

MODELOS DE CALIFICACIÓN CUANTITATIVA (SCORING) PARA AGILIZAR CRÉDITOS Y MITIGAR EL RIESGO



**EMILIO VILLEGAS MOLINA
JOHN ALFREDO RESTREPO LLANOS**

**Christian Lochmüller
MSc.**

**ESCUELA DE INGENIERÍA DE ANTIOQUIA
Ingeniería Administrativa**

**ENVIGADO
Mayo de 2013**

AGRADECIMIENTOS

Un agradecimiento especial al profesor Christian Lochmueller, por todo su apoyo y dedicación en el desarrollo del proyecto. Sin sus conocimientos y experiencia no se habría podido alcanzar los objetivos. También se les agradece a los empleados y colaboradores de la empresa Sistecrédito SAS por su gestión y excelente disposición por la causa.

CONTENIDO

| | pág. |
|---|------|
| INTRODUCCIÓN..... | 9 |
| 1 PRELIMINARES..... | 11 |
| 1.1 Planteamiento del problema | 11 |
| 1.2 Objetivos del proyecto | 11 |
| 1.2.1 Objetivo General..... | 11 |
| 1.2.2 Objetivos Específicos | 11 |
| 1.3 Marco De Referencia..... | 12 |
| 1.3.1 Antecedentes | 12 |
| 1.3.2 Que es el Scoring? | 14 |
| 1.3.3 Modelos No Paramétricos | 17 |
| 1.3.4 Modelos Paramétricos | 18 |
| 1.3.5 Curva ROC..... | 24 |
| 1.3.6 El Análisis Descriptivo | 24 |
| 2 METODOLOGÍA..... | 26 |
| 3 DESARROLLO DEL MODELO ESTADISTICO PARA MITIGAR EL RIESGO Y AGILIZAR EL ESTUDIO DE CREDITO | 27 |
| 3.1 Base de datos..... | 27 |
| 3.2 Técnicas para Desarrollar el Scoring | 38 |
| 3.3 Construcción del modelo | 39 |
| 3.3.1 Codificación de variables categóricas..... | 39 |
| 3.3.2 Corridas para construir un modelo LOGIT | 41 |
| 3.4 Validación deL Modelo LOGIT (conStruido con SPSS)..... | 51 |

| | | |
|-------|--|----|
| 3.4.1 | Validación con R PROJECT y RATTLE | 51 |
| 3.4.2 | Validación con PALISADE DECISION TOOLS | 57 |
| 3.5 | Perfil de Riesgo de la empresa | 59 |
| 4 | DISCUSION DE RESULTADOS | 63 |
| 5 | CONCLUSIONES Y CONSIDERACIONES FINALES | 67 |
| | BIBLIOGRAFÍA..... | 68 |

LISTA DE FIGURAS

| | |
|---|----|
| Figura 1. Modelos de Calificación. Fuente: Elaboración Propia. ¡Error! Marcador no definido. | 20 |
| Figura 2. Función de distribución. Fuente: (Abuín, 2007)..... | 20 |
| Figura 4. Distribución para Edad. Fuente: Elaboración Propia..... | 31 |
| Figura 5. Distribución para Cantidad Créditos. Fuente: Elaboración Propia..... | 32 |
| Figura 6. Distribución para Promedio. Fuente: Elaboración Propia..... | 33 |
| Figura 7. Distribución para Referencias Comerciales. Fuente: Elaboración Propia..... | 34 |
| Figura 8. Distribución para Referencias Personales. Fuente: Elaboración Propia..... | 35 |
| Figura 9. Distribución para Años laborando. Fuente: Elaboración Propia..... | 36 |
| Figura 10. Distribución para Ingresos. Fuente: Elaboración Propia..... | 37 |
| Figura 11. Variables Dummy. Fuente: Elaboración Propia..... | 40 |
| Figura 12. Matriz de Correlación. Fuente: Elaboración Propia..... | 40 |
| Figura 13. Variables SPSS. Fuente: Elaboración Propia..... | 41 |
| Figura 14. . Variables. Fuente: Elaboración propia..... | 41 |
| Figura 15. Tabla de Clasificación. Fuente: Elaboración Propia..... | 42 |
| Figura 16. Tabla de Clasificación. Fuente: Elaboración Propia..... | 43 |
| Figura 17. Tabla de Hosmer. Fuente: Elaboración Propia..... | 44 |
| Figura 18. Tabla de Clasificación. Fuente: Elaboración Propia..... | 44 |
| Figura 20. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia..... | 45 |
| Figura 21. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia..... | 46 |
| Figura 22. Tabla de Hosmer. Fuente: Elaboración Propia..... | 47 |
| Figura 23. Tabla de Clasificación. Fuente: Elaboración Propia..... | 47 |
| Figura 24. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia..... | 48 |

| | |
|---|----|
| Figura 25. Curva ROC. Fuente: Elaboración Propia. | 49 |
| Figura 26. Boxplot Edad. Fuente: Elaboración Propia..... | 52 |
| Figura 27. Boxplot Salario. Fuente: Elaboración Propia..... | 53 |
| Figura 28. Matrices de Error arrojado por Rattle en R. Fuente: Elaboración Propia. | 55 |
| Figura 29. Resultados Curva ROC. Fuente: Elaboración Propia..... | 55 |
| Figura 30. Summary StatTools. Fuente: Elaboración Propia..... | 57 |
| Figura 31. Tabla de Clasificación. Fuente: Elaboración Propia. | 58 |
| Figura 32. Impacto de las Variables. Fuente: Elaboración Propia..... | 58 |
| Figura 33. Tabla ejemplo perfil de riesgo. Fuente: Elaboración Propia. | 59 |
| Figura 34. Tabla ejemplo orden Deciles. Fuente: Elaboración Propia..... | 60 |
| Figura 35. Análisis Perfil de riesgo. Elaboración Propia..... | 61 |
| Figura 36. Validación de Perfil de Riesgo. Elaboración Propia..... | 61 |

RESUMEN

La génesis de este trabajo de grado surge al tratar de estudiar una solicitud de crédito de manera aleatoria estadísticamente, teniendo en cuenta el peso que tienen las variables en la probabilidad de que un crédito se pague o se incumpla. Por esta razón, la empresa Sistecredito SAS ha facilitado su base de datos para realizar el análisis estadístico de este problema y desarrollar un modelo que permita medir el riesgo con base a las variables explicativas.

Inicialmente, se analiza toda la información disponible de los clientes de manera que se puedan definir las variables apropiadas para realizar el modelo. Se tienen en cuenta las características de cada cliente como el municipio donde vive, el estado civil, entre otros. Examinando la base de datos, se definen cuales variables utilizar, para tratar de pronosticar el comportamiento del cliente.

Una vez identificadas las variables, que se puedan expresar de manera numérica o categórica, se procede a analizar las diferentes técnicas que permitan realizar el modelo. Se analizarán los posibles modelos que se ajusten a la base de datos para realizar el análisis estadístico apropiado, analizando las ventajas y desventajas de cada uno.

Luego de escoger la técnica a desarrollar, se realiza una depuración de la base de datos para eliminar registros atípicos que puedan afectar la validez. Se procede a realizar las primeras corridas para encontrar el modelo. Inicialmente, no se encuentran resultados significativos, por lo cual se procede a realizar varios cambios en la base de datos y se desarrolla un modelo alterno paralelamente al inicialmente escogido.

Por último, se logra encontrar la estructura adecuada de la base de datos que permite ejecutar un modelo con dos técnicas escogidas, la regresión logística y la red neuronal. Estos últimos modelos encontrados tienen la capacidad de predecir cierto porcentaje de clientes de manera correcta, y se valida a través de la curva ROC y las matrices de clasificación. Paso a paso se obtiene el perfil de riesgo de la empresa y se muestra como analizar este concepto para maximizar las ganancias. Finalmente se comparan las dos técnicas desarrolladas y se realizan las recomendaciones apropiadas de acuerdo a las necesidades de la empresa. Se concluye que el modelo propuesto logra mitigar el riesgo y agilizar el proceso de estudio de crédito.

PALABRAS CLAVE: Riesgo Crédito, Modelos estadísticos, Calificación, Incumplimiento.

ABSTRACT

The genesis of the thesis arises when trying to study a credit application in a random way, without the consideration of the impact of each variable on the probability that a loan is paid or defaults. For this reason, the company Sistecredito SAS has facilitated a database for the statistical analysis of this problem and the development of a model to measure risk based on the explanatory variables.

Initially, all the information available from clients is analyzed to define the appropriate variables for the model. Taking into account the characteristics of each client like the municipality where they live or the marital status, among others, the authors define the variables to predict customer's behavior.

Having identified the variables that can be expressed in a numerical or categorical manner, the authors proceed to analyze the different techniques to construct the model. The project takes into account possible models that adjust the database structure to perform the appropriate statistical analysis, analyzing the advantages and disadvantages of each technique.

After choosing the model to develop, a thorough purification of the database is carried out to eliminate outliers that may affect the validity of the model. Subsequently, several runs are carried out to find the precise model. Initially, results are not significant, leading to several changes in the database and develop an improved model.

Finally, a structure of the database that confirms the validity with the two chosen techniques is found. This last model is validated through analyzing the ROC curve and classification matrices and the ability of the model to predict a certain percentage of customers correctly. Step by step the authors find the risk profile of the company and show how to analyze this concept to maximize profits. Finally, the results of the two techniques developed are compared and recommendations for future work are made that are based on the company's needs. The work concludes that the proposed model helps in mitigating risk and improves the study process of a credit application.

KEY WORDS: Credit Risk, Statistical Models, Scoring, Default.

INTRODUCCIÓN

En Colombia los préstamos han venido aumentando y el registro histórico se ha vuelto cada vez más robusto, por lo cual se nota una falencia en cuanto a modelos estadísticos para gestionar las solicitudes de crédito y la cobranza de una manera rápida y eficaz. En el cuarto trimestre de 2012 “la demanda de crédito percibida por los intermediarios financieros aumentó, luego de haber presentado una desaceleración durante el resto del año. Al analizar la demanda de las distintas modalidades se observa que las carteras de consumo y comercial presentan el mayor dinamismo, de acuerdo con lo indicado por los tres grupos de establecimientos de crédito” (Castaño, Estrada, & Patiño, 2012).

Es importante que los bancos conozcan sus clientes para determinar el nivel de impago o mora y así colocar sus dineros en préstamos menos riesgosos. Se debe tener presente el tratado de Basilea III que consiste básicamente en legislar y regular el sector bancario, esta medida se toma a raíz de la crisis financiera del 2008, dado que por no estar lo suficientemente regulado y supervisado, se desarrolló una de las peores crisis económicas de la historia. Previamente, el sector financiero, sobretodo en E.U., se encontraba relativamente sin supervisión por lo cual los bancos aprovecharon para colocar los dinero captados de manera fácil y desmedida por medio de hipotecas subprime (riesgosas), formando una burbuja especulativa de tal magnitud que llevó a la quiebra a bancos, instituciones financieras y aseguradoras por todo el mundo, repercutiendo en la economía global. Por esta razón el tratado de Basilea III, el cual dicta que el sistema financiero debe ser mucho más riguroso en cuanto a la gestión del riesgo. Para esto se tienen tres pilares (Marasca, Figueroa, Stefanelli, & Indri, 2003):

1. Requerimientos mínimos de capital: Este pilar se enfoca en los riesgos teniendo como principales riesgo de crédito, riesgo de mercado y riesgo operativo.
2. Proceso de supervisión bancaria: Este pilar es más riguroso con los procesos con los que deben contar los bancos y tienen supervisores
3. Disciplina de mercado: Aparecen requerimientos de obligación de divulgación con el objetivo de que el cliente pueda evaluar el riesgo del banco donde va a depositar o a prestar el dinero.

La mayoría de las empresas poseen información valiosa que permite realizar diversos tipos de análisis, pero se detecta que muy pocas empresas medianas y pequeñas aprovechan la información disponible. En el contexto actual, se puede ver claramente como las empresas grandes en general son las que explotan la información utilizando diversas técnicas y programas de software, pero también existen casos de empresas medianas y pequeñas que pueden realizar un trabajo similar teniendo en cuenta las restricciones de recursos y talento humano. En este caso, partiendo de la base de datos de Sistecredito SAS, se alcanza a ver como la empresa posee una información la cual no se está explotando adecuadamente, siendo esta la génesis de este proyecto.

Se ve la necesidad de explotar la base de datos disponible para mejorar el otorgamiento de crédito. Dado que se trata de una empresa mediana, surge la misión de realizar una minería de datos con la información disponible y encontrar la solución para efectuar un estudio de crédito más acertado.

Lo esencial en este estudio es obtener un modelo estadístico que permita mitigar el riesgo al momento de otorgar crédito y además que permita disminuir el tiempo del estudio del mismo.

Inicialmente se presentara un informe sustentando el estudio de: cuales son las variables o parámetros más significativos para estudiar un cliente y así predecir su comportamiento. Identificando cuales variables tienen más valor que otras y cuáles son las críticas para que el modelo funcione de manera apropiada. Se presentara un breve estudio de diferentes técnicas en diferentes software (red neuronal-Excel, Logit-SPSS, Logit y red neuronal-R), identificando ventajas y desventajas de cada una, y se concluirá decidiendo, cual es el más apropiado. De esta forma se logra disminuir tiempo de respuesta y se disminuye el riesgo al otorgar créditos (S.A.S).

El análisis de la base de datos que se presenta en las siguientes secciones, consiste en una mirada a las características del cliente y la identificación de las variables que permitan predecir una condición, o un estado de "salida". En este caso, gravita en el análisis para modelar si un cliente pagara o no un crédito. Este análisis se puede generalizar para todas las empresas que poseen diferentes problemáticas y gozan de información para estudiar al cliente.

El modelo presentado en los siguientes capítulos consiste en un análisis cuantitativo de la base de datos, utilizando las variables explicativas para modelar la probabilidad de que el cliente sea "buena paga". Se pretende encontrar patrones subyacentes a la base de datos para modelar el comportamiento del cliente, utilizando técnicas que se explicaran posteriormente, de manera que se puedan utilizar estos conocimientos para modelar cualquier tipo de problema.

Para realizar dicho modelo se realizó un análisis de la base de datos, identificando la información relevante para la construcción del modelo. Posteriormente se analizan las diferentes técnicas para realizar el análisis estadístico apropiado y por último se procede a la construcción del modelo que permita pronosticar el comportamiento del cliente. Por último se valida el modelo con parte de la base de datos que no fue utilizada para la construcción del mismo, y se discuten los resultados analizando las diferentes técnicas y los diferentes programas (software) para dicho análisis.

1 PRELIMINARES

1.1 PLANTEAMIENTO DEL PROBLEMA

El problema planteado es básicamente hallar la forma de disminuir el riesgo y a su vez facilitar el proceso de préstamo. Esto favorecerá tanto al prestamista como el prestatario, por otro lado, volviéndose muy importante hacerle un seguimiento a los clientes, según la calificación dada (bueno: pago a tiempo y pago total; malo: pago en mora o impago). Para llevar a cabo este estudio se evaluarán varios métodos estadísticos para hallar un modelo que converja a calificar de manera acertada el cliente y de igual forma se agilice el proceso de estudio de crédito. Esto se hará comparando el modelo más innovador conocido en scoring: Redes Neuronales y el modelo tradicional: Logit.

Existe la necesidad de profundizar en el manejo y administración de la información de los clientes, para gestionar el riesgo utilizando modelos estadísticos, y a partir de una base de datos con clientes antiguos, disminuir el tiempo de estudio, de crédito y predecir el comportamiento de clientes potenciales.

La idea, es utilizar uno o varios modelos que permitan darle una calificación cuantitativa y estadísticamente objetiva al prestatario, utilizando los datos del cliente (edad, sexo, estrato, profesión, etc.) y las bases de datos históricas disponibles, las cuales se trabajarán con la empresa Sistecrédito S.A.S para así proponer un modelo estadístico confiable.

1.2 OBJETIVOS DEL PROYECTO

1.2.1 Objetivo General

Desarrollar un modelo apropiado para mitigar el riesgo, aumentando la velocidad en la entrega de los créditos, y, para identificar clientes potenciales (historia) por medio de este modelo.

1.2.2 Objetivos Específicos

- Identificar las variables y parámetros necesarios para definir los criterios que permitan desarrollar el modelo más apropiado para calificar el cliente.
- Analizar las alternativas de los diferentes modelos estadísticos, basados en las técnicas scoring y en las redes neuronales.
- Construir el modelo basado en la información disponible.
- Validar el modelo que más se acomode al perfil de riesgo de empresas de financiamiento.

1.3 MARCO DE REFERENCIA

1.3.1 Antecedentes

Con respecto al problema planteado: Modelos de Calificación cuantitativa (scoring) para agilizar créditos y mitigar el riesgo, se logró hallar un Libro (Medición Integral del riesgo de crédito) y una Tesis de la Universidad EIA (Medición del riesgo crédito en Colombia hacia Basilea III), que representan publicaciones e investigaciones relacionados con este trabajo de grado.

Medición integral del riesgo de crédito por Alan Elizondo. Este libro fue publicado en 2004, en México, en el primer capítulo de este libro (Los métodos de calificación de cartera y su importancia para los paradigmas de medición de riesgo crédito), en este capítulo el autor se centra en el estudio de modelos teóricos, basado en modelos ya existentes para un análisis en profundidad de éstos mismos, los más importantes a criterio son: Modelos basados en la ponderación de factores de riesgo: Éste está basado en diferentes técnicas cualitativas y cuantitativas por ejemplo: Redes neuronales, Estadísticas de discriminación y “Z Score” de E.Altman 1968. Y el modelo de Merton para determinar probabilidades de impago: Éste fue realizado en 1974 y tiene en cuenta factores como la liquidez, pasivos y activos del cliente.

Se examinaron los esquemas de calificación de crédito con el propósito de determinar la calidad de los créditos, teniendo una valoración subjetiva como objetiva. Se puede decir que un sistema en “estado de arte” de calificación de riesgo crediticio, debe incluir los siguientes componentes (Elizondo, 2004):

- 1) Metodología Adecuada para los créditos
- 2) Definición objetiva y fácil de aplicar
- 3) Mecanismo eficiente de probabilidades de impago
- 4) Formalización del proceso que refleja la forma en que migran los créditos de una calificación a otra
- 5) Obtener estadísticas sobre tasas de recuperación de créditos.

VAR. Valor en Riesgo (abreviado VAR a partir de su expresión en inglés, Value at Risk) es un método para cuantificar el riesgo de un mercado, esto se puede lograr por medio de estadísticas tradicionales. Este concepto está muy relacionado con la posible pérdida, (dado que si no hay pérdida no hay riesgo), el riesgo se puede dividir en riesgo operativo y riesgo crédito, siendo este último el que interesa y se pretende estudiar (Noesis, 2005).

Las instituciones financieras manejan el scoring para reducir, no para eliminar el riesgo de impago, dado que es imposible eliminar completamente este riesgo, además, lo que se busca es saber si será un buen cliente (que paga y que lo hace a tiempo), a través del uso de parámetros (si es persona natural: sexo, edad, estado civil, grado de estudios, ocupación, etc.), arrojando como resultado un riesgo crediticio aceptable o negativo. Es muy usado por las instituciones financieras sobre todo para otorgar créditos a las

pequeñas y medianas empresas, lo más importante para el desarrollo de modelos es conocer perfectamente el público objetivo, para esto se necesita tener una base de datos extensa y antigua, después de tener esto el paso a seguir es tener un patrón con el que se va a calificar este riesgo crédito. Las grandes instituciones manejan modelos muy avanzados y cada entidad acomoda su modelo de acuerdo a su perfil de riesgo (Briceño, 2010).

Modelos. Existen varios modelos estadísticos de gran utilidad en la administración del riesgo de crédito bancario, como principal se encontró los modelos tradicionales (sistemas expertos y sistemas de calificación). Aquí se identifican dos corrientes en el tema: 1. Fundamental: parten de la proyección de variables económicas y financieras variando en el tiempo, esto involucra el criterio subjetivo de cada analista. 2. Ponderación de factores: se identifican como determinantes del incumplimiento de las obligaciones. Por otro lado los modernos (KMV = Kecholfer, Macqown y Vasicec) y el capital de riesgo crédito de países emergentes el cual fué creado en México CyRCE, tratan de captar la intuición de los expertos y sistematizarla aprovechando la tecnología, pues su campo de dominio es la inteligencia artificial por medio de la cual intentan crear sistemas expertos y redes neuronales.

Los principales factores para el otorgamiento de un crédito, son las cinco C's (Saavedra García & Saavedra García, 2010),

- Capacidad: la capacidad de pago del acreditado, este es el factor más importante en la decisión de un banco.
- Capital: Valores invertidos en el negocio del acreditado, así como obligaciones, un estudio de las finanzas.
- Colateral: Son todos aquellos elementos con los que dispone el acreditado para garantizar el cumplimiento del pago del pago en el crédito.
- Carácter: Son las cualidades de honorabilidad y solvencia moral que tiene el deudor para responder al crédito
- Condiciones: Son los factores exógenos que pueden afectar la marcha del negocio del acreditado.

En la tesis encontrada el enfoque principal el surgimiento de Basilea III y cómo se llegó a este nuevo acuerdo, que en general, busca fortalecer la solidez de los sistemas financieros, siendo la principal causa, la crisis del 2008 (Burbuja Inmobiliaria). Esta crisis de las hipotecas "subprime", ocasionó grandes impactos en los sistemas financieros de las economías mundiales especialmente en la Estadounidense, que fue, donde todo esto comenzó. Ahora, si se revisa cuidadosamente factor del riesgo crédito en este episodio mundial, se encuentra que tuvo un papel fundamental, porque define la posibilidad de que una entidad financiera tenga pérdidas, disminuyendo así el valor de sus activos, como consecuencia de que los clientes paguen morosamente o no paguen. Con el tratado de Basilea II, se tenían propuestas tres metodologías método IRB (Internal Rating Based), este método IRB dispone de modelos de calificación para la estimación de calificación de impago y aplica estándares (SBS) y el método IRB avanzado (similar al básico), éstos dos

anteriores se basan en los modelos teóricos: análisis discriminante, de respuesta binaria, VaR, Camel, sistemas expertos, árboles de decisión, scoring, entre otros (Osorio & Álvarez, 2011).

1.3.2 Que es el Scoring?

El scoring estadístico ha sido una técnica utilizada desde hace varias décadas para el sector financiero, utilizando la información histórica de las organizaciones para medir probabilísticamente el “score” o la calificación de los prestatarios. El scoring consiste básicamente en diseñar modelos probabilísticos utilizando las bases de datos con comportamientos de clientes pasados para tratar de determinar, cuantitativamente, el comportamiento de clientes futuros. En las microfinanzas es una técnica antigua. Hoy en día no solo se utiliza para empresas del sector financiero, sino en otro tipo de empresas que proporcionen productos y servicios diferentes, como por ejemplo servicios de telecomunicaciones. Surge la necesidad de las empresas de calificar a sus clientes, para disminuir la exposición a clientes riesgosos. De este modo se puede calificar clientes de manera cuantitativa utilizando los registros anteriores, para tratar de pronosticar el comportamiento del cliente dadas sus características (Schreiner, 2002).

