Clasificación de créditos de libranza negociados en el mercado secundario colombiano, aplicando técnicas de aprendizaje supervisado

Juan Camilo Gomez Betancur

jcgomezb@eafit.edu.co

Asesor
Nicolas Alberto Moreno Reyes

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2024

Tabla de Contenido

Lista de figuras

Lista de tablas

Resumen

El riesgo crediticio, exacerbado por eventos como la crisis financiera de 2008, sigue siendo una preocupación para entidades bancarias y no bancarias. Este estudio aborda la necesidad de mejorar la clasificación de créditos de libranza en Colombia mediante técnicas tradicionales y de aprendizaje automático. El objetivo de este estudio es desarrollar un modelo capaz de identificar los créditos con mayor probabilidad de incumplimiento mediante el uso de algoritmos de aprendizaje supervisado. Este enfoque ha demostrado una eficacia superior en la clasificación del riesgo crediticio. Como resultado, se espera optimizar la adquisición de créditos de libranza y fortalecer la cartera de los portafolios de inversión.

Palabras clave: Clasificación Binaria, Puntuación crediticia, Crédito, Riesgo Crédito, Clasificador de créditos, Crédito de Libranza, XGBoost.

Abstract

Credit risk, exacerbated by events such as the 2008 financial crisis, remains a concern for both banking and non-banking entities. This study addresses the need to improve the classification of payroll loans in Colombia using both traditional and machine learning techniques. It highlights the superior effectiveness of supervised learning algorithms in credit risk classification, with the ultimate goal of developing a model capable of identifying loans with a higher probability of default. This would optimize the acquisition of payroll loans and strengthen investment portfolios.

Keywords: Binary Classification, Credit Scoring, Credit, Credit Risk, Credit Classifiers, Libranza, XGBoost.

1. Introducción

El riesgo crédito existe hace décadas, las instituciones bancarias y no bancarias han hecho múltiples esfuerzos por brindar un análisis más detallado y perfeccionar las técnicas alrededor de la administración del riesgo crédito y el impacto en sus resultados financieros; esto se ha convertido en una prioridad, precedentes como la crisis financiera en 2008 han llevado a desarrollar modelos que permitan realizar la administración de riesgo crediticio (Louzada, Ara, y Fernandes, 2016). En este documento aplicaremos diferentes técnicas: tanto tradicionales como de aprendizaje automático para la clasificación de créditos de libranza, los cuales son negociados por medio del mercado secundario no estandarizado en Colombia, donde el objetivo es identificar los créditos de libranza que pueden presentar un evento de incumplimiento.

En los mercados secundarios se lleva a cabo la compra y venta de activos que fueron emitidos antes de la fecha de negociación. En este mercado, se negocian los créditos y los flujos futuros esperados asociados a ellos. En Colombia, una de las modalidades de crédito que se negocia en este mercado es el crédito de libranza, dada sus características particulares. Tradicionalmente, para clasificar los créditos se han utilizado técnicas como el Análisis Discriminante Lineal (LDA, por sus siglas en inglés) y la Regresión Logística. Sin embargo, este documento busca contrastar estas técnicas con métodos de aprendizaje automático, específicamente algoritmos de aprendizaje supervisado. Estudios recientes sobre el análisis del riesgo crediticio han demostrado que estos algoritmos ofrecen mejores resultados que las técnicas tradicionales de clasificación de créditos (Barboza, Kimura, y Altman, 2017).

A lo largo del documento se examina cómo los algoritmos de aprendizaje supervisado mejoran la clasificación en comparación con las técnicas tradicionales para evaluar las libranzas, especialmente aquellas que pueden conllevar eventos de incumplimiento. Esto tiene como objetivo optimizar la adquisición de créditos de libranza, lo que a su vez permite construir una cartera más sólida. Este enfoque robusto y estandarizado en la inversión busca obtener una mayor rentabilidad asumiendo un menor riesgo.

1.1. Planteamiento del problema

El mercado secundario de créditos de libranza en Colombia es amplio, pero poco estandarizado. Esto significa que los créditos de libranza se compran y venden de manera personalizada, adaptándose a

las necesidades de las partes involucradas. En un intento reciente por estandarizar el mercado, se ha promovido la negociación de pagarés desmaterializados a través de Deceval. Sin embargo, las estructuras de estas negociaciones suelen ser complejas y con requisitos adicionales al pagaré, lo que dificulta la homogeneización del mercado. Cada contraparte presenta condiciones particulares que definen las características de las negociaciones.

Por esta razón, surge la necesidad de seleccionar cuidadosamente los créditos de libranza que serán adquiridos e integrados en un portafolio. El objetivo es mantener una cartera en la que los deudores realicen sus pagos puntualmente. En el caso de los créditos de libranza, esto implica adquirir aquellos créditos en los que las cuotas son descontadas mensualmente del salario o pensión del deudor sin que se produzcan retrasos. De lo contrario, podríamos terminar con una cartera que presenta altos índices de morosidad, lo que se traduciría en pérdidas para la inversión.

Este documento analiza la cartera de un Fondo de Inversión Colectiva que invierte en operaciones de crédito en Colombia. El objetivo es clasificar los créditos de libranza según la probabilidad de que se produzca un incumplimiento por parte del deudor o la Pagaduría. Según las condiciones de la operación entre el Fondo de Inversión Colectiva y el Originador de las libranzas (quien actúa como vendedor en la negociación), se considerará un evento de incumplimiento cuando el crédito supere un periodo de mora de 90 días. Esto significa que el crédito de libranza ha superado los 90 días sin recibir pagos por parte de la Pagaduría, que puede ser el empleador o el administrador de la pensión del deudor.

Con el objetivo de mejorar la calidad de la cartera adquirida por el Fondo de Inversión Colectiva, se llevará a cabo la evaluación de diferentes modelos de clasificación de créditos de libranza. Se compararán tanto técnicas tradicionales, como LDA y la regresión logística, como también técnicas de aprendizaje automático, entre las que se incluyen Support Vector Machines, Boosting, Bagging, Random Forest y XGBoost. La evaluación de los modelos se llevará a cabo con el propósito de identificar el clasificador óptimo de créditos de libranza. Este clasificador deberá ser capaz de predecir de manera eficiente los eventos de incumplimiento, utilizando como medida de desempeño el F1-Score.

1.2. Justificación

En Colombia, existen diversos tipos de originadores de créditos de libranza, algunos de los cuales son entidades no bancarias. En ocasiones, estas entidades carecen de los recursos financieros necesarios para expandir la cartera administradas. Muchas de estas compañías suelen surgir con respaldo de fondeadores privados, que generalmente son recursos propios o de familiares y amigos. Sin embargo, estas entidades pueden llegar a un punto en el que la liquidez necesaria para mantener el crecimiento del negocio y continuar con la operación, que consiste principalmente en la colocación de créditos, no puede ser cubierta con los recursos de los fondeadores.

Cuando los fondeadores iniciales no pueden cubrir las necesidades de liquidez para la colocación de créditos, las compañías buscan nuevas formas de financiamiento. Una de las modalidades más comunes es la venta de cartera. La liquidez proporcionada por los fondeadores iniciales se destina a la colocación de créditos, es decir, el dinero inicialmente aportado se presta a diversos pagadores/deudores. Posteriormente, las entidades originadoras venden los créditos, los cuales incluyen los flujos futuros derivados de los pagos recibidos, con el fin de obtener más liquidez para financiar nuevos créditos y mantener la continuidad del negocio.

Existen Fondos de Inversión Colectiva, como el que se evaluará en este documento, que se especializan en realizar inversiones en activos de naturaleza crediticia, tales como bonos, CDTs, descuento de facturas, compra de créditos y otros títulos relacionados con el crédito que puedan ser objeto de negociación. La adquisición de cartera compuesta por créditos de libranza es un activo muy atractivo para estos fondos y para todo tipo de fondeadores en general. Esto se debe a que los créditos de libranza tienen la particularidad de que los pagos no son realizados directamente por el deudor, sino que las cuotas son descontadas automáticamente del salario o la pensión por parte del empleador o el administrador del fondo de pensiones (Pagadurías). Este mecanismo de descuento automático reduce el riesgo crediticio en cierta medida, debido a la prioridad que se otorga al pago de estos créditos.

La necesidad de analizar el riesgo crediticio surge desde los primeros días del comercio, cuando comenzaron a surgir transacciones que involucraban el endeudamiento y los préstamos de dinero (Louzada y cols., 2016). Para los Fondos de Inversión Colectiva, es crucial evaluar el nivel de riesgo al que está expuesta la cartera de su portafolio. Este riesgo proviene de las inversiones realizadas por el Fondo en activos crediticios y se traduce principalmente en una exposición al riesgo de crédito. Al optar por adquirir este

riesgo, se espera obtener una rentabilidad esperada a cambio. Sin embargo, es fundamental controlar y mitigar esta exposición, ya que cualquier evento de incumplimiento podría resultar en pérdidas potenciales tanto para el Fondo como para sus inversionistas.

Después de los acontecimientos de la crisis financiera de 2008, la gestión del riesgo crediticio se elevó a la categoría de prioridad (Barboza y cols., 2017). Para un Fondo que se especializa en inversiones en activos de naturaleza crediticia, esta gestión se convierte en el eje central de sus operaciones, dado que el riesgo predominante en este tipo de inversiones es el riesgo crediticio. Evaluar el grado de exposición al riesgo que representa cada crédito en el portafolio se vuelve crucial para obtener una cartera óptima, mediante la selección de créditos con menor probabilidad de incumplimiento.

