



**Modelo de Contacto en Casas de Cobro bajo la Ley 2300 de 2023:
Mejoramiento y Condiciones de Aplicabilidad.**

GERMÁN DARÍO MACÍAS GALÍNDEZ
HÉCTOR MAURICIO HERRERA GALARZA

Universidad Ean
Facultad de Ingeniería
Maestría en Ciencia de Datos
Bogotá, Colombia
01/Marzo/2026

**Modelo de Contacto en Casas de Cobro bajo la Ley 2300 de 2023: Mejoramiento y
Condiciones de Aplicabilidad.**

**GERMÁN DARÍO MACÍAS GALÍNDEZ
HÉCTOR MAURICIO HERRERA GALARZA**

Trabajo de grado presentado como requisito para optar al título de:

Maestría en Ciencia de Datos

Director (a):

JOHN JAIRO PORRAS VEGA

Modalidad:

Trabajo Dirigido

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

01/Marzo/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Ciudad, día/mes/año

A nuestros padres, que con paciencia y amor nos enseñaron a caminar por la vida.

Agradecimientos

A nuestro tutor por su dedicación y constantes acompañamiento en el desarrollo del proyecto.

Resumen

Las empresas de gestión de cobro enfrentan ineficiencias en la contactabilidad con clientes morosos, especialmente en carteras con más de 120 días de mora. Promociones y Cobranzas Beta, filial de Davivienda, requiere optimizar su modelo de contacto bajo los lineamientos de la Ley 2300 de 2023.

El objetivo es presentar una propuesta de diseño de un modelo de priorización de contacto basado en analítica predictiva, que identifique el momento óptimo y el cliente adecuado para contactar dentro del marco normativo vigente. El trabajo constituye un enfoque complementario a los modelos existentes en la empresa, orientado a que Promociones y Cobranzas Beta pueda integrarlo y desplegarlo en sus propios sistemas.

La metodología es de predominancia cuantitativa, aplicando técnicas de machine learning sobre datos históricos de gestión de cobranza de Promociones y Cobranzas Beta. Se evaluaron seis modelos de clasificación supervisada (Regresión Logística, Random Forest, Extra Trees, XGBoost, LightGBM y CatBoost) sobre una muestra de 50.000 registros, con partición 80/20 estratificada para entrenamiento y prueba.

Los resultados obtenidos muestran que CatBoost logró el mejor equilibrio entre precisión y recall (F1-score: 0.655, ROC-AUC: 0.967) en un dataset con alta asimetría de clases (5.87% de contactos efectivos). Con base en este modelo, se diseñó un esquema híbrido de priorización que combina el score predictivo con reglas operativas y las restricciones de la Ley 2300 de 2023, generando un archivo de segmentación operativa que asigna a cada cliente una acción sugerida (llamada, WhatsApp o SMS) y una franja horaria de contacto. La validación del modelo se realizó sobre datos históricos con partición train/test; la validación operativa en producción queda fuera del alcance de este trabajo. Las conclusiones buscan demostrar que la analítica avanzada mejora la recuperación de cartera mientras garantiza cumplimiento normativo y trato ético al deudor.

Palabras clave: gestión cobranza, contactabilidad, Ley 2300/2023, recuperación cartera, optimización recursos, segmentación clientes.

Abstract

Debt collection companies face inefficiencies in contacting delinquent clients, particularly in portfolios more than 120 days overdue. Promociones y Cobranzas Beta, a subsidiary of Davivienda, needs to optimize its contact model under the guidelines of Law 2300 of 2023.

The objective is to present a design proposal for a contact prioritization model based on predictive analytics that identifies the optimal moment and the appropriate client to contact within the current regulatory framework. The work constitutes a complementary approach to the models already in use at the company, intended for Promociones y Cobranzas Beta to integrate and deploy it within its own systems.

The methodology is predominantly quantitative, applying machine learning techniques to historical debt-collection management data from Promociones y Cobranzas Beta. Six supervised classification models (Logistic Regression, Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost) were evaluated on a sample of 50,000 records, using a stratified 80/20 train/test split.

The results show that CatBoost achieved the best balance between precision and recall (F1-score: 0.655; ROC-AUC: 0.967) on a dataset with high class imbalance (5.87% effective contacts). Based on this model, a hybrid prioritization scheme was designed that combines the predictive score with operational rules and the constraints of Law 2300 of 2023, producing an operational segmentation file that assigns each client a suggested action (call, WhatsApp, or SMS) and a contact time slot. The model was validated on historical data using a train/test split; operational validation in production is beyond the scope of this work. The conclusions demonstrate that advanced analytics improves portfolio recovery while ensuring regulatory compliance and ethical treatment of the debtor.

Keywords: debt collection management, contactability, Law 2300/2023, portfolio recovery, resource optimization, customer segmentation

Contenido

	Pág.
Lista de Figuras	10
Lista de Tablas	11
Introducción	12
Objetivos	14
<i>Objetivo General</i>	<i>14</i>
<i>Objetivos Específicos</i>	<i>14</i>
Justificación.....	15
Marco Institucional	18
Análisis del sector	24
<i>Factores legales y regulatorios.....</i>	<i>24</i>
<i>Factores tecnológicos.</i>	<i>25</i>
<i>Factores Políticos.....</i>	<i>26</i>
<i>Factores Sociales.....</i>	<i>27</i>
<i>Factores Ambientales.....</i>	<i>28</i>
Diagnóstico interno de la empresa.....	30
<i>Análisis interno (DOFA).....</i>	<i>30</i>
<i>Fortalezas.</i>	<i>30</i>

Modelo de contacto en casas de cobro bajo la Ley 2300 de 2023: Mejoramiento y Condiciones de Aplicabilidad.	9
<i>Debilidades.</i>	31
Análisis externo	32
<i>Oportunidades.</i>	32
<i>Amenazas.</i>	32
Marco Teórico	34
Diseño metodológico	45
<i>Población y muestra</i>	46
<i>Variables de análisis</i>	47
<i>Metodología para el desarrollo del Objetivo 1</i>	48
<i>Metodología para el desarrollo del Objetivo 2</i>	48
<i>Metodología para el desarrollo del Objetivo 3</i>	49
<i>Metodología para el desarrollo del Objetivo 4</i>	50
Contribuciones originales esperadas	52
Fundamentación Teórica	54
Plan de intervención	63
Conclusiones y Recomendaciones	85
Referencias	91

Lista de Figuras

Figura 1	21
Figura 2	51
Figura 3	69
Figura 4	70
Figura 5	71
Figura 6	72
Figura 7	72
Figura 8	73
Figura 9	73
Figura 10	73
Figura 11	74
Figura 12	75
Figura 13	82
Figura 14	83

Lista de Tablas

	Pág.
Tabla 1.....	16
Tabla 2.....	67
Tabla 3.....	77
Tabla 4.....	78
Tabla 5.....	78
Tabla 6.....	88

Introducción

Las empresas dedicadas a la gestión de cobro de las carteras por deuda financiera presentan un inconveniente, en especial a las carteras que tienen por mora más de 120 días y es el contacto con el cliente para dar a conocer el beneficio que le puede otorgar para solventar la deuda que tiene presente con alguna entidad financiera donde por motivos varios se dejó de pagar.

En Promociones y Cobranzas Beta, existe un sistema de contactabilidad implementado, pero aún hay problemas accionables con la eficiencia en la recuperación de cartera, el servicio al cliente y la optimización de recursos. El modelo actual también puede verse afectado por diversos aspectos como el enfoque tecnológico, la segmentación de clientes y la metodología de gestión de contactos.

Ahora bien, a pesar de contar con infraestructura operativa estable para la gestión de todos los portafolios de compra de deuda que tienen, donde se dispone diferentes tipos de datos para hacer un primer contacto con los clientes que presentan deuda, aún existen brechas en la optimización del modelo de contactabilidad que requieren un análisis detallado. Uno de los principales temas a tratar es la falta de priorización eficiente para la gestión de saldos por cobrar. De igual modo se puede apoyar con técnicas modernas para tener un rendimiento más adecuado al momento de segmentar los clientes, estudios han demostrado que la implementación de modelos de aprendizaje automático puede mejorar significativamente la priorización de clientes con mayor probabilidad de pago, lo que optimiza los recursos de cobranza y mejora la efectividad en la recuperación de cartera (Appel et al., 2019). Sin embargo, la falta de integración de información estructurada y modelos predictivos en muchos procesos de cobranza sigue representando un obstáculo para su modernización.

En este sentido, la optimización del modelo de contacto debe centrarse en estrategias basadas en datos y segmentación inteligente de clientes. Diversos enfoques han demostrado que un modelo de contactabilidad eficiente debe considerar factores como el comportamiento de pago histórico, la clasificación del cliente y el tiempo óptimo de contacto para aumentar las probabilidades de recuperación de la deuda (Zeng et al., 2008). No obstante, la realidad en muchas empresas de cobranzas es que los procesos se gestionan con metodologías tradicionales que no aprovechan plenamente la analítica avanzada para mejorar la toma de decisiones. Este proyecto tiene como objetivo desarrollar un modelo

para la contactabilidad basado en técnicas innovadoras, con el fin de incrementar la eficiencia en la recuperación de cartera y optimizar la asignación de recursos dentro del proceso de cobranza en Promociones y Cobranzas Beta.

En este contexto, el proyecto se aborda como un proceso de análisis y mejora del modelo de contactabilidad, orientado a identificar las principales deficiencias del esquema actual y a proponer soluciones aplicables al contexto operativo de dicha organización. El trabajo parte de un diagnóstico de la gestión de contacto vigente, continúa con el análisis de la información disponible y culmina con la formulación de un plan de mejoramiento que integra criterios analíticos, operativos y normativos, en concordancia con los lineamientos establecidos por la Ley 2300 de 2023.

El documento se estructura de la siguiente manera: inicialmente se presenta el marco institucional y el análisis del sector, que contextualizan la problemática desde el ámbito organizacional y regulatorio; posteriormente se desarrolla el marco teórico que sustenta la propuesta; a continuación se expone el diseño metodológico para el desarrollo del proyecto; seguidamente se presenta el diagnóstico organizacional y el plan de intervención propuesto; y finalmente se exponen las conclusiones y recomendaciones derivadas del estudio.

Objetivos

Objetivo General

Desarrollar una propuesta de diseño de un modelo de priorización de contactos basado en Machine Learning para la gestión de cobranza de cartera de difícil recaudo en Promociones y Cobranzas Beta, mediante la identificación del momento óptimo y los clientes prioritarios a contactar, en el marco del cumplimiento de la Ley 2300 de 2023.

Objetivos Específicos

1. Diagnosticar el estado actual de la gestión de contactabilidad de Promociones y Cobranzas Beta a partir del análisis exploratorio de los datos históricos de gestión y del conocimiento operativo del equipo, con el fin de identificar las variables con mayor incidencia en el contacto efectivo.
2. Evaluar comparativamente algoritmos de Machine Learning de clasificación supervisada para la predicción de contactabilidad, mediante métricas de desempeño sobre datos históricos de gestión, con el fin de seleccionar el de mejor resultado para el contexto de cartera de difícil recaudo.
3. Proponer un esquema híbrido de priorización de contacto que integre el algoritmo de mejor desempeño con reglas operativas y normativas de la Ley 2300 de 2023, generando como entregable un archivo de segmentación que asigne a cada cliente una acción sugerida y una franja horaria de contacto, como insumo accionable para la gestión de cobranza de Promociones y Cobranzas Beta.

Justificación

Cuando se trata del sector de gestión de deuda, contactar a los clientes de la manera más eficiente es la clave para recuperar documentos en la cartera. Sin embargo, muchas corporaciones luchan por priorizar a los clientes, segmentar efectivamente y ejecutar acciones que promuevan el contacto con el deudor. Esta deficiencia se magnifica cuando la cartera cumple 120 días de atraso, mientras que la probabilidad de rehabilitar al suscriptor en mora se desploma si no se aplican las medidas de gestión adecuadas y las metodologías de contacto correctas.

Así, todavía hay organizaciones con una infraestructura operativa estable, pero que no logran optimizar su modelo de contacto. En general, la ausencia de una estrategia basada en datos, acompañada del uso limitado de herramientas tecnológicas especializadas, impacta directamente en la efectividad de las cobranzas. Como resultado, los costos operativos son altos y la capacidad para recuperar la cartera se reduce. Además, debido a la naturaleza de los métodos tradicionales, los clientes con mayor disposición o capacidad de pago no siempre se identifican con precisión, lo que lleva a deficiencias en los esfuerzos y a una asignación ineficiente de recursos. Para el caso de Promociones y Cobranzas Beta esta situación no es ajena y les ha conducido en la búsqueda y aplicación de nuevas estrategias o modelos analíticos que permitan minimizar el número de intentos de contacto innecesarios para optimizar la experiencia del cliente y aumentar la tasa de recuperación efectiva de los portafolios comprados, lo que conlleva en última instancia a mejores opciones y más sostenibles a largo plazo.

Por último, este proyecto ayudaría a detectar ineficiencias en cuanto a la gestión de deuda, además de proporcionar un enfoque práctico para optimizar el proceso. Esto apoyará a la empresa en el diseño de mejores estrategias para la recuperación de cartera, apoyadas en tecnologías modernas, en busca de ofrecer una experiencia más

fluida a sus clientes a través de una segmentación adecuada, una asignación prudente de recursos y procesos de contacto eficientes.

Viabilidad del proyecto

Tabla 1

Criterios de factibilidad del proyecto

Criterio	Factibilidad (siendo 1 menor y 5 mayor)
Acceso a la información	5
Apoyo e interés de la alta dirección	5
Disponibilidad de recursos requeridos	5
Probabilidad de avance en el tiempo establecido	4
Tamaño de la empresa para soportar y desarrollar el plan de mejora a proponer	4
Promedio	4.6

De acuerdo con los criterios evaluados en la Tabla 1, el proyecto presenta un alto nivel de viabilidad, reflejado en un promedio de 4.6 sobre 5. Este resultado evidencia que se cuenta con condiciones favorables para su ejecución.

En primer lugar, se dispone de acceso a la información necesaria, respaldado por la autorización formal de la empresa Promociones y Cobranzas Beta, la cual ha permitido el uso de los datos bajo estrictos protocolos de protección de la información y en cumplimiento de los lineamientos éticos y legales vigentes. Asimismo, se destaca el apoyo e interés de la alta dirección, lo cual constituye un factor clave para el desarrollo exitoso del proyecto.

Adicionalmente, el equipo de trabajo cuenta con la experiencia requerida, particularmente en procesos de migración de bases de datos a la nube, lo que facilita la

comprensión del negocio, su estructura y el tratamiento adecuado de la información. Esto incluye la gestión segura de datos sensibles mediante mecanismos de cifrado, garantizando su integridad y confidencialidad.

En cuanto a los recursos, se dispone de las bases de datos necesarias, tanto de entidades externas como del Banco Davivienda, como de carteras propias de la empresa, lo que permite abordar el análisis desde diferentes perspectivas del negocio.

Por otra parte, existe una alta probabilidad de avance dentro del tiempo establecido, dado que el proyecto cuenta con un periodo definido desde 2024 para la recolección de información y el inicio del proyecto. Además, la empresa dispone de una estructura organizacional sólida, con 587 colaboradores distribuidos en Colombia, y con acceso directo a la sede principal en Bogotá, lo que facilita la interacción con los directivos y el área encargada de la protección de datos.

En conjunto, estos factores confirman que el proyecto no solo es factible, sino que cuenta con las condiciones necesarias en términos de apoyo institucional, disponibilidad de información, experiencia del equipo y recursos organizacionales para su adecuada ejecución.

Marco Institucional

Promociones y Cobranzas Beta S.A. es una empresa colombiana especializada en la gestión de cobranzas y recuperación de cartera vencida, con más de 34 años de experiencia en el sector financiero (Cobranzas Beta, 2025). Como filial de Davivienda, su principal objetivo es recuperar cartera en mora, aplicando estrategias basadas en el análisis de datos y el cumplimiento de regulaciones financieras (Cobranzas Beta, 2025).

La empresa brinda soluciones tecnológicas y personalizadas en la gestión de cobranzas, garantizando procesos ágiles y adaptados a las necesidades del cliente. Su enfoque está alineado con la Ley 2300 de 2023, la cual regula la comunicación con deudores para evitar el acoso financiero y promover la transparencia en la cobranza (Congreso de Colombia, 2023). Entre sus disposiciones centrales, la ley establece que el contacto solo puede realizarse de lunes a viernes entre las 7:00 a.m. y las 7:00 p.m., y los sábados hasta la 1:00 p.m., prohibiendo cualquier contacto en domingos y días festivos. Adicionalmente, limita la frecuencia de intentos de cobro por día y por semana, define los canales autorizados de comunicación, prohíbe el contacto con terceros no autorizados y exige que en todo momento se respete la dignidad e intimidad del deudor. Estas restricciones se integraron como parámetros de diseño en el modelo propuesto: la jornada operativa modelada va de 8:00 a.m. a 5:00 p.m., se respetó el horario de almuerzo de los asesores con capacidad reducida en las franjas de 12:00 a 2:00 p.m., se estableció máximo un intento de contacto por cliente por día y se evitó el contacto multicanal en la misma semana.

Promociones y Cobranzas Beta opera como una empresa especializada en la gestión y recuperación de cartera vencida, con un modelo de negocio enfocado en la compra y administración de portafolios de deuda provenientes principalmente del sector financiero. En el caso analizado, la operación se concentra en cartera castigada, compuesta por

obligaciones con altos niveles de mora y mayor dificultad de recuperación, adquiridas en su mayoría a entidades como Davivienda y, en menor medida, a otros bancos. Este tipo de cartera representa un reto operativo importante, dado que se trata de clientes que han transitado por distintos segmentos de mora sin lograr una normalización efectiva de su obligación, lo que reduce progresivamente la probabilidad de contacto y recuperación.

La gestión de cobranza dentro de la empresa se organiza según el tramo de mora de los clientes. En etapas tempranas, como mora de 0 a 30 días o de 31 a 60 días, la recuperación puede resultar más sencilla, ya que en muchos casos el incumplimiento responde a situaciones puntuales, como un olvido de pago, una dificultad transitoria de liquidez o un inconveniente operativo en el débito. Sin embargo, a medida que la mora avanza hacia rangos de 61 a 90 días, 91 a 120 días y posteriormente cartera castigada, la gestión se torna más compleja. Esto ocurre tanto por la prolongación del incumplimiento como por la acumulación de intentos previos de contacto realizados en campañas anteriores, lo que implica que los clientes que llegan a estos tramos suelen ser más difíciles de localizar, menos propensos a responder y, en muchos casos, más sensibles a su situación financiera particular.

