Vector Soporte.*

3.2.4.5 Modelos Locales.

Por lo que respecta a la pérdida logística local, los estimadores locales se definen en función del riesgo empírico local en una vecindad $V_{mh}\left(x_0\right)$ de un punto $x_0\in\mathbb{R}^p$ especificada por una métrica $m_h\left(x,x_0\right)$ en la forma siguiente:

Definición 3.15.- Sean (X,Y), $(X_1,Y_1),...,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p \times \mathbb{R}$. Si $m_h(x,x_0)$ es una métrica que especifica la vecindad $V_{mh}(x_0)=\{x\in\mathbb{R}^p / m_h(x,x_0)\le h\}$ de $x_0 \in \mathbb{R}^p$,

a) Se define el *estimador de Mínimos Cuadrados Ordinarios Local de la función de calificación de acreditados* $S(X)$, *en la vecindad* $V_{mh}(x_0)$, como el valor $\hat{S}(x_0)\in\mathbb{R}$ que minimiza el riesgo empírico local en la vecindad $V_{mh}(x_0)$ con la función de pérdida cuadrática en dicha vecindad,

$$\hat{S}(x_0)=\underset{S}{Mín}\ -\sum_{i=1}^{N} m_h(x_i,x_0)\left(y_i-S(x_i)\right)^2 \tag{3.50}$$

El problema de optimización (3.50) se conoce como *Problema Local de Mínimos Cuadrados Ordinarios en la vecindad* $V_{mh}(x_0)$.

b) Se define el *estimador logístico local de la función de calificación de acreditados* $S(X)$ *en la vecindad* $V_{mh}(x_0)$, como el valor $\hat{S}(x_0)\in\mathbb{R}$ que minimiza el riesgo empírico local en la vecindad $V_{mh}(x_0)$ con la función de pérdida logística en dicha vecindad,

$$\hat{S}(x_0)=\underset{S}{Mín}\ -\sum_{i=1}^{N} m_h(x_i,x_0)\left[y_i S(x_i)-\log\left(1+e^{S(x_i)}\right)\right] \tag{3.51}$$

El problema de optimización (3.51) se conoce como *Problema Logístico Local en la vecindad* $V_{mh}(x_0)$.

c) Se define el *estimador Probit local de la función de calificación de acreditados* $S(X)$ *en la vecindad* $V_{mh}(x_0)$, como el valor $\hat{S}(x_0)\in\mathbb{R}$ que minimiza el riesgo empírico local en la vecindad $V_{mh}(x_0)$ con la función de pérdida Probit en dicha vecindad,

$$\hat{S}(x_0)=\underset{S}{Mín}\ -\sum_{i=1}^{N} m_h(x_i,x_0)\left[y_i \log\left(\Phi\left(S(x_i)\right)\right)+(1-y_i)\log\left(1-\Phi\left(S(x_i)\right)\right)\right] \tag{3.52}$$

El problema de optimización (3.52) se conoce como *Problema Probit Local en la vecindad* $V_{mh}(x_0)$.

d) Se define el *estimador Vector Soporte Local de la función de calificación de acreditados en la vecindad* $V_{mh}(x_0)$, como el valor $\hat{S}(x_0)\in\mathbb{R}$ que minimiza el riesgo empírico local en la

vecindad $V_{mh}(x_0)$ con la función de pérdida bisagra en dicha vecindad,

$$\hat{S}(x_0) = \underset{S}{Mín} \; -\sum_{i=1}^{N} m_h(x_i, x_0)\left[1 - y_i S(x_i)\right]_+ \tag{3.53}$$

El problema de optimización (3.53) se conoce como *Problema Vector Soporte Local en la vecindad* $V_{mh}(x_0)$.

3.3 REGULARIZACIÓN DE MODELOS DE CREDIT SCORING.

3.3.1 Introducción.

El principio de inducción, a pesar de ser un método usado con mucha frecuencia, presenta dos serios problemas:

1.- El primer problema consiste en la no unicidad de la solución. Como puede verse en la figura siguiente, en principio la solución de (3.31) no tiene por que ser necesariamente única.

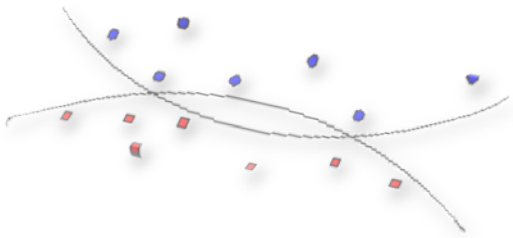


Figura 3.1.- Representación de dos funciones que separan dos clases con riesgo empírico cero.

La figura 3.1 indica que puede haber más de una función con el mismo riesgo empírico sobre los mismos datos muestrales. Además, estas funciones pueden tomar valores arbitrarios en otros puntos de X, puesto que la solución que minimiza el riesgo empírico no garantiza minimizar el verdadero riesgo esperado. Por tanto, sin más información es imposible decidirse por una o por otra. Esto ocurre, por ejemplo, cuando para ciertas combinaciones de familias de modelos paramétricos, de funciones de riesgo empírico y de datos muestrales, el problema de optimización paramétrica puede ser no acotado con respecto a los parámetros del modelo $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$. Así, si consideramos el modelo lineal definido por $S(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, y los datos muestrales son linealmente separables entonces se puede aumentar la magnitud de $\boldsymbol{\beta}$ indefinidamente para reducir el riesgo empírico.

2.- El segundo problema lo constituyen el sobreajuste y el infraajuste, que restan capacidad de generalización al modelo, pues una función compleja puede describir muy bien los datos muestrales con un riesgo empírico muy pequeño y, sin embargo, no generalizar bien los resultados a la población en estudio, también puede darse lo contrario.

Para evitar los dos problemas anteriores es necesario dotar al principio de inducción de la capacidad de desarrollar modelos que, por un lado, describan bien los datos y, por otro posean un adecuado grado de generalización. Para ello se introduce el concepto de regularización del modelo.

La generalización se consigue principalmente "suavizando el modelo", o, en otros términos, consiguiendo modelos de tendencia antes que modelos muy ajustados localmente, a través de una *funcional de suavizado*, $J\left(S\left(X\right)\right)$, también llamada de *penalización o regularización*, de tal modo que bajos valores de la funcional corresponden a funciones suavizadas.

Puesto que miramos hacia una función estrechamente ajustada a los datos y a la vez suave, es natural elegir como solución del problema de aproximación la función que haga mínimo un criterio aditivo con componentes el riesgo empírico, para ajustar, y el término de regularización, para suavizar.

En el caso paramétrico el término de regularización $\lambda J\left(S\left(X\right)\right)$ es una función de los parámetros del modelo que devuelve un alto valor para modelos "inverosímiles" o complejos que, aparte de ser difíciles de interpretar, son responsables de malas generalizaciones.

3.3.2 Estimadores Regularizados de Modelos de Credit Scoring.

Optimizando la expresión (3.54), conocida como *funcional de riesgo regularizado o estructural,* suma del regularizador y del riesgo empírico, se consigue un equilibrio entre la simplicidad del modelo y la minimización del riesgo empírico

$$\sum_{i=1}^{N}\ell\left(y_i,S\left(x_i\right)\right)+\lambda J\left(S\left(X\right)\right) \tag{3.54}$$

donde $\sum_{i=1}^{N}\ell\left(y_i,S\left(x_i\right)\right)$ es el término asociado a la función de pérdida $\ell\left(Y,S\left(X\right)\right)$, llamado *término de ajuste*, puesto que mide el ajuste de los datos, y $J\left(S\left(X\right)\right)$ es la

funcional de regularización, penalización o suavizado, definida sobre un conveniente espacio de Hilbert de funciones $\mathcal{H}$. $\lambda J\left(S(X)\right)$ es el *término de regularización* que toma valores grandes para estimadores complejos de $S(X)$. La cantidad $\lambda > 0$ se llama *parámetro de suavización o regularización* y su misión es equilibrar los dos términos en la ecuación (3.54) para hallar *un estimador óptimo de la desconocida* $S(X)$.

Si el equilibrio es elegido adecuadamente se puede mejorar significativamente la ejecución de la generalización, en comparación con la simple minimización del riesgo empírico. Este equilibrio, conocido en estadística como *regularización,* es lo que ha dado nombre a estas técnicas, (SAHARON, 2003).

Por tanto, puesto que en la construcción de modelos de probabilidad y funciones de calificación de acreditados estamos interesados en una buena estimación del verdadero riesgo sobre todas las posibles puntuaciones de los datos, en ciertas situaciones nuestro objetivo, condicionado por los requerimientos de Basilea II, es más ambicioso que encontrar una función con riesgo empírico minimal, razón por la que necesitamos introducir un término para penalizar soluciones muy complejas.

Definición 3.16.- Para una muestra aleatoria $\tau = \left\{\left(x_i, y_i\right) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N}$, el problema de optimización

$$\underset{S(X)}{Mín} \sum_{i=1}^{N} \ell\left(y_i, S\left(x_i\right)\right) + \lambda J\left(S\left(X\right)\right) \tag{3.55}$$

se conoce como *Problema de Minimización de la Pérdida Empírica Regularizada.*

Los modelos regularizados se clasifican en primer lugar según la función de pérdida, por lo que tendremos modelos logísticos regularizados, modelos probit regularizados y modelos de vector soporte regularizados. Los estimadores correspondientes se pueden definir formalmente en la forma siguiente:

Definición 3.17.- Sean (X,Y), $(X_1,Y_1),\ldots,(X_N,Y_N)$ vectores aleatorios independientes, idénticamente distribuidos, de valores de $\mathbb{R}^p \times \mathbb{R}$.

a) Se dice que $\hat{S}(x) = \text{logit}\left(\hat{P}\left(Y = 1 / X = x\right)\right)$, $x \in \mathbb{R}^p$, es el *estimador logístico regularizado de la función* $S(X)$ si es la solución del *problema logístico regularizado* siguiente

$$\underset{S(X)}{mín}\left\{-\sum_{i=1}^{N}\left[y_i S(x_i)-\log\left(1+e^{S(x_i)}\right)\right]+\lambda J\left(S(X)\right)\right\} \tag{3.56}$$

b) Se dice que $\hat{S}$ (x)=logit$\left[\hat{P}\left[Y=1/X=x\right]\right]$, $x\in\mathbb{R}^p$, es el *estimador probit regularizado de la función* $S(X)$ si es la solución del *problema probit regularizado* siguiente

$$\underset{S(X)}{mín}\left\{-\sum_{i=1}^{N}\left[y_i\log\left(\Phi\left(S\left(x_i\right)\right)\right)+\left(1-y_i\right)\log\left(1-\Phi\left(S\left(x_i\right)\right)\right)\right]+\lambda J\left(S(X)\right)\right\} \tag{3.57}$$

c) Se dice que $\hat{S}$(x), $x\in\mathbb{R}^p$ es el *estimador vector soporte regularizado de la función* $S(x)$ si es la solución del *problema vector soporte regularizado* siguiente

$$S(x)=\underset{S}{mín}\left\{-\sum_{i=1}^{N}\left[1-y_i S(x_i)\right]_+ +\lambda J(S)\right\} \tag{3.58}$$

La forma del regularizador depende de algunas extensiones sobre la forma del modelo.

Si consideramos los modelos más sencillos y fáciles de interpretar en el dominio del credit scoring, los modelos paramétricos de dimensión p, $S(X,\boldsymbol{\beta})$, $\boldsymbol{\beta}=\left(\beta_0,\beta_1...,\beta_p\right)^T$, se puede definir una familia de regularizadores parametrizados muy popular, tal como se recoge en la siguiente definición.

Definición 3.18.- Dada la función de calificación de acreditados $S(X,\boldsymbol{\beta})$, con vector de parámetros $\boldsymbol{\beta}=\left(\beta_0,\beta_1...,\beta_p\right)^T$, para un entero no negativo q y un vector de números reales α, se define el *regularizador de Minkowski de orden* q *para* α en la forma siguiente:

$$J_{q,\alpha}\left(S\left(\boldsymbol{X},\boldsymbol{\beta}\right)\right)=\boldsymbol{\alpha}^T\left|\boldsymbol{\beta}\right|^q=\sum_{j=0}^{p}\alpha_j\left|\beta_j\right|^q \tag{3.59}$$

donde $\boldsymbol{\alpha}^T=\left(\alpha_0,...,\alpha_p\right)$ y $\left(\left|\boldsymbol{\beta}\right|^q\right)^T=\left(\left|\beta_0\right|^q,...,\left|\beta_p\right|^q\right)$.

Los miembros de la familia de regularizadores representados por (3.59) corresponden a diferentes clases de normas de Minkowski del vector de parámetros, por lo que es usual referirse a $J_{q,\alpha}\left(S\left(X,\boldsymbol{\beta}\right)\right)$ como el l_q-*regularizador* . En la práctica suele elegirse $\alpha_i\in\{0,1\}$, lo que nos permite incluir o excluir ciertos elementos de $\boldsymbol{\beta}=\left(\beta_0,\beta_1,...,\beta_p\right)^T$ en la regularización, (por ejemplo, en un modelo lineal se busca normalmente excluir el

término constante o intercepto). *La esencia de estos regularizadores es que penalizan grandes valores de* β_j *cuando* $\alpha_j > 0$.

De acuerdo con el teorema de Girosi, Jones y Pogio (GIROSI et al., 1995) encontrar soluciones para la minimización no restringida de

$$\sum_{i=1}^{N} \ell\left(y_i, S\left(x_i, \boldsymbol{\beta}\right)\right) + \lambda J_{q,\alpha}\left(S\left(\boldsymbol{X}, \boldsymbol{\beta}\right)\right) \tag{3.60}$$

donde $J_{q,\alpha}\left(S\left(\boldsymbol{X}, \boldsymbol{\beta}\right)\right)$ es el regularizador de Minkowki de parámetros $\left(q, \alpha\right)$, es equivalente a encontrarlas para el problema de optimización con restricciones siguiente:

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{mín} \sum_{i=1}^{N} \ell\left(y_i, S\left(x_i, \boldsymbol{\beta}\right)\right) \\ &restingido\ a \sum_{j=0}^{p} \alpha_j \left|\beta_j\right|^q \leq \gamma \end{aligned} \tag{3.61}$$

Existe una correspondencia uno a uno entre los parámetros λ y γ. Los miembros más interesantes de la familia de reguladores que se corresponden con normas de Minkowski son aquellos para los que $q \in \{0, 1, 2\}$.

Las cuestiones de consistencia, sobreajuste e infraajuste están estrechamente relacionadas con el concepto de regularización, (TIKHONOV y ARSENIN, (1977), MOROZOV, (1984)), y con el principio de minimización del riesgo estructural (VAPNIK, 1998).

Existen varias posibilidades para elegir λ y J en orden a deducir un principio inductivo consistente. (1) Se pueden seguir las pautas inspiradas por los trabajos de VAPNIK, (1998), (2) se puede elegir los criterios de información de AKAIKE (1974) y los de MALLOWS (1973), dentro de las metodologías estadísticas clásicas o bien, (3) la regularización por splines, (WABHA, 1980), CART (BREIMAN et al., 1984), la regularización wavelet, (DONOHO et al. 1996), o Vectorial Support Machine, (BARLETT y MENDELSON, 2002), y cualesquiera otras aproximaciones modernas. Una fundamentación general para la selección del modelo de regularización viene dado en BARRÓN et al. (1999).

Una interesante interpretación de la función de penalización y de los métodos de regularización fue dada por HASTIE et al. (2009), afirmando que *"expresan*

nuestra creencia a priori de que el tipo de funciones que se busca exhibe cierta de conducta de alisado, e indica que puede ser incluida en un armazón Bayesiano".

En la subsección siguiente mostraremos como se pueden construir medidas de serpenteo de las funciones que sean fáciles de interpretar y que puedan obtenerse por métodos computacionales directos y de bajo coste.

3.3.3 SPLINES.

3.3.3.1 Concepto y Definición.

Es generalmente aceptado que *el término "spline" hace referencia a una amplia clase de funciones que son utilizadas en aplicaciones que requieren la interpolación de datos, o un suavizado de curvas*. Las funciones para la interpolación por splines normalmente se determinan como minimizadores de la rugosidad sometidas a una serie de restricciones.

La primera referencia a los splines, con el suavizado o aproximación polinomial a trozos, se encuentra en un artículo de SCHOEMBERG (1946), inicialmente el concepto se utilizó sobre todo en el diseño de aeronaves y pronto fue ganando terreno en los más diversos dominios de la actividad humana. BOOR (1978) presenta un tipo de splines con una gran capacidad de adaptación a las técnicas de computación, los *B_Splines*, una década más tarde STONE y KOO (1986) presentaron los *Splines de Regresión Cúbicos Restringidos*, Splines_RCS; los splines más utilizados actualmente son los *Splines Cúbicos*, según la versión de HASTIE y TIBSHIRANI (1990). De difícil aplicación actual pero con gran potencial de futuro fueron presentados por WAHBA (1990) los *Splines de Lamina Fina* que fueron mejorados por GREEN y SILVERMAN (1994). La relación de pioneros se cierra con la presentación de los *Splines Penalizados* de EILERS y MARX (1996).

El término esplines penalizado ha emergido como un descriptor para el ajuste general de splines sujeto a penalización. O'SULLIVAN (1986) introdujo una clase de splines penalizados, basados en las funciones de base B-splines de Boor, que poseen la atractiva característica de las condiciones naturales de frontera (GREEN Y SILVERMAN (1994). Estos splines se han convertido en la clase más usada de splines penalizados en el análisis estadístico, como resultado de su implementación en la función *smooth.spline()* del popular S-PLUS (INSIGHTFUL CORPORATION, 2007) y en el software asociado a los modelos aditivos generalizados (por ejemplo, en la librería GAM en *R*, HASTIE (2006). WAND Y ORMEROD (2008) desarrollaron una matriz exacta para los splines penalizados de O'Sullivan que puede ser implementada en unas pocas líneas de lenguaje de computación basado en matrices.

Excelentes sinopsis sobre splines se incluyen en EUBANK (1999), GU (2002), RUPPERT et al. (2003), DENISON et al. (2002) y en los más recientes WAND y ORMEROD (2008) y HASTIE et al. (2009).

Hasta mediados de la década de 1990 la mayor parte de la literatura sobre regresión no paramétrica se basaba en splines de suavizado y su extensión multivariante, splines de lámina fina, en los cuales la penalización toma una forma particular y el número de nudos es aproximadamente igual al tamaño muestral (por ejemplo, WAHBA (1990), GREEN y SILVERMAN (1994)). En los últimos años, sin embargo, existe una gran cantidad de investigaciones con estrategias basadas en splines penalizados más generales que los debidos a EILERS y MARX (1996), muchas de las cuales usan un número considerablemente menor de nudos.

Los esfuerzos actuales incluyen:

(i) Modelos complejos, frecuentemente con varias funciones de suavizado.

(ii) Grandes conjuntos de datos, en los cuales los splines de suavizado y lámina fina resultan computacionalmente intratables.

(iii) Modelos mixtos y representaciones bayesianas de suavizadores, para los que existe software tal como: BRUGS, 0.4, R-Package, R *Development Core Team,* 2008, LIGGES et al., (2007), y PROC MIXED en SAS (SAS INSTITUTE, INC., 2007), en los cuales el número de nudos es relativamente pequeño.

Al concepto de spline se llega de forma inmediata a través del concepto de función polinomial troceada cuya definición se da a continuación:

Definición 3.19.- Supuesto que X es unidimensional, una *función polinomial troceada* se obtiene dividiendo el dominio de X en intervalos consecutivos con fronteras contiguas llamadas nudos y definiendo polinomios de grado bajo, funciones de base h_r, que interpolan a la función en dos nudos consecutivos $\left(x_r, y_r\right)$ y $\left(x_{r+1}, y_{r+1}\right)$. El conjunto de funciones $h(x) = \{h_r\}_{r=1}^{q}$ forma la *curva polinomial a trozos*.

En el panel izquierdo de la figura 3.2 se representa una polinomial constante a trozos con tres funciones de base:

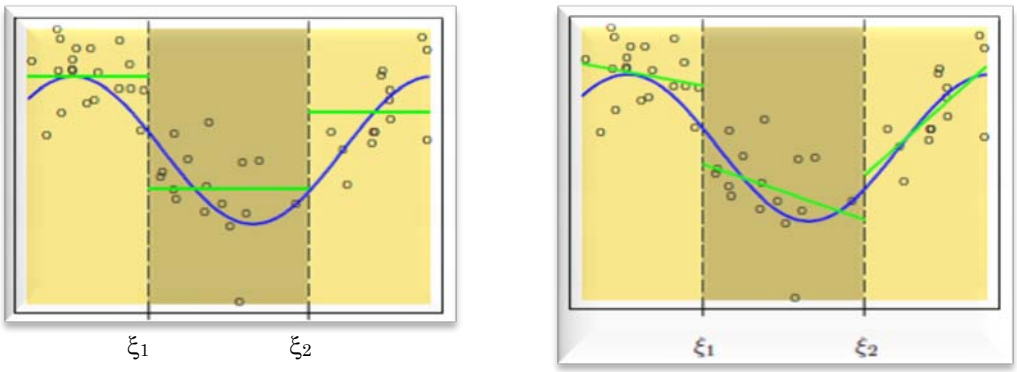
$$h_1(X) = I(X < \xi_1)$$

$$h_2(X) = I(\xi_1 \le X < \xi_2) \qquad (3.62)$$

$$h_3(X) = I(\xi_2 \le X)$$

Puesto que estas funciones de base son positivas sobre regiones disjuntas, el estimador de mínimos cuadrados del modelo $S(X) = \sum_{r=1}^{3} \beta_r h_r(X)$ es $\hat{\beta}_r = \overline{Y}_r$, la media de Y en la m-ésima región.

a. Costante a trozos **b. Lineal a trozos**



Fuente: Hastie et al. (2009).

Figura 3.2.- Curvas polinomiales a trozos.

Podemos obtener una polinomial a trozos más "refinada" que la constante a trozos sin más que considerar 3 funciones de base adicionales,

$$h_{r+3}(X) = h_r(X)X, \quad r = 1,2,3 \qquad (3.63)$$

en este caso la curva polinomial obtenida es lineal a trozos, panel derecho de la figura 3.2. Más "suave", en el sentido de ajuste a una curva "lisa", será la función polinomial resultante de considerar la lineal a trozos con la restricción de continuidad en los dos nudos. Continuidad que podemos conseguir a través de dos alternativas:

a) Introducir funciones lineales sobre cada intervalo

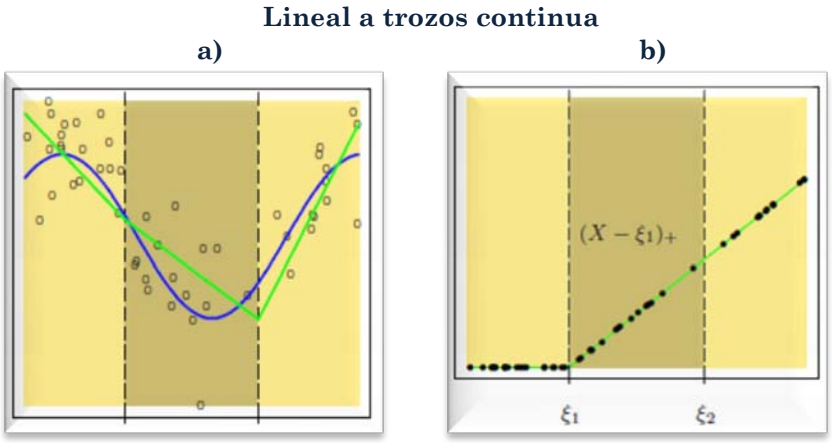
$$\begin{cases} y_1 = \alpha_1 x + \beta_1 \\ y_2 = \alpha_2 x + \beta_2 \\ y_3 = \alpha_3 x + \beta_3 \end{cases} \quad \begin{cases} y_1(\xi_1) = y_2(\xi_1) \\ y_2(\xi_2) = y_3(\xi_2) \end{cases} \qquad (3.64)$$

Esta restricción de continuidad conduce a restricciones lineales sobre los parámetros , por ejemplo, $f(\xi_1^-) = f(\xi_1^+)$ implica que $\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$. En este caso, por tanto, existen dos restricciones. Nótese que se parte con cuatro parámetros libres.

b) Una vía más directa para proceder en este caso es usar bases que incorporen las restricciones, es decir usar expansiones de base (4 parámetros libres), como las siguientes:

$$h_1(X) = 1,\ h_2(X) = X,\ h_3(X) = (X - \xi_1)_+,\ h_4(X) = (X - \xi_2)_+ \qquad (3.65)$$

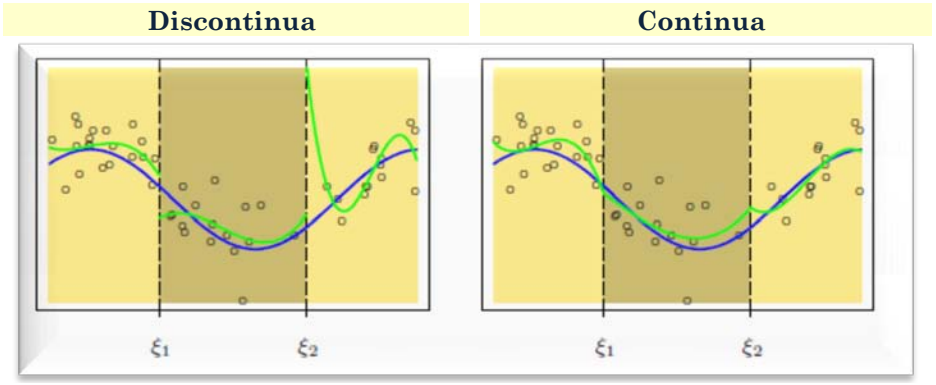
donde Z_+ denota la parte positiva. La función h_3 se muestra en la figura 3.3 b).



Fuente: Hastie et al. (2009).

Figura 3.3.- Polinomial Lineal a trozos Continua.

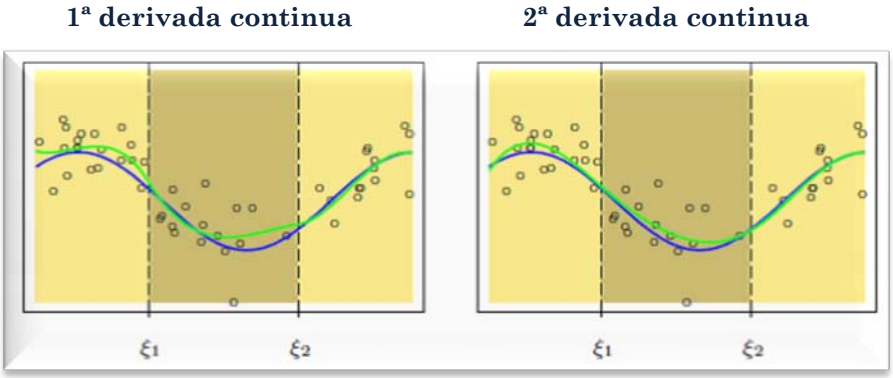
Lo habitual es intentar ajustar a funciones "suaves", lo que se consigue aumentando el orden del polinomio local. En las figuras siguientes, 3.4 y 3.5 se muestran una serie de ajustes polinomiales a los mismos datos, con ordenes crecientes de continuidad y de diferenciabilidad.



Fuente: Hastie et al. (2009).

Figura 3.4.- Polinomial Lineal a trozos Continua y Discontinua.

Si, además, exigimos derivadas de orden suficiente nuestra función de ajuste presentará sin duda una apetecible curva suave como las representadas en la figura 3.5, donde se ha exigido 1ª derivada continua en los nudos, panel izquierdo de la figura, o 2ª derivada continua en los nudos, panel derecho de la figura.



Fuente: Hastie et al. (2009).

Figura 3.5.- Polinomiales con 1ª y 2ª Derivadas Continuas.

A partir de las figuras 3.2, 3.3, 3.4 y 3.5 es fácil intuir el papel central que los polinomiales troceados juegan en la regresión no paramétrica y semiparamétrica.

Definición 3.20.- Se dice que una polinomial troceada $h(x)=\{h_r\}_{r=1}^{q-1}$ con q nudos, para los que $\xi_1<\xi_2<...<\xi_q$, es un *spline de orden* $M\geq 0$ si satisface las siguientes condiciones:

(i) En cada intervalo $[\xi_{i-1},\xi_i)$ $h(x)$ es un polinomio de grado menor o igual que $M-1$.

(ii) $h(x)$ tiene una derivada de orden $M-2$ continua en $[\xi_1,\xi_q)$

$$h(x)=\begin{cases} h_1(x)=a_1+b_1x+...+d_1x^{M-1} & x\in[\xi_1,\xi_2) \\ h_2(x)=a_2+b_2x+...+d_2x^{M-1} & x\in[\xi_2,\xi_3) \\ \quad\quad ... & \\ h_{q-1}(x)=a_{q-1}+b_{q-1}x+...+d_{q-1}x^{M-1} & x\in[\xi_{q-1},\xi_q) \end{cases} \tag{3.66}$$

Los polinomios h_{r-1} y h_r interpolan el mismo valor en el punto ξ_r

$$h_r(\xi_r)=y_r=h_r(\xi_{r+1}),\quad (1\leq r<q-1) \tag{3.67}$$

Los splines de orden 1 son funciones constantes por zonas. Una forma explícita de presentar un spline de orden 1 es la siguiente:

$$h(x)=\begin{cases} h_1(x)=a_1 & x\in[\xi_1,\xi_2) \\ h_2(x)=a_2 & x\in[\xi_2,\xi_3) \\ \cdots & \\ h_{q-1}(x)=a_{q-1} & x\in[\xi_{q-1},\xi_q) \end{cases} \tag{3.68}$$

Los intervalos $[\xi_{i-1},\xi_i)$ no se intersecan entre si, por lo que no hay ambigüedad en la definición de la función en los nudos.

Definición 3.21.- Se definen los *splines de orden 2* en los términos siguientes:

$$h(x)=\begin{cases} h_1(x)=a_1+b_1 & x\in[\xi_1,\xi_2) \\ h_2(x)=a_2+b_2 & x\in[\xi_2,\xi_3) \\ \cdots & \\ h_q(x)=a_q+b_q & x\in[\xi_q,\xi_{q+1}) \end{cases} \tag{3.69}$$

Definición 3.22.- La forma general para el *conjunto de bases potencia-truncada* puede ser, para orden M y q+1 nudos

$$\begin{aligned} &h_r(X)=(X-\xi_l)_+^{M-1},\ r=1,\ldots,q \\ &h_{q+l}(x)=x^{l-1},l=1,\ldots,M \end{aligned} \tag{3.70}$$

De acuerdo con (3.70), las siguientes funciones de base representan un spline cúbico con nudos en ξ_1 y ξ_2:

$$\begin{aligned} &h_1(X)=(X-\xi_1)_+^3,\ h_2(X)=(X-\xi_2)_+^3 \\ &h_3(X)=1,\ h_4(X)=X,\ h_5(X)=X^2,\ h_6(X)=X^3 \end{aligned} \tag{3.71}$$

Existen seis funciones de base correspondientes a un espacio lineal de funciones de dimensión 6. Un rápido chequeo confirma la cuenta de parámetros:

$$(3\ \textit{regiones})\times(4\ \textit{parámetros por región})-(2\ \textit{nudos})\times(3\ \textit{restricciones por nudo})=\ 6.$$

3.3.3.2 Splines de Regresión.

Los splines de regresión son aquellos en que implícitamente se asume que tanto el orden como el número de nodos no está fijado a priori, al igual que su ubicación. En este caso se necesita seleccionar el orden del spline, el número de nodos y su emplazamiento (para controlar la suavidad de la función ajustada) e imponer restricciones para que los trozos

de polinomio se unan de forma suave. Una vez hecha la elección, el modelo se ajusta de acuerdo con la función de pérdida seleccionada.

Una simple aproximación consiste en parametrizar una familia de splines por el número de funciones de base o grados de libertad, y a partir de las observaciones determinar la posición de los nudos. Un método muy usual para evitar el problema del emplazamiento de los nodos consiste en usar *splines suavizadores* o *esplines penalizados*, veremos estos splines en las subsección 3.3.3.3, mientras a continuación veremos los dos tipos de splines de regresión más notables por su utilización en las técnicas de ajuste y clasificación binarios, los *splines cúbicos* y los *splines cúbicos naturales*.

3.3.3.2.1 Splines Cúbicos.

Los splines de regresión más utilizados son los de orden $M=4$ y se los conoce como *splines cúbicos* (HASTIE y TIBSHIRANI (1990). En estos splines se consigue la 2ª derivada continua en todos los nodos, por esta razón proporcionan un excelente ajuste a los datos, además, su cálculo no es excesivamente complejo.

$$h(x)=\begin{cases} h_1(x)=a_1+b_1x+c_1x^2+d_1x^3 & x\in[\xi_1,\xi_2) \\ h_2(x)=a_2+b_2x+c_2x^2+d_2x^3 & x\in[\xi_2,\xi_3) \\ \cdots & \\ h_{q-1}(x)=a_{q-1}+b_{q-1}x+c_{q-1}x^2+d_{q-1}x^3 & x\in[\xi_{q-1},\xi_q) \end{cases} \tag{3.72}$$

Aplicando las condiciones de continuidad del spline $h(x)$ y de la primera y segunda derivadas, $h'(x)$ y $h''(x)$, es posible encontrar su expresión analítica. La expresión resultante viene dada por

$$\begin{aligned} h_r(x)=&\frac{z_r}{6v_r}(\xi_{r+1}-x)^3+\frac{z_{r+1}}{6v_r}(x-\xi_r)^3 \\ &+\left(\frac{y_{r+1}}{v_r}+\frac{z_{r+1}v_r}{6}\right)(x-\xi_r)+\left(\frac{y_r}{v_r}-\frac{z_rv_r}{6}\right)(\xi_{r+1}-x) \end{aligned} \tag{3.73}$$

donde $v_r=\xi_{r+1}-\xi_r$ y los elementos del conjunto $\{z_r\}_{r=1}^{q}$ son incógnitas. Para determinar los valores de las incógnitas, se utilizan las condiciones de continuidad que deben cumplir estas funciones, de donde resulta

$$v_{r-1}z_{r-1}+2(v_r+v_{r-1})z_r+v_rz_{r+1}=\frac{6}{v_{r-1}}(\xi_{r+1}-\xi_r)-\frac{6}{v_{r-1}}(\xi_r-\xi_{r-1}) \tag{3.74}$$

La ecuación (3.74) con $r=1,...,q-2$ genera un sistema de $q-2$ ecuaciones lineales con q incógnitas, $z_1, z_2, ..., z_q$. Podemos elegir z_1 *y* z_q de forma arbitraria y resolver el sistema de ecuaciones resultante para obtener $z_2, ..., z_{q-1}$. Si en la ecuación (3.73) se realiza una elección especialmente adecuada de los elementos z_1 *y* z_q , $z_1 = z_q = 0$, la función resultante se llama *spline cubico natural,* que estudiaremos en el apartado siguiente.

3.3.3.2.2 Splines Cúbicos Naturales

Se conoce que el comportamiento del ajuste de datos por polinomiales tiende a ser errático cerca de las fronteras, y la extrapolación puede ser peligrosa. Estos problemas se acrecientan con splines. El ajuste de polinomiales más allá de las fronteras de los nodos se manejan más delirantemente que los correspondientes a los polinomiales globales en la región. Esto puede ser convenientemente summarizado en términos de la varianza con respecto a la puntuación de las funciones splines ajustadas por mínimos cuadrados. Por eso se consideran muy importantes los llamados *splines cúbicos naturales,* que son splines de regresión cúbicos restringidos, pues corrigen estos problemas.

Un spline cubico natural añade restricciones adicionales, a saber, que la función es lineal cerca de los nodos frontera o, equivalentemente, que el spline tiene segunda derivada cero fuera del intervalo $\left[x_1, x_q\right]$, requerimiento conveniente por cuanto reduce el peligro asociado con la extrapolación.

Una base elemental de splines cúbicos está formada por las funciones siguientes:

$$h_r(x) = |x - \xi_r|^3, \quad r = 1,...,q, \; h_{q+1}(x) = 1 \text{ y } h_{q+2}(x) = x \tag{3.75}$$

En este caso un splin cúbico "natural", que representaremos por N(x), viene dado por

$$N(x) = \sum_{r=1}^{q+2} \beta_r h_r(x) = \sum_{r=1}^{q} \beta_r |x - \xi_r|^3 + \beta_{q+1} + \beta_{q+2} x \tag{3.76}$$

donde se han impuesto a los coeficientes las siguientes restricciones "naturales": $\sum_{r=1}^{q} \beta_r = 0$ y $\sum_{r=1}^{q} \beta_r x_r = 0$, que significan que el spline tiene segunda derivada cero fuera del intervalo $\left[x_1, x_q\right]$.

Por otro lado, la restricción de linealidad en las fronteras libera cuatro grados de libertad (dos restricciones cada uno en ambas regiones frontera), puesto que pueden ser más provechosos unos pocos nudos en la región interior, pero existe un precio a pagar en el sesgo cerca de las fronteras, a pesar de lo cual asumir la linealidad cerca de las mismas se considera bastante razonable.

Los splines cúbicos naturales con q nudos, conocidos como *splines de regresión q-anudados*, están representados por q funciones de base. Se puede comenzar desde una base de splines cúbicos y deducir la base reducida imponiendo las restricciones en la frontera. Por ejemplo partiendo de (3.72) se llega a

$$\begin{aligned} N_r(X) &= \frac{(X-\xi_r)_+^3 - (X-\xi_q)_+^3}{\xi_q - \xi_r} - \frac{(X-\xi_{q-1})_+^3 - (X-\xi_q)_+^3}{\xi_q - \xi_{q-1}}, \quad r = 1,\ldots,q \\ N_l(x) &= x^{l-1}, l = 1,\ldots,M-2 \end{aligned} \tag{3.77}$$

Cada una de estas funciones de base tienen 1ª y 2ª derivada cero para $x \geq \xi_{q+1}$.

Los splines cúbicos son los splines de mayor orden para los cuales la discontinuidad en los nodos es visible al ojo humano. Existen muy pocas razones para considerar splines de orden mayor a 4, salvo que tengamos interés en ajustar derivadas. En la práctica los órdenes más usuales son $M = 1,\ 2\ y\ 4$.

Como ejemplo, supongamos que queremos ajustar nuestra función de calificación, $S(X)$, para una muestra de (X,Y), $\{(x_i, y_i)\}_{i=1}^N \in (X \times Y)^N$, al spline cúbico natural de regresión expresado en (3.76), minimizando cierta discrepancia media entre los valores observados y los pronosticados, el modelo a estimar viene entonces dado por

$$S(X) = N(X) = \sum_{r=1}^{q+2} \beta_r h_r(X) = \sum_{r=1}^{q} \beta_r |X - \xi_r|^3 + \beta_{q+1} + \beta_{q+2} X \tag{3.78}$$

Las funciones de base del spline cúbico natural son $h_r(X) = |X - \xi_r|^3$, $r = 1,\ldots,q$, $h_{q+1}(x) = 1$ y $h_{q+2}(x) = x$. El spline se ajusta buscando los β_r que minimizan la discrepancia media entre los valores observados y los pronosticados, obviamente función de β_r, sujeto a $\boldsymbol{C\beta} = 0$ donde

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_{q+2} \end{bmatrix}^T, \quad \boldsymbol{C} = \begin{bmatrix} 1 & 1 & \ldots & 1 & 0 & 0 \\ \xi_1 & \xi_2 & \ldots & \xi_q & 0 & 0 \end{bmatrix}$$

y

$$\boldsymbol{H}(\mathbf{X}) = \begin{bmatrix} |X_1-\xi_1|^3 & |X_1-\xi_2|^3 & \dots & |X_1-\xi_q|^3 & 1 & X_1 \\ |X_2-\xi_1|^3 & |X_2-\xi_2|^3 & \dots & |X_2-\xi_q|^3 & 1 & X_2 \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot & \cdot \\ |X_N-\xi_1|^3 & |X_N-\xi_2|^3 & \dots & |X_N-\xi_q|^3 & 1 & X_N \end{bmatrix} \tag{3.79}$$

Modelizar splines de regresión es agradablemente simple, pero en la practica la elección de los emplazamientos de los nudos ξ_r puede tener una sustancial influencia sobre los resultados de la modelización. Por tanto, la necesidad de elegir la localización de los nudos puede ser un factor de complicación serio en la modelización por splines de regresión. Una amplia discusión de estos problemas puede verse en HASTIE y TIBSHIRANI (1990).

El problema más importante que presenta el ajuste por splines de regresión y, en particular los splines cúbicos, es la elección del conjunto de nudos, lo que influye notablemente en los resultados del ajuste. Nótese que si el número de nudos es cero el ajuste es lineal, mientras que valores del número de nudos próximos a N conducen al sobreajuste. Este problema puede evitarse, por un lado, utilizando *splines penalizados o de suavizado*, que veremos a continuación, 3.3.3.3 y 3.3.3.4, cuando la complejidad del modelo lo requiere, o, en otro caso, los *splines cúbicos restringidos* de STONE y KOO (1985).

A efectos prácticos está resultando muy eficaz la idea de STONE y KOO (1985) de utilizar los splines cúbicos restringidos a que sean lineales en las colas, como, por ejemplo, los splines de regresión q-anudados, para especificar variables explicativas continuas dentro de componentes no lineales en modelos generales con cualesquiera términos, además poseen la ventaja de un número pequeño de nudos, q, por lo que son expansiones lineales de $q-1$ funciones de base de las variables explicativas. En el capítulo 6, se analizará brevemente la propuesta de STONE y KOO (1985) junto con su recomendación de situar los nudos en ciertos cuartiles. Utilizaremos las funciones de base correspondientes a este tipo de splines en la especificación de variables de la componente no lineal del modelo HLLM proactivo que se propone en esta Tesis Doctoral.

3.3.3.3 Splines de Suavizado.

El problema más importante que presenta el ajuste por splines de regresión se puede evitar mediante un planteamiento alternativo del ajuste, tal como el que se basa en los residuales penalizados, en decir, en el riesgo empírico penalizado, que analizaremos en este apartado.

En la línea anterior están los *splines de suavizado*, (WAHBA, (1990), GREEN y SILVERMAN, (1994)), que resuelven el problema de la localización de los nudos usando un número relativamente grande de ellos lo que puede conducir a modelos sobreajustados, pero puede evitarse el problema aplicando una penalización "ondulante" a la función objetivo del ajuste. Un gran número de nudos significa que el modelo ajustado es bastante insensible a la elección de la localización exacta de los nudos.

Por razones de simplicidad y sin pérdida de generalidad asumiremos en esta apartado, mientras no indiquemos lo contrario, que X es unidimensional.

Si la función de calificación $S(X)$ es una combinación lineal de funciones de base de las variables explicativas del riesgo de crédito, $S(X)=\sum_{r=1}^{q}\beta_r h_r(X)$, sus derivadas e integrales son lineales en β_r:

$$S'(X)=\sum_{r=1}^{q}\beta_r h_r'(X)\ ,\ S''(X)=\sum_{r=1}^{q}\beta_r h_r''(X)\ ,\ \int_{\mathbb{R}} S(X)\ dx=\sum_{r=1}^{q}\beta_r\int_{\mathbb{R}} h_r(X)\ dx \qquad (3.80)$$

A partir de esta linealidad, es posible construir una penalización sobre $S(X)$ que sea grande si $S(X)$ es muy serpenteante y pequeña si es casi plana, lo que puede ser representado convenientemente en términos de los β_r. La posibilidad más usual, consiste en considerar una penalización en la segunda derivada de la curva, que fue introducida por O'SULLIVAN (1986), con el fin de evitar obtener una curva que interpola los datos cuando el número de nudos coincide con el número de observaciones:

$$J(S(X))=\int_{\mathbb{R}}\left(S''(x)\right)^2 dx \qquad (3.81)$$

La segunda derivada de $S(X)$ en un punto $x \in \mathbb{R}$ puede expresarse en términos matriciales en la forma

$$S''(x) = \sum_{r=1}^{q} \beta_r h_r''(x) = \boldsymbol{\beta}^T \boldsymbol{H}''(x) \tag{3.82}$$

donde $\boldsymbol{H}''(x)$ es el vector de segundas derivadas de las funciones de base evaluadas en x. Por ser $S''(x)$ escalar y, por tanto, igual a su traspuesta, se tiene

$$\left(S''(x)\right)^2 = \boldsymbol{\beta}^T \boldsymbol{H}''(x)^T \boldsymbol{H}''(x)\boldsymbol{\beta} = \boldsymbol{\beta}^T \boldsymbol{D}_H''(x)\boldsymbol{\beta} \tag{3.83}$$

siendo $\boldsymbol{D}_{\boldsymbol{H}}''(x)$ la matriz funcional de derivadas segundas de la matriz de las funciones de base evaluadas en x, $\boldsymbol{H}(X)$.

$$\boldsymbol{D}_{\boldsymbol{H}}''(x) = \begin{bmatrix} h_1''(x)^2 & h_1''(x)h_2''(x) & h_1''(x)h_3''(x) & . & . & h_1''(x)h_q''(x) \\ h_2''(x)h_1''(x) & h_2''(x)^2 & h_2''(x)h_3''(x) & . & . & h_2''(x)h_q''(x) \\ h_3''(x)h_1''(x) & h_3''(x)h_2''(x) & h_3''(x)^2 & . & . & h_3''(x)h_q''(x) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_q''(x)h_1''(x) & h_q''(x)h_2''(x) & h_q''(x)h_3''(x) & . & . & h_q''(x)^2 \end{bmatrix} \tag{3.84}$$

Por tanto $$J\left(S(X)\right) = \boldsymbol{\beta}^T \left[\int_{\mathbb{R}} \boldsymbol{D}_H''(x)\, dx\right]\boldsymbol{\beta} = 2\boldsymbol{\beta}^T \boldsymbol{\Omega}_H \boldsymbol{\beta} \tag{3.85}$$

dónde $\boldsymbol{\Omega}_{\boldsymbol{H}} = \int_R \boldsymbol{D}_{\boldsymbol{H}}''(x)\, dx$. De este modo, dadas las funciones de base $\{h_r\}_{r=1}^{q}$, se pueden evaluar siempre los elementos de la matriz $q \times q$ $\boldsymbol{\Omega}_{\boldsymbol{H}}$, lo que permite que la penalización pueda expresarse como una forma cuadrática en el vector de parámetro $\boldsymbol{\beta}$ (donde la matriz $\boldsymbol{\Omega}_{\boldsymbol{H}}$ no depende de $\boldsymbol{\beta}$). En el caso de splines base los elementos de $\boldsymbol{\Omega}_{\boldsymbol{H}}$ pueden encontrarse en GREEN y SILVERMAN (1994) y WAHBA (1990).

Una utilidad importante de (3.85) es que proporciona una vía práctica para aplicar penalizaciones al serpenteo como parte del modelo de ajuste, como se muestra en el apartado 3.3.4, dedicado a los *splines penalizados*.

Se pueden desarrollar otras medidas del serpenteo usando la misma aproximación general. Por ejemplo, $\int_{\mathbb{R}}\left[S'(x)\right]^2 dx$, $\int_{\mathbb{R}}\left[S'''(x)\right]^2 dx$ pueden ser tratadas en la misma vía que $J\left(S(X)\right) = \int_{\mathbb{R}}\left[S''(x)\right]^2 dx$.

Una cuestión clave en la estimación semiparamétrica de la función de calificación de acreditados es que se puede hacer el modelo más flexible agrandando el espacio usando combinaciones lineales de funciones de base para las variables originales, en este caso splines de suavizado.

Desde el punto de vista de la clasificación, se puede ver un spline de suavizado ajustado como una frontera lineal en el espacio agrandado con el objetivo de separar a las observaciones default de las no default. La separación es mejor que en el espacio original de los datos de entrenamiento, y la traslada a la frontera no lineal en el espacio original.

Tanto los *Splines de Suavizado* como los *Splines de Regresión* se utilizan en técnicas de suavizado, los primeros para modelos regularizados. Ambos métodos están orientados a la construcción de modelos con funciones de base, por lo que es posible controlar el serpenteo de la función ajustada controlando el número de funciones de base usadas, aunque como hemos visto esto puede crear dificultades. Ambas técnicas presentan inconvenientes: en los splines de regresión, la suavidad de la función ajustada depende de la elección de los nudos, y ha de hacerse, en general, mediante complicados algoritmos que son difíciles de extender al caso multidimensional. En el caso de los splines de suavizado, los problemas son de tipo computacional, ya que este tipo de splines, utilizan tantos nudos (y, por tanto, parámetros) como observaciones. Como de costumbre la virtud está en un punto intermedio de equilibrio. Este equilibrio lo aportan los S*plines con Penalización* que pasamos a ver a continuación.

3.3.3.4 Splines Penalizados.

Los *splines con penalización* combinan lo mejor de los *splines de regresión* y de los *splines de suavizado, utilizan menos nudos que los splines de suavizado que poseen tantos nudos como observaciones, pero la penalización provoca que la selección de los nudos no sea tan determinante como en los splines de regresión.*

En general, los splines penalizados son splines de rango bajo, en el sentido de que el número de nudos es mucho menor que la dimensión de los datos, al contrario de lo que ocurre en el caso de los splines de suavizado, lo que hace que sean computacionalmente eficientes, sobre todo cuando se trabaja con gran cantidad de datos. Además, la introducción de penalizaciones relaja la importancia de la elección del número y la localización de los nudos, cuestión que es de gran importancia en los splines de rango bajo sin penalizaciones (RICE y WU (2001)).

Aplicando una penalización "ondulante" a la función objetivo del ajuste se evitan modelos excesivamente serpenteantes. Un gran número de nudos significa que el modelo ajustado es bastante insensible a la elección de la localización exacta de los nudos, pero puede usarse la penalización para evitar el peligro de sobreajuste que con frecuencia acompaña al uso de muchos nudos, (WABHA (1980), BREIMAN et al. (1984), O'SULLIVAN (1986), EILERS y MARX (1996), BARLETT y MENDELSON (2002), WAND y ORMEDOR (2008)).

Definición 3.23.- De forma general, se define un *spline penalizado* por la expresión

$$\hat{\boldsymbol{\beta}}^T \boldsymbol{H}(\boldsymbol{X}) \tag{3.86}$$

donde $\boldsymbol{H}(\boldsymbol{X})$ es la matriz de diseño del vector de funciones de base $\left(h_1(X),...,h_q(X)\right)$, y $\hat{\boldsymbol{\beta}} = \left(\beta_1,...,\beta_q\right)^T$ es la solución del *Problema de Minimización de la pérdida Regularizada* siguiente

$$\underset{\boldsymbol{\beta}}{Mín}\ \frac{1}{N}\sum_{i=1}^{N}\ell(y_i,\boldsymbol{\beta}^T\boldsymbol{H}(x_i)) + \lambda\boldsymbol{\beta}^T\Omega_H\boldsymbol{\beta} \tag{3.87}$$

Siendo Ω_H una matriz semidefinida positiva, que penaliza los coeficientes de forma suave, y $\lambda \in \mathbb{R}$, $\lambda > 0$ el parámetro de suavizado. Según la función de pérdida que se utilice tendremos splines logísticos penalizados, probit penalizados etc.

3.4 BONDAD DE AJUSTE.

3.4.1 Introducción.

Una vez que se ha estimado un modelo de predicción, minimizando cierta pérdida media desde una muestra aleatoria simple de entrenamiento, es necesario evaluar la bondad de ajuste, evaluación que consiste en medir las discrepancias entre los valores pronosticados por el estimador de la función de calificación de acreditados, $\hat{S}(x)$, y los valores observados de la variable aleatoria respuesta Y.

De hecho, $\hat{S}(x)$ se ha obtenido por el principio de inducción, minimizando la suma total de las discrepancias de todos los elementos de la muestra de entrenamiento,

$\sum_{i=1}^{N} \ell\left(y_i, S(x_i)\right)$. Por tanto, una medida de la bondad de ajuste adecuada podría ser el siguiente estadístico

$$\sum_{i=1}^{N} \ell\left(y_i, \hat{S}(x_i)\right) \tag{3.88}$$

Por ejemplo, en el caso logístico, para una muestra aleatoria simple de entrenamiento se tiene

$$\ell_{\Lambda}\left(y_i, \hat{S}(x_i)\right) = -\left[y_i \hat{S}(x_i) - \log\left(1 + e^{\hat{S}(x_i)}\right)\right] = -\log\left(\Lambda\left(\hat{S}(x_i)\right)^{y_i}\left[1 - \Lambda\left(\hat{S}(x_i)\right)\right]^{1-y_i}\right)$$
$$= -\log\left(\hat{P}\left(Y = y_i / X = x_i\right)^{y_i}\left(1 - \hat{P}\left(Y = y_i / X = x_i\right)^{1-y_i}\right)\right) = -\log\left(\hat{P}\left(Y = y_i / X = x_i\right)\right)$$

Por tanto, la discrepancia total para todos los pares $\left(y_i, \hat{S}(x_i)\right)$, $i = 1,...,N$, viene dada por

$$\sum_{i=1}^{N} \ell_{\Lambda}\left(y_i, \hat{S}(x_i)\right) = -\log L\left(Y, \hat{S}(X)\right) = NR_{emp\ell_{\Lambda}}\left(Y, \hat{S}(X)\right) \tag{3.89}$$

donde $L\left(Y, S\left(\hat{X}\right)\right)$ *es la función de verosimilitud* y $-\log L\left(Y, \hat{S}(X)\right)$ es la *log-verosimilitud negativa.*

Para el modelo probit con el mismo razonamiento se llega a un resultado similar,

$$\sum_{i=1}^{N} \ell_{\Phi}\left(y_i, \hat{S}(x_i)\right) = -\log L\left(Y, \hat{S}(X)\right) = NR_{emp\ell_{\Phi}}\left(Y, \hat{S}(X)\right) \tag{3.90}$$

Un primer estadístico de bondad de ajuste para modelos de respuesta binaria viene dada por

$$-2\log ver = -2\log L\left(Y, \hat{S}(X)\right) = -2\sum_{i=1}^{N} \log\left(\hat{P}\left(Y = y_i / X = x_i\right)\right) \tag{3.91}$$

o , en términos del riesgo empírico,

$$-2\log ver = 2NR_{emp\ell_{\Phi}}\left(Y, \hat{S}(X)\right) \tag{3.92}$$

Sea M el modelo ajustado con todas las variables explicativas consideradas y M_0 el modelo con sólo el término intercepto, llamado *modelo nulo,* notaremos la *log-verosimilitud negativa* para ambos modelos respectivamente por $-2\log ver(M)$ y $-2\log ver(M_0)$.

La *log-verosimilitud negativa* es muy útil para comprobar si un modelo M se ajusta mejor a los datos que el modelo nulo, M_0. Esta comprobación puede realizarse definiendo un contraste con la hipótesis nula

H_0: *El modelo* M *no proporciona resultados significativamente diferentes a los que proporciona* M_0

frente a la hipótesis alternativa

H_1: *El modelo* M *proporciona resultados significativamente diferentes a los que proporciona* M_0.

El estadístico de contraste, *chi_cuadrado*, es la diferencia $T = -2\left(\text{logver}(\text{M}) - \text{logver}(\text{M}_0)\right)$, llamada *diferencia chi-cuadrado*. El test será, por tanto,

$$H_0 : \text{-}2\left(\text{logver}(\text{M}) - \text{logver}(\text{M}_0)\right) = 0 \quad \textit{frente a} \quad H_1 : \text{-}2\left(\text{logver}(\text{M}) - \text{logver}(\text{M}_0)\right) \neq 0,$$

Después de algunas pinceladas algebraicas, el test se puede escribir formalmente:

$$H_0 : \frac{\textit{verosimilitud}(M)}{\textit{verosimilitud}(M_0)} = 1, \qquad H_1 : \frac{\textit{verosimilitud}(M)}{\textit{verosimilitud}(M_0)} \neq 1 \tag{3.93}$$

Por otra parte, si M es el modelo investigado y M_{-1} es el modelo reducido en una variable explicativa o en una función de base, puede construirse un test similar al anterior, aunque más general, llamado *test razón de verosimilitud* o *test de log verosimilitud* en la forma

$$H_0 : \frac{\textit{verosimilitud}(M)}{\textit{verosimilitud}(M_{-1})} = 1, \qquad H_1 : \frac{\textit{verosimilitud}(M)}{\textit{verosimilitud}(M_{-1})} \neq 1 \tag{3.94}$$

El uso del estadístico chi-cuadrado como medida de bondad de ajuste ha sido muy criticado, puesto que cuando se dispone de muestras de tamaño grande son mucho más fáciles de aceptar, o menos difíciles de rechazar, modelos más complejos debido a que los tests estadísticos chi-cuadrado están diseñados para detectar cualquier desviación entre el modelo y los datos observados. Añadir más términos a un modelo siempre mejora el ajuste y con muestras grandes se hace más difícil distinguir una mejora "real" del ajuste de una mejora trivial.

Ante este problema ha surgido un grupo de medidas de la bondad de ajuste de modelos muy populares. Estas medidas son muy similares al *coeficiente de determinación* R^2 de la regresión por mínimos cuadrados ordinarios de modelos lineales y se conocen como *pseudo coeficientes de determinación generalizada.*

3.4.2 Pseudo Coeficientes de Determinación Generalizada.

La popularidad del coeficiente R^2 se debe a que es una medida calculada a partir de un modelo ajustado, que toma valores entre 0 y 1, que es mayor cuanto mejor es el ajuste del modelo y que proporciona una clara y simple interpretación. Para los modelos lineales generales estimados por máxima verosimilitud no existe una estadística equivalente al coeficiente R^2 para la bondad de ajuste, por lo que se propusieron medidas pseudo-R^2, (McFADDEN (1974, 1978), MADDALA (1983), COX y SNELL (1989), NAGELKERKE (1991), MITTLBOCK y SCHEMPER (1996)). El nombre de "pseudo" se debe al hecho de que estos coeficientes "se parecen" a R^2 en el sentido de que están en una escala similar, que va de 0 a 1, y donde valores más altos indican un mejor ajuste del modelo.

Existen varias aproximaciones de R^2 en mínimos cuadrados ordinarios, OLS: R^2 como *tasa de variabilidad explicada,* es decir, el cuadrado del coeficiente de correlación y la tasa R^2 *como una mejora del modelo nulo al modelo ajustado.* Estas diferentes aproximaciones conducen por "extensión" a distintas formas de calcular el pseudo-R^2 en regresiones de respuesta categórica.

En 1974 McFadden presentó su *pseudo-coeficiente de determinación,* $R^2(U)$, basado en la log-verosimilitud, según la formulación siguiente:

$$R^2(U) = 1 - \frac{\text{logver}(\text{M})}{\text{logver}(\text{M}_0)} \tag{3.95}$$

La expresión (3.95) se llama también *índice razón de verosimilitud* (GREENE, 2003). La aproximación de McFadden pretende recoger para modelos lineales generalizados los conceptos de R^2 como *tasa de variabilidad explicada* y como *mejora del modelo nulo al modelo ajustado,* respecto de este segundo aspecto, la razón de las verosimilitudes sugiere el nivel de mejora ofrecido por el modelo ajustado (con las variables explicativas) sobre el modelo con solo el intercepto. Por otro lado, se verifica que

$$0 \le R^2(U) < 1 \tag{3.96}$$

Obsérvese que $R^2(U)$ no llega a alcanzar el máximo de 1.

Dado que una verosimilitud se sitúa entre 0 y 1, su logaritmo es menor o igual a cero. Si un modelo tiene una verosimilitud muy baja, entonces el logaritmo de la verosimilitud será mayor que el logaritmo de un modelo más verosímil. Por lo tanto, una proporción de

logaritmo de verosimilitud muy baja indica que el modelo se ajusta mejor que el modelo con sólo intercepto. Si se comparan dos modelos sobre los mismos datos, el coeficiente $R^2(U)$ de McFadden será más alto para el modelo con la mayor verosimilitud.

El coeficiente $R^2(U)$ de McFadden adolece del problema que afecta a las log verosimilitudes, aumentan con la introducción de complejidad en el modelo, por lo que para evitar problemas de sobreajuste se ajusta, como es habitual, de modo que se penalice la complejidad. El coeficiente $R^2(U)\_Ajustado$ de McFadden adopta entonces la forma

$$R^2(U)\_Ajustado = 1 - \frac{\text{logver}(M) - K}{\text{logver}(M_0)} \tag{3.97}$$

O, equivalentemente,

$$R^2(U)\_Ajustado = R^2(U) - \frac{K}{\text{logver}(M_0)} \tag{3.98}$$

Si las variables explicativas en el modelo son efectivas, entonces la penalización será pequeña en relación con la información ganada por la introducción de las variables explicativas. Sin embargo, si se contemplan variables explicativas que no añaden nada al modelo, entonces la penalización se hace sentir y el coeficiente $R^2(U)\_Ajustado$ puede disminuir con la adición de una variable explicativa, aunque el pseudo-R^2 aumente ligeramente.

Como se deduce de (3.98) son posibles valores negativos del $R^2(U)\_Ajustado$ de McFadden.

COX y SNELL (1989), sobre una idea original de MADDALA (1983), presentaron una versión de $R^2(U)$ muy popular, notada por R^2, basada en la medida de verosimilitud, $verosim(M)$. Por definición, $verosim(M)$ es la probabilidad de la variable dependiente dadas las variables independientes. Si M= $\hat{S}(X)$ entonces

$$\begin{aligned} L\left(Y, \hat{S}(X)\right) &= \hat{P}\left((Y = y_1 / X = x_1), \ldots, (Y = y_N / X = x_N)\right) \\ &= \prod_{i=1}^{N} \hat{P}(Y = y_i / X = x_i) \end{aligned} \tag{3.99}$$

Por tanto, la raíz N-ésima del producto de probabilidades, $\left(\prod_{i=1}^{N}\hat{P}\left(Y=y_i / X=x_i\right)\right)^{\frac{1}{N}}$, proporciona una estimación de cada valor de Y.

Cox y Snell definieron su pseudo coeficiente de determinación R^2, según la expresión siguiente:

$$R^2=1-\left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}} \tag{3.100}$$

La relación de las verosimilitudes, $\left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}$, refleja la mejora del modelo completo sobre el modelo de intercepto, de forma que cuanto menor sea esta cantidad, mayor será la mejora y tanto más se aproximará el pseudo-R^2 de Cox y Snell a 1. Podemos decir, por tanto, que la aproximación de Cox y Snell se centra, para modelos lineales generalizados, en medir la *mejora del modelo nulo al modelo ajustado.*

Por otro lado, el pseudo-R^2 de Cox y Snell se puede escribir en la forma

$$R^2=1-\exp\left(\log\left(\left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}\right)\right)=1-\exp\left(-\frac{2}{N}\left(\text{logver}(M)-\text{logver}(M_0)\right)\right) \tag{3.101}$$

es decir,

$$-\log\left(1-R^2\right)=\frac{2}{N}\left(\text{logver}(M)-\text{logver}(M_0)\right) \tag{3.102}$$

Como se muestra en SHTATLAND y BARTON (1998) la parte derecha de (3.102) puede interpretarse como *la cantidad de información ganada cuando se incluyen variables explicativas en comparación con el modelo con solo el término intercepto.*

Por simplificación de la notación, en adelante nos referiremos al pseudo-R^2 de Cox y Snell como

$$R^2=1-\text{antilog}\left(\frac{2}{N}\left(\text{logver}(M)-\text{logver}(M_0)\right)\right) \tag{3.103}$$

Una seria desventaja del pseudo-R^2 de Cox y Snell es que su valor máximo no es 1, como demuestra el hecho de que incluso si el modelo ajustado pronostica el resultado a la perfección con probabilidad 1, entonces $R^2=1-\left(verosim(M_0)\right)^{\frac{2}{N}}<1$. Por ejemplo, es posible

que el modelo ajuste perfectamente y los residuos sean cero y R^2 de Cox y Snell = 0.75. El rango de este estadístico es el que se muestra en la expresión siguiente

$$0 \le R^2 \le 1 - \left(verosim(M_0)\right)^{\frac{2}{N}} \tag{3.104}$$

El hecho de que el máximo del R^2 de Cox y Snell pueda ser (y normalmente lo es) inferior a 1 conlleva que este estadístico sea difícil de interpretar. NAGELKERKE (1991), sobre una idea de CRAGG Y UHLER (1970), elaboró una versión ajustada del coeficiente R^2 de Cox y Snell para asegurarse de que varia entre 0 y 1. La idea de Nagelkerke consiste en dividir el R^2 de Cox y Snell por su máximo, $1-\left(verosim(M_0)\right)^{\frac{2}{N}}$, a fin de lograr una medida que oscile entre 0 y 1. Por tanto, el pseudo-R^2 de Nagelkerke, que notaremos $\tilde{R}^2$, será normalmente mayor que el pseudo-R^2 de Cox y Snell, y adopta la forma:

$$\tilde{R}^2 = \frac{1-\left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}}{1-\left(verosim(M_0)\right)^{\frac{2}{N}}} \tag{3.105}$$

O, equivalentemente,

$$\tilde{R}^2 = \frac{R^2}{R^2_{máx}} = \frac{1-\text{antilog}\left(\frac{2}{N}\left(\text{logver}(M)-\text{logver}(M_0)\right)\right)}{1-\text{antilog}\left(\frac{2}{N}\text{logver}(M_0)\right)} \tag{3.106}$$

Si el modelo ajustado predice perfectamente los resultados y tiene verosimilitud 1, el pseudo-R^2 de Nagelkerke es $\tilde{R}^2$ = 1. Sin embargo, si el modelo completo no mejora al modelo con sólo intercepto, el pseudo-R^2 de Nagelkerke es mayor que cero, $\tilde{R}^2$ >0, por lo que el rango total [0,1] de los mínimos cuadrados ordinarios no está todavía cubierto.

Ninguno de los estadísticos pseudo-R^2 considerados aquí puede ser interpretado de forma independiente o comparando distintos conjuntos de datos, son válidos y útiles en la evaluación de varios modelos de predicción sobre la misma muestra de datos con la misma variable respuesta. En esta situación, un pseudo R-cuadrado más alto indica que el modelo predice mejor la respuesta.

3.4.3 Criterios de Información, AIC y BIC.

Los test de bondad de ajuste que hemos visto hasta ahora basados exclusivamente en la log-verosimilitud negativa, conducen con frecuencia a rechazar modelos aceptables a la vez que a aceptar algunos que resultan menos parsimoniosos de lo que debieran, lo que va en contra de la capacidad de generalización y facilidad explicativa requerida al modelo.

Cuanto más complejo es el modelo (por ejemplo, más parámetros) mejor es el ajuste y, por tanto, más alto es el valor de la verosimilitud que se obtiene, o, en otros términos, a mayor verosimilitud mayor sobreajuste, lo que hace necesarios otros criterios alternativos para medir la bondad del ajuste. Los primeros criterios alternativos que se propusieron se basaron en la *log-verosimilitud negativa penalizada*, Criterio de Información de AKAIKE (1973), AIC, y Criterio de Información Bayesiano de SCHWARZ (1978), BIC, en los que el término de penalización es el encargado de corregir la complejidad del modelo, por lo que se han ido haciendo cada vez más populares.

Con frecuencia la pérdida empírica, $L_{emp\ell}\left(Y,S(X)\right)$, tiende a ser excesivamente optimista y para aislar este factor es necesario definir el concepto de *pérdida empírica dentro de la muestra de entrenamiento,* lo que nos conducirá a los criterios AIC y BIC.

Definición 3.24.- Se llama *pérdida empírica dentro de la muestra de entrenamiento* $\boldsymbol{\tau_e} = \left\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N} \in \left(X \times Y\right)^N$ a la cantidad

$$L_{empin\ell}\left(Y,S(X)\right) = \sum_{i=1}^{N}\left[\ell\left(y_i^{nuevo}, \hat{S}(x_i)\right)/\tau\right] \tag{3.107}$$

donde y_i^{nuevo} indica que se observan nuevos valores respuesta en cada uno de los elementos de la muestra de entrenamiento x_i, $i = 1,\ldots,N$.

Definición 3.25.- Se llama optimismo de la *pérdida empírica a la diferencia*

$$Op \leftarrow L_{empin\ell}\left(Y,S(X)\right) - L_{emp\ell}\left(Y,S(X)\right) \tag{3.108}$$

Normalmente el optimismo es positivo, puesto que $L_{emp\ell}\left(Y,S(X)\right)$ es habitualmente sesgado hacia abajo, debido al ruido cuando se remuestrean observaciones de entrenamiento.

El optimismo medio es $\overline{Opt} = E_Y\left(Opt\right)$ y, por tanto, de (3.108) se sigue que

$$\overline{Opt} = E_Y\left[Opt\right] = E_Y\left[L_{empin\ell}\left(Y, S\left(X\right)\right)\right] - E_Y\left[L_{emp\ell}\left(Y, S\left(X\right)\right)\right] \tag{3.109}$$

Como se puede demostrar, para las pérdidas cuadráticas, logística, probit, etc., se tiene

$$\overline{Opt} = 2\sum_{i=1}^{N} C\,\mathrm{ov}\left(\hat{y}_i, y_i\right) \tag{3.110}$$

donde Cov significa covarianza. Cuanto más difícilmente ajustemos los datos mayor será $C\,\mathrm{ov}\left(\hat{y}_i, y_i\right)$ por lo que aumentará el optimismo. En resumen, se verifica una importante igualdad

$$E_Y\left[L_{empin\ell}\left(Y, S\left(X\right)\right)\right] = E_Y\left[L_{emp\ell}\left(Y, S\left(X\right)\right)\right] + 2\sum_{i=1}^{N} C\,\mathrm{ov}\left(\hat{y}_i, y_i\right) \tag{3.111}$$

Nótese que la expresión (3.111) se simplifica si los y_i se obtienen por un ajuste lineal con p variables explicativas o funciones de base. El optimismo crece linealmente con el número p de entradas o funciones de base utilizadas pero decrece cuando el tamaño de la muestra de entrenamiento crece.

Para modelos lineales y datos binarios se tiene

$$E_Y\left[L_{empin\ell}\left(Y, S\left(X\right)\right)\right] \simeq E_Y\left[L_{emp\ell}\left(Y, S\left(X\right)\right)\right] + 2\,p\sigma_\varepsilon^2 \tag{3.112}$$

AKAIKE (1973) propuso el criterio AIC, basado en la teoría de la información, para estimar el riesgo empírico para modelos paramétricos cuando se utiliza como función de pérdida la log-verosimilitud. En este caso, cuando $N \to \infty$, se verifica asintóticamente que la expresión (3.111) es equivalente a

$$-2E\left[N \log P_{\hat{\beta}}\left(Y\right)\right] \simeq -2E\left[\mathrm{logver}\right] + 2\,p \tag{3.113}$$

donde $P_\beta\left(Y\right)$ es una familia de densidades para Y, (conteniendo la verdadera densidad), $\hat{\beta}$, es el estimador de máxima verosimilitud de β, p es el número de variables explicativas de entrada y logver es la log-verosimilitud maximizada

$$\mathrm{logver} = \sum_{i=1}^{N} \log P_{\hat{\beta}}\left(y_i\right) \tag{3.114}$$

Para el modelo de regresión logística, usando la log-verosimilitud binomial, se tiene

$$AIC = -2\mathrm{logver} + 2\,p \tag{3.115}$$

Para modelos no lineales, es necesario reemplazar p por una medida de la complejidad del modelo.

AIC es asintótico por lo que se requieren muestras grandes, además, el número máximo de parámetros no puede exceder $2pN$, donde N es el número de observaciones y, por último queremos resaltar que existen casos en los que AIC decrece monótonamente, es decir, no existe solución, (en la mayoría de los casos el culpable es la mala selección del tipo de modelo).

Por otra parte, el Criterio de Información Bayesiana, (BIC), o Criterio de Schwarz, SCHWARZ (1978), al igual que AIC y también basado en la teoría de la información, es aplicable también en escenarios donde el ajuste se realiza a través de la maximización de la log-verosimilitud. La fórmula genérica de BIC es

$$BIC = -2 \ \text{logver} + \log(N)\, p \tag{3.116}$$

Cuando $N > e^2 \approx 7.4$, BIC tiende a penalizar modelos complejos, dando preferencia a los modelos simples en la selección.

Para AIC y BIC cuanto menor es su valor, mejor es el ajuste. Dos de las mayores fortalezas de estas dos medidas son:

- Se puede comparar el ajuste de diferentes modelos, incluso cuando los modelos no están anidados. Esto es particularmente útil cuando se tienen teorías que son muy diferentes. La idea básica es comparar la verosimilitud relativa de los dos modelos en vez de analizar la desviación absoluta de los datos observados de un modelo particular.

- Como medidas de información que son, penalizan la inclusión de variables que no mejoran significativamente el ajuste. En particular, con grandes muestras, las medidas de información pueden conducir a modelos más parsimoniosos.

3.5 EVALUACIÓN Y SELECCIÓN DEL MODELO.

La *evaluación del modelo* supone valorar si el modelo ajustado en la etapa de estimación y ajuste es un modelo válido, más allá de que presente un ajuste adecuado a los datos. La evaluación o diagnóstico del modelo se refiere a la adecuación de los aspectos implicados en la etapa de especificación, en este sentido se han de evaluar posibles errores de especificación de la componente sistemática, de la distribución de probabilidad de la componente aleatoria y de la relación asumida entre ambas componentes del modelo en la fase de especificación.

Otra cuestión a evaluar es el error de predicción sobre la muestra de validación. Una vez ajustado los datos, se obtiene los valores calculados de la función de calificación y se estima el error de predicción sobre dicha muestra. Será preferible, en lo que respecta a este apartado, el modelo con menor error de predicción sobre la muestra de validación.

En la misma línea que para la bondad de ajuste, para evitar el optimismo propio del error empírico sobre la muestra de validación se pueden utilizar los criterios de información AIC y BIC calculados sobre la muestra de validación. Para propósitos de selección del modelo, no existe una clara superioridad de un indicador sobre el otro. BIC es asintóticamente consistente como criterio de selección, lo que significa que si se da una familia de modelos, incluyendo al verdadero, la probabilidad de que BIC pueda seleccionar el correcto se aproxima a 1, cuando el tamaño de la muestra $N \to \infty$. Este no es el caso para AIC, que tiende a elegir modelos que son muy complejos cuando $N \to \infty$. Por otro lado, para muestras finitas, BIC elige frecuentemente modelos que son muy simples, a causa de su fuerte penalización de la complejidad. Tanto AIC como BIC se han revelado muy útiles en la evaluación y selección de modelos de credit scoring.

Por otro lado, en credit scoring es fundamental la capacidad discriminante del modelo, que en modelos de respuesta binaria coincide con su capacidad predictiva. Un modelo con alto poder discriminante es, sin duda alguna, un potente instrumento de predicción sobre la probabilidad de que un solicitante de crédito cumpla o no con sus obligaciones de pago, en un horizonte temporal previamente fijado. En otras palabras, es un modelo con un alto porcentaje de aciertos frente a un bajo porcentaje de fallos. Para la selección de modelos es habitual, por su demostrada eficacia, utilizar como medida de poder discriminante el Área bajo la Curva ROC, AUC, donde tanto la curva como su estadístico asociado se obtiene sobre la muestra de validación. Analizaremos la curva ROC y el estadístico AUC en la sección 4.3 del capítulo 4.

3.6 CAPACIDAD DE GENERALIZACIÓN DEL MODELO. ERROR TEST.

3.6.1 Introducción.

Dado un modelo de predicción $\hat{S}(x)$ que ha sido estimado minimizando cierta pérdida media desde una muestra aleatoria simple de entrenamiento

$$\boldsymbol{\tau} = \left\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N} \subset \left(X \times Y\right)^N \tag{3.117}$$

la capacidad de generalización de $\hat{S}(x)$, cualidad clave en cualquier modelo de credit scoring, se mide a través del error test y del error test esperado.

Definición 3.26.- Dada una muestra test $\boldsymbol{\tau}_t$ aleatoria e independiente de la muestra de entrenamiento, procedente de la población con función de distribución conjunta F_{XY}.

a) Se llama *error test, (o de generalización),* a la pérdida media para el modelo $\hat{S}(x)$ sobre la muestra test $\boldsymbol{\tau}_t$, o, en otras palabras, al error de predicción sobre una muestra test independiente de la muestra de entrenamiento:

$$Error_T = E\left[\ell\left(y, \hat{S}(x)\right) / \tau\right] = \frac{1}{N_t}\sum_{i=1}^{N_t} \ell\left(y_i, \hat{S}(x_i), \tau_t\right) \tag{3.118}$$

que, equivalentemente, puede definirse como:

$$Error_T = E_{x^{nuevo} y^{nuevo}}\left[\ell\left(y^{nuevo}, \hat{S}(x^{nuevo})\right) / \tau\right] \tag{3.119}$$

donde $\left(x^{nuevo},\ y^{nuevo}\right)$ es un nuevo punto de la muestra test independiente $\boldsymbol{\tau}_t$.

Promediando, (por ejemplo, bootstraping), sobre un número "suficiente" de muestras de entrenamiento se llega al error test esperado.

b) Se llama *error test esperado*, (o *error de predicción esperado*), a la esperanza matemática del error de generalización

$$E\left[Error_T\right] = E\left[E\left[\ell\left(y, \hat{S}(x)\right) / \tau_t\right]\right] \tag{3.120}$$

o, equivalentemente,

$$E\left[Error_T\right] = E_\tau\left[E_{x^{nuevo} y^{nuevo}}\left[\ell\left(y^{nuevo}, \hat{S}(x^{nuevo})\right) / \tau\right]\right] \tag{3.121}$$

que es más manejable desde el punto de vista estadístico. En la práctica muchos métodos estiman el error test esperado en lugar del error de generalización.

Nótese que la muestra de entrenamiento τ está fijada tanto en la expresión (3.118) como en la expresión (3.119) y el error test o de generalización se refiere al error de generalización del modelo finalmente elegido para esta muestra específica, es decir, es el riesgo de clasificación errónea de las observaciones de la muestra τ_t por el modelo ajustado $\hat{S}(x)$ sobre la muestra de entrenamiento τ .

En nuestro caso de respuesta cualitativa, el error de generalización, $Error_T$, es el error de clasificación errónea de la población que presenta el clasificador entrenado sobre τ, y el error de predicción esperado, es el error de clasificación errónea esperada.

Si bien nuestro objetivo es la estimación de $Error_T$ para el modelo estimado $\hat{S}(x)$, desde el punto de vista estadístico es más manejable $E\left[Error_T\right]$, por lo que muchos métodos estiman efectivamente el error esperado. De momento no es posible estimar el error condicional de manera efectiva contando sólo con la información que procede del mismo conjunto que la información de entrenamiento, HASTIE et al. (2009). Cuanto más complejo es el modelo, más datos de entrenamiento son necesarios, y cuantos mas datos de entrenamiento se usen mayor será la capacidad del modelo para adaptarse a estructuras subyacentes más complicadas, lo que implica que se obtenga un decrecimiento en sesgo pero con el contrapunto del crecimiento en varianza. Aunque podemos afirmar que *existe algún modelo intermedio con error test esperado mínimo, desgraciadamente el error de entrenamiento no es un buen estimador del error test*.

El error de entrenamiento decrece con la complejidad del modelo, llegando a caer a cero cuando la complejidad del modelo crece demasiado. Además, un modelo con error de entrenamiento igual a cero está sobreajustado a los datos de entrenamiento, con la incapacidad para la generalización que ello comporta.

Un estimador del error test esperado podría ser la media de los errores test de un número suficiente de muestras test, su obtención no es fácil debido a la escasez de datos en credit scoring, afortunadamente contamos con los métodos bootstraps.

3.6.2.- Estimación del error test por métodos bootstrap.

El método más usual para estimar el error test o de generalización en modelos estadísticos consiste en reservar parte de los datos como un conjunto "test", los cuales no

deben usarse durante el entrenamiento. El conjunto test debe ser una muestra representativa de los casos que se busca generalizar después del entrenamiento. Ejecutar el modelo sobre el conjunto de entrenamiento y el error sobre el conjunto test proporciona un estimador insesgado del error de generalización, a condición de que el conjunto test sea elegido aleatoriamente.

Los métodos basados en la partición de la muestra estiman directamente el error test esperado, $Error = E\left[Error_T\right] = E\left[\ell\left(y, \hat{S}(x)\right)\right]$, es decir, la media del error test o de generalización cuando el modelo $\hat{S}(x)$ es aplicado a una muestra test independiente desde la distribución conjunta de X e Y, (WEISS y KULIKOWSKI (1991)). Los dos métodos más usuales son:

(1).- *Validación Cruzada.* La validación cruzada es una mejora en la validación muestra-partida que permite utilizar todos los datos para el entrenamiento. La desventaja de la validación cruzada es que el modelo debe aprender habilidades nuevas muchas veces.

(2).- ***Métodos Bootstrap***. El bootstrap es una herramienta general para validar el ajuste estadístico.

En 1979, Bradley Efron, (EFRON, 1979), publicó el análisis formal del M*étodo Booststrap,* término que procede de la expresión inglesa "to pull oneself up by one´s bootstrap" (que podría traducirse por: *levantarse mediante el propio esfuerzo*), tomada de una de las aventuras del Barón Manchausen, personaje del siglo XVIII creado por Rudolph Erich Raspe, en el cual el barón había caído al fondo de un lago profundo y, cuando creía que todo estaba perdido, tuvo la idea de ir subiendo tirando hacia arriba de los cordones de sus propias botas, (bootstrap).

El método bootstrap se basa en la analogía entre la muestra y la población de la cual se extrae la muestra. De acuerdo con EFRON y TIBSHIRANI (1986), dada una muestra con N observaciones, el estimador no paramétrico de máxima verosimilitud de la distribución poblacional es la función de densidad de probabilidad que asigna una masa de probabilidad $\frac{1}{N}$ a cada una de las observaciones. La idea central es que *muchas veces puede ser mejor extraer conclusiones sobre las características de la población a partir de los datos obtenidos en la muestra, que haciendo supuestos poco realistas sobre la población.*

La esencia del método bootstrap consiste en que en ausencia de otra información, los valores de una muestra aleatoria son la mejor representación de la distribución de la población y remuestrear la muestra nos proporciona la mejor información sobre lo que sucedería si remuestreamos la población, (EFRON Y TIBSHIRANI, (1993) y MANLY, (1997)).

Los procedimientos basados en los métodos bootstrap implican obviar los supuestos sobre la distribución teórica que siguen los estadísticos. En su lugar, la distribución del estadístico se determina simulando un número elevado de muestras aleatorias construidas directamente a partir de los datos observados. Es decir utilizamos la muestra original para generar a partir de ella nuevas muestras que sirvan de base para estimar inductivamente la forma de la distribución muestral de los estadísticos, en lugar de partir de una distribución teórica asumida a priori.

Este enfoque tiene un antecedente inmediato en la técnicas de Monte Carlo, las cuales consisten en extraer un número elevado de muestras aleatorias de una población conocida, para calcular a partir de ellas el valor del estadístico cuya distribución muestral pretende ser estimada.

Sin embargo, en la práctica no solemos conocer la población y lo que manejamos es una muestra extraída de ella. El investigador parte de un conjunto de datos observados, que constituyen una muestra extraída de la población que pretende estudiar. Cuando las técnicas de Montecarlo son aplicadas a la resolución de problemas estadísticos, partiendo de datos observados en una muestra, reciben más apropiadamente la denominación de *técnicas de remuestreo.*

El método bootstrap es simple y directo para calcular los sesgos aproximados, desviaciones estándar, intervalos de confianza, etc., en casi cualquier problema de estimación no paramétrico. Debido a que el sustento teórico matemático del bootstrap es bastante complejo, hasta finales de la década de 1980, la eficiencia del método era probada de manera empírica, es decir, en el terreno de la práctica.

Los procedimientos de remuestreo en general, han comenzado a centrar la atención de los estadísticos a partir de la década de los años ochenta, cuando el desarrollo de la informática allanó los obstáculos prácticos unidos a la simulación de un número elevado de muestras. A finales de esta década, la utilización del bootstrap para el contraste de hipótesis empezaba a ser considerada una alternativa a los test paramétricos y no paramétricos convencionales (NOREEN, 1989).

Formalmente la idea básica de la estimación Bootstrap, en síntesis, consiste en tratar la(s) muestra(s) como si fuera la población, (debido a la analogía entre muestra y población). Y a partir de ella extraer con reposición un gran número de muestras de tamaño N. De este modo, aunque cada "remuestra" tendrá el mismo número que la muestra original, mediante el remuestreo con reposición cada una podría incluir algunos de los datos originales más de una vez.

Como resultado cada remuestra será, muy probablemente, algo diferente de la muestra original; con lo cual un estadístico $\hat{\theta}^*$, calculado a partir de una de estas remuestras tomará un valor diferente del que produce otra remuestra y del $\hat{\theta}$ observado. *La afirmación fundamental del bootstrap es que una distribución de frecuencias de esos* $\hat{\theta}^*$ *calculados a partir de las remuestras es una estimación de la distribución muestral de* $\hat{\theta}$ (MONEY y DUVAL, 1993).

Cuando el valor del parámetro θ de una población, sobre la que se ha definido un vector aleatorio $(X,Y) \in \mathbb{R}^{p+1}$, es desconocido, tradicionalmente se obtiene una muestra aleatoria de tamaño N, $\tau = \{(x_i, y_i)\}_{i=1}^{N}$, que por simplificación notaremos $Z = (z_1, z_2, ..., z_N)$ donde $z_i = (x_i, y_i)$, y se utiliza un estimador del mismo $\hat{\theta} = f((x_1, y_1), ..., (x_N, y_N), \theta)$, la distribución de $\hat{\theta}$ se aproxima a través de la muestra Z. En el caso de boostrap, se generan B remuestras independientes a partir de Z, $\{Z_b\}_{b=1}^{N}$, calculando para cada remuestra el estadístico $\hat{\theta}_j^*$, que se utiliza como estimador del parámetro poblacional θ, en cuyo valor aproximado estamos interesados. Se trata de tener un número elevado de estimaciones $\hat{\theta}_j^*$. Aunque para obtener el número total de todas las posibles muestras bootstrap $(N)^N$ el tiempo requerido de ordenador puede ser considerable, en la práctica no es necesario extraer tal número total de muestras ya que a veces se logra convergencia con aproximadamente 1000 muestras, o incluso con menos.

De acuerdo con la idea central en que se basa el método bootstrap, el procedimiento supone utilizar la muestra considerando que en si misma contiene la información básica sobre la población. Por tanto, la adecuación de este método será mayor, cuanta más información aporte la muestra sobre la población.

Una consecuencia directa es que a medida que aumenta el tamaño de la muestra mejor es la estimación que podemos hacer sobre la distribución muestral de un estadístico. No obstante, incluso con muestras pequeñas, entre 10 y 20 casos, el método booststrap puede ofrecer resultados correctos. Con un tamaño suficientemente grande, el incremento en el número de muestras procurará una mejora en la estimación de la distribución muestral.

Resumiendo, en términos generales*, los métodos bootstrap son aquellos que se basan en el muestreo con reemplazamiento de una muestra para estudiar las propiedades estadísticas de los estimadores derivados de esa muestra.*

Método bootstrap paramétrico: se supone un modelo paramétrico predeterminado a partir del cual se realiza la simulación, es decir, se crean nuevos datos; los datos de entrada en el modelo, son sustituidos por su función de densidad. El modelo se repite un número suficientemente grande de veces y las propiedades estadísticas de las salidas del modelo se analizan a través de su distribución. Su efectividad depende de la suposición sobre que distribuciones estadísticas son las que mejor se ajustan a los parámetros o variables que deseamos simular.

Método bootstrap no-paramétrico: se lleva a cabo por medio de la distribución obtenida directamente de los datos. La idea consiste en generar observaciones a partir de la distribución de una muestra aleatoria independiente obtenida de la población en estudio.

La diferencia está en función de que el remuestreo se produzca sobre una distribución teórica o una distribución empírica. También puede estar condicionado al ajuste del modelo, es decir, se remuestrean los residuos del modelo en vez de los datos observados, entonces es llamado *boostrap condicionado.*

En resumen, las diferentes versiones de bootstrap se distinguen por el estimador $\hat{F}$ que utilizan:

- *Bootstrap paramétrico*, si se supone que $\hat{F}$ pertenece a un modelo paramétrico $\{F_\theta : \theta \in \Theta\}$, entonces $\hat{F} = F_{\hat{\theta}}$.
- *Bootstrap no-paramétrico*, si no se hace ninguna hipótesis sobre F, entonces $\hat{F} = F_N$, donde F_N es la función de distribución empírica.

El método bootstrap proporciona una vía de cálculo directo de la incertidumbre evaluando el muestreo de datos de entrenamiento. En esencia el bootstrap es una implementación de

computador de máxima verosimilitud paramétrica o no paramétrica. La ventaja del bootstrap sobre la fórmula de máxima verosimilitud es que este nos permite computar estimadores de máxima verosimilitud de errores estándar y otras cantidades en entornos donde no se dispone de fórmulas.

Los métodos bootstrap pueden proporcionar en situaciones complicadas errores estándar e intervalos de confianza que de una manera analítica serían intratables, sin embargo no siempre funcionan correctamente. Es muy conocido que la varianza y la covarianza bootstrap pueden estar sesgadas en un factor de $\frac{N}{N-1}$; este sesgo es inapreciable con tamaños de muestra mayores de 20 aunque con muestras pequeñas debe ser tenido en cuenta, a parte de esto, el bootstrap puede fallar debido a sus propiedades asintóticas, inexactitud inherente de la muestra y presencia de casos atípicos (DIXON, 2001).

Las propiedades asintóticas se refieren a la facultad del método de converger hacia un determinado valor según aumenta el número de replicas; se ha observado que en determinadas situaciones esta convergencia se produce más lentamente de lo deseable y el método falla si se detiene en un número insuficiente de replicas.

En situaciones de incertidumbre sobre la distribución de un parámetro, los métodos bootstrap proporcionan estimadores más robustos (EFRON y TIBSHIRANI, 1993). También pueden ser útiles en situaciones en que se conoce el modelo del error pero el parámetro a estimar implica procesos complejos y el cálculo analítico de su error no es sencillo. Por otra parte las desventajas de los métodos bootstrap son: la necesidad de desarrollar programas de ordenador adecuados a las circunstancias particulares de cada caso y el tiempo que se emplea en los cálculos, que depende de la complejidad del problema y del número de réplicas.

Puesto que *el único supuesto de los métodos bootstrap es que la distribución de la muestra conserva las propiedades estadísticas de la distribución de la población*; *el bootstrap fallará cuando la distribución muestral no sea representativa de la distribución poblacional; esta última característica no hace que el método sea inferior a otros ya que no hay ninguno suficientemente robusto para este problema, que constituye todo un clásico en la inferencia estadística.*

Utilizaremos el método bootstrap como herramienta general para estimar el error de generalización. El método boostrap pretende estimar el *error test condicional a una muestra de entrenamiento* τ , también llamado *error de generalización*, es decir el error de

predicción sobre una muestra test independiente $Error_T = E\left[\ell\left(y,\hat{S}(x)\right)/\tau\right]$ pero usualmente sólo estima bien el error de predicción esperado $Error = E\left[E\left[\ell\left(y,\hat{S}(x)\right)/\tau\right]\right] = E\left[\ell\left(y,\hat{S}(x)\right)\right]$, donde ambos x e y son obtenidos aleatoriamente desde su distribución conjunta, (población).

En el caso de pocos datos, una aproximación para aplicar el bootstrap como estimador del error de predicción puede ser ajustar el modelo en cuestión sobre un conjunto de remuestras bootstrap obtenidas de la muestra de entrenamiento y entonces comprobar si predice bien la muestra original. Si $\hat{S}^*\left(x_i\right)$ es el valor pronosticado en x_i, desde el modelo ajustado a la b -ésima re-muestra, nuestro estimador es

$$\widehat{Error}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N}\ell\left(yi,\hat{S}^{*b}\left(x_i\right)\right) \tag{3.122}$$

Sin embargo es fácil ver que $\widehat{Error}_{boot}$ no proporciona un buen estimador en general, la razón es que las remuestras están actuando como muestras de ajuste, mientras que la muestra original actúa como muestra test, y estas dos muestras tienen observaciones en común. Este solapamiento puede hacer predicciones sobreajustadas que se verán como irrealmente buenas, y esta es la razón de que la validación cruzada use explícitamente datos no solapados para las muestras de ajuste y test. BREIMAN et al. (1984) muestran este hecho con un ejemplo que se discute en HASTIE et al (2009).

CAPÍTULO 4

VALIDACIÓN Y CALIBRACIÓN DE MODELOS DE CREDIT SCORING.

4.1 INTRODUCCIÓN.

El enfoque IRB de Basilea II permite a las Entidades Financieras construir los modelos del riesgo para sus carteras de crédito a partir de información propia, "con la obligación de estos de utilizar toda la información pertinente de que dispongan para asignar calificaciones a los prestatarios, con el fin determinar el riesgo de sus carteras de crédito; cuanto menor sea la cantidad de información con la que se cuente, más conservadora deberá ser esta asignación", (BCBS (2006), § III, 411).

Toda la información relevante disponible en diferentes fuentes dentro de la Entidad Financiera, tras un preanálisis que la convierta en apta para los cálculos estadísticos, se fusionará en un conjunto de datos sobre los que se construirán los modelos estadísticos para determinar el riesgo asociado a sus carteras. Una vez que un modelo de riesgo de crédito se aplica en la gestión de riesgos de la Entidad este proceso ha de repetirse de forma periódica (por ejemplo, una vez al año).

Por otro lado, los acuerdos de Basilea II, en el enfoque IRB, requieren la validación del proceso anterior (BCBS (2006), § III 500): "*Las Entidades Financieras deberán contar con un buen sistema para validar la precisión y coherencia de los sistemas de calificación, los procesos y la estimación de todos los componentes de riesgo pertinentes. Asimismo, deberán demostrar a sus supervisores que su proceso de validación interna les permite evaluar, de forma consistente y significativa, el funcionamiento de los sistemas de calificación interna y de estimación de riesgos*".

En el Nuevo Acuerdo de Capital de Basilea II se establecen los requisitos que debe de cumplir un modelo de probabilidad de default, PD. Validar un modelo de PD significa verificar en qué medida el modelo cumple los requisitos mínimos de Basilea II. Se distinguen dos formas de validación de un modelo: *la validación teórica y la validación estadística.*

La validación teórica viene requerida por los acuerdos de Basilea II, (BCBS (2006), § III, 402), puesto que se exige la revisión de las teorías y las hipótesis que sustentan el modelo propuesto en un esquema detallado de la teoría y de los supuestos que se requieren: "*En el caso de que la Entidad Financiera utilice modelos estadísticos en su proceso de calificación, deberá documentar sus respectivas metodologías. Dicho material deberá:*

a) Ofrecer una descripción detallada de la teoría, los supuestos y/o las bases matemáticas y empíricas de la asignación de estimaciones a grados, deudores

individuales, posiciones o conjuntos de posiciones, y la(s) fuente(s) de datos utilizada(s) en la estimación del modelo;

b) Establecer un proceso estadístico riguroso (que compruebe el ajuste del modelo tanto fuera de la muestra como fuera del periodo muestral) al objeto de validar el modelo.

c) *Indicar cualesquiera circunstancias que impidan el funcionamiento eficaz del modelo*".

Sin duda alguna, un aspecto clave de la validación teórica de un sistema de calificación de una Entidad Financiera lo constituye la calidad del sistema. Si bien son objeto de esta Tesis Doctoral fundamentalmente los aspectos cuantitativos, es necesario dejar patente aquí que estos no conseguirán la eficacia deseada si se descuida el aspecto cualitativo del sistema de calificación en su conjunto.

Basilea estableció de forma clara ciertos requerimientos de control del riesgo de crédito relacionados con los aspectos cualitativos de los modelos, (BCBS (2006), § III, 441, 442 y 443):

441: La Entidad Financiera *deberá contar con unidades independientes de control del riesgo de crédito, encargadas de diseñar o seleccionar, aplicar y controlar sus sistemas internos de calificación. La unidad o unidades deberán ser funcionalmente independientes del personal y de las unidades administrativas responsables de generar las posiciones. Sus ámbitos de actuación deberán incluir:*

1. *Comprobación y seguimiento de los grados internos;*
2. *Elaboración y análisis de informes sobre el sistema de calificación de la Entidad, abarcando datos históricos de incumplimiento ordenados según su calificación en el momento del incumplimiento y un año antes, análisis de la migración entre grados y seguimiento de las tendencias de los criterios básicos de calificación;*
3. *Aplicación de procedimientos destinados a comprobar que las definiciones de las calificaciones se aplican de manera coherente en los distintos departamentos y áreas geográficas;*
4. *Examen y documentación de cualquier cambio en el proceso de calificación,*
5. *Incluyendo las razones que lo motivaron; y*
6. *Examen de los criterios de calificación al objeto de evaluar si siguen prediciendo el riesgo. Las modificaciones introducidas en el proceso de calificación, en sus criterios*

o en los parámetros individuales utilizados deberán documentarse y conservarse para su examen por parte de las autoridades supervisoras.

442: *La unidad de control del riesgo de crédito deberá participar activamente en el desarrollo, selección, aplicación y validación de los modelos de calificación, y será la responsable de vigilar y supervisar estos modelos, siendo en última instancia responsable de su continua revisión y de los cambios que pudieran introducirse en ellos.*

443: *El departamento de auditoría interna u otra unidad igualmente independiente deberá examinar, al menos anualmente, el sistema de calificación de la Entidad Financiera y su funcionamiento, incluida la operativa de la unidad de créditos y la estimación de las PD, LGD y EAD. Los ámbitos de examen incluirán la observancia de todos los requisitos mínimos aplicables, debiendo documentar sus conclusiones. Algunos supervisores nacionales podrán además solicitar una auditoría externa del proceso de asignación de calificaciones de la Entidad, así como de su estimación de las características de pérdida".*

La validación teórica está formada por la revisión de, entre otros aspectos, los supuestos que subyacen en el modelo teórico, de la representatividad e integridad de los datos sobre los que se construirán el modelo, la idoneidad de las variables explicativas y, por último, la capacidad de reproducir todo el proceso de construcción del modelo. Todos estos aspectos los analizaremos en la sección 4.2.

El aspecto de mayor interés en esta Tesis Doctoral es sin duda el estadístico. Como ya hemos apuntado, en la introducción y en el capítulo 1, el análisis estadístico de los sistemas de clasificación y calificación se basa en la suposición de que hay dos categorías de acreditados en una Entidad Financiera: *los acreditados que incurrirán en default en un horizonte de tiempo predefinido, y los acreditados que no incurrirán en default antes de este límite temporal*.

Por lo general, no se sabe de antemano si un deudor pertenece a la primera o a la segunda categoría. Las Entidades Financieras se enfrentan pues a un problema de clasificación dicotómica (o binaria) ya que tienen que evaluar el estado futuro de un deudor, usando solamente las características que en la actualidad dispone sobre él. Como hemos visto a lo largo de todos los capítulos anteriores, los modelos de calificación pueden ser considerados herramientas de clasificación en el sentido de que proporcionan indicaciones estimadas de la situación previsible respecto del default del deudor. El procedimiento de aplicación de

cualquier instrumento de clasificación para la evaluación de la situación futura de un deudor se llama comúnmente discriminación.

El principio director de la construcción de funciones de calificación puede ser descrito como: "*un sistema de calificación discrimina mejor cuanto más difiere la distribución de la calificación de los acreditados default de la distribución de las calificaciones de los acreditados no default*".

El primer aspecto cuantitativo a validar es el poder discriminante, entendida esta validación como la corrección de la calificación de los acreditados, el requisito de validación consiste en rendimiento de la calificación interna.

El poder discriminante de un modelo es su capacidad para separar los acreditados buenos de los malos, (HARRELL, 2001).

El segundo aspecto cuantitativo a validar es la calibración, entendida la calibración como la corrección de la PD, LGD y EAD, el requisito de validación es la estimación de todos los componentes de riesgo relevantes.

La calibración es la habilidad del modelo para hacer estimaciones insesgadas de las probabilidades de default, (HARRELL, 2001). Decimos que un modelo está bien calibrado, si la fracción de sucesos que ocurren al final del horizonte temporal, se estima en forma no sesgada por el estimador de la PD de estos sucesos al principio de dicho horizonte.

Ambas, discriminación y calibración comparan las probabilidades de default estimadas con la frecuencia observada de default en el conjunto de datos. Usualmente, tal como refleja la literatura sobre los sistemas de calificación, la PD se valida a través del poder discriminante y la calibración. Sin embargo, se puede ampliar el rigor de la *validación, validando también los parámetros del modelo,* $\left(\hat{\beta}\right)$*, en el sentido de comprobar su estabilidad tanto en el tiempo como en los grupos*, en este sentido es importante resaltar que los modelos están destinados a ser utilizados para hacer predicciones, y que las predicciones sólo serán validas si los parámetros son estables en el tiempo, MEDEMA et al. (2009). De hecho Basilea II, (BCBS (2006), § III, 503), requiere que las Entidades Financieras demuestren que los análisis cuantitativos y otros métodos de validación no varíen de forma sistemática con el ciclo económico.

Un rasgo característico de un sistema de clasificación estable es que los modelos contemplen de forma adecuada la relación causal entre los factores de riesgo y el estado de

default, evitando dependencias espurias derivadas de correlaciones empíricas. A diferencia de los sistemas estables, los sistemas inestables con frecuencia muestran un nivel muy decreciente de la exactitud del pronóstico en el tiempo.

En general, un modelo no es capaz de reproducir exactamente los datos en que se basa su construcción. Para determinar la exactitud del modelo la literatura dispone de varias pruebas estadísticas, nosotros estudiaremos las que son más usuales y que se recogen en HARRELL (2001), donde se describe muy claro cómo validar un modelo logit, (con aplicación en las ciencias médicas), la colección de estudios sobre los métodos de validación, recomendados por el grupo de análisis cuantitativo del Comité de Basilea II, BCBS (2005a, 2005b), y en ENGELMANN y RAUHMEIER (2006), donde se proporciona un conjunto de artículos sobre la probabilidad de default, PD, sobre la pérdida en caso de default, LGD, y sobre la exposición al default, EAD.

Existe una colección importante de medidas estadísticas para analizar el poder discriminante, algunas de los cuales se describen en la sección 4.3 de este capítulo. Sin embargo, la medida absoluta del poder discriminante de un sistema de clasificación sólo tiene una significación limitada. Una comparación directa de sistemas de clasificación diferentes sólo se puede realizar si se tiene en cuenta el "ruido estadístico". En general, tanto mayor será el problema del ruido cuanto menor sea el tamaño de la muestra de acreditados default disponible. Por esta razón es muy importante considerar herramientas estadísticas para la comparación de sistemas de calificación. Algunas de las herramientas, en particular, las tasas de precisión, AR, y curvas ROC, tienen en cuenta explícitamente el tamaño de la muestra de default.

Por otra parte, el test del poder discriminante debe hacerse no sólo en el conjunto de datos de entrenamiento, sino también en un conjunto de datos independientes, la muestra test, de lo contrario existe el peligro de que el poder discriminante pueda ser exagerado por el sobreajuste en el conjunto de datos de entrenamiento. En el caso de sobreajuste, el sistema de calificación con frecuencia presenta un poder discriminante relativamente bajo en la muestra test, a pesar de que esta muestra es estructuralmente similar a la muestra de entrenamiento, lo que significa que el sistema de calificación tiene una baja estabilidad.

En la práctica, los sistemas de calificación no se utilizan solamente para la toma de decisiones sobre sí conceder o no un crédito, sino que, además, constituyen la base para el cálculo de los precios de los créditos, de las primas de riesgo y de las cargas de capital. A estos efectos, *cada puntuación de crédito o categoría de calificación debe estar asociada con*

una probabilidad de default que da una valoración cuantitativa de la probabilidad con la que los deudores clasificados de esta manera podrían entrar en default. Además, bajo ambos enfoques IRB, los requerimientos de capital de una Entidad Financiera se determinan por cálculos internos de los parámetros de riesgo para cada exposición. Estos se derivan a su vez de los resultados de calificación interna de la Entidad. El conjunto de parámetros incluye la Probabilidad de Incumplimiento del prestatario, PD, así como en algunos casos la Pérdida Esperada dado el Default, LGD y la Exposición en el momento del Default, EGD. En este sentido se habla también de la calibración del sistema de calificación. Como los parámetros de riesgo se pueden determinar por la propia Entidad Financiera, la calidad de la calibración es un importante criterio prudencial para evaluar los sistemas de calificación.

Comprobar el poder discriminante y verificar la calibración son tareas diferentes. Como la capacidad de la discriminación depende de la diferencia entre las distribuciones de los acreditados default y de los acreditados no default, respectivamente, en las categorías de calificación, algunas medidas de poder discriminante resumen las diferencias de las densidades de probabilidad de estas distribuciones. Por otra parte, se puede medir la variación de las probabilidades de incumplimiento que se asignan a las diferentes categorías.

Por el contrario, la calibración correcta de un sistema de calificación significa que las estimaciones de la PD son "suficientemente correctas". Por lo tanto, en el examen de la calibración se deben considerar de alguna manera las discrepancias entre los pronósticos de default y las tasas de default efectivamente observadas. Esto se puede hacer de forma simultánea para todas las categorías de calificación en una prueba conjunta o por separado para cada categoría de calificación, dependiendo de si se pretende la evaluación global o un examen en detalle.

Para las carteras de crédito, las actuales propuestas de Basilea II van en la dirección de que una serie de 5 años de tasas de impago es suficiente. La independencia estándar de pruebas basadas en el poder discriminante y en la calibración pudiera ser parcial cuando se aplica a las carteras de crédito. En particular, con respecto a la calibración se trata de un problema importante al que el grupo de validación de Basilea II prestó una atención considerable.

En este capítulo se analizan las actuales técnicas de validación estadística, tanto para el poder discriminante, sección 4.3, como para la calibración (o cuantificación de la PD-) de

un sistema de calificación, sección 4.4, y se evalúa su utilidad a efectos de supervisión. Una consecuencia importante de las conclusiones del grupo es que cualquier aplicación de una técnica estadística debe ser complementada por los controles de calidad. Esto es importante ya que el uso acrítico de las técnicas pueden conducir a resultados engañosos.

Por otra parte, la elección de una técnica específica para aplicar a la validación debe depender de la naturaleza de la cartera en cuestión. Las carteras al por menor o las carteras de pequeñas y medianas empresas con grandes registros de datos de default son mucho más fáciles de estudiar con métodos estadísticos que, por ejemplo, las carteras de los soberanos o las instituciones financieras donde los datos de default son escasos.

Para terminar esta introducción queremos presentar un resumen de los requerimientos más importantes que respecto de la validación de los sistemas de calificación, y, por tanto, de los modelos estadísticos se contemplan en los acuerdos de Basilea II:

1. Las Entidades Financieras periódicamente deben comparar las tasas de default con la PD estimada para cada grado de calificación y han de ser capaces de demostrar que las tasas de default están dentro del rango esperado para cada grado.
2. Las Entidades Financieras que utilicen el método IRB avanzado deberán completar el análisis de este tipo para sus estimaciones de LGD y de EADS.
3. Las Entidades Financieras también deberán emplear otras herramientas de validación cuantitativa y hacer comparaciones con fuentes de datos externas.
4. Las Entidades Financieras deben demostrar que los métodos de análisis cuantitativos y otros métodos de validación no varían de forma sistemática con el ciclo económico.
5. Las Entidades Financieras deben tener bien definidas las normas internas para situaciones en las que las desviaciones entre PD, LGD y EAD de sus valores esperados sean lo suficientemente importantes como para cuestionarse la validez de las estimaciones en cuestión. De acuerdo con (503), estas normas deben tener en cuenta los ciclos económicos y otras variaciones sistemáticas de índole similar.

4.2 VALIDACIÓN TEÓRICA.

4.2.1 Validación de los Supuestos Teóricos del Modelo

La revisión de los supuestos que subyacen en el modelo teórico es parte de la validación teórica, por lo que habrá de hacerse ésta considerando que las teorías asociadas con los modelos de la probabilidad de default han de ser pensadas como teorías económicas sobre los factores más importantes que afectan al riesgo de que se produzca un incumplimiento de las obligaciones de pago de los acreditados. De ahí que el epígrafe § 411 requiera que si falta algún factor importante de riesgo la entidad Financiera tenga que estimar una PD conservadora.

Un aspecto fundamental de los supuestos del modelo teórico lo constituye la elección de la forma funcional. Lo más habitual en los modelos de credit scoring consiste en utilizar el modelo de regresión logística lineal para estimar la función de calificación de acreditados y, por tanto, de la probabilidad de default. Los supuestos más importantes del modelo son la linealidad del modelo y la función de pérdida logística como transformación de enlace entre que las variables explicativas del riesgo de crédito y el estado de default, es decir, las variables explicativas del riesgo de crédito tienen un efecto lineal en $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)$. Sin embargo, en la práctica, esta relación también puede ser no lineal. Pero es conveniente realizar un test de especificaciones alternativas de la estructura funcional del modelo para comparar los resultados, de modo que si la discriminación o la calibración del modelo lineal no fuese mejor que una propuesta alternativa se tendrá que pensar que la forma funcional del modelo propuesto es demasiado restrictiva.

En general, la estimación de un modelo semiparamétrico podría exigir más datos de los que normalmente están disponibles, pues con frecuencia son pocos los default disponibles, el número efectivo de observaciones default es proporcionalmente muy pequeña en comparación con los acreditados asociados a una cartera (CRAMER, 2004). Por esta razón, actualmente las Entidades Financieras son un poco reacias a utilizar alternativas no paramétricas o semiparamétricas a un modelo LOGIT. La varianza de tales modelos de clasificación tiende a ser más alta (HASTIE et al., 2009). Por otra parte, el banco supervisor holandés, De Nederlandsche Bank N.V. estableció en el año 2005: " la validación también consta de un análisis cualitativo del modelo por medio de la evaluación de la transparencia (no deberá haber cajas negras) y la intuición del modelo " y, en este sentido, las redes neuronales, las técnicas SVM e incluso los árboles de clasificación ni son lo suficientemente transparentes ni proporcionan ninguna orientación sobre la

importancia de los diferentes factores de riesgo. Una posición de equilibrio la representa la utilización de modelo logísticos híbridos por expansiones lineales de funciones de base, HLLM, que presentamos en el capítulo 6, pues hacen compatible la transparencia e interpretación del modelo con el hecho de no renunciar a incorporar al modelo la no linealidad de algunas de las variables de ser necesario o, al menos, conveniente.

4.2.2 Validación de los Datos.

También es parte de la validación teórica la *revisión de la* representatividad e integridad del conjunto de datos. "*Las Entidades Financieras deberá contar con un proceso para examinar los datos que se incorporan como argumentos a los modelos estadísticos de predicción del incumplimiento. Dicho proceso deberá incluir un estudio de la precisión, exhaustividad e idoneidad de los datos utilizados específicamente al asignar una calificación aprobada. La Entidad Financiera deberá demostrar que los datos utilizados para construir el modelo son representativos del universo de sus prestatarios actuales*", (Basilea II, § III. H , 417, BCBS (2006)).

Las Entidades Financieras pueden utilizar datos internos o datos externos para estimar el modelo. Basilea II permite el uso de datos externos, (Basilea II, § III. H , 448, 463 BCBS (2006)), pero exige a estas Entidades que demuestren que los datos son representativos. Cuando la Entidad utiliza los datos internos sobre una cartera completa de créditos y un segmento completo de acreditados, los datos son claramente representativos. En la práctica, el conjunto de datos en una cartera completa puede ser demasiado grande para estimar el modelo, por lo que es usual en este caso utilizar en su lugar una muestra aleatoria. El procedimiento de muestreo tiene que ser revisado para determinar si el subconjunto es representativo de la población subyacente.

Con respecto a la *integridad de los datos* Basilea II exige que la duración del periodo de observación subyacente a los datos habrá de consistir por menos en cinco años, (Basilea II, § III. H , 463, BCBS (2006)): "*con independencia de que la Entidad Financiera utilice fuentes de datos externas, internas, o una combinación de ambas, al estimar la PD, la duración del periodo histórico de observación deberá ser como mínimo de cinco años para al menos una de las fuentes. Si el periodo de observación disponible es más largo en el caso de alguna de las fuentes y estos datos son relevantes y pertinentes, deberá utilizarse este periodo más dilatado".*

En la práctica, podría ocurrir que algunas Entidades Financieras dispongan todavía de información para un período inferior a cinco años. Esto significa que el conjunto de datos

es incompleto. Por supuesto, este problema de los datos incompletos se puede resolver con el tiempo a medida que se vaya disponiendo de más información. Cuándo el período de observación subyacente es menor de cinco años las Entidades Financieras están autorizados a utilizar datos externos para estimar el modelo. Allí dónde se utilicen datos externos debe agregarse un margen de cautela (Basilea II, § III. H, 451, 462, BCBS (2006)). La falta de datos puede presentarse de otras muchas formas. A menudo falta la información de algunas variables para una serie de observaciones, lo que significa que hay menos información disponible y por consiguiente, los resultados deben ser interpretados de forma conservadora (Basilea II, § 411). El conservadurismo puede implicar que los resultados de impago del modelo sean considerados como un límite inferior, un poco por encima del cual se puede establecer la estimación final de la PD. Desde un punto de vista estadístico, los datos que faltan son un problema, ya que la mayoría de los métodos estadísticos más usuales requieren conjuntos de datos completos. El método más comúnmente usado para manejar datos faltantes es analizar sólo casos completos. Los casos incompletos se quitan de la serie de datos, lo que puede dar lugar, en el mejor de los casos, a estimaciones insesgadas pero ineficientes, y, en el peor de los casos, a estimaciones sesgadas. Una buena referencia en el análisis de datos faltantes es LITTLE y RUBIN (2002), donde se discuten tanto enfoques clásicos como enfoques más actuales.

4.2.3 Validación de la Idoneidad de las Variables.

Las variables explicativas que se han de considerar en un modelo de credit scoring deberán referirse como mínimo a las características del acreditado, a las características de riesgo de la transacción y a la exposición al incumplimiento de pago por parte del acreditado, (Basilea II, § III. H, 402, BCBS (2006)). Ejemplos de las características del prestatario son la edad, ingresos, estado civil y ocupación. Las características del riesgo de la operación son, por ejemplo, el tipo de crédito, el nominal solicitado, la forma de pago, etc. Y, por último, ejemplos de características de exposición al riesgo son el historial de pago y el historial de incidencias.

Existen varios problemas que pueden surgir con las variables. En primer lugar, los valores de una variable pueden cambiar con el tiempo. En segundo lugar, algunas variables son difíciles de medir, por ejemplo, la medida del default es en sí misma difícil. Según los acuerdos de Basilea II, "*Se considera que el incumplimiento con respecto a un deudor concreto ocurre cuando acontece al menos una de las siguientes circunstancias*, (Basilea II, § III. H, 452, BCBS (2006))*:*

a) La Entidad Financiera considera probable que el deudor no abone la totalidad de sus obligaciones crediticias frente al grupo financiero, sin recurso por parte de la Entidad a soluciones tales como la realización de protecciones (si existieran).

b) *El deudor se encuentra en situación de mora durante más de 90 días con respecto a cualquier obligación crediticia significativa frente al grupo financiero. Se considerará que los descubiertos se encuentran en situación de mora cuando el cliente haya excedido un límite recomendado o cuando se le haya recomendado un límite inferior a su actual saldo deudor*".

Simplificando podríamos decir que se ha producido incumplimiento cuando es improbable que el deudor pague y/o el deudor está en mora más de 90 días. En la práctica es difícil medir cuando es poco probable que pague un deudor.

Los elementos que se utilizan como indicadores de la probabilidad de impago incluyen, (Basilea II, § III. H, 453, BCBS (2006)):

a) La Entidad Financiera asigna a la obligación crediticia la condición de no reditual.

b) La Entidad Financiera cancela la deuda o dota una provisión específica a consecuencia del significativo deterioro de la solvencia percibido tras tomar la posición 0.

c) La Entidad Financiera vende la obligación crediticia incurriendo en una pérdida económica significativa relacionada con la calidad crediticia.

d) La Entidad Financiera acepta una reestructuración forzosa de la obligación crediticia que es probable que resulte en una menor obligación financiera a consecuencia de la condonación o el aplazamiento del principal (capital), intereses o (cuando proceda) comisiones.

e) La Entidad Financiera ha solicitado la quiebra del deudor o una figura similar con respecto a la obligación crediticia del deudor frente al grupo financiero.

f) El deudor ha solicitado la quiebra o se le ha declarado en quiebra o en una situación de protección similar, lo que conlleva la imposibilidad o el aplazamiento del reembolso de la obligación crediticia al grupo financiero".

4.2.4 Reproducibilidad de la Construcción del Modelo.

Un cuarto aspecto, relacionado con la validación cualitativa, consiste en la *capacidad de reproducir todos y cada uno de los pasos dados desde la obtención de los datos, en su aspecto más primario, hasta que el modelo se implementa en la organización.*

La reproducibilidad, o la replicación, de cualquier investigación se define como la duplicación de los resultados de un estudio anterior (McCULLOUGH et al. 2006). La replicación positiva y negativa tiene un valor para el estudio replicado. La reproducibilidad positiva da más apoyo a los resultados de un estudio anterior. Cuando una réplica es negativa, es evidente que se han producido errores en la investigación. Por supuesto, la cuestión a continuación sigue siendo si el estudio original o la reproducción del estudio contienen errores. Para que un investigador sea capaz de reproducir un estudio, la documentación del estudio anterior debe de estar completa. En general, la documentación incompleta hará que sea imposible reproducir los resultados de un estudio. Un segundo problema que hace que sea difícil reproducir los resultados se asocia con los datos. Cuando los datos no se registran y documentan correctamente son completamente inútiles a otro investigador, (DEWALD et al. (1986)). Además, los datos suelen ser revisados cuando se dispone de nueva información por lo que, en este caso, la réplica exacta será imposible.

Según el documento de trabajo *nº 14 del Comité de Supervisión Bancaria del Banco Internacional de Pagos de Basilea II*, BCBS (2005a), un objetivo fundamental de la validación de un sistema calificación de solicitantes de crédito consiste en comprobar su eficiencia. Esta eficiencia se basará, entre otros, en el poder discriminante de los modelos de scoring y en su adecuada calibración. Ambas características de eficiencia del modelo se validan de forma cuantitativa y a ellas dedicaremos las dos siguientes secciones, 4.3 y 4.4

4.3 PODER DISCRIMINANTE.

4.3.1 Introducción.

La definición de HARREL (2001) del *poder de discriminación de un modelo, como su capacidad para separar a las poblaciones objeto de estudio*, que en los sistemas de calificación se traduciría como su *capacidad para separar a los acreditados buenos de los malos*, es muy general y desde luego muy válida para *las posiciones frente acreditados minoristas. La* definición de Harrel fue extendida por TASCHE, BCBS (2005a), (en aplicación de (BCBS (2006), § III, 403): "las Entidades Financieras *distribuirán las posiciones en diversas categorías de forma significativa, sin concentraciones excesivas, utilizando sus baremos de calificaciones*"), adaptándola al caso de *posiciones frente a empresas, soberanos y bancos*, para lo que se basó en *el principio fundamental en la construcción de las funciones de calificación de acreditados de exigir que el calificador asigne una mejor categoría de calificación cuanto menor sea la proporción de incumplidores y mayor sea la proporción de cumplidores*.

En esta Tesis Doctoral se adopta como definición general de *poder discriminante* la dada por TASCHE, BCBS (2005a):

Definición 4.1.- (Tasche, BCBS (2005)). Se dice que un modelo presenta *alto poder discriminante* cuando claramente distingue, a priori, una vez descartado el sobreajuste, entre los acreditados incumplidores, "malos" y los que si cumplen, "buenos".

Un modelo con alto poder discriminante es, sin duda alguna, un potente instrumento de predicción sobre la probabilidad de que un solicitante de crédito cumpla o no con sus obligaciones de pago, en un horizonte temporal previamente fijado. En otras palabras, es un modelo con un alto porcentaje de aciertos frente a un bajo porcentaje de fallos.

La puntuación $S(X)$ asignada a los solicitantes de crédito sintetiza la información proporcionada por un conjunto de variables seleccionadas por su influencia en el comportamiento de los solicitantes frente al incumplimiento de sus obligaciones de pago, por lo que *a cada solicitante se asocian dos variables, una* $S(X)$*, que representa una puntuación sobre una escala continua que le asigna el sistema scoring y que refleja el crédito que para este sistema merece el solicitante, la otra, Y, muestra el estado de cumplimiento o incumplimiento que el acreditado presenta al final del horizonte temporal que se haya fijado, normalmente 1 año.*

La intención con la variable $S(X)$ es "*pronosticar el futuro estado de Y para el acreditado, confiando en la información sobre el merecimiento de crédito que está contenida en* $S(X)$", con *la premisa de que una función de calificación razonable debe asignar bajas puntuaciones a los solicitantes que presenten alta probabilidad de incumplimiento.*

Una característica básica que han de presentar estas funciones es su alta eficiencia para separar los solicitantes en dos grupos, los buenos y los malos, por tanto, las medidas del poder discriminante pueden usarse como medidas de la eficacia del sistema de puntuación.

En esta sección se analizan métodos estadísticos que permitan juzgar si la función de calificación o puntuación es apropiada para discriminar entre los solicitantes *que probablemente cumplan, "buenos", y los que no, "malos"*

En los modelos de scoring es necesario establecer puntos de corte, reglas de decisión y dictámenes con el fin de objetivar al máximo el proceso de concesión de operaciones.

Definición 4.2 .- Se llama *punto de corte Pc* , "cut off", al número real que indica la frontera entre los niveles de calificación de las operaciones, de tal forma que las operaciones que tengan la variable (puntuación) por debajo de ese valor serán calificadas como "malas" y las que tengan un valor superior a Pc se califican como "buenas".

En la figura 4.1 se muestran como se distribuyen, según una función de frecuencias, cada una de las su muestras de acreditados "buenos" y "malos". En el eje X de abscisas se representa la puntuación, para ambos grupos de acreditados, los malos en rojo y los buenos en azul. Nótese que el punto de corte *Pc* es la abscisa del punto de intersección de las curvas de densidad de ambas poblaciones, frontera entre dos áreas de solapamiento de ambas curvas.

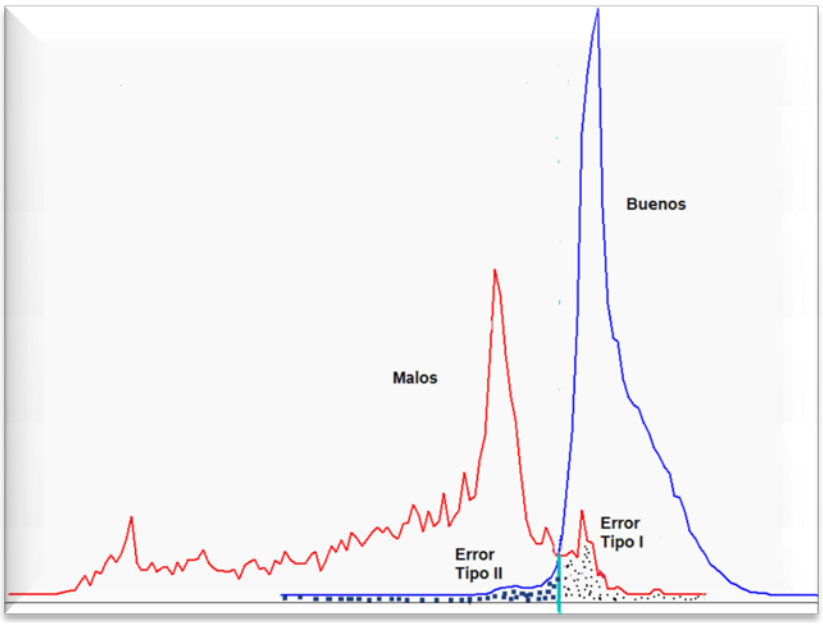


Figura 4.1. - Funciones de distribución de los acreditados "Buenos" y "Malos". Fijémonos como se solapan las dos funciones de frecuencia. Es evidente que si no existiera solapamiento y el área fuese nula, entonces el modelo de calificación sería perfecto.

Para el punto de corte Pc, existen dos Zonas de Error: la zona marcada con cuadrados azules a la izquierda de Pc y la zona marcada con puntos a la derecha de PC, ambas representadas en la figura 4.1, y cuyo significado numérico se recoge en la tabla cruzada 4.1, siendo

a: el número de clientes 'malos' para los que el modelo predice correctamente.

b: el número de clientes 'buenos' para los que el pronóstico es incorrecto.

c: el número de clientes 'malos' con pronóstico incorrecto.

d: el número de clientes 'buenos' para los que el modelo predice correctamente.

Tabla 4.1.- Estado de default estimado y observado para los acreditados.

		Estimados	
		Buenos	Malos
Observados	Buenos	**a** Verdadero	**b** Falsa Alarma
	Malos	**c** Fallo	**d** Éxito

Definición 4.3.- Se dice que se ha realizado *una buena predicción para un punto de corte y para un acreditado concreto* si la predicción de incumplimiento al comienzo del horizonte temporal considerado coincide con el estado de default del acreditado al final de dicho horizonte, en caso contrario se dirá que para ese punto de corte en tal horizonte temporal se ha realizado *una mala predicción*.

Como puede observarse en la tabla 4.1, existen dos casos para los que, dado un punto de corte, ocurre una buena predicción del comportamiento de un acreditado:

1. Al principio del horizonte temporal, normalmente un período de un año, el modelo predijo para un acreditado incumplimiento, y durante dicho período el acreditado incumple. En este caso se dice que se ha tenido un 'éxito'.
2. Cuando el modelo predice cumplimiento antes de comenzar el período y el acreditado no incumple a lo largo del período.

También puede observarse que existen dos casos de predicción errónea:

1. Una primera situación de *mala predicción* se presenta si, para un punto de corte, frente a la predicción de incumplimiento por parte del modelo, el acreditado no presenta actualmente ningún incumplimiento (Tasa de Falsa Alarma, **FAR**, False Alarm Ratio, Tasa β o Error Tipo II). Es el llamado *"Coste de Oportunidad"* o *"Coste de Pérdida de Negocio"* y se obtiene en la forma siguiente

$$Tasa\ de\ error\ tipo\ II = \frac{b}{b+d} \tag{4.1}$$

   Cuando existe una alta tasa de error tipo II, los 'buenos' clientes mal clasificados son rechazados, lo que implica un coste de oportunidad por la pérdida de un buen cliente. Si una institución crediticia mantiene elevada esta tasa durante largo tiempo, lo que significa que durante ese largo período adopta una política de crédito restrictiva, puede perder una importante cuota de mercado.

2. Una segunda situación de mala predicción se presenta cuando se estimó cumplimiento y en la actualidad el acreditado incumple con sus obligaciones de pago. Es el llamado Fallo, Error de Tipo I o Riesgo de Crédito.

$$Tasa\ de\ error\ tipo\ I = \frac{c}{a+c} \tag{4.2}$$

Definición 4.4.- a) La *proporción de éxitos, o tasa de aciertos* (**HR**, *Hit Ratio*), es el porcentaje de acreditados que estimó el modelo como default al principio del horizonte temporal y que, para un punto de corte dado Pc, resultaron default al final del horizonte temporal.

b) La *proporción de falsas alarmas,* (**FAR**, False Alarm Ratio), es el porcentaje de acreditados estimados por el modelo como default y que, para un punto de corte dado Pc, resultaron no default.

c) El *error total de clasificación* es la suma del error de tipo I y el error tipo II.

Ambas tasas, HR, error tipo I, y FAR, error tipo II, dependen obviamente del valor de corte Pc. La importancia de los dos tipos de error es obviamente diferente en los problemas de credit scoring, a causa de que la predicción correcta del riesgo de default es más importante debido al alto coste que suele conllevar la clasificación incorrecta de los acreditados default. Esto implica que el error total de clasificación no sea un criterio apropiado para medir el rendimiento de un modelo de clasificación de acreditados, aunque desgraciadamente se utiliza, por simplicidad y no con poca frecuencia, para comparar diferentes modelos de clasificación.

El poder discriminante de un modelo de credit scoring puede ser evaluado utilizando una serie de medidas estadísticas absolutas de la discriminación, algunas de las cuales veremos en esta sección.