Existen ciertos parámetros que permiten el cálculo estadístico de la calificación del cliente, como la edad, genero, ingresos, empleo, activos, etc. Estas variables son utilizadas para calificar al cliente utilizando estos mismos parámetros de clientes anteriores y tratar de pronosticar estadísticamente como serán los hábitos de pago del cliente en el caso de la solicitud de un crédito. Para realizar esto existen varias técnicas y modelos estadísticos para cuantificar el riesgo (Schreiner, 2002).

Probabilidad de default

La Probabilidad de default o de impago es una medida de calificación crediticia que se otorga internamente a un cliente con el objetivo de estimar su probabilidad de pago a mediano plazo (Inversores BBVA).

1.4.1.1 Ventajas del scoring estadístico

El scoring presenta ventajas para las organizaciones dado que, no solo cuantifica el riesgo como una probabilidad, sino que también evita los sesgos que puede tener la persona que estudia el crédito al momento de realizar un otorgamiento. El scoring es consistente en su calificación, y pronostica el comportamiento utilizando la información, por esto facilita el estudio de crédito y agiliza el otorgamiento. Dado que las técnicas no son subjetivas, la calificación resulta explícita al juzgamiento de la persona que estudia el crédito. El analista basa su criterio en el “score” y decide si otorgar el crédito de manera objetiva.

Además de agilizar el proceso, también ayuda a reducir el riesgo que toma la compañía al colocar dineros y mejora la calidad de la cartera. Como el scoring arroja un valor numérico, las políticas de cada empresa se ajustan a sus perfiles de riesgo y pueden tomar decisiones claras utilizando la calificación del modelo. También disminuye el tiempo

gastado en la cobranza dado que al disminuir el riesgo la institución asegura que se desgastara mucho menos en las gestiones de cobranza. Todas estas ventajas reflejan cómo se le puede dar un excelente uso a la información de toda la base de datos de clientes y generalizar los indicadores y parámetros para calificar a un solo individuo (Schreiner, 2002).

1.4.1.2 Desventajas del scoring estadístico

Existen varias desventajas en las técnicas de scoring, dado que si se utiliza de manera inadecuada esta herramienta, se puede aumentar la probabilidad de morosidad y de impago. Para modelar de manera adecuada una técnica de scoring se necesita, una base de datos amplia, lo cual es una desventaja para empresas medianas y pequeñas (ya que estas no cuentan con estas bases de datos amplias). Si la muestra para realizar el modelo no es lo suficientemente grande para modelar el riesgo crédito, no se alcanza a modelar de manera adecuada el riesgo que corre una institución al otorgar un crédito. El scoring funciona con probabilidades y no con certezas, esto quiere decir que el modelo supone que el futuro se comportara de manera idéntica al pasado y esto presenta cierta incertidumbre. El scoring también presenta un problema crítico para los departamentos de crédito y es el hecho de que el analista de crédito confíe demasiado en la calificación cuantitativa del modelo y se olvide de su análisis cualitativo, es por ello que esta técnica es muy susceptible al mal uso (Schreiner, 2002).

1.4.1.3 Técnicas para realizar el Scoring:

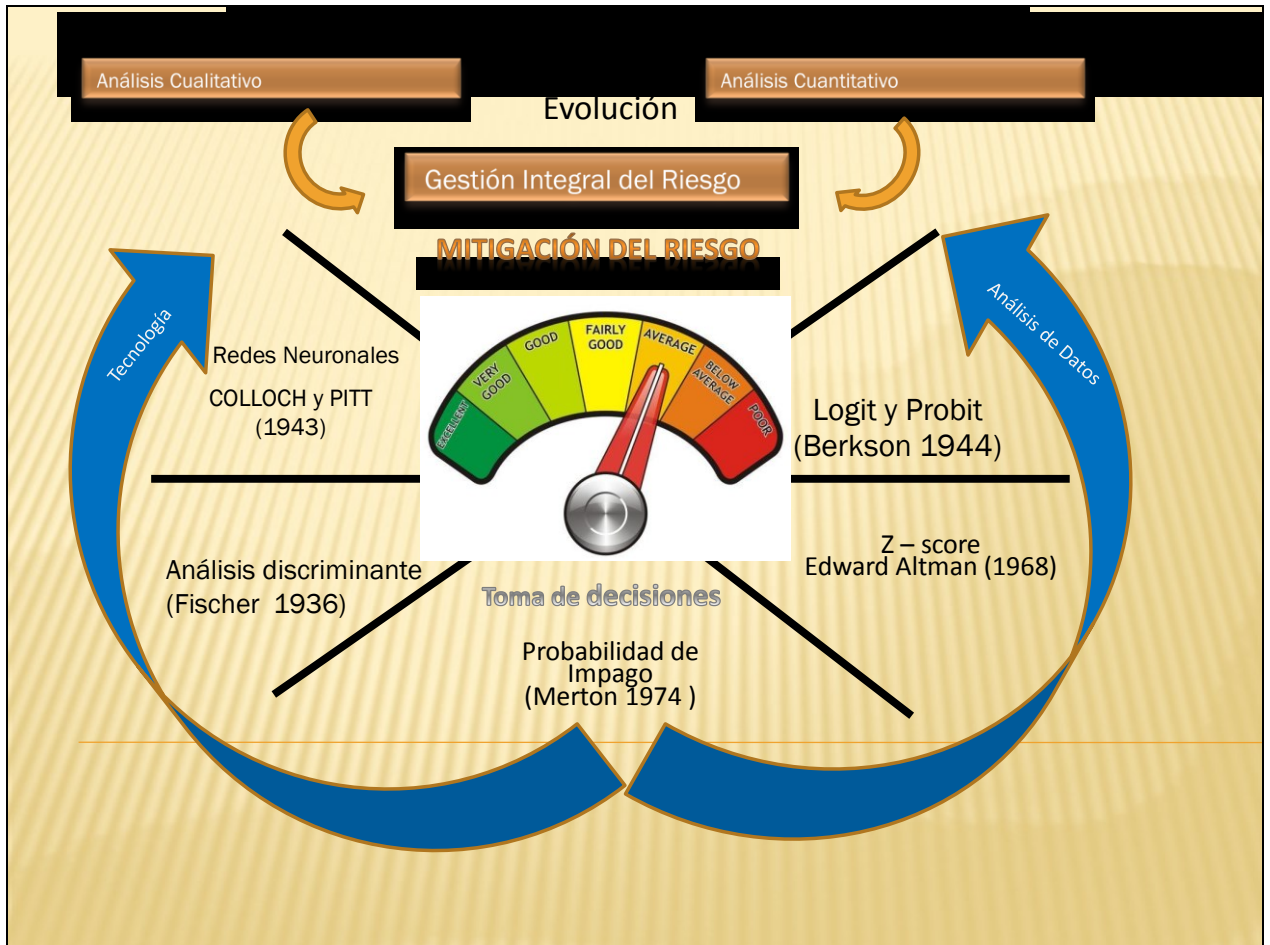


Figura 1. Modelos de Calificación. Fuente: Elaboración Propia.

Tradicionalmente existen varios métodos para emplear las técnicas estadísticas del scoring, pero la más reciente innovación en este campo corresponde al uso de redes neuronales para clasificar a los clientes. Las técnicas convencionales más utilizadas corresponden al análisis discriminante, análisis Probit y la regresión logística. Las técnicas se empezaron a utilizar inicialmente en los años 50s por los vendedores al por menor, pero a través de los años su mayor aplicación se encontró en el sector financiero que requerían clasificar los dineros colocados como mal crédito o buen crédito. La evaluación para los nuevos deudores es la aplicación más importante para los modelos de scoring, y se ha convertido en una de las herramientas más utilizadas por los bancos para minimizar el riesgo.

El análisis Probit consiste en asignarle a cada variable un coeficiente. El modelo utiliza una combinación lineal con un conjunto de variables independientes para transformarlas

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

en una probabilidad acumulada de una distribución normal. El modelo entrega una probabilidad de 0 a 100 % la cual es la clasificación del cliente.

El modelo de Logit, consiste en una combinación lineal de variables explicativas las cuales tienen diferentes cantidades de peso en el modelo. La variable de salida, o variable dependiente se ve expresada de forma logarítmica con una fórmula que entrega un valor con dos posibilidades, 0 o 1. En el caso de un análisis de crédito, 0 puede ser no otorgar el crédito y 1 puede ser si otorgar el crédito.

La técnica de análisis discriminante fue introducida por Fischer en 1936, y consiste en clasificar una población heterogénea en subconjuntos homogéneos y da ahí en adelante el proceso de decisión de estos subconjuntos. Se asume que por cada solicitud hay un número de variables explicativas y la idea del modelo es tratar de encontrar la combinación lineal que separe a los subconjuntos entre sí. Una vez separados los conjuntos se analiza cada uno por separado y el que más se ajuste a la solicitud, y de ahí se toma la decisión.

Las redes neuronales corresponden a un modelo no lineal que tiene la capacidad de modelar cualquier fenómeno. Este modelo nació inspirado por la manera cómo funciona el cerebro humano y sus neuronas, con su capacidad de adaptarse a sí mismo con la información entrante. El modelo consiste en varias capas que procesan la información y cada capa tiene las mismas variables, es por ello que se llama un modelo no lineal. La primera capa tiene una información de entrada y esta pasa por todas las capas hasta la capa de salida, es por ello que este modelo es multidimensional. Cada capa genera un “peso” para la información, el cual va pasando a través de las capas ocultas hasta que la capa de salida arroja el peso final el cual clasifica el cliente (Schreiner, 2002).

1.3.3 Modelos No Paramétricos

Red Neuronal

La investigación en redes neuronales comenzó en 1943 con las publicaciones de COLLOCH y PITT, el cual consistía en el principio matemático que simula la operación natural de una neurona. Las redes neuronales corresponden a un modelo no lineal que tiene la capacidad de modelar funciones complejas. Este modelo nació inspirado por la manera cómo funciona el cerebro humano y sus neuronas, con su capacidad de adaptarse con la información entrante. Inspirado en la sinapsis que hacen las neuronas del cerebro, el modelo consiste en varias “capas” que procesan la información y cada capa tiene las mismas variables, es por ello que se llama un modelo no lineal. La primera capa tiene una información de entrada y esta pasa por todas las capas hasta la capa de salida, es por ello que este modelo es multidimensional. Cada capa genera un “peso” para la información, el cual va pasando a través de las capas ocultas hasta que la capa de salida arroja el peso final el cual clasifica el cliente. La operación matemática de la red neuronal es relativamente simple. Al utilizar una función dada, la red procesa la información recibida de otras “neuronas”, y si la información entrante excede el “umbral de estímulo”, la información sigue fluyendo a través de la red.

La propiedad más importante de la red consiste en que está cambiando su función u operación interna, constantemente. Esto basado en la información de entrada de las otras

neuronas, por lo cual se dice que esta “aprendiendo”. La sinapsis juega un papel fundamental en este proceso de aprendizaje, a medida que es capaz de suprimir o amplificar la información entrante de otras neuronas. Esto quiere decir que la red se ajusta dándole “peso” a la información entrante, en el modelo neuronal esto se llama “aprender”.

La gran ventaja de este modelo, al ser no lineal, cuando faltan datos del cliente, los cuales se procesan como parámetros o variables, la red es capaz de acomodarse a la información y presentar un “score” para el cliente. Esto no pasa con las otras técnicas convencionales, en las cuales si se da el caso de que falta algún dato, la calificación se ve afectada y no se acomoda a la realidad del cliente.

Este modelo consiste de algoritmo, el cual está diseñado para minimizar los errores a través de las capas, pero la información empírica demuestra que no representa un gran avance para calificar los clientes dado que los modelos tradicionales presentan resultados similares o en algunos casos mejores. La red neuronal también tiene la desventaja de que dado la naturaleza de su funcionamiento, el modelo es capaz de predecir el comportamiento de un cliente pero en el caso de un rechazo a una solicitud de un crédito, no se puede distinguir cual fue el factor determinante que ocasiono el rechazo (Ochoa, Galeano , & Agudelo, 2010), (Ocaris & Fernández Horacio, 2007).

Programación Matemática

Es un campo en las matemáticas aplicadas, desde un punto de vista práctico, resolviendo problemas de optimización con recursos limitados (González, 2001).

Árboles de Clasificación (algoritmos recursivos partición)

Es una técnica binaria, el cual permite separar las observaciones que componen la muestra a grupos establecidos (Bonilla, Olmeda, & Puertas, 2003).

Los sistemas expertos

Es una técnica que simula el proceso de aprendizaje, lo cual le permite almacenar datos y conocimientos (Informática Integral, 2004).

1.3.4 Modelos Paramétricos

El modelo de probabilidad lineal

El modelo de probabilidad lineal es esencialmente un modelo de regresión, donde el valor de la variable dependiente es cero o uno, esto quiere decir que, mide la variación en la probabilidad de éxito, matemáticamente la regla de decisión puede ser:

$$y = b_1x_1 + b_2x_2 + \dots + b_kx_k + u \quad (1)$$

Dónde:

Y es la variable dependiente (el resultado de la decisión), x_i es la variable explicativa (criterio) i , b_i es el peso asignado a la variable explicativa, u es el error aleatorio, $P(u) = 0$.

Más adelante se definirá si cero es aprobado o es denegado.

En un vector de expresión la regla puede ser:

$$y = \mathbf{b}' \mathbf{x} + u \quad (2)$$

Dónde: \mathbf{x} es el vector de las variables explicativas, \mathbf{b}' la transpuesta del vector de parámetros de las variables explicativas.

Por consiguiente,

$$P(\mathbf{y}|\mathbf{x}) = \mathbf{b}' \mathbf{x} + u \quad (3)$$

La probabilidad condicional también puede interpretarse como la probabilidad de aprobación de la aplicación perteneciente al parámetro x del grupo. La probabilidad estimada de aprobación puede interpretarse de manera similar. Así, se obtiene la estimación de regresión de pesos, por lo que la probabilidad estimada de aprobación puede ser calculada para una nueva aplicación.

Cuando la decisión de préstamo se hace, y se obtiene la puntuación recibida, debe ser comparado con un límite de puntuación.

Limitaciones:

Por ser un modelo de probabilidad lineal la varianza de los errores no es constante y para ciertas combinaciones de las variables explicativas las probabilidades estimadas pueden ser mayores a 0 o menores a 1 (Halweb, 2003).

Modelo de Regresión Logística

El modelo de regresión Logit, parte de la hipótesis de que los datos, siguen la respectiva fórmula:

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{b}_0 + \mathbf{b}_1 x_1 + \mathbf{b}_2 x_2 \dots \mathbf{b}_k x_k + u = \mathbf{x}\mathbf{b} + u \quad (4)$$

Para simplificar la ecuación se nombra

$$Z = \ln\left(\frac{p}{1-p}\right) \quad (5)$$

Por lo tanto,

$$Z = b_0 + b_1x_1 + b_2x_2 \dots b_kx_k \quad (6)$$

Entonces:

$$\ln\left(\frac{p}{1-p}\right) = z + u \quad (7)$$

Siendo P la probabilidad de que el evento ocurra:

Despejando P, se obtiene

$$p = \left(\frac{e^z}{1+e^z}\right) \quad (8)$$

Obteniendo la función de distribución logística

$$F(x) = \left(\frac{e^x}{1+e^x}\right) \quad (9)$$

Al analizar esta ecuación se deduce que es un modelo de regresión no lineal, pero en realidad es lineal en escala logarítmica, partiendo de la ecuación inicial. Es decir, la resta de la probabilidad de que el evento ocurra, respecto a la probabilidad de que el evento no ocurra, es lineal en escala logarítmica.

$$\ln(p) - \ln(1 - p) = b_0 + b_1x_1 + b_2x_2 \dots b_k \quad (10)$$

En general, la gráfica de una función de distribución es:

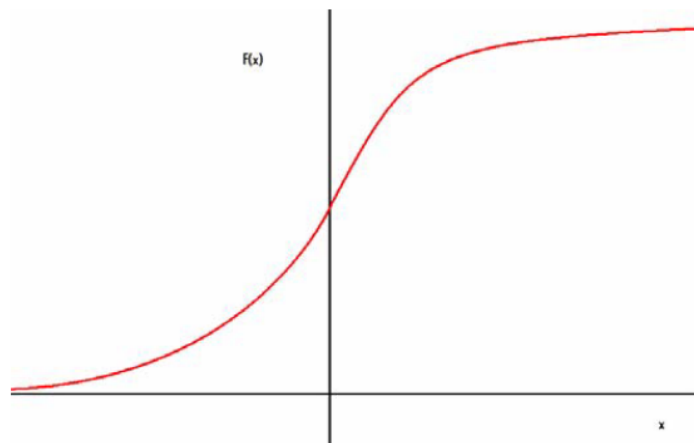


Figura 2. Función de distribución. Fuente: (Abuín, 2007)

Si utilizamos la función de distribución logística, el análisis se denomina Regresión Logística, y si utilizamos la función de distribución normal se denomina Regresión Probit (Abuín, 2007).

Modelo de Regresión Probit

El modelo Probit consiste en asignarle a cada variable un coeficiente. El modelo utiliza una combinación lineal con un conjunto de variables independientes para transformarlas en una probabilidad acumulada de una distribución normal. El modelo entrega una probabilidad de 0 a 100 %.

Debería adoptarse un modelo bajo el cual los valores de P_i estén restringidos al intervalo $[0,1]$. Una forma muy conveniente de restringir la forma funcional es la siguiente:

$$P_i(n_i) = F(b'\beta) + \varepsilon_i \quad (11)$$

En donde $F(b'\beta)$ es una función de distribución acumulada (FDA). La cual es una función diferenciable monótona creciente con dominio \mathbb{R} y rango $[0,1]$.

El modelo no lineal sería el siguiente:

$$Y_i = F(x_{ti}) + \varepsilon_i \quad (12)$$

Con:

$$\varepsilon_i = E(Y_i | X_i) - F(X_{ti}\beta) \quad (13)$$

Aunque son posibles varias alternativas de la FDA, sólo se considerarán dos: la normal y la logística.

Donde el modelo Probit se asocia con la distribución Normal

$$P = \phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz \quad (14)$$

Dónde: $\phi(z)$ es la función de densidad normal estándar, esto produce el modelo Probit. Si la función de distribución logística se selecciona para expresar la probabilidad p de aprobación, esta dará lugar al modelo Logit. En este caso:

$$P = \phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz = \frac{1}{1 + e^{-b'x}} \quad (15)$$

O alternativamente:

$$P = \frac{e^{b_1x_1+\dots+b_kx_k}}{e^{b_1x_1+\dots+b_kx_k}} \quad (16)$$

En contraste con la función de distribución normal, la función de distribución logística tiene una forma cerrada, lo que hace que el cálculo del modelo Logit sea mucho más simple que la del modelo Probit. Por lo general, los dos modelos se estiman utilizando el método de máxima verosimilitud, por lo que su aplicación informatizada es relativamente simple y de menor costo. Como estos modelos son ampliamente utilizados, un gran número de estudios han sido puestos en su aplicación y la experiencia adquirida en los consumidores, préstamos comerciales y agrícolas (Halweb, 2003), (Univeridad Tecnológica de la mixteca, 2009).

Máxima verosimilitud

El método de máxima verosimilitud se utiliza para estimar los coeficientes de un modelo logístico de regresión, en el que se calcula la probabilidad de que ocurra un determinado suceso mediante la siguiente ecuación:

$$P = 1/(1 + e^{-b_0 + b_1x_1 + \dots + b_kx_k})$$

Donde p es la probabilidad de que ocurra el suceso de interés y xi son los posibles factores (factores de riesgo) que se piensa que están relacionados con la probabilidad de que el suceso se produzca.

Ahora a partir de los datos de la muestra, para los que se ha observado si se ha producido o no el suceso, y a partir de los valores de los factores de riesgo en cada caso de la muestra, se trata de estimar los valores de los coeficientes (bi) en el modelo para cada factor de riesgo, lo que entre otras cosas permite calibrar el efecto de esos factores en la probabilidad de que el suceso ocurra. Si se denomina de forma compacta a esos coeficientes con la letra b (vector de valores), y dado que los valores de los factores x son conocidos para cada sujeto, la probabilidad p es función de los coeficientes b, y se representa como p=f (b).

Si p es la probabilidad de que ocurra el suceso, la de que NO ocurra será 1-p, y entonces en los sujetos en los que ocurrió el suceso vendrá dada por p(xi), mientras que para un sujeto en el que NO ocurre el suceso, se calcula como 1-p(xi). Siendo ambas expresiones función de b.

Si la muestra es aleatoria y las observaciones son independientes entre sí, la probabilidad de que un sujeto de la muestra experimente el suceso es independiente de lo que le ocurra a cualquier otro, por lo que la probabilidad conjunta se calcula como el producto de las probabilidades individuales y de esa forma se obtiene la función de verosimilitud, que tiene en cuenta todos los datos de forma global, y será función únicamente de los coeficientes. De igual manera que antes se calculará la derivada de esa función, se iguala

a cero y se obtienen los valores de los coeficientes que maximizan esa función. (Moliner, 2003).

Análisis discriminante

La técnica de análisis discriminante fue introducida por Fischer en 1936, y consiste en clasificar una población heterogénea en subconjuntos homogéneos y da ahí en adelante el proceso de decisión de estos subconjuntos. Se asume que por cada solicitud hay un número de variables explicativas, la idea del modelo es tratar de encontrar la combinación lineal que separe a los subconjuntos entre sí. Una vez separados los conjuntos se analiza cada uno por separado y del que más se ajuste a la solicitud, de ahí se toma la decisión.

La situación inicial se presenta con dos grupos, el primero, los solicitantes los cuales se les han aprobado el crédito, y la segunda, a los que se les ha negado el crédito. La tarea consiste en clasificar el solicitante al crédito con el vector $X = X_1 + X_2 + X_3 \dots + X_n$ representando las características del cliente. El análisis discriminante resuelve este problema al generar la llamada fórmula de discriminación γX , donde γ es el vector de coeficientes o "peso" asignado a cada X_n . El modelo determina estos valores al crear la mayor diferencia posible de los dos grupos de clientes. Se asume que el vector X , que contiene las características del cliente, es multivariado y tiene una distribución normal en los dos grupos. Cada grupo tiene asignados sus medias y sus covarianzas para cada parámetro. La p_i significa la probabilidad de que cada solicitante pertenezca al grupo i , mientras c_{ij} , representa el costo de clasificar mal a un cliente dentro de un grupo u otro.

Un solicitante con una combinación de datos X pertenece al grupo 1 si:

$$\gamma X \geq \alpha + \ln\left(\frac{c_{21} p_2}{c_{12} p_1}\right) \quad (17)$$

Donde,

$$\alpha = \ln(\gamma(\mu_1 + \mu_2)) / 2 \quad (18)$$

En todos los otros casos el solicitante debe estar clasificado en el grupo 2

La función de discriminación γX puede ser generada a través de medición lineal del vector X y la clasificación debe ser comparada con la calificación de corte

$$\text{"corte"} = \ln\left(\frac{c_{21} p_2}{c_{12} p_1}\right) \quad (19)$$

Si el solicitante está por encima del límite, será clasificado en el grupo G_1 , de manera contraria en el grupo G_2 .

