

Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

**Trabajo previo a la obtención de título de
Magister en Inteligencia Artificial Aplicada**

AUTORES:

Andrade Enclada José Oswaldo
Jiménez Benavides Darwin Wilson
Morales Malo Jairo Alexander

TUTORES:

Karla Estefanía Mora Cajas
Andrés Roberto Navas Castellanos

**OPTIMIZACIÓN DE ESTRATEGIAS DE RECUPERACIÓN DE
CARTERA BASADA EN MODELOS DE INTELIGENCIA
ARTIFICIAL Y COMPORTAMIENTO HISTÓRICO**

Certificación de autoría

Nosotros, **Andrade Encalada José Oswaldo, Jiménez Benavides Darwin Wilson y Morales Malo Jairo Alexander**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



**Jose Oswaldo
Andrade Encalada**



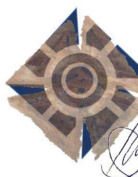
Andrade Encalada José Oswaldo



**Darwin Wilson
Jimenez Benavides**



Jiménez Benavides Darwin Wilson



Firmado digitalmente
por JAIRO ALEXANDER
MORALES MALO
Fecha: 2026.08.24
21:54:17 -05'00'

Morales Malo Jairo Alexander

Autorización de Derechos de Propiedad Intelectual

Nosotros, Andrade Encalada José Oswaldo, Jiménez Benavides Darwin Wilson y Morales Malo Jairo Alexander, en calidad de autores del trabajo de investigación titulado *Optimización de estrategias de recuperación de cartera basada en modelos de Inteligencia Artificial y comportamiento histórico*, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, (Julio 2026)



Jose Oswaldo
Andrade Encalada



José Oswaldo Andrade Encalada



Darwin Wilson
Jimenez Benavides



Darwin Wilson Jiménez Benavides

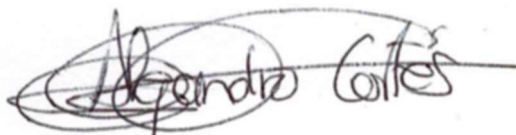


Firmado digitalmente
por JAIRO ALEXANDER
MORALES MALO
Fecha: 2026.08.24
21:54:36 -05'00'

Jairo Alexander Morales Malo

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Andrade Encalada José Oswaldo, Jiménez Benavides Darwin Wilson y Morales Malo Jairo Alexander** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

Dedicatoria

A mis hijas, Nayeli y Karen, quienes son el motor de mi vida y la mayor inspiración para alcanzar cada una de mis metas. Su amor, alegría y apoyo incondicional me han brindado la fortaleza necesaria para culminar este importante logro académico.

Asimismo, dedico este trabajo a todas las personas que me acompañaron y alentaron durante este proceso, contribuyendo de una u otra manera a que este objetivo se hiciera realidad.

— Oswaldo Andrade

Dedico este trabajo a mi familia, por ser mi apoyo incondicional, mi fuerza en los momentos difíciles y mi mayor motivación para seguir adelante. Su amor, paciencia y comprensión han sido fundamentales durante este proceso, y me han permitido llegar hasta esta meta con esperanza y determinación.

A mis padres, con especial cariño y gratitud, por enseñarme desde siempre el valor de la responsabilidad, la disciplina y el esfuerzo constante. Gracias por ser ejemplo de vida, por impulsarme a superarme cada día y por recordarme que siempre hay algo nuevo por aprender. Este logro también les pertenece a ustedes.

— Darwin Jiménez

Dedico este trabajo a Dios, mi fortaleza y mi roca. A mi amada esposa, Paola Alexandra, por su amor incondicional y su complicidad en los momentos más retadores de nuestras vidas, por su confianza e inspiración permanente. A nuestras hijas, Dana Romina y Ema Sofia, quienes impulsan cada uno de nuestros sueños. Que este logro, también suyo, sea inspiración de constancia y fe.

A mis padres, que plantaron en tierra fértil y son también fuente de inspiración; a mis hermanos, y a toda mi familia, que con su cariño y sus palabras de aliento me han cobijado a lo largo del camino.

— Jairo Morales Malo

Agradecimientos

Agradezco profundamente a Dios por acompañarme a lo largo de este camino académico y permitirme alcanzar esta meta.

Mi gratitud más especial es para mis hijas, Nayeli y Karen, quienes con su amor, alegría y confianza se convirtieron en la principal razón para perseverar y superarme cada día.

A mi familia, a mis profesores y a todas las personas que me brindaron su apoyo y orientación durante este proceso, les expreso mi más sincero agradecimiento por haber sido parte de este importante logro.

— Oswaldo Andrade

Agradezco profundamente a mi familia por su amor, paciencia y comprensión durante todo este tiempo de estudio. Su apoyo incondicional fue fundamental para mantenerme firme en los momentos de cansancio, duda y esfuerzo. Gracias por acompañarme en este proceso y por ser una fuente constante de ánimo y fortaleza.

A mis padres, les expreso mi más sincero agradecimiento por haberme formado con valores como la responsabilidad, la disciplina y la perseverancia. Su ejemplo ha sido una guía permanente en mi vida y una inspiración para seguir creciendo personal y profesionalmente. Gracias por enseñarme que cada meta se alcanza con esfuerzo y que siempre existe algo nuevo por aprender.

— Darwin Jiménez

Agradezco a Dios, por ser aliento en cada nuevo amanecer, por habitar tanto en el ruido como en el silencio; por la luz. Por quienes caminaron a mi lado durante esta travesía, acompañando los días de entusiasmo y también las jornadas de duda. Por mi esposa y mis amadas hijas, cómplices y testigos de este recorrido.

Por mis padres, quienes me dieron lecciones de entrega, perseverancia y amor; por mis hermanos, por mis maestros y amigos. Por cada una de las personas que de alguna u otra forma siempre han estado ahí y a aquellas que se fueron. Y también al cardo que nació entre las espinas.

— Jairo Morales Malo

Resumen

El presente proyecto propone el desarrollo de un sistema inteligente de apoyo a la toma de decisiones para la optimización de la gestión de cobranza en una empresa del sector retail con crédito directo en Ecuador. La investigación surge de la necesidad de incrementar la efectividad de recuperación de cartera mediante decisiones fundamentadas en datos, superando los enfoques tradicionales que aplican estrategias homogéneas sin considerar las características individuales y el comportamiento histórico de cada cliente.

Para ello, se utilizan datos históricos relacionados con comportamiento de pago, características crediticias, gestión de cobranza y riesgo de morosidad, con el objetivo de construir modelos predictivos capaces de estimar la probabilidad de mora, la probabilidad de pago, la probabilidad de regularización y la contactabilidad de los clientes. Los resultados de estos modelos son integrados en un esquema de priorización que permite identificar las cuentas con mayor potencial de recuperación y determinar el nivel de esfuerzo operativo más adecuado para cada gestión.

Sobre esta base, se desarrolla un modelo de integración para la definición de estrategias de cobranza, el cual utiliza como variables de decisión el Decil de Gestión, el Tipo de Gestión y el Nivel de Contactabilidad. Estas variables sintetizan los resultados obtenidos por los modelos predictivos y permiten asignar estrategias diferenciadas de acuerdo con la prioridad de gestión, la naturaleza preventiva o recuperativa de la intervención y la probabilidad estimada de contacto con el cliente. La integración de estos criterios transforma los resultados predictivos en recomendaciones operativas orientadas a optimizar la asignación de recursos y apoyar la toma de decisiones dentro del proceso de cobranza.

La propuesta demuestra cómo las técnicas de inteligencia artificial y aprendizaje automático pueden integrarse dentro de un esquema de priorización y definición de estrategias para apoyar la toma de decisiones en un entorno empresarial real. Como resultado, se obtiene

una herramienta capaz de transformar información histórica en recomendaciones operativas accionables, contribuyendo a mejorar la gestión de cobranza, optimizar la asignación de recursos y fortalecer la transformación analítica de los procesos de gestión crediticia.

Palabras Claves:

Inteligencia artificial; aprendizaje automático; modelos predictivos; análisis predictivo; gestión de cobranza; priorización de cartera; segmentación de clientes; recuperación de cartera; riesgo crediticio; comportamiento de pago; comportamiento histórico; contactabilidad; toma de decisiones; eficiencia operativa; cartera vencida; morosidad; sector retail; uso del crédito; exposición al riesgo; regularización de cartera.

Abstract

This research proposes the development of an intelligent decision-support system to optimize debt collection management in a retail company that provides direct consumer credit in Ecuador. The study arises from the need to improve debt recovery effectiveness through data-driven decision-making, overcoming traditional approaches that apply uniform collection strategies without considering the individual characteristics and historical behavior of each customer.

To achieve this objective, historical data related to payment behavior, credit characteristics, collection management, and delinquency risk are used to develop predictive models capable of estimating the probability of default, the probability of payment, the probability of regularization, and customer contactability. The outputs of these models are integrated into a prioritization framework that identifies the accounts with the greatest recovery potential and determines the most appropriate operational effort for each collection process.

Based on these results, an integration model is developed for the definition of collection strategies using the Management Decile, Management Type, and Contactability Level as decision variables. These variables summarize the outputs of the predictive models and enable the assignment of differentiated collection strategies according to management priority, the preventive or recovery-oriented nature of the intervention, and the estimated probability of successfully contacting the customer. The integration of these criteria transforms predictive results into operational recommendations aimed at optimizing resource allocation and supporting decision-making throughout the collection process.

The proposed approach demonstrates how artificial intelligence and machine learning techniques can be integrated into a prioritization and strategy-definition framework to support decision-making in a real business environment. As a result, the research provides a tool capable of transforming historical information into actionable operational

recommendations, contributing to improved collection management, optimized resource allocation, and enhanced analytical transformation of credit management processes.

Keywords:

Artificial intelligence; machine learning; predictive models; predictive analytics; debt collection management; portfolio prioritization; customer segmentation; debt recovery; credit risk; payment behavior; historical behavior; contactability; decision-making; operational efficiency; overdue portfolio; delinquency; retail sector; credit utilization; risk exposure; portfolio regularization.

Tabla de contenidos

Contenido	
Capítulo 1	1
Definición del proyecto	1
Justificación e importancia del trabajo de investigación	2
Alcance	3
Objetivos.....	6
Objetivo general	6
Hipótesis.....	6
Objetivo específico	6
Capítulo 2.....	8
Estado del Arte.....	8
Modelos de Predicción y Métricas de Evaluación	9
Marco Teórico	13
Fundamentos de credit scoring y comportamiento del cliente	13
Ventanas de tiempo en el análisis del comportamiento crediticio	14
Modelos de predicción en credit scoring.....	16
Métricas de clasificación y medición utilizadas en modelos predictivos.	19
Segmentación de clientes.....	22
Formulación de la cobranza como problema multiobjetivo.	22
Dominancia de Pareto y soluciones no dominadas.	23
Algoritmos de optimización multiobjetivo.	24
Generación de políticas de cobranza basadas en Pareto.	25
Capítulo 3.....	27
Desarrollo del Trabajo.....	27
Descripción Datasets.....	29
Dataset DSScoreCobranza.....	29
Dataset DSScoreContacto.....	35
Preparación y procesamiento de datos	38
Desarrollo del Modelo Predictivo de Mora	40
Objetivo del Modelo Predictivo de Mora.....	41
Construcción y validación del TargetMora	42
Preparación del conjunto de datos	45
Selección de variables.....	46
Metodología de entrenamiento	48
Implementación del modelo.....	50
Conclusión del desarrollo	51
Desarrollo del Modelo Predictivo de Pago.....	52
Objetivo del Modelo Predictivo de Pago	53
Construcción y validación de TargetPago	53
Preparación del conjunto de datos	55

Selección y depuración de variables	57
Metodología de entrenamiento	59
Implementación del modelo.....	61
Conclusión del desarrollo	63
Desarrollo del Modelo Predictivo de Regularización	64
Objetivo del Modelo Predictivo de Regularización	65
Construcción y validación del TargetRegularizacion	65
Preparación del conjunto de datos	67
Selección y depuración de variables	69
Metodología de entrenamiento	70
Implementación del modelo.....	72
Conclusión del desarrollo	74
Modelo de Priorización de Gestión de Cobranza	75
Objetivo del Modelo de Priorización de Gestión de Cartera	76
Integración de los modelos predictivos.....	77
Construcción del Score de Priorización de Gestión	78
Segmentación de la cartera mediante deciles de gestión	80
Implementación del Modelo de Priorización de Gestión de Cartera	81
Conclusión del desarrollo	81
Desarrollo del Modelo Predictivo de Contactabilidad.....	82
Objetivo del Modelo Predictivo de Contactabilidad	83
Construcción y validación del TargetContactabilidad.....	84
Análisis exploratorio por muestra temporal	85
Preparación del conjunto de datos	87
Selección y depuración de variables	88
Metodología de entrenamiento	91
Evaluación de modelos	91
Generación de probabilidad, score y segmentos de contactabilidad.....	92
Implementación del Modelo	94
Conclusión del desarrollo	95
Implementación del Modelo de Priorización de Gestión de Cobranza	96
Propósito de la implementación.....	97
Variables utilizadas en el modelo.....	98
Transformación operativa de DecilGestion.....	99
Formulación del esquema de decisión	99
Alternativas de estrategia.....	100
Motor de reglas de asignación	101
Integración en la tabla operativa	101
Trazabilidad, parametrización y control.....	102
Capítulo 4.....	103
Resultados de los modelos predictivos.....	103
Resultados del Modelo Predictivo de Mora	104
Resultados del Modelo Predictivo de Pago.....	109
Resultados del Modelo Predictivo de Regularización.....	114

Resultados del Modelo Predictivo de Contactabilidad	119
Resultados del Modelo de Priorización de Gestión de Cobranza	128
Resultados del Score de Priorización de Gestión	128
Segmentación de la cartera mediante deciles de gestión	130
Análisis de la exposición y priorización	131
Integración de Contactabilidad y Priorización de Gestión	136
Análisis de segmentos de gestión.....	138
Resultados de la generación de estrategias de gestión	140
Matriz de decisión estratégica.....	141
Composición por TipoGestion dentro de cada estrategia	143
Análisis económico de estrategias de cobranza.....	146
Supuestos de costo y construcción del escenario económico	147
Composición del gasto actual por canal.....	147
Comparación económica entre escenario actual y escenario propuesto	148
Impacto económico por estrategia propuesta	149
Ahorro por combinación operativa	151
Capítulo 5.....	153
Conclusiones.....	153
Conclusión 1. Cumplimiento del objetivo general	153
Conclusión 2. Construcción del conjunto de datos y aporte metodológico	153
Conclusión 3. Desarrollo y validación de los modelos predictivos	154
Conclusión 4. Modelo de Priorización de Gestión de Cobranza	154
Conclusión 5. Aplicabilidad de la solución propuesta.....	154
Conclusión 6. Contrastación de la hipótesis de investigación	155
Conclusión 7. Aporte del Modelo de Contactabilidad.....	155
Conclusión 8. Generación de estrategias diferenciadas de cobranza	155
Conclusión 9. Integración entre analítica predictiva y operación	156
Conclusión 10. Contribución académica y metodológica	156
Recomendaciones	156
Recomendación 1. Implementación gradual de la solución en producción	156
Recomendación 2. Establecimiento de un proceso formal de monitoreo y reentrenamiento ..	157
Recomendación 3. Explotación operativa del Modelo de Contactabilidad	157
Recomendación 4. Parametrización y evolución continua de las estrategias de cobranza.....	158
Recomendación 5. Evaluación continua del impacto de negocio.....	158
Recomendación 6. Evolución hacia un ecosistema de cobranza inteligente.....	158
Recomendación 7. Evaluación futura de técnicas de optimización multiobjetivo	159
Recomendación 8. Incorporación de experimentación controlada (A/B Testing)	159
Recomendación 9. Fortalecimiento de la gobernanza y trazabilidad de modelos.....	159
Recomendación 10. Líneas futuras de investigación.....	160
Referencias.....	161
Anexo.....	165
Repositorio digital de la investigación	165
Estructura del repositorio	165

Organización de la carpeta documentación	165
Organización de la carpeta modelos	166
Organización de la carpeta resultados	166
Repositorio Digital.....	166

Lista de Tablas

Tabla 1 Resumen estadístico del Dataset utilizado para el entrenamiento del Modelo Predictivo de Mora.....	43
Tabla 2 Resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Pago	54
Tabla 3 Resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Regularización	66
Tabla 4 Resumen descriptivo del Dataset utilizado para el Modelo Predictivo de Contactabilidad ..	84
Tabla 5 Evolución de la contactabilidad y de la intensidad de gestión por muestra temporal	86
Tabla 6 Principales pares de variables altamente correlacionadas identificadas en el análisis exploratorio	89
Tabla 7 Criterios metodológicos para seleccionar el modelo final.....	92
Tabla 8 Segmentos metodológicos de contactabilidad.....	93
Tabla 9 Campos de contactabilidad integrables al proceso de gestión	94
Tabla 10 Componentes utilizados para construir el ScoreContactabilidad.....	95
Tabla 11 Propósito funcional de la implementación del modelo.....	98
Tabla 12 Variables utilizadas para la asignación de estrategias de cobranza.....	98
Tabla 13 Transformación operativa de DecilGestion en GrupoValor.....	99
Tabla 14 Formulación del esquema de decisión para la asignación de estrategias.....	100
Tabla 15 Criterios considerados para la asignación de estrategias	100
Tabla 16 Alternativas de estrategia consideradas por el motor de decisión.....	100
Tabla 17 Integración metodológica de la estrategia en la tabla operativa	102
Tabla 18 Atributos metodológicos de la implementación propuesta	102
Tabla 19 Comparación de métricas de desempeño de los modelos predictivos de Mora	104
Tabla 20 Comparación de métricas de desempeño de los modelos predictivos de Pago.....	110
Tabla 21 Comparación de métricas de desempeño de los modelos predictivos de Regularización ..	115
Tabla 22 Distribución de la variable TargetContactabilidad	120
Tabla 23 Comparación de métricas de desempeño de los modelos predictivos de Contactabilidad	121
Tabla 24 Variables con mayor importancia en el modelo XGBoost de Contactabilidad	123
Tabla 25 Segmentación operativa de la ProbabilidadContactabilidad	127
Tabla 26 Resumen estadístico de las variables utilizadas en la construcción del Score de Priorización de Gestión	130
Tabla 27 Segmentación de la cartera mediante deciles de gestión	131
Tabla 28 Composición de la cartera según tipo de gestión	132
Tabla 29 Agrupación de DecilGestion en grupos de valor.....	137
Tabla 30 Segmentos definitivos de gestión según TipoGestion, GrupoValor y NivelContactabilidad	138
Tabla 31 Distribución definitiva de estrategias recomendadas y canales asociados.....	140
Tabla 32 Interpretación operativa de las estrategias de gestión	144
Tabla 33 Supuestos de costo unitario utilizados para la evaluación económica.....	147
Tabla 34 Distribución del gasto actual estimado por canal	148
Tabla 35 Comparación económica entre escenario actual, escenario propuesto y ahorro estimado	149
Tabla 36 Impacto económico por estrategia propuesta	150
Tabla 37 Principales combinaciones operativas según contribución al ahorro estimado	152

Lista de Figuras

Figura 1 Construcción del Dataset mediante ventanas temporales móviles	31
Figura 2 Proceso de preparación y procesamiento de datos para el entrenamiento de los modelos predictivos.....	39
Figura 3 Metodología para el desarrollo de los modelos predictivos.....	41
Figura 4 Distribución del TargetMora utilizado para el entrenamiento del Modelo Predictivo de Mora.....	44
Figura 5 Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Mora	45
Figura 6 Proceso de selección y depuración de variables para el desarrollo del Modelo Predictivo de Mora.....	47
Figura 7 Metodología de entrenamiento y evaluación de los modelos de clasificación del Modelo Predictivo de Mora.....	48
Figura 8 Proceso de implementación del Modelo Predictivo de Mora	50
Figura 9 Distribución del TargetPago utilizada para el entrenamiento del Modelo Predictivo de Pago	55
Figura 10 Metodología para el desarrollo del Modelo Predictivo de Pago	57
Figura 11 Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Pago	58
Figura 12 Metodología de entrenamiento y evaluación de los modelos de regresión del Modelo Predictivo de Pago.....	60
Figura 13 Proceso de implementación del Modelo Predictivo de Pago	62
Figura 14 Distribución del TargetRegularizacion utilizado para el entrenamiento del Modelo Predictivo de Regularización.....	67
Figura 15 Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Regularización	68
Figura 16 Proceso de selección y depuración de variables para el desarrollo del Modelo Predictivo de Regularización.....	69
Figura 17 Metodología de entrenamiento y evaluación de los modelos de clasificación del Modelo Predictivo de Regularización.....	71
Figura 18 Proceso de implementación del Modelo Predictivo de Regularización	73
Figura 19 Integración de los modelos predictivos para la construcción del conjunto de datos de gestión.....	77
Figura 20 Distribución del TargetContactabilidad utilizado para el entrenamiento del modelo.....	85
Figura 21 Evolución de la tasa de contactabilidad.....	86
Figura 22 Volumen de registros por muestra	86
Figura 23 Intensidad promedio de canales de contacto por muestra temporal	87
Figura 24 Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Contactabilidad	87
Figura 25 Mapa de correlación de variables numéricas principales.....	89
Figura 26 Proceso de selección y depuración de variables para el Modelo Predictivo de Contactabilidad	90
Figura 27 Importancia de variables del modelo XGBoost para contactabilidad	90
Figura 28 Salidas operativas generadas por el Modelo Predictivo de Contactabilidad	93
Figura 29 Proceso de implementación del Modelo Predictivo de Contactabilidad	95
Figura 30 Flujo metodológico de implementación del Modelo de Priorización de Gestión de Cobranza	97
Figura 31 Espacio de decisión definido por TipoGestion, DecilGestion y NivelContactabilidad	98
Figura 32 Jerarquía metodológica de reglas de asignación	101
Figura 33 Comparación de métricas de desempeño de los modelos predictivos de Mora	105
Figura 34 Variables con mayor importancia en el Modelo Predictivo de Mora mediante Random Forest.	107
Figura 35 Matriz de confusión del Modelo Predictivo de Mora mediante Random Forest.....	108
Figura 36 Comparación de métricas de desempeño de los modelos predictivos de Pago	111
Figura 37 Variables con mayor importancia en el Modelo Predictivo de Pago mediante Random Forest Regressor.	112

Figura 38 Comparación entre valores reales y valores predichos del Modelo Predictivo de Pago mediante Random Forest Regressor	113
Figura 39 Comparación de métricas de desempeño de los modelos predictivos de Regularización	116
Figura 40 Variables con mayor importancia en el Modelo Predictivo de Regularización mediante Random Forest	117
Figura 41 Matriz de confusión del Modelo Predictivo de Regularización mediante Random Forest	118
Figura 42 Distribución del TargetContactabilidad utilizado en la evaluación del modelo	120
Figura 43 Comparación de métricas de desempeño de los modelos predictivos de Contactabilidad	121
Figura 44 Balance de clasificación del modelo seleccionado XGBoost	122
Figura 45 Variables con mayor importancia en el Modelo Predictivo de Contactabilidad mediante XGBoost	123
Figura 46 Curvas ROC de los modelos evaluados.....	124
Figura 47 Curvas Precision-Recall de los modelos evaluados.....	124
Figura 48 Matriz de confusión del Modelo Predictivo de Contactabilidad mediante XGBoost.....	125
Figura 49 Lift por decil del modelo XGBoost.....	126
Figura 50 Distribución del score de contactabilidad por clase real.....	126
Figura 51 Segmentación operativa del ScoreContactabilidad.....	127
Figura 52 Exposición impaga total por decil de gestión	133
Figura 53 Comportamiento promedio de las variables predictivas por decil de gestión	134
Figura 54 Concentración acumulada de la exposición económica por decil de gestión (Análisis de Pareto).....	135
Figura 55 Distribución de clientes por decil según el tipo de gestión propuesto	136
Figura 56 Segmentos de gestión con mayor volumen de clientes.....	139
Figura 57 Distribución definitiva de estrategias recomendadas.....	141
Figura 58 Matriz de decisión estratégica con segmentos definitivos para tipo de gestión Preventiva	142
Figura 59 Matriz de decisión estratégica con segmentos definitivos para tipo de gestión Recuperativa.....	143
Figura 60 Composición por TipoGestion dentro de cada estrategia.....	144
Figura 61 Mapas de Decisión de Estrategias Preventivas	145
Figura 62 Mapas de Decisión de Estrategias Recuperativas	146
Figura 63 Participación del gasto actual por canal	148
Figura 64 Comparación del costo total actual y propuesto	149
Figura 65 Ahorro estimado por estrategia propuesta	151
Figura 66 Pareto de ahorro por combinación operativa	152

Capítulo 1

Definición del proyecto

Las empresas del sector retail que operan bajo esquemas de crédito directo enfrentan desafíos significativos en la gestión de cobranza, particularmente debido al incremento en el volumen de clientes, la diversidad de perfiles de comportamiento y la complejidad del riesgo crediticio. En la mayoría de los casos, los procesos de cobranza se apoyan en enfoques tradicionales caracterizados por estrategias homogéneas, reglas de negocio estáticas y segmentaciones predefinidas que no consideran de manera integral el comportamiento histórico ni la probabilidad de recuperación de los clientes (Khandani, Kim, & Lo, 2010; Lessmann, Baesens, Seow, & Thomas, 2015).

Si bien los sistemas ERP y plataformas de cobranza actuales permiten gestionar información de clientes, registrar historiales de pago y aplicar segmentaciones básicas, estos sistemas suelen operar bajo modelos de decisión unidimensionales, en los que el análisis se limita a indicadores individuales o reglas heurísticas. Este tipo de enfoque dificulta la toma de decisiones en contextos donde coexisten múltiples objetivos potencialmente conflictivos, tales como maximizar la recuperación, minimizar el riesgo de mora y optimizar el uso de los recursos operativos (Bellotti & Crook, 2013; Simumba, Okami, Kodaka, & Kohtake, 2021).

En este contexto, el presente proyecto propone un enfoque metodológico estructurado en dos niveles complementarios. En el primero, se desarrollan modelos predictivos basados en técnicas de Inteligencia Artificial para estimar la probabilidad de mora, la probabilidad de pago, la probabilidad de regularización y la probabilidad de contactabilidad de los clientes, utilizando información relacionada con su comportamiento histórico, uso del crédito y evolución del riesgo. Estos modelos permiten generar indicadores objetivos para

caracterizar el comportamiento esperado de cada cuenta. (Gicić & Đonko, 2025; Hayashi, 2022).

En el segundo nivel, los resultados de los modelos predictivos son integrados en un Modelo de Priorización de Gestión de Cobranza, complementado con el Modelo de Contactabilidad, para segmentar las cuentas y definir estrategias diferenciadas de gestión. Esta integración considera simultáneamente el nivel de prioridad, el tipo de gestión requerido y la probabilidad estimada de contacto, proporcionando un esquema de apoyo a la toma de decisiones orientado a optimizar la asignación de recursos y fortalecer la gestión operativa de cobranza (Sánchez, Maldonado, & Vairetti, 2022).

Justificación e importancia del trabajo de investigación

La gestión de cobranza constituye un proceso estratégico en organizaciones que administran cartera crediticia, especialmente en el sector retail, donde el volumen de clientes crece constantemente y los perfiles de riesgo son altamente heterogéneos. Servimax Servicios Globales S.A., empresa encargada de la gestión de crédito y cobranza de la cartera de Corporación Favorita, una de las principales compañías retail del país, aplica un enfoque tradicional basado en reglas de segmentación estática, lo cual dificulta la priorización de gestiones, incrementa los costos operativos y reducen la efectividad de las acciones de cobranza, evidenciando la relevancia de abordar este problema mediante enfoques analíticos más sofisticados.

Diversos estudios han demostrado que los modelos tradicionales de scoring presentan restricciones para capturar patrones complejos y relaciones no lineales en datos crediticios, lo que puede afectar la precisión de las estimaciones de riesgo (Lessmann, Baesens, Seow, & Thomas, 2015; Khandani, Kim, & Lo, 2010). Frente a ello, los modelos de Inteligencia Artificial han mostrado mejoras significativas en la capacidad predictiva, especialmente

cuando se dispone de información conductual rica y heterogénea (Hayashi, 2022; Mestiri, 2024).

Sin embargo, la mejora en el desempeño predictivo no resuelve por sí sola el problema de la toma de decisiones en cobranza. La definición de estrategias requiere integrar múltiples criterios relacionados con el riesgo, la probabilidad de pago, la exposición económica y la capacidad de contacto con los clientes (Simumba, Okami, Kodaka, & Kohtake, 2021). En este sentido, la importancia del proyecto radica en integrar modelos predictivos de Inteligencia Artificial dentro de un esquema de priorización y definición de estrategias de cobranza que apoye una asignación más eficiente de los recursos y una toma de decisiones basada en datos.

La resolución de este problema genera un impacto significativo a nivel empresarial y social, pues permite mejorar la eficiencia operativa y la focalización de estrategias de cobranza, además de mejorar la interacción con el cliente, a través de campañas más adecuadas y menos intrusivas, promoviendo prácticas responsables de endeudamiento. Esto impulsa la cultura financiera, el fortalecimiento de la capacidad de recuperación de cartera, y se traduce en la sostenibilidad del negocio.

A nivel académico, aporta evidencia aplicada sobre la integración de modelos de inteligencia artificial, priorización de cartera y criterios de contactabilidad para apoyar la definición de estrategias de cobranza en contextos reales de crédito retail, un campo aún poco desarrollado en la literatura. (Noriega, Rivera, & Herrera, 2023; Gicić & Đonko, 2025).

Alcance

El presente proyecto se desarrolla a partir del uso de datos históricos reales provenientes del Core operativo de Servimax Servicios Globales S.A., correspondientes al último año completo de operación. Este conjunto de datos integra información estructurada a nivel

cliente–periodo, incluyendo variables relacionadas con el comportamiento de pago, el uso y exposición al crédito, el riesgo de mora y las características operativas de la gestión de cobranza. La utilización de una ventana temporal reciente y homogénea permite capturar patrones actuales de comportamiento, asegurando la pertinencia de los resultados en el contexto del negocio.

Sobre esta base de datos, el proyecto contempla el desarrollo de modelos predictivos basados en técnicas de Inteligencia Artificial que permitan comprender y anticipar el comportamiento de los clientes dentro del proceso de cobranza. Para ello, se aprovecha la información histórica disponible sobre mora, pago esperado, regularización y características de riesgo, con el objetivo de transformar los datos en herramientas analíticas que faciliten una toma de decisiones más informada y eficiente.

El análisis no se limita a una única técnica de modelado, sino que contempla la comparación de distintos algoritmos de aprendizaje supervisado, entre ellos Regresión Logística, Random Forest y XGBoost, con el fin de identificar aquellos que ofrecen el mejor desempeño predictivo. A partir de esta evaluación, se seleccionan los modelos con mayor capacidad para estimar probabilidades de pago y generar scores que representen el comportamiento esperado de los clientes, convirtiéndose en insumos clave para la gestión de cobranza.

En un segundo nivel, el proyecto integra los resultados de los modelos predictivos en un esquema de priorización orientado a la definición de estrategias de cobranza. Sobre la base de las probabilidades y scores generados, se consideran criterios relevantes para la toma de decisiones, tales como la probabilidad de pago, el riesgo de mora, la exposición crediticia, el potencial de recuperación y el nivel de contactabilidad. La integración de estos criterios permite clasificar las cuentas según su prioridad de gestión y establecer estrategias diferenciadas que favorecen una asignación más eficiente de los recursos operativos.

Asimismo, el alcance comprende el análisis e interpretación de la segmentación obtenida mediante el Modelo de Priorización de Gestión de Cobranza, con el propósito de identificar perfiles diferenciados de riesgo, recuperación y contactabilidad. Esta segmentación permite definir estrategias preventivas y recuperativas acordes con las características de cada grupo de cuentas, proporcionando un mecanismo objetivo para apoyar la toma de decisiones operativas.

De esta manera, la propuesta trasciende el enfoque convencional de predicción aislada mediante la integración de modelos de Inteligencia Artificial, un modelo de priorización y un modelo de contactabilidad, conformando un esquema de apoyo a la toma de decisiones para la gestión de cobranza. Los resultados obtenidos aportan valor tanto desde la perspectiva analítica como operativa, al facilitar la priorización de cuentas, la asignación de estrategias diferenciadas y la utilización más eficiente de los recursos disponibles.

Finalmente, la implementación del modelo y la integración de los resultados de los modelos predictivos, del Modelo de Priorización y del Modelo de Contactabilidad se plantea como una prueba de concepto mediante la ejecución de procesos Batch, orientados a apoyar la toma de decisiones sin intervenir directamente en los sistemas operativos en tiempo real. Quedan explícitamente fuera del alcance del proyecto las implementaciones en tiempo real o basadas en servicios API, la automatización completa en entornos productivos, el desarrollo de nuevos algoritmos teóricos de Inteligencia Artificial, así como la incorporación de datos externos o de terceros. El score y los indicadores derivados serán incorporados a futuro en las bases de datos de cobranza para su análisis y utilización posterior. Esta integración está orientada a no modificar la arquitectura interna del Core operativo, garantizando estabilidad y trazabilidad en el manejo de la información.

Estas limitaciones permiten mantener un enfoque metodológico sólido, viable en el tiempo de ejecución establecido y alineado con los objetivos académicos del trabajo de

investigación. Además, el proyecto se delimita al contexto del sector retail con crédito directo en Ecuador, lo que asegura coherencia entre los datos utilizados, el modelo desarrollado y las condiciones operativas reales en las que se plantea su aplicación.

Objetivos

Objetivo general

Desarrollar un sistema de apoyo a la gestión de cobranza basado en la integración de modelos predictivos de comportamiento crediticio y contactabilidad, que permita segmentar las cuentas y definir estrategias diferenciadas de gestión, con el fin de fortalecer la toma de decisiones basada en datos.

Hipótesis

La integración de modelos predictivos de comportamiento crediticio y contactabilidad permite definir estrategias diferenciadas de gestión de cobranza, mejorando la priorización de las cuentas y apoyando una asignación más eficiente de los recursos mediante un enfoque basado en datos.

Objetivo específico

1. Recolectar y organizar los datos históricos correspondientes al último año provenientes del Core operativo de Servimax Servicios Globales S.A., garantizando su disponibilidad, consistencia y trazabilidad para el entrenamiento y evaluación del modelo.
2. Preprocesar y transformar el conjunto de datos mediante técnicas de limpieza, integración, selección de variables relevantes y preparación de características, asegurando su adecuación para los modelos a usar.
3. Desarrollar modelos de Inteligencia Artificial para estimar la probabilidad de mora, la probabilidad de pago, la probabilidad de regularización y la probabilidad de

contactabilidad de los clientes, utilizando variables de comportamiento histórico, uso del crédito y riesgo de cartera.

4. Evaluar el desempeño de los modelos de Inteligencia Artificial mediante métricas especializadas de clasificación y regresión, tales como Accuracy, Precision, Recall, F1-Score, ROC-AUC, MAE, RMSE y R^2 , con el fin de seleccionar los modelos con mejor capacidad predictiva para cada problema abordado.
5. Integrar los resultados de los modelos predictivos en un Modelo de Priorización de Gestión de Cobranza que permita segmentar las cuentas según su nivel de riesgo, potencial de recuperación y nivel de contactabilidad.
6. Desarrollar un esquema de priorización que integre los resultados del Modelo de Priorización y del Modelo de Contactabilidad para asignar estrategias diferenciadas de gestión de cobranza basadas en datos.
7. Validar la viabilidad de la solución propuesta mediante su integración como prueba de concepto en el Core operativo utilizando procesos Batch.

Capítulo 2

Estado del Arte

En los últimos años, la aplicación de técnicas de inteligencia artificial ha transformado de manera significativa el análisis del riesgo crediticio y la gestión de cartera. Inicialmente, los modelos estadísticos clásicos, como la regresión logística y el análisis discriminante, constituyeron la base del credit scoring debido a su simplicidad e interpretabilidad (Khandani, Kim, & Lo, 2010; Lessmann, Baesens, Seow, & Thomas, 2015). Sin embargo, estos enfoques presentan limitaciones importantes para capturar relaciones no lineales y patrones complejos en conjuntos de datos altamente heterogéneos.

La literatura reciente evidencia un crecimiento sostenido en el uso de técnicas de Machine Learning y Aprendizaje Profundo para la estimación de la probabilidad de pago e incumplimiento. Estudios comparativos muestran que modelos como redes neuronales artificiales, Gradient Boosting y otros enfoques avanzados superan consistentemente a los métodos tradicionales en términos de capacidad predictiva, especialmente cuando se incorporan variables conductuales y transaccionales (Mestiri, 2024; Ayari, Guetari, & Kraïem, 2025).

En particular, el Aprendizaje Profundo ha demostrado ser eficaz para modelar patrones no lineales y dinámicas temporales presentes en el comportamiento crediticio, reduciendo la dependencia de una ingeniería manual intensiva de variables y mejorando la generalización del modelo (Hayashi, 2022; Gicić & Donko, 2025). No obstante, gran parte de estos trabajos se concentra en la mejora del desempeño predictivo del scoring, manteniendo un enfoque predominantemente unidimensional.

Diversas investigaciones han extendido el uso del scoring predictivo hacia la gestión de cobranza, utilizando la probabilidad de pago para priorizar clientes o definir estrategias diferenciadas por nivel de mora. Estos enfoques han demostrado mejoras en eficiencia

operativa frente a esquemas puramente basados en reglas (Carrera & Benalcázar, 2025). Sin embargo, la literatura señala que la mayoría de estas aplicaciones continúan abordando la decisión desde un único criterio principal, sin integrar explícitamente otros factores relevantes como la exposición crediticia o las restricciones operativas.

En este contexto, se identifica la necesidad de integrar los resultados de los modelos predictivos dentro de esquemas de apoyo a la toma de decisiones que consideren simultáneamente distintos criterios relevantes para la gestión de cobranza. La literatura reciente destaca la importancia de combinar indicadores de riesgo, probabilidad de pago, exposición crediticia y capacidad de contacto para definir estrategias diferenciadas de gestión, permitiendo asignar de manera más eficiente los recursos disponibles y mejorar la efectividad operativa de los procesos de recuperación (Simumba, Okami, Kodaka, & Kohtake, 2021; Sánchez, Maldonado, & Vairetti, 2022).

A pesar de estos avances, aún existen vacíos relevantes en aplicaciones dirigidas al sector retail con crédito directo, particularmente en contextos latinoamericanos, donde la integración de múltiples modelos predictivos con esquemas operativos de priorización y definición de estrategias continúa siendo limitada (Noriega, Rivera, & Herrera, 2023; Jankowski & Paliński, 2024). El presente trabajo propone un enfoque que integra modelos predictivos de comportamiento crediticio y contactabilidad dentro de un Modelo de Priorización de Gestión de Cobranza, proporcionando un mecanismo objetivo para apoyar la segmentación de cuentas y la definición de estrategias diferenciadas de gestión (Sánchez, Maldonado, & Vairetti, 2022).

Modelos de Predicción y Métricas de Evaluación

El credit scoring constituye un componente fundamental en la gestión del riesgo crediticio, orientado a estimar la probabilidad de incumplimiento o recuperación de los clientes a partir de información histórica. Tradicionalmente, este problema ha sido abordado mediante

modelos estadísticos como la regresión logística, ampliamente utilizada debido a su interpretabilidad, estabilidad y capacidad para modelar probabilidades de eventos binarios (Khandani, Kim, & Lo, 2010; Lessmann, Baesens, Seow, & Thomas, 2015). No obstante, estos enfoques presentan limitaciones para capturar relaciones no lineales y estructuras complejas presentes en los datos financieros.

La evolución reciente del credit scoring ha estado marcada por la incorporación de técnicas de Machine Learning y Aprendizaje Profundo, las cuales han demostrado mejoras significativas en la capacidad predictiva frente a los modelos tradicionales. Diversos estudios destacan que algoritmos como Random Forest, Gradient Boosting (incluyendo XGBoost) y otros métodos de ensamble permiten modelar interacciones complejas entre variables, logrando un mejor desempeño en escenarios con datos heterogéneos y altamente dimensionales (Ayari, Guetari, & Kraïem, 2025; Mestiri, 2024; Noriega, Rivera, & Herrera, 2023)

En particular, los métodos basados en boosting han sido ampliamente adoptados en aplicaciones reales de scoring crediticio, debido a su capacidad para optimizar funciones de pérdida y capturar patrones no lineales en grandes volúmenes de datos. Estos enfoques han demostrado una mayor capacidad discriminativa en comparación con modelos lineales, especialmente en contextos donde el comportamiento del cliente depende de múltiples variables interrelacionadas (Kamimura, Pinto, & Nagano, 2023; Carrera & Benalcázar, 2025).

Dentro de los modelos supervisados también se encuentran los métodos basados en similitud, como K-Nearest Neighbors (KNN), los cuales realizan predicciones a partir del comportamiento de observaciones cercanas en el espacio de características. Si bien estos modelos pueden ofrecer resultados competitivos en ciertos escenarios, su aplicación

práctica se ve limitada por su costo computacional y su sensibilidad a la dimensionalidad de los datos, lo que restringe su escalabilidad en entornos empresariales de gran tamaño.

Por otro lado, el Aprendizaje Profundo ha ganado relevancia en el ámbito financiero debido a su capacidad para modelar patrones complejos, relaciones no lineales y dinámicas temporales en datos estructurados y conductuales. En particular, las redes neuronales tipo Multi-Layer Perceptron (MLP) han sido utilizadas en credit scoring con resultados favorables, especialmente cuando se dispone de información rica en comportamiento histórico. La literatura destaca que estos modelos pueden superar a los enfoques tradicionales en términos de desempeño predictivo, aunque introducen desafíos asociados a la interpretabilidad y la transparencia (Hayashi, 2022; Gicić & Đonko, 2025; Shi, Tang, & Yu, 2025).

En este contexto, también adquiere relevancia el estudio de aspectos como la explicabilidad, equidad y sesgo en modelos de inteligencia artificial aplicados al crédito. Investigaciones recientes resaltan la importancia de balancear el desempeño predictivo con la interpretabilidad, especialmente en entornos regulados donde las decisiones deben ser justificables y transparentes (Bahlool, Hewahi, & Elmedany, 2026; Yeo, y otros, 2025; Černevičienė & Kabašinskas, 2024).

Paralelamente a los modelos supervisados, los métodos de aprendizaje no supervisado han sido ampliamente utilizados para analizar la estructura subyacente del portafolio de clientes. En particular, algoritmos como K-Means permiten identificar agrupaciones naturales en función de las características de los clientes, facilitando la segmentación y la comprensión de patrones de comportamiento sin requerir una variable objetivo. Este tipo de técnicas resulta especialmente útil en escenarios donde se busca caracterizar perfiles de riesgo o identificar grupos homogéneos dentro de la cartera, aportando valor en la interpretación de los resultados del scoring y en la definición de estrategias diferenciadas. Sin embargo, estos

enfoques no permiten predecir directamente la probabilidad de recuperación o incumplimiento, sino que actúan como herramientas complementarias para el análisis exploratorio de la información (Bermúdez Vera, Mosquera Restrepo, & Manotas-Duque, 2025).

En cuanto a la evaluación de los modelos de credit scoring, la literatura coincide en el uso de métricas orientadas a medir la capacidad discriminativa y la calidad de clasificación.

Entre las más relevantes se encuentra el Área Bajo la Curva ROC (AUC-ROC), ampliamente utilizada para evaluar la capacidad del modelo de distinguir entre clases, particularmente cuando existe desbalance de datos (Yang, Dong, Lu, & Peng, 2022).

Adicionalmente, se emplean métricas como precisión y recall, para analizar el desempeño en términos de errores de clasificación, así como la matriz de confusión, la cual facilita una interpretación detallada de los resultados desde una perspectiva operativa.

Estas métricas comparan el desempeño de modelos tradicionales y enfoques basados en inteligencia artificial, proporcionando evidencia cuantitativa sobre las mejoras en la estimación del comportamiento crediticio (Ayari, Guetari, & Kraïem, 2025). No obstante, la literatura reciente señala que la evaluación del desempeño predictivo, aunque necesaria, no es suficiente para apoyar la toma de decisiones en procesos como la gestión de cobranza, donde intervienen múltiples objetivos en conflicto (Chehrazi, Glynn, & Weber, 2019), tales como la maximización de la recuperación, la minimización del riesgo de mora y la optimización del uso de recursos operativos.

En este sentido, diversos trabajos han avanzado hacia la formulación del problema de cobranza como un proceso de decisión multiobjetivo, integrando métricas predictivas con criterios de negocio que reflejan la complejidad real del entorno financiero (Simumba, Okami, Kodaka, & Kohtake, 2021; Jones & Florentino, 2022). Este enfoque reconoce que no existe una única solución óptima, sino un conjunto de soluciones eficientes que

representan distintos compromisos entre objetivos, concepto formalizado a través del Frente de Pareto.

Para abordar este tipo de problemas, se propone el uso de algoritmos de optimización evolutiva multiobjetivo, entre los cuales destaca el Non-Dominated Sorting Genetic Algorithm II (NSGA-II), uno de los métodos más utilizados para la generación de Frentes de Pareto en problemas complejos. Este algoritmo permite identificar un conjunto diverso de soluciones no dominadas, garantizando simultáneamente convergencia hacia la frontera óptima y diversidad en las alternativas encontradas (Ma, Zhang, Sun, Liu, & Shan, 2023).

De esta manera, el enfoque multiobjetivo complementa los modelos de scoring tradicionales al incorporar explícitamente los trade-offs entre criterios de decisión, proporcionando un marco más robusto para la definición de estrategias en la gestión de recuperación de cartera.

Marco Teórico

Fundamentos de credit scoring y comportamiento del cliente

El credit scoring puede entenderse como un conjunto de métodos analíticos orientados a estimar el nivel de riesgo de un cliente a partir de información histórica relacionada con su comportamiento financiero. Su propósito principal es transformar datos dispersos en una medida útil para la toma de decisiones, especialmente en organizaciones que administran grandes volúmenes de clientes y necesitan distinguir entre distintos perfiles de riesgo.

Aunque los primeros enfoques de scoring se apoyaban principalmente en técnicas estadísticas tradicionales, la evolución reciente del análisis financiero ha incorporado modelos de aprendizaje automático e inteligencia artificial, capaces de identificar patrones más complejos en los datos crediticios (Dastile, Celik, & Potsane, 2020).

Dentro de este análisis, el comportamiento histórico del cliente cumple un papel central. La forma en que una persona ha pagado en el pasado, la frecuencia de sus atrasos, el uso que hace del crédito, la evolución de su mora y su nivel de exposición financiera constituyen

señales relevantes para comprender su perfil. Por ello, el análisis histórico no se limita a describir hechos pasados, sino que permite anticipar posibles escenarios futuros y fortalecer procesos como la evaluación del riesgo, la priorización de cartera y la gestión de cobranza. En este sentido, el credit scoring se convierte en una herramienta estratégica porque permite pasar de decisiones generales o intuitivas hacia decisiones basadas en evidencia (Noriega, Rivera, & Herrera, 2023).

Los modelos de inteligencia artificial aplicados al credit scoring han ampliado las posibilidades de análisis, especialmente porque permiten trabajar con grandes volúmenes de información, variables heterogéneas y relaciones no lineales entre los datos. Sin embargo, también se reconoce que estos modelos deben ser evaluados no solo por su capacidad predictiva, sino también por su transparencia, equidad y posibilidad de explicación, debido a que sus resultados pueden influir en decisiones financieras sensibles (Černevičienė & Kabašinskas, 2024; Babaei, 2023; Kvamme, 2022).