Por otra parte, por razones análogas a las que conducen a realizar la prueba de bondad de ajuste del modelo no sólo sobre los datos de desarrollo sino también sobre una muestra test es necesario realizar la prueba del poder discriminante también en un conjunto de datos independientes (fuera de la muestra de validación). De lo contrario existe el peligro de que el poder discriminante pueda ser exagerado por el sobreajuste en el conjunto de datos de entrenamiento.

Que el modelo de credit scoring presente un poder discriminante relativamente bajo en un conjunto de datos independiente, aunque estructuralmente similar, del conjunto de datos de entrenamiento, significa que el modelo tiene una baja estabilidad. Un rasgo característico de un modelo estable es que recoge de forma adecuada la relación causal entre los factores de riesgo y el default, lo que evita dependencias espurias derivadas de correlaciones empíricas. En contraste con los modelos estables, los sistemas inestables con

frecuencia muestran una disminución considerable en el nivel de precisión de la estimación con el tiempo.

La medición del poder discriminante es lo suficientemente importante como para que el Grupo de Implementación de los Acuerdos del Comité de Basilea II , (CSCB, 2005a), haya recomendado, como estadísticos más significativos para comparar el poder discriminante de funciones de calificación de acreditados, los mostrados en la tabla 4.2.

Tabla 4.2.- Estadísticos de Discriminación, CSCB, (2005a).

Tasa de precisión, AR, (Acuracy Rate) o Coeficiente de GINI. y su herramienta gráfica : Curva de Ajuste Acumulativo, CAP , (Cumulative Accuracy Profile).
La Curva ROC, (Receiver Operating Characterístics) y su estadístico resumido: *Área Bajo la Curva, AUC (Area Under the Curve), o Coeficiente de Concordancia.*
Tasa de Error Bayesiana, (Bayesian Error Rate).
Entropía Condicional, (Conditional Entropy), Distancia de Kullback-Leibler, Tasa de Entropía de Información Condicional, CIER, (Conditional Information Entropy Rate).
Valor de Información (Divergencia, Índice de Estabilidad)
τ *de Kendall, D de Somers*
La puntuación de Brier (Brier score).

En la práctica, los Entidades Financieras utilizan para medir la discriminación de un modelo, en general, sólo una parte de las medidas anteriores, en concreto la *Tasa de Precisión*, estadístico resumen de la *Curva de Precisión Acumulativa,* CAP, el *Área Bajo la Curva ROC*, AUC, estadístico resumen de l*a Curva ROC, (Receiver Operating Characterístics)* y la *Puntuación de Brier*. El estadístico AUC y la Puntuación de Brier son los dos estadísticos considerados por HARREL (2001). ENGELMANN, (2006). Por su parte, ENGELMANN y RAUHMEIER (2006) y MEDEMA et al. (2009) consideran la *Puntuación de Brier*, la *Tasa de Precisión*, AR, la *Curva de Precisión Acumulativa*, CAP, la Curva ROC y AUC.

CLAVERO (2008) analiza en detalle, con elegante y riguroso formalismo matemático, todos los estadísticos y curvas anteriores a los que añade además el "*Perfil de la diferencia entre las funciones de distribución acumulativas*" de las poblaciones de buenos y malos, y su estadístico asociado, *estadístico de Kolmogorov-Smirnov, K*-S.

El propio Grupo de Implementación de los Acuerdos del Comité de Basilea II, (BCBS, 2005a) ha constatado que el coeficiente de precisión (AR) y el área bajo la curva (AUC)

parecen ser más significativos que los demás índices que hemos mencionado por sus propiedades estadísticas puesto que para ellos es posible calcular intervalos de confianza de una manera sencilla. La anchura del intervalo de confianza dependerá de la cartera de clientes de que se trate así como del número de acreditados default disponibles para llevar a cabo la estimación. Como regla general, cuanto más ancho sea el intervalo de confianza para AR (o AUC) tanto menor será la calidad de la estimación. Además ambos estadísticos tienen en cuenta el tamaño de la muestra. En consecuencia, estos instrumentos de medida reflejan tanto la calidad del sistema de calificación como del tamaño de las muestras en que este sistema se basa.

Tanto AR como AUC están muy bien considerados por los bancos Supervisores por cuanto estos pueden probar de forma fiable si un modelo de calificación sujeto a validación es significativamente diferente de una modelo sin poder discriminante.

Así mismo, el Grupo de Implementación de los Acuerdos de Basilea II también considera que la Puntuación de Brier puede ser útil en el proceso de desarrollo de varios modelos de credit scoring ya que también indica cuál de ellos tiene mayor poder discriminante. Sin embargo, debido a la falta de test estadísticos aplicables a la Puntuación de Brier, la utilidad de este indicador para la validación es limitada.

En esta Tesis Doctoral analizaremos y utilizaremos para la validación del poder discriminante del modelo de credit scoring proactivo finalmente seleccionado el "*Perfil de la diferencia entre las funciones de distribución acumulativas*" de las poblaciones de buenos y malos, y su estadístico asociado, *estadístico de Kolmogorov-Smirnov, K*-S, desde la perspectiva de CLAVERO, (2008). La razón es que el perfil de las funciones de distribuciones de las poblaciones de default y no default es un instrumento gráfico muy potente para visualizar las posibles diferencia entre ambas poblaciones, pero sobre todo por qué se cuenta con un test asociado para contrastar si las diferencias son o no significativas.

4.3.2 Perfil de la Diferencia entre la Distribuciones Acumulativas. Test de Kolmogorov-Smirnov.

El objetivo fundamental de un modelo de credit scoring es discriminar entre acreditados buenos y acreditados malos. Para comprobar la consecución de este objetivo se ha de validar si la distribución de las puntuaciones de los acreditados buenos se diferencia de la distribución de las puntuaciones de los acreditados malos. Tal validación puede fundamentarse en la siguiente idea:

"Cuando la diferencia entre la frecuencias relativas acumuladas de dos muestras aleatorias de datos es muy pequeña, las distribuciones de las dos poblaciones, origen de tales muestras, deben ser similares y recíprocamente, cuando la las distribuciones de las dos poblaciones no son similares, la diferencia entre las frecuencias relativas acumuladas de las muestras debe ser significativa".

Usualmente suele utilizarse el área de solapamiento de las densidades de probabilidad de los acreditados buenos y los acreditados malos para medir la distancia entre sus respectivas puntuaciones. Denotaremos, como es habitual, por $F_0(\bullet)$ y $F_1(\bullet)$ las funciones de distribución acumulada de $S(X)/(Y=0)$, la función de calificación o puntuación condicionada al grupo de los acreditados buenos), y de $S(X)/(Y=1)$, (la función de calificación puntuación condicionada al grupo de los acreditados malos), respectivamente y por $f_0(\bullet)$ y $f_1(\bullet)$ las correspondientes funciones de densidad, es decir

$$\begin{aligned} F_0(s) &= P(S \le s / Y=0) = \int_{-\infty}^{s} f_0(u)\,du \\ F_1(s) &= P(S \le s / Y=1) = \int_{-\infty}^{s} f_1(u)\,du \\ F(s) &= P(S \le s) = \int_{-\infty}^{s} f(u)\,du \end{aligned} \tag{4.3}$$

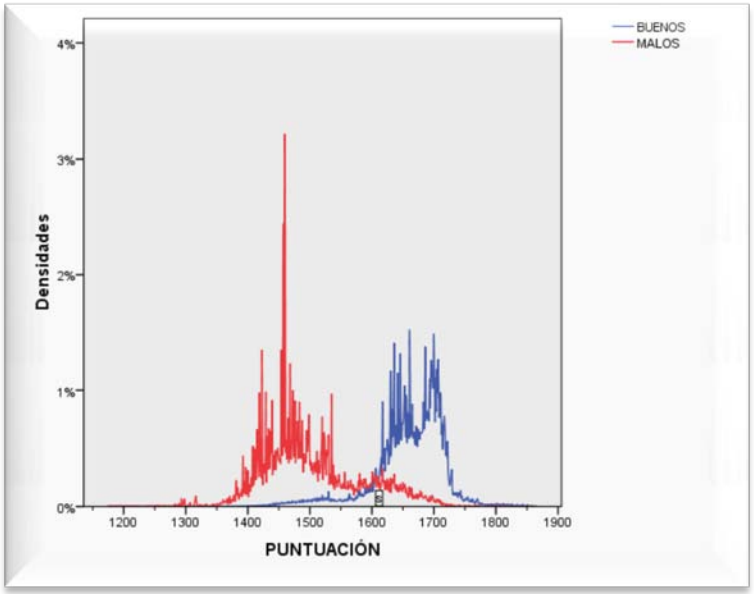


Figura 4.2.- Funciones de densidad de la puntuación de crédito en la poblaciones de default y no default.

Como puede observarse en la figura 4.2, las dos densidades tienen un punto de intersección en $S(X)=s$ que es la frontera de separación que *"permite predecir que todos los solicitantes con puntuación* $S(X)<s$ *serán incumplidores"*. En principio este punto de intersección podría ser el punto de corte seleccionado, si bien no tiene por qué.

Definición 4.5.- *La región de solapamiento* O está constituida por la zona bajo la densidad de buenos, a la izquierda del umbral s, y la zona bajo la densidad de malos, a la derecha de s.

Si no se hace ninguna hipótesis distribucional para $S(\bullet)$, entonces la definición de O puede ser fácilmente generalizada al caso no paramétrico

$$O=\int_{-\infty}^{+\infty}\min\{f_0(s),f_1(s)\}\,ds \tag{4.4}$$

en el cual puede haber cualquier número de puntos de intersección.

Si se asume un solo punto de intersección óptimo, para una relación monótona positiva entre la puntuación $S(\bullet)$ y la probabilidad de incumplimiento, el área de solapamiento está definida por

$$O_{pos}=\min_s\{F_1(s)+1-F_0(s)\} \tag{4.5}$$

Alternativamente, para una influencia monótona negativa de los valores de $S(\bullet)$ sobre Y, se tiene

$$O_{neg}=\min_s\{F_0(s)+1-F_1(s)\} \tag{4.6}$$

Para una relación monótona obviamente se tiene

$$O_{mon}=\min\{O_{pos},O_{neg}\} \tag{4.7}$$

Está claro que para que las distribuciones se separen perfectamente el área de la región de solapamiento O ha de ser cero y si ambas regiones son idénticas entonces el área de O es 1 (puesto que el área bajo la función de densidad es igual a 1)

$$\int_{-\infty}^{+\infty}f_k(u)\,du=1 \tag{4.8}$$

Una *medida del poder discriminante* puede entonces venir dada por

$$T=1-O_{mon}=\max\left|F_0(s)-F_1(s)\right| \tag{4.9}$$

El indicador del poder discriminante T toma valores en el intervalo [0,1], donde $T=1$ indica una separación total y $T=0$ significa que no existe separación alguna.

Las versiones monótona positiva y monótona negativa de T son respectivamente

$$T_{pos} = 1 - O_{pos} = máx\{F_0(s) - F_1(s)\} \tag{4.10}$$

$$T_{neg} = 1 - O_{neg} = máx\{F_1(s) - F_0(s)\} \tag{4.11}$$

en la práctica se tienen observaciones s_i e Y_i para el i-ésimo acreditado o solicitante de crédito, i=1,...,N.

Bajo hipótesis generales sobre la distribución de buenos y malos, los estadísticos O y T pueden calcularse, por ejemplo, por estimadores no paramétricos de la densidad, tales como histogramas, estimadores de la densidad por núcleos, etc.).

En el caso no paramétrico es suficiente contar con estimadores no paramétricos de las distribuciones acumuladas F_0 y F_1. Estos estimadores pueden ser fácilmente calculados como funciones de distribución empírica:

$$\hat{F}_k(s) = \frac{\sum_{i=1}^{n} I_{(s_i \le s, y_i = k)}}{\sum_{i=1}^{n} I_{(y_i = k)}}, \quad k = 0,1 \tag{4.12}$$

En este caso, la distribución del estadístico $T = 1 - O = 1 - mín\{F_1(s) + 1 - F_0(s)\}$ $= máx|F_0(s) - F_1(s)|$ se relaciona con la distribución de Kolmogorov. Por tanto, el test de Kolmogorov-Smirnov, que chequea la hipótesis $F_0 = F_1$, puede aplicarse para contrastar si la variable de puntuación $S(\bullet)$ influye sobre la probabilidad de incumplimiento.

Tabla 4.3.- Tests Estadísticos de Dominancia Estocástica.

Test	H_0	H_1	Test Estadístico	Rechazo
(1)	$F_1(s) = F_0(s)$	$F_1(s) > F_0(s)$	$\hat{T}_{pos} = \underset{s}{máx}\{\hat{F}_0(s) - \hat{F}_1(s)\}$	$T_{pos} > \Delta_{n_1,n_0,1-\alpha}$
(2)	$F_1(s) = F_0(s)$	$F_1(s) < F_0(s)$	$\hat{T}_{neg} = \underset{s}{máx}\{\hat{F}_1(s) - \hat{F}_0(s)\}$	$T_{neg} > \Delta_{n_1,n_0,1-\alpha}$
(3)	$F_1(s) = F_0(s)$	$F_1(s) \neq F_0(s)$	$\hat{T} = \underset{s}{máx}\{\lvert\hat{F}_0(s) - \hat{F}_1(s)\rvert\}$	$T > \Delta_{n_1,n_0,1-\alpha/2}$

Si se consideran las hipótesis de la tabla 4.3 entonces se pueden usar los test estadísticos (1) y (2) para chequear la dominancia estocástica de F_1 sobre F_0 y viceversa. Los valores críticos fueron dados por MILLER (1956), como sigue:

$$\Delta_{n_1,n_0,1-\alpha} = \Delta_{q,1-\alpha} \quad \text{con} \quad q = \frac{n_o\, n_1}{(n_o + n_1)} \tag{4.13}$$

donde n_o y n_1 son el número de default y no default respectivamente. Para n_1 decreciente los valores críticos son crecientes, pero esto no implica que los valores de los test estadísticos crezcan. Lo que significa que dado un valor del test estadístico puede ser más difícil rechazar la hipótesis nula si la proporción de incumplimientos es pequeña.

En la práctica:

1.- Se calcula la frecuencia acumulada empírica tanto para acreditados cumplidores como para incumplidores.

2.- Se obtienen las diferencias entre las frecuencias acumuladas empírica de los dos grupos.

3.- Se busca el valor K-S, que maximiza las diferencias calculadas en el paso anterior.

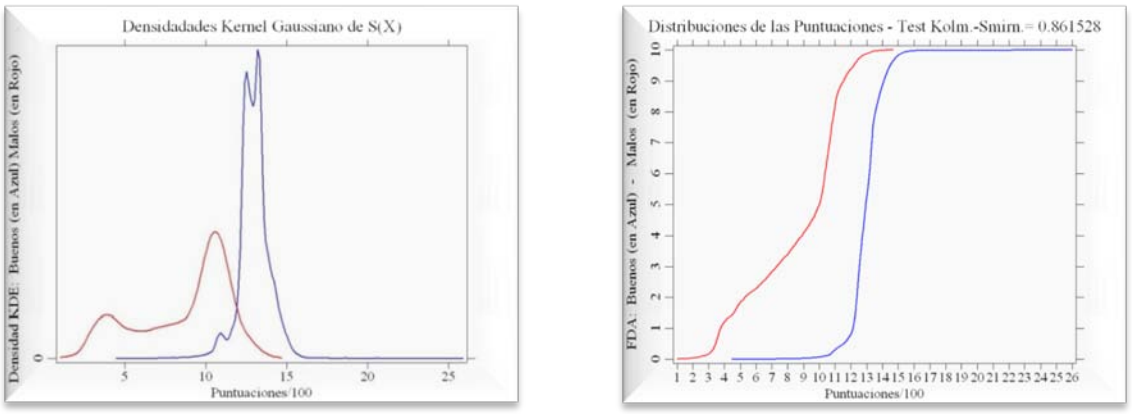


Figura 4.3.- Perfiles de las Funciones de Densidad y Distribución Acumulada de las Puntuaciones. Test de Kolmogorov-Smirnov.

Un valor del estadístico $\hat{T}$ alto significa que las puntuaciones de los dos grupos de acreditados no se distribuyen uniformemente, es decir, el modelo discrimina bien ambos grupos.

4.3.3 Precisión de un Modelo de Credit Scoring, CAP y AR.

Otra medida usada para la precisión de un Sistema de Calificación es la *Tasa de Precisión*, AR, (obtenida a partir del *Coeficiente de Gini*, G), basada en la *Curva de Lorenz. La* Curva de Lorenz, también conocida como *Perfil de Precisión Acumulada,* CAP*, Curva de Potencia o Curva de Selección*, representa gráficamente la distribución de función de calificación $S(\bullet)$ frente a las distribuciones de default $S(\bullet)/Y=1$, lo que posibilita la comparación gráfica de diferentes puntuaciones de crédito.

Para las probabilidades acumuladas, los porcentajes de solicitantes son ordenados desde las puntuaciones más bajas a las más altas. Es deseable que el sistema de calificación acumule un mayor porcentaje de deudores morosos en las puntuaciones más bajas, de este modo una curva de acumulación perfecta debe acumular primero a todos los deudores morosos. Una curva con estas características es la *Curva de Lorenz,* CAP*,* que permite visualizar la concentración de los valores de una variable cuantitativa. Utilizaremos una versión de esta curva en la que se visualiza la concentración de la puntuación de los incumplidores con respecto a la distribución del total de acreditados.

Para obtener la Curva de Lorenz, en primer lugar, se ordenan todos los acreditados por sus respectivas puntuaciones, desde un estado de riesgo hasta un estado de seguridad, es decir desde los acreditados de menor puntuación a los de mayor puntuación. Y en segundo lugar, se construye la curva de Lorenz asignando a cada fracción x del número total de acreditados el porcentaje $d(x)$ de los incumplidores cuyas puntuaciones son menores o iguales que la puntuación máxima de la fracción x, esto dota a x de un rango entre 0% y 100%.

Para construir formalmente la curva CAP, es necesario estimar las siguientes probabilidades:

Probabilidad de incumplimiento con puntuación s $\quad Ps_D = Ns_D / N_D$

Probabilidad de cumplimiento con puntuación s $\quad Ps_{ND} = Ns_{ND} / N_{ND}$

Probabilidad de tener puntuación s $\quad Ps_T = Ns_T / N$

donde

Ns_D : Número de incumplidores que tienen puntuación s.

N_D : Número total de incumplidores.

Ns_{ND} : Número de cumplidores que tienen puntuación s.

N_{ND} : Número total de cumplidores.

Ns_T : Número total de acreditados con puntuación s.

N : Número total de acreditados.

La probabilidad de incumplidores en la muestra es $p_1 = N_D / N$ y la de cumplidores $p_0 = N_{ND} / N = 1 - p_1$. La probabilidad de acreditados con puntuación s viene dada por:

$$Ps_T = p_1 Ps_D + (1 - p_1) Ps_{ND} \tag{4.14}$$

Las probabilidades acumuladas pueden ser estimada para cada puntuación s como:

$F_1(s) = Fs_D = \sum_{k=1}^{s} Ps_D$ (Probabilidad de incumplimientos con puntuación $\leq s$).

$F_0(s) = Fs_{ND} = \sum_{k=1}^{s} Ps_{ND}$ (Probabilidad de cumplimientos con puntuación $\leq s$).

$F(s) = Fs_T = \sum_{k=1}^{s} Ps_T$ (Probabilidad de acreditado con puntuación $\leq s$).

La curva CAP es la representación gráfica de los puntos de coordenadas $(1 - F(s), 1 - F_1(s))$ *para cada puntuación*. Lo que formalmente podemos representar por

$$\begin{aligned} L(s) &= \{L_1(s), L_2(s)\} = \{P(S > s), P(S > s / Y = 1)\}, \quad s \in (-\infty, \infty) \\ &= \{1 - P(S \leq s), 1 - P(S \leq s / Y = 1)\}, \quad s \in (-\infty, \infty) \\ &= \{1 - F(s), 1 - F_1(s)\}, \quad s \in (-\infty, \infty) \end{aligned} \tag{4.15}$$

Por tanto la curva de Lorenz puede representarse por la siguiente expresión

$$L(u) = 1 - F\left[(1 - F_1)^{-1}(u)\right], \quad u \in (0,1) \tag{4.16}$$

y puede estimarse por medio de las funciones empíricas de distribución acumulada.

La cantidad $F(s) = P(S \leq s)$ coincide con la *Tasa de Alarma*, (Alarm Rate), para la puntuación s y $F_1(s) = P(S \leq s / Y = 1)$ con la *Tasa de Aciertos*, (Hit Rate), para la misma puntuación. La cantidad $100 * L(s)$ indica el porcentaje de acreditados incumplidores que se hallan entre los $100 * s$ primeros acreditados (de acuerdo a sus puntuaciones). Además existe una relación directa entre las probabilidades de incumplimiento condicionales dadas las puntuaciones (vía la derivada de la curva):

$$\frac{dL(s)}{ds} = \frac{P\left[Y=1/s=F^{-1}(u)\right]}{p} \tag{4.17}$$

Una fortaleza de la curva CAP es el rápido crecimiento de $L(s)$ para s próximo a 0 y una debilidad el también rápido crecimiento de $L(s)$ para s próximo a 1, *"a más discrepancia entre las densidades condicionales mejor poder discriminante de la función de puntuación subyacente"* .

La *Curva de Lorenz Optimal* se corresponde con la puntuación que separa perfectamente a los acreditados cumplidores de los no cumplidores. Esta curva tiene la peculiaridad de que la ordenada $(1-F_1)(s)=1$ se alcanza para la abscisa $(1-F)(s)=P(Y=1)$, es decir, en la probabilidad de default, y está dada por

$$Lopt(s) = \left\{1-F(s),\ g\left(1-F_1(s)\right)\right\},\ s\in(-\infty,+\infty) \tag{4.18}$$

siendo

$$g(x)=\begin{cases}\dfrac{x}{P(Y=1)} & 0<x\le P(Y=1)\\ 1 & P(Y=1)<x\le 1\end{cases} \tag{4.19}$$

La concavidad de la curva CAP es equivalente a la propiedad de que las probabilidades de default moldean la función de calificación como una función decreciente de las calificaciones. Además, la no concavidad indica que no se hace un uso óptimo de la información en la especificación de la función de calificación de acreditados.

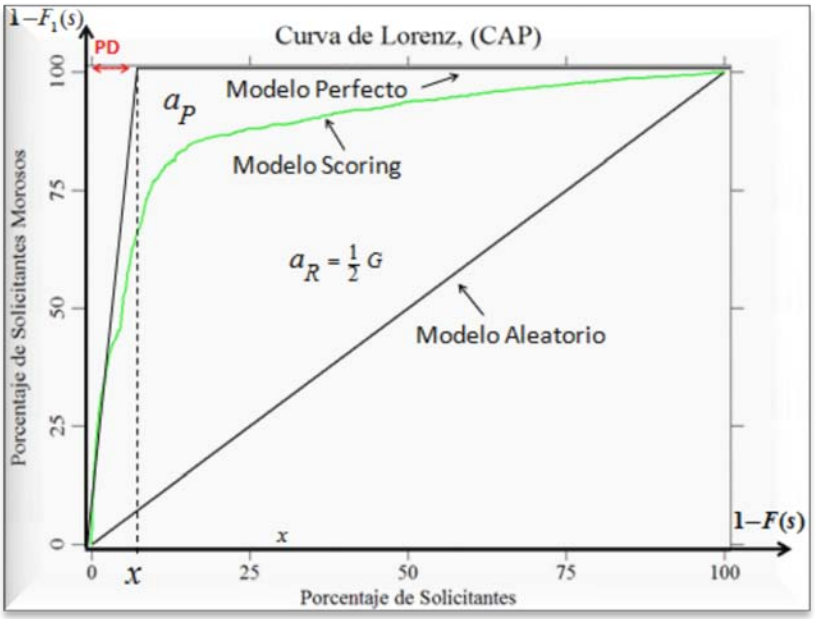


Figura 4.4.- Curva de Lorenz, CAP. Modelo perfecto. Modelo Aleatorio.

Una curva CAP idéntica a la diagonal corresponde a una puntuación que ordena a los solicitantes de crédito de forma totalmente aleatoria. El modelo aleatorio no tiene ningún poder discriminante, es decir, en este modelo cualquier fracción de deudores con puntuaciones bajas contendrá la misma fracción de deudores morosos. El modelo de calificación real está entre los dos extremos. Por tanto, la curva también puede usarse para comparar diferentes distribuciones de puntuación: las mejores distribuciones de puntuación están muy cercanas a la curva CAP y las peores se acercan más a la diagonal. Por tanto una comparación visual de las CAP de distintas carteras de crédito puede ser interesante. La forma de la CAP depende de la proporción de acreditados default y no default en la muestra.

El área de la región comprendida entre la curva CAP y la diagonal es una buena medida de la eficacia de la puntuación. Uno de los coeficientes más populares para medir tal eficacia es el *Coeficiente de Gini* , G, que consiste en dos veces esa área, es decir

$$G = 2\int_{+\infty}^{-\infty} L(S)\ ds = 2\int_{+\infty}^{-\infty} \left[(1-F_1)(s)\ d(1-F)(s)\right] - 1 = 1 - 2\int_{-\infty}^{+\infty} F_1(s)\ dF(s) \qquad (4.20)$$

En la práctica la última integral se estima por integración numérica de $\hat{F}_1$ sobre el rango de $\hat{F}$. Esta integral puede aproximarse matemáticamente por la ecuación:

$$\begin{aligned} G &= \sum_{k=1}^{S-1}\left(\widehat{(1-F)}(s_{k+1}) - \widehat{(1-F)}(s_k)\right)\left(\widehat{(1-F_1)}(s_{k+1}) + \widehat{(1-F_1)}(s_k)\right) 1 \\ &= 1 - \sum_{k=1}^{S-1}\left(\hat{F}(s_{k+1}) - \hat{F}(s_k)\right)\left(\hat{F}_1(s_{k+1}) + \hat{F}_1(s_k)\right) \end{aligned} \qquad (4.21)$$

Proposición 4.1.- *Para la Curva de Lorenz Optimal, se tiene que el coeficiente de Gini optimal está dado por*

$$G_{opt} = P(Y=0) = 1 - P(Y=1) \qquad (4.22)$$

Este resultado se debe al hecho de que el coeficiente optimal de Gini es dos veces el área del triangulo entre la curva optimal de Lorenz y la diagonal, lo que es lo mismo que calcular el área de un paralelogramo, en este caso, la base es $P(Y=0)$ y la altura $1 - P(Y=0) + P(Y=1)$. El coeficiente de Gini es en esencia una varianza usada para cuantificar la diferencia entre dos puntos.

Una medida cuantitativa muy popular, sobre la precisión de un modelo de credit scoring, alternativa al coeficiente de Gini, es la Tasa de Precisión, AR, Acuracy Ratio, también

basada en la curva CAP, (KEENAN y SOBEHART (1999) , ENGELMANN et al. (2003a, 2003b)).

La tasa de precisión viene dada por la relación del coeficiente de Gini de cada puntuación y el coeficiente de Gini de la curva de Lorenz optimal,

$$AR = \frac{G}{G_{opt}} = \frac{G}{P(Y=0)} = \frac{2\int_{+\infty}^{-\infty} L(s)\, d\, s1}{1 - P(Y=1)} = \frac{2\int_{+\infty}^{-\infty} \left[(1-F_1 L)(s)\, d(1-F)(s)\right] - 1}{1 - P(Y=1)} \tag{4.23}$$

es decir, se obtiene a partir de los estimadores de ambos coeficientes de Gini.

Si adoptamos la siguiente notación:

a_P : Área entre la curva CAP del modelo perfecto y la curva CAP del modelo aleatorio.

a_R : Área entre la curva CAP del modelo validado y la curva CAP del modelo aleatorio.

Se tiene la siguiente expresión para el Coeficiente de Gini

$$G = 2a_R \tag{4.24}$$

Por otro lado, el coeficiente de Gini Óptimo se puede expresar como

$$G_{opt} = 2a_P \tag{4.25}$$

Por tanto,

$$AR = \frac{G}{G_{opt}} = \frac{2a_R}{2a_P} = \frac{a_R}{a_P} \tag{4.26}$$

El valor de AR se sitúa entre 0 y 1 si la curva CAP es realmente cóncava, es decir, si existe una relación monótona positiva entre $S(X)$ e Y (altos valores de puntuación se corresponden con bajos valores de la probabilidad de default). El modelo de calificación óptimo es aquel para el que más se aproxime AR a 1, el peor es aquel para el que más se aproxime a 0, es decir, al modelo aleatorio, esto hace que AR sea una medida idónea para comparar diferentes puntuaciones.

Maximizar AR es equivalente a maximizar el poder discriminante en el sentido de maximizar el área bajo la CAP vía $\frac{d}{du}CAP(u) = P\left[Y=1/s=F^{-1}(u)\right]/p$, resultando altas probabilidades de default para pequeñas puntuaciones y bajas probabilidades de default para grandes puntuaciones. AR se puede estimar como:

$$A_R = 1 - 2\sum_{k=1}^{K-1} Ps_D\left(F_1(s_k) - F_1(s_{k+1})\right) = P\left[S_D < S_N\right] - P\left[S_D > S_N\right] \tag{4.27}$$

donde S_N y S_D son independientes y distribuidas de acuerdo con F_N y F_D, respectivamente.

La calidad del sistema de calificación viene dada por la tasa de precisión de sus modelos de credit scoring.

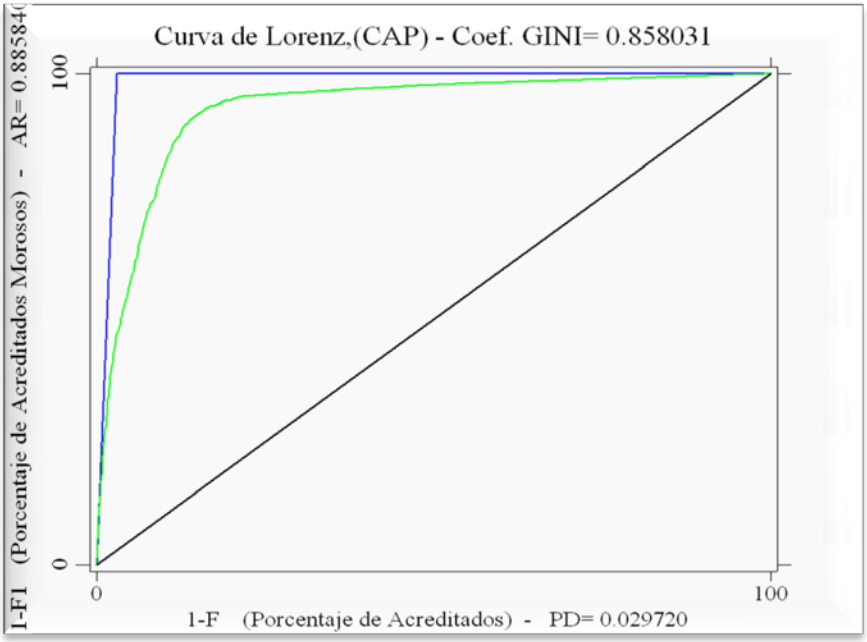
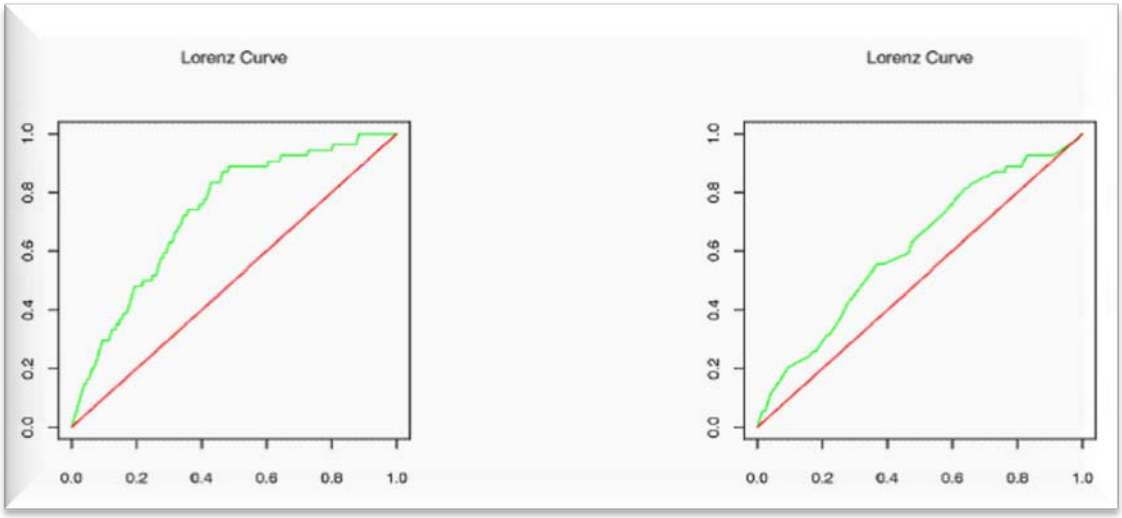


Figura 4.5.- Curva de Lorenz, Coeficiente de Gini y Tasa de Precisión, AR.

La experiencia práctica muestra que la Tasa de Precisión toma valores entre el 50% y el 80%, (TASCHE, BCBS, 2005a), véase figura 4.6



Fuente: Clavero, (2008).

Figura 4.6.- Curva de Lorenz para dos valores distintos de AR.

4. 3.4 Curva ROC y área bajo la curva ROC, AUC.

Una segunda curva que permite visualizar el poder discriminante de un modelo de calificación es la curva de *Características Operativas del Receptor,* ROC. Esta curva de aspecto muy similar a la curva CAP presenta frente a esta una importante diferencia, en el eje horizontal se sitúa $1-F_0(s)$, mientras en la primera se sitúa $1-F(s)$, siendo $F(s)$ la función de distribución acumulada del total de acreditados o solicitantes de crédito, se supone en ambos casos que los porcentajes de acreditados están ordenados desde puntuaciones malas a buenas. La Curva ROC viene dada, por tanto, por

$$1-F_0(s) \quad \textit{frente a} \quad 1-F_1(s),\ \forall\ s\in(-\infty,+\infty) \tag{4.28}$$

donde $F_0(s)=P(S\le s/Y=0)$ es la tasa de fallos para la puntuación s, es decir la curva ROC puede representarse formalmente por la expresión

$$ROC(u) = 1-F_0\left[(1-F_1)^{-1}(u)\right], \quad u\in(0,1) \tag{4.29}$$

y, por tanto, consiste en el conjunto de todos los pares

$$R(s)=\left\{1-F_0(s),\ 1-F_1(s)\right\} \tag{4.30}$$

El gran parecido entre la curva ROC y la Curva CAP se debe a que el número de incumplimientos es usualmente pequeño lo que implica que $F_0(s)\approx F(s)$.

Teóricamente se puede tener un continuo de pares $\{1-F_0(s),\ 1-F_1(s)\}$ pero en la práctica se tiene sólo una muestra finita de puntos sobre la curva ROC, estimados por medio de las funciones empíricas de distribución acumulada. Entonces la curva ROC completa se obtiene por interpolación lineal de este conjunto de puntos ó bien a través de la estimación no paramétrica de las funciones de densidad, por ejemplo por estimación de la densidad por núcleos, y a partir de estas obtener $\left\{1-\hat{F}_0(s),1-\hat{F}_1(s)\right\}$.

La curva ROC es ligeramente más compleja que la CAP, pero a cambio no requiere la composición muestral para reflejar la verdadera proporción de default y no default.

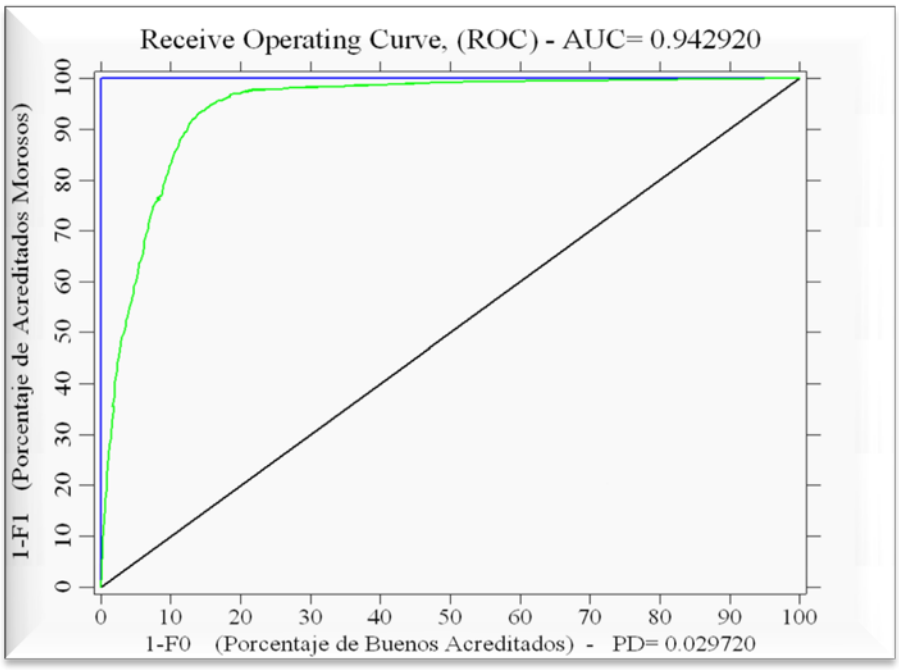


Figura 4.7.- Curva ROC y Área Bajo la Curva, AUC.

Los puntos $(0,0)$ *y* $(1,1)$ están contenidos en toda curva ROC, a causa de que para $s < mín\{Puntuaciones\}$ se tiene qué $F_0(s) = F_1(s) = 0$ y para $s > máx\{Puntuaciones\}$ se tiene que $F_0(s) = F_1(s) = 1$. La curva ROC óptima corresponde a la puntuación que separa con exactitud los cumplidores de los no cumplidores y viene determinada por los puntos $(0,0)$,$(0,1)$ *y* $(1,1)$.

Al igual que para CAP, la concavidad de la curva ROC es equivalente a la propiedad de que las probabilidades de default dan la forma de la función de calificación como una función decreciente de las calificaciones. Además, la no concavidad indica que no se hace un uso óptimo de la información en la especificación de la función de calificación de acreditados.

La segunda medida del poder discriminante, cada día más popular, es el *Área Bajo la Curva ROC*, AUC, a la que se le suele denominar también *Coeficiente de Concordancia* (C). Si nos fijamos en el área bajo la curva ROC óptima determinada por los puntos $(0,0)$,$(0,1)$ *y* $(1,1)$, figura 4.7, en la figura los puntos de referencia vienen representados por $(100,100)$,$(0,100)$ *y* $(100,100)$, puesto que tanto abscisas como ordenadas viene dadas en porcentajes, vemos que vale 1, es decir AUC vale 1 para el "modelo perfecto". Un modelo donde cumplidores e incumplidores se confundieran totalmente, "modelo

aleatorio", tendrá un área igual a 0.5. En una situación real y para un sistema de calificación razonable el "área bajo la curva" será una cantidad entre 0.5 y 1.

El área bajo la curva se obtiene según la siguiente expresión:

$$AUC = \int_{+\infty}^{-\infty} \left[\left(1 - F_1(s)\right) \; d\left(1 - F_0\right)(s) \right] \tag{4.31}$$

y toma valores entre 0 y 1, 0 para la menor desviación y 1 para la mayor, si bien un AUC por debajo de 0.5 no tiene significado.

El AUC es una transformación lineal de la *Tasa de Precisión* y puede interpretarse como *la habilidad media del modelo de credit scoring para clasificar exactamente acreditados o solicitantes de crédito buenos y malos*.

Cuando el valor de AUC es de 0,5 la curva ROC es igual a la diagonal lo que significa que el modelo hace predicciones al azar. Un valor de AUC igual a 1 indica que la curva ROC se encuentra en la esquina superior izquierda y las predicciones son perfectas. Una curva ROC cercana a la diagonal indica que el modelo es poco informativo. Cuanto más cercana se encuentre la curva de ROC a la esquina superior izquierda, mejor es el poder discriminante del modelo entre cumplidores e incumplidores. O, dicho de otra manera, *cuanto mayor sea el área bajo la curva ROC, mejor será el modelo.*

Entre las medidas del poder discriminante existe una relación lineal muy interesante, (CLAVERO, 2008), que viene dada por la expresión

$$AR = 2AUC - 1 \tag{4.32}$$

Como consecuencia, el uso de cualquiera de las dos medidas para comparar funciones de calificación de acreditados conducirá a las mismas conclusiones.

Las propiedades estadísticas de AUC son bien conocidas y coinciden con las del estadístico de Mann-Witney, en particular, existen potentes test para comparar el AUC de un modelo scoring con el de un modelo aleatorio y para comparar dos o más modelos entre sí.

Test U de Wilcoxon-Mann-Witney.

Dos de los test no paramétricos más clásicos para contrastar si dos distribuciones son o no idénticas son el test de la suma de rangos de Wilcoxon y su equivalente, el test U de Mann-Witney.

A continuación se deduce el test U para funciones de calificación de acreditados continuas en su forma más simple:

Si denotamos por s_{j0} todas las puntuaciones observadas de no default y por s_{i1} todas las puntuaciones observadas de default, el estadístico del test U viene dado por

$$\hat{U} = \#\left\{s_{i1} > s_{j0}\right\}, \text{ sobre todo } i, j. \tag{4.33}$$

Para una separación perfecta entre acreditados default y no default, se obtiene $\hat{U} = N_0 N_1$. Si S e Y no están totalmente relacionadas, entonces el suceso $s_{i1} > s_{j0}$ ocurre con probabilidad $1/2$, de forma que $U \approx \dfrac{1}{N_0 N_1}$. En consecuencia, una versión reescalada del estadístico $\hat{U}$, $\tilde{U} = \dfrac{\hat{U}}{N_0 N_1}$, es un estimador para el área bajo la curva que se obtiene según la siguiente expresión:

$$U = P\left\{(S/Y=1) > (S/Y=0)\right\} = \int_{+\infty}^{-\infty} \left[(1-F_1(s))\ d(1-F_0)(s)\right] = AUC \tag{4.34}$$

Y, por lo tanto, usando (4.32),

$$\tilde{U} = \left(\frac{\widehat{AR}+1}{2}\right) N_0 N_1 \tag{4.35}$$

La relación entre $\tilde{U}$ y $\widehat{AR}$ seguirá siendo válida aunque las distribuciones de las calificaciones no sean continuas. Sin embargo, puede ocurrir que para cualquier valor de calificación se observen tanto el default como el no default, es decir puede haber casos ligados, de ser así deberá de utilizarse una versión corregida del estadístico (añadir $1/2$ si $s_{i1} = s_{j0}$) para estimar

$$P\left\{(S/Y=0) > (S/Y=1)\right\} + P\left\{(S/Y=0) = (S/Y=1)\right\}$$

Se demuestra, (LEHMANN, 1975) que bajo la hipótesis $F_1(s) = F_0(s)$, para N_0, N_1 grandes U se distribuye aproximadamente normal. Si se consideran las hipótesis:

Tabla 4.4.- Test U de Wilcoxon-Mann-Witney.

Test	H_0	H_1	Test Estadístico	Rechazo
(1)	$F_1(s)=F_0(s)$	$F_1(s)>F_0(s)$	U	$U > k_{N_1,N_0,1-\alpha}$
(2)	$F_1(s)=F_0(s)$	$F_1(s)<F_0(s)$	U	$U < N_1N_0 - k_{N_1,N_0,1-\alpha}$

Siendo el valor crítico

$$k_{N_1,N_0,1-\alpha} = \frac{N_0N_1}{2} + U_{1-\alpha}\sqrt{\frac{1}{12}N_0N_1(N_0+N_1+1)} \tag{4.36}$$

El valor crítico y el test estadístico decrecen cuando decrece el número de acreditados default N_1. Por tanto, no se puede decir que sea más difícil rechazar la hipótesis nula para bajas tasas de default.

La medida AUC no depende de la probabilidad de default total de la cartera de crédito, de hecho puede ser estimada sobre muestras con proporciones de default y no default no representativas. Similarmente, pueden compararse representaciones de carteras de crédito bancario con diferentes proporciones de incumplidores.

No es posible definir para AUC un valor mínimo general, con significación estadística, en orden a decidir si el modelo de calificación tiene bastante poder discriminante. Como requerimiento mínimo del modelo de calificación, pueden servir el test Mann-Whitney o el de Kolmogorov-Smirnov, con un nivel de significación, por ejemplo, del 5%, para rechazar significativamente la hipótesis nula (el modelo no tiene más poder discriminante que el modelo aleatorio). Para la mayoría de los modelos de credit scoring usados en la industria bancaria los p-valores no se distinguen de cero y, como consecuencia, el uso de los p-valores como indicadores de la calidad del modelo parece ser limitado.

En la práctica, las curvas ROC no sólo se utilizan para determinar la discriminación, sino también para determinar un punto de corte en la función de calificación para la concesión de préstamos (BLÖCHLINGER Y LEIPPOLD, 2006; STEIN, 2005).

4.3.5 Puntuación de Brier.

La puntuación de Brier, (BRIER, 1950), es un estadístico para la validación de la calidad de las probabilidades pronosticadas por el modelo, con origen en el ámbito de las previsiones meteorológicas, que consiste en un estimador del promedio de la diferencia

entre el cuadrado de la probabilidad pronosticada por el modelo y el valor del estado de default observado

$$B = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i^2 - y_i\right)^2 \tag{4.37}$$

donde $\hat{p}_i$ es la probabilidad pronosticada e y_i el estado de default para el acreditado i. B se puede interpretar como la media de la suma de los cuadrados de los residuos de una regresión no lineal del estado de default sobre la probabilidad pronosticada. Es evidente que toma valores entre 0 y 1, y además un valor cercano a 0 indica que el modelo funciona bien. Como consecuencia, reducir al mínimo la puntuación de Brier es equivalente a maximizar el varianza de las probabilidades de default pronosticadas.

Resultados empíricos indican que la minimización de la puntuación de Brier conlleva la maximización del área bajo la curva ROC, en este sentido, la puntuación de Brier es una medida de poder discriminante que posee la debilidad de que hasta la fecha no existen test estadísticos efectivos que lo respalden como estadístico de decisión.

Otra desventaja de la puntuación de Brier es su escaso rendimiento para pequeñas probabilidades de incumplimiento. En este caso los pronósticos triviales, aquellas en las que a todos acreditados default se les asigna la tasa de default en la muestra, suelen tener una puntuación de Brier bastante pequeña. En este caso el valor esperado de la puntuación de Brier viene dado por $\overline{B} = (1-p)\,p$, donde p es la frecuencia de default de la muestra. Obviamente se verifica que $\lim_{p\to 0}\overline{B} = 0$.

La *puntuación de Brier* también puede ser utiliza para determinar el poder discriminante de un sistema de calificación del crédito con grados $g = 1,...,G$, (ENGELMANN y RAUHMEIER, 2006). En este caso la puntuación de Brier se define como sigue:

$$B = \frac{1}{N}\sum_{g=1}^{G}\left[n_{1g}\left(1-p_g\right)^2 + n_{0g}\left(p_g\right)^2\right] \tag{4.38}$$

donde p_g, $g = 1,...,G$, es la probabilidad de default pronosticada para la clase g del sistema de calificación, n_{1g} y n_{0g} representan, respectivamente, el número de acreditados default y no default de la clase g.

4.4 CALIBRACION DE LA PROBABILIDAD DE DEFAULT.

4.4.1 Introducción.

Las categorías de calificación de un modelo de credit scoring se construyen normalmente sobre la base de las probabilidades de default referidas a un horizonte temporal de un año. En la práctica, las probabilidades de default pronosticadas difieren de las tasas de impago finalmente observadas. El problema surge cuando estas desviaciones no se producen al azar, sino sistemáticamente, lo que indica que seguramente el modelo de credit scoring no sea el adecuado. La cuestión que abordaremos aquí es "*cómo la probabilidad de default pronosticada por el modelo de credit scoring al comienzo del horizonte temporal puede ser revisada dadas las tasas de default realmente observadas al final de dicho período*".

Cuando la probabilidad de default pronosticada por un modelo de credit scoring se desvía solo marginalmente de la que se observa se dice que el modelo está bien calibrado, es decir, l*a calibración es la capacidad del modelo para realizar estimaciones no sesgadas (objetivas) de las probabilidades de default.*

La calibración es un concepto que se originó en meteorología, donde se utilizan los modelos de probabilidad para las previsiones meteorológicas. En este contexto se originó la definición general siguiente (SEIDENFELD (1985)):

Definición. 4.6.- *Un conjunto de probabilidades son calibradas (o están bien calibradas),* si el p por ciento de todas las predicciones informadas en probabilidad p son verdaderas. Esta definición general puede ser aplicada en el marco de las probabilidades de default. Tradicionalmente, el ajuste de modelos binarios se analiza a través de una tabla de clasificación 2×2, donde las columnas son los dos valores pronosticados de la variable respuesta y las filas son los dos valores observados de dicha variable.

Los valores pronosticados se determinan mediante un punto de corte, (cut_off), basado en la probabilidad, que teóricamente suele ser 0,5. De este modo el valor previsto de la variable dependiente es igual a 1 si la probabilidad pronosticada está por encima de 0,5 y 0 en caso contrario. El modelo es perfecto si todos los casos están en la diagonal de la tabla de clasificación. Una tabla de clasificación da el porcentaje de predicciones correctas. En el caso de default los conjuntos de datos son muy desequilibrados en el sentido de que sólo una pequeña fracción incumple las obligaciones de pago contractuales, por ejemplo, el suceso default sólo ocurre el 2% de las veces. En general, cuando se usa una tabla de clasificación para determinar la bondad del ajuste se concluye que siempre será preferible

un modelo que conduzca a una probabilidad de default constante igual a cero. En el caso del riesgo de crédito, esta probabilidad de default cero es inútil para el cálculo de los requerimientos de capital. En otras palabras, en el marco de la determinación de los requerimientos de capital, una tabla de clasificación no es una herramienta de calibración útil, por lo que habremos de considerar alternativas adecuadas.

Dado que las tasas de default están sujetas a fluctuaciones estadísticas, es necesario desarrollar test estadísticos para contrastar la significación de las desviaciones de la frecuencia de default observada con las estimaciones (a través del modelo correspondiente) de la probabilidad de default dada la puntuación crediticia. Es decir, test estadísticos hacia atrás, (backtesting), para contrastar la hipótesis

$$\begin{aligned} \boldsymbol{H_0}: &\text{ La probabilidad de default pronosticada en un nivel de calificación es correcta} \\ \boldsymbol{H_1}: &\text{ La probabilidad de default pronosticada en un nivel de calificación es incorrecta} \end{aligned} \tag{4.39}$$

La cuestión que se trata de resolver con los test estadísticos para la calibración de sistemas de calificación es la siguiente:

¿Cómo de exactas son las estimaciones de la probabilidad de default condicionada a una puntuación de crédito dada ?. (4.40)

El grado de discrepancia entre la probabilidad de default pronosticada y la realmente observada puede indicar problemas potenciales y acciones que necesitan ser acometidas.

Para formalizar los test a los que conduce la cuestión (4.40), consideramos un sistema de credit scoring con N acreditados que se clasifican en R categorías de calificación diferentes de acuerdo con sus calificaciones crédito. Si N_r indica el número de acreditados que son clasificados en la clase de calificación $r \in \{1,...,R\}$, entonces se tiene $N = \sum_{r=1}^{R} N_r$ e indicando por

$0 < \hat{p}_r < 1$: *Probabilidad de default pronosticada* por el sistema de calificación para la categoría r.

$0 < p_r < 1$: *Probabilidad de default real*, (desconocida), para la categoría r.

$0 < p_r^{obs} < 1$: Tasa de default observada, para r, es decir, la proporción de acreditados en default de un total de N_r acreditados en la categoría de calificación r.

podemos diferenciar entre formulaciones de test de una cara y de dos caras.

Los *test de una cara* se caracterizan a través de la hipótesis:

$$\begin{aligned} H_0 &: \ P_1 = \hat{P}_1,..., P_R = \hat{P}_R \\ H_1 &: \ \exists r \in \{1,...,R\} \text{ con } P_r > \hat{P}_r \end{aligned} \tag{4.41}$$

lo que *está conforme con la percepción del Banco Supervisor*, al que concierne que el riesgo no sea infra estimado .

Los *test de dos caras* se caracterizan a través de la hipótesis:

$$\begin{aligned} H_0 &: \ P_1 = \hat{P}_1,..., P_R = \hat{P}_R \\ H_1 &: \ \exists r \in \{1,...,R\} \text{ con } P_r \neq \hat{P}_r \end{aligned} \tag{4.42}$$

lo que *está conforme con la percepción del Controlador del Riesgo* que está interesado en una estimación tan exacta como sea posible.

El test estadístico puede entonces usarse, para un cierto nivel de significación pre-especificado con su correspondiente p-valor, para tomar una decisión con ese nivel de significación. Altos p-valores indican que el test es significativo y, por tanto, no se rechaza la hipótesis nula de que la probabilidad de default es al nivel p significativamente infraestimada. Elegir el apropiado nivel de significación depende de la política conservadora que en mayor o menor grado adopte la Entidad Financiera.

Un aspecto importante a tener en cuenta en estos test es si en las estimaciones de la probabilidad de default se tiene o no en cuenta el estado de la economía, por ejemplo, por la inclusión de variables explicativas macroeconómicas. Desde este punto de vista si se tiene en cuenta el estado de la economía las estimaciones condicionadas a este estado se llaman estimaciones en un *punto del tiempo*, (PIT, *Point-In-Time*), y en caso de no considerar este estado se denominan estimaciones a *través del ciclo*, (TTC, *Through-The-Cycle*), que, lógicamente, son incondicionales y se calculan con datos de un ciclo económico completo.

En el caso de estimaciones en un punto del tiempo, dada una realización actual de las covariables puede ser adecuada la suposición de independencia de los sucesos de crédito, mientras que en el caso de estimaciones incondicionales no se puede hacer ningún supuesto de independencia. De acuerdo con este punto de vista, los métodos más utilizados para validar la probabilidad de default pueden agruparse en dos categorías:

1) Tests condicionados al estado de la economía, estimaciones en un *punto del tiempo*, (PIT):

   Test Binomial (y *aproximación normal*)

   Test de Hosmer-Lemeshow $\left(\chi^2\right)$

   Test de Spiegelhalter

2) Tests no condicionados al estado de la economía, estimaciones a *través del ciclo*, (TTC):

   Test Normal

   Semáforos (Extended Traffic Lights)

Bajo la hipótesis de sucesos de default estocásticamente independientes, las dos formulaciones de los test (4.41) y (4.42) están relacionadas con problemas estándar que pueden ser abordados por el *Test Binomial,* (ENGELMANN y RAUHMEIER, 2006) *,* por el *Test Chicuadrado,* (HOSMER y LEMESHOW, 2000), y el *Test de Spiegelhalter* , (SPIEGELHALTER, 1986), como veremos en esta sección. El principal problema es que los impagos de crédito no son estocásticamente independientes; como alternativa, se asumirá la estructura de dependencia de la IRB (Aproximación Basada en Rating Internos). Para esta estructura de dependencia, la tasa de default no converge a la probabilidad de default asociada, sino a una distribución de probabilidad no-degenerada en el intervalo [0, 1]. Con el fin de evaluar la calidad de las estimaciones de las probabilidades de default que varían en el tiempo, se consideran dos enfoques: el *Test Normal*, (BLOCHWITZ et al. 2006) *y* el enfoque de los *Semáforos de Tráfico*, (TASCHE, 2003, y BLOCHWITZ et al. 2004) .

Mientras el *test Binomial* sólo se puede aplicar a un simple grado de calificación sobre un único período de tiempo, el *test de Hosmer-Lemeshow* $\left(\chi^2\right)$ y el de *Spiegelhalter* proporcionan métodos más avanzados que pueden usarse para contrastar la adecuación de la predicción de la probabilidad de default sobre un único período de tiempo para varias categorías de calificación de acreditados.

4.4.2 Test basados en la hipótesis de independencia de los sucesos de default.

La construcción de contrastes de hipótesis en el supuesto de sucesos de default independientes se basa en dos hechos bien conocidos:

1. Bajo la hipótesis de independencia de los sucesos de default, el número de default observados en una categoría de calificación r con N_r acreditados, ($N_r p_r^{obs}$), y probabilidad de default p_r, se distribuye según una variable aleatoria Binomial de parámetros N_r y p_r:

$$N_r p_r^{obs} \sim Bin(N_r, p_r) \tag{4.43}$$

2. Para la tasa de default observada p_r^{obs} y $N_r \to \infty$ se verifica la ley fuerte de los grandes números,

$$p_r^{obs} \xrightarrow{c.s.} p_r \tag{4.44}$$

y el teorema central del límite

$$\sqrt{N_r}\,\frac{P_r^{obs} - P_r}{\sqrt{P_r(1-P_r)}} \xrightarrow{D} N(0,1) \tag{4.45}$$

3. Para R categorías de calificación se verifica

$$\sum_{r=1}^{R} N_r \frac{\left(P_r^{obs} - P_r\right)^2}{\sqrt{P_r(1-P_r)}} \xrightarrow{D} \chi^2(R) \tag{4.46}$$

donde $\chi^2(R)$ es la distribución chi-cuadrado con R grados de libertad.

4.4.2.1 Test Binomial.

Usualmente, el primer paso en la calibración de un modelo de probabilidad de default es a menudo realizar el test Binomial (ENGELMANN y RAUHMEIER (2006)). El test Binomial está diseñado para contrastar los pronósticos de la probabilidad de default estimada por el modelo, $\hat{P}$, frente a la tasa de default observada, P^{obs}, para una categoría de calificación r dada, usando el siguiente test:

$$\begin{aligned} &H_0: \ P_r = \hat{P}_r \quad \text{La PD coincide con la probabilidad de default estimada en la clase r.} \\ &H_1: \ P_r > \hat{P}_r \quad \text{La PD es infraestimada en la clase r.} \end{aligned} \tag{4.47}$$

Si se asume que los default ocurren independientemente**,** dada una categoría de calificación r con probabilidad de default P_r, si N_r y p_r^{obs} son, respectivamente, el número de acreditados en la categoría y la proporción de default observada en la misma, entonces $N_{1r} = N_r P_r^{obs}$, número de default en la categoría r, sigue una distribución Binomial de parámetros N_r y P_r y, por tanto,

$$\Pr(N_{1r}) = \binom{N_r}{N_{1r}} (P_r)^{N_{1r}} \left(1\text{-}P_r\right)^{N_r - N_{1r}} \tag{4.48}$$

Para cada categoría $r=1,...,R$ se busca contrastar si la probabilidad de default pronosticada para una categoría de calificación es correcta frente a la alternativa de que está infraestimada, es decir contrastar la hipótesis nula H_0 , $H_0: P_r = \hat{P}_r$, de que la verdadera probabilidad de default P_r es igual a $\hat{P}_r$, frente a la alternativa de una cara $H_1: \ P_r > \hat{P}_r$.

La hipótesis nula para un nivel de significación α se rechaza si el número de defaults $N_{1r} = N_r p_r^{obs}$ es mayor que un valor crítico k^* dado por:

$$k^* = \min\left\{k / \sum_{i=k}^{N_r} \binom{N_r}{i} \hat{p}_r (1-\hat{p}_r) \le \alpha\right\} \tag{4.49}$$

siendo N_r el número total de acreditados de la categoría r. Dado que para grandes valores de N_r el cálculo de k^*, (4.49), es muy costoso, se suele usar el hecho de que la distribución Binomial converge a la distribución normal cuando el número de pruebas crece

$$\mathrm{P}_r^{obs} \sim N\left(\hat{\mathrm{P}}_r \ , \ \frac{\hat{\mathrm{P}}_r\left(1-\hat{\mathrm{P}}_r\right)}{N_r}\right) \tag{4.50}$$

o, equivalentemente,

$$z = \frac{P_r^{obs} - \hat{P}_r}{\sqrt{\frac{\hat{P}_r\left(1-\hat{P}_r\right)}{N_r}}} \sim N(0,1) \tag{4.51}$$

Se puede comparar el valor calculado z, el cual se distribuye según la normal estandarizada, frente al cut-off, basado sobre el nivel de confianza deseado, y tomar la decisión sobre aceptar o rechazar H_0. O bien considerar que de acuerdo con (4.51), el valor crítico $k_{\alpha-1}$ puede ser aproximado como sigue:

$$k^* \approx \Phi^{-1}(1-\alpha)\sqrt{N_r \hat{p}_r (1-\hat{p}_r)} + N_r \hat{p}_r \tag{4.52}$$

siendo $\Phi^{-1}(\bullet)$ la inversa de la función de distribución de la variable aleatoria normal (0,1). Pero es preferible expresarlo en términos de tasa de default:

Se rechaza la hipótesis nula si la tasa de default observada $\hat{p}_r^{obs}$ *es mayor que* $p_{1-\alpha}$*:*

$$p_{1-\alpha} \approx \Phi^{-1}(1-\alpha)\left(\frac{\hat{p}_r(1-\hat{p}_r)}{N_r}\right)^{1/2} + \hat{p}_r \tag{4.53}$$

Esta aproximación a la $N(0,1)$ se aplica cuando $N_r > 1000$ y de no ser así se aproxima la Binomial a la distribución de Poisson, que está bastante experimentada en el caso de carteras de bajo default. Para un número bajo de observaciones se calcula efectivamente la distribución Binomial.

Para el test de dos caras:

$$\begin{aligned} &H_0: \ \mathrm{P}_r = \hat{\mathrm{P}}_r \quad \text{La PD coincide con la probabilidad de default estimada en la clase r.} \\ &H_1: \ \mathrm{P}_r \neq \hat{\mathrm{P}}_r \end{aligned} \tag{4.54}$$

siendo también $N_{1r} = N_r P_r^{obs}$ el estadístico de contraste, se tiene que la región crítica para p_r^{obs} y un nivel de significación asintótica α está dado por:

$$[0, p_{\alpha/2}) \cup (p_{1-\alpha/2}, 1] \tag{4.55}$$

Por tanto se rechazará la hipótesis nula (*para la clase de calificación r, la tasa de default observada coincide con la probabilidad de default pronosticada por el modelo*) si p_r^{obs} queda fuera del intervalo (4.55).

En ambas formulaciones del test, puede ser más difícil rechazar la hipótesis nula para un pequeño número de acreditados en la clase r, puesto que $p_{1-\alpha}$ crece cuando N_r decrece.

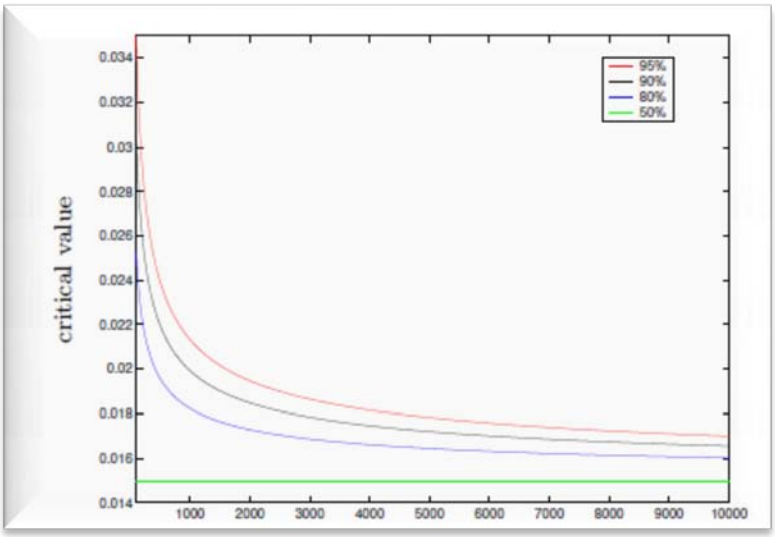


Figura 4.8.- *Valores críticos del test Binomial para una probabilidad de default de referencia de 0,015 asumiendo varios niveles de significación, (en el eje de ordenadas), y varios valores de N, (en el eje de abscisas).*

Como puede verse en la figura 4.8 los valores críticos decrecen cuando crece el número de observaciones, lo cual representa una motivación para ejecutar el Backtesting sobre un nivel agregado, (CASTERMANS et al. 2010).

4.4.2.2 Test Chi_cuadrado.

La prueba binomial (o su mencionada extensión normal) es principalmente adecuada para contrastar un único nivel de calificación, pero no de varios o todas las categorías de calificación de forma simultánea. El Chi_cuadrado o de Hosmer-Lemeshow es en esencia un test de conjunto para varios grados de calificación que se originó en el campo de la regresión categórica y es habitual referirse a él como *test de bondad de ajuste*, (HOSMER y LEMESHOW (2000)).

Ahora se busca contrastar si las probabilidades de default son correctas para todas las categorías de calificación de acreditados simultáneamente, es decir, contrastar:

$$\begin{aligned} &H_0: \ P_1 = \hat{P}_1, ..., P_R = \hat{P}_R \\ &H_1: \ \exists r \in \{1, ..., R\} \ \text{con} \ P_r \neq \hat{P}_r \end{aligned} \tag{4.56}$$

Aquí asumiremos las siguientes hipótesis:

a) Las probabilidades de default pronosticadas por el modelo, $\hat{p}_r$, y las tasas de default observadas, p_r^{obs}, son idénticamente distribuidas.

b) Todos los sucesos de default tanto dentro de cada categoría de calificación como entre las categorías son independientes.

El *test estadístico chi-cuadrado* se deduce del muy conocido *estadístico chi-cuadrado de Pearson* original, (véase D'ÁGOSTINO y STEPHEN, 1986), y viene dado por:

$$t_R = \sum_{r=1}^{R} N_r \frac{\left(p_r^{obs} - \hat{p}_r\right)^2}{\hat{p}_r\left(1-\hat{p}_r\right)} \tag{4.57}$$

Bajo las hipótesis a) y b), cuando $N_r \to \infty$ simultáneamente para todo $r = 1, ..., R$, por el *Teorema Central del Límite*, la distribución de t_R converge en distribución a una distribución χ^2 con R grados de libertad:

$$t_R = \sum_{r=1}^{R} N_r \frac{\left(P_r^{obs} - P_r\right)^2}{\sqrt{P_r\left(1-P_r\right)}} \xrightarrow{D} \chi^2\left(R\right) \tag{4.58}$$

Por tanto, se rechazará la hipótesis nula para un nivel de significación asintótico α, si t_R es mayor que el $(1-\alpha)\_cuantil$ de una distribución χ^2 con R grados de libertad.

El $p-valor$ del test χ^2 es una medida para validar la adecuación de las probabilidades estimadas, cuanto más se acerca el $p-valor$ a cero peor es la estimación. Sin embargo, si las probabilidades de incumplimiento estimadas son muy pequeñas, la tasa de la convergencia a la distribución- χ^2 puede ser muy baja también. Por otra parte, los $p-valores$ proporcionan una posible forma de comparar directamente los pronósticos con diferentes números de categorías de calificación.

Debemos tener en cuenta que puesto que la prueba de Hosmer-Lemeshow se basa en las hipótesis de independencia y de aproximación Normal, este test posiblemente subestime el verdadero error tipo I , (BLOCHWITZ et al., (2006). Para un pequeño número de acreditados en cada clase de calificación, la hipótesis nula es más difícil de rechazar.

4.4.2.3 Test de Spiegelhalter.

Normalmente se calculan individualmente las probabilidades de default pronosticadas por el modelo para cada acreditado. Puesto que el test Chi_Cuadrado de Hosmer-Lemeshow, al igual que el test Binomial, requieren que todos los acreditados asignados a una categoría de calificación tengan la misma probabilidad de default, lo que requiere promediar las probabilidades de default pronosticadas de los acreditados que han sido clasificados en la misma categoría de calificación, en los cálculos se pueden introducir algunos sesgos. Se puede evitar este problema usando el test de Spiegelhalter, SPIEGELHALTER, (1986**) y** BCBS, (2005c), que permite las variaciones en probabilidades de default dentro de la misma categoría de calificación.

Si consideramos N acreditados en el sistema de credit scoring y para el acreditado i, $i=1,...,N$, el modelo ha asignado una puntuación s_i y una estimación de la probabilidad de default $\hat{p}_i$ al comienzo de un periodo, observándose al final del período el default $(y_i=1)$ o el no default $(y_i=0)$ para cada acreditado, el *Error Cuadrático Medio*, MSE, o Puntuación de Brier en el contexto de validación, (BRIER, 1950), viene dado por

$$MSE=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{p}_i\right)^2 \tag{4.59}$$

El estadístico MSE constituye el punto de partida del *Test de Spiegelhalter*.

Dado que a los incumplidores les son asignados altos pronósticos de la probabilidad de default y bajos a los cumplidores, se deduce que el estadístico SME obtenido es pequeño; en general, un bajo MSE indica un buen modelo de credit scoring.

Se asume que los sucesos default son independientes tanto dentro de cada categoría de calificación diferente como entre todas las categorías de calificación.

MSE es una medida de exactitud, lo que significa que es una herramienta de pronóstico que trata de estimar la probabilidad de default para cada acreditado, $p_i = \hat{p}_i$. Puesto que las probabilidades $p_i = E[Y / X = x_i]$ son desconocidas, se reemplazan por sus realizaciones observables, p_i^{obs}, $i \in \{1,...,N\}$. MSE cuantifica la desviación de las estimaciones y las observaciones, por lo que puede interpretarse como una función de penalización ponderada, en la cual grandes discrepancias entre observaciones y estimaciones son ponderadas por su severidad a través de la función de pérdida cuadrática. Cuanto más alta sea la exactitud de la estimación, más pequeño es MSE. Dado que

$$y_i \sim Ber(p_i) = Bin(1, p_i),\ i = 1,...,N \tag{4.60}$$

con probabilidad de default real desconocida p_i, es decir, MSE es una variable aleatoria. El estimador minimiza el MSE esperado si se da un adecuado pronóstico de la probabilidad de default.

La hipótesis nula para el contraste es que "*todas las probabilidades de default estimadas,* $\hat{p}_i$*, coinciden exactamente con la verdadera, aunque desconocida, probabilidad de default* $P(y_i = 1 / X = x_i)$ *para todo acreditado* i"

$$H_0:\ \hat{p}_i = P(y_i = 1 / X = x_i) = E[Y / X = x_i],\quad i = 1,...,N \tag{4.61}$$

Entonces bajo la hipótesis nula el MSE tiene un valor esperado de

$$\begin{aligned} E\left[MSE_{\hat{p}_i = p_i}\right] &= E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2 / _{\hat{p}_i = p_i}\right] \\ &= E\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2\right] = \frac{1}{N}\sum_{i=1}^{N}E\left[\left([Y / X = x_i] - p_i\right)^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{E\left[(Y / X = x_i)^2\right] - 2E[Y / X = x_i]p_i + p_i^2\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{p_i - 2p_i^2 + p_i^2\right\} = \frac{1}{N}\sum_{i=1}^{N}p_i(1 - p_i) \end{aligned} \tag{4.62}$$

y

$$V\left[MSE\right]=\frac{1}{N^2}\sum_{i=1}^{N}p_i\left(1-p_i\right)\left(1-2p_i\right)^2 \tag{4.63}$$

Asumiendo la hipótesis de independencia y usando el teorema central del límite, se puede demostrar, (SPIEGELHALTER, 1986), que bajo la hipótesis nula el test estadístico

$$Z \;=\; \frac{MSE-E\left[MSE\right]}{\left(V\left[MSE\right]\right)^{1/2}} \;=\; \frac{\sum_{i=1}^{N}\left\{\left(y_i-\hat{p}_i\right)^2-\hat{p}_i\left(1-p_i\right)\right\}}{\left(\sum_{i=1}^{N}p_i\left(1-p_i\right)\left(1-2p_i\right)^2\right)^{1/2}} \tag{4.64}$$

sigue aproximadamente una distribución normal estándar que permite un test de decisión estándar (véase RAUHMEIER y SCHEULE (2005) para ejemplos prácticos). Puesto que las probabilidades $p_i=E\left[Y/X=x_i\right]$ son desconocidas se reemplazan por sus realizaciones estimadas, $\hat{p}_i$, $i\in\left\{1,...,N\right\}$. Por lo que el estadístico (4.64) queda

$$Z \;=\; \frac{MSE-E\left[MSE\right]}{\left(V\left[MSE\right]\right)^{1/2}} \;=\; \frac{\sum_{i=1}^{N}\left\{\left(y_i-\hat{p}_i\right)^2-\hat{p}_i\left(1-\hat{p}_i\right)\right\}}{\left(\sum_{i=1}^{N}\hat{p}_i\left(1-\hat{p}_i\right)\left(1-2\hat{p}_i\right)^2\right)^{1/2}} \tag{4.65}$$

CAPÍTULO 5

FUNCIONES DE BASE EN MODELOS DE CREDIT SCORING.

5.1 INTRODUCCIÓN.

Una fase esencial en la construcción de los modelos de credit scoring la constituye la especificación del modelo, fase de cuya importancia nos hemos ocupado ya en el capítulo 2, sección 2.7. Nuestro interés radica sobre todo en los *Modelos de Probabidad Generalizados*, GPM, que se caracterizan por que su estructura formal se basa en una función de enlace, $g(\bullet)$, entre la probabilidad, variable respuesta o explicada, y las variables explicativas X, $g(P(A)) = S(X)$, donde A es un suceso perteneciente a un álgebra de sucesos apropiada, y $g(\bullet)$ es una función con las propiedades necesarias para asegurar que $P(\bullet) = g^{-1}(S(X))$ es una función de probabilidad. Para el caso de la probabilidad de default el modelo viene dado por $g(P(Y = 1/\ X = x)) = S(X)$.

HASTIE y TIBSHIRANI (1996), sugirieron que una forma natural de generalizar los Modelos Logísticos Lineales consiste en reemplazar las variables explicativas por una versión no paramétrica de las mismas, superficies arbitrarias de regresión tales como las funciones sierra, superficies más estructuradas de alta dimensión como las funciones bisagra obtenidas por el procedimiento MARS de FRIEDMAN (1991), o modelos paramétricos de mayor complejidad como los modelos aditivos por *expansiones de funciones de base*.

La sugerencia de Hastie y Tibshirani es muy interesante porque, tal como veremos a lo largo de este capítulo, las distintas técnicas que con mayor o menor acierto se utilizan para construir *modelos de credit scoring,* a excepción de los estimadores no paramétricos de la probabilidad de default, *k-vecinos más próximos*, K_NN, y el *estimador de Nadaraya_Watson*, NWE, pueden abordarse desde la óptica de encontrar la estructura formal más adecuada para el modelo desde distintas estrategias respecto de construir *la expansión lineal de funciones de base, lo que nos proporciona una visión unificadora de estas técnicas,* tarea que acometemos en este capítulo.

Además, la idea central de esta Tesis Doctoral, que desarrollaremos en detalle en el capítulo 6, consiste en añadir la no linealidad a los modelos utilizando *expansiones lineales de las variables explicativas* que tienen como idea básica, sugerida por HASTIE y TIBSHIRANI, (1996), *aumentar o remplazar el vector de variables explicativas X con variables adicionales, funciones de base, las cuales son transformaciones de X, y entonces usar modelos lineales en este nuevo espacio.* Esta idea se puede formalizar según la definición siguiente:

Definición 5.1.- Denotando por $h_r(X):\mathbb{R}^p \to \mathbb{R}$ la r-ésima transformación de X, llamada r-ésima función de base de X, $r=0,...,q$,

a).- La expresión $\sum_{r=0}^{q}\beta_r h_r(X)=\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})$, donde $\boldsymbol{X}=\left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta}=\left(\beta_0,\beta_1,...,\beta_q\right)^T$ vector de q coeficientes desconocidos y $\boldsymbol{H}(\boldsymbol{X})=\left(h_0(X),h_1(X),...,h_q(X)\right)^T$, representa una *expansión por funciones de base del vector de variables explicativas* X. Si $\beta_0=0$ y $\left(\beta_1,...,\beta_q\right)=(1,...,1)$ entonces tenemos una expansión aditiva por funciones de base, y si $\beta_0\neq 0$ y $\left(\beta_1,...,\beta_q\right)\neq(1,...,1)$ tenemos una expansión es lineal.

b).- Diremos que un modelo general de probabilidad de default para el que la función de calificación de acreditados, $S(x)$, se representa como una expansión de las *funciones de base*, $h_r(\bullet)$

$$g\left(P\left(Y=1/\ X=x\right)\right)=S(X)=\sum_{r=0}^{q}\beta_r h_r(X)=\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X}) \qquad (5.1)$$

es un *Modelo de Probabilidad de Default Generalizado, GDPM*.

El concepto de expansión lineal del vector de variables explicativas por funciones base nos permite construir modelos más flexibles aumentando o reeplazando el vector de variables explicativas originales con variables adicionales, transformaciones de X. *Estos espacios "agrandados" de las expansiones por funciones de base de las variables explicativas son generalmente espacios de Hilbert.* El mejor acontecimiento en la especificación del modelo es encontrar las adecuadas funciones de base $\left(h_0(X),\ h_1(X),...,\ h_q(X)\right)^T$.

Suponer que la relación existente entre una conveniente transformación de la probabilidad de default y las variables de riesgo de crédito explicativas del estado de default es una expansión base lineal del vector de entradas X es sin duda alguna la forma más elegante, desde el punto de vista estadístico, de introducir la no linealidad en los modelos que relacionan el estado de default con las variables explicativas. Estimar $S(x)$ es ahora equivalente a encontrar los coeficientes β_r. Las funciones $h_r(\bullet)$, que en algún caso contiene parámetros desconocidos, son elegidas por poseer propiedades convenientes.

Algunas veces el problema de añadir la no linealidad, o, en general, *aumentar o reemplazar el vector de variables explicativas* X *con variables adicionales, transformaciones de X,* se resuelve con funciones de base h_r sencillas, tales como funciones potencias (5.2), (5.3), logaritmos, raíces cuadradas, interacciones de variables, (5.4), etc.

Si se utilizan las siguientes funciones de base

$$h_r(X) = X_r, \ r = 1,..., p \tag{5.2}$$

la expansión recupera el modelo lineal original. Los modelos logísticos lineales utilizan funciones de base de este tipo.

Por otro lado, funciones de base como las siguientes

$$h_r(X) = X_j^2, \ h_r(X) = X_j X_k, ..., \ r = 1,...,q, \ j,k = 1,..., p \tag{5.3}$$

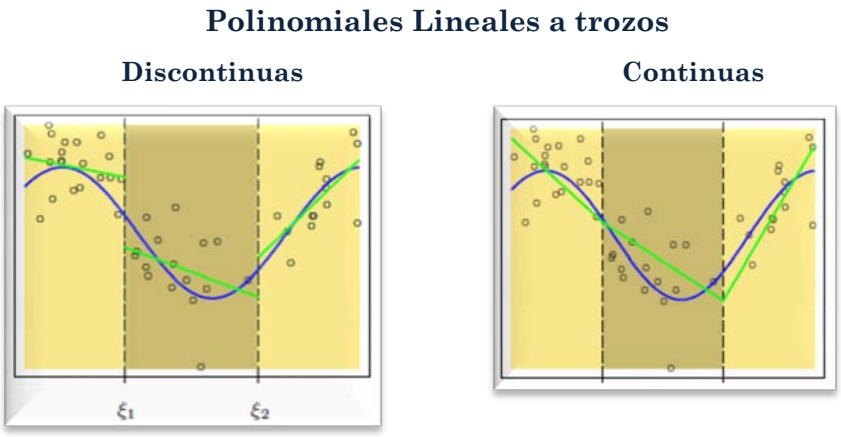
permiten aumentar las entradas con términos polinomiales para alcanzar expansiones de Taylor de altos órdenes. Nótese, sin embargo, que el número de variables crece exponencialmente en el grado del polinomio. Un modelo cuadrático completo en p variables requiere $O\left(p^2\right)$ términos cuadráticos y de productos cruzados, en general, $O\left(p^d\right)$ para un polinomio de grado p.

Por su parte, las funciones de base

$$h_r(X) = \log\left(X_j\right), \ \sqrt{X_j}, ..., \ r = 1,...,q, \ j = 1,..., p \tag{5.4}$$

proporcionan transformaciones no lineales simples a partir de las variables explicativas de entrada. De forma más general pueden usarse funciones con varias entradas, tales como $h_r(X) = \|X\|$.

En realidad lo más frecuente es el uso de expansiones básicas más complejas que las anteriores para alcanzar representaciones más flexibles para $S(X)$, sobre todo cuando se trata de especificar variables no lineales en el modelo. Tal es el caso de las *polinomiales* cuya representación se muestra en la figura 5.1.



Fuente: Hastie et al. (2009).

Figura 5.1.- Polinomiales Lineales a trozos

Pero las polinomiales presentan el problema de estar limitadas por su naturaleza global, "al retorcer los coeficientes para alcanzar una forma funcional en una región puede ocurrir que la función se agite furiosamente en regiones remotas", HASTIE et al. (2009), por lo que, como alternativa, se suelen utilizar *splines de regresión o suavizado*, *funciones sierra*, funciones *bisagra MARS*, *pesos de la evidencia* asignados a particiones recursivas o particiones estadísticas automáticas óptimas, *funciones radiales Gaussianas*, núcleos *definidos positivos*, *funciones de dimensión infinita* como (5.5) y (5.6) y otras.

$$h_r(X) = m(X), \ r = 1,...,q \tag{5.5}$$

$$h_r(X) = m_j\left(X_j\right) = \sum_{r=1}^{q_j} \beta_{jr} h_{jr}\left(X_j\right), \ j = 1,...,p \ , r = 1,...,q_j \tag{5.6}$$

donde $m(\bullet)$ y $m_r(\bullet)$ son funciones no paramétricas de dimensión infinita.

Las polinomiales fragmentadas y los splines que permiten representaciones locales de polinomiales, figura 5.2, y otras funciones de base, producen un diccionario *D* que consiste típicamente de un gran número de funciones de base que son susceptibles de ser ajustadas a nuestros datos.

Lógicamente acompañando al diccionario se requiere un método para controlar la complejidad del modelo que una técnica automática de este tipo pudiese proporcionar.

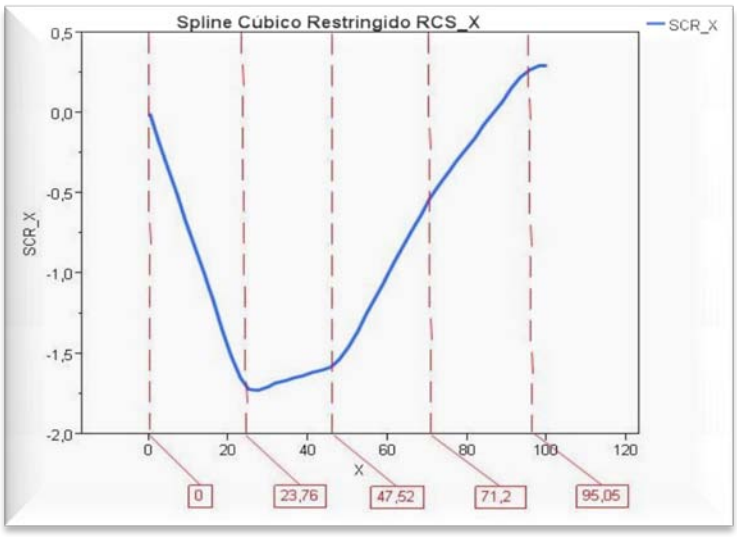


Figura 5.2. - Spline Cúbico Restringido, SCR_X, con 5 nudos en 0, 23.76, 47.52, 71.28 y 95.05.

Existen tres aproximaciones a este tipo de métodos:

Métodos de restricción, donde se decide a priori limitar la clase de funciones, la aditividad es un ejemplo donde se asume que nuestro modelo tiene la forma

$$S(X)=\sum_{j=1}^{p} m_j(X_j)=\sum_{j=1}^{p}\sum_{r=1}^{q_j}\beta_{jr}h_{jr}(X_j) \tag{5.7}$$

El tamaño del modelo está limitado por el número de funciones de base q_r usadas para cada función componente m_j.

Métodos de selección, los cuales escudriñan adaptativamente el diccionario e incluyen solo aquellas funciones de base h_r que contribuyen significativamente a ajustar el modelo. A esta categoría pertenecen técnicas tan importantes como CART, MARS, etc.

Métodos de regularización, donde se usa el diccionario entero y se restringen los coeficientes.

En el capítulo 6 se analiza de forma detallada un conjunto de funciones de base notables, algunas de las cuales se encuentran entre las más utilizadas actualmente en la construcción de modelos de credit scoring.

La estructura formal del modelo (5.1) quedará perfectamente fijada una vez que se fije la función de enlace, $g(\bullet)$, y la expansión de funciones de base, $\sum_{r=0}^{q}\beta_r h_r(X)$.

La selección de ambos elementos depende principalmente del grado de conocimiento que se posea sobre la relación de dependencia del estado de default con las variables explicativas del riesgo de crédito. Nuestro conocimiento obviamente se sitúa entre el *total desconocimiento* y el *conocimiento total*, situación que analizaremos a continuación a través de las técnicas más relevantes que se han venido utilizado en credit scoring desde la perspectiva de la relación de dependencia entre el estado de default y las variables explicativas de riesgo de crédito.

5.2 MODELOS BASADOS EN EL DESCONOCIMIENTO TOTAL DE $S(X)$.

Se supone que no se conoce la distribución poblacional ni ninguna estructura formal que refleje la relación de dependencia entre la variable estado de default y las variables explicativas del riesgo de crédito. Frente a este desconocimiento, los métodos que se recogen en la literatura sobre los modelos de credit scoring vienen navegando entre dos corrientes, por un lado los pertenecientes a la corriente que supone que simplemente se ha de actuar con la filosofía de "*dejar hablar a los datos*", encontrándonos así ante el modelo de probabilidad no paramétrico puro, $P(Y=1/X=x)=S(X)$ y los pertenecientes a la corriente que considera que, aceptado el principio de dejar hablar a los datos, no se ha de obviar que la estructura formal del modelo, por consideraciones sólidamente fundadas, adopta la forma

$$\text{logit}\left(P\left(Y=1/X=x\right)\right)=\text{logit}\left(p\right)+\log\left(LR\left(X\right)\right) \qquad (5.8)$$

independientemente de que se conozcan o no las verosimilitudes de las variables explicativas del default con respecto a las poblaciones de default y no default.

En ambos casos la forma usual de proceder consiste en captar las peculiaridades locales de los datos en vecindades especificadas por métricas apropiadas para estimar directamente en ellas la probabilidad de default en el primer caso, o las funciones de verosimilitud de las poblaciones de default y no default, $f_1(x)$ y $f_0(x)$ respectivamente, en el segundo caso.

En el capítulo 2, sección 2.5, hemos analizado la estimación directa de la probabilidad de default bajo el desconocimiento total de la distribución poblacional y utilizando la herramienta teórica básica reflejada en las expresiones (2.11) y (2.13). Como resultado de ese análisis concluimos que el *estimador de los k vecinos más próximos,* $k-NN$, es muy poco adecuado para estimar la probabilidad de default puesto que, por un lado, es

muy rugoso, seguramente el menos suave de los existentes, lo que conduce al sobreajuste y, por tanto, a un estimador muy poco generalizable, y, por otro, no cumple, entre otros, el requisito básico exigido por los acuerdos de Basilea II que consiste en que la probabilidad de default venga explicada por la contribución de cada variable de riesgo de crédito a la función de calificación de acreditados.

Por lo que respecta a la *estimación de las funciones de densidad condicional al default y no default por funciones núcleo*, este es un método que sin duda sería muy eficaz de no ser por la *maldición de dimensionalidad* que se cierne sobre él, practicamente inaplicable para más de 5 variables. *La solución alternativa a esta técnica, basada en la hipótesis ingénua de Bayes que supone la independencia de las variables explicativas condicionadas al estado de default, es absolutamente inaceptable en los sistemas de calificación de riesgo de crédito.*

5.3 MODELOS BASADOS EN EL DESCONOCIMIENTO CASI TOTAL DE $S(X)$.

Si disponemos de una muestra $\tau = \left\{(x_i, y_i \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1}^{N} \subset \left(X \times Y\right)^N$, de la que sólo sabemos que ha sido obtenida por muestreo aleatorio simple de algún espacio de variables aleatorias de riesgo de crédito sobre $\mathbb{R}^p$. El problema de estimar la función de probabilidad en base a ese conocimiento es irresoluble puesto que pueden existir infinitas soluciones, salvo que afrontemos el ajuste de forma local, con los importantes problemas que ello conlleva. Para estimar globalmente la probabilidad de default es necesario más conocimiento a priori que el que hasta ahora hemos supuesto. Podemos "aumentar" nuestro conocimiento a priori como resultado de asumir la hipótesis de que *la función de calificación de acreditados es "suave", en el sentido de que para valores similares de las variables explicativas corresponden estadísticamente respuestas similares.*

De asumir la hipótesis anterior contaríamos , aparte de con los datos de la muestra τ y con el conocimiento de que el nexo de unión entre la probabilidad de default y las variables explicativas es la transformación logística, con información a priori de suavizado. La información a priori sobre $S(X)$, puede ser, por ejemplo, para el caso unidimensional, que $S(X)$ es una función de "suavizado" continua, dos veces diferenciable, que puede estimarse a través del problema de optimización siguiente

$$\min_{S} \left\{ \sum_{i=1}^{N} \ell\left(y_i, S(x_i)\right) + \frac{1}{2}\lambda \int_{-\infty}^{+\infty} S''(t)^2 dt \right\} \tag{5.9}$$

El teorema de GREEN y SILVERMAN (1994) establece que la solución al problema logístico regularizado (5.9) con función de pérdida logística – espacio de dimensión infinita- es única y es un spline cúbico natural de dimensión finita con N nodos en los "valores únicos" de $\{x_i,\ i=1,...,N\}$, es decir, es una expansión lineal de N funciones de base $h_r(x)$, $S(X)=\sum_{r=1}^{N}\beta_r h_r(X)$.

La función de calificación de acreditados adopta la forma $S(X)=\sum_{r=1}^{N}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por:

$$h_r(X)=\begin{cases} |X-x_r|^3 & r=1,...,N \\ 1 & r=N+1 \\ X & r=N+2 \end{cases} \tag{5.10}$$

es decir el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\sum_{r=1}^{N}\beta_r |X-x_r|^3+\beta_{N+1}+\beta_{N+2}X \tag{5.11}$$

Con el fin de que el spline tenga segunda derivada cero fuera del intervalo $[x_1, x_N]$, se imponen a los coeficientes las restricciones "naturales" $\sum_{r=1}^{N}\beta_r=0$ y $\sum_{r=1}^{N}\beta_r x_r=0$.

Cuando se utilizan splines de suavizado seguramente el modelo resulta sobreparametrizado por cuanto existen N nudos, lo que implica N grados de libertad. *No obstante, el término de regularización traslada la penalización a los coeficientes de los splines, lo cual orienta el problema hacia el ajuste lineal.*

El modelo (5.11), spline de suavizado, se ajusta resolviendo el *problema de minimización de la pérdida empírica regularizada* (3.55), utilizando la pérdida logística, (3.14), y el funcional de regularización $\frac{1}{2}\int_{-\infty}^{+\infty} S''(t)^2 dt$, método que es conocido como ***Regresión Logística Regularizada por Splines Cúbicos,*** **CSRLR**.

5.4 MODELOS BASADOS EN EL CONOCIMIENTO CASI TOTAL DE $S(X)$.

Dos populares situaciones de conocimiento casi total de la distribución de las variables explicativas condicionadas al estado de default y de no default vienen dadas por el ***Análisis Discriminante Lineal, LDA***, ejemplo 2.1, y por el ***Análisis Discriminante Cuadrático, QDA***. En concreto:

Las distribuciones del vector aleatorio de variables explicativas X condicionadas al estado de default y no default son normales de distintas medias desconocidas e igual matriz de covarianza desconocida, $X/(Y=1)\sim N(\mu_1,\Sigma)$, $X/(Y=0)\sim N(\mu_0,\Sigma)$, $\mu_1\neq\mu_0$, en el caso lineal, y matrices de covarianzas distintas, también desconocidas, $X/(Y=1)\sim N(\mu_1,\Sigma_1)$, $X/(Y=0)\sim N(\mu_0,\Sigma_0)$, $\mu_1\neq\mu_0$, $\Sigma_1\neq\Sigma_0$, en el caso cuadrático.

En el caso lineal: $f_{X/Y=k}(x)=\dfrac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-1/2(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\}$, $k=0,1$,

La función de calificación de acreditados adopta la forma $S(X)=\sum_{r=0}^{1}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base se definen en la forma siguiente:

$$h_r(X)=\begin{cases}1 & r=0\\ X & r=1\end{cases} \qquad (5.12)$$

Además se conoce la formulación de los coeficientes:

$$\beta_0=\text{logit}(p)-\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) \text{ y } \beta_1=\left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T$$

es decir, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/X=x)\right)=\text{logit}(p)-\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)+\left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T X \qquad (5.13)$$

La función de calificación de acreditados $S(X)$, parte derecha de la igualdad (5.13), coincide con la *función discriminante de Fisher*, (FISHER (1936), LADD (1966), LACHENBRUCH, (1975)). Como se observa en (5.13), en LDA el logit de la probabilidad de default se relaciona con X a través de un modelo lineal que se puede escribir como expansión lineal de funciones de base en la forma (5.1); estamos ante un modelo paramétrico, de cuyos parámetros (p,μ_1,μ_0,Σ) se conoce la

formulación y tan sólo de han estimar sus valores a través de *estimadores de máxima verosimilitud* *calculados a partir de la muestra de entrenamiento*: $\hat{p}=\frac{1}{N}\sum_{i=1}^{N} I_{\{y_i=1\}}$

$\hat{\mu}_k=\sum_{i=1}^{N_k}\frac{x_i}{N_k}$, $\hat{\Sigma}=\sum_{i=1}^{N}\sum_{y_i=k}\frac{(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T}{N-k}$ con $k=0,1$.

<u>En el caso cuadrático</u>: $f_{X/Y=k}(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-1/2(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right\}$, k=0,1,

por lo que

$$\text{logit}\left(P(Y=1/X=x)\right)=X^TAX+B^TX+\left\{Logit(p)+C\right\},\ \forall X\in\mathbb{R}^p \qquad (5.14)$$

donde

$$A=\frac{1}{2}\left(\Sigma_0^{-1}-\Sigma_1^{-1}\right),\ B=\left(\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0\right),\ C=\frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1\right)-\frac{1}{2}log\frac{|\Sigma_1|}{|\Sigma_0|}$$

es decir,

La función de calificación de acreditados adopta la forma $S(X)=\sum_{r=0}^{p+p^2}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por:

$$h_r(\bullet):\mathbb{R}^p\rightarrow\mathbb{R}$$

$$X\ \rightarrow\ h_r(X)=\begin{cases}1 & si\ r=0\\ X_r & si\ r=1,\ldots,p\\ X_rX_l & si\ r=p+1,\ldots,p^2\end{cases}\qquad \forall X\in\mathbb{R}^p \qquad (5.15)$$

Además se conoce la formulación de los coeficientes:

$\beta_0=logit(p)+\frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1)-\frac{1}{2}Log\frac{|\Sigma_1|}{|\Sigma_0|}$, $(\beta_1,\ldots,\beta_p)^T=(B_j)^T_{j=1,\ldots,p}=\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0$ y

$(\beta_{p+1},\ldots,\beta_{p^2})^T=(A_{jl})^T_{\substack{j=1,\ldots,p\\ l=1,\ldots,p}}=\frac{1}{2}\left(\Sigma_0^{-1}-\Sigma_1^{-1}\right)$

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/X=x)\right)=\sum_{r=0}^{p+p^2}\beta_r h_r(X)=\underbrace{\beta_0+\sum_{r=1}^{p}\beta_rX_r}_{Parte\ Lineal}+\underbrace{\sum_{\substack{r=p+1\\ j=1,\ldots,p\\ l=1,\ldots,p}}^{p^2}\beta_rX_jX_l}_{Parte\ No\ Lineal} \qquad (5.16)$$

siendo la parte derecha de la igualdad (5.14), o, equivalentemente, (5.16), la *función discriminante cuadrática.*

Estamos también ante un modelo logístico, esta vez parcialmente lineal, cuadrático en la parte no lineal, que también se puede escribir como expansión lineal de funciones de base en la forma (5.1).

En ambas situaciones las distribuciones de las variables están perfectamente determinadas y para su conocimiento total tan sólo faltaría conocer el valor de los parámetros, es decir de las medias y las covarianzas, cuestión que la inferencia estadística clásica resuelve con toda eficacia, a través de la *estimación de máxima verosimilitud*.

Los estimadores de máxima verosimilitud de los parámetros $\left(p,\mu_1,\mu_0,\Sigma_1,\Sigma_0,\Sigma_1^{-1},\Sigma_0^{-1}\right)$ son

$$\hat{p}_k=\sum_{i=1}^{N_k}\frac{N_k}{N},\ \hat{\mu}_k=\sum_{i=1}^{N_k}\frac{x_i}{N_k},\ \hat{\Sigma}_k=\sum_{i=1}^{N}\sum_{y_i=k}\frac{\left(x_i-\hat{\mu}_k\right)\left(x_i-\hat{\mu}_k\right)^T}{N-k},\ \text{con}\ k=0,1.$$

Importantes referencia sobre análisis discriminante son MORRISON (1976), MARDIA et al. (1979), SEBER (1984), KRZONOWSKI (1988) CUADRAS (1991, 1992) y HASTIE et al. (2009), y referencias en las se utiliza la metodología estadística de análisis discriminante en el problema del *credit scoring* son ARTÍS *et al.* (1994), BONILLA *et al.* (2003), HAND y HENLEY (1997) y TRIAS *et al.* (2005 y 2008).

El LDA y el QDA son dos notables ejemplos de modelos logísticos basados en la razón de verosimilitud, donde se conocen las funciones de densidad condicional multivariantes. Estos dos modelos hubieran sido una estupenda solución en la mayoría de las situaciones que se plantean en la industria del credit scoring, puesto que cumplen a la perfección los requerimientos más importantes que sobre los modelos demandan los acuerdos de Basilea II, (entre otras cualidades poseen las siguientes: son buenos clasificadores, generalizan bastante bien, tienen una gran capacidad explicativa de la variable respuesta y son fácilmente interpretables), y, por último, el LDA conlleva la linealidad de todas las variables y el QDA conserva una componente lineal e incorpora la no linealidad a través de una forma cuadrática. El problema de ambas técnicas es que *requieren la hipótesis de normalidad multivariante que difícilmente se cumplirá en los datos de riesgos relacionados con el comportamiento de acreditados y solicitantes de crédito frente al default*.

La experiencia indica que la suposición sin más de normalidad multivariante en los vectores aleatorios de las variables explicativas, usualmente socio-económicas,

de los acreditados y solicitantes de crédito es enormemente arriesgada y carece de todo fundamento. Si ya es difícil que cada variable marginal se distribuya normalmente, mucho más lo será que los vectores aleatorios de las variables condicionadas a las poblaciones de default y no default, respectivamente, posean distribuciones poblacionales normales multivariante. La dificultad se acrecienta en el caso multivariante donde "*la normalidad marginal es una condición necesaria para la normalidad multivariante, pero no es suficiente*". Por tanto, antes de estimar la dependencia entre el estado de default y las variables explicativas observadas a través de un modelo de análisis discriminante habremos de asegurarnos de la normalidad multivariante de las variables condicionadas $X/Y=1$ y $X/Y=0$ y sólo entonces el uso de estos modelos daría resultados rigurosos y satisfactorios.

El requerimiento de normalidad ha provocado que las técnicas de análisis discriminante, LDA y QDA, utilizadas en ciertas épocas con profusión, principalmente el primero, como técnicas de puntuación crediticia, en multitud de casos erróneamente, por las razones antes apuntadas, estén ya prácticamente en desuso y, de hecho, son hoy en día muy escasos los trabajos donde se relaciona esta técnica con el riesgo de crédito.

Una técnica de Análisis Discriminante alternativa a las dos anteriores es el *Análisis Discriminante Basado en Distancias* de Individuos a Poblaciones que permite variables binarias, categóricas o mixtas, DBDA, técnica debida a CUADRAS et al. (1997), sobre sus trabajos pioneros, CUADRAS (1989, 1992), en a*nálisis discriminante y clasificación usando variables categóricas y continuas.*

5.5 MODELOS BASADOS EN EL CONOCIMIENTO TOTAL DE $S(X)$.

Estamos en el otro extremo de los modelos analizados en la sección 5.2, se supone que conocemos la distribución conjunta de las variables explicativas y la variable respuesta, $P(X,Y)$, o las funciones de verosimilitud condicionada a las poblaciones de default, $f_{Y=1}(X)$, y de no default $f_{Y=0}(X)$, o bien la razón de verosimilitud, $LR(X)$, así como la probabilidad de default a priori, $P(Y=1)$.

En este caso el modelo viene dado por (5.8), con todas sus componentes conocidas, siendo $S(X)=logit(p)+\log\left(LR(X)\right)$ la función de calificación de

acreditados y $p = P(Y=1)$ la probabilidad de default a priori. Como resultado del teorema de Bayes, la transformación resulta ser la logística y $S(X)$ se puede expresar como expansión lineal de funciones de base

La función de calificación de acreditados adopta la forma $S(X) = \sum_{r=0}^{1} \beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet) = \text{logit}(\bullet)$ y las funciones de base por:

$$h_r(X) = \begin{cases} 1 & r = 0 \\ \log(LR(X)) & r = 1 \end{cases} \qquad (5.17)$$

Además se conoce la formulación de los coeficientes:

$$\beta_0 = \text{logit}(p) \text{ y } \beta_1 = 1$$

es decir, el modelo presenta la siguiente estructura formal

$$\text{logit}(P(Y=1/X=x)) = \text{logit}(p) + \log(LR(X)) \qquad (5.18)$$

El modelo con conocimiento total podría también tener una función de base por variable o varias funciones de base formando una expansión lineal de todas o algunas variables. Lo importante en este caso, es que todas las funciones de base están siempre perfectamente determinadas y los coeficientes de la expansión lineal son conocidos, por tanto, no hay nada que estimar, por lo que en este caso no es necesario ningún criterio de optimización ni tiene sentido la hipótesis de regularización. Estamos ante el modelo logístico perfecto, error cero. Desgraciadamente esta situación no se da "casi nunca" en la práctica, pero no cabe duda de que es un ideal de referencia.

5.6 MODELOS BASADOS EN EL CONOCIMIENTO PARCIAL DE $S(X)$.

Por un lado, los supuestos de desconocimiento completo sobre las distribuciones conjunta o condicionada de las variables explicativas del riesgo de crédito y el estado de default son, por el momento, implanteables dentro de los requerimientos de Basilea II, y, por otro, el conocimiento total o casi total es absolutamente irreal. Por ello es necesario contar con técnicas, normalmente de regresión, o algoritmos, usualmente basados en la *teoría del aprendizaje*, que permitan estimar modelos generales de la probabilidad de default situados entre ambos extremos, es decir, en un *conocimiento parcial razonable*.

Así podremos especificar en el modelo, de manera fundamentada, algunas o todas las variables con estructura lineal, o con estructura no lineal, y dentro de la no linealidad seleccionar algunas de las innumerables formas en que esta puede manifestarse, todo ello estructurado sobre el armazón común de las expansiones lineales por funciones de base de las variables explicativas del estado de default.

Todos los modelos a los que nos hemos referido en las secciones 5.2, 5.3, 5.4, y 5.5 presentan estructura formal en términos de (5.1), por lo que para todos la función $S(X)$ es una expansión lineal de funciones de base del vector de variables explicativas X y en todos, a excepción de los estimadores $k-NN$, *la función* $g(\bullet)$ es *la logística*. En la literatura actual sobre sistemas de calificación de acreditados se proponen modelos donde $g(\bullet)$ también puede ser la *función de enlace probit* o, en el campo de la teoría del aprendizaje, la *función de enlace vector soporte*. En el resto de este capítulo nos referiremos a modelos con los tres tipos de transformación, es decir, veremos modelos *logísticos, probit* y *vector soporte*.

Los métodos orientados a estimar los modelos logísticos y probit, regularizados o no, son muy similares, de hecho ambos trabajan con modelos lineales de probabilidad generalizados y estiman la función de calificación de acreditados a través de la minimización de la pérdida empírica media que representan las discrepancias entre los valores pronosticados por el modelo y los observados, más un término de regularización, de ser necesario, a través del método iterativo de Newton Raphson. A partir de la función de calificación estimada, con la transformación correspondiente de esta, se calcula el estimador de la probabilidad de default y, a partir de cualquiera de los dos estimadores, el clasificador Bayes optimal de acreditados y solicitantes de crédito.

Modelos Logísticos: $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=S\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$, la función objetivo a minimizar, se obtiene a partir de la pérdida logística,

$$\ell_\Lambda\left(Y,S\left(X\right)\right)=-\left(Y\ S\left(X\right)-\log\left(1+e^{S(X)}\right)\right)$$

Modelos Probit: $\text{probit}\left(P\left(Y=1/X=x\right)\right)=S\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$, la función objetivo a minimizar se obtiene, en este caso, a partir de la pérdida probit,

$$\ell_{\Phi}\left(Y,S\left(X\right)\right)=-\left(Y\ \log\Phi\left(S\left(X\right)\right)+\left(1-Y\right)\log\left(1-\Phi\left(S\left(X\right)\right)\right)\right).$$

Las diferencias importantes se encuentran entre los métodos anteriores y los métodos de estimación de los *Modelos de Vector Soporte*, pues estos últimos, orientados al problema de la clasificación, aunque vienen revelándose como excelentes clasificadores en las aplicaciones prácticas, HÄRDLE et al. (2005, 2007, 2011), *presentan la insalvable debilidad de no proporcionar directamente estimadores de la probabilidad de default subyacente*. Los modelos vector soporte adoptan la forma

Modelos Vector Soporte: $\text{Sign}\left\{\left(P\left(Y=1/X=x\right)\right)-\frac{1}{2}\right\}=S\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$ y la función a optimizar se obtiene a partir de la pérdida bisagra VSM

$$\ell_{SVM}\left(Y,S\left(X\right)\right)=\left[1-Y\ S\left(X\right)\right]_+$$

Los modelos disponibles para representar la relación de dependencia entre la variable estado de default y las variables explicativas del riesgo de crédito y las técnicas para estimarlos se pueden clasificar en función de que el modelo contemple la linealidad o no linealidad de las variables que lo conforman.

5.6.1 Supuesto de Linealidad de Todas las Variables.

Cuando hay razones suficientes para suponer que todas las variables explicativas son lineales, o que al menos este supuesto puede recoger adecuadamente la relación de dependencia entre la variable estado de default y las variables explicativas, se establece una hipótesis muy simple, *la función de calificación de acreditados* $S(X)$ *es lineal en* $X=\left(X_1,...,X_p\right)^T$. Bajo la hipótesis de linealidad se tiene:

La función de calificación de acreditados adopta la forma $S\left(X\right)=\sum_{r=0}^{p}\beta_r h_r\left(X\right)$, la función de enlace viene dada por $g\left(\bullet\right)$ y las funciones de base por:

$$h_r\left(X\right)=\begin{cases}1 & r=0\\ X_r & r=1,...,p\end{cases} \qquad (5.19)$$

es decir, el modelo presenta la siguiente estructura formal

$$g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p}\beta_r X_r \qquad (5.20)$$

Ejemplos notables de este tipo de modelos son, por un lado, la función discriminante de Fisher, LDA, que hemos visto en la sección 5.4, y, por otro, los *Modelos Lineales de Probabilidad*, PLM, donde $g(\bullet)$ es la transformación identidad y que se ajustan por regresión lineal. Pero la hipótesis excesivamente restrictiva de normalidad multivariante de las variables explicativas constituyó, como ya hemos comentado, un serio inconveniente para que el LDA se consolidara como técnica de predicción y clasificación en muchos dominios científicos, principalmente en el campo económico financiero y, por otro lado, los modelos lineales de probabilidad constituyeron un fracaso desde su propia concepción conceptual, por cuanto el estimador pronostica valores fuera del intervalo cerrado de números reales $[0,1]$, por lo que difícilmente puede pronosticar una probabilidad.

Coincidente en el tiempo con la aparición de la *función discriminante lineal* de *Fisher* , GADDUM (1933) y BLISS (1934a, 1934b, 1935), sobre los trabajos de FECHNER (1860), en particular la transformación de frecuencias a desviaciones normales equivalentes, introdujeron el término *probit* y los conceptos básicos de este tipo de modelos**,** comenzando una serie de publicaciones sobre aplicaciones prácticas del probit, que Gaddum continúo hasta comienzos de la década de 1950.

La introducción de la función logística como alternativa a la función de probabilidad normal se debe a BERKSON (1944, 1950) que consolidó el concepto de *logit* y sobre él fundamentó la *Regresión Logística Lineal*, LLR, y fue el primero en utilizarla en aplicaciones prácticas. CORNFIELD (1951, 1956) aportó importantes contribuciones prácticas tales como la estimación de los valores del riesgo relativo a partir de las tasas odds y a mediados de la década de 1950 aparecieron las primeras aplicaciones de la regresión logística en el dominio económico empresarial, (investigación de mercados), FARREL (1954), AITCHISON y BROWN (1957) y ADAM (1958), pero la técnica no se popularizó hasta la publicación del libro "The Analysis of Binary Data" de (COX 1970).

Con la introducción de los *Modelos Lineales Generalizados*, GLM, por NELDER y WEDDERBURN (1972) se unificaron las técnicas logit y probit, puesto que la única diferencia entre ambas reside en la función de enlace entre la variable respuesta y las variables explicativas.

En la década de 1980 se introdujeron la regresión logística y la programación lineal en la construcción de *tarjetas de puntuación de acreditados*. La regresión logística llegó al credit scoring de la mano de OHLSON (1980) que fue el primero en aplicar la probabilidad condicional y en particular el modelo logístico para predecir la quiebra empresarial, modelo que por su estructura lineal pura es el más amigable y de más fácil interpretación de los modelos paramétricos de estimación de la probabilidad de default, y, de hecho, es el modelo más utilizado en credit scoring.

5.6.1.1 Regresión Logística Lineal, LLR o LOGIT.

La hipótesis de linealidad, $S(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}$, junto a la función de enlace logística entre la variable respuesta y las variables explicativas conducen al modelo lineal

La función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por (5.19), es decir, la función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=0}^{p}\beta_r X_r$, y por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p}\beta_r X_r=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X} \tag{5.21}$$

donde $\boldsymbol{X}=\left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta}=\left(\beta_1,...,\beta_p\right)^T$.

El modelo (5.21) es conocido como LOGIT y su ajuste se realiza a través de la ***Regresión Logística Lineal,*** **LLR**. Se estiman el parámetro β_0 y el vector de parámetros $\boldsymbol{\beta}$ resolviendo el *problema de estimación*:

$$\min_{\beta_0,\boldsymbol{\beta}} \ -\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}\right)-\boldsymbol{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}\right)\right)\right] \tag{5.22}$$

La solución para el logit viene dada por

$$\left(\beta_0,\boldsymbol{\beta}\right)^{nuevo}=\left[(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}\ (\boldsymbol{1},\boldsymbol{X})\right]^{-1}(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}\boldsymbol{z} \tag{5.23}$$

donde

$$\boldsymbol{z}=\left(\left(\beta_0,\boldsymbol{\beta}\right)^{anterior}\right)^T(\boldsymbol{1},\boldsymbol{X})+\boldsymbol{W}^{-1}(Y-P) \tag{5.24}$$

$\boldsymbol{z}$ es la variable respuesta ajustada y $(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{\beta}^{anterior}$ es la expansión de funciones de base estimada en el paso anterior de Newton-Raphson. $\boldsymbol{W}$ es la matriz diagonal con

elementos diagonales $p(x_i)(1-p(x_i))$, $\boldsymbol{W} = diag\left[p(x_i)(1-p(x_i))\right]_{N\times N}$ siendo $\boldsymbol{p}$ es el vector de las probabilidades estimadas tales que las respuestas son iguales a 1, es decir,

$$\boldsymbol{p} = \frac{e^{(\beta_0,\boldsymbol{\beta})^T \boldsymbol{X}}}{1+e^{(\beta_0,\boldsymbol{\beta})^T}} = \left(P\left(Y=1/\ X=x_1,(\beta_0,\boldsymbol{\beta})^{anterior}\right),...,P\left(Y=1/\ X=x_N,(\beta_0,\boldsymbol{\beta})^{anterior}\right)\right)$$

Usualmente se usa $(\beta_0,\boldsymbol{\beta})^{inicial} = (0,...,0)^T$ como valor inicial para el procedimiento iterativo. Aunque el algoritmo normalmente converge, puesto que la log-verosimilitud negativa es convexa, la convergencia no siempre está garantizada debido a que puede producirse sobreajuste.

La nueva expansión de las variables explicativas del default se expresa según la siguiente igualdad

$$S(X)^{anterior} = \beta_0 + \left(\boldsymbol{\beta}^{anterior}\right)^T \boldsymbol{X} = (\boldsymbol{1},\boldsymbol{X})\left[(\boldsymbol{1},\boldsymbol{X})^T \boldsymbol{W}\ (\boldsymbol{1},\boldsymbol{X})\right]^{-1}(\boldsymbol{1},\boldsymbol{X})^T \boldsymbol{W}\boldsymbol{z} = \boldsymbol{S}_{\lambda W m_h}\boldsymbol{z} \qquad (5.25)$$

Este algoritmo coincide con el IRLS, (*Iterative Reweighted Least Squares*), mínimos cuadrados reponderados iterativos, puesto que cada iteración resuelve el problema de mínimos cuadrados ponderados

$$(\beta_0,\beta)^{nuevo} \leftarrow \min_{(\beta_0,\beta)}\left(\boldsymbol{z}-\left(\beta_0+\beta^T\boldsymbol{X}\right)\right)^T \boldsymbol{W}\left(\boldsymbol{z}-\left(\beta_0+\beta^T\boldsymbol{X}\right)\right)$$

Como puede verse, se alterna entre la formación de una variable dependiente ajustada $\boldsymbol{z}$ (5.24) con varianza $\boldsymbol{W}^{-1}$ y regresionar $\boldsymbol{z}$ sobre $\boldsymbol{X}$ con pesos $\boldsymbol{W}$.

En este caso los grados de libertad efectivos son aproximados por la expresión, (HASTIE Y TIBSHIRANI, 1999),

$$df(\lambda) = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \qquad (5.26)$$

y la varianza de los coeficientes, (GRAY 1992),

$$\begin{aligned} Var(\beta) &= Var\left[\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}\right] \\ &= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1} Var\left[\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{p})\right]\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1} \\ &= \left(H(\beta)\right)^{-1} I(\beta)\left(H(\beta)\right)^{-1} \end{aligned} \qquad (5.27)$$

donde $\boldsymbol{W}$ es la obtenida en el paso final del algoritmo e $\boldsymbol{I}(\beta) = -E\left[\frac{\partial^2 L(\beta)}{\partial\beta\ \partial\beta^T}\right]$ *es la matriz de información de Fisher.*

La inferencia sobre este modelo se basa en el estadístico razón de verosimilitud (COX, (1970), THEIL, (1979), COX y SNELL (1989), COLLET (1991), KLEINBAUM (1994), HOSMER y LEMESHOW (2000), CRAMER (2003), HASTIE et al. (2009)).

Bajo condiciones relativamente generales, el estimador de máxima verosimilitud tiene las siguientes propiedades de interés (THEIL, 1979):

La teoría de verosimilitud dice que si el modelo es correcto, entonces

1.- $\hat{\beta}$ *es un estimador consistente de* β.

$$\hat{\beta} \xrightarrow{P} \beta \tag{5.28}$$

2.- $\hat{\beta}$ *distribuye asintóticamente normal.*

$$\hat{\beta} \overset{(a)}{\sim} N\left(\beta, \{\boldsymbol{I}(\beta)\}^{-1}\right) \tag{5.29}$$

donde $\boldsymbol{I}(\beta) = -E\left[\frac{\partial^2 L(\beta)}{\partial\beta\ \partial\beta^T}\right]$ *es la matriz de información de Fisher.*

o, equivalentemente, $\hat{\beta} \overset{(a)}{\sim} N\left(\beta, \left(X^T W X\right)^{-1}\right)$

3.- $\hat{\beta}$ *es asintóticamente eficiente.*

Resulta relativamente sencillo estimar $\boldsymbol{I}(\beta)$ sustituyendo el parámetro β por $\hat{\beta}$, debido que en este modelo $E\left[\boldsymbol{H}(\beta)\right] = \boldsymbol{H}(\beta)$, *por lo que* $\boldsymbol{I}(\beta) = -\frac{\partial^2 L(\beta)}{\partial\beta\ \partial\beta^T}$.

El estimador de máxima verosimilitud $\hat{\beta}$ satisface una relación consistente: sus componentes son los coeficientes del ajuste de mínimos cuadrados, donde las respuestas son

$$z_i = x_i^T \hat{\beta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)} \tag{5.30}$$

y los pesos son $w_i = \frac{\hat{p}_i}{(1-\hat{p}_i)}$, ambos dependientes de $\hat{\beta}$.

5.6.1.2 Regresión Logística Lineal L_2-Penalizada, LLR_ L_2.

Los dos posibles problemas que surgen de la minimización del riesgo empírico en la regresión logística, infinitas soluciones y sobre e infraajuste, pueden evitarse en gran medida activando el término de regularización en dicho riesgo, es decir, especificando el modelo bajo la formulación (5.1), con la transformación logística y resolviendo el problema de estimación con la función objetivo del problema (5.22) regularizada por el funcional de regularización $J\left((\beta_0,\boldsymbol{\beta})^T(1,X)\right)=\frac{1}{2}\|\boldsymbol{\beta}\|^2$, es decir, resolviendo el problema de minimización

$$\min_{(\beta_0,\boldsymbol{\beta})} -\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}\right)\right)\right]+\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{5.31}$$

Nos encontramos, de este modo, ante la ***Regresión Logística Lineal L_2-Penalizada*, LLR_L_2.**

Dado que

$$\boldsymbol{J}'(\boldsymbol{\beta})=\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{1}{2}\|\boldsymbol{\beta}\|^2\right)=2\boldsymbol{\beta}^{anterior}$$

$$\boldsymbol{J}''(\boldsymbol{\beta})=\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\left(\frac{1}{2}\|\boldsymbol{\beta}\|^2\right)=\frac{\partial}{\partial\boldsymbol{\beta}^T}(\boldsymbol{\beta})=\boldsymbol{D}$$

(donde las derivadas se evalúan en $\boldsymbol{\beta}^{anterior}$ y $\boldsymbol{D}$ es la matriz diagonal con elementos en la diagonal $\{0,1,...,1\}_{1\times(p+1)}$), la solución al problema (5.31) viene dada por

$$(\beta,\boldsymbol{\beta})^{nuevo}=\left[(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}\ (\boldsymbol{1},\boldsymbol{X})+\lambda\boldsymbol{D}\right]^{-1}(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}\boldsymbol{z} \tag{5.32}$$

donde la variable respuesta ajustada $\boldsymbol{z}$ viene dada por

$$\boldsymbol{z}=\beta_0+\left(\boldsymbol{\beta}^{anterior}\right)^T\boldsymbol{X}+\boldsymbol{W}^{-1}(Y-P)=S(\boldsymbol{X})^{anterior}+\boldsymbol{W}^{-1}(Y-P) \tag{5.33}$$

La nueva expansión de las variables explicativas del default se expresa según la siguiente igualdad

$$S(X)^{nueva}=S_{\lambda W}\boldsymbol{z} \tag{5.34}$$

donde

$$S_{\lambda W}=\left((\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}(\boldsymbol{1},\boldsymbol{X})+\lambda\boldsymbol{D}\right)^{-1}(\boldsymbol{1},\boldsymbol{X})^T\boldsymbol{W}(\boldsymbol{1},\boldsymbol{X})$$

es el operador de regresión.

La solución de(5.31) puede interpretarse en un contexto bayesiano, como un estimador de maxima probabilidad a posteriori (MAP) de β_0 y $\boldsymbol{\beta}$, cuando $\boldsymbol{\beta}$ tiene una distribución a priori Gaussiana sobre $\mathbb{R}^p$ con media cero y covarianza λI y β_0 tiene una distribución a priori uniforme sobre $\mathbb{R}$ (CHALONER y LARNTZ, (1989) y JAAKKOLA y JORDAN, (2000)).

5.6.1.3 Regresión Probit Lineal, LPR.

La hipótesis de linealidad, $S(X) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}$, junto a la función de enlace *probit* entre la variable respuesta y las variables explicativas conducen al modelo lineal siguiente:

La función de enlace viene dada por $g(\bullet) = \text{probit}(\bullet)$ y las funciones de base por (5.19), es decir, la función de calificación de acreditados adopta la forma $S(X) = \beta_0 + \sum_{r=0}^{p} \beta_r X_r$, y en este caso el modelo presenta la siguiente estructura formal

$$\text{probit}\left(P(Y=1/\ X=x)\right) = \beta_0 + \sum_{r=1}^{p} \beta_r X_r = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X} \tag{5.35}$$

donde $\boldsymbol{X} = (X_1,...,X_p)^T$, $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$.

El ajuste del modelo (5.35), llamado PROBIT, se realiza a través de la ***Regresión Probit Lineal*****, LPR**, que estima el parámetro β_0 y el vector de parámetros $\boldsymbol{\beta}$ resolviendo el problema de minimización

$$\min_{\beta_0,\boldsymbol{\beta}} -\left[\boldsymbol{Y}^T \log\left(\boldsymbol{\Phi}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}\right)\right) + (\boldsymbol{1}-\boldsymbol{Y})^T \log\left(1-\boldsymbol{\Phi}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}\right)\right)\right] \tag{5.36}$$

El proceso es análogo al de LLR y también se ha utilizado bastante en credit scoring, pero sin alcanzar los niveles de popularidad, y, según la mayoría de los autores, ni la eficacia del modelo logit. Es lógico que así sea, por cuanto la probabilidad de default y la esperanza condicional se relacionan con las características del estado de default a través de la función de distribución logística, tal como establece el teorema de Bayes.

El hecho de que el modelo probit haya funcionado aceptablemente bien se debe al extraordinario parecido entre las funciones de distribución logística y normal tipificada. Pero a pesar de este parecido, ambos modelos no pueden ser

comparados, puesto que tienen diferentes varianzas, 1 para la normal y $\pi^2/3$ para la función de distribución logística tipificada. Por tanto es necesario reescalar algunos de los parámetros en orden a comparar ambas distribuciones.

5.6.1.4 Métodos de Separación. Clasificador Vector Soporte

En la segunda mitad de la década de 1990, basados en la *Teoría Estadística de aprendizaje,* se introdujeron como alternativa a los *métodos logísticos* los *métodos de separación*, clasificadores que son extensiones del modelo *Perceptron,* ROSENBLATT (1958, 1962), cuyo ajuste se obtiene por optimización de la función de pérdida *"margen suave" o "vector soporte"*.

El modelo conocido en la literatura como *perceptron* desde hace más de 50 años, es un clasificador que computa una combinación lineal de las variables explicativas, el *Hiperplano Separador,* SH, y devuelve el signo. Estos algoritmos de aprendizaje, que intentan encontrar un hiperplano separador que minimice la distancia de los puntos mal clasificados a la frontera de decisión, establecieron los fundamentos de los modelos de redes neuronales de las décadas de 1980 y 1990.

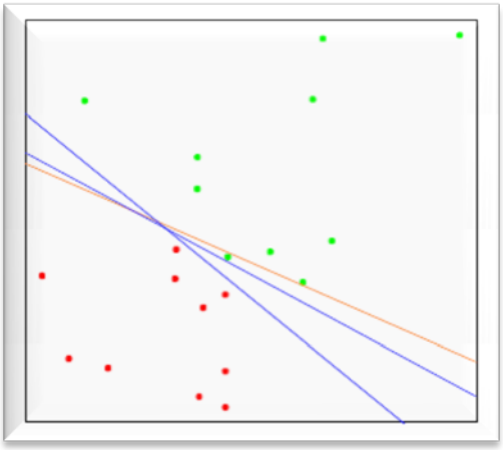
Dada una muestra de entrenamiento $\tau = \left\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1,\ldots,N} \in \left(X \times Y\right)^N$ con $y_i \in \{-1,+1\}$, se define un *hipeplano* en los términos siguientes:

Definición 5.2.- Un *hiperplano o conjunto afín en* $\mathbb{R}^p$ está definido por el conjunto

$$\left\{x \in \mathbb{R}^p \,/\, f(x) = \beta_0 + \beta^T x = 0\right\}$$

donde β es un vector unidad, $\|\beta\| = 1$.

Si observamos la figura (5.3), donde se han representado varios puntos pertencientes a dos poblaciones distintas en $\mathbb{R}^2$, vemos un claro ejemplo de dos clases que pueden ser separadas perfectamente por una frontera lineal. Incluidas en la figura se ven dos rectas de las infinitas que pueden separar perfectamente las clases, son dos *hiperplanos separadores*. La línea de color naranja es la solución de mínimos cuadrados al problema, obtenida por regresión de la respuesta Y $(-1/+1)$ sobre X (con intercepto), la línea está dada por el hiperplano $\left\{x \in \mathbb{R}^2 / \beta_0 + \beta^T x = 0\right\}$, al que llamaremos *hiperplano separador*.



Fuente: Hastie et al. (2009).

Figura 5.3.- *Un ejemplo con dos clases separables por un hiperplano.*

De entre los muchos problemas asociados con perceptron destaca sobre todo el hecho de que cuando los datos son separables, existen varias soluciones y cuál de ellas se encuentre depende de los valores de inicialización (RIPLEY, 1996).

Con el objetivo de resolver el problema de la no unicidad de la solución que presenta SH, VAPNIK, (1996) introdujo el *Hiperplano Separador Optimal,* (**OSH**, O*ptimal* ***S****eparating Hyperplane*), que separa las dos clases sin error y maximiza la distancia para los puntos más próximos de una clase a la otra, es decir maximiza una noción geométrica de margen M, lo que se consigue a base de añadir al hiperplano separador restricciones adicionales.

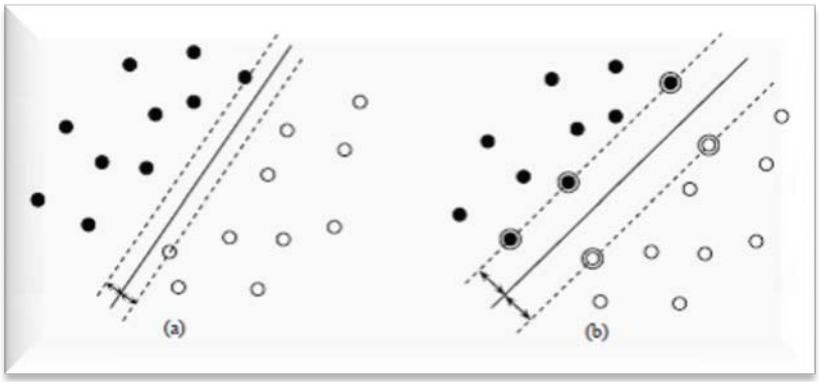


Figura 5.4.- (a) Hiperplano Separador con margen mínimo.
(b) Hiperplano Separador con margen óptimo.

Si los elementos de la muestra no son linealmente separables, no existe el Hiperplano Separador Optimal que pueda clasificar correctamente a estos elementos.

El hiperplano separador optimal produce una función $\hat{S}(X)=\hat{\beta}_0+\hat{\beta}^T X$ para clasificar nuevas observaciones, a través del clasificador Bayes optimal:

$$\Gamma_{OSH}(X)=sign\ \hat{S}(X) \tag{5.37}$$

A pesar de que ninguna de las observaciones de entrenamiento cae en el margen (por construcción), esto puede no ser necesariamente verdad para el caso de las observaciones test. La intuición indica que un margen grande sobre los datos de entrenamiento debe conducir a una buena separación sobre los datos test.

