MODELOS Y METODOLOGÍAS DE CREDIT SCORE PARA PERSONAS NATURALES: UNA REVISIÓN LITERARIA

Models and methodologies for credit scoring in personal banking: A literature review

David Esteban Rodríguez-Guevara
Magíster en Administración Financiera. Instituto Tecnológico Metropolitano, Medellín- Colombia, davidrodriguez@itm.edu.co

Jairo Alfonso Becerra-Arévalo
Magíster en Administración Financiera. Instituto Tecnológico Metropolitano. Medellín-Colombia, jairobecerra@itm.edu.co

Daniel Cardona-Valencia
Especialista en Finanzas y Mercado de Capitales. Instituto Tecnológico Metropolitano. Medellín-Colombia, danielcardona@itm.edu.co

Cómo citar / How to cite
Rodríguez-Guevara, D.E., Becerra-Arévalo, J.A. y Cardona-Valencia, D. (2017). Modelos y metodologías de credit score para personas naturales: una revisión literaria. *Revista CEA*, *3*(5), 13-28.

Recibido: 14 de agosto de 2016
Aceptado: 29 de septiembre de 2016

Resumen

Este trabajo pretende aportar literariamente una revisión de los modelos para la calificación del riesgo crediticio (modelos de Credit Score) utilizados en el otorgamiento de crédito personal; teniendo en cuenta los métodos de Abdou & Pointon (2011); Glennon, Kiefer, Larson, & Choi (2008); Saavedra-García & Saavedra-García (2010), se pretende crear un esquema de orden para explicar los múltiples modelos matemáticos y econométricos utilizados en el credit score, con el fin de generar un listado actualizado que esté sustentado por académicos y expertos en el tema.

Palabras clave: análisis discriminante, credit score, modelos paramétricos, modelos no paramétricos, modelos semi-paramétricos.

Abstract

This paper provides a literature review on risk scoring models for credit granting in personal banking. The methods by Abdou & Pointon (2011), Glennon, Kiefer, Larson, & Choi (2008), and Saavedra-García (2010) are considered. The aim is to create a sorting scheme to explain the multiple mathematical and econometrical models used for credit scoring and to produce an up-to-date list supported by scholars and experts in the field.

Keywords: discriminant analysis, credit score, parametric models, non-parametric models, semiparametric models.

1. INTRODUCCIÓN

Los credit score son metodologías estadísticas o matemáticas que se especializan en el pronóstico e identificación de un cliente que pueda tener o no riesgo de impago de un crédito (Rodríguez & Trespalacios, 2015); son usados principalmente para la aprobación de los créditos, determinación de clasificación de créditos, asignación de precio de los créditos, generación de alertas tempranas y estrategias de cobranza (García Sánchez & Sánchez Barradas, 2005). De lo anterior, resulta oportuno indicar que las metodologías de credit score usadas en entidades bancarias o prestantes son directamente aplicadas en personas naturales pero muy poco en personas jurídicas.

La función primordial del credit score es identificar el riesgo de impago de un cliente, «discriminando» a los clientes nuevos con la historicidad de impago de clientes antiguos; para ello, es indispensable un análisis de sus variables personales; a estos procesos se conocen como análisis discriminante; los desarrollos de dichos análisis pueden ser variados dependiendo de las necesidades de las entidades que los soliciten, requiriendo el tipo, forma de los datos obtenidos, la veracidad y la eficiencia de los modelos. Estas metodologías nacieron con el análisis lineal discriminante (LDA) descritos por Fisher (1936) como una variante de los modelos ANOVA y que fueron la entrada a los modelos de identificación de quiebra, y a través de los años han tomado vertientes de uso con modelos lineales, no lineales, paramétricos, no paramétricos, estadísticos y econométricos que buscan encontrar el medidor perfecto.

Así mismo, a modo de compilación, han existido autores que se han encargado de abordar varias teorías y compilarlas de acuerdo con su funcionalidad ; sea el caso de Abdou & Pointon (2011); Anderson & Narasimhan (1979); García Sánchez & Sánchez Barradas (2005); Gartner & Schiltz (2005); Glennon, Kiefer, Larson, & Choi (2008); Gutiérrez Girault (2007) y Saavedra-García & Saavedra-García, 2010), se encargaron de realizar una identificación teórica a través del tiempo, actualizando cada vez la base científica que permitía tomar una decisión de uso del mejor modelo o método para hallar credit score; por lo tanto, la intención de este trabajo es realizar una revisión completa de los autores anteriores e identificar bajo un esquema único la constitución de las metodologías usadas para credit score para personas naturales hasta 2015.

2. MARCO DE REFERENCIA

Basilea II y la administración del riesgo

Basilea II recomienda a las entidades crediticias realizar análisis de riesgo de cartera basándose en indicadores financieros de los posibles candidatos involucrando variables como: liquidez, rendimientos, deuda, servicio a la deuda, activos e IRB (Internal Rating-Based approach); y si bien, estas variables descritas son las mínimas requeridas, involucrar una gran cantidad de estas que permitan describir a un conjunto de personas son bienvenidas siempre y cuando tengan un contexto frente a los créditos. Cabe agregar que, frente a la metodología, Basilea no establece directamente una metodología estándar, ya que se propone que los estudios y modelos para su análisis sean directamente responsabilidad de las entidades, siempre y cuando puedan ser fieles a la determinación del riesgo crediticio implícito de los clientes (Rodríguez & Trespalacios, 2015).