Como este modelo toma como parámetro principal el vector x , por ello se le llama método de discriminación lineal. En el evento de que las covarianzas de los dos grupos sean diferentes ($\Sigma_2 \neq \Sigma_1$), la regla de clasificación será al cuadrado para X , es por ello que este modelo también es llamado análisis de discriminación cuadrática.

(Halweb, 2003) (Univeridad Tecnológica de la mixteca, 2009)

1.3.5 Curva ROC

La curva ROC por sus siglas en inglés (receiver operating characteristic o Característica operativa del receptor), es un procedimiento útil para evaluar el desempeño de esquemas de calificación en los cuales existe una variable con dos categorías, basado como una medida de bondad de ajuste, en la medida simultanea de sensibilidad, siendo esta los que verdaderamente son positivos y la especificidad siendo los que son verdaderos negativos, esto para cada punto de corte posible. Se calcula la especificidad y la sensibilidad porcentualmente.

Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo) (IBM, 2011).

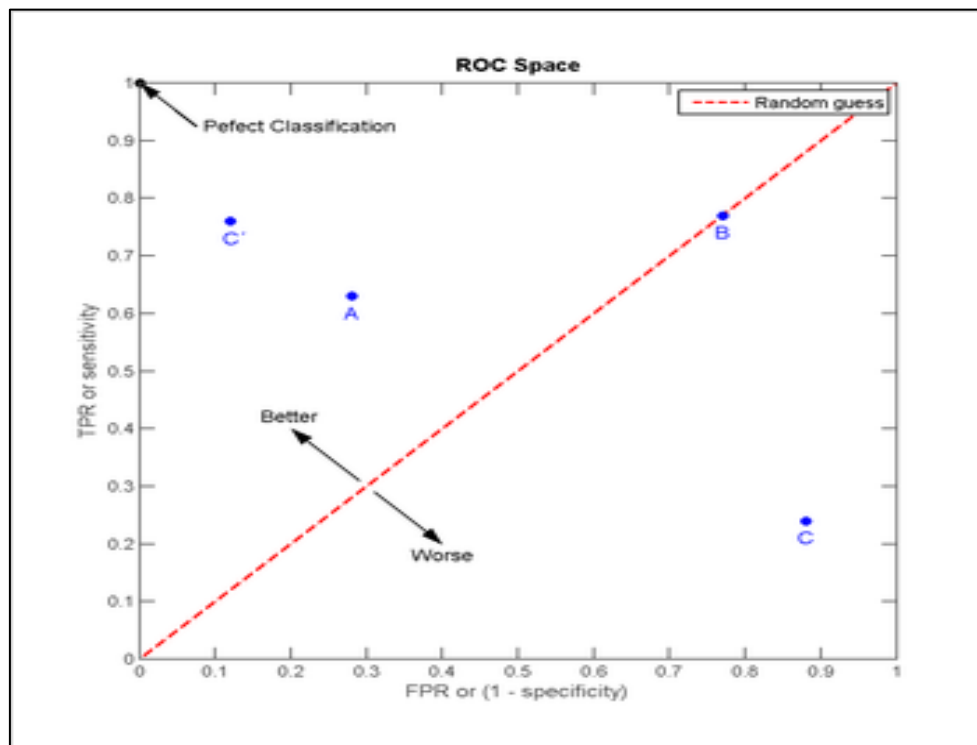


Figura 3. Curva Roc. Fuente: (IBM, 2011), (Wikipedia, 2009).

1.3.6 El Análisis Descriptivo

Corresponde al análisis que se le hace a la base de datos y las variables previamente establecidas. Este análisis además permite revisar y encontrar errores en la fase de inicio

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

al introducir los datos. También proporciona una idea de la forma que tienen los datos: su posible distribución de probabilidad con sus parámetros de centralización; media, mediana y moda; así como sus parámetros de dispersión; varianza, desviación típica, etc. (Dpto. de Matemática Aplicada, UCM, SF).

Las variables cualitativas

Corresponden a las que describen diferentes características o cualidades que describen la variable. Estas pueden ser ordinales o nominales; Las Nominales no admiten criterio de orden, ejemplo: Estado Civil = Soltero, Casado, Divorciado, Viudo; Las Ordinales tienen un orden ejemplo: puesto en una carrera = 1, 2, 3, etc. (Horra, SF).

Las variables cuantitativas

Las variables cuantitativas son las variables que pueden tomar valores numéricos, o cantidades medibles numéricamente. Estas pueden ser Discreta (Ordinal) o continua (Escala); La Discreta toma valores aislados, ejemplo: los hijos de 3 personas = 2, 0, 4; La continua admite un valor dentro de un intervalo por ejemplo: peso de 3 personas = 54,8kg, 65,3kg, 78,4 (Horra, SF).

Variables “Dummy” o Binarias

Son las variables cualitativas que solo pueden tomar los valores numéricos 0 o 1. Cualquier dato se puede codificar de esta manera, por ejemplo: si la persona entro en default se interpreta como un 0, y si la persona pago bien, se coloca un 1. (Martinez, 2011).

Correlación

Establece la relación que hay entre dos o más variables, esto quiere decir cómo afecta los cambios en una de las variables a la otra, si esto sucede se habla de correlación, hay diferentes tipos:

- **Directa**

Cuando al aumentar una de las variables la otra aumenta. La recta correspondiente a la nube de puntos de la distribución es una recta creciente.

- **Inversa**

La correlación inversa se da cuando al aumentar una de las variables, la otra disminuye

- **Nula**

Cuando no hay dependencia con ninguna de las variables. Dependiendo el grado de correlación entre las variables se clasifica: Fuerte (puntos en la gráfica cercanos), Débil (dispersión de puntos) o Nula (Ramón, SF).

2 METODOLOGÍA

La metodología de esta investigación es exploratoria y se desarrolló basado en una base de datos de una empresa del sector financiero. Esto con el objetivo de trabajar sobre datos reales. Para desarrollar este proyecto y alcanzar el objetivo general y los objetivos específicos, se ha planteado la siguiente metodología para ir paso a paso realizando las actividades y tener un orden de trabajo claramente definido. El proyecto se desarrollara en una serie de pasos.

Primero, se hará un análisis de la base de datos y de toda la información disponible suministrada por la empresa Sistecredito SAS. En esta fase se analizaran las características y variables que puedan describir el cliente y reflejar su comportamiento. Se realizara una depuración minuciosa de la base de datos para evitar datos atípicos y evitar sesgos en el modelo, esto para poder generalizar al máximo la conducta y los hábitos de pago del cliente teniendo en cuenta sus características y su tipología.

Una vez identificadas las variables y la información disponible, (inicialmente la base de datos de la empresa Sistecredito S.A.S contiene 140.000 datos) y luego de haber depurado y analizado la base de datos, se hará un análisis para identificar las técnicas que se ajusten a la información seleccionada, los modelos se entrenaran con el 70% de los datos y se validarán con el 30%. Se escogerán máximo dos técnicas a realizar para comparar los resultados y posteriormente realizar el análisis.

Luego se procederá a desarrollar el modelo utilizando los diferentes programas de software disponible para el estudio. Inicialmente se realizaran en SPSS de IBM versión 17, las primeras corridas. Se efectuaran varios intentos analizando la base de datos escogida hasta que se llegue a un modelo significativo. Documentando en cada caso como va evolucionando la construcción del modelo, e identificando las mejoras que se presentan con las medidas y correcciones que se hagan. Una vez se alcance algo significativo se construirá el modelo Logit con el mismo procedimiento utilizando el software de R Project versión 3.0 y el StatTools de Palisade versión 5.5, paralelamente se construirá la red neuronal utilizando el módulo de neuraltools versión 5.5 de Palisade y finalmente se compararan los resultados entre las técnicas.

Por último se comprobara la validez del modelo a través de la curva ROC y las matrices de error (matrices de clasificación). Se compararan los resultados arrojados por cada programa y por cada técnica. Luego de analizar estos resultados se desarrollara el modelo escogido en Excel, culminando así con el proyecto.

3 DESARROLLO DEL MODELO ESTADISTICO PARA MITIGAR EL RIESGO Y AGILIZAR EL ESTUDIO DE CREDITO

3.1 BASE DE DATOS

La base de datos de la empresa contiene toda clase de datos acerca del cliente, tanto datos explícitos, obtenidos desde la solicitud de crédito, como implícitos, obtenidos por medio de la minería de datos de la información histórica de los clientes.

Entrando en el análisis de la base de datos disponible, se analizan cuáles de los datos disponibles del cliente se podrían utilizar para realizar el análisis descriptivo identificando las variables que se utilizaran. La empresa Sistecredito SAS, ha facilitado la base de datos, con la cual se hace el análisis de que variables se pueden cuantificar en valores numéricos. La base de datos de clientes contiene toda la información, desde los ingresos, donde vive, si estudia o trabaja, o es independiente, si tiene historia crediticia, etc. En primera instancia, se escogieron los siguientes datos, que se lograron cuantificar en valores numéricos, codificándolos como valores numéricos y variables categóricas que permitan hacer el análisis.

| Numero | Variable | Descripción |
|--------|----------------|---|
| 1. | Edad | Se toma la fecha de nacimiento del cliente entregada en la solicitud, se compara con la fecha actual y se obtiene la edad de cliente. (Ejemplo: 28 Años, 42 Años....etc.) |
| 2. | Sexo | Si la persona es hombre, se interpreta como un 1, y si es mujer, se interpreta como un 0. (Ejemplo: Sandra=0, Manuel=1) |
| 3. | Municipio | Para este dato se utilizó el código interno utilizado por la empresa para cada municipio donde tiene presencia. (Ejemplo: Itagüí=63, Medellín=1) |
| 4. | Departamento | Se utilizó el código interno utilizado por la empresa para cada departamento donde tiene presencia. (Ejemplo: Antioquia=1, Risaralda=3) |
| 5. | Tiene historia | Esta variable se codifico de manera binaria, de |

| | | |
|-----|--|---|
| | | forma que si la persona tiene historia crediticia con la empresa, se interpreta como un 1, si no tiene, se interpreta como un 0. (Ejemplo: si tiene historia=1) |
| 6. | Cantidad de Créditos que ha tenido | Se toma el número entero mayor o igual a 0, como el número de créditos que ha tenido con la empresa. |
| 7. | Promedio de los créditos que ha tenido | se calcula como el valor total de crédito que ha solicitado, dividido por el número de créditos que ha tenido el cliente |
| 8. | Calificación interna del cliente entregada por la empresa (numero entre 1 y 5) | (Numero entre 1 y 5): corresponde a la calificación utilizada actualmente por la empresa para medir los hábitos de pago del cliente. |
| 9. | Cantidad de referencias comerciales | Este se muestra como un numero entero, mayor que 1 para la cantidad de referencias de este tipo que ha entregado el cliente, o se tiene de solicitudes anteriores. (Ejemplo: Cantidad de referencias comerciales=3) |
| 10. | Cantidad de referencias personales | Se muestra como un numero entero, mayor que 1 para la cantidad de referencias de este tipo que ha entregado el cliente, o se tiene de solicitudes anteriores. (Ejemplo: Cantidad de referencias personales =5) |
| 11. | Cantidad de referencias Laborales | Muestra como un numero entero, mayor que 1 para la cantidad de referencias de este tipo que ha entregado el cliente, o se tiene de solicitudes anteriores. (Ejemplo: Cantidad de referencias laborales = 2) |
| 12. | Estado civil | Se describe si la persona es casado, soltero, o divorciado, utilizando números para describir cada estado. (Ejemplo: Casado=0) |
| 13. | Labora | Describe por medio de una variable binaria si la persona labora o no. (Ejemplo: labora=1) |
| 14. | Años Labora | Cantidad de años que la persona ha laborado. Este dato se obtiene captando la fecha de ingreso al trabajo actual y comparándola con la fecha actual. (Ejemplo Años labora =3) |

| | | |
|-----|--------------------|--|
| 15. | Ingresos mensuales | La cantidad de dinero que le ingresa al cliente mensualmente, obtenida desde la solicitud. (Ejemplo: ingresos = 538,000 pesos) |
|-----|--------------------|--|

Tabla 1. Variables. Fuente: Elaboración Propia

Además, la base de datos contiene también información sobre el comportamiento de un cliente, indicando si éste pagó “bien” o “mal” su crédito (calificación), la cual sirve más adelante como variable de salida o variable de predicción para el modelo a construir.

Estos fueron los datos que se lograron cuantificar y codificar de manera numérica, de toda la información de los clientes encontrada en la base de datos disponible. A continuación se explica cómo se codificaron los datos para obtenerlos de manera numérica (S.A.S).

Análisis y depuración de la base de datos

Luego de tener en cuenta los diversos factores de cada variable, y los diferentes posibles modelos a desarrollar en el proyecto, se modificaron varias variables y datos encontrados en la base de datos completa, la cual consiste originalmente de 140 mil datos. Según la teoría de las diversas técnicas y metodologías de scoring, para que el modelo sea válido y describa correctamente el comportamiento del cliente, las variables no deben estar correlacionadas entre sí para que el modelo estadístico (lineal) funcione. Este criterio dio la pauta para eliminar las siguientes variables de la base de datos original

- Labora
- Tiene Historia
- Departamento

Estas variables se eliminaron debido a la relación que tienen con otras variables. En el caso de que la persona no labore, en Años Labora aparece un 0. Por esto la variable Años labora es suficiente para describir este aspecto del cliente. En el caso de si tiene historia, la variable calificación aparece como un -1 si la persona no tiene historia, por esto la calificación de si tiene historia tampoco se incluirá para efectos prácticos debido a que está incluida dentro de la variable calificación. Lo mismo sucede con la relación que existe entre departamento y municipio, por lo que solo se trabajará con el variable municipio.

Para evitar valores desfasados, se ordenó cada variable de mayor a menor, como por ejemplo los ingresos, promedio y edad. Esto para detectar valores extremos, como por ejemplo personas que registran con edad mayor a 100 años o personas con ingresos mayores a mil millones de pesos al mes. Estos valores se pueden dar por diversas causas como fraudes o errores en la digitación, por lo cual deben ser eliminados para no distorsionar el modelo.

Al eliminar los registros extraños, se pasó de tener 140 mil registros, a 90 mil registros. Esta gran disminución se da por la rigurosidad con la que se depuro la base de datos, al tratar de tener datos que se consideran dentro de lo “normal”.

Análisis de registros

Con el análisis de forma ordenada partiendo desde la base de datos, se hace la descripción del cliente acomodándose a las variables establecidas. Una vez introducidos los datos, se hace el análisis descriptivo para detectar posibles valores atípicos y errores que puedan afectar la validez del modelo. El análisis que se le hace a cada campo, o variable, de la base de datos, permite detectar valores fuera de rango, (como por ejemplo una persona con un ingreso de mil millones de pesos al mes). Este análisis inicial a la base de datos también proporciona la forma de los datos en cada campo, refiriéndose esto a la forma como posiblemente se distribuyan los mismos, ajustándose a las diferentes distribuciones de probabilidad. Cada distribución con sus parámetros de dispersión, como el promedio, varianza, desviación estándar, etc. (S.A.S).

Ajuste de distribución de probabilidad a los datos

Para realizar el análisis descriptivo, y posteriormente realizar la regresión logística para calcular el “score”, se debe determinar cómo se distribuyen los datos y analizar a cual distribución de probabilidad se ajusta mejor cada variable. Para este análisis se utilizara el programa @Risk, el cual utiliza por defecto el criterio de Chi-cuadrado para realizar los ajustes. Funciona solo para las variables que poseen valores en una escala que describan al cliente. Para el caso se utilizan las variables: Edad, Cantidad de créditos, Promedio de los créditos, referencias comerciales, referencias personales, Cantidad de años laborando, Ingresos y Calificación. Estas son las variables cuantitativas que se miden de manera discreta o continua.

Para las otras variables descriptivas del cliente, como Sexo, Estado Civil y municipio no se realizará este análisis, esto porque este tipo de variables son categóricas y se expresan de manera nominal y ordinal. Simplemente se hará un resumen de la cantidad de datos para cada uno de los posibles valores.

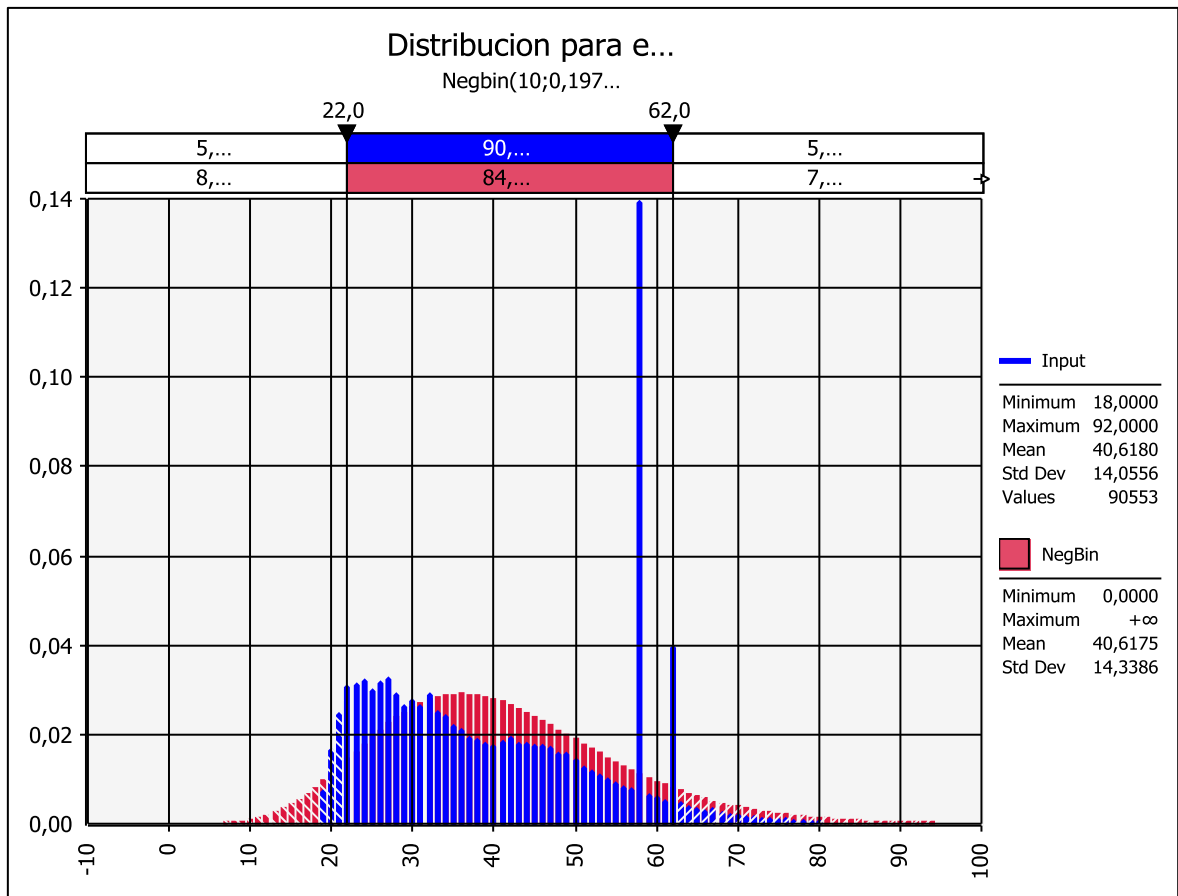


Figura 4. Distribución para Edad. Fuente: Elaboración Propia.

En el gráfico se puede ver que la mejor distribución que explica cómo se distribuyen los datos para la variable edad es la distribución binaria negativa. Con un mínimo de 18 años, un máximo de 92 años, una media de 40 años y una desviación estándar de 14 años. Según la distribución binaria negativa. En el gráfico, ajustando los datos a la distribución de probabilidad, se ve como el 84% de los datos están entre 22 y 62 años. Para los datos de entrada, el 90% de los datos se encuentran entre 22 y 62 años. Esta variable se analiza como una variable discreta, para efectos prácticos.

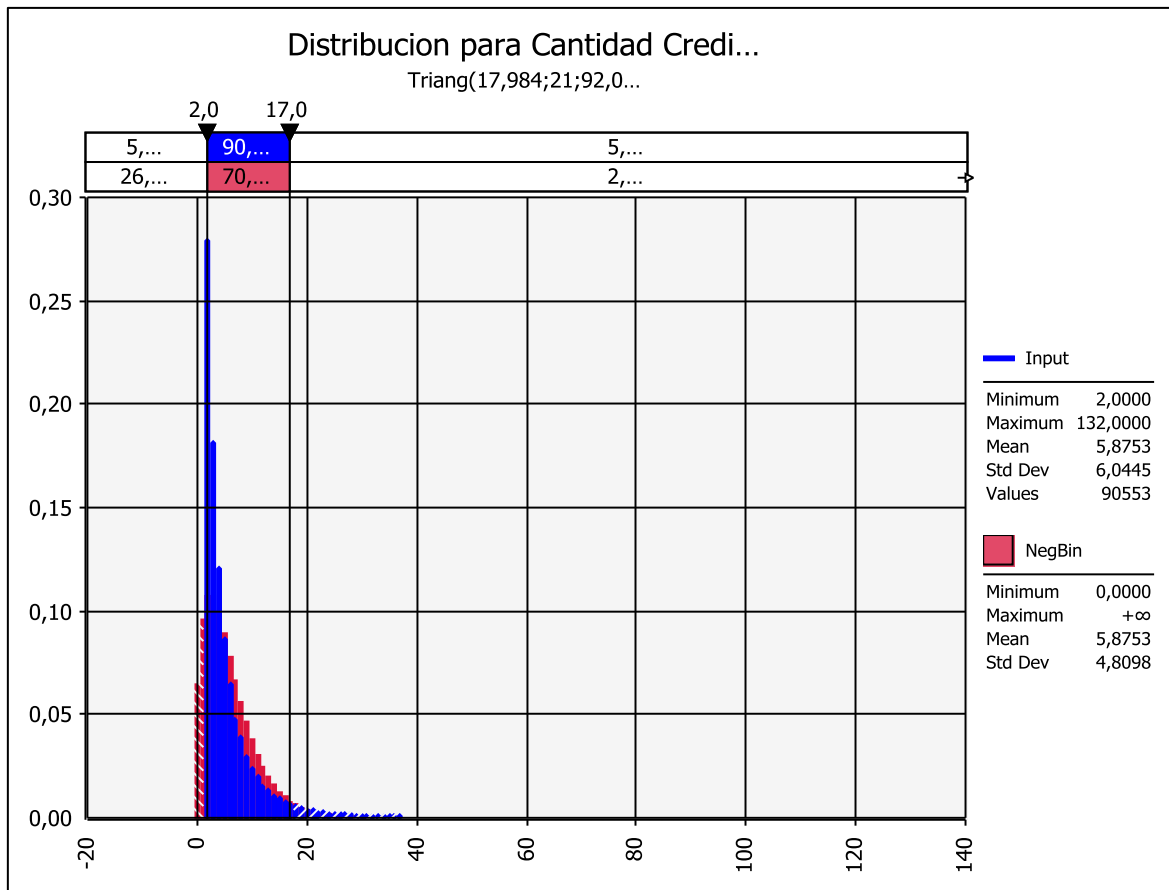


Figura 5. Distribución para Cantidad Créditos. Fuente: Elaboración Propia.