Cuando los datos no son separables no existe el OHS, por lo que es necesario recurrir a técnicas alternativas. De nuevo se puede agrandar el espacio usando expansiones por funciones de base, pero esto puede conducir a una separación artificial y, por tanto, al sobreajuste. Una alternativa más atractiva, extensión del OSH, permite solapamientos de clases en el espacio de las variables explicativas, pero minimiza una medida del alcance de este solapamiento. Una forma de tratar con los solapamientos es relajar la maximización del margen M permitiendo que algunos puntos estén en el lado equivocado del margen, lo que se consigue modificando la restricción $y_i\left[\beta_0+\beta^T x_i\right]\geq M,\ \ i=1,...,N$, en la forma siguiente:

$$y_i\left[\beta_0+\beta^T x_i\right]\geq M\left(1-\xi_i\right) \tag{5.38}$$

siendo $\xi=(\xi_1,...,\xi_N)$ una variable, llamada de "necesidad" que verifica que $\forall i,\ \xi_i\geq 0,\ \sum_{i=1}^{N}\xi_i\leq \text{constante}$. La hipótesis de linealidad, $S(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}$, junto a la función de enlace *vector soporte* entre la variable respuesta y las variables explicativas conducen al modelo lineal:

La función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=0}^{p}\beta_r X_r$, la función de enlace viene dada por $g(\bullet)=sign\left\{\bullet-\frac{1}{2}\right\}$ y las funciones de base por (5.19), por tanto, el modelo presenta la siguiente estructura formal

$$\text{sign}\left\{P(Y=1/\ X=x)-\frac{1}{2}\right\}=\beta_0+\sum_{r=1}^{p}\beta_r X_r=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X} \tag{5.39}$$

El ajuste del modelo (5.39), llamado ***Clasificador de Vector Soporte*****, SV**, se realiza a través del *problema del clasificador de vector soporte para el caso no separable* en la forma

$$\begin{aligned} &\underset{\beta_0,\beta}{mín}\ \frac{1}{2}\|\beta\|^2 \\ &\textit{sujeto a} \\ &y_i\left[\beta_0+\beta^T x_i\right]\geq 1-\xi_i,\quad \forall i=1,\ldots,N \\ &\xi_i\geq 0,\ \sum_{i=1}^{N}\xi_i\leq \text{constante} \end{aligned} \tag{5.40}$$

El problema de optimización convexa (5.40) es un *problema de programación matemática cuadrático con restricciones de desigualdad lineal*, que puede plantearse como un *problema de regresión lineal regularizada* para la función de pérdida vector soporte, (3.25), con la norma L_2, $\|\boldsymbol{\beta}\|^2$, como término de regularización:

$$\min_{(\beta_0\boldsymbol{\beta})}\left[\mathbf{1}^T\left(\mathbf{1}-\mathbf{Y}\left(\beta_0+\boldsymbol{\beta}^T\mathbf{X}\right)\right)\right]_+ +\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{5.41}$$

conocido como *Problema de Regresión Lineal SV***.** Es evidente que los problemas de optimización (5.39) y (5.41) son equivalentes, es decir, el problema de optimización VS adopta la forma de *pérdida* + *penalización*. Por tanto, VS resuelve el problema de ajustar la función de calificación con criterio basado en la *función de pérdida bisagra VS,* $\ell_{VS}\left(Y,\ \beta_0+\boldsymbol{\beta}^T X\right)=\left[1-Y\left(\beta_0+\boldsymbol{\beta}^T X\right)\right]_+$ y una forma de regularización particular con dos objetivos: *maximizar el margen* $\frac{2}{\|\beta\|}$ *y simultáneamente minimizar el grado de mala clasificación* $\sum_{i=1}^{N}\xi_i$, *y dos restricciones,* $\xi_i\geq 0$ e $y_i\left[\beta_0+\beta^T x_i\right]\geq 1-\xi_i$, $i=1,\ldots,N$.

Por la naturaleza de este problema, se tiene que aquellos puntos bien situados del lado de la frontera de su clase no juegan un gran papel en configurar la misma. Esta es una propiedad muy atractiva y representa una notable diferencia con el LDA (donde la frontera de decisión está determinada por la covarianza de las distribuciones de las clases y la posición de los centroides de las misma). Desde este punto de vista la regresión lineal es más parecida al clasificador de vector de soportes.

Dadas las soluciones $\hat{\beta}_0$ *y* $\hat{\beta}$, el modelo estimado adopta la forma

$$Sign\left\{\hat{P}\left(Y=1/\ X=x\right)-\frac{1}{2}\right\}=\hat{\beta}_0+\hat{\beta}^T X \tag{5.42}$$

Como ocurre en el OSH, desde la ecuación (5.42) no es posible conocer el estimador de la probabilidad de default, $\hat{P}\left(Y=1/\ X=x\right)$, sino simplemente si este es igual o superior a 0.50 o inferior, por tanto, *no es posible estimar directamente la probabilidad de default en los modelos VS, lo que constituye sin duda alguna la mayor debilidad del método, en orden a estimar la probabilidad de default según los requerimientos de Basilea II.*

Los métodos SV son muy populares por su buena ejecución en la clasificación binaria y según HÄRDLE et al. (2007) las primeras aplicaciones prácticas apuntan a que estos métodos tienen una eficacia superior en la clasificación de nuevos solicitantes de crédito en comparación con el análisis discriminante, LDA y modelo LOGIT.

Desde el punto de vista de *la minimización de la pérdida empírica regularizada*, el clasificador Vector Soporte y la Regresión Logística Lineal L_2-Penalizada resultan bastante similares:

$$VS: \qquad \underset{\beta_0,\boldsymbol{\beta}}{mín}\ \sum_{i=1}^{N}\left[1-y_i\left(\beta_0+\boldsymbol{\beta}^T x_i\right)\right]_+ +\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{5.43}$$

$$RLL\_L_2: \quad \underset{\beta_0,\boldsymbol{\beta}}{min}\ -\sum_{i=1}^{N} y_i\left(\beta_0+\boldsymbol{\beta}^T x_i\right)-\log\left(1+e^{\left(\beta_0+\boldsymbol{\beta}^T x_i\right)}\right)+\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{5.44}$$

Por otro lado se demuestra (ROSSET et al. 2004 y ZHU y HASTIE 2004) que para datos separables, $\lambda\to 0$, el estimador del modelo de regresión logística lineal regularizada converge al clasificador de margen máximal: $LLR-L_2 \xrightarrow[\lambda\to 0]{} VS$. Este es un resultado de capital importancia, por cuanto para modelos que necesiten regularización, por ejemplos los logísticos no supervisados con muestras de entrenamiento pequeñas, la clasificación obtenida por un modelo logístico ajustado por LLR_L2, tiende a ser tan buena como la clasificación SV.

Sintetizando este apartado 5.6.1, dedicado a los métodos lineales de estimación de la probabilidad de default, podemos afirmar que:

> La situación en que todas las variables explicativas se pueden suponer lineales es con mucho la situación más apetecible cuando el conocimiento sobre la relación de dependencia entre la variable estado de default y las variables explicativas del

riesgo de crédito no es completo, lo que casi siempre ocurre. Con ello se consiguen modelos paramétricos que fijan completamente la estructura del modelo, con lo que se alcanzan las cualidades idóneas para nuestro objetivo, cualidades ligadas a la triple calidad que Basilea II requiere a un modelo, calidad explicativa, predictiva y discriminante, lo que se traduce en equilibrio entre flexibilidad y dimensión, interpretabilidad y capacidad de generalización.

Respecto de la capacidad de generalización, debemos recordar que el modelo no solo ha de describir bien la relación entre el estado de default y las características de riesgo seleccionadas por los expertos y clasificar lo más correctamente posible a los acreditados que han servido de base para su construcción, sino que además es fundamental que extienda esta relación y clasifique correctamente a nuevos acreditados distintos a los de entrenamiento. Para ello es imprescindible que el modelo no esté sobreajustado, y, desde luego un modelo lineal dista bastante, con muestras suficientemente grandes, de estar sobreajustado, por el contrario, son las funciones complejas las propensas al sobreajuste. Un modelo complejo puede describir los datos muestrales muy bien y sin embargo no generalizar adecuadamente los resultados a la población en estudio por estar excesivamente sobreajustado o infraajustado.

Para finalizar queremos manifestar que, *en todo caso, una información que poseemos siempre es que la variable respuesta es binaria y, por tanto, que el nexo natural que une el verdadero modelo de las variables explicativas con la variable respuesta es la transformación logit.*

5.6.2 Supuesto de No Linealidad de Algunas de las Variables Explicativas.

Desde luego la linealidad es una característica muy deseable aunque no siempre alcanzable y cuando no se alcance, lo que ocurre con frecuencia en la práctica, serán necesarios métodos alternativos a los modelos lineales, *sean logísticos, probit* o *vector soporte*, pero para ello debemos de enfrentarnos al dilema de Occam con decisión, lo que significa obtener el modelo de la forma más sencilla que sea posible, salvaguardando la eficacia de los objetivos perseguidos.

En la década de 1980 se inició una febril actividad en la investigación para caracterizar la no linealidad de las variables explicativas en los modelos estadísticos, siendo los más relevantes los que expondremos a continuación.

5.6.2.1 Modelos Aditivos Generalizados, GAM.

Una familia de modelos para caracterizar la no linealidad de las variables explicativas a la vez que eliminar la maldición de la dimensionalidad es la de los ***Modelos Aditivos Generalizados,*** **GAM**. *Estos modelos son extensiones de* ***los*** **Modelos Lineales Generalizados, GLM**, *que combinan la precisión estadística típica de una variable explicativa unidimensional con la flexibilidad de los modelos semiparamétricos de variables explicativas multidimensionales, por lo que en estos modelos no está presente la maldición de la dimensionalidad. Los GAM, aparte de proporcionar eficaces reglas de predicción y clasificación así como herramientas para encontrar la importancia subyacente de las diferentes variables explicativas, contemplan métodos estadísticos automáticos más flexibles que los métodos logísticos ordinarios. Si bien en su contra está el hecho de que eluden la maldición de la dimensionalidad a costa de una hipótesis muy poco realista en credit scoring, la aditividad de los efectos de riesgo en su relación con el estado de default, (hipótesis aditiva).*

El primero en considerar los modelos aditivos fue LEONTIEF (1947), en el contexto del Análisis Input-Output, a los que le llamó ***Modelos Aditivos Separables,*** **SAM**. Deberían pasar casi 40 años hasta que, con el fin de evitar la maldición de la dimensionalidad, STONE (1985) propusiera los ***Modelos Aditivos*****, AM**, en los términos siguientes:

$$\begin{aligned} &Y = S(X) = \beta_0 + \sum_{j=1}^{p} S_j(X_j), \quad \beta_0 \in \mathbb{R} \\ &S_j(\bullet): \mathbb{R} \to \mathbb{R}, \text{ función de suavizado de dimensión infinita} \\ &E\left[S_j(X_j)\right] = 0 \quad (j = 1,\ldots,p) \end{aligned} \tag{5.45}$$

es decir, se supone que la función de calificación $S(X)$ es la suma de funciones de "suavizado" no paramétricas de dimensión infinita, no especificadas, $S_j(\bullet)$, que operan sobre un argumento unidimensional X_j, (la hipótesis $E\left[S_j(X_j)\right]=0$ conduce a un modelo identificable pues en otro caso se podría añadir o sustraer constantes en las funciones unidimensionales $S_j(\bullet)$).

STONE, (1986a, 1986b), extendió su propuesta a los ***Modelos Aditivos Generalizados*** **, GAM**, y HASTIE y TIBSHIRANI (1986, 1990) lo adaptaron a la forma conocida hoy en día.

Un modelo aditivo generalizado de respuesta binaria, particularizado a la probabilidad de default, se puede expresa como una expansión aditiva de funciones de base de suavizado de dimensión infinita, es decir

La función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=1}^{p}h_r(X)$, la función de enlace viene dada por $g(\bullet)$ y las funciones de base por

$$\begin{aligned} &h_r(X)=S_r(X_r),\ \text{función de suavizado de dimensión infinita}\\ &E\left[S_r(X_r)\right]=0\quad (r=1,...,p)\end{aligned} \tag{5.46}$$

por tanto, el modelo presenta la siguiente estructura formal

$$g\left(P(Y=1/\,X=x)\right)=\beta_0+\sum_{r=1}^{p}S_r(X_r),\quad \beta_0\in\mathbb{R} \tag{5.47}$$

Los beneficios de los modelos aditivos generalizados son al menos dos: primero, dado que cada uno de los términos aditivos se estima usando un suavizador univariante, se esquiva la maldición de la dimensionalidad. Segundo, las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas observadas, *lo que confiere a los modelos aditivos la habilidad de descubrir los patrones no lineales sin sacrificar la interpretabilidad.*

La facilidad de interpretación hace que los modelos aditivos generalizados, sobre todo los obtenidos a través de la pérdida logística, sean particularmente atractivos en los sistemas de calificación de acreditados, en orden al cumplimiento de los requerimientos de Basilea II. Además, la aditividad de los parámetros de entrada en estos modelos es consistente con la descentralización en la toma de decisiones de riesgo u optimización por etapas. En definitiva *los modelos aditivos generalizados, igual que ocurre con los modelos lineales, son fácilmente interpretables*. Por tanto, el modelo más interesante de la familia de los modelos aditivos generalizados, para la construcción de los sistemas de calificación de acreditados, es sin duda el **Modelo Logístico Aditivo, ALM**, $\left(g(\bullet)=\text{logit}(\bullet)\right)$, donde las funciones de base se definen como en (5.46).

La función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=1}^{p}h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por

$$\begin{aligned} &h_r(X):\ \mathbb{R}^p \rightarrow \mathbb{R} \\ &\qquad X \rightarrow h_r(X)=S_r(X_r),\ \text{función de suavizado de dimensión infinita} \\ &\qquad E\left[S_r(X_r)\right]=0\ \ (r=1,...,p) \end{aligned} \tag{5.48}$$

por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/\ X=x)\right)=\beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})=\beta_0+\sum_{r=1}^{p}S_r(X_r),\quad \beta_0\in\mathbb{R} \tag{5.49}$$

donde $\mathbf{1}^T=(1,...,1)_{1\times p}$, $\mathbf{S}(\mathbf{X})=\left(S_1(X_1),...,S_p(X_p)\right)^T$.

Las funciones $S_r(X_r)$ pueden estimarse de una manera flexible a través de cualquier método de suavizado no paramétrico.

La estimación de la función de calificación de acreditados $\beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})$, $\hat{S}(X)=\hat{\beta}_0+\mathbf{1}^T\hat{\mathbf{S}}(\mathbf{X})$, consiste en estimar β_0 y las funciones de base $S_r(X_r)$, resolviendo el problema de optimización

$$\min_{\beta_0,\mathbf{S}}\ L_{\ell_\Lambda}\left(Y,\ \beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})\right) \tag{5.50}$$

$$L_{\ell_\Lambda}\left(Y,\ \beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})\right)=-\frac{1}{N}\left[\mathbf{Y}^T\left(\beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\mathbf{1}^T\mathbf{S}(\mathbf{X})\right)\right)\right] \tag{5.51}$$

Resolver el problema (5.50), como *para cualquier modelo aditivo generalizado, GAM, con la función de enlace logística, es equivalente a minimizar la log-verosimilitud binomial negativa, ajuste global del modelo, por un algoritmo iterativo de puntuación local en combinación con el algoritmo backfitting, (un algoritmo iterativo con estructura de Gauss-Seidel introducido por* FRIEDMAN y STUETZLE (1981) y perfeccionado por HASTIE y TIBSHIRANI (1990)). El método se conoce como **Regresión Logística Aditiva, ALR**.

Una forma alternativa de plantear el problema, consiste en especificar cada función de suavizado de dimensión infinita $S_r(x_r)$ como una expansión de una

colección arbitrariamente larga de funciones de base, controlando la complejidad a través de un regularizador $J(S)$. En tal caso

La función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=1}^{p}S_r(X_r)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y cada función de base, de suavizado de dimensión infinita, $S_r(X_r)$, a su vez por una expansión arbitrariamente larga de K_r funciones de base $h_{r,k}$,

$$\begin{aligned} S_r(X_r)&=\sum_{k=1}^{K_r}\beta_{r,k}h_{r,k}(X_r) \\ E\left[S_r(X_r)\right]&=0,\quad (r=1,\ldots,p) \end{aligned} \tag{5.52}$$

por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/X=x)\right)=S(X)=\beta_0+\sum_{r=1}^{p}\sum_{k=1}^{K_j}\beta_{r,k}h_{r,k}(X_r),\quad \beta_0\in\mathbb{R} \tag{5.53}$$

Dado que se controla la previsible complejidad del modelo (5.53) a través de un regularizador $J(S)$, nos encontramos ante el modelo de ***Regresión Logística Aditiva Regularizada,*** **RALR,** cuyo proceso de estimación es similar al proporcionado por la regresión logística aditiva**.**

Este método constituye parte de la semilla origen de los *Modelos Logísticos Lineales Híbridos, HLLM*, (por expansiones lineales de funciones de base), que proponemos en esta Tesis Doctoral, si bien lo hacemos con una alternativa al uso de las funciones de suavizado infinito dimensional, acorde con los requerimientos de Basilea II.

Tanto ALR como RALR proporcionan directamente la probabilidad de default, $\text{logit}\left(\hat{P}(Y=1/X=x)\right)=\hat{\beta}_0+\boldsymbol{1}^T\hat{\boldsymbol{S}}(\boldsymbol{X})$, y, por tanto, la función de calificación correspondiente, $\hat{S}(X)=\hat{\beta}_0+\boldsymbol{1}^T\hat{\boldsymbol{S}}(\boldsymbol{X})$, y el clasificador Bayes optimal de acreditados y solicitantes de crédito. Ambos modelos cuentan además con las atractivas propiedades de aceptar de forma flexible la no linealidad y permitir conocer los efectos marginales de las distintas variables de entrada asociadas con el riesgo, (las funciones $S_r(X_r)$ son análogas a los coeficientes del modelo lineal generalizado), de hecho *el estimador de* $S_r(X_r)$ *explica cómo cambia la*

respuesta junto con la correspondiente variable explicativa, lo que es imprescindible para conocer la relación entre el estado de default y las variables observadas sobre los acreditados y solicitantes de crédito.

Si se utilizan como funciones de base splines estaremos ante la ***Regresión Logística Aditiva Regularizada por Splines,*** **SRALR**. Un caso particular consiste en utilizar un conjunto maximal de nodos, lo que evita los inconvenientes de la selección tanto del número de nodos de los splines como su ubicación.

A pesar de las ventajas de las técnicas aditivas que hemos mencionado, los dos mayores problemas que presentan son la dificultad de interpretación de la funciones no paramétricas de dimensión infinita que los convierte ya de por sí en poco satisfactorios para el credit scoring, y que la hipótesis de aditividad es muy comprometida.

5.6.2.2 Arboles de Decisión, TREE. Algoritmo CART.

Una técnica para especificar y caracterizar la no linealidad de las variables explicativas en los modelos estadísticos, con un planteamiento mucho menos comprometido que la exigencia de aditividad de los efectos de las variables explicativas y mucho más simple en su desarrollo teórico, consiste en dividir el espacio de características en un conjunto de híper-rectángulos, que constituyen una Partición Recursiva, RP, *y entonces ajustar un simple modelo (igual a una constante) en cada uno, obteniendo de este modo una regla de predicción de la probabilidad de default.*

El desarrollo teórico de los *árboles de decisión* se debió a BREIMAN et al. (1984), y a QUINLAN (1986, 1987), y el origen del uso de sus principales ideas se encuentra en las ciencias sociales con los trabajos de SONQUIST y MORGAN (1964) y MORGAN y MESSENGER (1973). El pilar sobre el que se asienta el modelo de los árboles de decisión es la *teoría de grafos*, específicamente en los *grafos conexos sin ciclos*.

BREIMAN, et al. (1984) construyeron el algoritmo CART (*Classification And Regression Trees*), muy popular en credit scoring, como una regla para predecir el comportamiento de la variable respuesta desde los valores de las variables explicativas.

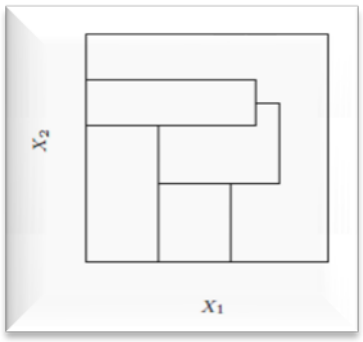
El objetivo de una partición recursiva es dividir el conjunto de observaciones en conjuntos disjuntos para incrementar la *homogeneidad* (en términos de clase) de los subconjuntos resultantes, o lo que es lo mismo, que éstos sean *más puros* que el conjunto originario. Cuando la partición se obtiene como resultado de evaluar una condición que

tiene dos únicas respuestas, por ejemplo, default o no default, se llama *Partición Recursiva Binaria*, BRP.

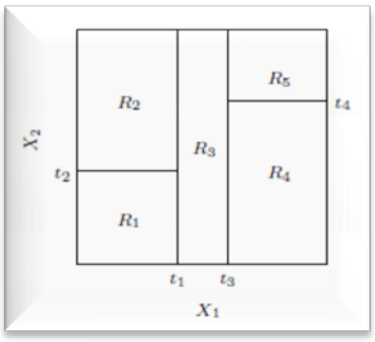
En el proceso de partición recursiva binaria se realiza una fractura de la variable explicativa, maximizando la reducción de varianza, lo que determina dos híper-rectángulos. El procedimiento de partición se aplica recurrentemente a cada uno de las regiones sucesivamente obtenidas. *La peculiaridad más sobresaliente de las particiones recursivas es que las fracturas se producen en paralelo a las proyecciones de cada coordenada de las variables explicativas en particular*.

En el panel a) de la figura 5.5 se muestra una partición del espacio de características por líneas paralelas a los ejes cartesianos. En cada elemento de la partición se puede modelizar Y con una constante diferente. Sin embargo, existe un problema, a pesar de que cada línea de partición tiene una simple descripción q, algunas de las regiones resultantes son de difícil interpretación.

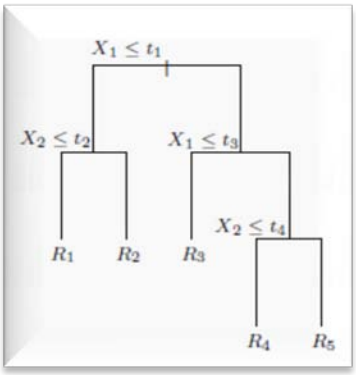
a) Partición Recursiva



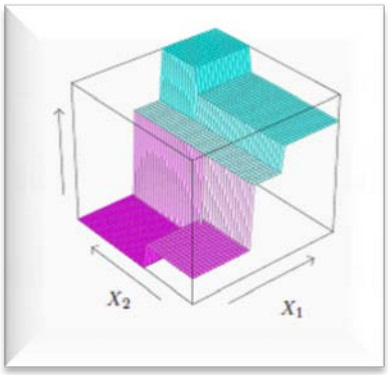
b) Partición Recursiva Binaria



c) Árbol de Decisión



d) Superficie de Regresión



Fuente: Hastie el al. 2009.

Figura 5.5.- Partición Recursiva, Partición Recursiva Binaria, Árbol de Decisión y Superficie de Regresión para las variables X_1 y X_2.

Para simplificar la cuestión, se restringe nuestra atención a particiones binarias recursivas tales como las del panel b) de la figura 5.5. Primero se divide el espacio en dos regiones y se modeliza la respuesta por medio de Y en cada región. Se elige la variable y el punto de división que consigue el mejor ajuste. Entonces una o las dos regiones son divididas en dos o más regiones cada vez más pequeñas, y este proceso continúa hasta que se aplica alguna regla de parada. El resultado de este proceso es una partición de q regiones homogéneas, $R_1, R_2, ..., R_q$, que entre sí son heterogéneas, como se muestra en el panel b) de la figura 5.5.

BREIMAN, et al. (1984) observaron que la estructura resultante de la partición recursiva admite una representación en un espacio métrico consistente en un *grafo conexo sin ciclos*, llamado *Árbol de Clasificación* . Es decir, del proceso de partición, a partir de la muestra de entrenamiento, resulta una partición recursiva del espacio de representación que se describe por una organización jerárquica cuya estructura formal viene dada por un árbol de clasificación como el del panel c) de la figura 5.5. Las observaciones que satisfacen la condición en cada disyunción se asignan a la rama izquierda y las que no a la rama derecha.

El pilar sobre el que se asientan los modelos de los árboles de decisión es la Teoría de Grafos, específicamente en los grafos denominados árboles. A continuación veremos algunos conceptos clave de la ***Estructura de Árbol*.**

Definición 5.3.- a) Un *grafo* es un par $G = (V, E)$, donde V es un conjunto finito no vacío (a cuyos elementos llamaremos vértices) y E es una familia finita de pares no ordenados de vértices de V (a cuyos elementos llamaremos *aristas o arcos*). Si $V = \{v_1, ..., v_n\}$, los elementos de E se representan de la forma $\{v_j, v_l\}$, $j \neq l$. Los elementos de una arista o arco se denominan *extremos* de dicha arista . Dos vértices v_j y v_l se dicen adyacentes sí $\{v_j, v_l\} \in E$.

b) Un *Grafo es Conexo* si para cada par de vértices v_j y v_l existe un camino de v_j a v_l .

c) Se llama *Ciclo* a un arco que sale de un vértice y entra en el mismo.

Definición 5.4.- *Se llama Árbol a un grafo* $G = (V, E)$ *conexo y sin ciclos.*

Teorema 5.1.- *Un grafo* $T=(V,A)$ *es un árbol* $\Leftrightarrow$ *dados* $v_j, v_l \in V$*, existe un único camino que los conecta.*

Un árbol de decisión puede interpretarse esencialmente como una serie de reglas compactadas representadas en forma de árbol. Los árboles constituyen una herramienta útil para describir estructuras que representan algún tipo de jerarquía.

Definición 5.5.- a) Se llama *árbol enraizado* a aquel en el que se destaca un vértice, llamado *vértice raíz*. El resto de vértices se disponen de modo que en el nivel k+1 se encuentran los vértices adyacentes a los de nivel k que no hubieran sido considerados antes.

b) El valor máximo k para el cual el nivel k es no vacio y el nivel k+1 es vacio se denomina *peso o altura del árbol.*

El conjunto completo de datos se sitúa al principio del árbol. Los nodos terminales o de salida del árbol corresponden a las regiones seleccionadas.

Como consecuencia de que la variable respuesta estado de default es binaria, nuestro interés radica en los á*rboles enraizados binarios,* BRT. Estos árboles nos permiten representar la estructura jerárquica que, respecto de las variables más discriminadoras del default, se obtiene como partición recursiva del conjunto de acreditados o solicitantes de crédito. De este modo, el árbol de decisión se constituye es un eficaz herramienta para asignar cada observación a uno de los dos grupos, tan homogéneos como sea posible, de default y no default.

En conclusión, podemos decir que una partición recursiva binaria de las variables de riesgo de crédito en una muestra de entrenamiento es un método estadístico de análisis multivariante que crea un árbol de decisión para *estimar la probabilidad de default y* clasificar correctamente a los acreditados o solicitantes de crédito basándose en variables de riesgo de crédito sobre dicha muestra. Desde la perspectiva de la clasificación, un árbol de decisión de respuesta binaria, particularizado a la probabilidad de default, es aquel en el que la probabilidad de default se expresa como *función constante a trozos, expansión lineal de funciones de base de las variables explicativas.*

La función de calificación de acreditados adopta la forma $S(X)=\sum_{r=1}^{q} h_r(X)$, la función de enlace viene dada por $g(\bullet)=1_{id}(\bullet)$ y las funciones de base por

$$\begin{aligned} & h_r(X):\mathbb{R}^p \to \{0,1\}, \quad r=1,..,q \\ & \qquad x \ \to h_r(X)=I_{[x\in R_r]}=\begin{cases} 1 & si\ x\in R_r \\ 0 & si\ x\notin R_r \end{cases} \\ & R_r \in \{R_1,...,R_q\},\ partición\ recursiva\ binaria\ del\ Rango(X) \\ & Rango(X)=\bigcup_{r=1}^{q} R_r,\ R_r\cap R_l=\varnothing,\ r\neq l. \end{aligned} \tag{5.54}$$

por tanto, el modelo presenta la siguiente estructura formal

$$P(Y=1/\ X=x))=\sum_{r=1}^{q}\beta_r I_{[x\in R_r]} \tag{5.55}$$

Las funciones $h_r(\bullet)$ definidas por (5.54) son funciones constantes, con valor 1 en la región R_r y cero en el resto.

El modelo (5.55) se estima a través de una función objetivo y un criterio de optimización muy peculiares debidos fundamentalmente a BREIMAN, (1984) y a QUINLAN (1986, 1987).

El objetivo de un árbol binario recursivo es la búsqueda de representaciones de los datos en espacios de dimensión reducida que se definen maximizando una función de utilidad determinada.

La parte más compleja en la construcción de un árbol de decisión consiste en realizar la partición recursiva binaria para lo que se utiliza usualmente el algoritmo CART, BREIMAN et al. (1984).

<u>Algoritmo para estructuras tipo árbol.</u>

1.- PARTICIÓN INICIAL

$$\begin{aligned} M &= 1 \\ P &= \{R\}=\mathbb{R}^p \end{aligned} \tag{5.56}$$

2.- REFINAMIENTO DE LA PARTICIÓN.

<u>Bucle</u> Para $r=1,...,q_{Máx}$ repetir los ciclos:

- Refinar la partición actual P, refinando la mejor región R en P en dos regiones $R_{izquierda}$ y $R_{derecha}$:

$$\begin{aligned} R_{izquierda} &= \mathbb{R}\times\mathbb{R}\times\ldots\times\left(-\infty, p_c\right]\times\mathbb{R}\times\ldots\times\mathbb{R} \\ R_{derecha} &= \mathbb{R}\times\mathbb{R}\times\ldots\times\left(p_c,\infty\right)\times\mathbb{R}\times\ldots\times\mathbb{R} \end{aligned} \tag{5.57}$$

  donde uno de los ejes se rompe en el punto de ruptura p_c perteneciente al conjunto finito de puntos medios entre los valores observados.

  La búsqueda del eje a romper, $j\in\{1,\ldots,p\}$, y del punto de ruptura p_c, $p_c\in\left\{\textit{puntos medios de los valores observados}\right\}$, se determina de forma que con el refinamiento *se maximice la reducción del máxima logverosimilitud negativa.*

- Se actualiza la partición

$$P^{nueva} = P^{anterior}\bigcup\left\{R_{izquierda}, R_{derecha}\right\} \tag{5.58}$$

- Como de grande ha de hacerse el árbol?. Claramente un árbol muy grande sobreajustará los datos, mientras que un árbol muy pequeño no capturará lo más importante de la estructura.

<u>Stop</u> Se para el refinamiento de la partición cuando algún tamaño mínimo de nodos sea rechazado.

La pregunta que podemos hacernos continuación es: *¿cuándo se debe declarar un nodo como terminal, o por el contrario continuar la división?*.

Para contestar a la cuestión anterior es necesario definir el concepto de *poder discriminante de un atributo* X_j que está estrechamente relacionado con la *heterogeneidad,* o *impureza de la distribución* Y/X_j. Es obvio que buena función de puntuación crediticia debe separar los default y no default en clases preferiblemente heterogéneas. La impureza debe ser grande cuando todos los sucesos son igualmente verosímiles y pequeña cuando sólo ocurre un suceso. Las referencia más generales, sobre el concepto de impureza, se encuentran en BREIMAN et al. (1984) y en FAHRMEIR et al. (1996) y las más actuales y en detalle en CLAVERO (2008).

De entre las diferentes medidas de la impureza del nodo r, en el árbol T $I_r(T)$, destacan las siguientes:

Error de clasificación incorrecta: $I_r(T) = ME_r(T) = 1 - máx\{p_r, 1-p_r\}$ (5.59)

Índice de Gini: $I_r(T) = GI_r(T) = \sum_{k=0}^{1} p_{rk}(1-p_{rk}) = 2p_r(1-p_r)$ (5.60)

Entropía Cruzada: $I_r(T) = CE_r(T) = -p_r \log(p_r) - (1-p_r)\log(1-p_r)$ (5.61)

donde $p_m = p_{m1}$ *y* $(1-p_m) = p_{m0}$, es decir, las proporciones de observaciones de las clases default y no default que se encuentran en el nodo r.

La cuestión puede ahora ser respondida, pues de forma intuitiva se podría utilizar el siguiente argumento "*detener el proceso de división cuando la mejor partición posible del nodo* R_r *(la que proporciona el máximo decrecimiento en impureza) produce un decrecimiento en impureza inferior al umbral* $\beta > 0$, *entonces estaremos ante un nodo terminal*".

Pero este criterio, aunque simple, no produce resultados satisfactorios en la práctica porque detiene el proceso de división prematuramente en algunos nodos y demasiado tarde en otros, siendo muy difícil hacer que el crecimiento se pare uniformemente en todas las ramas del árbol de forma óptima (BREIMAN et al. 1984).

El procedimiento usual resulta algo más complejo que establecer un umbral y detener el crecimiento: *se basa en un "criterio de poda", es decir, se trata de construir un árbol muy grande, incluso hasta que todos los nodos sean puros, y podar hacia la raíz de manera adecuada, (los subárboles que producen pequeños beneficios de bondad).*

Si no podásemos el árbol este se ajustaría mucho a los datos de entrenamiento disminuyendo su capacidad de generalización. Por lo que la poda generalmente mejora la capacidad predictiva del árbol (MINGERS, 1989a, 1989b; ESPÓSITO et al., 1997). En el proceso de poda se eliminan nodos de la zona terminal del árbol, donde las preguntas se han generado con menos elementos muestrales y por tanto se tiene menos certeza de su validez.

El tamaño del árbol es un parámetro de afinamiento que controla la complejidad del modelo, y el tamaño óptimo del árbol debe ser elegido adaptativamente desde los datos. La estrategia preferida es hacer crecer un gran árbol T_0, parando el proceso de subdivisión solo cuando algún tamaño mínimo de nodos sea rechazado, entonces este gran árbol es podado usando el *podado de coste-complejidad*.

Poda del árbol.

La poda consiste en que una vez obtenido $T_{Máx}$ *se van borrando hacia atrás regiones o subconjuntos muestrales hasta un tamaño razonable*, usualmente determinado vía *validación cruzada*, es decir, comenzando en $T_{Máx}$ se obtiene una secuencia decreciente y anidada de árboles $T_{máx} \succ T_2 \succ ... \succ_{\{r_1\}}$, de manera que el árbol $\{r_1\}$ consta de un único nodo.

BREIMAN et a. (1984) inciden en la importancia de la poda frente al de selección de particiones. Su argumento es que resulta más eficiente podar un árbol que detener su crecimiento. La poda permite que en el subárbol una rama de un de un nodo permanezca y la otra desaparezca, mientras que detener el crecimiento poda *ambas* ramas simultáneamente.

Definición 5.6.- Si T' se obtiene a partir de T por poda, T' *es un subárbol podado de* T y se escribe $T' \prec T$. Así, la secuencia decreciente de árboles podados y anidados será la siguiente:

$$T_{máx} \succ T_2 \succ ... \succ_{\{r_1\}} \tag{5.62}$$

Uno de estos árboles será el que se seleccione finalmente de acuerdo con la siguiente regla:

> *Para realizar esta selección se asocia una medida de error a cada árbol de la secuencia y se escoge aquel que tenga asociado el menor error.*

Se espera que árboles podados (más simples) produzcan mejores resultados que los obtenidos con árboles más grandes (más complejos) al clasificar observaciones independientes, esto es, los árboles podados tendrán más capacidad de generalización al no estar tan ajustados al conjunto de aprendizaje (el problema del *sobre aprendizaje*).

El procedimiento de poda más usual es el basado en el *criterio de mínimo coste-complejidad,* BREIMAN et al. (1984), *que* tiene en cuenta, además del error, la complejidad del árbol. La idea que subyace detrás de este criterio es que en general para árboles con un error similar en τ, tendrá mayor capacidad de generalización aquel con menor complejidad, y para árboles con complejidad similar, tendrá mayor capacidad de generalización aquél con un error menor en τ. Por tanto, **el objetivo es llegar a un compromiso entre error y complejidad**.

Una vez construido el árbol de clasificación, sólo resta decidir que procedimiento se debe seguir para hacer predicciones a cerca de las observaciones en los nodos terminales, o en términos de clasificación, asignar las clases pronosticadas a los nodos terminales. Usualmente se sigue *la regla de la pluralidad*, en la cual se pronostica la clase con el mayor número de observaciones en un nodo terminal. Es decir, si la clase 1 tiene más observaciones que la clase 0, se pronostica default, (este puede no ser el caso si las clasificaciones incorrectas no se ponderan igualmente, véase RIPLEY (1996)).

Definición 5.7.- Dado una muestra de entrenamiento $\tau = \{x_i, y_i\}_{i=1}^{N} \subset (X,Y)^N$, se define el *estimador de la probabilidad condicional* $P(Y = k / X = x \in R_m)$ como la proporción de observaciones de la clase k, $k = 0,1$, en el nodo r, que representa al subconjunto R_r con N_r observaciones, es decir,

$$\hat{p}_{rk} = \frac{1}{N_r} \sum_{i=1}^{N} I_{[y_i = k]} y_i I_{[x_i \in R_r]}, \quad k = 0,1 \tag{5.63}$$

En otras palabras, la proporción de observaciones de la clase k en el nodo r es un estimador de la probabilidad de que la respuesta tome el valor k bajo el supuesto de que la observación x pertenece a R_r.

Sea $T_{\hat{\alpha}}$ el árbol óptimo con $q = |T_{\hat{\alpha}}|$ nodos terminales indexados por $r = 1,...,q$ con r representando a la región R_r, $\bigcup_{r=1}^{q} R_r = \mathbb{R}^P$, $R_r \cap R_l = \varnothing \ \ \forall r \neq l$ y $N_r = \sum_{i=1}^{N} I_{[x_i \in R_r]}$ el número de observaciones en la región R_r, Para definir el estimador de la probabilidad de default nos basaremos, por un lado, en el criterio:

> *Se clasifican las observaciones del nodo r en la clase k, $(k = 0,1)$, tal que $k(r) = \underset{k=0,1}{máx}\ \hat{p}_{rk}$, es decir, en la clase a la que pertenecen la mayoría de las observaciones de la región R_r.*

Y, por otro lado, en el hecho de que q regiones forman una partición de $\mathbb{R}^p$ de regiones disjuntas, todo $x \in \mathbb{R}^p$ pertenece a una única región o nodo terminal R_m.

Definición 5.8.- Se define el *estimador de la probabilidad de default* $P(Y = 1 / X = x)$ como

$$\hat{P}(Y = 1 / X = x)) = \sum_{m=1}^{q} \hat{\beta}_m I_{[x \in R_m]}$$
$$\{R_1,...,R_q\} \ \textit{es una partición de} \ \mathbb{R}^p \tag{5.64}$$

Uno de los mayores problemas que presentan los árboles de decisión es su alta varianza. Con frecuencia un pequeño cambio en los datos puede devenir en una muy diferente secuencia de fracturas, lo que conlleva muchas veces a interpretaciones muy precarias. La mayor razón para esta inestabilidad procede de la naturaleza jerárquica del proceso: el efecto de un error en la división al comienzo de la construcción del árbol se propaga en cascada hacia todas las divisiones que le siguen. Podría aliviarse este problema en cierto grado usando un criterio de división más estable, pero la inestabilidad inherente no se elimina. Este es el precio a pagar por estimar la estructura subyacente en nuestros datos a través de una herramienta muy simple y de gran claridad interpretativa. Para reducir esta varianza se utiliza el método Bagging, que promedia varios árboles.

Los árboles de decisión, aunque son conceptualmente simples, son muy atractivos por cuanto están dotados de una gran facilidad de interpretación, lo que constituye su mejor baza de cara a los requerimientos de Basilea II. Entre sus cualidades destaca el hecho de que las ramas del árbol simulan bastante bien el proceso humano para la toma de decisiones, a la vez que definen directamente las reglas de asignación por lo que sus resultados son operativos inmediatamente. Una de sus mayores fortalezas es que detectan de forma automática estructuras complejas entre variables. Por otra parte, junto al hecho de que minimizan el pre análisis de los datos, puesto que tienen una gran capacidad para trabajar con un nivel de ruido relativamente alto y con datos faltantes, son computacionalmente muy eficientes.

Pero en la otra cara de la moneda las debilidades no son poca pues, aparte de su alta varianza, la probabilidad estimada es constante a trozos, como puede observarse en (5.55), lo que no es la forma usual de la función subyacente real. La aproximación por saltos a la función de respuesta también implica que la estimación de la probabilidad no se encuentra entre las mejores, (el error de predicción puede ser mayor que en otros modelos más flexibles).

Por tanto, a pesar de que la facilidad de interpretación de los árboles de clasificación los convierte en métodos muy usuales de estimación de la probabilidad de default y de la clasificación de acreditados y solicitantes de crédito, y que están relativamente bien vistos por las Instituciones Reguladoras de los Sistemas Financieros Internacionales, en particular del Banco de España, se ha de acompañar su uso de una gran dosis de cautela.

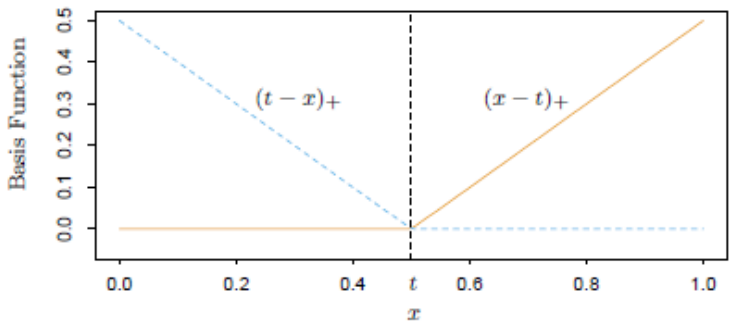
5.6.2.3 Splines de Regresión Adaptativos Multivariantes, MARS.

Otra forma de introducir la no linealidad de las variables explicativas en los modelos estadísticos de probabilidad de default viene de la mano de una técnica a caballo entre las particiones recursivas y los modelos logísticos aditivos, debida a STONE et al. (1997), que consiste en un híbrido llamado **BIMARS, *Modelos Logísticos de Splines de Regresión Adaptativos Multivariantes***, que incorporan la transformación logística para respuesta binaria a la estructura de los ***Modelos de Splines de Regresión Adaptativos Multivariantes*, MARS**, que había sido introducido por FRIEDMAN (1991), inspirado en la técnica de partición recursiva de CART y en los modelos GAM.

Las técnicas de construcción y estimación de los modelos de regresión MARS son técnicas semiparamétricas adaptativas, generalización de los modelos lineales paso a paso, que modelizan de forma automática relaciones no lineales e interacciones entre la variable respuesta y las variables explicativas. Estas aproximaciones suavizan el modelo en los nudos y uno de los objetivos de los algoritmos paso a paso hacia adelante es seleccionar estos nudos. El resultado es un modelo continuo muy flexible obtenido por ajuste de regresiones lineales a trozos, con derivadas continuas, muy eficaz para encontrar, a través de transformaciones óptimas de las variables y sus interacciones, la compleja estructura que frecuentemente se esconde en los datos.

MARS emplea una selección de pares reflejados de funciones bisagra, lineales a trozos, o sus productos tensoriales, definidos en la forma siguiente:

$$\boldsymbol{B} = \left\{ \left(X_j - t\right)_+ ,\ \left(t - X_j\right)_+ ;\ \ t \in \left\{x_{1j}, x_{2j}, ..., x_{Nj}\right\},\ \ j = 1,2,...,p \right\} \ \textit{donde}$$
$$\left(X_j - t\right)_+ = \begin{cases} X_j - t, & \textit{si } X_j > t \\ 0 & \textit{en otro caso} \end{cases} \quad y \quad \left(t - X_j\right)_+ = \begin{cases} t - X_j, & \textit{si } X_j < t \\ 0 & \textit{en otro caso} \end{cases} \tag{5.65}$$



Fuente: Hastie et al. (2009)

Figura 5.6.- Representación de las funciones bisagra $\left(x-0.5\right)_+$ y $\left(0.5-x\right)_+$

Cada función bisagra del tipo (5.65) es lineal a trozos con un nudo en $X_j = t$, es decir, son splines lineales a trozos y X_j, $(j = 1,..., p)$, es la j-ésima componente del vector $X \in \mathbb{R}^p$. Se dice que las dos funciones $(X_j - t)_+$ y $(t - X_j)_+$ forman un *par reflejado*. La idea es formar pares reflejados para cada entrada X_j con nudos en cada valor observado x_{ij} de tal entrada. La colección de pares reflejados es entonces un conjunto $\boldsymbol{B}$ de funciones de base (5.65) de las que, si todas las entradas son distintas, existe un total de $2Np$. Nótese que cada función básica depende solo de un X_j, por ejemplo, $h(X) = (X_j - t)_+$, a pesar de que es considerada como una función de $X \in \mathbb{R}^p$, es decir, sólo la j-ésima componente de X es relevante. En el caso univariante $\boldsymbol{B}$ extiende el espacio de splines lineales.

La función de calificación de acreditados en esta técnica se representa como una combinación lineal de funciones de base $h_r(\bullet)$ que son elementos o productos de elementos de $\boldsymbol{B}$.

La función de calificación de acreditados adopta la forma $S(X) = \sum_{r=1}^{q} \beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet) = Y$ y las funciones de base vienen dadas en la forma siguiente

$$h_r(\bullet): \mathbb{R}^P \to \mathbb{R}, \quad r = 1,..,q$$

$$X \to h_r(X) = \begin{cases} b(X) \in \boldsymbol{B} \\ o \\ \bigotimes_{k=1}^{K_r} b_k; \quad b_k \in \boldsymbol{B} \end{cases} \tag{5.66}$$

Siendo $\boldsymbol{B}$ el conjunto de funciones bisagra definido en (5.65).
Por tanto, el modelo presenta la siguiente estructura formal

$$Y = S(X) = \sum_{r=1}^{q} \beta_r h_r(X) \tag{5.67}$$

Dada una elección para las $h_r(\bullet)$, *los coeficientes* β_r *son estimados por la suma de cuadrados residual, es decir, por la regresión lineal estándar*. El arte de esta técnica,

está en la construcción del modelo, en decir, en el proceso de construcción de las funciones $h_r(\bullet)$.

La estrategia seguida por MARS para la construcción del modelo óptimo es una regresión lineal en dos pasos (un paso hacia adelante y uno hacia atrás), estrategia que coincide con la que se sigue en los árboles de partición recursiva como CART, pero en vez de usar las entradas originales, se usan funciones de base del tipo (5.65) y los productos tensoriales de estas. En el primer paso, paso hacia adelante, MARS construye un gran número de funciones de base que se seleccionan para sobreajustar los datos inicialmente, donde se permiten variables continuas, categóricas u ordinales y pueden interactuar unas con otras o restringirse sólo a componentes aditivas. En el segundo paso, paso hacia atrás, las funciones de base son removidas en el orden de mínima contribución usando el criterio de validación cruzada generalizada GCV.

Paso hacia adelante

MARS comienza con un modelo $\boldsymbol{M}_0$ que consiste sólo en el término intercepto, (la media de los valores respuesta), y todas las funciones del conjunto $\boldsymbol{B}$, (5.65), son candidatas. MARS entonces añade repetidamente pares de funciones de base al modelo $\boldsymbol{M}$. En cada paso encuentra el par de funciones de base que consigue la máxima reducción en el error suma de cuadrados residuales. Las dos funciones de base en el par son idénticas excepto que para cada función se usa un lado diferente de una función bisagra reflejada. Cada nueva función básica consiste en un término que ya está en el modelo, (que acaso puede ser el intercepto, es decir, una constante), multiplicado por una nueva función bisagra. Para añadir nuevas funciones de base, MARS debe buscar sobre todas las combinaciones de todas las funciones de base siguientes:

1) Términos existentes (llamados *términos padre en este contexto).*
2) Todas las variables, (para seleccionar una de las nuevas funciones de base).
3) Todos los valores de cada variable (para los nudos de las nuevas funciones bisagra, dado que una función bisagra se define por una variable y un nudo).

Se añade al modelo $\boldsymbol{M}$ el término de la forma

$$\hat{\beta}_{M+1} h_l(X).(X_t - t)_+ + \hat{\beta}_{M+2} h_l(X).(t - X_j)_+, \quad h_l \in M \tag{5.68}$$

que produce el mayor decrecimiento en el error de entrenamiento. Aquí $\hat{\beta}_{M+1}$ y $\hat{\beta}_{M+2}$ son coeficientes estimados por mínimos cuadrados, con todos los demás coeficientes en el

modelo. Entonces los productos ganadores son añadidos al modelo y el proceso de añadir términos continúa hasta que los cambios en el error residual son muy pequeños o hasta que se alcanza el máximo número de términos preestablecido por el usuario. Al final de este proceso se tiene un modelo, generalmente muy grande, de la forma

$$S_{Mars}(x) = \beta_0 + \sum_{m=1}^{q} \beta_m h_m(x) \tag{5.69}$$

Un aspecto clave de MARS es que, a causa de la naturaleza de las funciones bisagra, la búsqueda puede hacerse relativamente rápida usando una técnica rápida de actualización de mínimos cuadrados. Usualmente el procedimiento hacia delante que utiliza MARS sobreajusta los datos, lo implica que, si bien el modelo presenta un buen ajuste a los datos de entrenamiento, este puede no generalizar bien a nuevos datos. Se hace necesario, por tanto, un procedimiento de borrado hacia atrás, "poda" en términos de metodología CART, para limitar la complejidad del modelo reduciendo el número de sus funciones base.

El paso hacia atrás.

El sistema de podado de MARS remueve algunos términos uno a uno, borrando aquellas funciones de base asociadas con el menor crecimiento de la bondad de ajuste (mínimos cuadrados), hasta encontrar el mejor submodelo. A continuación, por V*alidación Cruzada Generalizada*, GCV, se computa una función del error de los mínimos cuadrados (inversa de la bondad de ajuste). Los submodelos obtenidos por podado de términos se comparan usando el criterio GCV.

El error de validación, GCV, es una medida de la bondad de ajuste que tiene en cuenta no solo el error residual SSR sino también la complejidad del modelo. Es decir, la utilización de GCV es una forma de regularización que equilibra la bondad de ajuste frente a la complejidad del modelo. La *suma de cuadrados residuales* (RSS) sobre los datos de entrenamiento es inadecuada para comparar modelos, a causa de que los RSS siempre crecen con el podado de términos MARS. En otras palabras, si se usa el RSS sobre los datos de entrenamiento para comparar modelos, el paso hacia adelante elegiría presumiblemente el modelo más grande.

La formulación del criterio GCV viene dada por

$$GCV = \frac{RSS}{N\left(1 - \frac{\text{Número efectivo de Parametros}}{N}\right)^2} \tag{5.70}$$

donde N es el número de observaciones que integran la muestra de entrenamiento.

El número efectivo de parámetros se define en el contexto MARS como

$$\text{Nº efectivo de Parametros} = \text{Nº de Términos Mars} + \text{Penalización}\,\frac{\left(\text{Nº de Términos Mars} - 1\right)}{2} \tag{5.71}$$

donde la penalización suele ser 2 o 3 (el software MARS permite a los usuarios fijar previamente la penalización). Nótese que $\frac{\left(\text{Nº de Términos Mars} - 1\right)}{2}$ es el número de nudos de función bisagra, por tanto la formula penaliza la adición de nudos. De este modo la formula incrementa el RSS de entrenamiento para tomar en cuenta la flexibilidad del modelo.

La Validación Cruzada Generalizada, llamada así porque esta fórmula se usa para aproximar el presunto error determinado por validación dejando uno fuera, fue introducida por CRAVEN Y WAHBA (1979) y extendida a MARS por FRIEDMAN (1991).

El paso hacia atrás tiene una ventaja sobre el paso hacia adelante: en cualquier paso se puede elegir cualquier término para borrar, mientras en el paso hacia adelante en cada paso solo se puede ver el último par de términos. El paso hacia adelante añade pares de términos reflejados, pero el paso hacia atrás usualmente descarta un lado del par y, por tanto, estos términos no se ven frecuentemente a pares en el modelo final.

Las restricciones que el usuario puede imponer sobre el modelo pueden consistir, por ejemplo, en el número máximo de términos a considerar en el paso hacia delante, restricción a la que ya nos hemos referido. Otra restricción en el paso hacia adelante puede ser el máximo grado de interacción permitido. Usualmente se permiten los grados uno o dos de interacción, pero puede ser mayor si los datos lo avalan.

También podrían plantearse otras restricciones en el paso hacia delante, como, por ejemplo, permitir interacciones solo para ciertas variables. Tales restricciones pueden tener sentido a causa del conocimiento del proceso que ha generado los datos.

Una ventaja clave de las funciones de la figura 5.6 es su habilidad para operar localmente, dado que son cero sobre parte de su rango. Cuando se multiplican unas con otras el resultado es distinto de cero sólo en pequeñas partes del espacio de características donde ambas funciones son distintas de cero. Como resultado la superficie de regresión se construye muy parsimoniosamente, usando componentes distintas de cero localmente – sólo donde esto es necesario. *Una segunda ventaja es la facilidad de computación de la funciones de base lineales a trozos* (HASTIE et al. 2009).

El modelo MARS planteado según (5.67), es un método bastante eficaz, pero con una debilidad importante para nuestros objetivos relacionados con Basilea II, no usa la información de que la variable respuesta estado de default es binaria, por lo que no restringe los valores ajustados entre 0 y 1. Para estimar las probabilidades de default hemos de considerar un método MARS Generalizado, en concreto la **Regresión Logística BIMARS**, que STONE et al. (1997) desarrollaron para una variable respuesta múltiple, POLYMARS, del cual el caso de respuesta binaria es un caso particular.

Con la idea de conseguir un método automático para seleccionar funciones de base del conjunto infinito $\boldsymbol{B}$ de funciones bisagra lineales a trozos adaptado a la pérdida logística, Stone y sus colaboradores desarrollaron un modelo paso a paso de forma progresiva parecido a MARS, pero en cada paso usan una aproximación cuadrática a la *verosimilitud binomial negativa* para buscar el siguiente par de funciones de base y *test de entrada y salida de funciones de base que utilizan la pérdida logística.*

Las restricciones impuestas sobre las funciones base a considerar son:

R1.- Las funciones lineales en una de las variables explicativas se permiten siempre.

R2.- Se permiten funciones de base de la forma

$$\left(x_j - t\right)_+, \quad t \in \left\{x_{1j}, x_{2j}, \ldots, x_{Nj}\right\}, \quad j = 1, 2, \ldots, p$$

solo cuando las funciones lineales correspondientes ya estén incluidas en el modelo.

R3.- Se permiten los productos tensoriales de funciones que abarcan dos variables explicativas diferentes que ya están en el modelo, excepto si tal producto tensorial conlleva un nudo en una o en ambas variables explicativas. Las funciones base correspondientes con términos lineales deben de estar ya en el modelo. Es decir,

para que se permita en el modelo un término de la forma $\left(x_j - t_k\right)_+ .\left(x_l - t_h\right)_+$, los términos $x_j x_l$, $x_j .\left(x_l - t_h\right)_+$ y $x_l .\left(x_j - t_k\right)_+$ han de estar ya dentro de él.

Al igual que en MARS se usa la adicción paso a paso para construir el modelo completo y el podado paso a paso para obtener el modelo definitivo, pero BIMARS a diferencia de MARS, utiliza el estadístico de RAO para la adicción y el estadístico de Wald para la selección de las funciones de base definitivas. La selección del modelo en BIMARS se hará usando AIC, un Conjunto de Test Independientes o la Validación Cruzada Generalizada. Una vez encontradas las funciones de base idóneas, el modelo agrandado se ajusta por máxima verosimilitud y el proceso se repite.

Basándose en datos muestrales $\tau = \left\{\left(x_i, y_i\right) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i=1}^{N} \subset \left(X \times Y\right)^N$, los coeficientes β_r, a diferencia de MARS, se estiman por máxima verosimilitud.

La estructura formal del modelo de STONE et al. (1997) viene caracterizada por:

La función de calificación de acreditados adopta la forma $S\left(X\right) = \sum_{r=1}^{q} \beta_r h_r\left(X\right)$, la función de enlace viene dada por $g\left(\bullet\right) = \text{logit}\left(\bullet\right)$ y las funciones de base vienen dadas como en (5.66) $h_r\left(X\right) = b\left(X\right) \in \boldsymbol{B}$, o bien, $h_r\left(X\right) = \bigotimes_{k=1}^{K_r} b_k$; $b_k \in \boldsymbol{B}$, siendo $\boldsymbol{B}$ el conjunto de funciones bisagra definido en (5.65), con la restricciones R1, R2 y R3.

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P\left(Y = 1 / X = x\right)\right) = S\left(X\right) = \beta_0 + \sum_{r=0}^{q} \beta_r h_r\left(X\right) \qquad (5.72)$$

Tampoco el método BIMARS estima satisfactoriamente la probabilidad de default, por cuanto se configura la estructura formal del modelo suponiendo que todas las variables pueden ser especificadas por splines bisagra definidos según (5.65) (5.66) y las restricciones R1, R2 y R3. No es necesario que nos extendamos en explicar que esa situación sería absolutamente excepcional. Sin embargo, es posible que la no linealidad de alguna de las variables pueda especificarse por una expansión lineal de funciones de base pertenecientes a $\boldsymbol{B}$.

Esta posibilidad junto con el hecho de que los modelos lineales generalizados pueden incorporarse a los modelos MARS aplicando una función de enlace después de que el modelo MARS se ha construido, nos sugiere como introducir funciones bisagra MARS en los modelos de probabilidad generalizada por expansiones lineales de funciones de base, idea que aprovecharemos en nuestra propuesta de modelos HLLM .

5.6.2.4 Proyecciones sobre Direcciones de Interés, PPR.

Una técnica para explorar ciertos aspectos de la estructura del modelo, en particular la no linealidad, con antecedentes en el Análisis de Componentes Principales, ACP*, consiste en proyectar los puntos sobre ciertas direcciones de interés,* PPR.

Una limitación obvia de las particiones recursivas, utilizadas en los árboles de decisión, es que las fracturas de las variables explicativas se producen solamente en paralelo a las proyecciones de cada coordenada en particular. Las funciones constantes a trozos consideradas para predecir el default en los árboles de decisión en un sistema de coordenadas diferente al anterior, por ejemplo girado y trasladado, no podrían aproximarse bien.

La idea de aproximar funciones de alta dimensión por funciones más simples a través de proyecciones se remonta al menos a KRUSKAL (1969). FRIEDMAN y TUKEY (1974) aplicaron esta idea de la "búsqueda de proyecciones de interés" en el análisis de datos de física de partículas, introduciendo así la formulación inicial de este método. La especialización de la técnica en regresión se debe a FRIEDMAN y STUETZLE (1981) y desde entonces la técnica se conoce como *Regresión de Búsqueda de la Proyección*, PPR, a estos autores se debe también el *algoritmo backfitting* básico. Una revisión de las técnicas de Búsqueda de la Proyección, con una variedad de ejemplos, también en la estimación de densidades, se encuentra en JONES y SIBSON (1987). Los aspectos teóricos de PPR se pueden encontrar en HALL (1989).

Un modelo PPR, al igual que CART y BIMARS consiste en sumas de elementos no lineales transformadas de modelos lineales. El objetivo de PPR es la búsqueda de representaciones de los datos en espacios de dimensión reducida que se definen maximizando una función de utilidad determinada. La idea básica puede ilustrarse con el ejemplo siguiente:

Una simple función como $S(X)=S(X_1,X_2)=X_1X_2-\sqrt{2}$ no puede ser representada por la regresión por partición recursiva, sin embargo, veamos que puede ser representada como una suma de funciones que operan sobre proyecciones, llamadas *funciones sierra:*

$$\begin{aligned} S(X)&=X_1X_2-\sqrt{2}=-\sqrt{2}+\frac{1}{4}(X_1+X_2)^2-\frac{1}{4}(X_1-X_2)^2 \\ &=\left(\left(\left(\frac{1}{2},\frac{1}{2}\right)^T(X_1,X_2)\right)^2-\frac{\sqrt{2}}{2\sqrt{}}\right)+\left(-\left(\left(\frac{1}{2},-\frac{1}{2}\right)^T(X_1,X_2)\right)^2-\frac{\sqrt{2}}{2\sqrt{}}\right) \end{aligned} \tag{5.73}$$

Usando terminología de funciones de base,

$$h_1\left(\alpha_1^TX\right)=\left(\frac{1}{2}X_1+\frac{1}{2}X_2\right)^2-\frac{\sqrt{2}}{2\sqrt{}},\quad h_2\left(\alpha_2^TX\right)=\left(\frac{1}{2}X_1-\frac{1}{2}X_2\right)^2-\frac{\sqrt{2}}{2\sqrt{}},$$
$$\alpha_1=\left(\frac{1}{2},\frac{1}{2}\right)^T,\quad \alpha_2=\left(\frac{1}{2},-\frac{1}{2}\right)^T$$

(5.73) se puede expresar en la forma:

$$S(X)=h_1\left(\alpha_1^TX\right)+h_2\left(\alpha_2^TX\right)=\sum_{r=1}^{2}h_r\left(\sum_{j=1}^{2}\alpha_{jr}X_j\right) \tag{5.74}$$

donde h_1 y h_2 son dos funciones que operan sobre las proyecciones:

$$\alpha_1^TX=\left(\frac{1}{2},\frac{1}{2}\right)^T\begin{pmatrix}X_1\\X_2\end{pmatrix}\quad\text{y}\quad\alpha_2^TX=\left(\frac{1}{2},-\frac{1}{2}\right)^T\begin{pmatrix}X_1\\X_2\end{pmatrix} \tag{5.75}$$

Obsérvese que las proyecciones se realizan en las direcciones definidas por los vectores

$$\alpha_1=\left(\frac{1}{2},\frac{1}{2}\right)^T\text{ y }\alpha_2=\left(\frac{1}{2},-\frac{1}{2}\right)^T$$

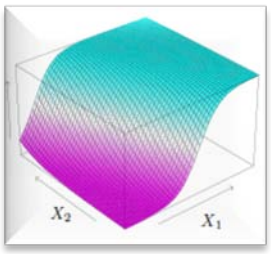
lo cual motiva una generalización de la regresión por partición recursiva, RPR, que consiste en aproximar la superficie de regresión por una suma de "*funciones caballete o sierra univariantes*" $h_r(\bullet)$ de proyecciones α_r^TX, determinadas empíricamente, en vez de usar funciones constantes de proyecciones a lo largo del eje de coordenadas.

Las *funciones sierra univariantes* se definen entonces como funciones reales de variable real siguientes:

$$\begin{aligned} h_r(\bullet):\mathbb{R}&\to\mathbb{R},\ r=1,\ldots,q \\ \alpha_r^TX&\mapsto h_r(\alpha_r^TX) \end{aligned} \tag{5.76}$$

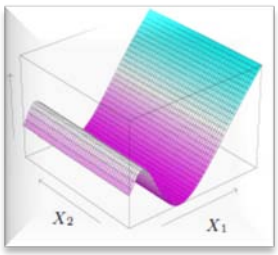
Obsérvese que las funciones $h_r(\bullet)$ sólo varían en la dirección de α_r, por esa razón se llaman también *funciones caballete* en $\mathbb{R}$, además pueden ser concebidas como

generalizaciones de funciones lineales, es decir, pueden ser constantes sobre hiperplanos, como se ve en la figura 5.7.



$$h_1 = \frac{1}{1+\exp\left(-5\left(\alpha_1^T x - \frac{1}{2}\right)\right)}$$

$$\alpha_1^T = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$



$$h_2 = \left(\alpha_2^T x + 0.1\right) sen\left(\frac{1}{3\left(\alpha_2^T x + 0.1\right)}\right)$$

$$\alpha_2^T = (1,0)$$

Fuente: Hastie et al. (2009)

Figura 5.7.- Gráficas de perspectiva de dos funciones sierra.

Las funciones $h_r(\bullet): \mathbb{R} \to \mathbb{R}$ *son totalmente no paramétricas, (suavizadores arbitrarios), no están especificadas y se estiman a lo largo de las direcciones* α_r *usando algún método flexible de alisado.* Cada combinación lineal $\alpha_r^T X = \sum_{j=1}^{p} \alpha_{jr} X_j$ es una proyección lineal de X en la dirección del vector $\alpha_r = \left(\alpha_{1r}, ..., \alpha_{pr}\right)^T$. Se busca α_r tal que el modelo ajuste bien, de ahí que el modelo se llame búsqueda de la proyección.

De acuerdo con lo anterior, el modelo PPR viene caracterizado por:

La función de calificación de acreditados adopta la forma $S(X) = \sum_{r=1}^{q} h_r(\alpha_r^T X)$, la función de enlace viene dada por $g(\bullet) = Y$ y las funciones de base viene dadas en la forma (5.76).

Por tanto, el modelo presenta la siguiente estructura formal

$$Y = \sum_{r=1}^{q} h_r\left(\sum_{j=1}^{p} \alpha_{jr} X_j\right) = \sum_{r=1}^{q} h_r\left(\alpha_r^T X\right) = 1^T \boldsymbol{H}\left(\boldsymbol{\alpha}^T X\right) \qquad (5.77)$$

donde $E\left[h_r\left(\alpha_r^T X\right)\right] = 0, \; r = 1, ..., q$.

El modelo (5.77) es un modelo aditivo, pero no en las variables explicativas X, sino en las características $\alpha_r^T X$. La representación puede no ser única, DIACONIS y SHAHSHAHANI (1984).

FRIEDMAN y STUETZLE (1981) propusieron el primer algoritmo para el modelo (5.77), algoritmo backfitting básico, que permite construir las proyecciones de PPR iterativamente por ciclos sobre los α_r y los h_r.

El modelo (5.77) ajusta bien y es muy general, puesto que la operación de sumas de funciones sierra genera una clase de modelos sorprendentemente grande. Además PPR tiene la propiedad de que si q toma valores arbitrariamente grandes, para elecciones apropiadas de h_r, el modelo puede aproximar cualquier función continua en $\mathbb{R}^p$ bastante bien, hasta el punto de que *llegó a pensarse que esta clase de modelos constituye una clase de estimadores universales,* (DIACONIS y SHAHSHAHANI, 1984). Pero esta generalidad conlleva un precio: "*la interpretación de los modelos PPR ajustados es normalmente difícil a causa de que cada variable explicativa entra en el modelo de forma compleja y multifacética*". Como resultado, el modelo PPR es más usado para labores de predicción que para conseguir un modelo comprensible de los datos.

Los modelos PPR son modelos aditivos y, por tanto, son generalizaciones de los modelos lineales que combinan la flexibilidad del modelado no paramétrico de entradas multidimensionales con la precisión estadística típica de una variable explicativa unidimensional, por lo que, la maldición de la dimensionalidad no está presente en estos modelos.

A pesar de que el procedimiento trabaja notablemente bien, (RIPLEY, 1996), ni en la estructura ni en la estimación del modelo se tiene en cuenta su naturaleza binaria, de este modo no se restringen los valores ajustados a estar entre cero y uno y el ajuste se efectúa minimizando la suma de los cuadrados residuales, es decir el riesgo empírico cuadrático. De este modo el método sólo es útil para clasificación y no verifica los requerimientos más básicos de Basilea II.

ROOSEN y HASTIE (1994) presentaron una formulación para el caso de respuesta binaria, a la que llamaron *Regresión Logística de Búsqueda de la Proyección,* PPLR, dentro de un contexto de Búsqueda de la Proyección Generalizada para modelos de la familia exponencial. Estos autores usaron el armazón de la regresión logística con el modelo PPR con la idea de que un procedimiento hecho a la medida específicamente para

respuesta binaria puede aproximar con más éxito la superficie subyacente que relaciona X e Y. La regresión logística de búsqueda de la proyección puede afrontarse desde dos puntos de vista:

a) Como *Modelo Logístico Aditivo de Funciones Sierra.*

La función de calificación de acreditados adopta la forma $S(X)=\sum_{r=1}^{q} h_r(\alpha_r^T X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por (5.76).

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\sum_{r=1}^{q} h_r\left(\alpha_r^T X\right)=1^T \boldsymbol{H}\left(\boldsymbol{\alpha}^T X\right) \tag{5.78}$$

donde $E\left[h_r\left(\alpha_r^T X\right)\right]=0,\ r=1,...,q$.

Como puede observarse el modelo (5.78) es un modelo logístico aditivo expansión aditiva de funciones sierra sobre las proyecciones $\alpha_r^T X$.

b) Como *Modelo Logístico Expansión Lineal de Funciones Sierra.*

ROOSEN y HASTIE (1994) formularon su modelo PPLR como expansión lineal de funciones sierra de proyecciones en la dirección de los vectores $\boldsymbol{\alpha}_r$. Para ilustrarlo, veamos que la función $S(X)=S(X_1,X_2)=X_1X_2-\sqrt{2}$ puede también representarse como una *combinación lineal de funciones sierra*

$$\begin{aligned} S(X)&=X_1X_2-\sqrt{2}=-\sqrt{2}+\frac{1}{4}(X_1+X_2)^2-\frac{1}{4}(X_1-X_2)^2 \\ &=-\sqrt{2}+\frac{1}{2}\left(\frac{1}{\sqrt{2}}X_1+\frac{1}{\sqrt{2}}X_2\right)^2-\frac{1}{2}\left(\frac{1}{\sqrt{2}}X_1-\frac{1}{\sqrt{2}}X_2\right)^2 \\ &=-\sqrt{2}+\frac{1}{2}\left(\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)^T(X_1,X_2)\right)^2-\frac{1}{2}\left(\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)^T(X_1,X_2)\right)^2 \end{aligned} \tag{5.79}$$

Siendo ahora,

$$\begin{aligned} &h_1\left(\alpha_1^T X\right)=\left(\frac{1}{\sqrt{2}}X_1+\frac{1}{\sqrt{2}}X_2\right)^2,\quad h_2\left(\alpha_2^T X\right)=-\left(\frac{1}{\sqrt{2}}X_1-\frac{1}{\sqrt{2}}X_2\right)^2, \\ &\alpha_1=\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)^T,\quad \alpha_2=\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)^T,\quad \beta_0=-\sqrt{2},\quad \beta_1=\frac{1}{2}\quad \beta_2=\frac{1}{2} \end{aligned} \tag{5.80}$$

(5.74) se puede expresar en la forma:

$$S(X)=\beta_0+\beta_1 h_1\left(\alpha_1^T X\right)+\beta_1 h_2\left(\alpha_2^T X\right)=\beta_0+\sum_{r=1}^{2}\beta_r h_r\left(\sum_{j=1}^{2}\alpha_{jr}X_j\right) \tag{5.81}$$

donde h_1 y h_2 son dos funciones que operan sobre las proyecciones:

$$\alpha_1^T X=\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)^T\begin{pmatrix}X_1\\X_2\end{pmatrix} \quad \text{y} \quad \alpha_2^T X=\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)^T\begin{pmatrix}X_1\\X_2\end{pmatrix} \tag{5.82}$$

Las proyecciones ahora se realizan en las direcciones definidas por los vectores unitarios, $\alpha_1=\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)^T$ y $\alpha_2=\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)^T$, $\|\alpha_1\|_2^2=\|\alpha_2\|_2^2=1$.

En este caso el modelo PPLR viene caracterizado en la forma

La función de calificación de acreditados se expresa $S(X)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(\alpha_r^T X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base vienen dadas en la forma (5.76)

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/\ X=x)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(\alpha_r^T X)=\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{\alpha}^T X\right) \tag{5.83}$$

con $\alpha_r^T X=1$ y $E\left[h_r\left(\alpha_r^T X\right)\right]=0,\ r=1,...,q$.

y $\boldsymbol{H}\left(\boldsymbol{\alpha}^T X\right)=\left(h_1\left(\boldsymbol{\alpha}_1^T X\right),...,h_q\left(\boldsymbol{\alpha}_q^T X\right)\right)^T$, $\boldsymbol{\alpha}^T=\left(\alpha_1,...,\alpha_q\right)$, $\alpha_r^T=\left(\alpha_{r1},...,\alpha_{rp}\right)$ y $\boldsymbol{\beta}^T=\left(\beta_1,...,\beta_q\right)$.

Tenemos, por tanto, un modelo logístico, expansión lineal de funciones de base sobre, funciones sierra, proyecciones definidas por los vectores unitarios (con norma Euclídea la unidad), $\alpha_r^T X=1,\ r=1,...,q$.

Pero en ambos casos, al igual que el método BIMARS, PPLR tampoco aborda satisfactoriamente la estimación de la probabilidad de default, puesto que también *en este método se configura la estructura formal del modelo suponiendo que todas las variables pueden ser especificadas homogéneamente, por funciones caballete o sierra* definidas según (5.76). Si añadimos ese inconveniente a la dificultad de interpretar las

funciones sierra, *a causa de que cada variable explicativa entra en el modelo de forma compleja y multifacética, para que concluyamos que a diferencia de MARS, cuyas funciones bisagra son de fácil explicación, las funciones sierra no parecen útiles para construir expansiones lineales por funciones de base de variables pertenecientes a la componente no lineal de un modelo de probabilidad generalizada por expansiones lineales de funciones de base.*

La primera cuestión a plantearnos es cuanto de grande o de pequeño ha de ser q. Una forma de resolver esta cuestión fue introducida por BREIMAN et al. (1984) en el algoritmo CART, y es utilizada por ROOSEN y HASTIE (1994) trasladando una metodología similar a la poda en los árboles de decisión a su método PPLR.

La aproximación general para la estimación del modelo propuesto por Roosen y Hastie consiste en ajustar los términos paso a paso, con un cantidad de ajuste hacia atrás entre la adición de términos. Se ajustan un total de $q_{máx} \geq q$ términos, para usar a continuación un procedimiento de selección hacia atrás que consiste en un proceso de poda del ajuste por debajo de q términos. Estamos, por tanto, ante otro método adaptativo automático. El algoritmo conduce a una estimación de la respuesta dada por

$$\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)=\hat{\beta}_0+\sum_{r=1}^{q}\hat{\beta}_r\hat{h}_r(\hat{\alpha}_r^T X) \tag{5.84}$$

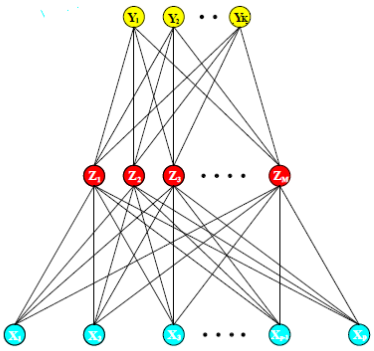
5.6.2.5 Modelo Perceptron de Capa Simple Oculta, SLPM.

En la misma línea que PPLR y con una estructura funcional muy similar se mueve el Modelo Perceptron de Capa Simple de transmisión de información hacia adelante y una sola capa oculta, SLPM**,** *una técnica muy especial de la familia de Redes Neuronales Artificiales,* ANNs*, que conecta a esta familia con los Modelos Lineales de Funciones de Base.*

Desde la óptica de la Inteligencia Artificial, las Redes de Neuronas Artificiales, ANNs**,** (BISHOP, 1995; HAYKIN, 1998), *son un paradigma de aprendizaje adaptativo y procesamiento automático inspirado en la forma en que funcionan los sistemas nerviosos biológicos. Desde la óptica estadística las ANNs son técnicas de modelización enormente flexibles fundamentadas en la emulación de tales sistemas. Igual que sucede con muchas otras técnicas estadísticas de construcción de modelos que aproximan la realidad, las*

ANNs proponen una forma concreta para la función $S(X,\theta)$ *y una serie de procedimientos concretos para estimar los parámetros* θ.

Las redes neuronales son modelos estadísticos no lineales muy parecidos al modelo de regresión de búsqueda de la proyección, PPR. De hecho *una red neuronal es una regresión en dos pasos o un modelo de clasificación típicamente representado por un diagrama de red* como en la figura siguiente:



Fuente: Hastie et al. (2009).

Figura 5.8.- Diagrama de Red Neuronal Artificial con una Capa Oculta.

El término redes neuronales, ANNs, ha evolucionado hasta abarcar una gran clase de modelos y métodos de aprendizaje y sus propiedades y características han hecho de ellas una herramienta usada con frecuencia en la resolución con éxito de problemas reales de gran complejidad, como por ejemplo, el diagnóstico médico, la decodificación del ADN y también, con bastante menos éxito en la calificación de créditos y acreditados en las Entidades Financieras.

Los modelos neuronales se construyen usando una arquitectura de capas fijas (*capa de entrada, oculta y de salida*) y los modelos de interés para problemas de regresión y clasificación son redes feedforward con una capa de entrada con las variables independientes, una capa oculta que puede poseer distintos tipos de nodos y una capa de salida lineal, siendo el número de salidas igual a uno si el problema es de regresión y mayor o igual que uno si nos encontramos ante la clasificación.

En esta Tesis Doctoral tenemos interés en las redes neuronales más usadas en los Sistemas de Calificación de Acreditados, llamadas algunas veces redes de Capa Simple Oculta con Propagación Hacia Atrás o retropropagación del error, (the hidden layer backpropagation network), o *Perceptron de Capa Simple*.

Desde el punto de vista formal, las ANNs de transmisión hacia adelante y una sola capa oculta no son más que modelos de regresión generalizada que considera una combinación lineal de transformaciones no lineales $h_r(X, w_r)$ de las variables explicativas X; cada función de base tiene como salida el resultado de una *función de transferencia* $\sigma(\bullet)$ *cuya entrada es el valor obtenido al aplicar una función de activación* $\phi(\bullet)$*, sobre las entradas de la red.* La función de excitación más utilizada es la siguiente:

$$\phi(X, \boldsymbol{w}_r) = w_o + \sum_{j=1}^{p} w_{rj} X_j \tag{5.85}$$

siendo $\boldsymbol{w}_r = (w_{r1}, ..., w_{rp})$ el vector de coeficientes asociados al nodo r-ésimo y w_{r0} el valor umbral de activación o sesgo de dicho nodo. Las funciones de base $h_r(X, w_r)$ tienen como salida el resultado de una función de activación o transferencia $\sigma(\bullet)$ aplicada a $w_o + \sum_{j=1}^{p} w_{rj} X_j$, es decir,

La función de calificación de acreditados adopta la forma $S(X) = \beta_0 + \sum_{r=1}^{q} \beta_r h_r(X, w_r)$, la función de enlace viene dada por $g(\bullet)$ y las funciones de base por

$$h_r(X, w_r) = \sigma\left(w_o + \sum_{j=1}^{p} w_{rj} X_j \right) \tag{5.86}$$

Por tanto, el modelo presenta la siguiente estructura formal

$$g\big(P(Y = 1 / X = x)\big) = S(X) = \beta_0 + \sum_{r=1}^{q} \beta_r \sigma\left(w_o + \sum_{j=1}^{p} w_{rj} X_j \right) = \boldsymbol{\beta}^T \boldsymbol{H}(X, \boldsymbol{w}) \tag{5.87}$$

siendo $\boldsymbol{H}(X, \boldsymbol{w}) = \left(\sigma\left(w_o + \sum_{j=1}^{p} w_{1j} X_j \right), ..., \sigma\left(w_o + \sum_{j=1}^{p} w_q X_{jj} \right) \right)^T$, $\boldsymbol{w} = (w_1, ..., w_q)^T$ y $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_q)$.

Donde q el número de transformaciones no lineales, $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{w})$ el conjunto de parámetros asociados al modelo, β_0 y $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^T$ los parámetros asociados a la parte lineal del modelo, $h_r(X, \boldsymbol{w}_r)$ cada una de las funciones de base y $\boldsymbol{w}_r$ su vector de parámetros asociados.

El modelo (5.87) se engloba dentro de los *modelos lineales de funciones de base* (HASTIE et al., 2009; BISHOP, 1995), es decir son modelos del mismo tipo que, por ejemplo, la regresión polinómica en la que existe una única variable y cada función de base es una potencia de esta variable, $h_r(X) = X^r$, $S(X,\boldsymbol{\theta}) = \beta_0 + \sum_{r=1}^{q} \beta_r X^r$, o, como, una extensión de esta aproximación, en la que se divide el espacio de entrada en diferentes regiones y se aproxima cada región por un polinomio distinto, *regresión por splines* (HASTIE et al. , 2009).

El modelo *Perceptron Logístico de Capa Oculta Simple,* **SLPM**, es un caso particular del modelo (5.87) donde la relación de dependencia entre el estado de default Y y las variables de riesgo de crédito X se expresa mediante un modelo logístico por expansiones lineales de funciones de base $h_r(X, w_r) = \sigma\left(w_o + \sum_{j=1}^{p} w_{rj} X_j\right)$. Este modelo se obtiene como sigue:

Las expansiones lineales, T_k, de las funciones de base $h_{rk}(\bullet)$, $k = 0,1$, viene dada por

$$T_k = \beta_{0k} + \boldsymbol{\beta}_k^T \boldsymbol{H}(\boldsymbol{X},\boldsymbol{w}) = \beta_{0k} + \sum_{r=1}^{q} \beta_{rk} h_r(\boldsymbol{X},\boldsymbol{w}_r) = \beta_{0k} + \sum_{r=1}^{q} \beta_{rk} \sigma\left(w_{r0} + \sum_{j=1}^{p} w_{rj} X_j\right) \qquad (5.88)$$
$$k = 0,1$$

Siendo $T = (T_0, T_1)$, los objetivos Y_k se modelizan como una función de T.

$$f_k(X) = g_k(T) \qquad (5.89)$$

a) La función de salida $g_k(T)$ permite una transformación final del vector de salidas T. Para la regresión se elige típicamente la función identidad $g_k(T) = T_k$. Al principio los trabajos en clasificación con K clases también usaron la función identidad, pero más tarde se abandonó a favor de la *función softmax*:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{k} e^{T_k}} \qquad (5.90)$$

La *función softmax* es exactamente la transformación logística, usada en los modelo logísticos y produce estimadores positivos que suman uno, $g_0(T) = 1 - g_1(T)$, es decir, el enfoque adoptado a la hora de interpretar las salidas de los nodos de la capa de salida es un enfoque probabilístico que considera la función softmax en cada uno de dichos nodos,

siendo $g_k(T)$ la probabilidad de que el patrón X pertenezca al nodo k, y por tanto, $g_1(T)=P(Y=1/X=x)$ y $g_0(T)=P(Y=0/X=x)$.

Por otro lado, dado que

$$\frac{g_1(T)}{g_0(T)}=\frac{e^{T_1}}{e^{T_0}}\Rightarrow \log\left(\frac{g_1(T)}{g_0(T)}\right)=T_1-T_0$$

Se tiene que la estructura de la red neuronal logística, SLPM, adopta la forma

$$\text{logit}\left(P(Y=1/X=x)\right)=(\beta_{01}-\beta_{00})+(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)^T\boldsymbol{H}(\boldsymbol{X})=\beta_0+\sum_{r=1}^{q}\beta_r\sigma\left(w_{0r}+\sum_{j=1}^{p}w_{jr}X_j\right)$$

es decir, el modelo *Perceptron Logístico de Capa Oculta Simple,* **SLPM**, viene caracterizado por

La función de calificación de acreditados adopta la forma $S(X)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(X,w_r)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por (5.86).

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/X=x)\right)=\beta_0+\sum_{r=1}^{q}\beta_r\sigma\left(w_o+\sum_{j=1}^{p}w_{rj}X_j\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X,\boldsymbol{w}) \qquad (5.91)$$

siendo $\boldsymbol{H}(X,\boldsymbol{w})=\left(\sigma\left(w_o+\sum_{j=1}^{p}w_{1j}X_j\right),...,\sigma\left(w_o+\sum_{j=1}^{p}w_q X_{jj}\right)\right)^T$, $\boldsymbol{w}=(w_1,...,w_q)^T$ y $\boldsymbol{\beta}^T=(\beta_1,...,\beta_q)$.

La elección de la función de transferencia da lugar a distintos tipos de neuronas o nodos aditivos, entre los cuales cabe destacar las neuronas umbral o perceptron (McCULLOH y PITTS, 1943) (que utilizan una función de tipo escalón), las USs (que utilizan funciones de transferencia logísticas sigmoidales, tangente hiperbólica o arco tangente) y los nodos lineales (cuya función de transferencia es la función identidad). Algunas veces se usan funciones de base radiales Gausianas como función de activación, produciendo lo que se conoce como redes de *función de base radial*, RBF .

Si, como es habitual, se elige como función de transferencia o activación, $\sigma(\bullet)$, la función de distribución de la variable aleatoria logística de parámetros 0 y 1, $\sigma(Z)=\frac{1}{1+e^{-Z}}$,

entonces $\sigma\left(w_{0r}+w_r^T X\right)=\dfrac{1}{1+e^{-\left(w_{0r}+w_r^T X\right)}}$ y se tiene el modelo **SLPM sigmoide** con la siguiente estructura funcional:

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)=\beta_0+\sum_{r=1}^{q}\beta_r\frac{1}{1+e^{-\left(w_{0r}+\sum_{j=1}^{p}w_{jr}X_j\right)}} \tag{5.92}$$

Aquí, como en PPLR, tienen que determinarse las direcciones w_r y los sesgos β_r y su estimación es el objetivo de la computación. Las funciones de base se eligen adaptativamente de un diccionario D de funciones de base candidatas entre las que elegir y el modelo es construido usando como mecanismo de búsqueda el algoritmo de *retropropagación, mecanismo de aprendizaje supervisado para redes neuronales artificiales multicapa* muy popular para resolver problemas de clasificación y pronóstico. El algoritmo de retropropagación es un método de entrenamiento general no lineal que requiere diferenciabilidad en la capa de salida, como es el caso de la sigmoide en (5.92) (WERBOS, (1974), PARKER, (1985), RUMELHART et al., (1986)).

La estimación de los parámetros, en lenguaje estadístico, o el aprendizaje de la ANN, en la teoría de aprendizaje máquina, consiste en estimar un valor para el conjunto de parámetros $\boldsymbol{\theta}=\left(\beta_0,\boldsymbol{\beta},\boldsymbol{w}\right)$ y una arquitectura para la red (es decir, el número de transformaciones no lineales q y el número de conexiones entre los distintos nodos de la red). Suponiendo una arquitectura fija, el aprendizaje de los parámetros ha sido tradicionalmente llevado a cabo mediante algoritmos de optimización basados en *gradiente descendente* de la función de error, tales como el Algoritmo de RetroPropagación (BackPropagation, BP).

Respecto del ajuste de la red (5.92), si se denota por θ el conjunto completo de los parámetros se tiene

$$\theta=\left\{w_{0r},w_r,\quad r=1,2...,q\right\}\cup\left\{\beta_{0k},\beta_k,\quad k=1,2\right\}$$

$$\textit{donde}\quad w_r=\left(w_{r1},...,w_{rp}\right),\quad \textit{con}\quad q\times(p+1)\quad \textit{elementos}$$
$$\beta_k=\left(\beta_{k1},...,\beta_{kq}\right),\quad \textit{con}\quad 2\times(q+1)\quad \textit{elementos}$$

por tanto, θ consta de $q\times(p+1)+2\times(q+1)$ parámetros. Se buscan los valores de los parámetros desconocidos que mejor ajustan el modelo a los datos de entrenamiento, es decir que minimizan el error, llamado entropía cruzada o desvianza,

$$R(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{k=0}^{1} y_{ik}\log\left[g_k(x_i)\right] \tag{5.93}$$

Podemos concluir que *las Redes Neuronales feedforward, o con transmisión de información hacia adelante, con una sola capa oculta, con la transformación logística y funciones de base de Unidad Sigmoide (US), cuyos modelos son ajustados por el criterio de pérdida de entropía cruzada son exactamente modelos de regresión logística lineal en las capas ocultas, y todos los parámetros son estimados por máxima verosimilitud.*

Las unidades en el centro de la red se llaman ocultas por que los valores $h_r(\bullet)$ no son observados directamente. Se puede pensar en las $h_r(\bullet)$ como una expansión de funciones de base de las entradas originales, la red neuronal es entonces un modelo lineal estándar, o un modelo logit lineal, usando estas transformaciones como entradas. Existe, sin embargo, una importante mejora con respecto a las técnicas de expansión de funciones de base normales, en las ANNs los parámetros de las funciones de base aprenden de los datos.

Nótese que si $\sigma(\bullet)$ es la función identidad entonces el modelo total se derrumba a un modelo lineal en las entradas. Por tanto una red neuronal puede ser entendida como una generalización del modelo lineal, tanto para la regresión como para la clasificación. Introduciendo transformaciones no lineales $\sigma(\bullet)$ se agranda la clase de los modelos lineales. En la figura anterior se ve que la tasa de activación de la sigmoide depende de la norma de w_r, y, si $\|w_r\|$ es muy pequeño, la unidad puede realmente operar en la parte lineal de sus funciones de activación.

La red neuronal con una capa oculta, SLPM, tiene exactamente la misma forma que el *modelo en* PPLR, la diferencia es que PPLR usa funciones no paramétricas $\varphi_r(v)$, mientras que la red neuronal usa ordinariamente una función basada en $\sigma(v)$, con tres parámetros libres en sus argumentos. En detalle viendo la red neuronal como un modelo PPLR, se identifica

$$\begin{aligned}\varphi_r\left(\alpha_r^T X\right) &= \beta_r\sigma\left(w_{0r} + w_r^T X\right)\\ &= \beta_r\sigma\left(w_{0r} + \|w_r\|\left(\alpha_r^T X\right)\right)\\ &= \beta_r\sigma\left(w_{0r} + \left(\|w_r\|\alpha_r^T X\right)\right)\end{aligned}$$

donde $\alpha_r = \frac{w_r}{\|w_r\|}$ es el m-ésimo vector unidad.

Puesto que $\sigma_{\beta,\alpha_0,s}(v) = \beta\sigma(\alpha_0 + sv)$ tiene mayor complejidad que una función no paramétrica más general $g(v)$, no es sorprendente que una red neuronal use de 20 a 100 de tales funciones, mientras que el modelo PPLR usualmente utiliza solamente entre $q = 5$ y $q = 10$.

Las redes neuronales no gozan en general de buena fama en el campo de la predicción, por cuanto ha existido un gran número de de redes neuronales de tipo publicitario que han conseguido que estos modelos sean vistos como cajas negra mágicas y misteriosas y en la mayor parte de los casos así es.

Muchos de los modelos de redes neuronales que se han usado en los sistemas de calificación de acreditados están más orientados a la decisión sobre la concesión de un crédito que a estimar un modelo de probabilidad concebido desde la óptica de los acuerdos de Basilea II, en ese sentido, aparte de que no proporcionan en muchos casos la probabilidad de default, la familia logística a la que pertenece el modelo (5.98) es una de las pocas excepciones, constituyen siempre una "caja negra" incapaz de explicar la relación entre el default y las variables explicativas del riesgo de crédito, lo que como hemos reiterado en multitud de ocasiones es imprescindible desde los requerimientos de Basilea II.

5.6.2.6 Modelos Regularizados por Núcleos, KRM.

A través de las expansiones lineales por funciones de base de las variables explicativas se trasladan los datos a un espacio agrandado, generalmente de Hilbert, $\mathcal{H}$. Con la frontera lineal en el espacio agrandado se consigue una mejor separación de los datos de entrenamiento, separación que se traslada a la frontera no lineal en el espacio original.

El peligro está en que con las suficientes expansiones de funciones de base los datos pueden hacerse separables de forma artificial, resultando sobre ajustados. Una alternativa consiste en extender la idea anterior a espacios de alta dimensión y controlar la complejidad del modelo ajustando la función con el criterio de minimización de la pérdida empírica asociada a una cierta función de pérdida regularizada; los métodos que estiman el modelo de acuerdo con esta filosofía se llaman *Métodos Regularizados por Núcleos,* KRM.

Los métodos KRM *se basan en la idea de agrandar el espacio usando para la función de calificación de acreditados,* $S(X)$*, expansiones por funciones de base de las variables originales y aprovechan el hecho de que la formulación del riesgo empírico admite una gran flexibilidad para* $S(X)$*, con generalizaciones no lineales como* $S(X)\in\mathcal{H}$*, siendo* $S(X)$ *una función arbitraria y* $\mathcal{H}$ *un espacio de Hilbert.*

Las *técnicas no paramétricas de estimación por núcleos*, utilizan la idea anterior expandiendo las variables originales a espacios de alta dimensión y ajustando la función con el criterio de minimización del riesgo empírico asociado a una determinada función de pérdida regularizada.

Si los elementos de la muestra no son linealmente separables, no existe el Hiperplano Separador Optimal que pueda clasificar correctamente a estos elementos, por lo que será necesario hallar una hipersuperficie en su lugar, lo que se puede conseguir usando una correspondencia no lineal

$$\begin{aligned} h:\ \mathbb{R}^p &\to \mathcal{H} \\ X &\to h(X) \end{aligned} \tag{5.94}$$

siendo $\mathcal{H}$ un espacio de Hilbert de características.

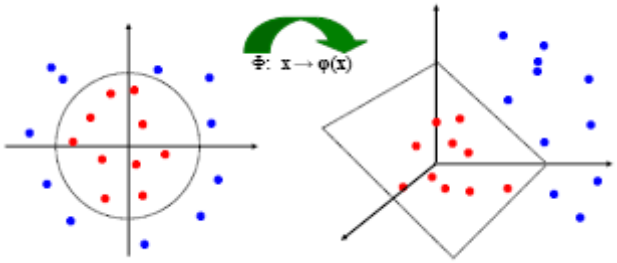


Figura 5.9.- Como puede observarse, los datos no son linealmente separables en el espacio de entrada en $\mathbb{R}^2$, sin embargo, existe un hiperplano en $\mathbb{R}^3$ que los separa perfectamente.

de este modo se introduce la transformación, función de base, $X\to h(X)$, llamada *función núcleo definida positiva*, desde el espacio de entrada en el espacio de alta dimensión de Hilbert $\mathcal{H}$, y es posible que el nuevo conjunto muestral transformado en el espacio de Hilbert, $\{h(x_i),\ y_i\}_{i=1}^N$, se vuelva separable, figura 5.9. Como consecuencia *el*

problema de hallar una hipersuperficie en el espacio de entrada se convierte en encontrar un hiperplano en el espacio de Hilbert $\mathcal{H}$.

Definición 5.9.- Sea $X \subset \mathbb{R}^N$ un conjunto cerrado y $N \in \mathbb{N}$, un *núcleo definido positivo* es una función simétrica $K(\bullet,\bullet): X \times X \to \mathbb{R}$, tal que para cualquier conjunto finito de puntos $\{x_i\}_{i=N}$ de X y para cualesquiera números reales $\beta_1,...,\beta_N$ se verifica

$$\sum_{i=1}^{N}\sum_{r=1}^{N}\beta_i\beta_r K(x_i,x_r) \geq 0 \tag{5.95}$$

o, en forma matricial, $\beta^T K\beta \geq 0$.

En otras palabras, $K(\bullet,\bullet)$ *representa un producto escalar en algún espacio de Hilbert,* (WEYL, 1928).

Los núcleos aquí no deben ser confundidos con las funciones núcleo de estimación de densidades, que representan métricas para especificar vecindades locales, puesto que en este caso los núcleos computan productos interiores en espacios de características de alta dimensión y son usados para modelos no lineales regularizados.

Con base en las ideas expuestas en los dos párrafos anteriores se construyen métodos alternativos a los que hemos visto en lo que va de capítulo, en concreto, *consisten en expandir el espacio original de entrada a un espacio de Hilbert, de forma que la carencia de un hiperplano separador en el espacio original sea suplida por una hipersuperficie en el espacio agrandado de Hilbert,* $\mathcal{H}_K$. Esta especial expansión se consigue a través del *teorema representer no paramétrico de* KIMELDORF y WAHBA (1971) y COX y O'SULLIVAN (1990) y SCHÖLKOPF et al. (2001), que asegura que la solución al problema de optimización

$$\underset{S\in H_K}{mín}\left[\sum_{i=1}^{N}\ell(y_i,S(x_i))+\lambda\|S\|^2_{\mathcal{H}_K}\right] \tag{5.96}$$

es finito dimensional y adopta la forma $S(X)=\beta_0+\sum_{r=1}^{N}\beta_r K(X,x_r)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{K}$ donde $\boldsymbol{\beta}=(\beta_1,...,\beta_N)^T$ y $\boldsymbol{K}$ es la matriz $N\times N$ con la $ij-ésima$ entrada igual a $K(x_i,x_r)$, siendo $K(\bullet,\bullet)$ un núcleo con valores reales definido positivo, es decir, el *teorema representer* permite representar $S(X)$, para $S(X)$ optimal, como una combinación lineal de *funciones núcleo* $K(\bullet,\bullet)$, *o, de otro modo,* $S(X)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(X)$ *es una expansión lineal donde las funciones de base* $h_r(X)$ *son funciones núcleo,* $h_r(X)=K(X,x_r)$.

Por tanto, el problema infinito dimensional (5.96) se reduce al criterio de dimensión finita

$$\min_{\beta_0,\beta_1,\ldots,\beta_N}\left[\sum_{i=1}^{N}\ell\left(y_i,\beta_0+\sum_{r=1}^{N}\beta_r K\left(x_i,x_r\right)\right)+\lambda\sum_{i=1}^{N}\sum_{r=1}^{N}\beta_r K\left(x_i,x_r\right)\beta_i\right] \tag{5.97}$$

que en notación matricial se expresa

$$\min_{\beta_0,\boldsymbol{\beta}}\ell\left(Y,\beta_0+\boldsymbol{\beta}^T K\right)+\lambda\boldsymbol{\beta}^T K\boldsymbol{\beta} \tag{5.98}$$

La maquinaria anterior está impulsada por la elección del núcleo K y la función de pérdida ℓ. Las dos funciones de pérdida más habituales utilizadas en la regularización por núcleos, en el entorno de los sistemas de calificación de acreditados, son la pérdida logística y la pérdida bisagra SVM, que respectivamente darán lugar a las técnicas *Regresión Logística Regularizada por Núcleos,* KLR, *y Maquinas de Vector Soporte,* SVM.

La *Regresión Logística Regularizada por Núcleos,* KLR, contempla modelos que responden a la formulación (5.1) con la transformación $g(\bullet)$ logística y bajo las hipótesis del *teorema representer* en la versión de SCHÖLKOPF et al. (2001), y su estructura formal viene caracterizada por

La función de calificación de acreditados se expresa $S(X)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por

$$\begin{aligned} &h_r(X)=K(X,x_r)\\ &\textit{donde}\quad K(X,x_r)\ \text{es un núcleo con valores reales definido positivo}\end{aligned} \tag{5.99}$$

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P\left(Y=1/\,X=x\right)\right)=\beta_0+\sum_{r=1}^{N}\beta_r K\left(X,x_r\right)=\beta_0+\boldsymbol{\beta}^T K \tag{5.100}$$

donde $\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_N\right)^T$ y $\boldsymbol{K}$ es la matriz $N\times N$ con la $ij-ésima$ entrada igual a $K\left(x_i,x_r\right)$, siendo $K(\bullet,\bullet)$ una función núcleo.

Como es habitual en los modelos con la transformación logística, para los propósitos de estimar la función de calificación de acreditados y la probabilidad de default asociada se

utiliza la función de pérdida logística, y, por tanto, el estimador del modelo (5.100) es la solución del problema de minimización

$$\underset{\beta_0\boldsymbol{\beta}}{mín}\ -\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{K}\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{K}\right)\right)\right]+\frac{\lambda}{2}\boldsymbol{\beta}^T K\boldsymbol{\beta} \qquad (5.101)$$

donde $\boldsymbol{\beta}=(\beta_1,...,\beta_N)^T$ es el vector de parámetros a estimar, y $K=\left(K(x_i,x_r)\right)_{N\times N}$ la matriz de valores muestrales de la función núcleo $K(\bullet,\bullet)$, λ es un hiperparámetro de regularización que puede ser obtenido por validación cruzada.

Generalmente el núcleo más usado en los modelos logísticos regularizados por núcleos es la *función de base radial Gaussiana* $K(x,x')=\exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$.

El principal inconveniente de la Regresión Logística Regularizada por Núcleos es que todos los vectores del conjunto de entrenamiento están envueltos en la solución final, lo que no es aceptable para grandes conjuntos de datos, debido al elevado número de parámetros distintos de cero, como a veces ocurre en los sistemas de calificación de acreditados. Para el problema de optimización (5.101) las condiciones de optimalidad de primer orden para $S(X)$ implican

$$S(x)=\beta_0+\sum_{i=1}^{N}\beta_i K(x,x_i)$$

$$con \qquad (5.102)$$

$$\beta_i=\frac{1}{2}\left(\frac{y_i+1}{2}-P\left(Y=1/\ X=x_i\right)\right)$$

Por otro lado, dado que $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=S\left(X\right)$, se tiene $0<\hat{P}\left(Y=1/\ X=x\right)<1$, por tanto, para todas las observaciones del conjunto de entrenamiento, $\beta_i\neq 0$, es decir la expansión exacta requiere N coeficientes.

Por lo que respecta a las *Maquinas de Vector Soporte,* SVM, ya nos hemos referido a que cuando los datos no son separables se recurre a la técnica *vector soporte*, SV, extensión de OSH, que permite solapamientos de clases en el espacio de las variables explicativas, si bien minimizando el alcance de este solapamiento, a los que trata relajando la maximización del margen M para permitir que algunos puntos estén en el lado equivocado del margen.

Bajo la misma filosofía que el método SV y con el mismo objetivo, si bien desde una óptica distinta que *suple la carencia de un hiperplano separador en el espacio*

original con una hipersuperficie en el espacio agrandado de Hilbert $\mathcal{H}_K$ *a través del teorema representer no paramétrico*, analizaremos ahora la técnica SVM.

El ***Modelo Maquina de Vector Soporte, por núcleos, SVM***, también responde a la formulación (5.1) donde la *transformación* $g(\bullet)$ es la *vector soporte* bajo las hipótesis del *teorema representer* en la versión de SCHÖLKOPF et al. (2001). Por tanto, su estructura formal viene caracterizada por

La función de calificación de acreditados se expresa $S(X)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=sign\left\{\bullet-\frac{1}{2}\right\}$ y las funciones de base por (5.99).

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{Sign}\left(P(Y=1/X=x)-\frac{1}{2}\right)=\beta_0+\sum_{r=1}^{N}\beta_r K(X,x_r)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{K} \qquad (5.103)$$

donde $\boldsymbol{\beta}=(\beta_1,...,\beta_N)^T$ y $\boldsymbol{K}$ es la matriz $N\times N$ con la $ij-ésima$ entrada igual a $K(x_i,x_r)$, siendo $K(\bullet,\bullet)$ una función núcleo.

Como es habitual en los modelos con la transformación vector soporte, para los propósitos de estimar la función de calificación de acreditados y la probabilidad de default asociada se utiliza la función de pérdida vector soporte, y, por tanto, el estimador del modelo (5.103) es la solución del problema de minimización con un término de ajuste, $\left[\mathbf{1}^T\left(\mathbf{1}-\boldsymbol{Y}\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}\right)\right)\right]_+$, y un término de regularización, $\boldsymbol{J}\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right)=\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta}$, *con* $\boldsymbol{J}'(\boldsymbol{\beta})=\boldsymbol{\beta}^T\boldsymbol{K}$, que adopta la forma

$$\underset{\beta_0,\boldsymbol{\beta}}{mín}\ \left[\mathbf{1}^T\left(\mathbf{1}-\boldsymbol{Y}\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{K}\right)\right)\right]_+ +\frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta} \qquad (5.104)$$

λ es el hiperparámetro de adaptación que puede ser obtenido por validación cruzada o técnicas análogas.

Resolver el problema de optimización (5.104) es equivalente al siguiente *problema de optimización con restricciones de desigualdad lineal*,

$$\begin{aligned} &\min_{\beta} \ \frac{1}{\lambda} \mathbf{1}^T \boldsymbol{\xi} + J\left(\boldsymbol{\beta}^T \mathbf{K}\right) \\ &\text{sujeto a} \ \ \boldsymbol{\xi} \geq \vec{0} \ \ y \ \ Y \circ \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{K}\right) \geq \mathbf{1} - \boldsymbol{\xi} \end{aligned} \tag{5.105}$$

donde $A \circ B$ es el producto de Hadamard de matrices.

El problema (5.105) conduce a la siguiente *funcional de Lagrange para el problema primal*:

$$L_P = \frac{1}{\lambda} \mathbf{1}^T \boldsymbol{\xi} + J\left(\boldsymbol{\beta}^T \mathbf{K}\right) - \left(\left(\boldsymbol{\alpha} \circ \mathbf{Y}\right)^T \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{K}\right) - \boldsymbol{\alpha}^T \left(\mathbf{1} - \boldsymbol{\xi}\right)\right) - \boldsymbol{\mu}^T \boldsymbol{\xi} \tag{5.106}$$

donde los vectores $\boldsymbol{\alpha}$, $\alpha_i \geq 0$, y $\boldsymbol{\mu}$, $\mu_i \geq 0$, son los multiplicadores de Lagrange.

La funcional objetivo Lagrangiano dual de Wolfe viene dada en la forma,

$$L_D = \boldsymbol{\alpha}^T \mathbf{1} - J'(\boldsymbol{\beta}) + J\left(\boldsymbol{\beta}^T \mathbf{H}\right) = \boldsymbol{\alpha}^T \mathbf{1} - \boldsymbol{\beta}^T \mathbf{K} + \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} \tag{5.107}$$

La cual da un límite inferior sobre la función objetivo para cualquier punto factible. Por tanto, se resuelve el problema dual:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \ \boldsymbol{\alpha}^T \mathbf{1} - \boldsymbol{\beta}^T \mathbf{K} + \boldsymbol{\beta}^T \mathbf{K} \boldsymbol{\beta} \\ &\textit{Sujeto a}: \ \ \boldsymbol{\alpha}^T Y = 0 \\ &\qquad\qquad 0 \leq \alpha_i \leq \frac{1}{\lambda}, \quad i = 1, \ldots, N. \end{aligned} \tag{5.108}$$

donde α_i, $i = 1, \ldots, N$, son los multiplicadores de Lagrange.

Dadas las soluciones $\hat{\beta}_0$ y $\hat{\beta}$, se tiene el estimador de la función de calificación lineal

$$\hat{S}(X) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{K} \tag{5.109}$$

de donde

$$Sign\left\{\hat{P}\left(Y = 1 / X = x\right) - \frac{1}{2}\right\} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{K} \tag{5.110}$$

Desde la ecuación (5.110) no es posible conocer el estimador de la probabilidad de default, $\hat{P}\left(Y = 1 / X = x\right)$, sino simplemente si este es o no igual o superior a 0.50, por tanto con el modelo SVM no es posible estimar directamente la probabilidad de default, lo que es un serio inconveniente desde la perspectiva de los requerimientos de Basilea II.

Como consecuencia de la naturaleza lineal por tramos de la primera parte del criterio (5.108), ocurre con frecuencia que para una fracción de los α_i su valor es cero (los puntos que no son soporte), cuanto menos se solapen las clases mayor será esta fracción.

Reduciendo λ se reduce generalmente el solapamiento (consiguiendo una $S(X)$ más flexible). Un pequeño número de puntos soporte significa que $\hat{S}(X)$ puede ser evaluada más rápidamente, lo cual es importante para el tiempo de computación. De todas formas ha de tenerse en cuenta que una fuerte reducción del solapamiento puede conducir a una pobre generalización.

Maximización del margen y Vector Import.

Los métodos de vector soporte se concibieron para estimar el margen, M, que se define como la distancia Euclídea minimal entre cualquier elemento de entrenamiento x_i, para ambas clases, y la frontera del hiperplano separador, es decir, $min_i\ y_i S(x_i)$, $y_i \in \{-1,+1\}_{i=1}^{N}$. Así, por ejemplo, para el caso OSH, bajo el supuesto de que el margen existe, se tiene $\frac{y_i S(x_i)}{\|\beta\|} \geq M \equiv y_i\left(\beta_0 + \beta^T x\right) \geq M\|\beta\|$ y si se desea una franja vacía alrededor de la frontera de ancho $\frac{1}{\|\beta\|}$ ha de verificarse $\frac{y_i\left[\beta_0 + \beta^T x_i\right]}{\|\beta\|} \geq \frac{1}{\|\beta\|}$, $\forall i = 1,...,N$, por lo que se impone al OSH en forma canónica la restricción $y_i\left[\beta_0 + \beta^T x_i\right] \geq 1,\ \ i = 1,...,N$. Se busca β_0 y β que hagan máxima esa franja vacía alrededor de la frontera. Intuitivamente, el margen mide la eficacia del hiperplano en la separación entre las dos clases.

El vector solución de los coeficientes β para el hiperplano separador optimal está definido en términos de combinaciones lineales de los *puntos soporte* x_i, puntos que se definen sobre la frontera de la franja vía $\alpha_i > 0$. Para estos mecanismos de clasificación, los vectores soporte son los elementos críticos del conjunto de datos de entrenamiento.

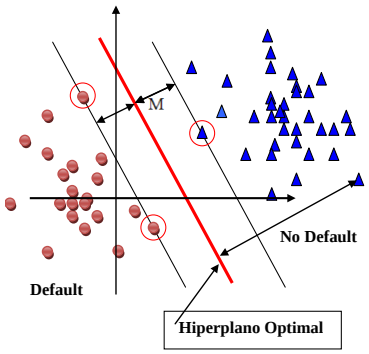


Figura 5.10.- Hiperplano separador Optimal.

En la figura 5.10 se muestra un hiperplano separador optimal, una franja de margen M y tres puntos soporte, representados dentro de un circulo, dos de default y uno de no default.

A pesar de la fuerte debilidad que caracteriza a los modelos de vector soporte, OSH, SV y SVM, al no ser posible estimar directamente la probabilidad de default, sus defensores resaltan sus dos grandes fortalezas como métodos de clasificación, maximizan el margen y la frontera de clasificación está definida en términos de combinaciones lineales de los *puntos soporte* x_i, puntos que se definen sobre la frontera de la franja vía los multiplicadores de Lagrange $\alpha_i > 0$.

Apoyados en los dos puntos fuertes mencionados, que convierten a los modelos de *vector soporte* en magníficos clasificadores y restando importancia tanto al cálculo directo de la probabilidad de default como a la capacidad explicativa de los modelos no son pocos los artículos aparecidos, hasta bien avanzada la primera década de este siglo, en que estos métodos resultan claros vencedores frente a los modelos logísticos.

Sin embargo la victoria no fue más que un espejismo, puesto que, por un lado, ZHU y HASTIE (2004) y ROSSET et al. (2004a) demostraron que KLR es también una técnica maximizadora del margen, por lo que la ventaja por ese hecho se desvaneció. Y, por otro lado, el principal inconveniente de la *Regresión Logística Regularizada por Núcleos*, todos los vectores del conjunto de entrenamiento están envueltos en la solución final, lo que se traduce en un número, con frecuencia inaceptable, de parámetros distintos de cero, fue resuelto un año más tarde también por ZHU y HASTIE (2005) con su propuesta de evitar ese inconveniente usando un algoritmo de selección hacia delante que aproxima la expansión completa con un número fijado de parámetros α_i distintos de cero. Esta aproximación puede ser interpretada como añadir al criterio de optimización un término extra penalizando el número de coeficientes no cero, en esto consistió su propuesta de ***Maquina de Vector Import,*** **IVM**, *o Regresión Logística por Núcleos con Vector Importado,* KLR_{IV}.

El algoritmo IVM busca un submodelo, $S_{M_S}(X) = \sum_{x_i \in M_S} \beta_i K(x, x_i)$, para aproximar el modelo completo, donde $S_{M_S} \in H_K$, espacio de funciones generado por el núcleo K, M_S es un subconjunto del conjunto de entrenamiento $\{x_1, x_2, ..., x_N\}$ y los puntos de M_S se llaman puntos import. La propuesta de ZHU y HASTIE, (2005) logró obtener una solución rala

de la expansión logística por núcleos considerando sólo funciones de base correspondientes a un subconjunto M_S del conjunto de entrenamiento $\boldsymbol{\tau}$ con q observaciones $S(X)=\sum_{i=1}^{q}\beta_i K(x,x_i)$, $q<<N$.

ZHU y HASTIE (2004) usaron un ejemplo de simulación para ilustrar la semejanza entre el funcionamiento de KLR y SVM con datos simulados de una mixtura de distribución Gaussiana $N\left((0,1)^T,I\right)$ usando el núcleo Gaussiano de base radial y seleccionando el parámetro de regularización λ de tal forma que se alcance un buen error de mala clasificación. Comprobaron que para su ejemplo el modelo ajustado por regresión logística por núcleos es bastante similar como clasificador al del SVM, ambos dan una eficaz frontera de clasificación bastante parecida, pero puesto que KLR estima los odds de las probabilidades de las clases, se pueden adicionalmente realizar contornos de probabilidad, lo que constituye una importante ventaja de KLR sobre SVM en la construcción de sistemas de calificación de acreditados.

Desde el punto de vista de Basilea II, a pesar de que las técnicas KLR y KLR_{IV} resuelven el fundamental problema de estimar directamente la probabilidad de default que presentan las técnicas de vector soporte, no nos parecen convenientes para construir modelos de credit scoring por cuanto no se resuelve el problema de la interpretabilidad, pues estos dos modelos describen la relación de dependencia de la variable estado de default con los acreditados y no con las variables de riesgo de crédito. Tal como hemos expuesto KLR_{IV} *es un clasificador que maximiza el margen y además se construye sólo sobre los puntos vector soporte igual que* SVM*, que además a diferencia de este último proporciona la probabilidad de default y no sólo su signo, por lo que entendemos que como clasificador es más adecuado en credit scoring que* SVM*, pero esta ventaja, por lo que hemos expuesto, no es suficiente para que sea idóneo para nuestros objetivos*.

5.6.2.7 Modelos Parcialmente Lineales, LPM.

En una situación intermedia entre los modelos totalmente lineales, 5.6.1, también casi siempre poco realistas, y los descritos hasta ahora en este apartado 5.6.2, se encuentran los Modelos Parcialmente Lineales, LPM. Estos modelos se configuran considerando como hipótesis de partida que se tiene información fundada sobre el hecho de que

una o varias de las variables de interés tienen influencia lineal sobre el comportamiento frente al default y el resto influencia no lineal.

Los modelos parcialmente lineales se formalizan en los siguientes términos: Supongamos que tenemos p_1 variables lineales $(X_1,...,X_{p_1})$ pero se desconoce el tipo de influencia que ejercen las otras $p_2 = p - p_1$ variables. Dado que $X^T = (X_1,...,X_{p_1},X_{p_1+1},...,X_{p_2})$, adoptando la siguiente notación $U^T = (U_1,...,U_{p_1})$, $U_j = X_j$, $j = 1,...,,..., p_1$, y $V^T = (V_1,...,V_{p_2})$, $V_j = X_{p_1+j}$, $j = 1,..., p_2$, se tiene

$$X^T = (U^T,\ V^T) = \left((U_1,...,U_{p_1}),\ (V_1,...,V_{p_2})\right)$$

Por tanto, la estructura formal más general del *modelo parcialmente lineal*, LPM, como *extensión semiparamétrica de los modelo lineales*, se expresa como la suma de una componente lineal, combinación lineal de las variables lineales $U = (U_1,...,U_{p_1})$, y una componente no lineal, que se diseña como una expansión no paramétrica $h(\boldsymbol{V})$, donde $h(\bullet)$ es una función no paramétrica infinito dimensional. Es decir, $h(\bullet)$ es una función de suavizado no paramétrica de dimensión infinita que opera sobre un argumento multidimensional de variables $V^T = (V_1,...,V_{p_2})$ y se calcula de una manera flexible, por ejemplo, cualquier método de suavizado no paramétrico. El modelo queda caracterizado en la forma

La función de calificación de acreditados se expresa $S(X) = \beta_0 + \sum_{r=1}^{p} \beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)$ y las funciones de base por

$$h_r(\bullet): \mathbb{R}^P \to \mathbb{R}, \quad r = 1,.., p$$

$$X \to h_r(X) = \begin{cases} U_r & r = 1,..., p_1 \\ h(V) & r = p_1 + 1 \end{cases} \tag{5.111}$$

Por tanto, el modelo presenta la siguiente estructura formal

$$g\left(P(Y = 1/\ X = x)\right) = \underbrace{\beta_0 + \sum_{r=1}^{p_1} \beta_r U_r}_{Lineal} + \underbrace{h\left(V_1,...,V_{p_2}\right)}_{No\ Lineal} = \boldsymbol{\beta}^T \boldsymbol{U} + h(\boldsymbol{V}) \tag{5.112}$$

donde $h(\boldsymbol{V})$ es una función no paramétrica infinito dimensional y $\beta_{p_1+1} = 1$.

Los modelos parcialmente lineales, LPM, con estructura formal (5.112), permiten un trato más flexible, para un subconjunto de las variables explicativas, que los modelos lineales.

MÜLLER y HÄRDLE (2003) propusieron un modelo con la estructura formal determinada por (5.113) donde, además, por un lado, se exige a las variables $V^T=\left(V_1,...,V_{p_2}\right)$ que sean absolutamente continuas, puesto que en su planteamiento la función $h(\boldsymbol{V})$ se determina por un método de suavizado no paramétrico que requiere la *continuidad absoluta de las variables* $V^T=\left(V_1,...,V_{p_2}\right)$, y, por otro lado, consideraron la *transformación logística* como función de enlace entre la probabilidad de default y las función de calificación de acreditados. Su aproximación se basa en que los modelos logísticos parcialmente lineales extienden la "facilidad de interpretación" de la estructura del modelo LOGIT para componentes no paramétricas. Estamos ante el **Modelo Logístico Parcialmente Lineal, LPLM**, *extensión semiparamétrica del modelo LOGIT,* caracterizado por

La función de calificación de acreditados se expresa $S(X)=\beta_0+\sum_{r=1}^{p}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por (5.12).

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/\ X=x)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{\textit{Lineal}}+\underbrace{h\left(V_1,...,V_{p_2}\right)}_{\textit{No Lineal}}=\boldsymbol{\beta}^T\boldsymbol{U}+h(\boldsymbol{V}) \qquad (5.113)$$

donde $h(\boldsymbol{V})$ es una función no paramétrica infinito dimensional y $\beta_{p_1+1}=1$.

Esta técnica presenta una interesante característica particular del LOGIT que consiste en que se estiman simultáneamente las puntuaciones de crédito y las probabilidades de default. Esto provocó un creciente interés en los LPLM para rediseñar los sistemas de calificación del crédito y de los acreditados de acuerdo con los requerimientos del NBCA de Basilea II.

Para la estimación de estos modelos no es necesario recurrir a técnicas totalmente no paramétricas, de mayor complejidad técnica y con resultados de más difícil interpretación, si no que se estiman a través de la **Regresión**

Logística Parcialmente Lineal, LPLR, (SEVERINI y WONG (1992), SEVERINI y STANISWALIS (1994), CHEN (1995), MÜLLER (2001), MÜLLER y HÄRDLE (2003), PENG (2004), PENG y WANG (2004), HÄRDLE et al. (2004a, 2004b)).

El modelo se estima en los términos siguientes:

Para una muestra de (X,Y), $\tau=\{(x_i,y_i\}_{i=1}^{N}\in\left(X\times\mathrm{Y}\right)^{N}$, y la función de pérdida logística empírica, $\ell(y_i,f(x_i))=-\left(yf(x_i)-\mathrm{lo}\left(\mathrm{g}+e^{f(x_i)}\right)\right)$, el *Problema de la Regresión Logística Parcialmente Lineal* para el modelo (5.112) se expresa en la forma

$$\min_{\beta_0,\boldsymbol{\beta},h}\ -\sum_{i=1}^{N}\left[y_i\left(\left(\beta_0+\boldsymbol{\beta}^T U\right)+h(V)\right)-\log\left(1+e^{\left(\beta_0+\boldsymbol{\beta}^T U\right)+h(V)}\right)\right] \qquad (5.114)$$

Los métodos existentes para resolver (5.114) se pueden clasificar en dos categorías:

1. *Tipo perfil de Verosimilitud,* categoría que comprende la aproximación de SEVERINI & WONG (1992) Y SEVERINI y STANISWALIS (1994) que usan dos funciones de verosimilitud diferentes para la estimación de la componente paramétrica y la componente no paramétrica y el método de Speckman Generalizado, SPECKMAN (1988).

2. *Tipo Backfitting,* cuyos métodos se basan alternativamente entre métodos paramétricos y no paramétricos, SEVERINI y STANISWALIS (1994), HÄRDLE et al. (2004a). Puesto que los Modelos Parcialmente Lineales Generalizados pueden verse como formados por dos componentes aditivas, la idea central del método backfitting, FRIEDMAN y STUETZLE (1981) y HASTIE y TIBSHIRANI (1990), puede aplicarse perfectamente aquí.

Los Modelos Logísticos Parcialmente Lineales, a pesar de que resuelven algunos importantes problemas planteados en la estimación de la probabilidad de default, tales como: se plantean en términos del logit de la probabilidad, permiten captar la linealidad paramétricamente, consideran la componente no lineal y no son excesivamente complejos, siguen sin resolver un problema importante desde el enfoque de Basilea II, la parte no lineal se estima a través de una función no paramétrica, infinito dimensional, que generalmente no permite explicar la aportación de cada variable explicativa del riesgo de crédito al estado

de default, además, como ocurre con casi todos los métodos no paramétricos, estas técnicas están aquejadas de la maldición de la dimensionalidad. Por esta razón *no los consideramos modelos satisfactorios para nuestro objetivo de estimar la probabilidad de default, calificar a los acreditados y clasificar a nuevos solicitantes de crédito de acuerdo con los requerimientos de Basilea II.*

Con el fin de resolver el problema anterior, surgió una peculiar familia de modelos logísticos parcialmente lineales, los ***Modelos Logísticos Aditivos Parcialmente Lineales,*** **LPALM**, que pueden considerarse desde dos ópticas: como *Modelos Logísticos Aditivos con una Componente Lineal* o bien como una extensión de los *Modelos Logísticos Lineales con una Componente No Lineal Aditiva.*

La estructura formal de un LPALM viene dada por la expresión (5.113), donde la función no paramétrica $h(\boldsymbol{V})$ de dimensión infinita consiste en una expansión aditiva de funciones de base $h_r(V_r)$ de las variables no lineales, es decir,

La función de calificación de acreditados se expresa $S(X)=\beta_0+\sum_{r=1}^{p}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por

$$h_r(\bullet):\mathbb{R}^P\to\mathbb{R},\quad r=1,..,p$$

$$X\to h_r(X)=\begin{cases} U_r & si\ \ r=1,...,p_1 \\ h_r(V_r) & si\ \ r=p_1+1,...,p \end{cases} \qquad (5.115)$$

con $h_r(V_r)$, $r=p_1+1,...,p$, función no paramétrica infinito dimensional.

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/\ X=x)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{Lineal}+\underbrace{\sum_{r=p_1+1}^{p}h_r(V_r)}_{No\ \ Lineal}=\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\mathbf{1}^T\boldsymbol{H}(\boldsymbol{V}) \qquad (5.116)$$

donde $h(\boldsymbol{V})$ es una función no paramétrica infinito dimensional y $\beta_{p_1+1}=...=\beta_p=1$ y $1^T=(1,...,1)_{1\times p_2}$.

La expresión (5.116) se corresponde con la formulación general de los *Modelos Logísticos Aditivos Parcialmente Lineales,* LPALM, que pueden verse como una extensión semiparamétrica del modelo LOGIT y que se estiman a través de la ***Regresión Logística Aditiva Parcialmente Lineal,*** **LPALR,** método cuyo núcleo operacional es

una variante del método backfitting aplicado en el Problema General de Estimación de los Modelos Logísticos Aditivos.

De nuevo se recurre a la aditividad para resolver el problema de maldición de la dimensionalidad y para conseguir un modelo más interpretable. Decíamos en la descripción de los modelos GAM, que los modelos aditivos presentan dos muy buenas cualidades, por un lado, dado que cada uno de los términos aditivos se estima usando un suavizador univariante, se esquiva la maldición de la dimensionalidad y, por otro, las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas del riesgo observadas sobre los acreditados. Pues bien estas cualidades son aplicables a la parte no lineal del modelo, la componente lineal las tiene garantizadas por construcción.

Esta segunda propiedad hace a los modelos aditivos, sobre todo a los obtenidos a través de la pérdida logística particularmente atractivos en los sistemas de calificación de acreditados en orden al cumplimiento del requerimiento de Basilea II, respecto de la interpretabilidad del modelo. De todas formas la interpretabilidad del modelo dependerá fundamentalmente de las funciones de base $h_r(V_r)$ utilizadas. Sin embargo, los modelos LPALM presentan la debilidad de que la hipótesis de aditividad es muy comprometida.

5.6.2.8 Modelos Logísticos Basados en Funciones de Proximidad, PFBLM.

Los modelos *Logísticos Basados en Funciones de Proximidad*, PFBLM, son modelos del tipo (5.1) donde la función de enlace entre la probabilidad de default y la expansión lineal de funciones de base es la *transformación logística* y las funciones de base son *funciones de proximidad* entre los acreditados y las poblaciones de default y no default.

Las funciones de proximidad entre individuos y poblaciones fueron desarrolladas y aplicadas al análisis discriminante por CUADRAS et al. (1997). Estos autores construyeron las funciones de proximidad entre un individuo y una población utilizando la matriz de distancias entre todos los individuos de la población y la *varianza geométrica*, una variante del *coeficiente de diversidad DIVC* (RAO, 1982a, 1982b).

La motivación subyacente en la construcción de las funciones de proximidad, según sus propios autores, es proporcionar un marco teórico para la regla de decisión para discriminación, *regla DB*, propuesta por CUADRAS (1989). Esta regla aplicada a

nuestro caso de dos poblaciones, Π_1, población de default, y Π_0, población de no default, se plantea en los siguientes términos:

Sea $\phi_k^2(\bullet)$, $k=0,1$, la función de proximidad para Π_k calculada usando el *cuadrado de las disimilaridades entre las observaciones de* Π_k, $\delta_k^2(\bullet,\bullet)$, *la regla BD para asignar un individuo* para el que se ha observado X_0 es

$$Se\ asigna\ X_0\ a\ \Pi_k\ si\ \phi_k^2(X_0)=mín\left\{\phi_0^2(X_0),\phi_1^2(X_0)\right\} \tag{5.117}$$

Estos autores demostraron que la regla DB, aunque se calcula sobre las disimilaridades entre observaciones, está, de hecho, basada sobre la distancia entre una observación y una "media" de la población, al menos cuando la representación Euclídea existe, es decir, es una regla Matusita, la cual asigna la observación a la población más cercana, MATUSITA (1956), KRZANOWSKI (1987).

Si como es habitual, capítulo 2, sección 2.2, se representa la población Π por un vector aleatorio X, definido sobre un espacio muestral Ω, con valores en $S\subset\mathbb{R}^p$, para algún $p\geq 1$, con función de densidad de probabilidad con respecto a una conveniente medida μ, siendo $\delta_k^2(\bullet,\bullet)$ una disimilaridad definida sobre S, CUADRAS y FORTIANA (1995) definen la V*ariabilidad Geométrica* de S con respecto a una medida adecuada μ como

$$V_\delta(X)=\frac{1}{2}\int_{S\times S}\delta^2(X,Z)f(X)f(Z)d\mu(X)d\mu(Z) \tag{5.118}$$

Cuando δ es la distancia Euclídea, se tiene $V_\delta(X)=Var(X)$, para $p=1$ y en general $V_\delta(X)=tr\left(Var(X)\right)$. Para otras disimilaridades δ, $V_\delta(X)$ es una medida generalizada de dispersión de X.

Dado $x_0\in\mathbb{R}^p$, se define la proximidad de x_0 a la población Π con respecto a δ según la expresión

$$\phi_\delta^2(X_0,\Pi)=\int_S\delta^2(X_0,X)f(X)d\mu(X)-V_\delta(X) \tag{5.119}$$

La variabilidad geométrica y la función de proximidad son cantidades que se refieren a una población. En el contexto del análisis discriminante ambas pueden obtenerse usando diferentes disimilaridades para cada una de las poblaciones.

