Cuadernos de Economía (2020) 43, 175-190

Asociación Cuadernos de economía

Cuadernos de economía

www.cude.es

ARTÍCULO

Tratamiento de clases desbalanceadas con el método del cubo en problemas de credit scoring a través de la minería de datos

Mauricio Beltrán Pascual[a], Francisco Javier Martínez de Pisón Ascacíbar[b] y Juan Antonio Vicente Vírseda[c]

[a] *Grupo EDMANS. Universidad de La Rioja. España*
[b] *Grupo EDMANS. Universidad de La Rioja. España*
[c] *Universidad Nacional de Educación a Distancia. España*

JEL CODES
C11

KEYWORDS:
Cube method; Credit scoring; Data mining; Classification cost

Abstract: This article discusses how to apply the sampling method called the "Cube Method" in credit scoring problems to improve the precision of the predictive models obtained. This method ensures an optimal balance of the samples when working with databases whose classes of the dependent variable are highly unbalanced. Using two samples of real banking data, we provide a comparative study of the best models obtained with different data mining methods when these are applied to the original databases and the balanced ones. Finally, we conclude that when the samples are balanced the predictive capacity of the classification algorithms is more precise and the models used reduce the economic cost of the classification.

CÓDIGOS JEL
C11

PALABRAS CLAVE:
Método del cubo; Credit scoring; Minería de datos; Coste de clasificación

Resumen: En este artículo se aborda la forma de aplicar el método de muestreo denominado "Método del cubo" en problemas de credit scoring con la finalidad de poder mejorar la precisión de los modelos predictivos que se obtengan. Este método permite garantizar un óptimo equilibrio de las muestras cuando se trabaja con bases de datos cuyas clases de la variable dependiente están altamente desbalanceadas. Utilizando dos muestras de datos bancarios reales, se realiza un estudio comparativo de los mejores modelos obtenidos con diversos métodos de minería de datos aplicados a las bases de datos originales frente a las balanceadas. Finalmente, se concluye que la capacidad predictiva de los algoritmos de clasificación es más precisa y que los modelos utilizados reducen el coste económico de la clasificación cuando se equilibran las muestras.

Correo electrónico: beltranpascual@gmail.com; javicente@cee.uned.es

https://doi.org/10.32826/cude.v42i122.137

1. Introducción

El sector bancario y en general toda la industria bancaria es, sin lugar a dudas, uno de los principales actores de la economía. La función de intermediación bancaria que realizan las instituciones financieras, entre otras actividades, la llevan a cabo a través de la inversión crediticia. Al conceder créditos, estas entidades están asumiendo riesgos y, si se quiere generar rentabilidad, tienen que gestionarlos adecuadamente.

Es obvia la necesidad de comprender y, por supuesto, de administrar los diferentes tipos de riesgo que surgen de la variabilidad de los diferentes resultados financieros. En Jorion (2000) se define el riesgo como la volatilidad de los resultados esperados, generalmente el valor de los activos o pasivos de interés. Atendiendo al tipo de factores que lo generan, podemos encontrar cuatro grandes grupos: riesgo de mercado, riesgo de crédito, riesgo de negocio o estratégico y riesgo operacional. A su vez, en el riesgo de crédito podemos identificar cuatro componentes: riesgo de default o de impago, riesgo de mercado, riesgo de liquidez y riesgo país.

Podemos definir el *credit scoring (CS)* como los métodos estadísticos utilizados para clasificar a los solicitantes de crédito, sean o no clientes de la entidad evaluadora, entre las clases de riesgo bueno o malo, Hand y Henley (1997). Por otro lado, Bessis (2002) define el riesgo de crédito como aquellas pérdidas asociadas al evento fallido del prestatario o al evento del deterioro de la calidad crediticia. El CS es, por tanto, un sistema o un método que, a través de predicciones, mide el riesgo inherente al mismo. Otros nombres con los que se conoce al CS son: calificación de riesgo de insolvencia o morosidad.

La cuantificación de la probabilidad de que ocurra una determinada solución es de vital importancia en la toma de decisiones económicas. En CS, tomar buenas decisiones determina la diferencia entre el éxito o el fracaso de la empresa, ya que la principal actividad de una entidad bancaria es ofrecer créditos a clientes y asegurarse de que éstos sean devueltos. Dado que las malas decisiones aumentan la posibilidad de quiebra de dicha entidad, la disponibilidad de un buen mecanismo que prediga la probabilidad de que un cliente devuelva un crédito es de vital importancia. A este respecto, en Caouette et al. (1998) se afirma que "El próximo gran reto de los mercados financieros es el desarrollo de nuevos métodos y técnicas para valorar el riesgo de crédito".

El CS lleva efectuándose varias décadas y se erige como una metodología ya plenamente aceptada por el Comité de Basilea para la supervisión bancaria y también por los sistemas financieros europeos y norteamericano, donde a través de un sistema de rating interno se clasifica a los clientes de la institución financiera como buenos o malos. A la hora de valorar el riesgo, encontramos que los determinantes del mismo son: la Probabilidad de incumplimiento (default), la Exposición y la Severidad o Tasa de recuperación.

El proceso de CS puede ser abordado mediante un proceso de minería de datos cuyo objetivo principal es el de descubrir estructuras subyacentes, potencialmente útiles para la toma de decisiones, que puedan encontrarse escondidas en las bases de datos. El conjunto de herramientas incorporadas en la minería de datos es muy extenso y se ha revelado muy eficiente. A nivel práctico, la minería de datos es un proceso interactivo e iterativo que combina la experiencia y conocimiento sobre un problema dado con la aplicación de técnicas tradicionales y avanzadas basadas en métodos de estadística, bases de datos, aprendizaje automático e inteligencia artificial.

Uno de los principales problemas a los que se enfrenta el CS es que los clasificadores predictivos conseguidos suelen predecir correctamente un porcentaje más elevado de la clase mayoritaria, pues la muestra de peticionarios que suelen devolver un crédito es mucho mayor que la de aquellos que no lo embolsan. Este problema, habitualmente denominado como "un caso con clases desbalanceadas", existe en muchos otros ámbitos, como por ejemplo: la detección de clientes que pueden abandonar una compañía, la detección de productos defectuosos en un proceso industrial, la identificación de comportamientos delictivos en la red, el uso fraudulento de tarjetas bancarias, etcétera; donde el número de clases de un tipo es mucho mayor que el del otro. En el caso del CS, si el desbalanceo es considerable, descubrir regularidades inherentes a la clase minoritaria se convierte en una tarea ardua y de poca fiabilidad, Japkowicz y Stephen (2002). En definitiva, el equilibrio entre las muestras de cada clase juega un papel determinante a la hora de desarrollar modelos eficientes.

El objetivo fundamental de este trabajo se ha centrado en la aplicación del "Método del Cubo" para el equilibrado de muestras desbalanceadas que permita el desarrollo de modelos predictivos eficientes en problemas de CS.

2. Equilibrado de la muestra

2.1. Introducción al balanceo de muestras

Cuando la base de datos con la que trabajamos tiene la variable dependiente desequilibrada, en nuestro caso representada por los clientes que devuelven el crédito y los morosos, los algoritmos de clasificación tienden a predecir de forma correcta un porcentaje más elevado de la clase dominante.

El tema de muestras desbalanceadas se ha tratado extensamente y se han utilizado muchas estrategias, aunque se puede afirmar que no existe una solución definitiva sobre cuál es la mejor. Hulse et al. (2007) concluyen que la decisión sobre la mejor técnica está influenciada en gran medida por la naturaleza del clasificador y la medida de efectividad. Otros autores como López et al. (2013) hacen una revisión de las principales cuestiones planteadas con conjuntos de datos desequilibrados y presentan los principales enfoques llevados a cabo por diversos investigadores para hacer frente a este problema desarrollando una profunda discusión sobre el efecto de las características intrínsecas que poseen las bases de datos. Estos investigadores llegan a la conclusión de que no sólo es la relación de desequilibrio la que tiene el efecto más significativo en el rendimiento de los clasificadores, sino que influyen otras características intrínsecas a los datos como la presencia de zonas con pequeñas desuniones, la falta de densidad y de

información en los datos de entrenamiento, el problema de la superposición de las clases, el impacto de los datos con ruido, el correcto tratamiento y gestión de los datos de la frontera y las posibles diferencias en la distribución de los datos de entrenamiento y prueba.

Las soluciones más usadas para tratar el desbalanceo se pueden encuadrar en dos grandes grupos: soluciones a nivel de datos y a nivel de algoritmos.

Las técnicas dirigidas a modificar los datos tratan de remuestrear la base de datos de entrenamiento, bien sea a través del sobremuestreo de la clase minoritaria o del submuestreo de la clase mayoritaria. Aunque estas técnicas han demostrado su efectividad, no dejan de tener ciertos inconvenientes, pues pueden eliminar ejemplos útiles, falsear el proceso de validación de los modelos, e incrementar los costes. Otra crítica a esta estrategia se refiere al cambio que se realiza en la distribución original del conjunto de entrenamiento de los datos, lo que dificulta enormemente el cálculo realista de predicciones basadas en probabilidad.

