Modelo de machine learning para predecir la carga financiera máxima de clientes en Nequi

Esteban Caro Peláez

Trabajo de Grado presentado para optar al título de Ingeniero Industrial

Modalidad de Práctica

Semestre de Industria

Asesor

Claudia Sofía Correa Puerta, Especialista (Esp) en Gerencia de Proyectos

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Industrial

Medellín, Antioquia, Colombia

2025

Cita	(Caro Peláez, 2024)
Referencia **Estilo APA 7 (2020)**	Caro Peláez, E. (2024). *Modelo de machine learning para predecir la carga financiera máxima de clientes Nequi*. [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.

Dedicatoria

A Dios, quien instruyó y guió mi camino. A mi familia, por acobijarme en un hogar sereno y amoroso. A mis amigos, por hacer que las cargas fuesen más llevaderas.

Agradecimientos

Agradecimientos al equipo de riesgo de crédito y analítica de riesgos de Nequi, por la asesoría y la compañía en este proceso.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

CRISP – DM	Cross-Industry Standard Process for Data Mining
MRM	Model Risk Management
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
KPI	Key Performance Indicator
AUC	Area Under the Curve
ICV	Índice de cartera vencida
IM	Indicador de mora
IND	Ingreso Neto Disponible
PD	Probabilidad de Default
CMP	Cuota Máxima Pagable
PSI	Population Stability Index
OOT	Out of Time

Resumen

El otorgamiento de créditos es un pilar fundamental en la rentabilidad de las entidades financieras. La lectura acertada de la realidad de los clientes es un factor clave en la aprobación de créditos, razón por la cual, en el presente proyecto se propone un modelo de machine learning para la estimación de la carga financiera máxima otorgada a un cliente en la compañía Nequi. La variable de respuesta definida refleja directamente la cuota máxima estimada para cada cliente, lo que permite establecer límites claros dentro de los cuales un cliente podría comprometerse financieramente sin incurrir en un riesgo significativo

Este modelo se desarrolla a través de la metodología CRISP – DM (Cross-Industry Standard Process for Data Mining), una forma de trabajo estándar en el ámbito de la tecnología para ciencia de datos, la cual va desde el conocimiento del negocio, etapa que se enfoca en comprender las necesidades específicas de la compañía que requiere la implementación del modelo predictivo, hasta el paso a producción de los modelos. La metodología CRISP – DM facilita la planificación y ejecución de las tareas requeridas para desarrollar un modelo analítico teniendo en cuenta que los proyectos de ciencia de datos son susceptibles a modificaciones a medida que se realizan iteraciones en los diferentes modelos.

El alcance del proyecto se limita a llegar hasta la evaluación de la precisión del modelo mas no incluye implementación. El resultado esperado es un modelo que permita maximizar la aprobación de créditos a clientes dentro del apetito de riesgo crediticio establecido por el área de riesgos de Nequi.

Palabras clave*: capacidad de pago, riesgo de crédito, modelo de regresión lineal, egresos financieros, carga financiera.*

Abstract

The granting of credit is a fundamental pillar of profitability for financial institutions. Accurately assessing clients' financial realities is a key factor in credit approval. In this context, the present project proposes a machine learning model to estimate the maximum financial burden that can be allocated to a client at Nequi. The defined response variable directly reflects the maximum estimated installment for each client, allowing the establishment of clear limits within which a client could commit financially without incurring significant risk.

This model was developed following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a standard framework in data science technology. This process spans from business understanding—focused on comprehending the specific needs of the company requiring the predictive model—to deploying the model in production. The CRISP-DM methodology facilitates the planning and execution of the tasks required to develop an analytical model, considering that data science projects are often subject to modifications as iterations are made on different models.

The scope of the project is limited to evaluating the accuracy of the model and does not include implementation. The expected outcome is a model that maximizes credit approvals for clients within the credit risk appetite defined by Nequi's risk management department.

Keywords*: repayment capacity, credit risk, linear regression model, financial expenses, financial burden.*

1. Introducción

Nequi es una compañía de financiamiento enfocada en ofrecer servicios financieros completamente digitales. En el año 2019, la compañía incursionó en el mundo de los préstamos bancarios. El área de riesgos de crédito de Nequi es responsable de determinar a cuáles clientes se les prestará dinero y qué montos se desembolsarán para dichos clientes. El riesgo de crédito está definido como la posibilidad de que Nequi incurra en pérdidas dado que un cliente no paga sus obligaciones para con la compañía. Uno de los conceptos fundamentales para evaluar la viabilidad de un desembolso crediticio es la capacidad de pago del cliente, que mide la relación entre sus ingresos y egresos.

Actualmente, al interior del área de riesgos de crédito se está haciendo un esfuerzo por mejorar el cálculo de la capacidad de pago de sus usuarios, dado que ésta es la causa principal de rechazo de créditos en el 80% de los clientes que entran al flujo de preaprobados, lo cual representa una pérdida de oportunidad para Nequi en cuanto a la administración de sus activos financieros.

En el marco del proceso de refinamiento de la capacidad de pago de los usuarios, se ha identificado la necesidad de estimar la carga financiera máxima aplicable a cada cliente. A pesar de que actualmente existen metodologías para la estimación de la capacidad de pago de los clientes, se ha identificado la necesidad de abordar este problema con un enfoque más analítico que permita tomar decisiones data – driven. El proceso de estimación de esta carga financiera es fundamental y crítico, ya que, se puede llegar a subestimar los egresos financieros de los clientes, lo cual podría resultar en la aprobación de créditos que excedan la su capacidad de pago. En este caso, los clientes podrían sobre endeudarse, lo que implicaría riesgos regulatorios para la compañía ante la Superintendencia Financiera. Por otro lado, la sobreestimación de dichos egresos podría llevar al rechazo injustificado de créditos viables y la pérdida de oportunidad sobre los activos financiero de Nequi.

En el presente trabajo se presenta el desarrollo de un modelo de machine learning como alternativa estimación de carga financiera máxima haciendo uso de la metodología CRISP – DM.

Esta es una metodología ampliamente utilizada en proyectos de análisis de datos y machine learning que proporciona un marco estructurado para abordar y resolver problemas de negocio mediante el uso de datos, comprende fundamentalmente comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Con este proyecto se espera refinar la estimación de la capacidad de pago de cada cliente, de manera tal que se refleje su realidad financiera y permita a la compañía ofrecerles créditos que se ajusten a sus necesidades individuales, optimizando la gestión del riesgo.

2. Objetivos

2.1 Objetivo general

Desarrollar un modelo de machine learning para estimar la carga financiera máxima de los egresos financieros de clientes Nequi, con el fin de mejorar la precisión en la evaluación de su capacidad de pago en el flujo de preaprobados al interior del área de riesgos de crédito.

2.2 Objetivos específicos

- Establecer las métricas de negocio y los requerimientos del área de riesgos de crédito en relación con la estimación de cuotas de tarjetas de crédito, para definir el impacto del proyecto en el proceso de flujo de preaprobados
- Explorar y analizar los datos históricos comprados a la central de información financiera, con el objeto de identificar variables relevantes, evaluar calidad de los datos y detectar tendencias que puedan afectar el rendimiento del modelo.
- Implementar el modelo de machine learning de regresión para la estimación de la carga financiera máxima
- Evaluar el mejor modelo de machine learning en función de las métricas de desempeño, con el objetivo de cumplir con los requerimientos mínimos exigidos por la mesa de riesgo de crédito
- Seleccionar el mejor modelo en función de las métricas de desempeño de modelos de machine learning y de negocio, a fin de garantizar que la solución propuesta dé respuesta a las necesidades específicas del negocio
- Documentar los hallazgos encontrados y los hiperparámetros del modelo de machine learning, para facilitar el despliegue del modelo en caso de ser aprobado por la mesa de riesgos de Nequi
- Documentar el modelo y los hallazgos encontrados durante la ejecución del proyecto, a fin de tener información que soporte la toma de decisiones relacionadas a capacidad de pago ante la Superintendencia Financiera y MRM de Bancolombia.

3. Marco teórico

3.1 Introducción al machine learning en Finanzas

3.1.1 Definición de Machine Learning

El machine learning es un campo de la inteligencia artificial que usa modelos estadísticos para realizar predicciones basadas en datos y características de los clientes. A través de las variables dependientes, se explica una independiente, así, los modelos capturan la información importante de los datos (Mitchell, 1997). Esto hace que los algoritmos de machine learning capturen patrones que de otra forma no sería posible observar, mediante las cuales se prediga un valor específico basado en características registro a registro.

La diferencia principal entre el machine learning y la programación tradicional radica en que en bajo el enfoque tradicional, se debe definir explícitamente las reglas y los pasos que el programa debe ir ejecutando para dar solución a un problema dado. Por otro lado, el machine learning adopta un enfoque diferente, bajo el cual existen unos algoritmos estadísticos previamente codificados que permiten que el sistema aprenda a partir de datos, así, en lugar de programar reglas manualmente, el algoritmo se alimenta de grandes cantidades de datos lo cual le permite identificar patrones de ellos (Géron, 2019).

Para comprender mejor el machine learning, es importante conocer sus principales categorías y cómo cada una aborda el proceso de aprendizaje de manera única. Al respecto Chollet expone:

> El machine learning se divide en tres categorías principales: el aprendizaje supervisado que opera con datos etiquetados, donde el algoritmo aprende a partir de ejemplos con resultados conocidos, como clasificar correos electrónicos como spam basándose en ejemplos previamente etiquetados. Por otro lado, el aprendizaje no supervisado trabaja con datos sin etiquetar, buscando patrones y estructuras inherentes en los datos, como agrupar clientes por

> comportamientos de compra similares sin categorías predefinidas. Finalmente, el aprendizaje por refuerzo se asemeja al proceso de aprendizaje humano mediante prueba y error, donde el algoritmo aprende a tomar decisiones a través de la interacción con un entorno, recibiendo recompensas o penalizaciones según sus acciones, similar a cómo un programa aprende a jugar al ajedrez mejorando con cada partida. (Chollet, 2021, p. 45)

3.1.2 Aplicaciones en el Sector Financiero

La inteligencia artificial ha tenido un impacto fuerte el sector financiero, ya que ha puesto sobre la mesa nuevas formas de mejorar la toma de decisiones (Aziz & Dowling, 2019). Algoritmos como los Random Forest, Decision Tree y XG – Boost permiten identificar tanto patrones lineales como no lineales en los datos, mejorando significativamente la precisión de las predicciones financieras (He et al., 2023).

> Uno de los aspectos que se ha tenido una mejora importante desde la llegada de los modelos de machine learning es la detección de fraudes financieros, puesto que estos algoritmos permiten identificar patrones sospechosos en los datos, incluso en tiempo real. Los algoritmos de machine learning pueden analizar grandes volúmenes de transacciones, detectando anomalías y comportamientos inusuales que podrían indicar actividades fraudulentas, lo que ha permitido a las instituciones financieras reducir sus pérdidas por fraude y mejorar la seguridad de sus clientes (West & Bhattacharya, 2016).

Por otro lado, en el ámbito del análisis de riesgos y evaluación crediticia, el machine learning ayuda a evaluar la solvencia de sus clientes teniendo en cuenta diferentes variables transaccionales. Esto, sumado a los análisis de tasas de morosidad con la información que se compra en los burós de crédito, ha permitido optimizar los procesos de toma de decisiones crediticias (Khandani et al., 2018).

Por último, otra rama financiera que ha tenido una mejoría ante la llegada del machine learning es la optimización de carteras de inversión mediante machine learning permite una mejor

gestión de activos. Los algoritmos de aprendizaje automático pueden analizar patrones históricos del mercado, correlaciones entre activos y factores macroeconómicos para generar estrategias de inversión más eficientes y flexibles, teniendo en cuenta volatilidades de los mercados. Esto permite una mejor diversificación del riesgo y la identificación de oportunidades de inversión (López de Prado, 2020).