Las funciones de proximidad pueden venir definidas por transformaciones que generan representaciones del espacio (S,δ) *en espacios de Hilbert,* $\left(H,\langle\bullet,\bullet\rangle\right)$, *es decir por*

funciones $\psi:(S,\delta)\to(H,\langle\bullet,\bullet\rangle)$. Si se asume que se verifica $\delta^2(x,z)=\|\psi(x)-\psi(z)\|^2$ y el valor esperado $E\left(\|\psi(X)\|^2\right)$ es finito, entonces se verifica

$$\begin{aligned} V_\delta(X) &= E\left(\|\psi(X)\|^2\right)-\|E(\psi(X))\|^2 \\ \phi_\delta^2(x_0,\Pi) &= \|\psi(x_0)-E(\psi(X))\|^2 \end{aligned} \tag{5.120}$$

Un ejemplo notable de función de proximidad del tipo anterior se obtiene a partir de la distancia de Mahalanobis:

Si $X\sim N(\mu,\Sigma)$ y $\delta^2(x,z)=(x-z)^T\Sigma^{-1}(x-z)$ es la distancia de Mahalanobis, entonces la transformación $\psi(x)=\Sigma^{-1}x$ que toma valores en $\mathbb{R}^p$ con su producto escalar Euclídeo ordinario, proporciona la siguiente función de proximidad

$$\phi_\delta^2(x_0,\Pi)=(x_0-\mu)^T\Sigma^{-1}(x_0-\mu) \tag{5.121}$$

A partir de (5.121) se obtiene, por ejemplo, la *clásica regla discriminante lineal:* Si Π_k es $N(\mu_k,\Sigma_k)$, $k=0,1$, con $\Sigma_0=\Sigma_1$, si consideramos la distancia de Mahalanobis, entonces

$$\begin{aligned} L(x) &= \frac{1}{2}\left[\phi_{\delta_1}^2(x,\Pi_1)-\phi_{\delta_0}^2(x,\Pi_0)\right] \\ &= \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) \\ &= A^TX+B \end{aligned} \tag{5.122}$$

donde

$$A=\left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T,\quad B=-\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$$

La función $L(x)$ es igual a la *función discriminante lineal* , LDF, (LACHENBRUCH 1975), por tanto, la regla BD (5.122) coincide en este caso con la regla LDF.

A partir de la regla LDF y de (2.14) y (2.15) se "eleva" la regla de decisión a un modelo de predicción en los términos

$$logit\left(P(Y=1/\ X=x)\right)=\text{logit}(p)+\frac{1}{2}\phi_{\delta_1}^2(x,\Pi_1)-\frac{1}{2}\phi_{\delta_0}^2(x,\Pi_0) \tag{5.123}$$

El modelo (5.123) es un modelo del tipo (5.1), que puede expresarse en la forma $logit\left(P(Y=1/\ X=x)\right)=\sum_{r=1}^{3}\beta_r h_r(X)$, donde el enlace entre la probabilidad de default y la función de calificación es la transformación logística, además la función de calificación es una expansión lineal con tres funciones de base

$$\begin{aligned} h_1(X) &= 1 \\ h_2(X) &= \phi^2_{\delta_1}(X,\Pi_1) = (X-\mu_1)^T \Sigma^{-1}(X-\mu_1) \\ h_3(X) &= \phi^2_{\delta_0}(X,\Pi_0) = (X-\mu_0)^T \Sigma^{-1}(X-\mu_0) \end{aligned} \qquad (5.124)$$

los parámetros de la expansión lineal vienen dados por $\beta_1 = \text{logit}(p)$, $\beta_2 = \frac{1}{2}$ y $\beta_3 = -\frac{1}{2}$.

Por tanto, la regla BD (5.117) coincide en este caso con la regla LDF, que viene determinada por los parámetros $(p, \mu_1, \mu_0, \Sigma)$ cuya formulación es conocida y tan sólo se han estimar sus valores a través de *estimadores de máxima verosimilitud* $\left(\hat{p}, \hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}\right)$, calculados a partir de la muestra de entrenamiento, tal como se ha visto en la sección 5.4.

Sobre la base de este enfoque del análisis discriminante a través de funciones de proximidad entre individuos y poblaciones, CUADRAS et al. (1997), es posible obtener, sin el conocimiento de la distribución de las poblaciones de default y de no default, los Modelos Logísticos Basados en Funciones de Proximidad, PFBLM, cuya estructura formal vendrá dada por la siguiente expresión

$$logit\left(P(Y=1/\ X=x)\right) = \sum_{r=1}^{3} \beta_r h_r(X) \qquad (5.125)$$

Es decir, la función de calificación es una expansión lineal con tres funciones de base. Pretendemos construir la estructura de este modelo utilizando como funciones de base funciones de proximidad, para lo cual se puede generalizar (5.123) en la forma siguiente:

Si se considera un conjunto de N individuos, y $\delta : \Omega \times \Omega \to \mathbb{R}^+$, $\delta(i,j) = \delta_{ij}$, tal que

$$\delta_{ij} \geq 0,\ \ \delta_{ij} = \delta_{ji}\ \ y\ \ \delta_{ij} \leq \delta_{ik} + \delta_{kj} \qquad (5.126)$$

y $\Delta = \left(\delta(i,j)\right)$ la matriz $N \times N$ de interdistancias; se dice que Δ es *Euclídea* si, para algún entero r se puede hallar $x_1, ..., x_N \in \mathbb{R}^r$, tal que $\left(x_i - x_j\right)^T \left(x_i - x_j\right) = \delta_{ij}$, $1 \leq i, j \leq N$.

Sean δ_0 y δ_1 dos funciones de distancia con la *propiedad Euclídea*, que pueden coincidir o no para cada población, si para $x \in \mathbb{R}^p$, se define la proximidad de x a la población Π con respecto a δ_k, $k = 0,1$, en función de la variabilidad geométrica $V_\delta(X)$, se tienen las dos funciones de proximidad $\phi^2_{\delta_0}(x,\Pi_0)$ y $\phi^2_{\delta_1}(x,\Pi_1)$, es decir, *se tienen dos*

funciones de proximidad definidas por transformaciones que generan representaciones del espacio (S,δ) *en espacios de Hilbert,* $\left(H,\ \langle\bullet,\bullet\rangle\right)$.

A partir de la funciones de proximidad $\phi^2_{\delta_0}(x,\Pi_0)$ y $\phi^2_{\delta_1}(x,\Pi_1)$, nuestra propuesta consiste en definir la probabilidad de default en los términos siguientes:

$$P(Y=1/\ X=x)=P(Y=1)\frac{e^{-\phi^2_{\delta_1}(x,\Pi_1)}}{e^{-\phi^2_{\delta_0}(x,\Pi_0)}+e^{-\phi^2_{\delta_1}(x,\Pi_1)}} \tag{5.127}$$

donde $p=P(Y=1)$ es la probabilidad de default a priori, lo que es equivalente a asumir la hipótesis de que la razón de verosimilitud, (2.6), viene dada por

$$LR(X)=\frac{f_1(X)}{f_0(X)}=e^{\left(\phi^2_{\delta_0}(x,\Pi_0)-\phi^2_{\delta_1}(x,\Pi_1)\right)} \tag{5.128}$$

Por tanto los **Modelos Logísticos Basados en Funciones de Proximidad, PFBLM**, están caracterizado por

La función de calificación de acreditados se expresa $S(X)=\sum_{r=1}^{3}\beta_r h_r(X)$, la función de enlace viene dada por $g(\bullet)=\text{logit}(\bullet)$ y las funciones de base por

$$h_r(\bullet):\mathbb{R}^P\to\mathbb{R},\quad r=1,2,3$$
$$X\to h_r(X)=\begin{cases}1 & si\ \ r=1\\ \phi^2_{\delta_0}(X,\Pi_0) & si\ \ r=2\\ \phi^2_{\delta_1}(X,\Pi_1) & si\ \ r=3\end{cases} \tag{5.129}$$

Por tanto, el modelo presenta la siguiente estructura formal

$$\text{logit}\left(P(Y=1/\ X=x)\right)=\text{logit}(p)+\phi^2_{\delta_0}(x,\Pi_0)-\phi^2_{\delta_1}(x,\Pi_1) \tag{5.130}$$

$\beta_1=\text{logit}(p),\ \beta_2=1\ y\ \beta_3=1.$

Por lo que respecta al clasificador Bayes optimal, bajo el supuesto de que un acreditado o solicitante de crédito se clasifica como default si $P(Y=1/\ X=x)\geq 0.5$, se tiene

$$\Gamma_{Bayes}(x)=\begin{cases}default & si\ \ \text{logit}(p)\geq\phi^2_{\delta_1}(x,\Pi_1)-\phi^2_{\delta_0}(x,\Pi_0)\\ no\ default & en\ \ otro\ caso\end{cases} \tag{5.131}$$

El problema se reduce entonces a obtener los estimadores

$$\hat{p},\ \hat{h}_2(X)=\hat{\phi}^2_{\delta_0}(X,\Pi_0)\ \ y\ \hat{h}_3(X)=\hat{\phi}^2_{\delta_1}(X,\Pi_1) \tag{5.132}$$

El modelo (5.129) es un modelo logístico basado en funciones de proximidad, y, por tanto, en distancias. Utilizar como funciones de base las funciones de proximidad nos permite de modo natural utilizar como variables explicativas del riesgo de crédito variables numéricas y categóricas, así como incorporar la no linealidad al modelo. Además, este modelo tiene la ventaja de que siempre es posible definir una distancia entre las observaciones correspondientes a los acreditados, por lo que siempre es planteable. Algunos autores se refieren a este tipo de modelos como no paramétricos, desde nuestra perspectiva es un modelo semiparamétrico, por cuanto, estamos haciendo suposiciones estructurales adicionales tan importantes como: que el nexo de unión entre el estado de default, variable respuesta Y, es *la transformación logística* y que, además, *las funciones de distancia que generan las funciones de proximidad son Euclídeas* en el sentido de los Escalogramas Multidimensionales Métricos, (GOWER 1982).

El modelo está orientado a la clasificación dada la dificultad de interpretación de las funciones de proximidad en términos de las variables explicativas del riesgo. Además para un número grande de acreditados, lo más usual en credit scoring , *está técnica está afectada por la maldición de la dimensionalidad*. Así, por ejemplo, en el sistema de calificación de proactivo que desarrollamos en el capítulo 7 de esta Tesis Doctoral, consideramos una muestra de entrenamiento estratificada por la variable de *incumplimiento* con sobremuestreo, con un total de 36.703 acreditados, de los cuales 2.755 pertenecen a la población de default y 33.852 son acreditados no default. Pues bien la matriz de inter-distancias entre los acreditados no default tiene 36.703^2 elementos, de los que es necesario computar $\frac{36.603^2}{2}-36.603\approx 670$ millones de términos.

BOJ et al. (2009a) aplicaron el *DBDA al credit scoring desde la perspectiva de la clasificación, utilizando la regla de asignación basada en distancias BD:* se asigna X_0 a Π_k sí $\hat{\phi}^2_k(X_0)=mín\left\{\hat{\phi}^2_0(X_0),\hat{\phi}^2_1(X_0)\right\}$, con los siguientes elementos:

1) Bajo los mismos supuestos en que CUADRAS et al. (1997) definen las funciones de proximidad, a partir de las variables explicativas $X=\left(X_1,...,X_p\right)$, calculan sobre las muestras de entrenamiento de las poblaciones de default y de no

default los estimadores de las *funciones de proximidad de un acreditado a cada una de dichas poblaciones*, $\hat{\phi}^2_{\delta_0}(x,\Pi_0)$ y $\hat{\phi}^2_{\delta_1}(x,\Pi_1)$.

2) *Estiman la probabilidad subyacente de pertenencia de un individuos a cada una de las poblaciones consideradas según la* formula

$$\hat{P}(Y=k/X=x)=\frac{e^{-\hat{\phi}^2_{\delta k}(x,\Pi_k)}}{\sum_{k=0,1} e^{-\hat{\phi}^2_{\delta_k}(x,\Pi_k)}},\quad k=0,1 \tag{5.133}$$

(realmente estos autores formulan su definición para g grupos que luego aplican a los dos usuales de credit scoring), la diferencia con nuestra propuesta (5.127) consiste en que no consideran la probabilidad de default a priori, que nosotros consideramos importante dado el papel que juega el logit de la probabilidad de default a priori, como constante aditiva, en el modelo teórico (2.10) que liga la probabilidad de default con la razón de verosimilitud.

3) C*onstruyen, de forma adaptativa, métricas bajo la hipótesis de independencia de las variables explicativas* y, alternativamente, también *familias de métricas adaptativas dependientes de parámetros*, (ESTEVE, 2003).

4) Proponen un *estadístico de bondad de ajuste* y unos *coeficientes de influencia* que permiten cuantificar la importancia relativa de los factores potenciales de riesgo.

El método posee un gran parte de las cualidades para situarse en la vía de las técnicas viables para ser utilizadas en credit scoring, de acuerdo con los requerimientos de Basilea II, pero todavía no son suficientes. Además, a pesar de que proporciona la probabilidad de default y un potente clasificador, la función de calificación y sobre todo su interpretación aún dista bastante de lo que Basilea II recomienda, sin embargo la posibilidad de cuantificar la importancia relativa de los factores potenciales de riesgo (BOJ et al. (2009b), parece apuntar en una buena dirección y no cabe duda de que merece la pena, dado el potencial del método, continuar investigando estos aspectos.

Téngase en cuenta que el análisis lineal discriminante, LDA, y el análisis discriminante cuadrático, QDA, no son más que la regresión logística en los supuestos de que las distribuciones de las funciones de verosimilitud sean normales con igual y distintas varianzas respectivamente, y que bajo los supuestos de normalidad, definidas las funciones de proximidad para la población de default y no default a partir de la distancia

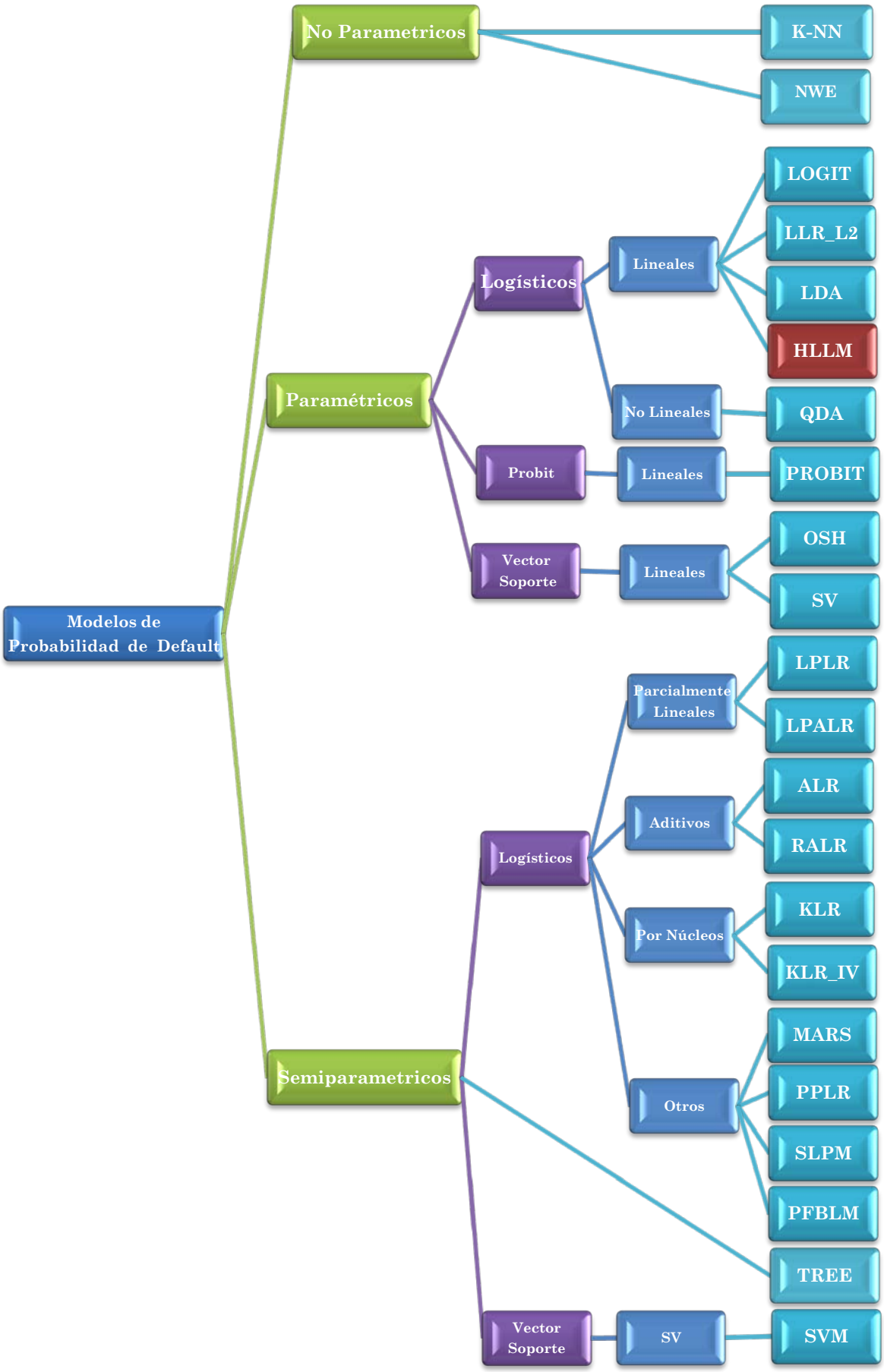
de Mahalanobis, el modelo PFBLM (5.130) coincide con la *función discriminante de Fisher* o *la función discriminante cuadrática* según que las varianzas sean iguales o distintas.

5.7 TAXONOMÍA DE LOS MODELOS DE CREDIT SCORING.

Los modelos de *credit scoring* pueden clasificarse en primer lugar de acuerdo con el conocimiento que se posee sobre su estructura funcional, (no paramétricos, paramétricos y semiparamétricos), en segundo lugar en función del nexo de unión entre el estado de default y las variables explicativas del riesgo de crédito, (Logística, Probit o Vector Soporte), en tercer lugar según el tipo de estructura lineal o no lineal de que se dote al modelo. El parámetro de regularización λ verifica $\lambda \neq 0$ o $\lambda = 0$ según que se pretenda regularizar o no el modelo. Los modelos de credit scoring tendrán distinta estructura de acuerdo con las funciones de base $h_r(X)$ que se consideren en la expansión

$$S(X) = \sum_{r=1}^{N} \beta_r h_r(X) = \boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X}).$$

Todos los modelos vistos en este capítulo tienen en común que <u>*su estructura viene caracterizada como una expansión de funciones de base*</u>, es decir su estructura funcional viene dada por (5.1). Como hemos visto, según los valores de la transformación $g(\bullet)$, tendremos distintos tipos de modelos: de ***probabilidad, logísticos, probit*** *y* ***vector soporte***.


Modelos de Probabilidad de Default
No Parametricos
K-NN
NWE
Paramétricos
Logísticos
Lineales
LOGIT
LLR_L2
LDA
HLLM
No Lineales
QDA
Probit
Lineales
PROBIT
Vector Soporte
Lineales
OSH
SV
Semiparametricos
Logísticos
Parcialmente Lineales
LPLR
LPALR
Aditivos
ALR
RALR
Por Núcleos
KLR
KLR_IV
Otros
MARS
PPLR
SLPM
PFBLM
TREE
Vector Soporte
SV
SVM

CAPÍTULO 6

MODELOS LOGÍSTICOS LINEALES HÍBRIDOS, HLLM.

6.1 INTRODUCCIÓN.

Sin duda alguna la estructura funcional más utilizada y popular en la construcción de modelos estadísticos es la lineal, es decir, aquella donde la variable respuesta o alguna de sus transformaciones se expresan en función de una combinación lineal de las variables explicativas o independientes, $g\left(P\left(Y=1/\ X=x\right)\right)=\sum_{j=1}^{p}\beta_j X_j$.

En credit scoring, la linealidad de la función de calificación de acreditados, $S(x)$, es habitualmente conveniente principalmente debido a su fácil interpretación, y además, porque es la aproximación de Taylor de primer orden de $S(x)$. Por otro lado, algunas veces es necesario, a causa de que con N grande y/o p pequeño, un modelo lineal podría tener toda la capacidad de ajustar los datos sin sobrejuste. Igualmente, en la clasificación de nuevos acreditados, una frontera de decisión Bayes optimal lineal implica que alguna transformación monótona de la probabilidad de default, $P\left(Y=1/\ X=x\right)$, es lineal en X .

No obstante los analistas de riesgos necesitan conocer a menudo de forma más precisa el modo en el que las variables explicativas X se asocian con el estado de default, y es poco realista pensar que la función de calificación sea lineal para todas las variables explicativas, en estos casos raramente se conoce una forma paramétrica de $S(x)$; por lo que se hace necesario recurrir a modelos no paramétricos o semiparamétricos.

Los métodos no paramétricos de "suavizado", o "alisamiento", proporcionan una herramienta eficaz para explorar la forma de $S(x)$ y, por tanto, de $P\left(Y=1/\ X=x\right)$, (*dejando hablar a los datos observados puede darse el ajuste adecuado para el modelo utilizando métodos de estimación que contemplen la no linealidad*), pero el precio a pagar es muy alto, pues, por un lado, las funciones de ajuste no paramétrico suelen ser de dimensión infinita fuertemente penalizadas por la maldición de la dimensionalidad y con tiempos de computación inasumibles y, por otro, la complejidad del modelo resultante, aparte de propiciar el sobreajuste perjudica la interpretabilidad del modelo, lo que va en contra de los requerimientos de Basilea II.

La predisposición a introducir en el modelo términos no lineales razonablemente interpretables viene motivada porque con frecuencia variables importantes para describir y predecir la probabilidad de default presentan realmente una importante relación no lineal con el estado de default. Si es necesario deberán introducirse estos

términos en el modelo, a pesar de que eso lo hace más complejo, por lo que, para lograr un razonable equilibrio entre la necesidad de esos términos y los requerimientos de Basilea II, no se ha de perder de vista el dilema de Occam: "*el modelo deberá ser tan complejo como sea necesario, pero no más*". Precisamente, la consecución de ese equilibrio es lo que motiva una de las aportaciones más importantes de esta Tesis Doctoral, la formalización de los **Modelos Logísticos Lineales por Expansiones Lineales Híbridas de funciones de base, HLLM**, (a los que en adelante llamaremos *Modelos Logísticos Lineales Híbridos*) modelos obtenidos como consecuencia de expandir la componente no lineal de Modelos Logísticos Parcialmente Lineales, LPLM, a través de expansiones lineales de funciones de base de las variables explicativas, de acuerdo con la sugerencia de HASTIE y TIBSHIRANI (1996).

A la hora de introducir funciones de base para captar factores no lineales es importante no olvidar que, por un lado, el problema de la no linealidad no es más que el problema de nuestro desconocimiento sobre la relación entre la variable respuesta y las variables independientes, es decir sobre la estructura del modelo y, por otro, el modelo no sólo es una herramienta de explicación, predicción y clasificación, sino que además ha de servir para calificar a los acreditados de forma comprensible, aislando el efecto que cada variable relevante para el riesgo de crédito provoca sobre tal calificación y sobre el estado de default pronosticado.

Suponer que la relación existente entre una conveniente transformación de la probabilidad de default y las variables de riesgo de crédito explicativas del estado de default es una expansión base lineal del vector de entradas X es sin duda alguna la forma más elegante, desde el punto de vista estadístico, de introducir la no linealidad en los modelos que relacionan el estado de default con las variables explicativas. Estimar $S(x)$ es ahora equivalente a encontrar los coeficientes β_r. Las funciones $h_r(\bullet)$, que pueden o no contener parámetros desconocidos, son elegidas por poseer propiedades convenientes.

La belleza de esta aproximación es que una vez que las funciones de base $h_r(x)$ *han sido determinadas, los modelos son lineales en este nuevo espacio expansionado de las características de entrada resultante, y el ajuste se hace como para los modelos lineales,* (HASTIE et al. 2009).

A lo largo de las secciones 5.2 a 5.6 del capítulo 5 se ha realizado un recorrido por los distintos tipos de modelos que se han usado o se usan con mayor o menor acierto para representar la relación de dependencia entre la variable estado de default y las variables relevantes de riesgo de crédito. De este modo hemos analizado las características más sobresalientes, sus fortalezas y debilidades, sobre todo en relación con la estimación de la probabilidad de default desde la óptica de los acuerdos de Basilea II, de un buen número de técnicas que, sin constituir un conjunto exhaustivo, creemos que representan adecuadamente todo o casi todo el espectro de técnicas que se utilizan para construir *modelos de credit scoring*.

Cabe destacar, en primer lugar, que a excepción de los estimadores no paramétricos de la probabilidad de default, *k-vecinos más próximos*, K_NN, y el *estimador de Nadaraya_Watson*, NWE, los métodos que hemos analizado en el capítulo 5, abordan el problema de encontrar la estructura formal más adecuada para el modelo desde distintas estrategias respecto de construir la expansión lineal de funciones de base, pero la mayoría de ellas tienen en común que las funciones de base utilizadas poseen todas la misma estructura. Sin embargo, *no hay ninguna razón para el supuesto de homogeneidad en las funciones de base*. Es natural pensar que algunas variables provocan efectos lineales sobre la expansión lineal y otras efectos no lineales, y debemos tener presente que *la no linealidad puede manifestarse de infinitas formas en los espacios de representación más usuales, como son los espacios de Hilbert. Parece por tanto razonable que busquemos para cada variable con linealidad no significativa y con no linealidad confirmada por las técnicas apropiadas para ello, la estructura funcional específica más adecuada*.

Por otra parte, todas las técnicas desde las que producen los modelos más rugosos y sobre ajustados, k_NN, hasta los modelos más parsimoniosos, amigables y fáciles de interpretar, los modelos lineales, LOGIT y PROBIT, tienen en mayor o menor medida sus fortalezas y debilidades, tal como hemos comentado. Así, por ejemplo, los primeros sobre ajustan los datos, además de no tener en cuenta la distribución de la variable respuesta, de Bernouilli de parámetro la probabilidad de default a priori, y los segundos no tienen en cuenta la posible *no linealidad* de algunas de las variables explicativas del riesgo de crédito, defecto importante en todas las técnicas que trabajan sobre la hipótesis de linealidad de todas las variables, LLR o LOGIT, LLR_L2, LDA, LPR o PROBIT, OSH y VS.

Muchas de las técnicas proporcionan excelentes clasificadores de los acreditados y de los solicitantes de crédito, como es el caso del LDA, cuyo problema es que exige normalidad multivariante, hipótesis de cumplimiento muy improbable en observaciones de riesgo de crédito, y de los métodos de Vector Soporte, lineal, SV, o por núcleos, SVM, cuyo defecto insalvable es que no son capaces de estimar la probabilidad subyacente, aspecto capital requerido por los acuerdos de Basilea II. Incluso las técnicas propuestas para corregir el problema que presentan las técnicas vector soporte, la Regresión Logística por Núcleos, KLR, y KLR_IV, que tienen en su haber una capacidad de clasificación similar a SVM , maximizan el margen igual que esta, tienen la debilidad de construir el modelo a partir de los acreditados y no de las variables del riesgo de crédito, aspecto este fundamental para conocer la relación entre estas y el estado de default, tal como requieren los acuerdos Basilea II.

Además, por un lado, las técnicas TREE y ALM proporcionan una gran facilidad interpretativa, aunque a costa de estimar una probabilidad constante a trozos la primera, situación que dista mucho de un apropiado concepto de probabilidad, y la segunda lo consigue a costa de la hipótesis muy ingenua y poco realista de la aditividad de las funciones de base. Por otro lado, las técnicas PPLR y BIMARS proporcionan estructuras muy complejas y SLPM forma parte de la familia de "*modelos caja negra*", para nada compatibles con Basilea II.

Ante esta situación, es necesario contar con técnicas alternativas con al menos las siguientes cualidades, combinación de propiedades adecuadas desde el punto de vista del riesgo de crédito, desde los requerimientos de Basilea II y desde el rigor estadístico:

a) Desde el punto de vista del riesgo de crédito y los requerimientos de Basilea II, *el modelo deberá responder con la suficiente calidad al triple objetivo perseguido en credit scoring: estimar las probabilidades de default, estimar la función de calificación de los acreditados, (calidad predictiva), clasificar a los nuevos solicitantes de crédito dentro de una de las poblaciones, default y no default (calidad discriminante) y poseer capacidad explicativa.*

   La combinación de los tres requerimientos del párrafo anterior determina todas las propiedades que ha de tener el modelo más idóneo para cada situación concreta.

- Equilibrio entre la capacidad del modelo para describir situaciones de naturaleza diferente, (por ejemplo, linealidad y no linealidad total o parcial de la relación entre el estado de default y las variables de riesgo de crédito).
- La complejidad, (el modelo ha de ser tan sencillo como sea posible, pero no más), y la dimensión del mismo.
- El modelo ha de extender bien tal relación de dependencia (generalización) y clasificar correctamente a nuevos acreditados o solicitantes de crédito distintos a los de entrenamiento.

b) El rigor estadístico nos induce a considerar prioritariamente la función de distribución logística como *enlace entre la probabilidad de default y la función de calificación de acreditados.*

Las condiciones anteriores nos sitúan en una clase de modelos donde se combinan ideas de Modelos Logísticos Lineales, LLM, (mejor aproximación a la relación de dependencia teórica entre la variable respuesta y las variables explicativas y alto grado de interpretabilidad), Modelos Logísticos Aditivos Parcialmente Lineales, LPALM, (que conservando la interpretabilidad introducen la flexibilidad necesaria para, por ejemplo, contemplar la no linealidad), y de modelos logísticos expansiones lineales de funciones de base de X, cuya característica más destacable es que es posible asignar a cada variable no lineal combinaciones lineales de funciones de base para especificar la no linealidad. Una vez que las funciones de base, nuevas variables, han sido determinadas, los modelos son lineales en estas nuevas variables, resultando una familia de modelos logísticos lineales dentro de la cual situamos nuestra propuesta de *Modelos Logísticos Lineales por expansiones lineales Híbridas de funciones de base,* HLLM.

En este capítulo desarrollamos la formalización de la estructura funcional de los modelos HLLM, así como de la estimación de los mismos junto al análisis, con cierto nivel de detalle, de algunas de las funciones base que nos parecen más apropiadas para describir la no linealidad en modelos de credit scoring, desde la óptica de Basilea II.

6.2 ESTRUCTURA FUNCIONAL Y ESTIMACIÓN DE LOS MODELOS DE PROBABILIDAD LINEAL GENERALIZADOS, GLPM.

Comenzamos por establecer tres hipótesis muy generales sobre las que asentar los *Modelos de Probabilidad Lineales Generalizados por expansión de funciones de base,* GLPM, familia a la que pertenecen los modelos de nuestro interés, HLLM. En base a esas tres hipótesis y los objetivos que perseguimos plantearemos la estructura funcional de dichos modelos y una metodología apropiada para estimarlos. Estas hipótesis están inspiradas, por un lado, en los Modelos de Probabilidad Generalizados, GPM, por otro, en las expansiones lineales por funciones de base y, por último, en el hecho de que para estimar los modelos de credit scoring se cuenta casi siempre con información limitada e imperfecta, lo que conduce al *principio de inducción.*

Hipótesis H1.- *El modelo expresa la relación existente entre una conveniente transformación de la probabilidad de default,* $g(\bullet)$ *con las propiedades necesarias para asegurar que* $P(\bullet)=g^{-1}(S(X))$ *es una función de probabilidad, y la función de calificación de acreditados* (*función de las variables de riesgo de crédito explicativas del estado de default*):

$$g\left(P\left(Y=1/\ X\right)\right)=S\left(X\right) \tag{6.1}$$

Con esta hipótesis se pretende que nuestro modelo pertenezca a la familia de los modelos de probabilidad generalizados, GPM, puesto uno de nuestros objetivos consiste en que el modelo nos proporcione la probabilidad de default.

Hipótesis H2.- L*a función de calificación de acreditados,* $S(X)\in F$ *, es una expansión lineal de funciones de base del vector de variables explicativas* X .

Siendo $h_r(X):\mathbb{R}^p\rightarrow\ \mathbb{R}$, $r=1,..,q$, una función de base de X, entonces el modelo se modifica a la forma más general

$$g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right) \tag{6.2}$$

donde $\boldsymbol{X}=\left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta}=\left(\beta_1,...,\beta_q\right)^T$ y $\boldsymbol{H}\left(\boldsymbol{X}\right)=\left(h_1\left(X\right),...,h_q\left(X\right)\right)^T$.

Desde este punto de vista, el mejor acontecimiento en la especificación del modelo consiste en encontrar las adecuadas funciones de base $\left(h_1\left(X\right),...,\ h_q\left(X\right)\right)^T$.

El objetivo de esta hipótesis consiste en construir modelos más flexibles, por ejemplo, que contemplen la no linealidad, aumentando o reeplazando el vector de variables explicativas originales con variables adicionales, transformaciones de X, tales como, pesos de la evidencia, polinomiales, splines cúbicos restringidos, RCS, funciones constantes a trozos resultado de particiones recursivas, funciones lineales a trozos, funciones bisagra MARS, funciones de base radial Gaussiana, etc. y usar modelos lineales en este nuevo espacio expansionado de las variables de entrada resultantes.

Esta hipótesis es clave, según los requerimientos de Basilea II, por cuanto facilita la interpretación del modelo, y además nos permite establecer una metodología común para una amplia variedad de modelos que recorre todo el espectro desde los más antíguos y amigables *paramétricos*, FISHER (1936), los más estudiados en el siglo XX, pasando por los menos "suaves", los *no paramétricos*, FIX Y HODGES (1951), a los *semiparamétricos*, los más realistas que, iniciados por LEONTIEF (1947), se vienen desarrollando en toda su plenitud en esta primera década del siglo XXI.

De acuerdo con el planteamiento que venimos haciendo, abordaremos directamente la estimación de los modelos generales del tipo (6.2), es decir, *abordaremos el problema de estimación en los espacios "agrandados" de las expansiones por funciones de base de las variables explicativas, que generalmente son espacios de Hilbert*.

Es en presencia de no linealidad cuando la *hipótesis* H2 se hace más necesaria y adquiere más significado, puesto que el hecho de que *la función de calificación* $S(X)$ *sea una expansión lineal de funciones de base de las variables explicativas del riesgo de crédito* permite seguir manteniendo la estructura del modelo bajo el cumplimiento de los requerimientos de Basilea II.

Hipótesis H3.- *Para ajustar el modelo, consideraremos como hipótesis general que los datos son finitos e imperfectos y la información que nos proporcionan es limitada, por lo que el Principio de Inducción constituye un método adecuado de estimación del modelo.*

Una vez fijada la estructura formal del modelo, hipótesis H1 y H2, $g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(X\right)$, es necesario establecer un método para estimar la expansión lineal $\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(X\right)$. Si se conoce alguna de distribuciones del vector aleatorio de variables explicativas del riesgo X y la variable estado de default Y, $P(X,Y)$,

$P(Y/X)$, o bien $\log(LR(X))$, el problema de estimar con función de pérdida $\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$el modelo (6.2) puede resolverse fácilmente, pero en la realidad lo frecuente es no conocer tales funciones de probabilidad, lo que implica que no podamos computar directamente el riesgo esperado asociado a $\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$, (esperanza matemática de la función de pérdida) y, por tanto, para resolver el siguiente problema de optimización

$$\min_{\beta_0,\boldsymbol{\beta}} E_{Y/X}\left[\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)/X\right] \tag{6.3}$$

donde minimizar la expresión $E_{Y/X}\left[\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)/X\right]$ es equivalente a minimizar $\mathrm{R}_\ell\left(\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)\right)$, será necesario disponer de una muestra aleatoria del vector (X,Y), $\boldsymbol{\tau}_e=\left\{(x_i,y_i)\in\mathbb{R}^p\times\mathbb{R}\right\}_{i=1,\ldots,N}\in(X\times Y)^N$.

Es importante configurar la muestra de entrenamiento teniendo en cuenta que el proceso completo que vamos a realizar consta fundamentalmente de tres fases: *fase de estimación del modelo, fase de evaluación y selección del modelo* y, por último, *fase fijación del modelo o validación de su generalización.*

La solución al problema (6.3) se obtiene resolviendo el *Problema de Minimización del Riesgo Empírico* sobre la muestra de entrenamiento $\boldsymbol{\tau}_e$, que adopta la forma

$$\underset{\beta_0,\boldsymbol{\beta}}{Mín}\left[\sum_{i=1}^{N}\ell\left(y_i,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(x_i)\right)/\tau_e\right] \tag{6.4}$$

Cuando la información es incompleta es razonable pensar que con las suficientes expansiones por funciones de base, los datos pueden ser casi perfectamente ajustados, aunque sería más apropiado decir interpolados, pero esto puede conducir a un ajuste artificial y al sobreajuste. Por esta razón, un punto fundamental en el empleo de expansiones de funciones de base consiste en controlar la complejidad del modelo. Como hemos visto en la sección 3.3, una forma de conseguir tal control es definiendo una *funcional de regularización o suavizado* $J\left(\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$ *en una vía tal que valores pequeños de la funcional correspondan a funciones suavizadas.*

Puede lograrse un punto de equilibrio con una expansión de funciones de base estrechamente ajustada a los datos y a la vez suave. De acuerdo con nuestra hipótesis, el *estimador optimo de* $S(\boldsymbol{X})=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)$, para el término de regularización

$$\lambda J\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right) \tag{6.5}$$

se obtiene minimizando la expresión (3.54),

$$\underset{\beta_0,\boldsymbol{\beta}}{Mín}\left[\sum_{i=1}^{N}\ell\left(y_i,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{x}_i)\right)\right]+\lambda J\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right) \tag{6.6}$$

Una forma alternativa de plantear la estimación de $S(\boldsymbol{X})$ ante el desconocimiento de su estructura, sobre todo cuando el desconocimiento es total, por lo que nos encontramos ante un modelo no paramétrico puro, consiste en estimarlo localmente, es decir, utilizando funciones de pérdida local. En este caso el *Problema de Minimización del Riesgo Empírico Local*, (3.29), para una métrica $m_h(x,x_0)=\|x-x_0\|$, que especifica una vecindad local en un punto x_0, $V_{mh}(x_0)=\{x\in\mathbb{R}^p/\|x-x_0\|\le h\}$, siendo h la llamada ventana de la vecindad viene dada en la forma

$$\underset{\beta_0(\boldsymbol{x}_0),\boldsymbol{\beta}(\boldsymbol{x}_0)}{Mín}\left[\sum_{i=1}^{N}m_h(\boldsymbol{x}_i,\boldsymbol{x}_0)\ell\left(y_i,\beta_0(\boldsymbol{x}_0)+\boldsymbol{\beta}(\boldsymbol{x}_0)^T\boldsymbol{H}(\boldsymbol{x}_i)\right)\right] \tag{6.7}$$

Se puede intentar suavizar la necesariamente rugosa estructura de una expansión de funciones de base estimada localmente introduciendo en la función objetivo del problema de optimización una funcional de suavizado, lo que conducirá al problema

$$\underset{\beta_0(\boldsymbol{x}_0),\boldsymbol{\beta}(\boldsymbol{x}_0)}{Mín}\left[\sum_{i=1}^{N}m_h(\boldsymbol{x}_i,\boldsymbol{x}_0)\ell\left(y_i,\beta_0(\boldsymbol{x}_0)+\boldsymbol{\beta}(\boldsymbol{x}_0)^T\boldsymbol{H}(\boldsymbol{x}_i)\right)\right]+\lambda J\left(\boldsymbol{\beta}(\boldsymbol{x}_0)^T\boldsymbol{H}(X)\right) \tag{6.8}$$

Según la *transformación elegida*, $g\left(P(Y=1/X=x)\right)$, las *funciones de base* $h_r(X)$ *seleccionadas*, la *función de pérdida considerada* $\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$, *la métrica* $m_h(x,x_0)$ *en el caso de modelos locales* y, si procede, el término de regularización $\lambda J\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right)$ utilizado tendremos distintos estimadores de los que recogen la hipotética relación entre el estado de default y las variables explicativas.

Por tanto, de acuerdo con las tres hipótesis anteriores, para fijar la estructura y estimar generales de probabilidad de default del tipo (6.2), deberemos hacer dos elecciones en

relación con la estructura de la expansión de funciones de base y tres sobre la función objetivo a optimizar por el principio de inducción. A continuación comentaremos brevemente los cinco elementos.

1) **En relación con la estructura funcional del modelo**.

1a).- *Elección de la transformación* $g(\bullet)$, nexo de unión entre la probabilidad de default y la expansión lineal de funciones de base $S(\boldsymbol{X})$. Las tres transformaciones más utilizadas en credit scoring, logística, probit y vector soporte, se muestran en la tabla 2.2, y dan lugar a los modelos logísticos, probit y vector soporte por expansión lineal de las variables originales a través de funciones de base,

$$\left.\begin{array}{ll} \text{Probabilidad} & P(Y=1/\,X=x) \\ \text{Logístico} & \text{logit}\left(P(Y=1/\,X=x)\right) \\ \text{Probit} & \text{probit}\left(P(Y=1/\,X=x)\right) \\ \text{Vector Soporte} & Sign\left\{P(Y=1/\,X=x)-\frac{1}{2}\right\} \end{array}\right\} = \beta_0 + \sum_{r=1}^{q}\beta_r h_r(X) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X}) \qquad (6.9)$$

donde $h_r(X)$, $r=1,..,q$, es una función de base, $\boldsymbol{X}=\left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta}=\left(\beta_1,...,\beta_q\right)^T$ y $\boldsymbol{H}(\boldsymbol{X})=\left(h_1(X),...,h_q(X)\right)^T$.

A diferencia de los modelos logísticos y probit, los modelos vector soporte no son modelos GLPM, puesto que $P(\bullet)=Sign^{-1}(\bullet)$ *no es una función de probabilidad,* es decir, los modelos vector soporte no proporcionan la probabilidad de default.

Como venimos insistiendo, debido a que la variable respuesta es binaria el nexo natural que une el modelo "teórico" de las variables explicativas con la variable respuesta es la *transformación logística*. Por esta razón *primaremos en esta Tesis Doctoral los modelos logísticos en sentido amplio y utilizaremos los modelos probit y los de vector soporte como técnicas de apoyo o contraste*. En consonancia con este planteamiento analizaremos de forma destacada los modelos logísticos.

1b).- *Selección de las funciones de base* $h_r(X)$ que formarán parte de la expansión lineal por funciones de base de X para obtener el modelo. Estas funciones son elegidas, en primer lugar, en función del conocimiento que se posee sobre la relación entre el estado de default y las variables explicativas del riesgo y, en segundo lugar, por mostrar

propiedades convenientes, tanto estadísticas como desde el punto de vista del riesgo de crédito. En la sección (6.7) se presentan de forma detallada algunas de las funciones de base más notables que utilizaremos en la construcción de nuestro modelo de credit scoring proactivo.

Generalmente el problema consiste en añadir o no la no linealidad, es decir, *aumentar o reemplazar el vector de variables explicativas X con variables adicionales, transformaciones de X.* Algunas veces, las menos, el problema se resuelve con sencillas funciones de base $h_r(X)$, tales como funciones logaritmo o potencia. Más frecuentemente, sin embargo, se usan las expansiones por funciones de base como un apoyo para alcanzar representaciones más flexibles para una parte del modelo. La menor o mayor necesidad de representaciones flexibles viene dada en función del grado de conocimiento sobre la relación de las variables explicativas con el estado de default.

2) **En relación con la función objetivo a optimizar.**

2.a).- <u>*Selección de una conveniente función de pérdida*</u> $\ell\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$.

Esta es otra elección fundamental por cuanto condicionará el método de ajuste del modelo a los datos, resultando técnicas distintas en función de esta elección.

La función de pérdida ha de ir pareja a la elección de la transformación que liga la probabilidad de default con la función de calificación de acreditados, por ejemplo con la función de enlace en el caso de los modelos lineales generales. De hecho, hemos clasificado los modelos de mayor interés en sistemas de calificación de acreditados como, logísticos, probit y vector soporte, de acuerdo con la función de pérdida, tabla 3.1, clasificación que coincide con la realizada con la transformación $g(\bullet)$: <u>*modelos logísticos*</u>, <u>*modelos probit*</u> y <u>*modelos vector soporte*</u> .

La función de pérdida es, por tanto, una herramienta clave para el ajuste del modelo, puesto que a partir de ella se obtiene la función objetivo a minimizar para ajustar el modelo a los datos de entrenamiento.

Una vez fijada la función de pérdida queda determinado el riesgo empírico, que para una muestra τ y las funciones de pérdida más notables, se recoge en la tabla 3.2, en el caso de los modelos globales.

2.b).- *En el caso de optar por modelos locales se deberá seleccionar una métrica adecuada,* $m_h(\boldsymbol{x},\boldsymbol{x}_0)$, *para la formulación del Problema de Minimización del Riesgo Regularizado Local cuya solución corresponde al modelo buscado.*

Cuando se deconoce totalmente la estructura formal del modelo, se puede optar por el ajuste local, aunque no es la opción más usual, minimizando el riesgo empírico local, que para las funciones de pérdida más notables se muestra en la tabla 3.3.

2.c).- *Elección del término de regularización* $\lambda J(\boldsymbol{S})$.

Una forma de solucionar el problema del desconocimiento de la estructura de la función de calificación de acreditados $S(\boldsymbol{X})$ consiste en plantear el problema desde el principio variacional que subyace en la teoría de la regularización, la cual además del ajuste de datos contempla información a priori de suavizado. Con ello, además, resolvemos problemas como el sobreajuste o la no unicidad de la solución, y suele dar muy buenos resultados cuando el número de variables es muy elevado en relación con el número de acreditados. En el caso de optar por modelos no regularizados, el parámetro de regularización λ será cero.

En la tabla 6.1 se muestra la función objetivo a minimizar para estimar los modelos de expansiones lineales de funciones de base de X para las funciones de pérdida más notables, incluidos los modelos no regularizados sin más que asignar al parámetro de regularización λ el valor cero.

Tabla 6.1- Función objetivo a optimizar en modelos regularizados para las funciones de pérdida más usuales, con $\mathrm{S}(X) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)$.

Pérdida	Función Objetivo de Optimización
Cuadrática	$\left(\boldsymbol{Y} - \left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right)^T \left(\boldsymbol{Y} - \left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right) + \lambda J\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right)$
Logística	$-\left[\boldsymbol{Y}^T \left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right) - \mathbf{1}^T \log\left(1 + \exp\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right)\right] + \lambda J\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right)$
Probit	$-\left[\boldsymbol{Y}^T \log\left(\boldsymbol{\Phi}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right) + (\mathbf{1} - \boldsymbol{Y})^T \log\left(1 - \boldsymbol{\Phi}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right)\right] + \lambda J\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right)$
SVM	$\left[\mathbf{1}^T \left(1 - \boldsymbol{Y}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)\right)\right]_+ + \lambda J\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right)$

donde $\boldsymbol{X} = \left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta} = \left(\beta_1,...,\beta_q\right)^T$, $\boldsymbol{H}\left(\boldsymbol{X}\right) = \left(h_1\left(X\right),...,h_q\left(X\right)\right)^T$, $\mathbf{1} = (1,...1)^T$ e

$\mathbf{Y} = (y_1,...y_N)^T$.

6.3 ESTRUCTURA FUNCIONAL DE LOS MODELOS HLLM.

Fijar la estructura funcional del modelo consiste en 1a) *elegir la transformación* $g(\bullet)$ y 2b) s*eleccionar las funciones de base* $h_r(X)$ que formarán parte de la expansión lineal por funciones de base de X $\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}\left(\boldsymbol{X}\right)$. En el capítulo 2, sección 2.3, como consecuencia de (2.12) se ha destacado el hecho de que *la probabilidad de default enlaza con las características de riesgo de crédito* $\left(X_1,...,X_p\right)$*, a través de la función de distribución logística.* Tal resultado, procedente del teorema de Bayes, nos orienta claramente sobre *la transformación* $g(\bullet)$ para configurar la estructura de los modelos de estimación de la probabilidad de default, y eso es justo lo que queremos, un modelo que estime la probabilidad de default. Por tanto nuestra primera elección consiste en que *la función de enlace entre la probabilidad de default y la función de calificación de acreditados,* $\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}\left(\boldsymbol{X}\right)$*, es la transformación logística.* Por tanto, por la hipótesis H2 , tenemos

$$\text{logit}\left(P\left(Y = 1 / X = x\right)\right) = \beta_0 + \sum_{r=1}^{q} \beta_r h_r\left(X\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}\left(\boldsymbol{X}\right) \qquad (6.10)$$

La estructura funcional (6.10) corresponde a un *modelo logístico de probabilidad por expansiones lineales de funciones de base y* quedará perfectamente determinada en cuanto se resuelva 1b), es decir en cuanto se especifiquen las *funciones de base* $h_r(X)$, componentes de $\boldsymbol{H}\left(\boldsymbol{X}\right)$.

Dado que nuestro interés principal se centra en que el modelo contemple la no linealidad de las variables que proceda, partimos del supuesto de que se dispone de $p = p_1 + p_2$ variables, de las cuales p_1 tienen una influencia lineal sobre el comportamiento frente al default y se desconoce el tipo de influencia que ejercen las restantes p_2 variables, la estructura formal más general del modelo (6.10), en este caso, corresponde al ***modelo logístico parcialmente lineal,*** **LPLM**, extensión semiparamétrica del modelo LOGIT y suma de una componente lineal y una componente no lineal, que se expresa en la forma, (5.113),

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{\textit{Lineal}}+\underbrace{h\left(V_1,...,V_{p_2}\right)}_{\textit{No Lineal}}=\boldsymbol{\beta}^T\boldsymbol{U}+h\left(\boldsymbol{V}\right) \tag{6.11}$$

donde $h\left(\boldsymbol{V}\right)$ es una función no paramétrica de dimensión infinita.

En el modelo (6.11) la parte lineal es la combinación lineal de las variables con significación lineal $U=\left(U_1,...,U_{p_1}\right)^T$ y la componente no lineal es una expansión no paramétrica de las variables no lineales $V=\left(V_1,...,V_{p_2}\right)^T$, $h\left(\boldsymbol{V}\right)$, donde $h\left(\bullet\right)$ es una función de suavizado no paramétrica de dimensión infinita, y se calcula de una manera flexible, por ejemplo, cualquier método de suavizado no paramétrico.

Los modelos del tipo (6.11) son más flexibles que los modelos logísticos lineales y, a pesar de tener menor complejidad técnica y más facilidad interpretativa que los modelos totalmente no paramétricos, distan mucho de cumplir el requerimiento de que el modelo sea interpretable, por cuanto la parte no lineal es completamente no paramétrica.

Una familia de modelos notables con la estructura (6.11) es la propuesta por MÜLLER (2000, 2001) y MÜLLER y HÄRDLE (2003), **Modelos Logísticos Parcialmente Lineales, LPLM**, apartado 5.6.2.7 del Capítulo 5, a la que también nos hemos referido en el apartado 1.3.2.1 del Capítulo 1, que, como decíamos allí, sus autores propusieron como una herramienta exploratoria de la no linealidad más que como un modelo predictivo del estado de default. La razón es que la parte no lineal se estima a través de una función no paramétrica, infinito dimensional, que generalmente no permite explicar la aportación de cada variable explicativa del riesgo de crédito al estado de default, además, como ocurre con casi todos los métodos no paramétricos, estas técnicas están aquejadas de la maldición de la dimensionalidad, por lo que no es una técnica acorde con los requerimientos de Basilea II.

Con el fin de de resolver los problemas anteriores surgió una peculiar subfamilia de los modelos logísticos parcialmente lineales, los ***Modelos Logísticos Aditivos Parcialmente Lineales*****, LPALM**, , cuya estructura formal viene dada por la expresión (5.116), donde la función no paramétrica $h\left(\boldsymbol{V}\right)$ de dimensión infinita se sustituye por una expansión aditiva de funciones de base $h_r\left(V_r\right)$, es decir,

$h(\boldsymbol{V}) = \sum_{r=1}^{P_2} h_r(V_r)$. La componente no lineal se constituye así en una estructura aditiva de los efectos no lineales. El modelo logístico en este caso, también semiparamétrico, adopta entonces la forma

$$\text{logit}\left(P(Y=1/\ X=x)\right) = \underbrace{\beta_0 + \sum_{r=1}^{p_1} \beta_r U_r}_{\textit{Lineal}} + \underbrace{\sum_{r=p_1+1}^{p} h_r(V_r)}_{\textit{No Lineal}} = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{H}(\boldsymbol{V}) \tag{6.12}$$

donde $\boldsymbol{H}(\boldsymbol{V}) = \left(h_1(V_1),...,h_{p_2}(V_{p_2})\right)$ y $1^T = (1,...,1)_{1\times p_2}$.

Bajo la formulación general (6.12) se pueden construir modelos tan complejos como se necesite, puesto que en un planteamiento general las funciones de base que forman la componente no lineal son cualesquiera funciones de X, por lo que muy bien podrían jugar un papel adicional las variables consideradas en la componente lineal, por ejemplo, mediante interacciones entre sí o con las variables con influencia no lineal. Un ejemplo podría ser el modelo siguiente

$$\text{logit}\left(P(Y=1/\ X=x)\right) = \underbrace{\beta_0 + \sum_{r=1}^{p_1} \beta_r U_r}_{\textit{Lineal}} + \underbrace{\sum_{r=p_1+1}^{p_2} S_{r-p_1}(V_{r-p_1}) + \sum_{\substack{r=p_2+1 \\ j,l=1,...,p_2 \\ j\neq l}}^{q} \beta_r (U_j U_l)}_{\textit{No Lineal}} \tag{6.13}$$

donde,

$$h_r(\bullet): \mathbb{R}^p \to \mathbb{R}, \quad r=1,..,q$$

$$X \to h_r(X) = \begin{cases} U_r & si\ r=1,...,p_1 \\ S_{r-p_1}(V_{r-p_1}) & si\ r=1,...,p_2 \\ U_j U_l\ ,\ j\neq l & si\ r=p_2+1,...,q\ ,\ j=1,...p_1,\ l=1,...,p_1 \end{cases} \tag{6.14}$$

donde $S_r(V_r)$ son splines y el vector $\boldsymbol{\beta}$ viene dado por $\boldsymbol{\beta} = \left(\beta_1,...,\beta_{p_1},\beta_{p_2},....,\beta_q\right)$. Estamos de este modo ante un particular modelo logístico semiparamétrico parcialmente lineal, la parte lineal es paramétrica y la no lineal tiene una parte no paramétrica, los splines, y otra paramétrica, la combinación lineal de la interacciones entre las variables con especificación lineal. Pero estos modelos no responden a nuestro concepto de sencillez e interpretabilidad en línea con los requerimientos de Basilea II referidos a la complejidad del modelo.

Como también comentamos en el Capítulo 1, apartado 1.3.2.5, LIU y CELA (2007) compararon los *resultados del modelo LPALM con el modelo LLR*. Su planteamiento es

también un caso particular de (6.12), donde la componente no lineal es una *expansión aditiva de splines cúbicos naturales,* (uno por cada variable no lineal),

$$\sum_{r=p_1+1}^{p} h_r(X) = \sum_{j=p_1+1}^{p}\left(\sum_{k=1}^{N+2}\beta_k h_{jk}(X)\right) \tag{6.15}$$

donde $\beta_r = 1,\ \ r = p_1+1,...,p$, N es el número de observaciones y las funciones base se expresan en la forma

$$\begin{aligned} &h_{jk}(\bullet):\mathbb{R}^p \to \mathbb{R}, \quad j = p_{1+1},..,p_1+p_2 \\ &X \to h_{jk}(X) = \begin{cases} \left|X_j - \xi_k\right|^3 & si\ k = 1,...,N \\ 1 & si\ k = N+1 \\ X_j & si\ k = N+2 \end{cases} \end{aligned} \tag{6.16}$$

donde se dota a los coeficientes de las restricciones apropiadas para que el spline tenga segunda derivada cero fuera del intervalo $\left[x_1, x_N\right]$.

La componente no lineal es entonces

$$\sum_{j=p_1+1}^{p}\left(\sum_{k=1}^{N}\beta_k\left|X_j - \xi_k\right|^3 + \beta_{j(N+1)} + \beta_{j(N+2)}X_j\right) \tag{6.17}$$

Nótese que el spline cúbico natural añade el término intercepto, $\beta_0 = \sum_{j=p_1+1}^{p}\beta_{j(N+1)}$, y un término lineal para la variable X_j, $\sum_{j=p_1+1}^{p}\beta_{j(N+2)}X_j$. De (6.17) se sigue que el modelo (6.12) se expresa entonces

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right) = \sum_{j=p_1+1}^{p}\beta_{j(N+1)} + \sum_{j=1}^{p_1}\beta_r X_j + \sum_{j=p_1+1}^{p}\beta_{N+2}X_j + \sum_{j=p_1+1}^{p}\left(\sum_{k=1}^{N}\beta_k\left|X_j - \xi_k\right|^3\right) \tag{6.18}$$

Si bien es cierto que utilizar splines cúbicos naturales resuelve el problema de la elección de la localización de los nodos, puesto que fijan los nodos en los N puntos $\{x_i\}_{i=1,...,N}$, no es menos cierto que esto puede conllevar problemas de sobreajuste, debido al elevado número de nudos, que no siempre se resuelven satisfactoriamente aplicando una penalización "ondulante" a la función objetivo del ajuste.

Pero como los mismos autores reconocen, el modelo (6.18) generaliza pobremente, como suele ocurrir en las técnicas que consiguen un elevado ajuste en el entrenamiento, con el consabido riesgo de admisión de solicitudes de crédito de alto riesgo de incumplimiento de

las obligaciones de pago, a la vez que no es un modelo adecuado para estimar la probabilidad de default, lo que afectará a los requerimientos de capital económico según Basilea II.

Como alternativa, los autores aconsejan utilizar el tramado de variables para aproximar el efecto no lineal una vez descubierto por LPALM, en este sentido se suman al planteamiento de MÜLLER et al. (2003) de considerar esta técnica como un método exploratorio y especificar la no linealidad con otro método que permita una interpretación más sencilla de la importancia relativa de cada factor de riesgo en la relación con el estado de default. Por ejemplo, un efecto no lineal en LPALM se puede categorizar incorporando conocimiento de riesgo de crédito y a continuación una vez sustituidas estas variables por las nuevas categóricas ajustar un modelo LLR. Puntualizan que "al hacerlo así, la interpretabilidad puede ser mejorada, si bien pagando el precio de un menor rendimiento predictivo". Es decir, proponen utilizar el procedimiento de FAIR ISAAC y SIDDIQI pero sólo para aquellas variables para las que la técnica LPALM descubra no linealidad subyacente.

El procedimiento de FAIR ISAAC – SIDDIQI (SIDDIQI, 2006), Capítulo 1, apartado 1.3.2.4, consiste en el desarrollo de todo el modelo, componentes lineal y no lineal, usando como funciones de base de las características *agrupaciones óptimas de atributos* de las mismas, (automáticas o supervisadas con criterios de riesgos), y estimar el modelo por el problema general de estimación de modelos logísticos expansión lineal de funciones de base. El modelo se plantea en términos de la transformación logística y su estructura funcional viene dada por

$$\operatorname{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{j=1}^{p}\beta_j\left(\sum_{k=1}^{K_j}I_{\left[X_{ij}\in R_{jk}\right]}WOE\left(R_{jk}\right)\right) \tag{6.19}$$

donde $1+\sum_{j=1}^{p}K_j=q$.

Esta método encaja bien con los requerimientos de Basilea II, en cuanto a conseguir un modelo parsimonioso, no sobreajustado y, por tanto, adecuadamente generalizable, y, además, muy explicativo, por cuanto, el peso de las variables viene dado por los coeficientes de la combinación lineal, hecho muy intuitivo y por tanto fácilmente explicable al cliente, facilidad muy conveniente sobre todo en el caso de tener que explicarle el porqué se le denegó el crédito solicitado.

A pesar de la importante fortaleza que presenta el método, puesto que *el tramado de las variables permite modelar dependencias no lineales a través de modelos lineales,* SIDDIQI, (2006), presenta dos importantes debilidades, por un lado al sustituir las variables originales por las correspondientes variables tramadas es necesario aceptar un relativo coste de posible pérdida de información y la renuncia a matices que en algunos casos podrían revelarse como importantes para la eficacia del modelo, y, por otro, *no se consideran las distintas formas de no linealidad que pueden presentar las variables consideradas con información relevante para la construcción del modelo, sino solamente aquella que el tramado es capaz de especificar*.

Es necesario, por tanto, contemplar modelos logísticos con estructuras funcionales capaces de especificar tanto la *linealidad* como las *distintas manifestaciones de la no linealidad*.

En el artículo de LIU y CELA (2009), Capitulo 1, apartado 1.3.2.7, los autores, reconociendo que *la estructura funcional parcialmente no paramétrica de los modelos LPALM hace difícil la calificación de nuevos solicitantes de crédito directamente desde los datos, lo que dificulta que se utilicen en entornos de negocios y se implementen en escenarios de producción,* presentan una combinación de los modelos LPALM con los Árboles de Regresión y Clasificación, TREE, y con los Splines de Regresión Adaptativos Multivariantes, MARS, para mejorar la interpretabilidad de los LPALM pretendiendo que estas dos nuevas técnicas de construcción de modelos se derrumben dentro del armazón de los Modelos Lineales Generales, GLM, que, según sus propias palabras, "son más familiares a los constructores de modelos" de calificación de acreditados.

LIU y CELA (2009) como solución al problema de estimar los parámetros de cada término no lineal deducidos del modelo LPALM proponen un modelo basado en la utilización de la técnica TREE, para abordar el problema con aproximaciones constantes a trozos, y en la utilización de la técnica MARS que es capaz de abordar el mismo problema con aproximaciones lineales a trozos.

Utilizan la técnica TREE a través del algoritmo CART, (BREIMAN et al. 1984), con una sola variable independiente y una variable dependiente, que son el correspondiente término no lineal y estado de default respectivamente. Después del desarrollo de CART se utiliza la *aproximación constante a trozos resultante como una variable categórica.*

Los autores, en la misma forma en que han utilizado TREE, utilizan la técnica MARS, FRIEDMAN, (1991). Después del desarrollo de MARS utilizan la *aproximación lineal a trozos resultante como una variable numérica.* Tras el desarrollo de MARS, se utilizan las funciones base obtenidas para reemplazar los términos no lineales de LPALM y estimar de nuevo un modelo con la inclusión de estas funciones base.

Sin embargo, los trabajos de LIU y CELA (2007, 2009) tropiezan con un serio inconveniente, se especifica la no linealidad para todas las variables que lo requieran del mismo modo. En el primero, por funciones de base consistentes en splines cúbicos naturales y en el segundo según dos versiones, por funciones constantes a trozos obtenidas de CART unidimensional para la variable en cuestión o por funciones lineales a trozos obtenidas por aplicación de MARS en la misma forma. Pero nada hay que asegure a priori que la no linealidad de todas las variables no lineales deba especificarse de la misma forma, excepto que se compruebe que efectivamente así es, lo que sería un hecho excepcional.

Si bien es un reto importante encontrar la estructura formal que especifique de forma rigurosa la no linealidad, el reto puede rozar el filo de lo imposible si se hace de forma que el modelo no pierda generalidad. El punto de equilibrio consistirá, sin ninguna duda, en describir la no linealidad de la forma más rigurosa posible sin complicar innecesariamente el modelo. En la línea de los modelos (6.12) se pueden considerar modelos tan complejos como se necesite, puesto que en un planteamiento general las funciones de base que forman la componente no lineal son cualesquiera funciones de X, $X^T = \left(U^T, V^T\right)$, pero no podemos olvidar los requerimientos de Basilea II, referidos a la complejidad del modelo.

Por lo que respecta al tramado de variables, (SIDIQI, 2006), nuestra propuesta consiste en aceptar como principio general el hecho de no renunciar a ninguna información si no es estrictamente necesario, es decir, de trabajar, siempre que sea posible, con las variables originales y, por tanto, de no transformar ninguna variable si no está totalmente justificado, aparte de que reiteramos que nada hay que justifique especificar todas las variables del mismo modo.

Ante la problemática anterior, con las soluciones aportadas por los distintos métodos, que consideramos que no son totalmente satisfactorias, con todos los condicionantes que nos venimos imponiendo, guiados por la necesidad de un modelo estadísticamente adecuado,

sujeto a los requerimientos de Basilea II, hecho a la medida de las políticas en materia de riesgos de la entidad financiera y de los comportamientos frente al default de sus acreditados, es evidente que nos vemos avocados a *modelos logísticos, parcialmente lineales e híbridos en la componente no lineal, es decir utilizando funciones de base específicas para cada variable, en función de sus características. Además los modelos habrán de ser supervisados tanto por los estadísticos como por los expertos en análisis del riesgo*.

De todas las consideraciones anteriores, y de las expuestas a lo largo de toda esta memoria, con el objetivo de conseguir modelos de credit scoring lo más parsimoniosos posibles con el suficiente poder predictivo y fácilmente interpretables, surge nuestra propuesta, basada en las hipótesis H1, H2 y H3,

El modelo, (cuya estructura funcional informa de la relación de dependencia entre la probabilidad de default y las variables de riesgo de crédito, $\boldsymbol{X}^T=\left(\boldsymbol{U}^T,V^T\right)$, con $\boldsymbol{U}^T=\left(U_1,...,U_{p_1}\right)$ vector de variables lineales y $\boldsymbol{V}^T=\left(V_1,...,V_{p_2}\right)$ vector de variables no lineales)*, viene dado por* (6.10), *donde además la expansión lineal de funciones de base de las variables explicativas del riesgo de crédito,* $\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$, *es la suma de una componente lineal,* $\beta_0+\sum_{r=1}^{p_1}\beta_r U_r$, *y una estructura aditiva de* p_2 *funciones de las variables no lineales,* $Z_r\left(V_r\right)$, *una para cada variable* V_r, $r=1,...,p_2$, $\sum_{r=1}^{p_2}Z_r\left(V_r\right)$.

El número de funciones de base que integran la combinación lineal $Z_r\left(V_r\right)$ *dependerá del tipo de la relación de dependencia no lineal entre el estado de default y la variable* V_r *y el método utilizado para captarla, por lo que para cada* $r=1,..,p_2$, $Z_r\left(V_r\right)$ *se puede expresar en la forma siguiente:*

$$Z_r\left(V_r\right)=\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r),\ r=1,...,p_2 \tag{6.20}$$

siendo K_r *el número de funciones de base de la combinación lineal* $Z_r\left(V_r\right)$ *y* $h_{r,k}(\cdot)$ *su k-ésima función de base.*

Por tanto, el modelo (6.10) en nuestra propuesta se expresa en la forma

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r+\sum_{r=1}^{p_2}\left(\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right) \qquad (6.21)$$

donde $p_1+\sum_{r=1}^{p_2}K_r=q$ y $p_1+p_2=p$, $\boldsymbol{X}^T=\left(\boldsymbol{U}^T,\boldsymbol{V}^T\right)$, $\boldsymbol{U}^T=\left(U_1,...,U_{p_1}\right)$, $\boldsymbol{V}^T=\left(V_1,...,V_{p_2}\right)$, $\boldsymbol{\beta}^T=\left(\beta_1,...,\beta_{p_1}\right)$, $\boldsymbol{H}\left(\boldsymbol{V}\right)=\left(\boldsymbol{H}_1\left(V_1\right),...,\boldsymbol{H}_{p_2}\left(V_{p_2}\right)\right)$, $\boldsymbol{H}_r^T\left(V_r\right)=\left(h_{r1}\left(V_r\right),...,h_{rK_r}\left(V_r\right)\right)$, con $r=1,...,p_2$, $\boldsymbol{\theta}^T=\left(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{p_2}\right)$, $\boldsymbol{\theta}_r^T=\left(\theta_{r1},...,\theta_{rK_r}\right)$.

A los modelos con estructura formal (6.21), cuyo método de construcción se enmarca dentro de los *métodos de restricción supervisada,* les llamaremos ***Modelos Logísticos Lineales Híbridos,*** **HLLM**. El calificativo "híbrido" se debe a que cada variable no lineal V_r, $r=1,...,p_2$, se puede expresar como una combinación lineal de funciones de base específicas diferentes, como por ejemplo, transformaciones polinómicas, logarítmicas, pesos de la evidencia obtenidas por un proceso de tramado óptimo de la variable V_r, funciones constantes a trozos resultados de particiones recursivas de árboles de decisión, obtenidas por el algoritmo CART sobre V_r, funciones lineales a trozos, splines, ya sean de regresión, de suavizado o de penalización, funciones sierra obtenidas por el método PPR sobre la variable V_r, funciones bisagra obtenidas por el procedimiento MARS sobre V_r, funciones de base radial Gaussiana, etc.

Una vez que se han determinado para cada variable no lineal V_r, las funciones de base que forman la combinación lineal (6.20), $Z_r\left(V_r\right)=\sum_{k=1}^{K_r}\theta_{rk}h_{rk}(V_r)$, *el complejo modelo no lineal* (6.11) *se derrumba dando paso al nuevo modelo lineal en el espacio expansionado de las funciones de base,* HLLM.

El correspondiente *Modelo Probit Lineal Híbrido por expansiones lineales de funciones de base,* HLPM adopta la forma:

$$\text{probit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r+\sum_{r=1}^{p_2}\left(\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right) \qquad (6.22)$$

6.4 PROBLEMA GENERAL DE ESTIMACIÓN DE LOS MODELOS LOGÍSTICOS.

A continuación, desarrollamos la solución del *Problema General de Minimización del Riesgo Empírico* para la *función de pérdida logística*, que se obtiene resolviendo el sistema ecuaciones normales a través del algoritmo de Newton_Raphson, que requiere el gradiente y la matriz de segundas derivadas o *matriz Hessiana*.

La función objetivo a minimizar para la estimación de los modelos logísticos, (6.10) se plantea, a través de las funciones de base que configuran la estructura de la función de calificación, $S(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)$, de la función de pérdida logística, $\ell_\Lambda\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)$, y de la funcional de regularización, $J\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right)$, en los siguientes términos:

$$\min_{\beta_0,\boldsymbol{\beta}} \; L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right) \tag{6.23}$$

$$\begin{aligned} L_{\ell_\Lambda}\left(Y,\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right) = -\Big[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)\right)\Big] \\ +\lambda J\left(\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})\right) \end{aligned} \tag{6.24}$$

Como es habitual, el procedimiento a seguir para resolver el problema de optimización (6.23), es decir, obtener los estimadores $\hat{\beta}_0$ y $\hat{\boldsymbol{\beta}}$, consiste en calcular el vector de derivadas parciales de primer orden de la función de log-verosimilitud negativa con respecto de los parámetros a estimar, β_0 y $\boldsymbol{\beta}$, *vector gradiente,* igualarlas a cero y resolver el sistema de *ecuaciones normales* resultante. El sistema de ecuaciones normales, representa un sistema de $q+1$ ecuaciones no lineales con $q+1$ incógnitas $\left(\beta_0,\beta_1,.....,\beta_q\right)$, por lo que será necesario aplicar un sistema iterativo que permita la convergencia de los estimadores. Este sistema puede resolverse por el algoritmo de Newton_Raphson, que, además del *vector gradiente,* requiere la *matriz Hessiana.*

Por tanto, para resolver el problema de optimización (6.23) es necesario establecer las dos siguientes hipótesis:

(1).– $J\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right)$, es una función convexa.

(2).– $\exists\ \dfrac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{J}\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right) = \boldsymbol{J}'(\boldsymbol{\beta})$ (6.25)

y

$\exists\ \dfrac{\partial^2}{\partial \boldsymbol{\beta}\ \partial \boldsymbol{\beta}^T} \boldsymbol{J}\left(\boldsymbol{\beta}^T \boldsymbol{H}(\boldsymbol{X})\right) = \boldsymbol{J}''(\boldsymbol{\beta})$

La primera hipótesis se requiere con el fin de que la función objetivo $L_{\ell_\Lambda}\left(Y,\ \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)$ sea convexa, (puesto que el primer sumando de (6.24) ya es una función convexa). La segunda exigencia es necesaria para que existan las dos primeras derivadas de la expresión anterior.

Respecto del gradiente se tiene

$$\frac{\partial L_{\ell_\Lambda}\left(Y,\ \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta})} = -\left((\boldsymbol{1}, \boldsymbol{H}(X))^T Y - (1, \boldsymbol{H}(X))^T \frac{e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}}{1 + e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}} \right) + \lambda\left(\boldsymbol{0}, \boldsymbol{J}'(\boldsymbol{\beta})\right)$$

$$= -(\boldsymbol{1}, \boldsymbol{H}(X))^T \left(Y - \frac{e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}}{1 + e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}} \right) + \lambda\left(\boldsymbol{0}, \boldsymbol{J}'(\boldsymbol{\beta})\right)$$

donde $\dfrac{e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}}{1 + e^{\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)^T}} = \left(P\left(x_1, \beta^{anterior}\right), ..., P\left(x_i, \beta^{anterior}\right), ..., P\left(x_N, \beta^{anterior}\right)\right)$ término que notaremos por $\boldsymbol{p}$, donde $P\left(x_i, \beta^{anterior}\right) = P\left(Y = 1 / X = x_i, \beta^{anterior}\right)$. Por tanto, $\boldsymbol{p}$ es el vector de las probabilidades estimadas tales que las respuestas son iguales a 1. Las derivadas se evalúan en $\beta^{anterior}$.

El gradiente de $L_{\ell_\Lambda}\left(Y,\ \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)$ se puede entonces expresar en la forma

$$\frac{\partial L_{\ell_\Lambda}\left(Y,\ \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta})} = -(\boldsymbol{1}, \boldsymbol{H}(\boldsymbol{X}))^T (Y - P) + \lambda\left(\boldsymbol{0}, \boldsymbol{J}'(\boldsymbol{\beta})\right) \quad (6.26)$$

Respecto de la matriz de segundas derivadas parciales,

$$\frac{\partial^2 L_{\ell_\Lambda}\left(Y,\ \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta})\ \partial(\beta_0, \boldsymbol{\beta})^T} = -(\boldsymbol{1}, \boldsymbol{H}(X))^T \frac{\partial}{\partial(\beta_0, \boldsymbol{\beta})^T} \left(y - \frac{e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X)}}{1 + e^{\beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X))}} \right) + \lambda \boldsymbol{J}''(\boldsymbol{\beta})$$

y puesto que

$$\frac{\partial}{\partial \boldsymbol{\beta}^T}\left(y - \frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}\right) = \frac{(\boldsymbol{1},\boldsymbol{H}(X))^T\, e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}{\left(1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}\right)^2} = (\boldsymbol{1},\boldsymbol{H}(X))^T \frac{e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}\frac{1}{1+e^{\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)}}$$

$$= (\boldsymbol{1},\boldsymbol{H}(X))^T \frac{1}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}\frac{e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right))}}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}$$

siendo

$$\frac{e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}} = \left(1-P\left(x_1,\beta^{anterior}\right),...,1-P\left(x_i,\beta^{anterior}\right),...,1-P\left(x_N,\beta^{anterior}\right)\right) = 1-\boldsymbol{p}$$

y

$$\frac{1}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}} = \frac{e^{\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}{1+e^{\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}} = \boldsymbol{p}$$

se llega a

$$\frac{e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}}\frac{1}{1+e^{-\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}} = \boldsymbol{p}(1-\boldsymbol{p}) = \boldsymbol{W}$$

donde $\boldsymbol{W}$ es la matriz diagonal con elementos diagonales $p(x_i)\left(1-p(x_i)\right)$, $\boldsymbol{W} = diag\left(p(x_i)\left(1-p(x_i)\right)\right)_{N\times N}$.

En conclusión

$$\frac{\partial^2 L_{\ell_\lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})\ \partial(\beta_0,\boldsymbol{\beta})^T} = -(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\,\boldsymbol{W}\ \ (\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X})) + \lambda\left(\boldsymbol{0},\ \boldsymbol{J}''(\boldsymbol{\beta})\right) \tag{6.27}$$

Los estimadores de los parámetros $\hat{\beta}_0$ y $\hat{\boldsymbol{\beta}}$ se obtienen del sistema de $q+1$ ecuaciones normales

$$\frac{\partial L_{\ell_\lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})} = -(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\left(y-P\right) + \lambda\left(\boldsymbol{0},\ \boldsymbol{J}'(\boldsymbol{\beta})\right) = 0 \tag{6.28}$$

Un procedimiento común para estimar β_0 y $\boldsymbol{\beta}$ en un **Modelo Lineal Generalizado, GLM**, es el de máxima verosimilitud en el que, para obtener el óptimo de la función de verosimilitud, se utiliza habitualmente un método de optimización local basado en un *algoritmo iterativo de tipo Newton-Raphson*. Variantes de este algoritmo son el conocido como *mínimos cuadrados iterativamente reponderados, o de mínimos cuadrados con reasignación de pesos* (IRLS, iteratively reweightd least squares), McCULLAG y NELDER, (1989), HASTIE et al. (2009), y el *algoritmo scoring de Fisher*, que reemplaza el Hessiano por sus esperanzas. Este último coincide en los modelos logísticos con el algoritmo de Newton Raphson debido a que el Hessiano del modelo logístico general

no incluye a y_i, por lo que se verifica $E\left[\boldsymbol{H}(\boldsymbol{\beta})\right]=\boldsymbol{H}(\boldsymbol{\beta})$, (GREEN, 2003). El método que usaremos aquí será el de Newton-Raphson.

El sistema de ecuaciones (6.28) se resuelve, por tanto, utilizando iterativamente el algoritmo de Newton_Raphson que, comenzando con $(\beta_0,\boldsymbol{\beta})^{anterior}$, actualiza en cada paso iteративamente el vector de coeficientes $(\beta_0,\boldsymbol{\beta})^{nuevo}$ en la forma siguiente

$$(\beta_0,\boldsymbol{\beta})^{nuevo}=(\beta_0,\boldsymbol{\beta})^{anterior}-\left(\frac{\partial^2 L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})\ \partial(\beta_0,\boldsymbol{\beta})^T}\right)^{-1}\frac{\partial L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})}$$

$$=(\beta_0,\boldsymbol{\beta})^{anterior}-\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ \ (\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\boldsymbol{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right)^{-1}\left(-(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T(y-P)+\lambda(\boldsymbol{0},\boldsymbol{J}'(\boldsymbol{\beta}))\right)$$

Notando por $A=\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ \ (\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\boldsymbol{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right)^{-1}$ se tiene

$$(\beta_0,\boldsymbol{\beta})^{nuevo}=(\beta_0,\boldsymbol{\beta})^{anterior\,anterior}-A\left(-(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T(y-P)\right)-A\lambda(\boldsymbol{0},\boldsymbol{J}'(\boldsymbol{\beta}))$$

$$=A\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}A^{-1}(\beta_0,\boldsymbol{\beta})^{anterior}+A(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\left(\boldsymbol{W}^{-1}(y-P)\right)$$

$$-A\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}\lambda(\boldsymbol{0},\boldsymbol{J}'(\boldsymbol{\beta}))$$

$$=A\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)\left[\begin{array}{l}\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}A^{-1}(\beta_0,\boldsymbol{\beta})^{anterior}\\ +\boldsymbol{W}\left(\boldsymbol{W}^{-1}(y-P)\right)-\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}\lambda(\boldsymbol{0},\boldsymbol{J}'(\boldsymbol{\beta}))\end{array}\right]$$

Sustituyendo la matriz A por la expresión $\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ \ (\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\boldsymbol{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right)^{-1}$ se tiene

$$(\beta_0,\boldsymbol{\beta})^{nuevo}=\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ \ (\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\boldsymbol{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right)^{-1}(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}$$

$$\left[(\beta_0,\boldsymbol{\beta})^{anterior}(\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))+\boldsymbol{W}^{-1}(y-P)+\lambda\left((\boldsymbol{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}\left((\boldsymbol{0},\boldsymbol{J}''(\boldsymbol{\beta}))\boldsymbol{\beta}^{anterior}-(\boldsymbol{0},\boldsymbol{J}'(\boldsymbol{\beta}))\right)\right]$$

por lo que

$$(\beta_0,\boldsymbol{\beta})^{nuevo}=\left[(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ (\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\mathbf{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right]^{-1}(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\boldsymbol{z} \tag{6.29}$$

donde

$$\begin{aligned}\boldsymbol{z}=&\left((\beta_0,\boldsymbol{\beta})^{anterior}\right)^T(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))+\boldsymbol{W}^{-1}(y-P)\\&+\lambda\left((\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\right)^{-1}\left[(\mathbf{0},\boldsymbol{J}''(\boldsymbol{\beta}))(\beta_0,\boldsymbol{\beta})^{anterior}-(\mathbf{0},\boldsymbol{J}'(\boldsymbol{\beta}))\right]\end{aligned} \tag{6.30}$$

es la variable respuesta ajustada o variable respuesta de trabajo.

Obsérvese que en el primer sumando, $(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T(\beta_0,\boldsymbol{\beta})^{anterior}$ es la expansión de funciones de base estimada en el paso anterior de Newton-Raphson. En el procedimiento de Newton-Raphson, para $\lambda=0$ se alterna la formación de una variable dependiente ajustada $\boldsymbol{z}$ con varianza $\boldsymbol{W}^{-1}$, y la regresión $\boldsymbol{z}$ sobre $\boldsymbol{H}(\boldsymbol{X})$ con pesos $\boldsymbol{W}$.

$\boldsymbol{\beta}^{anterior}=(0,...,0)^T$ es un buen valor inicial para el procedimiento iterativo y, si bien es importante resaltar que la convergencia no está nunca garantizada, normalmente el algoritmo converge puesto que la log-verosimilitud negativa es convexa, pero puede producirse el sobreajuste. En los raros casos en que la log-verosimilitud negativa decrece, el tamaño de paso se parte en dos garantizando la convergencia.

La nueva expansión de las variables explicativas del default se expresa según la siguiente igualdad:

$$\left((\beta_0,\boldsymbol{\beta})^{nuevo}\right)^T(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))=\boldsymbol{S}_{\lambda W m_h}\boldsymbol{z} \tag{6.31}$$

siendo

$$\boldsymbol{S}_{\lambda W}=(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))\left[(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W}\ (\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))+\lambda(\mathbf{0},\boldsymbol{J}''(\boldsymbol{\beta}))\right]^{-1}(\mathbf{1},\boldsymbol{H}(\boldsymbol{X}))^T\boldsymbol{W} \tag{6.32}$$

el operador de regresión.

Como puede observarse en (6.30) y (6.32), tanto la variable respuesta, $\boldsymbol{z}$, como el operador de regresión, $\boldsymbol{S}_{\lambda W m_h}$, no sólo dependen de la expansión de funciones de base de las variables explicativas sino también de la funcional de regularización y de la transformación y pérdida logísticas, reflejados los dos últimos elementos en la matriz diagonal $\boldsymbol{W}$ cuyos N elementos diagonales son los productos de la probabilidad de default y la probabilidad de no default para cada acreditado, valorados sobre la expansión de las funciones de base de las variable explicativas del paso Newton-Raphson anterior.

HASTIE y TIBSHIRANI, (1999), usando los valores del paso Newton-Raphson final del algoritmo anterior para el caso de la regresión logística lineal, caso particular del general que estamos analizando, estimaron los grados de libertad efectivos del modelo. Una *extensión a los grados de libertad efectivos de los modelos logísticos generales* puede aproximarse por

$$df(\lambda) = traza\left(\left[\left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right) + \lambda\left(\mathbf{0}, \mathbf{J}''(\boldsymbol{\beta})\right)\right]^{-1} \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)\right) \quad (6.33)$$

donde $\boldsymbol{W}$ es la obtenida en el paso final del algoritmo.

Por lo que respecta a la varianza de las estimaciones de los coeficientes, GRAY, (1992), también para el caso de la regresión logística lineal, Hastie y Tibshirani la estimaron desde la iteración final. La extensión a los modelos logísticos generales viene dada por:

$$\begin{aligned} Var(\beta_0, \boldsymbol{\beta})) &= Var\left(\left[\left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right) + \lambda\left(\mathbf{0}, \mathbf{J}''(\boldsymbol{\beta})\right)\right]^{-1} \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W}\mathbf{z}\right) \\ &= \left[\left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right) + \lambda\left(\mathbf{0}, \mathbf{J}''(\boldsymbol{\beta})\right)\right]^{-1} I(\boldsymbol{\beta}) \\ &= \left[\left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right) + \lambda\left(\mathbf{0}, \mathbf{J}''(\boldsymbol{\beta})\right)\right]^{-1} \\ &= \left(Hessiano\left(L(\boldsymbol{\beta})\right)\right)^{-1} I(\boldsymbol{\beta}) \left(Hessiano\left(L(\boldsymbol{\beta})\right)\right)^{-1} \end{aligned} \quad (6.34)$$

donde $\mathbf{I}(\beta_0, \boldsymbol{\beta}) = -E\left[\dfrac{\partial^2 L_{\ell_\lambda}\left(Y, \ \beta_0 + \boldsymbol{\beta}^T \mathbf{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta}) \ \partial(\beta_0, \boldsymbol{\beta})^T}\right]$ *es la matriz de información de Fisher.*

Debido a que el Hessiano del modelo logístico general no incluye a y, se verifica que $E\left[\dfrac{\partial^2 L_{\ell_\lambda}\left(Y, \ \beta_0 + \boldsymbol{\beta}^T \mathbf{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta}) \ \partial(\beta_0, \boldsymbol{\beta})^T}\right] = \dfrac{\partial^2 L_{\ell_\lambda}\left(Y, \ \beta_0 + \boldsymbol{\beta}^T \mathbf{H}(X)\right)}{\partial(\beta_0, \boldsymbol{\beta}) \ \partial(\beta_0, \boldsymbol{\beta})^T}$, (GREEN, 2003), por tanto, de (6.34) se obtiene.

$$Var(\beta_0, \boldsymbol{\beta}) = \left[\left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right)^T \mathbf{W} \ \left(\mathbf{1}, \mathbf{H}(\mathbf{X})\right) + \lambda\left(\mathbf{0}, \mathbf{J}''(\boldsymbol{\beta})\right)\right]^{-1} \quad (6.35)$$

Respecto de las propiedades del estimador de los coeficientes $(\beta_0,\boldsymbol{\beta})$ de la combinación lineal de funciones de base, $\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)$, l*a teoría de verosimilitud dice que si el modelo está correctamente especificado, entonces*

1.- $\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)$***es un estimador consistente de*** $(\beta_0,\boldsymbol{\beta})$,

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)\xrightarrow{P}(\beta_0,\boldsymbol{\beta}) \tag{6.36}$$

es decir, $\lim_{N\to\infty}\left(P\left|\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)_N-(\beta_0,\boldsymbol{\beta})\right|>\varepsilon\right)=0$, o, $\lim_{N\to\infty}\left(P\left|\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)_N-(\beta_0,\boldsymbol{\beta})\right|\le\varepsilon\right)=1$.

2.- $\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)$ ***se distribuye asintóticamente normal****.*

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)\overset{(a)}{\sim}N\left((\beta_0,\boldsymbol{\beta}),\left\{\boldsymbol{I}((\beta_0,\boldsymbol{\beta}))\right\}^{-1}\right) \tag{6.37}$$

Dado que $E\left[\dfrac{\partial^2 L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})\ \partial(\beta_0,\boldsymbol{\beta})^T}\right]=\dfrac{\partial^2 L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})\ \partial(\beta_0,\boldsymbol{\beta})^T}$, se tiene *la siguiente matriz de información de Fisher*

$$\boldsymbol{I}((\beta_0,\boldsymbol{\beta}))=-\frac{1}{N}E\left[\frac{\partial^2 L_{\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(X)\right)}{\partial(\beta_0,\boldsymbol{\beta})\ \partial(\beta_0,\boldsymbol{\beta})^T}\right]$$

Resulta relativamente sencillo estimar $\boldsymbol{I}(\beta_0,\boldsymbol{\beta})$ sustituyendo el parámetro $(\beta_0,\boldsymbol{\beta})$ por $\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)$. La expresión (6.37) se puede escribir en otra forma:

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)\overset{(a)}{\sim}N\left((\beta_0,\boldsymbol{\beta}),\left(\left(\mathbf{1},\boldsymbol{H}(X)\right)^T\boldsymbol{W}\ \left(\mathbf{1},\boldsymbol{H}(X)\right)+\lambda\left(\mathbf{0},\boldsymbol{J}''(\boldsymbol{\beta})\right)\right)^{-1}\right)$$

3.- $\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)$ ***es asintóticamente eficiente****, es decir,*

$$E\left[\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)\right]=0$$
$$Var\left(\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)\right)\le Var\left(\left(\tilde{\beta}_0,\tilde{\boldsymbol{\beta}}\right)\right) \tag{6.38}$$

siendo $\left(\tilde{\beta}_0,\tilde{\boldsymbol{\beta}}\right)$ *cualquier otro estimador insesgado de* $\left(\beta_0,\boldsymbol{\beta}\right)$.

El estimador de máxima verosimilitud $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$ satisface una relación consistente: sus componentes son los coeficientes del ajuste de mínimos cuadrados, donde las respuestas son

$$z_i = \left[\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^T (\mathbf{1}, \mathbf{H}(X_i)) + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}\right] + \lambda\left((\mathbf{1}, \mathbf{H}(X_i))^T \hat{p}_i(1-\hat{p}_i)\right)^{-1}\left[\left(\mathbf{0}, \mathbf{J}''\left(\hat{\boldsymbol{\beta}}\right)\right)\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) - \left(\mathbf{0}, \mathbf{J}'\left(\hat{\boldsymbol{\beta}}\right)\right)\right] \tag{6.39}$$

y los pesos son $w_i = \dfrac{\hat{p}_i}{(1-\hat{p}_i)}$, ambos dependientes de $\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)$.

Esta conexión con los mínimos cuadrados, aparte de proporcionar un conveniente algoritmo, ofrece más ventajas:

- La suma de cuadrados residuales ponderados es el familiar estadístico Chi-cuadrado de Pearson, $\sum_{i=1}^{N} \dfrac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$, una aproximación cuadrática a la desviación.

- La construcción del modelo de regresión logística general puede ser muy costosa, a causa de que cada modelo ajustado requiere iteración. Existen algunos atajos populares como el *test score de Rao* que contrasta la inclusión de un término y el test de Wald que puede usarse para la exclusión de un término. Ninguno de los dos requiere ajuste iterativo, y se basan en el ajuste de máxima verosimilitud del modelo actual. Así resulta que se puede quitar o añadir términos desde el ajuste de mínimos cuadrados ponderados, usando las mismas ponderaciones, con lo cual es posible hacer eficientes cálculos computacionales sin necesidad de recalcular el ajuste de mínimos cuadrados entero.

6.5 ESTIMACIÓN DE LOS MODELOS HLLM.

Por lo respecta a la función objetivo a optimizar para estimar el modelo HLLM, que proponemos, hemos de decir que la función de pérdida binomial negativa o pérdida logística es la función de pérdida más idónea para estimar el modelo de respuesta binaria.

La función objetivo asociada a la pérdida logística empírica regularizada, (6.24), para el modelo (6.22) viene dado por

$$
\begin{aligned}
&L_{emp\ell_\Lambda}\left(Y,\left(Y,\ \beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)\right)\\
&\quad=-\left[\mathbf{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)\right)\right]\\
&\quad+\lambda J\left(\beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)
\end{aligned}
\tag{6.40}
$$

Por lo que el modelo se estima resolviendo el *problema de optimización siguiente:*

$$
\underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}}{Min}\ L_{emp\ell_\Lambda}\left(Y,\ \beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)
\tag{6.41}
$$

En el caso de que se considere el caso regularizado, es decir $\lambda>0$, habremos de especificar $J\left(\beta_0+\boldsymbol{\beta}^T\mathbf{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\mathbf{H}(\mathbf{V})\right)$ como una función convexa y tal que existan su primera y segunda derivas respectos de los vectores de parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$, $\boldsymbol{J'}(\boldsymbol{\beta},\boldsymbol{\theta})$ y $\boldsymbol{J''}(\boldsymbol{\beta},\boldsymbol{\theta})$. En tal caso estamos ante el problema general de estimación de los modelos logísticos por expansiones lineales de funciones de base y la solución, que coincide con la del problema (6.23), viene dada por

$$
(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{nuevo}=\left[(\mathbf{1},\mathbf{U},\mathbf{H}(V))^T\mathbf{W}\ (\mathbf{1},\mathbf{U},\mathbf{H}(V))+\lambda(\mathbf{0},\boldsymbol{J''}(\boldsymbol{\beta},\boldsymbol{\theta}))\right]^{-1}(\mathbf{1},\mathbf{U},\mathbf{H}(V))^T\mathbf{W}\mathbf{z}
\tag{6.42}
$$

donde

$$
\begin{aligned}
\mathbf{z}&=\left((\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{anterior}\right)^T(\mathbf{1},\mathbf{U},\mathbf{H}(V))+\mathbf{W}^{-1}(y-P)\\
&+\lambda\left((\mathbf{1},\mathbf{U},\mathbf{H}(V))^T\mathbf{W}\right)^{-1}\left[(\mathbf{0},\boldsymbol{J''}(\boldsymbol{\beta},\boldsymbol{\theta}))(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{anterior}-(\mathbf{0},\boldsymbol{J'}(\boldsymbol{\beta},\boldsymbol{\theta}))\right]
\end{aligned}
\tag{6.43}
$$

es la variable respuesta ajustada o variable respuesta de trabajo.

La nueva expansión lineal de funciones de base se expresa según la siguiente igualdad

$$
\left((\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{nuevo}\right)^T(\mathbf{1},\mathbf{U},\mathbf{H}(V))=\beta_0^{nuevo}+\left(\boldsymbol{\beta}^{nuevo}\right)^T U+\left(\boldsymbol{\theta}^{nuevo}\right)^T\mathbf{H}(V).
$$

Dado que nos encontramos ante un *método restrictivo*, donde decidimos a priori limitar la clase de funciones base, la complejidad del modelo está controlada por el número de variables lineales y por la expansión por funciones de base que fijemos para cada variable no lineal. *Si se siguen con rigor los requerimientos de Basilea II, lo recomendable será no complicar en exceso las expansiones por funciones de base de las variables* V_r, $Z_r(V_r)=\sum_{k=1}^{K_r}\theta_{rk}h_{rk}(V_r)$, *por lo que en general no será necesario regularizar el modelo, es decir,* $\lambda=0$, lo que implica que tanto el modelo como el método de estimación y, por ello, la solución se simplifique bastante, (*esta es una más de las ventajas de utilizar modelos logísticos híbridos*

supervisados en lugar de utilizar modelos logísticos adaptativos). De este modo el problema de optimización general de los modelos logísticos expansión lineal de funciones de base se reduce a

$$\underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}}{Min} -\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)-\mathbf{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)\right)\right] \tag{6.44}$$

y la solución viene dada por

$$(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{nuevo}=\left[(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}(V))^T\boldsymbol{W}\ (\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}(V))\right]^{-1}(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}(V))^T\boldsymbol{W}\boldsymbol{z} \tag{6.45}$$

donde

$$\boldsymbol{z}=\left((\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{anterior}\right)^T(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}(V))+\boldsymbol{W}^{-1}(y-P)$$

es la variable respuesta ajustada o variable respuesta de trabajo.

Para el caso del modelo probit la función objetivo asociada a la pérdidas empírica regularizada vienen dadas por

$$\begin{aligned}L_{emp\ell_\Lambda}(Y,S(X))=&-\boldsymbol{Y}^T\log\left(\Phi\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)\right)\\&-(\mathbf{1}-\boldsymbol{Y})^T\log\left(1-\Phi\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)\right)\\&+\lambda J\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)\end{aligned} \tag{6.46}$$

La optimización de (6.46) para el caso general, $\lambda\geq 0$, se resuelve de forma análoga al caso logístico, (6.32). Habremos de especificar $J\left(\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)$ como una función convexa para la que existan su primera y segunda derivadas respecto de los vectores de parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$, $\boldsymbol{J}'(\boldsymbol{\beta},\boldsymbol{\theta})$ y $\boldsymbol{J}''(\boldsymbol{\beta},\boldsymbol{\theta})$, tras lo cual se procede como en el problema general de estimación de los modelos logísticos por expansiones lineales de funciones de base, con la salvedad de haber sustituido la función de distribución logística de parámetros 0 y 1, $\Lambda(\bullet)$, por la función de distribución normal estandarizada, $\Phi(\bullet)$.

6.6 SELECCIÓN DE LAS FUNCIONES DE BASE PARA LA COMPONENTE NO LINEAL DEL MODELO *HLLM*.

En general, para la construcción de un modelo de credit scoring se contará con p_1 variables con influencia lineal sobre la variable estado de default, $U_1,...,U_{p_1}$, y p_2 variables, $V_1,...,V_{p_2}$, cuyo tipo de influencia sobre el default es desconocido. Con

este hecho como hipótesis de partida, el modelo logístico parcialmente lineal inicial, semiparamétrico, adopta la forma estructural dada por (6.11). Nuestro objetivo consiste en "convertir" la estructura (6.11) en la correspondiente al modelo (6.21), modelo logístico lineal, paramétrico, a través de expansiones lineales de funciones de base $\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r)$. La estructura (6.21) quedará perfectamente especificada una vez que para cada variable V_r, $r=1,...,p_2$, hayan sido seleccionadas las funciones de base $h_{r,k}(V_r)$, $k=1,...,K_r$. Estas funciones de base se especificarán, en primer lugar, de acuerdo con nuestro conocimiento sobre la relación entre el estado de default y las variables explicativas del riesgo y, en segundo lugar, por poseer unas propiedades convenientes.

La metodología que proponemos para la selección de las funciones de base que formarán parte de la expansión lineal de las variables de la componente no lineal del modelo (6.21) *consiste en el siguiente método constructivo:*

1.- En primer lugar se considera un modelo de partida, $M_{inicial}$, modelo logístico lineal ,

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r \qquad (6.47)$$

con todas las variables explicativas linealmente significativas que se consideran de interés en riesgo de crédito, donde la función de calificación de acreditados es una expansión lineal de funciones identidad, $h_r(\boldsymbol{U})=V_r$, $\boldsymbol{U}=\left(U_1,...,U_{p_1}\right)$. Este punto es de necesidad, de acuerdo con los requerimientos de Basilea II, pues difícilmente cualquier otra estructura resulta más fácilmente interpretable que la lineal.

2.- En segundo lugar se hace uso de un método constructivo para seleccionar la combinación lineal de funciones de base "más prometedora" para cada variable de un conjunto de candidatas e incorporarla al modelo. Este método consiste en ir añadiendo, en cada paso, al modelo inicial las combinaciones lineales cuyas funciones de base resulten significativas, utilizando el *test chi_cuadrado*, en el modelo logístico conseguido en el paso anterior. En cada paso se irá analizando si la incorporación de funciones de base incide en la significación el modelo de otras variables ya incorporadas, a la vez que se comprobará el *ajuste del modelo a los datos de entrenamiento*, a través de los *pseudo-coeficientes de determinación*, *McFadden y Nagelkerke*, del *Error Empírico* y de los *criterios de Información de Akaike*, AIC, y

Schwarz, BIC, y el *poder discriminante del modelo*, a través del *área bajo la curva ROC*, AUC.

Antes de proceder a la selección de las funciones de base más prometedoras en cada paso, se realiza un proceso de evaluación sobre la muestra de validación, obtenida de forma aleatoria e independiente de la muestra de entrenamiento para tal fin. Este proceso conlleva la evaluación de los siguientes elementos*: el Ajuste del Modelo*, su *Error de Validación*, (error empírico sobre la muestra de validación), los C*riterios de Información* AIC y BIC, *la Validación del Poder Discriminante* del nuevo modelo sobre la muestra de validación, AUC, y la *Tasa de Clasificación Errónea*.

Se procede secuencialmente según el proceso anterior para todas y cada una de las variables de la componente no lineal. A la vez que se construirán modelos alternativos considerando expansiones lineales por funciones base que aunque no sean las más prometedoras parezcan, en principio, adecuadas, o, al menos, de interés a efectos comparativos.

3.- Como resultado del proceso de construcción expuesto en el apartado 2 se llega a un conjunto de modelos con estructura inicialmente valida pero que podría, en principio, sobreajustar los datos, no poseer la cualidad de facilidad interpretativa etc. Para evitar tales inconvenientes, en una tercera fase se procede a aplicar técnicas de poda o regularización para reducir el número de funciones de base.

6.7 FUNCIONES DE BASE NOTABLES.

Con el fin de especificar la componente no lineal a través de funciones de base adecuadas para el vector de las variables no lineales $V=\left(\mathrm{V}_1,...,\mathrm{V}_{p_2}\right)$, para las que no se observa significación en el modelo logístico lineal (5.23), $M_{inicial}$, consideraremos funciones de base de varios tipos diferentes:

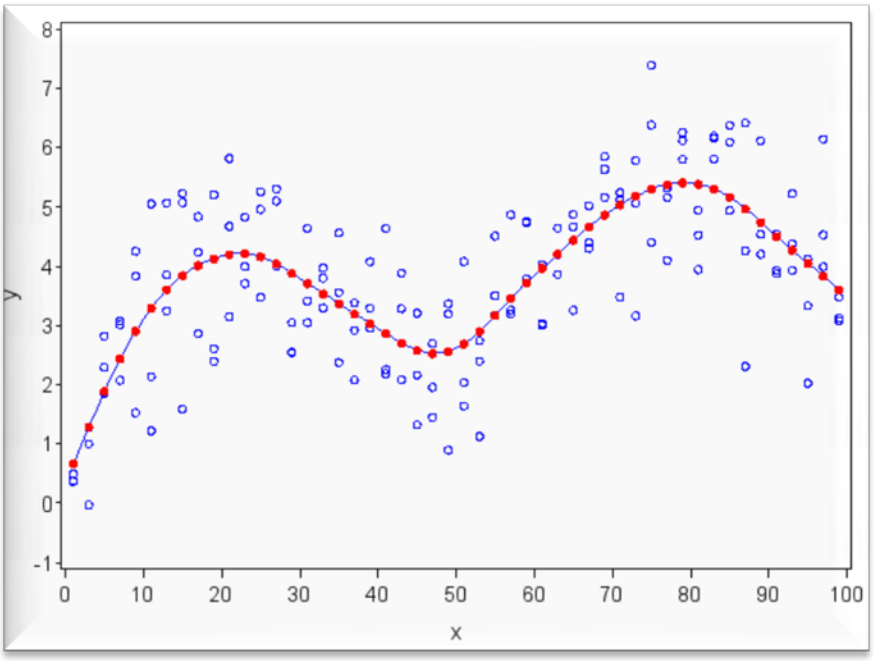
Funciones de base h_r *muy sencillas*, tales como *potencias* $h_r\left(X\right)=X_j^d$, logaritmos, $h_r\left(X\right)=\log\left(X_j\right)$, raíces cuadradas, $h_r\left(X\right)=\sqrt{X_j}$, interacciones de variables, $h_r\left(X\right)=X_jX_k,.....,\ r=1,...,q$ y $j,k=1,...,p$; *funciones de base polinómicas de orden* p, *pesos de la evidencia asignados a tramados estadísticos automáticos o por criterios de riesgos o por particiones recursivas; splines cúbicos restringidos de Stone y Koo, RCS,*

funciones de base bisagra obtenidas por MARS univariante; funciones de base radial, RBF.

6.7.1 Funciones de Base Polinómicas.

Las funciones de base polinómicas de orden igual o superior a 2 y las funciones de base del tipo $h_r(V)=V_rV_k,....,$ permiten aumentar las entradas con términos polinómicos para alcanzar expansiones de Taylor de altos órdenes, pero poseen el inconveniente de que el número de variables crece exponencialmente con el grado del polinomio. Un modelo cuadrático completo en p variables requiere $O\left(p^2\right)$ términos cuadráticos y de productos cruzados, en general $O\left(p^d\right)$ para un polinomio de grado p.

Por otro lado, las funciones de base polinómicas presentan también el problema de que *según aumenta su número crece su colinealidad,* lo que conduce a estimadores de los parámetros de alta varianza y a numerosos problemas de cálculo numérico. Además, los polinomios de alto orden tienen tendencia a oscilar descontroladamente si existen huecos grandes entre los $x_i's$. Para no obtener modelos excesivamente complejos, que muy probablemente conducirían al sobreajuste, lo usual es añadir la potencia de orden 2 de la variable, $h_r(V)=V_r^2$.



Fuente: SAS/STAT 9.2 User's Guide.SAS®.

Figura 6.1.- Ajuste a una Polinomial con 1ª y 2ª derivadas continuas.

6.7.2 Funciones Constantes a Trozos obtenidas a partir de Indicadores de Particiones Recursivas.

Troceando apropiadamente el rango de la variable X_j en regiones disjuntas se obtiene una partición del rango de V_r en K_r regiones disjuntas, $\{R_{j1},...,R_{jK_j}\}$, donde $rango(V_r)=\bigcup_{k=1}^{K_j} R_{rk}$, $R_{rk}\cap R_{rl}=\varnothing, k\neq l$. Definiendo las funciones de base como indicadores de cada una de las regiones de la partición del rango de X_j resulta un modelo con contribuciones constantes a trozos para la variable.

$$h_r(X)=I_{I_{[X_j\in R_{jk}]}},\ r=1,...,q,\ \ j=1,...,p \tag{6.48}$$

Según el método utilizado para obtener la partición de las regiones que componen el rango de la variable tendremos diferentes funciones pero siempre constantes a trozos. Las particiones más habituales en la industria de los sistemas de calificación del riesgo son las *particiones estadísticas automáticas óptimas*, aquellas obtenidas de forma automática con criterios estadísticos únicamente, las *particiones estadísticas con criterios de riesgos*, que partiendo inicialmente de una *partición óptima obtenida por métodos estadísticos es modificada con criterios de riesgo de crédito* (SIDDIQI, 2006) y las *particiones recursivas binarias* conseguidas por el algoritmo CART de los árboles de decisión y de clasificación, TREE, (BREIMAN, et al., 1984).

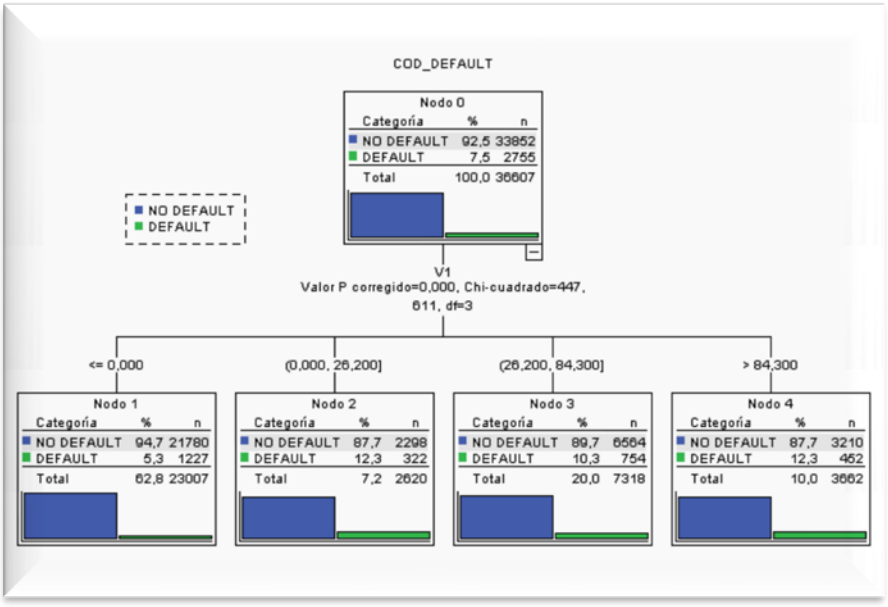


Figura 6.2.- Árbol de clasificación de la variable V1, obtenido por el algoritmo CART, con el método de agrupación CHAID. Se utiliza el estadístico chi-cuadrado de Pearson, para partir nodos y unir categorías, al nivel de significación $\alpha=0,05$.

Las particiones recursivas, tal como vimos en el capítulo 5, se representan de forma natural a través de grafos conexos sin ciclos, es decir sobre árboles de decisión. En la figura 6.2 se muestran los resultados para la variable de riesgo de crédito V_1, *mínimo tasa de saldo pendiente a fin de mes actual frente al importe formalizado en préstamos con garantía personal a fecha de visión,* representados en forma de árbol.

6.7.3 Pesos de la evidencia asignados a tramados de las variables o a particiones recursivas.

Una familia de funciones de base muy popular en credit scoring se obtiene definiendo para cada variable no lineal original V_r una función de base en la forma

$$\begin{aligned} & h_r(V_r): \mathbb{R} \to \mathbb{R} \\ & V_r \ \to h_r(V_r) = \sum_{k=1}^{K_r} I_{[V_r \in R_{rk}]} WOE(R_{rk}) \end{aligned} \tag{6.49}$$

donde $\{R_{r1},...,R_{rK_r}\}$ es una partición del rango de V_r en K_r regiones disjuntas, $rango(V_r) = \bigcup_{k=1}^{K_j} R_{rk}$, $R_{rk} \cap R_{rl} = \varnothing, k \neq l$. y $WOE(R_{rk}) = Ln\left(\frac{P_{\text{Buenos en R}_{rk}}}{P_{\text{Malos en R}_{rk}}}\right)$ el peso de la evidencia del default para la subregión R_{rk} del rango de la variable V_r.

$P_{Buenos\ en\ R_{rk}} = \frac{\text{Número de Clientes Buenos en R}_{rk}}{\text{Número Total de Clientes Buenos}}$, es la proporción de clientes buenos en la región R_{jk} y

$P_{Malos\ en\ R_{rk}} = \frac{\text{Número de Clientes Malos en R}_{rk}}{\text{Número Total de Clientes Malos}}$, es la proporción de clientes malos en la región R_{rk}.

Es decir, cada función de base $h_r(X)$ se define como un indicador para el WOE de cada una de las subregiones R_{rk} del rango de V_r. Sustituyendo la variable V_r en el modelo por la variable $\sum_{k=1}^{K_r} I_{[V_r \in R_{rk}]} WOE(R_{rk})$, resulta un modelo con contribuciones constantes a trozos para V_r, siendo la constante en cada región o tramo del rango de la variable el peso de la evidencia del default para la variable en esa región, a x_{ir} se le asigna $\sum_{k=1}^{K_r} I_{[x_{ir} \in R_{rk}]} WOE(R_{rk})$.

Dado que la partición del rango de V_r es disjunta, x_{ir} sólo puede pertenecer a una de las subregiones R_{rk}, en la que $I_{[x_{ir} \in R_{rk}]}$ vale 1 y vale 0 en las restantes subregiones, por lo que el valor de la variable V_r sobre el acreditado i-ésimo se sustituye por el peso de la evidencia de la subregión del rango de V_r a la que pertenece el acreditado.

Asignar el peso de la evidencia a los acreditados que integran una región del rango de la variable parece muy razonable, por cuanto esa cantidad integra toda la información que sobre el default y no default aporta la pertenencia de un acreditado a una región del rango de la variable.

A partir de la expresión general (6.49) es posible obtener distintas familia de funciones de base cuyos elementos vienen caracterizados por la forma en que para la variable V_r se obtiene la partición de su rango en K_r regiones disjuntas, $\{R_{r1},...,R_{rK_r}\}$.

Si consideramos la variable de riesgo de crédito V_1 y construimos dos funciones de base W_V_1 y WT_V_1, donde W_V_1 es una función del tipo (6.49) para una *partición estadística automática óptima con la asignación a cada acreditado de los pesos de la evidencia correspondientes a tal partición, y* WT_V_1 se obtiene como W_V_1 con la diferencia de que se utiliza una *partición recursiva binaria* obtenida de la aplicación del algoritmo CART sobre la variable con respuesta el estado de default se obtienen las representaciones gráficas mostradas en la figura 6.3, que como puede observarse son constantes a trozos.

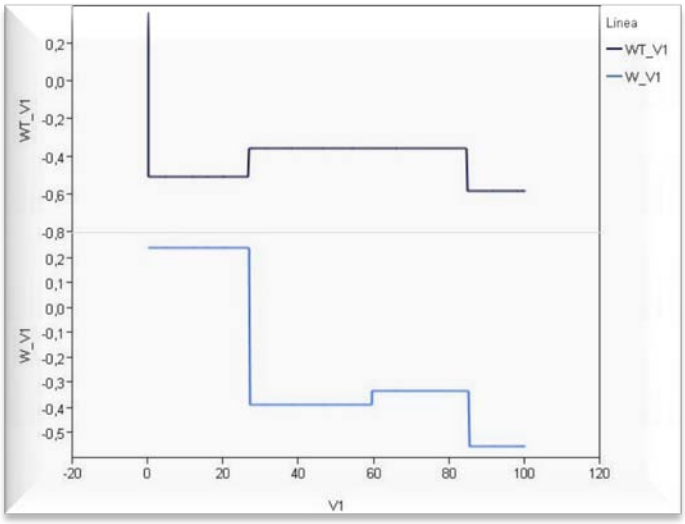


Figura 6.3.- Representación gráfica de las funciones de base W_V_1 y WT_V_1. Ambas funciones son constantes a trozos.

Una función de base del tipo (6.49) es muy fácil de explicar a un cliente, puesto que la puntuación del mismo en una variable de este tipo es directamente proporcional al peso del default frente al no default en la región en que el cliente se sitúa para la variable en cuestión.

Si consideramos dos modelos HLLM construidos ambos sobre las variables $\{V_1,\ldots,V_4,V_6,\ldots,V_{16}, WT\_V_{17}, W\_V_{19}\}$ más la variable V_1 para el primero y WT_V_1 para el segundo, como consecuencia del ajuste de los dos por LLR se tienen los resultados que se muestran en la tabla 6.2.

Tabla 6.2.- Parámetros estimados y test chi-cuadrado, para V_1 y WT_V_1.

Termino	Estimador	Error Std	χ^2	$Prob>\chi^2$
V_1	0,00179403	0,0015933	1,27	0,2602
WT_V_1	0,65554867	0,2015572	10,58	0,0011*

Si se consideran los logit de la probabilidad de default obtenidos en el ajuste LLR y se representan gráficamente frente a V_1 y WT_V_1 respectivamente, se obtienen las gráficas mostradas en la figura 6.4.

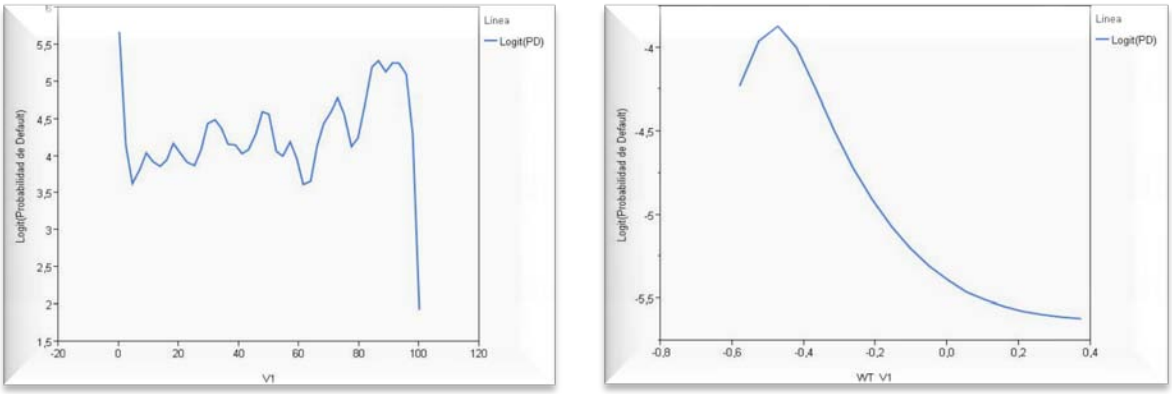


Figura 6.4.- Representación gráfica del logit de la probabilidad de default frente a la variable V1, panel de la izquierda, y frente a la variable WT_V1, panel de la derecha.

La interpretación de los pesos de la evidencia asignados a cada acreditado es muy sencilla, así por ejemplo, para un acreditado con $0<V_1\leq 26,200$, es decir, perteneciente a la región correspondiente al nodo 2, R_{12}, figura 6.2, se tiene $\dfrac{P_{\text{Buenos en R}_{12}}}{P_{\text{Malos en R}_{12}}}=0,5807$ lo que significa que en el tramo al que pertenece el acreditado considerado la ventaja de no

default sobre los default es de 0,58 : 1. El peso de la evidencia que se asigna a este acreditado es, por tanto, $WOE\left(R_{12}\right) = Ln\left(\dfrac{P_{\text{Buenos en R}_{12}}}{P_{\text{Malos en R}_{12}}}\right) = Ln\left(0,5807\right) = -0,5450$.

6.7.4 Splines Cúbicos Restringidos de Stone y Koo, RCS.

Otra importante familia de funciones de base la constituyen los *splines de regresión* que, como hemos visto en el apartado 3.3.3.2 del capítulo 3, poseen magnificas propiedades teóricas y prácticas.

Las funciones para la interpolación por splines se determinan normalmente como suavizadores sometidos a una serie de restricciones. Si bien inicialmente el uso de splines en aplicaciones prácticas conllevó modelos muy complicados con un gran número de funciones de base de suavizado, *Splines de Suavizado y Lamina Fina*, aplicados a grandes conjuntos de datos que resultaban con frecuencia computacionalmente intratables, en la actualidad los *Splines Cúbicos Restringidos*, RCS, en los cuales el número de funciones de base es relativamente pequeño, se están revelando como muy eficaces. A esta última categoría pertenecen los *splines de regresión cúbicos k-anudados de Stone*, (STONE, 1986b), o también de Stone-Koo, (STONE y KOO, 1985), una clase particular de splines cúbicos de regresión restringidos, RCS, que tienen primeras y segundas derivadas continuas en los nodos (HASTIE y TIBSHIRANI, 1990) para suavizado visual (DURRLEMAN y SIMON, 1989).

La idea de Stone y Koo consiste en ajustar un efecto principal continuo por una segmentación de suavizadores polinomiales de grado 3, para lo que se generan fórmulas para funciones base, que permiten ajustar splines restringidos a la linealidad en las colas, es decir, linealidad por encima del último nudo y por debajo del primero.

Los splines k *-anudados de Stone y Koo*, además de ser conservadores comparados con las polinomiales puras, en el sentido de que la extrapolación fuera del rango de los datos es una línea recta, en vez de una curva polinomial, se construyen con un número relativamente pequeño de funciones de base de cálculo sencillo, lo que conduce a modelos menos complejos.

Si V_r es una *variable continua* que forma parte de la componente no lineal del modelo y k es el número de nudos, introducidos sobre el eje V_r, localizados en $\xi_1 < \xi_2 < ... < \xi_k$, las

funciones de base generadas por el método de Stone y Koo para V_r, $h_1(V_r),...,\ h_{k-1}(V_r)$, se obtienen en la forma siguiente:

$$h_1(V_r) = V_r$$

$$h_l(V_r) = \left(V_r - \xi_{r(l-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(k-1)}\right)_+^3 \left(\xi_{rk} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} + \frac{\left(V_r - \xi_{rk}\right)_+^3 \left(\xi_{r(k-1)} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}},\ l = 2,...,k-1$$

$$\text{donde } (Z)_+ = \begin{cases} Z & \text{si } Z > 0 \\ 0 & \text{si } Z \le 0 \end{cases}$$

y la expansión de V_r combinación lineal de tales funciones de base se expresa entonces en la forma:

$$RCS\_V_r = \beta_r V_r + \sum_{l=2}^{k-1} \theta_{r+l-1} \left(\left(V_r - \xi_{r(l-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(k-1)}\right)_+^3 \left(\xi_{r,k} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} + \frac{\left(V_r - \xi_{rk}\right)_+^3 \left(\xi_{r(k-1)} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} \right)$$

es decir, se trata de expandir las variable V_r a través de splines de regresión cúbicos restringidos, RCS, con k nudos, $\{\xi_{r1},...,\xi_{rk}\}$. Nos referiremos a la expansión correspondiente a la variable V_r como RCS_V_r.

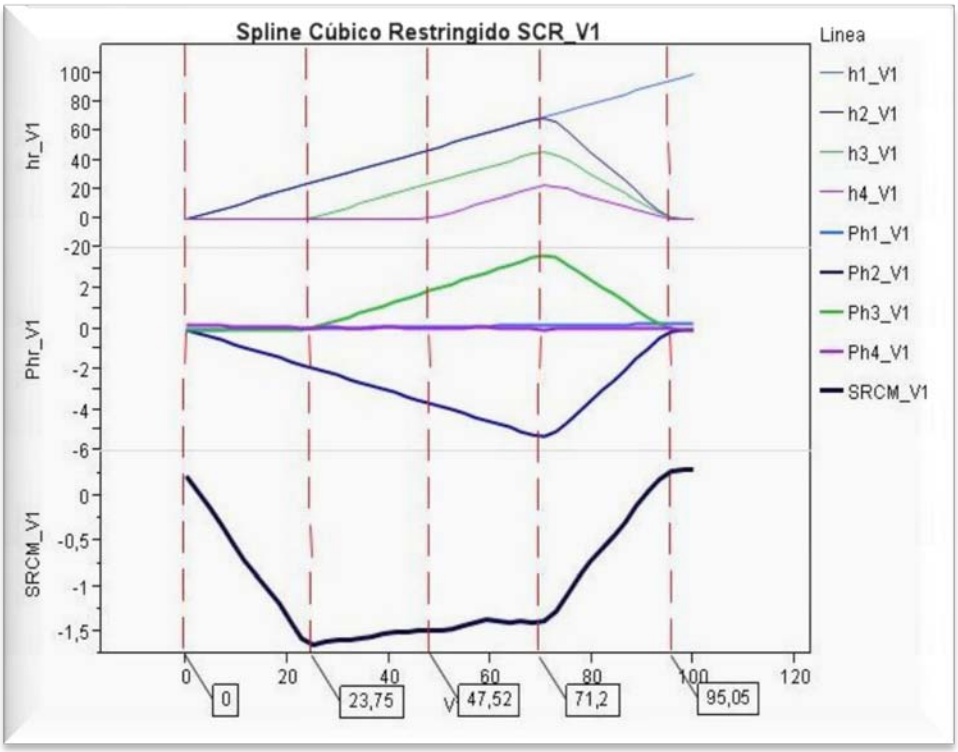


Figura 6.5.- Spline Cúbico Restringido, de Stone y Koo, con 5 nudos para la variable V_1.

Nótese que las cuatro funciones de base $h_r\_V_1$ son funciones de V_r y de los nudos, pero son independientes de Y. Por no tener razones suficientes para suponer una ubicación concreta de los nudos, los fijamos de acuerdo con las siguientes reglas empíricas:

Por regla general los nudos suelen situarse entre 3 y 7. Menos de 5 para un número pequeño de acreditados, $N \leq 100$, y 5 o más para un número suficientemente grande, $N > 100$. De existir un número suficientemente grande (N>100) de acreditados, se colocan el primero y el último nudos en los percentiles 5 y el 95 respectivamente. Criterios alternativos pueden verse en HARREL (2001).

En el spline cúbico mostrado en la figura 6.5, correspondiente a la variable V_1 y que en el capítulo 7 usaremos para conseguir la especificación del modelo de proactivo más adecuado, se consideran 5 nodos, puesto que contamos con una muestra de 36.607 acreditados. El nodo correspondiente al percentil 5 es $k_1 = 0$ y el correspondiente al percentil 95 es $k_5 = 95,05$. Los otros tres nudos se obtienen dividiendo el rango de V_1 entre los nudos $k_1 = 0$ y $k_5 = 95,05$ en cuatro partes iguales y considerando las fronteras de las 4 regiones obtenidas que se corresponden con $k_2 = 23,75$, $k_3 = 47,52$ y $k_4 = 71,2$.

Para ilustrar el método, ajustamos a nuestros datos de entrenamiento por regresión logística el modelo HLLM siguiente

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right) = \beta_0 + \sum_{r=1}^{16}\beta_r U_r + \beta_{17}\left(WT\_Vr\right) + \beta_{19}\left(W\_U_{19}\right) + \sum_{l=1}^{4}\theta_{1,l}h_{1,l}(V_1) \qquad (6.50)$$

Como resultado se obtiene que sólo las funciones de base $h_{1,2}(V_1)$ y $h_{1,3}(V_1)$ resultan linealmente significativas. Por tanto, repetimos el ajuste LLR del modelo (6.50) pero ahora el último término adopta la forma $\sum_{k=2}^{3}\theta_{1,k}h_{1,k}(V_1)$. De este modo el modelo resulta totalmente lineal y la expansión por funciones de base de V_1 es

$$RCS\_V_1 = \sum_{l=2}^{3}\theta_{1l}h_{1l} = \ -0.076079\ h_{12}\ + 0,097176\ h_{13}$$

Los coeficientes de la expresión anterior se corresponden con los coeficientes de máxima verosimilitud estimados por LLR sobre el modelo logístico lineal híbrido (6.50), donde se sustituye el último término por $\sum_{k=2}^{3}\theta_{1,k}h_{1,k}(V_1)$).

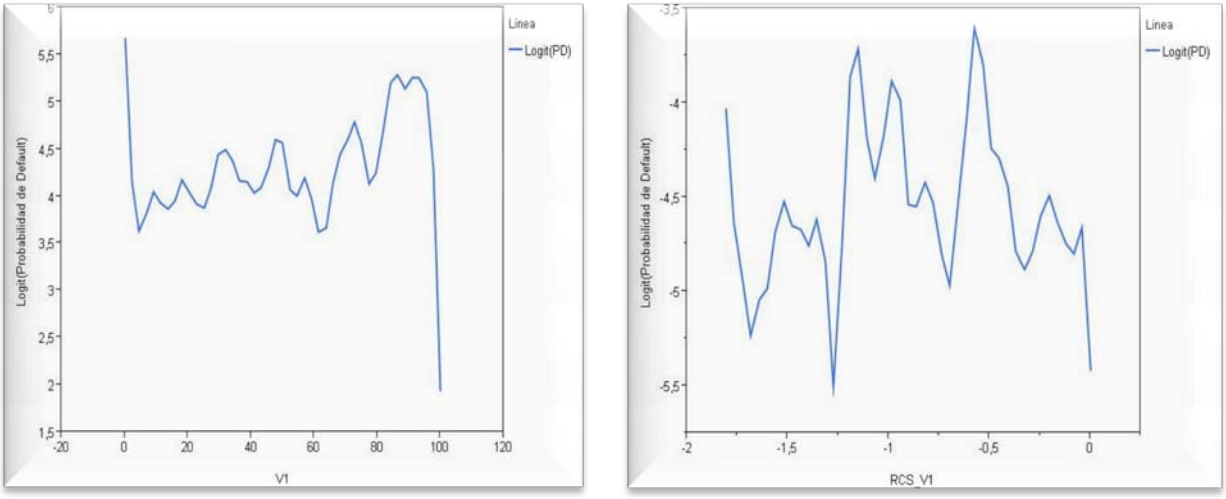


Figura 6.6.- Representación gráfica del logit de la probabilidad de default frente a la variable V_1, panel de la izquierda, y frente a la expansión RCS_ V_1, panel de la derecha.

En la figura 6.6 se muestra la representación gráfica de los logit de la probabilidad de default obtenidos por ajuste LLR de las *funciones de base* h_r *linealmente significativas en el* modelo HLLM (6.50), $h_{1,2}(V_1)$ y $h_{1,3}(V_1)$, frente a V_1 y la expansión de esas funciones base, RCS_ V_1.

6.7.5 Funciones de base bisagra obtenidas por MARS univariante.

Otra familia importante de expansiones lineales por funciones de base para una variable X_j es aquella en que las funciones de base son *funciones bisagra* resultado de aplicar el método MARS a modelos univariantes con variable respuesta el estado de default y variable independiente X_j.

