Predicción de Riesgo de Mora en Créditos Usando Machine Learning

Un Enfoque Basado en Segmentación y Modelos Predictivos

Henny Rocío Carrillo

Juan Sebastián Ballen Villalba

Michael Stiven Escobar Rodríguez

Wilber Alexander Rodríguez Castro

Universidad Ean

Facultad de ingeniería - Especialización en Machine Learning - Seminario Investigación

Mauricio Bolívar Rodríguez

Bogotá, Colombia

16/03/2025

Tabla de contenido

Resumen 5
Problema de investigación 6
Antecedentes de problema 6
Descripción del Problema 8
Causas y Síntomas del Problema 10
Falta de segmentación efectiva 10
Ausencia de análisis de factores de riesgo 10
Deficiencias en la gestión de cobranza 10
Diagnóstico de la Situación 12
Propuesta de Solución 13
Pregunta de investigación. 15
Objetivos 16
Objetivo general 16
Objetivos específicos. 16
Justificación 18
Conveniencia 19
Relevancia social 19
Implicaciones prácticas 20
Valor Teórico 20
Utilidad metodológica 20
Marco Teórico 22
Conducta 22
Capacidad de pago histórica 22
Capacidad de endeudamiento 23
Condiciones macroeconómicas 23
Capacidad de pago proyectada 23
Machine Learning aplicado a finanzas 25
Regresión logística 28
Random Forest 30
XGBoost (Extreme Gradient Boosting). 32
Redes neuronales 32

Redes Neuroales LSTM (Long Short-Term Memory) 33
Análisis exploratorio de datos (EDA) 33
Análisis de Componentes Principales (PCA) 34
Marco institucional 35
Proceso o tema de estudio 38
Metodología 39
Enfoque de la investigación 39
Diseño de la investigación 39
Diseño no experimental, correlacional, explicativo – aplicado: 39
Definición de Variables 40
Población y Muestra 44
Selección de métodos o instrumentos para recolección de información 44
Técnicas de análisis de datos 49
Instrumentos de recolección y origen de datos 50
Justificación de las técnicas 56
Justificación técnica de las técnicas seleccionadas 56
Justificación técnica de las técnicas no implementadas 57
Análisis y discusión de los resultados 59
Resultados 59
Exploración inicial 59
Análisis de calidad de datos 59
Distribución de variables numéricas 59
Distribución de variables categóricas 61
Análisis bivariado: mora por categorías 62
Modelos predictivos 65
Arbol de clasificación XGBoost multiclase 66
Resultados de la matriz de confusión 66
Interpretación estratégica de los resultados 68
Red neuronal LSTM binaria 68
Interpretación estratégica 71
Red neuronal LSTM multiclase 71
Interpretación estratégica 74
Análisis Actual Power BI Segmentación de Cartera de clientes 74

Recomendaciones de tipo estratégico....76
Patrones y Tendencias Claves....79
Recomendaciones Estratégicas....79
Patrones Clave por Nivel Educativo....83
Ajustes de Política....85
Análisis de Concentración de Cartera Vencida por Puntaje ACIERTA....85
Propuesta y estrategias para la prevención de la mora y recuperación de cartera....87
Prevención y alerta temprana....87
Segmentación de clientes según riesgo y comportamiento....87
Diseñar estrategias diferenciadas por grupo etario:....88
Segmentar por tipo de cliente....88
Ajustar política según nivel educativo....88
Gestión de cobranza con enfoque predictivo....88
Ajustes en políticas de crédito y scoring....89
Monitoreo, aprendizaje y mejora continua....90
Conclusiones y recomendaciones....91
Anexos de Entrega....93
Lista de referencias....95

Resumen

El aumento en la cartera vencida de Confirmeza S.A.S. pone en riesgo su estabilidad financiera y revela la necesidad de implementar herramientas analíticas avanzadas que fortalezcan la gestión del riesgo crediticio. Este estudio propone un enfoque predictivo que combina técnicas de análisis exploratorio de datos (EDA) y visualización mediante Power BI, junto con la aplicación de modelos de Machine Learning, específicamente XGBoost y redes neuronales LSTM. El objetivo es anticipar el comportamiento de pago de los clientes y clasificar su estado crediticio futuro. Los modelos desarrollados permiten segmentar la cartera, priorizar acciones de cobranza y diseñar estrategias preventivas más eficaces, contribuyendo a una toma de decisiones más informada y oportuna dentro de la organización.

Palabras clave: riesgo crediticio, morosidad, Machine Learning, modelos predictivos, segmentación de clientes, redes neuronales, XGBoost, recuperación de cartera.

Problema de investigación

Antecedentes de problema

En los últimos años, el deterioro de la cartera de crédito en las instituciones financieras ha experimentado un crecimiento notable, lo que ha afectado la estabilidad y rentabilidad de las mismas. En el caso específico de la empresa Confirmeza, este deterioro ha sido más evidente en el segmento de la cartera con mora superior a 90 días. Para abordar este fenómeno, de acuerdo con León y Espinoza (2023), es crucial analizar las causas subyacentes que han generado el problema de la morosidad y el aumento en la cartera vencida, las cuales se relacionan principalmente con factores económicos, sociales y laborales que afectan el comportamiento de pago de los clientes.

Por una parte, nos encontramos con las condiciones económicas de los clientes, una de las principales causas del aumento en la cartera vencida. Muchos de ellos enfrentan dificultades económicas derivadas del desempleo o la precarización laboral, lo que impacta directamente en su capacidad para cumplir con los pagos. En sectores de bajos ingresos o en zonas con altas tasas de desempleo, los clientes tienen un acceso limitado a empleos formales, lo que genera inestabilidad en sus ingresos y dificulta la regularidad en los pagos de sus obligaciones financieras. La volatilidad económica también contribuye a la incertidumbre financiera, lo que lleva a que los clientes prioricen otros gastos sobre sus compromisos crediticios.

Otro factor relevante es la existencia de malos hábitos de pago por parte de los clientes, que se ven reflejados en la tendencia creciente de la mora. Algunos clientes, aunque tengan capacidad económica, no mantienen una disciplina financiera adecuada, lo que provoca que las cuotas de sus préstamos no se paguen a tiempo. Esta falta de responsabilidad puede estar

relacionada con una cultura de consumo basada en el uso excesivo de crédito sin un manejo adecuado, o con la falta de educación financiera para gestionar sus obligaciones.

Siendo lo anterior parte de la problemática también nos encontramos con la informalidad laboral en el que, haciendo hincapié, contribuye de manera significativa a la morosidad en la cartera de crédito. En contextos donde una gran parte de la población trabaja en la informalidad, los ingresos de los prestatarios son irregulares o no están formalmente registrados, lo que complica la verificación de su capacidad de pago y genera un alto riesgo para las entidades financieras. La informalidad implica también una falta de acceso a los beneficios sociales y de seguridad laboral, lo que, a su vez, aumenta la vulnerabilidad económica de los clientes.(Crowe Colombia, s.f.)

Los factores sociales también juegan un papel importante en el deterioro de la cartera. En algunos casos, las dificultades en la educación financiera y la falta de asesoría sobre los impactos del sobreendeudamiento generan una mala toma de decisiones por parte de los clientes. Además, el acceso fácil al crédito en ciertos sectores puede incentivar una cultura de endeudamiento excesivo, sin una adecuada evaluación del impacto a largo plazo de esos compromisos financieros.

En resumen, el aumento en la cartera vencida de Confirmeza no es únicamente el resultado de una mala gestión interna de los riesgos crediticios, sino que se ve influenciado por un conjunto de factores socioeconómicos y comportamentales que afectan la capacidad de pago de los clientes. Estos incluyen la inestabilidad económica, los malos hábitos de pago, la informalidad laboral y los déficits en la educación financiera. Por lo tanto, es fundamental que cualquier estrategia de recuperación de cartera no solo considere el perfil de los clientes, sino también las condiciones externas que inciden en su comportamiento de pago, lo que permitirá desarrollar enfoques más efectivos para mitigar el riesgo de mora.

En el ámbito financiero, la gestión del riesgo crediticio es un pilar fundamental para garantizar la estabilidad y sostenibilidad de las instituciones que ofrecen servicios de financiamiento. En este sentido, el presente proyecto de investigación se enfoca en analizar los factores que influyen en el deterioro de la cartera de crédito de la empresa Confirmeza, con el objetivo de desarrollar un modelo predictivo basado en técnicas de Machine Learning. Este modelo permitirá identificar a aquellos clientes con mayor probabilidad de caer en mora superior a 120 días, brindando herramientas para la toma de decisiones preventivas y la optimización de estrategias de recuperación.

Mediante comunicación interna el día 5 de enero de 2025, se compartió un documento que identificaba el desempeño de la cartera entre los años 2023 y 2024, respectivamente, en concordancia con las funciones asociadas al área de riesgo. Este informe buscaba resaltar las principales variables de comportamiento y distribución socio-demográfica según el perfil de otorgamiento para los clientes durante este período (N. Rincón, comunicación personal, 5 de enero de 2025).

El área de riesgo financiero de Confirmeza ha observado un deterioro significativo en la calidad de su cartera en los últimos dos años. Los datos indican que:

La concentración de cartera con mora de 30 a 90 días sigue una tendencia estable, pero la cartera con mora superior a 90 días ha mostrado un crecimiento sostenido.

Dado este panorama, es fundamental analizar el comportamiento de pago y las características sociodemográficas de los clientes en mora con el objetivo de implementar estrategias preventivas que mitiguen el riesgo de castigo y mejoren la eficiencia en la recuperación de cartera.

Descripción del Problema

Confirmeza S.A.S., una empresa especializada en el ámbito de la financiación de vehículos nuevos, usados y pólizas, enfrenta un creciente desafío en la administración de sus créditos debido al aumento sostenido en la cantidad de clientes que entran en mora. Este problema se refleja en la evolución de la cartera vencida mayor a 30 días, la cual ha alcanzado los $10.730 millones, con un aumento del 30.62% interanual. Este incremento evidencia un deterioro progresivo de la cartera, afectando la estabilidad financiera y la capacidad operativa de la empresa.

Además, el segmento de cartera vencida con más de 90 días de mora ha crecido de manera significativa, pasando de representar el 3.27% del total en 2024 al 6.75% en 2025, lo que representa un crecimiento del 106.18% en términos relativos. Este incremento sugiere que un número importante de clientes no solo está ingresando a estados de mora temprana, sino que permanece en esa condición sin una estrategia efectiva de recuperación. A pesar de los esfuerzos en la gestión de cobranza. En comparación con enero del año 2023 y enero del 2025 se ha observado un incremento de 3 puntos porcentuales en la cartera vencida, pasando del 8,19 % al 11,17 %. Este aumento impacta negativamente el flujo de efectivo y reduce la capacidad de la empresa para otorgar nuevos créditos, lo que podría afectar su crecimiento y sostenibilidad financiera. En este contexto, el Código Civil Colombiano establece el marco legal para la recuperación de los créditos a través de la acción ejecutiva, la cual se activa cuando el acreedor no recibe el pago de su crédito. En este sentido, el artículo 516 del Código de Procedimiento Civil establece que: "La acción ejecutiva procede para la obtención de la obligación que se exija en virtud de título ejecutivo, sea este de naturaleza judicial o extrajudicial." (Código Civil Colombiano, Artículo 516).

Este fenómeno plantea preguntas fundamentales sobre qué factores llevan a los clientes a incurrir en mora y cómo se pueden anticipar estos eventos para evitar el deterioro de la cartera. En la actualidad, Confirmeza S.A.S. no cuenta con herramientas de segmentación

avanzadas ni con modelos predictivos, que permitan identificar a tiempo a los clientes con alto riesgo de incumplimiento. Como resultado, la empresa adopta una estrategia reactiva en la gestión de cartera, en lugar de implementar mecanismos preventivos que reduzcan la entrada de nuevos clientes en mora.

Causas y Síntomas del Problema

El análisis de la cartera de crédito ha permitido identificar diversos factores que inciden en la morosidad de los clientes:

Falta de segmentación efectiva

La empresa no cuenta con una clasificación detallada de los clientes en función de su nivel de riesgo. Actualmente, los esfuerzos de cobranza se aplican de manera general, sin priorizar aquellos clientes con mayor probabilidad de incumplimiento.

Ausencia de análisis de factores de riesgo

No se ha implementado un estudio detallado para identificar las variables que más influyen en la entrada en mora. Un análisis de Componentes Principales (PCA) permitiría determinar qué factores, como el nivel de ingresos, el historial crediticio o el tipo de producto financiero, entre otros, tienen mayor impacto en la probabilidad de incumplimiento.

Crecimiento de la mora sin intervención temprana: Se ha observado un aumento del 11.18% en la cartera vencida mayor a 30 días y un incremento alarmante del 121.25% en la cartera con más de 90 días de mora. Esto sugiere que los clientes no solo están cayendo en mora, sino que permanecen en esa condición sin una estrategia efectiva de recuperación. (Informe de desempeño de cartera).

Deficiencias en la gestión de cobranza

Más del 60% de los clientes en mora no tienen abogado asignado, lo que retrasa los procesos de recuperación judicial. Adicionalmente, la estrategia de cobranza no está alineada con un sistema de alertas tempranas que permita actuar antes de que los clientes acumulen varios meses de incumplimiento de pago. (Informe de desempeño de cartera).

En este contexto, el deterioro de la cartera de Confirmeza no solo es el resultado de una gestión de riesgos inadecuada, sino también de factores socioeconómicos y comportamentales que afectan la capacidad de pago de los clientes. Los déficits en educación financiera, la facilidad para acceder a crédito sin una evaluación adecuada de sus implicaciones, y la falta de asesoría en torno al sobreendeudamiento son elementos que contribuyen a este problema.

En este sentido, como lo señala el profesor Jairo Parra Quijano, "la función cautelar constituye, al lado de la función cognoscitiva y ejecutiva, una manifestación de la actividad jurisdiccional. Y el derecho a una medida provisoria deriva del derecho a una protección jurídica efectiva"(Tamayo Lombana, 2004, p. 31). Así, cuando la mora de los créditos alcanza niveles elevados y no se puede recuperar la deuda a través de mecanismos convencionales, Confirmeza, a través de su área jurídica, se ve en la necesidad de recurrir a procesos judiciales para recuperar los bienes empeñados, como vehículos, utilizando medidas cautelares. Estas medidas, conforme al artículo 513 del Código de Procedimiento Civil, permiten solicitar el embargo y secuestro de bienes, con el objetivo de salvaguardar el patrimonio del deudor y garantizar que el acreedor pueda hacer efectivo su crédito.

Si bien este proceso judicial ha sido ocasional en el pasado, con el tiempo ha comenzado a repetirse con mayor frecuencia, dado que la morosidad ha aumentado considerablemente. Según lo establecido en el Código Civil, en estos casos las medidas cautelares culminan con la subasta pública de los bienes embargados, cuyo producto se destina a pagar el crédito, intereses y costos procesales. Este enfoque se ha convertido en una herramienta cada vez más relevante para Confirmeza en su intento de mitigar el riesgo de mora

y recuperar los fondos adeudados, evidenciando la necesidad de un sistema de recuperación de cartera más eficaz y adaptado a las nuevas condiciones del entorno socioeconómico.

Estos síntomas reflejan una tendencia preocupante, cada vez más clientes están entrando en mora y permaneciendo en esa condición por períodos prolongados, lo que incrementa la probabilidad de castigo de la cartera.

Diagnóstico de la Situación

Las preocupaciones sobre la liquidez y la solvencia de entidades financieras como Confirmeza están vinculadas directamente a la crisis de la deuda, la cual ocurre después de un periodo de aumento acelerado de la cartera de crédito y concentración de riesgo en un sector específico, como lo son los préstamos para la compra de vehículos. Esto es especialmente relevante cuando factores macroeconómicos, como la caída del Producto Interno Bruto, el alza de la inflación, el aumento de las tasas de interés y la devaluación del tipo de cambio, incrementan la cartera morosa. En este contexto, Confirmeza, al igual que otros actores del sector financiero, debe incrementar sus provisiones y, después de un cierto tiempo, castigar los créditos morosos. Este proceso no solo afecta el resultado económico de la empresa, sino que erosiona su patrimonio y podría reducir su capital por debajo del nivel mínimo legal, lo que pone en riesgo su estabilidad financiera. La combinación de estos factores puede generar una presión significativa sobre su capacidad de otorgar nuevos créditos y comprometer su salud financiera (Abisetti, 2021, p. 61).

Si la empresa no implementa estrategias basadas en el análisis de datos, es probable que la tasa de morosidad continúe en ascenso. Actualmente, el índice de deterioro de la cartera ha aumentado del 7% en 2023 al 9% en 2024, lo que representa un incremento de $2.885 millones en cartera deteriorada (Informe de desempeño de cartera). Aunque la tasa de

recuperación mejoró del 35% en 2023 al 44% en 2024, el volumen de nuevos clientes en mora sigue superando la capacidad de recuperación, generando un efecto acumulativo negativo.

El análisis del rodamiento de cartera indica que el mayor crecimiento de la morosidad se da en clientes que recientemente adquirieron créditos, lo que evidencia la necesidad de mejorar los modelos de evaluación del riesgo. Además, la falta de segmentación y análisis de causas impide la personalización de estrategias de cobranza, lo que reduce la efectividad de las acciones para mitigar el problema.

En este contexto, la empresa enfrenta dos grandes desafíos: comprender las razones por las cuales los clientes entran en mora y desarrollar herramientas que permitan anticipar estos eventos. Para ello, es fundamental implementar un enfoque basado en el análisis de datos y modelos predictivos que faciliten la identificación temprana de clientes en riesgo.

Propuesta de Solución

Se compartió una propuesta inicial para trabajar mediante una iniciativa en referencia al análisis de datos, el documento “Proyecto Machine Learning”, con el fin de desarrollar un modelo que permita evaluar y segmentar los créditos de vehículos, optimizando los procesos de financiación y mejorando la estrategia de asignación de recursos. Estos objetivos específicos buscan identificar patrones en los datos históricos de créditos otorgados, agrupar a los clientes según criterios como riesgo, capacidad de pago y tipo de vehículo, establecer perfiles de clientes para ajustar las estrategias de financiamiento y facilitar la evaluación de créditos, así como la predicción del comportamiento futuro de los clientes.

Dentro de este proyecto, se propone un análisis descriptivo para identificar correlaciones y distribuciones relevantes, además de la visualización de datos clave para determinar tendencias iniciales. La compañía busca identificar patrones y concentraciones en los clientes actuales con el fin de analizar acciones que le permitan mejorar los indicadores de

cartera, así como optimizar la asignación de recursos presupuestados para la gestión de cartera. Con la implementación, se busca identificar patrones de acuerdo con el origen del préstamo, el número de cuotas restantes, y ubicar a los clientes en segmentos específicos que puedan indicar comportamientos de pago, riesgo o necesidades de manejo de cartera (A. Rodríguez, comunicación personal, 24 de enero de 2025).

Para abordar este problema, se propone una estrategia basada en análisis de datos y machine learning, con los siguientes componentes:

Segmentación de clientes en función de su riesgo de mora: Implementación de gráficos que permitan visualizar la cartera segmentada por categorías de riesgo, antigüedad de mora y tipo de cliente.