La selección de créditos es crucial para optimizar la estructura de costos (Maldonado, Peters, y Weber, 2018). En estas transacciones crediticias, la estructura juega un papel fundamental, ya que define los términos entre el Originador y el Fondeador. El Fondeador busca una estructura robusta que mitigue los posibles incumplimientos y riesgos en la cartera del portafolio, mientras que el Originador busca una estructura ágil que permita un ciclo rápido desde la originación del crédito hasta su venta al Fondeador. Por ello, la clasificación de créditos de libranza puede actuar como un factor de mitigación del riesgo crediticio, permitiendo la creación de estructuras más flexibles que mejoren las condiciones de negociación para ambas partes.

Ohlson fue pionero en el análisis del riesgo crediticio mediante la regresión logística (Altman, 1968). Este proyecto tiene como objetivo identificar el mejor modelo de clasificación para los créditos adquiridos por un fondo de inversión, con el fin de mejorar la calidad de su cartera, reducir el índice de cartera vencida y generar mayor confianza entre los inversionistas. Además, busca negociar nuevas estructuras que sean más flexibles y proporcionen mejores condiciones para el Originador. Estas mejoras pueden traducirse en un mayor retorno para el Fondeador y sus acreedores, al tiempo que se mitiga el riesgo crediticio mediante el uso de técnicas tradicionales y de aprendizaje automático.

2. Objetivos

2.1. Objetivo General

Desarrollar un modelo para la clasificación de créditos de libranza negociados en el mercado secundario colombiano, el cual permita predecir los créditos de libranza que van a presentar eventos de incumplimiento.

2.2. Objetivos específicos

- Evaluar el desempeño de diferentes modelos de clasificación para créditos de libranza negociados en el mercado secundario colombiano, para identificar el modelo más efectivo en la predicción de incumplimientos, utilizando el F1-Score como métrica principal de evaluación.

- Seleccionar el modelo con el mejor rendimiento para el conjunto de datos, para optimizar la precisión en la predicción de incumplimientos de créditos de libranza, Comparando y analizando los resultados de los modelos evaluados en la etapa de diagnóstico.

- Mejorar la calidad y el comportamiento del portafolio de créditos de libranza adquiridos en el mercado secundario colombiano, para reducir el riesgo de incumplimiento y aumentar la rentabilidad del portafolio, aplicando el modelo seleccionado y validado a los datos del portafolio, monitoreando y ajustando según los resultados obtenidos.

3. Marco Teórico

El crédito es el mecanismo por el cual personas o entidades con exceso de liquidez realizan préstamos a otras personas o entidades con necesidades de liquidez, y a cambio del préstamo realizado se espera un porcentaje de rentabilidad sobre el monto entregado. Sin embargo, existe la posibilidad de que este dinero no sea retornado de nuevo al prestatario por parte del deudor en los tiempos o condiciones pactadas desde el inicio, y a esto se le conoce como riesgo de crédito.

El riesgo de crédito es la piedra angular de la evaluación del riesgo (Mushava y Murray, 2022), lo que ha llevado a la necesidad de desarrollar técnicas y modelos para el análisis y administración del riesgo crediticio.

3.1. Crédito de Libranza

El crédito de libranza se destina a empleados y pensionados, a quienes descuentan mensualmente del monto de su nómina o mesada pensional para pagar las cuotas crédito (Bogotá, 2024). Este tipo de crédito permite a los deudores utilizar los fondos como deseen, con la facilidad de pagar las cuotas a través de descuentos automáticos en su salario o pensión (BBVA, 2024). En esta dinámica, el empleador o fondo de pensión del deudor es el encargado de realizar los pagos a las entidades prestamistas o bancos. Se utiliza para financiar diversos proyectos e incluso consolidar deudas mediante la compra de cartera (Bogotá, 2024). Para asegurar el correcto funcionamiento de estos créditos, se establecen convenios entre la entidad pagadora (empleador o administrador de fondos pensionales) y la entidad prestamista, permitiendo el descuento directo de las cuotas del crédito de la nómina o mesada pensional.

Con el fin de mantener esta característica fundamental, los originadores de créditos registran los recaudos en las pagadurías a través de un Patrimonio Autónomo, segregando así el patrimonio y permitiendo que los fondeadores reciban los pagos de las cuotas directamente desde dicho Patrimonio Autónomo. Además del riesgo crediticio inherente a este tipo de activos, las libranzas enfrentan constantemente el riesgo de incorporación, que se materializa cuando una libranza no completa adecuadamente el proceso de registro en la pagaduría para los pagos de las cuotas.

3.2. Riesgo Crédito

El riesgo de crédito se refiere a la posibilidad de pérdida que ocurre cuando un prestatario no cumple con sus obligaciones crediticias. Este riesgo es un aspecto crucial para la gestión financiera, dado que puede afectar significativamente la estabilidad económica de una institución. Para mitigar este riesgo, se emplean diversas garantías y seguros en las operaciones de crédito, diseñados para reducir la exposición al incumplimiento. Estas herramientas son fundamentales en la evaluación y gestión del riesgo crediticio, permitiendo a las instituciones anticipar y manejar posibles pérdidas derivadas de la falta de pago por parte de los prestatarios. La comprensión y aplicación efectiva de estas medidas son esenciales para proteger el capital invertido y asegurar la estabilidad financiera a largo plazo.

El riesgo asociado a un prestatario varía considerablemente según el tamaño de la empresa. Generalmente, las grandes empresas presentan un riesgo mayor comparado con las medianas o pequeñas (Bohn y Stein, 2009). Esto se debe a que las grandes corporaciones suelen enfrentar más variables económicas y operativas que pueden afectar su capacidad de cumplir con las obligaciones crediticias. Por otro lado, concentrarse en un menor número de prestamistas puede incrementar la exposición al riesgo crediticio. En contraste, diversificar el portafolio de préstamos a través de una mayor cantidad de prestatarios con créditos menores puede disminuir la exposición al riesgo al repartir la carga crediticia entre múltiples entidades. Esta estrategia ayuda a equilibrar el riesgo global y a minimizar el impacto negativo de un incumplimiento individual.

La evaluación del riesgo crediticio es una práctica esencial para prevenir crisis financieras que pueden surgir a partir de incumplimientos. No obstante, ser excesivamente estricto al conceder créditos puede limitar la competitividad y la oportunidad de obtener rendimientos atractivos (Maldonado y cols., 2018). Encontrar el equilibrio adecuado entre la prudencia en la evaluación de riesgos y la necesidad de mantener una cartera de créditos rentable es crucial. Una estrategia demasiado conservadora puede llevar a una reducción en la capacidad de crecimiento y expansión del negocio, mientras que una política de crédito demasiado laxa puede incrementar las pérdidas por impagos. Por lo tanto, es vital ajustar las políticas de crédito de manera que se proteja el capital mientras se aprovechan las oportunidades de negocio.

Las herramientas de medición del riesgo de crédito juegan un papel fundamental en la toma de decisiones financieras al permitir a las instituciones rechazar créditos con alto riesgo de incumplimiento (Verbraken, Bravo, Weber, y Baesens, 2014). Estas herramientas incluyen modelos analíticos y métodos

estadísticos que proporcionan una evaluación precisa del riesgo asociado con cada solicitud de crédito. Implementar estas herramientas puede resultar en una gestión más eficiente del riesgo y, a largo plazo, en mayores ganancias, al evitar la aceptación indiscriminada de créditos que podrían resultar en pérdidas. Al utilizar técnicas avanzadas de medición, las instituciones financieras pueden optimizar su portafolio de préstamos y mejorar su rendimiento general, garantizando una mayor estabilidad y éxito en sus operaciones.

3.3. Mercado Secundario

El mercado secundario es un componente esencial dentro de los mercados financieros, donde se lleva a cabo la negociación de activos que han sido previamente emitidos en el mercado primario. En contraste con el mercado primario, en el que se crean y emiten nuevos valores, el mercado secundario proporciona una plataforma para la compra y venta de estos valores ya existentes, como acciones, bonos y otros instrumentos financieros. Este proceso permite a los inversores intercambiar activos y ajustar sus carteras conforme a las condiciones cambiantes del mercado. La capacidad de negociar activos en el mercado secundario contribuye significativamente a la liquidez general del sistema financiero, facilitando así la transferencia de propiedad y ayudando a mantener un flujo constante de inversiones.

En el mercado secundario, los inversores tienen la posibilidad de comprar y vender valores entre ellos, lo que proporciona una valiosa liquidez y facilita la transferencia de la propiedad de los activos financieros. Esta capacidad de realizar transacciones de manera rápida y eficiente es crucial para que los inversores puedan ajustar sus estrategias de inversión y diversificar sus riesgos. Además, la posibilidad de negociar activos de forma ágil permite a los inversores responder a las fluctuaciones del mercado y a los cambios en las condiciones económicas, manteniendo así una flexibilidad necesaria para la gestión efectiva de sus carteras y para la optimización de sus rendimientos financieros.

Existen diversos tipos de mercados secundarios que incluyen tanto bolsas de valores organizadas como sistemas de negociación electrónicos. Las bolsas de valores organizadas proporcionan un entorno regulado y transparente para la ejecución de transacciones, asegurando la integridad y la equidad en la negociación de valores. Por otro lado, los mercados over-the-counter (OTC) permiten la negociación directa entre compradores y vendedores sin la intermediación de una bolsa centralizada. Ambos tipos de mercados juegan roles importantes en el funcionamiento del mercado secundario, ofreciendo diferentes opciones pa-

ra la compra y venta de activos y contribuyendo a la eficiencia y la accesibilidad en el comercio de valores.

El mercado secundario cumple un rol crucial en la eficiencia y el funcionamiento saludable de los mercados financieros en su conjunto. Actúa como una plataforma para la fijación de precios de activos financieros, reflejando de manera precisa la oferta y la demanda en tiempo real. Esta función es fundamental para mantener la transparencia en el proceso de formación de precios y fomentar la competencia entre los participantes del mercado. Sin un mercado secundario eficiente, la valoración de activos sería menos precisa y la liquidez del sistema financiero se vería comprometida, afectando la capacidad de los inversores para realizar ajustes necesarios en sus carteras y para gestionar sus riesgos financieros de manera efectiva.