MARS usa expansiones en funciones de base lineales a trozos de la forma $\left(X_j - t\right)_+$ y $\left(t - X_j\right)_+$ donde "+" significa parte positiva. Es decir la combinación lineal MARS univariante para la variable X_j, viene dada por

$$MARS\_X_j = \sum_{l=1}^{k_j} \theta_{jl} h_{jl}(X_j) \tag{6.51}$$

donde

$$h_{jl}(\boldsymbol{x}) = \begin{cases} b(x) \in \boldsymbol{B} \\ o \\ \bigotimes_{l=1}^{K_j} b_l; \quad b_l \in \boldsymbol{B} \end{cases} \tag{6.52}$$

y

$$\boldsymbol{B} = \left\{ \left(x_j - t\right)_+ ,\ \left(t - x_j\right)_+ ;\ \ t \in \left\{x_{1j}, x_{2j}, ..., x_{Nj}\right\},\ \ j = 1, 2, ..., p \right\}$$

$$\left(X_j - t\right)_+ = \begin{cases} X_j - t, & si\ X_j > t \\ 0 & en\ otro\ caso \end{cases} \qquad (6.53)$$

$$\left(t - X_j\right)_+ = \begin{cases} t - X_j, & si\ X_j < t \\ 0 & en\ otro\ caso \end{cases}$$

es decir, splines lineales a trozos y sus productos tensoriales seleccionados.

Por ejemplo, para la variable V_1 y nuestros datos de entrenamiento, las funciones de base "bisagra" seleccionadas como resultado de aplicar el método MARS al modelo univariante con variable respuesta el estado de default y variable independiente V_1, son

$$h_{11} = máx\left(0,\ V_1 - 98{,}9\right),\ h_{12} = máx\left(0,\ 98{,}9 - V_1\right),\ h_{13} = máx\left(0,\ V_1 - 1{,}5\right),$$

$$h_{14} = máx\left(0,\ V_1 - 8{,}36\right),\ h_{15} = máx\left(0,\ V_1 - 78{,}74\right),\ h_{16} = máx\left(0,\ V_1 - 73{,}8\right),$$

$$h_{17} = máx\left(0,\ V_1 - 79{,}94\right),\ h_{18} = máx\left(0,\ V_1 - 68{,}97\right).$$

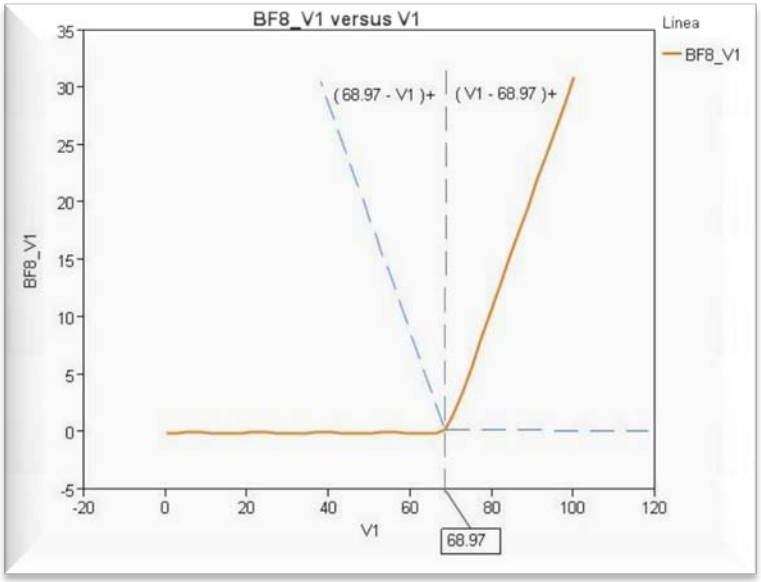


Figura 6.7.- Representación gráfica de la función bisagra $(V_1 – 68{,}97\ )_+$ frente a V_1, en naranja sólido. En azul punteado se representa la función bisagra espejo de la anterior, $(68{,}97\text{-}V_1\ )_+$, que MARS no seleccionó para expandir V_1. Como puede observarse poseen forma de "bisagra".

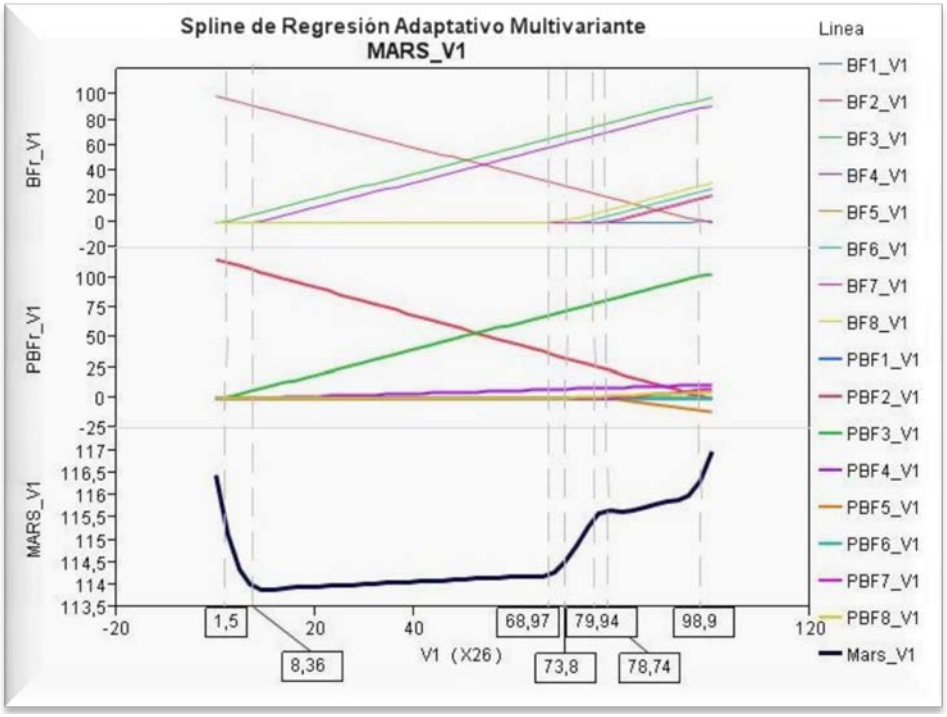


Figura 6.8.- Combinación lineal MARS para la variable V_1.

Con el fin de distinguir las funciones bisagra MARS de los splines cúbicos restringidos notaremos la función bisagra l-ésima para la variable X_j por la notación habitual de estas funciones en la literatura sobre MARS, $BF_l\_X_j$.

En la figura 6.8 se muestra, en la parte superior, $BF_r\_V_1$, la representación gráfica de las ocho funciones bisagra MARS frente a la variable V_1, en el eje de abscisas. En la parte central, $PBF_r\_V_1$, también frente a V_1, se muestra la representación gráfica de las mismas funciones ponderadas por los coeficientes estimados por regresión logística lineal, LLR. Por último, en la parte inferior, $MARS\_V_1$, se muestra la combinación lineal de todas las funciones FBr seleccionadas por el método MARS ponderadas por los coeficientes estimados.

Para ilustrar el método, ajustamos a nuestros datos de entrenamiento por regresión logística el modelo (6.50) donde el último término adopta inicialmente la forma

$$\begin{aligned}\sum_{l=1}^{8}\theta_{1,l}BF_{1,l} = \; & \theta_{1,1}\,máx\left(0,\,V_1-98{,}9\right)+\theta_{1,2}\,máx\left(0,\,98{,}9-V_1\right)+\theta_{1,3}\,máx\left(0,\,V_1-1{,}5\right)\\ & +\theta_{1,4}\,máx\left(0,\,V_1-8{,}36\right)+\theta_{1,5}\,máx\left(0,\,V_1-78{,}74\right)+\theta_{1,6}\,máx\left(0,\,V_1-73{,}8\right)\\ & +\theta_{1,7}\,máx\left(0,\,V_1-79{,}94\right)+\theta_{1,8}\,máx\left(0,\,V_1-68{,}97\right)\end{aligned}$$

Como resultado del ajuste por LLR de este modelo se obtiene que sólo son linealmente significativas las funciones de base $BF_{1,2}$, $BF_{1,3}$ y $BF_{1,4}$, por lo que considerando la siguiente combinación lineal de funciones base

$$MARS\_V_1 = \sum_{l=2}^{4} \theta_{1l} BF_{1l} = \; 0{,}945315\ máx\left(0,\ 98{,}9 - V_1\right) + \ 0{,}766475\ \ máx\left(0,\ V_1 - 1{,}5\right)$$
$$+ \ 0{,}194814\ \ máx\left(0,\ V_1 - 8{,}36\right)$$

el modelo resulta totalmente lineal y los coeficientes de la expresión anterior se corresponden con los coeficientes de máxima verosimilitud estimados por LLR sobre el modelo logístico lineal híbrido (5.50), donde se sustituye el último término por $\sum_{l=2}^{4} \theta_{1l} BF_{1l}$.

En la figura 6.9 se muestra la representación gráfica de los logit de la probabilidad de default obtenidos por ajuste LLR de las *funciones de base FBr linealmente significativas* en el modelo HLLM, (6.50), BF2, BF3 y BF4, frente a V_1 y la expansión de esas funciones base, MARS_ V_1.

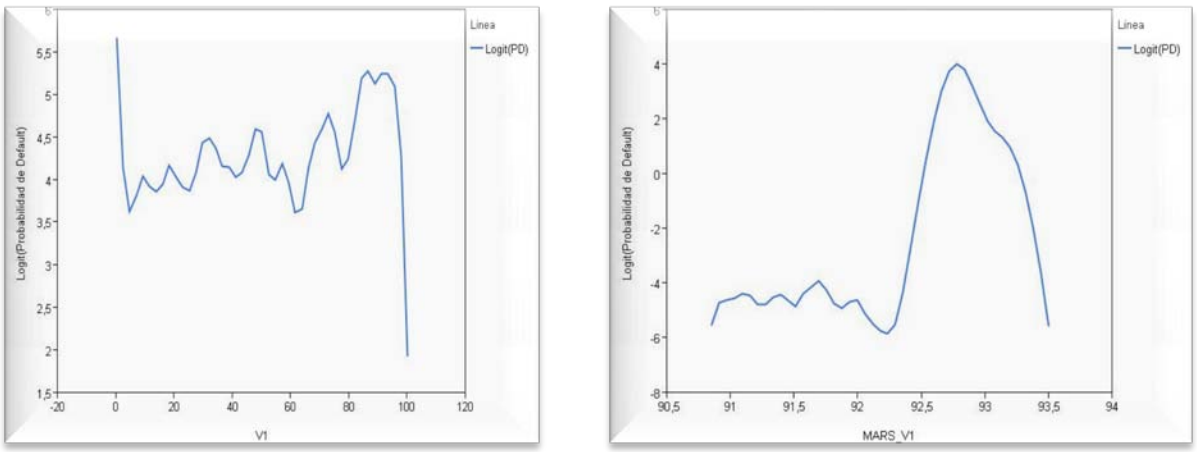


Figura 6.9.- Representación gráfica del logit de la probabilidad de default frente a la variable V_1, panel de la izquierda, frente a la combinación lineal de las funciones de base MARS_ V_1, panel de la derecha.

6.7.6 Funciones de Base Radial, RBF.

Una Función de Base Radial, RBF*, es una función de valores reales que dependen solo de la distancia al origen, es decir,*

$$\phi(X) = \phi(\|X\|) \tag{6.54}$$

o, alternativamente de la distancia a algún otro punto c, llamado centro, es decir,

$$\phi(X, c) = \phi(\|X - c\|) \tag{6.55}$$

Usualmente la norma es la Euclídea, aunque pueden usarse otras normas.

Típicamente se usan sumas de funciones RBF para aproximar funciones dadas.

Las funciones radiales son funciones cuya respuesta cambia monótonamente (crece o decrece) con la distancia a un punto central, tal como la función Gaussiana. Los parámetros del modelo son el centro, la métrica y la forma de la función radial.

La forma más general para una función de base RBF es:

$$h_r(X) = \varphi\left((X-c)^T M (X-c)\right)$$

donde

$\varphi(\cdot)$ *es la función usada (Gaussiana, Multicuadrática,..)*

c es el centro y M es la métrica (6.56)

La cantidad $(X-c)^T M(X-c)$ es la distancia entre la observación X de las variables explicativas y el centro definida por la métrica M.

Pueden usarse múltiples tipos de funciones, las más usuales en la estadística del aprendizaje máquina son:

Cauchy $\varphi(z) = \dfrac{1}{1+z}$

Gaussiana $\varphi(z) = e^{-Z}$

Splines Poliharmónicos $\varphi(z) = \begin{cases} Z^k, & k = 1,3,5,..., \\ Z^k Ln(Z), & k = 2,4,6,..., \end{cases}$

Splines Lamina Fina $\varphi(z) = Z^2 Ln(Z)$

Nosotros utilizaremos en esta Tesis Doctoral la versión de RBF Gaussiana, RBFG, que se expresa en la forma

$$h_r(X) = \exp\left(\frac{(X-c)^2}{r^2}\right) \tag{6.57}$$

donde la métrica viene definida por la esferas de centro c y radio r y es monótona decreciente desde el centro.

Para la variable V1 y nuestros datos de entrenamiento, la función de base radial RBFG, sobre la esfera unidad, es decir centro cero y radio 1, viene dada por

$$RBFG\_V_1 = \exp\left(-V_1^2\right) \tag{6.58}$$

Su representación gráfica se muestra en la figura 6.110.

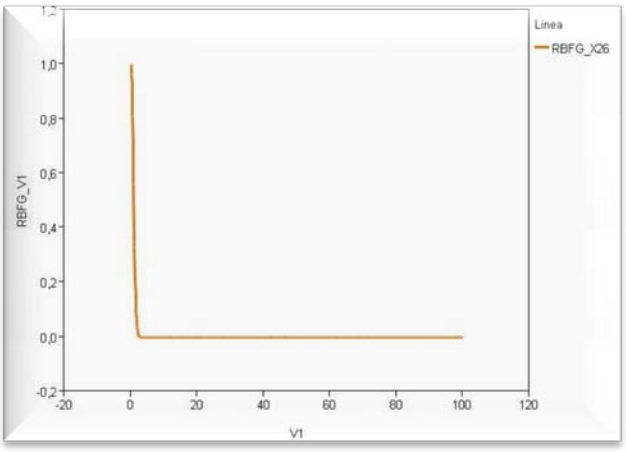


Figura 6.10.- Representación gráfica de la función radial RBFG con centro en cero y radio la unidad. Obsérvese que RBFG decrece monótonamente desde el centro, en cero.

Como viene siendo habitual ilustraremos el método, ajustando el modelo HLLM (6.50) a nuestros datos de entrenamiento por regresión logística lineal. La combinación lineal de funciones de base radial Gaussiana para V_1 resultó ser

$$Z_1(V_1)\_RBFG = \theta_{11} RBFG\_V_1 = \ 0.842356 \ \ \exp\left(-V_1^2\right)$$

El coeficiente de la expresión anterior se corresponde con el coeficiente de máxima verosimilitud estimado por LLR sobre el modelo logístico lineal híbrido (6.50), donde se sustituye su último término por la expansión por funciones de base radial $Z_1(V_1)\_RBFG$.

En la figura 6.11 se muestra, panel de la izquierda, la representación gráfica del logit de la probabilidad de default obtenido por ajuste LLR frente a V_1 y frente a $Z_1(V_1)\_RBFG$ en el panel de la derecha.

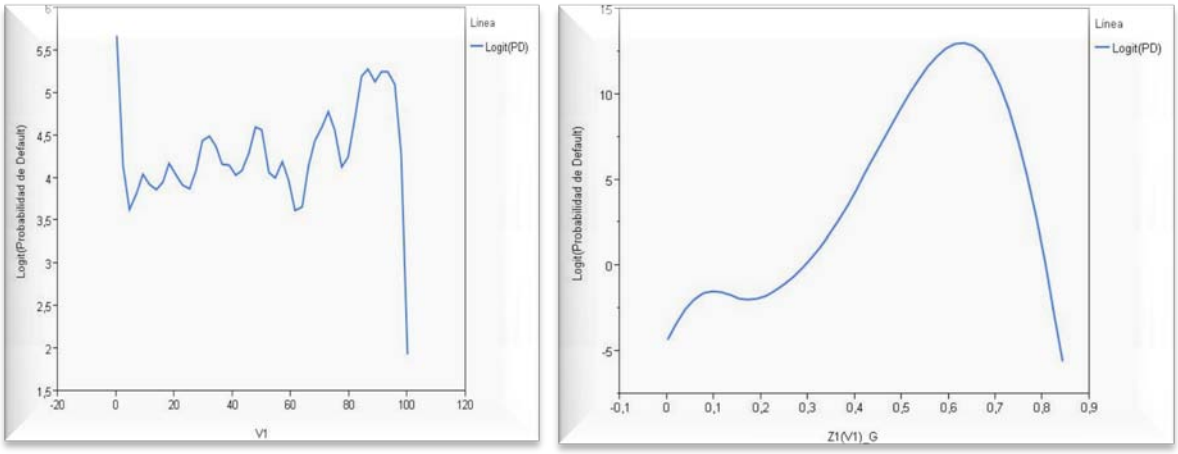


Figura 6.11. Representación gráfica del logit de la probabilidad de default frente a la variable V_1 panel de la izquierda, frente a la combinación lineal de funciones de bas $Z_1(V_1)\_RBFG$, panel de la derecha.

CAPÍTULO 7

CONSTRUCCIÓN DE UN MODELO DE CREDIT SCORING PROACTIVO HLLM.

7.1. INTRODUCCIÓN.

Este capítulo está dedicado a la *construcción de un sistema proactivo de calificación de acreditados utilizando Modelos Logísticos Lineales Híbridos por Expansiones Lineales de Funciones de Base,* HLLM, propuestos en el capítulo 6 de esta Tesis Doctoral, a través de las 8 fases de que consta generalmente el desarrollo de un modelo de credit scoring, mostradas en el esquema representado en la figura 7.1.

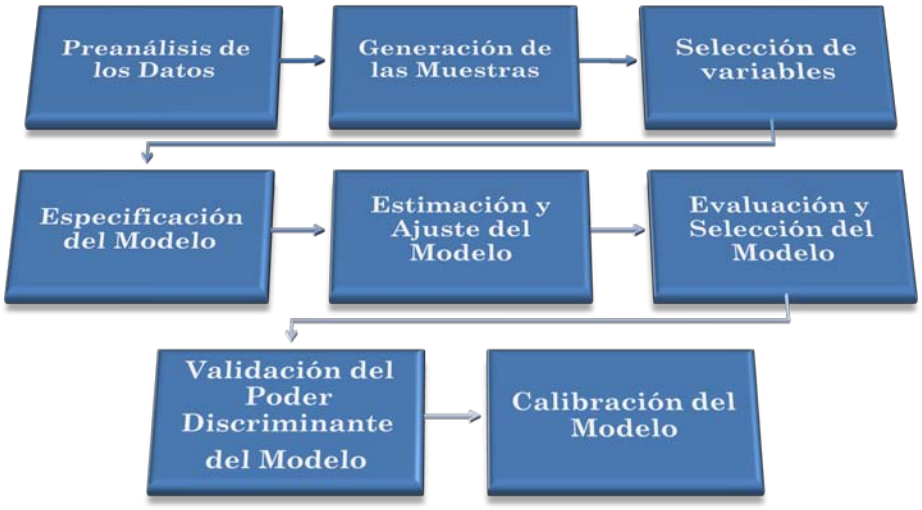


Figura 7.1.- Fases típicas del desarrollo de un modelo de Credit Scoring.

Sin duda alguna, *una de las cuestiones más crítica de cualquier tesis de esta naturaleza la constituye el proceso de extracción de la información* que está disponible en las Entidades Financieras y que, por tanto, no es impedimento para abordar el proyecto si se cuenta con la colaboración de alguna de ellas. Este es el caso de esta Tesis Doctoral, puesto que hemos contado con la colaboración de una Entidad Financiera española que puso a nuestra disposición la información de su magnifico datamart de riesgos y la larga experiencia tanto de sus especialistas informáticos en procesos de extracción de la información como la de sus expertos analistas en detección de la información relevante para el riesgo de crédito.

Para la construcción de nuestro modelo proactivo de credit scoring se considera el *segmento de clientes particulares con préstamos de la Entidad con fechas de visión 30 de noviembre de 2007 y 2008.* El segmento está constituido por clientes particulares de la Entidad Financiera, que a fecha de observación tienen únicamente préstamos, con garantía personal o hipotecaria, como productos de activo, que no tienen préstamos con finalidad fuera del

grupo de financiación a particulares, es decir, carecen de productos típicos de empresa, pudiendo ser además titulares de productos de pasivo.

Respecto a la distribución por tipo de producto, más del 60% de los clientes es titular de algún préstamo con garantía hipotecaria, de estos el 10% tiene además un préstamo personal. Prácticamente el 10% de los clientes es titular de algún producto de pasivo no a la vista a fecha de observación.

Respecto del plazo, el promedio para préstamos de consumo se sitúa ligeramente por encima de los 4 años, y en torno a los 18 años para las hipotecas.

En lo referente al importe formalizado, el promedio en préstamos con garantía personal se acerca a los 10.000 euros y en préstamos con garantía hipotecaria supera ligeramente los 75.000 euros.

A fecha 30 de noviembre de 2007 el segmento constaba de 93.761 clientes distribuidos respecto del incumplimiento como se muestra en la tabla 7.1.

Tabla 7.1.- Distribución del indicador de incumplimiento de los acreditados particulares con préstamos.

Incumplimiento	Frecuencia	Porcentaje
Buenos	71.312	76,1%
Malos	5.750	6,1%
Indeterminados	16.699	17,8%
Total	93.761	100,00%

Tras aplicar, en primer lugar, los criterios de selección establecidos por la política de riesgos de la Entidad y las recomendaciones que al respecto se incluyen en el Nuevo Acuerdo de Capital de Basilea II y después de eliminar los clientes con *comportamiento intermedio* o *indeterminado*, con objeto de construir un modelo que discrimine al máximo entre buenos y malos pagadores, el segmento se fijó en 77.062 clientes, de los cuales 71.312, el 92.50%, presentan un estado de cumplimiento de sus obligaciones de pago, estado de default, considerado como *"bueno"* mientras que 5.750, el 7.50% restante, presenta un estado *"malo"*. Se prescindió de los clientes indeterminados puesto que en la construcción de modelos de *credit scoring*, se catalogan como *indeterminados* aquellos clientes que desde el punto de vista de sus comportamiento frente al cumplimiento de sus obligaciones de pago, no pueden ser asignados ni al grupo de buenos ni al de malos de forma clara, ya que éstos aportan

perfiles intermedios de comportamiento que dificultarían nuestro objetivo de conseguir modelos que separen lo mejor posible a los clientes con "*buen comportamiento"* de aquellos que presenten un "*mal comportamiento".*

Como consecuencia, se cuenta con 63 variables observadas sobre el segmento de 77.062 clientes de referencia, relativas a la *relación global del cliente con la Entidad*, a la *relación a través de productos de pasivo y de riesgo*, a su *relación a través del consumo de servicios ofrecidos por la Entidad*, a su *vinculación a través de la domiciliación de la nómina y/o recibos* y al *comportamiento del cliente con la Entidad en los productos de riesgo contratados* así como a *la calificación, según la definición de incumplimiento del Nuevo Acuerdo de Capital de Basilea II, que la Entidad asigna al cliente de acuerdo con su comportamient*o.

Una vez que se ha construido una base de datos con la información proporcionada por la Entidad Financiera, que incluye el conjunto de 63 características de riesgo de crédito, $X=\left(X_1,...,X_{63}\right)^T$, relacionadas con el estado de default, variable respuesta, Y, el siguiente paso consiste realizar un preanálisis que permita obtener un conocimiento preliminar de las características particulares del cliente y seleccionar y preparar, si es necesario, la adecuada información relevante para el proceso de modelización.

L*a forma en que se acometen la mayor parte de las fases de construcción de un modelo de credit scoring depende fundamentalmente de su estructura formal y de la metodología que se adopte para la estimación, evaluación, validación y calibración del mismo*. Existen varios métodos para este cometido, y tenemos interés en aquellos que no sólo implican establecer y cuantificar la relación entre las características de riesgo de crédito observadas sobre los acreditados y el comportamiento de estos frente al default a través de un modelo estadístico o de un algoritmo de aprendizaje, sino que además permitan, de acuerdo con los requerimientos de Basilea II, la evaluación y validación necesarias para asegurar la generalización del modelo, así como su calibración para detectar a posteriori la eficacia del modelo en sus pronósticos. Además, el modelo deberá explicar suficientemente la relación de dependencia entre el estado de default y las variables, que relacionadas con el riesgo de crédito, hayan sido seleccionadas para formar parte del mismo.

La consideración de todas las características anteriores junto con un primer análisis exploratorio de la información disponible, fase de preanálisis, y la conveniente concurrencia de conocimiento experto en riesgo de crédito habrán de guiarnos hacia la familia de modelos y la metodología más adecuada a nuestros objetivos.

7.2 PREANÁLISIS DE LOS DATOS POBLACIONALES.

7.2.1 Introducción.

En la fase de preanálisis se realiza <u>*en primer lugar*</u> una *clasificación conceptual de las variables en visiones del cliente, una General y cinco específicas relacionadas con el Pasivo, el Activo, los Servicios, las Incidencias y el Incumplimiento*. En esta tarea es fundamental profundizar en las características más importantes de las variables, tanto desde su especificación, tipo de datos, tipo de variable, rol específico, etc., como desde su significado e importancia para medir o explicar el riesgo de crédito.

<u>*En segundo lugar*</u>, se realiza un *análisis estadístico individual para cada variable*, abarcando los aspectos siguientes:

a) Resolución del problema de *datos faltantes* (en el caso de pocas faltas, a través de los métodos clásicos, y en el resto asignando el peso de la evidencia, WOE, a las categorías resultantes de un proceso de discretización optima, inicialmente automático, MDLP, basado en la *entropía de Shannon*, (FAYYAD e IRANI (1993), SPSS, IBM®, a partir del cual, para obtener el tramado definitivo, donde se incluirá la clase de valores faltantes, se realiza una discretización supervisada usando el algoritmo *Interactive Binning* del programa *Enterprise Miner* de SAS®.

   Resolveremos el problema de número elevado de datos faltantes sustituyendo las variables afectadas por las correspondientes variables tramadas.

b) Eliminación de los *valores extremos* que pudieran distorsionar la construcción del modelo.

<u>*En tercer lugar*</u>, se acomete un examen cuidadoso de la relación y su significado, en términos de poder de explicativo, de todas y cada una de las variables con las demás. Esto conlleva analizar cuestiones como la *correlación*, la *multicolinealidad* o la *asociación parcial* con el estado de default.

a) Con respecto a la detección de la *correlación*, además del *análisis de la correlación bivariante* utilizando la matriz de correlaciones, utilizaremos una variante del *Análisis de Componentes Oblicuas* (OCA), el procedimiento PROC VARCLUS, im
b) plementado en SAS® / STAT V9.2, para agrupar las variables explicativas en clases de variables caracterizadas por su interrelación.

c) Con el fin de evitar, entre otros problemas, que pequeñas variaciones de los datos puedan provocar cambios significativos en los coeficientes del modelo, se analiza y elimina la *multicolinealidad entre las variables* analizadas. Para ello se utiliza el *Factor de Inflación de la Varianza*, VIF, (BELSEY et al. (1980) y KLEINBAUM et al. (1988)), y los *Índices de Condición*, CI, y *Proporción de Varianza*, VP. Los dos últimos indicadores se obtienen a partir de los valores propios de la matriz de correlaciones entre las variables independientes. Los tres estadísticos se obtienen con una macro rutina de confección propia en SAS® V9.2.

d) Un aspecto clave consiste en *detectar las variables con mayor potencial de información en relación con el estado de default de los acreditados*, tanto en cantidad como en calidad, es decir, detectar aquellas variables con *alto poder explicativo del estado de default o que presenten suficiente asociación con el mismo*.

El *poder explicativo de la variable* se medirá a través de la cantidad de información proporcionada por ésta, para lo que se utilizarán los *estadísticos de Gini y Valor de la Información,* IV, (KULBACK, 1959), y *la asociación entre la variable explicativa y el estado de default* se mide utilizando los adecuados test de asociación con la variable respuesta, en función de que las variables sean de intervalo, (eta, R_cuadrado), o nominales, (Chi_cuadrado y V de Cramer), descartando aquellas variables para las que la cantidad de información no sea suficiente y/o los test de asociación no indiquen asociación. En el primer caso se utilizó el algoritmo Interactive Binning del programa Enterprise Miner de SAS®, y en el segundo el módulo Crosstabs de SPSS, IBM®.

Como consecuencia de a), b) y c) y conocimiento experto de riesgo de crédito se realiza una primera selección de las variables más adecuadas a nuestros objetivos. Así, por ejemplo, una vez aplicado PROC VARCLUS y obtenidos los distintos grupos de variables y detectada la indeseable colinealidad, algunas variables en cada grupo pueden ser consideradas como variables de pronóstico en base a su Valor de la Información (IV) o correlación con la variable respuesta, análisis R^2, en función de si las variables se han agrupado o no. Este proceso implica una reducción de variables, descartando aquellas que mostrando colinealidad con otras poseen escaso poder explicativo sobre el estado de default o que presenten escasa asociación con el mismo. Además seguimos criterios razonables respecto del riesgo de crédito, como no eliminar todas las variables de la misma visión de la relación

del cliente con la Entidad Financiera y en equilibrio con criterios estadísticos se descartamos aquellas que aportan información redundante dentro de una visión concreta.

El objetivo de la reducción de variables en los términos anteriores consiste en mantener un conjunto compacto de variables candidatas a explicar y predecir el modelo que pueden ayudar en la construcción del mismo, sin perder capacidad potencial de predicción.

En *cuarto lugar*, tras la exclusión de aquellas variables que por razones estadísticas y de riesgo de crédito no resulten adecuadas para la especificación del modelo, se acometerá la tarea de *obtener una muestra representativa de la población objetivo*, de forma que los resultados obtenidos en ésta sean aplicables a toda la población. Con ello se consigue, además de trabajar con un número de clientes más reducido, por un lado, maximizar el número de clientes malos en relación con el de buenos, con el fin de poder capturar más fácilmente los distintos perfiles de comportamiento, y, por otro, eliminar posibles sesgos que puedan distorsionar la relación de dependencia capturada por el modelo. La muestra se obtiene por el método *sobre muestreo estratificado*, que básicamente consiste en construir la muestra estratificada por la variable *incumplimiento de las obligaciones de pago* y aplicar el sobre muestreo, selección deliberada de acreditados default, para obtener estimaciones razonablemente precisas de las propiedades de estos, para evitar que, al construir el modelo, las características de los malos queden ocultas por la gran proporción de clientes buenos de la población.

A partir de esta muestra se obtienen tres submuestras: (1) *muestra de entrenamiento*, que se utiliza para ajustar el modelo con el criterio de minimizar el error de intervalo, (2) *muestra de validación*, a través de la cual se estima el error de predicción esperado con el fin de seleccionar el modelo adecuado y (3) *muestra test*, con la que se calcula el error de generalización del modelo finalmente elegido.

7.2.2 Clasificación conceptual de las variables.

La información de detalle proporcionada por la Entidad Financiera, relevante para el desarrollo y explotación de los modelos de credit scoring proactivos, se estructura en una serie de *"visiones" parciales*, cada una de las cuales proporciona una perspectiva concreta de la relación del cliente con la Entidad a través de un determinado tipo de información.
Usualmente se suelen crear las siguientes visiones parciales:

(1) *Visión General del Cliente*: proporciona una visión del cliente y/o de su relación con la Entidad Financiera, desde una perspectiva global a fecha de observación, que contiene

información socio demográfica, información sobre la relación de actividad del cliente con la Entidad, información sobre los productos de activo y pasivo del cliente.

(2) *Visión de Pasivo:* la visión de pasivo proporciona la visión de la relación del cliente con Entidad a través de los productos de pasivo que tenga contratados. Por tanto, por un lado se consideran diversas variables que reflejan la operativa en cuentas de pasivo a la vista y, por otro, consumo y saldos de otros productos de ahorro e inversión, pasivo no a la vista.

(3) *Visión de Activo*, que proporciona la visión de la relación del cliente con la Entidad a través de los productos de riesgo con los que trabaje. Se consideran variables de activo como medida de comportamiento histórico en el cumplimiento de las obligaciones crediticias. Los bloques de préstamos con garantía personal, con garantía hipotecaria, total préstamos y el de total activo son especialmente relevantes para la modelización del segmento, a*horro en descubierto, préstamos de finalidad particular con garantía personal, préstamos de finalidad particular con garantía hipotecaria, Importe dispuesto en Cirbe.*

(4) *Visión de Servicios* formada por variables que miden aspectos relacionados con el grado de vinculación, como la *domiciliación de nómina, domiciliación de recibos y el uso de los diversos servicios ofrecidos por la Entidad, como tarjeta de débito, banca electrónica, etc.*

(5) *Visión de Incidencias* que proporciona la perspectiva del comportamiento del cliente con la Entidad Financiera en los productos de riesgo que tiene contratados. Es el porcentaje de días en incidencia del cliente, en distintos tipos de préstamos durante un periodo determinado, usualmente, mes, trimestre, semestre. Las incidencias en este segmento son originadas por los préstamos y el descubierto en ahorro vista.

(6) *Visión de Incumplimiento*: proporciona la calificación del Cliente dentro de la Entidad Financiera, en función de su comportamiento y según la definición de incumplimiento manejada en Basilea II.

Como resultado de esta fase, a cada una de las variables del conjunto de desarrollo se le ha asignado una clase dentro de los tres niveles de la estructura conceptual de información previamente descrita.

En total se consideran 63 variables referidas a las 6 visiones anteriores. En muchos casos, una misma variable, referida a un concepto de riesgo relacionado con el estado de default del cliente, da origen a 4 nuevas variables que miden el mismo concepto en distintos períodos de tiempo: mes, trimestre, semestre o año.

El paso siguiente en la fase de preanálisis consiste en analizar *la calidad de la información*, identificando variables que, ya sea por causas estadísticas o por criterios de riesgos de crédito, no reúnen los requisitos necesarios para ser tenidas en cuenta en el análisis posterior. Para ello será necesario un análisis estadístico individual estudiando la distribución de cada variable, observando la presencia de valores extremos y el número de valores ausentes, la existencia de concentraciones en determinados valores, etc. De este proceso detallaremos en el apartado siguiente los aspectos más relevantes.

7.2.3 Análisis Estadístico Individual de cada variable.

7.2.3.1 Eliminación de observaciones extremas.

En algunas ocasiones se pone de manifiesto que algunas variables presentan valores extremos, que pueden ser debidos a errores producidos en el momento de la extracción de los datos (lo más habitual), o a características excepcionales de los registros seleccionados (por ejemplo un cliente que tenga más de 90 años) que presenten valores muy poco frecuentes. Estos casos han de eliminarse de la muestra de análisis, porque pueden distorsionar el modelo final: si se incluyera un número elevado de valores extremos, el modelo trataría de ajustar lo mejor posible la variable objetivo a todos los valores de las variables de entrada; como consecuencia de ello, el modelo resultante podría estar *sobre ajustado*, es decir, funcionaría bien sobre la muestra en particular sobre la que se ha desarrollado pero no generalizaría bien, es decir, no se aplicaría bien sobre cualquier nueva muestra seleccionada. Nosotros hemos eliminado los casos extremos bajo el principio general de que *no se deben quitar demasiados casos malos ni demasiados casos buenos de la muestra.*

7.2.3.2 Tratamiento de datos faltantes.

Como suele ser habitual con este tipo de información, en nuestro conjunto de datos sobre los acreditados se observa que algunas de las variables presentan un número importante de valores faltantes, este es el caso de las variables X_{12}, X_{46}, X_{49}, X_{54}, X_{55} y X_{56} cuyo número de observaciones faltantes se sitúa entre el 40% y más del 60% del total de 73.207 acreditados.

Por su parte la variable X_{47}, aunque en menor medida que anteriores, 1,28%, también presenta datos faltantes.

Tabla 7.2.- Tabla de estadísticos y porcentajes de datos faltantes en las siete variables que los presentan.

		% Faltas				
	N		**Mínimo**	**Máximo**	**Media**	**Desv. típ.**
X_{12}	41716	43,02%	0,01	200069,47	6042,3623	11352,29266
X_{46}	42294	42,23%	0,00	60193,40	2953,6914	4970,07356
X_{47}	72270	1,28%	0,00	1996,49	384,0501	292,84481
X_{49}	28321	61,44%	0,00	129,00	22,9188	23,54738
X_{54}	38404	47,54%	0,00	100,00	49,4392	43,41272
X_{55}	41716	43,02%	0,00	100,00	16,4035	33,62197
X_{56}	41716	43,02%	0,00	998,61	81,1975	174,75649
N válido	**15015**					

Un modelo paramétrico ajusta los parámetros correspondientes a partir de casos para los que la información está completa, es decir, casos donde ninguna de las variables analizadas contiene valores ausentes, de no tomar medidas correctoras del problema sólo podríamos considerar para nuestro análisis 15.015 acreditados. Para maximizar el número de registros completos se aplican técnicas de imputación que dependen del porcentaje de datos faltantes y del tipo de variable que se esté analizando (continua ó discreta).

Es evidente que el porcentaje de datos faltantes admisible está en función de la naturaleza de los datos que se analizan, en nuestro caso *el segmento de clientes con préstamos a particulares*. Así, *aunque habitualmente se prescinde de variables con un porcentaje de valores faltantes superior al 40%, parece que debiera relajarse este umbral cuando la presencia de valores faltantes sea justificable por las características intrínsecas del segmento, ya que entonces dichos valores aportan "cierta información" sobre el cliente. Esto ocurrirá cuando el valor faltante no es debido a la imposibilidad de obtener información por parte de la Entidad respecto a determinado aspecto de la relación con el cliente, sino a que dicha información "no existe".* Así por ejemplo, en la mayoría de los registros de los clientes de préstamos no existe información sobre *pasivo no a la vista*, al estar presentes solamente en alrededor de un 30% de la cartera, pero los valores faltantes en estas variables son informativos, (por ejemplo el "número de contratos en ahorro no a la vista" sería equivalente a cero). En cualquier caso la ausencia de observaciones posee un significado especial desde el punto de vista del riesgo de crédito.

Al hilo del razonamiento anterior cuando nos encontramos con un porcentaje de valores ausentes superior al 5% consideramos el valor "*faltante*" como una categoría más de la variable de cara a la construcción del modelo, ya que puede constituir un grupo homogéneo frente al incumplimiento. El tramado de variables nos proporciona la posibilidad de integrar los valores faltantes en un grupo con un comportamiento similar ante el default o bien de mantenerlos definitivamente como un grupo independiente. En caso contrario (nº de datos faltantes inferior a 5%) optamos por el criterio de analizar si el dato faltante era equivalente a un cero, cuando es así se le imputa este valor.

La solución que hemos dado al problema datos faltantes discurre por dos vías:

En el caso de pocas faltas, a través de los métodos clásicos.

Para variables continuas a los datos faltantes equivalentes a cero se le imputa el valor medio de los valores informados o la mediana, (que está menos influenciada por valores extremos). Cuando la variable es categórica no tiene sentido analizar su equivalencia con el valor cero, así que les imputamos la moda de los valores informados.

En otro caso asignando el *peso de la evidencia*, WOE, a las categorías resultantes de un proceso de discretización optima, inicialmente automático y posteriormente adaptado con criterios de riesgos.

Tramado de variables.

Una forma eficaz de resolver problemas de datos faltantes, de datos extremos y de clases *"raras"* en variables explicativas del riesgo de crédito X_j, cuando estos abundan, consiste en sustituir las variables afectadas por estos problemas por funciones de base definidas en la forma siguiente.

$$h_j(X) = \sum_{k=1}^{K_j} I_{\left[X_j \in I_{jk}\right]} WOE\left(I_{jk}\right) \tag{7.1}$$

donde los K_j intervalos, I_k, constituyen una agrupación óptima de la variable X_j y

$$WOE\left(I_{jk}\right) = ln\left(\frac{\text{Proporción de clientes buenos en el intervalo } I_{jk}}{\text{Proporción de clientes malos en el intervalo } I_{jk}}\right) \tag{7.2}$$

son los pesos de la evidencia en el intervalo I_k de la variable X_j. Las variables construidas del modo anterior serán etiquetadas como W_X_j.

Una cuestión fundamental la constituye *la construcción de los intervalos o tramos* I_k. Para construir los tramos existen muchos algoritmos, algunos de ellos implementados en el abundante software actual de análisis estadísticos. En esta Tesis Doctoral utilizamos el *algoritmo de tramado óptimo* implementado en el *Nodo Interactive Binning* del *programa Enterprise Mining de SAS®*, que es una herramienta que computa clases iniciales por cuantiles que podemos partir o combinar interactivamente. Esta herramienta permite, además, seleccionar características con fuerte valor explicativo base al estadístico de Gini y el Valor de la Información.

El *Nodo Interactive Binning* obtiene de forma iterativa una partición recursiva utilizando un algoritmo tipo árbol de decisión. En primer lugar obtiene una partición inicial del rango de la variable según una de dos opciones posibles: el m*étodo de cuantiles*, que genera grupos con aproximadamente la misma frecuencia en cada grupo y del método del cubo que genera grupos resultantes de dividir los datos en intervalos iguales, sobre la base de la diferencia entre los valores máximo y mínimo del rango de la variable, nosotros hemos optado por el método de cuantiles. En segundo lugar aplica un algoritmo tipo árbol de decisión, para obtener la partición recursiva final del rango de la variable. De los tres métodos de agrupamiento disponibles (tasa de sucesos óptima, cuantiles y tasa de sucesos monótona), hemos seleccionado el método de la tasa de sucesos óptima con el criterio de reducción de la medida de la entropía de Shannon, método de discretización supervisada debido a FAYYAD e IRANI, (1993) y DOUGHERTY et al. (1995), sustentado sobre la base de la *entropía de la información de clase*, que mejora el algoritmo de discretización de RISSANEN (1989).

Una vez construidas las clases, se asigna como valor de cada tramo I_{Xk} de la partición del rango de X el *peso de la evidencia* correspondiente, $WOE(I_{Xk})$.

Habitualmente en variables X_l en las cuales el número de observaciones faltantes es elevado y se considera que las faltas son informativas se sustituye X_l por una transformación del tipo (7.1), pero nosotros daremos un paso más allá de considerar la variable W_X_l *construida tan sólo por criterios estadísticos*, puesto que *abordaremos la segmentación del rango de estas variables en distintos grupos combinando criterios de riesgos y criterios estadísticos*. El método puede resultar útil para elaborar datos que representen las tendencias de clasificación del riesgo en vez de datos obtenidos tan sólo a partir de técnicas estadísticas de optimización que podrían conducir a modelos sobre

ajustados. Bajo la doble perspectiva de criterios estadísticos y criterios de riesgos de crédito, pretendemos que los grupos formados posean las siguientes características:

1. *Han de ser fácilmente interpretables y deben ajustarse, en la medida de lo posible, a la realidad del análisis de riesgos de la Entidad.*
2. Han de ser *homogéneos*, es decir, con una suficiente representación de la población en cada grupo.
3. Deberán presentar *tendencias de morosidad suaves.*
4. Se *comportarán, en relación a la morosidad, de forma significativamente distinta en cada tramo.*

Para conseguir las cuatro características anteriores se tendrán en cuenta los siguientes requisitos:

- ✓ Aunque se buscan tramos de frecuencia homogénea, se admitirán grupos pequeños si son significativamente distintos en cuanto a su comportamiento.
- ✓ Se aislarán valores especiales dentro del rango de variación de la variable (missing, cero, etc.) aunque la frecuencia del grupo sea muy pequeña. Posteriormente se analizará la conveniencia de incluirlos en algún grupo con una tasa de incumplimiento similar.
- ✓ Los límites de los tramos se redondearán para que sean justificables desde el punto de vista de negocio, aunque ello suponga un pequeño empeoramiento del Valor de la Información.

Existen tres requisitos para un buen tramado, PORATH. (2006):

i) Cada clase debe tender a tener un mínimo de buenos y malos acreditados, de lo contrario la estimación de los parámetros del modelo tiende a ser imprecisa.

ii) Los acreditados agrupados en cada clase deben presentar un perfil de riesgo similar.

iii) La clasificación resultante debe revelar un plausible patrón de riesgo (como el indicado por el *peso de la evidencia*) y una alta eficacia (como la indicada por un alto *valor de la información*).

Cada partición realizada puede ser evaluada por el estadístico *valor de la información* construido a partir de los pesos de la evidencia. En los primeros tiempos del Credit Scoring, la consultoría de riesgo americana Fair Isaac adoptó esta medida a la que apodaron "*the*

información value" (el valor de la información, IV), para medir el poder explicativo de una característica. Este estadístico es una medida de divergencia debida a KULBACK (1959) y se usa para medir la diferencia entre dos distribuciones.

Definición 7.1.- Se define el estadístico *Valor de la Información de Kulback* en la forma siguiente:

$$IV = \sum_{Tramos} (P_{Buenos\ en\ el\ Tramo} - P_{Malos\ en\ el\ Tramo})\ WOE_{Tramo} \qquad (7.3)$$

El *valor de la información* es siempre positivo y cuanto mayor sea la información aportada por una variable tanto mayor será el valor de este estadístico.

Usualmente se considera que las características con VI menor que 0,01 proporcionan una débil información, mientras que *las que presentan valores mayores que 0,3 son las más idóneas para figurar en los modelos de credit scoring.*

La siguiente escala de valores permite identificar las variables según su poder explicativo:

$$\begin{array}{ll} IV < 0.02 & \text{sin capacidad predictiva.} \\ 0.02 \le IV < 0.10 & \text{capacidad de predicción débil.} \\ 0.10 \le IV < 0.30 & \text{capacidad predictiva media.} \\ 0.30 \le IV & \text{fuerte capacidad predictiva.} \end{array} \qquad (7.4)$$

Fijar un tramado es complejo a causa de que, por un lado, el valor de la información tiende a crecer con un número creciente de clases, y por otro lado, la estimación de los parámetros del modelo tiende a mejorar cuando el número de clases decrece.

En orden a conseguir un adecuado tramado final es necesario construir una serie de tablas cruzadas diferentes y calcular los correspondientes pesos de la evidencia y valores de información. El tramado final es, por tanto, el resultado de un proceso heurístico fuertemente determinado por el conocimiento y la experiencia del analista.

A continuación se acomete el tramado con criterios de riesgos de las variables X_{12}, X_{46}, X_{47}, X_{49}, X_{54}, X_{55} y X_{56}, utilizando de nuevo el modulo *agrupación interactiva* del programa Enterprise Miner de SAS®, resultando las variables WR_X_{12}, WR_X_{46}, WR_X_{47}, WR_X_{49}, WR_X_{54}, WR_X_{55} y WR_X_{56}, (se ha utilizado el prefijo WR_X_l para distinguirlas de las obtenidas por el tramado automático W_X_l. En las figura 7.2 se muestra el tramado estadístico realizado sobre X_{47} por el algoritmo de tramado optimo con el criterio de

minimización de la entropía de Shannon. Dadas las característica de la variable X47, la transformación W_X_{47} obtenida por tramado automático parece excesivamente artificiosa.

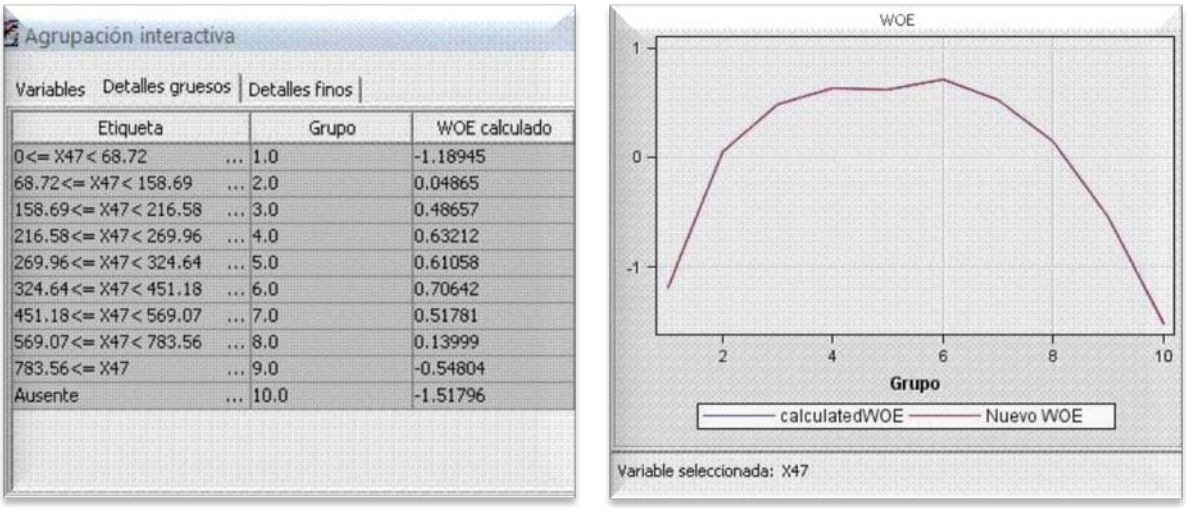


Figura 7.2.- Tramado Automático y Cálculo de WOE de X_{47} con SAS Enterprise Miner.

El tramado de la figura 7.3, realizado con criterios de riesgos parece mucho más adecuado, y dado que existen 937 acreditados para los que X_{47} presenta un dato faltante, se ha creado un tramo donde se recogen estos casos, es el tramo con etiqueta: "ausente" en el panel izquierdo de la figura 7.3.

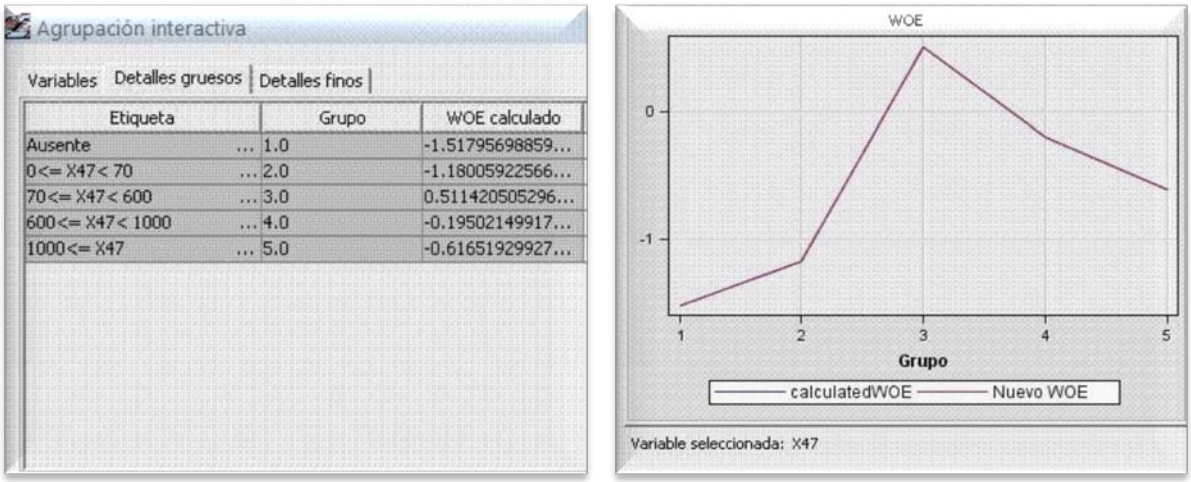


Figura 7.3.- WR_X_{47} – Tramado con criterios de riesgos y WOE de X_{47}, con SAS Enterprise Miner.

En la tabla 7.3 se muestra el valor de la información de las variables tramadas con criterios de riesgos, como puede observarse en dicha tabla, todas presentan fuerte capacidad explicativa del estado de default.

Tabla 7.3.- Valor de la información de las variables tramadas con criterios de riesgos.

Variable	Valor de la Información
WR_X12	0,909356
WR_X46	2,172027
WR_X47	0,460801
WR_X49	0,305135
WR_X54	0,385362
WR_X55	0,411351
WR_X56	0,501170

Otra de las características sobresalientes de *las variables tramadas* WR_X_l *es que en general poseen una distribución más suave que las variables originales* X_l. Las herramientas más utilizadas para explorar la distribución de una variable y, por tanto, de su correspondiente variable tramada son el *histograma y la densidad de probabilidad estimada no paramétricamente por funciones núcleo univariantes*, PARZEN (1962), en el caso de las variables métricas. Aquí utilizamos estimadores no paramétricos de la densidad por núcleos Gaussianos.

El *estimador por núcleos Gaussianos de la densidad* $f(x)$ de la variable aleatoria X, $\hat{f}(x)$, con *función núcleo* $K(\bullet)$, *viene dado por* (2.47)

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{7.5}$$

donde h es la amplitud de ventana y $K(\bullet)$ el núcleo Gaussiano

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2), \quad (-\infty < t < \infty) \tag{7.6}$$

Para la obtención del estimador (7.5) hemos utilizado el PROC KDE de SAS® V9.2, que usa el método SJPI, (Sheater-Jones Plug-In), JONES et al. (1996), para computar la amplitud de ventana h, método que resuelve la ecuación de punto fijo siguiente:

$$h = \left[\frac{\int_{-\infty}^{+\infty} K^2(x)dx}{N\int_{-\infty}^{+\infty}\left(f''_{g(h)}(x)\right)^2 dx\left(\int_{-\infty}^{+\infty} x^2 K(x)dx\right)^2}\right]^{1/5} \tag{7 .7}$$

donde N es el tamaño de la muestra y $g(h)$ es una función de la ventana de Parzen h.

Los métodos Plug-In de selección de ventana tratan de encontrar un h que minimice la expresión

$$\mathrm{MISE}\left(\hat{f}_h\right) = E\left[\int_{-\infty}^{+\infty}\left(\hat{f}_h(x) - f(x)\right)^2 dx\right] \tag{7.8}$$

La idea de los métodos Plug-In, debida a WOODROOFE (1970), fieles al principio subyacente: *si se tiene una expresión que contiene un parámetro desconocido, se sustituye el parámetro desconocido con una estimación del mismo,* consiste en sustituir en la expresión

$$h_{optimo} = \left[\frac{\int_{-\infty}^{+\infty} K^2(x)\,dx}{N\int_{-\infty}^{+\infty} f''(x)^2 dx\left(\int_{-\infty}^{+\infty} x^2 K(x)\,dx\right)^2}\right]^{1/5} \tag{7.9}$$

el valor $\int_{-\infty}^{+\infty} f''(x)^2 dx$ a través de una muestra piloto. O, en otras palabras, se trata de estimar la curvatura de f usando un estimador de la densidad por núcleos preliminar con algún ancho de banda piloto $g(h)$.

En la figura 7.4 se muestran las representaciones gráficas de los *histogramas y las densidades de probabilidad estimadas por funciones núcleo Gaussiano univariantes* para las variables W_X47 y WR_X47. Como puede observarse la variable tramada por criterios estadísticos y de riesgos, WR_X47, presenta una densidad más "suave" que la variable tramada únicamente por criterios estadísticos, W_X47.

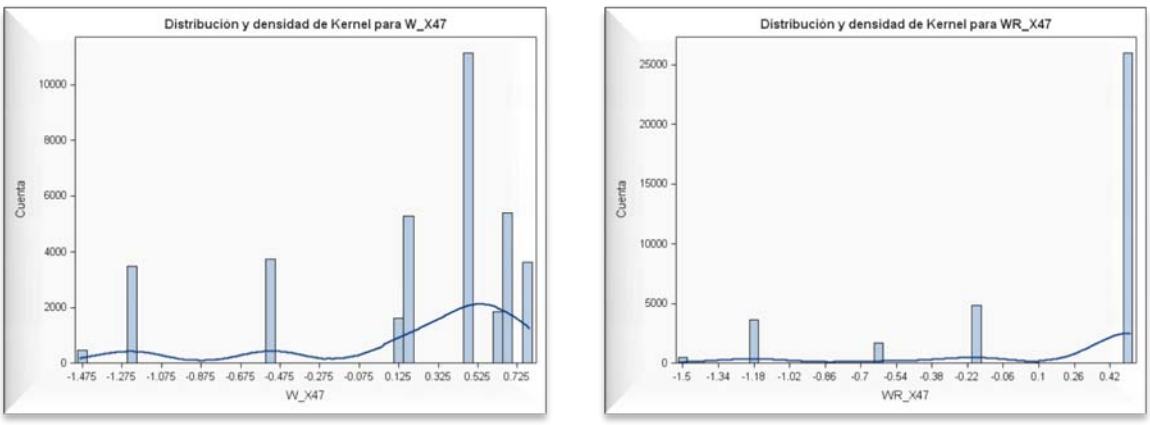


Figura 7.4.- Densidades núcleo para W_X47 y WR_X47.

En la figura 7.5 se representan gráficamente los histogramas y la densidad estimada no paramétricamente por funciones núcleo Gaussiano, para las variables tramadas con criterios de riesgos, WR_X_{12}, WR_X_{46}, WR_X_{49}, WR_X_{54}, WR_X_{55} y WR_X_{56}, WR_X12.

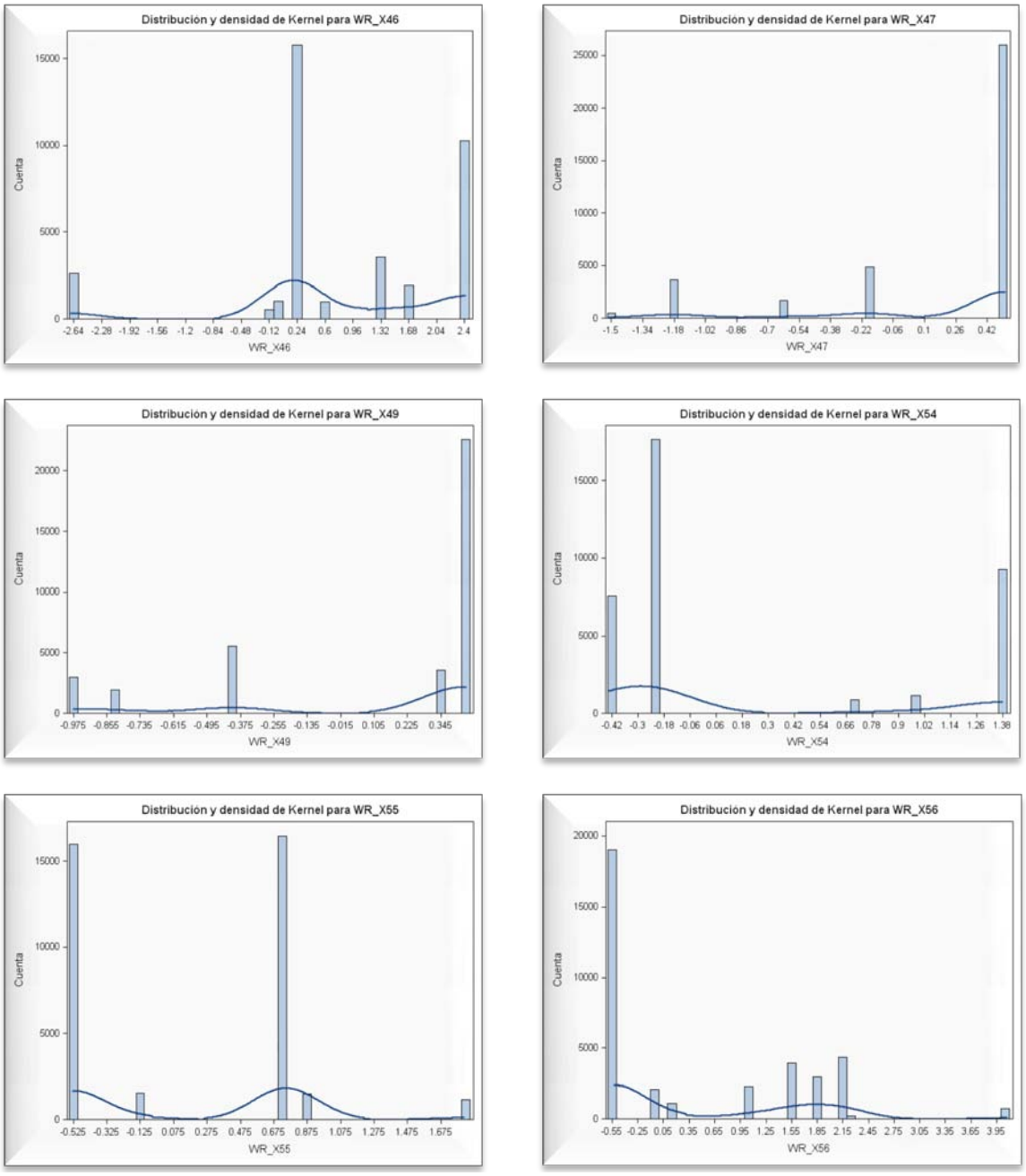


Figura 7.5.- Distribución por densidad núcleo de las variables tramadas con criterios de riesgos.

Tras la imputación de valores faltantes y la eliminación de valores extremos de acuerdo con los criterios expuestos anteriormente concluimos esta fase con 75.001 acreditados a considerar en nuestro análisis, de los que 69.538, el 92,7%, son acreditados con buen

comportamiento frente al default y los 5.463, el 7,3%, son acreditados que presentan un mal comportamiento frente a sus obligaciones de pago y 63 variables de las cuales para siete, X_{12}, X_{46}, X_{47}, X_{49}, X_{54}, X_{55} y X_{56}, se ha sustituido la variable original por una variable tramada con criterios de riesgos, WR_X_{12}, WR_X_{46}, WR_X_{47}, WR_X_{49}, WR_X_{54}, WR_X_{55} y WR_X_{56}.

7.2.4 Correlación, multicolinealidad y poder explicativo.

Una vez que se ha resuelto el problema de los datos faltantes y se han eliminado los valores extremos acometemos el estudio cuidadoso de la relación, en términos de poder explicativo, existente entre las variables, para lo que en primer lugar se estudia la correlación de cada variable con todas las demás así como la de grupos de variables entre sí.

7.2.4.1 Correlación entre las variables explicativas.

A pesar de que el análisis de la matriz de correlación junto con el criterio de Güilford nos permite conocer la intensidad de la relación entre los pares de variables consideradas, no es un análisis adecuado para abordar el estudio de la relación entre grupos de variables. Para analizar esta relación se utiliza una variante del *Análisis de Componentes Oblicuas*, OCA, el *Análisis de Componentes Principales Oblicuas*, OPCA, que nos permite agrupar las variables explicativas en clases de variables caracterizadas por su interrelación.

A diferencia del Análisis de Componentes Principales clásico, PCA, que diagonaliza la matriz de varianzas-covarianzas, o la matriz de correlaciones, por el vector aleatorio subyacente, el OPCA intenta diagonalizar por bloques la matriz de varianzas-covarianzas o la matriz de correlaciones utilizando únicamente permutaciones de las filas. Como resultado, las variables en cada bloque (clúster) retienen tanto mejor la similaridad cuanto más se minimicen las correlaciones entre los bloques. Dado que en credit scoring la interpretación del modelo es muy importante el OPCA puede ser de gran ayuda para mantener la máxima interpretabilidad de las variables.

El OPCA divide un conjunto de variables numéricas en un número cualquiera de clústeres jerárquicos disjuntos. Asociado con cada grupo existe una combinación lineal de las variables del grupo, que podrá ser la primera componente principal o la componente cancroide. La primera componente principal es una media ponderada de las variables que explica de la varianza tanto como sea posible.

OPCA, de modo similar a PCA, si se utiliza la opción de matriz de covarianzas maximiza la suma de la varianza que es explicada por los componentes del clúster, por lo que las

variables con mayor varianza tienen más importancia en el análisis; si se utiliza la matriz de correlación todas las variables tienen la misma importancia. *Este tipo de conglomeración de variables pretende encontrar grupos de variables que sean tan correlacionadas entre sí como sea posible y tan incorrelacionadas como sea posible con las variables de otros clúster.*

Para realizar el análisis utilizamos el procedimiento PROC VARCLUS de SAS® V9.2, en el cual la reasignación de las variables se produce en dos fases:

La primera fase consiste en la clasificación de la componente más cercana, muy similar en principio al algoritmo de clasificación de la más cercana al centroide descrito por ANDERBERG (1973), en cada iteración se calculan los componentes del clúster y cada variable es asignada a la componente con la que tiene la mayor correlación al cuadrado (SAS Institute Inc. (2009), Guía de usuarios SAS / STAT, páginas 1642-1643).

La segunda fase consiste en un algoritmo de búsqueda en la que cada variable a su vez es contrastada para ver si se asigna a un grupo diferente aumentando la cantidad de varianza explicada. Si una variable es reasignada durante la fase de búsqueda, los componentes de los dos grupos involucrados se vuelven a calcular antes de probar la siguiente variable.

Si los clústers están bien separados, el valor del coeficiente de determinación, R^2, de la variable con todas las demás del clúster más cercano debe ser bajo.

Dado que una variable seleccionada de cada grupo debe tener una alta correlación con su propio grupo y una baja correlación con los otros grupos, se puede utilizar la tasa $1-R^{*2}$ para seleccionar este tipo de variables. La fórmula de esta tasa es la siguiente:

$$1-R^{*2}=\frac{1-R^2\ \left(del\ \ propio\ \ cluster\right)}{1-R^2\ \left(del\ \ cluster\ \ más\ \ cercano\right)} \tag{7.10}$$

Una pequeña tasa $1-R^{*2}$ indica una buena conglomeración.

Obsérvese en la tabla 7.4 que las variables $W\__{23}$ y $W\__{32}$ no están asignadas, la razón es que ningún conglomerado cumple las condiciones de división.

Tabla 7.4.- Clústeres jerárquicos disjuntos de las variables numéricas obtenidos por el procedimiento VARCLUS.

			R^2 de cada variable con		
Cluster	*Descripción*	*Variables*	*su propio Cluster*	*el cluster más cercano*	*Tasa 1-R*²*
Cluster 1	***Solvencia del cliente***	WR_X_{12}	0.8974	0.5677	0.2374
		WR_X_{46}	0.7215	0.3262	0.4133
		WR_X_{56}	0.8327	0.5084	0.3402
Cluster 2	***Riesgo del cliente***	X_{15}	0.7418	0.3205	0.3800
		X_{16}	0.9835	0.5061	0.0335
		X_{17}	0.9835	0.5061	0.0335
		X_{18}	0.9769	0.4934	0.0457
		X_{45}	0.9545	0.5119	0.0931
		X_{60}	0.9806	0.5038	0.0390
Cluster 3	***Incidencias en contratos de activo***	X_{23}	0.6005	0.2918	0.5641
		X_{36}	0.8878	0.3558	0.1741
		X_{43}	0.9335	0.3996	0.1107
		X_{53}	0.8134	0.2471	0.2479
		X_{61}	0.8926	0.3641	0.1689
Cluster 4	***Pasivo no vista***	X_4	0.9259	0.2811	0.1030
		X_5	0.8769	0.6019	0.3091
		X_9	0.7136	0.0369	0.2974
		X_{11}	0.9021	0.2797	0.1359
		X_{44}	0.8769	0.6019	0.3091
Cluster 5	***Contratos operativos***	X_{27}	0.9928	0.0271	0.0074
		X_{39}	0.9982	0.0270	0.0018
		X_{41}	0.9982	0.0270	0.0018
Cluster 6	***Antigüedad del cliente en activo y pasivo***	X_{19}	0.9886	0.8399	0.0712
		X_{62}	0.9886	0.8798	0.0947
Cluster 7	***Ingresos del cliente***	X_8	0.7637	0.4545	0.4331
		X_{10}	0.8691	0.3145	0.1909
		X_{34}	0.8308	0.3370	0.2552
		WR_X_{54}	0.7995	0.2983	0.2857
		X_{59}	0.9214	0.3305	0.1174
		X_{63}	0.5673	0.3425	0.6580
Cluster 8	***Antigüedad en préstamos con garantía personal***	X_{38}	0.9253	0.1563	0.0885
		WR_X_{49}	0.9253	0.2702	0.1024

Cluster	Descripción	Variables	R^2 de cada variable con su propio Cluster	R^2 de cada variable con el cluster más cercano	Tasa $1-R^{*2}$
Cluster 9	**Saldo a la vista**	X_7	0.9863	0.5231	0.0287
		X_{13}	0.9863	0.5467	0.0302
Cluster 10	**Antigüedad del cliente en pasivo**	X_{20}	0.9455	0.2668	0.0743
		X_{21}	0.9455	0.2088	0.0688
Cluster 11	**Pasivo vista**	X_2	0.9335	0.5263	0.1404
		X_3	0.9335	0.4864	0.1295
Cluster 12	**Edad y antigüedad del cliente en la Entidad**	X_{37}	0.6745	0.2181	0.4163
		X_{58}	0.6745	0.0528	0.3436
Cluster 13	**Recibos**	X_{42}	0.9965	0.3082	0.0051
		X_{52}	0.9965	0.3115	0.0051
Cluster 14	**Ingresos en el último mes**	X_{14}	1.0000	0.0800	0.0000
Cluster 15	**Pagos último mes**	WR_X_{47}	1.0000	0.0712	0.0000
Cluster 16	**Contrat Incidencia/contrat vista**	X_1	1.0000	0.0068	0.0000
Cluster 17	**Tiempo excedido en límite de crédito.**	X_{32}	1.0000	0.0036	0.0000
Cluster 18	**Tasa Riesgo Caja/Riesgo Total**	X_{57}	1.0000	0.0098	0.0000
Cluster 19	**Actividad en pasivo**	X_{29}	0.8420	0.6942	0.5168
		X_{33}	0.9079	0.6003	0.2303
		X_{35}	0.5880	0.2740	0.5674
		X_{51}	0.8328	0.3762	0.2680
		WR_X_{55}	0.8863	0.5467	0.2507
Cluster 20	**Actividad del cliente ultimo año**	X_{50}	1.0000	0.2706	0.0000
Cluster 21	**Valor medio red/saldo pasivo**	X_{64}	1.0000	0.3410	0.0000
Cluster 22	**Nº contratos vista deudores**	X_{28}	0.8135	0.2166	0.2381
	Descubierto en ahorro vista	X_{48}	0.8135	0.4558	0.3428
Cluster 23	**Saldo pendiente/formalizado**	X_{26}	0.7177	0.2987	0.4025
	Actividad en active último año	X_{30}	0.7177	0.3745	0.4512
Cluster 24	**Tasa de saldo pendiente sobre saldo formalizado en préstam. Hipotec.**	X_{22}	0.9885	0.5074	0.0234
		X_{25}	0.9885	0.4912	0.0227
Clúster 25	**Ingresos medios mensuales en el último año**	X_6	1.0000	0.0990	0.0000
Cluster 26	**Antiguedad en préstamos hipotecarios**	X_{31}	1.0000	0.8696	0.0000
Cluster 27	**Plazo pendiente/Plazo en origen**	X_{24}	1.0000	0.2315	0.0000

Tabla 7.5.- Proporción de variación explicada por los clúster, (Procedimiento VARCLUS).

Número de clúster	*Variación total explicada por los clúster*	*Proporción de variación explicada por clúster*	*Proporción mínima explicada por un clúster*	*Autovalor secundario máximo en un clúster*	*Mínimo R^2 para una variable*	*Máxima tasa $1-R^{*2}$ para una Variable*
1	12.824608	0.2036	0.2036	8.045701	0.0001	
2	20.474250	0.3250	0.3072	4.998203	0.0004	0.9997
3	25.235409	0.4006	0.3581	4.187067	0.0004	1.0004
4	28.909544	0.4589	0.3581	3.299464	0.0004	1.0362
5	31.850240	0.5056	0.4121	2.985017	0.0004	1.0383
6	34.169577	0.5424	0.4499	2.092855	0.0007	4.0427
7	36.049806	0.5722	0.4726	1.838953	0.0007	4.0427
8	37.622923	0.5972	0.4726	1.639325	0.0007	4.0427
9	38.986396	0.6188	0.4941	1.443589	0.0007	4.0427
10	40.424402	0.6417	0.5048	1.371489	0.0007	4.0427
11	41.400763	0.6572	0.5048	1.263135	0.0007	4.0427
12	42.639963	0.6768	0.4938	1.217802	0.0007	4.0427
13	43.809108	0.6954	0.4938	1.192011	0.0007	4.0427
14	44.834736	0.7117	0.4938	1.176972	0.0007	4.0427
15	46.002185	0.7302	0.4938	1.000386	0.0007	1.5559
16	47.001732	0.7461	0.4938	0.998583	0.0050	1.5570
17	47.997714	0.7619	0.4938	0.990721	0.0155	1.5570
18	48.985400	0.7775	0.4938	0.975816	0.0883	1.5570
19	49.715231	0.7891	0.4938	0.957982	0.0883	1.5570
20	50.670586	0.8043	0.4938	0.872460	0.3200	1.5570
21	51.538283	0.8181	0.6143	0.867540	0.3618	1.5570
22	52.337908	0.8308	0.6143	0.850619	0.4219	1.5570
23	53.187860	0.8443	0.6143	0.818105	0.4219	2.8521
24	53.997825	0.8571	0.6143	0.771418	0.4219	2.8521
25	54.769243	0.8694	0.6282	0.743579	0.4219	2.8521
26	55.512822	0.8812	0.6745	0.720922	0.4219	0.7522
27	56.233735	0.8926	0.6745	0.650948	0.5673	0.6580

Variación total explicada = 56.2385 Proporción = 0.8927

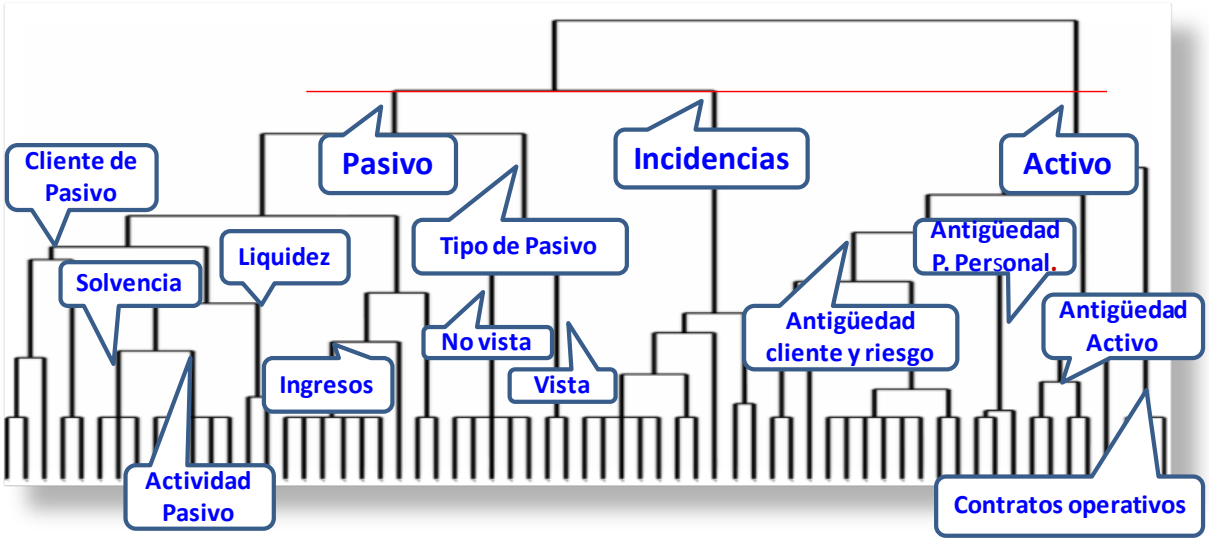


Figura 7.6.- Dendograma de representación de la distancia ultramétrica entre las variables, (Procedimiento VARCLUS).

7.2.4.2 Multicolinealidad entre las variables explicativas.

Un punto importante a tener en cuenta en la selección de las variables consiste en evitar altas correlaciones entre las variables de entrada, es decir, evitar la *colinealidad*, puesto que esto implica que pequeñas variaciones en los datos pueden provocar grandes cambios en los coeficientes del modelo. Las varianzas se inflan demasiado en detrimento de la regresión debido a que algunas variables añaden muy poco o incluso ninguna información nueva e independiente al modelo (BELSLEY et al., 1980).

La colinealidad es un problema cuando se tiene como objetivo no sólo la mera predicción sino además, (caso Basilea), la explicación de la variable respuesta a través de las variables independientes (VAUGHAN y BERRY, 2005), puesto que la colinealidad hace más difícil alcanzar la significación de los parámetros colineales. Pero si esas estimaciones son estadísticamente significativas, son tan fiables como cualquier otra variable en el modelo, y, a veces, aunque no sean significativos, la suma de los coeficientes es probable que lo sea. En este caso, el aumento del tamaño de la muestra es un remedio posible para resolver el problema de colinealidad cuando la meta es la predicción en lugar de la explicación (LEAHY, 2001). Sin embargo, si el objetivo es la explicación, son necesarias otras medidas.

A pesar de que SCHROEDER et al. (1986) afirmaron que no existe una prueba estadística que pueda determinar si la multicolinealidad es o no realmente un problema, eso no impide que se usen algunos métodos empíricos de la detección de la multicolinealidad. La primera opción es usar la *matriz de correlaciones bivariantes*. Este método puede servirnos para

explorar la correlación más sobresaliente por pares de variables, pero carece de sensibilidad a las correlaciones múltiples , puesto que observando la matriz de correlación no se detecta de forma correcta si se tienen variables de entrada combinaciones lineales de otras. Una segunda opción consiste en considerar la regresión de cada variable independiente sobre las demás variables (BERRY y FELDMAN, 1985), este enfoque no suele decir mucho sobre la influencia de los regresores en las varianzas. Es necesario, por tanto, recurrir a otros métodos. Un enfoque mejor que los dos anteriores consiste en utilizar el Factor de Inflación de la Varianza (VIF).

De acuerdo con MENARD (2002), mucha de la información necesaria para pronosticar la multicolinealidad puede obtenerse calculando modelos de regresión por mínimos cuadrados ordinarios, MCO, usando la misma variable dependiente y las mismas independientes que se vayan a usar para el modelo de regresión logística. A causa de que lo que concierne a la colinealidad está en la relación entre las variables independientes, "*la forma funcional para la variable dependiente es irrelevante a la estimación de la colinealidad*". En otras palabras, se puede estudiar la colinealidad sobre el modelo de regresión lineal obtenida por MCO, razón por la cual podemos estudiar la colinealidad entre las variables preseleccionadas considerando el modelo de regresión lineal ajustado por el método MCO

$$DEFAULT = \beta_0 + \sum_{j=1}^{63} \beta_j X_j \tag{7.11}$$

Para detectar todas las combinaciones lineales posibles, se considera un parámetro que se conoce como *Factor de Inflación de la Varianza* (VIF).

Definición 7.2.- Para cada variable X_j, el *Factor de Inflación de la Varianza* de esta variable viene dada por

$$VIF_j = \frac{1}{1-R_j^2} \tag{7.12}$$

donde R_j^2 es el coeficiente de determinación múltiple de la regresión con intercepto de X_j sobre las otras $(p-1)$ variables.

Si el parámetro de ajuste R_j es próximo a 1 el factor VIF_j aumenta y, por tanto, la varianza de los coeficientes también aumenta, en otras palabras, si hay colinealidad, entonces R_j^2 se aproxima a 1, y por, tanto, la cantidad VIF_j es grande. Como no se conoce la ley de

distribución de este coeficiente, BELSLEY et al. (1980) definieron un límite único de forma empírica, estableciendo la regla de que "*un valor de* VIF_j *más grande que 10 revela un problema de colinealidad*"; para KLEINBAUM et al., (1988) el valor aceptable del VIF es que sea menor o igual a 5 (estas reglas han de usarse con ciertas cautela, revisando, además, las cargas de los autovectores correspondientes a los autovalores de menor valor del Ánalisis de componentes principales, ACP).

Definición 7.3.- Para cada variable X_j, se llama *tolerancia de* X_j a la inversa de la inflación de varianza de X_j

$$Tol_j = \frac{1}{VIF_j} = 1 - R_j^2 \tag{7.13}$$

La inflación de varianza y la tolerancia son muy sensibles a la influencia muestral, por lo que utilizaremos la técnica bootstrapping para su estimación.

Definición 7.4.- Siendo $M_1,\ldots,M_B$, las B muestras bootstraps, $VIF_j^b = \frac{1}{1-\left(R_j^b\right)^2}$ y $\frac{1}{VIF_j^b} = 1-\left(R_j^b\right)^2$ la inflación de la varianza y la tolerancia, respectivamente, para la variable X_j y la muestra M_b

a) Se define la *Inflación de la Varianza Bagging para* X_j, como

$$VIF_{jBag} = \frac{1}{B}\sum_{b=1}^{B} VIF_j^b = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{1-\left(R_j^b\right)^2} \tag{7.14}$$

donde R_j^2 es el coeficiente de correlación múltiple de la regresión con intercepto de X_j sobre las otras $(p-1)$ variables en la re-muestra M_b.

b) Se define la *Tolerancia Bagging para* X_j, como

$$TOL_{jBag} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{VIF_j^b} = \frac{1}{B}\sum_{b=1}^{B}\left(1-\left(R_j^b\right)^2\right) \tag{7.15}$$

Con el fin de obtener la inflación de varianza y la tolerancia de las 63 variables disponibles se ajustó el modelo (7.11) por regresión MCO, utilizando la técnica Bootstrapping con 200 muestras, con una macrorrutina de desarrollo propio, sobre la base del procedimiento REG de SAS®, (Regresión por mínimos Cuadrados Ordinarios).

Para cada variable X_j, la *Inflación de Varianza* se calcula como la media aritmética de la Inflación de Varianza de X_j, expresión (7.14), para cada ajuste de regresión MCO del modelo (7.11) realizado sobre cada una de las 200 muestras Bootstrap, obtenidas a partir de la muestra de entrenamiento de forma aleatoria simple con reemplazamiento, Tabla 7.6.

Tabla 7.6.- Variables para las que se verifica $VIF\_Bag \geq 10$, en azul.

Variable	*NumBoot*	*Valort_bag*	*Tolerancia_bag*	*VIF_bag*
X_{17}	200	0.9631	0.000003	380860.49278
X_{16}	200	-0.8329	0.000003	380636.91701
X_{60}	200	5.6034	0.003549	282.647650
X_{11}	200	0.4448	0.007738	129.995186
X_{2}	200	0.2778	0.010610	94.615245
X_{9}	200	0.1673	0.012143	82.854169
X_{52}	200	0.1674	0.013591	73.607212
X_{42}	200	-0.2853	0.013651	73.285731
X_{62}	200	0.5799	0.014339	69.787668
X_{39}	200	-0.3497	0.015936	63.290652
X_{27}	200	0.8354	0.016089	62.698677
X_{31}	200	-3.2613	0.019507	51.304170
X_{4}	200	-4.0149	0.021959	45.938455
X_{5}	200	1.6194	0.025292	39.672290
X_{7}	200	-1.1813	0.027403	36.605200
X_{45}	200	-1.9783	0.032486	30.946853
X_{22}	200	-3.9123	0.033184	30.152774
X_{25}	200	1.1575	0.038374	26.077816
X_{19}	200	3.2054	0.041312	24.264309
X_{3}	200	-0.6937	0.049507	20.276782
X_{20}	200	5.4005	0.057414	17.425885
X_{21}	200	-4.2264	0.059699	16.758622
X_{59}	200	2.1772	0.066385	15.066296
X_{43}	200	158.3409	0.068966	14.504431
X_{33}	200	-9.8526	0.077028	12.983882
X_{29}	200	-19.4262	0.088600	11.288045
X_{10}	200	-1.8732	0.097773	10.230663
X_{38}	200	-1.7706	0.103274	9.690546
X_{36}	200	50.4053	0.112420	8.897809
X_{61}	200	5.6972	0.119781	8.351893
WR_X_{12}	200	2.9199	0.120579	8.294074
WR_X_{49}	200	4.1446	0.126861	7.883431
WR_X_{55}	200	1.1993	0.135942	7.356691
X_{51}	200	5.7529	0.159785	6.258866
WR_X_{54}	200	1.3750	0.161006	6.211698
X_{34}	200	0.9016	0.177693	5.628437
X_{8}	200	-5.7842	0.180032	5.555374
X_{53}	200	-0.4244	0.202230	4.946457
WR_X_{56}	200	0.4862	0.234439	4.265874
X_{15}	200	14.3878	0.248495	4.027437
X_{30}	200	2.9177	0.295651	3.383027
X_{26}	200	-3.1542	0.305825	3.270293
X_{64}	200	-0.1557	0.322034	3.105378
X_{48}	200	-28.3173	0.327675	3.052580
WR_X_{46}	200	-9.1625	0.341260	2.930491
X_{35}	200	3.5894	0.349661	2.860037
X_{24}	200	14.4926	0.360887	2.774750

Variable	*NumBoot*	*Valort_bag*	*Tolerancia_bag*	*VIF_bag*
X_{63}	200	-2.5531	0.421267	2.373878
X_{23}	200	59.3594	0.455546	2.197392
X_{50}	200	-3.3580	0.498343	2.006846
X_{28}	200	7.3670	0.513231	1.948717
X_{14}	200	-1.2898	0.558433	1.791933
X_{37}	200	-0.8443	0.662529	1.509435
WR_X_{47}	200	-1.3173	0.674871	1.481842
X_{6}	200	4.2666	0.705798	1.416923
X_{58}	200	2.6942	0.739193	1.352847
X_{57}	200	-4.9521	0.947095	1.055863
X_{1}	200	-13.3959	0.953292	1.049011
X_{32}	200	5.8817	0.986544	1.013647

Decíamos que el límite superior empírico de BELSLEY et al. (1980), fijado en $VIF_j = 10$, o el de KLEINBAUM et al. (1988), fijado en $VIF_j = 5$ deben usarse con cierta cautela, pues con frecuencia en aplicaciones prácticas se ha detectado colinealidad en variables con VIF_j inferior a tales límites, por lo que es necesario utilizar herramientas estadísticas alternativas que nos permitan determinar el problema de la colinealidad en toda su amplitud. Un potente método para obtener información sobre la colinealidad es el *Análisis en Componentes Principales* (ACP), que consiste en transformar las variables originales para obtener otras variables ortogonales, que son combinaciones lineales de las primeras, llamadas *componentes principales*, MALLO (1985). Con el fin de completar nuestro diagnóstico de colinealidad calculamos los *Índices de Condición y las Proporciones de Varianza*, estadísticos que se obtienen en función de los autovalores y autovectores propios asociados a las componentes principales como veremos a continuación.

La edición de los valores propios asociados a las componentes principales dará información sobre la existencia de colinealidad, por cuanto el autovalor es la cantidad de varianza que dicha dimensión explica en exclusiva. Autovalores próximos a cero indican poca capacidad de explicación, es decir, gran colinealidad con otras dimensiones. De forma general, se calculan los valores propios de la matriz $X^T X$ previamente transformada para tener únicamente el valor 1 sobre los elementos de la diagonal. Dado que trabajaremos sobre un modelo con las variables explicativas centradas y reducidas, $\left(X^T X\right)$ es la matriz de las correlaciones entre las p variables independientes.

Definición 7.5.- *Siendo* $\left(X^T X\right)$ la matriz de las correlaciones entre las p variables explicativas y $\left(\lambda_1,...,\lambda_p\right)$ el vector de autovalores resultantes de la diagonalización de la matriz $\left(X^T X\right)$, ordenados de menor a mayor, se define

a) El *Índice de Condición para el autovalor k-ésimo,* CI_k, según la siguiente expresión

$$CI_k = \sqrt{\frac{\lambda_1}{\lambda_k}} \tag{7.16}$$

b) Para cada valor propio λ_k y cada CI_k, la *proporción de varianza para el coeficiente de la variable j* es el vector, normado para que la suma de sus componentes sea igual a 1,

$$\text{VP}_{jk} = \left(\frac{U_{jk}^2}{\lambda_k}\right), \quad j,k = 1,...,p \tag{7.17}$$

donde U_{jk}^2 es el cuadrado de la componente j-ésima del autovector correspondiente al autovalor λ_k.

Los conceptos de la definición 7.5 se aplican al estudio de la colinealidad en dos pasos:

1) Se editan los valores propios del más grande λ_1 al más pequeño λ_p. <u>*Un valor propio nulo revela la existencia de dependencia lineal entre las columnas de X*</u>, es decir colinealidad. Para cada X_j, $j = 1,...,p$, calculamos el índice de c*ondición,* $CI_k = \sqrt{\frac{\lambda_1}{\lambda_k}}$.

   Un valor propio muy pequeño, al que corresponde un índice de condición muy grande, evidencia un problema de colinealidad. BELSLEY et al. (1980), fijan el límite inferior empírico para el índice de condición en 30.

2) A continuación, para cada valor propio λ_k y cada CI_k se calculan las *proporciones de varianza,* que indican cuales son las variables responsables de la colinealidad revelada por este valor propio. Puesto que la matriz de varianzas-covarianzas de los coeficientes $\hat{\beta}$ de la regresión sobre las variables centradas y reducidas es $Var\left(\hat{\beta}\right) = UVar\left(\hat{c}\right)U^T$, y para un coeficiente j, $Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{n}\sum_{k=1}^{p}\frac{U_{jk}^2}{\lambda_k}$, la proporción de varianza para el coeficiente

k de una variable X_j es el vector $VP_{jk} = \left(\frac{U_{jk}^2}{\lambda_k} \right)$, $j,k = 1,...,p$, normado para que la suma de sus componentes sea igual a 1.

Según BELSLEY et al. (1980), *si las proporciones de varianza de varias variables son mayores que 0.50*, para un *índice de condición* grande, las variables correspondientes tienen un problema de colinealidad entre ellas.

Con el fin de obtener los *índices de condición* y las *proporciones de varianza* de las 63 variables consideradas se ajustó el modelo (7.11) utilizando la técnica bootstrapping con 200 muestras, con una macrorrutina de desarrollo propio, también sobre la base del procedimiento REG de SAS© V9.2.

Definición 7.6.- Siendo $M_1,\ldots,M_B$, las B muestras bootstraps, $CI_k^b = \sqrt{\frac{\lambda_1^b}{\lambda_k^b}}$ el *Índice de Condición para el autovalor k-ésimo de la muestra* M_b, CI_k^b, y $VP_{jk}^b = \left(\frac{\left(U_{jk}^b\right)^2}{\lambda_k^b} \right)$, $j,k = 1,...,p$, la P*roporción de Varianza para el coeficiente de la variable* X_j del autovalor k-ésimo de la muestra M_b,

a) Se define el Í*ndice de Condición* Bagging *para el autovalor k-ésimo* como la media de los Í*ndices de Condición* para *el autovalor k-ésimo de las* B muestras

$$CI_{kBag} = \frac{1}{B}\sum_{b=1}^{B} CI_k^b = \frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{\lambda_1^b}{\lambda_k^b}} \tag{7.18}$$

b) Se define la P*roporción de Varianza Bagging para el coeficiente de la variable* X_j del autovalor k-ésimo, como la media de las P*roporciones de Varianza Bagging para el coeficiente de la variable* X_j del autovalor k-ésimo

$$VP_{jk} = \frac{1}{B}\sum_{b=1}^{B} VP_{jk}^b = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{\left(U_{jk}^b\right)^2}{\lambda_k^b} \right) \tag{7.19}$$

Una vez realizado un análisis de componentes principales sobre la matriz de correlación se tendrán en cuenta los valores propios más pequeños y las correlaciones de las variables con las componentes principales asociadas, las llamadas cargas de las variables. De esta forma,

en combinación con los valores de inflación de la varianza, podremos eliminar de análisis posteriores las variables causantes de la indeseable colinealidad.

Los índices de condición y las proporciones de varianza, para CI > 30, vienen recogidos en la tabla 7.7. Dado que los *autovalores-bag* están ordenados por orden inverso de debilidad, los autovalores-bag que presenta mayor debilidad, autovalor_bag ≈ 0 e índice de condición_bag = 3576440.950 son, y por este orden, los numerados 63, 62, 61 y 60. En la última columna se relacionan las cargas de las variables asociadas a las Componentes Principales correspondientes.

Tabla 7.7.- Índices de Condición de las Componentes Principales y Cargas de las Variables, para índices de condición > 30.

	Componentes Principales		Cargas de las Variables	
Nº C.P.	**Autovalor_bag**	**Índice Condición_bag**	**Variables**	**Carga_bag**
54	0.01329161	31.05203842	X_7	0.77903929
55	0.01098374	34.21115846	X_{27}	0.84004023
56	0.00825866	39.36419995	X_{31} X_{62}	0.74747244 0.84995695
57	0.00685499	43.20227120	X_{42} X_{52}	0.98888452 0.98893126
58	0.00285329	67.07525465	X_{60}	0.99655773
59	0.00000221	2601.872415	X_{16}	0.99992153
60	0.00000000	3576440.950	X_{17} X_{18}	0.99999980 0.99999982
61	0.00000000	3576440.950	X_5 X_{44}	1.00000000 1.00000000
62	0.00000000	3576440.950	X_{39} X_{41}	1.00000000 1.00000000
63	0.00000000	3576440.950	X_2 X_9 X_{11} X_{13}	0.99504341 0.99504341 0.99481097 0.99481097

7.2.4.3 Poder Explicativo de las variables de riesgo de crédito.

Decíamos en 7.2.1 que un aspecto clave consiste en *detectar las variables con mayor potencial de información en relación con el estado de default de los acreditados*, tanto en cantidad como en calidad, es decir, detectar aquellas variables con *alto poder explicativo sobre el estado de default o que presenten suficiente asociación con el mismo.*

El poder explicativo de la variable se medirá a través de la cantidad de información proporcionada por la variable, para lo que se utilizarán los *estadísticos de Gini y Valor de la Información,* IV, definición 7.3, (KULBACK, 1959), y *la asociación entre la variable explicativa y el estado de default* se mide utilizando los adecuados test de asociación con la

variable respuesta, en función de que esta sea de intervalo, (eta, R_cuadrado), o nominal, nuestro caso, (Chi_cuadrado y V de Cramer), descartando aquellas variables para las que la cantidad de información no sea suficiente y/o los test no indiquen asociación. En el primer caso se utilizó el algoritmo Interactive Binning del programa Enterprise Miner de SAS®, y en el segundo el módulo Crosstabs de SPSS, IBM®.

Cuando se trabaja con variables discretizadas por tramos óptimos, ya sea en sentido estadístico y/o desde el enfoque de riesgo de crédito, el valor de información permite hacer una primera selección de variables susceptibles de ser utilizadas en los modelos. Ahora bien, el análisis llevado a cabo es univariante y, por tanto, sólo evalúa la capacidad que tiene cada variable, de forma individual, a la hora de explicar el incumplimiento. Así, una variable muy predictiva de forma individual no tiene por qué ser necesariamente significativa en el modelo multivariante, es decir, no tiene por qué aportar nueva información relevante teniendo en cuenta otras variables ya incluidas en el modelo.

Un alto valor de VI indica un alto poder discriminatorio de una variable específica, pero dado que VI tiene una cota inferior pero no una cota superior del valor absoluto de VI no se puede deducir si el poder discriminante es satisfactorio o no, de hecho el valor de la información se calcula principalmente con el propósito de comparación con otras variables o para la clasificación con las mismas variables y distintas carteras.

El estadístico valor de la información tiene la gran ventaja de ser independiente del orden de las categorías de las variables. Esto es extremadamente importante cuando se analizan datos con distribución de riesgo desconocida.

En la tabla 7.8 se muestra el valor de la información y el índice de Gini de las 63 variables consideradas, cuya interpretación se deduce de 7.4.

Tabla 7.8.- Rango según el Valor de la Información, IV, e Índice de Gini de las 63 variables con WOE calculado por tramos óptimos.

Variable	Valor de Información	Estadístico de Gini
X43	12,606	97,354
X36	9,085	97,324
X61	8,374	66,891
X53	3,264	62,675
X48	2,458	58,449
X46	1,917	64,491
X5	1,337	54,400
X44	1,336	54,372
X28	1,143	35,609
X3	1,106	46,738
X7	1,044	41,140
X2	1,019	42,637
X13	1,001	42,247
X29	0,991	41,935
X8	0,951	39,911
X12	0,891	42,449
X56	0,772	37,183
X63	0,609	30,138
X33	0,576	35,974
X15	0,547	37,304
X59	0,509	25,588
X10	0,462	24,873
X34	0,445	24,849
X47	0,442	32,568
X22	0,420	30,481
X25	0,408	31,610
X64	0,399	24,579
X42	0,399	19,884
X52	0,393	19,884
X60	0,374	31,201
X18	0,371	31,209
X54	0,370	25,066
X55	0,357	30,202
X31	0,319	28,581
X21	0,311	30,160
X38	0,308	27,561
X49	0,282	26,907
X19	0,278	27,589
X16	0,256	27,123
X17	0,256	27,123
X37	0,252	27,144
X45	0,248	26,948
X62	0,246	26,608
X20	0,209	23,582
X26	0,156	19,862
X24	0,120	16,958
X4	0,113	7,850
X58	0,109	18,050
X11	0,107	7,714
X35	0,055	10,515
X30	0,045	6,672
X14	0,035	7,428
X51	0,033	8,864
X27	0,020	5,544
X39	0,019	5,348
X41	0,019	5,348
X50	0,001	0,727
X1	0,000	0,000
X23	0,000	0,000
X32	0,000	0,000

7.2.5 Selección de las variables explicativas y de la muestra poblacional.

7.2.5.1 Exclusión de variables.

En los dos apartados anteriores 7.2.3 y 7.2.4 hemos identificado variables que, desde el punto de vista estadístico, presentan una calidad inferior a la deseable para ser tenidas en cuenta en el análisis posterior. Las causas son diversas:

a) Elevado porcentaje de datos faltantes

b) Concentración de la distribución en pocos valores

c) Valores extraños que ponen de manifiesto errores en la construcción de la variable

d) Existencia de colinealidad

e) No existencia de poder predictivo y/o baja asociación con la variable objetivo

Sin embargo, no se descartan automáticamente las variables desde las técnicas estadísticas afectadas, para evitar que un listado construido con criterios estrictamente estadísticos deje fuera de la modelización algún aspecto relevante desde el punto de vista de negocio. El descarte se hará teniendo en cuenta criterios del riesgo de crédito.

Como consecuencia de a), b), c), d) y e) y conocimiento experto de riesgo de crédito se realiza una primera selección de las variables más adecuadas a nuestros objetivos. Así, por ejemplo, una vez aplicado el proc VARCLUS y obtenidos los distintos grupos de variables y detectada la indeseable colinealidad, algunas variables en cada grupo pueden ser consideradas como variables de pronóstico en base a su valor de la información (IV) o correlación con la variable respuesta, análisis R^2, en función de si las variables se han agrupado o no. Este proceso conlleva implícita una reducción de variables, descartando aquellas que mostrando colinealidad con otras poseen escaso valor explicativo sobre el estado de default o que presenten escasa asociación con el mismo. Además seguiremos criterios razonables respecto del riesgo de crédito, como no eliminar todas las variables de la misma visión de la relación del cliente con la Entidad Financiera y en equilibrio con criterios estadísticos descartar aquellas que aportan información redundante dentro de una visión concreta.

El objetivo de la reducción de variables es mantener un conjunto compacto de variables candidatas a explicar y predecir el comportamiento del acreditado frente al default a través

del modelo, sin que esta reducción provoque pérdida de capacidad potencial de predicción del mismo.

En el apartado 7.2.4.1 hemos analizado la agrupación de variables bajo el principio de que las variables de un grupo han de poseer elevada correlación entre si y baja correlación con las variables de otro grupo. De este modo se han obtenido 27 grupos, algunos con únicamente una variable.

En el caso de variables tramadas, en la construcción de los test de asociación entre la variable estado de default y cada una de las variables explicativas del riesgo nos encontramos con las siguientes situaciones:

1.- *En muchos casos no se pudo calcular el test V de Cramer porque toda la distribución se concentra en uno o unos pocos valores.* Algunas variables categóricas presentan concentraciones muy altas en un único valor. Para las variables continuas, se analizó el porcentaje de clientes para los que la variable toma un número de valores distintos de aquel en el que se concentra la distribución. En función de este porcentaje y del resto de variables disponibles para explicar el mismo concepto, se decidió rechazar o no la variable.

2.- *En los casos en que el test indicó la no existencia de asociación se descartó la variable.*
Cuando se utilizan variables no agrupadas, entre los estadísticos apropiados para evaluar la fuerza predictiva se incluyen *R_Cuadrado* y *Chi_Cuadrado*. Ambos métodos son criterios de bondad de ajuste. La técnica R_cuadrado usa un método de selección paso a paso para rechazar características que no cumplen con aumento incremental de R_cuadrado en los puntos de corte. Un punto de corte típico para R_cuadrado paso a paso es 0,005. Los puntos de corte se deben incrementar si se retienen demasiadas características en el modelo.

La aproximación clásica para evaluar la significación de una variable para su inclusión en el modelo consiste en el bien conocido procedimiento de contrastar la significación de la hipótesis nula, que se basa en la reducción en el error de predicción (beneficio actual menos beneficio previsto) asociado con la variable en cuestión.

Como es habitual en la construcción de modelos de credit scoring en el proceso de eliminación de variables seguiremos los siguientes criterios generales:

- No prescindimos de ningún bloque completo de información, puesto que ninguno resultó no significativo por las características del segmento de clientes a analizar.

- Respecto de la selección de variables desde la perspectiva cliente/producto, dado que estamos analizando el segmento de Préstamos (constituido por los clientes que a fecha de visión tienen un préstamo, independientemente de su tenencia o no de alguna tarjetas de crédito o cuenta de crédito), se seleccionaron en primer lugar las variables del grupo préstamos, (las representativas de su estructura). A continuación se analizaron los grupos cuentas de crédito, tarjetas de crédito, pasivo a la vista y cliente (edad, operatividad, etc.). Pusimos especial cuidado en que no hubiese redundancia en el conjunto final, realizando la elección de variables atendiendo al número de datos faltantes y a su capacidad explicativa.
- En los casos de información en varios periodos (habitualmente para el último mes, trimestre, semestre y año), se ha seleccionado, en general, el periodo más reciente y de entre los restantes se eligió la variable con menor número de datos faltantes y mayor poder explicativo.

Un enfoque muy popular del proceso de eliminación de variables aplicando criterios de riesgos se asienta en la clasificación conceptual de variables:

Visión general del Cliente: El preanálisis de estas variables se centró en la fuerza de la asociación pues los porcentajes de no informados en las variables de esta visión son bajos (inferiores al 5%).

Visión de pasivo: Las variables de esta visión presentan mayores tasas de no informados que las otras, ya que el segmento está formado por clientes particulares que a fecha de observación tienen préstamos de finalidad particular como único producto de activo. Así se han considerado tanto los porcentajes de datos faltantes como la fuerza en la asociación para analizar este conjunto de variables.

Visión de Activo: Los bloques de préstamos con garantía personal, garantía hipotecaria, total préstamos y el de total activo son especialmente relevantes para la modelización del segmento; por otro lado estas variables se solapan entre sí por lo que la selección de variables se centra tanto en la calidad estadística de las variables como en recoger la información más completa del cliente sin que las variables sean redundantes.

Visión de incidencias: Las incidencias en este segmento son originadas por los préstamos y el descubierto en ahorro vista. Así se han seleccionado variables de incidencias centrándose en estos productos y aplicando los criterios habituales promulgados por el Banco Regulador y según criterios de Basilea II, y se ha completado la información con las variables de incidencias totales.

Mediante el anterior enfoque conjuntamente con el análisis de la matriz de correlación, de los resultados del análisis VARCLUS, Tablas 7.4, 7.5 y la figura 7.6, del análisis de la colinealidad, tablas 7.6 y 7.7, del análisis del poder de predicción y asociación de las variables independientes con la variable respuesta, tabla 7.8, y la colaboración de expertos en riesgos de crédito, se llegó a la *eliminación de las siguientes variables*:

$$\begin{aligned} &X_1, X_2, X_4, X_9, X_{10}, X_{11}, WR\_X_{12}, X_{13}, X_{14}, X_{16}, X_{17}, X_{18}, \\ &X_{21}, X_{22}, X_{23}, X_{27}, X_{28}, X_{29}, X_{30}, X_{31}, X_{33}, X_{34}, X_{36}, X_{38}, \\ &X_{39}, X_{41}, X_{43}, X_{44}, X_{50}, X_{51}, X_{52}, WR\_X_{54}, X_{59}, X_{60}, X_{61}, X_{62}. \end{aligned} \tag{7.20}$$

Por tanto, las *variables preseleccionadas* son:

$$\begin{aligned} &X_3, X_5, X_6, X_7, X_8, X_{15}, X_{19}, X_{20}, X_{24}, X_{25}, X_{26}, X_{32}, X_{35}, \\ &X_{37}, X_{42}, X_{45}, WR\_X_{46}, WR\_X_{47}, X_{48}, WR\_X_{49}, X_{53}, \\ &WR\_X_{55}, WR\_X_{56}, X_{57}, X_{58}, X_{63} \text{ y } X_{64}. \end{aligned} \tag{7.21}$$

Tras un preanálisis exhaustivo de las 63 variables iniciales, combinando los criterios de menor entropía, mayores índices de Gini y Valor de la Información, eliminando la colinealidad, a través de la técnica de inflación de la varianza bootstrap y los estadísticos índice de condición y proporción de varianza obtenidos a través de un análisis de componentes principales, utilizando el PCA bootstrap con selección hacia adelante y contando con opiniones subjetivas expertas, llegamos a la preselección de 27 variables que se describen brevemente en la tabla 7.12, observadas sobre 73.207 acreditados particulares que poseen préstamos a particulares, libres de colinealidad y alta calidad de información.

Tabla 7.9- Descripción de las 27 Variables Preseleccionadas.

Variable	Descripción
X_3	Saldo medio del Cliente en productos de pasivo líquido para los últimos 12 meses
X_5	Saldo medio del Cliente en productos de pasivo para los últimos 12 meses
X_6	Importe medio que el Cliente ha percibido de manera recurrente en los últimos 12 meses.
X_7	Saldo mínimo mensual del Cliente en Pasivo a la Vista para el último mes.
X_8	Rango mínimo de saldo (importe máximo – importe mínimo) del Cliente en Pasivo a la Vista para los últimos 12 meses.
X_{15}	Importe total que el Cliente debe pagar a fecha de observación para hacer frente a los requerimientos de pago del cliente con la Caja correspondientes a productos de Activo.
X_{19}	Máxima antigüedad en el último año del cliente en productos de Activo
X_{20}	Máxima antigüedad en el año del cliente en productos de Pasivo.
X_{24}	Mínimo ratio del plazo pendiente a fin de mes frente al plazo original de la operación en préstamos a particulares
X_{25}	Mínimo ratio de saldo pendiente a fin de mes actual frente al importe formalizado en productos de activo con garantía hipotecaria
X_{26}	Mínimo ratio de saldo pendiente a fin de mes actual frente al importe formalizado en préstamos con garantía personal a fecha de visión
X_{32}	Número de meses que el Cliente ha tenido saldo dispuesto por encima del límite en productos de crédito en los últimos 12 meses.
X_{35}	Número de meses, en los últimos 6, que el Cliente ha percibido ingresos recurrentes
X_{37}	Número de meses de antigüedad de relación del Cliente en la Caja
X_{42}	Número de recibos básicos cargados en la cuenta en los dos últimos meses
X_{45}	Importe total que el Cliente tiene en productos de Activo en los últimos 12 meses.
W_X_{46}	WOE_ Saldo mínimo mensual medio del Cliente en Pasivo a la Vista para los últimos 6 meses.
W_X_{47}	WOE_ Cuota máxima de los productos de préstamos para financiación a particulares vigentes a pagar a fecha de visión.
X_{48}	Número de meses, en los últimos 12, que el Cliente ha tenido descubierto (saldo deudor) en Ahorro Vista
W_X_{49}	WOE_ Número de meses de antigüedad desde que el Cliente se dio de alta en el primer préstamo con garantía personal.
X_{53}	Porcentaje de contratos del cliente en incidencia, sobre el total de contratos que hayan estado operativos en algún momento del mes actual.
W_X_{55}	WOE_ Ratio del saldo en pasivo no vista frente al pasivo total en el último mes.
W_X_{56}	WOE_ Ratio de requerimientos de pago totales sobre el saldo en pasivo del Cliente en el mes actual.
X_{57}	Porcentaje que representa el importe en riesgo del cliente en la Caja a fin de mes frente su importe en riesgo total (considerando tanto la Caja como otras Entidades Financieras recogidas en la CIRBE) en el mes de fin del periodo de visión
X_{58}	Edad del Cliente para la fecha de visión.
X_{63}	Número de recibos no básicos cargados en la cuenta en los tres últimos meses.
X_{64}	Ratio de valor medio de red frente al saldo medio total en pasivo en los 3 últimos meses.

7.2.5.2 Selección de la muestra poblacional.

Por lo que respecta al número de acreditados del segmento de particulares con préstamos a fecha 30 de noviembre de 2007, inicialmente constaba de 93.761 clientes distribuidos respecto del incumplimiento como se muestra en la tabla 7.1, pues bien, tras la fase de preanálisis, asignación de valores faltantes y eliminación de extremos, se cuenta con 73.207 acreditados que se distribuyen respecto del default tal como se muestra en la tabla 7.10.

Tabla 7.10.- Distribución del estado de default de los acreditados particulares con préstamos tras el preanálisis.

Incumplimiento	Frecuencia	Porcentaje
Buenos	67.698	92,47%
Malos	5.509	7,53%
Total	73.207	100,00%

La siguiente fase a acometer tras el preanálisis consiste en obtener una muestra representativa de la población objetivo, de forma que los resultados obtenidos en ésta sean aplicables a toda la población. Los objetivos que se pretenden alcanzar al trabajar con muestras en vez de con los conjuntos originales, además del general de trabajar con un número de clientes más reducido, son dos:

1. Maximizar el número de clientes malos en relación con el de buenos, con el fin de poder capturar más fácilmente los distintos perfiles de comportamiento.

2. Eliminar posibles sesgos que puedan haberse producido en el proceso de extracción del conjunto de desarrollo.

Como siempre el procedimiento utilizado para obtener la muestra viene usualmente condicionado por la composición del conjunto de desarrollo.

Nosotros hemos utilizado el *sobre muestreo estratificado*, selección deliberada de acreditados default, para obtener estimaciones razonablemente precisas de las propiedades de estos, construyendo la muestra estratificada por la variable estado de default con el fin de evitar que al construir el modelo las características de los malos queden ocultas por la gran proporción de clientes buenos de la población.

Habitualmente en credit scoring proactivo se considera adecuada una muestra con tamaño entre 10.000 y 20.000 observaciones, con todos los clientes malos y una extracción aleatoria de, al menos, el doble de buenos de la población con el fin de garantizar la existencia de una muestra de modelización suficientemente grande. Construimos nuestra muestra con todos los clientes malos de la población y el doble de clientes buenos, estos últimos seleccionados aleatoriamente.

Creamos una muestra que contiene a todos los clientes malos, 5.509, y un poco más del doble de buenos, 11.283, resultando una muestra "real" de 16.792 acreditados, con una peculiaridad: "*consideramos cada observación de cliente bueno como representativa de una cantidad o peso de clientes buenos en la población total, por lo que utilizando el peso para ponderar los casos de clientes buenos podremos recuperar la tasa real de malos de la población. Construimos la variable peso como cociente del número total de clientes buenos de la población partido por la cantidad de buenos incluidos en la muestra, por lo que valor del peso resultó ser 6. La variable peso indica a* cuántos casos de la población real representa cada caso de la muestra, garantizando de este modo la representatividad de la selección. La muestra pesada consta de 73.210 observaciones con una tasa de malos de aproximadamente el 7,5% y representa el comportamiento de default real de la población.

Con el fin de seleccionar el modelo, medir su capacidad de generalización, validar su poder discriminante y calibrarlo, así como comparar los resultados obtenidos por distintas técnicas, y dado que nos encontramos en una situación rica en datos, siguiendo a HASTIE et al, (2009), obtendremos a partir de nuestro conjunto de datos tres muestras aleatorias simples, (esta partición estará estratificada por la variable indicador de incumplimiento de forma que la proporción de acreditados buenos y malos en cada conjunto sea similar):

(1) *La muestra de entrenamiento,* que usaremos para ajustar el modelo con el criterio de minimizar el error de entrenamiento, a la que asignamos un 40% del total de observaciones de la muestra, (2) la *muestra de validación,* a través de la cual estimaremos el error de predicción esperado con el fin de seleccionar el modelo adecuado, a la que asignamos un 30% de los acreditados de la muestra, y, por último, (3) la *muestra test,* en la que nos apoyaremos para fijar el error de generalización del modelo finalmente elegido, a la que asignamos el 30% restante de los acreditados de la muestra que consideramos.

Tabla 7.11.- Partición Muestral.

	Total		Entrenamiento		Validación		Test	
Default	5.509	(32,80%)	2.755	(32,80%)	1.377	(32,80%)	1.377	(32,80%)
No Default	11.283	(67,20%)	5.642	(67,20%)	2.821	(67,20%)	2.820	(67,20%)
Total	16.792	(100%)	8.397	(100%)	4.198	(100%)	4.197	(100%)

Tabla 7.12.-Partición Muestral Ponderada.

	Total		Entrenamiento		Validación		Test	
Default	5.509	(7,50%)	2.755	(7,50%)	1.377	(7,50%)	1.377	(7,50%)
No Default	67.698	(92,50%)	33.852	92,50%)	16.926	(92,50%)	16.920	(92,50%)
Total	73.207	(100,0%)	36.607	(100,0%)	18.303	(100,0%)	18.297	(100,0%)

7.3 EXPLORACIÓN DE LOS DATOS DE ENTRENAMIENTO.

7.3.1 Introducción.

Una vez construida y divida la muestra de trabajo en las tres submuestras, *de entrenamiento o ajuste*, *de validación* y *test,* la siguiente fase que acometemos consiste en explorar, analizar y preparar los datos de entrenamiento para conseguir modelos de predicción del default lo más generalizables y eficientes posible. Para ello hemos desarrollado una estrategia metodológica basada en los tres principios siguientes:

1.- *Dado que la relación teórica entre el estado de default y las variables explicativas se expresan a través de la función logística, (Teorema de Bayes), tenemos interés fundamentalmente en los modelos logísticos.*

2.- Con el fin de equilibrar la interpretabilidad impuesta por los requerimientos de Basilea II con la necesidad de una adecuada bondad de ajuste, tenemos interés en los modelos logísticos paramétricos y semiparamétricos ya sean completa o parcialmente lineales. Es decir, modelos logísticos con una componente lineal y otra componente no lineal, que podría ser paramétrica, como, por ejemplo, forma cuadrática etc., o no paramétrica, como, por ejemplo, una expansión lineal por splines cúbicos y cualesquiera otras clases de funciones base, siempre que puedan ser razonablemente interpretadas.

3.- *La función de pérdida binomial negativa o pérdida logística es la función de pérdida más idónea para estimar el modelo adecuado para describir, explicar y predecir el estado de*

default en base a apropiadas variables explicativas o funciones de base de las mismas del riesgo de crédito, y cumpliendo los requerimientos de Basilea II.

Habitualmente, la primera tarea en la exploración de la muestra de entrenamiento consiste en hacer una revisión de la distribución de las variables, (mediante medidas descriptivas, tanto de localización como de escala, gráficas de dispersión, histogramas y funciones de densidad de la variables, etc.), para comprobar que no existen diferencias significativas respecto a la distribución en el conjunto de desarrollo completo, de modo que sigan siendo válidas para la muestra de entrenamiento las conclusiones preliminares sobre la capacidad explicativa de las variables independientes.

Puesto que perseguimos modelos logísticos parcialmente lineales, el aspecto más importante de la exploración de los datos de entrenamiento a los efectos de construcción de modelos de credit scoring está relacionado con el efecto lineal o no lineal de las variables explicativas sobre la relación de dependencia de estas con la variable estado de default. Una vía que parece plausible para incorporar variables con relación lineal al modelo, sería utilizar las técnicas de selección automática basadas en la regresión logística lineal, que, por otra parte, es la práctica más usual entre los investigadores y constructores de modelos en distintos campos, en particular en credit scoring.

Queremos resaltar que nuestro enfoque no consiste en seleccionar un conjunto de variables explicativas cuya relación de dependencia con el logit de la probabilidad de default sea significativa para construir un modelo logístico lineal con todas ellas, sino en detectarlas para formar la componente lineal del modelo logístico así como detectar también a las que presenten una clara relación no lineal para formar una componente no lineal y con ambas conformar una expansión de funciones de base para estimar el logit de la probabilidad de default. Es decir, *no estamos interesados solamente, aunque si prioritariamente, en variables con relación lineal, sino también en las no lineales.*

La mayoría de las técnicas de selección automática de variables para el modelo logístico lineal, stepwise, forward, backward, etc., se basan en la significatividad de los coeficientes en el modelo lineal, usando el estadístico razón de verosimilitud o estadísticos deducidos a partir de éste. Todos estos procedimientos son intuitivamente atractivos y operan de una manera común salvando sus diferencias: *se seleccionan conjuntos de variables de forma secuencial.* Según EVERITT y DER (1996), la utilización secuencial de uno de los tres procedimientos es básicamente una cuestión de gusto y afirman: "*en el mejor de los mundos*

el modelo final elegido por cada de estos procedimientos sería el mismo. Pero esto, que sucede a menudo, no está garantizado de ninguna manera". La literatura sobre el proceso de selección de variables para la construcción de un modelo logístico es muy abundante, una visión muy completa puede verse en HOSMER y LEMESHOW (2000).

A pesar de que la selección de variables por métodos automáticos es más o menos sencilla, existe un *primer problema* con el que es habitual encontrarse y que se centra en **la elección de un p-valor crítico α que determine una regla de detención del proceso**. Usualmente la mayoría de los investigadores y de la mayor parte del software informático utilizan como p-valor por defecto $\alpha = 0,05$, sin ningún motivo aparente, solo porque la tradición estadística dice "si no se tienen suficiente opinión subjetiva sobre el asunto utilícese $\alpha = 0,05$". Este valor se ha utilizado con demasiada frecuencia, intencionadamente o no, y ha sido criticado por muchos autores.

La elección de un apropiado nivel de significación ha generado una cantidad importante de artículos, muchas veces contradictorios entre sí. Así, por ejemplo, HOSMER y LEMESHOW (2000) consideran que la elección de $\alpha = 0,05$ es demasiado estricta, ocurriendo a menudo que se excluyen variables importantes en el modelo, por lo que proponen el uso de un rango de 0,15 a 0,25 y hasta 0,30.

Por otro lado, muchos autores entienden que un $\alpha = 0,05$ es absolutamente insuficiente, tanto para la predicción como para la interpretación de los efectos, por lo que es necesario un $\alpha < 0,05$ (véase, por ejemplo, SHTATLAND et al. (2001).

En una tercera vía, LEE y KOVAL (1997) muestran, mediante el uso de *Simulaciones de Monte Carlo*, que el mejor valor para α varía entre 0,05 y 0,40. Al mismo tiempo, STEYERBERG et al. (2000) recomiendan usar $\alpha = 0,05$, de modo que incluyen todas las variables útiles para una mejor predicción.

De la combinación de las anteriores recomendaciones, se puede concluir que el intervalo para α debería ser $0,05 \leq \alpha \leq 0,50$. Pero es evidente que se trata de un intervalo muy amplio, por lo que es necesario *elegir el valor "correcto", tarea ardua y angustiosa que requiere de muchas pruebas, sobre todo porque no hay teoría alguna que respalde cualquier elección de* α. DERKSEN y KESELMAN (1992) describieron esta situación en la forma siguiente: "*si a los datos se les tortura durante el tiempo suficiente, al*

final van a confesar. Los datos siempre confiesan, incluso lo que nosotros queremos si presionamos lo suficiente".

Estas recomendaciones para el uso, por un lado, de p-valores mucho más pequeños que 0,05 y, por otro, mucho mayores que 0,05, parecen contradictorias. *Esta aparente contradicción puede resolverse si se rechaza la idea de la elección de un conjunto de variables orientadas a un modelo único como dogma. En realidad, no existe un "super modelo", que sea bueno para todos los propósitos. La elección del* p*-valor crítico* α *constituye así un aspecto difícil y crucial de la utilización de la regresión logística paso a paso.*

Para resolver el problema de la aparente arbitrariedad en la especificación del valor de α, desde una base más teórica, SHTATLAND et al. (2003) *propusieron una alternativa a los métodos anteriores consistente en un procedimiento totalmente automatizado de selección de variables, basado en la combinación de la regresión logística paso a paso, los criterios de información, Akaike,* (AIC)*, y Schwarz,* (BIC)*, y la selección del mejor subconjunto de variables. El enfoque hereda las mejores características de los tres componentes mencionados anteriormente y nos ayuda a evitar el agobiante proceso de elegir el p valor crítico "derecho" en la regresión automatizada. En palabras de sus propios autores "este enfoque se justifica si el objetivo de la construcción del modelo es la predicción, con un gran número de covariables y poca orientación teórica para seleccionarlas".*

Existe un <u>*segundo problema*</u>*,* aún más crucial que el anterior, de la selección automática de variables basada en la regresión logística paso a paso que no soluciona el procedimiento anterior: ***la inestabilidad y el sesgo que presentan las estimaciones de los coeficientes de regresión, sus errores estándar y los intervalos de confianza cuando se seleccionan con técnicas paso a paso variables entre cuyas interrelaciones existe demasiado "ruido",*** SHTATLAND et al. (2003), LIU y CELA (2007)**.** Por tanto, *los métodos de selección paso a paso sin más no son estadísticamente buenos en presencia de ruido.* Entre los trabajos más recientes donde se pretende dar solución a la problemática anterior destacan LIU y CELA (2007) y HAYDEN et al. (2009*).*

LIU y CELA (2007) obtienen un importante rango de variables explicativas en regresión logística usando las técnicas bootstrap sobre el modelo logit. El rango se establece sobre el porcentaje de modelos en los que cada variable ha dado significativo en el análisis de

máxima verosimilitud usando el estadístico Chi_cuadrado de Wald para los niveles de confianza del 99%, el 95% y el 90%, ($\alpha = 0{,}01$, $\alpha = 0{,}05$ y $\alpha = 0{,}10$).

HAYDEN et al. (2009*) evalúan el rendimiento de modelos logísticos lineales de riesgo de crédito cuyas variables son seleccionadas, por un lado, por medio de métodos de selección paso a paso, Forward y Backward , y, por otro, por Bayesian Model Averaging, BMA. En primer lugar validan internamente, sobre la muestra de entrenamiento, los 3 modelos con el test de* Hosmer-Lemeshow para a continuación validar el rendimiento en la predicción del default de los modelos, "fuera de muestra", utilizando la técnica bootstraping para obtener las medidas de rendimiento, tasa de ajuste, AR, la puntuación de Brier, BS, y la puntuación logarítmica, LS. La comparación se basa en las probabilidades de default predichas.

En esta Tesis Doctoral *proponemos explorar la linealidad de las variables a través de la línea desarrollada por* SHTATLAND et al. (2003), *el uso de la regresión logística lineal paso a paso combinada con los criterios de información de Akaike (AIC) y Schwarz (BIC)*, pero con notables diferencias metodológicas, *orientadas a la obtención de la componente lineal del modelo HLLM* con criterios de riesgo de crédito y el cumplimiento estricto de los requerimientos de Basilea II:

- Selección automática sólo en la fase exploratoria.
- Con el fin de eliminar la inestabilidad de los procedimientos automáticos, *obtener las variables por selección automática por el método backward con bootstraping* (los resultados no difieren de las técnicas forward y stepwise usando remuestras bootstrap).
- Complementación con criterios de riesgo.
- Observancia de los requerimientos de Basilea II.

Nos parece muy interesante el planteamiento de SHTATLAND et al. (2003), pero no como alternativa en la selección de variables para un modelo logístico lineal, con lo que dejaríamos fuera las componentes no lineales relevantes, sino para seleccionar entre varios modelos logísticos lineales candidatos.

Una vez preseleccionadas las variables explicativas por todos los métodos anteriores para conformar la componente lineal, podemos decir que las variables no significativas no se relacionan linealmente con la variable estado de default, pero lógicamente estamos interesados en cualquier tipo de relación lineal o no lineal y, dado que la *regresión logística lineal es evaluada en el marco de la estructura media lineal, los patrones no lineales pueden*

perderse, por lo que sigue siendo cuestionable si la forma funcional de nuestro modelo con todas las variables preseleccionadas se ha especificado correctamente o no. Por tanto, **es fundamental en esta fase de análisis exploratorio de la muestra de entrenamiento detectar la no linealidad**.

Dado que a priori el conocimiento sobre la estructura y forma de la distribución de las variables explicativas es muy escaso, antes de explorar la no linealidad de las variables para las que las técnicas de regresión logística lineal no detectaron linealidad significativa, analizaremos la estructura y forma de estas variables a través del histograma y la densidad de probabilidad estimada por funciones núcleo univariantes, PARZEN (1962). Detectar este hecho es muy importante para configurar la componente no lineal de un modelo puesto que las componentes no paramétricas requieren siempre variación continua, por ejemplo cuando se usan splines y otros suavizadores.

Por otra parte analizaremos la linealidad y no linealidad de las variables en los modelos logísticos lineales, en <u>*primer lugar*</u>, enfrentando el *logaritmo natural de los odds, logit*, a cada variable continua a través de los gráficos exploratorios de dispersión, gráficos utilizados por MÜLLER y HÄRDLE (2003) como una herramienta exploratoria para la construcción de su pionero modelo logístico parcialmente lineal, LPLM, para una situación práctica de credit scoring.

En <u>*segundo lugar*</u> utilizaremos el *test de Box-Tidwell* para aquellas variables con todos sus valores mayores que cero, ya que este test conlleva la inclusión de una expresión de la forma $X * Ln(X)$ en el modelo para cada variable independiente continua; este es un test especializado en detectar los casos más graves de no linealidad, BOX y TIDWELL, (1962).

En <u>*tercer y último lugar*</u> utilizaremos el más moderno y teóricamente fundamentado test de la *suma acumulada de los residuos* de LIN et al. (2002) para detectar la no linealidad de variables continuas en los modelos lineales generalizados.

Los residuos, definidos como la diferencia entre las observaciones y los valores ajustados de la variable respuesta, se han utilizado durante mucho tiempo en los exámenes gráficos y numéricos de la adecuación de los modelos de regresión, la razón es que proporcionan mucha información sobre la pertinencia de estos modelos. Si el modelo es correcto, los residuos se centran en cero y la gráfica de los residuos frente a cualquier coordenada, tal como las variables explicativas o los valores ajustados, no debe exhibir ninguna tendencia sistemática. La aparición de una tendencia sistemática puede indicar una forma funcional

errónea de la variable explicativa o falta de linealidad. Sin embargo, la determinación de cuando una tendencia observada en las gráficas de los residuos refleja una incorrecta especificación del modelo o una variación natural puede ser un gran reto.

Un problema en la interpretación de los gráficos de los residuos brutos es la naturaleza subjetiva de tal interpretación. La naturaleza subjetiva del análisis de los residuos se debe al hecho de que la variabilidad de los residuos individuales es desconocida. Esta dificultad se refleja en un comentario hecho por COOK y WEISBERG (1999, pág. 337): "Pueden encontrarse anomalías en todas las representaciones gráficas de residuos si se mira con suficiente rigor". Sentencias similares pueden encontrarse en ATKINSON (1985) y en MCCULLAGH y NELDER (1989), entre otros.

Como alternativa, LIN et al. (2002) presentaron una técnica de *chequeo de modelos lineales generalizados basado en la suma acumulada de residuos* sobre ciertas coordenadas, método que aplicaremos a las variables continuas cuyos coeficientes resulten no significativos en la regresión logística lineal.

La principal motivación para considerar sumas acumuladas de residuos es que sus variaciones naturales pueden ser verificadas. Específicamente, bajo la hipótesis nula de especificación correcta de la forma funcional del modelo, la distribución de la suma acumulada, cuando se ve como un proceso estocástico, puede ser aproximada por un proceso estocástico Gaussiano de media cero cuyas realizaciones pueden ser generadas por simulaciones de ordenador. Para apreciar si el patrón de residuos observado refleja alguna fluctuación aleatoria, se puede comparar la suma acumulada de residuos a un número suficientemente grande de realizaciones desde el proceso Gaussiano.

De acuerdo con los resultados obtenidos por LIN et. al. (2002), si la forma funcional de una variable explicativa lineal está correctamente especificada, la suma acumulada de los residuos sobre la variable debe estar centrada alrededor de cero y mostrar un patrón sin tendencia sistemática.

7.3.2 Exploración de la linealidad de las variables explicativas del riesgo en relación con el logit de la probabilidad de default.

