BONDAD DE AJUSTE Y ELECCIÓN DEL PUNTO DE CORTE EN REGRESIÓN LOGÍSTICA BASADA EN DISTANCIAS. APLICACIÓN AL PROBLEMA DE *CREDIT SCORING*.[†]

Teresa Costa Cor[1], Eva Boj del Val[2] y José Fortiana Gregori[3]

ABSTRACT

The goal of this paper is finding and evaluating criteria for choosing an adequate group assignation cut point from probabilities predicted by the distance-based logistic regression model, with application to credit scoring problems. Goodness-of-fit is assessed by means of the Kolmogorov-Smirnov and Gini index statistics, in concert with the ROC curve. Resulting misclassification probabilities and error cost functions are evaluated. Applications to real datasets, namely two credit risk portfolios, illustrate the procedure. Computations are performed with the *dbstats* package in the R environment.

KEY WORDS: Credit scoring; Distance-based logistic regression; cut-off point; ROC curve; Probability of default, dbstats.

RESUMEN

En este trabajo se estudian criterios para la elección de un punto de corte adecuado en el modelo de regresión logística basado en distancias. Todo ello con aplicación al problema de *credit scoring*. Los criterios de calidad de ajuste que se analizan son el coeficiente Kolmogorov-Smirnov y el índice de Gini, junto con la representación gráfica ROC. También se calculan las probabilidades de mala clasificación y unas funciones de coste del error. Se

[†] Trabajo financiado por el Ministerio de Educación y Ciencia, proyecto número MTM2010-17323, y por la Generalitat de Catalunya, AGAUR, proyecto número 2009SGR970.

[1] Autora de correspondencia. Profesora Titular de Escuela Universitaria. Departamento de Matemática Económica, Financiera y Actuarial. Facultad de Economía y Empresa. Universidad de Barcelona. Avenida Diagonal 690, 08034_Barcelona. España. E-mail: tcosta@ub.edu

[2] Profesora Titular de Universidad. Departamento de Matemática Económica, Financiera y Actuarial. Facultad de Economía y Empresa. Universidad de Barcelona. Avenida Diagonal 690, 08034_Barcelona. España. E-mail: evaboj@ub.edu

[3] Profesor Titular de Universidad. Departamento de Probabilidad, Lógica y Estadística. Facultad de Matemáticas. Universidad de Barcelona. Gran Vía de las Cortes Catalanas 595, 08007_Barcelona. España. E-mail: fortiana@ub.edu

Este artículo ha sido recibido en versión revisada el 24 de septiembre de 2012.

realiza la aplicación a dos carteras de datos reales de riesgo de crédito haciendo uso del paquete *dbstats* de R.

1. INTRODUCCIÓN

En trabajos anteriores (ver Boj *et al.*, 2009b, 2011) se explica con detalle la importancia para las Entidades Financieras de realizar un cálculo preciso de las primas por riesgo de crédito. Estas primas se calculan haciendo uso de las probabilidades de insolvencia de los riesgos a partir de un modelo de *credit scoring*. En Boj *et al.* (2009b) se propuso la aplicación de análisis discriminante basado en distancias (BD) en este problema y, posteriormente, en Boj *et al.* (2011) se propuso la aplicación de regresión logística BD (presentada inicialmente en Boj *et al.*, 2008), ésta última utilizando un punto de corte de 0.5 para el cálculo de las matrices de confusión.

Este trabajo se centra nuevamente en la técnica de regresión logística BD, ampliando su estudio para diferentes puntos de corte dentro del intervalo (0,1). La justificación de este objetivo teórico del modelo está en que no siempre es adecuado utilizar el punto de corte 0.5 si contamos con datos reales no balanceados. En *credit scoring* es el caso en que los individuos insolventes no suponen aproximadamente la mitad de la cartera.

Se analizan como medidas de calidad de ajuste el coeficiente Kolmogorov-Smirnov (K-S) y el índice de Gini, y se realiza la representación gráfica de la curva ROC obtenida con regresión logística BD. En la curva ROC se representan los resultados para diferentes puntos de corte teniendo en cuenta el coeficiente K-S, cuyo máximo genera un óptimo según dicho criterio. El índice de Gini se calcula únicamente como medida global del modelo, pues tiene en cuenta todos los puntos de corte adecuados o no.

Además, para completar el estudio se calculan las probabilidades de mala clasificación y las funciones de coste del error que se describieron en Boj *et al.* (2009b) como criterios de elección de modelo de *credit scoring*, aquí para diferentes puntos de corte.

El estudio de sensibilidad que se realiza en este artículo de las diferentes medidas ante cambios en el punto de corte con variaciones entre 0 y 1 para el modelo de regresión logística BD, tiene su motivación en las sugerencias recibidas durante el congreso *RISK2011* celebrado en Sevilla y en el cuál se presentó el trabajo de Boj *et al*. (2011). Se agradecen los comentarios

recibidos por parte de los participantes ya que han ayudado a completar este estudio.

Se realizan dos aplicaciones con conjuntos de datos reales de riesgo de crédito, los cuales pueden ser descargados gratuitamente junto con su descripción del repositorio *Machine Learning Repository*[4]. Ambas carteras pertenecen a Entidades Financieras, la primera, "*Statlog (Australian Credit Approval)*" [5], a una Financiera australiana y la segunda, "*Statlog (German Credit Data)*" [6], a una alemana.

