Trabajo Fin de Máster

Scoring No Clientes (Modelización con información Big Data)

Iria Cañete Quian

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo Fin de Máster

Título en galego: Scoring No Clientes (Modelización con información Big Data)

Título en español: Scoring No Clientes (Modelización con información Big Data)

English title: Scoring No Clients (Modeling with Big Data information)

Modalidad: Modalidad B

Autor/a: Iria Cañete Quian, Universidade de Santiago de Compostela

Director/a: Ricardo Cao Abad, Universidade da Coruña;

Tutor/a: Daniel López Souto, ABANCA;

Breve resumen del trabajo:

El scoring es una herramienta determinante para medir el riesgo de crédito en las entidades bancarias. En el presente trabajo se estudian las principales técnicas estadísticas que se utilizan para construir un modelo de scoring. Finalmente, se desarrolla un modelo de scoring comportamental para financiación no clientes a partir de bases de datos internas procedentes de sistemas de agregación de clientes.

Recomendaciones:

Otras observaciones:

Don Ricardo Cao Abad, Catedrático de Universidad de la Universidade da Coruña y don Daniel López Souto, Gestor técnico de ABANCA informan que el Trabajo Fin de Máster titulado

Scoring No Clientes (Modelización con información Big Data)

fue realizado bajo su dirección por doña Iria Cañete Quian para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 5 de Septiembre de 2018.

El director: El tutor:

Don Ricardo Cao Abad Don Daniel López Souto

La autora:

Doña Iria Cañete Quian

Agradecimientos

Me gustaría empezar por agradecer a mi director académico Ricardo Cao Abad por su asesoramiento y sugerencias durante la realización de este trabajo y su apoyo en la elaboración del mismo.

También quiero agradecer a ABANCA por darme la oportunidad de realizar las prácticas en la empresa. En especial a Daniel López Souto y a Cristina González Fragueiro, por toda la ayuda que me proporcionaron en estos meses, así como a todos los compañeros del departamento de Riesgos.

Índice general

Resumen xi

Prólogo xiii

1. Términos bancarios 1

1.1. Definiciones previas . . . 1

1.2. Credit scoring . . . 2

1.2.1. Etapas de construcción de un modelo de scoring . . . 4

2. Regresión logística 7

2.1. Modelo de regresión logística . . . 7

2.2. Estimación de los parámetros del modelo . . . 8

3. Técnicas de agrupamiento de variables 11

3.1. Árboles de decisión . . . 12

3.1.1. Paquete rpart . . . 14

3.2. Árboles de inferencia condicional . . . 17

3.2.1. Función ctree del paquete party . . . 18

4. Medidas de predicción 21

4.1. Valor de información y peso de la evidencia . . . 21

4.2. Ejemplo de cálculo de IV con R . . . 27

4.2.1. Paquete information . . . 27

4.2.2. Paquete smbinning . . . 30

4.3. Estadístico chi-cuadrado . . . 30

4.3.1. Ejemplo del cálculo del estadístico chi-cuadrado . . . 31

5. Técnicas de validación del modelo 35

5.1. Criterios de selección de variables y comparación de modelos . . . 35

5.1.1. Algoritmos de selección de variables paso a paso . . . 35

5.1.2. Criterios globales . . . 36

5.2. Pruebas de diferencias de dos poblaciones 37
5.2.1. Contraste de Kolmogorov-Smirnov 38
5.2.2. Validación mediante curvas ROC 39
5.2.3. Validación mediante curvas de Lorentz e índice de Gini 45

6. Construcción de la tabla de puntuaciones **47**

7. Aplicación con datos reales **49**
7.1. Tabla de solo cuenta corriente o Tabla 1 52
7.2. Tabla cuenta corriente y producto de riesgo o Tabla 2 65

8. Conclusiones **73**

A. Resumen de acrónimos **75**

Bibliografía **77**

Resumen

Resumen en español

El scoring es una herramienta determinante para medir el riesgo de crédito en las entidades bancarias. En la primera parte de este trabajo, se estudian las principales técnicas estadísticas utilizadas para la construcción de un modelo de scoring. En el primer paso para esta construcción, se agrupan las características en categorías y se calcula su capacidad predictiva utilizando medidas de predicción, como el valor de la información y el peso de la evidencia. Las variables que presentan malas capacidades predictivas son eliminadas del modelo, por lo que se realiza una selección de variables. Una vez obtenido un modelo cuyas variables presenten un buen poder predictivo, este modelo se valida utilizando las curvas ROC o el índice de Gini, entre otros.

En la segunda parte de trabajo, se aplican estas técnicas a datos reales y se desarrolla un modelo de scoring comportamental para financiación de no clientes a partir de bases de datos internas procedentes de sistemas de agregación de clientes.

English abstract

The scoring is a decisive tool to measure the credit risk at banks. In the first part of this paper, the characteristics are binned into categories and their predictive capacity is calculated using predictive measures, such as information value and the weight of evidence. Variables with poor predictive capabilities are removed from the model, so a selection of variables is made. Once a model is obtained whose variables have a good predictive power, it is validated using ROC curves or Gini index, among others.

In the second part of the project, these techniques are applied to real data and a scoring behavioural model is developed for financing of non-customers from internal databases coming from account aggregation systems.

Prólogo

La actividad principal de las entidades financieras es la intermediación bancaria. Esta intermediación consiste en tomar fondos de una serie de clientes para dejárselos a otros clientes. Esta actividad intermediadora conlleva un riesgo asociado pues las entidades deben asegurar que el dinero que prestan les es devuelto. Este riesgo se conoce como riesgo de crédito y se define como la posibilidad de sufrir una pérdida como consecuencia de un impago por parte de nuestra contrapartida en una operación financiera, es decir, el riesgo de que el cliente no pague. Este tipo de riesgo puede comprometer la estabilidad de las entidades financieras si la proporción de créditos en impago alcanza niveles elevados.

Para gestionar el riesgo de crédito de un modo eficiente las entidades calculan la pérdida esperada asociada a cada operación que conceden. Esta pérdida esperada se calcula de la siguiente forma:

$$PE = PD \times EAD \times LGD$$

donde

PE pérdida esperada.

PD probabilidad que una operación entre en morosidad/default (incumplimientos de más de 90 días) en un horizonte especificado.

EAD exposición de la operación en el momento de entrar en default.

LGD pérdida máxima que se produce una vez que la operación entra en mora (aquí se consideran las garantías que la operación pueda tener).

Estos términos se definirán con más detalle en el capítulo 1. Para calcular la probabilidad de entrada en morosidad se utilizan modelos de scoring, sobre todo pensados para las personas físicas, o modelos de rating en personas jurídicas (empresas).

Muchas entidades consideran el scoring una herramienta determinante para conceder o no crédito ya que posibilita trabajar con muchas peticiones de créditos de manera rápida, aparte de reducir la tasa de morosidad frente a la toma de decisiones humanas.

Adicionalmente, en función del momento de la utilización de los modelos, estos se pueden dividir en:

- Modelos de admisión. Se utilizan para determinar si una operación de crédito se concede o no.
- Modelos para el seguimiento. Se utilizan en operaciones ya concedidas para calcular la probabilidad de entrada en mora en un horizonte temporal. Esto es muy útil para el establecimiento de provisiones.

Otra clasificación de los modelos de scoring/rating se obtiene en función de la información que utilizan:

- Modelos reactivos: Utilizan información del cliente (edad, situación laboral, ingresos, ...) y de la operación que están solicitando (plazo, importe solicitado, ...) para calcular la probabilidad que la operación entre en morosidad.
- Modelos proactivos. Utilizan sobre todo información de operativa interna del cliente (historial de pagos, ingresos, gastos, ...) con la entidad para calcular la probabilidad de entrada en morosidad.

El presente trabajo se centra en el análisis y mejora de un scoring proactivo que permita discriminar los clientes con una menor probabilidad de impago de aquellos cuya probabilidad es mayor. El sistema está pensado para trabajar con no clientes y permitirá calificarlos en función de las posiciones bancarias que presenten en su banco origen.

Capítulo 1

Términos bancarios

En la actualidad, las entidades bancarias cada vez se ayudan más del modelo de scoring para tomar la decisión de aprobar o denegar un crédito. El scoring es un método que pronostica el riesgo futuro por el incumplimiento de pagos en un lapso de tiempo determinado. Muchas entidades consideran el scoring una herramienta determinante para conceder o no crédito, ya que posibilita trabajar con muchas peticiones de créditos de manera rápida, aparte de reducir la tasa de morosidad frente a la toma de decisiones humanas. En este capítulo, se definen algunos términos bancarios y se presentan los pasos para construir un modelo de scoring.

1.1. Definiciones previas

Es posible medir el riesgo de crédito a partir del estudio y de la estimación de un conjunto de parámetros. Este es el enfoque de la medición del riesgo de crédito basado en calificaciones internas, más conocido como enfoque de modelos IRB (Internal rating-based aproach).

Los parámetros fundamentales al riesgo de crédito que pueden calcular internamente las entidades son la probabilidad de mora, la exposición a la mora, la pérdida dada la mora y la madurez del crédito. Las definiciones de estos términos se recogen en [3] y [15].

Probabilidad de mora (PD)

La probabilidad de mora o de incumplimiento (probability of default o PD), también conocida como "frecuencia esperada de mora", es la probabilidad de que un acreditado se declare incapaz de hacer frente a sus compromisos con la entidad al cabo de 1 año desde la formalización de la deuda. Se considera que el cliente es moroso cuando se produce un impago por un período de entre 90 y 180 días.

Exposición en caso de mora (EAD)

La exposición en caso de mora (exposure at default), es la parte de la deuda que queda expuesta al riesgo de pérdida cuando se produce la mora.

Pérdida dada la mora (LGD)

La pérdida dada la mora o incumplimiento (loss given default), es la proporción de la deuda que la entidad espera perder si el acreditado incumple sus obligaciones y se expresa como porcentaje sobre la EAD. Cuando estos tres parámetros se multiplican ($PD \times LGD \times EAD$) dan lugar a la pérdida esperada.

Madurez (M)

La madurez del crédito (maturity) es el período de vencimiento de la operación. Este parámetro es

muy importante para validar los datos utilizados por las entidades, evitando que se produzcan fechas incoherentes.

1.2. Credit scoring

Como se comentó en el prefacio, el scoring abarca una serie de técnicas estadísticas que se utilizan para otorgar o no crédito en las instituciones bancarias. Es un método que pronostica el riesgo futuro por el incumplimiento de pagos en una ventana de tiempo determinada. Este procedimiento lleva utilizándose desde hace 50 años gracias a los avances de los ordenadores que permiten trabajar con grandes volúmenes de datos de manera rápida y eficaz. Estas técnicas tienen como objetivo asociar una puntuación de riesgo a las solicitudes de crédito o a cuentas. Para ello, se podría utilizar una tabla de puntuaciones o scorecard, que es una tabla que contiene las puntuaciones asignadas a cada atributo de cada una de las características usadas para construir el modelo. Esta puntuación puede determinar, por ejemplo, la probabilidad de pago de la deuda para un cliente cuando se le conceda una tarjeta de crédito. Mayores puntuaciones se corresponden con una mayor probabilidad de pago.

Características	Atributos	Score
Edad	Menor de 25 años	-10
	25-45	10
	45-65	20
	Mayor de 65	30
Estado civil	Casado	15
	Soltero	0
	Otro	-30
Antigüedad Empleo	0-1 año	-10
	2-5 años	5
	5-10	10
	Más de 10 años	20
Sexo	Masculino	-10
	Femenino	5

Cuadro 1.1: Tabla de puntuación.

Un ejemplo de una tabla de puntuaciones muy básica puede ser la que aparece en el Cuadro 1.1. En ella, se consideran 4 variables: edad, estado civil, antigüedad en el empleo y sexo. Una mujer de 30 años, casada y con 8 años de antigüedad en su trabajo tiene una puntuación en base a esta tabla de

40=10+15+10+5. Se puede observar que un individuo con mayor antigüedad en el trabajo no tiene necesariamente que tener una mayor probabilidad de obtener el préstamo, ya que se consideran varias variables explicativas.

Previamente a la construcción de un modelo de scoring, se debe definir lo que se considera la bandera de rendimiento, así como un buen y un mal cliente. Estas definiciones se recogen en [19]. Se define la bandera de rendimiento como un indicador de si una cuenta es o no morosa en un intervalo de tiempo determinado. La definición de lo que constituye una cuenta "mala" se basa en varias consideraciones:

- La definición debe estar en línea con los objetivos de la entidad. Si el objetivo es aumentar la rentabilidad, entonces se establece un punto de morosidad donde la cuenta no es rentable. Esto se puede complicar si hay cuentas que, por ejemplo, pagan sus cuotas siempre un mes tarde pero este retraso no llega a ser de dos o tres meses. En este caso, la bandera de rendimiento podría considerarse en 60 ó 90 días.
- La definición debe ir en consonancia con el propósito por el que se construye el modelo de scoring, que puede ser, detectar bancarrota (quiebra, ruina económica), fraude (vulneración de una norma tributaria con la que se pretende eludir mediante engaño el pago de un impuesto), reclamos y cobros.
- Una definición estricta de malo (por ejemplo 120 días de mora) proporciona una información más extrema y precisa, pero se corre el riesgo de que se produzcan tamaños de muestra muy bajos.
- Por el contrario, una definición flexible (por ejemplo 30 días de mora) considera tamaños muestrales mayores pero no diferencia bien entre cuentas buenas y malas.
- La definición debe ser fácilmente interpretable (por ejemplo, siempre 90 días de mora o siempre 60 días de mora). Pero definiciones como, por ejemplo, "tres veces 30 días de mora o dos veces 60 días de mora" pueden darnos información más precisa pero son más difíciles de interpretar y manejar, así como es más difícil tomar decisiones.

Una vez que se definen las cuentas malas, se busca definir lo que se considera una cuenta buena para la entidad. De nuevo, esta definición tiene que ir en línea con los objetivos y propósitos que busque la entidad financiera. Algunas características que puede considerar una cuenta clasificada como buena son que nunca haya mora, no tenga reclamos, no quiebre o no haya fraude.

Las cuentas buenas deben conservar su estado en toda la ventana de rendimiento, que es el intervalo de tiempo en el que se observa el rendimiento de una cuenta. Una cuenta se considera mala si alcanza la etapa de morosidad especificada por el banco en cualquier momento en el transcurso de la ventana.

Las cuentas indeterminadas son aquellas que no entran en las categorías de buena o mala, debido a que no tienen un historial de rendimiento suficiente para la clasificación o que tienen una leve morosidad que no es lo suficientemente baja para clasificarse como cuenta buena, ni lo suficientemente alta para clasificarse como mala. Se pueden considerar como cuentas indeterminadas:

- Las cuentas que llegan a entre 30 y 60 días de mora pero no sobrepasan los 60 días por lo que no llegan a ser consideradas malas.
- Las cuentas inactivas o canceladas voluntariamente.
- Las cuentas con uso insuficiente, por ejemplo, tarjetas de crédito con un saldo máximo de 20 euros.

Como regla general, el número de cuentas indeterminadas no debe exceder el 15 % de la cartera. Agregar cuentas indeterminadas al conjunto de datos de desarrollo del modelo crea un escenario de

clasificación errónea. En los casos en los que la proporción de indeterminados sea muy alta (por ejemplo, si hay una gran proporción de cuentas inactivas), es importante analizar sus causas (como podría ser, la presencia de otras tarjetas con mejores propuestas). Conocer las razones por las que los clientes de tarjetas de crédito no usan el producto hace que se puedan tomar decisiones apropiadas para remediar la situación, aumentando los límites o reduciendo las tasas de interés en algunos casos.

Para el desarrollo de un modelo de scoring solo se deben considerar las cuentas buenas y las cuentas malas. A continuación, se definen los conceptos de tasa de mora y ventana de muestreo que se siguen de la referencia [13].

Tasa de mora o tasa de morosidad

Se define la tasa de mora como el cociente del número de cuentas malas entre el número total de cuentas de la muestra.

Ventana de muestreo

La ventana de muestreo es un período de tiempo en el que se observa el comportamiento de las cuentas a partir de su apertura. Conforme aumenta la edad de las cuentas, la tasa de mora va variando hasta llegar a estabilizarse. En ese momento ya se puede clasificar a un cliente como bueno o malo. Cuando la tasa de morosidad se estabiliza se dice que las cuentas llegan a su madurez de comportamiento.

La muestra de datos tomada para elaborar el scoring debe ser representativa de la población actual. Si la muestra no es muy cercana, las características puede que no representen bien la realidad y, si se toman muestras demasiado cercanas, se reduciría el tamaño de la muestra. Normalmente se observan en un intervalo de tiempo de entre 12 y 18 meses anteriores a la fecha en la que se hace el estudio para asegurar que las cuentas hayan adquirido su madurez. En la subsección 1.2.1 se enuncian las etapas o los pasos que se deben seguir para obtener un modelo de scoring.

1.2.1. Etapas de construcción de un modelo de scoring

1. Conformar la base de datos.

   En esta etapa se realiza un vaciado de la información contenida en las solicitudes de los clientes. Se construye una base que contiene el comportamiento de mora de los clientes registrados en la base de las solicitudes. Además los datos se depuran, excluyendo variables con exceso de campos sin respuesta.

2. Agrupar los datos contenidos en la base.

   Una vez que se conformó la base de datos, se procede a agrupar las características (variables) en categorías o grupos. Estos grupos también son conocidos como atributos.

3. Determinar los clientes buenos y malos.

   En este punto se forma una base de datos con los clientes que se clasifican como clientes buenos y como clientes malos.

4. Determinar las variables que entran en el modelo.

   Se hace una selección de las características que tienen una mayor capacidad predictiva. Para ello se utiliza una medida llamada valor de la información, que se definirá en el capítulo 4. Con las variables que tienen un buen poder predictivo se obtiene un modelo de predicción utilizando la regresión logística. Las puntuaciones asignadas a los atributos de cada característica se calculan en función de las estimaciones de los parámetros obtenidos con la regresión logística y el peso de la evidencia de los atributos. El peso de la evidencia es otra medida de predicción que se definirá en el capítulo 4.

5. Medir la eficiencia del modelo.

   Se utilizan técnicas estadísticas como las curvas ROC o el índice de Gini para determinar cómo de bien clasifica el modelo propuesto.

6. Establecer el punto de corte.

   Se determina un punto de corte que separa las solicitudes nuevas en aceptadas o rechazadas.

Capítulo 2

Regresión logística

La regresión logística es un modelo de regresión que se utiliza muy habitualmente en el credit scoring. Nos permite discriminar a los solicitantes de crédito en buenos o malos. El modelo de regresión logística no requiere de los supuestos de la regresión lineal, como pueden ser el supuesto de normalidad de los errores de observación. Este modelo de regresión se utiliza cuando la variable respuesta, Y, es binaria y las variables explicativas son continuas. El modelo puede extenderse a situaciones en las que las variables explicativas son discretas o cualitativas o la variable respuesta es cualitativa con más de dos modalidades. La regresión lineal no es aplicable a este tipo de variables en los que la variable respuesta Y solo puede tomar dos posibles valores. El modelo debe ser capaz de clasificar a los individuos en los grupos, buenos o malos, basado en las variables que definen las características de los individuos. En este capítulo se presenta el modelo de regresión logística y se estiman los parámetros de este modelo como se puede ver en [7], [13] y [14].

2.1. Modelo de regresión logística

En este modelo, la variable respuesta, Y, únicamente puede tomar dos valores: 0 y 1. Se considera $Y = 1$ si se trata de un buen cliente e $Y = 0$ si se trata de un mal cliente. Las variables explicativas de este modelo pueden ser cualitativas o cuantitativas, tanto continuas como discretas o categóricas y pueden tomar dos o más valores.

Con la regresión logística se modela la probabilidad de éxito, es decir, de que $Y = 1$. Si la probabilidad estimada de que $Y = 1$ es mayor que 0.5, se clasifica al cliente como bueno y si esta probabilidad es menor que 0.5 como malo.

Sea n el tamaño de la cartera de créditos y sea $X_i = (X_{i1}, \ldots, X_{ip})$ un vector aleatorio que caracteriza el perfil crediticio del i-ésimo solicitante, con $1 \leqslant i \leqslant n$ e Y_i la variable respuesta asociada. Sea $p_i = P(Y = 1|X_i)$, se define $g(p_i) = \beta_0 + \beta' X_i$ siendo $\beta' = (\beta_1, \ldots, \beta_p)$ un vector de parámetros de coeficientes de las variables explicativas del modelo. A esta función g se le llama función link o función de enlace. En situaciones donde la variable respuesta es dicotómica es habitual considerar como función link la función logística o función logit:

$$g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

La función logística consiste en efectuar un logaritmo al cociente entre la probabilidad de éxito y la probabilidad de fracaso. A este cociente se le conoce como odds, disparidad o desventaja.

$$Odds = \frac{P(Y=1)}{P(Y=0)} = \frac{p_i}{1-p_i}.$$

Los coeficientes de regresión pueden tomar signo positivo o negativo pero la odds solamente toma valores positivos. A partir de la probabilidad de éxito se puede averiguar el valor de la odds, pero también se puede averiguar el valor de la probabilidad de éxito a partir de la odds. Son dos formas de medir el parámetro desconocido de una variable dicotómica. Una de las diferencias entre estas medidas es que la probabilidad de éxito puede tomar valores en el intervalo $[0,1]$ y la odds en el intervalo $[0,\infty)$, es decir, puede tomar cualquier valor real positivo. No puede tomar valores negativos puesto que es el cociente entre dos valores positivos. Si a la odds se le aplica un logaritmo se tiene una cantidad que se mueve en el intervalo $(-\infty,\infty)$.

Utilizando la función inversa de g, dada por $g^{-1}(t) = \frac{\exp(t)}{1+\exp(t)}$ se tiene:

$$p_i = \frac{e^{\beta_0+\beta' X_i}}{1+e^{\beta_0+\beta' X_i}}.$$

Si $\beta_0 + \beta' X_i = 0$, entonces la probabilidad de éxito que predice el modelo es de 0.5. Si $\beta_0 + \beta' X_i$ es positivo y grande, la probabilidad de éxito aumenta, pero lo hace cada vez menos teniendo como asíntota el 1. Si $\beta_0 + \beta' X_i$ es negativo, la probabilidad es menor de 0.5 y cuanto más pequeño sea menor es la probabilidad, llegando a tener una asíntota en cero. El modelo logístico es adecuado para representar el comportamiento de la función de regresión cuando la variable respuesta es dicotómica. Se tiene también que $1 - p_i = \dfrac{1}{1+e^{\beta_0+\beta' X_i}}$.

La función link también se puede expresar como:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

y $\beta_j, 1 \leq j \leq p$, sirve para analizar la cantidad de cambio del ratio de probabilidades cuando se incrementa en una unidad la variable X_j. Esta cantidad es denominada odds ratio (OR) y tiene la siguiente expresión:

$$OR = e^{\beta_j}.$$

Para el modelo de regresión $g(X_i) = \beta_0 + \beta' X_i$ se pueden utilizar las mismas técnicas de selección de variables que en el modelo de regresión lineal como pueden ser los métodos backward y forward que se explicarán con más detalle en el capítulo 5.