Cuando se plantea la reducción de la muestra de la clase mayoritaria, hemos de acudir a un método de submuestreo. Una de las primeras propuestas para editar o filtrar las muestras de entrenamiento fue el algoritmo de Edición de Wilson (1972), también conocido como la regla del vecino más cercano editado (*Edited Nearest Neighbor*). Actualmente, existen muchas formas de proceder, algunas de ellas son las siguientes:

- Submuestreo aleatorio: Esta forma de proceder ha sido utilizada por muchos investigadores.
- Submuestreo dirigido: algoritmo *One-sides selection* de Kubat y Matwin, (1997).
- Técnicas de vecindad: algoritmo *Neighbordhood Cleaning Rule* de Laurikkala, (2002).
- Submuestreo aplicando algoritmos genéticos: Kuncheva y Jain, 1999.
- Submuestreo por distancia: Zhanng y Mani, (2003).
- Submuestreo por clustering: Cohen *et al*. (2006).
- Aprendizaje activo de Provost, (2003).

Pero de entre todos los métodos existentes en la literatura estadística, el denominado "submuestreo equilibrado del Cubo", propuesto por Deville y Tillé (2004), es el único que nos permite seleccionar una muestra equilibrada sobre variables auxiliares con probabilidades de inclusión que pueden ser iguales o no. La característica principal del método del cubo es que selecciona únicamente las muestras cuyos estimadores de Horvitz-Thompson son iguales a los totales de las variables auxiliares conocidas.

Una técnica inteligente para aumentar los ejemplos de la clase minoritaria, es la utilización del algoritmo SMOTE (Synthetic Minoritary Over-sampling TEchnique) originario de Chawla et al. (2002). En este método, la creación de nuevas muestras se origina a través de la interpolación. En un primer paso elegimos los K vecinos más cercanos que pertenezcan a su misma clase. Posteriormente elegimos el número de muestras artificiales que se generarán y, finalmente, para generar una nueva muestra, se calcula la diferencia entre el vector de atributos bajo consideración y uno de los vecinos más cercanos de los k vecinos elegidos al azar. El resultado de la diferencia se multiplica por un valor aleatorio entre cero y uno.

El algoritmo SMOTE ha sido modificado de diferentes formas para ajustarse mejor a muchos ejemplos. Algunas de estas aportaciones son las efectuadas por Han et al (2005) que proponen el algoritmo Borderline-SMOTE para generar ejemplos positivos cercanos a una frontera. Wang et al (2006) presentan el algoritmo LLE-SMOTE (Locally Linear Embedding) que proyecta conjuntos de alta dimensionalidad a otro de menor dimensionalidad. En este espacio de reducida dimensionalidad es donde se aplica SMOTE, y después los ejemplos generados son transformados a su espacio de representación original.

Existen otras formas de obtener una representación mayor de la clase minoritaria basadas en técnicas de agrupamiento. Por ejemplo, Japkowicz (2001) emplea el algoritmo de clustering k-medias sobre cada clase por separado. Los clusters resultantes se sobremuestrean aleatoriamente hasta conseguir un equilibrio entre las clases. Otro trabajo en esta línea de investigación es el de Cohen et al. (2006), que también explora la generación de nuevas instancias a través de algoritmos de clustering, pero en este caso los centroides de los clusters se obtienen a través de un algoritmo aglomerativo jerárquico.

Respecto a los métodos de clasificación en entornos no balanceados que no cambian la distribución a priori de las clases, nos encontramos con las siguientes soluciones:

- Algoritmos de aprendizaje sensible al coste.
- Algoritmos de clasificación con sesgo hacia la clase minoritaria.
- Clasificadores de una clase.

En este trabajo, se han utilizado dos bases de datos: una proveniente de una Caja de Ahorros actualmente integrada en un banco español y una segunda base de datos ampliamente extendida, denominada "German Credit". En la base de datos que proviene de los clientes de la Caja de Ahorros, los resultados de los diferentes clasificadores que se presentan se aplican a un conjunto de datos que se han balanceado a través de un método mixto, donde se aplica el método SMOTE a la clase minoritaria y se reduce la muestra de la clase mayoritaria a través del método de submuestreo equilibrado del Cubo. En los datos de German Credit sólo se aplica el submuestreo equilibrado con el Método del Cubo, cuyo resumen teórico se presenta a continuación.

2.2. El Método del Cubo

La construcción de diseños estratificados es, a menudo, un ejercicio difícil, especialmente cuando se pretende estratificar usando un número elevado de variables cualitativas.

En muchos casos, se tiende a proceder cruzando todos los estratos de todas las variables, lo que hará que muchas de las celdas sean demasiado pequeñas para seleccionar muestras en ellas. Para solucionar este problema actualmente se utiliza el denominado muestreo equilibrado, que puede ser

visto como un tipo de calibración directamente integrada en el diseño muestral.

Considérese S una muestra de tamaño n, definida como un subconjunto de una población finita U de tamaño N. Esta muestra será equilibrada si para un vector de variables auxiliares $x_k=(x_{k1},\ldots,x_{kj},\ldots,x_{kp})'$ se verifica:

$$\frac{1}{n}\sum_{k\in S}x_k=\frac{1}{N}\sum_{k\in U}x_k \equiv \frac{N}{n}\sum_{k\in S}x_k=\sum_{k\in U}x_k \quad (1)$$

Lo cual significa que las medias (totales) en la muestra de las x-variables coinciden con las de la población. En otras palabras, en una muestra equilibrada el total de las x-variables son estimadas sin error.

Si la muestra S es aleatoria, se deben además satisfacer las siguientes ecuaciones de equilibrio:

$$\sum_{k\in S}\frac{x_k}{\pi_k}=\sum_{k\in U}x_k \quad (2)$$

Donde π_k son las probabilidades de inclusión en la muestra asociadas a cada elemento de la población.

En una muestra equilibrada y aleatoria, han de cumplirse ambas ecuaciones de equilibrio.

Si $x_k=\Delta k\epsilon U$, entonces la ecuación de equilibrio nos queda:

$$\sum_{k\in S}\frac{1}{\pi_k}=\sum_{k\in U}1=N \quad (3)$$

Donde $\sum_{k\in S}\frac{1}{\pi_k}$ es el estimador de Horvitz-Thompson del tamaño de la población N. Esta igualdad se cumple en los muestreos equiprobabilísticos, pero es evidente que no en aquellos con probabilidades desiguales.

Cabe señalar, que el muestreo estratificado no deja de ser un caso particular de muestreo equilibrado. Así, supóngase que la población U está particionada en L estratos, U_h,h=1,…,L, con tamaños poblacionales N_h, y que en cada uno de ellos seleccionamos una muestra aleatoria simple de tamaño n_h. En este caso, las variables auxiliares (o variables de equilibrio) serían los indicadores de pertenencia al estrato, es decir:

$$y_{kh}=\begin{cases}1 & si\ k\in U_h\\ 0 & en\ otro\ caso\end{cases} \quad (4)$$

Bajo este diseño, los estimadores de Horvitz-Thompson de los tamaños poblacionales de los estratos coinciden con el valor real N_h, lo cual en esencia es la condición de equilibrio tomando como variables auxiliares al vector y_k. Las ecuaciones de equilibrio serán entonces:

$$\sum_{k\in S}\frac{N_h y_{kh}}{n_h}=\sum_{k\in U}y_{kh} \quad con\ h=1,\ldots,L \quad (5)$$

En resumen, el equilibrado de la muestra se realiza sobre los totales marginales y no sobre cada una de las celdas contenidas en una tabla de contingencia. Sin embargo, la teoría habitual de estratificación permite estratos superpuestos, ya que la estratificación debe ser una partición de la población. El Método del Cubo permite además equilibrar directamente en los totales de superposición de estratos, simplemente utilizando los indicadores de los estratos como variables de equilibrio.

El Método del Cubo se configura como una clase de algoritmos de muestreo que seleccionan muestras equilibradas considerando el conjunto de probabilidades de inclusión definidas. Se basa en una transformación aleatoria del vector de probabilidades de inclusión hasta que se obtiene una muestra tal que:

- Se cumplen exactamente las probabilidades de inclusión.
- Se cumplen lo más exactamente posible las ecuaciones de equilibrio en las variables auxiliares (expresión 2).

Las ecuaciones definidas en la expresión 2 raramente podrán cumplirse de forma exacta. Esto se conoce como problema de redondeo y viene derivado de que la selección de la muestra es un problema entero. Para ver esto, considérese una muestra aleatoria extraída de una población tal que:

- x_k son números enteros $\Delta k\epsilon U$
- $\pi_k=\frac{1}{2}\ \Delta k\epsilon S\rightarrow\frac{N}{n}=2$
- $\sum_{k\epsilon U}x_k$ es un número impar

Entonces, según las ecuaciones de equilibrio:

$$2\sum_{k\in S}x_k=\sum_{k\in U}x_k \quad (6)$$

Como vemos, el lado izquierdo de la ecuación será siempre un número par, mientras que el lado derecho es, por definición, un número impar, lo cual implica que no existirían muestras de tamaño N/2 exactamente balanceadas.