Ventanas de tiempo en el análisis del comportamiento crediticio

En los modelos de credit scoring y gestión de cobranza, las ventanas de tiempo constituyen un elemento fundamental para comprender la evolución del comportamiento financiero del cliente. Una ventana de tiempo puede entenderse como el periodo histórico definido para observar, medir y transformar variables relacionadas con pagos, mora, uso del crédito, exposición financiera y regularización. En lugar de analizar únicamente una fotografía del cliente en un momento específico, este enfoque permite construir una visión más completa de su trayectoria crediticia y de los cambios que pueden anticipar riesgo o recuperación. (Shi, Tse, Luo, D'Addona, & Pau, 2022)

El uso de ventanas de tiempo es relevante porque el comportamiento de un cliente no siempre es estable. Una persona puede presentar un atraso puntual en un mes, pero mantener un historial positivo en periodos anteriores; de igual manera, otro cliente puede no

registrar una mora severa en el momento actual, pero mostrar una tendencia progresiva de deterioro. Por esta razón, variables como recencia, frecuencia de pago, eventos de mora, máximo de días de mora y tendencia de mora adquieren mayor valor cuando se analizan dentro de un periodo histórico definido. La información temporal permite diferenciar entre clientes con comportamientos ocasionales y clientes con patrones persistentes de riesgo. (Mestiri, 2024; Noriega, Rivera, & Herrera, 2023)

Desde el punto de vista metodológico, la selección de una ventana temporal adecuada permite organizar los datos de manera coherente para el entrenamiento de modelos predictivos, en este sentido, la ventana de tiempo no solo cumple una función descriptiva, sino también predictiva, porque permite que el modelo aprenda patrones anteriores y los utilice para estimar la probabilidad de recuperación futura (Shi, Tse, Luo, D'Addona, & Pau, 2022).

Las ventanas de tiempo también permiten construir variables agregadas que resumen la conducta del cliente durante un periodo determinado. Estas variables ayudan a transformar datos transaccionales en indicadores interpretables para el modelo. Los modelos de Machine Learning aplicados al riesgo crediticio se fortalecen cuando incorporan información histórica y variables conductuales que permiten representar mejor la complejidad del cliente (Noriega, Rivera, & Herrera, 2023; Xu, y otros, 2024; Dastile, Celik, & Potsane, 2020).

En el ámbito de la cobranza, las ventanas de tiempo permiten distinguir entre diferentes momentos del comportamiento crediticio. Una ventana corta puede ser útil para identificar señales recientes de riesgo o recuperación, mientras que una ventana más amplia permite observar patrones de comportamiento más estables. Esta combinación ayuda a reducir decisiones basadas en hechos aislados y permite una segmentación más equilibrada (Mestiri, 2024; Bijak & Thomas, 2012).

Además, el análisis por ventanas temporales favorece la construcción de modelos más sensibles a la evolución del riesgo. Los modelos de Aprendizaje Profundo pueden capturar patrones complejos en datos financieros, especialmente cuando se dispone de información histórica suficiente para representar relaciones temporales y no lineales (Hayashi, 2022). Para proyectos de cartera, la ventana de tiempo se convierte en un criterio clave para generar modelos predictivos de gestión, su incorporación fortalece la investigación, ya que conecta el análisis histórico del cliente con la predicción mediante inteligencia artificial y con la posterior toma de decisiones en cobranza (Noriega, Rivera, & Herrera, 2023).

Modelos de predicción en credit scoring.

Los modelos de predicción en credit scoring son métodos analíticos que permiten estimar la probabilidad de que un cliente presente un determinado comportamiento financiero, como caer en mora, incumplir una obligación, mantenerse al día o regularizar una deuda. Estos modelos trabajan a partir de información histórica y buscan identificar patrones que permitan anticipar eventos futuros. En el ámbito financiero, su utilidad radica en transformar datos de comportamiento, pago, exposición crediticia y mora en estimaciones que apoyen la toma de decisiones de riesgo. El uso de modelos de Machine Learning y Deep Learning ha permitido mejorar la capacidad predictiva frente a métodos tradicionales, especialmente en escenarios con alta dimensionalidad y comportamientos crediticios complejos (Mestiri, 2024; Dastile, Celik, & Potsane, 2020).

La regresión logística es uno de los modelos estadísticos más utilizados en credit scoring debido a su sencillez, estabilidad e interpretación directa. Conceptualmente, este modelo estima la probabilidad de ocurrencia de un evento binario, por ejemplo, si un cliente incumplirá o no una obligación. Su principal fortaleza es que permite comprender la relación entre las variables explicativas y la probabilidad estimada. No obstante, su limitación aparece cuando los datos presentan relaciones no lineales o interacciones

complejas entre variables, situaciones frecuentes en carteras crediticias con comportamientos heterogéneos. Por esta razón, aunque continúa siendo una referencia importante, suele complementarse o compararse con modelos de aprendizaje automático más flexibles (Dastile, Celik, & Potsane, 2020).

El modelo Random Forest pertenece a los métodos de aprendizaje automático basados en árboles de decisión. Su funcionamiento consiste en construir múltiples árboles a partir de diferentes subconjuntos de datos y variables, para luego integrar sus resultados en una predicción final. A diferencia de un árbol de decisión individual, Random Forest no depende de una sola estructura de reglas, sino de la combinación de varios árboles, lo que aumenta la robustez del modelo y reduce el riesgo de sobreajuste. Este tipo de modelo resulta útil en credit scoring porque puede manejar datos complejos, variables conductuales y relaciones no lineales sin requerir supuestos estadísticos estrictos. En el material revisado también se señala que Random Forest reduce la varianza y mejora la generalización al combinar árboles entrenados con subconjuntos aleatorios de características (Dastile, Celik, & Potsane, 2020; Goh & Lee, 2019).

El modelo XGBoost, o Extreme Gradient Boosting, también se basa en árboles de decisión, pero utiliza una lógica secuencial. Esto significa que cada nuevo árbol se construye para corregir los errores cometidos por los árboles anteriores. Su fortaleza está en que mejora progresivamente la predicción y suele obtener un alto rendimiento en datos tabulares, que son comunes en bases financieras y crediticias. En credit scoring, XGBoost permite capturar relaciones complejas entre variables como historial de pago, uso del crédito, nivel de mora, frecuencia de incumplimiento y comportamiento de regularización. La literatura reciente lo reconoce como uno de los modelos más competitivos para problemas de clasificación crediticia, especialmente cuando se busca mayor precisión predictiva (Noriega, Rivera, & Herrera, 2023).

Las redes neuronales artificiales son modelos inspirados en el procesamiento de información mediante capas de unidades o neuronas artificiales. Estas redes reciben variables de entrada, las transforman mediante funciones matemáticas y generan una salida predictiva. En credit scoring, su principal aporte consiste en aprender relaciones no lineales entre variables financieras, transaccionales y conductuales. Esto permite representar patrones que pueden ser difíciles de identificar mediante modelos estadísticos tradicionales. Sin embargo, su desempeño depende de la calidad de los datos, del diseño de la arquitectura y del proceso de entrenamiento (Shi, Tse, Luo, D'Addona, & Pau, 2022).

Las redes neuronales profundas, también conocidas como modelos de Deep Learning, amplían la lógica de las redes neuronales artificiales mediante el uso de varias capas ocultas. Estas capas permiten aprender representaciones más complejas de los datos y descubrir relaciones profundas entre variables. En el análisis crediticio, los modelos profundos son relevantes porque pueden trabajar con grandes volúmenes de información y capturar comportamientos financieros heterogéneos. Arquitecturas como Deep Belief Networks, redes convolucionales adaptadas a datos tabulares y redes profundas han mostrado ventajas frente a modelos tradicionales en distintos escenarios de credit scoring (Shi, Tse, Luo, D'Addona, & Pau, 2022).

Finalmente, los modelos híbridos combinan dos o más técnicas predictivas con el propósito de aprovechar las fortalezas de cada una. Por ejemplo, pueden integrar modelos de árboles con redes neuronales, técnicas de selección de variables con modelos de clasificación o enfoques de Machine Learning con métodos de optimización. En credit scoring, los modelos híbridos son relevantes porque permiten mejorar la capacidad predictiva, reducir limitaciones individuales de cada algoritmo y adaptarse mejor a problemas financieros complejos. La literatura reciente señala que los enfoques híbridos han mostrado resultados

favorables en la predicción del incumplimiento, especialmente cuando se trabaja con datos de alta dimensionalidad y múltiples variables de comportamiento (Mestiri, 2024).

En síntesis, los modelos de predicción en credit scoring han evolucionado desde métodos estadísticos interpretables, como la regresión logística, hacia modelos más flexibles y complejos, como Random Forest, XGBoost, redes neuronales artificiales, redes profundas y enfoques híbridos. Cada modelo ofrece ventajas particulares: la regresión logística facilita la interpretación; Random Forest aporta robustez; XGBoost destaca por su precisión en datos tabulares; las redes neuronales permiten modelar relaciones no lineales; y los modelos profundos e híbridos amplían la capacidad de aprendizaje en contextos financieros complejos (Dastile, Celik, & Potsane, 2020; Brown & Mues, 2012).

Métricas de clasificación y medición utilizadas en modelos predictivos.

Las métricas de clasificación y medición son indicadores que permiten evaluar el desempeño de un modelo predictivo. Su propósito es determinar qué tan bien el modelo logra comparar sus predicciones con los resultados reales. En los modelos de clasificación, estas métricas ayudan a identificar si el algoritmo asigna correctamente una categoría a cada caso; mientras que en los modelos de regresión permiten medir qué tan cercana es la predicción numérica al valor real observado. En el ámbito del credit scoring, la evaluación de modelos es fundamental porque no basta con generar una predicción; es necesario comprobar que dicha predicción sea confiable, interpretable y útil para la toma de decisiones financieras (Noriega, Rivera, & Herrera, 2023; Dastile, Celik, & Potsane, 2020). La matriz de confusión es una herramienta básica para evaluar modelos de clasificación. Consiste en una tabla que compara las clases reales con las clases predichas por el modelo. A partir de ella se identifican los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Esta matriz permite comprender no solo cuántos aciertos tuvo el modelo, sino también qué tipo de errores cometió. En problemas de riesgo crediticio, esta

información es importante porque no todos los errores tienen el mismo impacto; por ejemplo, clasificar como no riesgoso a un cliente que sí representa riesgo puede afectar directamente la calidad de la decisión crediticia (Noriega, Rivera, & Herrera, 2023).

El Accuracy o exactitud mide la proporción total de predicciones correctas realizadas por el modelo. Es una métrica general y fácil de interpretar, porque indica el porcentaje de casos bien clasificados sobre el total evaluado. Sin embargo, debe utilizarse con cuidado cuando existe desbalance entre clases, ya que un modelo puede obtener una exactitud alta clasificando correctamente la clase mayoritaria, pero fallando en la identificación de la clase más crítica. Esta situación es frecuente en datos crediticios, donde los casos de incumplimiento o mora severa suelen ser menos numerosos que los casos de cumplimiento normal (Yang, Dong, Lu, & Peng, 2022).

La Precision o precisión mide qué proporción de los casos que el modelo clasificó como positivos realmente pertenecen a esa clase. En otras palabras, indica qué tan confiables son las predicciones positivas del modelo. Esta métrica es útil cuando se desea reducir los falsos positivos, es decir, cuando se busca evitar que el modelo clasifique incorrectamente a un caso como perteneciente a una categoría de interés. En credit scoring, una baja precisión puede llevar a tomar decisiones innecesarias sobre clientes que no corresponden al perfil de riesgo identificado (Yang, Dong, Lu, & Peng, 2022).

El Recall o sensibilidad mide la capacidad del modelo para detectar correctamente los casos positivos reales. Responde a la pregunta: de todos los casos que realmente pertenecían a la clase positiva, ¿cuántos logró identificar el modelo? Esta métrica es especialmente importante cuando el costo de no detectar un caso positivo es alto, como ocurre en contextos donde se necesita identificar riesgos, incumplimientos o eventos críticos. Por ello, en modelos de riesgo crediticio, el recall ayuda a evaluar si el modelo realmente está

capturando los casos que requieren mayor atención (Noriega, Rivera, & Herrera, 2023; Dastile, Celik, & Potsane, 2020).

El F1-score combina la precisión y el recall en una sola medida. Se utiliza cuando se necesita equilibrar la capacidad del modelo para detectar los casos positivos y, al mismo tiempo, evitar clasificaciones incorrectas. Esta métrica es útil en escenarios donde existe desbalance de clases, porque ofrece una evaluación más equilibrada que el accuracy. En consecuencia, el F1-score permite valorar el desempeño del modelo considerando tanto la calidad de las predicciones positivas como la capacidad de identificar correctamente los casos relevantes (Yang, Dong, Lu, & Peng, 2022).

El ROC_AUC mide la capacidad del modelo para diferenciar entre clases. La curva ROC representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos en distintos umbrales de decisión. El AUC corresponde al área bajo esa curva. Mientras más cercano sea su valor a 1, mayor será la capacidad del modelo para distinguir correctamente entre una clase y otra. En credit scoring, esta métrica es ampliamente utilizada porque permite evaluar el poder discriminativo del modelo sin depender de un único punto de corte (Mestiri, 2024; Yang, Dong, Lu, & Peng, 2022).

En modelos de regresión se utilizan métricas como MAE, RMSE y R^2 . El MAE o error absoluto medio mide el promedio de las diferencias absolutas entre los valores reales y los valores predichos; por tanto, indica cuánto se equivoca el modelo en promedio. El RMSE o raíz del error cuadrático medio también mide el error de predicción, pero penaliza con mayor fuerza los errores grandes. El R^2 o coeficiente de determinación indica qué proporción de la variabilidad de la variable objetivo logra explicar el modelo. Estas métricas son útiles cuando el objetivo no es clasificar una categoría, sino estimar un valor numérico asociado al comportamiento financiero del cliente (Mestiri, 2024).

En síntesis, las métricas de clasificación y medición permiten evaluar si un modelo predictivo funciona de manera adecuada. La matriz de confusión muestra los aciertos y errores; el accuracy mide el rendimiento general; la precisión evalúa la confiabilidad de las predicciones positivas; el recall mide la capacidad de detección; el F1-score equilibra precisión y sensibilidad; el ROC_AUC mide la capacidad discriminativa; y las métricas MAE, RMSE y R^2 permiten evaluar modelos que predicen valores numéricos. Estas métricas son esenciales para determinar si un modelo no solo predice, sino si sus predicciones son útiles, confiables y técnicamente válidas en contextos financieros (Noriega, Rivera, & Herrera, 2023; Dastile, Celik, & Potsane, 2020).

Segmentación de clientes.

La segmentación de clientes es un proceso analítico que permite agrupar a los clientes de acuerdo con características comunes, tales como comportamiento de pago, nivel de riesgo, uso del crédito, exposición financiera o probabilidad de recuperación. En el ámbito del credit scoring y la gestión de cobranza, esta técnica permite comprender que no todos los clientes presentan el mismo perfil ni requieren la misma estrategia de gestión. Por ello, la segmentación contribuye a una toma de decisiones más diferenciada, eficiente y sustentada en evidencia, al transformar los datos históricos del cliente en grupos útiles para el análisis y la acción operativa (Carrera & Benalcázar, 2025).

Formulación de la cobranza como problema multiobjetivo.

La gestión de cobranza no puede reducirse únicamente a la intención de recuperar dinero. En la práctica, este proceso implica equilibrar distintos objetivos al mismo tiempo: maximizar la recuperación de cartera, reducir el riesgo de deterioro, optimizar el uso de recursos operativos y evitar esfuerzos innecesarios en cuentas con baja probabilidad de respuesta. Estos objetivos no siempre avanzan en la misma dirección, por lo que la cobranza

puede entenderse como un problema de decisión complejo y multiobjetivo (Chehrazi, Glynn, & Weber, 2019).

Desde esta perspectiva, una estrategia de cobranza puede ser efectiva en términos de recuperación, pero costosa en términos operativos. De igual manera, una estrategia menos intensiva puede reducir costos, pero disminuir la probabilidad de recuperación en ciertos segmentos. Esta tensión obliga a buscar equilibrios entre criterios que pueden entrar en conflicto. Por ello, la optimización multiobjetivo resulta adecuada, ya que permite analizar varias metas de manera simultánea sin imponer una única variable como criterio dominante (Simumba, Okami, Kodaka, & Kohtake, 2021; Bellotti & Crook, 2013).

Aplicada a la cobranza, esta lógica permite responder preguntas más estratégicas: qué clientes deben priorizarse, qué nivel de esfuerzo operativo conviene asignar, qué segmentos representan mejor oportunidad de recuperación y qué políticas permiten equilibrar riesgo, recuperación y recursos. La combinación de Machine Learning e investigación de operaciones puede contribuir a optimizar procesos de recuperación de deuda, especialmente cuando se busca asignar recursos de forma más eficiente (Shoghi, 2019; Chehrazi, Glynn, & Weber, 2019).

Dominancia de Pareto y soluciones no dominadas.

Uno de los conceptos centrales en la optimización multiobjetivo es la dominancia de Pareto. Una solución domina a otra cuando logra un mejor desempeño en al menos un objetivo sin empeorar el desempeño en los demás. A partir de este principio se identifican las soluciones no dominadas, es decir, aquellas que representan alternativas eficientes porque no pueden mejorar en un criterio sin sacrificar otro. El conjunto de estas soluciones conforma el denominado frente de Pareto (Jones & Florentino, 2022).

En términos prácticos, el frente de Pareto permite comprender que en muchos problemas reales no existe una única respuesta óptima, sino varias alternativas razonables, cada una

con ventajas y costos distintos. En la gestión de cobranza, esto significa que pueden existir diferentes políticas válidas según se priorice la recuperación, la reducción del riesgo o la eficiencia operativa. Por ejemplo, una política puede enfocarse en clientes con alta probabilidad de pago; otra, en clientes con mayor exposición crediticia; y otra, en casos con riesgo creciente de deterioro (Jones & Florentino, 2022).

Este enfoque es relevante porque evita imponer una decisión única antes de analizar los datos. En lugar de elegir arbitrariamente un solo objetivo, el análisis de Pareto permite observar distintos equilibrios posibles y seleccionar la alternativa más adecuada según la estrategia de la organización. En el contexto del credit scoring, resaltan que los enfoques multiobjetivo son útiles porque permiten considerar simultáneamente criterios de desempeño, riesgo y rentabilidad (Jones & Florentino, 2022; Xu, et al., 2024).

Algoritmos de optimización multiobjetivo.

Los algoritmos de optimización multiobjetivo han sido diseñados para encontrar soluciones eficientes en problemas donde varios objetivos entran en conflicto. Dentro de este grupo, los algoritmos evolutivos multiobjetivo han tenido una amplia aceptación por su capacidad para explorar espacios de solución complejos y generar conjuntos diversos de alternativas. Su lógica se inspira en procesos evolutivos, combinando mecanismos de selección, variación y supervivencia de soluciones con mejor desempeño relativo (Ma, Zhang, Sun, Liu, & Shan, 2023).

Uno de los métodos más conocidos en este ámbito es NSGA-II, un algoritmo ampliamente utilizado para ordenar soluciones según dominancia y preservar diversidad dentro del frente de Pareto. Su importancia radica en que no busca una sola respuesta final, sino un conjunto de soluciones eficientes que puedan apoyar al decisor. Este continúa siendo uno de los algoritmos de referencia en optimización multiobjetivo debido a su aplicabilidad en

problemas complejos y a su capacidad para generar frentes de Pareto diversos (Ma, Zhang, Sun, Liu, & Shan, 2023).

En problemas como la cobranza, donde intervienen varios objetivos simultáneos, este tipo de algoritmos ofrece una herramienta valiosa para explorar combinaciones de decisión. Por ejemplo, puede ayudar a identificar segmentos de clientes que maximizan recuperación esperada, reducen riesgo de mora o equilibran el uso de recursos operativos. De esta manera, la optimización multiobjetivo permite traducir la complejidad del problema en alternativas concretas de acción (Ma, Zhang, Sun, Liu, & Shan, 2023; Jones & Florentino, 2022).

Generación de políticas de cobranza basadas en Pareto.

La utilidad práctica de la optimización multiobjetivo se observa cuando las soluciones eficientes se convierten en políticas de gestión. En el caso de la cobranza, el uso del frente de Pareto permite transformar los resultados analíticos en estrategias operativas diferenciadas. En lugar de depender de una sola regla general para todos los clientes, la organización puede construir distintas políticas según el equilibrio deseado entre recuperación, riesgo y esfuerzo operativo (Jones & Florentino, 2022).

Estas políticas pueden orientar acciones más intensivas para segmentos con alta exposición y probabilidad de recuperación, estrategias preventivas para clientes con señales tempranas de deterioro o gestiones menos costosas para clientes con bajo riesgo. Los modelos basados en reglas y aprendizaje automático pueden mejorar la eficiencia de la cobranza de carteras masivas, al reducir acciones innecesarias y apoyar decisiones más sostenibles (Jankowski & Paliński, 2024).

En este sentido, las políticas basadas en Pareto permiten una mejor utilización de los recursos disponibles, ya que ayudan a seleccionar alternativas eficientes según las necesidades y restricciones de la organización. La cobranza deja de ser una gestión

uniforme y pasa a convertirse en un proceso más inteligente, donde cada segmento puede recibir una estrategia proporcional a su riesgo, probabilidad de recuperación y valor operativo.

Capítulo 3

Desarrollo del Trabajo

El desarrollo del presente trabajo de titulación tuvo como propósito diseñar e implementar una solución basada en Inteligencia Artificial y Aprendizaje Automático para optimizar la gestión de cobranza de una fintech, mediante la construcción de modelos predictivos capaces de apoyar la toma de decisiones, priorizar la gestión de cartera y contribuir a una asignación más eficiente de los recursos operativos.

La solución propuesta se estructuró de forma secuencial, iniciando con la consolidación y preparación del conjunto de datos históricos de cobranza provenientes de la operación de la organización. Sobre esta información se realizó un proceso de validación de calidad de datos, análisis de completitud, tratamiento de valores nulos, generación de variables derivadas y construcción de los targets necesarios para el entrenamiento de los modelos predictivos. De manera complementaria, se desarrolló un modelo de contactabilidad, este componente incorporó información histórica relacionada con canales de comunicación, resultados de contacto y características operativas de la gestión.

Posteriormente se desarrollaron cuatro modelos de Inteligencia Artificial orientados a resolver diferentes necesidades del proceso de cobranza. El primero corresponde al Modelo Predictivo de Mora, formulado como un problema de clasificación binaria para estimar la probabilidad de incumplimiento de pago. El segundo corresponde al Modelo Predictivo de Pago, planteado como un problema de regresión cuyo objetivo consiste en estimar el porcentaje esperado de recuperación económica de cada cuenta. El tercer modelo desarrolló el Modelo Predictivo de Regularización, también formulado como un problema de clasificación, destinado a estimar la probabilidad de que una obligación sea regularizada mediante pago o refinanciamiento. Finalmente, se desarrolló el Modelo de Contactabilidad,

formulado como un problema de clasificación, cuyo propósito fue estimar la probabilidad de establecer contacto efectivo con cada cliente.

Para cada uno de estos modelos se aplicó una metodología homogénea que incluyó la validación del target, el análisis exploratorio de datos, la selección de variables, la evaluación de multicolinealidad, el entrenamiento de diferentes algoritmos de aprendizaje supervisado, la comparación de métricas de desempeño y la selección del modelo con mejor capacidad predictiva. Una vez identificado el algoritmo con mejor desempeño para cada problema, se realizó el reentrenamiento utilizando la totalidad de los registros etiquetados, obteniendo los modelos definitivos destinados a su utilización en ambientes productivos. Como etapa posterior, las predicciones generadas por los tres modelos fueron integradas en un único conjunto de datos denominado `dfGestion`, sobre el cual se construyó un `ScorePrioridadGestion`. Este score constituye un índice compuesto diseñado para priorizar la gestión de cobranza considerando simultáneamente el riesgo de mora, el potencial de recuperación, la probabilidad de regularización y la exposición económica de cada cuenta. A partir de este indicador se segmentó la cartera en deciles de gestión y se clasificó cada cuenta según su tipo de gestión.

Los indicadores generados por los modelos predictivos fueron utilizados como insumo para la construcción del Score de Priorización de Gestión y la posterior definición de estrategias de cobranza. Para ello, se integraron criterios relacionados con la recuperación esperada, el riesgo de mora, la probabilidad de regularización y la exposición económica de cada cuenta. Posteriormente, estos resultados se complementaron con el Tipo de Gestión, el Decil de Gestión y el Nivel de Contactabilidad, permitiendo asignar estrategias diferenciadas orientadas a mejorar la priorización de clientes y la utilización de los recursos operativos vinculados al proceso de cobranza.

Como resultado final, se implementó una solución integral capaz de transformar información histórica de clientes en recomendaciones operativas y estrategias diferenciadas de recuperación de cartera. Los resultados fueron almacenados en estructuras de SQL Server, complementados con modelos persistidos en archivos de tipo PKL, reportes analíticos y visualizaciones de soporte, constituyendo una prueba de concepto funcional para futuras integraciones dentro de los sistemas operativos de la organización. La solución desarrollada proporciona una base analítica para mejorar la focalización de recursos y fortalecer la gestión de cobranza mediante la integración de modelos predictivos, priorización de cartera y contactabilidad.

Descripción Datasets

La presente investigación utilizó dos conjuntos de datos principales construidos específicamente para apoyar el desarrollo de modelos predictivos y mecanismos de optimización aplicados a la gestión de cobranza: DSScoreCobranza y DSScoreContacto. Ambos datasets fueron obtenidos a partir de información histórica almacenada en SQL Server y diseñados para representar diferentes dimensiones del comportamiento de los clientes dentro del proceso de recuperación de cartera.

Dataset DSScoreCobranza

El dataset DSScoreCobranza fue concebido para modelar fenómenos relacionados con riesgo de mora, recuperación económica y regularización de obligaciones. Este conjunto consolidó información histórica de comportamiento de pago, exposición económica, utilización de crédito, severidad de mora, cumplimiento financiero y otras variables derivadas orientadas al desarrollo de modelos predictivos supervisados. El dataset estuvo conformado por 573,328 registros y 47 variables, constituyendo la principal fuente de información para los modelos de Mora, Pago, Regularización y para la construcción del Score de Priorización de Gestión.

Análisis del Dataset construido para Mora, Pago, Regularización y Score de priorización de gestión. Con el fin de garantizar la reproducibilidad del proceso y la trazabilidad de la información utilizada durante la investigación, se documentó detalladamente cada grupo de variables incorporadas al Dataset. Esta clasificación facilita la comprensión del aporte de cada conjunto de características dentro del proceso de entrenamiento de los modelos predictivos y permite justificar su inclusión desde una perspectiva metodológica y de negocio.

El Dataset fue diseñado con el propósito de construir modelos predictivos orientados a la gestión inteligente de cobranza. Para ello, se empleó una estrategia de ventanas temporales móviles, mediante la cual se construyeron seis muestras consecutivas. Cada muestra estuvo conformada por una ventana histórica de seis meses utilizada para la generación de las variables explicativas (features) y una ventana posterior de tres meses destinada a la definición de las variables objetivo (targets). Posteriormente, las seis muestras fueron consolidadas en un único Dataset para el entrenamiento y evaluación de los modelos predictivos. Esta metodología garantiza que toda la información utilizada por los modelos corresponde a datos disponibles antes del período de predicción, evitando fuga de información (data leakage), incorporando diferentes momentos temporales del comportamiento de los clientes y reproduciendo un escenario más cercano a las condiciones reales de operación.

Figura 1 Construcción del Dataset mediante ventanas temporales móviles

	Ventana de Variables Explicativas (6 meses)	Ventana de Variables Objetivo (3 meses)
Muestra 1	Feb 2025 → Jul 2025	Ago 2025 → Oct 2025
Muestra 2	Mar 2025 → Ago 2025	Sep 2025 → Nov 2025
Muestra 3	Abr 2025 → Sep 2025	Oct 2025 → Dic 2025
Muestra 4	May 2025 → Oct 2025	Nov 2025 → Ene 2026
Muestra 5	Jun 2025 → Nov 2025	Dic 2025 → Feb 2026
Muestra 6	Jul 2025 → Dic 2025	Ene 2026 → Mar 2026

Nota. Elaboración Propia

El conjunto de variables fue diseñado para capturar múltiples dimensiones del comportamiento financiero del cliente: historial de cumplimiento, severidad y recurrencia de la mora, exposición monetaria, calidad de los pagos, utilización del crédito y señales recientes de deterioro. Las variables fueron diseñadas para representar no solo el estado actual de la cartera, sino también los patrones históricos de comportamiento con capacidad predictiva sobre el riesgo de mora, el potencial de recuperación y la probabilidad de regularización de cada cuenta. La combinación de estas características proporciona una representación integral del riesgo y de la capacidad de recuperación de cada cuenta, permitiendo entrenar modelos predictivos robustos que posteriormente servirán como insumo para los procesos de priorización de cartera, definición de estrategias diferenciadas de cobranza y evaluación operativa de la solución propuesta.

Variables objetivo

Target	Descripción	Justificación
Mora	Identifica si una cuenta alcanza una mora igual o superior a 45 días dentro de la ventana objetivo.	Permite estimar el riesgo futuro de deterioro de la cartera y constituye un indicador fundamental para priorizar acciones preventivas de cobranza. El umbral de 45 días fue seleccionado debido a que el umbral inicialmente considerado (90 días) presentaba una frecuencia insuficiente para el entrenamiento de modelos predictivos robustos.
Pago	Representa la proporción de recuperación monetaria efectiva durante la ventana objetivo, calculada considerando exclusivamente pagos reales y excluyendo refinanciamientos.	Este target no solo permite determinar si un cliente realiza pagos, sino también cuantificar el porcentaje del monto recuperado, proporcionando una medida continua del potencial de recuperación económica.
Regularizacion	Indica si una cuenta logró regularizar su situación durante la ventana objetivo mediante un pago efectivo o un refinanciamiento.	Aunque no representa recuperación de efectivo, refleja la capacidad de normalización de una obligación y constituye una métrica complementaria útil para el análisis estratégico de la cartera.

Consideración metodológica. El TargetRegularizacion se definió con valor igual a uno

cuando, durante el horizonte objetivo, la cuenta registró un pago efectivo o un

refinanciamiento que permitió regularizar administrativamente la obligación. Esta

definición responde a que el objetivo del modelo es estimar la probabilidad de

normalización de la cartera mediante cualquiera de los mecanismos aceptados por la

fintech.

Variables de comportamiento histórico

Variable	Descripción	Justificación
NumPeriodos	Número de períodos históricos considerados para cada cuenta.	Proporciona una medida de antigüedad y disponibilidad de información histórica, permitiendo diferenciar clientes con trayectorias extensas de aquellos con escaso historial.
RatioCumplimiento	Proporción de estados cancelados mediante pagos efectivos respecto del total de estados analizados	Resume el comportamiento histórico de pago y constituye uno de los principales indicadores de disciplina financiera.
RatioCumplimiento3m	Calcula el mismo indicador anterior restringido a los tres meses más recientes.	Captura cambios recientes en el comportamiento del cliente, permitiendo detectar deterioros o mejoras de corto plazo.
DeltaCumplimiento	Diferencia entre el cumplimiento reciente y el cumplimiento histórico.	Mide la tendencia del comportamiento de pago. Valores negativos indican deterioro reciente, mientras que valores positivos reflejan mejoras en el hábito de pago.
RatioRefinanciamiento	Proporción de estados regularizados mediante refinanciamiento.	Permite identificar clientes que dependen frecuentemente de procesos de reestructuración en lugar de pagos efectivos.

NumRefinanciamientos	Número de refinanciamientos registrados durante la ventana histórica.	Complementa el indicador anterior cuantificando la recurrencia de este tipo de operaciones.
MesesConsecutivosSinPago	Cantidad de estados consecutivos pendientes al cierre de la ventana histórica.	Refleja persistencia en el incumplimiento y constituye un indicador temprano de deterioro financiero.
DiasDesdeUltimoPago	Número de días transcurridos desde el último pago registrado hasta la fecha de corte.	Permite medir la recencia de la actividad de pago del cliente.
FlagSinPagosHistoricos	Identifica cuentas que no registran pagos efectivos durante toda la ventana histórica.	La ausencia total de pagos constituye una señal de riesgo por sí misma y aporta información adicional a los modelos predictivos.

Consideración metodológica. En los casos que un cliente no registró pagos históricos, la variable `DiasDesdeUltimoPago` tomó el valor 999 como indicador explícito de ausencia de pagos previos. Esta decisión evitó imputaciones artificiales como valores cero o promedios y permitió conservar una condición que potencialmente aporta capacidad predictiva al modelo.

Variables relacionadas con la mora

Variable	Descripción	Justificación
MaxMora	Máximo número de días de mora observado en el historial.	Resume el peor comportamiento histórico registrado.
MoraPromedio	Promedio de días de mora considerando estados efectivamente pagados.	Representa el patrón habitual de atraso del cliente.
AnticipoPromedio	Promedio de días de anticipación en pagos realizados antes del vencimiento.	Permite identificar clientes con hábitos de pago anticipado.
RatioMora	Proporción de estados que presentaron mora.	Resume la frecuencia relativa de incumplimientos.
RatioPuntualidad	Proporción de estados cancelados puntualmente o antes del vencimiento.	Mide disciplina de pago desde una perspectiva positiva.
VolatilidadMora	Desviación estándar de los días de mora.	Evalúa la estabilidad o variabilidad del comportamiento histórico.
PeorMoraHistorica	Mayor nivel de atraso registrado.	Refleja eventos extremos de incumplimiento.
MaxMoraPagada	Máximo atraso registrado entre estados finalmente cancelados.	Diferencia clientes que eventualmente regularizan sus obligaciones pese a presentar atrasos significativos.
MaxMora3m	Máxima mora registrada en los últimos tres meses.	Capturan el comportamiento reciente, generalmente más predictivo que el histórico completo.
RecenciaUltimoEstado	Representa el número de días transcurridos entre la fecha del último estado de cuenta considerado en la ventana de características y la fecha de referencia utilizada para la construcción del dataset.	Utilizada como una aproximación del nivel de mora vigente al momento de realizar la predicción.
DeterioroReciente	Variable binaria que identifica la existencia de mora significativa durante los últimos tres meses.	Actúa como señal temprana de incremento del riesgo.

FlagDefaultHistorico	Indica si la cuenta alcanzó previamente una mora igual o superior a 45 días.	El comportamiento pasado suele constituir uno de los mejores predictores del riesgo futuro.
-----------------------------	--	---

Variables monetarias

Variable	Descripción	Justificación
ExposicionTotal	Monto acumulado facturado durante la ventana histórica.	Representa la magnitud económica administrada para cada cuenta.
ExposicionImpaga	Monto pendiente correspondiente exclusivamente a obligaciones no canceladas y no refinanciadas.	Refleja exposición efectiva al riesgo de incumplimiento.
RatioImpagoReal	Proporción de exposición pendiente respecto del total facturado.	Normaliza la exposición para hacer comparables cuentas con distintos niveles de utilización.
RatioPagoParcial	Proporción de estados con pagos parciales.	Identifica clientes que muestran intención de pago, aunque sin cancelar completamente sus obligaciones.
RatioPagoReal	Proporción monetaria efectivamente recuperada mediante pagos reales respecto del total facturado.	Constituye uno de los principales indicadores históricos de recuperación y excluye refinanciamientos para representar únicamente flujo efectivo de caja.
ScoreCalidadPago	Indicador ponderado que asigna mayor valor a pagos completos y un valor intermedio a pagos parciales.	Resume la calidad del comportamiento de pago en una única métrica continua.
RatioPagosConMora	Proporción de pagos realizados con atraso respecto del total de pagos.	Distingue entre clientes que pagan oportunamente y aquellos que lo hacen de manera recurrentemente tardía.

Variables asociadas al cliente

Variable	Descripción	Justificación
EsClienteNuevo	Identifica cuentas con escaso historial disponible.	Los clientes recientes suelen presentar mayor incertidumbre por ausencia de comportamiento histórico suficiente.

Variables relacionadas con el cupo de crédito

Variable	Descripción	Justificación
CupoMaximo	Límite máximo que se puede asignar a un cliente / cuenta.	Estas variables reflejan el nivel de utilización del crédito y la capacidad financiera remanente del cliente, siendo indicadores relevantes de exposición y potencial riesgo.
CupoAprobado	Monto efectivamente aprobado para utilización.	
CupoUtilizado	Saldo utilizado por el cliente.	
CupoDisponible	Diferencia entre el cupo aprobado y el cupo utilizado.	
PorcUtilizado	Relación entre cupo utilizado y cupo aprobado.	
MargenDisponible	Proporción de cupo aún disponible respecto del cupo aprobado.	Mide la holgura financiera existente al momento de la evaluación.
RatioExposicionCupo	Relación entre la exposición impaga y el cupo aprobado.	Permite medir el nivel relativo de endeudamiento del cliente respecto de su capacidad crediticia disponible.
FlagSobreUtilizado	Indica si el saldo utilizado supera el cupo aprobado.	Detecta situaciones excepcionales de sobreexposición crediticia.

FlagCupoSuspendido	Identifica cuentas cuyo cupo aprobado fue reducido a cero mientras mantienen saldo utilizado.	Constituye una señal explícita de políticas internas de mitigación de riesgo aplicadas por la entidad y puede anticipar problemas futuros de recuperación.
---------------------------	---	--

Variables auxiliares para tratamiento de datos faltantes

Variable	Descripción	Justificación
RatioPagoRealNull RatioCumplimientoNull VolatilidadMoraNull	Estas variables indican si determinados atributos originales fueron imputados debido a la ausencia de información.	Permiten conservar la información implícita contenida en los valores faltantes y evaluar posteriormente si dicha ausencia tiene capacidad predictiva para los modelos de inteligencia artificial.

Dataset DSScoreContacto

El dataset DSScoreContacto fue desarrollado para analizar y modelar la probabilidad de contacto efectivo con los clientes. Este conjunto incorporó información relacionada con campañas de comunicación, historial de contacto, uso de canales, intensidad de gestión, frecuencia de interacción y variables operativas asociadas a la contactabilidad. El dataset estuvo compuesto por 518,887 registros y 114 variables, proporcionando una visión complementaria al componente financiero de la cobranza al permitir estimar la factibilidad de establecer contacto con cada cliente y apoyar la selección de estrategias operativas más eficientes.

Análisis del Dataset construido para Contactabilidad. Contiene la caracterización multicanal de la contactabilidad de los clientes. Cada fila representa un cliente con su historial de interacciones por teléfono, e-mail, push y visitas a local. Se utiliza para el modelo de contactabilidad que optimiza el canal y momento de contacto.

Variables objetivo

Variable	Descripción	Justificación
TargetContactabilidad	Variable binaria que indica si se logró establecer contacto efectivo con el cliente durante el horizonte objetivo.	Constituye la variable principal para el entrenamiento del Modelo Predictivo de Contactabilidad.
TargetIntensidad	Variable numérica que representa la intensidad de gestión requerida para lograr contacto.	Permite complementar el análisis de contactabilidad incorporando una perspectiva operativa sobre el esfuerzo necesario para generar interacción.

Variables de actividad telefónica

Variable	Descripción	Justificación
ContactosTelefonia7d, ContactosTelefonia30d, ContactosTelefonia90d	Número de intentos de contacto telefónico realizados durante los últimos 7, 30 y 90 días.	Permiten medir la intensidad histórica de gestión telefónica y analizar su relación con la probabilidad de contacto efectivo.
RatioTelefonia7d, RatioTelefonia30d, RatioTelefonia90d	Proporción de interacciones realizadas mediante telefonía respecto del total de contactos efectuados.	Identifican la relevancia relativa del canal telefónico dentro de la estrategia de comunicación utilizada con cada cliente.
ExitoTelefonia7d, ExitoTelefonia30d, ExitoTelefonia90d	Número de contactos telefónicos exitosos registrados en cada ventana temporal.	Constituyen una medida directa de la efectividad histórica del canal telefónico.
TasaExitoTelefonia7d, TasaExitoTelefonia30d, TasaExitoTelefonia90d	Relación entre los contactos telefónicos exitosos y el total de intentos efectuados.	Permiten estimar la probabilidad histórica de éxito del canal y aportan señales predictivas al modelo de contactabilidad.

Las variables telefónicas permiten medir simultáneamente volumen, frecuencia y efectividad del canal más utilizado dentro de los procesos de cobranza. Su incorporación facilita identificar clientes con mayor predisposición a responder llamadas y contribuye a optimizar la asignación de recursos operativos.

Variables de actividad por correo electrónico

Variable	Descripción	Justificación
ContactosEmail7d, ContactosEmail30d, ContactosEmail90d	Número de correos electrónicos enviados al cliente en distintos horizontes temporales.	Permiten medir la intensidad de comunicación digital mediante correo electrónico.
RatioEmail7d, RatioEmail30d, RatioEmail90d	Participación relativa del correo electrónico dentro del total de canales utilizados.	Facilitan identificar preferencias históricas de comunicación por parte del cliente.
ExitoEmail7d, ExitoEmail30d, ExitoEmail90d	Número de interacciones exitosas generadas mediante correo electrónico.	Constituyen un indicador de efectividad operativa del canal.
TasaExitoEmail7d, TasaExitoEmail30d, TasaExitoEmail90d	Proporción de correos exitosos respecto del total de envíos realizados.	Permiten evaluar la capacidad del canal para generar respuestas o interacciones exitosas.

Estas variables permiten evaluar el comportamiento histórico del cliente frente a campañas digitales y aportan información relevante para estimar la efectividad potencial de futuras comunicaciones.

Variables de actividad push

Variable	Descripción	Justificación
ContactosPush7d, ContactosPush30d, ContactosPush90d	Número de notificaciones push enviadas durante cada ventana temporal.	Capturan la intensidad de uso de canales móviles de contacto.
RatioPush7d, RatioPush30d, RatioPush90d	Participación relativa de notificaciones push respecto del total de contactos efectuados.	Permiten identificar clientes con mayor exposición a este canal de comunicación.
ExitoPush7d, ExitoPush30d, ExitoPush90d	Número de interacciones exitosas generadas mediante notificaciones push.	Reflejan la efectividad histórica de este canal digital.
TasaExitoPush7d, TasaExitoPush30d, TasaExitoPush90d	Relación entre contactos exitosos y total de notificaciones enviadas.	Permiten identificar patrones de respuesta y aceptación del canal móvil.

Las variables push complementan la visión omnicanal del cliente y permiten identificar segmentos con mayor afinidad hacia interacciones móviles.

Variables de interacción en locales

Variable	Descripción	Justificación
VisitasLocal7d, VisitasLocal30d, VisitasLocal90d	Número de visitas registradas a establecimientos físicos en diferentes horizontes temporales.	Proporcionan evidencia de interacción presencial con la organización.
UltimoLocal	Último local o establecimiento visitado por el cliente.	Permite identificar patrones recientes de interacción física.
LocalPredominante	Local con mayor frecuencia de visitas durante el período histórico analizado.	Facilita caracterizar hábitos de comportamiento geográfico y operacional.
CambioLocalReciente	Indica si el cliente modificó recientemente su patrón habitual de visita.	Puede reflejar cambios de comportamiento relevantes para la estrategia de contacto.

La incorporación de variables físicas complementa la información obtenida mediante canales digitales y permite construir una visión más completa de la interacción histórica del cliente.

Variables de intensidad y comportamiento multicanal

Variable	Descripción	Justificación
IndiceIntensidadPonderado30d	Indicador compuesto que resume el nivel de interacción acumulada durante los últimos 30 días.	Permite estimar la presión de contacto reciente ejercida sobre el cliente.
IndiceIntensidadPonderado90d	Indicador compuesto equivalente calculado sobre un horizonte de 90 días.	Captura patrones de interacción de mediano plazo.

Estas variables resumen la actividad operacional de múltiples canales en indicadores únicos, facilitando la captura de patrones complejos de interacción.

Variables de canal preferente

Variable	Descripción	Justificación
CanalPrincipal	Canal de comunicación predominante utilizado con el cliente.	Permite identificar la vía de contacto históricamente más utilizada o efectiva.

La identificación del canal principal constituye un elemento clave para determinar estrategias personalizadas de contacto y optimizar la asignación de recursos de cobranza.

La construcción de los Dataset siguió un proceso iterativo de validación y depuración, en el que se verificó la coherencia lógica de las variables derivadas, la ausencia de fugas de información temporal (*data leakage*) y el cumplimiento de restricciones de dominio. Se realizaron controles sobre rangos válidos, consistencia entre indicadores y correcta definición de las variables objetivo. Este proceso permitió garantizar que el conjunto de datos utilizado para el entrenamiento de los modelos representara fielmente el comportamiento histórico de los clientes y fuera adecuado para tareas predictivas de cobranza.

La utilización conjunta de ambos datasets permitió construir una visión integral del proceso de cobranza. Mientras DSScoreCobranza aportó información para responder quiénes deben ser priorizados en función de su riesgo y potencial de recuperación, DSScoreContacto permitió determinar cómo gestionar a dichos clientes considerando su probabilidad de contacto, los canales disponibles y la intensidad de gestión más adecuada. Esta integración constituye uno de los principales aportes metodológicos de la investigación, al combinar variables financieras, conductuales y operativas dentro de un esquema unificado de toma de decisiones basado en Inteligencia Artificial.

Preparación y procesamiento de datos

La etapa de preparación y procesamiento de datos tuvo como objetivo transformar la información consolidada en un conjunto de datos consistente, íntegro y adecuado para el

entrenamiento de los modelos de Inteligencia Artificial. Esta fase constituye una de las etapas más importantes del proceso analítico, ya que la calidad de los datos influye directamente en la capacidad predictiva de los modelos desarrollados.

Figura 2 *Proceso de preparación y procesamiento de datos para el entrenamiento de los modelos predictivos*



Nota. Elaboración Propia

Como parte de este proceso se realizó una revisión integral de la estructura del Dataset, verificando la consistencia de los tipos de datos, la identificación de valores nulos, la completitud de la información y la coherencia lógica entre las variables. Adicionalmente, se efectuaron controles sobre rangos válidos y relaciones entre indicadores financieros con el propósito de detectar posibles inconsistencias antes del entrenamiento de los modelos.

Posteriormente se construyeron las variables objetivo correspondientes a los problemas de clasificación y regresión definidos en la investigación. Para evitar fuga de información (*data leakage*), la construcción de los targets se realizó utilizando exclusivamente información perteneciente al horizonte de predicción, mientras que las variables

explicativas fueron generadas únicamente a partir de la ventana histórica definida para cada cuenta.

Con el propósito de preservar información potencialmente relevante para los modelos predictivos, no todos los valores faltantes fueron tratados mediante imputación. En aquellos casos donde la ausencia de información representaba una característica propia del comportamiento del cliente, se diseñaron variables indicadoras (*flags*) que permitieron conservar dicha condición como información predictiva. Asimismo, determinadas variables utilizaron valores convencionales previamente definidos cuando estos representaban situaciones de negocio claramente identificables, como la ausencia total de pagos históricos. Una vez completada la construcción del Dataset, se generó un proceso de documentación técnica orientado a garantizar la trazabilidad de la información utilizada durante la investigación. Esta documentación incluyó la descripción de variables, análisis de completitud, evaluación de valores nulos, caracterización estadística de los targets y registro de los principales indicadores descriptivos del conjunto de datos, constituyendo una evidencia metodológica que respalda la reproducibilidad del proceso desarrollado. Como resultado de esta etapa se obtuvo un Dataset preparado para el entrenamiento de los diferentes modelos predictivos, garantizando consistencia temporal, calidad de la información y representatividad de las características utilizadas durante el proceso de aprendizaje supervisado.

