Rodríguez Guevara, David Esteban; Rendón Garcia, Juan Fernando; Trespalacios Carrasquilla, Alfredo; Jiménez Echeverri, Edwin Andrés

Article

Modelación de riesgo de crédito de personas naturales. Un caso aplicado a una caja de compensación familiar colombiana

Revista de Métodos Cuantitativos para la Economía y la Empresa

Provided in Cooperation with:
Universidad Pablo de Olavide, Sevilla

Suggested Citation: Rodríguez Guevara, David Esteban; Rendón Garcia, Juan Fernando; Trespalacios Carrasquilla, Alfredo; Jiménez Echeverri, Edwin Andrés (2022) : Modelación de riesgo de crédito de personas naturales. Un caso aplicado a una caja de compensación familiar colombiana, Revista de Métodos Cuantitativos para la Economía y la Empresa, ISSN 1886-516X, Universidad Pablo de Olavide, Sevilla, Vol. 33, pp. 29-48, https://doi.org/10.46661/revmetodoscuanteconempresa.5146

This Version is available at:
https://hdl.handle.net/10419/286260

REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (33). Páginas 29-48.
Junio de 2022. ISSN: 1886-516X. D.L: SE-2927-06.
www.upo.es/revistas/index.php/RevMetCuant/article/view/5146

Modelación de riesgo de crédito de personas naturales. Un caso aplicado a una caja de compensación familiar colombiana

RODRÍGUEZ GUEVARA, DAVID ESTEBAN
Instituto Tecnológico Metropolitano de Medellín (Colombia)
Correo electrónico: davidrodriguez@itm.edu.co

RENDÓN GARCÍA, JUAN FERNANDO
Instituto Tecnológico Metropolitano de Medellín (Colombia)
Correo electrónico: juanrendon@itm.edu.co

TRESPALACIOS CARRASQUILLA, ALFREDO
Instituto Tecnológico Metropolitano de Medellín (Colombia)
Correo electrónico: alfredotrespalacios@itm.edu.co

JIMÉNEZ ECHEVERRI, EDWIN ANDRÉS
Instituto Tecnológico Metropolitano de Medellín (Colombia)
Correo electrónico: edwinjimenez@itm.edu.co

RESUMEN

Los modelos de tipo *Credit Score* permiten a los analistas de crédito la cuantificación de los riesgos que implican las operaciones de crédito, la segmentación de afiliados y la recomendación de decisiones de otorgamiento o rechazo de un crédito para personas naturales. Estos modelos buscan entregar la información necesaria para inferir sobre las probabilidades de impago de un afiliado, mediante la aplicación de técnicas paramétricas o no paramétricas. En este trabajo se busca identificar cuáles de los siguientes modelos pueden ser más apropiados para medir el riesgo de crédito de personas naturales en una caja de compensación familiar ubicada en Colombia: Logit, Probit, Redes Neuronales o Linear Support-Vector Machine. Los resultados obtenidos muestran que, si bien los Linear Support Vector Machine pueden tener mejor desempeño, los modelos Probit-Stepwise son igualmente útiles y tienen como ventaja la posibilidad de interpretar los parámetros calibrados.

Palabras clave: riesgo de crédito, Logit, Probit, red neuronal, support vector machine.
Clasificación JEL: G21; C35; C45.
MSC2010: 91G40; 91G70.

Artículo recibido el 17 de julio de 2020 y aceptado el 2 de agosto de 2021.

Natural People Credit Risk Modeling. An applied case in a Colombian Family Benefit Fund

ABSTRACT

Credit score models quantify the risks in credit operations, customer segmentation, and approve or reject requests to credit customers. These models provide the necessary information to calculate the probabilities of default of any customer through the application of parametric or non-parametric techniques. This work identifies which model (Logit, Probit, Neural Networks, or Linear Support-Vector Machine (L-SVM)) may be more appropriate to measure the credit risk of individuals in a Family Benefit Fund located in Colombia. The results show Linear Support Vector Machine produces better performance, but Probit - Stepwise models are equally useful and they have the advantage of being interpreting the calibrated parameters.

Keywords: Credit Risk, Logit Model, Probit Model, Neural Network, Support Vector Machine.
JEL classification: G21; C35; C45.
MSC2010: 91G40; 91G70

1. Introducción

Las entidades financieras que tienen dentro de sus objetos sociales la colocación de créditos destinados al desarrollo de actividades de inversión o necesidades de gastos de sus clientes, que pueden ser personas naturales o jurídicas, se exponen al riesgo de crédito que Rodríguez y Trespalacios (2015) describen como "la pérdida monetaria generada por la posibilidad de impago en la cartera de los clientes" (p.2). Dentro del proceso de gestión del riesgo de crédito, se requiere establecer metodologías para la toma de decisiones en la asignación de créditos a clientes, que tengan en cuenta las características socioeconómicas de los clientes que puedan ser determinantes de posibles eventos de impago. Los análisis por modelación discriminatoria de clientes basada en *Credit Score*, han sido utilizados ampliamente para la medición del riesgo de crédito de un solicitante de crédito. Este tipo de análisis parte de la selección y medición de las variables socioeconómicas determinantes del riesgo de crédito que sirven como entrada a modelos de *Credit Score* que tienen como salida una variable dicotómica asociada al evento de impago.

Los modelos de *Credit Score* se pueden aplicar en diferentes contextos que tienen en común la necesidad de medir la probabilidad del evento impago a partir de cierta información socioeconómica conocida del solicitante. Rayo et al. (2010) que tiene un enfoque sobre las entidades microfinancieras, o el de Támara et al. (2018) que tiene un enfoque en el sector salud y los impagos en un hospital, son ejemplos de aplicación de modelos de *Credit Score* en diferentes contextos. Dentro del sector financiero - cooperativo enfocado en las cajas de compensación familiar, que a través de los años han podido tener como productos de beneficio a los trabajadores "afiliados", el otorgamiento de créditos. Según la legislación colombiana son entidades que administran y pagan el subsidio familiar (Ley 21, 1982).

Para los casos de aplicación de modelos probabilísticos en entidades de carácter financiero, se realiza una revisión teórica de los conceptos y metodologías aplicables al Credit score con los trabajos de Saavedra-García y Saavedra-García (2010), Abdou y Pointon (2011) y Rodríguez et al. (2017), que muestran una clasificación teórica de los tipos de modelos utilizados, dentro de los que se encuentran: 1) modelos paramétricos, 2) modelos no paramétricos y 3) modelos mixtos o metodologías mezcladas. Adicionalmente, Melo-Velandia y Becerra-Camargo (2005), Saavedra-García y Saavedra-García (2010), Abdou y Pointon (2011), Rodríguez et al. (2017) y Támara et al. (2018) indican que los métodos usados con mayor frecuencia por su efectividad y facilidad de realización son los modelos paramétricos; asimismo, Moreno y Melo (2011) demuestran que los modelos paramétricos son más fáciles de leer, pero pueden tener problemas de ajuste en el pronóstico, en comparación con los modelos no paramétricos que muestran una mejor respuesta de ajuste, pero mayor dificultad en su interpretación.

En contraste, los modelos no paramétricos son usados y recomendados por autores como Desai et al. (1996), Zhai y Russell (1999), Pérez y Fernández (2007), Sustersic et al. (2007), Zhou et al. (2009), Martens et al. (2010), Soydaner y Kocadağlı (2015) y Millán-Solarte y Caicedo-Cerezo (2018), quienes en la mayoría de los casos usan redes neuronales que procesan la información con el fin de minimizar los errores en las discriminaciones, validándose bajo los mismos criterios que los modelos paramétricos para medir su desempeño. Justamente, este criterio ha llevado a la pregunta de cuál puede ser la metodología óptima para aplicar en una medición de Credit score, y es que los autores anteriores no identifican objetivamente cuáles de estos modelos son más apropiados para utilizar, como lo explican Melo y Granados (2011), todas las metodologías tienen pros y contras en el desarrollo y en la comprensión de los resultados haciendo ambigua esta decisión; pero Thomas et al. (2002), Anderson (2007), Hosmer et al. (2013), Hilbe (2015) y Trejo et al. (2017) destacan la inclusión de pruebas de bondad de ajuste comparables bajo ciertos criterios de evaluación, con el fin de identificar el mejor modelo. Un ejemplo de estas pruebas lo constituyen el AUC-ROC y las tablas de confusión.