Así mismo, García Sánchez & Sánchez Barradas (2005) indican:

«La gestión de riesgos es fundamental para cualquier empresa cuya rentabilidad de negocio esté íntimamente ligada a los riesgos que asume. Cualquier entidad económica necesita identificar, valorar y cuantificar su exposición al riesgo, optimizando al mismo tiempo la rentabilidad, que se traslada directamente al cliente mediante unos precios más competitivos y la generación de mayores beneficios».

Credit Score y su uso para empresas

Según Thomas, Edelman & Crook (2002), las entidades financieras al momento de evaluar la capacidad crediticia de sus clientes siempre tienen en cuenta variables de tipo cualitativo y cuantitativo, pero, dependiendo de las necesidades y del uso que se deba dar dichas variables desentenderán del tipo de empresa a la que se le hará el estudio.

Los estudios de personas naturales toman entonces variables enfocadas a las características de la persona, la descripción del ser resultando en una infinidad de variables cualitativas (por ejemplo: genero, estrato, tipo de vivienda, municipalidad, raza), o cuantitativas (por ejemplo: ingresos, egresos, patrimonio, salario, número de personas a cargo, número de vehículos); mientras que un estudio de credit score para una persona jurídica, como lo muestra Gonçalves & Braga (2008) se usa variables de indicadores financieros: la razón corriente, la liquidez, activos, encaje bancario, cobertura en pasivos, provisión, rentabilidad de activos y patrimonio entre otros.

Anteriores revisiones literarias

Para identificar el proceso de referenciación de los modelos usados para el credit score, es importante tener en cuenta la explicación de algunos autores sobre los métodos usados discriminando los usos y los enfoques en que se desenvuelven.

García Sánchez & Sánchez Barradas (2005) explican que los modelos de credit score se pueden referenciar en: modelos tradicionales y en los modelos de enfoque moderno.

Los modelos tradicionales, son caracterizados por análisis de expertos que se enfocan en el «Carácter, Capital, Capacidad, Colateral y Ciclo», y si bien estos modelos son normalmente usados de manera intuitiva, el autor indica que pueden llegar a ser difícilmente aplicables entre los clientes, definiendo el hecho que las características por cada cliente hacen difícil su aplicación. Mientras, que los modelos de enfoque moderno, son de aplicación matemática, como los modelos Z-Score, modelos Z, modelos de respuesta binaria, Creditmetrics y VaR (Value at Risk).

Gutiérrez Girault (2007) realiza una relación totalmente directa para identificación de credit score mostrando que los modelos más usados para las entidades bancarias son: análisis discriminante, regresión lineal, regresión logística, modelos Probit, modelos Logit, métodos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, algoritmos de particioamiento recursivo (árboles de decisión), sistemas expertos, algoritmos genéticos, redes neuronales y, finalmente, el «juicio humano», dándole a su explicación un factor favorable a los modelos de respuesta bivariada (modelos Probit).

Glennon et al. (2008), en su estudio, muestra una explicación más característica desde el ámbito estadístico, explica en su documento el trato de los datos, teniendo especial cuidado en el tipo de información (datos transversales y de series de tiempo), el desarrollo del score-card para las observaciones, y los tipos de

modelos que se pueden utilizar para este fin; en este apartado se subdividen los modelos de forma paramétrica, donde solamente se muestra al modelo Logit como el único modelo paramétrico, semi-paramétrica, mostrando una función combinada de un modelo Logit y una red neuronal, y no paramétrica como el modelo CHAID (Chi-squared Automatic Interaction Detector).

Saavedra-García & Saavedra-García (2010), muestran bajo un sistema similar al de García Sánchez & Sánchez Barradas (2005), donde se parten los modelos de análisis de crédito en modelos tradicionales y modelos modernos, los primeros identifican modelos matemáticos paramétricos que analizan las variables con modelos Logit, Lineales y hasta sistemas expertos; los segundos utilizan modelos empresariales especiales: modelo MKV (Kealhofer, McQuown and Vasicek) y el modelo CyRCE (Capital y Riesgo de Crédito en Países Emergentes).

Abdou & Pointon (2011) separan su resultado en métodos avanzados estadísticos y métodos tradicionales estadísticos. Los primeros son modelos que usan programación genética y redes neuronales para determinar los resultados de una discriminación bivariada o multivariada, y los segundos, los que usan métodos clásicos de análisis paramétrico. Igualmente, los modelos descritos por Abdou & Pointon (2011) tienen en cuenta modelos lineales, análisis discriminante, análisis probabilístico (Probit) y logístico (Logit), árboles de decisión, sistemas expertos (Modelos Bayesianos), redes neuronales, y programación genética; a esta investigación el autor expone también los métodos usuales de verificación de resultados y bondad de ajuste (curva ROC, el costo de error en la clasificación, y la matriz de confusión).

3. RESULTADOS

Teniendo en cuenta el método de descripción de los modelos propuesto por Glennon et al., (2008) y Saavedra-García & Saavedra-García (2010), se establece que los modelos más representativos de credit score hasta el año 2015 se parten en tres tipos de modelos generales: 1) Los modelos paramétricos, 2) modelos no paramétricos, 3) modelos semi-parámetricos. En general responden a la relación matematica f(Y,X), el cual busca identificar el resultado de riesgo o no riesgo (variable dependiente) frente a una serie indeterminada de variables socioeconómicas (variables independientes); dando una descripción de su uso y la formulación matemática que desarrolla el método:

Modelos paramétricos

Estos modelos tienen la condición matemática de describir la sensibilidad de discriminación por medio de parámetros o estimaciones obtenidas bajo el proceso MCO (Mínimos Cuadrados Ordinarios); permiten identificar la sensibilidad o cambio presentado por las variables usadas según el tipo de modelo; operando de dos formas: Univariantes y Multivariantes.