Para la cantidad de créditos que han tenido las personas, se observa que la mejor distribución de probabilidad para ajustarse es la binomial negativa, con un mínimo de 2 créditos y un máximo de 132 créditos. Además se ve como en los valores de entrada, el 90 % de los datos están entre 2 y 17 créditos. En el gráfico, ajustando los datos a la distribución de probabilidad, se ve como el 70,3% de los datos están entre 2 y 17 créditos.

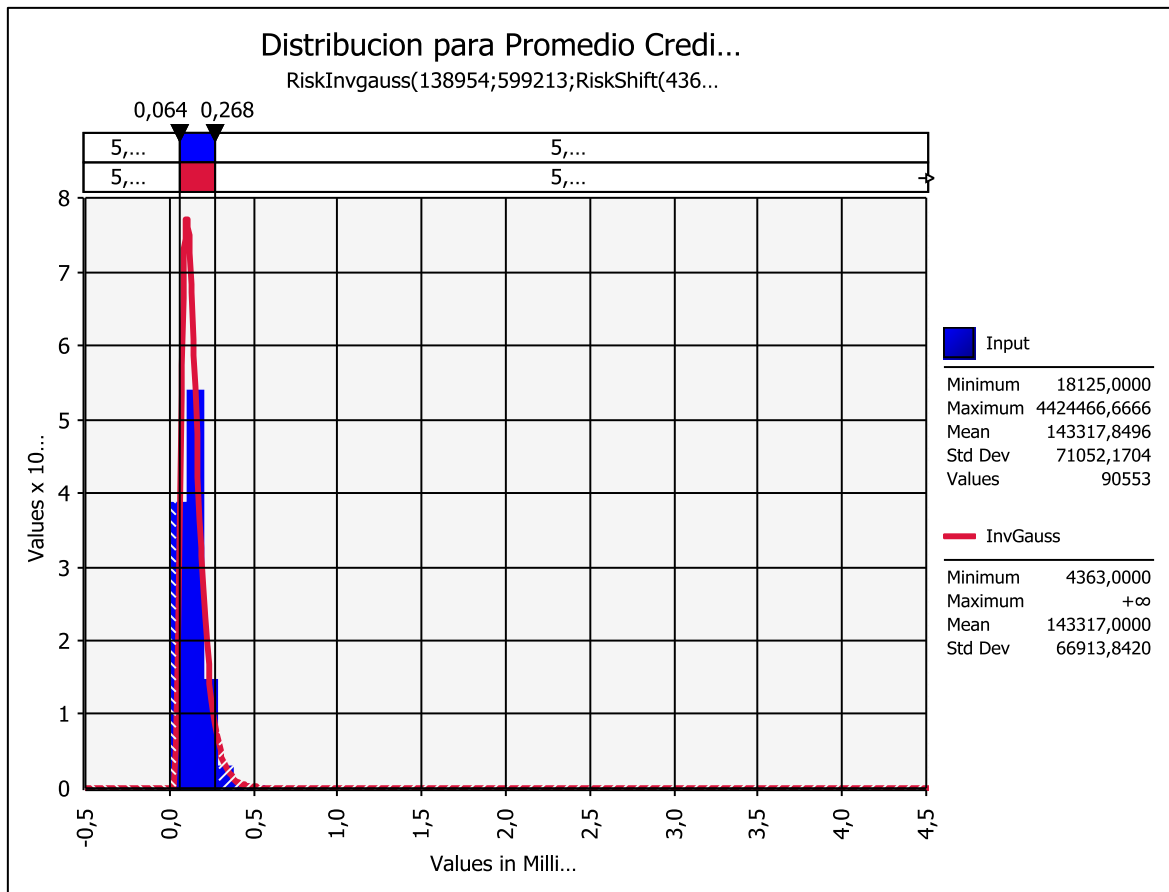


Figura 6. Distribución para Promedio. Fuente: Elaboración Propia.

Para la variable, promedio de créditos, se puede ver como la distribución que mejor se ajusta a los datos es una Gauss Inversa. El ajuste a la distribución, arroja una media de 143.317 pesos y una desviación de 66.913 pesos. Se puede ver como el 90 % de los datos están entre 64 000 pesos y 268 000 pesos, resaltando de esta forma que el objetivo de la empresa es, otorgar créditos de consumo de “bajo monto”.

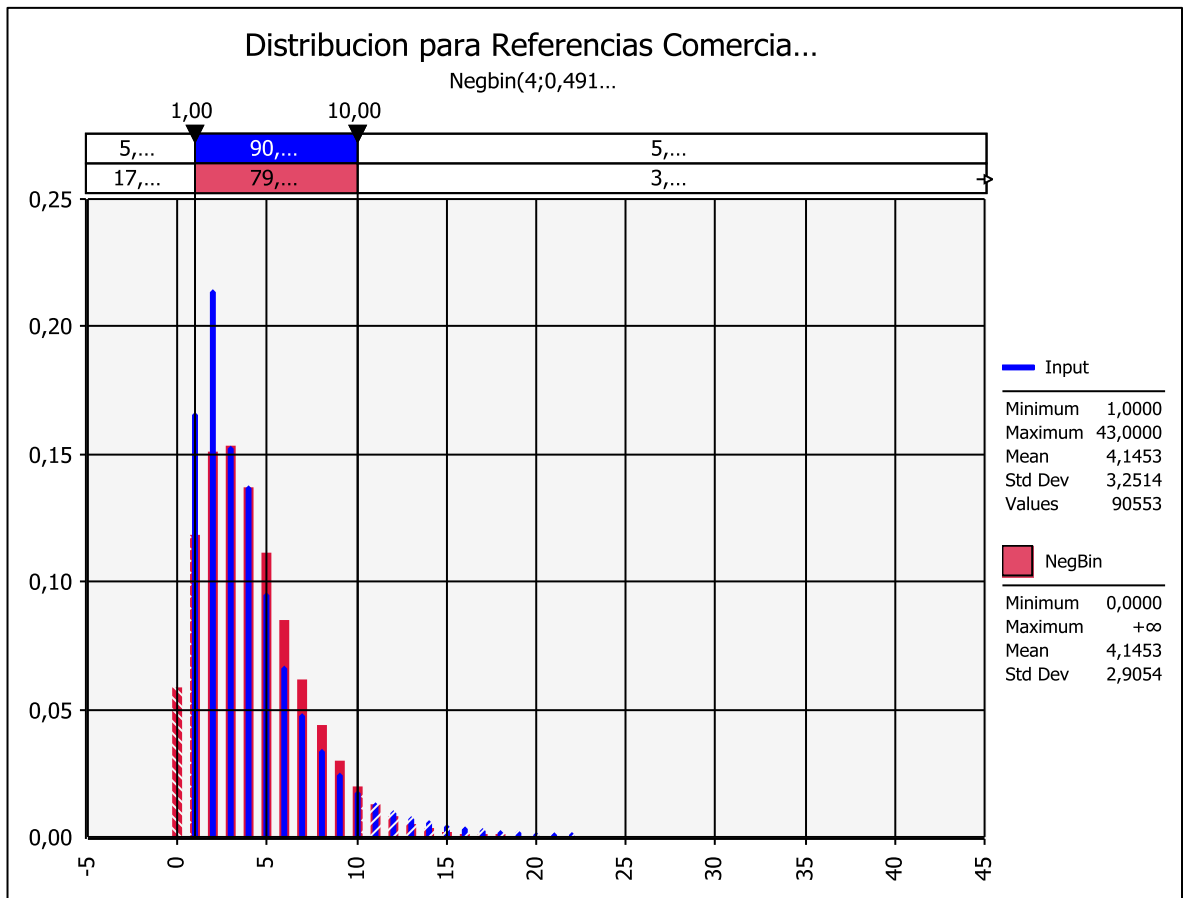


Figura 7. Distribución para Referencias Comerciales. Fuente: Elaboración Propia.

Para las referencias comerciales, la distribución que mejor se ajusta es la binomial negativa. Se observa como el 90% de los datos se encuentran entre 1 referencia y 10 referencias. Según esta distribución, la media es de 4 referencias y la desviación es de 3 referencias redondeando hacia arriba.

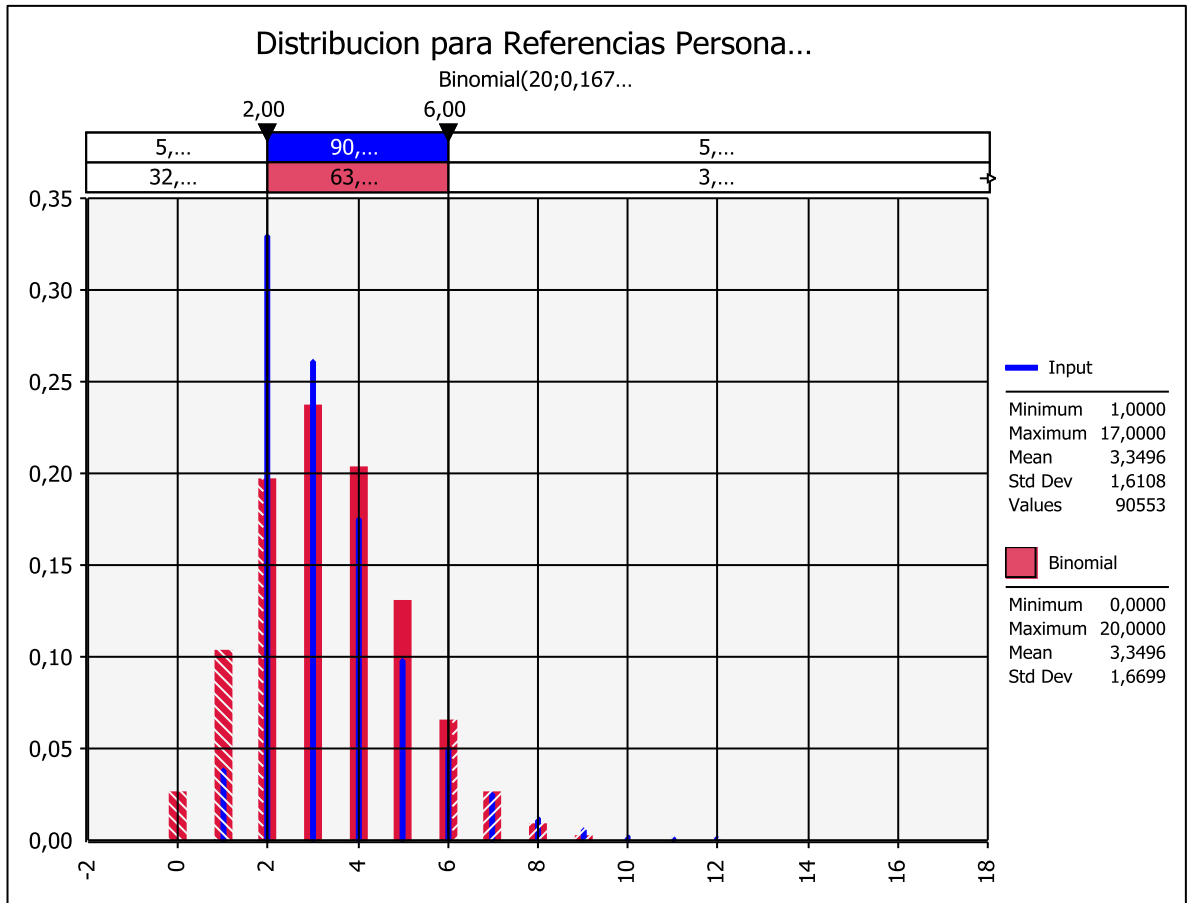


Figura 8. Distribución para Referencias Personales. Fuente: Elaboración Propia.

La distribución que mejor se ajusta para las referencias personales es la binomial. Con un mínimo de 1 referencia, y un máximo de 17 referencias, los datos arrojan una media de 3 referencias y una desviación estándar de 2 redondeando hacia arriba, según la distribución binomial.

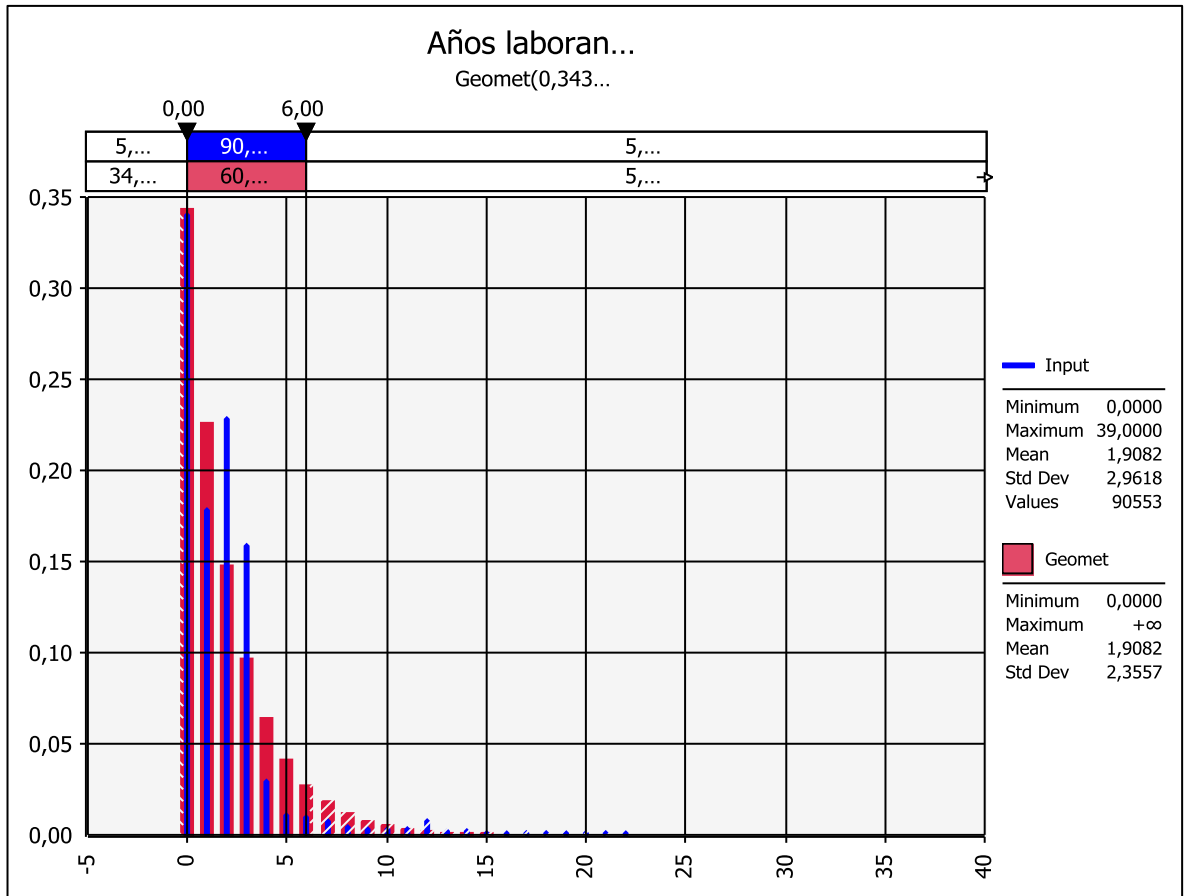


Figura 9. Distribución para Años laborando. Fuente: Elaboración Propia.

La cantidad de años que lleva laborando cada cliente, se distribuye de la mejor manera con la distribución geométrica. Por defecto se entiende que si la persona tiene 0 años laborado quiere decir que no labora. La distribución geométrica arroja una media de 1,9 años laborados, y una desviación estándar de 2,3 años.

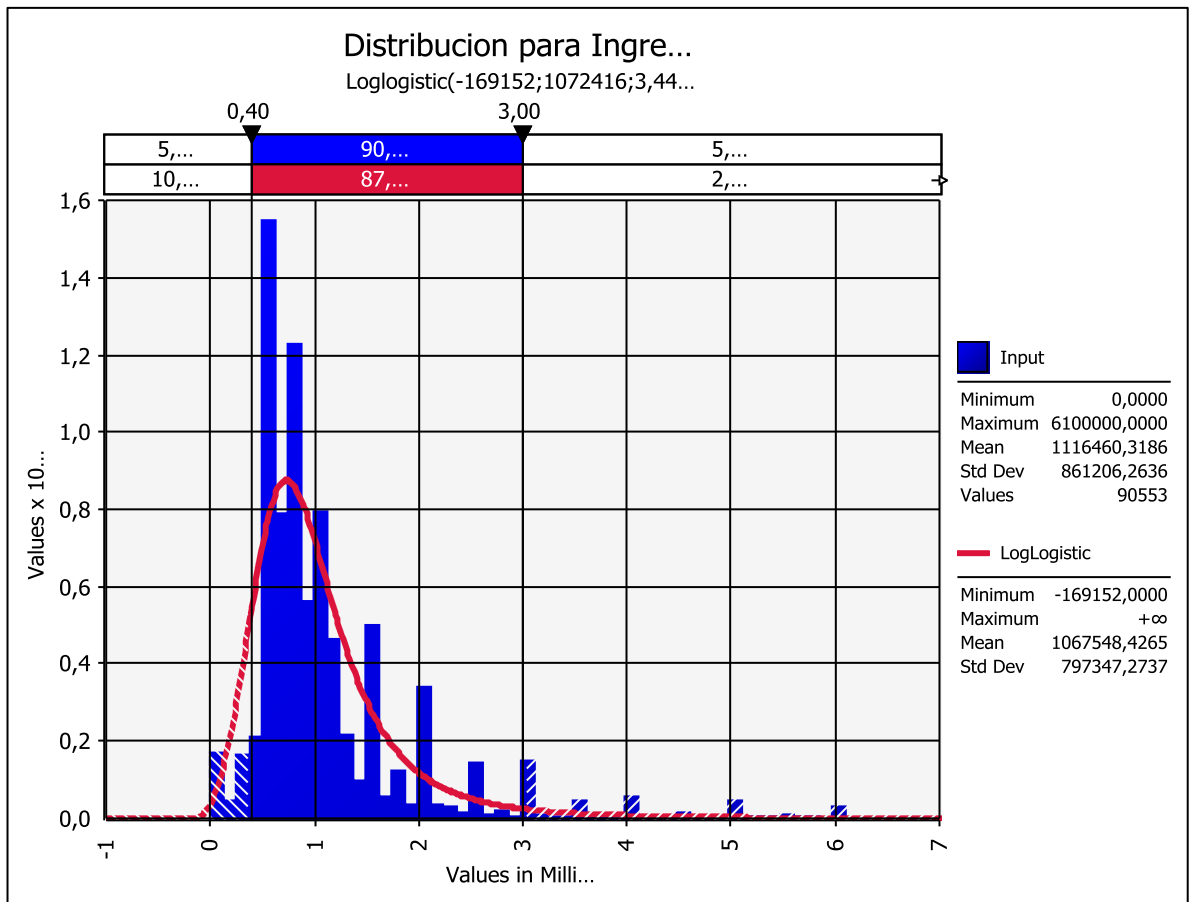


Figura 10. Distribución para Ingresos. Fuente: Elaboración Propia.

Para los ingresos, la distribución que mejor se ajusta es la Log logística. Los datos de entrada tiene en un mínimo de 0 ingresos y un máximo de 6,100,000 pesos. Una media de \$1.116.460 pesos, y una desviación de 861,206 pesos.

3.2 TÉCNICAS PARA DESARROLLAR EL SCORING

Como se revisó en el marco teórico cada técnica tiene sus ventajas y sus desventajas según el problema a desarrollar y los datos disponibles.

Luego de analizar cada una de las técnicas para desarrollar el modelo, y teniendo en cuenta varios criterios para la selección, se decidió escoger en primera instancia el modelo Logit. Se optó por la regresión logística dado que es el modelo tradicional utilizado por bancos y entidades financieras según los informes del Banco de la República. También se detecta que la estructura de la base de datos es adecuada para realizar este modelo dado que los datos se presentan de manera numérica y categórica, y existen muy pocos registros que tengan vacíos en alguna de las variables (González Arbelaez, 2012).

La técnica Logit resulta ser conveniente por la facilidad al momento de ejecutar el modelo y a su vez la agilidad para realizar cambios. Además, presenta la probabilidad de “éxito” del suceso, en este caso, la probabilidad de que un cliente pague su crédito. Esta probabilidad permite realizar diversos análisis como el perfil de riesgo que se ajusta a la empresa. Por estas razones, la técnica Logit resulta ser la opción más lógica y conveniente por encima de otras técnicas, teniendo en cuenta el fenómeno que se pretende estudiar.

Además, se desarrollará paralelamente una red neuronal para ir comparando resultados y escoger cuál de estos es más apropiado para mitigar el riesgo y el más óptimo según la información que se capta en el momento de realizar la solicitud del crédito. Se escoge la red neuronal para tratar de estudiar un modelo relativamente moderno y compararlo, en términos de los resultados, con la regresión logística, la cual es una técnica convencional.

3.3 CONSTRUCCION DEL MODELO

Para iniciar la construcción del modelo a desarrollar, se empezara por codificar las variables categóricas, luego se analizara la correlación entre las variables numéricas y por último se realizaran varias corridas del modelo, realizando ajustes hasta encontrar un modelo significativo.

3.3.1 Codificación de variables categóricas

Las variables escogidas, de carácter categóricas como son sexo, municipio y estado civil, se codifican de la siguiente manera para estructurar el modelo de manera adecuada. Para cada valor que pueda tomar una de estas variables, se genera una variable “dummy”, en la cual aparece un 1 o un 0 para cada valor.