En resumen, el mercado secundario es un componente esencial de los mercados financieros que permite la compra y venta de activos financieros ya existentes. Su existencia es crucial para proporcionar liquidez, eficiencia y transparencia al sistema financiero global. Al facilitar la transferencia de capital y la asignación eficiente de recursos, el mercado secundario contribuye significativamente al funcionamiento estable y dinámico del sistema financiero, permitiendo a los inversores operar en un entorno donde los precios se ajustan continuamente en respuesta a las condiciones del mercado.

Los mercados en los que se realizan operaciones de compra y venta de créditos se dividen en dos categorías principales: Mercado Primario y Mercado Secundario. Esta clasificación se basa en si se trata de la emisión inicial de valores por el emisor o de la negociación de valores ya emitidos y en circulación entre inversionistas (Chen, Härdle, y Moro, 2011). En el caso de los créditos de libranza, el mercado primario es aquel en el que el crédito es emitido por primera vez; aquí, el originador realiza el desembolso y el cliente firma la libranza. Por otro lado, el mercado secundario se refiere a la negociación de estos créditos entre inversionistas después de su emisión inicial. Este mercado facilita la transferencia de estos activos financieros entre diferentes participantes, permitiendo una mayor flexibilidad y dinámica en la gestión de estos créditos.

3.4. Regresión Logística

La Regresión Logística es un modelo de aprendizaje supervisado utilizado para problemas de clasificación binaria, donde la probabilidad de ocurrencia de un evento (éxito o fracaso) se da en función de las variables independientes. La distribución condicional de la variable dependiente sigue una distribución

Bernoulli, donde la función de probabilidad de esta variable es desconocida. La regresión logística modela la probabilidad transformada por logit como una relación lineal con las variables predictoras.

Sea $X = (x_1, x_2, \ldots, x_n)$ una muestra aleatoria en un espacio de 12 dimensiones, donde cada x_i representa un crédito de libranza negociado en el mercado secundario.

Además, sea Y una variable aleatoria que indica si un crédito de libranza ha presentado default. En este contexto, Y es la respuesta para un clasificador binario y toma valores en el conjunto $\{0, 1\}$, donde 1 indica que el crédito presentó default y 0 indica que no lo presentó. Así, Y es utilizado para etiquetar los créditos en la muestra X según su estatus de default.

$$\text{logit}(P(Y=1)) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{1}$$

donde:

- $P(Y=1)$ es la probabilidad del evento que ocurre.
- $\beta_0, \beta_1, \ldots, \beta_k$ son los coeficientes de regresión.
- $X_1, X_2, \ldots, X_k$ son las variables independientes.

La ecuación para obtener la probabilidad del riesgo de fracaso es:

$$P(Y=1) = \frac{1}{1+\exp -(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)} \tag{2}$$

Una vez que se obtienen las predicciones de probabilidad del modelo, se utiliza un umbral de decisión para clasificar las instancias en las clases correspondientes. Por ejemplo, si el valor predicho es mayor que el umbral, se clasifica como la clase positiva (éxito); de lo contrario, se clasifica como la clase negativa (fracaso).

Umbral de Decisión:

$$\text{Predicción} = \begin{cases} 1 & \text{si } h_\theta(x) \geq U \\ 0 & \text{si } h_\theta(x) < U \end{cases} \tag{3}$$

donde $U \in (0,1)$

Durante el entrenamiento del modelo, se optimiza la función de costo para ajustar los coeficientes de regresión y encontrar los parámetros óptimos que minimizan el error de predicción. La función de costo comúnmente utilizada en la regresión logística es la función de pérdida logarítmica (entropía cruzada binaria), que mide la discrepancia entre las predicciones del modelo y los valores reales. El objetivo del entrenamiento es encontrar los coeficientes de regresión que minimicen esta función de costo.

Función de Costo (Entropía Cruzada):

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \tag{4}$$

Donde:

- $J(\theta)$ es la función de costo.
- m es el número de ejemplos de entrenamiento.
- $y^{(i)}$ es la etiqueta real del i-ésimo ejemplo.
- $h_\theta(x^{(i)})$ es la predicción del modelo para el i-ésimo ejemplo.

La regresión logística se basa en una serie de investigaciones previas. Altman, en su trabajo de 1981 (Hua, Yumeng, Siwen, Jianbin, y Yutong, 2020), proporciona una revisión exhaustiva de la regresión logística y su aplicación en problemas financieros. También se ha explorado la aplicación de la regresión logística en la predicción de riesgos crediticios, como se discute en el estudio de Doumpos et al. (M., K., G., y C., 2002) sobre la evaluación del riesgo crediticio utilizando un enfoque de discriminación jerárquica multicriterio.

3.5. Support Vector Machines

El algoritmo de Support Vector Machines (SVM) ha ganado popularidad en los últimos años en el ámbito del Machine Learning, especialmente en el reconocimiento de patrones. En el contexto del riesgo

crediticio, se utiliza para identificar la distancia entre los deudores y clasificarlos en diferentes clases (default o no default). SVM no solo minimiza el riesgo empírico, sino que también maximiza la separación marginal entre las clases, logrando un equilibrio entre ambos objetivos (Chen y cols., 2011). La optimización del modelo SVM se basa en la transformación de una función matemática llamada 'kernel', que ayuda a identificar las distancias óptimas entre observaciones con diferentes clasificaciones (Barboza y cols., 2017). Esta transformación permite encontrar un hiperplano en un espacio de alta dimensión que maximiza el margen entre las clases, lo que lo convierte en un algoritmo eficaz para la clasificación binaria.

La función de decisión de SVM está dada por:

$$f(x) = \text{sign}(\mathbf{w}^T\mathbf{x} + b) \tag{5}$$

donde:

- $f(x)$ es la función de decisión.
- $\mathbf{w}$ es el vector de pesos.
- $\mathbf{x}$ es el vector de características de la instancia.
- b es el término de sesgo (bias).

La función de decisión clasifica una instancia x como perteneciente a una clase si el resultado de $\mathbf{w}^T\mathbf{x} + b$ es mayor que cero, y a la otra clase si es menor que cero.

Durante el entrenamiento del modelo, SVM busca el hiperplano óptimo que maximiza el margen entre las clases, lo cual se formula como un problema de optimización convexa. La función de costo asociada con SVM es la función de pérdida hinge, que penaliza las predicciones incorrectas y ayuda a ajustar los parámetros del modelo para mejorar la precisión de la clasificación. La función de costo asociada con SVM es la función de pérdida hinge:

$$J(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b)) \tag{6}$$

donde:

- $J(\mathbf{w}, b)$ es la función de costo.
- n es el número de ejemplos de entrenamiento.
- $y^{(i)}$ es la etiqueta real del i-ésimo ejemplo.
- $\mathbf{x}^{(i)}$ es el vector de características del i-ésimo ejemplo.

En resumen, SVM es un algoritmo eficaz para la clasificación de datos, especialmente en problemas de clasificación binaria donde se busca maximizar la separación entre las clases. Su capacidad para encontrar un hiperplano óptimo en un espacio de alta dimensión lo hace especialmente útil en aplicaciones como el riesgo crediticio, donde se necesita una precisión alta y una buena generalización del modelo.

3.6. Bagging

El Bagging, o Bootstrap Aggregating, es una técnica ampliamente utilizada en aprendizaje automático, para mejorar la precisión y la estabilidad de los modelos predictivos mediante el uso de ensamblajes. Consiste en la creación de múltiples conjuntos de datos de entrenamiento mediante muestreo con reemplazo del conjunto de datos original. Cada conjunto de datos se utiliza para entrenar un modelo base independiente, y luego se combinan las predicciones de estos modelos para producir una predicción final más robusta y precisa (Barboza y cols., 2017). La idea central detrás del Bagging es introducir variaciones en los modelos al construirlos sobre diferentes subconjuntos de datos de entrenamiento. Esto ayuda a reducir el sesgo y la varianza de las predicciones finales, mejorando así la capacidad de generalización del modelo y reduciendo el riesgo de sobreajuste.

Yu, Lean, Wang, Shouyang y Lai, Kin Keung propusieron un enfoque de aprendizaje de ensamble basado en redes neuronales, utilizando un método de muestreo de bagging para generar conjuntos de entrenamiento diversos y robustos (Lean, Shouyang, y Keung, 2008). El Bagging es particularmente eficaz cuando se aplican modelos base con alta varianza. Al promediar las predicciones de múltiples modelos base, el Bagging puede reducir la varianza total del modelo final y mejorar su rendimiento predictivo (Breiman, 1996). Este enfoque es especialmente beneficioso en conjuntos de datos ruidosos o con alta dimensionalidad, donde los modelos individuales pueden tener dificultades para generalizar adecuadamente.

Algorithm 1 Algoritmo de Bagging con Ponderación de Modelos

procedure BAGGING
    Entrada: Conjunto de entrenamiento D, número de modelos base B
    Salida: Modelo Bagging M
    Inicializar una lista vacía para almacenar los modelos base: models $\leftarrow []$
    for $b = 1$ **hasta** B **do**
        Seleccionar aleatoriamente n muestras con reemplazo del conjunto de entrenamiento D: $D_b \leftarrow$ muestrear_con_reemplazo(D, n)
        Entrenar un modelo base en el conjunto D_b: $M_b \leftarrow$ entrenar_modelo(D_b)
        Agregar el modelo base entrenado a la lista de modelos: models.append(M_b)
    end for
    Calcular la predicción final ponderando los modelos base entrenados M_b, con pesos iguales: $\hat{y} = \frac{1}{B}\sum_{b=1}^{B} M_b(x)$
    Retornar $\hat{y}$
end procedure

3.7. Random Forest

La técnica Random Forest se basa en árboles de decisión, los cuales se construyen utilizando subconjuntos aleatorios de datos y características. En este proceso, se crean múltiples árboles de decisión, cada uno operando sobre un subconjunto distinto de datos, y luego se combina la predicción de cada árbol para determinar la etiqueta final del registro (Barboza y cols., 2017; Lean y cols., 2008; Kruppa, Schwarz, Arminger, y Ziegler, 2013). Random Forest es esencialmente un conjunto de árboles de decisión que utiliza el método de bagging para mejorar la precisión y evitar el sobreajuste. Este método generaliza el algoritmo de clasificación y regresión de árboles (CART), que divide los datos utilizando diferentes porciones para identificar grupos de clientes con características crediticias similares. Además, se pueden utilizar árboles de decisión para estimar probabilidades individuales, lo que se conoce como árboles de estimación de probabilidad (PETs) (Kruppa y cols., 2013).

