Recibido en 26.05.2011. Aprobado en 05.09.2012
Evaluado por el sistema *double blind review*
Editor Científico: Ricardo Ratner Rochman

ANALISIS DEL *CREDIT SCORING*

ANÁLISE DO CREDIT SCORING

CREDIT SCORING ANALYSIS

RESUMEN

El problema de la morosidad está cobrándose una gran importancia en los países desarrollados. En este trabajo realizamos un análisis de la capacidad predictiva de dos modelos paramétricos y uno no paramétrico abordando, en este último, el problema del sobreaprendizaje mediante la validación cruzada que, muy habitualmente, se obvia en este tipo de estudios. Además proponemos la distinción de tres tipos de solicitudes dependiendo de su probabilidad cumplimiento: conceder, no conceder (de forma automática), y dudoso y, por consiguiente, proceder a su estudio manual por parte del personal bancario.

PALABRAS CLAVE: Clasificación crediticia, logit, análisis discriminante, árboles de clasificación, validación cruzada.

Rosa Puertas Medina rpuertas@esp.upv.es
Profesora do Departamento de Economía y Ciencias Sociales, Grupo de Economía Internacional, Universidad Politécnica de Valencia – Valencia, España

Maria Luisa Martí Selva mlmarti@esp.upv.es
Profesora do Departamento de Economía y Ciencias Sociales, Grupo de Economía Internacional, Universidad Politécnica de Valencia – Valencia, España

Resumo O problema dos atrasos em pagamentos vem adquirindo grande importância nos países desenvolvidos. Neste trabalho, realizamos uma análise da capacidade preditiva de dois modelos paramétricos e de um não paramétrico, abordando, neste último, o problema da sobreaprendizagem mediante a validação cruzada, o qual é muito frequentemente negligenciado nesse tipo de estudo. Além disso, propomos a distinção de três tipos de pedido conforme sua probabilidade de cumprimento – conceder, não conceder (de forma automática) e duvidoso – e, por conseguinte, realizar seu estudo manual por parte do pessoal bancário.
Palavras-chave: Classificação de crédito, logit, análise discriminante, árvores de classificação, validação cruzada.

Abstract *The problem of unpaid bank debts is becoming increasingly important in developed countries. Many empirical works are being published in an attempt to find a model capable of determining as accurately as possible whether an individual requesting a loan will be able to pay it back. This paper analyses the predicting capability of one non-parametric and two parametric models. As regards the former, the often-overlooked problem of overlearning is also tackled using the cross-validation technique. Furthermore, a three-level grading of loan applications is proposed depending on their likely performance: grant, refuse, or doubtful hence subject to manual consideration by bank staff.*
Keywords: *Credit scoring, logit, discriminant analysis, classification trees, cross-validation.*

INTRODUCCIÓN

La concesión de créditos es uno de los principales negocios de las instituciones bancarias, que a su vez puede ocasionar la quiebra de las mismas. Tal es el caso de numerosos bancos europeos que, en la actualidad, está pasando por una delicada situación debido a la creciente tasa de morosidad, obligando así a dichas entidades a incrementar la provisión por insolvencia, y eliminado cualquier posibilidad de beneficio e incluso llegando a tener que soportar importantes pérdidas.

Cada vez más se reclaman sistemas automáticos de concesión de créditos que aseguren con alta probabilidad que el cliente será capaz de hacer frente a sus obligaciones crediticias. Las entidades precisan incorporar calidad a sus créditos, utilizando para ello distintos modelos que faciliten y mejoren el proceso de aprobación de los mismos.

Se denomina *credit scoring* a todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. Riesgo que estará en función de la solvencia del deudor, del tipo de crédito, de los plazos, y de otras características propias del cliente y de la operación, que van a definir cada observación, es decir, cada solicitud de crédito. Únicamente no existirá riesgo en una operación de crédito cuando la entidad que los instrumenta actúe como mediadora o intermediaria, o bien cuando el crédito se conceda con la garantía del Estado.

Los créditos de clientes que no se pagan a su vencimiento no sólo generan costes financieros, sino que además producen costes administrativos de gestión para su recuperación. Por este motivo, las entidades financieras están prestando, cada vez más, especial atención a estas partidas que deterioran considerablemente su cuenta de resultados. Los modelos automáticos de clasificación crediticia pretenden evitar, en la medida de lo posible, la concesión de créditos a clientes que posteriormente puedan resultar fallidos, ocasionando un cuantioso quebranto a la entidad emisora del mismo. Las principales ventajas que incorpora el *credit scoring* en los procesos de concesión son: el proceso de aprobación o no del crédito no va a depender de la discrecionalidad del personal, al tratase de un sistema objetivo y, además, al ser un sistema automático, no precisa de mucha dedicación de tiempo y personal, permitiendo reducir costes y tiempo de tramitación.

El *credit scoring* constituye, por lo tanto, un problema de clasificación propiamente dicho, ya que dado un conjunto de observaciones cuya pertenencia a una determinada clase es conocida a priori, se busca una regla que permita clasificar nuevas observaciones en dos grupos: los que con alta probabilidad podrán hacer frente a sus obligaciones crediticias, y aquellos que, por el contrario, resultarán fallidos. Para ello se tendrá que realizar un análisis de las características personales del solicitante (profesión, edad, patrimonio...) y de las características de la operación (motivo del crédito, porcentaje financiado,...), que permitirá inducir las reglas que posteriormente se aplicarán a nuevas solicitudes, determinando así su clasificación.

En este ámbito consideramos pionero el trabajo de Durand (1941) en el que se elaboró un modelo discriminatorio de clasificación de créditos personales, demostrando que la estabilidad en el trabajo seguida de la propiedad de la vivienda eran las características más determinantes en la probabilidad de devolución. Más recientemente, la utilización de nuevas metodologías ha logrado innumerables mejoras en los resultados como recogen los estudios de Greene (1992), Altman y otros (2008) y Gordy (2000), entre otros.