En R [16], la función básica para la regresión logística es la función *glm*. Ésta tiene una sintaxis similar a la función *lm* para modelos lineales. La variable respuesta está separada de las variables explicativas por el símbolo $\sim$. Para indicar que se quiere hacer una regresión logística es necesario especificar *family=binomial* y como función *link* la función logit. A continuación, se muestra un ejemplo de código para el caso concreto de 3 variables explicativas.

```
glm(Y~X_1+X_2+X_3,family=binomial(link=logit))
```

2.2. Estimación de los parámetros del modelo

La estimación de los parámetros del modelo se realiza mediante el método de máxima verosimilitud. Como cada Y_i es una Bernoulli con parámetro p_i, la función de verosimilitud adopta la forma:

$$P(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i}$$

y su logaritmo será:

$$\log P(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} [Y_i \log p_i + (1 - Y_i) \log(1 - p_i)]$$

Si se considera $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ y $X_i' = (1, X_{i1}, \ldots, X_{ip})$ se puede escribir:

$$\log \left(\frac{p_i}{1 - p_i} \right) = X_i' \beta.$$

Sustituyendo esta expresión en la función log verosimilitud se obtiene la función de verosimilitud de logaritmos en términos de β como:

$$L(\beta) = \sum_{i=1}^{n} Y_i X_i' \beta - \sum_{i=1}^{n} \log(1 + e^{X_i' \beta}).$$

Si se deriva la expresión $L(\beta)$ con respecto a los parámetros β_j y se iguala a cero se obtienen las ecuaciones:

$$\frac{\delta L(\beta)}{\delta \beta} = \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \left(\frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}} \right)$$

$$\sum_{i=1}^{n} Y_i X_i = \sum_{i=1}^{n} X_i \left(\frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}} \right) = \sum_{i=1}^{n} X_i p_i.$$

Existen varios métodos iterativos para resolver las ecuaciones de verosimilitud del modelo de regresión logística. Como las ecuaciones no son lineales en los parámetros β, es habitual usar el método de Newton-Raphson.

Para medir la capacidad discriminatoria de un modelo de regresión logística se puede utilizar la deviance. Se define deviance como $D(\beta) = -2L(\beta)$, es decir,

$$D(\beta) = 2 \sum_{i=1}^{n} [\log(1 + e^{X_i' \beta}) - Y_i X_i' \beta]$$

o lo que es lo mismo,

$$D(\beta) = -2 \sum_{i=1}^{n} [Y_i \log p_i + (1 - Y_i) \log(1 - p_i)].$$

Para estudiar si se puede prescindir de algún parámetro de la regresión logística, es decir, contrastar la hipótesis nula $H_0 : \beta_i = 0$ y la hipótesis alternativa $H_1 : \beta_i \neq 0$, se puede considerar el estadístico de Wald. Este estadístico sigue una distribución normal bajo H_0 y tiene la siguiente expresión:

$$\frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)} \sim N(0, 1)$$

Capítulo 3

Técnicas de agrupamiento de variables

Un buen modelo debe satisfacer dos condiciones, la primera es que tenga una fuerte capacidad predictora y la segunda es que la estimación de sus parámetros tenga una alta precisión. Además, es preferible que el modelo sea lo más sencillo posible, ya que así es más fácil de interpretar. Esto quiere decir que es preferible aquel modelo que contenga el mínimo número de variables explicativas satisfaciendo las dos condiciones. Por lo tanto, es necesario hacer un análisis de las variables explicativas necesarias para construir el modelo, ya que puede ser que alguna se excluya por su baja capacidad predictiva.

El primer paso para realizar este análisis es llevar a cabo una agrupación inicial de las variables y ordenarlas por alguna medida de poder de predicción. Entre estas medidas, las más utilizadas en el contexto del credit scoring son el valor de información (IV, del inglés information value) y el peso de la evidencia (WOE, del inglés weight of evidence), medidas que serán explicadas en el capítulo 4.

Existen varias técnicas de agrupamiento de las características. Una manera de realizar estas agrupaciones es mediante lo que se denomina un "binning", para luego calcular las medidas de predicción para las características agrupadas. Los árboles de decisión también se usan a menudo para agrupar variables, aunque los usuarios los suelen utilizar para generar ideas iniciales.

En R se pueden hacer estas agrupaciones mediante el paquete llamado *smbinning*. Éste realiza las agrupaciones óptimas y calcula los valores de WOE e IV. Los detalles de este paquete pueden consultarse en [10] y se verán con más profundidad en el capítulo 4.

La agrupación óptima categoriza una característica numérica en atributos para su posterior uso en el modelo de puntuación. Este proceso, también conocido como discretización supervisada, utiliza el particionado recursivo para categorizar la característica numérica. El algoritmo específico que utiliza este paquete es el de árboles de inferencia condicional. Inicialmente excluye los valores perdidos (NA) para calcular los puntos de corte y los agrega más tarde en el proceso, para el cálculo del valor de información.

En este capítulo se definen los conceptos de árboles de decisión, particionado recursivo y árboles de inferencia condicional. Para más información se pueden consultar las referencias [12], [21] y [22] acerca de los árboles de decisión y, para los árboles de inferencia condicional, las referencias [6] y [8]. El objetivo de la clasificación es encontrar un modelo para predecir la clase a la que pertenecería cada conjunto de datos. Esta asignación debe realizarse con la mayor precisión posible. Existen dos tipos de clasificación, la supervisada y la no supervisada. La clasificación supervisada (que es la que utiliza el binning óptimo) es aquella en la que variable respuesta es conocida, mientras que en la clasificación no supervisada no se tienen las posibles respuestas.

Los métodos de clasificación supervisada permiten especificar las categorías para el algoritmo de clasificación, pero tienen el inconveniente de que son necesarias grandes cantidades de datos de en-

trenamiento. Los métodos de clasificación no supervisada generan ellos mismos las categorías, pero tienen el inconveniente de que el usuario tiene que especificar el número de categorías en las que se van a agrupar las variables y para ello es necesario establecer un criterio para determinar ese número. Además, interpretar los resultados de la clasificación no supervisada suele ser complejo. Un ejemplo típico de método no supervisado son los algoritmos de clustering.

El particionamiento de los datos permite estimar la precisión de las predicciones de un modelo antes de aplicarlo a un conjunto de datos no observado aún, por lo que tiene bastante importancia. Este método ayuda a validar el modelo ajustado. Se toman datos que ya están clasificados y se dividen en dos grupos, datos de entrenamiento y datos de prueba. Para que se pueda saber como va a actuar un modelo en la práctica con datos nuevos, no se recomienda utilizar el total de datos como datos de entrenamiento. Se pueden tomar, por ejemplo, el 80 % de los datos para entrenamiento y el 20 % para prueba.

Con el grupo de entrenamiento se estiman los parámetros del modelo y se utiliza este modelo ajustado para predecir la clasificación de los datos de prueba. Como los datos de prueba ya estaban clasificados, se puede estudiar si el modelo entrenado se equivoca en un porcentaje elevado o no, es decir, determinar la precisión del modelo. Si el modelo tiene pocos errores de predicción en los datos de entrenamiento pero no predice bien los de prueba existe sobrepredicción. Es decir, el clasificador no aprendió lo suficiente de los datos y no es capaz de generalizar a los de prueba.

3.1. Árboles de decisión

El árbol de decisión es una técnica de clasificación en la que se realizan particiones binarias de los datos de forma recursiva. Es un método usado en múltiples campos como modelo de predicción. Son similares a los diagramas de flujo, en los que se toman decisiones al llegar a un punto de acuerdo con una regla o criterio. El resultado se puede representar con un árbol.

Hay distintas maneras de obtener árboles de decisión. Una de las más conocidas son los CART: classification and regression trees. Esta es una técnica no paramétrica de predicción con la que se pueden obtener árboles de clasificación y de regresión y es uno de los métodos más eficaces de clasificación supervisada. Su carácter no paramétrico hace que no requiera ninguna hipótesis relativa a la distribución de las variables dependientes e independientes, ni a la relación entre ellas ni a sus posibles interacciones. Se tiene una variable dependiente y lo que se pretende es obtener una función que permita predecir, a partir de las variables independientes, el valor de la variables dependiente para casos desconocidos.

Se obtendrán árboles de clasificación cuando la variable dependiente sea discreta o categórica como, por ejemplo, rechazo o no de un tratamiento, ser o no moroso. Se obtendrán árboles de regresión cuando esta es continua. En ambos casos, el objetivo es identificar las variables que mejor identifican la variable dependiente.

Una de las implementaciones de CART es conocida como recursive partitioning and regression trees o simplemente RPART. Este algoritmo encuentra la variable independiente que mejor separa los datos en grupos, que se corresponden con las categorías de la variable dependiente. Esta mejor separación es expresada con una regla y a cada regla le corresponde un nodo. Para obtener estas reglas se suele utilizar el índice de Gini, que es una medida de la varianza total en los grupos.

El índice de Gini de una muestra se calcula como:

$$I_G(P) = \sum_{i=1}^{m} P_i(1 - P_i) = 1 - \sum_{i=1}^{m} P_i^2$$

donde P_i representa la probabilidad de que un elemento de la muestra escogido al azar pertenezca

a la clase i en el nodo correspondiente. Es la probabilidad de estar en la clase i estando en el nodo t. El criterio a seguir es el de encontrar el corte que supone un menor índice de Gini tras realizar la división. Este índice toma valores próximos a cero cuando hay mucha igualdad dentro de los grupos y puede llegar al valor 1 a medida que va creciendo la desigualdad.

Por ejemplo, si se tiene una variable dependiente que puede tomar dos posibles valores: moroso y no moroso, y se tienen como variables explicativas ingreso mensual, saldo máximo a fin de mes y número de préstamos, el algoritmo encontraría la variable que mejor separe los datos. Suponiendo que esta variable sea el ingreso mensual, se obtendría una regla para separar los datos según si este ingreso alcanza cierto umbral o no. Entonces, los datos que la regla clasifique como no morosos tendrían más probabilidades de pertenecer al grupo de clientes no morosos que al de morosos.

Una vez que los datos sean particionados en grupos a partir de una regla se repite el proceso para cada uno de los grupos resultantes. Es decir, se vuelve a buscar la variable que mejor separe los datos en grupos, se obtiene una regla y se separan los datos. Se realiza el proceso de forma recursiva hasta que sea imposible obtener una separación mejor, probablemente por no haber un número mínimo de observaciones. En este momento, el algoritmo se para. Los grupos que ya no se pueden seguir particionando reciben el nombre de nodo terminal u hoja.

Una observación digna de mención sobre este método es que una vez que una variable ha sido elegida para separar los datos, esta ya no vuelve a ser utilizada en los nuevos grupos creados. Por lo tanto, se buscan variables distintas que mejoren la separación de los datos. Otro factor a tener en cuenta es que en los dos grupos formados a partir de una partición puede ocurrir que para el primer grupo la variable que mejor separe a los datos sea diferente que la del segundo grupo. Es decir, una vez hecha la partición, los grupos son independientes entre sí, por lo que pueden tener diferentes reglas.

El árbol que se ha construido generalmente está sobreajustado, es decir, contiene gran cantidad de niveles. Si el árbol construido es muy grande se suele "podar" para disminuir su complejidad, ya que no necesariamente una mayor complejidad significa una mejor clasificación. Al aumentar la complejidad de un modelo siempre se obtiene un mejor ajuste pero puede que ese ajuste se deba a que se recogen peculiaridades de los datos que no son útiles para describir el comportamiento de nuevos datos (datos de prueba). La poda consiste en eliminar las últimas particiones que no representen un aumento considerable de la capacidad de predicción total.

Conforme se va podando el árbol se va reduciendo la complejidad, pero el número de individuos mal clasificados puede aumentar. El mejor árbol será aquel que tenga la proporción óptima entre la tasa de mala clasificación (cociente entre las observaciones mal clasificadas y el número total de observaciones) y la complejidad del árbol.

La selección del mejor árbol se realiza por validación. Esto consiste en separar una parte de la muestra que no se utiliza en la construcción del árbol, sino únicamente para la predicción. Es decir, los datos de entrenamiento y de prueba que se mencionaron con anterioridad. Pero también se puede realizar la validación mediante el método de validación cruzada, que consiste en dividir la muestra en varios grupos excluyentes de aproximadamente igual tamaño, y hacer el árbol dejando cada vez un grupo fuera, que es utilizado para la predicción y para medir el porcentaje de aciertos de clasificación. Este proceso se repite hasta usar todos los grupos y se selecciona finalmente el árbol con la menor tasa agregada de mala clasificación.

El resultado se presenta como un árbol boca abajo, donde la raíz representa a toda la población y está en la parte superior y las hojas o nodos terminales en la inferior. A partir de la raíz, el árbol se va dividiendo en ramas y estas a su vez en más ramas hasta llegar a las hojas. Por este motivo, una de las ventajas de este método es su fácil interpretabilidad, pues nos da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Otra de las ventajas es que suele tener una buena capacidad predictiva para muchos tipos de datos a pesar de no requerir gran poder de cómputo. Entre sus desventajas está que en este tipo de clasificación puede tener resultados que varían mucho dependiendo de las muestra de datos utilizada en los datos de entrenamiento. Además es fácil que

haya sobreajustes en los modelos, lo que significa que clasifiquen con un muy buen poder predictivo los datos que conocemos pero que sean deficientes para datos desconocidos. Para esta sección se han seguido las referencias [12] y [21].

3.1.1. Paquete rpart

El paquete *rpart* de R [21] usa este algoritmo de particiones recursivas basado en el modelo CART considerando dos etapas en su procedimiento. Primero se encuentra la mejor variable que particiona los datos en dos grupos. Este procedimiento es aplicado a cada uno de los subgrupos y así recursivamente hasta que los subgrupos tengan el mínimo tamaño de muestra predefinido (mínimo de 5 observaciones en cada subgrupo) o hasta que no se puedan hacer mejoras. Luego, el modelo construido en la primera etapa se poda mediante validación cruzada, eligiendo el modelo cuyo subárbol tenga la estimación del riesgo más alta.

Existen otros paquetes en R que permiten realizar estos modelos CART. Uno de ellos es el paquete *tree* [17], pero *rpart* ofrece una mejor solución a los valores perdidos del conjunto de datos, ya que crea la categoría NA para los predictores discretos y continuos.

Ejemplo

El siguiente ejemplo muestra cómo se construye el árbol de decisión con R utilizando el paquete *tree* y el paquete *rpart* para la base de datos de iris de R. Esta base de datos fue recolectada durante varios años por Edgar Anderson. Quería demostrar que la longitud y el ancho de los sépalos y los pétalos podrían utilizarse para diferenciar entre especies de plantas iris. Contiene 3 especies de tipos de planta iris (Iris Setosa, Iris Versicolour e Iris Virginica) con 50 casos cada una.

Las variables que contiene esta base de datos son:

- Longitud del sépalo en cm (Sepal.Length)
- Ancho del sépalo en cm (Sepal.Width)
- Longitud del pétalo en cm (Petal.Length)
- Ancho del pétalo en cm (Pepal.Width)

El código R para construir el árbol de decisión con el paquete *tree* es el siguiente:

```
> library(tree)
> iristree <- tree(Species ~., iris)
> iristree
node), split, n, deviance, yval, (yprob)
* denotes terminal node

1) root 150 329.600 setosa ( 0.33333 0.33333 0.33333 )
2) Petal.Length < 2.45 50   0.000 setosa ( 1.00000 0.00000 0.00000 ) *
3) Petal.Length > 2.45 100 138.600 versicolor ( 0.00000 0.50000 0.50000 )
6) Petal.Width < 1.75 54  33.320 versicolor ( 0.00000 0.90741 0.09259 )
12) Petal.Length < 4.95 48   9.721 versicolor ( 0.00000 0.97917 0.02083 )
24) Sepal.Length < 5.15 5   5.004 versicolor ( 0.00000 0.80000 0.20000 ) *
25) Sepal.Length > 5.15 43   0.000 versicolor ( 0.00000 1.00000 0.00000 ) *
13) Petal.Length > 4.95 6   7.638 virginica ( 0.00000 0.33333 0.66667 ) *
7) Petal.Width > 1.75 46   9.635 virginica ( 0.00000 0.02174 0.97826 )
```

```
14) Petal.Length < 4.95 6   5.407 virginica ( 0.00000 0.16667 0.83333 ) *
15) Petal.Length > 4.95 40   0.000 virginica ( 0.00000 0.00000 1.00000 ) *
> summary(iristree)

Classification tree:
tree(formula = Species ~ ., data = iris)
Variables actually used in tree construction:
[1] "Petal.Length" "Petal.Width"  "Sepal.Length"
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
```

Se puede representar el árbol de decisión para facilitar la visualización.

```
> plot(iristree)
> text(iristree)
```

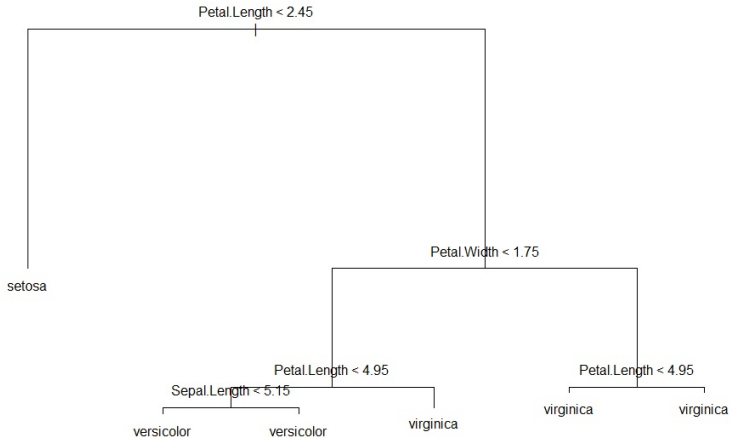

Figura 3.1: Ejemplo de árbol de decisión

```
misclass.tree(iristree, detail=TRUE)
predict(iristree)
predict(iristree,type="class")
table(predict(iristree))
arb<- snip.tree(iristree, nodes = c(12, 7))
plot(iris[, 3],iris[, 4], type="n",
xlab="petal length", ylab="petal width")
text(iris[, 3], iris[, 4], c("s", "c", "v")[iris[, 5]])
partition.tree(arb, add = TRUE, cex = 1.5)
```

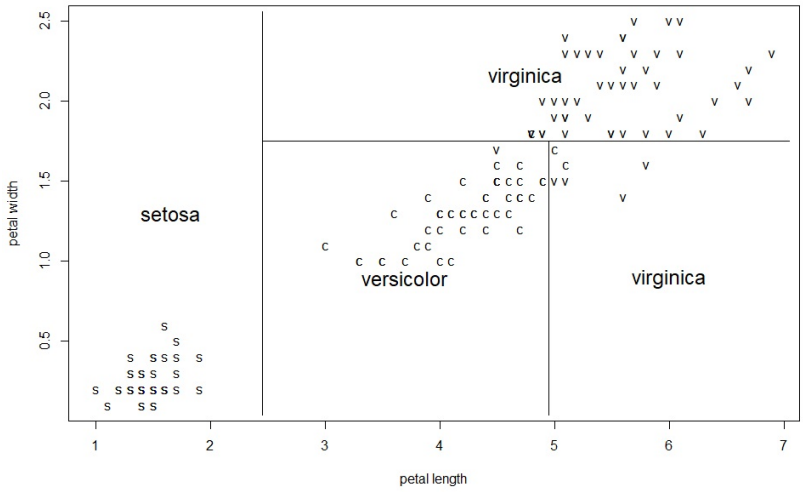

Figura 3.2: Representación de árbol de decisión

En las figuras 3.1 y 3.2 se observa que hay una diferencia bastante clara en las longitudes de los pétalos entre las diferentes especies, siendo las longitudes de los pétalos de la especie setosa menores a 2.45 cm. Si se considera el ancho de los pétalos también se observa una diferencia clara entre las especies virgínica y versicolor, siendo el ancho de la especie versicolor menor a 1.75 cm.

Ahora se verá el código R para este mismo ejemplo con el paquete *rpart*, que, como se comentaba antes, ofrece una mejor solución a los valores perdidos. En este caso se va a considerar el 80 % de la muestra como datos de entrenamiento y el 20 % restante como datos de prueba.

```
> library(rpart)
> muestra<-sample(2, nrow(iris), replace=TRUE, prob=c(0.8, 0.2))
> entrena <- iris[muestra==1,]
> prueba <- iris[muestra==2,]
> arb <- rpart(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
data=entrena)
> print(arb)
n= 122

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 122 79 versicolor (0.33606557 0.35245902 0.31147541)
2) Petal.Length< 2.45 41  0 setosa (1.00000000 0.00000000 0.00000000) *
3) Petal.Length>=2.45 81 38 versicolor (0.00000000 0.53086420 0.46913580)
6) Petal.Width< 1.75 46  4 versicolor (0.00000000 0.91304348 0.08695652) *
7) Petal.Width>=1.75 35  1 virginica (0.00000000 0.02857143 0.97142857) *
```

El paquete *rpart.plot* proporciona una mejor visualización del árbol, como se puede ver en la Figura 3.3.

```
> library(rpart.plot)
```

```
> rpart.plot(arb)
```

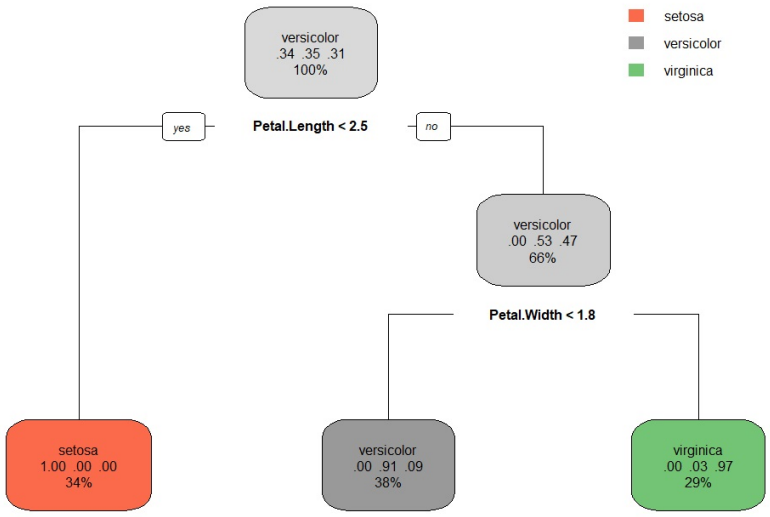

Figura 3.3: Árbol de decisión creado con el paquete rpart.

En la Figura 3.3 cada uno de los rectángulos representa un nodo del árbol, con su regla de decisión. Cada nodo está coloreado de acuerdo a la categoría mayoritaria entre los datos que agrupa. Esta es la categoría que ha predicho el modelo para ese grupo. Dentro del rectángulo de cada nodo se muestra la proporción de casos que pertenecen a cada categoría y la proporción de datos que han sido agrupados allí. Por ejemplo, el rectángulo verde de la gráfica tiene 0 % de los casos de la especie setosa, el 3 % de la especie versicolor y el 97 % de la especie virgínica, y estos representan al 29 % de todos los datos. En este modelo se clasificaron correctamente al 100 % de la especie setosa, al 91 % de la especie versicolor y al 97 % de la especie virgínica. Se puede usar la función *predict()* con los datos de prueba para generar un vector con los valores predichos por el modelo con los datos de entrenamiento.

3.2. Árboles de inferencia condicional

Los árboles de inferencia condicional constituyen un método de clasificación que trata de solucionar los problemas asociados al método CART: árboles de clasificación y árboles de regresión descritos en la sección anterior. En especial, tratan de resolver el problema del sobreajuste y el sesgo de selección de variables hacia aquellas que presentan más opciones de corte.