El nombre del método proviene de la representación geométrica de un diseño de muestreo. En efecto, una muestra puede ser representada por un vector de indicadores muestrales de la siguiente manera:

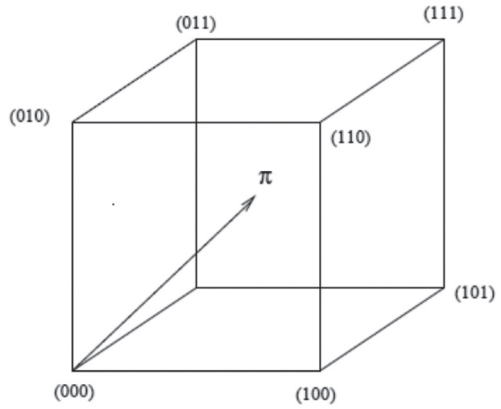
$$s=\left(I\left[1\epsilon s\right],I\left[2\epsilon s\right],\ldots,I\left[k\epsilon s\right],\ldots,I\left[N\epsilon s\right]\right)'$$

donde

$$I\left[k\epsilon s\right]=\begin{cases}1 & si\ k\epsilon s\\ 0 & si\ k\notin s\end{cases}$$

Entonces, una muestra puede ser vista como un vértice de un cubo N-dimensional tal como se muestra en la siguiente figura:

Figura 1. Posibles muestras en una población con N=2.



Fuente: Deville y Tillé (2004).

Sea p(s) = Pr(S=s) el diseño de muestreo o probabilidad de que la muestra s sea seleccionada, donde S es la muestra aleatoria y n(S) el tamaño de S. Bajo este esquema, la esperanza matemática de s se define como:

$$E(s) = \sum_{s \in S} p(s)s = \pi \tag{7}$$

donde π representa el vector de probabilidades de inclusión. Teniendo en cuenta esto, las ecuaciones de equilibrio definidas en la expresión 2 pueden escribirse como:

$$\frac{1}{n}\sum_{k \in S} x_k = \frac{1}{N}\sum_{k \in U} x_k \equiv \frac{N}{n}\sum_{k \in S} x_k = \sum_{k \in U} x_k \tag{8}$$

donde

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k$$

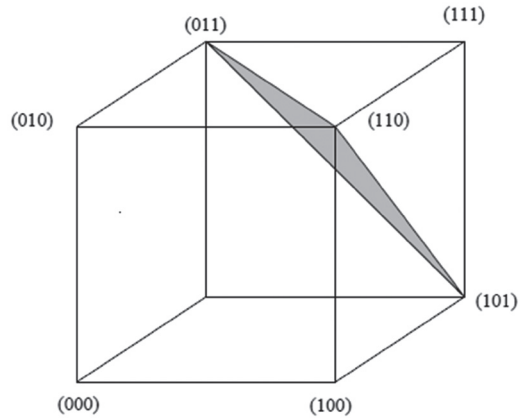
El sistema de ecuaciones definido en la expresión 8, con valores desconocidos s_k, define el siguiente subespacio afín en R^n de dimensión N - p:

$$Q = \left\{ u \epsilon \mathbb{R}^N / \sum_{k \epsilon U} \check{x}_k u_k = \sum_{k \epsilon U} x_k \right\} \tag{9}$$

Esto nos permite reformular el problema de selección de una muestra balanceada como la elección de un vértice de un cubo N-dimensional sobre el subespacio lineal Q.

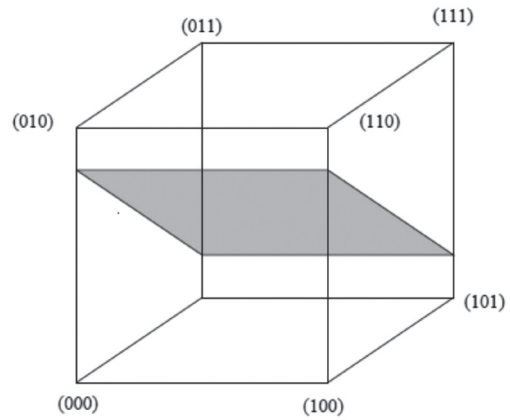
La Figura 2 muestra la representación gráfica del problema de la elección de una muestra de tamaño n=2 sobre una población N=3 y la Figura 3 el problema de redondeo descrito anteriormente.

Figura 2. Posibles muestras en una población de tamaño N=3 con una restricción de tamaño de muestras n=2.



Fuente: Deville y Tillé (2004).

Figura 3. Posibles muestras en una población de tamaño N=3 con una restricción que genera un problema de redondeo.



Fuente: Deville y Tillé (2004).

El método del cubo, Deville y Tille, (2004) se divide en dos fases: la fase de vuelo y la fase de aterrizaje.

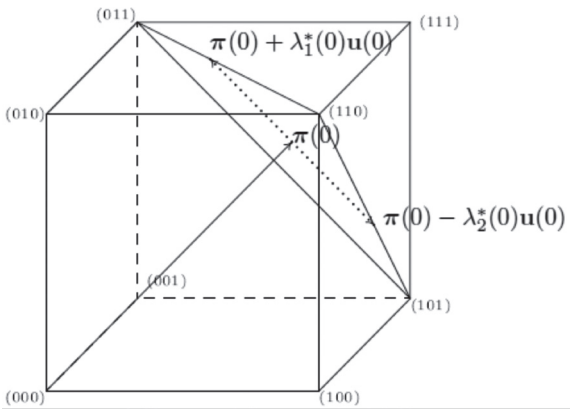
La fase de vuelo es un paseo aleatorio que comienza en el vector de probabilidades de inclusión y permanece en la intersección del cubo y el subespacio de restricciones. Este paseo aleatorio se detiene en un vértice de la intersección del cubo y el subespacio de restricciones. Al final de la fase de vuelo, si no se obtiene una muestra, la fase de aterrizaje determinará una muestra que esté tan cerca como sea posible del subespacio de restricciones.

2.2.1. Fase de vuelo

En primer lugar, debemos elegir un vector π de tal manera que $\pi + u(0)$ permanezca en el subespacio de restricciones. De hecho, el Método del Cubo es realmente una familia de métodos que dependen de la manera en que $u(0)$ sea elegido, pudiendo hacerse aleatoriamente o no.

Si desde π avanzamos en dirección de $u(0)$ cruzaremos necesariamente una cara del cubo (representado por $\pi(0) - \lambda_1^*(0)u(0)$ en la figura 4 y, análogamente, si avanzamos en dirección opuesta también cruzaremos otra cara del cubo ($\pi(0) - \lambda_2^*(0)u(0)$ en figura 4.

Figura 4. Fase de vuelo en una población de tamaño N=3 con una restricción de tamaño de muestra n=2.



Fuente: Deville y Tillé (2004).

En el primer paso, el vector $\pi(0) = \pi$ será modificado aleatoriamente a:

$$\pi(1) = \begin{cases} \pi(0) + \lambda_1^*(0)u(0) & con\ prob = p(0) \\ \pi(0) - \lambda_2^*(0)u(0) & con\ prob = 1 - p(0) \end{cases} \quad (10)$$

donde p(0) es determinado de tal manera que $E[\pi(1)] = \pi(0)$, es decir, no se modifican las probabilidades de inclusión establecidas a priori en base al diseño de la muestra.

Una vez finalizado este primer paso de la fase de vuelo, hemos saltado a una de las caras del cubo, lo que significa que al menos una de las componentes de $\pi(1)$ es igual a 0 ó 1, es decir, el problema de muestreo se ha reducido de una población N=3 a un problema con N=2.

Analíticamente la forma de proceder sería la siguiente:

- Consideramos $\pi(0) = \pi$.
- Para t = 0 hasta t=T,
  generamos un vector $u(t) = [u_k(t)] \neq 0$ de tal forma que:
  $u(t)$ está en el núcleo de la matriz
  $A = (x_1/\pi_1, \ldots, x_k/\pi_k, \ldots, x_N/\pi_N)$, es decir, $Au(t) = 0$.
  $u_k(t) = 0$ si $\pi_k(t)$ es un número entero (0 ó 1).
- Calcular $\lambda_1^*(t)$ y $\lambda_2^*(t)$ como los mayores valores tales que:
  $0 \leq \pi(t) + \lambda_1(t)u(t) \leq 1$.
  $0 \leq \pi(t) + \lambda_2(t)u(t) \leq 1$.
- Calcular $\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)u(t) & \text{con prob} = p(t) \\ \pi(t) - \lambda_2^*(t)u(t) & \text{con prob} = 1 - p(t) \end{cases}$
  donde $p(t) = \dfrac{\lambda_2^*(t)}{\lambda_1^*(t) + \lambda_2^*(t)}$

La fase de vuelo finalizará cuando no sea posible encontrar un vector $u(t) \neq 0$.