Es importante destacar que el tamaño de los árboles en Random Forest juega un papel crucial en su rendimiento. Los árboles más pequeños son más comprensibles, pero pueden no capturar las complejidades de los datos, mientras que los árboles más grandes pueden sobreajustarse y perder capacidad de generalización. Por lo tanto, el proceso de construcción de árboles se puede detener cuando el tamaño del nodo alcanza un umbral predefinido, como el 5 % o el 10 % de todas las muestras disponibles (Kruppa y cols., 2013; Malekipirbazari y Aksakalli, 2015). En resumen, Random Forest es una técnica poderosa que utiliza múltiples árboles de decisión construidos sobre subconjuntos aleatorios de datos y características

para realizar predicciones. Esta técnica es especialmente útil para evitar el sobreajuste y mejorar la precisión en problemas de clasificación y regresión. Además, la capacidad de estimar probabilidades individuales mediante árboles de estimación de probabilidad permite una mayor flexibilidad en la modelización de ciertos problemas.

3.8. Boosting

Boosting es una técnica de Machine Learning que se basa en el uso secuencial de modelos simples. Consiste en aplicar repetidamente una regla o función de predicción base en diferentes subconjuntos del conjunto de datos inicial (Barboza y cols., 2017), este enfoque aprovecha la dependencia que se genera entre los modelos simples. Las predicciones de los modelos simples se combinan mediante ponderación para obtener el resultado final de la clasificación. El método genera una secuencia de clasificadores base a través de muestras obtenidas mediante la ponderación de los datos de entrenamiento en múltiples iteraciones (Louzada y cols., 2016). Inicialmente, todos los conjuntos de entrenamiento tienen el mismo peso, pero a medida que avanza el algoritmo, se asigna mayor ponderación a las predicciones incorrectas antes de entrenar el siguiente modelo en la secuencia.

En este enfoque, se utilizan modelos de aprendizaje débiles, como árboles de decisión poco profundos.

El peso de la instancia se calcula mediante la siguiente fórmula:

$$D_t(i) = \frac{D_{t-1}(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{7}$$

Donde error_t es el error ponderado del clasificador débil h_t en la etapa t y Z_t es un factor de normalización.

La actualización del peso se realiza mediante la siguiente fórmula:

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\text{error}_t}{\text{error}_t}\right) \tag{8}$$

Esta descripción del algoritmo de Boosting coincide con la metodología mencionada en el artículo de García, Marqués y Sánchez (García, Marqués, y Sánchez, 2019). Se hace especial hincapié en el proceso de asignación de pesos a los ejemplos de entrenamiento en cada iteración y en la utilización de modelos débiles como clasificadores base en el proceso de Boosting. Además, se menciona la actualización del peso de la instancia ($D_t(i)$) y el cálculo del coeficiente de aprendizaje (α_t), que son componentes fundamentales del algoritmo de Boosting.

3.9. Extreme Gradient Boosting (XGBoost)

XGBoost es una implementación optimizada de árboles de decisión que utiliza el método de boosting para mejorar la precisión del modelo. Ha ganado popularidad debido a su eficacia en la mayoría de los conjuntos de datos y su capacidad para manejar problemas de regresión y clasificación (Hua y cols., 2020; Xiaojun y cols., 2018).

Función Objetivo:

$$\mathscr{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{9}$$

Donde $l(y_i, \hat{y}_i)$ es la función de pérdida y $\Omega(f_k)$ es la regularización del árbol, como se describe en (Xiaojun y cols., 2018).

Regularización (Penalización del Árbol):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{10}$$

Donde γ y λ son hiperparámetros que controlan la penalización del árbol y la complejidad del modelo, respectivamente, según lo mencionado en (Xiaojun y cols., 2018).

4. Metodología

La preparación meticulosa de los datos constituye un paso esencial en cualquier investigación, particularmente cuando se trata de analizar datos complejos como los que se encuentran en el mercado secundario de créditos de libranza en Colombia. Este proceso implica una serie de etapas fundamentales para garantizar que los datos recolectados y analizados sean de la más alta calidad, fiabilidad y robustez. En primer lugar, es necesario recopilar los datos relevantes de manera sistemática, lo que incluye identificar las fuentes de datos más pertinentes y asegurar su integridad. A continuación, se realiza un análisis exhaustivo para comprender la naturaleza y las características de los datos, identificando posibles inconsistencias o errores. Finalmente, se procede a la preparación de los datos, que incluye la limpieza y transformación de la información para que sea adecuada para el análisis. Estos pasos son cruciales para asegurar que los resultados obtenidos a partir del análisis sean precisos y representen de manera fiel la realidad del mercado, facilitando así la identificación de patrones, tendencias y conclusiones relevantes.

El mercado secundario de compra y venta de créditos de libranza en Colombia es un entorno complejo que involucra a diversos actores, como los Fondeadores y Originadores. Cada Originador opera con sus propias estructuras y procedimientos para llevar a cabo las actividades de originación, administración y recaudación de los créditos. En este contexto, es común observar estructuras de garantías asociadas a los créditos adquiridos. Estas garantías sirven como una especie de salvaguarda en caso de incumplimiento por parte de los prestatarios, proporcionando una capa adicional de seguridad para los actores involucrados en las transacciones. Sin embargo, es importante señalar que el establecimiento de estas garantías puede resultar costoso para los Originadores, ya que implica mantener ciertos activos en garantía o satisfacer necesidades de liquidez, lo que puede impactar en la eficiencia y la flexibilidad operativa de las entidades involucradas.

Dada la complejidad del negocio y las garantías asociadas, así como el riesgo operacional involucrado en la incorporación de créditos en las pagadurías, es fundamental considerar la mora superior a 90 días como un indicador crucial de incumplimiento en los créditos de libranza adquiridos por el fondo. Este enfoque es necesario para gestionar adecuadamente el riesgo y garantizar la estabilidad financiera de las operaciones. No obstante, la industria enfrenta el desafío adicional de adaptar sus modelos predictivos a las múltiples variables del mercado, que pueden influir en la estabilidad de las proyecciones de mora. La capacidad de ajustar y recalibrar estos modelos en respuesta a cambios en el entorno económico y financiero es

vital para mantener la precisión en las predicciones y asegurar una gestión eficaz del riesgo en el mercado secundario de créditos de libranza.

4.1. Recolección de los Datos

Los datos utilizados en este estudio provienen de una base de datos SQL que contiene información detallada sobre los créditos de libranza originados por entidades no bancarias y adquiridos por un fondo de inversión colectiva en el mercado secundario colombiano. En total, la base de datos contiene registros de 133,184 créditos adquiridos durante la operación del fondo. A continuación, se detalla el proceso de recolección de datos y la naturaleza de la información recopilada. La mayoría de los créditos están dirigidos a un nicho muy específico, que corresponde a las personas que no tienen historial bancario o que han sido reportadas en centrales de riesgo. En general, las características de cada crédito son muy similares, con poca variabilidad.

4.2. Entendimiento de los Datos

A continuación, se proporciona una descripción detallada de las variables utilizadas en el análisis y se discute su relevancia para el estudio. Estas variables proporcionan información crucial sobre cada crédito, su negociación y el pagador asociado. A través de un análisis exhaustivo de estas variables, se busca comprender mejor la dinámica del mercado de créditos de libranza y su impacto en los eventos de incumplimiento.

Nombre	**Descripción**
Contraparte	Originador del crédito, quien realiza el desembolso inicial del crédito.
Cuota del Crédito	Monto mínimo que debe pagar el deudor mensualmente según el compromiso de pago en el momento de la originación.
Número de Cuotas	Cantidad total de cuotas que debe pagar el deudor desde la originación del crédito, si el plan de amortización se lleva según las condiciones iniciales del crédito.

Nombre	**Descripción**
Tasa	Tasa en términos efectivos anuales a la cual se prestaron los recursos al deudor en el momento de la originación del crédito.
Mes de Desembolso	Mes en el cual se realizó el desembolso del crédito al deudor.
Empresa	Pagaduría encargada de realizar los giros de las cuotas a los patrimonios autónomos, corresponde al empleador o administrador de pensiones del deudor.
Salario	Salario fijo devengado por el deudor en el momento de solicitar el crédito.
Sexo	Género del deudor, con valores masculino o femenino según la información proporcionada por el deudor.
Ciudad	Ciudad en la cual se realiza la solicitud de crédito por parte del deudor.
Edad	Edad actual del deudor.
Score	Puntaje crediticio otorgado al deudor en el momento de la solicitud del crédito.
Tipo de Deudor	Categoría que clasifica al deudor en: • Pensionado: Pagaduría corresponde a un administrador de pensiones, puede ser pensión o renta vitalicia. • Privada: Empleado de una empresa privada, con la pagaduría siendo una compañía privada. • Pública: Empleado de una empresa pública, con la pagaduría siendo una entidad estatal. • FFAA: Deudores relacionados con alguna rama de las Fuerzas Armadas colombianas.

Nombre	Descripción
Default	Variable booleana que indica si el crédito se considera en incumplimiento, con 1 representando incumplimiento y 0 representando buen comportamiento y al día.

