ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

EVALUACIÓN DE MODELOS DE MACHINE LEARNING APLICADOS AL CÁLCULO DE PÉRDIDAS ESPERADAS EN ENTIDADES DE MICROFINANZAS

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

RENE ALEJANDRO SANGACHA AVILA

renealejandrosangacha@gmail.com

DIRECTOR: MSC. DIEGO PAÚL HUARACA SAGÑAY

diego.huaracas@epn.edu.ec

DMQ, FEBRERO 2024

CERTIFICACIONES

Yo, RENE ALEJANDRO SANGACHA AVILA, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

RENE ALEJANDRO SANGACHA AVILA

Certifico que el presente trabajo de integración curricular fue desarrollado por RENE ALEJANDRO SANGACHA AVILA, bajo mi supervisión.

MSC. DIEGO PAÚL HUARACA SAGÑAY
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

RENE ALEJANDRO SANGACHA AVILA

MSC. DIEGO PAÚL HUARACA SAGÑAY

RESUMEN

El proyecto busca comparar el efecto de los modelos de machine learning no paramétricos frente a la metodología tradicional de regresión logística, para estimar las pérdidas esperadas de una cartera de crédito de una institución financiera ecuatoriana. Los modelos de Credit Scoring entrenados fueron: Regresión Logística (RGL), Random Forest (RF) y Extreme Gradient Boosting (XGB), los cuales estiman las probabilidades de ser mal pagador, considerando como tales a las personas con 61 o más días de vencimiento en la ventana de desempeño, según los resultados del análisis de Roll-Rate. Debido al desbalance de categorías BUENO/MALO, fue necesario rebalancear la muestra con la finalidad de capturar de mejor manera las características que predominan en los malos pagadores.

Considerando que para el desarrollo del proyecto se contó con un total de 805 variables, fue necesario emplear medidas de divergencia que permitan seleccionar las variables a ingresar en los modelos. Se utilizaron el test de Kolmogorov-Smirnov (KS) para variables cuantitativas y el test de Valor de Información (VI) para variables cualitativas, buscando diversidad de la información y una representación más precisa de la realidad financiera.

Los tres modelos entrenados tienen buenas métricas de rendimiento, tanto en poder de discriminación como en estabilidad predictiva, lo que se evidencia en las tablas performance obtenidas. Por los resultados de validación, el modelo con las mejores métricas de rendimiento resultó ser la metodología tradicional, seguido del XGB. Para el cálculo de las pérdidas esperadas, se utilizó el enfoque básico según Basilea II, es decir, se tomó la severidad de pérdida igual al 45 %. En base a los resultados de validación, el modelo que redujo el aprovisionamiento fue XGB, indicando que se debe aprovisionar en promedio el 3.51 % de la cartera mensualmente.

Palabras clave: **Credit Scoring, Regresión Logística, Random Forest, Extreme Gradient Boosting, Tablas Performance, Pérdida Esperada**

ABSTRACT

The project aims to compare the effect of non-parametric machine learning models against the traditional methodology of logistic regression, to estimate the expected losses of a credit portfolio of an Ecuadorian financial institution. The trained Credit Scoring models were: Logistic Regression (RGL), Random Forest (RF), and Extreme Gradient Boosting (XGB), which estimate the probabilities of being a defaulter, considering individuals with 61 or more days past due in the performance window, based on Roll-Rate analysis results. Due to the imbalance of GOOD/BAD categories, it was necessary to rebalance the sample in order to better capture the characteristics predominant in bad payers.

Considering that the project had a total of 805 variables, it was necessary to employ divergence measures to select the variables to be included in the models. The Kolmogorov-Smirnov (KS) test was used for quantitative variables, and the Value of Information (VI) test for qualitative variables, seeking diversity of information and a more accurate representation of the financial reality.

The three trained models have good performance metrics, both in discrimination power and predictive stability, as evidenced in the performance tables obtained. According to the validation results, the model with the best performance metrics turned out to be the traditional methodology, followed by XGB. For the calculation of expected losses, the basic approach according to Basel II was used, meaning the loss severity was taken as 45 %. Based on the validation results, the model that reduced provisioning was XGB, indicating that an average of 3.51 % of the portfolio should be provisioned monthly.

Keywords: **Credit Scoring, Logistic Regression, Random Forest, Extreme Gradient Boosting, Performance Tables, Expected Loss**

DEDICATORIA

A mi Mami y a mis hermanos, Kevin y Fernando, que la vida nos de el tiempo para seguir cuidandonos, apoyandonos y verles cumplir sus sueños, los amo.

AGRADECIMIENTO

Agradezco a Mama Esthela y Papa Zabala, mis abuelitos, sin su apoyo no habría podido terminar esta carrera, mi inspiración y motivación, sus decisiones y acciones me han mostrado la persona que deseo ser, gracias por todo, les debo todo, ahora estoy un paso más cerca de la promesa que les prometí gracias a ustedes.

A mi Ñaña, Juan y mis primos, por estar no solo para mi sino para mi familia, gracias por no dejarme caer.

A mi padre, por su apoyo a pesar de todo, gracias, he cumplido esta meta gracias a usted, por su compromiso conmigo y mis hermanos.

A mi tutor Diego Huaraca, no solo por el conocimiento compartido en la TIC sino a lo largo de toda la carrera, todo lo que aprendí de usted se que tendrá utilidad en mi vida profesional, muchas gracias.

A la directiva de la ASOiMAT, pay por haber confiado en mí, una parte por la cual acabe esta TIC fue gracias a que puede desarrollarla en la Aso.

A toda la gente que conocí gracias a la carrera, compañeros, amigos y profesores, en especial a la profe Kathia Pinzón, sus palabras me ayudaron a aclarar mi camino, también a Melany, alguien a quien quiero mucho y que fue parte de este proceso, pay Mel, te deseo lo mejor.

A Mabe Y Erik, pay por las noches de Polibus, las recordaré con cariño.

A la Sapoimath: Dianita, Ammy, Jorge, Agus, Cris, Jeremy, Luis y Erik, sin duda la carrera fue especial gracias a ustedes, sé que triunfarán, su talento en todo sé que les dará el futuro que desean y espero estar ahí para verlo. Estaré feliz de verles lograr sus sueños.

A mi Mami y a mis hermanos, no ha sido fácil, llegar hasta aquí es gracias a ustedes, pay por el tiempo, por comprenderme, su esfuerzo, son mi motivación y el porque continuo, gracias por todo, más de lo que las palabras pueden decir.

Y Finalmente, gracias a la vida por dejarme seguir.

Índice general

1. Descripción del componente desarrollado **1**

1.1. Descripción del proyecto 1

1.2. Objetivo general 2

1.3. Objetivos específicos 2

1.4. Alcance 3

2. Marco Teórico **4**

2.1. Modelos Estadísticos 4

2.1.1. Modelos de Aprendizaje No Supervisado 4

2.1.2. Modelos de Aprendizaje Supervisado 5

2.2. Regresión Logística (RGL) 6

2.3. Árbol de Decisión (AD) 7

2.4. Random Forest (RF) 14

2.5. Extreme Gradient Boosting (XGB) 15

2.5.1. Descripción Base del Algoritmo XGB 15

2.5.2. XGB: Método Greedy 19

2.6. Medidas de Divergencia 22

2.6.1. Test de Kolmogorov-Smirnov (KS) 22

2.6.2. Valor de Información (VI) 22

2.7. Índice de Condicionamiento 23

2.8. Métricas de Desempeño 23
2.8.1. Tabla Performance 24
2.8.2. KS 25
2.8.3. Curva ROC 26
2.8.4. GINI 26
2.8.5. Indice de Estabilidad Poblacional (IPS) 27
2.9. Pérdidas Esperadas 28

3. Metodología **32**
3.1. Credit Scoring para el calculo de pérdidas esperadas 32
3.2. Generación de la Información 33
3.3. Imputación Información Crediticia 34
3.4. Análisis Inicial de la Información 35
3.5. Definición de la Variable Dependiente 36
3.5.1. Análisis de Población Inicial 36
3.5.2. Análisis de Roll-Rate 37
3.5.3. Variable Dependiente 39
3.5.4. Especificación de la población de estudio 41
3.6. Muestra de Modelamiento y Validación 41
3.7. Balanceo de Categorías 42
3.8. Análisis de las Variables 43
3.9. Selección de Sistemas de Crédito 44
3.10. Creación de Variables 46
3.10.1. Variables de Marca 46
3.10.2. Ratios 47
3.10.3. Variables de Probabilidad 49
3.11. Selección de Variables 56
3.12. Modelos de Credit Scoring 58

3.12.1. Modelo Logit 58
3.12.2. Modelo Random Forest 62
3.12.3. Modelo XGBoost 67

4. Resultados **73**
4.1. Tablas Performance 74
4.1.1. Modelo Regresión Logística 74
4.1.2. Modelo Random Forest 77
4.1.3. Modelo XGBoost 80
4.1.4. Comparativa Métricas de Rendimiento 82
4.2. Pérdidas Esperadas 83

5. Conclusiones y recomendaciones **86**
5.1. Conclusiones 86
5.2. Recomendaciones 87

A. Anexos **90**
A.1. Código, modelos y bases de datos 90
A.2. Tablas Roll Rate 91
A.3. Selección de Variables para los modelos 95
A.4. Grilla de Hiperparámetros 96
A.4.1. Modelo Random Forest 96
A.4.2. Modelo XGBoost 97

Bibliografía **98**

Índice de figuras

2.1. Distribución Logística Estándar 7
2.2. Ejemplo Partición del Nodo - Caso: Clasificación 11
2.3. Ejemplo Partición del Nodo - Caso: Regresión 12
2.4. Interpretación KS 26

3.1. Generación de la Información 33
3.2. División de Sistemas de Crédito del Ecuador 36
3.3. Esquema: Variable Dependiente 40
3.4. Gráfico de Anillo: Marca Vencido 46
3.5. Esquema: Ejemplo Ratio de Deuda 48
3.6. Distribución Categorizada: Tipo Vivienda 52
3.7. Distribución Categorizada: Deuda Vencida SCE 12M 54
3.8. Importancia Variables - Modelo: RGL 60
3.9. Matriz de Correlaciones - Modelo: RGL 61
3.10 Importancia Variables - Modelo: RF 65
3.11 Matriz de Correlaciones - Modelo: RF 66
3.12 Importancia Variables - Modelo: XGB 71
3.13 Matriz de Correlaciones - Modelo: XGB 72

4.1. Razón de pérdida en la Ventana de Desempeño 84

Capítulo 1

Descripción del componente desarrollado

1.1. Descripción del proyecto

La principal actividad del Sector Financiero Ecuatoriano es la intermediación, la cual, por sus características, permite generar mayores beneficios y, a su vez, conlleva a la presencia de mayores riesgos financieros, destacándose de sobremanera el riesgo de crédito.

La Superintendencia de Bancos, como ente regulador nacional, supervisa y controla a las entidades del Sistema Financiero para preservar su seguridad, estabilidad, solidez y transparencia, generando la normativa necesaria para la gestión de los riesgos y la determinación de los elementos mínimos que deben observar las entidades financieras en la administración del riesgo, en conformidad con los estándares internacionales, acorde a su naturaleza y escala de actividades.

Para la gestión del riesgo de crédito, existen normas tanto nacionales como internacionales (Comité de Basilea) que proporcionan los lineamientos para la adecuada gestión, adoptando políticas y procedimientos relacionados con el desarrollo de metodologías que permitan la identificación y medición, así como el establecimiento de los límites y mecanismos de monitoreo, control y mitigación de los niveles de exposición a este riesgo.

En este sentido, las entidades financieras se ven en la necesidad de cuantificar el riesgo de crédito, haciendo uso de diferentes enfoques y metodologías desarrolladas en este ámbito. Comúnmente, la técnica estadística adoptada para este propósito corresponde a la regresión logística; sin embargo, en los últimos años se ha prestado una atención creciente a los algoritmos de aprendizaje automático (Machine Learning) para desafiar a los modelos tradicionales y explorar nuevas soluciones en la estimación de la probabilidad de incumplimiento de los deudores al momento de adquirir un crédito.

El marco regulatorio que rige a las instituciones financieras establece que se constituyan las provisiones necesarias que permitan cubrir las eventuales pérdidas adquiridas por el incumplimiento de los pagos de los deudores de créditos.

En este proyecto, nos centraremos en comparar la capacidad predictiva de tres algoritmos de Machine Learning, siendo estos: Regresión Logística, Random Forest y Extreme Gradient Boosting, para la clasificación de deudores en entidades de microfinanzas.

1.2. Objetivo general

Construir modelos analíticos de Machine Learning que permitan estimar la probabilidad de incumplimiento de una persona natural para hacer frente a sus obligaciones crediticias en una entidad financiera.

1.3. Objetivos específicos

1. Comparar, por medio de estadísticos, la capacidad predictiva de los algoritmos de Machine Learning con la metodología tradicional de regresión logística.

2. Evaluar el impacto sobre el cálculo de las pérdidas esperadas que tienen los algoritmos de Machine Learning respecto a la metodología tradicional.

1.4. Alcance

Para alcanzar nuestro objetivo principal, es necesario adquirir conocimiento en modelos lineales generalizados y modelos de clasificación no paramétrica, así que se comenzará por estudiar estos tipos de modelos. Se consolidará, analizará y depurará la información crediticia, sociodemográfica y socioeconómica disponible para el desarrollo del proyecto (información con fecha de corte septiembre de 2021).

Se calcularán al menos 2 medidas de divergencia en función de cada tipo de variable, de manera que se pueda generar un ranking entre las variables candidatas a formar parte de cada uno de los modelos.

Posteriormente, se entrenarán los modelos tradicionales y los algoritmos de Machine Learning que permitan estimar la probabilidad de incumplimiento de una persona natural, y se evaluará el poder predictivo y el error de ajuste de cada uno de los modelos candidatos en la población de estudio.

Finalmente, se realizarán estimaciones del cálculo de las pérdidas estimadas (provisiones) con cada uno de los modelos, a fin de medir el beneficio que podrían generar este tipo de modelos.

Capítulo 2

Marco Teórico

En este capítulo se explica el fundamento teórico de la metodología utilizada. Se inicia explicando los modelos entrenados para el cálculo del Score crediticio. Luego, se detallan estadísticas y técnicas útiles tanto para la selección de variables como para la validación de los modelos, incluida la tabla performance, herramienta útil para calcular las pérdidas esperadas. Finalmente, se concluye con la explicación de como se estiman estas últimas.

2.1. Modelos Estadísticos

2.1.1. Modelos de Aprendizaje No Supervisado

Son modelos en los cuales, en la base de datos no se tiene una variable de interés predeterminada, en la que se conozca su valor para cada individuo, en cambio, estos modelos buscan identificar patrones y estructuras en los datos por sí mismos. Un ejemplo común de estos algoritmos se encuentra en la recomendación de contenido en plataformas de streaming como Netflix o Spotify, donde, en base al historial de consumo, se sugiere contenido. Estos modelos sirven para realizar clustering, jerarquizar y reducir la dimensionalidad de los datos. Ejemplos de estos modelos son K-means, T-SNE y Análisis de Componentes Principales (ACP).

2.1.2. Modelos de Aprendizaje Supervisado

Son modelos que buscan predecir una variable de interés predefinida. Es decir, para cada individuo se conoce su valor para esta variable, y en base a otras se intenta predecirla. Estos algoritmos se clasifican en modelos de regresión cuando buscan predecir una variable numérica, como por ejemplo estimar el número de crímenes a cometerse en una región, y clasificación, como cuando se busca determinar si una persona puede caer o no en morosidad en sus pagos. Ejemplos de estos modelos son la Regresión Lineal y Logística, Random Forest y XGBoost.

Modelos Paramétricos

Son modelos en los cuales se hacen suposiciones acerca de las distribuciones de los datos, así como sobre la relación entre las variables. Además, asumen hipótesis para los errores generados por estos. En donde, estos modelos se centran en estimar los parámetros que ayudan a describir la relación especificada. Ejemplos de estos modelos son la Regresión Logística, los modelos ARIMA, los modelos ETS y el modelo Lee-Carter Poisson.

Modelos No Paramétricos

Son modelos que no hacen suposiciones sobre la relación entre las variables, ni sobre las distribuciones de los datos ni sobre los errores generados por estos, y por lo tanto, pueden descubrir relaciones no observadas entre las variables. Ejemplos de estos modelos son el Random Forest, XGBoost, Redes Neuronales y DBSCAN.

2.2. Regresión Logística (RGL)

Es un modelo a través del cual se busca clasificar a un individuo de manera binaria. Para el enfoque de este proyecto, clasificar a una persona como buena o mala pagadora. En donde, el modelo estima la probabilidad de pertenecer a una categoría y en base a ella, clasifica al individuo en una u otra. Para el enfoque del proyecto, si denotamos la variable dependiente como:

$$Y = \begin{cases} 1 & \text{Si el individuo es mal pagador} \\ 0 & \text{Si el individuo es buen pagador} \end{cases} \tag{2.1}$$

Entonces, para cada individuo i de la base, el modelo buscará a través de k variables predictoras $X_1, \ldots, X_k$, estimar las probabilidades:

$$\begin{aligned} p_i &= P(y_i = 1 \mid x_{i1}, \ldots, x_{ik}) \\ q_i &= P(y_i = 0 \mid x_{i1}, \ldots, x_{ik}) = 1 - p_i \end{aligned}$$

Estas son las probabilidades de que el individuo sea malo y bueno, respectivamente, dadas unas características de interés como sus días de vencimiento, el saldo vencido, el saldo total, entre otras. La relación $q_i = 1 - p_i$ se tiene porque estamos en un problema binario.

Ahora bien, la forma funcional que asume el modelo para estimar estas probabilidades es la distribución logística estándar. Es decir, para cada individuo i de la base:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik})}}$$

Donde, el término elevado a la exponencial se puede ver como una regresión lineal múltiple:

$$z_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik} \tag{2.2}$$

Así, la distribución de probabilidad de malos pagadores tiene el comportamiento de la distribución logística estándar:

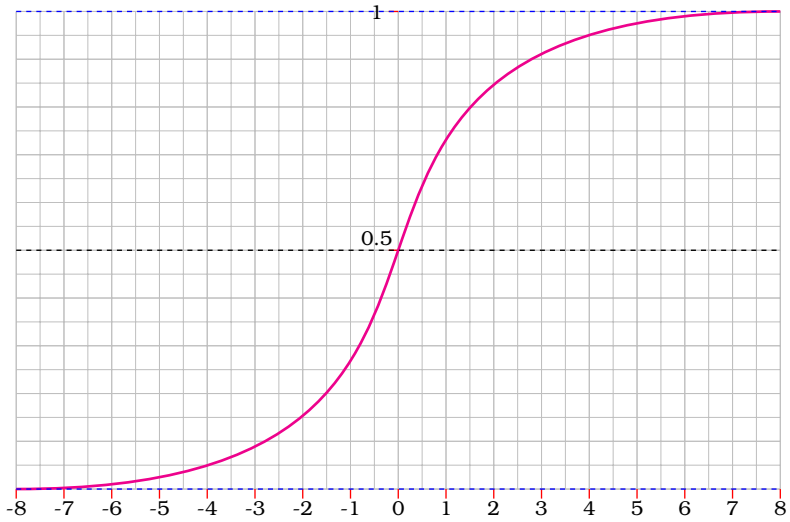

Figura 2.1: Distribución Logística Estándar
Elaboración: El autor

Donde los valores z_i de (2.2) representan:

$$z_i = ln\left(\frac{p_i}{1-p_i}\right) = ln(w_i)$$

Donde, w_i representa la odds o la razón de probabilidad, indicando que por cada M individuos buenos con las características $x_{i1}, \ldots, x_{ik}$, habrá $M \cdot w_i$ individuos malos con las mismas características. Los coeficientes del modelo se estiman mediante máxima verosimilitud, donde un coeficiente positivo indica que la variable asociada a ese coeficiente aumenta la probabilidad de ser mal pagador (castiga), mientras que un coeficiente negativo indica que la variable asociada disminuye la probabilidad de ser mal pagador (premia). Para más detalles acerca de este método, se puede revisar [19].

2.3. Árbol de Decisión (AD)

El árbol de decisión es un método no paramétrico que busca predecir una variable de interés, ya sea categórica (clasificación) o numérica (regresión), mediante la segmentación de la población en función de sus

características. Este modelo sirve como base para los otros dos modelos desarrollados en este trabajo. Donde, para el Random Forest, se utilizó el enfoque de clasificación, mientras que para el modelo XGBoost se utilizó el enfoque de regresión, los cuales se explicarán en sus respectivas secciones. Además, a través de el se realizó la categorización de variables. A continuación, se explica el árbol de decisión.

El árbol de decisión busca predecir la variable dependiente Y, mediante la segmentación de k variables predictoras, $X_1, \ldots, X_k$. Para llevar a cabo este proceso, se construye un grafo con estructura de árbol, donde cada nodo padre del árbol tiene al menos 2 hijos y aquellos nodos que carecen de hijos se denominan nodos hojas.

El modelo inicia con un nodo raíz que mediante una variable divide la población inicial en grupos (nodos). Este procedimiento se realiza de manera recursiva, es decir, en cada nuevo nodo se vuelve a seleccionar una variable para particionar la población, no necesariamente en la misma cantidad de grupos que en el paso anterior. Este proceso se repite hasta que se alcance algún criterio de parada previamente definido.

Ahora bien, para seleccionar la variable que particiona el nodo padre t, se elige aquella cuya segmentación contribuye a mejorar las predicciones. Para ello, se evalúa el poder de segmentación de la variable en cada nodo generado, el cual puede medirse a través de:

- **Problemas de Clasificación**
  Índice de Gini:

$$Gini(v) = 1 - \sum_{i=1}^{2} (p_i)^2 \tag{2.3}$$

  Donde v representa el nodo en análisis, y p_1 y p_2 son las probabilidades de que una persona en el nodo v sea un mal y buen pagador, respectivamente. Este índice mide la probabilidad de seleccionar aleatoriamente 2 individuos del nodo v y que estos sean distintos. Por lo tanto, es de interés elegir variables X que generen nodos v con un índice Gini bajo, ya que indicaría que la segmentación generada por la variable ayuda a diferenciar entre un buen y un mal pagador, dado que la probabilidad de tomar aleatoriamente 2 individuos, uno bueno y otro malo, es muy baja.

- **Problemas de Regresión**

  Suma de Errores Cuadrados (SSE):

$$SSE(v) = \sum_{i=1}^{n_v} (y_i - \hat{y}_v)^2 \tag{2.4}$$

  Donde v representa el nodo en análisis, n_v es el número de individuos en el nodo v, y $\hat{y}_v$ es la predicción de la variable Y en el nodo v, la cual es la misma para todos los individuos del nodo. El índice anterior mide el error de predicción, por lo que nos interesa seleccionar variables X que generen nodos v con un SSE bajo, ya que esto indicaría que la segmentación generada por la variable ayuda a tener una buena estimación de Y.

Así, se obtiene la medida de poder de segmentación en cada nodo hijo s y luego se procede a ponderarla. Esto implica calcular el índice Gini ponderado y el SSE ponderado para la variable analizada en el nodo padre t, dependiendo de si el problema es de clasificación o regresión:

$$Gini_{Ponderado}(t) = \sum_{s=1}^{r} p_s \cdot Gini_t(s)$$
$$SSE_{Ponderado}(t) = \sum_{s=1}^{r} p_s \cdot SSE_t(s)$$

En lo anterior, r es el número de grupos (nodos) creados por la variable, p_s representa el porcentaje de individuos del nodo padre t que se encuentran en el nodo hijo s. Los términos $Gini_t(s)$ y $SSE_t(s)$ son los índices de Gini y la suma de errores cuadrados (SSE) calculados para el nodo s de padre t, dependiendo del enfoque del árbol.

Y finalmente, se obtiene la métrica de poder de discriminación de la variable en el nodo padre t, según el enfoque del árbol:

- **Problemas de Clasificación:**

$$\Delta(t) = Gini(t) - Gini_{Ponderado}(t) \tag{2.5}$$

- **Problemas de Regresión:**

$$\Delta(t) = SSE(t) - SSE_{Ponderado}(t) \tag{2.6}$$

La cual me indica que si al particionar el nodo padre t, las predicciones de la variable Y han mejorado o no. Un valor positivo de la anterior métrica indicaría que las estimaciones en los nodos hijos son mejores que la estimación en el nodo padre, mientras que un valor negativo indicaría que la segmentación empeoró las predicciones, las del nodo padre son mejores. Donde, esta métrica permite comparar las diferentes particiones generadas por las variables.

