ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELO DE CLASIFICACIÓN BINARIA PARA UNA POBLACIÓN BANCARIZADA EMPLEANDO METODOLOGÍA DE ENSAMBLE

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

PROYECTO DE INVESTIGACIÓN

GISSELA FERNANDA VARGAS PACHECO
gis.selavargas95@gmail.com

DIRECTOR: MSc. Diego Paúl Huaraca Shagñay
diego.huaracas@epn.edu.ec

CODIRECTOR: MSc. Ménthor Oswaldo Urvina Mayorga
menthor.urvina@epn.edu.ec

Quito, Septiembre 2021

DECLARACIÓN

Yo GISSELA FERNANDA VARGAS PACHECO, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Gissela Fernanda Vargas Pacheco

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Gissela Fernanda Vargas Pacheco, bajo mi supervisión.

Diego Huaraca, MSc.
DIRECTOR DE PROYECTO

Ménthor Urvina, MSc..
CO-DIRECTOR DE PROYECTO

AGRADECIMIENTOS

A Dios por ser luz, guía y fortaleza en cada uno de los pasos de mi vida.

A mis padres Nelson y Luz Angélica por su amor, paciencia, cariño y arduo trabajo en busca de un mejor porvenir y a mi querida hermana Lisbeth por su compañía, cariño y complicidad en todo momento.

A Diego Huaraca y Ménthor Urvina por su confianza, apoyo y dirección en el desarrollo del presente trabajo.

DEDICATORIA

Nunca debes perder: tu forma de ser, tu esencia, la humildad de tu alma, la voz de tu corazón, el respeto a ti mismo y el valor por la vida.

Leo Pavoni

A mis padres por sus enseñanzas, esfuerzo continuo y ser mi pilar fundamental en la vida.
A mi hermana por su amor y amistad incondicional.
A mis familiares que partieron y hoy en día me acompañan de manera espiritual.

Índice

Índice de figuras

Índice de Tablas

Resumen

En el presente trabajo se realiza una investigación empírica del rendimiento de varios modelos de clasificación basados en la metodología de ensamble homogéneo y heterogéneo para la predicción del comportamiento de pago de las obligaciones crediticias adquiridas por los individuos en una entidad financiera; el propósito es reducir el coste de análisis crediticio manual, tomar decisiones más tempranas y disminuir el posible riesgo. La construcción de modelos más robustos y con mayor precisión en la predicción del incumplimiento se basa en la combinación de los clasificadores: Regresión Logística (RL), Random Forest (RF), Gradient Boosting Machine (GBM) y Naïve Bayes (NB). Adicionalmente, el rendimiento de los resultados del modelo obtenidos con la metodología de ensamble es superior que el rendimiento de los resultados obtenidos con los clasificadores individuales.

Palabras claves: método de ensamble, clasificación, calificación crediticia, probabilidad de incumplimiento, aprendizaje automático.

Abstract

In this project an empirical investigation of the performance of several classification models based on the methodology of homogeneous and heterogeneous ensemble for the prediction of the payment behavior of credit obligations acquired by individuals in a financial institution is carried out; the purpose is to reduce the cost of manual credit analysis, make earlier decisions and reduce the potential risk. The construction of more robust models with higher accuracy in the prediction of default is based on the combination of classifiers: Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM) and Naïve Bayes (NB). Additionally, the performance of the model results obtained with the ensemble methodology is superior to the performance of the results obtained with the individual classifiers.

Keywords: ensemble method, classification, credit scoring, probability of default, machine learning.

Capítulo 1

Introducción

1.1 Antecedentes

El asignar objetos a una de varias categorías definidas previamente se lo denomina clasificación, y el término minería de datos, según (Han et al., 2012) se lo expresa como el proceso de extraer, descubrir y analizar patrones ocultos a partir de los datos y por último obtener un resumen con información oportuna para la toma de decisiones (Suman et al., 2012).

Así pues, en minería de datos (DM) desde un conjunto de datos de entrada podemos construir un modelo de clasificación capaz de predecir una clase para una instancia dada a partir de la información preliminar aprendida de los datos. (Altman, 1968), (Beaver, 1966) y otros hasta la actualidad han venido desarrollando la investigación de los modelos de clasificación crediticia que fueron propuestos inicialmente por (Fisher, 1936).

(Zurada y Barker, 2007) señalan que las personas y empresas necesitan financiamiento, es decir, requieren acceder a un crédito para el buen funcionamiento entre el mercado y la sociedad. De modo que, las entidades financieras juegan un papel crucial en la economía actual y comprenden que la estrategia para una buena competitividad es prevenir la estafa, disminuir el riesgo crediticio y conservar a los buenos clientes. Una tarea difícil para las instituciones financieras es la predicción del comportamiento de pago del futuro prestatario, es decir, identificar si un cliente será moroso o no moroso. La predicción del incumplimiento de pago es un problema de clasificación binaria y uno de los métodos que utiliza una entidad para determinar si se otorga o no un servicio financiero es la calificación crediticia.

La literatura refleja el estudio de modelos contemporáneos de clasificación crediticia basados en métodos de minería que incluyen la regresión logística estudiado por (Henley,

1995); los modelos no paramétricos, árboles de decisión (Davis et al., 1992), (Zhou et al., 2008); Random Forest (Breiman, 2001); redes neuronales artificiales (ANN) (Jensen, 1992), (West et al., 2005); k-vecino más cercano (Henley, 1996); Naïve Bayes (Lewis, 1998); algoritmo genético (GA) (Desai et al., 1997), (Zhang et al., 2007); y máquinas de soporte vectorial (SVM) (Baesens et al., 2003), (Huang et al., 2007) que llegaron hacer métodos para estimar la probabilidad que un individuo no cumpla con sus obligaciones adquiridas, evaluar los préstamos y mejorar la precisión de la calificación crediticia, ya que permiten enriquecer los patrones encontrados en perfiles de clientes solventes como lo mencionan (Thomas et al., 2004). Sin embargo, las técnicas de aprendizaje más recientes tienen la capacidad de superar a las mencionadas; (Kittler et al., 1998) y (Kuncheva, 2014) señalan que un ensamble clasificador (también denominado sistema clasificador múltiple) consiste en un conjunto de clasificadores entrenados individualmente (clasificadores base) cuyas decisiones se combinan de alguna manera, generalmente mediante votación ponderada o no ponderada al clasificar nuevos vectores de características con el objetivo de predecir el comportamiento de los clientes.

Las técnicas estadísticas contemporáneas y recientes mencionadas asocian una puntuación de riesgo a los individuos que solicitan un producto financiero o a los clientes de la entidad financiera con el fin de calcular la probabilidad de pago de la obligación adquirida.

(Apampa, 2016), (Dahiya et al., 2015), (Patil et al., 2016), (Ala'raj y Abbod, 2016) y (Nanni y Lumini, 2009) son algunas de las investigaciones de modelos para la calificación crediticia realizadas en países extranjeros que emplean la metodología de ensamble siendo viables, confiables y con resultados favorables en las áreas de banca y finanzas, en consecuencia, se realiza un estudio empírico en el país con el propósito de aportar al conocimiento existente sobre las metodologías utilizadas para la clasificación de clientes de una cartera de clientes dentro de una entidad financiera, mediante el desarrollo y aplicación de herramientas basadas en teorías y técnicas precedentes.

1.2 Justificación

Las entidades financieras como por ejemplo los bancos comprenden que las pérdidas de oportunidades o pérdidas financieras se derivan de decisiones erróneas. Por lo tanto, el desarrollo de la investigación ayuda a resolver uno de los problemas que afrontan las entidades financieras que es el riesgo de crédito, es decir, la probabilidad de pérdida que asume la entidad como consecuencia del incumplimiento de las obligaciones de los prestatarios.

En virtud de que el sector bancario se caracteriza por recopilar grandes bases de datos con atributos particulares de cada solicitante como cliente, (Junqué de Fortuny et al., 2013) mencionó que pueden ser aprovechadas de manera eficiente ya sea para la realización de un análisis predictivo tradicional o para que los bancos privados obtengan una mayor ventaja de competitividad al adoptar técnicas de aprendizaje automático con el fin de mejorar la capacidad de diferenciar la calidad de los clientes internos y predecir el comportamiento de los solicitantes.

Además, (Abdou y Pointon, 2011) sostuvieron que la calidad de los servicios bancarios es el determinante clave para la supervivencia, competencia y rentabilidad de una institución bancaria; por esto, el desarrollo de modelos de clasificación eficientes contribuyen a que las entidades financieras aumenten su tolerancia al riesgo y ofrezcan productos a un segmento más amplio de clientes.

Así, las investigaciones realizadas durante los últimos años han prestado considerable atención en una nueva técnica atractiva y exitosa de aprendizaje automático denominado ensamble por su gran habilidad de reconocer patrones estadísticos que aplicados principalmente en las áreas de banca y de finanzas demuestran mayor poder de discriminación de clientes a comparación de los modelos tradicionales, pues se enfocan en reunir la visión de un grupo de clasificadores capacitados en el mismo problema para después combinar las decisiones de cada uno y finalmente llegar a una única clasificación efectiva y favorable para la institución financiera.

(Dietterich, 1997) demostró que la metodología de ensamble es una técnica de predicción de riesgo de crédito efectiva al tiempo que extiende la posibilidad de brindar un crédito a los solicitantes de préstamos, también, (Huang et al., 2004) identifican una diferencia significativa: el método de ensamble tiene la capacidad de aprender varias estructuras y patrones del modelo a partir de los mismos datos mientras que los métodos tradicionales dependen del investigador para utilizar una estructura particular, como la linealidad, mediante la estimación de parámetros que se ajusten a los datos.

Además, (Nanni y Lumini, 2009) examinaron el desempeño de varios modelos de análisis basados en métodos de ensamble para la calificación crediticia. Los resultados revelaron que las técnicas de ensamble se pueden utilizar para mejorar el rendimiento del clasificador autónomo.

Más aún la teoría presentada por (Ala'raj y Abbod, 2016) manifiestan que entre los modelos de sistemas de clasificadores múltiples desarrollados en la literatura, se ha dado

poca consideración a:

1. Combinar clasificadores de distintos algoritmos, ya que la mayoría se ha concentrado en construir clasificadores del mismo algoritmo.
2. Explorar diferentes técnicas de combinación para múltiples sistemas clasificadores de varios algoritmos de clasificación.

Por este motivo, se incentiva el estudio de las distintas técnicas de ensamble como bagging, boosting y stacking; relevantes en la construcción de la metodología.

Tomando en consideración lo expuesto, el propósito es descubrir que modelos individuales son adecuados para la estrategia de ensamble en el área de calificación crediticia, por lo cual se efectúan varios experimentos en el conjunto de datos de información operativa interna de tarjetas de crédito y operaciones del cliente que forman parte de los productos y servicios que brinda una institución financiera perteneciente al Sistema Crediticio Ecuatoriano.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un modelo de clasificación binaria basado en una metodología de ensamble a través de la combinación eficiente de modelos individuales que tenga mejor capacidad para identificar individuos que incumplan las obligaciones adquiridas en una entidad financiera.

1.3.2 Objetivos Específicos

- Identificar potenciales solicitantes de crédito en función al riesgo, protegiendo de esta manera la entidad financiera de tentativos clientes con impagos eventuales.
- Describir las variables que mejor explican las características de la población bancarizada.
- Encontrar las técnicas estadísticas de clasificación adecuadas para la estrategia de ensamble propuesta en modelos de clasificación de crédito personal.
- Realizar un análisis empírico del desempeño y calidad del método desarrollado.
- Comparar rendimientos de técnicas nuevas sobre modelos tradicionales de análisis de datos y recomendar modelos con mejor capacidad de identificación de clientes buenos y malos.

Capítulo 2

Marco Teórico

Este capítulo contempla los fundamentos teóricos de los conceptos, términos clave y conocimientos necesarios para identificar, comprender y expresar la metodología propuesta en la construcción del modelo de clasificación binaria de una cartera de clientes. En primer lugar, se incluye una descripción de estadísticos e índices basados en analizar la diferencia entre las distribuciones de probabilidad de dos variables aleatorias, se detallan las técnicas estadísticas tradicionales utilizadas en la construcción de modelos scoring y finalmente, para fundamentar la investigación del modelo de clasificación binaria a través de la metodología de ensamble se estudian las diversas técnicas de combinación como *bagging*, *boosting* y *stacking* para múltiples conjuntos de modelos predictivos.

En la práctica comparar y detectar la divergencia entre distribuciones es de utilidad en los siguientes escenarios:

1. Conocer si dos variables tienen similar distribución, incluso analizar posibles evidencias de que una misma variable se distribuya de forma distinta entre dos grupos;
2. Monitorizar el comportamiento de variables empleadas al construir modelos predictivos supervisados y no supervisados;
3. Identificar variables que generan distintos resultados a causa de su comportamiento en un proceso específico.

En nuestro estudio, se presentan múltiples métodos como estrategias de aproximación a fin de brindar respuesta al segundo caso expuesto, esto es, analizar el comportamiento de variables en un modelo de clasificación binaria.

2.1 Medidas de separación

2.1.1 Prueba Kolmogorov-Smirnov para dos muestras (*KS*)

Se define la función de distribución empírica acumulada con la intención de caracterizar la prueba *KS* para dos muestras.

Definición 1 *Sea la variable aleatoria X donde,* $x_1, x_2, \ldots, x_n$ *son observaciones de una muestra de tamaño n y* $F(x)$ *la función de distribución acumulada teórica subyacente de los datos. Definimos la distribución acumulada empírica de X como:*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x) \tag{2.1}$$

La prueba estadística *KS* cuantifica la distancia vertical máxima entre dos funciones de distribución acumulada empíricas de dos muestras aleatorias independientes disponibles, siendo la más conocida y comúnmente utilizada como prueba de bondad de ajuste no paramétrica dado que es sensible a diferencias de escala y forma de la función de distribución acumulada.

Ahora, consideremos las muestras: $x_1, \ldots, x_{n_1}$ de tamaño n_1 de la variable aleatoria continua X y $y_1, \ldots, y_{n_2}$ de tamaño n_2 de la variable aleatoria continua Y, F_1 y F_2 funciones de distribución acumulada teórica de las variables X e Y, respectivamente. Se expone la prueba Kolmogorov-Smirnov que tiene por objetivo responder si las dos muestras de datos provienen de la misma distribución continua hipotética, sin especificar cual es la distribución en común. La hipótesis a contrastar es la siguiente:

$$\begin{cases} H_0 : F_1(x) = F_2(x) \\ H_1 : F_1(x) \neq F_2(x) \end{cases} \tag{2.2}$$

En la ecuación (2.2) se contrasta la hipótesis nula H_0 que las dos muestras se derivan de una distribución en común y la hipótesis alternativa H_1 señala que las dos muestras no se derivan de la misma distribución teórica.

El siguiente paso es presentar el estadístico para contrastar la hipótesis nula H_0

$$KS = \max_x |\hat{F}_{n_1}(x) - \hat{F}_{n_2}(x)| \tag{2.3}$$

donde $\hat{F}_{n_1}(x)$ es el valor de la función de distribución empírica de X en la observación x y $\hat{F}_{n_2}(x)$ es el valor de la función de distribución empírica de Y en la observación x.

Ahora, si las muestras $x_1, \ldots, x_{n_1}$ y $y_1, \ldots, y_{n_2}$ proceden de la misma población el contraste es siempre de una cola, pues sus funciones de acumulación empíricas $\hat{F}_{n_1}$ y $\hat{F}_{n_2}$ no pueden ser tan diferentes. Finalmente, para un nivel de significancia α se rechaza la hipótesis nula H_0 de igual distribución si el valor del estadístico KS es superior a KS_α. Para ello, KS_α es un valor crítico para la prueba Kolmogorov-Smirnov de dos muestras que se obtiene de tabla de valores, ver (Massey Jr, 1951).

2.1.2 Prueba de Anderson Darling para dos muestras (AD)

La prueba AD para dos muestras tiene el mismo propósito que la prueba de Kolmogorov-Smirnov: contrastar la hipótesis (2.2) y presenta las siguientes ventajas:

1. Sensibilidad a la forma y escala de la distribución acumulada (Anderson y Darling, 1954).
2. Aplicabilidad a las muestras pequeñas (Pettitt, 1976).
3. Sensibilidad al comportamiento de las colas de las distribuciones.
4. En muestras grandes, identifica diferencias muy pequeñas.

Las pruebas de Anderson Darling y de Kolgomorov-Smirnov comparten las dos primeras ventajas, sin embargo, la prueba AD generalmente es más potente que la prueba KS dado que las dos últimas ventajas expuestas son exclusivas de la prueba AD.

(Pettitt, 1976) generaliza la prueba Anderson Darling para dos muestras presentada por (Darling, 1957) en la siguiente fórmula:

$$AD = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1+n_2} (N_i Z_{(n_1+n_2-n_1 i)})^2 \frac{1}{i Z_{(n_1+n2-i)}} \tag{2.4}$$

en la que $Z_{(n_1+n_2)}$ es la combinación de la muestra $X_{n_1} = x_1, \ldots, x_{n_1}$ y la muestra $Y_{n_2} = y_1, \ldots, y_{n_2}$, donde las observaciones de cada muestra son ordenadas de mayor a menor. N_i representa en número de observaciones en $X_{(n_1)}$ que son menores o iguales que la i-ésima observación de $Z_{(n_1+n_2)}$.

La hipótesis nula H_0 de la expresión (2.2) que $X_{(n_1)}$ y $Y_{(n_2)}$ procede de la misma distribución continua es rechazada si el estadístico AD es mayor que el valor crítico AD_α, para un nivel de significancia α. Para los valores críticos dependiendo del nivel de significancia α y el tamaño de la muestra ver la tabla en (Pettitt, 1976).

Por último, para probar la homogeneidad en múltiples muestras, (Scholz y Stephens, 1987) generalizaron aún más la prueba Anderson Darling para k muestras, sin embargo, esta versión no se aborda en este estudio.

2.2 Medidas de asociación

El objetivo de esta sección es mostrar varios conceptos estadísticos que tienen por fin testear el vínculo existente entre variables categóricas nominales, es decir, aquellas variables cuyas categorías no tienen un orden natural. En nuestro caso nos enfocamos en la distribución condicional de la variable respuesta y la forma de cambio que tiene a medida que varía la categoría de la variable explicativa, ya que su tratamiento es distinto a las variables cuantitativas. Para estudiar dicha asociación, el primer paso es exponer una tabla cruzada que resume la relación entre variables categóricas denominada tabla de contingencia (ver Tabla 2.1). Sean Y la variable dependiente binaria (Bueno/Malo) y una variable polinómica cualitativa arbitraria X con k categorías, cada observación debe pertenecer a una sola categoría, es decir, las k categorías deben ser exhaustivas y excluyentes entre sí.

	X						
Y	C_1	C_2	...	C_i	...	C_k	***Total***
Bueno	b_1	b_2	...	b_i	...	b_k	B
Malo	m_1	m_2	...	m_i	...	m_k	M
Total	n_1	n_2	...	n_i	...	n_k	n

Tabla 2.1: Tabla de contingencia $2 \times k$.

Para los individuos que se clasifican en una de las dos categorías de Y, b_i y m_i denotan las frecuencias observadas de Bueno y Malo, respectivamente, en una de las C_j categorías de X, $j = 1 \dots k$, además, B es el número total de individuos Buenos, M es el número total de sujetos etiquetados como Malos, $n_i = b_i + m_i$ y $n = B + M$.

2.2.1 Prueba de independencia Ji-Cuadrado (χ^2)

Para conocer la relación de dependencia entre dos variables cualitativas en muestras suficientemente grandes se utiliza la prueba de independencia estadística Ji-Cuadrado.
El objetivo en nuestro estudio es determinar si la variable respuesta Y dicotómica es independiente de una variable cualitativa X con $C_1, \dots, C_k$ categorías, entonces se contrastan las siguientes hipótesis:

$$\begin{cases} H_0 : X, Y \textit{ son independientes} \\ H_1 : X, Y \textit{ están relacionados} \end{cases} \tag{2.5}$$

Luego, el estadístico vinculado a la prueba es calculado a partir de la expresión (2.6), donde $\widehat{b_i}$ y $\widehat{m_i}$ son las frecuencias esperadas de los individuos de las etiquetas Bueno y Malo, respectivamente, en cada categoría $C_i, i = 1, \dots, k$, definidas en las expresiones

(2.7) y (2.8).

$$\chi^2 = \sum_{i=1}^{k} \frac{(b_i - \widehat{b_i})^2}{\widehat{b_i}} + \frac{(m_i - \widehat{m_i})^2}{\widehat{m_i}} \tag{2.6}$$

$$\widehat{b_i} = \frac{(b_i + m_i)B}{n} \tag{2.7}$$

$$\widehat{m_i} = \frac{(b_i + m_i)M}{n} \tag{2.8}$$

Al comparar el resultado del estadístico χ^2 y el valor crítico χ^2_α, para un nivel de significancia α y $(2-1) \times (k-1)$ grados de libertad, si χ^2 es mayor que χ^2_α entonces se rechaza la hipótesis nula H_0 a favor que la variable dependiente y variable explicativa están relacionadas.

Sin embargo, el estadístico χ^2 no se encuentra acotado superiormente, en consecuencia, es difícil de interpretar, puesto que se calcula mediante frecuencias absolutas y no relativas. Para evitar estos inconvenientes se emplea el **Coeficiente de contingencia de Pearson** definido desde la fórmula (2.9) y se deduce la desigualdad $0 \leq C < 1$, donde valores cercanos a uno indican relación entre variables y a medida que se acercan al valor de cero señalan independencia entre variables.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{2.9}$$

2.2.2 Valor de Información *(VI)*

En nuestro proyecto se aborda el concepto de valor de información, debido a que en la Sección 3 es utilizado para filtrar variables cualitativas en función de su importancia para la modelización de la probabilidad de incumplimiento de pago y de esta manera garantizar la construcción de un modelo scoring con mayor poder predictivo.

A partir de la Tabla (2.1), el VI se calcula empleando la siguiente expresión:

$$VI = \sum_{i=1}^{k} \left(\frac{m_i}{M} - \frac{b_i}{B} \right) ln \left(\frac{m_i/M}{b_i/B} \right) \tag{2.10}$$

mientras mayor sea el valor de VI, entonces existe una fuerte relación entre el predictor y la razón de probabilidades de Buenos/Malos, es decir, el nivel de predicción es fuerte.

A continuación se presenta una regla general de interpretación del valor resultante de la expresión (2.10):

- La fuerza de relación entre la variable explicativa y la variable dependiente binaria es alta si $VI > 0.3$.
- La fuerza de relación entre la variable explicativa y la variable dependiente binaria es media si $0.1 < VI \leq 0.3$.
- La fuerza de relación entre la variable explicativa y la variable dependiente binaria es baja si $0.02 \leq VI \leq 0.1$.
- La fuerza de relación entre la variable explicativa y la variable dependiente binaria es no predecible si $VI < 0.02$.

Adicionalmente, las siguientes consideraciones son importantes en el concepto del valor de información:

- El valor resultante de la expresión (2.10) incrementa cuando aumentan las categorías de la variable explicativa.
- En cada una de las categorías de una variable cualitativa con más de 20 categorías debemos prestar atención a la representatividad de sujetos buenos y sujetos malos, debido a que, el número de clientes con mora y clientes sin mora debe ser estricto mayor a cero.

2.2.3 Peso de la Evidencia *(WOE)*

Una de las medidas utilizadas para cuantificar la relación entre una variable explicativa categórica y la variable respuesta es el peso de la evidencia (*WOE* por sus siglas en inglés de *Weight of Evidence*).

En nuestro estudio, el *WOE* se emplea en la medición del poder predictivo para discriminar sujetos buenos y sujetos malos de cada una de las categorías de una variable cualitativa. En base a la Tabla (2.1) se procede a describir la forma de cálculo del peso de evidencia para una variable categórica en la siguiente fórmula:

$$WOE = ln\left(\frac{Db_i}{Dm_i}\right) \times 100 \tag{2.11}$$

donde Db_i es la estimación empírica de pertenecer a la categoría i de la variable cualitativa X condicionado a formar parte de la categoría Bueno de la variable dependiente Y y Dm_i es la estimación empírica de pertenecer a la categoría i de la variable cualitativa X condicionado a formar parte de la categoría Malo de la variable dependiente Y. Los términos Db_i y Dm_i se definen como la distribución de Bueno y distribución de Malo de la variable X, respectivamente.

Si el valor resultante de la fórmula (2.11) es negativo indica que la proporción de clientes

malos es superior al número de clientes buenos en la categoría i de la variable cualitativa X, además, la variable cualitativa tendrá mayor capacidad de predicción si la brecha de valores del WOE entre categorías es amplia.

Finalmente, el valor de información y peso de la evidencia son aplicados en el análisis predictivo del riesgo de crédito porque consideran la contribución de cada variable explicativa en el modelo scoring final, visualizan las correlaciones entre la variable respuesta y las variables explicativas y permiten comparar el poder de predicción de las variables explicativas cualitativas entre sí.

2.3 Técnicas estadísticas en la calificación crediticia (Credit Scoring)

La esencia del modelo de clasificación es comparar diversas características de un solicitante que requiere un servicio financiero con el perfil de clientes anteriores que adquirieron obligaciones con la institución financiera. Si los atributos del solicitante son semejantes a un cliente con la capacidad de cumplir el pago de sus obligaciones sin arriesgar la estabilidad financiera se define como solicitante solvente, entonces el requerimiento del servicio será otorgado de otra manera rechazado. Generalmente, las técnicas utilizadas en el país para este proceso son la evaluación del oficial de crédito y la calificación crediticia.

Básicamente, puede existir un modelo correcto para el propósito correcto y cualquier modelo de calificación crediticia cuando se desarrolla debe contemplar la definición del problema y forma de ajuste en función de la disponibilidad de datos como la capacidad computacional. Un modelo debe ser fácil y comprensible en su funcionamiento, adaptable a la variabilidad de condiciones a través del tiempo y también debe extenderse para ser rentable y detectar problemas (Alaraj et al., 2014).

(Ravi et al., 2015) señalan que la toma de decisiones para la evaluación del riesgo de crédito a asumir es objetiva, consistente y estandarizada al utilizar un modelo de calificación; también, (Abdou y Pointon, 2011) apuntan al Credit Scoring como uno de los modelos más exitosos utilizados en las áreas de banca, finanzas y negocios.

Generalmente, el proceso para diseñar una herramienta de calificación crediticia se describe en los siguientes pasos:

1. Encontrar el problema de estudio y definir la variable dependiente.
2. Recopilar y transformar los datos empleados en el desarrollo del modelo.
3. Construcción y desarrollo del modelo de calificación.

La Figura 2.1 presenta un esquema general del modelo de Credit Scoring.

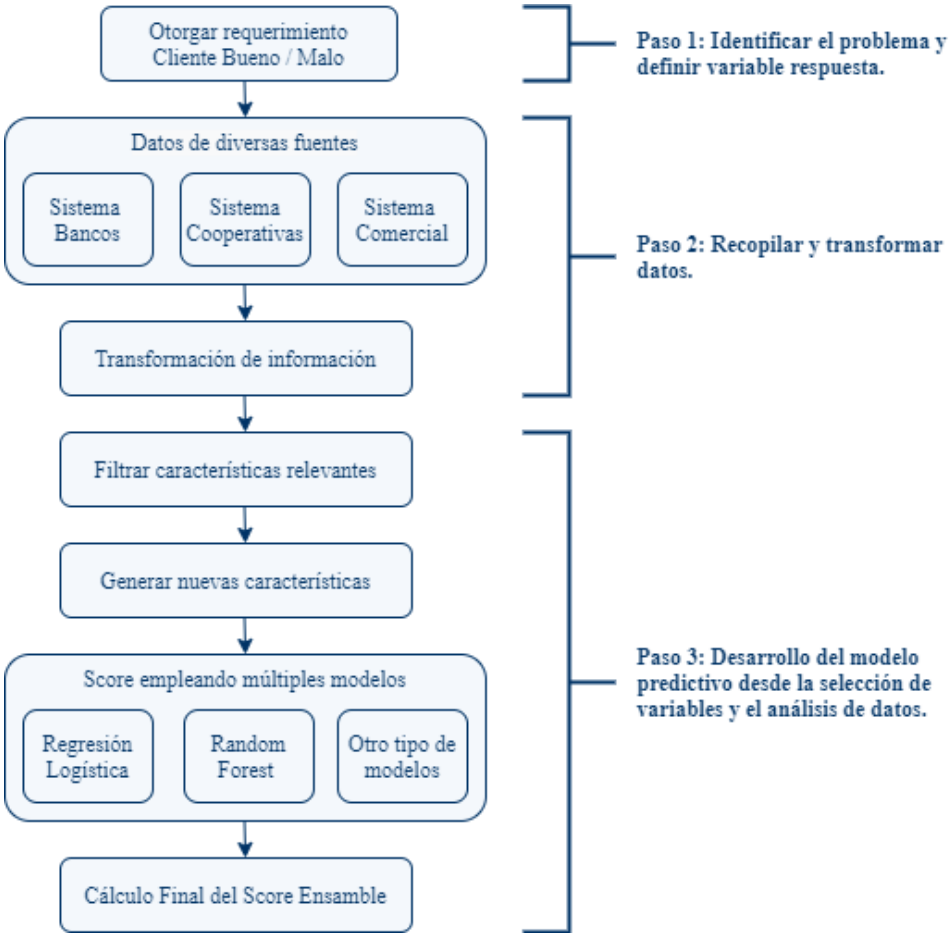


Figura 2.1: Ciclo para el desarrollo de un modelo Credit Scoring

Por lo tanto, es posible afirmar que un modelo Credit Scoring permite a las instituciones financieras evaluar la capacidad de cumplimiento del solicitante de forma oportuna y con menor costo. Además, la calificación crediticia brinda a las instituciones financieras la oportunidad de mejorar sus servicios y mantener a los buenos clientes.

Existe una variedad de técnicas estadísticas para la construcción de modelos Scoring, la mayoría de estos modelos estadísticos demuestran que son efectivos y eficientes en fines predictivos. Seguidamente, se detallan algunas de las técnicas más conocidas, como la regresión logística, Random Forest, Gradient Boosting Machine y Naïve Bayes que son utilizadas para la construcción de modelos de calificación crediticia por investigadores, analistas de crédito, prestamistas y desarrolladores de software.

2.3.1 Regresión Logística *(LR)*

La regresión logística es una de las técnicas estadísticas paramétricas más conocidas de aprendizaje automático y pertenece al conjunto de Modelos Lineales Generalizados (GLM, del inglés *Generalized Linear Model*). En minería de datos es adecuada para predecir una variable respuesta categórica, el caso más habitual es tener una variable dicotómica.

La regresión logística es la generalización de la regresión lineal, sin embargo, no se puede construir directamente desde la regresión lineal ya que el resultado de la misma es un valor discreto (resultado 0/1 en problemas de clasificación binaria). A pesar de ello, los métodos empleados en un análisis que utiliza regresión logística siguen los mismos principios generales que se utilizan en la regresión lineal (Hosmer et al., 1989).

Una de las principales aplicaciones de la regresión logística es la calificación crediticia, debido a que permite clasificar los solicitantes y clientes de la institución financiera dentro de las categorías Bueno o Malo de la variable dependiente en base a las variables explicativas cuantitativas continuas como discretas o variables cualitativas que caracterizan el comportamiento de pago de los individuos.

Para una cartera de clientes de tamaño n definimos el valor Y_i de la variable dependiente binomial Y como sigue:

$$Y_i = \begin{cases} 0 & \text{si el individuo } i \text{ es etiquetado como Bueno} \\ 1 & \text{si el individuo } i \text{ es etiquetado como Malo} \end{cases} \tag{2.12}$$

De la expresión (2.12) se tiene que la probabilidad de malo asociado al sujeto i se denota por $\pi_i = \mathbb{P}(Y_i = 1)$ y $1 - \pi_i = \mathbb{P}(Y_i = 0)$ en las etiquetas de malo y bueno de la variable respuesta, respectivamente. El propósito es relacionar funcionalmente π_i y un vector de variables explicativas $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})$ que describen el comportamiento de pago del i-ésimo sujeto, es decir:

$$f(\pi_i) = \beta_0 + \beta_1 Xi1 + \cdots + \beta_1 Xik \tag{2.13}$$

donde $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son coeficientes de las variables explicativas que deben ser estimadas por máxima verosimilitud.

Generalmente, la función f de la ecuación (2.13) es conocida como función de enlace. Si la variable dependiente es dicotómica usualmente se utiliza la función logística como

función de enlace y se obtiene el conocido modelo logit:

$$f(\pi_i) = logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi}\right) = \beta_0 + \beta_1 Xi1 + \cdots + \beta_1 Xik \tag{2.14}$$

La razón $ln\left(\frac{\pi_i}{1-\pi_i}\right)$ toma valores en el intervalo $(-\infty, \infty)$, se denomina razón de probabilidades y se interpreta en términos de riesgo.

Adicionalmente, la definición de la función inversa de f está dada por:

$$f^{-1}(t) = \frac{exp(t)}{1 + exp(t)} \tag{2.15}$$

Si utilizamos la fórmula (2.15) se tiene que:

$$\pi_i = \frac{exp(\omega)}{1 + exp(\omega)} \tag{2.16}$$

$$1 - \pi_i = \frac{1}{1 + exp(\omega)} \tag{2.17}$$

donde $\omega = \beta_0 + \beta' X_i$. En la ecuación 2.16 si ω es negativo, π_i es un valor con tendencia a 0 y si ω es positivo, π_i es un valor con tendencia a 1.