Estos datos fueron analizados en Boj *et al.* (2009b) y en Boj *et al.* (2011) con análisis discriminante BD y regresión logística BD para un punto de corte de 0.5, obteniendo como resultado que el ajuste de ambas técnicas es competitivo frente al de otros modelos de *credit scoring* [7] (ver Tablas 1, 2, 3 y 4 en ambas referencias).

Cabe destacar como novedad que en este estudio los cálculos se realizan haciendo uso de la función "dbglm" del paquete *dbstats* de R (Boj *et al.*, 2012).

El trabajo está estructurado del siguiente modo: en el apartado 2 se presentan las medidas de calidad de ajuste y la curva ROC resultante con regresión logística BD y se aplica a los dos conjuntos de datos de riesgo de crédito; en los apartados 3 y 4 se calculan respectivamente las probabilidades de mala clasificación y los costes del error para diferentes puntos de corte; finalmente, en el apartado 5, se exponen las principales conclusiones y aportaciones del trabajo.

2. CALIDAD DEL MODELO: CÁLCULO DEL COEFICIENTE KOLMOGOROV-SMIRNOV E ÍNDICE DE GINI. REPRESENTACIÓN GRÁFICA DE LA CURVA ROC.

El modelo de regresión logística BD (ver Boj *et al.*, 2008 y 2011 para el detalle sobre el algoritmo iterativo de estimación por mínimos cuadrados

[4] http://archive.ics.uci.edu/ml/datasets.html
[5] http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
[6] http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
[7] Métodos no-paramétricos como las redes neuronales, el método de los *k* vecinos más próximos, el método de la estimación núcleo de la densidad y el árbol de clasificación *classification and regression trees* (CART); y Métodos paramétricos como el análisis discriminante lineal y la regresión logística clásica.

ponderados) es una versión de la regresión logística clásica en el ámbito no paramétrico y BD. Es no paramétrico y BD puesto que la única información requerida en el espacio de los predictores es una matriz de distancias al cuadrado, *D2*, calculada mediante una función de distancias a partir de los predictores originales, usualmente de tipo mixto en el caso del riesgo de crédito. La variable respuesta en *credit scoring* se construye codificando con 1 a los individuos que han resultado insolventes en el periodo de estudio y con 0 a los que no. La predicción de las probabilidades de insolvencia para cada uno de los n individuo en la población ω se estima como:

$$\hat{\pi}(\omega) = \hat{\mu}(\omega) = \frac{e^{\hat{\eta}(\omega)}}{1 + e^{\hat{\eta}(\omega)}}.$$

Finalmente, dado un punto de corte *s*, la matriz de confusión se calcula como:

		Estimada		
		Buenos riesgos	Malos riesgos	Total
Real	Buenos riesgos	n_{11}^s	n_{21}^s	$n_{11}^s + n_{21}^s$
	Malos riesgos	n_{12}^s	n_{22}^s	$n_{12}^s + n_{22}^s$
	Total	$n_{11}^s + n_{12}^s$	$n_{21}^s + n_{22}^s$	n

Con estos resultados calcularemos los criterios de calidad de ajuste con los que elegir un punto de corte "óptimo" para unos datos determinados.

En el caso de riesgo de crédito, para la determinación del punto de corte o valor del *score s* a partir del cual decidir si un cliente es un mal riesgo de crédito, podemos utilizar la denominada curva ROC (*Receiver Operating Characteristic o Característica Operativa del Receptor*). La curva ROC fue desarrollada inicialmente por ingenieros para la estimación de errores en la transmisión de mensajes y se ha aplicado posteriormente en áreas como la medicina y la estadística.

A partir de la matriz de confusión calculada con el modelo de regresión logística BD se calculan las razones de verdaderos positivos y falsos positivos. En nuestro caso, los verdaderos positivos son aquellos malos riesgos predichos como malos (en la matriz de confusión el elemento n_{22}^s) y los falsos positivos son aquellos buenos riesgos predichos como malos (en la

matriz de confusión el elemento n_{21}^s). Gráficamente, en un espacio ROC, se pueden representar los intercambios entre verdaderos positivos (eje de ordenadas) y falsos positivos (eje de abcisas).

Otra interpretación posible de la curva ROC es la de representar la Sensibilidad (eje de ordenadas) frente a (1 – Especificidad) (eje de abcisas), como se indica en Reyes *et al*. (2007).

Para su cálculo, en función de los elementos de la matriz de confusión, tenemos:

- Sensibilidad: $\frac{n_{22}^s}{n_{12}^s + n_{22}^s}$ representa la proporción de malos riesgos predichos como malos.
- Especificidad: $\frac{n_{11}^s}{n_{11}^s + n_{21}^s}$ representa la proporción de buenos riesgos predichos como buenos.

En el modelo de regresión logística BD, según se van variando los puntos de corte o frontera, *s*, se obtienen los distintos puntos que conforman la curva ROC.