Así, una vez definida la métrica que ayudará a seleccionar la variable para segmentar el nodo t, se lleva a cabo el siguiente procedimiento: para cada variable, se analizan distintos puntos de corte en caso de ser numérica, o diferentes agrupaciones de categorías si la variable es categórica. Posteriormente, se seleccionará para dividir el nodo la variable y sus puntos de corte o agrupaciones que maximicen el poder de discriminación $\Delta(t)$. Este proceso se realizará recursivamente en los nodos generados, donde no necesariamente se dividen los nuevos nodos en la misma cantidad de grupos que en el paso anterior. Este proceso se llevará a cabo hasta llegar al criterio de parada predefinido.

Para profundizar en esta parte, se presentan los siguientes ejemplos:

- **Problema de Clasificación**

  Supongamos que deseamos predecir si un individuo es buen o mal pagador, y nos encontramos en el nodo padre visto en la figura 2.2, al cual deseamos particionar. Entonces, para elegir la variable óptima que dividirá el nodo, se debe analizar cada variable. Comencemos con la variable Estado Civil y la agrupación Casado-Unión Libre. Así, tendríamos la siguiente partición:

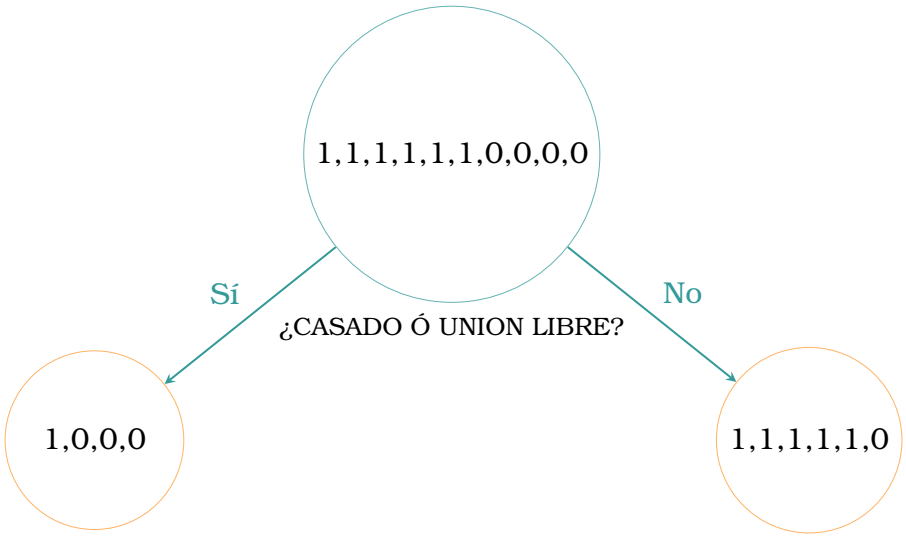

Figura 2.2: Ejemplo Partición del Nodo - Caso: Clasificación
Elaboración: El autor

En donde, en el nodo izquierdo y derecho se concentran el 40 % y 60 % de la información del nodo padre, respectivamente. Así, esta partición me indicaría que si una persona es casada o tiene unión libre, su probabilidad de ser mala y buena son $\frac{1}{4}$ y $\frac{3}{4}$, respectivamente. Mientras que una persona cuyo estado civil no es ninguna de las dos anteriores, tiene probabilidades de ser mala y buena de $\frac{5}{6}$ y $\frac{1}{6}$.

Luego, se analiza el índice de Gini dentro de cada nodo generado, utilizando la ecuación (2.3). Así, el índice de Gini en el nodo izquierdo es 0.375, mientras que en el nodo derecho es 0.277. Luego, se calcula el Gini Ponderado, que resulta en 0.3166 y finalmente, se determina el valor de la métrica de discriminación, ecuación (2.5). Dado que el índice de Gini del nodo padre t es 0.48, esta métrica tiene un valor de 0.1633.

Luego, para la misma variable se probarían otras combinaciones de agrupaciones. Por ejemplo, realizar 3 categorías: Viudo-Divorciado, Soltero-Unión Libre y Casado, o una de dos: Viudo-Divorciado y los que no. Así se probarían otras combinaciones y, para cada una de ellas, se obtendría su métrica de discriminación asociada. Luego, se seguiría el mismo proceso para las demás variables y al final, se seleccionaría la variable con los puntos de corte o agrupaciones que tuvieron la máxima métrica de poder de discriminación, ecuación (2.5), para particionar el nodo.

Supongamos que la mejor partición fue la mostrada en la figura 2.2. Hasta ese momento, las probabilidades estimadas de ser un mal pagador y buen pagador serían de $\frac{1}{4}$ y $\frac{3}{4}$, respectivamente, para todos los individuos del nodo izquierdo. Para el nodo derecho, las probabilidades serían $\frac{5}{6}$ y $\frac{1}{6}$, respectivamente. Luego, se repetiría el proceso para cada uno de los nodos hijos generados, donde no necesariamente se divide en la misma cantidad de nodos, y así sucesivamente de manera recursiva hasta alcanzar algún criterio de parada.

- **Problema de Regresión**

  Para el problema de regresión, el proceso es similar, cambiando únicamente la métrica que mide el poder de discriminación y la predicción a tomar dentro de cada nodo. Observemos el siguiente ejemplo: supongamos que deseamos estimar el número de autos robados por semana en una ciudad y tenemos el nodo padre visto en la figura 2.3, y se desea particionarlo. Así, de manera similar, se debería analizar cada variable y, para cada una de ellas, evaluar distintas agrupaciones o puntos de corte, y determinar cual proporciona el mejor poder de discriminación, ecuación (2.6).

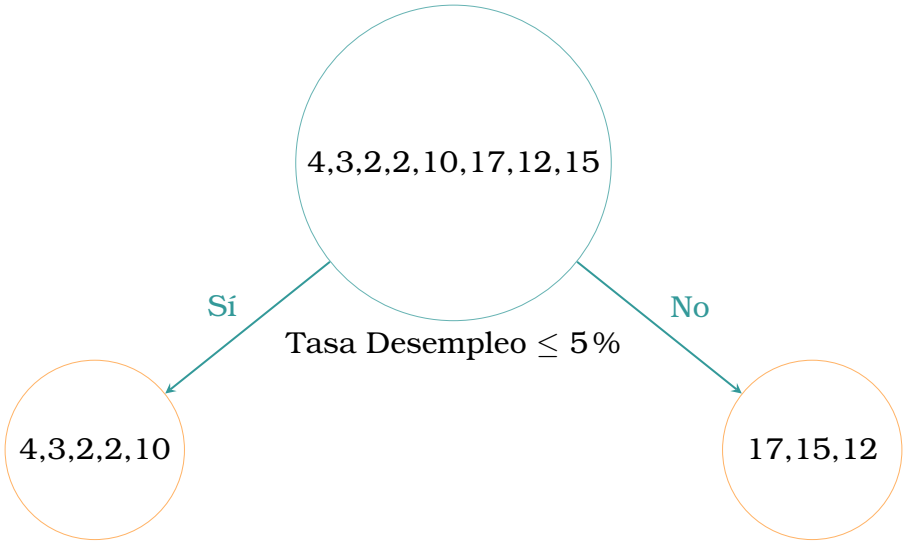

Figura 2.3: Ejemplo Partición del Nodo - Caso: Regresión
Elaboración: El autor

En este caso, se comenzó probando la tasa de desempleo en la ciudad con un umbral del 5%, y así se tiene la partición descrita en la figura 2.3. En ella, el 62.5% y el 37.5% de la información del nodo

padre se encuentran en el nodo izquierdo y derecho, respectivamente. Para estimar el número de autos robados a la semana, se calcula el promedio de la variable Y dentro de cada nodo hijo. Así, para todas las ciudades del nodo izquierdo se predice que se robarían semanalmente 4.2 carros, mientras que para todas las ciudades del nodo derecho se estima que se robarían 14.67 carros por semana.

Luego, se calcula el SSE por nodo para evaluar el error de predicción, utilizando la ecuación (2.4). Los valores serían de 44.8 y 12.67 para el nodo izquierdo y derecho, respectivamente. Posteriormente, se calcula el SSE ponderado, que en este caso sería de 32.75. Finalmente, se obtiene la métrica de discriminación descrita en (2.6). Dado que el SSE del nodo padre t es 262.88, el valor de la métrica de poder de discriminación sería 230.13.

Luego, para la misma variable, se van probando otras agrupaciones, como por ejemplo, una tasa de desempleo menor o mayor al 7.5%, o analizar la tasa de desempleo en 3 grupos: menor al 4%, entre 4% y 8%, y otro mayor al 8%. Así, se probarían otras combinaciones y para cada una de ellas se obtendría su métrica de poder de discriminación, según la ecuación (2.6). Posteriormente, se realizaría el mismo proceso para las demás variables y al final se seleccionaría la variable con los puntos de corte o agrupaciones que tuvieron la mayor métrica de discriminación, para dividir el nodo.

Supongamos que la mejor partición fue la mostrada en la figura 2.3. Hasta ese momento, los robos de autos estimados serían de 4.2 autos por semana para todas las ciudades del nodo izquierdo y de 14.67 autos por semana para todas las ciudades del nodo derecho. Luego, se volvería a repetir el proceso para cada uno de los nodos hijos, donde no necesariamente se dividen los nodos hijos en la misma cantidad, y así sucesivamente, hasta algún criterio de parada.

Donde, es importante aclarar que la división de los nodos no se realiza indefinidamente sino hasta cumplir un criterio de parada. Este criterio puede ser que no se puede dividir un nodo si no se tiene al menos un porcentaje mínimo de individuos en los nodos hijos o si no existe una variable tal que la métrica de discriminación descrita en (2.5) o (2.6) sea

positiva, porque en ese caso, las predicciones en los hijos generados serían peores que las que ya se tienen en el nodo padre y por lo tanto ya no tiene sentido particionar más la población.

Así, estos nodos que no tienen hijos son los denominados nodos hoja y son, en base a su información, que se determina el valor de la variable objetivo Y. Es decir, si, para clasificación, un nodo hoja tiene 93 registros malos y 7 buenos, la probabilidad de ser malo entregado por el árbol de decisión sería del 97 % para todos los individuos cuyas características los llevaron hasta ese nodo. Mientras que, en regresión, si el promedio del nodo hoja es de 3 autos robados por semana, ese será el valor estimado por el árbol de decisión para la variable Y para todas las ciudades cuyas características las llevaron hasta ese nodo hoja en específico.

2.4. Random Forest (RF)

El modelo de Random Forest es un modelo de ensamble que se basa en el modelo de árbol de decisión. El método consiste en construir una cantidad determinada de árboles, donde para cada uno se estima la variable dependiente Y para cada individuo. Donde, para estimar el valor final de Y, si el enfoque es de regresión, para cada individuo, se promedian los valores obtenidos por cada árbol. En cambio, si el modelo es de clasificación, el modelo entrega la categoría más votada para cada individuo como categoría predicha. En situaciones en las que se necesitan las probabilidades de pertenecer a una u otra categoría para cada individuo, la probabilidad entregada por el modelo es el promedio de las probabilidades dadas por cada árbol.

Ahora bien, sea Y la variable dependiente a predecir, con k variables predictoras $X_1, \ldots, X_k$, el proceso de construcción de cada árbol de decisión sigue el enfoque descrito en la sección 2.3, con una ligera variación: al dividir un nodo, en lugar de analizar todas las variables predictoras, se elige aleatoriamente un número específico de ellas. Al número de variables seleccionadas aleatoriamente para analizar la partición de un nodo se le denomina $\#mtry$, que junto con el número total de árboles a construir, $\#ntrees$, se consideran 2 hiperparámetros del modelo de Random Forest.

Donde, la probabilidad de seleccionar una variable para el análisis de partición de un nodo es $\frac{1}{k}$ y los valores usuales de $\#mtry$, dependiendo del enfoque del árbol, como se indica en [10], son:

- **Problema de Regresión:**

$$\#mtry = \frac{k}{3}$$

- **Problema de Clasificación:**

$$\#mtry = \sqrt{k}$$

Otro hiperparámetro de este modelo es $\#min.node.size$, que indica el número mínimo de registros que debe tener cada nodo hoja de cada árbol generado. Este hiperparámetro, como se mencionó en la sección del árbol de decisión, actúa como criterio de parada de división de nodos. Estos tres fueron los hiperparámetros considerados en la implementación del modelo. Para revisar otros hiperparámetros, se recomienda consultar [15]. El ajuste de hiperparámetros es importante para evitar tanto el subajuste como el sobreajuste del modelo.

2.5. Extreme Gradient Boosting (XGB)

El modelo XGBoost es un modelo secuencial basado en árboles de decisión binarios. Es secuencial porque cada árbol generado aprende de los errores del anterior, y es binario porque los nodos padres de cada árbol generado tienen exactamente dos hijos (nodo hijo izquierdo y nodo hijo derecho).

2.5.1. Descripción Base del Algoritmo XGB

Sea Y la variable dependiente a predecir, con k variables predictoras $X_1, \ldots, X_k$. El modelo inicia con una estimación inicial de Y, F_0, la cual comparten todos los individuos i de la población:

$$F_0(x_i) = \begin{cases} \overline{Y} & \text{Si } Y \text{ es numérica} \\ \dfrac{\#\text{Total de Malos}}{\#\text{Total de Individuos}} & \text{Si } Y \text{ es categórica} \end{cases} \tag{2.7}$$

Donde, el valor inicial del modelo de clasificación es de esa forma, ya que para este proyecto se consideró a los individuos malos con 1 en la variable dependiente Y, ecuación (2.1). Luego, se obtienen los residuos iniciales r_0:

$$r_0 = Y - F_0(x)$$

Y lo que se hace es estimar los residuos r_0 a través de las k variables predictoras mediante un árbol de decisión binario con enfoque de regresión, $r_0 \sim X_1, \ldots, X_k$. Supongamos que los valores predichos por el árbol binario son w_1. Entonces, se actualiza la predicción de Y como:

$$F_1(x) = F_0(x) + \eta \cdot w_1$$

Donde η representa la tasa de aprendizaje del modelo, indicando cuánto contribuye el árbol creado en cada iteración a la estimación de Y, varía entre 0 y 1. Al anterior valor se le denomina eta, que, junto al número de iteraciones (árboles) a realizar $\#nrounds$ y al número mínimo de registros en los nodos hojas de cada árbol generado, $\#min\_child\_weight$, resultan ser tres hiperparámetros del modelo.

Luego, se vuelven a obtener los nuevos residuos del modelo:

$$r_1 = Y - F_1(x)$$

Y se vuelve a realizar el mismo proceso iterativamente, es decir, estimar r_1 en función de las k variables predictoras mediante un AD binario con enfoque de regresión, obtener los residuos predichos w_2 y actualizar la estimación de Y:

$$F_2(x) = F_1(x) + \eta \cdot w_2$$

Después, se vuelven a obtener los residuos del modelo y se vuelve a realizar otra vez el mismo proceso, y así sucesivamente. Así, la predicción $m-$ésima de Y y su respectivo residuo serán:

$$F_m(x) = F_{m-1}(x) + \eta \cdot w_m \tag{2.8}$$

$$r_m = Y - F_m(x) \tag{2.9}$$

Este proceso se realiza hasta un criterio de parada, como por ejemplo, un máximo número de iteraciones o parar el entrenamiento si las predicciones no mejoran en una cantidad determinada de iteraciones. Ahora bien, el proceso de construcción de cada árbol de decisión binario con enfoque de regresión sigue lo descrito en la sección 2.3, con dos variaciones importantes.

En primer lugar, cada árbol no se estima necesariamente con las k variables predictoras. Inicialmente, se selecciona un número aleatorio de predictores para construir cada árbol, con una probabilidad de selección de $\frac{1}{k}$ para cada variable. A la proporción de variables tomadas aleatoriamente se le denomina $colsample\_bytree$, otro hiperparámetro del modelo.

Cabe destacar que, al igual que el modelo de Random Forest, es posible tomar aleatoriamente una cantidad de variables para el análisis de partición de un nodo. Sin embargo, en la implementación del modelo, no se trabajó con esta opción; es decir, para el análisis de partición de un nodo, se consideran todas las variables tomadas al principio para construir el árbol. Para obtener más información sobre otras opciones de hiperparámetros, se recomienda revisar [7].

La segunda variación y, sobre todo, fundamental, es que en cada iteración m, se busca construir un árbol que :

$$\min_{w_m} \sum_{i=1}^{n} \mathcal{L}[y_i, F_{m-1}(x_i) + \eta \cdot w_m(x_i)] + \Omega(w_m) \tag{2.10}$$

Donde n representa el número de individuos, $\mathcal{L}$ es una función de pérdida que varía según el enfoque del problema, y y_i, $F_{m-1}(x_i)$, $w_m(x_i)$ representan el valor original de la variable dependiente Y, la predicción de Y en la iteración anterior y el residuo predicho en la iteración actual para

el individuo i, respectivamente. Mientras que η es la tasa de aprendizaje del modelo. En resumen, la función de pérdida mide el error de predicción en la iteración m, ver (2.8). La función de pérdida a utilizar en este proyecto, dado que estamos en un enfoque de clasificación binario, será la pérdida logarítmica:

$$\mathcal{L}[y_i, t_i] = -[y_i \cdot \ln(t_i) + (1 - y_i) \cdot \ln(1 - t_i)] \tag{2.11}$$

Donde t_i es la probabilidad de ser malo para el individuo i. Así, esta función de pérdida penaliza cuando las predicciones de ser malo t_i son bajas si el individuo es malo. De manera similar, penaliza las probabilidades de ser bueno $1 - t_i$ bajas si el individuo es bueno. Pueden observarse más funciones de pérdida para todos los enfoques en [7].

Mientras que:

$$\Omega(w_m) = \gamma \cdot T + \frac{1}{2}\lambda \cdot \sum_{j=i}^{T} w_{m_j}^2 \tag{2.12}$$

Añade regularización al algoritmo y penaliza la complejidad del modelo para evitar el sobreajuste mediante la magnitud de los errores predichos w_{m_j}. Donde, T es el número de nodos hoja en el árbol de la iteración m, w_{m_j} representa el error predicho en la iteración m para el nodo hoja j del árbol. Mientras γ representa la pérdida mínima requerida para realizar una partición adicional en un nodo hoja del árbol, y λ es el coeficiente de regularización $L2$ de los pesos del árbol, ya que penaliza los errores predichos w_{m_j} más grandes. Valores altos de γ y λ hacen que el modelo sea más conservador.

Donde, inicialmente, se espera que a medida que pasan las iteraciones, la función objetivo se reduzca, siendo esta la forma de evaluar la mejora en las predicciones. Si no se observa una disminución en la función objetivo después de un cierto número de iteraciones, se detiene el entrenamiento. Además, dentro de cada árbol, la variable y sus puntos de corte o agrupaciones que se toman para particionar un nodo t ya no se eligen en función de las métricas de poder de discriminación $\Delta(t)$ descritas en la sección 2.3, ecuaciones (2.5) o (2.6). En cambio, se seleccionan en base a la función objetivo vista en (2.10). Es decir, para particionar un

nodo, se elige la variable con su punto de corte o agrupaciones que ayudan a reducir la función objetivo, siempre y cuando esta sea menor que la que se tiene en el nodo padre, ya que caso contrario no tiene sentido realizar una partición.

Así, este sería el proceso a través del cual se construyen los árboles de decisión binarios en cada iteración del algoritmo. El valor entregado por el modelo sería la estimación de Y alcanzada en la última iteración, correspondiendo a $\#nrounds$ con $F_{\#nrounds}$ si nunca se alcanzó el criterio de parada. También podría ser la predicción de la iteración m en la cual se detuvo el algoritmo o la predicción en la iteración en la que se alcanzó el menor valor de la función objetivo (2.10). Sí, la variable Y es numérica, se entrega un valor, mientras que si Y es categórica, se proporciona la probabilidad de la categoría de interés, en el caso del proyecto, la probabilidad de ser malo, esto por la descripción de la variable dependiente Y hecha en (2.1).

2.5.2. XGB: Método Greedy

En teoría, la forma en que se implementaría el algoritmo XGB sería la de la subsección anterior; sin embargo, al utilizar propiedades del cálculo en la función objetivo, se puede encontrar una forma simplificada de construir el árbol de regresión en cada iteración, la cual tiene sobre todo ventajas computacionales. Esta es la que se implementó en el proyecto y se basa en los resultados de [5].

Al utilizar aproximaciones de Taylor de Segundo Orden en la segunda componente de la función de pérdida, $\mathcal{L}$, se tiene que:

$$\sum_{i=1}^{n} \mathcal{L}[y_i, F_{m-1}(x_i) + \eta \cdot w_m(x_i)] \approx \sum_{i=1}^{n} \left[\cancelto{0}{\mathcal{L}[y_i, F_{m-1}(x_i)]} + g_i \eta w_m(x_i) + \frac{1}{2} h_i^2 w_m^2(x_i) \eta^2 \right]$$

Con:

$$g_i = \partial_{F_{m-1}(x_i)} \mathcal{L}[y_i, F_{m-1}(x_i)], \quad h_i = \partial^2_{F_{m-1}(x_i)} \mathcal{L}[y_i, F_{m-1}(x_i)] \tag{2.13}$$

Donde g_i y h_i representan la primera y segunda derivada parcial de

$\mathcal{L}$, respectivamente, con respecto a $F_{m-1}(x_i)$. Estas derivadas dependen únicamente de los valores de la estimación anterior y no de la actual. La función objetivo, ecuación (2.10), junto con su componente de regularización, ecuación (2.12) , se simplifica a:

$$\sum_{i=1}^{n} \mathcal{L}[y_i, F_{m-1}(x_i) + \eta \cdot w_m(x_i)] + \gamma \cdot T + \frac{1}{2}\lambda \cdot \sum_{j=i}^{T} w_{m_j}^2$$
$$= \sum_{i=1}^{n} \left[g_i \eta w_m(x_i) + \frac{1}{2} h_i^2 w_m^2(x_i) \eta^2 \right] + \gamma \cdot T + \frac{1}{2}\lambda \cdot \sum_{j=i}^{T} w_{m_j}^2$$

Donde, dado que w_{m_j} es constante en un mismo nodo hoja para todos los individuos, el problema a minimizar se reduce a:

$$\min_{w_m} \sum_{j=1}^{T} \left[\eta G_j w_{j_m} + \frac{1}{2} \left(H_j^2 \eta^2 + \lambda \right) w_{m_j}^2 \right] + \gamma T \tag{2.14}$$

El cual será ahora el problema a optimizar y es equivalente a (2.10). Donde:

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \tag{2.15}$$

Donde I_j es el conjunto de individuos del nodo hoja j del árbol de la iteración m. El mínimo del problema (2.14) se alcanzaría en:

$$w_{m_j} = -\frac{G_j \eta}{H_j \eta^2 + \lambda}, \forall j = 1, \ldots, T \tag{2.16}$$

Así, el mínimo del problema (2.14) sería:

$$-\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \frac{\lambda}{\eta^2}} + \gamma T \tag{2.17}$$

Este valor depende de las estimaciones anteriores, como se observa en las ecuaciones (2.13) y (2.15). Así, en la nueva forma de crear árboles, se supone que ya se ha alcanzado el mínimo de la función de pérdida en los nodos padres, y su valor es el descrito en (2.17). Por lo tanto, para realizar

una partición en un nodo, el valor de la función de pérdida debería ser menor a (2.17).