Tabla 4.1: *Descripción de Variables*

4.3. Preparación de los Datos

Para asegurar la calidad y coherencia de los análisis, se implementaron una serie de pasos para limpiar y preparar los datos. En esta sección, se detallan los procesos de limpieza y ajuste llevados a cabo, así como las razones detrás de cada decisión. Estos pasos son fundamentales para eliminar posibles errores y sesgos en los datos, garantizando la integridad de los análisis posteriores.

4.3.1. Codificación de los Datos

Para preservar la confidencialidad de las operaciones realizadas por el fondo de inversión, se aplicó un proceso de codificación a todas las variables categóricas en el conjunto de datos. Este paso asegura que la información sensible y específica no sea directamente identificable, al tiempo que permite mantener la integridad y la utilidad de los datos para el análisis y modelado subsiguiente.

4.3.2. Análisis del Negocio

Se realizó un análisis exhaustivo de las variables más relevantes del negocio, considerando las condiciones de negociación establecidas por el fondo con sus contrapartes. A continuación, se detallan las acciones tomadas para cada variable:

Nombre	Descripción
Valor del crédito	Se eliminaron registros que contenían información sobre créditos cuyos montos superaban los límites autorizados para operar en el fondo en la actualidad.
Número de cuotas	Se eliminaron registros que mostraban información sobre créditos con un número de cuotas que excedía los límites autorizados en el fondo en la actualidad.
Tasa	Se eliminaron registros que presentaban información sobre créditos con tasas que estaban claramente fuera de los estándares del mercado.
Salario	Se eliminaron registros que incluían información sobre créditos cuyos salarios eran desproporcionados en relación con el público objetivo de los Originadores.

Tabla 4.2: *Descripción de Eliminación de Registros*

Después de esta depuración, se obtuvo un conjunto de datos compuesto por 66,299.

4.3.3. Análisis Univariante

El análisis univariante permite explorar y comprender la distribución y las características de cada variable individual en los datos. A través de una serie de gráficos y estadísticas descriptivas, se examina detalladamente cada variable en busca de patrones, tendencias y valores atípicos. En esta sección, se presentan los hallazgos preliminares y se discute su relevancia para el estudio.

1. **Contraparte:** La Figura 4.1 muestra que más del 80 % de los créditos se concentran en un solo Originador, el cual es la entidad no bancaria con mayor colocación de libranzas en Colombia y con la cual el fondo tiene un amplio historial de negociaciones desde el inicio de sus operaciones.

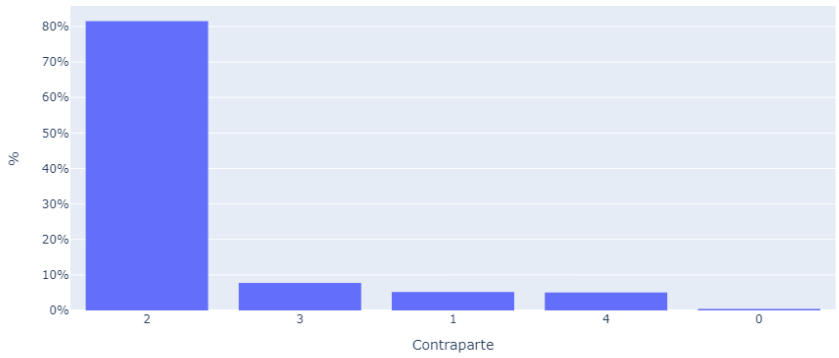

Figura 4.1: *Histograma de la cantidad de créditos comprados a cada contraparte.*

2. **Número de cuotas:** Después de la limpieza según la lógica del negocio, la Figura 4.2 muestra que el rango de cuotas es razonable para el negocio, con un mínimo de 12 cuotas desde la originación y un máximo de 152 cuotas.

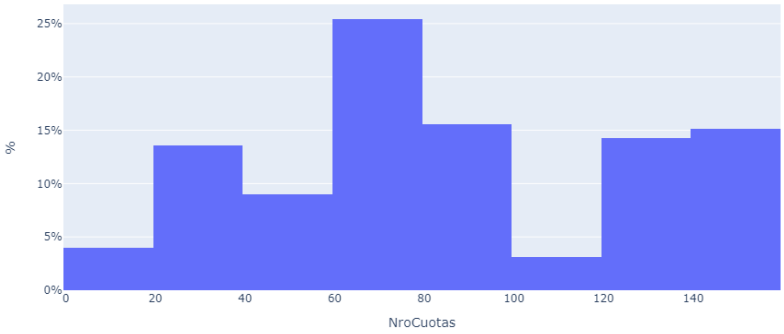

Figura 4.2: *Histograma del número de cuotas pactadas en la originación del crédito.*

Además, en el diagrama de caja y bigotes de la Figura 4.3 no se observan valores atípicos.

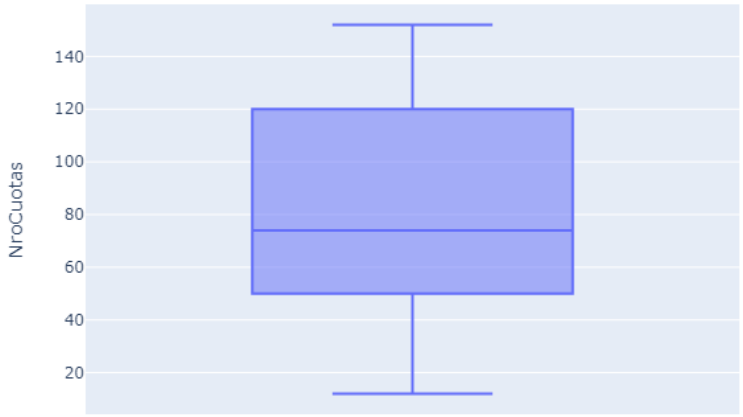

Figura 4.3: *Diagrama de caja y bigotes del número de cuotas pactadas en la originación del crédito.*

3. **Tasa:** Después de ajustar las tasas a términos efectivos anuales y eliminar las que estaban clara-

mente fuera de los estándares del mercado, se eliminó el 5 % de los datos más extremos (2.5 % en ambos lados de la distribución).

Una vez eliminados estos valores, el histograma de la variable se muestra en la Figura 4.4 y el diagrama de caja y bigotes en la Figura 4.5.

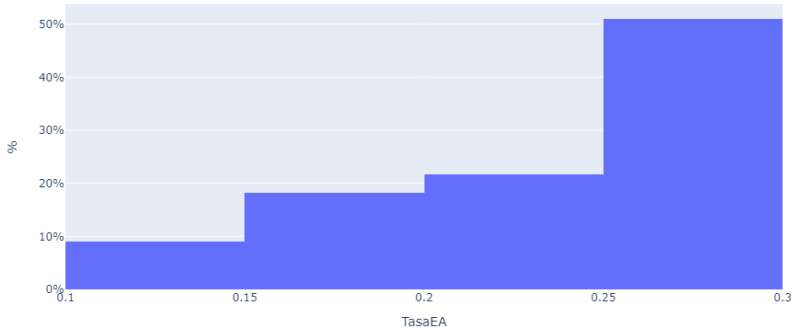

Figura 4.4: *Histograma de las tasas en términos efectivos anuales pactadas en la originación del crédito.*

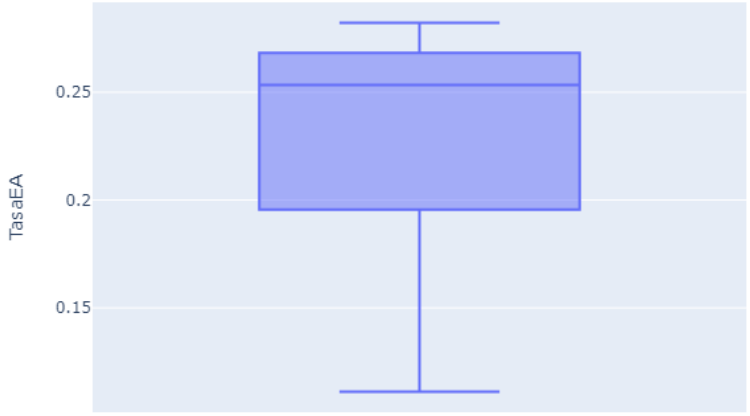

Figura 4.5: *Diagrama de caja y bigotes de las tasas en términos efectivos anuales pactadas en la originación del crédito.*

Las Figuras 4.4 y 4.5 muestran el funcionamiento de los Originadores no bancarios, los cuales tienen como objetivo principal originar créditos a personas no bancarizadas o a personas reportadas en centrales de crédito. Las personas suelen acceder a créditos con tasas de interés altas con el objetivo de ingresar al sistema bancario, hacer un buen historial y poder realizar una compra de cartera a tasas bajas.

4. **Mes de Desembolso:** La Figura 4.6 muestra que el mes con la mayor cantidad de desembolsos es octubre, mientras que los demás meses son relativamente estables.

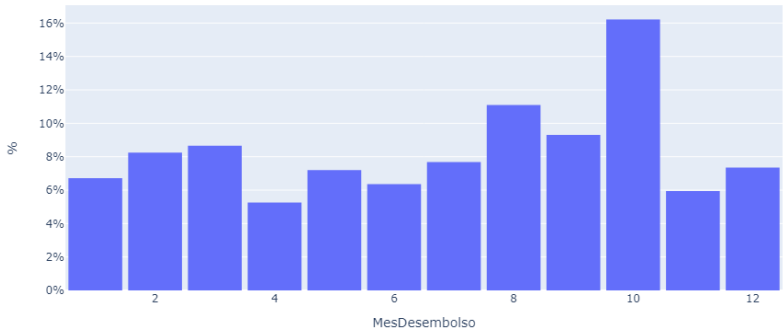

Figura 4.6: *Histograma del número de desembolsos por mes.*

5. **Empresa:** La variable que indica la empresa o entidad pagadora de los créditos presenta altas concentraciones en administradoras de fondos de pensiones, compañías del estado y algunas entidades de las fuerzas armadas, como se observa en la Figura 4.7.