Desarrollo del Modelo Predictivo de Mora

Con el propósito de proporcionar una visión general de la metodología aplicada para el desarrollo del Modelo Predictivo de Mora, la Figura 3 resume las principales etapas ejecutadas durante su construcción, desde la definición de la variable objetivo hasta la implementación del modelo definitivo. Esta metodología constituye la base del proceso desarrollado y fue replicada posteriormente en los modelos predictivos de Pago y

Regularización, realizando únicamente las adaptaciones correspondientes a la naturaleza de cada variable objetivo.

Figura 3 Metodología para el desarrollo de los modelos predictivos



Nota. Elaboración Propia

Objetivo del Modelo Predictivo de Mora

El Modelo Predictivo de Mora fue desarrollado con el propósito de estimar la probabilidad de que una cuenta incurra en mora dentro del horizonte de predicción definido para la investigación. Este modelo constituye el primer componente de la solución propuesta y proporciona una medida cuantitativa del riesgo de incumplimiento, información fundamental para priorizar las acciones de cobranza y optimizar la asignación de recursos operativos.

Desde una perspectiva de aprendizaje automático, el problema fue formulado como una tarea de clasificación binaria, donde cada cuenta fue etiquetada de acuerdo con la ocurrencia o no de mora durante el período objetivo. Esta formulación permitió estimar, para cada

registro, la probabilidad de pertenecer a la clase correspondiente a clientes con riesgo de incumplimiento.

La construcción del modelo se realizó utilizando variables históricas relacionadas con el comportamiento de pago, antecedentes de mora, utilización de cupo, características financieras y variables derivadas obtenidas a partir de la información operacional de la fintech. Con el propósito de garantizar la validez temporal del proceso de entrenamiento, las variables explicativas fueron construidas exclusivamente a partir de información histórica disponible antes del horizonte de predicción, evitando la incorporación de información futura que pudiera introducir sesgos en el proceso de aprendizaje.

El desarrollo del Modelo Predictivo de Mora siguió una metodología estructurada que comprendió la construcción y validación del target, la preparación del conjunto de datos, la selección de variables, el entrenamiento de diferentes algoritmos de clasificación y la implementación del modelo definitivo para su utilización en ambientes productivos. Cada una de estas etapas se describe en los apartados siguientes.

Construcción y validación del TargetMora

La construcción del TargetMora constituyó la primera etapa para el desarrollo del modelo predictivo, ya que definió la variable objetivo sobre la cual se entrenaron los algoritmos de clasificación supervisada. El objetivo consistió en identificar aquellas cuentas que, durante el horizonte de predicción establecido para la investigación, alcanzaron una condición de mora igual o superior a cuarenta y cinco días.

La variable fue definida como un indicador binario, asignando el valor de **1** a las cuentas que registraron una mora mayor o igual a 45 días dentro del período objetivo y el valor de **0** a aquellas que no alcanzaron dicha condición. Esta definición permitió transformar el problema de negocio en un problema de clasificación binaria, donde el modelo estima la probabilidad de incumplimiento futuro de cada cuenta.

La selección del umbral de 45 días respondió a criterios operativos establecidos por la fintech, considerando que dicho nivel de mora representa el punto a partir del cual la gestión de cobranza requiere una intervención más intensiva y diferenciada. Esta decisión metodológica permitió alinear el modelo predictivo con los procesos reales de gestión implementados por la organización, garantizando que las predicciones obtenidas fueran directamente aplicables a la operación.

Una vez construido el target, se realizó un proceso de validación orientado a verificar su consistencia y distribución dentro del conjunto de datos. Para ello se analizaron la frecuencia de cada clase, el balance entre cuentas con y sin mora y los principales indicadores descriptivos del Dataset utilizado para el entrenamiento. Esta validación permitió confirmar que la variable objetivo representaba adecuadamente el fenómeno de estudio y proporcionaba una base consistente para el desarrollo del modelo predictivo.

La información obtenida durante esta etapa fue documentada mediante estadísticas descriptivas y resúmenes del Dataset, los cuales constituyen evidencia del proceso de validación realizado y sirven como soporte para la reproducibilidad de la investigación.

Con el propósito de verificar la consistencia del Dataset utilizado para el entrenamiento del Modelo Predictivo de Mora, se elaboró un resumen descriptivo del Dataset y de la distribución del TargetMora. Este análisis permitió confirmar la integridad de la información y la adecuada representación del fenómeno de estudio.

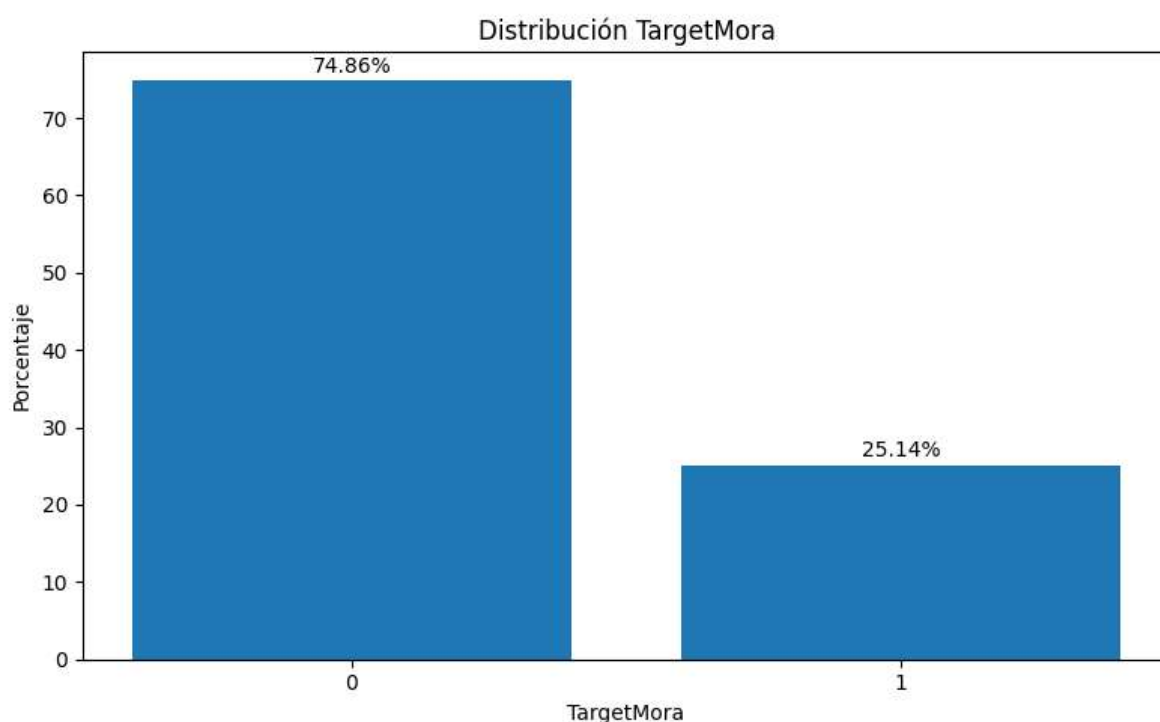
Tabla 1 *Resumen estadístico del Dataset utilizado para el entrenamiento del Modelo Predictivo de Mora*

Indicador	Valor
Total registros dataset completo	573,328
Total columnas dataset completo	47
Registros con TargetMora disponible	496,879
Registros sin TargetMora	76,449
Proporción TargetMora = 0	74.86%
Proporción TargetMora= 1	25.14%

Nota. Elaboración Propia

Como se observa en la tabla, el Dataset original estuvo conformado por 573,328 registros y 47 variables. De este conjunto, 496,879 registros disponían de información suficiente para la construcción del TargetMora, mientras que los 76,449 registros restantes fueron reservados inicialmente para el proceso de predicción. La distribución de la variable objetivo evidencia que el 25.14% de las observaciones corresponden a cuentas que alcanzaron una mora igual o superior a 45 días durante el horizonte de predicción, proporción que constituye una base adecuada para el entrenamiento de modelos de clasificación supervisada.

Figura 4 Distribución del TargetMora utilizado para el entrenamiento del Modelo Predictivo de Mora



Nota. Elaboración Propia

La figura evidencia la distribución de la variable objetivo, confirmando la naturaleza binaria del problema y justificando la utilización de modelos de clasificación supervisada para estimar la probabilidad de mora de cada cuenta.

Preparación del conjunto de datos

Una vez validado el TargetMora, se procedió a preparar el conjunto de datos que sería utilizado durante el entrenamiento del Modelo Predictivo de Mora. Esta etapa tuvo como finalidad construir una matriz de variables explicativas consistente con la variable objetivo, garantizando la calidad de la información y evitando la incorporación de datos que pudieran introducir sesgos en el proceso de aprendizaje.

Figura 5 *Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Mora*



Nota. Elaboración Propia

Como primer paso, se seleccionaron únicamente los registros que disponían de un valor conocido para el TargetMora, obteniéndose el conjunto de observaciones utilizado para el entrenamiento y evaluación de los algoritmos de clasificación. Los registros sin información del target fueron excluidos temporalmente de esta etapa, ya que posteriormente serían utilizados durante el proceso de implementación para generar las predicciones correspondientes.

Posteriormente se definió la matriz de variables explicativas (XMora) eliminando aquellas variables que no debían intervenir en el entrenamiento del modelo. Entre ellas se incluyeron el identificador único de la cuenta (IdCuenta), la variable Muestra y la variable FechaObservacion, así como las variables objetivo correspondientes a los demás modelos predictivos desarrollados en la investigación. La variable IdCuenta se utilizó únicamente

como identificador de cada registro, mientras que Muestra y FechaObservacion permitieron controlar y mantener la trazabilidad temporal durante la construcción del Dataset. Ninguna de estas variables fue utilizada como predictor durante el entrenamiento de los modelos, con el propósito de evitar fuga de información (data leakage) y garantizar que el aprendizaje se realizara exclusivamente a partir de información disponible antes del horizonte de predicción.

Finalmente, la variable TargetMora fue definida como variable dependiente (yMora), conformando el conjunto de datos que sirvió como punto de partida para las etapas posteriores de selección de variables, entrenamiento y evaluación de los diferentes algoritmos de aprendizaje supervisado.

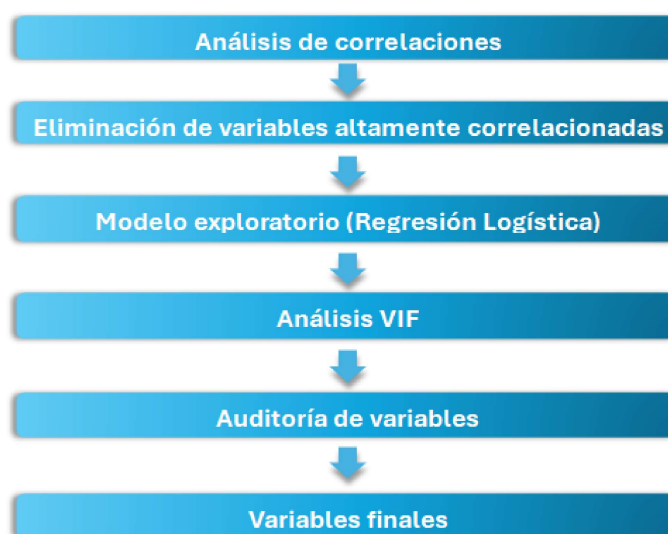
Selección de variables

La selección de variables constituyó una etapa fundamental dentro del desarrollo del Modelo Predictivo de Mora, ya que permitió identificar el conjunto de características con mayor capacidad para representar el comportamiento de pago de los clientes, reduciendo simultáneamente la redundancia de información y el riesgo de sobreajuste durante el entrenamiento de los modelos.

El proceso de selección se desarrolló de forma secuencial y combinó criterios estadísticos, analíticos y de conocimiento del negocio. En una primera etapa se evaluó la relación existente entre las variables mediante análisis de correlación, con el propósito de identificar posibles dependencias lineales y detectar variables altamente relacionadas entre sí.

Posteriormente se desarrolló un modelo exploratorio de Regresión Logística, cuyo objetivo no fue seleccionar el modelo predictivo definitivo, sino analizar el comportamiento inicial de las variables explicativas mediante la evaluación de sus coeficientes. Esta etapa permitió identificar variables con bajo aporte predictivo y complementar el proceso de selección antes del entrenamiento de los modelos finales.

Figura 6 *Proceso de selección y depuración de variables para el desarrollo del Modelo Predictivo de Mora*



Nota. Elaboración Propia

Como complemento al análisis anterior, se evaluó la presencia de multicolinealidad mediante el cálculo del Factor de Inflación de la Varianza (Variance Inflation Factor, VIF). Los resultados del VIF fueron utilizados como un criterio de diagnóstico para identificar posibles redundancias entre las variables explicativas. La decisión de conservar o eliminar una variable no se basó exclusivamente en el valor del VIF, sino en una evaluación conjunta que consideró su importancia obtenida mediante el modelo exploratorio, su relación con la variable objetivo y su coherencia con el proceso de cobranza. Este procedimiento permitió construir un conjunto de variables con alta capacidad predictiva, minimizando la redundancia de información sin afectar el desempeño de los modelos de aprendizaje automático.

Finalmente, los resultados obtenidos fueron sometidos a un proceso de auditoría técnica y validación desde la perspectiva del negocio, verificando que las variables seleccionadas mantuvieran coherencia con el proceso de cobranza y aportaran información relevante para la estimación del riesgo de mora. Como resultado de esta metodología se obtuvo el conjunto definitivo de variables utilizado durante el entrenamiento del Modelo Predictivo de Mora.

Metodología de entrenamiento

Una vez definido el conjunto final de variables, se procedió al entrenamiento de los modelos de clasificación supervisada considerados en la investigación. Con el propósito de garantizar una evaluación homogénea y facilitar la comparación posterior de los algoritmos, se desarrolló una función genérica de evaluación que automatizó el proceso de entrenamiento y cálculo de las principales métricas de desempeño.

Figura 7 Metodología de entrenamiento y evaluación de los modelos de clasificación del Modelo Predictivo de Mora



Nota. Elaboración Propia

El conjunto de datos fue dividido en datos de entrenamiento y prueba mediante la técnica Train-Test Split, utilizando el 80% de las observaciones para el entrenamiento y el 20% restante para la evaluación del modelo. Previamente, el Dataset fue consolidado a partir de las seis muestras temporales construidas mediante la estrategia de ventanas móviles. La partición se realizó empleando un valor de `random_state = 42` para garantizar la reproducibilidad de los resultados y la opción `stratify`, preservando la distribución original de las clases del TargetMora en ambos subconjuntos.

La construcción de las muestras mediante ventanas temporales móviles garantizó que las variables explicativas fueran generadas exclusivamente con información disponible antes del período utilizado para la definición de las variables objetivo, evitando de esta manera la fuga de información temporal (*data leakage*). La partición aleatoria del Dataset consolidado fue utilizada específicamente para comparar los algoritmos bajo condiciones experimentales homogéneas, por lo que las métricas obtenidas en esta etapa corresponden a una validación interna. De manera complementaria, la evaluación comparativa del esquema de gestión actual frente a la propuesta se realizó utilizando información correspondiente a una ventana temporal posterior a la empleada durante el desarrollo de los modelos, proporcionando una evaluación adicional sobre información no utilizada en su construcción.

Posteriormente se entrenaron de manera independiente tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost. Todos los modelos fueron evaluados bajo las mismas condiciones experimentales, utilizando exactamente el mismo conjunto de entrenamiento y prueba, con el propósito de garantizar una comparación objetiva de su desempeño.

Como parte del proceso de evaluación se calcularon las métricas de Accuracy, Precision, Recall, F1-Score y ROC-AUC, además de la correspondiente matriz de confusión para cada algoritmo. Estas métricas permitieron caracterizar el comportamiento predictivo de cada modelo y constituyeron la base para el análisis comparativo presentado en el capítulo de resultados.

Adicionalmente, una vez entrenados los modelos, se obtuvo la importancia de las variables mediante los mecanismos de interpretación propios de cada algoritmo, con el propósito de analizar el aporte relativo de las características utilizadas durante el proceso de aprendizaje. Toda la información generada durante esta etapa fue documentada mediante reportes de

métricas, matrices de confusión e indicadores de importancia de variables, garantizando la trazabilidad y reproducibilidad del proceso metodológico desarrollado.

Implementación del modelo

Una vez concluido el proceso de entrenamiento y evaluación de los modelos candidatos, se procedió a la implementación del Modelo Predictivo de Mora con el propósito de generar predicciones sobre la totalidad de la cartera de clientes. Esta etapa describe el procedimiento aplicado una vez identificado el algoritmo con mejor desempeño durante la etapa de evaluación, permitiendo trasladar el modelo desde un entorno de desarrollo hacia un esquema orientado a su utilización en producción.

Figura 8 *Proceso de implementación del Modelo Predictivo de Mora*



Nota. Elaboración Propia

Como primer paso, el algoritmo identificado con el mejor desempeño durante la etapa de evaluación fue reentrenado utilizando la totalidad de los registros que disponían de un valor conocido para el TargetMora, con el objetivo de maximizar el aprovechamiento de la

información disponible antes de generar las predicciones finales. Posteriormente, el modelo resultante fue almacenado para su reutilización durante los procesos de inferencia y despliegue.

A continuación, se construyó el conjunto de datos de predicción a partir del Dataset completo, aplicando las mismas transformaciones realizadas durante la etapa de entrenamiento y conservando únicamente las variables explicativas utilizadas por el modelo definitivo. Con el fin de garantizar la consistencia estructural entre ambos conjuntos de datos, se verificó la correspondencia exacta de las variables mediante un proceso de validación previo a la generación de las predicciones.

Una vez validada la estructura del conjunto de predicción, el modelo estimó la probabilidad de mora para cada cuenta de la cartera. Posteriormente, dichas probabilidades fueron transformadas en un ScoreMora normalizado en una escala de 0 a 1000 puntos, facilitando su interpretación e integración con los procesos operativos de cobranza.

Como complemento al score predictivo, cada cuenta fue clasificada en un nivel de riesgo (Bajo, Medio, Alto y Crítico) de acuerdo con umbrales previamente definidos. Esta clasificación permitió convertir la probabilidad estimada por el modelo en una variable de fácil interpretación para los procesos de gestión y priorización de la cartera.

Finalmente, se consolidó el conjunto de resultados correspondiente al Modelo Predictivo de Mora, integrando el identificador de la cuenta, la probabilidad estimada, el score predictivo y el nivel de riesgo asignado. Este conjunto de información constituye la salida operativa del modelo y representa el insumo utilizado posteriormente para la construcción del Modelo de Priorización de Gestión desarrollado en esta investigación.

Conclusión del desarrollo

El desarrollo del Modelo Predictivo de Mora se ejecutó siguiendo una metodología estructurada que comprendió la construcción y validación de la variable objetivo, la

preparación del conjunto de datos, la selección y depuración de variables, el entrenamiento de diferentes algoritmos de clasificación y la implementación del modelo para la generación de predicciones sobre la totalidad de la cartera.

Cada una de estas etapas fue documentada mediante procedimientos reproducibles y evidencia técnica que garantizan la trazabilidad del proceso desarrollado. Asimismo, la implementación del modelo permitió obtener una salida estandarizada compuesta por la probabilidad de mora, un score predictivo normalizado y la clasificación del nivel de riesgo para cada cuenta, información que constituye la base para la integración posterior con los demás modelos predictivos desarrollados en la investigación.

De esta manera, el Modelo Predictivo de Mora representa el primer componente de la solución integral propuesta para la optimización de la gestión de cobranza, aportando una estimación objetiva del riesgo de incumplimiento y proporcionando información de alto valor para los procesos posteriores de priorización y definición de estrategias de gestión.

Desarrollo del Modelo Predictivo de Pago

El Modelo Predictivo de Pago fue desarrollado con el propósito de estimar el porcentaje esperado de recuperación de cada cuenta durante el horizonte de predicción definido en la investigación. A diferencia del Modelo Predictivo de Mora, cuya finalidad consiste en clasificar la probabilidad de incumplimiento, este modelo aborda un problema de regresión supervisada orientado a estimar una variable continua que representa el nivel esperado de pago de cada cliente.

El desarrollo del Modelo Predictivo de Pago siguió la misma metodología presentada en la Figura 3, adaptando únicamente la definición de la variable objetivo y las métricas de evaluación propias de un problema de regresión, comprendiendo las etapas de construcción de la variable objetivo, preparación del conjunto de datos, selección y depuración de

variables, entrenamiento de diferentes algoritmos de regresión, evaluación de su desempeño e implementación del modelo definitivo para la generación de predicciones.

Objetivo del Modelo Predictivo de Pago

El objetivo del Modelo Predictivo de Pago fue estimar el porcentaje esperado de recuperación de cada cuenta dentro del horizonte de análisis definido para la investigación. Esta estimación permitió cuantificar la probabilidad de recuperación económica de la cartera, proporcionando un indicador continuo que complementa la información generada por el Modelo Predictivo de Mora.

La construcción de este modelo respondió a la necesidad de no solo identificar qué clientes presentan mayor riesgo de incumplimiento, sino también estimar el nivel de pago esperado en aquellos casos donde exista posibilidad de recuperación. Esta información constituye un insumo relevante para la definición posterior de prioridades y estrategias de gestión de cobranza, al incorporar una estimación objetiva del potencial de recuperación de cada cuenta.

Al tratarse de un problema de regresión supervisada, el modelo fue diseñado para predecir una variable continua comprendida entre 0 y 1, correspondiente al porcentaje esperado de pago de cada cliente durante el período de evaluación. Posteriormente, este valor fue transformado en un score estandarizado y en una clasificación operacional que facilita su interpretación dentro de los procesos de gestión desarrollados por la institución.

Construcción y validación de TargetPago

La construcción del TargetPago constituyó la primera etapa para el desarrollo del Modelo Predictivo de Pago, definiendo la variable objetivo sobre la cual fueron entrenados los algoritmos de regresión supervisada. A diferencia del TargetMora, cuya naturaleza corresponde a un problema de clasificación binaria, el TargetPago fue definido como una

variable continua que representa el porcentaje esperado de recuperación de cada cuenta durante el horizonte de predicción establecido para la investigación.

El TargetPago fue calculado como la relación entre el valor efectivamente pagado por el cliente y el valor total de la obligación registrada durante el período de evaluación, obteniéndose valores comprendidos entre 0 y 1. Esta definición permitió cuantificar el nivel de recuperación esperado de cada cuenta, proporcionando una medida continua del comportamiento de pago que posteriormente fue utilizada para el entrenamiento de los modelos predictivos.

Una vez construido el target, se realizó un proceso de validación orientado a verificar su consistencia y distribución dentro del Dataset. Para ello se analizaron las principales estadísticas descriptivas de la variable objetivo, así como su distribución de frecuencias, con el propósito de confirmar que el conjunto de datos representara adecuadamente el comportamiento de pago observado durante el horizonte de predicción.

La tabla presenta un resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Pago, mientras que la figura muestra la distribución del TargetPago, permitiendo verificar la naturaleza continua de la variable objetivo y su idoneidad para la aplicación de técnicas de regresión supervisada.

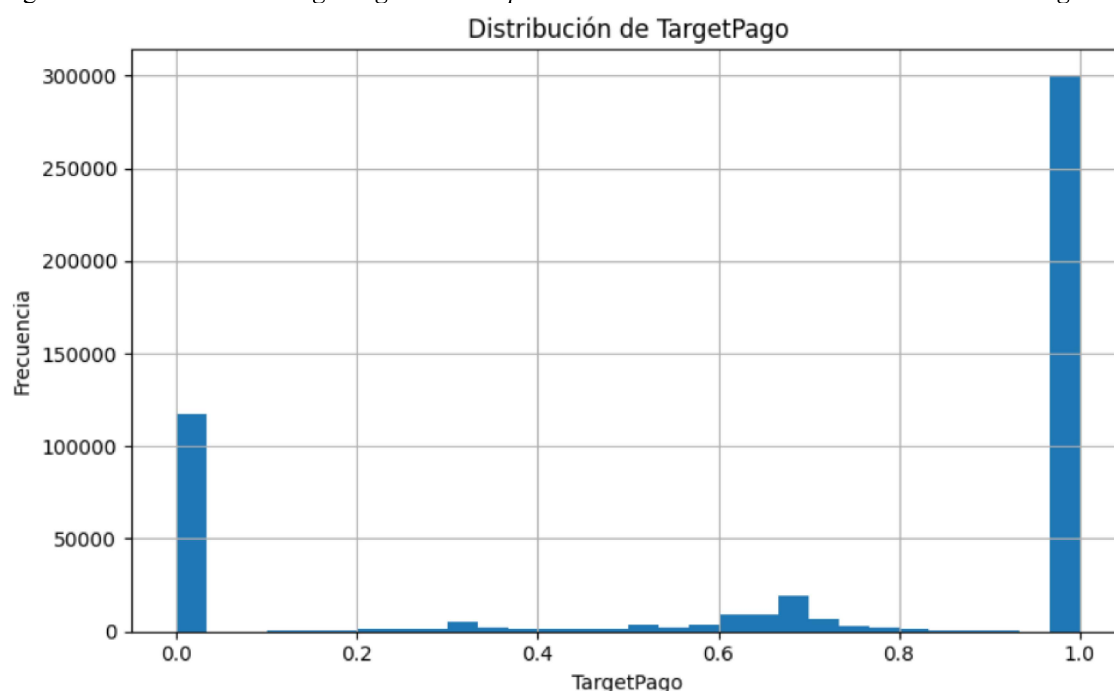
Tabla 2 *Resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Pago*

Indicador	Valor
Total registros dataset completo	573,328
Total columnas dataset completo	47
Registros con TargetPago disponible	496,879
Registros sin TargetPago	76,449
Valor mínimo TargetPago	0
Valor máximo TargetPago	1
Promedio TargetPago	0.70
Mediana TargetPago	1
Desviación estándar TargetPago	0.42

Nota. Elaboración Propia

Como se observa en la tabla, el Dataset completo estuvo conformado por 573,328 registros y 47 variables. De este conjunto, 496,879 registros disponían de información suficiente para calcular el TargetPago, mientras que 76,449 registros no contaban con dicho target y fueron reservados para la etapa de implementación y generación de predicciones. Los estadísticos descriptivos evidencian que el TargetPago presenta valores entre 0 y 1, con un promedio de 0.70, una mediana igual a 1.00 y una desviación estándar de 0.42, lo que confirma su naturaleza continua y justifica la utilización de modelos de regresión supervisada para estimar el porcentaje esperado de recuperación de cada cuenta. Por su parte, la figura evidencia la distribución de la variable objetivo, confirmando su naturaleza continua y justificando la utilización de modelos de regresión supervisada para estimar el porcentaje esperado de recuperación de cada cuenta.

Figura 9 Distribución del TargetPago utilizada para el entrenamiento del Modelo Predictivo de Pago



Nota. Elaboración Propia

Preparación del conjunto de datos

Una vez validado el TargetPago, se procedió a preparar el conjunto de datos que sería utilizado durante el entrenamiento del Modelo Predictivo de Pago. Esta etapa tuvo como

finalidad construir una matriz de variables explicativas consistente con la variable objetivo, garantizando la calidad de la información y la aplicación uniforme de los procedimientos de preparación utilizados durante la investigación.

Como primer paso, se seleccionaron únicamente los registros que disponían de un valor conocido para el TargetPago, conformando el conjunto de observaciones destinado al entrenamiento y evaluación de los modelos de regresión. Los registros que no disponían de información del target fueron reservados para la etapa de implementación y generación de predicciones.

Posteriormente se definió la matriz de variables explicativas (XPago) eliminando aquellas variables que no debían intervenir en el entrenamiento del modelo. Entre ellas se incluyeron el identificador único de la cuenta (IdCuenta), la variable Muestra y la variable FechaObservacion, así como las variables objetivo correspondientes a los demás modelos predictivos desarrollados en la investigación. La variable IdCuenta se utilizó únicamente como identificador de cada registro, mientras que Muestra y FechaObservacion permitieron controlar y mantener la trazabilidad temporal durante la construcción del Dataset. Ninguna de estas variables fue utilizada como predictor durante el entrenamiento de los modelos, evitando la incorporación de información que pudiera generar dependencia entre modelos o introducir sesgos durante el proceso de aprendizaje.

Como parte de la preparación del conjunto de datos, se verificó la existencia de valores faltantes en las variables explicativas, aplicando procesos de imputación en aquellas variables donde fue necesario preservar la integridad del Dataset. Asimismo, las variables de tipo booleano fueron transformadas a formato numérico para garantizar su compatibilidad con los algoritmos de aprendizaje supervisado empleados durante el entrenamiento.

Finalmente, la variable TargetPago fue definida como variable dependiente (yPago), obteniéndose el conjunto de datos definitivo que sirvió como base para las etapas posteriores de selección de variables, entrenamiento y evaluación de los modelos de regresión.

Figura 10 Metodología para el desarrollo del Modelo Predictivo de Pago



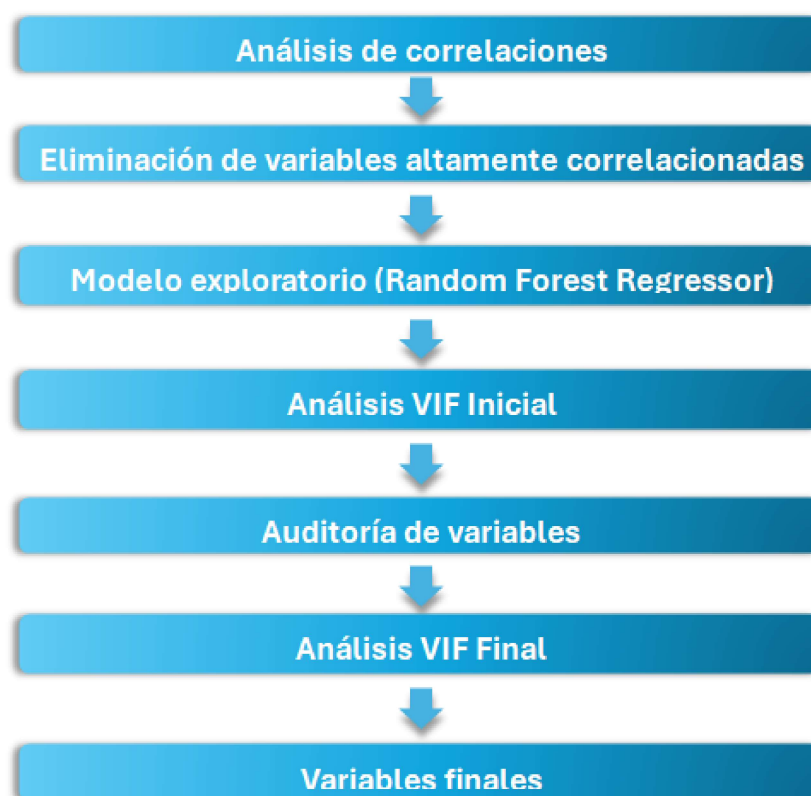
Nota. Elaboración Propia

La figura resume la metodología aplicada durante el desarrollo del Modelo Predictivo de Pago, la cual mantiene la misma estructura metodológica empleada en los demás modelos predictivos desarrollados en esta investigación, adaptándose únicamente a las particularidades propias de un problema de regresión.

Selección y depuración de variables

La selección y depuración de variables constituyó una etapa fundamental para el desarrollo del Modelo Predictivo de Pago, permitiendo identificar el conjunto de variables con mayor capacidad para explicar el porcentaje esperado de recuperación de cada cuenta y reducir la presencia de información redundante dentro del modelo.

Figura 11 *Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Pago*



Nota. Elaboración Propia

El proceso inició con el análisis de correlación entre las variables explicativas y el TargetPago, con el propósito de identificar relaciones relevantes y detectar variables altamente correlacionadas que pudieran afectar la estabilidad del modelo. Como resultado de este análisis, se eliminaron aquellas variables cuya información resultaba redundante, favoreciendo la construcción de un conjunto de datos más consistente.

Posteriormente, se desarrolló un modelo exploratorio utilizando Random Forest Regressor, el cual permitió obtener una primera estimación de la importancia relativa de las variables explicativas. La utilización de este algoritmo en la etapa exploratoria respondió a su capacidad para capturar relaciones no lineales entre las variables independientes y la variable objetivo, proporcionando un criterio robusto para priorizar aquellas características con mayor aporte predictivo.

Como complemento al análisis anterior, se evaluó la presencia de multicolinealidad mediante el cálculo del Factor de Inflación de la Varianza (Variance Inflation Factor, VIF). Inicialmente se obtuvo un diagnóstico del conjunto completo de variables, identificando posibles relaciones de dependencia entre ellas. Posteriormente, la información obtenida fue integrada en un proceso de auditoría técnica que combinó los resultados del análisis de correlaciones, la importancia de variables obtenida mediante el modelo exploratorio, los indicadores de multicolinealidad y el conocimiento del negocio. En consecuencia, la decisión de conservar o eliminar una variable no se basó exclusivamente en el valor del VIF, sino en una evaluación conjunta de todos estos criterios, permitiendo construir un conjunto de variables con alta capacidad predictiva y reducida redundancia de información. Finalmente, una vez realizada la depuración del conjunto de datos, se recalculó el VIF para verificar que las variables conservadas presentaran niveles adecuados de independencia, obteniéndose el conjunto definitivo de variables utilizado durante el entrenamiento de los modelos de regresión.

Metodología de entrenamiento

Una vez definido el conjunto final de variables, se procedió al entrenamiento de los modelos de regresión supervisada considerados en la investigación. Con el propósito de garantizar una evaluación homogénea y facilitar la comparación posterior de los algoritmos, se desarrolló una función genérica de evaluación que automatizó el proceso de entrenamiento y cálculo de las principales métricas de desempeño.

Figura 12 Metodología de entrenamiento y evaluación de los modelos de regresión del Modelo Predictivo de Pago



Nota. Elaboración Propia

El conjunto de datos fue dividido en datos de entrenamiento y prueba mediante la técnica Train-Test Split, utilizando el 80% de las observaciones para el entrenamiento y el 20% restante para la evaluación del modelo. Previamente, el Dataset fue consolidado a partir de las seis muestras temporales construidas mediante la estrategia de ventanas móviles, permitiendo que el proceso de partición aleatoria incluyera observaciones construidas a partir de diferentes ventanas temporales. La partición se realizó empleando un valor de `random_state = 42` para garantizar la reproducibilidad de los resultados obtenidos durante el proceso de experimentación.

Al igual que en el Modelo Predictivo de Mora, esta partición se realizó sobre el Dataset consolidado de las seis muestras temporales, por lo que corresponde a una validación aleatoria interna y no a una validación temporal estrictamente fuera de muestra.

Posteriormente se entrenaron de manera independiente tres algoritmos de regresión supervisada: Regresión Lineal, Random Forest Regressor y XGBoost Regressor. Todos los modelos fueron evaluados bajo las mismas condiciones experimentales, utilizando

exactamente el mismo conjunto de entrenamiento y prueba, con el propósito de realizar una comparación objetiva de su capacidad predictiva.

Como parte del proceso de evaluación se calcularon las métricas de Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y Coeficiente de Determinación (R^2).

Estas métricas permitieron cuantificar el nivel de precisión alcanzado por cada algoritmo y constituyeron la base para el análisis comparativo presentado posteriormente en el capítulo de resultados.

Adicionalmente, para cada uno de los modelos entrenados se generaron gráficos de comparación entre los valores reales y los valores predichos, así como el análisis de importancia de variables utilizando los mecanismos de interpretación correspondientes a cada algoritmo. Esta información permitió complementar la evaluación cuantitativa de los modelos y facilitar la interpretación de los factores con mayor influencia sobre la estimación del porcentaje esperado de pago.

Finalmente, los modelos entrenados fueron almacenados junto con sus respectivas métricas y metadatos, garantizando la trazabilidad del proceso de entrenamiento y permitiendo su posterior utilización durante la etapa de implementación del modelo definitivo.

Implementación del modelo

Una vez concluido el proceso de entrenamiento y evaluación de los modelos candidatos, se procedió a la implementación del Modelo Predictivo de Pago con el propósito de generar predicciones sobre la totalidad de la cartera de clientes. Esta etapa describe el procedimiento aplicado una vez identificado el algoritmo con mejor desempeño durante la etapa de evaluación, permitiendo trasladar el modelo desde un entorno de desarrollo hacia un esquema orientado a su utilización en producción.

Figura 13 *Proceso de implementación del Modelo Predictivo de Pago*



Nota. Elaboración Propia

Como primer paso, el algoritmo identificado con el mejor desempeño durante la etapa de evaluación fue reentrenado utilizando la totalidad de los registros que disponían de un valor conocido para el TargetPago, con el objetivo de maximizar el aprovechamiento de la información disponible antes de generar las predicciones finales. Posteriormente, el modelo resultante fue almacenado para su reutilización durante los procesos de inferencia y despliegue.

A continuación, se construyó el conjunto de datos de predicción a partir del Dataset completo, aplicando las mismas transformaciones realizadas durante la etapa de entrenamiento y conservando únicamente las variables explicativas utilizadas por el modelo definitivo. Con el fin de garantizar la consistencia estructural entre ambos conjuntos de datos, se verificó la correspondencia exacta de las variables mediante un proceso de validación previo a la generación de las predicciones.

Una vez validada la estructura del conjunto de predicción, el modelo estimó el porcentaje esperado de pago para cada cuenta de la cartera. Debido a que se trata de un problema de

regresión, las predicciones obtenidas fueron acotadas al intervalo comprendido entre 0 y 1, garantizando que los valores estimados representaran porcentajes válidos de recuperación. Posteriormente, el porcentaje esperado de pago fue transformado en un ScorePago normalizado en una escala de 0 a 1000 puntos, facilitando su interpretación y posterior integración con los demás componentes del modelo de priorización desarrollado en la investigación.

Como complemento al score predictivo, cada cuenta fue clasificada en cinco categorías de pago esperado (Muy alto, Alto, Medio, Bajo y Muy bajo), permitiendo traducir la predicción continua del modelo en una clasificación de fácil interpretación para los procesos de análisis y gestión.

Finalmente, se consolidó el conjunto de resultados correspondiente al Modelo Predictivo de Pago, integrando el identificador de la cuenta, el porcentaje esperado de pago, el score predictivo y la clasificación del nivel de pago esperado. Este conjunto de información constituye la salida operativa del modelo y representa uno de los insumos utilizados posteriormente para la construcción del Modelo de Priorización de Gestión de Cartera.

Conclusión del desarrollo

El desarrollo del Modelo Predictivo de Pago se ejecutó siguiendo una metodología estructurada que comprendió la construcción y validación de la variable objetivo, la preparación del conjunto de datos, la selección y depuración de variables, el entrenamiento de diferentes algoritmos de regresión y la implementación del modelo para la generación de predicciones sobre la totalidad de la cartera.

Cada una de estas etapas fue desarrollada mediante procedimientos reproducibles y documentada a través de evidencia técnica, garantizando la trazabilidad del proceso y la consistencia metodológica de la investigación. La implementación del modelo permitió

obtener una salida estandarizada compuesta por el porcentaje esperado de pago, un score predictivo normalizado y la clasificación del nivel de pago esperado para cada cuenta. La información generada por este modelo complementa el análisis realizado por el Modelo Predictivo de Mora, incorporando una estimación cuantitativa del potencial de recuperación de cada cliente. De esta manera, el Modelo Predictivo de Pago constituye el segundo componente de la solución integral propuesta para la optimización de la gestión de cobranza, proporcionando información relevante para la priorización de cuentas y la definición de estrategias de gestión que serán integradas posteriormente en el Modelo de Priorización de Gestión de Cobranza.

Desarrollo del Modelo Predictivo de Regularización

El Modelo Predictivo de Regularización fue desarrollado con el propósito de estimar la probabilidad de que una cuenta logre regularizar su situación durante el horizonte de predicción definido para la investigación, ya sea mediante un pago efectivo o a través de un proceso de refinanciamiento. Este modelo complementa la información generada por los modelos predictivos de Mora y Pago, incorporando una estimación de la probabilidad de recuperación futura de aquellas cuentas que presentan incumplimientos en sus obligaciones. El desarrollo del Modelo Predictivo de Regularización siguió la misma metodología presentada en la Figura 3, adaptando únicamente la definición de la variable objetivo correspondiente al proceso de regularización. La metodología comprendió las etapas de construcción de la variable objetivo, preparación del conjunto de datos, selección y depuración de variables, entrenamiento de diferentes algoritmos de clasificación, evaluación de su desempeño e implementación del modelo definitivo para la generación de predicciones.

Objetivo del Modelo Predictivo de Regularización

El objetivo del Modelo Predictivo de Regularización fue estimar la probabilidad de que un cliente regularice su obligación financiera durante el horizonte de predicción considerado en la investigación. Para este propósito, se definió como regularización tanto el pago total o parcial de la obligación como la formalización de un proceso de refinanciamiento, considerando que ambas situaciones representan mecanismos válidos para la recuperación de la cartera.

La construcción de este modelo respondió a la necesidad de complementar la información proporcionada por los modelos predictivos de Mora y Pago, incorporando una estimación probabilística de la capacidad de regularización de cada cuenta. Esta información constituye un insumo relevante para la definición posterior de prioridades y estrategias de gestión, permitiendo identificar aquellos clientes que presentan mayores posibilidades de normalizar su situación financiera.

Al tratarse de un problema de clasificación supervisada, el modelo fue diseñado para estimar la probabilidad de ocurrencia de un evento binario, correspondiente a la regularización o no de la obligación durante el período de evaluación. Posteriormente, dicha probabilidad fue transformada en un score estandarizado y en una clasificación operacional que facilita su interpretación dentro de los procesos de gestión de cobranza.

Construcción y validación del TargetRegularizacion

La construcción del TargetRegularizacion constituyó la primera etapa para el desarrollo del Modelo Predictivo de Regularización, definiendo la variable objetivo sobre la cual fueron entrenados los algoritmos de clasificación supervisada. Al igual que el Modelo Predictivo de Mora, este modelo aborda un problema de clasificación binaria, donde el objetivo consiste en estimar la probabilidad de que una cuenta regularice su situación durante el horizonte de predicción.

El TargetRegularizacion fue definido asignando el valor 1 a aquellas cuentas que, durante el período de evaluación, realizaron al menos un pago o formalizaron un proceso de refinanciamiento, mientras que las cuentas que no presentaron ninguno de estos eventos fueron clasificadas con el valor 0. Esta definición permitió representar el concepto de regularización desde una perspectiva operativa, considerando tanto la recuperación efectiva de recursos como la formalización de acuerdos que contribuyen a la normalización de la cartera.

Una vez construido el target, se realizó un proceso de validación orientado a verificar su consistencia y distribución dentro del Dataset. Para ello se analizaron las principales estadísticas descriptivas de la variable objetivo y la distribución de las clases, con el propósito de confirmar que el conjunto de datos representara adecuadamente el fenómeno de regularización observado durante el horizonte de predicción.

La tabla a continuación presenta un resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Regularización, mientras que la figura muestra la distribución del TargetRegularizacion, permitiendo verificar la proporción de cuentas que lograron regularizar su situación frente a aquellas que no registraron eventos de pago o refinanciamiento.

Tabla 3 *Resumen descriptivo del Dataset utilizado para el entrenamiento del Modelo Predictivo de Regularización*

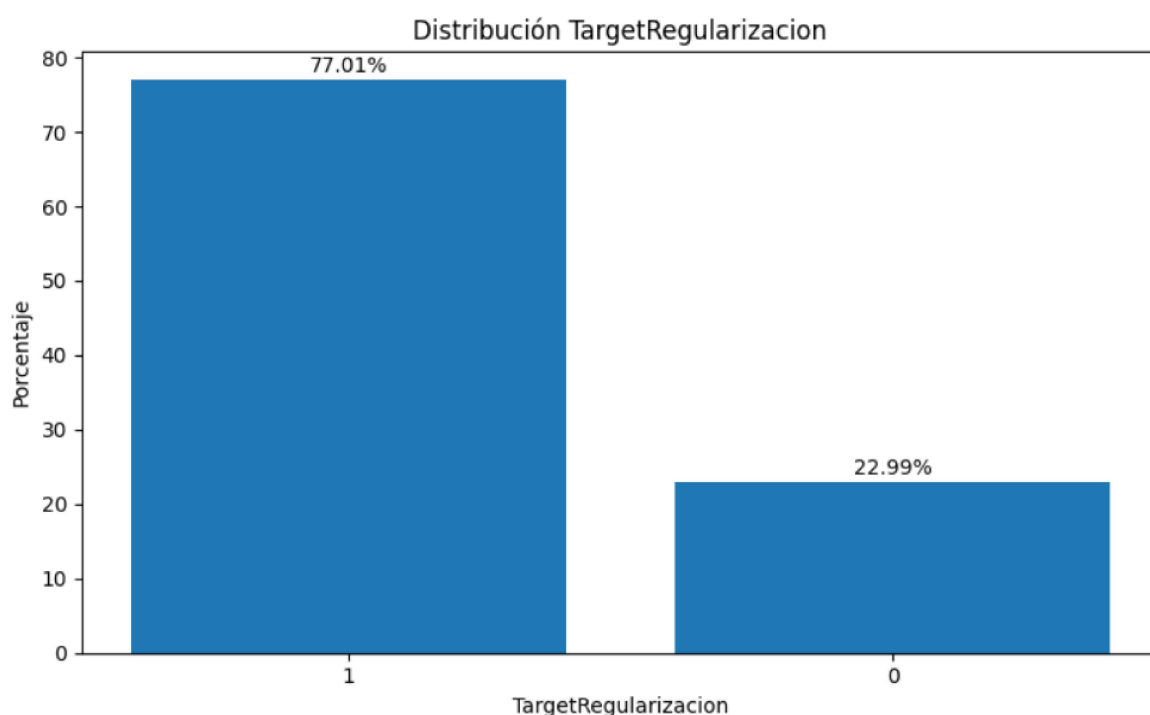
Indicador	Valor
Total registros dataset completo	573,328
Total columnas dataset completo	47
Registros con TargetRegularizacion disponible	496,879
Registros sin TargetRegularizacion	76,449
Proporción TargetRegularizacion = 0	22.99%
Proporción TargetRegularizacion = 1	77.01%

Nota. Elaboración Propia

Como se observa en la tabla, el Dataset utilizado para el entrenamiento del Modelo Predictivo de Regularización estuvo conformado por 496,879 registros con información

suficiente para calcular el TargetRegularizacion, mientras que 76,449 registros fueron reservados para la etapa de implementación y generación de predicciones. Por su parte, la figura evidencia una distribución donde el 77.01% de las cuentas lograron regularizar su situación durante el horizonte de evaluación, frente a un 22.99% que no registró eventos de pago o refinanciamiento. Esta distribución confirma la naturaleza binaria del problema y justifica la utilización de modelos de clasificación supervisada para estimar la probabilidad de regularización de cada cuenta.

Figura 14 *Distribución del TargetRegularizacion utilizado para el entrenamiento del Modelo Predictivo de Regularización*



Nota. Elaboración Propia

Preparación del conjunto de datos

Una vez validado el TargetRegularizacion, se procedió a preparar el conjunto de datos que sería utilizado durante el entrenamiento del Modelo Predictivo de Regularización. Esta etapa tuvo como finalidad construir una matriz de variables explicativas consistente con la variable objetivo, garantizando la calidad de la información y la aplicación uniforme de los procedimientos de preparación utilizados durante la investigación.

Figura 15 *Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Regularización*



Nota. Elaboración Propia

Como primer paso, se seleccionaron únicamente los registros que disponían de un valor conocido para el TargetRegularizacion, conformando el conjunto de observaciones destinado al entrenamiento y evaluación de los modelos de clasificación. Los registros que no disponían de información del target fueron reservados para la etapa de implementación y generación de predicciones.