3.1.3 Tendencias y desafíos en el uso de Machine Learning en Finanzas

Las innovaciones más recientes en machine learning se enfocan principalmente Modelos de Deep learning, Procesamiento del Lenguaje Natural e Inteligencia Artificial Generativa (GenAI). Estas tecnologías permiten la implementación de sistemas de trading algorítmico más sofisticados apalancándose del desarrollo de redes neuronales. La integración de técnicas de aprendizaje por refuerzo y redes neuronales profundas ha abierto nuevas posibilidades para la automatización de decisiones financieras complejas, aunque su implementación sigue siendo un reto significativo dada la multiplicidad de datos que se generan (Huang et al., 2020).

Con respecto a los desafíos éticos y de privacidad en el contexto del machine learning en finanzas, se han convertido en un tema importante tanto para las instituciones financieras como para las entidades públicas reguladoras. La recopilación masiva de datos personales para modelos predictivos plantea interrogantes sobre el consentimiento informado y la protección de la privacidad del consumidor. Además, existe una creciente preocupación sobre el sesgo algorítmico en las decisiones financieras automatizadas, que podría perpetuar o amplificar desigualdades existentes en el acceso a servicios financieros, especialmente en procesos de evaluación crediticia y gestión de riesgos (O'Dwyer & Vallor, 2023).

Por otro lado, la interpretabilidad y transparencia de los modelos de machine learning representa uno de los mayores desafíos en su implementación en el sector financiero. Los modelos de aprendizaje automático tienen una complejidad natural dado a que están fundamentados en el álgebra lineal, el cálculo vectorial, la estadística y la probabilidad, lo cual implica que la tecnología detrás de los modelos sea difícilmente comprensibles para agentes no técnicos en el tema. Esta complejidad de interpretabilidad es particularmente problemática en el sector financiero, donde las

decisiones algorítmicas pueden tener impactos significativos en la vida de las personas; sin embargo, actualmente se están enfocando están desarrollando técnicas llamadas "XAI" (Explainable Artificial Intelligence), que son metodologías que permitan mantener el rendimiento de los modelos complejos mientras se mejora su interpretabilidad (Minh, Wang, Li, & Nguyen, 2022).

3.2 Evaluación de capacidad de pago en finanzas

3.2.1 Modelos tradicionales de evaluación de crédito

Los puntajes de crédito han sido una herramienta fundamental en el sector financiero a lo largo de la historia. El modelo tradicional más conocido es el FICO Score, que asigna una puntuación basada el historial de pagos, la deuda total y la duración del historial crediticio. La metodología detrás de estos puntajes se basa en la correlación histórica entre características del solicitante y su comportamiento crediticio futuro, ofreciendo a las instituciones una herramienta cuantitativa para la toma de decisiones (Feldman & Schmidt, 2016).

También, el análisis del historial financiero es otro criterio clave en los modelos tradicionales de evaluación de crédito. Este enfoque considera el comportamiento pasado del solicitante, analizando su capacidad para cumplir con obligaciones financieras anteriores. A través de este método, las entidades pueden identificar patrones de pago, incumplimientos o atrasos que son indicativos de posibles riesgos en el futuro.

En suma, los criterios tradicionales de evaluación de riesgo suelen incluir, además de los factores mencionados, la relación deuda-ingreso (carga financiera) y variables transaccionales como movimientos, extractos, entre otros. Estos criterios se utilizan para complementar el puntaje de crédito y proporcionar una visión más amplia de la capacidad de pago del cliente. Algo importante que cabe mencionar es que estos modelos se implementan de manera estandarizada, lo que puede limitar la capacidad de personalizar la evaluación según el perfil específico del cliente (Jagtiani & Lemieux, 2018).

3.2.1 Importancia de estimar cuotas de tarjetas de crédito

La estimación de las cuotas de tarjetas de crédito juega un papel importante en la gestión del riesgo financiero. Al predecir correctamente los pagos mensuales que un cliente deberá realizar por concepto de cuotas de tarjetas de crédito, las instituciones financieras pueden anticipar la carga financiera que este soportará y ajustar su capacidad de endeudamiento. Esto no solo mitiga el riesgo de impago, sino que también permite tomar decisiones más informadas sobre la aprobación de líneas de crédito. La incorporación de estimaciones precisas sobre las cuotas mejora la precisión en la evaluación del riesgo crediticio (Huang & Tang, 2019).

Además, una estimación adecuada de las cuotas tiene un impacto importante en la rentabilidad de las instituciones crediticias. Un cálculo preciso permite ajustar las tasas de interés y otras condiciones crediticias de acuerdo con el perfil de riesgo del cliente, optimizando así los márgenes de ganancia. Las entidades que manejan esta información eficientemente tienden a mejorar sus indicadores de rentabilidad a largo plazo, ya que logran mantener un equilibrio entre riesgo y retorno. Estimaciones erróneas pueden llevar a la sobreexposición al riesgo o a la infravaloración del cliente, afectando negativamente el rendimiento de la cartera de crédito (Kim & Yoon, 2020).

Por último, un ejemplo de cómo una estimación precisa puede mejorar las decisiones crediticias es el caso de el caso de los clientes que se estiman de manera acertada y así se tiene una visión clara de cómo son estos clientes en términos de sus ingresos y egresos, puesto que la cuota se toma como un egreso y si el egreso se calcula de manera adecuada, se puede entender cuánto de sus ingresos le queda restante al cliente para poder adquirir una obligación más.

3.2.3 Factores que afectan las cuotas de tarjeta de crédito

Las tasas de interés y los límites de usura emitidas y reguladas por el Banco de la República son factores clave que afectan directamente las cuotas de las tarjetas de crédito. Las tasas de interés determinan el costo del crédito, y cuando estas son elevadas, el monto de las cuotas mensuales también aumenta. Además, los límites de usura establecidos por el Bando Central buscan proteger a los consumidores de tasas excesivamente altas. Las fluctuaciones en las tasas de interés influyen

significativamente en la morosidad, ya que cuotas más altas pueden generar dificultades financieras para los clientes, incrementando el riesgo de impago (Mendoza & Pérez, 2018).

Además de esto, el saldo de la tarjeta y el comportamiento de gasto del cliente también son determinantes críticos en la definición de las cuotas mensuales de un cliente. A medida que un cliente acumula más deuda en su tarjeta de crédito, el saldo a pagar cada mes aumenta. Adicional a esto, un comportamiento de gasto excesivo puede resultar aumentando las cuotas y, por ende, el riesgo de sobreendeudamiento. Los clientes con altos niveles de gasto y saldos pendientes suelen enfrentar cuotas más elevadas y son más propensos a incumplir con sus pagos (Lee & Chan, 2020).

Por último, los cambios en las políticas crediticias y las condiciones económicas también influyen en las cuotas de las tarjetas de crédito. En periodos de recesión económica las instituciones financieras suelen modificar sus políticas de crédito, lo que puede resultar en un aumento en las cuotas para mitigar el riesgo percibido. Estos cambios terminan siendo condiciones más estrictas para los consumidores y mayores requisitos para la aprobación de créditos. Las modificaciones en las políticas económicas y crediticias son un reflejo de la necesidad de adaptarse a escenarios macroeconómicos fluctuantes, lo que afecta directamente el costo y las cuotas de los productos crediticios (García & Morales, 2021).

3.3 Métricas de desempeño en Modelos de Machine Learning

3.3.1 Métricas de evaluación de modelos de regresión

El Error Cuadrático Medio (MSE) es una de las métricas más comunes para evaluar la calidad de un modelo de regresión. Mide el promedio de los cuadrados de los errores o diferencias entre los valores reales y los valores predichos. Al penalizar los errores grandes de manera más severa que los pequeños, el MSE da mayor peso a las predicciones que se alejan mucho de los valores reales, lo que permite identificar si el modelo está cometiendo grandes errores de predicción (Chai & Draxler, 2014).

La Raíz del Error Cuadrático Medio (RMSE) es una métrica comúnmente utilizada para evaluar el desempeño de modelos de predicción. Al calcular la raíz cuadrada del error cuadrático medio, el RMSE permite obtener una medida del error promedio en las mismas unidades que la variable objetivo, facilitando así su interpretación (Seber, 2003).

El Error Absoluto Medio (MAE) es otra métrica de evaluación de modelos de regresión que calcula el promedio de las diferencias absolutas entre los valores reales y los valores predichos. El MAE es especialmente útil cuando los errores grandes no son considerados problemáticos o cuando se busca un error promedio sencillo de interpretar (Willmott & Matsuura, 2005).

El Coeficiente de Determinación (R^2) mide la proporción de la varianza en la variable dependiente que es explicada por el modelo de regresión. Un valor de cercano a 1 indica que el modelo explica bien la variabilidad de los datos, mientras que un valor cercano a 0 sugiere que el modelo no tiene capacidad predictiva. Sin embargo, el R^2 puede ser engañoso en algunos contextos, ya que tiende a aumentar con la adición de variables independientes al modelo, independientemente de si estas mejoran realmente el ajuste (Nagelkerke, 1991).

3.3.2 Comparación entre modelos

La evaluación de diferentes algoritmos de regresión es esencial para determinar cuál se ajusta mejor a un conjunto de datos específico. Entre los algoritmos más comunes se encuentran la regresión lineal, la regresión polinómica y los métodos basados en árboles como los bosques aleatorios. Mientras que la regresión lineal es fácil de interpretar, los métodos basados en árboles pueden capturar relaciones no lineales más complejas, pero corren el riesgo de sobreajustarse a los datos si no se ajustan adecuadamente (James et al., 2013). La comparación entre estos algoritmos generalmente se basa en métricas como el MSE, RMSE o MAE, con el fin de seleccionar el modelo que minimiza el error de predicción.

El desempeño de un modelo de regresión depende en gran medida de su capacidad para generalizar a nuevos datos. Como señalan Hastie, Tibshirani y Friedman (2009), el sesgo y la

varianza son dos componentes clave del error de predicción. Un modelo con alto sesgo tiende a subajustarse a los datos, mientras que uno con alta varianza tiende a sobreajustarse. Por lo tanto, el análisis de sesgo y varianza es fundamental para encontrar un modelo que ofrezca un buen equilibrio entre ambos. Así pues, el análisis de sesgo y varianza son una herramienta clave para entender el desempeño de un modelo de regresión.

Del mismo modo, el uso de técnicas de validación cruzada es fundamental para comparar el desempeño de diferentes modelos de regresión y garantizar que los resultados no dependan exclusivamente de un conjunto de datos específico. La validación cruzada divide el conjunto de datos en *k* subconjuntos, entrenando el modelo con k-1 subconjuntos y evaluándolo en el subconjunto restante. Este proceso se repite varias veces para obtener una estimación promedio del desempeño del modelo. La validación cruzada ayuda a mitigar el riesgo de sobreajuste y proporciona una estimación más robusta de la capacidad de generalización del modelo (Stone, 1974).

3.3.3 Impacto en el negocio

Las métricas de desempeño de los modelos de machine learning deben de estar alineadas con los objetivos comerciales, ya que permiten evaluar la eficacia de las soluciones predictivas en términos que dan solución para el negocio. Por ejemplo, en un entorno comercial, optimizar métricas como el MAE o RMSE no solo busca mejorar la precisión del modelo, sino que se traduce en decisiones más acertadas, como la reducción de pérdidas crediticias o la mejora en la asignación de recursos financieros. Al mejorar la exactitud de las predicciones, los modelos pueden contribuir a una mejor segmentación de clientes, colocar créditos de una forma menos riesgosa, lo que impacta directamente en la rentabilidad de la organización (Provost & Fawcett, 2013). El uso eficiente de las métricas garantiza que los modelos no solo funcionen bien desde una perspectiva técnica, sino que también se alineen con los KPI, como el aumento de ingresos o la disminución de costos.

El balance entre la precisión del modelo y los costos operativos es un desafío común en la implementación de modelos predictivos, especialmente en infraestructuras en la nube como AWS.