Sin embargo los avances no se han centrado exclusivamente en la modelización del problema, así Majnoni y otros (2004), con información de Argentina, Brasil y México, demuestran la mayor capacidad predictiva de los modelos al sustituir la información negativa (incumplimientos y atrasos) por positiva y descriptiva de las deudas. En el ámbito de las PYMEs se tiende a combinar información personal del titular y del negocio, como en el trabajo de Miller y Rojas (2005) los cuales hacen *credit scoring* para pequeñas empresas de México y Colombia, mientras que Milena y otros (2005) lo aplican a microfinancieras de Nicaragua.

Igualmente el establecimiento de un *cut off* o punto de corte que determina las solicitudes aceptadas por las entidades financieras ha sido objeto de estudio. En el sistema financiero argentino encontramos el trabajo de Pailhé (2006) en el que los solicitantes de créditos con un *score* inferior al *cut off* son rechazados automáticamente, mientras que aquellos con valores superiores deben superar otras etapas de análisis antes de ser aprobada la financiación solicitada.

El interés del estudio aquí realizado es doble. En primer lugar, realizamos un estudio comparativo entre dos modelos paramétricos (Análisis Discriminante y Logit) y uno no paramétrico (Árboles de Clasificación) con objeto de determinar la precisión de los mismos en el problema analizado. Analizando, además, la mejora obtenida cuando se diferencia entre los tres estados que deben existir en todo proceso de aprobación de

un crédito: conceder o rechazar de forma automática la solicitud presentada, o estudiar de forma manual. En este último caso, el programa informático no otorga un nivel de probabilidad de cumplimiento o incumplimiento que permita aceptar o rechazar la solicitud con un alto grado de certeza. Por lo que se recomienda al analista estudiar individualmente la información aportada por el cliente, con objeto de determinar la conveniencia o no de conceder dicho crédito.

En segundo lugar, pensamos que muchos de los estudios realizados no abordan con suficiente cuidado el dilema del aprendizaje-generalización característico de los modelos no paramétricos, es decir, muestran los resultados en una situación particular para la cual el modelo en cuestión ofrece una buena capacidad predictiva. Como es lógico, esta no es la situación real a la que uno se enfrenta, en la que un determinado decisor debe elegir el modelo más adecuado antes de disponer de las observaciones que empleará para validarlo. En el presente trabajo acometemos tal problema mediante el empleo de la validación cruzada.

La estructura seguida en el desarrollo de nuestro estudio es la siguiente: en la sección segunda se explican los fundamentos teóricos de las técnicas paramétricas y no paramétricas. En la sección tercera se presenta la metodología utilizada en un problema de *credit scoring*: Análisis Discriminante (AD), Logit, Árboles de Clasificación (CART) y validación cruzada. En la sección cuarta presentamos el trabajo empírico realizado, para finalizar en la sección quinta con las principales conclusiones obtenidas.

MODELOS DE CLASIFICACIÓN PARAMÉTRICOS Y NO PARAMÉTRICOS

Como es sabido, los problemas de estimación y predicción pueden ser tratados por una gran variedad de técnicas estadísticas que, dependiendo del conocimiento o no de la forma funcional que explica la variable dependiente, se clasifican en paramétricos y no paramétricos. Los modelos paramétricos parten de una función de distribución conocida, y reducen el problema a estimar los parámetros que mejor ajusten las observaciones de la muestra. Dichos modelos resultan muy potentes cuando el proceso generador de datos sigue la distribución propuesta, aunque pueden llegar a ser muy sensibles frente a la violación de las hipótesis de partida cuando se utilizan muestras de tamaño reducido.

Con objeto de salvar ésta y otras limitaciones, se emplean los denominados modelos no paramétricos, conocidos también como métodos de distribución libre, debido a que no se encuentran sujetos a ninguna forma funcional. Dichos modelos presentan pocas restricciones, por lo que en ocasiones resultan más fáciles de aplicar que los paramétricos y permiten «reconstruir» la función de distribución en todo tipo de situaciones, incluidas aquellas en las que la forma funcional sea sencilla y conocida.

Así pues, la diferencia fundamental entre los modelos paramétricos y no paramétricos es la siguiente. Supongamos que la variable dependiente Y puede ser explicada mediante la expresión:

$$Y = f(x_1, x_2, \ldots, x_k) + \varepsilon \qquad (1)$$

Donde: x_i son las variables explicativas
ε es la perturbación aleatoria
f=(x) determina la relación existente entre las variables utilizadas

Los modelos paramétricos suponen conocida la forma funcional de $f(x)$, reduciéndose el problema a determinar los parámetros que la definen. Por su parte, los modelos no paramétricos emplean formas funcionales flexibles que permiten formular una función $\hat{f}(x)$, de manera, que sea una buena aproximación de $f(x)$. Es decir, el problema consiste en calcular los parámetros de una función $\hat{f}(x)$, y no los parámetros de una función conocida. En ambos casos es necesario estimar los parámetros de los que depende la forma funcional. Sin embargo, en el caso de los modelos paramétricos, la elección de dicha forma funcional se establece a priori, por lo que una elección inadecuada se traducirá en un modelo que no ajuste los datos (por ejemplo, supuesta una relación lineal entre las variables, dicha función presentará un mal ajuste cuando la respuesta es, entre otras, cuadrática).

Dadas las características del problema que nos proponemos analizar, donde es difícil suponer una relación funcional clara entre las variables del problema, los modelos paramétricos podrían parecer, a priori, que no poseen la flexibilidad suficiente para ajustarse a todo tipo de situaciones. Por otra parte, y en lo que respecta a su capacidad predictiva, existen algunos estudios que demuestran su inferioridad frente a los modelos no paramétricos (TAM Y KIANG, 1992;

ALTMAN y otros, 1994). Como modelos no paramétricos encontramos el algoritmo CART, C4.5, MARS y las redes neuronales, entre otros. En el trabajo de Bonilla y otros (2003) se realiza un análisis comparativo entre diversos modelos de clasificación crediticia, demostrando que las redes neuronales mejoran significativamente los resultados. Sin embargo, uno de los mayores inconvenientes de esta técnica es la gran cantidad de tiempo necesaria para correcto desarrollo de la aplicación, pues en caso contrario podría alcanzarse un mínimo local que desvirtuaría los resultados.