Posteriormente, se definió la matriz de variables explicativas (XRegularizacion), excluyendo el identificador único de la cuenta (IdCuenta), la variable Muestra, la variable FechaObservacion y las variables objetivo correspondientes a los demás modelos predictivos desarrollados en la investigación. La variable Muestra fue utilizada únicamente para identificar el origen temporal de cada observación dentro de las seis ventanas móviles, mientras que FechaObservacion permitió mantener la trazabilidad temporal del registro. Ninguna de estas variables participó en el entrenamiento de los modelos, evitando la incorporación de información que pudiera introducir sesgos o generar dependencia entre las observaciones.

Como parte de la preparación del conjunto de datos, se verificó la existencia de valores faltantes en las variables explicativas, aplicando procesos de imputación en aquellas

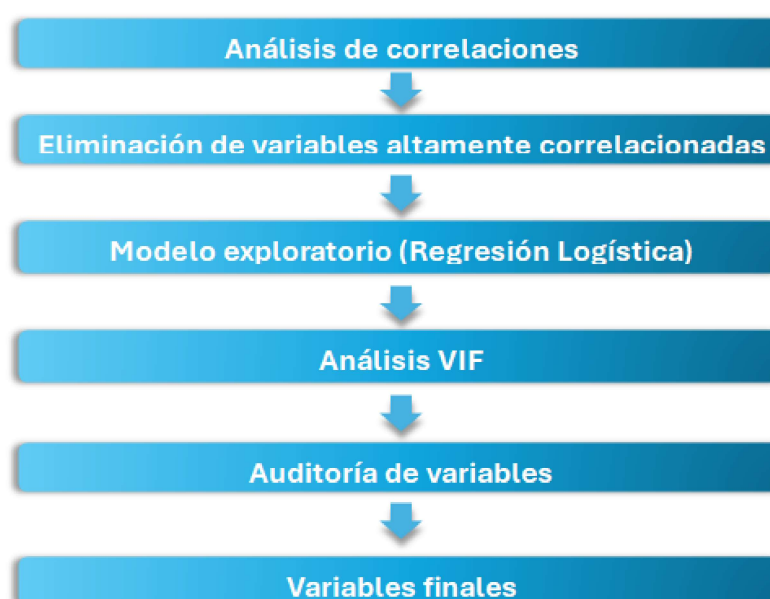
variables donde fue necesario preservar la integridad del Dataset. Asimismo, las variables de tipo booleano fueron transformadas a formato numérico para garantizar su compatibilidad con los algoritmos de aprendizaje supervisado empleados durante el entrenamiento.

Finalmente, la variable TargetRegularizacion fue definida como variable dependiente (yRegularizacion), obteniéndose el conjunto de datos definitivo que sirvió como base para las etapas posteriores de selección de variables, entrenamiento y evaluación de los modelos de clasificación.

Selección y depuración de variables

La selección y depuración de variables constituyó una etapa fundamental para el desarrollo del Modelo Predictivo de Regularización, permitiendo identificar el conjunto de variables con mayor capacidad para explicar la probabilidad de que una cuenta regularice su situación financiera y reducir la presencia de información redundante dentro del modelo.

Figura 16 *Proceso de selección y depuración de variables para el desarrollo del Modelo Predictivo de Regularización*



Nota. Elaboración Propia

El proceso inició con el análisis de correlación entre las variables explicativas y el TargetRegularizacion, con el propósito de identificar relaciones relevantes y detectar variables altamente correlacionadas que pudieran afectar la estabilidad del modelo. Como resultado de este análisis, se eliminaron aquellas variables cuya información resultaba redundante, favoreciendo la construcción de un conjunto de datos más consistente.

Posteriormente, se desarrolló un modelo exploratorio utilizando Regresión Logística, cuyo objetivo no fue seleccionar el modelo predictivo definitivo, sino analizar el comportamiento inicial de las variables explicativas mediante la evaluación de sus coeficientes. Esta etapa permitió identificar variables con bajo aporte predictivo y complementar el proceso de selección antes del entrenamiento de los modelos finales.

Como complemento al análisis anterior, se evaluó la presencia de multicolinealidad mediante el cálculo del Factor de Inflación de la Varianza (VIF). Inicialmente se obtuvo un diagnóstico del conjunto completo de variables, identificando posibles relaciones de dependencia entre ellas. Posteriormente, la información obtenida fue integrada en un proceso de auditoría técnica que combinó los resultados del análisis de correlaciones, la importancia de las variables y los indicadores de multicolinealidad, permitiendo fundamentar la eliminación de variables redundantes.

Finalmente, una vez realizada la depuración del conjunto de datos, se recalculó el VIF para verificar que las variables conservadas presentaran niveles adecuados de independencia, obteniéndose el conjunto definitivo de variables utilizado durante el entrenamiento de los modelos de clasificación.

Metodología de entrenamiento

Una vez definido el conjunto final de variables, se procedió al entrenamiento de los modelos de clasificación supervisada considerados en la investigación. Con el propósito de garantizar una evaluación homogénea y facilitar la comparación posterior de los algoritmos,

se utilizó una función genérica de evaluación que automatizó el proceso de entrenamiento y cálculo de las principales métricas de desempeño.

Figura 17 Metodología de entrenamiento y evaluación de los modelos de clasificación del Modelo Predictivo de Regularización



Nota. Elaboración Propia

El conjunto de datos fue dividido en datos de entrenamiento y prueba mediante la técnica Train-Test Split, utilizando el 80% de las observaciones para el entrenamiento y el 20% restante para la evaluación del modelo. Previamente, el Dataset fue consolidado a partir de las seis muestras temporales construidas mediante la estrategia de ventanas móviles, permitiendo que el proceso de partición aleatoria incluyera observaciones construidas a partir de diferentes ventanas temporales. La partición se realizó empleando un valor de `random_state = 42` para garantizar la reproducibilidad de los resultados y la opción `stratify`, preservando la distribución original de las clases del TargetRegularizacion en ambos subconjuntos.

Posteriormente se entrenaron de manera independiente tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost. Todos los modelos fueron

evaluados bajo las mismas condiciones experimentales, utilizando exactamente el mismo conjunto de entrenamiento y prueba, con el propósito de garantizar una comparación objetiva de su desempeño.

Como parte del proceso de evaluación se calcularon las métricas de Accuracy, Precision, Recall, F1-Score y ROC-AUC, además de la correspondiente matriz de confusión para cada algoritmo. Estas métricas permitieron caracterizar el comportamiento predictivo de cada modelo y constituyeron la base para el análisis comparativo presentado en el capítulo de resultados.

Las métricas obtenidas corresponden a una validación interna realizada sobre el Dataset consolidado mediante la partición Train-Test Split. La construcción previa de las muestras mediante ventanas temporales móviles garantizó la separación entre la información utilizada para generar las variables explicativas y el período posterior empleado para construir el TargetRegularizacion, evitando la fuga de información temporal (*data leakage*).

Adicionalmente, una vez entrenados los modelos, se obtuvo la importancia de las variables mediante los mecanismos de interpretación propios de cada algoritmo, con el propósito de analizar el aporte relativo de las características utilizadas durante el proceso de aprendizaje. Toda la información generada durante esta etapa fue documentada mediante reportes de métricas, matrices de confusión e indicadores de importancia de variables, garantizando la trazabilidad y reproducibilidad del proceso metodológico desarrollado.

Implementación del modelo

Una vez concluido el proceso de entrenamiento y evaluación de los modelos candidatos, se procedió a la implementación del Modelo Predictivo de Regularización con el propósito de generar predicciones sobre la totalidad de la cartera de clientes. Esta etapa describe el procedimiento aplicado una vez identificado el algoritmo con mejor desempeño durante la

etapa de evaluación, permitiendo trasladar el modelo desde un entorno de desarrollo hacia un esquema orientado a su utilización en producción.

Figura 18 *Proceso de implementación del Modelo Predictivo de Regularización*



Nota. Elaboración Propia

Como primer paso, el modelo seleccionado durante la etapa de evaluación será reentrenado utilizando la totalidad de los registros que dispongan de un valor conocido para el `TargetRegularizacion`, con el objetivo de maximizar el aprovechamiento de la información disponible antes de generar las predicciones finales. Posteriormente, el modelo definitivo será almacenado para su reutilización durante los procesos de inferencia y despliegue. A continuación, se construirá el conjunto de datos de predicción a partir del Dataset completo, aplicando las mismas transformaciones realizadas durante la etapa de entrenamiento y conservando únicamente las variables explicativas utilizadas por el modelo definitivo. Con el fin de garantizar la consistencia estructural entre ambos conjuntos de datos, se verificará la correspondencia exacta de las variables mediante un proceso de validación previo a la generación de las predicciones.

Una vez validada la estructura del conjunto de predicción, el modelo estimará la probabilidad de regularización para cada cuenta de la cartera. Posteriormente, dichas probabilidades serán transformadas en un ScoreRegularizacion, normalizado en una escala de 0 a 1000 puntos, con el propósito de facilitar su interpretación e integración con los procesos operativos de cobranza.

Como complemento al score predictivo, cada cuenta será clasificada en niveles de probabilidad de regularización de acuerdo con umbrales previamente definidos. Esta clasificación permitirá convertir la probabilidad estimada por el modelo en una variable de fácil interpretación para los procesos de gestión y priorización de la cartera.

Finalmente, se consolidará el conjunto de resultados correspondiente al Modelo Predictivo de Regularización, integrando el identificador de la cuenta, la probabilidad estimada, el score predictivo y el nivel de clasificación asignado. Este conjunto de información constituirá la salida operativa del modelo y representará uno de los insumos utilizados posteriormente para la construcción del Modelo de Priorización de Gestión desarrollado en esta investigación.

Conclusión del desarrollo

El desarrollo del Modelo Predictivo de Regularización se ejecutó siguiendo una metodología estructurada que comprendió la construcción y validación de la variable objetivo, la preparación del conjunto de datos, la selección y depuración de variables, el entrenamiento de diferentes algoritmos de clasificación y la implementación del modelo para la generación de predicciones sobre la totalidad de la cartera.

Cada una de estas etapas fue desarrollada mediante procedimientos reproducibles y documentada mediante evidencia técnica, garantizando la trazabilidad del proceso y la consistencia metodológica de la investigación. La implementación del modelo permitió obtener una salida estandarizada compuesta por la probabilidad de regularización, un score

predictivo normalizado y la clasificación de la oportunidad de regularización para cada cuenta.

La información generada por este modelo complementa los resultados obtenidos por los modelos predictivos de Mora y Pago, incorporando una estimación de la probabilidad de que un cliente normalice su situación financiera mediante un pago o un refinanciamiento.

De esta manera, el Modelo Predictivo de Regularización constituye el tercer componente de la solución integral propuesta para la optimización de la gestión de cobranza, proporcionando información relevante para la construcción del Modelo de Priorización de Gestión de Cartera y la definición posterior de estrategias diferenciadas de gestión.

Modelo de Priorización de Gestión de Cobranza

El Modelo de Priorización de Gestión de Cobranza fue desarrollado con el propósito de integrar las salidas de los modelos predictivos de Mora, Pago y Regularización en un único esquema de decisión orientado a apoyar la gestión operativa de cobranza. A diferencia de los modelos anteriores, cuyo objetivo fue estimar eventos o comportamientos específicos de cada cuenta, este componente busca consolidar dichas predicciones en un indicador compuesto que permita ordenar la cartera según su prioridad de gestión.

Para su construcción se integraron las probabilidades, scores y clasificaciones generadas por los tres modelos predictivos previamente desarrollados. Adicionalmente, se incorporó la variable *ExposicionImpaga*, con el fin de considerar el impacto económico asociado a cada cuenta dentro del proceso de priorización. Esta integración permitió construir un conjunto de datos denominado *dfGestion*, el cual constituye la base para el cálculo del *ScorePrioridadGestion* y la segmentación por deciles.

El modelo fue diseñado bajo los objetivos de maximizar la recuperación de cartera, reducir la morosidad y optimizar los costos operativos de cobranza. En este sentido, la priorización no se basa únicamente en el riesgo de incumplimiento ni en la exposición económica, sino

en la combinación de variables predictivas y financieras que permiten diferenciar las cuentas según su potencial de recuperación, probabilidad de regularización, riesgo de mora e impacto económico.

Finalmente, los resultados del modelo fueron consolidados en una estructura denominada resultadoGestionCartera, preparada para su almacenamiento en SQL Server y posterior integración con los sistemas operativos de la fintech.

Objetivo del Modelo de Priorización de Gestión de Cartera

El objetivo del Modelo de Priorización de Gestión de Cobranza fue construir un mecanismo de decisión que permitiera ordenar las cuentas de la cartera en función de su prioridad operativa de cobranza. Para ello, se integraron las predicciones generadas por los modelos de Mora, Pago y Regularización, junto con la exposición económica de cada cuenta, obteniendo un score compuesto orientado a apoyar la asignación eficiente de recursos de gestión.

Este modelo responde a la necesidad de superar esquemas homogéneos de cobranza, en los cuales todos los clientes reciben tratamientos similares independientemente de su riesgo, potencial de pago o nivel de exposición. A través del ScorePrioridadGestion, se busca proporcionar una visión integral de cada cuenta, permitiendo diferenciar entre gestiones recuperativas y preventivas, así como establecer niveles de prioridad mediante deciles de gestión.

La finalidad del modelo no es reemplazar las reglas operativas de cobranza, sino proporcionar un insumo analítico para su parametrización. De esta manera, los deciles y el tipo de gestión constituyen una base objetiva para que la Fintech pueda definir una estrategia de gestión y acciones diferenciadas de contacto, frecuencia e intensidad de gestión, alineadas con los objetivos de maximizar la recuperación, reducir la mora y optimizar los costos operativos.

Integración de los modelos predictivos

Figura 19 Integración de los modelos predictivos para la construcción del conjunto de datos de gestión



Nota. Elaboración Propia

La primera etapa del Modelo de Priorización de Gestión de Cobranza consistió en integrar la información generada por los tres modelos predictivos desarrollados durante la investigación. Para ello, se consolidaron las salidas de los Modelos Predictivos de Mora, Pago y Regularización utilizando como llave de integración las variables IdCuenta, Muestra y FechaObservacion. Esta estructura permitió asociar correctamente las predicciones correspondientes a cada observación generada mediante la estrategia de ventanas temporales móviles, preservando la trazabilidad de la información durante todo el proceso de priorización.

Posteriormente, a esta estructura se incorporó la variable ExposicionImpaga, correspondiente al valor económico pendiente de recuperación para cada cuenta. La inclusión de esta variable permitió complementar las predicciones de los modelos con una

medida del impacto financiero asociado a cada cliente, proporcionando una visión integral para el proceso de priorización.

Como parte del proceso de integración se verificó la consistencia de la información consolidada, comprobando la inexistencia de valores faltantes en las variables utilizadas para la construcción del modelo de priorización. Posteriormente, la exposición económica fue normalizada mediante una transformación lineal, obteniendo valores comprendidos entre 0 y 1. Esta normalización permitió integrar la exposición económica con las probabilidades generadas por los modelos predictivos, evitando que las diferencias de escala influyeran sobre el cálculo del indicador compuesto.

El resultado de esta etapa fue la construcción del conjunto de datos `dfGestion`, el cual reúne las variables predictivas, financieras y operativas necesarias para el cálculo del `ScorePrioridadGestion`, constituyendo la base del modelo integral de priorización desarrollado en esta investigación.

Construcción del Score de Priorización de Gestión

Una vez integrado el conjunto de datos `dfGestion`, se procedió a la construcción del `ScorePrioridadGestion`, indicador diseñado para sintetizar en un único valor la información generada por los tres modelos predictivos y el impacto económico asociado a cada cuenta. La construcción del score respondió al objetivo de establecer un criterio cuantitativo que permitiera ordenar la cartera según su prioridad de gestión, considerando simultáneamente el potencial de recuperación, la probabilidad de regularización, el riesgo de mora y la exposición económica pendiente. De esta manera, la priorización dejó de depender de un único indicador, incorporando una visión integral del comportamiento esperado de cada cliente.

Para ello se definió una combinación lineal ponderada de cuatro componentes: el `PagoEsperado`, la `ProbabilidadRegularizacion`, la `ProbabilidadMora` y la

ExposicionNormalizada. Las ponderaciones asignadas fueron establecidas de acuerdo con los objetivos operativos del modelo, otorgando un mayor peso al potencial de recuperación económica, seguido de la probabilidad de regularización, el riesgo de mora y, finalmente, la exposición económica normalizada.

La formulación del indicador se expresa mediante la siguiente ecuación:

$$\begin{aligned} \text{ScorePrioridadGestion} = & 0.40(\text{PagoEsperado}) \\ & + 0.30(\text{ProbabilidadRegularizacion}) \\ & + 0.20(\text{ProbabilidadMora}) \\ & + 0.10(\text{ExposiciónNormalizada}) \end{aligned}$$

Las ponderaciones asignadas reflejan la importancia relativa de cada componente dentro del proceso de priorización, privilegiando el potencial de recuperación económica de la cuenta, seguido de la probabilidad de regularización, el riesgo de mora y la exposición económica normalizada. Esta distribución de pesos responde al objetivo de equilibrar la maximización de la recuperación de cartera, la reducción del riesgo de mora y la optimización de los recursos operativos destinados a la gestión de cobranza. La justificación de estas ponderaciones y su comportamiento sobre la priorización de la cartera se analizan en el Capítulo 4.

Una vez obtenido el score compuesto, sus valores fueron normalizados a una escala comprendida entre 0 y 1000 puntos, generando el indicador ScorePrioridadGestion1000, el cual facilita su interpretación y utilización dentro de los procesos operativos de la organización. Esta estandarización permite integrar el modelo con sistemas de información y herramientas de gestión que emplean escalas normalizadas para la priorización de clientes.

El ScorePrioridadGestion constituye el principal resultado metodológico del modelo de priorización, ya que resume en un único indicador la información predictiva generada durante toda la investigación y sirve como base para la segmentación posterior de la cartera mediante deciles de gestión.

Segmentación de la cartera mediante deciles de gestión

Una vez calculado el ScorePrioridadGestion para las observaciones disponibles, se construyó el conjunto operativo final seleccionando la última observación registrada para cada cuenta. Para ello, los registros fueron ordenados por IdCuenta y FechaObservacion, conservando únicamente el corte temporal más reciente de cada cliente. Este procedimiento permitió pasar del conjunto longitudinal utilizado durante el desarrollo de los modelos a una estructura con una única observación por cuenta, conformada por 108.910 cuentas únicas. Sobre este conjunto se realizó la segmentación mediante deciles de gestión utilizando el ScorePrioridadGestion. Previamente, los valores del score fueron ordenados mediante un ranking para resolver posibles empates y posteriormente se aplicó una segmentación por cuantiles, generando diez grupos de tamaño equivalente. Como resultado, cada decil quedó conformado por aproximadamente el 10 % de las cuentas de la cartera.

La numeración de los deciles se estableció de acuerdo con el orden ascendente del ScorePrioridadGestion. En consecuencia, el Decil 1 concentra las cuentas con los valores más bajos del score y mayor criticidad relativa para la gestión, mientras que el Decil 10 agrupa las cuentas con los valores más altos y menor prioridad relativa de intervención. Esta interpretación se encuentra respaldada por el comportamiento de las variables predictivas y de exposición económica observado en cada segmento.

Posteriormente, cada cuenta fue clasificada según su TipoGestion. Las cuentas con ExposicionImpaga > 0 fueron clasificadas como Recuperativas, mientras que aquellas sin exposición impaga fueron clasificadas como Preventivas. La combinación de DecilGestion y TipoGestion constituyó la salida operativa del Modelo de Priorización y sirvió posteriormente como insumo para su integración con el Modelo de Contactabilidad y la definición final de las estrategias de cobranza.

Implementación del Modelo de Priorización de Gestión de Cartera

Una vez concluido el proceso de priorización y segmentación de la cartera, se consolidaron los resultados obtenidos en el conjunto resultadoGestionCartera, conformando una estructura operativa con una única observación vigente por cuenta.

El conjunto resultante integra las variables de identificación y fecha de observación, las probabilidades, scores y clasificaciones obtenidas por los Modelos Predictivos de Mora, Pago y Regularización, la exposición económica, el ScorePrioridadGestion, el DecilGestion, el TipoGestion y la fecha de generación del modelo. Esta estructura permite conservar en un único registro por cuenta los principales indicadores analíticos generados durante el proceso de priorización.

En esta etapa, el resultado del modelo permite establecer la prioridad relativa de cada cuenta y determinar la naturaleza preventiva o recuperativa de la gestión. La asignación de una estrategia específica de cobranza no forma parte de este proceso, ya que se realiza posteriormente mediante la integración del Modelo de Priorización con los resultados del Modelo Predictivo de Contactabilidad.

Finalmente, el conjunto de resultados fue preparado para su almacenamiento en SQL Server, permitiendo su utilización como insumo para las etapas posteriores de integración y definición de estrategias, así como para su futura explotación dentro de los sistemas operativos de la organización.

Conclusión del desarrollo

El Modelo de Priorización de Gestión de Cobranza integró los resultados generados por los modelos predictivos de Mora, Pago y Regularización en un único esquema de apoyo a la toma de decisiones para la gestión de cobranza. Mediante la combinación del potencial de recuperación, la probabilidad de regularización, el riesgo de mora y la exposición

económica normalizada, fue posible construir un indicador compuesto que permitió priorizar las observaciones de la cartera de acuerdo con su nivel de atención requerido.

La segmentación mediante deciles y la clasificación entre gestión recuperativa y preventiva, constituyeron una propuesta metodológica orientada a optimizar la asignación de recursos de cobranza, proporcionando una base objetiva para la planificación de acciones de gestión de acuerdo con las características predictivas y financieras de cada observación.

Finalmente, la implementación del modelo permitió consolidar los resultados en una estructura estandarizada preparada para su integración con los sistemas de información de la organización. El resultado final corresponde a una única clasificación por cuenta para el período de evaluación, incorporando el Score de Prioridad de Gestión, el decil asignado y el tipo de gestión, constituyendo uno de los principales insumos para la ejecución de las acciones de cobranza dentro del proceso operativo de la fintech.

Desarrollo del Modelo Predictivo de Contactabilidad

El Modelo Predictivo de Contactabilidad fue desarrollado con el propósito de estimar la probabilidad de que un cliente pueda ser contactado exitosamente mediante los canales de gestión disponibles dentro del proceso de cobranza. Esta estimación permite transformar la información histórica de interacciones en una señal probabilística útil para priorizar esfuerzos, definir canales de contacto y reducir la intensidad operativa en aquellos casos donde la probabilidad de contacto efectivo es mayor.

A diferencia de los modelos predictivos orientados a estimar mora, pago o regularización, el modelo de contactabilidad se enfoca en la dimensión operativa de la gestión. Su objetivo no es predecir de forma directa la recuperación de cartera, sino estimar la factibilidad de establecer comunicación con el cliente a partir de registros históricos de contacto, uso de canales, frecuencia de gestión, antigüedad de interacciones y comportamiento observado en diferentes ventanas temporales.

El desarrollo del modelo siguió una metodología estructurada similar a la empleada en los modelos predictivos anteriores. Esta metodología incluyó la construcción y validación del target, la preparación del conjunto de datos, la depuración de variables, el entrenamiento de modelos de clasificación supervisada, la evaluación comparativa mediante métricas de desempeño y la implementación del modelo final para la generación de puntajes, segmentos y recomendaciones operativas de contacto.

Objetivo del Modelo Predictivo de Contactabilidad

El objetivo del Modelo Predictivo de Contactabilidad fue estimar la probabilidad de contacto efectivo de cada cuenta, utilizando como insumo la información histórica disponible en el conjunto de datos DSScoreContacto. Esta probabilidad permite identificar aquellas cuentas con mayor posibilidad de ser contactadas mediante canales como telefonía, SMS o correo electrónico.

Desde el punto de vista operativo, el modelo permite incorporar una dimensión adicional al proceso de priorización de gestión. Una cuenta con alta probabilidad de recuperación, pero baja probabilidad de contacto puede requerir una estrategia distinta a una cuenta con probabilidad de contacto alta y disponibilidad de canales digitales efectivos. Por esta razón, la variable de contactabilidad se convierte en un insumo complementario dentro del modelo integral de priorización de cartera.

El resultado del modelo se expresa en una probabilidad de contactabilidad, un segmento de interpretación, un decil de priorización y variables operativas adicionales que recomiendan el tipo de canal y el número de iteraciones de contacto. Esta salida facilita su integración posterior con los procesos de gestión y con la tabla ResultadoGestionCartera utilizada para consolidar los resultados finales.

Construcción y validación del TargetContactabilidad

La primera etapa del desarrollo consistió en la definición de la variable objetivo TargetContactabilidad. Esta variable fue construida como una variable binaria, donde el valor True representa los casos en los que se identificó contacto efectivo y el valor False representa los casos en los que no se observó dicho evento dentro del horizonte de información considerado.

La validación inicial del target permitió verificar la distribución de clases y confirmar la naturaleza desbalanceada del problema. En el dataset se observaron 518,887 registros, con 149,812 registros positivos y 369,075 negativos. Esto implica una proporción de contacto efectivo de 28.87%, frente a una proporción de 71.13% correspondiente a registros no contactables.

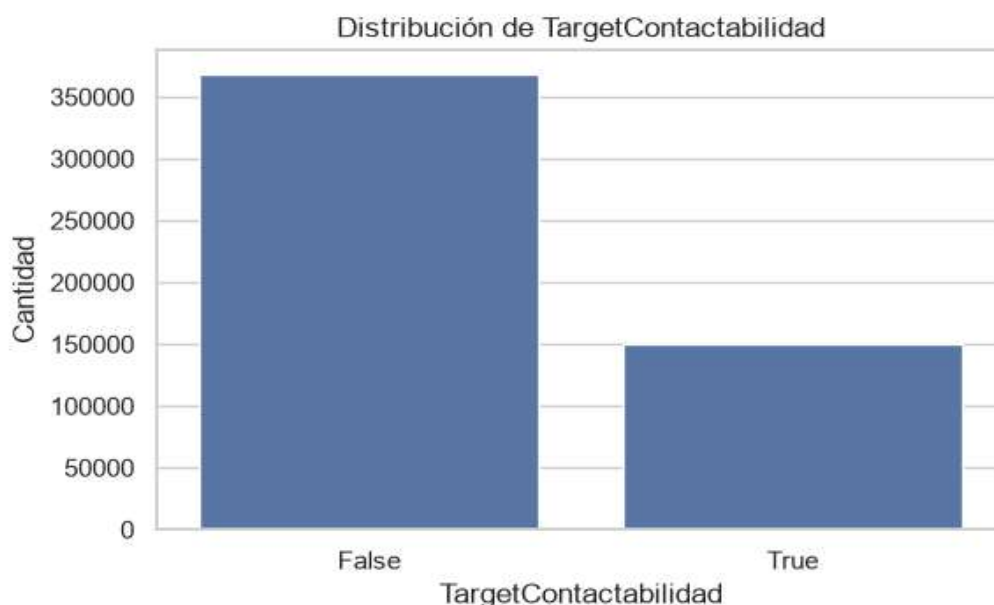
Este comportamiento justifica la utilización de métricas complementarias a la exactitud global, especialmente ROC-AUC, PR-AUC, Recall y F1-Score, debido a que el objetivo del modelo es discriminar correctamente aquellos casos con mayor probabilidad de contacto efectivo en un contexto donde la clase positiva no constituye la mayoría de las observaciones.

Tabla 4 *Resumen descriptivo del Dataset utilizado para el Modelo Predictivo de Contactabilidad*

Indicador	Valor
Total registros dataset completo	518,887
Total columnas predictoras luego de preparación	577
Registros con TargetContactabilidad = False	369,075
Registros con TargetContactabilidad = True	149,812
Proporción TargetContactabilidad = False	71.13%
Proporción TargetContactabilidad = True	28.87%

Nota. Elaboración propia.

Figura 20 Distribución del TargetContactabilidad utilizado para el entrenamiento del modelo



Nota. Elaboración propia.

Análisis exploratorio por muestra temporal

Como parte de la validación exploratoria, se analizó el comportamiento de la tasa de contactabilidad por muestra temporal. Este análisis permitió observar cambios importantes en la proporción de contactos efectivos a lo largo de las seis muestras consideradas.

Los resultados muestran que la tasa de contactabilidad presenta un incremento progresivo a partir de la tercera muestra. Mientras las muestras 1 y 2 registran tasas inferiores al 4%, las muestras 4, 5 y 6 presentan niveles superiores al 42%, alcanzando 52.59% en la muestra 6. Este comportamiento evidencia que la variable objetivo posee una estructura temporal relevante y que la muestra constituye una variable con alta capacidad explicativa dentro del modelo.

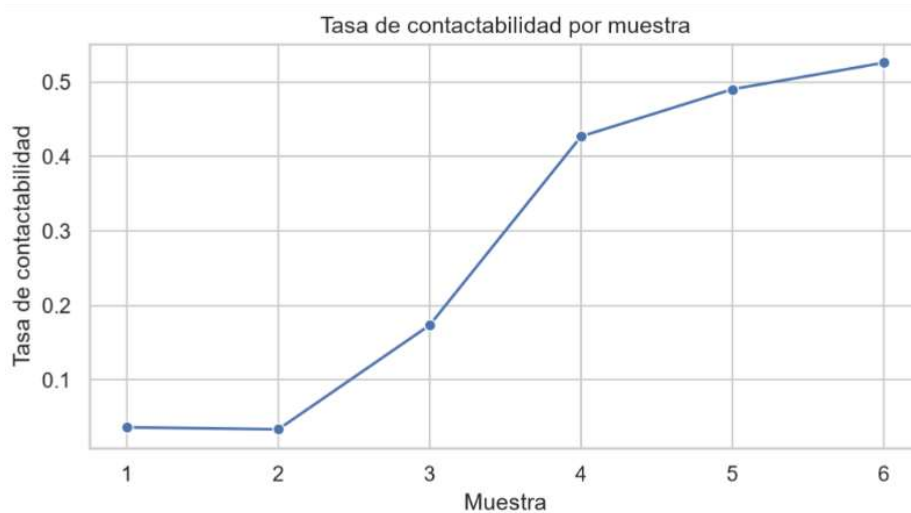
Adicionalmente, el análisis por muestra permitió revisar la intensidad promedio de los canales de contacto. En términos generales, el canal SMS presentó la mayor intensidad promedio frente a Email y Telefonía, lo cual posteriormente fue considerado para la recomendación de canal y la construcción de variables operativas asociadas al modelo.

Tabla 5 Evolución de la contactabilidad y de la intensidad de gestión por muestra temporal

Muestra	Registros	Cuentas	Tasa Contactabilidad	Prom SMS	Prom Email	Prom Telefonía	Prom Contactos	Ratio llamada exitosa
1	81,585	81,585	3.64%	13.23	4.05	2.69	19.97	0.0390
2	83,606	83,606	3.37%	15.53	4.87	2.73	23.13	0.0340
3	85,586	85,586	17.36%	15.43	5.33	2.68	23.44	0.0260
4	87,488	87,488	42.70%	14.75	5.32	2.72	22.80	0.0245
5	89,524	89,524	49.03%	13.92	5.32	2.76	22.00	0.0246
6	91,098	91,098	52.59%	13.75	5.34	2.85	22.12	0.0231

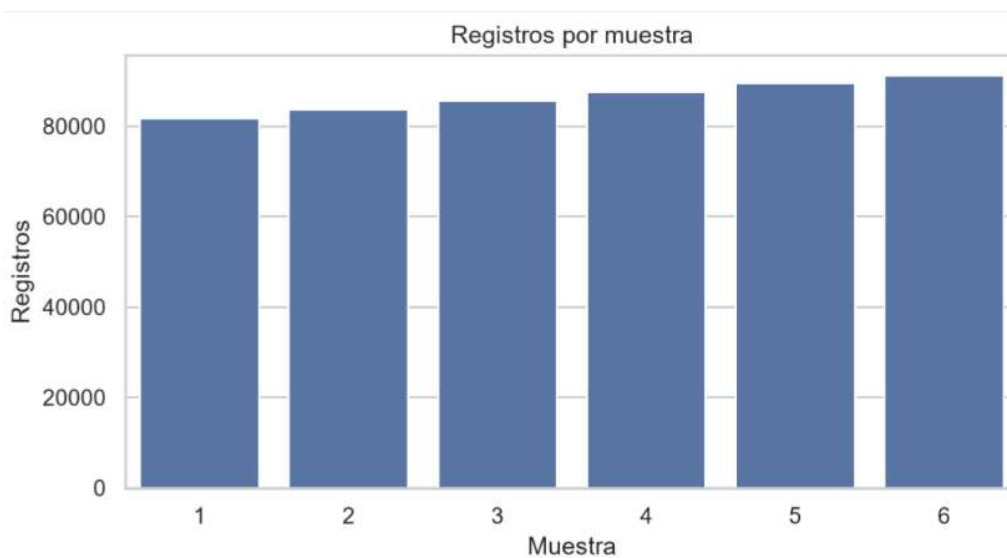
Nota. Elaboración propia.

Figura 21 Evolución de la tasa de contactabilidad

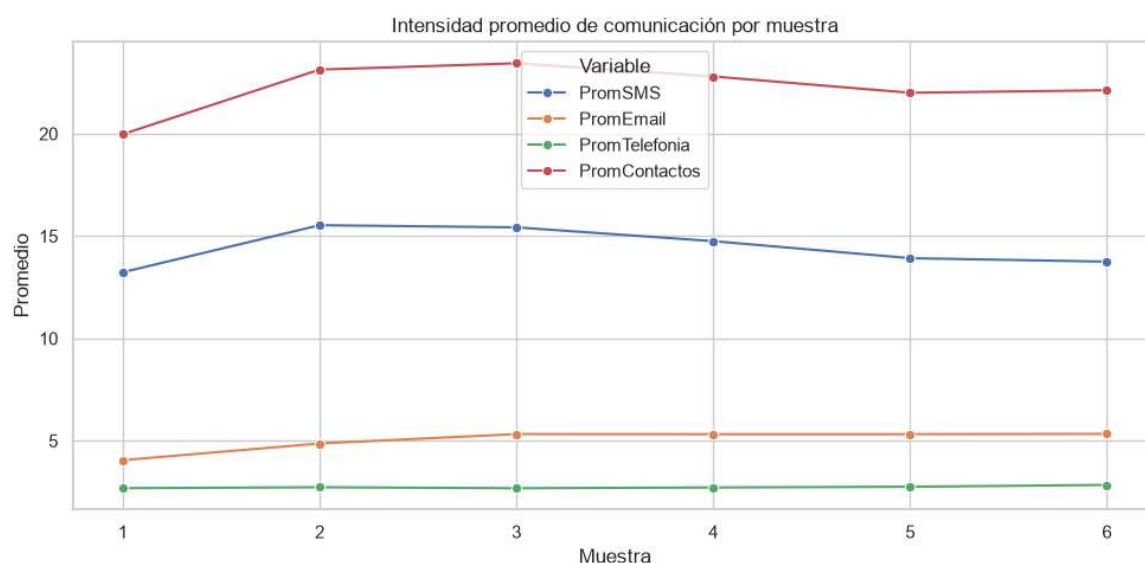


Nota. Elaboración propia.

Figura 22 Volumen de registros por muestra



Nota. Elaboración propia.

Figura 23 *Intensidad promedio de canales de contacto por muestra temporal*

Nota. Elaboración propia.

Preparación del conjunto de datos

Una vez validada la variable objetivo, se procedió a preparar la matriz de variables explicativas. Para ello se partió del dataset DSScoreContacto, excluyendo el identificador único de cuenta y las variables que no debían participar directamente en el aprendizaje supervisado. Posteriormente, se separó la variable TargetContactabilidad como variable dependiente y se construyó la matriz XContactabilidad.

Figura 24 *Preparación del conjunto de datos para el entrenamiento del Modelo Predictivo de Contactabilidad*

Nota. Elaboración propia.

El conjunto de datos fue sometido a procesos de imputación para garantizar que los algoritmos pudieran entrenarse de manera consistente ante la presencia de valores faltantes. En los modelos basados en árboles se aplicó imputación por mediana, mientras que para Regresión Logística se utilizó un pipeline con imputación y escalamiento, con el propósito de estabilizar el entrenamiento y mantener condiciones homogéneas de evaluación. La preparación también incluyó la depuración de variables no numéricas y la validación de dimensiones antes de la partición de entrenamiento y prueba. Como resultado, se obtuvo una matriz predictiva con 577 variables y 518,887 registros disponibles para el desarrollo del modelo.

Selección y depuración de variables

La selección y depuración de variables tuvo como propósito reducir la presencia de información redundante y mejorar la estabilidad del proceso de entrenamiento. Se analizaron correlaciones entre variables, identificando pares con relaciones prácticamente perfectas, especialmente entre variables de frecuencia mensual, número de contactos y variables derivadas de canales específicos.

La identificación de estas relaciones permitió fundamentar la eliminación o control de variables redundantes, evitando que múltiples variables representaran la misma señal de comportamiento. Este proceso resulta especialmente relevante en modelos con un número elevado de variables, donde la redundancia puede afectar la interpretación del modelo y elevar innecesariamente la complejidad de entrenamiento.

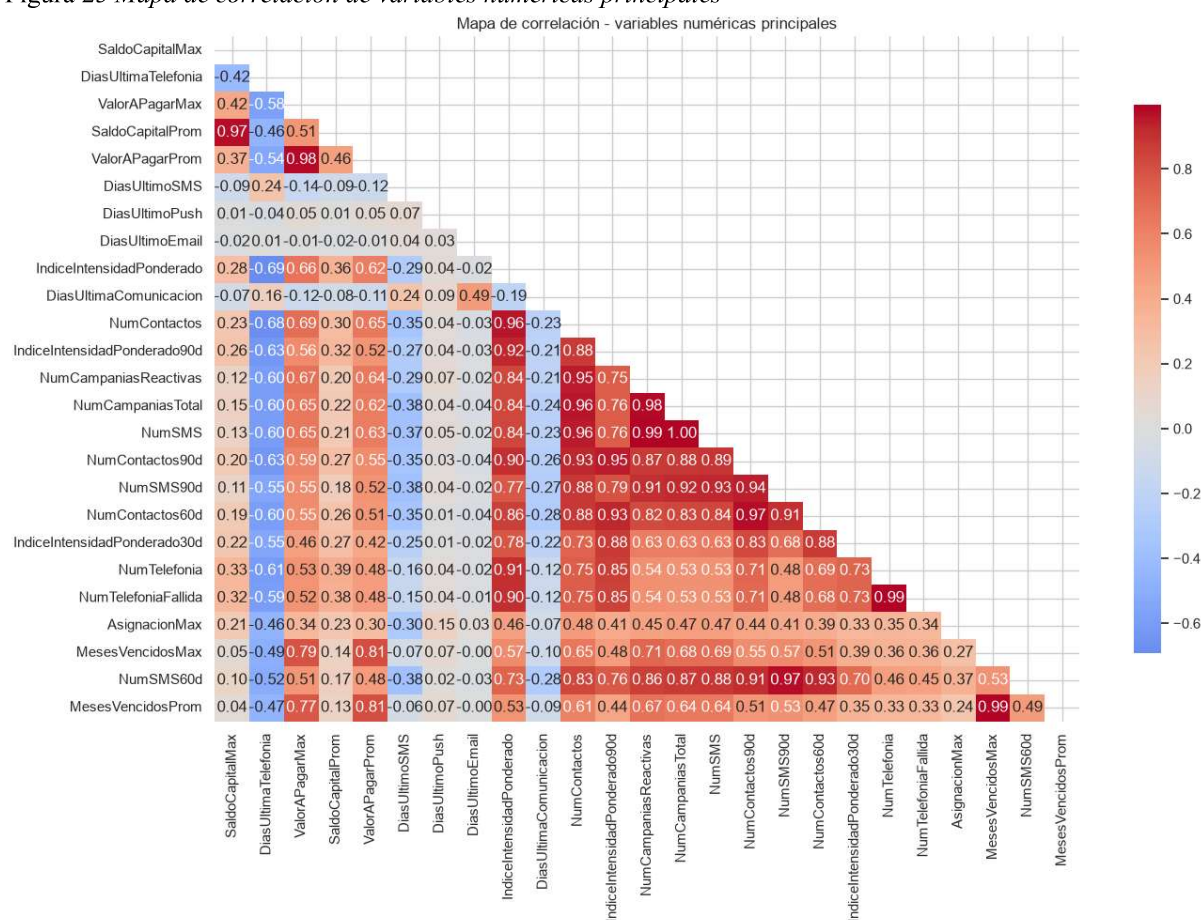
Además del análisis de correlación, se revisó la importancia de variables generada por XGBoost. Este análisis permitió identificar que la variable Muestra fue la característica con mayor aporte relativo, seguida por variables asociadas a vencimiento, antigüedad del último email, campañas preventivas, asignación y comportamiento omnicanal.

Tabla 6 Principales pares de variables altamente correlacionadas identificadas en el análisis exploratorio

Variable 1	Variable 2	Correlación
NumEmail	NumCampaniasInformativas	1.0000
UltimoLocal TERMINAL EXPRESS	LocalPredominante TERMINAL EXPRESS	1.0000
UltimoLocal	LocalPredominante	1.0000
NumContactos30d	FrecuenciaMensualContacto30d	1.0000
NumContactos60d	FrecuenciaMensualContacto60d	1.0000
UltimoLocal CARPA ESMERALDAS	LocalPredominante CARPA ESMERALDAS	1.0000
NumContactos90d	FrecuenciaMensualContacto90d	1.0000
NumContactos	FrecuenciaMensualContacto	1.0000
NumSMS	FrecuenciaMensualSMS	1.0000
NumTelefonia	FrecuenciaMensualTelefonia	0.9999

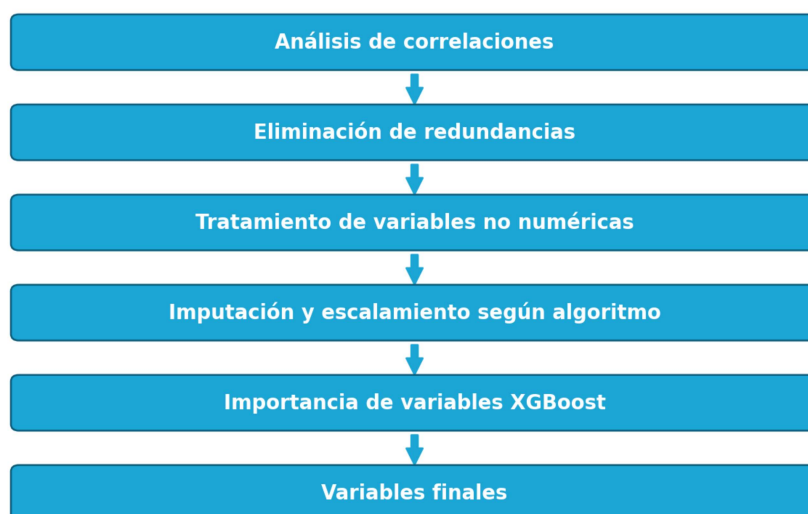
Nota. Elaboración Propia

Figura 25 Mapa de correlación de variables numéricas principales



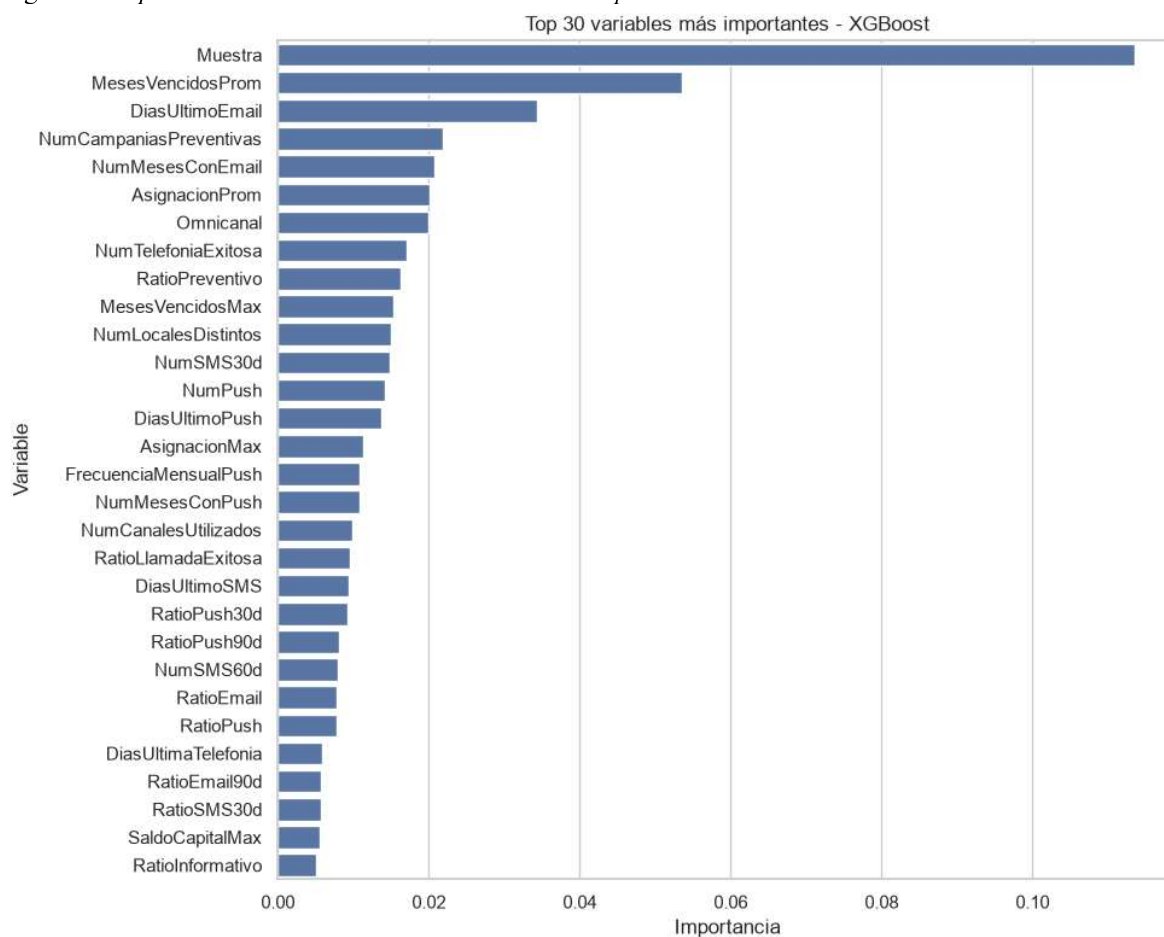
Nota. Elaboración propia.

Figura 26 *Proceso de selección y depuración de variables para el Modelo Predictivo de Contactabilidad*



Nota. Elaboración propia.

Figura 27 *Importancia de variables del modelo XGBoost para contactabilidad*



Nota. Elaboración propia.

Metodología de entrenamiento

El entrenamiento del Modelo Predictivo de Contactabilidad se plantea como un problema de clasificación supervisada. Para ello, se dividen los datos en conjuntos de entrenamiento y prueba, procurando conservar la proporción de clases mediante una partición estratificada. Se consideran modelos candidatos con características complementarias: un modelo lineal como línea base, un modelo de ensamble basado en árboles y un modelo de boosting. Esta comparación permite evaluar tanto alternativas simples e interpretables como algoritmos no lineales con mayor capacidad para capturar relaciones complejas.

Dado que la contactabilidad suele presentar desbalance entre clases, la evaluación debe considerar métricas robustas para clasificación, priorizando no solo la capacidad global de discriminación sino también el comportamiento sobre la clase positiva.

Evaluación de modelos

La evaluación comparativa de modelos consideró métricas complementarias orientadas a medir tanto la capacidad de discriminación como el desempeño sobre la clase positiva. Se utilizaron ROC-AUC y PR-AUC para evaluar el comportamiento probabilístico del modelo, así como Precision, Recall y F1-Score para analizar el desempeño de clasificación con umbral igual a 0,5.