Para medir la calidad del modelo de *credit scoring* es usual utilizar índices cuantitativos como el índice de Gini o el coeficiente K-S, que se basan en la función de distribución o probabilidades acumuladas. El índice de Gini es únicamente una medida global de calidad del modelo, mientras que en el coeficiente K-S, a parte de medir la calidad de ajuste, identifica el valor del *score* para el cual se maximiza dicho coeficiente, y es ”ideal” si el punto de corte “esperado” es cercano a dicho *score* (Řezáč y Řezáč, 2011).

En este trabajo se representa la curva ROC para obtener gráficamente el punto de corte que permite maximizar el coeficiente K-S. El procedimiento de construcción de la curva es el siguiente (Íñiguez y Morales, 2009):

a) Ordenar los valores de los puntos de corte, *s*, de manera ascendente.

b) Calcular la proporción de buenos y malos riesgos que comparten el mismo punto de corte, $p_b(s)\ y\ p_m(s)$, siendo:

$$p_b(s) = \frac{n_{11}^s + n_{12}^s}{\sum_s n_{11}^s + n_{12}^s} \qquad \text{y} \qquad p_m(s) = \frac{n_{21}^s + n_{22}^s}{\sum_s n_{21}^s + n_{22}^s}\ .$$

c) Calcular la proporción acumulada de buenos y malos riesgos $P_b(s)$ y $P_m(s)$:

$$P_b(s) = \sum_{S \le s} p_b(S) \qquad \text{y} \qquad P_m(s) = \sum_{S \le s} p_m(S).$$

d) Calcular las diferencias entre proporciones acumuladas por punto de corte entre buenos y malos riesgos: $\left|P_m(s) - P_b(s)\right|$.

e) Identificar el punto de corte $\hat{s}^*$ que proporciona la máxima diferencia absoluta del coeficiente K-S: $K-S = \underset{s}{máx}\left\{\left|P_m(s) - P_b(s)\right|\right\}$.

Por otro lado, como medida de calidad global del modelo se calcula el índice de Gini, que se puede calcular a partir de la siguiente expresión (Íñiguez y Morales, 2009):

$$Gini = 1 - \sum_{i=1}^{n} \left(P_m(s_i) - P_m(s_{i-1})\right) \cdot \left(P_b(s_i) + P_b(s_{i-1})\right),$$

$$P_m(s_0) = 0,\ P_b(s_0) = 0,$$

donde:

$P_m(s_i)$: proporción acumulada de malos riesgos para un *score* s_i

$P_m(s_{i-1})$: proporción acumulada de malos riesgos para el *score* anterior a s_i

$P_b(s_i)$: proporción acumulada de buenos riesgos para un score s_i

$P_b(s_{i-1})$: proporción acumulada de buenos riesgos para el *score* anterior a s_i

El modelo "ideal", es decir, que predice exactamente los buenos y malos riesgos, tendría un índice de Gini igual a 1; en caso contrario, el modelo asignaría un *score* aleatorio al cliente y tendría un índice de Gini igual a 0 (Řezáč y Řezáč, 2011).

En resumen, en este trabajo se maximiza el coeficiente K-S para la obtención de un punto de corte "óptimo", $\hat{s}^*$, según dicho criterio. El coeficiente K-S se representa gráficamente en una curva ROC. Por otro lado, el índice de Gini se calcula como medida global de calidad de ajuste, pues tiene en cuenta los mismos datos que el coeficiente K-S para todo el repertorio de puntos de corte entre 0 y 1. El índice de Gini no ofrece un punto de corte óptimo, pero puede resultar útil para comparar distintos modelos para un mismo conjunto de datos. En las aplicaciones de este trabajo sólo se indica

su valor numérico como dato ilustrativo del procedimiento de cálculo para el modelo BD.

2.1. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO AUSTRALIANOS

Estos datos hacen referencia al riesgo asociado a tarjetas de crédito de una Entidad Financiera. Para mantener la confidencialidad, el autor no cedió los nombres de los factores de riesgo ni lo que significan sus clases y valores. La base de datos es de especial interés porque el conjunto de predictores es de tipo mixto y el número de datos faltantes es reducido. Tal y como se explica en Boj *et al*. (2009b), para las variables continuas los datos faltantes fueron re-emplazados por la media de la variable correspondiente, y para las variables categóricas y binarias éstos fueron re-emplazados por la moda. En total contiene $n = 690$ individuos, de los cuales 307 fueron buenos riesgos y 383 malos. Los factores potenciales de riesgo son 14, de los cuales 6 son continuos, 4 categóricos y 4 binarios. A partir de los 14 predictores mixtos, calculamos la matriz de distancias, *D2*, como la suma pitagórica dada por la fórmula (1) de Boj *et al.* (2009b) teniendo en cuenta el índice de similaridad de Gower (Gower, 1971) para cada una de las variables individualizadas.

Estimamos la regresión logística BD haciendo uso de la función "dbglm" del paquete de R *dbstats* (Boj *et al.*, 2012) especificando que la distribución del error es Binomial y que el link es el canónico *logit* mediante la instrucción:

```
> dbglmaus <- dbglm(D2, y, family = binomial (link = "logit"), maxiter =
50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99); dbglmaus

Call:  dbglm.D2(D2 = D2, y = y, family = binomial(link = "logit"), maxiter =
50,     eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99)

family: binomial

Degrees of Freedom: 689 Total (i.e. Null);  556 Residual
Null Deviance:      948.2
Residual Deviance: 248.4        AIC: 516.4
```

y obtenemos la estimación en la variable dbglmaus$fitted.values, con la que se calculan las matrices de confusión para diferentes puntos de corte.