Los árboles de inferencia condicional introducen el uso de un estadístico que permite decidir si existe relación entre cada una de las variables explicativas y la respuesta, de manera que, llegado al punto de que dicha relación se encuentre por debajo de un valor, se detiene la construcción del árbol, evitando así el sobreajuste.

El método utilizado en R consiste en comprobar primero si la variable dependiente está relacionada con las explicativas, para lo que se aplica un test de independencia. Si detecta dependencia, selecciona la variable explicativa con mayor asociación a la respuesta, medida con el p-valor de un test para cada variable explicativa, y a continuación realiza una clasificación binaria con esa variable, repitiendo de forma recursiva los pasos anteriores para formar clases en las que los valores de la variable dependiente sean distintos.

Los árboles de inferencia condicional son árboles individuales al igual que los CART, pero se diferencian de estos en que no seleccionan para la partición la variable que maximiza el valor de un indicador (como por ejemplo el coeficiente de Gini), sino que llevan a cabo varias pruebas de contraste al poner en marcha el algoritmo para comprobar la importancia de cada una de las variables explicativas. Con los árboles de inferencia condicional se consigue evitar el sobreajuste y la tendencia a seleccionar las variables que presentan un mayor número de divisiones posibles que se da en los CART cuando las variables toman rangos de valores muy diferentes o tienen distinto número de niveles.

3.2.1. Función ctree del paquete party

El paquete *party* [8] utiliza el algoritmo de partición recursiva con la implementación del procedimiento de los árboles de inferencia. El paquete *rpart* que se describió en la sección anterior, emplea el algoritmo CART y realiza una búsqueda exhaustiva sobre todas las divisiones posibles maximizando una medida de información (basada en IV o Gini) de la impureza del nodo seleccionando con la covariable que muestra la mejor división. Este método tiene dos grandes problemas que son el sobreajuste y el sesgo de selección de las variables con muchas divisiones posibles. El paquete *party* soluciona estos problemas a través de la implementación de la prueba de permutación desarrollada por Strasser y Weber [20].

Ejemplo

Siguiendo con el ejemplo de datos de iris de R, se puede contruir el árbol de inferencia condicional con la función *ctree* del paquete *party* como se muestra en el siguiente código:

```
> library(party)
> irisct <- ctree(Species ~ .,data = iris)
> irisct

 Conditional inference tree with 4 terminal nodes
Response:  Species
Inputs:  Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Number of observations:  150

1) Petal.Length <= 1.9; criterion = 1, statistic = 140.264
  2)*  weights = 50
1) Petal.Length > 1.9
  3) Petal.Width <= 1.7; criterion = 1, statistic = 67.894
    4) Petal.Length <= 4.8; criterion = 0.999, statistic = 13.865
      5)*  weights = 46
    4) Petal.Length > 4.8
      6)*  weights = 8
  3) Petal.Width > 1.7
    7)*  weights = 46
```

Se puede representar el árbol de inferencia condicional para facilitar la visualización.

```
> plot(irisct)
```

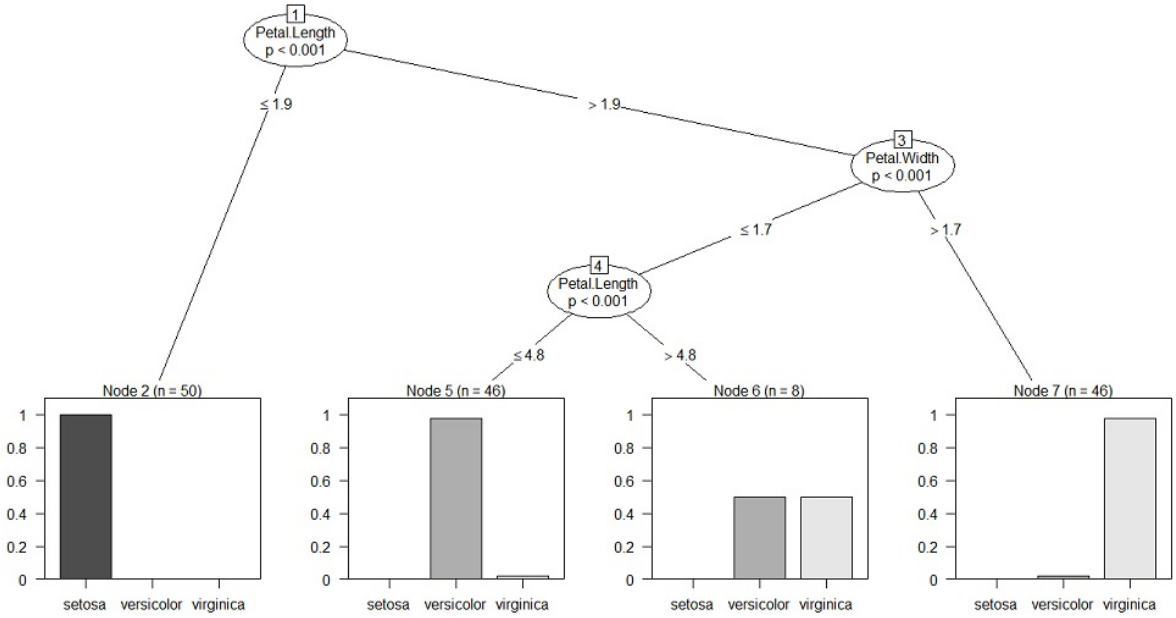

Figura 3.4: Ejemplo de árbol de inferencia condicional

En la Figura 3.4 se observa que la primera variable que mejor separa a los datos en grupos es la longitud de los pétalos y la regla a tener en cuenta es si la longitud es menor o mayor que 1.9 cm. Se aplica un test de independencia y para p-valores menores que 0.001 esta regla discrimina muy bien a la especie setosa, ya que los 50 datos de la muestra de la especia setosa tienen longitudes de pétalo menores a 1.9 cm. La segunda variable que mejor separa a los datos es el ancho del pétalo. La regla a considerar es que el ancho sea menor o mayor a 1.7 cm. De los 50 datos de la muestra de la especie virgínica, 46 de ellos tienen un ancho de pétalo superior a 1.7 cm.

Por lo tanto, el árbol de inferencia condicional permite discriminar las especies setosa, virgínica y versicolor en función de sus longitudes de pétalos y sépalos. La especie setosa tiene las longitudes de los pétalos menores a 1.9 cm, y las especies virgínica y versicolor tienen valores mayores. Para discriminar entre estas dos especies se utiliza la variable del ancho del pétalo, ya que la especie virgínica presenta valores superiores a 1.7 cm y la especie versicolor valores menores en general.

Capítulo 4

Medidas de predicción

Una vez que se tienen las características agrupadas en categorías, el siguiente paso es analizar si estas características así agrupadas tienen un buen poder predictivo. En este capítulo se estudiarán algunas medidas comunes de poder de predicción que se pueden ver en [13] y [19].

La capacidad predictiva de una característica es calibrada mediante cuatro criterios principales:

- El peso de la evidencia (WOE) es una medida del poder predictivo para cada atributo (cada categoría de la característica).
- El rango y la tendencia del peso de la evidencia en atributos agrupados dentro de una característica.
- El valor de información (IV) es una medida del poder predictivo de cada característica.
- Consideraciones operacionales y de negocio.

El valor de la información (IV, por su sigla en inglés de Information Value) es un concepto muy útil para la selección de variables durante la construcción del modelo. Las raíces del valor de la información están en la teoría de la información propuesta por Claude Shannon. Algunos analistas utilizan otros algoritmos para la selección de variables como pueden ser el valor chi cuadrado o el R-cuadrado antes de agrupar características. Esto les da una indicación del poder de las características, y también los alerta en casos donde el valor de información es alto o bajo en comparación con otras medidas. El valor chi cuadrado es una medida muy usada en estadística y es un buen sustituto del valor de la información. Sin embargo, el IV es una medida que está muy extendida y es muy popular en el scoring.

4.1. Valor de información y peso de la evidencia

El estadístico del valor de información es el más utilizado a la hora de seleccionar variables para construir un modelo de scoring. En esta sección se definen los conceptos de peso de la evidencia y valor de la información.

Los modelos de clasificación binaria son quizás el caso de uso más común en el análisis predictivo. La razón es que muchas acciones de los clientes son de naturaleza binaria, como puede ser incumplir un préstamo.

Antes de construir un modelo de clasificación binaria, un paso común es realizar una selección de las variables y un análisis exploratorio de datos. En este paso se conocen los datos y se descartan las variables que no contienen información que nos ayude a predecir la acción de interés. El objetivo

de este paso no debe confundirse con el de las técnicas de selección de variables múltiples, como las técnicas stepwise que se verán más adelante, dónde se seleccionan las variables que entran en el modelo final. Al contrario, este es un paso previo diseñado para garantizar que el modelo tenga un buen poder predictivo.

El peso de la evidencia (WOE) y el valor de la información (IV) se han utilizado en el mundo del riesgo de crédito durante varias décadas, y la teoría subyacente se remonta a la década de 1950. Sin embargo, todavía no se usan demasiado en otros campos.

Los análisis WOE e IV nos permiten:

- Considerar la contribución independiente de cada variable al modelo final.
- Detectar relaciones lineales y no lineales.
- Clasificar las variables en términos de poder predictivo "univariante".
- Visualizar las correlaciones entre las variables predictivas y la variable binaria.
- Comparar el poder de predicción de las variables continuas y categóricas sin crear variables ficticias.
- Manejar los valores perdidos.
- Evaluar el poder predictivo de los valores perdidos.

El peso de la evidencia (WOE, del inglés Weight of Evidence) es una medida del poder de predicción de buenos y malos de cada atributo, o de cada grupo de atributos. Es decir, mide la diferencia entre la proporción de clientes morosos (llamados también clientes malos) y no morosos (clientes buenos) en cada atributo (mide la odds de una persona de ser bueno o malo con ese atributo). El peso de la evidencia indica el poder predictivo de una variable independiente en relación con la variable dependiente y se basa en el logaritmo del cálculo de la odds:

$$\frac{Db_{ij}}{Dm_{ij}}, \text{ siendo}$$

$$Db_{ij} = \frac{\text{número de buenos en el atributo } j \text{ de la característica } i}{\text{número de buenos en la característica } i}$$

la distribución de buenos en la característica i, y siendo

$$Dm_{ij} = \frac{\text{número de malos en el atributo } j \text{ de la característica } i}{\text{número de malos en la característica } i}$$

la distribución de malos en la característica i. Estos términos son estimaciones empíricas de dos probabilidades condicionadas. Db_{ij} es la estimación empírica de estar en el atributo j de la característica i condicionado a no ser moroso y Dm_{ij} es la estimación empírica de estar en el atributo j de la característica i condicionado a ser moroso.

Una forma más habitual de calcular el WOE es la siguiente:

$$WOE = [\log\left(\frac{Db_{ij}}{Dm_{ij}}\right)] \times 100. \tag{4.1}$$

Si este valor es un número negativo indica que hay una mayor proporción de malos que de buenos en el atributo. Se denotan b_i y m_i el número de cuentas buenas y malas para la característica i. Para

el atributo j de la característica i se obtiene el número de cuentas buenas y malas y se denotan b_{ij} y m_{ij}. Se tiene que $b_i = b_{i1} + b_{i2} + \ldots + b_{in_i}$ y que $m_i = m_{i1} + m_{i2} + \ldots + m_{in_i}$, donde n_i es el número de atributos para la característica i.

El valor de la información, viene de la teoría de la información propuesta por Shannon, y es una medida que se calcula a partir de la siguiente fórmula:

$$IV = \sum_{i=1}^{n} \left(\frac{b_{ij}}{b_i} - \frac{m_{ij}}{m_i} \right) \times log \left(\frac{b_{ij} m_i}{m_{ij} b_i} \right). \tag{4.2}$$

El WOE y el IV juegan dos papeles distintos. Por un lado, el WOE describe la relación entre una variable predictiva y una variable objetivo binaria, mientras que el IV mide la fuerza de esa relación. Se deben de tener en cuenta algunas consideraciones para el cálculo de estas dos medidas.

- Los datos "perdidos" se agrupan por separado.
- Para que el análisis sea significativo, se aplica la regla de que al menos tiene que haber un mínimo del 5 % de los datos en cada grupo. La cantidad de grupos determina la cantidad de suavizado: cuantos menos grupos, más suavizado. Los grupos con menos del 5 % de casos pueden conducir a la inestabilidad del modelo.
- No puede haber grupos donde el número de buenos o el número de malos sea cero.
- Tanto la tasa de malos como el valor del WOE deben ser lo suficientemente diferentes de un grupo a otro, esto es, la agrupación se debe realizar de forma que se maximice la diferencia entre buenos y malos. Cuanto mayor sea la diferencia entre el WOE de los grupos mayor será la capacidad predictiva de esta característica.
- El WOE sigue una tendencia lógica creciente, no teniendo en cuenta los valores faltantes como se puede ver en la Figura 4.2.

Se siguen los siguientes pasos para calcular los valores de WOE e IV:

1. Para una variable continua, se realiza un binning para agrupar las características, teniendo en cuenta las consideraciones descritas anteriormente. Si la variable es cualitativa no es necesario este paso.
2. Se calcula el número de buenos y malos que caen en cada grupo (bin).
3. Se calcula el porcentaje de la distribución de buenos y malos en cada grupo.
4. Se calcula el WOE mediante el logaritmo de la fórmula (4.1) y el IV con la fórmula (4.2).

A continuación, se muestra con un ejemplo simulado cómo se calculan tanto el valor de la información como el peso de la evidencia para la variable edad agrupada en clases.

Edad	Nº Total Clientes	Distribución	Buenos	Distribución Buenos	Malos	Distribución Malos	Tasa de Malos	WOE	WOE Tanto por 1	Diferencia Distribuciones	IV parciales
Perdidos	1.000	2,94%	840	2,75%	160	4,60%	16,00%	-51,31218	-0,5131218	-1,85%	0,00946919
18–23	3.000	8,82%	2.050	6,72%	950	27,30%	31,67%	-140,2217	-1,4022168	-20,58%	0,28860348
24–30	8.000	23,53%	6.800	22,28%	1.200	34,48%	15,00%	-43,67489	-0,4367489	-12,20%	0,05329335
30–35	12.000	35,29%	11.100	36,37%	900	25,86%	7,50%	34,09557	0,34095571	10,51%	0,03582601
35–42	6.000	17,65%	5.800	19,00%	200	5,75%	3,33%	119,5946	1,19594592	13,26%	0,15854422
42+	4.000	11,76%	3.930	12,88%	70	2,01%	1,75%	185,655	1,85654955	10,87%	0,20171982
Total	34.000	100,00%	30.520	100,00%	3.480	100,00%					0,74745607

Figura 4.1: Ejemplo del cálculo de WOE e IV.

Las columnas Distribución, Distribución Buenos y Distribución Malos se refieren al porcentaje de distribución de cada grupo del total, de los buenos y de los malos, respectivamente. Por ejemplo, para el grupo de edades entre 30 y 35 años, hay 12000 clientes y estos representan el 35.29 % de los clientes de la muestra. De estos clientes, 11100 son buenos y representan el 36.37 % de los clientes buenos de la muestra y 900 son malos, los cuales representan al 25.86 % de los malos. El WOE para este grupo de edad se calcula así:

$$WOE = [\log\left(\frac{36.37}{25.86}\right)] \times 100 = 34.09.$$

La componente de IV para este grupo se edad es

$$IV = (0.3637 - 0.2586) \times log\left(\frac{0.3637}{0.2586}\right) = 0.0358.$$

Para este conjunto de datos el valor del IV es la suma de la columna de IV parciales y tiene el valor total de 0.7474.

Una vez calculado el IV, surge la pregunta de ¿cómo interpretar este valor de IV? La respuesta viene dada por la regla general que se describe en [19].

- IV < 0.02: impredecible, no es útil para modelar (no discrimina morosos y no morosos).
- $0.02 \leqslant$ IV $\leqslant 0.1$: el poder de predicción es débil.
- $0.1 \leqslant$ IV $\leqslant 0.3$: el poder de predicción es medio.
- IV > 0.3: el poder de predicción es fuerte.

Las características con un valor de IV superior a 0.5 deben examinarse ya que podría haber una sobrepredicción. En este caso, se deben quedar fuera del modelo o utilizarlas de forma controlada. Cuando la scorecard se desarrolla usando características que no están agrupadas, los estadísticos para evaluar el poder predictivo más utilizados son el R-cuadrado y el chi cuadarado. Ambos métodos usan criterios de bondad de ajuste para evaluar las características. La técnica R-cuadrado usa el método de selección por pasos que rechaza las características que no cumplen el aumento de cuadrados. Un punto de corte habitual para el R-cuadrado por pasos es 0.005. El estadístico chi-cuadrado funciona de manera similar, pero normalmente se utiliza el mínimo punto de corte de 0.5. Los puntos de corte pueden aumentar si hay demasiadas características. De la misma manera que con el IV, estas técnicas que no agrupan las variables, buscan seleccionar las características para la regresión.

Es importante tener en cuenta que estas técnicas, no tienen en cuenta las asociaciones parciales ni las interacciones entre las características. Se produce una asociación parcial cuando el efecto de una característica cambia en presencia de otra. En este caso, los métodos multivariados que consideran subconjuntos pueden ser mejores a la hora de hacer el modelo. En cualquier caso, el objetivo es elegir un conjunto de variables para la regresión.

Tendencia Lógica

El poder de predicción medido en términos de WOE y de IV no es el único factor en la elección de una característica para el análisis posterior. Cuando los datos están agrupados, deben tener un orden lógico y tener sentido económico. Esto es, las tasas de malos deben ser mayores en los grupos de las características que tengan menos puntuación e ir disminuyendo a medida que esta puntuación toma valores más grandes.

Si se quitan los valores faltantes, las agrupaciones de una característica tienen una relación monótona creciente con el valor del WOE; es decir, existe una relación monótona creciente y lógica entre las clases de la variable edad y la proporción de malos. Esto confirma la experiencia comercial en el

sectores de crédito de que los jóvenes tienden a tener, en general, un riesgo más alto de ser morosos que la población de mayor edad.

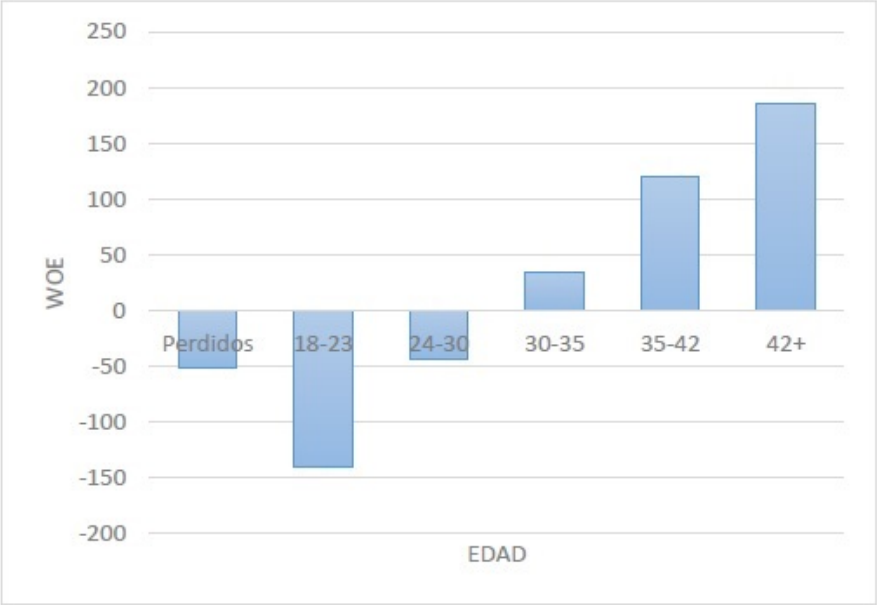

Figura 4.2: Ejemplo del cálculo de WOE e IV.

Se puede experimentar con diferentes agrupaciones para eliminar las inversiones (donde la tendencia se invierte) y otras relaciones ilógicas. Las tendencias generales se pueden ver al observar la relación entre WOE y los atributos desagrupados, ya que la agrupación suaviza la curva.

Sin embargo, en algunos casos estas inversiones pueden estar reflejando un comportamiento de los datos, y eliminar esta inversión puede reducir el poder predictivo de la característica.

Estos casos primero deben ser investigados, para ver si hay un explicación comercial para tal comportamiento. En general, la agrupación sirve para reducir el "sobreajuste". Una curva convexa podría ser un ejemplo de estas inversiones con respecto al WOE y debe mantenerse así si la relación puede ser explicada. Si las cuentas tienen una muy baja utilización presentan un mayor riesgo, y este va disminuyendo hasta cierto punto. Después, el riesgo comienza a aumentar a medida que aumenta la utilización.

Las variables nominales se agrupan para poner atributos con WOE similar juntos, y, como con las variables continuas, para maximizar la diferencia de un grupo a otro.

A continuación se presenta un par de ejemplos. El primero es un ejemplo de una tendencia que no sigue un patrón lógico. El segundo es un ejemplo donde se tienen dos tendencias lógicas y se puede razonar cuál de ellas tiene un poder predictivo más fuerte observando las gráficas de los valores del WOE.

La Figura 4.3 ilustra un ejemplo de una tendencia que no es lógica. En este conjunto de datos, esta característica es débil y no muestra ninguna relación lógica entre la edad y la bandera de rendimiento moroso/no moroso.

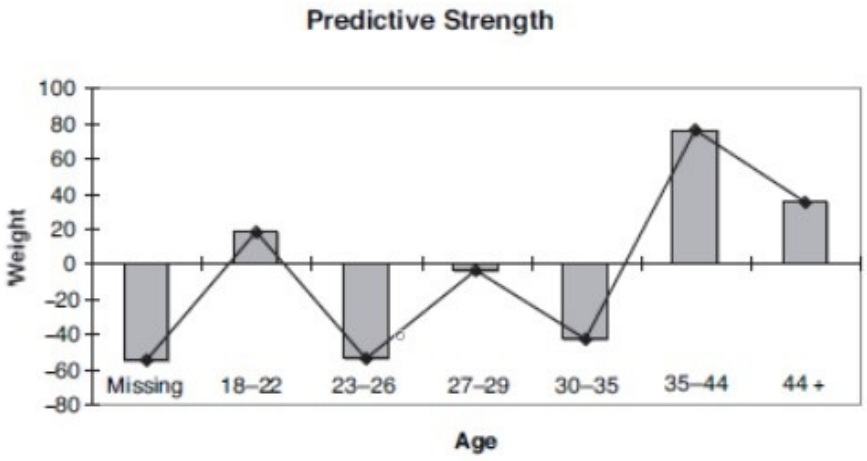

Figura 4.3: Ejemplo de tendencia ilógica.

La Figura 4.4 muestra dos relaciones de WOE donde ambas son lógicas. La línea que tiene los marcadores cuadrados tiene más pendiente y representa un poder predictivo más fuerte. Esto se reflejará en su valor de IV.

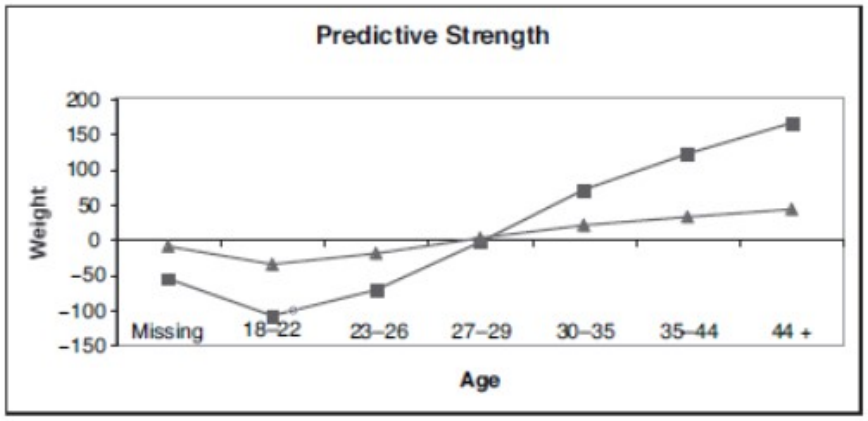

Figura 4.4: Ejemplo de dos tendencias lógicas.