Así, para particionar un nodo, se realiza un análisis local, comparando el nodo padre t (un árbol de una sola hoja, $T = 1$) con la partición generada por la variable y su punto de corte o agrupación, teniendo ahora un árbol con 2 hojas (nodo izquierdo y derecho, $T = 2$), y se asume que se ha alcanzado el mínimo con esta partición, es decir, el valor descrito en (2.17). Así, la métrica de poder de discriminación compara los valores de la función objetivo antes y después de la partición:

$$\begin{aligned}\Delta(t) &= \text{Función Objetivo}_{\text{Padre}} - \text{Función Objetivo}_{\text{Partición}} \\ &= -\frac{1}{2}\frac{(G_I + G_D)^2}{H_I + H_D + \dfrac{\lambda}{\eta^2}} + \gamma - \left[\frac{1}{2}\frac{G_I^2}{H_I + \dfrac{\lambda}{\eta^2}} + \frac{1}{2}\frac{G_D^2}{H_D + \dfrac{\lambda}{\eta^2}} + 2\gamma\right]\end{aligned}$$

Así, si el valor de lo anterior es positivo, con el nodo padre no estábamos en el mínimo, pero al realizar la partición sí. Mientras que si es negativo, entonces sí estábamos en el mínimo en el nodo padre. Por lo tanto, la métrica de discriminación para una variable y su punto de corte o agrupación sería, en base a (2.17):

$$\Delta(t) = \frac{1}{2}\left[\frac{G_I^2}{H_I + \dfrac{\lambda}{\eta^2}} + \frac{G_D^2}{H_D + \dfrac{\lambda}{\eta^2}} - \frac{(G_I + G_D)^2}{H_I + H_D + \dfrac{\lambda}{\eta^2}}\right] - \gamma \tag{2.18}$$

De este modo, para particionar un nodo, se busca la variable con su punto de corte o agrupación que maximice la métrica de discriminación (2.18), la cual indica que se han mejorado las predicciones. Así, el proceso de construcción del árbol de decisión binario con enfoque de regresión para los residuos sería similar al descrito en la sección 2.3, utilizando esta nueva métrica, (2.18), en lugar de la suma de errores cuadrados (SSE) y tomando aleatoriamente una cantidad de variables de las k disponibles para construir todo el árbol.

2.6. Medidas de Divergencia

Las métricas para seleccionar las variables de los modelos son:

2.6.1. Test de Kolmogorov-Smirnov (KS)

Es un test estadístico que permite comparar dos distribuciones, ya sea una muestral con una teórica o dos muestrales. En este proyecto, se utilizará el segundo enfoque, ya que a través de este estadístico se observará que tan diferentes son las distribuciones de buenos y malos pagadores para una variable predictora X . Para ello, primero se construyen las distribuciones empíricas de X tanto para los individuos buenos como para los malos, digamos F_X^B y F_X^M, respectivamente. Luego se construye el estadístico:

$$D = \max_z \left| F_X^B(z) - F_X^M(z) \right| \tag{2.19}$$

Para obtener más información, revisar [18]. Así, nos interesa variables con un estadístico (2.19) alto, ya que esto indica que las distribuciones de la variable X son diferentes al pasar de buenos a malos pagadores y, por lo tanto, tienen variación que permite identificar a unos de otros. Se puede revisar un ejemplo del funcionamiento de este test en [9].

2.6.2. Valor de Información (VI)

Es una técnica que permite observar el poder predictivo de una variable con respecto a una variable objetivo, [8]. Así, para modelos de crédito, el Valor de Información de una variable categórica o numérica categorizada X es:

$$VI = \sum_{i=1}^{n} (p_i - q_i) \cdot \ln\left(\frac{p_i}{q_i}\right) \tag{2.20}$$

Donde:

- p_i: Es el porcentaje de individuos malos en la categoría i de X con

respecto al total de malos.

- q_i: Es el porcentaje de individuos buenos en la categoría i de X con respecto al total de buenos.
- n: Es el número de categorías de X.

Así, a mayor VI, mayor es el nivel predictivo de la variable, ya que como se observa en la ecuación (2.20), el estadístico calcula como las diferentes categorías de X separan a los individuos buenos de los malos, donde el logaritmo ayuda a amplificar o penalizar esas diferencias.

2.7. Índice de Condicionamiento

El índice de condicionamiento de una matriz de correlaciones permite evaluar la presencia de multicolinealidad en los datos, [3]. Este índice se calcula como:

$$IC = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

donde λ_{max} y λ_{min} son los valores propios máximo y mínimo de la matriz de correlaciones, respectivamente. Su interpretación es la siguiente:

IC	**Interpretación**
< 10	No se evidencian problemas significativos de multicolinealidad en los datos
$[10, 30)$	Existe multicolinealidad moderada en los datos
≥ 30	Indica una multicolinealidad fuerte en los datos

Cuadro 2.1: Interpretación del Índice de Condicionamiento
Fuente: Bertoli et al, [3]

2.8. Métricas de Desempeño

Las métricas a través de las cuales se realizará la validación de los modelos, tanto en poder discriminativo como en estabilidad poblacional, son:

2.8.1. Tabla Performance

La tabla performance nos permite evaluar la capacidad discriminativa de un modelo de Credit Scoring. En esta tabla, la población analizada se divide en deciles según su Score, que representa la probabilidad de ser considerado bueno en una escala del 1 al 999. A continuación, se presenta un ejemplo de una tabla performance:

KS	ROC	GINI
56.9	85.02	70.3

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
986	999	963	10 %	10 %	11	1 %	1 %	1.14 %	1.14 %
975	986	964	10 %	20 %	13	1 %	1 %	1.35 %	1.25 %
965	975	963	10 %	30 %	20	1 %	3 %	2.08 %	1.52 %
951	965	963	10 %	40 %	44	3 %	5 %	4.57 %	2.28 %
923	951	963	10 %	50 %	61	4 %	9 %	6.33 %	3.09 %
880	923	964	10 %	60 %	100	6 %	15 %	10.37 %	4.31 %
821	880	963	10 %	70 %	130	8 %	23 %	13.50 %	5.62 %
675	821	963	10 %	80 %	266	16 %	39 %	27.62 %	8.37 %
507	675	964	10 %	90 %	393	24 %	63 %	40.77 %	11.97 %
1	506	963	10 %	100 %	617	37 %	100 %	64.07 %	17.18 %
Total		9,633			1,655				

Cuadro 2.2: Ejemplo Tabla Performance
Elaboración: El autor

La tabla anterior representa el formato en el cual se presentarán las tablas performance para cada modelo entrenado en cada base, modelamiento y validación. Al inicio de la tabla, se encuentran tres medidas: KS, ROC, GINI, que permiten evaluar la capacidad discriminativa del modelo entrenado y comparar los modelos entre sí, las cuales serán descritas posteriormente. A continuación, se proporciona la información que describe cada columna:

- **Score Min:** El Score mínimo del rango.
- **Score Max:** El Score máximo del rango.
- **Total Int#:** El número de individuos en el rango.
- **Total Int %:** El porcentaje de individuos en el rango respecto al total de la población.

- **Total Cum %:** El porcentaje acumulado de individuos hasta ese rango.
- **Malo Int#:** El número de individuos malos en ese rango.
- **Malo Int %:** El porcentaje de individuos malos en el rango respecto al total de malos en la población.
- **Malo Cum %:** El porcentaje acumulado de individuos malos hasta ese rango respecto al total de malos de la población.
- **Razón de Malo Int %:** El porcentaje de individuos malos en el rango respecto al número de individuos en el rango.
- **Razón de Malo Cum %:** El porcentaje acumulado de individuos malos respecto al número de individuos acumulados hasta ese rango.

La columna Razón de Malo Int % indica la probabilidad de incumplimiento del crédito y se utiliza para el cálculo de las pérdidas esperadas. Así, para la tabla performance 2.2, las personas con el 10 % del Score más alto tienen una probabilidad de incumplimiento (PD) del 1.14 %, mientras que aquellas que se encuentran en el quinto decil, por ejemplo, tienen una PD del 6.33 %.

Además, la columna Razón de Malo Cum % ayuda a la institución a definir el riesgo que desea asumir. Para el ejemplo de la tabla 2.2, si la institución desea asumir un riesgo menor o igual al 4.50 % debería aprobar al 60 % de las personas con el Score más alto.

2.8.2. KS

La aplicación del test de Kolmogorov-Smirnov en modelos de crédito consiste en evaluar que tan diferentes son las funciones de distribución de buenos y malos pagadores para cada nivel de puntaje crediticio, con el propósito de analizar la capacidad discriminativa del modelo. El valor del test oscila entre 0 y 100 puntos [13]. La interpretación de este indicador, proporcionada por TransUnion, una empresa líder en la gestión de riesgo crediticio en la región, es la siguiente:

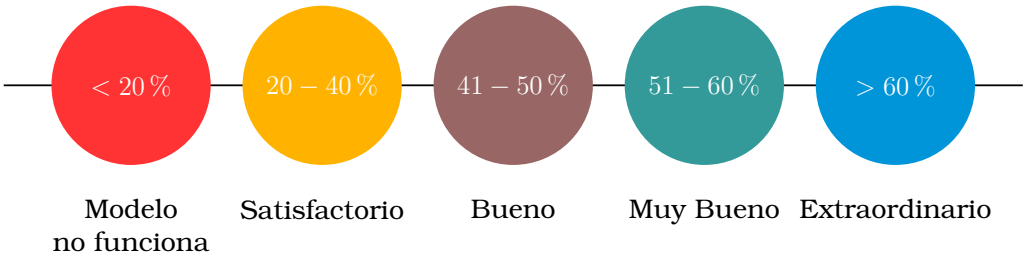

Figura 2.4: Interpretación KS
Fuente: TransUnion, [17]

2.8.3. Curva ROC

Representa el área bajo la curva ROC, la cual ilustra la relación entre la sensibilidad y 1-especificidad al variar el punto de corte en un modelo de clasificación binario, [11]. En un modelo de Credit Scoring:

- **Sensibilidad**: Probabilidad de que un buen pagador sea clasificado correctamente.
- **1-especificidad**: Probabilidad de que un mal pagador sea clasificado incorrectamente.

Su valor varía entre 0 y 1, si este valor es igual o inferior a 0.5, el modelo no tiene poder discriminativo y es equivalente a realizar predicciones de manera aleatoria. En el ámbito del riesgo crediticio, se considera que un área bajo la Curva ROC igual o superior a 0.75 cumple con el estándar aceptado por la industria, y a mayor valor, mejor es el nivel de discriminación del modelo [2].

2.8.4. GINI

Este indicador describe la propiedad de clasificación del modelo mediante un número entre 0 y 1, donde cuanto más cerca de 1 esté el valor, mejor será el poder predictivo del modelo [12]. Está relacionada con el ROC a través de la ecuación:

$$GINI = 2 \cdot ROC - 1$$

Dado que vamos a realizar un Score de Originación, es decir, aprobar o no un crédito, la interpretación del GINI es la siguiente:

$GINI$	**Poder Discriminativo del Modelo**
< 0.25	Bajo
$0.25 - 0.45$	Promedio
$0.45 - 0.60$	Bueno
> 0.60	Muy Bueno

Cuadro 2.3: Interpretación GINI
Fuente: Plug & Score, [6]

2.8.5. Indice de Estabilidad Poblacional (IPS)

El Índice de Estabilidad Poblacional (IPS) es una herramienta que analiza los cambios entre dos muestras. En el contexto de modelos, su enfoque se centra en comparar los resultados de la muestra con la que se entrenó, con la información de validación. Este proceso tiene como objetivo evaluar la relevancia de la información empleada e identificar posibles sobreajustes o subajustes en el modelo.

Una vez que el modelo está en operación, el IPS permite llevar a cabo un monitoreo continuo. Su función consiste en evaluar los cambios entre la población inicial y la actual. Con base en los resultados obtenidos, se toman decisiones, como llevar a cabo ajustes o incluso reemplazar el modelo existente. Este enfoque se realiza con el objetivo de preservar la validez del modelo a lo largo del tiempo, [4].

Donde el IPS para variables categóricas o numéricas categorizadas, como el puntaje crediticio, se calcula de la siguiente manera:

$$IPS(Y_M, Y_N) = \sum_{i=1}^{B} (p_{M,i} - p_{N,i}) \cdot \ln\left(\frac{p_{M,i}}{p_{N,i}}\right)$$

Donde Y_M es la variable de interés categorizada en la base de modelización, Y_N es la variable de interés categorizada en la nueva base, B es el número de categorías de la variable Y, y $p_{M,i}$ y $p_{N,i}$ representan los porcentajes de individuos en la categoría i en la muestra de modelización y la nueva, respectivamente. Para revisar un ejemplo, se puede consultar [4].

Este indicador lo utilizaremos para evaluar si hay cambios en el poder predictivo de los modelos al pasar de la base de modelización a la de validación. El análisis del IPS es el siguiente:

IPS	**Interpretación**
$< 10\,\%$	El modelo no presenta cambios significativos y puede seguir siendo utilizado
$[10, 25)\,\%$	El modelo necesita ajustes
$\geq 25\,\%$	El modelo presenta cambios significativos y ya no debe ser utilizado

Cuadro 2.4: Interpretación IPS
Fuente: Yurdakul, [4]

2.9. Pérdidas Esperadas

La Superintendencia de Bancos del Ecuador, en LIBRO I.- NORMAS GENERALES PARA LAS INSTITUCIONES DEL SISTEMA FINANCIERO, TITULO X.- DE LA GESTIÓN Y ADMINISTRACIÓN DE RIESGOS, CAPITULO II.- DE LA ADMINISTRACIÓN DEL RIESGO DE CRÉDITO (incluido con resolución No JB-2003-602 de 9 de diciembre de 2003), SECCIÓN I.- ALCANCE Y DEFINICIONES, Artículo 2, Numeral 2.1, define al Riesgo de crédito como: "*Es la posibilidad de pérdida debido al incumplimiento del prestatario o la contraparte en operaciones directas, indirectas o de derivados que conlleva el no pago, el pago parcial o la falta de oportunidad en el pago de las obligaciones pactadas*".

Mientras que en el numeral 2.2, se define el Incumplimiento como: "*No efectuar el pago pactado dentro del período predeterminado; o efectuarlo con posterioridad a la fecha en que estaba programado, o en distintas condiciones a las pactadas en el contrato*".

Así, la Pérdida Esperada se define en el mismo artículo, Numeral 2.7, como: "*Es el valor esperado de pérdida por riesgo crediticio en un horizonte de tiempo determinado, resultante de la probabilidad de incumplimiento, el nivel de exposición en el momento del incumplimiento y la severidad de la pérdida*", es decir:

$$PE = EAD \cdot LGD \cdot PD \tag{2.21}$$

Donde, en los numerales 2.3, 2.4 y 2.6 del mismo artículo se define:

- Probabilidad de incumplimiento (PD): "*Es la posibilidad de que ocurra el incumplimiento parcial o total de una obligación de pago o el rompimiento de un acuerdo del contrato de crédito, en un período determinado*".
- Nivel de exposición del riesgo de crédito (EAD): "*Es el valor presente (al momento de producirse el incumplimiento) de los flujos que se espera recibir de las operaciones crediticias*".
- Severidad de la pérdida (LGD): "*Es la medida de la pérdida que sufriría la institución controlada después de haber realizado todas las gestiones para recuperar los créditos que han sido incumplidos, ejecutar las garantías o recibirlas como dación en pago*".

Como indica Lucía Acurio en [1], las instituciones financieras deberían ser capaces de modelar las tres componentes de la pérdida esperada, según Basilea II. Dependiendo de su nivel de desarrollo, Basilea II propone tres esquemas distintos para la LGD:

- Esquema Básico: En este enfoque, se asigna generalmente una severidad de pérdida del 45 % para todos los créditos.
- Esquema Estándar: En este método, el regulador establece porcentajes estándares para calcular el capital mínimo requerido, teniendo en cuenta factores como el tipo de garantía y el nivel de cobertura.
- Esquema Avanzado: En este enfoque, las entidades financieras utilizan sus propias estimaciones internas para determinar la Pérdida Dada la Defecto (LGD).

Así, en el caso del proyecto, se seguirá un enfoque básico para el cálculo de las pérdidas esperadas, es decir, se tomará una severidad de pérdida igual al 45 %. Cabe destacar que existen trabajos desarrollados, como [14] y [16], en los que se explica cómo obtener la severidad de pérdida (LGD) de la institución a lo largo del tiempo. El primero de estos trabajos está adaptado al caso ecuatoriano, donde se explica cómo la LGD puede variar en función de factores exógenos a la institución, como el PIB o el IPC. Se discute cómo la inclusión de información propia de cada institución puede mejorar y personalizar las estimaciones. Este enfoque no

se desarrolló en el presente trabajo, pero podría ser objeto de análisis en trabajos posteriores.

Además, la exposición al riesgo (EAD) se considerará como el monto del crédito solicitado, mientras que la probabilidad de incumplimiento (PD) se determinará mediante la columna razón de malos Int % de las tablas performance, como se detalló en la sección 2.8.1. Por ejemplo, para un individuo con 940 puntos de Score en la tabla performance 2.2, y con un crédito de 5,000$, su pérdida esperada según (2.21) es:

$$PD = 5,000 \cdot 45\,\% \cdot 6.33\,\% \tag{2.22}$$

$$PD = 142.43\$ \tag{2.23}$$

Donde la PD es del $6.33\,\%$ debido a que una persona con 940 puntos de Score se encuentra en el quinto decil de Score de la tabla performance 2.2.

Este dinero es el que la institución debe aprovisionar en el Banco Central del Ecuador (BCE). Por lo tanto, se buscan montos bajos sustentados en criterios técnicos, ya que para el caso anterior, si se lograra desarrollar un modelo en el que la PE fuera de 120$, entonces la institución tendría 22.43$ más para la colocación de créditos, los cuales podría otorgar a una tasa de interés efectiva mayor que la ofrecida por el BCE para el aprovisionamiento de pérdidas.

Además, se obtiene el ratio de pérdida de la cartera:

$$r_{\text{Pérdida}} = \frac{\sum_{i=1}^{n} PE_i}{\sum_{i=1}^{n} EAD_i}$$

Donde PE_i y EAD_i representan la pérdida esperada y la exposición al riesgo del individuo i de los n analizados. También se puede calcular para un grupo de individuos, G, con una característica de interés, de la siguiente manera:

$$r_{\text{Pérdida}} = \frac{\sum_{i \in I_G}^{n} PE_i}{\sum_{i \in I_G}^{n} EAD_i}$$

Donde I_G contiene los individuos del grupo G de interés a analizar, como aquellos con terrenos, salarios mayores al SBU o aquellos en el primer mes de la ventana de desempeño. Este análisis puede ser útil para la institución en la definición de políticas de colocación y gestión de crédito.

Capítulo 3

Metodología

En este capítulo, se llevará a cabo el entrenamiento de tres modelos de Credit Scoring: Regresión Logística, Random Forest y XGBoost, con el objetivo de calcular las pérdidas esperadas de una cartera de crédito.

El proyecto fue desarrollado en el lenguaje de programación R, utilizando datos provenientes de una institución financiera del país, la cual, por confidencialidad, no se mencionará.

3.1. Credit Scoring para el calculo de pérdidas esperadas

Un modelo de Credit Scoring constituye una metodología estadística y analítica empleada en el sistema financiero que permite establecer una política de colocación de créditos minimizando los riesgos de la operación.

El modelo se basa en la información histórica de pago del individuo y en sus datos socioeconómicos con dos objetivos principales. En primer lugar, evalúa si se debe otorgar o no el crédito a una persona. Posteriormente, una vez concedido el crédito, determina la probabilidad de incumplimiento de pago, permitiendo a la institución estimar el monto que la persona no abonará de la deuda adquirida. Esta suma debe ser aprovisionada por la institución en el Banco Central del Ecuador, a la

cual esta entidad le aplica una tasa de interés pasiva.

Así, es de interés que este monto sea lo más bajo posible pero sustentado en criterios técnicos, para no perder oportunidades de negocio. Esto se debe a que una parte de ese dinero aprovisionado se podría destinar a créditos, con una tasa mayor a la que ofrece el Banco Central por aprovisionar la pérdida esperada.

3.2. Generación de la Información

La información de la institución con la que trabajaremos se recopila de manera transversal. En nuestro caso, contamos con 12 puntos de observación (datos mensuales igualmente espaciados), los cuales son:

Mes de Observación	
1	Ago 2020
2	Sep 2020
⋮	⋮
11	Jun 2021
12	Jul 2021

Cuadro 3.1: Puntos de Observación
Elaboración: El autor

En donde, para cada punto de observación, se obtiene la cartera total de socios activos al final del mes, es decir, las personas que abrieron crédito en ese mes. Donde, se dispone de una ventana histórica y una ventana de desempeño para cada punto de observación, como se ilustra a continuación:

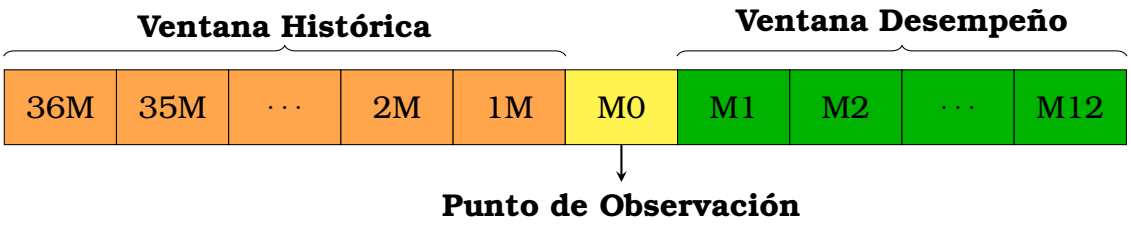

Figura 3.1: Generación de la Información
Elaboración: El autor

Donde, para cada individuo, se recopila su información histórica en el sistema crediticio ecuatoriano durante los 36 meses previos al punto de observación, que es la ventana máxima de tiempo establecida por la Superintendencia de Bancos del Ecuador para la extracción de información.

Adicionalmente, se obtiene información relevante para la institución en el punto de observación, incluyendo datos socioeconómicos de la persona y detalles sobre el producto otorgado. Finalmente, se recopila información sobre el desempeño de pago del crédito para cada individuo durante los 12 meses siguientes a su concesión. Esta información resulta fundamental, ya que nos permite determinar si el individuo es un buen o mal pagador.

Cabe resaltar que, la información de la ventana histórica y del punto de observación proviene tanto de la institución como del buró de crédito, al cual se recurrió para obtener información de mercado relacionada con cada individuo en cada punto de observación. Por otro lado, la información de la ventana de desempeño es exclusivamente suministrada por la institución.

3.3. Imputación Información Crediticia

Se llevó a cabo la primera imputación de datos, enfocándonos específicamente en la información crediticia. Durante este proceso, cuando se encontraba una variable que representaba el comportamiento crediticio y tenía un valor nulo (NA), se optó por imputar dicho valor con cero. La justificación detrás de esta elección fue la interpretación de que la ausencia de datos indicaba la falta del comportamiento crediticio correspondiente.

Un ejemplo concreto de esta imputación se refleja en casos donde la información sobre la Deuda Total o el número de entidades vigentes estaba ausente para una persona. En tales situaciones, se consideró que la persona no tenía deuda total o no contaba con entidades con créditos vigentes. De esta manera, se llevó a cabo el proceso de imputación para las variables de carácter crediticio.

3.4. Análisis Inicial de la Información

Ahora bien, es importante notar que un individuo puede abrir más de un crédito en un mes. Por ende, resulta crucial reducir la información de nivel de operación a nivel de persona, ya que los modelos estiman el comportamiento de pago por persona y no por operación. Para asegurar registros únicos por individuo en cada punto de observación, se decidió que, si un individuo tiene más de un crédito en un mes, por ejemplo la deuda vencida de ese individuo sería la suma de las deudas vencidas de cada uno de los créditos abiertos en ese periodo.

De esta manera, se trabajó con información sobre el número de operaciones, cantidad de acreedores, variables de deuda en términos monetarios, entre otras. Respecto a la información temporal, como los días de vencimiento, se eligió considerar como valor de vencimiento en esa fecha la cantidad más alta de días de vencimiento entre los créditos abiertos en dicho periodo.

Es importante aclarar que no solo se extrajo información del valor máximo para variables temporales, sino también para otras variables de interés, como la deuda vencida. A continuación, se presenta un ejemplo que ilustra lo mencionado anteriormente. Supongamos que en el mes de febrero de 2021, un individuo abrió 3 créditos con el siguiente comportamiento de pago en el mes siguiente:

Crédito	Deuda Vencida ($)	Días de Vencido
1	100	12
2	0	0
3	40	23

Cuadro 3.2: Ejemplo de tratamiento inical de la información
Elaboración: El autor

Así, con el objetivo de contar con registros únicos por individuo en cada punto de observación, se establece que la deuda vencida para este individuo asciende a 140$, mientras que tiene 23 días de vencido, en el mes siguiente al otorgamiento del crédito.

Donde, la información crediticia para cada individuo en cada punto de observación se obtiene de cada uno de los sistemas que se presentan en la figura 3.2.