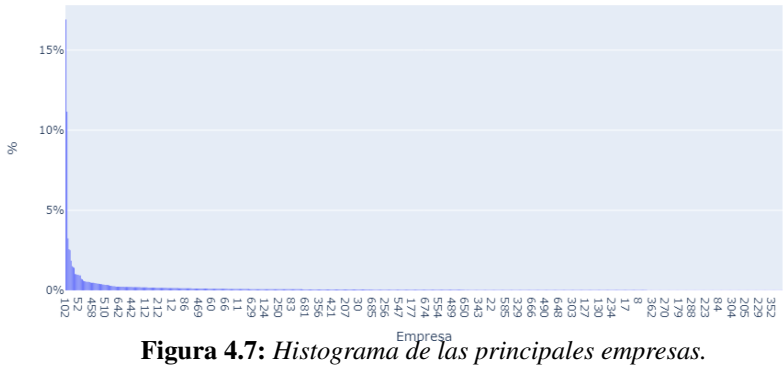

Figura 4.7: *Histograma de las principales empresas.*

6. **Valor del crédito:** Después de eliminar el 1 % de los valores extremos de la distribución, el valor de los créditos se encuentra en un rango de 980,000 a 20,000,000 pesos, como se muestra en el diagrama de caja y bigotes de la Figura 4.8.

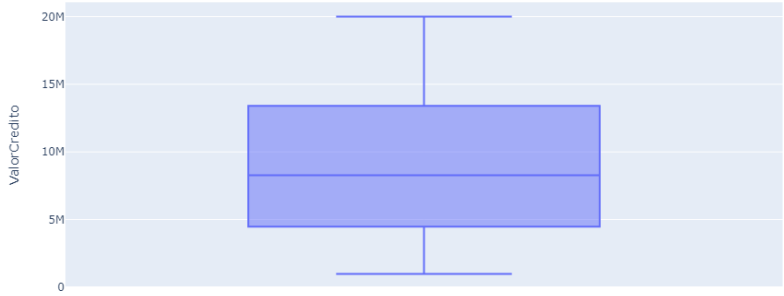

Figura 4.8: *Diagrama de caja y bigotes del valor de los créditos.*

De acuerdo con la naturaleza del negocio, el 75 % de los créditos tienen un valor menor a 13,400,000 pesos. El fondo busca optimizar el riesgo crediticio a través de la diversificación.

7. **Valor de la cuota:** Después de eliminar el 1 % de los valores extremos de la distribución, el valor de las cuotas se encuentra en un rango de 29,000 a 677,000 pesos, como se muestra en el diagrama de caja y bigotes de la Figura 4.9.

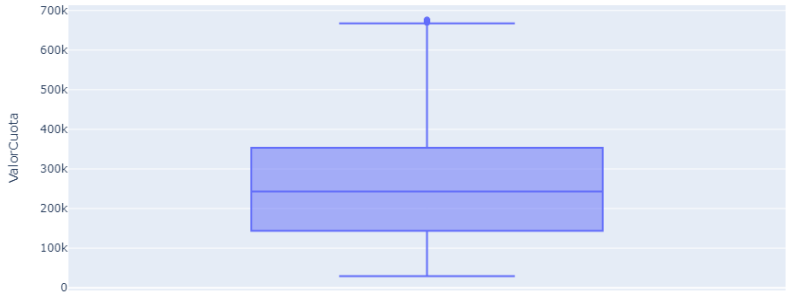

Figura 4.9: *Diagrama de caja y bigotes del valor de la cuota de los créditos.*

En línea con el perfil del público objetivo, el 75 % de los créditos tienen cuotas menores a 354,000 pesos.

8. **Salario:** Después de eliminar el 5 % de los valores extremos de la distribución de los salarios, estos se encuentran en un rango de 600,000 a 2,340,000 pesos, como se muestra en el diagrama de caja y bigotes de la Figura 4.10.

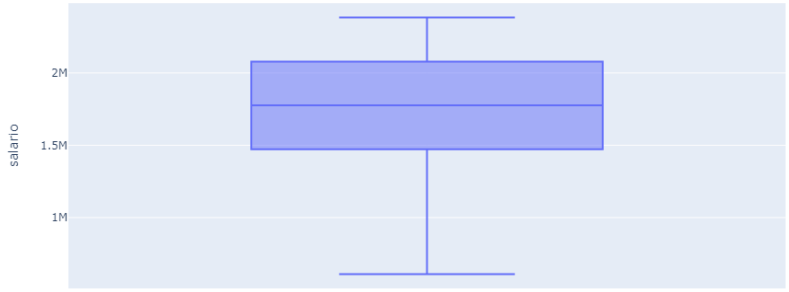

Figura 4.10: *Diagrama de caja y bigotes de los salarios de los deudores.*

En línea con el perfil del público objetivo, el 75 % de los créditos tienen un salario fijo menor a 2,080,000 pesos.

9. **Sexo:** Existe una desproporción entre hombres y mujeres, como se puede observar en la Figura 4.11, donde la base de datos tiene más del doble de registros de hombres en comparación con los registros de mujeres.

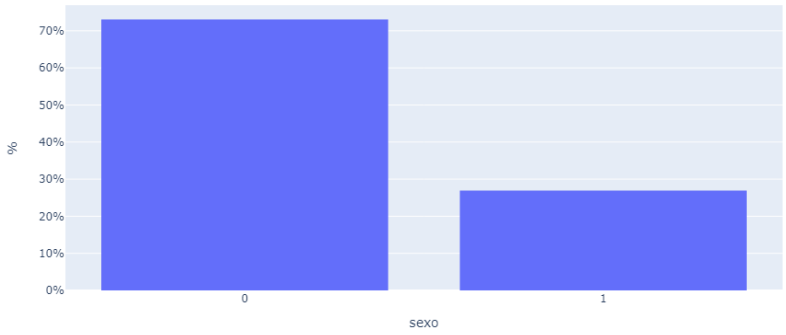

Figura 4.11: *Histograma del sexo de los deudores.*

10. **Ciudad:** La variable que indica la ciudad de solicitud del crédito muestra 605 ciudades de Colombia, de las cuales se seleccionan las 15 ciudades con mayor frecuencia y el resto se agrupan en una

categoría, como se observa en la Figura 4.12.

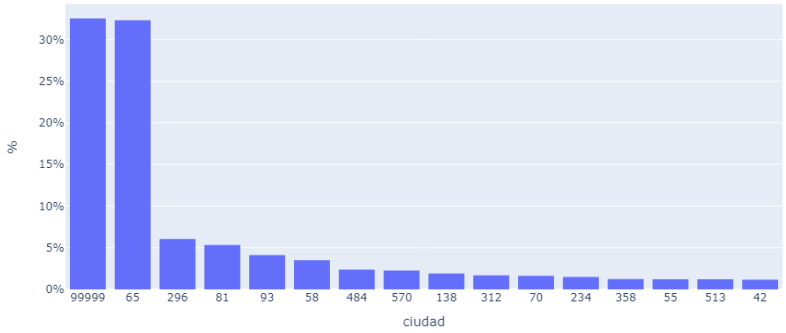

Figura 4.12: *Histograma de las ciudades de residencia de los deudores.*

Además, se eliminan las ciudades con una frecuencia menor a 30.

11. **Tipo deudor:** La concentración en pensionados, seguida de las pagadurías del estado o las fuerzas armadas, y luego las entidades privadas, es evidente en la Figura 4.13.

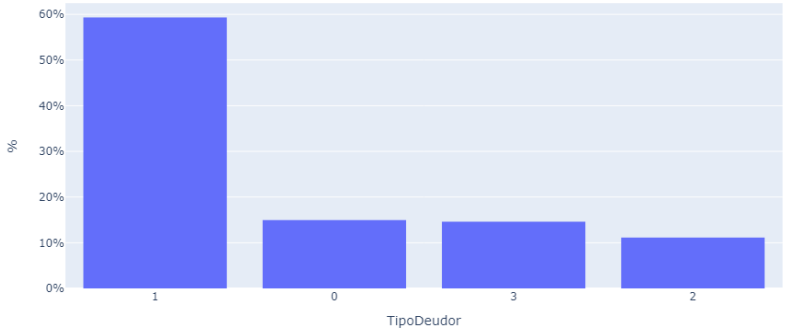

Figura 4.13: *Histograma de los tipos de deudores.*

El fondo tiene un interés particular en los deudores pensionados debido a su mejor historial de

comportamiento, donde los incumplimientos son principalmente por errores en la incorporación para pagos en la Pagaduría o por demandas de alimentos.

12. **Score:** Se observa una alta concentración de puntajes entre 150 y 500 puntos, lo cual es esperado dada la naturaleza del deudor de libranza en Colombia, como se muestra en la Figura 4.14.

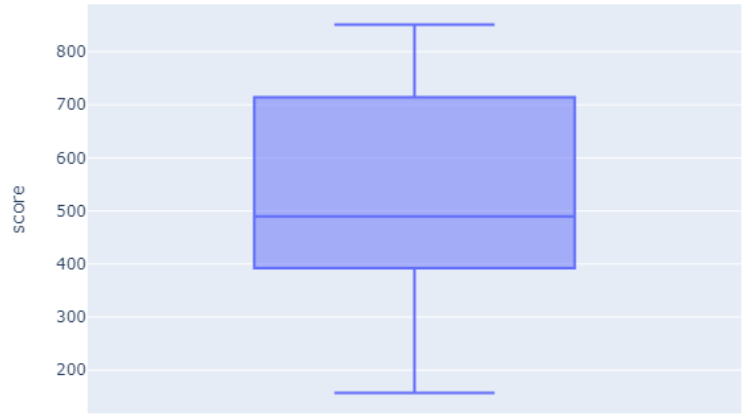

Figura 4.14: *Histograma del puntaje crediticio de los deudores.*

13. **Default:** Esta variable booleana toma el valor de 1 para indicar que el registro corresponde a un crédito de libranza que ha excedido los 90 días de mora. Cuando su valor es 0, indica que el crédito aún no ha alcanzado este nivel de mora.