Finalmente, en base a la literatura, la regresión logística es un método apropiado para problemas de clasificación como predicción del incumplimiento de pago mencionado por (Hand y Henley, 1997). Hasta ahora, la regresión logística se usa ampliamente en aplicaciones de calificación crediticia (por ejemplo, (Abdou y Pointon, 2011); (Lessmann et al., 2015); (Crook et al., 2007)).

2.3.2 Random Forest *(RF)*

Definición 2 *Random Forest es un clasificador que consiste en una colección de clasificadores estructurada de árboles, cada árbol se construye con respecto a un vector aleatorio Θ_k, donde $\Theta_k, k = 1, \ldots, L$ son independientes e idénticamente distribuidos. Cada árbol emite un voto unitario para la clase más popular en la entrada $\boldsymbol{x}$.*

La precedente definición del algoritmo RF fue propuesta por (Breiman, 2001), su idea fue utilizar árboles de decisión de bajo sesgo y alta varianza como clasificadores base para construir el grupo de modelos predictivos, por lo cual, la combinación forma un modelo ensamble de menor varianza. El árbol de decisión (DT sigla en inglés *decision trees*) es un clasificador no paramétrico, es decir, no necesita que se cumpla una distribución en específico, y se divide en dos tipos: árbol de regresión que pronostica variables de respues-

ta cuantitativa y el árbol de clasificación que predice variables de respuesta cualitativa.

En la construcción del modelo Random Forest se puede realizar cualquier combinación de estas fuentes de diversidad:

- Muestras del conjunto de características (*feature sampling*).
- Muestras del conjunto de datos (*tree bagging*).
- Modificar algunos de los parámetros del árbol de clasificación.

En los experimentos realizados por (Breiman, 2001) se relacionan las dos primeras fuentes de diversidad para el desarrollo del modelo Random Forest conforme la siguiente secuencia: se extraen múltiples submuestras con reemplazo del conjunto de entrenamiento original, seguido, en cada una de las submuestras de entrenamiento se crea un árbol mediante la selección aleatoria de características.

En la Figura 2.2 se presenta el clásico modelo RF que se construye en base a la selección aleatoria de un subconjunto de variables para cada nodo del árbol con el fin de obtener resultados menos correlacionados entre sí, y el número de muestras con reemplazo empleadas para entrenar cada árbol creado. Generalmente, en la práctica el modelo Random Forest es fácil de utilizar, pues cuenta con dos parámetros que no son tan sensibles a sus valores: el número de árboles y la cantidad de variables en el subconjunto aleatorio en cada nodo. En la Tabla 2.2 se observa el entrenamiento y operación del modelo para problemas de clasificación y regresión:

RANDOM FOREST

Entrada:
Colección de muestras con reemplazo desde el conjunto de datos $T = T_1, T_2, \ldots, T_N$
Subconjunto de variables desde $P = P_1, P_2, \ldots, P_k$

Operación: para cada nuevo vector de características x
1. Construir un árbol de decisión sin podar empleando un subconjunto de variables en cada muestra.
2. Clasificar x en cada árbol de decisión $D_1, D_2, \ldots, D_N$
3. Asignar a x la clase con el mayor número de votos, denominamos "voto" a la etiqueta que asigna cada clasificador D_i

Salida:
Retornar la etiqueta del ensamble del nuevo objeto

Tabla 2.2: Modelo Random Forest

Finalmente, (Liaw y Wiener, 2001) sugieren las siguientes observaciones para el uso prác-

tico del modelo Random Forest:

- Para un idóneo rendimiento del RF, el número de árboles necesarios aumenta con el número de predictores. Un procedimiento óptimo a fin de determinar cuántos árboles son necesarios es comparar las predicciones realizadas por un Forest frente a las predicciones efectuadas en un subconjunto del Forest.
- Testar el número de variables seleccionadas m en cada nodo del árbol. Aconseja los valores siguientes: $m \approx \sqrt{p}$, $m \approx \frac{\sqrt{p}}{2}$, $m \approx 2 \times \sqrt{p}$, donde p corresponde el número total de predictores.
- En los problemas de clasificación que abarcan clases sumamente desequilibradas, es conveniente cambiar la regla de predicción a otra que no sea el voto mayoritario.

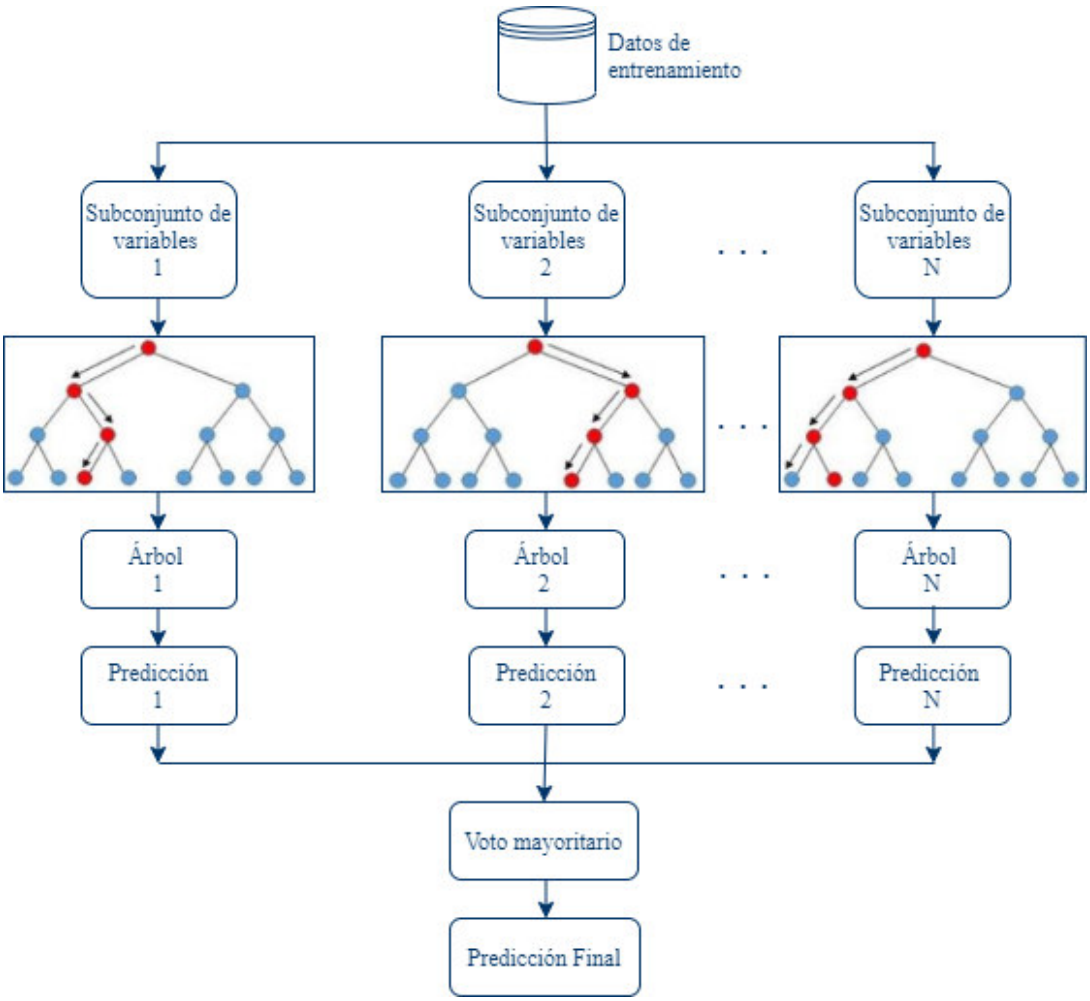


Figura 2.2: Esquema Random Forest

2.3.3 Gradient Boosting Machine (*GBM*)

Este apartado presenta esencialmente la derivación de boosting descrita en la subsección 2.4.2. Utilizando el trabajo de (Friedman, 2001) y (Friedman, 2002) procedemos a describir el algoritmo de clasificación que tiene la idea intuitiva de conectar el método boosting y la optimización de la función de pérdida durante el ajuste del modelo.

Se introduce el algoritmo gradient boosting como la combinación secuencial de múltiples modelos débiles (weak learners) para obtener un modelo predictivo fuerte de clasificación o regresión. En esta combinación, cada nuevo modelo usualmente basado en árboles trata de corregir los errores de los modelos anteriores. Ahora, en la Tabla (2.3) se observa el algoritmo propuesto por Friedman cuyo razonamiento evita problemas de error de generalización, ya que en cada iteración determina la dirección y el gradiente en el que se necesita mejorar el ajuste en el conjunto de datos como la selección de un modelo de un grupo de funciones que utilizan información covariable (generalmente un árbol de regresión) acorde a su dirección.

GRADIENT BOOSTING MACHINE

Entrada:
Conjunto de datos $D = (x_1, y_1), (x_2, y_2), (x_m, y_m)$
Función de pérdida $\Psi(y, \rho)$
Función de regresión $\hat{f}(x) = \arg\min_{\rho} \sum_{i=1}^{N} \Psi(y_i, \rho)$
Número de modelos base T

Operación: para cada nuevo vector de características x
Para $t = 1, \ldots, T$:
1. Calcular el gradiente negativo
   $z_i = -\frac{\partial}{\partial f(x_i)} \Psi(y_i, f(x_i))|_{f(x_i)=\hat{f}(x_i)}$
2. Predecir z_i desde las covarianzas de x_i
3. Ajustar el modelo de regresión $g(x)$
4. Elegir el tamaño de paso óptimo de descenso del gradiente
   $\rho = \arg\min_{x} \sum_{i=1}^{N} \Psi(y_i, \hat{f}(x_i) + \rho g(x_i))$
5. Actualizar la estimación de $f(x)$ agregando el nuevo weak learner
   $\hat{f}(x) \leftarrow \hat{f}(x) + \rho g(x)$

Salida:
Retornar la etiqueta del ensamble del nuevo objeto

Tabla 2.3: Procedimiento general del algoritmo GBM

En la Tabla (2.3) $\hat{f}(x)$ es una función de regresión que intenta minimizar la esperanza de una función de pérdida $\Psi(y, f)$. La predicción final del algoritmo GBM es la agregación de los valores predichos de cada modelo integrante del conjunto. En general, la agregación para problemas de clasificación es la moda y la media en problemas de regresión.

En resumen, GBM aprende relaciones no lineales entre la variable respuesta y variables explicativas, así como aplica el método de descenso del gradiente en la función de pérdida para ajustar el modelo.

2.3.4 Clasificador Naïve Bayes *(NB)*

Entre los algoritmos de aprendizaje automático supervisados se encuentra el clasificador de Naïve Bayes, mismo que se fundamenta en el teorema de probabilidades condicionales o también conocido como el Teorema de Bayes.

Dado que una instancia no es más que un vector de características; un enfoque para clasificar la instancia X es formular un modelo que determine la probabilidad posterior $\mathbb{P}(y|X)$ de distintos y, luego predecir el de mayor probabilidad posterior (regla de máximo a posterior MAP). Por el teorema de Bayes se describe la probabilidad posterior como:

$$\mathbb{P}(y|X) = \frac{\mathbb{P}(X|y)\mathbb{P}(y)}{\mathbb{P}(X)} \tag{2.18}$$

Donde $\mathbb{P}(y)$ de la expresión (2.18) se estima desde el conjunto de entrenamiento por medio del conteo de la proporción de la clase y en Y. La $\mathbb{P}(X)$ generalmente se omite, pues se utiliza la misma instancia X para comparar distintas y.

Ahora, para obtener un buen clasificador a partir de los datos de entrenamiento con una tasa de error pequeña se debe estimar $\mathbb{P}(X|y)$ de la fórmula 2.18 con la siguiente suposición: en el problema de clasificación una vez que tenemos las etiquetas de clase (bueno/malo), en cada clase cada variable debe ser independiente una de otra.

$$\mathbb{P}(X|y) = \prod_{i=1}^{n} \mathbb{P}(x_i|y) \tag{2.19}$$

La expresión (2.19) señala que sin necesidad de calcular las probabilidades conjuntas,

podemos estimar la probabilidad condicional desde el cálculo del valor de cada característica de la instancia x en cada clase.

Por último, para la etapa de validación, dada una nueva instancia X con etiqueta de clase y, el clasificador Naïve Bayes predice si conduce al mayor valor de la expresión (2.20) entre todas las etiquetas de clase.

$$\mathbb{P}(y|X) \propto \mathbb{P}(y) \prod_{i=1}^{n} \mathbb{P}(x_i|y) \tag{2.20}$$

2.4 Modelos de Ensamble

La finalidad de la sección es presentar varias nociones de la metodología ensamble, mostrar las claves para comprender como se utilizarán los métodos para diseñar una solución a nuestro problema de clasificación de clientes en una cartera crediticia.

El método de ensamble es parte de los algoritmos de aprendizaje supervisado que consiste en combinar múltiples modelos clasificadores (denominados clasificadores base) generados mediante algoritmos de aprendizaje que se entrenan de forma individual, comúnmente en un conjunto de datos de entrenamiento y resuelven un problema específico; luego se combinan los resultados mediante uno o múltiples métodos en un clasificador para obtener un único resultado con mejor rendimiento predictivo.

A partir de los datos originales, se generan conjuntos diversificados de datos en los cuales el ensamble clasificador aprende y se entrena, así, conduciendo a una mejor precisión en la predicción. La Figura 2.3 representa la estructura común de un ensamble.

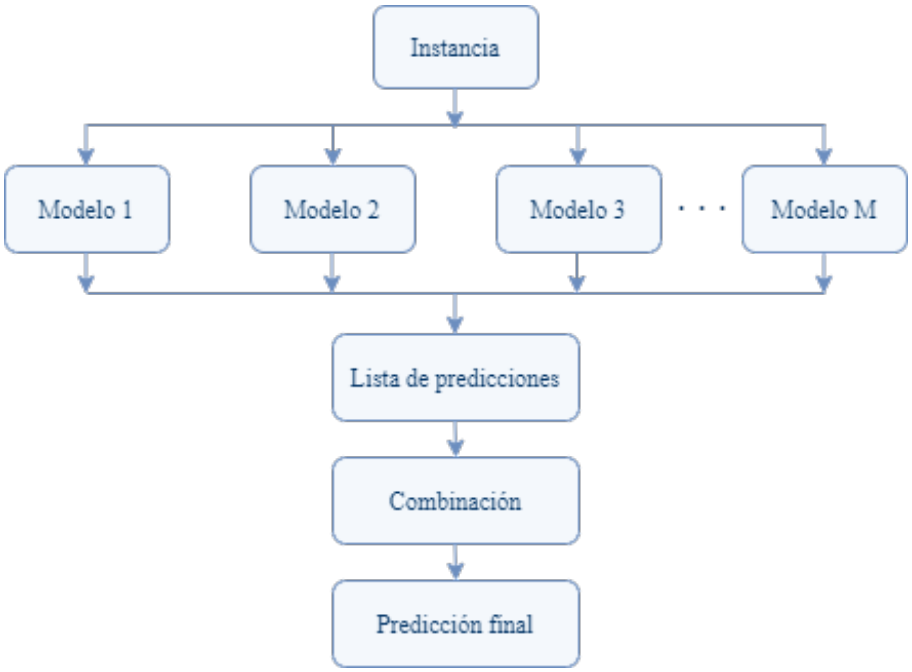


Figura 2.3: Arquitectura de un ensamble

(Lessmanna et al., 2013) expresan que la construcción de modelos clasificadores de ensamble se puede lograr de dos maneras:

- **Ensambles homogéneos:** utilizan un único algoritmo de aprendizaje base para generar modelos base del mismo tipo, esto es, modelos homogéneos. Finalmente, combina las predicciones en un solo clasificador.
- **Ensambles heterogéneos:** utilizan diferentes algoritmos de aprendizaje base para generar modelos individuales de distinto tipo, esto es, modelos heterogéneos. Por último, se combinan las predicciones en un clasificador exclusivo.

Los modelos de ensamble (homogéneos o heterogéneos) tiene mejor capacidad de generalización de un problema debido a que estimulan a los modelos base para que puedan hacer predicciones más precisas. Además, construir un buen ensamble clasificador requiere de modelos base diversos y lo más precisos posible.

En el trabajo de (Goyal y Kaur, 2016) se señalan que los principales beneficios de utilizar modelos de ensamble son:

- Mejor pronóstico.
- Modelo con mayor robustez.
- Resultados más óptimos.
- Reducción del error de predicción.

La elección de clasificadores base y la manera de formar el conjunto son clave para realizar un correcto modelo de ensamble, de manera que, el mínimo sesgo y varianza sean características esenciales del modelo. Si los clasificadores seleccionados son de sesgo bajo y varianza alta, cada clasificador es agregado al conjunto con un método que tienda a reducir la varianza, mientras que, si los modelos base corresponden a clasificadores de varianza baja y sesgo alto, se incorpora al ensamble con un método que reduce el sesgo.

En nuestro estudio, se construirá un modelo que combina ensambles clasificadores homogéneos y heterogéneos basados en cuatro técnicas de clasificación utilizadas principalmente en la literatura como son: Regresión Logística, Random Forest, Gradient Boosting y Clasificador Naïve Bayes. A continuación, se describen tres métodos de aprendizaje supervisado: bagging, boosting y stacking para la construcción de ensambles aplicados a los datos con el fin de reconocer patrones estadísticos.

2.4.1 Bagging

En primer lugar, se presenta la definición de muestra bootstrap que será de utilidad para describir los procedimientos de los distintos enfoques de ensamble.

Definición 3 *Sea un conjunto de observaciones de tamaño N, una muestra bootstrap consiste en una muestra con reemplazo de tamaño B, a partir de la selección aleatoria de observaciones del conjunto original.*

Las hipótesis para que se verifiquen las dos principales propiedades estadísticas de las muestras bootstrap (aproximadamente representatividad e independencia) de la verdadera distribución de datos respectivamente son:

- El tamaño del conjunto de observaciones debe ser lo suficientemente grande para capturar la complejidad de la distribución subyacente.
- El tamaño del conjunto de observaciones debe ser suficientemente grande en contraste al tamaño de cada muestra con el fin de evitar correlación de las muestras bootstrap.

Utilizando el trabajo (Breiman, 1996), procedemos a describir el método de ensamble bagging. El término *bagging* es acrónimo para ***B****ootstrap* ***AGG****regat****ING***. El método es simple, atractivo y con un buen rendimiento: el ensamble se constituye de clasificadores homogéneos construidos sobre muestras bootstrap del conjunto de entrenamiento, es decir, cada clasificador aprende al mismo tiempo, uno independiente del otro y se combinan las respuestas por el voto de pluralidad (también conocido en la literatura como el voto mayoritario) para obtener un modelo de baja varianza y mejor precisión.

En la práctica, se dispone de un único conjunto de entrenamiento, $T = T_1, T_2, \ldots, T_N$, por lo cual, se simula la generación aleatoria de L conjuntos de entrenamiento de longitud N empleando un procedimiento bootstrap (Egmont-Petersen et al., 1999).

Adicionalmente, para obtener una colección diversa de clasificadores, pequeños cambios en el conjunto de entrenamiento deberán conducir a grandes cambios en el resultado y rendimiento del clasificador, pues se intenta reducir el error causado por la variación de cada clasificador base.

La Tabla 2.4 muestra el entrenamiento y operación del bagging, mientras que la arquitectura del ensamble se observa en la Figura 2.4.

2.4.2 Boosting

Boosting es un algoritmo supervisado secuencial de tipo ensamble, es decir, se construye a partir de un conjunto de clasificadores débiles homogéneos (débilmente correlacionado con la clasificación correcta) con distinto peso en función de la exactitud de sus predictores para obtener un clasificador robusto.

ENSAMBLE BAGGING

Entrada:
Conjunto de datos $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
Algoritmo de aprendizaje base ξ
Número de modelos base T
Distribución bootstrap $\mathcal{D}_{bs}$

Operación: para cada nuevo vector de características x
1. Para $t = 1, \ldots, T$:
2. $h_t = \xi(D, \mathcal{D}_{bs})$
3. Fin

Salida:
Retornar la etiqueta del ensamble del nuevo objeto

Tabla 2.4: Modelo Bagging

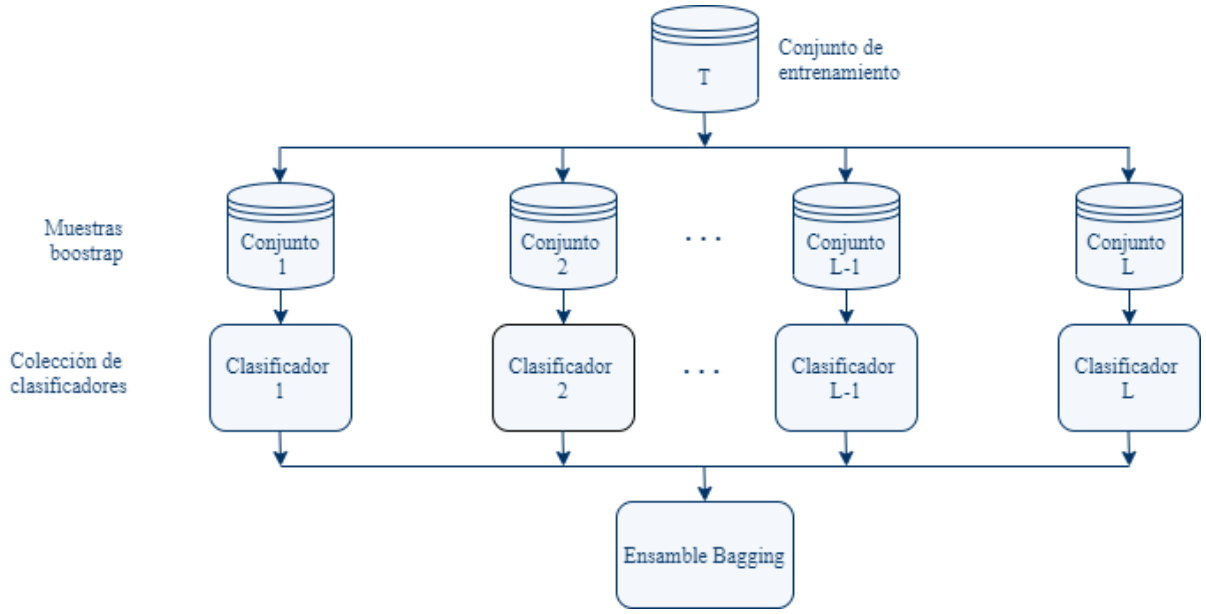


Figura 2.4: Ensamble Bagging

Los clasificadores que pertenecen al conjunto aprenden de forma secuencial, esto es, cada nuevo modelo enfoca sus esfuerzos en entrenar los datos más difíciles de manejar procedentes de los resultados de modelos ajustados en iteraciones previas del algoritmo, finalmente combina los resultados de manera determinística para generar un modelo de ensamble con menor sesgo a los clasificadores base que lo componen. Uno de los clasificadores más conocidos de la familia de métodos de boosting es el algoritmo *Adaptive Boosting* (también conocido como Ada-Boost) propuesto por (Freund et al., 1996).

Formalmente, la Tabla 2.5 describe el procedimiento general del ensamble *boosting*, para ello, se proporciona un conjunto de entrenamiento etiquetado $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ donde y_i es la etiqueta asociada a la instancia x_i. En nuestro estudio, x_i son los datos observables de un solicitante para un servicio financiero, mientras que y_i es la etiqueta binaria de cliente moroso o cliente no moroso.

ENSAMBLE BOOSTING

Entrada:
Distribución de la muestra $\mathcal{D}$
Algoritmo de aprendizaje base ξ
Número de ejecuciones del modelo T

Operación: para cada nuevo vector de características x
1. Inicializar la distribución: $\mathcal{D}_1 = \mathcal{D}$
2. Para $t = 1, \ldots, T$:
3. Entrenar un clasificador débil para $\mathcal{D}$
   $h_t = \xi(\mathcal{D}_t)$
4. Evaluar el error de h_t
   $\epsilon_t = P_{x \sim \mathcal{D}_t}(h_t(x) \neq f(x))$
5. Distribución ajustada $\mathcal{D}_{t+1}$
   $\mathcal{D}_{t+1} = dist.ajustada(\mathcal{D}_t, \epsilon_t)$

Salida:
Retornar la etiqueta del ensamble del nuevo objeto
$H(x) =$ combinación$(\{h_1(x), \ldots, h_t(x)\})$

Tabla 2.5: Proceso general Boosting

2.4.3 Stacking

El procedimiento de ensamble *stacking* en problemas de clasificación hace referencia a los modelos individuales como clasificadores de primer nivel y el algoritmo que se utiliza para su combinación es conocido como modelo de segundo nivel o metaclasificador. Stacking crea clasificadores de primer nivel desde un grupo de modelos que se entrenan en paralelo, uno independiente del otro utilizando muestras bootstrap en el conjunto de datos de entrenamiento, por ejemplo, bagging. Se entrena un segundo clasificador a partir de los resultados de los clasificadores de nivel uno para obtener una sola predicción. (Wolpert, 1992)

El punto clave del método de ensamble subyacente es saber si los datos se entrenan o

no correctamente, para que posterior el metaclasificador pueda corregir las entradas de entrenamiento inadecuadas y reducir el error; así, desarrollar un modelo fuerte menos sesgado que los clasificadores individuales.

Para la construcción del ensamble se deben definir L clasificadores individuales y el metaclasificador, a continuación, se muestran los pasos a seguir:

- Dividir los datos en dos subconjuntos.
- En el primer subconjunto ajustar cada clasificador L.
- El segundo subconjunto emplea la predicción realizada de cada clasificador para un mismo problema de clasificación.
- En el segundo subconjunto el metaclasificador realiza la predicción final en base a las predicciones de los clasificadores base.

Para que se entrene el metamodelo en todas las observaciones se utiliza un enfoque de "entrenamiento cruzado de k-folds". Así, se entrena $k-1$ veces y posterior se hacen predicciones en el subconjunto de datos restante y eso de forma iterativa para obtener predicciones de observaciones en cualquier subconjunto, finalmente el metamodelo se capacita en todas estas predicciones.

Para evitar un sobreajuste, se recomienda utilizar un procedimiento de validación cruzada o exclusión, de tal manera se genera el nuevo conjunto de datos con las instancias excluidas en el entrenamiento de los modelos de primer nivel. En nuestro estudio utilizaremos validación cruzada para evitar sobreajuste.

El procedimiento del *stacking* comienza dividiendo el conjunto de entrenamiento original D en k partes casi iguales $D_1, \ldots, D_k$. Para cada j-ésimo fold se define el conjunto de validación D_j y $D_{(-j)} = D \setminus D_j$ como el conjunto de entrenamiento. Dados T algoritmos de aprendizaje, el modelo de primer nivel $h_t^{(-j)}$ se genera en función del t-ésimo algoritmo de aprendizaje de $D_{(-j)}$. Para cada instancia x_i en el conjunto de validación D_j del j-ésimo fold, se define z_{it} como el resultado del modelo $h_t^{(-j)}$ de x_i. Luego del proceso de validación cruzada, se genera el nuevo conjunto de datos a partir de los T modelos de primer nivel como:

$$D' = \{(z_{i1}, z_{i2}, \ldots, z_{iT}, y_i)\}_{i=1}^{m} \tag{2.21}$$

En el conjunto de la expresión (2.21) se aplica el metaclasificador y su modelo resultante h' es una función de $(z_1, \ldots, z_T)$ para la etiqueta y de la instancia dada. Después de generar el nuevo conjunto de datos, en general, los modelos finales de primer nivel se vuelven a generar mediante el entrenamiento en $D_{(-j)}$.

La Tabla 2.6 describe el procedimiento general de un ensamble tipo *stacking* para un modelo de clasificación.

ENSAMBLE STACKING

Entrada:
Conjunto de datos $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
Algoritmo de aprendizaje de primer nivel $\xi_1, \xi_2, \ldots, \xi_T$
Algoritmo de aprendizaje de segundo nivel ξ

Operación: para cada nuevo vector de características x
1. Entrenar el modelo individual aplicando el algoritmo de primer nivel ξ_t
   Para $t = 1, \ldots, T$:
   $h_t = \xi_t(D)$
   Fin
2. Generar un nuevo conjunto de datos $D' = \emptyset$
   Para $i = 1, \ldots, m$:
   Para $t = 1, \ldots, T$:
   $z_{it} = h_t(x_i)$
   Fin
   $D' = D' \cup \{(z_{i1}, z_{i2}, \ldots, z_{iT}), y_i\}$
Fin
3. Entrenar el metaclasificador aplicando el algoritmo de segundo nivel
   $h' = \xi(D')$

Salida:
Retornar la etiqueta del ensamble del nuevo objeto
$H(x) = (h'(h_1(x), h_2(x), \ldots, h_T(x))$

Tabla 2.6: Proceso general del ensamble stacking

Finalmente, *bagging* y *boosting* son métodos habituales para crear ensambles. El enfoque de ensambles clasificadores realizado por la literatura relacionada se centra usualmente en conjuntos homogéneos, mientras que los conjuntos heterogéneos pocas veces se consideran. De cada diez estudios se mencionaron siete orientados en la construcción de conjuntos clasificadores homogéneos, cuatro se enfocan en clasificadores heterogéneos, y un estudio se centró en ambos (Ala'raj y Abbod, 2015).

Por esta razón, el presente proyecto propone un modelo que integre conjuntos homogéneos y heterogéneos utilizando varios métodos de ensamble y combinando estrategias para aprovechar al máximo los diversos enfoques de diferentes clasificadores sobre el conjunto de datos.

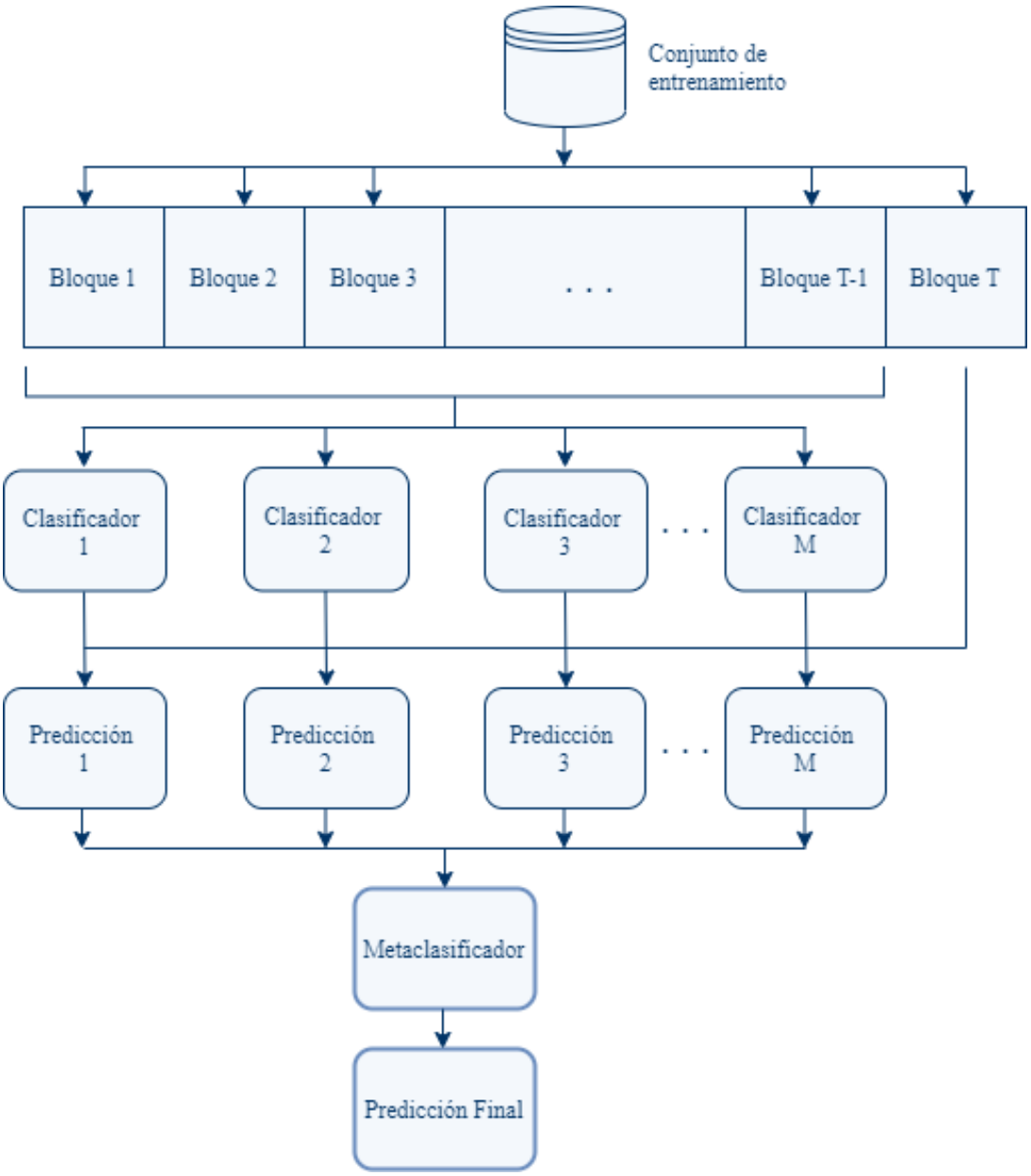


Figura 2.5: Arquitectura del ensamble stacking

Capítulo 3

Marco Metodológico

El propósito principal de este capítulo es describir la metodología utilizada para la construcción del modelo de clasificación binaria basado en un procedimiento de ensamble; para ello, detallaremos la cartera de clientes empleada, definiremos la variable respuesta empleada en el modelo, generaremos y seleccionaremos las variables explicativas con mayor poder de predicción del comportamiento de pago del cliente, entrenaremos el modelo desarrollado mediante las técnicas estadísticas detalladas en la sección 2.3 y 2.4, por último, validaremos el modelo scoring predictivo.