En la siguiente tabla, Tabla 1, se presentan los cálculos del coeficiente K-S para representar la curva ROC (Figura 1), para distintos puntos de corte con los datos australianos:

Punto de corte	**Cálculo del coeficiente K-S**		
	Proporción de buenos riesgos acumulados	**Proporción de malos riesgos acumulados**	**K-S**
0.05	0.0152	0.0392	0.0240
0.1	0.0341	0.0755	0.0414
0.15	0.0557	0.1095	0.0538
0.2	0.0783	0.1428	0.0645
0.25	0.1024	0.1750	0.0726
0.3	0.1278	0.2060	0.0782
0.35	0.1537	0.2368	0.0831
0.4	0.1801	0.2671	0.0870
0.45	0.3163	0.4151	0.0988
0.5	0.4563	0.5603	0.1040
0.55	0.6015	0.7013	0.0998
0.6	0.7504	0.8394	0.0890
0.65	0.7817	0.8658	0.0841
0.7	0.8139	0.8914	0.0775
0.75	0.8468	0.9166	0.0698
0.8	0.8815	0.9403	0.0588
0.85	0.9178	0.9627	0.0449
0.9	0.9566	0.9831	0.0265
0.95	1	1	0

Tabla 1. Cálculo del coeficiente K-S para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito australianos.

Analizando con más detalle en la Tabla 2 los valores del K-S para los puntos de corte del intervalo [0.45, 0.55] , donde se observan los valores más altos de la Tabla 1, se obtiene que el punto de corte que maximiza el K-S es igual a 0.51.

Punto de corte	Cálculo del K-S		
	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	K-S
0.45	0.3163	0.4151	0.0988
0.46	0.3439	0.4445	0.1006
0.47	0.3717	0.4737	0.1020
0.48	0.3996	0.5028	0.1032
0.49	0.4279	0.5316	0.1037
0.5	0.4563	0.5603	0.1040
0.51	0.4848	0.5889	**0.1041**
0.52	0.5134	0.6174	0.1040
0.53	0.5426	0.6455	0.1029
0.54	0.5720	0.6734	0.1014
0.55	0.6015	0.7013	0.0998

Tabla 2. Cálculo del coeficiente K-S para diferentes puntos de corte en el intervalo [0.45, 0,55], con los datos de riesgo de crédito australianos.

En el siguiente gráfico, Figura 1, se representa la curva ROC y el punto de corte que maximiza el K-S, que se corresponde con el punto en la curva ROC cuya distancia horizontal al eje es máxima (Balzarotti y Castelpoggi, 2009):

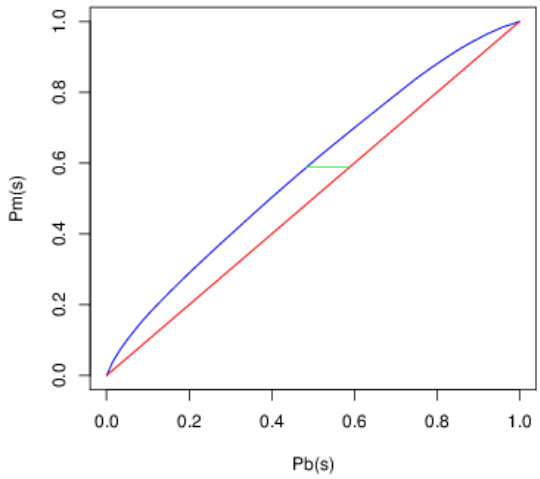

Figura 1. Curva ROC para los datos de riesgo de crédito australianos. Se obtiene un punto de corte óptimo de 0.51.

En cuanto al índice de Gini, para su cálculo, tal como hemos indicado en el apartado 1, son necesarias las proporciones acumuladas de buenos y malos riesgos, es decir, los mismos datos que se han utilizado para el cálculo del coeficiente K-S. En los datos de riesgo de crédito australianos el valor obtenido en el índice de Gini es 0.16. Este es un valor pequeño, pero similar al que se obtiene con regresión logística clásica[8] que es de 0.20. Recordemos que para estos datos si estudiamos con detalle los resultados obtenidos en las Tablas 1 y 2 de Boj *et al*. (2011) o en las Tablas 3 y 4 de Boj *et al.* (2009b) la regresión logística en sus dos versiones es uno de los métodos competitivos, por lo que aunque los valores obtenidos no son elevados, éstos son un indicador del ajuste global para dichos modelos.