En el análisis empírico hemos utilizado el AD, el Logit y el algoritmo CART, modelos que, recientemente, han sido utilizados con éxito en problemas de clasificación (CARDONA, 2004; DE SOUZA y OLIVIERA, 2009; XU y ZHANG, 2009, entre otros). En el artículo pretendemos estudiar la precisión de los mismos con objeto de determinar su potencia en la concesión o no de los créditos solicitados.

METODOLOGÍA

Los modelos de *credit scoring*, como hemos indicado, tratan de obtener, a partir la relación existente entre diversas variables que definen tanto al solicitante como a la operación, una regla general que permita determinar, con rapidez y fiabilidad, la probabilidad de fallido de una determinada solicitud. Por tanto, resulta imprescindible estudiar las relaciones existentes entre la información recogida de cada una de los créditos concedidos en el pasado y los impagos observados.

Realizado este análisis, y utilizando un sistema de puntuación establecido en función de las características del cliente, se podrá determinar la probabilidad de que éste pueda o no afrontar sus obligaciones de pago. Así, el problema al que nos enfrentamos puede especificarse mediante la siguiente expresión:

$$P = f(x_1, x_2, \ldots, x_k) + \varepsilon \qquad (2)$$

Donde:
x_i serán los atributos del sujeto
ε la perturbación aleatoria
$f(x)$ la función que determina la relación existente entre las variables utilizadas
P la probabilidad de que el crédito resulte fallido.

El objetivo principal de los modelos de clasificación se centra en estimar la función que permita ajustar con la máxima exactitud las observaciones de la muestra, de manera que el error incurrido en la predicción sea mínimo. Dependiendo de que la forma funcional, $f(x)$, sea conocida o desconocida estaremos ante modelos paramétricos o no paramétricos, como hemos indicado anteriormente. El problema que estamos analizando conlleva una decisión no estructurada, ya que no existe ningún patrón estandarizado que establezca qué variables utilizar, a lo que se añade la dificultad de tener que especificar a priori una forma funcional.

A pesar de esta gran limitación, y de las inherentes a cada uno de los modelos que analizaremos a continuación, los modelos estadísticos ofrecen, generalmente, buenos resultados, por lo que estas técnicas estadísticas, tanto paramétricas como no paramétricas, son consideradas herramientas de gran utilidad para la adecuada toma de decisiones en la empresa.

Análisis discriminante

El análisis discriminante (FISHER, 1936) es una técnica estadística multivariante que permite estudiar de forma simultánea el comportamiento de un conjunto de variables independientes, con objeto de clasificar un colectivo en una serie de grupos previamente determinados y excluyentes. El método presenta la gran ventaja de poder contemplar conjuntamente las características que definen el perfil de cada grupo, así como las distintas interacciones que pudieran existir entre ellas.

Las variables independientes representan las características diferenciadoras de cada individuo, siendo éstas las que permiten realizar la clasificación. Indistintamente se denominan variables clasificadoras, discriminantes, predictivas, o variables explicativas.

De este modo se puede establecer que el objetivo del AD es doble:

- Obtener las mejores combinaciones lineales de variables independientes que maximicen la diferencia entre los grupos. Estas combinaciones lineales reciben el calificativo de funciones discriminantes,
- Predecir, en base a las variables independientes, la pertenencia de un individuo a uno de los grupos establecidos a priori. De este modo se evalúa la potencia discriminadora del modelo.

Para el logro de estos objetivos, la muestra de observaciones se divide aleatoriamente en dos

submuestras: una primera, conocida como muestra de entrenamiento, que se utilizará para la obtención de las funciones discriminantes, y una segunda, denominada muestra de test, que servirá para determinar la capacidad predictiva del modelo obtenido.

Por tanto, el objetivo del AD consiste en encontrar las combinaciones lineales de variables independientes que mejor discriminen los grupos establecidos, de manera que el error cometido sea mínimo. Para ello será necesario maximizar la diferencia entre los grupos (variabilidad entre grupos) y minimizar las diferencias en los grupos (variabilidad intragrupos), obteniendo así el vector de coeficientes de ponderación que haga máxima la discriminación.

Con objeto de asegurar la potencia discriminadora del modelo es necesario establecer fuertes hipótesis de partida que van a suponer una limitación para el análisis de cualquier problema de clasificación que se presente. Éstas son:

- Las *K* variables independientes tiene una distribución normal multivariante.
- Igualdad de la matriz de varianzas-covarianzas de las variables independientes en cada uno de los grupos.
- El vector de medias, las matrices de covarianzas, las probabilidades a priori, y el coste de error son magnitudes todas ellas conocidas.
- La muestra extraída de la población es una muestra aleatoria.

Tan sólo bajo estas hipótesis la función discriminante obtenida será óptima. Las dos primeras hipótesis (la normalidad y de igualdad de la matriz de varianzas y covarianzas) difícilmente se verifican en muestras de carácter financiero, cuestión que no impide al AD obtener buenas estimaciones, aunque realmente éstas no puedan considerarse óptimas.

Modelo Logit

El modelo Logit permite calcular la probabilidad de que un individuo pertenezca o no a uno de los grupos establecidos a priori. La clasificación se realizará en función del comportamiento de una serie de variables independientes que son características de cada individuo. Se trata de un modelo de elección binaria en el que la variable dependiente tomará valores 1 ó 0. En nuestro problema el valor dependerá de que el individuo haya hecho o no frente a sus obligaciones crediticias. Si se presentara una situación en la que el sujeto tuviera que elegir entre tres o más alternativas mutuamente excluyentes (modelos de elección múltiple), tan sólo se tendría que generalizar el proceso.

El modelo Logit queda definido por la siguiente función de distribución logística obtenida a partir de la probabilidad a posteriori aplicada al AD mediante el teorema de Bayes,

$$P_i = P(Y = 1/X) = F(Z_i) = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta Z_i)}} \quad (3)$$

Donde:
β_0 representa los desplazamientos laterales de la función logística
β es el vector de coeficientes que pondera las variables independientes y del que depende la dispersión de la función
X es la matriz de variables independientes.