Identificación de las principales causas de mora mediante Análisis de Componentes Principales (PCA): Aplicación de técnicas de reducción de dimensionalidad para determinar los factores que más influyen en la morosidad y diseñar estrategias de mitigación basadas en estos hallazgos (Wang S, 2024).

Clasificación de clientes en riesgo de incumplimiento: Desarrollo de modelos predictivos como Regresión logística, Random Forest y XGBoost para calcular la probabilidad de que un cliente entre en mora en los próximos meses. Esto permitirá priorizar los esfuerzos de cobranza en función del nivel de riesgo.

Este enfoque permitirá a Confirmeza S.A.S. optimizar su gestión de cartera, reducir la morosidad y mejorar su liquidez, garantizando una estrategia más eficiente en la recuperación de créditos y en la prevención del deterioro de la cartera.

De acuerdo conÓskarsdóttir et al. (2019), una herramienta para evaluar a los clientes según sus características, para decidir si otorgar un crédito o no, y con qué condiciones, es la

calificación crediticia. No obstante, esta calificación debe estar acompañada de una estrategia definida para la gestión de las cuentas por cobrar.

Es una pieza fundamental para el bienestar financiero de las empresas definir una política de crédito y cobranza eficientes, que permitan seleccionar a los buenos clientes, que son aquellos que adquieren grandes volúmenes de bienes y los pagan conforme a lo convenido, así como definir un proceso eficiente de cobranza, permita convertir las cuentas por cobrar en dinero en efectivo Davenport (2014)

Como menciona Davenport (2014), las herramientas analíticas de calificación crediticia han evolucionado para apoyar la toma de decisiones, considerando Business Intelligence, Business Analytics, Big Data y finalmente, el Big Data Analytics.

Pregunta de investigación.

¿Cuáles son las principales variables que explican la entrada en mora de los clientes de Confirmeza S.A.S. y cómo pueden utilizarse modelos predictivos para anticipar el incumplimiento?

Objetivos

Objetivo general

Desarrollar un modelo predictivo basado en machine learning, utilizando datos históricos de cartera, con el fin de diseñar una herramienta eficaz para anticipar la entrada en mora de los clientes de Confirmeza S.A.S. y apoyar la formulación de estrategias de prevención de incumplimiento.

Objetivos específicos.

I. Explorar y segmentar la cartera de clientes de Confirmeza S.A.S., clasificándolos según su estado de mora, características sociodemográficas etc., a través de técnicas de análisis de datos y técnicas de visualización, con el propósito de identificar patrones de comportamiento de pago.

II. Determinar las variables más relevantes en la entrada en mora, aplicando técnicas de análisis exploratorio, para comprender los factores de mayor impacto en el incumplimiento de pagos.

III. Desarrollar un modelo de clasificación para la predicción de mora, empleando algoritmos de machine learning XGBoost y redes neuronales LSTM, con el objetivo de anticipar qué clientes tienen mayor riesgo de incumplimiento.

IV. Evaluar el desempeño del modelo predictivo, mediante métricas de validación como precisión, recall y curva AUC-ROC, con el fin de garantizar su efectividad y confiabilidad en la detección temprana del riesgo crediticio.

V. Proponer estrategias de prevención y recuperación de cartera, basadas en los resultados obtenidos del análisis y la clasificación, con el propósito de optimizar la toma de decisiones y reducir el impacto de la morosidad en la empresa.

Justificación

En un entorno financiero cada vez más dinámico y competitivo, la adecuada gestión del riesgo crediticio se ha convertido en un pilar fundamental para la sostenibilidad de las empresas que operan en el sector de financiamiento automotriz. Confirmeza S.A.S., como entidad especializada en la colocación y financiamiento de vehículos, enfrenta desafíos constantes en la identificación y mitigación de riesgos asociados a la mora de sus clientes.

Este proyecto busca aportar valor mediante el análisis de las principales variables que explican la entrada en mora de los clientes de Confirmeza S.A.S. y la implementación de modelos predictivos que permitan anticipar posibles incumplimientos. El uso de herramientas de analítica de datos se convierte en un aliado estratégico para mejorar la toma de decisiones, optimizar la gestión de cartera y reducir la exposición a pérdidas financieras.

A través de un enfoque basado en datos, se explorarán patrones de comportamiento en los clientes, considerando factores como historial crediticio, capacidad de pago, tipo de financiamiento y características del vehículo adquirido. La aplicación de modelos estadísticos y de machine learning permitirá desarrollar una visión más precisa del perfil de riesgo de cada cliente, lo que facilitará la implementación de estrategias preventivas y la optimización de los procesos de cobranza.

Desde una perspectiva académica y profesional, este estudio no solo contribuirá a la mejora operativa de la empresa, sino que también reforzará la importancia de la analítica de datos como una herramienta clave en la gestión financiera. El proyecto permitirá demostrar cómo el uso adecuado de modelos predictivos puede traducirse en una mayor estabilidad para la compañía, una relación más confiable con los clientes y una reducción significativa en los índices de morosidad.

En conclusión, esta investigación representa una oportunidad para integrar el conocimiento teórico sobre riesgo financiero con aplicaciones prácticas en el sector automotriz, generando soluciones innovadoras que puedan ser replicables en otras organizaciones con dinámicas similares. Además, contribuirá al fortalecimiento de estrategias que fomenten la sostenibilidad del negocio a largo plazo, promoviendo una cultura de toma de decisiones basada en datos y en la optimización de recursos.

La presente investigación sobre la predicción de morosidad en la empresa Confirmeza S.A.S. es relevante y necesaria por su contribución a la estabilidad financiera de la empresa, la mejora de la toma de decisiones y el desarrollo de modelos predictivos en el ámbito del riesgo crediticio. Los siguientes son los principales criterios que respaldan la importancia de este estudio:

Conveniencia

La creciente morosidad en la cartera de crédito de Confirmeza S.A.S. tiene un impacto directo en su estabilidad financiera y operativa. La tasa de morosidad de la compañía a más de 90 días ha aumentado un 106,18% en los últimos años, dejando una cartera comprometida con un déficit de más de 3.800 millones. Dado que la compañía no cuenta con herramientas avanzadas de gestión del riesgo crediticio, el desarrollo de modelos predictivos basados en machine learning ayudará a identificar a los clientes con mayor probabilidad de incumplimiento y mejorar las estrategias de cobro, asegurando una gestión más eficiente de la cartera.

Relevancia social

Los impactos que la empresa tiene en cuanto a la morosidad crediticia también afectan a los clientes y al sistema financiero en general. Un alto índice de incumplimiento disminuye el acceso al crédito para nuevos clientes. La implementación de modelos predictivos para evitar la

morosidad permitirá a las empresas ofrecer mejores condiciones de financiación y apoyar el crecimiento de los clientes, fomentando así una cultura de pago responsable.

Implicaciones prácticas

La gestión de cobranza en Confirmeza S.A.S. no cuenta con estrategias de cobranza del todo efectivas lo cual ha limitado la recuperación de cartera. Teniendo un modelo basado en análisis de datos y técnicas de Machine Learning se podrá mejorar el flujo de caja y fortalecer el área de financiamiento de la empresa. Con dicho modelo se podrá tener una segmentación de clientes según su perfil de riesgo, predecir incumplimientos, optimizar la cobranza y reducir pérdidas.

Valor Teórico

La aplicación de modelos predictivos en la gestión de riesgo crediticio es un área de creciente interés en ciencia de datos y finanzas. Este estudio contribuirá al campo al evaluar el impacto de variables sociodemográficas, comportamiento de pago y factores económicos en la predicción de morosidad.

Utilidad metodológica

La investigación proporcionará una metodología estructurada para la implementación de modelos de análisis de riesgo crediticio en empresas financieras. La combinación de técnicas estadísticas y de Machine Learning permitirá generar un enfoque replicable para otras organizaciones con desafíos similares. Además, la integración de herramientas de visualización facilitará la interpretación de los resultados y su aplicación en la toma de decisiones empresariales.

Aplicando el modelo basado en análisis de datos e implementando la técnica de Machine Learning en la empresa Confirmeza S.A.S., permitirá mejorar su gestión de riesgo crediticio y optimizar la recuperación de cartera.

Como parte del Programa PAT - Seminario de Investigación en Ingeniería, dentro de la Especialización - Grupo 10 - FIN - Primer Semestre de 2025, este proyecto es desarrollado por Henny Rocío Carrillo, Wilber Alexander Rodríguez Castro, Juan Sebastián Ballén Villalba y Michael Stiven Escobar Rodríguez, quienes, a través de un enfoque analítico y aplicado, buscamos aportar soluciones innovadoras a la gestión del riesgo financiero en el sector de financiamiento automotriz, utilizando modelos predictivos y herramientas de analítica de datos para optimizar la toma de decisiones y fortalecer la sostenibilidad del negocio.

Marco Teórico

Los créditos han sido otorgados a diferentes actores de la sociedad y su figura se puede evidenciar desde épocas muy antiguas, remontándose a Mesopotamia, la Grecia clásica, y Roma republicana.

El crédito se define como un préstamo en dinero, donde la persona se compromete a devolver la cantidad solicitada en el tiempo o plazo definido según las condiciones establecidas para dicho préstamo, más los intereses devengados, seguros y costos asociados si los hubiere. (Morales & Morales, 2014).

Hay un concepto denominado las 5 del crédito, de acuerdo con Morales & Morales (2014), las 5 c del crédito contemplan los factores de riesgo que deberán ser evaluados al realizar un análisis de crédito.

Conducta

El objetivo de evaluar la conducta es determinar la calidad moral y capacidad administrativa de los clientes, a través de un análisis cualitativo del riesgo del deudor, que incluye evaluar la calidad y veracidad de la información del cliente, el desempeño en el pago de sus obligaciones con los bancos y con otros acreedores, liderazgo y las consecuencias en su operación.

Capacidad de pago histórica

El objetivo de analizar la capacidad de pago histórica es evaluar la habilidad del cliente de haber generado, en el pasado, los recursos suficientes para cumplir con sus compromisos financieros a través de un análisis cuantitativo de su riesgo financiero. Aquí se contempla el análisis de ventas netas, márgenes de utilidad y generación de flujo neto para cubrir el pago de intereses, capital, dividendos e inversiones y sus tendencias y comparación.

Capacidad de endeudamiento

El objetivo de este factor es medir la solidez de la estructura financiera de la empresa, evaluando la congruencia de los recursos solicitados acordes con su giro principal; todo ello a través de un análisis cuantitativo del riesgo financiero del deudor. Aquí se contempla el análisis de tendencias y comparación con la industria de los índices de liquidez, apalancamiento, rentabilidad y eficiencia.

Condiciones macroeconómicas

El objetivo de este factor es determinar el comportamiento de la industria en su conjunto, para determinar la influencia que tiene en la capacidad y fortaleza financiera del deudor.

Capacidad de pago proyectada

El objetivo de este factor es analizar la capacidad que tiene un cliente para generar efectivo suficiente en el futuro, y cumplir sus compromisos financieros, con base en la viabilidad de su negocio. dentro de la industria.

Es importante señalar que cuando se solicita un crédito, adicional a las 5 deben considerarse factores adicionales como: factores gerenciales, factores financieros, factores industriales, factores de negocios, y factores de seguimiento de la cuenta.

En Colombia, existen diferentes tipos de crédito:

- Crédito hipotecario.
- Crédito de nómina o libranza.
- Crédito de libre inversión.
- Tarjetas de crédito.

- Créditos comerciales.
- Créditos educativos
- Compra de cartera.
- Leasing

Junto con el concepto crédito, se encuentra el concepto de cobranza, el cual inicia después de que ha otorgado el crédito y el cliente debe pagarlo. Una empresa puede verse envuelta en problemas financieros al no tener la capacidad de convertir sus cuentas por cobrar en efectivo.

“Una actividad fundamental es la prevención, a través del conocimiento mejor de los clientes, y teniendo cuidado especial en el otorgamiento de créditos, para que la administración de la cobranza sea eficiente. Otra medida es reaccionar de manera inmediata y atinada a la situación ya existente, para ello se debe tener una administración óptima de cartera de clientes y haber determinado estrategias para las situaciones en que los clientes no cumplen con sus pagos, las cuales deben ser consistentes y adecuadas a la situación muy particular del mercado, a la economía y, sobre todo, a las peculiaridades del tipo de cliente; todo esto debe conducir a una cobranza eficiente y oportuna. Para una adecuada administración de la cartera de crédito, es importante conocer a los clientes de la empresa, sus hábitos de compra, qué estímulos los hacen reaccionar, y además qué factores sirven para medir riesgo y de qué manera se pueden evitar las pérdidas como consecuencia de la presencia de esos riesgos en el proceso de cobranza.”(Morales & Morales, 2014, pp 145).

De acuerdo con Turing (2012) la inteligencia artificial se define como la imitación de la inteligencia de los seres humanos haciendo uso de algoritmos a través de un sistema informático o un ordenador. En su artículo "Computing Machinery and Intelligence", incluyó la pregunta "¿Pueden las máquinas pensar?" y estableció a su vez un criterio para evaluar la

inteligencia de una máquina conocida en la actualidad como el “Test de Turing”. Por otra parte, planteó el siguiente enfoque: "En lugar de tratar de producir un programa que simule la mente adulta, ¿por qué no más bien intentar producir uno que simule la mente de un niño? Si esto se hiciera, entonces podría someterse a un proceso de educación.", lo anterior sugiere que la IA debería aprender y desarrollarse a través de la experiencia, una idea fundamentada en el aprendizaje automático moderno. Así las cosas, la Inteligencia Artificial busca desarrollar sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de patrones, el aprendizaje y la toma de decisiones. (Russell & Norvig, 2021).

El Machine Learning por su parte es una rama de la inteligencia artificial enfocado en el desarrollo de algoritmos y modelos que pueden aprender y mejorar a partir de datos, en lugar de ser programados explícitamente para realizar una tarea específica. De acuerdo con la definición de, Samuel & Gabel (1959), quien fue pionero en el campo de aprendizaje “El campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas.”

Samuel & Gabel (1959), explica que el aprendizaje automático se divide en cinco grandes categorías: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado, aprendizaje autosupervisado y aprendizaje por refuerzo.

Machine Learning aplicado a finanzas

El crédito es imprescindible en los sistemas financieros. Todas las instituciones que asignan créditos, tienen la necesidad de comprender el riesgo que esto conlleva y tomar la mejor decisión entre conceder un crédito y negarlo, para ello se apoyan en una de las aplicaciones más exitosas de la estadística y la investigación operativa, que es el score crediticio.

El desarrollo del mercado de entrega de crédito ha generado un gran interés por los modelos cuantitativos de riesgo de crédito, por lo que la modelización del riesgo de crédito es un subcampo muy activo de las finanzas cuantitativas y la gestión de riesgos. Los modelos de gestión del riesgo de crédito se utilizan para determinar la distribución de pérdidas de una cartera de préstamos, a lo largo de un periodo de tiempo fijo, y para calcular medidas de riesgo basadas en la distribución de pérdidas. (Scagliola, 2022, pp 16).

Los métodos tradicionales de evaluación del riesgo crediticio pueden tardar mucho tiempo y estar sesgados a los juicios de cada persona, ya que se basan generalmente en la evaluación manual y en reglas predeterminadas. Esta situación experimenta un cambio gracias al crecimiento exponencial de los datos y los avances en las técnicas de aprendizaje automático, lo que permite una evaluación del riesgo crediticio mediante la aplicación de algoritmos automatizados.

La demanda de crédito genera una información considerable, con la que, analizada mediante Big Data, se diseñan nuevos productos, modelos de machine learning y métodos de evaluación del riesgo de crédito. Consecuentemente, en escenarios de aumento de la demanda, los riesgos de crédito también escalan considerablemente, de forma no lineal, considerando el nivel de riesgo, la tasa y los plazos del crédito. Del mismo modo, existe la expectativa de un aumento del fraude en el año siguiente (Noriega et al., 2023)

En el contexto financiero, la inteligencia artificial (IA) y el Machine learning (ML) han sido ampliamente utilizados para mejorar la gestión del riesgo de crédito, optimizar carteras de inversión y detectar fraudes financieros. De acuerdo con Wang et al. (2020), los algoritmos de aprendizaje automático permiten a las entidades bancarias analizar grandes volúmenes de datos y predecir la probabilidad de incumplimiento de pagos, lo que facilita la toma de decisiones estratégicas.

Por su parte, Sandeep Shinde & Satish Kale (2023) en el contexto de la evaluación del riesgo crediticio, los algoritmos de aprendizaje automático ofrecen la posibilidad de mejorar la precisión y la eficiencia del proceso de evaluación, permitiendo a los prestamistas evaluar la solvencia con mayor precisión y agilidad.

Por otra parte, los modelos de redes neuronales y aprendizaje profundo han demostrado ser eficaces en la identificación de patrones anómalos en transacciones financieras, contribuyendo a la prevención del lavado de dinero y el fraude, según, Wang et al. (2020). La aplicación de IA conforme con expuesto por Baesens et al. (2016), en la gestión de cartera permite la optimización en estrategias de cobranza, priorizando clientes con mayor riesgo así mismo reducir la morosidad, mejorando la predicción y prevención de incumplimientos.

Actualmente, se presentan retos y enfoques en la predicción de riesgo crediticio mediante modelos de Maching Learning, de acuerdo como lo mencionan Herrera, Rivera, y Noriegua (2023) manifestando las dificultades en la implementación de métodos como el modelo de la caja negra, la necesidad de inteligencia artificial explicativa, la importancia de seleccionar características relevantes, y el problema de desequilibrio en los datos de entrada. Sin embargo, identifican la más grande limitación con la representatividad de la realidad, y las variables principalmente usadas en la industria del microcrédito son datos relacionados con el comportamiento demográfico, de operación y de pago.

Para Shi et al. (2022) los métodos de aprendizaje profundo son más potentes que el aprendizaje automático tradicional y los enfoques estadísticos, De igual manera, han demostrado mediante su estudio que los conjuntos de varios métodos superan a uno solo, lo que también ha quedado demostrado en investigaciones relacionadas.

Para la realización de la presente investigación se hará uso de modelos híbridos métodos de aprendizaje supervisado y no supervisado, con el fin de poder identificar comportamientos con factores de riesgo basados en datos históricos, a continuación, se presenta explicación de algunos modelos supervisados, los cuales serán considerados para finalmente establecer los modelos a implementar.

Regresión logística.

La regresión logística es una técnica estadística utilizada para modelar la probabilidad de un evento binario, es decir, un resultado que puede tener dos posibles valores, como "éxito" o "fracaso", usa a su vez una función logística que permite modelar una variable dependiente que puede ser binaria, multinomial u ordinal.

La clase binaria se utiliza en situaciones en las que el resultado obtenido tiene dos categorías, expresadas generalmente como “0” y “1”. Se presenta con la siguiente función:

$$P(Y=1) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+..+\beta_n X_n)}}$$

Por su parte la clase Multinomial, se usa cuando la variable dependiente tiene tres o más categorías sin orden. La clase ordinal se aplica cuando la variable dependiente tiene tres o más categorías con un orden jerárquico.

La elección del tipo de regresión depende de si la variable dependiente tiene dos, tres o más categorías, y si estas tienen un orden jerárquico o no. A diferencia de la regresión lineal, que se utiliza para estimar valores continuos, la regresión logística se aplica cuando la variable objetivo es de tipo categórico. Además, esta técnica permite trabajar con variables independientes que pueden ser tanto continuas como categóricas.

