Universidad Andina Simón Bolívar

Sede Ecuador

Área de Gestión

Maestría en Gestión Financiera y Administración de Riesgos Financieros

Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito

Cristhian Oswaldo Montalván Acaro

Tutor: Felipe Alexander Andrade Cóndor

Quito, 2019

Trabajo almacenado en el Repositorio Institucional UASB-DIGITAL con licencia Creative Commons 4.0 Internacional

Reconocimiento de créditos de la obra
No comercial
Sin obras derivadas

creative commons

Para usar esta obra, deben respetarse los términos de esta licencia

Cláusula de cesión de derecho de publicación de tesis

Yo, Cristhian Oswaldo Montalván Acaro, autor de la tesis intitulada ***Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito*** mediante el presente documento dejo constancia de que la obra es de mi exclusiva autoría y producción, que la he elaborado para cumplir con uno de los requisitos previos para la obtención del título de Máster en Gestión Financiera y Administración de Riesgos Financieros en la Universidad Andina Simón Bolívar, Sede Ecuador.

1. Cedo a la Universidad Andina Simón Bolívar, Sede Ecuador, los derechos exclusivos de reproducción, comunicación pública, distribución y divulgación, durante 36 meses a partir de mi graduación, pudiendo por lo tanto la Universidad, utilizar y usar esta obra por cualquier medio conocido o por conocer, siempre y cuando no se lo haga para obtener beneficio económico. Esta autorización incluye la reproducción total o parcial en los formatos virtual, electrónico, digital, óptico, como usos en red local y en internet.

2. Declaro que en caso de presentarse cualquier reclamación de parte de terceros respecto de los derechos de autor/a de la obra antes referida, yo asumiré toda responsabilidad frente a terceros y a la Universidad.

3. En esta fecha entrego a la Secretaría General, el ejemplar respectivo y sus anexos en formato impreso y digital o electrónico.

Fecha: …….

Firma: …….

Resumen

Esta investigación pretende contrastar la hipótesis de sí el uso de redes neuronales, para la modelización del *credit scoring* de una cartera de microcrédito, logra un mejor performance que utilizar una metodología de regresión logística.

Para esto se hace uso de la información de una cartera del producto de microcrédito, proporcionada por una institución financiera ecuatoriana de mediano tamaño, desembolsada entre los periodos de enero 2014 a diciembre 2017. La información corresponde al momento del desembolso de los créditos, por lo cual se construye modelos scoring de originación.

Los dos modelos por desarrollar son construidos con el programa estadístico R y una vez se cuenta con estos, se procede a comparar su poder predictivo, capacidad para diferenciar entre buenos y malos clientes, adaptación a los datos, etc.; a partir de distintos estadísticos como KS, coeficiente de Gini, matriz de confusión, AUROC, criterio de información de Akaike.

Además, se aplica los modelos en un periodo de información distinto al que se usó en su construcción y de esta forma establecer si son generalizables, al ser funcionales con información distinta al de su desarrollo.

El modelo de redes neuronales se ajusta de mejor forma a los datos, pues se obtiene un criterio de información Akaike menor al de regresión logística, con una diferencia de 8.029,2 puntos. De igual forma, los estadísticos KS, coeficiente de Gini y curva ROC evidencia que hacer uso de las redes neuronales logra una mejor clasificación de los clientes, con 5,19, 5,84 y 2,92 puntos porcentuales por encima de los estadísticos del modelo de regresión logística, respectivamente. Finalmente, la matriz de confusión muestra un menor error con el modelo de redes neuronales, al compararlos con un mismo punto de corte óptimo.

Los resultados obtenidos evidencian que la metodología de redes neuronales proporciona un modelo scoring más robusto que al usar una regresión logística, pudiendo corroborar que la hipótesis planteada es verdadera, bajo el proceso de modelización empleado.

Palabras clave: regresión logística, redes neuronales, score, originación, modelos, proceso de modelización

Dedicado a todos aquellos que me quieren y quiero.

Sí, este trabajo está dedicado a ti.

Agradecimientos

Han sido muchas personas que, a lo largo de este sueño, me han apoyado e incentivado a cumplirlo; a todas aquellas personas expreso mi agradecimiento:

- A Dios, que ha guiado mis pasos en este camino de la vida.
- A mi padre y madre, que, sin su esfuerzo, constancia y amor, no sería la persona que me he esforzado en ser.
- A mis hermanos Oscar y Diana, que sin sus locuras y constante presión me hubiera vuelto loco y no de la forma agradable, los adoro.
- A mi mejor amiga de la vida, Jessica, que siempre me incentiva a ser mejor, y seguir adelante.
- Priscila, Samanta y Santiago, personas en las que pude confiar y puedo decir, sin miedo, que aprecio.
- A personas que, a pesar de la distancia se hicieron presentes, los quiero chicos, Alex y Eli.
- Y por qué no, algunas menciones especiales, Daniel, Cecilia, Gonzalo y Paúl, espero se pregunten por que los incluí, pero tengo mis razones.

Tabla de contenidos

Capítulo primero: Planteamiento del problema 17

1. Descripción de la realidad problemática 17
2. Hipótesis 18
3. Objetivo general y específicos 18
4. Justificación 19
5. Delimitación del problema 20

Capítulo segundo: Marco teórico 21

1. Naturaleza de los modelos 24
2. Modelo de regresión logística - logit 24
   - 2.1 Estimación de los parámetros modelo logit 25
   - 2.2 Interpretación coeficientes de una regresión logística. 27
3. Modelo de Redes Neuronales 28
   - 3.1 Modelo biológico 28
   - 3.2 Elementos de una red neuronal artificial 30
   - 3.3 Arquitectura de las redes neuronales 33
   - 3.4 Modos de operación de una red neuronal 34
   - 3.5 Clasificación de los modelos neuronales 36
   - 3.6 Algoritmo Backpropagation 37
   - 3.7 Algoritmo RPROP+ 38
4. Conceptos del proceso de modelización 39
   - 4.1 Prueba Dickey – Fuller aumentada 39
   - 4.2 Muestreo aleatorio simple 39
   - 4.3 Roll Rate 39
   - 4.4 Valor de Información (IV) 40
   - 4.5 Backward Stepwise y criterio de información de Akaike (AIC) 40
   - 4.6 Estadístico de Wald 41
   - 4.7 GVIF: Factor de inflación de la varianza generalizado 42
   - 4.8 Estadístico de Kolmogorov – Smirnov (KS) 42

4.9 Coeficiente de GINI 44
4.10 Matriz de confusión: error, sensibilidad y especificidad 44
4.11 Estadístico AUROC 46

Capítulo tercero: Proceso de modelación 47

1. Selección y consistencia de la muestra 49
2. Determinación del periodo de modelización 50
3. Definición de buen y mal cliente 54
4. Análisis de las variables independientes 58
4.1 Análisis exploratorio de las variables 59
4.2 Categorización de las variables 59
4.3 Elección de variables predictoras 60

Capítulo cuarto: Construcción y contraste de modelos 63

1. Modelo de Regresión Logística 63
1.1 Significancia de las variables independientes 66
1.2 Consistencia de signos y coeficientes de las variables del modelo 66
1.3 Resultados del factor de inflación generalizado de la varianza 70
1.4 Estadísticos del modelo de regresión logística 70
2. Modelo de Redes Neuronales 71
2.1 Estadísticos del modelo de redes neuronales 75
3. Comparación Modelos Scoring 76
3.1 Poder predictivo en el periodo de Testing 80

Capítulo quinto: Conclusiones y recomendaciones 83

1. Conclusiones 83
2. Recomendaciones 85

Bibliografía 87

Anexos 91

Anexo 1: Descripción del total de variables 91

Anexo 2: Categorización de variables independientes – árboles 94

Anexo 3: Código R 106

Lista de tablas

Tabla 1 Matriz de Confusión 45
Tabla 2 Indicadores de eficiencia 45
Tabla 3 Tamaño muestra periodo de modelo 54
Tabla 4 Roll Rate – Definición de Cliente bueno, malo e indeterminado 55
Tabla 5 Definición del tipo de cliente 57
Tabla 6 IV – Variable Num_Inst_Adeuda 61
Tabla 7 Modelo Scoring Originación – Metodología logit 64
Tabla 8 Estadísticos del modelo logit 70
Tabla 9 Coeficientes – Capa 1 73
Tabla 10 Coeficientes – Capa 2 74
Tabla 11 Coeficientes – Capa 3 75
Tabla 12 Coeficientes – Capa 4 75
Tabla 13 Estadísticos del modelo redes neuronales 75
Tabla 14 Comparación estadísticos modelos – Periodo modelo 77
Tabla 15 Matrices de confusión: Punto de corte 0,8680071 78
Tabla 16 Matrices de confusión: Punto de corte 0,890068781 78
Tabla 17 Periodo Testing 80
Tabla 18 Estadísticos - Periodo Testing 80

Lista de Figuras

Figura 1 Función Logística 25
Figura 2 Modelo de una neurona biológica 29
Figura 3 Modelo de neurona artificial inspirada en una neurona biológica 30
Figura 4 Funciones de activación usuales 32
Figura 5 Estructura de una red neuronal artificial multicapa 34
Figura 6 Clasificación de las redes neuronales 36
Figura 7 Regla valor de la información IV 40
Figura 8 Estándar estadístico KS 43
Figura 9 Esquema de modelización de un *scoring* 48
Figura 10 Definición del performance 49
Figura 11 Análisis de cosechas 51
Figura 12 Prueba ADF – Raíz Unitaria 52
Figura 13 Roll Rate: Atraso mayor a 90 días luego de 12 meses del desembolso 56
Figura 14 Árbol de decisión para la variable Num_Inst_Adeuda 60
Figura 15 Curva ROC del modelo logit 71
Figura 16 Arquitectura del modelo de redes neuronales 72
Figura 17 Red Neuronal 73
Figura 18 Curva ROC del modelo de redes neuronales 76
Figura 19 Comparación curvas ROC 79

Capítulo primero: Planteamiento del problema

1. Descripción de la realidad problemática

Evaluar el riesgo de crédito es de gran importancia para una institución financiera, pues busca evitar pérdidas que pueden estar relacionadas a cualquier tipo de decisión de concesión de crédito inapropiada. Además, la competencia empresarial para obtener más ganancias y participación en el mercado se vuelve cada vez más agresiva con el paso de los años, por lo que algunas instituciones asumen un mayor riesgo para lograr una ventaja competitiva en el mercado, provocando que muchas sufran grandes pérdidas debido a un aumento en el incumplimiento y préstamos incobrables (Yu 2008, 17).

En el proceso de otorgamiento de créditos, una institución no puede tomar decisiones a partir de su juicio experto para cada una de las solicitudes recibidas, pues con el aumento en el número de solicitantes y la competencia intensa en la industria crediticia, este método no puede satisfacer las demandas en los aspectos económicos y de eficiencia (Yu 2008, 17); sino que intentará adoptar sistemas de calificación de créditos para facilitar y acelerar los procesos en la toma de decisiones. Por tal motivo, nace el concepto de modelos de *credit scoring* o modelos de calificación de créditos (Islam, Zhou, y Li 2009, 7).

Muchos algoritmos son usados para la construcción de un *credit scoring*, sin embargo, cada vez se deberá buscar alternativas más efectivas para tomar decisiones más precisas, por ejemplo, las redes neuronales (Islam, Zhou, y Li 2009, 7).

En el Ecuador, la cartera productiva, en la cual se considera la cartera comercial y microcrédito, han tenido un incremento del 30,2% entre los periodos marzo 2015 y marzo 2019, pasando de 11.285,4 millones de dólares a 14.698,8 millones de dólares, respectivamente (Arias 2016, 3). De igual forma, la cartera vencida más cartera que no devenga intereses para estos productos, que la Superintendencia de Bancos del Ecuador llama cartera problemática, tuvo un incremento del 17,9% pasando de 192,32 millones dólares a 226,76 millones de dólares, en los mismos periodos. Más preocupante son las cifras de la cartera con calificación C, D y E pues pasan de 981,5 millones de dólares en marzo del 2015 a 14.698,8 millones de dólares en marzo del 2019, teniendo un incremento del 235,2%, representando el 22,4% del total de cartera productiva a marzo del 2019 (Paredes y Cóndor 2018, 2).

Varios factores pueden incidir en el crecimiento de la cartera vencida o el incremento de la mora de una cartera, como malas prácticas en la concesión de créditos, metas de créditos agresivas por parte de las instituciones, deterioro del empleo, recesión económica, etc. Sin embargo, el anterior escenario evidencia la necesidad de contar con nuevas herramientas para la gestión de riesgo de crédito que ayuden a minimizar la probabilidad de pérdida de una institución, buscando alcanzar la eficiencia de la gestión de riesgos a partir de mejores herramientas estadísticas e informáticas.

El presente trabajo será desarrollado a partir de información proporcionada por una institución financiera ecuatoriana de la cual por razones de privacidad y seguridad no se revelará el nombre, pero cuando sea necesario, será llamada Entidad Financiera X.

La Entidad Financiera X, evidenció un incremento del 1.6% en su saldo en mora mayor a 90 días y un crecimiento del 4.4% en el saldo en mora mayor a 180 días, para su cartera de microcrédito, comparando los meses junio 2017 y junio 2018. Frente a esta situación, la institución se ha visto en la necesidad de contar con herramientas que apoyen de forma eficaz y eficiente los procesos de concesión de crédito, y buscar contrastar metodologías distintas a las que cuenta al momento, pues únicamente, hace uso de metodologías de regresión logística.

2. Hipótesis

El hacer uso de redes neuronales, para la modelización del *credit scoring* de una cartera de microcrédito, logra un mejor performance que utilizar una metodología de regresión logística.

3. Objetivo general y específicos

Objetivo general:

Comparar la metodología de *credit scoring* de regresión logística y redes neuronales en una cartera de microcrédito, de una institución financiera ecuatoriana.

Objetivos específicos:

- Conocer la teoría que respalda los métodos de regresión logística y de redes neuronales, para la obtención de un modelo *credit scoring*.

- Detallar el proceso de modelización al usar metodologías de regresión logística y de redes neuronales.
- Identificar el modelo de *credit scoring*, regresión logística o redes neuronales, que tiene un mejor desempeño.

4. Justificación

Para las instituciones financieras es de gran importancia conocer con anticipación la probabilidad de que un solicitante de crédito vaya a cumplir con su obligación, y poder tomar decisiones más objetivas y eficientes sobre el otorgamiento de los préstamos y sobre los términos que establecerán para sus diferentes productos.

La necesidad de modelos cuantitativos confiables que discriminen a los buenos y malos clientes con precisión es crucial para el negocio de muchas instituciones de crédito, pues podrán tomar medidas preventivas o correctivas en el proceso de concesión y seguimiento de créditos. Es evidente que la modelización y análisis de riesgo se vuelven muy importantes para la sostenibilidad y beneficio de las empresas convirtiendo el análisis y modelización del riesgo de crédito en un tema valioso tanto para la comunidad académica y de la industria (Yu 2008, 5).

Para cumplir este objetivo, se cuenta con los modelos matemáticos, *credit scoring* (TransUnion 2007, 4). Los modelos *credit scoring* son algoritmos o métodos que pueden ayudar a obtener la probabilidad de incumplimiento de un solicitante de crédito, permitiendo evaluar el riesgo en el origen de la financiación (Gutierrez Girault 2007, 2).

Existe una gran variedad de metodologías disponibles para evaluar el riesgo de crédito, los modelos de regresión logística se encuentran en el grupo de los más usados en la industria por ser intuitivos, de buen poder predictivo y de fácil implementación; sin embargo, existen otros métodos menos explorados, con un mayor nivel de complejidad que podrían mejorar el poder predictivo, entre estos se encuentra los métodos de redes neuronales (Bonilla, Olmeda, y Puertas 2003, 864).

Esta investigación pretende contrastar las historias de superioridad de la metodología de redes neuronales sobre los métodos tradicionales como es la regresión logística, con información de una institución financiera ecuatoriana, cotejando su poder predictivo y estableciendo diferencias entre cada uno. Esto será muy útil para las empresas financieras pues se impulsa el uso de algoritmos avanzados que pueden mejorar la identificación de buenos y malos clientes.

Además, de comparar los resultados de los dos modelos a desarrollar, se busca detallar cada uno de los pasos del proceso de modelización, identificando las diferencias por hacer uso de dos metodologías distintas, sirviendo de guía para las instituciones financieras que tratan con microcréditos, y aquellas que busquen desarrollar modelos de *credit scoring*.

5. Delimitación del problema

En el presente proyecto se generan dos modelos *scoring* para la cartera de microcrédito de una institución financiera ecuatoriana haciendo uso de dos metodologías distintas, regresión logística y redes neuronales. Los dos modelos se enfocarán en el proceso de otorgamiento de créditos, por lo cual se los categoriza como modelos de originación. La Entidad Financiera X que facilita la información para el desarrollo del presente trabajo cuenta con más de 1.000 colaboradores en todo el país a diciembre del 2017 y un valor en activo de más de 600 millones de dólares.

Capítulo segundo: Marco teórico

En el presente capítulo se expone las metodologías de regresión logística y redes neuronales, con el fin de conocer la estructura tras cada modelo y los distintos conceptos que serán empleados en el proceso de modelización y comparación de los scores resultantes.

Una de las principales actividades de una institución financiera es la de intermediación financiera, permitiendo transferir los recursos de aquellos individuos con exceso de estos, a aquellos que lo necesitan, para consumirlo o invertirlo. Cumplir con esta actividad, conduce a la institución a exponerse a un conjunto de riesgos, que conlleva a la presencia de pérdidas.

Por lo cual, es de suma importancia, para la institución, contar con herramientas eficientes para la identificación, medición, monitoreo, control, mitigación y divulgación de los riesgos, a la cual está sujeta; es decir, contar con metodologías y políticas que lleven a una correcta gestión de los riesgos (Bolivia et al. 2008, 40).

Así nacen las metodologías conocidas como *credit scoring*, que cuentan con algunas definiciones, dependiendo de la perspectiva del autor. D. J. Hand y Henley (1997, 523) lo definen como un método estadístico formal, usado para clasificar a solicitantes de crédito en grupos de buenos y malos; menciona que estos métodos han llegado a ser muy importantes, con un dramático crecimiento a partir de los años 70, pues en aquellos años, los métodos tradicionales para decidir si otorgar o no un crédito, para un individuo en particular, era el uso del juicio humano, basado en la experiencia de decisiones tomadas anteriormente.

Sin embargo, con el incremento de la demanda de crédito, alineado con una competencia comercial más grande y una mejora de los recursos estadísticos y tecnológicos, la industria financiera se vio en la necesidad de hacer un más eficaz y eficiente proceso de otorgamiento de créditos, así como de la evaluación de su portafolio.

Es así, que la utilización de modelos *credit scoring*, para la evaluación del riesgo de crédito, comienzan a ser usados en los años 70’s pero se generalizan a partir de los 90’s (Gutierrez Girault 2007, 2).

Pese a todo, tal como lo menciona Gutierrez Girault (2007, 2), a pesar de la diversidad de modelos scoring, el juicio humano o del analista, no ha quedado rezagado, continúa siendo utilizado, afirmando que en la práctica ambos métodos de evaluación

coexisten y complementan; siendo los basados en la experiencia denominados como expertos y aquellos que se basan en un análisis estadístico como estadísticos.

Autores como Thomas, Edelman, y Crook (2002, 16) definen a un *credit scoring* como un conjunto de modelos de decisión y sus técnicas implícitas que ayudan a los prestamistas en la concesión de un crédito. Shanmugapriya (2012, 34) los considera como un proceso en el cual se analiza el comportamiento pasado de los clientes para diferenciar clientes en bancarrota y no bancarrota.

Existen algunas definiciones para un *credit scoring*; sin embargo, en este trabajo, se considera a un *credit scoring* como un conjunto de metodologías que permitirán evidenciar la probabilidad que un cliente cumpla o no con sus obligaciones, a partir de la información que se conozca del mismo.

Los modelos credit scoring tienen muchas funciones dentro del ciclo de vida de un crédito que comprende el otorgamiento, seguimiento, cobranza y recuperación (Bolivia et al. 2008, cap. 2); sin embargo, este proyecto se enfocará en las metodologías que sirven de apoyo, en la toma de decisiones del proceso de otorgamiento de créditos; tales metodologías deben permitir medir los riesgos de los potenciales clientes, a fin de precisar quiénes podrían ser sujetos de crédito, estableciendo variables significativas que ayuden a identificar los individuos cuyo riesgo se ajuste al perfil de riesgo de la institución.

En la literatura, se cuenta con diversas metodologías para modelar un *credit scoring*, no obstante, se abordarán los modelos de regresión logística y de redes neuronales; los primeros, considerados como tradicionales y comúnmente usados por las instituciones por su poder de predicción y facilidad de interpretación y los segundos, métodos menos explorados, que podrían mejorar el poder predictivo obtenido por los modelos tradicionales (Bonilla, Olmeda, y Puertas 2003, 864).

En general cada método tiene ventajas y desventajas para la construcción de un modelo de riesgo de crédito. Yu (2008, 17) expresa que modelos como las redes neuronales tienen como ventaja el construir un modelo sin la necesidad que el usuario tenga una comprensión profunda del problema, pudiendo utilizarlos independientemente del cumplimiento de supuestos teóricos relativos a técnicas estadísticas o paramétricas, además de funcionar mejor desde el punto de vista de la precisión en la clasificación. Mientras que métodos más tradicionales, como la regresión logística, necesitan construir el modelo en base al cumplimiento de algunos supuestos. Las ventajas de los métodos tradicionales es que son relativamente simples, fáciles de interpretar e implementar, que

es algo que no se tiene en un modelo de redes neuronales pues su compleja estructura no permite lograr resultados óptimos. Por lo cual, es difícil para alguien determinar cuál es "el mejor" modelo en general, para todas las posibles situaciones.

Trabajos de Pitarque, Ruiz, y Roy (2000, 459) buscan apoyar, con sus resultados, la idea que las redes neuronales han mostrado una capacidad clasificatoria igual o superior que las técnicas estadísticas. Trabajan con una serie de simulaciones y aplicaciones sobre datos reales, comparando las redes neuronales con la regresión logística y el análisis discriminante. Si bien sus resultados muestran que las redes hacen un mejor trabajo en la clasificación de la variable objetivo, plantean la pregunta si vale la pena ganar en capacidad de clasificación a costa del costo computacional y de recursos involucrados en el entrenamiento de las redes neuronales, concluyendo que ambas técnicas deberían complementarse. Islam, Zhou, y Li (2009, 29) llega a conclusiones similares.

En cambio, Croall y Mason (1992, 74), Ladino (2014, 27), Larasati, DeYong, y Slevitch (2011, 304) no muestran inclinación alguna hacia una u otra metodología, estableciendo que no se puede demostrar la superioridad de metodologías como las redes neuronales, sobre las tradicionales al tener cada una sus ventajas en distintos escenarios.

Bonilla, Olmeda, y Puertas (2003, 859) evalúan la capacidad predictiva de dos modelos paramétricos (análisis discriminante y regresión logística) y cinco no paramétricos, entre los que se encuentran las redes neuronales, sobre una base de 690 observaciones de solicitantes de tarjetas de crédito. Concluye que las redes neuronales resultan ser el modelo de mayor capacidad predictiva, superando a todos los demás. Sin embargo, la escasez de datos dificulta la adecuada comparación de los modelos. Compara los mismos a partir del error de predicción, obteniendo un error del 12,50% y 12,33% en el periodo de entrenamiento y el 12,22% y 10% en el periodo de testing para los modelos logit y redes neuronales, respectivamente.

En cambio Trujillano et al. (2003, 504) compara la capacidad de predicción de un modelo de regresión logística y redes neuronales a partir de la curva ROC, evidenciando mejores valores con las redes neuronales, que se mantienen tanto en la muestra de desarrollo como testing.

Zhang (2004, 272) expone que uno de los métodos comúnmente usados para evaluar el poder de clasificación de los modelos es el estadístico de Kolmogorov – Smirnov (K-S). El mismo compara un modelo de regresión logística y redes neuronales a partir de los estadísticos K-S, curva ROC y matriz de confusión. El estadístico K-S fue superior con 14 puntos porcentuales en las redes neuronales, mientras que para la curva

ROC y tasa de clasificación global fue superior el modelo de regresión logística con 4,41 y 10,57 puntos porcentuales, respectivamente. Zhang concluye que ningún método de evaluación representa una panacea para los investigadores o analistas, sino mas bien es necesaria una comprensión de los datos y el problema para la correcta elección de no solo la metodología de modelación sino también del método de evaluación (Zhang 2004, 289).

En el presente trabajo se busca contrastar, si el hacer uso de redes neuronales, para la modelización del *credit scoring*, logra un mejor performance que utilizar una metodología de regresión logística, para una cartera de microcrédito de una institución financiera ecuatoriana, comparando estadísticamente cada metodología.

1. Naturaleza de los modelos

El objetivo principal de los dos modelos es estimar la probabilidad de incumplimiento de un solicitante de crédito; para esto es necesario contar con una variable que identifique si el solicitante es un buen o mal cliente, esta variable será representada con la letra **Y**, que es la variable dependiente del modelo.

La variable dependiente es una variable dicotómica (binaria) que toma los siguientes valores:

$$Y = \begin{cases} 1: Si\ el\ solicitante\ de\ crédito\ es\ definido\ como\ un\ buen\ cliente \\ 0: Si\ el\ solicitante\ de\ crédito\ es\ definido\ como\ un\ mal\ cliente \end{cases} \quad (1)$$

La definición de cliente bueno (1) y malo (0), para la variable dependiente Y, es construida a partir de su información demográfica, comportamiento en el buró y/o dentro de la institución.

Los modelos buscan estimar la probabilidad que la variable dependiente Y tome el valor de 0 o 1, a partir de un conjunto de variables denominadas independientes, las cuales serán representadas con la letra X, que pueden ser cualitativas o cuantitativas. Las variables independientes son obtenidas a partir de diversas fuentes de información tanto internas y externas del individuo, como puede ser su información crediticia, de buró, demográfica, etc. Muchas de las cuales dependerán de las características del crédito que se esté considerando.

2. Modelo de regresión logística - logit

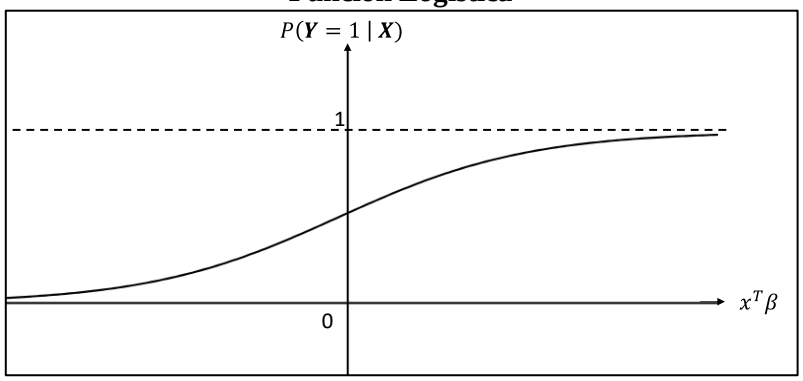
Los modelos logit pertenecen al grupo de modelos de regresión con respuesta cualitativa, en este caso binaria; mientras que las variables independientes pueden ser cualitativas o cuantitativas, o una mezcla de ambas (Flórez y Rincón 2002, 124).