La facilidad de interpretación de los modelos lineales para explicar el estado de default, además de su gran capacidad de generalización, debido a su estructura, nos conducen a que sea la linealidad del modelo y, por tanto, de las variables explicativas, un tema de fundamental importancia.

Ppor tanto, es razonable que empecemos por ajustar nuestros datos de entrenamiento a un modelo logístico lineal con todas las variables de la preselección. De esta forma, además de profundizar en la selección de variables explicativas del riesgo de crédito podremos explorar la estructura formal, sobre todo su linealidad, que las liga al comportamiento frente al default de los acreditados. Con ese objetivo ajustaremos nuestros datos a un modelo lineal través de la regresión logística, para una vez comprobada la bondad del ajuste, sea por *log verosimilitud, -2logver, criterios de Información, AIC o BIC*, podremos, con la ayuda de los *test chi-cuadrado, chi-cuadrado de Wald y razón de verosimilitud*, decidir sobre la significación de cada variable en la estructura lineal del modelo.

Ajustaremos los datos de entrenamiento utilizando una técnica bootstrap, el *Bagging*, un sobrenombre para la agregación bootstrap media *(Bootstrap aggregating). Bagging* es un método de conjunto, introducido por BREIMAN (1996), para mejorar la estimación inestable de sistemas de clasificación. Breiman motivó el bagging como una técnica de reducción de varianza para un procedimiento de base dado, como, por ejemplo, árboles de decisión o métodos de selección de variables o de ajuste a modelos lineales. Dado que el promedio bootstrap es una media aposteriori, HASTIE et al. (2009), existe una clara relación entre los métodos bootstrap y la aproximación Bayesiana), de hecho, BREIMAN (1996) explotó esta conexión para la creación del método Bagging.

El Bagging ha sido ampliamente aplicado a las técnicas de aprendizaje máquina, pero rara vez ha sido aplicada a herramientas estadísticas tales como la regresión logística, (una interesante y excepcional aplicación puede verse en PERLICH et al. (2003)). No faltan razones para esto, ya que Bagging está diseñado para hacer frente a la variabilidad de un método cuando esta variabilidad es grande, y los modelos de regresión logística lineal son en general mucho más estables que los producidos por las herramientas de aprendizaje máquina, como, por ejemplo, los árboles de regresión o clasificación. Sin embargo, eso no significa que Bagging no se pueda aplicar a métodos como la regresión logística lineal, de hecho su aplicación en esta técnica es muy sencilla y aporta considerables ventajas.

Utilizaremos la *Regresión Logística Lineal Bagged*, LLR_Bag, considerándola bajo el enfoque de que su objetivo será mejorar el rendimiento en la predicción de un aprendizaje estadístico o de una técnica de ajuste de modelos, es decir como un *método de conjunto*.

El principio general de los *métodos de conjunto* es la construcción de una combinación lineal de modelos ajustados, ya sea por el mismo o por distintos métodos de ajuste, en lugar de utilizar un único modelo.

En el escenario de ajuste del modelo lineal $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=S\left(x\right)=\sum_{j=1}^{p}\beta_j X_j$ a la muestra de datos $\boldsymbol{Z}=\left\{z_i\right\}_{i=1}^{N}$, $z_i=\left(x_i,y_i\right)\in\mathbb{R}^{p+1}$, donde, como viene siendo habitual, $x_i=\left(x_{i1},...,x_{ip}\right)$ es una observación del vector aleatorio p-dimensional de las variables explicativas correspondiente al i-ésimo acreditado e y_i es una observación de la variable respuesta unidimensional correspondiente a x_i, siendo

$$\text{logit}\left(\hat{P}^{*b}\left(Y=1/\ X=x\right)\right)=\hat{S}\left(x\right)=\hat{\beta}_0^{*b}+\sum_{j=1}^{p}\hat{\beta}_j^{*b}X_j \tag{7.22}$$

la predicción en el punto x, se define el *estimador bagging de* $S\left(x\right)$ en los términos siguientes:

$$\begin{aligned}\text{logit}\left[\hat{P}^{bag}\left(Y=1/\ X=x\right)\right]&=\frac{1}{B}\sum_{b=1}^{B}\text{logit}\left[\hat{P}^{*b}\left(Y=1/\ X=x\right)\right]\\&=\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_0^{*b}+\sum_{j=1}^{p}\hat{\beta}_j^{*b}X_j\right),\qquad b=1,...,B\\&=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_0^{*b}+\sum_{j=1}^{p}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^{*b}X_j\right)=\hat{\beta}_0^{bag}+\sum_{j=1}^{p}\hat{\beta}_j^{bag}X_j\end{aligned} \tag{7.23}$$

La expresión (7.23) es un estimador de Montecarlo del verdadero estimador bagging, alcanzándolo cuando $B\to\infty$. El número de remuestras, B, gobierna, en la práctica, el ajuste de la aproximación de Montecarlo, por lo demás, no debe ser visto como un parámetro de ajuste para el bagging.

La principal desventaja de bagging, y otros algoritmos de conjunto, es la falta de interpretación. Una combinación lineal de árboles de decisión es mucho más difícil de interpretar que un solo árbol. Del mismo modo aplicando el bagging a la selección de variables - algoritmo de ajuste de modelos lineales (por ejemplo, la selección de las variables utilizando el criterio de AIC en el marco de la estimación por pérdida logística) da pocas pistas sobre si las variables explicativas son realmente importantes.

A pesar de lo anterior, EFRON y TIBSHIRANI (1993) justifican el bootstrap para juzgar la importancia de variables seleccionadas automáticamente al mirar el aspecto relativo de las frecuencias en la ejecución del bootstraping. El estimador bagging es la media de las funciones ajustadas por bootstrap, mientras que las frecuencias de aparición de determinadas variables seleccionadas o interacciones pueden servir para la interpretación.

La agregación promedio de la predicción sobre una colección de muestras bootstrap reduce la varianza, dado que

$$Var\left(\hat{S}_{bag}(x)\right) = \frac{1}{N}\hat{S}(x) \tag{7.24}$$

En la práctica los modelos estarán correlacionados por lo que la reducción de la varianza vendrá dada por un factor menor que $\frac{1}{N}$.

Para especificar un modelo inicial basado sobre la calidad del riesgo de crédito y en conocimiento experto tendremos en cuenta la ventaja que supone comenzar seleccionando las variables apoyándonos en técnicas objetivas basadas en LLR_Bag. Esta ventaja consiste en que *en la estimación de las probabilidades de default estamos sobre todo interesados, por razones teóricamente fundamentadas, en los métodos de ajuste logístico en su sentido más general, donde la función de pérdida para la minimización del error empírico es la log-verosimilitud binomial negativa, y dentro de estos, por exigencias del nuevo acuerdo de capital de Basilea II, en los modelos más sencillos posible, los lineales y, si la información disponible lo requiere, los parcialmente lineales.*

La predisposición a introducir en el modelo términos no lineales razonablemente interpretables viene motivada por que con frecuencia variables importantes para describir y predecir la probabilidad de default presentan realmente una importante relación no lineal con el estado de default, por lo que es aplicable la célebre sentencia de Albert Einstein, "*el modelo deberá ser tan sencillo como sea necesario, pero no más*", y también el dilema de Occam que se expresa en términos relativamente parecidos, pero más contundentes: "*el modelo deberá ser tan complejo como sea necesario, pero no más*" .

Ejemplos de diferentes tipos de relación que pueden detectarse entre la variable dependiente y una variable explicativa pueden verse en la figura 7.7.

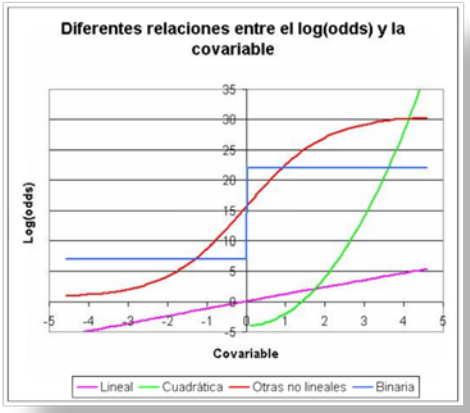


Figura 7.7.- Diferentes relaciones entre el $\log(odds)$ y la variable explicativa.

En *primer lugar* y con el objetivo final de obtener el "mejor modelo" dentro del contexto de credit scoring proactivo, abordamos la tarea más inmediata de configurar la componente lineal que formará parte del modelo semiparamétrico parcialmente lineal hacia el que se orienta nuestro objetivo.

Tal como decíamos en la introducción de esta sección, pretendemos detectar tanto las variables explicativas cuya relación de dependencia con el logit de la probabilidad de default sea significativa como las que presenten una clara relación no lineal. Para ello utilizaremos, en primera instancia, la propuesta de LIU y CELA (2007), basada en EFRON y TIBSHIRANI (1993): la construcción de un rango de las 27 variables explicativas preseleccionadas sobre los niveles de significación $\alpha = 0.01$, $\alpha = 0.05$ y $\alpha = 0.10$, aplicando la regresión logística lineal un gran número de veces sobre submuestras bootstrap de la muestra de entrenamiento, (para lo que usamos una rutina de confección propia basada en los procedimientos Logistic y Surveyselect de SAS® V9.2, para el bootstraping). El rango se establece sobre el porcentaje de modelos en los que cada variable ha dado significativo en el análisis de máxima verosimilitud usando el estadístico Chi_Cuadrado de Wald para los niveles de significación 99%, 95% y 90% .

Como alternativa a la propuesta de LIU y CELA (2007), con el fin de conseguir un método que participe de las cualidades especializadas como método de selección automático de la regresión logística lineal backward a la vez que solucione los problemas de inestabilidad y sesgo que presentan las estimaciones en presencia de ruido, *proponemos utilizar una de las técnicas más usuales de selección automática, la regresión logística lineal backward, pero con la variante de la utilización de muestreo bootstrap*, lo que es una novedad respecto a los

trabajos que hemos visto ahora. El rango se establece exactamente igual que para el método de LIU y CELA (2007).

En *segundo lugar*, tras la orientación que nos proporcionan los métodos anteriores sobre la posible estructura del modelo, sobre todo si la componente lineal no es suficiente para describir el estado de default de los acreditados, usamos otros métodos orientados a detectar y confirmar la no linealidad. Entre los métodos de detección de la no linealidad destacan el más sencillo y conocido que consiste en la utilización de *Gráficos Exploratorios de Dispersión*, otro clásico es el *Test de Box-Tidwell* , BOX Y TIDWELL (1962), que veremos en segundo lugar, y, por último, el *Método de los Residuos Acumulados*, LIN et al. (2002), que consiste en una técnica de *chequeo de modelos lineales generalizados basado en la suma acumulada de residuos*, complementada con el *test del supremo de Kolmogorov*.

7.3.2.1 Exploración de la linealidad usando la regresión logística lineal con muestreo bootstrap, LLR_Bag y BLLR_Bag.

Con el fin de detectar que variables deberán integrar la componente lineal y cuales la componente no lineal del modelo logístico inicial con las 27 variables preseleccionadas, M_{27}, se aplicó la regresión logística lineal bagging, LLR_Bag, sobre 200 submuestras tomadas con reemplazo, del mismo tamaño que la muestra original. El análisis se ha realizado utilizando una macrorutina de confección propia, en lenguaje SAS® sobre la base de los procedimientos Logistic y Surveyselect de SAS® V9.2 , ejecutándose 200 veces y contabilizando la frecuencia de los niveles de significación de 1%, 5% y el 10%, $\Pr > chi2$, del test chi-cuadrado de Wald de los efectos.

Para medir la bondad de ajuste de cada modelo que, en el sentido dado al concepto en la subsección 3.4.1, consiste en medir las discrepancias entre los datos observados y los datos pronosticados, sobre la muestra de entrenamiento, de modo que a menor discrepancia mejor será el ajuste, se utilizaran tres estadísticos basados en $-2\log ver$ (3.91), con formulación $-2\log ver = -2\log L\left(Y, \hat{S}\left(X\right)\right) = -2\sum_{i=1}^{N}\log\left(\hat{P}\left(Y = y_i \,/\, X = x_i\right)\right)$:

1.- El *pseudo coeficiente de determinación de NAGELKERKE* (1991), (3.106), que adopta la forma

$$\tilde{R}^2 = \frac{R^2}{R^2_{máx}} = \frac{1 - \text{antilog}\left(\frac{2}{N}\left(\text{logver}(M) - \text{logver}(M_0)\right)\right)}{1 - \text{antilog}\left(\frac{2}{N}\text{logver}(M_0)\right)} \qquad (7.25)$$

donde M representa el modelo con todas las variables explicativas consideradas y M_0 es el modelo con sólo el término intercepto, llamado *modelo nulo*.

Entonces si el modelo ajustado predice perfectamente los resultados y tiene verosimilitud 1, el pseudo coeficiente de Nagelkerke es $\tilde{R}^2 = 1$, sin embargo, si el modelo completo no mejora al modelo consistente en sólo el coeficiente intercepto es mayor que cero, $\tilde{R}^2 > 0$, por lo que el rango total $[0,1]$ de los mínimos cuadrados ordinarios no está cubierto.

El *pseudo coeficiente de determinación* $\tilde{R}^2$ no puede ser interpretado de forma independiente o comparando distintos conjuntos de datos. Su validez y utilidad se centra en la evaluación de varios modelos de predicción sobre la misma muestra de datos con la misma variable respuesta, nuestro caso. En esta situación, un $\tilde{R}^2$ más alto indica que el modelo predice mejor la respuesta.

Tal como se ha visto en la subsección 3.4.2, el pseudo coeficiente de Nagelkerke, $\tilde{R}^2$, se obtiene como estadístico de contraste del test, (3.93),

$$H_0 : \frac{verosimilitud\,(M)}{verosimilitud\,(M_0)} = 1, \qquad H_1 : \frac{verosimilitud\,(M)}{verosimilitud\,(M_0)} \neq 1$$

es decir, el modelo ajustado con todas las variables explicativas consideradas, M, se ajusta mejor a los datos que el modelo *nulo,* M_0. Este test de bondad de ajuste, por estar basado en $-2\log ver$, conduce con frecuencia a rechazar modelos aceptables a la vez que a aceptar algunos que resultan menos parsimoniosos de lo que debieran, lo que va en contra de la capacidad de generalización y facilidad explicativa requerida al modelo. Cuanto más complejo es el modelo (por ejemplo, más parámetros) mejor es el ajuste y, por tanto, más alto es el valor de la verosimilitud que se obtiene, o, en otros términos, a mayor verosimilitud mayor sobreajuste.

Por tanto, se han propuesto otros criterios para medir la bondad del ajuste, basados en la $-2\log ver$ *penalizada*, el Criterio de Información de AKAIKE (1974), AIC, y el Criterio de Información Bayesiano de SCHWARZ (1978), BIC, también conocido como SBC. En estas medidas de información el término de penalización es el encargado de corregir la complejidad del modelo o, en otros términos, el sobreajuste, por lo que se han ido haciendo cada vez más populares.

2.- Para el modelo de regresión logística, se tiene que el *Criterio de Información de AKAIKE* (1973), AIC, viene dado por, (3.115),

$$AIC = -2\text{logver} + 2p \tag{7.26}$$

donde p es el número de variables explicativas del riesgo de crédito.

AIC es asintótico por lo que se requieren muestras grandes, además, el número máximo de parámetros no puede exceder $2pN$, donde N es el número de observaciones.

3.- Para el modelo de regresión logística, el *Criterio de Información de Bayesiano, (BIC),* SCHWARZ (1978), se expresa en la forma, (3.116),

$$BIC = -2\ \text{logver} + \log(N)\, p \tag{7.27}$$

Asumiendo que $N > e^2 \approx 7.4$, BIC tiende a penalizar modelos complejos, dando preferencia a los modelos simples en la selección.

Para AIC y BIC cuanto menor es su valor, mejor es el ajuste. Dos de las mayores fortalezas de estas dos medidas son:

- Se puede comparar el ajuste de diferentes modelos, incluso cuando los modelos no están anidados. La idea básica es comparar la verosimilitud relativa de los dos modelos en vez de analizar la desviación absoluta de los datos observados de un modelo particular.

- Como medidas de información que son, penalizan la inclusión de variables que no mejoran significativamente el ajuste. En particular, con grandes muestras, las medidas de información pueden conducir a modelos más parsimonias.

Como resultado de la aplicación de la *regresión logística lineal bagging*, LLR_Bag, sobre las 200 remuestras bootstrap se obtuvieron 200 vectores de estimaciones de los parámetros correspondientes a los 200 ajustes del modelo y, a continuación, se construyó la media de los parámetros, los resultados se muestran en la tabla 7.13.

Tabla 7.13 Coeficientes Estimados por Regresión Logística Lineal Bagged y rango de las variables explicativas, ordenadas por el número de veces que cada variable alcanza cada grado de significación, 1%, 5% y el 10% .

Regresión Logística Lineal Bagged (LLR_Bag)							
Variable	*NumBoot*	*Estimador_bag*	*ErrorSTd_bag*	*$\alpha \le 0.01$*	*$\alpha \le 0.05$*	*$\alpha \le 0.10$*	*Otros*
X_{53}	200	0.052374	0.001019	200	200	200	0
X_{48}	200	1.932490	0.071222	200	200	200	0
WR_X_{47}	200	-0.587860	0.076289	200	200	200	0
WR_X_{55}	200	1.103073	0.141772	200	200	200	0
X_{32}	200	-2.316509	0.303730	200	200	200	0
WR_X_{46}	200	-0.276614	0.043766	200	200	200	0
X_{8}	200	-0.001937	0.000279	199	200	200	0
X_{24}	200	-0.013228	0.002050	199	199	199	1
X_{25}	200	0.010966	0.001993	199	199	199	1
X_{57}	200	-0.009843	0.001627	197	198	198	2
X_{64}	200	0.000846	0.000165	196	196	196	4
WR_X_{56}	200	-0.386382	0.084480	194	197	197	3
X_{63}	200	-0.143368	0.024809	193	195	196	4
WR_X_{49}	200	-0.529985	0.108963	191	195	198	2
X_{42}	200	-0.399180	0.090151	188	193	193	7
X_{6}	200	0.000109	0.000027	180	186	188	12
X_{45}	200	0.000001	0.000000	179	187	187	13
X_{58}	200	0.012590	0.003616	163	173	178	22
X_{15}	200	-0.000667	0.000237	150	169	182	18
X_{26}	200	0.004105	0.001693	125	140	154	46
X_{35}	200	-0.043208	0.032386	84	116	129	71
X_{37}	200	-0.000592	0.000415	79	103	123	77
X_{19}	200	-0.000694	0.001420	62	87	103	97
X_{7}	200	0.000022	0.000039	52	77	91	109
X_{20}	200	-0.000530	0.000972	35	61	82	118
X_{5}	200	-0.000010	0.000019	22	57	83	127
X_{3}	200	-0.000026	0.000035	15	47	72	138

En la tabla 7.16 se muestran los estadísticos de bondad de ajuste, $\tilde{R}^2_{bag}$, AIC_{bag} y BIC_{bag}, todos ellos calculados como el promedio de $\tilde{R}^2$, AIC y BIC para los 200 modelos ajustados (por ello a todos estos estadísticos los notamos con el sufijo "bag").

Para obtener el rango de las variables explicativas, ordenadas por el número de veces que cada variable alcanza cada grado de significación, se contabilizaron las frecuencias de los niveles de significación de 1%, 5% y el 10% de todos los análisis para cada variable explicativa que se muestran en la Tabla 7.13. Por ejemplo, de las 200 veces que se ejecutó el análisis, X_{48} es significativa 200 veces al 1%, mientras X15 es significativa 150 veces al nivel del 1%, 169 a niveles menores o iguales al 5% y 182 a niveles inferiores o iguales al 10%. Las variables X_{53}, X_{48}, WR_X_{47}, WR_X_{55}, X_{32}, X_8, X_{24} y X_{25} son las variables más importantes con respecto a la significación lineal.

Hemos coloreado en azul la etiqueta de las variables que no han llegado a mantenerse en el modelo logístico completamente lineal al menos 150 veces para alguno de los 3 niveles de significación, de las 200 que se ha ejecutado la LLR. Las variables X_{26}, X_{35}, X_{37}, X_{19}, X_7, X_{20}, X_5, X_3 son las menos importantes en la significación lineal y la mayoría de ellas se relacionan de forma no lineal con el logit de la probabilidad de default, como veremos en el apartado siguiente.

Como alternativa a la técnica LLR-Bag aplicada por LIU y CELA (2007) *cabe plantearse la posibilidad de que tal vez sea más adecuado considerar un método de selección automática de variables a través de la regresión logística lineal y ejecutarlo de acuerdo con técnicas bootstraping*. En principio este planteamiento parece razonable, por cuanto esta técnica participaría de las cualidades especializadas de los métodos automáticos de selección de variables y el botstraping podría solucionar los problemas de inestabilidad y sesgo que presentan las estimaciones en presencia de ruido. Por esta razón *proponemos utilizar una de las técnicas más usuales de selección automática, la Regresión Logística Lineal Backward, pero con la variante de la utilización de muestreo bootstrap, Regresión Logística Lineal paso a paso Backward Bagged, BLLR_Bag*.

Hemos implementamos el método con una rutina de confección propia basada en los procedimientos LOGISTIC *con método de selección backward* y SURVEYSELECT de SAS® V9.2, para el bootstraping. Los resultados se muestran en la tabla 7.14, (los resultados no difieren de las técnicas forward y stepwise usando remuestras bootstrap).

Se consideraron los niveles de significación $\alpha = 0.01$, $\alpha = 0.05$ y $\alpha = 0.10$ del test Chicuadrado de Wald para detener la eliminación hacia atrás de variables explicativas del modelo en el método Backward. En la tabla 7.14 se muestra una parte de los resultados, los coeficientes estimados para cada variable, más el término intercepto, y el rango de importancia de las variables por su significación lineal. El resto de estadísticos de ajuste, los mismos que fueron obtenidos para LLR_Bag se muestran en la tabla 7.16.

Tabla 7.14 Parámetros Estimados para las variables explicativas por Regresión Logística Lineal Pasos a Paso Backward* Bagged y rango de las variables explicativas, ordenadas por el número de veces que cada variable alcanza cada grado de significación, 1%, 5% y el 10% .

Regresión Logística Lineal Backward Bagged (BLLR_Bag)							
Variable	*NumBoot*	*Estimador_bag*	*ErrorSTd_bag*	$\alpha \leq 0.01$	$\alpha \leq 0.05$	$\alpha \leq 0.10$	*Otros*
X_{53}	200	0.052230	0.001003	200	200	200	0
X_{48}	200	1.924682	0.069846	200	200	200	0
WR_X_{55}	200	1.016381	0.123527	200	200	200	0
X_{8}	200	-0.002053	0.000271	200	200	200	0
WR_X_{47}	200	-0.591856	0.073606	200	200	200	0
WR_X_{46}	200	-0.277641	0.041376	200	200	200	0
X_{24}	200	-0.012516	0.001837	200	200	200	0
X_{32}	200	-2.322839	0.290556	200	200	200	0
X_{64}	200	0.000829	0.000154	200	200	200	0
X_{25}	200	0.010918	0.001935	200	200	200	0
X_{63}	200	-0.141464	0.024233	199	200	200	0
WR_X_{49}	200	-0.555167	0.103834	200	200	200	0
X_{57}	200	-0.009300	0.001625	199	199	199	1
WR_X_{56}	200	-0.373653	0.080857	196	194	194	6
X_{42}	200	-0.409458	0.088586	183	192	193	7
X_{45}	200	0.000001	0.000000	176	183	187	13
X_{6}	200	0.000110	0.000026	169	178	183	17
X_{58}	200	0.012231	0.003476	172	176	182	18
X_{26}	200	0.005220	0.001615	161	172	175	25
X_{15}	200	-0.000703	0.000236	140	164	171	29
X_{35}	200	-0.085223	0.031610	150	162	169	31
X_{37}	200	-0.001041	0.000409	138	152	161	39
X_{20}	200	-0.001990	0.001020	133	149	160	40
X_{19}	200	-0.003052	0.001388	123	155	159	41
X_{7}	200	-0.000023	0.000032	92	124	140	60
X_{5}	200	-0.000028	0.000015	77	103	108	92
X_{3}	200	-0.000056	0.000026	88	101	123	77

* El método backward comienza añadiendo todas las variables en el modelo y va eliminando las menos significativas de forma iterativa.

Por otro lado, con fines comparativos, se realizó una *Regresión Logística Lineal*, LLR, con la obtención de los mismos indicadores de bondad de ajuste. Los resultados del análisis LLR se muestran en la tabla 7.15, e incluyen los estimadores de los coeficientes de las variables en el modelo, junto con sus errores típicos y los valores del estadístico de significación lineal chi-cuadrado de Wald con su p-valor correspondiente. El resto de estadísticos de ajuste, los mismos que los obtenidos para LLR_Bag, se muestran en la tabla 7.16 conjuntamente con los de este modelo.

Tabla 7.15 Parámetros Estimados para las variables explicativas por Regresión Logística Lineal y p-valor chi-cuadrado para el estadístico de significación lineal de Chicuadrado de Wald.

Variable	Estimador	ErrorSTd	Chi2Wald	Chi2Prob
X_{53}	0.052162	0.001007	2682.1238	0.0000
X_{48}	1.915420	0.070112	746.3615	0.0000
X_{32}	-2.301197	0.282839	66.1957	0.0000
WR_X_{55}	1.106086	0.140233	62.2126	0.0000
WR_X_{47}	-0.579903	0.076114	58.0464	0.0000
X_{8}	-0.001891	0.000276	47.0366	0.0000
WR_X_{46}	-0.277227	0.043450	40.7097	0.0000
X_{24}	-0.012982	0.002041	40.4598	0.0000
X_{57}	-0.009641	0.001621	35.3695	0.0000
X_{63}	-0.144044	0.024780	33.7897	0.0000
X_{25}	0.010942	0.001980	30.5345	0.0000
X_{64}	0.000839	0.000163	26.4090	0.0000
WR_X_{49}	-0.540711	0.108597	24.7911	0.0000
WR_X_{56}	-0.381557	0.083385	20.9386	0.0000
X_{42}	-0.391851	0.089026	19.3734	0.0000
X_{6}	0.000108	0.000027	16.4180	0.0001
X_{58}	0.012577	0.003596	12.2347	0.0005
X_{45}	0.000001	0.000000	12.1200	0.0005
X_{15}	-0.000659	0.000237	7.7177	0.0055
X_{26}	0.003963	0.001687	5.5163	0.0188
X_{35}	-0.045755	0.032003	2.0441	0.1528
X_{37}	-0.000554	0.000411	1.8167	0.1777
X_{3}	-0.000027	0.000033	0.6718	0.4124
X_{7}	0.000024	0.000037	0.4302	0.5119
X_{20}	-0.000445	0.000945	0.2219	0.6376
X_{19}	-0.000575	0.001414	0.1655	0.6841
X_{5}	-0.000006	0.000018	0.1281	0.7204

Para la regresión logística lineal el p-valor chi-cuadrado de Wald es menor que $\alpha = 0.05$, para las variables X_{15} y X_{26}, por lo que al nivel de significación del 5% estas variables resultan linealmente significativas, no así , X_{35} que como puede apreciarse en la tabla 7.15 no resulta linealmente significativa ni siquiera al nivel de significación del 10%.

Tabla 7.16.- Estadísticos de "bondad de ajuste" para los modelos LLR_Bag , BLLR_Bag y LLR.

	Log Verosimilitud	Desvianza	GL.	$\tilde{R}^2_{bag}$	AIC_{bag}	BIC_{bag}
LLR_Bag	-2261.433	4522.864	27	0.8136	4576.862	4815.088
BLLR_Bag	-2267.646	4535.292	27	0.8129	4589.292	4838.376
LLR	-2272.369	4544.739	27	0.8127	4589.739	4838.962

Como ya apuntamos, para LLR_Bag y BLLR_Bag los estadísticos de bondad de ajuste son los promedios de los correspondientes estadísticos obtenidos para todas las remuestras bootstrap.

Lo primero a destacar observando la tabla de estadísticos de bondad de ajuste, Tabla 7.16, son los **altos valores que alcanza el coeficiente de determinación de Nagelkerke** $\tilde{R}^2 \geq 0.80$, tanto en el ajuste de los modelos LLR_Bag y BLLR_Bag como en el ajuste del modelo LLR, **lo que apunta a modelos sobre ajustados,** como consecuencia de utilizar demasiadas variables. Podríamos proceder, por tanto, a un proceso de eliminación de algunas de las variables con un método combinado de conocimiento experto y herramientas estadísticas. Pero nos parece más conveniente realizar esta labor una vez se haya especificado el papel lineal o no lineal que las variables juegan en la estructura formal del modelo. La razón es que siempre será preferible eliminar una variable con efecto no lineal, por su complejidad en la interpretación, siempre que razones fundadas no aconsejen su permanencia en el modelo.

Podemos concluir que *BLLR_Bag y LLR dan resultados muy parecidos, pero es evidente que tanto a efectos predictivos, criterios de información AIC y BIC más bajos, como a efectos de clasificación, coeficiente de determinación de Nagelkerke más alto, es "ligeramente" mejor LLR_Bag que BLLR_Bag y LLR, pero las diferencias son inapreciables.*

Además, *queda claro que el modelo no está correctamente especificado en la formulación logística lineal,* $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}$, para las 27 variables preseleccionadas.

En los dos métodos LLR_Bag y BLLR_Bag las variables X7, X3, X5, X20, X19, X35 y X37 han sido seleccionadas, a nivel de significación $\alpha \leq 0.10$, menos del 75% de las veces y las variables X15 y X26, han sido seleccionadas un número de veces un poco inferior al 75% de las veces en LLR_Bag y un número de veces superior al 75% en BLLR_Bag, tablas 7.13 y 7.14. Esta diferencia sugiere que el método LLR_Bag es más propenso a rechazar la significatividad de los coeficientes de las variables explicativas que el método BLLR_Bag.

La propuesta de LIU y CELA (2007), LL_Bag, y la nuestra, BLLR_Bag son más costosas computacionalmente, por cuanto para cada remuestra bootstrap hemos de considerar la complejidad del proceso iterativo de exclusión de variables frente al proceso más simple de una regresión lineal sobre tal remuestra, pero, como podemos comprobar en el apartado siguiente, puede ser una buena herramienta para detectar la linealidad de las variables

explicativas, en presencia de ruido. En todo caso, si bien es muy útil utilizar estos métodos, en los sistemas de calificación de acreditados es fundamental hacerlo conjuntamente con criterios de riesgos.

7.3.2.2 Exploración de la estructura de la distribución de las variables con linealidad no significativa.

Antes de explorar la no linealidad de las variables para las que las técnicas LLR_Bag, BLLR_Bag y LLR no detectaron linealidad significativa, X_3, X_5, X_7, X_{19}, X_{20}, X_{26}, X_{35} y X_{37}, analizamos en detalle si presentan estructura continua, o bien, discreta o casi-discreta. Detectar este hecho es muy importante para configurar la componente no lineal de un modelo, puesto que las componentes no paramétricas requieren siempre variación continua, por ejemplo cuando se usan splines y otros suavizadores. Para ello observamos el histograma y la densidad de probabilidad estimada por funciones núcleo univariantes, usando el núcleo Gaussiano y una amplitud de ventana h obtenida por la regla de "a dedo" SJPI, Sheater-Jones Plug-In, que obtenemos usando de nuevo el procedimiento KDE de SAS V9.2, y cuya salida parcial se muestra en las figuras 7.8, para X_{35}, y en 7.9, para el resto.

Como puede observarse en la figura 7.9, la mayoría de las distribuciones de densidad de las variables que consideramos son muy sesgadas a la izquierda, hecho que se puede apreciar para el conjunto de las 27 variables preseleccionadas.

Para todas las variables se observa que un valor cubre la mayoría de los casos, además las variables X_{19}, X_{20}, X_{35} y X_{37} presentan estructura discreta o casi-discreta, por ejemplo X_{35}:

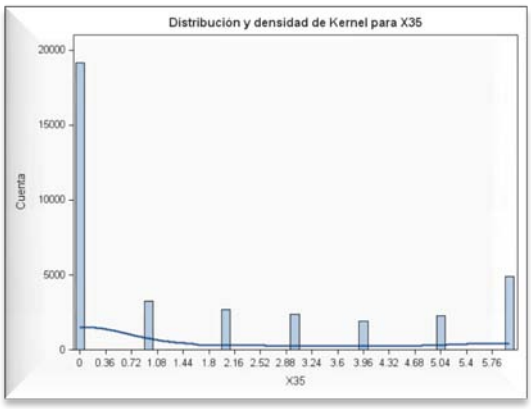


Figura 7.8.- Representación de la distribuciones de la variable X_{35}, utilizando la técnica de *estimación de la densidad por núcleos* (Procedimiento KDE de SAS V9.2).

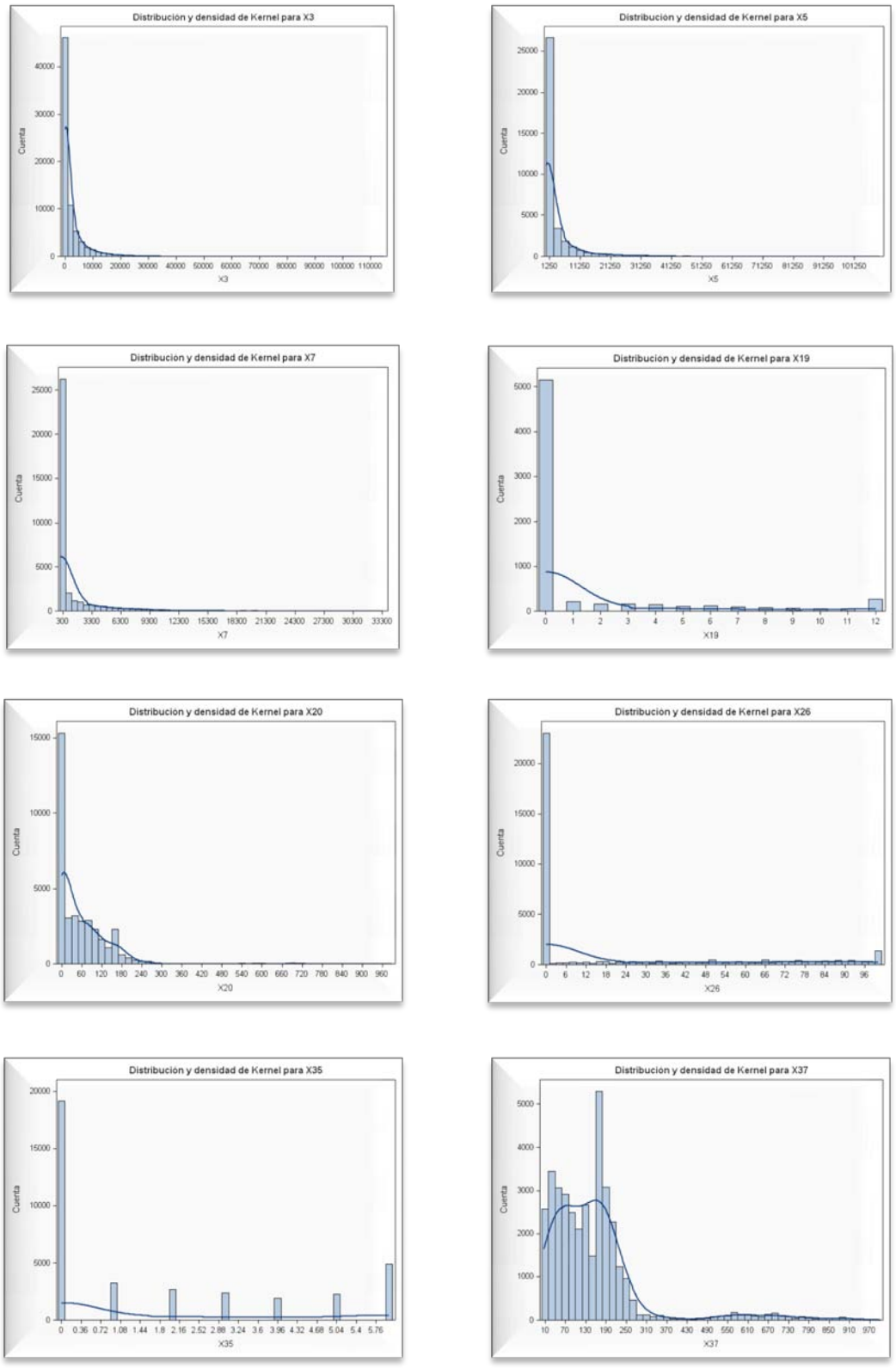


Figura 7.9.- Representación de las distribuciones de e las variables X_3, X_5, X_7, X_{19}, X_{20}, X_{26}, X_{35} y X_{37}, utilizando la técnica de *estimación de la densidad por núcleos Gaussianos* (Procedimiento KDE de SAS V9.2).

7.3.2.3 Exploración de la no linealidad por *diagramas de dispersión del logit de la probabilidad de default frente a las variables explicativas.*

A continuación exploramos la no linealidad de las variables en modelos logísticos, enfrentando el logaritmo natural de los odds, es decir, el logit de la probabilidad de default, $\ln\left(\frac{P(Y=1/X=x)}{1-P(Y=1/X=x)}\right)$, a cada variable a través de los gráficos exploratorios de dispersión, gráficos utilizados por MÜLLER y HÄRDLE (2003) como una herramienta exploratoria para la construcción de su pionero *modelo logístico parcialmente lineal*, LPLM, para una situación práctica de credit scoring. Exploramos la linealidad de las variables explicativas X_{19}, X_{26}, X_{35} y X_{37}:

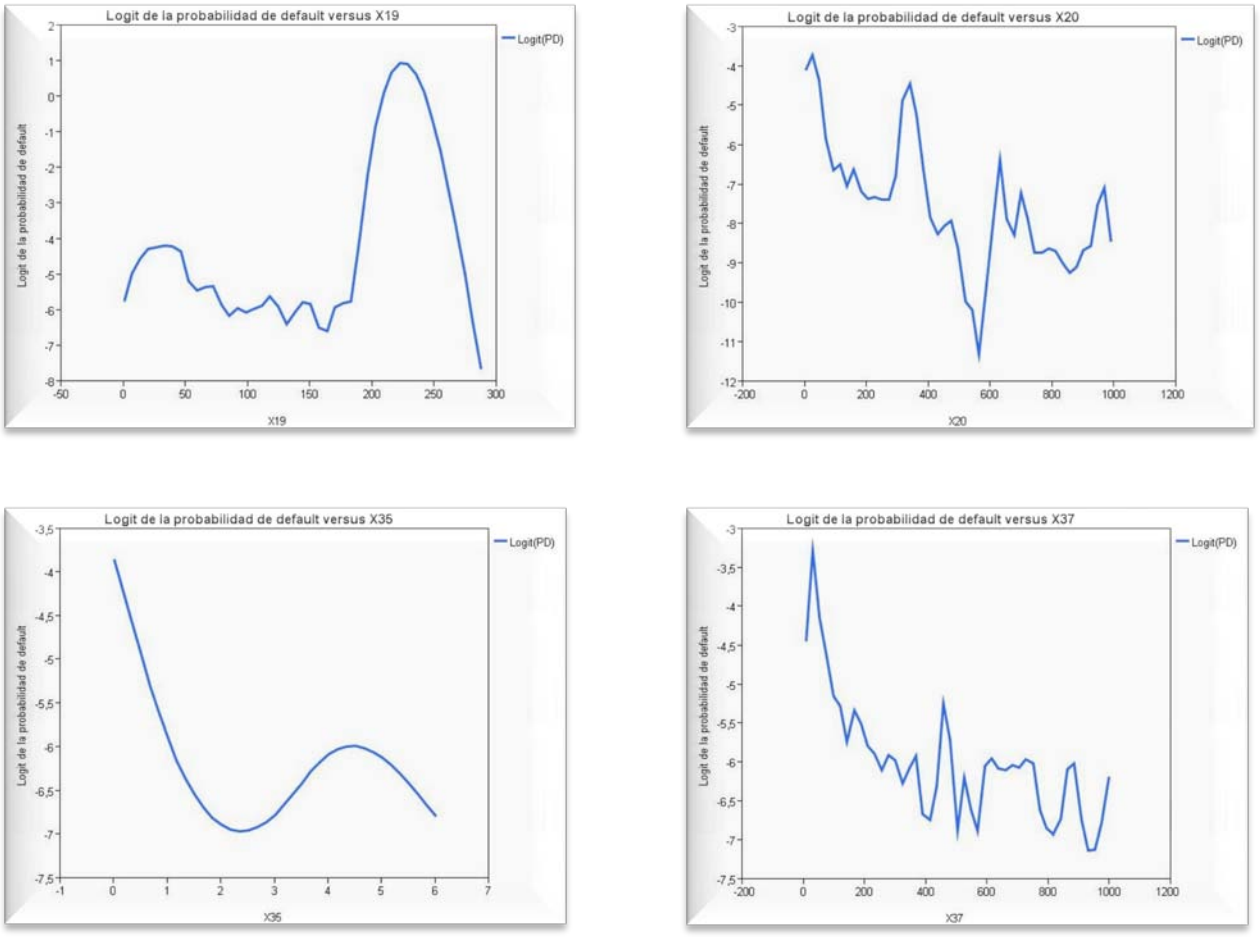


Figura 7.10.- Representación gráfica de las variables explicativas X_{19}, X_{20}, X_{35} y X_{37} frente al logit de la probabilidad de default.

Como puede observarse en la figura 7.10 el logit de la probabilidad de default no se relaciona linealmente con ninguna de las cuatro variables consideradas.

Sin embargo, transformaciones adecuadas, funciones de base, de estas variables pueden relacionarse linealmente con el logit de la probabilidad de default; por ejemplo, WT_X_{37}, función de base de X_{37}, construida por la asignación de pesos de la evidencia a un tramado óptimo por partición recursiva. Como puede observarse en la figura 7.11 la transformación WT_X_{37} rebaja fuertemente la no linealidad, aparte de ser fácilmente interpretable.

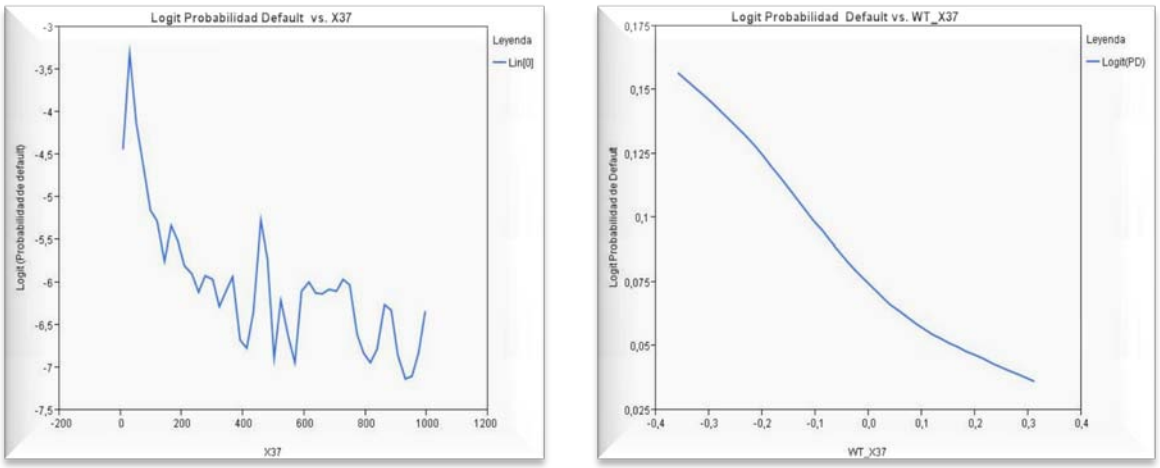


Figura 7.11.- Representación gráfica de las variables explicativas X_{37} y WT_X_{37} frente al logit de la probabilidad de default.

El diagrama de dispersión del logit de la probabilidad de default frente a las variables explicativas es un instrumento de exploración de la linealidad en modelos logísticos lineales adecuado para todo tipo de variables, ya sean numéricas, (continuas o discretas), o cualitativas. Lo consideramos aquí, por cuanto es una útil alternativa al método de los residuos acumulados para la exploración de variables numéricas discontinuas no significativas en modelos lineales que está limitado a variables continuas, por lo que no es aplicable para X_{19}, X_{20}, X_{35} y X_{37}.

7.3.2.4 Test de Box-Tidwell para la Exploración de la No Linealidad.

El clásico test de Box-Tidwell conlleva la inclusión en el modelo de términos de interacción que son productos de cada variable independiente X_j y su logaritmo natural, $X_j ln\left(X_j\right)$. Una vez introducidos estos términos y ejecutada la regresión logística lineal, una

estimación estadísticamente significativa para el término añadido indica que la relación entre X_j y $\text{logit}\left[P(Y=1/X)\right]$ no es lineal.

El test no sólo detecta la no linealidad de la variable X_j, sino que proporciona una posible expansión por funciones de base de la variable, $h(X_j)=X_j ln\left(X_j\right)$.

El principal problema del test de Box-Tidwell es que sólo permite identificar los casos más graves de no linealidad, puesto que no es sensible a pequeñas desviaciones de la linealidad. Y, además, en el caso de que se observe en una variable X_j muchas veces el valor cero, el uso de $ln\left(X_j\right)$ es impracticable para esa variable, por lo que es poco práctico para la mayoría de las variables usuales en un credit scoring proactivo, donde el cero abunda. Sin embargo, lo consideramos porque allí donde pueda ser aplicado constituye un buen refuerzo a otros métodos de detección de la no linealidad.

Para detectar la no linealidad de la variable X_{37}, sobre la que no se observa el valor cero, hemos ejecutado la regresión logística en SPSS sobre un modelo compuesto de todas las variables no significativas en BLLR_Bag apartado 7.3.2.1, tabla 7.14, al que hemos incorporado la transformación de Box-Tidwell, $X_{37}*ln\left(X_{37}\right)$. Obsérvese en la tabla 7.17 que la interacción $X_{37}*ln\left(X_{37}\right)$ es significativa, por lo que se puede concluir que *la relación entre el logit y* X_{37} *no es lineal*. Este resultado confirma los obtenidos para X_{37} con las técnicas LLR, LLR_Bag y BLLR _Bag.

Tabla 7.17.- Test Global del Modelo LLR y de los Coeficientes de las Variables $X37$ y $X37*ln\left(X37\right)$.

Log Verosimilitud	Desviance	GL.	R^2	$\tilde{R}^2$	AIC	BIC
-2240.8024	4481,6048	29	0.339	0.8190	4539,605	4496.336

Variable	Estimador	ErrorSTd	Chi2Wald	Chi2Prob
.....	.....	.....	.....	.....
X_{37}	-0,0315012	0,0042945	53,81	<,0001*
LN(X_{37})	3,39353392	0,4593131	54,59	<,0001*
(X_{37}-154,439)*(Ln(X_{37})-4,647)	0,0113484	0,0016115	49,59	<,0001*
.....	.....	.....	.....	.....

7.3.2.5 Detección de la No Linealidad por el Método de los Residuos Acumulados.

El siguiente método para detectar la no linealidad, debido a LIN et al. (2002), consiste en una técnica de chequeo de modelos lineales generalizados basado en la suma acumulada de residuos.

Los residuos se definen como la diferencia entre las observaciones y los valores ajustados de la variable respuesta. Los análisis de los residuos convencionales basados en la representación gráfica de los residuos brutos o con el suavizado de sus curvas son muy subjetivos, mientras que la mayoría de las pruebas numéricas de bondad de ajuste proporcionan poca información sobre la naturaleza de los errores que son consecuencia de la mala especificación del modelo.

Como ilustración, consideremos un modelo que contiene todas las variables con linealidad significativa inferior o igual al 5% en LLR o que hayan sido seleccionadas más de 150 veces en LLR_Bag y en LLR_Backward_Bag al nivel del 5%, tablas 7.13 y 7.14 respectivamente, añadiéndole la variable X_{20} que no tiene linealidad significativa y que, por tanto, queremos analizar:

$$\begin{aligned} &\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X} \\ &\text{donde} \\ &Y\ =\ \text{Estado de Default} \\ &\boldsymbol{X}=(\boldsymbol{X_{20}},\ X_{53},\ X_{48},\ WR\_X_{47},WR\_X_{55},\ X_{32},\ WR\_X_{46}, \\ &\qquad X_8,\ X_{24},\ X_{25},\ X_{57},\ X_{64},\ WR\_X_{56},\ X_{63},\ WR\_X_{49}, \\ &\qquad X_{42},\ X_6,X_{45},\ X_{58},\ X_{15}) \end{aligned} \tag{7.28}$$

El modelo lineal general (7.28) se ajustó a los datos de entrenamiento a través del procedimiento GENLIN del programa PASW Statistics IBM®, con *función de enlace logit* y *distribución binomial.* Los residuos de desvianza tipificados representados frente a X_{20}, se muestran en la Figura 7.12.

Los residuos brutos indicados por los puntos en la Figura 7.12, no parecen mostrar ningún patrón en particular. Un problema en la interpretación de estos gráficos es la naturaleza subjetiva de tal interpretación.

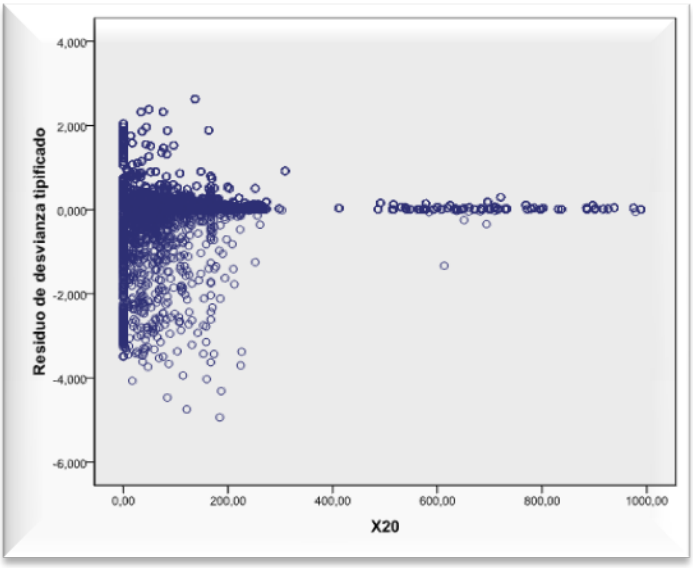


Figura 7.12.- Representación de los residuos de desvianza tipificados frente a X_{20}.

Como alternativa, LIN et al. (2002) presentaron la técnica de *chequeo de modelos lineales generalizados basados en la suma acumulada de residuos* sobre ciertas coordenadas. La técnica desarrollada por estos autores es una técnica de control objetivo del modelo mediante las sumas acumuladas de residuos en las variables explicativas. El método, de aplicación general, puede ser muy útil en los modelos lineales generalizados, las distribuciones de estos procesos estocásticos bajo el modelo asumido pueden aproximarse por la distribución de ciertos procesos de Gauss de media cero, cuyas realizaciones pueden ser fácilmente generadas por simulación con ordenador. Cada proceso observado puede entonces se comparado, tanto gráfica como numéricamente, con un número de realizaciones del proceso de Gauss. Estas comparaciones nos permiten evaluar de manera objetiva si una tendencia observada en una gráfica residual refleja una especificación incorrecta del modelo o una variación natural. Las técnicas propuestas por LIN et al. (2002), son particularmente útiles en el chequeo de la forma funcional de una variable explicativa y de la función de enlace.

Expondremos a continuación unas breves pinceladas sobre la base teórica del método para modelos lineales generalizados.

Si consideramos un modelo lineal generalizado con respuesta binomial y función de enlace logit, tal como $\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}$, donde Y es la variable respuesta,

$\boldsymbol{X} = \left(X_1,...,X_p\right)^T$ es el vector de covariables, β_0 el término intercepto y $\boldsymbol{\beta} = \left(\beta_1,...,\beta_p\right)^T$ el vector de coeficientes de regresión, se definen *los residuos para el modelo* como

$$e_i = y_i - \Lambda\left(\beta_0 + \hat{\boldsymbol{\beta}}^T x_i\right) = y_i - \frac{1}{1+e^{-\left(\beta_0+\hat{\boldsymbol{\beta}}^T x_i\right)}} \tag{7.29}$$

Para chequear la forma funcional para la j-ésima componente del vector de variables explicativas, se considera la suma acumulada de los e_i sobre los x_{ij}, es decir,

$$W_j(x) = \frac{1}{N^{1/2}} \sum_{i=1}^{N} I_{\left[x_{ij} \le x\right]} e_i \tag{7.30}$$

donde $x \in \mathbb{R}$ e $I_{[\bullet]}$ es la función indicador. Nótese que $W_j(\bullet)$ es una función escalera con posibles saltos en los distintos valores de x_{ij}. La curva azul en la figura 7.13 es un ejemplo de (7.30). Puede verse $w_j(\bullet)$ como un proceso estocástico indexado por x. LIN et al. (2002), demostraron que si la forma funcional de una variable explicativa lineal está correctamente especificada, la suma acumulada de los residuos sobre la variable, mostrada en la expresión (7.30), debe estar centrada alrededor de cero y mostrar un patrón sin tendencia sistemática.

Nuestro interés está en el examen gráfico de la asunción del modelo específico, tal como la forma funcional de cada variable explicativa. La distribución nula de $W_j(x)$ puede ser aproximada simulando un proceso Gaussiano de media cero, $\hat{W}_j(x)$. Para fijar como es de excepcional el proceso observado $w_j(\bullet)$ bajo la hipótesis nula H_o de que *el modelo está bien especificado*, se puede representar gráficamente el proceso $w_j(\bullet)$ con unas pocas realizaciones del proceso $\hat{w}_j(\bullet)$, tal como se muestra en las figuras 7.13 o 7.14.

De acuerdo con los trabajos de LIN et al. (2002), para fijar la linealidad del modelo (7.28) y más generalmente la función de enlace logit

$$\log it\left(P\left(Y=1/\boldsymbol{X}\right)\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X} \tag{7.31}$$

se considera la suma móvil de los residuos sobre los valores ajustados siguiente:

$$W_{\Lambda}(x,b) = \frac{1}{N^{1/2}} \sum_{i=1}^{N} I_{\left[x-b \le \beta_0 + \hat{\boldsymbol{\beta}}^T x_i \le x\right]} e_i \tag{7.32}$$

La distribución nula de $W_{\Lambda}(x,b)$ puede aproximarse por la distribución condicional $\hat{W}_{\Lambda}(x,b)$ que se obtiene de $\hat{W}_j(x)$ reemplazando $(x-b \le x_{ij} \le x)$ con $\left(x-b \le \beta_0 + \hat{\boldsymbol{\beta}}^T x_i \le x\right)$. Como en el caso de W_j, se puede representar gráficamente el proceso observado $W_{\Lambda}(\bullet,b)$ con una pocas realizaciones de $\hat{W}_{\Lambda}(\bullet,b)$ y suplementar la visualización gráfica con un estimado p-valor para el llamado *test supremo,* también llamado test de la función de enlace logística,

$$S_{\Lambda}(b) \equiv \sup_x \left|W_{\Lambda}(x,b)\right| \tag{7.33}$$

Las anomalías en W_{Λ} pueden reflejar una especificación errónea de la función de enlace, de la forma funcional de la variable respuesta o de la linealidad de la variable explicativa.

SU y WEI (1991), concentran su atención en el *test ómnibus* $S_{\Lambda}(b) \equiv \sup_x \left|W_{\Lambda}(x,b)\right|$, también conocido como *test supremo de Kolmogorov*. Ellos demuestran que este test es consistente frente a la alternativa general de que no existe un vector $(\beta_0, \boldsymbol{\beta})$ tal que (7.31) se verifica para todo x en el espacio generado por X. Extendiendo sus argumentos a los de LIN et al. (1993), se puede demostrar que $S_{\Lambda}(b)$ es consistente frente a la alternativa general de que la función de enlace no está bien especificada en el modelo (7.31). Además, $S_j(b)$ es consistente frente a cualquier alternativa bajo la cual $W_j(x,b)$ es no centrado en cero para todo x, es decir, $\lim_{N\to\infty} \frac{1}{\sqrt{N}} W_j(x,b)$ es no cero para algún x. En general $W_j(x,b)$ puede no estar centrada en cero para todo x si la forma funcional de X_j está incorrectamente especificada. Esto es particularmente verdad para la regresión lineal cuando no existe ningún modelo adicional erróneamente especificado y X_j es independiente de las demás covariables.

Las irregularidades vistas en las gráficas de los residuos frente a X_j pueden ser un artificio causado por una forma funcional defectuosa para otra covariable estrechamente correlacionada con X_j. Además, una forma funcional defectuosa para una covariable puede manifestarse en el plot de los residuos frente a los valores ajustados, sugiriendo una elección incorrecta de la función enlace. Es decir, todos los métodos propuestos chequean el

ajuste del modelo entero especificado por (7.31). No obstante, en general W_j da más información acerca de la forma funcional de X_j, y W_Λ a cerca de la función de enlace.

En síntesis se tiene, bajo la *hipótesis nula* de *especificación correcta de la forma funcional del modelo,* la distribución de la suma acumulada de los residuos puede ser aproximada por un proceso estocástico Gaussiano de media cero cuyas realizaciones pueden ser generadas por simulaciones de ordenador. *Según los resultados obtenidos por LIN et al. (2002), si la estructura funcional de una variable explicativa lineal está correctamente especificada, la suma acumulada de los residuos sobre la variable debe estar centrada alrededor de cero y mostrar un patrón sin tendencia sistemática.*

Para analizar la linealidad de nuestras variables de referencia, X_3, X_5, X_7 y X_{26}, empezamos por considerar el modelo lineal general con todas las variables explicativas que presentan una linealidad significativa con $\alpha \leq 0.10$, a las que se añade en cada caso una variable para la que la linealidad no es significativa. Cada modelo resultante se ajusta a los datos de entrenamiento a través del procedimiento GENMOD de SAS V9.2, que tiene implementado el contraste y representación de los residuos acumulados de LIN (2002) para detectar la no linealidad de las variables, con función de *enlace logit* y *distribución binomial*. Comenzamos, por ejemplo, por la variable X_7, para la que usando este procedimiento, comprobamos gráfica y numéricamente la suma acumulada de residuos a través de 1.000 rutas de simulación. La representación gráfica de los residuos acumulados se muestra en la figura 7.13.

La línea de color azul sólido en la figura 7.13 es la suma acumulada de los residuos observados de X_7 para el modelo (7.28), para cualquier valor x sobre el eje horizontal, el valor correspondiente del eje vertical es la suma de los residuos asociados con los valores de la covariable menores o iguales que x. Las líneas de puntos más suaves muestran las rutas de simulación del proceso estocástico de Gauss. Los p-valores de la hipótesis nula de que la forma funcional de las variables independientes se ha especificado correctamente se recogen en la esquina inferior derecha del gráfico del panel.

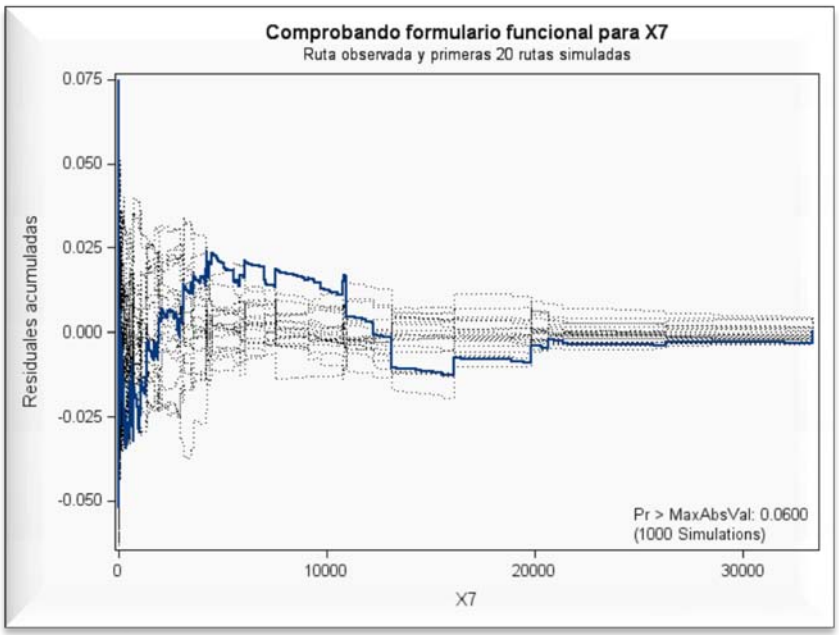


Figura 7.13.- Residuos Acumulados de la variable X_7.

En la gráfica de la figura 7.13 se observa que las curvas generadas desde la distribución nula tienden a situarse de una manera similar que la línea azul observada y cortan al eje x con la misma frecuencia que dicha línea, es decir, la línea que representa *la suma acumulada de los residuos está centrada en cero, como cabe esperar de ser correcta la especificación del modelo logístico lineal con la incorporación de* X_7 *al modelo.* Por esta razón el test de linealidad arroja el resultado de $\left(\Pr > MáxAbsVal\right) < 0.0600$, tabla 7.19, es decir, se rechaza, al nivel de significación $\alpha = 0.050$, la hipótesis de que la especificación estructural del modelo lineal logístico, con la incorporación de X_7 al modelo (7.28), no es correcta. *Por lo tanto* X_7*, en principio, puede pasar a formar parte de la componente lineal del modelo* (7.28).

Por el contrario para la variable X_{26}, como se puede apreciar en la figura 7.14 las curvas generadas desde la distribución nula tienden a situarse más próximas a la vez que cortan al eje x más frecuentemente que la línea azul observada, es decir, la línea que representa *la suma acumulada de los residuos no está centrada en cero, como debería de ocurrir de ser correcta la especificación del modelo logístico lineal con la incorporación de* X_{26} *al modelo.*

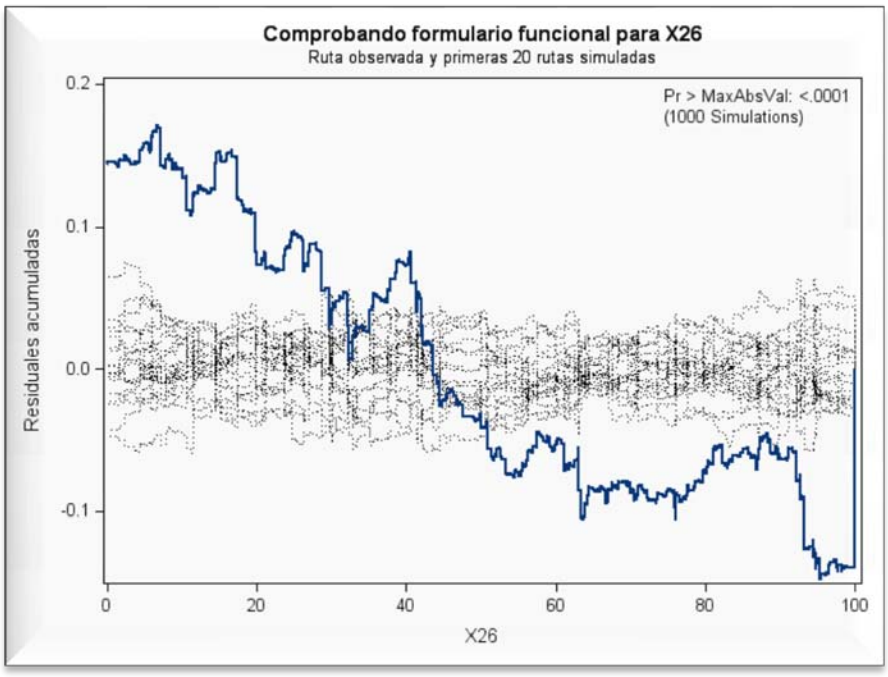


Figura 7.14.- Residuos Acumulados de la variable X_{26}.

Por otro lado, en la tabla 7.19, donde se muestra el resumen de valoración de los residuos acumulados del modelo (7.28) para las variables analizadas, se observa que, para la variable X_{26}, el máximo valor absoluto de la suma acumulada de residuos observados es 0,1705, lo que significa que después de 1.000 realizaciones desde la distribución nula menos de un 0,01% tiene un máximo mayor que 0,1705, o, en términos equivalentes, el valor crítico p para el test de linealidad de los residuos acumulados, *test supremo de Kolmogorov*, es $\left(\Pr > MáxAbsVal\right) < 0.0001$, de ahí la asimetría tan pronunciada de su correspondiente curva de representación de la suma acumulada de los residuos.

Tanto los resultados numéricos como los resultados gráficos sugieren que el modelo (7.31), *especialmente la forma funcional* $logit\left(P\left(Y=1/X=x\right)\right)=\beta_0+\beta^T X$ *puede no ser apropiada para la variable* X_{26}. Lo mismo que para la variable X_{26} ocurre para las variables X_3 y X_5.

Las representaciones gráficas de las funciones de densidad, de los histograma de frecuencia y de los residuos acumulados de las variables X_3, X_5, X_7 y X_{26} se muestran en las figuras 7.16 y 7.17. Para cada variable se muestra un panel con la representación gráfica de los residuos acumulados y, en alguna esquina de la gráfica correspondiente el p valor crítico del *test del supremo de* Kolmogorov. El resumen de valoración de los residuos acumulados del modelo (7.28) se muestra en la tabla 7.18.

Tabla 7. 18.- Resumen de valoración de los residuos acumulados del modelo (7.28).

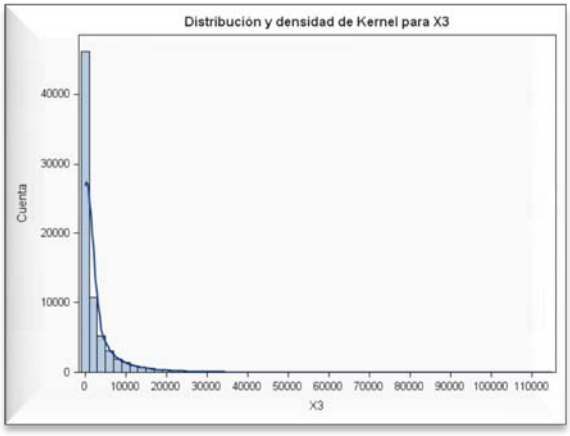
Variable de valoración	*Máximo valor absoluto*	*Reiteración*	*Pr>MaxAbsVal*
X_3	0.2062	1000	<.0001*
X_5	0.1862	1000	<.0001*
X_7	0.0746	1000	0.0600
X_{26}	0.1705	1000	<.0001*

*Significativo al nivel $\alpha = 0.05$.

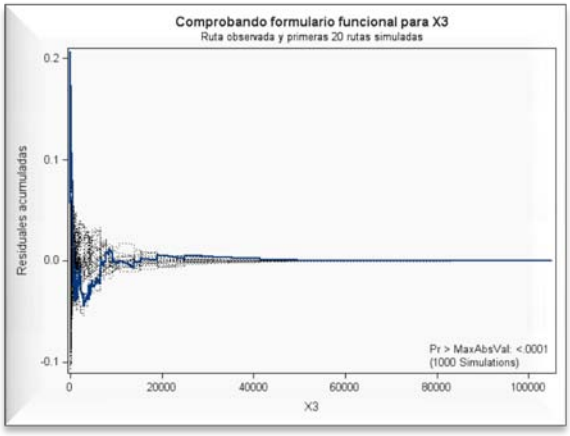
Según lo indicado por los p-valores críticos, podrían ser necesarias formas funcionales más apropiadas, no lineales, para X_3, X_5 y X_{26}.

Tanto la gráfica de residuos acumulados, como el test supremo de Kolmogorov indican que X_7*, al nivel* $\alpha = 0.05$*, podría formar parte de la componente lineal del modelo.*

Una vez detectadas las variables no lineales en el modelo logístico lineal es lógico abordar la cuestión sobre ***cuál es la forma funcional adecuada para las variables cuya especificación lineal es incorrecta****. En la siguiente sección* 7.4*, dedicada a la especificación del modelo, se profundiza en esta cuestión abordando el análisis de los patrones de no linealidad más utilizados en los últimos años dentro del credit scoring, a la vez que haremos nuestra propuesta y la compararemos con las más habituales.*

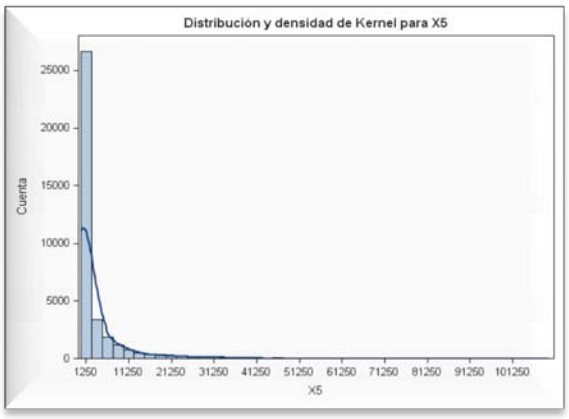


7.15.a

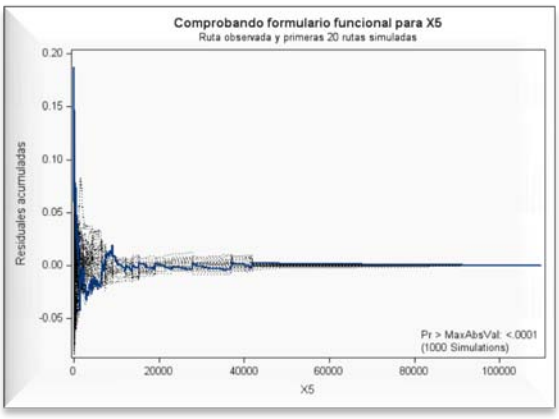


7.15.b

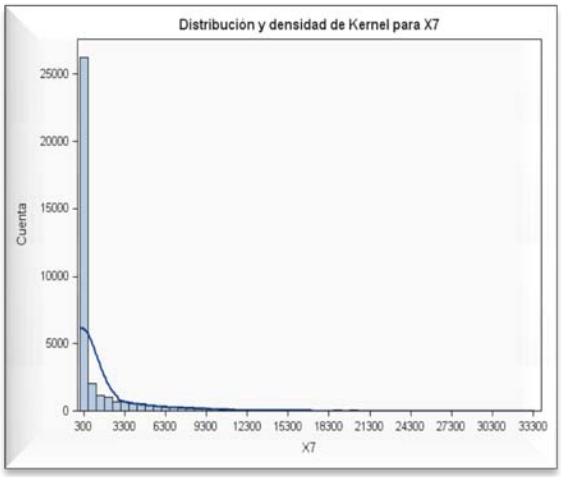
Figura 7.15.- Función de densidad estimada por núcleos y representación de residuos acumulados para la variable X3.



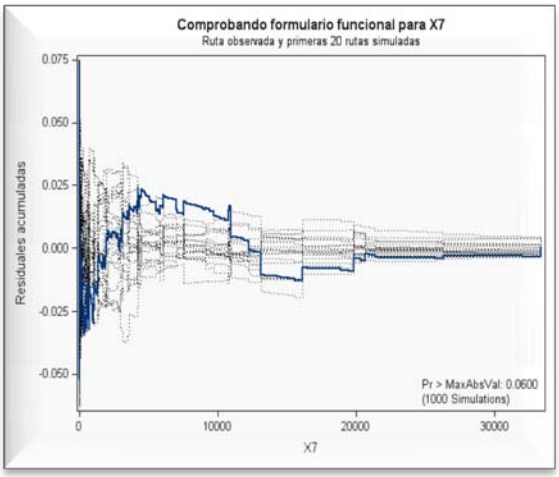
7.16.a



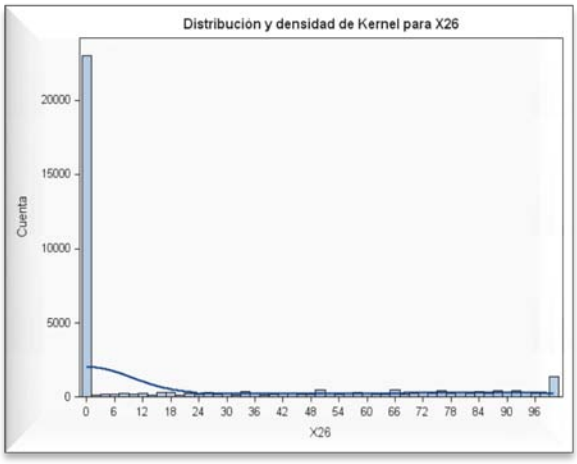
7.16.b



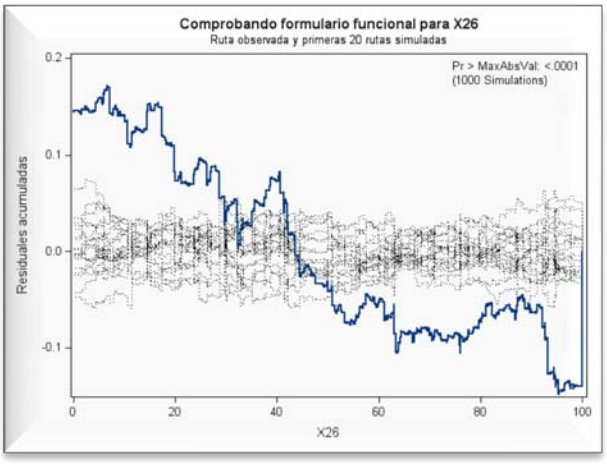
7.16.c



7.16.d



7.16.e



7.16.f

Figura 7.16.- Función de densidad estimada por núcleos y representación de residuos acumulados para las variables X_5, X_7 y X_{26}.

7.4 ESPECIFICACIÓN Y AJUSTE DEL MODELO.

7.4.1 Introducción.

En esta sección, nos proponemos identificar el modelo más adecuado para nuestro objetivo combinando toda la información obtenida de las etapas de preanálisis y de exploración de los datos de entrenamiento con la reflexión sobre las relaciones entre los elementos fundamentales de un modelo de credit scoring aprovechando el conocimiento experto del riesgo de crédito y ajustarlo a los *datos de entrenamiento*.

El objetivo de la especificación del modelo estadístico consiste en seleccionar de entre un conjunto de estructuras funcionales posibles aquellas más relevantes para describir las principales características de la variable estado de default y que se espera generalice mejor, lo que se alcanzará a través de la ejecución de diferentes modelos en orden a elegir el mejor, usando los datos de validación disponibles. Este proceso implica tomar decisiones que conciernen a la formulación de la componente sistemática, los supuestos sobre la componente aleatoria y cómo se combina las dos componentes en el modelo.

La especificación de la estructura funcional es una de las cuestiones fundamentales en la construcción de un modelo estadístico, por cuanto una especificación incorrecta puede dar lugar a estimaciones sesgadas o a coeficientes ineficientes. Por tanto, las primeras cuestiones que hemos de plantearnos son:

a) Si la forma funcional del modelo es correcta.

b) Si todas las variables explicativas pertinentes están incluidas en el modelo y no se ha incluido ninguna de las irrelevantes.

es decir, *si el modelo se ha especificado correctamente*.

La idoneidad del modelo dependerá del grado de conocimiento que se tenga tanto de la distribución conjunta de la variable respuesta y las variables explicativas como de las distribuciones conjuntas de las variables explicativas condicionadas al default y al no default y además deberá obtenerse en función de los datos disponibles, que no siempre son los más adecuados para especificar los mejores modelos.

Una buena formulación de un modelo deberá satisfacer las condiciones siguientes:

a) El modelo explica de forma adecuada el comportamiento del acreditado frente al default.

b) El modelo predice de forma significativamente correcta la probabilidad de default.

c) El modelo clasifica correctamente a nuevos acreditados distintos a los utilizados en su entrenamiento.

d) La puntuación otorgada por el modelo de calificación a un acreditado es fácilmente interpretable en función del "peso" de cada variable en el modelo.

La combinación de las cuatro condiciones del párrafo anterior determina todas las propiedades que ha de tener el modelo más idóneo para cada situación concreta. Sin embargo, existen algunas pautas generales, casi siempre exigibles y en todo caso deseables, que ha de cumplir cualquier modelo con el que se pretendan obtener rendimientos adecuados. Entre ellos destacan los ligados a la cuádruple calidad que Basilea II requiere a un modelo, Capítulo 2, sección 2.6: *flexibilidad, dimensión, facilidad de interpretación y generalización.*

En el Capítulo 6, sección 6.2 se propusieron tres hipótesis, H_1, (6.1), H_2, (6.2) y H_3, (6.3) sobre las que fundamentar tanto la construcción como la estimación de los modelos de probabilidad de default y, por tanto, la función de calificación de acreditados y del clasificador Bayes optimal de nuevos solicitantes de crédito que, en forma resumida, vienen a decir

El modelo expresará la relación existente entre una conveniente transformación de la probabilidad de default, $g\left(P\left(Y=1/\ X=x\right)\right)$, *y una expansión lineal de funciones de base de las variables explicativas,* $\boldsymbol{\beta}^T \boldsymbol{H}\left(\boldsymbol{X}\right)$,

$$g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right) \tag{7.34}$$

donde $\boldsymbol{X}=\left(X_1,...,X_p\right)^T$, $\boldsymbol{\beta}=\left(\beta_1,...,\beta_q\right)^T$ y $\boldsymbol{H}\left(\boldsymbol{X}\right)=\left(h_1\left(X\right),...,h_q\left(X\right)\right)^T$, y, para $r=1,...,q$, $h_r(X):\mathbb{R}^p\rightarrow\ \mathbb{R}$ *es la r-ésima función de base de* $\boldsymbol{X}$.

Además, el conocimiento sobre la distribución poblacional de los acreditados o sobre las distribuciones condicionadas al default y no default es generalmente escaso y los datos son finitos e imperfectos por lo que la información que nos proporcionan es limitada, por lo que el principio de inducción constituye un método adecuado de estimación del modelo.

El objetivo de estas suposiciones se orienta a la construcción de modelos más flexibles, aumentando o reemplazando el vector de variables explicativas originales con variables adicionales, transformaciones de X.

En todo caso, $S(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X})$ es una función sobre la que estableceremos las hipótesis necesarias para obtener el "*mejor modelo posible*" para alcanzar nuestro triple objetivo de predicción del default, calificación de acreditados y clasificación de solicitantes de crédito.

Especificaremos modelos generales del tipo (7.40), es decir, modelos con representación en los *espacios "agrandados" de las expansiones por funciones de base de las variables explicativas, que generalmente son espacios de Hilbert*. Tal como ya expusimos en el Capítulo 6, suponer que la relación existente entre una conveniente transformación de la probabilidad de default y las variables de riesgo de crédito explicativas del estado de default es una expansión lineal de funciones de base del vector de entradas X es la forma más elegante, desde el punto de vista estadístico, de introducir la no linealidad en los modelos que relacionan el estado de default con las variables explicativas, (HASTIE et al., 2009).

Los modelos de credit scoring por expansiones lineales de las variables originales a través de funciones de base que utilizaremos en esta Tesis Doctoral son

$$g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r(X)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}(\boldsymbol{X}) \tag{7.35}$$

Tabla 7.19.- Transformaciones de la probabilidad de default, $g(\bullet)$, más notables.

Probabilidad	$P\left(Y=1/\ X=x\right)$
Logístico	$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)$
Probit	$\text{probit}\left(P\left(Y=1/\ X=x\right)\right)$
Vector Soporte	$Sign\left\{P\left(Y=1/\ X=x\right)-0.5\right\}$

La estructura de un miembro de cualquiera de las cuatro familias de modelos definidos por (7.35) y la tabla 7.19 quedará perfectamente determinada una vez que se fije la matriz de funciones de base $\boldsymbol{H}(\boldsymbol{X})$.

En lo que sigue, utilizaremos sobre todo el modelo logístico, si bien compararemos sus resultados con las estructuras *probit* y *vector soporte*.

Una vez que las funciones de base $h_r(x)$ han sido determinadas, los modelos, en principio parcialmente lineales, son lineales en estas nuevas variables. De hecho, como se ha visto en el Capítulo 5 los modelos lineales participan de la estructura (7.40) para funciones de base muy sencillas.

7.4.2 Aspectos básicos de la Especificación del Modelo.

La especificación del modelo habrá de realizarse contando con al menos dos aspectos claves: los requerimientos de Basilea II y los datos disponibles junto al conocimiento sobre los mismos.

7.4.2.1.- Los requerimientos de Basilea II.

Con respecto a la especificación del modelo, los requerimientos de Basilea se enfocan especialmente en tres aspectos, (Capítulo 2, sección 2.7):

1.- Calidad de la información, es decir, de las variables explicativas del riesgo de crédito a utilizar en la estimación de la probabilidad de default.

2.- Capacidad del modelo para describir el comportamiento del default.

7.4.2.2.- Los datos disponibles y el conocimiento sobre los mismos.

Después del desarrollo de las dos fases de preanálisis y exploración de los datos de entrenamiento nos encontramos con:

a) *Una población de referencia objeto de estudio de 73.207 clientes pertenecientes al segmento de clientes particulares con préstamos*, de los cuales 5.509 son malos y 67.698 son buenos, de acuerdo con su estado de default o de cumplimento de sus obligaciones de pago a fecha 31 de noviembre de 2007. Con una *muestra estratificada con sobre muestreo formada por 16.792 clientes de la población de referencia cuya distribución en default y no default se muestra en la tabla 7.11, así como las submuestras de entrenamiento, validación y test.*

b) Un conjunto de *27 variables preseleccionadas, explicativas del estado de default, más la variable código de incumplimiento*, variable respuesta binaria que, para cada acreditado, recoge el estado de default, observadas sobre la población de clientes considerada. Todas ellas están *exentas de elementos faltantes*, condición que hemos

recodificado en variables transformadas en pesos de la evidencia obtenidos con criterios expertos de riesgo de crédito, *exentas de elementos extremos* y *exentas de colinealidad*. Las 27 variables explicativas son numéricas a excepción de X32 que es cualitativa binaria.

Sobre la muestra de entrenamiento hemos desarrollado una estrategia de exploración de la linealidad y no linealidad de las variables partiendo de tres hipótesis clave:

1.- *La relación teórica entre el estado de default y las variables explicativas se expresa a través de la función logística, (Teorema de Bayes).*

2.- *Tenemos interés en los modelos logísticos paramétricos y semiparamétricos ya sean completa o parcialmente lineales, con componente lineal y componente no lineal,* formadas por funciones de base que puedan ser razonablemente interpretadas.

3.- *La función de pérdida binomial negativa o pérdida logística es la función de pérdida más idónea para estimar el modelo logístico adecuado para describir, explicar y predecir el estado de default en base a apropiadas variables explicativas del riesgo de crédito, y cumpliendo los requerimientos de Basilea II.*

Como consecuencia de la estrategia seguida resultaron dos grupos diferenciados de variables explicativas, por un lado, un grupo donde *la linealidad de la relación de cada variable con el default es altamente significativa*, con un nivel de significación igual o inferior a $\alpha = 0{,}10$:

Tabla 7.20.- Variables con relación lineal significativa, $\alpha \leq 0.10$.

X_{53}, X_{48}, WR_X_{47}, WR_X_{55}, X_{32}, WR_X_{46}, X_8, X_{24}, X_{25}, X_{57}, X_{64}, WR_X_{56}, X_{63}, WR_X_{49}, X_{42}, X_6, X_{45}, X_{58}, X_{15}.

Y, por otro, para los mismos niveles de significación, un segundo grupo donde *las variables no muestran nítidamente una linealidad significativa*:

Tabla 7.21.- Variables que no muestran linealidad significativa, $\alpha = 0.10$.

X_3, X_5, X_7, X_{19}, X_{20}, X_{26}, X_{35} *y* X_{37}.

Además, para las variables continuas sin significación lineal, X_3, X_5, X_7, X_{26}, una vez contrastada la forma funcional por aplicación del *test supremo de Kolmogorov de los residuos acumulados*, se llegó a la conclusión de que *podrían necesitarse formas funcionales más apropiadas que la linealidad para todas ellas a excepción de* X_7:

$$X_3,\ X_5\ y\ X_{26} \tag{7.36}$$

Por su parte, la no significación de X_7 en el modelo lineal parece tener más que ver con su relativa poca importancia para explicar, a través del modelo, el estado de default, que con la especificación lineal, puesto que se tiene $\Pr > MaxAbsVal = 0,0600$.

Por lo que respecta a las variables no continuas X_{19}, X_{20}, X_{35}, X_{37}, que no son linealmente significativas en el modelo M_{27}, como se desprende del diagrama suavizado de dispersión de estas variables frente al logit de la probabilidad de default, figura 7.10, la especificación lineal en el modelo parece no ser correcta para todas ellas, por lo que debería especificarse también de forma no lineal.

La relación de variables consideradas junto con la visión parcial de riesgo de crédito a la que pertenecen se muestra en la tabla 7.22.

Tabla 7.22.- Variables seleccionadas según la clasificación conceptual por visiones parciales de las relaciones del cliente con la Institución Financiera.

Visión General Cliente:	
Socio_Demográficas:	X_{58}
Activo Total:	X_{15}, X_{45}, X_{19}
Relación Activo/Pasivo:	X_{64}, W_X_{56}
Antiguedad Relación Caja:	X_{37}, WR_X_{49}
Ingresos Recurrentes:	X_{35}
Visión de Pasivo:	
PasivoVista:	X_3, X_6, X_8, X_{42}, WR_X_{47}
Pasivo NoVista:	WR_X_{46}, WR_X_{55}
Pasivo Total:	X_5, X_{20}
Visión de Activo:	
Descubiertos:	X_{48}
Excedidos Credito:	X_{32}
Garantía Hipotecaria:	X_{25}
Garantia Personal:	X_{26}
Activo Total:	X_{57}, X_{24}
Visión de Servicios:	
Recibos no básicos:	X_{63}
Visión de Incidencias:	X_{53}

- Las variables en azul poseen linealidad significativa con nivel de significación $\alpha \leq 0{,}10$.

- Las variables en verde requieren formas funcionales más apropiadas que la linealidad.

En la tabla 7.23 se muestra la descripción de las variables seleccionadas junto con las nuevas etiquetas se les han asignado para facilitar la notación en el resto del capítulo. Hemos etiquetado con U_j las variables con linealidad significativa y con V_j las que no presentan tal significación.

Tabla 7.23.- Descripción de las variables seleccionadas.

Etiqueta de Variables		Descripción de Variables
Nueva	**Anterior**	
U_1	X_{53}	Porcentaje de contratos del cliente en incidencia, sobre el total de contratos que hayan estado operativos en algún momento del mes actual.
U_2	X_{48}	Número de meses, en los últimos 12, que el Cliente ha tenido descubierto (saldo deudor) en Ahorro Vista
U_3	X_{32}	Número de meses que el Cliente ha tenido saldo dispuesto por encima del límite en productos de crédito en los últimos 12 meses.
U_4	WR_X_{55}	WOE_ Porcentaje que representa el importe de la nómina / pensión de los últimos 3 meses frente al total de ingresos percibidos en los 3 últimos meses.
U_5	WR_X_{47}	WOE_ Cuota máxima de los productos de préstamos para financiación a particulares vigentes a pagar a fecha de visión.
U_6	X_8	Rango mínimo de saldo (importe máximo – importe mínimo) del Cliente en Pasivo a la Vista para los últimos 12 meses.
U_7	WR_X_{46}	WOE_ Saldo mínimo mensual medio del Cliente en Pasivo a la Vista para los últimos 6 meses.
U_8	X_{24}	Mínimo ratio del plazo pendiente a fin de mes frente al plazo original de la operación en préstamos para financiación a particulares
U_9	X_{57}	Porcentaje que representa el importe en riesgo del cliente en la Caja a fin de mes frente su importe en riesgo total (considerando tanto la Caja como otras Entidades Financieras recogidas en la CIRBE) en el mes de fin del periodo de visión
U_{10}	X_{63}	Número de recibos no básicos cargados en cuenta en los tres últimos meses.
U_{11}	X_{25}	Mínimo ratio de saldo pendiente a fin de mes actual frente al importe formalizado en productos de activo con garantía hipotecaria
U_{12}	X_{64}	Ratio de valor medio de red frente al saldo medio total en pasivo en los 3 últimos meses.
U_{13}	WR_{X49}	WOE_ Número de meses de antigüedad desde que el Cliente se dio de alta en el primer préstamo con garantía personal.
U_{14}	WR_X_{56}	WOE_Ratio de requerimientos de pago totales sobre el saldo en pasivo del Cliente en el mes actual
U_{15}	X_{42}	Número de recibos básicos cargados en cuenta en los dos últimos meses
U_{16}	X_6	Importe medio que el Cliente ha percibido de manera recurrente en los últimos 12 meses.
U_{17}	X_{58}	Edad del Cliente para la fecha de visión.
U_{18}	X_{45}	Importe total que el Cliente tiene en productos de Activo en los últimos 12 meses.
U_{19}	X_{15}	Importe total que el Cliente debe pagar a fecha de observación para hacer frente a los requerimientos de pago del cliente con la Caja correspondientes a productos de Activo.
V_1	X_{26}	Mínimo ratio de saldo pendiente a fin de mes actual frente al importe formalizado en préstamos con garantía personal a fecha de visión
V_2	X_{35}	Número de meses, en los últimos 6, que el Cliente ha percibido ingresos recurrentes
V_3	X_{37}	Número de meses de antigüedad de relación del Cliente en la Caja
V_4	X_3	Saldo medio del Cliente en pasivo líquido para los últimos 12 meses
V_5	X_{20}	Máxima antigüedad en el año del cliente en productos de Pasivo.
V_6	X_{19}	Máxima antigüedad en el último año del cliente en activo
V_7	X_5	Saldo medio del Cliente en productos de pasivo para los últimos 12 meses

7.4.2.3.- Fijación de la estructura funcional apropiada para el modelo.

*

De acuerdo con la metodología general que hemos propuesto en el Capítulo 6, la fijación de la estructura de la expansión de funciones de base (7.34), conlleva establecer en primer lugar la transformación $g(\bullet)$ y en segundo lugar las funciones de base $h_r(X)$.

La elección de la transformación que relaciona la probabilidad de default con una expansión lineal de funciones de base ya la hemos comprometido en el desarrollo de nuestra estrategia de exploración de la linealidad y no linealidad de las variables explicativas sobre la muestra de entrenamiento, pues allí partimos de la base de que, con respecto a la estructura del modelo, *la relación teórica entre el estado de default y las variables explicativas se expresa a través de la función logística, es decir, tenemos interés en los modelos logísticos.*

Por otro lado, hemos encontrado un grupo de 19 variables cuya relación lineal con el default es altamente significativa, con nivel de significación menor o igual a $\alpha = 0{,}10$, tabla 7.20, para las 8 restantes variables la linealidad no es nítidamente significativa, tabla 7.21, y de estas últimas, 7 de ellas podrían necesitar formas funcionales más apropiadas que la linealidad (7.36). En resumen, contamos con 19 variables que presumiblemente ejercen una influencia lineal sobre la variable estado de default, pero desconocemos el tipo de influencia que ejercen las otras 7 variables, razón por la cual *tenemos interés en los modelos parcialmente lineales.*

Por último, por (7.34), *la función de calificación de acreditados se expresa como una expansión lineal de funciones de base que, de acuerdo con Basilea II, puedan ser razonablemente interpretadas.*

De los tres párrafos anteriores se deduce que nuestro objetivo consiste en *encontrar un modelo "adecuado" de la familia de Modelos Logísticos Parcialmente Lineales por Expansión de Funciones de Base*, LPLM.

Antes de proceder a fijar la estructura del modelo a nuestro caso concreto hacemos una breve descripción de los *Modelos Logísticos Parcialmente Lineale*s, LPLM, y sobre todo de la subfamilia integrada por los *Modelos Logísticos Lineales Híbridos por Expansiones Lineales de Funciones Base,* HLLM, que proponemos.

El exhaustivo análisis de la mayor parte de los elementos que configuran un Modelo de Probabilidad de Default desde la perspectiva de los requerimientos de Basilea II y de las reflexiones que sobre los mismos hemos ido haciendo a lo largo de los 6 primeros capítulos de esta Tesis Doctoral nos induce a concluir que la estructura formal más general y

apropiada para un *modelo de credit scoring*, en general, y de un *modelo de credit scoring proactivo*, en particular, es la correspondiente a los *modelos logísticos parcialmente lineale*s, LPLM, extensiones semiparamétricas del *modelo LOGIT*, y cuya estructura se expresa en la forma siguiente

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{Lineal}+\underbrace{h\left(V_1,...,V_{p_2}\right)}_{No\ Lineal}=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{U}\right)+h\left(\boldsymbol{V}\right) \tag{7.37}$$

donde $X^T=\left(U^T,V^T\right)$, $U^T=\left(U_1,...,U_{p_1}\right)$ vector de p_1 variables lineales, $V^T=\left(V_1,...,V_{p_2}\right)$, vector de p_2 variables no lineales y $\boldsymbol{H}\left(\boldsymbol{U}\right)=\left(h_1\left(U_1\right),...,h_{p_1}\left(U_{p_1}\right)\right)$.

Por un lado, la parte lineal del modelo (7.37) es la combinación lineal de las variables de linealidad significativa $U=\left(U_1,...,U_{p_1}\right)^T$, que puede verse como una expansión lineal de funciones de base, $\sum_{r=1}^{p_1}\beta_r U_r=\sum_{r=1}^{p_1}\beta_r h_r\left(X\right)$, donde las funciones de base verifican $h_r\left(X\right)=U_r$, para $r=1,..,p_1$.

Por otro lado, la componente no lineal, se diseña como una expansión no paramétrica $h(\boldsymbol{V})$, donde $h(\bullet)$ es una función no paramétrica infinito dimensional, es decir, una función de suavizado no paramétrica de dimensión infinita que opera sobre un argumento multidimensional de variables $V^T=\left(V_1,...,V_{p_2}\right)$, y se calcula de una manera flexible, por ejemplo, cualquier método de suavizado no paramétrico.

Los modelos generales (7.37), más flexibles que los modelos logísticos lineales y con más facilidad interpretativa y menos complejidad técnica que los modelos totalmente no paramétricos, se estiman a través de la *Regresión Logística Parcialmente Lineal, LPLR*.

Con el objetivo de conseguir modelos de credit scoring lo más parsimoniosos posibles, con el suficiente poder predictivo y fácilmente interpretables, *nuestra propuesta, capítulo 6, consiste en los Modelos Logísticos Lineales Híbridos por Expansiones Lineales de Funciones Base*, HLLM, *modelos similares a* (7.37) *pero con la diferencia fundamental de cómo se especifica la componente no lineal, cuyo método de construcción se enmarca dentro de los métodos de restricción supervisada, con la siguiente estructura formal:*

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{Lineal}+\underbrace{\sum_{r=1}^{p_2}Z_r\left(V_r\right)}_{No\ Lineal} \tag{7.38}$$

La componente no lineal, se diseña como una componente aditiva de p_2 *de funciones de las variables no lineales,* $Z_r(V_r)$*, una para cada variable* V_r*,* $r=1,...,p_2$.

Para cada $r=1,..,p_2$, $Z_r(V_r)$ se expresa en la forma siguiente:

$$Z_r(V_r)=\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r),\ r=1,...,p_2 \tag{7.39}$$

siendo K_r el número de funciones de base de la expansión de V_r, $h_{r,k}(\cdot)$ su k-ésima función de base y $p_1+\sum_{r=1}^{p_2}K_r=q$.

La estructura funcional del modelo adopta entonces la forma

$$\text{logit}\left(P(Y=1/X=x)\right)=\underbrace{\beta_0+\sum_{r=1}^{p_1}\beta_r U_r}_{\textit{Lineal}}+\underbrace{\sum_{r=1}^{p_2}Z_r(V_r)}_{\textit{No Lineal}} \tag{7.40}$$

El calificativo "híbrido" se debe a que las variables no lineales V_r, $r=1,...,p_2$, *se expresan como combinaciones lineales de funciones de base específicas diferentes, que dependen de la no linealidad subyacente implícita en* V_r, *por lo que la componente no lineal resultará híbrida. Las funciones de base podrán ser, por ejemplo, transformaciones polinómicas o logarítmicas, pesos de la evidencia obtenidas por un proceso de tramado óptimo de la variable* V_r*, funciones constantes a trozos resultados de particiones recursivas de árboles de clasificación, obtenidas, por ejemplo, por el algoritmo CART sobre* V_r*, funciones lineales a trozos, splines, ya sean de regresión, de suavizado o de penalización, funciones sierra obtenidas por el método PPR sobre una sola variable,* V_r*, funciones bisagra obtenidas por el procedimiento MARS sobre* V_r*, funciones de base radial Gaussiana, etc.*

La clave de estos modelos consiste en que *una vez que se han determinado para cada variable no lineal* V_r *las funciones de base que forman la expansión lineal* (7.39), $Z_r(V_r)=\sum_{k=1}^{K_r}\theta_{rk}h_{rk}(V_r)$, *el complejo modelo no lineal* (7.37) *se derrumba dando paso a un modelo lineal en el nuevo espacio expansionado, MLLH,* (7.40), *que se puede estimar por el método general de estimación de los modelos logísticos,* propuesto en el Capítulo 6, subsección 6.5.

La función objetivo a optimizar para estimar un modelo MLLH, viene determinada por *la función de pérdida binomial negativa o pérdida logística*, y se expresa en la forma

$$L_{emp\,\ell_\Lambda}\left(Y,S\left(X\right)\right)=-\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)-\boldsymbol{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)\right)\right]+\lambda J\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right) \tag{7.41}$$

Dado que nos encontramos ante un *método de restricción*, donde decidimos a priori limitar la clase de funciones base, la complejidad del modelo está controlada por el número de variables lineales y por la expansión por funciones de base que fijemos para cada variable no lineal. Si se siguen con rigor los requerimientos de Basilea II, lo recomendable será no complicar en exceso las expansiones por funciones de base de las variables V_r, $Z_r\left(V_r\right)=\sum_{k=1}^{K_r}\theta_{rk}h_{rk}(V_r)$, por lo que en nuestro caso de riqueza de datos y método supervisado no será necesario regularizar el modelo, es decir, fijamos $\lambda=0$. De este modo el problema de optimización de los modelos logísticos híbridos por expansión lineal de funciones de base, (7.41), se reduce a (6.44)

$$\underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}}{Min}\ -\left[\boldsymbol{Y}^T\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)-\boldsymbol{1}^T\log\left(1+\exp\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)\right)\right] \tag{7.42}$$

y la solución viene dada por (6.45)

$$\left(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}\right)^{nuevo}=\left[\left(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}\left(V\right)\right)^T\boldsymbol{W}\ \left(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}\left(V\right)\right)\right]^{-1}\left(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}\left(V\right)\right)^T\boldsymbol{W}\boldsymbol{z} \tag{7.43}$$

donde *variable respuesta ajustada* o *variable respuesta de trabajo se expresa*

$$\boldsymbol{z}=\left(\left(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}\right)^{anterior}\right)^T\left(\boldsymbol{1},\boldsymbol{U},\boldsymbol{H}\left(V\right)\right)+\boldsymbol{W}^{-1}\left(y-P\right)$$

7.4.2.4 Selección de las funciones de base para la componente no lineal del modelo.

7.4.2.4.1 Introducción.

Tal como hemos visto, para la construcción del modelo de credit scoring proactivo que proponemos, contamos con 19 variables con influencia lineal sobre la variable estado de default, $U_1,...,U_{19}$, pero desconocemos el tipo de influencia que ejercen las otras 7 variables, $V_1,...,V_7$. Si consideramos este hecho como hipótesis de partida y establecemos la siguiente notación, tal como se muestra en la tabla 7.23,

$X^T=\left(U^T,V^T\right)$ donde $U^T=(U_1,...,U_{19})$ y $V^T=(V_1,...,V_7)$, entonces nos encontramos ante un modelo logístico parcialmente lineal, semiparamétrico, con estructura (7.37),

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{19}\beta_r U_r+h\left(V_1,...,V_7\right) \tag{7.44}$$

Nuestro objetivo consiste en "convertir" el modelo (7.37) en un modelo con estructura formal (7.40), donde $Z_r\left(V_r\right)=\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r),\ r=1,...,7$, es decir,

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{19}\beta_r h_r\left(U_r\right)+\sum_{r=1}^{7}\left(\sum_{k=1}^{K_r}\theta_{r,k}h_{r,k}(V_r)\right) \tag{7.45}$$

estructura que quedará perfectamente especificada una vez que para cada variable $V_r,\ r=1,...,7$, hayan sido seleccionadas las funciones de base $h_{r,k}(V_r),\ k=1,...,K_r$.

Para seleccionar las funciones de base que formarán parte de la componente no lineal del modelo de credit scoring proactivo, utilizaremos la metodología que para tal fin propusimos en la sección 6.6 de esta Tesis Doctoral. Recordemos que el método constructivo consta de los tres pasos siguientes:

1.- En primer lugar se considera un modelo de partida con todas las variables explicativas linealmente significativas que se consideran de interés en riesgo de crédito.

2.- En segundo lugar se hace uso de un método constructivo para seleccionar la combinación lineal de funciones de base "más prometedora" para cada variable no lineal de un conjunto de candidatas e incorporarla al modelo. Este método consiste en ir añadiendo, en cada paso, al modelo inicial las combinaciones lineales cuyas funciones de base resulten significativas, utilizando el *test chi_cuadrado*, en el modelo logístico conseguido en el paso anterior. En cada paso se irá analizando si la incorporación de funciones de base incide en la significación del modelo de otras variables ya incorporadas, a la vez que se comprobará el *ajuste del modelo a los datos de entrenamiento*, a través de los *pseudo-coeficientes de determinación*, *MCFadden y Nagelkerke*, del *Error Empírico* y de los *criterios de Información de Akaike*, AIC, y *Schwarz*, BIC, y el *poder discriminante del modelo*, a través del *área bajo la curva ROC*, AUC.

Antes de proceder a la selección de las funciones de base más prometedoras en cada paso, se realiza un proceso de evaluación sobre la muestra de validación, obtenida de forma

aleatoria e independiente de la muestra de entrenamiento para tal fin. El proceso sobre la muestra de validación conlleva prácticamente los mismos estadísticos que para la muestra de entrenamiento. Se procede secuencialmente según el proceso anterior para todas y cada una de las variables de la componente no lineal. A la vez que se construirán modelos alternativos considerando expansiones lineales por funciones base que aunque no sean las más prometedoras parezcan en principio adecuadas, o al menos de interés a efectos comparativos.

3.- Como resultado del proceso de construcción expuesto en el punto 2 anterior se llega a un conjunto de modelos con estructura inicialmente valida pero que podría, en principio, sobre ajustar los datos, no poseer la cualidad de facilidad interpretativa etc. Para evitar tales inconvenientes, en una tercera fase se procede a aplicar técnicas de poda o regularización para reducir el número de funciones de base.

Antes de comenzar a construir el modelo debemos de fijar que funciones de base pueden ser apropiadas para construir las expansiones lineales para cada variable. En el Capítulo 6, sección 6.7, se hace una exposición detallada de aquellas funciones de base más habituales en modelos de credit scoring o que nos consta se han utilizado alguna vez, f*unciones de base polinómicas de 2º orden, pesos de la evidencia asignados a tramados estadísticos automáticos o a particiones recursivas, splines cúbicos restringidos de Stone y Koo, RCS, funciones de base bisagra obtenidas por MARS univariante, funciones de base radial, RBF.*

En esta sección expondremos brevemente las principales características de las mismas, y sus debilidades y fortalezas, en relación con los modelos de riesgo de crédito.

1.- Las primeras funciones de base que se han utilizado para expandir la no linealidad a espacios agrandados de Hilbert posiblemente sean los **funciones de base polinómicas**. Las funciones de base polinómicas de orden superior a 2 y las funciones de base del tipo $h_r(V) = V_r V_k, \ldots.$, tienen el inconveniente de que el número de variables crece exponencialmente con el grado del polinomio. Por otro lado, *según aumenta su número crece su colinealidad,* lo que conduce a estimadores de los parámetros de alta varianza y a numerosos problemas de cálculo numérico. Además, los polinomios de alto orden tienen tendencia a oscilar descontroladamente si existen huecos grandes entre los valores de la variable. Por estas razones expandiremos la variable V_r a través de la variable $(V_r)^2$, es decir funciones de base polinómicas de orden 2. La representación

grafica de $(V_r)^2$ frente a V_1 se muestra en el panel superior izquierdo de la figura 7.17.

2.- (BREIMAN, et al., 1984), utilizaron en su algoritmo CART para arboles de regresión y clasificación ***funciones de base indicadores de regiones pertenecientes a particiones recursivas binarias***. Una de las fortalezas de los arboles de clasificación es su capacidad para detectar de forma automática estructuras complejas entre las variables, para trabajar con un nivel de ruido relativamente alto y con datos faltantes y su eficiencia computacional. Aprovecharemos esta fortaleza sustituyendo en el modelo logístico parcialmente lineal cada variable no lineal por la función $Z_r(V_r)=\sum_{k=1}^{K_r} I_{[V_r\in R_{rk}]}WOE(R_{rk})$, siendo $WOE(R_{rk})=Ln\left(\frac{P_{\text{Buenos en R}_{rk}}}{P_{\text{Malos en R}_{rk}}}\right)$, tal como se expone en el capítulo 6, subsección 6.7.3, donde los K_r regiones $\{R_{r1},...,R_{rK_j}\}$ representan una partición recursiva binaria optima de la variable V_r y los $WOE(R_{rk})=Ln\left(\frac{P_{\text{Buenos en R}_{rk}}}{P_{\text{Malos en R}_{rk}}}\right)$ son los pesos de la evidencia en la región R_{rk} de la variable V_r.

Una alternativa a la partición recursiva del rango de la variable puede consistir en un *método automático* que proporcione una partición optima, incluso un tramado que partiendo inicialmente de una *partición óptima obtenida por métodos estadísticos sea modificada con criterios de riesgo de crédito* (SIDDIQI, 2006). De hecho, se sustituyeron, ya desde el preanálisis, las variables con un número elevado de datos faltantes por expansiones de este tipo asignando el peso de la evidencia, (WOE), a las categorías resultantes de un proceso de tramado óptimo supervisado, MDLP, basado en la entropía de Shannon, (FAYYAD e IRANI (1993), SPSS, IBM®, entre las cuales se encuentra la clase de valores faltantes.

Para cada variable V_r se considera la función $Z_r(V_r)=W\_V_r=\sum_{k=1}^{K_j} I_{[X_r\in I_{rk}]}WOE(I_{rk})$, donde los K_r intervalos o tramos representan una agrupación óptima de la variable V_r y los $WOE(I_{rk})$ son los pesos de la evidencia en el intervalo I_k de la variable V_r. La sustitución de las variables explicativas del riesgo de crédito por las funciones $Z_r(V_r)$ resuelve eficazmente los problemas de datos faltantes, de datos extremos y de clases "raras" en estas variables. Estas características nos hacen pensar en estas funciones de base como interesantes para expandir la no linealidad de una variable explicativa del riesgo.

Las funciones de base obtenidas por asignación de pesos de la evidencia a particiones sean recursivas binarias u obtenidas por optimización estadística automática, se revisen o no con criterios de riesgos, son constantes a trozos y, por tanto, muy fáciles de explicar a los clientes, puesto que la puntuación de los mismos en una expansión lineales de estas funciones de base es directamente proporcional al peso del default frente al no default en la región en que los clientes se sitúan para la expansión en cuestión.

La representación grafica de W_V_1 y WT_V_1 frente a V_1 se muestra en los paneles superior derecho y central izquierdo de la figura 7.17, respectivamente.

3.- Otro importante tipo de expansiones por funciones de base que utilizaremos lo conforman los **Splines Cúbicos Restringidos k-anudados de Stone y Koo**, **RCS**, (STONE Y KOO, 1985), que son un caso particular de los *splines de regresión* que, como hemos visto en la subsección 6.7.4 del Capítulo 6, poseen magnificas propiedades teóricas y prácticas. Los RCS poseen la particularidad de que tienen primeras y segundas derivadas continuas en los nudos, (HASTIE y TIBSHIRANI, 1990), para suavizado visual, (DURRLEMAN Y SIMON, 1989) y, además, las expansiones por splines cúbicos restringidos se construyen con un número relativamente pequeño de funciones de base de cálculo sencillo, lo que conduce a modelos más parsimoniosos, por lo que se están revelando como muy eficaces en campos como la Medicina y *creemos que pueden ser igualmente eficaces en la construcción de modelos de credit scoring con variables no lineales*.

La idea de Stone y Koo consiste en ajustar un efecto principal continuo por una segmentación de suavizadores polinomiales de grado 3, para lo que se generan fórmulas para funciones base, que permiten ajustar splines restringidos a la linealidad en las colas, es decir, linealidad por encima del último nudo y por debajo del primero, por lo que los *splines* k *-anudados de Stone y Koo son* conservadores comparados con las polinomiales puras, en el sentido de que la extrapolación fuera del rango de los datos es una línea recta, en vez de una curva polinomial.

Si V_r es una *variable continua* que forma parte de la componente no lineal del modelo y k es el número de nudos, introducidos sobre el eje V_r, localizados en $\xi_1 < \xi_2 < ... < \xi_k$, las funciones de base generadas por el método de Stone y Koo para V_r, $h_1(V_r),..., h_{k-1}(V_r)$, se obtienen en la forma siguiente:

$$h_1(V_r) = V_r$$

$$h_l(V_r) = \left(V_r - \xi_{r(l-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(k-1)}\right)_+^3 \left(\xi_{rk} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} + \frac{\left(V_r - \xi_{rk}\right)_+^3 \left(\xi_{r(k-1)} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}},\ l = 2,...,k-1$$

$$\text{donde } (Z)_+ = \begin{cases} Z & \text{si } Z > 0 \\ 0 & \text{si } Z \le 0 \end{cases}$$

y la expansión de V_r combinación lineal de tales funciones de base se expresa entonces en la forma (7.46):

$$RCS\_V_r = \beta_r V_r + \sum_{l=2}^{k-1} \theta_{r+l-1} \left(\left(V_r - \xi_{r(l-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(k-1)}\right)_+^3 \left(\xi_{r,k} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} + \frac{\left(V_r - \xi_{rk}\right)_+^3 \left(\xi_{r(k-1)} - \xi_{r(l-1)}\right)}{\xi_{rk} - \xi_{r(k-1)}} \right)$$

es decir, se trata de expandir las variable V_r a través de splines de regresión cúbicos restringidos, RCS, con k nudos, $\{\xi_{r1},...,\xi_{rk}\}$. Nos referiremos a la expansión correspondiente a la variable V_r como RCS_V_r.

Las cuatro funciones de base $h_r\_V_1$ son funciones de V_1 y de los nudos, pero son independientes de Y.

Por no tener razones suficientes para suponer una ubicación concreta de los nudos, los fijamos de acuerdo con las siguientes reglas empíricas, Capítulo 6, subsección 6.7.4:

> Por regla general el número de nudos suele oscilar entre 3 y 7. Menos de 5 para un número pequeño de acreditados, $N \le 100$, y 5 o más para un número suficientemente grande, $N > 100$. De existir un número suficientemente grande de acreditados, se colocan el primero y el último nudos en los percentiles 5 y 95 respectivamente.
>
> En el spline cúbico restringido con 5 nudos, N=36.607, correspondiente a la variable V1, que se detalla en el Capítulo 6, subsección 6.7.4 , se muestra en el panel central derecho de la figura 7.17. El nudo correspondiente al percentil 5 es $k_1 = 0$ y el correspondiente al percentil 95 es $k_5 = 95{,}05$. Los otros tres nudos se obtienen dividiendo el rango de V_1 entre los nudos $k_1 = 0$ y $k_5 = 95{,}05$ en cuatro partes iguales, considerando las fronteras de las 4 regiones obtenidas que se corresponden con $k_2 = 23{,}75$, $k_3 = 47{,}52$ y $k_4 = 71{,}2$.

4.- Para expandir linealmente las variables no lineales de nuestro modelo logístico parcialmente lineal inicial (7.37), utilizaremos también para cada variable continua no lineal V_r las funciones de base conocidas como ***funciones bisagra,*** Capítulo 6, subsección

6.7.5, que obtendremos como resultado de aplicar el método **MARS** , (*Splines de Regresión Adaptativos Multivariantes*, Capítulo 5, apartado 5.6.2.3), a modelos univariantes con variable respuesta el estado de default y variable independiente V_r .

MARS usa expansiones lineales de funciones base lineales a trozos de la forma $(V_r - t)_+$ y $(t - V_r)_+$ donde "+" significa parte positiva. Es decir la expansión lineal MARS univariante que utilizaremos para la variable V_r , se viene dada por

$$Z(V_r) = MARS\_V_r = \sum_{l=1}^{K_r} \theta_{rl} h_{rl}(V_r) \tag{7.47}$$

donde $h_{rl}(V_r) \in \boldsymbol{B}$ y

$$\boldsymbol{B} = \left\{ (V_r - t)_+ ,\ (t - V_r)_+ ;\ \ t \in \{v_{1r}, v_{2r}, ..., v_{Nr}\},\ \ r = 1,2,..., p_2 \right\}$$

$$(V_r - t)_+ = máx(0, X_j - t) = \begin{cases} V_r - t, & si\ V_r > t \\ 0 & en\ otro\ caso \end{cases}$$

$$(t - V_r)_+ = máx(0, t - X_j) = \begin{cases} t - V_r, & si\ V_r < t \\ 0 & en\ otro\ caso \end{cases}$$

es decir, splines lineales a trozos.

Con el fin de distinguir las funciones bisagra MARS de los splines cúbicos restringidos notaremos la función bisagra l-ésima para la variable V_r por la notación habitual de estas funciones en la literatura sobre Splines de Regresión Multivariantes Adaptativos, $BF_l\_V_r$.

Para la variable V_1 y nuestros datos de entrenamiento, las funciones de base "bisagra" seleccionadas como resultado de aplicar el método MARS al modelo univariante con variable respuesta el estado de default y variable independiente V_1, son

$$h_{11} = máx(0,\ V_1 - 98{,}9),\ h_{12} = máx(0,\ 98{,}9 - V_1),\ h_{13} = máx(0,\ V_1 - 1{,}5),$$

$$h_{14} = máx(0,\ V_1 - 8{,}36),\ h_{15} = máx(0,\ V_1 - 78{,}74),\ h_{16} = máx(0,\ V_1 - 73{,}8),$$

$$h_{17} = máx(0,\ V_1 - 79{,}94),\ h_{18} = máx(0,\ V_1 - 68{,}97).$$

Obviamente nos interesan aquellas funciones de base que tengan relación lineal significativa con el estado de default, por lo que ajustamos nuestros datos de entrenamiento por regresión logística a un modelo donde variable V_1 adopta la forma

$$\sum_{l=1}^{8}\theta_{1,l}BF_{1,l} = \theta_{1,1}\ máx\left(0,\ V_1-98{,}9\right)+\theta_{1,2}\ máx\left(0,\ 98{,}9-\ V_1\right)+\ \theta_{1,3}\ \ máx\left(0,\ V_1-1{,}5\right)$$
$$+\ \theta_{1,4}\ \ máx\left(0,\ V_1-8{,}36\right)+\theta_{1,5}\ \ máx\left(0,\ V_1-78{,}74\right)+\ \theta_{1,6}\ máx\left(0,\ V_1-73{,}8\right)$$
$$+\ \theta_{1,7}\ \ máx\left(0,\ V_1-79{,}94\right)+\ \theta_{1,8}\ \ máx\left(0,\ V_1-68{,}97\right)$$

Como resultado del ajuste por regresión logística del modelo se obtiene que sólo son linealmente significativas las funciones de base $BF_{1,2}$, $BF_{1,3}$ y $BF_{1,4}$ y la expansión lineal de funciones base para V_1 viene dada por

$$MARS\_V_1=\sum_{l=2}^{4}\theta_{1l}BF_{1l} = 0{,}945315\ máx\left(0,\ 98{,}9-\ V_1\right)+\ 0{,}766475\ \ máx\left(0,\ V_1-1{,}5\right)$$
$$+\ 0{,}194814\ \ máx\left(0,\ V_1-8{,}36\right)$$

Según la metodología que proponemos, la variable V_1 pasa a formar parte del modelo en cuestión a través de la expansión $MARS\_V_1$, cuya representación grafica se muestra en el panel inferior izquierdo de la figura 7.17.

5.- Por último utilizaremos también para expandir la no linealidad de las variables continuas V_1 y V_4 las **Funciones de Base Radial, RBF**, capitulo 6, apartado 6.7.6, caracterizadas por ser funciones *reales que dependen solo de la distancia a un punto* c, llamado centro, es decir, $\phi\left(V_r,c\right)=\phi\left(\left\|V_r-c\right\|\right)$. Usualmente la norma utilizada es la Euclídea, aunque pueden usarse otras normas, así como sumas de funciones RBF para aproximar funciones dadas.

Las funciones radiales son funciones cuya respuesta cambia monótonamente (crece o decrece) con la distancia a un punto central, tal como la función Gaussiana. Los parámetros del modelo son el centro, la métrica y la forma de la función radial.

La forma más general para una función de base RBF es:

$$h_r\left(X\right)=\varphi\left(\left(X-c\right)^T M\left(X-c\right)\right)$$

donde

$\varphi(\cdot)$ *es la función usada (Gaussiana, Multicuadrática,..)*

c es el centro y M es la métrica (7.48)

La cantidad $\left(X-c\right)^T M\left(X-c\right)$ es la distancia entre la observación X de las variables explicativas y el centro definida por la métrica M.

Pueden usarse múltiples tipos de funciones de este tipo, en esta Tesis Doctoral utilizaremos una de las más usuales en la estadística del aprendizaje máquina, la RBF Gaussiana que se expresa en la forma

$$h_r(V_r) = \exp\left(\frac{(V_r - c)^2}{d^2}\right) \tag{7.49}$$

donde la métrica viene definida por la esferas de centro c y radio d y es monótona decreciente desde el centro.

Para la variable V_1, que venimos considerando, y nuestros datos de entrenamiento, la función de base radial RBFG, sobre la esfera unidad, es decir centro cero y radio 1, viene dada por

$$RBFG\_V_1 = \exp\left(-V_1^2\right) \tag{7.50}$$

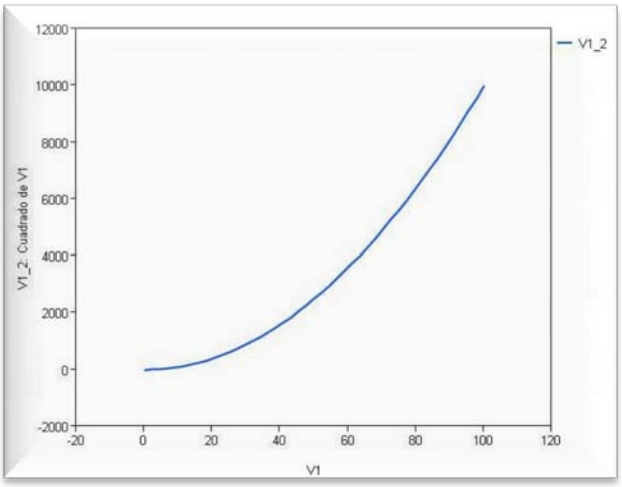
su representación gráfica se muestra en el panel inferior derecho de la figura 7.17.

Un gran reto que presentan los modelo HLLM está en que al presentarse la no linealidad en infinitas formas seguramente existan otras tantas familias de funciones base capaces de aproximarla expandiendo las variables de riesgo de crédito originales a espacios agrandados de Hilbert, ahí sin duda alguna nos espera un enorme y sin duda apasionante campo a explorar, aquí hemos considerado las más notables.

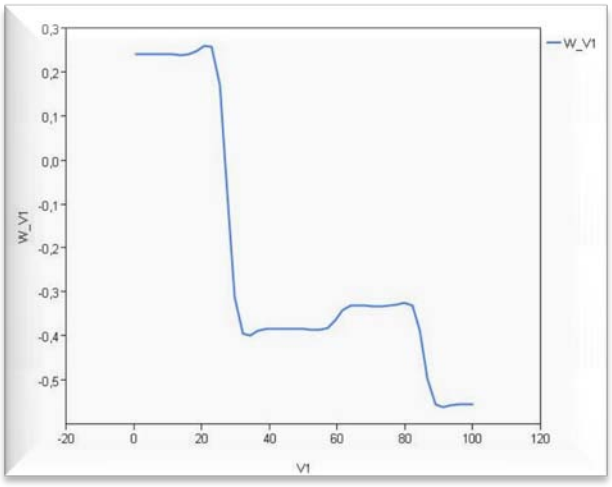
Una tarea fundamental consiste en encontrar las funciones de base que mejor especifican la no linealidad de las variables de riesgo de crédito, campo este apenas explorado y del que depende el éxito de los modelos basados en la expansión lineal de funciones de base. En la siguiente sección exponemos nuestra propuesta de un método de construcción de la componente no lineal del modelo.

7.4.2.4.2 Asignación de las expansiones lineales de funciones de base a las variables de la componente no lineal.

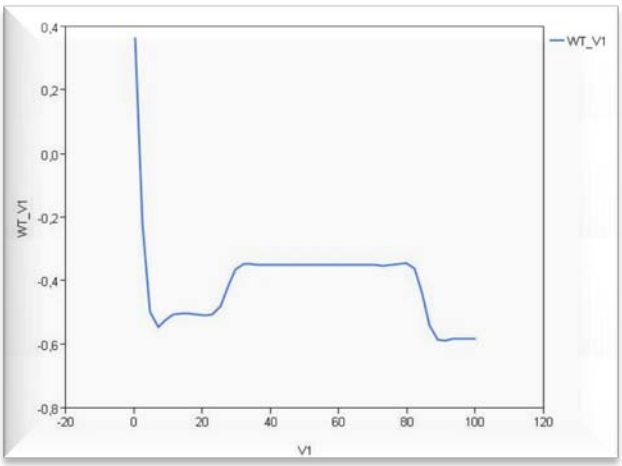
De acuerdo con el primer punto del método de construcción de la estructura de un modelo HLLM, expuesto en el apartado anterior, el modelo de partida, que notaremos $M_{inicial}$, deberá estar formado, en principio, por las 19 variables, $\{U_1,...,U_{19}\}$, tabla 7.20, que como consecuencia de la estrategia desarrollada en las secciones 7.1, 7.2 y 7.3 mostraron una relación de linealidad con el estado de default altamente significativa, con niveles de significación iguales o inferiores a $\alpha = 0{,}10$.



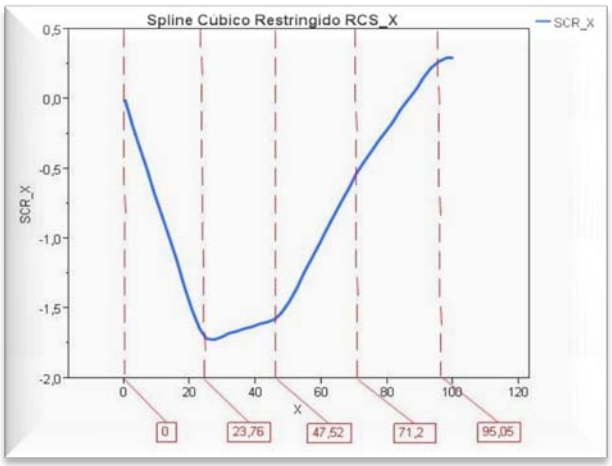
Función de base V_1^2 frente a V_1.



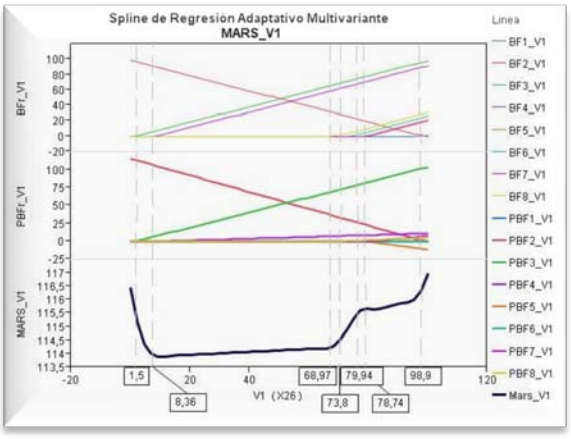
Función de base W_V_1 frente a V_1.



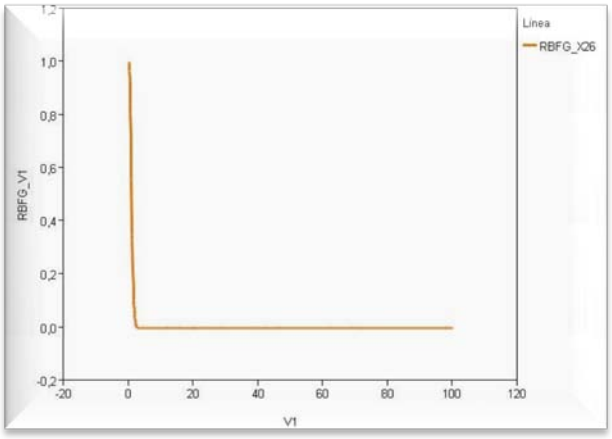
Función de base WT_V_1 frente a V_1.



Spline Cúbico Restringido con 5 nudos para V1.



Expansión lineal MARS_V1.



RBFG_V1 con centro en cero y radio la unidad.

Figura 7.17.- Representación gráfica de la variable no lineal V1 frente a seis funciones de base notables.

Se ha observado que la significación lineal de la variable U_{17}, *edad del cliente en la fecha de visión de los datos*, mejora sustancialmente si consideramos para esta variable la transformación WT_U_{17}, y además que el ajuste y poder discriminante del modelo también mejoran si se prescinde de la variable U_{18}, *importe total que el cliente posee en productos de activo en los últimos 12 meses*, variable de la que tras consultar con expertos en riesgo de crédito decidimos prescindir, puesto que consideran que la visión de activo queda perfectamente representada por la variable U_{19}, *importe total que el cliente debe pagar a fecha de observación para hacer frente a los requerimientos de pago del cliente con la Entidad Financiera correspondientes a productos de activo.* Al igual que ocurre con U_{17} el modelo mejora sustancialmente si se considera W_U_{19}.

Tras las consideraciones anteriores nuestro modelo $M_{inicial}$ adopta la forma

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right) \tag{7.51}$$

De acuerdo con el segundo punto de la metodología constructiva propuesta, se procede a introducir en el modelo distintas expansiones de funciones de base para cada una de las variables de la componente no lineal del modelo, $\{V_1,\ V_2,\ V_3,\ V_4,\ V_5,\ V_6,\ V_7\}$. *Se ajustan los modelos alternativos por LLR y se selecciona la expansión que con criterios de bondad de ajuste, poder predictivo y discriminante y eficacia clasificadora en combinación con los requerimientos de Basilea II y con la política de riesgos de la Entidad Financiera parezca más adecuada.*

Asignación de las funciones de base a la variable continua V1 de la componente no lineal.

En primer lugar expandimos la primera de las variables con linealidad no significativa y presumible no linealidad, V_1. Dado que V_1 es continua, es posible expandirla linealmente a través de funciones de base de, al menos, los tipos considerados en el apartado anterior, 7.4.2.4.1. Por tanto, deberemos estimar por LLR los 6 modelos sintetizados en la expresión siguiente:

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+Z_1\left(V_1\right) \tag{7.52}$$

Cada uno de los 6 modelos en particular se obtiene según la estructura asignada a la expansión de funciones de base $Z_1(V1)=\sum_{k=1}^{K_1}\theta_{1,k}h_{1,k}(V1)$:

1.- $Z_1\left(\boldsymbol{V}_1\right)=\theta_{1,1}h_{1,1}(V)=\theta_{1,1}V_1^2$ (7.53)

2.- $Z_1(\boldsymbol{V}_1) = h_{1,1}(V) = \sum_{k=1}^{4} I_{\left[V_1 \in I_{j,k}\right]} WOE\left(I_{1,k}\right) = \left(W\_V_1\right)$ (7.54)

$\left\{I_{1,1},\ \mathrm{I}_{1,2},\ \mathrm{I}_{1,3},\ \mathrm{I}_{1,4}\right\}$, Partición Automática de V_1 formada por cuatro intervalos.

La partición automática de V_1, $\left\{I_{1,1},\ \mathrm{I}_{1,2},\ \mathrm{I}_{1,3},\ \mathrm{I}_{1,4}\right\}$, se obtuvo utilizando el *algoritmo de tramado óptimo* implementado en el *Nodo Interactive Binning* del *programa Enterprise Mining de SAS®*.