Consideraciones

1. El valor de información aumenta a medida que el número de grupos de una variable independiente aumenta. Se debe tener cuidado cuando haya más de 20 grupos, ya que podría ser que haya un número pequeño de morosos o de no morosos en alguno de estos grupos.

2. El valor de la información no se debe utilizar como un método de selección de características cuando se está construyendo un modelo de clasificación que no sea la regresión logística binaria (por ejemplo, bosque aleatorio) ya que está diseñado solo para el modelo de regresión logística binaria.

El modelo de regresión logística es una de las técnicas estadísticas más utilizadas para resolver el problema de clasificación binaria. Estos dos conceptos, IV y WOE, se propusieron a partir de la técnica de regresión logística. Estos dos términos han existido en el mundo de la calificación crediticia durante más de cinco décadas. Se han utilizado tanto como punto de referencia para detectar variables en los proyectos de modelos de riesgo de crédito como para calcular la probabilidad de mora.

4.2. Ejemplo de cálculo de IV con R

Existen varios paquetes en R que nos ayudan a calcular los valores de WOE y de IV. Los mencionaremos a continuación.

4.2.1. Paquete information

El paquete *information* [11] está diseñado para realizar análisis WOE e IV para modelos de clasificación binaria. Para que el paquete se pueda utilizar con la máxima eficiencia posible, previamente es necesario cargar el paquete *data.table* que permite trabajar con grandes volúmenes de datos.

Para entender cómo funciona este paquete se presenta un ejemplo. Los datos con los que se trabaja en el ejemplo se descargan automáticamente cuando se instala el paquete *information* y provienen de una campaña de marketing histórica de una empresa de seguros. Los datos se almacenan en dos archivos .R, uno para el conjunto de datos de entrenamiento y otro para el conjunto de datos de validación. Cada archivo tiene 68 variables predictivas y 10000 registros.

Los conjuntos de datos contienen dos variables binarias:

- Purchase: Esta variable toma el valor 1 si el cliente aceptó la oferta y 0 en caso contrario.
- Treatment: Esta variable es igual a 1 si el cliente estaba en el grupo de prueba (recibió la oferta) y 0 en caso contrario.

Las funciones principales de este paquete son:

- CreateTables(): crea tablas y calcula los valores de IV y de WOE para todas las variables de los datos de entrada.
- Plot(): representa el vector WOE para una variable.
- MultiPlot(): representa varios vectores WOE en una página para poder compararlos.

A través del código siguiente se carga la base de datos y se calculan los valores de IV de las variables.

```
> library(Information)
> library(gridExtra)
> library(data.table)
> library(compareGroups)
> options(scipen=10)

> data(train, package="Information")
> data(valid, package="Information")

> train <- subset(train, TREATMENT==1)
> valid <- subset(valid, TREATMENT==1)
```

```
> IV <- create_infotables(data=train,
+                         valid=valid,
+                         y="PURCHASE")
[1] "Variable TREATMENT was removed because it has only 1 unique level"
> grid.table(head(IV$Summary), rows=NULL)
```

Las seis variables con valores más altos de IV son las que se pueden ver en la Figura 4.5.

Variable	IV	PENALTY	AdjIV
N_OPEN_REV_ACTS	1.0107695	0.08385690	0.9269126
TOT_HI_CRDT_CRDT_LMT	0.9345902	0.10269736	0.8318929
RATIO_BAL_TO_HI_CRDT	0.8232539	0.06544355	0.7578104
D_NA_M_SNC_MST_RCNT_ACT_OPN	0.6355466	0.07667477	0.5588718
M_SNC_OLDST_RETAIL_ACT_OPN	0.5573438	0.07840106	0.4789427
M_SNC_MST_RCNT_ACT_OPN	0.5026402	0.06044698	0.4421932

Figura 4.5: Variables con mayor IV.

A continuación se analiza si el WOE presenta una tendencia lógica. El objeto IV$Tables es un data.frame que contiene las tablas WOE para todas las variables en el conjunto de datos de entrada. En la Figura 4.6 se muestra la salida para la variable con mayor IV, que en este caso es N_OPEN_REV_ACTS.

```
grid.table(IV$Tables$N_OPEN_REV_ACTS, rows=NULL)
```

N_OPEN_REV_ACTS	N	Percent	WOE	IV	PENALTY
[0,0]	1469	0.29545455	-2.0465968	0.6401443	0.05703080
[1,2]	958	0.19267900	-0.5900120	0.6958705	0.06226262
[3,3]	310	0.06234916	0.2033085	0.6986029	0.06514553
[4,5]	583	0.11725664	0.4419768	0.7244762	0.06767437
[6,8]	632	0.12711183	0.6148243	0.7810611	0.07159274
[9,11]	453	0.09111022	0.8815772	0.8692672	0.07683238
[12,48]	567	0.11403862	0.9883818	1.0107695	0.08385690

Figura 4.6: Tabla de valores WOE e IV para la variable N_OPEN_REV_ACTS.

La tabla muestra que las probabilidades de purchase = 1 aumentan a medida que aumentan los valores de WOE de la variable N_OPEN_REV_ACTS, aunque la relación no es lineal.

El paquete *information* intenta crear grupos de tamaño uniforme. Sin embargo, no siempre es posible debido a las relaciones existentes entre los datos, como pasa con la variable N_OPEN_REV_ACTS en [0,0]. Si la variable considerada es categórica, sus distintas categorías aparecerán como filas en la tabla WOE. Además, si la variable tiene valores perdidos, la tabla WOE contendrá una fila "NA"separada que se puede usar para medir el impacto de los valores perdidos. Por lo tanto, el paquete trabaja sin problema con los valores faltantes y las variables categóricas sin ningún tipo de codificación ficticia o imputación. Este paquete también permite representar la tendencia del WOE con la función *plot_infotables*, lo que nos facilita la visualización del la misma, como se puede observar en la Figura 4.7.

```
plot_infotables(IV, "N_OPEN_REV_ACTS")
```

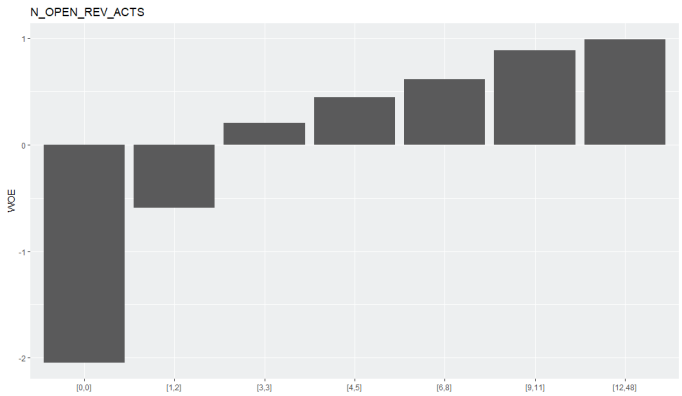

Figura 4.7: Tendencia del WOE de la variable N_OPEN_REV_ACTS.

Se pueden representar varias gráficas de diferentes características a la vez gracias al comando MultiPlot y así poder comparar los patrones del WOE de cada característica. Con el código de R expuesto a continuación se representan las primeras nueve variables en la Figura 4.8 .

```
MultiPlot(IV, IV$Summary$Variable[1:9])}
```

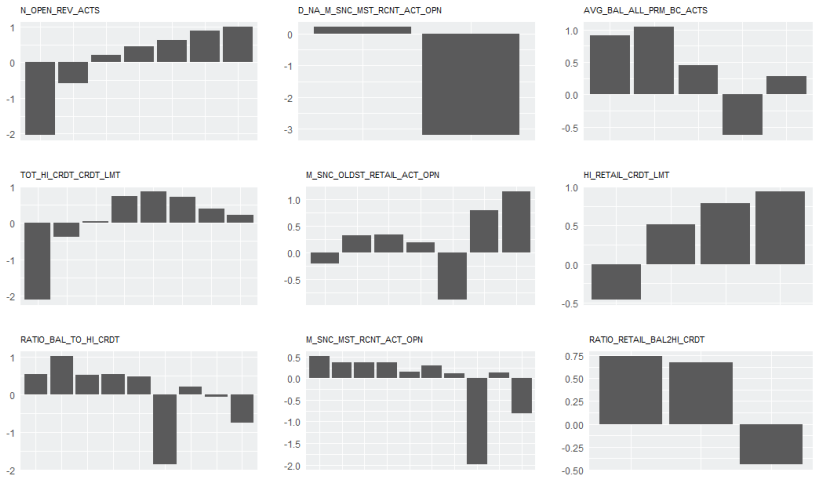

Figura 4.8: Tendencias del WOE de las nueve primeras variables.

4.2.2. Paquete smbinning

El paquete *smbinning* [10], además de clasificar las características en atributos, también permite calcular los valores de WOE e IV. El algoritmo específico que utiliza este paquete es el de árboles de inferencia condicional, descrito en el capítulo 3. Inicialmente, excluye los valores perdidos (NA) para calcular los puntos de corte y los añade posteriormente en el proceso para el cálculo del IV.

En la Figura 4.9, se muestra un ejemplo de código R y salida de este paquete para los datos que se explicarán y se utilizarán en el capítulo 7. En ella, se calculan los valores de IV, el árbol de inferencia condicional y los puntos de corte para categorizar la variable score.

```
> result=smbinning(df=tabla1,y="band_rend_30_2",x="score") ;result
$ivtable
  Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec GoodRate BadRate       Odds  LnOdds     WoE     IV
1   <= 625   5172    3616   1556       5172       3616      1556 0.0510   0.6991  0.3009     2.3239  0.8433 -3.1118 2.4285
2   <= 640   6190    6056    134      11362       9672      1690 0.0611   0.9784  0.0216    45.1940  3.8110 -0.1441 0.0014
3   <= 655   6433    6349     84      17795      16021      1774 0.0635   0.9869  0.0131    75.5833  4.3252  0.3702 0.0073
4   <= 678  22717   22613    104      40512      38634      1878 0.2242   0.9954  0.0046   217.4327  5.3819  1.4269 0.2466
5   <= 697  26189   26163     26      66701      64797      1904 0.2584   0.9990  0.0010  1006.2692  6.9140  2.9590 0.7382
6    > 697  34639   34638      1     101340      99435      1905 0.3418   1.0000  0.0000 34638.0000 10.4527  6.4977 2.2601
7  Missing      0       0      0     101340      99435      1905 0.0000      NaN     NaN         NaN     NaN     NaN    NaN
8    Total 101340   99435   1905         NA         NA        NA 1.0000   0.9812  0.0188    52.1969  3.9550  0.0000 5.6821

$iv
[1] 5.6821

$ctree

Model formula:
band_rend_30_2 ~ score

Fitted party:
[1] root
|   [2] score <= 625: 0.699 (n = 5172, err = 1087.9)
|   [3] score > 625
|   |   [4] score <= 655
|   |   |   [5] score <= 640: 0.978 (n = 6190, err = 131.1)
|   |   |   [6] score > 640: 0.987 (n = 6433, err = 82.9)
|   |   [7] score > 655
|   |   |   [8] score <= 678: 0.995 (n = 22717, err = 103.5)
|   |   |   [9] score > 678
|   |   |   |   [10] score <= 697: 0.999 (n = 26189, err = 26.0)
|   |   |   |   [11] score > 697: 1.000 (n = 34639, err = 1.0)

Number of inner nodes:    5
Number of terminal nodes: 6

$bands
[1] 539 625 640 655 678 697 729

$x
[1] "score"

$col_id
[1] 2

$cuts
[1] 625 640 655 678 697
```

Figura 4.9: Código del paquete smbinning

4.3. Estadístico chi-cuadrado

Para la definición del estadístico chi-cuadrado, se seguirá utilizando la notación de b_i y m_i como el número de cuentas buenas y malas para la característica i que se introdujo en la Sección 4.1.

El número esperado de buenos y malos en el atributo j de la característica i se definen por:

$\hat{b}_{ij} = \dfrac{(b_{ij}+m_{ij})b_i}{b_i+m_i}$ y $\hat{m}_{ij} = \dfrac{(b_{ij}+m_{ij})m_i}{b_i+m_i}$, suponiendo que en cada atributo de la misma ca-

racterística, la respectiva proporción de buenos y malos es igual que la proporción en el total de la característica.

El estadístico χ^2 viene dado por:

$$\chi^2 = \sum_{j=1}^{n_i} \left(\frac{(b_{ij} - \hat{b}_{ij})^2}{b_{ij}} + \frac{(m_{ij} - \hat{m}_{ij})^2}{m_{ij}} \right) \sim \chi^2_{k-1}.$$

Este estadístico toma valores pequeños si $b_{ij} \simeq \hat{b}_{ij}$ y grandes si ambos valores son muy diferentes. El estadístico chi-cuadrado se usa para medir lo diferentes que son las odds en cada clase. Cuanto más grande sea el valor del estadístico mayores diferencias hay en las odds, por lo que si se comparan dos agrupaciones distintas es preferible el valor más alto del estadístico.

4.3.1. Ejemplo del cálculo del estadístico chi-cuadrado

Se tiene una característica i que fue agrupada en tres categorías distintas. En este ejemplo se calcula el estadístico chi cuadrado.

Atributos	b_{ij}	m_{ij}	$b_{ij} + m_{ij}$	$\hat{b}_{ij}$	$\hat{m}_{ij}$
Grupo 1	50000	2500	52500	49218.75	3281.25
Grupo 2	6500	400	6900	6468.75	431.50
Grupo 3	17000	2000	19000	17812.50	1187.50
Total	73500	4900	78400	73499.80	4900.25

$$\chi^2 = \frac{(50000 - 49218.75)^2}{49218.75} + \frac{(2500 - 3281.25)^2}{3281.25} + \frac{(6500 - 6468.75)^2}{6468.75}$$
$$+ \frac{(400 - 431.50)^2}{431.50} + \frac{(17000 - 17812.50)^2}{17812.50} + \frac{(2000 - 1187.50)^2}{1187.50} = 793.84.$$

Estos cálculos también se pueden hacer con R usando la función chisq.test como se muestra a continuación:

```
> tabla<-matrix(c(50000,6500,17000, 2500,400,2000),ncol=2,byrow=F)
> colnames(tabla)=c("buenos","malos")
> rownames(tabla)=c("grupo1","grupo2", "grupo3")
> tabla
       buenos malos
grupo1  50000  2500
grupo2   6500   400
grupo3  17000  2000

> chisq.test(tabla)
Pearson's Chi-squared test
```

```
data:  tabla
X-squared = 793.81, df = 2, p-value < 2.2e-16
```

El valor del estadístico es 793.81.

Si se tiene otra clasificación diferente y se quiere ver cuál de ellas es mejor, se puede utilizar el test chi cuadrado. Será preferible aquella agrupación con un valor mayor del estadístico.

Atributos	b_{ij}	m_{ij}	$b_{ij} + m_{ij}$
Grupo 1	50000	2500	52500
Grupo 2	14000	2300	16300
Grupo 3	9500	100	9600
Total	73500	4900	78400

```
> tabla<-matrix(c(50000,14000,9500, 2500,2300,100),ncol=2,byrow=F)
> colnames(tabla)=c("buenos","malos")
> rownames(tabla)=c("grupo1","grupo2", "grupo3")
> tabla
       buenos malos
grupo1  50000  2500
grupo2  14000  2300
grupo3   9500   100
> chisq.test(tabla)

Pearson's Chi-squared test

data:  tabla
X-squared = 2361.7, df = 2, p-value < 2.2e-16
```

En este caso, el valor del estadístico es 2361.7 > 793.81, por lo que se prefiere esta agrupación. Se puede calcular el estadístico del IV para comprobar que efectivamente esta clasificación tiene una mejor capacidad predictiva.

Para la primera clasificación se obtiene un valor de IV de 0.15.

Atributo	Nº Total Clientes	Distribución	Buenos	Distribución Buenos	Malos	Distribución Malos	Tasa de Malos	WOE	WOE Tanto por 1	Diferencia Distribuciones	IV parciales
Grupo 1	52.500	66,96%	50.000	68,03%	2.500	51,02%	4,76%	28,76821	0,28768207	17,01%	0,04892552
Grupo2	6.900	8,80%	6.500	8,84%	400	8,16%	5,80%	8,004271	0,08004271	0,68%	0,00054451
Grupo 3	19.000	24,23%	17.000	23,13%	2.000	40,82%	10,53%	-56,7984	-0,567984	-17,69%	0,10045976
Total	78.400	100,00%	73.500	100,00%	4.900	100,00%					0,14992979

Figura 4.10: Cálculo del IV para la primera agrupación.

Para la segunda agrupación se obtiene un IV de 0.50.

Atributo	Nº Total Clientes	Distribución	Buenos	Distribución Buenos	Malos	Distribución Malos	Tasa de Malos	WOE	WOE Tanto por 1	Diferencia Distribuciones	IV parciales
Grupo 1	52.500	66,96%	50.000	68,03%	2.500	51,02%	4,76%	28,76821	0,28768207	17,01%	0,04892552
Grupo2	16.300	20,79%	14.000	19,05%	2.300	46,94%	14,11%	-90,1902	-0,901902	-27,89%	0,2515509
Grupo 3	9.600	12,24%	9.500	12,93%	100	2,04%	1,04%	184,5827	1,84582669	10,88%	0,20090631
Total	78.400	100,00%	73.500	100,00%	4.900	100,00%					0,50138273

Figura 4.11: Cálculo del IV para la segunda agrupación.

De nuevo, según los valores del IV se obtiene que la mejor clasificación es la segunda.

Capítulo 5

Técnicas de validación del modelo

Los modelos de puntuación crediticia se utilizan como herramientas de clasificación y de predicción de la solvencia de los clientes de las entidades bancarias. Una vez que se construye el modelo, es necesario validar que discrimina adecuadamente. Esta validación se realiza mediante técnicas estadísticas que permiten garantizar que los modelos poseen un buen poder predictivo. Estas técnicas permiten medir y comparar la bondad del ajuste, la exactitud de las predicciones y el poder discriminante del modelo.

Existen varias técnicas de validación del modelo, en este capítulo se analizarán las técnicas de validación según los criterios Akaike y Bayes ([1], [3] y [9]) y el análisis de las curvas ROC ([5],[23]) y los valores del área bajo la curva y el índice de Gini ([13],)

5.1. Criterios de selección de variables y comparación de modelos

En los capítulos anteriores se explica como se construye un modelo de riesgo de crédito. Para ello, es necesario analizar las características mediante técnicas como el estadístico del valor de la información o el estadístico chi-cuadrado. Aquellas características que tienen unos valores aceptables de IV, que tienen sentido económico y que presenten un patrón de tendencia del WOE tienen un buen poder predictivo y se seleccionan como variables del modelo de regresión logística que se construye posteriormente.

Ahora lo que se pretende es seleccionar qué variables quedan en el modelo, puesto que puede haber interacciones entre las mismas. Incluir variables poco significativas o con información redundante puede distorsionar la capacidad predictiva. Por otro lado, no incluir variables explicativas altamente significativas da una pobre estimación de los parámetros. También se busca comparar varios modelos y decidir cuál de ellos es mejor en cuanto a ciertos criterios.

5.1.1. Algoritmos de selección de variables paso a paso

El método de selección de variables del tipo stepwise o paso a paso se debe a Efroymson [4]. Estos algoritmos actúan incluyendo o excluyendo una sola variable explicativa crediticia en cada paso del algoritmo, de acuerdo a cierto criterio, como por ejemplo el criterio de información de Akaike (AIC) o el criterio de información de Bayes (BIC) que se verán más adelante. Los procedimientos más utilizados son:

- **Backward (hacia atrás)**

Este método empieza con la regresión de Y en función de todas las variables independientes potencialmente influyentes. Para cada variable independiente se realiza un contraste F sobre si su coeficiente es 0 controlando por las demás variables dentro del modelo. Se elimina aquella variable con coeficiente no significativo, cuyo nivel crítico o p-valor se aproxime más a 1. Se realiza la regresión con las variables restantes y se vuelve a repetir este procedimiento de eliminación hasta que todas las variables sean significativas. La ventaja de este algoritmo es que se evita la exclusión de variables significativas pero tiene el inconveniente de que exige una gran capacidad de cálculo y su ejecución puede dar problemas cuando existe multicolinealidad.

- **Forward (hacia delante)**

  Este método empieza eligiendo una única variable y va incluyendo una a una hasta obtener el modelo definitivo. En cada paso se elige la variable más significativa en el contraste F que compara el modelo con esa nueva variable explicativa añadida respecto al modelo que considera las variables previamente introducidas. El algoritmo se detiene cuando no hay variables con coeficientes significativos. Tiene la ventaja de que requiere poca capacidad de cálculo pero el inconveniente de que conduce a problemas de error de especificación porque no es capaz de eliminar variables, cuando una vez introducidas otras, su presencia no sea necesaria.

- **Pasos sucesivos**

  Es una modificación del procedimiento hacia delante, en el sentido de que cada vez que se incluye una variable nueva, el papel de las ya presentes es reevaluado mediante un contraste F, pudiendo eliminarse alguna de las ya incluidas. Una variable introducida en una etapa anterior es eliminada, si al repetir el análisis siendo introducida en último lugar no resulta significativa. Este procedimiento es muy utilizado, sin embargo, debe ser visto tan sólo como una ayuda. Se debe trabajar con niveles de F no muy restrictivos para evitar excluir variables relevantes en el modelo.

5.1.2. Criterios globales

Se pueden utilizar los siguientes criterios globales que nos ayudan a elegir un modelo de modo que se tengan en cuenta el ajuste y el número de parámetros del modelo. Se escoge el modelo cuya medida global sea mejor. Algunos de los criterios globales más utilizados son:

- **Criterio R^2 ajustado**

  El coeficiente de determinación lineal R^2 es la medida de bondad de ajuste por excelencia cuando se estudia el ajuste lineal de un modelo de regresión. En el contexto de riesgo de crédito, el coeficiente R^2 ha sido utilizado, por ejemplo, para evaluar la calidad del ajuste de modelos basados en el enfoque lineal multivariante de Altman.

  Debido a que el modelo utilizado es de regresión logística y por tanto, no es un modelo de regresión lineal, la aplicación del coeficiente R^2 puede no ser adecuada en este caso. Para corregir este problema existen medidas de bondad de ajuste que se conocen como coeficientes pseudo-R^2 y se utilizan en el ajuste de modelos lineales generalizados. Dos de las medidas más utilizadas en la validación del ajuste de modelos de regresión logística son el coeficiente pseudo-R^2 de Cox-Snell y el coeficiente pseudo-R^2 de Nagelkerke.

  Con estos criterios se comparan modelos y se elige aquel que tenga un valor de R^2 ajustado mayor.

- **Criterio de información de Akaike (AIC)**

  El criterio de Akaike [1] sirve para cuantificar la distancia en información entre el modelo teórico y el modelo estimado y por ello es una medida apropiada para discriminar entre los modelos de puntuación crediticia.