2.2.2. Fase de aterrizaje

Si al final de la fase de vuelo las ecuaciones de equilibrio no son exactamente satisfechas, entonces debemos aplicar la fase de aterrizaje.

Sea π^* el último vector $\pi(t+1)$ obtenido en la fase de vuelo. Si definimos el subesespacio:

$$U^* = \{k \epsilon U / 0 < \pi_k^* < 1\} \quad (11)$$

Es posible demostrar, Deville y Tille (2004), que $card(U^* \leq p)$

Donde p es el número de variables de equilibrio. El objetivo de la fase de aterrizaje es encontrar una muestra s casi-equilibrada, tal que:

$$E\left(s / \pi^*\right) = \pi^* \quad (12)$$

Existen dos formas de proceder:

- La fase de vuelo de programación lineal, que consiste en considerar todas las posibles muestras de U^*. Esto exige asignar un coste a cada muestra. Lo más lógico sería considerar la distancia entre la muestra y el subespacio de restricciones. A continuación, se busca un diseño de muestra en U^* que minimice el coste esperado y que satisfaga las probabilidades de inclusión obtenidas en π^*. Este problema puede ser resuelto porque el número de muestras a considerar es razonable debido al pequeño tamaño de U^*.
- La fase de vuelo por supresión de variables se puede utilizar cuando el número de variables de equilibrio es demasiado grande para que el programa lineal pueda ser resuelto por un algoritmo simplex (p>20). Con este método, al final de la fase de vuelo se elimina una variable auxiliar. A continuación, retornamos a la fase de vuelo hasta que no sea posible "moverse" en el subespacio de restricciones. Las restricciones son entonces "relajadas" sucesivamente de acuerdo a un orden de preferencia.

La entropía del diseño de la muestra depende de la forma en que los vectores u(t) son elegidos durante la fase de vuelo. Con el fin de aumentar la entropía, el vector u(t) puede ser elegido aleatoriamente, o la población puede reordenarse de forma aleatoria antes de seleccionar la muestra.

3. Evaluación de modelos de credit scoring

En cualquier etapa de un proceso de minería de datos resulta fundamental el poder estimar los niveles de calidad obtenidos por el modelo que hayamos construido. En función del problema y del modelo existen bastantes mecanismos de evaluación.

En CS los dos grupos que deben de ser discriminados son los cumplidores y los morosos. Existirán algunos buenos pagadores que se clasificarán como "malos" y también habrá registros de personas morosas en la base de datos que serán clasificadas como "buenos pagadores". Evaluar un modelo es encontrar aquel modelo que cometa el error mínimo en la clasificación de clientes. A continuación, se exponen los métodos de evaluación más habituales utilizados en la clasificación y que se pueden agrupar en tres grupos:

- Métodos basados en métricas.
- Métodos basados en curvas ROC.
- Métodos que incorporan una matriz de costes.

3.1. Métodos basados en métricas

La evaluación de modelos de clasificación basados en métricas se realiza con los resultados de la matriz de confusión definiendo diferentes métricas en función de la disparidad entre los resultados obtenidos y los datos reales.

Una matriz de confusión estándar tiene la estructura de la tabla 1. Una primera métrica es la exactitud o *accuracy* que corresponde con el porcentaje de acierto, definido como el número de casos acertados entre el número de casos totales.

Es decir,

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{13}$$

Esta es la métrica más sencilla y habitual, pero resulta muy engañosa, pues no toma en consideración la distribución de las clases. Por ejemplo, si tenemos un base de datos altamente desbalanceada con una población del 90% de individuos de clase A y un 10% restante de clase B, una exactitud cercana al 90% puede ser alcanzada con un modelo que simplemente estime que todos los casos son de la clase A.

Tabla 1. Matriz de confusión

		Clase clasificada como:		
		A+ (SI)	A- (NO)	Total casos
Estado real	A+ (SI)	Verdaderos positivos (VP) $FVP = \frac{VP}{TCP}$	Falsos negativos (FN) $FFN = \frac{FN}{TCP}$	TCP=VP+FN 1
	A- (NO)	Falsos positivos (FP) $FFP = \frac{FP}{TCN}$	Verdaderos negativos (VN) $FVN = \frac{VN}{TCN}$	TCN=VN+FP 1

La *precisión* es otra métrica que determina la proporción de casos positivos correctos, definida como el número de los casos de esa clase clasificados correctamente dividido por el número total de casos identificados como esa misma clase. Es decir:

$$Precisión_{SI} = \frac{VP}{VP + FP} \tag{14}$$

Según el ejemplo de la tabla 1, para la clase SI, la precisión correspondería con el cociente de los ejemplos clasificados correctamente (verdaderos positivos) entre todos los elementos clasificados de esa clase (verdaderos positivos + falsos positivos).

La exhaustividad o cobertura (*recall*) es otra medida muy usada y se define como el número de casos de esa clase predichos correctamente entre el número de casos de esa clase que existen en la base de datos. Por ejemplo, para calcular la cobertura de la clase SI se divide el número de casos correctamente clasificados de SI (verdaderos positivos) por la suma de los verdaderos positivos más los falsos negativos (que realmente corresponden con los positivos de la base de datos que han sido clasificados como negativos). Es decir:

$$Exhaustividad_{SI} = \frac{VP}{VP + FN} \tag{15}$$

Igualmente, puede utilizarse estas métricas para la clase NO.

La precisión habla de la "bondad" de la predicción de esa clase, mientras que la cobertura indica la capacidad del modelo para identificar los casos correctos, de esa clase, que existen en la base de datos. En muchos casos se buscan modelos que maximicen ambas métricas para una o varias clases.

FMeasure combina ambas medidas mediante una media armónica que es más restrictiva que una media aritmética. Se define como:

$$FMeasure = \frac{2}{\frac{1}{Precisión} + \frac{1}{Exhaustividad}} \tag{16}$$

En algunas disciplinas científicas, son mucho más utilizados los conceptos de "sensibilidad" y "especificidad". La sensibilidad (*S*) es la probabilidad de clasificar correctamente una instancia cuyo estado real sea la presencia de la condición de interés, y corresponde con la precisión de los casos positivos: $\frac{VP}{VP + FN}$.

La especificidad (*E*) se define como la probabilidad de clasificar correctamente a un individuo cuyo estado real sea la ausencia de la condición, y corresponde con la precisión de la clase negativa: $\frac{VN}{VN + FP}$.

Existen otras medidas de exactitud como el índice de *Youden*, que se calcula a través de las diferencias entre las proporciones de respuestas positivas correctas e incorrectas. Se define como:

$$J = especifidad + sensibilidad - 1 \tag{17}$$

La tasa o razón de verosimilitud (*LR*) es otra medida muy utilizada como forma de discriminar los clasificadores. Esta tasa de verosimilitud representa el grado de evidencia de una respuesta del clasificador a favor de la presencia de la condición con respecto a la ausencia de la condición.

Otras métricas muy utilizadas son el estadístico *Kappa* y el Coeficiente de Correlación de *Matthews* (*MCC*).

Kappa se define como:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (18)$$

donde *P(A)* corresponde con la *Exactitud* o porcentaje de casos acertados, $P(A) = \frac{VP + VN}{VP + VN + FP + FN} = \frac{VP + VN}{N}$, y *P(E)* corresponde con el porcentaje de casos cambiados y se define como:

$$P(E) = \frac{VP + FP}{N} \times \frac{VP + FN}{N} + \frac{VN + FN}{N} \times \frac{VN + FP}{N}$$

Y el coeficiente de correlación de Matthews:

$$MCC = \frac{VPxVN - FPxFN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \quad (19)$$

Ambas métricas toman en cuenta el porcentaje de casos correctamente clasificados entre los casos cambiados obtenidos a partir de la matriz de confusión, de forma que cuanto más se aproximen las métricas a uno, el clasificador mejor capacidad generalizadora tendrá el estimador para ambas clases.

3.2. Métodos basados en la curva ROC

La evaluación de modelos de clasificación basados en curvas ROC (*Receiver Operating Characteristic*) representan de forma gráfica el rendimiento de un clasificador que muestra la distribución de las fracciones de verdaderos positivos y la fracción de falsos negativos La curva ROC nos proporciona una herramienta visual para examinar la capacidad que dispone un clasificador para detectar correctamente a los individuos con presencia de la condición de interés en el análisis y su incapacidad para identificar los individuos del grupo de ausencia.