2.2. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO ALEMANES

Estos datos clasifican a un conjunto de individuos como buenos o malos riesgos en función de una serie de predictores de tipo mixto. La cartera contiene datos cedidos en fecha 17-11-1994. En total contiene n = 1000 individuos, de los cuales 700 han sido buenos riesgos y 300 malos. Los factores potenciales de riesgo considerados son 20, de los cuales 7 son continuos, 11 categóricos y 2 binarios. Al igual que en la primera aplicación, se utiliza el índice de similaridad de Gower en el cálculo de *D2* y se estima el modelo haciendo uso de la función "dbglm" del paquete *dbstats*:

```
> dbglmger <- dbglm (D2, y, family = binomial(link = "logit"), maxiter =
50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99); dbglmger

Call:  dbglm.D2(D2 = D2, y = y, family = binomial(link = "logit"), maxiter =
50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99)

family: binomial

Degrees of Freedom: 999 Total (i.e. Null);  860 Residual
Null Deviance:      1222
Residual Deviance: 780.1        AIC: 1060
```

obteniendo la estimación en la variable dbglmger$fitted.values, y con ella las matrices de confusión para diferentes puntos de corte.

[8] Calculado con la función "dbglm" introduciendo como input la matriz de distancias *D2* calculada a partir de la función Eclídea y tratando a los predictores categóricos y binarios como factores.

En este apartado se detalla el cálculo del coeficiente K-S y su representación gráfica en la curva ROC y se obtiene el índice de Gini.

En la siguiente tabla, Tabla 3, se presentan los cálculos del coeficiente K-S para representar la curva ROC (Figura 2), para distintos puntos de corte con los datos alemanes:

Punto de corte	**Cálculo del K-S**		
	Proporción de buenos riesgos acumulados	**Proporción de malos riesgos acumulados**	**K-S**
0.05	0.0075	0.0828	0.0753
0.1	0.0199	0.1531	0.1332
0.15	0.0356	0.2150	0.1794
0.2	0.0546	0.2681	0.2135
0.25	0.0761	0.3150	0.2389
0.3	0.0994	0.3571	0.2577
0.35	0.1244	0.3948	0.2704
0.4	0.1510	0.4286	0.2776
0.45	0.2872	0.5885	0.3013
0.5	0.4317	0.7271	0.2954
0.55	0.5856	0.8416	0.2560
0.6	0.7445	0.9432	0.1987
0.65	0.7781	0.9588	0.1807
0.7	0.8125	0.9721	0.1596
0.75	0.8481	0.9826	0.1345
0.8	0.8847	0.9906	0.1059
0.85	0.9223	0.9959	0.0736
0.9	0.9609	0.9987	0.0378
0.95	1	1	0

Tabla 3. Cálculo del coeficiente K-S para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito alemanes.

Punto de corte	Cálculo del K-S		
	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	K-S
0.40	0.1510	0.4286	0.2776
0.41	0.1776	0.4623	0.2847
0.42	0.2045	0.4950	0.2905
0.43	0.2319	0.5268	0.2949
0.44	0.2544	0.5581	0.2987
0.45	0.2872	0.5885	0.3013
0.46	0.3153	0.6183	0.3030
0.47	0.3437	0.6473	**0.3036**
0.48	0.3725	0.6753	0.3028
0.49	0.4018	0.7020	0.3002
0.50	0.4317	0.7271	0.2954

Tabla 4. Cálculo del coeficiente K-S para diferentes puntos de corte en el intervalo [0.4, 0.5], con los datos de riesgo de crédito alemanes.

Si calculamos los valores del coeficiente K-S para los puntos de corte del intervalo [0.4, 0.5], donde los valores de la Tabla 3 son mayores, se obtiene que, en este caso, el valor máximo del K-S está en el punto de corte 0.47 (ver Tabla 4).

La representación gráfica, Figura 2, de la curva ROC y del valor que maximiza el K-S para estos datos es:

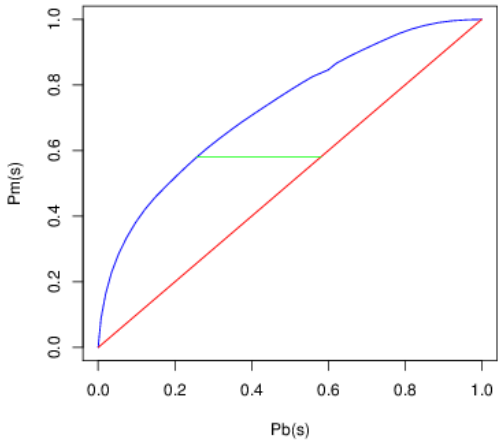

Figura 2. Curva ROC para los datos de riesgo de crédito alemanes. Se obtiene un punto de corte óptimo de 0.47.

El índice de Gini que se obtiene para los datos de riesgo de crédito alemanes es igual a 0.44.

3. ESTUDIO DE LAS PROBABILIDADES DE MALA CLASIFICACIÓN EN FUNCIÓN DEL PUNTO DE CORTE

En este apartado se estudia el comportamiento de las probabilidades de mala clasificación: de buenos riesgos, de malos riesgos y global para los distintos puntos de corte entre 0 y 1. En Boj *et al*. (2009b y 2011) se utilizaron dichas probabilidades para la comparación de modelos de *credit scoring*.

Se calculan, para los dos conjuntos de datos, las probabilidades de mala clasificación con puntos de corte que toman valores desde 0.05 hasta 0.95 con incrementos de 0.05. De este modo es posible observar la evolución de dichas probabilidades en función de los diferentes puntos de corte.