Tiendo en cuenta las anteriores consideraciones, este trabajo realiza una evaluación y posterior selección de metodologías paramétricas y no paramétricas para la estimación de la probabilidad de impago de personas naturales en una caja de compensación familiar en Colombia con una base de datos de 13.091 afiliados con información del 2018. Los modelos que se evalúan en este estudio son modelos Logit, Probit, Redes Neuronales y Maquinas de Soporte Vectorial lineales (L-SVM, por su sigla en

inglés Linear-Support Vector Machine), usando metodologías de Salazar (2013), Ochoa et al. (2010), Moreno (2013) y Moreno y Melo (2011). Los resultados obtenidos muestran que los modelos Logit, Probit, Redes Neuronales y L-SVM estimados no presentan diferencias significativas en los criterios de bondad de ajuste mostrando valores superiores a 90% en AUC-ROC y valores superiores a 80% en Gini, lo que implica un alto desempeño, mientras que el modelo de Red Neuronal presenta desempeño inferior en términos de ajuste con un criterio de AUC-ROC inferior al 90% y un valor GINI inferior al 80%. Si bien el modelo L-SVM tuvo el mejor desempeño, los modelos Probit, con selección de variables mediante proceso de optimización Stepwise, son igualmente útiles y tienen como ventaja la posibilidad de interpretar los parámetros calibrados. Teniendo en cuenta los criterios de Moreno y Melo (2011) de flexibilidad e interpretación se hace la elección final enfocada en la necesidad de la entidad.

Para entender el procedimiento en este trabajo se hace una descripción del estado del arte tanto en las metodologías de cálculo como en la definición de riesgo de crédito. Luego se muestra la metodología aplicada y la especificación de cada uno de los modelos matemáticos analizados. Posteriormente se realiza una descripción del conjunto de datos facilitados por la caja de compensación familiar colombiana. Finalmente se presentan los resultados obtenidos, las conclusiones y las recomendaciones para trabajos futuros.

2. Evaluación de riesgo de crédito por modelos probabilísticos

La gestión de riesgo de crédito permite a las entidades asegurar de la mejor forma su posible inversión enfocada en los clientes reduciendo el riesgo de castigo de cartera, como lo describen Anderson (2007) y Rodríguez et al. (2017) esta gestión requiere de la implementación de un proceso que va desde la evaluación crediticia de los individuos, hasta la recuperación y la pérdida de la cartera; para ello, el marco general de los procesos de riesgo de crédito involucra la vinculación, la medición, la cobranza, la cobertura y el seguimiento. Y es justamente en el proceso de medición que la modelación del *Credit Score* permite dar un valor de identificación en la mayoría de las veces porcentual a un cliente por sus variables socioeconómicas que identifican si puede o no generar riesgo de impago (basado en respuestas binarias) al mismo portafolio.

La implementación, los parámetros y las condiciones de los modelos de *Credit Score* como lo describen Melo-Velandia y Becerra-Camargo (2005), Saavedra-García y Saavedra-García (2010), Abdou y Pointon (2011), Rodríguez et al. (2017) están sujetos a la naturaleza de la base de datos de los clientes o afiliados; son estos últimos los que determinan qué signo o qué camino van a tomar los análisis de variables según se constituya la información de las entidades analizadas, e incluso, el hecho que existan tipos de personas (personas naturales o jurídicas) en los análisis, puede afectar los resultados esperados, pero que pueden ser justificados ampliamente con previo estudio de dichas variables. Es por esto por lo que Basilea II y Basilea III dan libre albedrío al uso de la metodología de análisis a fin de llegar a un resultado que favorezca tanto a la entidad como al afiliado, siempre y cuando se tengan en cuenta las condiciones de las variables, las políticas empresariales y los objetivos del análisis de riesgo que se desea desarrollar.

Sobre su evolución metodológica, los análisis de *Credit Score* se pueden distribuir en tres etapas de tiempo. El primero comprende los inicios de los modelos de análisis discriminatorio (desde los años 30 hasta los años 60), que son modelos de riesgo post guerra aplicados a la medición de probabilidad de bancarrota de entidades bancarias tomando en cuenta variables de rubro empresarial, como así lo demuestran los trabajos de Fisher (1936) y Myers y Forgy (1963).

La segunda etapa se inicia con Altman (1968) que abordó mejoras en los procesos anteriores, identificando por medio de LDA (Linear Discriminatory Analysis) un modelo lineal que entrega un puntaje de referencia para quiebras (Z-Score) por medio de análisis parámetrico. Este fue el inicio de los modelos de *Credit Score* que se convirtieron con el tiempo en modelos específicos para la medición de riesgo de crédito para clientes. Posteriormente Orgler (1970), Apilado et al. (1974) y Altman (1980)

demostraron que dichos modelos lineales son eficientes en la revisión de los parámetros adicionando la probabilidad de éxito (usando variables dicotómicas) y variables socioeconómicas a los modelos para determinar el riesgo de impago de un afiliado de carácter natural a los análisis.

El tercer escenario se puede mencionar en los estudios que se enfocan en los procesos de exploración de modelos paramétricos similares, como lo demuestran Rodríguez et al. (2017), desarrollados en su mayoría por Glorfeld (1990), Lipovetsky y Conklin (2004), Roszbach (2004), Gonçalves y Braga (2008), Olagunji & Ajiboye (2010), Constangioara (2011), Melo & Granados (2011), Mures et al. (2011), Palacio et al. (2011), Webster (2011), Chaudhuri y Cherical (2012), Puertas y Marti (2012), Moreno (2013), Rodríguez y Trespalacios (2015) y Trejo et al. (2017) los cuales trabajan con modelos Logit, Probit, Tobit, Multinomiales, Logit Mixtos, Modelos Lineales Probabilísticos (MLP), Modelos Lineales Discriminatorios (LDA), Modelos Least - Absolute - Value (LAV), que tienen como elemento práctico ser fáciles de recrear, analizar y operar.

Así mismo, como una variante de este último escenario a su vez, los modelos no paramétricos muestran una diferencia sustancial con los trabajos de Desai et al. (1996), Zhai & Russell (1999), Pérez y Fernández (2007), Sustersic et al. (2007), Zhou et al. (2009), Martens et al. (2010) y Soydaner y Kocadağlı (2015) que usan métodos de minimización de error por iteración y creación generacional a fin de encontrar el mejor modelo aplicado. Para estos trabajos se encuentran modelos de redes neurales de tipo genético, modelos de Fuzzy Logic, Support Vector Machines, arboles binomiales y modelos híbridos como modelos CART o modelos de segmentación CHAID.

3. Metodología aplicada

La modelación econométrica de riesgo de crédito aplicada está basada en el desempeño histórico de una canasta de afiliados que en términos de Anderson (2007) y Hosmer et al. (2013) se pueden realizar con tiempo estático asumiendo que todos los usuarios de dicha base de información son clientes activos inmediatos de una entidad por evaluar. Los enfoques elegidos para determinar un modelo apropiado discriminatorio binomial se soportarán de modelos Logit, Probit, redes neuronales y Linear - Support Vector Machine como se observan en trabajos de estilo comparativo de Akkoç (2012) o de Moreno y Melo (2011) que usan criterios estadísticos de bondad de ajuste para medir qué modelo fue óptimo comparativamente. Para esta selección, se tendrán en cuenta los criterios de bondad de ajuste adecuados por Hosmer et al. (2013), Anderson (2007) y Moreno y Melo (2011) para el manejo y elección del modelo óptimo.