Los modelos más complejos pero precisos requieren más recursos computacionales, lo que se traduce en mayores costos de procesamiento y almacenamiento. Los modelos más complejos pero precisos, requieren más recursos computacionales, lo que se traduce en mayores costos de procesamiento y almacenamiento. Estos costos pueden escalar rápidamente en servicios en la nube, donde el uso de instancias de alto rendimiento implica gastos significativos (García-Martín et al., 2019). Por ejemplo, ejecutar entrenamientos en instancias GPU puede ser costoso, y el almacenamiento de grandes volúmenes de datos en servicios como Amazon S3 también incrementa los costos. Por lo tanto, las organizaciones deben considerar los trade-offs entre la necesidad de un modelo más preciso y los costos computacionales asociados.

3.4 Análisis y preparación de datos

3.4.1 Importancia de la calidad de datos

La calidad de los datos se refiere a la confiabilidad de los datos en el contexto de su uso. Implica que los datos sean consistentes, completos, precisos, actualizados y relevantes la organización (Jiang et al., 2019). La calidad de los datos es fundamental para la toma de decisiones informadas, ya que afecta directamente la efectividad de las estrategias de negocio y los resultados de análisis. Un conjunto de datos de alta calidad no solo apoya el análisis efectivo, sino que también mejora la confianza en los resultados y las recomendaciones derivadas de los mismos (Redman, 2018).

Los datos incompletos, incorrectos o sesgados pueden tener consecuencias graves en cualquier análisis y toma de decisiones. Por ejemplo, la falta de datos relevantes puede llevar a conclusiones erróneas. Por lo tanto, mantener altos estándares de calidad en los datos es crucial para asegurar decisiones basadas en información precisa y justa.

3.4.2 Exploración de datos

La exploración de datos es un proceso fundamental en el análisis de datos, que busca comprender y resumir las características de un conjunto de datos a través de análisis descriptivos, visualizaciones y la identificación de patrones y tendencias. Los análisis descriptivos incluyen el

cálculo de estadísticas básicas como la media, la mediana, la moda y la desviación estándar, así como la evaluación de distribuciones que ayudan a entender la dispersión y la forma de los datos (Tukey, 1977). La visualización de datos es una herramienta esencial en este proceso, ya que permite representar gráficamente la información de manera que se facilite la comprensión; por ejemplo, los histogramas pueden mostrar la distribución de una variable, mientras que los gráficos de dispersión permiten observar la relación entre dos variables (Wilkinson, 2005). Esta etapa es crítica para la toma de decisiones, ya que proporciona una base sólida sobre la cual construir análisis más profundos y aplicar técnicas de modelado predictivo.

3.4.3 Preparación y transformación de datos

La limpieza de datos es un paso crucial en la preparación de datos que garantiza la calidad y la integridad de la información utilizada para el análisis. Este proceso incluye el manejo de valores nulos, donde se decide si se deben imputar, eliminar o dejar como están, dependiendo de la cantidad y la importancia de los datos faltantes. Un conjunto de datos limpio y sin duplicados no solo mejora la precisión de los análisis, sino que también optimiza el rendimiento de los algoritmos de modelado, permitiendo una interpretación más clara y confiable de los resultados (Karr et al., 2014).

La normalización y estandarización de datos son técnicas utilizadas para transformar las variables a una escala común, lo cual es especialmente importante cuando se trabaja con algoritmos de aprendizaje automático sensibles a las magnitudes de las variables (Han et al., 2011). La normalización implica ajustar los valores a un rango específico, como [0, 1], mientras que la estandarización transforma los datos para que tengan una media de cero y una desviación estándar de uno. Ambas técnicas facilitan la comparación entre diferentes características de los algoritmos de aprendizaje automático, asegurando que todas las variables contribuyan de manera equitativa en el proceso de modelado.

La creación de características implica generar nuevas variables a partir de las existentes. Esto puede incluir la combinación de variables, la extracción de información de fechas o el uso de transformaciones matemáticas (Domingos, 2012). Por otro lado, la selección de variables relevantes es fundamental para reducir la dimensionalidad del conjunto de datos, eliminando características irrelevantes o redundantes que pueden introducir ruido y afectar negativamente el rendimiento del modelo (Guyon & Elisseeff, 2003).

3.5 Modelos de machine learning para regresión

3.5.1 Regresión lineal

La regresión lineal es una técnica estadística utilizada para modelar la relación entre una variable dependiente continua y otras independientes. El modelo se formula como una ecuación lineal, donde la variable dependiente (Y) se expresa como una combinación lineal de las variables independientes (X) más un término de error. Esta técnica permite no solo predecir valores de la variable dependiente, sino también entender la relación entre las variables y la fuerza de esta relación a través de los coeficientes estimados.

Para que los resultados de la regresión lineal sean válidos y confiables, se deben cumplir ciertos supuestos. Primero, se asume que la relación entre las variables independientes y la dependiente es lineal. Segundo, los errores deben ser independientes y estar distribuidos de manera normal con media cero y varianza constante. Tercero, no debe haber multicolinealidad significativa entre las variables independientes, lo que significa que no deben estar altamente correlacionadas entre sí (Muggeo & Adelfio, 2011).

La implementación de un modelo de regresión lineal generalmente involucra varios pasos, incluyendo la preparación de los datos, la selección de variables y el entrenamiento del modelo. Una vez que los datos están limpios y listos, se pueden dividir en conjuntos de entrenamiento y prueba. El entrenamiento del modelo implica el uso de algoritmos para estimar los coeficientes que minimizan la suma de los errores al cuadrado (Hastie et al., 2009). Después de entrenar el modelo, se evalúa su rendimiento utilizando métricas como el error cuadrático medio (MSE) o el coeficiente de determinación, que indican la proporción de la variación en la variable dependiente que se explica por las variables independientes.

3.5.2 Regresión no lineal

La regresión polinómica permite modelar relaciones no lineales entre la variable dependiente y las variables independientes mediante la inclusión de términos polinómicos en el modelo. En este modelo no se trata de ajustar una línea recta, sino que con este enfoque se trata de modelar a una curva incluyendo términos polinómicos. Sin embargo, aunque la regresión polinómica puede mejorar el ajuste a los datos, también conlleva el riesgo de sobreajuste, especialmente con polinomios de alto grado, lo que puede reducir la capacidad de generalización del modelo.

Las técnicas de regularización, como Lasso (Least Absolute Shrinkage and Selection Operator) y Ridge, se utilizan para abordar el problema de sobreajuste en modelos de regresión. Lasso añade una penalización al valor absoluto de los coeficientes de las variables en la función de pérdida, lo que no solo ayuda a reducir la magnitud de los coeficientes, sino que también puede llevar a la eliminación de variables no significativas (Tibshirani, 1996). Por otro lado, Ridge añade una penalización al cuadrado de los coeficientes, lo que tiende a distribuir la carga entre todas las variables en lugar de eliminar algunas por completo. Ambos métodos mejoran la interpretabilidad del modelo y su capacidad de generalización al restringir la complejidad del modelo y prevenir el sobreajuste (Hastie et al., 2009).

Ahora bien, para abordar otro tipo de problemas más complejos de regresión no lineal, se pueden emplear algoritmos avanzados como las máquinas de soporte vectorial (SVM) y las redes neuronales. Las SVM son técnicas de aprendizaje supervisado que encuentran un hiperplano óptimo para separar diferentes clases en los datos, y pueden adaptarse a la regresión no lineal mediante el uso de núcleos (kernels) que transforman los datos a un espacio de mayor dimensión (Cortes & Vapnik, 1995). Por otro lado, las redes neuronales son modelos compuestos por capas de nodos interconectados que pueden aprender representaciones complejas de los datos a través de múltiples capas ocultas, permitiendo la captura de relaciones no lineales profundas (LeCun et al., 2015). Ambos enfoques son poderosos para modelar relaciones no lineales, pero requieren un ajuste cuidadoso de sus hiperparámetros.

3.5.3 Implementación y entrenamiento del modelo

La preparación de datos para el entrenamiento de un modelo es un paso crítico que implica varias etapas, como la limpieza, transformación y división del conjunto de datos. Durante la limpieza, se identifican y manejan valores nulos, duplicados y errores en los datos. Posteriormente, las variables pueden ser transformadas a formatos adecuados, como la normalización o estandarización, para garantizar que todas las características contribuyan de manera equitativa al modelo (Han et al., 2011). Además, se debe dividir el conjunto de datos en grupos de entrenamiento y prueba, lo que permite entrenar el modelo en un subconjunto de los datos y evaluar su rendimiento en un conjunto diferente, evitando así el sobreajuste y asegurando que el modelo sea capaz de generalizar a nuevos datos (Kohavi, 1995).

Por otro lado, el ajuste de hiperparámetros es un proceso esencial en la implementación de modelos de machine learning, ya que estos parámetros influyen en el comportamiento del modelo durante el entrenamiento. Los hiperparámetros pueden incluir el número de capas en una red neuronal, la tasa de aprendizaje, el coeficiente de regularización, entre otros. Para encontrar los valores óptimos de estos parámetros, se utilizan técnicas como la búsqueda en cuadrícula (grid search) y la búsqueda aleatoria (random search), que permiten explorar diferentes combinaciones de hiperparámetros (Bergstra & Bengio, 2012).

Finalmente, la evaluación de modelos durante el entrenamiento es muy importante para monitorear su rendimiento. Se utilizan métricas de evaluación, como la precisión, el recall, la F1-score o el error cuadrático medio (MSE), dependiendo del tipo de problema (clasificación o regresión). La validación cruzada se aplica comúnmente para evaluar el modelo en diferentes subconjuntos de datos, proporcionando una visión más completa de su rendimiento y ayudando a detectar problemas de sobreajuste (Kohavi, 1995).

3.6 Evaluación y selección del modelo

3.6.1 Criterios de selección del modelo

La comparación de rendimiento entre modelos es un paso fundamental en el proceso de selección del modelo, ya que permite determinar cuál de los modelos probados se adapta mejor a los datos y al problema específico. Esta comparación se lleva a cabo evaluando varios modelos utilizando el mismo conjunto de datos de prueba. La visualización de los resultados a través de gráficos comparativos y tablas de rendimiento ayuda a identificar rápidamente el modelo que logra la mejor combinación de precisión y generalización (Hastie et al., 2009).

Paralelamente, la evaluación de modelos basada en métricas de desempeño es esencial para cuantificar la efectividad de cada modelo en función de su capacidad para hacer predicciones precisas. Dependiendo del tipo de problema, se pueden utilizar métricas diferentes: para problemas de regresión, el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación son utilizados, mientras que, para problemas de clasificación, métricas como la precisión, el recall, la F1-score y el área bajo la curva (AUC) son más relevantes (Sokolova & Lapalme, 2009).

En último lugar, al seleccionar un modelo, es crucial considerar no solo su rendimiento, sino también factores prácticos como los tiempos de ejecución y la complejidad del modelo. Modelos más complejos, como redes neuronales profundas, pueden ofrecer un alto rendimiento, pero también requieren tiempos de entrenamiento más largos y mayor poder computacional, lo que puede ser una limitación en entornos con recursos limitados o cuando se necesitan resultados en tiempo real (Chollet, 2018).

3.6.2 Validación cruzada

La validación cruzada es una técnica esencial para evaluar la capacidad de generalización de un modelo al dividir el conjunto de datos en múltiples subconjuntos. Existen varias técnicas de validación cruzada. En la validación cruzada, el conjunto de datos se divide en varios subconjuntos,

y el modelo se entrena con esos subconjuntos. Esta técnica ofrece una estimación robusta del rendimiento del modelo, ya que se aprovecha al máximo el conjunto de datos (Stone, 1974).

3.7 Documentación y comunicación de resultados

3.7.1 Documentación técnica

La documentación técnica es fundamental en el desarrollo de modelos de machine learning, ya que proporciona un registro claro y sistemático de los procesos y decisiones tomadas a lo largo del proyecto. Este registro incluye la justificación de la selección de modelos, las técnicas de preprocesamiento de datos utilizadas y las evaluaciones realizadas durante el entrenamiento (Pérez et al., 2021).

La documentación de los hiperparámetros y configuraciones del modelo es crucial para asegurar que el modelo pueda sea reproducible. Esto incluye detallar los valores específicos asignados a los hiperparámetros. Además, es importante incluir el proceso de ajuste de hiperparámetros, como las técnicas de validación cruzada utilizadas y los criterios de selección para los valores óptimos. Esta información no solo ayuda a replicar los resultados, sino que también proporciona una base para la evaluación de mejoras y ajustes en iteraciones futuras del modelo (Bengio et al., 2015).