En este contexto, la operación de la empresa no se limita a contactar clientes de forma masiva, sino que exige una gestión constante, segmentada y estratégica según el nivel de mora y las posibilidades reales de recuperación. Aunque la dificultad aumenta a medida que la cartera envejece, siguen existiendo casos en los que es posible recuperar al cliente, facilitar acuerdos de pago y contribuir incluso a una eventual mejora de su vida crediticia. Por ello, el contacto continuo y bien priorizado sigue siendo una función central del negocio. Esta lógica hace necesario optimizar la distribución diaria de la cartera entre asesores, canales y franjas horarias, de manera que la gestión sea más eficiente, trazable y alineada tanto con la operación real como con las restricciones normativas vigentes.

Referentes Estratégicos

Promociones y Cobranzas Beta estructura su gestión en torno a los siguientes elementos clave:

Misión

Nuestro objetivo es ser la gestión líder de cobranza de deudas en Colombia (desde la cobranza de carteras vencidas sin generar molestias a los clientes), a través de soluciones innovadoras y eficientes. Nos enfocamos en el uso de sistemas de alta gama actualizados, técnicas de pronóstico y segmentación para formular procesos articulados y personalizados. También aspiramos a proporcionar un servicio transparente y ético que examine las regulaciones existentes, mejore el bienestar financiero de los clientes y apoye la sostenibilidad del sector bancario (Cobranzas Beta, 2026).

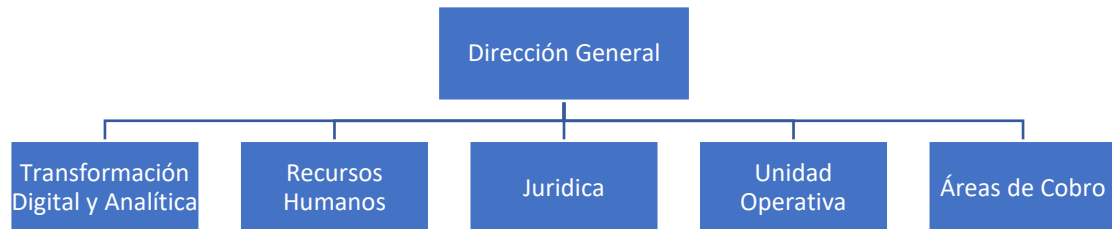
Visión

Promociones y Cobranzas Beta busca ser considerada por las autoridades en la región como la mejor empresa de cobranza de deudas en América Latina, enfocada en la eficiencia de la recuperación de carteras y la experiencia del cliente a través de herramientas tecnológicas innovadoras. A través de estos cambios, la empresa aspira a revolucionar el sector introduciendo automatización, omnicanalidad y aprendizaje automático, para facilitar el contacto con los consumidores de una manera que sea tanto efectiva como ética, en línea con los cambios en la regulación financiera y la digitalización del mercado (Cobranzas Beta, 2026).

Estructura Organizacional¹

Figura 1

Organigrama Promociones y Cobranzas Beta S.A.



Nota. Creación propia

El organigrama está en actualización teniendo en cuenta el proceso de restructuración que se adelanta al momento del proyecto.

La estructura de la empresa está diseñada para garantizar eficiencia en la gestión de cobranza:

- Dirección general: encabezada por la directora de la empresa en supervisión del vicepresidente del área de cobranzas en Davivienda, también se va dividiendo por articuladores encargados de diferentes áreas o segmentos de carteras.
- Área de transformación digital y analítica: es un área transversal encargada de optimizar las demás áreas encargadas de la operatividad o las de gestión de cobro de carteras.
- Recursos Humanos: Maneja la capacitación y gestión del talento.
- Área jurídica: encargada de cobro junto con los procesos legales según sea el caso
- Unidad operativa: encargada de todos los temas operativos que requiera la empresa como contabilidad o gestión de procesos
- Áreas de cobro: son encargadas de gestionar la recuperación monetaria de los portafolios de deuda, segmentado por las fechas de mora.

¹ Por movimientos internos de restructuración de áreas (unión de áreas e integración de otras) aún no hay organigrama actualizado.

Productos o servicios ofertados

Promociones y Cobranzas Beta opera bajo un modelo de negocio especializado en la compra y gestión de cartera vencida del sector financiero. Para ejecutar esta operación, la empresa despliega un portafolio integral de soluciones de cobranza que incluyen contactabilidad estratégica, negociación de acuerdos de pago, refinanciamiento de deudas y análisis predictivo del comportamiento del deudor, asegurando equidad y transparencia para los deudores durante el proceso.

Sus principales servicios incluyen:

a. Gestión de Cobros de Siguiente Nivel

Se especializa en el desarrollo comercial en el campo de recuperaciones de cuentas y recuperación de portafolios ya dados de baja. Además, se desarrollan estrategias personalizadas para cada cliente a través de un modelo de contactabilidad y análisis de datos, permitiendo que los esfuerzos de recolección prioricen a quienes tienen mayor probabilidad de pagar de manera segura.

b. Cómo negociar y refinanciar deudas

Promociones y Cobranzas Beta realiza procesos de negociación de deudas, planes de descuentos, y refinanciamiento acorde con la capacidad de pago del cliente. Estas estrategias tienen como objetivo evitar acciones legales sobre la deuda y mejorar las tasas de recuperación.

c. Análisis y Segmentación de Deudores

Optimización de Procesos de Cobranza a través de la Automatización

Chatbots, sistemas de gestión omnicanal, y sistemas automatizados de recordatorios son solo algunos ejemplos de las soluciones tecnológicas que se utilizan para minimizar costos operativos y mejorar la contactabilidad con los deudores.

d. Cumplimiento Normativo y Transparencia en Cobranza

Está al día con todas las regulaciones vigentes, que establece que la comunicación con los clientes debe ser transparente, justa, y no acosadora financieramente (Ley 2300 de 2023).

Análisis del sector

La gestión de cobranza en el sistema financiero se desarrolla en un entorno dinámico que busca equilibrar el acceso al crédito con la recuperación eficiente de la cartera. En los últimos años, este entorno ha evolucionado de manera significativa como resultado de cambios regulatorios, avances tecnológicos y transformaciones en el comportamiento de los deudores. Al considerar que este proyecto se desarrolla en el área de analítica de datos de Promociones y Cobranzas Beta, se hace necesario analizar de forma estructurada los factores externos que inciden en ella. Para tal propósito, se aplica el modelo PESTEL como herramienta de análisis del entorno, con un enfoque particular en aquellos factores políticos, económicos, sociales, tecnológicos, ambientales y legales que afectan la capacidad del área analítica para diseñar, implementar y sostener un modelo predictivo de contactabilidad alineado con las restricciones normativas y las condiciones operativas de la empresa.

Factores legales y regulatorios.

Las leyes que gobiernan los procesos de cobranza han experimentado transformaciones orientadas a la protección integral del consumidor y la salvaguarda de su privacidad. En Colombia, este marco regulatorio se fundamenta en la Ley 1581 de 2012 (Ley de Protección de Datos Personales), que establece los principios para el tratamiento de información sensible y el derecho de *habeas data*, y se complementa con la Ley 2300 de 2023, la cual limita la frecuencia, horarios y canales de contacto para promover una comunicación menos invasiva. A nivel internacional, normativas como la *Fair Debt Collection Practices Act* (FDCPA) sirven como referencia para la adopción de estas buenas prácticas. Para el equipo de analítica, este entorno legal no es solo un marco restrictivo, sino un determinante del diseño técnico, ya que obliga a integrar

protocolos de seguridad, cifrado de datos sensibles y reglas de cumplimiento normativo directamente en el núcleo de los modelos predictivos.

Factores tecnológicos.

La industria de cobranza ha incorporado de manera progresiva tecnologías como la inteligencia artificial y la automatización de procesos para optimizar la gestión operativa. Estudios previos señalan que el uso de modelos predictivos puede mejorar la efectividad de recuperación entre un 30% y un 50%, al permitir una mejor priorización de acciones y recursos (Zeng et al., 2008). No obstante, estos avances también plantean retos asociados a la calidad de los datos, la integración de sistemas de información, posibles sesgos derivados del entrenamiento de los modelos y la capacidad organizacional para adoptar soluciones analíticas avanzadas.

Entre las herramientas tecnológicas que sostienen la operación de las casas de cobranza, los sistemas de gestión de relaciones con el cliente (Customer Relationship Management, CRM) ocupan un lugar central. Un CRM es una plataforma tecnológica que centraliza la información de los clientes y registra de manera estructurada cada interacción —llamadas, mensajes, promesas de pago, acuerdos y resultados de gestión—, integrando los procesos de mercadeo, ventas y servicio en una única fuente de datos (Payne & Frow, 2005). Desde una perspectiva conceptual, el CRM comprende tres dimensiones complementarias: una operativa, que automatiza y estandariza los flujos de contacto y seguimiento; una analítica, que explota la información histórica para segmentar clientes y anticipar comportamientos; y una colaborativa, que articula los distintos canales de comunicación con el cliente (Buttle, 2009).

En el ámbito específico de la cobranza, los CRM especializados gestionan la cartera asignada, controlan la trazabilidad de los intentos de contacto, administran las reglas de frecuencia y horarios permitidos, y consolidan el estado de cada obligación a lo largo del

embudo de recuperación (contacto, promesa de pago y pago efectivo). Esta capacidad de registro estructurado es precisamente la que provee los datos históricos sobre los que se entrenan los modelos predictivos de contactabilidad. Para el presente proyecto, el CRM cumple un doble rol: por una parte, es la fuente de los registros de gestión utilizados en el modelamiento; por otra, constituye el sistema operativo donde, en una eventual implementación, se consumiría el ranking de priorización generado por el modelo, traduciendo las probabilidades de contacto en acciones concretas (canal y franja horaria sugeridos) dentro de las restricciones de la Ley 2300 de 2023. De este modo, el CRM no es un componente accesorio, sino la plataforma que articula la analítica predictiva con la ejecución operativa de la estrategia de contacto.

Factores Políticos

Los factores políticos inciden indirectamente en la dinámica de la cartera morosa a través de decisiones macroeconómicas y reformas estructurales que afectan el ingreso disponible de los hogares y las condiciones del crédito. En Colombia, decisiones gubernamentales relacionadas con el incremento del salario mínimo, ajustes al Índice de Precios al Consumidor (IPC), reformas laborales y políticas de protección social inciden en la capacidad de pago de los deudores. Un aumento significativo en el salario mínimo puede mejorar transitoriamente la liquidez de ciertos segmentos poblacionales; sin embargo, también puede generar presiones inflacionarias que deterioren el poder adquisitivo real, afectando la estabilidad financiera de los hogares.

Adicionalmente, las decisiones de política monetaria adoptadas por el Banco de la República, particularmente aquellas relacionadas con el incremento de tasas de interés para controlar la inflación, influyen en el costo del crédito y en el crecimiento de la cartera vencida. Un entorno de tasas altas tiende a incrementar el riesgo de mora, especialmente

en créditos de consumo y microcrédito, lo cual impacta directamente la gestión de cobranza y la efectividad de los modelos de contactabilidad.

En este contexto, la tasa de usura certificada periódicamente por la Superintendencia Financiera de Colombia constituye un referente regulatorio relevante, pues establece el límite máximo permitido para el cobro de intereses (Superintendencia Financiera de Colombia, 2023). Aunque en etapas avanzadas de mora la estructura de la deuda puede incluir costos adicionales distintos al interés corriente, el entorno de tasas elevadas incrementa el saldo adeudado y, por tanto, la complejidad del proceso de recuperación. Esto refuerza la necesidad de estrategias analíticas que prioricen de manera eficiente los contactos y optimicen la probabilidad de pago dentro del marco regulatorio vigente.

Factores Sociales

Desde el punto de vista social, la gestión de la cartera morosa se encuentra influenciada por la percepción ciudadana frente al endeudamiento, la cultura de pago y el nivel de educación financiera. En contextos de tasas de interés elevadas, los deudores pueden experimentar mayor presión financiera, lo que impacta tanto su disposición como su capacidad real de pago. Sin embargo, estrategias de negociación que contemplen descuentos, reducción de intereses o acuerdos de pago flexibles pueden generar efectos positivos en la regularización de la deuda y en la rehabilitación del historial crediticio del cliente.

La normalización de obligaciones en mora no solo tiene un impacto financiero individual, sino también un efecto social más amplio, al permitir que los usuarios recuperen acceso al sistema financiero formal y mejoren su perfil de riesgo crediticio. En este sentido, un modelo de contactabilidad basado en analítica avanzada no solo busca incrementar la eficiencia operativa, sino también facilitar procesos de regularización más oportunos y menos invasivos, alineados con una gestión responsable del crédito.

Factores Ambientales

En el contexto de la intervención en el área de analítica de datos, los factores ambientales adquieren relevancia técnica a través del elevado consumo energético derivado de las capacidades de cómputo requeridas para las fases de procesamiento, entrenamiento y ejecución de modelos predictivos complejos, tales como Random Forest, CatBoost entre otros. Esta demanda energética, que se manifiesta tanto en la infraestructura local (servidores y estaciones de trabajo) como en la utilización de centros de datos en la nube, constituye el principal impacto ambiental indirecto de las operaciones del equipo analítico. Asimismo, la digitalización de la cobranza implica una dependencia tecnológica sostenida del uso en otras áreas de herramientas esenciales para la organización, como la gestión de CRMs, el mantenimiento de bases de datos masivas, sistemas de automatización y tableros de visualización, lo que incrementa la huella de carbono institucional. El factor ambiental no constituye un determinante crítico en la viabilidad del modelo de contactabilidad; sin embargo, su consideración permite ampliar la visión estratégica del proyecto hacia un enfoque más integral. En este sentido, vale precisar que el entrenamiento de los modelos se realizó íntegramente en Google Colab, sin requerir infraestructura de cómputo especializada ni centros de datos propios, lo que implica un impacto computacional acotado frente a proyectos de mayor escala.

A modo de referencia cuantitativa, Strubell et al. (2019) estimaron que el entrenamiento de un único modelo de procesamiento de lenguaje natural de gran escala, incluyendo la búsqueda de arquitecturas e hiperparámetros, puede llegar a emitir alrededor de 284 toneladas de CO₂ (cerca de 600.000 libras), una cifra comparable a las emisiones generadas por cinco automóviles durante toda su vida útil. Estas magnitudes corresponden a escenarios de cómputo intensivo y contrastan con el alcance del presente proyecto, en el que se entrenaron seis algoritmos sobre una muestra de 50.000

registros en Google Colab, sin búsqueda masiva de arquitecturas ni infraestructura dedicada, lo que implica una huella considerablemente menor. No obstante, el dato evidencia que la dimensión ambiental del modelamiento no es despreciable y que su magnitud depende de factores como la duración del entrenamiento, el tipo de hardware y, de manera determinante, la matriz energética de la región donde se ejecuta el cómputo (Lacoste et al., 2019). En el caso colombiano, la alta participación de fuentes hidroeléctricas en la generación eléctrica reduce la intensidad de carbono por kWh frente a referentes como el promedio estadounidense empleado en dichos estudios, lo que matiza el impacto ambiental atribuible al proyecto. Esta consideración refuerza la pertinencia de privilegiar modelos con una buena relación entre desempeño y costo computacional, criterio que orientó la selección final de CatBoost, como práctica alineada con un enfoque de inteligencia artificial sostenible.

Diagnóstico interno de la empresa

El análisis interno para el área de analítica de datos de Promociones y Cobranzas Beta se desarrolla a partir de la revisión documental de informes internos, y la observación directa de los procesos operativos relacionados con la gestión de cobranzas. Este análisis permite identificar las principales fortalezas, debilidades, oportunidades y amenazas que inciden en la viabilidad y alcance de la intervención propuesta.

Análisis interno (DOFA)

Es importante precisar que el análisis DOFA desarrollado en este proyecto se centra específicamente en el área de analítica de datos de Promociones y Cobranzas Beta, en tanto constituye el núcleo de la intervención propuesta. Si bien las dinámicas organizacionales generales influyen en el desempeño de la empresa, el alcance del presente proyecto se delimita a aquellas fortalezas, debilidades, oportunidades y amenazas que impactan directamente la capacidad del área analítica para diseñar, implementar y sostener un modelo predictivo de contactabilidad alineado con los objetivos estratégicos y las exigencias normativas de la organización.

Fortalezas.

Promociones y Cobranzas Beta presenta una estructura organizacional sólida, evidenciada en la reciente creación de un área de analítica conformada por un equipo multidisciplinario integrado por un ingeniero de datos, científicos de datos, dos analistas BI y un líder de área. A esto se suma la incorporación de un equipo de programadores en formación, lo cual refleja una apuesta institucional por el fortalecimiento técnico y analítico. Esta capacidad instalada facilita el acceso a fuentes de información confiables, el desarrollo de modelos analíticos y la validación colaborativa de resultados, constituyéndose en un soporte clave para la ejecución del proyecto.

Debilidades.

Como principal debilidad se identifica la dificultad para integrar los procesos analíticos con áreas operativas más tradicionales, situación que se ve acentuada por la estructura jerárquica y los procesos burocráticos derivados de la relación de la empresa con el Banco Davivienda, del cual es filial directa. Adicionalmente, persisten retos asociados a la dispersión en la gestión de los datos y a la falta de una estandarización completa de la información relevante para el análisis analítico.

Análisis externo

Oportunidades.

El avance regulatorio representado por la Ley 2300 de 2023 genera una necesidad urgente de reformular las estrategias de contactabilidad, lo cual puede actuar como un catalizador para cambios positivos en la gestión de cobranza. Asimismo, la madurez tecnológica de herramientas de inteligencia artificial y visualización de datos abre oportunidades para rediseñar modelos de segmentación y priorización más eficientes, alineados con criterios de cumplimiento normativo y mejora en la experiencia del deudor.

Amenazas.