3.1 Selección y consistencia de la cartera de clientes

La Figura 3.1 indica las fechas en las cuáles se levantó la información empleada para la construcción del modelo scoring y serán denominadas *puntos de observación*. Notemos que los puntos de observación son excluyentes entre sí (Ver Tabla 3.1), puesto que tomamos en consideración el último dígito del documento de identificación.

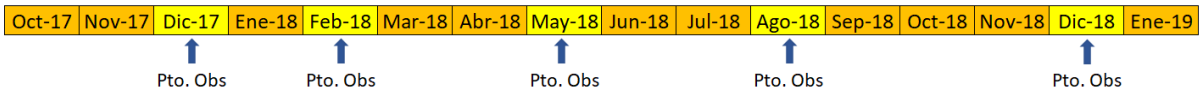


Figura 3.1: Esquema de generación de información

Punto de observación	Dígito
Dic - 17	3
Feb - 18	5
May - 18	7
Ago - 18	9
Dic - 18	2

Tabla 3.1: Generación de muestras excluyentes

Adicionalmente, en la Tabla 3.2 presentamos la distribución de la cartera de clientes de acuerdo al punto de observación en la entidad financiera; es importante mencionar que para la construcción de los modelos consideramos únicamente personas naturales con tipo de documento de identificación: Cédula y Extranjero que presentan información en algunos de los siguientes sistemas: Sistema de Bancos, Sistema de Cooperativas y Sistema Comercial.

Punto de observación	Sujetos	Porcentaje
Dic - 17	105.822	18,71 %
Feb - 18	109.982	19,44 %
May - 18	113.256	20,02 %
Ago - 18	115.436	20,41 %
Dic - 18	121.197	21,42 %
Total	**565.693**	**100,00 %**

Tabla 3.2: Cartera total de clientes

La Figura 3.2 expone el esquema de la ventana histórica, punto de observación y ventana de desempeño que serán descritas una a una.

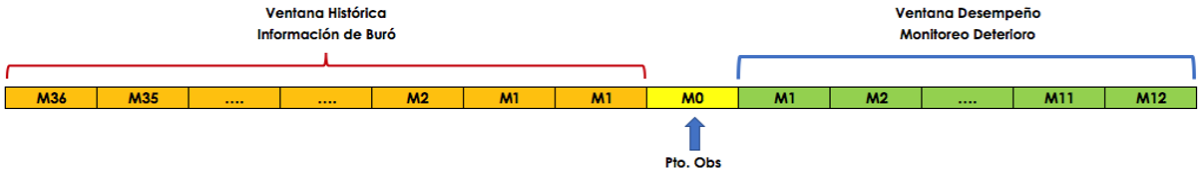


Figura 3.2: Esquema de requerimiento de información

La ventana histórica se genera a través de los últimos 36 meses consecutivos al punto de observación, se establecen 36 pues son los meses de historial que permite emplear la Superintendencia de Bancos (SB) para la construcción de modelos scoring y contiene información de Buró, es decir, historial crediticio y financiero de cada sujeto de la cartera de clientes a nivel de todo el mercado, la información de esta ventana permite la generación de variables asociadas al comportamiento crediticio del sujeto. A continuación, se presentan las principales variables de la ventana histórica:

1. Número de operaciones refinanciadas
2. Número de operaciones vigentes por vencer
3. Número de operaciones vigentes en deuda vencida
4. Número de operaciones vigentes que no devengan intereses
5. Número de operaciones vigentes en demanda judicial
6. Número de operaciones vigentes en cartera castigada
7. Numero de operaciones vigentes aperturadas

8. Mayor valor vencido
9. Mayor valor en demanda judicial
10. Mayor valor en cartera castigada
11. Suma de deuda total de las operaciones
12. Número de entidades en que registra vencidos
13. Promedio del máximo numero de días vencidos
14. Antigüedad del sujeto
15. Promedio de la deuda por vencer
16. Promedio de la deuda no devenga intereses
17. Promedio de la deuda vencida
18. Promedio de la deuda demanda judicial
19. Promedio de la deuda cartera castigada
20. Promedio de la deuda por vencer
21. Promedio de la deuda no devenga intereses
22. Promedio de la deuda vencida
23. Promedio de la deuda demanda judicial
24. Promedio de la deuda cartera castigada

Ahora, la información de la cartera total de clientes vigentes y la información socio demográfica del sujeto como por ejemplo: estado civil, género, grado de instrucción, profesión, provincia, cantón, parroquia y número de hijos se obtiene en los puntos de observación. Las principales variables que se encuentran en los puntos de observación son las siguientes:

1. Deuda total de las operaciones de consumo
2. Deuda total de las operaciones de microcrédito
3. Deuda total de las operaciones de comercial
4. Deuda total de las operaciones de vivienda
5. Deuda total de las operaciones de productivo
6. Deuda total de las operaciones de otras operaciones
7. Deuda vencida
8. Deuda de demanda judicial
9. Deuda castigada
10. Número de operaciones de consumo
11. Número de operaciones de microcrédito
12. Número de operaciones de comercial
13. Número de operaciones de vivienda
14. Número de operaciones de productivo
15. Número de operaciones de otras operaciones
16. Máximo número de días vencidos

17. Número de entidades en los que registra vencidos

La ventana de desempeño constituida por los 12 meses posteriores al punto de observación permite monitorear el deterioro de conducta de pago de cada sujeto, esta información permite definir las categorías de la variable dependiente del modelo, pues asumimos el supuesto: el comportamiento futuro es un reflejo del comportamiento de pago del pasado de cada uno de los sujetos de la cartera. Las principales variables en la ventana de desempeño son las siguientes:

1. Saldo de deuda
2. Saldo por vencer
3. Saldo vencido
4. Saldo que no devenga intereses
5. Saldo de cartera castigada
6. Saldo de demanda judicial
7. Número de días vencidos

Dentro de la información correspondiente a la ventana de desempeño se identifican dos categorías de la variable dependiente: Sin Desempeño y No Bancarizado, las cuales facilitan determinar la población de modelamiento.

La categoría Sin Desempeño de la variable respuesta comprende los sujetos que disponen información menor a 6 meses en la ventana posterior al punto de observación y se maneja de forma particular ya que no se puede tomar una decisión de la forma de comportamiento de pago en los siguientes 12 meses, si un individuo presenta un número menor a 6 saldos en deuda de operaciones o tarjetas de crédito.

La Tabla 3.3 expone el número de meses en los que un individuo presenta información dentro de la ventana de desempeño.

Se observa que existen 10.636 registros para los cuales no se tiene ningún saldo en la ventana de desempeño, es decir, los individuos solo obtuvieron saldo en el punto de observación; posiblemente dichos individuos terminaron de pagar la deuda de operación o tarjeta de crédito. Luego, el desempeño 1 se interpreta de la siguiente manera: si se analiza una ventana de 12 meses posteriores al punto de observación los 8.915 individuos presentan únicamente un saldo en un mes de los 12 meses que se analizan. La descripción de desempeño para los siguientes meses se realiza de forma análoga al desempeño 1.

Meses	Sujetos	Porcentaje
0	10.636	1,88 %
1	8.915	1,58 %
2	9.024	1,60 %
3	9.037	1,60 %
4	10.389	1,84 %
5	10.684	1,89 %
6	10.684	1,89 %
7	10.548	1,86 %
8	10.602	1,87 %
9	12.346	2,18 %
10	13.406	2,37 %
11	21.564	3,81 %
12	427.858	75,63 %
Total	**565.693**	**100,00 %**

Tabla 3.3: Desempeño de los individuos

En tanto que si la deuda total en el Sistema de Bancos, Sistema de Cooperativas o Sistema Comercial en los 12 meses anteriores al punto de observación es igual a cero, es decir, si el individuo no presenta créditos en el último año es etiquetado como No Bancarizado en la variable dependiente. La cartera de clientes empleada consta de 9.539 personas etiquetadas como No Bancarizadas que representan el 1.69 % del total de clientes. Es importante distinguir este grupo de personas, puesto que, el modelo predictivo se basa en la información histórica, por lo cual, se exige que el individuo tenga información histórica al menos en el último año anterior al punto de observación.

Finalmente, se observa que la cartera de clientes seleccionada para el estudio es madura y estable, puesto que el período de tiempo empleado es adecuado para conocer de forma correcta el comportamiento de pago del individuo y se mantiene la tasa de morosidad en el curso del tiempo.

3.2 Metodología Roll Rate

La metodología Roll Rate permite analizar la morosidad de los sujetos mediante una matriz de transición que identifica el punto a partir del cual la tasa de deterioro crediticio se incrementa y se estabiliza, generalmente se consideran los tramos vencidos de un mes versus el tramo de vencido al mes siguiente.

En este caso para aplicar Roll Rate y con el propósito de ampliar la discriminación entre clientes excluiremos: los sujetos que presentan cartera castigada y/o demanda judicial al

punto de observación porque son clientes que difícilmente puedan recuperarse, solicitantes no bancarizados, es decir, personas que no presentan deuda en el sistema crediticio en los últimos 12 meses anteriores al punto de observación, finalmente, los individuos que presentan información menor a seis meses dentro de la ventana de desempeño. El período de evaluación será de 12 meses correspondientes a la ventana de desempeño y la Tabla 3.4 indica los estados empleados en la metodología.

Rango de Vencido	**Descripción al final del período**
Sin vencido	Al día y tiene máximo 0 días de morosidad durante el período.
De 1 a 30 días	Presenta entre 1 a 30 días de morosidad durante el período.
De 31 a 60 días	Presenta entre 31 a 60 días de morosidad durante el período.
De 61 a 90 días	Presenta entre 61 a 30 días de morosidad durante el período.
De 91 a 120 días	Presenta entre 91 a 120 días de morosidad durante el período.
De 121 a 150 días	Presenta entre 121 a 150 días de morosidad durante el período.
De 151 a 180 días	Presenta entre 151 a 180 días de morosidad durante el período.
Más de 180 días	Presenta más de 180 días de morosidad durante el período.

Tabla 3.4: Estados de los rangos de vencido de Roll Rate

En la Tabla 3.5 presentamos la distribución promedio de los 12 meses por rango de vencido, con el propósito de identificar la tasa de rotación de los clientes, etiquetamos como Si Avanza a aquellos individuos que tienen un incremento en la tasa de deterioro, es decir, hallándose en un rango de vencido específico, transitan a un diferente tramo de vencido al mes siguiente. Además, el procedimiento para definir cada zona coloreada y posterior establecer las categorías de la variable dependiente se encuentra sujeta al nivel de riesgo que asume la institución financiera, comúnmente se opera de la siguiente manera:

- **Zona Verde:** el porcentaje de la etiqueta Si Avanza es menor al 10 %, es decir, de cada 100 personas existen a lo más 10 sujetos que empeoran su estatus (rango) de morosidad.
- **Zona Amarilla:** la etiqueta Si Avanza alcanza un porcentaje mayor o igual al 10 % y menor al 40 %, esto es, que la tasa de deterioro del sujeto es alta para pertenecer a la zona verde, y al mismo tiempo, la tasa de deterioro del sujeto es baja para pertenecer a la zona roja.
- **Zona Roja:** la etiqueta Si Avanza presenta un porcentaje mayor o igual al 40 %, es decir, de cada 100 personas existen más de 40 que empeoran su estatus inicial (rango de vencido), por lo tanto, la brecha en los porcentajes para la distinción entre la zona roja y zona verde es amplia.

Rangos de Vencido	Roll Rate				
	No Avanza		Si Avanza		Total
	Sujetos	Porcentaje	Sujetos	Porcentaje	Sujetos
Sin vencido	311.482	92,98 %	23.500	7,02 %	334.982
De 1 a 29 días	39.365	85,48 %	6.689	14,52 %	46.054
De 30 a 59 días	6.459	61,87 %	3.981	38,13 %	10.440
De 60 a 89 días	2.279	45,98 %	2.677	54,02 %	4.956
De 90 a 119 días	2.000	46,19 %	2.330	53,81 %	4.330
De 120 a 149 días	676	26,77 %	1.849	73,23 %	2.525
De 150 a 179 días	446	22,05 %	1.577	77,95 %	2.023
Más de 180 días	2.297	6,09 %	35.404	93,91 %	37.701
Total	**365.004**	**82,39 %**	**78.007**	**17,61 %**	**443.011**

Tabla 3.5: Análisis Roll Rate

A partir de la Tabla 3.5 se establece 60 días en los rangos de vencido analizados para considerar que el comportamiento del cliente tiene una alta probabilidad de empeorar su rango de morosidad y por tal razón, se considerarán como sujetos malos a aquellos que presentan 60 o más días de morosidad.

3.3 Definición de la variable dependiente

La definición de la variable dependiente es crítica para asegurar que el modelo clasificador pronostique el evento de interés, por lo cual, la construcción de la variable respuesta empleada considera un criterio objetivo en función de la definición de mora y los saldos de vencido que presenta el cliente.

A continuación, analizamos los saldos de vencido y los tramos de días de morosidad con el fin de mejorar la distribución de cada categoría de la variable dependiente, puesto que, los sujetos que pertenecen a la cartera pueden presentar saldos de vencido relativamente bajos en los 60 días de mora y una cantidad de días de vencido alta. Por ejemplo: un sujeto con monto medio vencido de $5 y días de vencido superior a 90 días.

Para este análisis separamos los individuos que presentan saldo vencido, cartera castigada y/o demanda judicial al punto de observación y sujetos no bancarizados.

Considerando la Tabla 3.6 correspondiente al promedio de deuda vencida mensual (incluye demanda y castigo) versus el rango de días morosidad en la ventana de desempeño evidenciamos que para saldos de vencido mayor a $19,13 la tasa de representatividad entre el saldo vencido y la deuda total da un salto considerable de 3,06 % a 17,90 %.

Promedio Saldo Vencido 12M	Sin Vencido	De 1 a 30	De 31 a 60	De 61 a 90	De 91 a 120	De 121 a 150	De 151 a 180	Más de 180
0	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %
(0 a 0,007]	0,00 %	0,00 %	0,00 %	0,00 %	0,00 %		0,00 %	20,00 %
(0.007 a 0,03]	0,01 %	0,00 %	0,65 %	0,00 %	0,01 %	0,00 %	0,00 %	5,27 %
(0.03 a 0,11]	0,05 %	0,00 %	0,00 %	3,76 %	0,02 %	0,00 %	0,00 %	26,14 %
(0,11 a 0,49]	1,09 %	0,04 %	0,51 %	0,46 %	10,13 %	0,01 %	0,01 %	32,05 %
(0,49 a 0,81]	0,21 %	0,10 %	0,52 %	4,60 %	14,35 %	52,30 %	0,01 %	17,55 %
(0,81 a 1,59]	0,37 %	0,15 %	0,82 %	4,14 %	16,73 %	9,74 %	0,04 %	40,72 %
(1,59 a 2,84]	0,50 %	0,27 %	0,71 %	5,05 %	8,09 %	10,31 %	27,30 %	51,33 %
(2,84 a 4,76]	0,49 %	0,40 %	1,54 %	4,26 %	16,73 %	16,80 %	5,84 %	58,50 %
(4,76 a 7,85]	0,64 %	0,65 %	1,49 %	3,05 %	4,40 %	8,31 %	61,18 %	66,85 %
(7,85 a 12,34]	1,87 %	0,82 %	2,10 %	6,33 %	6,28 %	15,70 %	13,59 %	81,24 %
(12,34 a 19,13]	3,06 %	1,03 %	1,72 %	6,11 %	6,95 %	16,93 %	12,34 %	94,29 %
(19,13 a 32,08]	17,90 %	1,11 %	1,93 %	4,29 %	8,29 %	15,59 %	33,43 %	79,50 %
(32,08 a 53,27]	31,45 %	2,73 %	2,14 %	4,86 %	8,25 %	18,18 %	27,41 %	78,11 %
(53,27 a 84,44]	100,00 %	2,26 %	2,36 %	4,50 %	7,91 %	11,03 %	17,90 %	87,95 %
(84,44 a 131,4]	86,75 %	2,00 %	2,39 %	4,95 %	7,68 %	12,15 %	19,02 %	91,04 %
(131,4 a 218,8]	91,49 %	3,58 %	2,62 %	4,82 %	7,01 %	11,40 %	16,04 %	81,42 %
(218,8 a 390,8]	96,70 %	4,60 %	4,11 %	4,81 %	7,15 %	11,06 %	16,22 %	80,77 %
(390,8 a 745,2]	95,54 %	7,64 %	3,73 %	5,22 %	7,42 %	11,43 %	15,70 %	71,23 %
(745,2 a 1.723]	96,87 %	24,48 %	6,65 %	5,83 %	8,81 %	12,47 %	17,42 %	62,72 %
Más de 1.723	94,42 %	48,52 %	23,13 %	8,82 %	11,67 %	17,42 %	27,55 %	62,93 %

Tabla 3.6: Porcentaje Saldo Vencido / Deuda Total

Luego, como el objetivo es predecir el riesgo total en las operaciones sobre la población empleada mediante el comportamiento de pago de cada individuo en la institución financiera, definimos las categorías de la variable dependiente en base al análisis combinado del número de días mora y el monto medio vencido en la ventana de desempeño conjuntamente con la metodología Roll Rate como sigue:

- **Bueno**: Si no presenta días de morosidad y el promedio de deuda vencida no supera los $19 ó si el promedio de deuda vencida es cero en la ventana de desempeño.
- **Malo**: Si alcanza una mora mayor a 60 días y el promedio de deuda vencida es mayor a cero ó si el promedio de deuda vencida es superior a $19 en la ventana de desempeño.
- **Malo observado**: Si presenta cartera castigada y demanda judicial en el mes de observación.
- **Sin desempeño**: Si registra deuda en menos de 6 meses de la ventana de desempeño.
- **No Bancarizado**: Sujetos con historial crediticio inferior a un año.
- **Indeterminado**: Caso contrario.

A continuación, en la Tabla 3.7 se presenta el resumen de la definición y distribución de la variable dependiente aplicada a la cartera total de clientes.

Categoría	Valor	Sujetos	Porcentaje
Bueno	0	279.663	49,44 %
Malo	1	100.786	17,82 %
Indeterminado	2	62.561	11,06 %
Malo Observado	3	57.271	10,12 %
Sin desempeño	4	55.873	9,88 %
No Bancarizado	5	9.539	1,69 %
Total	—	**565.693**	**100,00 %**

Tabla 3.7: Distribución de la Variable Dependiente

Para la construcción de un buen modelo de acuerdo a las buenas prácticas y en base al estudio realizado por (Siddiqi, 2012) el porcentaje máximo de indeterminados debe ser inferior al 10 % o 15 % de la cartera, analizando nuestra distribución se observa un porcentaje de indeterminados del 11,06 % respecto de la población total. Además, la proporción de clientes buenos y clientes malos en la cartera es por cada cliente con etiqueta de Malo, alrededor de 2 clientes son etiquetados como Buenos.

3.4 Segmentación

Para mejorar el poder predictivo del modelo clasificador binario, se emplea un criterio de segmentación de la cartera total de clientes con el propósito de identificar grupos homogéneos, tal que, dentro de cada grupo el comportamiento relacionado al incumplimiento de pago de los clientes sea similar.

La segmentación que genera resultados óptimos es la agrupación de clientes de acuerdo al promedio del máximo número de días vencidos en los últimos 12 meses anteriores al mes de observación en el Sistema Crediticio Ecuatoriano (SCE), de esta manera la población bancarizada se divide en los segmentos **Clean** y **Dirty**.

Denominamos segmento Clean al conjunto de sujetos que registran un promedio del máximo número de días vencidos en los últimos 12 meses anteriores al mes de observación en SCE menor a 30 días y denominaremos segmento Dirty al conjunto de sujetos que registran un promedio del máximo número de días vencidos en los últimos 12 meses anteriores al mes de observación en SCE mayor o igual a 30 días. En la Tabla 3.8 se presenta el criterio de segmentación para las categorías Bueno y Malo de la variable dependiente definida en la sección 3.3.

Segmento	Good Bad 60					
	Bueno	**% Bueno**	**Malo**	**Tasa Malo**	**Total**	**% Sujetos**
Clean	270.444	86,58 %	41.903	13,42 %	312.347	82,10 %
Dirty	9.219	13,54 %	58.883	86,46 %	68.102	17,90 %
Total	**279.663**	**73,51 %**	**100.786**	**26,49 %**	**380.449**	**100,00 %**

Tabla 3.8: Distribución de segmentación Clean/Dirty

A partir de lo expuesto en la Tabla 3.8 se concluye lo siguiente: el segmento Clean consta de 312.347 sujetos, donde el 13,42 % del total del segmento son clientes etiquetados como Malos y el 86,58 % del total del segmento representa los clientes etiquetados como Buenos; el 86,46 % del total de sujetos que pertenecen al segmento Dirty corresponden a clientes que debido a su comportamiento de pago son considerados Malos mientras que el 13,54 % del total de clientes del segmento corresponden a clientes catalogados como Buenos.

En la construcción de modelos de clasificación binaria, generalmente, la segmentación de la cartera de clientes es parte importante si existe una amplia diferencia entre las tasas de malo de los diferentes grupos generados, por lo tanto, en base a la Tabla 3.8 se concluye que el criterio de segmentación es significativo y representativo. Además, dicho criterio nos permite desarrollar dos modelos independientes para los segmentos Clean y Dirty obteniendo resultados satisfactorios.

En el presente documento cada uno de los modelos score de riesgos individuales como el modelo score de riesgos de ensamble para una población bancarizada constan de dos segmentos; en la Tabla 3.9 exponemos el nombre y la forma que definimos cada uno de los segmentos.

Segmento	Condición
Clean	Si $prom\_max\_dven\_sce\_12M < 30$
Dirty	Si $prom\_max\_dven\_sce\_12M \geq 30$

Tabla 3.9: Definición de los segmentos

A continuación se detalla la construcción de la variable empleada en la Tabla 3.9.

prom_max_dven_sce_12M: Promedio del máximo número de días vencidos de las operaciones y tarjetas de crédito que registra el sujeto en el Sistema Crediticio Ecuatoriano en los últimos 12 meses anteriores al punto de observación.

$$prom\_max\_dven\_sce\_12M = max(prom\_max\_dven\_sbs\_op\_12M,$$
$$prom\_max\_dven\_sc\_op\_12M, prom\_max\_dven\_sicom\_op\_12M,$$
$$prom\_max\_dven\_sbs\_tc\_12M, prom\_max\_dven\_sc\_tc\_12M,$$
$$prom\_max\_dven\_sicom\_tc\_12M)$$

donde:

- *prom_ max_ dven_ sbs_ op_ 12M:* Promedio del máximo número de días vencidos de las operaciones que registra el sujeto en el Sistema Regulado por la SB en los últimos 12 meses anteriores al punto de observación.
- *prom_ max_ dven_ sc_ op_ 12M:* Promedio del máximo número de días vencidos de las operaciones que registra el sujeto en el Sistema Regulado por la SEPS en los últimos 12 meses anteriores al punto de observación.
- *prom_ max_ dven_ sicom_ op_ 12M:* Promedio del máximo número de días vencidos de las operaciones que registra el sujeto en el Sistema Comercial en los últimos 12 meses anteriores al punto de observación.
- *prom_ max_ dven_ sbs_ tc_ 12M:* Promedio del máximo número de días vencidos de las tarjetas de crédito que registra el sujeto en el Sistema Regulado por la SB en los últimos 12 meses anteriores al punto de observación.
- *prom_ max_ dven_ sc_ tc_ 12M:* Promedio del máximo número de días vencidos de las tarjetas de crédito que registra el sujeto en el Sistema Regulado por la SEPS en los últimos 12 meses anteriores al punto de observación.
- *prom_ max_ dven_ sicom_ tc_ 12M:* Promedio del máximo número de días vencidos de las tarjetas de crédito que registra el sujeto en el Sistema Comercial en los últimos 12 meses anteriores al punto de observación.

3.5 Muestra de modelamiento y validación

Una vez que se realiza la segmentación de grupos homogéneos, procedemos a dividir la cartera total de clientes, cada subdivisión está constituida por sujetos con comportamiento de pago variado (bueno, malo, indeterminado, malo observado, sin desempeño y no bancarizado) los cuales representan posibles sujetos que aplicarán a una solicitud de operación crediticia o tarjeta de crédito en un futuro por cada segmento.

Emplearemos muestreo aleatorio simple para la división de la cartera de clientes; el primer porcentaje corresponde al 50 % del total de la población (282.846 individuos). La submuestra es denominada muestra de modelamiento y es empleada para el desarrollo

de la metodología y construcción del modelo, su distribución de acuerdo a cada grupo homogéneo se presenta en la Tabla 3.10.

Categoría	Segmento				
	Clean	**Porcentaje**	**Dirty**	**Porcentaje**	**Total**
Bueno	135.011	59,07 %	4.627	8,52 %	139.638
Malo	20.828	9,11 %	29.684	54,66 %	50.512
Indeterminado	28.128	12,31 %	3423	6,30 %	31.551
Malo Observado	16.455	7,20 %	12.020	22,14 %	28.475
Sin desempeño	23.407	10,24 %	4.548	8,38 %	27.955
No bancarizado	4.715	2,06 %	0	0,00 %	4.715
Total	**228.544**	**100,00 %**	**54.302**	**100,00 %**	**282.846**

Tabla 3.10: Muestra de modelamiento

Secuencialmente, con el propósito de verificar la estabilidad y robustez del modelo clasificador validamos la metodología desarrollada en un segundo porcentaje correspondiente al 50 % de la cartera original (282.847 individuos) conocido como muestra de validación, su distribución de acuerdo a la segmentación Clean/Dirty se presenta en la Tabla 3.11.

Categoría	Segmento				
	Clean	**Porcentaje**	**Dirty**	**Porcentaje**	**Total**
Bueno	135.433	59,15 %	4.592	8,52 %	140.025
Malo	21.075	9,20 %	29.199	54,20 %	50.274
Indeterminado	27.739	12,11 %	3.271	6,07 %	31.010
Malo Observado	16.471	7,19 %	12.325	22,88 %	28.796
Sin desempeño	23.431	10,23 %	4.487	8,33 %	27.918
No bancarizado	4.824	2,11 %	0	0,00 %	4.824
Total	**228.973**	**100,00 %**	**54.302**	**100,00 %**	**282.847**

Tabla 3.11: Muestra de validación

Consideramos importante mencionar que en la fase de entrenamiento de los modelos se contempla únicamente personas en las dos categorías de la variable dependiente: Bueno y Malo, con el propósito de desarrollar los clasificadores binarios que identifiquen las variables explicativas que realizan una buena distinción del comportamiento de pago de los clientes etiquetados como buenos y malos, mientras que en la fase de validación se emplearán todas las categorías de la variable dependiente.

3.6 Generación de variables explicativas

La generación de variables explicativas es de importancia para discriminar a un cliente bueno o malo en base a información del comportamiento de pago en operaciones y tarjetas de crédito dentro de la institución como en un entorno externo a la entidad financiera. El objetivo de la sección es generar variables explicativas que determinen de

manera individual o conjunta las características de la cartera de clientes que inducen un comportamiento de impago en las operaciones adquiridas por los sujetos.

En nuestro estudio esencialmente se incluyeron variables de comportamiento del individuo en el sistema regulado por la Superintendencia de Bancos, el sistema controlado por la Superintendencia de Economía Popular y Solidaria y el sistema comercial, variables de condiciones de crédito y variables financieras generadas desde la información de la ventana histórica, también, variables sociodemográficas obtenidas al punto de observación, por ejemplo: edad, género, estado civil, instrucción, provincia, entre otras.

Considerando la información disponible de las operaciones adquiridas por el sujeto proporcionada por la institución financiera privada como las herramientas y técnicas estadísticas matemáticas, construiremos y definiremos las variables explicativas finales que integran el modelo clasificador final.

Habitualmente en la práctica se generan variables acumuladas, ratios y logaritmos en función de variables que identifican características de deterioro de pago del cliente mediante la medición de los siguientes factores:

- La mora o incumplimiento de obligaciones del individuo en meses anteriores al punto de observación.
- El endeudamiento adquirido por el individuo en los productos crediticios a nivel interno y externo de la institución financiera.
- Hábito y preferencias de pago de obligaciones adquiridas de los individuos.
- Antigüedad de operaciones adquiridas (último refinanciamiento, última operación aperturada, etc).
- Número de obligaciones adquiridas en cada uno de los sistemas.

Generamos las variables acumuladas para los últimos 3, 6, 12, 24 y 36 meses en base a las variables primarias de los sistemas: SB, SC y SICOM que ofrezcan productos de operaciones y/o tarjetas de crédito a los clientes. A continuación, presentamos un ejemplo de la construcción de la variable acumulativa: $nope\_apert\_sce\_3M$.

$$\begin{aligned} nope\_apert\_sce\_3M = & nope\_apert\_sbs\_op\_3M + nope\_apert\_sc\_op\_3M + \\ & nope\_apert\_sicom\_op\_3M \end{aligned} \tag{3.1}$$

En la ecuación (3.1) observamos que la variable acumulada es generada de la suma de tres variables pertenecientes a brindar características del número de operaciones aperturadas por parte del cliente en el SB, SC y SICOM en los 3 últimos meses anteriores al punto

de observación.

Adicionalmente, se construyeron ratios de variación, ratios de representatividad y ratios entre sistemas, como ejemplo en la ecuación (3.2) tenemos la generación del ratio de representatividad de la deuda total en el SB respecto a la deuda total en el sistema crediticio ecuatoriano presentado en los últimos 6 meses.

$$r\_deuda\_total\_sbsssce\_6M = \begin{cases} \frac{deuda\_total\_sbs\_6M}{deuda\_total\_sce\_6M} & \text{Si } deuda\_total\_sce\_6M > 0 \\ 0 & \text{Si } deuda\_total\_sce\_6M = 0 \end{cases} \quad (3.2)$$

donde la $deuda\_total\_sbs\_6M$ es la suma de las variables $deuda\_total\_sbs\_op\_6M$ y $deuda\_total\_sbs\_tc\_6M$ que indican la deuda total tanto en operaciones como tarjetas de crédito en el sistema bancario en los últimos 6 meses anteriores al punto de observación.

Finalmente realizamos transformaciones logarítmicas de variables que identifican el mayor valor vencido, saldos de deuda promedio y saldos de deuda total del individuo en operaciones y tarjetas de crédito en los sistemas de bancos, cooperativas y comercial en los últimos 3, 6, 12, 24 y 36 meses. La ecuación (3.3) presenta un ejemplo de generación de una variable mediante la transformación logarítmica de la deuda total en operaciones dentro del sistema de cooperativas en los últimos 6 meses anteriores al punto de observación.

$$ln\_deuda\_total\_sc\_op\_6M = \begin{cases} log(deuda\_total\_sc\_op\_6M) & \text{Si } deuda\_total\_sc\_op\_6M > 1 \\ 0 & \text{Si } deuda\_total\_sc\_op\_6M \leq 1 \end{cases} \quad (3.3)$$

Construimos alrededor de 1.780 variables entre variables acumuladas, ratios y logaritmos para el desarrollo del modelo clasificador binario.

3.7 Filtrado de variables explicativas

El propósito de la sección es identificar un subconjunto de variables generadas que expliquen la variable dependiente de manera eficaz en cada uno de los segmentos de la cartera de clientes, es decir, las variables que establezcan mayor poder de discriminación entre un cliente bueno y un cliente malo. Para esto emplearemos las medidas de separación o medidas de asociación resumidas en las secciones 2.1 y 2.2 respectivamente. Por último obtendremos una colección de variables para el segmento Clean y otra colección de variables explicativas para el segmento Dirty las cuales serán candidatas a formar parte de los modelos preliminares.

Únicamente empleando los sujetos bancarizados que pertenezcan a las categorías Bueno

o Malo de la variable dependiente en la muestra de modelamiento, se procede a filtrar las variables explicativas de la siguiente manera:

1. Etiquetamos el conjunto de datos para cada segmento (Clean/Dirty) de la cartera de clientes;
2. Identificamos las variables explicativas numéricas continuas y variables categóricas con un porcentaje menor o igual al 30 % de valores perdidos para cada cliente;
3. Identificamos y excluimos las variables explicativas con un porcentaje de observaciones iguales a una constante mayor o igual a 90 % para cada cliente; y,
4. Diferenciamos las variables cuantitativas de las variables cualitativas de la base de modelamiento para un filtrado específico de cada tipo de variable.

3.7.1 Filtrado de variables cuantitativas

Realizamos la selección de variables cuantitativas continuas de forma análoga para el segmento Clean y el segmento Dirty empleando los criterios del estadístico de Kolgomorov-Smirnov (KS) y el estadístico Anderson Darling (AD).

A continuación, se describe el proceso para la selección de variables que explican de mejor manera la variable dependiente definida:

1. Sea una variable cuantitativa arbitraria X_i de la base de modelamiento definiremos dos nuevas variables:
   - X_M: Variable formada por los valores de X_i de individuos con etiqueta Malo en la variable dependiente.
   - X_B: Variable formada por los valores de X_i de individuos pertenecientes a la categoría Bueno de la variable dependiente.
2. Comparamos las distribuciones de las variables explicativas cuantitativas X_M y X_B empleando las medidas de separación de la sección (2.1).
3. Seleccionamos las variables cuantitativas X_i que mejor expliquen la variable dependiente conjuntamente con el valor de los estadísticos KS, AD y un indicador del poder predictivo (IND) generado por la combinación convexa de KS y AD, es decir, $IND = 0,40KS + 0,60AD$.