3.- $Z_1(\boldsymbol{V}_1) = h_{1,1}(V) = \sum_{k=1}^{4} I_{\left[V_1 \in R_{1,k}\right]} WOE\left(R_{1,k}\right) = \left(WT\_\boldsymbol{V}_1\right)$ (7.55)

$\left\{R_{1,1}, R_{1,2}, R_{1,3}, R_{1,4}\right\}$ *es una Partición Recursiva de* V_1 formada por 4 regiones en los nodos terminales del algoritmo CART.

La partición recursiva $\left\{R_{1,1},\ R_{1,2},\ R_{1,3},\ R_{1,4}\right\}$ se obtuvo por el procedimiento Árbol de Clasificación de PASW® Statistics de IBM®, con el método de agrupación CHAID. El estadístico de contraste, tanto para partir los nodos como para unir categorías, es el chi-cuadrado de Pearson y el nivel de significación utilizado $\alpha = 0{,}05$; los valores de significación se ajustaron usando el método de Bonferroni.

4.- $Z_1(V_1) = RCS\_V_1 = \theta_{1,1} h_{1,1}(V_1) + \sum_{l=2}^{4} \theta_{1,l} h_{1,l}(V_1)$ (7.56)

donde

$$h_{1,1}(V_1) = V_1$$

$$h_{1,l}(V_1) = \left(V_1 - \xi_{1,l-1}\right)_+^3 - \frac{\left(V_1 - \xi_{1,4}\right)_+^3 \left(\xi_{1,5} - \xi_{1,l-1}\right)}{\xi_{1,5} - \xi_{1,4}} + \frac{\left(V_1 - \xi_{1,5}\right)_+^3 \left(\xi_{1,4} - \xi_{1,l-1}\right)}{\xi_{1,5} - \xi_{1,4}}, \quad l = 2,\ldots,4$$

$$\text{donde } (u)_+ = \begin{cases} u & \text{si } u > 0 \\ 0 & \text{si } u \le 0 \end{cases}$$

es decir, $Z_1(V_1)$ es el spline cúbico restringido con 5 nudos, $\left\{\xi_{1,1} = 0,\ \xi_{1,2} = 23{,}75,\ \xi_{1,3} = 47{,}52,\ \xi_{1,4} = 71{,}2,\ \xi_{1,5} = 95{,}05\right\}$.

Esta expansión lineal ha sido obtenida utilizando software de producción propia confeccionado sobre el Sistema XPLORE, MD&TECH© (Method and Data Technologies).

5.- $Z_1(V_1) = MARS\_V_1 = \sum_{l=1}^{8} \theta_{1,l} FB_{1,l}$ (7.57)

donde

$$\sum_{l=1}^{8}\theta_{1,l}BF_{1,l} = \theta_{1,1}\ máx\left(0,\ V_1-98{,}9\right)+\theta_{1,2}\ máx\left(0,\ 98{,}9-\ V_1\right)+\ \theta_{1,3}\ \ máx\left(0,\ V_1-1{,}5\right)$$
$$+\ \theta_{1,4}\ \ máx\left(0,\ V_1-8{,}36\right)+\theta_{1,5}\ \ máx\left(0,\ V_1-78{,}74\right)+\ \theta_{1,6}\ máx\left(0,\ V_1-73{,}8\right)$$
$$+\ \theta_{1,7}\ \ máx\left(0,\ V_1-79{,}94\right)+\ \theta_{1,8}\ \ máx\left(0,\ V_1-68{,}97\right)$$

Esta expansión lineal se obtuvo utilizando el módulo MARS del Sistema *Salford Predictive Miner* de Salford Systems®.

6.- $Z_1(V_1)=\theta_{11}(RBF\_V_1)$ (7.58)

donde $(RBF\_V1)=\ \exp\left(-V_1^2\right)$, es decir, la función de base radial Gaussiana de centro 0 y radio 1.

Se obtienen de este modo 6 *modelos que ajustaremos por LLR y aquella expansión lineal por funciones base de* V_1 *para la que se consigan mejores estadísticos de ajuste, de poder discriminante, de capacidad clasificatoria, así como cualidades convenientes respecto de los requerimientos de Basilea II y de la política sobre riesgo de crédito de la Entidad Financiera, será la que sustituya a* V_1 en el modelo *logístico lineal híbrido*

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+Z_1\left(V_1\right) \quad (7.59)$$

Contamos por tanto con 6 modelos en el espacio agrandado por la expansión lineal $Z_1(V_1)$:

a) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(V_1\right)^2 \quad (7.60)$$

b) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(W\_V_1\right) \quad (7.61)$$

c) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(WT\_V_1\right) \quad (7.62)$$

d) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+RCS\_V_1 \quad (7.63)$$

e) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+MARS\_V_1 \quad (7.64)$$

f) $$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+RBF\_V_1 \quad (7.65)$$

Los seis modelos, (7.60) a (7.65), con origen en el modelo logístico lineal híbrido (7.59), se estiman mediante la solución del problema general de estimación de modelos logísticos cuya estructura viene dada por una expansión lineal de funciones de base, (6.44). Para encontrar

la solución se utilizó el modulo FIT MODEL del programa JMP de SAS®, en su opción de *Regresión Logística Nominal.*

Tabla 7.24.- Estadísticos de ajuste y poder discriminante para distintas expansiones lineales de V_1 sobre la muestra de entrenamiento. Modelos $M_{inicial} + Z_1(V_1)$.

	V_1^2	W_V_1	WT_V_1	RCS_V_1	$MARS\_V_1$	RBF_V_1
$-LogVer\_Dif$	7533,2371	7530,841	7536,1853	7559,8115	7586,2694	7540,9408
$-LogVer\_Tot$	2242,0854	2244,4815	2239,1372	2215,511	2189,0531	2234,3817
$-LogVer\_\mathrm{Red}$	9775,3225	9775,3225	9775,3225	9775,3225	9775,3225	9775,3225
$R^2(U)$	0,7706	0,7704	0,7709	0,7734	0,7761	0,7714
R^2	0,3374	0,3373	0,3375	0,3384	0,3393	0,3377
$\tilde{R}^2$	0,8154	0,8152	0,8156	0,8177	0,8200	0,8161
$-2L\mathrm{og}Ver$	4484,1708	4488,9630	4478,2744	4431,0220	4378,1062	4468,7634
Error Empírico	0,1225	0,1226	0,1223	0,1210	0,1196	0,1221
AIC	4522,1708	4526,9630	4516,2744	4475,0220	4434,1062	4506,7634
BIC	4683,8227	4688,6149	4677,9263	4662,1979	4672,3301	4668,4153
AUC	0,9784	0,9784	0,9786	0,9780	97,89	0,9786
χ^2	5,03	0,17	10,58			20,3741
$\mathrm{Prob} > \chi^2$	0,0249*	0,6763	<0,0001*	0,6045	0,9162	<0,0001*

En la tabla 7.24 se muestran los estadísticos de bondad de ajuste y discriminación de los 6 modelos considerados. En las columnas de la tabla se muestran los resultados para los seis modelos y por lo que respecta al significado de los elementos mostrados en las filas, distinguiremos en primer lugar entre la información referida al modelo, mostrada en las once primeras y la información referida a la significación en el modelo logístico lineal híbrido de las funciones de base considerada en el modelo analizado, filas 12 y 13.

I.- En las tres primeras filas se muestran las log verosimilitudes negativas, del modelo completo, $-LogVer\_Tot$, del modelo con solo el término intercepto, $-LogVer\_\mathrm{Red}$, y del modelo con los términos de las variables consideradas sin el término intercepto, $-LogVer\_Dif$. Los tres estadísticos son necesarios para calcular los *pseudo coeficientes de determinación del modelo, que se muestran en las filas* 4, 5 y 6.

Los pseudo coeficientes de determinación del modelo, 3.95, 3.103 y 3.106, se formulan en la forma siguiente :

$$\text{Coeficiente de } McFadden,\ R^2(U) = \frac{-LogVer\_Dif}{-LogVer\_\mathrm{Red}} \tag{7.66}$$

Coeficiente de Cox y Snell $R^2 = 1 - \text{antilog}\left(\frac{2LogVer\_Dif}{N}\right)$ (7.67)

Coeficiente de Nagelkerke $\tilde{R}^2 = \frac{1 - \text{antilog}\left(\frac{2LogVer\_Dif}{N}\right)}{1 - \text{antilog}\left(\frac{2LogVer\_\text{Re}d}{N}\right)}$ (7.68)

N es el número de acreditados que forman parte de la muestra de entrenamiento.

II. En las filas 7 y 8 se muestran los estadísticos más utilizado para medir la bondad de ajuste en modelos de respuesta binaria, *-2logver*, (3.91), y el riesgo empírico (3.6), es decir ,

$$\textit{-2logver} = -2\log L\left(Y, \hat{S}(X)\right) = -2\sum_{i=1}^{N}\log\left(\hat{P}\left(Y = y_i / X = x_i\right)\right) \quad (7.69)$$

$$\textit{Error Empírico} = \frac{\textit{-2logver}}{N} \quad (7.70)$$

III. En la fila 9 se muestra el *Criterio de Información de Akaike* AIC, (3.115), y en la fila 10 el *Criterio de Información Bayesiana de Schwarz* BIC , (3.116). Ambos estadísticos miden la bondad de ajuste en modelos de respuesta binaria, pero corrigiendo el optimismo de *-2logver*, lo que conlleva a penalizar la complejidad del modelo y por ello conduce a la selección de modelos más sencillos.

IV. En la fila 11 se muestra la *medida del poder discriminante del modelo*, más popular en credit scoring, el Área Bajo la Curva ROC, AUC, (4.31), figura (4.7).

Una curva que permite visualizar el poder discriminante de un modelo de calificación es la *curva ROC,* que es la que usualmente se utiliza en las aplicaciones prácticas de credit scoring cuando el modelo proporciona la función de probabilidad, como es el caso del modelo HLLM. La Curva ROC consiste en la representación gráfica de los puntos de coordenadas $\{1 - F_0(s), 1 - F_1(s)\}$ para cada puntuación s, lo que formalmente podemos representar por la expresión

$$ROC(u) = 1 - F_0\left[(1 - F_1)^{-1}(u)\right], \quad u \in (0,1) \quad (7.71)$$

donde $F_0(s) = P(S \le s / Y = 0)$ es la tasa de fallos para la puntuación s, llamada *especificidad, y* $F_1(s) = P(S \le s / Y = 1)$, siendo $1 - F_1(s)$ *la sensibilidad.*

En la figura 7.18 se representa gráficamente un ejemplo de curva ROC para la función de calificación de acreditados de modelo de credit scoring.

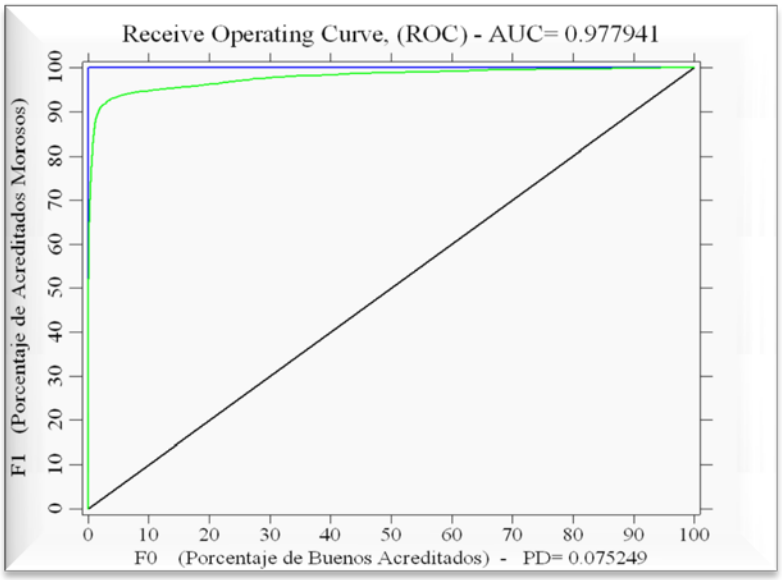


Figura 7.18.- Curva ROC y su estadístico asociado Área Bajo la Curva, AUC, para la función de calificación de acreditados del modelo M_{11_I}.

La medida del poder discriminante asociada a la Curva ROC, es el *Área Bajo la Curva ROC*, AUC. El área bajo la curva ROC óptima, "modelo perfecto", viene determinada por los puntos $(0,0)$,$(0,1)$ y $(1,1)$, (figura 7.18), por lo que vale 1. Un modelo donde cumplidores e incumplidores se confundan totalmente, "modelo aleatorio", tendrá un área igual a 0.5. En una situación real y para un sistema de calificación razonable el "área bajo la curva" será una cantidad entre 0.5 y 1.

El área bajo la curva se obtiene según la siguiente expresión:

$$AUC = \int_{+\infty}^{-\infty} \left[\left(1 - F_1(s)\right) \; d\left(1 - F_0\right)(s) \right] \tag{7.72}$$

y toma valores entre 0 y 1, 0 para la menor desviación y 1 para la mayor, si bien un AUC por debajo de 0.5 no tiene significado.

El AUC es una transformación lineal de la Tasa de Precisión, $AR = 2AUC - 1$, y puede interpretarse como la habilidad media del modelo de credit scoring para clasificar exactamente acreditados o solicitantes de crédito buenos y malos.

Cuando el valor de UAC es de 0,5 la curva ROC es igual a la diagonal lo que significa que el modelo discrimina al azar. Un valor de AUC igual a 1 indica que la curva ROC se

encuentra en la esquina superior izquierda y la discriminación son perfectas. Una curva ROC cercana a la diagonal indica que el modelo es poco informativo. Cuanto más la curva de ROC se encuentre cercana a la esquina superior izquierda, mejor es el poder discriminante del modelo entre cumplidores e incumplidores. O, dicho de otra manera, cuanto mayor sea el área bajo la curva ROC, mejor será el modelo.

V. Por último, en la fila 12 se muestra el *estadístico* χ^2 *de Wald* y en la fila 13 *la significación* del *test de los efectos de Wald,* $\text{Prob} > \chi^2$. *El test de Wald se usa para contrastar la significación estadística de cada coeficiente* β *en el modelo*, es decir, prueba si la variable respuesta tiene una relación de dependencia significativa con cada variable explicativa, el *estadístico de contraste* viene dado por

$$Z=\frac{\hat{B}}{SE} \tag{7.73}$$

Varios autores han identificado problemas con el uso del estadístico de Wald. MENARD (1995) advierte que para grandes coeficientes se infla el error estándar y se reduce el valor chi-cuadrado de Wald. AGRESTI (2002) afirma que la prueba de razón de verosimilitud es más fiable que la prueba de Wald para tamaños de muestra pequeños.

Con el mismo contenido conceptual y estructura, usaremos tablas como la 7.24 para la asignación de las funciones de base para el resto de variables de la componente no lineal. Como puede observarse en las dos últimas filas correspondientes a las columnas RCS_V_1 y MARS_V_1, en la primera de ellas no figura el valor χ^2, la razón es que ambas expansiones están formadas por varias funciones de base con sus correspondientes χ^2 y en la segunda el valor $\text{Prob} > \chi^2$ que figura se refiere al nivel de significación más bajo de los presentados por todas la funciones de base que integran la expansión. Los valores correspondientes a ambos estadísticos particularizados para cada función de base de las dos expansiones en cuestión se muestran en las tablas 7.25 y 7.26.

Tabla 7.25.- Funciones de base para el splin cúbico restringido RCS_V_1

	χ^2	$\text{Prob} > \chi^2$
$h_1\_V_1$	0,81	0,3673
$h_2\_V_1$	37,69	<0,0001*
$h_3\_V_1$	15,21	<0,0001*
$h_4\_V_1$	0,2700	0,6050

Tabla 7.26.- Funciones de base para el Splin de Regresión Adaptativo Univariante MARS_V_1.

	χ^2	Prob > χ^2
$BF_1\_V_1$	0,42	0,5169
$BF_2\_V_1$	7,81	0,0052*
$BF_3\_V_1$	4,08	0,0434*
$BF_4\_V_1$	3,88	0,0487*
$BF_5\_V_1$	0,22	0,6378
$BF_6\_V_1$	0,01	0,9162
$BF_7\_V_1$	0,08	0,7771
$BF_8\_V_1$	0,65	0,4210

De la observación de las tablas 7.25 y 7.26 se pueden extraer la siguientes conclusiones:

La función de base WT_V_1 resulta significativa a nivel $\alpha < 0.0001$ en el modelo logístico lineal (7.62), lo mismo ocurre con RBF_V_1 en el modelo (7.65). Con un nivel de significación más alto, $\alpha = 0.0249$, V_1^2 resulta significativa en el modelo (7.60). Por lo que respecta a las expansiones lineales de funciones base $Z_1(V_1) = SCR\_V_1$ y $Z_1(V_1) = MARS\_V_1$, para la primera, dos de las cuatro funciones de base que la integran resultan significativas en el modelo (7.63), $h_{12}(V_1)$ y $h_{13}(V_1)$, ambas con nivel $\alpha < 0.0001$, y para la segunda, tres de las ocho funciones de base que la integran son significativas en el modelo (7.64), $BF_2\_V_1$, $BF_3\_V_1$, $BF_4\_V_1$, con niveles de significación $\alpha = 0,0052$ d, $\alpha = 0,0434$ y $\alpha = 0,0487$ respectivamente.

Antes de pasar a explorar que expansión es la más adecuada para especificar la no linealidad de V_1 procedemos a replantear los modelos (7.63) y (7.64) con las expansiones formadas sólo por las funciones de base significativas en el análisis anterior. Es decir, $SCR\_V_1\_2 = \sum_{l=2}^{3} \theta_{1,l} h_{1,l}(V_1)$ y $MARS\_V_1\_2 = \sum_{l=2}^{4} \theta_{1,l} BF_{1,l}(V_1)$. Ajustamos ambos modelos por LLR. Los resultados se muestran en la Tabla 7.27.

Tabla 7.27.- Estadísticos de ajuste y poder discriminante de los 6 modelos considerados. Los modelos (7.63) y (7.64) se han modificado con las expansiones $SCR\_V_1\_2$ y $MARS\_V_1\_2$ respectivamente.

	V_1^2	W_V_1	WT_V_1	$RCS\_V_1\_2$	$MARS\_V_1\_2$	RBF_V_1
$-LogVer\_Dif$	7533,2371	7530,841	7536,1853	7559,2651	7582,1967	7540,9408
$-LogVer\_Tot$	2242,0854	2244,4815	2239,1372	2216,0574	2193,1258	2234,3817
$-LogVer\_\mathrm{Red}$	9775,3225	9775,3225	9775,3225	9775,3225	9775,3225	9775,3225
$R^2(U)$	0,7706	0,7704	0,7709	0,7733	0,7756	0,7714
R^2	0,3374	0,3373	0,3375	0,3383	0,3392	0,3377
$\tilde{R}^2$	0,8154	0,8152	0,8156	0,8177	0,8197	0,8161
$-2L\,\mathrm{og}Ver$	4484,1708	4488,9630	4478,2744	4432,1148	4386,2516	4468,7634
Error Empírico	0,1225	0,1226	0,1223	0,1211	0,1198	0,1221
AIC	4522,1708	4526,9630	4516,2744	4472,1148	4424,2516	4506,7634
BIC	4683,8227	4688,6149	4677,9263	4642,2747	4585,9035	4668,4153
AUC	0,9784	0,9784	0,9786	0,9780	0,9789	0,9786
χ^2	5,03	0,17	10,58			20,3741
Prob > χ^2	0,0249*	0,6763	<0,0001*	<0,0001*	0,0008*	<0,0001*

Los estadísticos mostrados en la tabla 7.27 conducen a las siguientes conclusiones:

- De los 4 modelos estimados por LLR para los que se ha conseguido la linealidad total, (7.62), WT_V_1, (7.63), $RCS\_V_1\_2$, (7.64), $MARS\_V_1\_2$, (7.65), RBF_V_1, destaca por los indicadores de bondad de ajuste aquel al que se ha añadido la expansión $MARS\_V_1\_2$, con un coeficiente de determinación generalizado de Nagelkerke de 0,8197, que viene a indicar, más o menos, que el 81,97% de la variabilidad del estado de default viene explicada por las 18 variables del modelo inicial y las tres funciones de base de la expansión $MARS\_V_1\_2$, $BF_2\_V_1$, $BF_3\_V_1$, $BF_4\_V_1$. Este es el modelo que posee el menor riesgo empírico, y ,más importante aún, por cuanto tienen en cuenta el número de variables del modelo, los coeficientes AIC y BIC menores. *Pero este hecho ha de ser tratado con mucho cuidado puesto que el sobreajuste es un inconveniente muy conocido de MARS.*

- Con una situación muy similar para los indicadores de ajuste se sitúa en segundo lugar el modelo inicial con las dos funciones de base $h_{12}(V_1)$ y $h_{13}(V_1)$ del spline cúbico restringido $RCS\_V_1\_2$, por delante de los modelos correspondientes al modelo inicial con las funciones de base RBF_V_1 y de WT_V_1 respectivamente y por este orden.

- Por lo que respecta al poder discriminante sobre los acreditados de entrenamiento, también el modelo correspondiente a $MARS\_V_1\_2$ presenta la mayor *área bajo la Curva ROC*, AUC, lo que indica que es el que mejor discrimina el default del no default sobre los acreditados sobre los que se entrenó dicho modelo.

- En cuanto al poder discriminante para los otros tres modelos, se invierte la situación respecto a lo ocurrido con los indicadores de la bondad de ajuste: los modelos correspondientes a la función de base radial Gaussiana, RBF_V_1, y a la función de base de los pesos de la evidencia observados sobre una partición recursiva, WT_V_1, poseen AUC ligeramente mayor que el modelo correspondiente a $RCS\_V_1\_2$.

- Por otro lado, Basilea II requiere *modelos "tan poco complejos y fáciles de interpretar como sea posible"*. En ese sentido es evidente que WT_V_1 y $MARS\_V_1\_2$ poseen una clara ventaja sobre los splines cúbicos restringidos y sobre las funciones de base radial Gaussiana.

- No es difícil explicarle a un cliente, sobre una *tarjeta de puntuación,* que el hecho de que el valor de la variable V_1 observado sobre él se encuentre dentro de un determinado tramo de dicha variable donde la ventaja de no default sobre default es, por ejemplo, 1:10 le resta puntuación y por tanto posibilidades de concesión de crédito, pues su probabilidad de default es alta, tanto como lo sea la contribución de la variable al modelo que expresa formalmente la relación entre el estado de default y las variables relevantes del riesgo de crédito.

- Otro tanto ocurre con funciones las de base "bisagra" de los splines de regresión adaptativos univariantes tales como la que integran la expansión lineal

$$MARS\_V_1 = \sum_{l=2}^{4} \theta_{1l} BF_{1l} = \; 0{,}945315\; máx\left(0,\; 98{,}9 - \; V_1\right) + \; 0{,}766475\; máx\left(0,\; V_1 - 1{,}5\right)$$
$$+ \; 0{,}194814\; máx\left(0,\; V_1 - 8{,}36\right)$$

donde si el valor observado para un acreditado, (*mínima tasa de saldo pendiente frente al importe formalizado con garantía personal a fecha de visión*), es menor del 98,9% la contribución de la función de base BF1 al modelo es la diferencia ponderada por 0,945315; si además está por encima del 1,5% a la contribución anterior se le suma la proporcionada por BF2, la diferencia a 1,5 ponderada por

0,766475, y, si además de estar por encima del 1,5%, está por encima del 8,36% se le suma la contribución de BF3, es decir, la diferencia entre valor observado y 8,36 ponderada por 0,194814.

- Una de las claves del punto anterior estriba en el significado que se le quiera dar a la expresión "*tan poco complejos y fáciles de interpretar como sea posible*".

Tal como ya procedimos en la exploración de la linealidad usando la regresión logística lineal con muestreo bootstraap, LLR_Bag y BLLR_Bag, apartado 7.3.2.2, antes de considerar adecuada una expansión lineal por funciones de base de una variable para conseguir la significación en el modelo logístico lineal procedemos a la evaluación del modelo resultante, de este modo podremos valorar si el modelo ajustado es un modelo válido y, por tanto, lo es también la expansión lineal por funciones de base considerada, más allá de que presente un ajuste adecuado a los datos.

Una vez ajustados los datos, se obtienen los valores calculados de la función de calificación de acreditados y se estima el error de predicción sobre la muestra de validación. Será preferible, en lo que respecta a este apartado, el modelo con menor error de predicción sobre la muestra de validación.

Tal como hemos expuesto en el Capítulo 3, subsección 3.4.3, de esta Tesis Doctoral, los test basados en $-2\log ver$ *conducen con frecuencia a rechazar modelos adecuados aceptando algunos que resultan menos parsimoniosos de lo que debieran.* Por lo que será necesario utilizar aquí los criterios de Información de Akaike, AIC, y de Información Bayesiano de Schwarz, BIC, con el fin de que el término de penalización corrija la complejidad del modelo y se pueda evitar el sobreajuste.

Por otro lado, dado que las funciones de base se seleccionan con la intención de construir modelos de predicción y clasificación es evidente que AIC, BIC , AUC y el porcentaje de clasificación incorrecta deben evaluarse sobre la muestra de validación. Esta evaluación puede contribuir eficazmente a conocer la solidez de los indicadores calculados inicialmente sobre la muestra de entrenamiento y, por tanto, pueden constituir una herramienta de inestimable ayuda en la selección de las funciones de base más prometedoras.

Tabla 7.28.- Validación del ajuste y poder discriminante para los 4 modelos seleccionados (columnas de la tabla). (N=18.303).

	WT_V_1	$RCS\_V_1\_2$	$MARS\_V_1\_2$	RBF_V_1
$-LogVer\_Tot$	1074,3	1067,3	1060	1072,4
$-2LogVer$	2148,6000	2134,6000	2120,0000	2144,8000
Error de Validación	0,1174	0,1166	0,1158	0,1172
AIC	2186,6000	2174,6000	2162,0000	2182,8000
BIC	2335,0816	2330,8964	2326,1112	2331,2816
AUC	0,9785	0,9787	0,9791	0,9783
% Clasificación Errónea	1,66	1,60	1,67	1,66

Como se muestra en la tabla 7.28, para la muestra de validación el menor error empírico, así como los menores AIC y BIC, los presenta el modelo logístico lineal inicial con la incorporación de la expansión lineal $MARS\_V_1\_2$, seguido muy de cerca por el modelo inicial con la incorporación de la expansión $RCS\_V_1\_2$, además ambos modelos, y en el mismo orden anterior, presentan las áreas bajo la curva más altas, por lo que son los de mejor ajuste y poder discriminante. Respecto del porcentaje de clasificados incorrectos, el menor porcentaje, 1,60%, lo presenta en la muestra de validación la incorporación al modelo inicial de la funciones de base del spline cúbico restringido, $RCS\_V_1\_2$, linealmente significativas, este porcentaje es igual, 1,66%, para los modelos con las incorporaciones WT_V_1 y RBF_V_1, y el mayor porcentaje lo presenta $MARS\_V_1\_2$, 1,67%.

Los modelos de mayor AIC y BIC son los que incorporan respectivamente, y por este orden, a WT_V_1 y a RBF_V_1. Sin embargo la incorporación de WT_V_1 proporciona un AUC más alto, tabla 7.29, y, por tanto, mejor poder discriminante que RBF_V_1 y muy próximo al proporcionado por el spline cúbico restringido $RCS\_V_1\_2$.

Tabla 7.29.- Área bajo la curva ROC – AUC. Muestra de **Validación**. N=18.303.

				Intervalo de confianza asintótico al 95%	
	AUC	Error típ.[a]	Sig. asintótica[b]	Límite inferior	Límite superior
WT_V_1	0,9785	0,0026	0,0000	0,9734	0,9835
$RCS\_V_1\_2$	0,9787	0,0026	0,0000	0,9737	0,9837
$MARS\_V_1\_2$	0,9791	0,0026	0,0000	0,9740	0,9841
RBF_V_1	0,9783	0,0026	0,0000	0,9732	0,9834

a. Bajo el supuesto no paramétrico b. Hipótesis nula: área verdadera = 0.5.

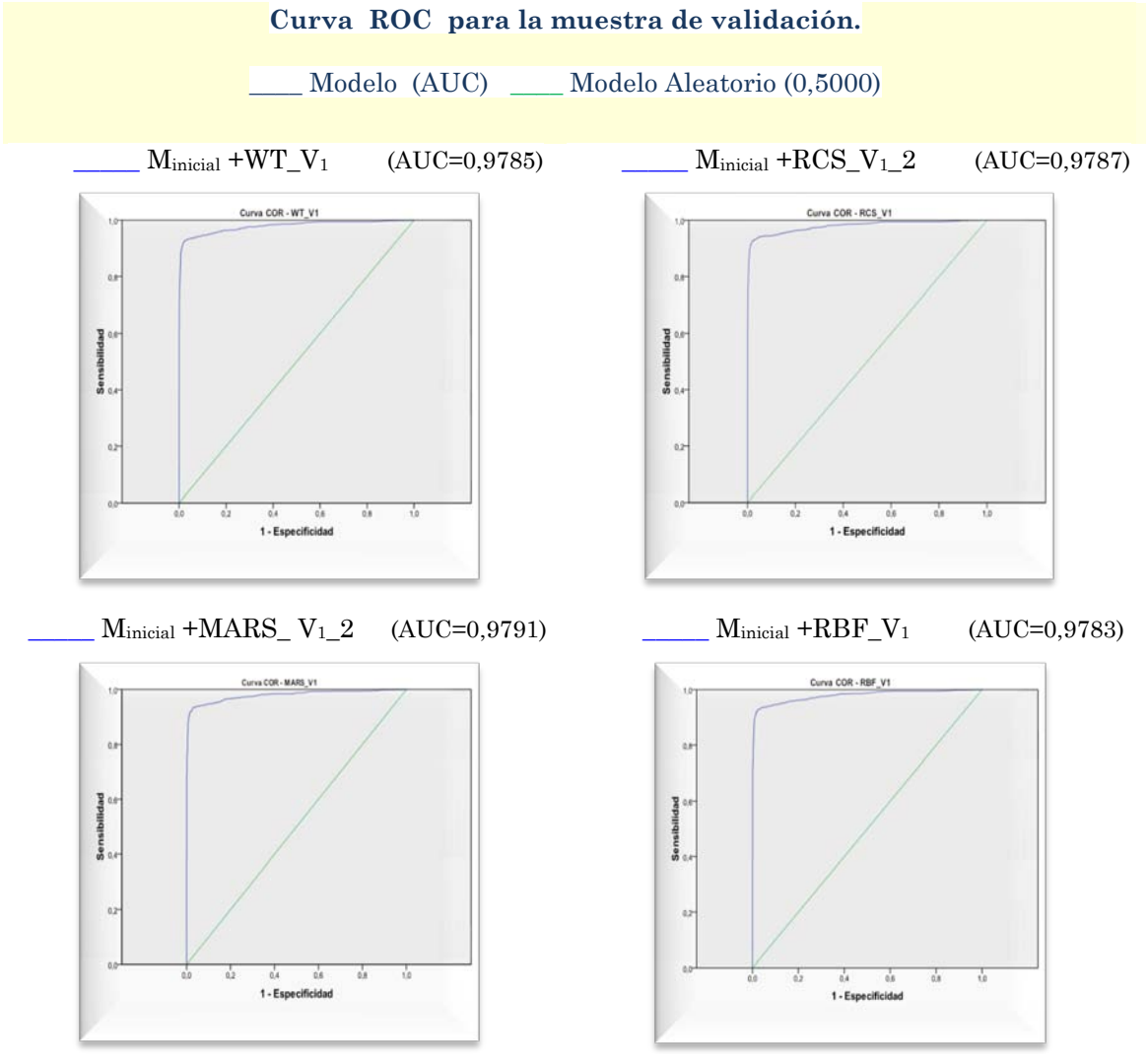


Figura 7.19.- Representación gráfica de la curva ROC para el modelo inicial al incorporarle cada una de las expansiones analizadas, WT_ V_1, RCS_ V_1_2, MARS_ V_1_2 y RBF_ V_1. Como puede observarse a simple vista las curvas ROC son prácticamente iguales, como corresponde al hecho de que lo sean las áreas bajo la curva.

Todo parece indicar, tablas 7.28 y 7.29 y figura 7.19, que la pequeña superioridad de MARS y de los splines cúbicos restringidos en ajuste y poder discriminante no compensa su mayor dificultad de interpretación frente a los pesos de la evidencia en particiones recursivas.

Ante la situación planteada recurrimos al criterio de sencillez del modelo y facilidad interpretativa, en línea con los requerimientos de Basilea II, por lo *que especificamos en nuestro modelo la no linealidad de V_1 a través de la función de base WT_V_1*, que además por su ligero peor comportamiento ante el ajuste a los datos de entrenamiento puede contribuir a prevenir el sobreajuste, por lo que el modelo, que notaremos M_{19_1}, contempla las

siguientes variables {U1,…,U16}, WT_U17, W_U19, WT_ V1, y adopta la siguiente estructura funcional:

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(WT\_V_1\right) \qquad (7.74)$$

Asignación de las funciones de base a la variable V2 de la componente no lineal.

Nuestra siguiente tarea consiste en especificar la no linealidad de V_2 con la misma metodología seguida con V_1, si bien con una diferencia importante con respecto a la situación anterior, V_2 es una variable discreta, por lo que no tiene sentido expandirla o aproximarla por funciones continuas, es decir prescindimos en este caso de los Splines Cúbicos Restringidos, de las funciones bisagra MARS y de las funciones de base radiales Gaussianas. Por tanto, las opciones son este caso V_2^2, W_V_2, y WT_V_2.

Al ajustar por regresión logística lineal el modelo (7.74) con la introducción de las funciones de base V_2^2, W_V_2, y WT_V_2 observamos que en los tres casos la variable U_5 pierde su significación lineal en el modelo. Dado que los expertos en riesgo de crédito consideran que la variable V2, *número de meses, en los últimos 6, que el cliente ha percibido ingresos recurrentes,* debe introducirse en el modelo si es posible y que la variable U_5 forma parte de la *visión de activo con garantía personal*, al igual que V_1 y ésta ya está en el modelo, y además este concepto está recogido en la variable U_{19}, también ya en el modelo, optamos por prescindir del término $\beta_5 U_5$, de donde los tres modelos a estimar vienen dados por

a)
$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(WT\_V_1\right)+\theta_{2,1}\left(V_2\right)^2 \qquad (7.75)$$

b)
$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(WT\_V_1\right)+\theta_{2,1}\left(W\_V_2\right) \qquad (7.76)$$

c)
$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{1,1}\left(WT\_V_1\right)+\theta_{2,1}\left(WT\_V_2\right) \qquad (7.77)$$

De los resultados mostrados en tabla 7.30 se deduce que, en principio, la especificación más adecuada para la no linealidad de V_2 viene dada por WT_V_2, pues ninguna de las otras dos

resulta significativa en el modelo logístico lineal resultado de incorporarlas al modelo (7.74).

El comportamiento de la incorporación de la función de base WT_V_2 respecto del ajuste y el poder discriminante en la muestra de entrenamiento es bastante bueno, por cuanto el coeficiente de determinación de Nagelkerke supera a 0,81 y el área bajo la curva ROC alcanza casi el valor 0,98.

Tabla 7.30.- Estadísticos de ajuste y poder discriminante de los modelos (7.75), (7.76) y (7.77). Distintas funciones de base para la no linealidad de V_2.

	V_2^2	W_V_2	WT_V_2
$-LogVer\_Dif$	7534,9662	7535,0704	7538,4867
$-LogVer\_Tot$	2240,3563	2240,2521	2236,8358
$-LogVer\_\mathrm{Re}d$	9775,3225	9775,3225	9775,3225
$R^2(U)$	0,7708	0,7708	0,7712
R^2	0,3375	0,3375	0,3376
$\tilde{R}^2$	0,8155	0,8155	0,8158
$-2L\mathrm{og}Ver$	4480,7126	4480,5042	4473,6716
Error Empírico	0,1224	0,1224	0,1222
AIC	4518,7126	4518,5042	4511,6716
BIC	4680,3645	4680,1561	4673,3235
AUC	0,9787	0,9788	0,9787
χ^2	1,08	1,30	7,93
$\mathrm{Prob} > \chi^2$	0,2982	0,9788	0,0049*

Por lo que respecta al comportamiento respecto de la muestra de validación se tiene:

Tabla 7.31.- Validación del ajuste y poder discriminante para el modelo seleccionado. N=18.303.

	WT_V_2
$-LogVer\_Tot$	1070,5
$-2L\mathrm{og}Ver$	2141,0000
Error de Validación	0,1170
AIC	2179,0000
BIC	2327,4816
AUC	0,9785
% Clasificación Errónea	1,66

Tabla 7.32.- Validación del área bajo la curva ROC – AUC. (N=18.303).

	AUC	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95% Límite inferior	Límite superior
WT_V_2	0,9788	0,0025	0,0000	0,9739	0,9838

. Bajo el supuesto no paramétrico

. Hipótesis nula: área verdadera = 0,5

Curva ROC para la muestra de validación.

____ Modelo (AUC) ____ Modelo Aleatorio (0,5000)

_____ M_{19_1} + WT_V2 (AUC=0,9788)

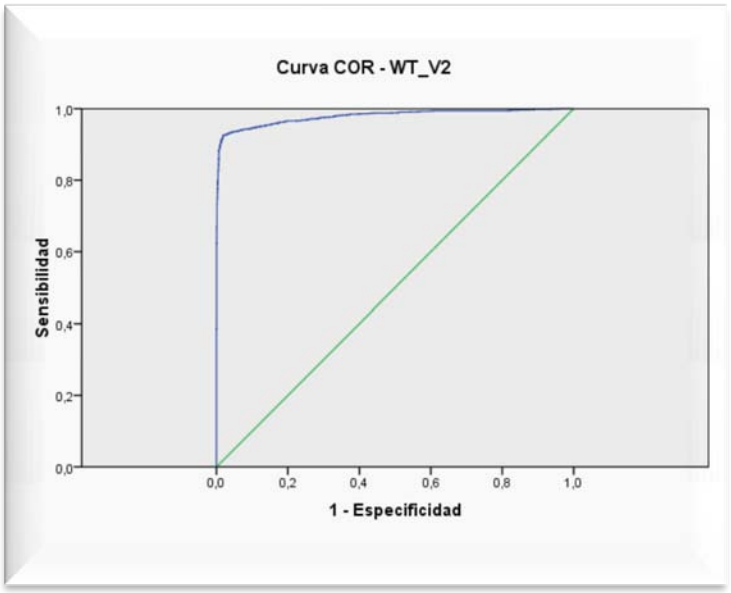


Figura 7.20.- Representación gráfica de la curva ROC para el modelo M_{19_1} al incorporarle la expansión WT_V2.

En la muestra de validación, el área bajo la curva ROC se acerca casi a 0,98, lo que indica un buen rendimiento discriminante de este modelo, y el porcentaje de clasificados incorrectos se sitúan por debajo del 1,7%. Los resultados nos conducen a *especificar en nuestro modelo la no linealidad de* V_2 *a través de la función de base* WT_V_2, por lo que el modelo, que notaremos M_{19_2}, contempla las siguientes variables {U_1,..., U_4, U_6,...,U_{16}}, WT_U_{17}, W_U_{19}, WT_V_1, y WT_V_2 y adopta la siguiente estructura funcional:

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\sum_{r=1}^{2}\theta_{r,1}\left(WT\_V_r\right) \tag{7.78}$$

Este mismo análisis se realizó para las restantes cinco variables no lineales resultando el modelo final $\boldsymbol{M}_{21\_I}$ que contempla las 21 "variables" siguientes:

$$\begin{aligned}&\text{U1, U2, U3, U4, U6, U7, U8, U9, U10, U11, U12, U13, U14,}\\&\text{U15, U16, WT\_U17, W\_U19, WT\_V1, WT\_V2, } BF1\_V4,\\&\text{WT\_V6 .}\end{aligned} \tag{7.79}$$

que se corresponden, respectivamente, con las siguientes variables originales

$$\begin{aligned}&\text{X53, X48, X32T, WR\_X55, X8, WR\_X46, X24, X57, X63,}\\&\text{X25, X64, WR\_X49, WR\_X56, X42, X6, WT\_X58, W\_X15,}\\&\text{WT\_X26, WT\_X35, BF1\_X3, WT\_X19 .}\end{aligned} \tag{7.80}$$

El modelo $\boldsymbol{M}_{21\_I}$ viene dado por la siguiente estructura funcional

$$\begin{aligned}\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)\\ +\theta_{1,1}WT\_V_1+\theta_{2,1}WT\_V_2+\theta_{4,1}BF_1\_V_4+\theta_{6,1}WT\_V_6\end{aligned} \tag{7.81}$$

7.4.2.4.3 Selección de Modelos Alternativos

El método supervisado de selección de funciones de base que hemos propuesto posee sin duda alguna notables ventajas frente a los modelos adaptativos automáticos, sobre todo porque la supervisión se orienta fundamentalmente a la importancia de las variables para explicar el estado de default y su facilidad interpretativa, ambos requerimientos importantes de Basilea II. Así, por ejemplo, la razón para no seleccionar, en algunos casos, los splines cúbicos restringidos a pesar de que su comportamiento en la bondad de ajuste y el poder de discriminación es mejor que la alternativa de intervalos óptimos y particiones recursivas binarias es su más difícil interpretación, lo que choca frontalmente con los requerimientos de Basilea II.

Es necesario, y posible, establecer un equilibrio entre los requerimientos de Basilea II y las cualidades estadísticas de los modelos de credit scoring. De hecho, si se cuida con rigor la selección supervisada de las funciones de base, fase en la que la colaboración entre estadísticos expertos en construcción de modelos y expertos en riesgo de crédito se hace imprescindible, parece razonable pensar que la ganancia en facilidad interpretativa será ventajosa frente al coste estadístico que implica el no seleccionar aquellas componentes con mejores indicadores tanto de ajuste, como de discriminación y clasificación, lo que no evita

que el método pueda conllevar importantes dosis de subjetividad. Es por ello que se hace necesario conocer como incide esa subjetividad sobre la selección de las funciones de base finalmente seleccionadas. Una forma de aproximarnos al grado de esa incidencia consiste en considerar varios modelos logísticos lineales alternativos, uno con las funciones de base seleccionadas y otros con funciones de base no seleccionadas por razones, sobre todo, de interpretabilidad.

Los modelos alternativos a M_{21_I} que analizaremos son:

II. M_{21_II}, en el que se contemplan los splines cúbicos en el caso de su superioridad en bondad de ajuste y/o poder discriminante, cuya estructura formal viene dada por

$$\begin{aligned}\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{2,1}\left(WT\_V_2\right)\\ +\ \theta_{6,1}\left(WT\_V_6\right)+RCS\_V1\_2+RCS\_V4\_2\end{aligned} \tag{7.82}$$

donde las expansiones lineales de funciones base de V_1 y V_4 vienen dadas por

$$RCS\_V_1\_2=\sum_{l=2}^{3}\theta_{1l}h_{1l}\left(V_1\right)\ \text{ y }\ RCS\_V_4\_2=\sum_{l=3}^{4}\theta_{4l}h_{4l}\left(V_4\right)$$

III. M_{21_III}, este modelo logístico se construye con las mismas transformaciones de las variables explicativas que M_{21_II} con la excepción de las expansiones lineales $RCS\_V_1\_2$ y $RCS\_V_4\_2$, que son sustituidas por las expansiones lineales $MARS\_V_1\_2$ y $MARS\_V_4$. Su estructura formal es la siguiente:

$$\begin{aligned}\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{2,1}\left(WT\_V_2\right)\\ +\ \theta_{6,1}\left(WT\_V_6\right)+MARS\_V1\_2+MARS\_V4\end{aligned} \tag{7.83}$$

donde

$$MARS\_V_1\_2=\sum_{l=2}^{4}\theta_{1l}\left(BF_{1l}\_V_1\right)\ \text{ y }\ MARS\_V_4=\left(BF_1\_V_4\right)$$

IV. M_{21_IV}, en este caso las expansiones lineales por funciones base de V_1 y V_4 que se incorporan son RBF_V_1 y RBF_V_4, por lo que su estructura formal es la siguiente:

$$\begin{aligned}\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{\substack{r=1\\ r\neq 5}}^{16}\beta_r U_r+\beta_{17}\left(WT\_U_{17}\right)+\beta_{19}\left(W\_U_{19}\right)+\theta_{2,1}\left(WT\_V_2\right)\\ +\ \theta_{6,1}\left(WT\_V_6\right)+\theta_{1,1}\left(RBF\_V_1\right)+\theta_{4,1}\left(RBF\_V_4\right)\end{aligned} \tag{7.84}$$

Los cuatro modelos logísticos lineales híbridos considerados, M_{21_I}, M_{21_II}, M_{21_III} y M_{21_IV}, son modelos logísticos lineales, donde las variables no lineales se expanden por combinaciones lineales de funciones de base. Ninguno de ellos requiere, en principio, regularización y, por tanto, pueden ser estimados resolviendo el problema general de estimación de los modelos logísticos lineales no regularizados. Los resultados del ajuste se muestran en la tabla 7.33, en la columna correspondiente al modelo.

Tabla 7.33.- Modelos logísticos lineales híbridos, HLLM, alternativos.

	M_{21_I}	M_{21_II}	M_{21_III}	M_{21_IV}
$-LogVer\_Dif$	7590,0279	7621,7765	7633,1363	7592,3773
$-LogVer\_Tot$	2185,2946	2153,546	2142,1862	2182,9452
$-LogVer\_Red$	9775,3225	9775,3225	9775,3225	9775,3225
$R^2(U)$	0,7764	0,7797	0,7809	0,7767
R^2	0,3394	0,3406	0,3410	0,3395
$\tilde{R}^2$	0,8203	0,8231	0,8241	0,8205
$-2LogVer$	4370,5892	4307,0920	4284,3724	4365,8904
Error Empírico	0,1194	0,1177	0,1170	0,1193
AIC	4412,5892	4353,0920	4330,3724	4407,8904
BIC	4591,2571	4548,7759	4526,0563	4586,5583
AUC	0,9801	0,9797	0,9803	0,9799
% Clasificación Incorrecta	1,82%	1,82%	1,82%	1,81%

Tabla 7.34.- Test de bondad de ajuste de Hosmer y Lemeshow.

Modelo	Chi-cuadrado	GL	Pr > ChiSq
M_{21_I}	108,3144	8	<.0001
M_{21_II}	138,6106	8	<.0001
M_{21_III}	117,6734	8	<.0001
M_{21_IV}	120,3484	8	<.0001

El test de bondad de ajuste de Hosmer y Lemeshow para todos los modelos, $\Pr > \text{ChiSq} < 0.001$, indica un buen ajuste para todos ellos.

Tabla 7.35.- Tabla de Clasificación (para un punto de corte de 0,50).

Modelo	Correcto		Incorrecto		Porcentajes				
	Default	No Default	Default	No Default	Correcto	Sensi-bilidad	Especi-ficidad	Falso POS	Falso NEG
M_{21_I}	2347	33588	264	408	98,16	85.2	99.2	10.1	1.2
M_{21_II}	2344	33588	264	411	98,16	85,1	99,2	10,1	1,2
M_{21_III}	2353	33582	270	402	98,19	85,4	99,2	10,3	1,2
M_{21_IV}	2343	33600	252	412	98,19	85.0	99.3	9.7	1.2

Tabla 7.36.- Probabilidad Pronosticada, (área bajo la curva ROC).

Modelo	UAC	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95%	
				Límite inferior	Límite superior
M_{21_I}	0,9801	0,0016	0,0000	0,9767	0,9833
M_{21_II}	0,9797	0,0017	0,0000	0,9764	0,9829
M_{21_III}	0,9803	0,0016	0,0000	0,9771	0,9835
M_{21_IV}	0,9799	0,0017	0,0000	0,9767	0,9832

a. Bajo el supuesto no paramétrico.

b. Hipótesis nula: área verdadera = 0,05.

Desde el punto de vista estadístico los 4 modelos presentan un buen ajuste, coeficientes de Nagelkerke superiores a 0.80, un alto poder discriminante y explicativo, AUC muy próximos a 0,98, tablas 7.33, 7.34 y 7.36, y alta eficacia como clasificadores, porcentajes de clasificados correctamente superiores a 98 %, tabla 7.35. Es decir, *los cuatro modelos logísticos ajustados parecen ser modelos correctos para explicar el estado de default de los acreditados.*

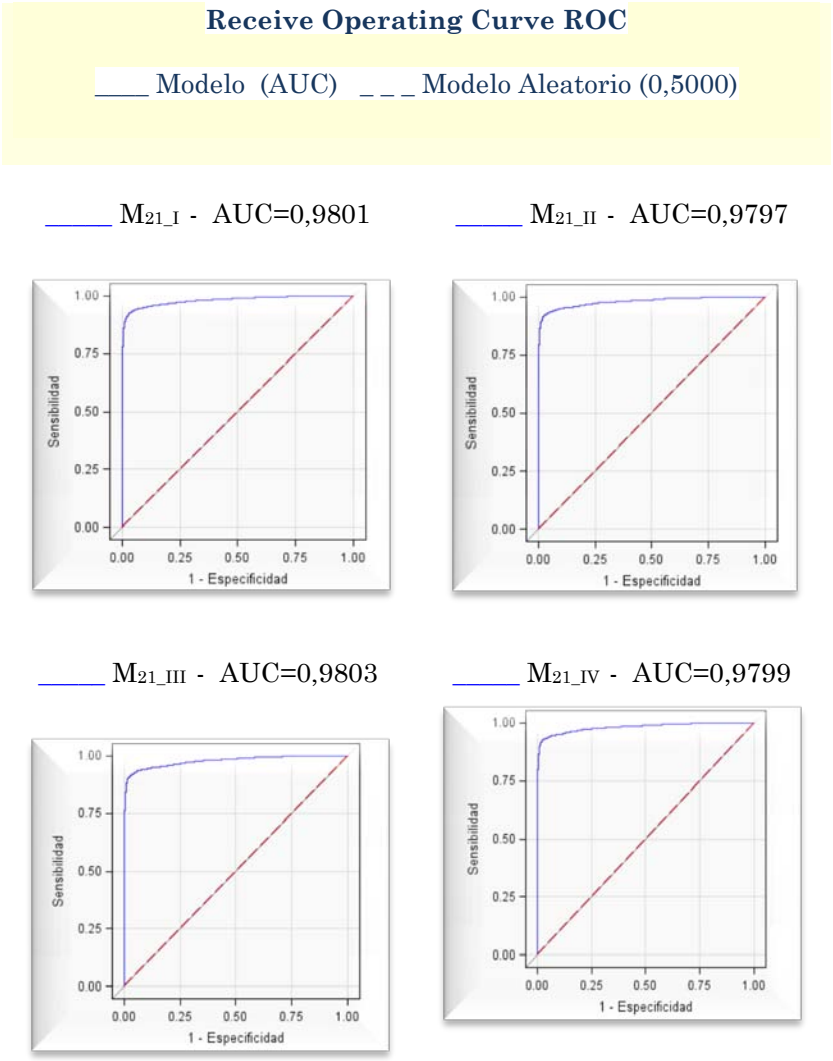


Figura 7.21.-Curva ROC y área bajo la curva, AUC, de los 4 modelos considerados.

Con la metodología seguida, el modelo LPLM, (7.37)*, se ha derrumbado como un castillo de naipes, dando lugar al modelo HLLM,* (7.81), (o alternativamente a los modelos (7.82), (7.83). y (7.84)*, herramienta de predicción y clasificación capaz de explicar la relación de dependencia del estado de default con las variables de riesgo de crédito más relevantes, estimar la función de probabilidad de default, calificar a los acreditados, explicar la puntuación que se les otorga y clasificar una nueva solicitud de crédito con suficiente rigor; todo ello dentro del más estricto cumplimiento de los requerimientos de Basilea II, que en lo que respecta a los modelos estadísticos se mantiene en los acuerdos de Basilea III.*

7.5 EVALUACIÓN, GENERALIZACIÓN Y SELECCIÓN DEL MODELO.

7.5.1 Introducción.

Antes de proceder a la selección del modelo final y fijar su error de predicción o generalización esperado es necesario evaluar los distintos modelos alternativos en orden a detectar cuál de ellos reúne las cualidades idóneas. La *evaluación del modelo* supone valorar si el modelo ajustado en la etapa de estimación es un modelo válido, más allá de que presente un ajuste adecuado a los datos. La evaluación o diagnóstico del modelo se refiere a la adecuación de los aspectos implicados en la etapa de especificación.

En este sentido, aparte de estimar el error empírico del modelo sobre la muestra de validación, *error de evaluación*, se han de evaluar posibles errores de especificación de la componente sistemática, de la distribución de probabilidad de la componente aleatoria y de la relación asumida entre ambas componentes del modelo en la fase de especificación.

Es necesario también valorar las posibles pérdidas de eficacia estadística al priorizar la facilidad explicativa de las transformaciones de tramado de variables ya sea por intervalos óptimos automáticos o por particiones recursivas binarias sobre la "mayor eficacia" de los splines cúbicos restringidos, de las funciones bisagra o de las funciones de base radial Gaussiana.

Como resultado del proceso anterior se obtiene una aproximación al modelo más prometedor para nuestros objetivos. Nos referimos a una aproximación por cuanto en el proceso seguido hasta ahora no hemos analizado a fondo la posible complejidad del modelo, tarea que acometeremos más adelante para obtener el modelo "óptimo" para nuestro triple objetivo: *estimar la probabilidad de default*, *estimar la función de calificación de acreditados* y *estimar el clasificador de nuevas solicitudes de créditos*, todos ellos desde la estricta observación de los requerimientos de Basilea II.

7.5.2- Evaluación de los modelos HLLM Preseleccionados.

En este apartado *evaluaremos los modelos a través del Criterio de Información Bayesiano,* BIC, puesto que seleccionar el modelo con mínimo BIC es equivalente a elegir el modelo con mayor probabilidad a posteriori, y *analizaremos el poder discriminante de los modelos a través el área bajo la curva* ROC, AUC. Ambos estadísticos, que se obtendrán sobre la muestra de validación compuesta por 18.303 acreditados, *son muy adecuados para seleccionar el mejor modelo final* (HASTIE et al. 2009).

Tabla 7.37.- Criterio de Información Bayesiana, BIC, de M_{21_I}, M_{21_II}, M_{21_III}, M_{21_IV}, sobre la muestra de validación.

	–LogVer _Tot	BIC
M_{21_I}	1059,9	2325,9112
M_{21_II}	1053,7	2333,1402
M_{21_III}	1048,1	2323,1409
M_{21_IV}	1049,8	2305,7112

Tabla 7.38.- Poder Discriminante de M_{21_I}, M_{21_II}, M_{21_III}, M_{21_IV}, sobre la muestra de validación, Área bajo la curva ROC, AUC.

				Intervalo de confianza asintótico al 95%	
Modelo	AUC	Error típ.[a]	Sig. asintótica[b]	Límite inferior	Límite superior
M_{21_I}	0,9792	0,0026	0,0000	0,9742	0,9843
M_{21_II}	0,9793	0,0025	0,0000	0,9743	0,9843
M_{21_III}	0,9795	0,0026	0,0000	0,9746	0,9846
M_{21_IV}	0,9794	0,0026	0,0000	0,9744	0,9845

a. Bajo el supuesto no paramétrico. b. Hipótesis nula: área verdadera = 0,05.

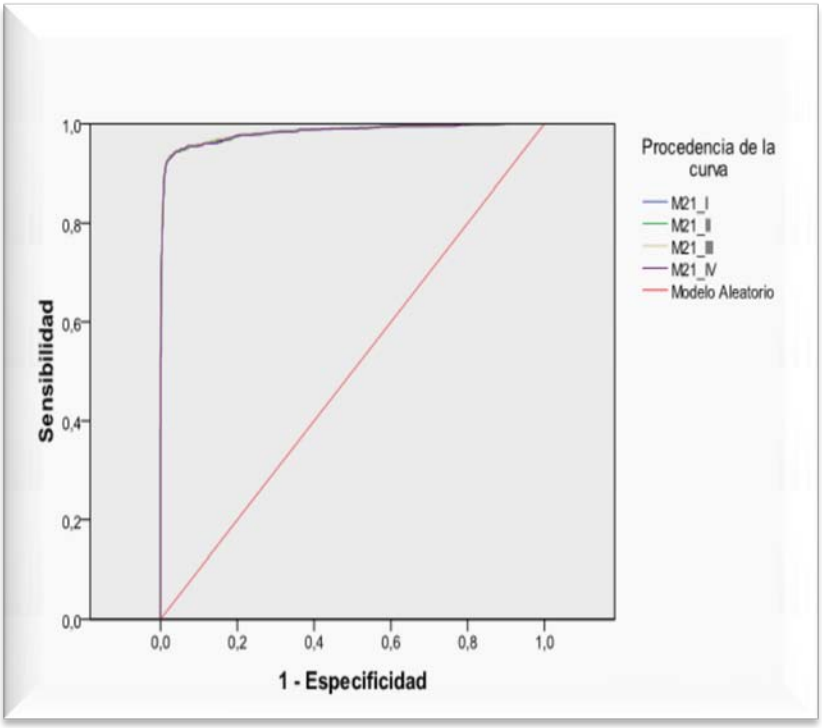


Figura 7.22.-Curvas ROC de evaluación. HLLM.

Analizaremos los resultados de las tablas 7.37 y 7.38 en la subsección 7.5.4, una vez que analizemos la generalización de los cuatro modelo HLLM en lasiguiente subsección 7.5.3.

7.5.3- Generalización. Error Test de los modelos HLLM preseleccionados.

Con el objetivo de conocer si los cuatro modelos alternativos seleccionados son suficientemente generalizables, así como establecer su ranking respecto de esta característica, analizamos en este apartado el error test, es decir el error de predicción esperado sobre la muestra test, para cada uno de ellos. Para ello utilizamos la muestra, independiente de las muestras de entrenamiento y validación, compuesta por 18.297 acreditados.

El error test empírico no es un buen estimador del error de generalización, sin embargo la ejecución del modelo sobre el conjunto de entrenamiento y el error sobre el conjunto test proporciona un estimador insesgado del error de generalización, a condición de que el conjunto test sea elegido aleatoriamente. Estimamos el error de predicción esperado por bootstraping sobre la muestra test, puesto que nos encontramos en una situación de riqueza de datos.

En la segunda fila de la tabla 7.39 se muestra el e*stimador del error de predicción esperado* por bootstraping, $\text{Error}_G\_\text{Bag}=\text{E}[\text{Error}_T\_Bag]$, estimador que calculamos como la media de los errores de generalización obtenidos a partir de 1000 remuestras bootstraps para cada modelo, y en la primera fila de la misma tabla, a efectos comparativos, el estimador *error empírico de entrenamiento por bootstraping,* $\overline{\text{Error}}\_Bag$, valor medio de los errores empíricos correspondientes a 1000 remuestras bootstraps..

Tabla 7.39.- Error Empírico_Bag (1000 remuestras bootstraping obtenidas de la muestra de entrenamiento) y Error de Predicción Esperado, (sobre la muestra test), de M_{21_I}, M_{21_II}, M_{21_III} y M_{21_IV}.

	M_{21_I}	M_{21_II}	M_{21_III}	M_{21_IV}
$\overline{\text{Error}}\_Bag$	0,1183	0,1170	0,1170	0,1185
Error_G_Bag	0.1194	0.1180	0.1171	0.1187

En la tabla 7.39 se observa que los errores empíricos y de generalización son muy similares para los modelos M_{21_III} y M_{21_IV}. El modelo M_{21_III} es el que presenta menores valores para ambos errores. Recuérdese que en este modelo se sustituyen las funciones de base

$\mathrm{WT\_V_1}$ y $BF1\_V_4$ del modelo $\mathrm{M_{21\_I}}$ por $MARS\_V_1\_2$ y $MARS\_V_4$ respectivamente, donde las combinaciones lineales de funciones de base son $MARS\_V_1\_2 = \sum_{l=2}^{4} \theta_{1,l} BF_{1,l}(V_1)$ y $MARS\_V_4 = \theta_{4,1} BF_{4,1}$, siendo las funciones bisagra para cada una de las combinaciones lineales $BF_{1,2} = máx(0,\ 98{,}9 - V_1)$, $BF_{1,3} = máx(0,\ V_1 - 1{,}5)$, $BF_{1,4} = máx(0,\ V_1 - 8{,}36)$ y $BF_{4,1} = Máx(0,\ 789{,}83 - V4)$. Obviamente esta estructura es más complicada que $\mathrm{WT\_V1}$ y $BF1\_V4$ del modelo $\mathrm{M_{21\_I}}$, y dado que los errores se diferencian en milésimas deberá tenerse en cuenta esta información en el momento de seleccionar el mejor modelo.

7.5.4 Selección del modelo de credit scoring proactivo.

Tabla 7.40.- Ranking del Ajuste, BIC, y Poder Discriminante, AUC, sobre la muestra de validación, y del Error Esperado de Generalización, sobre la muestra test, de los modelos $\mathrm{M_{21\_I}}$, $\mathrm{M_{21\_II}}$, $\mathrm{M_{21\_III}}$ y $\mathrm{M_{21\_IV}}$.

Estadístico	Rango
BIC de Evaluación	(2.305,71) $\mathrm{M_{21\_III}} \overset{+16,43}{\prec} \mathrm{M_{21\_IV}} \overset{+3,77}{\prec} \mathrm{M_{21\_I}} \overset{+7,23}{\prec} \mathrm{M_{21\_II}}$ (2.333,14)
AUC de Evaluación	(0,9795) $\mathrm{M_{21\_III}} \overset{-0,0001}{\succ} \mathrm{M_{21\_IV}} \overset{-0,0001}{\succ} \mathrm{M_{21\_II}} \overset{-0,0001}{\succ} \mathrm{M_{21\_I}}$ (0,9792)
$\mathrm{Error_G\_Bag}$, Error de Predicción Esperado Bootstraping	(0,1171) $\mathrm{M_{21\_III}} \overset{+0,0009}{\prec} \mathrm{M_{21\_II}} \overset{+0,0007}{\prec} \mathrm{M_{21\_IV}} \overset{+0,0007}{\prec} \mathrm{M_{21\_I}}$ (0,1194)

Como puede observarse en la tabla 7.40 el "mejor modelo desde el punto de vista estadístico" es $\mathrm{M_{21\_III}}$, puesto que presenta el menor BIC, la mayor Área Bajo la Curva ROC, AUC, es decir, se ajusta y discrimina mejor entre las poblaciones de default y no default, a la vez que presenta el Error de Predicción Esperado más pequeño. En este sentido el "peor" modelo es $\mathrm{M_{21\_I}}$, pero las diferencias reflejadas en la tabla 7.40, son muy pequeñas. Respecto del poder discriminante la diferencia entre el "mejor" y el "peor" modelo es de 3 diezmilésimas, $\left| AUC_{\mathrm{M_{21\_III}}} - AUC_{\mathrm{M_{21\_I}}} \right| = 0{,}0003$. Respecto de la capacidad de generalización la diferencia entre ambos modelo, $\left| Error_G\_\mathrm{Bag\_M_{21\_III}} - Error_G\_\mathrm{Bag\_M_{21\_II}} \right| = 0{,}0023$, es poco más de 2 milésimas. En consecuencia, desde el punto de vista estadístico cualquiera de los

cuatro modelos es en principio "bueno", pues para todos ellos, el coeficiente de determinación de Nagelkerke es superior a 0,80, lo que significativa un alto grado de ajuste, la medida AUC sobre la muestra de validación es casi 0,98 y el error de predicción esperado presenta tan sólo una ligera discrepancia respecto del error empírico de entrenamiento, ambos obtenidos por técnicas bootstrap.

Dado que la única diferencia en la estructura de los cuatro modelos radica en la expansión lineal por funciones de base utilizada para especificar la no linealidad de las variables V_1 y V_4, aplicaremos el requerimiento de facilidad de interpretación sobre estas dos variables para seleccionar nuestro modelo final.

I.- M_{21_I} : $\begin{cases} Z_1(V_1) = 0,4418(WT\_V_1) \\ Z_4(V_4) = 0,0009(BF_1\_V_4) \end{cases}$

II.- M_{21_II} : $\begin{cases} Z_1(V_1) = RCS\_V_1\_2 = -0,0849h_{12}(V_1) + 0,1077h_{13}(V_1) \\ Z_4(V_4) = RCS\_V_4\_2 = 0,0005h_{43}(V_4) - 0,0011h_{44}(V_4) \end{cases}$

III.- M_{21_III} : $\begin{cases} Z_1(V_1) = MARS\_V_1\_2 = 0,7559(BF_{12}\_V_1) + 0,5546(BF_{13}\_V_1) + 0,2194(BF_{14}\_V_1) \\ Z_4(V_4) = MARS\_V_4 = 0,0008(BF_1\_V_4) \end{cases}$

IV.- M_{21_IV} : $\begin{cases} Z_1(V_1) = 0,6997(RBF\_V_1) \\ Z_4(V_4) = 0,4832(RBF\_V_4) \end{cases}$

De los cuatro pares de expansiones lineales por funciones de base anteriores las dos más fáciles de interpretar son sin duda los correspondientes a los modelos M_{21_I}, (WT_V_1, $BF_1\_V_4$) , y M_{21_III}, ($MARS\_V_1\_2$, $MARS\_V_4$), siendo el primero el modelo de menor complejidad.

Por tanto, ***el modelo M_{21_I} parece ser la mejor opción como modelo de credit scoring proactivo para pronosticar la relación de dependencia entre el estado de default y las 21 variables explicativas del riesgo de crédito consideradas.***

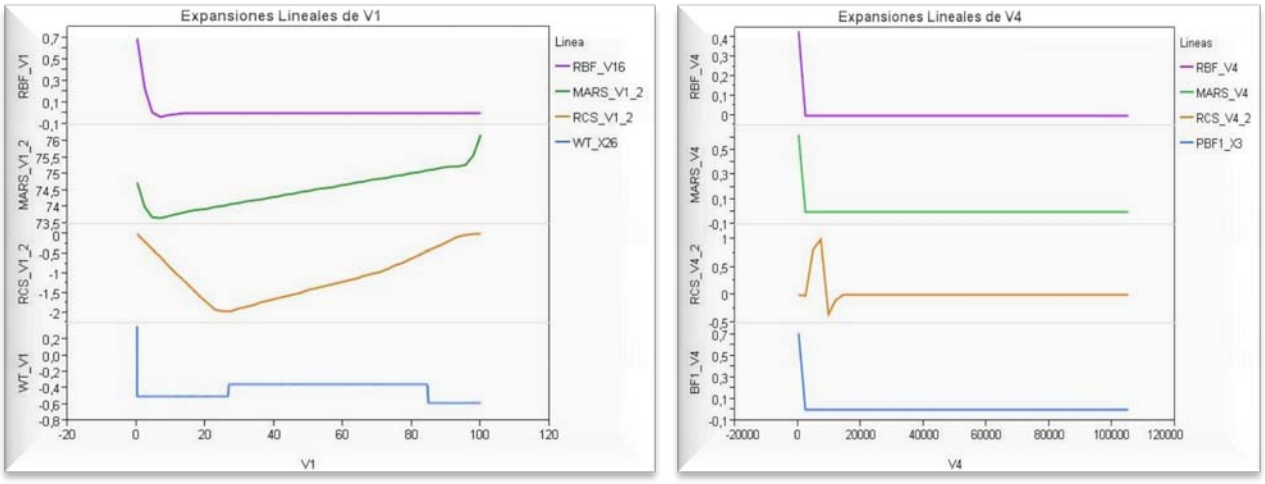


Figura 7.23.- Expansiones lineales por funciones de base de las variables V_1 y V_4 en los 4 modelos seleccionados.

En el panel derecho de la figura 7.23 se observa que $BF_1\_V_4$, $MARS\_V_4$ y $RBF_1\_V_4$ coinciden, la coincidencia entre $BF_1\_V_4$ y $MARS\_V_4$ es trivial por cuanto son idénticas y se han repetido en la figura por simetría con el panel de la izquierda, sin embargo la coincidencia gráfica entre $MARS\_V_4$ y $RBF_1\_V_4$ es harina de otro costal,

$$MARS\_V_4 = 0{,}0008\ máx\left(0,\ 789{,}83 - V4\right) = RBF_1\_V_4 = 0{,}4832\ \exp\left(-V_4^2\right)$$

por tanto, el modelo final viene dado por la estructura (7.81)

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right) = \beta_0 + \sum_{\substack{r=1 \\ r\neq 5}}^{16} \beta_r U_r + \beta_{17}\left(WT\_U_{17}\right) + \beta_{19}\left(W\_U_{19}\right) + \sum_{r=1}^{2} \theta_{r,1}\left(WT\_V_r\right)$$
$$+\theta_{4,1} BF_1\_V_4 + \theta_{6,1} WT\_V_6$$

Una cuestión que surge de forma natural una vez que se han obtenido las funciones de base más prometedoras para construir el modelo logístico lineal híbrido M_{21_I} es si el método de obtención de las mismas es o no excesivamente sensible a la configuración de la muestra de entrenamiento. Antes de comenzar la fase de búsqueda del modelo más parsimonioso, es necesario comprobar previamente que no se produce tal extremo, cuestión que abordamos en la siguiente subsección, 7.5.5.

7.5.5 Sensibilidad de la selección de las funciones de base a la configuración de la muestra de entrenamiento

Una forma intuitiva de abordar la cuestión de si el método de obtención de las funciones de base es o no excesivamente sensible a la configuración de la muestra de entrenamiento consiste en analizar la estabilidad de los parámetros y estadísticos de bondad de ajuste, poder discriminante y clasificación correcta frente al muestreo de entrenamiento, lo que podemos realizar a través del ajuste por regresión logística bagged del modelo M_{21_I}, sobre 200 remuestras bootstraping de la muestra de entrenamiento.

Una medida que proponemos para medir la sensibilidad del método puede basarse en las discrepancias entre los estimadores de los coeficientes del modelo y los diferentes indicadores de la bondad del ajuste, del poder discriminante y de la eficacia de clasificación del mismo ajustado sobre la muestra de entrenamiento, HLLR, y los promedios de los 200 ajustes sobre la remuestras bootstraps, HLLR_Bag:

$$\begin{aligned}&\hat{\beta}_0\\&\hat{\beta}_r,\ r=1,...4,\ 6,...,17,19\\&\hat{\theta}_{r,1},\ r=1,2,4,6\end{aligned}$$

$$\begin{aligned}\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)=\hat{\beta}_0+\sum_{\substack{r=1\\r\neq 5}}^{16}\hat{\beta}_r U_r+\hat{\beta}_{17}\left(WT\_U_{17}\right)+\hat{\beta}_{19}\left(W\_U_{19}\right)+\sum_{r=1}^{2}\hat{\theta}_{r,1}\left(WT\_V_r\right)\\+\ \hat{\theta}_{4,1}\left(BF_1\_V_4\right)+\hat{\theta}_{6,1}\left(WT\_V_6\right)\end{aligned}\qquad (7.85)$$

Los estimadores bagging de los coeficientes de las funciones base en el modelo son las medias muestrales de los coeficientes correspondientes para el conjunto de coeficientes de los modelos estimados para las distintas muestras bootstrap:

$$\begin{aligned}&\hat{\beta}_0\_bag=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_0^{*b},\ b=1,...,B\\&\hat{\beta}_r\_bag=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_r^{*b},\ b=1,...,B,\ r=1,...,4,\ 6,...,17,19\\&\hat{\theta}_{r,1}\_bag=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{r,1}^{*b},\ b=1,...,B,\ r=1,2,4,6\end{aligned}\qquad (7.86)$$

donde $b\in\{1,...,B\}$ indica la remuestra bootstrapping b-ésima de la muestra de entrenamiento.

El estimador del logit de la probabilidad de conjunto, $\text{logit}\left[\hat{P}^{bag}\left(Y=1/\ X=x\right)\right]$, se obtiene, por tanto, según la expresión siguiente:

$$\text{logit}\left[\hat{P}^{bag}\left(Y=1/\ X=x\right)\right]=\frac{1}{B}\sum_{b=1}^{B}\text{logit}\left[\hat{P}^{*b}\left(Y=1/\ X=x\right)\right]$$
$$=\hat{\beta}_0^{bag}+\sum_{\substack{r=1\\r\neq 1}}^{16}\hat{\beta}_r^{bag}U_r+\hat{\beta}_{17}^{bag}\left(WT\_U_{17}\right)+\hat{\beta}_{19}^{bag}\left(W\_U_{19}\right)+\sum_{r=1}^{2}\hat{\theta}_{r,1}^{bag}\left(WT\_Vr\right) \qquad (7.87)$$
$$+\hat{\theta}_{4,1}^{bag}\left(BF_1\_V_4\right)+\hat{\theta}_{6,1}^{bag}\left(WT\_V_6\right)$$

La expresión (7.87) es un estimador de Montecarlo del verdadero estimador bagging, alcanzándolo cuando $B\to\infty$. El número finito B , número de remuestras, en la práctica gobierna el ajuste de la aproximación de Montecarlo.

Tabla 7.41.- Análisis del estimador de máxima verosimilitud de los parámetros del modelo M_{21_I} y del modelo de conjunto M_{21_I}_Bag obtenido por bootstraping de 200 remuestras.

	M_{21_I}	M_{21_I}_Bag	Diferencia
Intercept	-3,3275885	-3,309577	0,018011
U_1	0,05329182	0,053568	-0,000276
$BF1\_V_4$	0,00093881	0,000967	-0,000028
W_U_{19}	-0,793538	-0,795541	0,002003
WT_V_6	-1,9951853	-2,023706	0,028521
U_7	-0,2320306	-0,225944	-0,006087
U_{13}	-0,4132248	-0,41208	-0,001145
U_4	1,22535696	1,229101	-0,003744
U_{14}	-0,2676023	-0,273059	0,005457
WT_V_1	0,44182872	0,440397	0,001432
WT_V_2	-0,6876707	-0,713744	0,026073
WT_U_{17}	0,95874415	0,935538	0,023206
U_8	-0,0126164	-0,012633	0,000017
U_{11}	0,00797753	0,007948	0,000030
U_3	3,2644472	3,22053	0,043917
U_{15}	-0,3468924	-0,373859	0,026967
U_2	1,92645781	1,95133	-0,024872
U_9	-0,0092989	-0,009264	-0,000035
U_{16}	0,00014296	0,000141	0,000002
U_{10}	-0,1359273	-0,139864	0,003937
U_{12}	0,00082724	0,000824	0,000003
U_6	-0,0017517	-0,001773	0,000021

Como puede observarse en la tabla 7.41 las discrepancias entre los estimadores de máxima verosimilitud de los coeficientes de ambos modelos M_{21_I} y M_{21_I}_Bag son muy pequeñas, y, por otro lado, lo mismo ocurre con los estadísticos de bondad de ajuste, tabla 7.42, y con que los estadísticos de clasificación, tabla 7.43, lo que indica que la estabilidad de nuestro modelo es satisfactoria.

Tabla 7.42.- Comparación de los estadísticos de bondad de ajuste, error empírico y poder discriminante del modelo M_{21_I} y del modelo de conjunto M_{21_I}_Bag obtenido por bootstraping de 200 remuestras bootstrap.

	M_{21_I}	$M_{21\_I}\_Bag$
$-LogVer\_Dif$	7590,0279	7603,7805
$-LogVer\_Tot$	2185,2946	2171,5420
$-LogVer\_\mathrm{Red}$	9775,3225	9775,3225
$R^2(U)$	0,7764	0,7779
R^2	0,3394	0,3399
$\tilde{R}^2$	0,8203	0,8215
$-2log(L)$	4370,5892	4343,0840
Error Empírico	0,1194	0,1186
AIC	4412,5892	4385,0840
BIC	4591,2571	4563,7519
AUC	0,9800	0,9802

Tabla 7.43.- Comparación de los estadísticos de clasificación de los modelos M_{21_I} y el modelo de conjunto M_{21_I}_Bag obtenido por bootstraping de 200 remuestras bootstrap.

Tabla de clasificación (Para un nivel de probabilidad de 0.50)									
	Correcto		Incorrecto		Porcentajes				
Modelo	Default	No Default	Default	No Default	Correcto	Sensi-bilidad	Especi-ficidad	Falso POS	Falso NEG
M_{21_I}	2347	33588	264	408	98,16	85.2	99.2	10,01	1.2
M_{21_I}_Bag.	2346	33597	257	407	98,19	85.2	99.2	9.9	1.2

7.6 REDUCCIÓN DE LA COMPLEJIDAD DEL MODELO.

7.6.1 Introducción.

Para evitar posibles inconvenientes como el de sobreajuste y, en general, la complejidad del modelo, se procede a aplicar técnicas de regularización o "poda" para reducir el número de funciones de base.

El "proceso de poda" que utilizamos para reducir la complejidad del modelo M_{21_I} consiste en un *Proceso de Selección Limitada de Funciones de Base hacia Atrás* hasta conseguir un modelo óptimo desde el punto de vista estadístico y desde el enfoque de los requerimientos de Basilea II.

7.6.2 Proceso de poda de funciones de base.

El proceso de poda utilizado consiste en un *ajuste paso a paso hacia atrás,* utilizando el procedimiento *Logistic*, con selección Backward, implementado en SAS® / STAT V9.2.

Tabla 7.44.- Ajuste paso a paso hacia atrás del M_{21_I}.

Paso	Parámetro	Acción	L-R ChiCuadrado	"Sig Prob"	p
1	WT_V_1	Poda	0,169357	0,6807	21
2	U_{13}	Poda	7,094592	0,0077	20
3	WT_V_2	Poda	8,541140	0,0035	19
4	WT_U_{17}	Poda	9,857960	0,0017	18
5	U_{14}	Poda	14,406380	0,0001	17
6	U_7	Poda	17,837130	0,0000	16
7	U_{12}	Poda	17,859700	0,0000	15
8	U_{10}	Poda	27,700700	0,0000	14
9	U_{11}	Poda	27,284330	0,0000	13
10	U_3	Poda	63,135640	0,0000	12

Como puede observarse en la tabla 7.44, el procedimiento usado poda 10 variables, 9 de ellas de forma muy significativa, la función de base WT_V_1 es la más dudosa, pero dado que pertenece a la visión de activo total, visión de la que se conservan las variables U_2, U_8 y U_9 en el modelo, no se crea ningún problema con su poda. El resto de visiones quedan convenientemente representadas.

Tabla 7.45.- Variables seleccionadas para el modelo proactivo final.

U_1	X_{53}	Porcentaje de contratos del cliente en incidencia, sobre el total de contratos que hayan estado operativos en algún momento del mes actual.
U_2	X_{48}	Número de meses, en los últimos 12, que el Cliente ha tenido descubierto (saldo deudor) en Ahorro Vista
U_4	WR_X_{55}	WOE_ Porcentaje que representa el importe de la nómina / pensión de los últimos 3 meses frente al total de ingresos percibidos en los 3 últimos meses.
U_6	X_8	Rango mínimo de saldo (importe máximo – importe mínimo) del Cliente en Pasivo a la Vista para los últimos 12 meses.
U_8	X_{24}	Mínimo ratio del plazo pendiente a fin de mes frente al plazo original de la operación en préstamos para financiación a particulares
U_9	X_{57}	Porcentaje que representa el importe en riesgo del cliente en la Caja a fin de mes frente su importe en riesgo total (considerando tanto la Caja como otras Entidades Financieras recogidas en la CIRBE).
U_{15}	X_{42}	Número de recibos básicos cargados en cuenta en los dos últimos meses
U_{16}	X_6	Importe medio que el Cliente ha percibido de manera recurrente en los últimos 12 meses.
U_{19}	X_{15}	Importe total que el Cliente debe pagar para hacer frente a los requerimientos de pago de productos de Activo.
V_4	X_3	Saldo medio del Cliente en pasivo líquido para los últimos 12 meses
V_6	X_{19}	Máxima antigüedad en el último año del cliente en activo

La estructura formal del modelo con las 11 variables seleccionadas mostradas en la tabla 7.45, al que notaremos por M_{11_I}, se expresa en la forma

$$\begin{aligned}\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\beta_1U_1+\beta_2U_2+\beta_4U_4+\beta_6U_6+\beta_8U_8+\beta_9U_9+\beta_{15}U_{15}\\+\beta_{16}U_{16}+\beta_{19}W\_U_{19}+\theta_{4,1}BF_1\_V_4+\theta_{6,1}WT\_V_6\end{aligned}\tag{7.88}$$

Tabla 7.46.- Análisis del EVM de los parámetros del modelo M_{11_I}.

Término	Estimador	Error STD	χ^2	Prob> χ^2
Intercepto	-3,768957	0,2503807	226,59	<,0001*
U_1	0,05360776	0,0009956	2899,2	0,0000*
U_2	1,76712216	0,0601029	864,46	<,0001*
U_4	0,61125742	0,0986894	38,36	<,0001*
U_6	-0,0025053	0,0002492	101,08	<,0001*
U_8	-0,0105305	0,0013724	58,88	<,0001*
U_9	-0,009027	0,001581	32,60	<,0001*
U_{15}	-0,5316419	0,0860046	38,21	<,0001*
U_{16}	0,00012988	2,3184e-5	31,39	<,0001*
W_U_{19}	-0,7578914	0,0594901	162,30	<,0001*
BF1_V_4	0,00163612	0,0002581	40,19	<,0001*
WT_V_6	-2,1590795	0,1935004	124,50	<,0001*

El modelo resulta, por tanto,

$$\begin{aligned}\text{logit}\left(\hat{P}\left(Y=1/\ X=x\right)\right)=-3,7789+0,0536\ U_1+1,7671\ U_2+0,6113\ U_4-0,0025\ U_6\\-0,0105\ U_8-0,0090\ U_9-0,5316\ U_{15}+0,0001\ U_{16}\\-0,7579\ W\_U_{19}+0,0016\ BF_1\_V_4-2,1591\ WT\_V_6\end{aligned}\tag{7.89}$$

Obviamente es importante verificar que el *modelo reducido* presenta, por un lado, buenas condiciones estadísticas, buen ajuste, por ejemplo coeficiente de Nagelkerke superior a 0.80, un alto poder discriminante y predictivo, ejemplo, AUC superior a 0,95, alta eficacia como clasificador, porcentaje de clasificados correctamente superior a 98 %, y que generaliza bien, y, por otro lado, que presenta adecuadas características desde el punto de vista del riesgo de crédito, así, por ejemplo, de ser posible que todas las visiones parciales del cliente estén representadas y de los requerimientos de Basilea II.

Tabla 7.47.- Cuadro de ajuste de los modelos logísticos híbridos por expansiones lineales de funciones de base M_{21_I} y M_{11_I}.

	M_{21_I}	M_{11_I}
$-LogVer\_Dif$	7590,0279	7493,0845
$-LogVer\_Tot$	2185,2946	2282,2380
$-LogVer\_Red$	9775,3225	9775,3225
$R^2(U)$	0,7764	0,7665
R^2	0,3394	0,3359
$\tilde{R}^2$	0,8203	0,8119
-2LogVer	4370,5892	4564,4760
Error Empírico	0,1194	0,1247
AIC	4412,5892	4586,4760
BIC	4591,2571	4680,0639
AUC	0,9801	0,9789
% Clasificación Incorrecta	1,82%	1,91%

Tabla 7.48.- Test de bondad de ajuste de Hosmer y Lemeshow para M_{11_I} .

Modelo	Chi-cuadrado	GL	Pr > ChiSq
M_{11_I}	45,1432	8	<.0001

El test de bondad de ajuste de Hosmer y Lemeshow, $Pr > ChiSq < 0.001$, indica un buen ajuste para el modelo reducido M_{11_I}.

Tabla 7.49.- Probabilidad Pronosticada, (Área bajo la curva ROC) sobre la muestra de entrenamiento para M_{11_I}.

	AUC				
				Intervalo de confianza asintótico al 95%	
Modelo	AUC	Error típ.[a]	Sig. asintótica[b]	Límite inferior	Límite superior
M_{11_I}	0,9789	0,0017	0,0000	0,9756	0,9822

a. Bajo el supuesto no paramétrico.

b. Hipótesis nula: área verdadera = 0,05.

Receive Operating Curve ROC

____ Modelo M_{11_I} (**AUC=0,9789**)____ Modelo Aleatorio (AUC=0,5000)

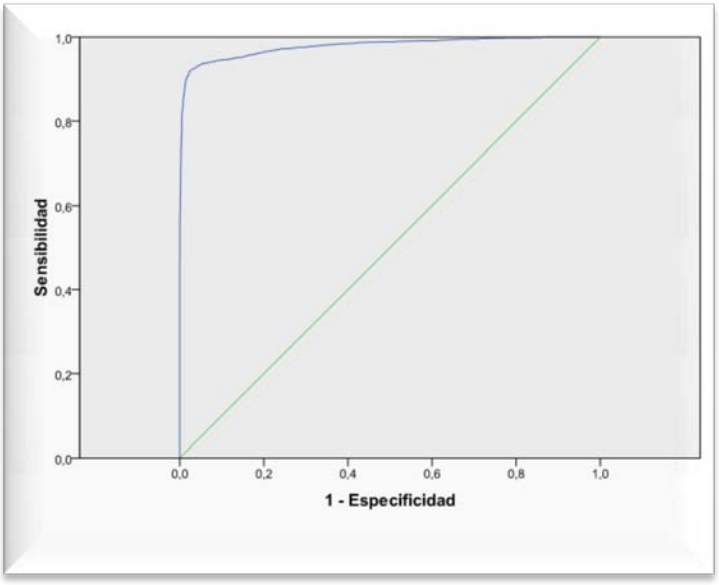


Figura 7.24.- Curva ROC y área bajo la curva, AUC, del modelo M_{11_I} para la muestra de entrenamiento.

Tabla 7.50.- Tabla de Clasificación (para un punto de corte de 0,50).

Modelo	Correcto		Incorrecto		Porcentajes				
	Default	No Default	Default	No Default	Correcto	Sensi-bilidad	Especi-ficidad	Falso POS	Falso NEG
M_{11_I}	2324	33582	270	431	98,09	84.4	99.2	10.4	1.3

Desde el punto de vista estadístico, como se deduce de las tablas 7.47, 7.48, 7.49 y 7.50, *el modelo reducido* M_{11_I} *presenta un buen ajuste, coeficiente de Nagelkerke superior a 0.80, un alto poder discriminante y predictivo, AUC muy próximo a 0,98, y alta eficacia como clasificador, porcentaje de clasificados correctamente 98,1 %.*

Tabla 7.51.- Error Empírico_Bag (1000 remuestras bootstraping de entrenamiento) y Error de Predicción Esperado_Bag, o Error Test_Bag, (1000 remuestras bootstraping obtenidas de la muestra test), de M_{21_I} y M_{11_I}.

	M_{21_I}	M_{11_I}
$\overline{\text{Error}}\_Bag$	0,1183	0,1248
Error_G_Bag	0,1194	0,1228

Respecto de la *capacidad de generalización del modelo* M_{11_I}, como se muestra en la tabla 7.51, el *error de predicción esperado_bag* es muy parecido al *error empírico_bag* y se observa también que la diferencia de estos estimadores entre ambos modelos M_{21_I} y M_{11_I} es del orden de las milésimas.

Además en M_{11_I} *están representadas las visiones más importantes de riesgos a fecha 30 de noviembre de 2007* consideradas en el modelo no reducido M_{21_I}: *visión general del cliente*, W_U19 y WT_V6; *visión de pasivo*, BF_1_V_4, U_{16}, U_6, U_{15} y U_4; *visión de activo*, U_2, U_8 y U_9; *visión de incidencias*, U_1.

En adelante nos referiremos al modelo M_{11_I} como HLLM.

7.7 PODER DISCRIMINANTE DEL MODELO HLLM.

7.7.1 Introducción.

Según el documento de trabajo *nº 14,* BCBS (2005a), un objetivo fundamental de la validación de un sistema calificación de solicitantes de crédito consiste en comprobar su eficiencia que se basará, entre otros, en el *poder discriminante* de los modelos de scoring. Por tanto, dedicamos esta sección a validar el poder discriminante del modelo HLLM. El objetivo será juzgar si la función de calificación o puntuación asociada a este modelo es apropiada para discriminar entre los solicitantes *"buenos" y "malos"*, por lo que *se califica a través del modelo* HLLM *a todos los acreditados de los que se dispone en el año 2007, y se valida si los estadísticos de poder discriminante son suficientemente significativos*.

La primera tarea a acometer será transformar la función de calificación de acreditados para que cumpla la premisa de que *debe asignar bajas puntuaciones a los solicitantes que presenten alta probabilidad de incumplimiento,* principio que facilitará sin duda la interpretación de tal función a la vez que su explicación a los clientes.

En *segundo lugar* analizamos el "*Perfil de la diferencia entre las funciones de distribución acumulativas*" de las poblaciones de buenos y malos, y su *estadístico de Kolmogorov-Smirnov, K*-S, asociado. La conveniencia del uso de estos dos instrumentos estadísticos radica, por un lado, en que el perfil de las funciones de distribuciones de las poblaciones de default y no default es un instrumento gráfico muy potente para visualizar las posibles diferencia entre ambas poblaciones, y, por otro, lo que aún es más importante, en que se cuenta con un test asociado para contrastar si las diferencias son o no significativas.

Como se recoge en el apartado 4.3.2 del capítulo 4: *"cuando la diferencia entre la frecuencias relativas acumuladas de dos muestras aleatorias de datos es muy pequeña, las distribuciones de las dos poblaciones, origen de tales muestras, deben ser similares y recíprocamente, cuando la las distribuciones de las dos poblaciones no son similares, la diferencia entre las frecuencias relativas acumuladas de las muestras debe ser significativa".* La cuestión, por tanto, se reduce a medir la distancia entre las poblaciones de default y no default a partir de la función de calificación de acreditados. Para medir la distancia entre las funciones de densidad de las puntuaciones se utiliza el área de solapamiento de las densidades de probabilidad de los acreditados buenos y los acreditados malos.

En tercer lugar analizaremos otras medidas del poder discriminante tales como la *Tasa de Precisión* (AR), que se obtiene a partir del *Índice de Gini*, resumen de la *Curva de Ajuste Acumulativo* (CAP), también conocida como Curva de Lorenz, que consiste en la representación gráfica de los puntos de coordenadas $(1-F(s),1-F_0(s))$ para cada puntuación, y puede estimarse por medio de las funciones empíricas de distribución acumulada.

La curva que permite visualizar el poder discriminante de un modelo de calificación es la *curva ROC,* (7.71) y figura 7.18 y la medida del poder discriminante asociada es el á*rea bajo la curva ROC*, AUC, (7.72).

Por último, dado que las propiedades estadísticas de AUC coinciden con las del estadístico de Mann-Witney, podemos aplicar el potente Test U de Wilcoxon-Mann-Witney para comparar el AUC del modelo HLLM con el del modelo aleatorio, $AUC = 0{,}50$. Dos de los test no paramétricos más clásicos para contrastar si dos distribuciones son o no idénticas son el test de la suma de rangos de Wilcoxon y su equivalente, el test U de Mann-Witney.

7.7.2 Cambio de localización y escala de la función de calificación de acreditados.

Como se muestra en la figura 7.25, la relación entre la puntuación asignada por el modelo HLLM y la probabilidad de default no cumple el principio básico de *asignar bajas puntuaciones a los solicitantes que presenten alta probabilidad de incumplimiento*, ya que asigna mayor puntuación a los acreditados con mayor probabilidad de default.

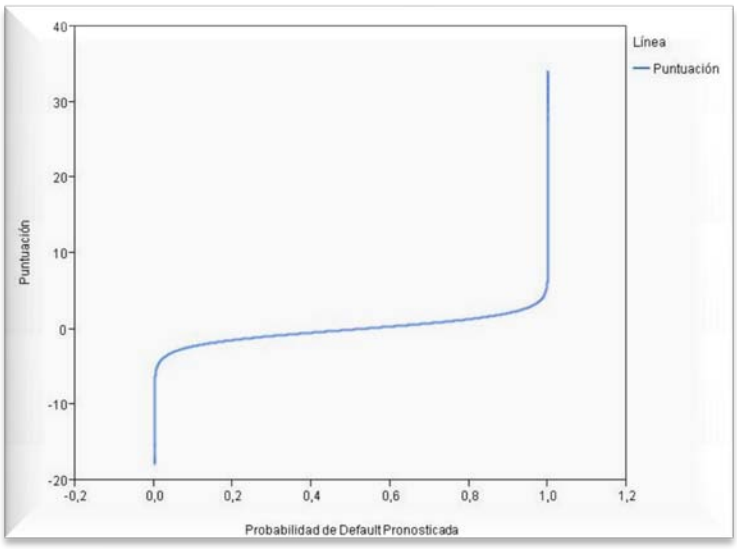


Figura 7.25.- Puntuación frente a la probabilidad de default pronosticada del modelo M_{11_I}.

Con el fin de que la función de calificación de acreditados sea monótona decreciente respecto de la probabilidad de default es usual realizar un *cambio de localización y escala* para conseguir que los clientes calificados por el modelo como malos (con una probabilidad de incumplimiento elevada) les corresponda una puntuación muy baja y, recíprocamente, a los clientes mejores (con una probabilidad de incumplimiento pequeña) les corresponda una puntuación elevada, es decir:

$$P(Y=0/X=x_1)<P(Y=0/X=x_2)\Rightarrow S(x_1)\leq S(x_2) \tag{7.90}$$

La escala será única e independiente del modelo que proporcione la estimación de la probabilidad de *mal pagador* lo que permitirá comparar la calidad crediticia de clientes pertenecientes a distintos segmentos de la cartera, en el caso de que proceda.

La puntuación resultante se llama *puntuación escalada* y en su obtención se utilizan habitualmente los siguientes parámetros de localización y escala, (SIDDIQI, 2006):

- Inicialmente se asigna una puntuación concreta, usualmente 600 puntos, cuando $odds\left(P(Y=0/X=x)\right)=\dfrac{P(Y=0/X=x)}{P(Y=1/X=x)}=50$, es decir, 50:1, lo que significa que un cliente que recibe 600 puntos es 50 veces más probable que sea Bueno a que sea Malo, (o, en lenguaje coloquial, la probabilidad de no default presenta una ventaja de 50 a 1 frente a la probabilidad de default).

- Cada vez que se dobla esta proporción en los odds se incrementan los puntos en 20.

Por la linealidad de los modelos que estamos considerando, y del mismo modo que se calcula la probabilidad de incumplimiento del cliente como suma de las contribuciones de las variables intervinientes, es inmediato descomponer la puntuación final del cliente como suma de las puntuaciones correspondientes a cada característica o expansión por funciones de base de estas. Análogamente al reparto de los puntos totales entre las variables, los puntos correspondientes a una característica se distribuyen entre sus atributos según la siguiente fórmula:

$$Puntuación = offset + factor \times ln\left(odds\left(P\left(Y=0/X=x\right)\right)\right) \tag{7.91}$$

Y dado que

$$\begin{aligned} ln\left(odds\left(P\left(Y=0/X=x\right)\right)\right) &= -ln\left(odds\left(P\left(Y=1/X=x\right)\right)\right) \\ &= -\text{logit}\left(P\left(Y=1/X=x\right)\right) \end{aligned} \tag{7.92}$$

Para el *modelo logístico híbrido por expansiones lineales de funciones de base* M_{11_I}, sustituyendo (7.130) en (7.131) se tiene

$$Puntuación = offset + factor \times \left(-\left(\begin{array}{l} \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_4 U_4 + \beta_6 U_6 + \beta_8 U_8 \\ +\beta_9 U_9 + \beta_{15} U_{15} + \beta_{16} U_{16} + \beta_{19} W\_U_{19} \\ +\theta_{4,1}\, BF_1\_V_4 + \theta_{6,1}\, WT\_V_6 \end{array} \right) \right) \tag{7.93}$$

De acuerdo con el planteamiento de SIDDIQI (2006), los parámetros de localización, *offset*, y escala, *factor*, se obtienen según el siguiente sistema de ecuaciones:

$$\begin{array}{r} 620 = offset + ln(100)\ factor \\ 600 = offset + ln(50)\ factor \\ \hline 20 = \left(ln(100) - ln(50)\right) factor \end{array}$$

de donde,

$$\begin{aligned} factor &= \frac{20}{ln(100)} = 28{,}8539 \\ offset &= 600 - ln(50)\ factor = 487{,}123 \end{aligned} \tag{7.94}$$

Pero esta transformación afín puede seguir dando puntuaciones negativas lo que no es conveniente por imagen del cliente. Este inconveniente puede resolverse relocalizando de nuevo la función de calificación de acuerdo con el siguiente parámetro de localización

$$offset^{nuevo} = offset - pm + pmi \tag{7.95}$$

donde pm es la puntuación mínima otorgada por la función de calificación anterior y pmi es la puntuación mínima que con la nueva función de calificación se pretende otorgar a los acreditados, a la que llamaremos *puntuación mínima de imagen*.

Para el modelo HLLM con $factor = 28{,}8539$ y $offset = 487{,}123$, la puntuación mínima asignada por la función de calificación es -501, por lo que, asignando una puntuación mínima a los acreditados de 500 puntos, el nuevo offset alcanza el valor de 1488,123, y la nueva función de calificación, a la que añadiremos el calificativo de *re escalada*, viene dada por,

$$Puntuación = 1488{,}123 + 28{,}8539 \begin{pmatrix} -3{,}7789 + 0{,}0536\ U_1 + 1{,}7671\ U_2 \\ +0{,}6113\ U_4 - 0{,}0025\ U_6 - 0{,}0105\ U_8 \\ -0{,}0090\ U_9 - 0{,}5316\ U_{15} + 0{,}0001\ U_{16} \\ -0{,}7579\ W\_U_{19} + 0{,}0016\ BF_1\_V_4 \\ -2{,}1591\ WT\_V_6 \end{pmatrix} \tag{7.96}$$

Como puede observarse en la figura (7.26) la relación entre la probabilidad de default estimada por el modelo y la puntuación escalada es ahora monótona decreciente.

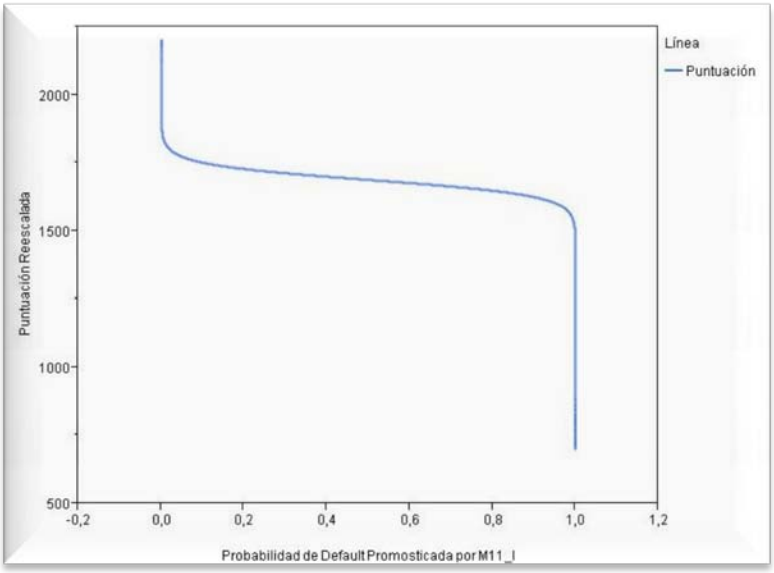


Figura 7.26- Probabilidad de Default Pronosticada frente a la Puntuación Escalada para el Modelo HLLM.

7.7.3 Perfil de la diferencia entre las funciones de distribución acumulativas. Test estadístico de Kolmogorov-Smirnov asociado.

Denotaremos, como es habitual, por $F_0(\bullet)$ y $F_1(\bullet)$ las funciones de distribución acumulada de $S(X)/(Y=0)$, (la función de calificación o puntuación condicionada al grupo de los acreditados buenos), y de $S(X)/(Y=1)$, (la función de calificación puntuación condicionada al grupo de los acreditados malos), respectivamente y por $f_0(\bullet)$ y $f_1(\bullet)$ las correspondientes funciones de densidad, es decir

$$\begin{aligned} F_0(s) &= P(S \le s / Y=0) = \int_{-\infty}^{s} f_0(u)\,du \\ F_1(s) &= P(S \le s / Y=1) = \int_{-\infty}^{s} f_1(u)\,du \\ F(s) &= P(S \le s) = \int_{-\infty}^{s} f(u)\,du \end{aligned} \tag{7.97}$$

En el panel izquierdo de la figura 7.27 se muestra, para el modelo HLLM, la representación gráfica de las funciones de densidad de las puntuaciones asignadas por la función de acreditados de las poblaciones de buenos acreditados, en azul, y de malos acreditados, en rojo, estimadas a través de la estimación no paramétrica por el núcleo Gaussiano, y en el panel de la derecha las funciones de distribución acumuladas correspondientes.

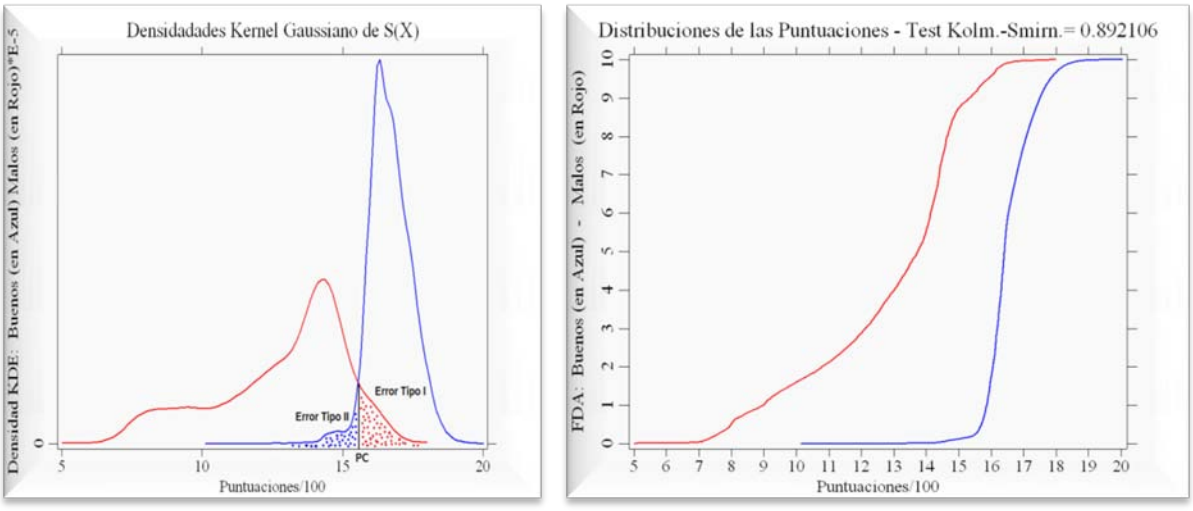


Figura 7.27.- Perfiles de las Funciones de Densidad y Distribución Acumulada de las Puntuaciones para las poblaciones de buenos y malos acreditados, para el modelo HLLM y Test de Kolmogorov-Smirnov. Conjunto de acreditados default en rojo y conjunto de acreditados no default en azul.

Como puede observarse en la figura 7.27, las dos densidades tienen un punto de intersección en $S(X)=s$ que es la frontera de separación que permite predecir que *"todos los solicitantes con puntuación* $S(X)<s$ *serán incumplidores"*. En principio este punto de intersección podría ser el punto de corte seleccionado, pero en principio no tiene por qué.

Los valores pronosticados se determinan mediante un punto de corte, Pc (cut_off), basado en la probabilidad, que teóricamente suele ser 0,5. De este modo el valor previsto de la variable dependiente es igual a 1 si la probabilidad pronosticada está por encima de 0,5 y 0 en caso contrario. Para el punto de corte Pc, existen dos Zonas de Error cuyo significado numérico se recoge en la tabla cruzada 7.52, *matriz de confusión*, siendo

Verdadero: número de clientes 'buenos' para los que el modelo predice correctamente.

Falsa Alarma: número de clientes 'buenos' para los que el pronóstico es incorrecto.

Fallo: número de clientes 'malos' con pronóstico incorrecto.

Éxito: número de clientes 'malos' para los que el modelo predice correctamente.

Tabla 7.52.- Matriz de Confusión. Estado de default estimado y observado para los acreditados a 30 de noviembre de 2007.

			Estado de Default Pronosticado en 2007		
			Buenos	Malos	Total
Estado de Default Observado en 2007		Buenos	**67.008** Verdadero	**693** Falsa Alarma	67.701
		Malos	**803** Fallo	**4.706** Éxito	5.509
		Total	67.811	5.399	73.210

La matriz de confusión recoge el porcentaje de predicciones correctas. El modelo es perfecto si todos los casos están en la diagonal de la tabla de clasificación. En el caso de default los conjuntos de datos son muy desequilibrados en el sentido de que sólo una pequeña fracción incumple las obligaciones de pago contractuales, por ejemplo, el suceso default sólo ocurre el 2% de las veces.

Si, a partir de la matriz de confusión se calculan las dos situaciones de mala predicción, Error tipo I y Error tipo II, con un punto de corte con probabilidad pronosticada de 0,50, se tiene

$$\textit{Riesgo de Crédito}\ (\text{Error Tipo I}) \quad : \quad \frac{803}{67811} = 0{,}0118 \quad (1{,}18\%)$$

$$\textit{Coste de pérdida de Negocio}\ (\text{Error Tipo II}) \quad : \quad \frac{693}{5399} = 0{,}1284 \quad (12{,}84\%)$$

El coste de pérdida de negocio parece bastante elevado, sin embargo, tratar de rebajarlo conlleva situar el punto de corte, por política de riesgos de la entidad financiera, hacia la derecha haciendo una política más flexible, lo que obviamente va en contra del riesgo de crédito por cuanto este aumenta. Por tanto el punto de corte habrá de situarse de acuerdo con la política que necesite la entidad financiera en cada momento equilibrando y adaptando estos riesgos a sus necesidades.

En general, cuando se usa una matriz de confusión para determinar la bondad del ajuste se concluye que siempre será preferible un modelo que conduzca a una probabilidad de default constante igual a cero. En el caso del riesgo de crédito, esta probabilidad de default cero es inútil para el cálculo de los requerimientos de capital. Por otro lado, la importancia de los dos tipos de error es obviamente diferente en los problemas de credit scoring, a causa de que la predicción correcta del riesgo de default es más importante debido al alto coste que suele conllevar la clasificación incorrecta de los acreditados incumplidores. Ambas razones implican que el error total de clasificación (*error total de calificación* suma del Error de Tipo I y el Error Tipo II), no es un criterio apropiado para medir el rendimiento de un modelo de clasificación de acreditados, su *poder discriminante*, aunque desgraciadamente se utiliza, por simplicidad y no con poca frecuencia, para comparar diferentes modelos de clasificación. Por tanto, es necesario utilizar otras medidas alternativas para medir el poder discriminante del modelo conjuntamente con los test estadísticos correspondientes. Una medida importante se basa en la región de solapamiento de las funciones de densidad, panel izquierdo de la figura 7.27.

La región de solapamiento O está constituida por la zona bajo la densidad de buenos, a la izquierda del umbral s, y la zona bajo la densidad de malos, a la derecha de s. Dado que hay un solo punto de intersección óptimo, y la relación entre la puntuación $S(\bullet)$ y la probabilidad de incumplimiento es monótona positiva el área de solapamiento está definida por

$$O = \min_s \{F_1(s) + 1 - F_0(s)\} \tag{7.98}$$

Por lo que se puede definir una *medida del poder discriminante* como

$$\left|D_{N_0N_1}\right| = 1 - O = \max \left|F_0(s) - F_1(s)\right| \tag{7.99}$$

El indicador del poder discriminante $|D|$ toma valores en el intervalo [0,1], donde $\left|D_{N_0N_1}\right| = 1$ indica una separación total y $\left|D_{N_0N_1}\right| = 0$ significa que no existe separación alguna.

Si se considera el test de hipótesis donde la hipótesis nula $F_1(s) = F_0(s)$, que indica que las dos muestras resultan de la misma población, es decir, la función de calificación de acreditados no discrimina entre default y no default, o, de forma más general, la variable de calificación $S(\bullet)$ no influye sobre la probabilidad de incumplimiento, entonces *se rechaza la hipótesis nula a nivel de significación* α *si*

$$T = \left|D_{N_0N_1}\right| \sqrt{\frac{N_0 N_1}{N_0 + N_1}} > \kappa_{N_0,N_1,1-\alpha/2} \tag{7.100}$$

donde K_α se obtiene de $P(K \le x) = 1 - \alpha$. , donde κ es una variable aleatoria con función de distribución acumulada

$$P(\kappa \le x) = 1 - \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x} e^{\frac{-(2i-1)^2\pi^2}{8x^2}} \tag{7.101}$$

El nivel probabilidad de dos colas se estima usando los tres primeros términos de la formula de SMIRNOV (1948):

$$\begin{aligned}
&si \;\; 0 \le T \le 0{,}27 \Rightarrow \; \text{Prob}{>}\kappa_{1-\alpha/2} = 1 \\
&si \;\; 0{,}27 \le T < 1 \;\Rightarrow \text{Prob}{>}\kappa_{1-\alpha/2} = 1 - \frac{2{,}506628}{T}\left(Q + Q^9 + Q^{25}\right), \;\; donde \; Q = e^{-1{,}233701\; T^{-2}} \\
&Si \;\; 1 \le T \le 3{,}1 \;\; \Rightarrow \text{Prob}{>}\kappa_{1-\alpha/2} = 2\left(Q - Q^4 + Q^9 - Q^{16}\right), \qquad donde \; Q = e^{-2T^2} \\
&si \;\; T \ge 3{,}1 \;\; \Rightarrow \; \text{Prob}{>}\kappa_{1-\alpha/2} = 0
\end{aligned} \tag{7.102}$$

Tabla 7.53.- Test bilateral de Kolmogorov_Smirnov.

H0	H1	Test Estadístico	Rechazo
$F_1(s) = F_0(s)$	$F_1(s) \neq F_0(s)$	$\left\|\hat{D}_{N_0N_1}\right\| = \max_s \left\{\left\|\hat{F}_0(s) - \hat{F}_1(s)\right\|\right\}$	$T = \left\|D_{N_0,N_1}\right\| \sqrt{\frac{N_0N_1}{N_0+N_1}} > \kappa_{N_0N_1,1-\alpha/2}$

Como puede observarse en la tabla 7.54, se obtiene un *alto valor del estadístico de Kolmogorov-Smirnov,* $\hat{T}=63{,}6744$, $\hat{T}>3{,}1$, por lo que según (7.102 la significación asintótica bilateral es $\text{Prob}>\kappa<0.0001$, es decir, las puntuaciones de los dos grupos de acreditados no se distribuyen uniformemente y, por tanto, podemos afirmar que *el modelo logístico híbrido por expansiones lineales de funciones de base,* HLLM*, discrimina bien ambos grupos.*

Tabla 7.54.- Resultados del Tests de Kolmogorov_Smirnov sobre HLLM.

Diferencias más extrema absoluta	$\left\|\hat{D}_{N_0N_1}\right\|$	0,892106
T de Kolmogorov-Smirnov	$\hat{T}=\left\|\hat{D}_{N_0N_1}\right\|\sqrt{\frac{N_0N_1}{N_0+N_1}}$	63,674421
Sig. asintót. (bilateral)		<0,0001

a. Variable de agrupación: Estado de Default

7.7.4 Curva de Ajuste Acumulativo, CAP, y Tasa de Precisión, AR.

La curva de Lorenz es la representación gráfica de los puntos de coordenadas $\{1-F(s),1-F_1(s)\}$ para cada puntuación . La cantidad $F(s)=P(S\le s)$ coincide con la *Tasa de Alarma* para la puntuación s y $F_1(s)=P(S\le s/Y=1)$ con la *Tasa de Aciertos* para la misma puntuación. La cantidad $100\times L(s)$ indica el porcentaje de acreditados incumplidores que se hallan entre los $100\times s$ primeros acreditados (de acuerdo a sus puntuaciones).

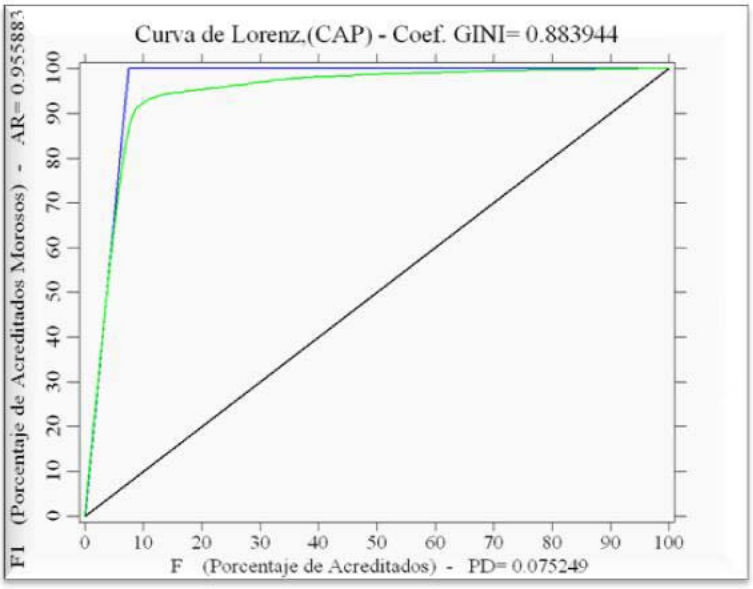


Figura 7.28.- Curva de Lorenz, (CAP), y sus estadísticos asociados Índice de Gini y Tasa de Precisión, (AR), para la función de calificación de acreditados del modelo HLLM.

En la figura 7.28 se representa gráficamente la curva de Lorenz para las puntuaciones que otorga función de calificación del modelo HLLM a los acreditados a 30 de noviembre de 2007.

Una curva de Lorenz idéntica a la diagonal corresponde a una puntuación que ordena a los solicitantes de crédito de forma totalmente aleatoria. El modelo aleatorio no tiene ningún poder discriminante, es decir, en este modelo cualquier fracción de deudores con puntuaciones bajas contendrá la misma fracción de deudores morosos. El modelo de calificación real está entre los dos extremos y, por tanto, el área de la región comprendida entre la curva de Lorenz CAP y el modelo aleatorio es una buena medida de la eficacia de la puntuación. Uno de esas medidas es el *Coeficiente de Gini* , G, que consiste en dos veces esa área.

El índice de Gini para las calificaciones del modelo HLLM sobre el total de los acreditados de 2007 es $G = 0{,}8839$.

La medida más usual y popular sobre la precisión de un modelo de credit scoring, alternativa al coeficiente de Gini, es la *Tasa de Precisión,* (AR), también basada en la curva CAP, (KEENAN y SOBEHART (1999) , ENGELMANN et al. (2006)). El propio Grupo de Implementación de los acuerdos de Basilea II afirma, documento de trabajo *nº 14 del* BCBS (2005a), que "*la calidad de un sistema de calificación viene dada por la tasa de precisión de sus modelos de credit scoring*".

La tasa AR se es cociente entre el índice de Gini de la curva CAP y coeficiente de Gini de una curva CAP muy especial, la *Curva de Lorenz Optimal* que se corresponde con la puntuación que separa perfectamente a los acreditados cumplidores de los no cumplidores. Esta curva tiene la peculiaridad de que la ordenada $(1-F_1)(s)= 1$ se alcanza para la abscisa $(1-F)(s)= P(Y=1)$, por lo que para la Curva de Lorenz Optimal, se tiene que el *coeficiente de Gini optimal*

$$G_{opt} = P(Y=0) = 1-P(Y=1) \tag{7.103}$$

La Tasa de Precisión, AR, viene dada por la relación del coeficiente de Gini de cada puntuación y el coeficiente de Gini de la Curva de Lorenz Optimal.

$$AR = \frac{G}{G_{opt}} \tag{7.104}$$

Es decir, se obtiene a partir de los estimadores de ambos coeficientes de Gini.

Para nuestro modelo M_{11_I}, se tiene $P(Y=1)=0{,}075249$, y, por tanto, se verifica $G_{opt} = P(Y=0) = 1-0{,}075249 = 0{,}92751$, de donde $AR = \dfrac{G}{G_{opt}} = \dfrac{0{,}8839}{0{,}9275} = 0{,}9559$.

El valor de AR se sitúa entre 0 y 1 si la Curva de Lorenz es realmente cóncava, es decir, si existe una relación monótona positiva entre $S(X)$ e Y, como es nuestro caso. El modelo de calificación óptimo es aquel para el que más se aproxime AR a 1. El peor modelo de calificación es aquel para el que más se aproxime AR a 0, es decir, el que más se aproxime al modelo aleatorio, esto hace que AR sea, no sólo una medida idónea para medir el poder discriminante de un modelo, sino también para comparar diferentes puntuaciones.

Dado que la Tasa de Precisión de la Curva de Lorenz asociada al modelo de credit scoring HLLM alcanza un valor muy alto, $AR=0{,}9559$, se tiene que *el modelo posee una alta calidad como función discriminante entre acreditados default y no default.*

7.7.5 Curva ROC y Área bajo la Curva ROC, AUC.

Una segunda curva que permite visualizar el poder discriminante de un modelo de calificación es la *curva ROC*, (7.71), que es la que usualmente se utiliza en las aplicaciones prácticas de credit scoring cuando el modelo proporciona la función de probabilidad, como es el caso del modelo HLLM. Esta curva de aspecto muy similar a la curva CAP presenta frente a ésta una importante diferencia, en el eje horizontal se sitúa $1-F_0(s)$, mientras en la primera se sitúa $1-F(s)$, siendo $F(s)$ la función de distribución acumulada del total de acreditados o solicitantes de crédito; se supone en ambos casos que los porcentajes de acreditados están ordenados desde puntuaciones malas a buenas.

La Curva ROC consiste en la representación gráfica de los puntos de coordenadas $\{1-F_0(s), 1-F_1(s)\}$ para cada puntuación s.

En la figura 7.29 se representa gráficamente la curva ROC para la función de calificación de acreditados del modelo HLLM.

Como puede observarse, la curva ROC es ligeramente más compleja que la CAP, pero a cambio no requiere la composición muestral para reflejar la verdadera proporción de default y no default.

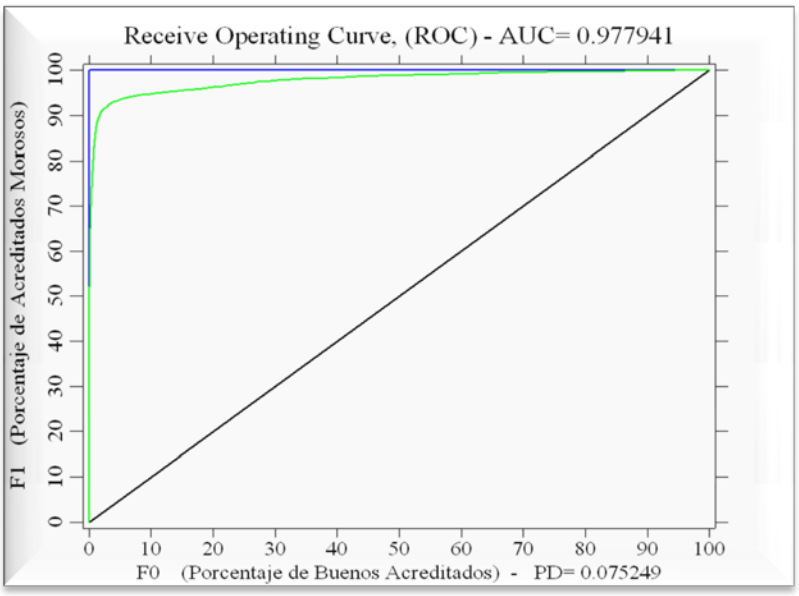


Figura 7.29.- Curva ROC y su estadístico asociado Área Bajo la Curva, AUC, para la función de calificación de acreditados del modelo HLLM.

La medida del poder discriminante asociada a la Curva ROC, es el á*rea bajo la curva ROC*, AUC; cuanto mayor sea el área bajo la curva ROC, mejor será el modelo.

En nuestro caso, modelo HLLM , el área bajo la curva ROC es $AUC = 0{,}9779$, que puede considerarse un valor alto como ocurre con el estadístico de Kolmogorov-Smirnov y la Tasa de Precisión, es decir *el modelo* HLLM *presenta un alto valor discriminante*.

7.7.6 Test U de Mann-Witney.

Dado que no existe un valor mínimo para AUC, con significación estadística, que nos permita decidir si el modelo de calificación tiene bastante poder discriminante, se puede utilizar el test U de Mann-Whitney para rechazar significativamente la hipótesis nula de que el modelo no tiene más poder discriminante que el modelo aleatorio.

El test *U* para funciones de calificación de acreditados continuas en su forma más simple se puede deducir en la forma siguiente:

Si denotamos por s_{j0} todas las puntuaciones observadas de no default y por s_{i1} todas las puntuaciones observadas de default, el estadístico del test U viene dado por

$$\hat{U} = \#\left\{s_{i1} > s_{j0}\right\}, \text{ sobre todo } i, j. \tag{7.105}$$

Para una separación perfecta entre acreditados default y no default, se obtiene $\hat{U} = N_0 N_1$. Si S e Y no están totalmente relacionadas, entonces el suceso $s_{i1} > s_{j0}$ ocurre con

probabilidad $1/2$, de forma que $U \approx \frac{1}{N_0 N_1}$. En consecuencia, una versión reescalada del estadístico $\hat{U}$, $\tilde{U} = \frac{\hat{U}}{N_0 N_1}$, es un estimador para el área bajo la curva que se puede obtener según la siguiente expresión:

$$U = P\{(S/Y=1) > (S/Y=0)\} = \int_{+\infty}^{-\infty} \left[(1 - F_1(s))\ d(1 - F_0)(s) \right] = AUC \tag{7.106}$$

Y, por lo tanto, dado que $AR = 2AUC - 1$, se tiene,

$$U = \left(\frac{AR+1}{2} \right) N_0 N_1 \tag{7.107}$$

Se demuestra, (LEHMANN, 1975) que bajo la hipótesis $F_1(s) = F_0(s)$, para N_0, N_1 grandes U se distribuye aproximadamente normal, con media $\mu_U = \frac{N_0 N_1}{2}$ y desviación típica $\sigma_U = \sqrt{\frac{N_0 N_1 (N_0 + N_1 + 1)}{12}}$, por tanto

$$Z = \frac{U - \frac{N_0 N_1}{2}}{\sqrt{\frac{N_0 N_1 (N_0 + N_1 + 1)}{12}}} \sim N(0,1) \tag{7.108}$$

El test U de Wilcoxon-Mann-Witney se construye entonces con los elementos de la tabla 7.55.

Tabla 7.55- Test U de Wilcoxon-Mann-Witney.

Test	H_0	H_1	Test Estadístico	Rechazo
(1)	$F_1(s) = F_0(s)$	$F_1(s) > F_0(s)$	U	$U > k_{N_1,N_0,1-\alpha}$
(2)	$F_1(s) = F_0(s)$	$F_1(s) < F_0(s)$	U	$U < N_1 N_0 - k_{N_1,N_0,1-\alpha}$

Siendo el valor crítico

$$k_{N_1,N_0,1-\alpha} = \frac{N_0 N_1}{2} + Z_{1-\alpha} \sqrt{\frac{1}{12} N_0 N_1 (N_0 + N_1 + 1)} \tag{7.109}$$

En nuestro caso los resultados del test de U de Wilcoxon-Mann-Witney se muestran en la tabla 7.56.

Tabla 7.56.- Resultados del Tests U de Wilcoxon-Mann-Witney sobre HLLM.

	Score
U de Mann-Whitney	8228194,000
W de Wilcoxon	2,341E7
Z	-118,170
Sig. asintót. (bilateral)	<0,0001

Variable de agrupación: Estado de Default.

De acuerdo con los resultados de la tabla 7.56, a partir del valor del estadístico de Mann-Whitney, $\hat{U} = 8228194$, se obtiene, por (7.108), un valor del estadístico $\hat{Z} = -118{,}170$ que proporciona una significación asintótica $\text{Prob}>Z_{1-\alpha} < 0.0001$, por lo que se rechaza la hipótesis nula de la igualdad de las distribuciones acumuladas de las poblaciones de default y no default, de este modo contamos con un instrumento de contraste más que nos permite afirmar que *el modelo logístico lineal híbrido por expansiones lineales de funciones de base* M_{11_I} *discrimina bien ambos grupos*.

La medida AUC tiene dos interesantes propiedades, en primer lugar, no depende de la probabilidad de default total de la cartera de crédito, de hecho puede ser estimado sobre muestras con proporciones de default/ y no default no representativas, y, en segundo lugar, pueden compararse representaciones de carteras de crédito bancario con diferentes proporciones de incumplidores.

No es posible definir para AUC un valor mínimo general, con significación estadística, en orden a decidir si el modelo de calificación tiene bastante poder discriminante. Como requerimiento mínimo del modelo de calificación, pueden servir el test Mann-Whitney o el de Kolmogorov-Smirnov, con un nivel de significación, por ejemplo, del 5%, para rechazar significativamente la hipótesis nula (el modelo no tiene más poder discriminante que el modelo aleatorio).

7.8 CALIBRACIÓN DEL MODELO HLLM.

7.8.1- Introducción.

En la sección 7.7 hemos analizado la capacidad del modelo para separar a los acreditados buenos de los malos, es decir su poder discriminante, análisis que hemos realizado utilizando los datos observados sobre el total de acreditados a 30 de noviembre de 2007. El segundo aspecto cuantitativo a validar es la *calibración*, entendida la calibración como la *corrección de la probabilidad de default,* PD, y como consecuencia de la *pérdida dado el default,* LGD, y la *exposición al default*, EAD.

Se dice que *un modelo está bien calibrado, si la fracción de sucesos que ocurren al final del horizonte temporal, se estima en forma no sesgada por el estimador de la PD de estos sucesos al principio de dicho horizonte*. La calibración compara las probabilidades de default pronosticadas al comienzo del horizonte temporal de referencia con las tasas de default observadas al final del período, analizando las discrepancias entre unas y otras en orden a discernir si las mismas se deben a factores sistemáticos o aleatorios. En definitiva, se plantea el análisis de si los hechos acaecidos a posteriori respaldan los pronósticos a priori, de no ser así seguramente el modelo no sea el más adecuado.

La cuestión que abordaremos en esta sección es comprobar si la probabilidad de default pronosticada por el modelo de credit scoring HLLM a 30 de noviembre de 2007 debe ser revisada dadas las tasas de default realmente observadas un año después.

Para la calibración es necesario, en primer lugar, asignar categorías de calificación de los acreditados, cuestión sobre la que se pronunció el Comité de Supervisión Bancaria de Basilea II en su tercer documento, BCBS (2003), en el que sugiere que los bancos deben tener un mínimo de 7 categorías de calificación para los prestatarios cumplidores y 1 categoría para no cumplidores, y que los acreditados deben ser razonablemente distribuidos a través de estas categorías, sin concentraciones excesivas.

Aparte de la recomendación BCBS (2003), no hay en la literatura de la industria del credit scoring, ni consenso sobre el número de las categorías de calificación para la partición, ni un método único para lograrlo. La mayoría de las Entidades Financieras (aproximadamente el 85%) usan escalas de calificación numérica.

Las categorías de calificación se obtienen usualmente a través de alguna regla con respecto a la probabilidad de default, por ejemplo, probabilidad de default media por categoría, o por

división en intervalos de determinado tamaño, o también utilizando el algoritmo CART para generar las categorías, tal como hemos hecho en esta Tesis Doctoral.

En la tabla 7.57 se muestran las distribuciones de default y no default así como las probabilidades pronosticadas en 2007 por el modelo HLLM y las tasas de default observadas un año después.

Si la probabilidad de default pronosticada por el modelo de credit scoring HLLM se desvía solo marginalmente de las tasas de default que han sido observadas, se dice que el modelo está bien calibrado. Utilizamos en este contexto la definición 4.6 recogida en el capítulo 4, debida a SEIDENFELD (1985), aplicada al marco de las probabilidades de default: "*Un conjunto de probabilidades de default son calibradas (o están bien calibradas), si el p por ciento de todas las predicciones informadas en probabilidad p son verdaderas*".