Al igual que el modelo discriminante, el Logit es un modelo multivariante paramétrico en el que existen variables categóricas tanto en el conjunto de variables explicativas como en de las variables dependientes. Frente al AD presenta la gran ventaja de que no va a ser necesario establecer ninguna hipótesis de partida: no plantea restricciones ni con respecto a la normalidad de la distribución de variables, ni a la igualdad de matrices de varianzas-covarianzas. Ahora bien, corresponde señalar que, en caso de verificarse dichas hipótesis, el modelo discriminante obtendría mejores estimadores que el Logit, pues según afirma Efron (1975) «...bajo estas circunstancias, los estimadores logísticos resultan bastante menos eficientes que los de la función discriminante».

La mayoría de los problemas financieros con los que nos enfrentamos utilizan alguna variable cualitativa, imposibilitando de este modo el cumplimiento de la hipótesis de normalidad, siendo el modelo Logit con los estimadores de máxima verosimilitud claramente preferible. En este sentido, Press y Wilson (1978) enumeran los distintos argumentos existentes en contra de la utilización de los estimadores de la función discriminante, presentando, asimismo, dos problemas de clasificación cuyas variables violan dicha restricción. Ambos problemas se resolvieron mediante el AD y el Logit quedando claramente demostrada la superioridad de este último.

A pesar de estas limitaciones, la literatura sigue avalando la utilización de ambos modelos linea-

les (LENNOX, 1999; CALVO-FLORES y otros, 2006; BEAVER y otros, 2008).

Método de partición recursiva: árboles de clasificación y regresión (CART)

Los árboles de decisión son una técnica no paramétrica que reúne las características del modelo clásico univariante y las propias de los sistemas multivariantes. Originariamente fueron propuestos para separar las observaciones que componen la muestra asignándolas a grupos establecidos a priori, de forma que se minimizara el coste esperado de los errores cometidos. Esta técnica fue presentada por Friedman en 1977, pero originariamente sus aplicaciones a las finanzas no fueron muy numerosas, si bien corresponde destacar dos estudios pioneros: Frydman y otros (1985) en el que utilizan el modelo para clasificar empresas, comparando su capacidad predictiva con el AD, y Marais y otros (1984) que, por el contrario, lo aplican a préstamos bancarios. En ambos trabajos se ha llegado a demostrar la gran potencia que presenta este algoritmo como técnica de clasificación.

El modelo CART supone esencialmente que las observaciones son extraídas de una distribución ϕ en $L \times X$, donde L es el espacio de categorías, y X el espacio de características. Las densidades condicionales $\phi(x|l)$ difieren al variar l, y las probabilidades marginales $\phi(l)$ son conocidas. El proceso utiliza la muestra S como conjunto de entrenamiento para la estimación no paramétrica de una regla de clasificación que permita particionar directamente el espacio X de características. Para cada l de L, el subconjunto St del conjunto de entrenamiento S constituye una muestra aleatoria de la distribución condicional $\phi(x|l)$ en X.

Así pues, el proceso consiste en dividir sucesivamente la muestra original en submuestras, sirviéndose para ello de reglas univariantes que buscarán aquella variable independiente que permita discriminar mejor la división. Con ello, se pretende obtener grupos compuestos por observaciones que presenten un alto grado de homogeneidad, incluso superior a la existente en el grupo de procedencia (denominado nodo madre).

Con objeto de encontrar la mejor regla de división, el algoritmo estudiará cada una de las variables explicativas, analizando los puntos de corte para, de este modo, poder elegir aquella que mayor homogeneidad aporte a los nuevos subgrupos. El proceso finaliza cuando resulte imposible realizar una nueva división que mejore la homogeneidad existente.

El modelo, como vemos en la Figura 1, se estructura en forma de árbol compuesto de una sucesión de nodos y ramas que constituyen, respectivamente, los grupos y divisiones que se van realizando de la muestra original. Cada uno de los nodos terminales representa aquel grupo cuyo coste esperado de error sea menor, es decir, aquellos que presenten menor riesgo. El riesgo total del árbol se calcula sumando los correspondientes a cada uno de los nodos terminales.

En definitiva, el algoritmo de partición recursiva puede resumirse en los siguientes cuatro pasos:

1. Estudiar todas y cada una de las variables explicativas para determinar para cuál de ellas y para qué valor es posible incrementar la homogeneidad de los subgrupos. Existen diversos criterios para seleccionar la mejor división de cada nodo, todos ellos buscan siempre aquella división que reduzca más la impureza del nodo, definida ésta mediante la siguiente expresión,

$$i(t) = -\sum p(j/t) \cdot \log[p(j/t)] \qquad (4)$$

siendo $p(j/t)$ la proporción de la clase j en el nodo t. Como medida de la homogeneidad o impureza se utiliza una extensión del índice de Gini para respuestas categóricas. El algoritmo optará por aquella división que mejore la impureza, mejora que se mide comparando la que presenta el nodo de procedencia con la correspondiente a las dos regiones obtenidas en la partición.
2. El paso anterior se repite hasta que, o bien resulte imposible mejorar la situación realizando otra división, o bien el nodo obtenido presente el tamaño mínimo. En esta fase del algoritmo se obtiene el árbol máximo en el cual cada uno de sus nodos interiores es una división del eje de características.
   Ahora bien, este procedimiento, tal y como ha sido expuesto, presenta un grave problema, el sobreaprendizaje: el modelo memoriza las observaciones de la muestra siendo incapaz de extraer las características más importantes, lo que le impedirá «generalizar adecuadamente», obteniendo resultados erróneos en los casos no contemplados con anterioridad. Para evitarlo Friedman (1977) propuso la siguiente solución: desarrollar el árbol al máximo, y posteriormente ir podándolo eliminando las divisiones y, por lo tanto, los nodos que presenten un mayor coste de complejidad, hasta

encontrar el tamaño óptimo, que será aquel que minimice el coste de complejidad.

3. Calcular la complejidad de todos y cada uno de los subárboles podando aquellos que verifiquen la siguiente expresión,

$$R_K(T_1) = \min_{T' \leq T} R_K(T') \quad (5)$$

siendo el coste de complejidad,

$$R_K(T) = [R(T) + K \cdot |T| \quad (6)$$

donde $R_K(T)$ es el coste de complejidad del árbol T para un determinado valor del parámetro K, $R(T)$ es el riesgo de errar (K se denomina parámetro de complejidad que penaliza la complejidad del árbol y siempre será positivo) y $|T|$ es el número de nodos terminales.