Swets y Pickett (1982) señalan tres importantes propiedades de las curvas ROC:

- Representan un índice de exactitud intrínseco: es un índice de la capacidad del clasificador para discriminar estados y seleccionan el estado correcto, independientemente del criterio de selección. Son curvas que reflejan las probabilidades subjetivas junto con las utilidades que usualmente determina este criterio, por lo que podemos decir que es un índice del criterio de decisión.
- Utilizan las probabilidades a priori de los posibles estados de forma más precisa, y también los costes de las decisiones correctas y equivocadas para determinar el criterio óptimo de decisión de un clasificador.
- En tercer lugar, las curvas ROC contienen las estimaciones de las probabilidades de los distintos tipos de resultados para todos y cada uno de los criterios de decisión.

La curva ROC es el patrón de oro en muchas áreas de análisis de modelos, ya que representan de forma compacta muchísima información del rendimiento de un clasificador.

Muchos autores consideran a las curvas ROC como herramientas fundamentales en la fase de evaluación de un modelo, entre ellos Zweig y Campbell (1993).

El área bajo la curva ROC (AUC) es equivalente al valor del estadístico suma de rangos de Wilcoxon, tal y como señalan Bamber (1988) y Hanley y McNeil (1982), lo que permite trasladar las propiedades del estadístico de Wilcoxon a las medidas de exactitud global del AUC. También se interpreta como un promedio de la sensibilidad para todos los valores de especificidad y, de forma análoga, como un promedio de la especificidad para todos los posibles valores de sensibilidad. Breiman et al. (1984) relacionan el AUC con el coeficiente de Gini.

3.3 Métodos basados en una matriz de costes

La evaluación de modelos de clasificación basados en costes es otra alternativa disponible para comparar modelos. Estos métodos se basan en establecer una matriz de costes asociados a la clasificación. Cuando utilizamos el porcentaje de acierto o de error para evaluar el desempeño de nuestros modelos de clasificación, estamos suponiendo que ambos tipos de errores son equivalentes.

Los factores de riesgo de los modelos de CS, es decir, los factores que están detrás de los errores tipo I (admitir como sana una operación insolvente) y tipo II (rechazar como insolvente una operación sana) no son los mismos. No es igual, en términos de coste económico, clasificar a un cliente como bueno, concederle el crédito y que luego no nos lo devuelva que no conceder el crédito a una persona que es cliente. En el primer caso estamos expuestos a un caso de riesgo de crédito y, en el otro caso, incurrimos en un coste de oportunidad por la pérdida potencial de buenos clientes.

La mayoría de algoritmos de aprendizaje, por su propia naturaleza, buscan minimizar el número de errores del clasificador generado. Sin embargo, son múltiples los problemas de aprendizaje automático en los que los errores cometidos por el clasificador no tienen la misma importancia, Provost y Fawcett, (2001).

Una función de coste esperado, siguiendo la notación de West (2000) y de Boj et al. (2009), es aquella que pondera el porcentaje de los que devuelven el crédito y los que no por sus respectivos costes. Si llamamos C_e al coste esperado, la función es la siguiente:

$$C_e = C_{12}\,\pi_2 \frac{n_2}{N_2} + C_{21}\,\pi_1 \frac{n_1}{N_1} \quad (20)$$

Donde $\frac{n_2}{N_2}$ y $\frac{n_1}{N_1}$ son la proporción de buenos y malos pagadores, C_{12} y C_{21} son los costes asociados a los errores de tipo I y II y, π_1 y π_2 son las probabilidades a priori de buenos y malos riesgos. En el artículo publicado por West se propone estudiar las probabilidades a priori con dos valores: π_2 = 0.144 y π_2 = 0.249. Estos valores fueron obtenidos por Gopinathan y O'Donnell (1998) y Jensen (1992) de una experiencia real.

La complejidad existente en el cálculo de los costes asociados a los dos tipos de errores es considerable, dado que los factores que los afectan son difíciles de cuantificar.

A los costes asociados a la pérdida del monto del crédito otorgado (C_{12}) hay que restarle los ingresos recibidos antes

de pasar a la situación de moroso u otros ingresos recibidos por valores de la propiedad asegurada en el momento de la liquidación, y sumar aquellos gastos que se deriven de costes legales, costes administrativos, etcétera.

Los costes asociados al error de tipo II (C_{21}) están asociados a la pérdida de los intereses que se generarían si se hubiera concedido el préstamo al buen pagador más la pérdida o beneficio de destinar este crédito no concedido a otro cliente. Estos costes se pueden llamar coste de oportunidad. Por otra parte, si el solicitante del crédito es un cliente del banco y no se le concede, muy probablemente deje de ser cliente. Si el peticionario del crédito no es cliente y no se le concede el crédito, casi con toda seguridad ese demandante no llegue a ser cliente de la entidad financiera a quien ha dirigido su solicitud de dinero y, desde un punto de vista más práctico, para cuantificar estos costes deberíamos contar con información de todos los productos financieros que dejaría de consumir a lo largo del ciclo de vida del cliente. Estos costos es muy probable que cambien con el tiempo, por lo que se puede concluir que, aunque se pueda establecer unos rangos entre los que probablemente estén, es prácticamente improbable el cálculo exacto.

Una de las escasas referencias que se disponen de una matriz de coste se encuentra en los datos del banco alemán, uno de los dos ejemplos utilizados en este artículo, que se descarga del repositorio de la UCI, y que es la siguiente:

$$C_{ij} = \begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix}$$

En esta matriz, los costes asociados al error de tipo I se suponen cinco veces mayores a los costes que involucra el error de tipo II, sin embargo, algunos investigadores mantienen que entre los errores de ambos tipos existe una diferencia mayor. Las conversaciones mantenidas con varios responsables de Cajas de Ahorro y de Bancos Comerciales indican que el establecimiento de esta matriz de costes para modelos de CS es de una complejidad considerable y la relación entre los costes asociados a los errores de tipo I y de tipo II puede abarcar un abanico muy amplio, dependiendo del tipo de crédito concedido y de la vinculación del prestatario con el banco, fundamentalmente. Esta complejidad se agrava aún más en épocas de crisis económicas.

En general, se puede afirmar que los métodos de aprendizaje sensibles al coste suelen ser adaptaciones de algoritmos existentes, como los árboles de decisión (Ting, 1998). Sin embargo, existen estrategias que son independientes del algoritmo de aprendizaje utilizado. Estas estrategias, corrientemente denominadas meta-esquemas de aprendizaje, toman como entrada un algoritmo de aprendizaje, una colección de datos de entrenamiento y una distribución de costes, y generan un clasificador. Entre los trabajos pioneros de clasificación sensible al coste se encuentran: Turney (1995 y 2000), Ting (1998), Elkan (2001), Zadrozny y Elkan (2001) y Lizotte (2003).

Para los modelos con costes asimétricos, podemos encontrar diferentes estrategias para abordar una correcta clasificación:

- Basadas en un umbral. Witten y Frank (1999) proponen un modelo aplicable a todo algoritmo cuya salida sea un clasificador que predice valores numéricos (como probabilidades, similitudes, etc.).
- Ponderando las instancias modificando los pesos asignados a cada clase, de manera que se asigna más peso a los ejemplos cuyos errores puedan ser más costosos. Ting (1998) sugiere dar pesos a cada cliente (*Instance Weighting*), pero un peso mayor a los individuos de una clase (por ejemplo, a la clase de los que no devuelven el crédito), con el objetivo de que el algoritmo se fije especialmente en clasificar correctamente estos ejemplares, minimizando el error sobre ellos. También se encuentra dentro de este ámbito el algoritmo *MetaCost* de Domingos (1999), que tiene como objetivo re-etiquetar cada muestra de entrenamiento por la estimación del riesgo de Bayes. Finalmente, el clasificador se entrena con un método no basado en costes con el conjunto que ya ha sido re-etiquetado. Este método es aplicable a cualquier algoritmo de aprendizaje.

Algunas aportaciones interesantes en los métodos de clasificación a través de los costes las podemos encontrar en Ling et al. (2004). Estos autores emplean árboles de decisión con costes mínimos. La idea que subyace es introducir un factor de costes mientras se va construyendo el árbol de acuerdo con los criterios de división que minimizan el coste total, en lugar de minimizar la entropía. En este sentido, los árboles de decisión con costos mínimos y *MetaCost* son similares, aunque hay una gran diferencia: en los árboles de decisión con un coste mínimo la parte más sensible a los costos se construye directamente en el clasificador, mientras que el algoritmo *MetaCost* puede utilizar no sólo los árboles de clasificación, sino cualquier método de clasificación: redes neuronales, máquinas de vectores soporte, redes bayesianas, etcétera.

Otro enfoque importante es el de López et al. (2010) que utilizan reglas difusas para problemas de bases de datos no balanceados. Desde esta perspectiva, el aprendizaje sensible al coste en las bases de datos que analizan, alcanza un buen equilibrio entre las clases, mejorando la clase positiva (sensibilidad) y no perjudicando la precisión de la clase considerada negativa (especificidad).