- *Modelos paramétricos Univariantes – Modelos ANOVA*

Palacio, Lochmúller, Murillo, Pérez & Vélez (2011), explican que usando datos de encuestas realizadas a los clientes, y asumiendo un proceso similar al bayesiano que usa promedios de expertos, se revisan las variables más representativas que pueden representar el riesgo de impago de crédito con un proceso de un solo factor o ANOVA; si bien este estudio toma varias variables en el estudio, el proceso estadístico utilizado revisa una a una de las variables para identificar

cuales presentan efecto en la variable dependiente.

➢ *Modelos paramétricos multivariantes*

- *Análisis discriminante*

Fisher (1936), Puertas & Marti (2012) y Mures, García, & Vallejo (2011), explican que el análisis discriminante es una relación de comportamiento de múltiples variables identificadas frente a una variable dependiente, la cual estructura varios posibles resultados o grupos de ecuaciones lineales que identifican la combinación eficiente para separar o discriminar a los buenos o malos pagadores.

El proceso distingue una serie de modelos lineales tal que así (1):

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_i x_i \qquad (1)$$

Donde:

λ_i= Representa los parámetros de la relación de combinaciones de variables.

X, representa una cantidad n de modelos posibles que nacen de las combinaciones entre variables. Cuando se tuviesen los parámetros hallados por cada grupo de datos, se toma la serie de datos en diferencia entre las medias (desviaciones estándar) quedando (2):

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \cdots + \lambda_i d_i \qquad (2)$$

De esto, se busca encontrar una función de varianzas y covarianzas que determine el modelo con menos valores residuales entre ellos.

$$S = \sum_{p=1}^{n} \sum_{q=1}^{n} \lambda_p \lambda_q S_{pq}$$

Cuando se obtengan las estimaciones, la predicción depende de la puntuación de corte optimo que discrimina los valores 0 y 1 de forma perfecta.

- *Modelos LDA (Linear Discriminant Analysis) - Modelo Z – Score*

De la misma forma que Fisher (1936) obtuvo el proceso discriminante, estos modelos fueron usados por Altman (1968), Altman (1980) y Elliott, Siu & Fung (2014) en sus estudios para identificar en el primer caso la probabilidad de quiebra de una entidad, que después utilizo para estudiar la probabilidad de impagos en una entidad bancaria, a dichos trabajos se les conoció como modelos LDA, Modelos Z y modelos Z – score; la ecuación del modelo Z es (3):

$$Z = \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \qquad (3)$$

Donde:

$Z =$ índice general
$\beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago
$X_{ik} =$ variables cualitativas o cuantitativas.

- *Modelos Least – Absolute – Value (LAV)*

Glorfeld (1990) demuestra que el modelo lineal discriminante tiene amplios fallos, debido a la violación de los supuestos lineales; este autor propone hacer un análisis de credit score tomando un modelo OLS (Ordinary Least Squares), condicionando el vector de residuales contrastados contra el valor absoluto de la diferencia entre los datos reales y los pronosticados, siendo este el determinante del mínimo error expresado.

Estos modelos garantizan que los valores de los parámetros tengan distribuciones más ajustadas, haciendo que la media de los errores sea asintótica mostrando entonces que la

afectación de los datos por outliers sea menor en LAV que en modelos LDA o MDA.
Para ello, es requerido minimizar la sumatoria de los errores tal que (4):

$$\min z = \sum_{i}^{n} (\varepsilon_i^+ - \varepsilon_i^-), i = n \quad (4)$$

Y, dado que el modelo lineal requiere un valor cercano a cero, es imprescindible que dichos errores sean totalmente positivos (5).

$$y_i - \left(\alpha + \sum_{j}^{n} \beta_j x_{ij} \right) + \varepsilon_i^+ - \varepsilon_i^- = 0 \quad (5)$$

Donde:

$y_i =$ es la respuesta bivariada de riesgo de impago
$\alpha, \beta_j =$ son los parámetros de cambio que afectan la probabilidad de impago
$x_{ik} =$ variables cualitativas o cuantitativas
ε_i = residuales de la función

- *Modelos Lineales Probabilísticos (MLP)*

Como los describen Hardy & John (1985), Santos & Famá (2007), Bumacov, Ashta & Singh (2014), Puertas & Marti (2012), los modelos lineales probabilísticos son modelos en donde la mejor combinación lineal de las variables en estudio pueden arrojar una respuesta bivariada eficiente, pero son sensibles a presentar problemas de especificación si violasen los supuestos de linealidad (Gujarati, 2004) (6).

$$Y_i = \alpha + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad (6)$$

Donde:
$Y_i =$ es la respuesta bivariada de riesgo de impago
$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago
$X_{ik} =$ Variables cualitativas o cuantitativas

La diferencia entre los modelos MLP y los modelos discriminantes, radica en la construcción de varias combinaciones lineales, en cambio toma los preceptos del modelo lineal usando la bondad de ajuste para definir el modelo optimo, y así obtener un valor porcentual que defina el valor de impago.