El modelo está basado en una función de distribución logística, cuya estructura se presenta a continuación:

$$P(Y = 1|\, X) = F(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z})}{1 + \exp(\boldsymbol{z})}, \qquad -\infty < z < \infty \quad (2)$$

Con $z = X^T\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$.

Figura 1
Función Logística



Fuente: Flórez y Rincón (2002, 128)
Elaboración: Propia

Donde:

- Y**:** Es la variable dependiente, binaria, que puede tomar dos posibles valores, que se etiquetará con 0 (cliente malo) y 1 (cliente bueno).
- X**:** Es el conjunto de n variables independientes $(x_1, x_2, \dots, x_n)$ relacionadas con la información propia del solicitante, tomadas con el fin de explicar y/o predecir el valor de Y.
- $F(z)$: Es la función de probabilidad, que depende de un vector de parámetros $\beta = (\beta_0, \beta_1, \dots, \beta_n)$, que permitirán relacionar las variables independientes X, con la dependiente Y. Esta función tiene un rango entre [0, 1] y se conoce como función de distribución logística.

El objetivo del modelo es hallar los coeficientes β que mejor se ajusten a la expresión (2).

2.1 Estimación de los parámetros modelo logit

La estimación de los coeficientes β puede realizarse a partir del método de máxima verosimilitud (Gujarati, Guerrero, y Medina 2005, 574).

Supóngase se cuenta con un conjunto de k individuos; de tal forma que, catalogarles como buenos o malos clientes será definido por la variable Y_i, considerando $i = 1, 2, \dots, k$.

En vista que cada Y_i es una variable aleatoria de Bernoulli, por tomar dos valores, 0 o 1, podemos expresar la probabilidad que suceda uno u otro evento, como sigue:

$$P(Y_i = 1) = P_i$$
$$P(Y_i = 0) = 1 - P_i$$

Su función de probabilidad será:

$$f_i(Y_i) = P_i^{Y_i} \times (1 - P_i)^{1-Y_i}, \qquad con\ i = 1, 2, 3, \dots, k \quad (3)$$

Es decir, la función $f_i(Y_i)$ denota la probabilidad que $Y_i = 1\ o\ 0$.

Como cada observación es independiente, la probabilidad conjunta de observar los k valores de la variable Y, se expresa como:

$$f(Y_1, Y_2, \dots, Y_k) = \prod_{i=1}^{k} f_i(Y_i) = \prod_{i=1}^{k} P_i^{Y_i} \times (1 - P_i)^{1-Y_i} \quad (4)$$

Donde $\prod$ es el operador producto. A esta probabilidad conjunta se la conoce como función de verosimilitud (FLV). Al tomar el logaritmo de esta función se tiene:

$$\begin{aligned} \ln\left(f(Y_1, Y_2, \dots, Y_k)\right) &= \sum_{i=1}^{k} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)] \\ &= \sum_{i=1}^{k} [Y_i \ln P_i - Y_i \ln(1 - P_i) + \ln(1 - P_i)] \\ &= \sum_{i=1}^{k} \left[Y_i \ln\left(\frac{P_i}{1 - P_i}\right)\right] + \sum_{i=1}^{k} \ln(1 - P_i) \quad (5) \end{aligned}$$

Tal como se expuso en (2), la probabilidad que un individuo sea un buen o mal cliente es representado por:

$$P_i = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)} \quad (6)$$

De aquí se puede, fácilmente, demostrar que:

$$1 - P_i = \frac{1}{1 + \exp(X_i^T \beta)} \quad (7)$$

De igual forma,

$$\ln\left(\frac{P_i}{1 - P_i}\right) = X_i^T \beta \quad (8)$$

Considerando (7) y (8) en (5), se puede expresar el logaritmo de la función de verosimilitud, como sigue:

$$\ln\left(f(Y_1, Y_2, \dots, Y_k)\right) = \sum_{i=1}^{k} Y_i(X_i^T\beta) - \sum_{i=1}^{k} \ln(1 + \exp(X_i^T\beta)) \quad (9)$$

Podemos observar que (9) es una función que depende de los coeficientes β, pues Y_i y X_i se conocen.

El método de máxima verosimilitud consiste en maximizar la expresión (9), para buscar la máxima capacidad predictiva. Para esto se deriva parcialmente, respecto a cada una de las incógnitas; es decir, respecto a cada β_j, con $j = 1, \dots, n$. Obteniendo un sistema de *n* ecuaciones no lineales, que deberán resolverse por procedimientos numéricos.

Una vez obtenidos los valores β se verifica que en verdad maximicen la función de verosimilitud a partir de la condición de maximización de segundo orden. Luego de este proceso, se obtiene los coeficientes, necesarios para estimar la probabilidad de incumplimiento de un individuo, a partir de la ecuación (2).

2.2 Interpretación coeficientes de una regresión logística.

Una de las razones, por las cuales se utiliza con mayor frecuencia un modelo de regresión logística es que su interpretación es relativamente sencilla. Para apreciar este beneficio, es de ayuda entender el significado de *odds.* Tal como lo expresa (Allison 2012, 15) muchas personas consideran a una probabilidad como la forma "natural" de cuantificar que un evento ocurra, considerando valores que se mueven entre 0 y 1. Sin embargo, existen otras formas de representar un cambio natural en algún evento, esto son los *odds ratios*.

El mismo autor, define los *odds*, como la relación entre el número esperado de veces que un evento ocurra y el número esperado de veces que este no ocurra. De esta forma, la relación entre el *odds* y la probabilidad es:

$$Odds = \frac{Probabilidad\ que\ un\ evento\ ocurra}{1 - Probabilidad\ que\ un\ evento\ ocurra}$$

Esta expresión tiene relevancia en un modelo de regresión logística, pues si se considera (6) y (7) se tiene,

$$\frac{P_i}{1 - P_i} = \frac{\frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}}{\frac{1}{1+\exp(X_i^T\beta)}} = \exp(X_i^T\beta) \quad (10)$$

A esta expresión se la considera como transformación logit de la probabilidad P_i, cuya parte izquierda es una razón de probabilidades u *odds* (Flórez y Rincón 2002, 128). Al considerar el logaritmo natural de (10) se obtiene el logaritmo de la razón de

probabilidades conocido como logit y es por este término que al modelo de regresión logística se lo conoce, también, como modelo logit. Así, se llega a la ecuación (8).

$$\mathrm{L} = \ln\left(\frac{P_i}{1-P_i}\right) = X_i^T\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (11)$$

De esta forma, la interpretación del modelo está dada por esta expresión logit (L); por ejemplo, β_2 mide el cambio en L ocasionado por un cambio unitario en x_2, suponiendo constantes el resto de variables explicativas (Gujarati, Guerrero, y Medina 2005, 575).

La interpretación del modelo también puede darse a partir del *odds ratio*, la cual es una medida de la magnitud de asociación entre dos variables; en este caso, cada una de las independientes con la dependiente. Un *odds ratio* mayor a 1, muestra que existe una relación positiva o directa entre las dos variables, mientras que un *odds ratio* menor a 1, establece una relación negativa o inversa. Cuando el *odds ratio* es igual a 1, significa que no existe una relación entre las mismas (Salas 1996, 11).

El *odds ratio* puede calcularse a partir de la estimación de los parámetros del modelo:

$$odds\ ratio = e^{\beta} \quad (12)$$

3. Modelo de Redes Neuronales

Las redes neuronales artificiales (RNA) son modelos matemáticos-computacionales que intentan imitar el funcionamiento del cerebro y la forma como este procesa la información. Se cataloga dentro de las técnicas no paramétricas de *credit scoring*, como sistemas con la capacidad de aprender a través de entrenamiento, también conocido como la interpretación que ellas hacen de la información que reciben (Bahamón 2013, 37).

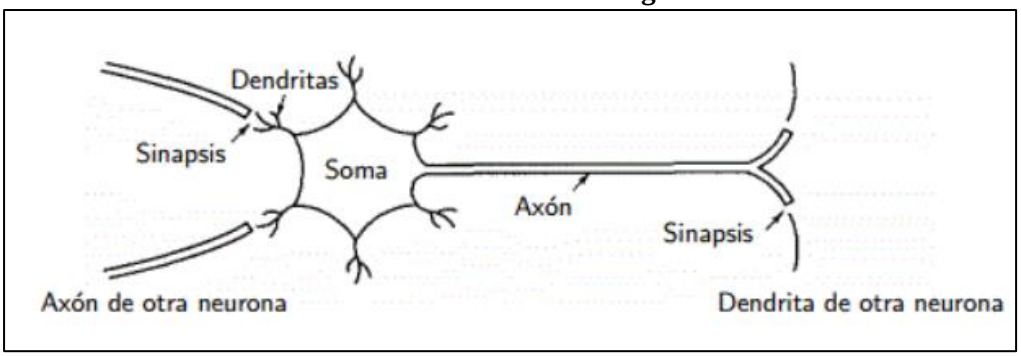
3.1 Modelo biológico

La figura 2 muestra un tipo común de neurona biológica, que está compuesta principalmente por:

- Un cuerpo central, que contiene el núcleo celular, denominado Soma.
- La conexión entre neuronas se establece a partir de una prolongación del Soma, llamada Axón, que también se ramifica en su extremo final para establecer conexión con otras neuronas, estas ramificaciones son conocidas como terminales axónicos.

- Las dendritas que son ramificaciones del cuerpo central, con las cuales se logra la conexión sináptica.

Figura 2
Modelo de una neurona biológica



Fuente y elaboración: Tablada y Torres (2009, 23)

Se estima que alrededor de cien mil millones de neuronas son las que conforman el sistema nervioso; estas se diferencian del resto de células vivas en el hecho que poseen capacidad de comunicarse.

En general, las dendritas y el Soma reciben las señales de entrada, el cuerpo celular las combina y emite señales de salida; a continuación, el Axón transmite esta señal a sus terminales, que se encargan de distribuir la información a otro conjunto de neuronas (Brío y Molina 2002, 5).

Un aspecto muy importante en el proceso de comunicación entre neuronas es el término conocido como sinapsis. Brío y Molina (2002, 6) lo definen como la unión entre dos neuronas, en el proceso de generación y transmisión de la señal nerviosa.

En cambio González y Hernando (2000, 46) lo precisa como contactos especiales, a través de los cuales fluyen neurotransmisores que llevan las señales que se transmiten entre los terminales axónicos de una neurona y las dendritas de la neurona siguiente.

Con relación al tipo de sinapsis, se puede establecer dos tipos de neuronas, las neuronas presinápticas que son las que envían las señales y las postsinápticas las cuales reciben. Así, se define dos tipos de sinapsis:

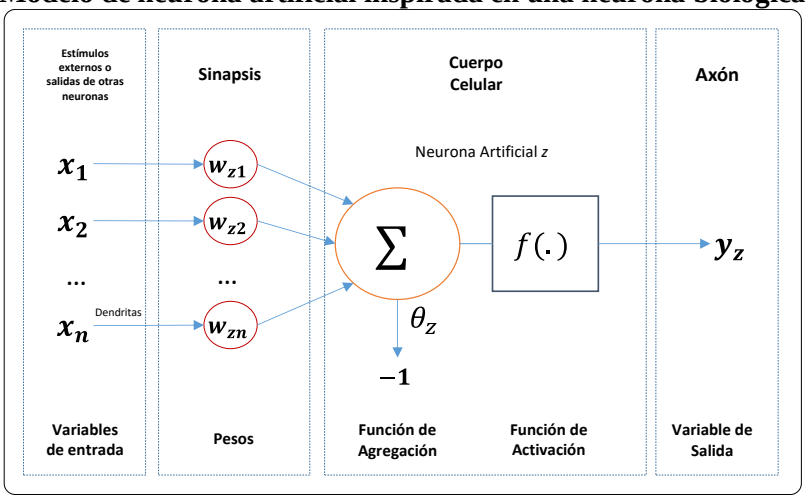
- Sinapsis excitadora: es aquella cuyos neurotransmisores facilitan la generación de impulsos a mayor velocidad.
- Sinapsis inhibidoras: Sus neurotransmisores dificultan la emisión de impulsos.

A continuación, se establece una similitud directa entre las redes neuronales artificiales y el proceso sináptico.

3.2 Elementos de una red neuronal artificial

Para introducir los elementos de una red neuronal artificial, se hará uso de la figura 3, que cuenta con una sola neurona; es decir, una pequeña parte de un sistema de red neuronal artificial. Además, se puede evidenciar que esta tiene una forma similar a la neurona biológica de la figura 2.

Figura 3
Modelo de neurona artificial inspirada en una neurona biológica



Fuente: Rico, Paredes, y Fernández (2009, 73)
Elaboración: Propia

Los aspectos que se buscan representar son los que se exponen a continuación. Se puede observar (figura 3) que se tiene un elemento central, llamado *neurona artificial* la cual recibe información del exterior o de otras neuronas. Las *dendritas* son la estructura a través de la cual la neurona artificial recibe información, que luego es procesada de acuerdo con la intensidad asignada al nexo entre la unión de las entradas y las neuronas; este nexo se denomina *sinapsis* y a la intensidad descrita se le conoce como *peso sináptico* ($\boldsymbol{W}$).

Las redes neuronales intentan reproducir el comportamiento del cerebro, por lo cual, cualquier modelo de red neuronal consta de dispositivos elementales de proceso, denominados neuronas; las cuales son definidas por Brío y Molina (2002, 12) como un dispositivo que a partir de un conjunto de entradas que proceden ya sea del exterior u otras neuronas, proporcionan una salida o respuesta.

Los elementos que constituyen a una neurona *z* son los siguientes (figura 3):

- Conjunto de entradas $x_i(t),\ i = 1, \dots, n$: las cuales pueden ser binarias o continuas, dependiendo del tipo de modelo y aplicación.

- Pesos sinápticos de la neurona *z*: catalogados como w_{zi}, que representa la intensidad de interacción entre la entrada *i* y la neurona *z.*

  Dependiendo de los pesos, se puede obtener la salida necesaria, considerando entradas específicas. Cuanto más grande sea el peso, más fuerte y relevante será el nodo de entrada.
- Función de agregación también llamada regla de propagación: Se denomina función de agregación a aquella regla que relaciona las entradas y los pesos para obtener el valor de la señal postsináptica h_z, conocido como potencial postsináptico:

$$h_z(t) = \sigma_z\big(w_{zi}, x_i(t)\big) \quad (13)$$

  La función más habitual es lineal y consiste en la suma ponderada de las entradas con los pesos sinápticos,

$$h_z(t) = \sum_{i=1}^{n} w_{zi} x_i \quad (14)$$

  Dado una entrada positiva, si el peso también es positivo, entonces este tenderá a excitar a la neurona, si el peso es negativo, tenderá a inhibirla (Brío y Molina 2002, 16).
- Función de activación**:** La misma proporciona el estado de activación, en el tiempo t, $a_z(t)$, a partir del potencial postsináptico $h_z(t)$ y del estado de activación anterior $a_z(t-1)$.

$$a_z(t) = f_z\big(a_z(t-1), h_z(t)\big) \quad (15)$$

  Sin embargo muchos modelos de redes neuronales consideran que el estado actual de la neurona (tiempo t) no depende de su estado anterior (Brío y Molina 2002, 17), por lo cual

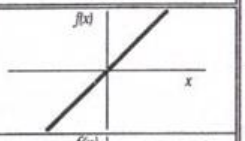
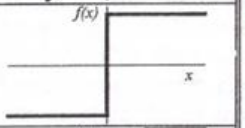
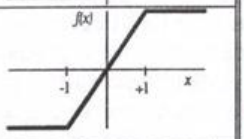
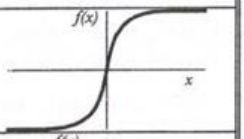
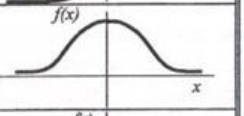
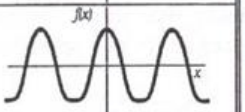
$$a_z(t) = f_z\big(h_z(t)\big) \quad (16)$$

  En general se puede establecer dos estados posibles, reposo y excitado, a los cuales se les asigna un valor que puede ser continuo o discreto (González y Hernando 2000, 52).

  En la mayor parte de modelos la función de activación $f(.)$ es monótona creciente y continua. En la figura 4, se expone las funciones de activación más usuales, en donde: $\boldsymbol{x}$ representa el potencial postsináptico y el estado de activación.

Figura 4

Funciones de activación usuales

	Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$	
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$	
Lineal a tramos	$y = \begin{cases} -1, & si\ x < -l \\ x, & si\ +l \leq x \leq -l \\ +1, & si\ x > +l \end{cases}$	$[-1, +1]$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	$[0, +1]$ $[-1, +1]$	
Gaussiana	$y = Ae^{-Bx^2}$	$[0,+1]$	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	$[-1,+1]$	

Fuente y elaboración: Brío y Molina (2002, 18)

Muchas veces se adiciona al grupo de pesos, un parámetro adicional θ_z, el cual se resta del potencial postsináptico, y representa características propias de la neurona, de tal forma que no es igual en todas ellas.

Por ejemplo, en el caso de neuronas todo-nada, el parámetro representa el nivel mínimo que debe lograr el potencial postsináptico para que la neurona se active. De tal forma, el argumento de la función de activación se expresa se la siguiente forma

$$\sum_{i=1}^{n} w_{zi} x_i - \theta_z \quad (17)$$

- Función de salida: Es la función que proporciona la salida de la neurona $y_z(t)$, que depende del estado de activación $a_z(t)$. Por lo general la función de salida es la identidad ($F(x) = x$) por lo cual la salida es considerada con el estado de activación de la neurona

$$y_z(t) = F_z\big(a_z(t)\big) = a_z(t) \quad (18)$$

Finalmente, el modelo neuronal que Brío y Molina (2002, 19) denomina como estándar queda como se muestra a continuación

$$y_z(t) = f_z\left(\sum_{i=1}^{n} w_{zi}x_i - \theta_z\right)$$

$$= f_z\left(\sum_{i=0}^{n} w_{zi}x_i\right) \quad (19)$$

Con $w_{z0} = \theta_z$ y $x_0 = -1$.

3.3 Arquitectura de las redes neuronales

En la sección anterior se mostró los principales componentes de una red neuronal. A continuación, se expone las características de cada nodo de la red, así como la organización de esta.

Generalmente, se puede encontrar tres tipos de neuronas:

- Las que toman la información de entrada, de las fuentes externas de la red.
- Las que procesan la información y generan cualquier tipo de representación interna de la misma. A estos se los denomina unidades ocultas pues no tienen relación directa con la información de entrada o de salida.
- Cuando ya se tiene procesada la información, esta pasa a los nodos de salida, los cuales dan una respuesta al sistema.

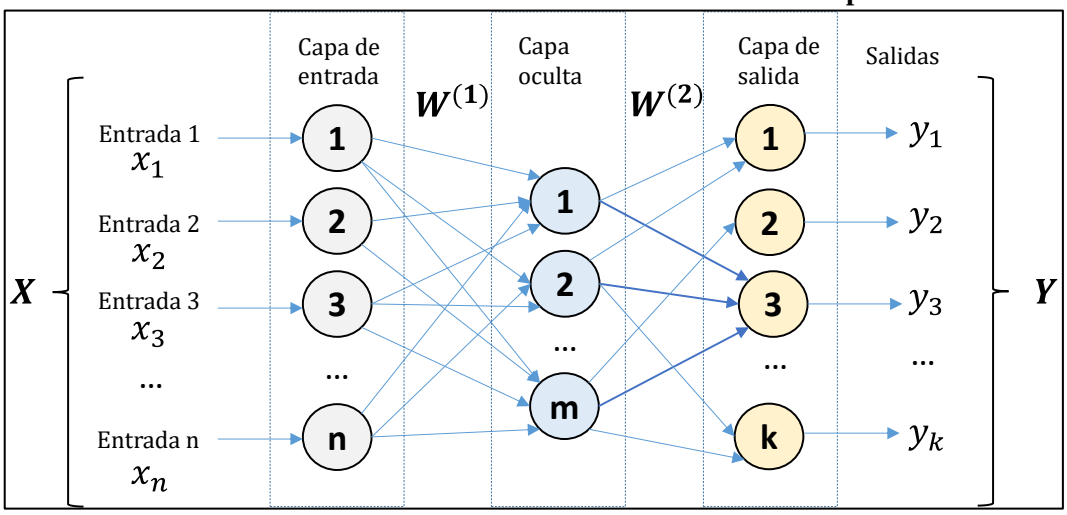
La distribución de estas neuronas está dada formando niveles o capas de un número determinado de neuronas cada una. Así, se puede determinar tres tipos de capas: de entrada, ocultas y de salida, conformadas por los tipos de neuronas ya descritas.

El número de capas ocultas puede estar entre cero y un número elevado y pueden estar interconectadas de diversas formas, estos dos aspectos determinan las distintas tipologías de redes neuronales.

Otro aspecto importante en la arquitectura de una red neuronal es la forma en la que se realizan las conexiones entre las neuronas, es decir, la forma en la que las salidas de las neuronas están encaminadas para convertirse en las entradas de otras neuronas. Incluso se puede dar que la salida de un nodo sea la entrada de sí misma, llamando a este tipo de conexión como auto recurrente.

González y Hernando (2000, 67) mencionan dos tipos de conexiones, *propagación hacia atrás* que es cuando las salidas de los nodos pueden conectarse con capas previas o del mismo nivel, incluso con sí mismos. Y *propagación hacia delante*, cuando la salida de los nodos se conecta únicamente con nodos de capas posteriores.

Figura 5
Estructura de una red neuronal artificial multicapa



Fuente: Tudela y Nardó (2011, 59)
Elaboración: Propia

Es así, que la arquitectura de las redes neuronales se basa en la forma en que se organizan y disponen las neuronas formando capas más o menos alejadas de la entrada y salida de la red, tal como se muestra en la figura 5.

En general, no se cuenta con una regla que determine el número óptimo de neuronal ocultas que ayudan a resolver un problema; sino más bien, es a base de prueba y error, realizando cambios en el que se sume o reste el número de neuronas ocultas hasta alcanzar la estructura que mejor se ajuste a la solución de un problema dado (Tudela y Nardó 2011, 59).

Se suele distinguir entre redes con una sola capa o un solo nivel de neuronas denominadas como *redes monocapa* y, con múltiples capas. Es así que, los principales parámetros de una red neuronal serían el número de capas, el número de neuronas en cada capa, el grado de conectividad y el tipo de conexión entre cada neurona (Carranza Bravo 2010, 12).

3.4 Modos de operación de una red neuronal

Se considera dos tipos de operación en un sistema neuronal: el modo recuerdo o ejecución y el modo aprendizaje.

- **Fase de aprendizaje**

Se tiene un especial interés en esta fase pues una de las principales características de una red neuronal es que son sistemas entrenables, es decir, son capaces de llevar a cabo

un específico procesamiento aprendiendo de un grupo de patrones de aprendizaje o ejemplos.

Puede definirse al aprendizaje como el proceso en el cual se modifica los pesos de la neurona en respuesta a la información de entrada. Tal como expresa, (González y Hernando 2000, 75) en el proceso de aprendizaje se destruye, modifica y crea conexiones entre las neuronas; que una conexión se destruya significa que su peso pasa a tener el valor de cero y que se cree significa que toma un valor diferente de cero.

Como ya se mencionó, en el proceso de aprendizaje se modifican los pesos de las conexiones. Por lo cual, se puede establecer que la red neuronal ha terminado su fase de aprendizaje una vez que los pesos logren estabilidad en el tiempo. Generalmente, se modifica los pesos sinápticos siguiendo cierta regla de aprendizaje, que es construida a partir de una función de error. Este proceso es iterativo, es decir, los pesos van actualizándose una y otra vez hasta que la red neuronal logra un rendimiento deseado (Brío y Molina 2002, 26).

Es importante conocer las reglas de aprendizaje de la red; que son los criterios para cambiar los pesos de las conexiones, con el objetivo que esta aprenda. Se considera dos tipos de reglas, aprendizaje supervisado y no supervisado, cuya diferencia principal radica en la existencia o no de un agente externo que controle el proceso.

Redes neuronales con aprendizaje supervisado: En este tipo de aprendizaje se tiene la participación de un agente externo o supervisor que establece la respuesta que debería tener la red a partir de una entrada específica. Este supervisor comprueba la salida de la red y si no se da la coincidencia con la deseada, se modifica los pesos de las conexiones, hasta que la salida se aproxime al valor requerido (González y Hernando 2000, 76).

Redes neuronales con aprendizaje no supervisado: Este tipo de aprendizaje no requiere de un agente externo para ajustar los pesos de las conexiones, la red no recibe información que le indique la salida deseada en función de una determinada entrada. Esto significa que no se conoce si la salida de la neurona es correcta o no, se dice que estas redes son capaces de autoorganizarse (González y Hernando 2000, 82).

Un criterio a tener en cuenta en las reglas de aprendizaje es lo que se conoce como aprendizaje *on line* y *off line*. González y Hernando (2000, 76) establecen que en el aprendizaje *on line* los pesos varían dinámicamente siempre que se ingrese nueva información al sistema; mientras que, en el aprendizaje *off line*, una vez que la red a aprendido, los pesos se mantienen fijos.

- **Fase de Recuerdo**

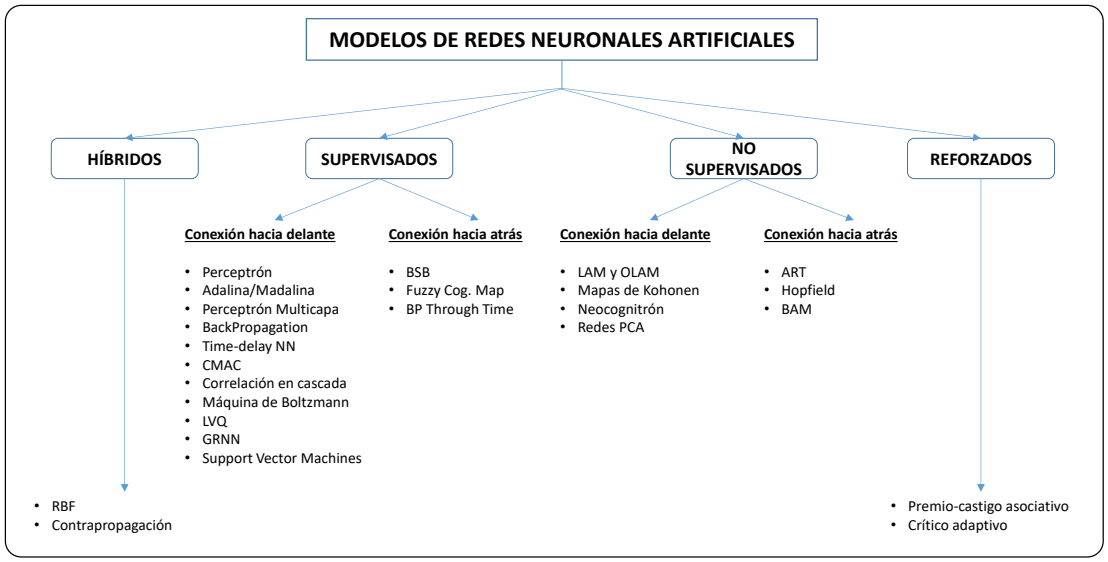
Por lo general, una vez que la red a concluido su fase de aprendizaje esta se "apaga o desconecta"; es decir, pasa a un estado *off line*, por lo cual, los pesos, conexiones y estructura de la red se mantiene fijos y esta puede procesar nueva información.