3.1 Análisis estadístico previo

Para garantizar una calibración sin sesgo, es indispensable usar los criterios descritos por Anderson (2007), donde se realiza una previa revisión estadística de las variables socioeconómicas presentes en la base de datos con la que se realiza el modelo. Para ello, se requiere como insumo básico la información de los clientes o afiliados, verificando que exista una cantidad de clientes que identifiquen el criterio actual de calificación de crédito, ejecutado con la mayor cantidad de variables que lo describan; esta información proviene de la institución financiera analizada. Una vez obtenida la información, se realiza una revisión de estadística descriptiva en variables cuantitativas y cualitativas, a fin de establecer si las series presentan datos atípicos que deban ser corregidos, o es posible identificar el tipo de distribución que posea cada variable; adicionalmente, estos análisis son útiles para hacer transformaciones a las series de variables cuantitativas para contemplar reducción de información o cambios de estructura en su análisis y pasarlas a variables cualitativas. Frente a los análisis de series cualitativas, solo se requiere identificar el porcentaje que corresponde cada categoría de la variable analizada frente a la variable Y_i. Estos análisis en conjunto permiten identificar la naturaleza del signo esperado en la modelación según la naturaleza de la base de información.

Posteriormente, es necesario realizar un análisis de correlación o colinealidad de todas las variables incluyendo a la variable Y_i, para ello, Anderson et al. (2008) sugieren realizar análisis de correlación de Pearson para variables cuantitativas; si las variables presentan alto nivel de correlación se sugiere eliminar dichas variables. Respecto a las variables cuantitativas Taylor y Chappell (1980) sugieren una variante de análisis de correlación para variables categóricas conocida como lambda (λ) de Goodman- Kruskal. Una vez identificadas las variables con bajo nivel de similitud, se identifica por medio de gráficas de análisis jerárquico y dendogramas la importancia de dichas variables según su respuesta frente a la variable Y_i,, una herramienta útil para verificar la idoneidad en la significancia estadística de cada variable previa a la calibración de los modelos.

3.2 Modelo LOGIT

Los modelos Logit realizan la transformación de un modelo lineal probabilístico multivariado de tal forma que la probabilidad de impago del afiliado esté acotada entre valores de cero y uno; esto se logra a través de una distribución logística estándar o sigmoide. Si P_i es la probabilidad de impago del afiliado i y puede ser descrita por un modelo tipo Logit, puede ser calculada con la ecuación [1], con $x_{1i}, x_{2i}, \dots, x_{ki}$ las variables que explican el comportamiento de pago del afiliado.

$$P_i(Y = 1 \mid X_{ik}) = \frac{1}{1 + e_i^{-z_i}};\ z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \qquad [1]$$

Los parámetros $\beta_1, \beta_2, \dots, \beta_k$ explican la sensibilidad de incremento o decremento del riesgo que tiene el movimiento de cada variable explicativa sobre la probabilidad de pago. Un valor positivo de β_i implica que, cuando la variable x_i aumenta para el individuo i, aumenta la probabilidad de impago del individuo i.

3.3 Modelo PROBIT

Como los modelos Logit, los modelos Probit transforman un modelo lineal probabilístico a través de una transformación de distribución de probabilidad acumulada en una distribución normal estándar. Si P_i es la probabilidad de impago del afiliado i y puede ser descrita por un modelo tipo Probit y puede ser calculada con la ecuación [2], con $x_{1i}, x_{2i}, \dots, x_{ki}$ las variables que explican el comportamiento de pago del afiliado.

$$\begin{aligned} P_i(Y = 1 \mid X_{ik}) = F(x) &= \frac{1}{\sqrt{2\pi\sigma_I^2}} \int_{-\infty}^{I_i} e^{\frac{Z^2}{2\sigma^2}}\, dZ;\ I_i \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \end{aligned} \qquad [2]$$

Similar a la condición en los modelos tipo Logit, los parámetros $\beta_1, \beta_2, \dots, \beta_k$, en el modelo Probit, explican el aporte que tiene el movimiento de cada variable explicativa sobre la probabilidad de impago. Un valor positivo de β_i implica que, cuando la variable x_i aumenta, para el individuo i, aumenta la probabildad de impago del individuo i.

3.4 Modelo de Red Neuronal (RN)

Este modelo hace uso de una red con centroide logístico con número de capas y número de nodos variables. En la ecuación [3] se presenta la especificación matemática para una red neuronal de una capa que define la probabilidad de impago de un afiliado $f_k(x)$, a partir de N variables exógenas x_i. En la ecuación [3], α representa el cambio de nivel, $\mathrm{w_{ik}}$ son los parámetros de peso que y f_0 la función de transformación.

$$f_k(x) = f_0\left(\alpha + \sum_{i=1}^{N} \mathrm{w_{ik}}\, x_i\right) \quad [3]$$

3.5 Linear Support-vector Machine - L-SVM

Esta representación, que se muestra en la ecuación [4], se corresponde con un clasificador. Este algoritmo entrega el hiperplano óptimo para segmentar la muestra. En la ecuación [4], w es un vector que contiene los pesos, x es una matriz (que tiene tantas filas como observaciones y tantas columnas como variables) que contiene las N variables exógenas con todas sus observaciones y b elvector de intercepto. El resultado obtenido del hiperplano y del signo define la posición de la variable dicotómica ajustada: se define $+1$ para identificar las personas que generan impago, -1 para identificar las personas que no lo generan.

$$y(x) = sign(w^T x + b);\ \begin{cases} w^T x + b \geq +1, si\ y_k = +1 \\ w^T x + b \leq -1, si\ y_k = -1 \end{cases} \quad [4]$$

3.6 Estimación y calibración de los modelos

El proceso de estimación utiliza dos conjuntos de datos, el primero se denomina datos de entrenamiento (Train data) y el segundo, datos de verificación (Test data), coherente con las propuestas realizadas por Hosmer et al. (2013) y Moreno (2013). El conjunto de entrenamiento que corresponde con el 80% de las observaciones disponibles, es seleccionado de forma aleatoria y utilizado para estimar los parámetros del modelo; mientras que el 20% restante de la información pertenece al conjunto de verificación, con el que se hace la validación del desempeño del modelo estimado.

Con el conjunto de entrenamiento, para los modelos Logit, Probit se selecciona el conjunto de variables explicativas utilizando el proceso "*Stepwise - backforward*" que permite agregar cada variable e internamente se analiza con criterios de información AIC y BIC (Akaike y Schwartz) cuyo resultado es que el mejor modelo corresponde al de menor valor. En el caso de la red neuronal y el L-SVM, el conjunto de entrenamiento es utilizado no solo para estimar los pesos del modelo sino también para obtener el numero óptimo de nodos y número de capas, mediante un proceso de análisis iterativo que permita así mismo encontrar el menor valor de criterio de información.

Los modelos óptimos presentan variables totalmente significativas y que no muestran coeficientes excesivamente sensibles; si esto se presenta significa que la calibración del modelo todavía presenta variables colineales con la variable Y_i. En algunos casos, las variables cualitativas, al tener varias categorías, pueden presentar categorías con más datos que otras, afectando su significancia individual en la prueba de Wald, pero no afectando su significancia en el modelo; para ello se somete al análisis de criterios de información AIC - BIC con el fin de conocer el impacto de la variable en el modelo y posteriormente ordenar las categorías evidenciando su real significancia individual.