4. Casos de estudio

4.1. Descripción de las bases de datos utilizadas

El primer conjunto de datos tiene 1.786 registros, que representan a los clientes de una Caja de Ahorros de la Rioja que demandaron un crédito entre los años 2010 y 2011. Del total de los casos, 1.609 devuelven el crédito frente a los 177 que no reingresan el dinero prestado. La base de datos original entregada contiene diecisiete variables y sus atributos son tanto numéricos como nominales. Los atributos de cada cliente informan sobre diversas cuestiones: estado civil, sexo, edad, tipo de trabajo, código de profesión, situación de la vivienda, nacionalidad, etcétera, así como de otra información relacionada con el crédito: finalidad, importe solicitado, importes pendientes en su entidad bancaria y en otras, patrimonio, valor neto de la vivienda, situación de ingresos, cuotas y gastos de alquiler y préstamos, etcétera. También existe una variable que indica si el crédito se ha concedido o se ha denegado.

La segunda base de datos, "German Credit", procede del repositorio de la Universidad de California del directorio "Statlog Databases" cedido por Ross D. King y Hans Hofman. Este conjunto de datos está formado con la información de mil clientes de una entidad financiera bancaria que solicitaron un crédito. Contiene veinte variables de la que siete son numéricas y trece nominales. La variable clase es binaria y nos indica si el cliente es fiable para concederle un crédito o bien denegárselo.

4.2. Calidad de la muestra a través del Método del Cubo para ambos modelos

En las tablas 2 y 3 se encuentran los resultados del submuestreo mediante el Método del Cubo. En la primera columna están los totales para cada variable de clasificación que interviene en el procedimiento de selección, la segunda columna es el valor que presenta el estimador del total de Horvitz-Thompson (que depende de la muestra). Las dos últimas columnas representan la desviación absoluta y relativa respecto al valor real, informando sobre la calidad del ajuste efectuado por el Método del Cubo.

Para la base de datos de la Caja de Ahorros, de los 1.609 ejemplos disponibles que devolvieron el crédito, se han seleccionado 312 registros a través del Método del Cubo. Para esta selección de los individuos las variables auxiliares de equilibrio utilizadas han sido: el estado civil, la nacionalidad, las condiciones de la casa y el tipo de trabajo de las personas que solicitan el crédito.

La base de datos que finalmente se utiliza es la combinación resultante de la aplicación del método del cubo y del aumento de los registros de la clase más desfavorecida a través del método SMOTE.

Para la extracción de la muestra en el caso de la base de datos *German Credit* se han empleado las siguientes variables: situación actual de la cuenta corriente, historia del crédito, cuenta de ahorro, propiedades, tiempo de empleo y propósito del crédito.

Tabla 2. Calidad del submuestreo equilibrado. Método del Cubo. Caja de Ahorros.

VARIABLES	Totales	Estimador Horvitz Thompson	Desviación absoluta	Desviación relativa
UNO	1.609	1.609	0,00	0,00
CASADO	884	882	2,42	-0,27
SEPARADO	86	87	-0,71	0,83
SOLTERO	639	641	-1,71	0,27
ESPAÑOL	1.445	1.445	-0,21	0,01
EXTRANJERO	164	164	0,21	-0,13
LIBRE	497	496	0,81	-0,16
HIPOTECA	609	607	2,01	-0,33
ALQUILER	138	135	3,11	-2,26
FAMILIA	300	303	-3,49	1,16
OTRAS	65	67	-2,44	3,76
TECNICO	435	434	1,44	-0,33
OBRERO_FIJO	476	472	3,90	-0,82
OBRERO_TEMPORAL	159	159	0,03	-0,02
OBRERO_ESP_FIJO	161	164	-2,79	1,73
OBRERO_ESP_TEMPORAL	28	29	-0,90	3,23
AUTONOMO	155	154	0,84	-0,54
JUBILADO_RENTISTA	105	106	-0,98	0,94
NO_ACTIVO	90	92	-1,53	1,70

4.3. Análisis comparativo de modelos

En la fase de la minería de datos relacionada con la selección de modelos se han utilizado una amplia variedad de algoritmos como: árboles de clasificación, redes neuronales, máquinas de vectores soporte, metaclasificadores, regresión logística y redes bayesianas.

La búsqueda de los parámetros óptimos de los modelos para cada método se ha realizado mediante un proceso iterativo de búsqueda aleatoria (*random search*), reduciéndose en cada iteración el área de búsqueda de los parámetros óptimos hasta conseguir minimizar al máximo los costes asociados al problema de clasificación.

Tabla 3. Calidad del submuestreo equilibrado. Método del Cubo. German Credit.

VARIABLES	Totales	Estimador Horvitz Thompson	Desviación absoluta	Desviación relativa
UNO	700	700	0,00	0,00
< 0 DM	139	140	-1,00	0,72
0 < 200 DM	164	163	0,67	-0,41
>= 200 DM	49	49	0,00	0,00
No checking account	348	348	0,33	-0,10
No credits taken	15	14	1,00	-6,67
All credits at this bank paid back duly	21	21	0,00	0,00
Existing credits paid back duly till now	361	362	-0,67	0,18
Delay in paying off in the past	60	58	1,67	-2,78
Critical account	243	245	-2,00	0,82
< 100 DM	386	385	1,00	-0,26
< 500 DM	69	70	-1,00	1,45
< 1000 DM	52	51	0,67	-1,28
>= 1000 DM	42	42	0,00	0,00
Unknown/ no savings account	151	152	-0,67	0,44
Real estate	222	224	-2,00	0,90
If not A121 : building society savings agreement/	161	163	-2,00	1,45
If not A121/A122 : car or other, not in attribute	230	226	4,00	-1,59
Unknown / no property	87	86	0,67	-0,77
Unemployed	39	40	-0,67	1,71
< 1 year	102	105	-3,00	2,94
< 4 years	235	233	1,67	-0,71
< 7 years	135	133	2,00	-1,48
>= 7 years	189	189	0,00	0,00
Car (new)	145	145	0,33	-0,23
Car (used	86	89	-2,67	3,10
Furniture/equipment	123	124	-0,67	0,54
Radio/television	218	217	1,00	-0,46
Domestic appliances	8	7	1,00	-12,50
Repairs	14	14	0,00	0,00
Education	28	26	2,33	-8,33
Retraining	8	7	1,00	-12,50
Business	63	65	-2,33	3,70
Others	7	7	0,00	0,00
Extranjero	667	670	-2,67	0,40
No extranjero	33	30	2,67	-8,08

Los resultados de los distintos modelos de clasificación, así como el método del Cubo, que aparecen en las siguientes tablas han sido obtenidos con el software estadístico R.

En los cuadros siguientes se presentan, para diversos métodos de clasificación, el porcentaje correctamente clasificado, tanto para el total como para cada una de las clases, el área bajo la curva ROC y la valoración de los costes suponiendo una matriz de costes como la comentada en el epígrafe anterior y para los dos valores de las probabilidades a priori de clientes morosos propuestos por West (2000).

En la estimación de los modelos se ha aplicado un esquema de validación denominado "validación cruzada-*k*", que consiste en dividir el conjunto de entrenamiento en k particiones o pliegues (folds) y repetir el procedimiento de entrenamiento y validación k veces, de forma que en cada una de ellas se entrene el modelo con k-1 particiones y se evalúe con la partición restante. Las métricas finales de validación cruzada se obtienen como promedio de las validaciones parciales de cada una de las particiones no usadas en el proceso de entrenamiento.

Tabla 4. Muestra desbalanceada Caja de Ahorros (1.609 instancias clase SI y 177 clase NO).

Técnica	CLASE SI (%)	CLASE NO (%)	TOTAL (%)	AREA ROC	COSTE 1 π1=0,144	COSTE 2 π2=0,249
Regresión Logística	97,5	31,3	90,9	0,777	0,516	0,654
C 4.5	97,9	35,2	91,7	0,885	0,485	0,614
Maq. Vect. Soporte	99,9	0,1	89,9	0,500	0,720	0,928
Red neuronal MLP	94,6	35,8	88,7	0,817	0,508	0,631
Red neuronal RBF	100	0	90	0,825	0,720	0,852
Naïve Bayes	66,4	81,6	68,0	0,832	0,420	0,388
Red Bayesiana(K2)	88,3	69,8	86,4	0,884	0,318	0,356
Metaclasificadores						
Bagging	99,3	20,7	91,4	0,879	0,577	0,741
Adaboost	97,3	39,7	91,5	0,893	0,457	0,577
Random Forest	98,4	33,5	91,9	0,846	0,492	0,628
STAKING C (5 modelos)	97,7	24,6	90,4	0,772	0,563	0,715
Decorate	97,1	37,4	91,1	0,860	0,476	0,600
Metacost 9/1	86,9	75,4	85,8	0,828	0,289	0,313

En las tablas 4 y 5 se muestran los resultados obtenidos con la base de datos correspondiente a la "Caja de Ahorros" sin balancear y balanceada, respectivamente. En cada una de las tablas se muestra el porcentaje de clases clasificadas correctamente, el área bajo la curva ROC y, los costes obtenidos según la ecuación 21 asociados con las diferentes probabilidades a priori propuestas (π_2 = 0.144 y π_2 = 0.249).