3.5 Clasificación de los modelos neuronales

Por lo expresado hasta el momento se puede deducir que dependiendo del modelo de neurona que se utilice, su arquitectura, tipo de conexión, y algoritmo de aprendizaje se obtendrán distintos modelos de redes neuronales.

En la figura 6 se expone, a modo de resumen, la clasificación de las redes neuronales por tipo de aprendizaje y arquitectura ya expuestas anteriormente.

Figura 6
Clasificación de las redes neuronales



Fuente: Brío y Molina (2002, 31).
Elaboración: Propia.

En el caso de un *credit scoring*, los nodos de entrada representan las variables independientes X, que son las características propias del solicitante de crédito. La respuesta de la neurona producirá una salida que representa la variable dependiente Y, descrita anteriormente (Jiménez-Caballero y Martínez 2000, 3).

El método que se utilizará para encontrar los pesos de la red neuronal, del presente proyecto, es el algoritmo RPROP+ (*resilient backpropagation with weight backtracking*).

Tal como lo expresa Riedmiller y Braun (1993, 586) el algoritmo Backpropagation es el más extensamente usado para aprendizaje supervisado y redes neuronales multicapa. Desafortunadamente este puede ser muy lento para aplicaciones prácticas (Schiffmann, Joost, y Werner 1994, 3). Para superar esta dificultad se propone diversas variantes a este método como el método RPROP+, el cual es usado en este trabajo. Tanto el algoritmo Backpropagation como su variante RPROP+, son descritos a continuación.

3.6 Algoritmo Backpropagation

El algoritmo de backpropagation es el más ampliamente usado para modelos con aprendizaje supervisado multicapa. La idea básica de este algoritmo es la siguiente:

En un espacio de N+1 dimensiones, donde N es el número de pesos de la red, se representa una superficie que muestre el error que se genera en la red neuronal, para un determinado valor en los pesos de esta.

El algoritmo backpropagation hace que se vaya bajando por la superficie del error hasta lograr un mínimo, es por esta razón que la variación de un peso w_{ij} de la red, en una iteración, al procesar un conjunto de patrones p_i, es proporcional al gradiente descendente (González y Hernando 2000, 146).

$$\Delta w_{ji} = -\alpha \frac{\partial E_p}{\partial w_{ji}} \quad (20)$$

Considerando a E la función de error, se tiene que:

$$\frac{\partial E_p}{\partial w_{ji}} = \frac{\partial E_p}{\partial y_{pj}} \times \frac{\partial y_{pj}}{\partial Net_j} \times \frac{\partial Net_j}{\partial w_{ji}} \quad (21)$$

Donde w_{ji} representa el peso de la neurona j a la neurona i, y_{pj} es la salida de la neurona j en el patrón p, y Net_j es la suma ponderada de las entradas a la neurona j, es decir:

$$Net_j = \sum_{i=1}^{k} w_{ji} \times y_i \quad (22)$$

$$y_{pj} = f\left(Net_j\right)$$

Siendo k, en número de entradas de la neurona j y f la función de activación, derivable.

Es así, que la actualización de los pesos está dada por la siguiente ecuación:

$$w_{ji}(t+1) = w_{ji}(t) - \alpha \frac{\partial E_p}{\partial w_{ji}}(t+1) \quad (23)$$

Es evidente que la elección de α tiene un efecto importante en el tiempo de convergencia del algoritmo, el cual se detiene cuando el error resulte aceptablemente pequeño para cada uno de los patrones aprendidos.

3.7 Algoritmo RPROP+

Una alternativa al algoritmo de backpropagation es el resiliente backpropagation (RPROP+) en el que, en lugar de usar la magnitud de la derivada $\frac{\partial E_p}{\partial w_{ji}}$, presente en la ecuación 20, se considera únicamente su signo multiplicado por una constante. Este algoritmo tiene la ventaja de ser uno de los algoritmos de aprendizaje más rápidos (Almeida et al. 2009, 4).

El algoritmo RPROP+ consiste en los siguiente (Igel y Hüsken 2000, 116): Para cada peso se introduce su valor de actualización Δ_{ji}, que determina el tamaño de la actualización del peso.

$$\Delta_{ji}^{t} = \begin{cases} \eta^{+} \times \Delta_{ji}^{t-1}, & si \ \frac{\partial E_p}{\partial w_{ji}}^{(t-1)} \times \frac{\partial E_p}{\partial w_{ji}}^{(t)} > 0 \\ \eta^{-} \times \Delta_{ji}^{t-1}, & si \ \frac{\partial E_p}{\partial w_{ji}}^{(t-1)} \times \frac{\partial E_p}{\partial w_{ji}}^{(t)} < 0 \\ \Delta_{ji}^{t-1}, & caso\ contrario \end{cases} \quad (24)$$

Donde $0 < \eta^{-} < 1 < \eta^{+}$. Δ_{ji} están acotados por dos parámetros Δ_{min} y Δ_{max}.

Una vez obtenido los tamaños de actualización es necesario obtener el valor de variación de los pesos Δw_{ji}, distinguiendo dos casos.

Si el signo de la derivada parcial no ha cambiado se tiene:

$$Si \ \frac{\partial E_p}{\partial w_{ji}}^{(t-1)} \times \frac{\partial E_p}{\partial w_{ji}}^{(t)} \geq 0 \ entonces \ \Delta w_{ji}^{t} = -sign\left(\frac{\partial E_p}{\partial w_{ji}}^{(t)}\right) \times \Delta_{ji}^{t} \quad (25)$$

Donde el operador signo retorna el valor de +1 si el argumento es positivo, -1 si es negativo y 0 en otro caso. En caso de que el signo de la derivada parcial cambia, se tiene:

$$Si \ \frac{\partial E_p}{\partial w_{ji}}^{(t-1)} \times \frac{\partial E_p}{\partial w_{ji}}^{(t)} < 0 \ entonces \ \Delta w_{ji}^{t} = -\Delta w_{ji}^{t-1} \ y \ \frac{\partial E_p}{\partial w_{ji}}^{(t)} = 0 \quad (26)$$

Finalmente se actualiza los nuevos pesos, los cuales están dados por,

$$w_{ji}(t+1) = w_{ji}(t) + \Delta w_{ji}(t) \quad (27)$$

La función de error que será usada para optimizar los pesos es la suma de los errores cuadráticos definida de la siguiente forma (Ladino 2014, 14):

$$E = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (28)$$

Donde n es el número de datos en entrenamiento, observaciones, y_i es la salida deseada y $\hat{y}_i$ es la salida de la red.

4. Conceptos del proceso de modelización

Para la construcción de los modelos scoring, ya sea con el uso de regresión logística o redes neuronales, es necesario definir algunos conceptos:

4.1 Prueba Dickey – Fuller aumentada

La prueba de Dickey Fuller aumentada permite corroborar la estacionariedad de una serie de tiempo. Cuando una serie no es estacionaria presenta al menos una raíz unitaria, por lo que se busca contrastar la siguiente hipótesis (Santos 2008, 197):

- Ho: la serie tiene raíz unitaria.
- Ha: la serie es estacionaria.

Se rechaza Ho si:

Valor absoluto (Estadístico de la prueba (ADF)) > valores críticos de la prueba (1%, 5%, 10%)

4.2 Muestreo aleatorio simple

El tamaño de muestra necesario para estimar *p* (proporción poblacional) con un límite para el error de estimación *B*, está dado por (Scheaffer, Mendenhall, y Ott 2013, 100):

$$n = \frac{Npq}{(N-1)D + pq} \quad \text{donde: } q = 1 - p \quad y \quad D = \frac{B^2}{4} \quad (29)$$

4.3 Roll Rate

Implica comparar el peor atraso en un X número de meses, con los posteriores X meses. Y así, calcular el porcentaje de operaciones o clientes que mantienen, mejoran o empeoran su comportamiento, en distintos rangos de atraso. El propósito es identificar el

punto de no retorno; es decir, el nivel de atraso en la cual, una operación es considerada como insalvable (Siddiqi 2006, 40). [1]

4.4 Valor de Información (IV)

El valor de información es un valor numérico que permite cuantificar el poder de predicción de una variable independiente. El mismo, funciona con variables categóricas (Zeng 2013, 6). Se calcula a partir de la siguiente expresión,

$$IV = \sum_{i=1}^{n} \left(\frac{b_i}{b} - \frac{m_i}{m} \right) \times \ln \left(\frac{b_i / b}{m_i / m} \right) \quad (30)$$

Donde, n es el número de categorías en la variable independiente, b_i y m_i es el número de buenos y malos clientes dentro de la categoría *i.* Y *b* y *m:* el número total de buenos y malos clientes en el periodo de modelo, respectivamente.

Intuitivamente, mientras más grande sea el valor del IV, más predictiva será la variable independiente categorizada. Sin embargo las variables con valores superiores a 0.5 deberían ser revisadas pues puede darse el caso de sobreestimación (Siddiqi 2006, 82).

Una regla, propuesta por (Siddiqi 2006, 81), para los valores aceptados del IV es:

Figura 7
Regla valor de la información IV

Valor IV	Nivel de Predicción
Menor que 0,02	No predictivo
Entre 0,02 y 0,1	Débil
Entre 0,1 y 0,3	Medio
Mayor que 0,3	Fuerte

Fuente: Siddiqi (2006, 81)
Elaboración: Propia

4.5 Backward Stepwise y criterio de información de Akaike (AIC)

La técnica de pasos hacia atrás (Backward Stepwise) consiste en introducir en el modelo todas las variables e ir excluyendo una tras otra según algún criterio de evaluación (Moral 2006, 200). En este caso, se utiliza el AIC como criterio de selección de variables, considerando aquellas que logren el menor valor AIC.

[1] Es decir, no se recupera el monto que la entidad otorga a la persona.

El criterio de información de Akaike penaliza los modelos con muchos parámetros y busca determinar la significancia de incluir algunos parámetros en el modelo (Sánchez 2012, 70). Este criterio se define como sigue:

Para un conjunto de *m* modelos, de tal forma que,

$$M_j^m \supset M_{j-1}^{m-1} \supset \cdots \supset M_{j-i}^1 \quad (31)$$

Donde, $j > i, i > 0$ y los subíndices denotan el número de variables en cada modelo. Se busca elegir aquel valor de j que minimice la siguiente expresión,

$$AIC = -2\log L + 2k \quad (32)$$

Siendo, L es la función de máxima verosimilitud, que se define como:

$$L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2}\sum_{t=1}^{n}\frac{e_t^2}{\sigma^2} \quad (33)$$

Donde, k es igual al número de parámetros del modelo, n corresponde al número de datos para la construcción de este y σ^2 el promedio de los residuales e_t al cuadrado. Como se puede observar, AIC penaliza modelos según su desviación de los datos reales, siendo el mejor modelo aquel que manifieste el valor más pequeño entre todos los modelos evaluados, siendo este el que mejor ajusta a los datos (Sánchez 2012, 71).

4.6 Estadístico de Wald

Para corroborar la significancia de los coeficientes de un modelo logit, se contrasta la siguiente hipótesis nula:

$$H_0: \beta_i = 0$$

Considerando a β_i como el coeficiente estimado correspondiente a la variable independiente x_i.

Así, se cuenta con el estadístico de Wald (Wasserman 2010, 183): Sea θ un parámetro escalar, $\hat{\theta}$ un estimador de θ y sea $\widehat{se}$ el error estándar estimado de $\hat{\theta}$. Considerar probar $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. Se asume que $\hat{\theta}$ es asintóticamente normal,

$$\frac{\sqrt{n}\,(\hat{\theta} - \theta_0)}{\widehat{se}} \longmapsto N(0,1). \quad (34)$$

Con un nivel de significancia de α, la prueba de Wald establece: Se rechaza H_0 cuando $|W| > z_{\alpha/2}$ donde,

$$W = \frac{\hat{\theta} - \theta_0}{\widehat{se}} \quad (35)$$

En este caso, $\theta_0 = 0$, por lo cual el estadístico de Wald se simplifica a,

$$W = \frac{\hat{\beta}}{\widehat{se}} \quad (36)$$

4.7 GVIF: Factor de inflación de la varianza generalizado

GVIF es una medida de cuanto de la varianza del coeficiente estimado β_j está inflado por la existencia de correlación entre las variables independientes del modelo. Se usa comúnmente cuando se tiene variables categóricas con más de dos valores posibles o con variables polinomiales (Fox 2003, 16), está dado por,

$$GVIF_i = \frac{\det R_i \times \det R_{-i}}{\det R} \quad (37)$$

Siendo $\det R_i$ el determinante de la matriz de correlación sobre las columnas de la variable i, $\det R_{-i}$ el determinante de la matriz de correlación sobre las columnas del resto de variables del modelo, distintas a la variable i y $\det R$ el determinante de la matriz completa de correlación. Cuando el número de coeficientes de cada variable es uno, el GVIF coincide con VIF.[2] A pesar, de no tener una regla formal, generalmente se acepta que valores de GVIF superiores a 10 puede ser perjudicial (Yoo et al. 2014, 3).

4.8 Estadístico de Kolmogorov – Smirnov (KS)

El estadístico de Komogorov Smirnov consiste en medir cuan distintas son las funciones de distribución de buenos y malos clientes para cada valor de puntaje score.

El valor del estadístico está dado por (Rezac y Řezáč 2011, 4):

$$KS = \ Max_i\big(P_b(i) - P_g(i)\big) \quad (38)$$

Donde:

i: es el valor de score, en el rango L - H, que es, $L \leq i \leq H$.

$\boldsymbol{P_g(i)}$, $\boldsymbol{P_b(i)}$: Proporción de buenos y malos con score menor o igual a i, en la población.

$$P_g(i) = \frac{N_g(i)}{N_g} = \sum_{j=L}^{i} p_g(j)$$

[2] Factor de inflación de la varianza (VIF): $= \frac{1}{1-R_i^2}$ siendo R_i^2 el coeficiente de determinación, que representa la proporción de la varianza de la variable independiente *i*, que se asocia con las otras variables independientes en el modelo.

$$P_b(i) = \frac{N_b(i)}{N_b} = \sum_{j=L}^{i} p_b(j) \quad (39)$$

Con:

N_g: El total de buenos en la población.

N_b: El total de malos en la población.

$N_g(i),\ N_b(i)$: El número de buenos y malos en la población con scores menores o iguales a *i*.

$$N_g(i) = \sum_{j=L}^{i} n_g(j) \ \ y$$

$$N_b(i) = \sum_{j=L}^{i} n_b(j) \quad (40)$$

$p_g(i),\ p_b(i)$: Es la proporción de buenos con score *i* y malos con score *i* en la población.

$$p_g(i) = \frac{n_g(i)}{N_g} p_b(i) = \frac{n_b(i)}{N_b} \quad (41)$$

Finalmente, $n_g(i)\ \ y\ \ n_b(i)$ el número de buenos casos y malos con score *i*, en una población.

El estadístico puede ser usado para medir la capacidad de clasificación de un modelo, tomando valores entre 0 y 1. Se considera que un modelo con un KS menor a 20% debe ser cuestionado y mayor a 70% sea, probablemente, muy bueno para ser cierto (Anderson 2007, 196). TransUnion, empresa líder en gestión de la información de crédito, mide la efectividad de sus scores de riesgos de acuerdo con los siguientes estándares internacionales.

Figura 8
Estándar estadístico KS

Fuente y elaboración: TransUnion (2012, 3)

4.9 Coeficiente de GINI

El coeficiente de Gini es un estadístico usado para medir cuan bien el modelo *scoring* distingue entre los buenos y malos clientes, toma valores entre 0 y 1, considerando que, si el coeficiente de Gini es igual a 1, entonces el modelo separa perfectamente a buenos y malos (Anderson 2007, 204).

El estadístico está dado por la siguiente ecuación:

$$Gini = 1 - \sum_{i=L}^{H'} (P_b(i+1) - P_b(i))(P_g(i+1) + P_g(i)) \quad (42)$$

Donde:

i: es el valor de score, en el rango L - H, que es, $L \leq i \leq H$.

$\boldsymbol{P_g(i)}$, $\boldsymbol{P_b(i)}$: Proporción de buenos y malos con score menor o igual a *i*, en la población, respectivamente.

Se considera que para modelos de originación, un coeficiente de Gini menor a 35% es sospechoso y mayor o igual a 50% es más que satisfactorio (Anderson 2007, 205).

4.10 Matriz de confusión: error, sensibilidad y especificidad

Una forma simple de evaluar la predictividad de un modelo, es calculando el porcentaje de clientes que este clasifica de forma correcta. Este porcentaje de clasificación es derivado, de la matriz de confusión, la cual es construida de la siguiente forma:

- Elegir un punto de corte, para los valores de score obtenidos.
- Clasificar a todas las operaciones o clientes con un valor de score por debajo del punto de corte, como malos esperados. Y a los que tiene un valor de score por encima del punto de corte, como buenos esperados.
- Construir una tabla cruzada entre la clasificación real u original de bueno/malo y la de bueno/malo esperado, obtenida en el paso anterior.
- Calcular las diferentes ratios que pueden ser obtenidos del modelo, como error, valor de sensibilidad y especificidad.

Los casos correctamente clasificados son denominados como verdaderos positivos (buenos) y verdaderos negativos (malos). Si no son bien clasificados se tiene a los falsos positivos (malos que son clasificados como buenos) y a los falsos negativos (buenos que son clasificados como malos).

Además, se define como sensibilidad y especificidad, a la habilidad del modelo de catalogar correctamente al cliente bueno y malo, respectivamente. La forma general

de la matriz de confusión es mostrada en la tabla 1. Y en la tabla 2 se expone los indicadores de eficiencia, que serán usados para comparar los dos modelos desarrollados, construidos a partir de la matriz de confusión.

Tabla 1
Matriz de Confusión

Matriz de Confusión		Real	
		0: Malo	1: Bueno
Estimado	0: Malo	A	B
	1: Bueno	C	D

Fuente y elaboración propias

Tabla 2
Indicadores de eficiencia

Nombre de Indicador	Definición	Fórmula
Tasa de aciertos	Cociente entre el número de predicciones correctas y el total.	$\frac{A+D}{A+B+C+D}$
Error	Cociente entre el número de predicciones incorrectas y el total.	$\frac{B+C}{A+B+C+D}$
Sensibilidad	Cociente entre el número de buenos clasificados correctamente y el total de buenos.	$\frac{D}{B+D}$
Especificidad	Cociente entre el número de malos clasificados correctamente y el total de malos.	$\frac{A}{A+C}$

Fuente y elaboración propias

Una forma de obtener el punto de corte que permita nivelar la predicción correcta de buenos y malos es hacer uso del Índice de Youden, que está dado por:

$$Índice\ de\ Youden\ (YI) = \max(Sensibilidad + Especificidad - 1) \quad (42)$$

En una curva ROC, que será definida en el siguiente apartado, el índice de Youden es la distancia vertical máxima entre la curva y la diagonal. Siendo el punto de corte óptimo, aquel en el cual se alcanza el valor de YI (Cadarso Suárez 2010, 19).

4.11 Estadístico AUROC

Es una medida de performance cercanamente relacionada al coeficiente de Gini y mide el área entre la curva y la diagonal conformada en la gráfica de $sensibilidad\ vs\ (1 - especificidad)$, comúnmente llamada Curva ROC, para el total de los valores de probabilidad. Permite medir la eficacia predictiva de un modelo, evaluando gráficamente la capacidad de este para discriminar entre buenos y malos.

La relación entre el estadístico AUROC y el coeficiente de Gini es el que se muestra a continuación (D. Hand y Till 2001, 173):

$$AUROC = \frac{Gini + 1}{2} \quad (43)$$

El valor de AUROC varía entre 0 y 1. Un valor de 0,5 implicaría que el modelo es igual a hacer una clasificación aleatoria para catalogar a un cliente como bueno o malo, un valor de 1 significa que las clasificaciones son perfectamente correctas y un valor de 0 implica que las mismas con totalmente incorrectas (Anderson 2007, 207). Tal como expresa (Siddiqi 2006, 124) el valor del estadístico AUROC debe estar sobre 0,5, y un valor superior a 0,7 se considera adecuado.

Capítulo tercero: Proceso de modelación

En el presente capítulo se busca exponer los resultados obtenidos en el proceso de modelación de un modelo *credit scoring*, previo a aplicar las metodologías de regresión logística y redes neuronales.

Se aplicará las dos metodologías, en el proceso de modelización de un scoring, para una cartera de microcrédito de una institución financiera ecuatoriana de mediano tamaño, del sector privado; la cual será nombrada Institución X. Mencionada cartera de Microcrédito, corresponde a créditos desembolsados entre enero del 2014 y diciembre del 2017.

Además, es importante mencionar que las operaciones seleccionadas corresponden a clientes bancarizados, es decir, que fueron reportados en la central de riesgos al menos una vez durante los últimos treinta y seis meses, anteriores al correspondiente desembolso; esto con el fin de hacer uso de variables de central de riesgos para explicar el comportamiento del cliente.

Los modelos *credit scoring* pueden ser aplicados en cualquier momento durante el ciclo de vida de un crédito; diferenciándose principalmente por

a) El objetivo para el que se lo construye, y
b) La disponibilidad de información con la que se cuenta.

Tal como lo expresa Bolivia et al. (2008, 56) por lo general se clasifica los modelos scoring en dos tipos:

- **Scoring de originación:** Interviene en la fase de aprobación o evaluación de solicitudes de nuevos créditos.
- **Scoring de comportamiento:** Se realiza para dar seguimiento a los clientes ya incorporados a la institución.

En el presente estudio se busca desarrollar dos modelos *credit scoring* de originación. Es decir, los modelos buscarán determinar la probabilidad de incumplimiento asociado a solicitudes de nuevos créditos, permitiendo:

a) Segmentar la población de solicitantes de crédito asociando al solicitante una probabilidad de incumplimiento, la cual puede ser expresada como un puntaje (score) de tal forma que a mayor puntaje menor probabilidad de incumplimiento, y a menor puntaje, mayor probabilidad de incumplimiento; esto, dependiendo de

cómo haya sido definida la variable dependiente bueno/malo, la cual se explicará más adelante.

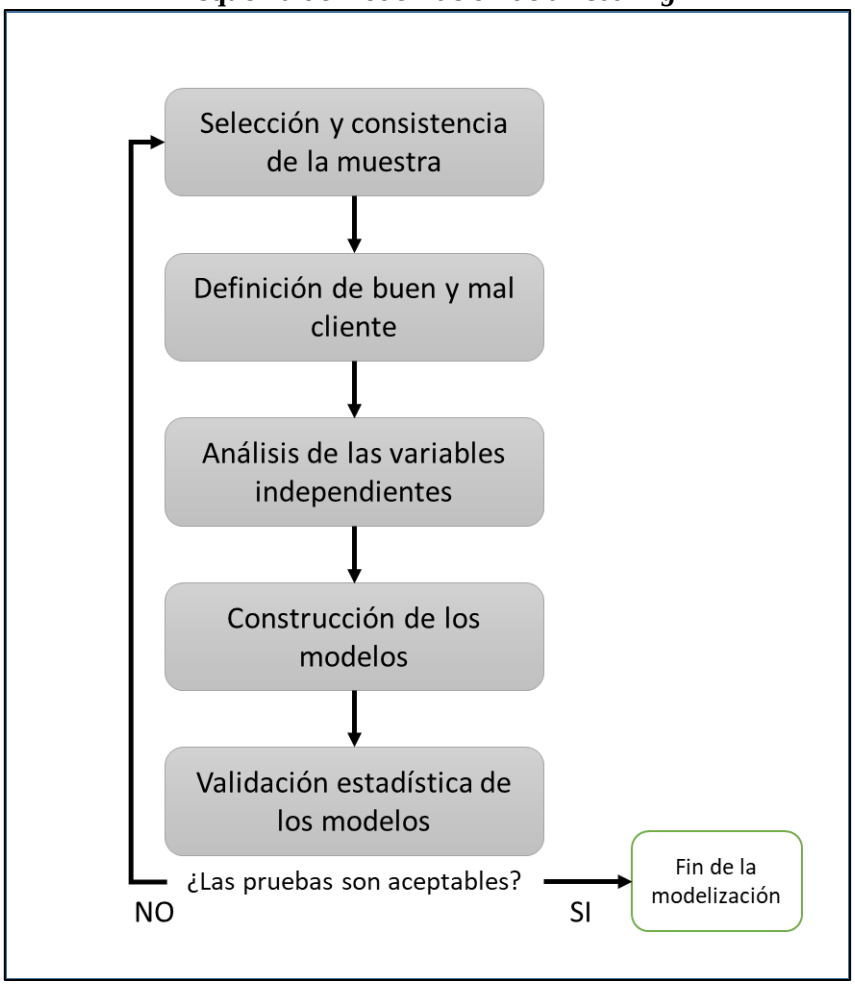
b) Apoyar en la toma de decisiones de aprobación o rechazo de las mismas.

c) Establecer estrategias comerciales diferenciadas por punto de corte, pues se puede discriminar a solicitantes que deben ser rechazados (puntaje más bajo) de aquellos cuya aprobación de crédito debería ser inmediata (puntaje más alto).

De esta forma, se plantea las siguientes etapas para el desarrollo de los modelos *scoring* de originación:

Figura 9

Esquema de modelización de un *scoring*



Fuente y elaboración propias

En este capítulo, se desarrollará los puntos correspondientes a:

- Selección y consistencia de la muestra.
- Definición de un buen y mal cliente.
- Análisis de las variables independientes.

1. Selección y consistencia de la muestra

Para el desarrollo de un modelo *scoring* es necesario contar con una muestra de clientes previos, y su respectiva información crediticia; la cual permitirá explicar su comportamiento. Tal muestra, debe cumplir dos principios fundamentales:

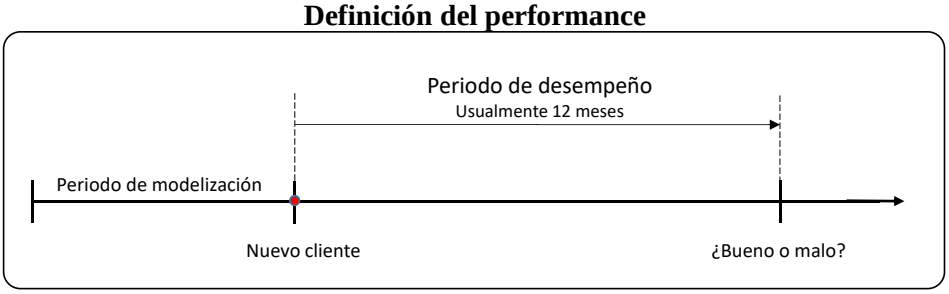
- La muestra debe ser representativa: Representativa en el hecho que esta debe representar a aquellos solicitantes que aplicarán a un crédito en un futuro.
- Y debe contar con información suficiente: Este punto se refiere a que la muestra debe incorporar casos de diferente tipo de comportamiento de pago (bueno y malo), de tal forma que sea posible identificar las características, más representativas, que reflejen el comportamiento de pago de los clientes de la muestra (Thomas, Edelman, y Crook 2002, 137).

