UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Proyecto de Grado
Enero 2020

Comparación de dos modelos para predecir la probabilidad deincumplimiento (Modelo Logit-Modelo GAM)

María Alejandra Gualteros Barbosa [a]
maria.gualteros@usantotomas.edu.co

Gil Robert Romero[b]
gilromero@usantotomas.edu.co

Resumen

Una de las necesidades más importantes de las instituciones crediticias es tener criterios confiables para determinar a quienes deben otorgar el crédito. En este sentido, se debe implementar metodologías como lo son los modelos de otorgamiento de créditos como una solución a estas necesidades. En este proyecto se compara dos metodologías que pueden utilizarse en el Riesgo de Crédito, los datos proporcionados a este proyecto son de una entidad financiera Colombiana, como primera parte se describe los estudios anteriores que tuvieron respuesta al problema, las metodologías utilizadas para desarrollar un modelo de originacíon, como primer paso para esta construcción es la agrupación de caracteristicas en categorías y se calcula su capacidad predictiva utilizando medidas de predicción, como el peso de la evidencia y el valor de la información, las variables que presentan poca capacidad de predicción son eliminadas del modelo. Una vez obtenidos los modelos cuyas variables presentan un buen nivel predictivo se sigue con la etapa de selcción de modelo utilizando técnicas como el criterio AUC y la matriz de confusión y finalmente se hacen algunas las recomendaciones con el fin de implemntar el mejor modelo en esta entidad.

Palabras clave: Modelo Loístico, Modelos Aditivos Generalizados, Probabilidad de Incumplimiento, Riesgo de Crédito.

Abstract

One of the most important needs of lending institutions is to have reliable criteria to determine who should grant the credit. In this sense, methodologies such as credit granting models should be implemented as a solution to these needs. In this project, two methodologies that can be used in Credit Risk are purchased, the data provided to this project are from a Colombian financial institution, as part one describes the previous studies that responded to the problem, the methodologies used to develop a model of origin, as a first step for this construction is the grouping of characteristics into categories and its predictive capacity is calculated using prediction measures, such as the weight of the evidence and the information value, the variables that have little predictive capacity are eliminated of the model. Once the models whose variables have a good predictive level are obtained, the model selection stage is followed using techniques such as the AUC criterion and the confusion matrix and finally some recommendations are made in order to implement the best model in this entity.

Keywords: Logistic Model, Generalized Additive Models, Probability of Default, Credit risk.

[a]Estudiante pregrado Estadística, U. Santo Tomás, sede Bogotá
[b]Docente Facultad de Estadística, U. Santo Tomás, sede Bogotá

1. Introducción

El Riesgo de Crédito (RC) puede definirse como la posibilidad de que una entidad vea afectado su patrimonio y utilidades debido al incumplimiento de las obligaciones por parte de los clientes, por ello es importante para las entidades financieras analizar el RC ya que ayuda a mitigar las pérdidas. Igualmente la Superintendencia Financiera de Colombia (SFC) entidad gubernamental encargada de regular y supervisar los sistemas financieros y bursátil, mediante la circular 31 del 2002 define los lineamientos generales para la administración del riesgo de crédito de las entidades vigiladas, donde estipula que para la correcta administración del portafolio crediticio compuesto por tres etapas (Otorgamiento, Seguimiento, y Control) cada entidad financiera debe diseñar metodologías que le permita evaluar los respectivos riesgos y ayudar a la toma de decisiones (Colombia, 2002).

La finalidad de las instituciones financieras es la captación de dinero para después prestarlo por medio de diferentes productos como créditos rotativos o tarjetas de crédito y ganar por medio de la diferencia de las tasas de interés, pero uno de los puntos importantes en los que las instituciones deben centrarse para llegar a dicho objetivo es calcular las provisiones de capital, las cuales ayudarán a mitigar las pérdidas provocadas por el impago de sus productos financieros, y el desarrollo de metodologías para la teoría de riesgos ha permitido establecer las pérdidas a las que las entidades están expuestas. (Fernandez Castaño Perez Ramírez , 2005).

En la etapa de otorgamiento de créditos, las entidades deben optimizar la toma de decisiones y en su gran mayoría escogen por automatizar dichas decisiones por medio de metodologías como Credit Scoring (Ladino, 2014), estos métodos se especializan en pronosticar e identificar la posibilidad de que un cliente incumpla con sus obligaciones crediticias, dichas metodologías consisten en clasificar los clientes en dos categorías ("Buenos" y "Malos") a partir de su comportamiento histórico con las obligaciones financieras, contemplando sus variables sociodemográficas y financieras. Dentro de los modelos más usados en las entidades bancarias se encuentran análisis discriminantes, regresiones lineales y logísticas, métodos no paramétricos de suavizamiento, cadenas de Markov, entre otros.

Es por esto que el objetivo de este documento es comparar dos metodologías (Modelo Logístico y Modelo Aditivo Generalizado) usadas en el RC, utilizando métodos de discriminación, las cuales serán definidas más adelante y así escoger el modelo que mejor predicción tenga para la evaluación de probabilidad de incumplimiento. La base utilizada para dicho análisis fue proporcionada por la gerencia de Riesgo de Crédito de una entidad financiera de Colombia y contiene la información de un conjunto de clientes cuyos créditos pertenecen al segmento de microfinanzas.

Este documento consta de 7 secciones adicionales a esta introducción, la primera parte contiene los antecedentes y estudios realizados por otros autores que abarcan el tema de probabilidad de incumplimiento. En la segunda parte se contextualiza sobre el RC y las metodologías utilizadas en este tema para proceder a describir la base de datos con la que se realiza el estudio y luego mostrar los resultados y discutirlos en la última sección las conclusiones.

2. Antecedentes

El propósito del sistema de medición de riesgo de crédito es determinar los factores que hacen que la cartera vencida de un banco crezca, y con el objetivo de prevenir pérdidas mayores se crean métodos que ayudan a evaluar el RC, dentro de los elementos que componen el riesgo crediticio la probabilidad de incumplimiento es la más importante y en la que la mayoría de estudios se centran, las metodologías para estimar esta probabilidad surgieron desde los años treinta, ya que se iniciaron estudios dedicados a las razones financieras, pero fue hasta 1968 en el que Altman estudió los determinantes de la probabilidad de quiebra de las empresas, dónde demostraba que las variables que explicaban las quiebras eran las variables financieras como el apalancamiento , el flujo de efectivo y la rentabilidad. En este estudio se inicia el análisis discriminante como metodología para el análisis de riesgo de crédito. (Altman , 1968)

Un estudio en 1999 utilizó un análisis discriminante y los modelos logit y probit para determinar las causas de bancarrota en las empresas, tomando una muestra de 949 empresas y concluyó que estos modelos son más precisos para el análisis de interés y que las variables más importantes que determinan la bancarrota son la solvencia, la rentabilidad, el flujo de caja y el ciclo económico (Lennox, 1999). En Colombia existen diversos estudios que utilizan modelos de variables discretas para estimar la probabilidad de quiebra o de incumplimiento en las empresas, un ejemplo de esto es un estudio que se realizó en 2003 donde se estiman los determinantes de insolvencia para las empresas colombianas basándose en la información de los estados financieros de 9000 empresas y utilizando técnicas de regresión probit el cual identifica al 82 % de las empresas propensas a la quiebra (Martinez A, 2003).

Otro estudio realiza ejercicios de pruebas de estrés para estimar el comportamiento del sistema financiero ante situaciones externas, buscando determinar el incremento de la cartera vencida ante situaciones económicas impredecibles y los resultados muestran la importancia de abarcar la regulación del RC (Amaya G, 2005). En 2007 utilizaron matrices de transición para estimar la calidad de créditos, calculan la probabilidad de que un crédito pase de una calificación a otra, y encontraron que los factores que determinaban si un crédito migra de una calificación a otra son la composición de la deuda, la liquidez y el tamaño de la empresa, también se tienen en cuenta variables macroeconómicas como el PIB y el promedio trimestral de la tasa de cambio, las cuales dan significativas y la conclución es que el ciclo económico es determínante del cambio de calificación de cartera (Gómez Gonzáles, Morales Acevedo, Pineda Garcia, Zamudio Gómez, 2007).

En 2007 una investigación desarrolló un modelo logit multinomial ordenado para calcular la probabilidad de incumplimiento, la variable dependiente es el vencimiento de los créditos y el resultado demuestra que la liquidez, el plazo del crédito y el tipo de garantía son determinantes para la probabilidad de incumplimiento (Zamudio Gómez, 2007).

Un estudio en el 2009 con metodologías como redes neuronales, análisis discriminante y regresión logística para evaluar el riesgo crediticio en tarjetas de crédito arrojó que la red neuronal es la que mejor clasifica comparado con las otras dos metodologías, ya que obtiene una mejor capacidad de predicción, también se sugiere como futuro alcance de investigación que una actualización en la base de datos se puede utilizar para mejorar el rendimiento de la red neuronal y los resultados se pueden comparar con los resultados anteriores (Islam, Zhou, Li, 2009).

En el 2018 en el Master en Técnicas Estadísticas de la universidad de Coruña en España se realizó una revisión metodológica de las técnicas que se utilizan generalmente para construir un modelo para RC, los resultados evidencian que las variables que ayudan a medir el riesgo de crédito son el comportamiento de pago y los movimientos en cuenta, también se determina que es mejor construir dos diferentes modelos en función de que el cliente tenga o no algún producto de riesgo contable, y el modelo que se expone en el trabajo permite discriminar a los clientes morosos de los no morosos descargando la información de su banca electrónica y utilizando la información de sus movimientos bancarios con su previa autorización, lo cual es una gran ventaja para otorgar financiación a nuevos clientes (Cañete, 2018).

Los modelos de medición de RC buscan cuantificar el riesgo ante algún incumplimiento por parte de los clientes, las instituciones financieras deben establecer ideas para la administración y el control del riesgo de crédito al que se expone la entidad, las metodologías estadísticas son una gran herramienta para cumplir dicho objetivo, por lo tanto, es importante que cada entidad desarrolle sus métodos que ayuden a mitigar y controlar el riesgo y cumplan con los márgenes reglamentarios emitidos por los entes de regulación.

3. Objetivos

3.1. Objetivo General

Comparar la eficiencia de dos metodologías en la predicción del Riesgo de Crédito de una entidad financiera en el producto de Microfinanzas con el fin de seleccionar el de mejor predicción y así implementarlo en la entidad en el año 2020.

3.2. Objetivos Específicos

- Identificar variables cuantitativas y cualitativas con un alto nivel discriminativo que ayuden a predecir la probabilidad de incumplimiento.
- Analizar los resultados de validación para cada uno de los modelos implementados en este proyecto.
- Seleccionar el mejor modelo que ayude a predecir la probabilidad de incumplimiento crediticio.

4. Marco Teórico

En los años setenta se empezó a ver cambios en las instituciones financieras por la globalización, innovación en productos financieros y la volatilidad de variables exógenas y endógenas, que obligaron al mercado financiero a optar por cambios. En 1974 se crea el Comité de Supervisión Bancaria de Basilea con los presidentes de los bancos centrales del G10, (grupo de países que accedieron a participar en el

Acuerdo General de Préstamos en 1962) por causa de la bancarrota de dos importantes bancos, Bankhaus Herstatt en Alemania y Franklin en los EE.UU. El principal objetivo era realizar sugerencias para la regulación de entidades financieras a nivel mundial, la recomendación más importante es la de limitar el apalancamiento o el efecto multiplicativo de la inversión de las entidades financieras en 12,5 veces el valor de los recursos propios en sus balances, otra recomendación es la definición de capital regulatorio que se divide en dos categorías Tier I y Tier II, y el primer elemento a tener en cuenta es el Riesgo de Crédito. Este organismo hace público el acuerdo de Basilea I en 1988 donde se hacen recomendaciones necesarias dada la importancia de asegurar la estabilidad del sistema y mantener un capital mínimo con el que se cubran los fondos sujetos al riesgo de impago.