    Después de la limpieza de las variables, la distribución de la variable de incumplimiento se encuentra muy balanceada, como se muestra en la Figura 4.15.

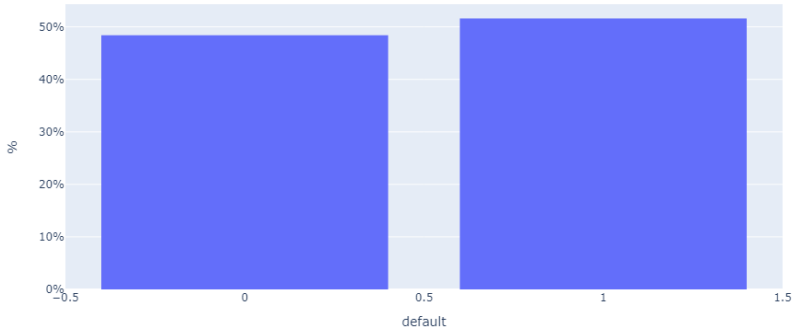

Figura 4.15: *Histograma de créditos en incumplimiento.*

4.4. Nuevas Variables Generadas

4.4.1. Proporción Cuota/Salario

Esta variable calcula la proporción del salario fijo del deudor que se destina al pago de la cuota del crédito, proporcionando una medida de la carga financiera en relación con los ingresos fijos. Se crea con el propósito de obtener una mejor comprensión del compromiso de liquidez de cada deudor, mediante el análisis de la proporción del salario destinada al pago de las cuotas.

4.4.2. Ratio Salario/Valor del Crédito

Esta variable indica cuántas veces el salario del deudor cubre el valor total del crédito desembolsado, proporcionando una perspectiva sobre la relación entre el salario y la magnitud del crédito adquirido. Al crear esta variable, se busca evaluar cuántas veces la deuda equivale al salario del deudor, lo que permite analizar la capacidad de pago de la deuda.

4.5. Modelado

Para la construcción de modelos de selección de créditos, se exploraron varias técnicas de aprendizaje supervisado, considerando la naturaleza compleja y dinámica del riesgo crediticio. Para definir el mejor modelo, se llevó a cabo un proceso de búsqueda de hiperparámetros para optimizar cada uno de los modelos de selección de créditos, los cuales fueron posteriormente comparados. Los modelos analizados incluyen regresión logística, métodos de ensamble, máquinas de aprendizaje y métodos de refuerzo.

4.5.1. Búsqueda de Hiperparámetros

Para cada modelo considerado, se realizó una exhaustiva búsqueda de hiperparámetros con el objetivo de encontrar la combinación óptima que maximizara su desempeño predictivo. Se utilizó una técnica de búsqueda en la cuadrícula (*Grid Search*), la cual exploró sistemáticamente un conjunto predefinido de valores para cada hiperparámetro.

4.5.2. Modelos Considerados

Se procedió a implementar los modelos teóricamente fundamentados anteriormente. Se han desarrollado y evaluado diversos enfoques de aprendizaje automático, incluyendo SVM, regresión logística, boosting, bagging, XGBoost y random forest. Estos modelos fueron seleccionados por su capacidad demostrada para abordar la predicción de riesgo crediticio. Además de los modelos que se desarrollarán, se comparará la eficiencia de un modelo basado en la puntuación crediticia otorgada por el originador a los deudores en el momento del estudio de crédito previo al desembolso. En este contexto, los créditos con puntuaciones más bajas son aquellos que tienen una mayor propensión a presentar un default.

4.5.3. Evaluación y Comparación de Modelos

Una vez completada la búsqueda de hiperparámetros, se evaluó cada modelo utilizando métricas de rendimiento estándar para problemas de clasificación, incluyendo precisión, recall, puntaje F1 y matriz de confusión. Además, se compararon los resultados obtenidos por cada modelo para identificar aquellos con el mejor rendimiento predictivo para el conjunto de datos.

5. Resultados y Conclusiones

5.1. Métricas de los modelos

Para determinar el modelo más adecuado para la clasificación de los nuevos créditos adquiridos por el fondo, se utilizó la métrica F1-score. A continuación, se presentan los resultados de las métricas analizadas:

Modelo	**Precisión**	**Puntuación F1**	**Recall**
Regresión Logística	0.51	0.46	0.51
BaggingClassifier	0.85	0.85	0.85
Random Forest	0.73	0.72	0.72
GradientBoostingClassifier	0.79	0.80	0.79
SVM	0.53	0.42	0.51
XGBoost	0.74	0.74	0.76
Score de Crédito	0.84	0.64	0.52

Tabla 5.1: *Resultados de Evaluación de Modelos*

Como se puede observar en la 5.1, el mejor F1 score alcanzado por un modelo de clasificación fue de 0.85. Este valor indica un buen desempeño del modelo en términos de equilibrio entre precisión y recall, reflejando un adecuado equilibrio entre los verdaderos positivos y los falsos positivos. Es decir, el modelo mantiene un equilibrio entre las predicciones de créditos que no deberían presentar un default y aquellos que realmente no lo presentan, así como entre los créditos que se predicen como propensos al default y aquellos que realmente presentan un default. En otras palabras, el modelo muestra un equilibrio en la identificación de créditos sanos y aquellos que no lo son, reduciendo tanto los errores de clasificación de créditos saludables como de créditos con problemas de pago adquiridos por el Fondo.

5.2. Matriz de confusión

A continuación, se muestran los resultados de la matriz de confusión para cada uno de los modelos implementados.

1. **Regresión Logistica**

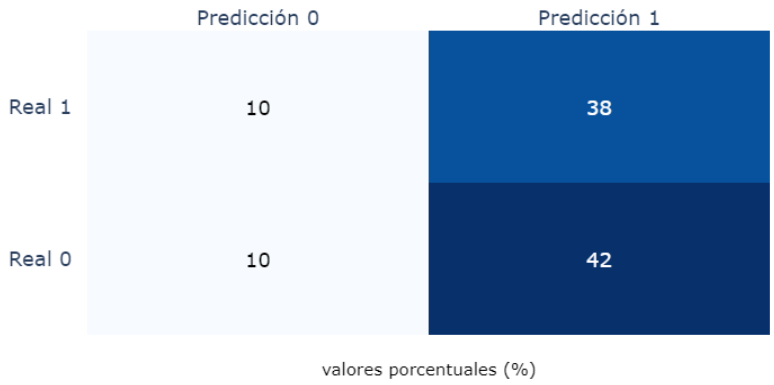

Figura 5.1: *Matriz de confusión Regresión Logistica.*

La regresión logística arrojó que el 80 % de los créditos presentarían un evento de default. Este resultado revela un sesgo en el modelo, que tiende a clasificar un alto porcentaje de créditos como propensos al default. Para el negocio del Fondo, esto representa un problema significativo, ya que solo se aceptaría el 10 % de la base de datos recibida, mientras que el 50 % de los créditos aceptados presentarían eventos de default.

2. **Bagging**

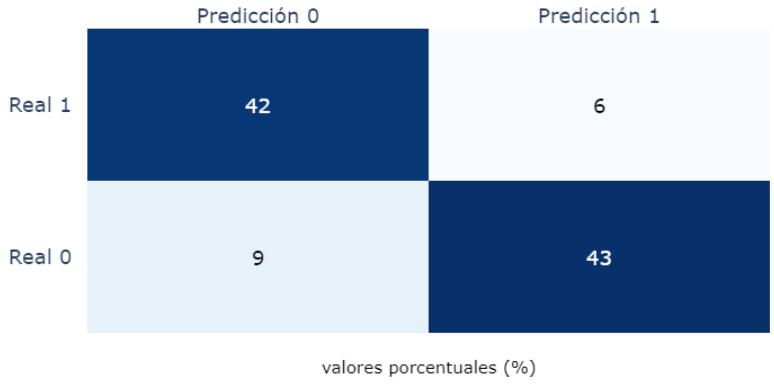

Figura 5.2: *Matriz de confusión Bagging.*

El modelo de Bagging obtuvo el mejor resultado, lo cual se refleja claramente en la matriz de confusión, donde se observa que el 85 % de los datos se encuentran en la diagonal. Esto indica un alto nivel de precisión en las predicciones del modelo.

3. **Random Forest**

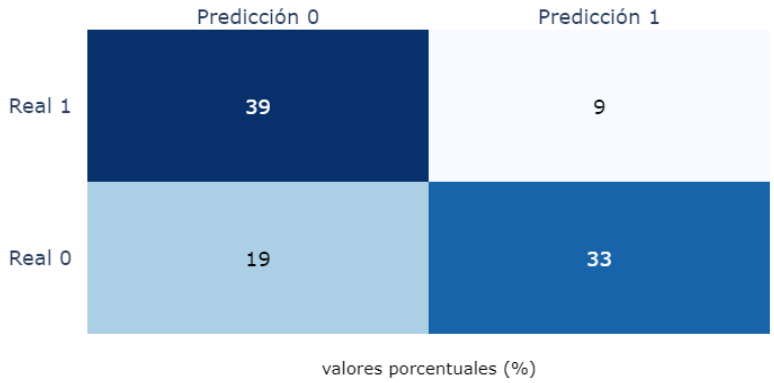

Figura 5.3: *Matriz de confusión Random Forest.*

El modelo basado en Random Forest mostró un desempeño aceptable, con aproximadamente el 70 % de las observaciones ubicadas en la diagonal de la matriz de confusión. Esto sugiere una buena precisión en las predicciones realizadas por el modelo.