La función logística se basa en la función sigmoide, la cual genera una curva en forma de "S". Su principal característica es que transforma cualquier valor real en un resultado

comprendido entre 0 y 1, lo que la hace ideal para modelar probabilidades en problemas de clasificación.

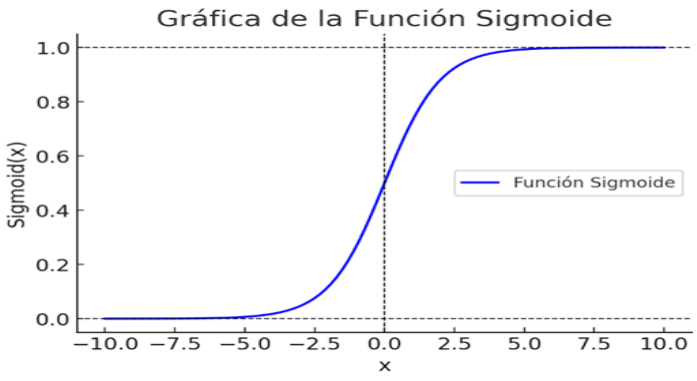
En el ámbito del aprendizaje automático (Machine Learning), la función sigmoide establece una relación entre la variable dependiente y las variables independientes. Su comportamiento se representa mediante una curva continua y suave que restringe sus valores dentro del rango de 0 a 1, sin exceder estos límites. La ecuación matemática que describe esta función es la siguiente:

$$f(x) = \frac{1}{1 + e^{-x}}$$

En esta ecuación, x representa un número real, y su comportamiento puede analizarse en función de sus límites. Cuando x se aproxima a menos infinito, el valor del cociente se acerca progresivamente a cero. Por el contrario, cuando x tiende a infinito, el cociente converge hacia uno.

Figura 1

Gráfico de una función sigmoide



Nota. Elaboración propia basada en la ecuación de la función sigmoide.

Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado basado en árboles de decisión que pertenece a la categoría de métodos ensemble, lo que significa que combina múltiples modelos de Machine Learning para mejorar la precisión de las predicciones. Cada modelo genera una predicción independiente, y estas se combinan para obtener un resultado final. Para la combinación de predicciones se emplean diferentes estrategias, como votación por mayoría, bagging, boosting y stacking (Zhang & Ma, 2012). Los modelos de Random Forest fueron inicialmente propuestos por Ho (1998) y posteriormente generalizados y denominados "Random Forest" por (Breiman, 2001). Estos modelos tienen una amplia variedad de aplicaciones, incluyendo la clasificación de imágenes, motores de recomendación y selección de características (Louppe, 2014). Además, se utilizan en áreas como la detección de fraudes, la predicción de enfermedades y la clasificación de solicitantes de préstamos, y se caracterizan por construir múltiples árboles de decisión durante el entrenamiento y combinar sus predicciones para mejorar la precisión y reducir el sobreajuste (James et al., 2013).

El método de Random Forest ópera en tres etapas principales:

- Construcción de múltiples árboles de decisión: Se generan varios árboles de decisión, cada uno entrenado con una muestra aleatoria del conjunto de datos original (Bootstrap Sampling). Cada árbol aprende de un subconjunto diferente, lo que introduce diversidad en el modelo.

- Generación de predicciones individuales: Cada árbol del bosque emite una predicción basada en los datos con los que fue entrenado. En problemas de clasificación, cada árbol vota por una categoría; en problemas de regresión, se generan estimaciones numéricas.

- Agregación de predicciones y resultado final: para clasificación se aplica una votación por mayoría, seleccionando la categoría con más votos como la predicción final.

Para regresión: Se calcula el promedio de las predicciones de todos los árboles para obtener un resultado más preciso y estable.

Los modelos de Random Forest ofrecen múltiples beneficios, destacándose por su flexibilidad y facilidad de uso en problemas tanto de clasificación como de regresión. Su capacidad para generar predicciones precisas y robustas se debe a la combinación de múltiples árboles de decisión, lo que reduce la varianza del modelo y mejora su capacidad de generalización. Uno de sus mayores atributos es su resistencia al sobreajuste. A medida que aumenta el número de árboles en el bosque, el modelo se vuelve más estable, evitando la excesiva adaptación a los datos de entrenamiento. De hecho, tiene un límite teórico de sobreajuste (Hastie et al., 2009), lo que lo hace ideal para conjuntos de datos complejos y ruidosos. Además, puede manejar una gran cantidad de variables de entrada, lo que lo convierte en una opción poderosa para el análisis de datos de alta dimensión (Hastie et al., 2009). Otro beneficio importante es su capacidad para manejar datos faltantes, permitiendo realizar predicciones sin que la precisión se vea significativamente afectada.

A pesar de sus ventajas, Random Forest también tiene algunas limitaciones. Su costo computacional puede ser elevado, ya que el entrenamiento de múltiples árboles puede ser lento y demandante en términos de memoria, lo que lo hace menos eficiente en escenarios donde la rapidez de respuesta es prioritaria(Louppe, 2014). Asimismo, uno de los desafíos más importantes es su falta de interpretabilidad. A diferencia de un árbol de decisión individual, donde es posible visualizar fácilmente la estructura de las decisiones, Random Forest opera como un modelo en conjunto, dificultando la comprensión de cómo se toman las decisiones específicas por Breiman (2001), debido a esto, se considera un modelo predictivo más que

descriptivo, lo que puede ser una desventaja en aplicaciones donde la interpretabilidad es clave.

XGBoost (Extreme Gradient Boosting).

Es un algoritmo de aprendizaje automático basado en árboles de decisión que emplea la técnica de Gradient Boosting para optimizar el rendimiento predictivo (Chen & Guestrin, 2016). Se caracteriza por ser altamente eficiente, flexible y escalable, lo que lo ha convertido en una de las herramientas más utilizadas en competencias de ciencia de datos y aplicaciones del mundo real.

XGBoost se desarrolló con un enfoque en la optimización computacional, utilizando técnicas como la paralelización del cálculo de árboles, el uso eficiente de memoria y la poda estratégica de árboles para evitar el sobreajuste. Además, incorpora regularización L1 y L2, lo que lo hace más robusto frente a datos ruidosos en comparación con otros algoritmos de boosting.

Fue desarrollado por Tianqi Chen como parte de su trabajo de investigación en la Universidad de Washington y se presentó en 2016 en la conferencia Knowledge Discovery and Data Mining (KDD). Su código fue lanzado como software de código abierto y rápidamente ganó popularidad en la comunidad de Machine Learning.

Redes neuronales

Las redes neuronales son modelos computacionales inspirados en el funcionamiento del cerebro humano, diseñados para reconocer patrones, aprender de datos y tomar decisiones. Están hechas por programas que pueden aprender de la experiencia, es decir, de los datos, y con el tiempo se vuelven mejores para reconocer cosas, tomar decisiones o hacer predicciones, como si “aprendieran” por sí solas. De acuerdo con Aurélien Géron (2022) las redes neuronales son modelos computacionales inspirados en la estructura del cerebro

humano, compuestos por capas de nodos (neuronas artificiales) interconectados que procesan información mediante funciones de activación no lineales. Estas redes son capaces de aprender representaciones complejas a partir de datos, lo que las hace especialmente útiles para tareas como clasificación de imágenes, reconocimiento de voz y procesamiento del lenguaje natural.

Redes Neuroales LSTM (Long Short-Term Memory)

Las redes LSTM son un tipo de arquitectura de red neuronal que permite aprender patrones en secuencias largas al mantener una memoria interna que regula qué información conservar, olvidar o actualizar con cada nuevo dato que recibe. Gracias a esta capacidad, son especialmente útiles en tareas donde el contexto previo es clave, como la traducción automática, el análisis de sentimientos o la predicción de series temporales. Aurélien Géron (2022) introduce las redes LSTM (Long Short-Term Memory) como una arquitectura especializada de redes neuronales recurrentes (RNN) diseñada para manejar secuencias de datos y superar las limitaciones de las RNN tradicionales, especialmente en lo que respecta a la retención de información a largo plazo.

Algunas de las técnicas utilizadas para el desarrollo de los modelos supervisados antes mencionados son:

Análisis exploratorio de datos (EDA)

Es considerado un enfoque estadístico para analizar conjuntos de datos, de acuerdo con Aurélien Géron (2022) el EDA se presenta como una etapa fundamental en el desarrollo de proyectos de aprendizaje automático, implica examinar los datos para identificar patrones, detectar valores atípicos, comprender las relaciones entre variables y evaluar la calidad de los datos. Este proceso es esencial para tomar decisiones informadas sobre la selección de características, la ingeniería de variables y la elección de algoritmos adecuados.

Análisis de Componentes Principales (PCA)

El PCA es una técnica estadística utilizada para reducir la dimensionalidad de conjuntos de datos, transformando variables posiblemente correlacionadas en un conjunto de variables no correlacionadas llamadas componentes principales. Esta transformación facilita la identificación de patrones y estructuras subyacentes en los datos, mejorando su interpretabilidad sin perder información significativa. Una vez reducida la dimensionalidad se pueden emplear métodos de machine learning con mayor eficiencia. Pearson (1901) introdujo el PCA como un método para describir la variabilidad en un espacio multidimensional mediante variables no correlacionadas, facilitando la comprensión de estructuras de datos complejas.

Para la aplicación del análisis de componentes principales (PCA) es importante realizar una limpieza y estandarización de los datos, así mismo realizar aplicación de la matriz de covarianza o correlación, que de acuerdo con Van Kampen (1981), es la generalización de la varianza para múltiples dimensiones, proporcionando una medida de cómo las variables aleatorias varían conjuntamente. Una vez calculada la matriz de covarianza se obtienen los valores propios, los cuales indican la cantidad de varianza por cada componente y vectores propios, que representan las direcciones de los componentes principales. Por último, la selección del número de componentes se realiza a través del método del codo o seleccionado el número de componentes que explican un porcentaje alto de varianza acumulada.

Marco institucional

Confirmeza SAS es una empresa colombiana con más de 12 años de experiencia en la financiación de vehículos nuevos, usados y pólizas. Su misión y visión son:

Misión: Proponer soluciones y servicios, ofreciendo el apoyo financiero a los sueños de familias y compañías en el territorio nacional.

Visión: Ser uno de los principales actores en la financiación de cartera de consumo y comercial, estableciendo sinergias con nuestros stakeholders. Comprometidos con el medio ambiente y la sociedad.

Ubicación Confirmeza SAS tiene su sede principal en la ciudad de Bogotá, ubicada en la Calle 224 No. 9 -60, Costado Oriental. La compañía forma parte del conglomerado ecuatoriano Corporación Eljuri, con presencia en diversos sectores económicos y amplia experiencia en servicios financieros, comerciales y automotrices en Latinoamérica.

Sector económico al CIIU La compañía está registrada bajo la actividad económica CIIU 6494 - “Otras actividades de distribución de fondos”. Adicionalmente, su operación también puede estar relacionada con el CIIU 6492 - “Actividades de financiación mediante contratos de arrendamiento (leasing)”. Su participación en este sector le permite atender una demanda creciente en el financiamiento de vehículos, ofreciendo productos especializados para diferentes perfiles de clientes.

Nicho de mercado Confirmeza SAS se especializa en el sector financiero automotriz, con un enfoque en:

- Clientes individuales y corporativos que buscan adquirir vehículos nuevos mediante crédito o leasing, brindando planes de financiamiento adaptados a sus necesidades y capacidades de pago.

- Concesionarios de vehículos que requieren aliados financieros para facilitar la venta de unidades y garantizar el acceso a créditos rápidos y efectivos para sus clientes. Entre los principales se destacan toda la red de vitrinas Metrokia, Autocom y Greccomotors a nivel nacional.
- Empresas aseguradoras y brokers que ofrecen pólizas de seguro obligatorias y voluntarias ligadas al financiamiento vehicular, asegurando una protección integral para los clientes y la institución financiera.
- Fabricantes y ensambladoras de automóviles que buscan mecanismos de financiamiento para impulsar sus ventas, garantizando mayor rotación de inventario y accesibilidad a sus productos.

Principales productos y procesos Confirmeza SAS ofrece productos financieros especializados, tales como:

- Colocación de crédito para vehículos comerciales y de consumo.
- Arrendamiento financiero (leasing) de vehículos, facilitando opciones de pago flexibles y adaptadas a diferentes perfiles de clientes.
- Ofrecimiento de pólizas de seguro vinculadas a la financiación de vehículos, asegurando la protección tanto del bien como del cliente.
- Evaluación y gestión del riesgo crediticio en la colocación de financiamientos, minimizando el riesgo de mora y asegurando la sostenibilidad del negocio.
- Estructura organizacional Confirmeza SAS cuenta con una estructura organizacional dividida en diversas áreas clave:
- Área Comercial: Encargada de la recepción, perfilamiento y asignación de negocios a entidades financieras mediante modelos de F&I o colocación directa.

Su labor es fundamental en la generación de nuevas oportunidades y alianzas estratégicas.

- Área de Crédito: Evalúa la documentación de los clientes, determina la viabilidad del negocio y capacidad de pago, asegurando la colocación de créditos sostenibles y rentables.
- Área de Operaciones: Gestiona las condiciones del negocio, aprobaciones de tasas, verificaciones de seguros y otras garantías, garantizando el cumplimiento normativo y la protección de activos financieros.
- Área de Servicio al Cliente: Monitorea la calidad del servicio, fidelización y resolución de casos especiales, asegurando una experiencia satisfactoria para los clientes.
- Área de Gestión de Cartera y Jurídico: Realiza el seguimiento de hábitos de pago, cobranza y procesos jurídicos en casos de mora, protegiendo la estabilidad financiera de la organización.
- Área de Tesorería: Administra los fondos y flujo de caja, garantizando la liquidez y eficiencia en el manejo de recursos.
- Área Contable: Gestiona los estados financieros y resultados económicos de la compañía, asegurando la transparencia y cumplimiento normativo.
- Elementos del área Cada área de la compañía cuenta con procesos estructurados para garantizar una eficiente gestión del financiamiento de vehículos. Estos incluyen:
- Perfilamiento y evaluación crediticia de clientes, asegurando condiciones óptimas para la colocación de créditos.

- Procesamiento y verificación de documentación para garantizar la confiabilidad de los datos y la seguridad de las operaciones.
- Administración de seguros y garantías asociadas a los créditos.
- Seguimiento de cartera y gestión de cobranza, minimizando el riesgo de mora y mejorando la recuperación de cartera.
- Optimización de flujos de efectivo y administración financiera, asegurando la sostenibilidad del negocio.

Proceso o tema de estudio

Dando un enfoque particular a este estudio, se tiene como objetivo analizar patrones de comportamiento de pago y desarrollar modelos de predicción que permitan a la compañía optimizar su gestión de cartera, reduciendo la morosidad y mejorando así la liquidez. A través del uso de técnicas de minería de datos, machine learning y análisis financiero, se espera identificar factores determinantes en el incumplimiento y diseñar estrategias proactivas para mitigar riesgos.

La implementación de estrategias basadas en estos modelos permitirá una toma de decisiones más efectiva en la concesión de créditos, facilitando la prevención del deterioro de la cartera. Además, se busca desarrollar herramientas predictivas que ayuden a establecer políticas de crédito más eficientes y alineadas con la realidad del mercado.

Este enfoque no solo beneficiará a Confirmeza SAS, sino que también fortalecerá la confianza de sus clientes y aliados estratégicos, consolidando su posición como un referente en el financiamiento automotriz en Colombia.

Metodología

Enfoque de la investigación

Enfoque cuantitativo: en el contexto de este estudio, se plantea una investigación con un **enfoque cuantitativo**, dado a que se busca analizar datos de clientes, con el fin de identificar patrones de morosidad, analizando la relación entre diversas variables, lo anterior con el fin de predecir el riesgo de incumplimiento mediante modelos de ML.

Diseño de la investigación

Diseño no experimental, correlacional, explicativo – aplicado:

Diseño no experimental: no se manipularán las variables involucradas en el estudio, los datos recolectados corresponden a la realidad de las colocaciones efectuadas por la empresa y sus hábitos de pago, lo que permitirá observar la relación entre las diferentes variables sin intervenir en su evolución natural. Este enfoque es adecuado para estudios en los que no se requiere la manipulación directa de las variables, sino que se parte de los datos ya existentes.

Transversal: los datos se recolectarán en un solo momento, haciendo uso de una base de transacciones que identifican el estado de la cartera con corte al mes de marzo del año 2025. Lo que permite obtener una instantánea de la situación de los deudores y sus préstamos en ese momento específico.

Correlacional y aplicado: se pretende identificar la relación a través del análisis de las distintas variables (factores socioeconómicos, comportamiento financiero, historial de pagos) y predecir la ocurrencia mediante modelos de ML. Lo anterior ayudará a identificar tendencias en el comportamiento de los deudores, que podrían ser útiles para futuras intervenciones o estrategias dentro de una organización o para la optimización del modelo de crédito. Así mismo, los resultados serán utilizados para mejorar la gestión de cartera en Confirmeza S.A.S.

Definición de Variables

A continuación, se presentan las variables estratégicas del estudio:

Tabla 1.

Definiciones conceptuales y operacionales de las variables del estudio

Nº	Variable	Definición conceptual	Definición operacional	Dimensiones	Tipo de campo
1	Número de la colocación	Identificador único del crédito desembolsado.	Código numérico asignado por el sistema al momento del desembolso.	Único por cliente-producto.	Cualitativa nominal
2	Número de identificación deudor	Identificación oficial del cliente en el sistema financiero.	Número de documento de identidad registrado.	Único por cliente.	Cualitativa nominal
3	Número del producto	Código que representa el producto financiero contratado.	Valor numérico o alfanumérico asociado al producto.	Relacionado con el tipo de producto.	Cualitativa nominal
4	Producto	Tipo de producto financiero contratado.	Se registra como categoría según la denominación del producto en la entidad.	Crédito de consumo, tarjeta, libranza, etc.	Cualitativa nominal
5	Fecha apertura del préstamo	Fecha en la que se originó el préstamo.	Campo tipo fecha registrado en el sistema.	Día/Mes/Año.	Fecha
6	Valor original del préstamo	Monto inicial otorgado en el crédito.	Valor monetario registrado al momento del desembolso.	Valor positivo en moneda.	Numérico (decimal)
7	Saldo del préstamo	Monto pendiente por pagar a la fecha.	Valor monetario restante del préstamo.	Valor positivo en moneda.	Numérico (decimal)
8	Días de mora	Número de días de atraso desde la fecha límite de pago.	Diferencia en días entre fecha de vencimiento y fecha actual o de pago.	Entero ≥ 0.	Numérico (entero)

9	Nro plan pagos vigente	Número de acuerdos de pago activos.	Número entero registrado por el sistema.	Entero ≥ 0.	Numérico (entero)
10	Plazo en días	Duración total del crédito en días calendario.	Diferencia entre fecha de apertura y fecha de vencimiento.	Valor numérico positivo.	Numérico (entero)
11	Número de pagos pactados	Total de cuotas acordadas.	Entero registrado en el cronograma de amortización.	Entero ≥ 1.	Numérico (entero)
12	Capital pendiente por facturar	Monto de capital aún no facturado.	Valor restante no cobrado del capital.	Valor positivo o cero.	Numérico (decimal)
13	Valor cuota mes	Monto mensual pactado del crédito.	Valor monetario correspondiente a la cuota mensual.	Valor positivo en moneda.	Numérico (decimal)
14	Ingresos	Ingreso mensual reportado por el cliente.	Monto monetario mensual declarado.	Valor positivo en moneda.	Numérico (decimal)
15	Sexo	Género del cliente.	Valor categórico (M/F).	Masculino, Femenino.	Cualitativa nominal (binaria)
16	Estado civil	Situación conyugal del cliente.	Categorizado como soltero, casado, unión libre, etc.	Soltero, Casado, Unión libre, Viudo, etc.	Cualitativa nominal
17	Edad del cliente	Edad del deudor en años.	Se calcula desde la fecha de nacimiento.	Entero ≥ 18.	Numérico (entero)
18	Nº de personas a cargo	Personas que dependen del cliente.	Número entero declarado.	Entero ≥ 0.	Numérico (entero)
19	Nivel de estudios	Máximo nivel educativo alcanzado.	Clasificación ordinal del nivel académico.	Primaria, Secundaria, Técnico, Universitario, Posgrado.	Cualitativa ordinal
20	Acierta	Score de riesgo crediticio del cliente.	Puntaje numérico (100 a 1000).	Entero en rangos establecidos.	Numérico (entero)
21	Tipo de vivienda	Forma de tenencia de la vivienda.	Clasificación nominal: propia, arriendo, etc.	Propia, Arriendo, Familiar, Cedida, etc.	Cualitativa nominal

22	Antigüedad laboral	Tiempo en el empleo actual.	Valor numérico en años o meses.	Entero o decimal ≥ 0.	Numérico (entero o decimal)

Nota. Esta tabla resume las definiciones conceptuales y operacionales de las variables utilizadas en el estudio, así como sus dimensiones y el tipo de campo según su naturaleza estadística.