Debido a la rápida innovación de los productos financieros y la cantidad de vacíos regulatorios las sugerencias propuestas por el Comité de Supervisión Bancaria de Basilea se tuvieron en cuenta por parte de los entes regulatorios en la mayoría de los países del mundo, e inmediatamente las entidades implementaron mejoras en la administración de riesgos, para 1999 el Comité de Basilea se reúne de nuevo para redactar un acuerdo nuevamente (Basilea II), publicado en 2004, donde se amplía el tratamiento de los riesgos teniendo en cuenta el riesgo operacional y el riesgo de mercado, por lo que el nuevo acuerdo queda con elementos del acuerdo pasado y la definición de tres pilares (Ochoa P, Galeano M, Agudelo V, 2010):

- *Requerimientos mínimos de capital:* El primer pilar plantea unos requisitos mínimos de capital regulatorio con respecto a los riesgos de mercado, crédito y operacional, conservando el requisito mínimo de Basilea I (8 %) de capital económico frente a los activos ponderados por riesgo.
- *Proceso de examen supervisor:* Define cuatro principios generales para la actividad supervisora, con el fin de garantizar que los bancos estén cumpliendo con sus coberturas de riesgo.
- *La disciplina de mercado:*acceso y transparencia de la información suministrada por las entidades financieras.

Por consiguiente, la Superintendencia Financiera (SFC) mediante la circular Externa 11 y la Carta Circular 31 de marzo de 2002 modifica el capítulo II de la Circular Externa 100 de 1995 con la normatividad de regulación bancaria, ajustándo los cambios mundiales e implementando el Sistema Administrativo de Riesgo de Crédito(SARC). El sistema de Administración de Riesgo de Crédito, induce la gestión interna de todos los riesgos que deben afrontar las entidades financieras, en cuanto a la implementación de modelos que permitan vigilar la cartera de créditos hasta el momento de su último pago. Debe contar con los siguientes componentes básicos:

- Políticas de Administración de RC como son otorgamiento, recuperación, seguimiento y control.
- Procesos de administración del RC.
- Modelos internos o de referencia para la estimación o cuantificación de pérdidas esperadas.
- Sistema de provisiones para cubrir el RC.
- Procesos de control interno.

Las entidades vigiladas por la SFC, deben diseñar e implementar un SARC que les permita monitorear de manera eficaz su cartera de créditos, por lo cual se vuelve indispensable los modelos que ayuden a estos fines. Actualmente el ente regulador de Colombia solo tiene modelos de referencia para el seguimiento de cartera y el cálculo de provisiones, pero para el otorgamiento de créditos, los modelos aunque deben cumplir con lineamientos exigidos por SFC, esta no cuenta con un modelo de referencia, ya que cada entidad tiene un mercado particular y por consiguiente las características de los clientes a los que están dirigidos los créditos difiere en cada una de ellas. Sin embargo, el diseño de los modelos de otorgamiento debe contar con unas pautas específicas, dentro de los cuales se encuentra la implementación de información cuantitativa y cualitativa de los clientes, al igual de información previa al otorgamiento, capacidad de pago del cliente, garantías que respalden la deuda, entre otras. La Superintendencia Financiera en la Circular 22 de 2008 expone: "... todas las entidades obligadas a implementar el SARC que tengan cartera de consumo, deben establecer un modelo de otorgamiento de crédito que permita clasificar y calificar según el riesgo a los potenciales sujetos de crédito ...", por lo cual es importante la implementación de un modelo de otorgamiento de crédito para entidades financieras que cumpla con las condiciones dadas por SFC.

Partiendo de lo que expone la SFC y para evaluar el RC se han diseñado modelos estadísticos con ciertas características financieras y demográficas que permiten cuantificar la probabilidad de que un cliente incumpla una obligación financiera en un futuro, dentro de los métodos más usados para la evaluación del riesgo de crédito se encuentra la regresión logística debido a su poder predictivo y a su facilidad de interpretación, pero actualmente existen numerosas metodologías que también ayudan a estimar la probabilidad de incumplimiento.

En los años sesenta se crearon modelos donde se contemplaba la información financiera para estimar el riesgo de crédito, dentro de los trabajos más destacados se encuentran Tamari (1966), Altman (1968), para los setenta se empiezan a usar técnicas discriminantes y para los ochenta y noventa se emplean modelos con máxima verosimilitud y redes neuronales (Leal Fica , Aranguiz Casanova, Gallegos Mardones , 2018).

En 2005 plantearon un modelo de regresión logística binaria para estimar las perdidas en las que puede incidir una compañía financiera al otorgar créditos, este trabajo se realizó mediante varios pasos entre los cuales se encontraba análisis de componentes principales, regresión logística, entre otros, al fin concluían que la definición de default debía hacerse dependiendo la finalidad de cada negocio para así tener una buena estimación de perdida esperada, también proponían calcular el default dependiendo el ciclo económico, ya que las probabilidades de incumplimiento se veían afectadas por la coyuntura económica que influyen en la recesión o desaceleración de la economía (Aguas Gonzales Castillo H, 2005).

Un estudio en redes neuronales para la estimación de riesgo de crédito se publicó en 2007 en la revista Ingeniería Universidad de Medellín, en este artículo se trabajó con tres diferentes redes neuronales en donde la última, una red neuronal probabilística mostro los más altos porcentajes de clasificación, además este tipo de metodología permite discriminar ampliamente las probabilidades de incumplimiento y de cumplimiento (Fernandez Castaño Perez Ramirez , 2005).

En 2008 se realizó un estudio con una cooperativa de servicios financieros, se utilizó un Sistema de Inferencia Difuso para evaluar la capacidad de los solicitantes de crédito de la cooperativa, el resultado

fue un modelo que tiene en cuenta toda la información en la etapa de evaluación de un crédito para así minimizar el riesgo operativo y el riesgo de crédito, dentro de sus conclusiones resaltan que el sistema de inferencia difuso al aplicarse al análisis de crédito es una opción para la evaluación crediticia, pero dentro de las dificultades que se evidencian son la obtención de la base de conocimiento a partir de expertos y como evaluar la información (Medina Paniagua, 2008).

Por lo cual, todas estas metodologías utilizadas para RC han surgido como una herramienta útil en la evaluación y seguimiento de crédito, a continuación, se describen todas las metodologías que se implementaron para realizar este trabajo, desde las técnicas que se usan para agrupar y crear nuevas variables hasta las metodologías para comparar y definir el mejor modelo.

Las primeras metodologías que se implementaron en este trabajo fue el gráfico de barras que es una forma de agrupar un conjunto de datos por categorías, es usado para mostrar valores en diferentes momentos, también se usó tablas de frecuencia que es una herramienta en la que se muestra la distribución de los datos mediante sus frecuencias, se puede utilizar para variables cuantitativas o cualitativas (Murray R. Larry J., 2009).

4.1. Técnicas de Agrupamiento de Variables

La selección de variables es una etapa muy importante porque es indispensable que el modelo seleccionado sea parsimonioso, es decir, que tenga un gran poder predictivo y a su vez sea lo más sencillo posible, para cumplir esto es preferible que el modelo tenga el menor número de variables explicativas, esta etapa se puede realizar con infinitas metodologías como análisis multivariado, análisis discriminante, criterio de experto, entre otras, en esta ocasión realizaremos un análisis bivariado para conocer detalladamente y así poder tener una mejor selección de variables.

En el análisis bivariado se busca la asociación y causalidad entre dos variables, el objetivo es conocer las características y diferencias más básicas de la población y de los dos grupos que la componen (Buenos y Malos). Pero realizando este análisis se puede analizar la proporción de casos de la variable objetivo, es decir si existe una gran proporción de datos en alguna categoría de la variable dependiente puede generar problemas a la hora de clasificar, ya que se clasifican todos en la clase que contiene la mayoría de datos, a esta situacion se le conoce como desbalanceo de la base, y existen algunos metodos que ayudan a solucionar este problema. se puede intentar equilibrar las frecuencias de las clases, hay métodos de muestreo que pueden mitigar el efecto del desequilibrio en el conjunto de entrenamiento. Dos métodos de muestreo son Downsampling y Upsampling. Downsampling es una técnica que reduce el tamaño de la muestra para mejorar el equilibrio de dichas clases y Upsampling es una técnica que simula o atribuye datos adicionales para mejorar el equilibrio de las clases. Despúes de investigar estas tecnicas si se puede proseguir con las técnicas utilizadas para el agrupamiento de variables, en este trabajo se estudian los Odds y WOE, las cuales se describirán a continuación.

- *Odds:* es una medida estadística utilizada principalmente en epidemiología, y básicamente se define como la razón entre la probabilidad de que un evento ocurra y la probabilidad de que no ocurra, (Aedo, Pavlov, Clavero, 2010) los Odds se calculan para cada una de las categorías de una variable, entre mayor sea este, la categoría es mejor. Esta es un herrmamienta para re-categorizar la variable

dado que las categorías con Odds similares se pueden agrupar teniendo encuenta que la distribución entre buenos y malos es parecida.

A través de esta metodología, se seleccionan las variables que mejor discriminan la calidad del cliente dentro del conjunto de las variables mejor distribuidas, obteniendo un conjunto de variables regresoras con mayor potencial para el modelo.

- *Woe:* son sus siglas en inglés (Weight of Evidence), y es una medida de poder predictivo que mide la diferencia proporcional entre clientes incumplidos (malos) y cumplidos (buenos) en cada categoría de cada variable. Ayuda a considerar la contribución de independencia de cada una de las variables al modelo final, descartar linealidad entre variables(Cañete, 2018).

  La forma más frecuente para calcular el WOE es:

$$WOE = [log\,(Dm_{ij}/Db_{ij})] \times 100 \tag{1}$$

  Donde

  Dm_{ij} Distribución de malos en la característica ı

  Db_{ij} Distribución de buenos en la característica ı

  Si este valor es positivo significa que hay una mayor proporción de malos que de buenos en esa categoría, por lo cual esa categoría se clasificara como mala. Dentro de las características que se deben tener en cuenta al implementar esta metodología se encuentra el manejo que le dan a los valores atípicos ya que los agrupa en una categoría aparte, para que el análisis sea significativo se considera que en cada categoría debe haber un 5 % de los datos, si los grupos cuentan con menos del 5 % puede ocasionar inestabilidad a la hora de modelar. La número de grupos determina la cantidad de suavizados, es decir entre menos grupos más suavizados. Otra regla que se debe considerar es que en ninguno de los grupos debe haber cero personas "buenas" o "malas". La agrupación se debe hacer de tal manera que maximice la diferencia entre buenos y malos, cuanto mayor sea el valor del woe de los grupos, mayor será la capacidad predictiva de esta característica.

  Para calcular los valores de WOE's se siguen los siguientes pasos:

  - Si es una variable continua, se realiza un binning para agrupar las características, si la variable es cualitativa no es necesario este paso.
  - Se calcula el número y porcentaje de buenos y malos que cae en cada grupo (Bin).
  - Después de calcular el WOE mediante el la fórmula descrita anteriormente.