A la larga, la documentación de resultados y análisis es esencial para comunicar efectivamente los hallazgos y el rendimiento del modelo a las partes interesadas. Una documentación clara y accesible no solo facilita la toma de decisiones informadas, sino que también contribuye al aprendizaje organizacional y a la mejora continua en proyectos de machine learning.

3.7.2 Informes para la Superintendecia Financiera y MRM

La elaboración de informes para la Superintendencia Financiera requiere un formato estructurado que refleje la información relevante de manera efectiva. Generalmente, los informes deben incluir secciones como una introducción que contextualice el objetivo del informe, un resumen que resalte los hallazgos clave y las conclusiones. Además, es importante incluir tablas,

gráficos y visualizaciones que faciliten la comprensión de los datos presentados (Superintendencia Financiera de Colombia, 2021).

Adicional a esto, la presentación de hallazgos debe tener un enfoque narrativo. También, los hallazgos clave deben destacarse con formato especial, como negritas o cuadros, y los gráficos deben ser etiquetados adecuadamente con títulos y leyendas que expliquen su contenido (Tufte, 2001). La organización lógica de la información es crucial. Al finalizar, se deben incluir recomendaciones prácticas basadas en los hallazgos, lo que proporciona valor agregado al informe.

Finalmente, la Ley 1870 de septiembre de 2017 es la guía para el cumplimiento con regulaciones y requisitos de la Superintendencia Financiera. Esta ley dicta las normas para fortalecer la regulación y supervisión de los conglomerados financieros, garantizando una gestión más eficiente de los riesgos dentro del sector financiero y promoviendo la transparencia en la información que reportan estas entidades. (Congreso de la República, 2017)

4. Metodología

A continuación, se detalla la metodología empleada para la estructuración del modelo de estimación de carga financiera máxima para clientes en Nequi.

4.1 Establecer las métricas de negocio y los requerimientos del área de riesgos de crédito

A fin de tener claridad de las necesidades de negocio con respecto a la estimación de la carga financiera máxima de los clientes y establecer un mínimo producto viable esperado, se detallan las actividades realizadas para llevar a cabo la determinación de las métricas y requerimientos del área de riesgos de crédito con respecto al modelo.

4.1.1 Revisar documentación riesgos de crédito

En esta etapa se realizó una revisión de los documentos que contienen las políticas de riesgo de crédito de Nequi con miras a comprender el negocio, primeramente. Los documentos analizados muestran una variedad de información relacionada con la gestión y evaluación del riesgo crediticio. Los documentos analizados fueron:

- Políticas de riesgo de crédito
- Políticas de originación

En el documento de políticas de riesgo de crédito se establecen las directrices para evaluar y gestionar el riesgo crediticio, garantizando que las decisiones de crédito se tomen de manera prudente y que los riesgos estén controlados. Por otro lado, el documento de políticas de originación contiene los lineamientos para la concesión de nuevos créditos, asegurando que se cumplan los criterios de calidad y se minimicen riesgos en la incorporación de nuevos clientes.

4.1.2 Entender indicadores de mora

Esta actividad implicó analizar y comprender las métricas que reflejan el comportamiento de pago de los clientes y el estado de los créditos otorgados. En esta etapa se revisó el documento "Indicadores riesgo de crédito" en el cual se recopilan los diferentes indicadores que se usan para la gestión de riesgo de crédito en sus etapas de originación, monitore y recuperación. A continuación, se muestran los indicadores analizados:

- Indicador de cartera vencidad (ICV)
- Indicador de mora (IM)

4.1.3 Comprender indicadores de cosechas

Ligada a la actividad anterior, en esta etapa se analizó el indicador de cosechas, el cual se entiende como el conjunto de nuevos créditos o desembolsos colocados u originados en un período de tiempo determinado. Se analizaron las variables que son requeridas para el cálculo de cosechas, las cuales son: fecha de desembolso, valor desembolsado, fecha de corte, altura de vida del crédito, saldo capital vencido no castigado, saldo capital vencido no castigado, saldo capital castigado, número de créditos castigados, número de créditos vencidos.

4.1.4 Realizar reuniones con equipo funcional a cerca de expectativas del modelo

Las reuniones con equipos funcionales fueron espacios cruciales para establecer las expectativas del negocio sobre el modelo. Durante estas sesione, se discutió el impacto esperado en el flujo de preaprobados, los segmentos de clientes prioritarios y otras necesidades funcionales.

4.1.4 Realizar reuniones con equipo técnica a cerca de la viabilidad técnica del proyecto en función de los datos disponibles

En las reuniones con el equipo técnico se analizaron las fuentes de datos disponibles para la construcción del modelo y de la variable de respuesta, se determinaron posibles obstáculos y se

definió un plan de trabajo que contemplaba la inclusión de las políticas de riesgo de crédito en los datos con los cuales se entrenará el modelo.

4.1.5 Revisar el proceso de flujo de preaprobados

La revisión del proceso de flujo de preaprobados se enfocó en comprender cómo se toman las decisiones. Esto incluyó examinar las fuentes de datos que alimentan el flujo de preaprobados, tales como las tablas transaccionales y los datos comprados al buró de crédito, los criterios de filtrado que existen y los puntos donde se establecen reglas duras que dejan a los clientes fuera del flujo de preaprobados.

4.1.6 Definir métrica de negocio para aceptación del modelo con equipo funcional

La definición de la métrica de negocio con equipo funcional constituyó la inclusión de objetivos específicos y de mejora en la capacidad de pago. Se realizaron varias reuniones con los implicados en el flujo de preaprobados y se establecieron las relaciones en cuanto a IND (Ingreso Neto Disponible) y PD (Probabilidad de Default) con CMP (Cuota Máxima Pagable).

4.1.7 Definir métrica de desempeño estadística para aceptación del modelo con equipo técnico

En la etapa de definición de las métricas de desempeño estadístico con el equipo técnico se establecieron los umbrales mínimos aceptables para el modelo. Esto incluyó el análisis de estabilidad poblacional (PSI) y la estabilidad de la variable objetivo (CMP). Con estas métricas se garantizó que el modelo no solo sea estadísticamente robusto, sino también operativamente viable.

4.2 Exploración y análisis de datos históricos comprados a la central de información financiera

4.2.1 Realizar consulta SQL para la extracción de los datos históricos desde la central de información financiera

La consulta para la extracción de los datos históricos desde la información comprada al buró de crédito se realizó en Athena AWS. A través de lenguaje SQL, se estructuró la ETL (Extraction, Transformation, Load) que permitió la observación de un cliente en varios periodos de tiempo con el objetivo de analizar las cuotas pagadas por cada cliente en diferentes fechas de corte y así estructurar la variable de respuesta. Se construyeron queries optimizados que recuperaban la información sociodemográfica de la tabla de perfilación, comportamientos de pago y cuotas pagables.

4.2.2 Cargar los datos preparados desde Athena a SageMaker

La transferencia de datos desde Athena a SageMaker se hizo a través de un script que mediante un comando UNLOAD genero los datos del query extraído en Athena a un bucket en S3. Posteriormente, mediante la URI de S3 y con una función construida en Python para leer datos de formato. parquet, se leyeron los datos en un Notebook de SageMaker.

4.2.3 Análisis exploratorio de datos

En la fase de análisis exploratorio se examinaron la distribución de las cargas financieras, los clientes, la relación con respecto a la variable de respuesta. También se construyó la matriz de correlación y se hizo un ejercicio de clustering para comprender la estabilidad de la variable de respuesta en el tiempo por cada segmento de clientes.

Los gráficos utilizados para esta etapa fueron de percentiles de variable de respuesta, matriz de correlación, boxplots, diagramas de dispersión, entre otras visualizaciones relevantes.

4.2.4 Implementar métricas de calidad de datos

La implementación de métricas de calidad de datos se realizó mediante una función realizada por el desarrollador en la cual se evaluaba la consistencia, integridad y validez de los datos. Esto implicó un análisis de duplicados, nulos y valores atípicos.

4.2.5 Identificar y documentar tendencias

La fase de identificación y documentación de tendencias en los datos históricos se enfocó en el análisis de la variable de respuesta (CMP) en la cual se graficó los percentiles de la mediana de la variable de respuesta por clúster a lo largo del tiempo mediante un gráfico de líneas. Este proceso se enfocó en la comprensión de patrones por segmentos de clientes. Así mismo, se documentaron todos los gráficos realizados en el Notebook de SageMaker.

4.3 Implementación del modelo de machine learning de regresión para la estimación de la carga financiera máxima

4.3.1 Implementar técnicas de selección de variables RFE, backward elimination y métodos embebidos

La implementación de técnicas de selección de variables requirió un enfoque sistemático que combinara diferentes metodologías para asegurar la robustez de la selección de características. Para esta selección de variables se apalancó de la librería Scikit-learn. El proceso inició con la aplicación de un RFE que evaluaba iterativamente la importancia de las variables eliminando las menos relevantes en cada paso. Paralelamente, se implementó un backward elimination, el cual comenzó con todas las variables y se eliminaron sistemáticamente aquellas que no tenían relevancia.

Por otro lado, también se realizó un análisis de varianza explicada mediante PCA, determinando las variables que más peso tenían sobre las variables latentes generadas por la reducción de la dimensionalidad PCA.

4.3.2 Seleccionar las variables más importantes que podrían influir en el rendimiento del modelo

Esta etapa consistió en analizar los resultados arrojados por las diferentes estrategias de selección de variables y a criterio de experto, determinar cuál estrategia apunta más los objetivos y las métricas determinadas tanto por el equipo funcional como por el equipo técnico. Esta actividad, por tanto, implicó la evaluación del poder predictivo de cada variable, su estabilidad a través del tiempo y su interpretabilidad desde una perspectiva de negocio.

4.4 Evaluar el mejor modelo de machine learning en función de las métricas de desempeño

4.4.1 Evaluar diferentes algoritmos de regresión

La fase de evaluación de diferentes modelos siguió una estrategia en la cual se entrenaban diferentes algoritmos bajo las mismas condiciones, y se analizaba mediante gráficos boxplots las métricas estadísticas de desempeño para los diferentes modelos a fin de comparar el comportamiento de los diferentes modelos. Se inició con modelos simples como regresión lineal, avanzando hacia técnicas más complejas como ridge y lasso para manejar multicolinealidad. Adicionalmente, también se analizaron árboles de decisión y random forest para capturar relaciones no lineales.

4.4.2 Seleccionar el modelo que mejor se ajuste a los datos, considerando el MAPE

En esta fase, se analizó cada modelo no solo en función de una sola métrica, sino considerando un balance entre ellas. Para esto, se compararon los resultados obtenidos de cada algoritmo utilizando gráficos de boxplot, que proporcionan una visualización clara de la distribución de las métricas de desempeño para cada modelo al implementar una técnica de

validación cruzada. Esta herramienta permitió identificar rápidamente el modelo con mayor estabilidad y consistencia en su rendimiento.

4.4.3 Ajustar los hiperparámetros del modelo seleccionado

Una vez seleccionado el modelo con el mejor desempeño según las métricas de evaluación (como el error cuadrático medio (MSE), el R-cuadrado (R^2) y el error absoluto medio (MAE)), el siguiente paso fue optimizar los hiperparámetros del modelo para maximizar su rendimiento. Esta fase se hizo utilizando la librería Optuna.

4.5 Seleccionar el mejor modelo en función de las métricas de desempeño de modelos de machine learning y de negocio

En esta fase, se validó la viabilidad del modelo de machine learning desarrollado en el contexto del proyecto, asegurando que estuviera alineado con los objetivos estratégicos, operativos y regulatorios de la organización. El proceso involucra reuniones de trabajo con los equipos funcionales y con el equipo de Model Risk Management (MRM) de Bancolombia para revisar diversos aspectos clave que puedieran afectar la implementación y el uso del modelo en un entorno de producción.