3.7 Análisis de bondad de ajuste de los modelos estimados

Para el análisis de bondad de ajuste de los modelos *Credit Score*, Hosmer et al. (2013) y Anderson (2007) sugieren el uso de la curva AUC-ROC (Area Under the Curve of Reciever Operating Characteristic), el coeficiente de GINI y la matriz de confusión ("*confusión matrix*") y determinan los valores de aceptación críticos. Por ejemplo, en términos de desempeño de la función discriminante de un modelo de *Credit Score.* Hosmer et al. (2013) establece que el valor mínimo del AUC para que el modelo sea aceptado es del 80%. Un criterio similar para la evaluación del desempeño de los modelos de *Credit Score* es el coeficiente de GINI, entendido como una métrica de variable real definida en el intervalo [0,1], donde 0 se refiere a un modelo totalmente aleatorio en la discriminación y 1 un modelo que discrimina la información de forma perfecta. Por otro lado, la matriz de confusión permite medir la capacidad que tiene el modelo para clasificar verdaderos positivos y verdaderos negativos a través de

la sensibilidad y la especificidad respectivamente. Cuanto más se aproximen estas métricas al 100%, mejor desempeño presenta el modelo estimado.

4. Aplicación y resultados

El siguiente apartado muestra el proceso de desarrollo de los modelos estimados desde la elección de variables, hasta la resolución de elección del modelo óptimo para la base de clientes analizada.

4.1 Descripción de los datos

La información obtenida para este estudio corresponde a la cartera de afiliados del año 2018, compuesta por 13.091 registros de personas naturales con clientes activos hasta ese año, de una Caja de Compensación Familiar del departamento de Caldas en la República de Colombia. Con el ánimo de identificar la relevancia de las variables sobre el comportamiento de cultura de pago de los afiliados, se trabaja con las siguientes variables: el plazo de la deuda, la tasa de interés pactada, el tiempo laboral, la edad de la persona, el score interno, el índice de rotación y el número de personas a cargo, que aparecen en la Tabla 1. Respecto a las variables cualitativas se encuentran: producto, tipo de trabajo, tamaño de la empresa a la que está vinculado el afiliado, nivel educativo, estado civil, score en CIFIN (entidad que evalúa la historia crediticia de las personas en todo el sistema financiero colombiano), nivel de salario, franja de morosidad, agencia, tipo de vivienda, tipo de contrato, calificación de provisión y estrato. La asignación de categorías a cada una de estas variables se presenta en la Tabla 2.

Tabla 1. Estadística descriptiva información cuantitativa.

	Variable	min	max	mean	std.dev	coef.var
Información total	Plazo	1	96	32.84	20.7	0.63
	Tasa de Interés	5.2	25.2	16.98	4.15	0.24
	Tiempo Laboral	0	415.6	76.5	75.16	0.98
	Edad	19	82	39.42	9.07	0.23
	Score Interno	0	95	65.46	17.3	0.26
	Índice Rotación	0	1	0.21	0.17	0.82
	Personas a Cargo	0	8	1.76	1.33	0.76
	Tipo de Afiliación	0	443	69.89	74.12	1.06
Información de afiliados que generaron riesgo	Plazo	1	96	23.02	15.59	0.67
	Tasa de Interés	5.8	24	17.42	3.49	0.2
	Tiempo Laboral	0	361.3	30.99	53.2	1.71
	Edad	20	82	38.85	9.38	0.24
	Score Interno	0	90	36.21	32.44	0.89
	Índice Rotación	0	1	0.34	0.28	0.81
	Personas a Cargo	0	6	0.57	1.11	1.94
	Tipo de Afiliación	0	387	45.77	51.6	1.12

Nota: Los siguientes son las descripciones de las variables en su orden. 1. Plazo, tiempo acordado por las partes para amortizar el crédito (en meses). 2. Tasa de Interés, tasa acordada por la entidad (efectiva anual). Tiempo Laboral, tiempo de trabajos consecutivos en la empresa (en meses). 4. Edad, años de vida del afiliado. 5. Score Interno, calificación de créditos anteriores de la entidad del afiliado. 6. Índice de Rotación, es la rotación laboral descrita por la entidad del afiliado. 7. Personas a Cargo, cantidad de personas que dependen del afiliado. 8. Tipo de Afiliación del afiliado a la entidad. Estas son variables de real importancia para el estudio y se omiten las estadísticas de las variables descartadas posteriormente en la figura 1b.

Fuente: Elaboración propia.

Tabla 2. Estadística descriptiva información cualitativa y porcentaje de impagos por categoría.

Categoría	Total	Si R	%	No R	%	Categoría	Total	Si R	%	No R	%
Producto						Franja de Morosidad					
Familiares	6955	494	7	6461	93	Al día	11782	0	0	11782	100
Educación	2422	237	10	2185	90	1_30	367	0	0	367	100
Vivienda	1450	78	5	1372	95	31_60	89	89	100	0	0
Compra cart	1386	33	2	1353	98	61_90	55	55	100	0	0
Salud	440	21	5	419	95	91_180	90	90	100	0	0
Recreación	183	8	4	175	96	181_360	122	122	100	0	0
(Otro)	255	71	28	184	72	360_mas	586	586	100	0	0
Tamaño Empresa						Tipo de Vivienda					
Grande	6281	349	6	5932	94	Alquiler	3866	376	10	3490	93
Mediana	3245	168	5	3077	95	Familiar	4915	346	7	4569	91
Pequeña	2783	274	10	2509	90	Propia	4269	181	4	4088	85
Micro	782	151	19	631	81	Sin_info	41	39	95	2	89
Nivel Educativo						Agencia					
Ninguno	21	1	5	20	95	Manizales	10689	706	7	9983	93
Primaria	756	100	13	651	87	Chinchiná	537	49	9	488	91
Secundaria	5790	497	9	5235	91	La Dorada	534	82	15	452	85
Pregrado	3230	166	5	3048	95	Riosucio	383	44	11	339	89
Tecni/tecn	2817	133	5	2666	95	Villamaría	255	13	5	242	95
Postgrado	536	8	1	527	99	Supia	198	21	11	177	89
Sin_info	39	37	95	2	5	(Other)	495	27	5	468	95
Estado Civil						Tipo Contrato					
Casado	3271	178	5	3093	95	Carre_adm	636	20	3	616	93
Separado	332	17	5	315	95	Obra labor	2131	400	19	1731	91
Soltero	5476	361	7	5115	93	Otros	230	49	21	181	85
Unión libre	3976	383	10	3593	90	Term fijo	5275	346	7	4929	89
Viudo	36	3	8	33	92	Term ind	4819	127	3	4692	85
Score CIFIN						Tipo de Empresa					
< 400	1301	176	14	1125	86	Publica	1586	80	5	1506	95
400-699	900	121	13	779	87	Privada	10986	733	7	10253	93
> 700	2063	109	5	1954	95	Mixta	263	2	1	261	99
sin_info	8827	536	6	8291	94	sin_info	256	127	50	129	50
Calificación provisión crediticia						Estrato Social					
A	12145	0	0	12145	93	e0	126	14	11	112	93
B	91	87	96	4	91	e1	1691	122	7	1569	91
C	56	56	100	0	85	e2	4751	301	6	4450	85
D	91	91	100	0	89	e3	4127	182	4	3945	89
E	708	708	100%	0	85	e4	641	17	3	624	85
Categoría Salario Afiliado						e5	142	3	2	139	83
< 2 SMLV	9833	865	9	8968	91	e6	106	3	3	103	81
2-4 SMLV	2552	62	2	2490	98	sin_info	1507	300	20	1207	80
> 4_SMLV	706	15	2	691	98						

Nota: Los siguientes son las descripciones de las variables en su orden. 1. Producto, refiere al producto crediticio obtenido. 2. Franja de Morosidad, muestra el tiempo que ha tenido un afiliado i en demora de pago. 3. Tamaño Empresa, se refiere al tamaño empresarial decreto 957/2019 en donde trabaja el afiliado i. 4. Tipo Vivienda, indica el tipo de residencia del afiliado i. 5. Nivel Educativo, muestra el nivel educativo alcanzado por el afiliado. 6. Agencia, sede donde se firma el contrato de préstamo del afiliado. 7. Estrato Civil, estrato civil del afiliado. 8. Tipo de Contrato, Tipo de contrato actual del afiliado. 9. Score CIFIN, muestra el calificativo manejado por CIFIN (Central de Información Financiera). 10. Tipo de Empresa. Tipo de sociedad según propiedad de capital. 11. Calificación Provisión Crediticia, se refiere a la calificación crediticia dada por una entidad de riesgos. 12. Estrato, estrato social en el que reside el afiliado.