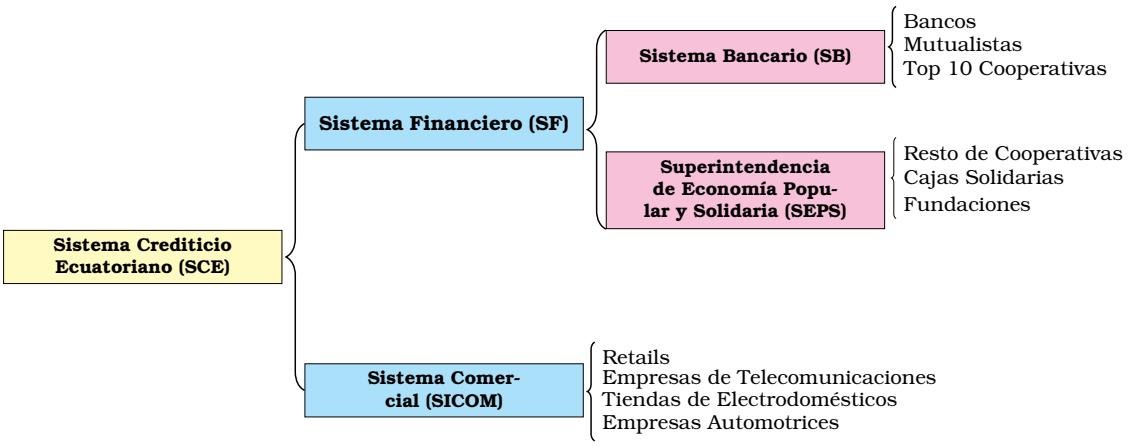

Figura 3.2: División de Sistemas de Crédito del Ecuador
Elaboración: El autor

Así, tras obtener registros únicos por individuo en cada punto de observación, se dispone de una base inicial de 45,159 registros.

3.5. Definición de la Variable Dependiente

3.5.1. Análisis de Población Inicial

Es importante recordar que los puntos de observación establecidos en el proyecto abarcan desde agosto de 2020 hasta julio de 2021. No obstante, la institución financiera proporcionó más puntos de observación de los solicitados, incluyendo información de los créditos abiertos en cada mes desde agosto de 2020 hasta marzo de 2022. Al limitarnos exclusivamente a los registros de los puntos de observación detallados en la tabla 3.1, obtenemos un total de 27,106 registros.

Adicionalmente, excluiremos del análisis a los créditos de carácter inmobiliario, debido a sus características particulares en el mercado. A su vez, excluiremos los créditos autoliquidables, ya que, en conjunto, estos dos tipos de créditos representan el 1.34 % de la información total y el objetivo principal es desarrollar modelos para operaciones de microcrédito y consumo, las cuales son las más representativas para la institución. Con esta selección, nos quedamos con 26,740 registros en la base de datos, que serán utilizados en los análisis posteriores.

3.5.2. Análisis de Roll-Rate

El análisis de Roll-Rate permite a la institución mantener una visión de la evolución de la morosidad a lo largo del tiempo, facilitando así el control del riesgo. Se centra en observar el deterioro en los días de vencimiento al pasar de un mes a otro, con el objetivo de definir una tolerancia de deterioro que clasifique a un individuo como mal pagador. Este análisis se realiza mensualmente dentro de la ventana de desempeño, generando 11 ventanas de comparación: m1 vs m2, m2 vs m3, . . . , m11 vs m12, las cuales se muestran en la sección de anexos, A.2.

Este análisis se realizó con aquellos individuos que tienen deuda en la ventana histórica del sistema crediticio ecuatoriano y aquellos con comportamientos de pago en al menos 6 meses de la ventana de desempeño. Esto se debe a que los modelos se construirán utilizando información histórica crediticia, por lo que es fundamental que las personas dispongan de datos de deuda en los últimos 3 años. Además, es necesario que cuenten con información suficiente en la ventana de desempeño para evitar subestimar las probabilidades de deterioro.

A continuación, se presenta la tabla general, la cual es el promedio de las tablas roll rate en las que aparece el rango de vencimiento a lo largo de las ventanas de comparación:

Rango Vencido	**Estado**	
	No Avanza (%)	**Avanza (%)**
Sin Vencido	94.22	5.78
De 1 a 30 días	87.60	12.40
De 31 a 60 días	54.71	45.29
De 61 a 90 días	33.76	66.24
De 91 a 120 días	18.52	81.48
De 121 a 150 días	14.20	85.80
De 151 a 180 días	8.10	91.90
Más de 180 días	3.51	96.49

Cuadro 3.3: Tabla Roll-Rate general
Elaboración: El autor

Así, para el rango de días de vencimiento de 61 a 90 días, se utiliza la información de las últimas 9 tablas, mientras que para el rango de 121 a 150 días, se toma la información de las últimas 7 tablas. Donde, los resultados de la tabla general nos indican la probabilidad de deterioro

o no del estado actual de los días de vencimiento. Donde, el análisis de la tabla 3.3 es el siguiente: El 12.40 % de las personas con 1 a 30 días de vencimiento en el mes de análisis continúan sin pagar el crédito en el siguiente mes, mientras que el 81.48 % de las personas en el rango vencido de 91 a 120 días en el mes de observación empeoran su situación al mes siguiente. Es decir, de cada 10 personas en este rango en el mes de observación, 8 continúan sin pagar su deuda en el siguiente mes.

Es interesante destacar los extremos de la tabla, la cual indica que de cada 100 clientes sin días de vencimiento en el punto de observación, aproximadamente 6 incumplen su deuda en el mes siguiente. En contraste, de cada 100 clientes con más de 180 días de vencimiento, aproximadamente 96 continúan sin pagar su deuda, lo cual tiene sentido ya que son clientes que llevan más de seis meses sin pagar su crédito y es muy probable que continúen con ese comportamiento negativo.

En donde, en la tabla 3.3, se observa que el umbral a partir del cual se registra un 50 % o más de deterioro es en el intervalo de 61 a 90 días. Por lo tanto, se establece como criterio para identificar a un mal pagador a aquella persona que presenta 61 o más días de vencimiento en la ventana de desempeño.

Ahora bien, es importante notar que el rango de 31 a 60 días ya muestra un avance del 45.29 %; sin embargo, se opta por la categoría de 61 a 90 días. Esto se debe a que, como se muestra en los anexos, A.2, de las 11 ventanas de comparación, únicamente en la primera no se encuentra un rango con un avance superior al 50 %. En las tres siguientes, se observa que a partir de 31 días de vencimiento ya se experimenta un 50 % de deterioro en el pago. No obstante, en las últimas siete ventanas, el rango que supera el 50 % de avance se registra a partir de los 61 días de vencimiento o más.

Por este motivo, se decide considerar como indicador de mal pagador a aquellos individuos con 61 o más días de vencimiento, ya que en la mayoría de las ventanas de comparación se observa el avance del 50 % a partir de los 61 días de vencido. Además, estas ventanas resultan ser las últimas, donde los patrones de pago tienden a estabilizarse.

Resultaría interesante observar cómo influye en el cálculo de las pérdidas esperadas que se considere un rango de días de vencimiento menor para clasificar a un mal pagador, siempre y cuando esté respaldado por fundamentos técnicos. Sin embargo, este enfoque va más allá de los objetivos establecidos para este proyecto y podría ser objeto de análisis en proyectos posteriores.

3.5.3. Variable Dependiente

En el proyecto, es importante predecir si un individuo es un buen o mal pagador, ya que esto determina su probabilidad de incumplimiento. En función de los resultados del análisis de Roll-Rate previo, sección 3.5.2, se detallan a continuación las variables que permiten clasificar un individuo:

- **MARCA BANCARIZADO**: Me indica si una persona tuvo deuda o no en el sistema crediticio ecuatoriano en los últimos 36 meses previos al punto de observación.
- **DESEMPEÑO 6M**: Me indica si una persona tiene o no comportamientos de pago en al menos 6 meses de la ventana de desempeño.
- **MAX NDIAS MOROSIDAD 12M**: Me indica el máximo número de días de morosidad alcanzado en toda la ventana de desempeño por el individuo.

Así, la identificación de un individuo se rige por el esquema mostrado en la figura 3.3.

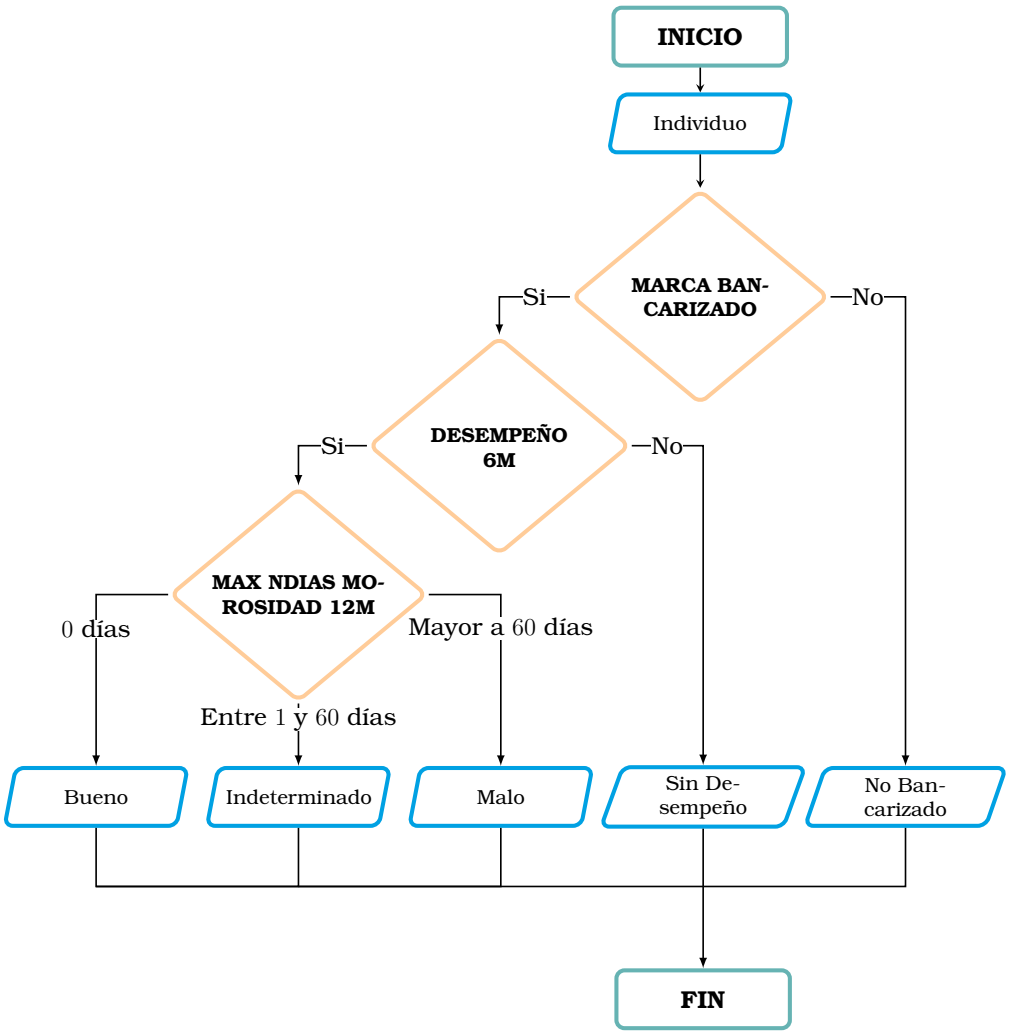

Figura 3.3: Esquema: Variable Dependiente
Elaboración: El autor

Donde se utilizó la siguiente notación para cada tipo de cliente en la variable dependiente, VarDep, con la subsiguiente distribución de registros:

Tipo Cliente	VarDep	Registros	%
Bueno	0	10,308	38.54
Malo	1	1,274	4.76
Indeterminado	2	5,182	19.37
Sin desempeño	4	7,528	28.15
No Bancarizado	5	2,455	9.18
Total		**26,747**	**100**

Cuadro 3.4: Distribución de Registros
Elaboración: El autor

3.5.4. Especificación de la población de estudio

Para la población de estudio final, se excluyó a los individuos no bancarizados de la base de datos, es decir, aquellos sin historial crediticio en los 36 meses anteriores a la entrega del crédito en el sistema crediticio ecuatoriano. Estos individuos no pueden ser segmentados, ya que los modelos a construir dependen de información crediticia, la cual no poseen. Así, disponemos de 24,292 registros para el horizonte de tiempo previsto. A continuación, se presenta la distribución de registros con la que trabajaremos.

Tipo Cliente	VarDep	Registros	%
Bueno	0	10,308	42.43
Malo	1	1,274	5.25
Indeterminado	2	5,182	21.33
Sin desempeño	4	7,528	30.99
Total		**24,292**	**100**

Cuadro 3.5: Distribución de registros final
Elaboración: El autor

3.6. Muestra de Modelamiento y Validación

Debido a la cantidad de registros disponibles, se decidió tomar el 50 % de los registros para el modelamiento y otro 50 % para la validación. Este proceso se realizó mediante un muestreo aleatorio simple sin reemplazo. En la tabla 3.6 se presenta la distribución resultante de los registros en las bases de Train y Test.

VarDep	Información Modelamiento		Información Validación	
	Registros	%	Registros	%
0	5,148	42.38	5,160	42.48
1	623	5.13	651	5.37
2	2,608	21.47	2,574	21.19
4	3,767	31.02	3,761	30.96
Total	**12,146**	**100**	**12,146**	**100**

Cuadro 3.6: Distribución de la Información de las bases de Train & Test
Elaboración: El autor

Donde, para entrenar los modelos, se tomarán únicamente los registros clasificados como buenos o malos de la información de modelamiento. Sin embargo, para generar las tablas performance, se tendrán en cuenta todas las categorías, tanto en la base de modelamiento como en la de validación. A continuación, se presenta la distribución de la información con la cual se entrenarían los modelos.

VarDep	Información Entrenamiento	
	Registros	%
0	5,148	89.20
1	623	10.80
Total	**5,771**	**100**

Cuadro 3.7: Distribución de la Información de Entrenamiento
Elaboración: El autor

Con lo cual contaríamos con 5,771 registros para entrenar los modelos, cantidad que supera los 4,000 registros estándar para entrenar un modelo de Credit Scoring.

3.7. Balanceo de Categorías

En la tabla 3.7, se evidencia un desbalanceo entre las categorías de buenos y malos pagadores. Para abordar esto, se optó por rebalancear la variable dependiente, tomando aleatoriamente un registro por cada 5 registros en la base de entrenamiento actual. Este enfoque incrementa el tamaño de la muestra de entrenamiento inicial en un 20%. El proceso de rebalanceo se llevó a cabo mediante un muestreo aleatorio simple con reemplazo, buscando obtener una proporción del 80-20 entre buenos y malos pagadores, respectivamente. A continuación, se muestra la distribución de la información con la que finalmente se entrenó los modelos:

VarDep	Información Rebalanceada	
	Registros	%
0	5,540	80
1	1,385	20
Total	**6,925**	**100**

Cuadro 3.8: Distribución de la Información Rebalanceada
Elaboración: El autor

3.8. Análisis de las Variables

En la base de datos inicial, contamos con un total de 579 variables. Sin embargo, se opta por excluir las variables correspondientes a la ventana de desempeño en el análisis subsiguiente. Estas variables específicas no contribuyen al desarrollo de modelos, ya que su función se limita a la definición de la variable dependiente, VarDep. Al realizar esta exclusión, nos quedamos con un conjunto de 445 variables que siguen en su mayoría el formato genérico:

XXX_S_tM

En esta notación, XXX representa el nombre de la variable. La letra S identifica uno de los sistemas descritos en la Figura 3.2, al cual pertenece la información. Por último, t indica el número de meses anteriores al punto de observación al que corresponde la información. Por ejemplo, la variable NOPE_VENC_SF_3M indica el número de operaciones vencidas en el sistema financiero en los últimos 3 meses anteriores al punto de observación.

Es importante destacar que la información de periodos más cortos está contenida en la información de periodos más grandes. Por ejemplo, el número de operaciones vencidas en el sistema financiero en los últimos 3 meses será menor o igual al número de operaciones vencidas en el sistema financiero en los últimos 6 meses.

Ahora bien, en cuanto a los periodos considerados para la recopilación de información, se seleccionaron intervalos de 3, 6, 12, 24 y 36 meses. Esta elección se basa en la capacidad de estos periodos para reflejar los patrones de comportamiento de pago. Al abarcar tanto el corto plazo como el largo plazo, estos intervalos nos permitirán examinar posibles cambios en el comportamiento crediticio.

Donde la mayoría de la información a trabajar se clasifica en una de las categorías descritas en la tabla 3.9.

Tipo	SubTipo
Sociodemográficas	Vivienda, Provincia Actividad, Sexo, Estado Civil, ...
Estado Financiero	Total Activos, Total Patrimonio, Total Ingresos, Total Egresos, ...
Deuda	Información monetaria de Deuda: Total, Vencida, No Devenga Intereses, Por Vencer, Cartera Castigada, Demanda Judicial, Refinanciada
Días de Vencido	Máximo número de días Vencido
Operaciones	Número de Operaciones: Vencidas, Refinanciadas, Demanda Judicial, Cartera Castigada, Aperturadas, Vigentes
Entidades	Número de entidades con créditos: Vigentes, Demanda Judicial, Cartera Castigada, Vencidas
Acreedores	Número de Acreedores

Cuadro 3.9: Descripción de Información
Elaboración: El autor

Algo importante a aclarar es que si la variable tiene la palabra OP en su nombre, como en el caso de DEUDA_TOTAL_OP_12M, esto indica que la información proviene exclusivamente de la institución. Por lo tanto, la variable anterior nos proporciona datos sobre la deuda total en la institución durante los últimos 12 meses anteriores al punto de observación.

3.9. Selección de Sistemas de Crédito

Considerando los distintos tipos de sistemas de crédito presentes en la figura 3.2, se ha determinado que la información numérica discreta, como el número de operaciones vencidas, se analizará del SB o SF (SB+SEPS). Por otro lado, la información numérica continua, como los valores de deuda, se estudiará para todo el SCE (SICOM+SF).

Esta elección se fundamenta en la siguiente razón: al considerar las variables discretas de todo el SCE, se estarían colocando al mismo nivel, por ejemplo, el número de operaciones vencidas en el SF y en SICOM. En la práctica, los montos de deuda vencida suelen ser más pequeños en SICOM en comparación con el SF. Por ejemplo, una persona puede tener una operación vencida de 20$ en una empresa de televisión por cable,

mientras que otra persona podría tener una igualmente, pero en una cooperativa por un monto de 400$. Al considerar todo el SCE, estaríamos poniendo a estas 2 operaciones vencidas como comparables.

Esta equiparación podría perjudicar o beneficiar a distintas personas y resultar en interpretaciones incorrectas en los modelos. Un caso similar se presenta con el número de acreedores; comparar, por ejemplo, un acreedor de SICOM (como un retail de ropa) con uno del SF (como un Banco) podría ser problemático, ya que los créditos en un Banco son más altos en general que los otorgados por un retail. Este enfoque se aplicó de manera análoga a las variables creadas, con el objetivo de evitar distorsiones en la interpretación de los datos y buscando una representación más precisa de la realidad financiera.

3.10. Creación de Variables

Ahora bien, usando la información disponible, se decidió crear las siguientes variables en función de lo explicado en la sección anterior:

3.10.1. Variables de Marca

Las Variables de Marca son indicadores binarios que permiten determinar si una persona posee o no una característica o comportamiento crediticio de interés. Por ejemplo, la variable MARCA_DVEN_SB_3M, indica si una persona tiene o no días de vencimiento en el sistema bancario en los últimos 3 meses previos al punto de observación. Su creación se describe de la siguiente manera:

$$\text{MARCA\_DVEN\_SB\_3M} = \begin{cases} 1 & \text{si MAX\_DVEN\_SB\_3M} > 0 \\ 0 & \text{si MAX\_DVEN\_SB\_3M} = 0 \end{cases}$$

Así, se observa el siguiente comportamiento en la variable creada:

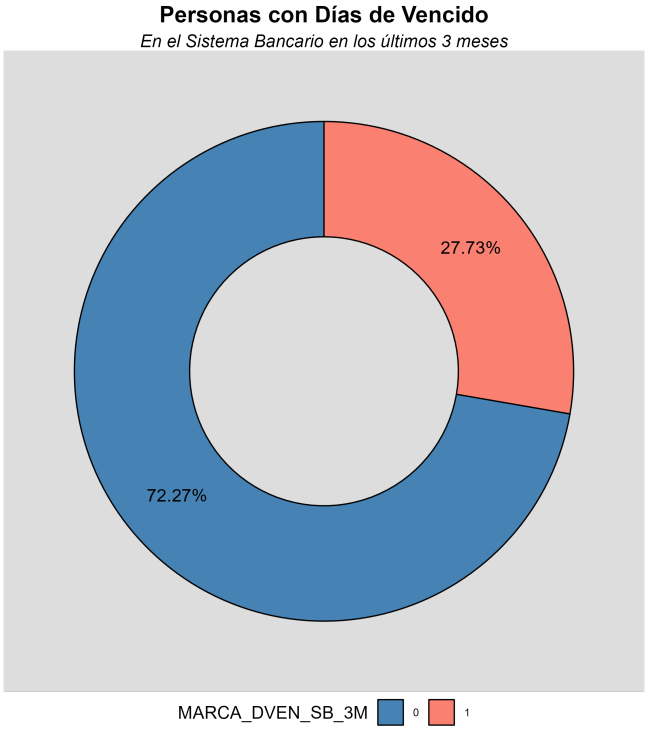

Figura 3.4: Gráfico de Anillo: Marca Vencido
Elaboración: El autor

En la figura 3.4 se indica que, en toda la población de estudio, aproximadamente 28 de cada 100 personas tienen días de vencido en el sistema bancario en los últimos 3 meses. Así, es más probable que un individuo al azar esté al día con sus créditos en los últimos 3 meses en el sistema bancario.

Así, se han creado otras variables de Marca para diversas características de interés. Por ejemplo, se diseñó una variable para identificar a personas que tenían o no deuda en cartera castigada o demanda judicial en el sistema crediticio ecuatoriano durante toda la ventana histórica, con el fin de identificar a individuos con muy malos hábitos de pago.

Por lo cual, la relevancia de las variables de Marca reside en su capacidad para identificar individuos con características específicas de interés para la institución. Esto posibilita la definición de políticas de colocación de créditos. Por ejemplo, se podría optar por no otorgar créditos a personas con deuda en Cartera Castigada o con Demanda Judicial, mientras que se podrían establecer políticas de gestión del riesgo, como ofrecer montos de crédito más elevados a aquellos individuos sin días de vencimiento.

3.10.2. Ratios

Se ha decidido construir ratios con el propósito de evaluar la variación en la información crediticia a lo largo del tiempo. Los ratios calculados comparan la información de distintos periodos: los últimos 3 meses con los últimos 6, los últimos 3 con los últimos 12, los últimos 6 con los últimos 12, los últimos 12 con los últimos 24 y los últimos 12 con los últimos 36 meses al punto de observación. La elección de estos periodos se fundamenta en su capacidad para reflejar cambios tanto a corto como a largo plazo en la situación crediticia, capturando así más información que si se analizara cada variable de manera individual. Por ejemplo, se ha creado el siguiente ratio:

$$\text{r\_DEUDA\_TOTAL\_SF\_3a6M} = \begin{cases} \dfrac{\text{DEUDA\_TOTAL\_SF\_3M}}{\text{DEUDA\_TOTAL\_SF\_6M}} & \text{si: DEUDA\_TOTAL\_SF\_6M} > 0 \\ 0 & \text{si: DEUDA\_TOTAL\_SF\_6M} = 0 \end{cases}$$

Este ratio ofrece información sobre la variación en la deuda total de un individuo en el sistema financiero entre los últimos 3 y 6 meses, permitiéndonos evaluar si la persona ha aumentado o reducido su endeudamiento en ese período. La comparación se realiza considerando siempre la información de un periodo más corto en relación con uno más extenso. Dado que la información de los periodos cortos está contenida en la de los periodos largos, los valores de estos ratios estarán entre 0 y 1. Un ratio cercano a 1 indicaría un empeoramiento en la situación crediticia de la persona, ya que ha adquirido este comportamiento en los periodos más recientes. En cambio, un ratio cercano a 0 señalaría que la persona tenía ese comportamiento crediticio en los periodos más distantes, sugiriendo una mejora en su situación financiera.