4.6 Documentar los hallazgos encontrados y los hiperparámetros del modelo de machine learning

4.6.1 Crear documento consolidado con los hallazgos encontrados durante EDA

La fase de documentación de los hallazgos encontrados durante el EDA garantiza que los resultados sean comprendidos y puedan ser utilizados para guiar futuras decisiones sobre la selección de modelos y estrategias. Dicha documentación se hizo teniendo a la mano los gráficos generados en etapa posteriores y los Notebooks en los cuales se generaron las visualizaciones.

4.6.1 Crear documento consolidado con los hallazgos encontrados selección de variables e hiperparámetros

En esta sección, se documentó de manera detallada los hallazgos obtenidos durante el proceso de selección de variables y ajuste de hiperparámetros en el desarrollo del modelo de machine learning. Dicha documentación se hizo en Word, apalancándose de las etapas anteriormente realizadas, con los scripts de SQL y los códigos de Python en donde se realizaron la selección de variables y el afinamiento de hiperparámetros.

4.7 Documentación del modelo y los hallazgos encontrados durante la ejecución del proyecto

4.7.1 Gestionar permisos para acceder a la wiki

El proceso de gestión de permisos se realizó a través de un pedido por el software Jira, dicho pedido debe se aprobó por el jefe de área.

4.7.2 Cargar documentos importantes a la wiki

Una vez se tuvo acceso a la wiki, se cargaron las notas agregadas en el JupyterNotebook de manera secuencial, los gráficos de tendencia, los códigos de entrenamiento del modelo, los códigos de selección de características, los códigos de afinamiento de hiperparámetros y finalmente las consideraciones a tener en cuenta en el modelo.

4.7.3 Cargar gráficos importantes del modelo

En esta fase, se cargaron los gráficos clave que se generaron durante el proceso de desarrollo del modelo, para que los equipos puedan visualizar fácilmente los resultados. Estos gráficos incluyen boxplots.

4.7.4 Consolidar información del modelo

En esta fase, se compiló toda la información relevante sobre el modelo en un documento único y organizado. Esto incluyó los detalles sobre los datos utilizados, las características seleccionadas, los hiperparámetros ajustados, las métricas de desempeño, los gráficos generados, consideraciones y limitaciones del modelo.

5. Análisis de resultados

5.1 Establecer las métricas de negocio y los requerimientos del área de riesgos de crédito

5.1.1 Revisar documentación riesgos de crédito

La revisión del documento de políticas de riesgo de crédito dio como resultado que las políticas de riesgo de crédito al interior de Nequi se alinean mediante Circular Básica Contable y Financiera, la cual es una circular emitida por la Superintendencia de Sociedades. En ese sentido, el estudio de crédito al interior de Nequi se realiza en función de dos fuentes de información importantes: central interna y central externa. Por otro lado, la capacidad de pago es medida mediante la siguiente ecuación, y para seguir en el proceso de aprobación debe tener un IND mayor a 55.000 COP.

$$IND = Ingresos - Egresos$$

En cuanto al documento de políticas de originación, se realiza una serie de cálculos en función de los ingresos y egresos de los clientes, con los cuales se da el proceso de estimación del cupo, el cual es diferenciado dependiendo del tipo de crédito al cual puede acceder un cliente basados en su IND. Nequi tiene dos principales líneas de créditos: crédito salvadidas y crédito propulsor. El crédito salvavidas sostiene cupos de entre 100.000 COP y 500.000 COP. Por otro lado, el crédito propulsor ofrece créditos de entre 500.000 COP hasta 10.000.000 COP.

5.1.2 Entender indicadores de mora

Al interior del área de riesgos de créditos se tienen tres principales indicadores: indicador de cartera vencida o calidad de cartera (ICV), indicador de cartera C (ICC+) e indicador de cartera B (ICB+). Por motivos de confidencialidad, la compañía se reserva la metodología bajo la cual estos indicadores son calculados.

5.1.3 Comprender indicadores de cosechas

Las cosechas se entienden como un conjunto de nuevos créditos o desembolsos colocados (a cierta población) en un periodo de tiempo determinado. El análisis de cosechas permite evaluar el rendimiento de grupos de préstamos a lo largo del tiempo. El indicador de cosechas es medido se la siguiente manera:

$$Cosechas = \frac{Saldo\ vencido + Saldo\ castigado}{Valor\ desembolsado}$$

5.1.4 Realizar reuniones con equipo funcional a cerca de expectativas del modelo

Durante las reuniones con el equipo funcional con respecto a las expectativas del modelo, se determinó que la principal necesidad de negocio que se tienen con respecto a la estimación de la carga financiera máxima se debe a que la mayoría de los clientes al interior de la compañía son rechazados por su capacidad de pago. Por esta razón, las métricas de evaluación del modelo será la cantidad de créditos aprobados de manera porcentual con el modelo durante el desarrollo del backtesting, con un mínimo de 40% de aprobación. Dicha métrica será medida de la siguiente manera:

$$\frac{Cantidad\ de\ clientes\ con\ IND > 55.000}{Cantidad\ de\ clientes\ evaluados\ en\ el\ modelo}$$

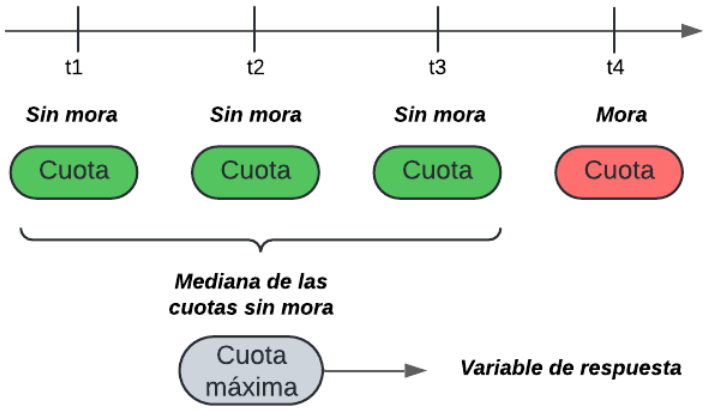
5.1.5 Realizar reuniones con equipo técnica a cerca de la viabilidad técnica del proyecto en función de los datos disponibles

Las diversas reuniones con el equipo de científicos de datos resultaron en la definición de la variable de respuesta. Dicha variable de respuesta está definida en un vector de 13 compras en información al buró, como la mediana de los próximos tres meses previos a la caída en mora en cualquiera de las obligaciones financieras tal como se observa en la siguiente Figura (ver Figura 1).

Para los clientes buenos se define como la máxima cuota observada históricamente. Las métricas para la evaluación del modelo de estimación de carga financiera máxima serán el MAPE y el SMAPE.

Figura 1

Definición de la variable de respuesta



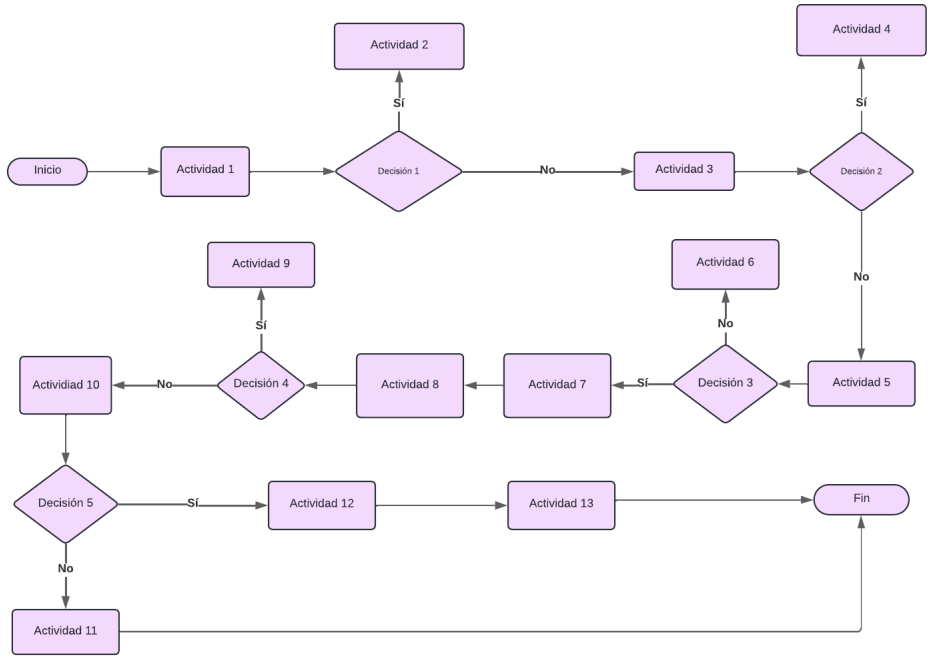
Nota. La figura 1 muestra la lógica bajo la cual fue construida la variable de respuesta

4.1.5 Revisar el proceso de flujo de preaprobados

La revisión del flujo de preaprobados resultó en la creación del diagrama de evaluación en el flujo de preaprobados. El proceso de flujo de preaprobados se detalla en el siguiente diagrama (ver Figura 2), el cual organiza las actividades y decisiones en un formato visual simplificado. Por motivos de confidencialidad, las actividades y decisiones en los nodos del diagrama de proceso se anonimizan.

Figura 2

Diagrama anonimizado del proceso de flujo de preaprobados



Nota. La figura 2 muestra el diagrama del flujo de preaprobados

Dicho proceso inicia con fuentes de datos de los clientes basados información interna y de proveedores externos. Se siguen una serie de actividades y decisiones que resultan en la aprobación o en el rechazo de créditos a los clientes.

5.1.6 Definir métrica de negocio para aceptación del modelo con equipo funcional

La métrica de negocio definida para la aceptación del modelo es la cantidad de clientes con una diferencia entre CMP y la cuota actual observada mayor a 55.000 COP.

5.1.7 Definir métrica de desempeño estadística para aceptación del modelo con equipo técnico

La métrica de desempeño definida para la aceptación del modelo con el equipo técnico es un MAPE < 30% en evaluación y máximo 40% en OOT.

5.2 Exploración y análisis de datos históricos comprados a la central de información financiera

5.2.1 Realizar consulta SQL para la extracción de los datos históricos desde la central de información financiera

Durante el desarrollo del proyecto se realizaron aproximadamente 18 consultas en SQL, a continuación, se muestra la consulta final que dio como resultado la base con la cual se entrena el modelo.

Figura 3

Script SQL extracción de datos

```
UNLOAD (
WITH base_carga AS (
        SELECT *
        FROM fuente_datos_anonimizada1),

ingresos AS (
        SELECT
        num_doc AS num_doc3,
        fecha_cruce,
        mediana_1_7,
        sum_valor_transaccion_nomina,
        ingreso_pic,
        tipo_cotizante_empleador_1,
        ROW_NUMBER () OVER (PARTITION BY num_doc ORDER BY fecha_cruce DESC) AS rn3,
        ingestion_year,
        ingestion_month
FROM fuente_datos_anonimizada2),

ingresos_filtrado AS (
        SELECT *
        FROM ingresos
        WHERE rn3 = 1)

SELECT *
FROM base_carga b
LEFT JOIN ingresos_filtrado c
ON CAST(b.num_doc AS varchar) = CAST (c.num_doc3 AS varchar)
AND CAST(b.f_ingestion_year_month_cleaned AS varchar) < CAST(c.fecha_cruce AS varchar)
```

Nota. La figura 3 contiene el código SQL que genera la tabla con los datos para entrenar el modelo

5.2.2 Cargar los datos preparados desde Athena a SageMaker

En la porción de código detallado en la Figura 3 se observa el uso del comando UNLOAD para arrojar los datos extraídos mediante la consulta a un bucket en S3 en formato. parquet.