Fuente: Elaboración propia.

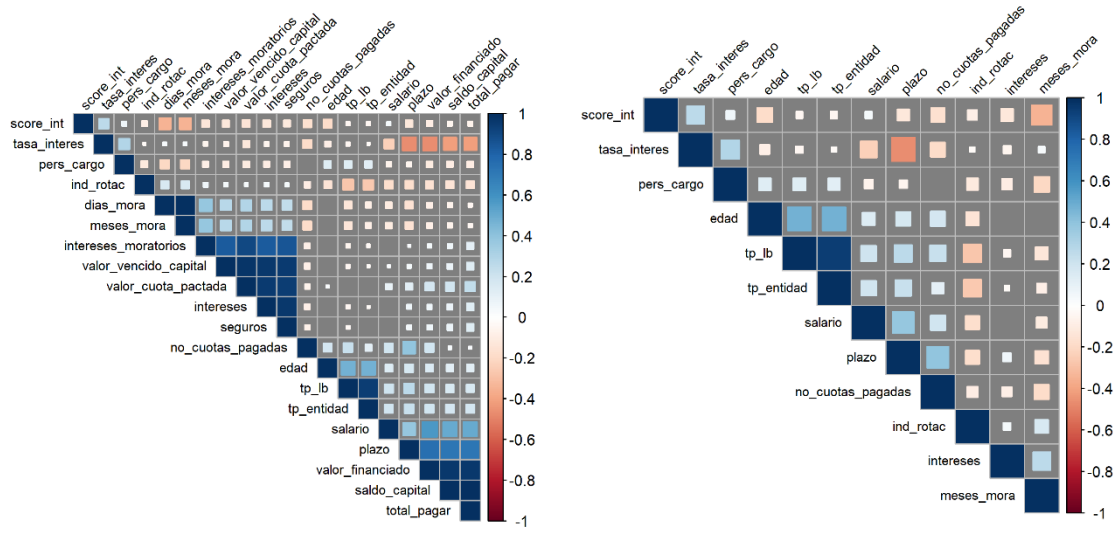
4.2 Análisis estadístico previo (elección de variables optimas)

Para el desarrollo de este apartado, se secciona el análisis en dos estadíos, siendo el estadío 1) el análisis devariables cuantitativas. Para este procedimiento se efectúa el análisis de correlación de Pearson de variables descrito en la metodología y cuyos resultados se aprecian en la Figura 1.a: las variables que presentaran valores superiores a |0.5| fueran nuevamente revisados y posteriormente descartados verificando su relación entre sí. Las variables que no tienen amplios niveles de correlación se presentan en la Figura 1.b.

Se debe aclarar entonces que las variables meses de mora y score interno, solo presentan correlación nominal, no estructural, debido a que trabajan inversamente según el tiempo que tenga el afiliado en vigencia en la entidad, lo que las hace similares y, por tanto, no se deben excluir mutuamente; lo mismo sucede con el plazo y la tasa de interés, que parece mostrar una correlación negativa (a mayor plazo, menor tasa de interés).

El estadío número 2) hace referencia al análisis de variables cualitativas. Para determinar la correlación en las variables cualitativas se utiliza la lambda (λ) de Goodman - Kruskal, que consiste en un método similar de correlación basándose en tablas de contingencia, usando los mismos valores de correlación [−1,1]. Las variables que se descartan por su alta correlación entre ellas son: franja de morosidad, calificación de provisión, calificación interna afiliado, forma de pago, estado del saldo y agencia y zona, como se muestra en la Tabla 3.

Figura 1. Graficas de correlaciones entre variables cuantitativas.



(a) (b)

Fuente: Caja de compensación Familiar (2019).

Para finalizar el proceso de elección de variables, se vuelve a establecer una prueba de correlación de lambda (λ) de Goodman – Kruskal, para las variables cuantitativas y cualitativas depuradas para verificar una posible correlación entre los tipos de variables. A partir de este se encuentra que las variables óptimas para los modelos son: Agencia, Clasificación salarial afiliado, Edad, Educación, Estado civil, Estado de la empresa, Estrato, Forma de pago, Género, Índice de rotación, Personas a Cargo, Plazo, Producto, Score CIFIN, Score Interno, Tamaño de empresa, Tiempo laboral, Tipo de contrato, Tipo de empresa y Tipo de vivienda.

Tabla 3. Correlación entre variables cualitativas.

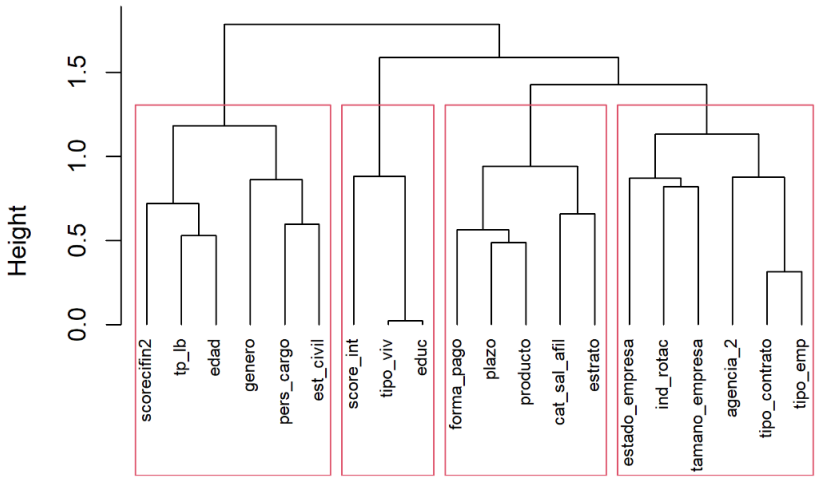
Variables	Zona	Franja Morosidad	Calificación provisión	Calificación interna	Estado Saldo	Forma Pago	Y_i
Franja Morosidad			99%	58%	93%		100%
Calificación provisión		62%		58%	92%		100%
Calificación interna		40%	64%		72%		74%
Estado Saldo		45%	74%	53%			75%
Agencia	100%						
Rango Financiado						60%	

Nota: Las celdas en blanco referencian valores inferiores a 50% de correlación. Solo se muestran las variables que reflejaron valores superiores a 50%.

Fuente: Elaboración propia.

Por medio del dendograma de jerarquización de información, que permite visualizar el nivel de importancia de las variables, en su orden el primer grupo de variables se compone de tipo de contrato, tipo de vivienda, educación, score interno y estado de la empresa; el segundo grupo se compone de estado civil, tamaño de la empresa, tiempo laboral, edad y score CIFIN; el tercer grupo incluye agencia / zona, estrato social, personas a cargo y género; y el cuarto grupo se compone de forma de pago, plazo, producto, índice de rotación, categorial salarial del afiliado y tipo de empresa, esto con el fin de identificar qué variables tienen un impacto importante en los parámetros del modelo y así descartar los que estadísticamente no cumplan condiciones de calibración.