Para ilustrar esto, examinemos el ratio de deuda para dos individuos, cuya información se muestra en el siguiente esquema:

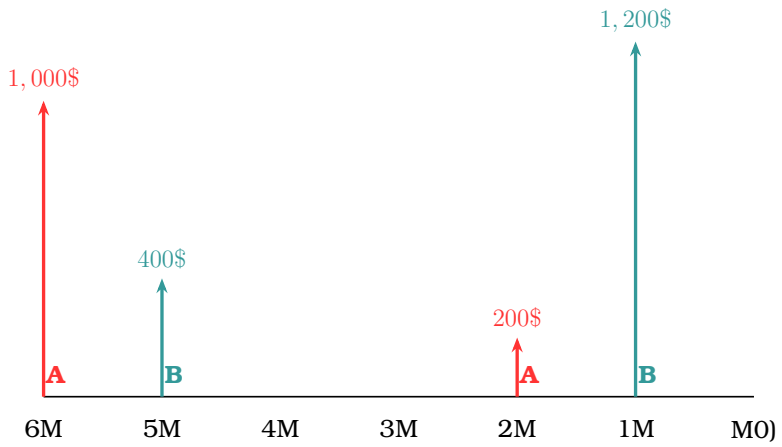

Figura 3.5: Esquema: Ejemplo Ratio de Deuda
Elaboración: El autor

Los resultados de estos individuos se presentan en la siguiente tabla:

Individuo	Deuda 3M ($)	Deuda 6M ($)	r_Deuda_3a6M
A	200	1,200	0.17
B	1,200	1,600	0.75

Cuadro 3.10: Ejemplo Ratios de Deuda
Elaboración: El autor

En este ejemplo, observamos lo descrito anteriormente. Mientras el ratio sea más cercano a 0, implica una mejora en la situación de la persona. En este caso, la persona A contrajo menos deuda en los últimos 3

meses en comparación con los anteriores 3 . Por otro lado, cuanto más cercano a 1 sea el ratio, significa que la situación de la persona ha empeorado, como en el caso de la persona B, que ha adquirido más deuda en el último trimestre en comparación con el anterior.

La importancia de estos ratios reside en su capacidad para registrar variaciones en la información crediticia a lo largo del tiempo. Este análisis facilita la identificación de individuos cuya situación financiera ha empeorado o mejorado, permitiendo a la institución establecer estrategias de control del riesgo y cobranza. Por ejemplo, realizar seguimiento de pagos a personas con indicadores significativamente altos en ratios de Deuda, ya que este grupo demuestra una propensión elevada a endeudarse, aumentando el riesgo de descuidar sus pagos.

3.10.3. Variables de Probabilidad

Se optó por agrupar variables, tanto numéricas como categóricas, para crear categorías relacionadas al riesgo de ser considerado mal pagador. Esta estrategia facilita la identificación de patrones, permitiendo determinar, por ejemplo, en qué tipo de vivienda o rango de deuda existe mayor probabilidad de ser clasificado como mal pagador. Las categorías creadas se obtuvieron a través de un árbol de decisión $Y \sim X$. Donde, Y es la variable descrita en (2.1) y X la variable a categorizar. Es importante destacar que el análisis se realizó utilizando la información de buenos y malos de la base de modelamiento descrita en la tabla 3.7.

La implementación de esta agrupación se realizó utilizando la biblioteca `scorecard`. Este enfoque simplifica la creación de nuevas variables que describen el riesgo crediticio asociado a diversas características. A continuación, se presenta un ejemplo de la línea de código en el lenguaje de programación R para construir estas agrupaciones, en este caso, para el Tipo de Vivienda:

```
library(scorecard)
fc_tree_malos<-mod|> woebin(y='VarDep',x="TIPO_VIVIENDA",positive = 1,
   method = 'tree',bin_num_limit = 10)
```

En el código anterior:

- `mod`: Representa la base de datos, en nuestro caso, la base descrita en la tabla 3.7.
- `y`: Indica la variable categórica de interés, en nuestro caso, `Vardep`, la variable dependiente donde se registran los buenos y malos.
- `x`: Indica la variable de interés que queremos categorizar, en este caso, `TIPO_VIVIENDA`.
- `positive`: Indica la categoría de interés de la variable `y` en base a la cual se realizará la agrupación de la variable `x`. En nuestro ejemplo, es la categoría 1 de `VarDep`, que corresponde a la notación para los individuos malos.
- `method`: Indica el método utilizado para la agrupación. En este caso, se lleva a cabo mediante el método de árboles, `tree`, descrito en 2.3.
- `bin_num_limit`: Representa el número máximo de categorías a crear para la variable `x`. Es importante señalar que el algoritmo estima el número óptimo de categorías a crear. Si el número óptimo es menor al especificado en este parámetro, se construirán las agrupaciones consideradas óptimas.

Si se desea analizar más de una variable, se puede ingresar un vector de caracteres con las nombres de las variables de interés en el parámetro `x`. Además, si se desea agrupar todas las variables de la base de datos, se puede utilizar la línea anterior de código sin incluir el parámetro `x`. Para acceder a las agrupaciones creadas para cada variable, se puede hacer, por ejemplo, a través de `fc_tree_malos$TIPO_VIVIENDA`.

Cabe aclarar que se revisaron detenidamente los resultados generados por el algoritmo con el fin de asegurar que las agrupaciones fueran coherentes y reflejaran los comportamientos observados en la realidad. Por ejemplo, en la información socioeconómica, se descartaron agrupaciones que no tenían sentido, como aquellas que incluían personas con vivienda propia hipotecada y vivienda propia no hipotecada cuando se analizaba el tipo de vivienda. Esto se debe a que las segundas ya están pagando un crédito por su vivienda, y agregar otro crédito más podría llevar a posibles

incumplimientos de pago, beneficiándose así al estar en el mismo grupo que las personas con vivienda propia sin hipotecar. Asimismo, se verificó que, para los grupos de ingresos, la probabilidad de ser mal pagador disminuyera a medida que aumentaban los ingresos.

En relación a la información crediticia, se corroboró, por ejemplo, que a medida que aumentaba la deuda total, la probabilidad de ser mal pagador incrementara. Una vez realizadas las nuevas agrupaciones, se creó una variable asociada a la probabilidad de ser mal pagador según la pertenencia a cierta categoría. Los resultados obtenidos en el modelamiento se llevaron a toda la población. A continuación, se presentan dos ejemplos del proceso de creación de estas variables:

TIPO VIVIENDA

La recategorización del Tipo de Vivienda, una vez revisada, presenta la siguiente distribución de registros:

Tipo de Vivienda	Grupo	Registros	%
Propia No Hipotecada	1	3,326	57.63
Vive con familiares	2	1,874	32.47
Arrendada o Propia Hipotecada	3	568	9.84
NA	-	3	0.06
Total		**5,771**	**100**

Cuadro 3.11: Registros Tipo de Vivienda Recategorizada
Elaboración: El autor

Así, la probabilidad de ser catalogado como bueno o malo, dado el tipo de vivienda en la que reside la persona, se presenta a continuación:

Tipo de Vivienda	Grupo	Bueno (%)	Malo (%)
Propia No Hipotecada	1	90.74	9.26
Vive con familiares	2	88.15	11.85
Arrendada o Propia Hipotecada	3	83.80	16.20
NA	-	66.67	33.33

Cuadro 3.12: Probabilidades condicionales al Tipo de Vivienda
Elaboración: El autor

Así, según la tabla 3.12, la probabilidad de que una persona sea un mal pagador dado que vive con familiares es del 11.85 %. En contraste, de cada 100 personas que poseen una vivienda propia hipotecada o que arriendan, el 16.20 % son malos pagadores.

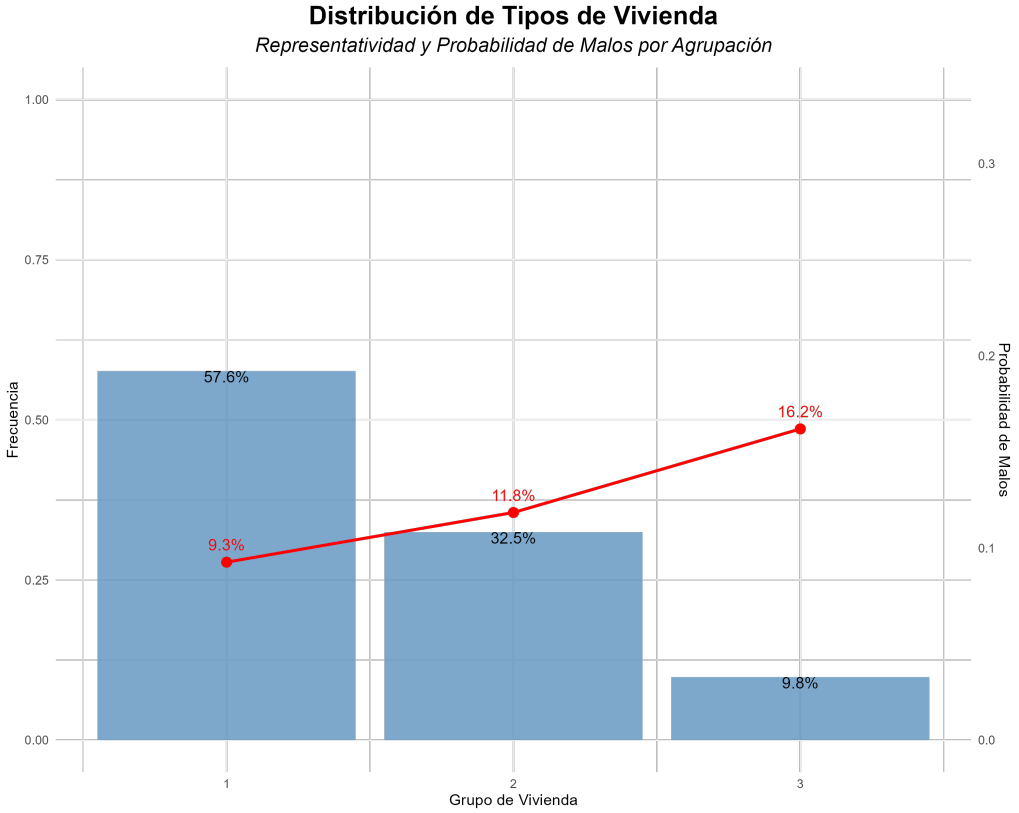

Figura 3.6: Distribución Categorizada: Tipo Vivienda
Elaboración: El autor

En la figura 3.6, se presenta la distribución visual de los grupos, destacando la representatividad de cada categoría mediante barras, y la probabilidad de ser mal pagador dada la categoría, representada por líneas. Se observan diferencias significativas en estas probabilidades entre los distintos grupos. Esta visualización simplifica la identificación de los grupos con mejores y peores pagadores, proporcionando información crucial para ajustar estratégicamente las políticas de colocación de créditos.

Así, se creó la variable prbm_VIVIENDA en base a la información de la tabla 3.12. Esta variable indica la probabilidad de ser mal pagador dado el tipo de vivienda en la que habita la persona. En caso de que exista una categoría no representativa, es decir, que no aparezca en la base de modelización y, por tanto, no forme parte del análisis anterior, se

le asignará la probabilidad de ser mal pagador del peor caso, adoptando una postura conservadora. Este enfoque también se aplicó a los valores NA. La misma postura se ha adoptado para las demás variables creadas a partir de variables sociodemográficas. La variable resultante de este proceso es:

$$\text{prbm\_VIVIENDA} = \begin{cases} 0.0926 & \text{Si la vivienda es Propia No Hipotecada} \\ 0.1185 & \text{Si la persona vive con familiares} \\ 0.1620 & \text{Si la vivienda es de Otro tipo o NA} \end{cases}$$

DEUDA_VENCIDA_SCE_12M

Para las variables numéricas, se procedió de la misma manera. Además, se analiza lo que sucede cuando el individuo no tiene ese comportamiento crediticio, es decir, la información crediticia de interés es cero. Por ejemplo, el primer intervalo que proporciona el árbol de decisión para la deuda vencida en el sistema crediticio en los últimos 12 meses es el intervalo de personas que tienen deuda entre $[0, 100)$ dólares. Este intervalo tiene una probabilidad de ser malo del 6.2 %. Al analizar lo que sucede por separado con los individuos sin deuda vencida y aquellos con deuda entre $(0, 100)$ dólares, obtuvimos los siguientes resultados, que fueron con los que nos quedamos:

Rango de Deuda Vencida ($)	Grupo	Registros	%
0	1	3,920	67.92
$(0, 100)$	2	531	9.20
$[100, 260)$	3	428	7.42
≥ 260	4	892	15.46
Total		**5,771**	**100**

Cuadro 3.13: Registros Deuda Vencida SCE últimos 12 meses
Elaboración: El autor

Así, las probabilidades de ser catalogado como bueno o malo, dado el rango de deuda en el que se encuentra la persona, se detallan en 3.14. Por lo cual, se puede ver que si poníamos a los individuos del grupo 1 y 2 juntos, como recomendaba el algoritmo, se iba a beneficiar a aquellos individuos que tenían deuda vencida pero era inferior a 100$.

Rango de Deuda Vencida ($)	Grupo	Bueno (%)	Malo (%)
0	1	94.08	5.92
$(0, 100)$	2	92.09	7.91
$[100, 260)$	3	82.48	17.52
≥ 260	4	69.28	30.72

Cuadro 3.14: Probabilidades Deuda Vencida SCE últimos 12 meses
Elaboración: El autor

Así, en el gráfico 3.7 se puede observar cómo, a medida que la deuda vencida aumenta, la probabilidad de ser malo también aumenta, lo cual va acorde con la lógica del comportamiento crediticio. Esto se verificó para todas las variables de probabilidad creadas en base a información crediticia.

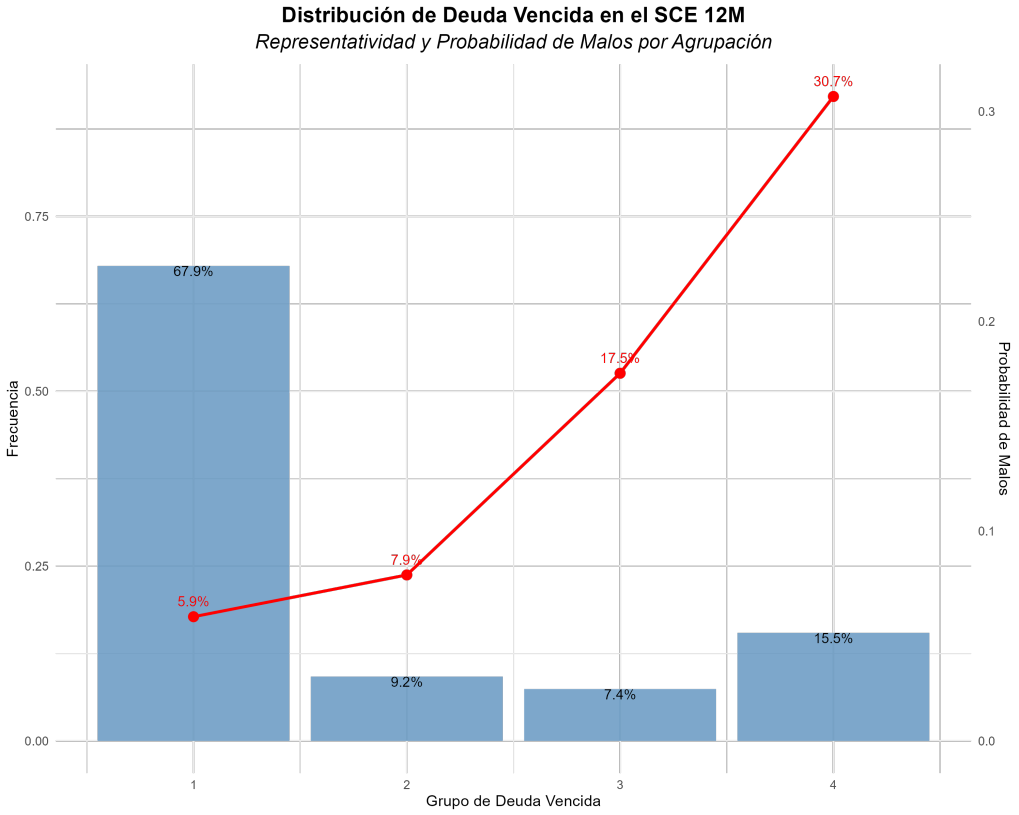

Figura 3.7: Distribución Categorizada: Deuda Vencida SCE 12M
Elaboración: El autor

Así, la variable que se creó en base a la información de la tabla 3.14 fue:

$$\text{prbm\_DEUDA\_VENCIDA\_SCE\_12M} = \begin{cases} 0.0592 & \text{Si: DEUDA\_VENCIDA\_SCE\_12M } = 0 \\ 0.0791 & \text{Si: } 0 < \text{DEUDA\_VENCIDA\_SCE\_12M} < 100 \\ 0.1752 & \text{Si: } 100 \leq \text{DEUDA\_VENCIDA\_SCE\_12M} < 260 \\ 0.3072 & \text{Si: DEUDA\_VENCIDA\_SCE\_12M} \geq 260 \end{cases}$$

Es importante señalar que en ocasiones se trabaja con la probabilidad de ser catalogado como bueno, dado que se pertenece a una categoría específica. Esta probabilidad se obtiene, por ejemplo, como:

$$\text{prbb\_DEUDA\_VENCIDA\_SCE\_12M} = 1 - \text{prbm\_DEUDA\_VENCIDA\_SCE\_12M}$$

La cual indicaría la probabilidad de ser catalogado como bueno, dado el rango de deuda vencida al que pertenezco en el sistema crediticio ecuatoriano en los últimos 12 meses.

Por lo tanto, las variables de probabilidad se diseñan con el propósito de identificar grupos de riesgo en lugar de analizar individualmente cada categoría o el espectro completo de valores de la variable. Este enfoque tiene varias ventajas, entre ellas facilitar la toma de decisiones, por ejemplo en la implementación de políticas de colocación para individuos con bajas probabilidades de ser malos.

Además, la creación de estas variables de probabilidad tiene el beneficio adicional de reducir los grados de libertad en los modelos. Como ejemplo, para agregar a los modelos los grupos generados para la deuda vencida en el Sistema Crediticio Ecuatoriano en los últimos 12 meses, requeriríamos la inclusión de tres variables binarias asociadas a los rangos creados. En cambio, mediante el uso de las variables de probabilidad, podemos lograr el mismo análisis con una sola variable, simplificando así la complejidad del modelo.

Otro beneficio de utilizar estas variables es que permite realizar una imputación indirecta a las variables originales, como se realizó en el caso de la variable prbm_VIVIENDA. En este caso, a los datos con NA en el tipo de vivienda se les asignó la probabilidad de ser mal pagador del peor caso. Así, tras completar este proceso, la base de datos resultante cuenta con 805 variables y los 24,292 registros mencionados anteriormente.

3.11. Selección de Variables

El análisis para la selección de variables se llevó a cabo con la información de buenos y malos de la base de modelamiento, descrita en la tabla 3.7. En este proceso, se siguieron los siguientes pasos: en primer lugar, se excluyeron las variables con más del 30 % de valores faltantes, ya que imputar estos valores podría introducir sesgos debido a su elevado porcentaje de NA's. De esta manera, nos quedamos con 799 variables, lo que representa un descarte del 0.75 % del total de variables.

Una vez identificadas las variables con un bajo porcentaje de valores faltantes, se excluyeron del análisis aquellas variables numéricas constantes o aquellas variables categóricas en las cuales el 99 % o más de la información se concentraba en una sola categoría. La razón detrás de esta exclusión es la necesidad de contar con variables que presenten variación, ya que estas son las que pueden distinguir entre los comportamientos de un buen y un mal pagador. Luego de llevar a cabo este proceso, nos quedamos con 760 variables. En consecuencia, de las 805 variables iniciales, se ha reducido la información en un 5.59 %.

Ahora bien, es importante destacar que, entre las 760 variables, únicamente 4 presentan valores faltantes: PROVINCIA DE ACTIVIDAD, CANTON DE ACTIVIDAD, NUMERO DE CARGAS y TIPO DE VIVIENDA. Estas variables fueron imputadas de manera indirecta durante la creación de variables, así como otras variables sociodemográficas. Por otro lado, la información crediticia ya no contiene valores faltantes debido a consideraciones iniciales. En dicha consideración, se estableció que la ausencia de un valor (NA) en una variable de este tipo indicaba la falta de comportamiento crediticio correspondiente, y, por ende, se imputaron dichos casos con cero.

Así, una vez identificadas las variables válidas, aquellas sin un porcentaje alto de valores faltantes y no constantes, se procedió a ejecutar el test KS y VI para variables numéricas y categóricas, respectivamente. Inicialmente se seleccionaron las 5 variables con mayor valor de KS o VI para cada tipo de variable, descritas en la tabla 3.9. Esta elección se realizó con la finalidad de obtener información diversa sobre diferentes aspectos del individuo, evitando centrarnos exclusivamente en un solo

tipo de información. Se tomaron las 5 variables de acuerdo a lo mencionado en la sección 3.9, siendo la información de deuda monetaria de todo el Sistema de Crédito Ecuatoriano (SCE), mientras que otro tipo de información crediticia se obtuvo solo del Sistema Financiero (SF) o Sistema Bancario (SB).

Así, las variables finalmente seleccionadas en los modelos resultan de la combinación encontrada de variables de diferente índole, las cuales proporcionaron el índice de condicionamiento más bajo de la matriz de correlaciones, a la vez que redujeron el índice de estabilidad poblacional (IPS) del modelo. Estas variables se detallan en la sección de anexos, A.3. Así, el modelo de Regresión Logística se entrenó con 8 variables mientras que los modelos de Random Forest y XGBoost se entrenan con 11 variables.

3.12. Modelos de Credit Scoring

Como se explicó en la sección 3.1, el primer paso consiste en determinar si una persona es buena o mala pagadora, para luego establecer su probabilidad de incumplimiento. Los modelos entrenados estiman estas probabilidades de ser un buen o mal pagador. A continuación, se describe el proceso mediante el cual se llevó a cabo el entrenamiento de cada modelo utilizando la base de datos detallada en la tabla 3.8.

3.12.1. Modelo Logit

Dada la notación utilizada para los tipos de clientes (0 para los individuos buenos y 1 para los individuos malos) en la variable Vardep, el modelo de regresión logística estima la probabilidad de ser un mal pagador. Dado que este modelo no tiene hiperparámetros, procedemos a explicar su implementación utilizando las variables descritas en la tabla A.12.

Implementación del modelo

La implementación del modelo de credit scoring basado en el modelo logit se realizó en el lenguaje estadístico R utilizando la biblioteca `stats`, que es una biblioteca base de este lenguaje. La modelización se llevó a cabo mediante la siguiente línea de código:

```
modelo <- glm(formula = as.formula(formula), family = binomial("logit")
    , data = rmod)
```

En el código anterior:

- `formula`: Describe la relación entre las variables del modelo; en nuestro caso, será algo similar a `"VarDep ~ x + z"`, donde `VarDep` es la variable dependiente, y `x, z` son variables independientes.
- `family`: Determina la distribución y la función de enlace del modelo. Aquí, la distribución binomial con función de enlace logit.
- `data`: Especifica la base de datos con la que se entrena el modelo; en nuestro caso, la muestra rebalanceada descrita en la tabla 3.8.

Tabla de Coeficientes

Variable	Estimación	Std. Error	Z valor	Sign
Intercepto	-0.0993	0.3858	-0.2573	0.7970
prbm_PROVINCIA_ACTIVIDAD_DES	5.7769	0.5749	10.0486	0.0000
prbm_TOTAL_INGRESOS	18.3105	1.7350	10.5536	0.0000
MARCA_DVEN_SF_3M	1.6043	0.0773	20.7553	0.0000
r_NUM_OPE_VIG_SF_12s36M	0.7997	0.1308	6.1122	0.0000
prbm_NOPE_APERT_OP_24M	2.3842	0.7632	3.1239	0.0018
prbb_DEUDA_VENCIDA_SCE_36M	-6.6538	0.3717	-17.8989	0.0000
prbm_DEUDA_TOTAL_OP_12M	3.0972	0.5069	6.1096	0.0000
prbm_NUM_ENT_VIG_SB_12M	1.9366	0.6672	2.9024	0.0037

Cuadro 3.15: Tabla de coeficientes del Modelo Logit
Elaboración: El autor

En la tabla 3.15, se observa que todas las variables, excepto el intercepto, son significativas al 95 % de confianza. Un valor del intercepto cercano a cero es deseable, ya que representa todo lo que nuestro modelo no captura. Esto sugiere que estamos capturando la información necesaria para clasificar a individuos entre buenos y malos pagadores. El modelo consta de un total de 8 variables, donde una de ellas premia (con signo negativo) y reduce la probabilidad de ser un mal pagador, mientras que las otras 7 castigan (con signo positivo) y aumentan la probabilidad de ser un mal pagador. Además, se observa consistencia en los signos de las variables incluidas en el modelo. También se indica que la información abarca varios periodos, aunque mayormente del último año.