- *Modelo Logístico (Logit)*

Según, Constangioara (2011); Gonçalves & Braga (2008); Lipovetsky & Conklin (2004); Majer (2006); Rayo, Lara, & Camino (2010); Rodríguez & Trespalacios (2015); Támara, Aristizábal, & Velásquez (2010), el modelo logístico es una respuesta para el incorrecto funcionamiento de los modelos MLP, este usa la función logística matemática para determinar un crecimiento condicionado en donde los valores más cercanos a cero (incluyendo $-\infty$) sean iguales a cero, y a medida de crecen al $+\infty$, tomarán la forma del valor 1, por lo tanto son modelos con parámetros no lineales que nacen de la máxima verosimilitud. Su funcionamiento como un modelo econométrico dependerá del uso de la bondad de ajuste, tablas de confusión y curvas ROC para determinar el nivel de discriminación propuesto (7).

$$z_t = \alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \quad (7)$$

Siendo el modelo en la función logística (8):

$$P_i = E(y = 1|X) = \frac{1}{1 + e^{-(\alpha+\beta_2 X_{t2}+\cdots+\beta_k X_{tk})}} + \varepsilon_i = \frac{1}{1 + e^{-z_t}} + \varepsilon_i \quad (8)$$

Donde:

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ Variables cualitativas o cuantitativas.

- *Modelos Probabilísticos (Probit)*

Así mismo, Melo & Granados (2011); S. Moreno (2013); Rayo et al. (2010); Roszbach (2004); Támara et al. (2010) y Webster (2011) muestran que los modelos Probit también han sido ampliamente usados por su condición de establecer normalidad al proceso de predicción, dándoles a las variables mayor estabilidad en el proceso de ajuste.

Su uso dependerá entonces de la función normal (9):

$P_i = P(Y = 1|X) = P(I_t^* \leq I_t)$, que es igual a

$P_i = P(Z_t \leq \alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$, igual a

$$P_i = F(\alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \quad (9)$$

Se dará entonces la CDF de la distribución normal para el modelo lineal, quedando (10):

$$F(I) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta_2 X_{i2}+\cdots+\beta_k X_{ik}} e^{-\frac{z^2}{2}} dz + \varepsilon_i \quad (10)$$

Donde:

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ variables cualitativas o cuantitativas

$z =$ es la distribución normal estandarizada

- *Modelos Tobit*

Autores como, Fernández & Pérez (2005); Olagunji & Ajiboye (2010), Tan, Yen & Loke (2011) y Roszbach (2004) indican que dichos modelos proporcionan un mejor ajuste de la realidad de los datos al tener comprometida la información cuando existen con problemas de especificación de normalidad, condicionando la variabilidad de los errores. Si bien, estos modelos son similares al modelo Probit, su uso es menos comercial por su complejidad en el uso y que no siempre son tan generales para la construcción de un modelo de predicción de riesgo de crédito.

$$y_i^* = \alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \quad (11)$$

Siendo y_i^* una variable inobservable, la cual puede tener dos resultados

$y_i = y_i^*, si\ y_i > 0$

$y_i = 0, si\ y_i \leq 0$

Para todos los casos la función para obtener el resultado óptimo sería (12):

$$lnL = \sum_{y_i>0} [-ln\sigma + ln\phi(y_i - \alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})] + \sum_{y_i=0} \ln\left[1 - \Phi\left(\frac{\alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}}{\sigma}\right)\right] \quad (12)$$

Donde:

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ variables cualitativas o cuantitativas

$\phi\ y\ \Phi$= son funciones de densidad acumulada para una distribución normal estándar

- *Modelos Logit Multinomiales*

Una propuesta alternativa al modelo Logit y Probit, Chaudhuri & Cherical (2012) y Gonçalves & Braga (2008) tienen en cuenta el uso que propone Basilea II en la forma que se debe analizar los créditos en rangos de maduración (A,B,C,D,E, siendo A, un buen pagador, y B,C,D,E pagadores morosos), por lo tanto, el cambio dirigido para los autores radica en que la variable Y, puede tener 5 respuestas posibles.

$P_i = Prob(Y_i = j)$, que es igual a (13):

$$P_i = \frac{e^{(\alpha+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}{\sum_{k=0}^{J} e^{(\alpha+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}, j = 0,1,2,3,4 \quad (13)$$

Donde:

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ variables cualitativas o cuantitativas

- *Modelos Logit Mixtos*

Kukuk & Rönnberg (2013) y Moreno (2013), propone una alternativa para los modelos Logit y Probit convencionales, esta autora recomienda su uso por su flexibilidad a la hora de predecir el nivel de crédito sin caer en los errores convencionales de los modelos Logit y Probit, ella muestra que usando un condicional binomial se obtendrá la siguiente ecuación (14), donde:

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ variables cualitativas o cuantitativas

$\tilde{\beta}, \tilde{\theta}$ = parámetros hallados por máxima verosimilitud

$P(y_i = 1|x_i, \tilde{\beta}, \tilde{\theta}) = \int_{\beta_i} \Lambda(\alpha + \beta_2 X_{t2} + \cdots + \beta_k X_{tk}) f(\beta_i|\tilde{\beta}, \tilde{\theta}) d\beta_i$, que es igual a:

$$P(y_i = 1|x_i, \beta_k) = \Lambda(\alpha + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) = \frac{e^{(\alpha+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}}{\sum e^{(\alpha+\beta_2 X_{i2}+\cdots+\beta_k X_{ik})}} \quad (14)$$

La diferencia encontrada frente al Logit convencional, es que los modelos permiten utilizar parámetros simulados, cuando se establece esta simulación, se pone también a favor el hecho que los residuos son simulados, estandarizándolos y creando una mejor predicción.