Figura 2. Dendograma de las variables, proceso de jerarquización de variables.



Fuente: Caja de compensación Familiar (2019).

4.3 Estimación y resultados de los modelos

De la información suministrada con 13,091 registros, con el 7.19% de estos representando afiliados en condiciones de no pago de sus obligaciones -*Default*. La Tabla 4 muestra cómo esta información se divide en proporción 70%/30% para identificar la serie de información aleatoria del "Data Train" y del "Data Test" para calcular y probar la eficacia de los modelos.

Tabla 4. Datos usados para realizar el estudio de los modelos.

Data	No Default	Default	Total
N	12149	942	13091
Train Data	8513	649	9162
Test Data	3636	293	3929

Fuente: Caja de compensación Familiar (2019).

Respecto a los resultados de los modelos, la Tabla 5 muestra los resultados de los parámetros de las variables evaluadas en cada tipo de modelo. Se evidencia que, en su mayoría las variables identificadas previamente para los modelos paramétricos fueron eficientes, identificando que las variables: educación, tipo de vivienda y zona no son significativas y se pueden descartar; sin embargo, en el caso de los modelos No Paramétricos, todas las variables muestran efecto de manejo.

Tabla 5. Modelos Calibrados con Train Data.

Variables	Categoría	Logit		Probit		L-SVM	Neural Network	
		$\hat{\beta}$	P-value	$\hat{\beta}$	P-value	W	Hidden1	Hidden2
Intercept		5.095	0.000	2.088	0.000	-0.13	1.423	0.739
Categoría Salarial	Hasta_2_smlv	-----	-----	0.222	0.027	2904.94	1.343	0.253
Afiliado	Mas_4_smlv	-----	-----	-0.079	0.722		-11.915	0.061
Edad		-0.014	0.041	-----	-----	57035.06	-6.144	-0.185
Educación	Posgrado	-----	-----	-----	-----	6891.65	3.383	2.685
	Pregrado	-----	-----	-----	-----		9.461	-1.712
	Primaria	-----	-----	-----	-----		16.235	-2.024
	Secundaria	-----	-----	-----	-----		-6.056	-1.855
	Técnico/tecnólogo	-----	-----	-----	-----		-0.081	-1.832
Estado civil	Casado	-0.247	0.145	-0.145	0.072	4581.49	4.519	0.231
	Soltero	-0.834	0.000	-0.440	0.000		-5.591	0.314
	Separado	-1.151	0.006	-0.698	0.001		15.747	0.106
	Viudo	-0.332	0.709	-0.271	0.547		8.622	-0.124
Estrato social	0	-0.156	0.735	-0.104	0.658	5246.04	-5.138	-1.636
	1	-0.810	0.000	-0.440	0.000		6.394	-0.04
	2	-0.904	0.000	-0.456	0.000		-0.726	0.307
	3	-0.981	0.000	-0.512	0.000		5.872	0.292
	4	-1.050	0.005	-0.497	0.005		-9.65	0.396
	5	-1.558	0.145	-0.866	0.112		-5.847	0.453
	6	-1.199	0.124	-0.605	0.138		6.427	0.41
Género	Masculino	0.621	0.000	0.302	0.000	2390.03	12.415	-0.443
Índice Rotación		0.705	0.019	0.412	0.008	433.78	21.006	2.983
Personas a cargo		-0.926	0.000	-0.449	0.000	1354.14	8.468	2.042
Plazo		-0.060	0.000	-0.028	0.000	32441.2	-10.791	-3.67

Producto	Compra cartera	-0.427	0.224	-0.229	0.209	4888.56	-2.831	-0.08
	Educación	-0.638	0.017	-0.370	0.009		-12.471	0.259
	Familiares	-0.868	0.000	-0.490	0.000		-6.7	0.284
	Recreación	-2.212	0.003	-1.108	0.003		4.74	-0.135
	Salud	-1.175	0.004	-0.660	0.002		1.096	0.677
	Vivienda	-1.022	0.001	-0.599	0.000		5.97	0.213
Score Interno		-0.055	0.000	-0.028	0.000	96137.45	-7.12	4.424
Score CIFIN	Menor_400	1.000	0.000	0.488	0.000	2470.03	6.074	-0.442
	400_699	1.012	0.000	0.523	0.000		-11.394	-0.422
	Mayor_700	0.707	0.000	0.340	0.000		-2.14	-0.021
Tamaño Empresarial	Pequeña	-0.498	0.011	-0.268	0.008	3453.98	-0.411	0.088
	Mediana	-0.723	0.000	-0.416	0.000		0.603	0.034
	Grande	-0.540	0.004	-0.288	0.003		-13.474	-0.179
Tasa de Interés		0.090	0.000	0.045	0.000	26341.6	-4.914	0.346
Tipo Contrato	Obra labor	1.988	0.000	1.015	0.000	4901.13	34.197	2.98
	Carrera admón.	0.503	0.220	0.323	0.105		1.755	-0.601
	Término fijo	0.928	0.000	0.445	0.000		1.949	-0.035
	Otros	0.794	0.083	0.361	0.104		3.65	-0.19
Tipo Empresa	Público	-0.622	0.064	-0.295	0.099	4110.12	-1.516	0.498
	Privado	-1.250	0.000	-0.684	0.000		8.456	-0.008
	Mixto	-1.871	0.057	-0.969	0.039		18.983	0.101
Tipo Vivienda	Alquiler	-----	-----	-----	-----	4272.61	-6.028	0.375
	Familiar	-----	-----	-----	-----		-2.257	2.671
	Propia	-----	-----	-----	-----		13.453	2.562
Tiempo en Entidad		0.012	0.000	0.006	0.000	62140.57	-14.484	-0.365
Tiempo Laboral		-0.023	0.000	-0.011	0.000	55861.3	-3.64	-2.751
Zona	Centro sur	-----	-----	-----	-----	4603.82	-12.737	-0.13
	Manizales	-----	-----	-----	-----		4.197	0.331
	Norte	-----	-----	-----	-----		-22.269	0.449
	Occidente	-----	-----	-----	-----		15.817	0.478
	Oriente	-----	-----	-----	-----		18.231	-0.136
	Sur occidente	-----	-----	-----	-----		-10.083	0.542
AIC		2470.2		2504.9		2387.3	2540.3	
BIC		2755.1		2796.9		2101.6	2674.5	

Nota: Las variables están en orden alfabético y no por tipo de variable (cuantitativa/cualitativa). Para los modelos Logit y Probit se presenta el modelo estimado por "Stepwise Brack-forward". Los parámetros con "-----" muestran que la variable no tuvo significancia alguna en el modelo.

Fuente: Elaboración propia.

Respecto a los parámetros de los modelos Logit y Probit se encuentra que algunos signos parecen ser contraintuitivos como lo fueron las variables plazo solicitado o tiempo de afiliación en la entidad,

pero estos signos los determina la base de información de afiliados en la entidad. Por ejemplo, el plazo de pago de los créditos es inverso: por evidencia empírica (sustentada en la Tabla 1) los clientes de esta entidad tienen mejores comportamientos de pago a largo plazo que a corto plazo. Así mismo, el tiempo de afiliación en la entidad muestra dos elementos propios de este tipo de entidades: 1) Al ser una entidad del sector cooperativo y no comercial tiene preferencias sobre los afiliados más antiguos y pueden tomar medidas laxas al momento de crear un crédito. 2) El sector cooperativo no visiona a sus afiliados como "clientes" mostrando políticas diferenciales sobre el proceso de la toma de decisión del comité de crédito.