4.2. Técnicas para selección de variables

Como ya se mencionó anteriormente uno de los factores que ayuda a la eficacia de un modelo es el número de variables que este contenga. Sin embargo, las variables que se seleccionen deben ser las más robustas.

- *Information Value:* es una de las técnicas más utilizadas para seleccionar variables importantes en el modelo, técnica que viene de la teoría de la información que fue propuesta por el matemático

Claude Elwood Shannon y ayuda a clasificar las variables en función de su importancia. Se calcula de la siguiente manera:

$$IV = \sum_{i=1}^{n} (DistribucióndeIncumplidos - DistribucióndeCumplidos) \times WOE \tag{2}$$

$$IV = \sum_{i=1}^{n} \left(\frac{b_{ij}}{b_i} - \frac{m_{ij}}{m_i} \right) \times log \left(\frac{b_{ij} m_{ij}}{b_i m_i} \right) \tag{3}$$

El IV mide la fuerza de la relación que es medida por el WOE, y se debe tener en cuenta si las variables son continuas o no, ya que, al serlo se realizara un binning para agrupar las características y cuando la variable es cualitativa no es necesario contemplarlo. Para interpretar el valor de información IV se considera los siguientes rangos.

- $IV < 0.02$: Impredecible, la variable no es útil para modelar, no discrimina entre buenos y malos.
- $0.02 \leq IV \leq 0.1$: El poder de la predicción es débil.
- $0.1 \leq IV \leq 0.3$: El poder de predicción es medio.
- > 0.3 : El poder de predicción es fuerte.

Las variables con un valor de IV superior a 0.5 deben ser examinadas ya que indica que debe haber un problema de sobrepredicción y deberían quedar por fuera del modelo.

4.3. Metodologías utilizadas para la construcción del modelo de RC (Modelo Logit – Modelo GAM)

4.3.1. Modelo Logit

El modelo logit es un modelo de regresión que esta creado específicamente para ajustar una variable dependiente de tipo binario y las variables independientes pueden ser cuantitativas o categóricas, este modelo es usado principalmente para riesgo de crédito, donde ajustan la variable binaria como 1 si incumplió y 0 no incumplió. (Ladino, 2014) Con la regresión logística se quiere modelar la probabilidad de éxito (y=1), si la probabilidad estimada de que y=1 es mayor que 0.5 se clasificará al cliente como malo y si la probabilidad es menor a 0.5 como bueno.

Sea n el tamaño de cartera de créditos de una entidad finaciera y $x_i = (x_{i1} \dots x_{ip})$ un vector aleatorio el cual caracteriza el perfil de crédito del ı-ésimo solicitante, $1 \leq i \leq n$e y_i la variable respuesta asociada.

Sea $p_i = P(Y = 1|x_i)$, se define $g(p_i) = \beta_0 + \beta' X_i$ siendo $\beta' = (\beta_1, \dots, \beta_p)$ un vector de parámetros de coeficientes de las variables explicativas del modelo. En situaciones donde la variable dicotómica es habitual considerar como la función de la siguiente manera: (Cañete, 2018)

$$g(p_i) = log(\frac{p_i}{1 - p_i}) \tag{4}$$

El modelo logístico tiene la siguiente forma:

$$E(y) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} \tag{5}$$

Donde x es el vector de variables explicativas y β es el vector de parámetros, que también se puede expresar de la siguiente manera:

$$E(y) = \frac{1}{1 + e^{-x'\beta}} \tag{6}$$

O sea:

$$\pi_i = \frac{1}{1 + e^{-x'\beta}} \tag{7}$$

Que es equivalente a :

$$1 - \pi_i = \frac{1}{1 + e^{-x'\beta}} \tag{8}$$

Con lo que se tiene:

$$\frac{\pi_i}{1 - \pi_i} = \frac{1 + e^{x'\beta}}{1 + e^{-x'\beta}} = e^{x'\beta} \tag{9}$$

A esta transformación se le conoce como transformación logit de la probabilidad π_i y la relación $\pi_i/(1 - \pi_i)$ una razón de probabilidades o ventajas (odds ratio).

Si se toma el logaritmo se obtiene:

$$Ln\left(\frac{\pi_i}{1 - \pi_i}\right) = x'\beta \tag{10}$$

Con lo cual se tiene que el logaritmo de la razón de probabilidad es lineal, tanto en las variables como en los parámetros. Las estimaciones de estos pueden realizarse mediante el método de máxima verosimilitud. (Moscote Flórez Arley Rincón, 2012)

La estimación de los parámetros del modelo se realiza mediante el método de máxima verosimilitud. Como la variable Y_i tiene una distribución Bernoulli con parámetro π_i, la función de verosimilitud de forma:

$$P(Y_1, ..., Y_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{11}$$

y su logaritmo será:

$$logP(Y_1, ..., Y_n) = \sum_{i=1}^{n} [Y_i log(p_i) + (1 - Y_i) log(1 - p_i)] \tag{12}$$

Si se considera $\beta' = (\beta_0, \beta_1, ..., \beta_p)$ y $X_i' = (1, X_{1i}, ..., X_{ip}$ se puede escribir:

$$log(\frac{p_i}{1 - p_i}) = X_i'\beta \tag{13}$$

Sustiyendo esta expresión en la funcion log verosimilitud se obtiene la funcion de verosimilitud de logaritmos en términos de β como:

$$L(\beta) = \sum_{i=1}^{n} Y_i X_i' \beta - \sum_{i=1}^{n} log(1 + e^{X_i' \beta}) \tag{14}$$

Si se deriva la expresión $L(\beta)$ con respecto a los parámetros β_j y se iguala a cero se obtienen:

$$\sum_{i=1}^{n} = \sum_{i=1}^{n} X_i' \left(\frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right) = \sum_{i=1}^{n} X_i p_i \tag{15}$$

Existen varios métodos iterativos para resolver las ecuaciones de verosimilitud del modelo de regresión logística. Como las ecuaciones no son lineales en los parámetros β, se usa el método de Newton-Raphson.

La interpretación de los coeficientes estimados depende de la función de enlace, un cambio de unidad en un factor se refiere a una comparación de un determinado nivel de referencia. El enlace logit ofrece la interpretación más natural de los coeficientes estimados y, la interpretación utiliza el hecho de que las probabilidades de un evento de referencia son $P(evento)$ / $P(noevento)$. Cuanto mayor sean las probabilidades logarítmicas, más probable será el evento de referencia. Por lo tanto, los coeficientes positivos indican que el evento se vuelve más probable y los coeficientes negativos indican que el evento se vuelve menos probable.

4.3.2. Modelo Aditivo Generalizado

Los Modelos Aditivos Generalizados son una extensión de la regresión lineal múltiple, ya que estos modelos combinan la no linealidad y la regresión no paramétrica, fue propuesto por Hastie y Tibshirani en 1990. Este modelo está formado por la suma de funciones suaves de la variable predictora más conocidas como spline, una de las ventajas de este modelo es que no es necesario probar independencia ni normalidad de las variables y se representa de la siguiente manera:

$$Y = b_o + f_1(x_1) + f_2(x_2) + ... + f_m(x_m) + \varepsilon \tag{16}$$

Donde $f_i(x_i)$ son funciones polinómicas por tramos que tratan de explicar el cambio de la variable explicativa con las variables independientes sin tener en cuenta la parte explicada por las otras variables (Quintas, 2015).

Estos modelos fueron introducidos a la literatura por Marx and Eilers para el caso unidimensional en 1998 y por Durbán et al. En 2002 y Currie en 2004 para elc aso bidimensional(Durbán , 2015). Una de las grandes ventajas de estos modelos es que no es necesario sugerir ningún tipo de función existente entre las variables, el modelo es quien define la forma de la relación entre las variables, por lo que la forma de la función queda en términos de los datos y un parámetro de suavizamiento que ayuda a determinar qué tan cerca se ajusta la función a los puntos. (Quintas, 2015)

Permite el ajuste de variables con distribuciones como Poisson, Normal, Binomial o Gamma. Las estimaciones de estos modelos se realizan mediante un algoritmo llamado backfitting, el cual ajusta cada componente del modelo mediante suavizamientos parciales a los residuos.

Si se considera que el modelo de regresión lineal normal y el modelo de regresión lineal generalizado asumen una relación lineal para formar covariables $X_1, X_(2,)X_3, \ldots, X_p$ desde una regresión lineal usando un modelo con solo una variable predictora, que está dada por:

$$Y_i = \beta_0 + \beta X_{ij} + \varepsilon_i \tag{17}$$

Donde $\varepsilon \sim N(0, \sigma^2)$, Y_i es la variable respuesta proveniente de una distribución de la familia exponencial, α es el intercepto, β_j es la pendiente, $X_i j$ es la variable predictora y ε_i es el residuo. Se asume teóricamente en las regresiones lineales, que la distribución de la variable respuesta es normal y la relación entre la variable respuesta y predictiva es lineal.

En el caso de los modelos de clase aditivos generalizados (GAM), se reemplaza a la forma lineal $sum\ \beta_j X_i j$ por una suma de funciones suavizadas $\sum S_j(X_i j)$ los $s_j(.)s$ son funciones no especificadas que se estiman utilizando un diagrama de dispersión más suave, en un proceso iterativo que llamamos el algoritmo de calificación local.

El GAM expande el componente sistemático de un modelo lineal generalizado basado en la función de suavización de las covariables. Para poder comprender los modelos aditivos generalizados, la forma más fácil de iniciar es desde una regresión lineal usando solo una variable predictora, la primera y más notable diferencia es que se usa una función para encontrar la relación entre $_i$ y $_j$ donde:

$$Y_i = \beta_0 + f(X_{ij}) + \varepsilon_i \tag{18}$$

Donde $\varepsilon \sim N(0, \sigma^2)$

A simple vista la única diferencia entre los dos modelos es el reemplazo de $\beta_j X_i$ por $f(X_i)$ de la ecuación anterior. Sin embargo, este cambio es muy importante. La regresión lineal da como resultado una fórmula y la relación entre las variables respuestas y predictoras es cuantificada por un solo parámetro de regresión. En cambio, en el GAM se tiene una función que relaciona la variable respuesta con las variables predictoras mediante suavizadores. Donde $f(X_i)$ es calculada usando una función base:

$$f(X_i) = \sum_{j=1}^{p} \beta_j \times b_j(X_i) \tag{19}$$

Donde β son los parámetros y los b_j (X_i) son los suavizadores. Dentro de los suavizadores se encuentra los spline regresivos penalizados que están expresados por:

$$\|Y - X\beta\|^2 = \lambda \int f''(x)^2 dx \tag{20}$$

En la ecuación anteriormente descrita al estimar los parámetros β también se estiman los suavizadores . La primera parte de la Ecuación, $\|Y - X\beta\|^2$, se refiere al ajuste de la curva mediante el método de mínimos cuadrados. En la suma de cuadrados se obtiene los parámetros β_j, y es útil

para medir la cercanía de las estimaciones de los datos. La suma de los cuadrados a minimizar está dada por:

$$S = \sum_{j=1}^{n} (Y_j - \mu)^2 = \|Y - \mu\|^2 = \|Y - X\beta\|^2 \tag{21}$$

Donde Y_j es el valor observado o valor estimado para cada observación, y j es el número de observaciones. En esta ecuación la notación $\|\|$ es debido a la norma Euclideana; Y contiene todos los valores observados en formato de vector, β son los parámetros y X son las funciones base o suavizadores.