4. Encontrar todos los valores críticos de K, y utilizar la técnica de validación cruzada para cada uno de ellos con objeto de estimar $R(T(K))$, eligiendo aquella estructura que presente mejor valor estimado de $R(T(K))$.

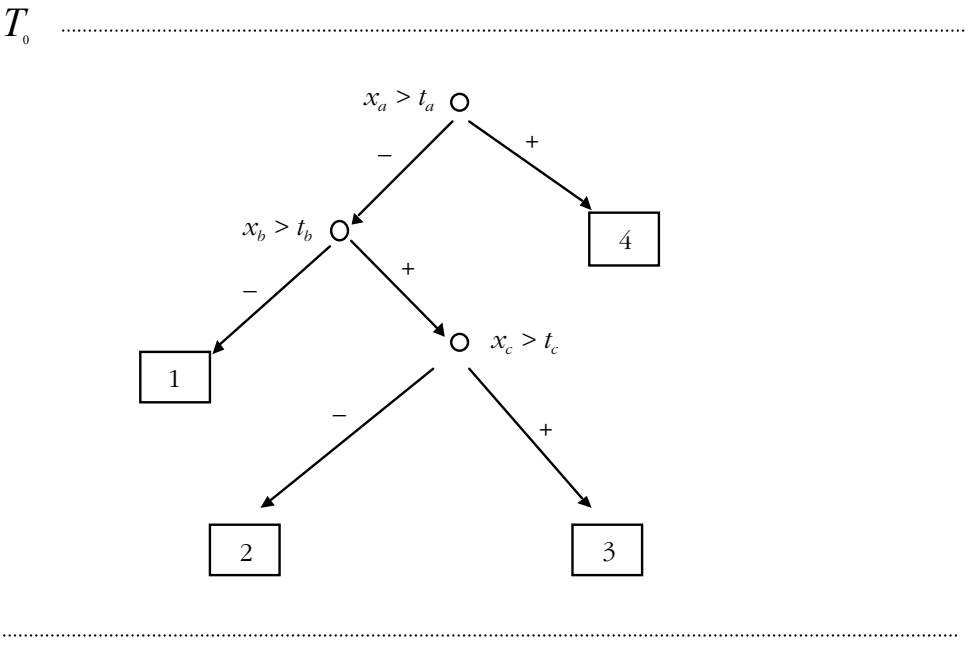
Por tanto, el principal problema con el que se enfrenta este modelo es la complejidad de su estructura que, como ya hemos indicado, fácilmente puede desembocar en el sobreaprendizaje del modelo. De ahí que no sólo se persiga crear conjuntos homogéneos con bajo riesgo, sino también obtener aquella estructura que presente una complejidad óptima. Bajo este doble objetivo, resulta necesario penalizar la excesiva complejidad del árbol.

Método de validación cruzada

Así pues el problema que presenta el algoritmo CART, y además común al resto los métodos no paramétricos, es el sobreaprendizaje. La aparición de este fenómeno puede atribuirse fundamentalmente a dos causas (Figura 2):

1. La sobreparametrización, el modelo presenta una estructura más compleja de la necesaria para tratar el problema en cuestión.
2. La escasez de datos que impide al modelo extraer en la fase de entrenamiento las características más relevantes de la muestra y, posteriormente, en la fase de test, verificar la capacidad predictiva del modelo con otra muestra de datos distinta a la utilizada en el entrenamiento.

Figura 1 – Árboles de Decisión



Supongamos que disponemos de un conjunto de observaciones y lo dividimos en dos: un conjunto de entrenamiento, que servirá para ajustar el modelo, y un conjunto de test que será empleado para validarlo. En el eje de abscisas hemos representado el número de parámetros de un determinado modelo (siendo el modelo más complejo, es decir, el de mayor número de parámetros, el más alejado del origen), y en el eje de ordenadas se cuantifica el error cometido sobre los conjuntos de aprendizaje y test.

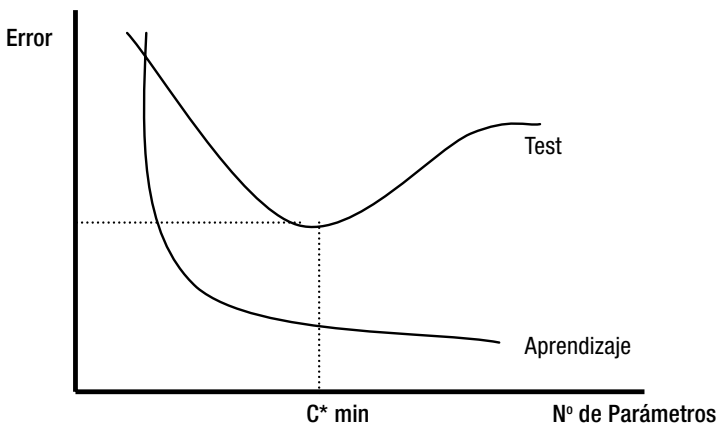
Cuando la estructura del modelo es muy simple, éste es incapaz de capturar la relación subyacente entre los atributos y la variable respuesta, por lo que cometerá un elevado porcentaje de fallos tanto sobre el conjunto de entrenamiento como sobre el de test. A medida que el número de parámetros aumenta, va adquiriendo suficiente potencia o flexibilidad, lo que le permitirá aprender la relación existente entre las variables independientes y dependiente, relación que debe verificarse sobre ambos conjuntos, por lo que el error cometido irá decreciendo.

Si incrementamos sucesivamente la complejidad, el error a lo largo del conjunto de entrenamiento seguirá disminuyendo progresivamente, es decir, el modelo se irá acomodando a las características peculiares de los ejemplos propios de dicho conjunto que no tienen porque estar presentes en el de test. Por este motivo, llegados a un punto, *C**, el error incurrido sobre el conjunto de test, que es el que determina la potencia predictiva del modelo, se incrementará considerablemente. Por consiguiente, a partir de *C** la estructura es tan compleja que el modelo ha memorizado la muestra, lo que se traduce en una débil capacidad de generalización.