A nivel corporativo, se identifica como una amenaza la alta dependencia de la compra de cartera al Banco Davivienda, lo que limita la autonomía estratégica de la organización. Esta amenaza se ve reducida por la preferencia del banco por darle prioridad de la venta de sus portafolios de deuda de sus carteras castigadas a Promociones y Cobranzas Beta.

Igualmente, esta situación no representa una amenaza directa para la continuidad técnica del área de analítica de datos. Al ser esta una unidad transversal orientada a la optimización operativa, esta posee la capacidad de aplicar sus modelos predictivos y de priorización a diversos portafolios, lo que garantiza la viabilidad y el valor de la intervención analítica independientemente del origen de los activos financieros. A nivel de entorno competitivo, la irrupción de Fintechs y plataformas digitales que emplean modelos de segmentación automatizada y estrategias de contacto dinámico eleva el estándar de eficiencia en el sector de cobranzas. Estas entidades operan con estructuras ágiles que fuerzan una transformación digital acelerada, lo cual representa un desafío para organizaciones que aún dependen de metodologías tradicionales.

El presente proyecto responde directamente a este desafío, pues busca dotar al área de capacidades analíticas (Machine Learning) que permitan no solo alcanzar los niveles de eficiencia de los nuevos competidores, sino superarlos al integrar de forma robusta el rigor ético y el cumplimiento de la Ley 2300 de 2023 en el núcleo del modelo.

Marco Teórico

Transformación de la cobranza: del contacto masivo a la decisión informada.

La gestión de cobro en las casas de cobranza en el sector financiero ha tenido una mezcla de operaciones lógicas tradicionales tanto estadísticas como manuales, pero buscando un enfoque en el cual se pueda apoyar con la tecnología o nuevas metodologías. A pesar del avance de la tecnología y la disponibilidad de datos, muchas organizaciones aún no han migrado sus bases de datos locales a entornos de computación en la nube que faciliten el tratamiento de los datos, independiente del proveedor. Esta brecha entre potencial tecnológico y realidad operativa genera ineficiencias importantes. Como lo plantea Montoya Yepes (Montoya, 2019), la falta de sistematización puede llevar a que procesos que deberían tardar segundos se prolonguen por semanas o se preste para malas prácticas tanto por error humano o tendencia de “echar mano” para que las cifras cuadren. Por lo que la transición a nuevas tecnologías y ajuste de metodologías en la cobranza basada en analítica no es solo una mejora operativa, sino una reconfiguración de la lógica de toma de decisiones.

Analítica avanzada en cobranzas: más allá del pronóstico, una estrategia

Nuevas herramientas avanzadas como el uso de machine learning y sus variantes como redes neuronales y procesamiento de lenguaje natural (NLP) puede permitir el anticipo en comportamientos de pago de los clientes y optimizar la asignación de recursos en la gestión, donde lo menciona en estudios como el de (Cifuentes et al., 2023), se evidencia que modelos como Random Forest, XGBoost o redes neuronales logran estimar con precisión superior al 60% la probabilidad de que un cliente en mora aumente su tiempo de impago. No se trata solo de predecir, sino de intervenir con

criterios objetivos (Cifuentes et al., 2023). Un modelo analítico efectivo permite establecer qué cliente tiene propensión mayor o menor al pago en calidad de deterioro.

Modelos de clasificación, score de cobranza y priorización inteligente

En términos técnicos, los modelos predictivos aplicados a cartera en mora emplean algoritmos supervisados (clasificación binaria) para estimar la probabilidad de que un cliente "ruede" a un tramo más avanzado de mora. Como lo demuestran Cifuentes et al. (2023), la integración de variables de comportamiento transaccional, perfil de riesgo y uso del crédito permite construir un score de cobranza útil para segmentar clientes por prioridad. Esta aproximación permite apoyar a un experto del negocio a la mejora de toma de decisiones con estrategias basadas en datos reales y actualizados, ya sea reafirmando su hipótesis o planteando alternativas que a priori no serían tomadas en cuenta.

Visualización y automatización: analítica que se ve y se usa

Tener datos es necesario para “no ir a ciegas” al momento de tomar decisiones, sin embargo, no es suficiente. Para que la analítica transforme la operación, debe estar integrada a las decisiones diarias, haciendo un manejo de buenas prácticas tanto de ETL(Extracción, Transformación y Carga) como de estar alimentando la base de datos para hacer pruebas y no se estanque. Montoya Yepes (Montoya, 2019) diseñó un ecosistema en el que dashboards interactivos, automatización de scripts en R y Python, y procesos de ETL en Bash permitieron no solo monitorear resultados, sino intervenir sobre ellos en tiempo real. Siendo a tiempo real como se ve los problemas para así tener una reacción visible para poder actuar y resolver el inconveniente que se presente con paneles personalizados para los expertos en el área o supervisores teniendo más control de la operación.

La Ley 2300 de 2023: cumplimiento normativo con enfoque de datos

La legislación colombiana ha evolucionado hacia la protección de los derechos del consumidor financiero. La Ley 2300 regula aspectos como los horarios permitidos de contacto, la frecuencia de llamadas y los medios autorizados. Una estrategia de cobranza que no contemple estas restricciones puede ser legalmente riesgosa y reputacionalmente costosa. Integrar estas reglas dentro del modelo analítico, filtrando por franja horaria, canal permitido o número máximo de interacciones permite garantizar cumplimiento desde el diseño. En palabras simples: el modelo no solo debe predecir bien, debe hacerlo dentro de lo permitido.

Segmentación basada en riesgo y valor: más allá de la mora

No todos los clientes en mora son iguales, y tratarlos de forma homogénea es ineficiente. Estudios como los de Rojas Romero (Romero, 2023) y Montoya Yepes (Montoya, 2019) muestran que, al combinar propensión al impago con saldo adeudado, es posible construir matrices de segmentación que orienten la intensidad de contacto. Por ejemplo, un cliente con alta probabilidad de deterioro y un saldo relevante debería ser gestionado por canales intensivos (como el equipo interno), mientras que un cliente con baja probabilidad y saldo marginal puede ser gestionado por medios automatizados sin contacto humano. Esta lógica permite maximizar el retorno operativo sin sacrificar cumplimiento ni calidad.

Retos operativos y culturales en la adopción de modelos analíticos

La transición hacia una cobranza basada en datos no es exclusivamente técnica. Como lo señala Mesa (Mesa, 2022), muchas casas de cobranza enfrentan limitaciones estructurales: equipos no formados en analítica, baja calidad de datos, resistencia al

cambio y cultura centrada en la acción operativa más que en la estrategia. Superar estos retos requiere una visión institucional que entienda la analítica no como un lujo, sino como un nuevo estándar de eficiencia y transparencia. El modelo debe ser útil, replicable y comprensible para quien lo opera.

La gestión de cartera en etapas tempranas de mora depende directamente de la oportunidad y calidad del contacto con el deudor. Tradicionalmente, las entidades de cobranza han priorizado el número de llamadas realizadas o la intensidad del seguimiento; sin embargo, la literatura evidencia que estas estrategias no garantizan mejores niveles de recaudo si no están acompañadas de criterios de priorización basados en probabilidad de contacto, probabilidad de promesa y probabilidad de pago efectivo (Tran et al., 2025). En este sentido, el proceso deja de ser reactivo y pasa a ser un sistema de decisión, donde cada acción (llamar, enviar mensaje, esperar, escalar de canal o segmentar) se realiza con base en la probabilidad de generar una respuesta del cliente y no solamente en la carga operativa disponible.

La efectividad del proceso de cobranza se mide principalmente en tres dimensiones: contactabilidad, promesa de pago y pago efectivo, que se comportan como un embudo de conversión (Przybyłek et al., 2025). En el contexto del presente proyecto, el concepto de contactabilidad adquiere una precisión específica: hace referencia a la capacidad de localizar y establecer comunicación con el cliente, independientemente del resultado de esa interacción. No implica necesariamente un acuerdo de pago ni una promesa; el objetivo en esta primera etapa es simplemente que el cliente sea ubicado y atienda el contacto. Esta distinción es especialmente relevante en carteras con más de 180 días de mora, donde los clientes presentan baja calificación crediticia y han transitado por múltiples intentos de contacto previos, lo que hace de su localización en sí misma el principal reto operativo. A partir de ese primer contacto exitoso, el proceso avanza hacia la promesa de pago o "promise to pay (PTP)", cuando el cliente se compromete a un

pago futuro, y finalmente hacia el pago efectivo, que se concreta cuando dicha promesa es cumplida. El objetivo del modelo propuesto es maximizar la probabilidad de esa primera etapa: localizar al cliente en el momento y canal más adecuado, respetando los límites normativos vigentes.

La literatura plantea además que el orden y el momento en el que se contacta al deudor influye directamente en los resultados. Según (Tran et al., 2025), asignar a cada deudor el gestor óptimo en el momento adecuado puede incrementar la recuperación entre un 15 % y 70 %, dependiendo del segmento de mora. Este modelo, denominado Agent Matching Model (AMM), se basa en combinar modelos predictivos de probabilidad de pago con técnicas de optimización matemática que asignan cada cuenta al gestor que maximiza la recuperación esperada, considerando tiempo disponible, carga de trabajo y restricciones legales.

En una perspectiva complementaria, (Przybyłek et al., 2025) modelan la cobranza como un Proceso de Decisión de Markov (MDP), donde la situación de cada cliente cambia dinámicamente en función de las acciones tomadas (contacto, espera, mensaje, visita) y de los resultados obtenidos. El sistema propone acciones óptimas mediante aprendizaje por refuerzo (reinforcement learning), pero añade un componente esencial: todas las decisiones deben respetar las restricciones legales de contacto y los derechos del consumidor. Esto es relevante en el contexto colombiano, donde la Ley 2300 de 2023 establece horarios autorizados, canales permitidos y frecuencia máxima de contacto.

Con base en lo anterior, la investigación plantea que no basta con construir un modelo que prediga si el cliente pagará o no; es necesario desarrollar un sistema que determine cuándo contactar, a través de qué canal, con qué prioridad y sin vulnerar la normativa legal ni generar presión indebida. Esta visión alinea el propósito del presente proyecto con tendencias internacionales de cobranza inteligente (smart collections systems),

donde el contacto deja de ser masivo y pasa a ser estratégico, centrado en datos, optimización y cumplimiento normativo.

El comportamiento del deudor no responde siempre a decisiones racionales financieras. Diversos estudios evidencian que factores como la aversión a la pérdida, el sesgo de presente (preferencia por beneficios inmediatos sobre costos futuros) y la ansiedad asociada a la deuda influyen directamente en la disposición a pagar o evitar el contacto con la entidad (Ibañez et al., 2023). Deudores con capacidad de pago pueden evitar llamadas por miedo a ser juzgados o por la percepción de que aceptar la llamada implica reconocer una obligación imposible de cumplir en el corto plazo.

En consecuencia, el uso de presión excesiva o contacto reiterado puede generar efectos contrarios a los esperados: aumenta la evasión del deudor, incrementa las quejas ante los entes regulatorios y deteriora la reputación de la entidad (SSRN, 2023). Esto fundamenta que los modelos de cobranza modernos no deben enfocarse únicamente en insistir más, sino en contactar mejor: en el momento adecuado, en el canal correcto y con un mensaje que favorezca la disposición del cliente al cumplimiento.

En el contexto colombiano, la Ley 2300 de 2023 regula las relaciones entre consumidores y entidades financieras o de cobranza, estableciendo límites claros respecto al contacto. La ley determina horarios permitidos (lunes a viernes entre 7 a. m. y 7 p. m., sábados hasta la 1 p. m.), prohíbe totalmente el contacto los domingos y festivos, y exige respetar la dignidad del deudor, evitando prácticas de acoso, presión indebida o contacto con terceros no autorizados. Adicionalmente, limita el número de intentos de cobro por día y semana, promoviendo el uso responsable de los datos del cliente.

Estos elementos se alinean con la propuesta de este trabajo, en la medida en que cualquier sistema de priorización de contacto no solo debe maximizar indicadores de eficiencia como PTP (Promesa de Pago) o recaudo, sino hacerlo desde un enfoque de cobranza ética, informada y legalmente sostenible. Tal como lo plantean (Przybyłek et al.,

2025), los modelos predictivos y de optimización en cobranzas deben combinarse con reglas explícitas de cumplimiento normativo dentro del algoritmo, de modo que el sistema nunca sugiera una acción que viole la ley. Incorporar el comportamiento del deudor, la normatividad vigente y principios éticos no es un complemento opcional, sino un componente esencial para que un modelo de contacto sea aplicable en contextos reales, especialmente en entidades bancarias vigiladas y casas de cobranza como Promociones y Cobranzas Beta.

El análisis del riesgo crediticio ha evolucionado desde modelos tradicionales como logistic regression y árboles de decisión hacia técnicas basadas en aprendizaje automático (machine learning), capaces de capturar relaciones no lineales, interacciones entre variables y patrones de comportamiento del deudor. En el contexto de cobranza, esta transición no se limita a predecir la probabilidad de incumplimiento, sino que permite estimar con mayor precisión la probabilidad de contacto efectivo, promesa de pago y pago real, elementos fundamentales para orientar estrategias de gestión (Martínez, 2024).

Uno de los enfoques más relevantes en la literatura es el uso de modelos neuro-difusos (ANFIS), donde se integran redes neuronales y lógica difusa para construir sistemas de scoring interpretables y adaptativos. Autores como Kovalenko (Kovalenko et al., 2021) desarrollaron un modelo de evaluación de deudas vencidas en bancos comerciales utilizando este enfoque, logrando clasificar a los clientes en niveles de riesgo con alta precisión y manteniendo reglas comprensibles para analistas y áreas de riesgos. Este tipo de modelos resulta pertinente cuando la entidad requiere decisiones explicables y auditables, como ocurre bajo regulaciones financieras o normativas como la Ley 2300 de 2023, que exige transparencia en el tratamiento de los datos del cliente y en las decisiones asociadas a la gestión de cobranza.

De manera complementaria, otros autores han utilizado algoritmos como Gradient Boosting, Random Forest, Support Vector Machines y Redes Neuronales Artificiales, demostrando mejoras significativas frente a métodos tradicionales. (Martínez, 2024), en una empresa de cobranza latinoamericana, compararon distintos modelos predictivos bajo la metodología CRISP-DM, identificando que el modelo de Gradient Boosting alcanzó un AUC superior al 0,98 para predecir pagos posteriores a una promesa. Este estudio aporta evidencia de que los modelos predictivos permiten priorizar casos con mayor probabilidad de recuperación utilizando variables operativas como días de mora, intentos de contacto previos, saldo de deuda, frecuencia de pago histórico y tipo de producto.

Además del scoring, la literatura reciente destaca la importancia de combinar modelos predictivos con técnicas de investigación de operaciones (Operations Research) para tomar decisiones más allá de la clasificación. Tran y Tran (2025) plantean un modelo mixto denominado Agent Matching Model (AMM), que integra machine learning para estimar la probabilidad de pago y programación lineal para asignar cada cuenta al gestor más adecuado, maximizando la recuperación bajo restricciones de carga laboral y cumplimiento legal. De modo similar, Chehrazi y Weber (Chehrazi & Weber, 2015) aplican modelos dinámicos de valoración de cuentas morosas de tarjeta de crédito, utilizando aprendizaje por refuerzo y optimización secuencial para decidir si contactar, esperar o cerrar el caso.

Estos desarrollos sugieren que el potencial de los modelos predictivos en cobranza no se limita a determinar quién pagará, sino que pueden alimentar sistemas que decidan a quién contactar, cuándo, por qué canal y de qué forma, optimizando recursos y evitando el contacto innecesario o ilegal. En consecuencia, incorporar esta perspectiva en el presente proyecto permite trascender el enfoque descriptivo y avanzar hacia la

construcción de un modelo de contacto basado en datos, aprendizaje automático y decisiones operativas alineadas con el marco legal colombiano.

La literatura contemporánea en gestión de cobranza ha evolucionado desde el simple uso de modelos de scoring hacia sistemas inteligentes que integran predicción, optimización operativa y restricciones legales. Estos sistemas no solo estiman la probabilidad de pago, sino que recomiendan acciones específicas como cuándo contactar, por qué canal, con qué prioridad, qué gestor asignar y cuándo detener gestiones para evitar presión indebida o incumplimiento normativo (Przybyłek et al., 2025).

Uno de los enfoques más robustos en esta línea es el modelo Agent Matching Model (AMM) propuesto por Tran y Tran (2025). Este sistema combina técnicas de machine learning para predecir la probabilidad de pago de cada deudor con algoritmos de programación lineal entera, con el fin de asignar a cada cliente el gestor más eficaz disponible. El objetivo es maximizar la recuperación esperada, sujeto a restricciones como disponibilidad de asesores, volumen de cuentas y reglas de equidad en la distribución de carga operativa. Los resultados reportados evidencian incrementos de entre 1,5 y 1,7 veces en la recuperación, en comparación con modelos estáticos o asignación manual, lo que demuestra el impacto de integrar predicción con optimización.

Por otra parte, autores como Przybyłek(Przybyłek et al., 2025) plantean el desarrollo de un Smart Debt Collection System (SDCS) basado en el enfoque de Design Science Research (DSR). En este sistema, el proceso de cobranza es representado como un Proceso de Decisión de Markov (MDP), en el cual cada cliente se encuentra en un “estado” según su nivel de mora, historial de contacto, comportamiento pasado y probabilidad de pago. A partir de estos estados, el sistema define acciones posibles como “llamar”, “esperar”, “enviar mensaje”, “escalar a jurídico” o “finalizar gestión”,

utilizando Reinforcement Learning (RL) para aprender qué acción genera la mayor recompensa en términos de promesa o pago.

Una característica distintiva del SDCS es que no permite que el modelo actúe sin control normativo. A diferencia de enfoques tradicionales de reinforcement learning, aquí se integra explícitamente un componente de restricciones legales, éticas y de ciberseguridad, lo cual es directamente aplicable al marco de la Ley 2300 de 2023 en Colombia. En el modelo propuesto, estas restricciones operan como filtros fijos (hard constraints): el sistema evita sugerir llamadas fuera del horario permitido (lunes a viernes 7:00 a.m. - 7:00 p.m., sábados hasta la 1:00 p.m.), bloquea contactos en domingos y festivos, limita el número de intentos por cliente por día y evita el contacto multicanal en la misma semana, todo en cumplimiento de la norma. Desde la perspectiva ética, el diseño privilegia una gestión no invasiva que evita la sobre-gestión de clientes con baja probabilidad de contacto, respetando el principio de trato digno al deudor que establece la regulación. En cuanto a la ciberseguridad, el dataset utilizado corresponde a información de clientes con obligaciones en mora, tratada bajo autorización de la empresa y en cumplimiento de la Ley 1581 de 2012 de protección de datos personales, garantizando que el análisis se realizó sobre datos anonimizados y bajo los protocolos de acceso y confidencialidad vigentes en la organización. Esto convierte el sistema en un modelo híbrido donde la optimización basada en datos convive con reglas normativas, éticas y de seguridad fijas.