Importancia de las Variables

Al analizar la importancia de las variables, basada en el valor absoluto del estadístico Z de la tabla 3.15, se observa el comportamiento descrito en la figura 3.8. En dicha figura, se destaca que la variable que más contribuye al modelo es MARCA_DVEN_SF_3M .Esto implica que tener o no días de vencimiento en el sistema financiero en los últimos 3 meses previos al punto de observación es determinante para estimar la probabilidad de que alguien sea un mal pagador. Le sigue la variable que favorece a las personas que tienen un monto menor de deuda vencida en el sistema crediticio ecuatoriano en los últimos 3 años.

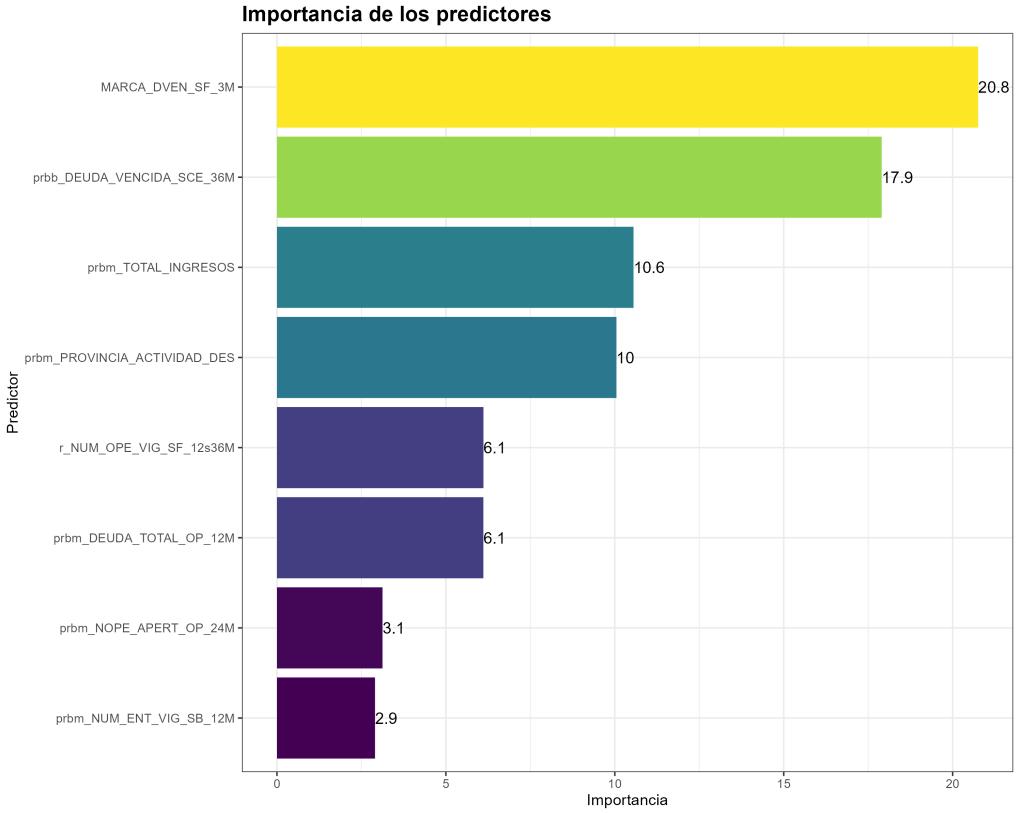

Figura 3.8: Importancia Variables - Modelo: RGL
Elaboración: El autor

Análisis de Correlación

Se puede observar que en la matriz del modelo ajustado, gráfico 3.9, la máxima correlación en valor absoluto es de 0.6, y la tienen las variables r_NUM_OPE_VIG_SF_12s36M y prbm_NUM_ENT_VIG_SB_12M. Esto podría deberse a que ambas variables consideran comportamientos crediticios vigentes relacionados a los últimos 12 meses y trabajan al menos con información del sistema bancario. Sin embargo, dado que el índice de condicionamiento para la matriz de correlación es 2.62, se puede descartar problemas de multicolinealidad.

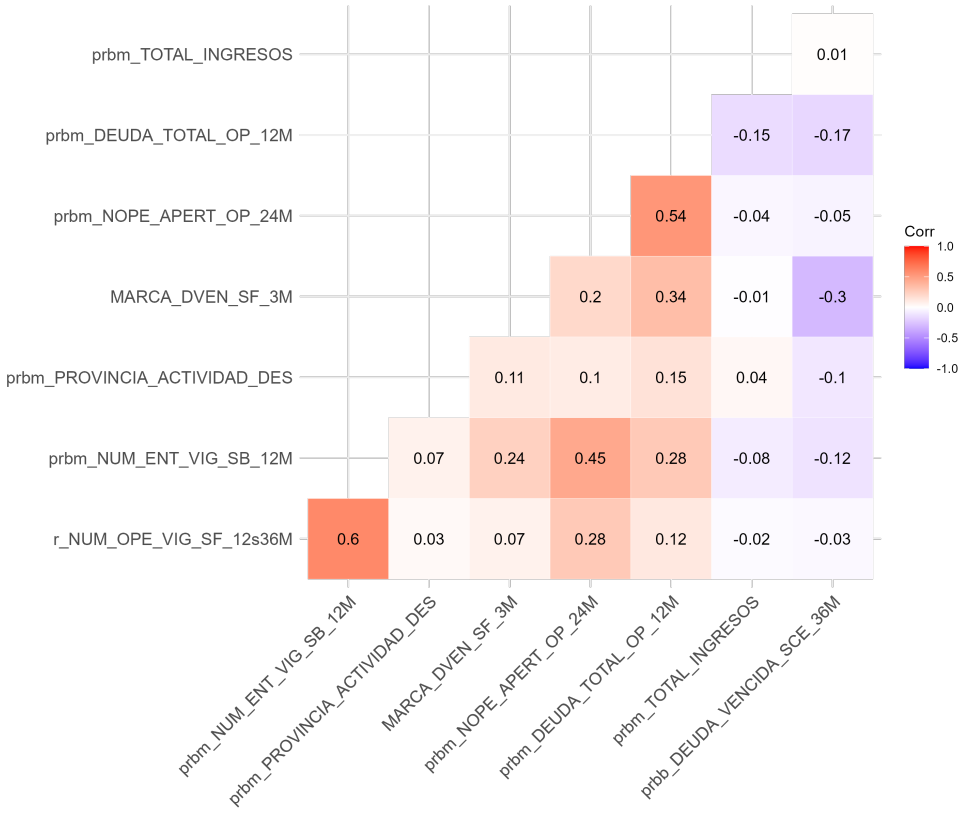

Figura 3.9: Matriz de Correlaciones - Modelo: RGL
Elaboración: El autor

3.12.2. Modelo Random Forest

El modelo de Random Forest estima simultáneamente la probabilidad de ser un mal y buen pagador. Dado que este modelo cuenta con hiperparámetros, primero procederemos a explicar el proceso de selección de los mismos:

Selección de Hiperparámetros

Para seleccionar los hiperparámetros del modelo, se construyó una grilla basada en: $\#mtry$, que indica el número de predictores seleccionados aleatoriamente para el análisis de partición de un nodo; $\#ntrees$, que establece la cantidad total de árboles a generar; y $\#min.node.size$, que define el número mínimo de registros requeridos en cada nodo hoja de cada árbol. Para modelos de clasificación, una elección común para $\#mtry$ es:

$$\#mtry = \sqrt{\#predictores} \tag{3.1}$$

Dado que el modelo se estima con 11 variables, descritas en la tabla A.13, $\#mtry \approx 3.23$, lo que llevó a la decisión de variar el número de predictores entre 3 y 4. Al considerar $\#mtry = 3$ y $\#mtry = 4$, se podrían tener $C_3^{11} = 165$ y $C_4^{11} = 330$ conjuntos de variables diferentes para analizar en cada nodo, respectivamente. Por lo tanto, en el peor de los casos, se tendrían 165 y 330 árboles diferentes, si para cada árbol se escogiera el mismo conjunto de variables en cada nodo. Así, para evitar árboles repetidos y prevenir problemas de sobreajuste, en la grilla construida, se optó por variar el número de árboles, $\#ntrees$, entre el 50 % y el 90 % de 165, con el objetivo de lograr decorrelación entre los árboles.

Además, dado que el modelo se entrenó con la base remuestreada descrita en la tabla 3.8, para asegurar la representatividad del nodo hoja, se varió $\#min.node.size$ entre el 4 % y el 10 % de 6,925, que es la cantidad de registros utilizada para entrenar los modelos.

Después de definir la grilla de hiperparámetros, se procedió a estimar los modelos mediante validación cruzada para cada combinación posible:

$$(\#mtry, \#ntrees, \#min.node.size)$$

Esto se realizó empleando el método de *k*-fold con *k* igual a 5, lo que implica dividir de manera aleatoria la base de entrenamiento en 5 particiones denominadas *folds*. Para cada combinación única de hiperparámetros, se ajustó un modelo cinco veces, utilizando cada uno de los cinco *folds* como conjunto de prueba una vez y los restantes como conjunto de entrenamiento. Luego, se evaluó la precisión (accuracy) en los *folds* de prueba, y al final se obtuvo el promedio de la precisión en los *folds* de validación para cada combinación única de hiperparámetros, la cual será la métrica de rendimiento mediante la cual se determina qué tan buena es una tripleta respecto a otra.

Este proceso generó un total de 70 combinaciones posibles de hiperparámetros, cada una con su respectivo *accuracy promedio*. Para seleccionar el modelo final, se optó por analizar las tripletas que se encontraban alrededor de la mediana de los *accuracy promedio* con el fin de evitar los sobreajustes que se tenían al analizar los modelos con las mejores métricas de rendimiento, estás combinaciones se encuentran en la sección A.4.1. Una vez completado este análisis, se eligió la tripleta que redujo el IPS del modelo, la cual fue:

$$(\#mtry, \#ntrees, \#min.node.size) = (3, 132, 485)$$

Implementación del modelo

La aplicación del modelo de credit scoring, basado en el algoritmo de random forest, se ejecutó utilizando el lenguaje estadístico R y la biblioteca `ranger`. El proceso de modelado se llevó a cabo mediante la siguiente línea de código:

```
modelo <- ranger(formula   = as.formula(formula),
  data       = rmod,
  classification = TRUE,
  probability = TRUE,
  importance = "impurity",
  replace = TRUE,
  num.trees =132,
  mtry       =3 ,
  min.node.size =485,
  seed       = 12345)
```

En el código anterior:

- `formula`: Representa la relación entre las variables del modelo; en nuestro caso, tomará una forma similar a `"VarDep ~ x + z"`, donde `VarDep` es la variable dependiente, y `x, z` son las variables independientes.
- `data`: Indica la base de datos utilizada para entrenar el modelo; en nuestro caso, se refiere a la muestra rebalanceada detallada en la tabla 3.8.
- `classification`: Especifica si estamos trabajando con un problema de clasificación (`TRUE`) o de regresión (`FALSE`).
- `probability`: Indica, en caso de trabajar con un modelo de clasificación, si se deben entregar las probabilidades de clasificación (`TRUE`) o la clase predicha (`FALSE`) para cada individuo.
- `importance`: Indica el método a través del cual se va a calcular la importancia de las variables. En este caso, `impurity` hace referencia al índice de impureza de Gini.
- `replace`: Indica si al momento de muestrear los datos se permiten reemplazos (`TRUE`) o no (`FALSE`).
- `num.trees`: Indica el número de árboles a generar en el bosque.
- `mtry`: Indica el número de predictores a tomar aleatoriamente para el análisis de división de un nodo.
- `min.node.size`: Especifica el número mínimo de registros que debe tener cada nodo hoja en cada árbol generado.
- `seed`: Permite fijar la semilla de aleatorización.

Importancia de las Variables

Al examinar la importancia de las variables en el modelo, fundamentada en el índice de impureza de Gini de cada variable para evaluar la homogeneidad en términos de clases y observar que variables aportan más en la reducción de la impureza de los nodos, se observa el siguiente patrón:

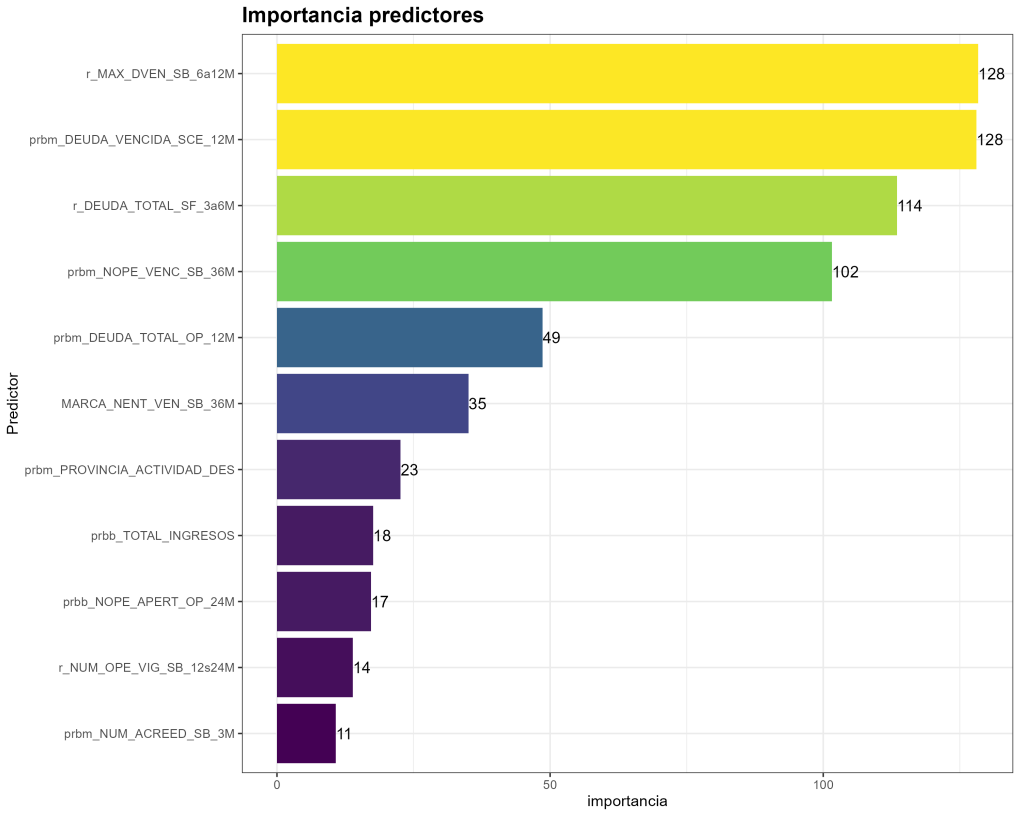

Figura 3.10: Importancia Variables - Modelo: RF
Elaboración: El autor

En el gráfico 3.10, se evidencia que hay dos variables igual de influyentes en el modelo: r_MAX_DVEN_SB_6a12M y prbm_DEUDA_VENCIDA_SCE_12M. Esto indica que ambas variables son de suma importancia para generar nodos en los cuales se puede identificar de manera clara a los individuos buenos de los malos, y viceversa. Es decir, estas variables contribuyen de manera significativa a la construcción de nodos que presentan proporciones distintivas de buenos o malos pagadores, en lugar de nodos con proporciones relativamente iguales.

Análisis de Correlación

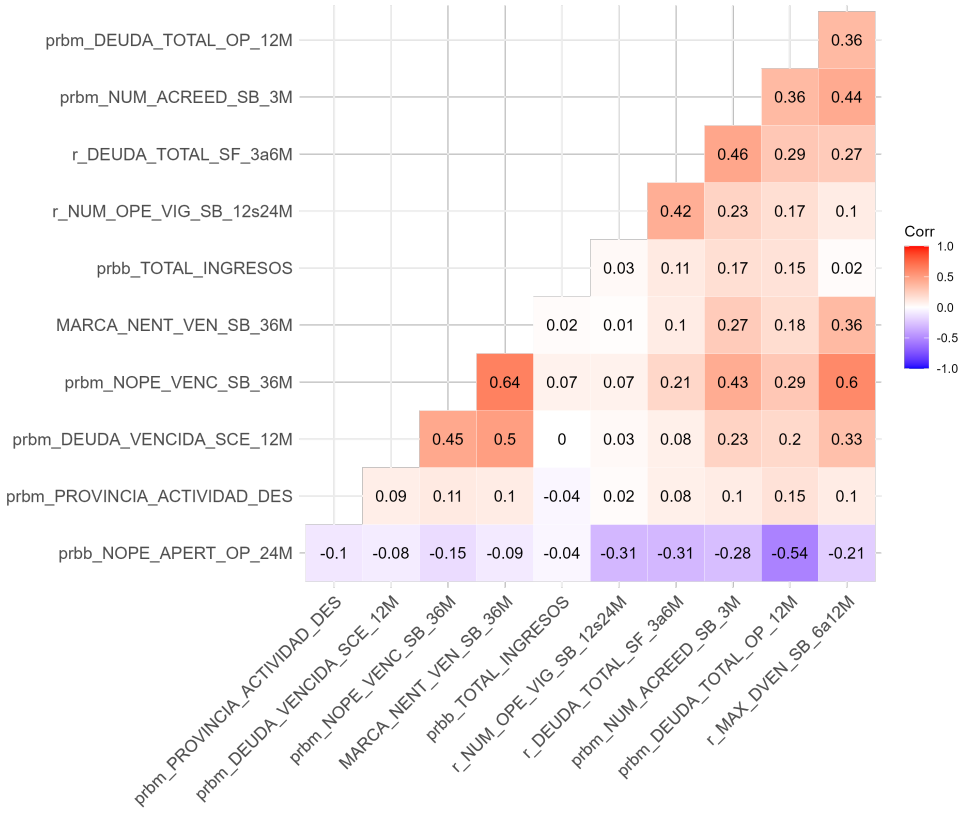

Figura 3.11: Matriz de Correlaciones - Modelo: RF
Elaboración: El autor

Resulta interesante notar, en la matriz de correlaciones del modelo ajustado, que la correlación entre las variables:

r_MAX_DVEN_SB_6a12M y prbm_DEUDA_VENCIDA_SCE_12M

Que resultan ser de mayor e igual importancia para el modelo, es de 0.36. Por lo tanto, no se observan problemas de multicolinealidad entre estas variables.

Además, se observa que la máxima correlación en valor absoluto es de 0.64, y esta se comparte entre las variables: prbm_NOPE_VENC_SB_36M y MARCA_NENT_VEN_SB_36M. La similitud podría deberse a que ambas variables consideran comportamientos crediticios vencidos relacionados con los últimos 36 meses y trabajan con información del sistema bancario. Sin embargo, dado que el índice de condicionamiento para la matriz de correlación es de 3.50, se descartan problemas de multicolinealidad.

3.12.3. Modelo XGBoost

El modelo de XGBoost estima la probabilidad de ser un mal pagador. Dado que este modelo cuenta con hiperparámetros, primero procederemos a explicar el proceso de selección de los mismos:

Selección de Hiperparámetros

Este modelo sigue un enfoque iterativo, donde cada iteración representa la construcción de un árbol. En otras palabras, si el modelo realiza, por ejemplo, 23 iteraciones, implica que se han creado 23 árboles, y cada uno aprende de los errores del anterior. En cada iteración, el objetivo es reducir la función de pérdida. En el caso de un modelo de clasificación binaria, se emplea la pérdida logarítmica, la cual penaliza las predicciones incorrectas. Por lo tanto, se busca reducir esta pérdida en cada iteración, ya que esto reflejaría una mejora en las predicciones del modelo.

Así, los hiperparámetros a estudiar para este modelo son: eta, la tasa de aprendizaje del modelo que controla la contribución de cada árbol (iteración) a la predicción final; $\#min\_child\_weight$, que define el número mínimo de registros requeridos en cada nodo hoja de cada árbol; $colsample\_bytree$, que indica el porcentaje de predictores seleccionados aleatoriamente en cada árbol; y $\#nrounds$, que indica el número de árboles (iteraciones) a realizarse.

Donde, para el ratio de aprendizaje del algoritmo, eta, se decidió variarlo entre 0.1 y 0.3, con el fin de controlar la contribución de cada árbol en el modelo final y evitar el sobreajuste. También, dado que el modelo se entrenó con la base remuestreada descrita en la tabla 3.8, para asegurar la representatividad del nodo hoja, se varió $\#min\_child\_weight$ entre el 4 % y el 10 % de 6,925, que es la cantidad de registros utilizada para entrenar los modelos.

Ahora bien, se utilizó la raíz cuadrada del número de predictores para determinar la cantidad de variables a seleccionar aleatoriamente para construir cada árbol. Dado que contamos con 11 predictores, descritos en la tabla A.14, según la ecuación (3.1), se debería probar entre 3 y 4 predictores para cada árbol. Sin embargo, debido a que este modelo re-

quiere un alto poder computacional, se decidió seleccionar directamente 3 predictores de manera aleatoria en cada árbol. Esta decisión se fundamenta también en los resultados obtenidos en el modelo de random forest, de esta manera:

$$colsample\_bytree = \frac{3}{11}$$

Por lo tanto, se podrían generar $C_3^{11} = 165$ árboles distintos, y es necesario no generar más árboles (iteraciones) en el modelo, ya que hacerlo podría resultar en árboles repetidos, lo que podría llevar a posibles sobreajustes.

Ahora bien, recordando que el modelo es iterativo y se basa en minimizar la función de pérdida, se decidió que el modelo realice todas las iteraciones (árboles) posibles, es decir, 165, para cada tripleta:

$$\left(eta, \#min\_child\_weight, colsample\_bytree = \frac{3}{11}\right)$$

Donde, se seleccionará el número de árboles, $\#nrounds$, en el cual se obtuvo la menor reducción de la función de pérdida para cada tripleta. Este proceso se lleva a cabo mediante validación cruzada, utilizando el método de *k*-fold con *k* igual a 5. En esta técnica, la base de entrenamiento se divide de manera aleatoria en 5 particiones denominadas *folds*. Para cada combinación única de hiperparámetros, se ajusta un modelo cinco veces. En cada iteración, se utiliza uno de los cinco *folds* como conjunto de prueba y los restantes como conjunto de entrenamiento.

Luego, para cada combinación de hiperparámetros, en cada iteración del algoritmo XGB, se calculará el promedio de la función de pérdida en todos los conjuntos de validación. Posteriormente, se seleccionará la *mínima pérdida promedio* y la iteración en la que se alcanzó dicho mínimo, la cual sera la métrica de rendimiento que se utilizará para determinar cuándo una combinación de hiperparámetros es mejor a otra.

Este proceso generó un total de 21 combinaciones posibles de hiperparámetros, cada una con su respectiva menor pérdida promedio y la iteración en la que se alcanzó. Para seleccionar el modelo final, se optó por analizar los conjuntos de parámetros que se encontraban alrededor

de la mediana de la *mínima pérdida promedio*, con el objetivo de evitar los sobreajustes que se obtenían al analizar los modelos con las mejores métricas de rendimiento, estás combinaciones se encuentran en la sección A.4.2. Una vez completado este análisis, se eligió el conjunto de hiperparámetros que redujo el IPS del modelo, el cual fue:

$$(eta, \#min\_child\_weight, colsample\_bytree, \#nrounds) = \left(0.2, 346, \frac{3}{11}, 72\right)$$

Respecto a los parámetros de regularización, se procedió a trabajar con los valores por defecto: $\gamma = 0$, lo que significa que se realizarán divisiones en los nodos sin que estos tengan un umbral mínimo de pérdida, y $\lambda = 1$.