Con objeto de evitar el sobreaprendizaje se viene utilizando, entre otros, el método de validación cruzada propuesto por Stone (1974) que, como veremos en los resultados del trabajo, hemos empleado para elegir la estructura idónea de los Árboles de Regresión, es decir, aquella que facilite la obtención de una adecuada generalización del problema que estamos analizando. El proceso se estructura en los siguientes pasos:

- Las observaciones que componen la muestra se dividen en dos: un conjunto de entrenamiento, que sirve para ajustar el modelo, y un conjunto de test que es empleado para validarlo.
- El conjunto de entrenamiento se divide aleatoriamente en 10 particiones distintas.
- Por rotación, un conjunto de 9 particiones se utiliza para estimar el modelo con un número de parámetros determinado, y la décima partición para contrastar su capacidad predictiva.
- El paso 2 se repite diez veces, de forma que el algoritmo, utilizando distintas estructuras, va a ser entrenado y testeado con 10 pares distintos de conjuntos de entrenamiento y test, siendo la estructura óptima aquélla que minimice el error de predicción a lo largo de los 10 conjuntos de test

Figura 2: Proceso de sobreaprendizaje de los modelos



(a este error se le denomina error de validación cruzada, EVC).

- Elegida la estructura óptima, se utilizará toda la muestra para reentrenar el modelo, de manera que se entrenará y testeará con los conjuntos totales para obtener el error de predicción (EP). Debido a que el EVC es un estimador insesgado del error de predicción del modelo elegido (EP), el modelo seleccionado tendrá también una capacidad de generalización óptima cuando sea empleado con observaciones no presentes en el conjunto de entrenamiento.

En concreto, en nuestro análisis, la muestra formada por 1446 observaciones se ha dividido aleatoriamente en dos conjuntos de entrenamiento y test, cada uno de ellos formado por el 50% de las observaciones. La submuestra de test (723 observaciones) ha sido reservada con objeto de testear la capacidad generalizadora del modelo (EP). Las observaciones restantes se han utilizado como conjunto de entrenamiento para elegir aquel modelo cuya estructura presente el menor EVC. Lo conforman 231 individuos calificados como fallidos y 492 como no fallidos. La magnitud de tales submuestras no se corresponden con ningún criterio *ad hoc*, pero están en línea con la literatura al respecto.

Las 723 observaciones destinadas al entrenamiento se han dividido a su vez en conjuntos de entrenamiento y test, representando el conjunto de test el 10% de la muestra. Las observaciones de estas dos submuestras se han combinado de tal forma que disponemos de 10 pares no solapados de conjuntos de entrenamiento y test formados por 651 y 72 observaciones respectivamente, que se utilizarán en la obtención del EVC.

De manera que, el modelo CART ha sido entrenado y testeado con estos 10 pares de conjuntos utilizando distintas estructuras, para, de este modo, poder determinar su estructura óptima. Dicha estructura será aquella que presente el menor EVC, calculado éste como una media de los errores cometidos a lo largo de los 10 conjuntos de test validados. El parámetro que determina la complejidad de los Árboles de Clasificación es el número de nodos, por lo que la selección de su estructura óptima consiste en determinar el número óptimo de nodos.

Para finalizar, el modelo elegido será entrenado y testeado con la muestra total (723 observaciones de entrenamiento y 723 de test) con objeto de obtener el EP que nos permitirá comparar la potencia predictiva del algoritmo en cada uno de los sectores analizados.

Este proceso de división se repitió 10 veces hasta obtener 100 pares de conjuntos de entrenamiento y test distintos, con objeto de eliminar la posible incidencia que la división de la muestra podía tener en los resultados de nuestro análisis. Además, en todos los conjuntos se mantuvo la misma proporción de morosos y no morosos existente en la muestra original.

ANALISIS DE LOS RESULTADOS

En el desarrollo del trabajo empírico hemos utilizado una base de datos de préstamos al consumo facilitada por una de las principales entidades financieras de nuestro país (España), de la que no se aporta el nombre por motivos de confidencialidad. Dicha base de datos está formada por 1446 individuos, de los cuales 462 fueron calificados como morosos y 984 como no morosos. El individuo era calificado como moroso si se retrasaba más de dos meses en el pago del mismo, con independencia de que posteriormente hiciera o no frente a sus obligaciones con la entidad.

Cada caso viene definido por 13 variables explicativas, así como por el comportamiento crediticio posterior. La variable dependiente se ha denotado con 1 si al individuo al que se le concedió el préstamo resultó moroso y 0 en caso contrario. Las variables independientes que caracterizan la solicitud de cada cliente son las siguientes: finalidad del préstamo (por ejemplo, compra de una motocicleta, de una vivienda, gastos por estudios, etc.), documentación aportada (si tan sólo se presentó el D.N.I. o se aportaron otro tipo de documentos, tales como avales, certificado del registro, etc.), estado civil, ingresos anuales, edad del solicitante, antigüedad de la cuenta en años, saldo de la cuenta, vinculación con la caja (una variable subjetiva proporcionada por la entidad, según fuera considerado el cliente), tenencia de otros préstamos, importe del préstamo solicitado, y plazo del préstamo solicitado.

Para el algoritmo CART la muestra se ha dividido 10 veces en 10 pares de conjuntos de entrenamiento y test, es decir, el proceso de validación cruzada se ha repetido 10 veces con objeto de evitar que las posibles divisiones de la muestra puedan afectar a los resultados. De manera que, ha exigido un intenso esfuerzo computacional, dado que se han utilizado una gran variedad de estructuras para obtener la que realmente

minimice el ECV, es decir aquella que permita obtener generalizaciones adecuadas.

En el caso de los modelos paramétricos, la estructura del mismo viene establecida, por lo que no es necesario utilizar la validación cruzada. Sin embargo, e igualmente para evitar que la división de los conjuntos de entrenamiento y test pueda afectar a la precisión de los modelos, la muestra ha sido dividida diez veces en dos conjuntos de entrenamiento y test, cada uno de ellos formado por el 50% de las observaciones.

En las Tablas 1, 2 y 3 se facilitan los resultados obtenidos a lo largo de las 10 simulaciones utilizando los algoritmos propuestos.

Resultados: fallidos y no fallidos

Los modelos no paramétricos resultan mucho más flexibles que los paramétricos pues, como hemos indicado anteriormente, a priori no precisan establecer ninguna forma funcional, sino que ajustarán aquella que mejor aproximen las variables del modelo.