Además de los enfoques basados en MDP y programación lineal, existen estudios como los de (Chehraz & Weber, 2015) y (Abe, 2010) que utilizan modelos de valoración dinámica de cuentas morosas, donde se calcula el valor futuro esperado de continuar gestionando a un cliente o detener las gestiones. Esto se relaciona con decisiones del tipo: “¿Conviene llamar nuevamente o es mejor esperar?”, “¿Esta cuenta debe

priorizarse o dejarse en automatizado?”, introduciendo la noción de decisiones
secuenciales con horizonte temporal y costo acumulado.

Diseño metodológico

Este proyecto se enmarca en un enfoque metodológico de predominancia cuantitativa, complementado con conocimiento organizacional derivado de la experiencia directa de uno de los integrantes del equipo en la empresa objeto de estudio y de reuniones con líderes operativos de la organización. Esta elección responde a la naturaleza del problema identificado: la baja eficiencia y limitada efectividad del modelo actual de contactabilidad en procesos de cobranza dentro de Promociones y Cobranzas Beta. El componente cuantitativo se trabajó sobre grandes volúmenes de datos históricos que reflejan el comportamiento operativo y la respuesta de los clientes a distintas estrategias de contacto. Si bien se planteó inicialmente la incorporación de un componente cualitativo formal, este no pudo ejecutarse dentro del periodo del proyecto; el conocimiento del negocio y sus dinámicas internas se obtuvo a través de la experiencia del equipo y de las reuniones con los líderes operativos.

Este enfoque, de predominancia cuantitativa pero complementado con conocimiento contextual del negocio, encuentra respaldo en autores como Hernández (Hernández Sampieri et al., 2014), quienes señalan la pertinencia de integrar la medición de variables con la interpretación del contexto en escenarios de intervención empresarial, donde se busca tanto diagnosticar como proponer acciones de mejora. En este trabajo, dicha interpretación contextual se obtuvo a través de la experiencia directa del equipo en la organización y de las reuniones con líderes operativos, y no mediante instrumentos cualitativos formales.

El tipo de estudio corresponde a un estudio de caso con carácter exploratorio-descriptivo y una orientación aplicada. El componente exploratorio se justifica en la necesidad de mapear y comprender en profundidad los factores que afectan la

efectividad del contacto con clientes en mora, dado que, si bien existen datos históricos, no se cuenta aún con una sistematización completa ni con modelos predictivos validados. Por su parte, la dimensión descriptiva permitirá organizar la información recogida y analizar relaciones causales entre variables. Finalmente, el carácter aplicado se evidencia en la intención práctica del estudio: generar un plan de mejoramiento del modelo de contactabilidad, basado en resultados empíricos, viable y alineado con el cumplimiento de la Ley 2300 de 2023, que regula las prácticas de cobranza en Colombia.

La unidad de análisis corresponde al conjunto de procesos operativos de contactabilidad asociados a la gestión de cartera castigada, particularmente aquella con más de 120 días de mora, gestionada desde la sede principal de Bogotá. Esta unidad incluye tanto los datos estructurados que documentan el historial de contacto, como las estrategias utilizadas por los asesores, los canales empleados, los horarios de contacto, cliente (exposición mora, monto proyectado) y las características del asesor.

Población y muestra

La población objeto de estudio corresponde a los registros históricos de gestión de cobranza asociados a la cartera propia de Promociones y Cobranzas Beta, caracterizada por presentar mayores niveles de dificultad en los procesos de recuperación. Esta cartera fue seleccionada de manera intencional por tratarse de un portafolio gestionado directamente por la empresa, lo que garantiza mayor autonomía operativa, disponibilidad de información y control sobre las estrategias de contactabilidad implementadas.

Dado que el enfoque del estudio es analítico y se basa en el uso de información secundaria, no se contempla la aplicación de instrumentos directos a los clientes en mora. En su lugar, la muestra se conforma a partir de los registros históricos de contacto generados en la gestión de dicha cartera, los cuales incluyen información relevante sobre canal de contacto, franja horaria, resultado de la gestión, tramo de mora, monto

adeudado, antigüedad de la deuda y demás variables operativas necesarias para el desarrollo del modelo de contactabilidad.

Variables de análisis

La variable dependiente principal será la efectividad del contacto, operacionalizada a partir del resultado de la gestión registrada (contacto exitoso, promesa de pago, acuerdo firmado, normalización o pago parcial). Las variables independientes incluirán: canal utilizado, horario del contacto, número de intentos previos, monto de deuda, tramo de mora, perfil del asesor, edad del cliente, días transcurridos desde la asignación del caso, y variables de comportamiento históricas del cliente. Estas variables serán cruzadas mediante técnicas de análisis multivariado para determinar patrones y propensiones al éxito del contacto.

El archivo utilizado “Base.xlsx”, es un dataset detallado y anonimizado de cartera castigada, de Promociones y Cobranzas Beta. Este dataset está asociado a productos financieros, como tarjetas de crédito y marcas aliadas, de entidades como Davivienda, organizada por clientes, productos, saldos y gestión de cobranza.

Incluye información de identificación del cliente, segmentaciones por tipo de producto (mono/multiproducto), rangos de saldo (menores, medios y altos), rentabilidad, días de mora con predominio de más de 180 días, y estados de contacto tales como “sin contacto” o “sin pago reciente”.

Asimismo, registra canales y resultados de gestión (WhatsApp, llamadas, correo, SMS, agente virtual), fechas clave (apertura, castigo, última gestión, promesas), riesgos y situaciones especiales (cartera jurídica, insolvencia, cliente ajustado), y variables operativas como pagos, promesas, número de gestiones y responsables, lo que lo hace útil para análisis, seguimiento y toma de decisiones en procesos de recuperación de cartera.

Metodología para el desarrollo del Objetivo 1

Para el desarrollo de este objetivo se realizará un diagnóstico integral del modelo actual de contactabilidad utilizado por Promociones y Cobranzas Beta, enfocado en la gestión de la cartera propia con más de 120 días de mora. Este diagnóstico se sustenta principalmente en el análisis cuantitativo de los registros históricos de gestión, complementado con el conocimiento operativo del equipo y de los líderes del área, con el fin de identificar brechas operativas, limitaciones técnicas y oportunidades de mejora.

En el componente cuantitativo se realizará un análisis exploratorio de los registros históricos de gestión, considerando variables como canal de contacto, franja horaria, número de intentos, resultado de la gestión, tramo de mora, monto adeudado y antigüedad de la deuda.

El diagnóstico organizacional se complementó con el conocimiento del negocio aportado por uno de los integrantes del equipo, quien cuenta con experiencia directa en la empresa y participó en reuniones con líderes operativos de Promociones y Cobranzas Beta. Este conocimiento informó decisiones clave del diseño, como la estructura de franjas horarias, el respeto al horario de almuerzo de los asesores y las capacidades operativas diarias por canal. Se reconoce que, aunque se planteó un componente cualitativo formal, este no pudo implementarse dentro del periodo de desarrollo del proyecto, por lo que el enfoque metodológico real es de predominancia cuantitativa, sustentado en el análisis de los registros históricos de gestión.

Metodología para el desarrollo del Objetivo 2

El segundo objetivo se orienta a la evaluación comparativa de herramientas analíticas, modelos predictivos y enfoques metodológicos aplicables a la optimización de la contactabilidad en contextos de mora avanzada.

Para ello se realizará una revisión estructurada de literatura académica y técnica relacionada con modelos de clasificación supervisada, scoring de cobranza, modelos de propensión al contacto y segmentación basada en riesgo y valor. Esta revisión permitirá identificar alternativas metodológicas aplicables al contexto de la empresa.

Con base en este análisis se definirá un marco de selección de modelos y herramientas, considerando criterios de desempeño predictivo, interpretabilidad, viabilidad operativa, escalabilidad y cumplimiento normativo.

El resultado de este objetivo será una justificación técnica que sustente la elección de los enfoques metodológicos a utilizar en el desarrollo del modelo predictivo.

Metodología para el desarrollo del Objetivo 3

Con base en el dataset consolidado y en los lineamientos definidos en los objetivos anteriores, se formulará el problema de modelado orientado a estimar la probabilidad de contacto efectivo en la cartera objeto de estudio.

Se definirá la variable objetivo y se realizará la partición del conjunto de datos en subconjuntos de entrenamiento y prueba, garantizando consistencia temporal y representatividad operativa. Se construirá una línea base (baseline) que permita comparar el desempeño de diferentes modelos supervisados, tales como: LogisticRegression, RandomForest, ExtraTrees, XGBoost, LightGBM y CatBoost. La selección del modelo final se realizará mediante la evaluación de métricas como: ROC-AUC, precisión, recall, Matriz de Confusión y F1-score, priorizando no solo el desempeño predictivo sino también la interpretabilidad y estabilidad del modelo en contextos operativos reales.

El resultado de este objetivo será un modelo predictivo validado y documentado, que estime la probabilidad de contacto efectivo y que pueda ser integrado posteriormente al esquema de priorización operacional.

Metodología para el desarrollo del Objetivo 4

Una vez validado el modelo predictivo, se procederá al diseño de un esquema híbrido de priorización que combine el score generado por el modelo con reglas operativas y normativas.

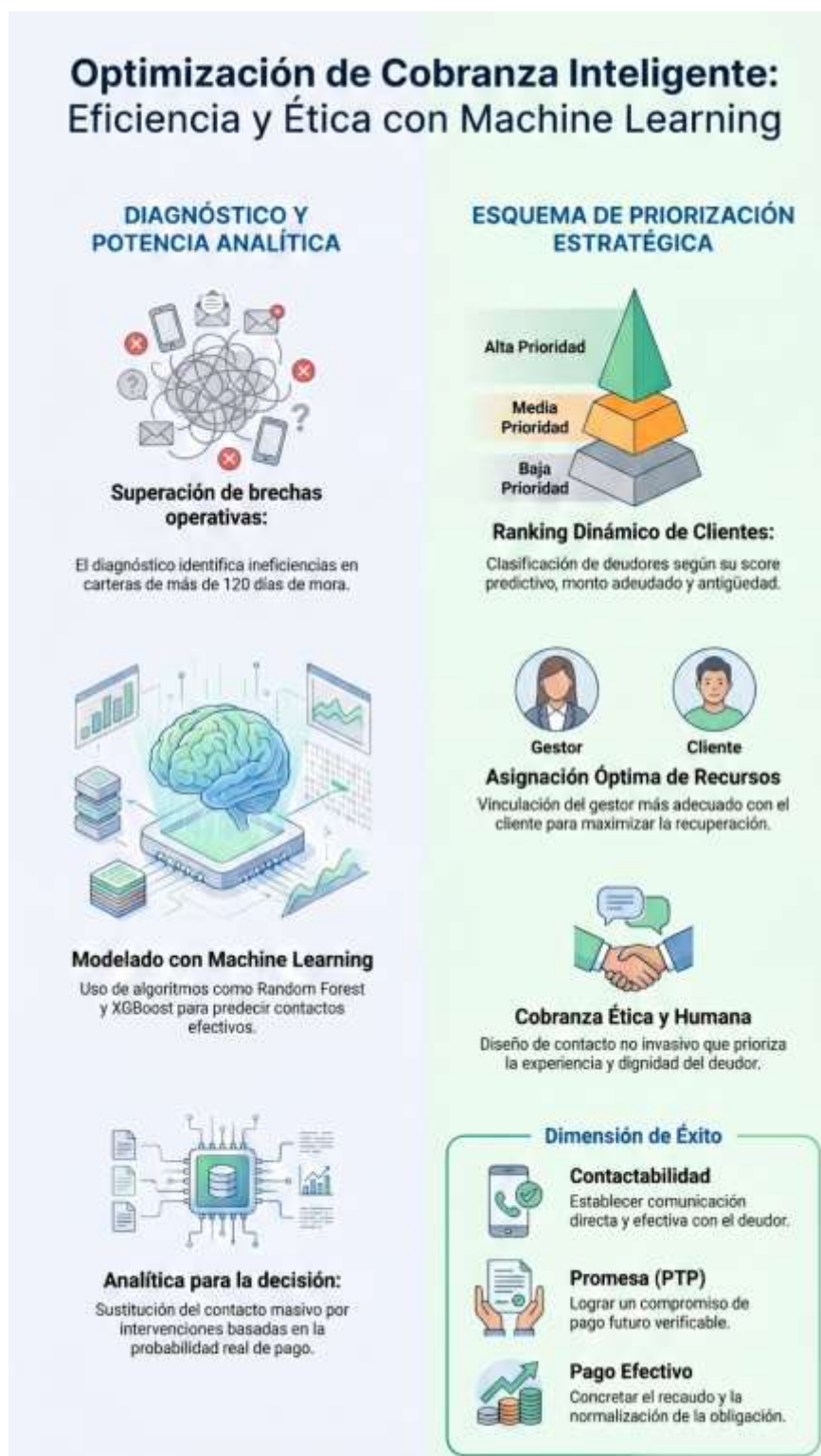
En este proceso se incorporarán explícitamente las restricciones establecidas por la Ley 2300 de 2023, incluyendo límites de frecuencia, horarios permitidos y canales autorizados de contacto. Estas restricciones actuarán como filtros previos al ranking, asegurando que ninguna recomendación derivada del modelo implique acciones contrarias al marco legal vigente.

El esquema de priorización integrará la probabilidad estimada de contacto efectivo con variables de negocio relevantes, tales como tramo de mora, monto adeudado y antigüedad de la obligación, generando un ranking dinámico de clientes que oriente la asignación eficiente de recursos operativos.

El resultado será un diseño documentado del sistema de priorización, trazable y auditable, que sirva como guía para la gestión diaria de cobranza.

Figura 2

Optimización de Cobranza Inteligente



Contribuciones originales esperadas

El proyecto busca aportar una propuesta concreta de mejoramiento para el modelo de contactabilidad de Promociones y Cobranzas Beta, integrando técnicas de analítica avanzada, aprendizaje automático y visualización de datos, bajo el marco de la Ley 2300 de 2023. Las contribuciones que se esperan pueden analizarse desde varios ángulos.

Aportes al conocimiento académico

En el ámbito académico, el trabajo representa un ejercicio aplicado de analítica en el sector de cobranza colombiano, donde aún hay pocos estudios que combinen modelos predictivos, criterios normativos y aspectos organizativos en un solo modelo operativo. Aunque existen investigaciones internacionales que emplean machine learning para mejorar los procesos de cobranza (Appel et al., 2019; Zeng et al., 2008), esta propuesta adapta dichos modelos al contexto local, considerando los retos propios de la normativa y las prácticas empresariales en Colombia. El enfoque, de predominancia cuantitativa y apoyado en el conocimiento de los procesos organizativos, permite una comprensión más amplia de los factores que inciden en la efectividad del contacto.

Aportes a Promociones y Cobranzas Beta

En lo empresarial, se espera que los modelos desarrollados ayuden a priorizar los contactos con los deudores, identificando de forma más certera aquellos casos con mayor probabilidad de pago. Esto permitirá a la empresa optimizar los recursos humanos y operativos, reduciendo los intentos innecesarios de contacto, mejorando los indicadores de efectividad y, en última instancia, aumentando los ingresos por recuperación de cartera.

De igual forma, la incorporación directa de las restricciones de la Ley 2300 dentro de los modelos ayudará a la empresa a operar bajo un marco de cumplimiento normativo

desde el diseño mismo de los procesos. Esto no solo reducirá riesgos legales, sino que proyectará una imagen responsable frente a los clientes, entidades financieras y entes regulatorios.

Aportes al sector y la sociedad

Este trabajo también puede servir como referencia para otras empresas del sector de cobranza que buscan actualizar sus modelos de operación. Al demostrar cómo integrar analítica avanzada dentro de los límites regulatorios, el proyecto ofrece un modelo replicable que promueve mejores prácticas en la industria. Desde una mirada social, el uso más eficiente de los recursos de cobranza podría derivar en negociaciones más oportunas y realistas para los deudores, ayudando a disminuir la presión financiera sobre los hogares y promoviendo acuerdos de pago más sostenibles.

Aporte profesional

Finalmente, el desarrollo de este proyecto representa una aplicación práctica de los conocimientos adquiridos durante la maestría, combinando habilidades de manejo de bases de datos, modelado predictivo, visualización de datos y análisis organizacional, todo dentro de un contexto ético y normativo. Esta experiencia enriquecerá tanto las competencias técnicas como las habilidades de consultoría aplicada en procesos reales.

Fundamentación Teórica

La fundamentación teórica constituye un elemento esencial en una tesis basada en ciencia de datos, ya que establece el marco conceptual y científico que sustenta todo el estudio. A través de esta sección se definen los conceptos clave, permitiendo contextualizar el problema de investigación dentro del estado del arte. De este modo, se garantiza un uso adecuado del lenguaje técnico y se evita la ambigüedad en los términos empleados.

Asimismo, la fundamentación teórica respalda las decisiones metodológicas adoptadas, como la selección de modelos, técnicas de análisis y métricas de evaluación, demostrando que estas no son arbitrarias sino basadas en conocimiento previo validado. Esto contribuye a otorgar rigor académico, credibilidad y coherencia al estudio, además de facilitar la interpretación de los resultados y su comparación.

Fundamentación Teórica de los Modelos Evaluados

En el presente trabajo se evaluaron seis algoritmos de Machine Learning, que abarcan desde enfoques paramétricos tradicionales hasta métodos avanzados de ensamble basados en árboles de decisión. A continuación, se presenta una descripción académica de cada uno de ellos:

1. Regresión Logística (Logistic Regression)

La regresión logística es un modelo estadístico paramétrico ampliamente utilizado para problemas de clasificación binaria. Su objetivo es modelar la probabilidad de pertenencia de una observación a una clase específica mediante la función logística o sigmoide. (Hastie et al., 2001). Desde una perspectiva matemática, el modelo estima los coeficientes de una combinación lineal de variables predictoras que maximizan la

verosimilitud de los datos observados. En este estudio, la regresión logística actúa como un modelo base (baseline), proporcionando un punto de referencia interpretable y lineal frente a modelos no lineales de mayor complejidad. (Hosmer et al., 2013).