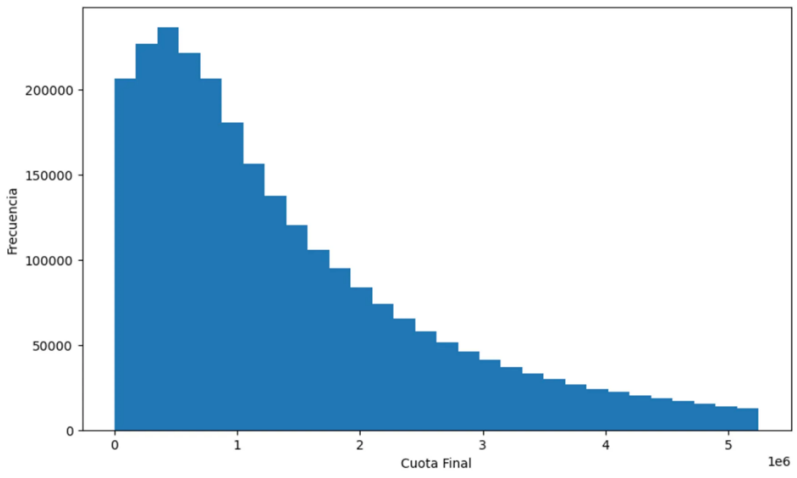
5.2.3 Análisis exploratorio de datos

El análisis exploratorio se realizó sobre una muestra representativa de la base final. En esta etapa se analizaron diferentes variables que a criterio funcional del equipo de riesgo de crédito pueden ser representativas para explicar la carga financiera máxima de los clientes. A continuación de detallará cada variable analizada y una breve interpretación de esta.

En primer lugar, se analizó la variable más representativa en la carga financiera de los clientes que es la cuota pagable mes a mes. En el histograma a continuación (ver Figura 4) se observa que la mayoría de clientes analizados en la muestra tiene cuotas concentradas mayoritariamente hasta 1.000.000 COP, con un pico alto cerca a los 600.000 COP.

Figura 4

Histograma cuotas financieras

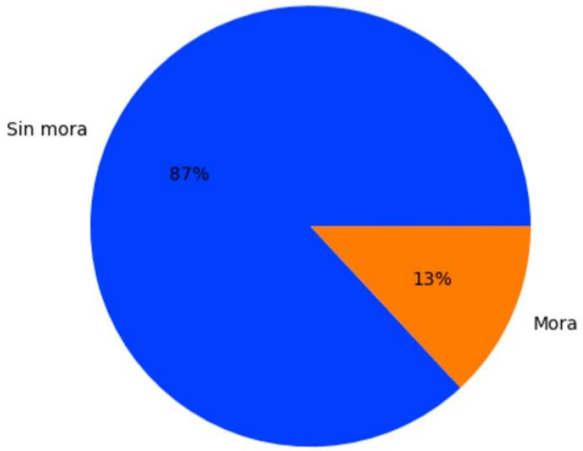


Nota. La figura 4 representa la distribución de la variable de respuesta

Por otro lado, en la Figura 5 se logra observar que la base de clientes tiene un desbalance de clases en cuanto a clientes con mora vs clientes sin mora. La mayoría de clientes no tienen mora en la base de entrenamiento, lo cual es necesario resaltar dado que, aunque la solución no es un algoritmo de clasificación, se debe tener en cuenta este aspecto a la hora de generalizar el modelo para todos los clientes.

Figura 5

Distribución de moras



Nota. La figura 5 contiene el gráfico circular con la distribución de clientes con caídas en mora y sin mora

La distribución de la carga financiera de los clientes es un aspecto fundamental en la construcción del modelo. En la Figura 6 se visualiza que los clientes con experiencia crediticia alta tienden a tener cuotas más altas, seguidos de aquellos que tienen experiencia media. Nótese que la cola de la distribución para clientes con experiencias crediticia baja está más achatada hacia la izquierda, lo cual implica cuotas más bajas para este segmento de clientes. Tiene mucho sentido este fenómeno y da indicios de posibles clústeres en la base de entrenamiento.

Figura 6

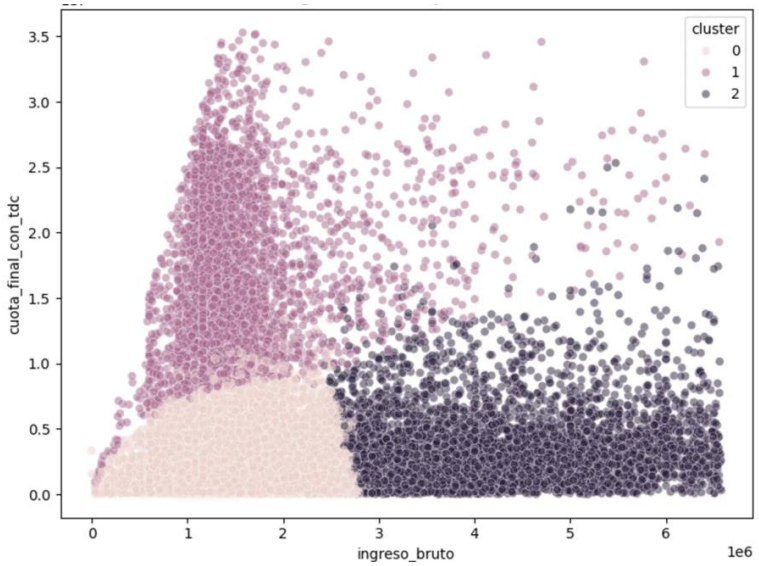
Distribución de carga financiera por experiencia crediticia

Nota. La figura 6 muestra las distribuciones de la carga financiera por experiencia crediticia

Dado que se identificó la posible existencia de clústeres dentro la base de entrenamiento se procedió a implementar un modelo no supervisado K-Means para analizar si efectivamente se hallaban dichos clústeres. En el diagrama de dispersión de ingresos vs cuotas pagables (ver Figura 7) es posible identificar que existen tres segmentos de clientes importantes: clientes con ingresos bajos y cuotas altas (clúster 1), clientes con ingresos y cuotas bajos (clúster 0) y clientes con ingresos altos y cuotas bajas (clúster 2).

Figura 7

Diagrama de dispersión ingresos vs cuotas por clúster



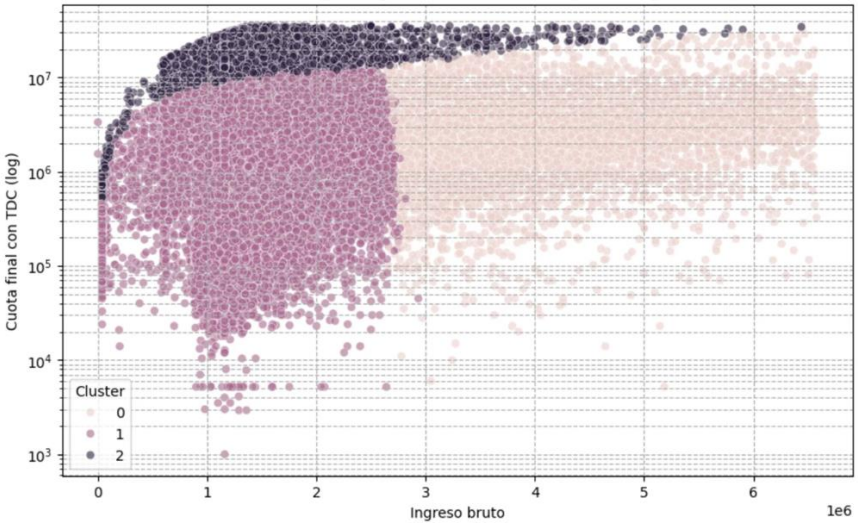
Nota. El diagrama de dispersión ingreso vs cuota por clúster se observa en la figura 7

Si bien el diagrama anterior es diciente con respecto a la pertenencia de clúster de los clientes, puede estar sesgado por valores atípicos. Nótese que el eje y tiene una multiplicación por 10^7. Esto podría dar la errónea conclusión de que la mayoría de los clientes tienen cuotas extremadamente altas.

Para mitigar esto, se escalaron los datos con una transformación logarítmica. En la Figura 8 se lo logra identificar la distribución verdadera de los ingresos vs la cuota de los clientes. En el eje y están las cuotas en base de notación científica. Así, se logra identificar que hay muchos clientes en el clúster 1, y la mayoría de dichos clientes tienen cuotas inferiores al 1.000.000 COP. Por otro lado, la mayoría de los clientes de clúster 0 tienen cuotas superiores a 1.000.000 COP e inferiores a los 10.000.000 COP. Finalmente, los clientes de clúster 2 (clientes morosos) tienen cuotas superiores a los 10.000.000 COP.

Figura 8

Diagrama de dispersión ingresos vs cuotas escalada logarítmica

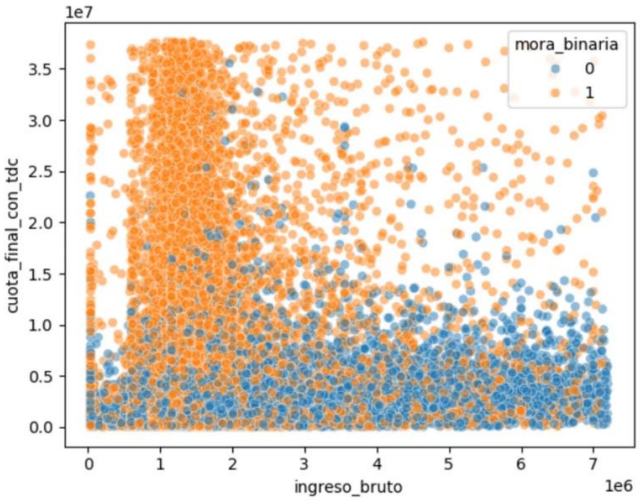


Nota. El diagrama de dispersión ingreso vs cuota en escala logarítmica por clúster se observa en la figura 8

Posteriormente, se realizó un análisis de clústeres discriminando por caídas o no caída en mora para confirmar si dichos clústeres son significativos en cuanto a la mora. En el diagrama de dispersión ingresos vs cuota por mora (ver Figura 8), se identifica que, en clientes con ingresos bajos y cuotas altas, la mora es mayor, mientras que los clientes con ingresos altos y cuotas bajas tienen menos niveles de morosidad.

Figura 9

Diagrama de dispersión ingresos vs cuotas por mora



Nota. En la figura 9 se muestra la relación ingreso vs cuota por nivel de mora

5.2.4 Implementar métricas de calidad de datos

Figura 10

Función para evaluación calidad de datos

```
def check_df(dataframe, head=10):
        display(Markdown('**Dimensiones base general**'))
        display(dataframe.shape)

        display(Markdown('**Número de duplicados**'))
        display(dataframe.duplicated().sum())

        display(print(" \n "))

        display(Markdown('**Tipos**'))
        display(dataframe.dtypes)

        display(Markdown('**Primeros Registros**'))
        display(dataframe.head(head))

        display(Markdown('**Nulos**'))
        display(dataframe.isnull().sum())

        display(Markdown('**Percentiles**'))
        display(dataframe.describe([0, 0.05, 0.50, 0.75, 0.95, 0.99, 1]).T)
```

Nota. En la figura 10 se muestra el código de evaluación de calidad de datos en Python

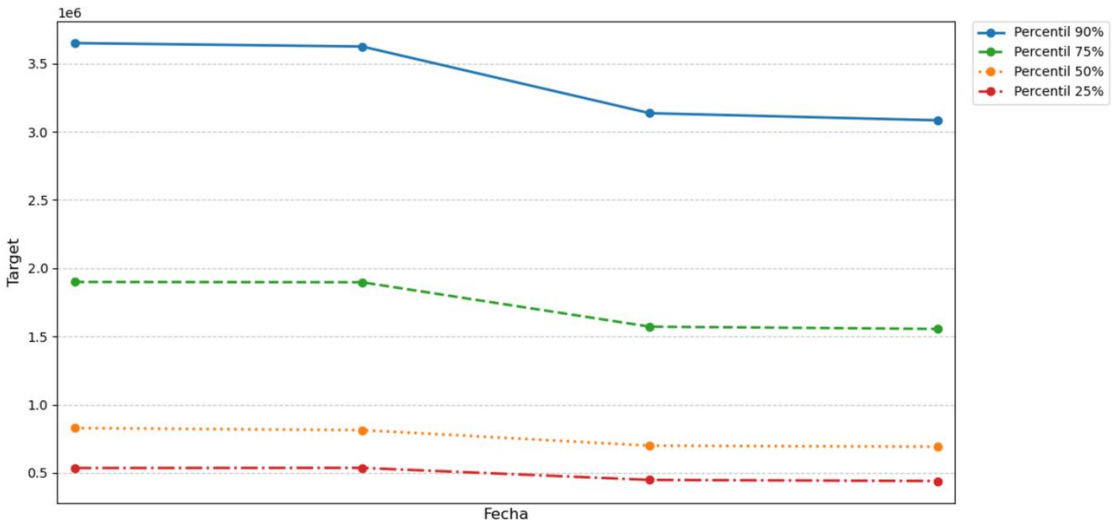
A continuación, se evidencia la función diseñada para analizar calidad de datos. Esta función permite evaluar: dimensiones, duplicados, tipos de datos, nulidad y análisis de diferentes percentiles.