Para el segmento Dirty de la base de modelamiento calculamos los estadísticos KS y AD para 1.780 variables[1]. Observamos en la Figura 3.3 que el valor del indicador IND después de la variable 250 es inferior a 0.10 y tiende a ser cero, por lo tanto analizamos las

[1]El Top 75 de las variables cuantitativas con el valor de los estadísticos KS y AD respectivo del segmento Dirty se observan en el ANEXO B.

primeras 250 variables y aquellas que presenten mayor divergencia serán empleadas en la construcción de clasificadores individuales y ensambles clasificadores, variables dummy y probabilidades de malo con la técnica de árboles de decisión.

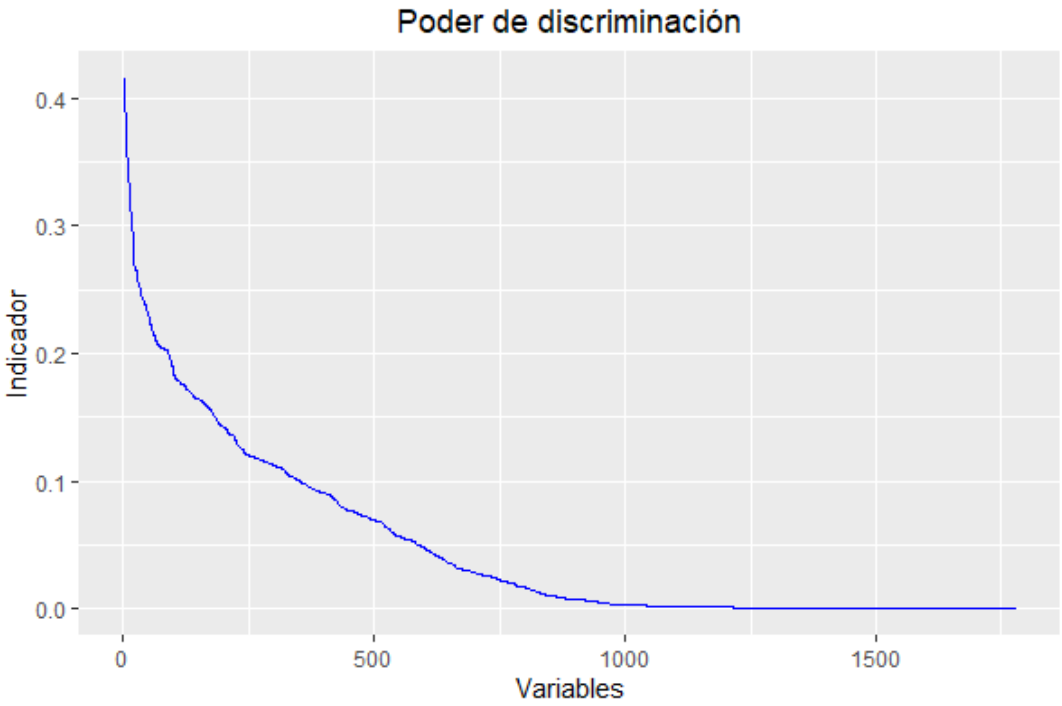


Figura 3.3: Poder de discriminación de variables en el segmento Dirty

Análogamente, en el segmento Clean calculamos los estadísticos *KS* y *AD* para 1.780 variables[2] y después de seleccionar las primeras 100 variables con un valor del indicador *IND* significativo analizamos la posibilidad de incluir una colección de variables que mejor expliquen la variable dependiente en los modelos de clasificación o utilizarlas en la construcción de variables dummy y probabilidades de malo con la técnica de árboles de decisión. La Figura 3.4 presenta el gráfico de sedimentación para el indicador *IND* en el segmento Clean.

3.7.2 Filtrado de variables cualitativas

El propósito es identificar las variables cualitativas con mayor poder predictivo en la distinción entre individuos etiquetados como buenos o malos de la variable dependiente utilizando medidas de asociación explicadas en la sección 2.2.

[2]El Top 75 de las variables cuantitativas con el valor de los estadísticos KS y AD respectivo del segmento Clean se observan en el ANEXO B.

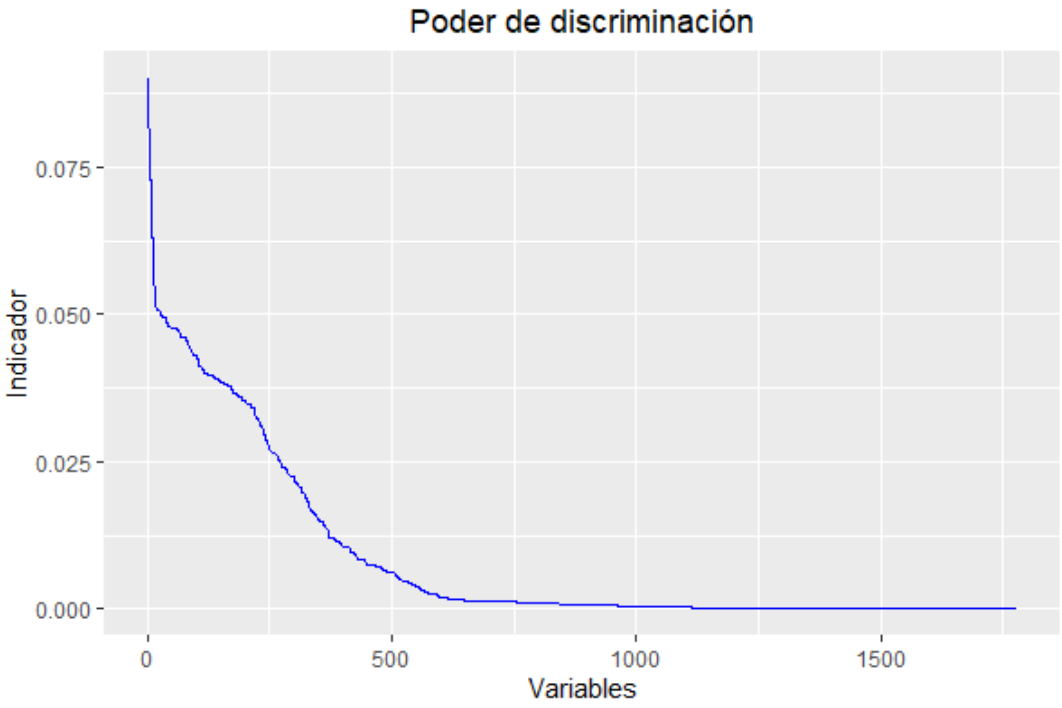


Figura 3.4: Poder de discriminación de variables en el segmento Clean

Para una variable cualitativa arbitraria X_i con k categorías de la base de modelamiento comparamos las distribuciones de los individuos etiquetados como Malo en cada categoría de la variable cualitativa y medimos la diferencia del número de individuos Malos entre las categorías empleando el índice de valor de información (VI).

Luego calculamos el índice de valor de información para las variables cualitativas en el segmento Clean y el segmento Dirty[3] observando que el número de variables cualitativas es inferior al número de variables cuantitativas, más aún, se reduce la cantidad de variables categóricas con valor de información significativo en cada segmento. Así, los gráficos de sedimentación para el índice de valor de información de cada segmento (ver Figuras 3.5 y 3.6) reflejan que las primeras 5 variables tienen importancia para la construcción de variables dummy y probabilidades de malo mediante la técnica de árboles de decisión.

[3]La lista de variables cualitativas y el índice de VI respectivo para los segmentos Clean y Dirty están ubicados en el ANEXO C.

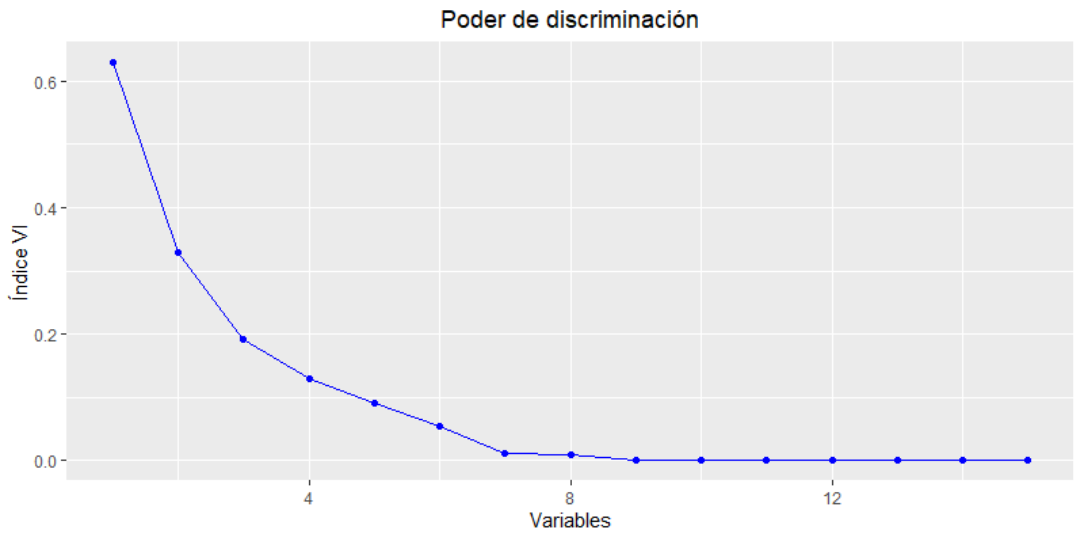


Figura 3.5: Poder predictivo de variables en el segmento Dirty

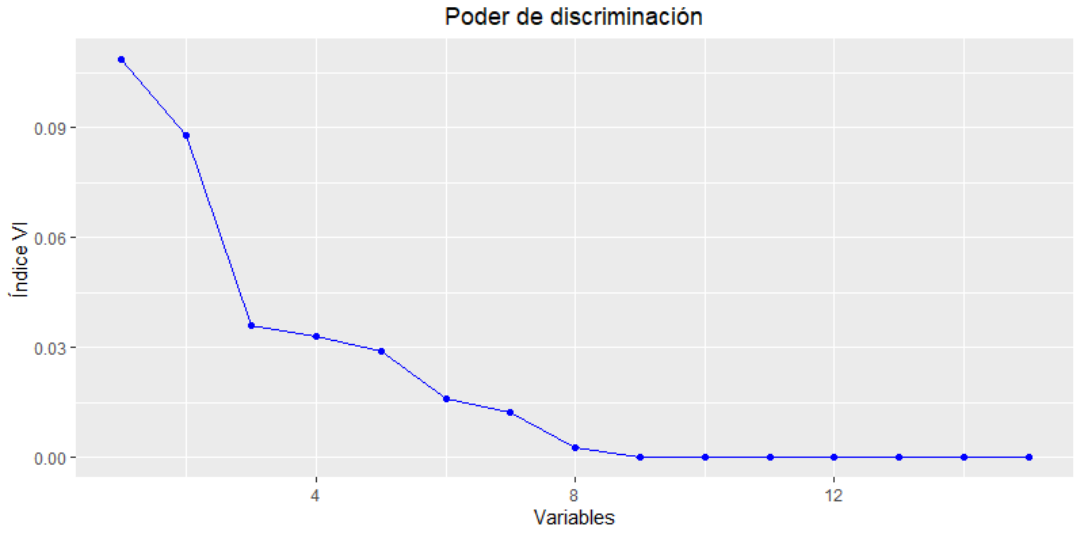


Figura 3.6: Poder predictivo de variables en el segmento Clean

3.8 Desarrollo del modelo de clasificación

El objetivo de esta sección es describir la metodología de construcción del modelo de clasificación binaria final implementado bajo el lenguaje de programación del software estadístico R[4] y la interfaz fundamentada en java H2O[5] que permite combinar distintos algoritmos de aprendizaje automático con el Big Data mediante la paralelización distribuida.

[4]La información del software estadístico está disponible en la página web: `http://www.r-project.org`

[5]La documentación completa del producto H2O creado por la compañía H2O.ai se encuentra en la página web: `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html`

El flujo del proceso para construir el modelo de clasificación binaria se representa en la Figura 3.7, se describe desde la generación y filtrado de variables explicativas de una base de datos hasta la selección del modelo final que mejor estima la probabilidad de malo conjuntamente con los resultados para su respectiva validación.

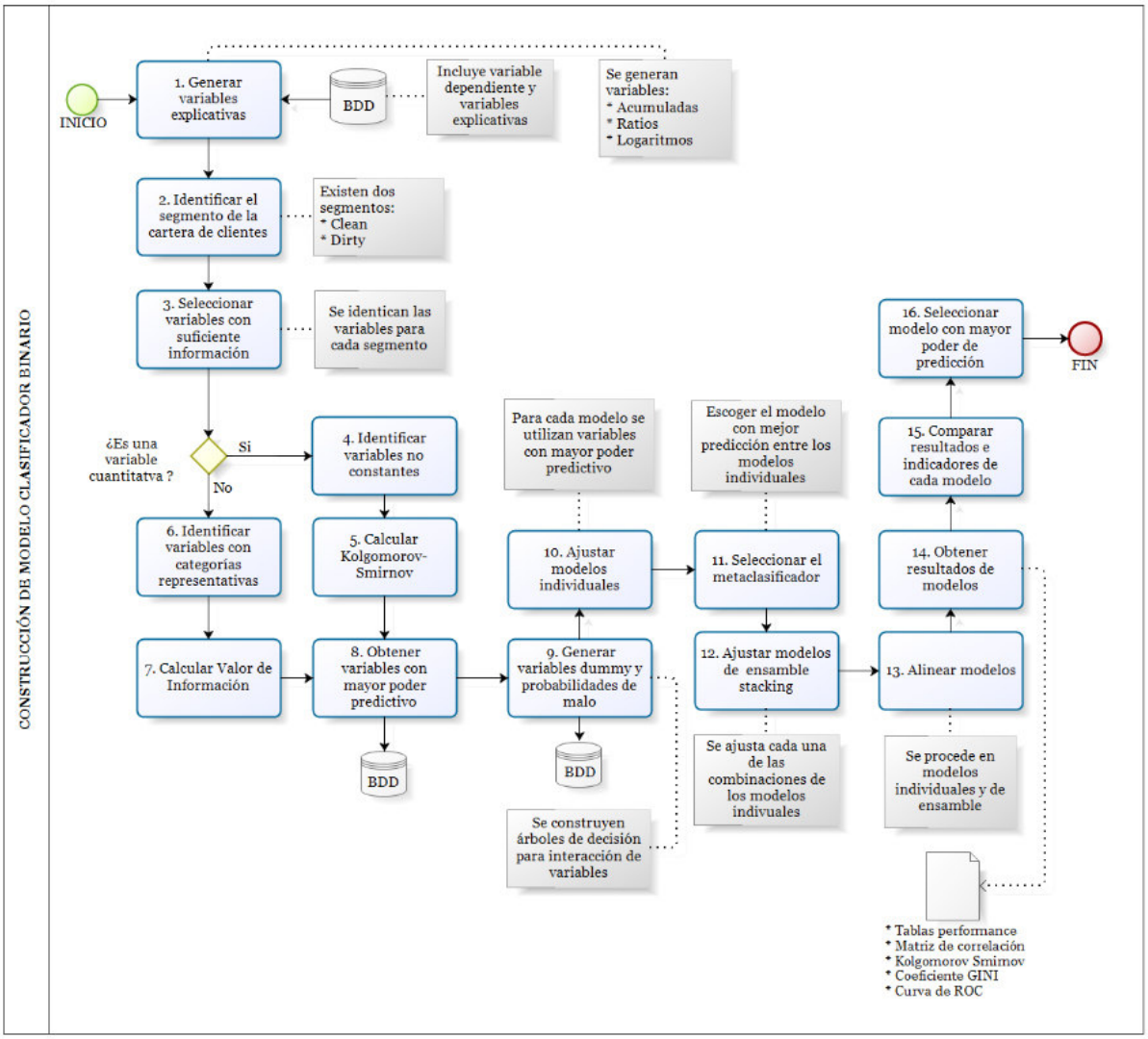


Figura 3.7: Flujograma de la construcción del modelo clasificador binario

3.8.1 Selección de variables del segmento Clean

La Tabla 3.12 presenta las variables explicativas que tienen mayor relación con la variable dependiente y son utilizadas en la construcción de los modelos clasificadores de primer nivel para el modelo de ensamble en el segmento Clean.

	Segmento Clean				
N°	**Variables explicativas**	**Modelos**			
		RF	**GBM**	**NB**	**RL**
1	*r_nope_apert_sicomssce_op_36M*	X	X	X	X
2	*porc_uso_cupo*	X	X	X	X
3	*d_nent_ven_sce_op_36M*	X	X	X	X
4	*antiguedad_sce*	X	X	X	X
5	*d_ntc_ndi_tc_36M*	X	X	X	X
6	*r_nope_apert_sce_12a24M*	X	X	X	X
7	*r_mvalven_sbs_tcsmalven_sbs_op_36M*	X	X	X	
8	*prbb_nope_xven_op_12M*	X	X	X	X
9	*PromLocalesCom*	X	X	X	X
10	*r_deuda_total_scssce_3M*	X	X	X	
11	*r_prom_ven_sbssprom_deuda_total_sbs_tc_36M*	X	X	X	
12	*prbb_ntc_apert_sce_24M*	X	X	X	X
13	*prbb_deuda_total_sce_3M*	X	X	X	
14	*r_nope_apert_sbssce_op_24M*				X

Tabla 3.12: Variables explicativas en el segmento Clean

A continuación detallamos cada variable explicativa que se muestra en la Tabla 3.12, además, la tercera, quinta, octava, décima segunda y décima tercera variables fueron construidas mediante la técnica de árboles de decisión utilizando un programa adicional conocido como IBM SPSS Modeler[6].

1. **r_nope_apert_sicomssce_op_36M:** Razón entre el número de operaciones vigentes aperturadas por el sujeto en el Sistema Comercial respecto al número de operaciones vigentes aperturas en el Sistema Crediticio Ecuatoriano en los últimos 36 meses anteriores al punto de observación.

Condición	**r_nope_apert_sicomssce_op_36M**
Si $nope\_apert\_sce\_36M = 0$	0
Si $nope\_apert\_sce\_36M > 0$	$\frac{nope\_apert\_sicom\_op\_36M}{nope\_apert\_sce\_36M}$

   donde:

   - *nope_apert_sicom_op_36M:* Número de operaciones vigentes aperturadas por el sujeto en los últimos 36 meses anteriores al punto de observación en el Sistema Comercial.
   - *nope_apert_sce_36M:* Número de operaciones vigentes aperturadas por el sujeto en los últimos 36 meses anteriores al punto de observación en el Sistema Crediticio Ecuatoriano.

[6]La información correspondiente a la instalación y manejo de SPSS Modeler se encuentra en la página web oficial: `https://www.ibm.com/products/spss-modeler`

2. **porc_uso_cupo:** Porcentaje del cupo total que el sujeto utiliza al punto de observación.

3. **d_nent_ven_sce_op_36M:** Variable binaria que toma el valor de 1 si en más de una entidad del Sistema Crediticio Ecuatoriano el sujeto registra vencidos en los últimos 36 meses anteriores al punto de observación y 0 caso contrario.

$$d\_nent\_ven\_sce\_op\_36M = \begin{cases} 1 & \text{Si } nent\_ven\_sce\_op\_36M > 0 \\ 0 & \text{Si } nent\_ven\_sce\_op\_36M = 0 \end{cases}$$

4. **antiguedad_sce:** Número de meses transcurridos desde que el sujeto apertura su primera operación en las entidades financieras pertenecientes al Sistema Comercial, Sistema Regulado por la SB y Sector Cooperativas (SEPS).

$$\begin{aligned} antiguedad\_sce = & mx(antiguedad\_op\_sbs, antiguedad\_tc\_sbs, \\ & antiguedad\_op\_sc, antiguedad\_tc\_sc, \\ & antiguedad\_op\_sicom, antiguedad\_tc\_sicom)/12 \end{aligned}$$

donde:

- *antiguedad_op_sbs:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en alguna entidad financiera del Sistema Regulado por la SBS.
- *antiguedad_tc_sbs:* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en alguna entidad financiera del Sistema Regulado por la SBS.
- *antiguedad_op_sc:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en una entidad financiera perteneciente al Sector de Cooperativas.
- *antiguedad_tc_sc:* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en una entidad financiera perteneciente al Sector de Cooperativas.
- *antiguedad_op_sicom:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en una entidad financiera perteneciente al Sistema Comercial.
- *antiguedad_tc_sicom* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en una entidad financiera perteneciente al Sistema Comercial.

5. **d_ntc_ndi_tc_36M:** Variable binaria que toma el valor de 1 si más de una

tarjeta vigente no devenga intereses del sujeto en los últimos 36 meses anteriores al punto de observación y 0 caso contrario.

$$d\_ntc\_ndi\_tc\_36M = \begin{cases} 1 & \text{Si } ntc\_ndi\_tc\_36M > 0 \\ 0 & \text{Si } ntc\_ndi\_tc\_36M = 0 \end{cases}$$

6. **r_nope_apert_sce_12a24M:** Razón entre el número de operaciones vigentes aperturadas en los últimos 12 meses respecto al número de operaciones vigentes aperturadas en los últimos 24 meses anteriores al punto de observación en las entidades pertenecientes al sistema regulado por la SB, SEPS y Sistema Comercial.

Condición	**r_nope_apert_sce_12a24M**
Si $nope\_apert\_sce\_24M = 0$	0
Si $nope\_apert\_sce\_24M > 0$	$\frac{nope\_apert\_sce\_12M}{nope\_apert\_sce\_24M}$

7. **r_mvalven_sbs_tcsmalven_sbs_op_36M:** Razón entre el mayor valor vencido de las tarjetas de crédito respecto al mayor valor vencido de las operaciones crediticias que registra el sujeto en las entidades financieras pertenecientes al Sistema Regulado por la SB en los últimos 36 meses anteriores al punto de observación.

Condición	**r_mvalven_sbs_tc y mvalven_sbs_op_36M**
Si $mvalven\_sbs\_tc\_36M = 0$ y $mvalven\_sbs\_op\_36M = 0$	0
Si $mvalven\_sbs\_tc\_36M > 0$ y $mvalven\_sbs\_op\_36M = 0$	1
Si $mvalven\_sbs\_op\_36M > 0$	$\frac{mvalven\_sbs\_tc\_36M}{mvalven\_sbs\_op\_36M}$
Caso Contrario	0

donde:

- *mvalven_sbs_tc_36M:* Mayor valor vencido de las tarjetas de crédito que registra el sujeto en las entidades financieras pertenecientes al Sistema Regulado por la SB en los últimos 36 meses anteriores al punto de observación.
- *mvalven_sbs_op_36M:* Mayor valor vencido de las operaciones de crédito que registra el sujeto en las entidades financieras pertenecientes al Sistema Regulado por la SB en los últimos 36 meses anteriores al punto de observación.

8. **prbb_nope_xven_op_12M:** Probabilidad de bueno del número de operaciones vigentes por vencer del sujeto en los últimos 12 meses anteriores al punto de observación.

Condición	prbb_nope_xven_op_12M
Si $nope\_xven\_op\_12M = 0$	0.90342
Si $nope\_xven\_op\_12M \leq 3$	0.87158
Si $nope\_xven\_op\_12M \leq 5$	0.83028
Caso contrario	0.79541

La probabilidad de bueno de la variable *nope_xven_op_12M* construimos en base al árbol de decisión de la Figura 3.8 y observamos que el porcentaje de sujetos con etiqueta de bueno decrece desde el Nodo 1 hasta el Nodo 4, lo cual nos indica que un sujeto con menor número de operaciones vigentes por vencer tiene mayor probabilidad de bueno.

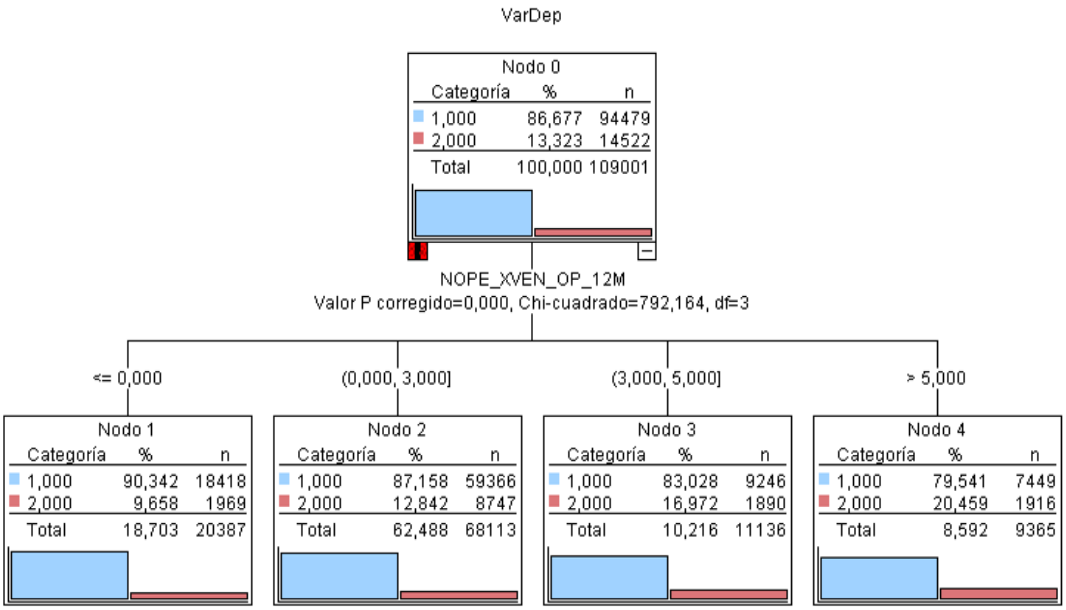


Figura 3.8: Árbol de decisión para la variable explicativa *nope_xven_op_12M*

9. **PromLocalesCom:** Promedio de locales comerciales que registra el sujeto a nivel parroquial; esta variable forma parte del conjunto de variables de indicadores agregados a nivel parroquial (INEC) y es asignada a cada sujeto respecto a su información de de provincia, cantón y parroquia levantada en el punto de observación.

10. **r_deuda_total_scssce_3M:** Razón de la suma de la deuda total de las operaciones y tarjetas de crédito que registra el sujeto en las entidades financieras reguladas por la SEPS respecto a la suma de la deuda total total de las operaciones y tarjetas de crédito del sujeto en las entidades financieras del SCE en los últimos 3 meses anteriores al punto de observación.

Condición	r_deuda_total_scssce_3M
Si $deuda\_total\_sce\_3M = 0$	0
Si $deuda\_total\_sce\_3M > 0$	$\frac{deuda\_total\_sc\_op\_3M + deuda\_total\_sc\_tc\_3M}{deuda\_total\_sce\_3M}$

donde:

- *deuda_total_sc_op_3M:* Suma de la deuda total de las operaciones del sujeto en las instituciones financieras reguladas por la SEPS en los últimos 3 meses anteriores al punto de observación.
- *deuda_total_sc_tc_3M:* Suma de la deuda total de las operaciones de tarjeta de crédito del sujeto en las instituciones financieras reguladas por la SEPS en los últimos 3 meses anteriores al punto de observación.
- *deuda_total_sce_3M:* Suma de la deuda total en operaciones que registra el sujeto en las entidades financieras del SCE.

11. **r_prom_ven_sbssprom_deuda_total_sbs_tc_36M:** Razón del promedio de la deuda vencida respecto al promedio de la deuda total en tarjetas de crédito que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 36 meses anteriores al punto de observación.

Condición	r_prom_ven_sbs y prom_deuda_total_sbs_tc_36M
Si $prom\_deuda\_total\_sbs\_tc\_36M = 0$	0
Si $prom\_deuda\_total\_sbs\_tc\_36M > 0$	$\frac{prom\_ven\_sbs\_tc\_36M}{prom\_deuda\_total\_sbs\_tc\_36M}$

donde:

- *prom_ven_sbs_tc_36M:* Promedio de la deuda vencida en tarjetas de crédito que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 36 meses anteriores al punto de observación.
- *prom_deuda_total_sbs_tc_36M:* Promedio de la deuda total en tarjetas de crédito que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 36 meses anteriores al punto de observación.

12. **prbb_ntc_apert_sce_24M:** Probabilidad de bueno del número de tarjetas vigentes aperturadas por el sujeto en los últimos 24 meses anteriores al punto de observación en el SCE.

Condición	prbb_ntc_apert_sce_24M
Si $ntc\_apert\_sce\_24M \leq 0$	0.88525
Si $ntc\_apert\_sce\_24M \leq 1$	0.85284
Caso contrario	0.81080

La construcción de la variable *ntc_apert_sce_24M* se realizó mediante el árbol de decisión de la Figura 3.9, en la cual observamos que la probabilidad de bueno disminuye con el número de tarjetas de crédito aperturadas en las entidades pertenecientes al Sistema Crediticio Ecuatoriano.

En el Nodo 5 del árbol de decisión se encuentran los sujetos que no han aperturado tarjetas de crédito en los últimos 24 meses anteriores al punto de observación y tienen una probabilidad de bueno de 0.88525, seguido, el Nodo 6 comprende los sujetos que aperturaron una tarjeta de crédito en los últimos 24 meses anteriores al punto de observación y tienen una probabilidad de bueno de 0.85284, por último, los sujetos que aperturaron más de dos tarjetas de crédito en los últimos 24 meses anteriores al punto de observación (Nodo 7) tienen una probabilidad de bueno de 0.81080.

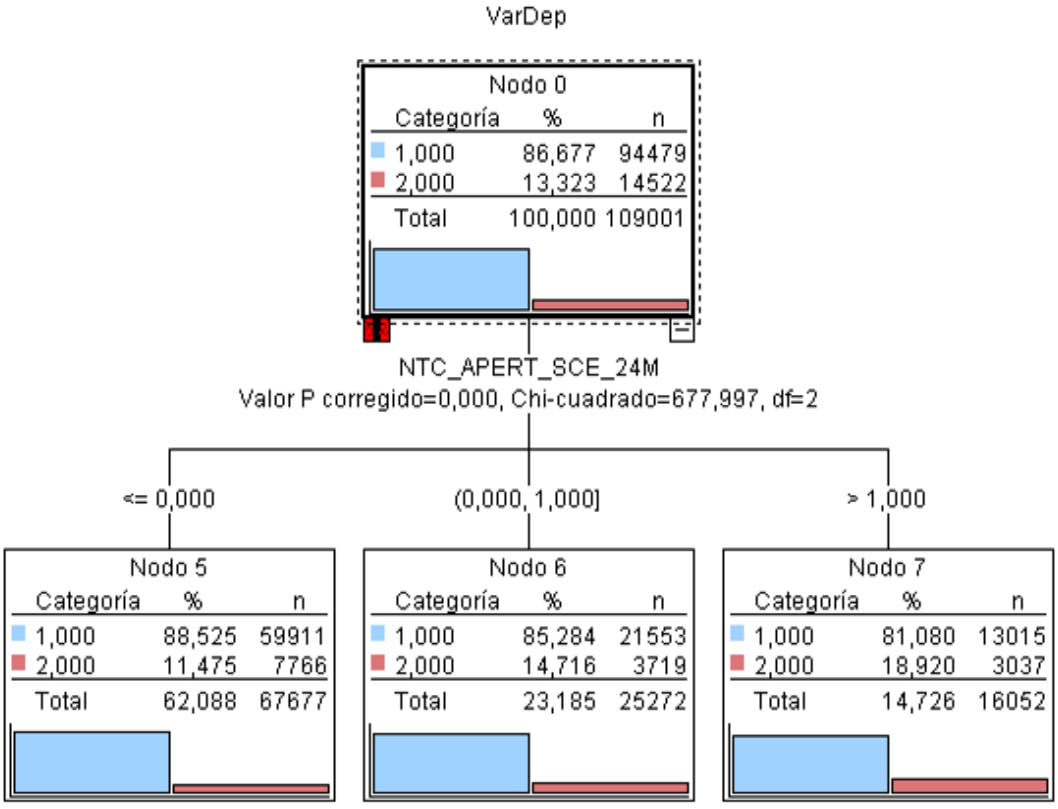


Figura 3.9: Árbol de decisión para la variable explicativa ntc_apert_sce_24M

13. **prbb_deuda_total_sce_3M:** Probabilidad de bueno de la suma de la deuda total vigente que registra el sujeto en las entidades financieras del SCE en los últimos 3 meses anteriores al punto de observación.

Condición	prbb_deuda_total_sce_3M
Si $deuda\_total\_sce\_3M \leq 44259.160$	0.87575
Si $deuda\_total\_sce\_3M \leq 72503.060$	0.85110
Caso contrario	0.81064

El árbol de decisión de la Figura 3.10 se utiliza para la construcción de la probabilidad de bueno de la variable $deuda\_total\_sce\_3M$, donde se observa que ha medida que el valor de la deuda total incrementa la probabilidad de bueno disminuye.