Por ejemplo, para el estado civil, existen 6 diferentes estados en los cuales el cliente puede estar como son;

- Casado
- Soltero
- Unión libre
- Separado
- Viudo
- Otro

Esto significa que para codificar la base de datos, esta variable se convierte en 5 variables dummy. En la muestra de la base de datos aparecerá un 1 si la persona está en el estado soltero y 0 en todos los otros campos. Se convierte solo en 5 campos (n-1), siendo n la cantidad de posibles valores que puede tomar la variable. Aun cuando son 6 alternativas, si en los 5 campos aparecen 0, por defecto será la sexta alternativa, en este caso es el estado civil “otro”. Esto se hace de esta manera para simplificar el modelo y tener el menor número posible de columnas para no arriesgar la validez de los datos. De la misma forma se organiza la hoja con las otras 2 variables; sexo y municipio. A continuación se ilustra cómo queda la hoja de Excel para la variable Estado civil en amarillo.

| | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV | AW |
|----|---------|--------------|-----------|---------|-------|--------|---------|----------|-------|------------|---------|------------|----|
| 1 | Cantida | Promedio | Calificac | RefeCor | RefPt | Casado | Soltero | Separado | Viudo | UnionLibre | AñosLat | Ingresos | |
| 2 | 6 | \$ 19.786,67 | 3 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | \$ - | |
| 3 | 3 | \$ 25.445,00 | 0,44444 | 4 | 4 | 1 | 0 | 0 | 0 | 0 | 3 | \$ 600.000 | |
| 4 | 3 | \$ 26.502,00 | 4,8 | 4 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | \$ 250.000 | |
| 5 | 3 | \$ 26.718,33 | 3,05128 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 3 | \$ 100.000 | |
| 6 | 2 | \$ 28.400,00 | 3,55556 | 4 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | \$ - | |
| 7 | 2 | \$ 28.447,50 | 3,875 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 12 | \$ 800.000 | |
| 8 | 18 | \$ 28.612,50 | 4,40741 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 3 | \$ 309.000 | |
| 9 | 4 | \$ 29.669,50 | 1,69231 | 6 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | \$ 600.000 | |
| 10 | 3 | \$ 29.800,00 | 5 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | \$ - | |
| 11 | 2 | \$ 29.950,00 | 5 | 1 | 4 | 0 | 1 | 0 | 0 | 0 | 2 | \$ 750.000 | |
| 12 | 3 | \$ 29.966,67 | 4,5 | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 3 | \$ 600.000 | |
| 13 | 12 | \$ 30.468,33 | 4,63889 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 3 | \$ 600.000 | |
| 14 | 3 | \$ 30.633,33 | 4,83333 | 5 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | \$ 680.000 | |
| 15 | 3 | \$ 31.218,67 | 4,1 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | \$ 800.000 | |
| 16 | 6 | \$ 31.278,00 | 4,52778 | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 3 | \$ 795.000 | |
| 17 | 2 | \$ 31.395,00 | 5 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | \$ - | |
| 18 | 2 | \$ 31.447,50 | 4,5 | 3 | 2 | 0 | 1 | 0 | 0 | 0 | 3 | \$ 900.000 | |
| 19 | 4 | \$ 31.596,25 | 4,625 | 1 | 5 | 1 | 0 | 0 | 0 | 0 | 3 | \$ 200.000 | |

Figura 11. Variables Dummy. Fuente: Elaboración Propia.

Correlación entre variables escogidas

Correlations

| | | Edad | salario | cantidad |
|----------|---------------------|--------------------|--------------------|--------------------|
| Edad | Pearson Correlation | 1 | .211 ^{**} | .170 ^{**} |
| | Sig. (2-tailed) | | .000 | .000 |
| | N | 14847 | 14847 | 14847 |
| salario | Pearson Correlation | .211 ^{**} | 1 | .111 ^{**} |
| | Sig. (2-tailed) | .000 | | .000 |
| | N | 14847 | 14847 | 14847 |
| cantidad | Pearson Correlation | .170 ^{**} | .111 ^{**} | 1 |
| | Sig. (2-tailed) | .000 | .000 | |
| | N | 14847 | 14847 | 14847 |

** . Correlation is significant at the 0.01 level (2-tailed).

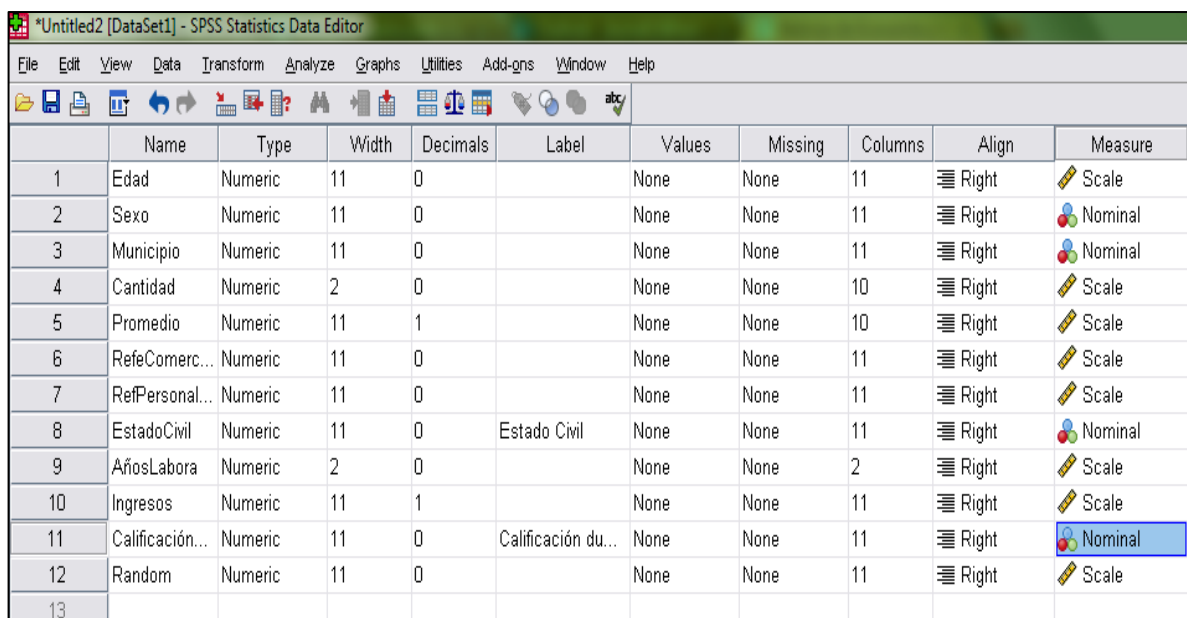
Figura 12. Matriz de Correlación. Fuente: Elaboración Propia

En esta tabla, se analiza la correlación que existe entre las variables numéricas de la base de datos. Se puede ver la matriz con la diagonal con los valores 1, y la correlación entre cada una con el resto de variables. La correlación es cero entre ellas como se observa en la tabla, por lo tanto las variaciones en cualquiera de estas variables no afectan las otras, lo cual es bueno para el modelo dado que cada variable de estas es independiente y servirá para medir la probabilidad de default.

3.3.2 Corridas para construir un modelo LOGIT

En primera instancia se utilizara el programa SPSS de IBM para iniciar con la construcción del modelo. Este programa permite al usuario realizar diferentes análisis con las técnicas estadísticas convenientes. Para esto se definen de qué tipo son las variables y se configuran varias características para el modelo. El output del modelo Logit es se expresa como la probabilidad de que la variable de salida tome el valor de 1, o en otras palabras, la probabilidad de que el suceso ocurra y el cliente pague bien. Cuando se programan las variables categóricas, se configura un campo llamado “measure”, o medida, en la cual se les asigna a estas variables la forma de medición, esta puede ser nominal, de escala u ordinal. El programa automáticamente genera las variables dummy con las distintas opciones para cada una, como se muestra a continuación:

CASO #1: Variables: Se configuraron de la siguiente manera en SPSS, ilustrada a continuación, y se definen por categoría (cualitativa o cuantitativa).



The screenshot shows the SPSS Statistics Data Editor window with a table of variables. The table has columns for Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, and Measure. The variables listed are: 1. Edad (Numeric, Width 11, Decimals 0, Measure Scale); 2. Sexo (Numeric, Width 11, Decimals 0, Measure Nominal); 3. Municipio (Numeric, Width 11, Decimals 0, Measure Nominal); 4. Cantidad (Numeric, Width 2, Decimals 0, Measure Scale); 5. Promedio (Numeric, Width 11, Decimals 1, Measure Scale); 6. RefeComerc... (Numeric, Width 11, Decimals 0, Measure Scale); 7. RefPersonal... (Numeric, Width 11, Decimals 0, Measure Scale); 8. EstadoCivil (Numeric, Width 11, Decimals 0, Label 'Estado Civil', Measure Nominal); 9. AñosLabora (Numeric, Width 2, Decimals 0, Measure Scale); 10. Ingresos (Numeric, Width 11, Decimals 1, Measure Scale); 11. Calificación... (Numeric, Width 11, Decimals 0, Label 'Calificación du...', Measure Nominal); 12. Random (Numeric, Width 11, Decimals 0, Measure Scale); 13. (Empty row).

| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure |
|----|-----------------|---------|-------|----------|--------------------|--------|---------|---------|-------|---------|
| 1 | Edad | Numeric | 11 | 0 | | None | None | 11 | Right | Scale |
| 2 | Sexo | Numeric | 11 | 0 | | None | None | 11 | Right | Nominal |
| 3 | Municipio | Numeric | 11 | 0 | | None | None | 11 | Right | Nominal |
| 4 | Cantidad | Numeric | 2 | 0 | | None | None | 10 | Right | Scale |
| 5 | Promedio | Numeric | 11 | 1 | | None | None | 10 | Right | Scale |
| 6 | RefeComerc... | Numeric | 11 | 0 | | None | None | 11 | Right | Scale |
| 7 | RefPersonal... | Numeric | 11 | 0 | | None | None | 11 | Right | Scale |
| 8 | EstadoCivil | Numeric | 11 | 0 | Estado Civil | None | None | 11 | Right | Nominal |
| 9 | AñosLabora | Numeric | 2 | 0 | | None | None | 2 | Right | Scale |
| 10 | Ingresos | Numeric | 11 | 1 | | None | None | 11 | Right | Scale |
| 11 | Calificación... | Numeric | 11 | 0 | Calificación du... | None | None | 11 | Right | Nominal |
| 12 | Random | Numeric | 11 | 0 | | None | None | 11 | Right | Scale |
| 13 | | | | | | | | | | |

Figura 13. Variables SPSS. Fuente: Elaboración Propia.

| | Edad | Sexo | Municipio | Cantidad | Promedio | RefeComerciales | RefPersonales | EstadoCivil | Año sLa bora | Ingresos | Calificacióndummy | Random |
|----|------|------|-----------|----------|----------|-----------------|---------------|-------------|--------------|-----------|-------------------|--------|
| 1 | 18 | 1 | 1 | 2 | 64342.0 | 1 | 4 | 1 | 0 | 550000.0 | 1 | 1 |
| 2 | 18 | 1 | 63 | 2 | 79350.0 | 1 | 2 | 1 | 0 | 510000.0 | 1 | 10 |
| 3 | 18 | 1 | 88 | 2 | 96910.0 | 1 | 3 | 1 | 0 | 500000.0 | 1 | 3 |
| 4 | 18 | 1 | 1 | 2 | 113070.0 | 3 | 4 | 1 | 0 | 800000.0 | 1 | 9 |
| 5 | 18 | 1 | 1 | 4 | 154250.0 | 2 | 3 | 1 | 0 | 5630000.0 | 1 | 9 |
| 6 | 19 | 1 | 1 | 3 | 43233.0 | 1 | 2 | 1 | 0 | 350000.0 | 1 | 9 |
| 7 | 19 | 1 | 63 | 2 | 43662.0 | 2 | 3 | 1 | 0 | 600000.0 | 1 | 10 |
| 8 | 19 | 1 | 1 | 2 | 44795.0 | 1 | 3 | 1 | 0 | 566700.0 | 1 | 10 |
| 9 | 19 | 1 | 2 | 3 | 49133.0 | 1 | 3 | 1 | 0 | 260000.0 | 1 | 6 |
| 10 | 19 | 1 | 134 | 2 | 51500.0 | 1 | 3 | 1 | 0 | 600000.0 | 1 | 8 |
| 11 | 19 | 1 | 1 | 5 | 52168.0 | 1 | 2 | 1 | 0 | 700000.0 | 1 | 2 |
| 12 | 19 | 1 | 46 | 2 | 53825.0 | 1 | 3 | 0 | 0 | 650000.0 | 1 | 5 |
| 13 | 19 | 1 | 1 | 7 | 54541.0 | 1 | 2 | 1 | 0 | 700000.0 | 1 | 9 |
| 14 | 19 | 1 | 4 | 11 | 55043.0 | 2 | 3 | 1 | 0 | 580000.0 | 1 | 7 |
| 15 | 19 | 1 | 133 | 2 | 55350.0 | 1 | 3 | 1 | 0 | 566700.0 | 1 | 5 |
| 16 | 19 | 1 | 59 | 4 | 57855.0 | 1 | 2 | 4 | 0 | 560000.0 | 1 | 3 |
| 17 | 19 | 1 | 1 | 5 | 58190.0 | 1 | 2 | 1 | 0 | 300000.0 | 1 | 9 |
| 18 | 19 | 1 | 59 | 3 | 58618.0 | 1 | 2 | 1 | 0 | 400000.0 | 1 | 3 |
| 19 | 19 | 1 | 1 | 2 | 58900.0 | 3 | 5 | 1 | 0 | 600000.0 | 1 | 4 |
| 20 | 19 | 0 | 67 | 2 | 61306.0 | 1 | 3 | 1 | 0 | 539000.0 | 1 | 10 |
| 21 | 19 | 1 | 67 | 2 | 61400.0 | 1 | 2 | 1 | 0 | 850000.0 | 1 | 1 |
| 22 | 19 | 1 | 63 | 3 | 61550.0 | 1 | 2 | 1 | 0 | 300000.0 | 1 | 7 |
| 23 | 19 | 0 | 1 | 2 | 62350.0 | 2 | 3 | 4 | 0 | 400000.0 | 1 | 5 |

Resultados SPSS:

Se toma la tabla de Hosmer y Lemeshow, porque esta prueba indica la significancia que tiene el modelo, ordenando el contenido en 10 grupos con un valor observado (valor real) y un valor esperado (valor que arroja el modelo). La relación entre estos valores, da la significancia en la prueba de Hosmer y Lemeshow. Por lo cual entre mayor diferencia haya entre lo observado y lo esperado, menor será la significancia.

| Hosmer and Lemeshow Test | | | |
|--------------------------|------------|----|------|
| Step | Chi-square | df | Sig. |
| 1 | 23.285 | 8 | .003 |

Figura 15. Tabla de Hosmer. Fuente: Elaboración Propia.

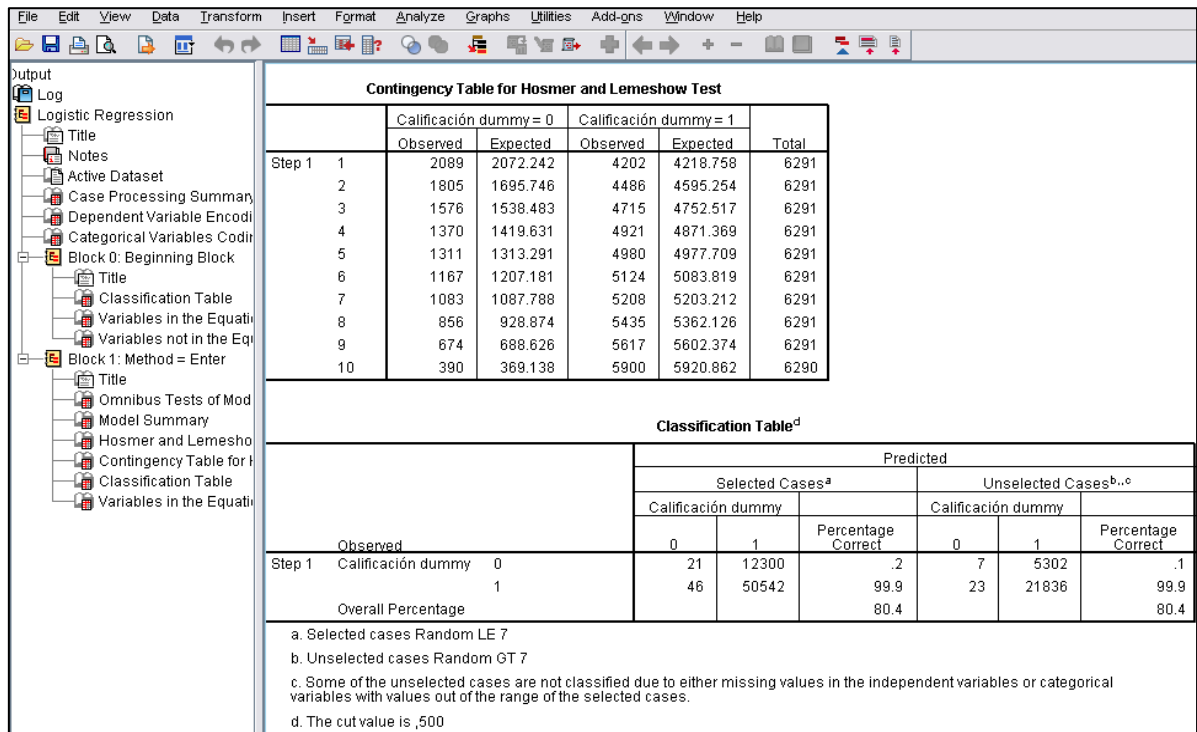


Figura 16. Tabla de Clasificación. Fuente: Elaboración Propia.

Análisis de Resultados Caso # 1:

Se toma el 70% de los datos para construir el modelo del análisis logístico (logit) y el 30% para el “test” o prueba de la base de datos. Es decir, el modelo se aplicará a datos (30%) que no participaron en la construcción del mismo.

Tabla de Clasificación: se divide en dos partes, en casos seleccionados y en casos no seleccionados. En esta tabla se observan los casos no seleccionados.

Tabla de Hosmer y Lemeshow: el valor Sig: (significancia), mide el desempeño del modelo, debe ser mayor al nivel de confianza (generalmente 5%), Entre mayor sea el valor Sig., mejor.

En este caso tenemos una significancia del 3%, por lo tanto el desempeño del modelo no es confiable.

Classification Table – tabla de clasificación: Esta tabla nos indica de los casos seleccionados y no seleccionados, que valores están calificados como “buenos” y “malos”, y el porcentaje calificado correctamente. “Bueno” y “malo” se refiere a clientes que pagaron “bien” o mal” su crédito.

En este caso, de los registros seleccionados el 99.99% originalmente era “bueno” y el modelo logit lo calificó como “bueno” y el 0.2% de los originalmente “malos”, los calificó como “malos”.

CASO # 2: Por la baja significancia obtenida en el caso anterior, se decidió eliminar las variables: referencias comerciales y referencias personales,

Estas variables se eliminan porque: eliminan al observarse que los valores son pocos confiables, lo que demuestra que a mayor número de referencias, no necesariamente implica que el cliente pague mejor.

Resultados SPSS:

| Step | Chi-square | df | Sig. |
|------|------------|----|------|
| 1 | 11.487 | 8 | .176 |

Figura 17. Tabla de Hosmer. Fuente: Elaboración Propia

| Observed | | Predicted | | | | | | |
|--------------------|--------------------|-----------------------------|----|--------------------|----------------------------------|----|--------------------|------|
| | | Selected Cases ^a | | | Unselected Cases ^{b, c} | | | |
| | | Calificación dummy | | Percentage Correct | Calificación dummy | | Percentage Correct | |
| | | 0 | 1 | | 0 | 1 | | |
| Step 1 | Calificación dummy | 0 | 19 | 12302 | .2 | 9 | 5300 | .2 |
| | | 1 | 50 | 50538 | 99.9 | 23 | 21836 | 99.9 |
| Overall Percentage | | | | | 80.4 | | | 80.4 |

a. Selected cases Random LE 7
b. Unselected cases Random GT 7
c. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.
d. The cutvalue is ,500

Figura 18. Tabla de Clasificación. Fuente: Elaboración Propia.

Análisis de Resultados Caso # 2:

Para este caso se logra evidenciar como mejora la significancia pasando de un 3% a un 17.6%, sin embargo la tabla de clasificación no mejora.

Caso # 3: Para este caso las variables cuantitativas (Edad, Ingresos, Promedio), se toman en rangos, formando grupos homogéneos. Estos rangos son agrupaciones de datos para convertir variables numéricas en variables categóricas. Los rangos se construyen de manera que las personas en cada grupo (rango) sean lo más parecidas entre sí. Los rangos ayudan a disminuir la desviación estándar entre las variables, y para correr el modelo se toman las variables nominales anteriormente mencionadas.

| Edad | Codigo | Ingreso | Codigo | Promedio | Codigo |
|---------|--------|-----------|--------|-----------|--------|
| 18-24 | 1 | 0-299 | 1 | 0 - 50 | 1 |
| 25-27 | 2 | 300-499 | 2 | 51 - 100 | 2 |
| 28-34 | 3 | 500-599 | 3 | 101 - 150 | 3 |
| 35-42 | 4 | 600-749 | 4 | 151 - 200 | 4 |
| 43-49 | 5 | 750-799 | 5 | 201 - 250 | 5 |
| 50-60 | 6 | 800-999 | 6 | 251 - 300 | 6 |
| 61-70 | 7 | 1000-1199 | 7 | 301 - 500 | 7 |
| 71----- | 8 | 1200-1499 | 8 | 501 ----- | 8 |
| | | 1500-1999 | 9 | | |
| | | 2000-2499 | 10 | | |
| | | 2500---- | 11 | | |

Figura 19. Tabla de Rangos. Fuente: Elaboración Propia.

Por ejemplo, como se ve en la tabla anterior, se asume que una persona de 18 años tendrá el mismo comportamiento que una de 24 años, y así mismo se define para ingreso y para promedio.

Resultados SPSS:

| Hosmer and Lemeshow Test | | | |
|--------------------------|------------|----|------|
| Step | Chi-square | df | Sig. |
| 1 | 7.888 | 8 | .444 |

| Contingency Table for Hosmer and Lemeshow Test | | | | | | |
|--|----|------------------------|----------|------------------------|----------|-------|
| | | Calificación dummy = 0 | | Calificación dummy = 1 | | Total |
| | | Observed | Expected | Observed | Expected | |
| Step 1 | 1 | 2010 | 2033.430 | 4281 | 4257.570 | 6291 |
| | 2 | 1710 | 1673.858 | 4581 | 4617.142 | 6291 |
| | 3 | 1560 | 1522.435 | 4732 | 4769.565 | 6292 |
| | 4 | 1383 | 1407.380 | 4908 | 4883.620 | 6291 |
| | 5 | 1297 | 1307.563 | 4996 | 4985.437 | 6293 |
| | 6 | 1238 | 1208.309 | 5053 | 5082.691 | 6291 |
| | 7 | 1112 | 1096.678 | 5180 | 5195.322 | 6292 |
| | 8 | 897 | 947.188 | 5394 | 5343.812 | 6291 |
| | 9 | 718 | 718.250 | 5573 | 5572.750 | 6291 |
| | 10 | 396 | 405.908 | 5890 | 5880.092 | 6286 |

| Classification Table ^c | | | | | | | |
|-----------------------------------|--------------------|-----------------------------|-------|--------------------|-------------------------------|-------|--------------------|
| | | Predicted | | | | | |
| | | Selected Cases ^a | | | Unselected Cases ^b | | |
| | | Calificación dummy | | Percentage Correct | Calificación dummy | | Percentage Correct |
| Observed | Calificación dummy | 0 | 1 | | 0 | 1 | |
| Step 1 | Calificación dummy | 0 | 12317 | 0 | 5309 | 0 | |
| | | 4 | 50584 | 100.0 | 21858 | 100.0 | |
| | Overall Percentage | 4 | | 80.4 | 1 | 80.5 | |

a. Selected cases Random LE 7
b. Unselected cases Random GT 7
c. The cut value is ,500

Figura 20. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Análisis de Resultados Caso # 3:

Para este caso se logra evidenciar como mejora la significancia pasando de un 17.6% a un 44.4%, sin embargo la tabla de clasificación no mejora.

CASO # 4: Para este caso se decide dejar las variables: Edad, Sexo, Cantidad, Estado Civil y Crédito, que según la empresa Sistecrédito S.A.S son las más “confiables”.