Las métricas obtenidas corresponden a una validación interna realizada sobre el Dataset disponible mediante la partición de entrenamiento y prueba. La construcción temporal de las variables explicativas y de la variable objetivo permitió mantener la separación entre la información histórica utilizada por el modelo y el resultado posterior de contactabilidad, evitando la incorporación de información futura durante el entrenamiento.

La selección del modelo final se realiza a partir de la comparación de los modelos candidatos bajo las métricas definidas. El criterio de selección debe considerar el equilibrio

entre capacidad discriminativa, sensibilidad sobre la clase positiva, estabilidad del modelo y facilidad de integración operativa.

En esta sección se describe el criterio metodológico de selección, pero no se presentan valores específicos de desempeño. Las métricas comparativas y la justificación cuantitativa del modelo seleccionado se documentan en el Capítulo 4.

Una vez seleccionado el modelo definitivo, se preservan los elementos necesarios para la inferencia posterior: variables finales, mecanismo de imputación y objeto del modelo entrenado.

Tabla 7 *Criterios metodológicos para seleccionar el modelo final*

Criterio	Descripción
Capacidad discriminativa	Preferencia por modelos con mejor separación entre contactables y no contactables.
Sensibilidad operativa	Priorización del Recall cuando se requiere identificar la mayor cantidad de cuentas contactables.
Balance de clasificación	Uso de F1-Score para revisar equilibrio entre Precision y Recall.
Integración técnica	Disponibilidad de variables, imputador y modelo para inferencia posterior.

Nota. Elaboración propia.

Generación de probabilidad, score y segmentos de contactabilidad

El modelo final genera una probabilidad individual de contactabilidad para cada registro evaluado, esta probabilidad representa la estimación de que una cliente pueda ser contactado efectivamente y constituye la base para construir variables operativas de gestión. Para permitir su uso en reglas de negocio y tableros operativos, la probabilidad se transforma en un nivel de contactabilidad. La segmentación permite clasificar las cuentas en categorías interpretables, facilitando la definición de intensidad de contacto y priorización de esfuerzos.

Adicionalmente, la probabilidad puede integrarse en un ScoreContactabilidad, junto con variables de canal e iteraciones, para disponer de un puntaje operativo normalizado.

Tabla 8 Segmentos metodológicos de contactabilidad

NivelContactabilidad	Interpretación metodológica
Muy Baja	Segmento con menor probabilidad relativa de contacto efectivo.
Baja	Segmento con probabilidad limitada y necesidad de refuerzo operativo.
Media	Segmento de contactabilidad intermedia.
Alta	Segmento con condiciones favorables de contacto.
Muy Alta	Segmento con mayor probabilidad relativa de contacto efectivo.

Nota. Elaboración propia.

Figura 28 Salidas operativas generadas por el Modelo Predictivo de Contactabilidad



Nota. Elaboración propia.

La probabilidad de contactabilidad se complementa con variables operativas que permiten recomendar el canal y la intensidad de contacto. El canal sugerido se deriva de la disponibilidad e historial de interacción observado para cada cliente, mientras que el número de iteraciones se orienta de forma inversamente proporcional a la probabilidad estimada.

Desde una perspectiva metodológica, los clientes con menor probabilidad de contacto requieren mayor esfuerzo de interacción, mientras que aquellos con mayor contactabilidad pueden gestionarse con menor cantidad de iteraciones. Esta regla permite ajustar los recursos de gestión al esfuerzo esperado para lograr contacto efectivo.

La lógica de canal e iteraciones debe mantenerse parametrizable, de forma que pueda ajustarse posteriormente a restricciones operativas, disponibilidad de canales o políticas internas de cobranza.

Implementación del Modelo

El Modelo Predictivo de Contactabilidad se procesa sobre la totalidad del dataset, utilizando el imputador final y las variables conservadas durante el proceso de entrenamiento. La salida del mismo se integra con la estructura final de gestión de cartera para complementar el ScorePrioridadGestion y las variables provenientes de los modelos de Mora, Pago y Regularización.

Esta integración permite que la organización no solo identifique qué cuentas deben ser priorizadas, sino también cuáles presentan mayor factibilidad de contacto y qué intensidad de interacción podría ser más adecuada. De esta manera, la contactabilidad se convierte en una dimensión operativa del modelo de gestión.

La actualización operativa se realiza incorporando las variables de contactabilidad a la tabla de resultados de gestión, preservando la relación con el identificador de cuenta y facilitando su uso posterior en reglas de estrategia.

Tabla 9 *Campos de contactabilidad integrables al proceso de gestión*

Campo integrado	Uso esperado
ProbabilidadContactabilidad	Insumo para estimar la factibilidad de contacto.
NivelContactabilidad	Variable usada para segmentación y reglas de estrategia.
TipoContacto	Canal propuesto para la gestión.
NumeroIteraciones	Intensidad sugerida de interacción.
ScoreContactabilidad	Puntaje de apoyo para la priorización operativa.

Nota. Elaboración Propia

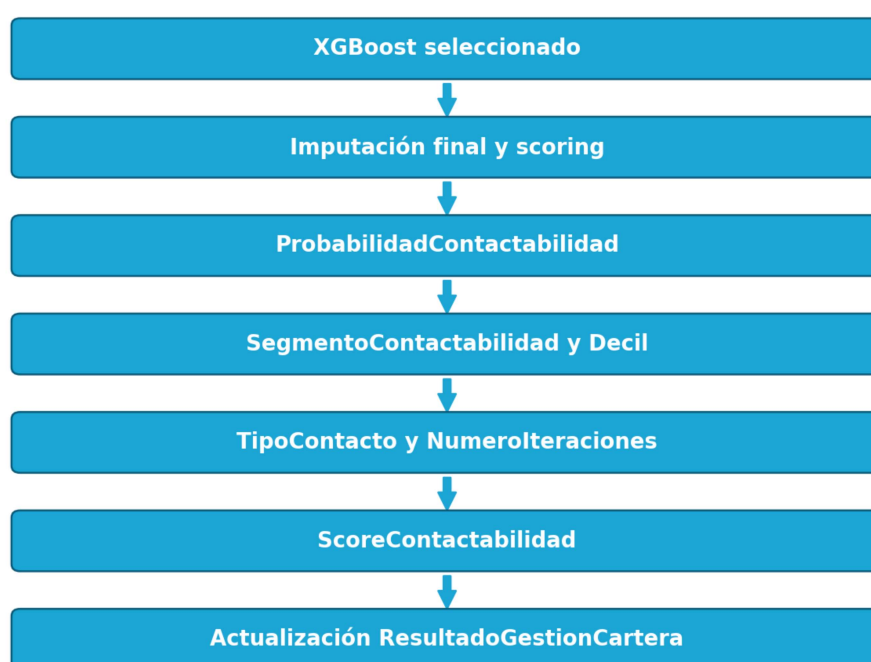
La construcción del Modelo Predictivo de Contactabilidad debe interpretarse como una fase metodológica orientada a incorporar factibilidad de contacto dentro del sistema de decisión de cobranza.

Tabla 10 Componentes utilizados para construir el ScoreContactabilidad

Componente	Descripción
ProbabilidadContactabilidad	Probabilidad estimada por el modelo, multiplicada por 700 puntos.
TipoContacto	Componente asociado al canal recomendado: Telefonía = 150, SMS = 100, Email = 80, Sin información = 50.
NumeroIteraciones	Componente operativo calculado como $(6 - \text{NumeroIteraciones}) * 30$, acotado entre 1 y 5 iteraciones.
ScoreContactabilidad	Suma de los componentes anteriores, acotada entre 0 y 1,000 puntos.

Nota. Elaboración propia.

Figura 29 Proceso de implementación del Modelo Predictivo de Contactabilidad



Nota. Elaboración Propia

Conclusión del desarrollo

El desarrollo del Modelo Predictivo de Contactabilidad permitió incorporar una dimensión operativa clave dentro de la solución integral de recuperación de cartera. A partir de la información histórica de contactos, canales, frecuencia de gestión y comportamiento temporal, el modelo estima la probabilidad de que un cliente pueda ser contactado exitosamente.

El análisis exploratorio evidenció una distribución desbalanceada del target, así como una evolución temporal significativa de la tasa de contactabilidad por muestra. Estos hallazgos

justificaron la preparación cuidadosa del dataset, la evaluación de variables correlacionadas y el uso de métricas robustas para comparar modelos de clasificación.

La comparación de algoritmos permite seleccionar el mejor modelo definitivo, debido a su desempeño en ROC-AUC, PR-AUC, Recall y F1-Score. La implementación de dicho modelo permite generar probabilidades, segmentos, deciles, recomendaciones de canal, número de iteraciones y un ScoreContactabilidad estandarizado, todos ellos integrables al proceso operativo de gestión de cartera.

En conjunto, el Modelo Predictivo de Contactabilidad complementa los modelos de Mora, Pago y Regularización, aportando información sobre la viabilidad de interacción con el cliente. De esta manera, contribuye a que el modelo de priorización de gestión no solo considere la probabilidad de incumplimiento o recuperación, sino también la factibilidad operativa de ejecutar acciones de contacto efectivas.

Implementación del Modelo de Priorización de Gestión de Cobranza

La implementación del Modelo de Priorización de Gestión de Cobranza corresponde a la fase metodológica en la que las salidas analíticas previamente consolidadas son transformadas en una decisión operativa de gestión de cobranza. A diferencia de los modelos predictivos desarrollados en secciones anteriores, esta etapa no busca estimar una nueva probabilidad, sino establecer una lógica de decisión que permita asignar una estrategia a cada cuenta.

Para mantener la trazabilidad metodológica con el notebook de implementación, la formulación del modelo se realiza únicamente con las variables utilizadas en dicha fase: TipoGestion, DecilGestion y NivelContactabilidad. Estas variables definen el espacio de decisión sobre el cual se construye la estrategia de cobranza recomendada.

Figura 30 *Flujo metodológico de implementación del Modelo de Priorización de Gestión de Cobranza*



Nota. Elaboración Propia

Propósito de la implementación

El propósito de esta implementación es convertir la información segmentada de cartera en una recomendación operativa de cobranza. El modelo actúa como una capa de decisión que toma como entrada variables categóricas y ordinales ya disponibles en la base consolidada y retorna una estrategia de gestión.

La implementación se concibe como un componente intermedio entre los modelos predictivos y el sistema de Gestión de Cobranza. Por esta razón, su valor metodológico se encuentra en la formalización de reglas, la definición del espacio de decisión, la trazabilidad de las condiciones utilizadas y la posibilidad de parametrizar posteriormente estas reglas dentro de los sistemas de la organización.

Tabla 11 *Propósito funcional de la implementación del modelo*

Elemento	Descripción
Entrada metodológica	Variables de decisión consolidadas por cuenta.
Proceso	Evaluación secuencial de reglas de decisión.
Salida	Asignación de <i>EstrategiaGestion</i> para cada cuenta.
Uso previsto	Parametrización operativa de estrategias diferenciadas de cobranza.

Nota. Elaboración Propia

Variables utilizadas en el modelo

Para generar la Estrategia recomendada se toman en cuenta tres variables fundamentales: la variable *TipoGestion* distingue la naturaleza de la intervención, *DecilGestion* ordena la cartera en función de prioridad operativa y *NivelContactabilidad* representa la accesibilidad esperada para establecer contacto con el cliente.

Aunque la tabla consolidada puede contener otras variables provenientes de los modelos predictivos anteriores, estas no forman parte directa de la regla de asignación en esta fase.

Por lo tanto, no se incorporan directamente como variables de decisión en esta fase de asignación de estrategias.

Tabla 12 *Variables utilizadas para la asignación de estrategias de cobranza*

Variable	Dominio / valores	Rol en el modelo
<i>TipoGestion</i>	Preventiva / Recuperativa	Define la naturaleza de la acción de cobranza.
<i>DecilGestion</i>	1 a 10	Representa la prioridad operativa de la cuenta.
<i>NivelContactabilidad</i>	Muy Alta, Alta, Media, Baja, Muy Baja, Sin Contacto	Orienta la intensidad y factibilidad de contacto.

Nota. Elaboración Propia

Figura 31 *Espacio de decisión definido por TipoGestion, DecilGestion y NivelContactabilidad*

Nota. Elaboración propia.

Transformación operativa de DecilGestion

La variable DecilGestion se usa como variable de prioridad. Para facilitar la lectura operativa de las reglas, esta variable se transforma en una agrupación denominada GrupoValor. Esta agrupación no reemplaza a DecilGestion ni introduce una variable adicional de formulación, únicamente resume los deciles en tres rangos de interpretación. La transformación permite que las reglas sean más comprensibles para el negocio y más fáciles de parametrizar en ambientes operativos. De esta forma, los deciles de mayor prioridad quedan asociados al grupo Alto, los deciles intermedios al grupo Medio y los deciles de menor prioridad al grupo Bajo.

Tabla 13 *Transformación operativa de DecilGestion en GrupoValor*

DecilGestion	GrupoValor derivado	Interpretación
1 - 2	Alto	Segmento de mayor prioridad operativa.
3 - 6	Medio	Segmento de prioridad intermedia.
7 - 10	Bajo	Segmento de menor prioridad relativa.

Nota. Elaboración Propia

Formulación del esquema de decisión

El modelo define un vector de decisión compuesto por tres elementos: $x_i = (TipoGestion_i, DecilGestion_i, NivelContactabilidad_i)$ La salida esperada corresponde a una estrategia de cobranza y_i , seleccionada dentro de un conjunto finito de alternativas operativas.

La asignación de la estrategia considera simultáneamente tres criterios: la naturaleza preventiva o recuperativa de la gestión, la prioridad operativa representada por el decil y la factibilidad de contacto determinada por el nivel de contactabilidad. La combinación de estos criterios permite establecer una regla de decisión consistente para asignar estrategias diferenciadas de cobranza de acuerdo con las características de cada cuenta.

Tabla 14 *Formulación del esquema de decisión para la asignación de estrategias*

Elemento	Notación	Descripción
Vector de entrada	$x_i = (\text{TipoGestion}_i, \text{DecilGestion}_i, \text{NivelContactabilidad}_i)$	Representa la cuenta dentro del espacio de decisión.
Salida del modelo	$y_i = \text{EstrategiaGestion}_i$	Estrategia de cobranza asignada.
Conjunto de estrategias	$S = \{s_1, s_2, \dots, s_7\}$	Alternativas disponibles para la gestión.
Función de decisión	$f(x_i) \rightarrow y_i$	Regla que asigna una estrategia según la combinación de variables.

Nota. Elaboración Propia

Tabla 15 *Criterios considerados para la asignación de estrategias*

Criterio de decisión	Interpretación metodológica
Coherencia con TipoGestion	La estrategia debe ser compatible con una gestión preventiva o recuperativa.
Alineación con DecilGestion	La estrategia debe reflejar la prioridad operativa asignada por el decil.
Adecuación a NivelContactabilidad	La estrategia debe considerar la factibilidad de contacto.

Nota. Elaboración propia.

Alternativas de estrategia

El conjunto de alternativas representa las posibles acciones de cobranza que puede asignar el motor de decisión. Cada estrategia corresponde a un nivel distinto de intensidad, automatización o intervención humana. Estas alternativas son tratadas como categorías de salida del modelo y no como resultados cuantitativos en esta sección.

La definición de estas alternativas permite que la decisión sea interpretable y pueda ser trasladada posteriormente a reglas de negocio, campañas, motores de cobranza o parametrización en sistemas transaccionales.

Tabla 16 *Alternativas de estrategia consideradas por el motor de decisión*

Estrategia	Propósito metodológico
Preventiva Alto Valor	Acción preventiva para segmentos prioritarios.
Gestor Especializado	Intervención humana especializada para casos recuperativos prioritarios y contactables.
Recuperación Alto Valor	Gestión recuperativa intensiva orientada a cuentas de alto valor.
Recuperación Digital	Gestión mediante canales digitales para segmentos con condiciones favorables de contacto.
Recuperación Bajo Costo	Acción de bajo costo para segmentos de baja contactabilidad.
Monitoreo y Seguimiento	Estrategia de seguimiento operativo de menor intensidad.
No Gestionar	Decisión de no asignar recursos activos cuando la combinación de variables lo justifica.

Nota. Elaboración propia.

Motor de reglas de asignación

La implementación se realiza mediante un motor de reglas secuenciales. Cada regla evalúa la combinación de TipoGestion, GrupoValor derivado de DecilGestion y NivelContactabilidad. El orden de evaluación es relevante, porque permite resolver primero los casos más específicos y posteriormente los casos generales.

La lógica de reglas conserva la trazabilidad de la decisión, dado que cada estrategia puede explicarse a partir de una condición explícita. Esta característica es importante para la adopción del modelo en procesos de negocio, ya que permite auditar, ajustar o reemplazar reglas sin alterar la estructura general de la solución.

Figura 32 Jerarquía metodológica de reglas de asignación



Nota. Elaboración propia.

Integración en la tabla operativa

Una vez determinada la estrategia recomendada, la implementación contempla la actualización de la columna EstrategiaGestion en la tabla ResultadoGestionCartera. Esta operación permite que la decisión generada por el modelo quede disponible para etapas posteriores de análisis, parametrización o ejecución operativa.

La integración se realiza conservando el identificador de cuenta como clave de actualización. De esta forma, la estrategia asignada puede ser trazada desde la cuenta

original hasta la salida final del modelo, manteniendo consistencia entre el proceso analítico y la base de datos operativa.

Tabla 17 *Integración metodológica de la estrategia en la tabla operativa*

Elemento	Descripción
Tabla destino	ResultadoGestionCartera
Clave de actualización	IdCuenta
Campo actualizado	EstrategiaGestion
Fuente del valor	EstrategiaRecomendada generada por el motor de reglas
Propósito	Disponer de la estrategia final para parametrización o análisis posterior

Nota. Elaboración propia.

Trazabilidad, parametrización y control

Una ventaja de la implementación basada en reglas explícitas es que cada decisión puede ser justificada mediante la combinación de las variables de entrada. Esto permite explicar por qué una cuenta fue asignada a una determinada estrategia sin depender de una caja negra algorítmica en esta fase de decisión.

Asimismo, la lógica implementada puede ser parametrizada en sistemas de cobranza mediante tablas de reglas, catálogos de estrategia o motores de decisión. Si la operación requiere modificar umbrales o redefinir estrategias, la estructura permite ajustar reglas sin rediseñar los modelos predictivos previos.

Finalmente, esta capa constituye un punto de control entre la analítica predictiva y la ejecución operativa. El modelo **multiobjetivo** no reemplaza la gestión de negocio, sino que proporciona un marco consistente para traducir los segmentos analíticos en acciones diferenciadas.

Tabla 18 *Atributos metodológicos de la implementación propuesta*

Criterio	Aporte metodológico
Explicabilidad	Cada estrategia se deriva de reglas explícitas.
Auditabilidad	La decisión puede rastrearse desde TipoGestion, DecilGestion y NivelContactabilidad.
Parametrización	Las reglas pueden implementarse como catálogo operativo.
Escalabilidad	La asignación puede aplicarse de forma masiva sobre la tabla consolidada.

Nota. Elaboración propia.

Capítulo 4

En este capítulo se presentan los resultados obtenidos durante el desarrollo de la investigación, evaluando el desempeño de los modelos predictivos de Mora, Pago y Regularización, así como del Modelo de Priorización de Gestión de Cobranza y el Modelo de Contactabilidad. El análisis comprende la interpretación de las métricas de desempeño, la comparación entre los algoritmos evaluados y la valoración del modelo integral propuesto para apoyar la priorización y definición de estrategias de gestión de cobranza. Los resultados corresponden a la metodología desarrollada utilizando seis ventanas temporales móviles para el entrenamiento de los modelos predictivos y a la generación de un conjunto final de resultados por cuenta para su aplicación operativa.

Adicionalmente, se analiza la capacidad del Modelo de Priorización y el Modelo de Contactabilidad para segmentar la cartera de clientes y proponer estrategias diferenciadas de gestión, considerando simultáneamente el riesgo de mora, el potencial de pago, la probabilidad de regularización, la exposición económica de cada cuenta y la probabilidad de contactabilidad. Finalmente, se compara la propuesta desarrollada con el esquema tradicional de gestión utilizado por la organización, identificando las principales ventajas operativas derivadas de la utilización de modelos predictivos e inteligencia artificial.

Resultados de los modelos predictivos

En esta sección se presentan los resultados obtenidos durante el entrenamiento, evaluación e implementación de los modelos predictivos desarrollados en la investigación. Para cada variable objetivo se analizan las métricas de desempeño alcanzadas por los algoritmos evaluados, la importancia de las variables predictoras, los principales casos de estudio y la justificación de la selección del modelo definitivo implementado en la solución propuesta. Los resultados corresponden al proceso metodológico desarrollado utilizando el conjunto de datos consolidado a partir de seis ventanas temporales móviles. Sobre este conjunto se

realizó la evaluación comparativa de los algoritmos bajo condiciones experimentales homogéneas, permitiendo seleccionar, para cada problema predictivo, el modelo con mejor desempeño de acuerdo con las métricas definidas para cada tipo de problema.

El análisis se desarrolla de forma independiente para los modelos predictivos de Mora, Pago, Regularización y Contactabilidad, permitiendo comparar el comportamiento de los diferentes algoritmos y justificar la elección del modelo con mejor desempeño para cada uno de los problemas abordados.

Resultados del Modelo Predictivo de Mora

El Modelo Predictivo de Mora fue evaluado mediante la comparación de tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost. Todos los modelos fueron entrenados utilizando el mismo conjunto de datos y las mismas condiciones experimentales, permitiendo realizar una comparación objetiva de su desempeño predictivo. Para la evaluación se emplearon las métricas de Accuracy, Precision, Recall, F1-Score y ROC-AUC, las cuales permiten valorar la capacidad de los modelos para identificar correctamente las cuentas con probabilidad de alcanzar una mora igual o superior a 45 días dentro del horizonte de predicción definido para la investigación.

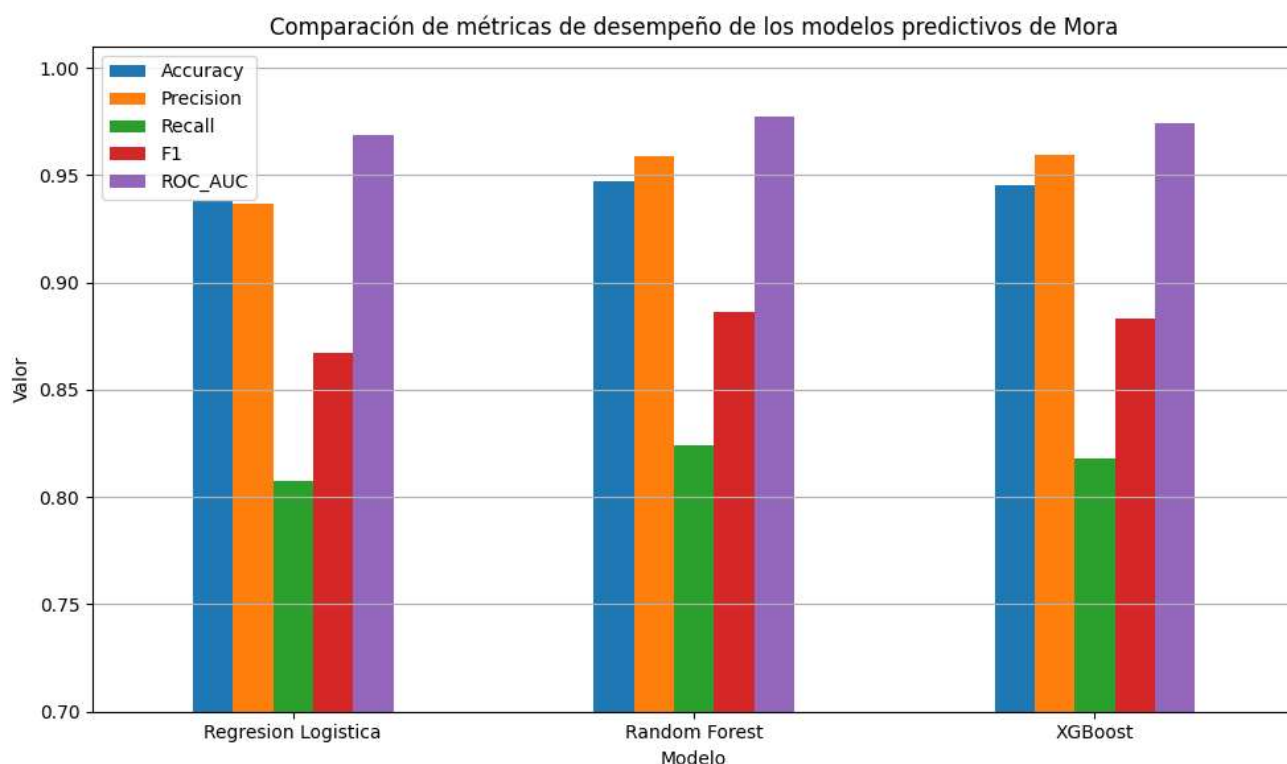
En la tabla se presenta la comparación de las métricas obtenidas por cada uno de los algoritmos evaluados, mientras que la figura resume gráficamente su desempeño. Estos resultados constituyen la base para la selección del modelo definitivo implementado en la solución propuesta.

Tabla 19 *Comparación de métricas de desempeño de los modelos predictivos de Mora*

Modelo	Accuracy	Precision	Recall	F1	ROC_AUC
Regresion Logistica	0.9379	0.9368	0.8075	0.8674	0.968388
RandomForest	0.9470	0.9589	0.8243	0.8865	0.977584
XGBoost	0.9455	0.9594	0.8177	0.8829	0.974107

Nota. Elaboración propia.

Figura 33 Comparación de métricas de desempeño de los modelos predictivos de Mora



Nota. Elaboración propia.

Como se observa en la tabla y la figura anexas, los tres algoritmos evaluados alcanzaron un desempeño adecuado para la predicción de mora, con valores de ROC-AUC superiores a 0,96 en todos los casos. Este resultado evidencia que las variables explicativas construidas a partir del Dataset longitudinal poseen capacidad discriminativa para diferenciar entre cuentas con y sin probabilidad de alcanzar mora durante el horizonte de predicción.

La Regresión Logística obtuvo un Accuracy de 0.9379, un F1-Score de 0.8674 y un ROC-AUC de 0.968388, constituyendo una línea base sólida para el problema de clasificación.

No obstante, su menor valor de Recall (0.8075) evidencia una capacidad inferior para identificar cuentas que efectivamente alcanzan la condición de mora.

El modelo Random Forest presentó el mejor desempeño global, alcanzando un Accuracy de 0.9470, un Recall de 0.8243, un F1-Score de 0.8865 y el mayor ROC-AUC (0.977584).

Estos resultados indican que el algoritmo logró el mejor equilibrio entre capacidad discriminativa, detección de casos positivos y estabilidad general del modelo.

Por su parte, XGBoost obtuvo resultados competitivos, con un Accuracy de 0.9455, una Precision de 0.9594, un F1-Score de 0.8829 y un ROC-AUC de 0.974107. Si bien alcanzó la mayor precisión entre los modelos evaluados, fue superado por Random Forest en las métricas de Accuracy, Recall, F1-Score y ROC-AUC.

Considerando que el objetivo del Modelo Predictivo de Mora consiste en identificar adecuadamente cuentas con riesgo de incumplimiento, se priorizó el equilibrio entre capacidad discriminativa y sensibilidad del modelo. Bajo este criterio, Random Forest fue seleccionado como modelo definitivo para la predicción de mora, al presentar el mayor valor de ROC-AUC y el mejor equilibrio entre Accuracy, Recall y F1-Score y su posterior integración en el Modelo de Priorización de Gestión de Cobranza.

Como complemento al análisis de desempeño, se evaluó la importancia de las variables utilizadas por el modelo Random Forest seleccionado como algoritmo definitivo. La figura presenta las quince variables con mayor contribución al proceso de clasificación.

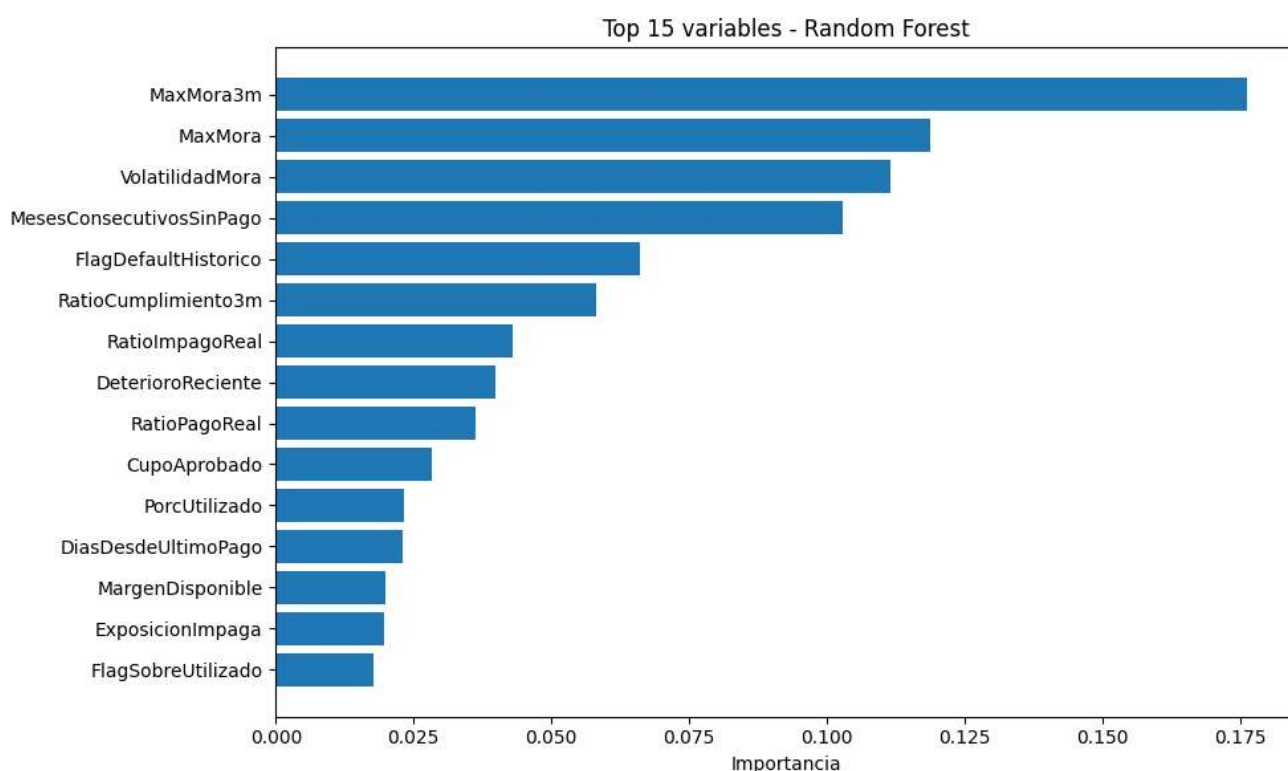
Los resultados evidencian que las variables relacionadas con el comportamiento histórico de la mora constituyen los predictores de mayor relevancia para la estimación del riesgo de incumplimiento. La variable MaxMora3m presentó la mayor importancia, seguida por MaxMora, VolatilidadMora y MesesConsecutivosSinPago, lo que indica que el comportamiento reciente y la severidad histórica de la mora representan los principales factores asociados a la ocurrencia de incumplimiento futuro.

Asimismo, variables como FlagDefaultHistorico, RatioCumplimiento3m, RatioImpagoReal y DeterioroReciente también mostraron una contribución significativa al modelo, confirmando que tanto el historial de pago como los indicadores recientes de deterioro financiero aportan información relevante para la identificación de cuentas con mayor riesgo de mora. Finalmente, variables relacionadas con la utilización del crédito y la exposición económica, como CupoAprobado, PorcUtilizado, ExposicionImpaga y MargenDisponible,

complementan la capacidad predictiva del modelo al incorporar información sobre el nivel de endeudamiento y la capacidad financiera de cada cuenta.

La distribución de importancia obtenida es consistente con la lógica del proceso de cobranza y respalda la capacidad del modelo para identificar los factores que mejor explican el riesgo de mora.

Figura 34 Variables con mayor importancia en el Modelo Predictivo de Mora mediante Random Forest.

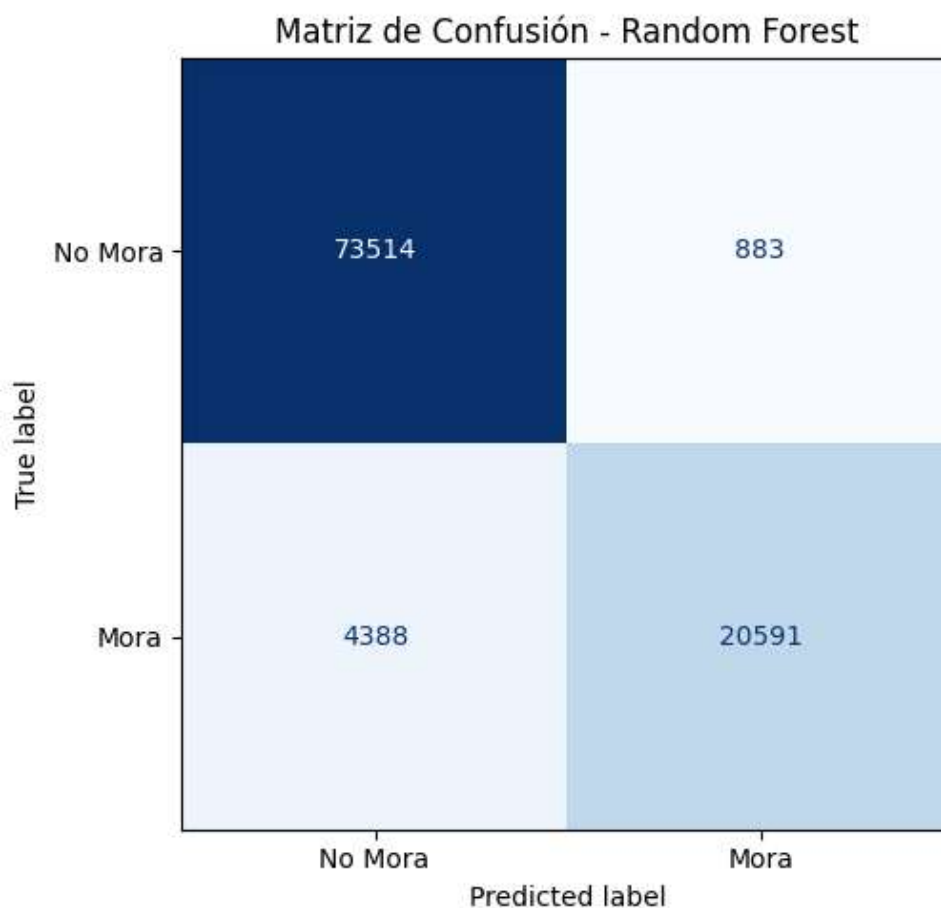


Nota. Elaboración propia.

La figura presenta la matriz de confusión correspondiente al modelo Random Forest. El modelo clasificó correctamente 73,514 cuentas sin mora (verdaderos negativos) y 20,591 cuentas con mora (verdaderos positivos). Adicionalmente, registró 883 cuentas clasificadas erróneamente como casos de mora (falsos positivos) y 4,388 cuentas con mora no fueron identificadas por el modelo (falsos negativos), reflejando un adecuado equilibrio entre la identificación de clientes con riesgo de mora y la minimización de clasificaciones incorrectas.

Estos resultados son consistentes con las métricas obtenidas durante la evaluación del modelo, particularmente con el Recall de 0.8243 y el ROC-AUC de 0.9776, confirmando una elevada capacidad discriminativa y una adecuada generalización sobre datos no utilizados durante el entrenamiento.

Figura 35 *Matriz de confusión del Modelo Predictivo de Mora mediante Random Forest*



Nota. Elaboración propia.

Con el propósito de ilustrar el comportamiento del modelo en diferentes escenarios de riesgo, se seleccionaron tres cuentas representativas correspondientes a los niveles Bajo, Medio y Crítico, de acuerdo con el ScoreMora obtenido por el modelo Random Forest. El primer caso corresponde a un cliente clasificado con riesgo Bajo, cuyo modelo estimó una probabilidad de mora de 0.0014, equivalente a un ScoreMora de 1. Este resultado

refleja un comportamiento histórico favorable, caracterizado por ausencia de mora, cumplimiento oportuno de las obligaciones y un reducido nivel de exposición al riesgo.

El segundo caso representa un cliente con riesgo Medio, para el cual el modelo estimó una probabilidad de mora de 0.5000, equivalente a un ScoreMora de 500. En este escenario se observa la presencia de algunos indicadores de deterioro, como períodos recientes de mora y antecedentes de refinanciamiento, que incrementan la incertidumbre sobre el comportamiento futuro del cliente.

Finalmente, el tercer caso corresponde a un cliente clasificado con **riesgo Crítico**, cuya probabilidad estimada de mora fue 0.999986, equivalente a un ScoreMora de 1000. Este cliente presenta múltiples señales de deterioro financiero, entre ellas elevados niveles de mora histórica, incumplimientos consecutivos y restricciones en el uso del crédito, características que justifican la máxima prioridad dentro de las estrategias de gestión de cobranza.

Estos casos evidencian que el modelo no solo proporciona una clasificación binaria sobre la ocurrencia de mora, sino que genera una estimación probabilística que posteriormente es transformada en un ScoreMora comprendido entre 0 y 1000, facilitando la interpretación de los resultados y su utilización en los procesos de priorización de la gestión de cobranza.

Resultados del Modelo Predictivo de Pago

El Modelo Predictivo de Pago fue evaluado mediante la comparación de tres algoritmos de regresión: Regresión Lineal, Random Forest Regressor y XGBoost Regressor. La evaluación se realizó utilizando las métricas Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y Coeficiente de Determinación (R^2), las cuales permiten valorar la precisión de las predicciones y la capacidad explicativa de cada modelo.

La tabla a continuación presenta la comparación de las métricas obtenidas por los algoritmos evaluados, mientras que la figura resume gráficamente su desempeño.

Como se observa en los resultados, la Regresión Lineal presentó el menor desempeño predictivo, alcanzando un MAE de 0.140593, un RMSE de 0.221250 y un R^2 de 0.723308. Estos resultados evidencian que, aunque el modelo logra capturar parcialmente la relación entre las variables explicativas y el porcentaje de pago esperado, su capacidad para representar el comportamiento de la cartera es inferior a la obtenida mediante modelos basados en árboles de decisión.

El algoritmo XGBoost Regressor mejoró significativamente el desempeño respecto a la Regresión Lineal, alcanzando un MAE de 0.117688, un RMSE de 0.204040 y un R^2 de 0.764679. Sin embargo, el mejor resultado fue obtenido por Random Forest Regressor, con un MAE de 0.113711, un RMSE de 0.201274 y un coeficiente de determinación (R^2) de 0.771015, evidenciando una mayor capacidad para modelar las relaciones no lineales presentes en el comportamiento de pago de la cartera.

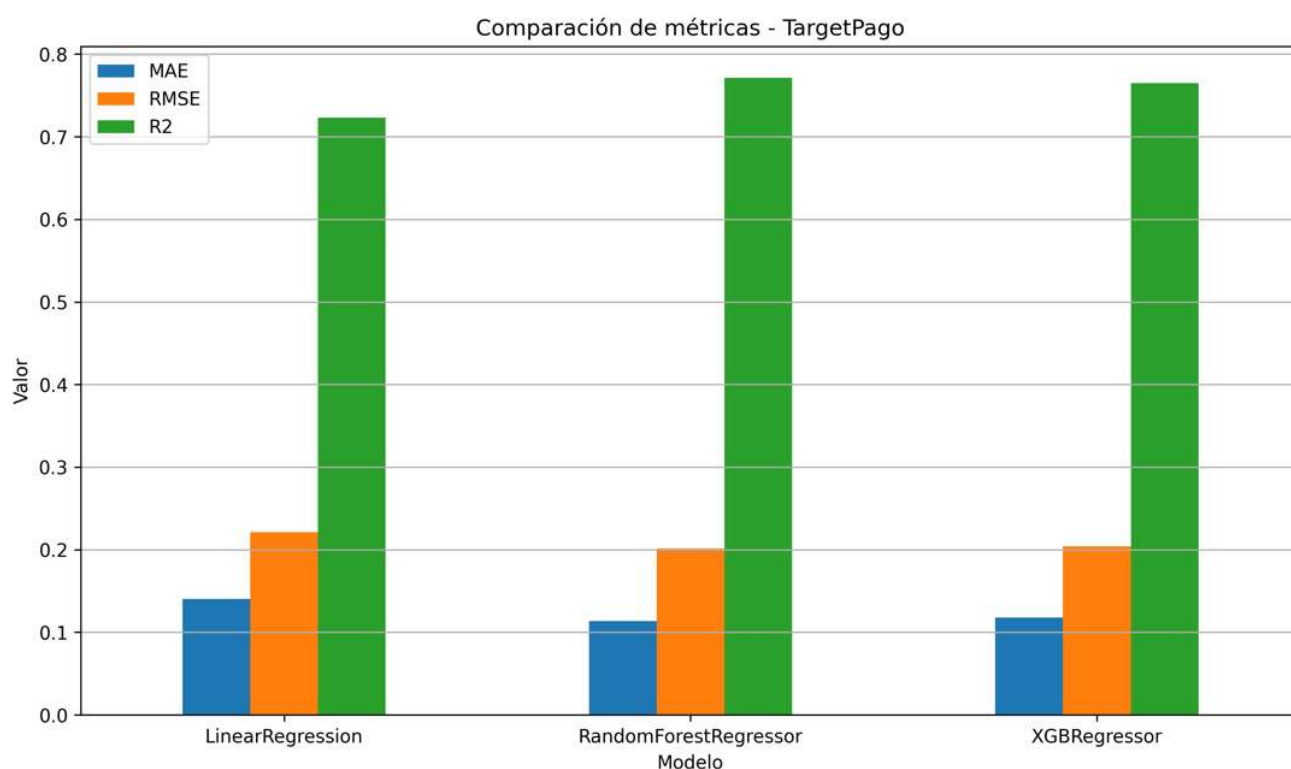
Considerando que el objetivo del modelo consiste en maximizar la capacidad explicativa y minimizar el error de predicción, se seleccionó Random Forest Regressor como modelo definitivo para la estimación del PagoEsperado. Este algoritmo obtuvo simultáneamente el mayor valor de R^2 y los menores valores de MAE y RMSE entre los modelos evaluados, constituyéndose en el segundo componente del Modelo de Priorización de Gestión de Cobranza.

Tabla 20 *Comparación de métricas de desempeño de los modelos predictivos de Pago*

Modelo	MAE	RMSE	R2
LinearRegression	0.140593	0.221250	0.723308
RandomForestRegressor	0.113711	0.201274	0.771015
XGBRegressor	0.117688	0.204040	0.764679

Nota. Elaboración propia.

Figura 36 Comparación de métricas de desempeño de los modelos predictivos de Pago



Nota. Elaboración propia.

Como complemento al análisis de desempeño, se evaluó la importancia de las variables del modelo seleccionado. La figura con los Top de las variables presenta las veinte variables con mayor contribución al proceso de predicción del porcentaje de pago esperado mediante el algoritmo Random Forest Regressor.

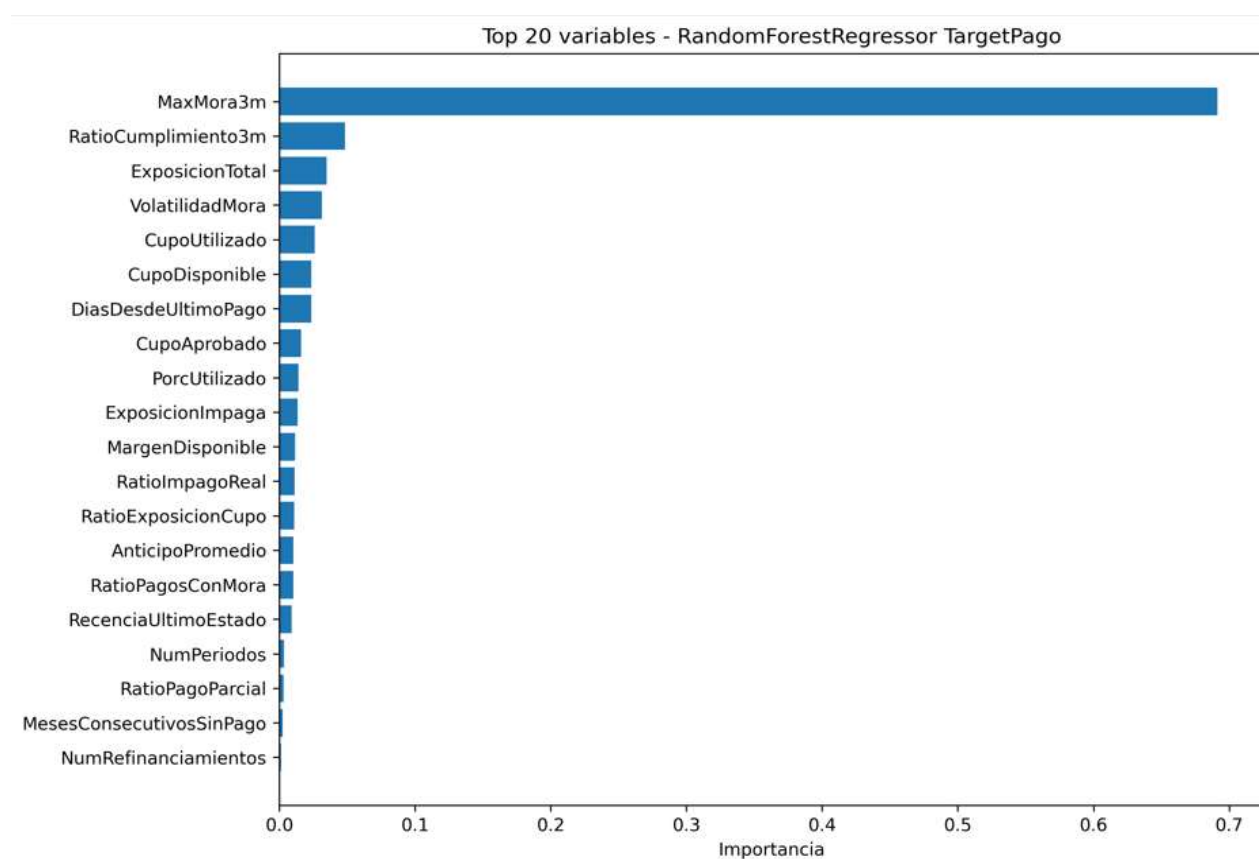
Los resultados muestran un predominio de la variable MaxMora3m, la cual concentra la mayor contribución al proceso de estimación del porcentaje esperado de pago. Este comportamiento evidencia que el nivel máximo de mora registrado durante los tres meses previos constituye el principal indicador para anticipar la capacidad de recuperación de una cuenta.

En un segundo nivel de importancia se ubican variables relacionadas con el comportamiento histórico de pago y la exposición financiera del cliente, entre las que destacan RatioCumplimiento3m, ExposicionTotal, VolatilidadMora, CupoUtilizado y CupoDisponible. Asimismo, variables asociadas con la utilización del crédito, la exposición

impaga y la antigüedad del último pago complementan la capacidad predictiva del modelo, permitiendo incorporar tanto el historial de cumplimiento como la situación financiera reciente de cada cuenta.

En conjunto, la distribución de importancia de las variables evidencia que el algoritmo concentra la mayor capacidad predictiva en indicadores relacionados con el comportamiento reciente del cliente, aspecto coherente con la naturaleza del problema de regresión abordado en esta investigación.