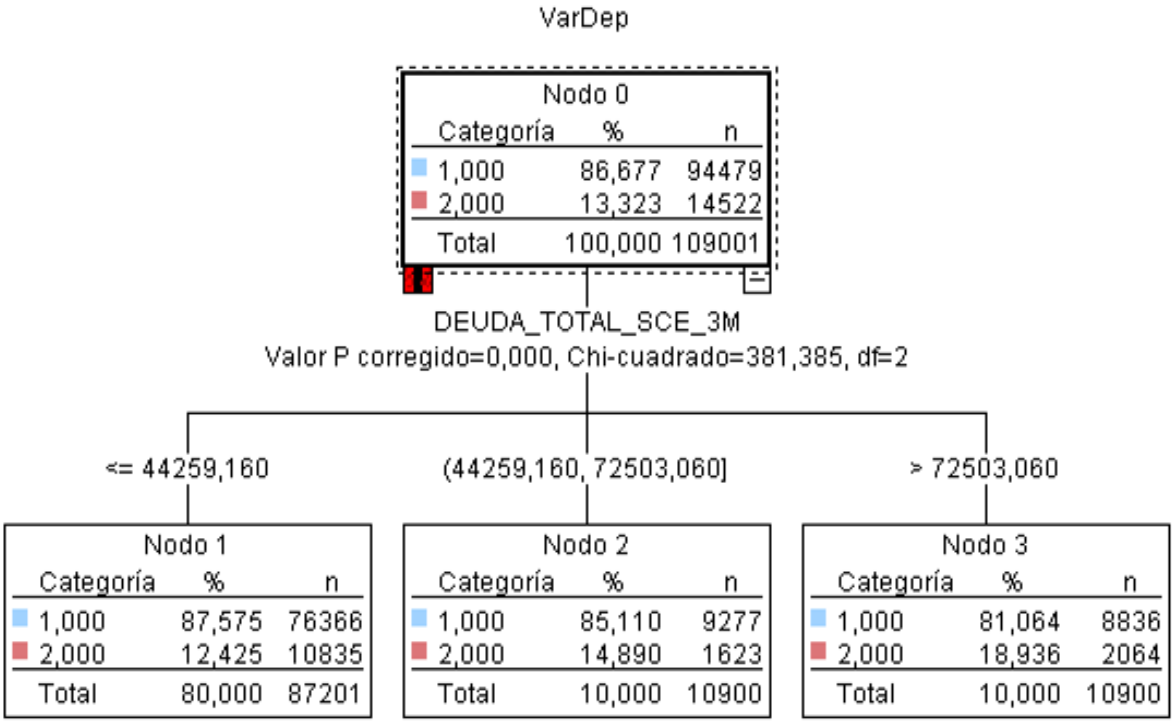


Figura 3.10: Árbol de decisión para la variable explicativa deuda_total_sce_3M

14. **r_nope_apert_sbssce_op_24M:** Razón del número de operaciones vigentes aperturadas por el sujeto en las entidades financieras reguladas por la SB respecto al número de operaciones vigentes aperturadas por el sujeto en las entidades del SCE en los últimos 24 meses anteriores al punto de observación.

Condición	r_nope_apert_sbssce_op_24M
Si $nope\_apert\_sce\_op\_24M = 0$	0
Si $nope\_apert\_sce\_op\_24M > 0$	$\frac{nope\_apert\_sbs\_op\_24M}{nope\_apert\_sce\_op\_24M}$

donde:

- *nope_apert_sbs_op_24M:* Número de operaciones vigentes aperturadas por el sujeto en las entidades financieras reguladas por la SB en los últimos 24 meses anteriores al punto de observación.

- *nope_apert_sce_op_24M:* Número de operaciones vigentes aperturadas por el sujeto en las entidades financieras en el SCE en los últimos 24 meses anteriores al punto de observación.

3.8.2 Selección de variables del segmento Dirty

Las variables explicativas con mayor poder discriminativo entre clientes buenos y clientes malos y son empleadas en la construcción de los modelos clasificadores de primer nivel para el modelo de ensamble en el segmento Dirty se presentan en la Tabla 3.13.

	Segmento Dirty				
N°	**Variables explicativas**	**Modelos**			
		RF	**GBM**	**NB**	**RL**
1	*prbb_estadocivil*	X	X	X	
2	*antiguedad_sce*	X	X	X	X
3	*prbb_nope_apert_sbs_op_12M*	X	X	X	X
4	*d_nope_ndi_op_12M*	X	X	X	
5	*r_nope_venc_op_3s12M*	X	X	X	
6	*r_mvalven_sicomsdeuda_total_sbs_3M*	X	X	X	X
7	*r_nope_apert_sc_op_24s36M*	X	X	X	
8	*r_mvalven_sbssdeuda_total_sbs_6M*	X	X	X	
9	*r_prom_ven_scsprom_deuda_total_sbs_36M*	X	X	X	
10	*PorcCuentaCoop*	X	X	X	X
11	*prbb_mvalven_sbs_12M*	X	X	X	
12	*r_prom_max_dven_n_op_3s12M*				X
13	*r_nope_xven_op_3s12M*				X
14	*r_prom_ven_sbssprom_deuda_total_sbs_6M*				X
15	*prbb_nent_ven_sc_12M*				X
16	*ln_prom_ven_sbs_tc_12M*				X
17	*r_nope_apert_scssce_op_12M*				X

Tabla 3.13: Variables explicativas en el segmento Dirty

Seguidamente, se realiza la descripción de cada variable explicativa que se muestra en la Tabla 3.13, adicionalmente, la primera, tercera, décimo primera y décimo quinta variable fueron construidas mediante la técnica de árboles de decisión utilizando el programa IBM SPSS Modeler.

1. **prbb_estadocivil:** Probabilidad de bueno de la variable estado civil construida mediante el árbol de decisión de la Figura 3.11, donde se observa que el porcentaje de sujetos buenos en el Nodo 1 (17,88 %) es mayor que el porcentaje de sujetos buenos en el Nodo 2 (15,06 %) y Nodo 3 (9,00 %).

Condición	prbb_estadocivil
Si $estadocivil = Casado$ o $estadocivil = UnionHecho$	0.17882
Si $estadocivil = Divorciado$ o $estadocivil = Viudo$	0.15057
Caso Contrario	0.09001

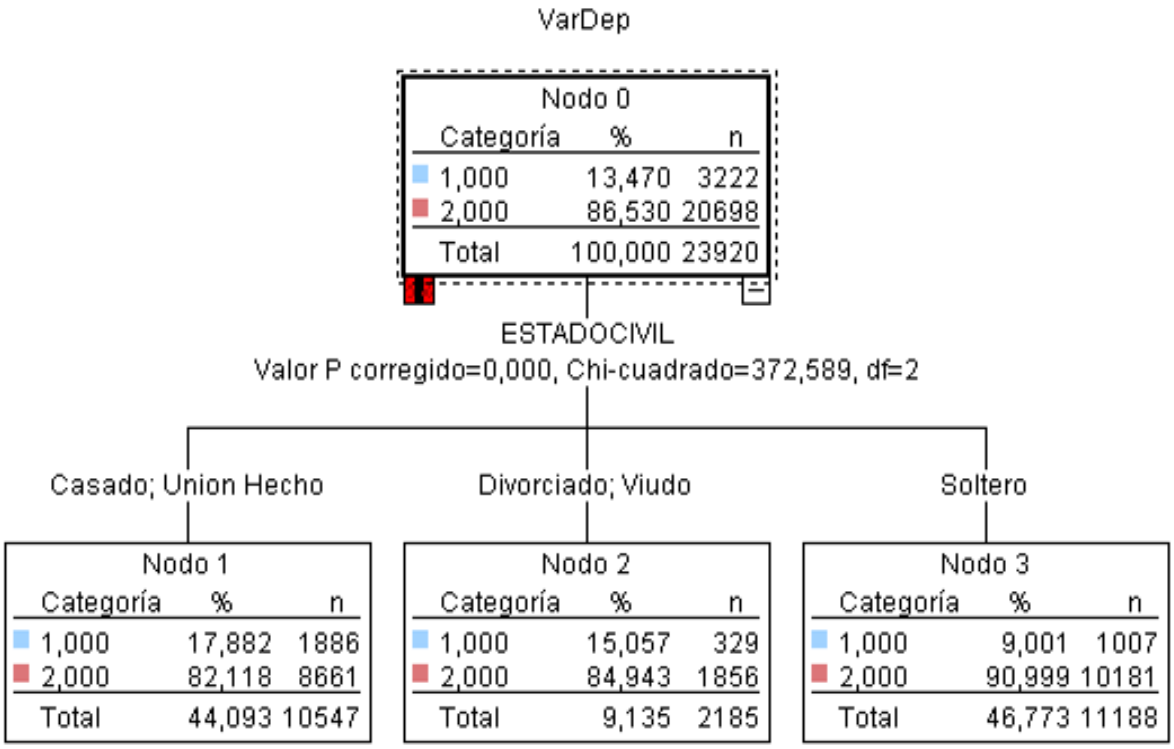


Figura 3.11: Árbol de decisión para la variable explicativa *estadocivil*

2. **antiguedad_sce:** Número de meses transcurridos desde que el sujeto apertura su primera operación en las entidades financieras pertenecientes al Sistema Comercial, Sistema Regulado por ls SB y Sector de Cooperativas (SEPS).

$$\begin{aligned} antiguedad\_sce = & mx(antiguedad\_op\_sbs, antiguedad\_tc\_sbs, \\ & antiguedad\_op\_sc, antiguedad\_tc\_sc, \\ & antiguedad\_op\_sicom, antiguedad\_tc\_sicom)/12 \end{aligned}$$

donde:

- *antiguedad_op_sbs:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en alguna entidad financiera del Sistema Regulado por la SB.
- *antiguedad_tc_sbs:* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en alguna entidad financiera del Sistema

Regulado por la SB.

- *antiguedad_op_sc:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en una entidad financiera perteneciente al Sector de Cooperativas.
- *antiguedad_tc_sc:* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en una entidad financiera perteneciente al Sector de Cooperativas.
- *antiguedad_op_sicom:* Número de meses transcurridos desde que el sujeto aperturó su primera operación en una entidad financiera perteneciente al Sistema Comercial.
- *antiguedad_tc_sicom:* Número de meses transcurridos desde que el sujeto aperturó su primera tarjeta de crédito en una entidad financiera perteneciente al Sistema Comercial.

3. **prbb_nope_apert_sbs_op_12M:** Probabilidad de bueno del número de operaciones vigentes aperturadas en las entidades financieras reguladas por la SB en los últimos 12 meses anteriores al punto de observación. En el árbol de decisión de la Figura 3.12 podemos observar que la probabilidad de bueno aumenta con el incremento de número de operaciones aperturadas.

Condición	**prbb_nope_apert_sbs_op_12M**
Si $nope\_apert\_sbs\_op\_12M = 0$	0.11233
Si $nope\_apert\_sbs\_op\_12M \leq 1$	0.19551
Caso Contrario	0.2202

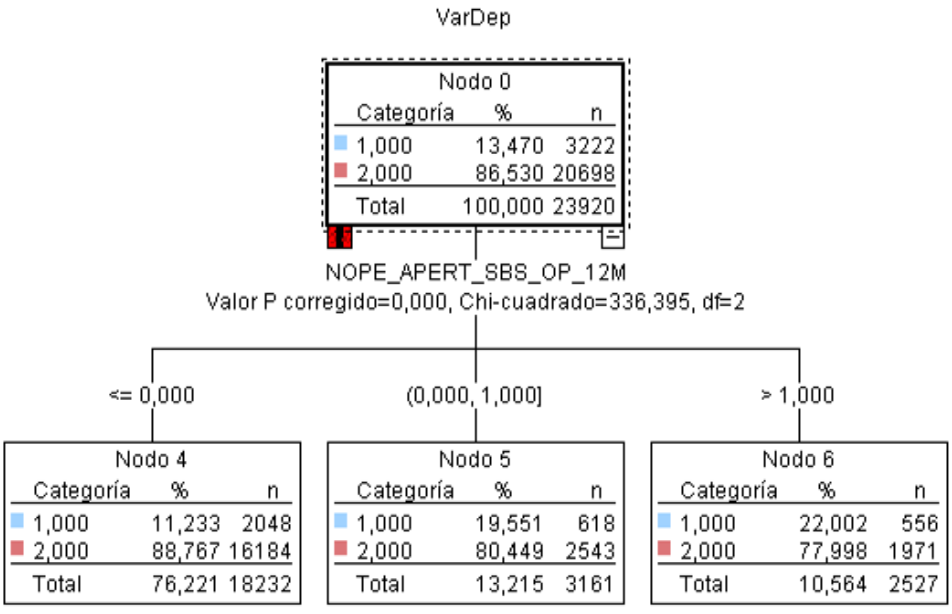


Figura 3.12: Árbol de decisión para la variable explicativa *nope_apert_sbs_op_12M*

4. **d_nope_ndi_op_12M:** Variable binaria que toma el valor de 1 si más de una operación que registra el sujeto no devenga intereses en los últimos 12 meses anteriores al punto de observación y 0 caso contrario.

$$d\_nope\_ndi\_op\_12M = \begin{cases} 1 & \text{Si } nope\_ndi\_op\_12M > 0 \\ 0 & \text{Si } nope\_ndi\_op\_12M = 0 \end{cases}$$

5. **r_nope_venc_op_3s12M:** Ratio del número de operaciones vigentes en deuda vencida en los últimos 3 y 12 meses anteriores al punto de observación.

Condición	**r_nope_venc_op_3s12M**
Si $nope\_venc\_op\_12M = 0$	0
Si $nope\_venc\_op\_3s12M \geq 0$	$\frac{nope\_venc\_op\_3M}{nope\_venc\_op\_12M}$

donde:

- *nope_venc_op_3M:* Número de operaciones vigentes en deuda vencida en los últimos 3 meses anteriores al punto de observación.
- *nope_venc_op_12M:* número de operaciones vigentes en deuda vencida en los últimos 12 meses anteriores al punto de observación.

6. **r_mvalven_sicomsdeuda_total_sbs_3M:** Razón del mayor valor vencido en el Sistema Comercial respecto a la suma de la deuda total en las entidades financieras reguladas por la SB que registra el sujeto en los últimos 3 meses anteriores al punto de observación.

Condición	**r_mvalven_sicom y deuda_total_sbs_3M**
Si $mvalven\_sicom_3M = 0$ y $deuda\_total\_sbs\_3M = 0$	0
Si $mvalven\_sicom\_3M > 0$ y $deuda\_total\_sbs\_3M = 0$	1
Si $deuda\_total\_sbs\_3M > 0$	$\frac{mvalven\_sicom\_3M}{deuda\_total\_sbs\_3M}$
Caso Contrario	0

donde:

- *mvalven_sicom_3M:* Mayor valor vencido que registra el sujeto en el Sistema Comercial en los últimos 3 meses anteriores al punto de observación.
- *deuda_total_sbs_3M:* Suma de la deuda total que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 3 meses anteriores al punto de observación.

7. **r_nope_apert_sc_op_24s36M:** Ratio del número de operaciones vigentes aperturadas por el sujeto en las entidades financieras reguladas por la SEPS en los últimos 24 y 36 meses anteriores al punto de observación.

Condición	r_nope_apert_sc_op_24s36M
Si $nope\_apert\_sc\_op\_36M = 0$	0
Si $nope\_apert\_sc\_op\_36M > 0$	$\frac{nope\_apert\_sc\_op\_24M}{nope\_apert\_sc\_op\_36M}$

donde:

- *nope_apert_sc_op_24M:* Número de operaciones vigentes aperturadas por el sujeto en las entidades financieras reguladas por la SEPS en los últimos 24 meses anteriores al punto de observación.
- *nope_apert_sc_op_36M:* Número de operaciones vigentes aperturadas por el sujeto en las entidades financieras reguladas por la SEPS en los últimos 36 meses anteriores al punto de observación.

8. **r_mvalven_sbssdeuda_total_sbs_6M:** Razón del mayor valor vencido respecto a la suma de la deuda total en las entidades financieras reguladas por la SBS que registra el sujeto en los últimos 6 meses anteriores al punto de observación.

Condición	r_mvalven_sbssdeuda_total_sbs_6M
Si $deuda\_total\_sbs\_6M = 0$	0
Si $deuda\_total\_sbs\_6M > 0$	$\frac{mvalven\_sbs\_sbs\_6M}{deuda\_total\_sbs\_6M}$

donde:

- *mvalven_sbs_sbs_6M:* Mayor valor vencido que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 6 meses anteriores al punto de observación.
- *deuda_total_sbs_6M:* Suma de la deuda total en las entidades financieras reguladas por la SB que registra el sujeto en los últimos 6 meses anteriores al punto de observación.

9. **r_prom_ven_scsprom_deuda_total_sbs_36M:** Razón del promedio de la deuda vencida en las entidades financieras reguladas por la SEPS respecto al promedio de la deuda total en las entidades financieras reguladas por la SB que registra el sujeto en los últimos 36 meses anteriores al punto de observación.

Condición	r_prom_ven_sc_36M y prom_deuda_total_sbs_36M
Si $prom\_deuda\_total\_sbs\_36M = 0$ y $prom\_ven\_sc\_36M = 0$	0
Si $prom\_deuda\_total\_sbs\_36M = 0$ y $prom\_ven\_sc\_36M > 0$	1
Si $prom\_deuda\_total\_sbs\_36M > 0$	$\frac{prom\_ven\_sc\_36M}{prom\_deuda\_total\_sbs\_36M}$

donde:

- *prom_ven_sc_36M:* Promedio de la deuda vencida de las operaciones que registra el sujeto en las entidades financieras pertenecientes al sistema regulado por la SEPS en los últimos 36 meses anteriores al punto de observación.
- *prom_deuda_total_sbs_36M:* Promedio de la deuda total en las entidades financieras reguladas por la SB que registra el sujeto en los últimos 36 meses anteriores al punto de observación.

10. **PorcCuentaCoop:** Porcentaje de sujetos que poseen una cuenta vigente en Cooperativas a nivel parroquial. Esta variable forma parte del conjunto de variables de indicadores agregados a nivel parroquial (INEC) y es asignada a cada sujeto respecto a su información de de provincia, cantón y parroquia levantada en el punto de observación.

11. **prbb_mvalven_sbs_12M:** Probabilidad de bueno del mayor valor vencido que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 12 meses anteriores al punto de observación. Comenzamos la construcción de la variable explicativa de la siguiente manera:

$$mvalven\_sbs\_12M = mvalven\_sbs\_op\_12M + mvalven\_sbs\_tc\_12M$$

donde:

- *mvalven_sbs_op_12M:* Mayor valor vencido de las operaciones de crédito que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 12 meses anteriores al punto de observación.
- *mvalven_sbs_tc_12M:* Mayor valor vencido de las tarjetas de crédito que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 12 meses anteriores al punto de observación.

Luego, el árbol de decisión de la Figura 3.13 se utiliza para la construcción de la probabilidad de bueno de la variable *mvalven_sbs_12M* , identificando que a un incremento del valor de deuda vencida la probabilidad de bueno disminuye.

Condición	prbb_mvalven_sbs_12M
Si $mvalven\_sbs\_12M \leq 357.49$	0.16639
Si $mvalven\_sbs\_12M \leq 900.90$	0.11079
Si $mvalven\_sbs\_12M \leq 1774.35$	0.08612
Caso contrario	0.07588

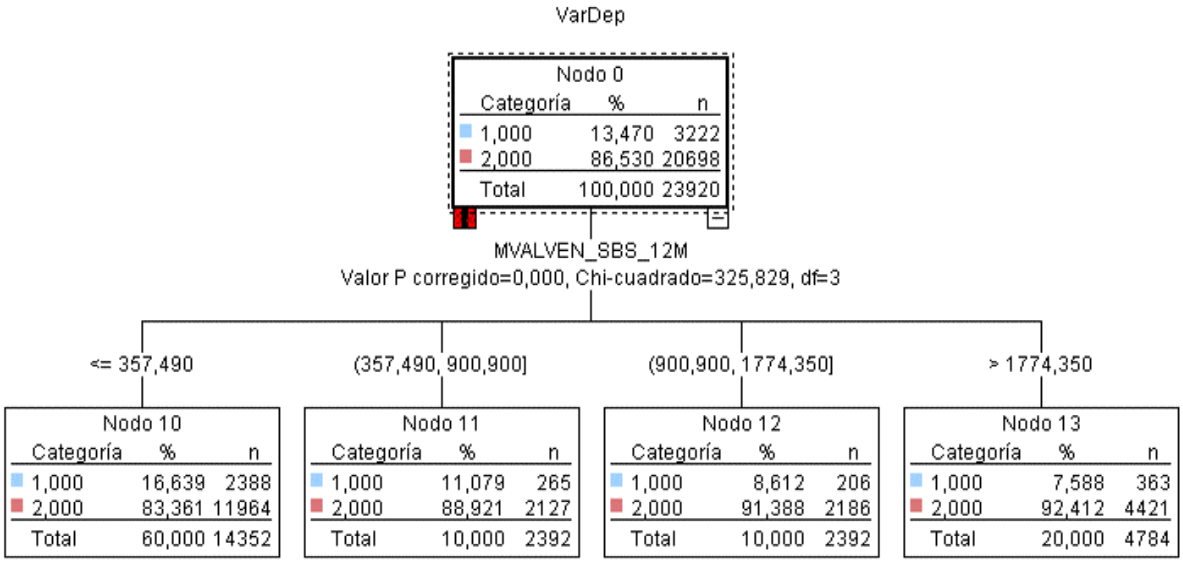


Figura 3.13: Árbol de decisión para la variable explicativa *mvalven_sbs_12M*

12. **r_prom_max_dven_n_op_3s12M:** Ratio del promedio del máximo número de días vencidos de las operaciones vigentes de consumo que registra el sujeto en los últimos 3 y 12 meses anteriores al punto de observación.

Condición	r_prom_max_dven_n_op_3s12M
Si $prom\_max\_dven\_n\_op\_12M = 0$	0
Si $prom\_max\_dven\_n\_op\_12M \geq 0$	$\frac{prom\_max\_dven\_n\_op\_3M}{prom\_max\_dven\_n\_op\_12M}$

donde:

- *prom_max_dven_n_op_3M:* Promedio del máximo número de días vencidos de las operaciones vigentes de consumo que registra el sujeto en los últimos 3 meses anteriores al punto de observación.
- *prom_max_dven_n_op_12M:* Promedio del máximo número de días vencidos de las operaciones vigentes de consumo que registra el sujeto en los últimos 12 meses anteriores al punto de observación.

13. **r_nope_xven_op_3s12M:** Ratio del número de operaciones vigentes por vencer del sujeto en los últimos 3 y 12 meses anteriores al punto de observación.

Condición	r_nope_xven_op_3s12M
Si $nope\_xven\_op\_12M = 0$	0
Si $nope\_xven\_op\_12M \geq 0$	$\frac{nope\_xven\_op\_3M}{nope\_xven\_op\_12M}$

donde:

- *nope_xven_op_3M:* Número de operaciones vigentes por vencer del sujeto en los últimos 3 meses anteriores al punto de observación.
- *nope_xven_op_12M:* Número de operaciones vigentes por vencer del sujeto en los últimos 12 meses anteriores al punto de observación.

14. **r_prom_ven_sbssprom_deuda_total_sbs_6M:** Razón del promedio de la deuda vencida respecto al promedio de la deuda total en las entidades financieras reguladas por la SB que registra el sujeto en los últimos 6 meses anteriores al punto de observación.

Condición	r_prom_ven_sbs_6M y prom_deuda_total_sbs_6M
Si $prom\_deuda\_total\_sbs\_6M = 0$	0
Si $prom\_deuda\_total\_sbs\_6M \geq 0$	$\frac{prom\_ven\_sbs\_6M}{prom\_deuda\_total\_sbs\_6M}$

donde:

- *prom_ven_sbs_6M:* Promedio de la deuda vencida que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 6 meses anteriores al punto de observación.
- *prom_deuda_total_sbs_6M:* Promedio de la deuda total que registra el sujeto en las entidades financieras reguladas por la SB en los últimos 6 meses anteriores al punto de observación.

15. **prbb_nent_ven_sc_12M:** Probabilidad de bueno del número de entidades reguladas por la SEPS en las cuales el sujeto registra vencidos en los últimos 12 meses anteriores al punto de observación. Comenzamos la construcción de la variable explicativa de la siguiente manera:

$$nent\_ven\_sc\_12M = nent\_ven\_sc\_op\_12M + nent\_ven\_sc\_tc\_12M$$

- *nent_ven_sc_op_12M:* Número de entidades reguladas por la SEPS en las cuales el sujeto registra vencidos en las operaciones en los últimos 12 meses anteriores al punto de observación.
- *nent_ven_sc_tc_12M:* Número de entidades reguladas por la SEPS en las cuales el sujeto registra vencidos en las tarjetas de crédito en los últimos 12 meses anteriores al punto de observación.

Luego, el árbol de decisión de la Figura 3.14 se utiliza para la construcción de la probabilidad de bueno de la variable *nent_ven_sc_12M*.

Condición	prbb_nent_ven_sc_12M
Si $nent\_ven\_sc\_12M = 0$	0.11204
Si $nent\_ven\_sc\_12M \geq 0$	0.21395

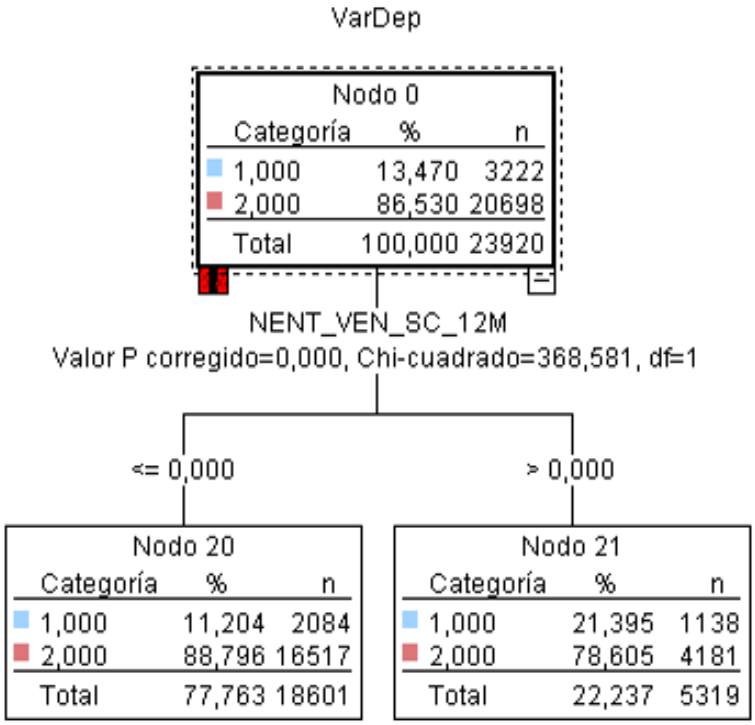


Figura 3.14: Árbol de decisión para la variable explicativa *nent_ven_sc_12M*

16. **ln_prom_ven_sbs_tc_12M:** Logaritmo natural del promedio de la deuda vencida de las tarjetas de crédito que registra el sujeto en los últimos 12 meses anteriores al punto de observación.

Condición	ln_prom_ven_sbs_tc_12M
Si $prom\_ven\_sbs\_tc\_12M \leq 1$	0
Si $prom\_ven\_sbs\_tc\_12M > 1$	log(prom_ven_sbs_tc_12M)

17. **r_nope_apert_scssce_op_12M:** Razón del número de operaciones vigentes aperturadas en las entidades reguladas por la SEPS respecto al número de operaciones vigentes aperturadas en el SCE en los últimos 12 meses anteriores al punto de observación.

Condición	r_nope_apert_scssce_op_12M
Si $nope\_apert\_sce\_op\_12M = 0$	0
Si $nope\_apert\_sce\_op\_12M \geq 0$	$\frac{nope\_apert\_sc\_op\_12M}{nope\_apert\_sce\_op\_12M}$

donde:

- *nope_apert_sc_op_12M:* Número de operaciones vigentes aperturadas por el sujeto en las entidades reguladas por la SEPS en los últimos 12 meses anteriores al punto de observación.
- *nope_apert_sce_op_12M:* Número de operaciones vigentes aperturadas por el sujeto en el SCE en los últimos 12 meses anteriores al punto de observación.

3.8.3 Validación cruzada

La validación cruzada[7] se realiza con el fin de validar internamente cada uno de los modelos que formarán el ensamble sin comprometer una división adicional para este procedimiento y comprobar que los resultados obtenidos son independientes de la partición. A continuación presentamos algunas características que se deben considerar en este proceso:

- El número de iteraciones para generar los datos de primer nivel para el ensamble debe ser el mismo en los modelos individuales; las buenas prácticas recomiendan alrededor de 5 a 10 iteraciones.
- El metaclasificador se entrena con las predicciones obtenidas por validación cruzada de todos los modelos individuales.

Para nuestro estudio realizamos una validación cruzada de 5 iteraciones en cada uno de los modelos individuales recopilando todos los valores predichos con validación cruzada.

3.8.4 Entrenamiento de algoritmos base

El apartado describe el procedimiento de entrenamiento de los algoritmos de primer nivel: Random Forest, Gradient Boosting Machine, Naïve Bayes y Regresión Logística, para ello, detallaremos el conjunto de valores para los hiperparámetros más importantes de los algoritmos empleados en la búsqueda de cuadrícula cartesiana, entrenaremos un algoritmo para cada combinación de valores de hiperparámetros y seleccionaremos el modelo con mayor poder de predicción respaldados de una métrica de rendimiento.

1. **Hiperparámetros de algoritmos individuales**
   Se presenta una lista de los valores utilizados para la búsqueda de cuadrícula carte-

[7]Para mayor detalle de la forma que opera la validación cruzada de H2O se puede encontrar en el siguiente enlace: `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/cross-validation.html`

siana de los principales hiperparámetros de cada modelo individual en los segmentos Clean y Dirty.

Los rangos de los valores de los hiperparámetros para cada uno de los algoritmos empleados en el presente estudio empírico son parte de los valores potenciales que considera el algoritmo de aprendizaje automático AutoML mediante la función *h2o.automl* que pertenece al paquete **H2O** del software estadístico **R**[8].

- **Random Forest** La Tabla 3.14 indica los valores de los hiperparámetros de Random Forest que se exploran al realizar la búsqueda de cuadrícula cartesiana.

Parámetro	**Valores de búsqueda**
mtries	$\lfloor \#predictores \times \{0.2, 0.4, 0.6, 0.8\} \rfloor$
ntrees	$\lfloor \#predictorees \times 10 \rfloor$
max_ depth	{5, 10, 15, 20}
min_ rows	$\lfloor \#observaciones \times \{0.01, 0.03, 0.05, 0.08, 0.1\} \rfloor$

Tabla 3.14: Valores de búsqueda de los hiperparámetros del algoritmo *RF*

  donde:
  - *mtries:* Número de predictores aleatorios seleccionados para cada nivel.
  - *ntrees:* Número de árboles ejecutados en el algoritmo, donde el valor predeterminado es 50.
  - *max_ depth:* Valor de la máxima profundidad a la que se construirá cada árbol, el valor predeterminado para el algoritmo es 20.
  - *min_ rows:* Número mínimo de observaciones de cada hoja para realizar una división.

- **Gradient Boosting Machine** La Tabla 3.15 indica los valores de los hiperparámetros de Gradient Boosting Machine que se exploran al realizar la búsqueda de cuadrícula cartesiana.
  donde:
  - *learn_ rate:* Valor de la tasa a la que el algoritmo aprende cuando construye un modelo, el rango permitido de valores es de 0.0 a 1.0.
  - *max_ depth:* Valor de la máxima profundidad a la que se construirá cada árbol, el valor predeterminado para el algoritmo es 5.

[8]Un mayor detalle de los valores de los hiperparámetros se puede encontrar en el siguiente enlace: `https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html`

Parámetro	Valores de búsqueda
learn_rate	0.1
max_depth	{5, 10, 15, 20}
ntrees	1000
min_rows	$\lfloor \#observaciones \times \{0.01, 0.03, 0.05, 0.08, 0.1\} \rfloor$

Tabla 3.15: Valores de búsqueda de los hiperparámetros del algoritmo *GBM*

  - *ntrees:* Número de árboles ejecutados en el algoritmo para la construcción del modelo.
  - *min_rows:* Número mínimo de observaciones de cada hoja para realizar una división.

- **Naïve Bayes** La Tabla 3.16 indica los valores de los hiperparámetros de Naïve Bayes que se exploran al realizar la búsqueda de cuadrícula cartesiana.

Parámetro	Valores de búsqueda
laplace	{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0}
eps_prob	{0.3, 0.4, 0.5, 0.6, 0.7, 0.8}
min_prob	{0.01, 0.07, 0.08, 0.1, 0.2}

Tabla 3.16: Valores de búsqueda de los hiperparámetros del algoritmo *NB*

  donde:
  - *laplace:* Valor para el factor de suavizado de Laplace, se utiliza para establecer la probabilidad condicional de una variable explicativa del modelo.
  - *eps_prob:* Valor mayor a cero para establecer un piso en la probabilidad calculada.
  - *min_prob:* Probabilidad mínima que debemos utilizar para los registros sin suficientes datos, su valor debe ser mayor o igual a $1exp\{-10\}$.

- **Regresión Logística** La Tabla 3.17 indica los valores del hiperparámetro de la Regresión Logística que se exploran al realizar la búsqueda de cuadrícula interna, donde *alpha* determina la distribución de la regularización entre L1 y L2[9].

Parámetro	Valores de búsqueda
alpha	Sucesión de valores entre 0 y 1 con incrementos de 0.01.

Tabla 3.17: Valores de búsqueda de los hiperparámetros del algoritmo *RL*

[9]La descripción completa del hiperparámetro *alpha* esta disponible en el siguiente enlace: `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/alpha.html`

2. **Selección de los mejores modelos de primer nivel**
   Una vez que se entrenan varias cuadrículas cartesianas de modelos de primer nivel procedemos a ordenar los modelos a través de la métrica de rendimiento AUC (Área bajo la curva de ROC) para seleccionar el clasificador binario de cada tipo de algoritmo con mayor capacidad de distinguir entre verdaderos positivos y los falsos positivos. Los valores de AUC cercanos a 1 indican una alta capacidad de clasificar y los valores de AUC menores o iguales a 0.5 indican un clasificador de capacidad baja.