3.1. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO AUSTRALIANOS

Los resultados para los datos australianos de riesgo de crédito se muestran en la Tabla 5.

Si nos fijamos en la probabilidad de mala clasificación global los mejores resultados, es decir, las menores probabilidades de mala clasificación, se encuentran en el punto 0.5.

Las probabilidades de mala clasificación de buenos riesgos, malos riesgos y global para el punto de corte de 0.51 son 0.081, 0.060 y 0.070 respectivamente.

Punto de corte	**Probabilidades estimadas de mala clasificación**		
	Buenos riesgos	**Malos riesgos**	**Global**
0.05	0.479	0.005	0.216
0.1	0.349	0.005	0.195
0.15	0.261	0.010	0.122
0.2	0.225	0.010	0.106
0.25	0.179	0.013	0.087
0.3	0.143	0.021	0.075
0.35	0.134	0.026	0.074
0.4	0.127	0.037	0.077
0.45	0.098	0.044	0.068
0.5	0.081	0.055	**0.067**
0.55	0.065	0.073	0.070
0.6	0.062	0.084	0.074
0.65	0.046	0.107	0.080
0.7	0.036	0.123	0.087
0.75	0.029	0.138	0.090
0.8	0.026	0.180	0.112
0.85	0.020	0.227	0.135
0.9	0.013	0.292	0.168
0.95	0	0.407	0.226

Tabla 5. Probabilidades estimadas de mala clasificación para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito australianos.

3.2. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO ALEMANES

Para los datos alemanes, los resultados de las probabilidades de mala clasificación considerando puntos de corte que varían desde 0.05 hasta 0.95 con incrementos de 0.05, quedan recogidos en la Tabla 6.

Las menores probabilidades de mala clasificación global se encuentran en el punto 0.55.

Las probabilidades de mala clasificación de buenos riesgos, malos riesgos y global para el punto de corte de 0.47 son 0.116, 0.323 y 0.178 respectivamente.

Punto de corte	**Probabilidades estimadas de mala clasificación**		
	Buenos riesgos	**Malos riesgos**	**Global**
0.05	0.733	0.010	0.516
0.1	0.567	0.030	0.406
0.15	0.463	0.063	0.343
0.2	0.361	0.110	0.286
0.25	0.287	0.143	0.244
0.3	0.239	0.183	0.222
0.35	0.186	0.203	0.191
0.4	0.159	0.263	0.189
0.45	0.127	0.303	0.180
0.5	0.086	0.380	0.174
0.55	0.059	0.420	**0.167**
0.6	0.043	0.477	0.173
0.65	0.031	0.563	0.191
0.7	0.017	0.607	0.194
0.75	0.014	0.690	0.217
0.8	0.010	0.763	0.236
0.85	0.006	0.840	0.256
0.9	0.001	0.913	0.275
0.95	0	0.957	0.287

Tabla 6. Probabilidades estimadas de mala clasificación para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito alemanes.

4. ESTUDIO DE LOS COSTES DEL ERROR EN FUNCIÓN DEL PUNTO DE CORTE

Por otro lado, utilizando las matrices de confusión se estudia el comportamiento de los costes del error dados por las fórmulas (5) y (6) de Boj *et al.* (2009b) para los distintos puntos de corte entre 0 y1. Recordemos que en Boj *et al*. (2009b y 2011) se aplicaron para la elección de un modelo predictivo en el cálculo de *scorings*.

4.1. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO AUSTRALIANOS

Para los datos australianos se calculan los costes del error utilizando puntos de corte que varíen entre 0.05 y 0.95 con incrementos de 0.05. Los resultados obtenidos se recogen en la Tabla 7.

Se observa que, en el escenario en que la probabilidad *a priori* es de 0.144, el mínimo coste se obtiene en el punto de corte 0.5, con un valor de 0.105, mientras que en el otro escenario, el mínimo coste es de 0.116 y se consigue cuando el punto de corte es de 0.3.

Los costes en los dos escenarios para el punto de corte de 0.51 son de 0.110 y 0.143 respectivamente.

Punto de corte	**Costes estimados**	
	$\pi_2 = 0.144$	$\pi_2 = 0.249$
0.05	0.247	0.224
0.1	0.195	0.177
0.15	0.162	0.152
0.2	0.144	0.136
0.25	0.122	0.120
0.3	0.111	**0.116**
0.35	0.111	0.119
0.4	0.118	0.134
0.45	0.106	0.129
0.5	**0.105**	0.135
0.55	0.110	0.151
0.6	0.116	0.163
0.65	0.122	0.182
0.7	0.130	0.201
0.75	0.131	0.208
0.8	0.156	0.252
0.85	0.178	0.294
0.9	0.207	0.347
0.95	0.243	0.419

Tabla 7. Costes estimados para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito australianos.

4.2. APLICACIÓN CON LOS DATOS DE RIESGO DE CRÉDITO ALEMANES

Para los datos alemanes se presentan los costes del error en los dos escenarios calculados para puntos de corte que varíen desde 0.05 hasta 0.95 con incrementos de 0.05. Los resultados se recogen en la Tabla 8.

Se obtiene que los puntos de corte en los que se minimiza el coste son de 0.95 en el primer escenario y de 0.7 en el segundo.