Modelos No Paramétricos

A diferencia de los modelos paramétricos, los modelos no paramétricos se estructuran en procesos matemáticos que ocultan el proceso interno y se especifican en las variables de entrada y la de salida; son normalmente usados con procesos de nodos o redes que asemejan al cerebro, encontrándose:

- *Redes Neuronales (Neuronal networks)*

Desai, Crook & Overstreet (1996);West (2000);Esteve (2007); Pérez & Fernández (2007) y Soydaner & Kocadağlı (2015), explican que el uso de las redes neuronales utilizan un sistema artificial que se asemeja al cerebro humano, y que es capaz de identificar comportamientos variables de varios individuos a la vez dando un resultado de predicción eficiente, su estructura se basa en neuronas de información que se interconectan y determinan bajo entrenamiento de información, el resultado más similar a la realidad. Para su desarrollo es requerido el uso de un perceptron multicapa, que se estimula con ecuaciones que definen una entrada y una salida y comportamiento de los datos a modo de entrenamiento; cuando esté totalmente entrenado es capaz de re direccionar nuevos datos a resultados predecibles.

- *SVM (Support – Vector – Machine)*

Martens et al. (2010); J. F. Moreno & Melo (2011) y Zhou, Lai & Yen (2009) identifican un proceso nacido de las redes neuronales que usa geometría euclidiana para discriminar correctamente los datos; la operación matemática se asemeja a una operación lineal que identifica los espacios que existen entre los datos, son llamados «hiperplanos», que muestran la distancia discriminante que hay en respuestas de tipo binomial [0,1]. El factor de entrenamiento de los datos sería (15):

$$y(x) = sign(w^T x + b) \quad (15)$$

Que identifican dos posibles relaciones de discriminación (16):

$$\begin{cases} w^T x + b \geq +1, si\ y_k = +1 \\ w^T x + b \leq -1, si\ y_k = -1 \end{cases} \quad (16)$$

Donde:

y_k = es un valor -1 o 1 que determina la clase a la que pertenece x, siendo este un vector real de carácter p dimensional.
w = es un vector normal en el hiperplano

Siendo esta una función lineal clásica que tomará el nombre de hiperplano cuya función es analizar las distancias de cada punto y se bifurca en posiciones de [-1, +1] mostrando la ecuación óptima que sirve para entrenar y discriminar la información, proporcionando una predicción óptima.

- *Modelos Bayesianos (Naive Models)*

Chang, Fung, Lucas, Oliver & Shikaloff (2000), Baesens, Castelo & Vanthienen (2002), Mileris (2010) y Webster (2011) usan modelos de predicción netamente probabilísticos que usan una función tal que (17):

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (17)$$

Donde $P(C_i|X)$ es la probablidad posterior condicionada de C_i en X; $P(Ci)$ es la probabilidad de C_i; $P(X|C_i)$ es la probabilidad de X condicionada en C_i; $P(X)$ es la

probabilidad de X; siendo esta una función maximizada para todos los casos de los clientes, sobre esta función es requerido hacer un entrenamiento de los datos quedando que (18):

$$P(X|C_i) = \prod_{k=1}^{n} P(x_1|C_i) * P(x_2|C_i) * \ldots * P(x_n|C_i) \quad (18)$$

Esta probabilidad será entonces una función normalizada para los errores que puedan proporcionar. Esta metodología trabaja esencialmente como un árbol de decisión, y puede ser amplificada si se tiene una variable de respuesta (A, B, C, D, E) mostrando una flexibilidad mayor en su uso.

Modelos Semi-paramétricos (Hybrid Models)

Abdou & Pointon (2011); Akkoç (2012); Mileris (2010) y Moreno & Melo (2011) identifican que los modelos paramétricos pueden tener problemas de especificación a la hora de verificar la eficiencia de los errores de los modelos; si bien pueden tener errores normales, pueden no ser del todo eficientes e insesgados, lo que produce problemas de especificidad.

Teniendo en cuenta lo anterior, el uso de modelos que involucren procesos de redes neuronales o de análisis genético o discriminador por medio de máquinas vectoriales potencializan los procesos de identificación de las variables eliminando el proceso de sesgamiento de los datos mostrando una serie de datos normalizada y eficiente, prometiendo que un uso de cualquier modelo paramétrico ofrezca la identificación de un modelo no solo insesgado, también altamente eficiente. Algunos ejemplos son:

- *Maximum score*

Bult (1993) hace un comparativo de los modelos semi-paramétricos contrastados con un modelo paramétrico estándar como lo es el modelo Logit; según el autor, la desventaja del uso de los modelos paramétricos se enfoca en que su estimación no proporciona un dato robusto por su amplio nivel de supuestos implícitos. Los modelos semi-paramétricos en cambio combinan el proceso de los estimadores con funciones no paramétricas que identifican una función de probabilidad de densidad normal para los errores. Para ello, la condición ecuacional será (19):

$$y_i^* = \begin{cases} 1, si\ y_i = \beta' x_i + u_i > 0 \\ 0, en\ otro\ caso \end{cases} \quad (19)$$

Donde y_i es una variable respuesta inobservable y y_i^* es el indicador de respuesta; cuando se obtiene dicho valor inobservable, se someterá a un proceso probabilístico tal que (20):

$$Prob(y_i^* = 1|x_i) = Prob(y_i > 0|x_i) = Prob(\beta' x_i + u_i > 0|x_i) = Prob(u_i > -\beta' x_i|x_i) = 1 - F(-\beta' x_i)$$

$$S(\beta) = \sum_{i=1}^{N} S_i(\beta) = \sum_{i=1}^{N} y_i^* sign(\beta' x_i) \quad (20)$$

$Y_i =$ es la respuesta bivariada de riesgo de impago

$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago

$X_{ik} =$ variables cualitativas o cuantitativas

- *Técnica CHAID (CHi-squared Automatic Interaction Detection)*

Antipov & Pokryshevskaya (2010), Espin-García & Rodríguez-Caballero (2013) explican en su tesis que los árboles de decisión son usados para optimizar el proceso que una regresión logística o paramétrica no puede contrastar fácilmente, y es que los errores presentados en las regresiones paramétricas presentan fallas demasiado amplias cuando la base de datos no es eficiente. Los autores manifiestan que el uso de árboles de decisión antes de operar los modelos paramétricos ofrece un uso prudente de las variables antes de operarlas, evidenciando cuáles serán los parámetros más eficientes e insesgados antes de cualquier operación econométrica.