La segunda parte de la ecuación (15) $(\lambda \int f''(x)^2 dx)$ es una penalización, por esta razón se llama spline penalizados, y mide el grado de fuerza en la estimación de la función. Esta parte de la ecuación contiene un parámetro suavizado λ, λ es una constante de ajuste que afecta el balance entre la precisión y el suavizado. Este parámetro es afectado por una integral de la segunda derivada de la función de suavizado f que muestra cuan suavizado es la curva.

Un alto valor de f significa que el suavizador f es altamente no lineal y la curva es muy flexible, mientras que un valor de cero en la segunda derivada es una línea recta. También si λ es muy grande, la penalización por no tener una curva no suavizada es grande y el suavizador resultante puede ser una línea y si λ es pequeño hay una baja penalización para la no suavización y probablemente se obtendrá una curva menos suavizada.

Este control en los dos sentidos hace necesario que se busque un punto intermedio entre baja penalización y optimo suavizado. Para alcanzar el valor optimo sin tener que calcular λ, se implementó el cálculo de GCV (Validacion Cruzada Generl) para estimar los parámetros de suavizado, en cambio de proveer un resultado de la validación cruzada en términos de λ, las modificaciones utilizan grados de libertad efectivos. Entre más alto sea el valor de los grados de libertar es menos lineal la función de suavizado utilizada.

En el GAM no se asume que las variables provengan de una distribución normal, la distribución puede cualquiera dentro de la familia exponencial, la elección de la distribución puede hacerse a priori basada en el conocimiento de las variables respuesta.

Otro aspecto muy importante es la relación entre la variable respuesta y las variables predictoras con una función de enlace, las funciones de enlace son las responsables de modelar la relación entre estas variables y existen varios tipos de funciones de enlace para la misma familia de distribución como se ve en el siguiente cuadro.

Distribución	Tipo de datos	Funciones de enlance
Normal	Continuos	identity, log, inverse
Poisson	Discretos (conteos)	log, identity, sqrt
Quasipoisson	Discretos (conteos)	logit, probit, cloglog, identity, inverse, log, sqrt
Negativa binominal	Discretos sobredispersos	log,sqrt, identity
Gamma	Continuos	inverse, identity, log
Binomial	Proporcionales	logit, probit, cauchy, log, cloglog
Bernoulli	Presencia y ausencia	logit, probit, cauchy, log, cloglog

Tabla 1: Funcion de Enlace por tipo de Distribución.

La estimación de los parametros se realiza por medio de maxima verosimilitud penalizada la cual se encarga de penalizar la curvatura excesiva de las estimaciones de los términos f_j, se puede imponer la penalización mediante:

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \beta^t S_j \beta, \tag{22}$$

La cual se puede escribir de forma matricial

$$l_p(\beta) = l(\beta) - \frac{1}{2}\beta^t S_j \beta \tag{23}$$

Esencialmente, puede presentar los resultados del modelo de un GAM como si fuera cualquier otro modelo lineal, la principal diferencia es que para los términos suaves, no hay un coeficiente único desde el que pueda hacer inferencia (es decir, negativo, positivo, tamaño del efecto, etc.) puede incluir términos lineales normales en el modelo (ya sea continuo o categórico, e incluso en un marco de tipo ANOVA) y hacer inferencia de ellos como lo haría normalmente. De hecho, los GAM a menudo son útiles para dar cuenta de un fenómeno no lineal que no es de interés directo, pero que debe tenerse en cuenta al hacer inferencia sobre otras variables.

Los GAM tiene aplicación en muchos campos de regresión siempre y cuando la variable respuesta pertenezca a la familia exponencial (Binomial, Gaussiana, Poisson, Binomial Negativa, etc.). Estos modelos pueden ser utilizados con fines exploratorios, predictivos o explicativos. En este caso lo utilizaremos para predecir la probabilidad de incumplimiento (Sior Alegre Norza, 2011).

4.4. Técnicas de selección del modelo

En esta sección explicaremos los criterios de selección que se implementaron en este trabajo para seleccionar el mejor modelo para la probabilidad de incumplimiento.

4.4.1. Probabilidad de mala clasificación

Es una técnica de predicción discriminativa que utiliza una matriz (matriz de confusión) donde las filas representan los datos reales y las columnas los datos predichos por el modelo para dar una

idea de cómo está clasificando el modelo a partir de un conteo de aciertos y errores de cada uno de los subgrupos como se muestran a continuación:

Valores Predichos / Valores Reales	Positivo	Negativo
Positivo	Verdaderos Positivos (a)	Falsos Negativos (b)
Negativo	Falsos Positivos (c)	Verdaderos Negativos (d)

Tabla 2: Matriz de Confusión.

Donde:

- a es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
- b es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
- c es la cantidad de negativos que fueron clasificados incorrectamente como positivos.
- d es la cantidad de negativos que fueron clasificados correctamente como negativos.

Se han definido varios indicadores para medir el desempeño del modelo.

Exactitud Porcentaje de datos clasificados correctamente. $\frac{a+d}{a+b+c+d}$

Tasa de error Porcentaje de datos clasificados incorrectamente. $\frac{b+c}{a+b+c+d}$

Sensibilidad Proporción de casos clasificados como buenos correctamente. $\frac{a}{a+b}$

Especificidad Proporción de casos clasificados como malos correctamente. $\frac{d}{c+d}$

4.4.2. Aréa bajo la curva (AUC) y Curva ROC

Dentro de las técnicas de análisis más utilizadas para comparar el poder discriminante de los modelos de clasificación se encuentran el AUC (por sus siglas en ingles Area Under the Curve) y la curva ROC, una de las ventajas que ofrecen estas medidas de validación son la flexibilidad de su construcción, ya que no dependen de supuestos sobre el método probabilístico que subyace en la regla discriminante. El AUC se utiliza para estimar la capacidad de discriminar entre morosos y no morosos. También se usa para comparar pruebas entre sí y determinar cuál es la más eficaz y se define como:

$$AUC = \int_0^1 (1 - Especificidad, Sesibilidad) \tag{24}$$

Puede tomar valores desde 0 hasta 1. Si el AUC diera 1 la prueba sería perfecta, es decir que clasificaría al 100 % de los morosos como morosos y al 100 % de los cumplidos como cumplidos y,

por tanto, los grupos estarían perfectamente diferenciados y bien clasiicados por la prueba. Si el área de la curva valiese 0.5, significaría que la probabilidad de clasificación incorrecta es de 0.5 lo que seria una prueba poco útil, la diagonal representa una situación como la de clasificación de 0.5. Si el área bajo la curva es menor al 0.5, significaría que la tiene baja capacidad de clasificación. Por lo tanto, cuánto mayor sea el valor del AUC mejor será la capacidad discriminante y se obtienen mejores resultados (Cañete, 2018).

Una interpretación de los valores de AUC es la siguiente:

- Baja capacidad discriminante: [0.5, 0.69).
- Capacidad discriminante útil: [0.7, 0.9).
- Alta capacidad discriminante: [0.9, 1].

En cuento a la curva ROC es la representacion gráfica en el que se observan la Sensibilidad y el complementario de la Especificidad, es decir (1-especificidad). Esta gráfica tienen una enorme cantidad de aplicaciones, siendo muy utilizadas en los ámbitos del RC y también en medicina entre otros.

4.4.3. Validación Cruzada o Cross-Validation

Es una técnica utilizada para la evaluación de resultados garantizando independencia entre subconjuntos de datos de entrenamiento y prueba, en la cual se mira el porcentaje de correcta clasificación, consta en clasificar nuevamente los clientes con el modelo obtenido, cuantificando los clientes bien clasificados para cada una de las categorías ("Buenos" y "Malos") Una de las ventajas es que es muy fácil a la hora de computar.

Consiste en dividir en dos conjuntos complementarios los datos de muestra, realiza el análisis de un subconjunto llamado entrenamiento y valida en otro conjunto llamado prueba, de forma que la función de aproximación solo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de prueba, existen principalmente tres tipos de validaciones cruzadas.

- **Validación cruzada aleatoria:** Este método consiste al dividir aleatoriamente el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Para cada división la función de aproximación se ajusta a partir de los datos de entrenamiento y calcula los valores de salida para el conjunto de datos de prueba. El resultado final se corresponde a la media aritmética de los valores obtenidos para las diferentes divisiones.
- **Validación cruzada dejando uno fuera:** La validación cruzada dejando uno fuera o Leave-one-out cross-validation (LOOCV) implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento. La evaluación viene dada por el error, y en este tipo de validación cruzada el error es muy bajo, pero en cambio, a nivel computacional es muy costoso, puesto que se tienen que realizar un elevado número de iteraciones.

- **Validación Cruzada de K iteraciones:** En este tipo de validación, la muestra se divide en k subconjuntos, uno de los subconjuntos se utiliza como datos de prueba y los demás (k-1) como datos de entrenamiento. Este proceso de validación cruzada es repetitivo durante k iteraciones con cada uno de los subconjuntos de datos de prueba, y asi finalmente se realiza la media aritmética de todos los resultados de cada una de las iteraciones para obtener un único resultado final, para este proyecto se utilizó este tipo de validación.

5. Datos

La base de datos con la que se construyeron los modelos fue aportada por una entidad financiera Colombiana, y se caracteriza por tener en sus registro clientes con activos menores a 145 SMMLV, pertenecientes al segmento de microfinanzas, dichos créditos se especifican por ser destinada a personas con establecimiento comercial y por lo general su destino no es agropecuario, el monto solicitado no debe superar los 40 salarios mínimos legales vigentes, esta base está compuesta por 25 variables y 321.261 registros comprendidos en un periodo de Marzo de 2014 hasta Febrero de 2018, es importante mencionar que el modelo que implementa en este trabajo son modelos de comportamiento, es decir modelos que ayudan a predecir el comportamiento de pago de los clientes que actualmente están vinculados con la entidad y los clasifica en clientes buenos o malos con el fin de crear un puntaje score con la ponderación de cada una de las variables seleccionadas en los modelos.

- Variable dependiente: La variable objetivo se llama $BMReal$, es una variable binaria donde 1 significa que el cliente incumplió con su obligación en algún momento dentro del periodo de observación y lo llamaremos Malo y 0 que no incumplió ningún día en el periodo de observación y los llamaremos Bueno, esta variable fue construida a partir de la variable días de mora la cual muestra el número de días en la que una persona dejo de pagar su obligación financiera, para el Banco y para este trabajo se contempla el incumplimiento después de los 30 días. grafica
- Variables Independientes: Las variables explicativas que se contemplan en los dos modelos son variables que principalmente describen características demográficas y comportamientos financieros, están divididas en diferentes tipos de variables (Discretas, continuas y dummys).

  A continuación, se muestra la descripción de algunas de las variables incluidas en la base.