Por la necesidad que la muestra sea representativa, se prefiere considerar grupos de muestras recientes, pues son los que presentarán características más similares a potenciales clientes.

Sin embargo, un *scoring* es construido bajo el supuesto que el comportamiento futuro será un reflejo del comportamiento pasado y para definir este comportamiento es necesario establecer una definición de buen o mal cliente. En este sentido, es necesario contar con un periodo de tiempo al cual se denomina periodo de desempeño, en el que se pueda observar el comportamiento de pago del cliente y de esta forma poder catalogarlo como bueno o malo.

Por otra parte, para definir a un cliente como bueno o malo se establece una muestra a la cual se denomina periodo de observación o periodo de modelización, en ella no se puede considerar grupos de créditos recientes pues no se contaría con el tiempo suficiente para observar el comportamiento de pago del cliente. De esta manera es importante establecer dos periodos, tal como se ilustra en la figura 10:

Figura 10
Definición del performance



Fuente: Siddiqi (2006, 33)
Elaboración: Propia

La figura 10, sintetiza los dos periodos que se deben determinar al momento de seleccionar la muestra.

- **Periodo de modelización:** Periodo que cumpla con los requisitos de ser madura y estable. Representa el tiempo en el cual se sitúa el modelador para observar el desempeño del cliente.
- **Periodo de desempeño:** Tiempo en el cual se puede definir el comportamiento de pago de un cliente, de tal forma que pueda ser catalogado como buen o mal cliente.

Tal como lo expresa Thomas, Edelman, y Crook (2002, 137), usualmente se fija un periodo de desempeño de 12 meses para un scoring de originación. Por lo cual, la interrogante es conocer cuál será el periodo de modelización que cumpla con las dos características descritas (madura y estable).

2. Determinación del periodo de modelización

Se busca determinar un periodo de tiempo que cumpla con las características de madurez y estabilidad en el tiempo.

a) Madurez

Se considera a una cartera como madura cuando cuenta con el tiempo suficiente para determinar el comportamiento de pago del cliente; es decir, que pudo ser observada en un periodo igual al periodo de desempeño. Es importante la madurez de la cartera pues permite no calificar como bueno a un cliente malo, ya que si no se cuenta con el tiempo suficiente de observación no se podrá establecer su comportamiento de pago, de forma correcta (Bambino Contreras 2005, 22).

b) Estabilidad

Es importante considerar un periodo en el cual la población de clientes presenta un comportamiento estable, para de esta forma alcanzar la madurez adecuada de la cartera. Un indicador comúnmente usado es la tasa de morosidad, la cual es calculada para cada mes de desembolso o colocación (cosecha) dentro del periodo de modelización (Bambino Contreras 2005, 23) ,

$$Tasa\ de\ morosidad = \frac{Número\ de\ clientes\ en\ mora}{Número\ total\ de\ clientes} \quad (44)$$

La tasa de morosidad busca representar, por mes de colocación, la relación entre los clientes en mora sobre el total de clientes, con el propósito de establecer los periodos

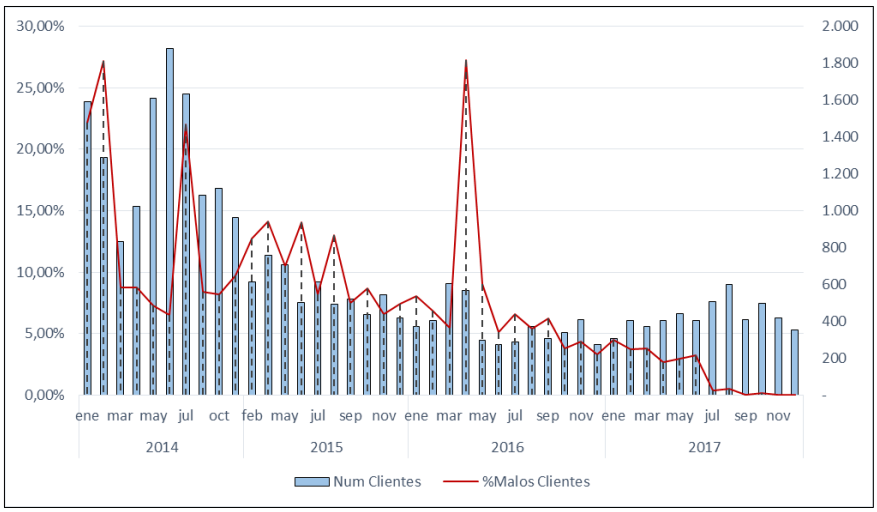
en los cuales esta relación se estabiliza, denotando un comportamiento estable de la cartera.

Se dice que una cartera es estable si mantiene su tasa de morosidad a lo largo del tiempo; es decir, la relación en la cosecha t, presenta pequeñas variaciones en el tiempo t+1 y t-1. Estadísticamente, se busca comprobar que no existe tendencia, en esta relación, en el periodo de tiempo considerado.

En la figura 11, se tiene la representación de la tasa de morosidad para las cosechas comprendidas entre enero 2014 y diciembre 2017, considerando como cliente en mora a aquel con un atraso mayor a treinta días en su periodo de desempeño, pues es el corte que la institución financiera X considera en sus procesos.

Como se puede observar, para las cosechas más cercanas a diciembre 2017 la tasa de morosidad decrece, esto no es necesariamente por que haya mejorado el comportamiento de los clientes sino más bien porque a estas cosechas les falta tiempo de maduración o de observación de su comportamiento.

Figura 11
Análisis de cosechas



Fuente y elaboración propias

En el presente caso de estudio, se considera un periodo de desempeño igual a 12 meses, que como ya se mencionó es el más común para scores de originación; por lo cual, únicamente se toma en cuenta a cosechas comprendidas entre enero 2014 y diciembre 2016, como candidatas para el periodo de modelización.

Es así como, se busca demostrar que la serie definida entre enero 2014 y diciembre 2016 es estacionaria, y de esta forma definir al periodo de modelización como estable.

Para corroborar la estacionariedad[3] de la serie conformada por la tasa de morosidad de las cosechas comprendidas entre enero 2014 y diciembre 2016, se hace uso de la prueba de Dickey - Fuller aumentada. La prueba es realizada en el programa estadístico Eviews 8, cuyos resultados son expuestos a continuación.

Figura 12
Prueba ADF – Raíz Unitaria

Hipótesis nula: La serie MICRO tiene una raíz unitaria
Exógena: Constante
Lag Length: 0 (Automatic - based on SIC, maxlag=7)

		Estadístico-t	Prob. *
Estadístico Dickey-Fuller Aumentado		-4,592719	0,0009
Valores críticos:	1% nivel	-3,661661	
	5% nivel	-2,960411	
	10% nivel	-2,619160	

*MacKinnon (1996) one-sided p-values.

Ecuación de la prueba Dickey-Fuller aumentado
Variable dependiente: D(MICRO)
Método: Mínimos cuadrado
Fecha: 02/16/19 Tiempo: 12:35
Muestra (ajustada): 2014M02 2016M08
Observaciones incluidas: 31 después de ajustes

Variable	Coeficiente	Std.. Error	Estadístico-t	Prob.
MICRO (-1)	-0,795826	0,173280	-4,592719	0,0001
C	0,076068	0,020800	3,657040	0,0010

R-cuadrado	0,421078	Media dependiente var.	-0,006074
R-cuadrado ajustado	0,401115	S.D. dependiente var.	0,076403
S.E. de la regresión	0,059127	Akaike info. criterio	-2,755928
Sum. Res. Cuadrados	0,101383	Schwarz criterio	-2,663412
Log likelihood	44,71688	Hannan-Quinn criter.	-2,725770
F-estadístico	21,09307	Durbin-Watson estat.	2,098933
Prob(F-estadístico)	0,000079		

Fuente y elaboración propias

Según lo evidenciado en la figura 12, el p-valor de la prueba es igual a 0,0009 por lo cual se rechaza la hipótesis nula al 95% de confianza. Es decir, la serie tasa de morosidad es estacionaria para el periodo de modelización; por lo cual se establece que el periodo de modelización es estable.

Con las consideraciones anteriores se estableció como periodo de modelización de enero 2014 a diciembre 2016, con un total de 23.138 clientes cuyo comportamiento es

[3] Un proceso se dice estacionario si cumple tres supuestos: 1. La varianza es finita y constante a lo largo del tiempo. 2. La media es constante a lo largo del tiempo. 3. La auto covarianza para dos puntos distintos de tiempo solo depende del lapso de tiempo entre estos (Santos 2008, 3).

maduro, estable y cumple con lo dispuesto por la Superintendencia de Bancos y Seguros quién establece contar con al menos tres años para el periodo de modelización.

De manera específica el Artículo 7, sección III.- "Metodología y procesos de administración de riesgo de crédito", del título X.- "De la gestión y administración de riesgos", del libro I, de la Codificación de Resoluciones de la Superintendencia de Bancos y Seguros y de la Junta Bancaria señala:

> "Metodologías y técnicas analíticas basadas en el comportamiento histórico de los portafolios de inversión y de las operaciones de crédito y contingentes, que permitan determinar la pérdida esperada sobre la base de la probabilidad de incumplimiento, el nivel de exposición y la severidad de la pérdida. Para el cálculo de estos componentes se deberá disponer de una base de datos mínima de tres (3) años inmediatos anteriores, que deberá contener elementos suficientes para el cálculo de los aspectos señalados" (SBS 2003).

En este punto se busca establecer la representatividad de la muestra y nace la pregunta ¿Qué muestra? Tal como lo expresa Siddiqi (2006, 63) en todo desarrollo de un *scoring* se debe segmentar a la muestra total del periodo de modelización en un conjunto de datos para el desarrollo y validación del mismo.

Hay varias formas para segmentar al periodo de modelización, para obtener un conjunto de datos que sirva en el desarrollo del scoring, al cual se denomina *Periodo de Modelo* y los datos que sirva para validar el scoring, que será denominado *Periodo de Testing.*

Normalmente, se considera un porcentaje del 70% al 80% para el periodo de Modelo y el 20% al 30% restante como periodo de Testing. Si el periodo de modelización fuera pequeño se podría considerar el 100% del mismo, para el desarrollo del scoring y tomar submuestras aleatorias de 50% a 80% para validar (Siddiqi 2006, 63).

Es así que, se debe corroborar que la submuestra del periodo de modelo sea representativa para el periodo de modelización, para esto se aplica el método de muestreo aleatorio simple, expuesto en el marco teórico, cuyos resultados se muestran a continuación.

Se estima la muestra para distintos niveles de error, con un número de clientes en el periodo de modelización de 23.138 (N) y una tasa de morosidad de 11,73% (p).

Tabla 3
Tamaño muestra periodo de modelo

B	n	% Muestra Modelo
0,50%	9.653	41,7%
0,30%	15.395	66,5%
0,25%	**17.315**	**74,8%**
0,20%	18.911	81,7%
0,19%	19.254	83,2%
0,18%	19.591	84,7%
0,17%	19.921	86,1%
0,16%	20.242	87,5%
0,15%	20.554	88,8%
0,14%	20.854	90,1%
0,13%	21.141	91,4%
0,12%	21.415	92,6%
0,11%	21.673	93,7%
0,10%	21.913	94,7%
0,09%	22.136	95,7%
0,08%	22.339	96,5%
0,07%	22.521	97,3%
0,06%	22.682	98,0%
0,05%	22.819	98,6%

Fuente y elaboración propias

A partir de lo expuesto se concluye lo siguiente:

- **Periodo de Modelo:** Se toma de forma aleatoria un total de 17.315 clientes del total comprendido entre las cosechas de enero 2014 a diciembre 2016, representando el 74,8% del total del periodo de modelización que equivale a un error del 0,25% (Ver tabla 3). En este periodo se desarrollan los dos modelos *scoring*.
- **Periodo de Testing:** Corresponde al 25,2% restante de los registros presentes en el periodo de modelización, pero no en el periodo de modelo (5.823 registros). Este periodo servirá para validar los modelos *scoring* que se desarrollen a partir de la información presente en el periodo de modelo.

3. Definición de buen y mal cliente

Un siguiente paso, en el proceso de modelización, y que va de la mano con la definición de los periodos de modelo y testing, corresponde a la definición de buen y mal cliente.

Para el caso de modelos de riesgo, la definición de un mal cliente es aquel que, de conocer su comportamiento, la institución, no le hubiese desembolsado; por lo cual, para

catalogar a un cliente como bueno o malo se hace uso de su comportamiento de pago, luego del desembolso.

Sin embargo que una cartera sea catalogada como mala, no significa que la restante sea buena, es por esta razón que se incluye la categoría de indeterminados, para asegurar que se logre una correcta distinción entre buenos y malos clientes (Thomas, Edelman, y Crook 2002, 138).

Tal como expresa Siddiqi (2006, 40) una de las metodologías para la definición de cliente es hacer uso de un análisis de *roll rate*, en donde se busca conocer la tasa de rotación de los clientes que hallándose en un rango de atraso específico, transitan a un distinto rango, luego de un periodo de tiempo.

Tabla 4
Roll Rate – Definición de Cliente bueno, malo e indeterminado

Rangos de atraso máximo luego de 6 meses del desembolso	Número de clientes en los rangos de atraso, luego de 6 meses del desembolso	Número de clientes con atraso mayor a 90 días, luego de 12 meses de su desembolso	Roll Rate
a: 0 - 16	20.386	121	0,6%
b: 17 - 30	1.425	189	13,3%
c: 31 - 90	602	422	70,1%
d: >90	725	725	100,0%
Total	**23.138**	**1.457**	**6,3%**

Fuente y elaboración propias.

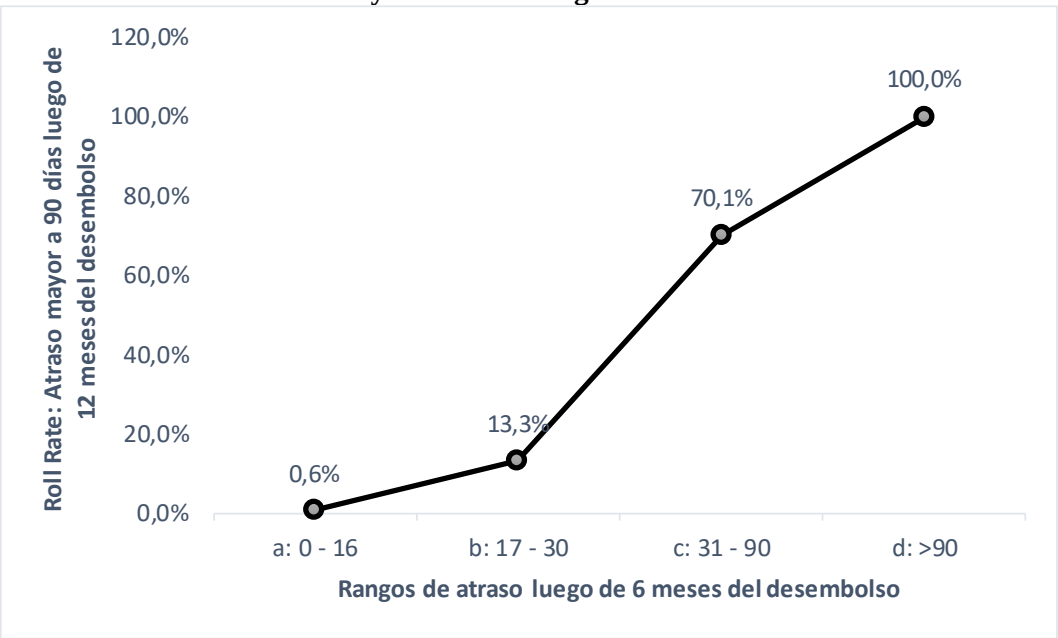
Según la definición de *roll rate*, es necesario analizar el atraso de los clientes en X número de meses y luego ver su transición transcurrido el mismo periodo. En este caso se considera el atraso máximo que tuvieron los clientes desembolsados en el periodo de modelización luego de seis meses del desembolso, al ser el tiempo intermedio en el que se analiza el comportamiento de una persona.

Para cada uno de los clientes, en los rangos de atraso analizados, se considera cuantos han pasado a un atraso máximo mayor a 90 días, luego de 12 meses de su desembolso.

Se establece 90 pues son los días de morosidad que considera la Superintendencia del Ecuador para catalogar a un crédito de Microcrédito en la categoría D (SBS 2016); categoría considerada como de dudoso recaudo; es decir, que es poco probable que salga de ese rango de morosidad.

Figura 13

Roll Rate: Atraso mayor a 90 días luego de 12 meses del desembolso



Fuente y elaboración propias

Para la categorización de los clientes en bueno, malo e indeterminado se considera los porcentajes de *roll rate*, en cada uno de los rangos de morosidad luego de los 6 meses del desembolso de la operación. Se puede observar, tanto en la tabla 4 como en la figura 13 que hay una fuerte diferenciación en los porcentajes de *roll rate* para los rangos de atraso a: 0-16 y c: 30-90, con 0,6% y 70,1%, respectivamente.

Por lo cual se define como cliente bueno a aquel con un atraso máximo, dentro del periodo de desempeño, menor a 16 días y cliente malo a aquel con un atraso mayor a 30 días. Para los indeterminados se considera los rangos de atraso entre 17 y 30 días que tiene un *roll rate* de 13,3%, que es un valor intermedio entre los valores concedidos a buen y mal cliente.

Es necesario recalcar, que las buenas prácticas establecen que el porcentaje de indeterminados dentro del periodo del modelo debería llegar a un máximo de 10% a 15% (Siddiqi 2006, 44). A continuación, se expone a modo resumen la definición de tipo cliente, aplicada en el periodo de modelo y testing.

Tabla 5
Definición del tipo de cliente

Y: Tipo Cliente	Atraso máximo luego de 12 meses del desembolso	Nro. Clientes	%Clientes
0: Malo	> 30	2.031	11,7%
1: Bueno	0 - 16	13.656	78,9%
2: Indeterminado	17 - 30	1.628	9,4%
Periodo modelo		**17.315**	**74,8%**
0: Malo	> 30	682	11,7%
1: Bueno	0 - 16	4.614	79,2%
2: Indeterminado	17 - 30	527	9,1%
Periodo testing		**5.823**	**25,2%**
Periodo modelización		**23.138**	**100,0%**

Fuente y elaboración propias

Podemos observar, en la tabla 5, que la proporción de buenos y malos para el periodo de modelo es de aproximadamente 6,7;[4] es decir, por cada cliente catalogado como malo, alrededor de 7 son catalogados como buenos. Esto se considera adecuado pues en la práctica no se hallará una relación homogénea de buenos y malos, es común que el porcentaje de malos sea pequeño pues de no darse el caso no existiera negocio crediticio sostenible en el tiempo.

Antes de pasar a la siguiente sección, se considera importante mencionar que para el análisis de las variables independientes y la construcción de los modelos scoring se considera únicamente a los clientes catalogados como buenos y malos, dejando de lado a los indeterminados, en el periodo de modelo. Esto con el objetivo de hacer una clara distinción entre un cliente de buen y mal comportamiento, de tal forma que el modelo *scoring* logre identificar de mejor forma a las variables que distingan entre los mismos; considerando así un total de 15.687 clientes, para el desarrollo de los modelos.

[4] El valor de 6,7 se obtiene dividiendo los 13.656 clientes catalogados como buenos, para los 2.031 clientes catalogados como malos, en el periodo de modelo.

4. Análisis de las variables independientes

Es importante mencionar que serán denominadas como variables independientes a aquellas características del cliente que ayudarán a discriminarle como bueno o malo; tales características pueden venir de tres distintas fuentes de información, las cuales serán expuestas a continuación (Bambino Contreras 2005, 29):

- Variables provenientes de la solicitud del crédito, catalogadas como demográficas.
- Variables provenientes de la central de riesgos o buró de crédito que denotan el comportamiento del cliente en un entorno externo a la institución.
- Variables de comportamiento interno del cliente, propias de la operación desembolsada.

A partir de este punto se denomina a la variable de tipo de cliente, definida en la sección 3, como variable dependiente, la cual será explicada por el conjunto de variables independientes.

Existen dos tipos de variables a analizar:

- **Cualitativas:** Son aquellas no numéricas, que constan de distintas categorías o atributos como el género, tipo de vivienda, actividad económica, etc.
- **Cuantitativas:** Son aquellas variables numéricas como el ingreso, edad, número de cargas familiares, etc.

Es necesario recordar que tanto para el modelo de regresión logística como de redes neuronales la variable dependiente es categórica y dicotómica. Además, se incluirá en los dos modelos a desarrollar, las variables, ya sean cualitativas o cuantitativas, una vez se las haya categorizado y dicotomizado[5], pues si bien el modelo de redes neuronales no calcula para cada variable independiente un coeficiente de regresión, como bien lo hace el de regresión logística; en el modelo logit se cuenta con el concepto de *odds ratio*, que representa una medida fundamental que evalúa la fuerza de asociación entre dos variables categóricas dicotómicas (Canela y Monge 2007, 5), alcanzando una mejor interpretación de las variables que se incluyen al modelo; pues toman el valor de 1 cuando se cumple la característica deseada y se considera al complemento como la categoría de referencia, logrando un mejor análisis de la variable.

[5] El término dicotomizar se refiere a convertir cada categoría de la variable independiente en análisis en una nueva variable que resultará dicotómica.

Para realizar la categorización, se hará uso de árboles de decisión, de forma que se logre la partición de la variable que mejor explique la variable dependiente Y.

Esta sección se divide en dos partes, en primer lugar, se hará un análisis exploratorio de las variables y se procederá a categorizar (empleando árboles de decisión) a aquellas que resulten adecuadas luego de este análisis. Como segunda parte, una vez categorizadas las variables se analizará su relación con la variable dependiente.

4.1 Análisis exploratorio de las variables

Se cuenta con un total de 79 variables construidas a partir de la información interna de la institución y del buró del cliente, cuyo listado y descripción se encuentran en el anexo 1. En el análisis exploratorio se busca identificar y cuantificar, lo siguientes puntos:

- **Valores perdidos:** Se considera como valores perdidos a todos aquellos valores, que por alguna razón desconocida no se cuenta dentro de la variable independiente. Si el porcentaje de valores perdidos para una variable, supera el 5%, no se la considerará dentro del grupo de variables independientes a incluirse en los modelos.
- **Inconsistencia en las variables:** Se busca que no existan valores que no concuerden con la naturaleza de la variable, por ejemplo, que dentro de la variable edad existan valores negativos. Además, se comprueba que las variables dentro de un determinado dominio no salgan del rango de valores permitido, por ejemplo, variables de tipo porcentual, deben estar dentro del rango de 0 a 1.

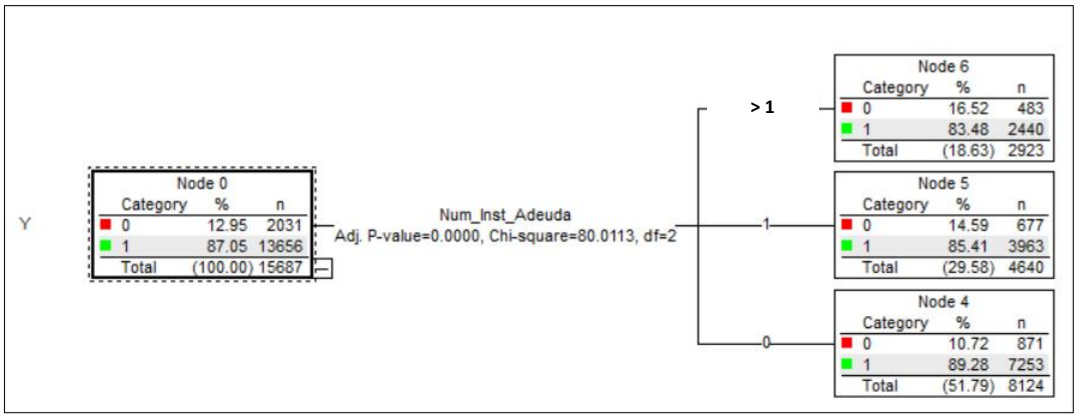
4.2 Categorización de las variables

Como ya se mencionó, se categorizó las variables a partir de árboles de decisión, tanto a variables cualitativas como cuantitativas, buscando crear grupos con al menos el 5% del total de registros. Una regla comúnmente usada para el número de observaciones en cada nodo, es considerar del 5% al 10% del total de la muestra (Weldon 2011, 2).

A modo de ejemplo, se muestra el árbol generado para la variable correspondiente al número de instituciones de central de riesgo, en las que el cliente tiene deuda al momento del desembolso, que denominaremos Num_Inst_Adeuda, para poder tratarla con facilidad. El resto de los árboles pueden ser observados en el anexo 2, los cuales son

generados haciendo uso del programa estadístico SPSS AnswerTree versión 3.0 y del método CHAID[6] para la categorización de la variable.

Figura 14
Árbol de decisión para la variable Num_Inst_Adeuda



Fuente y elaboración propias

Como se puede observar, en la figura 14, el árbol establece tres categorías para la variable Num_Inst_Adeuda, las cuales permiten explicar la variable dependiente Y, pues se rechaza la hipótesis nula de independencia entre estas dos variables, a un nivel de confianza del 95% (p-valor menor a 0,05).

Una vez categorizadas las variables y luego de realizar el análisis exploratorio de las mismas, se realiza el análisis de la relación de las variables independientes categorizadas con la variable dependiente Y, y su estructura. Para establecer variables que aporten al modelo y que logren su generalización y transferencia.

4.3 Elección de variables predictoras

Para la elección de las variables independientes que serán usadas en el desarrollo de los modelos se hace uso del estadístico IV, el cual fue construido en el programa R. A modo de ejemplo, se expone el valor de información (IV) para la variable Num_Inst_Adeuda (ver tabla 6), con la categorización obtenida por el árbol de decisión.

Como se puede observar, el valor IV para esta variable es superior a 0,02 e inferior a 0,1 (0,045), por lo cual entra en la categoría de predicción débil de la variable dependiente Y. En los modelos se incluirá aquellas variables que superen el valor de 0,02, es decir, que su nivel de predicción sea al menos débil.

[6] El método CHAID (del inglés Chi-squared Automatic Interaction Detection) usa el estadístico Chi - cuadrado para segmentar la variable, de tal forma que se logre el mayor grado de dependencia con la variable dependiente Y.

Tabla 6
IV – Variable Num_Inst_Adeuda

Nueva categorización obtenida por el árbol	**Significado de la categoría a partir de la variable original Num_Inst_Adeuda**	**Número de clientes catalogados como malos (*bi*)**	**Número de clientes catalogados como buenos (*mi*)**	$IV = \sum_{i=1}^{n} \left(\frac{b_i}{b} - \frac{m_i}{m}\right) \times \ln\left(\frac{b_i/b}{m_i/m}\right)$
0	0	871	7.253	2,2%
1	1	677	3.963	0,6%
2	>1	483	2.440	1,7%
Total		2.031	13.656	**4,5%**

Fuente y elaboración propias

Como ya se mencionó, inicialmente se cuenta con un total de 79 variables, sin embargo, luego de realizar el análisis exploratorio, considerando inconsistencias y tamaño de perdidos o nulos, se termina con 56 variables que superan este análisis.