A continuación, se presentan las siguientes entradas siendo estas complementarias y/o transversales al análisis del estudio:

Tabla 2.

Definiciones conceptuales y operacionales de las entradas del estudio

Nº	Variable	Definición conceptual	Definición operacional	Tipo de campo
23	Número de la Agencia	Identificador del punto de originación o trámite del crédito	Código único de agencia	Geográfica
24	Plazo Unidad Tiempo	Unidad temporal usada para definir el plazo (días, meses, años)	Categoría que acompaña al campo plazo	Tiempo
25	Nro de Utilizaciones	Cantidad de veces que el cliente ha usado una línea de crédito	Número total de desembolsos sobre una misma línea	Historial de uso
26	Calificación de Cartera	Estado crediticio según regulación financiera	A, B, C, D, E según la mora o riesgo	Riesgo
27	Fecha Cuota Más Antigua Pendiente de Pago	Fecha de la cuota vencida más antigua sin pagar	Fecha registrada en sistema de cartera	Mora
28	Fecha Próximo Pago	Fecha en la que se espera el próximo abono	Fecha registrada en plan de pagos	Tiempo / Calendario
29	Fecha Efectiva Último Pago	Fecha del último pago efectivamente realizado	Fecha registrada por sistema de recaudo	Pagos
30	Valor Último Pago	Monto del último pago realizado por el cliente	Valor registrado por sistema	Pagos
31	Valor Tasa	Porcentaje de interés aplicado al préstamo	Tasa efectiva anual o mensual	Financiera
32	Valor Mínimo de Pago	Monto mínimo exigido por cuota	Según plan de amortización	Pagos
33	Día Vencimiento Cuota	Día del mes en que vence la cuota	Número entre 1 y 31	Tiempo / Calendario

34	Número de Cuotas Vencidas	Cuotas vencidas sin pago a la fecha	Conteo de cuotas vencidas	Mora
35	Número de Cuotas Pagadas	Total de cuotas canceladas	Conteo de pagos realizados	Historial de pagos
36	Número de Cuotas Restantes	Cuotas que aún faltan por pagar	Total pactado - cuotas pagadas	Cartera activa
37	Fecha de Desembolso	Fecha en la que se desembolsó el crédito	Fecha registrada del desembolso inicial	Tiempo
38	Número de Solicitud Titular	Código interno del proceso de solicitud	Identificador único	Procesos
39	Perfil del Titular	Perfil crediticio o comercial del solicitante	Clasificación interna (Ej: Estándar, Premium, etc.)	Segmentación
40	Ciudad Vitrina	Ciudad de la agencia o vitrina	Localización geográfica	Geográfica
41	Marca	Marca del vehículo adquirido	Ej: KIA, Chevrolet	Vehículo
42	Línea	Línea o submodelo del vehículo	Ej: Picanto, Spark, etc.	Vehículo
43	Clase	Tipo de carrocería o clasificación técnica	Ej: Sedán, SUV, Pickup	Vehículo
44	Modelo	Año del modelo del vehículo	Año del vehículo	Vehículo
45	Servicio	Uso del vehículo	Público, Particular, Oficial	Vehículo
46	Valor Comercial	Valor estimado en el mercado del vehículo	Precio referencial o avaluado	Vehículo / Garantía
47	Fecha de Nacimiento Titular	Fecha de nacimiento del cliente	Usada para calcular edad	Demográfica
48	Valor Cuotas Vencidas	Monto total en mora	Suma de cuotas vencidas	Mora
49	Valor Capital Cuotas Vencidas	Valor en mora correspondiente solo a capital	Subcomponente de las cuotas vencidas	Mora
50	Valor Solicitado	Monto solicitado por el cliente inicialmente	Valor consignado en solicitud	Financiera
51	Asesor de Desembolso	Persona responsable del desembolso del crédito	Nombre o código del asesor	Operativa
52	Asesor de Ventas	Comercial que gestionó la venta	Nombre o código	Operativa / Comercial
53	Tasa de Mora	Tasa de interés por atraso en el pago	Porcentaje aplicado sobre saldo vencido	Financiera
54	Tipo de Solicitud	Categoría de solicitud (nuevo, usado, refinanciación, etc.)	Clasificación de solicitud	Producto
55	Tipo de ID	Tipo de identificación del cliente	Cédula, pasaporte, NIT, etc.	Demográfica
56	Ocupación	Profesión o actividad económica	Cargo u ocupación registrada en la solicitud	Socioeconómica
57	Tipo de Vehículo	Clasificación del vehículo según uso o estructura	Familiar, Comercial, etc.	Vehículo

58	Tipo de Servicio	Finalidad del vehículo	Público o Particular	Vehículo

Nota. Esta tabla resume las definiciones conceptuales y operacionales de las variables utilizadas en el estudio, así como sus dimensiones y el tipo de campo según su naturaleza estadística.

Población y Muestra

Base datos cartera: la población objetivo del estudio está compuesta por los clientes de la empresa que han adquirido créditos para la financiación de vehículos y pólizas, dichos créditos están abarcando un período desde el año 2012 hasta el 2025. La recopilación de los datos se obtiene mediante colaboración de la empresa Confirmeza e incluye información de montos, fechas, línea de crédito, edad de mora, entre otros.

Selección de métodos o instrumentos para recolección de información

Dado el enfoque del presente estudio, que combina un análisis descriptivo y correlacional de los datos de cartera crediticia, la recolección de la información no se basa en instrumentos tradicionales como encuestas o entrevistas, sino en la extracción y procesamiento de datos estructurados provenientes del sistema ERP de la compañía, denominado **SIIIF**, alojado en una infraestructura basada en **AS400**. Este sistema concentra información histórica de la gestión crediticia de los clientes, permitiendo acceder a un conjunto de variables clave relacionadas con los perfiles de crédito, estado de la cartera, características de los clientes, antigüedad, comportamiento de pago, entre otros.

El acceso a los datos se realiza mediante una conexión al **servidor FTP**, desde donde se descarga periódicamente un consolidado de cartera. Esta base de datos ha sido depurada y tratada para efectos del análisis exploratorio, preprocesamiento y posterior modelado, todo lo cual se ha documentado de forma transparente a través de notebooks disponibles en el repositorio de GitHub del proyecto.

A diferencia de estudios que requieren la construcción de instrumentos primarios como entrevistas estructuradas o formatos de observación directa, en este caso se ha optado por utilizar instrumentos digitales ya existentes, es decir, los sistemas de información que recogen y estructuran los datos operativos en tiempo real. Estos sistemas han sido diseñados previamente para fines administrativos y de control, lo cual garantiza coherencia en las variables medidas, consistencia en los datos, y precisión temporal**.**

No obstante, para asegurar la calidad y fiabilidad de los datos, se ha complementado el proceso con la construcción de un protocolo de extracción, transformación y carga de la información (**ETL**), que garantiza que los datos utilizados en el análisis cumplan con los criterios de integridad, consistencia y validez. Este protocolo ha sido definido en Python y se encuentra documentado en los notebooks del proyecto (1_1_exploracion_inicial.ipynb y 1_2_preprocesamiento.ipynb), donde se detalla paso a paso la transformación de las variables originales, su limpieza, codificación y preparación para los modelos de segmentación y clasificación.

En términos de métodos, se han adoptado técnicas propias del análisis cuantitativo de datos, incluyendo análisis descriptivos, correlacionales y modelos de clasificación supervisada. Estas técnicas se apoyan en herramientas de **Machine Learning** como regresión logística, Random Forest y XGBoost, permitiendo evaluar la probabilidad de mora y caracterizar a los clientes según su perfil de riesgo. La selección de estas técnicas responde a la necesidad de encontrar patrones complejos que no serían fácilmente visibles con técnicas estadísticas tradicionales.

Adicionalmente, como parte del proceso de visualización e interpretación de resultados, se hace uso de **Power BI** para construir dashboards interactivos que permiten visualizar las

variables clave del estudio, segmentar la cartera y facilitar la toma de decisiones por parte del equipo de gestión financiera y de riesgo.

A continuación, se relacionan las herramientas, librerías y sistemas usados en el modelo:

Tabla 3.

Herramientas, librerías y sistemas empleados en el flujo de trabajo analítico

Herramienta / Librería / Sistema	**Tipo / Tecnología**	**Descripción funcional**	**Rol / Importancia en el flujo de trabajo**
Python	Lenguaje de programación	Lenguaje interpretado, versátil y de alto nivel. Su sintaxis sencilla y su potente ecosistema de librerías lo convierten en el núcleo del flujo analítico.	Lenguaje integrador del proyecto: permite automatizar procesos, limpiar datos, construir modelos predictivos y visualizarlos.
SIIF (ERP Confirmeza)	Sistema de información	Plataforma ERP utilizada para la gestión financiera, contable y administrativa de la empresa.	Fuente primaria de datos contables y financieros relevantes para el análisis.
AS400 (G&G)	Sistema transaccional	Sistema heredado donde se registran operaciones históricas y transacciones de negocio.	Fuente secundaria de datos operativos y registros históricos.
Servidor FTP (10.60.1.37)	Almacenamiento intermedio	Repositorio de archivos CSV y planos extraídos de fuentes como SIIF y AS400.	Punto de integración y descarga automatizada de archivos de entrada para el modelo.

Herramienta / Librería / Sistema	Tipo / Tecnología	Descripción funcional	Rol / Importancia en el flujo de trabajo
Pandas	Librería Python	Manipulación y análisis de estructuras de datos (DataFrames).	Carga, transformación, limpieza y análisis exploratorio de los datos.
NumPy	Librería Python	Cálculo numérico y manejo eficiente de arrays y matrices.	Soporte matemático para cálculos, normalización y matrices de datos.
matplotlib.pyplot / seaborn	Visualización de datos	Generación de gráficos y análisis visual.	Exploración visual del comportamiento de variables, outliers, correlaciones, etc.
difflib.get_close_matches	Utilidad de Python estándar	Búsqueda de coincidencias aproximadas entre cadenas de texto.	Normalización de nombres y mapeo entre registros similares.
LabelEncoder / MinMaxScaler	Preprocesamiento (scikit-learn)	Transformación de variables categóricas y normalización de valores numéricos.	Preparación de datos para modelos de machine learning.
StandardScaler	Preprocesamiento (scikit-learn)	Escalado estándar (media 0, varianza 1) de los datos numéricos.	Normalización de entradas antes del modelado.
SelectKBest, f_classif	Selección de características	Técnica estadística para elegir las variables más relevantes.	Reducción de dimensionalidad y mejora del rendimiento del modelo.
PCA	Reducción de dimensionalidad	Análisis de componentes principales para	Evita sobreajuste y mejora la interpretación visual.

Herramienta / Librería / Sistema	Tipo / Tecnología	Descripción funcional	Rol / Importancia en el flujo de trabajo
		sintetizar información en menos variables.	
train_test_split	Utilidad (scikit-learn)	Divide el dataset en conjuntos de entrenamiento y prueba.	Validación cruzada del modelo.
SMOTE (imblearn)	Oversampling	Técnica para balancear datasets desbalanceados generando observaciones sintéticas.	Mejora la capacidad del modelo para aprender clases minoritarias.
XGBClassifier (XGBoost)	Modelo predictivo	Algoritmo de boosting eficiente para clasificación.	Modelo base de predicción supervisada de alta precisión.
compute_class_weight	Ajuste de clases (scikit-learn)	Calcula pesos para clases desbalanceadas.	Mejora la equidad en la predicción de clases.
classification_report, confusion_matrix	Métricas (scikit-learn)	Evalúan precisión, recall, F1 y matriz de confusión.	Evaluación detallada del rendimiento del modelo.
tensorflow.keras	Framework de Deep Learning	API de alto nivel para construir redes neuronales (LSTM, Dense, Dropout, etc.).	Implementación de modelos más complejos como redes neuronales recurrentes.
EarlyStopping, ReduceLROnPlateau	Callbacks de Keras	Técnicas para prevenir overfitting y ajustar dinámicamente el learning rate.	Mejoran el entrenamiento evitando sobreajuste y acelerando la convergencia.

Herramienta / Librería / Sistema	Tipo / Tecnología	Descripción funcional	Rol / Importancia en el flujo de trabajo
joblib	Serialización de modelos	Guarda y carga modelos entrenados o transformadores (scalers, PCA, etc.).	Persistencia del modelo para reutilización en producción.

Nota. Esta tabla presenta las principales herramientas, librerías y sistemas empleados en el proceso analítico, con su funcionalidad y el rol que desempeñan en el flujo de trabajo.

o4-mini

El proceso completo está diseñado de forma modular, partiendo de la disponibilidad de la información en el sistema ERP institucional (SIIF), pasando por un servidor de publicación (FTP), hasta llegar a un entorno controlado (OneDrive) donde se consolida el dataset. Esto permite tener una base sólida, confiable y estructurada para el entrenamiento y ejecución de modelos predictivos avanzados, alineados con las necesidades de riesgo y análisis de comportamiento.

En resumen, los "instrumentos" seleccionados para esta investigación no son formularios o cuestionarios, sino un conjunto de **fuentes de datos estructurados**, métodos de recolección digital a través de sistemas existentes, y herramientas analíticas que permiten transformar esos datos en conocimiento útil. Esta elección responde a la naturaleza del problema de investigación, los objetivos propuestos y la necesidad de contar con evidencia sólida y replicable que respalde las decisiones estratégicas de la organización.

Como anexo, se incluirán capturas de los dashboards elaborados, ejemplos del código de extracción y transformación de datos, y un resumen de las variables utilizadas en el análisis.

Técnicas de análisis de datos

En el desarrollo del presente proyecto de investigación, cuyo objetivo es optimizar la gestión de cartera y evaluar el riesgo crediticio en la empresa Confirmeza S.A.S., se han dispuesto diversas técnicas de análisis de datos, seleccionadas en función del enfoque cuantitativo del estudio. Como se mencionaba anteriormente los datos fueron obtenidos desde el sistema ERP SIIIF, basado en la plataforma AS400, el cual dispone de herramientas de extracción a través de un servidor FTP. Esta infraestructura permite consolidar información histórica de los créditos otorgados y las variables asociadas al comportamiento de los clientes. Dicha información fue organizada y transformada para su posterior análisis mediante un enfoque estadístico y de aprendizaje automático.

Instrumentos de recolección y origen de datos

El instrumento principal es el *consolidado de información de cartera*, el cual agrupa variables clave como: número de cuotas, estado del crédito, días en mora, valor financiado, tipo de cliente, marca y tipo de vehículo, entre otros. Esta base de datos fue modelada y estructurada en notebooks de Jupyter disponibles en el repositorio de GitHub del proyecto Seminario_investigacion.

La estructura del repositorio GitHub fue diseñada bajo un enfoque modular, facilitando la trazabilidad y reutilización del código y los datos en cada etapa del proceso. La carpeta raíz del proyecto se denomina **Modelo ML**, dentro de la cual se organizan las siguientes subcarpetas y archivos:

Carpeta data: contiene las fuentes de datos tanto en su estado original como procesado.

- **Subcarpeta raw:** incluye las bases de datos originales sin tratamiento:

cierres_cartera_consolidado.xlsx: histórico general de cartera.

cierres_cartera_marzo.xlsx: información correspondiente al corte de marzo.

- **Subcarpeta processed:** contiene los datasets resultantes del preprocesamiento para cada modelo desarrollado:

df_preprocesado_XGBoost.csv: base lista para modelos de clasificación XGBoost.

df_preprocesado_LSTM_binario.csv: base adaptada para clasificación binaria con redes LSTM.

df_preprocesado_LSTM.csv: base multiclase para redes neuronales LSTM.

Carpeta notebooks: organiza los notebooks utilizados en las etapas exploratorias, de limpieza y transformación de datos.

- Subcarpeta 01_EDA:

1_1_exploracion_inicial_marzo.ipynb: análisis exploratorio del conjunto de marzo, incluyendo revisión de valores nulos, estadísticos descriptivos y primeras visualizaciones.

1_2_preprocesamiento_consolidado_XGBoost.ipynb: preprocesamiento del histórico de cartera para uso en XGBoost, incluyendo codificación, normalización, revisión valores nulos, etc.

1_3_preprocesamiento_LSTM_Binario.ipynb: tratamiento específico para el modelo LSTM binario, orientado a clasificación en mora/sin mora.

1_3_preprocesamiento_LSTM_multiclase.ipynb: transformación y codificación de variables para clasificación multiclase.

Carpeta 02_modelos: contiene los notebooks correspondientes al entrenamiento, evaluación y validación de los modelos desarrollados.

2_1_modelo_XGBoost.ipynb: implementación del modelo XGBoost con ajuste de hiperparámetros y validación cruzada.

2_2_modelo_LSTM_binario.ipynb: arquitectura y entrenamiento de red neuronal LSTM para clasificación binaria.

2_2_modelo_LSTM_multiclase.ipynb: modelo LSTM multiclase, configurado para detectar la entrada, salida o permanencia de mora.

modelo_xgboost_optimizado.pkl: archivo .pkl que almacena el modelo XGBoost entrenado y listo para despliegue.

parametros_modelo_completo.txt: documento de texto donde se registran los mejores hiperparámetros seleccionados durante el tuning del modelo XGBoost.

Archivos adicionales en la raíz del proyecto:

README.md: documentación general del proyecto, que describe el objetivo, metodología y estructura del repositorio.

pyproject.toml, poetry.lock: archivos de configuración del entorno reproducible gestionado con poetry, que aseguran la instalación correcta de las dependencias.

.gitignore, .gitattributes: archivos para control de versiones y formato de archivos en Git.