Variable	Tipo	Descripción	Valores
BM Real	Dummy	Si alguna vez el cliente incumplió con sus obligaciones (dias mora >30)	0: No incumplió 1: Si incumplió
fo banco	Discreta	Numero de tramite del crédito	
Registro	Discreta	Numero de registro del crédito	
Regional	Nominal	Código de la regional en la que se solicita el crédito	10000 Antioquia 20000 Bogotá 30000 Cafetera

Variable	Tipo	Descripción	Valores
			40000 Costa 50000 Santander 60000 Occidente 70000 Oriente 80000 Sur
Endeudamiento Consolidado	Discreta	Valor de endeudamiento consolidado del cliente	
Edad	Discreta	Edad	
Género	Nominal	Género del Cliente	F: Femenino M: Masculino
Estado Civil	Nominal	Estado Civil del solicitante	CA: Casado SO: Soltero DI: Divorciado SE: Separado VI: Viudo UN: Unión Libre
Nivel de Educación	Nominal	Nivel educativo del Cliente	001: Primaria 002: Secundaria 003: Técnico 004: Tecnológico 005: Universitario 006: Especialización 007: Postgrado 008: Maestría 009: Doctorado 010: PHD 011: Otros 012: Ninguno
Tipo Actividad	Nominal	Actividad Económica del Cliente	0001: Empleado 0002: Independiente 0003: Comerciante 0004: Ganadero 0005: Agricultor 0006: Empresario 0007: Pensionado 0008: Servicios 0009: Pecuario

Variable	**Tipo**	**Descripción**	**Valores**
Tipo Vivienda	Nominal	Tipo de Vivienda del Cliente	0010: Producción Pro: Propio ARR: Arrendado FAM: Familiar ASI: Asignado por la empresa
Tiempo en la actividad (meses)	Discreta	Tiempo en meses que el cliente a dedicado a su actividad princial	
Número de personas	Discreta	Número de personas a cargo	
Número de empleados	Discreta	Número de empleados que cuenta el cliente	
Experiencia BAC	Nominal	Código que identifica si la persona tiene experiencia con el Banco	S: Si N: No
Nro obligaciones	Discreta	Número de obligaciones que cuenta el cliente	
Cuentas Embargadas	Nominal	Tiene Cuentas Embargadas	S: Si N: No
Calificacion Real	Nominal	Calificación de mayor riesgo otorgada al cliente en sus obligaciones con el sector real	A: Calificacion A B: Calificacion B C: Calificacion C D: Calificacion D E: Calificacion E K: Calificacion K N: No aplica
Endeudamiento	Discreta	Porcentaje de endeudamieto en el sector financiero	
Razón Corriente	Discreta	Activo Corriente/ Pasivo Corriente	
Indice Endeudamiento	Discreta	Total Pasivo / Total Activos	
Ingreso $_{E}greso$	Discreta	Ingreso / Egreso	
Margen Neto	Discreta	Utilidad Neta / Ventas Netas	
Fuentes Alternativas	Nominal	Tiene otras fuentes alternativas	S: Si

Variable	Tipo	Descripción	Valores
			N: No
Pasivo Patrimonio	Discreta	Pasivo / Patrimonio	

Tabla 3: Describción de Variables.

6. Resultados

En principio, se hizo un análisis exploratorio para saber si alguna variable no cuenta con el 100 % de la información o presenta algunos datos faltantes, erróneos o atípicos, es decir que muestren variaciones significativas, y con los cuales se pueden solucionar siempre y cuando no alteren la distribución de la variable. Para este caso se eliminaron 6.922 registros, que equivalen al 2 % de la información, que contaba con datos erróneos o vacíos o atípicos.

Luego de eso se empezó con el análisis univariado y exploratorio por medio de gráfico de barras y tablas de frecuencia el cual se muestra a continuación.

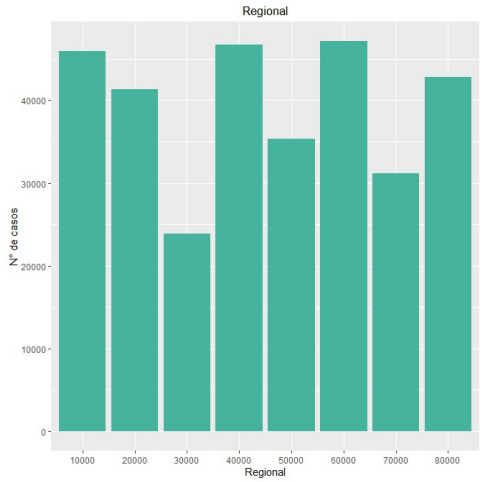

Tabla 4: Distribución de Regional

Regional	Frecuencia	%
10000	45925	15 %
20000	41351	13 %
30000	23842	7 %
40000	46708	15 %
50000	35316	11 %
60000	47145	15 %
70000	31181	10 %
80000	42865	14 %

Tabla 5: Tabla de Frecuencia por Regional.

Se observa que la Regional con menos clientes es Cafetera con 7 %, y la que presenta mayor número de clientes es Antioquia, Costa y Occidente con 15 %. Pero en general la proporcion de clientes por Rengional es equilibrada.

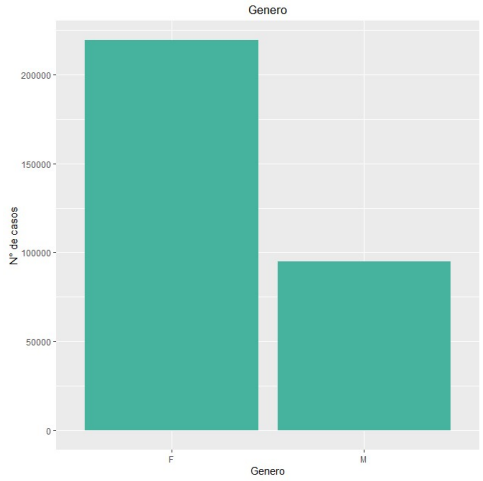

Tabla 6: Distribución de Género

Genero	**Frecuencia**	%
Femenino	219316	70 %
Masculino	95017	30 %

Tabla 7: Tabla de Frecuencia por Género.

Para el caso de la variable Género se muestra una diferencia, ya que el 70 % de la base son mujeres y el 30 % son hombres.

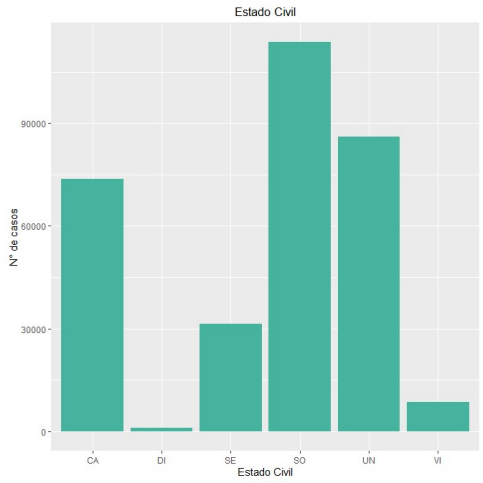

Tabla 8: Distribución de Estado civil

Estado Civil	**Frecuencia**	%
CA	73776	23 %
DI	930	3 %
SE	31374	9 %
SO	113666	36 %
UN	86111	27 %
VI	8476	2 %

Tabla 9: Tabla de Frecuencia por Estado Civil

En la variable Estado Civil, las categorías Divorciado y Viudo presenta pocos registros, ya que solo contiene el 5 % de la base, la categoría de Solteros contiene el 36 % y es la que contiene más datos.

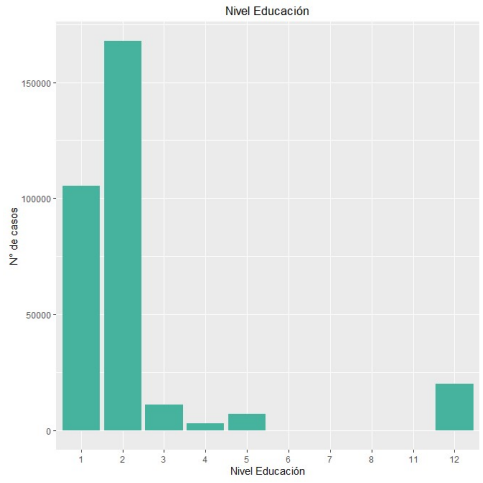

Tabla 10: Distribución de Nivel educativo

Nivel de Educación	Frecuencia	%
1	105339	33.51 %
2	167718	53.36 %
3	11071	3.52 %
4	3023	0.96 %
5	7078	2.25 %
6	85	0.03 %
7	62	0.02 %
8	6	0.01 %
11	41	0.01 %
12	19910	6.33 %

Tabla 11: Tabla de Frecuencia por Nivel de Educación

El 89 % de la base son clientes con un nivel de educación de primaria y secudaria, sólo el 1 % presentan posgrados.

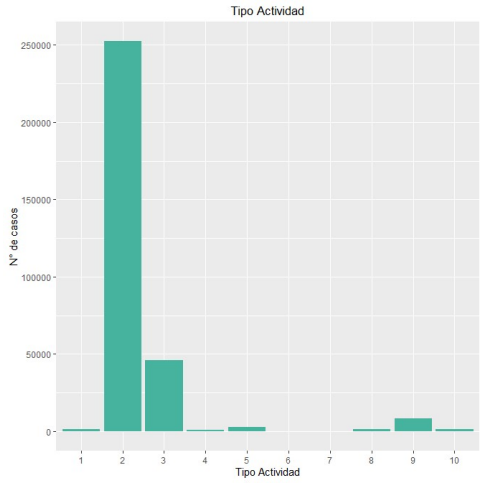

Tabla 12: Distribución de Tipo Actividad

Tipo Actividad	Frecuencia	%
1	1440	0.46 %
2	252342	80.28 %
3	45794	14.57 %
4	745	0.24 %
5	2586	0.82 %
6	6	0.00 %
7	36	0.01 %
8	1461	0.46 %
9	8395	2.67 %
10	1528	0.49 %

Tabla 13: Tabla de Frecuencia por Tipo Actividad

Para el caso de Tipo Acividad, es muy parecido a la variable Nivel de Educación, ya que el 80 % es Independiente.

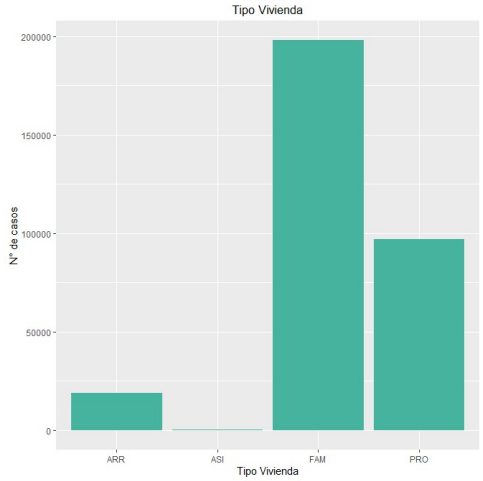

Tabla 14: Distribución de Tipo de Vivienda

Tipo Vivienda	**Frecuencia**	%
ARR	18851	6.00 %
ASI	469	0.15 %
FAM	197990	62.99 %
PRO	97023	30.86 %

Tabla 15: Tabla de Frecuencia por Tipo de Vivienda

La mayoría de los clientes de segmento de Microfinanzas tienen un tipo de vivienda familiar o propio, ya que son el 94 % de la población.