- *LS-SVM (Least Squares- Support Vector Machines)*

Un uso diferenciado mostrado por Tsai (2008), Zhou et al. (2009) del uso de un modelo mezclado entre las máquinas de soporte vectorial y los modelos lineales, estos procesos primero pueden discriminar la información de los datos por medio de un «kernel», que es un centro neurálgico de entrenamiento de los datos que permite identificar de forma no lineal el mejor proceso de discriminación, y sobre este determinar un modelo lineal que permite la relación matemática para un pronóstico eficiente de los datos. Si bien estos modelos no son del todo paramétricos y difíciles en algunos casos de crear, son muy potentes y tienen una aceptabilidad grande al momento de dar un resultado matemático.

- *Neuronal Network + Logit Models*

Sustersic, Mramor & Zupan (2007) presentan un uso sofisticado para los modelos logísticos muy similar al usado en la técnica CHAID, dando uso de redes neuronales para identificar las variables optimas usando un sistema de PCA (Análisis de Componentes Principales), acto seguido utiliza un proceso de modelos logísticos para identificar el modelo econométrico que permite estimar el discriminador de Credit Score más apropiado.

- *Decision trees – CART (Clasification and Regression Trees) models*

Zhang, Zhou, Leung & Zheng (2010),Baklouti (2014); Díaz Sepulveda (2012); Kočenda & Vojtek (2009) explican que los árboles de decisión basados en funciones bayesianas pueden ser usados como modelos de discriminación para el credit score, los árboles de decisión se comportan de alguna manera similar una red neuronal dirigida, es necesaria una técnica más sofisticada para no dejar en paralelo el hecho que los árboles de decisión solamente por medio de los nodos o variables no son capaces de dar un valor único de salida, por lo que el uso de regresión lineal es necesaria en dichos procedimientos (21).

$$i(\tau) = \phi(\{Y = 1|\tau\}) \tag{21}$$

Donde τ es el nodo y ϕ define la probabilidad de éxito del nodo. Para la relación lineal de esta expresión se encuentra entonces (22):

$$i(\tau) = \sum_{sujeto\ i \in \tau} (Y_i - \bar{Y}(\tau))^2 \tag{22}$$

Donde Y_i será (23):

$$Y = F(x_1, x_2, x_3, \dots, x_p) = \alpha + \sum_{j=1}^{p} \beta_k x_j = \alpha + \sum_{j=1}^{p} \beta_k g_j(x) = f(x) \tag{23}$$

$Y_i =$ es la respuesta bivariada de riesgo de impago
$\alpha, \beta_k =$ son los parámetros de cambio que afectan la probabilidad de impago
$X_{ik} =$ variables cualitativas o cuantitativas

4. CONCLUSIONES

Para los análisis de credit score en personas naturales se identificaron tres tipos de metodologías claves, modelos paramétricos, no paramétricos y semi-paramétricos, de los cuales se destacan en uso y mención los modelos Logit, los modelos LDA, los modelos de regresión censurada, los modelos LS-SVM, SVM, las redes neuronales, los árboles de decisión CART y los modelos bayesianos.

Los modelos pueden ser variados y totalmente abierto a las necesidades de análisis de las entidades financieras que los requieran; es de acotar que la amplia variedad de metodologías responde a una pregunta obvia de todo investigador, ¿cuál modelo es más eficiente?, y la respuesta es evidente, no existe un modelo o un método perfecto de identificación de credit score por varias situaciones, la primera de ellas es la calidad de los datos y las variables presentadas; la dependencia o el uso de un modelo dependerá de la calidad de la información y sobre que variables se deberán trabajar; pero tomando un punto estricto en la revisión bibliográfica, el modelo más utilizado por los autores es el modelo Logit o Probit, por su facilidad de interpretación, su facilidad de manejo en el proceso matemático, y si bien, no son los más óptimos si son un referente base para la discriminación de clientes.

Otra es la real necesidad de la medición y está sometida al criterio del investigador, muchos son los consultores que tienen preferencia por la metodología no paramétrica para identificar un modelo óptimo, pero también es de investigadores ortodoxos de la econometría el uso de modelos paramétricos o semi-paramétricos aduciendo su amplia eficiencia.

Para todos los modelos identificados es una premisa el uso de una regresión lineal o función lineal siendo esta la base de sus análisis, esto se debe que el aspecto de discriminación de la información en la variable dicotómica Y_i; lo cierto es que los modelos se refinan en su uso mejorando la característica lineal a no lineal asumiendo que la información y los datos no son perfectamente lineales ni tampoco son totalmente normales para poder teóricamente definir un elemento de simplicidad como tal.

REFERENCIAS

Abdou, H. A. & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. Intelligent Systems in Accounting, Finance and Management, 18(2–3), 59–88. http://doi.org/10.1002/isaf.325

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 222(1), 168–178. http://doi.org/10.1016/j.ejor.2012.04.009

Altman, E. I. (1968). the Prediction of Corporate Bankruptcy. The Journal of Finance, XXIII(September), 589–609.

Altman, E. I. (1980). Commercial Bank Lending: Process, Credit Scoring, And Costs Of Errors In Lending. Journal Of Financial And Quantitave Analysis, XV(4), 813–832.