Es un criterio de log-verosimilitud penalizada y el estadístico de Akaike viene dado por:

$$AIC = -2L(\hat{\beta}) + 2p$$

donde $l(\hat{\beta})$ es el logaritmo de la función de verosimilitud que viene dada a partir de la expresión

$$L(\beta) = \sum_{i=1}^{n} Y_i X_i' \beta - \sum_{i=1}^{n} \log(1 + e^{X_i' \beta})$$

vista en el capítulo 2 y el término de $2p$ es un valor de penalización debido a la parametrización del modelo que depende del número, p, de parámetros estimados.

Se basa en la idea de que un modelo será mejor cuánto mayor sea su verosimilitud y menor su número de parámetros, de forma que entre dos modelos de puntuación crediticia será preferible el de menor AIC.

- **Criterio de información de Bayes (BIC)**

  Este criterio también es conocido como criterio de información de Schwarz. Debido a que las propiedades de consistencia del estadístico AIC han sido criticadas por no tomar en cuenta el efecto del tamaño muestral, el criterio BIC ha mostrado ser una alternativa apropiada cuando el tamaño muestral incide negativamente sobre la dimensión del modelo, es decir, cuando la complejidad de este aumenta con el tamaño de la muestra.

  Al igual que el AIC es un criterio de verosimilitud penalizada, pero con una penalización más fuerte con respecto al número de variables. El estadístico de Bayes viene dado por:

$$BIC = -2L(\hat{\beta}) + \log(n)p$$

  siendo p el número de parámetros del modelo, n el tamaño de la muestra y el término $\log(n)p$ se introduce para penalizar el efecto que provocan n y p. Al igual que para el criterio de Akaike se debe elegir aquel modelo de puntuación crediticia cuyo valor de BIC es menor.

En general, si la complejidad del modelo estimado no aumenta de manera considerable con el tamaño muestral, basta con emplear el estadístico AIC para elegir el mejor modelo, de lo contrario, es preferible emplear el estadístico BIC.

La función *step* hace una regresión por pasos que utiliza el criterio de Akaike (AIC). El código genérico de esta función es el siguiente:

```
step (modelo_inicial,scope,direction=c("both", "backward", "forward"), k = 2,...)
```

Por defecto utiliza el método backward. En caso de querer utilizar el método forward se debe introducir como *modelo_ inicial* el modelo con la constante y en *scope* el modelo con todas las variables. Si se quiere utilizar el criterio BIC basta con especificar $k = \log(n)$.

5.2. Pruebas de diferencias de dos poblaciones

Cuanto mayor sea la diferencia de las puntuaciones del score entre los grupos de morosos y no morosos, mayor será la capacidad discriminante del modelo. Entre las técnicas para determinar la diferencia entre las puntuaciones de clientes buenos y malos destacan la prueba de Kolmogorov-Smirnov, el índice de Gini y las curvas ROC.

5.2.1. Contraste de Kolmogorov-Smirnov

Para esta sección se seguirán las referencias de [2] y [13]. El contraste Kolmogorov-Smirnov (KS) es una de las técnicas que se utilizan para validar un modelo de puntuación crediticia. Se emplea para probar si dos muestras independendientes provienen de la misma distribución. Es una prueba no paramétrica para la bondad de ajuste.

El estadístico se calcula como la máxima diferencia absoluta entre sus distribuciones empíricas y lo que se pretende es detectar las discrepancias entre las frecuencias relativas acumuladas de buenos y malos. Estas diferencias están determinadas no solo por las medias, sino también por la dispersión y, en general, la forma de las dos distribuciones.

Para este test se tienen las siguientes hipótesis:

H_0 : Las distribuciones de morosos y no morosos son iguales.

H_1 : Las distribuciones de morosos y no morosos son distintas.

Para explicar el estadístico de Kolmogorov-Smirnov es necesario definir previamente la función de distribución empírica. Sea $X_1, \ldots, X_n$ una muestra de una población con una función de distribución continua F. La función de distribución empírica tiene la expresión siguiente:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

donde $I(\cdot)$ es un indicador tal que

$$I(\cdot) = \begin{cases} 1 & \text{si se cumple la condición} \\ 0 & \text{en otro caso} \end{cases} \tag{5.1}$$

Tomando los valores ordenados de la muestra $X_{(1)}, \ldots, X_{(n)}$ la función de distribución puede ser escrita de la siguiente forma:

$$F_n(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ \dfrac{k}{n} & \text{si } X_{(k)} \leq x < X_{(k+1)} \\ 1 & \text{si } x \geq X_{(n)} \end{cases} \tag{5.2}$$

Se calculan las frecuencias relativas acumuladas $F_{n_1}(x)$ y $G_{n_2}(x)$ correspondientes a las muestras de no morosos y morosos con tamaños n_1 y n_2 respectivamente. A continuación, se calculan las diferencias relativas acumuladas $F_{n_1}(x) - G_{n_2}(x)$. El estadístico de Kolmogorov-Smirnov es el supremo de la diferencia de las distribuciones de frecuencias acumuladas

$$D_{n_1,n_2} = sup_x \mid F_{n_1}(x) - G_{n_2}(x) \mid .$$

Si las dos muestras proceden de la misma población, sus funciones de distribución empíricas no pueden ser muy distintas, y rechaza la hipótesis nula de que las distribuciones de morosos y no morosos son iguales si el estadístico toma un valor suficientemente grande. Para ello, se utilizan las tablas de este estadístico, para encontrar el valor crítico λ tal que $P(D_{n_1,n_2} \geq \lambda) = \alpha$. Dicho valor crítico puede aproximarse, cuando ambas muestras son grandes, por

$$\lambda = K\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

siendo K el valor obtenido de la tabla de Kolmogorv-Smirnov.

5.2.2. Validación mediante curvas ROC

Una técnica de análisis utilizada tradicionalmente para comparar el poder discriminante de los modelos de clasificación son las curvas ROC (del inglés, receiver operating characteristic). Una de las ventajas que ofrecen las curvas ROC como medidas de validación es la flexibilidad de su construcción, ya que no dependen de supuestos sobre el método probabilístico que subyace en la regla discriminante. Las curvas ROC tienen una enorme cantidad de aplicaciones, siendo muy utilizadas en los ámbitos del credit scoring y también en medicina entre otros. Es una prueba basada en una variable de decisión, y cuyo objetivo es clasificar a los individuos de una población en dos grupos: uno que presente un evento de interés y otro que no lo presente.

Nociones básicas

Sea D una distribución Bernouilli de parámetro p que se llamará prevalencia del evento sobre la población. La variable D es la condición real del sujeto y se denomina estado.

$$D = \begin{cases} 0 & \text{si el individuo no presenta el evento} \\ 1 & \text{si el individuo presenta el evento} \end{cases} \tag{5.3}$$

Por tanto, $p = P(D = 1) = P(\text{presentar el evento})$ y $1-p = P(D = 0) = P(\text{no presentar el evento})$.

En el contexto de riesgo de crédito, se tiene una cartera de n clientes en la que solo existen dos clases de créditos, los que provienen de la población de créditos morosos ($D = 1$) y los que provienen de la población de los créditos no morosos ($D = 0$).

Sea Y una variable aleatoria que se denomina prueba o marcador de la que se quiere estudiar su poder discriminante y que toma dos resultados posibles: Positivo P si el individuo se clasifica como moroso o Negativo N en caso de que se clasifique como no moroso.

$$Y = \begin{cases} 0 = \text{Negativo} & \text{si } x \geq c \\ 1 = \text{Positivo} & \text{si } x < c \end{cases} \tag{5.4}$$

siendo X una variable que mide una característica en cada individuo. El valor de c es el punto de corte o valor umbral por encima del cuál la prueba considera al individuo moroso aunque su estado real sea no moroso y viceversa.

Si se extrae una muestra de la población, un estimador de la prevalencia es:

$$p = \frac{\text{número de morosos de la muestra}}{\text{total de individuos de la muestra}}.$$

Para valorar la validez de la prueba se establece una comparación entre el resultado Y y el estado D, dando lugar a la división de la población en estos cuatro subgrupos.

	Moroso$\equiv D=1$	No Moroso$\equiv D=0$
Prueba+$\equiv Y=1$	Verdadero Positivo V_+	Falso Positivo F_+
Prueba-$\equiv Y=0$	Falso negativo F_-	Verdadero Negativo V_-

Cuadro 5.1: Posibles resultados de la prueba

En el Cuadro 5.1 se cruzan los posibles valores de la prueba frente al verdadero estado. Este resultado no tiene porqué tener la misma distribución en el grupo de morosos que de no morosos. En el caso de que las distribuciones fuesen iguales, entonces los morosos y los no morosos se comportarían igual ya que las distribuciones estarían solapadas. En este caso, la prueba sería inútil para detectar diferencias entre morosos y no morosos. Un ejemplo de esta situación se puede ver en la Figura 5.4. Cuánto más alejadas estén las distribuciones menos solapamiento habrá y, por tanto, menos falsos negativos y falsos positivos. Un caso de este tipo sería el representado en la Figura 5.2.

La evaluación de la prueba cuantifica la magnitud de aciertos y errores que puedan cometerse. Si un individuo no moroso se clasifica y predice correctamente como no moroso se llama verdadero positivo V_+ y si un individuo moroso se clasifica correctamente como moroso se llama verdadero negativo V_-. Puede haber dos tipos de errores, un resultado falso positivo F_+ (clasificar a un individuo no moroso como moroso, error tipo I) y un resultado falso negativo F_- (clasificar a un individuo moroso como no moroso, error tipo II).

El Cuadro 5.1 es una tabla de contingencia que permite calcular las probabilidades relevantes en el problema de clasificación:

Probabilidad de obtener un resultado negativo cuando el individuo es no moroso: $\dfrac{V_-}{V_- + F_+}$.

Probabilidad de obtener un resultado positivo cuando el individuo es moroso: $\dfrac{V_+}{F_- + V_+}$.

Proporción de resultados válidos entre los resultados negativos: $\dfrac{V_-}{V_- + F_-}$.

Proporción de resultados válidos entre los resultados positivos: $\dfrac{V_+}{V_+ + F_+}$.

Estos son los índices de la eficacia de una prueba: especificidad, sensibilidad, valor predictivo positivo y valor predictivo negativo. Por tanto, se definen los siguientes parámetros:

Sensibilidad

La sensibilidad es un parámetro que se mide en el grupo de individuos que verdaderamente son morosos. Es el cociente entre verdaderos positivos y el total de los individuos morosos. Por tanto, es la probabilidad de obtener un resultado positivo cuando el cliente es moroso, o la proporción de verdaderos positivos. La sensibilidad se denota entonces como TPF (True Positive Fraction). En este caso, $D=1, Y=1$.

$S = TPF = \dfrac{\text{Morosos correctamente clasificados}}{\text{Total de morosos}} = \dfrac{V_+}{F_- + V_+}$. El complementario de esta probabilidad es: $1 - S = FNF = \dfrac{F_-}{V_+ + F_-}$. Se puede observar que si $S = 1$ entonces $F_- = 0$.

Especificidad

La especificidad es un parámetro que se mide en el grupo de individuos no morosos. Es el cociente entre verdaderos negativos y el total de no morosos. Por tanto, es la probabilidad de obtener un resultado negativo cuando el individuo no es moroso. Se denota la especificidad como TNF (True

Negative Fraction). En este caso, $D = 0, Y = 0$.

$E = FPR = \dfrac{\text{No morosos incorrectamente clasificados}}{\text{Total de no morosos}} = \dfrac{V_-}{V_- + F_+}$. El complementario de esta probabilidad es: $1 - E = FPF = \dfrac{F_+}{V_- + F_+}$. Se puede observar que si $E = 1$ entonces $F_+ = 0$.

Para cuantificar la precisión del test se pueden utilizar $FPF = (1 - \text{Especificidad})$ y $TPF = \text{Sensibilidad}$. Si se representa el cociente de morosos correctamente clasificados contra los no morosos incorrectamente clasificados se obtiene la curva ROC, donde el eje de abscisas se corresponde con FPF y el eje de las ordenadas con TPF.

	Moroso$\equiv D = 1$	No Moroso$\equiv D = 0$
$Y = 1$	Sensibilidad $TPF = P(Y = 1/D = 1)$	$FPF = P(Y = 1/D = 0)$
$Y = 0$	$FNF = P(Y = 0/D = 1)$	Especificidad $TNF = P(Y = 0/D = 0)$

Curva ROC

La curva ROC es un gráfico en el que se observan los pares de Sensibilidad y el complementario de la Especificidad (1-especificidad). Representa el balance entre la capacidad e incapacidad de clasificar correctamente que posee un modelo discriminante. Denotando por c al punto de corte o umbral de decisión, se tiene:

$$FPF(c) = P(Y \geq c/D = 0)$$
$$TPF(c) = P(Y \geq c/D = 1)$$

Se define la curva ROC como: $ROC = \{(1 - E(c), S(c)\} = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}$ Notar que al aumentar c, $FPF(c)$ y $TPF(c)$ disminuyen. La curva ROC está contenida en el cuadrado $[0, 1] \times [0, 1]$. El paquete de R *pROC* que se describirá al final de esta sección permite representar la curva como se observa en la Figura 5.1. Este paquete representa la especificidad (en lugar de 1-especificidad) frente a la sensibilidad y para ello invierte la recta del eje X, empezando por el 1 hasta el 0.

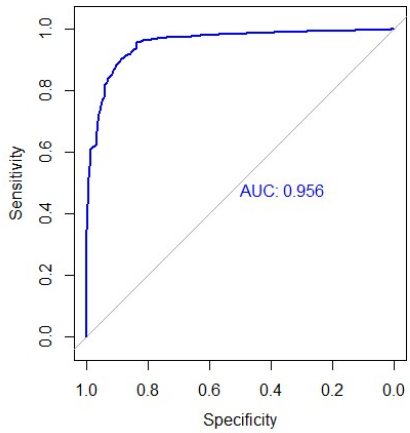

Figura 5.1: Ejemplo de una curva ROC.

Área bajo la curva

Se denomina AUC, o AUROC al área bajo la curva ROC (Area under the curve). Esta área se utiliza para estimar la capacidad de discriminar entre morosos y no morosos. También se usa para comparar pruebas entre sí y determinar cuál es la más eficaz. Se define como:

$$AUC = \int_0^1 ROC(p)dp.$$

Su rango de valores va desde 0 hasta 1. Si el área bajo la curva valiese 1 la prueba sería perfecta, clasificaría al 100 % de los morosos como morosos y al 100 % de los no morosos como no morosos y, por tanto, los grupos estarían perfectamente diferenciados por la prueba. Si el área de la curva valiese 0.5, significaría que la probabilidad de clasificación incorrecta es de 0.5 (sin prejuzgar cómo son de parecidas las probabilidades de clasificar a un individuo como moroso o no moroso). Esto sería una prueba inútil que no proporciona información y no ayuda a poder discriminar. Equivaldría a tirar una moneda al aire para clasificar a un cliente. Esta curva ROC sería aquella que tendría valores de sensibilidad iguales a 1-especificidad en todos los posibles puntos de corte. El área de una prueba inútil se corresponde a la diagonal del cuadrado por lo que se suele representar esta diagonal para ver en cuánto supera la prueba que se valora a lo que sería una prueba inútil. La diagonal representa una situación como la de clasificación al azar. Si el área bajo la curva valiese fuese menor que 0.5, esto significaría que la clasificación comete un error mayor que la mera clasificación al azar.

Una interpretación de los valores de AUC es la siguiente:

- Baja capacidad discriminante: $[0.5, 0.7)$.
- Capacidad discriminante útil: $[0.7, 0.9)$.
- Alta capacidad discriminante: $[0.9, 1]$.

Por tanto, cuánto mayor sea el valor del AUC mejor será la capacidad discriminante y se obtienen mejores resultados.

Dos curvas ROC iguales tienen el mismo área, pero sin embargo, dos áreas iguales no tienen porqué corresponderse con la misma curva ROC. La razón es que una puede ser más sensible y la otra tener más especificidad y esto depende del punto de corte. Un área menor de 0.5 aconseja considerar los negativos como positivos y viceversa. Esto provocaría que la nueva curva ROC resultante tendría un AUC mayor que 0.5.

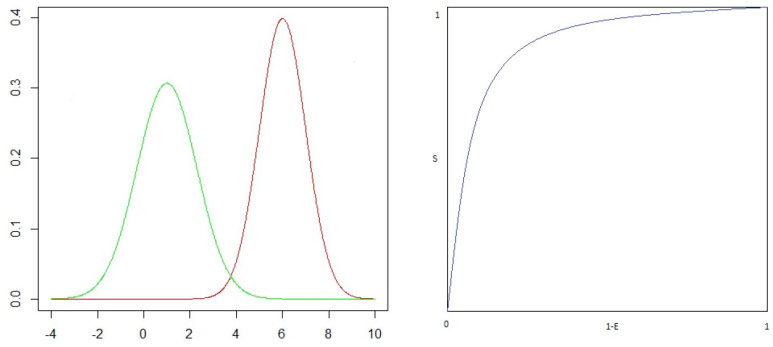

Figura 5.2: Ejemplo de densidades y curva ROC de dos poblaciones bien diferenciadas.

Las curvas ROC permiten describir cuanto de separadas están las distribuciones de especificidad y sensibilidad. En la Figura 5.5 se muestra un ejemplo de dos poblaciones que tienen una alta capacidad discriminante, en la Figura 5.3 se muestran dos poblaciones que tienen solapamiento pero presentan una capacidad discriminante útil y en la Figura 5.4 se muestra una prueba inútil, con dos poblaciones sin poder discriminante, dónde la curva ROC está muy cercana a la diagonal.

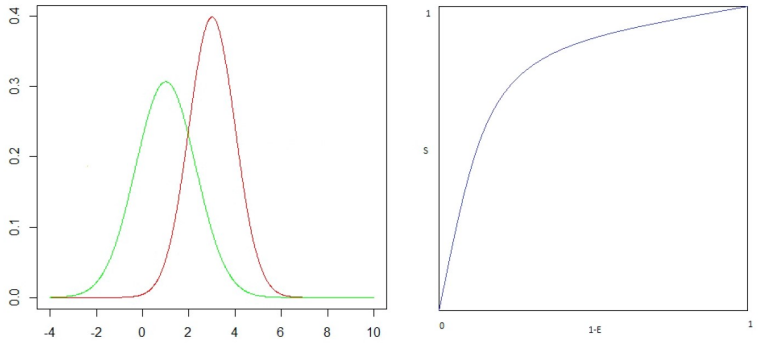

Figura 5.3: Ejemplo de densidades y curva ROC de dos poblaciones con cierto solapamiento.

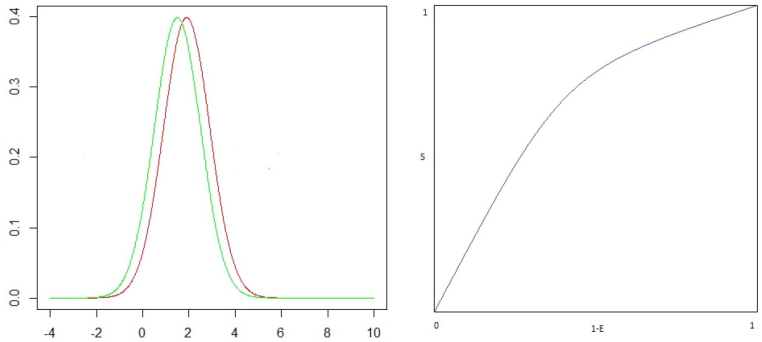

Figura 5.4: Ejemplo de densidades y curvas ROC de dos poblaciones totalmente solapadas.

Paquete pROC

El paquete *pROC* de R [18] permite representar la curva ROC (también su versión suavizada), así como calcular el área bajo la curva y comparar pruebas u obtener un intervalo de confianza para el área bajo la curva.

La función que representa la curva ROC es *roc()*. El siguiente código muestra la curva ROC de un modelo de scoring, donde la variable estado d es la bandera de rendimiento de 30 días y el marcador y es el score. Además, calcula el valor de AUC e intervalos de confianza para este valor. En el código, *percent* permite expresar los resultados en fracción o porcentaje, *na.rm* elimina o no aquellas cuentas a las que les falte algún dato, *direction* se utiliza cuando es necesario invertir la positividad de la curva para que sea cóncava o convexa, *smooth* suaviza o no la curvatura, *auc* permite calcular el área bajo la curva, *ci* obtiene el intervalo de confianza para el AUC y *smooth.method* especifica qué densidades se quieren ajustar los grupos de morosos y no morosos.

```
> ROC=roc(d,y,percent=FALSE,na.rm=TRUE,direction=c("auto","<",>"),smooth=FALSE,
auc=TRUE,ci=TRUE,plot=TRUE,smooth.method = "binormal", density=NULL)

> rocA1C <- with(tabla,roc(d,y))
> plot(rocA1C, col="blue", print.auc=TRUE)
> ci <- as.numeric(ci.auc(rocA1C))
> a<-plot.roc(smooth(rocA1C),col="blue")
> legend(0.85,.15, sprintf("AUC = %.3f;   95%%-CI = %.3f; %.3f",ci[2],ci[1],ci[3]),
bty="n",cex=1.15)
```

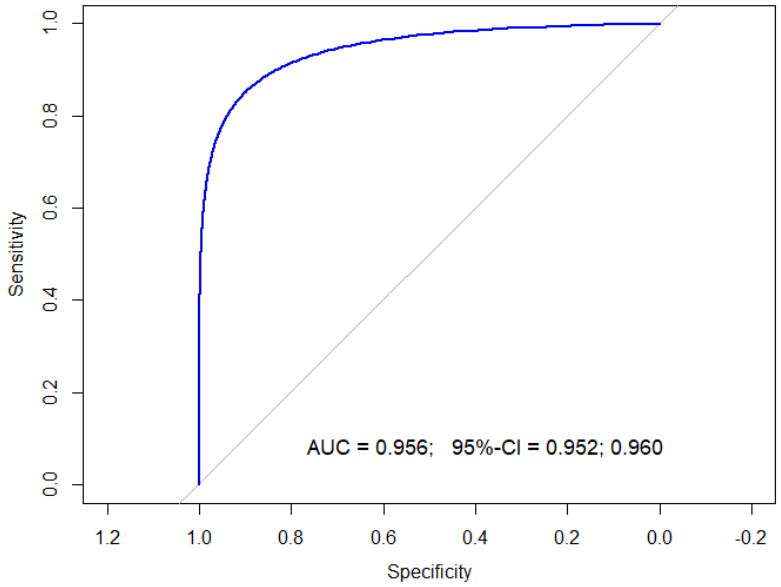

Figura 5.5: Ejemplo de curva ROC

Se pueden calcular el área bajo la curva y el índice de Gini (medida que se define en la sección 5.2.3) como aparece a continuación:

```
> AUC=auc(ROCtabla1);AUC
Area under the curve: 0.9563
> Gini= 2 * AUC - 1;Gini
[1] 0.9125462
```

La función coords permite encontrar aquel valor de la variable que, al ser usado como punto de corte para discriminar entre buenos y malos, maximiza la función de la sensibilidad y la especificidad dada por:

$$(1 - S(c))^2 + (1 - E(c))^2$$

El objetivo es conseguir que la sensibilidad y la especificidad se aproximen simultáneamente todo lo posible a 1.

```
> closest <- coords(rocA1C,"b",ret=c("threshold","specificity","sensitivity","npv",
"ppv"),best.method="closest.topleft")
> closest
threshold    specificity sensitivity       npv         ppv
640.5000000   0.8871391   0.9027304   0.1487414   0.9976105
```

Este paquete también permite realizar un contraste de hipótesis $H_0 : AUC_{y_1} = AUC_{y_2}$ con la función *roc.test*. Para p-valores próximos a cero se rechazaría la hipótesis nula y se aceptaría que las curvas ROC son diferentes.