Los costes en los dos escenarios para el punto de corte de 0.47 son 0.342 y 0.383 respectivamente.

Punto de corte	**Costes estimados**	
	$\pi_2 = 0.144$	$\pi_2 = 0.249$
0.05	0.554	0.495
0.1	0.515	0.469
0.15	0.493	0.462
0.2	0.466	0.451
0.25	0.433	0.428
0.3	0.414	0.421
0.35	0.371	0.385
0.4	0.369	0.397
0.45	0.349	0.386
0.5	0.318	0.371
0.55	0.279	0.343
0.6	0.264	0.339
0.65	0.267	0.357
0.7	0.230	**0.330**
0.75	0.249	0.360
0.8	0.256	0.377
0.85	0.257	0.389
0.9	0.234	0.378
0.95	**0.209**	0.362

Tabla 8. Costes estimados para diferentes puntos de corte en el intervalo (0, 1), con los datos de riesgo de crédito alemanes.

5. PRINCIPALES CONCLUSIONES Y APORTACIONES

En este trabajo se completa el estudio de aplicación del modelo de regresión logística BD en el problema de riesgo de crédito iniciado en Boj *et al*. 2011. Por un lado, de forma simétrica al procedimiento con el modelo clásico de regresión logística por mínimos cuadrados ordinarios, se analizan criterios teóricos de calidad de modelo para la elección de un punto de corte adecuado cuando los datos tratados no son balanceados y, por otro lado, se enfoca el estudio en aplicación al problema del cálculo del punto de corte "óptimo" en *credit scoring*.

Se propone el uso de la curva ROC para representar el punto de corte donde se maximiza el coeficiente K-S construida para el modelo de regresión logística BD y el índice de Gini como medidas de calidad de ajuste. El máximo en el coeficiente K-S da el punto de corte "óptimo" y el índice de Gini ofrece una medida de calidad global del modelo. Además, se analiza el comportamiento de las probabilidades de mala clasificación y de los costes del error (propuestos en Boj *et al*. 2009b y 2011 como criterio de elección de modelo en *credit scoring*) en función de los diferentes puntos de corte entre 0 y 1.

Con las dos carteras tratadas en este trabajo ya se habían obtenido buenos resultados con regresión logística BD y punto de corte 0.5 en comparación con otras técnicas de *credit scoring* (ver Tablas 1, 2, 3 y 4 de Boj *et al.*. 2011). Pero a raíz de las sugerencias recibidas en el congreso *RISK2011* celebrado en Sevilla, se mejora y amplía el estudio para estas carteras frente a diferentes puntos de corte. Puesto que ambos datos son bastante balanceados, se obtiene un resultado sobre el punto de corte "óptimo" con el criterio K-S que oscila el 0.5 como era de esperar. A continuación listamos las principales conclusiones para cada uno de los conjuntos por separado:

Para los datos australianos, el punto de corte "esperado" era de 0.55, ya que $n = 690$, de los cuales 307 eran buenos riesgos y 383 malos. Se obtienen los siguientes resultados:

- El punto de corte que maximiza el coeficiente K-S es de 0.51.
- El índice de Gini de calidad global del modelo es de 0.16.

Se obtiene que el punto de corte que minimiza las probabilidades globales de mala clasificación es de 0.5 con una probabilidad de 0.067. Las probabilidades de mala clasificación de buenos riesgos, malos riesgos y

global para el punto de corte 0.51 obtenido con el coeficiente K-S son 0.081, 0.060 y 0.070 respectivamente.

Los puntos de corte que minimizan el coste cuando las probabilidades *a priori* de malos riesgos son respectivamente de 0.144 y de 0.249 son de 0.5 y 0.3. Los costes calculados en los dos escenarios para el punto de corte de 0.51 obtenido con el coeficiente K-S son de 0.110 y 0.143 respectivamente.

Tanto si comparamos las probabilidades de mala clasificación, de malos riesgos y global, como si comparamos los costes del error en los dos escenarios para el punto de corte 0.51 obtenido con el coeficiente K-S, la regresión logística BD sigue siendo la técnica con menores valores y por lo tanto la mejor en las Tablas 1 y 2 de Boj *et al*. (2011).

Para los datos alemanes, el punto de corte "esperado" era de 0.3, ya que n = 1000, de los cuales 700 eran buenos riesgos y 300 malos, unos datos algo menos balanceados que los anteriores. Se obtienen los siguientes resultados:

- El punto de corte que maximiza el coeficiente K-S es de 0.47.
- El índice de Gini de calidad global del modelo es de 0.44.

Se obtiene que el punto de corte que minimiza las probabilidades globales de mala clasificación es de 0.55 con una probabilidad de 0.167. Las probabilidades de mala clasificación de buenos riesgos, malos riesgos y global para el punto de corte de 0.47 son 0.116, 0.323 y 0.178 respectivamente.

Los puntos de corte que minimizan el coste cuando las probabilidades *a priori* de malos riesgos son respectivamente de 0.144 y de 0.249 son de 0.95 y 0.7. Los costes calculados en los dos escenarios para el punto de corte de 0.47 son de 0.342 y 0.383 respectivamente.