Esta estructura permitió un manejo sistemático y reproducible del flujo de trabajo. La trazabilidad de los datos y del código se garantizó mediante la documentación en el archivo README.md y el uso de poetry para la gestión del entorno. La elección de este enfoque no solo asegura la integridad del proceso, sino que permite su futura escalabilidad o replicación.

Técnicas de análisis utilizadas

A continuación, se describen las técnicas implementadas en el desarrollo del proyecto, agrupadas por instrumento, técnica de análisis, descripción del procedimiento y nombre del notebook donde se implementó:

Tabla 4.

Descripción de técnicas de análisis y notebooks asociados

Técnica de análisis	**Descripción**	**Notebook**
Visualización inicial en Power BI	Se construyó un dashboard para observar tendencias por región, tipo de cliente, riesgo y tipo de vehículo. Permitió un diagnóstico preliminar del comportamiento de la cartera.	Power BI (externo)
Estadística descriptiva	Se calcularon medias, medianas, desviación estándar y rango para variables numéricas, caracterizando el comportamiento de la cartera.	1_1_exploracion_inicial_marzo.ipynb
Análisis univariado y bivariado	Se exploraron variables de forma individual y combinada para identificar relaciones básicas entre características y morosidad.	1_1_exploracion_inicial_marzo.ipynb
Revisión de datos nulos, negativos y duplicados	Se identificaron valores inconsistentes, faltantes y registros duplicados.	1_1_exploracion_inicial_marzo.ipynb

Técnica de análisis	Descripción	Notebook
Análisis de variables categóricas	Se examinaron las principales variables cualitativas por frecuencia y combinación con el riesgo de cartera.	1_1_exploracion_inicial_marzo.ipynb
Generación de etiquetas de clasificación	Se construyó la variable objetivo (target) según el modelo: binario o multiclase.	1_2 y 1_3 notebooks
Normalización y escalado de variables numéricas	Se aplicó StandardScaler y normalización.	2_1_modelo_XGBoost.ipynb 2_2_modelo_LSTM_binario.ipynb 2_2_modelo_LSTM_multiclase.ipynb
Codificación categórica (One-Hot Encoding)	Las variables categóricas fueron transformadas en variables binarias mediante codificación tipo One-Hot.	1_2_preprocesamiento_consolidado_XGBoost.ipynb 1_3_preprocesamiento_LSTM_Binario.ipynb 1_3_preprocesamiento_LSTM_multiclase.ipynb
Imputación de valores faltantes	Se imputaron valores nulos mediante estrategia basada en el tipo de variable (media, moda o eliminación).	1_2_preprocesamiento_consolidado_XGBoost.ipynb 1_3_preprocesamiento_LSTM_Binario.ipynb 1_3_preprocesamiento_LSTM_multiclase.ipynb

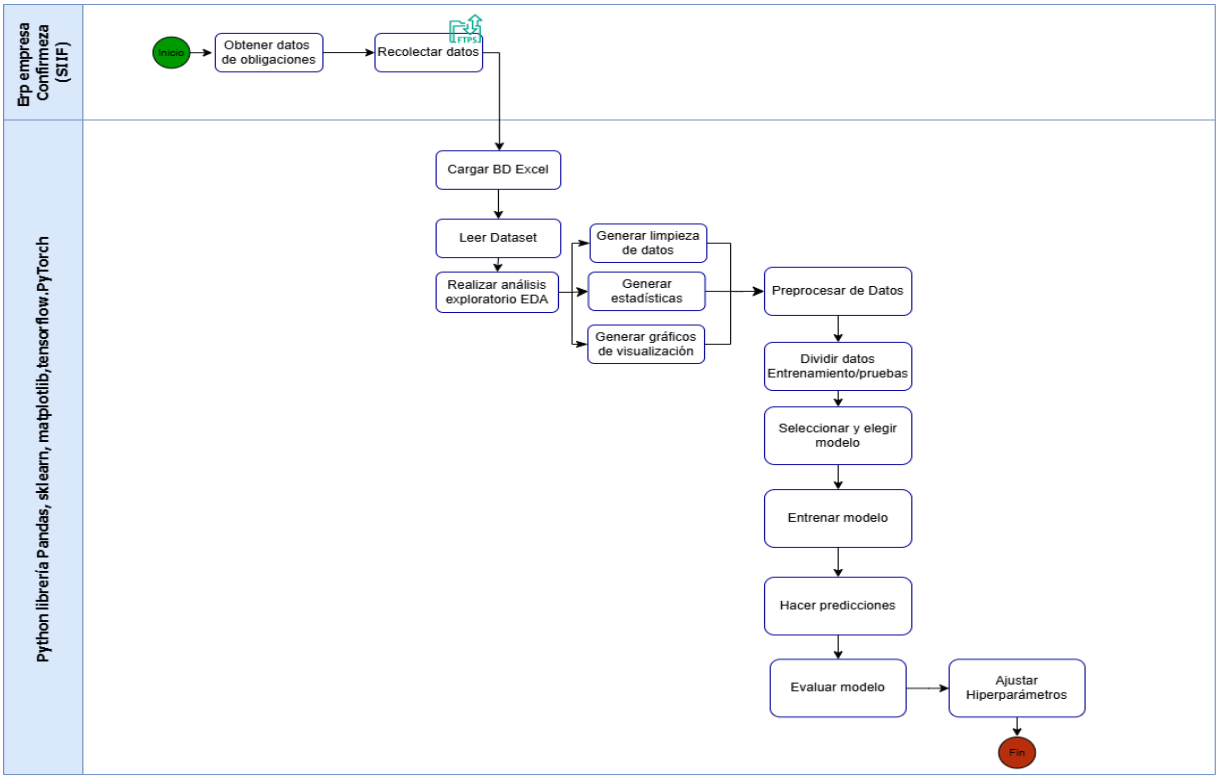
Técnica de análisis	Descripción	Notebook
Manejo de atípicos	Se identificaron y ajustaron valores extremos para evitar distorsión en el entrenamiento de modelos.	1_2_preprocesamiento_consolidado_XGBoost.ipynb 1_3_preprocesamiento_LSTM_Binario.ipynb 1_3_preprocesamiento_LSTM_multiclase.ipynb
Balanceo de clases (oversampling con SMOTE)	Se utilizó SMOTE para crear observaciones sintéticas de clases minoritarias y mejorar el rendimiento de modelos supervisados.	2_1_modelo_XGBoost.ipynb
Clasificación con XGBoost	Se implementó XGBoost con ajuste de hiperparámetros, pesos de clase, evaluación con matriz de confusión y métricas de desempeño.	2_1_modelo_XGBoost.ipynb
Modelado LSTM binario y multiclase	Se definieron secuencias temporales, arquitectura LSTM, pérdida personalizada (*focal loss*), entrenamiento y evaluación con métricas, curvas y matriz de confusión.	2_2_modelo_LSTM_binario.ipynb 2_2_modelo_LSTM_multiclase.ipynb
Ajuste de umbral de decisión	Se ajustó el umbral de clasificación para mejorar la detección de la clase minoritaria en modelos LSTM.	2_2_modelo_LSTM_binario.ipynb 2_2_modelo_LSTM_multiclase.ipynb

Nota. Esta tabla resume las principales técnicas de análisis empleadas, su descripción funcional y los cuadernos (notebooks) asociados al proceso.

A continuación, se describe el flujo del proceso seguido, que abarca desde la carga de los datos hasta la obtención de los resultados finales del modelo, incluyendo cada una de las etapas clave ejecutadas.

Figura 2

Diagrama de proceso



Nota. Fuente de elaboración propia

Justificación de las técnicas

Justificación técnica de las técnicas seleccionadas

Las técnicas empleadas fueron elegidas con base en la naturaleza del problema, el tipo de datos disponibles, los objetivos del proyecto y los criterios de eficiencia computacional y rendimiento del modelo.

El uso de **Power BI** permitió una visualización intuitiva y clara del comportamiento inicial de la cartera, lo cual fue clave para la comprensión general del problema y la identificación de patrones iniciales.

El análisis estadístico (descriptivo, univariado y bivariado) permitió explorar la estructura del dataset y evaluar la calidad de los datos antes del modelado.

El **preprocesamiento estructurado** (limpieza, codificación, imputación y escalado) garantizó que los modelos recibieran información coherente, sin sesgos derivados de registros erróneos o mal estructurados.

XGBoost fue seleccionado por ser un modelo robusto, eficiente con datasets de alta dimensionalidad, y con excelente desempeño en clasificación multiclase. Su capacidad de manejar directamente variables categóricas y su eficiencia computacional lo hicieron una elección óptima.

Los modelos **LSTM** (binario y multiclase) se utilizaron por su capacidad para aprender secuencias y dependencias temporales en los datos, aprovechando el comportamiento histórico de los clientes. Además, se integró una **función de pérdida personalizada (Focal Loss)** para mejorar el rendimiento frente a clases desbalanceadas.

Justificación técnica de las técnicas no implementadas

Aunque en el planteamiento teórico se mencionaron técnicas como Regresión Logística, Random Forest y Análisis de Componentes Principales (PCA), estas fueron descartadas por razones técnicas fundamentadas:

Regresión Logística: es un modelo paramétrico que requiere asumir relaciones lineales entre variables. La presencia de muchas variables categóricas y no linealidades complejiza su aplicación y reduce su rendimiento. Además, requeriría una reducción importante del número de variables, lo que podría implicar pérdida de información valiosa. Por ello, se optó por modelos de árbol más flexibles.

Random Forest: aunque efectivo, fue reemplazado por **XGBoost**, un modelo más moderno que ofrece mejor rendimiento, regularización y control sobre el ajuste. En pruebas internas, XGBoost mostró resultados superiores y menor tiempo de entrenamiento.

PCA (Análisis de Componentes Principales): no fue necesario implementar esta técnica de reducción de dimensionalidad, ya que los tiempos de entrenamiento fueron aceptables y no se presentaron problemas de sobreajuste ni colinealidad severa. Además, PCA habría afectado la interpretabilidad de las variables categóricas codificadas, lo cual era crucial para la toma de decisiones del área de riesgo.

Consideraciones cualitativas

Aunque el enfoque principal es cuantitativo, se incorporan algunos elementos cualitativos mediante el análisis interpretativo de los resultados y la construcción de perfiles de clientes a partir de los patrones encontrados. Esto permite definir estrategias de acción específicas para cada segmento, lo que se alinea con los objetivos estratégicos de la compañía.

Las técnicas de análisis adoptadas en este estudio no solo responden a una necesidad de evaluación objetiva de los riesgos de crédito, sino también a una estrategia de transformación digital en la gestión de cartera. El uso de modelos predictivos, herramientas de visualización y algoritmos de segmentación permitirá a la empresa mejorar sus políticas de financiamiento y reducir los niveles de morosidad, contribuyendo así a su estabilidad financiera.

Análisis y discusión de los resultados

Resultados

Exploración *inicial*

Como primera etapa del análisis, se realizó una exploración detallada del corte de cartera correspondiente al mes de marzo. Esta fase tuvo como objetivo evaluar la calidad de los datos, comprender la distribución de las variables relevantes y detectar patrones o relaciones preliminares entre las características de los clientes y el estado de mora. A continuación, se presentan los principales hallazgos:

Análisis de calidad de datos

Valores nulos: Se identificaron campos con ausencias significativas, especialmente en variables sociodemográficas y de identificación del vehículo. Las variables con mayor número de nulos fueron: PERFIL DEL TITULAR, CIUDAD VITRINA, MARCA, CLASE, FECHA DE NACIMIENTO TITULAR, VALOR SOLICITADO, entre otras.

Valores negativos: No se detectaron valores negativos en las variables numéricas, lo que indica que el registro financiero es coherente desde el punto de vista de integridad básica.

Registros duplicados: No se encontraron registros duplicados, lo que sugiere que la base de datos fue exportada sin errores de duplicación.

Distribución de variables numéricas

El análisis univariado permitió observar la estructura de las variables numéricas:

Variables como DÍAS DE MORA, NRO CUOTAS VENCIDAS y NRO CUOTAS PAGADAS presentan **distribuciones altamente sesgadas a la derecha**, lo que refleja una

alta concentración de clientes con atrasos bajos, pero también presencia de casos críticos con altos días de mora o acumulación de cuotas impagas.

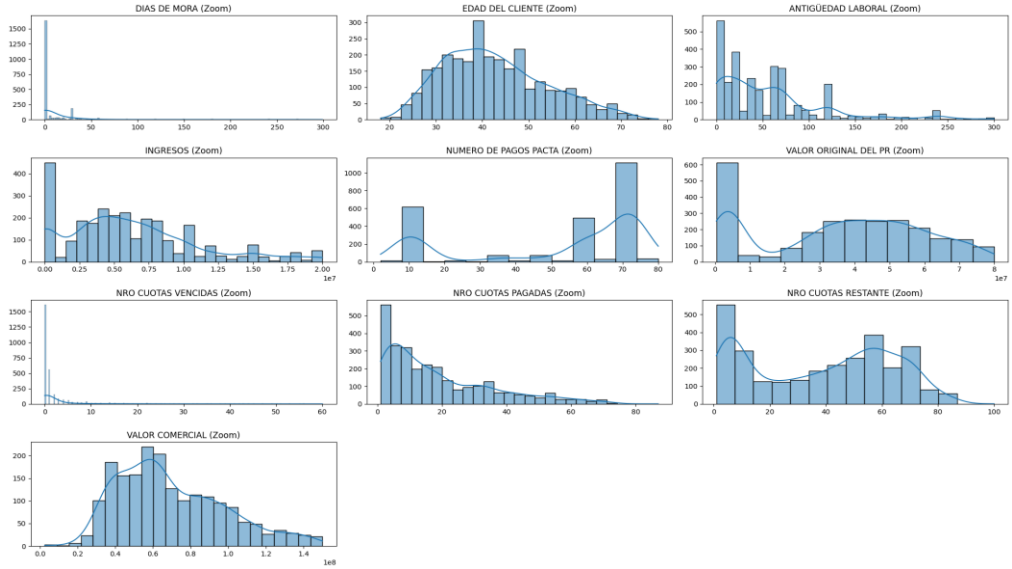
La variable EDAD DEL CLIENTE muestra una **distribución aproximadamente normal**, centrada en rangos entre los 30 y 45 años, lo que representa una población económicamente activa.

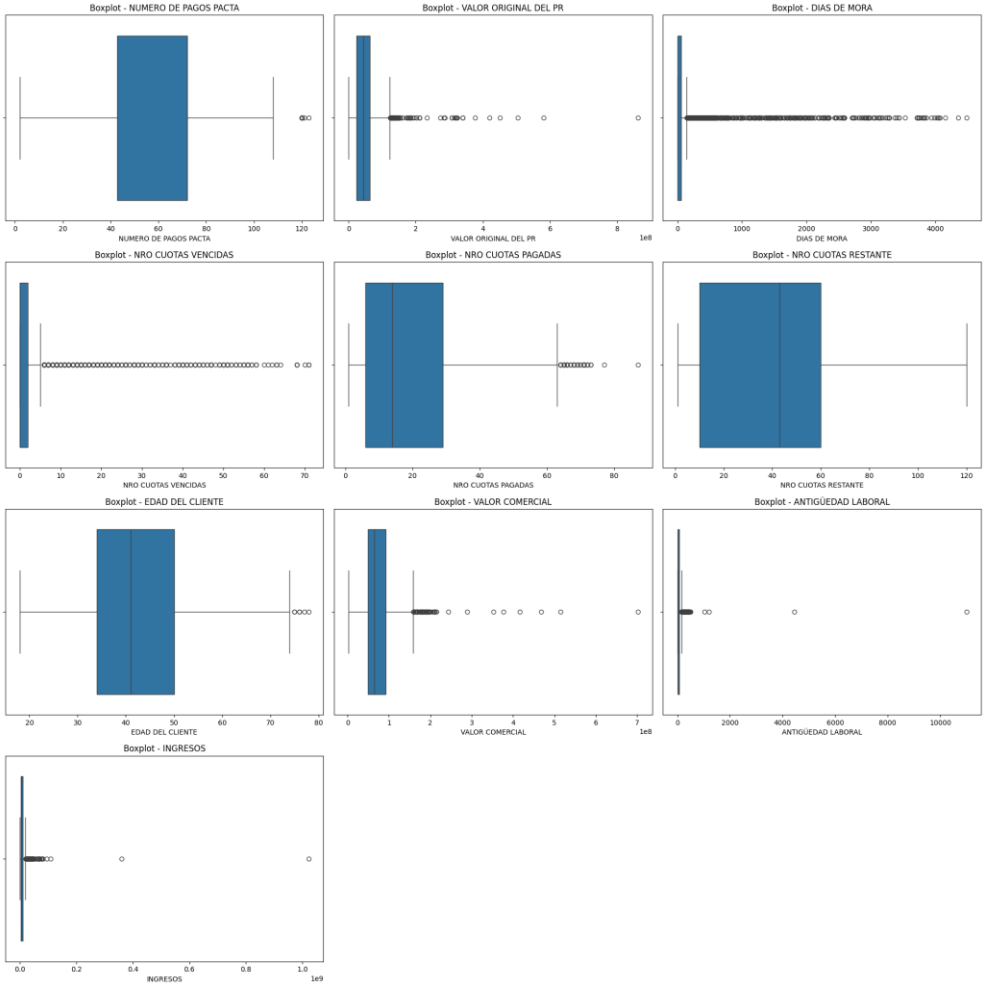
En contraste, variables como INGRESOS, VALOR COMERCIAL DEL VEHÍCULO y VALOR ORIGINAL DEL PRÉSTAMO presentan **distribuciones asimétricas**, lo cual es esperado en contextos de financiamiento automotriz donde coexisten vehículos de diferentes gamas y clientes de diversos estratos.

Mediante los **boxplots** se detectó la presencia de múltiples *outliers* en variables como VALOR COMERCIAL, VALOR ORIGINAL DEL PRÉSTAMO, INGRESOS y ANTIGÜEDAD LABORAL, indicando la necesidad de una futura imputación o transformación para evitar distorsiones en los modelos predictivos.