El resumen de la prueba IV para un total de 56 variables resultantes de haber realizado el análisis exploratorio es expuesto en el anexo 2. De estas, 40 superan el valor del 2% requerido para el estadístico IV. Es decir, 40 variables independientes tienen al menos un nivel de predicción débil.

De las 40 variables se obtiene un total de 64 categorías a ser consideradas como variables independientes, por las razones expuestas al inicio de la sección, sin embargo, en los modelos se incluyen N-1 categorías por cada variable,[7] al ser considerada una como referencia; trabajando con un conjunto de 53 variables dicotómicas para cada modelo.

[7] N: número de categorías en una variable.

Capítulo cuarto: Construcción y contraste de modelos

En el presente capítulo, se busca completar las dos etapas finales en el proceso de modelización, resumido en la figura 9, luego de seguir el proceso descrito en el mismo esquema.

- Construcción de los modelos
- Evaluación estadística de los modelos

Los dos modelos por desarrollar, se los implementa en el software libre R, haciendo uso de los paquetes *glm* y *neuralnet*, para los modelos de regresión logística y redes neuronales, respectivamente.

Una vez elegidas las variables independientes, que superan el análisis exploratorio y el estadístico IV, se hace uso de las metodologías de regresión logística y redes neuronales, con el fin de obtener dos modelos *scoring* que permitan cumplir los objetivos planteados en el proyecto.

Cabe recalcar que las variables independientes resultaron de categorizar un grupo de variables continuas y categóricas, a partir de árboles de decisión, considerando cada categoría como una variable dicotómica.

1. Modelo de Regresión Logística

Para obtener el mejor modelo, haciendo uso de la metodología logit, se emplea la función *step*, en R, que selecciona el modelo a partir del criterio de información de Akaike (AIC por sus siglas en inglés), haciendo uso de la eliminación hacia atrás de las variables (Backward Stepwise).

A continuación, se expone el modelo de regresión logística resultante y se explica las características que se busca cumpla el mismo:

1. Las variables que se incluyan sean significativas, con un nivel de confianza de al menos el 95%.
2. Consistencia de signos en los coeficientes de las variables.
3. Se cumpla la prueba de multicolinealidad.

Tabla 7
Modelo Scoring Originación – Metodología logit

Nombre Variable	Definición	Categoría considerada a partir del árbol	Coeficiente (B)	*Odds Ratio*	p – valor	GVIF
Constante			-0,38[8]		7,6%	
V1_Num_Inst_Adeuda_1	Indica el número de instituciones en las que el cliente tiene deuda en la central de Riesgos, al desembolso.	1 institución con deuda mayor a 0	0,43	1,53	0,0%	1,52
V9_NumVecesMora12	Número de veces que el sujeto ha caído en mora en los últimos 12 meses.	> 0	-0,39	0,68	0,1%	1,61
V27_CR_3	(Mayor valor de deuda vencida en los U36M + Mayor valor de deuda NDI en U36M) / Mayor deuda histórica U36M	> 0,159127	-0,43	0,65	0,0%	2,10
V32_TieneConsumo	Tiene créditos de consumo al desembolso.	1 (Tiene)	0,74	2,09	0,0%	1,24
V38_MayorDeudaHistorica_Comercial	Mayor deuda total de créditos comerciales últimos 36 meses.	> 0	-0,31	0,74	0,8%	1,05
V52_Por_End_U12M	Deuda promedio últimos 12 meses / Ingreso.	> 0,585435	-0,41	0,66	0,1%	5,25
V53_Por_End_Antes_U12M	Deuda promedio mes 13 a 36 meses / Ingreso.	> 0,039093	-0,49	0,61	0,0%	1,85
V54_EndeudamientoPromUltimo12M_SF	Deuda promedio últimos 12 meses.	> 444,20749	-0,34	0,72	0,6%	5,12
V55_EndeudamientoPromAntes12M_SF_0	Deuda promedio mes 13 a 36 meses.	> 3.158,7966	0,46	1,58	0,0%	4,23
V55_EndeudamientoPromAntes12M_SF_1		(182,0266 – 3.158,7966]	0,27	1,30	1,6%	4,17
V56_ValorTotalVencido	Suma de todos los valores en deuda Vencido (No castigos) del cliente a la fecha de corte de los últimos 36 meses.	> 135,2299	-0,31	0,73	0,6%	2,16
V63_Genero	Género del cliente.	Masculino	-0,33	0,72	0,0%	1,14
V64_Edad_0	Edad del cliente a la fecha de desembolso.	> 34	0,20	1,23	0,2%	1,11

[8] Al trabajar con muestras desproporcionadas es necesario realizar una corrección en el valor de la constante, la cual es explicada a detalle en (Maddala 1992, 330), que consiste en restar a la constante el término $ln(p_1) - ln(p_2)$, donde p_1 y p_2 son las proporciones de las observaciones elegidas de los grupos de buenos y malos, respectivamente.

V65_TipoVivienda_0	Tipo de vivienda.	Propia hipotecada - Propia no hipotecada	0,80	2,22	0,0%	2,69
V65_TipoVivienda_1		Prestada – Vive con familiares	0,24	1,27	0,7%	2,53
V66_EstadoCivil_0	Estado Civil.	Divorciado – Unión Libre – Viudo	0,46	1,59	0,0%	1,43
V66_EstadoCivil_1		Casado	0,38	1,46	0,0%	1,54
V67_NivelEstudios_0	Nivel de Estudios.	Postgrado – Universidad	0,48	1,61	0,0%	1,08
V67_NivelEstudios_1		Primaria – Sin estudios	0,43	1,54	0,0%	1,13
V68_Sector	Sector de vivienda.	Zona Rural	-0,26	0,77	0,0%	1,13
V72_LineaNegocio_0	Línea de negocio.	Riesgo bajo	1,07	2,91	0,0%	2,50
V72_LineaNegocio_1		Riesgo medio	0,97	2,65	0,0%	2,40
V72_LineaNegocio_2		Riesgo medio – alto	0,46	1,59	0,0%	2,11
V73_ActividadEconomica_0	Actividad económica.	Riesgo bajo	0,49	1,63	0,0%	1,23
V74_SectorEconomico	Sector económico al que pertenece la actividad económica.	Comercio – Servicios – Personal	0,19	1,22	0,8%	1,38
V79_CiudadDomicilio_0	Ciudad de domicilio.	Riesgo bajo	2,77	15,93	0,0%	1,07
V79_CiudadDomicilio_1		Riesgo medio	0,97	2,65	0,0%	1,11
V84_DestinoOperacion	Destino económico de la operación desembolsada.	Activos fijos intangibles – Mejoramiento otros activos – No productivas (Para operaciones de consumo y vivienda)	-1,76	0,17	0,0%	1,20

Fuente y elaboración propias

1.1 Significancia de las variables independientes

Para corroborar la significancia de los coeficientes del modelo, se busca contrastar la siguiente hipótesis nula:

$$H_0: \beta_i = 0$$

Considerando a β_i como el coeficiente estimado correspondiente a la variable independiente x_i. Se trata de contrastar la hipótesis nula que la variable x_i es independiente de la variable dependiente (Salas 1996, 204), a partir del estadístico de Wald.

En la tabla 7 se muestra el p-valor para la prueba de Wald, cuyo valor es menor a 0,05, para todas las variables y menor a 0,08 para la constante, lo cual implica que los coeficientes de cada variable, que ingresa al modelo, son significativos a un nivel de confianza del 95%.

1.2 Consistencia de signos y coeficientes de las variables del modelo

En este apartado se busca confirmar que las estimaciones de los parámetros presentan coherencia en la interpretación de signos. De acuerdo a la definición de la variable dependiente, donde 1 define a un cliente bueno y 0 a uno malo, un signo negativo en un coeficiente implica que la variable castiga al cliente, aumentando el riesgo de default o no pago (ver tabla 7).

- V1_Num_Inst_Adeuda_1: Variable explicativa, que toma el valor de 1, cuando el número de instituciones en las que el cliente tiene deuda en la central de riesgos, al desembolso es 1; caso contrario, 0. Su valor *odds ratio* es de 1,53, lo que indica que, si una persona tiene deuda en una sola institución, al desembolso, es 53% mejor[9] que si no.
- V9_NumVecesMora12: Variable explicativa, que toma el valor de 1, cuando el número de veces que el sujeto ha caído en mora en los últimos 12 meses es mayor a 0; caso contrario 0. Su valor *odds ratio* es de 0,68, lo que indica que, si una persona ha caído en mora en los últimos 12 meses es 32% peor que si no.
- V27_CR_3: Variable construida, cuyo valor *odds ratio* es de 0,65, lo que indica que, si la mayor deuda vencida junto con su mayor deuda que no devenga intereses en los últimos 36 meses ha superado el 15,91% de su máxima deuda en este mismo

[9] El término mejor o peor es referido a que el cliente tendrá un mejor o peor comportamiento, va empatado con la definición de buen o mal cliente.

periodo, entonces el comportamiento de la persona es 35% peor que si no se cumple esta característica.

- V32_TieneConsumo: Variable explicativa, que toma el valor de 1, si la persona tiene una operación de consumo al desembolso. Su valor *odds ratio* es de 2,09, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 2,09 veces o 109% mejor que si no cumple la misma.
- V38_MayorDeudaHistorica_Comercial: Variable explicativa, que toma el valor de 1 cuando la mayor deuda total de créditos comerciales, en los últimos 36 meses es mayor a 0; caso contrario, 0. Su valor *odds ratio* es de 0,74, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 26% peor que si no se cumple la misma.
- V52_Por_End_U12M: Variable explicativa, que toma el valor de 1 cuando la relación entre la deuda promedio en los últimos 12 meses y el ingreso del cliente es mayor a 0,5854435; caso contrario 0. Su valor *odds ratio* es de 0,66, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 34% peor que si no se cumple la misma.
- V53_Por_End_Antes_U12M: Variable explicativa, que toma el valor de 1 cuando la relación entre la deuda promedio entre los meses 13 y 36 anteriores al desembolso y el ingreso es mayor a 0,039093; caso contrario 0. Su valor *odds ratio* es de 0,61, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 39% peor que si no se cumple la misma.
- V54_EndeudamientoPromUltimo12M_SF: Variable explicativa, que toma el valor de 1 si la deuda promedio en los últimos 12 meses, en el sistema financiero es mayor a 444,20749 dólares; caso contrario 0. Su valor *odds ratio* es de 0,72, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 28% peor que si no se cumple la misma.
- V55_EndeudamientoPromAntes12M_SF_0: Variable explicativa, que toma el valor de 1 si la deuda promedio en los meses 13 y 36 anteriores al desembolso es mayor a 3.158,7966; caso contrario 0. Su valor *odds ratio* es de 1,58, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 58% mejor que si no se cumple la misma.
- V55_EndeudamientoPromAntes12M_SF_1: Variable explicativa, que toma el valor de 1 si la deuda promedio en los meses 13 y 36 anteriores al desembolso se encuentra en el intervalo (182,0266 – 3.158,7966]; caso contrario 0. Su valor *odds*

ratio es de 1,30, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 30% mejor que si no se cumple la misma.

- V56_ValorTotalVencido: Variable explicativa, que toma el valor de 1 si la suma de todos los valores vencidos del cliente en los últimos 36 meses anteriores al desembolso es mayor a 135,2299 dólares; caso contrario 0. Su valor *odds ratio* es de 0,73, lo que indica que, si se cumple esta característica, el comportamiento del cliente es 27% peor, que si no se cumpliera.
- V63_Genero: Variable explicativa, que toma el valor de 1 si el cliente es hombre. Su valor *odds ratio* es de 0,72, lo que significa que si el cliente es hombre tiene un comportamiento 28% peor que si no lo es.
- V64_Edad_0: Variable explicativa, que toma el valor de 1 si el cliente tiene una edad mayor a 34 años. Su valor *odds ratio* es de 1,23, por lo cual, si un cliente cumple esta característica, se espera un mejor comportamiento que alguien de una edad distinta, específicamente de un 23%.
- V65_TipoVivienda_0: Variable explicativa, que toma el valor de 1 si el cliente tiene un tipo de vivienda propia hipotecada o propia no hipotecada; caso contrario 0. Su valor *odds ratio* es de 2,22, por lo cual, si un cliente cumple esta característica es 2,22 veces mejor que alguien que no.
- V65_TipoVivienda_1: Variable explicativa, que toma el valor de 1 si el cliente tiene un tipo de vivienda prestada o vive con familiares; caso contrario 0. Su valor *odds ratio* es de 1,27, por lo cual, si un cliente cumple esta característica es 27% mejor que si no la cumple.
- V66_EstadoCivil_0: Variable explicativa, que toma el valor de 1 si el estado civil del cliente es divorciado, unión libre o viudo; caso contrario 0. Su valor *odds ratio* es de 1,59, por lo cual, si un cliente cumple esta característica es 59% mejor que si no la cumple.
- V66_EstadoCivil_1: Variable explicativa, que toma el valor de 1 si el estado civil del cliente es casado; caso contrario 0. Su valor *odds ratio* es de 1,46, por lo cual, si un cliente cumple esta característica es 46% mejor que si no la cumpliera.
- V67_NivelEstudios_0: Variable explicativa, que toma el valor de 1 si el nivel de estudios del cliente es de postgrado o universidad; caso contrario 0. Su valor *odds ratio* es de 1,61, por lo cual, si un cliente cumple esta característica es 61% mejor que si no la cumpliera.

- V67_NivelEstudios_1: Variables explicativa, que toma el valor de 1 si el nivel de estudios del cliente es de primaria o sin estudios; caso contrario 0. Su valor *odds ratio* es de 1,54, por lo cual, si un cliente cumple esta característica es 54% mejor que si no la cumple.
- V68_Sector: Variable explicativa, que toma el valor de 1 si el sector de vivienda del cliente es en una zona rural; caso contrario 0. Su valor *odds ratio* es de 0,77, por lo cual, si un cliente cumple esta característica es 23% peor, que si no la cumpliera.
- V74_SectorEconomico: Variable explicativa, que toma el valor de 1 si el sector económico al que pertenece la actividad del cliente es de comercio, servicios o personal; caso contrario 0. Su valor *odds ratio* es de 1,22, por lo cual, si un cliente cumple esta característica es 22% mejor que si no la cumpliera.
- V84_DestinoOperacion: Variable explicativa, que toma el valor de 1 si el destino económico de la operación desembolsada pertenece al grupo de riesgo bajo; caso contrario 0. Su valor *odds ratio* es de 1,20, por lo cual, si un cliente cumple esta característica es 20% mejor que si no la cumpliera.
- V72_LineaNegocio_0: Su valor *odds ratio* es de 2,91, por lo cual, si la línea de negocio del cliente pertenece al grupo de riesgo bajo, este será 2,91 veces mejor, que si no pertenece.
- V72_LineaNegocio_1: Su valor *odds ratio* es de 2,65, por lo cual, si la línea de negocio del cliente pertenece al grupo de riesgo medio, este será 2,65 veces mejor, que si no pertenece.
- V72_LineaNegocio_2: Su valor *odds ratio* es de 1,59, por lo cual, si la línea de negocio del cliente pertenece al grupo de riesgo medio alto, este será 1,59 veces mejor, que si no pertenece.
- V73_ActividadEconomica_0: Su valor *odds ratio* es de 1,63, por lo cual, si la actividad económica del cliente pertenece al grupo de riesgo bajo, este será 63% mejor, que si no perteneciese.
- V79_CiudadDomicilio_0: Su valor *odds rati*o es de 15,93, por lo cual, si la ciudad de domicilio del cliente pertenece al grupo de ciudades de riesgo bajo, este será 15.93 veces mejor, que si no.
- V79_CiudadDomicilio_1: Su valor *odds ratio* es de 2,65, por lo cual, si la ciudad de domicilio del cliente pertenece al grupo de ciudades de riesgo medio, este será 2,65 veces mejor, que si no.

Se ha descrito e interpretado, a partir del valor *odds ratio*, cada una de las variables que forman parte del modelo de regresión logística, teniendo como resultado un sentido económico en cada una de ellas, esto a partir de su coeficiente β.

1.3 Resultados del factor de inflación generalizado de la varianza

Cuando las variables que se incluyen en el modelo no son independientes entre sí, provoca el efecto de incrementar la varianza de los coeficientes, esto genera (O'Brien 2007, 673):

- Estimaciones de parámetros con signo incorrecto.
- Coeficientes con magnitud inverosímil.
- Crea situaciones donde pequeños cambios en los datos producen grandes variaciones en la estimación de los parámetros.
- Y en casos más extremos evita la solución numérica de un modelo.

Como se puede observar, en la tabla 7, los valores GVIF para cada una de las variables independientes del modelo logit, son menores a 10. De esta forma, se establece que el modelo no tiene problemas de correlación entre las variables.

Los valores GVIF fueron obtenidos con la función *vif* de la librería *car*, en el programa estadístico R.

1.4 Estadísticos del modelo de regresión logística

Para medir la capacidad predictiva del modelo se calcula los estadísticos de la tabla 8. Se puede observar que para el modelo y testing el estadístico KS se encuentra en el rango de *satisfactorio* de acuerdo con los estándares empleados en TransUnion. De igual forma, el coeficiente de Gini es más que satisfactorio al alcanzar un valor superior al 50%.

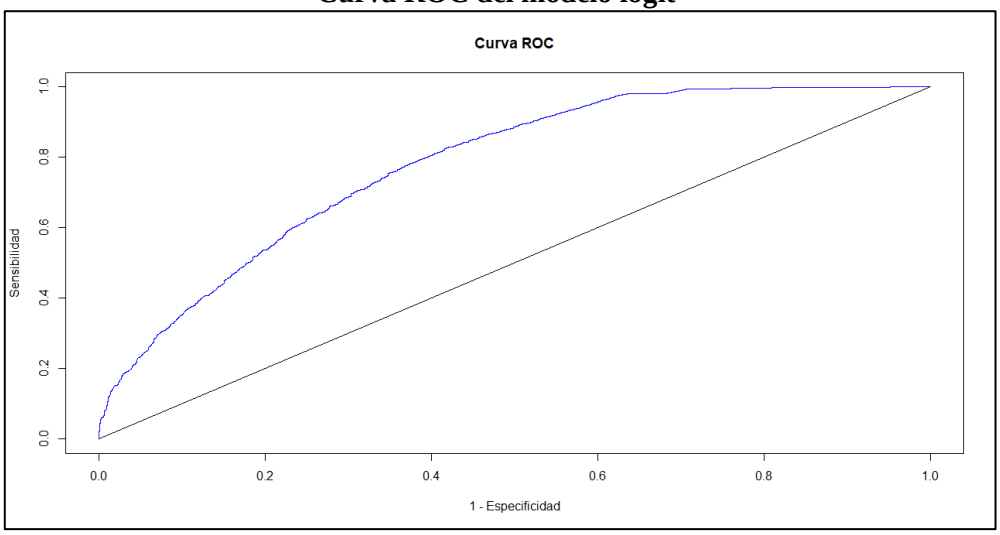
Tabla 8
Estadísticos del modelo logit

Estadístico	Valor periodo modelo	Valor periodo testing
KS	40,94%	39,92%
ROC	77,86%	76,85%
Coeficiente de Gini	55,72%	53,7%
Aciertos	75,71%	75,87%

Fuente y elaboración propias

La curva ROC para el modelo de regresión logística es expuesta en la figura 15 y como se puede observar (tabla 8) el estadístico AUROC es superior a 70% que es considerado adecuado según (Siddiqi 2006, 124).

Figura 15
Curva ROC del modelo logit



Fuente y elaboración propias

El porcentaje de aciertos se mantiene por encima del 75% tanto para modelo y testing, a partir de la matriz de confusión, en la cual se hizo uso del índice de Youden para obtener el punto de corte óptimo del modelo, cuyo valor es de 0,8680071.

A partir de los resultados expuestos, se comprueba que el modelo logit tiene un buen nivel predictivo, discriminando adecuadamente a buenos y malos clientes. Además, los valores para cada estadístico se mantienen en un rango adecuado para el periodo de testing, logrando un modelo con una fortaleza y eficiencia adecuada.

2. Modelo de Redes Neuronales

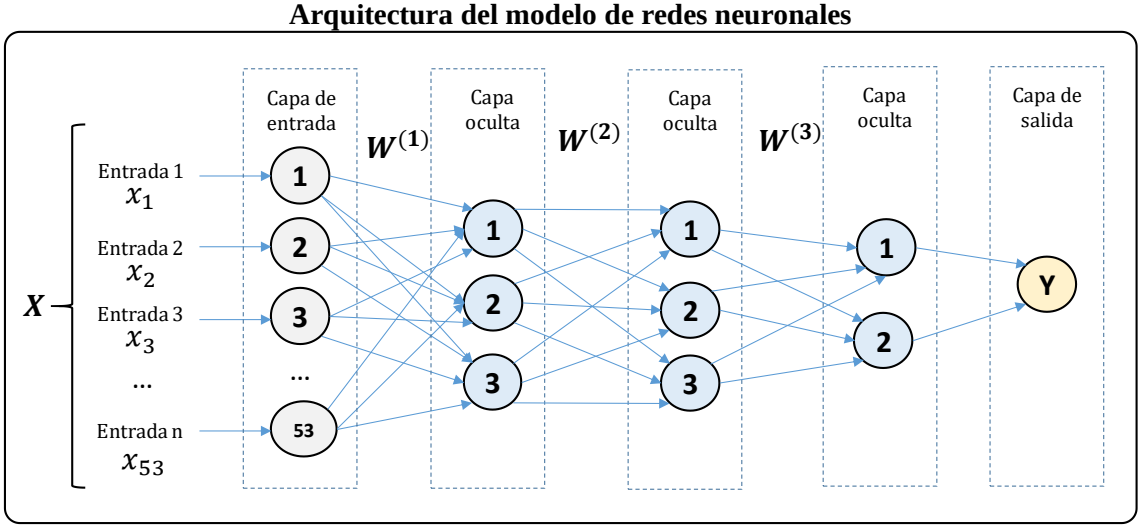
Como ya se mencionó, en el capítulo dos, existen varios elementos que conforman un modelo de redes neuronales; a continuación, se expone los que forman parte del modelo obtenido en este proyecto.

Niveles o capas de neuronas: Luego de realizar varias pruebas se decide que la mejor estructura para el modelo que se desea construir es la conformada por una capa de entrada, tres capas ocultas y una capa de salida. La capa de entrada le corresponden un total de 53 neuronas, una por cada variable a incluir en el modelo, las capas ocultas tienen tres, tres y dos neuronas cada una, tal como se muestra en la figura 16. Y la capa de salida

está formada por una sola neurona, que representa la variable dependiente que deseamos pronosticar.

Figura 16

Arquitectura del modelo de redes neuronales



Fuente y elaboración propias

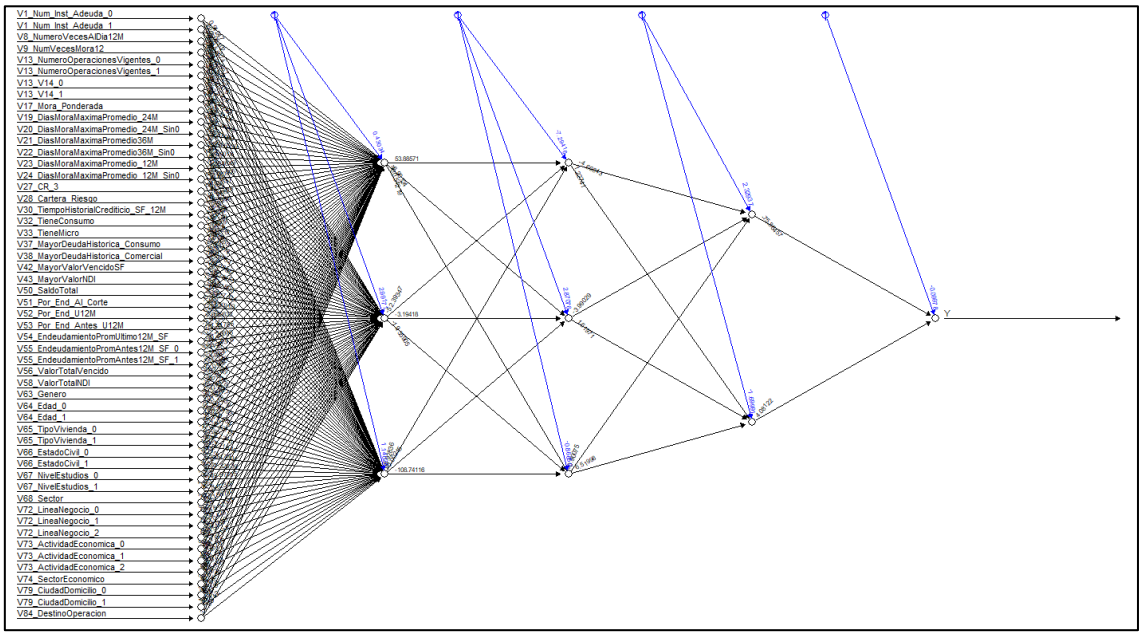
Función de agregación y activación: La función de agregación corresponde a una función lineal que relaciona las entradas y los pesos sinápticos. Como función de activación se hace uso de una función sigmoidea, pues tiene una salida en el intervalo $[0, 1]$, que es lo deseado al querer obtener un valor de probabilidad, que debe encontrarse entre estos valores.

Mecanismo de aprendizaje: En este caso se trabaja con un mecanismo de aprendizaje supervisado, pues se conoce la respuesta que debería generar la red a partir del conjunto de variables en la capa de entrada. El método que se utilizó para encontrar los pesos de la red neuronal es el algoritmo RPROP+ (*resilient backpropagation with weight backtracking*).

El modelo es obtenido en el programa R, a partir de la función *neuralnet*, cuya estructura es expuesta a continuación.

Figura 17

Red Neuronal



Fuente y elaboración propias

En la tabla 9, tabla 10, tabla 11 y tabla 12 se resumen los coeficientes resultantes de cada capa y nodo en la misma.