   Además, los segmentos Clean y Dirty disponen como metaclasificador al algoritmo *Gradient Boosting Machine* ya que tiene la mayor capacidad de predicción en base a los valores de la métricas del modelo: KS, ROC y Gini presentados en la Tabla 3.18.

	Segmento	
Métrica	**Clean**	**Dirty**
KS	37.8	55.9
ROC	75.3	85.7
GINI	50.7	71.5

Tabla 3.18: Valores de métricas del metaclasificador *GBM*

3.8.5 Entrenamiento del ensamble stacking

El objetivo del apartado es encontrar el modelo de ensamble stacking con mayor poder de predicción, para esto, entrenamos cada una de las combinaciones posibles de los clasificadores básicos, incorporando modelos que agreguen diversidad al conjunto de modelos de primer nivel como eliminando modelos base con un rendimiento de clasificación inferior a los otros algoritmos de primer nivel.

La Tabla 3.19 presenta el rendimiento de las distintas combinaciones de modelos individuales que se entrenaron para obtener varios modelos de ensamble en los segmentos Clean y Dirty en base a los valores del KS, índice GINI y la curva de ROC.

SEGMENTO DIRTY									
N	***ENSAMBLE***	***KS***	***ROC***	***GINI***	***N***	***ENSAMBLE***	***KS***	***ROC***	***GINI***
1	RF + GBM	55,2	85,4	70,8	7	RF + GBM + NB	55,3	85,5	71,0
2	RF + GLM	50,8	83,1	66,1	8	RF + GLM + NB	50,2	82,9	65,9
3	RF + NB	53,1	84,4	68,8	9	GBM + GLM + NB	53,9	84,4	68,8
4	GBM + NB	55,8	85,6	71,1	10	GBM + GLM + RF	53,3	84,2	68,5
5	GLM + NB	42,4	76,8	53,7	11	RF + GBM + GLM + NB	47,0	81,7	63,4
6	GBM + GLM	53,8	84,3	68,6					
SEGMENTO CLEAN									
N	***ENSAMBLE***	***KS***	***ROC***	***GINI***	***N***	***ENSAMBLE***	***KS***	***ROC***	***GINI***
1	RF + GBM	37,7	75,4	50,7	7	RF + GBM + NB	37,8	75,4	50,7
2	RF + GLM	34,9	73,2	46,3	8	RF + GLM + NB	34,9	73,4	46,8
3	RF + NB	34,6	73,2	46,4	9	GBM + GLM + NB	37,9	75,4	50,7
4	GBM + NB	37,8	75,3	50,7	10	GBM + GLM + RF	37,8	75,3	50,7
5	GLM + NB	32,0	71,3	42,6	11	RF + GBM + GLM + NB	35,6	73,9	47,8
6	GBM + GLM	37,8	75,3	50,7					

Tabla 3.19: Rendimiento del entrenamiento de modelos de ensamble (%)

Capítulo 4

Resultados de Modelos Estadísticos

En este capítulo presentamos varios de los resultados de entrenamiento y validación de los cuatro modelos individuales desarrollados: RF, GBM, NB y RL y del ensamble clasificador construido para el segmento Clean y el Segmento Dirty, respectivamente. Cada modelo fue construido en base a la definición de variable dependiente presentada en la Sección 3.3 y las variables explicativas de los apartados 3.8.1 y 3.8.2.

Las tablas de performance muestran los resultados de la calidad de discriminación de los modelos para cada uno de los 10 rangos distribuidos uniformemente con las siguientes notaciones:

- **Min:** Puntaje mínimo de score en cada intervalo.
- **Max:** Puntaje máximo de score en cada intervalo.
- **Int:** Número de sujetos en cada intervalo.
- **Int %:** Porcentaje de sujetos en cada intervalo.
- **Cum %:** Porcentaje acumulado de sujetos etiquetados como malos.

4.1 Resultados obtenidos para el segmento Clean

En primer lugar, la Tabla 4.1 presenta la distribución de sujetos en cada una de las categorías de la variable dependiente en la muestra de modelamiento y la muestra de validación; seguido, la Tabla 4.2 describe los valores de la importancia relativa en base a la reducción del error cuadrático medio de los predictores empleados en la construcción de los modelos Random Forest y Gradient Boosting Machine y su respectiva representación gráfica se observa en las Figuras 4.1 y 4.2.

		Modelamiento		Validación	
Categoría	Valor	Sujetos	Porcentaje	Sujetos	Porcentaje
Bueno	0	135.011	59,07 %	135.433	59,15 %
Malo	1	20.828	9,11 %	21.075	9,20 %
Indeterminado	2	28.128	12,31 %	27.739	12,11 %
Malo Observado	3	16455	7,20 %	16471	7,19 %
Sin desempeño	4	23.407	10,24 %	23.431	10,23 %
No bancarizado	5	4.715	2,06 %	4.824	2,11 %
Total	—	**228.544**	**100,00 %**	**228.973**	**100,00 %**

Tabla 4.1: Distribución de sujetos en el Segmento Clean

Variable	**Importancia relativa (GBM)**	**Importancia relativa (RF)**
r_nope_apert_sicomssce_op_36M	3120,16	28166,00
porc_uso_cupo	2988,45	27948,14
antiguedad_sce	1019,78	232,04
d_nent_ven_sce_op_36M	789,05	759,84
d_ntc_ndi_tc_36M	670,60	589,95
r_nope_apert_sce_12a24M	602,17	529,43
PromLocalesCom	543,21	483,61
prbb_deuda_total_sce_3M	502,04	122,76
prbb_nope_xven_op_12M	487,40	4459,03
prbb_ntc_apert_sce_24M	336,69	325,96
r_deuda_total_scssce_3M	315,02	263,32
r_prom_ven_sbssprom_deuda_total_sbs_tc_36M	250,28	353,01
r_mvalven_sbs_tcsmvalven_sbs_op_36M	58,60	23,55

Tabla 4.2: Importancia relativa de los predictores

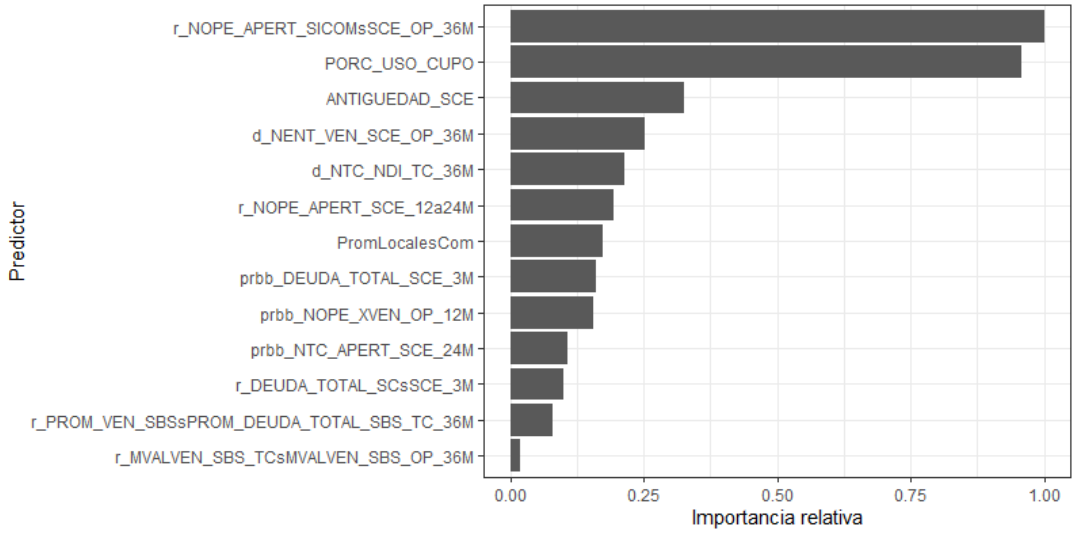


Figura 4.1: Importancia de predictores en el modelo GBM

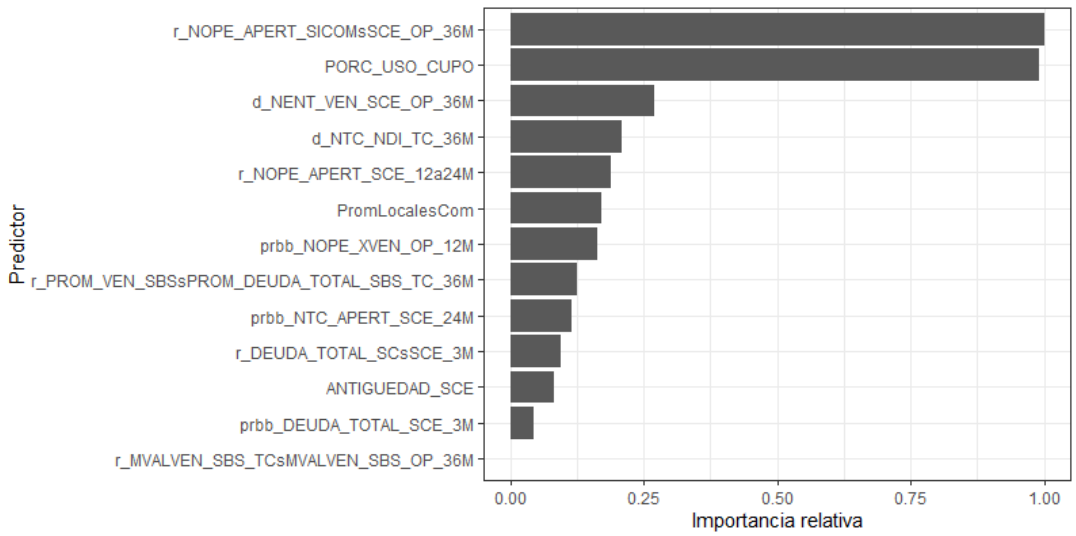


Figura 4.2: Importancia de predictores en el modelo RF

Podemos observar en la tabla performance del modelo Random Forest (ver Tabla 4.3) la distribución de los sujetos totales y sujetos con etiqueta de Malo para cada intervalo de la puntuación de score en la muestra de modelamiento.

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
931	**999**	15.584	10 %	10 %	405	2 %	2 %	2,60 %	2,6 %
921	**931**	15.584	10 %	20 %	770	4 %	6 %	4,94 %	3,8 %
912	**921**	15.584	10 %	30 %	857	4 %	10 %	5,50 %	4,3 %
897	**912**	15.584	10 %	40 %	1.137	5 %	15 %	7,30 %	5,1 %
881	**897**	15.584	10 %	50 %	1.428	7 %	22 %	9,16 %	5,9 %
865	**881**	15.583	10 %	60 %	1.717	8 %	30 %	11,02 %	6,8 %
845	**865**	15.584	10 %	70 %	2.252	11 %	41 %	14,45 %	7,9 %
815	**845**	15.584	10 %	80 %	2.678	13 %	54 %	17,18 %	9,0 %
766	**815**	15.584	10 %	90 %	3.590	17 %	71 %	23,04 %	10,6 %
1	**766**	15.584	10 %	100 %	5.994	29 %	100 %	38,46 %	13,4 %
Total		***155.839***			***20.828***				

Tabla 4.3: Tabla de performance del modelo Random Forest - Muestra de Modelamiento

El poder de discriminación del modelo Random Forest en la muestra de validación se presenta en la Tabla 4.4.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
930	**999**	22.415	10 %	10 %	602	2 %	2 %	2,69 %	2,7 %
919	**930**	22.415	10 %	20 %	1.087	4 %	6 %	4,85 %	3,8 %
907	**919**	22.415	10 %	30 %	1.295	5 %	11 %	5,78 %	4,4 %
890	**907**	22.415	10 %	40 %	1.675	6 %	18 %	7,47 %	5,2 %
875	**890**	22.414	10 %	50 %	2.078	8 %	26 %	9,27 %	6,0 %
854	**875**	22.415	10 %	60 %	2.325	9 %	35 %	10,37 %	6,7 %
832	**854**	22.415	10 %	70 %	2.918	11 %	46 %	13,02 %	7,6 %
804	**832**	22.415	10 %	80 %	3.476	13 %	59 %	15,51 %	8,6 %
761	**804**	22.415	10 %	90 %	3.957	15 %	74 %	17,65 %	9,6 %
1	**761**	22.415	10 %	100 %	6.741	26 %	100 %	30,07 %	11,7 %
Total		***224.149***			***26.154***				

Tabla 4.4: Tabla de performance del modelo Random Forest - Muestra de Validación

De igual manera en la Tabla 4.5 observamos los resultados del modelo obtenido mediante el algoritmo Gradient Boosting Machine, es decir, la distribución de los sujetos totales y sujetos malos para cada uno de los 10 intervalos de la puntuación de score en la muestra de modelamiento.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
963	**999**	15.584	10 %	10 %	322	2 %	2 %	2,07 %	2,1 %
950	**963**	15.584	10 %	20 %	605	3 %	4 %	3,88 %	3,0 %
939	**950**	15.584	10 %	30 %	829	4 %	8 %	5,32 %	3,8 %
925	**939**	15.584	10 %	40 %	1.093	5 %	14 %	7,01 %	4,6 %
905	**925**	15.584	10 %	50 %	1.280	6 %	20 %	8,21 %	5,3 %
881	**905**	15.583	10 %	60 %	1.626	8 %	28 %	10,43 %	6,2 %
849	**881**	15.584	10 %	70 %	2.012	10 %	37 %	12,91 %	7,1 %
803	**849**	15.584	10 %	80 %	2.630	13 %	50 %	16,88 %	8,3 %
706	**803**	15.584	10 %	90 %	3.697	18 %	68 %	23,72 %	10,0 %
1	**706**	15.584	10 %	100 %	6.734	32 %	100 %	43,21 %	13,4 %
Total		***155.839***			***20.828***				

Tabla 4.5: Tabla de performance del modelo GBM - Muestra de Modelamiento

La tabla performance del modelo Gradient Boosting Machine (ver Tabla 4.6) generado para la muestra de validación refleja la capacidad de identificar y clasificar los clientes clientes malos en cada uno de los rangos de puntuación score.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
960	**999**	22.415	10 %	10 %	531	2 %	2 %	2,37 %	2,4 %
947	**960**	22.415	10 %	20 %	958	4 %	6 %	4,27 %	3,3 %
934	**947**	22.415	10 %	30 %	1.304	5 %	10 %	5,82 %	4,2 %
915	**934**	22.415	10 %	40 %	1.578	6 %	16 %	7,04 %	4,9 %
891	**915**	22.414	10 %	50 %	1.944	7 %	24 %	8,67 %	5,6 %
862	**891**	22.415	10 %	60 %	2.404	9 %	33 %	10,72 %	6,5 %
826	**862**	22.415	10 %	70 %	2.933	11 %	44 %	13,08 %	7,4 %
779	**826**	22.415	10 %	80 %	3.497	13 %	57 %	15,60 %	8,4 %
673	**779**	22.415	10 %	90 %	4.352	16 %	73 %	19,42 %	9,7 %
1	**673**	22.415	10 %	100 %	7.252	27 %	100 %	32,35 %	11,9 %
Total		***224.149***			***26.753***				

Tabla 4.6: Tabla de performance del modelo GBM - Muestra de Validación

Así mismo, la tabla de performance del modelo basado en el algoritmo de Naïve Bayes (ver Tabla 4.7) presenta los rangos de puntuación de score y distribución de clientes para los individuos pertenecientes a la muestra de modelamiento.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
991	**999**	15.584	10 %	10 %	865	4 %	4 %	5,55 %	5,6 %
988	**991**	15.584	10 %	20 %	877	4 %	8 %	5,63 %	5,6 %
986	**988**	15.584	10 %	30 %	1.026	5 %	13 %	6,58 %	5,9 %
984	**986**	15.584	10 %	40 %	1.282	6 %	19 %	8,23 %	6,5 %
980	**984**	15.584	10 %	50 %	1.599	8 %	27 %	10,26 %	7,2 %
976	**980**	15.583	10 %	60 %	2.059	10 %	37 %	13,21 %	8,2 %
968	**976**	15.584	10 %	70 %	2.170	10 %	47 %	13,92 %	9,1 %
951	**968**	15.584	10 %	80 %	3.035	15 %	62 %	19,48 %	10,4 %
916	**951**	15.584	10 %	90 %	3.357	16 %	78 %	21,54 %	11,6 %
1	**916**	15.584	10 %	100 %	4.558	22 %	100 %	29,25 %	13,4 %
Total		***155.839***			***20.828***				

Tabla 4.7: Tabla de performance del modelo Naïve Bayes - Muestra de Modelamiento

De manera similar para la muestra de validación generamos la tabla de performance (ver Tabla 4.8) que presenta el poder discriminatorio del modelo basado en el algoritmo de Naïve Bayes para cada uno de los 10 rangos de la puntuación de score.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
991	**999**	22.415	10 %	10 %	1.266	5 %	5 %	5,65 %	5,6 %
988	**991**	22.415	10 %	20 %	1.083	4 %	9 %	4,83 %	5,2 %
986	**988**	22.415	10 %	30 %	1.464	6 %	15 %	6,53 %	5,7 %
983	**986**	22.415	10 %	40 %	1.669	7 %	22 %	7,45 %	6,1 %
980	**983**	22.414	10 %	50 %	2.132	8 %	30 %	9,51 %	6,8 %
975	**980**	22.415	10 %	60 %	2.632	10 %	41 %	11,74 %	7,6 %
968	**975**	22.415	10 %	70 %	2.690	11 %	51 %	12,00 %	8,2 %
950	**968**	22.415	10 %	80 %	3.537	14 %	66 %	15,78 %	9,2 %
915	**950**	22.415	10 %	90 %	3.914	16 %	81 %	17,46 %	10,1 %
1	**915**	22.415	10 %	100 %	4.748	19 %	100 %	21,18 %	11,2 %
Total		***224.149***			***25.135***				

Tabla 4.8: Tabla de performance del modelo Naïve Bayes - Muestra de Validación

En la Tabla 4.9 presentamos los nombres de los predictores para el modelo de Regresión Logística, dos tipos de coeficientes de regresión (parámetros estimados), importancia relativa en base a la reducción del error cuadrático medio y sistema al que pertenecen. El valor absoluto de los coeficientes estandarizados son utilizados para ordenar de forma descendente la influencia de los predictores del modelo de Regresión Logística como se observa en la Figura 4.3.

Variable	Coeficiente	Coeficiente estandarizado	Importancia relativa	Sistema
Constante	9,6116	-2,0436	—	—
r_nope_apert_sicomssce_op_36M	1,3759	0,2434	0,2434	Penaliza
porc_uso_cupo	0,9840	0,2762	0,2762	Penaliza
d_nent_ven_sce_op_36M	0,7957	0,2230	0,2230	Penaliza
antiguedad_sce	-0,0455	-0,1775	0,1775	Premia
d_ntc_ndi_tc_36M	1,0948	0,2382	0,2382	Penaliza
r_nope_apert_sce_12a24M	0,2646	0,1115	0,1115	Penaliza
r_nope_apert_sbsssce_op_24M	0,2554	0,1173	0,1173	Penaliza
PromLocalesCom	-8,8402	-0,1896	0,1896	Premia
prbb_ntc_apert_sce_24M	-6,2917	-0,1684	0,1684	Premia
prbb_nope_xven_op_12M	-6,9864	-0,2024	0,2024	Premia

Tabla 4.9: Predictores del modelo de regresión logística

Además, en la Figura 4.3 observamos que los signos de los predictores son los correctos para identificar el comportamiento de pago de un cliente y las variables con una importancia relativa mayor a 0.20 que mejor explican el problema a solucionar son:

- *porc_uso_cupo*
- *r_nope_apert_sicomssce_op_36M*
- *d_ntc_ndi_tc_36M*
- *d_nent_ven_sce_op_36M*
- *prbb_nope_xven_op_12M*

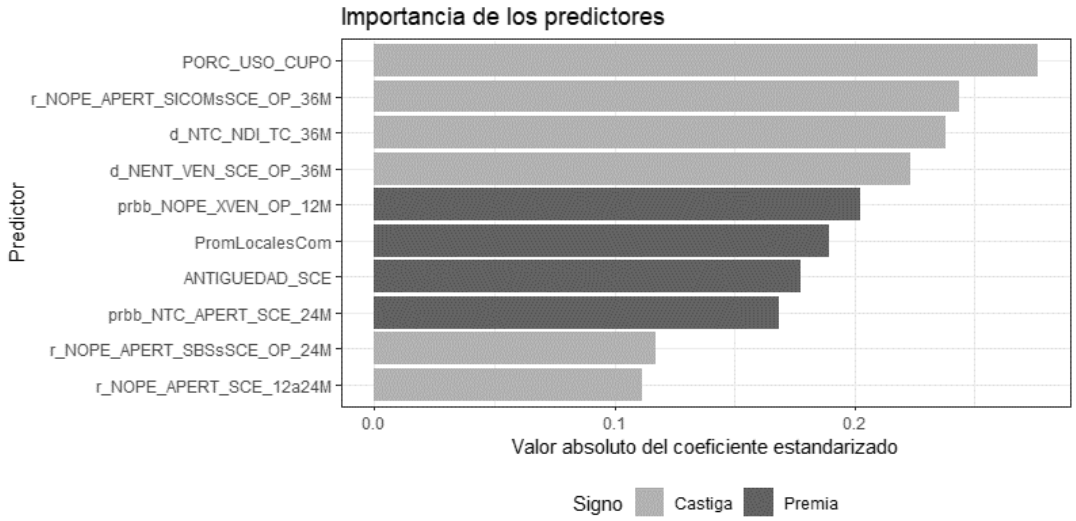


Figura 4.3: Importancia de las variables explicativas del modelo de RL

Ahora, se presenta el poder de predicción del modelo de regresión logística en la muestra de modelamiento a través de la tabla de performance (ver Tabla 4.10).

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
944	**999**	15.584	10 %	10 %	586	3 %	3 %	3,76 %	3,8 %
931	**944**	15.584	10 %	20 %	829	4 %	7 %	5,32 %	4,5 %
919	**931**	15.584	10 %	30 %	1.022	5 %	12 %	6,56 %	5,2 %
907	**919**	15.584	10 %	40 %	1.195	6 %	17 %	7,67 %	5,8 %
894	**907**	15.584	10 %	50 %	1.402	7 %	24 %	9,00 %	6,5 %
879	**894**	15.583	10 %	60 %	1.740	8 %	33 %	11,17 %	7,2 %
860	**879**	15.584	10 %	70 %	2.227	11 %	43 %	14,29 %	8,3 %
824	**860**	15.584	10 %	80 %	2.843	14 %	57 %	18,24 %	9,5 %
751	**824**	15.584	10 %	90 %	3.655	18 %	74 %	23,45 %	11,1 %
1	**751**	15.584	10 %	100 %	5.329	26 %	100 %	34,20 %	13,4 %
Total		***155.839***			***20.828***				

Tabla 4.10: Tabla de performance del modelo RL - Muestra de Modelamiento

Y de manera similar generamos la tabla de performance (ver Tabla 4.11) de 10 rangos uniformes del modelo de regresión logística para la muestra de validación.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
942	**999**	22.415	10 %	10 %	785	3 %	3 %	3,50 %	3,5 %
929	**942**	22.415	10 %	20 %	1.163	4 %	7 %	5,19 %	4,3 %
916	**929**	22.415	10 %	30 %	1.316	5 %	13 %	5,87 %	4,9 %
903	**916**	22.415	10 %	40 %	1.652	6 %	19 %	7,37 %	5,5 %
890	**903**	22.414	10 %	50 %	1.996	8 %	27 %	8,91 %	6,2 %
874	**890**	22.415	10 %	60 %	2.427	9 %	36 %	10,83 %	6,9 %
851	**874**	22.415	10 %	70 %	2.947	11 %	47 %	13,15 %	7,8 %
809	**851**	22.415	10 %	80 %	3.458	13 %	60 %	15,43 %	8,8 %
721	**809**	22.415	10 %	90 %	4.345	17 %	77 %	19,38 %	10,0 %
1	**721**	22.415	10 %	100 %	5.935	23 %	100 %	26,48 %	11,6 %
Total		***224.149***			***26.024***				

Tabla 4.11: Tabla de performance del modelo RL - Muestra de Validación

Finalmente, en el segmento Clean después de probar distintos modelos de ensamble y fundamentados en los estadísticos de Kolmogorov-Smirnov, el área bajo la curva ROC, el índice de GINI y la captación de malos, se consiguió el modelo final basado en los algoritmos de *Gradient Boosting Machine*, *Regresión Logística* y *Naïve Bayes*.

La tabla de performance (ver Tabla 4.12) presenta la distribución de los sujetos totales y sujetos malos de la muestra de modelamiento en cada uno de los 10 rangos uniformes basado en los modelos *GBM*, *GLM* y *NB*.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
958	**999**	15.584	10 %	10 %	310	1 %	1 %	1,99 %	2,0 %
949	**958**	15.584	10 %	20 %	621	3 %	4 %	3,98 %	3,0 %
937	**949**	15.584	10 %	30 %	850	4 %	9 %	5,45 %	3,8 %
922	**937**	15.584	10 %	40 %	1.052	5 %	14 %	6,75 %	4,5 %
906	**922**	15.584	10 %	50 %	1.299	6 %	20 %	8,34 %	5,3 %
885	**906**	15.583	10 %	60 %	1.623	8 %	28 %	10,42 %	6,2 %
848	**885**	15.584	10 %	70 %	1.994	10 %	37 %	12,80 %	7,1 %
810	**848**	15.584	10 %	80 %	2.640	13 %	50 %	16,94 %	8,3 %
710	**810**	15.584	10 %	90 %	3.673	18 %	68 %	23,57 %	10,0 %
1	**710**	15.584	10 %	100 %	6.766	32 %	100 %	43,42 %	13,4 %
Total		***155.839***			***20.828***				

Tabla 4.12: Tabla de performance del modelo de ensamble - Muestra de Modelamiento

Así mismo, la distribución uniforme en 10 rangos de los sujetos de la muestra de validación en la tabla de performance respecto a su puntuación de score se observa en la Tabla 4.13.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
956	**999**	22.415	10 %	10 %	527	2 %	2 %	2,35 %	2,4 %
946	**956**	22.415	10 %	20 %	975	4 %	6 %	4,35 %	3,4 %
934	**946**	22.415	10 %	30 %	1.320	5 %	11 %	5,89 %	4,2 %
915	**934**	22.415	10 %	40 %	1.558	6 %	16 %	6,95 %	4,9 %
890	**915**	22.414	10 %	50 %	1.962	7 %	24 %	8,75 %	5,7 %
865	**890**	22.415	10 %	60 %	2.380	9 %	33 %	10,62 %	6,5 %
832	**865**	22.415	10 %	70 %	2.950	11 %	44 %	13,16 %	7,4 %
787	**832**	22.415	10 %	80 %	3.489	13 %	57 %	15,57 %	8,5 %
665	**787**	22.415	10 %	90 %	4.301	16 %	73 %	19,19 %	9,6 %
1	**665**	22.415	10 %	100 %	7.319	27 %	100 %	32,65 %	11,9 %
Total		***224.149***			***26.781***				

Tabla 4.13: Tabla de performance del modelo de ensamble - Muestra de Validación

4.2 Resultados obtenidos para el segmento Dirty

Para empezar, la Tabla 4.14 muestra la distribución de sujetos en cada una de las categorías de la variable dependiente en las muestras de modelamiento y validación, respectivamente.

		Modelamiento		Validación	
Categoría	**Valor**	**Sujetos**	**Porcentaje**	**Sujetos**	**Porcentaje**
Bueno	0	4.627	8,52 %	4.592	8,52 %
Malo	1	29.684	54,66 %	29.199	54,20 %
Indeterminado	2	3.423	6,30 %	3.271	6,07 %
Malo Observado	3	12.020	22,14 %	12.325	22,88 %
Sin desempeño	4	4.548	8,38 %	4.487	8,33 %
No Bancarizado	5	0	0,00 %	0	0,00 %
Total	—-	**54.302**	**100,00 %**	**53.874**	**100,00 %**

Tabla 4.14: Distribución de sujetos en el Segmento Dirty

En la Tabla 4.15 presentamos los predictores conjuntamente con la importancia relativa en base a la reducción del error cuadrático medio de cada uno de ellos para los modelos generados por el algoritmo de Gradient Boosting Machine y Random Forest respectivamente.

Además, para una interpretación más clara, en la Figura 4.4 se observa la importancia de las variables en el modelo construido por el algoritmo Gradient Boosting Machine y en la Figura 4.5 presentamos la importancia de las variables en el modelo generado por el algoritmo Random Forest. Así, los predictores con mayor influencia en los dos modelos son las mismas: *prbb_mvalven_sbs_12M* y *r_nope_venc_op_3s12M*, es decir, son las variables que mayor aportan en el poder discriminatorio del modelo.

Variables	Importancia relativa (GBM)	Importancia relativa (RF)
prbb_mvalven_sbs_12M	2066,17	24960,62
r_nope_venc_op_3s12M	1950,81	36427,52
d_nope_ndi_op_12M	338,58	4943,78
antiguedad_sce	257,09	710,38
r_nope_apert_sc_op_24s36M	203,74	4221,92
prbb_nope_apert_sbs_op_12M	140,15	1642,46
PorcCuentaCoop	111,66	783,35
prbb_estadocivil	53,47	597,58
r_mvalven_sbssdeuda_total_sbs_6M	40,91	560,00
r_mvalven_sicomsdeuda_total_sbs_3M	5,99	30,430
r_prom_ven_scsprom_deuda_total_sbs_36M	3,34	0,32

Tabla 4.15: Importancia relativa de los predictores

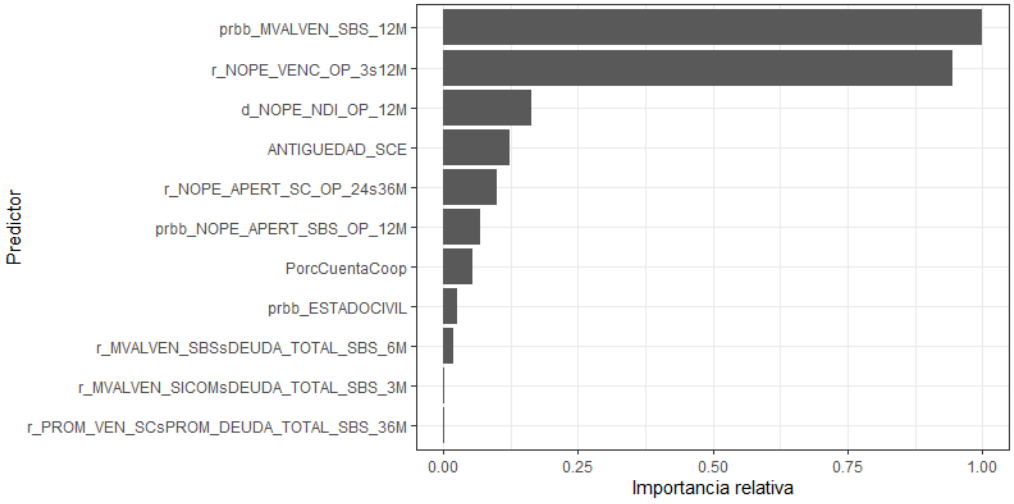


Figura 4.4: Importancia de las variables explicativas del modelo GBM

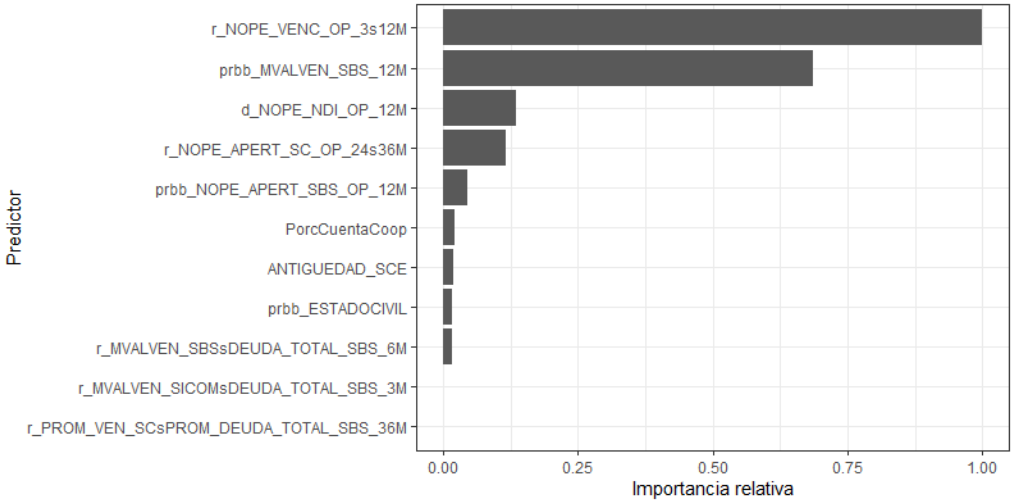


Figura 4.5: Importancia de las variables explicativas del modelo RF

Ahora, en la Tabla 4.16 observamos los resultados de clasificación del modelo aplicando el algoritmo Random Forest para los datos de la muestra de modelamiento.