El problema que tratamos de analizar establece la relación entre variables muy diversas, algunas de ellas cualitativas, por lo que consideramos difícil poder determinar a priori la relación funcional entre ellas. Todo ello nos conduce a suponer que los modelos no paramétricos, en nuestro caso el CART, presentarán un mejor comportamiento que los paramétricos. Sin embargo, a la vista de los resultados obtenidos en la Tabla 4, podemos afirmar que en el problema de clasificación crediticia que nos hemos propuesto analizar, el AD, aunque levemente, ha superado la capacidad predictiva del modelo CART y del Logit, lo que, en cierta medida, contradice la literatura encontrada al respecto.

El problema que surge cuando se realiza esta clasificación (conceder o no conceder) es que algunos de los individuos han sido calificados como aciertos, es

Tabla 1 – CART: Medias de los EVC y EP

	Porcentajes de Error CART										
Simulaciones	**Sim 1**	**Sim 2**	**Sim 3**	**Sim 4**	**Sim 5**	**Sim 6**	**Sim 7**	**Sim 8**	**Sim 9**	**Sim 10**	**Medias**
EVC											
Entrenamiento	24.95	23.11	23.94	23.55	24.65	22.26	24.26	24.52	23.52	22.36	**23.71**
Test	28.18	26.7	28.64	28.6	26.42	26.42	28.21	29.73	25.7	26.84	**27.54**
E. Predicción											
Entrenamiento	24.48	21.16	22.54	23.24	24.76	22.54	24.76	23.65	23.37	22.96	**23.34**
Test	25.59	27.25	30.71	28.91	29.05	27.66	26	29.6	26.28	27.94	**27.89**

Tabla 2 – Análisis Discriminante: Medias de los EP

Simulaciones	**Sim 1**	**Sim2**	**Sim 3**	**Sim 4**	**Sim 5**	**Sim 6**	**Sim 7**	**Sim 8**	**Sim 9**	**Sim 10**	**Medias**
E. Predicción											
Entrenamiento	26.69	27.38	26.95	27.80	27.38	27.10	26.55	26.55	27.93	27.80	**27.21**
Test	27.80	28.49	29.46	27.93	27.80	29.18	28.07	29.46	28.07	28.63	**28.48**

Tabla 3 – Logit: Medias de los EP

Simulaciones	**Sim 1**	**Sim 2**	**Sim 3**	**Sim 4**	**Sim 5**	**Sim6**	**Sim 7**	**Sim 8**	**Sim 9**	**Sim 10**	**Medias**
E. Predicción											
Entrenamiento	27.10	26.14	26.83	28.07	25.17	27.38	26.27	27.10	28.21	26.97	**26.92**
Test	27.10	27.52	28.76	27.38	26.14	28.07	27.38	26.83	27.66	27.66	**27.45**

decir, que dadas sus características el modelo aconseja aprobar/rechazar el crédito solicitado, simplemente porque su probabilidad de cumplimiento/incumplimiento es mayor/menor que 0,5, respectivamente. En este estudio se propone, conceder automáticamente tan sólo los que con alto grado de fiabilidad podrán atender a sus obligaciones crediticias, y rechazar, igualmente de forma automática, en caso contrario, por lo que sería necesario distinguir un tercer estado que hemos denominado como dudoso.

Resultados: fallidos, no fallidos y dudosos

En el estudio que presentamos se plantea la posibilidad de dividir el intervalo de acierto en tres estados:

- Fallidos: todos aquellos préstamos cuya probabilidad de devolución sea menor que el 30%
- No Fallidos: todos aquellos préstamos cuya probabilidad de devolución sea menor que el 70%
- Dudosos: todos aquellos préstamos cuya probabilidad de devolución esté comprendida entre el 30% y el 70%.

Con esta clasificación se pretende conseguir una perfecta discriminación entre los que con un alto grado de probabilidad podrán hacer frente a sus obligaciones crediticias, y los que por el contrario resultarán fallidos, recomendándose de este modo la no concesión del préstamo.

Todos ellos podrán ser automáticamente aceptados o rechazados dada la precisión que se le ha exigido al modelo. Sin embargo, el grupo de los dudosos deberá ser estudiado manualmente por el personal bancario, ya que el modelo no asegura el acierto de su decisión. Después de consultar a diversos expertos en riesgos bancarios, conocedores del funcionamiento del proceso de concesión en importantes entidades financieras de nuestro país, hemos llegado a la conclusión que no sólo se debe buscar aquel modelo que minimice el error sino que, igualmente, no contenga un conjunto de dudosos elevado. Las entidades suelen tener autonomía suficiente para poder conceder o rechazar discrecionalmente los créditos considerados como dudosos, por lo que el porcentaje de fracaso de estos dudosos suele ser elevado.

Aplicando nuevamente la validación cruzada diez veces, y distinguiendo entre los tres estados descritos anteriormente, hemos obtenido los siguientes resultados. En la Tabla 5 observamos que el modelo CART, aunque presenta un mayor porcentaje de error, el conjunto de los dudosos es mucho menor y el de aciertos mayor que los obtenidos mediante los modelos paramétricos. Los dudosos normalmente se convierten en fallidos, por lo que resulta interesante que dicho grupo sea lo menor posible. Así pues, apoyándonos en los resultados, podemos afirmar que el modelo CART ha resultado ser bastante más preciso que los modelos Logit y AD.

Comparando estos resultados con los obtenidos en el apartado 4.1. comprobamos que, aunque el porcentaje de acierto es menor, el de error se ha reducido considerablemente, que realmente es el que

Tabla 4 – Errores de predicción del modelo CART, Logit y AD

	Modelo CART		Modelo Logit		Modelo AD	
Créditos	Entrenamiento	Test	Entrenamiento	Test	Entrenamiento	Test
Aciertos	76,66	72,11	72,79	71,52	73,08	72,55
Errores	23,34	27,89	27,21	28,48	26,92	27,45

Tabla 5 – Errores de predicción de los modelos CART, Logit, AD

	Modelo CART		Modelo Logit		Modelo AD	
Créditos	Entrenamiento	Test	Entrenamiento	Test	Entrenamiento	Test
Dudosos	27,61	27,01	45,47	45,42	45,41	45,86
Aciertos	58,29	56,85	45,21	44,53	45,17	43,80
Errores	14,11	16,14	9,21	10,04	9,40	10,33

produce mayor quebranto a las entidades. Además los créditos concedidos presentan una mayor probabilidad de cumplir con sus obligaciones si se distinguen los tres estados que hemos considerado, ya que se le exige una mayor precisión a los algoritmos de clasificación utilizados.