Tabla 9

Coeficientes – Capa 1

Nodos - Capa de Entrada	**Nodo 1 - Capa 1**	**Nodo 2 - Capa 1**	**Nodo 3 - Capa 1**
Constante	0,4303	2,9917	1,1403
V1_Num_Inst_Adeuda_0	0,9425	1,2254	-0,0899
V1_Num_Inst_Adeuda_1	2,0540	-1,0566	-0,2653
V8_NumeroVecesAlDia12M	0,5837	0,3666	-0,6746
V9_NumVecesMora12	-1,9287	-0,6158	-1,3929
V13_NumeroOperacionesVigentes_0	-0,8352	1,1961	-0,5520
V13_NumeroOperacionesVigentes_1	0,1038	-0,9343	1,0589
V13_V14_0	-2,5102	-0,5159	2,4065
V13_V14_1	-0,9586	1,1093	0,0933
V17_Mora_Ponderada	-2,4714	1,3781	0,1167
V19_DiasMoraMaximaPromedio_24M	-0,3097	1,2475	-0,0214
V20_DiasMoraMaximaPromedio_24M_Sin0	3,3197	-1,3156	3,8435
V21_DiasMoraMaximaPromedio36M	-2,4656	0,5835	0,4016
V22_DiasMoraMaximaPromedio36M_Sin0	-4,9016	-0,8109	-1,9813
V23_DiasMoraMaximaPromedio_12M	-2,0465	-0,4088	0,3558
V24_DiasMoraMaximaPromedio_12M_Sin0	-1,0091	1,2595	0,1592
V27_CR_3	7,3334	2,0589	1,0779
V28_Cartera_Riesgo	-2,1783	-0,9140	-2,1643
V30_TiempoHistorialCrediticio_SF_12M	-1,3926	-0,4416	1,5325
V32_TieneConsumo	10,3403	-0,5908	49,7746

V33_TieneMicro	-2,9547	0,4241	0,8881
V37_MayorDeudaHistorica_Consumo	-1,4092	0,0066	0,4079
V38_MayorDeudaHistorica_Comercial	-2,2032	-0,0855	-0,0613
V42_MayorValorVencidoSF	-2,2381	-0,2825	0,8920
V43_MayorValorNDI	-0,6894	0,5280	2,9304
V50_SaldoTotal	1,4334	-1,2742	-0,0061
V51_Por_End_Al_Corte	-0,9576	3,3026	1,0238
V52_Por_End_U12M	0,3503	0,4083	-1,7130
V53_Por_End_Antes_U12M	-0,4412	1,1179	-1,4062
V54_EndeudamientoPromUltimo12M_SF	1,8278	0,2570	-2,2402
V55_EndeudamientoPromAntes12M_SF_0	4,1971	-1,7438	-1,1732
V55_EndeudamientoPromAntes12M_SF_1	1,7131	-1,6077	-0,7908
V56_ValorTotalVencido	0,2783	1,0587	0,3167
V58_ValorTotalNDI	1,3327	-0,3263	-4,1936
V63_Genero	-0,4626	0,1244	-0,3050
V64_Edad_0	1,6680	-0,7855	0,3088
V64_Edad_1	1,2582	0,0962	-0,5318
V65_TipoVivienda_0	2,3063	-1,0243	-0,7346
V65_TipoVivienda_1	1,0711	-0,3959	-1,7184
V66_EstadoCivil_0	19,4880	-0,9262	-3,4790
V66_EstadoCivil_1	2,2459	-1,3316	-1,3264
V67_NivelEstudios_0	1,0410	-0,1705	49,9774
V67_NivelEstudios_1	1,5501	-0,3066	3,0237
V68_Sector	3,8998	0,1128	-1,5870
V72_LineaNegocio_0	0,2757	-0,6849	48,5207
V72_LineaNegocio_1	-1,7578	-0,8049	1,0423
V72_LineaNegocio_2	-2,4262	0,6650	3,5469
V73_ActividadEconomica_0	9,8478	-0,0346	41,6214
V73_ActividadEconomica_1	-3,7681	0,4962	3,4627
V73_ActividadEconomica_2	-2,9942	-0,3256	-1,0250
V74_SectorEconomico	3,0466	-0,2867	-0,0071
V79_CiudadDomicilio_0	16,4786	-4,1225	-4,6857
V79_CiudadDomicilio_1	1,4984	-0,7057	1,5477
V84_DestinoOperacion	-11,6857	-0,5582	-1,9074

Fuente y elaboración propias

Tabla 10

Coeficientes – Capa 2

	Nodo 1 – Capa 2	**Nodo 2 – Capa 2**	**Nodo 3 – Capa 2**
Constante	-7,294159	2,87075998	-0,86659
Nodo 1 – Capa 1	53,885709	-2,903236537	400,54219
Nodo 2 – Capa 1	-52,393466	-3,194184085	-19,35905
Nodo 3 - Capa 1	68,928364	-0,002454327	-108,74116

Fuente y elaboración propias

Tabla 11
Coeficientes – Capa 3

	Nodo 1 - Capa 3	Nodo 2 - Capa 3
Constante	2,329373	-1,69991
Nodo 1 - Capa 2	-4,6684301	1,227409
Nodo 2 - Capa 2	-3,9902865	1,619711
Nodo 3 - Capa 2	-0,9007494	6,519976

Fuente y elaboración propiass

Tabla 12
Coeficientes – Capa 4

	Nodo - Capa de salida
Constante	-0,09978327
Nodo 1 - Capa 3	-75,86456849
Nodo 2 - Capa 3	4,08122061

Fuente y elaboración propiass

2.1 Estadísticos del modelo de redes neuronales

Para medir la capacidad predictiva del modelo se calcula los estadísticos de la tabla 13. Podemos observar que los valores del estadístico KS se encuentran por encima del 41% entrando en la categoría de *bueno* según los estándares considerados por TransUnion, para el periodo de modelo y testing. El coeficiente de Gini evidencia que el modelo de redes neuronales es más que satisfactorio al tomar valores por encima del 50%.

Tabla 13
Estadísticos del modelo redes neuronales

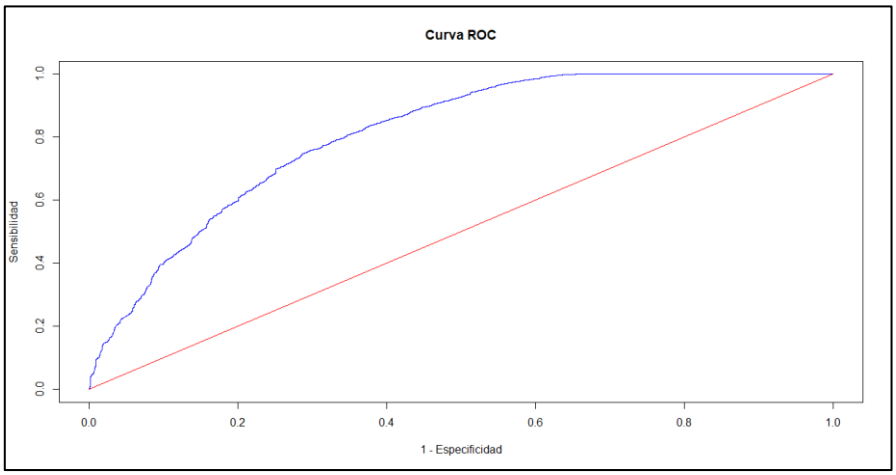
Estadístico	Valor periodo modelo	Valor periodo testing
KS	46,13%	42,43%
ROC	80,78%	77,76%
Coeficiente de Gini	61,56%	55,52%
Aciertos	74,58%	72,57%

Fuente y elaboración propias

A continuación se expone, en la figura 18, la curva ROC para el modelo de redes neuronales, como se puede observar el estadístico AUROC es superior al 80% para el periodo de modelo y supera el 77% para el testing; con lo cual se alcanza una correcta capacidad de discriminación entre buenos y malos clientes.

Figura 18

Curva ROC del modelo de redes neuronales



Fuente y elaboración propias

El porcentaje de aciertos se mantiene por encima del 70% tanto para modelo y testing, a partir de la matriz de confusión, en la cual se hizo uso del índice de Youden para obtener el punto de corte óptimo del modelo, cuyo valor es de 0,890068781.

A partir de los resultados expuestos, se comprueba que el modelo de redes neuronales tiene un buen nivel predictivo, discriminando adecuadamente a buenos y malos clientes. Además, los valores para cada estadístico se mantienen en un rango adecuado para el periodo de testing, logrando un modelo con una fortaleza y eficiencia adecuada.

3. Comparación Modelos Scoring

El objetivo principal de este proyecto es comparar dos modelos scoring, resultantes de hacer uso de las metodologías de regresión logística y redes neuronales. Para poder hacer esta comparación, se hace uso de los estadísticos, que se expone a continuación:

- Estadístico de Kolmogorov – Smirnov
- Coeficiente de GINI
- Matriz de confusión: error, sensibilidad y especificidad
- Estadístico AUROC
- Criterio de información de Akaike

Además, para poner a prueba el poder predictivo de los modelos, se emplea un periodo que no fue incluido en la construcción de los modelos y de esta forma, junto con los estadísticos propuestos, probar la hipótesis planteada.

Tabla 14

Comparación estadísticos modelos – Periodo modelo

Estadístico - Periodo Modelo	Regresión Logística	Redes Neuronales	¿Cuándo es mejor el estadístico?	¿Cuál modelo es mejor?
AUROC	77,86%	80,78%	Mayor valor	Redes neuronales
KS	40,94%	46,13%	Mayor valor	
Gini	55,72%	61,56%	Mayor valor	
Error Punto de Corte = 0.8680071	24,29%	20,86%	Menor valor	
Sensibilidad Punto de Corte = 0.8680071	77,58%	81,32%	Mayor valor	
Especificidad Punto de Corte = 0.8680071	63,12%	64,45%	Mayor valor	
Error Punto de Corte = 0.890068781	29,59%	25,42%	Menor valor	
Sensibilidad Punto de Corte = 0.890068781	70,68%	75,13%	Mayor valor	
Especificidad Punto de Corte = 0.890068781	68,59%	70,90%	Mayor valor	
Criterio de información de Akaike	9.541,776	1.512,553	Menor valor	

Fuente y elaboración propias

Se puede observar que,

- En el estadístico KS, el modelo de redes neuronales supera al de regresión logística en 5,19 puntos porcentuales y los valores se encuentran entre lo recomendado para un modelo scoring.
- Para el coeficiente de Gini, el modelo de redes neuronales supera al de regresión logística en 5,84 puntos porcentuales. Además, para los dos modelos el coeficiente supera el 50%, lo que muestra que ambos son satisfactorios.
- Los puntos de corte óptimos, calculados a partir del índice de Youden, para los modelos de regresión logística y redes neuronales son 0,8680071 y 0,890068781, respectivamente.

Para comparar los dos modelos se construye las matrices de confusión con los dos puntos de corte óptimos, hallados con el índice de Youden. Es decir, se tendrá cuatro matrices de confusión, dos por cada punto de corte aplicado en cada uno de los modelos construidos.

Tabla 15

Matrices de confusión: Punto de corte 0,8680071

Punto de Corte: 0,8680071 Modelo Logit		Real	
		0: Malo	1: Bueno
Estimado	0: Malo	1.282	3.062
	1: Bueno	749	10.594

Tasa de aciertos	75,71%
Error	24,29%
Especificidad	63,12%
Sensibilidad	77,58%

Punto de Corte: 0,8680071 Modelo Redes Neuronales		Real	
		0: Malo	1: Bueno
Estimado	0: Malo	1.309	2.551
	1: Bueno	722	11.105

Tasa de aciertos	79,14%
Error	20,86%
Especificidad	64,45%
Sensibilidad	81,32%

Fuente y elaboración propias

Tabla 16

Matrices de confusión: Punto de corte 0,890068781

Punto de Corte: 0,890068781 Modelo Logit		Real	
		0: Malo	1: Bueno
Estimado	0: Malo	1.393	4.004
	1: Bueno	638	9.652

Tasa de aciertos	70,41%
Error	29,59%
Especificidad	68,59%
Sensibilidad	70,68%

Punto de Corte: 0,890068781 Modelo Redes Neuronales		Real	
		0: Malo	1: Bueno
Estimado	0: Malo	1.440	3.396
	1: Bueno	591	10.260

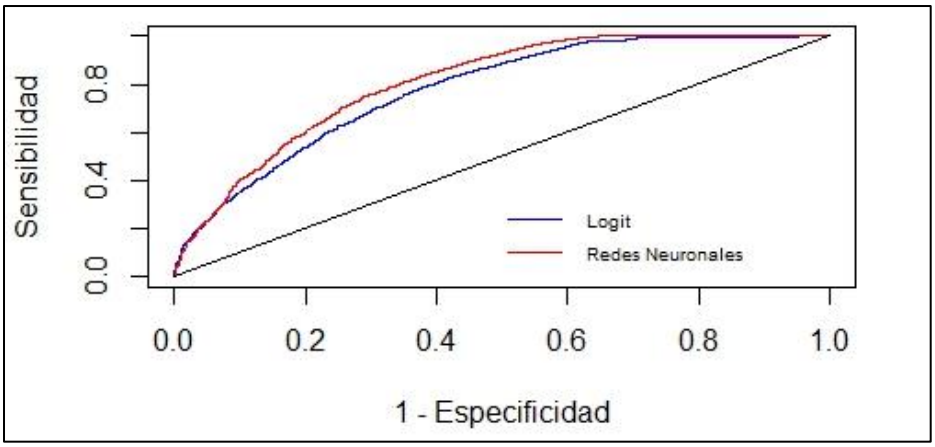
Tasa de aciertos	74,58%
Error	25,42%
Especificidad	70,90%
Sensibilidad	75,13%

Fuente y elaboración propias

Como se puede observar, en la tabla 15 y tabla 16, para los dos puntos de corte óptimos, el modelo de redes neuronales supera a los valores obtenidos para los cuatro indicadores de eficiencia.

- En la figura 19 se compara las curvas ROC obtenidas para los modelos de regresión logística (Roja) y redes neuronales (Azul). Podemos observar que la curva azul se encuentra por encima de la roja, por lo cual se evidencia que el modelo de redes neuronales es superior, respecto a su capacidad predictiva, frente al de regresión logística. El valor ROC para el modelo de redes neuronales es superior al de regresión logística en 2,92 puntos porcentuales.

Figura 19
Comparación curvas ROC



Fuente y elaboración propias

- El criterio de información de Akaike es usado para comparar modelos, penalizándolos según su desviación de los datos reales. El mejor modelo, según este criterio es el que presente el menor valor, siendo este el que mejor se ajusta a los datos (Sánchez 2012, 71). Se puede observar que el modelo de redes neuronales es mejor respecto al de regresión logística, por lo cual se puede establecer que este se ajusta de mejor forma a los datos.

3.1 Poder predictivo en el periodo de Testing

Para medir el poder predictivo de los modelos se hace uso del periodo de testing, que corresponde al 25,2% de la base (5.823 registros), los cuales están distribuidos de la siguiente forma:

Tabla 17
Periodo Testing

Categoría – Y	Número Clientes	%Clientes
0: Malos	682	11,71%
1: Buenos	4.614	79,24%
2: Indeterminados	527	9,05%

Fuente y elaboración propias

Los clientes que fueron catalogados como indeterminados son incluidos en el grupo de clientes buenos para poder aplicar los dos modelos *scoring*.

A continuación, se muestra los resultados obtenidos para los estadístico AUROC, KS, Coeficiente de Gini y Error para los dos puntos de corte óptimos obtenidos en la sección 2.1 del presente capítulo. Como se puede observar, para cada uno de los estadísticos mencionados, se obtiene mejores resultados al hacer uso de la metodología de redes neuronales, al compararlo con el modelo de regresión logística.

Tabla 18
Estadísticos - Periodo Testing

Estadístico	Regresión Logística	Redes Neuronales	Diferencia
AUROC	76,85%	77,76%	0,9%
KS	39,92%	42,43%	2,5%
Coeficiente de Gini	53,70%	55,52%	1,8%
Error Punto de Corte = 0.8680071	24,13%	21,72%	-2,4%
Error Punto de Corte = 0.8900687	29,59%	27,43%	-2,2%

Fuente y elaboración propias

Como se pudo notar, cada una de las pruebas estadísticas arroja un solo resultado. El hacer uso de la metodología de redes neuronales, proporciona un mejor modelo que se ajusta mejor a los datos, pronosticando y logrando un superior performance. Es por esto,

se puede concluir que la hipótesis planteada en este proyecto se considera como verdadera; es decir:

El hacer uso de redes neuronales, para la modelización del credit scoring de una cartera de microcrédito, logra un mejor performance que utilizar una metodología de regresión logística.

Capítulo quinto: Conclusiones y recomendaciones

1. Conclusiones

En el presente trabajo se busca contrastar la hipótesis: El hacer uso de redes neuronales, para la modelización del credit scoring de una cartera de microcrédito, logra un mejor performance que utilizar una metodología de regresión logística.

Para esto se construyó dos modelos scoring haciendo uso de dos metodologías distintas, regresión logística y redes neuronales, a partir de la información proporcionada por una institución financiera ecuatoriana.

La base de datos corresponde a la información al momento del desembolso, de créditos del producto microcrédito. Al contar con información al desembolso, los modelos scoring entran dentro de la categoría de originación, y se distingue tres grupos de datos: información demográfica, de central de riesgos e interna de la institución, del solicitante de crédito.

Para tener una base de datos sólida, previo a la construcción de los modelos se restringió la muestra bajo los siguientes criterios:

- Se consideró únicamente a aquellos clientes bancarizados, es decir, con al menos un registro en central de riesgos, en los últimos 36 meses.
- Se buscó que la información con la cual se construyó los modelos sea madura. Para esto se determinó que el periodo de desempeño en el cual se puede establecer que la cartera es madura es de 12 meses.
- El periodo de modelización también debe ser estable para lo cual se aplicó la prueba de Dickey – Fuller aumentada, comprobando que la serie de este periodo es estacionaria.
- Una vez determinado el periodo de modelización (enero 2014 a diciembre 2016), con un total de 23.138 registros y tres años de información, que es lo recomendado por la Superintendencia de Bancos y Seguros, se segmentó el periodo en lo que se denominó periodo de Modelo y Testing.

  El periodo de modelo que representó el 74,8% del total fue obtenido empleando técnicas de muestreo que aseguren que este sea representativo respecto al periodo de modelización.

El periodo de testing, que representó el 25,2% del total, fue usado en la validación de los modelos, pues se busca que los mismos funcionen en un periodo de información distinto al que se utilizó para su construcción.

Una vez establecido el periodo de análisis se estimaron variables independientes y una dependiente que aporten a los modelos para asegurar la funcionalidad de estos.

La variable dependiente fue construida con el máximo atraso del cliente en el periodo de 12 meses, luego de haber realizado el desembolso. Para asegurar el valor de información que aporta cada una de las variables independientes se hizo uso del estadístico IV; el cual permitió considerar aquellas variables que luego de ser categorizadas a partir de árboles de decisión tengan poder de predicción.

Una vez obtenido el periodo se procedió a aplicar las dos metodologías. Es importante aclarar que los dos modelos fueron desarrollados en el programa estadístico R, con las funciones *glm* y *neuralnet*.

Se logró el mejor modelo de regresión logística, usando la función *step* que considera el criterio de información de Akaike en la elección de este.

Para el modelo de redes neuronales se utilizó la función *neuralnet* con las siguientes características: [10]

- Niveles o capas de neuronas: Una capa de entrada con 53 nodos, tres capas ocultas con tres, tres y dos neuronas, y una capa de salida con una neurona.
- Función de agregación y activación: Corresponde a una función lineal y una sigmoidea, respectivamente.
- Mecanismo de aprendizaje: Aprendizaje supervisado con el algoritmo de RPROP+ para encontrar los pesos de la red neuronal.

Con los modelos construidos se procede a cotejarlos los resultados y capacidad predictiva con los estadísticos Kolmogorov – Smirnov (KS), Coeficiente de GINI, Matriz de confusión, AUROC y criterio de información de Akaike

Además de los estadísticos mencionados se aplica los modelos en el periodo de testing, el cual no fue empleado en la construcción de estos, y así establecer si los modelos son generalizables al funcionar en otro periodo de tiempo distinto al de su desarrollo. Para este periodo de tiempo también se emplea los estadísticos de KS, Gini, AUROC y Error.

[10] El tiempo que demoró el programa R, en obtener la red neuronal fue de 15 minutos, con un procesador Intel(R) Core(TM) i7-6500U y 16 GB RAM.

Se pudo observar, tanto en el periodo de modelo como testing, que se obtiene mejores resultados, en todos los casos, para el modelo de redes neuronales. Es por esto, se puede concluir que la hipótesis planteada en este trabajo es verdadera. Es decir, el hacer uso de redes neuronales, para la modelización del credit scoring de una cartera de microcrédito, logra un mejor performance que utilizar una metodología de regresión logística.

También se puede concluir que, cada modelo se debe adaptar o diferenciar de acuerdo con el tipo de información o segmento de mercado en el que será implementado, en este caso microcréditos. Se pudo observar que los modelos obtenidos tienen un fuerte poder de predicción, se ajustan correctamente a los datos con los que fueron construidos haciendo distinción entre buenos y malos clientes. Y pueden ser generalizables pues funcionan en un periodo distinto al de su desarrollo.

Respecto al tiempo de demora en obtener los modelos resultantes, las redes neuronales tomaron entre 10 a 15 minutos por cada prueba realizada respecto al número de capas y nodos que mejor se ajustara a la data en análisis, sin contar con el proceso de modelación descrito en el capítulo tercero. La función *step* obtuvo el mejor modelo de regresión logística en 12 minutos, sin embargo, luego se procedió a analizar cada variable obtenida para corroborar la consistencia en su signo, significancia de los coeficientes al 95% de confianza, presencia de multicolinealidad, hasta alcanzar un modelo que cumpliera con estos requisitos, tomando un tiempo de 45 minutos a 1 hora. En promedio, obtener el modelo final de redes neuronales tomó 40 minutos y el de regresión logística alrededor de 1 hora, evidenciando una ligera ganancia en el tiempo de modelación para la metodología de redes neuronales, al no ser necesario que cumpla algún supuesto.

2. Recomendaciones

Si bien los modelos de regresión logística son los comúnmente usados por su facilidad de interpretación e implementación, se tienen metodologías como redes neuronales, con la cual se logra mejores resultados, por lo cual se recomienda probar estas metodologías no tradicionales.

Si bien en el presente trabajo se construyen scores de originación, por el tipo de información que se dispone, se podría comparar las dos metodologías, regresión logística y redes neuronales, con otro tipo de scores, como de comportamiento.

Se puede probar distintos algoritmos para la obtención de los pesos o mecanismos de aprendizaje en la red neuronal, y comparar los distintos resultados obtenidos.

Bibliografía

Allison, Paul David. 2012. *Logistic regression using SAS: theory and application*. 2nd ed. Cary, N.C: SAS Pub.

Almeida, C, Carlton Baugh, C Lacey, C S. Frenk, G L. Granato, Laura Silva, y Alessandro Bressan. 2009. "Modelling the dusty universe I: Introducing the artificial neural network and first applications to luminosity and colour distributions". *Monthly Notices of the Royal Astronomical Society* 402 (junio). https://doi.org/10.1111/j.1365-2966.2009.15920.x.

Anderson, Raymond. 2007. *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press.

Arias, Xiomara. 2016. "Comportamiento crediticio sectorial". Subdirección de estadísticas y estudios: SBS.

Bahamón, Rodrigo. 2013. "MODELO PREDICTIVO NEURONAL PARA LA EVALUACIÓN DEL RIESGO CREDITICIO". Bogotá: Universidad Nacional de Colombia.

Bambino Contreras, Carlos. 2005. "Prestar como locos y obtener beneficios: ¿es realmente posible? (Un análisis logit multinomial para los determinantes del comportamiento de pago de una cartera de consumo)". Quito: FLACSO. http://repositorio.flacsoandes.edu.ec/handle/10469/61.

Bolivia, Superintendencia de Bancos y Entidades Financieras, Bolivia, e Intendencia de Estudios y Normas. 2008. *Guías para la gestión de riesgos*. La Paz, Bolivia: Superintendencia de Bancos y Entidades Financieras de Bolivia.

Bonilla, María, Ignacio Olmeda, y Rosa Puertas. 2003. "Modelos paramétricos y no paramétricos en problemas de credit scoring". *Revista Española de Financiación y Contabilidad* XXXII (118): 883–869.

Brío, Bonifacio Martín del, y Alfredo Sanz Molina. 2002. *Redes neuronales y sistemas difusos Bonifacio Martín del BRio, Alfredo Sanz Molina.* Bogotá (Colombia): Alfaomega Ra-Ma.

Cadarso Suárez, Carmen María. 2010. "Metodología ROC en la Evaluación de Medidas Antropométricas como Marcadores de la Hipertensión Arterial. Aplicación a Población Gallega Adulta". Máster en técnicas estadísticas, Santiago de Compostela: Universidad de Santiago de Compostela.

Canela, Mariano Aguayo, y Estrella Lora Monge. 2007. “Cómo hacer una regresión logística binaria ‘paso a paso’ (II): análisis multivariante”. Fundación Andaluza Beturia.

Carranza Bravo, Paola. 2010. “INTRODUCCIÓN A LAS TÉCNICAS DE INTELIGENCIA ARTIFICIAL APLICADAS A LA GESTIÓN FINANCIERA EMPRESARIAL”. *Fides et Ratio - Revista de Difusión cultural y científica de la Universidad La Salle en Bolivia* 4 (4): 8–15.

Croall, I. F., y J. P. Mason, eds. 1992. *Industrial applications of neural networks: project ANNIE handbook*. Research reports ESPRIT, vol. 1. Berlin; New York: Springer-Verlag.

Flórez, Orlando Moscote, y William Arley Rincón. 2002. “Modelo Logit y Probit: un caso de aplicación”, Comunicaciones en Estadística, 5 (diciembre): 123–33.

Fox, John. 2003. “Linear Models, Problems”. Canada: McMaster University.

González, José Ramón Hilera, y Víctor José Martínez Hernando. 2000. *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. México: Alfaomega: Ra-Ma.

Gujarati, Damodar N., Demetrio Garmendia Guerrero, y Gladys Arango Medina. 2005. *Econometría*. 4. ed. México: McGraw-Hill.

Gutierrez Girault, Matias Alfredo. 2007. “Modelos de Credit Scoring: Qué, Cómo, Cuándo y Para Qué”. MPRA Paper. octubre de 2007. https://mpra.ub.uni-muenchen.de/16377/.

Hand, D. J., y W. E. Henley. 1997. “Statistical Classification Methods in Consumer Credit Scoring: A Review”. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160 (3): 523–41.

Hand, David, y Robert Till. 2001. “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems | Request PDF”. *Machine Learning* 45: 171–86. https://doi.org/10.1023/A:1010920819831.

Igel, Christian, y Michael Hüsken. 2000. “Improving the Rprop Learning Algorithm”. En .

Islam, Samsul, Lin Zhou, y Fei Li. 2009. “Application of Artificial Intelligence (Artificial Neural Network) to Assess Credit Risk: A Predictive Model for Credit Card Scoring”. Msc, Blekinge Institute of Technology.

Jiménez-Caballero, José Luis, y Ramón Jesús Ruiz Martínez. 2000. “Las redes neuronales en su aplicación a las finanzas”. *Banca y finanzas: Revista profesional de gestión financiera*, núm. 54: 19–27.

Ladino, Becerra Iván Camilo. 2014. “Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales”. Bogotá: PONTIFICIA UNIVERSIDAD

JAVERIANA FACULTAD DE CIENCIAS ECONOMICAS Y ADMINISTRATIVAS MAESTRIA EN ECONOMÍA.

Maddala, G. S. 1992. *Introduction to econometrics*. 2nd ed. New York: Toronto: New York: Macmillan Pub. Co. ; Maxwell Macmillan Canada ; Maxwell Macmillan International.

Moral, Irene. 2006. "Modelos de regresión: lineal simple y regresión logística". *Revista Seden*, el 3 de diciembre de 2006.

O'Brien, Robert. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors". *Quality and Quality*, 673–90.