Figura 3

Distribuciones univariadas de variables clave en el análisis crediticio





Distribución de variables categóricas

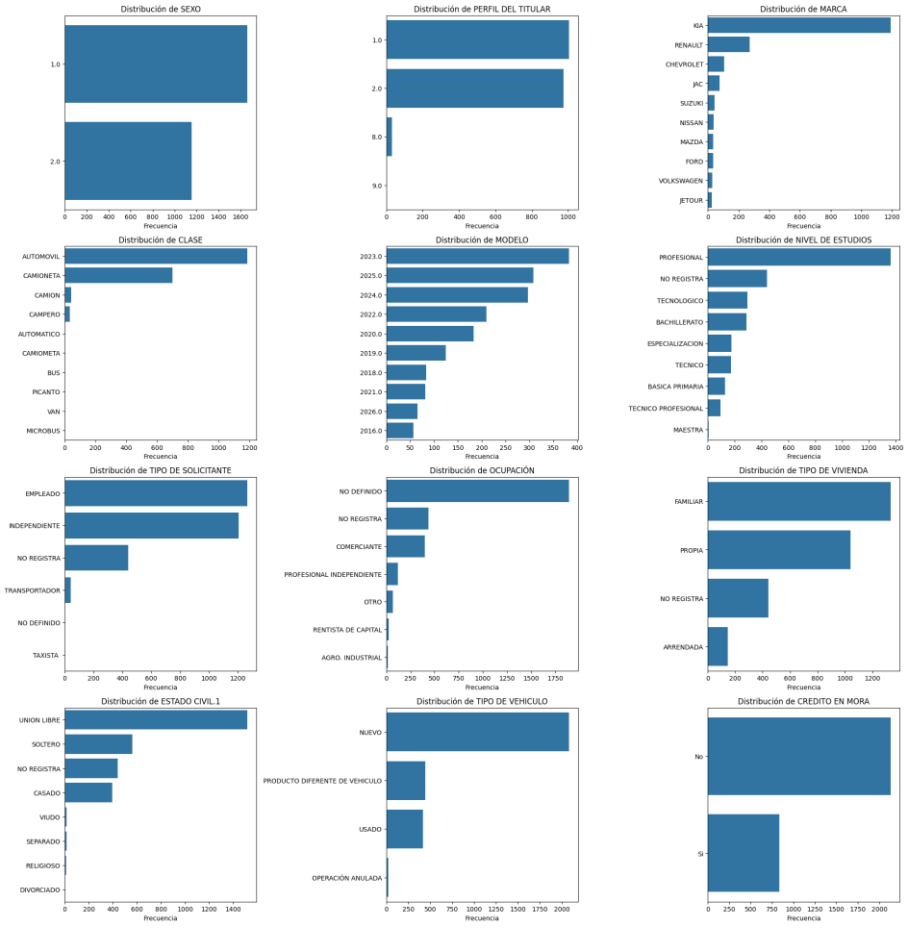
Se analizaron variables como SEXO, ESTADO CIVIL, OCUPACIÓN, TIPO DE SOLICITANTE, NIVEL DE ESTUDIOS, TIPO DE VIVIENDA, TIPO DE VEHÍCULO y MARCA.

Por ejemplo, el análisis muestra que la mayoría de los clientes son de sexo masculino, tienen vivienda familiar o propia, y laboran en condición de empleados o independientes.

La distribución de TIPO DE VEHÍCULO indica que el producto más frecuente es el vehículo usado, seguido por nuevo, lo cual es coherente con el enfoque comercial de la empresa.

Figura 4

Distribuciones de frecuencia de variables categóricas en el análisis de cartera de crédito



Análisis bivariado: mora por categorías

A través del análisis bivariado, se exploró la relación entre distintas variables categóricas y la variable objetivo CRÉDITO EN MORA. Se emplearon gráficos de barras apiladas que muestran la proporción de clientes en mora (rojo) y sin mora (verde) para cada categoría. Los principales hallazgos y su relevancia estratégica son los siguientes:

- Campos no registrados: una señal de alerta institucional

En múltiples variables como OCUPACIÓN, ESTADO CIVIL, TIPO DE VIVIENDA, TIPO DE SOLICITANTE y PERFIL DEL TITULAR, se observa que las categorías "No registra" o "No definido" presentan niveles significativamente altos de morosidad. Esto implica que los registros incompletos o no diligenciados correctamente no solo representan una falla operativa, sino un riesgo crediticio directo para la empresa.

- Perfil del titular y tipo de solicitante

El análisis muestra que ciertos perfiles presentan tasas de mora más elevadas, en particular aquellos con designación 3. Asimismo, solicitantes con categorías como "No registra" o "Taxista" tienden a presentar mayor probabilidad de incumplimiento.

- Nivel educativo y ocupación

Aunque el nivel educativo muestra morosidad distribuida de forma relativamente homogénea, se observa que clientes con **"**No registra" o "Básica primaria" tienden a tener una tasa de mora ligeramente mayor. En cuanto a ocupación, perfiles como "No registra", "Agroindustrial" y "Transportador" muestran mayor concentración de riesgo.

- Tipo de vivienda y estado civil

Los clientes que habitan en vivienda arrendada muestran mayor tendencia a la mora, lo cual es lógico dada la carga financiera adicional. De forma similar, estados civiles como "Separado" y "No registra" presentan mayor incidencia de incumplimiento.

- Tipo de vehículo y año del modelo

Se evidencia una mayor morosidad en créditos asociados a vehículos *más antiguos (modelos anteriores a 2018)*. Estos casos reflejan condiciones más riesgosas, posiblemente asociadas a clientes de perfil vulnerable o vehículos con menor valor de reventa.

- **Puntaje de riesgo ACIERTA**

El análisis de la variable ACIERTA_BIN (puntaje crediticio) arroja hallazgos relevantes:

- Clientes con puntaje superior a 900 presentan tasas muy bajas de mora, como se esperaría.
- Clientes con puntaje entre 600 y 800 tienen una tasa intermedia de mora.
- Curiosamente, algunos clientes con puntaje muy bajo no presentan mora significativa, lo cual indica que el puntaje por sí solo no es suficiente para clasificar el riesgo.
- Ingresos (rango binarizado) y morosidad

La distribución de morosidad según rangos de ingresos (INGRESOS_BIN) muestra que la proporción de clientes en mora se mantiene relativamente constante a lo largo de casi todos los rangos. Sin embargo, se evidencia un ligero aumento de morosidad en los clientes con ingresos entre 1 y 4 millones de pesos, donde la franja roja representa alrededor del 25-30% del total. En los extremos clientes con ingresos menores a 1 millón y mayores a 20 millones la morosidad no difiere sustancialmente del promedio general, lo que sugiere que el ingreso, por sí solo, no es un predictor fuerte de incumplimiento.

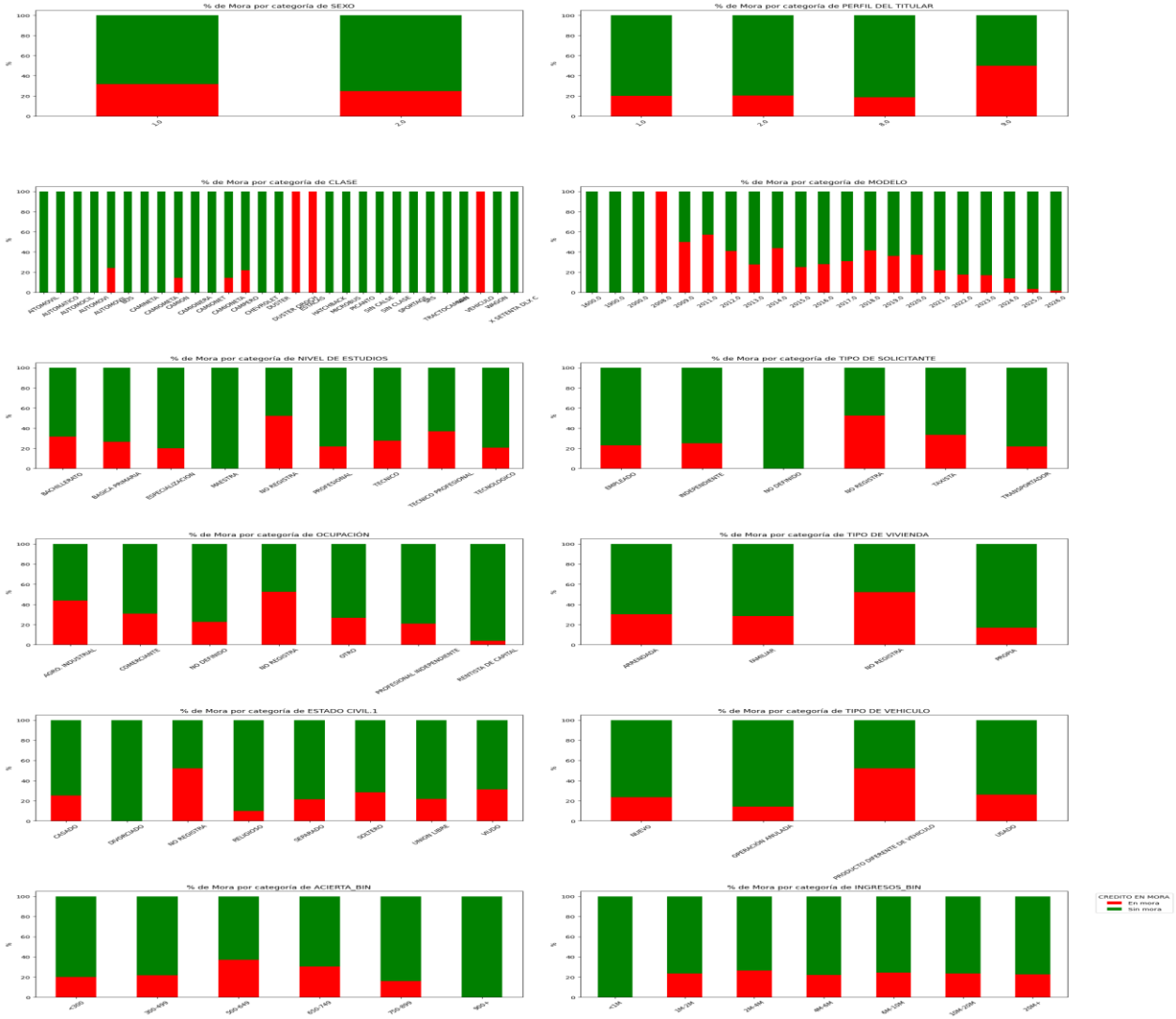
Este comportamiento puede deberse a que los ingresos más altos no necesariamente implican mejores hábitos de pago, y los ingresos más bajos podrían estar compensados por estructuras familiares compartidas, garantías adicionales o montos de crédito más pequeños.

Además, los clientes con ingresos medios pueden estar más expuestos al sobreendeudamiento, al acceder a múltiples productos financieros simultáneamente.

Este análisis bivariado permitió validar la relación preliminar entre las características del cliente y la morosidad, lo cual respalda la construcción de modelos predictivos más avanzados basados en dichas variables.

Figura 5

Porcentaje de clientes en mora y sin mora por categoría de variables clave



Modelos predictivos

Arbol de clasificación XGBoost multiclase

Como parte del objetivo de anticipar el comportamiento de los clientes frente a la mora, se implementó un modelo de clasificación multiclase utilizando el algoritmo **XGBoost**, el cual fue ajustado manualmente mediante una combinación de tuning de hiperparámetros y balanceo de clases. El modelo final permitió clasificar a los clientes en cuatro categorías:

0: Al día

1: Entra en mora

2: Sale de mora

3: Se mantiene en mora

Resultados de la matriz de confusión

La matriz de confusión muestra que:

El modelo predice correctamente a 5.566 clientes al día, con pocos falsos positivos (95 casos clasificados erróneamente como "entra en mora").

Para la clase 1: Entra en mora, solo 90 casos fueron correctamente identificados, mientras que 173 fueron erróneamente clasificados como “al día”.

En la clase 2: Sale de mora, la predicción es más equilibrada: 93 casos fueron correctamente identificados, aunque 96 fueron clasificados como "se mantiene en mora".

La clase 3: Se mantiene en mora se predice con alta precisión: 1.914 casos correctamente identificados, y solo 72 mal clasificados como “sale de mora”.

Este comportamiento refleja que el modelo tiene **una fuerte capacidad para identificar los extremos** (clientes al día o en mora persistente), pero aún presenta desafíos para detectar correctamente los casos **transicionales** (entra/sale de mora).

Figura 6

Matriz de confusión modelo XGboost

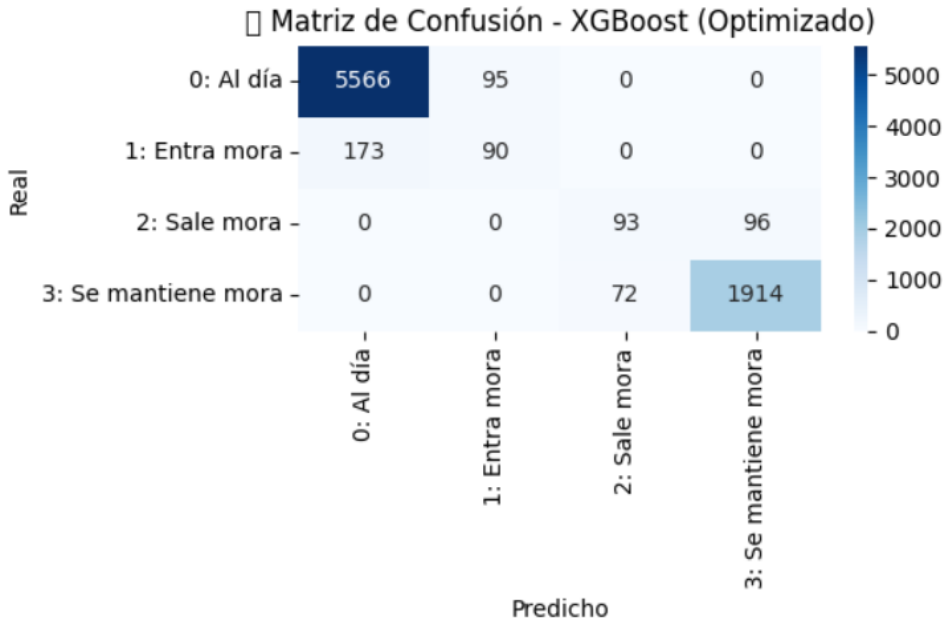


Tabla 5

Reporte de métricas

Clase	Precisión	Recall	F1-score
0: Al día	0.97	0.97	0.97
1: Entra en mora	0.41	0.38	0.39
2: Sale de mora	0.54	0.53	0.54
3: Se mantiene mora	0.96	0.96	0.96

Precisión global del modelo (accuracy): 94%

Macro promedio del F1-score: 0.71

Promedio ponderado del F1-score: 0.94

Estas métricas muestran que el modelo es **muy confiable para predecir las clases mayoritarias** (clientes estables), pero menos preciso en clases intermedias. Esto es común en escenarios de desbalance natural en los datos.

Interpretación estratégica de los resultados

Fortalezas del modelo: El modelo XGBoost es altamente eficaz para predecir clientes con comportamiento estable (al día o morosos persistentes), lo que es clave para implementar campañas de recompensa y seguimiento especializado.

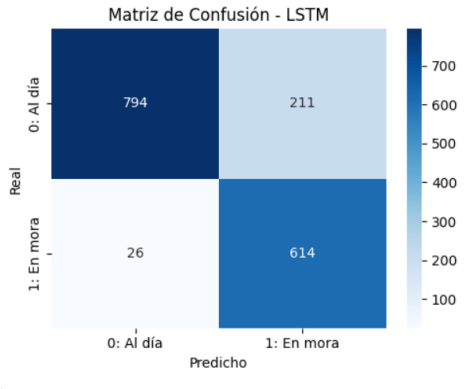
Limitaciones actuales: El rendimiento en clases de transición (entra/sale de mora) es más bajo. Esto sugiere que estas transiciones están condicionadas por variables que podrían no estar suficientemente representadas o tratadas en el dataset.

Red neuronal LSTM binaria

Como segunda estrategia de predicción del comportamiento crediticio, se implementó un modelo de clasificación binaria utilizando una red neuronal LSTM (Long Short-Term Memory). El objetivo fue identificar si un cliente se encontrará al día (clase 0) o en mora (clase 1) en el próximo período.

Figura 7

Resultados de la matriz de confusión



La matriz de confusión muestra un desempeño sólido del modelo:

El modelo identificó correctamente 794 clientes al día, pero clasificó erróneamente a 211 como “en mora”.

Identificó con gran precisión a 614 clientes efectivamente en mora, con apenas 26 falsos negativos.

Esto indica que el modelo es especialmente eficaz en la detección de clientes con riesgo real de incumplimiento, lo cual es altamente valioso para una política preventiva.

Tabla 6

Reporte de métricas

Clase	Precisión	Recall	F1-score	Soporte
Al día	0.97	0.79	0.87	1005
En mora	0.74	0.96	0.84	640

Precisión general (accuracy): 86%

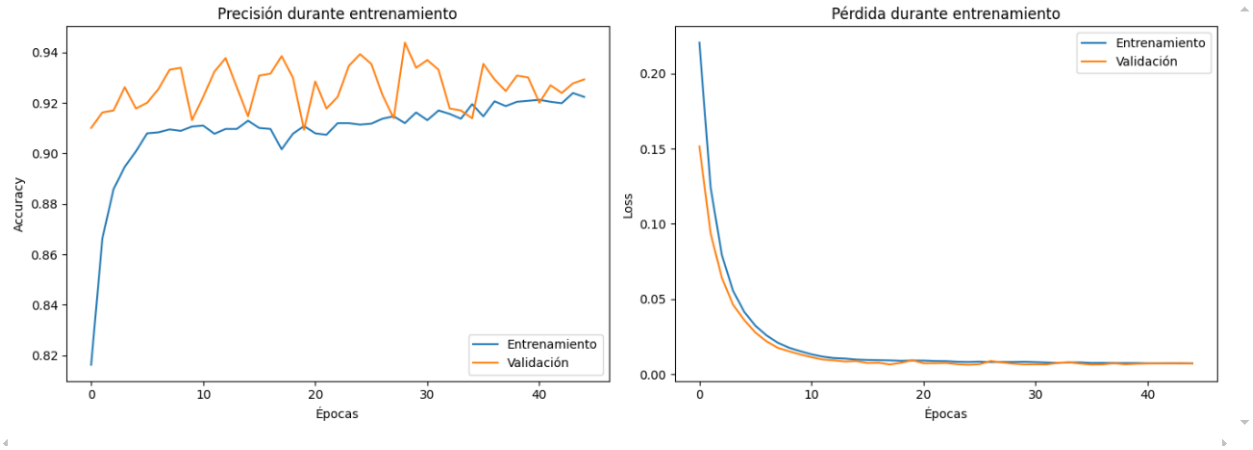
Macro F1-score: 0.85

Promedio ponderado (weighted): 0.86

Estos resultados reflejan un modelo equilibrado y con excelente recall en la clase minoritaria (en mora), lo cual es fundamental en escenarios de riesgo crediticio donde minimizar los falsos negativos es prioritario.

Figura 8

Curvas de entrenamiento



La precisión de entrenamiento y validación converge progresivamente hasta estabilizarse cerca del 92-94%, sin indicios fuertes de sobreajuste.

La curva de pérdida (loss) disminuye rápidamente en las primeras épocas y se mantiene en valores bajos, tanto en entrenamiento como validación.

Este comportamiento sugiere que el modelo fue entrenado de forma adecuada, y que la arquitectura y los hiperparámetros definidos fueron correctos.

Interpretación estratégica

El modelo LSTM binario tiene un rendimiento notablemente superior para anticipar la morosidad real (clase 1), con un recall del 96%, lo cual representa una herramienta potente para identificar a tiempo a los clientes con mayor probabilidad de incumplir.

Aunque la precisión para la clase “al día” es alta, su recall es menor (79%), lo que implica que algunos clientes que no caerán en mora podrían ser tratados como si lo hicieran (falsos positivos). Este efecto podría ser gestionado desde las áreas comerciales para evitar decisiones demasiado restrictivas.