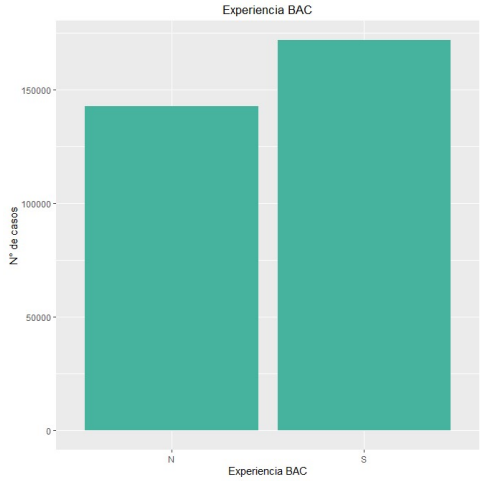

Tabla 16: Distribución de Experiencia BAC

Experiencia BAC	**Frecuencia**	%
Si	171603	45.41 %
No	142730	54.59 %

Tabla 17: Tabla de Frecuencia de Expeiencia BAC

Se puede observar que para el caso de Experiencia con el Banco,la distribucion de los datos es ma equilibrada, ya que el 45 % de los clientes si teiene experiencia con la entidad finnciera y el 55 % restanto no.

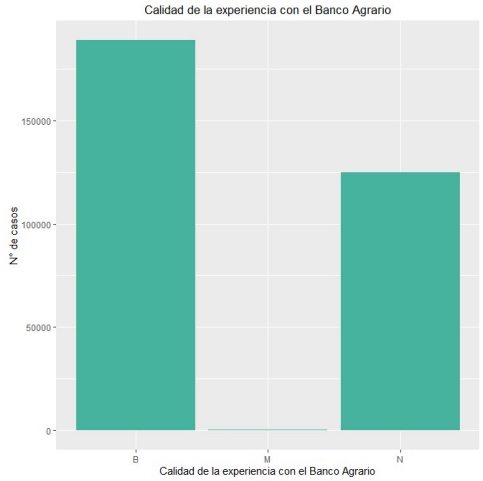

Tabla 18: Distribución de Calidad en la Experiencia con el Banco

Calidad Experiencia BAC	Frecuencia	%
Buena	188956	60.11 %
Mala	328	0.10 %
No tiene	125049	39.78 %

Tabla 19: Tabla de Frecuencia de Calidad de Expeiencia BAC

El 60 % de la base tiene una clificación buena con el Banco, lo que nos indica que son clientes catalogados como bueno, el 40 % no tiene calificación.

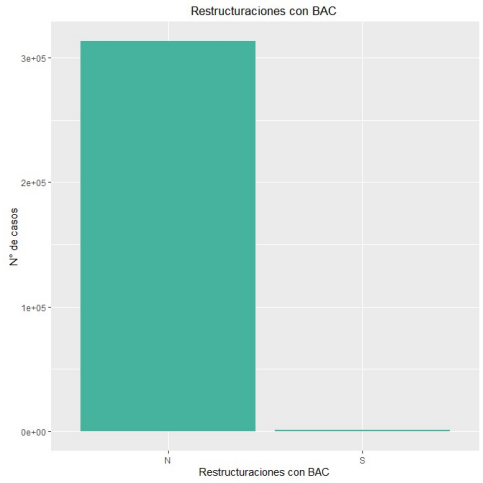

Tabla 20: Distribución de Restructuraciones con el Banco

Restructuraciones con BAC	Frecuencia	%
Si	1160	0.37 %
No	313173	99.63 %

Tabla 21: Tabla de Frecuencia de Restructuraciones con el Banco

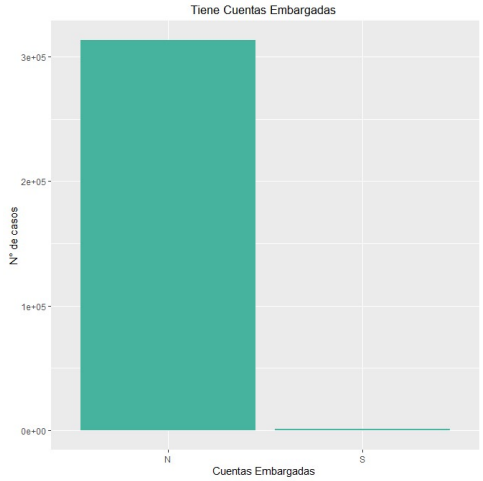

Tabla 22: Distribución de Cuentas Embargadas

Restructuraciones con BAC	Frecuencia	%
Si	1156	0.37 %
No	313177	99.63 %

Tabla 23: Tabla de Frecuencia de Cuentas Embargadas

El 99 % de la población no tiene Restructuraciones con el banco , significa que no ha cambiado las condiciones iniciles del crédito, tampóco tiene cuentas embargadas en el sector financiero.

En general se evidencia que variables como, Experiencia Bac, Tipo Actividad, Tipo Vivienda entre otras contienen más del 50 % de la base en una sola categoría, un ejemplo de este es "Tipo Actividad" el cual el 80 % de los datos está concentrado en la actividad 2 (Independiente), otra variable con las mismas condiciones es "Cuentas Embargadas" la cual el 99 % de la población no cuenta con cuentas embargadas. También se encuentran variables con una muy buena distribución como es el caso de "Regional" ya que todas las categorías contienen un porcentaje parecido, alrededor del 13 %, pero es importante analizar las variables de manera diferente y contemplando la distribución de nuestra variable objetivo (BM), por lo cual procedemos al análisis bivariado.

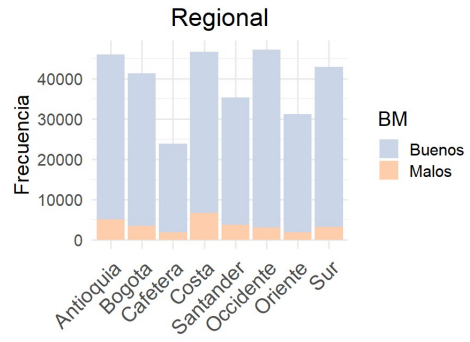

Tabla 24: Distribución de Regional por BM

Regional	Buenos	Malos	Odds
Antioquia	40856	5069	8.060
Bogotá	37871	3480	10.882
Cafetera	21964	1868	11.758
Costa	40047	6661	6.012
Santander	31665	3651	8.673
Occidente	44104	3041	14.503
Oriente	29363	1818	16.151
Sur	39680	3185	12.458

Tabla 25: Tabla de Frecuencia de Regional por BM

Para el caso del análisis bivariado se utilizó la técnica de agrupamiento *Odds*. En la variable Regional se observa que las categorías Costa y Santander contiene bastante número de clientes catalogados como mo-

rosos, esto significa que por cada persona incumplida sólo hay 6 y 8 personas cumplida respectivamente, en cambio para regionales como Oriente Occidente o Sur este valor es mayor.

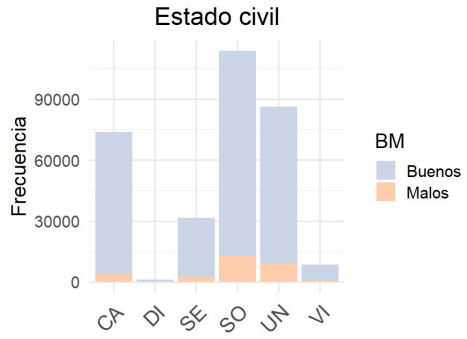

Tabla 26: Distribución de Estado Civil por BM

Estado Civil	Buenos	Malos	Odds
CA	69812	3964	17.612
DI	869	61	14.246
SE	28622	2747	10.419
SO	101083	12578	8.036
UN	77145	8966	8.604
VI	8019	457	17.547

Tabla 27: Tabla de Frecuencia de Estado Civil por BM

En el caso de Estado Civil las categorías Soltero y Unión Libre son las que presentan una mayor proporción de clientes morosos.

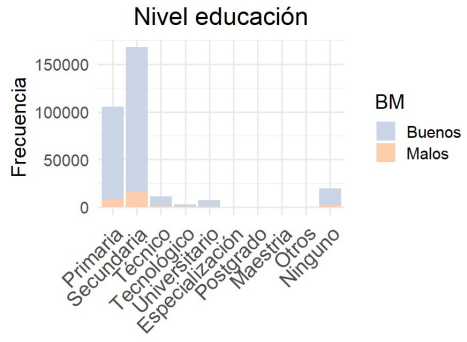

Tabla 28: Distribución de Nivel de Educación por BM

Nivel Educación	Buenos	Malos	Odds
Primaria	96673	8663	11.159
Secundaria	151426	16285	9.298
Técnico	10151	920	11.034
Tecnológico	2753	270	10.196
Universitario	6571	507	12.961
Especialización	81	4	20.250
Postgrado	60	2	30.000
Maestría	5	1	5.000
Otros	39	2	19.500
Ninguno	17791	2119	8.396

Tabla 29: Tabla de Frecuencia de Nivel de Educación por BM

En la variable Nivel de Educación se observa que las personas con alto nivel de esudios se caraterizan por ser personas cumplidas.

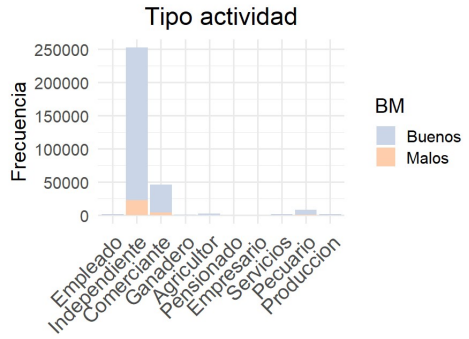

Tabla 30: Distribución de Tipo de Actividad por BM

Tipo Actividad	Buenos	Malos	Odds
Empleado	1288	149	8.644
Independiente	229301	23035	9.954
Comerciante	41851	3943	10.614
Ganadero	665	80	8.313
Agricultor	2280	306	7.451
Pensionado	5	1	5.000
Empresario	34	1	34.000
Servicios	1326	135	9.822
Pecuario	7419	976	7.601
Producción	1381	147	9.395

Tabla 31: Tabla de Frecuencia de Tipo Actividad por BM

En la actividad de los clientes, ocupaciones como independiente y comerciante presenta un mayor número de clientes no morosos.

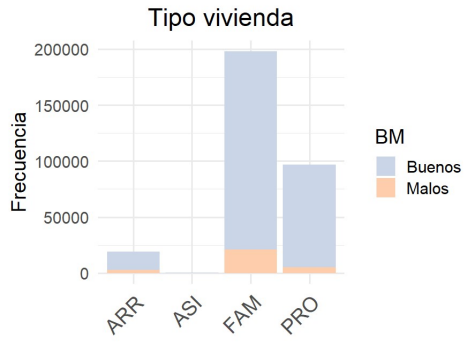

Tabla 32: Distribución de Tipo de Vivienda por BM

Tipo Vivienda	Buenos	Malos	Odds
ARR	16299	2552	6.387
ASI	431	38	11.342
FAM	177099	20881	8.481
PRO	91721	5302	17.299

Tabla 33: Tabla de Frecuencia de Tipo Vivienda por BM

Los clientes más cumplidos tiene viviendas propias, ya que por una persona incumplida hay 17 perosnas cumplidas para este tipo de vivienda.

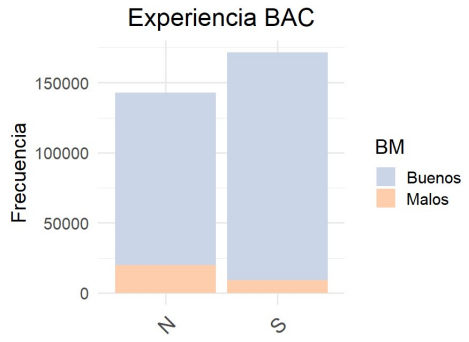

Tabla 34: Distribución de Experiencia BAC por BM

Experiencia BAC	Buenos	Malos	Odds
N	123024	19705	6.243
S	162526	9068	17.923

Tabla 35: Tabla de Frecuencia de Experiencia BAC por BM

En el caso de Experiencia BAC son más las personas morosas que no tiene experiencia con el Banco.