Resultados SPSS:

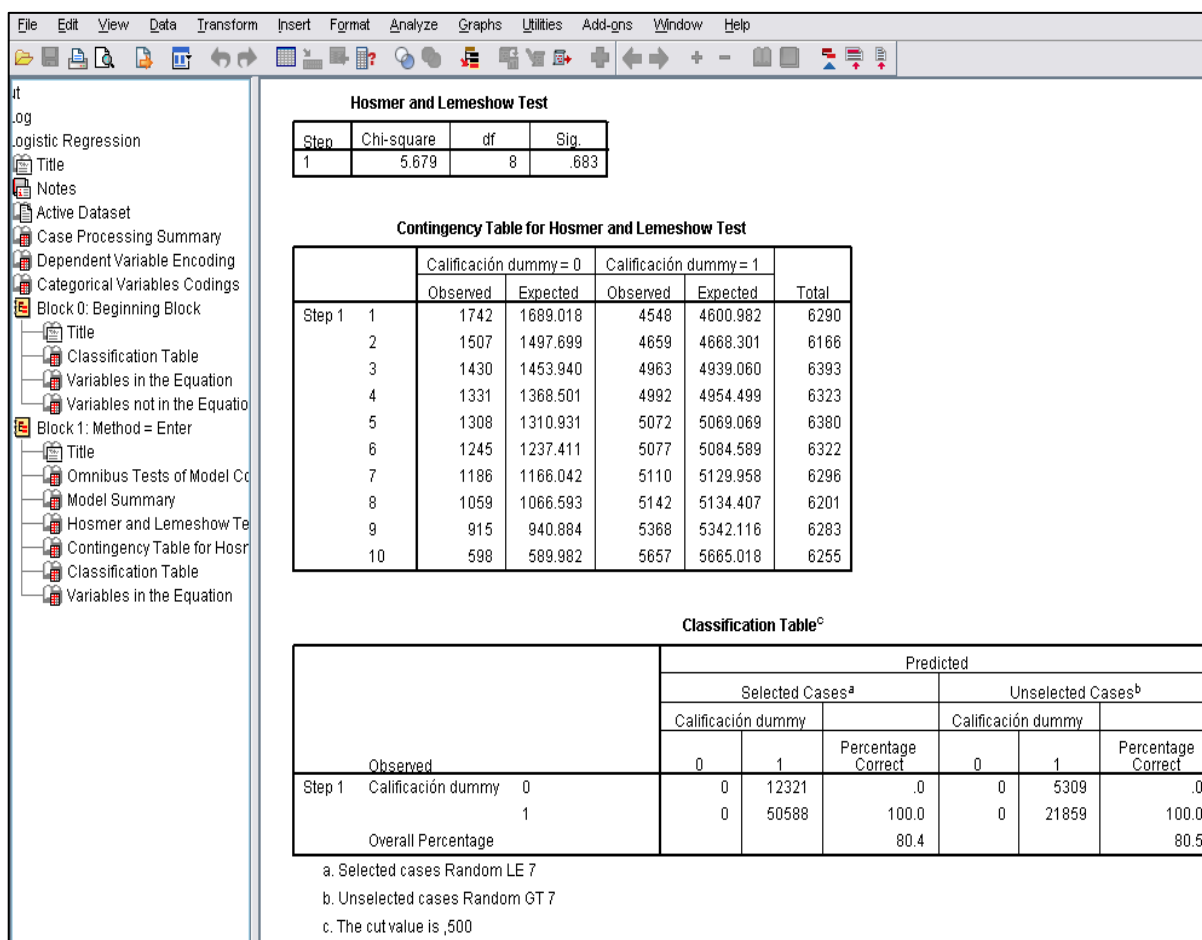


Figura 21. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia.

Análisis de Resultados Caso # 4:

Para este caso se logra evidenciar como mejora la significancia pasando de un 44.4% a un 68.3%, sin embargo la tabla de clasificación no mejora.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

CASO # 5: Como en los casos anteriores se puede observar que el modelo no tiene suficientes herramientas para calificar los créditos “malos”. Por tanto, se decidió cambiarla relación que existe entre los casos buenos y malos, es decir, como estaba la base de datos anteriormente, existían 1 caso malo por cada 7 casos buenos. Para equiparar esta relación se tomó una muestra aleatoria de casos buenos, donde la base quedara de menor tamaño con una relación de 1 caso malo por cada 2 casos buenos. Se decidió hacer esto para ver si el modelo sería capaz de calificar mejor a los casos malos, donde existía el problema en las anteriores corridas.

Resultados SPSS:

| Step | Chi-square | df | Sig. |
|------|------------|----|------|
| 1 | 6.983 | 8 | .539 |

Figura 22. Tabla de Hosmer. Fuente: Elaboración Propia.

| Observed | | | Predicted | | | | | |
|--------------------|--------------------|---|-----------------------------|-------|--------------------|-------------------------------|------|--------------------|
| | | | Selected Cases ^a | | | Unselected Cases ^b | | |
| | | | Calificación dummy | | Percentage Correct | Calificación dummy | | Percentage Correct |
| | 0 | 1 | | 0 | 1 | | | |
| Step 1 | Calificación dummy | 0 | 1860 | 10480 | 15.1 | 784 | 4506 | 14.8 |
| | | 1 | 1557 | 18790 | 92.3 | 671 | 7853 | 92.1 |
| Overall Percentage | | | | | 63.2 | | | 62.5 |

- a. Selected cases Random LE 7
 b. Unselected cases Random GT 7
 c. The cutvalue is ,500

Figura 23. Tabla de Clasificación. Fuente: Elaboración Propia.

Análisis de Resultados Caso # 5:

Una vez se realizó la regresión, se encontró que el entrenamiento con esta menor relación, alcanza a mejorar un poco la forma como se clasifican los casos malos, pero esto tiene un costo. Al haber eliminado los registros de casos buenos para alcanzar la relación deseada, también se alteró la forma como se clasificaban los casos buenos. Anteriormente calificaba el 100% de los casos buenos como buenos, mientras en esta corrida bajó a 92%. Además, también se puede ver como aumenta el porcentaje de clasificación de casos malos de 0% a 15%. Es decir, ahora el modelo es capaz de identificar de todos los registros que representan “malos”, un 15% como malos (calificación correcta) y un 85% como buenos (calificación incorrecta).

CASO #6: Analizando la base de datos se puede evidenciar que solo Medellín compone el 70% de los registros, por lo tanto se genera un sesgo, lo que quiere decir que el modelo interpreta todos los clientes por fuera de Medellín como “buenos”. Por esto se corre el modelo solo con el municipio Medellín. Se modifica la relación entre “buenos y malos “para un total de 60% de la base “buenos” y 40% de la base “malos”, obteniendo así los siguientes resultados:

| Hosmer and Lemeshow Test | | | |
|--------------------------|------------|----|------|
| Step | Chi-square | df | Sig. |
| 1 | 17.918 | 8 | .022 |

| Contingency Table for Hosmer and Lemeshow Test | | | | | | |
|--|----|-------------|----------|-------------|----------|-------|
| | | Credito = 0 | | Credito = 1 | | Total |
| | | Observed | Expected | Observed | Expected | |
| Step 1 | 1 | 846 | 814.505 | 189 | 220.495 | 1035 |
| | 2 | 710 | 701.255 | 325 | 333.745 | 1035 |
| | 3 | 575 | 591.228 | 460 | 443.772 | 1035 |
| | 4 | 475 | 496.257 | 560 | 538.743 | 1035 |
| | 5 | 404 | 423.162 | 631 | 611.838 | 1035 |
| | 6 | 339 | 350.961 | 696 | 684.039 | 1035 |
| | 7 | 272 | 282.518 | 763 | 752.482 | 1035 |
| | 8 | 246 | 216.757 | 789 | 818.243 | 1035 |
| | 9 | 149 | 149.600 | 886 | 885.400 | 1035 |
| | 10 | 88 | 77.757 | 943 | 953.243 | 1031 |

| Classification Table ^c | | | | | | | |
|-----------------------------------|--------------------|-----------------------------|------|--------------------|-------------------------------|------|--------------------|
| | Observed | Predicted | | | | | |
| | | Selected Cases ^a | | | Unselected Cases ^b | | |
| | | Credito | | Percentage Correct | Credito | | Percentage Correct |
| | | 0 | 1 | | 0 | 1 | |
| Step 1 | Credito 0 | 2241 | 1863 | 54.6 | 954 | 822 | 53.7 |
| | 1 | 1086 | 5156 | 82.6 | 524 | 2201 | 80.8 |
| | Overall Percentage | | | 71.5 | | | 70.1 |

a. Selected cases rand LE 7
b. Unselected cases rand GT 7
c. The cutvalue is ,500

Figura 24. Tabla de Hosmer y Tabla de Clasificación. Fuente: Elaboración Propia.

Análisis de Resultados Caso # 6

Considerando el caso anterior, se puede observar que la matriz de calificación mejoró considerablemente. El modelo califica un 54.6% correctamente los casos malos y un 82.6% correctamente los casos buenos. Estos resultados representan “algo significativo” y se puede decir que el modelo ya es capaz de juzgar el cliente como bueno y malo de una manera relativamente acertada.

Curva ROC caso #6

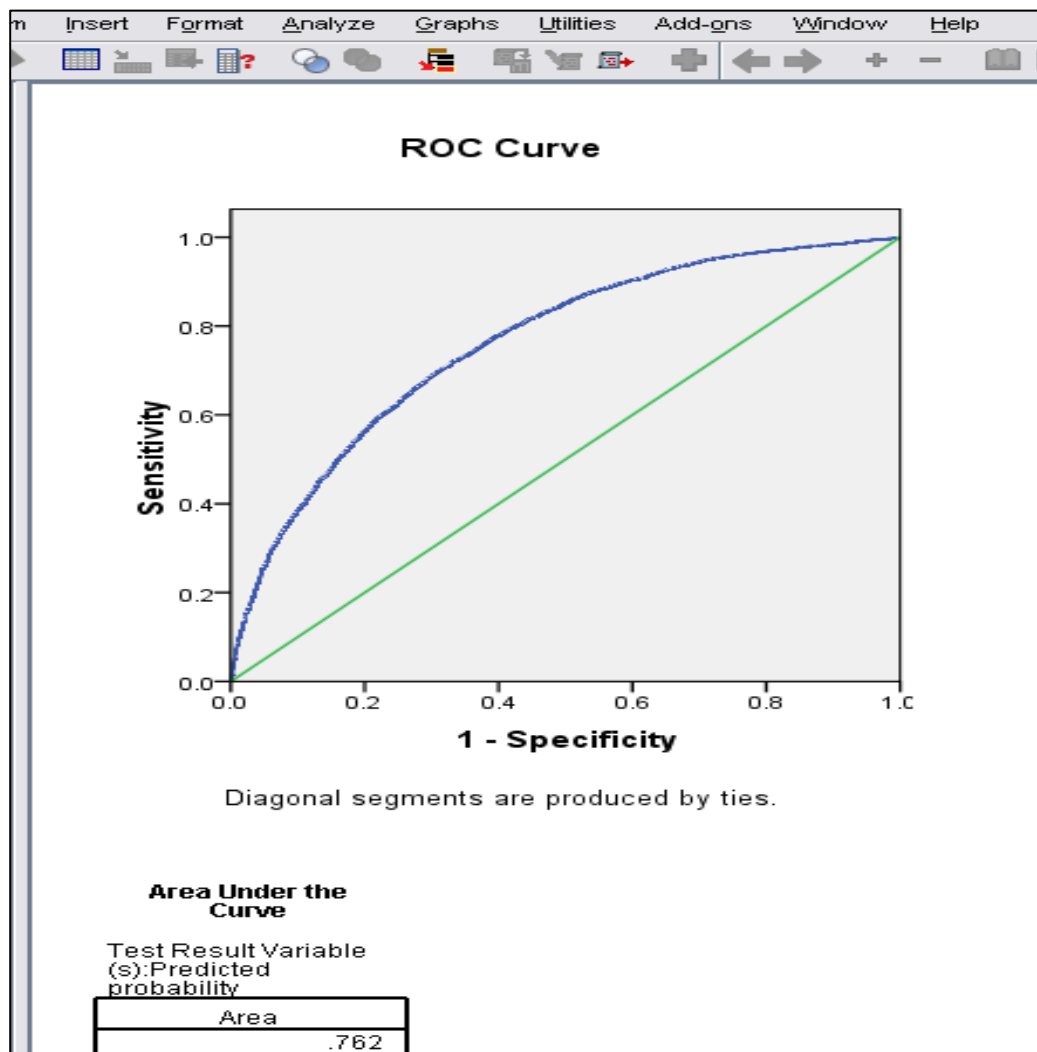


Figura 25. Curva ROC. Fuente: Elaboración Propia.

Análisis curva ROC Caso #6:

Sensibilidad (Sensitivity): Significa la proporción de actuales positivos, o “1” que fueron calificados como tal. En este caso significa la proporción de los clientes que son buena paga y el modelo califico como tal.

Especificidad (Specificity): significa la proporción de los actuales negativos o “0” que fueron calificados como tal. En este caso significa la proporción de los clientes que son mala paga, y el modelo califico como tal.

Estos dos conceptos sirven para estudiar la validez del modelo, entendiendo los dos posibles estados de la variable de salida. Sirven para construir la curva ROC utilizando la relación que existe entre estas dos proporciones de buenos y malos. Si el modelo no es capaz de calificar tanto los buenos como los malos, la relación entre la sensibilidad y la especificidad sería de 1 a 1, como se ve en la línea recta de la curva aleatoria. La relación entre estas proporciones por encima de la curva aleatoria quiere decir que el modelo tiene cierta capacidad explicativa.

Esta curva muestra en la diagonal las probabilidades de un modelo aleatorio (otorgamiento aleatorio de créditos), en la cual suponiendo que se tire una moneda la probabilidad de que caiga cara es de un 50%. Entre más se acerque a 1 el área bajo la curva, menor aleatoriedad representa el modelo. Se alcanza a ver que la curva azul, se aleja de la curva verde (aleatoria), por lo cual se puede decir que el modelo está presentando un desempeño “bueno” y podría ayudar a predecir de manera no-aleatoria si una persona pagará su crédito.

3.4 VALIDACIÓN DEL MODELO LOGIT (CONSTRUIDO CON SPSS)

Dado que se obtuvo la base de datos ideal y una configuración apropiada para construir el modelo y alcanzar los objetivos planteados, se procederá validando el modelo en dos tipos de software diferentes al utilizado inicialmente (SPSS), estos serán R Project y Palisade Decisión tools.

3.4.1 Validación con R PROJECT y RATTLE

R Project es un software estadístico gratuito que permite al usuario descargar los diferentes técnicas estadísticas y gráficas, por ejemplo: Rattle es módulo de R, tiene una interfaz fácil de utilizar, en la cual se ingresa la base de datos, se selecciona la clasificación de los mismos y lo más importante es que este módulo permite correr el modelo con varias técnicas al mismo tiempo.

Distribución de variables en R

Para analizar las distribuciones de las variables antes de correr los modelos que R presenta, se analizaran las variables utilizando una gráfica llamada diagrama de caja (boxplot). Este diagrama muestra la distribución de la variable para todos los casos y la distribución para cada caso en la variable de salida (0 y 1)

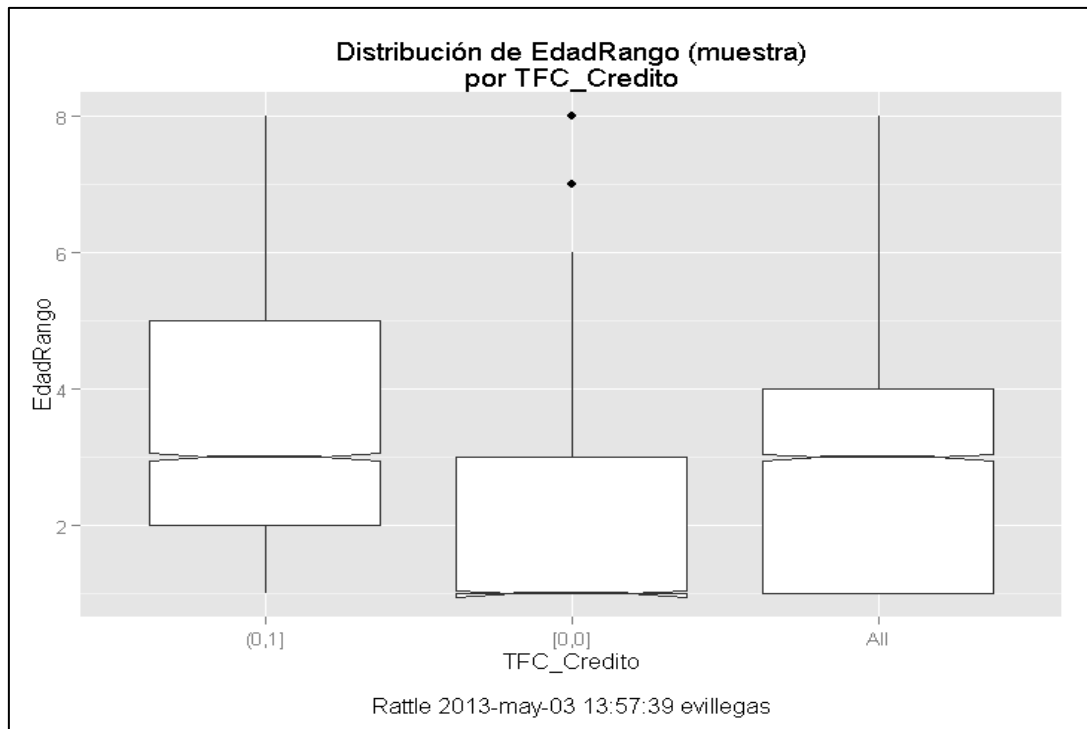


Figura 26. Boxplot Edad. Fuente: Elaboración Propia.

Como se alcanza a ver en el gráfico anterior, la distribución de la variable “EdadRango” muestra una distribución con una media justo debajo del rango número 4 para el boxplot que dice “all”, lo que quiere decir que la mayoría de los casos están entre el rango de 28 a 35 años de edad. También se puede ver como en el caso de los 0 (mala paga) la media está muy por debajo de la media total, lo que quiere decir que a medida que la edad disminuye, la probabilidad de default aumenta. En el caso de los 1, la media está prácticamente al mismo nivel que la media para todos los datos.

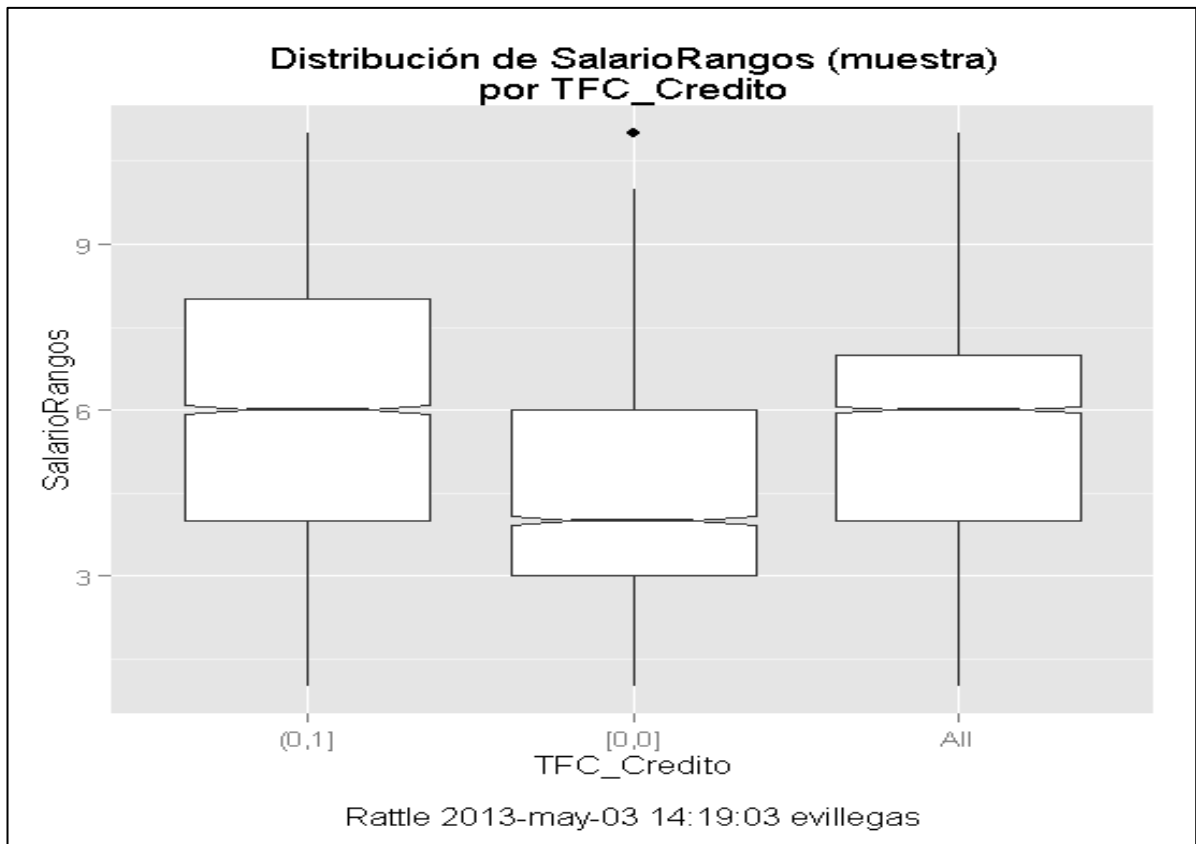


Figura 27. Boxplot Salario. Fuente: Elaboración Propia.

Para este caso se analiza la variable “SalarioRango” la cual presenta una media en el rango 6, lo que implica que la mayoría de los casos tienen un ingreso de alrededor de 800,000\$ pesos. Se puede ver claramente que para los casos de impago (0), la media está muy por debajo ubicándose cerca del rango 4. Esto quiere decir que las personas con un menor salario incrementan su probabilidad de impago.

Ejecución de los modelos utilizando R

Con la base de datos del caso número 6 de SPSS, se corre el modelo con las diferentes técnicas (de minería de datos) que ofrece el módulo de Rattle. Algunos de estos algoritmos no se presentan en el marco de referencia pero se ejecutan dada la facilidad del software R, a continuación se enlistan las técnicas disponibles:

- Árbol de Decisión,
- Ada Boost,

- Bosque Aleatorio
- SVM (Support Vector Machines),
- Regresión LÍneal (Logística)
- Red Neuronal.