Red neuronal LSTM multiclase

El modelo LSTM multiclase fue entrenado para clasificar a los clientes en una de las siguientes cuatro categorías de comportamiento crediticio:

0: Al día

1: Entra en mora

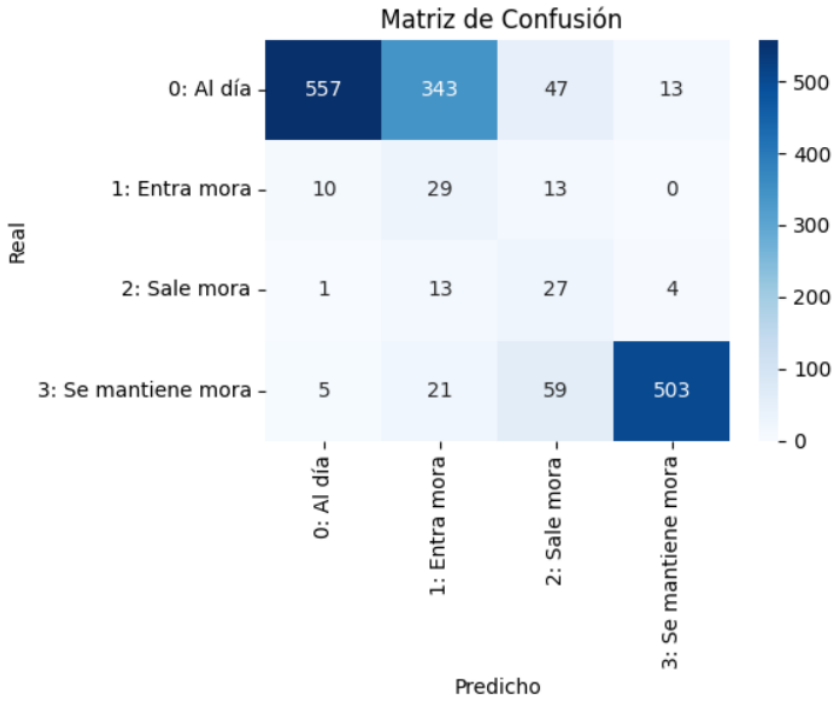
2: Sale de mora

3: Se mantiene en mora

Este modelo busca incorporar la dimensión secuencial de los pagos para predecir con mayor precisión el estado futuro del cliente, aprovechando la arquitectura de redes neuronales recurrentes.

Figura 9

Resultados de la matriz de confusión



La matriz de confusión refleja una alta capacidad del modelo para identificar correctamente las clases extremas, pero un bajo rendimiento para las clases transicionales:

557 clientes al día fueron correctamente clasificados, pero 343 fueron predichos incorrectamente como si fueran a entrar en mora, lo que indica una alta tasa de falsos positivos en esta clase.

Para la clase 1: Entra en mora, se identificaron 29 casos correctamente, mientras que 10 fueron confundidos como “al día” y 13 como “sale de mora”.

La clase 2: Sale de mora tuvo un desempeño moderado, con 27 casos correctos y 18 mal clasificados.

La clase 3: Se mantiene en mora fue la mejor predicha por el modelo, con 503 aciertos de 588 observaciones, y una mínima confusión con otras clases.

Tabla 8

Reporte de métricas

Clase	Precisión	Recall	F1-score
0: Al día	0.97	0.58	0.73
1: Entra en mora	0.07	0.56	0.13
2: Sale de mora	0.18	0.60	0.28
3: Se mantiene mora	0.97	0.86	0.91

Precisión general (accuracy): 68%

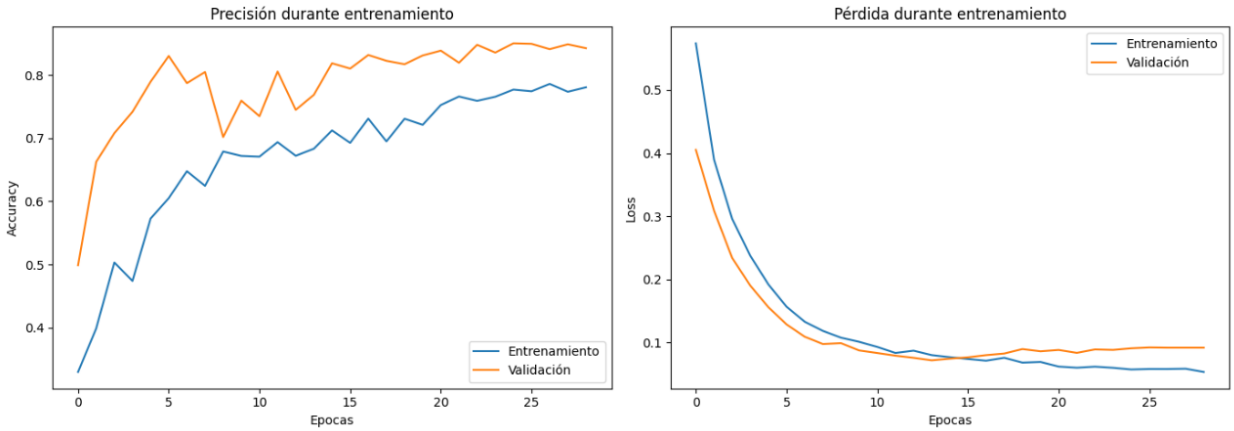
Macro F1-score: 0.51

Weighted F1-score: 0.76

Estas métricas indican que el modelo es eficaz en detectar comportamientos extremos, pero tiene dificultades claras para predecir correctamente las clases intermedias, especialmente la clase "Entra en mora", con un F1 de apenas 0.13.

Figura 10

Curvas de entrenamiento



La curva de **precisión** muestra una mejora sostenida tanto en entrenamiento como validación, con un comportamiento estable alrededor del 80-85% hacia las últimas épocas.

La **pérdida (loss)** se reduce de forma progresiva en ambas fases, sin evidencia de sobreajuste significativo.

Estas curvas respaldan que el modelo fue entrenado correctamente, aunque el desequilibrio de clases y la dificultad inherente de las clases transicionales afectan el rendimiento final.

Interpretación estratégica

El modelo LSTM multiclase es altamente fiable para predecir clientes que se mantienen en mora, lo que permite priorizar acciones de cobranza intensiva o legal.

Tiene un desempeño aceptable para identificar clientes que permanecen al día, aunque su alta tasa de falsos positivos podría generar acciones innecesarias sobre clientes que en realidad no representan un riesgo inminente.

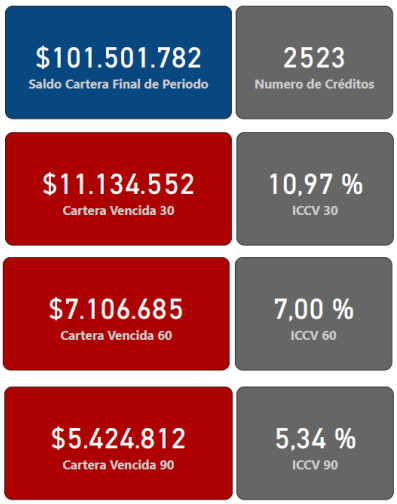
El bajo rendimiento en las clases “Entra en mora” y “Sale de mora” indica que los factores que determinan estas transiciones podrían requerir más profundidad histórica, variables externas (por ejemplo, comportamiento anterior en otros productos) o un mayor volumen de datos para entrenamiento.

Análisis Actual Power BI Segmentación de Cartera de clientes

En cuanto a la exposición total y de calidad de la cartera se evidencia un saldo en el último mes por valor de $101.501.782 (millones) para un total de 2523 créditos.

Figura 11

Indicadores clave de la cartera de crédito al cierre del periodo



Correspondiendo así que el 23.31% del portafolio presenta mora (suma de vencidos 30+), con un deterioro progresivo (10.97% - 5.34% a medida que aumenta la antigüedad del atraso). Esto sugiere problemas en cobranza temprana.

Figura 12

Comportamiento de cartera por ciudad

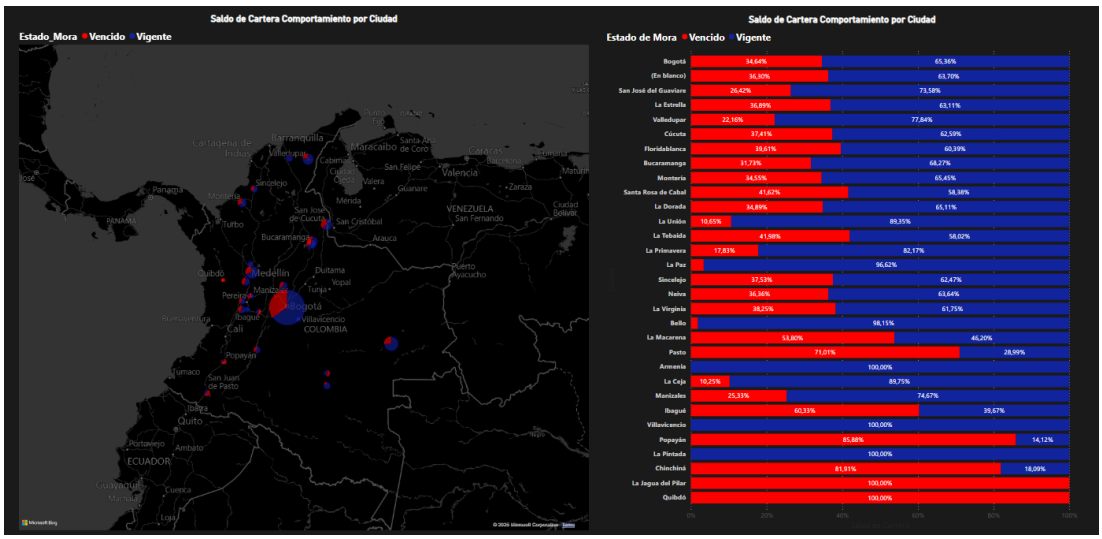


Figura 13

Número de préstamos y saldo de cartera por rangos de días de mora y vitrinas

Cartera con Castigos por Vitrina

Rangos en Días de Mora	1 a 30 días		121 a 150 días		151 a 180 días		181 días o más		31 a 60 días		61 a 90 días		91 a 120 días		Vigente		Total	
Fiki	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera	N° de Prestamos	Saldo en Cartera
⊞ METROKIA	158	$6.652.114.538	2	$116.039.619	7	$313.103.064	30	$912.781.982	33	$1.397.103.122	14	$551.666.502	8	$340.775.176	702	$30.552.026.787	954	$40.835.610.790
⊞ OTROS	141	$6.011.045.015	11	$570.283.151	7	$278.651.413	26	$962.328.091	39	$1.637.521.349	13	$506.996.907	5	$135.617.262	401	$19.186.450.747	643	$29.288.893.935
⊞ GRECCO	43	$1.732.228.614	0		1	$31.368.761	3	$197.715.599	5	$232.307.448	3	$87.270.827	4	$137.581.560	173	$8.968.941.575	232	$11.387.414.384
⊞ OFICINA PRINCIPAL	72	$1.798.050.275	4	$7.151.073	11	$108.456.280	60	$521.265.436	23	$86.971.120	13	$280.046.983	3	$79.148.958	351	$8.301.339.483	537	$11.182.429.608
⊞ AUTOCOM	29	$1.584.002.431	2	$92.971.489	1	$94.316	12	$532.623.000	10	$673.963.518	5	$255.891.593	2	$86.856.246	96	$5.581.030.974	157	$8.807.433.567
Total	443	$17.777.440.873	19	$786.445.332	27	$731.673.834	131	$3.126.714.108	110	$4.027.866.557	48	$1.681.872.812	22	$779.979.202	1723	$72.589.789.566	2523	$101.501.782.284

Recomendaciones de tipo estratégico

Acciones Prioritarias

Para METROKIA y OTROS (66.5% del riesgo):

- Implementar contacto preventivo a los 15 días de mora
- Asignar equipos especializados de cobranza

Para OFICINA PRINCIPAL:

- Auditoría urgente de garantías físicas
- Procesos judiciales acelerados para mora crítica

Para GRECCO y AUTOCOM:

- Replicar sus políticas crediticias exitosas en otras entidades
- Mantener monitoreo continuo

La cartera muestra:

- Un alto riesgo concentrado en METROKIA y OTROS
- Una oportunidad de mejora replicando modelos de GRECCO/AUTOCOM
- Se muestra una necesidad urgente de acción en OFICINA PRINCIPAL

- Actuar en los primeros 30 días de mora podría prevenir que una cartera en rangos entre los $6.000” – $7.000” COP migren a categorías críticas.

Comparativo por segmentación de variables principales de cartera vencida:

Figura 14

Tipo de Cliente en cartera vencida

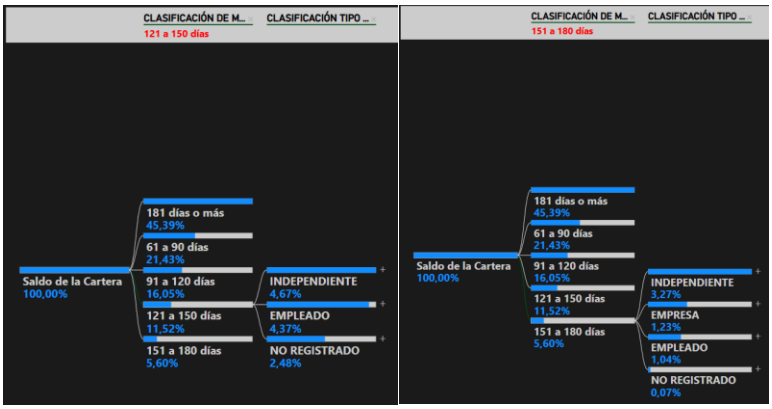


Figura 15

Saldo de cartera por tipo de cliente

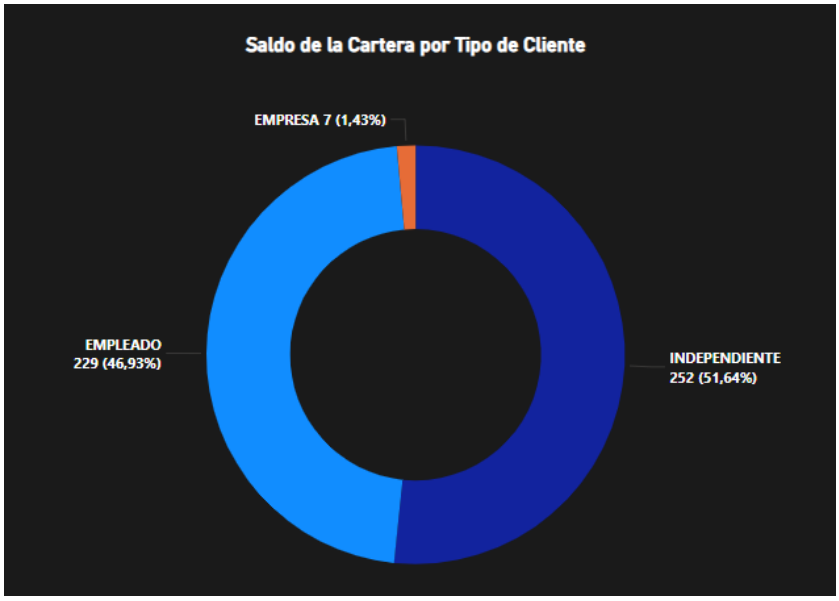


Tabla 14

Porcentaje de mora por rango de días de atraso y tipo de solicitante

Tabla 9

Mora por perfil laboral

Rango de Mora	Independiente	Empleado	Empresa	No Registrado	Tendencia Observada
61-90 días	10.46%	10.96%	-	0.01%	Empleados = Independientes
91-120 días	12.22%	3.79%	-	0.05%	Independientes 3.2x empleados
121-150 días	4.67%	4.37%	-	2.48%	Similar riesgo independiente/empleado
151-180 días	3.27%	1.04%	1.23%	0.07%	Independientes dominan la mora larga
181+ días	27.95%	11.40%	1.15%	4.89%	Independientes 2.5x empleados

Patrones y Tendencias Claves

Dominio de Independientes:

- Representan 27.95% de la mora extrema (181+ días)
- Mantienen liderazgo en todos los rangos excepto 61-90 días
- En mora media (91-120 días) triplican riesgo vs empleados

Comportamiento de Empleados:

- Pico en mora temprana (10.96% en 61-90 días)
- Mejoran significativamente en mora avanzada (solo 1.04% en 151-180 días)

Empresas vs No Registrados:

- Empresas aparecen solo en mora extrema (1.15%-1.23%)
- No registrados muestran patrón errático (4.89% en 181+ vs 0.01% en 61-90)

Recomendaciones Estratégicas

Para Independientes:

- Revisar flujos de efectivo declarados
- Exigir mayores garantías para nuevos créditos
- Programa especial de refinanciamiento

Para Empleados:

- Foco en prevención temprana (primeros 60 días)
- Descuento por nómina automático

Como conclusión se identifica que los independientes concentran el 58% del riesgo extremo (181+ días), mientras los empleados muestran mayor vulnerabilidad en etapas tempranas. Se requiere estrategias diferenciadas por tipo de cliente, intervención temprana para empleados y control estricto de independientes con mora >90 días.

Figura 16

Rangos de edad del cliente en cartera vencida

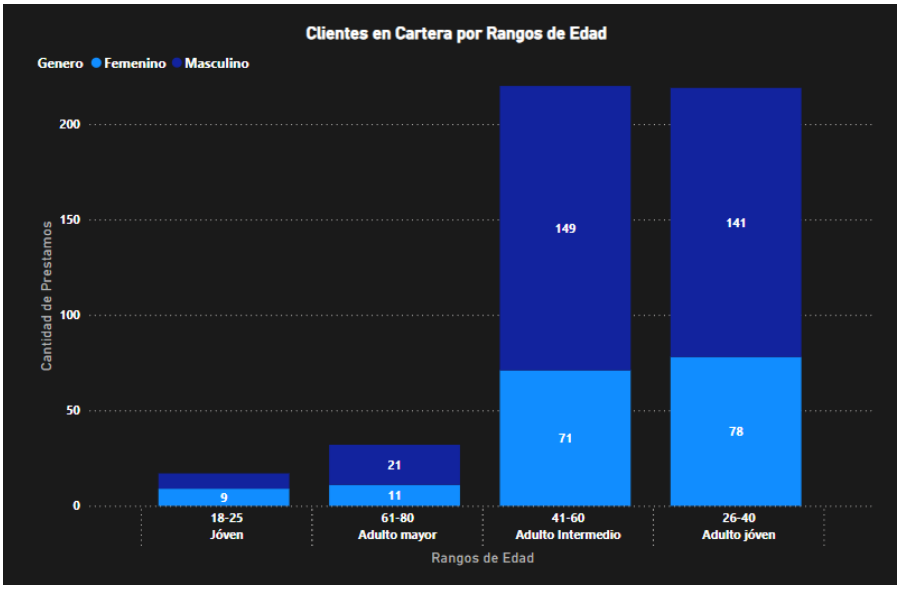


Tabla 10

Tabla mora por grupos de edad

Rango de Mora	18-25 años	26-40 años	41-60 años	61-80 años	No Registrado	Tendencia Observada
61-90 días	0.81%	12.01%	7.68%	0.92%	0.01%	Pico en adultos jóvenes (26-40)
91-120 días	0.42%	9.56%	4.52%	1.50%	-	26-40 años duplican a 41-60
121-150 días	0.76%	4.09%	4.19%	-	2.48%	Similar riesgo 26-40 y 41-60
151-180 días	-	2.57%	2.96%	-	0.07%	Adultos maduros dominan

Rango de Mora	18-25 años	26-40 años	41-60 años	61-80 años	No Registrado	Tendencia Observada
181+ días	-	22.21%	15.33%	-	4.89%	Jóvenes adultos (26-40) lideran mora extrema

Patrones Clave por Grupos de Edad

Grupo 26-40 años: máxima exposición en todos los rangos (12.01% en 61-90 días, 22.21% en 181+ días).