Se comprueba por medio de *Odds* la discriminación entre la variable objetivo y algunas variables como lo son Género, Tipo Vivienda, Experiencia BAC y Estado Civil lo cuál es bueno, las mismas variables del análisis univariado donde algunas categorías contenienen más del 50 % de la población en este análisis no resultan tan factibles analizarlas, lo cual da un indicio que son variables que el modelo no tendrá en cuenta para predecir la probabilidad de incumplimiento.

Al terminar con los análisis de variables el siguiente paso es separar la base en dos subconjuntos que se utiliza de la siguiente manera, la base de datos de entrenamiento es utilizada para estimar los parámetros del modelo y contiene el 70 % de los datos y el subconjunto de prueba 30 % o base de prueba ayuda a comprobar la eficacia del modelo. Para llevar a cabo este procedimiento se implementó el muestreo aleatorio simple con la función "createDataPartition" del software R-Studio, y así seguir con las técnicas de agrupamiento y selelcción de varaibles.

6.1. Técnicas de Agrupamiento de Variables

Se calcula el peso de evidencia (Woe), el cual mide la diferencia entre clientes incumplidos (malos) y cumplidos (buenos) en cada categoría de cada una de las variables.

Las columnas Distribución, Distribución Buenos y Distribución Malos se refieren al porcentaje de distribución de cada grupo del total, de los buenos y de los malos, respectivamente. Por ejemplo, para el grupo los que tienen experiencia BAC hay 120.139 clientes catalogados como malos, estos representan el 54 % de los clientes de la base. De estos clientes, 6.308 son malos y representan el 36 % de los clientes. El WOE para este grupo se calcula:

$$WOE = [log\,(31.3\,\%/56.9\,\%)] \times 100 = -59.8. \tag{25}$$

Como la variable objetivo esta definida como 1= Si incumplió en algun momento en la obligación y 0= si

Variable	Categoría	Total (Datos)	Distribución %	Total (1)	Total (0)	Distribución Malos	Distribución Buenos	Rate	WOE	IV
Experiencia BAC	S	120139	54.6 %	6308	113831	31.3 %	56.9 %	94.7 %	-59.8	0.153
	N	99885	45.5 %	13833	86052	68.7 %	43.1 %	86.2 %	46.7	0.120
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.273
Edad	<=29	49278	22.4 %	7761	41517	38.5 %	20.8 %	84.3 %	61.8	0.110
	>29	170746	77.6 %	12380	158366	61.5 %	79.2 %	92.7 %	-25.4	0.045
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.155
Tiempo Actividad (Meses)	<=60	78427	35.5 %	10718	67709	53.2 %	33.9 %	86.3 %	45.2	0.087
	>= 60	141597	64.4 %	9423	132174	46.8 %	66.1 %	93.3 %	-34.6	0.067
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.154
Pasivo Patrimonio	<0	130969	59.5 %	8682	122287	43.1 %	61.2 %	93.4 %	-35.0	0.063
	>0	20586	9.4 %	2549	18037	12.7 %	9.0 %	87.6 %	33.8	0.012
	0	68469	31.1 %	8910	59559	44.2 %	29.8 %	87.0 %	39.5	0.057
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.133
Género	F	153347	69.7 %	10867	142480	54.0 %	71.3 %	92.9 %	-27.8	0.048
	M	66677	30.3 %	9274	57403	46.0 %	28.7 %	86.1 %	47.2	0.082
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.130
Tipo Vivienda	Propio	68009	30.9 %	3668	64341	18.2 %	32.2 %	94.6 %	-57.0	0.080
	Por legalizar	332	0.2 %	25	307	0.1 %	0.2 %	92.5 %	-21.3	0.000
	Arrendado y Familiar	151683	68.9 %	16448	135235	81.7 %	67.7 %	89.2 %	18.8	0.026
	Total	220024	100.0 %	20141	199883	100.0 %	100.0 %	90.8 %	NA	0.106

no incumplió, entre mayor sea el porcentaje de WOE significa que esa categoría contiene mayor número de morosos.

Se evalúa en cada una de las variables si tiene coherencia el valor de WOE que da como resultado con el negocio, en caso de no tener una buena interpretación, no se tendrá encuenta esa variable en el momento de modelar. Veámos un ejemplo, la variable edad fue categorizada por el WOE como a 29 años y $>$ a 29 años, al observar el valor obtenido por los WOE's se interpreta que las personas menores o iguales a 29 años son más riesgosas que las personas mayores a 29, ya que el valor de WOE para esta categoría es mayor, lo que significa que es coherente ya que entre mayor edad, más experiencia en el sector financiero y menor morosidad.

6.2. Técnicas de Selección de Variables

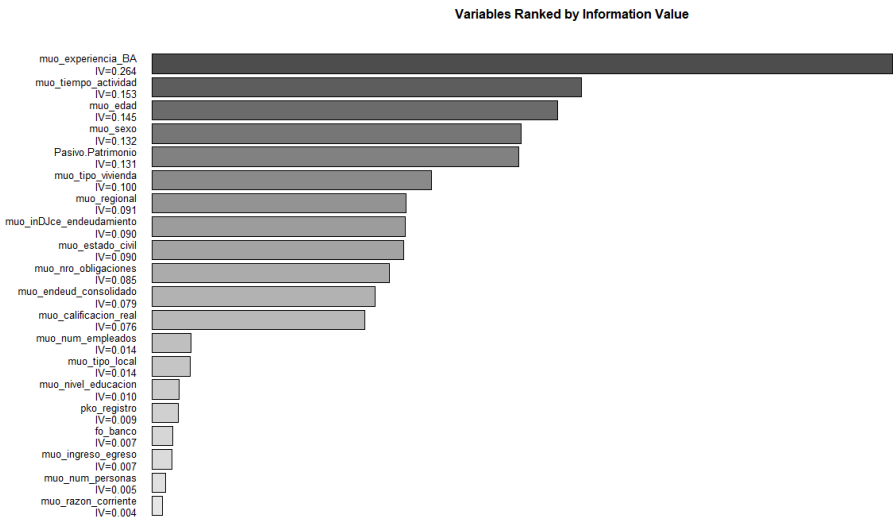

Figura 1: Ranking de Variables por el Valor de Información

Por medio del valor de información se analiza cuáles son las variables que más aportan al modelo según el criterio anteriormente descrito en el cual se debe seleccionar las variables con in IV mayor al 10 %, para este caso las variables son: Experiencia con el Banco, Edad, Tiempo en la actividad, Pasivo/Patrimonio, Género y Tipo Vivienda, todas estas variables ya fueron analizadas con sus valores de WOE y tienen una coherencia con el negocio, por lo cual no se elimina ninguna.

El paso a seguir fue balancear la base ya que la distribución de la variable dependiente (BM) es del 91 % categorizados como Buenos (0) y 9 % categorizados como malos (1) y esta situación de desbalanceo puede generar problemas a la hora de clasificar. Por regla de la entidad financiera el proceso de balanceo debe quedar 70 % de Buenos y 30 % de Malos, este proceso de balanceo se hizo por medio de la metodología de upsamplig, donde se aumentó los casos para la categoría de Malos (BM = 1). Despúes los valores de WOE previamente analizados se aplican a esta nueva base y se empieza con la fase de modelamiento.

6.3. Metodologías utilizadas para la construcción del modelo de RC (Mo-delo Logit – Modelo GAM

Se emplea a correr los dos modelos respectivamente, los betas mostrados a continuación no son los finales, ya que por medio de la validacion cruzada se realiza este cálculo:

- Modelo Logístico

```
> Mod = glm(B M ~ Experiencia BAC+
            Tiempo en la Actividad+Edad+
            Género+ Pasivo Patrimonio+
            Tipo de Vivienda,
            family = binomial, data = data)
```

Variable	β	$\mathbf{P}(>\|z\|)$
Intercepto	6.542E-01	2.00E-16
Experiencia BAC (S)	-7.862E-01	2.00E-16
Tiempo en la actividad	-1.516E-03	2.00E-16
Edad	-1.700E-02	2.00E-16
Género (M)	7.027E-01	2.00E-16
Pasivo / Patrimonio	1.341E-05	2.00E-16
Tipo de Vivienda (Asi)	2.735E-01	4.81E-04
Tipo de Vivienda (Fam+Arr)	2.636E-01	2.00E-16

Tabla 36: Estimación parámetros modelo logístico

En la regresión logística se observa que todas las variables son significativas (p-valores¡0.05), todas las variables tienen coherencia con el negocio, es decir su interpretación es adecuada. Para las variables edad y tiempo en la actividad a medida que aumenten la probabilidad de incumplimiento disminuye. En el caso de la variable Tipo Vivienda, es mas probable el incumpliento de una persona que cuenta con con una vivienda arrendada que una persona con vivienda propia. A medida que la razón pasivo/ patrimonio aumente la probabilidad de incumplimiento sera mayor, y si la persona cuenta con experiencia en el Banco es menos disminuye la probabilidad de incumplimiento

- Modelo GAM

```
> Mod = gam(B M ~ Experiencia BAC+ Género+
            Tipo de Vivienda+ s(Tiempo en la Actividad, k=3)
            +s(Edad, k=3)+ s(Pasivo Patrimonio, k=3)
            , family = binomial, data = data, method = "REML")
```

En la regresión GAM se observa que todas las variables son significativas (p-valores¡0.05), los betas parametricos tienen una coherencia con el negocio, un ejemplo de ellos es la variable Experiencia BAC, si

Variable	β	P(> |z|)
Paramétrica		
Intercepto	-2.087	2.00E-16
Experiencia BAC (S)	-0.725	2.00E-16
Género (M)	0.670	2.00E-16
Tipo de Vivienda (Asi)	0.273	2.81E-03
Tipo de Vivienda (Fam+Arr)	0.521	2.00E-16
No Paramétrica		
Edad	1.994	2.00E-16
Pasivo / Patrimonio	1.996	2.00E-16
Tiempo en la actividad	1.992	2.00E-16

Tabla 37: Estimación parámetros modelo GAM

la persona tiene Experiencia con el Banco disminuye la probabilidad de incumplimiento que unapersona que no cuente con experiencia en el Banco, en el caso de los betas no parametricos muestras una relacion directa con la probabilidad de incumplimiento.

6.4. Técnicas de selección del modelo

6.4.1. Aréa bajo la curva (AUC) y Curva ROC Base de entrenamiento Modelo Logístico

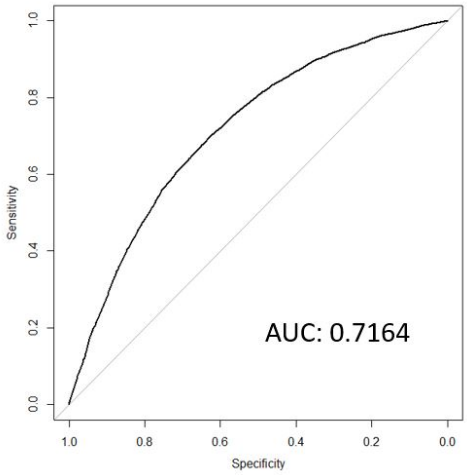

Tabla 38: AUC y Curva Roc de base de prueba -Modelo Logit

		Valores Predichos	
		Positivos	Negativos
Valores Reales	Positivos	143088	10928
	Negativos	52068	13939

Tabla 39: Matriz de Confusión de base de prueba - Modelo Logit

- Sensibilidad: 0.7332
- Especificidad: 0.5605
- Exactitud: 0.7137

El porcentaje de clasificación correcta del modelo logit en la base de entrenamiento es del 71.37 %, aunque en el grupo de score que son malos clientes solo clasifica bien el 56.05 %. Sin embargo, el modelo clasifica

bien al 73.32 % de los clientes buenos. El Criterio AUC da 0.7164, estos criterios dan un nivel aceptable para el modelo, solo falta realizar estas pruebas en la base de prueba que contiene el 30 % de los datos.