En la tabla 4, claramente se puede observar que en casi todos los métodos empleados existe una importante diferencia de aciertos entre clases. Así, tal y como era de esperar, los métodos clásicos de clasificación favorecen en general a la clase mayoritaria, salvo en el caso del clasificador bayesiano Naïves Bayes. Por otro lado, la red bayesiana es la que obtiene resultados más equilibrados entre ambas clases. Se da el caso extremo en el que el modelo basado en redes neuronales clasifica correctamente a toda la clase mayoritaria y a ningún caso de la minoritaria. Similar comportamiento se observa cuando se aplica el algoritmo basado en máquinas de vectores soporte. Tampoco los metaclasificadores estiman correctamente ambas clases. Solamente introduciendo un método cuyo aprendizaje es sensible al coste, Metacost 9/1, se logra equilibrar la precisión de los ejemplos bien clasificados. En este caso el coste que se ha introducido coincide con la ratio de desbalanceo de las clases.

En los resultados del modelo balanceado, tabla 5, se puede observar que todos los algoritmos utilizados realizan una clasificación más equilibrada de aciertos para ambas clases.

Para facilitar el análisis de los modelos utilizados se muestran las tablas 6 y 9, donde se recogen las diferencias de costes entre los modelos obtenidos con la base de datos balanceada frente a la original. Los mejores modelos, si consideramos esta función de costes, son aquellos que producen una mayor diferencia de costes entre las dos bases de datos.

En los datos de la Caja de Ahorros los mejores ajustes se producen, para los dos escenarios de costes considerados, cuando se utiliza las Máquinas de Vectores Soporte (-63,09% y -71,13%) y el multiclasificador Bagging (-61,15% y -70,96%).

Tabla 5. Muestra equilibrada Caja de Ahorros. Método Cubo y SMOTE.

Técnica	CLASE SI (%)	CLASE NO (%)	TOTAL (%)	AREA ROC	COSTE 1 π1=0,144	COSTE 2 π2=0,249
Regresión Logística	84,0	83,9	83,9	0,912	0,253	0,254
C 4.5	84,6	80,6	82,6	0,844	0,271	0,291
Maq. Vect. Soporte	83,3	82,9	83,1	0,831	0,266	0,268
Red neuronal MLP	83,0	84,2	83,6	0,896	0,259	0,253
Red neuronal RBF	75,6	82,6	79,1	0,852	0,334	0,300
Naïve Bayes	67,3	85,2	76,2	0,858	0,387	0,298
Red Bayesiana(K2)	84,9	84,5	84,7	0,925	0,240	0,243
Metaclasificadores						
Bagging	84,9	86,8	85,9	0,924	0,224	0,215
Adaboost	85,9	84,2	85,0	0,925	0,235	0,243
Random Forest	86,2	86,8	86,5	0,934	0,213	0,210
STAKING C (5 modelos)	83,0	86,1	84,6	0,936	0,245	0,230
Decorate	84,1	87,4	85,8	0,926	0,227	0,210
Metacost 9/1	82,9	84,7	83,8	0,922	0,257	0,248

Tabla 6. Diferencias de los costes entre el modelo balanceado y el desequilibrado. Caja de Ahorros.

Técnica	COSTE 1 π1=0,144	COSTE 2 π2=0,249	Var(%) COSTE 1	Var(%) COSTE 2
Regresión Logística	-0,263	-0,400	-50,91	-61,17
C 4.5	-0,213	-0,324	-44,06	-52,69
Maq. Vect. Soporte	-0,454	-0,660	-63,09	-71,13
Red neuronal MLP	-0,249	-0,378	-49,02	-59,84
Red neuronal RBF	-0,386	-0,552	-53,62	-64,84
Naïve Bayes	-0,033	-0,090	-7,95	-23,09
Red Bayesiana(K2)	-0,077	-0,114	-24,30	-31,96
Metaclasificadores				
Bagging	-0,353	-0,526	-61,15	-70,96
Adaboost	-0,223	-0,334	-48,71	-57,92
Random Forest	-0,279	-0,417	-56,71	-66,47
STAKING C (5 modelos)	-0,317	-0,485	-56,40	-67,83
Decorate	-0,249	-0,390	-52,30	-64,98
Metacost 9/1	-0,033	-0,065	-11,31	-20,85

En definitiva, el ahorro para la entidad financiera puede ser considerable al utilizar los datos balanceados, pues en algunos casos llega a reducirse a más de la mitad.

En relación con los datos del banco alemán, al igual que ocurría en la Caja de Ahorros analizada anteriormente, se presenta una variabilidad considerable en los porcentajes de acierto de las clases en gran parte de los modelos. En la tabla 7, se observa como los dos prototipos de redes neuronales son los que consiguen un mejor equilibrio entre las clases. Los porcentajes de acierto a nivel global están alrededor del 75% y no muestran muchas diferencias entre los diferentes algoritmos.

En la tabla 8, donde se presentan los resultados que han sido calculados con la base de datos equilibrada a través del método del cubo, los porcentajes de acierto entre los que devuelven el crédito y los morosos aparecen más igualados.

En la base de datos del banco alemán los modelos que reducen más el valor de la función de coste, en el escenario donde la probabilidad a priori de riesgo es de 0,144, son algunos de los metaclasificadores, el Metacost y el Naïve Bayes, destacando en primer lugar el Adaboost, - 14,75%. Cuando se considera un valor de la probabilidad a priori de no devolver el crédito de 0,249, el modelo que arroja los mejores resultados es el multiclasificador *Stacking* que, en este caso, ha sido construido con cinco modelos individuales, y cuyo ahorro de costes se cifra en un 34,85%.

Tabla 7. Muestra desbalanceada German Credit (700 instancias clase SÍ y 300 clase NO).

Técnica	CLASE SI (%)	CLASE NO (%)	TOTAL (%)	AREA ROC	COSTE 1 π1=0,144	COSTE 2 π2=0,249
Regresión Logística	87,0	50,0	75,9	0,792	0,471	0,720
CART	85,4	47,7	74,1	0,715	0,502	0,761
Maq. Vect. Soporte	88,4	49,0	76,6	0,694	0,466	0,722
Red neuronal MLP	76,7	62,7	72,5	0,725	0,468	0,639
Red neuronal RBF	75,0	66,7	72,5	0,723	0,454	0,602
K vecios próximos	89,1	39,3	74,2	0,727	0,530	0,838
Naïve Bayes	86,4	49,7	75,4	0,787	0,479	0,728
Red Bayesiana(K2)	85,9	51,3	75,5	0,780	0,471	0,712
Metaclasificadores						
Bagging	82,1	59,3	75,3	0,790	0,446	0,641
Adaboost	88,0	42,7	74,4	0,773	0,515	0,804
Random Forest	91,1	43,7	76,9	0,782	0,482	0,768
STAKING C (5 modelos)	93,9	32,0	75,3	0,794	0,542	0,892
Decorate	87,3	45,0	74,9	0,750	0,505	0,780
Metacost 5/1	58,0	78,0	64,0	0,675	0,518	0,589

Tabla 8. Muestra equilibrada German Credit (300 instancias clase SI y 300 clase NO).

Técnica	CLASE SI (%)	CLASE NO (%)	TOTAL (%)	AREA ROC	COSTE 1 π1=0,144	COSTE 2 π2=0,249
Regresión Logística	69,0	72,3	70,7	0,777	0,465	0,578
CART	65,3	71,7	68,5	0,678	0,501	0,613
Maq. Vect. Soporte	70,7	72,0	71,3	0,702	0,452	0,569
Red neuronal MLP	70,7	74,3	72,5	0,778	0,436	0,540
Red neuronal RBF	71,0	71,7	71,3	0,770	0,452	0,570
K vecios próximos	75,0	65,0	70,0	0,742	0,466	0,624
Naïve Bayes (kernel)	73,0	70,3	71,7	0,774	0,445	0,573
Red Bayesiana(K2)	69,7	70,7	70,2	0,760	0,470	0,592
Metaclasificadores						
Bagging	71,3	76,0	73,7	0,789	0,418	0,514
Adaboost	70,3	74,3	72,3	0,707	0,439	0,543
Random Forest	69,3	70,3	69,8	0,779	0,477	0,600
STAKING C (5 modelos)	69,0	72,0	70.5	0,781	0,467	0,581
Decorate	71,0	74,0	72,5	0,780	0,435	0,541
Metacost 1/1	67,0	77,3	72,2	0,787	0,446	0,530

Tabla 9. Diferencias de los costes entre el modelo balanceado y el desequilibrado. Banco German Credit.