Representan +45% del riesgo total en mora temprana/avanzada

Tendencia: Principal foco de riesgo crediticio

Grupo 41-60 años:

- Segundo lugar en mora (7.68%-15.33%)
- Mayor vulnerabilidad en mora avanzada (>121 días)
- Tendencia: Riesgo progresivo con antigüedad del crédito

Jóvenes (18-25):

Mínima participación (0.42%-0.81%)

Tendencia: Mejor comportamiento crediticio

Adultos mayores (61-80):

- Baja presencia (0.92%-1.50%)
- Tendencia: Perfil conservador

Tabla 11

Evaluación de los rangos de edad de clientes en mora

Grupo Edad	Problema	Medida Urgente	Meta 6 meses
26-40	Alto riesgo en todas las etapas	Revisión de capacidad de pago real	Reducir 25%
41-60	Mora avanzada persistente	Programas de reestructuración	Reducir 15%
18-25	Buen desempeño	Incentivar más créditos (bajo riesgo)	+10% volumen

Políticas Preventivas

- Para 26-40 años:

Análisis de deuda consolidada

Límites de crédito más conservadores

- Para 41-60 años:

Seguros de desempleo incluidos en créditos

Plazos más cortos

Para el caso de los prestatarios de 26-40 años debido a que la calidad de la cartera se concentra en:

- 55-60% del riesgo en mora temprana
- 48% de la mora crítica (>181 días)

Los requerimientos inmediatos sugeridos son:

- Estrategia focalizada para jóvenes adultos (26 – 40)
- Modelos predictivos específicos por edad
- Ajuste de scoring para rangos 26-40 y 41-60 años

Figura 16

Nivel de estudios cartera vencida

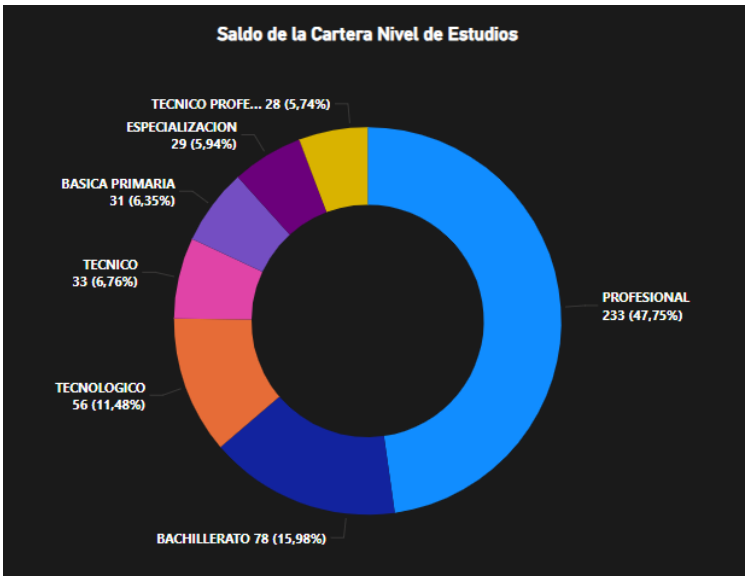


Tabla 12

Evaluación de niveles de estudio de clientes en mora

Rango de Mora	Profesional	Bachillerato	Técnico Profesional	Técnico	Tecnológico	Especialización	Básica Primaria	No Registrado
61-90 días	8.57%	3.01%	1.28%	1.85%	1.83%	2.87%	2.01%	0.01%
91-120 días	10.39%	1.23%	-	0.00%	3.97%	0.42%	-	0.05%
121-150 días	6.03%	2.08%	0.05%	-	-	0.88%	-	2.48%
151-180 días	2.22%	0.06%	0.93%	0.88%	0.87%	0.57%	-	0.07%
181+ días	18.03%	6.10%	4.21%	3.92%	2.60%	3.32%	2.32%	4.89%

Patrones Clave por Nivel Educativo

- Profesionales Universitarios

Lideran la mora en todos los rangos (8.57%-18.03%)

Representan 40% del riesgo total en mora extrema (>181 días)

Tendencia: Mayor exposición en mora avanzada

- Bachilleres

Segundo lugar en mora temprana (3.01%)

Alto riesgo en mora extrema (6.10%)

Tendencia: Riesgo consistente en todas las etapas

Técnicos y Tecnólogos:

Menor mora temprana (1.28%-1.85%)

Pico en 91-120 días (Tecnológico 3.97%)

Tendencia: Problemas selectivos en mora media

- No Registrados

Bajo riesgo inicial (0.01%)

Alto impacto en mora extrema (4.89%)

Tendencia: Gestión documental deficiente

Paradoja Educativa:

Los profesionales tienen 3x más mora que técnicos

La educación superior no correlaciona con mejor pago

Focos de Riesgo:

64.3% de la mora >181 días está en profesionales + bachilleres

Tecnológicos muestran pico inusual en mora media (91-120 días)

Ajustes de Política

Para Profesionales, establecer límites de endeudamiento basados en flujo de caja y seguro de desempleo obligatorio

En el caso de perfiles técnicos, promover incentivos por buen comportamiento y tasas preferenciales para mora <60 días.

Los profesionales universitarios generan el mayor riesgo crediticio (18.03% en mora extrema), se identifica en este caso que la educación técnica muestra mejor desempeño que la universitaria, para lo cual los casos en la categoría de no registrados escalan peligrosamente en mora avanzada, en este caso las acciones inmediatas son:

- Auditoría de capacidad de pago en profesionales
- Paquete educativo para bachilleres
- Regularización masiva de los “no registrados"

Análisis de Concentración de Cartera Vencida por Puntaje ACIERTA

Figura 17

Distribución de Riesgo por Puntaje de aprobación en el momento de desembolso.

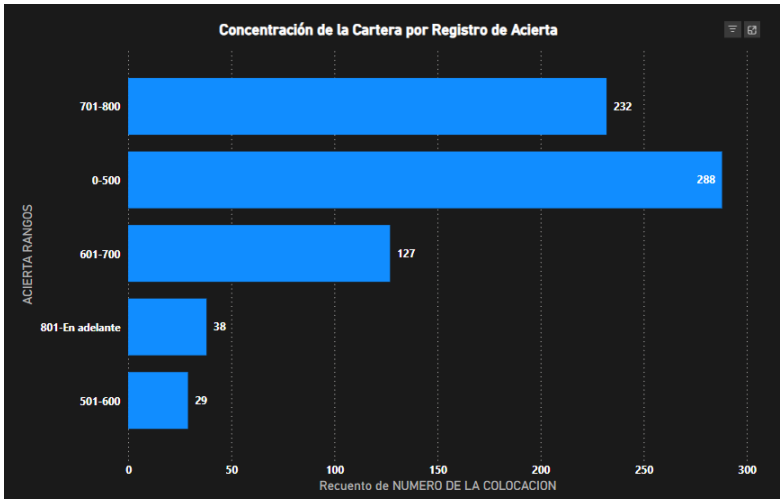


Tabla 13

Participación por rangos de ACIERTA

Rango ACIERTA	Participación en Mora
701–800	42.5%
0–500	30.4%
601–700	20.1%
801+	6.5%
501–600	0.5%

Concentración Principal:

El 72.9% de la mora se concentra en solo dos rangos. 701-800 puntos (42.5%) y 0-500 puntos (30.4%).

Los demas datos se distribución en 601-700 puntos: 20.1% 801+ puntos: 6.5% (mejor desempeño) y 501-600 puntos: marginal (0.5%)

Para lo relacionado con ACIERTA (Puntaje del cliente verificado en centrales de riesgo), el foco crítico se encuentra en el rango (701-800 puntos), para lo cual se sugiere revisión urgente de criterios de aprobación, análisis de causas de incumplimiento.

Para el caso del segmento 0-500 puntos, se sugiere mantener políticas restrictivas, estrategias de cobranza temprana

En el caso del ACIERTA que da un Buen Desempeño (801+), se sugiere validar y replicar factores de éxito, considerar flexibilización controlada.

Propuesta y estrategias para la prevención de la mora y recuperación de cartera

Con base en el análisis exploratorio, la segmentación de cartera y los modelos predictivos implementados, se propone un conjunto de estrategias estructuradas en cinco ejes: prevención, segmentación, gestión de cobranza, ajustes en políticas de crédito y mejora continua. Estas estrategias están orientadas a anticipar la entrada en mora, actuar con precisión sobre los perfiles más riesgosos y mejorar la eficiencia de la recuperación de cartera.

Prevención y alerta temprana

Implementar el modelo LSTM binario como herramienta de detección temprana: Su elevado recall (96%) para identificar clientes que efectivamente caerán en mora lo convierte en una herramienta clave para activar alertas automáticas de riesgo.

Generar alertas internas a los 15, 30 y 45 días de mora: Activar gestiones proactivas antes de que los clientes pasen a mora crítica, con foco en renegociación, recordatorios y atención personalizada.

Reforzar validación de datos en etapa de otorgamiento: La presencia de campos "No registra" se asocia a mayor morosidad. Se deben bloquear aprobaciones automáticas para formularios incompletos y penalizar estos casos en el score de riesgo interno.

Segmentación de clientes según riesgo y comportamiento

Utilizar el modelo XGBoost multiclase para segmentar cartera en cuatro grupos:

0: Al día: mantener condiciones actuales.

1: Entra en mora: activar alertas anticipadas.

2: Sale de mora: reforzar con incentivos y acompañamiento.

3: Se mantiene en mora: priorizar para cobranza intensiva o judicial.

Diseñar estrategias diferenciadas por grupo etario:

26–40 años: representan más del 45% del riesgo total. Aplicar scoring conservador, monitoreo continuo y programas educativos.

41–60 años: mostrar riesgo progresivo en mora avanzada. Enfocar en restructuración y seguimiento posterior al crédito.

18–25 años: buen comportamiento. Incentivar con condiciones preferenciales.

Segmentar por tipo de cliente

Independientes: presentar mayor riesgo en mora extrema (>181 días). Requieren garantías adicionales, scoring más estricto y plazos más cortos.

Empleados: mayor exposición en mora temprana. Automatizar descuento por nómina y ofrecer planes de pago flexibles.

Ajustar política según nivel educativo

Profesionales universitarios muestran mayor morosidad. Validar ingresos declarados y aplicar límites más estrictos.

Técnicos y tecnólogos presentan mejor comportamiento: ofrecer incentivos por cumplimiento.

Gestión de cobranza con enfoque predictivo

Aplicar el modelo LSTM multiclase para identificar clientes que se mantendrán en mora: Esta predicción debe usarse para priorizar procesos judiciales, auditorías de garantías y recuperación intensiva.

Jerarquizar recursos jurídicos y de cobranza según clasificación de riesgo

Clientes con alta probabilidad de permanecer en mora: asignación prioritaria de abogado y activación del proceso de embargo

Clientes con probabilidad de salir de mora: ofrecer acompañamiento intensivo y programas de rehabilitación financiera.

Automatizar asignación de gestores o agentes según perfil de riesgo: Clientes con predicción de entra en mora o mantiene mora deben tener contacto directo humano; clientes al día pueden ser gestionados de forma digital o semiautomatizada.

Aplicar campañas educativas o de refinanciamiento según perfil: En segmentos como bachilleres o independientes se debe combinar cobranza con programas de educación financiera y alivios de corto plazo.

Ajustes en políticas de crédito y scoring

Revisar criterios de aprobación para clientes con ACIERTA entre 701–800: Este grupo concentra el 42.5% de la mora y debe ser evaluado con mayor rigor a través del score interno.

Aplicar límites conservadores en ingresos entre 1 y 4 millones de pesos: Según el análisis bivariado, este grupo presenta la mayor proporción de mora; aplicar controles de carga financiera y validación documental adicional.

Ajustar condiciones de otorgamiento según el vehículo:

Modelos anteriores a 2018 presentan mayor riesgo. Limitar plazo, exigir garantías y monitorear el valor comercial.

Operaciones anuladas o productos no estándar deben pasar por filtros adicionales de análisis de riesgo.

Monitoreo, aprendizaje y mejora continua

Reentrenar mensualmente los modelos con datos actualizados: para mantener su precisión, los modelos deben adaptarse a cambios macroeconómicos, estacionales y operativos.

Simular escenarios de política con XGBoost: Evaluar cómo afectarían los cambios en variables clave (plazo, ingreso, perfil) al riesgo proyectado para afinar las decisiones comerciales.

Incorporar visualizaciones y dashboards en Power BI:

Mostrar mapas de riesgo.

Ver evolución de métricas de mora.

Identificar alertas tempranas por zona, perfil o producto.

Conclusiones y recomendaciones

Los modelos avanzados de machine learning (LSTM binario/multiclase y XGBoost multiclase) demostraron capacidad para predecir el comportamiento de mora con precisión. El modelo LSTM binario destacó con un *recall* del 96%, ideal para identificar clientes morosos reales.

Con el análisis realizado a los datos en los modelos XGBoost y redes secuenciales (LSTM) se pudieron identificar variables que aportan mayor cantidad de información nueva como son tipo de solicitante, modelo de vehículo y puntaje acierta, las cuales tienen un peso relevante en la clasificación del riesgo crediticio.

Los modelos presentaron dificultades para clasificar adecuadamente las clases relacionadas con transiciones del estado como entra en mora y sale de mora, lo que sugiere la necesidad de vincular variables más detalladas que permitan mejorar la representación del comportamiento dinámico de los clientes.

Una de las principales limitaciones se presentó el desequilibrio en las categorías del estado crediticio. Las clases “entra en mora” y “sale de mora” se encuentran subrepresentadas respecto a “se mantiene” o “no registra”, lo que afectó negativamente el desempeño de los modelos multiclase, especialmente en términos de precisión y recall para las clases minoritarias. Este desbalance reduce la capacidad del modelo para detectar transiciones clave en el comportamiento crediticio.

Dado su elevado poder de predicción de morosidad, se recomienda utilizar el modelo LSTM binario como herramienta de monitoreo preventivo, especialmente en la etapa previa a la mora, para activar alertas tempranas y permitir la intervención proactiva.

Para una visión más integral del ciclo de crédito, se sugiere emplear modelos multiclase que permitan anticipar no solo la ocurrencia de mora, sino también las transiciones entre estados.

Conforme con los resultados obtenidos se sugiere trabajar en diseño de estrategias de intervención basadas en el riesgo predicho, para cliente con alta probabilidad de entrar en mora, intervención inmediata, emplear herramientas de reestructuración de crédito. Riesgo medio hacer seguimiento constante y promover educación financiera, por último, para aquellos clientes que representan un riesgo bajo, promover campañas de fidelización.

Anexos de Entrega

Como parte del presente proyecto, se entregan los siguientes anexos que complementan y respaldan el análisis realizado:

1. Repositorio de Código y Documentación Técnica

Disponible en GitHub:

https://github.com/sebastiaanbv/Seminario_investigacion

Este repositorio contiene:

- Scripts de preprocesamiento, análisis exploratorio y modelado (LSTM y XGBoost).
- Documentación técnica del desarrollo, experimentación y evaluación de los modelos.
- Informes en Jupyter Notebooks con visualizaciones y justificaciones técnicas.
- Archivos de datos anonimizados utilizados para entrenamiento y validación.

Este repositorio permite replicar el flujo completo del análisis predictivo, promoviendo la transparencia y la reproducibilidad del estudio.

2. Dashboard Interactivo en Power BI

Disponible en línea:

Dashboard de Riesgo Crediticio - Power BI

El tablero muestra:

- Segmentación de mora por edad, tipo de solicitante, nivel educativo y puntaje ACIERTA.
- Análisis comparativo de riesgo y tendencias clave identificadas.
- Visualización de la distribución del saldo vencido y su participación relativa.

Este dashboard está diseñado para facilitar la interpretación ejecutiva de los resultados, permitiendo tomar decisiones basadas en datos de forma intuitiva y dinámica.

Lista de referencias

Albisetti, R. (2021). Finanza empresarial : Estrategia, mercados y negocios estructurados (2° ed.). In *Finanza empresarial*.

Aurélien Géron. (2022). Hands-on Machine Learning with Scikit-Learn, Keras, and Hands-On Machine Learning TensorFlow. In *O Reilly Media, Inc.*

Baesens, B., Rösch, D., & Scheule, H. (2016). Credit Risk Analytics. In *Credit Risk Analytics*. https://doi.org/10.1002/9781119449560

Crowe Colombia. (s.f). (n.d.). *Informalidad Laboral: Principal Causa de Desigualdad en Salarios de América Latina*.

Davenport, T. H. (2014). How strategists use “big data” to support internal business decisions, discovery and production. *Strategy and Leadership*, *42*(4). https://doi.org/10.1108/SL-05-2014-0034

Esborraz, D. F. (2021). El concepto, los principios generales y las fuentes en materia de obligaciones: observaciones a los artículos 432 a 434 del Proyecto de Código Civil de la Universidad Nacional de Colombia. *Revista de Derecho Privado*, *41*. https://doi.org/10.18601/01234366.n41.12

Izar Landeta, J. M., & Ynzunza Cortés, C. B. (2017). El Impacto del Crédito y la Cobranza en las Utilidades. *Poliantea*, *13*(24). https://doi.org/10.15765/plnt.v13i24.701

León-Vega, L. S., & Espinoza-Alcívar, E. I. (2023). Análisis de los factores que intervienen en el crecimiento de cartera vencida de empresas servicios financieros. *INNOVA Research Journal*, *8*(3.1). https://doi.org/10.33890/innova.v8.n3.1.2023.2342

Morales, A., & Morales, J. (2014). Crédito y Cobranza. In *Crédito y Cobranza*.

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, *8*(11). https://doi.org/10.3390/data8110169

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal*, *74*. https://doi.org/10.1016/j.asoc.2018.10.004

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11). https://doi.org/10.1080/14786440109462720

Russell, S., & Norvig, P. (2021). *Artificial Intelligence, Global Edition A Modern Approach*.

Samuel, A. L., & Gabel, F. (1959). Artificial Intelligence for Games: Seminar Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(1959).

Sandeep Shinde, & Satish Kale. (2023). Unleashing the Power of Machine Learning: A Comparative Study of Classification Algorithms for Credit Risk Assessment. *International Journal of Advanced Research in Science, Communication and Technology*. https://doi.org/10.48175/ijarsct-11139

Scagliola Martina. (2022). *Credit Risk Assessment Using Machine Learning Techniques (Doctoral dissertation, Politecnico di Torino)*. Politecnico di Torino.

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. In *Neural Computing and Applications* (Vol. 34, Issue 17). https://doi.org/10.1007/s00521-022-07472-2

Tamayo Lombana, A. (2004). *Las Principales Garantías del Crédito*.

Turing, A. M. (2012). Computing machinery and intelligence. In *Machine Intelligence: Perspectives on the Computational Model*. https://doi.org/10.7551/mitpress/6928.003.0012

Van Kampen, N. G. (1981). *Stochastic processes in physics and chemistry (1st ed.). Elsevier.*

Wang S, Z. X. (2024). *Research on Credit Default Prediction Model Based on TabNet-Stacking. Entropy (Basel)*.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. *Procedia Computer Science*, *174*. https://doi.org/10.1016/j.procs.2020.06.069