5.2.5 Identificar y documentar tendencias

Con el objetivo de identificar diferencias entre los clústeres construidos mediante el algoritmo de K – Means, se analizaron los percentiles de la variable de respuesta por clúster. En las ilustraciones 11, 12 y 13 se logra identificar la estabilidad de la variable de respuesta para los diferentes clústeres. Nótese que para el clúster 1 se nota mayor estabilidad que para los otros dos clústeres. Es de recalcar que el clúster 1 es el segmento de clientes que tienen cuotas altas, ingresos bajos y niveles de morosidad altos. Esto tiene sentido puesto que, si dichos clientes han alcanzado niveles de endeudamiento excesivamente altos, el sector bancario no desembolsará créditos a favor de ellos, lo cual explica la estabilidad de la variable de respuesta.

Figura 11

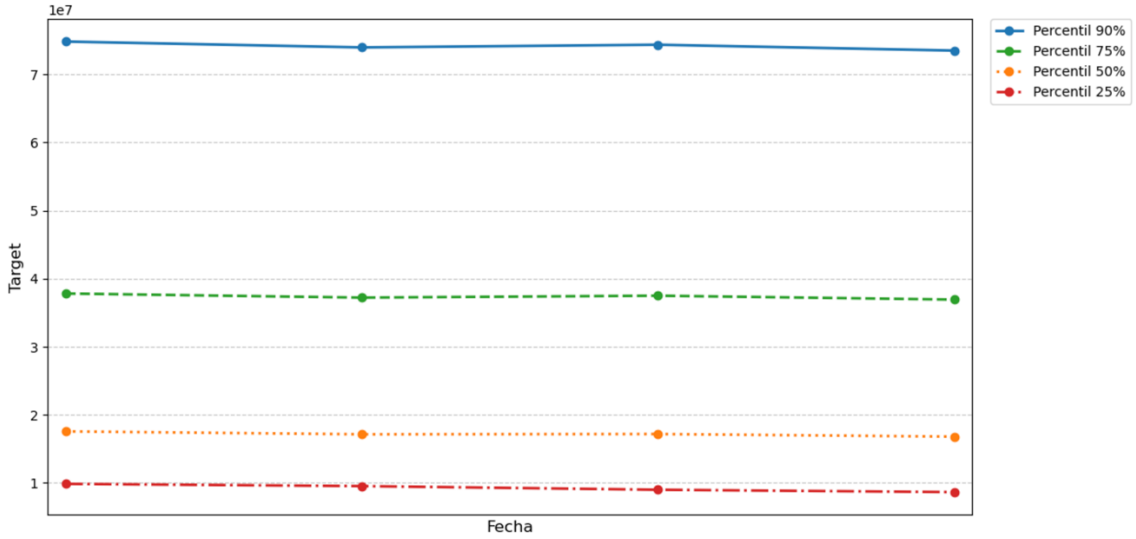
Tendencia clúster 0 en variable de respuesta



Nota. En la figura 11 se muestra el gráfico por percentiles de la variable de respuesta en cluster 0

Figura 12

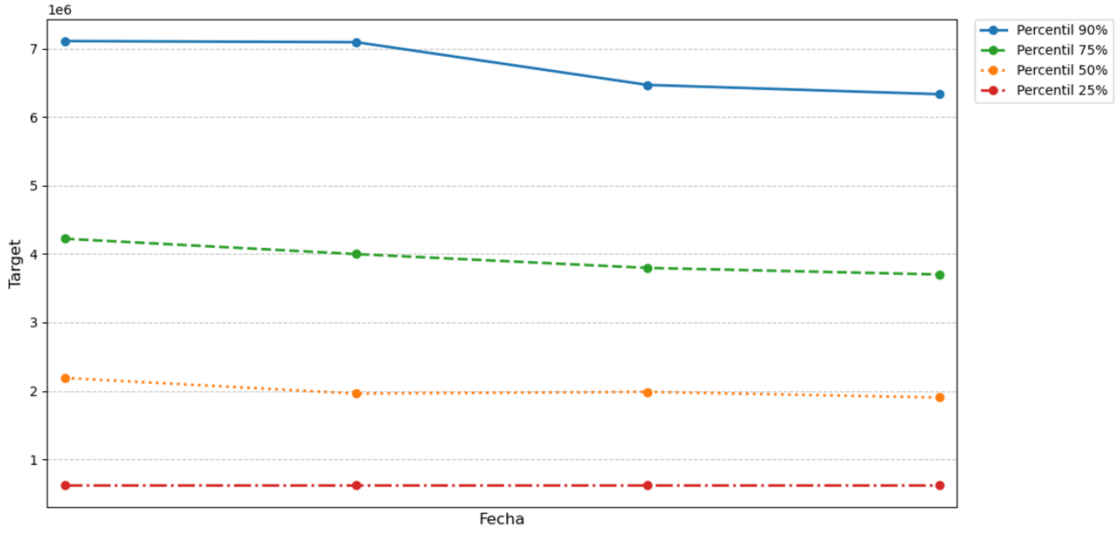
Tendencia clúster 1 en variable de respuesta



Nota. En la figura 12 se muestra el gráfico por percentiles de la variable de respuesta en clúster 1

Figura 13

Tendencia clúster 2 en variable de respuesta



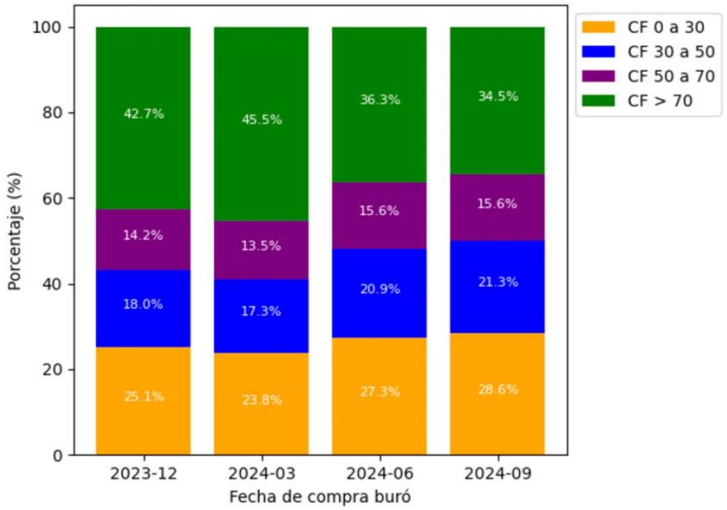
Nota. En la figura 13 se muestra el gráfico por percentiles de la variable de respuesta en clúster 2

Finalmente, se analizó la distribución de la carga financiera de los clientes por fecha de compra en el buró de crédito. En la Figura 14 se logra identificar que la mayoría de los clientes tienen cargas financieras mayores al 70%. Sin embargo, hay una tendencia de migración de

cargas financieras altas hacia cargas financiera más bajas, lo cual representa una oportunidad de originación de dichos clientes que están regazando sus cargas financieras.

Figura 14

Distribución de carga financiera por fecha de compra



Nota. En la figura 14 se observa la distribución de cargas financieras por fecha de compra en buró

5.3 Implementación del modelo de machine learning de regresión para la estimación de la carga financiera máxima

5.3.1 Implementar técnicas de selección de variables

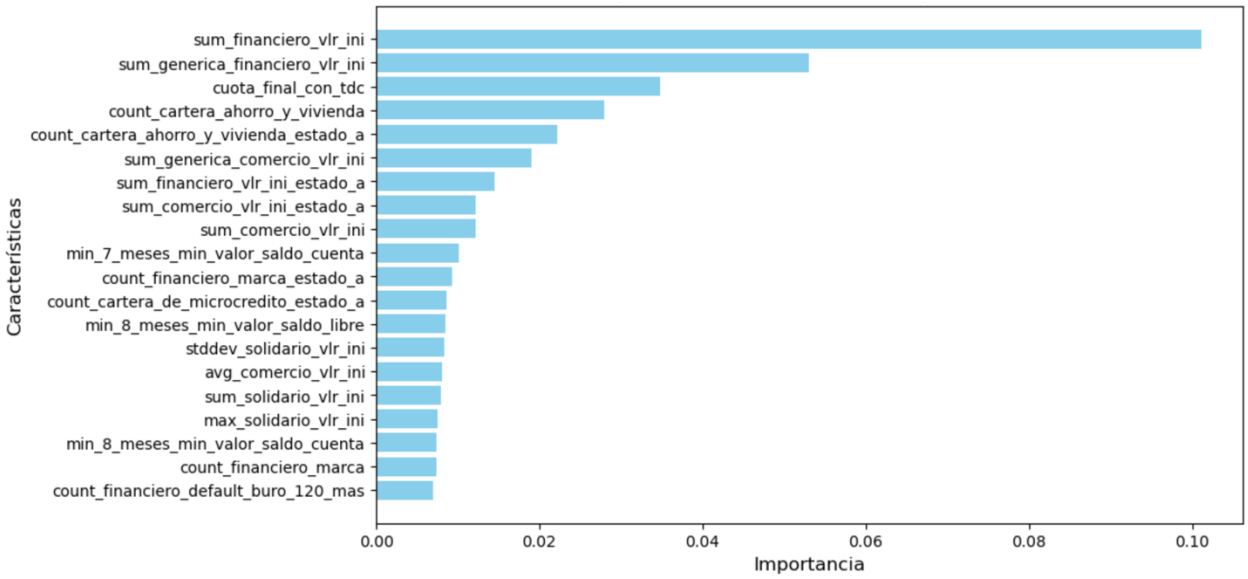
La implementación de técnicas de selección de variables se realizó mediante una variable aleatoria. Se generó una variable aleatoria y se añadió al dataset de entrenamiento, prueba y OOT (Out of Time). Posteriormente se entrenó un modelo LightGBM. Las variables que caen por debajo del nivel de importancia de la variable aleatoria salieron del modelo.

5.3.2 Seleccionar las variables más importantes que podrían influir en el rendimiento del modelo

Las variables más importantes arrojadas por el modelo se observan en la Figura 15. Las variables más importantes se alinean con el sentido del problema. En este caso, las variables observadas hacen referencia al módulo de variables del buró, en el cual reposa la información crediticia de los usuarios en el sector real servicio, real comercio, hipotecario y financiero.