Figura 37 Variables con mayor importancia en el Modelo Predictivo de Pago mediante Random Forest Regressor.



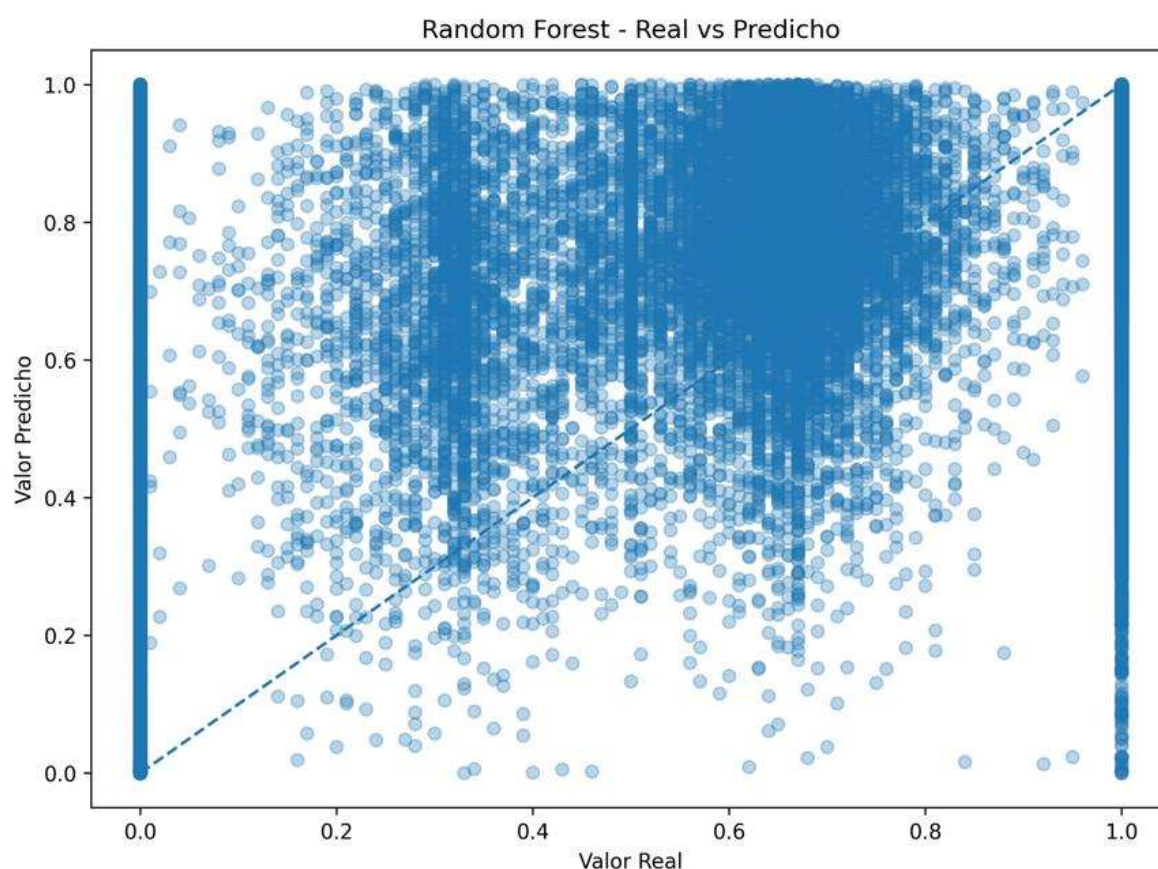
Nota. Elaboración propia.

La capacidad predictiva del modelo seleccionado fue evaluada adicionalmente mediante la comparación entre los valores reales y los valores estimados por Random Forest Regressor, presentada en la figura comparativa.

En un escenario ideal, las observaciones deberían concentrarse sobre la línea diagonal, indicando una correspondencia perfecta entre los valores reales y las predicciones generadas por el modelo. Los resultados obtenidos muestran que una proporción importante de las observaciones se distribuye alrededor de dicha línea de referencia, evidenciando una adecuada capacidad de estimación para diferentes niveles de pago esperado.

Se observa una concentración importante de predicciones en los extremos del intervalo $[0,1]$, comportamiento consistente con la distribución del TargetPago, donde predominan cuentas con recuperación nula o recuperación total. A pesar de la dispersión observable en los valores intermedios, el modelo logra capturar adecuadamente la tendencia general del comportamiento de pago, resultado coherente con el coeficiente de determinación obtenido durante la evaluación.

Figura 38 Comparación entre valores reales y valores predichos del Modelo Predictivo de Pago mediante Random Forest Regressor



Nota. Elaboración propia.

Los casos de estudio presentados permiten ilustrar el comportamiento del modelo Random Forest Regressor frente a diferentes perfiles de pago.

El primer caso corresponde a una cuenta con pago esperado nulo, caracterizada por antecedentes de mora, deterioro reciente y un nivel de utilización del crédito cercano al 68%, para la cual el modelo estimó correctamente un porcentaje de pago igual a cero.

El segundo caso representa una cuenta con un comportamiento de riesgo aún más crítico, evidenciado por seis meses consecutivos sin pago y un historial de mora más severo, obteniendo igualmente una predicción de pago nulo consistente con el comportamiento observado.

Finalmente, el tercer caso corresponde a una cuenta con historial de cumplimiento favorable, ausencia de mora reciente y una utilización moderada del crédito, para la cual el modelo estimó correctamente un pago esperado del 100%.

En los tres ejemplos el error absoluto fue igual a cero, evidenciando la capacidad del modelo para representar adecuadamente diferentes escenarios de recuperación de cartera.

Resultados del Modelo Predictivo de Regularización

El Modelo Predictivo de Regularización fue evaluado mediante la comparación de tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost. Para valorar su desempeño se utilizaron las métricas Accuracy, Precision, Recall, F1-Score y ROC-AUC, las cuales permiten medir la capacidad de los modelos para identificar correctamente las cuentas con probabilidad de regularizar su situación financiera durante el horizonte de predicción establecido.

La tabla presenta la comparación de las métricas obtenidas por los algoritmos evaluados, mientras que la figura resume gráficamente su comportamiento.

Como se observa en la tabla y la figura correspondiente, los tres algoritmos evaluados alcanzaron un desempeño sobresaliente para la predicción de la regularización de cartera,

obteniendo valores de ROC-AUC superiores a 0.97 en todos los casos. Estos resultados evidencian una elevada capacidad para diferenciar entre las cuentas que lograrán regularizar su situación y aquellas que no lo harán dentro del horizonte de análisis.

La Regresión Logística constituyó una línea base robusta, alcanzando un Accuracy de 0.9545 y un ROC-AUC de 0.974896. Por su parte, XGBoost obtuvo el mayor Recall (0.9968), mientras que Random Forest presentó el mejor equilibrio entre las métricas de evaluación, alcanzando un Accuracy de 0.9648, un Precision de 0.9595, un F1-Score de 0.9776 y un ROC-AUC de 0.980079.

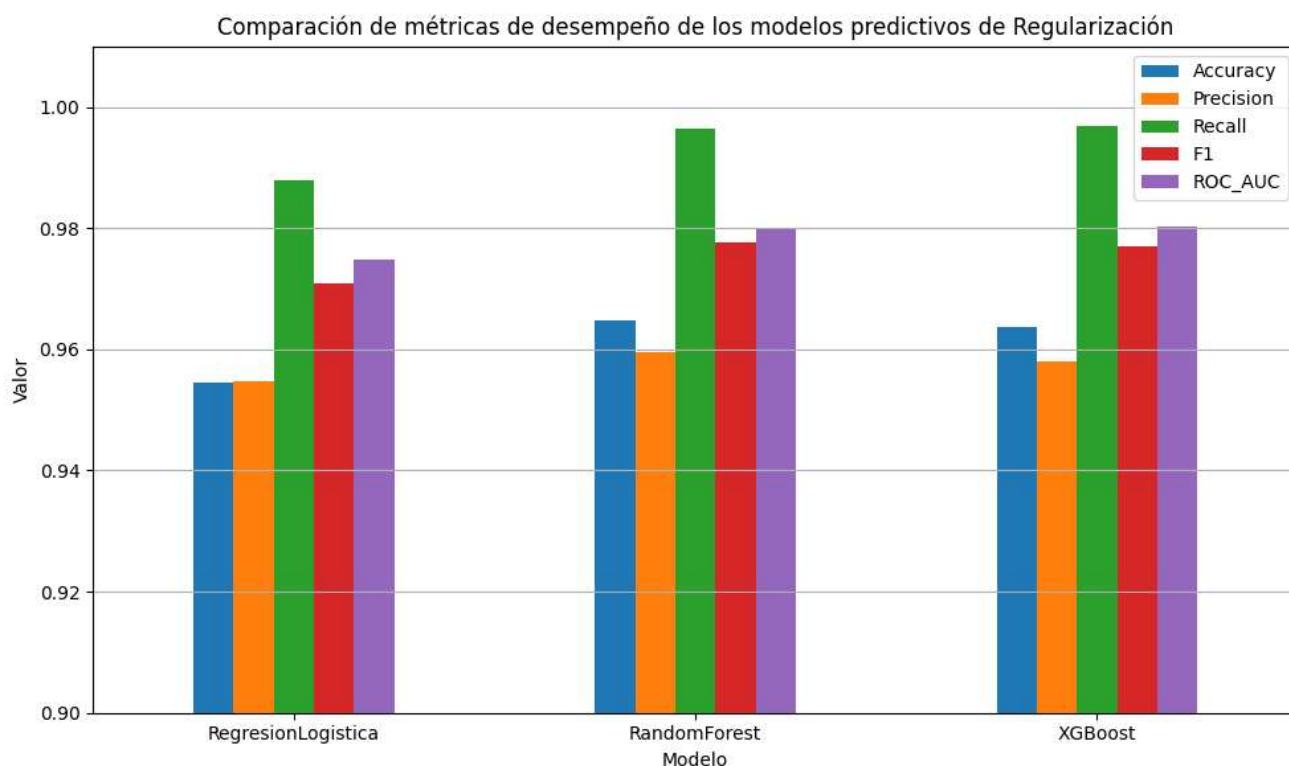
Considerando el comportamiento conjunto de las métricas de evaluación y el equilibrio alcanzado entre capacidad predictiva y generalización, se seleccionó Random Forest como modelo definitivo para la estimación de la Probabilidad de Regularización, constituyéndose en el tercer componente del Modelo de Priorización de Gestión de Cobranza.

Tabla 21 *Comparación de métricas de desempeño de los modelos predictivos de Regularización*

Modelo	Accuracy	Precision	Recall	F1	ROC_AUC
RegresionLogistica	0.9545	0.9546	0.9879	0.9710	0.974896
RandomForest	0.9648	0.9595	0.9963	0.9776	0.980079
XGBoost	0.9638	0.9579	0.9968	0.9770	0.980375

Nota. Elaboración propia.

Figura 39 Comparación de métricas de desempeño de los modelos predictivos de Regularización



Nota. Elaboración propia.

Como complemento al análisis de desempeño, se evaluó la importancia de las variables del modelo seleccionado. La figura presenta las quince variables con mayor contribución al proceso de clasificación mediante el algoritmo Random Forest.

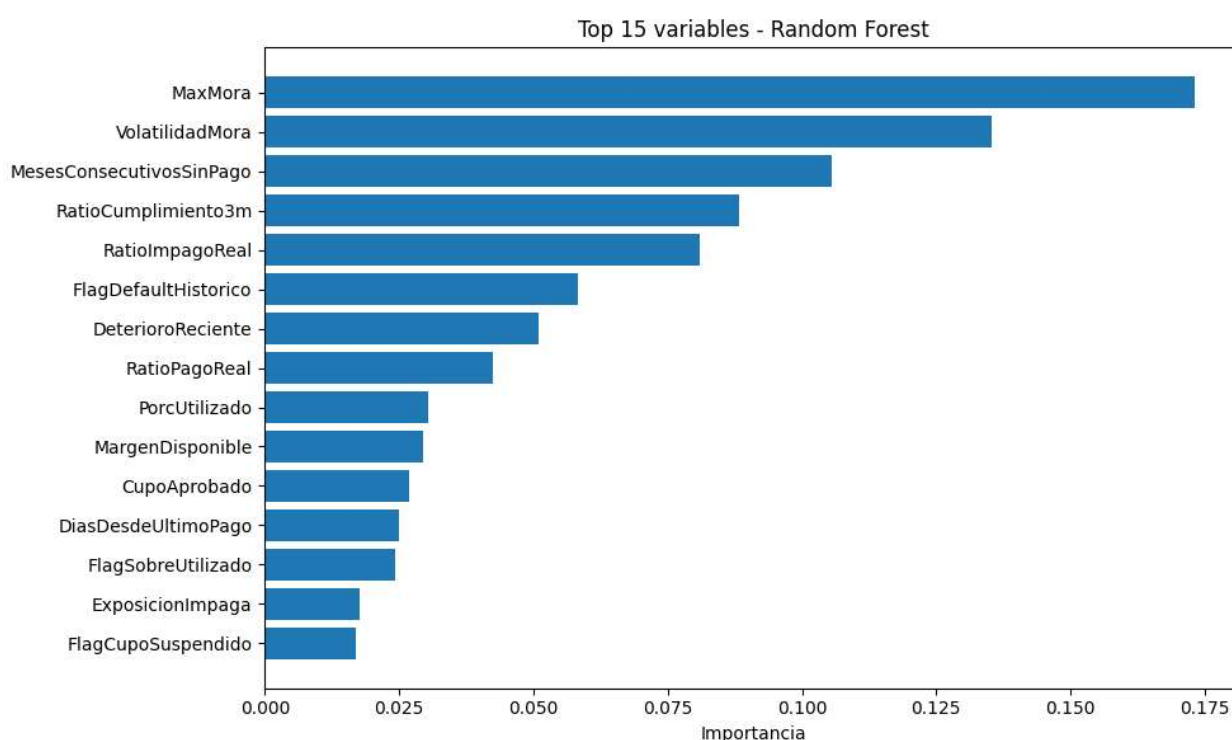
Los resultados muestran que las variables relacionadas con el comportamiento histórico de la mora constituyen los principales factores para estimar la probabilidad de regularización de una cuenta. En particular, MaxMora, VolatilidadMora y MesesConsecutivosSinPago registraron las mayores contribuciones al modelo, evidenciando que la severidad y persistencia del incumplimiento representan los principales determinantes del proceso de regularización.

Asimismo, variables como RatioCumplimiento3m, RatioImpagoReal, FlagDefaultHistorico y DeterioroReciente aportan información complementaria sobre el comportamiento de pago y el historial crediticio del cliente. Finalmente, indicadores relacionados con la utilización del crédito y la capacidad financiera, como PorcUtilizado, MargenDisponible,

CupoAprobado y ExposicionImpaga, fortalecen la capacidad predictiva del modelo al incorporar información relevante sobre la situación económica de cada cuenta.

La distribución de importancia de las variables confirma que el modelo incorpora información complementaria proveniente del historial de pagos, el nivel de mora y la exposición económica, proporcionando una estimación consistente de la probabilidad de regularización.

Figura 40 Variables con mayor importancia en el Modelo Predictivo de Regularización mediante Random Forest



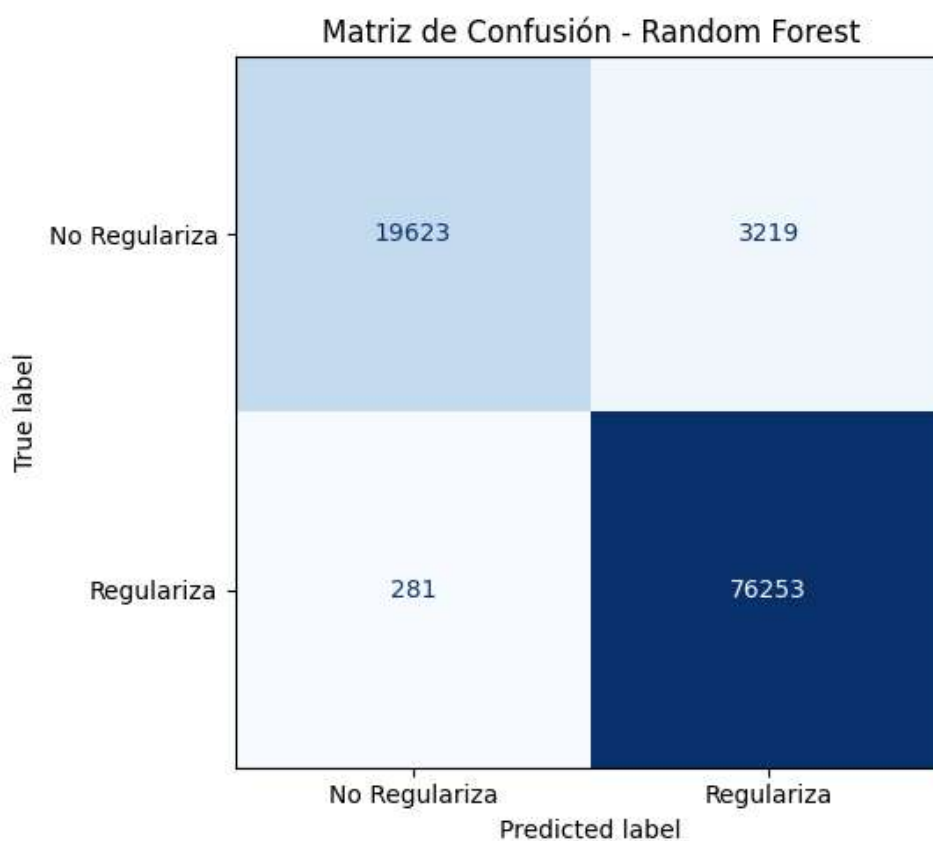
Nota. Elaboración propia.

La capacidad de clasificación del modelo seleccionado fue evaluada adicionalmente mediante la matriz de confusión.

El algoritmo clasificó correctamente 19,623 cuentas correspondientes a la clase "No Regularización" y 76,253 cuentas pertenecientes a la clase "Regularización". Asimismo, registró 3,219 falsos positivos y únicamente 281 falsos negativos, resultado que evidencia

una elevada capacidad para identificar correctamente las cuentas con probabilidad de regularizar su situación, manteniendo un equilibrio adecuado entre sensibilidad y precisión. Estos resultados evidencian que el modelo presenta una capacidad muy elevada para identificar las cuentas con mayor probabilidad de regularizar su situación financiera, privilegiando la detección de casos positivos. Este comportamiento es consistente con el elevado valor de Recall obtenido durante la evaluación (0.9801), lo que resulta especialmente conveniente en procesos de cobranza donde resulta preferible identificar la mayor cantidad posible de clientes con potencial de regularización.

Figura 41 *Matriz de confusión del Modelo Predictivo de Regularización mediante Random Forest*



Nota. Elaboración propia.

Con el propósito de validar el comportamiento del modelo en escenarios representativos, se analizaron casos de estudio correspondientes a diferentes niveles de probabilidad de regularización.

El primer caso corresponde a una cuenta clasificada con Baja Oportunidad de Regularización, obteniendo una probabilidad baja. Las características del cliente evidencian un elevado nivel de exposición impaga, múltiples indicadores de deterioro financiero y ausencia de señales recientes de recuperación, condiciones que justifican la baja probabilidad estimada por el modelo.

El segundo caso representa una cuenta clasificada con Oportunidad Media, cuya probabilidad de regularización se ubicó alrededor del punto de decisión. En este escenario el modelo identifica un cliente con antecedentes de cumplimiento parcial y características intermedias de riesgo, reflejando una posibilidad moderada de normalizar su comportamiento financiero.

Finalmente, el tercer caso corresponde a una cuenta clasificada con Muy Alta Oportunidad de Regularización, alcanzando una probabilidad cercana a la unidad y un ScoreRegularizacion de 1000 puntos. El cliente presenta un comportamiento financiero favorable, ausencia de exposición impaga y un historial reciente consistente, condiciones que respaldan la elevada probabilidad de regularización estimada por el modelo.

En conjunto, los casos de estudio demuestran que las predicciones generadas por Random Forest mantienen coherencia con las características observadas en los clientes analizados, aportando evidencia adicional sobre la capacidad del modelo para apoyar la definición de estrategias diferenciadas de cobranza.

Resultados del Modelo Predictivo de Contactabilidad

El Modelo Predictivo de Contactabilidad fue evaluado mediante la comparación de tres algoritmos de clasificación supervisada: Regresión Logística, Random Forest y XGBoost. Para valorar su desempeño se utilizaron las métricas ROC-AUC, PR-AUC, Precision, Recall y F1-Score, las cuales permiten medir la capacidad de los modelos para identificar

correctamente aquellas cuentas con mayor probabilidad de establecer un contacto efectivo mediante los canales de gestión disponibles.

Antes de analizar el desempeño predictivo, se revisó la distribución de la variable objetivo TargetContactabilidad. Esta variable fue construida como un indicador binario que diferencia entre cuentas con contacto efectivo y cuentas sin contacto efectivo dentro del horizonte de análisis. Los resultados muestran un desbalance moderado: el 71.13% de los registros corresponde a cuentas no contactadas y el 28.87 % a cuentas contactadas exitosamente.

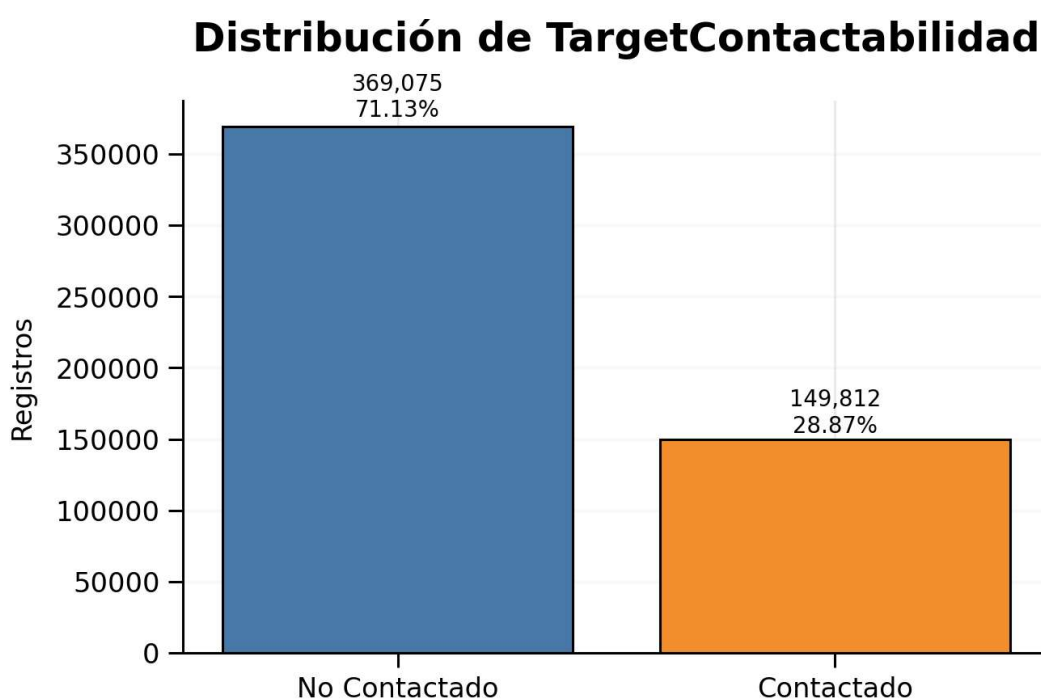
Este comportamiento justifica el uso de métricas complementarias al accuracy, especialmente PR-AUC y Recall, dado que el evento de interés corresponde a la clase contactable y no representa la mayoría de las observaciones.

Tabla 22 *Distribución de la variable TargetContactabilidad*

TargetContactabilidad	Cantidad	Porcentaje
No Contactado	369.075	71,13 %
Contactado	149.812	28,87 %

Nota. Elaboración propia.

Figura 42 *Distribución del TargetContactabilidad utilizado en la evaluación del modelo*



Nota. Elaboración propia.

Los resultados evidencian que los tres algoritmos alcanzaron niveles satisfactorios para la estimación de la contactabilidad; sin embargo, XGBoost presentó el mejor desempeño global, alcanzando un ROC-AUC de 0.912128 y un PR-AUC de 0.791855, lo que evidencia una mayor capacidad discriminativa en comparación con Regresión Logística y Random Forest. Asimismo, obtuvo el Recall más alto, con 0.892100, y el mayor F1-Score, con 0.730667, reflejando un mejor equilibrio entre sensibilidad y precisión para identificar cuentas efectivamente contactables.

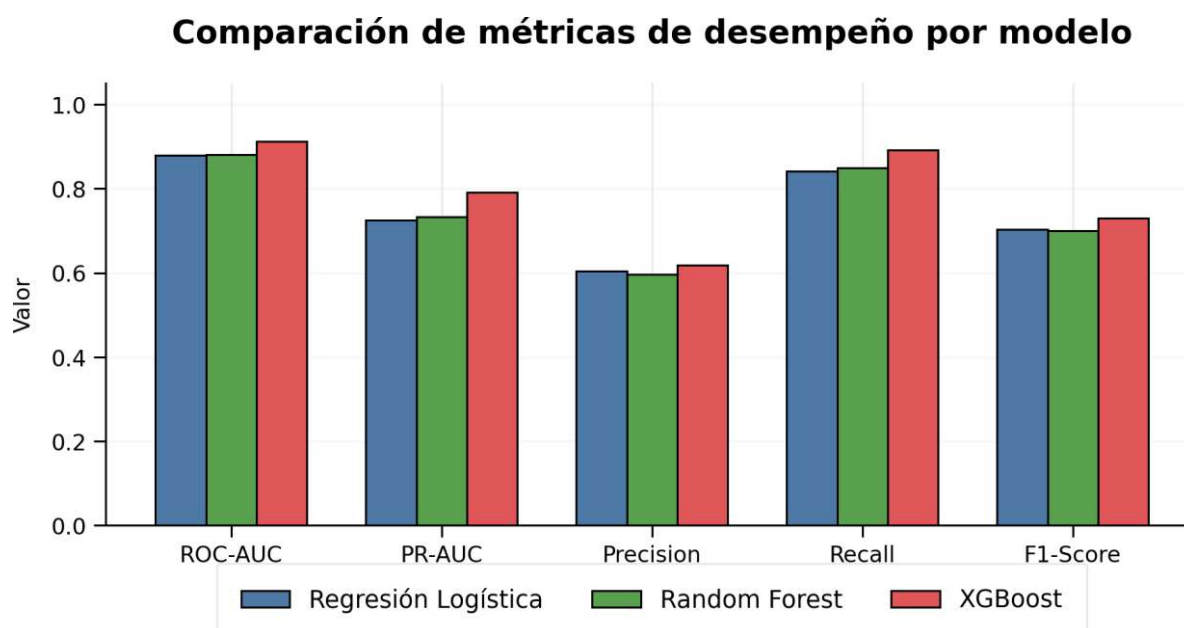
Considerando el comportamiento conjunto de las métricas de evaluación, se seleccionó XGBoost como modelo definitivo para la estimación de la Probabilidad de Contactabilidad. Esta salida constituye un insumo relevante para la posterior segmentación operativa y para la formulación de estrategias diferenciadas de cobranza.

Tabla 23 Comparación de métricas de desempeño de los modelos predictivos de Contactabilidad

Modelo	ROC-AUC	PR-AUC	Precision	Recall	F1-Score
Regresión Logística	0.879271	0.725501	0.603690	0.841838	0.703147
Random Forest	0.880419	0.732701	0.596012	0.849948	0.700682
XGBoost	0.912128	0.791855	0.618707	0.892100	0.730667

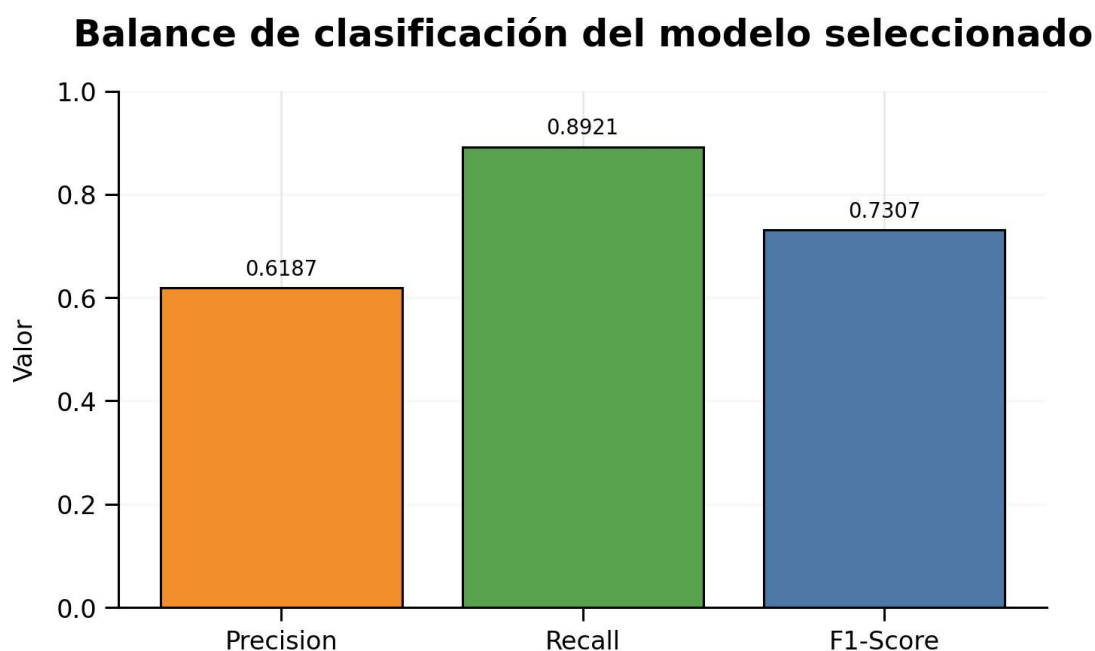
Nota. Elaboración propia

Figura 43 Comparación de métricas de desempeño de los modelos predictivos de Contactabilidad



Nota. Elaboración propia

Figura 44 Balance de clasificación del modelo seleccionado XGBoost



Nota. Elaboración propia.

Como complemento al análisis de desempeño, se evaluó la importancia de las variables del modelo seleccionado. Los resultados muestran que la variable **Muestra** constituye el principal factor explicativo del modelo, evidenciando la existencia de patrones temporales relevantes en el comportamiento de contacto de los clientes. Asimismo, variables asociadas a la antigüedad de la gestión, historial de comunicaciones e interacción omnicanal registran aportes significativos para la predicción de la contactabilidad.

Entre las variables de mayor relevancia se encuentran MesesVencidosProm, DiasUltimoEmail, NumCampaniasPreventivas, NumMesesConEmail, AsignacionProm, Omnicanal, NumTelefoniaExitosa y RatioPreventivo. En conjunto, estas variables reflejan tanto la intensidad histórica de gestión como la capacidad de interacción previa del cliente con los canales disponibles.

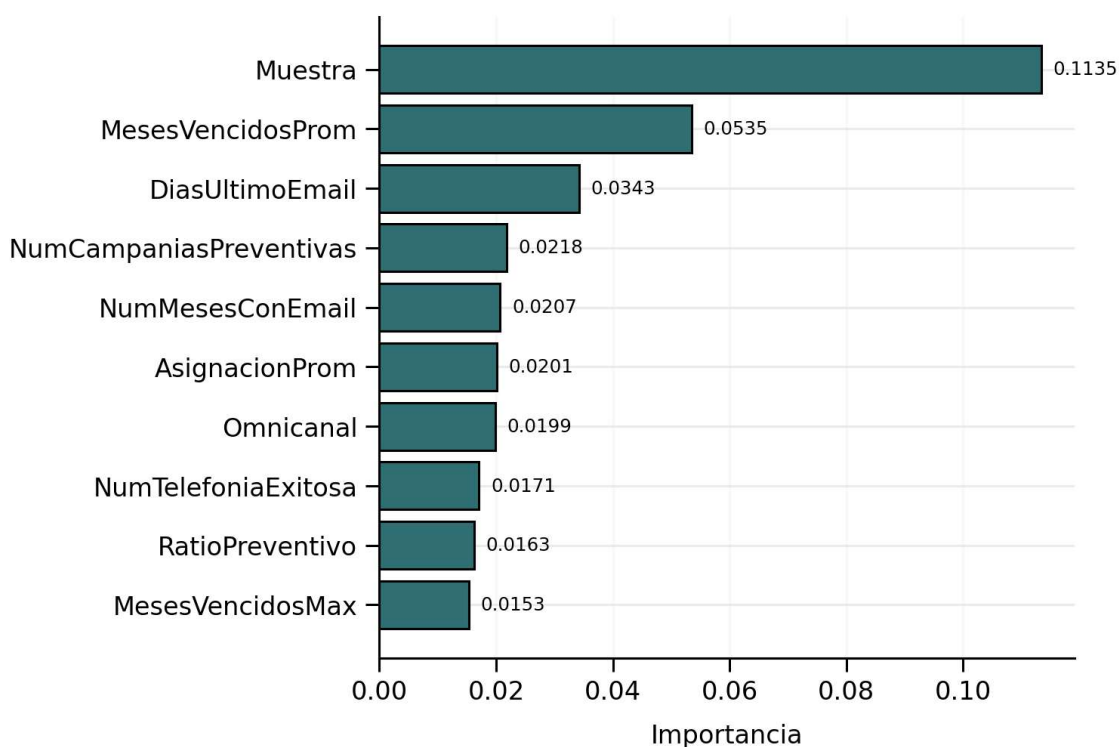
Tabla 24 Variables con mayor importancia en el modelo XGBoost de Contactabilidad

Variable	Importancia
Muestra	0.113453
MesesVencidosProm	0.053477
DiasUltimoEmail	0.034262
NumCampaniasPreventivas	0.021840
NumMesesConEmail	0.020711
AsignacionProm	0.020086
Omnicanal	0.019860
NumTelefoniaExitosa	0.017090
RatioPreventivo	0.016255
MesesVencidosMax	0.015289

Nota. Elaboración propia

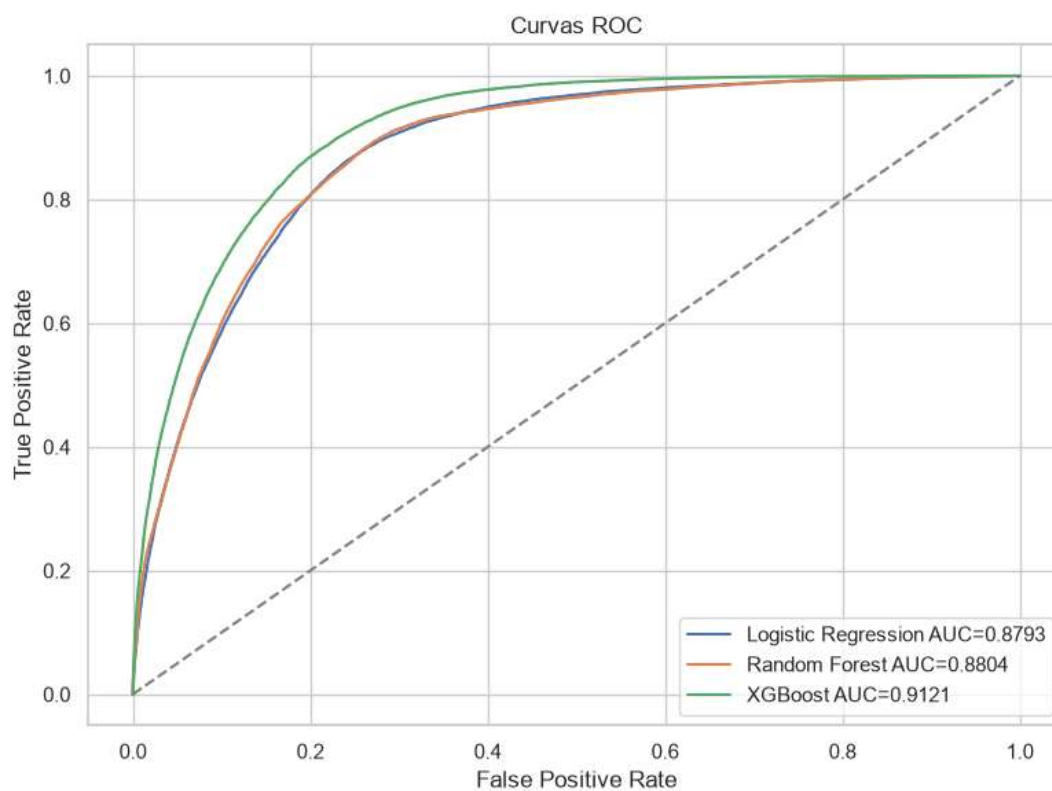
Figura 45 Variables con mayor importancia en el Modelo Predictivo de Contactabilidad mediante XGBoost

Importancia de variables del modelo XGBoost

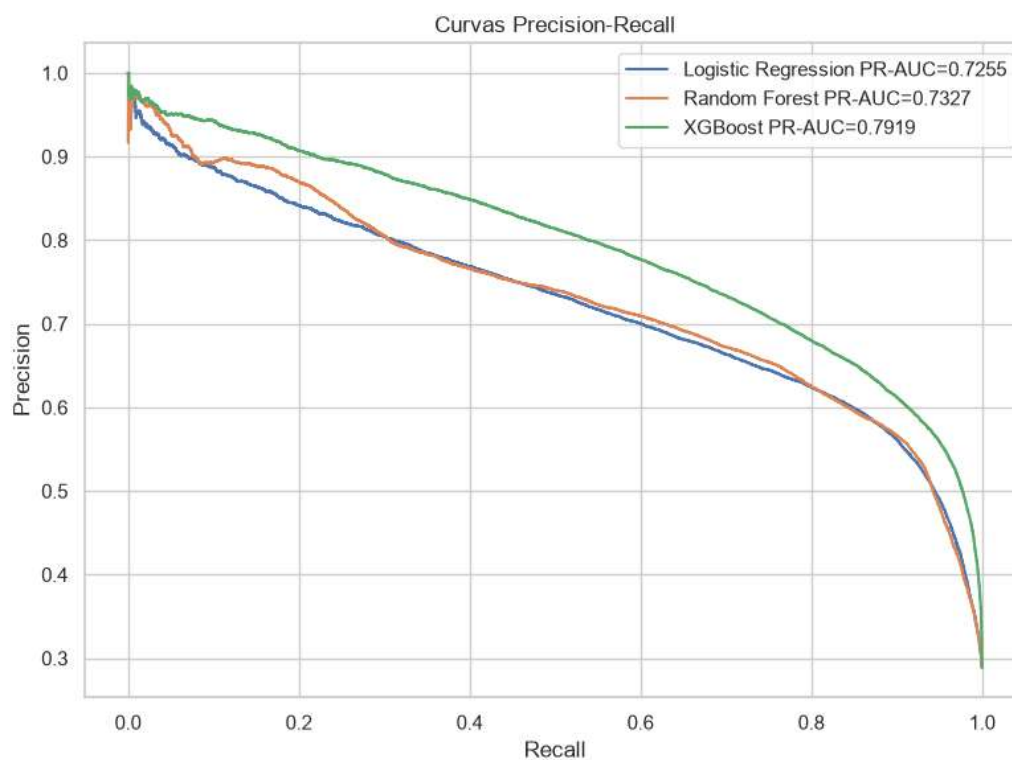


Nota. Elaboración propia.

La capacidad de clasificación del modelo seleccionado fue evaluada adicionalmente mediante la matriz de confusión. Esta matriz permite observar el comportamiento del modelo respecto a las clases No Contactado y Contactado, considerando un umbral de clasificación de 0.5.

Figura 46 *Curvas ROC de los modelos evaluados*

Nota. Elaboración propia.

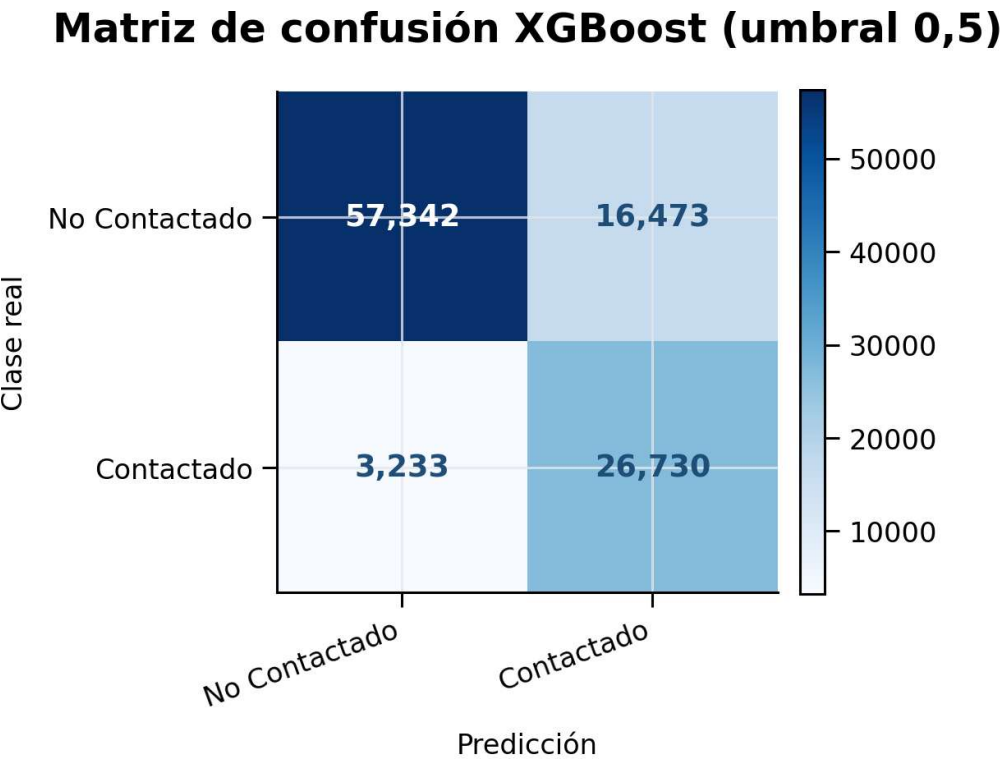
Figura 47 *Curvas Precision-Recall de los modelos evaluados*

Nota. Elaboración propia.

El elevado valor de Recall obtenido por XGBoost indica que el modelo identifica correctamente una proporción importante de cuentas efectivamente contactables. Este comportamiento resulta especialmente conveniente para procesos de cobranza, donde excluir cuentas con potencial de contacto podría reducir la efectividad de las campañas de gestión.

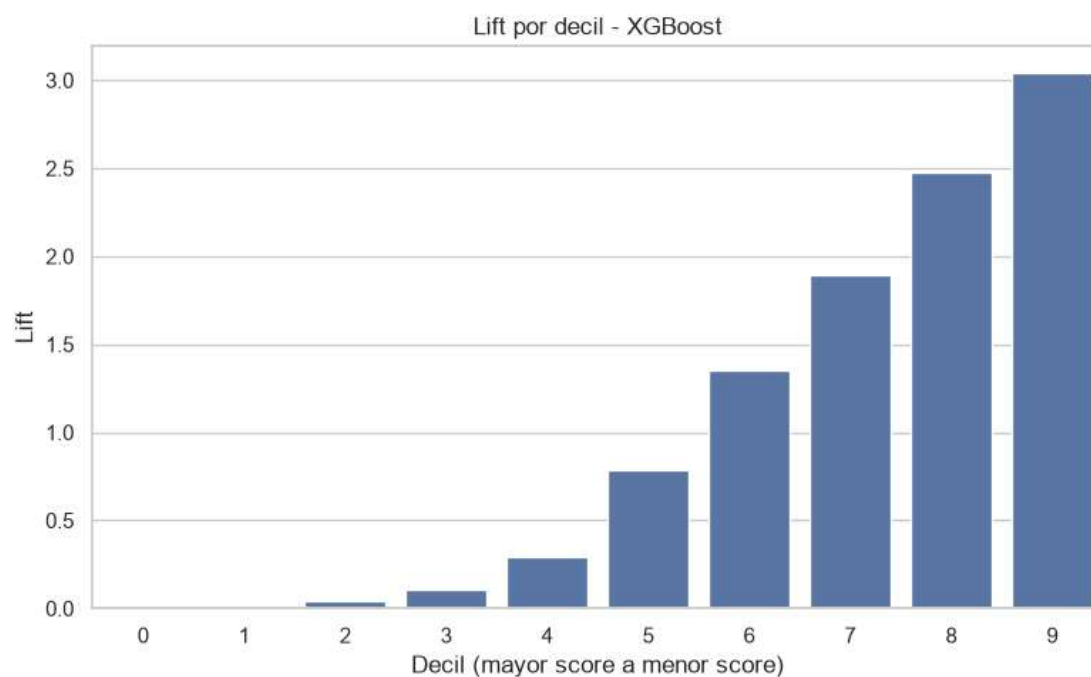
La matriz de confusión incluida se calculó a partir del tamaño del conjunto de prueba y de las métricas reportadas por el notebook. Por ello, los conteos deben interpretarse como una reconstrucción aproximada de apoyo visual; la matriz exacta debe sustituirse por la exportación directa del notebook si se requiere trazabilidad fila a fila.

Figura 48 *Matriz de confusión del Modelo Predictivo de Contactabilidad mediante XGBoost*

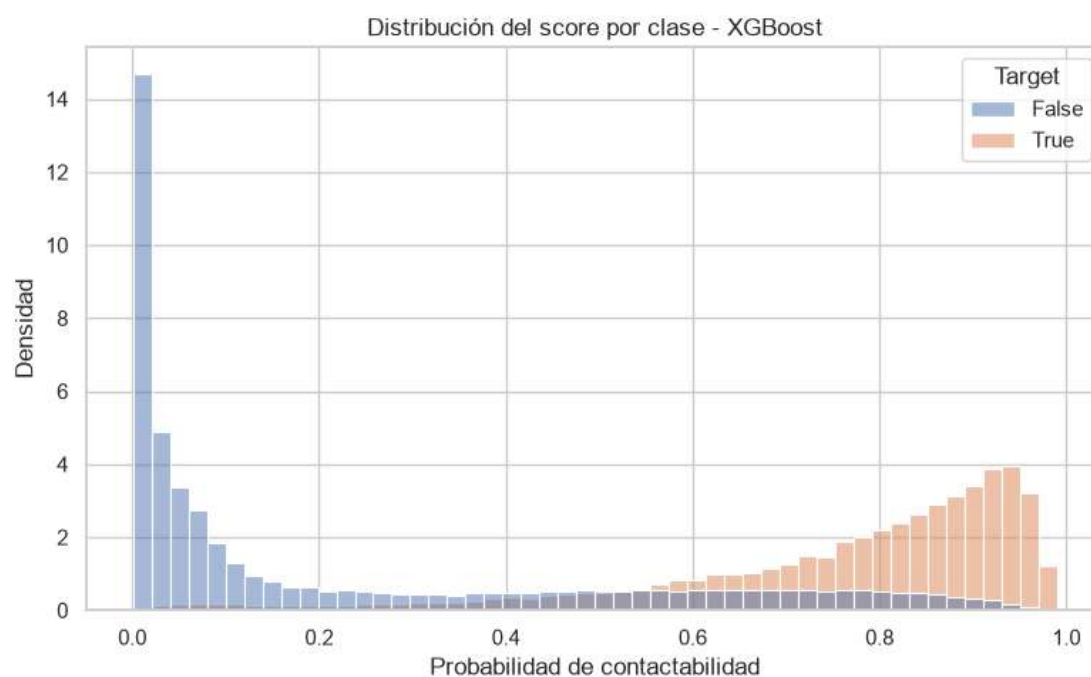


Nota. Elaboración propia.