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
449	**999**	3.431	10 %	10 %	1.607	5 %	5 %	46,84 %	46,8 %
175	**449**	3.431	10 %	20 %	2.280	8 %	13 %	66,45 %	56,6 %
130	**174**	3.431	10 %	30 %	2.871	10 %	23 %	83,68 %	65,7 %
103	**130**	3.431	10 %	40 %	3.062	10 %	33 %	89,25 %	71,6 %
78	**103**	3.432	10 %	50 %	3.136	11 %	44 %	91,38 %	75,5 %
55	**78**	3.431	10 %	60 %	3.232	11 %	55 %	94,20 %	78,6 %
35	**55**	3.431	10 %	70 %	3.294	11 %	66 %	96,01 %	81,1 %
13	**35**	3.431	10 %	80 %	3.358	11 %	77 %	97,87 %	83,2 %
5	**13**	3.431	10 %	90 %	3.414	12 %	88 %	99,50 %	85,0 %
1	**5**	3.431	10 %	100 %	3.430	12 %	100 %	99,97 %	86,5 %
Total		***34.311***			***29.684***				

Tabla 4.16: Tabla de performance del modelo Random Forest - Muestra de Modelamiento

La distribución de los sujetos totales y sujetos etiquetados como malos en cada uno de los 10 intervalos de la puntuación score en la muestra de validación empleando el modelo generado por el algoritmo Random Forest se presenta en la Tabla 4.17.

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
441	**999**	5.387	10 %	10 %	2.002	5 %	5 %	37,16 %	37,2 %
179	**441**	5.388	10 %	20 %	3.017	7 %	12 %	55,99 %	46,6 %
131	**179**	5.387	10 %	30 %	3.988	9 %	21 %	74,03 %	55,7 %
102	**131**	5.388	10 %	40 %	4.236	10 %	31 %	78,62 %	61,5 %
79	**102**	5.387	10 %	50 %	4.422	10 %	41 %	82,09 %	65,6 %
59	**79**	5.387	10 %	60 %	4.601	11 %	52 %	85,41 %	68,9 %
44	**59**	5.388	10 %	70 %	4.881	11 %	63 %	90,59 %	72,0 %
23	**44**	5.387	10 %	80 %	5.068	12 %	75 %	94,08 %	74,7 %
8	**23**	5.388	10 %	90 %	5.281	12 %	87 %	98,01 %	77,3 %
1	**8**	5.387	10 %	100 %	5.374	13 %	100 %	99,76 %	79,6 %
Total		***53.874***			***42.870***				

Tabla 4.17: Tabla de performance del modelo Random Forest - Muestra de Validación

Continuamos con la presentación de la tabla de performance de 10 intervalos uniformes (ver Tabla 4.18)generada mediante la aplicación del algoritmo Gradient Boosting Machine en los datos de la muestra de modelamiento.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
441	**999**	3.431	10 %	10 %	1.519	5 %	5 %	44,27 %	44,3 %
207	**441**	3.431	10 %	20 %	2.271	8 %	13 %	66,19 %	55,2 %
128	**207**	3.431	10 %	30 %	2.879	10 %	22 %	83,91 %	64,8 %
91	**128**	3.431	10 %	40 %	3.065	10 %	33 %	89,33 %	70,9 %
70	**91**	3.432	10 %	50 %	3.148	11 %	43 %	91,72 %	75,1 %
47	**70**	3.431	10 %	60 %	3.242	11 %	54 %	94,49 %	78,3 %
25	**47**	3.431	10 %	70 %	3.331	11 %	66 %	97,09 %	81,0 %
15	**25**	3.431	10 %	80 %	3.389	11 %	77 %	98,78 %	83,2 %
9	**15**	3.431	10 %	90 %	3.410	11 %	88 %	99,39 %	85,0 %
1	**9**	3.431	10 %	100 %	3.430	12 %	100 %	99,97 %	86,5 %
Total		***34.311***			***29.684***				

Tabla 4.18: Tabla de performance del modelo GBM - Muestra de Modelamiento

De igual forma, para los datos de la muestra de validación la tabla de performance (ver Tabla 4.19) del modelo basado en el algoritmo Gradient Boosting Machine presenta los 10 rangos de puntuación de score y la distribución de los clientes.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
429	**999**	5.387	10 %	10 %	1.983	5 %	5 %	36,81 %	36,8 %
199	**429**	5.388	10 %	20 %	3.094	7 %	12 %	57,42 %	47,1 %
123	**199**	5.387	10 %	30 %	3.923	9 %	21 %	72,82 %	55,7 %
89	**123**	5.388	10 %	40 %	4.318	10 %	31 %	80,14 %	61,8 %
70	**89**	5.387	10 %	50 %	4.468	10 %	41 %	82,94 %	66,0 %
51	**70**	5.387	10 %	60 %	4.592	11 %	52 %	85,24 %	69,2 %
33	**51**	5.388	10 %	70 %	4.919	11 %	63 %	91,30 %	72,4 %
20	**33**	5.387	10 %	80 %	5.128	12 %	75 %	95,19 %	75,2 %
10	**20**	5.388	10 %	90 %	5.296	12 %	88 %	98,29 %	77,8 %
1	**10**	5.387	10 %	100 %	5.370	12 %	100 %	99,68 %	80,0 %
Total		***53.874***			***43.091***				

Tabla 4.19: Tabla de performance del modelo GBM - Muestra de Validación

Ahora, el poder discriminatorio del modelo generado por el algoritmo Naïve Bayes en los datos de la muestra de modelamiento se presenta en la Tabla 4.20.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
956	**999**	3.431	10 %	10 %	1.914	6 %	6 %	55,79 %	55,8 %
847	**956**	3.431	10 %	20 %	2.391	8 %	15 %	69,69 %	62,7 %
714	**847**	3.431	10 %	30 %	2.744	9 %	24 %	79,98 %	68,5 %
570	**714**	3.431	10 %	40 %	2.963	10 %	34 %	86,36 %	73,0 %
350	**570**	3.432	10 %	50 %	3.115	10 %	44 %	90,76 %	76,5 %
71	**350**	3.431	10 %	60 %	3.137	11 %	55 %	91,43 %	79,0 %
8	**70**	3.431	10 %	70 %	3.280	11 %	66 %	95,60 %	81,4 %
3	**8**	3.431	10 %	80 %	3.338	11 %	77 %	97,29 %	83,4 %
2	**3**	3.431	10 %	90 %	3.403	11 %	89 %	99,18 %	85,1 %
1	**2**	3.431	10 %	100 %	3.399	11 %	100 %	99,07 %	86,5 %
Total		***34.311***			***29.684***				

Tabla 4.20: Tabla de performance del modelo NB - Muestra de Modelamiento

Así mismo, en la Tabla 4.21 presentamos los resultados del modelo Naïve Bayes, esto es, la distribución del total de sujetos y sujetos pertenecientes a la categoría de malos de la variable dependiente en cada uno de los 10 rangos de la puntuación score en la muestra de validación.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
949	**999**	5.387	10 %	10 %	2.434	6 %	6 %	45,18 %	45,2 %
830	**949**	5.388	10 %	20 %	3.168	7 %	13 %	58,80 %	52,0 %
709	**830**	5.387	10 %	30 %	3.591	8 %	21 %	66,66 %	56,9 %
574	**709**	5.388	10 %	40 %	4.078	10 %	31 %	75,69 %	61,6 %
411	**574**	5.387	10 %	50 %	4.411	10 %	41 %	81,88 %	65,6 %
128	**411**	5.387	10 %	60 %	4.632	11 %	52 %	85,98 %	69,0 %
25	**128**	5.388	10 %	70 %	4.940	12 %	64 %	91,69 %	72,3 %
6	**25**	5.387	10 %	80 %	5.085	12 %	75 %	94,39 %	75,0 %
2	**6**	5.388	10 %	90 %	5.205	12 %	88 %	96,60 %	77,4 %
1	**2**	5.387	10 %	100 %	5.325	12 %	100 %	98,85 %	79,6 %
Total		***53.874***			***42.869***				

Tabla 4.21: Tabla de performance del modelo NB - Muestra de Validación

En la Tabla 4.22 presentamos el nombre de los predictores con su respectivo parámetro estimado normal y estandarizado, importancia relativa en base a la reducción del error cuadrático medio y si la variable premia o castiga en el modelo, además, la importancia de los predictores en la regresión logística se representa de manera gráfica en la Figura 4.6.

Variable	Coeficiente	Coeficiente estandarizado	Importancia relativa	Sistema
antiguedad_sce	-0,0310	-0,1050	0,10503	Premia
prbb_nope_apert_sbs_op_12M	-1,7416	-0,0705	0,0705	Premia
r_mvalven_sicomsdeuda_total_sbs_3M	0,0001	0,0064	0,0064	Penaliza
PorcCuentaCoop	-0,7297	-0,0494	0,0494	Premia
r_prom_max_dven_n_op_3s12M	0,6532	0,3620	0,3620	Penaliza
r_nope_xven_op_3s12M	-0,2230	-0,0934	0,0934	Premia
r_prom_ven_sbssprom_deuda_total_sbs_6M	0,3745	0,1106	0,1106	Castiga
prbb_nent_ven_sc_12M	0,4354	0,0184	0,0184	Penaliza
ln_prom_ven_sbs_tc_12M	0,1508	0,3153	0,3153	Penaliza
r_nope_apert_scssce_op_12M	-0,4034	-0,1182	0,1182	Premia

Tabla 4.22: Predictores del modelo de regresión logística

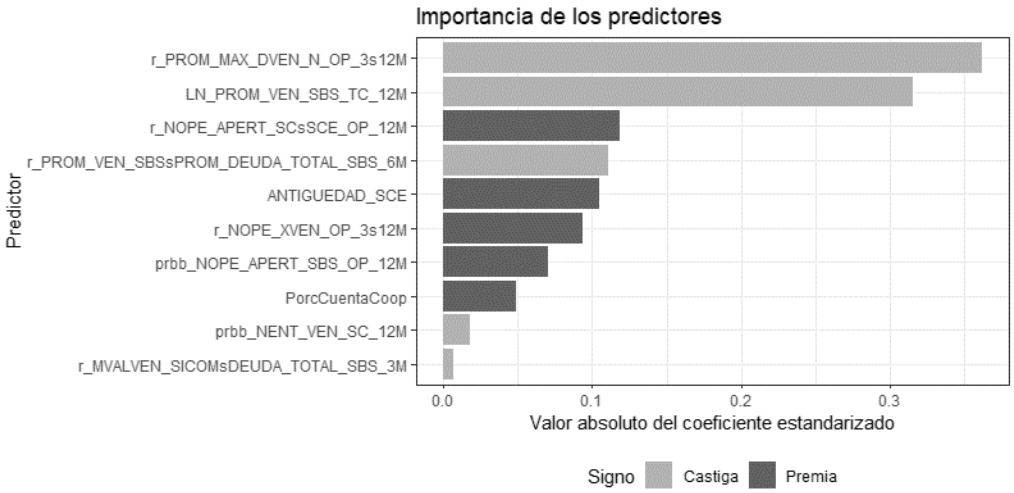


Figura 4.6: Importancia de las variables explicativas del modelo de RL

En cuanto a la tabla performance en base a los datos de modelamiento distribuidos uniformemente en 10 intervalos y el modelo de regresión, es presentada en la Tabla 4.23.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
672	**999**	3.431	10 %	10 %	2.076	7 %	7 %	60,51 %	60,5 %
629	**671**	3.431	10 %	20 %	2.292	8 %	15 %	66,80 %	63,7 %
581	**629**	3.431	10 %	30 %	2.620	9 %	24 %	76,36 %	67,9 %
530	**581**	3.431	10 %	40 %	2.986	10 %	34 %	87,03 %	72,7 %
485	**530**	3.432	10 %	50 %	3.079	10 %	44 %	89,71 %	76,1 %
446	**485**	3.431	10 %	60 %	3.169	11 %	55 %	92,36 %	78,8 %
418	**446**	3.431	10 %	70 %	3.326	11 %	66 %	96,94 %	81,4 %
393	**418**	3.431	10 %	80 %	3.355	11 %	77 %	97,78 %	83,4 %
338	**393**	3.431	10 %	90 %	3.373	11 %	89 %	98,31 %	85,1 %
1	**338**	3.431	10 %	100 %	3.408	11 %	100 %	99,33 %	86,5 %
Total		***34.311***			***29.684***				

Tabla 4.23: Tabla de performance del modelo de RL - Muestra de Modelamiento

De la misma forma generamos la tabla de performance (ver Tabla 4.24) de la muestra de validación distribuida uniformemente en 10 intervalos.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
668	**999**	5.387	10 %	10 %	2.682	6 %	6 %	49,79 %	49,8 %
623	**668**	5.388	10 %	20 %	3.169	7 %	13 %	58,82 %	54,3 %
574	**623**	5.387	10 %	30 %	3.600	8 %	22 %	66,83 %	58,5 %
518	**574**	5.388	10 %	40 %	4.071	9 %	31 %	75,56 %	62,7 %
475	**518**	5.387	10 %	50 %	4.400	10 %	41 %	81,68 %	66,5 %
436	**475**	5.387	10 %	60 %	4.784	11 %	52 %	88,81 %	70,2 %
410	**436**	5.388	10 %	70 %	5.153	12 %	64 %	95,64 %	73,9 %
374	**410**	5.387	10 %	80 %	5.159	12 %	76 %	95,77 %	76,6 %
331	**374**	5.388	10 %	90 %	5.252	12 %	88 %	97,48 %	78,9 %
1	**331**	5.387	10 %	100 %	5.304	12 %	100 %	98,46 %	80,9 %
Total		***53.874***			***43.574***				

Tabla 4.24: Tabla de performance del modelo de RL - Muestra de Validación

Por último, en el segmento Dirty después de probar diversos modelos de ensamble y fundamentados en los estadísticos de Kolmogorov-Smirnov, el área bajo la curva ROC, el índice de GINI y la captación de malos, se obtuvo el modelo final basado en los algoritmos de *Gradient Boosting Machine* y *Naïve Bayes*.

La tabla de performance (ver Tabla 4.25) reporta la distribución de los sujetos totales y sujetos malos de la muestra de modelamiento en cada uno de los 10 intervalos uniformes.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
458	**999**	3.431	10 %	10 %	1.540	5 %	5 %	44,88 %	44,9 %
190	**458**	3.431	10 %	20 %	2.252	8 %	13 %	65,64 %	55,3 %
128	**190**	3.431	10 %	30 %	2.879	10 %	22 %	83,91 %	64,8 %
87	**128**	3.431	10 %	40 %	3.091	10 %	33 %	90,09 %	71,1 %
77	**87**	3.432	10 %	50 %	3.138	11 %	43 %	91,43 %	75,2 %
51	**77**	3.431	10 %	60 %	3.231	11 %	54 %	94,17 %	78,4 %
30	**51**	3.431	10 %	70 %	3.320	11 %	66 %	96,76 %	81,0 %
11	**30**	3.431	10 %	80 %	3.392	11 %	77 %	98,86 %	83,2 %
6	**11**	3.431	10 %	90 %	3.412	11 %	88 %	99,45 %	85,0 %
1	**6**	3.431	10 %	100 %	3.429	12 %	100 %	99,94 %	86,5 %
Total		***34.311***			***29.684***				

Tabla 4.25: Tabla de performance del modelo de ensamble - Muestra de Modelamiento

De igual modo, la distribución uniforme en 10 rangos de los sujetos de la muestra de validación en la tabla de performance respecto a su puntuación de score se observa en la Tabla 4.26.

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
432	**999**	5.387	10 %	10 %	2.014	5 %	5 %	37,39 %	37,4 %
180	**432**	5.388	10 %	20 %	3.082	7 %	12 %	57,20 %	47,3 %
126	**180**	5.387	10 %	30 %	3.934	9 %	21 %	73,03 %	55,9 %
86	**126**	5.388	10 %	40 %	4.321	10 %	31 %	80,20 %	62,0 %
77	**86**	5.387	10 %	50 %	4.467	10 %	41 %	82,92 %	66,1 %
56	**77**	5.387	10 %	60 %	4.572	11 %	52 %	84,87 %	69,3 %
38	**56**	5.388	10 %	70 %	4.887	11 %	63 %	90,70 %	72,3 %
17	**38**	5.387	10 %	80 %	5.124	12 %	75 %	95,12 %	75,2 %
7	**17**	5.388	10 %	90 %	5.295	12 %	88 %	98,27 %	77,7 %
1	**7**	5.387	10 %	100 %	5.365	12 %	100 %	99,59 %	79,9 %
Total		***53.874***			***43.061***				

Tabla 4.26: Tabla de performance del modelo de ensamble - Muestra de Validación

Capítulo 5

Comparación de resultados obtenidos

El propósito de este capítulo es comparar los rendimientos y calidad de discriminación de clientes buenos y malos de un modelo scoring tradicional frente al modelo de clasificación construido con la metodología de ensamble.

La regresión logística es una de las técnicas estadísticas más utilizadas para la construcción de un clasificador crediticio donde la variable dependiente es binaria (bueno/malo), por tal motivo, la comparación de resultados se realiza con el modelo obtenido con logit y el modelo desarrollado con la metodología de ensamble.

Es preciso destacar si bien el modelo GBM en la Tabla 3.19 presenta estadísticos que indican una mejor discriminación entre sujetos buenos y malos, no se emplea para la comparación de resultados ya que el modelo clasificador *Gradient Boosting Machine* es un ensamble homogéneo (ver subsección 2.4.2) y el objetivo de esta sección es realizar una comparación de modelos que utilizan y no utilizan una metodología de ensamble realizando énfasis en el desempeño de clasificación de modelos basados en ensambles heterogéneos.

Finalmente, el clasificador desarrollado con la técnica de regresión logística lo denominaremos en adelante modelo tradicional y el modelo desarrollado con la técnica de ensamble lo denominaremos en adelante modelo de ensamble.

5.1 Medidas de calidad de discriminación

En el desarrollo de la sección se compara la calidad de discriminación de clientes buenos y malos del modelo tradicional versus la calidad de discriminación de clientes buenos y malos del modelo de ensamble.

5.1.1 Segmento Clean

El modelo tradicional utilizado para la comparación de medidas de calidad de discriminación de este segmento está construido con los predictores de la Tabla 4.9 en la sección 4.

En la Tabla 5.1 observamos que el estadístico KS, el índice ROC y el coeficiente de GINI del modelo de ensamble presentan un valor superior a comparación de los mismos indicadores del modelo tradicional. Esto se puede justificar debido a que el modelo de ensamble combina las distintas decisiones de cada uno de los clasificadores que conforman el conjunto, además, cada clasificador complementa los errores del otro clasificador lo que conduce a un mejor rendimiento del modelo en el segmento CLEAN.

INDICADOR	**VALOR**	
	MODELO TRADICIONAL	**MODELO DE ENSAMBLE**
KS	31,7 %	37,9 %
ROC	70,8 %	75,4 %
GINI	41,6 %	50,7 %

Tabla 5.1: Medidas de calidad de discriminación - Segmento Clean

5.1.2 Segmento Dirty

El modelo tradicional utilizado para la comparación de medidas de calidad de discriminación de este segmento está construido con los predictores de la Tabla 4.22 en la sección 4.

En la Tabla 5.2 observamos que el estadístico KS, el índice ROC y el coeficiente de GINI del modelo de ensamble presentan un valor superior a comparación de los mismos indicadores del modelo tradicional. Esto se puede justificar debido a que el modelo de ensamble combina las distintas decisiones de cada uno de los clasificadores que conforman el conjunto, además, cada clasificador complementa los errores del otro clasificador lo que conduce a un mejor rendimiento del modelo en el segmento DIRTY.

INDICADOR	VALOR	
	MODELO TRADICIONAL	MODELO DE ENSAMBLE
KS	47,9	55,8
ROC	80,4	85,6
GINI	60,8	71,1

Tabla 5.2: Medidas de calidad de discriminación - Segmento Dirty

5.2 Tablas performance

Esta sección presenta la comparación de tablas performance del modelo tradicional frente a las tablas performance del modelo de ensamble.

5.2.1 Segmento Clean

Comparando las tablas de performance del modelo realizado tradicional (Ver Tabla 4.10) y del modelo de ensamble (Ver Tabla 4.12), se puede observar que la brecha de la probabilidad de incumplimiento entre los deciles extremos del rango de score es más amplia en el modelo de ensamble a comparación del modelo tradicional, además, de presentar un alto rendimiento de discriminación a lo largo del rango de score.

5.2.2 Segmento Dirty

Comparando las tablas de performance del modelo realizado tradicional (Ver Tabla 4.23) y del modelo de ensamble (Ver Tabla 4.25), se puede observar que la brecha de la probabilidad de incumplimiento entre los deciles extremos del rango de score es más amplia en el modelo de ensamble a comparación del modelo tradicional, además, de presentar un alto rendimiento de discriminación a lo largo del rango de score.

5.3 Curvas ROC

En esta sección para comparar los modelos mediante el área bajo la curva utilizamos la representación gráfica de sensibilidad frente a (1 - especificidad) conocida como *curva ROC* a través de la evaluación de el rendimiento del modelo tradicional y el rendimiento del modelo de ensamble.

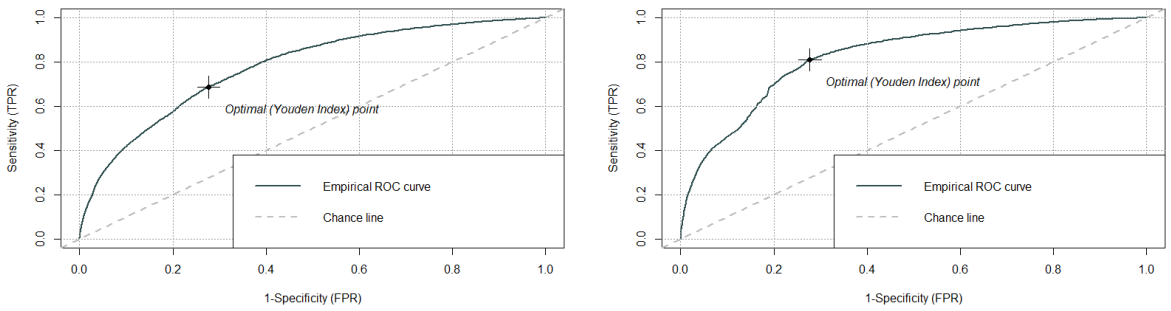
Denominamos sensibilidad a la probabilidad de clasificar correctamente a un cliente cuyo comportamiento de pago sea definido como bueno y la especificidad a la probabilidad de

clasificar correctamente a un cliente cuyo comportamiento de pago sea clasificado como malo.

En la representación gráfica entre más cercana se encuentra la curva ROC de la esquina superior izquierda del cuadrante, mejor será la predicción del modelo de clasificación y contiene los siguientes elementos:

- *Eje de las abscisas:* 1- especificidad.
- *Eje de las ordenadas:* sensibilidad.
- *Diagonal del gráfico:* divide la cuadrícula en dos mitades.
- *Área bajo la curva (AUC):* área de la cuadrícula bajo la curva de ROC.
- *Punto sobre la curva ROC:* índice de Youden definido como el máximo de la sensibilidad más especificidad menos uno.

Para el segmento Clean observamos que la curva de ROC del modelo de ensamble (Ver Figura 5.1(b)) diverge en mayor medida de la recta $Y = X$ y es más cercana de la esquina superior izquierda del cuadrante en comparación de la curva de ROC del modelo tradicional (Ver Figura 5.1(a)), por lo tanto, el modelo de ensamble tiene mejor rendimiento.

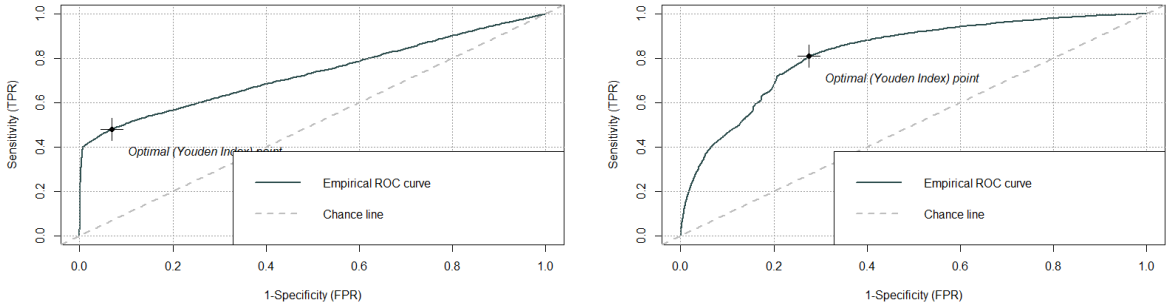


(a) Modelo Tradicional

(b) Modelo de Ensamble

Figura 5.1: Curvas ROC - Segmento Clean

Para el segmento Dirty observamos que la curva de ROC del modelo de ensamble (Ver Figura 5.2(b)) diverge en mayor medida de la recta $Y = X$ y es más cercana de la esquina superior izquierda del cuadrante en comparación de la curva de ROC del modelo tradicional (Ver Figura 5.2(a)), por lo tanto, el modelo de ensamble tiene mejor rendimiento.



(a) Modelo Tradicional

(b) Modelo de Ensamble

Figura 5.2: Curvas ROC - Segmento Dirty

Capítulo 6

Pruebas de ajuste

En el presente capítulo nos centramos en analizar el KS, ROC y GINI de los modelos de ensamble finales, el índice de condicionamiento (IC) permitirá estudiar la multicolinealidad de las variables explicativas incluidas en los modelos de ensamble y finalmente para probar que el modelo de ensamble no se encuentre sobre ajustado emplearemos el índice de estabilidad poblacional (PSI).

6.1 Análisis de KS, ROC y GINI

La Tabla 6.1 muestra que los valores obtenidos de KS, ROC y GINI se encuentran dentro del intervalo aceptable para considerar que el modelo desarrollado presenta características de ajuste, ordenamiento y discriminación adecuadas.

CRITERIO	**GBM GLM NB**		**GBM NB**		INTERVALO
	CLEAN	DIRTY	CLEAN	DIRTY	
KS	37,9 %	53,9 %	37,8 %	55,8 %	30,0 % - 100,0 %
ROC	75,4 %	84,4 %	75,3 %	85,6 %	60,0 % - 100,0 %
GINI	50,7 %	68,8 %	50,7 %	71,1 %	40,0 % - 100,0 %

Tabla 6.1: Medidas de discriminación

6.2 Análisis de multicolinealidad

El análisis de multicolinealidad lo realizamos para evitar el problema de mal condicionamiento, es decir, pequeñas variaciones en los datos implicarían grandes variaciones en el resultado y por lo tanto la estimación de cada parámetro se volvería inestable.

En el presente trabajo, el índice de condicionamiento (IC) es utilizado para estudiar el problema de multicolinealidad, el mismo que está definido por la expresión de la fórmula (6.1).

$$IC = \sqrt{\frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}}} \tag{6.1}$$

donde:

- $\lambda_{\text{máx}}$: es el valor propio máximo de la matriz de correlaciones de las variables explicativas del modelo de ensamble.
- $\lambda_{\text{mín}}$ es el valor propio mínimo de la matriz de correlaciones de las variables explicativas del modelo de ensamble.

Los valores de referencia del índice de condicionamiento de acuerdo a los estudios realizados por (Belsley et al., 2005) con datos observados y simulados son los siguientes:

- $IC < 10$: no hay presencia de multicolinealidad.
- *Entre* $10 \leq IC \leq 15$*:* existe una multicolinealidad moderada.
- $IC > 15$: existe una multicolinealidad fuerte.

6.2.1 Segmento Clean

Matriz de correlaciones para las variables de los modelos GBM, RF y NB

	prbb_estadocivil	antiguedad_sce	prbb_nope_apert_sbs_op_12m	d_nope_ndi_op_12m	r_nope_venc_op_3s12m	r_mvalven_sicomsdeuda_total_sbs_3m	r_nope_apert_sc_op_24s36m	r_mvalven_sbssdeuda_total_sbs_6m	r_prom_ven_scsprom_deuda_total_sbs_36m	porccuentacoop	prbb_mvalven_sbs_12m
prbb_estadocivil	1,00	0,25	0,13	0,16	-0,20	0,01	0,15	-0,18	0,00	0,24	-0,17
antiguedad_sce	0,25	1,00	0,02	-0,10	-0,30	0,01	-0,03	-0,06	0,01	0,04	-0,24
prbb_nope_apert_sbs_op_12m	0,13	0,02	1,00	0,22	-0,08	0,00	0,14	-0,06	-0,01	0,07	-0,17
d_nope_ndi_op_12m	0,16	-0,10	0,22	1,00	0,18	0,00	0,33	0,05	0,01	0,25	-0,20
r_nope_venc_op_3s12m	-0,20	-0,30	-0,08	0,18	1,00	0,00	-0,09	0,23	0,01	-0,09	0,26
r_mvalven_sicomsdeuda_total_sbs_3m	0,01	0,01	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,01	0,01
r_nope_apert_sc_op_24s36m	0,15	-0,03	0,14	0,33	-0,09	0,00	1,00	-0,11	0,00	0,26	0,03
r_mvalven_sbssdeuda_total_sbs_6m	-0,18	-0,06	-0,06	0,05	0,23	0,00	-0,11	1,00	-0,01	-0,18	-0,06
r_prom_ven_scsprom_deuda_total_sbs_36m	0,00	0,01	-0,01	0,01	0,01	0,00	0,00	-0,01	1,00	0,01	0,01
porccuentacoop	0,24	0,04	0,07	0,25	-0,09	0,01	0,26	-0,18	0,01	1,00	-0,04
prbb_mvalven_sbs_12m	-0,17	-0,24	-0,17	-0,20	0,26	0,01	0,03	-0,06	0,01	-0,04	1,00

Desde la expresión (6.2), se tiene que $IC = 1,93 < 10$, lo que implica que no existe presencia de multicolinealidad.

$$IC = \sqrt{\frac{2,0068745}{0,5409929}} = 1,926036 \tag{6.2}$$

Matriz de correlaciones para las variables del modelo GLM

	antiguedad_sce	prbb_nope_apert_sbs_op_12m	r_mvalven_sicomsdeuda_total_sbs_3m	porccuentacoop	r_prom_max_dven_n_op_3s12m	r_nope_xven_op_3s12m	r_prom_ven_sbssprom_deuda_total_sbs_6m	prbb_nent_ven_sc_12m	ln_prom_ven_sbs_tc_12m	r_nope_apert_scssce_op_12m
antiguedad_sce	1,00	0,02	0,01	0,04	-0,23	0,08	-0,05	-0,09	0,33	0,00
prbb_nope_apert_sbs_op_12m	0,02	1,00	0,00	0,07	-0,14	0,35	-0,15	0,06	0,00	-0,02
r_mvalven_sicomsdeuda_total_sbs_3m	0,01	0,00	1,00	0,01	0,01	0,00	0,00	0,00	0,00	0,01
porccuentacoop	0,04	0,07	0,01	1,00	-0,20	0,21	-0,22	0,30	0,00	0,21
r_prom_max_dven_n_op_3s12m	-0,23	-0,14	0,01	-0,20	1,00	-0,23	0,20	-0,17	-0,35	-0,16
r_nope_xven_op_3s12m	0,08	0,35	0,00	0,21	-0,23	1,00	-0,28	0,16	0,05	0,23
r_prom_ven_sbssprom_deuda_total_sbs_6m	-0,05	-0,15	0,00	-0,22	0,20	-0,28	1,00	-0,15	-0,18	-0,11
prbb_nent_ven_sc_12m	-0,09	0,06	0,00	0,30	-0,17	0,16	-0,15	1,00	-0,18	0,31
ln_prom_ven_sbs_tc_12m	0,33	0,00	0,00	0,00	-0,35	0,05	-0,18	-0,18	1,00	-0,04
r_nope_apert_scssce_op_12m	0,00	-0,02	0,01	0,21	-0,16	0,23	-0,11	0,31	-0,04	1,00

Desde la expresión (6.3), se tiene que $IC = 2,02 < 10$, lo que implica que no existe presencia de multicolinealidad.

$$IC = \sqrt{\frac{2,0334878}{0,5008231}} = 2,015017 \tag{6.3}$$

6.2.2 Segmento Dirty

Matriz de correlaciones para las variables de los modelos GBM, RF y NB

Desde la expresión (6.4), se tiene que $IC = 2,17 < 10$, lo que implica que no existe presencia de multicolinealidad.

$$IC = \sqrt{\frac{2,0222823}{0,4281415}} = 2,173338 \tag{6.4}$$

Matriz de correlaciones para las variables del modelo GLM

Desde la expresión (6.5), se tiene que $IC = 2,11 < 10$, lo que implica que no existe presencia de multicolinealidad.

$$IC = \sqrt{\frac{2,1849587}{0,4894248}} = 2,112898 \tag{6.5}$$

	r_nope_apert_sicomssce_op_36m	porc_uso_cupo	d_nent_ven_sce_op_36m	antiguedad_sce	d_ntc_ndi_tc_36m	r_nope_apert_sce_12a24m	r_mvalven_sbs_tcsmvalven_sbs_op_36m	prbb_nope_xven_op_12m	promlocalescom	r_deuda_total_scssce_3m	r_prom_ven_sbssprom_deuda_total_sbs_tc_36	prbb_ntc_apert_sce_24m	prbb_deuda_total_sce_3m
r_nope_apert_sicomssce_op_36m	1,00	0,04	0,05	-0,04	0,01	0,12	0,00	-0,12	-0,06	-0,06	0,00	-0,08	0,06
porc_uso_cupo	0,04	1,00	0,03	0,17	0,14	-0,13	0,00	0,14	-0,13	-0,12	-0,01	-0,27	-0,02
d_nent_ven_sce_op_36m	0,05	0,03	1,00	0,05	0,08	0,03	0,01	-0,09	0,00	0,02	0,05	0,04	0,00
antiguedad_sce	-0,04	0,17	0,05	1,00	0,18	-0,27	0,00	0,12	-0,08	-0,17	0,05	-0,09	-0,20
d_ntc_ndi_tc_36m	0,01	0,14	0,08	0,18	1,00	-0,06	0,02	0,06	-0,05	-0,06	0,15	-0,04	-0,04
r_nope_apert_sce_12a24m	0,12	-0,13	0,03	-0,27	-0,06	1,00	0,00	-0,34	0,05	0,17	0,01	0,12	-0,05
r_mvalven_sbs_tcsmvalven_sbs_op_36m	0,00	0,00	0,01	0,00	0,02	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00
prbb_nope_xven_op_12m	-0,12	0,14	-0,09	0,12	0,06	-0,34	0,00	1,00	-0,10	-0,18	0,00	-0,13	0,29
promlocalescom	-0,06	-0,13	0,00	-0,08	-0,05	0,05	0,00	-0,10	1,00	0,28	-0,01	0,11	-0,06
r_deuda_total_scssce_3m	-0,06	-0,12	0,02	-0,17	-0,06	0,17	0,00	-0,18	0,28	1,00	0,01	0,11	-0,09
r_prom_ven_sbssprom_deuda_total_sbs_tc_36m	0,00	-0,01	0,05	0,05	0,15	0,01	0,00	0,00	-0,01	0,01	1,00	0,03	0,00
prbb_ntc_apert_sce_24m	-0,08	-0,27	0,04	-0,09	-0,04	0,12	0,00	-0,13	0,11	0,11	0,03	1,00	0,04
prbb_deuda_total_sce_3m	0,06	-0,02	0,00	-0,20	-0,04	-0,05	0,00	0,29	-0,06	-0,09	0,00	0,04	1,00

	r_nope_apert_sicomssce_op_36m	porc_uso_cupo	d_nent_ven_sce_op_36m	antiguedad_sce	d_ntc_ndi_tc_36m	r_nope_apert_sce_12a24m	r_nope_apert_sbsssce_op_24m	promlocalescom	prbb_ntc_apert_sce_24m	prbb_nope_xven_op_12m
r_nope_apert_sicomssce_op_36m	1,00	0,04	0,05	-0,04	0,01	0,12	-0,16	-0,06	-0,08	-0,12
porc_uso_cupo	0,04	1,00	0,03	0,17	0,14	-0,13	-0,10	-0,13	-0,27	0,14
d_nent_ven_sce_op_36m	0,05	0,03	1,00	0,05	0,08	0,03	0,01	0,00	0,04	-0,09
antiguedad_sce	-0,04	0,17	0,05	1,00	0,18	-0,27	-0,15	-0,08	-0,09	0,12
d_ntc_ndi_tc_36m	0,01	0,14	0,08	0,18	1,00	-0,06	-0,05	-0,05	-0,04	0,06
r_nope_apert_sce_12a24m	0,12	-0,13	0,03	-0,27	-0,06	1,00	0,37	0,05	0,12	-0,34
r_nope_apert_sbsssce_op_24m	-0,16	-0,10	0,01	-0,15	-0,05	0,37	1,00	-0,08	0,10	-0,32
promlocalescom	-0,06	-0,13	0,00	-0,08	-0,05	0,05	-0,08	1,00	0,11	-0,10
prbb_ntc_apert_sce_24m	-0,08	-0,27	0,04	-0,09	-0,04	0,12	0,10	0,11	1,00	-0,13
prbb_nope_xven_op_12m	-0,12	0,14	-0,09	0,12	0,06	-0,34	-0,32	-0,10	-0,13	1,00

6.3 Análisis de estabilidad poblacional

Para probar que el modelo de ensamble no se encuentra sobre ajustado emplearemos el índice de estabilidad poblacional (*PSI* por sus siglas en inglés de *Population Stability Index*), el mismo que controla la estabilidad de la población en la cual se aplica el modelo (Muestra de Validación) respecto a la población con la cual se desarrolló el modelo (Muestra de Modelamiento).