Tanto si comparamos la probabilidad de mala clasificación global como los costes del error en los dos escenarios para el punto de corte 0.47 obtenido con el coeficiente K-S, la regresión logística BD sigue siendo la técnica con menores valores y por lo tanto la mejor en las Tablas 3 y 4 de Boj *et al*. (2011). Si comparamos las probabilidades de mala clasificación de malos riesgos para el punto de corte 0.47 obtenido con el coeficiente K-S, la regresión logística BD es la segunda mejor técnica después del análisis discriminante clásico y BD.

Si se comparan visualmente las curvas ROC (ver Figuras 1 y 2) obtenidas en ambas aplicaciones, se observa una más abombada en el caso alemán, en el que el índice de Gini era también superior, con valor de 0.44 frente a 0.16. Esto es debido a que dicho índice mide el área comprendida entre las funciones graficadas. El índice de Gini sirve para comparar la calidad global de ajuste de dos modelos calculados para los mismos datos. En este trabajo se propone como medida global pero no se llega a comparar con el obtenido mediante diferentes modelos.

En general, es apropiado y aconsejable realizar un estudio completo sobre cuál es el punto de corte "óptimo" si las poblaciones de la muestra original están desproporcionadas. De este modo se puede mejorar el ajuste del modelo logístico con punto de corte 0.5. Una ventaja del modelo de regresión logística BD es que por sí sólo permite flexibilidad en este sentido, pues la estimación resultante no es dicotómica, sino que produce probabilidades de insolvencia que están entre 0 y 1.

Se resalta que los cálculos se realizan con la función "dbglm", eligiendo distribución Binomial y link canónico *logit*, del paquete *dbstats* de R (Boj *et al*., 2012). Esto implica que el presente trabajo supone una herramienta de libre uso y de fácil utilización para el problema de *credit scoring* en el mercado asegurador.

BIBLIOGRAFÍA

Balzarotti, V y F. Castelpoggi (2009). Modelos de puntuación crediticia: la falta de información y el uso de datos de una central de riesgos. *Ensayos Económicos (Banco Central de la República Argentina)* 56, 95–156.

Boj, E., Claramunt, M. M. y J. Fortiana (2000). Una alternativa en la selección de los factores de riesgo a utilizar en el cálculo de primas. *Anales del Instituto de Actuarios Españoles*, Tercera Época 6, 11–35.

Boj, E., Claramunt, M. M. y J. Fortiana (2001). Herramientas estadísticas para el estudio de perfiles de riesgo. *Anales del Instituto de Actuarios Españoles*, Tercera Época 7, 59–89.

Boj, E., Claramunt, M. M. y J. Fortiana (2004). *Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación*. Cuadernos de la Fundación MAPFRE, 88. Fundación MAPFRE Estudios, Madrid.

Boj, E., Delicado, P. y J. Fortiana (2008). Logistic and local logistic distance-based regression. *Proceedings of the International Seminar on Nonparametric Inference ISNI 2008*, 66–70.

Boj, E., Claramunt, M. M., Esteve, A. y J. Fortiana (2009a). Credit Scoring basado en distancias: coeficientes de influencia de los predictores. En: Heras, A. y otros (2009). *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009*, pp. 15–22. Cuadernos de la Fundación MAPFRE, 136. Fundación MAPFRE Estudios, Madrid.

Boj, E., Claramunt, M. M., Esteve, A. y J. Fortiana (2009b). Criterios de selección de modelo en credit scoring, aplicación del análisis discriminante basado en distancias. *Anales del Instituto de Actuarios Españoles*, Tercera Época 15, 209–230.

Boj, E., Fortiana, J., Esteve, A., Claramunt, M.M. y T. Costa (2011). Aplicación de un modelo de regresión logística basado en distancias en el problema de credit scoring. En: Feria, J. M. y otros (2011). *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2011*, pp. 293–305. Cuadernos de la Fundación MAPFRE, 171. Fundación MAPFRE Estudios, Madrid.

Boj, E., Caballé, A., Delicado, P. y J. Fortiana (2012). *dbstats: Distance-based statistics (dbstats)*. R package version 1.0.2.

Gower J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–874.

Hosmer, D. W. y S. Lemeshow (2000). *Applied logistic regression, 2nd edition*. John Wiley & Sons, Inc. New York (USA).

Iñiguez, C. A.y M. G. Morales, M.G. (2009). *Selección de perfiles de clientes mediante regresión logística para muestras desproporcionadas, validación, monitoreo y aplicación en la proyección de provisiones*. Proyecto de Fin de Carrera, Escuela Politécnica Nacional, Ecuador.

Mures, M. J., García, A. y M. E. Vallejo (2005). Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. *Pecvnia* 1, 175–199.

Reyes, J., Escobar, C., Duarte, J. y P. Ramírez (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Estudios Pedagógicos* 23:2, 101–120.

Řezáč, M. y F. Řezáč (2011). How to Measure the Quality of Credit Scoring Models. *Journal of Economics and Finance* 61:5, .486–507

Siddiqi, N. (2006). Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring. John Wiley & Sons, Inc. New Jersey (USA).

West, D. (2000). Neural network credit scoring models. *Computer & Operations Research* 27, 1131–1152.