Técnica	COSTE 1 π1=0,144	COSTE 2 π2=0,249	Var(%) COSTE 1	Var(%) COSTE 2
Regresión Logística	-0,006	-0,142	-1,37	-19,78
CART	-0,001	-0,148	-0,15	-19,43
Maq. Vect. Soporte	-0,014	-0,153	-3,02	-21,25
Red neuronal MLP	-0,032	-0,099	-6,87	-15,54
Red neuronal RBF	-0,002	-0,032	-0,39	-5,35
Naïve Bayes	-0,064	-0,214	-12,13	-25,56
Red Bayesiana(K2)	-0,034	-0,156	-7,02	-21,40
Metaclasificadores				
Bagging	-0,028	-0,127	-6,23	-19,78
Adaboost	-0,076	-0,260	-14,75	-32,42
Random Forest	-0,005	-0,167	-1,02	-21,81
STAKING C (5 modelos)	-0,075	-0,311	-13,82	-34,85
Decorate	-0,069	-0,239	-13,73	-30,59
Metacost	-0,072	-0,059	-13,90	-9,99

5. Conclusiones

Entre los métodos existentes en la literatura estadística para la selección de submuestras, el "Método del Cubo" es el único que permite seleccionar una muestra equilibrada sobre variables auxiliares con probabilidades de inclusión que puedan ser iguales o no. Como ya se ha comentado, el método del Cubo selecciona únicamente las muestras cuyos estimadores de Horvitz-Thompson son iguales a los totales de las variables auxiliares conocidas.

Los resultados obtenidos con ambas bases de datos han demostrado que, cuando las bases de datos se balancean con el Método del Cubo, los algoritmos de clasificación utilizados generalmente equilibran los aciertos entre las clases y reducen el valor de la función de costes utilizada. En muchos de los métodos utilizados el decremento de los costes es considerable.

Si nos enfocamos en desarrollar modelos de *credit scoring* que estén de acuerdo a las exigencias de calcular la probabilidad de default que requieren los modelos de Basilea II y III, tenemos que reducir la búsqueda a dos de los modelos probabilísticos utilizados previamente: regresión logística y redes bayesianas, pues son los únicos que cumplen estos requisitos. En la base de datos de la Caja de Ahorros, es la regresión logística la que presenta un mejor desempeño, mientras que, en las estimaciones del banco alemán, las redes bayesianas disminuyen mucho más el coste asociado a la clasificación.

Como en este tipo de problemas el coste económico de la clasificación es diferente según las clases, incorporar la matriz de costes en los modelos se considera muy conveniente. Algunos métodos como el Metacost obtienen unos resultados muy aceptables ponderando la matriz de costes, ya que optimizan el análisis coste beneficio. Sin embargo, sugerimos como una futura línea de investigación, trabajar en encontrar una función de coste que esté consensuada por los expertos bancarios y basada en una teoría económica que la avale.

6. Bibliografía

Bamber, D.C., 1988. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol., 12, 387-415.

Bessis, J., 2002. Risk Management in Banking. Second edition. Chichester: John Wiley and sons, 496 pp.

Boj, E., Claramunt, M.M., Esteve, A. y Fortiana, J., 2009. Criterio de selección de modelo en credit scoring Aplicación del análisis discriminante basado en distancias. Anales del Instituto de Actuarios Españoles. 3:209–30.

Breiman, L., Friedman, J.H., Olshen, R.A. y Stone, C.J., 1984. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Book & Software.

Caouette, J., Altman, E. y Narayanan, P., 1998. Gestión del riesgo de crédito, el próximo gran desafío financiero. Wiley Frontiers in Finance, vol. Fronteras Wiley en Finanzas, Wiley & Sons, Inc., Nueva York.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S. y Geissbuhler, A., 2006. Learning from imbalancing Data in Surveillance of Nosocomial Infection. Artificial Intelligence in Medicine, pp. 7-18.

Chawla, N.V., Bowyer, K.W., Hall, L.O. y Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence research, pp.321-357.

Deville, J.C. y Tillé, Y., 2004. Eficient balanced sampling: The cube method. Biometrika, 91: pp 893_912.

Domingos, P., 1999. MetaCost A general method for making classifiers cost-sensitive. In: Fifth International Conference on Knowledge Discovery and Data Mining, pp.155-164.

Elkan, C., 2001. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Conference of Artificial Intelligence, 973-978. Seattle, Washington: Morgan Kaufmann.

Gopinathan, K., O'Donnell, D., 1998. Just in time risk management. Credit World, 2:10-2.

Hand, D.J. y Henley, W.E., 1997. Statistical Classification. Methods in Costumer Credit Scoring: A review. Journal of the Royal Statistical Association, 160(A/ Part3), 523-541.

Han, H., Wang, W. y Mao, B., 2005. Borderline-SMOTE: a new Over-Sampling Method in Imbalanced Data Sets Learning. En: D.-S. Huanng; X.-P.Zhzng y G.-B. Huang (Eds.), ICICS, volumen 3644 de LNCS, pp. 878-887.

Hanley, J.A. y McNeil, B.J., 1982. The meaning and use of the área under receiver operating characteristic (ROC) curve. Radilogy, 143, 29-36.

Hulse, J.V., Khoshgoftaar, T.M. y Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. En: Z. Ghahramani (Rd.), ICML volume 227 de ACM International Conference Proceeding series, pp. 935-942.

Japkowicz, N., 2001. Concept-Learning in the Presence of Between-Class and Within-Class Imbalances. En: E. Stroulia y S. Matwin (Eds.), Canadian Conference on AI, volume 2056 de LNCS, pp. 67-77.

Japkowicz, N. y Stephen, S., 2002. The Class Imbalance Problem: A Systematic Study Intelligent Data. Analysis, Journal, volume 6, issue 5, pp: 1-32.

Jensen HL., 1992. Using neural networks for credit scoring. Managerial Finance, 18:15-26.

Jorion, P., 2000. Valor en Riesgo, segundo. EDN, McGraw-Hill, Nueva York. Mester, L.J. (1997). What's the Point of Credit Scoring? Business Review, Set./Oct., pp. 3-16, Federal Reserve Bank of Philadelphia.

Kubat, M. y Matwin, S., 1997. Addressing the Course of Imbalanced Training Sets: One-Sided Selection. En: D.H. Fisher (Ed.), ICML, pp. 179-186.

Kuncheva, L. y Jain. L.C., 1999. Nearest neighbor classifier: Simultaneous editing and feature selection. Pattern Recognition Letters, pp. 1149-1156.

Laurikkala, J., 2002. Instance-based data reduction for improved identification of difficult small classes. Intelligent Data Analysis, pp.311-322.

Ling, C.X., Yang, Q., Wang, J. y Zhang, S., 2004. Decision trees with minimal costs. In ICML '04: Proceedings of the twenty-first international conference on Machine earning, page 69.

Lizotte, D., Madani, O. y Greiner, R., 2003. Budgeted Learning of Naïve-Bayes Classifiers. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence. Acapulco, Mexico: Morgan Kaufmann.

López, V., Fernández, A. y Herrera, F., 2010. Un primer estudio sobre el uso de aprendizaje sensible al coste con sistemas de clasificación basados en reglas difusas para problemas no balanceados. In Proceedings of the III Congreso Español de Informática (CEDI 2010). III Simposio sobre Lógica Fuzzy y Soft Computing, LFSC2010 (EUSFLAT), Valencia (Spain), 459-466.

López, V., Fernández, A., García, S., Palade, V., y Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

Provost, F., 2003. Machine learning from imbalanced data sets 101 (Extended Abstract). En: AAAI: Workshop on Learning with Imbalanced Data Sets.

Provost, F. y Fawcett. T., 2001. Robust classification for imprecise environments. Machine Learning Journal, 42(3):203-231.

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

Swets, J.A. y Pickett, R.M., 1982. Evaluation of Diagnostic system. Academic Press, Inc, New York.

Ting, K.M., 1998. Inducing cost-sensitive trees via instance weighting. En Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, pages 139-147.

Turney, P.D., 1995. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. Journal of Artificial Intelligence Research 2:369-409.

Turney, P.D., 2000. Types of cost in inductive concept learning. In Proceedings of the Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning, Stanford University, California.

Wang, J., Xu, M., Wang, H. y Zhang, J., 2006. Clasification of Imbalanced Data by Using the SMOTE Algorithm and locally Linear Embedding. En: ICSP, volume 3, pp. 16-20.

West, D., 2000. Neural network credit scoring models, Computers & Operations Research, vol. 27, pp. 1131–1152.

Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data, IEEE Transactions on Systems, Man and Cybernetics. IEEE Computer Society Press, Los Alamos.

Witten, I.H. y Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Zadrozny, B. y Elkan, C., 2001. Learning and Making Decisions When Costs and Probabilities are Both Unknown. In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, 204-213.

Zhang, J. y Mani, I., 2003. kNN approach to unbalanced data distributions: a case study involving information extraction. En ICML: Workshop on Learning from Imbalanced Dataset II.

Zweig, M.H. y Cambell, G., 1993. Receiver–Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical. Medicine. Clin. Chem., 39 (4), 561-577. [Correcciones en Clin. Chem., (1993), 39, 1589].