Implementación del modelo

La ejecución del modelo de Credit Scoring basado en el algoritmo XGBoost, se llevó a cabo utilizando el lenguaje estadístico R y su paquete `xgboost`. El entrenamiento del modelo se realizó mediante el siguiente código:

```
set.seed(12345)
modelo <- xgboost(
  data = RMOD,
  label = label,
  eta = 0.2,
  nthread = 4,
  nrounds = 72,
  objective = "binary:logistic",
  eval_metric = "logloss", maximize = FALSE,
  min_child_weight = 346,
  early_stopping_rounds = 50,
  colsample_bytree = 3/11,
  tree_method="exact", lambda=1, gamma=0)
```

En el código anterior:

- `set.seed()`: Fija la semilla de aleatorización.
- `data`: Indica la base de datos utilizada para entrenar el modelo; en

nuestro caso, se refiere a la muestra rebalanceada detallada en la tabla 3.8, la cual puede estar en el formato `matrix` de R o en el formato `DMatrix` propio de la librería `xgboost`.

- `label`: Indica la variable dependiente, VarDep, como vector numérico de R.
- `eta`: Indica la tasa de aprendizaje del modelo, que controla la contribución de cada árbol.
- `nthread`: Indica el número de hilos (threads) de la computadora a utilizar para el entrenamiento.
- `nrounds`: Indica el número total de iteraciones (árboles) a realizar.
- `objective`: Indica qué tipo de problema se realiza, en este caso, clasificación binaria.
- `eval_metric`: Especifica la métrica de evaluación que se utilizará durante el entrenamiento. En este caso, se utiliza la pérdida logarítmica.
- `maximize`: Indica si se debe maximizar (`TRUE`) o minimizar (`FALSE`) la métrica de evaluación.
- `min_child_weight`: Especifica el número mínimo de registros que debe tener cada nodo hoja en cada árbol generado.
- `early_stopping_rounds`: Detiene el entrenamiento si no hay mejora después de cierto número de iteraciones.
- `colsample_bytree`: Controla la proporción de predictores seleccionados aleatoriamente para construir cada árbol.
- `tree_method`: Indica el método utilizado para construir árboles. En este caso, exact hace referencia al método greedy visto en la sección 2.5.2.
- `lambda`: Parámetro de regularización L2 aplicado a los pesos del modelo. Ayuda a prevenir el sobreajuste.
- `gamma`: Parámetro que controla la reducción mínima necesaria para realizar una nueva partición en un nodo hoja.

Importancia de las Variables

Al examinar la importancia de las variables en el modelo, que se fundamenta en cómo cada una contribuye a la reducción total de la función de pérdida, se observa el siguiente comportamiento:

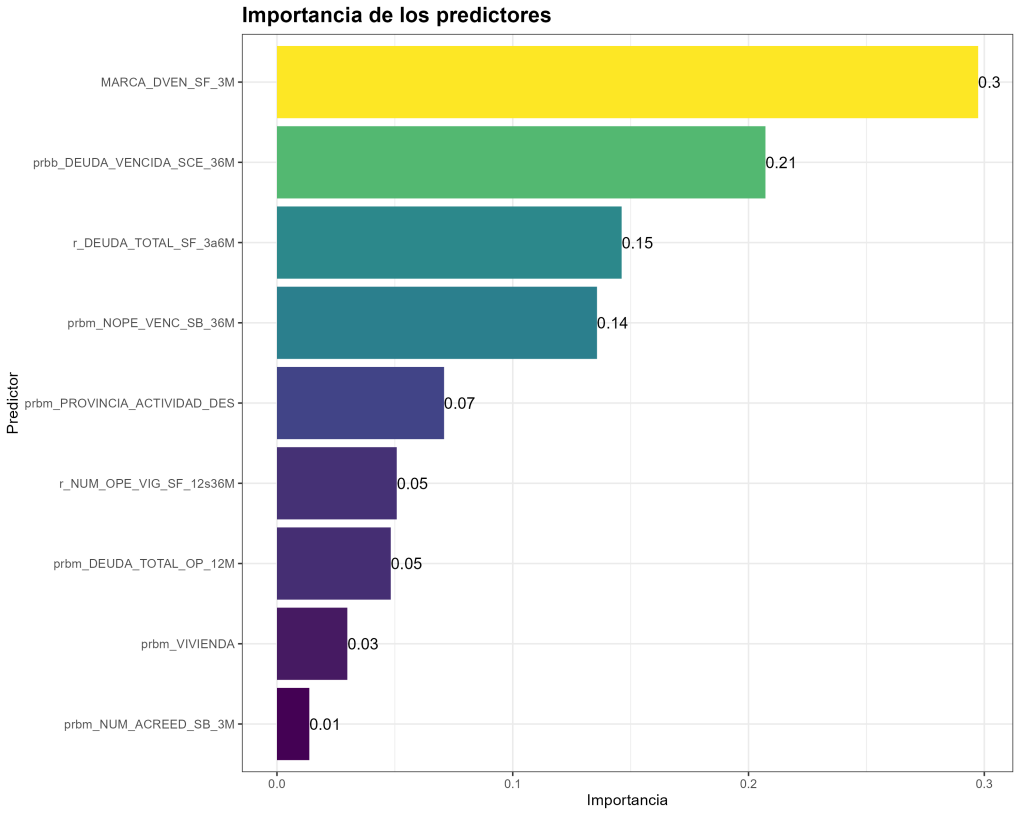

Figura 3.12: Importancia Variables - Modelo: XGB
Elaboración: El autor

En el gráfico 3.12, se observa que la variable más influyente en el modelo es MARCA_DVEN_SF_3M, seguida por la variable prbb_DEUDA_VENCIDA_SCE_36M, ambas contribuyendo con un 51 % a la reducción de la función de pérdida. Así, se evidencia que el comportamiento crediticio relacionado con la información vencida es determinante para clasificar si un individuo es bueno o malo bajo este modelo. Es importante señalar que las variables prbm_TOTAL_INGRESOS y MARCA_NENT_VENC_SB_24M no aparecen en el gráfico debido a su baja contribución en el modelo, esto se debe a los parámetros de regularización del algoritmo. Sin embargo, se mantienen en el modelo para realizar el muestreo de variables, asegurando así una mayor diversidad de árboles.

Análisis de Correlación

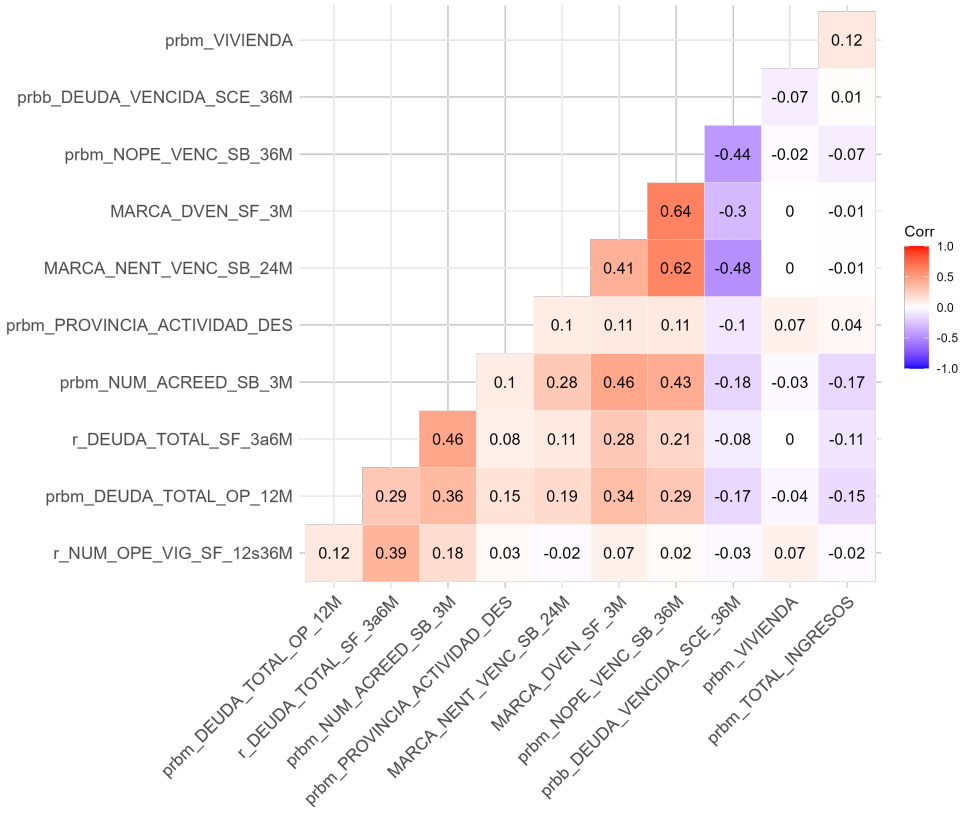

Figura 3.13: Matriz de Correlaciones - Modelo: XGB
Elaboración: El autor

Se puede observar que en la matriz del modelo ajustado, la máxima correlación en valor absoluto es de 0.64, y esta es compartida por las variables prbm_NOPE_VENC_SB_36M y MARCA_DVEN_SF_3M. La similitud podría deberse a que ambas variables consideran comportamientos crediticios vencidos y trabajan, como mínimo, con información del sistema bancario. Sin embargo, dado que el índice de condicionamiento para la matriz de correlación es de 3.33, se puede descartar problemas de multicolinealidad.

Capítulo 4

Resultados

En esta sección, se presentan las tablas performance para cada uno de los modelos entrenados, tanto en la base de modelización como en la de validación. Además, se incluyen las métricas de desempeño correspondientes que permitirán evaluar el poder discriminativo de cada modelo, las cuales permiten determinar si los modelos son adecuados para el cálculo de pérdidas esperadas. Además, a través de ellas se realiza la comparación entre los modelos de machine learning y la metodología tradicional, la regresión logística.

Una vez validada la eficacia de los modelos, se analiza el impacto de cada uno de ellos en el cálculo de las pérdidas esperadas para la cartera de créditos de la institución. Este análisis busca determinar si los modelos de machine learning ofrecen mejoras significativas con respecto a la metodología tradicional.

4.1. Tablas Performance

Para cada modelo entrenado, se tienen las siguientes tablas performance que permiten evaluar su capacidad discriminativa, y cuyo análisis se realizará en base a lo explicado en la sección 2.8.1.

4.1.1. Modelo Regresión Logística

Base de Modelamiento

KS	ROC	GINI
55.1	84.5	68.9

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int%	Cum%	Int#	Int%	Cum%	Int%	Cum%
970	999	1,215	10%	10%	10	1%	1%	0.82%	0.82%
958	970	1,214	10%	20%	12	1%	2%	0.99%	0.91%
941	958	1,215	10%	30%	20	2%	5%	1.65%	1.15%
922	941	1,214	10%	40%	29	3%	8%	2.39%	1.46%
888	922	1,215	10%	50%	45	5%	13%	3.70%	1.91%
851	888	1,215	10%	60%	57	6%	19%	4.69%	2.37%
780	851	1,214	10%	70%	65	7%	26%	5.35%	2.80%
660	780	1,215	10%	80%	120	13%	39%	9.88%	3.68%
457	660	1,214	10%	90%	177	19%	59%	14.58%	4.89%
1	457	1,215	10%	100%	376	41%	100%	30.95%	7.50%
Total		12,146			911				

Cuadro 4.1: Tabla Performance - Modelo: RGL - Base: Modelización
Elaboración: El autor

De esta manera, la información proporcionada por la tabla 4.1, a través de la columna Razón de Malo Int%, revela que en el 10% de la población con el Score más elevado, se podría esperar un 0.82% de malos pagadores. En contraste, en el 10% de la población con el Score más bajo se alcanzaría el 30.95%, representando la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona ubicada en el séptimo decil de Score tendría una PD del 5.35%, información a usar para el calculo de las pérdidas esperadas.

Adicionalmente, mediante la columna Razón de Malo Cum%, la institución podría establecer su política de riesgo en la aprobación de créditos. Por ejemplo, si la institución busca obtener un riesgo menor o igual al 3%, debería aprobar al 70% de la población con el Score más alto.

Base de Validación

KS	ROC	GINI
52.3	83.6	67.2

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
970	999	1,215	10 %	10 %	10	1 %	1 %	0.82 %	0.82 %
958	970	1,214	10 %	20 %	16	2 %	3 %	1.32 %	1.07 %
940	958	1,215	10 %	30 %	26	3 %	5 %	2.14 %	1.43 %
920	940	1,214	10 %	40 %	34	4 %	9 %	2.80 %	1.77 %
887	920	1,215	10 %	50 %	52	5 %	14 %	4.28 %	2.27 %
849	887	1,215	10 %	60 %	64	7 %	21 %	5.27 %	2.77 %
775	849	1,214	10 %	70 %	77	8 %	29 %	6.34 %	3.28 %
654	774	1,215	10 %	80 %	105	11 %	40 %	8.64 %	3.95 %
457	654	1,214	10 %	90 %	188	20 %	60 %	15.49 %	5.23 %
1	457	1,215	10 %	100 %	386	40 %	100 %	31.77 %	7.89 %
Total		12,146			958				

Cuadro 4.2: Tabla Performance - Modelo: RGL - Base: Validación
Elaboración: El autor

De esta manera, lo que nos comunica la tabla 4.2 mediante la columna Razón de Malo Int % es que dentro del 10 % de la población con el Score más elevado, se podría esperar un 0.82 % de malos pagadores, mientras que en el 10 % de la población con el Score más bajo se alcanzaría el 31.77 %, representando la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona con un Score de al menos 849 pero menor a 887 puntos tiene una probabilidad de default del 5.27 %.

Al analizar la columna Razón de Malo Cum %, se puede concluir que si la institución desea obtener un riesgo menor o igual al 3 %, debería aprobar al 60 % de la población con el Score más alto.

Métricas de Desempeño

	Métricas		
Base	KS	ROC	GINI
Modelización	55.1	84.5	68.9
Validación	52.3	83.6	67.2

Cuadro 4.3: Métricas de Discriminación - Modelo: RGL
Elaboración: El autor

La tabla 4.3 se construyó utilizando la información proporcionada por las tablas 4.1 y 4.2. Para ambas bases, el test KS es mayor a 50 y el GINI

es mayor a 65, lo que indica un buen rendimiento del modelo tanto en la modelización como en la validación. Además, al analizar el Índice de Estabilidad Poblacional (IPS) descrito en la tabla 4.4, se puede concluir que el modelo se mantiene estable y que el poder discriminante se conserva al pasar de la base de modelización a la de validación, ya que el IPS tiene un valor menor al 10 %. En resumen, según las métricas descritas, el modelo es sólido y adecuado para el cálculo de las pérdidas esperadas.

Decil	**Modelización Razón Malo Int % (w)**	**Validación Razón Malo Int % (x)**	**IPS [w-x]*ln[w/x]**
1	0.82	0.82	0.00
2	0.99	1.32	0.09
3	1.65	2.14	0.13
4	2.39	2.80	0.07
5	3.70	4.28	0.08
6	4.69	5.27	0.07
7	5.35	6.34	0.17
8	9.88	8.64	0.16
9	14.58	15.49	0.05
10	30.95	31.77	0.02
		IPS	0.85

Cuadro 4.4: IPS - Modelo: RGL
Elaboración: El autor

4.1.2. Modelo Random Forest

Base de Modelamiento

KS	ROC	GINI
56.2	86.1	72.3

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
968	999	1,215	10 %	10 %	3	0 %	0 %	0.25 %	0.25 %
937	968	1,214	10 %	20 %	6	1 %	1 %	0.49 %	0.37 %
910	937	1,215	10 %	30 %	22	2 %	3 %	1.81 %	0.85 %
881	910	1,214	10 %	40 %	30	3 %	7 %	2.47 %	1.26 %
854	881	1,215	10 %	50 %	39	4 %	11 %	3.21 %	1.65 %
810	854	1,215	10 %	60 %	60	7 %	18 %	4.94 %	2.20 %
759	810	1,214	10 %	70 %	72	8 %	25 %	5.93 %	2.73 %
689	759	1,215	10 %	80 %	102	11 %	37 %	8.40 %	3.44 %
559	689	1,214	10 %	90 %	175	19 %	56 %	14.42 %	4.66 %
1	559	1,215	10 %	100 %	403	44 %	100 %	33.17 %	7.51 %
Total		12,146			912				

Cuadro 4.5: Tabla Performance - Modelo: RF - Base: Modelización
Elaboración: El autor

De esta manera, la información proporcionada por la tabla 4.5, a través de la columna Razón de Malo Int %, revela que dentro del 10 % de la población con el Score más elevado, se podría esperar un 0.25 % de malos pagadores. En contraste, en el 10 % de la población con el Score más bajo se alcanzaría el 33.17 %, la cual representa la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona ubicada en el séptimo decil de Score tendría una PD del 5.93 %, información a usar para el calculo de las pérdidas esperadas.

Adicionalmente, mediante la columna Razón de Malo Cum %, la institución podría establecer su política de riesgo en la aprobación de créditos. Por ejemplo, si la institución busca obtener un riesgo menor o igual al 3 %, debería aprobar al 70 % de la población con el Score más alto.

Base de Validación

KS	ROC	GINI
51.2	83.6	67.3

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
968	999	1,215	10 %	10 %	6	1 %	1 %	0.49 %	0.49 %
935	968	1,214	10 %	20 %	13	1 %	2 %	1.07 %	0.78 %
910	935	1,215	10 %	30 %	26	3 %	5 %	2.14 %	1.23 %
881	910	1,214	10 %	40 %	42	4 %	9 %	3.46 %	1.79 %
852	881	1,215	10 %	50 %	47	5 %	14 %	3.87 %	2.21 %
809	852	1,215	10 %	60 %	71	7 %	21 %	5.84 %	2.81 %
755	809	1,214	10 %	70 %	90	9 %	31 %	7.41 %	3.47 %
687	755	1,215	10 %	80 %	118	12 %	43 %	9.71 %	4.25 %
563	687	1,214	10 %	90 %	168	17 %	60 %	13.84 %	5.32 %
1	562	1,215	10 %	100 %	384	40 %	100 %	31.60 %	7.95 %
Total		12,146			965				

Cuadro 4.6: Tabla Performance - Modelo: RF - Base: Validación
Elaboración: El autor

De esta manera, lo que nos comunica la tabla 4.6 mediante la columna Razón de Malo Int % es que dentro del 10 % de la población con el Score más elevado, se podría esperar un 0.49 % de malos pagadores, mientras que en el 10 % de la población con el Score más bajo se alcanzaría el 31.60 %, representando la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona con un Score de al menos 809 pero menor a 852 puntos tiene una probabilidad de default del 5.84 %.

Al analizar la columna Razón de Malo Cum %, se puede concluir que si la institución desea obtener un riesgo menor o igual al 3 %, debería aprobar al 60 % de la población con el Score más alto.

Métricas de Desempeño

	Métricas		
Base	KS	ROC	GINI
Modelización	56.2	86.1	72.3
Validación	51.2	83.6	67.3

Cuadro 4.7: Métricas de Discriminación - Modelo: RF
Elaboración: El autor

La tabla 4.7 se construyó utilizando la información proporcionada por las tablas 4.5 y 4.6. Para ambas bases, el test KS es mayor a 50 y el GINI es mayor a 65, lo que indica un buen rendimiento del modelo tanto en la modelización como en la validación. Además, al analizar el Índice de Estabilidad Poblacional (IPS) descrito en la tabla 4.8, se puede concluir que el modelo se mantiene estable y que el poder discriminante se conserva al pasar de la base de modelización a la de validación, ya que el IPS tiene un valor menor al 10 %. En resumen, según las métricas descritas, el modelo es sólido y adecuado para el cálculo de las pérdidas esperadas.

Decil	**Modelización Razón Malo Int % (w)**	**Validación Razón Malo Int % (x)**	**IPS [w-x]*ln[w/x]**
1	0.25	0.49	0.17
2	0.49	1.07	0.45
3	1.81	2.14	0.05
4	2.47	3.46	0.33
5	3.21	3.87	0.12
6	4.94	5.84	0.15
7	5.93	7.41	0.33
8	8.40	9.71	0.19
9	14.42	13.84	0.02
10	33.17	31.60	0.08
		IPS	1.90

Cuadro 4.8: IPS - Modelo: RF
Elaboración: El autor

4.1.3. Modelo XGBoost

Base de Modelamiento

KS	ROC	GINI
53.3	83.9	67.9

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
965	999	1,215	10 %	10 %	4	0 %	0 %	0.33 %	0.33 %
958	965	1,214	10 %	20 %	22	2 %	3 %	1.81 %	1.07 %
936	958	1,215	10 %	30 %	27	3 %	6 %	2.22 %	1.45 %
912	936	1,214	10 %	40 %	30	3 %	9 %	2.47 %	1.71 %
884	912	1,215	10 %	50 %	43	5 %	14 %	3.54 %	2.07 %
836	884	1,215	10 %	60 %	51	6 %	20 %	4.20 %	2.43 %
772	836	1,214	10 %	70 %	76	8 %	28 %	6.26 %	2.98 %
654	772	1,215	10 %	80 %	96	11 %	39 %	7.90 %	3.59 %
473	654	1,214	10 %	90 %	181	20 %	59 %	14.91 %	4.85 %
1	473	1,215	10 %	100 %	373	41 %	100 %	30.70 %	7.43 %
Total		12,146			903				

Cuadro 4.9: Tabla Performance - Modelo: XGB - Base: Modelización
Elaboración: El autor

De esta manera, la información proporcionada por la tabla 4.9, a través de la columna Razón de Malo Int %, revela que en el 10 % de la población con el Score más elevado, se podría esperar un 0.33 % de malos pagadores. En contraste, en el 10 % de la población con el Score más bajo se alcanzaría el 30.70 %, representando la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona ubicada en el séptimo decil de Score tendría una PD del 6.26 %, información a usar para el calculo de las pérdidas esperadas.

Adicionalmente, mediante la columna Razón de Malo Cum %, la institución podría establecer su política de riesgo en la aprobación de créditos. Por ejemplo, si la institución busca obtener un riesgo menor o igual al 3 %, debería aprobar al 70 % de la población con el Score más alto.

Base de Validación

KS	ROC	GINI
52.6	83.5	66.9

Score		Total			Malo			Razón de Malo	
Min	Max	Int#	Int %	Cum %	Int#	Int %	Cum %	Int %	Cum %
965	999	1,215	10 %	10 %	7	1 %	1 %	0.58 %	0.58 %
958	965	1,214	10 %	20 %	19	2 %	3 %	1.57 %	1.07 %
936	958	1,215	10 %	30 %	31	3 %	6 %	2.55 %	1.56 %
908	936	1,214	10 %	40 %	35	4 %	10 %	2.88 %	1.89 %
883	908	1,215	10 %	50 %	45	5 %	14 %	3.70 %	2.26 %
831	883	1,215	10 %	60 %	47	5 %	19 %	3.87 %	2.52 %
761	831	1,214	10 %	70 %	96	10 %	29 %	7.91 %	3.29 %
651	761	1,215	10 %	80 %	109	11 %	41 %	8.97 %	4.00 %
473	651	1,214	10 %	90 %	202	21 %	62 %	16.64 %	5.41 %
1	473	1,215	10 %	100 %	361	38 %	100 %	29.71 %	7.84 %
Total		12,146			952				

Cuadro 4.10: Tabla Performance - Modelo: XGB - Base: Validación
Elaboración: El autor

De esta manera, lo que nos comunica la tabla 4.10 mediante la columna Razón de Malo Int % es que dentro del 10 % de la población con el Score más elevado, se podría esperar un 0.58 % de malos pagadores, mientras que en el 10 % de la población con el Score más bajo se alcanzaría el 29.71 %, representando la probabilidad de incumplimiento (PD) del crédito. Así, por ejemplo, una persona con un Score de al menos 831 pero menor a 883 puntos tiene una probabilidad de default del 3.87 %.

Al analizar la columna Razón de Malo Cum %, se puede concluir que si la institución desea obtener un riesgo menor o igual al 3 %, debería aprobar al 60 % de la población con el Score más alto.

Métricas de Desempeño

	Métricas		
Base	KS	ROC	GINI
Modelización	53.3	83.9	67.9
Validación	52.6	83.5	66.9

Cuadro 4.11: Métricas de Discriminación - Modelo: XGB
Elaboración: El autor

La tabla 4.11 se construyó utilizando la información proporcionada por las tablas 4.9 y 4.10. Para ambas bases, el test KS es mayor a 50 y el GINI es mayor a 65, lo que indica un buen rendimiento del modelo tanto en la modelización como en la validación. Además, al analizar el Índice de Estabilidad Poblacional (IPS) descrito en la tabla 4.12, se puede concluir que el modelo se mantiene estable y que el poder discriminante se conserva al pasar de la base de modelización a la de validación, ya que el IPS tiene un valor menor al 10 %. En resumen, según las métricas descritas, el modelo es sólido y adecuado para el cálculo de las pérdidas esperadas.