La expresión (6.6) permite el cálculo del PSI para las muestras de modelamiento y validación denotadas por MOD y VAL, respectivamente.

$$PSI = \sum_{i=1}^{10} (VAL_i - MOD_i) \times ln\left(\frac{VAL_i}{MOD_i}\right) \tag{6.6}$$

donde:

- MOD_i: es el porcentaje de clientes de la población de modelamiento que tienen incumplimiento de pago en el rango i.
- VAL_i: es el porcentaje de clientes de la población de validación que tienen incumplimiento de pago en el rango i.

Los valores de referencia del índice de estabilidad poblacional son los siguientes:

- $PSI \leq 10\,\%$: las diferencias entre las distribuciones de las poblaciones comparadas no se consideran significativas, además, el modelo es bueno y presenta estabilidad en relación a la población de estudio.
- *Entre* $10\,\% < PSI \leq 25\,\%$: existen cambios en las distribuciones que requieren análisis, el modelo debe ser monitoreado.
- $PSI > 25\,\%$: las diferencias entre las distribuciones son significativas, el modelo presenta problemas.

Desde los datos de las tablas performance 4.12 y 4.13 del modelo de ensamble en el segmento CLEAN de la sección 4 se realizan los cálculos de la Tabla 6.2.

Decil	**Val (V)**	**Mod (M)**	**(V-M)**	**(V/M)**	**LN(V/M)**	**Índice**
1	0,0235	0,0199	0,0036	1,1819	0,1671	0,0006
2	0,0435	0,0398	0,0036	1,0916	0,0876	0,0003
3	0,0589	0,0545	0,0043	1,0797	0,0767	0,0003
4	0,0695	0,0675	0,0020	1,0297	0,0292	0,0001
5	0,0875	0,0834	0,0042	1,0501	0,0489	0,0002
6	0,1062	0,1042	0,0020	1,0195	0,0193	0,0000
7	0,1316	0,1280	0,0037	1,0286	0,0282	0,0001
8	0,1557	0,1694	-0,0137	0,9188	-0,0846	0,0012
9	0,1919	0,2357	-0,0438	0,8141	-0,2056	0,0090
10	0,3265	0,4342	-0,1076	0,7521	-0,2849	0,0307
					PSI	**4,25 %**

Tabla 6.2: Cálculo de PSI - Segmento Clean

Para el segmento CLEAN se obtiene que $PSI = 4,25\,\% \leq 10$, por lo tanto, no existen cambios significativos en la población, es decir, las poblaciones de modelamiento y validación son ordenadas y discriminadas por el modelo de ensamble de manera similar,

además, el modelo no presenta problemas de sobreajuste.

De manera análoga, los datos de las tablas performance 4.25 y 4.26 del modelo de ensamble en el segmento DIRTY de la sección 4 son empleados para los cálculos de la Tabla 6.3.

Decil	**Val (V)**	**Mod (M)**	**(V-M)**	**(V/M)**	**LN(V/M)**	**Indice**
1	0,3739	0,4488	-0,0750	0,8329	-0,1828	0,0137
2	0,5720	0,6564	-0,0844	0,8715	-0,1376	0,0116
3	0,7303	0,8391	-0,1088	0,8703	-0,1389	0,0151
4	0,8020	0,9009	-0,0989	0,8902	-0,1163	0,0115
5	0,8292	0,9143	-0,0851	0,9069	-0,0977	0,0083
6	0,8487	0,9417	-0,0930	0,9012	-0,1040	0,0097
7	0,9070	0,9676	-0,0606	0,9373	-0,0647	0,0039
8	0,9512	0,9886	-0,0375	0,9621	-0,0386	0,0014
9	0,9827	0,9945	-0,0117	0,9882	-0,0119	0,0001
10	0,9959	0,9994	-0,0035	0,9965	-0,0035	0,0000
					P.S.I	**7,54 %**

Tabla 6.3: Cálculo de PSI - Segmento Dirty

Para el segmento DIRTY se obtiene que $PSI = 7,54\,\% \leq 10$, por lo tanto, no existen cambios significativos en la población, es decir, las poblaciones de modelamiento y validación son ordenadas y discriminadas por el modelo de ensamble de manera similar, además, el modelo no presenta problemas de sobreajuste.

Capítulo 7

Alineación de Scores

El objetivo del presente capítulo es mostrar como se realiza la estandarización de los puntajes obtenidos al aplicar los distintos modelos estimados. Se detalla la alineación rango a rango que se aplica para los segmentos Clean y Dirty a partir de los resultados de la puntuación de score obtenidos en la muestra de validación.

7.1 Importancia de Alineación de Scores

Las tablas performance de cada uno de los modelos desarrollados muestran los resultados en 10 rangos distribuidos uniformemente, sin embargo, cada tabla performance de cada uno de los modelos tiene diferentes puntajes de score mínimo y máximo en cada intervalo, es por esto que surge la necesidad de traducir un valor de score o también denominado puntuación en función de un rango de un modelo a un valor de puntaje de otro modelo que usa un rango diferente con el fin de proporcionar la misma evaluación e interpretación del riesgo.

Se aplica la alineación de scores en el presente trabajo puesto que se desarrollaron modelos distintos para cada uno de los segmentos Clean y Dirty.

7.2 Proceso de Alineación de Scores

El proceso de la metodología de alineación de scores empleada en el presente trabajo se describe de manera general a continuación:

1. Seleccionamos un puntaje base, en este caso el puntaje del segmento CLEAN.

2. Ajustamos una regresión exponencial entre el puntaje medio del rango y la probabilidad de incumplimiento (PD) asociada al puntaje score como se expresa en la

ecuación (7.1)

$$PD = \frac{log(Score/a)}{b} \tag{7.1}$$

donde:

- *PD:* es la probabilidad de incumplimiento obtenida de la tabla performance del modelo de clasificación.
- *Score:* es el puntaje obtenido por el modelo de clasificación.
- *a:* constante real.
- *b:* constante real.

La Tabla 7.1 presenta las funciones para estimar la probabilidad de incumplimiento asociada a cada uno de los puntajes de score en los modelos de clasificación de ensamble finales.

MODELO		**FÓRMULA**
Ensamble GBM, GLM y NB	*Clean*	*PD_C = log(Score_C/2905.3725)/-4.2774*
	Dirty	*PD_D = log(Score_D/2905.3725)/-4.2774*
Ensamble GBM y NB	*Clean*	*PD_C = log(Score_C/3346.1998)/-4.4612*
	Dirty	*PD_D = log(Score_D/3346.1998)/-4.4612*

Tabla 7.1: Alineación de Probabilidad de Incumplimiento

3. Estimamos el score alineado mediante una transformación de tipo exponencial como se muestra en la ecuación (7.2).

$$Score = a \cdot exp\{-b \cdot PD\} \tag{7.2}$$

donde:

- *Score:* es el puntaje obtenido por el modelo de clasificación.
- *PD:* es la probabilidad de incumplimiento obtenida de la tabla performance del modelo de clasificación.
- *a:* constante real.
- *b:* constante real.

La Tabla 7.2 presenta las transformaciones empleadas para estimar el score alineado de cada uno de los modelos de clasificación de ensamble finales.

MODELO		FÓRMULA
Ensamble GBM, GLM y NB	*Clean*	*Score_Alineado_C = redondear(1074.3311 · exp{-2.1981 · PD_C})*
	Dirty	*Score_Alineado_D = redondear(1074.3311 · exp{-2.1981 · PD_D})*
Ensamble GBM y NB	*Clean*	*Score_Alineado_C = redondear(1073.6294 · exp{-2.1909 · PD_C})*
	Dirty	*Score_Alineado_D = redondear(1073.6294 · exp{-2.1909 · PD_D})*

Tabla 7.2: Alineación de Score

4. Alineamos los puntajes del segmento DIRTY respecto al puntaje base en cada uno de los modelos de clasificación como se muestra en la Tabla 7.3. Los rangos de puntuación del segmento CLEAN se mantienen invariantes ya que son considerados como score base.

MODELO	SCORE FINAL
(E1) Ensamble GBM, GLM y NB	*Score_Ali_E1 := Si(Clean_Dirty = DIRTY, fr(gr(Score_Dirty_E1)), Score_Clean_E1)*
(E2) Ensamble GBM y NB	*Score_Ali_E2 := Si(Clean_Dirty = DIRTY, fr(gr(Score_Dirty_E2)), Score_Clean_E2)*

Tabla 7.3: Alineación de puntajes respecto al segmento CLEAN

Las tablas performance obtenidas una vez que se realizó el proceso de alineación se presentan a continuación:

- **CLEAN**
  Las tablas performance del segmento CLEAN se mantienen invariantes, ya que se consideran como score base las puntuaciones del segmento.

– Modelo de Clasificación GBM, GLM y NB

KS	ROC	Gini
35,6	**73,9**	**47,8**

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
957	**999**	22.418	10 %	10 %	514	2 %	2 %	2,29 %	2,3 %
946	**957**	22.390	10 %	20 %	994	4 %	6 %	4,44 %	3,4 %
934	**946**	22.388	10 %	30 %	1.300	5 %	10 %	5,81 %	4,2 %
913	**934**	22.658	10 %	40 %	1.581	6 %	16 %	6,98 %	4,9 %
889	**913**	22.330	10 %	50 %	1.920	7 %	24 %	8,60 %	5,6 %
865	**889**	22.423	10 %	60 %	2.410	9 %	33 %	10,75 %	6,5 %
831	**865**	22.138	10 %	70 %	2.966	11 %	44 %	13,40 %	7,5 %
788	**831**	22.493	10 %	80 %	3.500	13 %	57 %	15,56 %	8,5 %
665	**788**	22.672	10 %	90 %	4.328	16 %	73 %	19,09 %	9,7 %
1	**665**	22.239	10 %	100 %	7.265	27 %	100 %	32,67 %	11,9 %
Total		***224.149***			***26.778***				

Tabla 7.4: Tabla de Performance Validación GBM,GLM y NB - Segmento Clean

– Modelo de Clasificación GBM y NB

KS	ROC	Gini
35,5	**73,9**	**47,8**

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
957	**999**	22.385	10 %	10 %	518	2 %	2 %	2,31 %	2,3 %
946	**957**	22.419	10 %	20 %	973	4 %	6 %	4,34 %	3,3 %
934	**946**	22.404	10 %	30 %	1.312	5 %	10 %	5,86 %	4,2 %
915	**934**	22.680	10 %	40 %	1.589	6 %	16 %	7,01 %	4,9 %
889	**915**	22.274	10 %	50 %	1.939	7 %	24 %	8,71 %	5,6 %
865	**889**	22.570	10 %	60 %	2.410	9 %	33 %	10,68 %	6,5 %
831	**865**	22.117	10 %	70 %	2.896	11 %	43 %	13,09 %	7,4 %
788	**831**	22.441	10 %	80 %	3.524	13 %	57 %	15,70 %	8,5 %
665	**788**	22.607	10 %	90 %	4.382	16 %	73 %	19,38 %	9,7 %
1	**665**	22.252	10 %	100 %	7.215	27 %	100 %	32,42 %	11,9 %
Total		***224.149***			***26.758***				

Tabla 7.5: Tabla de Performance Validación GBM y NB - Segmento Clean

- **DIRTY**

  Las tablas performance del segmento DIRTY se alinean respecto al segmento CLEAN con las pautas establecidas al principio del capítulo.

– Modelo de Clasificación GBM, GLM y NB

	ROC	Gini
54,6	**85,1**	**70,2**

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
421	**999**	5.387	10 %	10 %	1.967	5 %	5 %	36,51 %	36,5 %
262	**421**	5.388	10 %	20 %	3.036	7 %	11 %	56,35 %	46,4 %
206	**262**	5.387	10 %	30 %	3.925	9 %	20 %	72,86 %	55,2 %
177	**206**	5.388	10 %	40 %	4.331	10 %	30 %	80,38 %	61,5 %
149	**177**	5.387	10 %	50 %	4.550	10 %	41 %	84,46 %	66,1 %
117	**149**	5.387	10 %	60 %	4.869	11 %	52 %	90,38 %	70,2 %
97	**117**	5.388	10 %	70 %	5.094	12 %	64 %	94,54 %	73,6 %
67	**97**	5.387	10 %	80 %	5.112	12 %	75 %	94,90 %	76,3 %
41	**67**	5.388	10 %	90 %	5.337	12 %	88 %	99,05 %	78,8 %
1	**41**	5.387	10 %	100 %	5.368	12 %	100 %	99,65 %	80,9 %
Total		***53.874***			***43.589***				

Tabla 7.6: Tabla de Performance Validación GBM,GLM y NB - Segmento Dirty

– Modelo de Clasificación GBM y NB

KS	**ROC**	**Gini**
50,6	**83,1**	**66,2**

Score		**Total**			**Malo**			**Razón de Malo**	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
393	**999**	5.387	10 %	10 %	2.019	5 %	5 %	37,48 %	37,5 %
256	**393**	5.388	10 %	20 %	3.076	7 %	12 %	57,09 %	47,3 %
215	**256**	5.387	10 %	30 %	3.930	9 %	21 %	72,95 %	55,8 %
178	**215**	5.388	10 %	40 %	4.325	10 %	31 %	80,27 %	61,9 %
168	**178**	5.387	10 %	50 %	4.467	10 %	41 %	82,92 %	66,1 %
144	**168**	5.387	10 %	60 %	4.572	11 %	52 %	84,87 %	69,3 %
119	**144**	5.388	10 %	70 %	4.887	11 %	63 %	90,70 %	72,3 %
80	**119**	5.387	10 %	80 %	5.124	12 %	75 %	95,12 %	75,2 %
52	**80**	5.388	10 %	90 %	5.295	12 %	88 %	98,27 %	77,7 %
1	**52**	5.387	10 %	100 %	5.365	12 %	100 %	99,59 %	79,9 %
Total		***53.874***			***43.060***				

Tabla 7.7: Tabla de Performance Validación GBM y NB - Segmento Dirty

- **Población Total de Validación**

  A continuación, se detalla las tablas performance con el nuevo score alineado para toda la población de validación alineado respecto al segmento CLEAN.

– Modelo de Clasificación GBM, GLM y NB

KS	ROC	Gini
63,1	**88,5**	**77,0**

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
955	**999**	27.721	10 %	10 %	698	1 %	1 %	2,52 %	2,5 %
938	**955**	27.662	10 %	20 %	1.377	2 %	3 %	4,98 %	3,7 %
919	**938**	27.823	10 %	30 %	1.841	3 %	6 %	6,62 %	4,7 %
890	**919**	27.648	10 %	40 %	2.263	3 %	9 %	8,19 %	5,6 %
858	**890**	27.691	10 %	50 %	3.006	4 %	13 %	10,86 %	6,6 %
821	**858**	27.468	10 %	60 %	3.842	5 %	19 %	13,99 %	7,8 %
727	**821**	27.910	10 %	70 %	4.800	7 %	25 %	17,20 %	9,2 %
464	**727**	27.834	10 %	80 %	7.628	11 %	36 %	27,41 %	11,5 %
156	**464**	27.981	10 %	90 %	17.704	25 %	61 %	63,27 %	17,3 %
1	**156**	28.285	10 %	100 %	27.145	39 %	100 %	95,97 %	25,3 %
Total		***278.023***			***70.304***				

Tabla 7.8: Tabla de Performance GBM,GLM y NB - Validación

– Modelo de Clasificación GBM y NB

KS	ROC	Gini
62,7	**88,3**	**76,6**

Score		Total			Malo			Razón de Malo	
Min	**Max**	**Int#**	**Int %**	**Cum %**	**Int#**	**Int %**	**Cum %**	**Int**	**Cum**
956	**999**	27.701	10 %	10 %	726	1 %	1 %	2,62 %	2,6 %
937	**956**	27.648	10 %	20 %	1.351	2 %	3 %	4,89 %	3,8 %
917	**937**	27.889	10 %	30 %	1.812	3 %	6 %	6,50 %	4,7 %
890	**917**	27.621	10 %	40 %	2.313	3 %	9 %	8,37 %	5,6 %
860	**890**	27.713	10 %	50 %	2.963	4 %	13 %	10,69 %	6,6 %
821	**860**	27.441	10 %	60 %	3.858	6 %	19 %	14,06 %	7,8 %
726	**821**	27.900	10 %	70 %	4.809	7 %	26 %	17,24 %	9,2 %
448	**726**	27.827	10 %	80 %	7.589	11 %	36 %	27,27 %	11,5 %
171	**448**	27.998	10 %	90 %	17.604	25 %	62 %	62,88 %	17,2 %
1	**171**	28.285	10 %	100 %	26.635	38 %	100 %	94,17 %	25,1 %
Total		***278.023***			***69.660***				

Tabla 7.9: Tabla de Performance GBM y NB - Validación

Capítulo 8

Conclusiones y Recomendaciones

El incumplimiento de pago en las obligaciones adquiridas dentro de las instituciones financieras es más recurrente de lo que pensamos, por lo cual, se evidencia la necesidad de una búsqueda continua de un modelo de clasificación que permita estimar con mayor precisión el comportamiento de pago de los individuos, con el fin de que las instituciones financieras aumenten su tolerancia al riesgo mientras el segmento de individuos para otorgar un crédito sea más amplio sin una alteración negativa en sus márgenes de beneficio.

El modelo de clasificación binaria basada en clasificadores individuales, homogéneos y heterogéneos obtenido permite estimar la probabilidad de incumplimiento con mayor precisión que los modelos tradicionales, brindando a las instituciones financieras una nueva herramienta para el seguimiento del comportamiento de pago de los clientes de su cartera de crédito o identificando potenciales clientes que aún no pertenecen a la institución y son catalogados de bajo riesgo mejorando la rentabilidad de los prestamistas.

La investigación realizada de conceptos generales de riesgo de crédito, parámetros de ajuste de los clasificadores individuales y de ensamble, modelos scoring empleados en las instituciones financieras, proceso de construcción de modelos de ensamble homogéneos y heterogéneos e implementación en el software estadístico R fueron parte fundamental para el alcance los objetivos planteados en el trabajo.

A continuación, se enumera las principales conclusiones en base a los resultados obtenidos:

1. La información histórica del comportamiento de pago en una brecha de tiempo en operaciones crediticias y tarjetas de crédito resulta ser muy predictiva para medir el riesgo de crédito en cada uno de los clientes en base a sus características.

2. La metodología de ensamble para la construcción de un modelo scoring tiene la capacidad de aprender distintas estructuras y patrones estadísticos de múltiples

clasificadores, los algoritmos de ensamble *Gradient Boosting Machine* (Clasificador Homogéneo) y ensamble basado en los clasificadores *Gradient Boosting Machine*, *Naïve Bayes* y *Regresión Logística* (Clasificador Heterogéneo) proporcionan un mejor rendimiento y precisión para identificar individuos que incumplan las obligaciones adquiridas en una institución financiera.

3. El modelo clasificador basado en una metodología de ensamble cumple las siguientes desigualdades:

$$KS \geq \text{máx}\{KS_{X_1}, KS_{X_2}, \ldots, KS_{X_n}\} \tag{8.1}$$

$$KS \geq \text{máx}\{KS_{Y_1}, KS_{Y_2}, \ldots, KS_{Y_n}\} \tag{8.2}$$

donde:

- KS: es el valor del estadístico KS del modelo de clasificación binaria basado en una metodología de ensamble.
- $KS_{X_1}, KS_{X_2}, \ldots, KS_{X_n}$: son los valores del estadísitico KS de cada una de las variables explicativas $X_1, X_2, \ldots, X_n$ que integran los clasificadores base para el modelo final en el segmento CLEAN.
- $KS_{Y_1}, KS_{Y_2}, \ldots, KS_{Y_n}$: son los valores del estadísitico KS de cada una de las variables explicativas $Y_1, Y_2, \ldots, Y_n$ que integran los clasificadores base para el modelo final en el segmento DIRTY.

Podemos observar desde el ANEXO B el top de variables con mayor poder predictivo en el segmento CLEAN y DIRTY, la variable *antiguedad_ op_ sicom* (0,4483) tiene el mayor valor KS en el segmento CLEAN y del segmento DIRTY la variable con mayor valor KS es *porc_ uso_ cupo* (0,2003) que ha comparación de los valores de KS en los modelos de ensamble homogéneo y heterogéneos posterior a realizar la alineación de score en la población de validación son menores, cumpliéndose las desigualdades (8.1) y (8.2).

$$KS_{GBM} = 0,627 \geq \text{máx}\{KS_{X_1}, KS_{X_2}, \ldots, KS_{X_{75}}\} = 0,4483$$
$$KS_{GBM} = 0,627 \geq \text{máx}\{KS_{Y_1}, KS_{Y_2}, \ldots, KS_{Y_{75}}\} = 0,2003$$
$$KS_{GBM\_NB} = 0,627 \geq \text{máx}\{KS_{X_1}, KS_{X_2}, \ldots, KS_{X_{75}}\} = 0,4483$$
$$KS_{GBM\_NB} = 0,627 \geq \text{máx}\{KS_{Y_1}, KS_{Y_2}, \ldots, KS_{Y_{75}}\} = 0,2003$$
$$KS_{GBM\_GLM\_NB} = 0,631 \geq \text{máx}\{KS_{X_1}, KS_{X_2}, \ldots, KS_{X_{75}}\} = 0,4483$$
$$KS_{GBM\_GLM\_NB} = 0,631 \geq \text{máx}\{KS_{Y_1}, KS_{Y_2}, \ldots, KS_{Y_{75}}\} = 0,2003$$

4. Se analizó el comportamiento del modelo clasificador basado en la metodología de ensamble en una muestra distinta a la utilizada en el modelamiento y los resultados obtenidos en las tablas de performance de ordenamiento y discriminación en la muestra de validación fueron similares a los obtenidos en la muestra de modelamiento para el segmento CLEAN y DIRTY.

A continuación, se presentan algunas recomendaciones con el propósito de incentivar el uso del modelo score aquí desarrollado y la mejora continua en la calidad de predicción de los modelos de clasificación binaria:

1. El trabajo realizado en el documento es una investigación empírica, por lo que si se considera su implementación en alguna institución financiera debe ser de forma gradual con el fin de comparar en paralelo los resultados del modelo score que apliquen y el modelo propuesto para estimar el comportamiento de pago de las obligaciones.

2. Calibrar el modelo de clasificación binaria con la metodología de ensamble propuesto en función de las necesidades de las entidades financieras y posterior realizar ajustes periódicos.

3. Investigar el uso de algoritmos de clasificación binaria distintos a la regresión logística que es ampliamente empleada, con el fin de mejorar el rendimiento de la predicción de probabilidad de incumplimiento de pago de los clientes de las instituciones financieras y a la par el empleo de un software estadístico con alta capacidad de manipulación de grandes conjuntos de datos.

4. Con el fin de lograr un modelo de clasificación mucho más sólido, se recomienda investigar la posible incorporación de datos no tradicionales, como por ejemplo, la información de las redes sociales.

Bibliografía

Abdou, H. A. y Pointon, J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, vol. 18(2-3), páginas 59–88, 2011.

Ala'raj, M. y Abbod, M. A systematic credit scoring model based on heterogeneous classifier ensembles. En *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, páginas 1–7. IEEE, 2015.

Alaraj, M., Abbod, M. y Hunaiti, Z. Evaluating consumer loans using neural networks ensembles. En *International Conference on Machine Learning, Electrical and Mechanical Engineering*. 2014.

Ala'raj, M. y Abbod, M. F. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, vol. 104, páginas 89–105, 2016.

Altman, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, vol. 23(4), páginas 589–609, 1968.

Anderson, T. W. y Darling, D. A. A test of goodness of fit. *Journal of the American statistical association*, vol. 49(268), páginas 765–769, 1954.

Apampa, O. Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, vol. 25(4), página 6, 2016.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. y Vanthienen, J. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, vol. 54(6), páginas 627–635, 2003.

Beaver, W. H. Financial ratios as predictors of failure. *Journal of accounting research*, páginas 71–111, 1966.

Belsley, D. A., Kuh, E. y Welsch, R. E. *Regression diagnostics: Identifying influential data and sources of collinearity*, vol. 571. John Wiley & Sons, 2005.

Breiman, L. Bagging predictors. *Machine learning*, vol. 24(2), páginas 123–140, 1996.

Breiman, L. Random forests. *Machine learning*, vol. 45(1), páginas 5–32, 2001.

Crook, J. N., Edelman, D. B. y Thomas, L. C. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, vol. 183(3), páginas 1447–1465, 2007.

Dahiya, S., Handa, S. y Singh, N. P. Credit scoring using ensemble of various classifiers on reduced feature set. *Industrija*, vol. 43(4), 2015.

Darling, D. A. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, vol. 28(4), páginas 823–838, 1957.

Davis, R. H., Edelman, D. y Gammerman, A. Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, vol. 4(1), páginas 43–51, 1992.

Desai, V. S., Conway, D. G., Crook, J. N. y Overstreet Jr, G. A. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics*, vol. 8(4), páginas 323–346, 1997.

Dietterich, T. G. Machine-learning research. *AI magazine*, vol. 18(4), páginas 97–97, 1997.

Egmont-Petersen, M., Dassen, W. y Reiber, J. H. Sequential selection of discrete features for neural networks–a bayesian approach to building a cascade. *Pattern Recognition Letters*, vol. 20(11-13), páginas 1439–1448, 1999.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, vol. 7(2), páginas 179–188, 1936.

Junqué de Fortuny, E., Martens, D. y Provost, F. Predictive modeling with big data: Is bigger really better? *Big Data*, vol. 1(4), páginas 215–226, 2013.

Freund, Y., Schapire, R. E. et al. Experiments with a new boosting algorithm. En *icml*, vol. 96, páginas 148–156. Citeseer, 1996.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, páginas 1189–1232, 2001.

Friedman, J. H. Stochastic gradient boosting. *Computational statistics & data analysis*, vol. 38(4), páginas 367–378, 2002.

Goyal, A. y Kaur, R. A survey on ensemble model for loan prediction. *International Journal of Engineering Trends and Applications (IJETA)*, vol. 3(1), páginas 32–37, 2016.

Han, J., Kamber, M. y Pei, J. Data mining: concepts and techniques, waltham, ma. *Morgan Kaufman Publishers*, vol. 10, páginas 978–1, 2012.

Hand, D. J. y Henley, W. E. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160(3), páginas 523–541, 1997.

Henley, W. A k-nearest-neighbour classifier for assessing consumer credit risk. *The Statistician*, páginas 77–95, 1996.

Henley, W. E. *Statistical aspects of credit scoring*. Tesis Doctoral, The Open University, 1995.

Hosmer, D., Lemeshow, S. et al. Interpretation of the coefficients of the logistic regression model. *Applied logistic regression*, páginas 38–81, 1989.

Huang, C.-L., Chen, M.-C. y Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, vol. 33(4), páginas 847–856, 2007.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H. y Wu, S. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, vol. 37(4), páginas 543–558, 2004.

Jensen, H. L. Using neural networks for credit scoring. *Managerial finance*, 1992.

Kittler, J., Hatef, M., Duin, R. P. y Matas, J. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, vol. 20(3), páginas 226–239, 1998.

Kuncheva, L. I. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

Lessmann, S., Baesens, B., Seow, H.-V. y Thomas, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, vol. 247(1), páginas 124–136, 2015.

Lessmanna, S., Seowb, H., Baesenscd, B. y Thomasd, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. En *Credit Research Centre, Conference Archive*. 2013.

Lewis, D. D. Naive (bayes) at forty: The independence assumption in information retrieval. En *European conference on machine learning*, páginas 4–15. Springer, 1998.

Liaw, A. y Wiener, M. Classification and regression by randomforest, vol. 23. *Winston-Salem: Forest*, 2001.

Massey Jr, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, vol. 46(253), páginas 68–78, 1951.

Nanni, L. y Lumini, A. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, vol. 36(2), páginas 3028–3033, 2009.

Patil, P., Aghav, J. y Sareen, V. An overview of classification algorithms and ensemble methods in personal credit scoring. *IJCST*, vol. 7(2), páginas 183–187, 2016.

Pettitt, A. N. A two-sample anderson-darling rank statistic. *Biometrika*, vol. 63(1), páginas 161–168, 1976.

Ravi, V., Koparkar, S., Raju, N. P. y Sridher, S. Improving retail banking loans recovery via data mining techniques: a case study from indian market. *International Journal of Electronic Customer Relationship Management*, vol. 9(2-3), páginas 189–201, 2015.

Scholz, F. W. y Stephens, M. A. K-sample anderson–darling tests. *Journal of the American Statistical Association*, vol. 82(399), páginas 918–924, 1987.

Siddiqi, N. *Credit risk scorecards: developing and implementing intelligent credit scoring*, vol. 3. John Wiley & Sons, 2012.

Suman, M., Anuradha, T. y Veena, K. M. Direct marketing with the application of data mining. *International Journal of Engineering Research and Applications (IJERA)*, vol. 2(1), páginas 41–43, 2012.

Thomas, L. C., Edelman, D. B. y Crook, J. N. Readings in credit scoring: recent developments, advances, and aims. 2004.

West, D., Dellana, S. y Qian, J. Neural network ensemble strategies for financial decision applications. *Computers & operations research*, vol. 32(10), páginas 2543–2559, 2005.

Wolpert, D. H. Stacked generalization. *Neural networks*, vol. 5(2), páginas 241–259, 1992.

Zhang, D., Huang, H., Chen, Q. y Jiang, Y. A comparison study of credit scoring models. En *Third International Conference on Natural Computation (ICNC 2007)*, vol. 1, páginas 15–18. IEEE, 2007.

Zhou, X., Zhang, D. y Jiang, Y. A new credit scoring method based on rough sets and decision tree. En *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, páginas 1081–1089. Springer, 2008.

Zurada, J. y Barker, R. M. Using memory-based reasoning for predicting default rates on consumer loans. *Review of Business Information Systems (RBIS)*, vol. 11(1), páginas 1–16, 2007.