5.2.3. Validación mediante curvas de Lorentz e índice de Gini

El índice de Gini fue propuesto en 1960 y es uno de los más utilizados para medir la desigualdad entre dos poblaciones. En este caso, se mide la desigualdad de buenos y malos clientes. El índice de Gini se deriva de la curva de Lorentz. La curva de Lorentz es una medida similar a la curva ROC utilizada para comparar modelos y representar la distribución de los casos "malos" y casos totales por deciles en todos los rangos de puntuación. Esto es, mide la capacidad de clasificar los buenos y malos en deciles seleccionados.

La curva de Lorentz de las funciones de distribución F y G es el subconjunto del cuadrado unidad, $[0,1] \times [0,1]$, dado por

$$L(F,G) = \{(u,v)/u = F(x), v = G(x), x \in \mathbb{R}\}.$$

dónde F y G son las funciones de distribución teóricas asociadas a los clientes malos y buenos respectivamente, siendo x la puntuación. Si la puntuación para buenos es mayor que la puntuación para los malos, la curva de Lorentz es cóncava hacia abajo. Si $F(x) = G(x)$, entonces la curva se corresponde con la recta $u = v$ con $u \in (0,1)$. Cuanto más se separe la curva Lorentz de esta recta mayor será la diferencia entre las funciones de distribuciones F y G. El área que se encuentra entre la recta y la curva Lorentz es la medida de desigualdad entre las distribuciones, que se denota por A en la Figura 5.6. Se denota por T al triángulo delimitado por la recta $u = v$, el eje horizontal y la recta $u = 1$. El índice de Gini es el cociente entre el área de $T - A$ y el área de T.

Cuando se desconocen las funciones de distribución $F(x)$ y $G(x)$ pero se tienen muestras aleatorias de tamaños n_1 y n_2 cada una de estas dos distribuciones se pueden estimar la curva Lorentz y el índice de Gini. Se denotan los elementos de la primera muestra como $(a_1, a_2, \ldots, a_{n_1})$ y los de la segunda como $(b_1, b_2, \ldots, b_{n_2})$. Para estimar esta curva de Lorentz y el índice de Gini, es necesario hacer una partición $x_0 \leq x_1 \leq \ldots \leq x_k$ y se obtener los estimadores de F y G en los puntos x_i como sigue:

$$\hat{F}(x_i) = \frac{Card\{a_t \leq x_i, 1 \leq t \leq n_1\}}{n_1}$$

$$\hat{G}(x_i) = \frac{Card\{b_l \leq x_i, 1 \leq l \leq n_2\}}{n_2}.$$

Una aproximación poligonal de la estimación de la curva de Lorentz de $F(x)$ y $G(x)$ es igual a la unión de los segmentos de recta que unen los puntos $(\hat{F}(x_{i-1}), \hat{G}(x_{i-1}))$ y $(\hat{F}(x_i), \hat{G}(x_i))$. El área debajo la curva Lorentz se aproxima por la suma de las áreas de un conjunto de trapecios. El área del i-ésimo trapecio resulta:

$$A_i = \frac{(\hat{F}_i - \hat{F}_{i-1})(\hat{G}_i + \hat{G}_{i-1})}{2}.$$

El área total por debajo de la curva de Lorentz estimada es

$$A = \sum_{i=2}^{k} A_i.$$

El índice de Gini, es, por tanto:

$$Gini = \frac{1/2 - A}{1/2}.$$

El coeficiente de Gini también puede ser expresado como:

$$Gini = 2AUC - 1.$$

El rango de valores del coeficiente de Gini es $[0, 1]$. Cuanto más cerca esté el coeficiente a uno, mejor será la separación de los clientes morosos y de los no morosos.

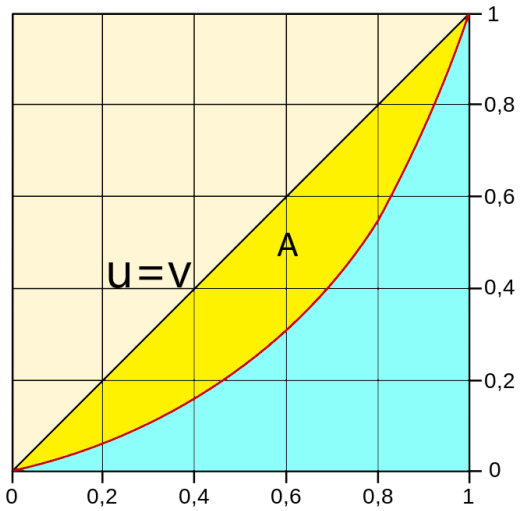

Figura 5.6: Curva de Lorentz

Capítulo 6

Construcción de la tabla de puntuaciones

Para este capítulo se siguen las referencias [13] y [19]. El modelo de regresión logística está dado, como se vio en el capítulo 2, por:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

donde $\beta_0, \beta_1, \ldots, \beta_k$ son parámetros desconocidos y $x_1, \ldots, x_k$ son variables explicativas cuyos valores están en función de la proporción de buenos y malos en cada atributo. El rango de valores de x_i es:

$$x_i = woe_{i1}, woe_{i2}, woe_{ini}.$$

siendo

$$woe_{ij} = \log\left(\frac{b_{ij} m_i}{m_{ij} b_i}\right)$$

De este proceso se obtienen los valores β_i que son necesarios para construir la tabla de puntuaciones.

Las puntuaciones del score se calculan por la ecuación:

$$\text{Score} = \text{Offset} + \text{Factor}\log(odds)$$

siendo Offset un término de traslación y Factor un término de reescalamiento. Los valores de la scorecard son el resultado de una transformación de los coeficientes β_i. Con esta transformación se quieren obtener valores enteros para cada atributo j de la característica i. Se suele calibrar la tabla de puntuaciones de manera que cada cierto aumento en la puntuación P_0, se obtengan el doble de buenos que de malos. Entonces es necesario resolver el sistema de ecuaciones siguiente:

$$\text{Score} = \text{Offset} + \text{Factor}\log(odds)$$

$$\text{Score} + P_0 = \text{Offset} + \text{Factor}\log(2 * odds).$$

Se obtienen así los valores de Offset y de Factor

$$\text{Factor} = \frac{P_0}{\log(2)}$$

$$\text{Offset} = \text{Score} - \text{Factor}\log(odds).$$

La relación del modelo de regresión logística con los WOE viene dada por la siguiente expresión:

$$\text{Score} = \text{Offset} + \text{Factor}\log(odds) = \text{Offset} + \text{Factor}\left(\hat{\beta}_0 + \sum_{i=1}^{p}\hat{\beta}_i woe_{ij}\right).$$

Por lo tanto, se tiene:

$$\text{Score} = \text{Offset} + \text{Factor}\hat{\beta}_0 + \text{Factor}\left(\sum_{j=1}^{p}\hat{\beta}_j woe_{ij}\right).$$

Las puntuaciones de la scorecard dependen de los parámetros que se utilicen.

Cuando llega a la entidad una nueva solicitud de crédito, se calcula su score y en base a ese valor se decide si se le otorga o no el crédito. Si el score es mayor que un punto a se concede el crédito y si es menor o igual a ese punto a se rechaza. A este punto a se le conoce como punto de corte.

Capítulo 7

Aplicación con datos reales

En este capítulo se verá la aplicación de alguno de los métodos expuestos en este trabajo, así como la construcción de un modelo de scoring, con un conjunto de datos reales proporcionados por una entidad bancaria. Toda la información utilizada en este trabajo ha sido proporcionada por parte de la entidad previo tratamiento de la misma para garantizar que no contuviese ningún dato que pudiese servir para identificar a los clientes. Se trata de bases de datos con identificadores de registros que permiten su uso a efectos de tratamiento estadístico pero en ningún caso es posible identificar a qué cliente corresponden los datos de informados en las distintas variables. Se asume que la información descargada de la entidad financiera del cliente ya se ha preprocesado y que la información de cuentas corrientes del cliente ha sido agregada mensualmente. Para construir el modelo de scoring se ha hecho un extracción en las fechas (Agosto 2016 y Febrero 2017). En ambos casos, los datos tienen un histórico de 12 meses. Se trata de dos muestras de 143133 y de 145545 clientes respectivamente, en las cuales se recogen las variables que se muestran en el Cuadro 7.1.

Variables	Explicación
Identificador de la cuenta	Número que identifica cada cuenta
Fecha de fin de mes	Mes de los datos
Saldo de la tarjeta	Cantidad positiva o negativa que hay en una tarjeta a fin de mes
Disposición cajero	Variable binaria que indica si se extrajo dinero o no de algún cajero
Valor de compras	Importe de las compras que se han hecho con la tarjeta durante el mes
Valor de amortizaciones	Importe de los pagos (amortizaciones) que se han hecho en la tarjeta durante el ciclo
Número de tarjetas	Son las tarjetas de crédito que tiene contradas el cliente
Débitos	Retiradas de dinero
Créditos	Ingresos que hacen los clientes
Días de descubierto	Días en los que la cuenta está en números rojos

Variables	Explicación
Indicador de nómina	Indicador de si en la cuenta se cobra una nómina o pensión
Saldo máximo	Cantidad máxima de dinero que tuvo la cuenta durante ese mes
Saldo mínimo	Cantidad mínima de dinero que tuvo la cuenta durante ese mes
Saldo medio	La media de las cantidades de dinero que tuvo la cuenta durante ese mes
Saldo a final de mes	Cantidad de dinero del que dispone la cuenta el último día del mes
Número de cuentas corrientes	Cantidad de cuentas corrientes del mismo cliente
Días consecutivos de descubierto	Días consecutivos en los que la cuenta está en números rojos
Días de impago de tarjetas	Días que la tarjeta está en impago o no en un determinado mes.
Capital concedido hipoteca	Cantidad de dinero concedida al cliente
Saldo pendiente hipoteca	Cantidad de dinero que el cliente que resta por pagar al banco de una hipoteca
Número de hipotecas	Cantidad de hipotecas que tiene el cliente
Capital concedido préstamo	Cantidad de dinero concedida al cliente
Saldo pendiente préstamo	Cantidad de dinero que el cliente resta por pagar al banco de un préstamo
Número préstamos	Cantidad de préstamos que tiene el cliente
Saldo en depósitos	Cantidad de dinero que tiene el cliente en depósitos
Días de impago resto	Días que el cliente está en impago
Bandera de rendimiento 90 días	Indicador de si el cliente ha estado en impago más de 90 días en los 12 meses siguientes a la extracción de la información
Bandera de rendimiento 30 días	Indicador de si el cliente ha estado en impago más de 30 días en los 12 meses siguientes a la extracción de la información

Cuadro 7.1: Tabla con las variables de los datos

La siguiente tabla muestra las poblaciones que se han utilizado:

Procesamiento	Tabla de puntuación	Nº Clientes	Nº Malos	Tasa de Malos
agosto-16	Solo Cuenta corriente	99.311	1.893	1,91%
agosto-16	Cuenta corriente y riesgo	43.822	2.143	4,89%
febrero-17	Solo Cuenta corriente	101.176	1.890	1,87%
febrero-17	Cuenta corriente y riesgo	44.369	1.963	4,42%

Lo primero que se hace es programar una serie de características. Para ello, se van a clasificar a los clientes en dos tablas dependiendo de si los clientes tienen solo cuenta corriente (cuya clasificación, en adelante, será la tabla 1) o si tienen cuenta corriente y un producto de riesgo (tabla 2).

Como se comentó en el capítulo 2, es necesario definir cuándo una cuenta se considera mala. Para este trabajo, las banderas de rendimiento (indicadores de morosidad) que se han considerado, han sido las siguientes:

- Tabla solo cuenta corriente: Bandera a 30 días de impago o si el cliente tenía una refinanciación en 12 meses.
- Tabla de cuenta corriente y producto de riesgo: Bandera a 90 días de impago o si el cliente tenía una refinanciación en 12 meses.

Como el modelo de score es un modelo proactivo, es necesario que los clientes tengan operatoria suficiente para poder ser calificados. Aquellos clientes que en el momento de la calificación no presentan una operatoria suficiente o que están en impago serán excluidos del proceso de calificación. A continuación se detallan cada una de las exclusiones.

- Exclusiones por falta de información.

  Los clientes que presentan exclusiones por falta de información no tienen operatoria suficiente en sus cuentas corrientes para poder otorgarles una calificación. Se excluye al cliente por este motivo si su cuenta corriente presenta alguna exclusión de este tipo. Presentan esta exclusión los clientes que se describen a continuación:

  - Aquellos clientes cuya cuenta corriente no está activa en los 12 meses de histórico.
  - Aquellos clientes cuya cuenta corriente fue recientemente activada (en los últimos 2 meses).
  - Aquellos clientes cuya cuenta corriente está inactiva en los últimos 6 meses.
  - Aquellos clientes que no tienen cuentas corrientes en la entidad.

- Exclusiones por malo.

  Los clientes presentan exclusiones por malo si en alguna de las cuentas tiene en el mes más reciente un descubierto consecutivo de más de 45 días. También se da si la línea de crédito tiene un exceso de más de 45 días en el mes más reciente.

Si un cliente el único producto de riesgo que tiene es una tarjeta, lo más normal es que se califique con la tabla de cuenta corriente más producto de riesgo. Si la tarjeta no se utiliza, se excluye del proceso de calificación y en ese caso el cliente se calificaría por la tabla de solo cuenta corriente pues la tarjeta no tiene operatoria suficiente para formar parte del proceso de calificación.

Una determinada tarjeta se excluye del proceso de calificación si cumple alguna de las siguientes condiciones:

- Tarjeta con menos de tres meses de antigüedad.
- Tarjeta cancelada. Si la tarjeta está cancelada no se considera para la calificación.
- Cuenta recientemente activada.
- Cuenta inactiva en los últimos 6 meses.

Una vez procesadas las exclusiones es el momento de calcular la puntuación del cliente. Si el cliente tiene tarjetas no excluidas, préstamos personales o hipotecarios se califica por la tabla de cuenta corriente y riesgo. En otro caso, se califica por la tabla de solo cuenta corriente.

7.1. Tabla de solo cuenta corriente o Tabla 1

Para los clientes que solo tienen cuenta corriente (en adelante tabla 1) se ha llevado a cabo la programación de las características 151, 176, 189, 194, 202, 210, 219, 224, 247 y 911 que constituyen la tabla. Por motivos de confidencialidad, no se describe en qué consisten dichas características, ya que forman parte de modelos internos de la entidad.

Una vez que se han calculado las características, se le asocia a cada una de ellas una puntuación en función de los valores obtenidos y posteriormente se suman cada una de estas puntuaciones obteniéndose el score final para cada cliente.

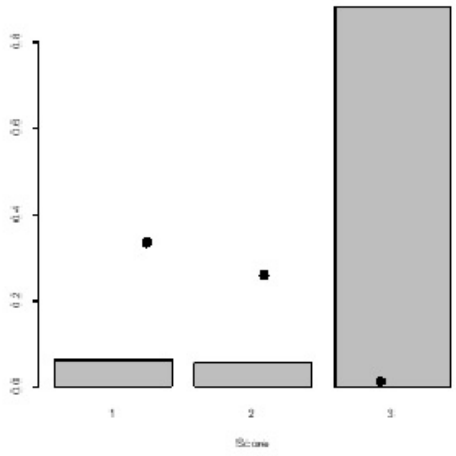

Figura 7.1: Ejemplo de una característica con sentido económico. Los puntos representan las tasas de mora asociadas a cada grupo de puntuaciones. El primer grupo tiene las puntuaciones más pequeñas y el tercero las puntuaciones más altas.

Cuando se tienen las características programadas y se agrupan en base a las puntuaciones, el siguiente paso es analizar la capacidad predictiva de esas variables. Para ello, se calcula el IV de cada una de las características. Las características con valores menores que 0.02 deben excluirse del modelo y las que tengan un valor superior a 0.5 deben examinarse por si hay sobrepredicción. En este caso hay varias variables con unos valores muy superiores a 0.5, por ejemplo las características 202 y 210 entre otras. Pero estas características es normal que tengan unos valores tan elevados de IV debido a que son características muy relacionadas con la mora y por tanto, deben tener una alta capacidad para discriminar entre morosos y no morosos.

Además de tener unos valores aceptables de IV, las características también tienen que tener sentido económico. Esto quiere decir que las puntuaciones más bajas deben tener las tasas de mora más elevadas y a medida que estas puntuaciones aumentan las tasas de mora deben disminuir como ocurre en la Figura 7.1.

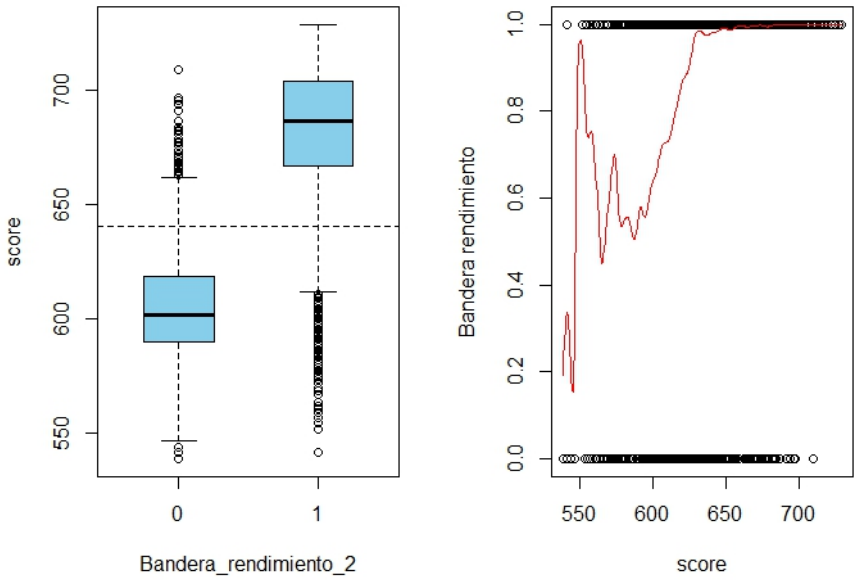

Figura 7.2: Diagrama de cajas y estimador Nadaraya Watson para los datos de la tabla 1.

En la Figura 7.2 se muestra que existe relación entre los clientes que son calificados como buenos y los que son considerados malos y la puntuación. En la gráfica de la izquierda la bandera de rendimiento 1 se corresponde con los clientes no morosos y bandera de rendimiento 0 con los clientes morosos. En el gráfico de caja se aprecia que la posición central de los valores de la puntuación es mayor para los clientes que son considerados buenos. En la gráfica de la derecha se muestra la estimación no paramétrica de la probabilidad de ser considerado bueno o malo en función del score mediante el estimador de Nadaraya-Watson.

En la Figura 7.3 en el primer gráfico se muestran las distribuciones de los clientes de la tabla 1 en función de su bandera de rendimiento (0 si el cliente es malo y 1 si el cliente es bueno). Se obtiene que los clientes con bandera de rendimiento 0 (malos) obtienen puntuaciones más bajas, con una media entorno a una puntuación de 600, y presenta algún dato atípico de clientes que son morosos y puntúan alto. En cambio, los clientes con bandera de rendimiento 1 (buenos) puntúan más alto, (puntuaciones superiores a 650), con algún dato atípico de clientes que son buenos y tienen una puntuación inferior a 600. Por lo tanto, se puede decir que el modelo discrimina bastante bien. Esto también se puede concluir con el valor del AUC=0.956, que es bastante elevado. Si se hace una agrupación del score mediante el paquete *smbinning* de R como se explicó en la sección 4.2.2 se obtiene un IV= 5.6821 y los puntos de corte para las puntuaciones que obtenemos son 625, 640, 655, 678, 697. En el segundo gráfico se representan los porcentajes de clientes que están en estas particiones. En todas las categorías cae al menos un 5 % de la población por lo que estas divisiones son representativas. En el tercer gráfico se representa la tasa de mora. Se puede ver que la tasa de mora es mayor en las puntuaciones más bajas y va disminuyendo a medida que estas puntuaciones puntúan más alto. Por lo tanto, tiene sentido económico. Por último, se representa el WOE en estas categorías. Para las puntuaciones más bajas el WOE es negativo y va creciendo a medida que aumenta el score, por lo que tiene tendencia. En la figura 7.4 se representa la curva ROC y el valor del AUC para la extracción de febrero para la tabla 1.

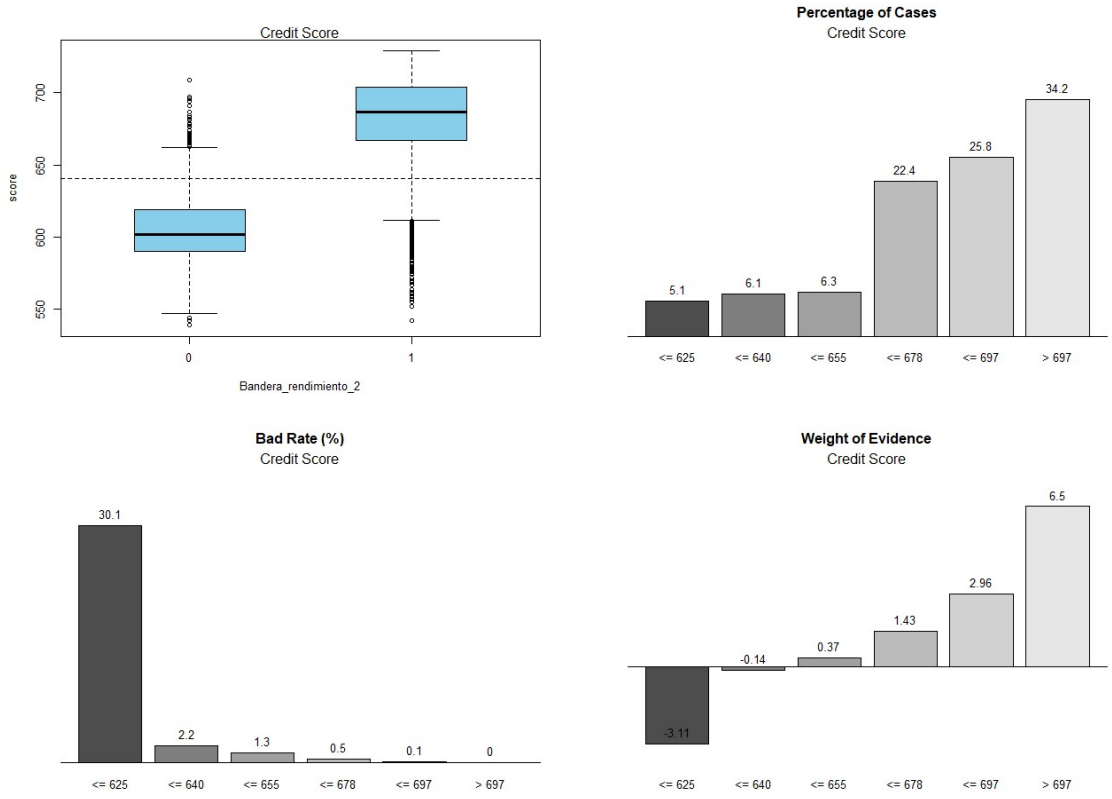

Figura 7.3: Diagrama de cajas, porcentajes de las puntuaciones, tasas de mora y tendencia de los valores de WOE asociados al modelo de la tabla 1.