CONCLUSIONES

En el estudio que presentamos hemos realizado un análisis comparativo entre dos técnicas paramétricas y una no paramétrica aplicadas a un problema real de clasificación crediticia. De manera que, dadas unas variables descriptivas del sujeto solicitante de un crédito, el modelo determinará con la mayor precisión posible si sería capaz o no de hacer frente a sus obligaciones crediticias. Igualmente hemos abordado, con suficiente minuciosidad, el problema del sobreaprendizaje que habitualmente se obvia en muchos estudios por el excesivo esfuerzo computacional que requiere.

Los modelos de clasificación crediticia, que habitualmente utilizan las entidades financieras, se alimentan de la propia información que van generando, de manera que si el modelo se equivoca muy frecuentemente, al cabo del tiempo el algoritmo deja de ser operativo porque los resultados que genera no son, en absoluto, fiables. Por ello, y aunque dicha solución en principio pueda parecer un incremento de coste para la entidad, proponemos la distinción de las solicitudes en tres estados: conceder automáticamente (porque la probabilidad de que el cliente devuelva el crédito es superior al 70%); rechazar automáticamente (dado que su probabilidad será inferior al 30%); y distinguir un tercer estado que hemos denominado como «dudosos», en el cual se aconseja su estudio manual por parte del personal bancario.

Lo que genera mayor quebranto a las entidades financieras es la insolvencia, es decir, que los clientes resulten fallidos. En el estudio se demuestra que se reduce considerablemente el error si, todas las solicitudes que no tengan una gran certeza de devolución o de incumplimiento, se procede a su análisis individual para determinar la conveniencia o no de rechazar/conceder dicho crédito. Ahora bien, sabiendo que ello eleva el coste de tramitación de las entidades, consideramos que dicho grupo no debe ser demasiado elevado.

Concluimos que el algoritmo CART obtiene mejores resultados que los modelos paramétricos, porque aunque el error es algo superior, el de aciertos también lo es. Además el porcentaje de calificados como dudosos no resultan excesivamente elevado, como sí ocurre con los otros métodos (AD y Logit).

REFERENCIAS

ALTMAN, E; MARCO, G; VARETTO, F. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. *Journal of Banking and Finance*, v. 18, n. 3, p. 505-529, 1994.

ALTMAN, E.; SABATO, G.; WILSON, N. The value of qualitative information in SME risk management. *Journal of Financial Services Research*, v. 40,n. 2, p. 15-55, 2008.

BEAVER, W.H; CORREIA, M; McNICHOLS, M. Have changes in financial reporting attributes impaired informativeness? Evidence from the ability of financial ratios to predict brnkruptcy. WORKING PAPER. Rock Center for Corporate Governance, Stanford University, December, 2008.

BONILLA, M.; OLMEDA, I.; PUERTAS, R. Modelos paramétricos y no paramétricos en problemas de *credit scoring*. *Revista Española de Financiación y Contabilidad*, v. 27, n. 118, p. 833-869, 2003.

CARDONA, P. Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de Estadística*, v. 27, n. 2, p. 139-151, 2004.

CLAVO-FLORES, A; GARCÍA, D.; MADRID, A. Tamaño, antigüedad y fracaso empresarial. WORKING PAPER. Universidad Politécnica de Cartagena, 2006.

DE SOUZA, A; OLIVIEIRA, W. Prevendo a insolvência de operadoras de planos de saúde, *RAE -Revista de Administração de Empresas*, v. 49, n.4, p. 459-471, 2009.

DURAND, D. Risk elements in consumer installment financing. WORKING PAPER, National Bureau of Economic Research, New York, 1941.

EFRON, B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, v. 70, n. 352, p. 892-898, 1975.

FISHER, R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 7, p. 179-188, 1936.

FRIEDMAN, J.H. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, v. 26, n. 4, p. 404-408, 1977.

FRYDMAN, H; ALTMAN, E; KAO, D. Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance,* v. XL, n. 1, p. 269-291, 1985.

GORDY, M.B. A comparative anatomy of credit risk models. *Journal of Banking & Finance*, v. 24, p. 119-149.

GREENE, W.H. A statistical model for credit scoring. WORKING PAPER, Stern School of Business, New York University, 1992.

LENNOX, C. Identifying failing companies: A Re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business*, v. 51, n. 4, p. 347-364, 1999.

MAJNONI, G.; MILLER, M.; MYLENKO, N. POWELL, A. Improving credit information, bank regulation and supervision: On the role and design of public credit registries. WORKING PAPER, World Bank Policy Research, November, 2004.

MARAIS, M.L; PATELL, J; WOLFSON, M. The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. *Journal of Accounting Research*, v. 22, n. 1, p. 87-114, 1984.

MILENA, E.; MILLER, N.; SIMBAQUEBA, L. The case for information sharing by microfinance institutions: Empirical Evidence of the value of credit bureau-type data in the Nicaraguan microfinance sector. WORKING PAPER, World Bank, 2005.

MILLER, M.; ROJAS, D. Improving access to credit for SMEs: An empirical analysis of the Feasibility of pooled data small business credit scoring models in Colombia y Mexico. WORKING PAPER, World Bank, Washington, 2005.

PAILHÉ, C. Sistemas de información para la administración del riesgo de crédito. Relevamiento en el sistema financiero argentino. WORKING PAPER, Central Bank, October, 2006.

PRESS, J; WILSON, S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association,* v. 73, n. 364, 699-705, 1978.

STONE, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, v. 36, n. 2, p. 11-144, 1974.

TAM, K; KIANG, M. Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, v. 38, n. 7, p. 926-947, 1992.

XU, M; ZHANG, C. Bankruptcy prediction: the case of Japanese listed companies. *Review of Accounting Studies*, v. 14, n. 4, p. 534-558, 2009.