Paredes, Maria Cecilia, y Jorge Cóndor. 2018. "Reporte comportamiento crediticio sectorial". Subdirección de estadísticas y estudios: SBS.

Pitarque, Alfonso, Juan Carlos Ruiz, y Juan Francisco Roy. 2000. "Las redes neuronales como herramientas estadísticas no paramétricas de clasificación". *Psicothema* 12: 459–63.

Rezac, Martin, y František Řezáč. 2011. "How to Measure the Quality of Credit Scoring Models". *Czech Journal of Economics and Finance (Finance a uver)* 61 (enero): 486–507.

Rico, Claudia, Mayerly Paredes, y Nelson Fernández. 2009. "Modeling of the Hierarchical Structure of Freshwater Macroinvertebrates Using Artificial Neural Networks". *Acta Biológica Colombiana* 14 (3): 71–96.

Riedmiller, Martin, y Heinrich Braun. 1993. "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm". En, 1:586–91 vol.1. https://doi.org/10.1109/ICNN.1993.298623.

Salas, Manuel. 1996. "La regresión logística. Una aplicación a la demanda de estudios universitarios". *Estadística Española*, 1996.

Sánchez, Paola. 2012. "Una nueva metodología de entrenamiento de redes neuronales y sus implicaciones en la selección de modelos". Doctorado en Ingeniería en Sistemas, Colombia: Universidad Nacional de Colombia.

Santos, Holger Capa. 2008. *Un primer curso en series temporales*. Primera. Quito.

SBS. 2003. *De la gestión y administración de riesgos*.

———. 2016. *Resolución No. 209-2016-F*.

Scheaffer, Richard L, William Mendenhall, y R. Lyman Ott. 2013. *Elementos de muestreo*. Madrid (España): Paraninfo.

Schiffmann, W, M Joost, y R Werner. 1994. “Optimization of the Backpropagation Algorithm for Training Multilayer Perceptrons”, diciembre.

Shanmugapriya, K. 2012. “Domain Driven Classification of Customer Credit Data for Intelligent Credit Scoring Using Fuzzy Set and MC2” 01 (02): 34–38.

Siddiqi, Naeem. 2006. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Hoboken, N.J: Wiley.

Tablada, Claudio Javier, y Germán Ariel Torres. 2009. “Redes Neuronales Artificiales” 24: 22–30.

Thomas, L. C., David B. Edelman, y Jonathan N. Crook. 2002. *Credit scoring and its applications*. SIAM monographs on mathematical modeling and computation. Philadelphia, PA: Society for Industrial and Applied Mathematics.

TransUnion. 2012. “Preguntas frecuentes sobre los modelos de score de TransUnion”.

Trujillano, J., J. March, M. Badia, A. Rodríguez, y A. Sorribas. 2003. “Aplicación de las redes neuronales artificiales para la estratificación de riesgo de mortalidad hospitalaria”. *Gac Sanit*, 2003.

Tudela, Sanjinés, y Gimmy Nardó. 2011. “Análisis y pronóstico de la demanda de potencia eléctrica en Bolivia: una aplicación de redes neuronales”. *Revista Latinoamericana de Desarrollo Económico*, núm. 15 (mayo): 45–77.

Wasserman, Larry. 2010. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Weldon, Gregg. 2011. “Decision Tree Analysis”. Georgia State University. 2011.

Yoo, Wonsuk, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua (Peter) He, y James W. Lillard. 2014. “A Study of Effects of MultiCollinearity in the Multivariable Analysis”. *International journal of applied science and technology* 4 (5): 9–19.

Yu, Lean, ed. 2008. *Bio-inspired credit risk analysis: computational intelligence with support vector machines*. Berlin: Springer Verlag.

Zeng, Guoping. 2013. “Metric Divergence Measures and Information Value in Credit Scoring”. Research article. Journal of Mathematics. 2013. https://doi.org/10.1155/2013/848271.

Zhang, G. Peter, ed. 2004. *Neural networks in business forecasting*. Hershey, Pa: Idea Group.

Anexos

Anexo 1: Descripción del total de variables

Tabla 1
Descripción de variables

Nombre de Variables	Descripción
V1_Num_Inst_Adeuda	Indica el número de instituciones en las que el cliente tiene deuda en la central de riesgos, a la fecha de desembolso.
V2_NumInstVencida	Indica el número de instituciones en las que el cliente tiene deuda vencida en la central de riesgos, a la fecha de desembolso
V3_NumeroInstitucionesVencido_36M	Indica el número de instituciones en las que el cliente tiene deuda vencida en la central de riesgos, en los últimos 36 meses.
V4_Num_Calificacion_A	Número de veces que el cliente tiene deuda con calificación en A, en los últimos 12 meses.
V5_Num_Calificacion_B	Número de veces que el cliente tiene deuda con calificación en B, en los últimos 12 meses.
V6_Num_Calificacion_CDE	Número de veces que el cliente tiene deuda con calificación en C, D o E, en los últimos 12 meses.
V7_Num_Calificacion_Blanco	Número de veces que el cliente no tiene deuda, en los últimos 12 meses.
V8_NumeroVecesAlDia12M	Número de veces que el cliente tiene deuda al día, en los últimos 12 meses.
V9_NumVecesMora12	Número de veces que el cliente tiene deuda en mora, en los últimos 12 meses.
V10_NumeroIncumplimientos_1_30_12M	Número de veces que el cliente presenta incumplimiento de 1 a 30 días, en los últimos 12 meses.
V11_NumeroIncumplimientos_30_60_12M	Número de veces que el cliente presenta incumplimiento de 30 a 60 días, en los últimos 12 meses.
V12_NumVecesMora_36M	Número de vece en mora en los últimos 36 meses.
V13_NumeroOperacionesVigentes	Número de operaciones con saldo o vigentes a la fecha de desembolso.
V14_NumeroOperacionesAlDia	Número de operaciones al día a la fecha de desembolso.
V13_V14	Variable construida a partir de V13 y V14. (Ver Anexo II)
V15_NumeroOperacionesDemandas	Número de operaciones con demanda judicial en los últimos 36 meses.
V16_NumeroOperacionesCastigo	Número de operaciones con castigo en los últimos 36 meses.
V17_Mora_Ponderada	Mora ponderada por el valor de la cartera, en los últimos 12 meses.
V18_DiasMora	Máximo días de mora a la fecha de desembolso.
V19_DiasMoraMaximaPromedio_24M	Máximo días de mora promedio en los últimos 24 meses.
V20_DiasMoraMaximaPromedio_24M_Sin0	Máximo días de mora promedio en los últimos 24 meses, sin considerar los valores en 0.

V21_DiasMoraMaximaPromedio36M	Máximo días de mora promedio en los últimos 36 meses.
V22_DiasMoraMaximaPromedio36M_Sin0	Máximo días de mora promedio en los últimos 36 meses, sin considerar los valores en 0.
V23_DiasMoraMaximaPromedio_12M	Máximo días de mora promedio en los últimos 12 meses.
V24_DiasMoraMaximaPromedio_12M_Sin0	Máximo días de mora promedio en los últimos 12 meses, sin considerar los valores en 0.
V25_CR_1	(V46 + V47 + V48 +V49) / V50
V26_CR_2	(V56 + V58) / (V56 + V57 + V58)
V27_CR_3	(V42 + V43) /V36
V28_Cartera_Riesgo	Cartera con más de un día de mora / Cartera total, en los últimos 12 meses.
V29_TiempoHistorialCrediticioSF	Tiempo de historial crediticio en los últimos 36 meses.
V30_TiempoHistorialCrediticio_SF_12M	Tiempo de historial crediticio en los últimos 12 meses.
V31_TieneComercial	Tiene un crédito comercial, al desembolso.
V32_TieneConsumo	Tiene un crédito de consumo, al desembolso.
V33_TieneMicro	Tiene un crédito de microcrédito, al desembolso.
V34_TieneOlla	Tiene una olla de oro, al desembolso.
V35_TieneVivienda	Tiene un crédito de vivienda, al desembolso
V36_MayorDeudaHistoricaSF	Mayor deuda en los últimos 36 meses, en el sistema financiero.
V37_MayorDeudaHistorica_Consumo	Mayor deuda de Consumo, en los últimos 36 meses, en el sistema financiero.
V38_MayorDeudaHistorica_Comercial	Mayor deuda de Comercial, en los últimos 36 meses, en el sistema financiero.
V39_MayorDeudaHistorica_Vivienda	Mayor deuda de Vivienda, en los últimos 36 meses, en el sistema financiero.
V40_MayorDeudaHistorica_Micro	Mayor deuda de Microcrédito, en los últimos 36 meses, en el sistema financiero.
V41_MayorDeudaHistorica_Olla	Mayor deuda de Olla de oro, en los últimos 36 meses, en el sistema financiero.
V42_MayorValorVencidoSF	Mayor valor de deuda vencida, en los últimos 36 meses.
V43_MayorValorNDI	Mayor valor de deuda que no devenga intereses, en los últimos 36 meses.
V44_MayorValorDemandaJudicial	Mayor valor de deuda en demanda judicial, en los últimos 36 meses.
V45_SaldoCarteraPorVencer	Saldo de cartera por vencer, a la fecha de desembolso.
V46_CarteraNoDevengaIntereses	Cartera que no devenga intereses, a la fecha de desembolso.
V47_ValorVencido	Cartera vencida, a la fecha de desembolso.
V48_ValorDemandaJudicial	Cartera en demanda judicial, a la fecha de desembolso.
V49_ValorCarteraCastigada	Valor de cartera castigada, a la fecha de desembolso.
V50_SaldoTotal	Saldo total de cartera, a la fecha de desembolso.
V51_Por_End_Al_Corte	V50 / V60
V52_Por_End_U12M	V54 / V60
V53_Por_End_Antes_U12M	V55 / V60

V54_EndeudamientoPromUltimo12M_SF	Endeudamiento promedio en los últimos 12 meses.
V55_EndeudamientoPromAntes12M_SF	Endeudamiento promedio entre los meses 13 y 36.
V56_ValorTotalVencido	Suma de todos los valores en deuda vencida, en los últimos 36 meses.
V57_ValortotalXVencer	Suma de todos los valores en deuda por vencer, en los últimos 36 meses.
V58_ValorTotalNDI	Suma de todos los valores en deuda que no devenga intereses, en los últimos 36 meses.
V59_Cargas_Familiares	Número de cargas familiares del cliente
V60_Ingresos_Mensual	Ingreso mensual del cliente.
V61_EsIndependiente	Variable que determina si el cliente es dependiente o independiente.
V62_AntiguedadResidencia_M	Antigüedad de residencia.
V63_Genero	Género.
V64_Edad	Edad del cliente a la fecha de desembolso.
V65_TipoVivienda	Tipo de vivienda.
V66_EstadoCivil	Estado Civil.
V67_NivelEstudios	Nivel de estudios.
V68_Sector	Sector de vivienda.
V69_TelefonoVivienda	Tiene teléfono de vivienda.
V70_TelefonoEmpresa	Tiene teléfono del lugar de trabajo.
V72_LineaNegocio	Línea de negocio.
V73_ActividadEconomica	Actividad económica.
V74_SectorEconomico	Sector económico.
V75_Cargo	Cargo que ocupa.
V79_CiudadDomicilio	Ciudad de domicilio.
V82_AntiguedadRelacionDependencia	Antigüedad de la relación de dependencia.
V83_TieneGarante	Tiene garante.
V84_DestinoOperacion	Destino de la operación.

Fuente: Institución financiera X
Elaboración: Propia

Anexo 2: Categorización de variables independientes – árboles

Tabla 1
Resumen categorización con árboles de decisión y valor de información

Nombre de variable	Grupos - Árbol	Nueva Categorización	IV	Observación
V79_CiudadDomicilio	Riesgo bajo, riesgo medio y riesgo alto	0; 1; 2	45,4%	Se consideran como variables independientes
V84_DestinoOperacion	Riesgo bajo y riesgo alto	0; 1	39,3%	
V72_LineaNegocio	Riesgo bajo, riesgo medio, riesgo medio - alto y riesgo alto	0; 1; 2; 3	36,8%	
V73_ActividadEconomica	Riesgo bajo, riesgo medio, riesgo medio - alto y riesgo alto	0; 1; 2; 3	26,4%	
V64_Edad	<=29; (29 - 34]; >34	2; 1; 0	9,6%	
V52_Por_End_U12M	<=0,585435; >0,585435	0; 1	9,1%	
V54_EndeudamientoPromUltimo12M_SF	<=444,20749; >444,20749	0; 1	8,7%	
V63_Genero	Femenino; Masculino (0; 1)	0; 1	8,4%	
V33_TieneMicro	0; 1	0; 1	7,7%	
V30_TiempoHistorialCrediticio_SF_12M	<=0; >0	0; 1	7,7%	
V65_TipoVivienda	Propia hipotecada -Propia no hipotecada; Arrendada; Prestada - Vive con familiares -	0; 2; 1	7,6%	
V8_NumeroVecesAlDia12M	<=0; >0	0; 1	7,5%	
V13_V14	(V13) <=0 - (V13 + V14) >1 + >2; 1; >1 + <=2	0; 1; 2	7,4%	
V55_EndeudamientoPromAntes12M_SF	<=182,0266; (182,0266 – 3.158,7966]; >3.158,7966	2; 1; 0	6,9%	
V9_NumVecesMora12	<=0; >0	0; 1	5,7%	
V17_Mora_Ponderada	<=0; >0	0; 1	5,7%	
V23_DiasMoraMaximaPromedio_12M	<=0; >0	0; 1	5,7%	
V24_DiasMoraMaximaPromedio_12M_Sin0	<=0; >0	0; 1	5,7%	
V28_Cartera_Riesgo	<=0; >0	0; 1	5,7%	

V74_SectorEconomico	PRODUCCIÓN - DEPENDENCIA - NO ASIGNADO; COMERCIO - SERVICIOS - PERSONAL	0; 1	5,7%	
V27_CR_3	<=0,159127; >0,159127	0; 1	5,3%	
V53_Por_End_Antes_U12M	<=0,039093; >0,039093	1; 0	5,2%	
V66_EstadoCivil	Soltero; Casado; Divorciado - Unión Libre - Viudo	2; 1; 0	5,2%	
V37_MayorDeudaHistorica_Consumo	<=1.999,96; >1.999,96	1; 0	4,5%	
V1_Num_Inst_Adeuda	0; 1; >=2	0; 1; 2	4,5%	
V68_Sector	Zona Urbana; Zona Rural	0; 1	4,4%	
V13_NumeroOperacionesVigentes	0; 1; >=2	0; 1; 2	4,3%	
V51_Por_End_Al_Corte	<=0; >0	0; 1	4,2%	
V50_SaldoTotal	<=0; >0	0; 1	4,2%	
V58_ValorTotalNDI	<=470,0899; >470,0899	0; 1	4,2%	
V42_MayorValorVencidoSF	<=99.01; >99.01	0; 1	4,0%	
V56_ValorTotalVencido	<=135,2299; >135,2299	0; 1	4,0%	
V67_NivelEstudios	Primaria - Sin estudios; Secundaria - Formación intermedia (técnica) ; Postgrado - Universidad	1; 2; 0	3,8%	
V19_DiasMoraMaximaPromedio_24M	<= 0,625; >0,625	0; 1	3,7%	
V20_DiasMoraMaximaPromedio_24M_Sin0	<=0; >0	0; 1	3,4%	
V21_DiasMoraMaximaPromedio36M	<=2,083; >2,083	0; 1	3,4%	
V32_TieneConsumo	0; 1	0; 1	3,4%	
V22_DiasMoraMaximaPromedio36M_Sin0	<=21; >21	0; 1	3,0%	
V43_MayorValorNDI	<=0; >0	0; 1	2,7%	
V38_MayorDeudaHistorica_Comercial	<=0; >0	1; 0	2,4%	
V26_CR_2	<=0; >0	0; 1	1,9%	
V60_Ingresos_Mensual	<=942,07; >942,07	1; 0	1,9%	

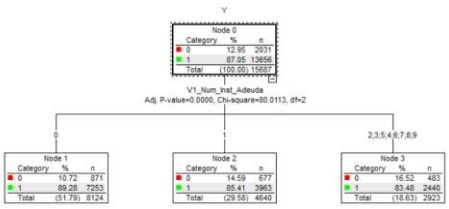
V3_NumeroInstitucionesVencido_36M	0; >=1	0; 1	1,9%	No predictiva por el estadístico IV
V12_NumVecesMora_36M	<=0; >0	0; 1	1,9%	
V83_TieneGarante	0; 1	0; 1	1,7%	
V29_TiempoHistorialCrediticioSF	<=29; >29	1; 0	1,6%	
V14_NumeroOperacionesAlDia	<=1; >1	0; 1	1,6%	
V62_AntiguedadResidencia_M	<=64; >64	1; 0	1,5%	
V57_ValortotalXVencer	<=11.982,29; >11.982,29	0; 1	1,4%	
V36_MayorDeudaHistoricaSF	<=4.120; >4.120	1; 0	1,2%	
V4_Num_Calificacion_A	<12; 12	1; 0	1,2%	
V45_SaldoCarteraPorVencer	<=0; >0	0; 1	1,0%	
V40_MayorDeudaHistorica_Micro	<=1.883,75; >1.883,75	0; 1	1,0%	
V59_Cargas_Familiares	0; >0	1; 0	0,2%	
V70_TelefonoEmpresa	1; 0	1; 0	0,1%	
V7_Num_Calificacion_Blanco	0; >=1	1; 0	0,0%	
V2_NumInstVencida		Es una constante		No supera el análisis exploratorio
V5_Num_Calificacion_B		Es una constante		
V6_Num_Calificacion_CDE		Es una constante		
V10_NumeroIncumplimientos_1_30_12M		Es una constante		
V11_NumeroIncumplimientos_30_60_12M		Es una constante		
V15_NumeroOperacionesDemandas		Es una constante		
V16_NumeroOperacionesCastigo		Es una constante		
V18_DiasMora		Es una constante		
V25_CR_1		Es una constante		
V31_TieneComercial		Es una constante		
V34_TieneOlla		Es una constante		
V35_TieneVivienda		Es una constante		
V39_MayorDeudaHistorica_Vivienda		Es una constante		
V41_MayorDeudaHistorica_Olla		Es una constante		

V44_MayorValorDemandaJudicial		Es una constante		
V46_CarteraNoDevengaIntereses		Es una constante		
V47_ValorVencido		Es una constante		
V48_ValorDemandaJudicial		Es una constante		
V49_ValorCarteraCastigada		Es una constante		
V61_EsIndependiente		Es una constante		
V69_TelefonoVivienda		Es una constante		
V75_Cargo		Es una constante		
V82_AntiguedadRelacionDependencia		Es una constante		