Ventajas:

- Alta interpretabilidad de los coeficientes.
- Bajo costo computacional.
- Funciona bien con relaciones lineales entre variables.
- Menor riesgo de sobreajuste en comparación con modelos complejos.

Desventajas:

- Limitada capacidad para capturar relaciones no lineales.
- Sensible a la multicolinealidad.
- Puede presentar bajo rendimiento en problemas complejos.

2. Bosques Aleatorios (Random Forest)

El algoritmo Random Forest es un método de ensamble fundamentado en el paradigma de bagging (Bootstrap Aggregating). Consiste en la construcción de múltiples árboles de decisión independientes a partir de submuestras aleatorias del conjunto de entrenamiento, generadas con reemplazo. En cada nodo, se selecciona aleatoriamente un subconjunto de variables para determinar la mejor división. La predicción final se obtiene mediante votación mayoritaria en problemas de clasificación. Este mecanismo introduce aleatoriedad controlada, lo que reduce la correlación entre árboles, disminuye la varianza del modelo y mitiga el sobreajuste en comparación con un árbol individual (Breiman, 2001).

Ventajas:

- Reduce la varianza y el sobreajuste.
- Maneja bien datos no lineales y de alta dimensionalidad.

- Robusto frente a ruido y valores atípicos.
- Permite estimar importancia de variables.

Desventajas:

- Menor interpretabilidad que modelos simples.
- Mayor costo computacional.
- Puede requerir ajuste de hiperparámetros.

3. Árboles Extremadamente Aleatorizados (Extra Trees)

Extra Trees es una variante del método Random Forest que incrementa el nivel de aleatoriedad en la construcción del modelo. A diferencia de Random Forest, donde se busca el umbral óptimo de división para cada característica, Extra Trees genera umbrales de forma completamente aleatoria y selecciona el mejor entre ellos. Este enfoque simplifica el proceso computacional y, en muchos casos, logra una mayor reducción de la varianza, aunque puede implicar un ligero incremento en el sesgo del modelo (Geurts et al., 2006).

Ventajas:

- Entrenamiento más rápido que Random Forest.
- Reducción adicional de la varianza.
- Menor riesgo de sobreajuste en algunos casos.

Desventajas:

- Ligero aumento del sesgo.
- Interpretabilidad limitada.
- Puede ser menos preciso si la aleatoriedad es excesiva.

4. XGBoost (eXtreme Gradient Boosting)

XGBoost constituye una implementación altamente optimizada del marco de Gradient Boosting. Este algoritmo construye secuencialmente árboles de decisión, donde cada nuevo modelo se entrena para corregir los errores residuales de los modelos anteriores mediante técnicas de descenso de gradiente. Una de sus principales fortalezas radica en la inclusión de términos de regularización (L1 y L2) en la función objetivo, lo que penaliza la complejidad del modelo y previene el sobreajuste. Adicionalmente, XGBoost está diseñado para manejar eficientemente datos dispersos y optimizar el uso de los recursos computacionales, lo que lo convierte en una herramienta altamente robusta y eficiente (Chen et al., 2016).

Ventajas:

- Alto rendimiento predictivo.
- Control del sobreajuste mediante regularización.
- Manejo eficiente de datos grandes y dispersos.
- Amplio soporte y optimización computacional.

Desventajas:

- Mayor complejidad en la configuración.
- Entrenamiento más lento en comparación con modelos simples.
- Sensible a la selección de hiperparámetros.

5. LightGBM (Light Gradient Boosting Machine)

LightGBM es un framework de Gradient Boosting desarrollado por Microsoft, orientado a lograr alta eficiencia y escalabilidad en grandes volúmenes de datos. Su principal innovación es el crecimiento de árboles basado en hojas (leaf-wise), en contraste con el crecimiento por niveles (level-wise) utilizado por otros algoritmos. Este enfoque permite reducir la pérdida de forma más rápida, aunque requiere mecanismos de control para

evitar el sobreajuste. Además, LightGBM emplea técnicas basadas en histogramas para discretizar variables continuas, lo que reduce significativamente el tiempo de entrenamiento y el consumo de memoria (Ke et al., 2017).

Ventajas:

- Alta velocidad de entrenamiento.
- Eficiencia en memoria.
- Escalabilidad para grandes volúmenes de datos.
- Buen rendimiento en problemas complejos.

Desventajas:

- Mayor riesgo de sobreajuste si no se controla.
- Sensible a datos desbalanceados.
- Menor interpretabilidad.

6. CatBoost (Categorical Boosting)

CatBoost, desarrollado por Yandex, es un algoritmo de Gradient Boosting especialmente diseñado para el manejo eficiente de variables categóricas. A diferencia de otros métodos que requieren transformaciones previas como One-Hot Encoding, CatBoost incorpora un enfoque nativo basado en estadísticas del objetivo ordenadas (Ordered Target Statistics), que minimiza el riesgo de fuga de información (target leakage). Asimismo, utiliza árboles simétricos u oblivious trees, en los cuales todos los nodos de un mismo nivel comparten la misma regla de división. Esta característica no solo acelera la inferencia, sino que también actúa como un mecanismo de regularización implícita, mejorando la generalización del modelo (Prokhorenkova et al., 2018).

Ventajas:

- Manejo eficiente de variables categóricas sin preprocesamiento intensivo.
- Reduce el riesgo de fuga de información.

- Buen equilibrio entre sesgo y varianza.
- Inferencia rápida gracias a árboles simétricos.

Desventajas:

- Mayor tiempo de entrenamiento en algunos escenarios.
- Complejidad en la interpretación del modelo.
- Menor adopción comparativa frente a XGBoost (aunque en crecimiento).

Fundamentación Teórica de Precision y Recall

Precision (Precisión) es una métrica fundamental en la evaluación de modelos de clasificación, especialmente en contextos donde los errores por falsos positivos resultan costosos. Se define como la proporción de instancias correctamente clasificadas como positivas respecto al total de instancias que el modelo predijo como positivas.

Matemáticamente, se expresa como la razón entre los verdaderos positivos (TP) y la suma de verdaderos positivos y falsos positivos (FP):

$$Precision = \frac{TP}{TP + FP}$$

Esta métrica evalúa la exactitud de las predicciones positivas del modelo, es decir, qué tan confiables son las clasificaciones positivas generadas (Encyclopaedia Britannica).

Desde una perspectiva conceptual, Precision se asocia con la calidad de las predicciones realizadas, ya que un valor alto indica que la mayoría de las instancias identificadas como positivas son realmente correctas. Por esta razón, es especialmente relevante en aplicaciones como detección de spam, fraude o sistemas de recomendación, donde un exceso de falsos positivos puede afectar significativamente la utilidad del sistema. No obstante, optimizar exclusivamente la precisión puede llevar a

modelos conservadores que evitan clasificar casos como positivos, lo que puede incrementar los falsos negativos (Inferensys).

Recall (sensibilidad o tasa de verdaderos positivos) es otra métrica clave utilizada para evaluar modelos de clasificación, particularmente en situaciones donde es crítico identificar la mayor cantidad posible de instancias positivas. Se define como la proporción de instancias positivas reales que el modelo logra identificar correctamente. Su formulación matemática es:

$$Recall = \frac{TP}{TP + FN}$$

donde FN corresponde a los falsos negativos (Google Developers).

En términos conceptuales, Recall mide la capacidad del modelo para capturar todos los casos relevantes dentro del conjunto de datos. Un valor alto de recall indica que el modelo tiene una baja tasa de omisión de positivos, lo que resulta esencial en dominios como el diagnóstico médico, la detección de fraudes o sistemas de seguridad, donde no identificar un caso positivo puede tener consecuencias graves. Sin embargo, maximizar Recall puede implicar un aumento de falsos positivos, evidenciando el trade-off inherente entre Precision y Recall, lo que hace necesario evaluar ambas métricas de manera conjunta (Inferensys).

Fundamentación Teórica de Validación Cruzada

La Validación Cruzada (cross-validation) es una técnica que se usa para evaluar qué tan bien funciona un modelo con datos que no ha visto. Es decir, en lugar de dividir los datos solo una vez en entrenamiento y prueba, la validación cruzada:

1. Divide los datos en varias partes (llamadas folds).
2. Entrena el modelo varias veces:
 - Cada vez usa una parte diferente como prueba.
 - Y el resto como entrenamiento.

3. Calcula el rendimiento promedio.

Si se usa 5 particiones (5-fold), como en nuestro caso:

1. El modelo se entrena 5 veces.
2. Cada vez cambia el grupo de datos de prueba.
3. Al final se promedian los resultados.

Las ventajas que obtenemos son:

- Evita depender de una sola división de datos.
- Da una evaluación más confiable del modelo.
- Reduce el riesgo de sobreajuste (overfitting).

Fundamentación Teórica del método RandomizedSearchCV

El RandomizedSearchCV es una técnica de optimización de hiperparámetros ampliamente utilizada en ciencia de datos y aprendizaje automático, especialmente dentro de la librería scikit-learn. Su propósito principal es encontrar la mejor combinación de hiperparámetros para un modelo, con el fin de mejorar su rendimiento predictivo. A diferencia de otros métodos como GridSearchCV, que evalúan exhaustivamente todas las combinaciones posibles, RandomizedSearchCV selecciona un número determinado de combinaciones de manera aleatoria a partir de distribuciones definidas por el usuario (Scikit-learn Developers, 2024).

El proceso funciona mediante la generación de múltiples configuraciones de hiperparámetros (controladas por el parámetro `n_iter`), las cuales son evaluadas utilizando validación cruzada. En cada iteración, se entrena el modelo con una combinación distinta y se mide su desempeño con una métrica definida (por ejemplo, accuracy o F1-score). Este enfoque permite explorar de manera más eficiente espacios

de búsqueda amplios o complejos, reduciendo significativamente el costo computacional sin sacrificar la calidad de los resultados (Bergstra et al., 2012)

Una de las principales ventajas de RandomizedSearchCV es que desacopla el tiempo de ejecución del tamaño total del espacio de búsqueda, ya que el usuario define explícitamente cuántas combinaciones evaluar. Además, investigaciones han demostrado que la búsqueda aleatoria puede ser tan efectiva o incluso más eficiente que la búsqueda en rejilla, especialmente cuando solo algunos hiperparámetros influyen de manera significativa en el rendimiento del modelo (Scikit-learn Developers, 2024).

Plan de intervención

Enfoque general del plan de intervención

El presente plan de intervención tiene como propósito central optimizar el modelo de contactabilidad utilizado por Promociones y Cobranzas Beta en la gestión de cartera con más de 120 días de mora, integrando criterios analíticos, operativos y normativos bajo los lineamientos establecidos por la Ley 2300 de 2023. La intervención se concibe como un proceso aplicado y progresivo, orientado no solo a identificar oportunidades de mejora, sino a proponer acciones concretas, viables y alineadas con la capacidad operativa real de la empresa.

El plan se estructura en fases secuenciales que permiten transitar desde el diagnóstico del modelo actual hacia la implementación de un esquema mejorado de priorización de contacto, soportado en analítica avanzada y mecanismos de seguimiento que garanticen su sostenibilidad en el tiempo. La propuesta no busca reemplazar abruptamente los procesos existentes, sino fortalecerlos mediante el uso estratégico de la información disponible, respetando las restricciones regulatorias y la dinámica organizacional propia de la empresa.

Fase 1. Diagnóstico operativo y analítico del modelo de contactabilidad

La primera fase del plan de intervención se orienta a realizar un diagnóstico integral del modelo de contactabilidad vigente. Este diagnóstico combina el análisis de la información histórica de gestión de cobranza con la revisión de los procesos operativos actualmente utilizados por los asesores y supervisores.

Desde el punto de vista analítico, se examinan los registros históricos de contacto para identificar patrones de efectividad asociados a variables como canal de contacto,

franja horaria, número de intentos, tramo de mora y tipo de cartera. Este análisis permite detectar ineficiencias como la sobre gestión de determinados segmentos, el uso reiterado de canales de bajo rendimiento o la realización de contactos en horarios con baja probabilidad de respuesta.

De manera complementaria, el diagnóstico incorpora una revisión de los procedimientos operativos, enfocada en comprender cómo se toman actualmente las decisiones de contacto, qué criterios utilizan los asesores para priorizar clientes y qué limitaciones enfrentan en su gestión diaria. Esta fase concluye con la identificación de brechas claras entre el modelo actual y un modelo de contactabilidad más eficiente, basado en datos y alineado con la normativa vigente.

El dataset 'Base.xlsx', que consolida tanto información sociodemográfica como el historial detallado de gestiones, canales y resultados operativos, constituye el insumo fundamental para la fase analítica del proyecto. No obstante, la transición de estos datos brutos hacia un motor de decisiones exige un proceso de refinamiento técnico que asegure la calidad de las predicciones.

Fase 2. Definición de criterios de priorización y segmentación de clientes

Con base en los hallazgos del diagnóstico, la segunda fase del plan de intervención se centra en la definición de criterios formales de priorización y segmentación de clientes. El objetivo es pasar de un enfoque de gestión homogénea a un esquema diferenciado, donde no todos los clientes en mora reciben el mismo nivel de intensidad ni el mismo tipo de contacto.

En esta fase se establecen segmentos de clientes considerando variables como probabilidad de contacto, probabilidad de promesa de pago, monto adeudado, antigüedad de la deuda y comportamiento histórico. Estos segmentos permiten clasificar

a los clientes según su valor esperado y su disposición potencial a responder a una gestión de cobranza, optimizando así la asignación de recursos humanos y tecnológicos.

Adicionalmente, se incorporan explícitamente las restricciones establecidas por la Ley 2300 de 2023, definiendo criterios claros sobre horarios permitidos, frecuencia máxima de contacto y canales autorizados. Estos criterios se integran como reglas operativas obligatorias dentro del esquema de priorización, garantizando que cualquier acción sugerida por el modelo respete el marco legal desde su diseño.

Este proceso permite transformar registros históricos (como fechas de castigo, montos de promesa y tipos de producto) en características (features) de alto valor predictivo. La definición de una variable objetivo (target) centrada en la efectividad del contacto permite que el modelado deje de ser un ejercicio estadístico descriptivo y se convierta en una herramienta de optimización operativa.

Bajo este enfoque, el desarrollo metodológico que se presenta a continuación describe la implementación de una línea base (baseline) y la evaluación competitiva de diversos algoritmos supervisados. El propósito es identificar el modelo que, partiendo de la estructura de datos de Promociones y Cobranzas Beta, logre el mejor equilibrio entre la precisión técnica y la viabilidad del negocio, garantizando siempre el cumplimiento de la Ley 2300 de 2023."

Se busca desarrollar el modelo predictivo basado en Machine Learning para ello se crea un cuaderno de en Colab de Google

<https://colab.research.google.com/drive/1EhYgpbXwmJsBzELI5y1bfj46cykGN9n3?usp=sharing>

Este ambiente de Colab permitió realizar la búsqueda del modelo del proyecto. A continuación, se explican las características más importantes de cada una de las secciones:

Las secciones “1. Funciones auxiliares” y “2. Carga y selección de columnas” definen varias funciones utilitarias que son esenciales para cargar, limpiar y realizar una exploración básica del archivo Excel “Base.xlsx.

La sección “3. Limpieza y tipificación” es fundamental para preparar los datos antes de cualquier análisis o modelado. Su objetivo principal es asegurar que cada columna tenga el tipo de dato correcto y que los valores sean consistentes, manejando los datos faltantes o inconsistentes.

La sección “4. Exploración visual básica”, se enfoca en la creación de gráficos sencillos para entender mejor la distribución y las características de las variables clave en el conjunto de datos.

La sección “5. Ingeniería de variables y target”, es crucial para transformar los datos brutos en características útiles para el modelo y definir la variable objetivo que el modelo intentará predecir. La variable objetivo, denominada `target_contacto_efectivo`, se define operacionalmente como el registro en que el cliente contestó la comunicación iniciada: es decir, que fue localizado y atendió el contacto (valor 1). Cualquier otro resultado, incluyendo llamada no contestada, número no disponible o cliente no localizado, se codifica como 0. Esta definición es distinta de la promesa de pago (PTP), que ocurre cuando el cliente que ya fue contactado se compromete a un pago futuro, y del pago efectivo, que se concreta cuando esa promesa es cumplida. El modelo predice únicamente la primera etapa de ese embudo: la probabilidad de que el cliente conteste, que en la cartera analizada (más de 180 días de mora) representa el principal reto operativo dado que estos clientes son difíciles de localizar.

La sección “6. Preparación del dataset para ML”, es donde se alistan los datos para ser consumidos por los modelos de Machine Learning. Aquí se definen las características (variables predictoras) y la variable objetivo, se manejan los valores categóricos y nulos, y se divide el conjunto de datos para entrenamiento y prueba.

La sección “7. Comparación de modelos”, se dedica a entrenar y evaluar múltiples algoritmos de Machine Learning, que ya fueron descritos al comienzo de esta sección, para identificar cuál ofrece el mejor rendimiento en la predicción de la contactabilidad efectiva.

Se solicitó al algoritmo la generación de una tabla comparativa de los diferentes modelos, obteniendo como resultado los siguientes valores:

Tabla 2

Comparación de modelos

	modelo	roc_auc	accuracy	precision	recall	f1
0	LightGBM	0,968815	0,9629	0,750000	0,551959	0,635918
1	XGBoost	0,96747	0,9624	0,718427	0,591141	0,648598
2	CatBoost	0,967189	0,9628	0,718941	0,601363	0,654917
3	RandomForest	0,963814	0,8827	0,325595	0,931857	0,482576
4	ExtraTrees	0,952603	0,8573	0,283505	0,936968	0,435299
5	LogisticRegression	0,753221	0,7801	0,146181	0,567291	0,232461

De esta comparación podemos sacar las siguientes conclusiones:

1. CatBoost se destaca como el mejor modelo en general, principalmente por haber logrado la puntuación F1 más alta (0.654917). Esto es crucial en problemas con clases desbalanceadas como la predicción de contactabilidad efectiva, ya que indica un excelente equilibrio entre precisión y recall.
2. Aunque LightGBM tuvo un ROC-AUC ligeramente superior (0.968815) que CatBoost (0.967189), la importancia del F1-score en este contexto lo posiciona justo después.