Tabla 7.57.- Categorías de calificación de HLLM. Probabilidades de Default Pronosticadas al 30 de noviembre de 2007 y Tasas de Default Observadas al 30 de noviembre de 2008.

	Categorías	Nº de Acreditados			Probabilidad	Tasa de Default
Nº	Intervalo de Calificación	No Default	Default	Total	Pronosticada 2007	Observada 2008
1	<= 1560,26	2.173	3.336	5.509	0,71849051	0,60555455
2	1560,27 - 1594,70	5.304	183	5.487	0,04004551	0,03335156
3	1594,71 - 1613,08	5.571	146	5.717	0,01797340	0,02553787
4	1613,09 - 1625,48	5.967	87	6.054	0,01061922	0,01437066
5	1625,49 - 1637,41	5.877	76	5.953	0,00685383	0,01276667
6	1637,42 - 1648,11	6.261	36	6.297	0,00473407	0,00571701
7	1648,12 - 1672,50	5.806	45	5.851	0,00273398	0,00769099
8	1672,51+	16.976	68	17.044	0,00046955	0,00398967

1 : Categoría de Calificación Default

Tradicionalmente, la capacidad de predicción de modelos binarios se analiza también a través de la matriz de confusión 2 × 2, tabla 7.58, donde las columnas son los dos valores pronosticados de la variable respuesta y las filas son los dos valores observados de dicha variable.

Tabla 7.58.- Matriz de Confusión. Estado de default pronosticado por el modelo HLLM, año 2007, y observado, año 2008.

			Estado de Default Pronosticado en 2007		
			Buenos	Malos	Total
Estado de Default Observado en 2008		Buenos	**52.851** Verdadero	**1.084** Falsa Alarma	53.935
		Malos	**887** Fallo	**3.090** Exito	3.977
		Total	53.738	4.174	57.912

Si se fija el punto de corte correspondiente a una probabilidad pronosticada de 0.5, el *Riesgo de Crédito,* (Error Tipo I), alcanza el valor 1,65% y el *Coste de Pérdida de Negocio,* (Error Tipo II), el 26,00%.

Por las mismas razones apuntadas para medir el poder discriminante del modelo, en el marco de la determinación de los requerimientos de capital, una tabla de confusión no es la herramienta más adecuada de calibración, e igual que allí habremos de proceder a considerar alternativas adecuadas.

Por otro lado, dado que las tasas de default están sujetas a fluctuaciones estadísticas, la cuestión que se trata de resolver para la calibración de modelos de credit scoring es la siguiente:

¿ Existen discrepancias entre las probabilidades de default estimadas al principio del horizonte temporal considerado condicionadas a una puntuación de crédito dada y las tasas de default realmente observadas para esa puntuación al final de dicho horizonte ?. (7.110)

El grado de discrepancia entre la probabilidad de default pronosticada y las tasas de default realmente observada puede indicar problemas potenciales y acciones que necesitan ser acometidas.

Para contrastar la significación de las discrepancias planteadas en la cuestión (7.110) es necesario utilizar test estadísticos hacia atrás (backtesting). El test estadístico puede entonces usarse, para un cierto nivel de significación pre especificado con su correspondiente p-valor, para tomar una decisión; *altos p-valores indican que el test es significativo y, por*

tanto, no se rechaza la hipótesis nula de que la probabilidad de default es al nivel p significativamente infraestimada. Elegir el apropiado nivel de significación depende de la política conservadora que en mayor o menor grado adopte la Entidad Financiera.

7.8.2.- Test basados en la hipótesis de independencia de los sucesos de default.

Un aspecto importante a tener en cuenta en estos test es si en las estimaciones de la probabilidad de default se tiene o no en cuenta el estado de la economía, por ejemplo, por la inclusión de variables explicativas macroeconómicas. Desde este punto de vista si se tiene en cuenta el estado de la economía las estimaciones condicionadas a este estado se llaman estimaciones en un *punto del tiempo*, (PIT, Point-In-Time), y en caso de no considerar este estado se denominan estimaciones a *través del ciclo*, (TTC, Through-The-Cycle), son lógicamente incondicionales y se calculan con datos de un ciclo económico completo. *Nuestro caso se refiere a estimaciones en un punto del tiempo.*

Los métodos más utilizados para calibrar la probabilidad de default sobre un único período de tiempo bajo la hipótesis de independencia de los sucesos de default son el *Test Binomial*, (ENGELMANN y RAUHMEIER, 2006) , el *Test Chicuadrado*, (HOSMER y LEMESHOW, 2000), y el *Test de Spiegelhalter* , (SPIEGELHALTER, 1986). Mientras el test Binomial sólo se puede aplicar a una simple categoría de calificación, los tests *Chi-cuadrado* de Hosmer-Lemeshow y el de Spiegelhalter pueden usarse para contrastar la adecuación de la predicción de la probabilidad de default para varias categorías de calificación.

En nuestro caso contamos con 57.912 acreditados a 30 de noviembre de 2007, y si N_r indica el número de acreditados que son clasificados en las clases de calificación $r \in \{1,...,8\}$, entonces se tiene $57.912 = \sum_{r=1}^{8} N_r$ e indicando por

$0 < \hat{p}_r < 1$: *Probabilidad de default pronosticada* por el sistema de calificación para la categoría r.

$0 < p_r < 1$: *Probabilidad de default real*, (desconocida), para la categoría r.

$0 < p_r^{obs} < 1$: Tasa de default observada, para r, es decir, la proporción de acreditados en default de un total de N_r acreditados en la categoría de calificación r.

7.8.2.1- Test Binomial.

El test Binomial está diseñado para contrastar los pronósticos de la probabilidad de default estimada por el modelo, $\hat{P}$, frente a la tasa de default observada, P^{obs}, para una categoría de calificación r dada usando test ya sean de una o dos caras.

Test de una cara:

$$\begin{aligned} H_0 &: \ \mathrm{P}_r = \hat{\mathrm{P}}_r \quad \text{La PD coincide con la probabilidad de default estimada en la clase r.} \\ H_1 &: \ \mathrm{P}_r > \hat{\mathrm{P}}_r \quad \text{La PD es infraestimada en la clase r.} \end{aligned} \tag{7.111}$$

Si se asume que los default ocurren independientemente**,** para cada categoría r, $r = 1,...,8$, *se busca contrastar si la probabilidad de default de una categoría de calificación es correcta frente a la alternativa de que está infra estimada*, es decir contrastar la hipótesis nula H_0 frente a la alternativa de una cara H_1. *Para un nivel de significación* α *se rechaza la hipótesis nula si el número de defaults,* $N_{1r} = N_r p_r^{obs}$, *es mayor que un valor crítico* $k^* = \min\left\{k \,/\, \sum_{i=k}^{N_r} \binom{N_r}{i} \hat{p}_r \left(1-\hat{p}_r\right) \le \alpha\right\}$:

$$N_r P_r^{obs} > k^* \tag{7.112}$$

Dado que la distribución Binomial converge a la distribución Normal cuando el número de pruebas crece, $\mathrm{P}_r^{obs} \sim N\left(\hat{\mathrm{P}}_r \ , \ \frac{\hat{\mathrm{P}}_r\left(1-\hat{\mathrm{P}}_r\right)}{N_r}\right)$, o, equivalentemente, $z = \frac{P_r^{obs} - \hat{P}_r}{\sqrt{\frac{\hat{P}_r\left(1-\hat{P}_r\right)}{N_r}}} \sim N(0,1)$.

en términos de la tasa de default, se tiene que s*e rechaza la hipótesis nula si la probabilidad de default observada* p_r^{obs} *es mayor que* $p_{1-\alpha}$*:*

$$P_r^{obs} > p_{1-\alpha} \tag{7.113}$$

donde $p_{1-\alpha} \approx \Phi^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_r\left(1-\hat{p}_r\right)}{N_r}} + \hat{p}_r$.

Esta aproximación a la $N(0,1)$ se aplica cuando $N_r > 1000$, nuestro caso.

En la tabla 7.59 se muestran los valores críticos y los p-valores para un nivel de significación $\alpha = 0.005$, para las 8 categorías de calificación de HLLM.

Tabla 7.59.- Test Binomial para el nivel de significación $\alpha = 0{,}005$.

	N_r	$\hat{P}_r$	P_r^{obs}	$\hat{P}_r\left(1-\hat{P}_r\right)$	$N_r P_r^{obs}$	$k^*_{r,0,995}$	$P_{r,0,995}$
1	5.509	0,71849051	0,60555455	0,2022619	3336,00002	4044,28603	0,734123439
2	5.487	0,04004551	0,03335156	0,03844187	183,00001	257,200208	0,046874468
3	5.717	0,0179734	0,02553787	0,01765036	146,000003	128,670678	0,022506678
4	6.054	0,01061922	0,01437066	0,01050645	86,9999756	84,8651372	0,014018027
5	5.953	0,00685383	0,01276667	0,00680686	75,9999865	57,2241678	0,00961266
6	6.297	0,00473407	0,00571701	0,00471166	36,000012	43,8635816	0,00696579
7	5.851	0,00273398	0,00769099	0,00272651	44,9999825	26,3012782	0,004495177
8	17.044	0,00046955	0,00398967	0,00046933	67,9999355	15,3000112	0,000897677

1 : Categoría de Calificación de Default. $Z_{0,995} = \Phi^{-1}(0{,}995) = 2{,}58$

Como puede observarse en la tabla 7.59, para la escala de calificación mostrada en la tabla 7.57, para la cual las 8 categorías de calificación verifican $N_r > 1000$, y para nuestro modelo M_{11_I} se verifica $N_r P_r^{obs} < k^*_{r,0.995}$ o, equivalentemente, $P_r^{obs} < P_{r,0.995}$, para $r = 1,2,6$, por lo que no se rechaza la hipótesis nula para esas categorías. En cambio para las categorías $r = 3,4,5,7,8$ se tiene $P_r^{obs} > P_{r,0.995}$, por tanto, en este caso se rechaza la hipótesis nula.

Por tanto, para $r = 1,2,6$ con un nivel de significación $\alpha = 0{,}005$ *"la probabilidad de default pronosticada no está infra estimada", aseveración que no podemos hacer para el resto de los grados de calificación.*

Evidentemente este resultado para $r = 1,2,6$ está conforme con la percepción del Banco Supervisor, sobre todo porque no se infra estiman las categorías 1, default, y la 2, no default, la más próxima a ella.

Test de dos caras:

$$H_0: P_r = \hat{P}_r \quad \text{La PD coincide con la probabilidad de default estimada en la clase r.}$$
$$H_1: P_r \neq \hat{P}_r \qquad (7.114)$$

siendo también $N_{1r} = N_r P_r^{obs}$ el estadístico de contraste, se tiene que la *región crítica para* p_r^{obs} *y un nivel de significación asintótica* α está dada por:

$$[0, p_{\alpha/2}) \cup (p_{1-\alpha/2}, 1] \tag{7.115}$$

y, por tanto, la *región de aceptación* es $\left[p_{\alpha/2} , p_{1-\alpha/2} \right]$.

Por tanto, se rechazará la hipótesis nula para la clase de calificación r, *la PD coincide con la probabilidad de default estimada por el modelo*, si p_r^{obs} cae dentro de la región crítica (7.115),

$$p_r^{obs} \in [0, p_{\alpha/2}) \cup (p_{1-\alpha/2}, 1] \tag{7.116}$$

Como puede observarse en la tabla 7.60, para todas las categorías de calificación, a excepción de 2 y 6, se verifica $p_r^{obs} \in \left[0,\ p_{0,0025}\right) \cup \left(p_{0,9975},\ 1\right]$, y $p_2^{obs}, p_6^{obs} \in \left[p_{0,0025},\ p_{0,9975}\right]$, es decir, *para las categorías 2 y 6 no se rechaza la hipótesis nula de que la probabilidad de default observada coincide con la probabilidad de default pronosticada un año antes pero se rechaza en el resto de categorías.* Esto lógicamente no puede tranquilizar al controlador del riesgo que habrá de estar interesado en una estimación tan exacta para todas las categorías como sea posible.

Tabla 7.60- Test Binomial de dos caras para el nivel de significación $\alpha = 0{,}005$.

	N_r	P_r^{obs}	$k^*_{r,0,0025}$	$k^*_{r,0,9975}$	$P_{0,0025}$	$P_{0,9975}$
1	5.509	0,60555455	3864,36489	4051,96355	0,70146395	0,735517072
2	5.487	0,03335156	178,918826	260,540601	0,03260777	0,047483251
3	5.717	0,02553787	74,5267693	130,981086	0,01303599	0,022910807
4	6.054	0,01437066	41,8780501	86,6994656	0,00691742	0,014321022
5	5.953	0,01276667	22,9134379	58,688262	0,00384906	0,009858603
6	6.297	0,00571701	14,5044964	45,1163811	0,0023034	0,007164742
7	5.851	0,00769099	4,77311427	27,2199197	0,00081578	0,004652182
8	17.044	0,00398967	0,05550136	15,950519	3,2564E-06	0,000935844

$Z_{0,9975} = \Phi^{-1}(0{,}9975) = 2{,}81$
$Z_{0,0025} = \Phi^{-1}(0{,}0025) = -2{,}81$

7.8.2.2 Test Chi_cuadrado.

La prueba binomial (o su extensión normal) es principalmente adecuada para contrastar un único nivel de calificación, pero no de varias o todas las categorías de calificación de forma simultánea. El test Chi_cuadrado de Hosmer-Lemeshow, (HOSMER y LEMESHOW (2000)), (Capítulo 4, 4.4.2.2), es en esencia un test de conjunto para varias categorías de calificación que se originó en el campo de la regresión categórica y es habitual referirse a él como *test de bondad de ajuste*.

Ahora se busca contrastar si las probabilidades de default son correctas para todas las categorías de calificación simultáneamente, es decir contrastar:

$$\begin{aligned} &H_0: \ P_1 = \hat{P}_1,..., P_8 = \hat{P}_8 \\ &H_1: \ \exists r \in \{1,...,8\} \text{ con } P_r \neq \hat{P}_r \end{aligned} \tag{7.117}$$

Aquí se asumen las siguientes hipótesis:

a) Las probabilidades de default pronosticadas por el modelo, $\hat{p}_r$, y las tasas de default observadas, p_r^{obs}, son idénticamente distribuidas.

b) Todos los sucesos de default tanto dentro de cada categoría de calificación como entre las categorías son independientes.

El test estadístico chi-cuadrado se deduce del estadístico chi-cuadrado de Pearson original, (véase D'ÁGOSTINO y STEPHEN, 1986), y viene dado en los siguientes términos:

Bajo las hipótesis a) y b), cuando $N_r \to \infty$ simultáneamente para todo $r = 1,...,8$, por el Teorema Central del Límite, se tiene:

$$t_8 = \sum_{r=1}^{8} N_r \frac{\left(P_r^{obs} - P_r\right)^2}{\sqrt{P_r\left(1-P_r\right)}} \xrightarrow{Distribución} \chi^2\left(8\right) \tag{7.118}$$

Es decir, la distribución de t_8 *converge en distribución* a una distribución χ^2 con 8 grados de libertad.

Por tanto, se rechazará la hipótesis nula para un nivel de significación asintótico α, si $t_8 = \sum_{r=1}^{8} N_r \frac{\left(p_r^{obs} - \hat{p}_r\right)^2}{\sqrt{\hat{p}_r\left(1-\hat{p}_r\right)}}$ es mayor que el $\left(1-\alpha\right)\_cuantil$ de una distribución $\chi^2\left(8\right)$.

El $p-valor$ del Test χ^2 es una medida para validar la adecuación de las probabilidades estimadas, cuanto más se acerca el $p-valor$ a cero peor es la estimación. Sin embargo, si las probabilidades de incumplimiento estimadas son muy pequeñas, la tasa de convergencia a la distribución- χ^2 puede ser muy baja también.

Debemos tener en cuenta que, al basarse la prueba de Hosmer-Lemeshow en las hipótesis de independencia y de aproximación Normal, este test posiblemente subestime el verdadero error tipo I, (BLOCHWITZ, 2006). Para un pequeño número de acreditados en cada categoría de calificación, la hipótesis nula es más difícil de rechazar.

A partir de los datos de la tabla 7.61, resulta $t_8 = 474{,}780625$ y, dado que el $(0{,}95)\_cuantil$ de la distribución χ^2 con 8 grados de libertad es $\chi^2_{8,0.950} = 15.51$, se tiene $t_8 > \chi^2_{8,0.950}$ por lo que se rechaza la hipótesis nula $H_0: P_1 = \hat{P}_1,..., P_8 = \hat{P}_8$, lo que implica que existe $r \in \{1,...,8\}$ con $P_r \neq \hat{P}_r$, tal como ya había resultado en los dos tests anteriores.

Tabla 7.61.- Test Chi_Cuadrado.

Categorías	N_r	$\hat{P}_r$	P_r^{obs}	$\left(p_r^{obs} - \hat{p}_r\right)^2$	$\hat{p}_r(1-\hat{p}_r)$	t_r
1	5.509	0,71849051	0,60555455	0,01275453	0,20226190	347,3947028
2	5.487	0,04004551	0,03335156	4,4809E-05	0,03844187	6,3958079
3	5.717	0,01797340	0,02553787	5,7221E-05	0,01765036	18,5341089
4	6.054	0,01061922	0,01437066	1,4073E-05	0,01050645	8,1092807
5	5.953	0,00685383	0,01276667	3,4962E-05	0,00680686	30,5760689
6	6.297	0,00473407	0,00571701	9,6617E-07	0,00471166	1,2912605
7	5.851	0,00273398	0,00769099	2,4572E-05	0,00272651	52,7306753
8	17.044	0,00046955	0,00398967	1,2391E-05	0,02166401	9,7487204
t_8						**474,7806250**

7.8.2.3 Test de Spiegelhalter.

Normalmente se calculan individualmente las probabilidades de default pronosticadas por el modelo para cada acreditado. Puesto que el test Chi_Cuadrado de Hosmer-Lemeshow, al igual que el test Binomial, requieren que todos los acreditados asignados a una categoría de calificación tengan la misma probabilidad de default, lo que requiere promediar las probabilidades de default pronosticadas de los acreditados que han sido clasificados en la misma categoría, en los cálculos se pueden introducir algunos sesgos. Se puede evitar este problema usando el test de Spiegelhalter, BCBS, (2005c), que permite las variaciones en probabilidades de default dentro de la misma categoría de calificación. Se asume que los sucesos default son independientes tanto dentro de cada categoría de calificación como entre todas la categorías.

El punto de partida del test de Spiegelhalter es el Error Medio Cuadrático, MSE, también conocido como Brier Score en validación, (BRIER, 1950), $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{p}_i\right)^2$.

El estimador $\hat{p}_i$ minimiza el MSE esperado si se da un adecuado pronóstico de la probabilidad de default. La hipótesis nula para el contraste es que *todas las probabilidades de default estimadas,* $\hat{p}_i$*, coinciden exactamente con la verdadera, aunque desconocida, probabilidad de default* $P\left(y_i = 1/\ X = x_i\right)$ *para todo i.*

$$H_0:\ \hat{p}_i = P\left(y_i = 1/\ X = x_i\right) = E\left[Y/\ X = x_i\right],\quad i = 1,...,N \tag{7.119}$$

Asumiendo la *hipótesis de independencia* y usando el *teorema central del límite*, se demuestra que *bajo la hipótesis nula*, (SPIGELHALTER, 1986),

$$Z_S = \frac{MSE - E\left[MSE\right]}{\sqrt{Var\left[MSE\right]}} = \frac{\sum_{i=1}^{N}\left\{\left(y_i - \hat{p}_i\right)^2 - \hat{p}_i\left(1-\hat{p}_i\right)\right\}}{\sqrt{\sum_{i=1}^{N}\hat{p}_i\left(1-\hat{p}_i\right)\left(1-2\hat{p}_i\right)^2}} \ \sim\ N\left(0,1\right) \tag{7.120}$$

En nuestro caso se tiene $Z_S = 24{,}9519$ y dado que el $(0{,}025)\_cuantil$ de la distribución N(0,1) es $Z_{0.025} = 1{,}96$, se tiene $Z > Z_{0.025}$. Por tanto, con nivel de significación $\alpha = 0{,}05$ se rechaza la hipótesis nula de que "*todas las probabilidades de default estimadas,* $\hat{p}_i$*, coinciden exactamente con la verdadera probabilidad de default* $P\left(y_i = 1/\ X = x_i\right)$ *para todo acreditado i*", lo que implica que existen acreditados con $\hat{P}_i \neq P\left(Y = 1/\ X = x_i\right)$, tal como ha resultado de los tres tests anteriores.

7.9 ANÁLISIS COMPARATIVO DE DISTINTAS TÉCNICAS DE CONSTRUCCIÓN DE MODELOS DE CREDIT SCORING PROACTIVO.

En esta sección compararemos el rendimiento del modelo HLLM con otros modelos construidos con las 11 funciones de base con las que se construyó M_{11_I}, U_1, U_2, U_4, U_6, U_8, U_9, U_{15}, U_{16}, W_U_{19}, $BF_1\_V_4$ y WT_V_6, que se corresponden con las variables originales $BF_1\_X_3$, X_6, X_8, W_X_{15}, WT_X_{19}, X_{24}, X_{42}, X_{48}, X_{53}, WR_X_{55} y X_{57}, y ajustados sobre la misma muestra de entrenamiento, 36.607 acreditados al 30 de noviembre de 2007, 2.755 acreditados default y 33.852 no default.

Los modelos a comparar con **HLLM**, *Modelo Logístico Lineal Híbrido, expansión lineal híbrida de funciones de base,* todos ellos analizados con cierto grado de detalle, sobre todo sus debilidades y fortalezas desde la óptica de Basilea II, en el capítulo 5 son:

HLPM: *Modelo Probit Lineal Híbrido, expansión lineal híbrida de funciones de base*, Capítulo 5, apartado 5.6.1.3.

TREE: *Árbol de Clasificación representación de Particiones Recursivas Binarias,* Capítulo 5, apartado 5.6.2.2.

SLPM: *Modelo Perceptron de Capa Simple de transmisión de información hacia adelante y una sola capa oculta,* Capítulo 5, apartado 5.6.2.5.

k-NN: *Modelo de los k Vecinos más Próximos en vecindades locales, Capítulo 2, subsección 2.5.2.*

SMV: *Maquinas de Vector Soporte,* . Capítulo 5, apartado 5.6.2.6.

A excepción de k-NN, las estructuras funcionales de los modelos anteriores se caracterizan por que una transformación se pueden expresar por expansiones lineales de funciones de base, (6.2), $g\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r h_r\left(X\right)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$, *por lo que pueden desarrollarse sobre las mismas funciones de base que se han utilizado para construir la estructura funcional del modelo HLLM* M_{11_I}.

Nuestra pretensión consiste en verificar si, aparte de las ventajas teóricas y las relacionadas con los requerimientos de Basilea II, el modelo HLLM presenta mejor rendimiento discriminante y predictivo, confirmados al final del horizonte temporal de

interés, que algunas de las técnicas utilizadas usualmente en los sistemas de calificación del riesgo de crédito.

Para cada uno de los modelos se analizará el *grado de ajuste a los datos de entrenamiento,* se validará el *poder discriminante* sobre la población total de acreditados a 30 de noviembre de 2007, población a partir de la cual se obtuvo la muestra de entrenamiento y, por último se analizará el *grado de concordancia* entre las probabilidades pronosticadas para el total de acreditados a 30 de noviembre de 2007 y las tasas de default realmente observadas un año después, 30 de noviembre de 2008, es decir, analizaremos *la habilidad del modelo para hacer estimaciones insesgadas de las probabilidades de default,* para los modelos que proporcionan tal probabilidad, HLLM, HLPM, TREE, SLPM, k-NN, o bien el *grado de acuerdo entre la distribución de default y no default observada y la pronosticada* para las *Maquinas de Vector Soporte*, SMV, que no proporcionan tal probabilidad, proporcionando únicamente la clasificación de los acreditados en default y no default.

7.9.1 Ajuste y Generalización de los modelos HLLM, HLPM, TREE, SLPM, k-NN y SVM.

7.9.1.1 Ajuste del modelo HLPM.

El ajuste del modelo HLPM, (5.35), $\text{probit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{p}\beta_r X_r=\beta_0+\boldsymbol{\beta}^T\boldsymbol{X}$, a los datos de entrenamiento se realiza por Regresión Probit Lineal, LPR, que proporciona los estimadores de los parámetros del modelo resolviendo el problema de optimización siguiente:

$$\min_{\beta_0,\boldsymbol{\beta}} -\left[-\boldsymbol{Y}^T\log\left(\boldsymbol{\Phi}\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)\right)-\left(\boldsymbol{1}-\boldsymbol{Y}\right)^T\log\left(1-\boldsymbol{\Phi}\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{U}+\boldsymbol{1}^T\boldsymbol{\theta}^T\boldsymbol{H}\left(\boldsymbol{V}\right)\right)\right)\right] \quad (7.121)$$

donde

$$\boldsymbol{U}^T=\left(U_1,\ U_2,\ U_4,\ U_6,U_8,\ U_9,\ U_{15},\ U_{16},\ W\_U_{19}\right),\ \ \boldsymbol{\beta}^T=\left(\beta_1,\ \beta_2,\ \beta_4,\ \beta_6,\beta_8,\ \beta_9,\ \phi_{15},\ \beta_{16},\ \beta_{19}\right)$$

$$\boldsymbol{\theta}^T=\left(\theta_{4,1},\theta_{6,1}\right)\ \text{y}\ \ \boldsymbol{H}\left(\boldsymbol{V}\right)^T=\left(BF_1\_V_4,\ WT\_V_6\right).$$

El problema (7.121) y su proceso de resolución son análogos a los correspondientes a la estimación de modelos HLLM.

El modelo ajustado adopta la expresión:

$$\begin{aligned}\text{Probit}\left(P\left(Y=1/\ X=x\right)\right)=&\ 2{,}0158+0{,}0283\ U_1+0{,}8360\ U_2+0{,}3249\ U_4-0{,}0010\ U_6\\&-0{,}0049\ U_8-0{,}0043\ U_9-0{,}1874\ U_{15}+0{,}0001\ U_{16}\\&-0{,}3310\ W\_U_{19}+0{,}0008\ BF_1\_V_4-0{,}9919\ WT\_V_6\end{aligned}\qquad (7.122)$$

7.9.1.2 Ajuste del modelo TREE.

El modelo TREE (5.55), $P\left(Y=1/\ X=x\right)=\sum_{r=1}^{q}\beta_r I_{[x\in R_r]}$, se estima utilizando el algoritmo CART con una función objetivo y un criterio de optimización muy peculiares debidos fundamentalmente a BREIMAN, (1984) y a QUINLAN (1986, 1987).

Dado un conjunto de entrenamiento $\tau=\left\{x_i,y_i\right\}_{i=1}^{N}\subset\left(X,Y\right)^N$ una vez obtenida una partición recursiva binaria óptima $\left\{R_1,...,R_q\right\}$, según se explica en detalle en la subsección 5.6.2.2, siendo $N_r=\sum_{i=1}^{N}I_{[x_i\in R_r]}$ el número de observaciones en la región R_r, *el estimador* $\hat{\beta}_r$ *de* β_r viene dado por la proporción de observaciones de la clase de default en el nodo r, que representa al subconjunto R_r, es decir, por $\hat{\beta}_r=\hat{p}_r=\frac{1}{N_r}\sum_{i=1}^{N}I_{[y_i=1]}I_{[x_i\in R_r]}$.

Por lo que nuestro modelo estimado viene dado por

$$\hat{P}\left(Y=1/\ X=x\right)=\sum_{r=1}^{q}\left(\frac{1}{N_r}\sum_{i=1}^{N}I_{[y_i=1]}I_{[x_i\in R_r]}\right)I_{[x\in R_r]}\qquad (7.123)$$

Para la estimación del modelo TREE, (7.123), M_{11_III}, se aplicó el procedimiento *Árbol de Clasificación* de PASW® Statistics de IBM®, con el método de agrupación, o crecimiento, CHAID, sobre las 11 funciones de base explicativas del riesgo de crédito seleccionadas. Como estadístico de contraste, tanto para partir los nodos como para unir categorías, se utilizó el chi-cuadrado de Pearson con nivel de significación $\alpha=0{,}05$; los valores de significación se ajustaron usando el método de Bonferroni.

Tabla 7.62.- Características del modelo TREE.

Variables independientes incluidas	U1, U2, BF1_V4, U6, U4
Número de nodos	27
Número de nodos terminales	17
Profundidad	3

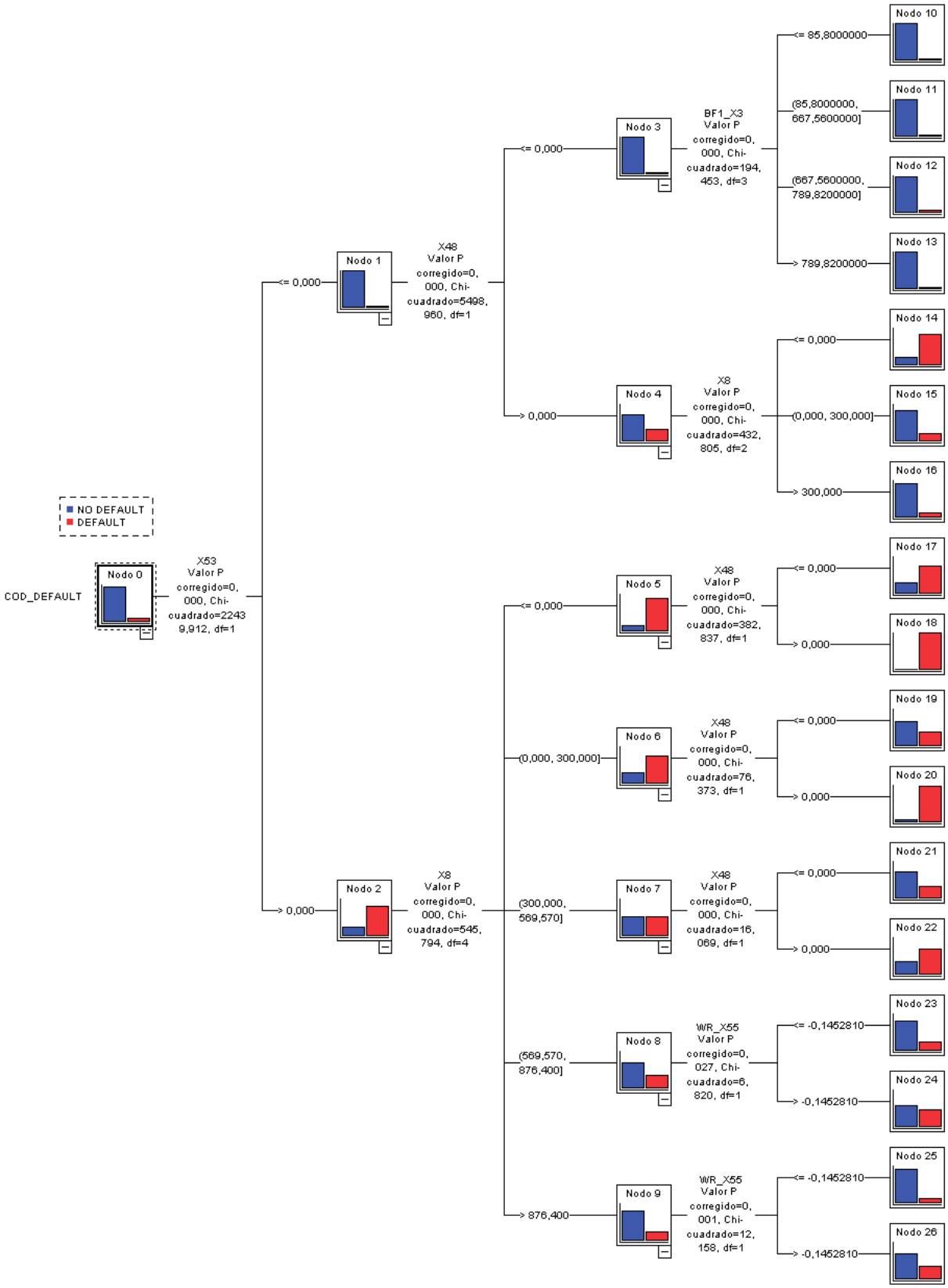


Figura 7.30.- Diagrama del árbol de clasificación correspondiente al modelo M_{11_III} .

El modelo ajustado adopta la expresión:

$$\begin{aligned} P(Y=1/\ X=x) = {} & 0{,}0011296\ I_{[x\in R_1]} + 0{,}0053262\ I_{[x\in R_2]} + 0{,}3166496\ I_{[x\in R_3]} \\ & +0{,}0108406\ I_{[x\in R_4]} + 0{,}8132295\ I_{[x\in R_5]} + 0{,}1870967\ I_{[x\in R_6]} \\ & +0{,}0887570\ I_{[x\in R_7]} + 0{,}7170305\ I_{[x\in R_8]} + 1{,}0000000\ I_{[x\in R_9]} \\ & +0{,}3684210\ I_{[x\in R_{10}]} + 0{,}9482758\ I_{[x\in R_{11}]} + 0{,}2941176\ I_{[x\in R_{12}]} \\ & +0{,}6619718\ I_{[x\in R_{13}]} + 0{,}2075471\ I_{[x\in R_{14}]} + 0{,}4444444\ I_{[x\in R_{15}]} \\ & +0{,}0958904\ I_{[x\in R_{16}]} + 0{,}3333333\ I_{[x\in R_{17}]} \end{aligned} \tag{7.124}$$

Cada una de las regiones de la partición binaria óptima obtenida por el algoritmo CART se define en función de relaciones de igualdad o desigualdad de determinadas variables, por ejemplo,

$$R_1 = \left\{ X \in \mathbb{R}^{11} / X53 \le 0\ \ y\ \ X48 \le 0\ \ y\ \ BF1\_X3 \le 85.80 \right\}$$

$$R_{17} = \left\{ X \in \mathbb{R}^{11} / X53 > 0\ \ y\ \ X48 > 8764\ \ y\ WR\_X55 > -0{,}145281 \right\}$$

donde R_1 es el nodo 10 y R_{17} es el nodo 26.

7.9.1.3.- Entrenamiento del modelo SLPM.

Respecto del ajuste de la red ANN (5.91), $\text{logit}\left(P(Y=1/\ X=x)\right) = \beta_0 + \sum_{r=1}^{q} \beta_r \sigma\left(w_o + \sum_{j=1}^{p} w_{rj} X_j \right)$,

el conjunto completo de los parámetros viene dado por

$$\vartheta = (\beta_0, \boldsymbol{\beta}, \boldsymbol{w}) = \left\{ w_{0r}, w_r,\ \ r = 1,2...,q \right\} \cup \left\{ \beta_{0k}, \beta_k,\ \ k = 0,1 \right\}$$

$$\begin{aligned} &donde\ \ w_r = \left(w_{r1},...,w_{rp} \right),\ \ con\ \ q\times(p+1)\ \ elementos \\ &\beta_k = \left(\beta_{k1},...,\beta_{kq} \right),\ con\ \ 2\times(q+1)\ \ elementos \end{aligned}$$

por tanto, ϑ consta de $q\times(p+1)+2\times(q+1)$ parámetros. Se buscan los valores de los parámetros desconocidos que mejor ajustan el modelo a los datos de entrenamiento, es decir que minimizan el error, llamado entropía cruzada o desvianza,

$$R(\vartheta) = -\sum_{i=1}^{N}\sum_{k=0}^{1} y_{ik} \log\left(g_k(x_i) \right) \tag{7.125}$$

El modelo a ajustar viene dado por

$$\text{logit}\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{q}\beta_r\frac{1}{1+e^{-\phi\left(X,\boldsymbol{w}_r\right)}} \tag{7.126}$$

donde

$$\begin{aligned}\phi\left(X,\boldsymbol{w}_r\right)=&\ w_{0r}+w_{1r}\ U_1+w_{2r}\ U_2+w_{3r}\ \mathrm{U}_4+w_{4r}\ \mathrm{U}_6+w_{5r}\ \mathrm{U}_8+\ w_{6r}\ \mathrm{U}_9\\&+w_{7r}\ \mathrm{U}_{15}+w_{8r}\ \mathrm{U}_{16}+w_{9r}\ \mathrm{W\_U}_{19}+w_{10r}\ \mathrm{BF}_{1}\mathrm{\_V}_4+w_{11r}\ \mathrm{WT\_V}_4\end{aligned}$$

Para la estimación del modelo, (7.126), *de transmisión de información hacia adelante y una sola capa oculta* SLPM que en este caso construimos sobre una expansión híbrida de funciones de base, se aplicó el procedimiento *Multilayer Perceptron* de PASW® Statistics de IBM®, *con una capa oculta, función de activación sigmoide, capa de salida con variable dependiente estado de default, función de activación softmax y función de error entropía cruzada.*

7.9.1.4 Ajuste del modelo k-NN.

Para la estimación de la probabilidad de default por los k vecinos más próximos, $k-NN$, se utilizó el estimador (2.27),

$$\hat{P}\left(Y=1/\ X=x\right)=\frac{k_1\left(x\right)}{k} \tag{7.127}$$

dónde $k_1\left(x\right)=\sum_{x_i\in V_k\left(x\right)}I_{\{y_i=1\}}=\sum_{x_i\in V_k\left(x\right)}y_i$.

Se aplicó el *Módulo MBR* del sistema de minería de datos *Enterprise Miner*® de SAS®, *para los k vecinos más próximos en el sentido de la métrica Euclídea.*

7.9.1.5 Entrenamiento del modelo SVM.

El entrenamiento del modelo SVM, (5.103), $sign\left(P\left(Y=1/\ X=x\right)\right)=\beta_0+\sum_{r=1}^{N}\beta_r K\left(X,x_r\right)$, se realizó resolviendo el problema de optimización con restricciones de desigualdad (5.105) con funcional de regularización $J\left(\boldsymbol{\beta}\right)=\frac{1}{2}\beta^T\beta$, con parámetro de regularización $\lambda=0.1$, es decir, constante de capacidad $C=1/\lambda=10$:

$$\min_{\boldsymbol{\beta}} \ \frac{1}{\lambda}\boldsymbol{1}^T\boldsymbol{\xi}+\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} \qquad (7.128)$$

$$\text{sujeto a } \boldsymbol{\xi}\geq\vec{0} \ \ e \ \ Y\circ\left(\beta_0+\boldsymbol{\beta}^T\boldsymbol{K}\right)\geq\boldsymbol{1}-\boldsymbol{\xi}$$

donde $A\circ B$ es el producto de Hadamard de matrices, $\boldsymbol{K}$ es el núcleo usado para expandir las variables explicativas del riesgo de crédito al espacio agrandado de características, y $\boldsymbol{\xi}$ es el vector de parámetros para "tratar" datos no separables. En otros términos, minimizando la función de error $\frac{1}{2}\|\boldsymbol{\beta}\|^2+\frac{1}{\lambda}\sum_{i=1}^{N}\xi_i$, sujeta a las restricciones $\xi_i\geq\vec{0} \ \ e \ \ y_i\left(\beta_0+\sum_{r=1}^{N}\beta_r K\left(x_i,x_r\right)\right)\geq 1-\xi_i, \ \ i=1,\ldots,N$, llamado problema de clasificación SVM tipo I en el Sistema Statistica® de StatSoft®. Nótese que cuanto mayor sea el parámetro de capacidad que utilicemos más se penaliza el error.

Como funciones núcleo hemos considerado, por un lado, la más popular, la función de base radial (RBF), $K\left(x_r,x_i\right)=exp\left(-\gamma\left\|x_r-x_i\right\|^2\right)$, y, por otro, la función sigmoide, $K\left(x_r,x_i\right)=\tanh\left(\gamma x_r x_i+c\right)$ (siendo tanh la tangente hiperbólica), en ambos casos con $\gamma=0{,}091$ y la constante $c=0$ en el núcleo sigmoide. En el entrenamiento del modelo se han obtenido claramente mejores resultados con el núcleo sigmoide, con una tasa de clasificación correcta del 96,968% frente al 80,072 % con RBF. Como consecuencia, utilizaremos el núcleo sigmoide para expandir las variables explicativas del riesgo de crédito al espacio agrandado de Hilbert de las características, para lo que se aplicó el procedimiento *Machine Learning (Support Vector)* del Sistema Statistica® de StatSoft®.

Tabla 7.63.- Especificaciones del modelo SVM.

Especificaciones del modelo	Valor
Tipo SVM	Clasificación tipo 1
Núcleo	Sigmoide
Nº de Vectores Soporte*	1395 (1387 frontera)
Nº de Vectores Soporte (-1)	698
Nº de Vectores Soporte (+1)	697
Constante de Decisión	$\beta_0=4{,}564892$

Software: Sistema Statistica® de StatSoft®.

* Puntos asociados a los multiplicadores de Lagrange $\alpha_i>0$.

El modelo resultante adopta la forma (parámetro $\hat{\beta}_0 = 4{,}564892$, $\gamma = 0{,}091$, $c = 0$):

$$Sign\left\{\hat{P}(Y=1/\ X=x)-\frac{1}{2}\right\} = \hat{\beta}_0 + \sum_{i=}^{n}\hat{\beta}_i K(x_i,x) = \hat{\beta}_0 + \sum_{i=1}^{n} y_i\alpha_i K(x_i,x)$$
$$= 4{,}564892 + \sum_{i=1}^{1395} y_i\alpha_i \tanh(0{,}091 x_i x) \qquad (7.129)$$

donde $\{(x_i,y_i)\in\mathbb{R}^p\times\{-1,+1\}\}_{i=1,\ldots,n}$, son los elementos de la muestra de entrenamiento para los puntos vector soporte, es decir, aquellos para los cuales los multiplicadores de Lagrange verifican $\alpha_i > 0$.

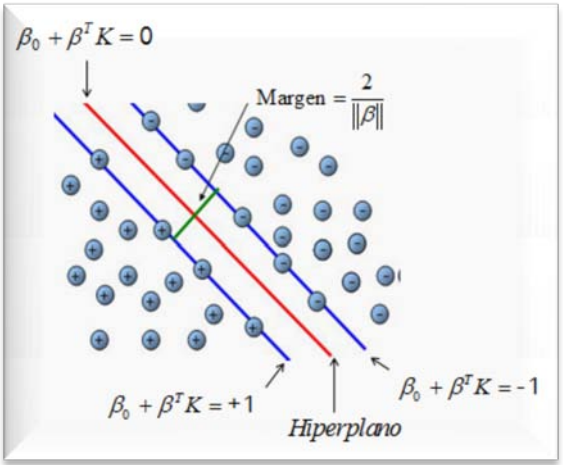


Figura 7.31.- Maquinas de Vector Soporte.

7.9.1.6 Estadísticos de Ajuste y Generalización.

En este apartado se muestran para todos los modelos analizados las tablas con los estadísticos de bondad de ajuste, el error empírico, los estadísticos de discriminación y los porcentajes de clasificación incorrecta, tabla 7.64, el área bajo la curva ROC, AUC, junto con la significación asintótica para hipótesis nula: área verdadera = 0,05 y los límites de confianza asintóticos con un nivel de significación α=0,05, tabla 7.65, las tasas de clasificación para los datos de entrenamiento y un punto de corte de Pc = 0,50, tabla 7.66, los ranking del poder discriminante y predictivo, AUC, y de la tasa de clasificación incorrecta sobre la muestra de entrenamiento, tabla 7.67, y, por último, el error test y la tasa de clasificación incorrecta para una muestra test independiente de 18.297 acreditados que no han intervenido en el entrenamiento, tabla 7.68.

Desde el punto de vista estadístico los modelos reducidos HLLM, y HLPM , presentan, como se muestra en la tabla 7.64, un buen ajuste, coeficiente de Nagelkerke superior a 0.80, un alto poder discriminante y predictivo, AUC muy próximo a 0,98, y alta eficacia como clasificadores, porcentaje de clasificados correctos superior a 98 % en ambos casos, en definitiva son modelos que ajustan bien a los datos y en el ajuste tienen un comportamiento muy similar.

Por otro lado, los modelos TREE, y SLPM, con un poder discriminante y predictivo inferior al de los dos modelos anteriores en los datos de entrenamiento, presentan además mayores tasas de clasificación incorrecta en esta muestra. Por su parte el modelo k-NN, presenta un alto valor discriminante y predictivo 0,9899 pero arroja sobre los acreditados de entrenamiento una tasa de clasificación errónea superior a los cuatro primeros modelos. Por último el modelo SVM, alcanza la mayor tasa de clasificación errónea de los acreditados de entrenamiento.

Tabla 7.64.- Estadísticos de bondad de ajuste, error empírico, estadísticos de discriminación y porcentajes de clasificación incorrecta sobre la muestra de entrenamiento para los 6 modelos analizados.

	HLLM	**HLPM**	**TREE**	**SLPM**	**K-NN**	**SVM**
$-LogVer\_Dif$	7433,0455	7485,0455				
$-LogVer\_Tot$	2342,2770	2290,2770				
$-LogVer\_\mathrm{Red}$	9775,3225	9775,3225				
$R^2(U)$	0,7665	0,7657				
R^2	0,3359	0,3356				
$\tilde{R}^2$	0,8116	0,8112				
$-2log(L)$	4564,4760	4580,5540				
Error Empírico	0,1247	0,1251	0,220*	242891,3**	0,0177***	
AIC	4586,4760	4602,5540			122192,250	
BIC	4680,0639	4696,1419			122098,660	
AUC	0,9789	0,9794	0,9721	0,9708	0,9899	
% Clasificación Incorrecta	1,91%	1,94%	2,21%	2,24%	2,34	3,03%

* Riesgo.

** Error de entropía cruzada.

*** Error medio cuadrático, SME.

Tabla 7.65.- Probabilidad Pronosticada, (área bajo la curva ROC) sobre la muestra de entrenamiento para HLLM, HLPM, TREE, SPLM y k-NN.

Modelo	AUC	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95%	
				Límite inferior	Límite superior
HLLM	0,9789	0,0017	0,0000	0,9756	0,9822
HLPM	0,9794	0,0017	0,0000	0,9762	0,9826
TREE	0,9721	0,0020	0,0000	0,9683	0,9758
SLPM	0,9708	0,0020	0,0000	0,9669	0,9746
k-NN	0,9899	0,0050	0,0000	0,9890	0,9909

a. Bajo el supuesto no paramétrico.
b. Hipótesis nula: área verdadera = 0,05.

Tabla 7.66.- Tabla de Clasificación, (para un punto de corte de 0,50), sobre la muestra de entrenamiento para HLLM, HLPM, TREE, SPLM , k-NN y SVM.

Modelo	Correcto		Incorrecto		Porcentajes				
	Default	No Default	Default	No Default	Correcto	Sensi-bilidad	Especi-ficidad	Falso POS	Falso NEG
HLLM	2324	33582	270	431	98,09	84.4	99.2	10.4	1.3
HLPM	2334	33564	288	421	98,06	84,7	99,1	11.0	1.2
TREE	2350	33450	405	402	97,80	85,4	98,8	14.7	1.2
SPLM	2268	33522	487	330	97,77	87,3	98,6	17,7	1.0
k-NN	2079	33672	676	180	97,66	92,0	98,0	8.0	1.2
SVM	2.131	33.366	624	486	96,97	81,4	98,2	22.7	1,4

Como puede observarse en la tabla de ranking del poder discriminante y predictivo, AUC, y de la tasa de clasificación incorrecta sobre la muestra de entrenamiento, tabla 7.67, los modelos que mejor se sitúan con respecto a ambas características en conjunto son HLLM y HLPM.

Tabla 7.67.- Ranking del poder discriminante y predictivo, AUC, y de la tasa de clasificación incorrecta, ambos sobre la muestra entrenamiento de los modelos HLLM, HLPM, TREE, SPLM , k-NN y SVM.

Estadístico	Rango
AUC de Entrenamiento	(0,9899) $M_{11\_IV} \overset{-0,0105}{\succ} M_{11\_II} \overset{-0,0005}{\succ} M_{11\_I} \overset{-0,0068}{\succ} M_{11\_IIII} \overset{-0,0013}{\succ} M_{11\_IV}$ (0,9708)
% de clasificación incorrecta	(1,91 %) $M_{11\_I} \overset{+0,03}{\prec} M_{11\_II} \overset{+0,27}{\prec} M_{11\_III} \overset{+0,03}{\prec} M_{11\_IV} \overset{+0,10}{\prec} M_{11\_V} \overset{+0,69}{\prec} M_{11\_VI}$ (3,03 %

Tabla 7.68.- Error de generalización y tasa de clasificación incorrecta para los modelos HLLM, HLPM, TREE, SPLM , k-NN y SVM. Muestra test N=18.297.

	HLLM	HLPM	TREE	SPLM	k-NN	SVM
Error Test	0,1227	0,1244				
% Clasificación Incorrecta	1,90%	1,95%	2,20%0	2,23%	2,34%	3,03%

Por lo que respecta al comportamiento de los 6 modelos sobre la muestra test, los modelos HLLM y HLPM poseen un error de generalización, muestra test, inferior al error empírico de entrenamiento y las tasas de clasificación incorrecta son similares a las de entrenamiento para los seis modelos, incluso conservan el mismo orden en el ranking por este concepto que para la muestra de entrenamiento.

7.9.2 Poder Discriminante de los modelos HLLM, HLPM, TREE, SLPM, k-NN y SVM.

Con el objetivo de validar cuantitativamente el poder discriminante de los modelos que venimos analizando, sobre la población total de acreditados del año 2007, una vez que se han obtenido los estimadores de los parámetros de cada modelo se calculan las probabilidades de default para cada acreditado y su puntuación correspondiente. De este modo podremos comparar los valores pronosticados con los valores realmente observados y

medir el rendimiento del modelo $\mathrm{g}\left(\hat{P}\left(Y=1/\ X=x\right)\right)=\hat{\beta}_0+\sum_{r=1}^{q}\hat{\beta}_r h\left(\boldsymbol{X}\right)=\hat{\beta}_0+\hat{\boldsymbol{\beta}}^T\boldsymbol{H}\left(\boldsymbol{X}\right)$. De forma análoga a como se procedió para el modelo HLLM, M_{11_I}, procederemos para los otros cinco modelos, HLPM, TREE, SLPM, k-NN y SVM.

En primer lugar, con el razonable fin de conseguir que a los clientes calificados por el modelo como malos (con una probabilidad de incumplimiento elevada) les corresponda una puntuación muy baja y, recíprocamente, a los clientes mejores (con una probabilidad de incumplimiento pequeña) les corresponda una puntuación elevada, es decir,

$$P\left(Y=0/\ X=x_1\right)<P\left(Y=0/\ X=x_1\right)\Rightarrow \mathrm{S}\left(x_1\right)\leq\ \mathrm{S}\left(x_2\right)$$

y , además, con los objetivos de posibilitar la comparación de la puntuación otorgada por diferentes modelos y no puntuar a ningún cliente con valores bajos, por cuestiones de imagen, se asigna una puntuación mínima. Para conseguir todo esto es necesario aplicar un *cambio de localización y escala a la función de calificación. Procediendo en forma análoga a como se hizo para el modelo* M_{11_I} se obtiene la puntuación reescalada obtenida según la expresión $S(x)=mi+c*\left(1-P\left(Y=1/\ X=x\right)\right)$.

En segundo lugar, puesto que el objetivo fundamental de un modelo de credit scoring es discriminar entre acreditados buenos y acreditados malos, una vez obtenida la nueva función de calificación re-escalada, procedemos a validar el *poder discriminante* del correspondiente modelo.

Para validar el poder discriminante de los modelos a comparar se calculan:

- A partir de las matrices de confusión, (*valores de default observados para todos los acreditados en noviembre de 2007 por valores de default pronosticados para los mismos acreditados por el correspondiente modelo en la misma fecha*), los *estadísticos:* **Error de Tipo I** o **Riesgo de Crédito**, (se estimó cumplimiento y en la actualidad el acreditado incumple con sus obligaciones de pago) y **Error Tipo II** o **Coste de Oportunidad o de Pérdida de Negocio**, (los 'buenos' clientes mal clasificados son rechazados). Ambos estadísticos se muestran para todos los modelos a comparar en las filas 1 y 2 de la tabla 7.69.
- A partir de las funciones de densidad y distribución acumulada de las poblaciones de "buenos" y "malos" clientes, la ***diferencia más extrema absoluta entre las funciones de distribución acumulativas*** de las poblaciones de buenos y malos, $\left|\hat{D}_{N_0N_1}\right|$, su

estadístico de Kolmogorov-Smirnov *asociado,* $\hat{T} = \left|\hat{D}_{N_0 N_1}\right| \sqrt{\frac{N_0 N_1}{N_0 + N_1}}$, la **significación asintótica bilateral del test sobre la distribución uniforme de ambas poblaciones** y la **tasa porcentual de clasificación incorrecta**, filas 3, 4, 5 y 6 de la tabla 7.69, respectivamente.

En los paneles izquierdos de las figuras 7.32 y 7.33, se muestra para cada modelo la representación gráfica de las funciones de densidad de las puntuaciones asignadas por la función de acreditados de las poblaciones de buenos acreditados, en azul, y de malos acreditados, en rojo, y en los paneles de la derecha las funciones de distribución acumuladas correspondientes.

- La **Tasa de Precisión**, **AR**, resumen de la Curva de Lorenz, asociada a cada modelo de credit scoring, que mide la *calidad como función discriminante entre acreditados default y no default,* fila 7 de la tabla 7.69.

- El **área bajo la curva ROC**, **AUC**, asociada a cada modelo, fila 8 de la tabla 7.69. En la tabla 7.69 se muestran los errores típicos, la significación asintótica y el intervalo de confianza asintótico a un nivel de confianza del 95%. La representación gráfica de las curvas ROC se muestra en la figura 7.34.

- El test no paramétrico ***suma de rangos de Wilcoxon*** **y** ***U de Mann-Whitney*** para contrastar si las distribuciones de las poblaciones de default y no default son o no idénticas. El test conlleva los estadístico **U de Mann-Whitney**, **W de Wilcoxon**, la **aproximación asintótica Z** y la **significación asintótica bilateral del test sobre la igualdad de las distribuciones acumuladas de ambas poblaciones** y la **tasa porcentual de clasificación incorrecta**, filas 9, 10, 11 y 12 de la tabla 7.69, respectivamente.

Dado que las curvas ROC y, por tanto, la medida AUC solo pueden obtenerse para modelos que producen una probabilidad o puntuación, para un modelo discreto como SVM, que sólo produce una decisión de clase ("default", "no default"), sólo contamos con la *matriz de confusión,* tabla de contingencia con dos efectos binarios, la clase pronosticada y la clase observada. No es posible, por tanto, realizar los tests d*e Kolmogorov-Smirnov y Mann-Whitney,* diseñados para variables continuas y este no es el caso del modelo SVM; para abordar la cuestión de si *el modelo discrimina bien las poblaciones de default y no default*, por lo que es necesario utilizar un test alternativo, *Kappa de Cohen,* (COHEN, 1960, 1968), que nos permite medir el *grado de acuerdo entre el efecto valor pronosticado por el modelo*

para los acreditados en el año 2007 y el valor realmente observado en dicha fecha, es decir, que nos permite medir el *poder discriminante y la precisión del modelo.*

El test *Kappa de Cohen* se basa en el hecho de que se puede concebir "el modelo", que pronostica el estado de default, y "la realidad", que proporciona el estado de default observado, como dos "observadores" que clasifican independientemente una muestra de individuos sobre la misma escala categórica. En este caso las clasificaciones conjuntas de ambos observadores se representan en la tabla de confusión con las mismas categorías en cada dimensión, , default y no default, cuya diagonal principal representa los casos en los que hay acuerdo.

El concepto de *acuerdo* es distinto del concepto de *asociación*, de hecho para que las clasificaciones de los dos observadores estén de acuerdo deben caer en categorías idénticas mientras que para que dos respuestas estén perfectamente asociadas solo se requiere que se pueda predecir la clasificación de un observador a partir de la del otro. Por lo tanto una tabla en la que hay asociación alta puede haber acuerdo alto o bajo.

Nosotros estamos interesados en verificar si las clasificaciones de los dos observadores caen en categorías idénticas, es decir, que *ambos estén de acuerdo.* Esta cuestión está directamente relacionada con el poder discriminante del modelo, puesto que el grado de acuerdo entre el pronóstico del modelo y el default realmente observado indica el poder del modelo para discriminar entre default y no default.

Cohen introdujo su estadístico *kappa* como una función de la diferencia entre la probabilidad de que los dos observadores estén de acuerdo y la probabilidad de acuerdo si sus clasificaciones conjuntas fuesen independientes en los términos siguientes:

Si denotamos por $p_a = \sum_{i=1}^{nc} p_{ii}$ la probabilidad de acuerdo, siendo *nc* el número de categorías, y por $p_a^I = \sum_{i=1}^{nc} p_{i.} p_{.i}$ la probabilidad de acuerdo bajo independencia (probabilidad de acuerdo por azar), el valor poblacional de la medida *kappa* es

$$\kappa = \frac{p_a - p_a^I}{1 - p_a^I} \tag{7.130}$$

La independencia implica claramente que $\kappa = 0$. El valor cero de esta medida implica que el acuerdo es igual al esperado por azar (ausencia de acuerdo). Si hay acuerdo perfecto, $p_a = \sum_{i=1}^{nc} p_{ii} = 1$, entonces $\kappa = 1$. Kappa podría tomar valores negativos raras veces cuando el acuerdo es menor que el debido al azar. El valor muestral de *kappa* es

$$\hat{\kappa} = \frac{n_c \sum_{i=1}^{n_c} n_{ii} - \sum_{i=1}^{I} n_{i.} n_{.i}}{n_c^{\,2} - \sum_{i=1}^{I} n_{i.} n_{.i}} \tag{7.131}$$

En la práctica raramente ocurre que el acuerdo no sea mayor que el esperado por azar. Por ello será más importante construir un intervalo de confianza para medir la longitud del acuerdo que contrastar si el acuerdo es cero.

En EVERITT (1992) se recogen algunas reglas empíricas útiles para evaluar el grado de acuerdo en base al valor de la medida *kappa*,

Kappa	Dimensión del Acuerdo
0	Pobre
0.00 - 020	Pequeño
0.21 - 0.40	Mediano
0.41 - 0.60	Moderado
0.61 - 0.80	Importante
0.81 - 1.00	Casi perfecto

El estimador $\hat{\kappa}$ del estadístico *kappa de Cohen* para el modelo SVM M_{11_VI} se muestra en la fila 13, columna correspondiente al modelo de la tabla 7.69.

Valor Kappa=0,746, (entre 0,61 y 0,80), según la regla empírica de Everitt el grado de acuerdo es *importante.*

La primer conclusión que se obtiene de la tabla 7.69, columnas 1 y 2, es que los *errores tipo I, riesgo de crédito*, son en todos los modelos suficientemente pequeños, mejores en los modelos TREE, 1,13%, HLPM, 1,14%, y HLLM, 1,18%, y algo peores en los modelos SLPM, 1,60%, k-NN, 1,76%, y SVM, 1,85%.

Por lo que se refiere al *error tipo II, pérdida de negocio*, los valores son bastante más altos que para el error tipo I. El error tipo II más pequeño se alcanza para el modelo k-NN, 11,80%, seguido muy de cerca por los modelos HLLM, 12,84%, y HLPM, 13,64%, TREE y SLPM se acercan casi al 16% y el mayor "pérdida de negocio" es SVM con el 24,19 que prácticamente dobla al error tipo II que presentan k-NN, HLLM y HLPM.

A pesar de que la mejor situación de ambos errores en conjunto la presentan los modelos HLPM y HLLM, debemos tener en cuenta que no es posible obtener un error total de clasificación debido a la desigual importancia de ambos tipos de error, por lo que es necesario para medir el poder discriminante utilizar la medida alternativa diferencia más extrema absoluta, $\left|\hat{D}_{N_0 N_1}\right|$, que lleva asociado el test de Kolmogorov-Smirnov con estadístico

de contraste $\hat{T} = \left|\hat{D}_{N_0 N_1}\right| \sqrt{\frac{N_0 N_1}{N_0 + N_1}}$. Ambos estadísticos y la significación asintótica se muestran en las filas 3 y 4 de la tabla 7.69. El test resulta significativo para todos los modelos con p-valor <0,0001, es decir, para todos ellos *se rechaza que las puntuaciones de las poblaciones de default y no default se distribuyan uniformemente,* lo que quiere decir que *todos los modelos analizados por el test discriminan bien ambas poblaciones*. Claro que unos mejor que otros, de hecho los modelos que presentan el estadístico de Kolmogorov-Smirnov, $\hat{T}$, mayor son HLLM y HLPM. En las figuras 7.32 y 7.33 se representan gráficamente los perfiles de las *funciones de densidad*, f.d., y *distribución acumulada,* f.d.a.*,* de las puntuaciones para las poblaciones de buenos y malos acreditados, para los modelos HLLM, HLPM, TREE, SPLM y k-NN.

Por lo que respecta a la *tasa de clasificación de acreditados incorrecta*, fila 6 de la tabla 7.69, los valores más bajos de nuevo se alcanzan para HLLM, 2,04%, y HLPM, 2,09%, y el más alto SVM, 3,57%.

Por otro lado, el indicador de poder discriminante *área bajo la curva ROC*, AUC, fila 8 tabla 7.69, está muy próximo a 0,98 en el caso de los modelos HLLM y HLPM y descendiendo y por este orden TREE, SLPM y k-NN, para el que AUC se sitúa en torno a 0,95 que comparativamente es un valor bajo, como cabía esperar de este modelo tendente al sobreajuste de los datos. Tampoco es bueno el AUC del modelo SLPM, un comportamiento mejor que este último modelo lo tiene TREE pero está en clara desventaja con respecto a los modelos lineales híbridos, logístico y probit. En la misma línea se comporta la tasa de precisión de los modelos, AR, fila 7 de la tabla 7.69.

En la tabla 7.70 se muestran los estadísticos AUC y los intervalos de confianza para el default pronosticado sobre los acreditados a 30 de noviembre de 2007 para los cinco modelos analizados que proporcionan probabilidades de default. La representación grafica conjunta de las curvas ROC para tales modelos se muestra en la figura 7.34.

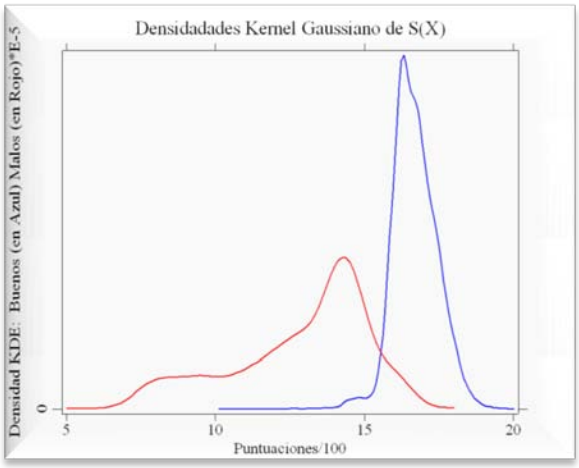
Para poder comparar el poder discriminante del modelo SVM con el de los modelos anteriores es necesario hacer uso del estadístico *kappa de Cohen* . Como puede observarse, última fila de la tabla 7.69, el grado de acuerdo entre la probabilidad de default pronosticada por el modelo y el estado de default observado es el menor para SVM, con notable diferencia con el de los demás modelo, sobre todo y en este orden, con HLLM, HLPM y TREE.

Por último utilizamos el potente test de Wilcoxon-Mann-Witney, herramienta idónea para comparar el poder discriminante de dos o más modelos entre sí. Este test contrasta si las distribuciones de las poblaciones de default y no default son o no idénticas.

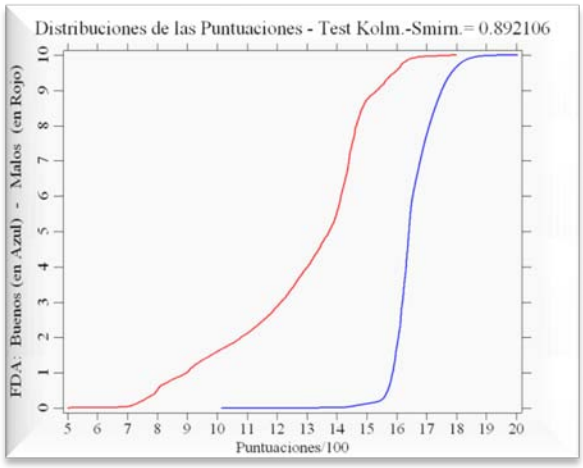
Los estadísticos U de Mann-Witney, W de Wilcoxon, Z de Mann-Witney-Wilcoxon y la significación asintótica de Z se muestran en las filas 9, 10, 11 y 12 de la tabla 7.69. También este test resulta significativo para todos los modelos con significación asintótica $\text{Prob}>Z_{1-\alpha} < 0.0001$, es decir, para todos ellos se rechaza la hipótesis nula de la *igualdad de las distribuciones acumuladas de las poblaciones de default y no default*, de este modo contamos con un instrumento de contraste más, posiblemente el más idóneo, que nos permite afirmar que *los modelos HLLM,HLPM, TREE, SPLM, k-NN discriminan bien ambas poblaciones*.

Tabla 7.69.- Estadísticos y test de validación del poder discriminante y precisión en la clasificación de los seis modelos analizados sobre la población total de acreditados a 30 de noviembre de 2011.

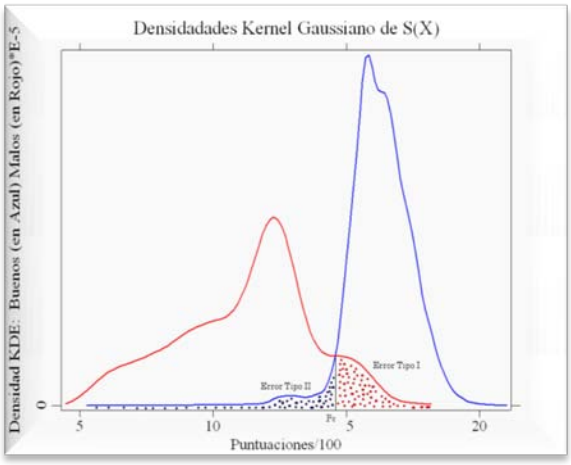
	HLLM	**HLPM**	**TREE**	**SLPM**	**K-NN**	**SVM**
Error Tipo I	1,18%	1,14%	1,13%	1,60%	1,76%	1,85%
Error tipo II	12,84%	13,64%	15,51%	15,42%	11,80%	24,19%
Test de Kolmogorov_Smirnov						
$\left\|\hat{D}_{N_0N_1}\right\|$	0,8921	0,8933	0,9028	0,8584	0,8743	
$\hat{T}=\left\|\hat{D}_{N_0N_1}\right\|\sqrt{\frac{N_0N_1}{N_0+N_1}}$	63,6744	63,7640	63,4208	61,2690	62,40	
Sig. Asintótica	<0,0001	<0,0001	<0,0001	<0,0001	<0,0001	
% Clasificación Incorrecta	2,04%	2,09%	2,24%	2,59%	2,43%	3,57%
AR	0,9558	0,9572	0,9460	0,9309	0,9018	
AUC	0,9779	0,9786	0,9730	0,9654	0,9509	
Test de Wilcoxon_Mann_Whitney						
U de M_W	8228194,00	9777242,00	1,008E7	1,289E7	0,548E7	
W de Wilcoxon	2,341E7	3,315E7	2,526E7	2,807E7	2,023E7	
Z	-118,170	-118,3370	-124,6020	-115,0790	-176,099	
Sig. Asintótica	<0,0001	<0,0001	<0,0001	<0,0001	<0,0001	
Test kappa de Cohen						
$\hat{\kappa}$	0,852	0,850	0,841	0,810	0,816	0,746



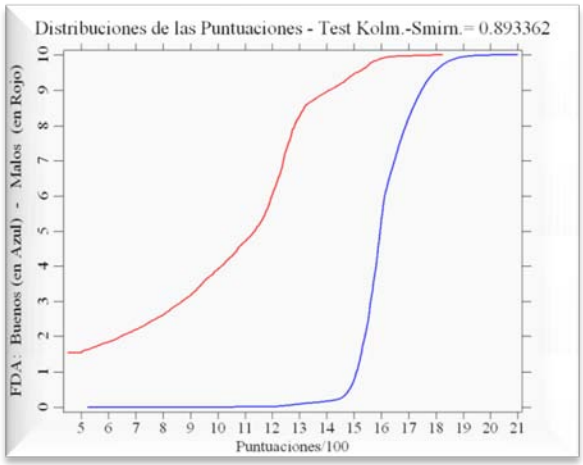
f.d M_{11_I} HLLM



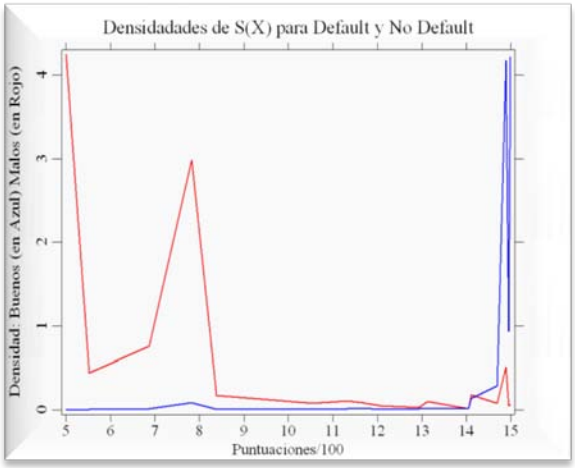
f.d.a M_{11_I} HLLM



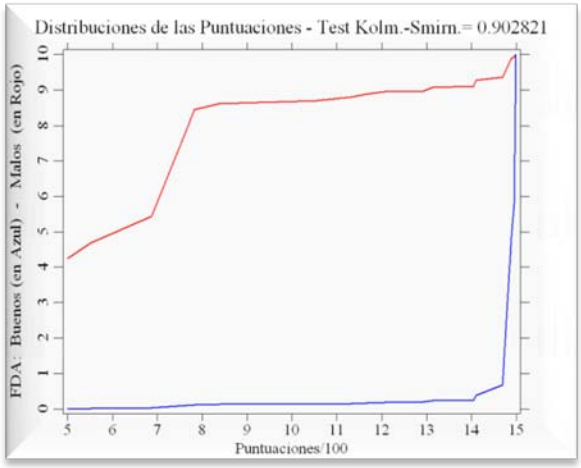
f.d HLPM



f.d.a HLPM

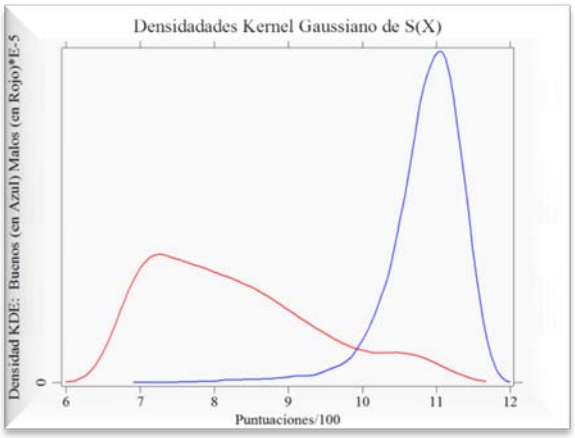


f.d TREE

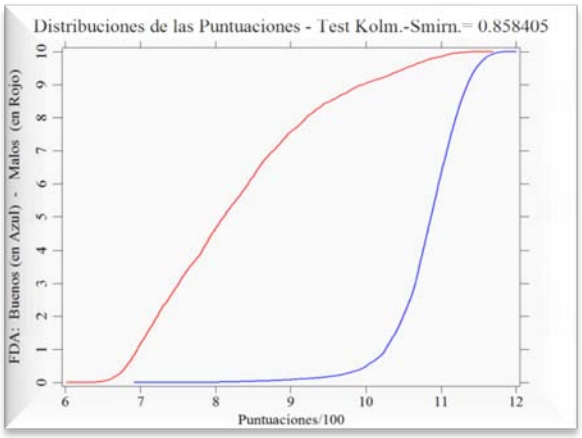


f.d.a TREE

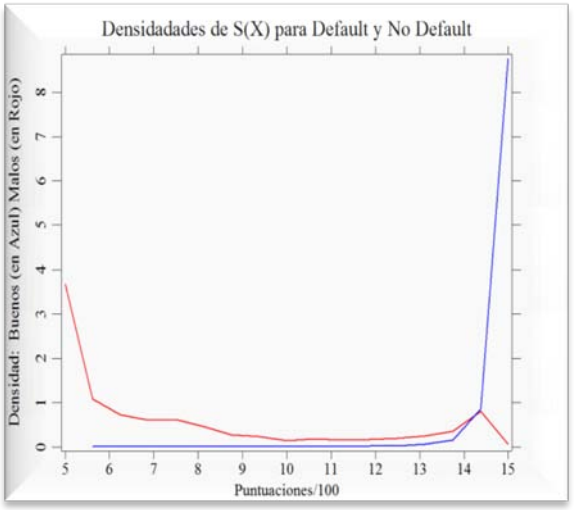
Figura 7.33.- Perfiles de las *funciones de densidad*, f.d., y *distribución acumulada,* f.d.a. de puntuaciones para las poblaciones de buenos y malos acreditados, para los modelos, HLLM, HLPM, TREE. Test de Kolmogorov-Smirnov. Conjunto de acreditados default en rojo y conjunto de acreditados no default en azul.



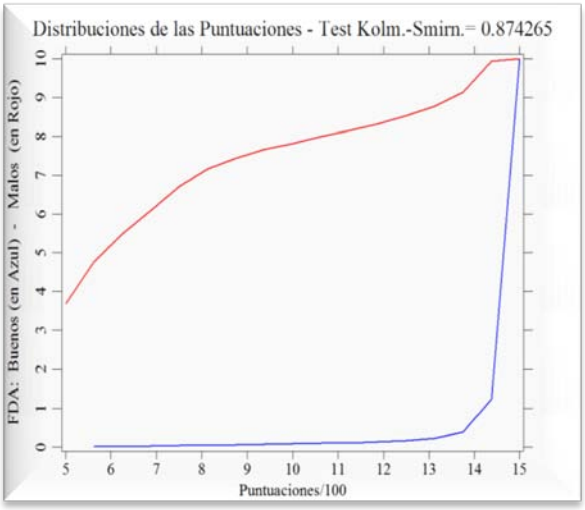
f.d SLPM



f.d.a SLPM



f.d k-NN



f.d.a k-NN

Figura 7.34.- Perfiles de las *funciones de densidad*, f.d., y *distribución acumulada,* f.d.a. de puntuaciones para las poblaciones de buenos y malos acreditados, para los modelos, SPLM y k-NN. Test de Kolmogorov-Smirnov. Conjunto de acreditados default en rojo y conjunto de acreditados no default en azul.

Salvo para el caso del modelo TREE y el modelo k-NN, las funciones de densidad representadas en las figuras 7.32 y 7.33 se estimaron por estimadores de la densidad por núcleos Gaussianos. Respecto de SVM al no proporcionarnos el modelo la probabilidad de default no disponemos de los perfiles de la densidad f.d..

Como síntesis del análisis de todos los estadísticos mostrados en la tabla 7.69 podemos decir que: **Todos los modelos presentan un importante poder discriminante, el menor lo presenta SVM y los mayores, y en ese orden, *HLPM, HLLM, cerca de 0.98. Por otro lado HLLM y HLPM presentan las menores tasas de clasificación incorrecta, en torno al 2%.***

Tabla 7.70.- Área bajo la curva ROC e intervalos de confianza para el default pronosticado sobre los acreditados a 30 de noviembre de 2007 para los cinco modelos analizados que proporcionan probabilidades de default.

				Intervalo de confianza asintótico al 95%	
Modelo	AUC	Error típ.[a]	Sig. asintótica[b]	Límite inferior	Límite superior
HLLM	0,9789	0,0017	0,0000	0,9756	0,9822
HLPM	0,9786	0,0012	0,0000	0,9762	0,9810
TREE	0,9730	0,0013	0,0000	0,9703	0,9756
LSPM	0,9654	0,0016	0,0000	0,9623	0,9688
k-NN	0,9509	0,0022	0,0000	0,9465	0,9552

a. Bajo el supuesto no paramétrico. Hipótesis nula: área verdadera = 0,05.

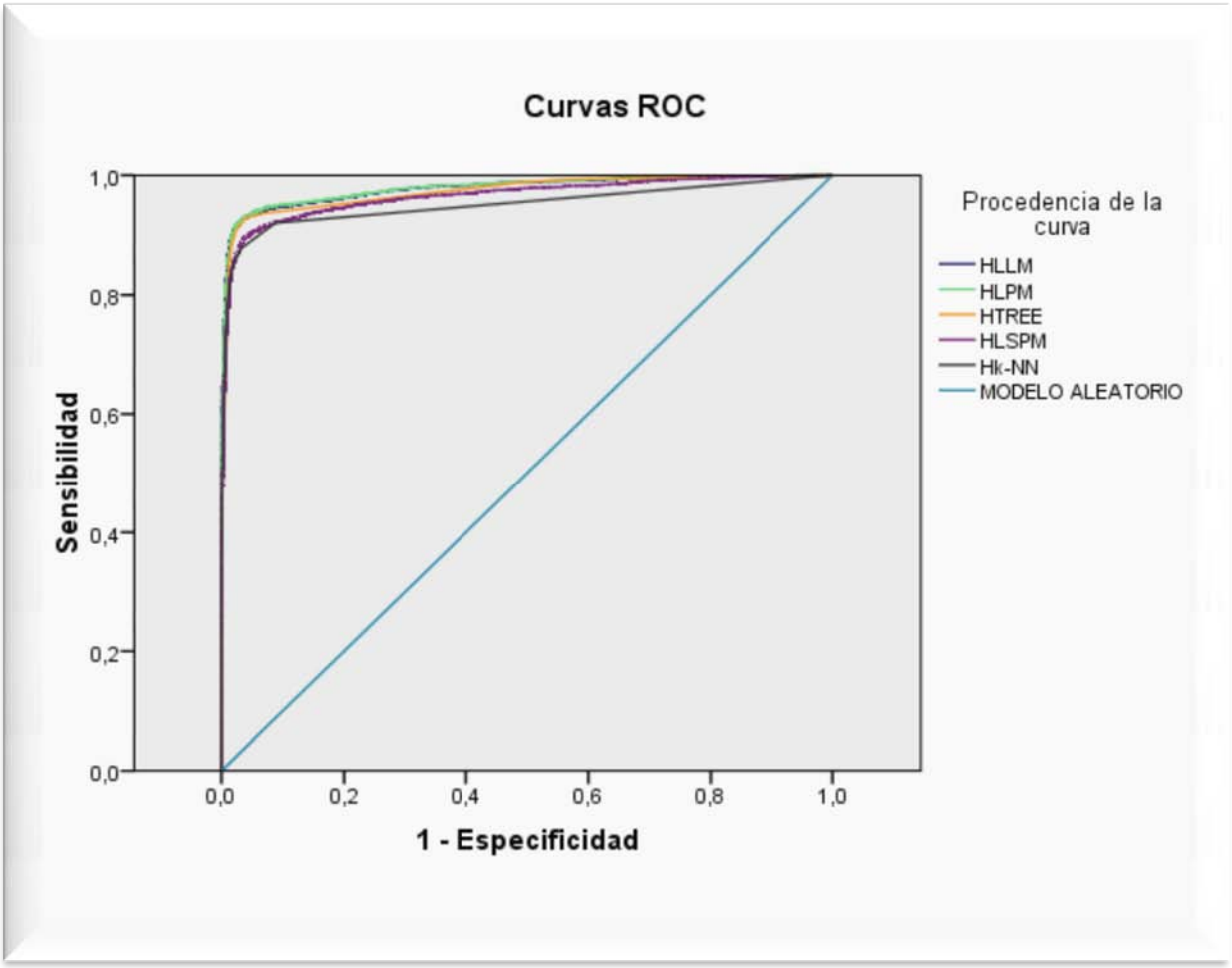


Figura 7. 34.- Curvas ROC para el default pronosticado sobre los acreditados a 30 de noviembre de 2007 para los cinco modelos analizados que proporcionan probabilidades de default.

7.9.3 Calibración de los modelos HLLM, HLPM, TREE, SLPM, k-NN y SVM.

En la sección anterior, hemos analizado la capacidad de los modelos a comparar para separar a los acreditados buenos de los malos, es decir su poder discriminante, requisito imprescindible para que el modelo sea lo suficientemente eficaz en la consecución de sus objetivos, análisis que hemos realizado utilizando los datos observados sobre el total de acreditados a 30 de noviembre de 2007. El segundo aspecto cuantitativo a validar es la *calibración*, entendida la calibración como la corrección de la *probabilidad de default,* PD.

La cuestión que abordaremos en esta subsección es si los hechos acaecidos a posteriori , (default observado a 30 de noviembre de 2008), respaldan los pronósticos a priori, default pronosticado a 30 de noviembre de 2007), es decir, comprobar si la probabilidad de default pronosticada por el modelo de credit scoring al inicio del período temporal debe ser revisada dadas las tasas de default realmente observadas un año después.

Tal como ya hicimos para el modelo HLLM, para cada uno de los cinco modelos restantes obtenemos en primer lugar las categorías de calificación de los acreditados utilizando el algoritmo CART para generar las clases y, a partir de estas se calculan las distribuciones de default y no default así como las probabilidades pronosticadas en 2007 para cada modelo. Por último se calculan las tasas de default observadas un año después.

Por las mismas razones apuntadas para medir el poder discriminante del modelo, en el marco de la determinación de los requerimientos de capital, una matriz de confusión no es una herramienta de calibración útil, e igual que allí habremos de proceder a considerar las adecuadas alternativas que, al igual que para HLLM, serán los tres tests condicionados al estado de la economía, que contrastan estimaciones en un *punto del tiempo*, (PITT), el *Test Binomial de una y dos caras (y su aproximación normal), el Test* χ^2 *de Hosmer-Lemeshow y el de Test de Spiegelhalter* .

Tabla 7.71.- Test binomial de una y dos caras, test chi-cuadrado, test de Spielhalter y kappa de Cohen para la **calibración** de los modelos HLLM, HLPM, TREE, SLPM, k-NN y SVM. Las probabilidades de default se pronosticaron a 30 de noviembre de 2007 y los default se observaron a 30 de noviembre de 2008.

	Test Binomial				Test Chi-Cuadrado			Test de Spiegelhalter		Kappa de Cohen
	Una Cara		Dos caras							
Modelo	$P_r^{obs} < k^*$	$P_r^{obs} \geq k^*$	$P_r^{obs} \in R_a$	$P_r^{obs} \in R_C$	$\hat{t}_R$	R	$\chi^2_{R,0.005}$	$\hat{Z}_S$	$Z_{0.05}$	$\hat{\kappa}$
HLLM	1,2,6	3,4,5,7,8	2,6	1,3,4,5,7,8	474,78	8	15,51	25,9519	1,96	0,852
HLPM	1,2	3,4,5,6,7,8		1,2,3,4,5,6,7,8	474,78	8	15,51	27,0192	1,96	0,850
TREE	1,2	3,4,5		1,2,3,4,5	497,22	5	11,07	29,0563	1,96	0,841
SPLM	1,2	3,4,5,6,7,8	2,4	1,3,5,6,7,8	2990,02	8	15,51	13,7043	1,96	0,810
k-NN	1,2,3,4,5	6	1,6,7	2,3,4,5	627,61	6	12,59	26,9278	1,96	0,816
SVM										0,657

1.- Si $P_r^{obs} \geq k^*$ se rechaza la hipótesis nula H_0: $P_r = \hat{P}_r$, es decir, se rechaza que la probabilidad de default coincide con la probabilidad de default estimada en la clase r, frente a la alternativa de que está infra estimada en dicha clase.

2.- $R_a = \left[p_{\alpha/2} , p_{1-\alpha/2} \right]$ región de aceptación del test chi-cuadrado.

3.- $R_C = [0, p_{\alpha/2}) \cup (p_{1-\alpha/2}, 1]$ región crítica del test chi-cuadrado. Si $P_r^{obs} \in R_C$ se rechaza la hipótesis nula H_0: $P_r = \hat{P}_r$, es decir, se rechaza que la probabilidad de default coincide con la probabilidad de default estimada en la clase r.

4.- Siendo R el numero de categorías de calificación, si $t_R > \chi^2_{R,\alpha}$ se rechaza, al nivel de significación α, la hipótesis nula de la igualdad simultánea de las probabilidades pronosticadas y los valores observados, H_0: $P_1 = \hat{P}_1, ..., P_R = \hat{P}_R$.

5.- Si $\hat{Z} > Z_{0.05}$ se rechaza la hipótesis nula de que todas las probabilidades de default estimadas, $\hat{p}_i$, coinciden exactamente con la verdadera probabilidad de default $P(Y = 1 / X = x_i)$ para todo acreditado i, H_0: $\hat{p}_i = P(y_i = 1 / X = x_i) = E[Y / X = x_i]$, $i = 1, ..., N$.

El *test binomial de una cara*, columnas 2 y 3 de la tabla 7.71, nos indica que para todos los modelos con un nivel de significación $\alpha = 0{,}005$ se rechaza que la probabilidad observada en 2008 coincide con la probabilidad de default estimada en las clases 1 y 2 un año antes,

es decir "*la probabilidad de default pronosticada para las categorías de calificación 1 y 2 no se infraestimó"*, afirmación *que no podemos hacer para el resto de las categorías*. Este es un hecho importante por cuanto la clase 1 es la clase de default y la clase 2 es la siguiente clase en el ranking por bajas puntuaciones lo que interesa a la Institución Reguladora.

Para HLLM los comentarios del párrafo anterior son válidos también para la clase 6, la tercera en el ranking de puntuaciones de acreditados ordenadas de mayor a menor . Por otra parte, para el modelo SLPM *tampoco se rechaza la hipótesis nula para las clases de calificación 3, 4 y 5.*

Desde el punto de vista de lo que interesa a los analistas del riesgo de crédito, *test binomial de dos caras*, todos los modelos, a excepción del modelo k-NN, coinciden en rechazar que para la clase de default, 1, las probabilidades observadas en 2008 coinciden con las pronosticadas un año antes, frente a la alternativa de que no coinciden, es decir, solamente para el modelo más rugoso de todos el test binomial de dos caras no rechaza que coinciden las probabilidades de default pronosticadas con las observadas para la categoría de default un año después.

Salvo algunas excepciones, como se aprecia en la columna 4 de la tabla 7.71, en todos los modelos hay coincidencia en que el test binomial de dos caras rechace la hipótesis nula para prácticamente todas las clases, lo que viene respaldado totalmente por el *test chi-cuadrado* , columnas 6, 7 y 8 de la tabla 7.71, que al nivel de significación $\alpha = 0{,}005$ rechaza, $t_R > \chi^2_{R,\alpha}$, la hipótesis nula de la igualdad simultánea de las probabilidades pronosticadas en 2007 y los valores observados en 2008, y por el *test de Spiegelhalter*, columnas 9 y 10 de la tabla 7.71, que al nivel de significación $\alpha = 0{,}005$, también rechaza la hipótesis nula de la igualdad simultánea para todas las clases de las probabilidades pronosticadas en 2007 y los valores observados en 2008.

Para el modelo SVM para el que no son posibles los test binomial, chi-cuadrado y Spiegelhalter, por lo que utilizamos el estadístico kappa de Cohen, fila correspondiente al modelo, columna 11, que indica que a pesar de que el grado de acuerdo entre los valores de default pronosticados en 2007 y los observados en 2008 es importante, $\hat{\kappa} = 0{,}67$, no se aproxima lo suficiente a 1 como para concluir que las probabilidades de default están calibradas en el período de referencia.

Del estadístico de acuerdo de Cohen, calculado para el resto de modelos, se puede extraer la misma conclusión, si bien en todos los casos, sobre todo en HLLM y HLPM, el valor de $\hat{\kappa}$ es más alto.

Podemos concluir que *ninguno de los 6 modelos está calibrado adecuadamente*, pero esto no puede ser imputado a la construcción de los modelos sino al profundo cambio de las condiciones en torno al riesgo de crédito existentes en el desarrollo de los mismos, cambio que fue el mayor desde hacía más de 80 años. De hecho el problema de las subprime americanas estalló en junio de 2007 y tuvo efectos adversos en toda la economía mundial, y, por tanto, en la española, arreciando la situación a finales de la primavera de 2008, entrando el sistema financiero en la crisis más profunda de los ùltimos años. En los sectores de la construcción e inmobiliario que habían sido un motor importante de la economía española se comenzó a destruir empleo, provocando el efecto contagio a otros sectores lo que afectó de manera muy importante a la capacidad de pago de las familias españolas, entorno geográfico de actividad financiera de la Entidad que nos cedió los datos, y, como consecuencia, al grado de cumplimiento de las obligaciones contractuales sobre préstamos a particulares.

Pero independientemente de las causas que lo provocaron, es evidente que los test de calibración indican que los modelos deben de ser reconstruidos, prácticamente en todas sus fases, por cuanto el comportamiento de los clientes frente al default ha cambiado, al igual que con toda seguridad las variables seleccionadas y, como consecuencia, la relación de dependencia entre el estado de default y las variables explicativas del riesgo de crédito. Precisamente el objetivo de la calibración es fundamentalmente la validación de los modelos para modificarlos si es necesario.

CONCLUSIONES.

La revisión de la bibliografía, sobre modelos de calificación del riesgo de crédito, pone de manifiesto que:

1.- La estructura funcional más utilizada, en la construcción de modelos de credit scoring, es la LOGIT, dada su facilidad de interpretación, requisito exigido en Basilea II. Sin embargo estos modelos no capturan de forma óptima la relación de dependencia entre las variables explicativas y el estado de default, cuando ésta relación es no lineal.

2.- Desde principios de los años 80 se han propuesto varias alternativas para intentar caracterizar la no linealidad de las variables explicativas, en los modelos estadísticos de probabilidad de default, pero ninguna de ellas ha dado una respuesta totalmente satisfactoria.

3.- En este trabajo proponemos una alternativa, basada en las ideas de HASTIE y TIBSHIRANI (1996), a la que llamamos Modelos Logísticos Lineales Híbridos de Expansiones Lineales por funciones de base, modelos HLLM, los cuales se obtienen al expandir la componente no lineal de Modelos Logísticos Parcialmente Lineales, LPLM, a través de expansiones lineales de funciones de base especificas para cada variable.

4.- Planteamos el Problema General de Estimación de los Modelos Logísticos por expansiones de funciones de base, cuya función objetivo se configura con la expansión lineal de las funciones de base, la función de pérdida logística como término de ajuste y la funcional de penalización como término de regularización. Obtenemos la solución resolviendo el sistema de ecuaciones normales a través del algoritmo de Newton_Raphson, que requiere el gradiente y la matriz de segundas derivadas o matriz Hessiana.

5.- Presentamos una novedosa visión general unificada de la gran mayoría de las técnicas más actuales de credit scoring, formalizando sus estructuras funcionales como expansión de funciones de base de las variables explicativas del riesgo de crédito, a partir de tres hipótesis generales. La diferencia entre las distintas técnicas viene dada por la función de enlace entre la probabilidad de default y la expansión lineal de funciones de base, identidad, logit, probit y vector soporte. De este modo clasificamos los modelos en cuatro grandes grupos: modelos de probabilidad, logísticos, probit y vector soporte por expansión lineal de funciones de base.

Hemos desarrollado una metodología para el proceso completo de construcción de un modelo proactivo de credit scoring HLLM, sobre los datos reales de una entidad financiera, desde la óptica IRB de Basilea II. Con respecto a esta parte del trabajo concluimos:

6.- Los estadísticos para detección de multicolinealidad y la cuantificación de la Inflación de varianza, son muy sensibles a la influencia muestral, lo cual hace imprescindible la utilización de técnicas bootstrap para estimar su significación y estabilidad, dado el volumen de datos con el que se trabaja. Se han formulado como estadísticos de conjunto, Bagging.

7.- Para evitar los problemas de inestabilidad y sesgo, que a menudo presentan las estimaciones de los coeficientes de regresión basadas en métodos Paso a Paso, ha sido necesario el uso de la Regresión Logística Lineal Backward con remuestreo Bootstrap, BLLR_Bag.

8.- Ha sido posible proponer un método constructivo supervisado para seleccionar las funciones de base que, con criterios de bondad de ajuste, poder explicativo y discriminante y eficacia clasificadora, han resultado esenciales para configurar la estructura del modelo HLLM.

9.- El modelo HLLM obtenido es parsimonioso y tanto en la fase de entrenamiento, como en la de validación y test, presenta características adecuadas no solo desde el punto de vista del riesgo de crédito, sino también desde el punto de vista estadístico: alta bondad de ajuste, un alto poder discriminante, alta eficacia como clasificador y un bajo error test esperado.

10.- Además de las ventajas teóricas y las relacionadas con los requerimientos de Basilea II, el modelo HLLM propuesto presenta mejor rendimiento discriminante, que las técnicas utilizadas usualmente en los sistemas de calificación del riesgo de crédito, TREE, SLPM, k-NN y SVM. *Los modelos HLLM y HLPM presentan las menores tasas de clasificación incorrecta.*

11.- Con respecto a la calibración, ninguno de los modelos está calibrado adecuadamente, si bien se tiene para todos ellos que la categoría de default y la de menor puntuación entre las de no default no están infra estimadas. Creemos que esto no puede ser imputado a la construcción de los modelos sino al profundo cambio de las condiciones en torno al riesgo de crédito acaecidas entre el año 2007 y 2008.

BIBLIOGRAFIA.

Adam, D. (1958). Les Reactions du Consommateur Devant les Prix. In *Observation Economique, **15***. Sedes; Paris.

Agresti A. (2002). *Categorical Data Analysis*. John Wiley and Sons, New York, (2ª Edición).

Aitchison, J., Brown, J.A.C. (1957). *The lognormal distribution*. Cambridge University Press, Cambridge.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-722.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics,* **16**, 125-127.

Altman E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, **23 (4)**, 589-609.

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* Academic Press.

Artís, M., Guillén, M., Martínez, J. M. (1994). A model for credit scoring: an application of discriminant anlysis. *QÜESTIIÓ ,* **18 (3)**, 385-395.

Ashford, J. R., Sowden, R. R. (1970). Multi-Variate Probit Analysis . *Biometrics*, **26**, 535-546.

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford University Press, Oxford.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical inference under order restrictions: the theory and application of isotonic regression.* Wiley, New York.

Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields,* **113**, 301-413.

Bartholomew, D.J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, **48**, 211-220.

Bartlett, P., Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research,* **3**, 463-482.

BCBS (2003), Basel Committee on Banking Supervision. Consultative Document: *The New Basel Capital Accord*. (Emitido en abril de 2003). http://www.bis.org/bcbs/bcbscp3.htm

BCBS (2004), Basel Committee on Banking Supervision. *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework* http://www.bis.org/publ/bcbs107.htm.

BCBS (2005a), Basel Committee on Banking Supervision. *Studies on the Validation of Internal Rating Systems (revised),* Technical Report Working Paper No. 14. http://www.bis.org/publ/bcbs_wp14.htm

BCBS (2005b), Basel Committee on Banking Supervision. *Validation of low-default portfolios in the Basel II Framework*. Newsletter nº 6. http://www.bis.org/publ/bcbs-n16.pdf.

BCBS (2005c), Basel Committee on Banking Supervision. *An explanatory note on the Basel II IRB Risk Weight Functions*. Technical report. http://www.bis.org/bcbs/irbriskweight.pdf.

BCBS (2006), Basel Committee on Banking Supervision. *Convergencia Internacional de Medidas y Normas de Capital. Marco Revisado*. Versión Integral. Junio 2006. http://www.bis.org/publ/bcbs128_es.pdf

Bellman, R. (1961). *Adaptative Control Process: A Guided Tour*. Princeton University Press.

Belsley, D. A., Kuh, E., Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*. John Wiley, New York.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, **39 (227)**, 357-65.

Berkson J (1950). Are there two regressions?. *Journal of the American Statistical Association,* **45 (250)**, 164-180.

Berry, W. D. y Feldman, S. (1985). *Multiple regression in practice*. Sage Publications, London.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Bliss, C. I. (1934a). The method of probits. *Science* **79**, 38-39.

Bliss, C. I. (1934b). The method of probits -- a correction. *Science,* **79**, 409-410.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology,* **22**, 134-167. *(Con un apéndice por R.A. Fisher).*

Blake, C. L., Merz, C. J. (1998). 'UCI Repository of Machine Learning Databases'. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Blöchlinger, A., Leippold, M. (2006). Economic Benefit of Powerful Credit Scoring. *Journal of Banking and Finance*, 30, 851–873.

Blochwitz, S. Hohl, S. Tasche, D. D., When, C.S. (2004). *Validating default probabilities on short time series*. Working paper, Deutsche Bundesbank. http://www-m4.ma.tum.de/pers/tasche/validating.pdf

Blochwitz, S., Martin, M. R. W., When, C. S. (2006). XIII. Statistical Approaches to PD Validation. En *The Basel II Risk Parameters*, 289-305. Springer, Berlin.

Bock, R. D., Gibbons, R. D. (1996). High-Dimensional Multivariate Probit Analysis. *Biometrics*, **52**, 1183–1194.

Boj, E., Claramunt, M. M., Esteve, A., Fortiana, J. (2009a). Credit Scoring basado en distancias: coeficientes de influencia de los predictores. En: Heras, A. y otros (2009). *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009*, 15-22. Cuadernos de la Fundación MAPFRE, **136**. Fundación MAPFRE Estudios, Madrid.

Boj, E., Claramunt, M. M., Esteve, A., Fortiana, J. (2009b). Criterios de selección de modelo en el credit scoring: aplicación del análisis discriminante basado en distancias. *Anales del Instituto de Actuarios Españoles*. **15**, 209-230.

Bonilla, M., Olmeda, I., Puertas, R. (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring. *Revista Española de Financiación y Contabilidad* **32, (118)**, 833-869.

Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

Box, G.E.P., Tidwell, P.W. (1962). Transformation of the Independent Variables. *Technometrics*, **4 (4)**, 531-550.

Box, G.E.P., Huntern, W.G. y Hunter, J.S. (1988). *Estadística para investigadores. Introducción al diseño de experimentos, análisis de datos y construcción de modelos.* Ed. Reverté, Barcelona.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall. Belmont, Wadsworth.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* **26**, 123-140.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science* **16**, 199-231**.** (Con comentarios y respuesta por el autor).

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78 (1)**, 1-3.

Castermans G, Martens D, Van Gestel T, Hamers B, Baesens B. (2010). An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society,* **61 (3**), 359-373.

Clavero Rasero, B. (2008). Credit Rating. *Statistical Aspects at Devolopment*. VDM Verlag. Saarbrücken.

Cleveland, W.S. (1979). Robust Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association,* **83**, 596-610.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* **20**, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* **70**, 213-220.

Collet, D. (1991). *Modelling Binary Data*. Chapman & Hall, London.

Cook, R. D., Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. *Journal of the National Cancer Institute,* **11**, 1269-1275.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. En J. Neyman (ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, Calif., 135-148. University of California Press.

Cox, D.R. (1970). *The Analysis of Binary Data*. Methuen, London.

Cox, D.R., Snell, E.J. (1989). *Analysis of Binary data*. Chapman & Hall. London.

Cox, D.R., O' Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Stat.*, 18, 1676-1695.

Cragg, J.G., Uhler, R. (1970). The demand for automobiles. *Can. J. Econ.,* **3**, 386-406.

Cramer, J.S. (2003). *Logit Models From Economics and Other Fields*. Cambridge University Press, Cambridge.

Cramer, J.S., (2004). Scoring bank loans that may go wrong: A case study. *Statistica Neerlandica,* **58**, 365-38.

Craven, p., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik,* **31**, 377-403.

Cuadras, C.M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. En Y. Dodge (ed.), *Statistical Data Analysis and Inference*, 459-473. Elsevier Science Publishers B. V. (North.Holland), Amsterdam.

Cuadras, C.M. (1991). *Métodos de Análisis Multivariante*. EUNIBAR, 2ª edición, PPU, Barcelona.

Cuadras, C.M. (1992). Some examples of distance based discrimination. *Biometrical Letters,* **29**, 3-20.

Cuadras, C. M. and Fortiana, J. (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis*, **52**, 1-14.

Cuadras, C. M., Fortiana, J. and F. Oliva (1997). The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification,* **14**, 117-136.

Culp, S., Nadal, J. (2010). Basilea III y el sistema bancario mundial. *Expansión.com*. Sección de Opinión, 12 de noviembre de 2010. http://www.expansion.com/2010/11/11/opinion/tribunas/1289512718.html?a=b8ba0230becdb5224c289ea609fc4bf4&t=1299440590

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian designs applied to logistic regression experiments. *J. Statist. Plann. Inference*, **21**, 191-208.

Chen, M. H. (1995). Asymptotically efficient estimation in Semiparametric Generalized Linear Models. *The Annals of Statistics*, **23 (4)**, 1102-1129.

Chen, M. H., Spiegelhalter, D., Thomas, A. y Best, N. (1996-2000): "The BUGS Project". http://www.mrc-bsu.cam.ac.uk/bug. (MRC Biostatistcs Unit, Cambrigde, UK).

Chen, S., Härdle, W., Moro, R. (2006). Estimation of Default Probabilities with Support Vector Machines, *SFB 649 Discussion Papers*. Economic Risk, Berlin. http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2006-077.pdf

D'Agostino, R.B., Stephens, M.A., eds. (1986). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Davis, L. (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York.

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002). *Bayesian methods for nonlinear classification and regression*. John Wiley, Chichester, West Sussex,

Derksen, S., Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noisy variables. *British Journal of Mathematical and Statistical Psychology,* **45**, 265-282.

Devroye, L, Wagner, T.J. (1977). The Strong Uniform Consistency of Nearest Neighbor Density Stimate. *The Annals of Statistics,* **5 (3)**, 536-540.

Dewald, W.G., Thursby, J.G., Anderson, R.G. (1986). Replication in empirical economics: The journal of money credit and banking project. *The American Economic Review,* **76**, 587-603.

Diaconis, P., Shahshahani, M. (1984). On nonlinear functions of linear combinations. *SIAM J. Sci. Stat Comput.* **5 (1)**, 175-191.

Dixon, P.M. (2001). The bootstrap and the jackknife: describing the precision of cological studies, in *Design and Analysis of Ecological Experiments*, 2nd Edition, S. Scheiner & J. Gurevitch, eds, Oxford University Press, Oxford.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, **24 (2)**, 508-539.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine Learning: Proceedings of the Twelfth International Conference.* 194-202.

Durrleman, S., Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine* **8 (5)**, 551–561.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics,* **7 (1)**, 1-26.

Efron B. (2004) "Large-Scale simultaneous hypothesis testing: the choice of a null hypothesis". *Journal of the American Statistical Association,* **99**, 96-104.

Efron, B., Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, **1**, 54-77.

Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing with *B*-splines and penalized likelihood. *Statist. Sci.*, **11 (2)**, 89-121.

Engelmann, B. (2006). Measures of a Rating's Discriminative Power - Applications and Limitations. En Engelmann, B., Rauhmeier, R. (eds.): *The Basel II Risk Parameters - Estimation, Validation and Stress Testing*. Springer, Berlin.

Engelmann, B., Hayden, E., Tasche, D. (2003a). Measuring the Discriminative Power of Rating Systems. En Discussion paper Nº 01/2003, *Banking and Financial Supervision - Deutsche Bundesbank* .
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.2643[1].pdf

Engelmann, B., Hayden, E., Tasche, D. (2003b). Testing rating accuracy. *Risk*, **1**, 82-86.

Engelmann, B., Rauhmeier, R. (2006). *The Basel II Risk Parameters Estimation Validation and Stress Testing*. Springer, Heidelberg.

Esposito, F., Malerba, D., Semeraro, G., Kay, J. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19 (5)**, 476-491.

Esteve, A. (2003). *Distancias estadísticas y relaciones de dependencia entre conjuntos de variables*. Tesis Doctoral. Universidad de Barcelona.

Eubank, R.L. (1988). *Spline smoothing and non parametric regression*. Marcel Dekker, New York.

Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York.

Everitt, B. S. (1992). *The Analysis of Contingency Tables*. Chapman Hall, London.

Everitt, B. S., Der, G. (1996). *A handbook of Statistical Analyses Using SAS*. Chapman & Hall, New York.

Evgeniou, T., Pontil, M., Poggio, T. (1999). Regularization networks and support vector machines. En A.J. Smola, P. Bartlett, B. Schölkopf & Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press.

Evgeniou, T., Pontil, M., Poggio, T. (2000). Regularization networks and support vector machines*, Advances in Computational Mathematics* 13(1):1-50.

Fahrmeir, L., Hamerle, A., Tutz, G. (1996). *Multivariate Statistiche Verfahren*. Walter de Gruyster, Berlin.

Fan, J. , Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. *Monographs on Statistics and Applied Probability*, **66**. Chapman and Hall, New York.

Farrell, M. J. (1954). The demand for motorcars in the United States. *Journal of the Royal Statistical Society, series A* **117**, 171-200.

Fayyad, U., Irani, K. (1993). Multi-interval discretization of continuous-value attributes for classification learning. En: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, **7**, 179-188.

Fix, E., Hodges., J.L. (1951). Discriminatory analisys, nonparametric discrimination, consistency properties. *Project 21-49-004, Report 4*, School Aviation Medicine, Randolph Field, Texas.

Fogel, D. B. (2006). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence,* IEEE Press, Piscataway, New York.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19 (1)**, 1-67.

Friedman, J. and Tukey, J. (1974), A projection pursuit algorithm for exploratory data analysis. *IEEE trans. on computers, Ser.,* **23**, 881-889.

Friedman, J. and Stuetzle, W. (1981), Projection pursuit regression, *Journal of the American Statistical Association,* **76**, 817-823.

Gaddum, J. H. (1933). *Reports on Biological Standard III. Methods of Biological Assay Depending on a Quantal Response.* Special Report Series of the Medical Research Council, **183**, Medical Research Council, London.

Girosi, F., Jones, M., Poggio, T. (1995). Regularization Theory and Neural Network Architectures. *Neural Computation,* **7**, 219-269.

Goldberg, G. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley.

Gower, J.C. (1982). Euclidean Distance Geometry. *Math. Scientist,* **7**, 475-493.

Gray, R. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Associations,* **87**, 942-951.

Green, P.J., Yandell, B. (1985). Semi-parametric generalized linear models. *Proceedings 2nd International GLIM Conference, Lancaster, Lecture Notes in Statistics,* **32**, 44-55. Springer-Verlag, New York.

Green, P.J., Silverman, B.W. (1994). Non Parametric Regression and Generalized Linear Models. *Vol. **58** of monographs on Statistics and Applied Probability*. Chapman and Hall. London.

Greene, W. H. (2003). *Econometric Analysis*. (5ª edición). Prenctice Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

Hall (1989). Polynomial Projection Pursuit. *Annals of Statistics,* **17**, 589-605.

Hand, D., Henley, W. (1997). Statistical classification in consumer credit scoring: a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* **160 (3)**, 523-541.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004a). *Nomparametric and Semiparametric Models*. Springer-Verlag, Berlin.

Härdle, W., Huet, S., Mammen, E., Sperlich, S. (2004b). Bootstrap Inference in *Semiparametric Models. Econometric Theory,* **20**, 265-300.

Härdle, W., Moro R. A., Schäfer, D. (2005). Predicting Bankruptcy with Support Vector Machines. *Statistical Tools for Finance and Insurance*. Springer Verlag, Berlin.

Härdle, W., Moro, R.A., Schäfer, D. (2007). Estimating probabilities of default with support vector machines. *Discussion Paper Series 2: Banking and Financial Studies 2007, 18, Deutsche Bundesbank*, Research Centre.

Härdle, W., Hoffmann, L., Moro, R. (2011). Learning Machines Supporting bankruptcy prediction. En *Statistical Tools in Finance and Insurance*. Springer Verlag, Berlin.

Harrell, J.F.E. (2001). *Regression Modeling Strategies*. Springer, New York.

Hastie, T. (2006). Gam: generalized additive models. R package version 0.98.

[CRAN]- http://cran.r-project.org/web/packages/gam/index.html

Hastie, T., Tibshirani, R. (1986). Generalized Additive Models. *Stat. Science,* **1**, 297-318.

Hastie, T., Tibshirani, R. (1990). Generalized Additive Models. *Monographs on Statistics and Applied Probability*, **43**. Chapmann and Hall, London.

Hastie, T. and Tibshirani, R. (1996). Nonparametric regression and classification. Part II- nonparametric classification. En V. Cherkassky, J. H. Friedmann y H. Wechsler (eds), *From Statistics to Neural Networks: Theory and Pattern Recognition Applications, Computer and System Sciences*, **136**, 70-82.

Hastie, T. and Tibshirani, R. (1999). *Generalized Additive Models*. Chapmann and Hall, London.

Hastie, T., Tibshirani, R., Friedmann, J. (2009). *Elements of Statistical Learning. Data Nining, Inference, and Prediction*. 2ª ed., Springer, New York. (1ª ed. 2001).

Hayden,E., Stomper, A., Westerkamp, A. (2009). *Selection vs. Averaging of Logistic Credit Risk Models*, Working Paper, Oesterreichische Nationalbank, MIT Sloan and University of Vienna.

http://homepage.univie.ac.at/evelyn.hayden/bma.pdf

Haykin, S. (1998). *Neural Networks: A comprehensive Foundation*. Prentice Hall, 2ª ed., Upper Saddle River, New York.

Hervás-Martínez, C., Martínez-Estudillo, F.J., Martínez-Estudillo, A. C., Gutiérrez, P.A., Fernández, J.C .(2007). Aprendizaje mediante la hibridación de técnicas heurísticas

y estadísticas de optimización en regresión logística binaria, V Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB07), Tenerife (España, 2007), 61-68.

Holland, J. H. (1987). Genetic algorithms and classifier systems: foundations and future directions. En *Proceedings 2nd International Conference on Genetic Algorithms*, Lawrence Erlbaum Associates, Inc., 82-89.

Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistig Regression*. 2ª ed., John Wiley, New York. (1ª ed. 1989).

Huber, P. (1985). Projection pursuit. *Annals of Statistics* , **13**, 435-525.

Jaakkola, T. S., Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing,* **10**, 25-37.

Jones, M.C., Sibson, R. (1987). What is Projection Pursuit?. *Journal of the Royal Statistical Siciety. Serie A.* **150**, 1-36.

Jones, M.C., Marron, J.S., Sheater, S.J. (1996). A Brief Survey of bandwidth Selection for Density Estimation. *Journal of the American Statistical Association*, **91**, 401-407.

Krzonowski, W.J. (1987). A Comparison Between Two Distance-Based Discriminant Principles. Journal of Classification, **4**, 73-84.

Krzonowski, W.J. (1988). *Principles of Multivariate Analysis: a user´s perspectiv*e. Clarendon Press, Oxford.

Keenan, S. y Sobehart, J. (1999). Performance measures for credit risk models. *Technical report, Moody's Investor service, Global Credit Research.*

Kimeldorf, G., Wahba, G. (1971). Some results on Tchebycheffian spline function. *J. Math. Anal. Applic.*, **33**, 82-95.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E. (1988). *Applied regression analysis and other multivariate methods*. PWS-Kent, Boston.

Kleimbaum, D.G. (1994). *Logistic regression: A Self-Learning Text*. Springer-Verlag, New York.

Kruskal, J.B. (1969). Toward a practical method wich helps uncover the structure of a set of observations by finding the line transformation which

optimizes a new "index of condensation". In Milton, R.C. and Nelder, J.A. Editors. *Statistical Computation*. 427-440. Academic press, New York.

Kullback, S. (1959). *Information Theory and statistics*. Wiley, New York.

Kumar, A.; Olmeda, I. (1999). A Study of Composite or Hybrid Classifiers for Knowledge Discovery, *INFORMS Journal of Computing,* **11**, 267-277.

Ladd, G.W. (1966). Linear probability functions and discriminant function. *Econometrica,* **34**, 873-885.

Lachenbruch, P.A. (1975). *Discriminant Analysis*. Hafner press, New York.

Larsen, K. (2005). ACM SIGKDD. *Explorations Newsletter,* **7 (1)**, 76-86.

Leahy, K. (2001). Multicollinearity: When the solution is the problem. En Olivia Parr Rud (ed.) *Data Mining Cookbook*, 106-108. John Wiley, New York.

Lee, K., Koval, J. J. (1997). Determination of the best significance level in forward logistic regression. *Communications in Statistics - Simulations*, **26**, 559-575.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. Holden-Day, Inc., San Francisco, y McGraw-Hill, New York.

Leontief, W. (1947). Introduction to a Theory of the Internal Structure of Functional Relationships. *Econometrica,* **15**, 361-73.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., Sturtz, S. (2007). BRugs 0.4. R package. R Foundation for Statistical Computing, Vienna, Austria. http://www.cran.r-project.org.

Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika,* **80,** 557–572.

Lin, D. Y., Wei, L. J., Ying, Z. (2002). Model-Checking Techniques Based on Cumulative Residuals. *Biometrics*, **58**, 1-12.

Lindsey, J. K. (1995). The uses and limits of linear models. *Statistics and Computig*, **5**, 87-89.

Little R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.(2ª Edición).

Liu, W., Cela, J. (2007). Improving Credit Scoring by Generalizad Additive Model. *Data Mining and Predictive Modeling. SAS Global Forum 2007*, *Paper 078-0076.* Cary, NC: SAS Institute Inc.

Liu, W., Cela, J. (2009). Generalizations of Generalized Additive Model (GAM): A Case of Credit Risk Modeling. Statistics and data Analysis. *SAS Global Forum 2009, Paper 113-2009.* SAS Institute Inc, Cary, North Carolina.

Loftsgaarden, D.O., Quesenberry, C.P. (1965). A Nonparametric Estimate of a multivariate Density Function. *Ann. Math. Statist.*, **36 (3)**, 1049-1051.

Lunneborg, C. E. (1994). *Modeling Experimental And Observational Data.* Duxbury Press, Belmont.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge.

Mallo, F. (1985). *Análisis de Componentes Principales y Técnicas Factoriales Relacionadas. Teoría, Computación y Aplicaciones*. Servicio de Publicaciones de la Universidad de León.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, **15 (4)**, 661-675.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* 2ª ed., Chapman and Hall, London.

Mardia, K., Kent, J., Bibby, J. (1979). *Multivariate Analysis.* Academic Press, London.

Matusita, K. (1956). Decisión rule, based on the distance, for the classification problem. *Annals of the Institute of Statistical Mathematics,* **8**, 67-77.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models*, 2ª Ed. Chapman and Hall, New York. (1ª ed. 1983).

McCulloch, W.S., Pitts, W.H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical Biophisics,* **5**, 115-133.

McCullough, B.D., McGeary, K.A., Harrison, T.D., (2006). Lessons from the JMCB archive. *Journal of Money, Credit, and Banking,* **38**, 1093–1107.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. En *Frontiers in Econometrics*, 105-142. P. Zarembka (ed.), Academic Press, New York.

McFadden, D. (1978). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. En D. Hensher and P. Stopher (eds.), B*ehavioural Travel Modelling*, 279-318. Croom Helm London, London.

Medema, L. Koning, R.H., Lensink, R. (2009). A practical approach to validating a PD model. *Journal of banking and finance*, **33(4)**, 701-708.

Menard, S. (2002). *Applied logistic regression analysis, 2ª ed.* Thousand Oaks, Sage Publications. Series: *Quantitative Applications in the Social Sciences*, 106.

Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance,* **29**, 449-470.

Michie, D., Spiegelhalter, D. J., Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, New York.

Miller, L.H. (1956). Table of percentage points of Kolmogorov Statistics. *Journal of the American Statistical Association, **51**, 111-121.*

Mingers, J. (1989a). An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, **4(2**), 227-243.

Mingers, J. (1989b). An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, **3(4)**, 319-342.

Mittlbock, M., Schemper, M. (1996). Explained variation for logistic regression. S*tatistics in Medicine,* **15**, 1987-1997.

Mooney, C. Z., Duval, R. D. (1993). *Bootstrapping. A nonparametric approach to statistical inference.* Newbury Park, Sage Plubications.

Morgan, J. N., Messenger, R. C. (1973). THAID: a sequential search program for the analysis of nominal scale dependent variables. *Ann Arbor*. Technical report, Institute for Social Research, Univ. of Michigan.

Moore, D. S., Henrichon, E. G. (1969). Uniform Consistency of Some Estimates of a Density Function. *Ann. Math. Statist.,* **40(4)**, 1499-1502.

Morozov, V.A. (1984). *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, New York.

Morrison , D.F. (1976). *Multivariate Statistical Meth*ods. 2ª ed. McGraw-Hill, New York.

Müller, M. (2000). *Semi-parametrics Extensions to Generalization Linear Models*. Habilitationschrift. Bajado el 2 de mayo de 2008 del sitio web http://www.marlenemueller.de/publications.html.

Müller, M. (2001). Estimation and testing in Generalized Partial Models – A comparative study. *Statistics and Computing*, **11**, 299-309.

Müller, M.; Härdle, W. (2003): Exploring Credit Data**.** En Bol, G.; Nakhaeizadeh, G.; Rachev, S.T.; Ridder, T.; Vollmer, K.-H. (eds.): *Credit Risk - Measurement, Evaluation and Management*, Physica-Verlag. (Proceedings Ökonometrie-Workshop *Kreditrisiko - Messung, Bewertung und Management*, Universität Karlsruhe)

Nadaraya, E.A. (1964). *On stimating regression . Theory of probability and its Applications,* vol. **10**. 186-190.

Nagelkerke, N.J.D. (1991). A Note on a General Definition of the Coeficient of Determination. *Biometrika,* **78**, 691-692.

Nelder, J.A., Wedderburn, R.W.N. (1972). Generalized Linear Models. *Journals of the Royall Statistical Society*, Series A, **135**, 370-384.

Noreen, E.W. (1989). *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley, New York.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research,* **18(1)**, 109-131.

Olmeda, I., Fernández, E. (1997). Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. *Computational Economics*, **10**, 317-335.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science,* **1**, 505-27.

Parker, D. (1985). *Learning Logic, Technical Report TR-87*, Cambridge MA: MIT Center for research in computacional Economics and Management Science.

Parzen, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statistics,* **33**, 1065-1076.

Peng, H. (2004). Efficient Inference in Semiparametric Generalized Linear Models, http://WWW.home.olemiss.edu/~mmpeng/preprints.

Peng, H., Wang, X. (2004). Moment estimation in semiparametric Generalized Linear Models. *Nonparametric Statistics*, 1-22.

Perlich, C., Provost, F., Simonoff, J.S. (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research,* **4**, 211-255.

Porath, D. (2006). Scoring Models for retail Exposures. En Engelmann, B. and Rauhmeier, R., eds. *The basel II Risk Parameters. Estimation, Validation and Stress testing*. Springer-Verlag, Heidelberg, Berlin.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, **1**, 81-106.

Quinlan, J. R. (1987). Simplifying decision trees. *Int. J. of Man-Machine Studies*, **27**, 221-234.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California.

Rao, C.R. (1982a). "Diversity and dissimilarity coefficient: a unified approach". *Theoretical Population Biology*, **21**, 24-43.

Rao, C.R. (1982b). "Diversity: its measurement, decomposition, apportionment and analysis". *Sankhya. The Indian Journal of Statistics*. Series A. **44**, 1-22.

Rauhmeier, R. , Scheule, H. (2005), Rating properties and their implications for Basel II capital. *Risk*, **18(3)**, 78-81.

Rice, J., Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics, **57**, 253-259.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Roosen, C., Hastie, T. (1994). Automatic smoothing spline projection pursuit. *Journal of Computational and Graphical Statistics,* **3**, 235-248.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics,* **27,** 832-837.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain, Cornell Aeronautical Laboratory. *Psychological Review* **65(6),** 386-408.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptron and Theory of Brain Mechanisms*. Spartan-Book, Washington.

Rosset, S., Zhu, J. Hastie, T. (2004). Margin maximizing loss function. Neural Information Processing Systems, 16.

Rumelhart, D., Hinton, G., Williams, R. (1986). Learning internal representations by error propagation. En D. Rumedlhart and J. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, The MIT Press, Cambridge, 318-362.

Saharon, R. (2003). "Topics in Regularization and Boosting". Ph. D. Dissertation, Standorf University.
http://www.stanford.edu/~hastie/THESES/saharon_rosset.pdf

SAS Institute Inc. (2007). *SAS/STAT ® R. Development Core Team*. 2007. SAS Institute Inc, Cary, North Carolina.

SAS Institute Inc. (2009). *SAS/STAT ® 9.2 .User's Guide, Second Edition*. SAS Institute Inc, Cary, North Carolina.

Schoenberg, I.J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions, *Quart. Appl. Math.*, **4**, 45-99 y 112-141.

Schölkopf, B., Herbrich, R., Smola, A. J. and Williamson, R. C. (2001). A generalized representer theorem. *Lecture Notes in Artificial Intelligence* , **2111,** 416-426. Springer, Berlin.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Scott, D.W. (1985). Average shifted histograms: effective nonparametric density stimators in several dimensions, *Ann. Statist.*, **13** , 1024-1040.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley, New York.

Seber, G.A.F. (1984). *Multivariate Observations*. John Wiley, New York.

Seidenfeld, T. (1985). Calibration, Coherence, and Scoring Rules. *Philosophy of Science*, **52(2)**, 274-294.

Severini, T.A., Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89,** 501-511.

Severini, T.A., Wong, W.H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics*, **20(4)**, 1768-1802.

Shtatland, E. S., Barton, M. B. (1998). An information-gain measure of fit in PROC LOGISTIC. *SUGI*98 *Proceedings*, SAS Institute Inc., Cary, North Carolina, 1194-1199.

Shtatland, E. S., Cain E., and Barton, M. B. (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System. *SUGI'26 Proceedings, Paper 222-26*. SAS Institute Inc., Cary, North Carolina.

Shtatland, E. S., Kleinman K., and Cain E. M. (2003). Stepwise methods in using SAS PROC LOGISTIC and SAS ENTERPRISE MINER for prediction. *SUGI '28 Proceeding, Paper 258-28*. SAS Institute Inc., Cary, North Carolina.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules. *Philosophy of Science* **52**, 274-294.

Siddiqi, N. (2006). *Credit Risk Scorecards. Developing and implementing Intelligent Credit Scoring*. Jhon Wiley, Hoboken, New Jersey.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Smirnov, N.V. (1948). *Tables for estimating the goodness of fit of empirical distributions*, *Annals of Mathematical Statistic*, 19, 279-281.

Sonquist, J.A. y Morgan, J.N. (1964): The Detection of Interaction Effects. *Monografía núm. 35. Survey Research Centre*, University of Michigan.

Speckman, P.E., (1988). Regression analysis for partially linear models. *Journal of the Royal of Statistical Society*, B **50,** 413-436.

Spiegelhalter, D. (1986). Probabilistic prediction in patient management and clinical trails. *Statistics in Medicine*, **5**, 421-433.

Schroeder, L. D., Sjoquist, D. L., Stephan, P. E. (1986). *Understanding regression analysis*. Sage Publications, Beverly Hills, California.

Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking and Finance*, **29**, 1213-1236.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell Jr, F. E., Habbema, J. D. F. (2000). Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, **19**, 1059-1079.

Stone, M . (1974). Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Roya1 Statistical Society,* B **36**, 111-147.

Stone, C.J. (1985). Additive Regression and Other Nonparametric Models. *Annals of statistics*, **13**, 689-705.

Stone, C,J. (1986a). Comment, 312-314, to paper by Hastie T. and Tibshirani R. (1986): Generalized additive models. *Statist Sciences*, **1**, 297-318.

Stone, C.J. (1986b). The dimensionality reduction principle for generalized additive models. *Annals of statistics*, **14(2)**, 590-606.

Stone, C.J., Koo, C.Y. (1985). Additive splines in statistics. *Proceedings of the Statistical Computing Section ASA, American Statistical Association*, 45-48. Washington .

Stone, C., Hansen, M., Kooperberg, C., Truong, Y. (1997). Polynomial Splines and their tensor products (con discusión). *Annals of Statistics,* **25(4)**, 1371-1470.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal* of *the American Statistical Association,* **86**, 420-426**.**

Tasche, D. (2003). A traffic lights approach to PD validation, Working Paper. Bajado el 18 de marzo de 2008 del sitio web: http://arxiv.org/PS_cache/cond-mat/pdf/0305/0305038v1.pdf

Theil, H. (1979). *Principles of Econometrics*. Wiley, New York.

Thomas, L. C., Edelman, D. B., Crook, J. N. (2002). *Credit Scoring and Its Applications,* (2ª edición), Elsevier Science Publishers, North-Holland, Amsterdam.

Tikhonov, A.N., Arsenin, V.Y. (1977). *Solutions of III-posed Problems*. W.H. Winston, Washington.

Tornabell, R. (2010). Basilea III: nuevos requisitos de capital y liquidez. *Expansión.com*. *Sección de Opinión*, 13 de setiembre de 2010. http://www.expansion.com/2010/09/13/opinion/tribunas/1284405984.html

Trias, R., Carrascosa, F., Fernández, D., Parés, Ll., G. Nebot (2005). *Riesgo de Créditos: Conceptos para su medición, Basilea II, Herramientas de Apoyo a la Gestión*. AIS Group - Financial Decisions. www.ais-int.com.

Trias, R., Carrascosa, F., Fernández, D., Parés, Ll., G. Nebot (2008). *El método RDF (Risk Dynamics into the Future). El nuevo estándar de stress testing de riesgo de crédito*. AIS Group - Financial Decisions. www.aisint. com.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Vaughn, T.S., Berry, K.E. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education,* [online]**, 13**(1). Bajado el 22 de abril de 2008 del sitio web www.amstat.org/publications/jse/v13n1/vaughan.html.

Wagner, T.J. (1973). Strong consistency of a nonparametric estimate of a density function. *IEEE trans. Systems, Man, and Cybernet,* **3**, 289-290.

Wahba, G. (1980). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. En *Approximation Theory III* (ed. W. Cheney), 905-912. Academic Press, New York.

Wahba, G. (1990). Splines models for observacional data. *CBMS-NSF Reg. Conf. Ser. Appl. Math*, **59,** SIAM. Philadelphia, Pennsylvania.

Wahba, G., Gu, C., Wang, Y., Chappel, R. (1995). Soft classification, a.k.a. Risk estimation, via penalized log-likelihood and smoothing spline analysis of variance. In D.H. Wolpert, (ed.), *The Mathematics of Generalization*. Santa Fe Institute in the Science of Complexity. Addisin-Wesley Publisher.

Wahba, G., Lin, Y., Zhang, H. (2000). GACV for support vector machines, En A. Smola, P. Bartlett, B. Schölkopf & Schuurmans, (eds.), *Advances in Large Margin Classifiers*. MIT Press. Cambridge, 297-311.

Wand, M.P., Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, **50**, 179-198.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

Watkins, A., Boggess, L. (2002). A new classifier based on resource limited artificial immune systems, in *Proceedings 2002 IEEE Congress on Evolutionary Computation (CEC2002),* **2**, 1546-1551.

Watkins, A., Timmis, J., Boggess, L. (2004). Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm. *Genetic Programming and Evolvable Machines*, **5**, 291-317.

Watson, G.S. (1964). Smooth regression analysis. *Shankya Series A*, vol. **26**, 359-372.

Weiss, S.M., Kulikowski, C.A. (1991). *Computer Systems That Learn*. Morgan Kaufmann Pub. Inc., San Francisco, California.

Werbos, P. (1974). *Beyond Regression*, PHD thesis, Harvard University.

Weyl, H. (1928). *Gruppentheorie und Quantenmechanik*. Hirzel, Leipzig.

Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser.*, B **20**, 234-343.

Widrow, B., Hoff, M. (1960). Adaptive switching circuits, *IRE WESCON convention record*, **4**, 66-104.

Woodroofe, M. (1970). On choosing a delta-sequence. *Ann. Math, Statist.,* **41**, 1665-1671.

Zhu, J., Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5(2)**, 427-443.

Zhu, J., Hastie, T. (2005). Kernel Logistic Regression and the Import Vector Machine,*J. of Computational and Graphical Statistics*, **14(1)**, 185-205.