Matriz de Error para cada una de la técnicas: Esta indica cómo se comportan los valores numéricos reales vs los predichos, por ejemplo en el árbol de decisión: 576 más 1149 son los datos reales calificados como 0 (“malos”) de los cuales 576 el árbol de decisión califica como 0 y 1149 como 1, de la misma forma 223 más 2506 son los datos reales calificados como 1 (“buenos”) de los cuales 223 calificó como 0 y 2506 como 1.

a) Árbol de Decisión

```
Matriz de error para el modelo Árbol de decisión en Best.csv [validar] (cuentas):

      Predicho
Real   0    1
0    576 1149
1    223 2506
```

b) Ada Boost

```
Matriz de error para el modelo Ada Boost en Best.csv [validar] (cuentas):

      Predicho
Real   0    1
0    749  976
1    382 2347
```

c) Bosque Aleatorio

```
Matriz de error para el modelo Bosque aleatorio en Best.csv [validar] (cuentas):

      Predicho
Real   0    1
0    809  916
1    447 2282
```

d) SVM (Support Vector Machines)

```
Matriz de error para el modelo SVM en Best.csv [validar] (cuentas):

      Predicho
Real   0    1
0    810  915
1    429 2300
```

e) Regresión (Logística)

Matriz de error para el modelo Lineal en Best.csv [validar] (cuentas):

| Real | Predicho | |
|------|----------|------|
| | 0 | 1 |
| 0 | 860 | 865 |
| 1 | 499 | 2230 |

f) Red Neuronal

Matriz de error para el modelo Red neural en Best.csv [validar] (cuentas):

| Real | Predicho | |
|------|----------|------|
| | 0 | 1 |
| 0 | 830 | 895 |
| 1 | 447 | 2282 |

Figura 28. Matrices de Error arrojado por Rattle en R. Fuente: Elaboración Propia.

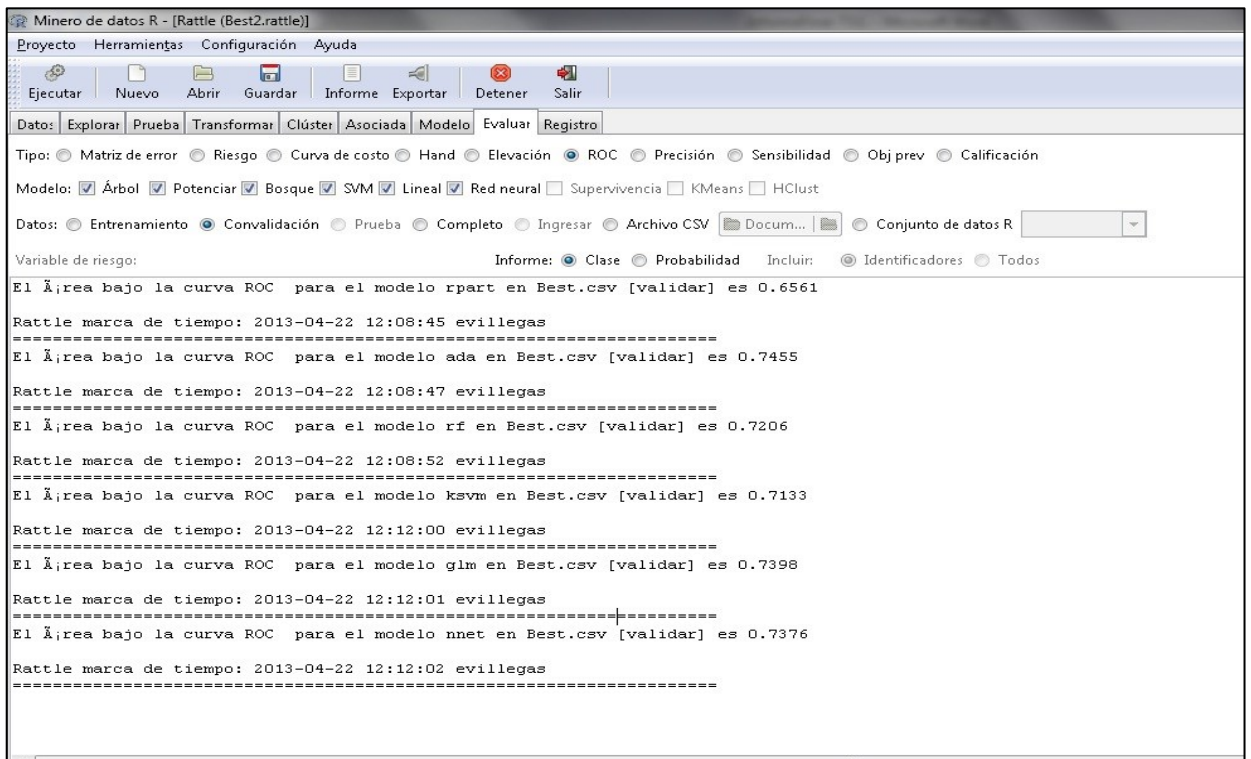


Figura 29. Resultados Curva ROC. Fuente: Elaboración Propia.

A continuación se muestra el resultado de las curvas ROC para cada técnica. Se puede ver como el área bajo la curva del modelo lineal es de 0.7398, lo cual indica que se aleja de la curva aleatoria, siendo esta (0.5). Entre más se acerca a 1 se dice que más capacidad explicativa tiene el modelo. En este caso la técnica más acertada es la ada Boost (Potenciar), y la peor es la del árbol de decisión, con un 0.6561 de área bajo la curva. Se concluye entonces que la técnica lineal es apropiada dado que es la segunda mejor en esta prueba.

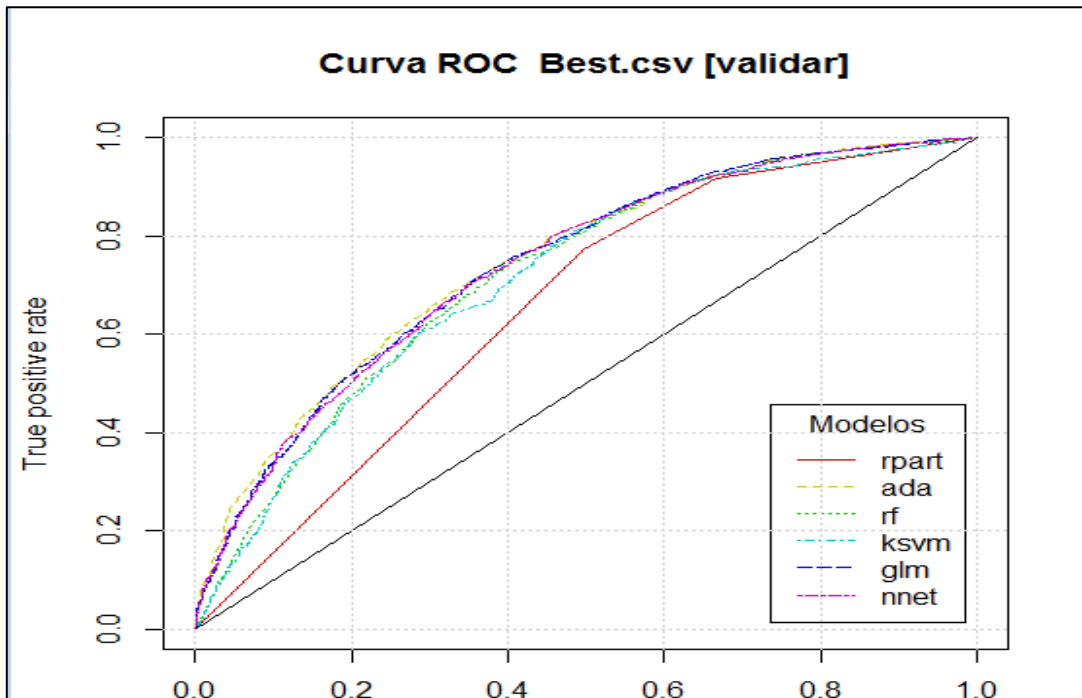


Figura 29. Curva ROC. Fuente: Elaboración Propia.

Dónde:

Rpart = Árbol de decisión

Ada = Boosting

rf= Random Forest (bosque aleatorio: un conjunto de árboles de decisión)

ksvm = Support Vector Machine

glm = Regresión logística

nnet= Red neuronal

Conclusiones R (Rattle): Como se puede observar en las matrices de error, una de las técnicas que mejor califica la base de datos es el modelo Logístico. En la gráfica anterior

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

se grafican todas las curvas ROC para todas las técnicas utilizando los datos de prueba (diferentes a los de entrenamiento), y se evidencia como la curva ROC muestra un buen desempeño del modelo de regresión logística, el cual está graficado en la línea azul (glm).

3.4.2 Validación con PALISADE DECISION TOOLS

Es un software que permite al usuario por medio de 7 módulos diferentes realizar las diferentes técnicas de decisión, en este caso nos centraremos en 2 módulos:

- a) **StatTools:** permite realizar la regresión logística de la misma forma que SPSS, pero arroja un resultado más claro.

| <i>Regression Coefficients</i> | Coefficient | Standard Error | Wald Value | p-Value | Lower Limit | Upper Limit | Exp(Coef) |
|--------------------------------|--------------|----------------|--------------|----------|-------------|-------------|-------------|
| Constant | -0,5512149 | 354688,8336 | -1,55408E-06 | 1,0000 | -695190,665 | 695189,5626 | 0,5762493 |
| Edad Rango | 0,306390127 | 0,013229331 | 23,15991001 | < 0.0001 | 0,280460637 | 0,332319616 | 1,358512195 |
| Cantidad Rango | 0,330527259 | 0,024851036 | 13,30034105 | < 0.0001 | 0,281819227 | 0,37923529 | 1,391701722 |
| Promedio | -4,8315E-06 | 2,53775E-07 | -19,03850606 | < 0.0001 | -5,3289E-06 | -4,3341E-06 | 0,999995169 |
| Salario Rangos | 0,126789714 | 0,008645496 | 14,66540589 | < 0.0001 | 0,109844541 | 0,143734886 | 1,13517828 |
| Sexo = 0 | 0,284260207 | #¡NUM! | | 1,0000 | | | 1,328778644 |
| Sexo = 1 | -0,835475107 | #¡NUM! | | 1,0000 | | | 0,433668394 |
| EstadoCivilCodigo = 0 | 0,373195611 | #¡NUM! | | 1,0000 | | | 1,452368411 |
| EstadoCivilCodigo = 1 | 0,100003649 | #¡NUM! | | 1,0000 | | | 1,10517495 |
| EstadoCivilCodigo = 2 | -0,202841954 | #¡NUM! | | 1,0000 | | | 0,816407261 |
| EstadoCivilCodigo = 3 | -0,532921977 | #¡NUM! | | 1,0000 | | | 0,58688759 |
| EstadoCivilCodigo = 4 | -0,227699387 | #¡NUM! | | 1,0000 | | | 0,796363621 |
| EstadoCivilCodigo = 5 | -0,06095084 | #¡NUM! | | 1,0000 | | | 0,940869492 |
| | 1 | 0 | Percent | | | | |
| <i>Classification Matrix</i> | | | Correct | | | | |
| 1 | 7392 | 1575 | 82,44% | | | | |
| 0 | 2694 | 3186 | 54,18% | | | | |
| | Percent | | | | | | |
| <i>Summary Classification</i> | | | | | | | |
| Correct | 71,25% | | | | | | |
| Base | 60,40% | | | | | | |
| Improvement | 27,40% | | | | | | |

Figura 30. Summary StatTools. Fuente: Elaboración Propia.

Conclusión StatTools: Como se puede observar en la matriz de clasificación, el resultado del modelo aparece relativamente igual a los arrojados por SPSS y por R, lo que indica que el modelo es significativo. En la figura anterior se puede ver la tabla donde se escriben los coeficientes de la ecuación de regresión, y se ve la tabla de calificación con los porcentajes que califico correctamente. 81,44% para los “buenos” y 54,18% para los

“malos”. Esto para un total de 71,25 % de correctos para toda la base de datos (buenos y malos).

- b) **Neural tools:** para este caso, se analizara el modelo alternativo que se trabajó durante el desarrollo del proyecto, este consiste en una red neuronal realizada con otro modulo del software de Palisade llamado Neural tools. Anteriormente se ejecutó una red neuronal con el software de R pero el resultado no fue tan significativo. El software de Palisade permite ver los resultados e interactuar con la base de datos de manera más sencilla. Estos son los resultados obtenidos luego de entrenar la red con la última base de datos configurada para el caso #6.

| Classification Matrix (for training cases) | | | |
|---|------|------|----------|
| | 0 | 1 | Bad (%) |
| 0 | 2648 | 2082 | 44,0169% |
| 1 | 1120 | 6028 | 15,6687% |

| Classification Matrix (for testing cases) | | | |
|--|-----|------|----------|
| | 0 | 1 | Bad (%) |
| 0 | 640 | 510 | 44,3478% |
| 1 | 317 | 1502 | 17,4272% |

Figura 31. Tabla de Clasificación. Fuente: Elaboración Propia.

| <i>Variable Impact Analysis</i> | |
|---------------------------------|----------|
| Promedio | 48,8104% |
| Edad Rango | 14,2552% |
| Cantidad Rango | 12,9347% |
| Sexo | 9,9176% |
| Salario Rangos | 9,7929% |
| EstadoCivilCodigo | 4,2892% |

Figura 32. Impacto de las Variables. Fuente: Elaboración Propia.

Conclusión Neural Tools: en la figura 31 se ve la matriz de clasificación para los casos de prueba (no para los de entrenamiento) que califico un 44,34% mal los ceros, y un 17,42% de los unos. Esto quiere decir que calificó bien el 55,6522% de los ceros y el 82,5728% de los unos. Es un resultado muy similar a los resultados de la técnica

Logística por lo cual se tendrá en cuenta este modelo para un posterior análisis. Neural Tools también arroja una tabla que mide el impacto de las variables en el modelo, como se puede ver en la figura. El promedio es la variable que más impacta el modelo, y el estado civil es la variable menos significativa para predecir si paga bien o mal un crédito.

Ventajas y desventajas de la red neuronal.

La mayor ventaja que tiene la red neuronal, es que al ser un modelo no lineal, tiene la capacidad de identificar errores en los datos atípicos y los datos faltantes. Esta es la mayor ventaja que posee sobre la regresión logística en la cual se descuadra totalmente el “score” si falta un dato o si el dato está desfasado. Una de las desventajas es que la red requiere de mucha capacidad de cómputo para entrenarse y esto crea un limitante en cuanto a flexibilidad para encontrar el mejor modelo. Además solo tiene un solo parámetro con el cual probar, este es el número de capas de la red para el entrenamiento, lo que vuelve muy rígida la técnica para realizar cambios e intentar mejorar el modelo.

3.5 PERFIL DE RIESGO DE LA EMPRESA

Cada compañía de financiamiento tiene su perfil de riesgo de acuerdo a las políticas de cada entidad, y se tienen en cuenta parámetros del tipo de crédito y el tipo de cliente. Existen grandes diferencias entre el riesgo de prestarle dinero a personas naturales y a empresas, como también hay diferencias entre el crédito de consumo o el crédito hipotecario. Es por esto que cada empresa tiene su perfil de riesgo y a continuación se explica cómo obtener el perfil de la empresa Sistecredito SAS en la cual se maximiza la ganancia de la compañía.

Cada institución financiera tiene diferente grado de tolerancia al riesgo, y en esta sección se explicara cual debe ser el nivel de riesgo que la empresa Sistecredito SAS debe aceptar para maximizar sus ganancias.

Con base en el modelo construido, se calculan las probabilidades de default de cada uno de los registros de la sección de la base de datos que no se utilizó para construir el modelo (sección de validación, 30% de la base). El modelo logit (SPSS) agrega otra columna donde se expresa la probabilidad de que el cliente pague el crédito (PRE_1), así:

| Edad Rango | Sexo | Cantidad Rango | Promedio | EstadoCivilCodigo | Salario Rangos | Crédito | PRE_1 |
|------------|------|----------------|-----------|-------------------|----------------|---------|--------|
| 6 | 0 | 4 | 101.738,0 | 1 | 11 | 1 | 98,16% |
| 6 | 0 | 5 | 182.028,0 | 1 | 11 | 1 | 98,01% |
| 5 | 0 | 4 | 62.646,0 | 0 | 9 | 1 | 97,89% |
| 7 | 0 | 4 | 147.952,0 | 5 | 10 | 0 | 97,89% |
| | | | | | | | |

Figura 33. Tabla ejemplo perfil de riesgo. Fuente: Elaboración Propia.

Se ordena esta columna con la probabilidad de mayor a menor, y se parte la base en grupos (deciles) del mismo tamaño (cantidad de registros). Luego se calcula para cada grupo cuantos buenos y cuantos malos hay. Dado que está en orden descendente, en los primeros grupos hay más buenos que malos, y en los últimos más malos que buenos.

| Decil | Buenos | Total | Malos |
|-------|--------|-------|-------|
| 1 | 39 | 44 | 5 |
| 2 | 40 | 44 | 4 |
| 3 | 42 | 44 | 2 |
| 4 | 40 | 44 | 4 |
| 5 | 39 | 44 | 5 |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| ... | ... | ... | ... |
| 92 | 11 | 44 | 33 |
| 93 | 8 | 44 | 36 |
| 94 | 9 | 44 | 35 |
| 95 | 8 | 44 | 36 |
| 96 | 11 | 44 | 33 |

Figura 34. Tabla ejemplo orden Deciles. Fuente: Elaboración Propia.

Para estos grupos se encuentra el acumulado de los buenos y malos hasta el final, y el porcentaje respectivo. Luego se hace un supuesto, que consiste en decir que por cada crédito bien otorgado se ganan 15 pesos, y por cada crédito mal otorgado se pierden 90 pesos. Posteriormente, para obtener la ganancia en cada grupo, se multiplican los buenos por la ganancia (15\$), y se resta el producto de los malos por la pérdida (90\$). Dado que están en orden, la ganancia neta de cada grupo va aumentando cierto punto y luego empieza a disminuir.

| Decil | good | Count of Decile | Bad | Cumulative good | Cumulative bad | Cum good % | cumulative bad % | Cumulative Bad Avoided | Profit | |
|-------|------|-----------------|-----|-----------------|----------------|------------|------------------|------------------------|--------|-------|
| 1 | | 39 | 44 | 5 | 39 | 5 | 1,43% | 0,28% | 99,72% | 135 |
| 2 | | 40 | 44 | 4 | 79 | 9 | 2,90% | 0,51% | 99,49% | 375 |
| 3 | | 42 | 44 | 2 | 121 | 11 | 4,44% | 0,62% | 99,38% | 825 |
| 4 | | 40 | 44 | 4 | 161 | 15 | 5,91% | 0,84% | 99,16% | 1065 |
| 5 | | 39 | 44 | 5 | 200 | 20 | 7,34% | 1,13% | 98,87% | 1200 |
| 6 | | 40 | 44 | 4 | 240 | 24 | 8,81% | 1,35% | 98,65% | 1440 |
| 7 | | 38 | 44 | 6 | 278 | 30 | 10,20% | 1,69% | 98,31% | 1470 |
| 8 | | 42 | 44 | 2 | 320 | 32 | 11,74% | 1,80% | 98,20% | 1920 |
| 9 | | 34 | 44 | 10 | 354 | 42 | 12,99% | 2,36% | 97,64% | 1530 |
| 10 | | 41 | 44 | 3 | 395 | 45 | 14,50% | 2,53% | 97,47% | 1875 |
| 11 | | 34 | 44 | 10 | 429 | 55 | 15,74% | 3,10% | 96,90% | 1485 |
| 12 | | 38 | 44 | 6 | 467 | 61 | 17,14% | 3,43% | 96,57% | 1515 |
| 13 | | 37 | 44 | 7 | 504 | 68 | 18,50% | 3,83% | 96,17% | 1440 |
| 14 | | 37 | 44 | 7 | 541 | 75 | 19,85% | 4,22% | 95,78% | 1365 |
| 15 | | 38 | 44 | 6 | 579 | 81 | 21,25% | 4,56% | 95,44% | 1395 |
| 16 | | 39 | 44 | 5 | 618 | 86 | 22,68% | 4,84% | 95,16% | 1530 |
| 17 | | 37 | 44 | 7 | 655 | 93 | 24,04% | 5,24% | 94,76% | 1455 |
| 18 | | 39 | 44 | 5 | 694 | 98 | 25,47% | 5,52% | 94,48% | 1590 |
| 19 | | 35 | 44 | 9 | 729 | 107 | 26,75% | 6,02% | 93,98% | 1305 |
| 20 | | 39 | 44 | 5 | 768 | 112 | 28,18% | 6,31% | 93,69% | 1440 |
| 21 | | 35 | 44 | 9 | 803 | 121 | 29,47% | 6,81% | 93,19% | 1155 |
| 22 | | 34 | 44 | 10 | 837 | 131 | 30,72% | 7,38% | 92,62% | 765 |
| 23 | | 36 | 44 | 8 | 873 | 139 | 32,04% | 7,83% | 92,17% | 585 |
| 24 | | 34 | 44 | 10 | 907 | 149 | 33,28% | 8,39% | 91,61% | 195 |
| 25 | | 35 | 44 | 9 | 942 | 158 | 34,57% | 8,90% | 91,10% | -90 |
| 26 | | 34 | 44 | 10 | 976 | 168 | 35,82% | 9,46% | 90,54% | -480 |
| 27 | | 32 | 44 | 12 | 1008 | 180 | 36,99% | 10,14% | 89,86% | -1080 |
| 28 | | 39 | 44 | 5 | 1047 | 185 | 38,42% | 10,42% | 89,58% | -945 |

Figura 35. Análisis Perfil de riesgo. Elaboración Propia.