Con el propósito de transformar la probabilidad estimada en una variable operativa, el modelo generó una segmentación del nivel de contactabilidad. Esta segmentación permite traducir la probabilidad continua en categorías de negocio fácilmente interpretables para la definición de canales y frecuencia de gestión.

Figura 49 *Lift por decil del modelo XGBoost*

Nota. Elaboración propia.

Figura 50 *Distribución del score de contactabilidad por clase real*

Nota. Elaboración propia.

Los clientes con probabilidad Muy Alta requieren menor esfuerzo de contacto, mientras que las cuentas con probabilidad Muy Baja requieren mayor número de iteraciones

recomendadas. Esta lógica permite asignar esfuerzos de gestión de forma proporcional a la probabilidad de contacto efectivo.

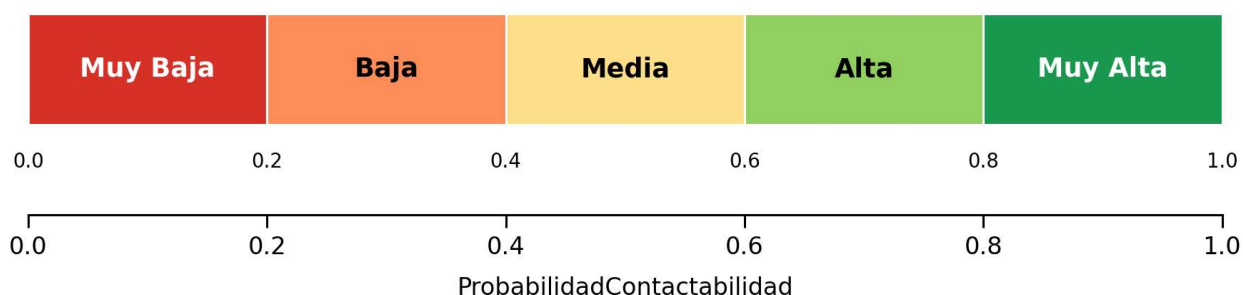
Tabla 25 Segmentación operativa de la ProbabilidadContactabilidad

Segmento	Rango de probabilidad	Interpretación operativa	Iteraciones
Muy Baja	0.00 a 0.20	Mayor esfuerzo de contacto	5
Baja	0.20 a 0.40	Gestión reforzada	4
Media	0.40 a 0.60	Gestión estándar	3
Alta	0.60 a 0.80	Gestión focalizada	2
Muy Alta	0.80 a 1.00	Contacto priorizado de bajo esfuerzo	1

Nota. Elaboración propia.

Figura 51 Segmentación operativa del ScoreContactabilidad

Segmentación operativa del ScoreContactabilidad



Nota. Elaboración propia.

En conjunto, los resultados demuestran que XGBoost proporciona una estimación robusta y consistente de la probabilidad de contacto efectivo. La combinación de ROC-AUC, PR-AUC, Recall y F1-Score evidenció un desempeño superior frente a las alternativas evaluadas, lo que sustentó su selección como modelo definitivo.

La incorporación del Modelo Predictivo de Contactabilidad fortalece la solución integral de cobranza, ya que permite complementar los modelos financieros de Mora, Pago y Regularización con una dimensión operativa asociada a la factibilidad de interacción con el cliente. De esta manera, el modelo no solo contribuye a priorizar cuentas, sino también a orientar la intensidad y el canal de gestión más adecuado para cada segmento.

Resultados del Modelo de Priorización de Gestión de Cobranza

Una vez validados los modelos predictivos de Mora, Pago y Regularización, se procedió a integrar sus resultados en el Modelo de Priorización de Gestión de Cobranza. El propósito de este modelo fue consolidar la información predictiva generada durante la investigación en un único esquema de apoyo a la toma de decisiones, permitiendo establecer prioridades objetivas para la gestión de cobranza.

A diferencia de los modelos predictivos individuales, cuyo objetivo consistió en estimar eventos específicos del comportamiento de los clientes, el Modelo de Priorización integra simultáneamente la probabilidad de mora, el porcentaje esperado de pago, la probabilidad de regularización y la exposición económica de cada cuenta. Como resultado, se obtuvo un indicador compuesto capaz de ordenar la cartera según su prioridad operativa, facilitando la asignación eficiente de recursos de cobranza.

En las siguientes secciones se analiza el comportamiento del ScorePrioridadGestion, la segmentación de la cartera mediante deciles y la distribución de la exposición económica.

Resultados del Score de Priorización de Gestión

El ScorePrioridadGestion fue construido integrando las probabilidades generadas por los modelos predictivos de Mora, Pago y Regularización, junto con la exposición económica normalizada de cada cuenta. Como resultado, se obtuvo un indicador compuesto orientado a establecer el nivel de prioridad de gestión para la totalidad de la cartera analizada.

Posteriormente, previo a la segmentación de la cartera, se consolidó un único registro por cuenta correspondiente a la última observación disponible, obteniéndose un conjunto final de **108,910 cuentas** utilizado para la generación de los deciles, el tipo de gestión y la implementación operativa del modelo.

Se presenta el resumen estadístico de las principales variables utilizadas para la construcción del score, así como del ScorePrioridadGestion y su versión normalizada en una escala de 0 a 1000 puntos.

Los resultados muestran que la Probabilidad de Regularización presentó el mayor promedio (0.7922), seguida por el Pago Esperado (0.7191), mientras que la Probabilidad de Mora registró un promedio de 0.2242. Este comportamiento evidencia que una proporción importante de la cartera mantiene una alta probabilidad de regularización y un elevado potencial de pago, aunque persiste un subconjunto de cuentas con mayor riesgo de incumplimiento que requiere una gestión diferenciada.

Por otra parte, la variable Exposición Normalizada presentó un promedio de 0.0207, confirmando que la mayor parte de las cuentas registra niveles reducidos de exposición económica respecto al valor máximo observado en la cartera. Este resultado justifica la aplicación del proceso de normalización previo a su integración en el cálculo del score.

Como consecuencia de la combinación ponderada de estas variables, el ScorePrioridadGestion alcanzó un valor promedio de 0.5722, con valores comprendidos entre 0.1074 y 0.8276. Posteriormente, el indicador fue transformado a una escala de 0 a 1000 puntos, obteniéndose un ScorePrioridadGestion1000 con un promedio de 691 puntos, facilitando su interpretación e implementación dentro del proceso de priorización de la gestión de cobranza.

La distribución obtenida evidencia que el modelo genera una adecuada diferenciación entre las cuentas de la cartera, proporcionando un indicador continuo que constituye la base para la segmentación mediante deciles y la definición de estrategias diferenciadas de gestión analizadas en los apartados siguientes.

Tabla 26 *Resumen estadístico de las variables utilizadas en la construcción del Score de Priorización de Gestión*

	count	mean	std	Min	25%	50%	75%	max
Probabilidad Mora	573,328	0.22	0.37	0.0000	0.0034	0.0189	0.1854	1.00
Pago Esperado	573,328	0.72	0.37	0.0000	0.6024	0.9302	0.9886	1.00
Probabilidad de Regularización	573,328	0.79	0.37	0.0000	0.8763	0.9837	0.9977	1.00
Exposición Impaga	573,328	58.03	130.50	0.00	0.00	0.00	58.59	2,803.03
Exposición Normalizada	573,328	0.02	0.05	0.00	0.00	0.00	0.02	1.00
Score de Prioridad de Gestión	573,328	0.57	0.18	0.1074	0.5440	0.6729	0.6959	0.83
Score de Prioridad de Gestión (0 – 1000)	573,328	691.38	218.48	130	657	813	841	1,000

Nota. Elaboración propia.

Segmentación de la cartera mediante deciles de gestión

Una vez calculado el Score de Prioridad de Gestión, la cartera fue segmentada en diez deciles con igual número de cuentas, permitiendo establecer diferentes niveles de prioridad para la gestión de cobranza. En la tabla se presenta el resumen estadístico obtenido para cada uno de los deciles generados por el modelo.

Los resultados muestran que cada decil quedó conformado por 10,891 cuentas, garantizando una distribución homogénea de la cartera y facilitando la comparación entre segmentos.

Esta segmentación permitió analizar el comportamiento conjunto de las principales variables predictivas y financieras consideradas por el modelo.

Se observa que los deciles 1 y 2 concentran la mayor exposición económica de la cartera, registrando además los mayores valores promedio de Probabilidad de Mora y los menores valores de Pago Esperado y Probabilidad de Regularización. A partir del decil 3 la exposición económica disminuye progresivamente, acompañando la reducción del riesgo de mora y el incremento del potencial de pago y de regularización. Este comportamiento evidencia que el modelo asigna las mayores prioridades a aquellas cuentas que combinan un mayor riesgo financiero con un mayor impacto económico para la organización.

En los deciles superiores la exposición económica continúa reduciéndose hasta alcanzar niveles mínimos, mientras que la Probabilidad de Mora disminuye progresivamente y tanto el Pago Esperado como la Probabilidad de Regularización presentan una tendencia creciente. En consecuencia, los deciles de menor prioridad concentran clientes con mejores perspectivas de pago, mayor probabilidad de regularización y menor riesgo de incumplimiento, características que justifican estrategias preventivas de menor intensidad y costo operativo.

En conjunto, la segmentación obtenida demuestra que el Score de Prioridad de Gestión logra ordenar la cartera de forma consistente, diferenciando claramente los distintos niveles de riesgo y oportunidad identificados por los modelos predictivos desarrollados durante la investigación. Esta capacidad de discriminación constituye la base para la asignación eficiente de recursos y la definición de estrategias diferenciadas de cobranza.

Tabla 27 Segmentación de la cartera mediante deciles de gestión

Decil Gestión	Cuentas	Score Promedio	Exposición Total	Exposición Promedio	ProbMora Promedio	PagoEsperado Promedio	ProbRegularizacion Promedio
10	10,891	0.7041	93,879.88	8.62	0.0470	0.9919	0.9921
9	10,891	0.6985	4,063.05	0.37	0.0069	0.9951	0.9969
8	10,891	0.6953	13,225.52	1.21	0.0114	0.9873	0.9937
7	10,891	0.6898	29,975.29	2.75	0.0173	0.9741	0.9888
6	10,891	0.6801	95,549.90	8.77	0.0279	0.9507	0.9799
5	10,891	0.6607	201,699.58	18.52	0.0477	0.9078	0.9580
4	10,891	0.6115	218,987.68	20.11	0.0540	0.7919	0.9442
3	10,891	0.5381	459,952.66	42.23	0.0958	0.6088	0.9129
2	10,891	0.2619	3,021,653.41	277.44	0.7479	0.0876	0.2247
1	10,891	0.2050	1,518,956.32	139.47	0.9929	0.0009	0.0036

Nota. Elaboración propia.

Análisis de la exposición y priorización

Con el propósito de evaluar la capacidad del modelo para priorizar la gestión de cobranza, se analizó la distribución de la exposición económica, el comportamiento de las variables predictivas y la segmentación de la cartera según el tipo de gestión propuesto.

La tabla a continuación presenta un resumen de la composición de la cartera de acuerdo con la existencia de exposición económica pendiente. Los resultados muestran que el conjunto final estuvo conformado por 108,910 cuentas, de las cuales 31,196 (28.64%) registraron exposición impaga y fueron clasificadas como candidatas para una gestión recuperativa. Las 77,714 cuentas restantes (71.36%) no presentaron exposición económica pendiente, por lo que fueron clasificadas para estrategias de gestión preventiva.

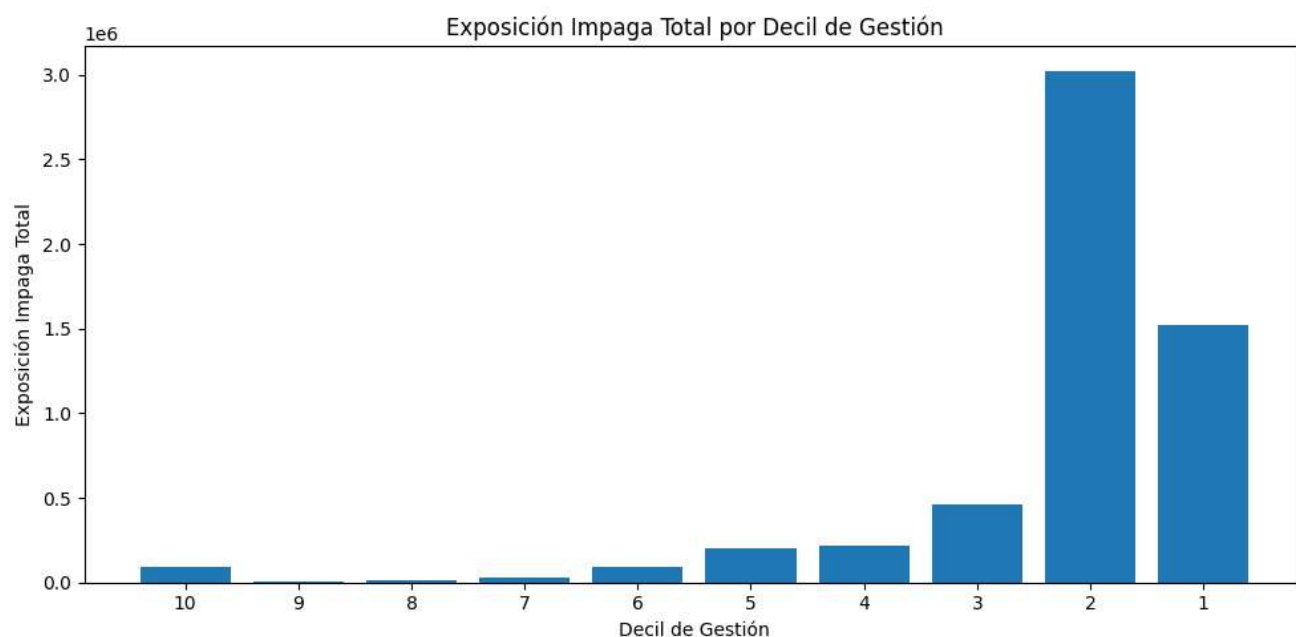
Tabla 28 *Composición de la cartera según tipo de gestión*

Indicador	Valor
Total Cuentas	108,910
Cuentas con Exposicion > 0	31,196
Cuentas con Exposicion = 0	77,714
% Exposicion > 0	28.64
% Exposicion = 0	71.36

Nota. Elaboración propia.

Estos resultados evidencian que el 28.64% de la cartera requiere acciones orientadas a la recuperación de deuda, mientras que el 71.36% restante corresponde a clientes cuyo seguimiento tiene un carácter preventivo, orientado a evitar el deterioro futuro del comportamiento de pago.

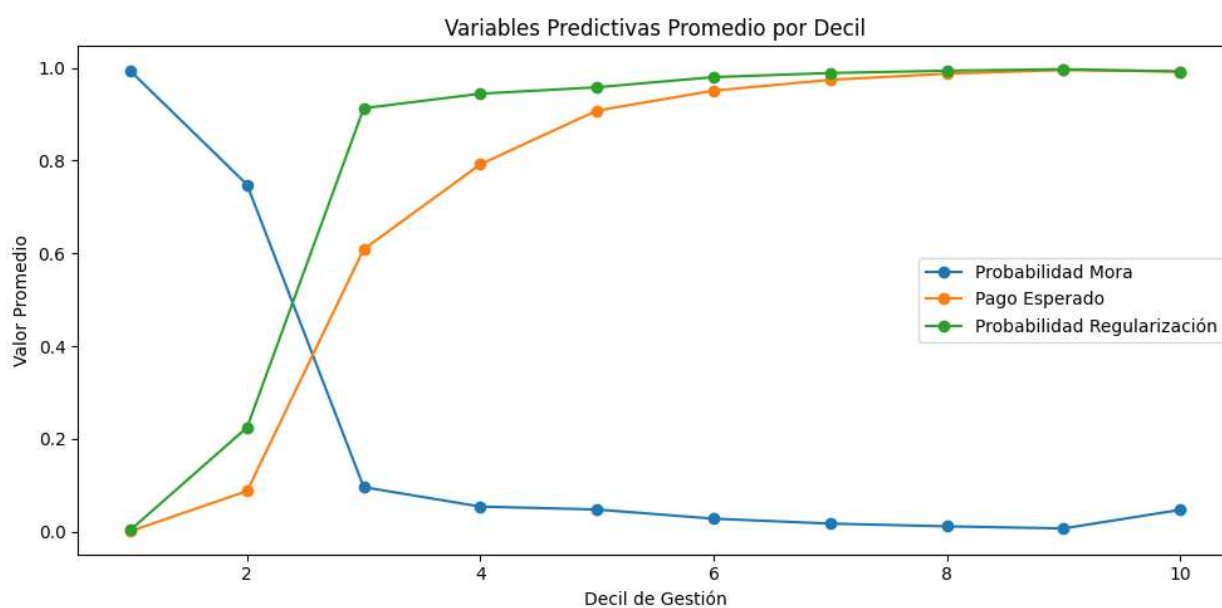
La figura presenta la distribución de la exposición impaga total por decil de gestión. Se observa que la mayor concentración de exposición económica se localiza en los deciles 1 y 2, los cuales reúnen la mayor parte del monto pendiente de recuperación. A partir del decil 3 la exposición disminuye progresivamente, alcanzando valores reducidos en los deciles superiores. Este comportamiento confirma que el modelo asigna las mayores prioridades a las cuentas con mayor impacto financiero para la organización, sin perder la capacidad de identificar casos recuperativos de menor exposición distribuidos en otros segmentos de la cartera.

Figura 52 *Exposición impaga total por decil de gestión*

Nota. Elaboración propia.

La figura muestra la evolución de las variables predictivas promedio a través de los deciles de gestión. Se observa una disminución progresiva de la Probabilidad de Mora, acompañada de un incremento sostenido tanto del Pago Esperado como de la Probabilidad de Regularización. Esta tendencia confirma que el modelo ordena la cartera de forma consistente, diferenciando gradualmente las cuentas de mayor riesgo de aquellas con mejores perspectivas de recuperación y cumplimiento.

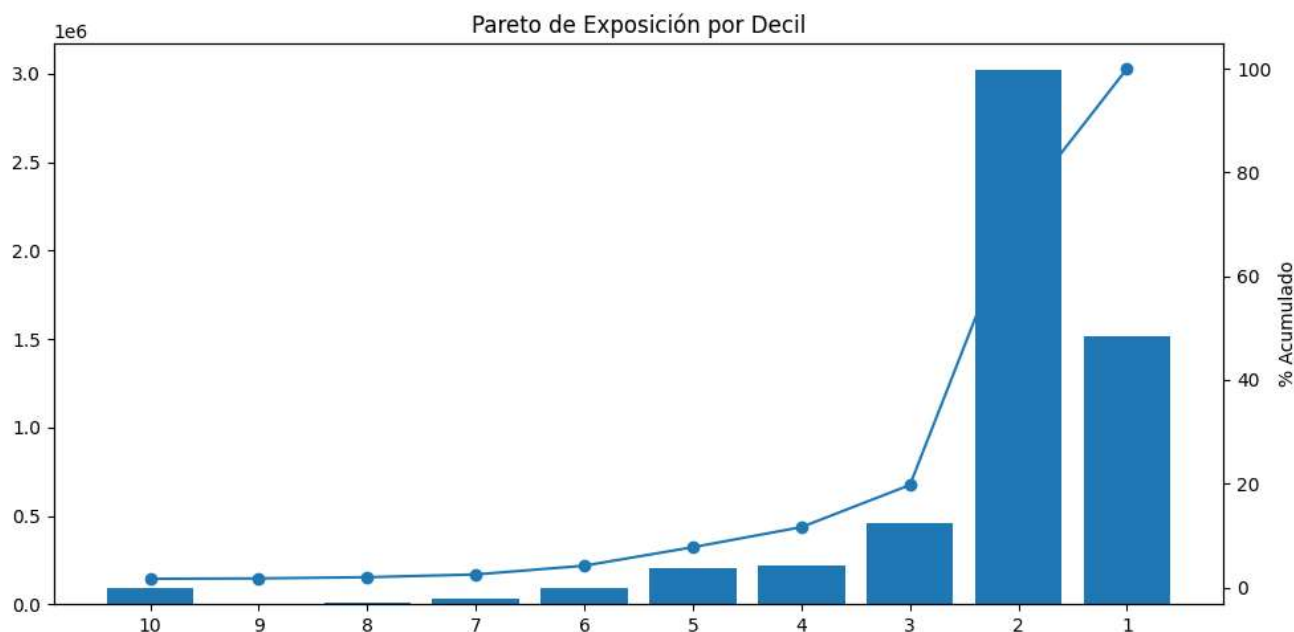
Figura 53 Comportamiento promedio de las variables predictivas por decil de gestión



Nota. Elaboración propia.

La concentración de la exposición económica puede apreciarse con mayor claridad en la figura correspondiente al análisis de Pareto. Los resultados muestran que la mayor parte de la exposición impaga se concentra en los primeros deciles de gestión, especialmente en los deciles 1 y 2, evidenciando que el modelo focaliza los esfuerzos de cobranza sobre el segmento de cuentas con mayor impacto económico, mientras que los deciles restantes representan una proporción significativamente menor de la exposición total.

Figura 54 Concentración acumulada de la exposición económica por decil de gestión (Análisis de Pareto)

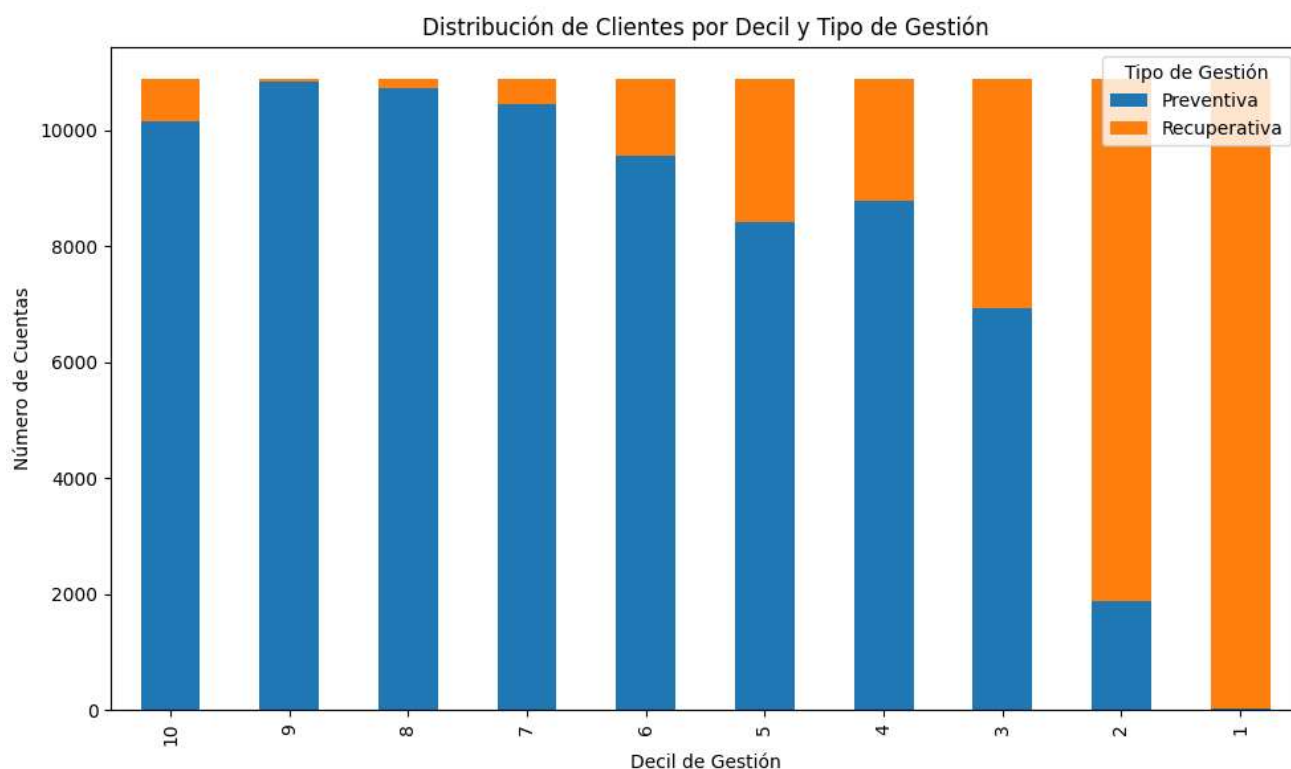


Nota. Elaboración propia.

Finalmente, la figura presenta la distribución de clientes según el tipo de gestión propuesto.

Se observa que el decil 1 está conformado casi en su totalidad por cuentas de gestión recuperativa, mientras que en el decil 2 continúa predominando este tipo de gestión, aunque con una mayor participación de cuentas preventivas. A partir del decil 3 se evidencia una transición gradual hacia estrategias preventivas, incrementándose progresivamente su participación en los deciles superiores. Este comportamiento refleja que el modelo no establece una separación estricta entre ambos tipos de gestión, sino una priorización continua que incorpora simultáneamente el riesgo, el potencial de recuperación y la exposición económica de cada cuenta.

Figura 55 Distribución de clientes por decil según el tipo de gestión propuesto



Nota. Elaboración propia.

En conjunto, estos resultados demuestran que el Modelo de Priorización de Gestión de Cobranza no solo ordena las cuentas según un indicador cuantitativo, sino que también establece una segmentación operativamente consistente, concentrando las mayores prioridades sobre las cuentas con mayor riesgo y exposición económica, sin dejar de identificar oportunidades de gestión recuperativa en otros segmentos de la cartera. Esta priorización permite asignar los recursos de cobranza de manera más eficiente y definir los tipos de gestión preventivas y recuperativas acordes con las características de cada grupo de clientes.

Integración de Contactabilidad y Priorización de Gestión

La integración entre el Modelo de Priorización y el Modelo de Contactabilidad permitió construir un espacio de decisión más completo para la asignación de estrategias de cobranza. Este espacio combina tres dimensiones principales: TipoGestion, GrupoValor y

NivelContactabilidad. La primera dimensión distingue entre cuentas preventivas y recuperativas, la segunda resume el valor relativo de la cuenta a partir del DecilGestion y la tercera representa la factibilidad de contacto estimada por el modelo.

Cada combinación de estas variables se traduce en una EstrategiaRecomendada. Esta lógica convierte la priorización analítica en una recomendación accionable, permitiendo diferenciar entre acciones preventivas, digitales, de recuperación de alto valor, recuperación de bajo costo, monitoreo y casos donde no se recomienda gestionar activamente.

El conjunto analizado contiene 108.910 clientes agrupados en 36 combinaciones de TipoGestion, GrupoValor y NivelContactabilidad. Esta integración permite observar tanto la concentración de la cartera por segmento como la distribución final de estrategias recomendadas.

En la implementación definitiva, el GrupoValor se construyó transformando el DecilGestion en tres categorías operativas: Alto para deciles 1 y 2, Medio para deciles 3 a 6 y Bajo para deciles 7 a 10. Esta agrupación corresponde a la codificación usada por el proceso y reduce la complejidad de la matriz de decisión, facilitando la generación de reglas de negocio aplicables en ambientes operativos.

Tabla 29 *Agrupación de DecilGestion en grupos de valor*

GrupoValor	Regla de agrupación	Interpretación
Alto	Deciles 1 y 2	Cuentas clasificadas por el notebook como segmentos de mayor valor relativo para reglas de alto valor.
Medio	Deciles 3, 4, 5 y 6	Cuentas con balance intermedio entre valor, riesgo y oportunidad.
Bajo	Deciles 7, 8, 9 y 10	Cuentas orientadas principalmente a monitoreo, gestión digital o costo controlado según contactabilidad.

Nota. Elaboración propia.

Análisis de segmentos de gestión

El análisis de segmentos muestra que los mayores volúmenes de clientes se concentran principalmente en combinaciones preventivas con niveles de contactabilidad altos y muy altos. En particular, los segmentos Preventiva | Bajo | Muy Alta (20,551 clientes) y Preventiva | Medio | Muy Alta (15,233 clientes) representan las mayores concentraciones de la cartera analizada. Estos resultados evidencian una oportunidad operativa relevante, ya que una proporción importante de cuentas presenta elevada probabilidad de contacto y puede ser gestionada mediante estrategias preventivas, digitales o de monitoreo antes de que se materialice un deterioro financiero.

Asimismo, destaca el segmento Recuperativa | Alto | Muy Baja, que concentra 12,167 clientes y registra una media de cinco iteraciones recomendadas. Este comportamiento sugiere la existencia de un grupo relevante de cuentas recuperativas con baja probabilidad de contacto, que requieren estrategias diferenciadas orientadas a balancear recuperación potencial y eficiencia operativa.

Los cuatro segmentos de mayor volumen acumulan 59,106 clientes, equivalentes al 54.3 % de los registros analizados. Esta concentración evidencia que una porción importante de la cartera puede representarse mediante un número reducido de combinaciones de gestión y contactabilidad.

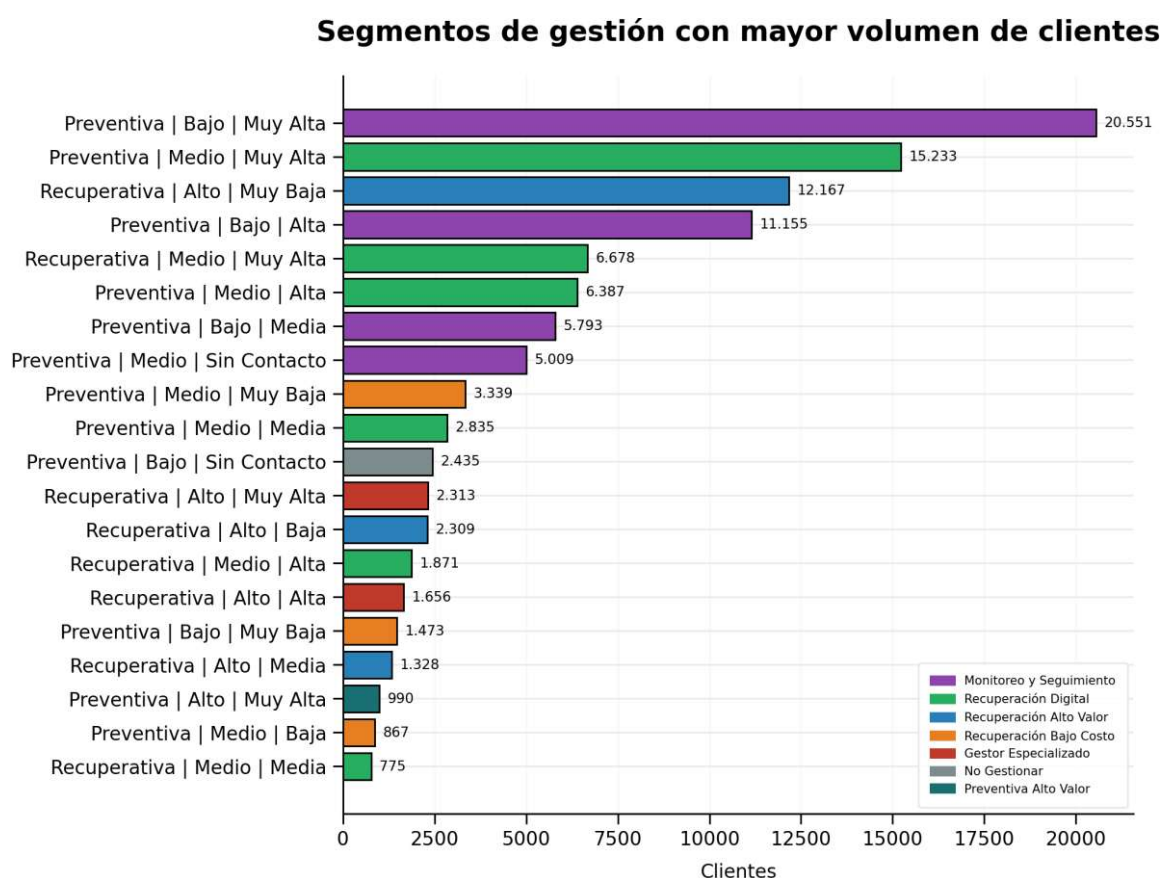
Tabla 30 *Segmentos definitivos de gestión según TipoGestion, GrupoValor y NivelContactabilidad*

Tipo Gestion	Grupo Valor	Nivel Contactabilidad	Clientes	Iteraciones Promedio	Estrategia Recomendada
Preventiva	Bajo	Muy Alta	20,551	1.0	Monitoreo y Seguimiento
Preventiva	Medio	Muy Alta	15,233	1.0	Recuperación Digital
Recuperativa	Alto	Muy Baja	12,167	5.0	Recuperación Alto Valor
Preventiva	Bajo	Alta	11,155	2.0	Monitoreo y Seguimiento
Recuperativa	Medio	Muy Alta	6,678	1.0	Recuperación Digital
Preventiva	Medio	Alta	6,387	2.0	Recuperación Digital
Preventiva	Bajo	Media	5,793	3.0	Monitoreo y Seguimiento
Preventiva	Medio	Sin Contacto	5,009	No aplica	Monitoreo y Seguimiento
Preventiva	Medio	Muy Baja	3,339	5.0	Recuperación Bajo Costo
Preventiva	Medio	Media	2,835	3.0	Recuperación Digital
Preventiva	Bajo	Sin Contacto	2,435	No aplica	No Gestionar
Recuperativa	Alto	Muy Alta	2,313	1.0	Gestor Especializado

Recuperativa	Alto	Baja	2,309	4.0	Recuperación Alto Valor
Recuperativa	Medio	Alta	1,871	2.0	Recuperación Digital
Recuperativa	Alto	Alta	1,656	2.0	Gestor Especializado
Preventiva	Bajo	Muy Baja	1,473	5.0	Recuperación Bajo Costo
Recuperativa	Alto	Media	1,328	3.0	Recuperación Alto Valor
Preventiva	Alto	Muy Alta	990	1.0	Preventiva Alto Valor
Preventiva	Medio	Baja	867	4.0	Recuperación Bajo Costo
Recuperativa	Medio	Media	775	3.0	Recuperación Digital
Preventiva	Bajo	Baja	726	4.0	Recuperación Bajo Costo
Recuperativa	Bajo	Muy Alta	568	1.0	Monitoreo y Seguimiento
Preventiva	Alto	Alta	459	2.0	Preventiva Alto Valor
Recuperativa	Medio	Baja	361	4.0	Recuperación Bajo Costo
Recuperativa	Bajo	Alta	322	2.0	Monitoreo y Seguimiento
Recuperativa	Bajo	Muy Baja	257	5.0	Recuperación Bajo Costo
Recuperativa	Medio	Muy Baja	193	5.0	Recuperación Bajo Costo
Preventiva	Alto	Sin Contacto	180	No aplica	Preventiva Alto Valor
Recuperativa	Bajo	Media	150	3.0	Monitoreo y Seguimiento
Preventiva	Alto	Muy Baja	142	5.0	Preventiva Alto Valor
Recuperativa	Bajo	Baja	132	4.0	Recuperación Bajo Costo
Preventiva	Alto	Media	102	3.0	Preventiva Alto Valor
Recuperativa	Alto	Sin Contacto	98	No aplica	Recuperación Alto Valor
Preventiva	Alto	Baja	38	4.0	Preventiva Alto Valor
Recuperativa	Medio	Sin Contacto	16	No aplica	Monitoreo y Seguimiento
Recuperativa	Bajo	Sin Contacto	2	No aplica	No Gestionar

Nota. Elaboración propia.

Figura 56 Segmentos de gestión con mayor volumen de clientes



Nota. Elaboración propia.

Resultados de la generación de estrategias de gestión

La generación de estrategias de gestión transforma los resultados analíticos en políticas accionables de cobranza. Para ello, cada combinación de TipoGestion, GrupoValor y NivelContactabilidad fue asignada a una estrategia recomendada mediante reglas de negocio, las mismas que son parametrizables.

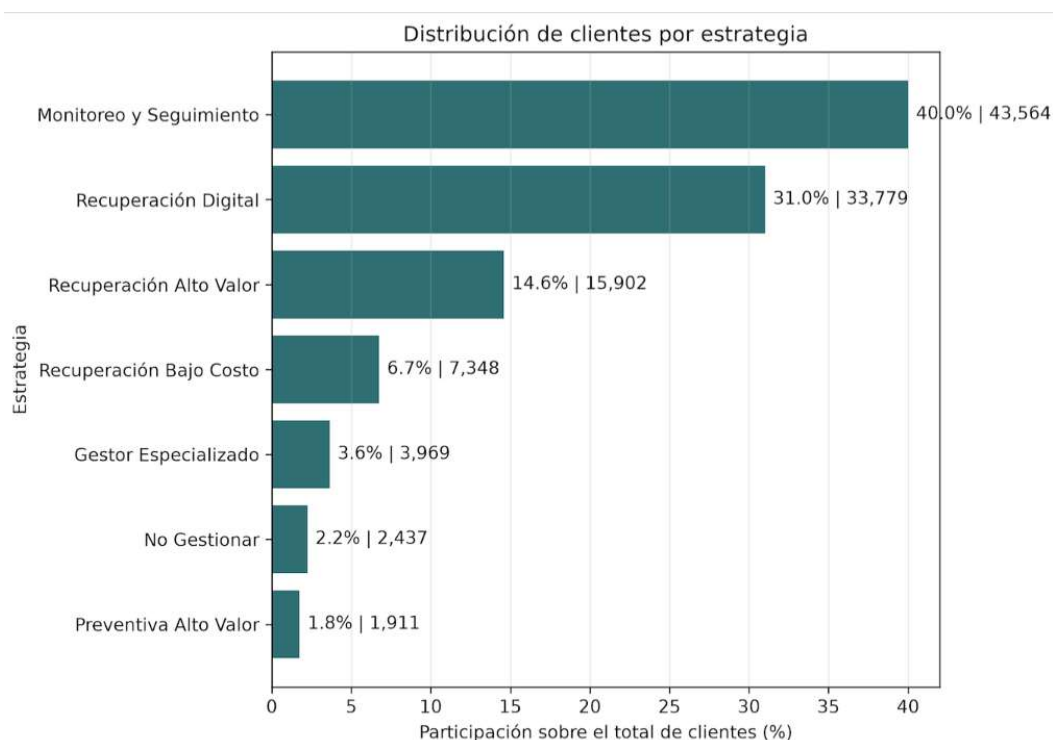
La distribución final evidencia que Monitoreo y Seguimiento concentra el mayor volumen, seguido de Recuperación Digital y Recuperación Alto Valor. Las dos primeras estrategias acumulan el 71.02% de los clientes, mientras que las tres primeras alcanzan el 85.62%. Este comportamiento refleja que la mayor parte de la cartera puede gestionarse mediante acciones de seguimiento, canales digitales o recuperación focalizada de alto valor.

Tabla 31 *Distribución definitiva de estrategias recomendadas y canales asociados*

Estrategia Recomendada	Canales	Clientes	Porcentaje	Porcentaje Acumulado
Monitoreo y Seguimiento	Email + PUSH/SMS	43,564	40.00%	40.00%
Recuperación Digital	WhatsApp + SMS + Email	33,779	31.02%	71.02%
Recuperación Alto Valor	WhatsApp + SMS + Telefonía	15,902	14.60%	85.62%
Recuperación Bajo Costo	SMS Masivo	7,348	6.75%	92.37%
Gestor Especializado	Telefonía + Gestor	3,969	3.64%	96.01%
No Gestionar	Ninguno	2,437	2.24%	98.25%
Preventiva Alto Valor	WhatsApp + SMS	1,911	1.75%	100.00%

Nota. Elaboración propia.

Figura 57 Distribución definitiva de estrategias recomendadas



Nota. Elaboración propia.

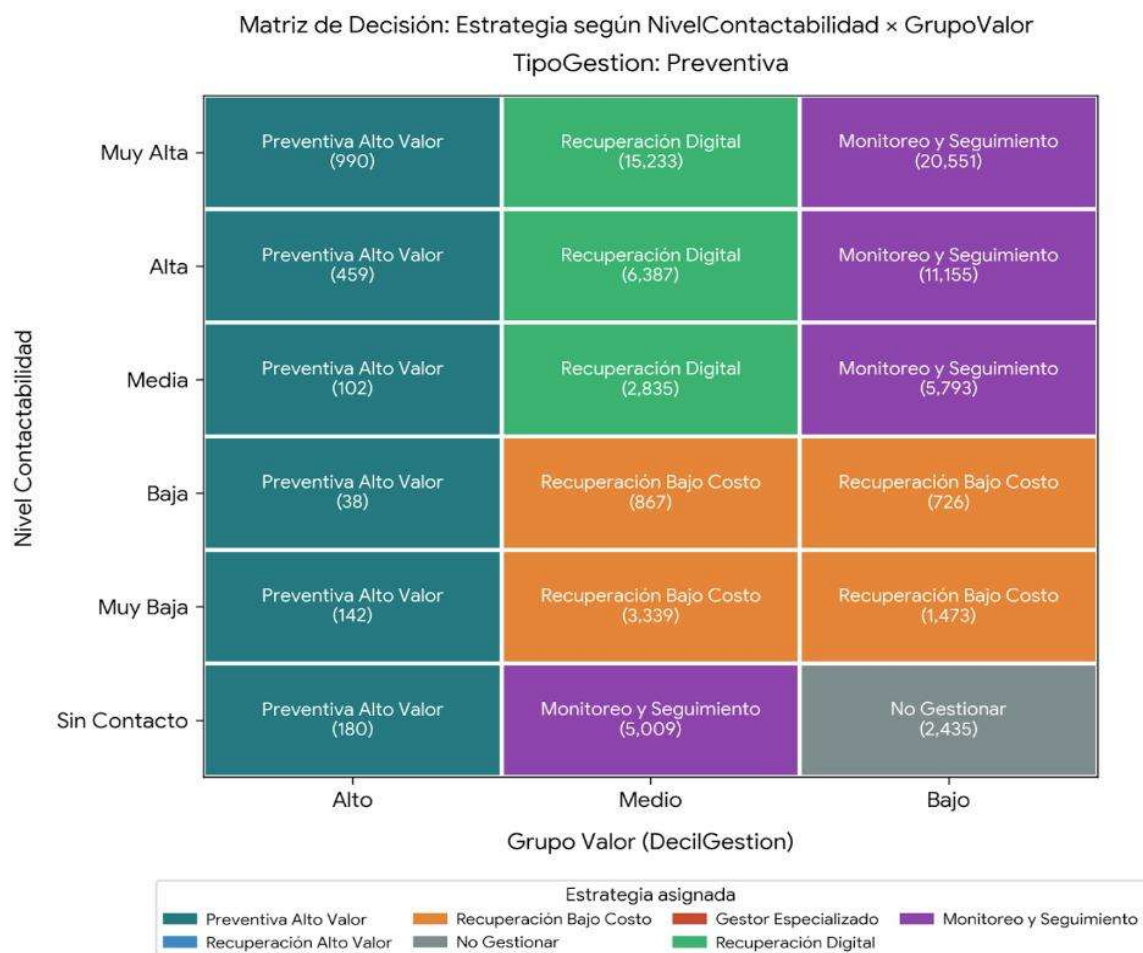
Matriz de decisión estratégica

La matriz de decisión resume la asignación de estrategias según NivelContactabilidad y GrupoValor, diferenciando entre TipoGestion Preventiva y TipoGestion Recuperativa. Esta visualización permite observar cómo cambia la estrategia cuando varía el nivel de contacto esperado o el grupo de valor derivado del DecilGestion.

En gestión preventiva, los segmentos de GrupoValor Alto se asignan a Preventiva Alto Valor, mientras que los grupos Medio y Bajo se distribuyen entre Recuperación Digital, Monitoreo y Seguimiento, Recuperación Bajo Costo o No Gestionar, dependiendo del nivel de contactabilidad. En gestión recuperativa, los casos de GrupoValor Alto con contactabilidad Alta o Muy Alta son enviados a Gestor Especializado, mientras que los restantes segmentos de alto valor se orientan a Recuperación Alto Valor.

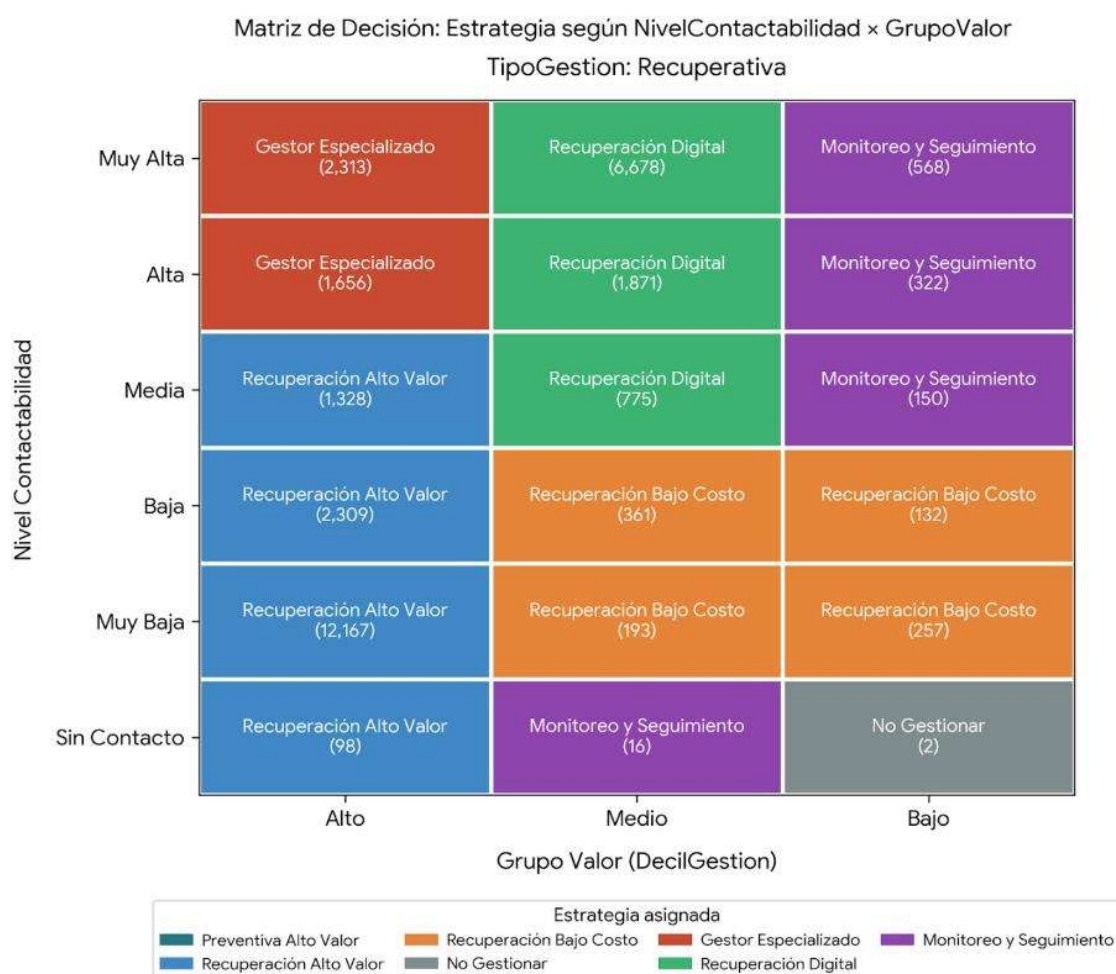
Esta matriz permite evidenciar que el modelo no aplica una única regla homogénea de cobranza, sino que ajusta la estrategia al perfil operativo de cada segmento.

Figura 58 Matriz de decisión estratégica con segmentos definitivos para tipo de gestión Preventiva



Nota. Elaboración propia.