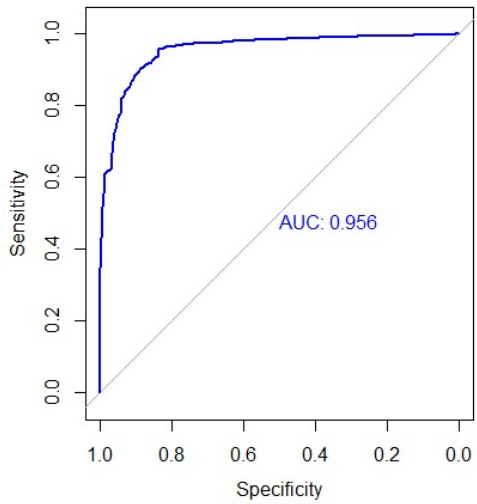

Figura 7.4: Representación de la curva ROC para la tabla 1.

También se obtienen las de estimaciones del AUC y del coeficiente de Gini para la puntuación final para los procesamientos de febrero y agosto, obteniéndose resultados muy similares en ambos procesamientos.

	Febrero	Agosto
AUC	0.9555382	0.9455008
Gini	0.9110763	0.8910015

Cuadro 7.2: Valores de AUC e índice de Gini para las dos muestras de datos de la tabla 1.

Se obtienen unos valores muy elevados de AUC e índice de Gini, por lo que parece que el modelo tiene una alta capacidad predictiva.

Puesto que el comportamiento parece bastante bueno, se va a ajustar un modelo de regresión logística. Para ello, se partirá del modelo con todas las características y se irán eliminando mediante la técnica stepwise.

Se ajusta un modelo de regresión logística y obtiene:

```
> m1<-glm(bandera~tabla1$s151+tabla1$s176+tabla1$s189+tabla1$s194
+tabla1$s202+tabla1$s210+tabla1$s219+tabla1$s224+tabla1$s247
+tabla1$s911,family=binomial(link=logit))
> summary(m1)

Call:
glm(formula = bandera ~ tabla1$s151 + tabla1$s176 + tabla1$s189 +
tabla1$s194 + tabla1$s202 + tabla1$s210 + tabla1$s219 + tabla1$s224 +
tabla1$s247 + tabla1$s911, family = binomial(link = logit))

Deviance Residuals:
Min       1Q   Median       3Q      Max
-3.9625   0.0324   0.0425   0.0584   2.0827

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -42.655419   4.373896  -9.752  < 2e-16 ***
tabla1$s151   0.358480   0.080902   4.431 9.38e-06 ***
tabla1$s176   0.049998   0.003418  14.628  < 2e-16 ***
tabla1$s189   0.024010   0.005566   4.313 1.61e-05 ***
tabla1$s194  -0.029219   0.007015  -4.165 3.11e-05 ***
tabla1$s202   0.051615   0.006837   7.550 4.37e-14 ***
tabla1$s210   0.061870   0.002841  21.779  < 2e-16 ***
tabla1$s219   0.049157   0.008434   5.828 5.60e-09 ***
tabla1$s224   0.014293   0.003253   4.394 1.11e-05 ***
tabla1$s247   0.137596   0.005775  23.828  < 2e-16 ***
tabla1$s911   0.022470   0.009728   2.310   0.0209 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 18914.9  on 101339  degrees of freedom
Residual deviance:  9249.4  on 101329  degrees of freedom
AIC: 9271.4

Number of Fisher Scoring iterations: 9
```

En la regresión logística se observa que todas las variables son significativas (p-valores$<$0.05), y todos los coeficientes son positivos y por lo tanto favorables al éxito 'el cliente es bueno', salvo la característica 194 (que tiene coeficiente negativo). Los intervalos de confianza para esta ventaja de ser bueno al 95 % se muestran a continuación:

```
> confint(m1)
Waiting for profiling to be done...
                   2.5 %       97.5 %
(Intercept) -51.217917908 -34.06130355
tabla1$s151   0.199915906   0.51729311
tabla1$s176   0.043312068   0.05671221
tabla1$s189   0.013124981   0.03494855
tabla1$s194  -0.042982269  -0.01548141
tabla1$s202   0.038272979   0.06507695
tabla1$s210   0.056343019   0.06748029
tabla1$s219   0.032259930   0.06535849
tabla1$s224   0.007918069   0.02067098
tabla1$s247   0.126366544   0.14901878
tabla1$s911   0.003770646   0.04194090

> round(exp(cbind(coef(m1), confint(m1))), digits=6)
Waiting for profiling to be done...
                       2.5 %   97.5 %
(Intercept) 0.000000 0.000000 0.000000
tabla1$s151 1.431153 1.221300 1.677481
tabla1$s176 1.051269 1.044264 1.058351
tabla1$s189 1.024300 1.013211 1.035566
tabla1$s194 0.971204 0.957928 0.984638
tabla1$s202 1.052970 1.039015 1.067241
tabla1$s210 1.063824 1.057961 1.069809
tabla1$s219 1.050385 1.032786 1.067542
tabla1$s224 1.014396 1.007949 1.020886
tabla1$s247 1.147512 1.134698 1.160695
tabla1$s911 1.022725 1.003778 1.042833
```

La ventaja de ser bueno se multiplica por 1.434453 por cada unidad que aumenta el score de la característica 151, siendo IC= (1.221300; 1.677481) un intervalo de confianza al 95 % para esa ventaja. El mismo razonamiento se haría para el resto de las características salvo para la característica 194 que no es favorable al éxito.

```
> t1<-table(bandera, round(m1$fitted, 0))
> addmargins(t1)

bandera      0      1    Sum
```

```
0           344   1561   1905
1           262  99173  99435
Sum         606 100734 101340

> prop.table(t1,1)

bandera           0           1
0           0.180577428 0.819422572
1           0.002634887 0.997365113

> (344+99176)/101340
[1] 0.9820407
```

El porcentaje de clasificación correcta del modelo es del 98.20 %, aunque en el grupo de score que son malos clientes solo clasifica bien el 18.05 %. Hay una tasa bastante elevada de clientes malos que son clasificados como buenos. Sin embargo, el modelo clasifica bien al 98.20 % de los clientes. Esto es debido a que la mayoría de los clientes son buenos y clasificando bien a estos ya se obtiene un alto porcentaje de clasificación del modelo.

Al tener una muestra tan descompensada en cuanto al tamaño muestral de morosos y no morosos, es lógico que acaben teniendo más importancia los no morosos y por eso el modelo clasifica muy bien los no morosos, pero bastante peor los morosos. Para ello, se selecciona una submuestra aleatoria de los no morosos para que esos representen el mismo tamaño muestral que los morosos, ya que así, ambos tipos de errores tendrán igual importancia.

```
> library(sampling)
> set.seed(1)
> estratos<-strata(tabla1, stratanames = c("band_rend_30_2"), size = c(1500,1500),
method="srswor")
> tabla1.muestreado<-getdata(tabla1, estratos)

> table(tabla1.muestreado$band_rend_30_2)
0    1
1500 1500
```

Se prueba con el mismo modelo para esta submuestra, la cual tiene 1500 clientes morosos y 1500 clientes no morosos y se comprueba si ahora todas las variables siguen siendo significativas o si por el contrario hay que eliminar alguna del modelo. Puesto que se trabaja con submuestras aleatorias, es necesario fijar la semilla para trabajar siempre con los mismos datos.

```
> bandera.muestreado=tabla1.muestreado$band_rend_30_2
> m2<-glm(bandera.muestreado~tabla1.muestreado$s151
+tabla1.muestreado$s176+tabla1.muestreado$s189
+tabla1.muestreado$s194+tabla1.muestreado$s202
+tabla1.muestreado$s210+tabla1.muestreado$s219
+tabla1.muestreado$s224+tabla1.muestreado$s247
+tabla1.muestreado$s911,family=binomial(link=logit))
> summary(m2)

Call:
glm(formula = bandera.muestreado ~ tabla1.muestreado$s151 + tabla1.muestreado$s176 +
tabla1.muestreado$s189 + tabla1.muestreado$s194 + tabla1.muestreado$s202 +
```

```
tabla1.muestreado$s210 + tabla1.muestreado$s219 + tabla1.muestreado$s224 +
tabla1.muestreado$s247 + tabla1.muestreado$s911, family = binomial(link = logit))

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.7580  -0.2264   0.0569   0.2946   3.1793

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -38.117452   8.657415  -4.403 1.07e-05 ***
tabla1.muestreado$s151   0.184171   0.153958   1.196 0.231604
tabla1.muestreado$s176   0.005253   0.009392   0.559 0.575970
tabla1.muestreado$s189   0.019941   0.014201   1.404 0.160253
tabla1.muestreado$s194   0.024580   0.015282   1.608 0.107747
tabla1.muestreado$s202   0.066231   0.019099   3.468 0.000525 ***
tabla1.muestreado$s210   0.079271   0.009600   8.258  < 2e-16 ***
tabla1.muestreado$s219   0.035924   0.012998   2.764 0.005713 **
tabla1.muestreado$s224   0.050620   0.009515   5.320 1.04e-07 ***
tabla1.muestreado$s247   0.106229   0.010245  10.369  < 2e-16 ***
tabla1.muestreado$s911   0.022664   0.019762   1.147 0.251443
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4158.9  on 2999  degrees of freedom
Residual deviance: 1314.5  on 2989  degrees of freedom
AIC: 1336.5

Number of Fisher Scoring iterations: 6
```

Se elimina manualmente hacia atrás con el test de razón de verosimilitudes aquella variable con coeficiente no negativo, cuyo p-valor se aproxime más a uno. En este caso se elimina del modelo la característica 176. A continuación, se realiza un test anova para estudiar si es preferible el modelo m2 con la variable 176 o el modelo m3 sin ella.

```
> m3<-glm(bandera.muestreado~tabla1.muestreado$s151
+tabla1.muestreado$s189+tabla1.muestreado$s194
+tabla1.muestreado$s202+tabla1.muestreado$s210
+tabla1.muestreado$s219+tabla1.muestreado$s224
+tabla1.muestreado$s247+tabla1.muestreado$s911,
family=binomial(link=logit))

> anova(m3,m2,test="Chisq")
Analysis of Deviance Table

Model 1: bandera.muestreado ~ tabla1.muestreado$s151 + tabla1.muestreado$s189 +
tabla1.muestreado$s194 + tabla1.muestreado$s202 + tabla1.muestreado$s210 +
tabla1.muestreado$s219 + tabla1.muestreado$s224 + tabla1.muestreado$s247 +
tabla1.muestreado$s911
Model 2: bandera.muestreado ~ tabla1.muestreado$s151 + tabla1.muestreado$s176 +
tabla1.muestreado$s189 + tabla1.muestreado$s194 + tabla1.muestreado$s202 +
```

```
tabla1.muestreado$s210 + tabla1.muestreado$s219 + tabla1.muestreado$s224 +
tabla1.muestreado$s247 + tabla1.muestreado$s911
      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2990     1314.8
2      2989     1314.5  1  0.31217   0.5764
```

El p-valor es mayor que 0.05, por lo tanto, es preferible el modelo m3 al m2. Repitiendo el proceso con el modelo m3 se van eliminando sucesivamente aquellas variables que no son significativamente distintas de cero.

```
> m4<-glm(bandera.muestreado~tabla1.muestreado$s151
+tabla1.muestreado$s189+tabla1.muestreado$s194
+tabla1.muestreado$s202+tabla1.muestreado$s210
+tabla1.muestreado$s219+tabla1.muestreado$s224
+tabla1.muestreado$s247,family=binomial(link=logit))

> m5<-glm(bandera.muestreado~tabla1.muestreado$s189
+         +tabla1.muestreado$s194+tabla1.muestreado$s202
+         +tabla1.muestreado$s210+tabla1.muestreado$s219+
+           tabla1.muestreado$s224+tabla1.muestreado$s247,
+         family=binomial(link=logit))

> m6<-glm(bandera.muestreado~tabla1.muestreado$s194+
tabla1.muestreado$s202+tabla1.muestreado$s210+tabla1.muestreado$s219+
+tabla1.muestreado$s224+tabla1.muestreado$s247,
family=binomial(link=logit))

> summary(m6)

Call:
glm(formula = bandera.muestreado ~ tabla1.muestreado$s194 + tabla1.muestreado$s202 +
tabla1.muestreado$s210 + tabla1.muestreado$s219 + tabla1.muestreado$s224 +
tabla1.muestreado$s247, family = binomial(link = logit))

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.6831  -0.2200   0.0669   0.2709   3.1643

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -25.672612   1.284315 -19.989  < 2e-16 ***
tabla1.muestreado$s194   0.035609   0.013788   2.583 0.009806 **
tabla1.muestreado$s202   0.065202   0.018935   3.444 0.000574 ***
tabla1.muestreado$s210   0.078996   0.009534   8.285  < 2e-16 ***
tabla1.muestreado$s219   0.040198   0.010872   3.697 0.000218 ***
tabla1.muestreado$s224   0.051903   0.009454   5.490 4.02e-08 ***
tabla1.muestreado$s247   0.108424   0.010104  10.731  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

```
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4158.9  on 2999  degrees of freedom
Residual deviance: 1320.2  on 2993  degrees of freedom
AIC: 1334.2

Number of Fisher Scoring iterations: 6
> exp(cbind(coef(m6),confint(m6)))
Waiting for profiling to be done...
2.5 %       97.5 %
(Intercept)            7.088045e-12 5.447618e-13 8.420500e-11
tabla1.muestreado$s194 1.036251e+00 1.008510e+00 1.064578e+00
tabla1.muestreado$s202 1.067375e+00 1.028616e+00 1.107974e+00
tabla1.muestreado$s210 1.082200e+00 1.062285e+00 1.102814e+00
tabla1.muestreado$s219 1.041017e+00 1.018949e+00 1.063371e+00
tabla1.muestreado$s224 1.053274e+00 1.034139e+00 1.073235e+00
tabla1.muestreado$s247 1.114521e+00 1.092942e+00 1.137145e+00
```

Ahora todas las variables son significativas. Por lo tanto, es mejor el modelo que contiene las características 194, 202, 210, 219, 224 y 247. El aumento de una unidad de la variable 194 supone que la ventaja de no ser moroso se multiplique por 1.036251, el aumento de una unidad de la variable 202 hace que esa ventaja se multiplique por 1.067375 y así con el resto de las variables.

```
> t6<-table(bandera.muestreado, round(m6$fitted, 0))
> addmargins(t6)

bandera.muestreado    0    1  Sum
0                  1379  121 1500
1                    92 1408 1500
Sum                1471 1529 3000

> prop.table(t6,1)

bandera.muestreado     0         1
0                  0.91933333  0.08066667
1                  0.06133333  0.93866667

> (1379+1408)/3000
[1] 0.929
```

Este modelo m6, que contiene las características 194, 202, 210, 219, 224 y 247 tiene un AIC de 1334, clasifica bien al 91.93 % de los clientes malos y al 93.86 % de los clientes buenos. Clasifica bien en total al 92.90 % de la población. Por lo tanto, la clasificación de los clientes malos mejora bastante y el modelo tiene un buen poder discriminante.

A continuación, se emplea la técnica stepwise pero con el comando de R *step* y se comprueba si obtiene el mismo modelo.

```
> step(m2)
```

En este caso, la función *step* elige el modelo m5 que considera las características 189, 194, 202, 210, 219, 224 y 247. Es decir, las mismas características que el modelo m6 y adicionalmente considera la 189.

```
> t5<-table(bandera.muestreado, round(m5$fitted, 0))
> addmargins(t5)

bandera.muestreado    0    1  Sum
0                  1378  122 1500
1                   95 1405  1500
Sum                1473 1527 3000

> prop.table(t5,1)

bandera.muestreado          0          1
0                     0.91866667 0.08133333
1                     0.06333333 0.93666667

> (1378+1405)/3000
[1] 0.9276667
```

Este modelo tiene unos porcentajes de clasificación ligeramente menores a los del modelo m6. Pero aun así tiene un poder discriminante muy elevado.

Ahora se comprueba si hay un modelo que sea mejor en términos del AIC. Un modelo será mejor cuanto mayor sea su verosimilitud y menor su número de parámetros, de forma que entre dos modelos será preferible el de menor AIC. Para ello, se utiliza el paquete de R *bestglm*.

```
> library(bestglm)
Loading required package: leaps
> a=tabla1.muestreado[, 4:13]
> b=tabla1.muestreado[, 29]
> union=data.frame(a,b)
> sub.Aic=bestglm(union, IC="AIC", family=binomial)
Morgan-Tatar search since family is non-gaussian.
> sub.Aic$BestModels
   s151  s176  s189 s194 s202 s210 s219 s224 s247  s911 Criterion
1 FALSE FALSE  TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  1332.135
2 FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  1332.160
3  TRUE FALSE  TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  1332.190
4 FALSE FALSE  TRUE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE  1332.445
5 FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE  1332.503
```

De nuevo, vuelve a salir como mejor modelo en cuanto al AIC el modelo m5, aunque utilizando esta función de R se obtiene un AIC ligeramente menor. El segundo modelo que propone R es el modelo m6. A continuación, se considera el criterio de información de Bayes (BIC) y se ven cuales serían los mejores modelos según este criterio. Este criterio tiene una penalización más fuerte al número de variables. Al igual que con el criterio Akaike se preferirá el modelo con menor BIC.

```
> sub.Bic=bestglm(union, IC="BIC", family=binomial)
Morgan-Tatar search since family is non-gaussian.
> sub.Bic$BestModels
   s151  s176  s189  s194  s202 s210 s219 s224 s247  s911 Criterion
1 FALSE FALSE FALSE FALSE  TRUE TRUE TRUE TRUE TRUE FALSE  1366.785
2 FALSE FALSE FALSE  TRUE  TRUE TRUE TRUE TRUE TRUE FALSE  1368.198
3 FALSE FALSE  TRUE FALSE  TRUE TRUE TRUE TRUE TRUE FALSE  1369.748
```

```
4 FALSE FALSE FALSE FALSE  TRUE TRUE TRUE TRUE TRUE  TRUE  1371.110
5 FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE  1371.408
```

Según este criterio el mejor modelo es aquel que tiene las características 202, 210, 219, 224 y 247 y el siguiente modelo añadiría además la 194, por lo que coincide con el modelo m6.

```
> m7<-glm(bandera.muestreado~tabla1.muestreado$s202
+tabla1.muestreado$s210+tabla1.muestreado$s219+
+tabla1.muestreado$s224+tabla1.muestreado$s247,
family=binomial(link=logit))

> t7<-table(bandera.muestreado, round(m7$fitted, 0))
> addmargins(t7)

bandera.muestreado    0    1  Sum
0                  1379  121 1500
1                    91 1409 1500
Sum                1470 1530 3000
> prop.table(t7,1)

bandera.muestreado          0          1
0                  0.91933333 0.08066667
1                  0.06066667 0.93933333
> (1379+1409)/3000
[1] 0.9293333
```

Este modelo tiene también porcentajes de clasificación similares a los anteriores, pero tiene un mayor AIC. A la vista de las tablas de clasificación y de los valores de AIC y BIC, puesto que son todos bastante similares, se seleccionan los modelos m6 o el m7, ya que son más sencillos que el m5 (tienen menos variables) y, en consecuencia, más fáciles de interpretar. Para decantarse entre el m6 o el m7 se puede calcular el $R^2 = 1 - \dfrac{D_m}{D_0}$.