3. XGBoost también mostró un rendimiento muy competitivo, situándose como el tercer mejor modelo, con un ROC-AUC de 0.967470 y un F1-score cercano al de CatBoost.
4. Los modelos basados en árboles (LightGBM, XGBoost, CatBoost, RandomForest, ExtraTrees) superaron significativamente a la Regresión Logística, lo cual es común en datasets con relaciones complejas entre las variables.

En resumen, CatBoost es el modelo recomendado por su robusto rendimiento y su capacidad para manejar el desbalance de clases, ofreciendo un buen balance entre la identificación de contactos efectivos y la minimización de falsos positivos.

Fase 3. Diseño del esquema de priorización de contacto

La tercera fase del plan de intervención corresponde al diseño del esquema de priorización de contacto, el cual actúa como puente entre el análisis de datos y la operación diaria de la cobranza. Este esquema combina los resultados del análisis predictivo con reglas operativas y normativas, generando un ranking de clientes que orienta las decisiones de contacto de manera estructurada.

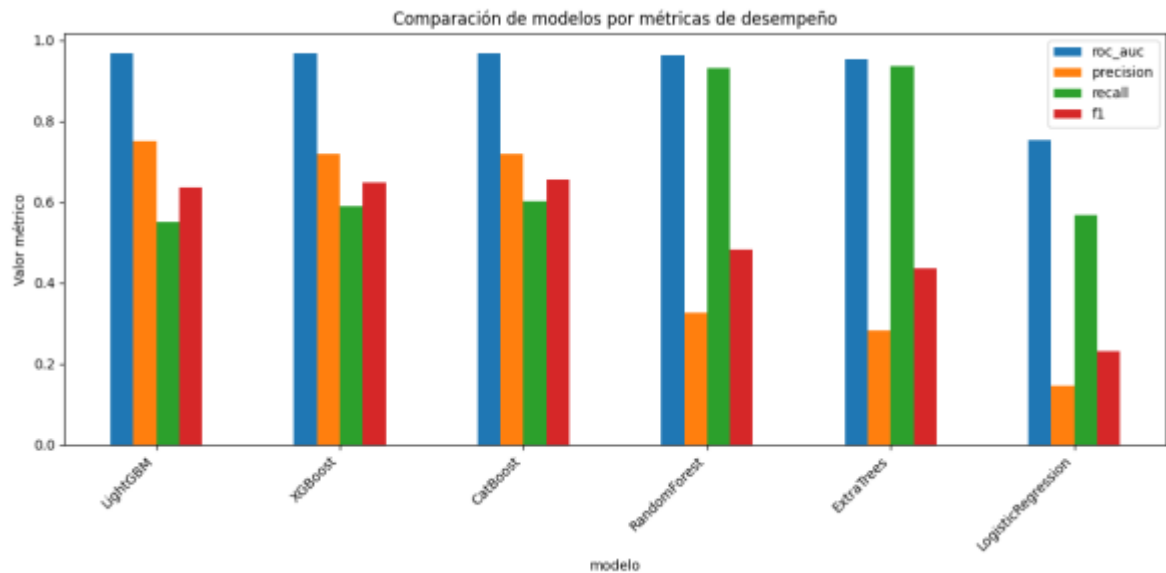
El esquema de priorización define qué clientes deben ser contactados primero, por qué canal y en qué franja horaria, considerando tanto la probabilidad de éxito como la viabilidad legal y operativa de la gestión. De esta manera, se evita la realización de contactos innecesarios o de bajo impacto, reduciendo la carga operativa de los asesores y mejorando la experiencia del cliente.

Este ranking no se concibe como una instrucción rígida, sino como una herramienta de apoyo a la toma de decisiones, que permite a los supervisores ajustar las estrategias de contacto según el comportamiento observado y las condiciones cambiantes de la operación.

La sección “8. Visualización de desempeño de modelos”, es fundamental para entender gráficamente cómo se comportaron los diferentes modelos que se entrenaron. Esto permite una comparación visual rápida y profunda de sus métricas. Podemos extraer las siguientes conclusiones:

Figura 3

Comparación de modelos por métricas de desempeño



1. Consistencia en los Modelos Superiores: Las visualizaciones (gráficos de barras de métricas, curvas ROC y curvas Precision-Recall) reafirman que LightGBM, XGBoost y CatBoost son consistentemente los modelos con mejor desempeño. Sus métricas son superiores y sus curvas ROC y Precision-Recall se encuentran más cercanas a la esquina superior izquierda, indicando una mejor capacidad de discriminación y un mejor equilibrio entre precisión y recall.
2. Importancia de Curvas Precision-Recall: Las curvas Precision-Recall son particularmente reveladoras en este escenario de clases desbalanceadas (pocos contactos efectivos). Modelos como Random Forest y Extra Trees, que en la curva ROC pueden parecer muy buenos, en la curva Precision-Recall muestran cómo su precisión puede caer rápidamente al intentar aumentar el recall, lo que se traduce en muchos falsos positivos. CatBoost, LightGBM y XGBoost mantienen un mejor balance en estas curvas.

Figura 4

Curvas ROC de los modelos

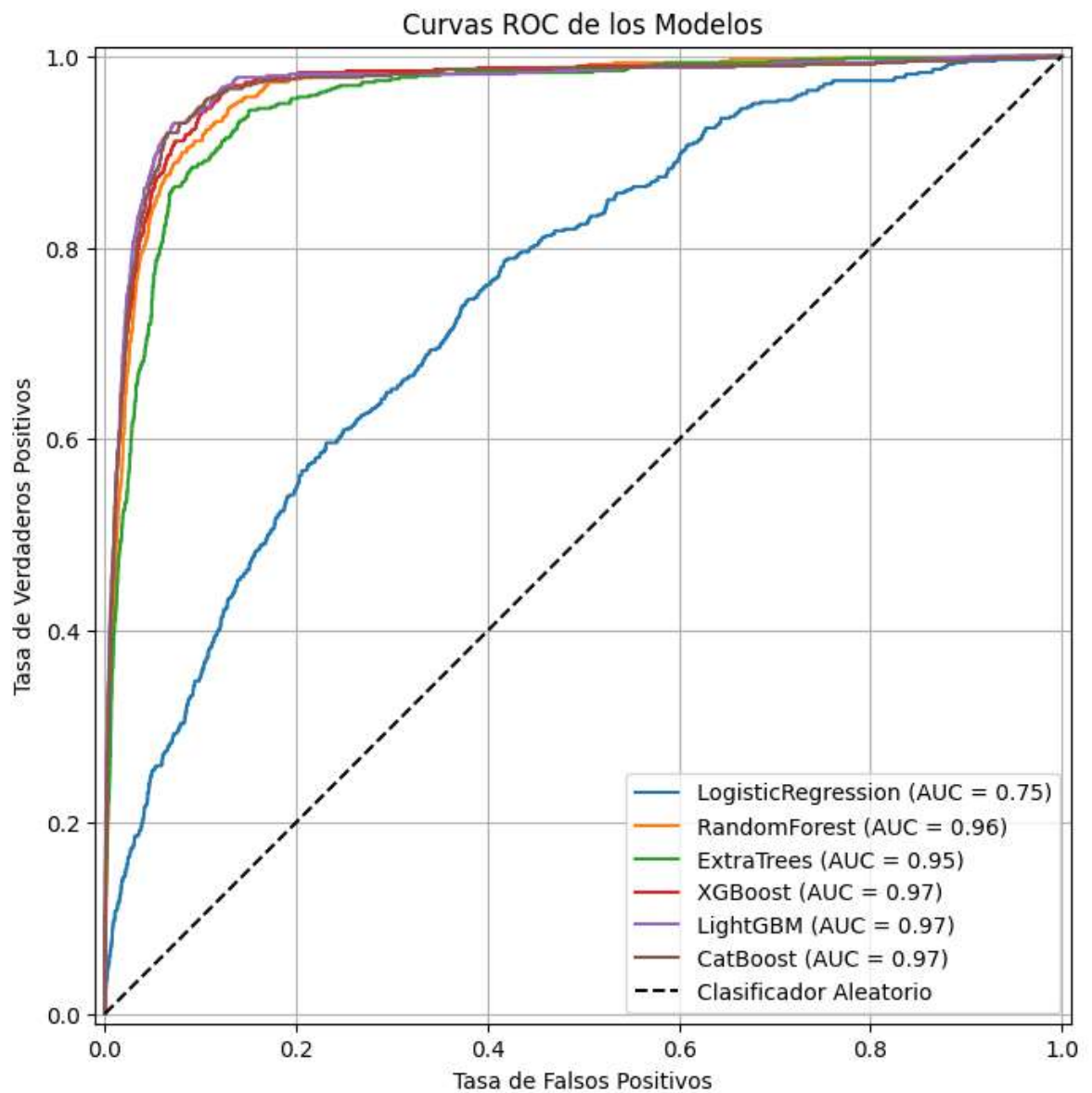
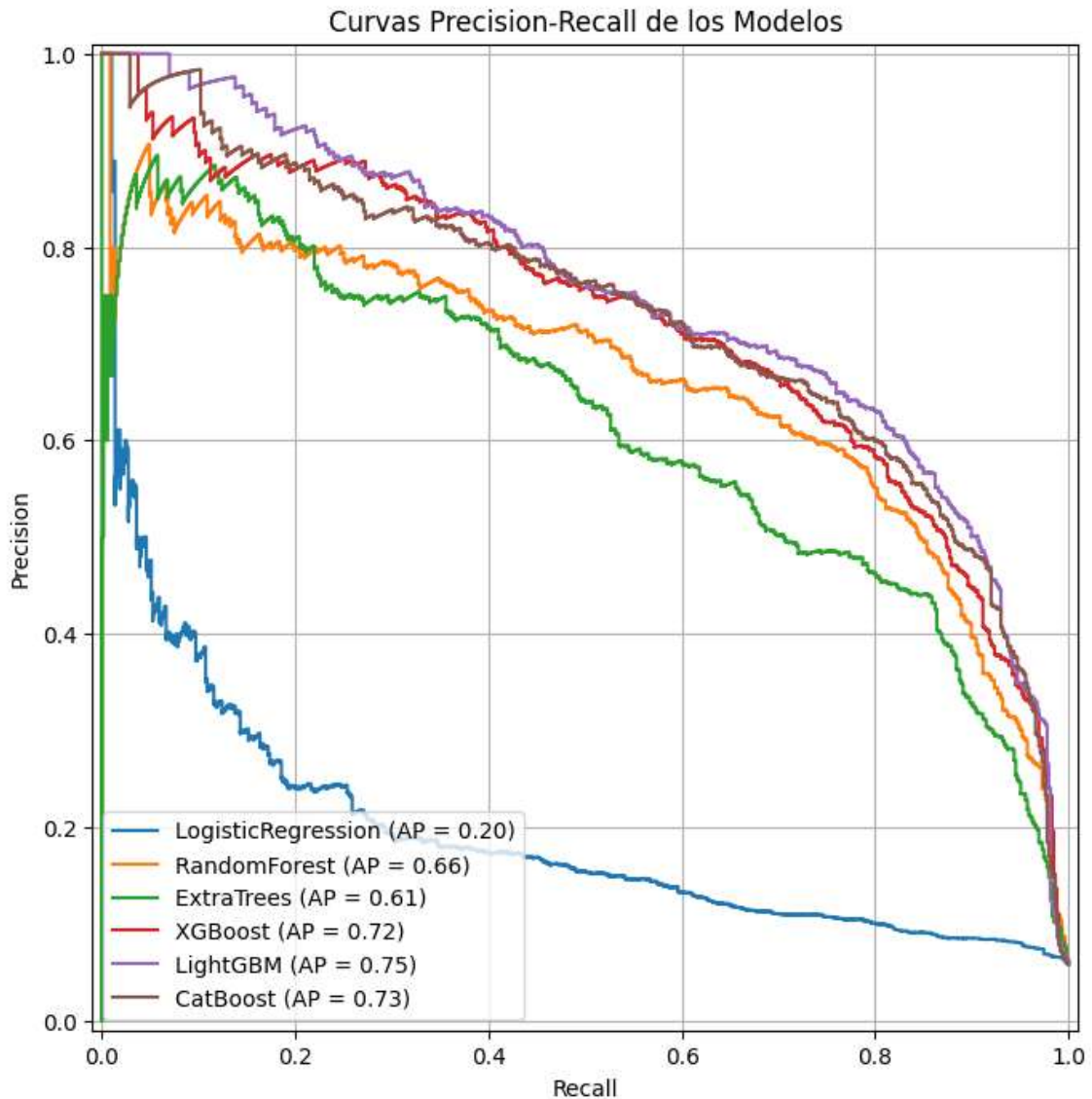


Figura 5

Curvas de Precisión Recall de los Modelos



3. **Análisis Granular con Matrices de Confusión:** Las matrices de confusión permiten un examen detallado de los tipos de errores que comete cada modelo. Por ejemplo, podemos ver el número exacto de verdaderos positivos (contactos efectivos correctamente identificados), falsos negativos (contactos efectivos perdidos) y falsos positivos (contactos inefectivos predichos como efectivos). Esto es crucial para entender las implicaciones operativas de cada modelo en el negocio.