Figura 15

Variables seleccionadas



Nota. En la figura 15 se muestran las 20 variables más relevantes para el modelo

5.4 Evaluar el mejor modelo de machine learning en función de las métricas de desempeño

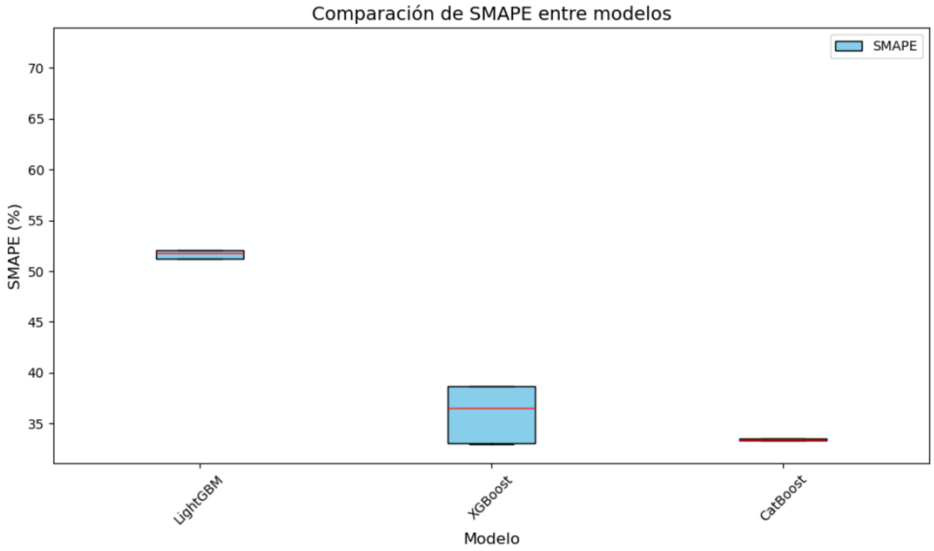
5.4.1 Evaluar diferentes algoritmos de regresión

Se evaluaron tres algoritmos de regresión: LightGBM, CatBoost y XGBoost mediante una función personalizada que iteraba en cada algoritmo, optimizando sus hiperparámetros en Optuna

en función del MAPE. En la figura 16 se observa el rendimiento de cada modelo en el MAPE en 5 iteraciones realizadas.

Figura 16

Comparación entre modelos



Nota. En la figura 16 se observa el rendimiento de los modelos en términos del SMAPE

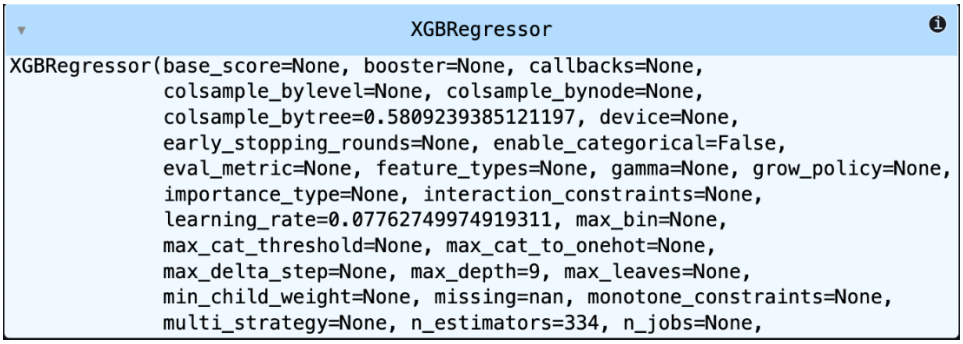
En la figura 16 se logra observar que XGBoost y CatBoost tienen mejores desempeños para el conjunto de datos de validación con un SMAPE promedio 33.22% y 33.31%, respectivamente.

4.4.2 Seleccionar el modelo que mejor se ajuste a los datos, considerando el MAPE

Teniendo en cuenta el desempeño en función del MAPE de los diferentes modelos, se selecciona un modelo XGBoost como el mejor modelo que se ajusta a los datos. En la figura 17 se observa el modelo XGBoost con sus hiperparámetros.

Figura 17

Hiperparámetros de XGBoost Regressor



Nota. En la figura 17 se observan los hiperparámetros del mejor modelo seleccionado

4.4.3 Ajustar los hiperparámetros del modelo seleccionado

Finalmente, se realizaron 5 iteraciones utilizando Optuna para minimizar el MAPE utilizando el modelo XGBooost. Dichas iteraciones quedan consignadas en la tabla 1. Nótese que la iteración 2 fue la que tuvo un mejor rendimiento con un MAPE de 33.316, lo cual implica que los hiperparámetros ajustados para este XGBoost en dicha iteración, son los que generalizan de manera adecuada los datos para predecir la variable de respuesta.

El hiperparámetro `max_depth` hace referencia a la profundidad máxima de cada árbol. Un valor de 11 indica que permite árboles relativamente profundos, lo que significa que el modelo puede capturar relaciones complejas en los datos. Para equilibrar el sobreajuste que la profundidad de los árboles pueda generar, el hiperparámetro `boosting_type` se ajusta en dart, es una variante del boosting que aplica técnicas de dropout similares a las de las redes neuronales. El hiperparámetro `num_leaves` es moderado y sugiere un balance con `max_depth`. Por otro lado, el hiperparámetro `learning_rate` en 0.0937 implica una tasa de aprendizaje lenta del modelo, lo que generalmente lleva a una mejor generalización. El hiperparámetro `feature_fraction` indica que en cada iteración se utiliza cerca del 69% de las características para construir árboles, lo cual introduce aleatoriedad al modelo y ayuda a prevenir el sobreajuste. Por último, el hiperparámetro `bagging_fraction` hiperparámetro ajustado en 0.57, implica que en cada iteración se entrena con aproximadamente el 57% de los datos, lo que también ayuda a prevenir el sobreajuste. Nótese que

se realizaron muchos ajustes para prevenir el sobreajuste, esto era necesario dado que el conjunto de datos de entrenamiento quedó con la gran masa de datos y en OOT se destinaron únicamente el 7% de la población total.

Esta combinación de hiperparámetros sugiere un modelo que busca un balance entre capacidad predictiva y generalización, con varios mecanismos para controlar el sobreajuste (DART, tasas de muestreo moderadas y learning rate bajo).

Tabla 1

Iteraciones XGB Regressor

Iteración	MAPE	Hiperparámetros
Iteración 1	33.952	{'max_depth': 11, 'boosting_type': 'dart', 'num_leaves': 26, 'learning_rate': 0.0938, 'feature_fraction': 0.6817, 'bagging_fraction': 0.6080, ...}
Iteración 2	33.316	{'max_depth': 11, 'boosting_type': 'dart', 'num_leaves': 26, 'learning_rate': 0.0937, 'feature_fraction': 0.6898, 'bagging_fraction': 0.5731, ...}
Iteración 3	34.362	{'max_depth': 11, 'boosting_type': 'dart', 'num_leaves': 26, 'learning_rate': 0.0849, 'feature_fraction': 0.6971, 'bagging_fraction': 0.6026, ...}
Iteración 5	33.407	{'max_depth': 11, 'boosting_type': 'dart', 'num_leaves': 26, 'learning_rate': 0.0151, 'feature_fraction': 0.6964, 'bagging_fraction': 0.6032, ...}
Iteración 5	51.746	{'max_depth': 11, 'boosting_type': 'dart', 'num_leaves': 25, 'learning_rate': 0.0907, 'feature_fraction': 0.6846, 'bagging_fraction': 0.5714, ...}

Nota. En la tabla 1 se observan las iteraciones con hiperparámetros del mejor modelo seleccionado

5.5 Seleccionar el mejor modelo en función de las métricas de desempeño de modelos de machine learning y de negocio

El modelo seleccionado fue un XGBoost Regressor dada la minimización del MAPE. Adicional a ello, dado que desde el negocio se espera un modelo que esté preciso en rango en cuanto a la predicción final, se selecciona el modelo que minimice el MAPE.

5.6 Documentar los hallazgos encontrados y los hiperparámetros del modelo de machine learning

La documentación de los hallazgos encontrados durante el análisis exploratorio y calibración del modelo se encuentran documentados en los repositorios de la compañía, y por motivos de confidencialidad, no se anexan en el presente proyecto

5.7 Documentación del modelo y los hallazgos encontrados durante la ejecución del proyecto

Este objetivo no se logró debido a que la búsqueda y definición de la variable de respuesta tomó más tiempo del previsto. Esto ocurrió porque la variable de respuesta juega un papel central en el desarrollo del modelo, ya que define el objetivo del análisis. Durante el proyecto, se identificaron retos técnicos y conceptuales que requirieron ajustes y validaciones adicionales, lo que extendió significativamente el tiempo necesario para su definición. Dado que la búsqueda y definición de la variable de respuesta tuvo prioridad debido a su importancia estratégica, la documentación del modelo y los hallazgos quedó pospuesta, siendo esta una actividad planificada para las etapas finales del proyecto. Sin embargo, al consumir más tiempo en las tareas críticas, no fue posible destinar los recursos necesarios para este objetivo dentro del tiempo establecido

6. Conclusiones y recomendaciones

- La estimación de la cuota financiera máxima de los clientes juega un papel fundamente en el proceso de flujo de preaprobados al tocar directamente el IND de los clientes.
- La creación de clústeres por carga financiera máxima creados en la etapa de exploración es una solución alternativa al modelo de aprendizaje supervisado.
- El machine learning ofrece múltiples modelos que por medio de diferentes caminos pueden dar solución al problema de cuota financiera máxima, tales como LightGBM, CatBoost, LSTM.
- La evaluación del modelo de *machine learning* en función de métricas de desempeño tales como el MAPE permitió identificar las variables más relevantes para la predicción de la cuota financiera máxima.
- La selección del mejor modelo está atado al módulo de información proveniente de las variables. Se observó que el módulo de información del buró es más explicativo para la variable de respuesta.
- El ajuste de hiperparámetros es una etapa fundamental en la calibración del modelo, y más específicamente cuando se particionan los datos por entrenamiento, testeo y OOT.
- La documentación del modelo requiere un profundo conocimiento de las guías de validación dictadas por MRM para la aprobación de modelos.

Referencias

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management Disrupting Finance: FinTech and Strategy in the 21st Century (pp. 33-50). Palgrave Pivot.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). Deep learning. Nature.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development.

Chollet, F. (2021). Deep learning with Python (2nd ed.). Manning Publications.

Congreso de la República. (2017). Ley 1870 de 2017. Por la cual se dictan normas para fortalecer la regulación y supervisión de los conglomerados financieros y se dictan otras disposiciones.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM.

Feldman, D., & Schmidt, S. (2016). The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. Oxford University Press.

García, M., & Morales, J. (2021). Políticas crediticias y condiciones económicas: Un análisis de su impacto en las cuotas de tarjetas de crédito. Revista de Economía Financiera.

García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing, 134(1), 75-88. https://doi.org/10.1016/j.jpdc.2019.07.007

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

He, Y., Zhang, S., & Li, J. (2023). Machine learning in finance: Recent developments and future directions. Journal of Finance and Data Science.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). Springer.

Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2020). SVM and SVM ensembles in breast cancer prediction. PloS one.

Huang, Y., & Tang, L. (2019). Credit risk assessment and tarjeta de crédito quota estimation: A machine learning approach. Expert Systems with Applications.

Jagtiani, J., & Lemieux, C. (2018). Do fintech lenders penetrate areas that are underserved by traditional banks? Journal of Economics and Business, 100, 43-54.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Jiang, L., Li, C., Wang, S., & Zhang, L. (2019). Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence.

Karr, A. F., Sanil, A. P., & Banks, D. L. (2014). Data quality: A statistical perspective. Statistical Methodology.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2018). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance.

Kim, J., & Yoon, H. (2020). Predicting credit card payment behavior: A machine learning approach. Expert Systems with Applications.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lee, T., & Chan, K. (2020). Understanding credit card spending patterns: A machine learning approach. Journal of Financial Services Research.

López de Prado, M. (2020). Machine learning for asset managers. Cambridge University Press.
Mendoza, R., & Pérez, J. (2018). Efectos de las tasas de interés en el comportamiento de pago de tarjetas de crédito. Estudios Económicos.

Minh, D. L., Wang, H., Li, Y., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Muggeo, V. M., & Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. Biometrika.

Seber, G. A. F., & Lee, A. J. (2003). Linear regression analysis (2nd ed.). John Wiley & Sons, Inc.
O'Dwyer, R., & Vallor, S. (2023). Ethical challenges in AI-driven financial services. Journal of Business Ethics.

Pérez, M., García, J., & López, R. (2021). Best practices in machine learning documentation: A systematic review. Journal of Systems and Software.

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Redman, T. C. (2018). If your data is bad, your machine learning tools are useless. Harvard Business Review.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological).

Superintendencia Financiera de Colombia. (2021). Circular Externa 029 de 2021. Instrucciones relativas a la gestión de riesgos de los establecimientos de crédito.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological).

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.
Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security.

Wilkinson, L. (2005). The grammar of graphics (2nd ed.). Springer.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research.