Los resultados de bondad de ajuste se presentan en la Tabla 6, teniendo los siguientes criterios. Respecto al indicador AUC-ROC el modelo que presenta el mejor criterio es el modelo L-SVM, pero, los modelos paramétricos no difieren mucho del valor presentado de este indicando que si bien presentan menor eficiencia son ciertamente similares. Solamente el resultado de la red neuronal muestra una eficiencia inferior a los demás modelos. Respectivamente el coeficiente de Gini muestra resultados similares al AUC-ROC; el modelo con mejor explicación corresponde al L-SVM, secundado por el modelo Probit, esto muestra que la diferencia entre ambos modelos en términos predictivos no es sensible y discriminativa.

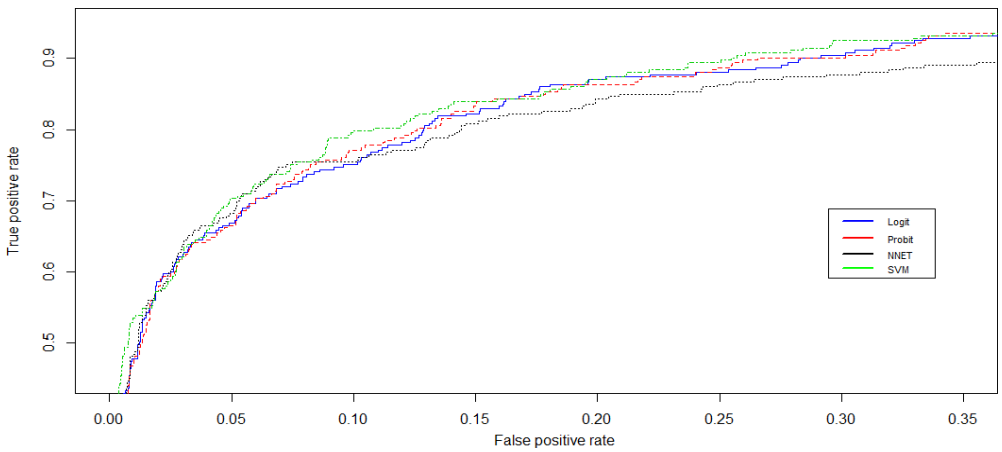
Los modelos paramétricos muestran una mayor estabilidad en la lectura de los indicadores de especificidad y sensibilidad venidos de la tabla de confusión; en contraste, los modelos no paramétricos presentan una mayor estimación sobre los valores de no riesgo (especificidad), si bien esto no representa un problema real, sí refleja que los modelos no paramétricos requieren de una serie de información más ecuánime para arrojar predicciones más estables. Este efecto también se puede evidenciar en la Figura 3 en la que los modelos L-SVM son mucho más eficientes, pero claramente no se diferencian de los modelos Logit o Probit.

Tabla 6. Indicadores de desempeño de los modelos.

	AUC-ROC	Gini	Especificidad	Sensibilidad	AIC	BIC
Logit	91.51	83.02	87%	82%	2470.2	2755.1
Probit	91.65	83.30	85%	84%	2504.9	2796.9
L-SVM	92.02	84.04	91%	79%	2387.3	2101.6
Red Neuronal	89.17	78.34	96%	78%	2540.3	2674.5

Fuente: Caja de compensación Familiar (2019).

Figura 3. Curvas AUC-ROC de los modelos.



Fuente: Caja de compensación Familiar (2019).

Respecto a la elección del modelo óptimo para la entidad, se establecen los criterios cualitativos del estudio de Moreno y Melo (2011); se encuentra que los modelos Logit y Probit son mucho más sencillos de leer debido a que estos se pueden transformar a lectura porcentual con el método de marginalización de parámetros, mientras que los pesos de los modelos Red Neuronal y L-SVM, cuyos valores solo identifican una capacidad de repetición, no tienen una interpretación definible o entendible.

Respecto al criterio de sencillez, los modelos probabilísticos son más prácticos y directos que los modelos Red Neuronal y L-SVM, tanto en su construcción como en su aplicación: para los modelos paramétricos estos valores son fácilmente replicables en software de uso común, en cambio los procesos de L-SVM y Red Neuronal mantienen un proceso oculto cuya programación se vuelve compleja en entornos sencillos. Finalmente, frente a la flexibilidad, los modelos Logit y Probit tienen desventaja debido a que su proceso es estático y deben ser reestimados según el corto o medio plazo, así mismo los modelos de redes neuronales poseen este mismo inconveniente y teniendo en cuenta que cuanta más información se suministre a futuro, estos tienden a sobreestimarse generando inconsistencia en los resultados. Ahora bien, los L-SVM son altamente flexibles al agregar nueva información, pero sigue siendo un proceso no visible y que requiere un reentrenamiento inmediato de los datos cada vez que se ingrese una unidad.

Con los anteriores criterios se infiere que los modelos L-SVM no son los más adecuados en aplicarse en los estudios de créditos de la entidad así sus resultados hayan sido óptimos estadísticamente, pues, no generaron criterios de especificidad y sensibilidad mejores que los modelos Logit o Probit y su manejo post calibración, por tanto, los modelos más estables para esta entidad pueden ser los modelos Logit o Probit. Estableciendo esto, se decide realizar un último análisis de estabilidad la serie Test Data para establecer cuál de estos puede ser más estable en diferentes niveles de series de información. En la Tabla 7 se muestra cómo los modelos Probit en una prueba de stress con porcentajes de datos aleatorio se ve cómo maneja mayor estabilidad que los modelos Logit, sosteniendo mejor la sensibilidad y especificidad.

Tabla 7. Análisis de estabilidad con testeo aleatorio de "Test Data" en porcentaje (%).

		5%	10%	15%	20%
Logit	AUC-ROC	94.48	95.21	89.24	91.95
	Especificidad	90.25	89.75	87.56	87.90
	Sensibilidad	86.96	86.92	77.27	82.99
Probit	AUC-ROC	94.31	94.96	89.18	92.18
	Especificidad	81.98	90.00	88.27	89.39
	Sensibilidad	93.48	86.92	78.03	80.93

Fuente: Caja de compensación Familiar (2019).

4.4 Discusión de los resultados obtenidos.

El anterior trabajo es una aplicación ejercida en una Caja de compensación Familiar en Colombia, dichas entidades que se enmarcan en el sector cooperativo son muy conocidas en el país, pero no tienen un símil perfecto en otros países. Los trabajos que se pueden acercar en términos del objeto de estudio son aquellos que desarrollan análisis de Credit score en cooperativas financieras basándose en el tratamiento de los "afiliados" como lo son los trabajos que se utilizaron como base para crear la metodología, tales como Salazar (2013), Ochoa et al. (2010), Moreno (2013) y Moreno y Melo (2011).

La primera consideración destacable en contraste de los trabajos revisados concierne a los resultados de los modelos estimados en los trabajos anteriores. En el caso de Salazar (2013) se presenta un enfoque a los modelos Logit, en el caso de Ochoa et al. (2010) se enfoca en un modelo MLP (Modelo Lineal Probabilístico), con Moreno y Melo (2011) el mejor modelo fue Support Vector Machine y

posteriormente con Moreno (2013) se identifica un modelo Logit como mejor modelo aplicable. Esto resalta el hecho de que las metodologías si bien pueden ser adecuadas bajo los contextos estadísticos de la bondad de ajuste, los datos y las empresas son las que determinan qué metodología es mejor. En el caso de este trabajo, estadísticamente un modelo L-SVM presenta una mejor calificación, pero el modelo Probit se ajusta mejor con las condiciones de uso y de políticas de la entidad a la que se hizo el proceso.