Figura 6

Matriz de Confusión del Modelo Logistic Regression

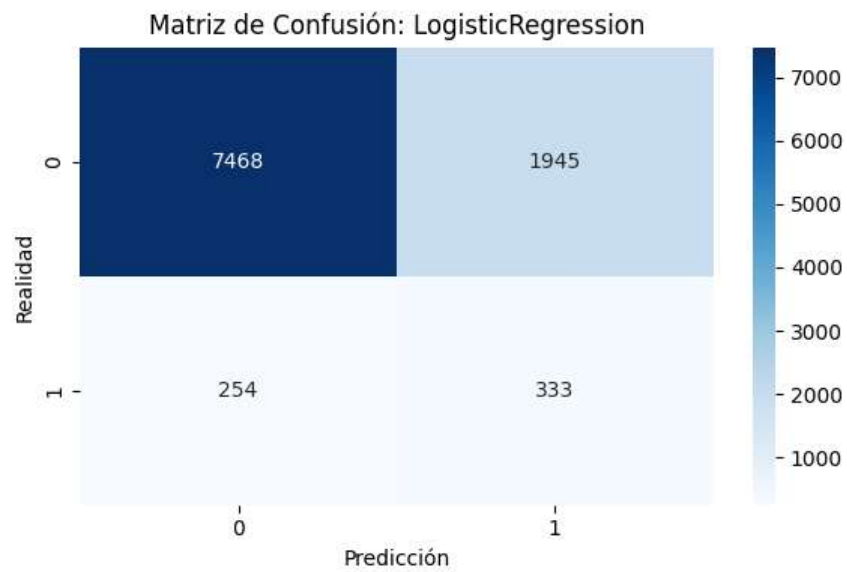


Figura 7

Matriz de Confusión del Modelo RandomForest

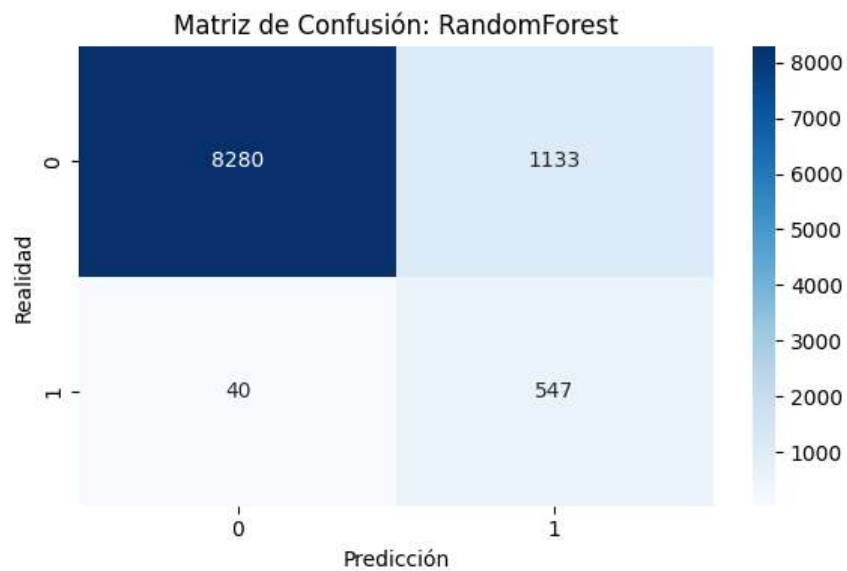


Figura 8

Matriz de Confusión del Modelo ExtraTrees

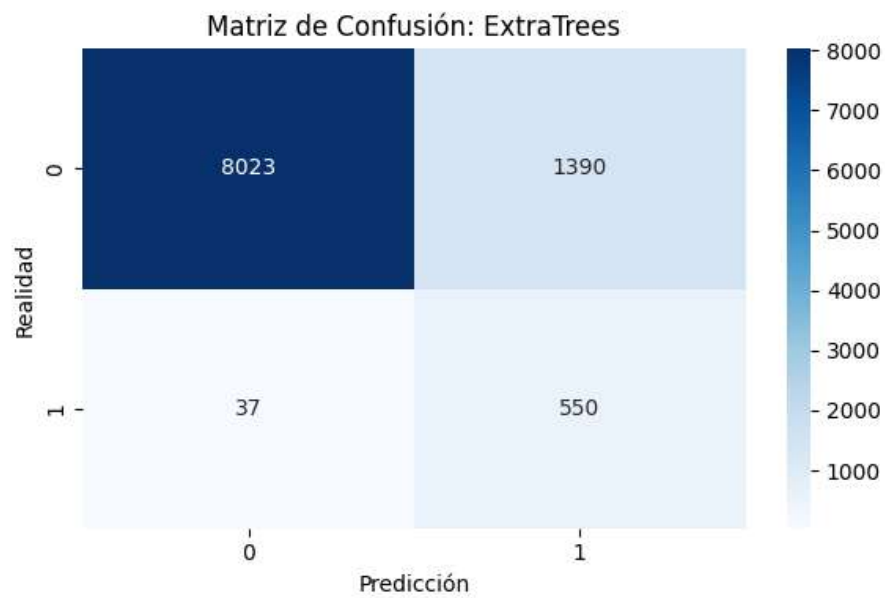


Figura 9

Matriz de Confusión del Modelo XGBoost

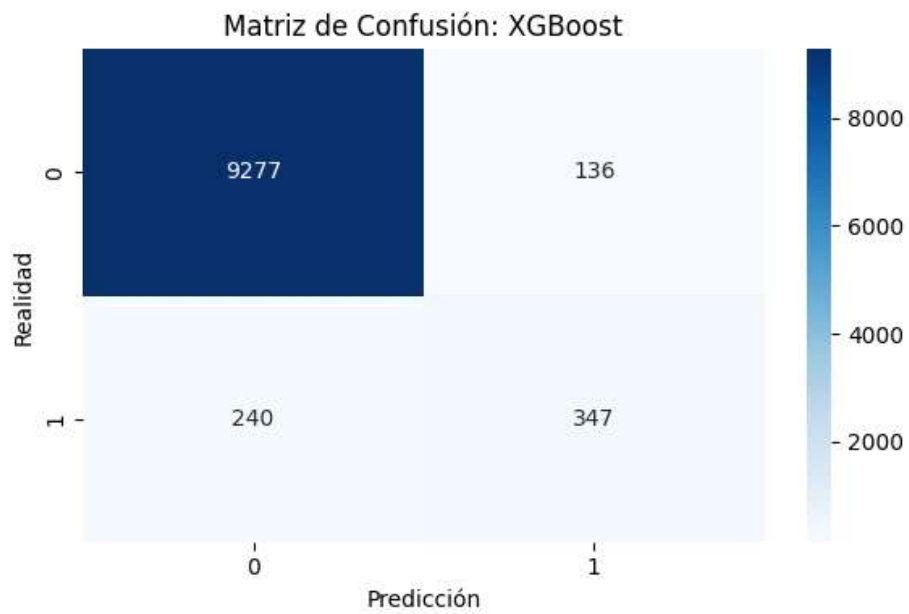


Figura 10

Matriz de Confusión del Modelo LightGBM



Figura 11

Matriz de Confusión del Modelo CatBoost

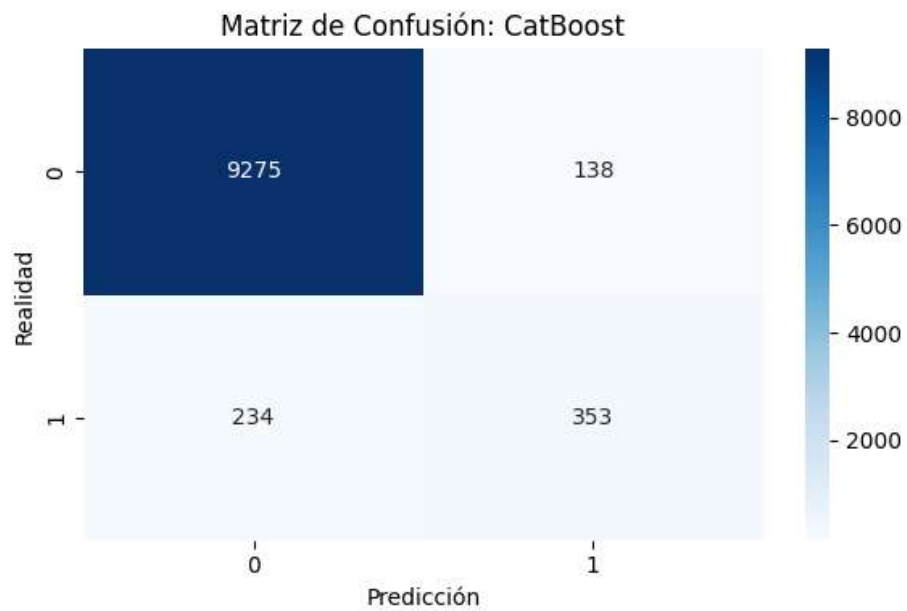
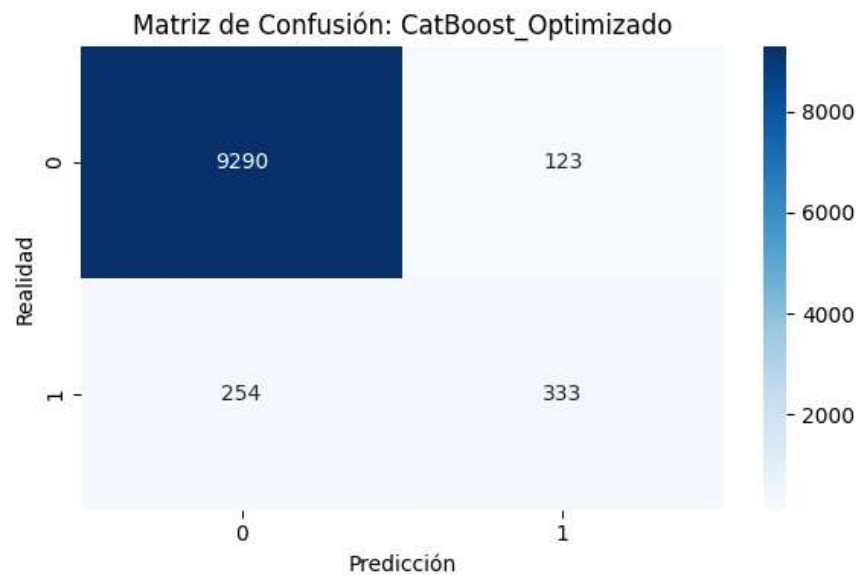


Figura 12

Matriz de Confusión del Modelo CatBoost Optimizado



4. Debilidad de Modelos Simples: La Regresión Logística muestra un rendimiento significativamente inferior en todas las métricas y visualizaciones, lo cual era esperable dado la complejidad del problema y el desbalance de clases.

En síntesis, las visualizaciones confirman la robustez y el buen equilibrio de CatBoost, LightGBM y XGBoost, con CatBoost destacando ligeramente por su F1-score, lo cual es vital para una estrategia de contactabilidad donde se busca eficiencia y efectividad simultáneamente

La sección “9. 'Selección del modelo base”, tiene como objetivo identificar y detallar el rendimiento del mejor modelo elegido tras la fase de comparación, basándose en métricas clave para el negocio. Podemos sacar las siguientes conclusiones clave:

1. Prioridad de Métricas: La sección demuestra la importancia de seleccionar un modelo basándose en métricas relevantes para el problema de negocio. Aunque LightGBM tuvo el ROC-AUC ligeramente más alto (0.968815), la elección final recayó en CatBoost.

2. CatBoost como Mejor Modelo por F1-score: CatBoost fue elegido como el 'mejor modelo' porque obtuvo la puntuación F1 más alta (0.654917). Para este problema de predecir la contactabilidad efectiva, donde la clase positiva es minoritaria, el F1-score es una métrica crucial ya que equilibra la precisión (cuántas de las predicciones positivas fueron correctas) y el recall (cuántos de los contactos efectivos reales fueron identificados).
3. Equilibrio entre Precisión y Recall: Las matrices de confusión y los reportes de clasificación para CatBoost muestran un buen balance. Aunque tiene una precisión del 72% y un recall del 60% para la clase positiva (contactos efectivos), este equilibrio es preferible en este contexto, ya que no se quiere sacrificar demasiado la precisión por un recall muy alto, o viceversa, para que las acciones operativas sean eficientes.
4. Rendimiento en Identificación de Contactos Efectivos: La matriz de confusión del modelo CatBoost muestra que identificó 353 verdaderos positivos (contactos efectivos) y tuvo 234 falsos negativos (contactos efectivos que no predijo), lo que indica una capacidad razonable para identificar la clase de interés. Tuvo 138 falsos positivos (contactos inefectivos que predijo como efectivos).

En resumen, la selección de CatBoost se justifica por su capacidad para ofrecer el mejor balance entre precisión y recall, lo que se refleja en el F1-score, haciéndolo el modelo más adecuado para optimizar las estrategias de contactabilidad.

La sección “10. Validación Cruzada (Cross-Validation)” garantiza la robustez metodológica y demuestra que el rendimiento de los modelos no depende de una partición aleatoria afortunada, se implementa Validación Cruzada Estratificada de 5 pliegues (Stratified 5-Fold Cross-Validation) sobre el conjunto de entrenamiento.

Para nuestro caso:

- Se utiliza Validación Cruzada Estratificada (Stratified Cross-Validation), la cual es una variante de la Validación Cruzada, y que es útil cuando se tiene problemas de clasificación desbalanceada (por ejemplo, detección de fraudes o enfermedades) donde hay muchos casos de una clase y pocos de otra. Dado el alto desbalance de la clase positiva (contactos efectivos ~6%), *StratifiedKFold*

asegura que cada uno de los 5 pliegues mantenga la misma proporción de clases que el dataset original.

- La Métrica evaluada es ROC-AUC, debido a que es la métrica más confiable para evaluar la capacidad de discriminación del modelo en escenarios desbalanceados.

Se solicitó al algoritmo la generación de una tabla comparativa de los diferentes modelos, que resumiera los valores obtenidos como resultado de la validación cruzada.

Tabla 3

Resumen de Validación Cruzada

	modelo	roc_auc_cv_media	roc_auc_cv_std
0	LightGBM	0,966827	0,004378
1	XGBoost	0,963787	0,005930
2	CatBoost	0,960859	0,005514
3	RandomForest	0,956689	0,006993
4	ExtraTrees	0,940638	0,007071
5	LogisticRegression	0,747674	0,012783

Los resultados de la validación cruzada, junto con sus respectivas desviaciones estándar (roc_auc_cv_std), permiten confirmar la estabilidad del modelo seleccionado. En particular, CatBoost presenta la segunda desviación estándar más baja, lo que indica una buena capacidad de generalización frente a distintos subconjuntos de datos. Esto respalda que las métricas obtenidas en el conjunto de prueba original (Test) sean consistentes y fiables.

La sección “11. Optimización de Hiperparámetros (Hyperparameter Tuning)”, atiende al rigor de la experimentación en Machine Learning, ya que no basta con probar los modelos con sus parámetros por defecto. Por ello, se realiza un proceso iterativo de

optimización de hiperparámetros para el modelo seleccionado (CatBoost) utilizando *RandomizedSearchCV*.

Como resultado de la optimización de los hiperparametros del modelo seleccionado, CatBoost, se obtuvieron los siguientes resultados:

Tabla 4

Mejores Hiperparámetros encontrados para CatBoost

Mejores Hiperparámetros encontrados para el modelo seleccionado Catboost	
learning_rate	0,1
l2_leaf_reg	7
iterations	200
depth	10

El Mejor ROC-AUC validado (CV) obtuvo un valor de 0,9666, y el ROC-AUC en conjunto de prueba (Test) con modelo optimizado fue de 0,9684.

La sección “12. Comparativa: CatBoost Base vs CatBoost Optimizado” verifica cómo cambiaron las métricas en el conjunto de prueba (Test) tras aplicar la búsqueda de hiperparámetros. Esto permitirá confirmar si el modelo mejoró su capacidad de generalización o si encontró un mejor balance entre Precisión y Recall.

Tabla 5

Tabla comparativa entre CatBoost y CatBoost Optimizado

modelo	roc_auc	accuracy	precision	recall	f1
CatBoost (Base)	0,967189	0,9628	0,718941	0,601363	0,654917
CatBoost (Optimizado)	0,968411	0,9623	0,730263	0,567291	0,638543

A partir de los resultados que muestra la tabla anterior, se pueden extraer las siguientes conclusiones tras la optimización de hiperparámetros:

1. Mejora en ROC-AUC: El modelo optimizado logra un ligero incremento en la métrica ROC-AUC (pasando de ~ 0.967 a ~ 0.968) sobre el conjunto de prueba. Esto indica una mejora en su capacidad general para discriminar correctamente entre la probabilidad de que un contacto sea efectivo frente a uno que no lo es.
2. Rendimiento Base Sólido: El modelo CatBoost con hiperparámetros por defecto ya presentaba un rendimiento excelente, por lo que el margen de mejora absoluta es estrecho. A pesar de ello, cualquier pequeña ganancia en el ROC-AUC es valiosa para la precisión en un escenario de priorización de llamadas a gran escala.
3. Estabilidad y Prevención de Sobreajuste (Overfitting): Gracias a la sintonización mediante RandomizedSearchCV, parámetros como la regularización L2 (`l2_leaf_reg`), la tasa de aprendizaje (`learning_rate`) y la profundidad del árbol (`depth`) se han optimizado en conjunto bajo validación cruzada. Esto nos da la garantía de que el modelo no solo es ligeramente mejor, sino significativamente más robusto y confiable para generalizar con datos nuevos en el entorno operativo final.

El proceso de optimización confirmó que la configuración base de CatBoost ya ofrecía el mejor equilibrio entre precisión y recall: la búsqueda de hiperparámetros mejoró marginalmente el ROC-AUC ($0,9672 \rightarrow 0,9684$) pero redujo el F1-score ($0,6549 \rightarrow 0,6385$), métrica priorizada en este trabajo por el desbalance de clases. En consecuencia, y de forma coherente con el criterio de selección aplicado, se conserva la configuración base de CatBoost para la generación de las predicciones del esquema de priorización. La experimentación con hiperparámetros queda como evidencia del proceso iterativo y como punto de partida para futuras iteraciones en la fase de implementación.

La sección “13. Importancia de variables”, es crucial para comprender qué características del conjunto de datos son las más influyentes en las predicciones de

nuestro modelo. Esto no solo ayuda a interpretar el modelo, sino también a obtener insights de negocio valiosos.

La sección “14. Scoring híbrido negocio + ML”, toma las predicciones del modelo “CatBoost” y las traduce en decisiones operativas concretas, qué hacer y cuándo, basándose en la lógica de negocio predefinida, creando así una estrategia de contactabilidad optimizada y lista para la implementación.

Esta sección es el puente entre el modelo predictivo de Machine Learning y la estrategia operativa real. Toma las probabilidades de contactabilidad efectiva predichas por el modelo CatBoost Optimizado, el mejor seleccionado, para cada cliente y las combina con reglas de negocio predefinidas. El objetivo es asignar a cada cliente una ACCION_SUGERIDA (llamada, WhatsApp o SMS) y un HORARIO_SUGERIDO para maximizar la eficiencia y efectividad del contacto, considerando las capacidades operativas.

El proceso se realiza en los siguientes pasos:

Cálculo de Probabilidades: Se utiliza el modelo CatBoost Optimizado para predecir la probabilidad de contacto (probabilidad_contacto) para todos los clientes.

Ordenamiento: Los clientes se ordenan de mayor a menor según su probabilidad de contacto.

Aplicación de Reglas de Negocio: Se aplican reglas jerárquicas para asignar la acción y el horario:

- **Prioridad 1 (Top 1.500):** Los 1.500 clientes con mayor probabilidad de contacto se destinan a llamada telefónica en la franja de 4:00 p.m. - 5:00 p.m. Las llamadas se generan automáticamente a través del sistema de campañas de la empresa, distribuyendo los clientes en franjas horarias según la segmentación propuesta por el modelo; cuando la llamada es contestada, es atendida por un asesor humano.

- Prioridad 2 (Siguiendo 10.500): Los siguientes 10.500 clientes se asignan a llamadas distribuidas en franjas horarias desde 8:00 a.m. hasta 4:00 p.m. El volumen total de 12.000 llamadas (1.500 + 10.500) responde a la capacidad operativa calculada por los líderes de la organización: número de asesores disponibles multiplicado por el promedio de gestiones por hora, descontando capacidad reducida en la franja de almuerzo (12:00 - 2:00 p.m.) y tiempos de descanso. Se considera además la rotación propia del canal: no todos los clientes contestan en el primer intento, por lo que los asesores alternan entre números no contestados y atendidos durante la jornada.
- Prioridad 3 (Siguiendo 10.000): Los próximos 10.000 clientes se asignan a contacto por WhatsApp durante el horario de 8:00 a.m. - 5:00 p.m. El canal opera de forma híbrida: inicia con un chatbot automatizado y, si el cliente avanza en la conversación o solicita atención personalizada, se transfiere a un asesor humano. Dado que solo una fracción de los contactos deriva en atención de asesor, el canal puede absorber un mayor volumen que el de llamadas. En caso de sobrecupo, los clientes entran en lista de espera para ser gestionados en la siguiente franja disponible.
- Prioridad 4 (Resto): Los clientes restantes se les asigna un mensaje de SMS para el horario de 8:00 a.m. - 5:00 p.m.

La importancia de esta sección radica en transformar una predicción numérica, la probabilidad, en una estrategia operativa tangible y accionable. Este "scoring híbrido" permite:

Optimización de recursos: Direccionar los canales más costosos y de mayor impacto como las llamadas, a los clientes con mayor probabilidad de éxito.

Eficiencia operativa: Distribuir la carga de trabajo del asesor comercial de manera inteligente a lo largo del día y entre los canales, evitando la saturación.

Maximización de la contactabilidad: Asegurar que todos los clientes reciban una acción de contacto adecuada a su perfil y a las capacidades operativas, incrementando las oportunidades de contacto efectivo.

Integración ML-Negocio: Unir el conocimiento del negocio (reglas de capacidad, horarios, preferencias de canal) con el poder predictivo del ML para tomar decisiones más inteligentes.