6.4.2. Aréa bajo la curva (AUC) y Curva ROC Base de entrenamiento Modelo GAM

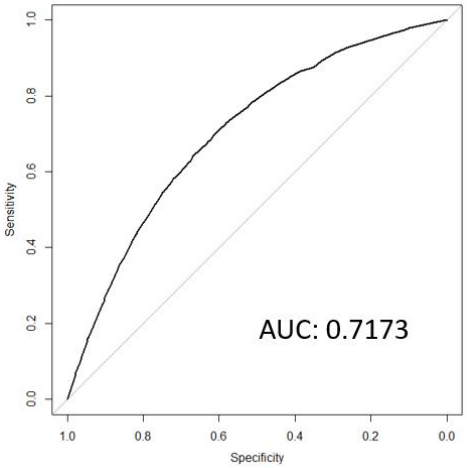

Tabla 40: AUC y Curva Roc de base de prueba -Modelo Gam

		Valores predichos	
		Positivo	Negativo
Valores Reales	Positivo	21527	44480
	Negativo	16075	137941

Tabla 41: Matriz de Confusión de base de entrenamiento -Modelo Gam

- Sensibilidad: 0.5725
- Especificidad: 0.7561
- Exactitud: 0.7248

El porcentaje de clasificación correcta del modelo GAM en la base de entrenamiento es del 72.48 %, en esta ocación el grupo de score que son malos clientes clasifica bien al 75.61 %, mejor que en el modelo logístico. Sin embargo, el modelo clasifica bien al 57.25 % de los clientes buenos. El Criterio AUC da 0.7173, al igual que en el modelo logit, ya que dan resultados muy parecidos.

Ahora se procede correr los dos modelos en las bases de prueba, y el que de mejor estimación sera escogido para implementarse en la Entidad Financiera.

6.4.3. Aréa bajo la curva (AUC) y Curva ROC Base de prueba (30 %) Modelo Logistico

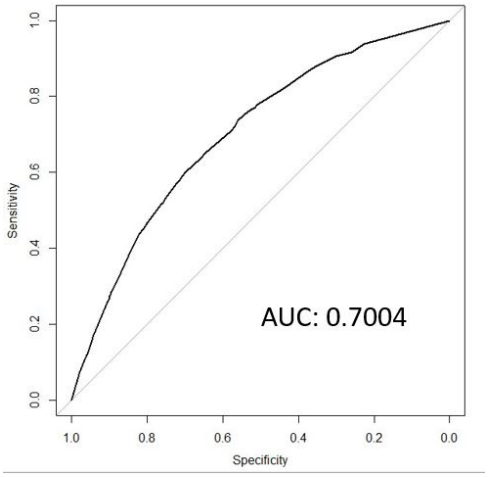

Tabla 42: AUC y Curva Roc de base de prueba-Modelo Logistico

		Valores predichos	
		Positivo	Negativo
Valores Reales	Positivos	7616	4968
	Negativos	20672	61037

Tabla 43: Matriz de Confusión de base de prueba - Modelo Gam

- Sensibilidad: 0.6052
- Especificidad: 0.7407
- Exactitud: 0.728

En esta ocasión el porcentaje de clasificación correcta del modelo Logit en la base de prueba es del 72.8 %, el grupo de score que son malos clientes clasifica bien al 74.07 %, mejor que en el modelo logístico en la base de prueba. Su criterio de AUC es de 0.70. Lo que concluye que es un modelo aceptable según estos criterios. Es de resaltar que el número de malos clasificados correctamente es mayor que el número de buenos clasificados correctamente, esto es un buen indicio, ya que la entidad busca es negar créditos a los que verdaderamente son morosos.

6.4.4. Aréa bajo la curva (AUC) y Curva ROC Base de prueba (30 %) Modelo GAM

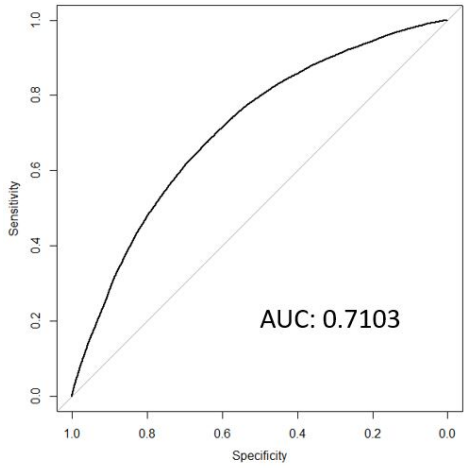

Tabla 44: AUC y Curva Roc de base de prueba-Modelo GAM

		Valores Predichos	
		Positvos	Negativos
Valores Reales	Positivos	4856	3464
	Negativos	23433	62542

Tabla 45: Matriz de Confusión de base de prueba - Modelo Gam

- Sensibilidad: 0.5836
- Especificidad: 0.7274
- Exactitud: 0.7147

El porcentaje de clasificación correcta del modelo GAM en la base de prueba es del 71.47 %, el grupo de score que son malos clientes clasifica bien al 72.74 %, da mejor el modelo GAM en la base de prueba. Su criterio de AUC es de 0.71. Lo que concluye que es un modelo aceptable según estos criterios. Al igual que en el modelo logístico, el número de malos clasificados correctamente es mayor que el número de buenos clasificados correctamente.

Si se compara los criterios de los dos modelos en la base de prueba (30 %) por crietrio de Especificidad y AUC es mejor el modelo GAM. Es de suma importancia comentar que la validación cruzada se hizo para los dos modelos, sin embargo, no es posible mostrar el resultado final de esta validación ya que el Banco utiliza este resultado para calcular un puntaje y por temas de privacidad no es posible publicar este resultado.

7. Conclusiones

En este proyecto de grado se presenta la comparación de un modelo logístico y un modelo aditivo generalizado, con el fin de seleccionar el de mejor predicción a l hora de estimar la probabilidad de incumplimiento de los clientes de microfinanzas de una entidad financiera colombiana, donde se observo que las variables edad, tipo de vivienda Experiencia con el Banco, tiempo en la actividad, género, y la razón pasivo/ patrimonio resultaron significativas en la estimación de los dos modelos y teniedo coherencia de negocio.

La utilización de técnicas estadísticas para el desarrollo y evaluación del riesgo de crédito permite analizar de forma estructurada y ordenada grandes volumenes de datos, todas las metodologías implementadas en este proyecto funcionan de forma correcta, y ayudan a responder a la necesidad de crear un modelo que calcule la probabilidad de incumplimiento para el segmento de microfinanzas de dicha entidad. Las técnicas utilizadas para el análisis exploratorio, la categorización y selección de variables fue útil, ya que las variables seleccionadas dieron significativas en los dos modelos y asi se se obtiene dos modelos para el RC con una gran capacidad predictiva respectivamente reflejados en los resultados del AUC, la matriz de confusión y la validación cruzada, los cuales ayudan a discriminar los clientes en buenos y malos de acuerdo con un puntaje obtenido por cada uno de las dos metodologías.

Se observó que el modelo GAM asume una distribución de valor extremo en la función link, presentaron un mejor desempeño predictivo respecto al modelo logit, teniendo mejor AUC tanto en la base de entrenamiento como en la base de prueba y permitiendo identificar mejor las personas con una probabilidad alta de incumplir sus obligaciones financieras. Sin embargo, dado que los resultados son muy parecidos tanto en la base de entrenamiento como en la base de prueba, los dos modelos son opcionales para implementarse en el Banco.

Referencias

Aedo, S., Pavlov, S., Clavero, F. (2010). Riesgo Relativo y Odds Ratio.

Aguas Gonzales , D. A., Castillo H, M. (2005). Modelo de administración del riesgo crediticio para la cartera comercial de una entidad financiera Colombiana. Bogotá.

Altman , E. (1968). Financial ratios, discriminant analysis and the prediction of corporate banckruptcy. The Journal of Finance.

Amaya G, C. A. (2005). Evaluación del riesgo de crédito en el sistema financiero colombiano. Reporte de estabilidad financiera, Banco de la República.

Cañete, I. (Septiembre de 2018). Scoring No Clientes (Modelización con información Big Data). Coruña.

Colombia, S. F. (2002). CARTA CIRCULAR 31 DE 2002.

Durbán , M. (2015). Modelos Aditivos Generalizados con P-splines. Madrid.

Fernandez Castaño , H., Perez Ramirez , F. O. (2005). El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito . Revista Ingenierías Universidad de Medellín, 55-75.

Gómez Gonzáles, J. E., Morales Acevedo, P., Pineda Garcia, F., Zamudio Gómez, N. (2007). An

alternative methodology for estimating credit quality transition matrices. Borradores de Economía Banco de la República.

Islam, S., Zhou, L., Li, F. (2009). Application of Artificial Intelligence (Artificial Neural Network) to Assess Credit Risk: A Predictive Model For Credit Card Score.

Ladino, I. C. (Marzo de 2014). COMPARACIÓN DE MODELOS DE RIESGO DE CREDITO: MODELOS LOGISTICOS Y REDES NEURONALES. COMPARACIÓN DE MODELOS DE RIESGO DE CREDITO: MODELOS LOGISTICOS Y REDES NEURONALES. Bogotá.

Leal Fica , A. L., Aranguiz Casanova, M. A., Gallegos Mardones , J. (2018). Análisis de riesgo crediticio, propuesta del modelo Credit Scoring. Revista Facultad de Ciencias Económicas: Investigación y Reflexión, vol. XXVI, núm. 1,, 181-207.

Lennox, C. (1999). Identifying Failing Companies: A Reevaluation of the Logit, Probit and DA Approaches. Jornal of Economics and Bussines, 347-364.

Martinez A, O. (2003). Determinantes de fragilidad en las empresas colombianas. Borradores de Economía, Banco de la República .

Medina, S., Paniagua, G. (2008). Modelo de Inferencia Difuso para estudio de Crédito. 215-229.

Moscote Flórez, O., Arley Rincón, W. (Diciembre de 2012). Modelo Logit y Probit: un caso de aplicación. pág. 2. Murray R., S., Larry J., S. (2009). Estadistica. En S. Murray R., S. Larry J..

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

Quintas, I. (2015). Modelo aditivo generalizado GAM: Regresión no lineal y no paramétrica.

Riascos Villegas , A. J., Benavides Gutiérrez , H. L. (2017). Modelo G con garantía hipotecaria.

Sior Alegre Norza, A. R. (2011). RELACIONES ONTOGÉNICAS Y ESPACIO-TEMPORALES EN LA DIETA DEL CALAMAR GIGANTE (Dosidicus gigas) EN PERÚ UTILIZANDO UN MODELO ADITIVO GENERALIZADO. Lima.

Zamudio Gómez, N. E. (2007). Determinantes de la Probabilidad de Incumplimiento de las Empresas Colombianas. Banco de la republica .