La segunda consideración respecto a los trabajos revisados indica que las variables utilizadas presentan amplias diferencias en los trabajos base, si bien muchos de ellos establecen sus criterios en variables socioeconómicas, estas están construidas con información única para cada entidad, indicando que no existe un criterio único de uso de variables para los estudios. En cada caso cada empresa tiene una base de datos única e incomparable, pero es claro que son las variables socioeconómicas las que destacan en todos los estudios. La consideración final que se encuentra es que los signos pueden variar en todos los estudios, y esto se da por el comportamiento de los clientes y como se refleja en las variables, en muchos de los casos, obedecen a la intuición financiera esperada, pero en algunos como el trabajo presentado tiene variables que contrastan con la realidad económica; esto se puede verificar fácilmente con análisis estadístico que puede sustentarse además con criterios de la entidad previamente consultada.

5. Conclusión

En este trabajo se ha evaluado el desempeño de cuatro tipos de modelos para la medición de probabilidad de impagos: Logit, Probit, Redes Neuronales y L-SVM. Como caso de estudio se consideró la información de una caja de compensación familiar en Colombia. Los resultados presentados permiten identificar que las redes neuronales son las menos eficientes para representar el problema planteado en la base de datos analizada; adicionalmente, aunque el modelo L-SVM obtuvo el mejor resultado en su estimación, sus relaciones de indicadores de Especificidad y Sensibilidad no fueron las mejores. Respecto a los modelos Logit y Probit, los resultados fueron muy similares a los del L-SVM, pero el modelo Probit evidenció una mayor fortaleza de manejo de la información que el modelo Logit, siendo más estable en identificación correcta de eventos y usando mejor las categorías que el primero. Por tanto, el modelo de riesgo de crédito que mejor puede considerarse para la base de datos analizada corresponde con el modelo Probit mejorado con procesos Stepwise para la selección de variables a utilizar.

Para trabajos futuros, se recomienda explorar este tipo de verificaciones para diferentes muestras de datos, de tal forma que los investigadores de riesgo de crédito puedan entregar afirmaciones categóricas sobre con qué tipos de modelos debe continuarse desarrollando aplicaciones en este campo. Así mismo, será conveniente explorar métodos de selección de variables explicativas que elimine el sesgo del investigador en la especificación del modelo final.

También es necesario realizar este tipo de estudios en sectores comerciales y en otras entidades del sector cooperativo o similares; si bien no es el objetivo central del estudio la realización de este estudio muestra de forma empírica que los modelos probabilísticos dependen de los datos suministrados y que son la conformación de las variables con los datos mismos los que dan la dirección de los análisis de los parámetros, pues, algunos resultados pueden ser contraintuitivos en sus variables frente a otros estudios y a la teoría económica, pero justamente es la naturaleza de la información de cada serie de clientes quienes las determinan.

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. https://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. https://doi.org/10.1016/j.ejor.2012.04.009

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.

Altman, E. (1968). The Prediction of Corporate Bankruptcy. *The Journal of Finance, XXIII*(4), 589-609. https://doi.org/https://doi.org/10.2307/2978933

Anderson, D.R., Sweeney, D.J., & Williams, T.A. (2008). *Estadística para Administración y Economía* (S. Cervantez (ed.); 10th ed.).

Anderson, R. (2007). *The Credit Scoring Toolkit*. (Oxford University Press, Ed.; 1st ed., Vol. 1). Oxford: Oxford University Press. Mexico DF: CENGAGE Learning.

Apilado, V.P., Warner, D.C., & Dauten, J.J. (1974). Evaluative Techniques In Consumer Finance - Experimental Result And Policy Implications For Financial Institutions. *Journal of Financial and Quantitave Analysis*, *9*(2), 275-284.

Chaudhuri, K., & Cherical, M.M. (2012). Credit rationing in rural credit markets of India. *Applied Economics*, *44*(7), 803-812. https://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. https://doi.org/10.1016/0377-2217(95)00246-4

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

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. https://doi.org/10.1590/S1415-65552008000400007

Hilbe, J. M. (2015). *Practical Guide to Logistic Regression*. New York, HY. Taylor and Francis Group. https://doi.org/10.1201/b18678

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression*. In Wiley (Ed.), *Wiley* (Third). Wiley. Hoboken (New Yersey): John Wiley & Sons. https://doi.org/10.2307/1270433

Ley 21, 1 (1982) (testimony of Senado de la República de Colombia). https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=4827

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. https://doi.org/10.1142/S0219622004001033

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. https://doi.org/10.1057/jors.2008.164

Melo-Velandia, L.F., & Becerra-Camargo, O.R. (2005). Medidas de riesgo, características y técnicas de medición: una aplicación del VaR y el ES a la tasa Interbancaria de Colombia. *Banco de La Republica*, 1-75.

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

Millán-Solarte, J., & Caicedo-Cerezo, É. (2018). Modelos para otorgamiento y seguimiento en la gestión de riesgo de crédito. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *1*(25), 23-41. https://www.upo.es/revistas/index.php/RevMetCuant/article/view/2370/2709

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. http://search.proquest.com/docview/818448211?accountid=10344

Myers, J.H., & Forgy, E.W. (1963). Comparison of Discriminant and Regression analysis for credit evaluation System. *Journal of the American Statistical Association*, *58*(303), 799-806.

Ochoa, J.C., Galeano, W., & Agudelo, L.G. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Coyuntura Económica*, *16*, 191-222.

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

Orgler, Y.E. (1970). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking*, *2*(4), 435-445. https://doi.org/10.2307/1991095

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., Rubio, J.L., & Blasco, D.C. (2010). A Credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance & Administrative Science*, *15*(28), 89-124.

Rodríguez, D.E., Becerra, J.A., & Cardona, D. (2017). Modelos y metodologías de credit score para personas naturales: una revisión literaria. *Revista CEA*, *3*(5), 13-28. https://doi.org/10.22430/24223182.645

Rodríguez, D., & Trespalacios, A. (2015). *Medición de valor en riesgo en cartera de clientes a través de modelos logísticos y simulación de Montecarlo*. Medellín (Colombia): Universidad EAFIT. https://repository.eafit.edu.co/handle/10784/7853.

Roszbach, K. (2004). Bank Lending Policy, Credit Scoring, and the Survival of Loans. *Review of Economics and Statistics*, *86*(4), 946-958. https://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. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-35922010000100013&lang=pt

Salazar, F.E. (2013). Cuantificación del riesgo de incumplimiento en créditos de libre inversión: un ejercicio econométrico para una entidad bancaria del municipio de Popayán, Colombia. *Estudios Gerenciales*, *29*(129), 416-427. https://doi.org/10.1016/j.estger.2013.11.007

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

Támara, A., Villegas, G., Leones, M., & Salazar, J. (2018). Modelación del riesgo de insolvencia en empresas del sector salud empleando modelos logit. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *1*(26), 128-145. https://www.upo.es/revistas/index.php/RevMetCuant/article/view/2757/3039

Taylor, K.W., & Chappell, N.L. (1980). Multivariate analysis of qualitative data. *Canadian Review of Sociology/Revue Canadienne de Sociologie*, *17*(2), 93-108. https://doi.org/10.1111/j.1755-618X.1980.tb00688.x

Thomas, L., Edelman, D., & Crook, J. (2002). *Credit Scoring and Its Applications*. Philadelphia: SIAM (Society For Industrial and Applied Mathematics).

Trejo, J.C., Martínez, M.Á., & Venegas, F. (2017). Credit risk management at retail in Mexico: An econometric improvement in the selection of variables and changes in their characteristics. *Contaduría y Administración*, *62*(2), 399-418. https://doi.org/10.1016/j.cya.2017.02.006

Webster, G. (2011). *Bayesian Logistic Regression Models for Credit Scoring* (Issue December). Grahamstown, South Africa: Rhodes University.

Zhai, H., & Russell, J.S. (1999). Stochastic modelling and prediction of contractor default risk. *Construction Management and Economics*, *17*(1), 563-576.

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. https://doi.org/10.1142/S0219622009003582