Resultados detallados

El principal producto es un DataFrame (df_final) que contiene, para cada cliente, la probabilidad de contacto predicha por el modelo, y las nuevas columnas ACCION_SUGERIDA y HORARIO_SUGERIDO. Las visualizaciones generadas muestran claramente la distribución de estas acciones:

Distribución de Acciones Sugeridas:

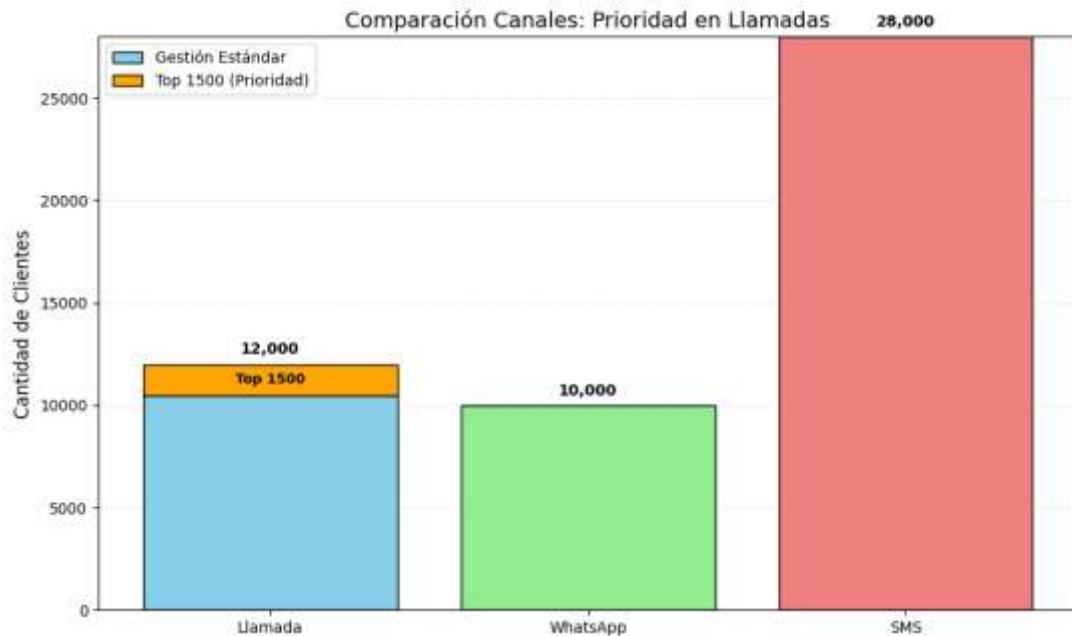
Llamadas: Un total de 12,000 clientes (1500 de alta prioridad y 10500 de prioridad media) son asignados a llamadas.

WhatsApp: 10,000 clientes son asignados a mensaje por WhatsApp.

SMS: Los clientes restantes son asignados a mensaje por SMS.

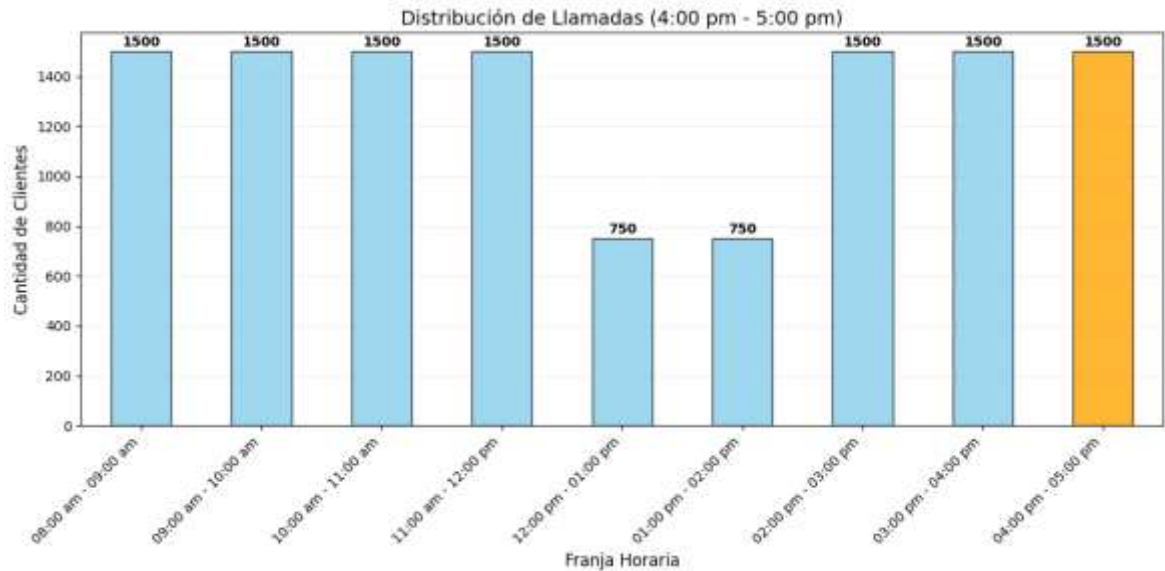
Figura 13

Distribución de Acciones Sugeridas



La asignación de los 1.500 clientes de mayor prioridad a la franja de 4:00 p.m. – 5:00 p.m. no implica relegar su gestión al final de la jornada, sino reservarles, precisamente, la franja con mayor probabilidad de contacto efectivo identificada en el análisis de los registros históricos. Dicho de otro modo, esta franja constituye el momento más favorable del día para lograr comunicación con el deudor; por ello, destinarla a los clientes con mayor score predictivo maximiza la probabilidad conjunta de contacto, al combinar al cliente con mayor propensión a responder con el horario de mayor tasa de respuesta histórica. Esta lógica es coherente con los enfoques de cobranza inteligente que conciben la decisión de cuándo contactar como una variable de optimización, y no como una acción uniforme a lo largo del día (Przybyłek et al., 2025).

Figura 14
Distribución Horaria de Llamadas



La decisión obedece, además, a una restricción de capacidad operativa: dado que el número de asesores no permite concentrar todas las llamadas en esa franja, se reserva de manera exclusiva para el segmento de mayor valor predictivo, mientras que los 10.500 clientes de la siguiente prioridad se distribuyen en bloques horarios desde las 8:00 a.m. hasta las 4:00 p.m. El tope de 1.500 clientes corresponde a la capacidad real de los equipos de asesores, definida por los líderes de la organización como parámetro de referencia, y puede ajustarse según la disponibilidad de cada campaña.

Las 10.500 llamadas restantes se distribuyen equitativamente (o según capacidades predefinidas) en bloques de 1500 (o 750 para el medio día) clientes por hora desde las 8:00 a.m. hasta las 4:00 p.m.

En resumen, la sección 14 produce un plan de acción detallado para el equipo de contactabilidad, priorizando a los clientes de mayor valor predictivo con los recursos operativos más adecuados.

Conclusiones y Recomendaciones

El desarrollo del presente proyecto permitió diseñar y validar, sobre datos históricos de Promociones y Cobranzas Beta, que variables operativas como la acción realizada y el canal de recuperación empleado son los predictores más relevantes de la contactabilidad efectiva (importancias de 21.83% y 11.18% respectivamente según CatBoost). Este resultado, junto con la confirmación de un desbalance significativo en el contacto exitoso que apenas alcanza un 5.87%, reforzó la necesidad de una propuesta de diseño técnica y estratégica basada en los datos disponibles. Es importante precisar que los hallazgos se sustentan en la validación interna del modelo sobre datos históricos con partición train/test; su desempeño en un entorno operativo real requeriría validación posterior en producción, lo cual queda fuera del alcance del presente trabajo.

Para dar respuesta a esta problemática, se realizó una evaluación sistemática de seis modelos de Machine Learning, utilizando métricas robustas para datos desbalanceados como el ROC-AUC y el F1-score, lo que permitió determinar que el algoritmo CatBoost ofrece el rendimiento óptimo para este contexto. Debido a su capacidad para equilibrar de manera fiable la precisión y el recall, este modelo se consolidó como el núcleo del esquema de priorización híbrido.

Conviene precisar el alcance del objetivo relativo al modelo predictivo. El presente trabajo no propone un nuevo algoritmo de aprendizaje automático, sino que aplica, optimiza y valida algoritmos ya existentes y ampliamente documentados en la literatura (Regresión Logística, Random Forest, Extra Trees, XGBoost, LightGBM y CatBoost) sobre los datos históricos de Promociones y Cobranzas Beta. En este sentido, el "desarrollo del modelo" se entiende en su acepción propia de la ciencia de datos aplicada: la construcción, el ajuste de hiperparámetros y la validación de un modelo

entrenado específicamente para el problema de contactabilidad de la empresa, y no la formulación de una técnica algorítmica inédita. Por tanto, el aporte diferenciador del proyecto no radica en el algoritmo en sí, CatBoost es un método preexistente, sino en el esquema híbrido de priorización que traduce las probabilidades estimadas por el modelo en decisiones operativas concretas (canal y franja horaria sugeridos), integrando las reglas de negocio y las restricciones de la Ley 2300 de 2023. Esa combinación de un modelo predictivo aplicado con un sistema de reglas y restricciones normativas constituye la contribución específica de la propuesta para Promociones y Cobranzas Beta.

Finalmente, este sistema integra las probabilidades predictivas con reglas operativas claras, traduciendo los datos en un ranking dinámico que asigna acciones concretas, como llamadas, mensajes de WhatsApp o SMS, en franjas horarias estratégicas. De este modo, el modelo propuesto orienta la asignación de la capacidad operativa hacia los clientes con mayor probabilidad de contacto efectivo, incorporando como filtros fijos las restricciones horarias y de frecuencia exigidas por la Ley 2300 de 2023. El cumplimiento efectivo de dichas restricciones en la operación real dependerá de que los sistemas que consuman el ranking respeten estos parámetros en producción, lo cual es parte del plan de implementación propuesto.

Partiendo de la solidez técnica de estos hallazgos y de la validación del modelo como una herramienta capaz de transformar la gestión operativa, se hace necesario establecer una hoja de ruta que garantice no solo su correcta puesta en marcha, sino también su evolución y efectividad a largo plazo. Con el propósito de convertir este potencial predictivo en resultados tangibles y sostenibles para la organización, se proponen las siguientes líneas de acción estratégica:

Para asegurar la sostenibilidad y el éxito a largo plazo del modelo desarrollado, es fundamental trascender la fase de diseño técnico e iniciar una estrategia de

implementación y mejora continua. Esto comienza con el monitoreo constante de las variables operativas críticas, específicamente la "Acción" y el "Canal de Recuperación", permitiendo identificar de manera proactiva cualquier cambio en el comportamiento del deudor que pueda afectar la validez de las predicciones. Complementario a este seguimiento, se debe explorar la optimización de los umbrales de clasificación del modelo CatBoost mediante el análisis de las curvas Precision-Recall y ROC, lo que facultará a la organización para ajustar la sensibilidad del modelo priorizando precisión o recall según los objetivos de negocio de cada campaña.

Como paso lógico hacia la operatividad total, resulta pertinente ejecutar un piloto controlado en un segmento real de la cobranza, lo cual servirá para validar el desempeño de la herramienta en un entorno vivo, cuantificar el retorno de inversión y medir indicadores clave de efectividad (como la Tasa de Contacto Efectivo) antes de un despliegue masivo. Finalmente, el éxito de este ecosistema dependerá de un proceso de refinamiento iterativo que nutra el esquema de priorización con la retroalimentación de los equipos de asesores de cobranzas, asegurando que el ajuste de franjas horarias, canales y capacidades operativas mantenga siempre el estricto cumplimiento de la Ley 2300 de 2023.

Cronograma de Implementación

Dado que el modelo predictivo constituye una propuesta inicial entregada en el marco de una asesoría, el valor de esta propuesta para Promociones y Cobranzas Beta se concreta cuando el modelo sea llevado a producción y sometido a un ciclo de refinamiento basado en indicadores de desempeño. Por ello se propone un cronograma de implementación posterior a esta asesoría, en el que se identifican las etapas, actividades y ruta crítica hacia la puesta en producción, con el fin de orientar a la empresa en la integración del modelo a sus procesos operativos.

Tabla 6

Cronograma Propuesta de implementación

Cronograma Propuesta de implementación (★ = ruta crítica)						
Etapas / Actividad	S1	S2	S3	S4	S5	S6
Etapas 1 – Preparación						
Aprobación, conformación del equipo (negocio, analítica, TI) y definición de KPIs y línea base	X					
★ Integración de datos y automatización del pipeline (conexión con el CRM y fuentes)	X	X				
Despliegue del modelo en ambiente de pruebas (staging)		X				
Etapas 2 – Piloto controlado						
★ Ejecución del piloto en un segmento real con grupo de control		X	X			
Medición de KPIs y comparación contra la línea base			X			
Etapas 3 – Ajuste y refinamiento						
Optimización de umbrales (curvas Precision-Recall y ROC) según KPIs			X	X		
Reentrenamiento con datos del piloto y ajuste de reglas operativas (frangas, canales, capacidad)				X		

Etapa / Actividad	S1	S2	S3	S4	S5	S6
Etapa 4 – Despliegue masivo y mejora continua						
★ Despliegue a producción a escala				X	X	
Monitoreo de KPIs, <i>data drift</i> y cumplimiento de la Ley 2300 de 2023					X	X
Ciclo de refinamiento iterativo y reentrenamiento periódico					X	X

La ruta crítica del cronograma de implementación está conformada por la secuencia: integración de datos y pipeline → ejecución del piloto controlado → despliegue a producción a escala. La integración de datos es el principal cuello de botella, ya que tanto el piloto como el despliegue dependen de contar con un flujo de información confiable y automatizado desde el CRM; cualquier retraso en esta actividad desplaza toda la fecha de puesta en marcha. El refinamiento del modelo se gobierna por un conjunto de KPIs definidos al inicio: la Tasa de Contacto Efectivo como indicador principal, complementada por la tasa de promesa de pago, la tasa de recaudo efectivo, el costo por contacto efectivo y el retorno de inversión, además de los indicadores técnicos del modelo (ROC-AUC, F1-score, precisión y recall) y un indicador de cumplimiento normativo (porcentaje de contactos dentro de los horarios y frecuencias permitidos por la Ley 2300 de 2023). El monitoreo continuo de estos KPIs frente a la línea base determina los ajustes de umbral y los reentrenamientos periódicos, cerrando el ciclo de mejora continua.

Referencias

Abe, M. (2010). *Customer lifetime value and dynamic treatment of customers*. Journal of Interactive Marketing, 24(2), 79–94. <https://doi.org/10.1016/j.intmar.2010.02.001>

Appel, A. P., Oliveira, V., Lima, B., Malfatti, G. L., Santana, V. F., & de Paula, R. (2019). Optimize cash collection: *Use machine learning to predicting invoice payment*. arXiv. <https://arxiv.org/abs/1912.10828>

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization*. Journal of Machine Learning Research, 13, 281–305.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>

Buttle, F. (2009). *Customer Relationship Management: Concepts and Technologies* (2.^a ed.). Butterworth-Heinemann.

Encyclopaedia Britannica. (s.f.). *Precision and recall*. Recuperado de <https://www.britannica.com/science/precision-and-recall>

Chehrizi, N., & Weber, T. A. (2015). *Dynamic valuation of delinquent credit-card accounts*. Management Science, 61(12), 3077–3096. <https://doi.org/10.1287/mnsc.2015.2203>

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Cifuentes, N., Gutiérrez, L., & Valencia, C. F. (2023). Modelo predictivo de la probabilidad de aumento de los días de mora para usuarios de tarjeta de crédito. Universidad de los Andes.

Cobranzas Beta. (2026). ¿Quiénes somos? Recuperado el 20 de enero 2026.

<https://www.cobranzasbeta.com.co/quienes-somos/>.

Cobranzas Beta. (2025). Obtenido de <https://www.cobranzasbeta.com.co/nuestras-regionales/>.

<https://www.cobranzasbeta.com.co/nuestras-regionales/>

Congreso de Colombia. (2023). Ley 2300 de 2023. <https://www.funcionpublica.gov.co/>

Financial Times. (2022). The Rise of Digital Debt Collection Strategies.

<https://www.ft.com/>

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>

Google Developers. (s.f.). *Classification: Accuracy, precision, recall*. Recuperado de <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall>

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación (6.ª ed.)*. McGraw-Hill Education.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley

Ibáñez, M., & Schneider, S. O. (2023). *Income risk, precautionary saving, and loss aversion: An empirical test*. SSRN Electronic Journal.

<https://doi.org/10.2139/ssrn.4495017>

Inferensys. (s.f.). *Precision and Recall: Definition, formulas, trade-off*. Recuperado de <https://inferensys.com>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems (NeurIPS)*.

Kovalenko, D., Afanasieva, O., Zabuta, N., Boiko, T., & Baltov, R. R. (2021). *Model of assessing the overdue debts in a commercial bank using neuro-fuzzy technologies*.

Journal of Risk and Financial Management, 14(5), Article 216.

<https://doi.org/10.3390/jrfm14050216>

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). *Quantifying the carbon emissions of machine learning*. arXiv:1910.09700. <https://arxiv.org/abs/1910.09700>

Martínez Quiroga, R. (2024). *Modelización del riesgo de crédito con técnicas de machine y deep learning* [Trabajo Fin de Máster, Universidad Carlos III de Madrid].

e-Archivo, Repositorio Institucional UC3M. <https://hdl.handle.net/10016/46862>

Mesa, S. (2022). La gestión de cobro como herramienta para el fortalecimiento del relacionamiento cliente-empresa. Universidad EAN.

Montoya Yepes, J. D. (2019). Analítica de datos aplicada a la cobranza de cartera. Universidad EAFIT.

Payne, A., & Frow, P. (2005). *A strategic framework for customer relationship management*. *Journal of Marketing*, 69(4), 167–176.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features*. *Advances in Neural Information Processing Systems (NeurIPS)*.

Przybyłek, M., Przybyłek, A., Wojciechowski, K., & Shkroba, I. (2025). *Towards a smart debt collection system: A design science research approach*. *Journal of Big Data*, 12, Article 193. <https://doi.org/10.1186/s40537-025-01252-0>

Rojas Romero, Y. V. (2023). *Optimización del procedimiento de cobro de cartera a clientes en una empresa del sector de elaboración de productos alimenticios*. Universidad Santo Tomás.

Scikit-learn Developers. (2024). *RandomizedSearchCV*. En *Scikit-learn: Machine Learning in Python*. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

Strubell, E., Ganesh, A., & McCallum, A. (2019). *Energy and policy considerations for deep learning in NLP*. En Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 3645–3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>

Superintendencia Financiera de Colombia. (2023). *Incertidumbre en Colombia: impactos macroeconómicos y en el mercado de crédito* (Documento de análisis del sistema financiero). <https://www.superfinanciera.gov.co>.

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., & Nguyen, H. D. (2025). *Multi-agent collaboration mechanisms: A survey of LLMs*. arXiv. <https://doi.org/10.48550/arXiv.2501.06322>

Zeng, S., Melville, P., Lang, C. A., Boier Martin, I., & Murphy, C. (2008). *Using predictive analysis to improve invoice to cash collection*. En *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 1043–1050. ACM. DOI: <https://doi.org/10.1145/1401890.1402014>