Fuente: Institución financiera X
Elaboración: Propia

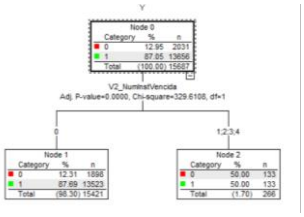
Árboles de decisión

Figura 1
V1_Num_Inst_Adeuda



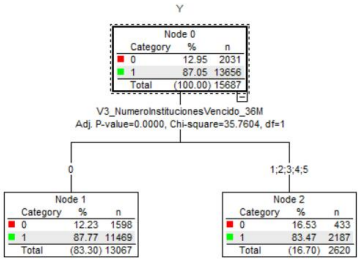
Fuente y elaboración propias

Figura 2
V2_NumInstVencida



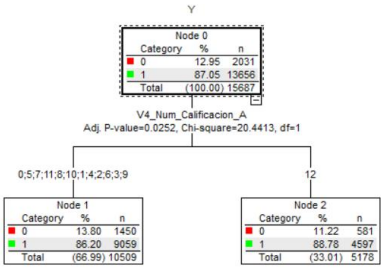
Fuente y elaboración propias

Figura 3
V3_NumeroInstitucionesVencido_36M



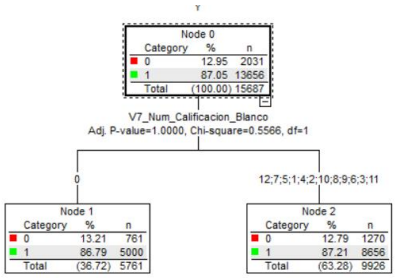
Fuente y elaboración propias

Figura 4
V4_Num_Calificacion_A



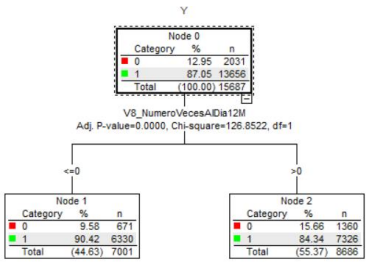
Fuente y elaboración propias

Figura 5
V7_Num_Calificacion_Blanco



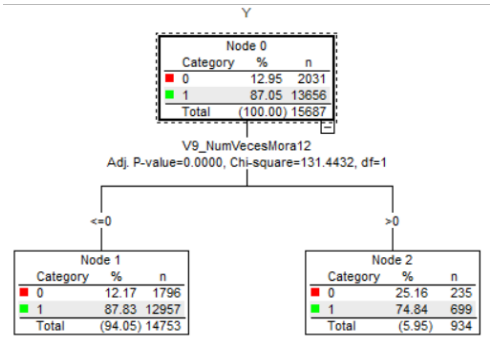
Fuente y elaboración propias

Figura 6
V8_NumeroVecesAlDia12M



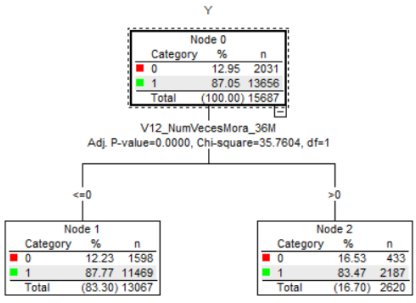
Fuente y elaboración propias

Figura 7
V9_NumVecesMora12



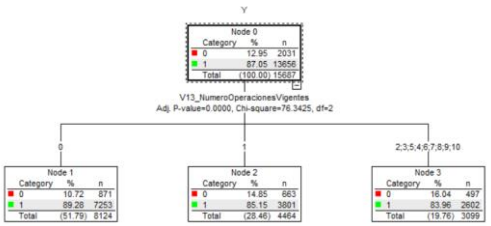
Fuente y elaboración propias

Figura 8
V12_NumVecesMora_36M



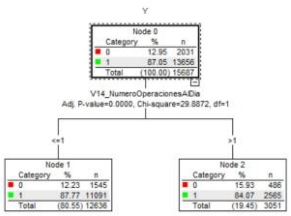
Fuente y elaboración propias

Figura 9
V13_NumeroOperacionesVigentes



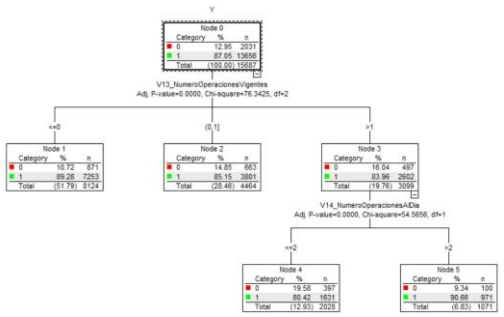
Fuente y elaboración propias

Figura 10
V14_NumeroOperacionesAlDia



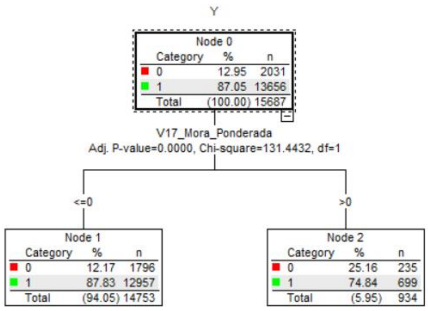
Fuente y elaboración propias

Figura 11
V13_V14



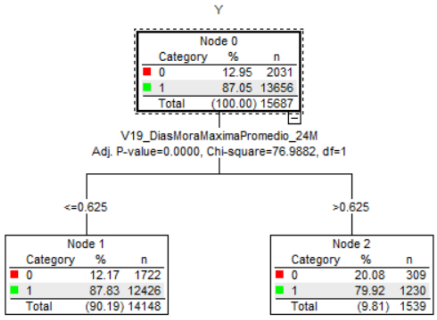
Fuente y elaboración propias

Figura 12
V17_Mora_Ponderada



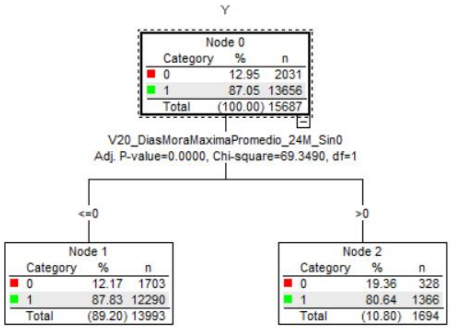
Fuente y elaboración propias

Figura 13
V19_DiasMoraMaximaPromedio_24M



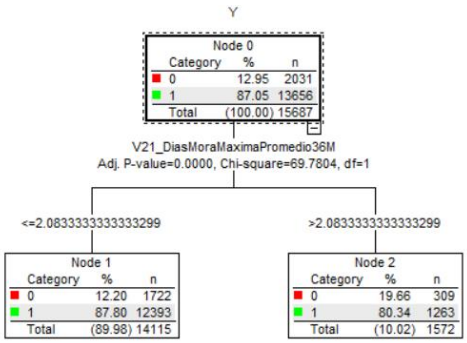
Fuente y elaboración propias

Figura 14
V20_DiasMoraMaximaPromedio_24M_Sin0



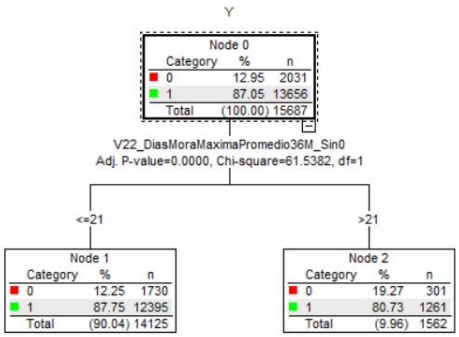
Fuente y elaboración propias

Figura 15
V21_DiasMoraMaximaPromedio36M



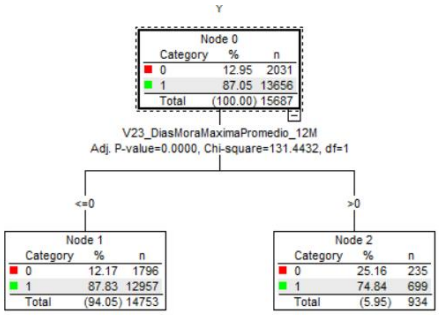
Fuente y elaboración propias

Figura 16
V22_DiasMoraMaximaPromedio36M_Sin0



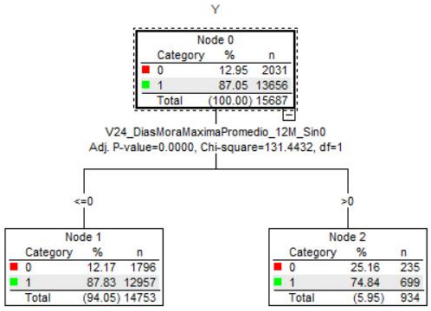
Fuente y elaboración propias

Figura 17
V23_DiasMoraMaximaPromedio_12M



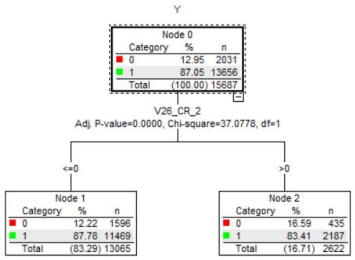
Fuente y elaboración propias

Figura 18
V24_DiasMoraMaximaPromedio_12M_ Sin0



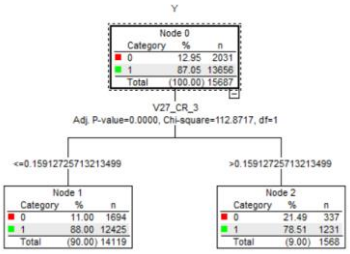
Fuente y elaboración propias

Figura 19
V26_CR_2



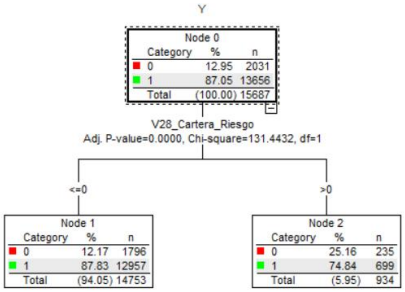
Fuente y elaboración propias

Figura 20
V27_CR_3



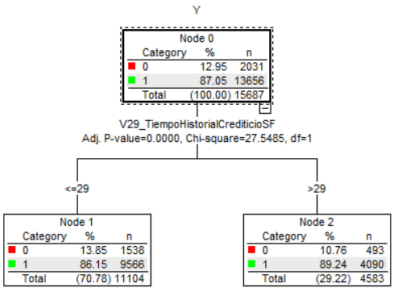
Fuente y elaboración propias

Figura 21
V28_Cartera_Riesgo



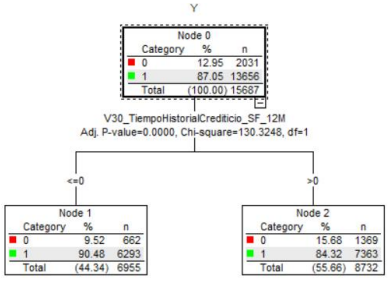
Fuente y elaboración propias

Figura 22
V29_TiempoHistorialCrediticioSF



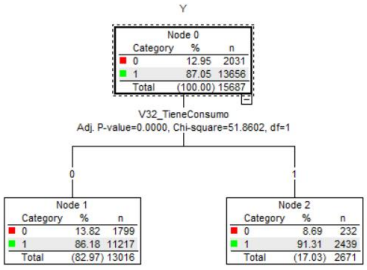
Fuente y elaboración propias

Figura 23
V30_TiempoHistorialCrediticio_SF_12M



Fuente y elaboración propias

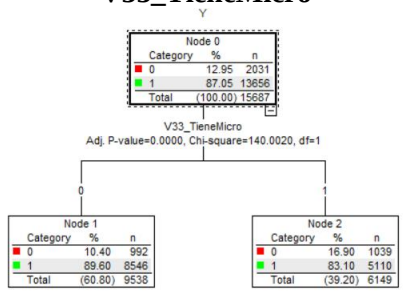
Figura 24
V32_TieneConsumo



Fuente y elaboración propias

Figura 25

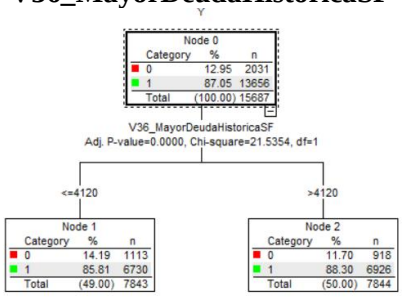
V33_TieneMicro



Fuente y elaboración propias

Figura 26

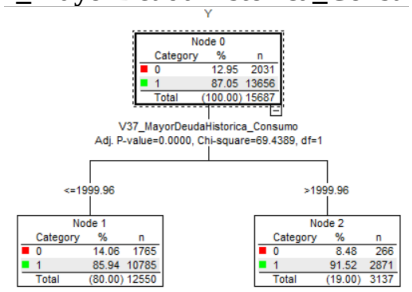
V36_MayorDeudaHistoricaSF



Fuente y elaboración propias

Figura 27

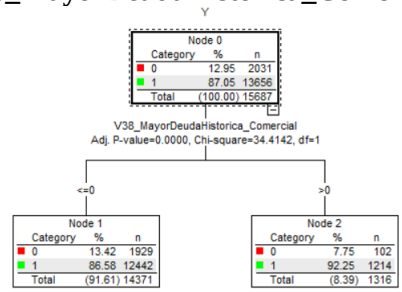
V37_MayorDeudaHistorica_Consumo



Fuente y elaboración propias

Figura 28

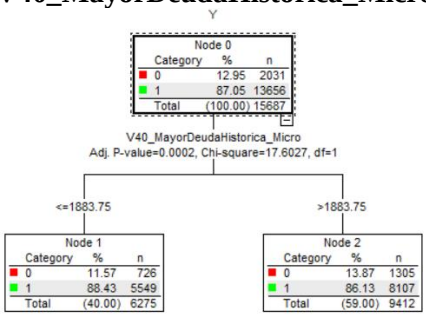
V38_MayorDeudaHistorica_Comercial



Fuente y elaboración propias

Figura 29

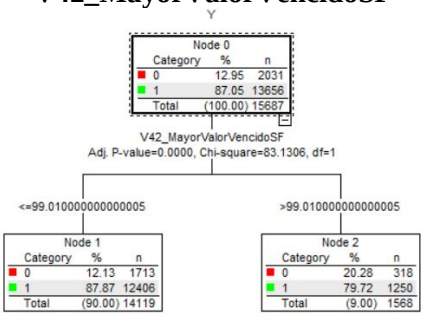
V40_MayorDeudaHistorica_Micro



Fuente y elaboración propias

Figura 30

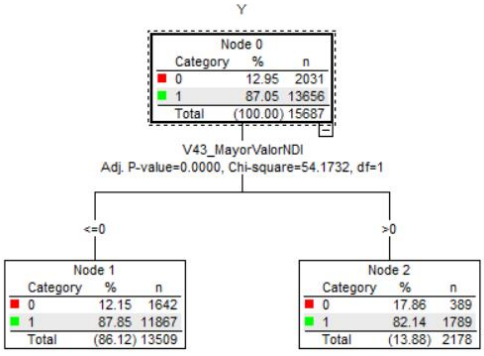
V42_MayorValorVencidoSF



Fuente y elaboración propias

Figura 31

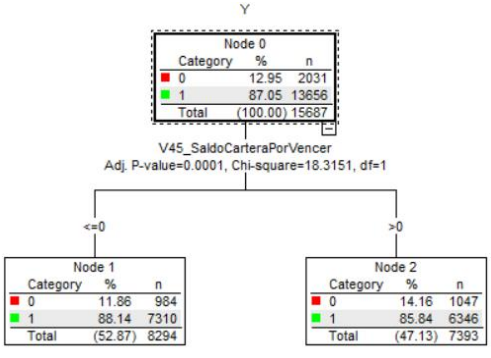
V43_MayorValorNDI



Fuente y elaboración propias

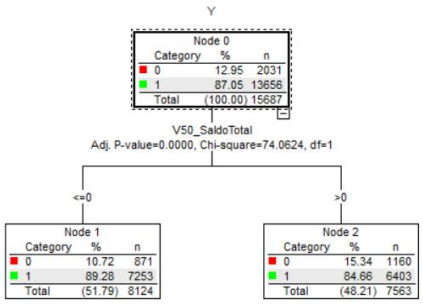
Figura 32

V45_SaldoCarteraPorVencer



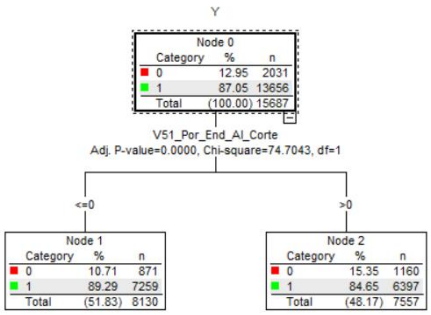
Fuente y elaboración propias

Figura 33
V50_SaldoTotal



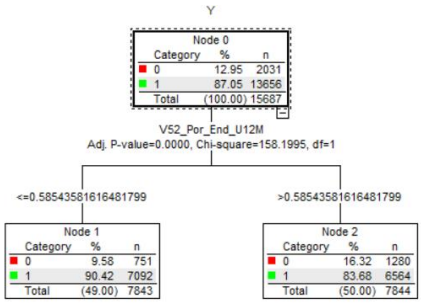
Fuente y elaboración propias

Figura 34
V51_Por_End_Al_Corte



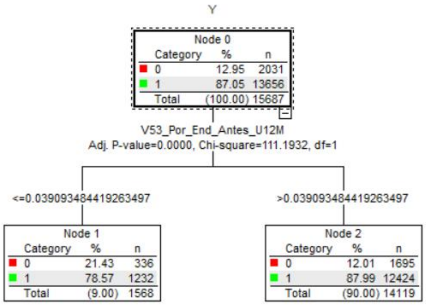
Fuente y elaboración propias

Figura 35
V52_Por_End_U12M



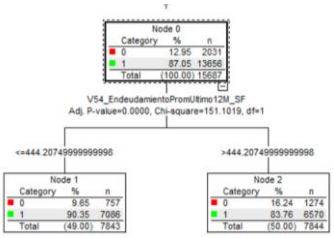
Fuente y elaboración propias

Figura 36
V53_Por_End_Antes_U12M



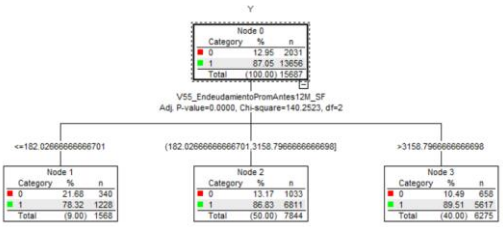
Fuente y elaboración propias

Figura 37
V54_EndeudamientoPromUltimo12M_SF



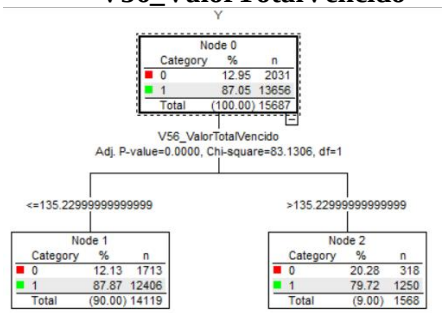
Fuente y elaboración propias

Figura 38
V55_EndeudamientoPromAntes12M_SF



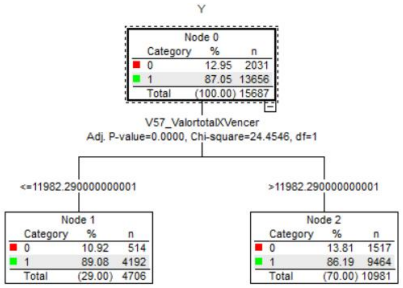
Fuente y elaboración propias

Figura 39
V56_ValorTotalVencido



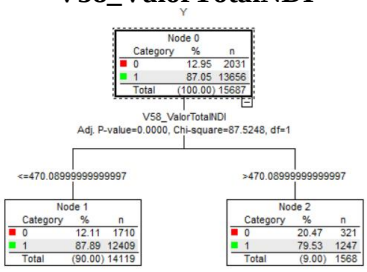
Fuente y elaboración propias

Figura 40
V57_ValortotalXVencer



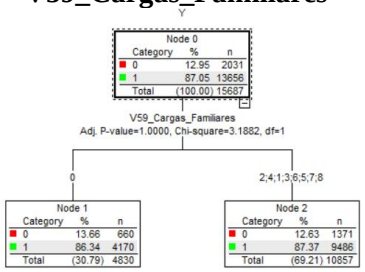
Fuente y elaboración propias

Figura 41
V58_ValorTotalNDI



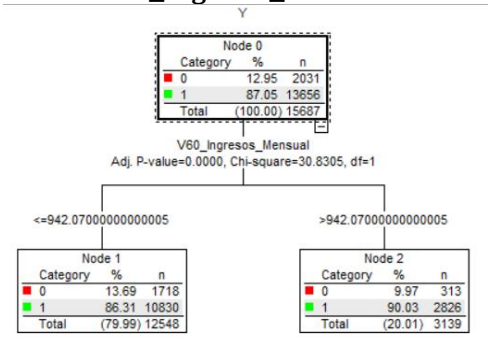
Fuente y elaboración propias

Figura 42
V59_Cargas_Familiares



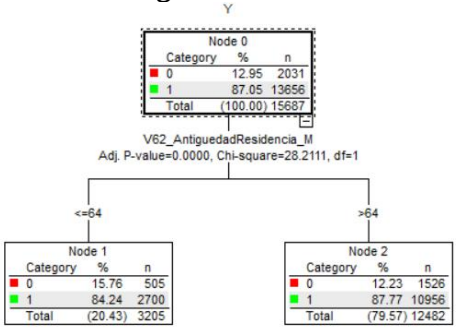
Fuente y elaboración propias

Figura 43
V60_Ingresos_Mensual



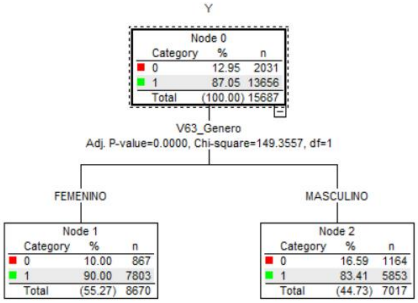
Fuente y elaboración propias

Figura 44
V62_AntiguedadResidencia_M



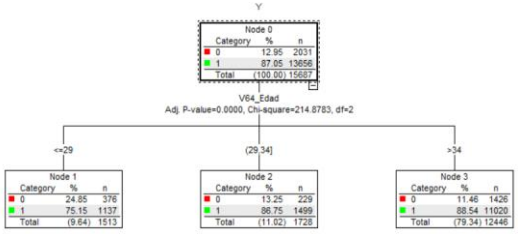
Fuente y elaboración propias

Figura 45
V63_Genero



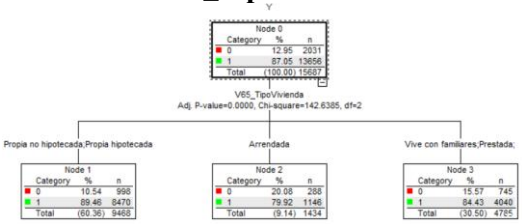
Fuente y elaboración propias

Figura 46
V64_Edad



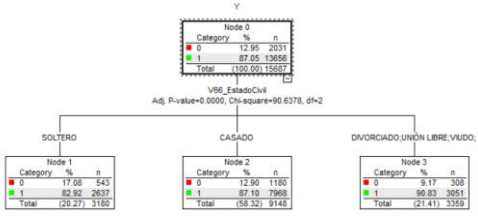
Fuente y elaboración propias

Figura 47
V65_TipoVivienda



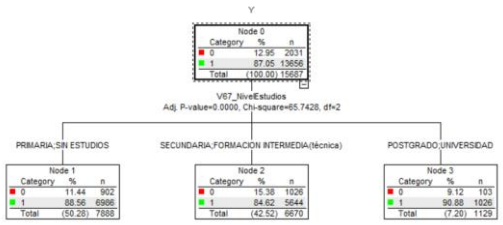
Fuente y elaboración propias

Figura 48
V66_EstadoCivil



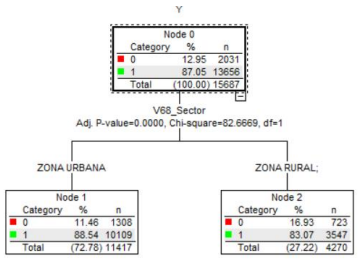
Fuente y elaboración propias

Figura 49
V67_NivelEstudios



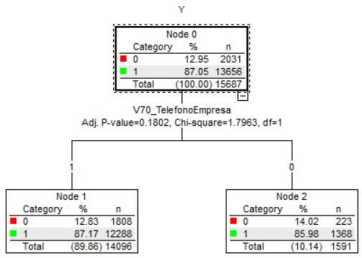
Fuente y elaboración propias

Figura 50
V68_Sector



Fuente y elaboración propias

Figura 51
V70_TelefonoEmpresa



Fuente y elaboración propias

Figura 52
V72_LineaNegocio

%Malos - Categorías			
8.20%	**11.88%**	**16.35%**	**36.58%**
0	1	2	3
Alimento animales	Alimentos	Agricultura	Admin publica
Alojamiento	Entretenimiento	Contruccion	Agencia viajes
Ama de casa	Quimicos	Cuero	Cemento / canteros
Animales	Tejidos	Empleado	Conservas / enlatados
Arinas		Entretenimiento / recreacion	Equipos
Arroz		Farmacos	Financieros
Art. Personales		Impresión	Industrial
Asesoria consultoria		Plastico	Informatica
Azucar		Restaurantes	Insumos
Bebidas		Telecomunicaciones	Madera
Calzado		Varios	Materias prmas
Caucho		Vegetales	Metales
Ceramica / vidrios			Muebles
Cereales			No asignado
Combustibles			Otros
Confites			Publicidad
Construccion			Video
Edificios			Viveres
Educacion			
Eventos			
Flores			
Frutas			
Hilados			
Hogar			
Inmobiliarios			
Intrumentos musicales			
Jubilado			
Juguetes			
Lacteos / huevos			
Maquinaria			
Maquinaria/ equipos / vehiculos			
Materia prima			
Materiales			
Medicos			
Metal			
Papel			
Papel / carton			
Pastos			
Pescado / mariscos			
Recreacion			
Reparacion			
Serv. Salud			
Servicios			
Transporte			
Vehiculos			
Vestuario			
Veterinarios			
Zapatos			

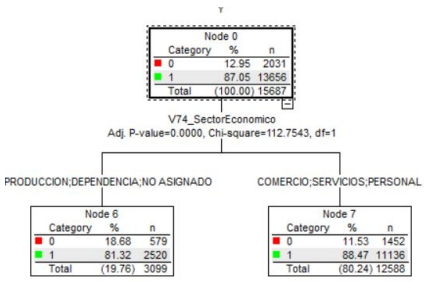
Fuente y elaboración propias

Figura 53
V73_ActividadEconomica

%Malos - Categorías			
6.24%	**10.70%**	**16.83%**	**29.27%**
0	**1**	**2**	**3**
Basicos	Agricultura	Empleado	Admin publica
Hogar / domestico	Alimentos / animales / bebidas	Informacion y comunicación	Alimentos
Inmobiliarios	Ama de casa	Plastico / caucho	Contruccion / materiales
Jubilado	Art metal / metales	Restaurantes / hoteles	Financieros
Papel / carton	Del hogar / personal / empresas		Maquinaria/ equipos / vehiculos
Serv. Salud	Educacion		Maquinaria/ equipos / vehiculos/artesanias
Transporte/ almacenamiento / courier	Entretenimiento /deportes		Muebles / art. Hogar
	Farmacos / art. Personales		No asignado
	Flores		Otros
	Materia primas /quimicos		
	Materia primas /quimicos / gas		
	Serv. Profesionales		
	Vestuario		

Fuente y elaboración propias

Figura 54
V74_SectorEconomico



Fuente y elaboración propias

Figura 55

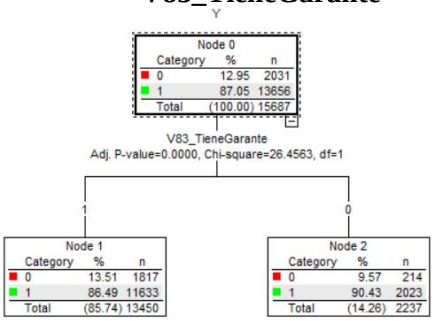
V79_CiudadDomicilio

%Malos - Categorías		
2.16%	**10.70%**	**33.17%**
0	**1**	**2**
Alfredo Baquerizo Moreno (Jujan)	24 De Mayo	Antonio Ante
Azogues	Ambato	Atacames
Bolivar	Babahoyo	Biblian
Chone	Baños De Agua Santa	El Empalme
Chordeleg	Buena Fe	El Triunfo
Coronel Marcelino Maridueña	Catamayo	Ibarra
Cotacachi	Cayambe	La Mana
Cumanda	Cevallos	Lomas De Sargentillo
Deleg	Chambo	Milagro
El Pan	Colta	Mira
Espejo	Cuenca	Mocache
Flavio Alfaro	Daule	Mocha
Giron	Duran	Naranjito
Guachapala	El Carmen	Otavalo
Gualaceo	El Guabo	Pimampiro
Isidro Ayora	Esmeraldas	Portoviejo
Jama	Guamote	Pujili
Jaramijo	Guano	Samborondon
Manta	Guayaquil	Santa Isabel
Montufar	Jipijapa	Sigchos
Naranjal	Junin	
Olmedo	La Concordia	
Pajan	La Libertad	
Pangua	Latacunga	
Pedernales	Loja	
Penipe	Machala	
Pichincha	Mejia	
Quinsaloma	Montecristi	
Rioverde	Nobol	
Rocafuerte	Pasaje	
San Miguel De Urcuqui	Patate	
San Vicente	Paute	
Santa Ana	Pedro Carbo	
Santa Lucia	Pedro Moncayo	
Santa Rosa	Playas	
Sigsig	Puerto Lopez	
Tulcán	Quero	
Ventanas	Quevedo	

Fuente y elaboración propias

Figura 56

V83_TieneGarante



Fuente y elaboración propias

Figura 57

V84_DestinoOperacion

%Malos - Categorías	
10.43%	**43.80%**
0	**1**
Activos fijos tangibles	Activos fijos intangibles
Capital de trabajo	Mejoramiento otros activos
Pago de deudas varias	No productivas (para operaciones de consumo y vivienda)

Fuente y elaboración propias

Anexo 3: Código R

Script desarrollo score logit

```
## Definición de variables y fórmula

Variable_Dependiente <- 'Y'

Nombres_Variables_Independientes <-
c(names(Matriz_Datos)[2:length(Matriz_Datos)])

reg <- as.formula(paste(Variable_Dependiente,
paste(Nombres_Variables_Independientes, collapse=" + "),
sep=" ~ "))

## Construcción del modelo

set.seed(7)
Modelo_Logistico <- step(glm(reg,data = Matriz_Datos,
family = binomial(link="logit")), trace=0)

### Prueba GVIF

vif(Modelo_Logistico)

### Cálculo del z, valor de probabilidad y Score

Matriz_Final_N_1$Z = predict.glm(Modelo_Logistico,
type='link', Matriz_Datos)
Matriz_Datos$prob <- predict(Modelo_Logistico,
Matriz_Datos,type = 'response')
Matriz_Datos$Score = round(1000 * Matriz_Datos $prob, 0)

m1_pred <- prediction(Matriz_Datos$prob , Matriz_Datos$Y)
m1_perf <- performance(m1_pred,"tpr","fpr")

### Cálculo de los estadísticos

#### KS

KS <- round(max(attr(m1_perf,'y.values')[[1]]-
attr(m1_perf,'x.values')[[1]])*100, 2)

#### RCO

ROC <- round(performance(m1_pred, measure =
"auc")@y.values[[1]]*100, 2)

Gini <- (2*ROC - 100)

#### Punto de corte óptimo y matriz de confusión
```

```
optCutOff <- optimalCutoff(Matriz_Datos$Y,
Matriz_Datos$prob, optimiseFor = "Both", returnDiagnostics
= TRUE)
PuntoCorte <- optCutOff$optimalCutoff

Error <- round(optCutOff$misclassificationError*100,2)

Matriz_Datos$Y_estimado <-
ifelse(Matriz_Datos$prob>PuntoCorte,1,0)
MatrizConfusion <- confusionMatrix(Matriz_Datos$Y,
Matriz_Datos$Y_estimado)

Sensibilidad_M  <- round(100*sensitivity(Matriz_Datos$Y,
Matriz_Datos$Y_estimado, threshold = PuntoCorte) ,2)
Especificidad_M <- round(100*specificity(Matriz_Datos$Y,
Matriz_Datos$Y_estimado, threshold = PuntoCorte),2)


## Gráfica curva ROC

plot(m1_perf, col='blue', main = 'Curva ROC', xlab = '1 -
Especificidad', ylab =  'Sensibilidad')
lines(c(0:1), c(0,1), col='black')
```

Script desarrollo score redes neuronales

```
## Definición de variables y fórmula

Variable_Dependiente <- 'Y'

Nombres_Variables_Independientes <-
c(names(Matriz_Datos)[2:length(Matriz_Datos)])

reg <- as.formula(paste(Variable_Dependiente,
paste(Nombres_Variables_Independientes, collapse=" + "),
sep=" ~ "))

## Construcción del modelo

set.seed(7896129)
library(neuralnet)

t <- proc.time()
Modelo_Redes<- neuralnet(reg, hidden = c(3,3,2) ,data =
Matriz_Datos,linear.output = FALSE,algorithm =
'rprop+',likelihood = TRUE,threshold = 0.1)
proc.time()-t
model_results3 <- compute(Modelo_Redes, Matriz_Datos
[2:length(Matriz_Datos)])
```

```
### Gráfica red neuronal

plot(Modelo_Redes,radius = 0.04,arrow.length =
0.15,fontsize = 8)

### Cálculo del z, valor de probabilidad y Score

Matriz_Datos$prob <- model_results3$net.result
Matriz_Datos$Score = round(1000 * Matriz_Datos$prob, 0)

detach(package:neuralnet,unload = T)
m1_pred <- prediction(Matriz_Datos$prob , Matriz_Datos$Y)
m1_perf <- performance(m1_pred,"tpr","fpr")

### Cálculo de los estadísticos

#### KS

KS <- round(max(attr(m1_perf,'y.values')[[1]]-
attr(m1_perf,'x.values')[[1]])*100, 2)

#### RCO

ROC <- round(performance(m1_pred, measure =
"auc")@y.values[[1]]*100, 2)

Gini <- (2*ROC - 100)

#### Punto de corte óptimo y matriz de confusión

optCutOff <- optimalCutoff(Matriz_Datos$Y,
Matriz_Datos$prob, optimiseFor = "Both", returnDiagnostics
= TRUE)
PuntoCorte <- optCutOff$optimalCutoff
Error <- round(optCutOff$misclassificationError*100,2)

Matriz_Datos$Y_estimado <-
ifelse(Matriz_Datos$prob>PuntoCorte,1,0)
MatrizConfusion <- confusionMatrix(Matriz_Datos$Y,
Matriz_Datos$Y_estimado)

Sensibilidad_M  <- round(100*sensitivity(Matriz_Datos$Y,
Matriz_Datos$Y_estimado, threshold = PuntoCorte) ,2)
Especificidad_M <- round(100*specificity(Matriz_Datos$Y,
Matriz_Datos$Y_estimado, threshold = PuntoCorte),2)

## Gráfica curva ROC

plot(m1_perf, col='blue', main = 'Curva ROC', xlab = '1 -
Especificidad', ylab =  'Sensibilidad')
lines(c(0:1), c(0,1), col='red')
```