Anderson, J. & Narasimhan, R. (1979). Assessing Project Implementation Risk: A

Methodological Approach. Management Science, 25(6), 512–521. http://doi.org/10.1287/mnsc.25.6.512

Antipov, E. & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement. Journal of Targeting, Measurement and Analysis for Marketing, 18(2), 109–117. http://doi.org/10.1057/jt.2010.3

Baesens, B.; Castelo, R. & Vanthienen, J. (2002). Learning Bayesian network classifiers fot credit scoring using Markov Chain Monte Carlo search. Ieee, (2), 2–5.

Baklouti, I. (2014). A Credit Scoring Model for Microfinance Bank Based on Fuzzy Classifier Optimized by a Differential Evolution Algorithm. IUP Journal Of Financial Risk Management, 11(2), 7–24.

Bult, J. R. (1993). Semiparametric versus Parametric Classification Models: An Application to Direct Marketing. Journal Of Marketing Research, 30(3), 380–390. http://doi.org/10.1007/978-3-642-21551-3

Bumacov, V.; Ashta, A. & Singh, P. (2014). The Use of Credit Scoring in Microfi nance Institutions and Their Outreach. Strategic Change, 23(1), 401–413. http://doi.org/10.1002/jsc

Chang, K. C.; Fung, R.; Lucas, A.; Oliver, R. & Shikaloff, N. (2000). Bayesian networks applied to credit scoring. IMA Journal of Management Mathematics, 11(1), 1–18. http://doi.org/10.1093/imaman/11.1.1

Chaudhuri, K. & Cherical, M. M. (2012). Credit rationing in rural credit markets of India. Applied Economics, 44(7), 803–812. http://doi.org/10.1080/00036846.2010.524627

Constangioara, A. (2011). Consumer Credit Scoring. Romanian Journal Of Economic Forecasting, 3, 162–178.

Desai, V. S.; Crook, J. N. & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, 95(1), 24–37. http://doi.org/10.1016/0377-2217(95)00246-4

Díaz, J. F. (2012). Comparación entré Árboles de Regresión CART y Regresión Lineal. Universidad Nacional de Colombia.

Elliott, R. J.; Siu, T. K. & Fung, E. S. (2014). A Double HMM approach to Altman Z-scores and credit ratings. Expert Systems with Applications, 41(4 PART 2), 1553–1560. http://doi.org/10.1016/j.eswa.2013.08.052

Espin-García, O. & Rodríguez-Caballero, C. V. (2013). Metodología para un scoring de clientes sin referencias crediticias. Cuadernos de Economía, 32(59), 139–164.

Esteve, E. M. (2007). Un modelo de credit scoring basado en el conocimiento de la aplicación de Basilea II y su papel innovador en el sector bancario. Asociación Española de Dirección y Economía de la Empresa.

Fernandez, H. & Pérez, F. O. (2005). El modelo logístico : una herramienta estadística para evaluar el riesgo de crédito. Revista Ingenierias Universidad de Medellín, 4, 55–75.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188. http://doi.org/10.1111/j.1469-1809.1936.tb02137.x

García Sánchez, M. & Sánchez, C. (2005). Antecedentes: modelos para estimar el riesgo de crédito. Riesgo de crédito en México: aplicación del modelo CreditMetrics. Universidad de las Américas Puebla.

Gartner, K. & Schiltz, E. (2005). What's Your Score? Educating College Students About Credit Card Debt. Legal Studies Research Paper Series WHAT'S, 24(1), 401–432.

Glennon, D.; Kiefer, N. M.; Larson, C. E. & Choi, H. (2008). Development and Validation of Credit Scoring Models. Journal of Credit Risk, Forthcoming, 1(1), 1–70. Retrieved from http://papers.ssrn.com/abstract=1180302

Glorfeld, L. W. (1990). A Robust Methodology for Discriminant Analysis Based on Least-absolute-value Estimation. Managerial and Desicion Economics, 11(1), 267–277.

Gonçalves, R. M. L. & Braga, M. J. (2008). Determinantes de risco de liquidez em cooperativas de crédito: uma abordagem a partir do modelo logit multinomial. Revista de Administração Contemporânea, 12(4), 1019–1041. http://doi.org/10.1590/S1415-65552008000400007

Gujarati, D. (2004). Econometría (4ta Edició). McGraw-Hill Interamericana. Retrieved from https://books.google.es/books?id=8RttQgAACAAJ

Gutiérrez, M. A. (2007). Modelos de Credit Scoring - Qué, Cómo, Cuándo y Para Qué.

Hardy, W. E. & John, I. (1985). A Linear Programming Alternative to Discriminant Analysis in Credit Scoring. Agribussiness, I(4), 285–292.

Kočenda, E. & Vojtek, M. (2009). Default Predictors and Credit Scoring Models for Retail Banking. CESIFO WORKING PAPER NO. 2862C. Retrieved from www.CESifo-group.org/wp

Kukuk, M. & Rönnberg, M. (2013). Corporate credit default models: A mixed logit approach. Review of Quantitative Finance and Accounting, 40(3), 467–483. http://doi.org/10.1007/s11156-012-0281-4

Lipovetsky, S. & Conklin, M. (2004). Decision Making By Variable Contribution in Discriminant, Logit, and Regression Analyses. International Journal of Information Technology & Decision Making, 3(2), 265–279. http://doi.org/10.1142/S0219622004001033

Majer, I. (2006). Application scoring: logit model approach and the divergence method compared (06 No. 10). Warsaw.