En el grupo donde la ganancia se maximiza, se observa en la gráfica que la mayor ganancia se da en el decil 8, con este punto identificado revisamos en la validación todos los deciles y vemos que el mínimo porcentaje que recibe el decil 8 es 89,75%.

| Edad | EdadRango | Sexo | cantidad | CantidadR | Cantidadprom | Promedio | EstadoCivilCo | salario | SalarioRai | Credito | rand | PRE_1 | Credito | Decile |
|------|-----------|------|----------|-----------|--------------|-----------|---------------|-------------|------------|---------|------|--------|---------|--------|
| 32 | 3 | 0 | 11 | 4 | 2.324.535,00 | 211.321,0 | 1 | 1.800.000,0 | 9 | 1 | 10 | 90,02% | 1 | 8 |
| 48 | 5 | 0 | 1 | 1 | 73.395,00 | 73.395,0 | 0 | 600.000,0 | 4 | 1 | 8 | 90,01% | 1 | 8 |
| 38 | 4 | 0 | 1 | 1 | 81.133,00 | 81.133,0 | 0 | 1.000.000,0 | 7 | 1 | 10 | 89,99% | 1 | 8 |
| 56 | 6 | 0 | 1 | 1 | 140.800,00 | 140.800,0 | 1 | 988.000,0 | 6 | 1 | 10 | 89,98% | 1 | 8 |
| 52 | 6 | 0 | 0 | 1 | 118.000,00 | 118.000,0 | 5 | 980.000,0 | 6 | 1 | 10 | 89,98% | 1 | 8 |
| 45 | 5 | 0 | 0 | 1 | 99.900,00 | 99.900,0 | 1 | 1.095.000,0 | 7 | 1 | 8 | 89,97% | 1 | 8 |
| 50 | 6 | 0 | 0 | 1 | 141.100,00 | 141.100,0 | 1 | 800.000,0 | 6 | 1 | 10 | 89,97% | 1 | 8 |
| 64 | 7 | 1 | 3 | 2 | 242.175,00 | 80.725,0 | 0 | 800.000,0 | 6 | 1 | 9 | 89,97% | 1 | 8 |
| 47 | 5 | 0 | 3 | 2 | 346.216,00 | 115.405,0 | 0 | 550.000,0 | 3 | 1 | 10 | 89,95% | 1 | 8 |
| 41 | 4 | 0 | 3 | 2 | 378.989,00 | 126.329,0 | 5 | 1.700.000,0 | 9 | 1 | 9 | 89,91% | 1 | 8 |
| 50 | 6 | 0 | 0 | 1 | 189.700,00 | 189.700,0 | 0 | 850.000,0 | 6 | 1 | 10 | 89,90% | 1 | 8 |
| 46 | 5 | 0 | 1 | 1 | 29.550,00 | 29.550,0 | 2 | 1.000.000,0 | 7 | 1 | 8 | 89,88% | 1 | 8 |
| 44 | 5 | 0 | 3 | 2 | 425.475,00 | 141.825,0 | 0 | 650.000,0 | 4 | 1 | 10 | 89,85% | 1 | 8 |
| 29 | 3 | 1 | 24 | 5 | 2.177.040,00 | 90.710,0 | 1 | 2.500.000,0 | 11 | 1 | 10 | 89,85% | 1 | 8 |
| 47 | 5 | 0 | 3 | 2 | 282.142,00 | 94.047,0 | 0 | 300.000,0 | 2 | 1 | 9 | 89,82% | 1 | 8 |
| 35 | 4 | 0 | 3 | 2 | 377.625,00 | 125.875,0 | 0 | 950.000,0 | 6 | 1 | 9 | 89,81% | 1 | 8 |
| 51 | 6 | 1 | 7 | 3 | 599.102,00 | 85.586,0 | 1 | 1.200.000,0 | 8 | 1 | 10 | 89,79% | 1 | 8 |
| 43 | 5 | 1 | 20 | 4 | 2.645.991,00 | 132.299,0 | 0 | 1.200.000,0 | 8 | 1 | 10 | 89,79% | 1 | 8 |
| 74 | 8 | 1 | 5 | 2 | 1.235.284,00 | 247.056,0 | 0 | 2.000.000,0 | 10 | 1 | 10 | 89,78% | 1 | 8 |
| 36 | 4 | 0 | 9 | 3 | 866.797,00 | 96.310,0 | 1 | 600.000,0 | 4 | 1 | 9 | 89,77% | 1 | 8 |
| 48 | 5 | 1 | 15 | 4 | 2.356.789,00 | 157.119,0 | 0 | 1.500.000,0 | 9 | 1 | 9 | 89,76% | 1 | 8 |
| 39 | 4 | 0 | 5 | 2 | 1.005.750,00 | 201.150,0 | 1 | 3.000.000,0 | 11 | 1 | 8 | 89,75% | 1 | 8 |
| 59 | 6 | 0 | 0 | 1 | 120.800,00 | 120.800,0 | 0 | 538.000,0 | 3 | 1 | 8 | 89,73% | 1 | 9 |

Figura 36. Validación de Perfil de Riesgo. Elaboración Propia.

La información presentada en este documento es de exclusiva responsabilidad de los autores y no compromete a la EIA.

Es donde la empresa maximiza la ganancia dado que el otorgamiento no es excesivamente cauteloso o conservador, pero tampoco es muy arriesgado. La probabilidad en ese grupo de personas es del 89,75%. Esta es la probabilidad mínima que debe arrojar el modelo para otorgar el crédito y maximizar las ganancias con respecto al riesgo.

Si se realiza este mismo ejercicio con la base de datos de cada empresa, y se tiene un supuesto de pérdida y ganancia para cada crédito, el perfil de riesgo de cada compañía arroja un valor distinto. Cada entidad puede maximizar sus ganancias teniendo en cuenta su perfil de riesgos. Pero no todas las empresas utilizan el valor de la información que poseen y simplemente fijan su tolerancia al riesgo teniendo en cuenta el perfil de los accionistas y las políticas de la empresa. Este se convierte en un buen ejemplo de la forma para explotar los datos y la información disponible para lograr una mejor gestión.

4 DISCUSION DE RESULTADOS

Tras el análisis de la información disponible, y una reflexión sobre el proceso de estudio de crédito de la empresa, surgen varias observaciones. El modelo estadístico obtenido en este proyecto se puede decir que alcanza a modelar el comportamiento del cliente, pero no parece ser tan preciso como se esperaba. Se detecta que el estudio de crédito de la empresa se enfoca principalmente en las referencias suministradas por el cliente, y no se alcanzan a validar la mayoría de las variables que se dan al momento de solicitar el crédito. Por ejemplo la edad, si la persona comete un error al llenar la solicitud, existe una gran probabilidad que el analista de crédito no detecte este error lo cual afecta la información y los datos para el análisis estadístico. Esto puede haber sucedido con muchos clientes, lo cual genera un sesgo en la información disponible y se afecta la validez del modelo. La gran mayoría de registros tienen cada variable correcta, pero se alcanza a ver que existen varios errores lo cual afectan la precisión del modelo.

En primera instancia, surgen dudas acerca de la calidad de la información de la base de datos. Las primeras corridas e intentos muestran unos resultados poco significativos por lo cual surge el interrogante si se optó por una técnica que no era la apropiada, o la base de datos tiene serios problemas en su estructura. A medida que se van descubriendo y corrigiendo los problemas que vienen afectando el modelo, los resultados van mejorando.

Después del tratamiento empleado a la base de datos y a la configuración del modelo, se llega a un resultado significativo el cual sirve para mitigar el riesgo. Los modelos no presentan la precisión esperada pero sirve para estudiar la solicitud con mayor cautela para cada cliente. Como se vio anteriormente, el modelo encontrado en última instancia tiene capacidad explicativa sobre la variable de salida. Aunque se mantiene la imprecisión de los valores en las variables, y no se tiene clara la forma como se debe juzgar el cliente para generar la base de datos inicialmente, pero se puede decir que se cumplió con los objetivos dentro del alcance del proyecto.

Incluso con la ambigüedad de la base de datos disponible, se logra llegar a un modelo mejor que el "aleatorio", por lo cual se puede decir que se cumplió con la meta propuesta inicialmente. Las técnicas escogidas fueron adecuadas al mostrar unos resultados relativamente parecidos por lo cual se concluye que fue una buena decisión optar por las técnicas Logit y la red neuronal.

Una de las desventajas de la regresión logística, es el hecho de que al ser un modelo lineal hace que se pueda afectar el "score" fácilmente al cometer un error en la solicitud de crédito. Si una persona se equivoca en el género (sexo) o en el estado civil, se puede afectar el score de manera que puede perjudicar el estudio de crédito. A diferencia de la red neuronal, la cual es capaz de interpretar este tipo de inconsistencias y no es tan sensible ante un error o un dato faltante. En este orden de ideas, se podría decir que la red neuronal es más apropiada para este caso que el modelo lineal a la hora de estudiar un crédito. Pero la red también tiene la desventaja de tener la necesidad de

entrenamiento y, en el software Neural Tools de la empresa Palisade, solo un parámetro principal con el cual jugar (el número de capas), por lo cual se concluye que en primera instancia, para la empresa es más conveniente trabajar con la regresión logística para contar con mayor flexibilidad y realizar ajustes fácilmente en el modelo o en los modelos para cada ciudad.

Una reflexión muy importante que se hace al analizar las dificultades que tuvo este proyecto es el hecho de que las mismas técnicas y el mismo tratamiento se pueden hacer para cualquier compañía de financiamiento. Por lo general las empresas del sector financiero se desgastan mucho más validando los datos del sujeto a crédito, por lo cual se da por hecho la veracidad de los datos. Si se realiza el mismo procedimiento utilizando la base de datos de cualquier empresa de financiamiento, los resultados deben ser mucho mejores y se puede realizar el mismo análisis en cuanto al perfil de riesgo. La base de datos y la historia de los clientes pueden explicar cuál es el perfil de riesgo de cada empresa y donde se maximizan sus utilidades. Este análisis se puede hacer para todo tipo de crédito, no solo para el crédito de consumo como se vio en este caso, por lo cual el riesgo es diferente para cada empresa y para cada tipo de cartera. Una compañía que coloque su cartera en crédito hipotecario se expone mucho menos dado que puede recuperar el inmueble en caso de impago, por lo cual el perfil de riesgo es muy distinto para este caso. Lo mismo sucede con los otros tipos de crédito como el crédito estudiantil, el crédito comercial o el microcrédito, teniendo cada uno su tratamiento y su análisis. En cuanto al alcance de este proyecto, se ve cual es el perfil de riesgo de la empresa Sistecredito SAS la cual se enfoca en el crédito de consumo.

Explícitamente, los resultados del modelo Logit son los siguientes, expresados en la ecuación para hallar la probabilidad.

La Ecuación en StatTools:

$$Z = -0,991069303512467 + 0,31156641475393 * x_1 + \mathbf{0,413627059181838} * x_2 + 0,0808899288068568 * x_3 + 0,0396757778452553 * x_4 + (-\mathbf{1,03074508125819}) * x_5 + 0,326739623428895 * x_6 + 0,02707267003054 * x_7 + (-0,309549741758799) * x_8 + (-\mathbf{0,623015251329825}) * x_9 + (-0,318631520913671) * x_{10} + (-0,0936850824718145) * x_{11} \quad (20)$$

Dónde:

- x_1 = Edad Rango
- x_2 = Cantidad Rango
- x_3 = Salario Rango
- x_4 = Sexo (masculino 1)
- x_5 = Sexo (femenino 0)
- x_6 = Estado Civil (Casado 0)
- x_7 = Estado Civil (Soltero 1)
- x_8 = Estado Civil (Separado 2)
- x_9 = Estado Civil (Viudo3)
- x_{10} = Estado Civil (Unión Libre 4)
- x_{11} = Estado Civil (Otro 5)

- En la ecuación se observa que las variables afectan de manera positiva o negativa el score, y cada coeficiente expresa el peso de la variable sobre la calificación, se observa resaltado las variables de mayor influencia.

La Ecuación en SPSS:

$$Z = -2,116 + 0,333 * x_1 + \mathbf{0,413} * x_2 + 0,074 * x_3 + \mathbf{1,138} * x_4 + 0,334 * x_5 + 0,071 * x_6 + (-0,319) * x_7 + (-\mathbf{0,591}) * x_8 + (-0,272) * x_9 \quad (21)$$

Dónde:

- x_1 = Edad Rango
- x_2 = Cantidad Rango
- x_3 = Salario Rango
- x_4 = Sexo
- x_5 = Estado Civil (Casado 0)
- x_6 = Estado Civil (Soltero 1)
- x_7 = Estado Civil (Separado 2)
- x_8 = Estado Civil (Viudo 3)
- x_9 = Estado Civil (Unión Libre 4)

Las diferencias que existen entre los modelos de regresión logística, realizados con la misma base de datos y la misma configuración, pero con los diferentes programas, no son tan significativas. Se puede concluir que cada software utiliza su propia implementación del algoritmo pero al tratarse de la misma técnica, las diferencias son poco relevantes. El programa de SPSS de IBM es probablemente el más conocido, pero no quiere decir que sea el mejor. Las diferencias entre los resultados son tan pequeñas que se puede escoger cualquiera para realizar la ecuación de regresión logística. El software de Palisade (StatTools) es el que arroja los resultados más claros (en términos de la presentación de los resultados) y tiene la ventaja de trabajar sobre la misma hoja de Excel. El SPSS y el R se prestan para realizar cambios de manera más fácil en las variables para “jugar” con el modelo, pero el resultado de salida (output) parece un poco más difícil de visualizar e interpretar. Lo mismo sucede entre la red neuronal ejecutada con el software de R, y la red neuronal construida con el software de Palisade Neuraltools.

Los resultados de Neuraltools se ven claramente y la versión industrial permite ejecutar la red en tiempo real sobre la misma hoja de Excel, por lo cual le sirve a la empresa al momento de realizar un estudio de crédito. Una vez se obtiene una red entrenada, el tiempo para estudiar un crédito es dependiendo de la capacidad de cómputo en promedio 3 segundos o menos (tiempo real), esto dado el caso que la compañía desee utilizar red neuronal. En caso de utilizar el modelo logit, simplemente se llenan las variables en la ecuación dada y esta arroja el score instantáneamente.

Sin importar el modelo que escojan utilizar, el score arroja un resultado por encima del perfil de riesgo de la empresa (89,75%), el analista de crédito no será tan riguroso con su

estudio lo cual agilizará sustancialmente cada solicitud. Si el modelo arroja resultados inferiores al perfil de riesgo, el analista de crédito ejecutará los pasos que se vienen realizando en la compañía de manera sistemática pero se tendrá un poco más de cautela para finalmente tomar la decisión de otorgar o negar el crédito. Actualmente la compañía estudia un promedio de 800 solicitudes de crédito diarias, con un promedio de 12 minutos por solicitud. Desde un punto de vista optimista, con el modelo la compañía logrará disminuir el tiempo a 10 minutos, lo cual representa un gran ahorro en recursos y esfuerzos.

Dado que StatTools arroja los resultados más claros, se opta por escoger este programa para concluir con la técnica Logit, al obtener la Z de la ecuación, por medio de la siguiente ecuación se obtiene la probabilidad,

$$P = e^z / (1 + e^z)$$

Dónde:

P= Probabilidad

e= Número de Euler

Z= Valor numérico que arroja la ecuación del modelo

Por último como se enunció anteriormente, la probabilidad mínima que debe arrojar el modelo es 89,75%, para otorgar el crédito y maximizar las ganancias con respecto al riesgo.

5 CONCLUSIONES Y CONSIDERACIONES FINALES

Los frutos de este proyecto no solo yacen en el modelo estadístico elaborado para la empresa Sistecredito SAS, sino en los conocimientos adquiridos, útiles para modelar cualquier fenómeno u ocurrencia. El caso de la empresa Sistecredito SAS se convierte en un buen ejemplo de una empresa mediana que tiene la posibilidad de explotar la data disponible para mejorar su gestión. Analizando el contexto de las empresas en la coyuntura actual, con la realización de este proyecto se puede ver claramente como la información se convierte en un activo fundamental para las organizaciones.

El estudio de las técnicas empleadas muestra cómo se puede plantear y estudiar cualquier problema de la misma naturaleza. Esto puede ser una sugerencia para proyectos posteriores a este, en donde se pueda modelar cualquier problema utilizando las técnicas empleadas en este, como por ejemplo, el cálculo de la probabilidad de recompra de un cliente. También quedan varias inquietudes sobre los otros modelos estadísticos disponibles que no se alcanzan a contemplar en detalle, y dan luz para un posterior análisis.

Luego de un análisis detallado del proceso de estudio de crédito, se encuentra que dentro del modelo de negocio de la empresa está la prioridad en la agilidad del estudio de crédito, por lo cual se puede decir que el modelo se implementará como una estrategia para la actividad “core” de la empresa.

Al construir el modelo y validarlo con la base de datos no utilizada, surgen varias estrategias que los autores proponen, en este caso para la empresa Sistecredito SAS. Dada la naturaleza del cliente y los perfiles que maneja la empresa, se concluye que todos los clientes merecen un estudio sistemático como se viene trabajando actualmente y se deben implementar estrategias que incorporen al modelo obtenido. La estrategia principal que se propone consiste en canalizar a los clientes con un “score” negativo a los agentes de crédito con mayor experiencia. La compañía posee un grupo de 7 personas especialistas en el análisis de crédito, y de otras 30 personas que gestionan la cobranza pero apoyan el área de créditos en los picos de solicitudes entrantes. Si se logra filtrar los clientes con alta probabilidad de default, solo a los agentes con mayor experiencia en el análisis de crédito, se puede mitigar de manera considerable la exposición al riesgo en este tipo de cartera.

La empresa seguirá creciendo y robusteciendo su base de datos, por lo cual se convierte una estrategia clave encontrar la manera de aprovechar la historia de la compañía para mejorar la gestión a futuro. Se concluye así que se lograron alcanzar los objetivos del trabajo, encontrando un modelo que consigue mitigar el riesgo de impago y agiliza el estudio de crédito. Es preciso decir, que se encontró la forma de utilizar la información existente en la base de datos, lo cual repercutirá en la gestión de la empresa Sistecredito.

Además, se concluye que la metodología que se aplicó en este trabajo se podría replicar en otras empresas, utilizando otras bases de datos.

BIBLIOGRAFÍA

- Abdou, H. P. (s.f.). Neural nets vs Conventional techniques in credit scoring.
- Abuín, J. M. (11 de 2007). *Regresión con Variable Dependiente Cualitativa*. Obtenido de Humanidades:
http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_variable_dependiente_dicotomica_3.pdf
- Bonilla, M., Olmeda, I., & Puertas, R. (2003). *Modelos paramétricos y no paramétricos*. Obtenido de Aeca: <http://aeca.es/pub/refc/articulos.php?id=0078>
- Briceño, P. L. (16 de 02 de 2010). *cómo las instituciones financieras evalúan otorgar préstamos*. Obtenido de Gestión:
<http://blogs.gestion.pe/deregresoalobasico/2010/02/la-caja-negra-o-como-las-insti-2.html>
- Castaño, J., Estrada, D., & Patiño, M. A. (2012). *Reporte de la situación del crédito en Colombia*. Obtenido de Banco de la República de Colombia:
http://www.banrep.gov.co/documentos/informes-economicos/encuestas/SCC/2012/escc_dic_2012.pdf
- Dpto. de Matemática Aplicada, UCM. (SF). *Introducción al SPSS*. Obtenido de e-estadística: http://e-stadistica.bio.ucm.es/web_spss/analisis_descriptivo.html
- Elizondo, A. (2004). *Medición Integral del Riesgo Crédito*.
- Franco, S. I., Lochmüller, C., & Betancur, A. O. (Diciembre de 2011). *La medición del riesgo crédito en Colombia y el Tratado de Basilea 3*. Obtenido de Postgrado EIA:
<http://revistapostgrado.eia.edu.co/Revista%20Edici%C3%B3n%20N%C2%BA.7/Soluciones%20N7%20art%203.pdf>
- GARDNER, M. J.–M. (1989). Evaluating the Likelihood of Default on Delinquency.
- González Arbelaez, Á. (2012). *Reporte de Estabilidad Financiera*. Obtenido de Banco de la República de Colombia:
http://www.banrep.org/documentos/publicaciones/report_estab_finan/2010/Determin_riesgo_credito_comercial_Colombia.pdf
- González, J. J. (2001). *Programación Matemática*. Tenerife; España.
- Halweb. (9 de Enero de 2003). *Modelo lineal de probabilidad*. Obtenido de Halweb:
http://halweb.uc3m.es/esp/Personal/personas/mcasas/esp/econometria/exercises/ejemplotema4_5_st.pdf

- Horra, J. d. (SF). *Estadística Descriptiva*. Obtenido de U.A.M: http://www.uam.es/personal_pdi/ciencias/horra/Estadistica-Apuntes/Descriptiva-Una-Variable.pdf
- IBM. (2011). *System User Guide*. Obtenido de IBM: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf
- Informatica Integral. (2004). *Sistemas Expertos*. Obtenido de Informatica Integral: <http://www.informaticaintegral.net/sisexp.html>
- Marasca, R., Figueroa, M., Stefanelli, D., & Indri, A. (Diciembre de 2003). *Basilea III*. Obtenido de Felaban: http://www.felaban.com/boletin_clain/basileall.pdf
- Martinez, A. (15 de 04 de 2011). *Variables Dummy*. Obtenido de Ciencias Empresariales: <http://cienciasempresariales.info/variables-dummy-en-modelos-de-regresion/>
- Mejía, A. T. (s.f.). Introducción ala historia económica en Colombia.
- Moliner, L. M. (Octubre de 2003). *Máxima Verosimilitud*. Obtenido de SEH: <http://www.seh-lelha.org/maxverosim.htm>
- Montoya, J. (SF). *Riesgo Crédito*. Obtenido de SAS: <http://www.sas.com/offices/europe/spain/prodsol/spotlights/riesgosas09.html>
- Noesis. (2005). *Teoría Var*. Obtenido de Noesis: <http://www.noesis.es/var/teoria.htm>
- Osorio, A., & Álvarez, S. I. (Mayo de 2011). Medición del Riesgo Crédito en Colombia - Hacia Basilea III. Medellín, Colombia.
- project, R. (2012). *Statistical Computing*. Recuperado el Abril de 2013, de R Project : <http://www.r-project.org/>
- Ramón, G. (SF). *Correlación de Variables* . Obtenido de Universidad de Antioquia: http://viref.udea.edu.co/contenido/menu_alterno/apuntes/ac36-correlacion-variables.pdf
- S.A.S, S. (s.f.). Compañía Sistecrédito S.A.S. 2013. Medellín, Colombia.
- Saavedra García, M. L., & Saavedra García, M. J. (21 de 05 de 2010). *Cuadernos Administración*. Obtenido de Javeriana: http://cuadernosadministracion.javeriana.edu.co/pdfs/Cnos_Admon_23-40_12_MSaavedra.pdf
- Schreiner, M. (11 de 09 de 2002). *Scoring*. Obtenido de Microfinance: http://www.microfinance.com/Castellano/Documentos/Scoring_Ventajas_Desventajas.pdf

Univeridad Tecnológica de la mixteca. (Septiembre de 2009). *Limitaciones del modelo lineal de probabilidad y alternativas de modelación microeconométrica*. Obtenido de UTM: http://www.utm.mx/edi_anteriores/Temas39/1ENSAYO%2039-1.pdf

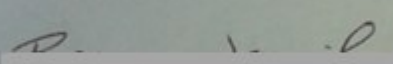
Wikipedia. (23 de 09 de 2009). *Curva ROC*. Obtenido de Wikipedia: http://es.wikipedia.org/wiki/Curva_ROC

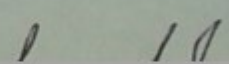
ESCUELA DE INGENIERÍA DE ANTIOQUIA

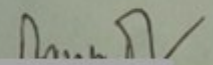
ACTA DE EVALUACIÓN FINAL DE TRABAJO DE GRADO

EIA
Escuela de Ingeniería de Antioquia
Ser, Saber y Servir

| | | |
|--|--|----------------------------------|
| Fecha: (dd/mm/aa) | 4 /7/2013 | |
| Nombre del proyecto: | <i>Modelos de calificación cuantitativa (scoring) para agilizar créditos y mitigar el riesgo</i> | |
| Director del proyecto: | <i>Christian Lochmüller</i> | |
| | Nombre del estudiante | Programa académico |
| | <i>John Alfredo Restrepo Llanos</i> | <i>Ingeniería Administrativa</i> |
| | <i>Emilio Villegas Molina</i> | <i>Ingeniería Administrativa</i> |
| Nombre del Jurado: | Javier Jaramillo Betancur | |
| Evaluación del proyecto: | | |
| <input type="checkbox"/> No aprobado <input checked="" type="checkbox"/> Aprobado | | |
| Espacio exclusivo para jurado | | |
| <input type="checkbox"/> Mención Pública <input type="checkbox"/> Mención honorífica <input type="checkbox"/> Trabajo laureado | | |
| <p>Justificación del reconocimiento: (Artículo 28 del Acuerdo 11: "El director del Programa presentará el acta final de evaluación al Consejo Académico, donde consta la solicitud de mención especial debidamente justificada y el Consejo determinará si se otorga o no")</p> | | |


 Camilo Sylva Sánchez
 Director del Programa


 Christian Lochmüller
 Director del Trabajo de Grado


 Javier Jaramillo Betancur
 Jurado