4. **Máquinas de Vectores de Soporte (SVM)**

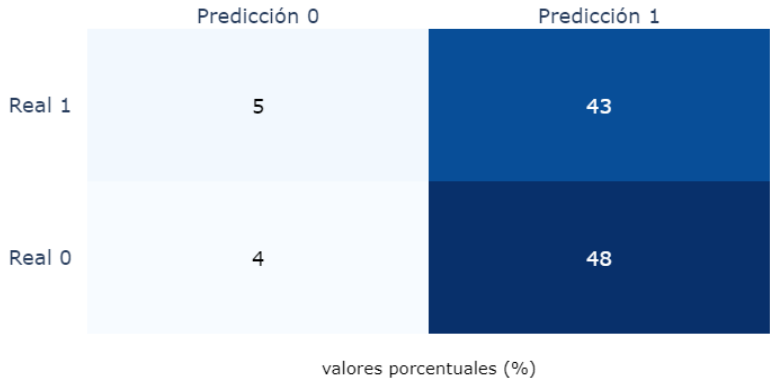

Figura 5.4: *Matriz de confusión svm.*

El modelo de SVM muestra un sesgo similar al de la regresión logística, con más del 90 % de las observaciones clasificadas como créditos que presentarán un evento de default. Esto indica una tendencia del modelo a predecir un alto porcentaje de créditos como propensos al default.

5. **Boosting**

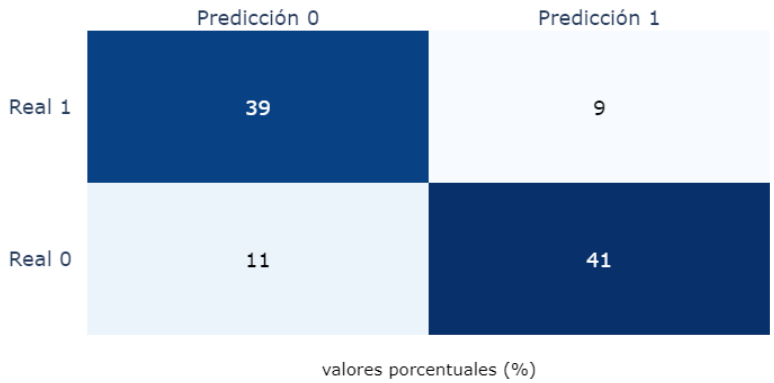

Figura 5.5: *Matriz de confusión Boosting.*

El modelo basado en Boosting mostró un desempeño destacado, con alrededor del 80 % de las observaciones ubicadas en la diagonal de la matriz de confusión. Esto sugiere una muy buena precisión en las predicciones realizadas por el modelo.

6. **XGBoost**

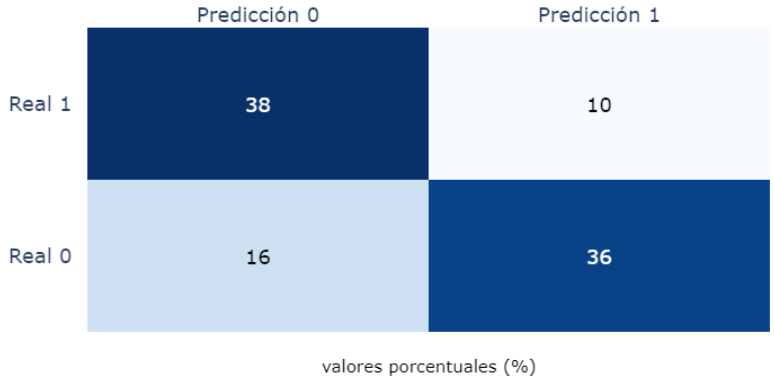

Figura 5.6: *Matriz de confusión XGBoost.*

El modelo XGBoost mostró un buen desempeño, con aproximadamente el 74 % de las observaciones ubicadas en la diagonal de la matriz de confusión. Esto indica una muy buena precisión en las predicciones realizadas por el modelo.

7. **Score**

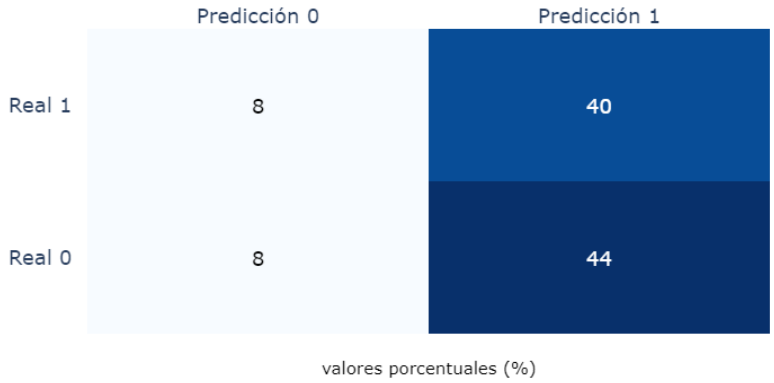

Figura 5.7: *Matriz de confusión selección score mayor a 350.*

Por último, un modelo basado en el puntaje crediticio otorgado a cada deudor por parte del originador muestra un claro sesgo hacia el rechazo de créditos debido a la posibilidad de default. Este sesgo es coherente con el enfoque del negocio de los originadores, que a menudo trabajan con clientes que están reportados en centrales de riesgo o que tienen un acceso limitado al crédito.

5.3. Conclusiones

Este estudio se centró en la evaluación de modelos de clasificación para predecir el riesgo de default en créditos de libranza adquiridos por un fondo de inversión en el mercado secundario colombiano. Se compararon varios modelos, incluyendo SVM, regresión logística, boosting, bagging, XGBoost y random forest, encontrando que el BaggingClassifier destacó como el más efectivo con un F1-Score de 0.85. El éxito del BaggingClassifier radica en su capacidad para reducir la varianza mediante la combinación de múltiples modelos entrenados en diferentes subconjuntos de datos. Este enfoque permite mejorar significativamente la generalización del modelo, especialmente dado la complejidad de los datos analizados.

Comparado con SVM, regresión logística y métodos de boosting, el BaggingClassifier demostró mayor precisión y robustez en la clasificación del riesgo crediticio. Sus resultados sugieren que representa una estrategia efectiva y confiable para mitigar riesgos en la gestión de créditos de libranza en mercados secundarios. Este estudio proporciona un marco robusto para la evaluación de modelos de clasificación en la predicción de riesgo crediticio, destacando la eficacia del BaggingClassifier como una herramienta prometedora para optimizar las operaciones financieras en el sector de créditos de libranza, especialmente

dirigidas a segmentos específicos de la población colombiana.

A continuación, se presentan las matrices de confusión de los diferentes modelos evaluados. Cada matriz de confusión proporciona una visión detallada del desempeño de cada modelo en términos de clasificación de las clases objetivo.

5.4. Recomendaciones

Para mejorar la precisión y la gestión del riesgo en la clasificación de créditos de libranza, se recomienda considerar la implementación de modelos probabilísticos específicos, como la probabilidad por días de mora, la curva de supervivencia y el análisis porcional de riesgos de Cox. Estos enfoques permiten una evaluación más dinámica y granular del comportamiento crediticio a lo largo del tiempo, facilitando una mejor estimación de la probabilidad de incumplimiento en diferentes puntos temporales. La utilización de estos modelos no solo podría mejorar la precisión de las predicciones, sino también proporcionar insights más detallados sobre los factores que contribuyen al riesgo crediticio en el contexto específico de créditos de libranza adquiridos en el mercado secundario colombiano. Además, la integración de estas técnicas podría fortalecer las estrategias de mitigación de riesgos y optimización de decisiones financieras, beneficiando tanto a los fondos de inversión como a los beneficiarios de los créditos.

Referencias

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of cor- porate bankruptcy. *The Journal of Finance*, *23*, 589-609.

Barboza, F., Kimura, H., y Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417.

BBVA, B. (2024). *¿qué es libranza?* Descargado de `https://www.bbva.com.co/personas/blog/educacion-financiera/prestamos/que-es-libranza.html` (Ingreso url en Junio 4, 2024)

Bogotá, B. (2024). *Crédito de libranza.* Descargado de `https://www.bancodebogota.com/wps/portal/banco-de-bogota/bogota/productos/para-ti/creditos-y-financiacion/credito-de-libranza` (Ingreso url en Junio 4, 2024)

Bohn, J. R., y Stein, R. (2009). *Active credit portfolio management in practice*. Wiley finance.

Breiman, L. (1996). Bagging predictors. *Kluwer Academic Publishers*, *24*, 123-140.

Chen, S., Härdle, W. K., y Moro, R. A. (2011). Modeling default risk with support vector machines. *Quantitative Finances*, *11*, 135-154.

García, V., Marqués, A. I., y Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, *47*, 88-101.

Hua, L., Yumeng, C., Siwen, L., Jianbin, Z., y Yutong, S. (2020). Xgboost model and its application to personal credit evaluation. *IEEE Intelligent Systems*, *35*, 52-61.

Kruppa, J., Schwarz, A., Arminger, G., y Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*, 5125-5131.

Lean, Y., Shouyang, W., y Keung, L. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, *34*, 1434-1444.

Louzada, F., Ara, A., y Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Manage-*

ment Science, *21*, 117-134.

M., D., K., K., G., B., y C., Z. (2002). Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis. *European Journal of Operational Research*, *138*, 392-412.

Maldonado, S., Peters, G., y Weber, R. (2018). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*, *507*, 700-714.

Malekipirbazari, M., y Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*, 4621-4631.

Mushava, J., y Murray, M. (2022). A novel xgboost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, *202*(117233).

Verbraken, T., Bravo, C., Weber, R., y Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measure. *European Journal of Operational Research*, *31*, 24-39.

Xiaojun, M., Jinglan, S., Dehua, W., Yuanbo, Y., Qian, Y., y Xueqi, N. (2018). Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 42-39.