Martens, D.; Van Gestel, T.; De Backer, M.; Haesen, R.; Vanthienen, J. & Baesens, B. (2010). Credit rating prediction using Ant Colony Optimization. Journal of the Operational Research Society, 61(4), 561–573. http://doi.org/10.1057/jors.2008.164

Melo, L. F. & Granados, J. C. (2011). Regulación y valor en riesgo. Ensayos Sobre Política Economica, 29(64), 110–177.

Mileris, R. (2010). Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier. Economics and Management, 15(1), 1078–1084. Retrieved from http://www.ktu.lt/lt/mokslas/zurnalai/ekovad/15/1822-6515-2010-1078.pdf

Moreno, J. F. & Melo, L. F. (2011). Pronóstico de incumplimientos de pago mediante

máquinas de vectores de soporte: una aproximación inicial a la gestión del riesgo de crédito. Boletín de Prensa DANE, 677, 1–33.

Moreno, S. (2013). El Modelo Logit Mixto para la construcción de un Scoring de Crédito. Universidad Nacional de Colombia.

Mures, J.; García, A. & Vallejo, E. (2011). Aplicación del análisis discriminate y regresión logística en el estudio de la morosidad en las entidades financieras comparación de resultados. Revista de La Facultad de Ciencias Económicas Y Empresariales, 1, 175–199. Retrieved from http://search.proquest.com/docview/818448211?accountid=10344

Olagunji, F. & Ajiboye, A. (2010). Agricultural lending decision: a tobit regression analysis. African Journal of Food Agriculture, Nutrition and Development, 10(5), 1–27. http://doi.org/10.4314/ajfand.v10i5.57897

Palacio, A. P.; Lochmúller, C.; Murillo, J. G.; Pérez, M. A. & Vélez, C. A. (2011). Modelo cualitativo para la asignación de créditos de consumo y ordinario - el caso de una cooperativa de crédito. Revista Ingenierias Universidad de Medellín, 10(19), 89–100.

Pérez, F. O. & Fernández, H. (2007). Las redes neuronales y la evaluación del riesgo de crédito. Revista Ingenierías, 6(10), 77–91.

Puertas, R. & Marti, M. L. (2012). Análisis Del Credit Scoring. Revista Administración de Empresas, 53(3), 303–315.

Rayo, S.; Lara, J. & Camino, D. (2010). Un modelo de credit scoring para instituciones de microfinanzas en el marco de Basile II. Journal of Economics, Finance & Administrative Science., 15(28), 89–124. Retrieved from http://eds.b.ebscohost.com/eds/detail?vid=2&sid=a4134bcc-0c4d-42d1-b74e-04f1be482687@sessionmgr112&hid=115&bdata=Jmxhbmc9ZXMmc2l0ZT1lZHMtbGl2ZQ==#db=bth&AN=51381543

Rodriguez, D. & Trespalacios, A. (2015). Medición de valor en riesgo en cartera de clientes a través de modelos logísticos y simulación de Montecarlo.

Roszbach, K. (2004). Bank Lending Policy, Credit Scoring, and the Survival of Loans. Review of Economics and Statistics, 86(4), 946–958. http://doi.org/10.1162/0034653043125248

Saavedra-García, M. L. & Saavedra-García, M. J. (2010). Modelos para medir el riesgo de crédito de la banca *. Cuadernos de Administración, 23(40), 295–319. Retrieved from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-35922010000100013&lang=pt

Santos, J. O. & Famá, R. (2007). Avaliação da aplicabilidade de um modelo de credit scoring com varíaveis sistêmicas e não-sistêmicas em carteiras de crédito bancário rotativo de pessoas físicas. Revista Contabilidade & Finanças, 18, 105–117. http://doi.org/10.1590/S1519707720070002000009

Soydaner, D. & Kocadağlı, O. (2015). Artificial Neural Networks with Gradient Learning Algorithm for Credit Scoring. Journal of the School of Bussiness Administration, 44(2), 3–12.

Sustersic, M.; Mramor, D. & Zupan, J. (2007). Consumer credit scoring models with limited data. Ljubljana Meetings Paper, 1(1),1–21.

http://doi.org/10.1016/j.eswa.2008.06.016

Támara, A. L.; Aristizábal, R. E. & Velásquez, H. (2010). estimación de las provisiones esperadas en una institución financiera utilizando modelos Logit. Revista Ciencias Estratégicas, 18(24), 259–270.

Tan, A. K. G.; Yen, S. T. & Loke, Y. J. (2011). Credit card holders, convenience users and revolvers: A tobit model with binary selection and ordinal treatment. Journal of Applied Economics, 14(2), 225–255. http://doi.org/10.1016/S1514-0326(11)60013-5

Thomas, L.; Edelman, D. & Crook, J. (2002). Credit scoring and its applications.

Tsai, C. F. (2008). Financial decision support using neural networks and support vector machines. Expert Systems, 25(4), 380–393. http://doi.org/10.1111/j.1468-0394.2008.00449.x

Webster, G. (2011). Bayesian Logistic Regression Models for Credit Scoring. Rhodes University.

West, D. (2000). Neural network credit scoring models. Computers and Operations Research, 27(11–12), 1131–1152. http://doi.org/10.1016/S0305-0548(99)00149-5

Zhang, D.; Zhou, X.; Leung, S. C. H. & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. Expert Systems with Applications, 37(12), 7838–7843. http://doi.org/10.1016/j.eswa.2010.04.054

Zhou, L.; Lai, K. K. & Yen, J. (2009). Credit Scoring Models With Auc Maximization Based on Weighted Svm. International Journal of Information Technology & Decision Making, 8(4), 677–696. http://doi.org/10.1142/S0219622009003582