```
> 1-m6$deviance/m6$null.deviance
[1] 0.6825686
> 1-m7$deviance/m7$null.deviance
[1] 0.6809832
```

El modelo m6 tiene una capacidad explicativa mayor que el m7, $R^2 = 68.26\,\%$ con un porcentaje global de clasificación $92.90\,\%$ ($91.93\,\%$ y $93.86\,\%$ en cada grupo). Por lo tanto, se elige el modelo m6 que contiene las características 194, 202, 210, 219, 224 y 247.

Si ahora se selecciona una submuestra aleatoria de manera que se tomen los tamaños submuestrales lo más grandes posibles para que se cumpla que el cociente entre morosos y no morosos sea 1. Para ello, se toma una submuestra de 1905 morosos (es decir, todos los morosos) y solo 1905 no morosos. Se repite el proceso unas 10 veces y se observan cuales son los 3 mejores modelos según los criterios AIC y BIC se obtienen los siguientes resultados.

Los modelos que más se repiten como mejores modelos según el criterio AIC son:

- Aquel en el que se elimina la variable 176.
- Aquel en el que se eliminan las variables 176, 219 y 911.

- Aquel en el que se eliminan las variables 176 y 911.

Se calcula el AUC para cada uno de estos mejores modelos:

- En la Figura 7.5 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar la variable 176, es decir, el modelo que cuenta con las variables 151, 189, 194, 202, 210, 219, 224, 247 y 911.

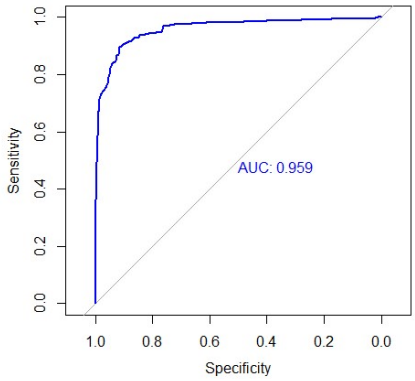

Figura 7.5: Curva ROC y valor de AUC para el modelo en el que se elimina la variable 176.

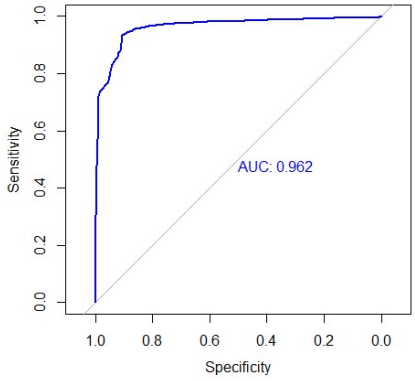

Figura 7.6: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 219 y 911.

- En la Figura 7.6 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar las variable 176, 219 y 911. Es decir, el modelo que cuenta con las variables 151, 189, 194, 202, 210, 224 y 247.

- En la Figura 7.7 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar las variables 176 y 911, es decir, el modelo que cuenta con las variables 151, 189, 194, 202, 210, 219, 224 y 247.

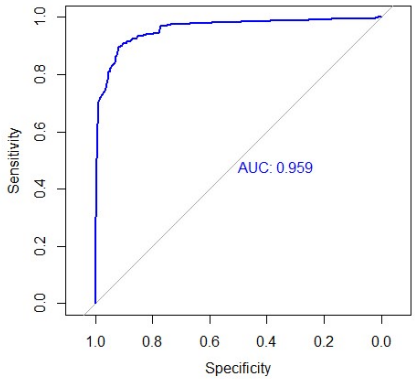

Figura 7.7: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176 y 911.

No hay mucha diferencia entre los valores del AUC en los tres modelos, siendo ligeramente mayor el del modelo que no considera las características 176, 219 y 911.

Los mejores modelos que más se repiten según el criterio BIC son:

- Aquel en el que se eliminan las variables 176, 189, 219 y 911.
- Aquel en el que se eliminan las variables 151, 176, 189 y 911.
- Aquel en el que se eliminan las variables 176, 219 y 911.

Si calculamos el AUC de estos modelos:

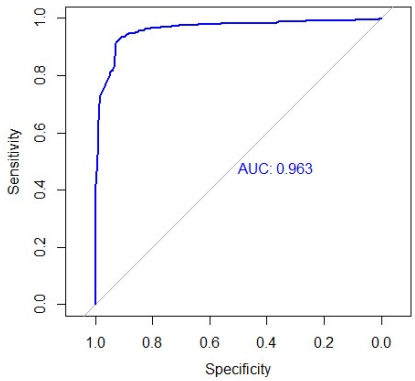

Figura 7.8: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 189, 219 y 911.

- En la Figura 7.8 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar las variables 176, 189, 219 y 911, es decir, el modelo que cuenta con las variables 151, 194, 202, 210, 224 y 247.
- En la Figura 7.9 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar las variables 151, 176, 189 y 911, es decir, el modelo que cuenta con las variables 194, 202, 210, 219, 224, 247.

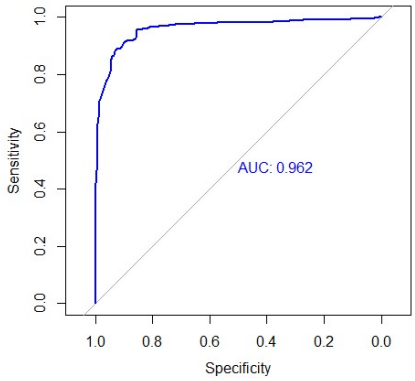

Figura 7.9: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 151, 176, 189 y 911.

- En la Figura 7.10 puede verse la curva ROC y el valor del AUC del modelo que resulta al eliminar las variable 176, 219 y 911. Es decir, el modelo que cuenta con las variables 151, 189, 194, 202, 210, 224 y 247.

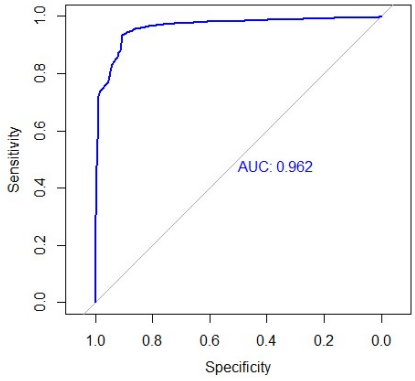

Figura 7.10: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 219 y 911.

Los valores del AUC son muy similares para estos modelos. Por lo tanto, teniendo en cuenta ambos criterios y observando los valores del AUC, un buen modelo podría ser aquel que elimina las variables 176, 219 y 911; ya que está entre los mejores según el criterio AIC y según el criterio BIC. Entonces, se concluye que para los clientes que solo tienen cuenta corriente, un modelo que tiene una alta capacidad predictiva es el que cuenta con las características 151, 189, 194, 202, 210, 224 y 247.

Las características de este modelo presentan unos valores aceptables de IV y tienen sentido económico. El valor de AUC para este modelo es de 0.962 y el índice de Gini es 0.924.

7.2. Tabla cuenta corriente y producto de riesgo o Tabla 2

Para los clientes que tienen cuenta corriente y producto de riesgo (tabla 2) se programan las características 130, 151, 176, 189, 202, 210, 224, 247, 312, 578, 707 y 807. Estas características no son

descritas por motivos de confidencialidad, puesto que forman parte de los modelos de la entidad.

Al igual que se hace en la tabla 1, se obtiene una puntuación para cada característica y la suma de estas puntuaciones conforma el score final para cada cliente.

Se calculan los valores de IV para las características y se eliminan del modelo aquellas características que tienen un IV menor que 0.02. En este caso todas las variables tienen valores de IV aceptables por lo que no es necesario eliminarlas del modelo.

A continuación, se hace un análisis exploratorio de los datos, así como la representación de la tasa de mora y de los valores de WOE.

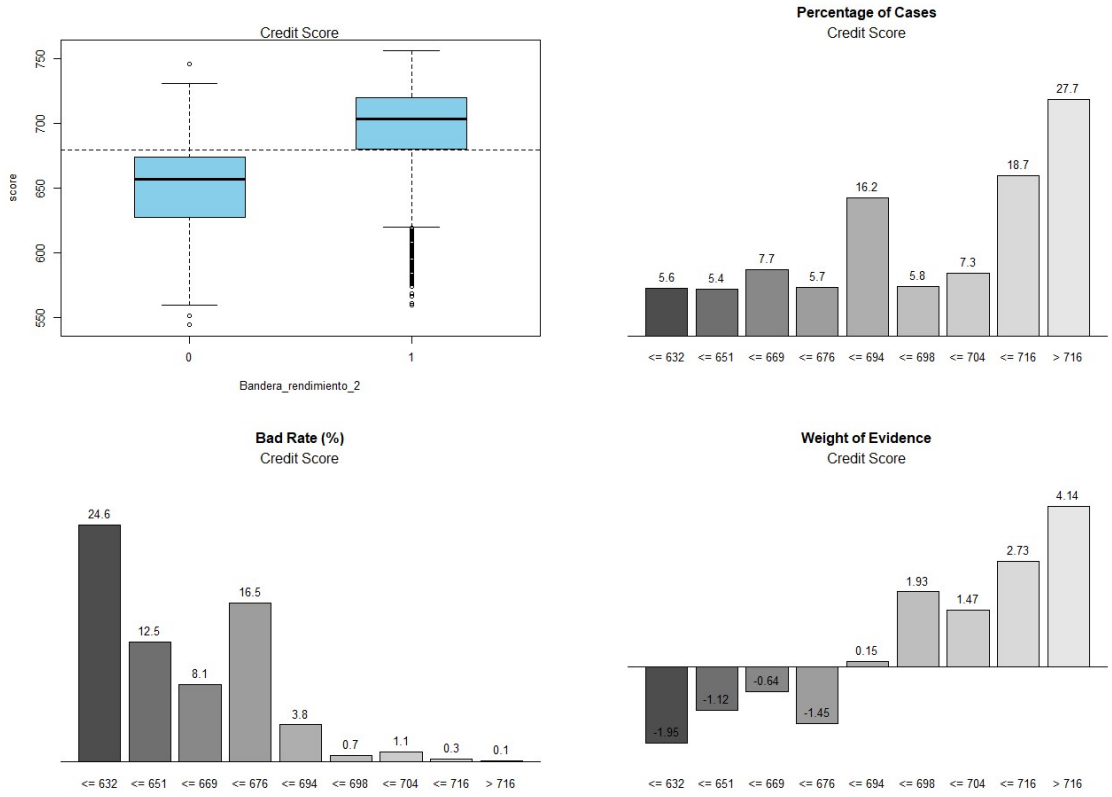

Figura 7.11: Diagrama de cajas, porcentajes de las puntuaciones, tasas de mora y tendencia de los valores de WOE asociados al modelo de la tabla 2.

Para la tabla 2, en la Figura 7.11 en el primer gráfico se muestran las distribuciones de los clientes en función de su bandera de rendimiento (0 si el cliente es malo y 1 si el cliente es bueno). Se obtiene que los clientes con bandera de rendimiento 0 (malos) obtienen puntuaciones más bajas (inferiores a 679.50), y presenta un dato atípico de un cliente moroso que puntúa alto (cerca de 750). En cambio, los clientes con bandera de rendimiento 1 (buenos) puntúan más alto, (puntuaciones superiores a 679.50), con algún dato atípico de clientes que son buenos y tienen una puntuación cercana o inferior a 600. Se puede concluir también que el modelo diferencia bastante bien entre buenos y malos. Esto también se ve con el valor del $AUC = 0.859$, que se puede considerar alto. Si se agrupa el score mediante el paquete de R *smbinning* se obtiene un $IV = 2.7782$ y los puntos de corte para las puntuaciones que se obtienen son 632, 651, 669, 676, 698, 704, 716. En el segundo gráfico se representan los porcentajes

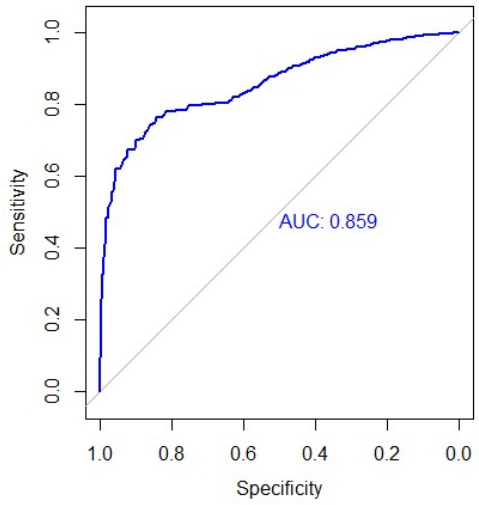

Figura 7.12: Representación de la curva ROC para la tabla 2.

de clientes que se encuentran las diferentes categorías. En todas las clases cae al menos un 5 % de la población por lo que son representativas. En el tercer gráfico se representa la tasa de mora. Se puede ver que la tasa de mora es mayor en las puntuaciones más bajas y va disminuyendo a medida que estas puntuaciones toman valores más altos. Sin embargo, hay un repunte en el intervalo (669, 676] y otro más ligero en el intervalo (698, 704]. Por último, se representa el WOE de las categorías. Para las puntuaciones más bajas el WOE es negativo y va creciendo a medida que aumenta el score. Igual que sucede con las tasas de malos, hay una bajada del WOE en los intervalos (669, 676] y (698, 704].

	Febrero	Agosto
AUC	0.8503179	0.8450490
Gini	0.7006359	0.6900981

Cuadro 7.3: Valores de AUC e índice de Gini para las dos muestras de la tabla 2.

Los valores de AUC y Gini son algo menores para el modelo de la Tabla 2 pero siguen teniendo una buena capacidad predictiva.

Considerando una submuestra aleatoria para la Tabla 2 de modo que haya el máximo número de morosos y que el cociente entre el número de morosos y no morosos sea 1 y repitiendo el proceso 10 veces, se seleccionan los 3 mejores modelos según los criterios AIC y BIC se obtienen los siguientes resultados.

Los modelos que más se repiten como mejores modelos según el criterio AIC y sus curvas ROC asociadas son:

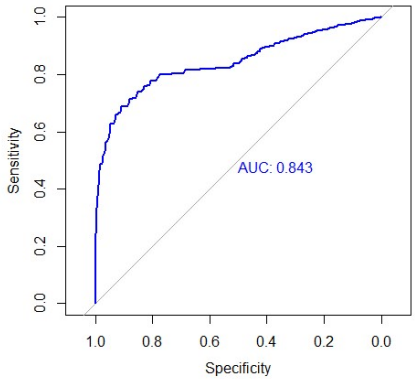

Figura 7.13: Curva ROC y valor de AUC para el modelo en el que se eliminan la variable 202.

- El modelo que resulta al eliminar la variable 202, es decir, el modelo que cuenta con las variables 130, 151, 176, 189, 210, 224, 247, 312, 578, 707 y 807. En la Figura 7.13 pueden verse la curva ROC y el valor del AUC correspondientes.

- El modelo que resulta al eliminar las variables 202 y 176, es decir, el modelo que cuenta con las variables 130, 151, 189, 210, 224, 247, 312, 578, 707 y 807. En la Figura 7.14 pueden verse la curva ROC y el valor del AUC correspondientes.

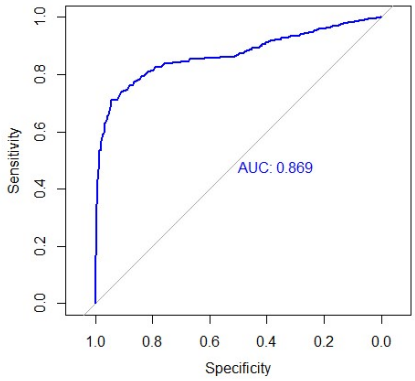

Figura 7.14: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176 y 202.

- El modelo que resulta al eliminar las variables 202 y 578, es decir, el modelo que cuenta con las variables 130, 151, 176, 189, 210, 224, 247, 312, 707 y 807. En la Figura 7.15 pueden verse la curva ROC y el valor del AUC correspondientes.

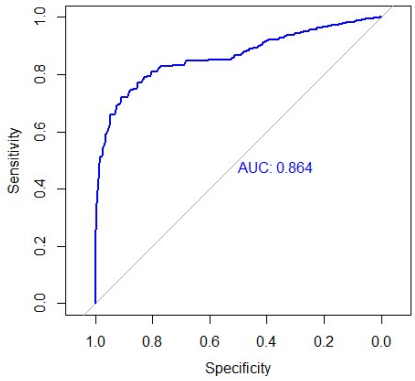

Figura 7.15: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 202 y 578.

- El modelo que resulta al eliminar la variable 247, es decir, el modelo que cuenta con las variables 130, 151, 176, 189, 202, 210, 224, 312, 707 y 807. En la Figura 7.16 pueden verse la curva ROC y el valor del AUC correspondientes.

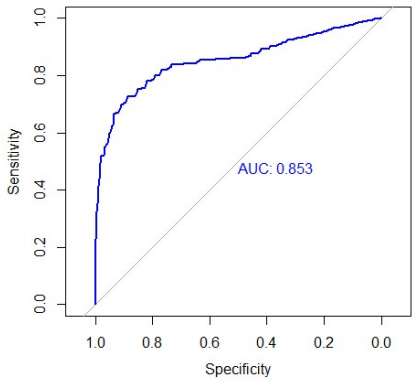

Figura 7.16: Curva ROC y valor de AUC para el modelo en el que se eliminan la variable 247.

A la vista de los valores AUC, el mejor modelo es aquel que elimina las variables 176 y 202.

Los mejores modelos que más se repiten según el criterio BIC y sus curvas ROC asociadas son:

- El modelo que resulta al eliminar las variables 176, 202, 247 y 578, es decir, el modelo que cuenta con las variables 130, 151, 189, 210, 224, 312, 707 y 807. En la Figura 7.17 pueden verse la curva ROC y el valor del AUC correspondientes.

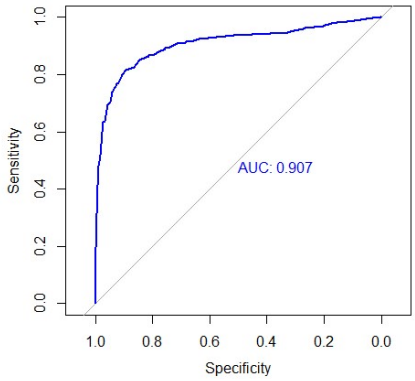

Figura 7.17: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 202, 247 y 578.

- El modelo que resulta al eliminar las variables 176, 202 y 578, es decir, el modelo que cuenta con las variables 130, 151, 189, 210, 224, 247, 312, 707 y 807. En la Figura 7.18 pueden verse la curva ROC y el valor del AUC correspondientes.

- El modelo que resulta al eliminar las variables 202 y 176, es decir, el modelo que cuenta con las variables 130, 151, 189, 210, 224, 247, 312, 578, 707 y 807. En la Figura 7.19 pueden verse la curva ROC y el valor del AUC correspondientes.

- El modelo que resulta al eliminar las variables 176, 202, 312 y 578, es decir, el modelo que cuenta con las variables 130, 151, 189, 210, 224, 247, 707 y 807. En la Figura 7.20 pueden verse la curva ROC y el valor del AUC correspondientes.

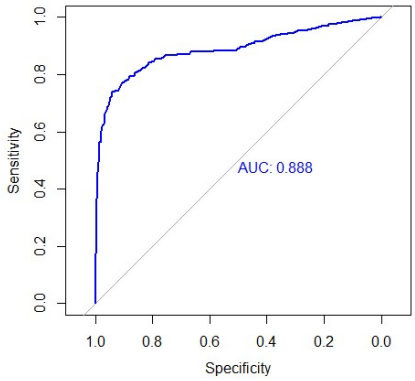

Figura 7.18: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 202 y 578.

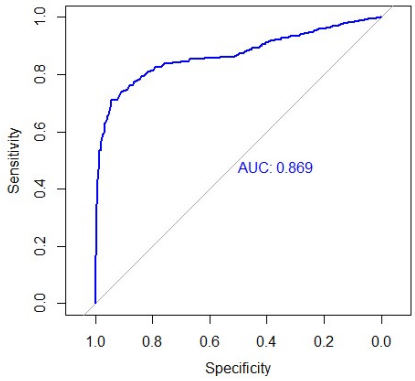

Figura 7.19: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176 y 202.

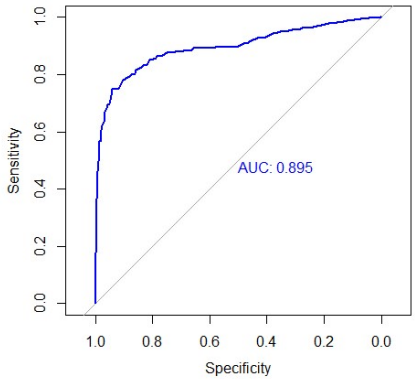

Figura 7.20: Curva ROC y valor de AUC para el modelo en el que se eliminan las variables 176, 202, 312 y 578.

El modelo que mejor AUC tiene es el que elimina las variables 176, 202, 247 y 578. Si se tienen en cuenta ambos criterios (AIC y BIC), un buen modelo podría ser aquel que elimina las variables 176 y 202. Como tiene mejor AUC el primero y además tiene dos variables menos y por lo tanto es más fácil de interpretar finalmente se selecciona este modelo que cuenta con las características 130, 151, 189, 210, 224, 312, 707 y 807.

El valor del AUC de este modelo es de 0.907 que se muestra en la Figura 7.17 y el índice de Gini es 0.814.

Adicionalmente, para construir estos modelos también se han programado otras características que no han sido incluídas, puesto que o bien los valores de IV son poco predictivos o bien no tienen sentido económico, ya que presentan tasas de mora más elevadas en clientes con puntuaciones más altas.

Capítulo 8

Conclusiones

Tras haber realizado una revisión metodológica de las técnicas que se utilizan generalmente para construir un modelo de scoring, se han aplicado dichas técnicas a un conjunto de datos reales. Se usa el esquema de datos que se obtienen en la descarga de detalle de movimientos a través de la banca electrónica. Una primera fase del trabajo consiste en el tratamiento y agregación a nivel cliente de dicha información para la construcción de una serie de variables o características candidatas a formar parte del modelo. Asimismo, se construye una variable de comportamiento de pago o variable respuesta que determina, durante los 12 meses siguientes a la fecha de observación, si el cliente resultó malo o bueno en términos de impago o necesidad de refinanciación.

Desde un punto de vista de negocio, este modelo serviría para poder evaluar a un no cliente utilizando la información de movimientos del banco en el que sí tenga vinculación, previa autorización del cliente para descargar el extracto de un periodo temporal a partir de la banca electrónica de su entidad de referencia. Esta descarga se puede hacer mediante los APIs de ciertas compañías que prestan el servicio de descarga de datos de la banca electrónica de entidades financieras.

Los resultados ponen de manifiesto que la información de movimientos en cuenta y comportamiento de pago en productos de riesgo como tarjetas de crédito o préstamos resulta muy predictiva para medir el riesgo de crédito.

Se determina que es mejor construir dos modelos diferentes en función de que el cliente tenga o no algún producto de riesgo contratado. Se obtiene un modelo de scoring con una capacidad predictiva elevada, que se refleja en las curvas ROC, así como en los valores de WOE e IV. Por lo tanto, este modelo permite discriminar a los no clientes morosos de los no morosos descargando los datos de su banca electrónica. Esto es una ventaja competitiva para otorgar financiación a nuevos clientes usando su información de movimientos de sus extractos con su entidad de referencia, previa autorización del cliente para descargar y hacer uso de esa información por parte de la que no es su entidad de referencia..

Un futuro desarrollo sería construir una tabla de puntuaciones para el modelo. Sería interesante continuar desarrollando este proyecto comprobando la adecuación del modelo utilizando datos del extracto bancario descargados de la banca electrónica de otras entidades que no fuesen ABANCA. También se podría ampliar el proyecto traduciendo el score a una probabilidad de mora o default PD, para obtener la probabilidad de que el cliente impague. Así, dicha PD se podría utilizar para establecer un punto de corte que determinase la concesión o no de crédito en función del apetito al riesgo de la entidad bancaria. Esta PD se puede complementar con una medición de la capacidad de pago mensual del cliente según los movimientos al debe y al haber registrados en sus cuentas. De esta manera, además de cuantificar la calidad crediticia del cliente a partir del modelo de scoring objeto de este trabajo, se podría llegar a determinar el importe a conceder a cada cliente.

Apéndice A

Resumen de acrónimos

AIC: Criterio de información de Akaike (Akaike information criterion)
AUC: Área bajo la curva ROC (Area under the curve)
BIC: Criterio de información de Bayes (Bayesian information criterion)
CART: Árboles de clasificación y regresión (classification and regression trees)
EAD: Exposición en caso de mora (Exposure at default)
IRB: Rating interno basado en la calificación (Internal rating-based aproach)
IV: Valor de información (Information value)
LGD: Pérdida de mora o incumplimiento (Loss given default)
PD: Probabilidad de mora o default (Probability of default)
ROC: Curva característica operativa del receptor (Receiver operating characteristic)
WOE: Peso de la evidencia (Weight of evidence)

Bibliografía

[1] Crawley MJ (2013). The R book. John Wiley & Sons. Segunda edición. England, UK.

[2] D'Agostino RB, Stephens MA, eds. (1986). Goodness-of-Fit techniques. Marcel Dekker, Inc. New York and Basel.

[3] Devia A (2015) Contribuciones al análisis estadístico de riesgo de crédito. Tesis, Universidade da Coruña.

[4] Efroymson, MA (1960) Multiple regression analysis, Mathematical Methods for Digital Computers, Ralston A. and Wilf,H S, (eds.), Wiley, New York.

[5] Franco N, Vivo Molina J.M (2013), Análisis de curvas ROC: Principios básicos y aplicaciones. Cuadernos de Estadística. La Muralla, S.A.

[6] Guisande C. Vaamonde A. (2013) Gráficos estadísticos y mapas con R. Díaz de Santos. Vigo.

[7] Hosmer DW, Lemeshow S (1989). Applied logistic regression. John Wiley & Sons. New York.

[8] Hothorn T, Hornik , Zeileis A (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674.

[9] James G, Witten D, Hastie T, Tibshirani R (2013). An Introduction to Statistical Learning with Applications in R. Springer. New York.

[10] Jopia H (2018). Scoring Modeling and Optimal Binning.

[11] Kim L (2016). Information: Data Exploration with Information Theory (Weight-of-Evidence and Information Value). R package version 0.0.9. https://CRAN.R-project.org/package=Information

[12] Mendoza JB (2018) https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

[13] Nieto S (2010) Crédito al consumo: La estadística aplicada a un problema de riesgo crediticio. Tesis, Universidad Autónoma Metropolitana. México.

[14] Ochirsukho M (2016) Application scorecard modelling: techniques and performance. Tesis, Vrije Universiteit Amsterdam.

[15] Ong MK (2005) The Basel Handbook. A guide for financial practitioners. Risk Books.

[16] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[17] Ripley B (2018). tree: Classification and Regression Trees. R package version 1.0-39. https://CRAN.R-project.org/package=tree

[18] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77.

[19] Siddiqi N (2006) Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring. John Wiley & Sons, Inc. New Yersey.

[20] Strasser H, Weber C (1999). On the Asymptotic Theory of Permutation Statistics. Mathematical Methods of Statistics, 8(2), 220-250. WU Vienna University of Economics and Business, Vienna.

[21] Therneau TM, Atkinson EJ, Founfation M (2018) An introduction to recursive partitioning using the rpart routines. http://www.cran.r-project.org/package=NPCirc.

[22] Therneau T, Atkinson B, Ripley B (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

[23] Torres A. (2010). Curvas ROC para Datos de Supervivencia. Aplicación a Datos Biomédicos. Trabajo final de master en Técnicas Estadísticas de la Universidad de Santiago de Compostela.