Decil	Modelización Razón Malo Int % (w)	Validación Razón Malo Int % (x)	IPS [w-x]*ln[w/x]
1	0.33	0.58	0.14
2	1.81	1.57	0.04
3	2.22	2.55	0.05
4	2.47	2.88	0.06
5	3.54	3.70	0.01
6	4.20	3.87	0.03
7	6.26	7.91	0.38
8	7.90	8.97	0.14
9	14.91	16.64	0.19
10	30.70	29.71	0.03
		IPS	1.06

Cuadro 4.12: IPS - Modelo: XGB
Elaboración: El autor

4.1.4. Comparativa Métricas de Rendimiento

Métrica	Modelos					
	RGL		RF		XGB	
	Mod	Val	Mod	Val	Mod	Val
KS	55.1	52.3	56.2	51.2	53.3	52.6
ROC	84.5	83.6	86.1	83.6	83.9	83.5
GINI	68.9	67.2	72.3	67.3	67.9	66.9
IPS	0.85 %		1.90 %		1.06 %	

Cuadro 4.13: Tabla Comparativa de las Métricas de Rendimiento
Elaboración: El autor

Ahora bien, al comparar las métricas de rendimiento entre los modelos, se observa que el modelo Logístico tiene mejores métricas de discriminación (KS, ROC, GINI) que el modelo XGBoost, pero no mejores que el modelo de Random Forest en la modelización. Sin embargo, al pasar a la base de validación, el modelo de XGBoost tiene un mejor KS, el ROC es casi igual entre los tres, y el GINI, basado en la métrica anterior, es similar, teniendo una diferencia de 0.4 puntos entre el mejor y el peor.

Al analizar el Índice de Estabilidad Poblacional (IPS) de los modelos, se observa que el que presenta la menor variación es el modelo RGL, seguido muy de cerca por el modelo XGBoost. Así, según lo anterior, se podría concluir que a nivel global y en base a los resultados de validación, los modelos de Regresión Logística y XGBoost presentan el mejor poder discriminativo y se mantienen al pasar de una base a otra. Más, cabe recalcar que los tres modelos son buenos y aplicables para el cálculo de pérdidas esperadas según los análisis anteriores. En la siguiente sección, se observará cual de estos 3 modelos reduce el aprovisionamiento de las pérdidas esperadas.

4.2. Pérdidas Esperadas

Una vez validados los modelos para el cálculo de las pérdidas esperadas, se procedió a estimarlas en función de lo explicado en la sección 2.9, con el fin de determinar el modelo que las reduce. Así, en la base de validación, se obtuvieron los siguientes resultados para la cartera en todo el horizonte de estudio:

Resultados	**Modelos**		
	RGL	**RF**	**XGB**
PE ($)	2,825,789	2,865,695	2,804,761
EAD ($)	79,615,050		
$r_{\textbf{Pérdida}}$ (%)	3.549315	3.599439	3.522903

Cuadro 4.14: Comparativas de resultados
Elaboración: El autor

La tabla 4.14 proporciona información sobre la pérdida esperada total en todo el horizonte de estudio, en la base de validación, en la fila PE. Mientras tanto, la fila de EAD indica el total de crédito otorgado en la

ventana de desempeño para la misma base. Por otro lado, la fila de $r_{\text{Pérdida}}$ refleja la razón de pérdida de la cartera de validación en toda la ventana de desempeño. Así, el modelo XGBoost presenta el menor ratio de pérdida total. Donde, la institución debería aprovisionar 3.52 dólares por cada 100 dólares otorgados en crédito durante los 12 meses analizados y es lo que se esperaría que suceda en el siguiente año si las características del mercado no cambian.

Ahora bien, al analizar la razón de pérdida, mes a mes, en la ventana de desempeño, se obtuvo el comportamiento descrito en la figura 4.1:

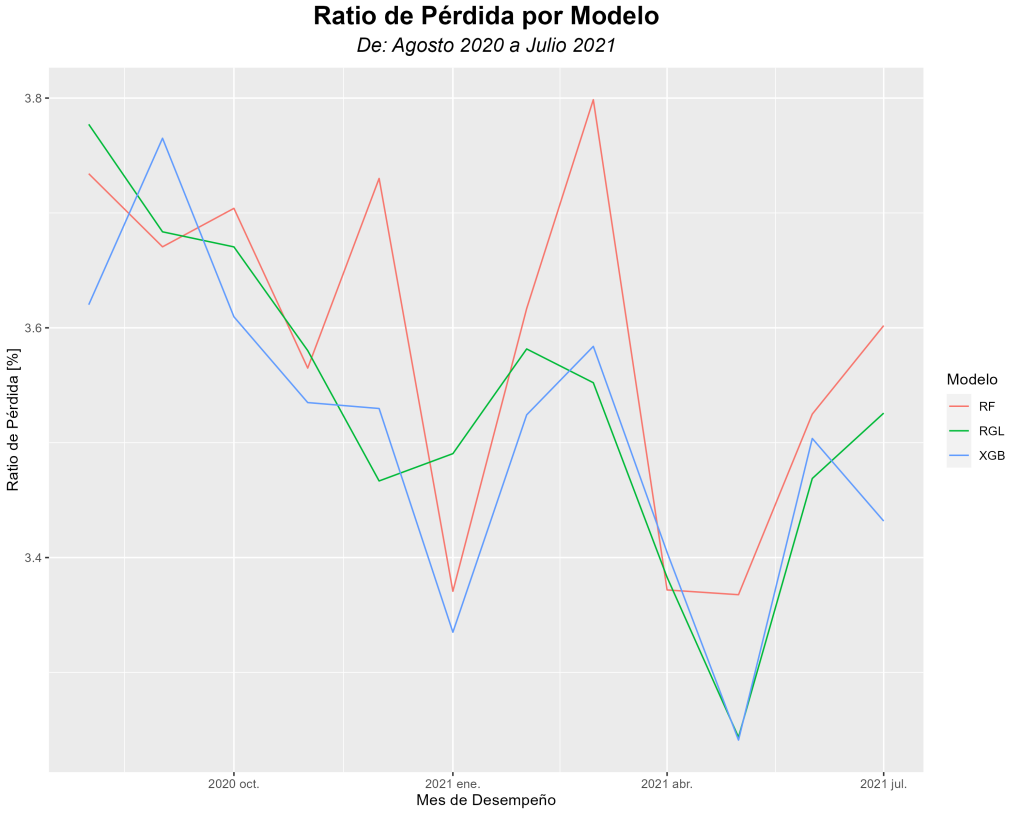

Figura 4.1: Razón de pérdida en la Ventana de Desempeño
Elaboración: El autor

La gráfica anterior muestra la evolución de la razón de pérdida en la ventana de desempeño para cada modelo, en la base de validación. Se puede observar que el modelo XGB tiene el menor ratio de aprovisionamiento en 7 de los 12 meses analizados, mientras que la Regresión Logística y el Random Forest tienen el menor ratio en 3 y 2 meses respectivamente. Al analizar, en la base de validación, la razón de pérdida promedio en las ventanas de desempeño, se obtuvieron los siguientes

resultados:

Modelo	razón de pérdida promedio (%)
RGL	3.535277
RF	3.587995
XGB	3.506968

Cuadro 4.15: Razón de pérdida promedio
Elaboración: El autor

Por lo tanto, el modelo que presenta el menor ratio de pérdida promedio es el XGBoost, según los resultados de validación, con una reducción del 0.80 % y del 2.25 % en el aprovisionamiento en comparación con los modelos de Regresión Logística y Random Forest, respectivamente. Así, con el modelo XGBoost, en promedio, se debería aprovisionar 3.51 dólares por cada 100 dólares otorgados en crédtio cada mes.

Capítulo 5

Conclusiones y recomendaciones

5.1. Conclusiones

1. Los tres modelos estimados para el cálculo de las pérdidas esperadas tienen buenas métricas de rendimiento, tanto en términos de poder discriminativo como de estabilidad, como se detalla en la Tabla 4.13. Por lo tanto, los tres son adecuados para calcular las pérdidas esperadas.

2. Basándonos en los resultados de validación descritos en la tabla 4.13, la metodología tradicional, regresión logística, supera en capacidad predictiva y estabilidad a los modelos de aprendizaje automático, con el modelo XGBoost estando muy cerca de su rendimiento.

3. El algoritmo XGBoost es el modelo que tiene el mayor impacto en el cálculo de las pérdidas esperadas de los tres modelos implementados, según los resultados de validación, con un 3.51 % de aprovisionamiento mensual de la cartera, en promedio, superando a la metodología tradicional.

4. La importancia de las variables permite identificar aquellas más significativas para la predicción de buenos y malos pagadores, información valiosa para la institución a la hora de establecer políticas crediticias.

5. El modelo de Random Forest presentó el mayor IPS, lo que indica un cambio en el poder discriminativo al pasar de la base de modelamiento a la de validación, lo cual podría deberse a la falta de variables más importantes en el modelo o a la necesidad de explorar otros hiperparámetros.

6. Las tablas performance son herramientas fundamentales para determinar el poder discriminativo de los modelos, a través de las cuales se puede calcular la probabilidad de incumplimiento de un individuo, así como la tasa de aprobación asociada al riesgo que la institución está dispuesta a asumir.

7. Si la institución desea mantener un riesgo igual o inferior al 3%, debería aprobar al 70% de las personas con el Score más alto en la base de modelización según todos los modelos.

8. Aprobando el 60% de los créditos con el Score más alto, se tiene un riesgo menor o igual al 3% en la base de validación para todos los modelos.

5.2. Recomendaciones

1. Se recomienda revisar el componente matemático de los modelos desarrollados para tener claro el proceso de modelado, así como los resultados entregados por cada uno, en especial del modelo XGBoost.

2. En el proyecto, se estableció que los individuos con 61 o más días de vencimiento en la ventana de desempeño serían clasificados como malos pagadores, según el análisis de Roll-Rate. Sin embargo, se observó que a partir de los 31 días de vencimiento, la probabilidad de deterioro rondaba el 45%. Por lo tanto, se sugiere revisar cómo el umbral de días de vencimiento para clasificar a un mal pagador puede influir en el cálculo de las pérdidas esperadas, siempre que esta revisión esté respaldada técnicamente.

3. En el proceso de construcción del modelo, se excluyeron a los individuos No Bancarizados debido a que no poseían historial crediticio

en los últimos 3 años anteriores al proceso de concesión, lo que imposibilitaba su segmentación a través de los modelos propuestos. Sin embargo, para no perder oportunidades de mercado, se recomienda construir modelos de Score Sociodemográfico para este tipo de individuos.

4. Se recomienda rebalancear las muestras para poder identificar patrones de comportamiento de malos pagadores y tener modelos que puedan identificarlos de mejor manera.

5. Se recomienda realizar ingeniería de variables para poder capturar y dar más valor a la información disponible, como se realizó en este trabajo.

6. En este proyecto, la selección de variables se realizó en base a los test KS y VI, buscando tener información de diferentes aspectos del individuo así como representar de la mejor manera su realidad financiera. Sin embargo, se recomienda revisar métodos adicionales para la selección de variables, como métodos Stepwise, Lasso o algoritmos genéticos, y observar su impacto en el cálculo de las pérdidas esperadas.

7. Se recomienda estudiar otros hiperparámetros a los ya expuestos para los modelos, como se mencionan en [15] y [7] para los modelos de Random Forest y XGBoost, respectivamente.

8. Si el ajuste de hiperparámetros se realiza a través de un proceso de grid-search, se recomienda realizar una segunda grilla de hiperparámetros donde estos se encuentren en un rango más cercano.

9. El ajuste de hiperparámetros se llevó a cabo mediante una grilla de hiperparámetros y validación cruzada. Sin embargo, en el proceso de revisión de literatura se encontró que una forma alternativa de ajustarlos es mediante optimización bayesiana, la cual trabaja de manera más eficiente que un proceso de grid-search. Por lo tanto, se recomienda revisar cómo este enfoque puede ayudar en el cálculo de las pérdidas esperadas en comparación con la metodología tradicional.

10. El proyecto siguió un enfoque estándar según Basilea II en términos de LGD. Sin embargo, se recomienda modelar la LGD según lo propuesto en [14] para mejorar las estimaciones de las pérdidas esperadas y que se ajusten al contexto de la institución.

11. Se recomienda realizar pruebas de backtesting, sensibilidad y de estrés a los modelos entrenados para la PD, para poder observar el poder de predicción de los modelos en el tiempo y al variar el comportamiento de las variables incluidas en los modelos, así como observar el rendimiento de estos en situaciones de crisis.

Capítulo A

Anexos

A.1. Código, modelos y bases de datos

Los códigos en R para reproducir los modelos, así como las bases de entrenamiento y validación, se pueden obtener en el siguiente enlace: https://github.com/RENE-2312/TIC-RENE-2023B

A.2. Tablas Roll Rate

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	95.01	4.99
De 1 a 30 días	93.62	6.38

Cuadro A.1: Tabla Roll-Rate m1 vs m2
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	94.33	5.67
De 1 a 30 días	89.73	10.27
De 31 a 60 días	37.74	62.26

Cuadro A.2: Tabla Roll-Rate m2 vs m3
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	94.70	5.30
De 1 a 30 días	87.80	12.20
De 31 a 60 días	44.44	55.56
De 61 a 90 días	17.65	82.35

Cuadro A.3: Tabla Roll-Rate m3 vs m4
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	93.98	6.02
De 1 a 30 días	86.57	13.43
De 31 a 60 días	41.18	58.82
De 61 a 90 días	24.69	75.31
De 91 a 120 días	6.67	93.33

Cuadro A.4: Tabla Roll-Rate m4 vs m5
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	92.60	7.40
De 1 a 30 días	86.25	13.75
De 31 a 60 días	53.82	46.18
De 61 a 90 días	20.53	79.47
De 91 a 120 días	6.56	93.44
De 121 a 150 días	6.90	93.10

Cuadro A.5: Tabla Roll-Rate m5 vs m6
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	93.73	6.27
De 1 a 30 días	83.13	16.87
De 31 a 60 días	54.91	45.09
De 61 a 90 días	33.56	66.44
De 91 a 120 días	13.01	86.99
De 121 a 150 días	13.56	86.44
De 151 a 180 días	7.69	92.31
Más de 180 días	0	100

Cuadro A.6: Tabla Roll-Rate m6 vs m7
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	94.57	5.43
De 1 a 30 días	87.44	12.56
De 31 a 60 días	51.97	48.03
De 61 a 90 días	33.51	66.49
De 91 a 120 días	18.52	81.48
De 121 a 150 días	16.82	83.18
De 151 a 180 días	3.85	96.15
Más de 180 días	7.69	92.31

Cuadro A.7: Tabla Roll-Rate m7 vs m8
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	93.97	6.03
De 1 a 30 días	87.40	12.60
De 31 a 60 días	64.09	35.91
De 61 a 90 días	29.74	70.26
De 91 a 120 días	26.81	73.19
De 121 a 150 días	8.60	91.40
De 151 a 180 días	6.59	93.41
Más de 180 días	2.63	97.37

Cuadro A.8: Tabla Roll-Rate m8 vs m9
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	95.02	4.98
De 1 a 30 días	88.55	11.45
De 31 a 60 días	66.58	33.42
De 61 a 90 días	48.52	51.48
De 91 a 120 días	13.21	86.79
De 121 a 150 días	19.09	80.91
De 151 a 180 días	10.84	89.16
Más de 180 días	4.88	95.12

Cuadro A.9: Tabla Roll-Rate m9 vs m10
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	94.29	5.71
De 1 a 30 días	87.84	12.16
De 31 a 60 días	65.22	34.78
De 61 a 90 días	46.20	53.80
De 91 a 120 días	35.19	64.81
De 121 a 150 días	17.99	82.01
De 151 a 180 días	10.20	89.80
Más de 180 días	2.13	97.87

Cuadro A.10: Tabla Roll-Rate m10 vs m11
Elaboración: El autor

Rango Vencido	Estado	
	No Avanza (%)	Avanza (%)
Sin Vencido	94.20	5.80
De 1 a 30 días	85.21	14.79
De 31 a 60 días	67.20	32.80
De 61 a 90 días	49.44	50.56
De 91 a 120 días	28.23	71.77
De 121 a 150 días	16.46	83.54
De 151 a 180 días	9.43	90.57
Más de 180 días	3.74	96.26

Cuadro A.11: Tabla Roll-Rate m11 vs m12
Elaboración: El autor

A.3. Selección de Variables para los modelos

Variables Escogidas - Modelo: Logit			
Cuantitativas		Categóricas	
Variable	KS	Variable	VI
r_NUM_OPE_VIG_SF_12s36M	0.2465	MARCA_DVEN_SF_3M	1.0408
		prbb_DEUDA_VENCIDA_SCE_36M	0.6840
		prbm_DEUDA_TOTAL_OP_12M	0.5094
		prbm_NUM_ENT_VIG_SB_12M	0.3018
		prbm_NOPE_APERT_OP_24M	0.2955
		prbm_PROVINCIA_ACTIVIDAD_DES	0.2918
		prbm_TOTAL_INGRESOS	0.0618

Cuadro A.12: Variables Escogidas para el Modelo: RGL
Elaboración: El autor

Variables Escogidas - Modelo: Random Forest			
Cuantitativas		Categóricas	
Variable	KS	Variable	VI
r_MAX_DVEN_SB_6a12M	0.4507	prbm_NOPE_VENC_SB_36M	0.9299
r_DEUDA_TOTAL_SF_3a6M	0.3567	prbm_DEUDA_VENCIDA_SCE_12M	0.6789
r_NUM_OPE_VIG_SB_12s24M	0.2389	prbm_DEUDA_TOTAL_OP_12M	0.5094
		MARCA_NENT_VEN_SB_36M	0.4799
		prbm_NUM_ACREED_SB_3M	0.3945
		prbb_NOPE_APERT_OP_24M	0.2955
		prbm_PROVINCIA_ACTIVIDAD_DES	0.2918
		prbb_TOTAL_INGRESOS	0.0617

Cuadro A.13: Variables Escogidas para el Modelo: RF
Elaboración: El autor

Variables Escogidas - Modelo: XGBoost			
Cuantitativas		Categóricas	
Variable	KS	Variable	VI
r_DEUDA_TOTAL_SF_3a6M	0.3567	MARCA_DVEN_SF_3M	1.0408
r_NUM_OPE_VIG_SF_12s36M	0.2465	prbm_NOPE_VENC_SB_36M	0.9299
		prbb_DEUDA_VENCIDA_SCE_36M	0.6840
		prbm_DEUDA_TOTAL_OP_12M	0.5094
		MARCA_NENT_VENC_SB_24M	0.4852
		prbm_NUM_ACREED_SB_3M	0.3945
		prbm_PROVINCIA_ACTIVIDAD_DES	0.2918
		prbb_TOTAL_INGRESOS	0.0617
		prbm_VIVIENDA	0.0462

Cuadro A.14: Variables Escogidas para el Modelo: XGB
Elaboración: El autor

A.4. Grilla de Hiperparámetros

A.4.1. Modelo Random Forest

En la siguiente tabla, se muestran las tripletas de hiperparámetros analizadas para el modelo Random Forest. La combinación de hiperparámetros que tiene la mediana del *acurracy promedio* está resaltada en amarillo, mientras que la seleccionada en el modelo está resaltada en rosa.

#mtry	*#min.node.size*	*#ntrees*	***acurracy promedio***
3	116	416	0.853285199
3	66	416	0.85299639
4	83	485	0.852851986
4	132	416	0.852851986
4	66	346	0.852707581
4	116	416	0.852707581
4	66	416	0.852563177
3	66	485	0.852418773
3	83	485	0.852274368
4	116	554	0.852274368
3	132	485	0.852274368
4	132	554	0.852129964
4	83	554	0.85198556
4	99	554	0.851841155
4	132	485	0.851841155
4	66	554	0.851552347
4	83	623	0.851552347
4	99	485	0.851552347
4	116	623	0.851407942
4	83	692	0.851407942

Cuadro A.15: Grilla de Hiperparámetros - Modelo: RF
Elaboración: El autor

A.4.2. Modelo XGBoost

En la siguiente tabla, se muestra la combinación de hiperparámetros analizados para el modelo XGBoost. La combinación de hiperparámetros que tiene la mediana de la *mínima pérdida promedio* está resaltada en amarillo, mientras que la seleccionada en el modelo está resaltada en rosa.

eta	*#min_child_weight*	*colsample_bytree*	*#nrounds*	***mínima pérdida promedio***
0.2	346	0.272727273	72	0.381613186
0.2	416	0.272727273	27	0.421744823
0.1	416	0.272727273	53	0.422311274
0.3	416	0.272727273	18	0.423446779
0.1	485	0.272727273	57	0.450535913
0.2	485	0.272727273	28	0.451981824
0.3	485	0.272727273	16	0.454044815
0.3	554	0.272727273	17	0.467340974
0.1	554	0.272727273	59	0.468448725
0.2	554	0.272727273	27	0.468794221
0.2	623	0.272727273	28	0.480935498

Cuadro A.16: Grilla de Hiperparámetros - Modelo: XGB
Elaboración: El autor

Referencias bibliográficas

[1] Lucia Acurio. Modelo de gestión para la implementación de los procesos de administración de riesgo de crédito de consumo por parte de las entidades del sistema bancario ecuatoriano. Master's thesis, Universidad Andina Simón Bolívar, 2015. Tomado de: https://repositorio.uasb.edu.ec/bitstream/10644/4848/1/T1859-MFGR-Acurio-Modelo.pdf.

[2] ALTAIR. Credit scoring series part five: Credit scorecard development, 06 2022. https://n9.cl/794fw.

[3] Claudia Bertoli, José Braccini Neto, and V. Roso. Comparing methodologies to estimate fixed genetic effects and to predict genetic values for an angus × nellore cattle population. *Journal of Animal Science*, 94:503–504, 02 2016. Tomado de: https://doi.org/10.2527/jas.2015-9344.

[4] Yurdakul Bilal. *Statistical Properties of Population Stability Index*. PhD thesis, Western Michigan University, 2018. Tomado de: https://scholarworks.wmich.edu/dissertations/3208.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. Tomado de: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785.

[6] Scorto Corporation. Scorecard validation. Tomado de: https://plug-n-score.com/learning/scorecard-validation.htm.

[7] Xgboost developers. Xgboost parameters, 2022. Tomado de: https://xgboost.readthedocs.io/en/latest/parameter.html.

[8] Harris Elshaddai. Understanding weight of evidence and information value, 2022. Tomado de: https://n9.cl/ocv8l.

[9] Farhad Khossro. Modelos estadísticos para la estimación de la capacidad de pago de personas naturales tarjetahabientes con información en el sistema de registro de datos crediticios, 2023. Tomado de: https://bibdigital.epn.edu.ec/handle/15000/25047.

[10] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. Tomado de: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf.

[11] Jayawant N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010. Tomado de: https://doi.org/10.1097/JTO.0b013e3181ec173d.

[12] Elvis. Mantilla. Aplicación de algoritmos genéticos en la selección de variables para la construcción de modelos de credit scoring, 2023. Tomado de: https://bibdigital.epn.edu.ec/handle/15000/24509.

[13] Cristhian Montalván. Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito. Master's thesis, Universidad Andina Simón Bolívar, 2019. Tomado de: https://repositorio.uasb.edu.ec/bitstream/10644/6872/1/T2962-MGFARF-Montalvan-Credit.pdf.

[14] María Rosero. Desafíos de la industria bancaria en la estimación de la severidad de pérdida (lgd - loss given default), 2023. Tomado de: https://bibdigital.epn.edu.ec/bitstream/15000/24843/1/CD%2013517.pdf.

[15] Sharoon Saxena. A beginner's guide to random forest hyperparameter tuning, 2023. Tomado de: https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/.

[16] Alejandro Sotomayor. Estimación de la pérdida esperada para una cartera de microcrédito basada en calificaciones internas, 2012. Tomado de: https://bibdigital.epn.edu.ec/bitstream/15000/4668/1/CD-4301.pdf.

[17] TransUnion. Preguntas frecuentes sobre los modelos de score de transunion, 2012. Tomado de: https://n9.cl/sxtgc.

[18] Wikipedia. Kolmogorov–smirnov test, 2024. Tomado de: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test.

[19] Jeffrey M. Wooldridge. *Introducción a la econometría, Un enfoque moderno*. CENGAGE Learning, 2009. Tomado de: https://n9.cl/nocp2.