Figura 59 Matriz de decisión estratégica con segmentos definitivos para tipo de gestión Recuperativa



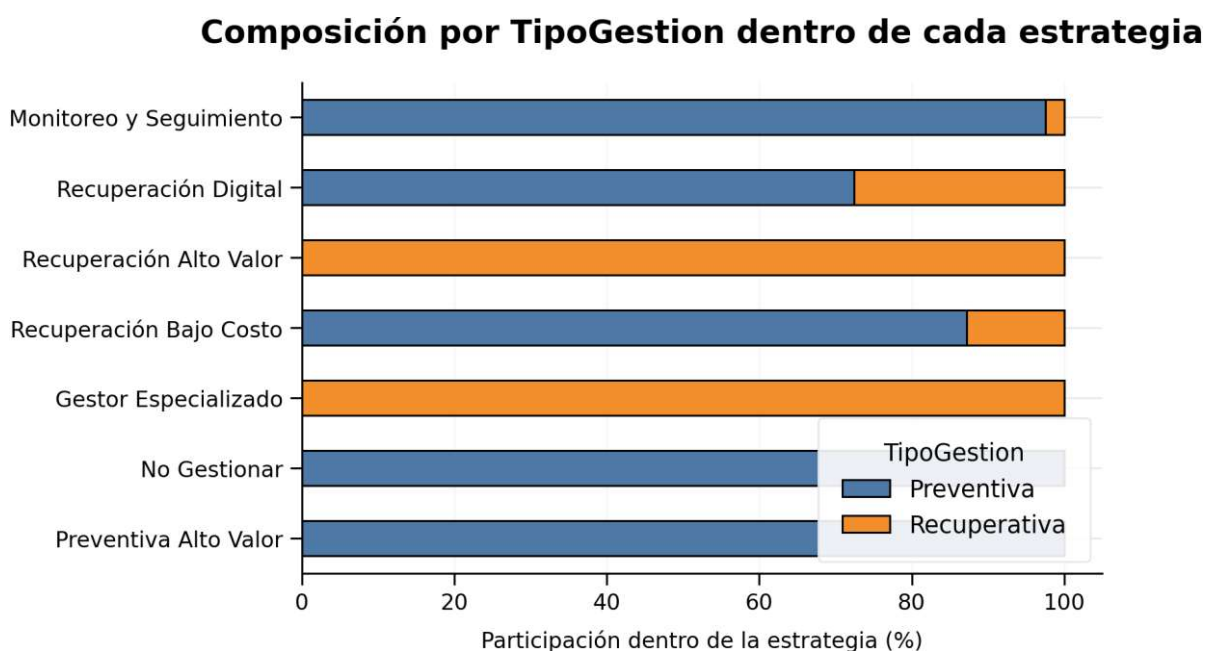
Nota. Elaboración propia.

Composición por TipoGestion dentro de cada estrategia

La composición por TipoGestion permite entender si cada estrategia se origina principalmente en cartera preventiva, recuperativa o en una combinación de ambas. Esta lectura aporta trazabilidad a la interpretación operativa, ya que una misma estrategia puede tener finalidades distintas según la naturaleza de la cuenta.

En la imagen siguiente se muestra la participación relativa de cuentas, preventivas y recuperativas, para cada estrategia. Esta vista complementa la distribución agregada y permite verificar que las estrategias intensivas o especializadas se concentran en los segmentos donde la regla de negocio lo define explícitamente.

Figura 60 Composición por TipoGestion dentro de cada estrategia



Nota. Elaboración propia.

Tabla 32 Interpretación operativa de las estrategias de gestión

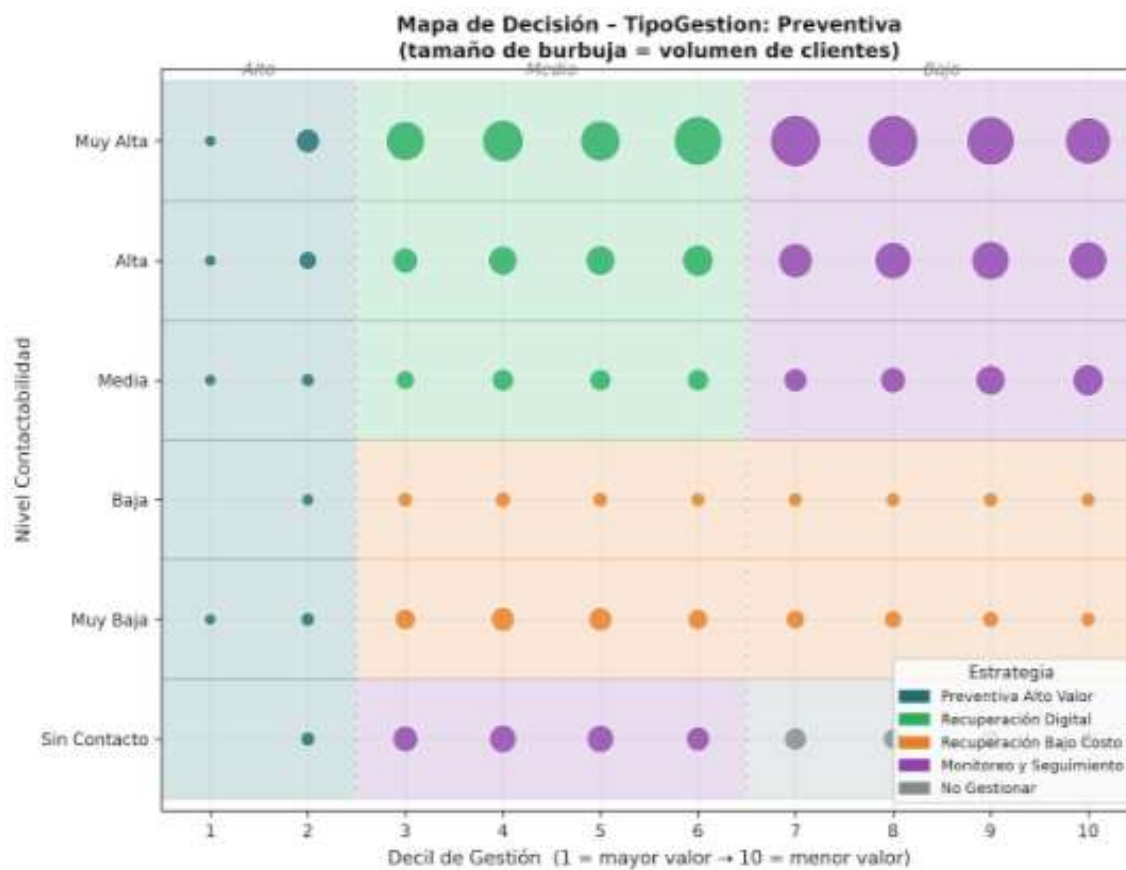
Estrategia	Descripción operativa	Justificación
Monitoreo y Seguimiento	Seguimiento operativo moderado por canales de bajo esfuerzo.	Concentra el mayor volumen de clientes y permite mantener presencia sin consumir recursos intensivos.
Recuperación Digital	Gestión escalable mediante WhatsApp, SMS y correo electrónico.	Adecuada para segmentos con contactabilidad media, alta o muy alta y valor intermedio.
Recuperación Alto Valor	Acción recuperativa focalizada para cuentas de alto valor.	Aplica en segmentos recuperativos de alto valor donde se requiere preservar potencial de recuperación.
Recuperación Bajo Costo	Acciones de bajo costo, principalmente SMS masivo.	Recomendada cuando la contactabilidad es baja o muy baja.
Gestor Especializado	Intervención humana prioritaria.	Orientada a cuentas recuperativas de alto valor con alta probabilidad de contacto.
No Gestionar	Exclusión temporal de campañas activas.	Aplicable en combinaciones de bajo valor y ausencia de contacto efectivo.
Preventiva Alto Valor	Acción preventiva priorizada mediante canales digitales.	Dirigida a cuentas preventivas clasificadas como alto valor por la regla de negocio.

Nota. Elaboración propia.

En conjunto, los resultados de la integración entre contactabilidad y priorización muestran que la estrategia final no depende únicamente del valor del cliente ni exclusivamente de la probabilidad de contacto, sino de la combinación de ambas dimensiones con el TipoGestion.

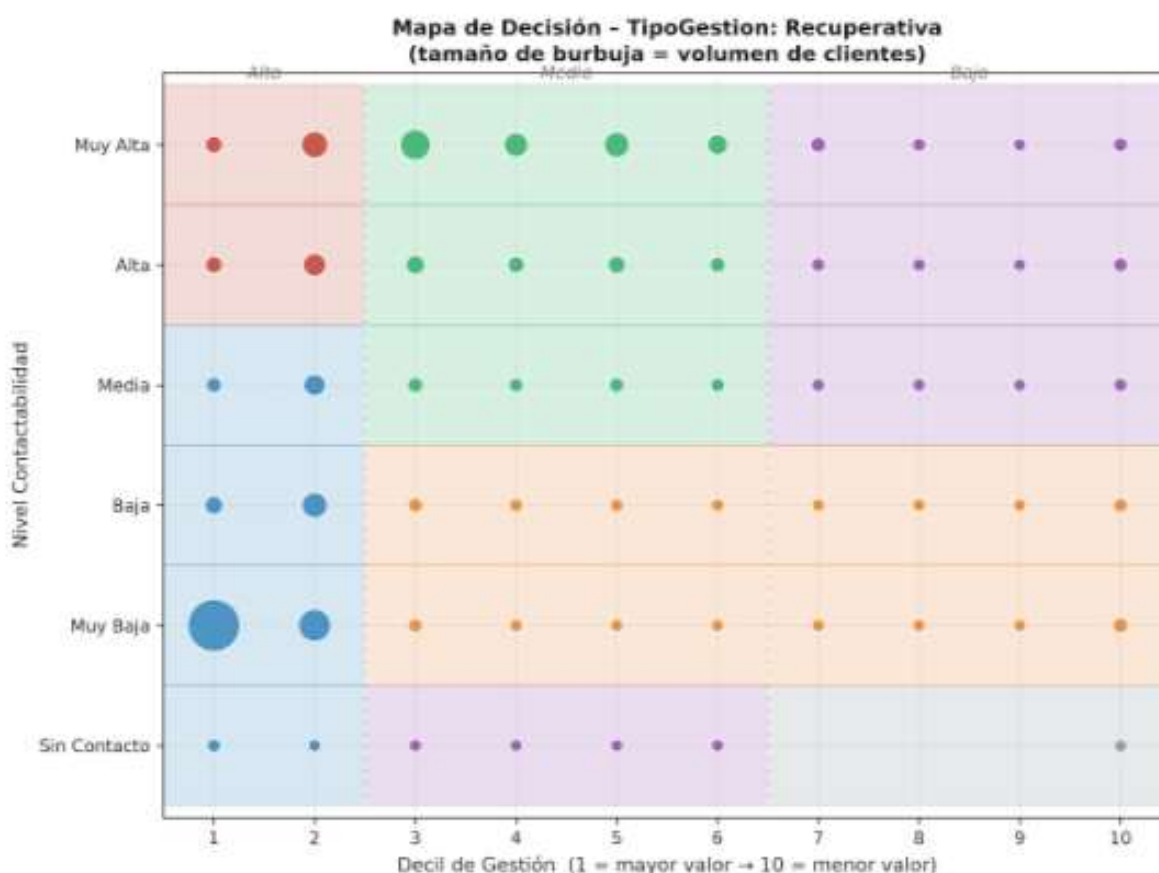
Esta integración permite construir una política de cobranza segmentada, trazable y orientada a la eficiencia operativa.

Figura 61 Mapas de Decisión de Estrategias Preventivas



Nota. Elaboración propia.

Figura 62 Mapas de Decisión de Estrategias Recuperativas



Nota. Elaboración propia.

La evidencia obtenida justifica la utilización del Modelo de Estrategia como capa final de decisión, debido a que transforma los resultados predictivos en reglas accionables que pueden ser parametrizadas en sistemas de Gestión de Cobranza.

Análisis económico de estrategias de cobranza

El análisis económico evalúa el impacto de convertir la priorización analítica y la contactabilidad en estrategias operativas de cobranza. A diferencia de una comparación basada únicamente en volumen de clientes, esta sección incorpora supuestos de costo por canal, compara el gasto actual frente al costo propuesto y cuantifica el ahorro esperado derivado de la asignación diferenciada de estrategias.

La estimación actualizada muestra que el costo total estimado del escenario actual asciende a 44,926.47, mientras que el escenario propuesto alcanza 28,699.64. La diferencia

representa un ahorro potencial de 16,226.83, equivalente al 36.12% del costo actual. Estos resultados evidencian que la asignación diferenciada de estrategias permite mejorar la focalización de la gestión y reducir el costo operativo estimado de la cobranza bajo los supuestos considerados en el análisis.

Supuestos de costo y construcción del escenario económico

Los costos unitarios se definieron como parámetros editables dentro del notebook de resultados. Esta parametrización permite actualizar los costos unitarios en caso de existir condiciones comerciales distintas para cada canal.

Tabla 33 *Supuestos de costo unitario utilizados para la evaluación económica*

Canal o concepto	Costo unitario
Telefonía	0.8500
SMS	0.0300
SMS Masivo	0.0300
WhatsApp	0.0100
Email	0.0025
PUSH	0.0010
Ninguno	0.0000
Sin información	0.0000

Nota. Elaboración propia.

El escenario actual se estimó a partir del comportamiento observado de envíos o contactos de los últimos treinta días por cliente y canal, consolidando por IdCuenta el promedio de las variables NumSMS30d, NumEmail30d y NumTelefonia30d. Este criterio es adecuado cuando DSScoreContacto contiene múltiples snapshots o cortes temporales por cliente. Por otra parte, el escenario propuesto utiliza la estrategia recomendada, los canales asociados a dicha estrategia y el número de iteraciones definido por el modelo de contactabilidad.

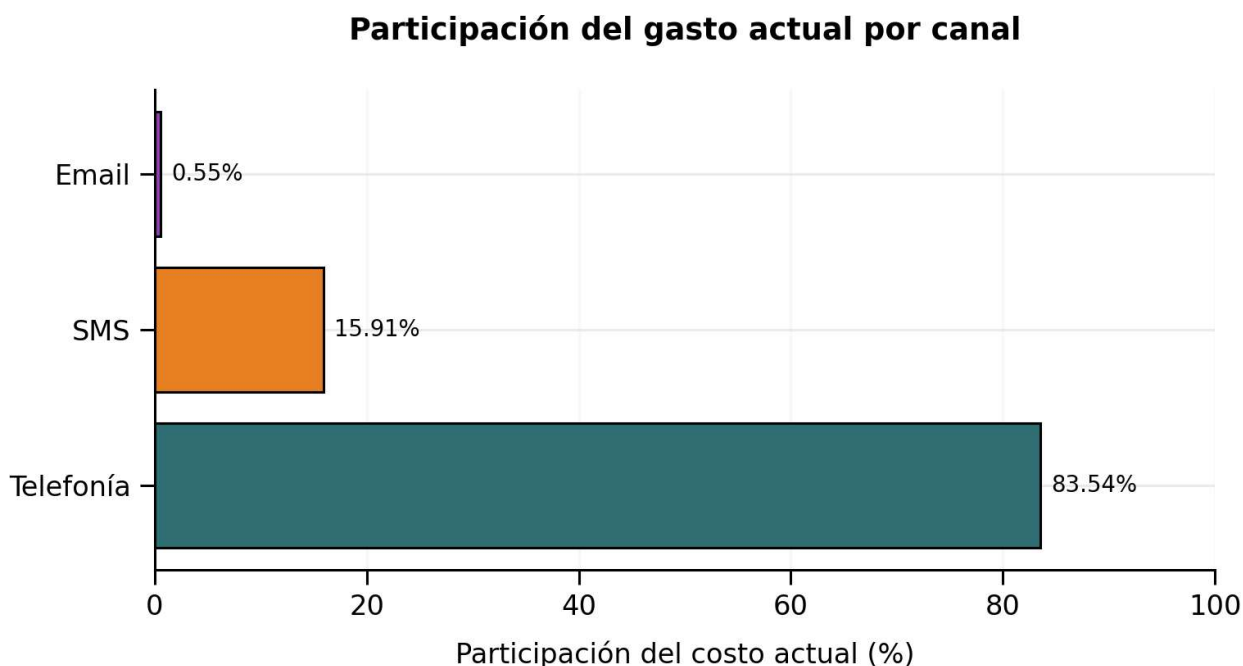
Composición del gasto actual por canal

La distribución del gasto actual evidencia que el componente de mayor peso corresponde a Telefonía, seguido por SMS y, en menor medida, Email. Este comportamiento es consistente con el costo unitario más alto asignado al canal telefónico y con la presencia de clientes con registros de contacto telefónico en DSScoreContacto.

Tabla 34 *Distribución del gasto actual estimado por canal*

Canal	Columna	Clientes	Clientes con envíos	Prom por cliente	Total promedios	Costo unitario	Gasto actual	Participación
Telefonía	NumTelefonia30d	101,170	20,141	0.4364	44,153.02	0.8500	37,530.06	83.54%
SMS	NumSMS30d	101,170	81,879	2.3548	238,240.20	0.0300	7,147.21	15.91%
Email	NumEmail30d	101,170	100,575	0.9859	99,680.65	0.0025	249.20	0.55%

Nota. Elaboración propia.

Figura 63 *Participación del gasto actual por canal*

Nota. Elaboración propia.

El canal Telefonía concentra el 83.54% del gasto actual, aunque el promedio de interacciones por cliente es menor que en SMS. Esto confirma que la eficiencia económica del modelo no depende únicamente de reducir el número de contactos, sino de reasignar los canales de mayor costo hacia segmentos donde su uso esté justificado por valor, oportunidad o probabilidad de recuperación.

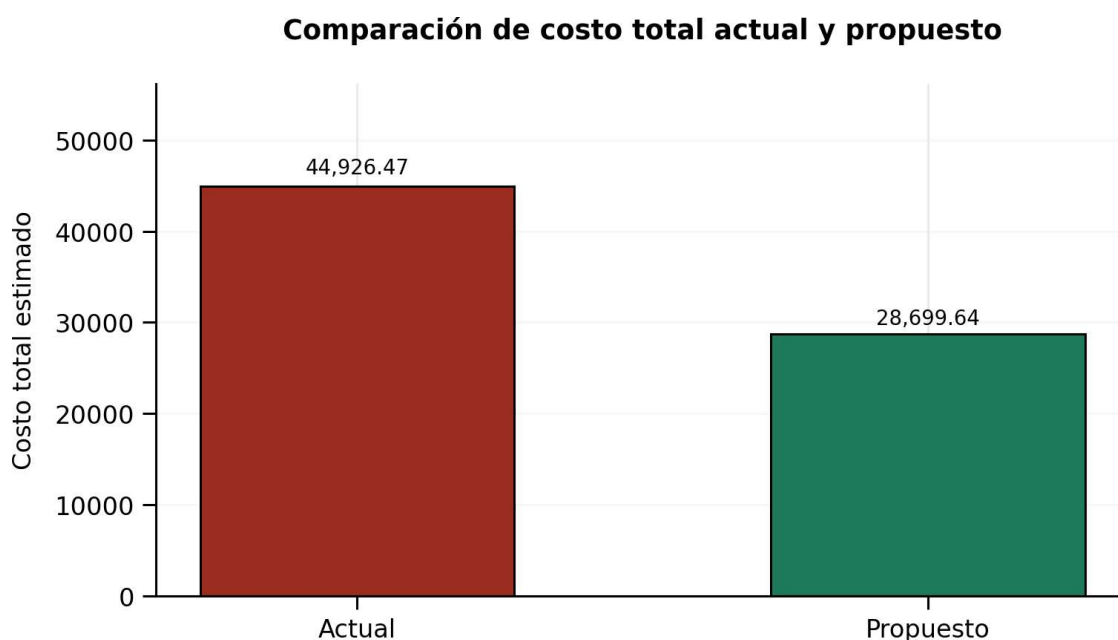
Comparación económica entre escenario actual y escenario propuesto

La comparación global muestra una disminución material del costo total estimado. El costo promedio por cliente disminuye de 0.41 a 0.26, lo que representa una reducción de 0.15 por cliente gestionado dentro del universo evaluado.

Tabla 35 *Comparación económica entre escenario actual, escenario propuesto y ahorro estimado*

Escenario	Costo total	Costo promedio por cliente	Variación vs. Actual
Actual	44,926.47	0.41	0.00%
Propuesto	28,699.64	0.26	-36.12%
Ahorro	16,226.83	0.15	36.12%

Nota. Elaboración propia.

Figura 64 *Comparación del costo total actual y propuesto*

Nota. Elaboración propia.

La diferencia estimada entre ambos escenarios no implica una reducción de la cobertura de gestión, sino una reasignación de los esfuerzos operativos según las características de cada segmento. Bajo los supuestos de costo utilizados, el escenario propuesto presenta una reducción estimada del gasto debido principalmente a la utilización de canales de menor costo en aquellos segmentos donde el valor, la prioridad o la probabilidad de contacto no justifican una intervención más intensiva. Este resultado debe interpretarse como una estimación económica del escenario propuesto y no como un ahorro efectivamente observado en producción.

Impacto económico por estrategia propuesta

El análisis por estrategia permite identificar qué políticas operativas presentan una mayor contribución a la diferencia de costo estimada entre el escenario actual y el escenario

propuesto. La estrategia Recuperación Digital genera el mayor ahorro absoluto, seguida por Recuperación Alto Valor y Monitoreo y Seguimiento. También se observa que Gestor Especializado presenta un ahorro negativo bajo los supuestos utilizados, lo que sugiere que su uso debe reservarse para casos donde el valor esperado de recuperación compense el mayor costo operativo.

Tabla 36 *Impacto económico por estrategia propuesta*

Estrategia	Clientes	Costo actual	Costo propuesto	Ahorro	Costo actual prom.	Costo prop. prom.	Iter. prom.	Ahorro %
Recuperación Digital	33,779	9,264.50	697.81	8,566.69	0.27	0.02	1.46	92.47%
Recuperación Alto Valor	15,902	26,062.15	21,998.72	4,063.43	1.64	1.38	4.66	15.59%
Monitoreo y Seguimiento	43,564	2,590.26	117.12	2,473.13	0.06	0.00	1.54	95.48%
Preventiva Alto Valor	1,911	885.20	65.12	820.08	0.46	0.03	1.70	92.64%
Recuperación Bajo Costo	7,348	1,499.78	1,039.62	460.16	0.20	0.14	4.72	30.68%
No Gestionar	2,437	0.00	0.00	0.00	0.00	0.00	1.00	N/A
Gestor Especializado	3,969	4,624.58	4,781.25	-156.67	1.17	1.20	1.42	-3.39%

Nota. Elaboración propia.

Figura 65 Ahorro estimado por estrategia propuesta



Nota. Elaboración propia.

La estrategia Recuperación Digital combina alto volumen con bajo costo marginal, por lo que explica una parte relevante del ahorro total. Monitoreo y Seguimiento también muestra eficiencia económica elevada debido al uso de canales de bajo costo. En contraste, Recuperación Alto Valor presenta un ahorro porcentual menor, pero mantiene importancia estratégica por el volumen de costo asociado y por su orientación hacia cuentas donde la intervención puede ser económicamente justificable.

Ahorro por combinación operativa

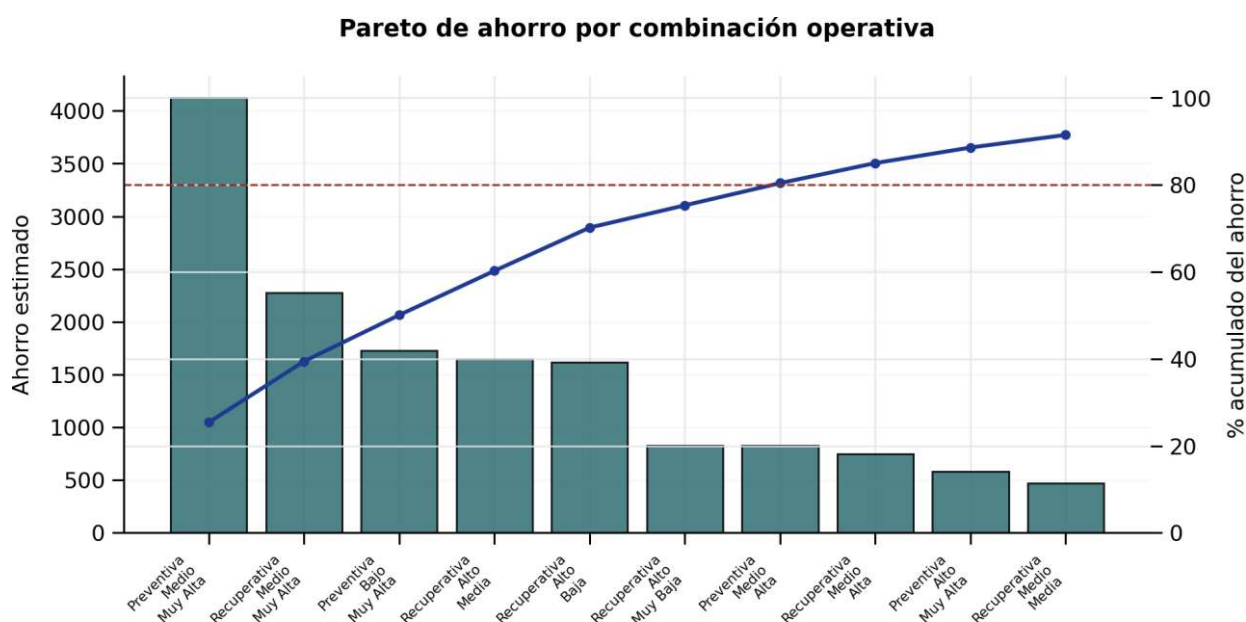
El análisis de concentración de Pareto permite determinar qué combinaciones de TipoGestion, GrupoValor y NivelContactabilidad concentran la mayor proporción de la diferencia económica estimada. Este análisis complementa la evaluación por estrategia, al permitir identificar los segmentos operativos que generan la mayor contribución a la reducción estimada del costo dentro del esquema de decisión propuesto.

Tabla 37 Principales combinaciones operativas según contribución al ahorro estimado

Rango	Combinación operativa	Estrategia	Clientes	Ahorro estimado	Ahorro acumulado
1	Preventiva Medio Muy Alta	Recuperación Digital	15,233	4,122.37	25.40%
2	Recuperativa Medio Muy Alta	Recuperación Digital	6,678	2,273.96	39.42%
3	Preventiva Bajo Muy Alta	Monitoreo y Seguimiento	20,551	1,728.38	50.07%
4	Recuperativa Alto Media	Recuperación Alto Valor	1,328	1,648.05	60.23%
5	Recuperativa Alto Baja	Recuperación Alto Valor	2,309	1,613.65	70.17%
6	Recuperativa Alto Muy Baja	Recuperación Alto Valor	12,167	830.81	75.29%
7	Preventiva Medio Alta	Recuperación Digital	6,387	829.33	80.40%
8	Recuperativa Medio Alta	Recuperación Digital	1,871	745.85	85.00%
9	Preventiva Alto Muy Alta	Preventiva Alto Valor	990	580.98	88.58%
10	Recuperativa Medio Media	Recuperación Digital	775	471.67	91.48%

Nota. Elaboración propia.

Figura 66 Pareto de ahorro por combinación operativa



Nota. Elaboración propia.

Las primeras siete combinaciones concentran aproximadamente el 80,40% del ahorro estimado, lo que evidencia que el beneficio económico no se distribuye de forma uniforme en toda la cartera. Esta concentración permite priorizar la implementación operativa en los segmentos de mayor impacto, reduciendo el riesgo de despliegues amplios sin retorno económico claro.

Capítulo 5

Conclusiones

Conclusión 1. Cumplimiento del objetivo general

La presente investigación permitió desarrollar un sistema integral de apoyo a la gestión de cobranza basado en la integración de modelos predictivos de comportamiento crediticio y contactabilidad. La solución integra los modelos de Mora, Pago, Regularización y Contactabilidad dentro de un marco analítico orientado a la priorización y definición de estrategias de cobranza. Como resultado, fue posible obtener para cada cuenta indicadores de riesgo, pago esperado, regularización, probabilidad de contacto y prioridad de gestión, proporcionando una base objetiva para mejorar la asignación de recursos y apoyar la toma de decisiones operativas y estratégicas.

Conclusión 2. Construcción del conjunto de datos y aporte metodológico

La metodología desarrollada permitió construir Datasets longitudinales basados en ventanas temporales móviles, incorporando múltiples observaciones históricas de cada cliente y preservando la separación entre la información utilizada como variables explicativas y los períodos posteriores empleados para la construcción de los targets. Esta estrategia permitió evitar la fuga de información temporal (data leakage) y representar diferentes momentos del comportamiento de los clientes. Para la comparación y selección de los algoritmos se utilizó una partición Train-Test sobre el Dataset consolidado, por lo que las métricas obtenidas corresponden a una validación interna bajo condiciones experimentales homogéneas. De manera complementaria, la evaluación comparativa del esquema de gestión actual frente a la propuesta se realizó utilizando información correspondiente a una ventana temporal posterior a la empleada durante el desarrollo, fortaleciendo la evaluación de la aplicabilidad de la solución propuesta en un período posterior.

Conclusión 3. Desarrollo y validación de los modelos predictivos

Los modelos predictivos desarrollados demostraron capacidad para estimar el riesgo de mora, el porcentaje esperado de pago, la probabilidad de regularización y la probabilidad de contacto efectivo con los clientes. Particularmente, el Modelo de Contactabilidad alcanzó métricas de desempeño que evidencian una adecuada capacidad de discriminación, permitiendo segmentar clientes según su probabilidad de interacción y aportar información complementaria a los modelos financieros tradicionales

Conclusión 4. Modelo de Priorización de Gestión de Cobranza

La integración de los modelos predictivos mediante el Score de Priorización permitió transformar múltiples indicadores individuales en un mecanismo único de apoyo a la decisión. La segmentación de la cartera en deciles evidenció una adecuada capacidad para diferenciar cuentas según su nivel de riesgo, potencial de recuperación y exposición económica, facilitando la definición de estrategias preventivas y recuperativas acordes con las características de cada segmento y favoreciendo una utilización más eficiente de los recursos operativos.

Conclusión 5. Aplicabilidad de la solución propuesta

La solución desarrollada fue implementada como una prueba de concepto preparada para su integración con el Core operativo mediante procesos batch. Como resultado, se generó una estructura consolidada que incorpora probabilidades predictivas, scores de priorización, nivel de contactabilidad y estrategias recomendadas de gestión. Este enfoque demuestra la viabilidad técnica de incorporar Inteligencia Artificial dentro de los procesos operativos de cobranza sin modificar la arquitectura transaccional existente.

Conclusión 6. Contrastación de la hipótesis de investigación

Los resultados obtenidos proporcionan evidencia favorable respecto de la hipótesis planteada, al demostrar que la integración de modelos predictivos de comportamiento crediticio y contactabilidad permite construir un esquema de gestión basado en criterios objetivos de priorización y asignación diferenciada de recursos. La comparación económica realizada entre el esquema actual y el escenario propuesto evidencia, bajo los supuestos de costo definidos, una reducción estimada del 36.12 % en el costo operativo de gestión. No obstante, al tratarse de una prueba de concepto, este resultado no constituye evidencia de recuperación incremental efectivamente observada. La validación definitiva de la mejora frente al proceso tradicional requerirá una implementación piloto que permita comparar ambos esquemas mediante resultados operativos reales.

Conclusión 7. Aporte del Modelo de Contactabilidad

La incorporación del Modelo de Contactabilidad permitió complementar la visión financiera tradicional de la gestión de cobranza mediante una estimación objetiva de la probabilidad de establecer contacto efectivo con cada cliente. Este componente demostró que la prioridad financiera junto con la probabilidad de contactabilidad permiten determinar la mejor estrategia de gestión. En consecuencia, la solución desarrollada permite responder simultáneamente las preguntas de qué clientes gestionar y cómo gestionarlos, fortaleciendo el proceso integral de toma de decisiones.

Conclusión 8. Generación de estrategias diferenciadas de cobranza

Los resultados obtenidos evidencian que es posible transformar las salidas de los modelos predictivos en políticas operativas concretas mediante una matriz de decisión basada en Tipo de Gestión, Grupo de Valor y Nivel de Contactabilidad. La segmentación desarrollada permitió asignar estrategias diferenciadas, demostrando que una gestión segmentada puede

representar una alternativa más eficiente que la aplicación de estrategias homogéneas a toda la cartera.

Conclusión 9. Integración entre analítica predictiva y operación

Uno de los principales aportes de la investigación consiste en la integración de múltiples componentes analíticos dentro de una única arquitectura de decisión. El proyecto logró conectar la preparación de datos, el modelado predictivo, la construcción de scores, la segmentación de contactabilidad y la asignación de estrategias operativas en un flujo analítico completo y reproducible. Esta integración demuestra que la Inteligencia Artificial puede ser utilizada no únicamente como una herramienta predictiva, sino también como un mecanismo de apoyo para la definición de acciones concretas de negocio.

Conclusión 10. Contribución académica y metodológica

Desde una perspectiva académica, el trabajo aporta una metodología replicable para abordar problemas de cobranza mediante la integración de aprendizaje automático, segmentación operativa, priorización de cartera y contactabilidad. La propuesta desarrollada puede ser adaptada a diferentes organizaciones que administran cartera crediticia, constituyendo una referencia metodológica para futuras investigaciones orientadas a integrar modelos predictivos con mecanismos de apoyo a la decisión. Además, el enfoque propuesto demuestra cómo la Inteligencia Artificial puede contribuir a una gestión más eficiente, objetiva y basada en evidencia en contextos financieros de alta complejidad.

Recomendaciones

Recomendación 1. Implementación gradual de la solución en producción

Se recomienda implementar la solución propuesta de manera progresiva dentro del Core operativo de la organización, iniciando mediante una fase piloto controlada que permita validar el comportamiento conjunto de los modelos predictivos, el score de priorización, el

modelo de contactabilidad y las estrategias recomendadas de cobranza. Esta implementación facilitará el monitoreo de indicadores operativos y financieros, permitiendo realizar ajustes antes de una adopción institucional a gran escala. La incorporación gradual reduce el riesgo operativo y favorece la aceptación de la solución por parte de las áreas usuarias.

Recomendación 2. Establecimiento de un proceso formal de monitoreo y reentrenamiento

Debido a la naturaleza dinámica del comportamiento de los clientes, las condiciones económicas y las estrategias comerciales de la organización, se recomienda establecer un esquema permanente de monitoreo y recalibración de los modelos predictivos. Este proceso debería incluir evaluaciones periódicas de desempeño, análisis de deriva de datos (data drift), seguimiento de estabilidad de variables y reentrenamiento programado cuando se identifiquen cambios significativos en los patrones de comportamiento. De esta manera, los modelos podrán mantener su capacidad predictiva y continuar generando resultados consistentes a lo largo del tiempo.

Recomendación 3. Explotación operativa del Modelo de Contactabilidad

Se recomienda incorporar formalmente el Modelo de Contactabilidad dentro de las reglas operativas de cobranza, utilizando variables como ProbabilidadContactabilidad, NivelContactabilidad, TipoContacto y Número de Iteraciones recomendadas como criterios de asignación de recursos. La utilización sistemática de estas variables permitirá mejorar la asignación de los canales utilizados, reducir intentos innecesarios de contacto y aumentar la efectividad de las gestiones realizadas, complementando las decisiones basadas en riesgo y potencial de recuperación. La estrategia final debe considerar simultáneamente el valor de la cuenta y la probabilidad real de interacción con el cliente.

Recomendación 4. Parametrización y evolución continua de las estrategias de cobranza

Se recomienda que las estrategias de cobranza definidas mediante la matriz de decisión permanezcan parametrizadas y sean ajustadas periódicamente según las políticas comerciales, capacidad operativa, costos de contacto y objetivos financieros de la organización. Las estrategias deben ser consideradas como una base inicial susceptible de evolucionar conforme se obtenga nueva evidencia operativa sobre su efectividad.

Recomendación 5. Evaluación continua del impacto de negocio

Una vez implementada la solución, se recomienda establecer un marco de medición orientado a evaluar periódicamente el impacto generado sobre indicadores clave de negocio. Entre los indicadores sugeridos se incluyen recuperación efectiva de cartera, reducción de mora, porcentaje de regularización, efectividad de contacto, productividad de gestores, utilización de canales, costo por gestión y retorno sobre la inversión operativa. Esta evaluación permitirá cuantificar objetivamente los beneficios obtenidos respecto al esquema tradicional de cobranza y sustentar futuras decisiones de inversión tecnológica.

Recomendación 6. Evolución hacia un ecosistema de cobranza inteligente

La arquitectura desarrollada facilita la incorporación de nuevos modelos, fuentes de información y mecanismos avanzados de toma de decisiones. Por esta razón, se recomienda continuar la evolución de la solución mediante la integración de nuevas variables comportamentales, datos transaccionales, información geográfica, patrones de interacción digital y otras fuentes relevantes que permitan enriquecer los modelos predictivos. Esta evolución puede conducir gradualmente hacia un ecosistema de cobranza inteligente capaz de adaptarse dinámicamente a las características de cada cliente y a las necesidades cambiantes de la organización.

Recomendación 7. Evaluación futura de técnicas de optimización multiobjetivo

Como línea de evolución de la solución desarrollada, se recomienda evaluar en futuras investigaciones la incorporación de técnicas formales de optimización multiobjetivo, como NSGA-II, para analizar simultáneamente criterios de recuperación, riesgo, contactabilidad y costo operativo. Esta extensión permitiría explorar soluciones no dominadas y Frentes de Pareto como complemento al esquema actual de asignación de estrategias basado en reglas de decisión, una vez que se disponga de evidencia operativa suficiente para definir y validar formalmente las funciones objetivo y sus restricciones.

Recomendación 8. Incorporación de experimentación controlada (A/B Testing)

Con el fin de validar cuantitativamente el beneficio de las estrategias recomendadas, se recomienda ejecutar pruebas piloto controladas que comparen el desempeño de la solución propuesta frente a las estrategias tradicionales actualmente utilizadas. La aplicación de metodologías A/B Testing permitirá medir el efecto real de la priorización, la contactabilidad y las estrategias diferenciadas sobre indicadores como recuperación monetaria, efectividad de contacto y costo operativo. Esta evidencia facilitará futuras decisiones de escalamiento y adopción institucional.

Recomendación 9. Fortalecimiento de la gobernanza y trazabilidad de modelos

Dado que la solución incorpora múltiples modelos predictivos e integra sus resultados dentro de procesos de negocio, se recomienda establecer mecanismos formales de gobernanza de modelos que incluyan documentación técnica, control de versiones, auditoría de variables, monitoreo de desempeño y procesos de aprobación para futuras modificaciones. Esto permitirá garantizar la trazabilidad, reproducibilidad y sostenibilidad de la solución en entornos productivos.

Recomendación 10. Líneas futuras de investigación

Desde una perspectiva académica, se recomienda profundizar la investigación mediante la evaluación de modelos avanzados de aprendizaje profundo, aprendizaje por refuerzo (Reinforcement Learning) y optimización dinámica de políticas de cobranza. Asimismo, futuras investigaciones podrían analizar la incorporación de restricciones operativas reales, capacidad de gestores, presupuestos por canal y métricas de satisfacción del cliente, con el fin de construir modelos de decisión aún más cercanos a la realidad operativa de las organizaciones financieras.

Referencias

- Ayari, H., Guetari, R., & Kraïem, N. (18 de Noviembre de 2025). Machine learning powered financial credit scoring: a systematic literature review. *Artificial Intelligence Review*, 59(13). Obtenido de <https://link.springer.com/article/10.1007/s10462-025-11416-2>
- Babaei, G. G. (2023). Explainable FinTech lending. *Journal of Economics and Business*(106126), 125-126. doi:10.1016/j.jeconbus.2023.106126
- Bahloul, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, 19(2), 104. doi:10.3390/jrfm19020104
- Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4), 563-574. Obtenido de <https://www.sciencedirect.com/science/article/abs/pii/S0169207013000502>
- Bermúdez Vera, I., Mosquera Restrepo, J., & Manotas-Duque, D. (2025). Data mining for the adjustment of credit scoring models in solidarity economy entities: A methodology for addressing class imbalances. *Risks*, 13(2), 20. Obtenido de <https://www.mdpi.com/2227-9091/13/2/20>
- Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 2433-2442. doi:10.1016/j.eswa.2011.08.093
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453. doi:10.1016/j.eswa.2011.09.033
- Carrera, A., & Benalcázar, M. (2025). Optimizing the collection process in credit risk management: A comparison of machine learning techniques. *Journal of Risk and Financial Management*, 18(11), 630. doi:10.3390/jrfm18110630

- Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review*, 57, 216.
doi:10.1007/s10462-024-10854-8
- Chehrizi, N., Glynn, P., & Weber, T. (2019). Dynamic Credit-Collections Optimization. *Management Science*, 65(6), 2737-2769. doi:10.1287/mnsc.2018.3070
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91(106263).
doi:10.1016/j.asoc.2020.106263
- Gicić, A., & Đonko, D. (2025). *Revolutionizing credit risk management: Opportunities and challenges in credit scoring with AI and deep learning*. IntechOpen.
doi:10.5772/intechopen.1011543
- Goh, R. Y., & Lee, L. S. (2019). Credit scoring: A review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019(1974794).
doi:10.1155/2019/1974794
- Hayashi, Y. (2022). Emerging trends in deep learning for credit scoring: A review. *Electronics*, 11, 3181. doi:10.3390/electronics11193181
- Jankowski, R., & Paliński, A. (2024). Debt collection model for mass receivables based on decision rules—A path to efficiency and sustainability. *Sustainability*, 16(14), 5885.
doi:10.3390/su16145885
- Jones, D., & Florentino, H. (2022). Multi-objective optimization: Methods and applications. En *The Palgrave Handbook of Operations Research*. Springer. doi:10.1007/978-3-030-96935-6_6
- Kamimura, E., Pinto, A., & Nagano, M. (2023). A recent review on optimisation methods applied to credit scoring models. *Journal of Economics, Finance and Administrative Science*, 28(56), 352-371. doi:10.1108/jefas-09-2021-0193

- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
doi:<https://doi.org/10.1016/j.jbankfin.2010.06.001>
- Kvamme, H. S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management*, 15(12)(556). doi:10.3390/jrfm15120556
- Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1), 124-136. doi:10.1016/j.ejor.2015.05.030
- Ma, H., Zhang, Y., Sun, S., Liu, T., & Shan, Y. (2023). A comprehensive survey on NSGA-II for multi-objective optimization and applications. *Artificial Intelligence Review*, 56, 15217-15270. doi:10.1007/s10462-023-10526-z
- Mestiri, S. (2024). Credit scoring using machine learning and deep learning-based models. *Data Science in Finance and Economics*, 4(2), 236-248. Obtenido de https://www.researchgate.net/publication/357449647_Credit_scoring_using_machine_learning_and_deep_Learning-Based_models
- Noriega, J., Rivera, L., & Herrera, J. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, 8(11), 169. doi:10.3390/data8110169
- Sánchez, C., Maldonado, S., & Vairetti, C. (2022). Improving debt collection via contact center information: A predictive analytics framework. *Decision Support Systems*, 159, 113812. doi:10.1016/j.dss.2022.113812
- Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, 34(14327-14339).
doi:10.1007/s00521-022-07472-2

- Shi, X., Tang, D., & Yu, Y. (2025). Credit scoring prediction using deep learning models in financial systems. *IEEE Access*. Obtenido de <https://ieeexplore.ieee.org/document/11087204>
- Shoghi, A. (2019). Debt Collection Industry: Machine Learning Approach. *Journal of Money and Economy*, 14(4), 453-473. Obtenido de <https://jme.mbri.ac.ir/article-1-469-en.html>
- Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2021). Comparison of Profit-Based Multi-Objective Approaches for Feature Selection in Credit Scoring. *Algorithms*, 14(9), 260. doi:10.3390/a14090260
- Xu, Y., Kou, G., Peng, Y., Ding, K., Ergu, D., & Alotaibi, F. S. (2024). Profit- and risk-driven credit scoring under parameter uncertainty: A multiobjective approach. *Omega*, 125(103004). doi:<https://doi.org/10.1016/j.omega.2023.103004>
- Yang, C., Dong, Y., Lu, J., & Peng, Z. (2022). Solving imbalanced data in credit risk prediction. *Proceedings of the 2022 International Conference on Machine Learning and Machine Intelligence*. Association for Computing Machinery. doi:10.1145/3568199.3568204
- Yeo, W., Van Der Heever, W., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2025). A comprehensive review on financial explainable AI. *Artificial Intelligence Review*, 58, 189. Obtenido de <https://link.springer.com/article/10.1007/s10462-024-11077-7>

Anexo

Repositorio digital de la investigación

Con el propósito de garantizar la trazabilidad, reproducibilidad y verificabilidad de los resultados obtenidos en la presente investigación, se entrega un repositorio digital que contiene todos los recursos desarrollados durante el proceso metodológico.

El repositorio reúne información de los datasets utilizados, notebooks de desarrollo, documentación técnica, modelos entrenados, archivos de resultados y recursos gráficos generados durante la construcción, entrenamiento, evaluación e implementación de los modelos de Inteligencia Artificial propuestos.

La organización del repositorio sigue las etapas metodológicas de la investigación, facilitando la consulta y reproducción de cada uno de los procedimientos realizados.

Estructura del repositorio

Carpeta	Descripción
documentacion	Contiene toda la documentación técnica generada durante el desarrollo de la investigación, organizada por componente metodológico. Incluye datasets descriptivos, análisis estadísticos, selección de variables, métricas, matrices de confusión, importancia de variables, gráficos y documentación utilizada para validar los modelos desarrollados.
modelos	Contiene los modelos de Inteligencia Artificial entrenados y almacenados en formato PKL para los modelos de Mora, Pago, Regularización y Contactabilidad, incluyendo los modelos finales seleccionados y los algoritmos evaluados durante la investigación.
notebook	Contiene los notebook del proyecto, donde se implementa el flujo completo de preparación de datos, entrenamiento, evaluación y producción de resultados, así como el archivo requirements.txt con las dependencias necesarias para reproducir el entorno de desarrollo.
resultados	Contiene los resultados finales obtenidos para cada uno de los modelos desarrollados, incluyendo comparación de algoritmos, métricas de desempeño, predicciones, metadatos, casos de estudio, informes finales y archivos utilizados para la elaboración de tablas y figuras presentadas en la tesis.

Organización de la carpeta documentación

Carpeta	Contenido
datasets	Resúmenes estadísticos y descriptivos de los datasets utilizados para el Modelo de Priorización y el Modelo de Contactabilidad.
Estrategia	Variables utilizadas para el cálculo del Score de Priorización, exposición económica, deciles de gestión, tipo de gestión y resultados de segmentación.
TargetMora	Correlaciones, análisis exploratorio, auditoría de variables, VIF, métricas, matrices de confusión, importancia de variables y gráficos correspondientes al Modelo Predictivo de Mora.

TargetPago	Correlaciones, análisis exploratorio, métricas de regresión, importancia de variables y gráficos correspondientes al Modelo Predictivo de Pago.
TargetRegularizacion	Correlaciones, análisis exploratorio, auditoría de variables, métricas, matrices de confusión, importancia de variables y gráficos correspondientes al Modelo Predictivo de Regularización.
TargetContacto	Documentación técnica, análisis exploratorio, métricas y gráficos correspondientes al Modelo de Contactabilidad.

Organización de la carpeta modelos

Carpeta	Descripción
Estrategia	Componentes asociados al Modelo de Priorización de Gestión de Cobranza.
TargetMora	Modelos entrenados para la predicción de Mora.
TargetPago	Modelos entrenados para la estimación del Pago Esperado.
TargetRegularizacion	Modelos entrenados para la predicción de Regularización.
TargetContacto	Modelos entrenados correspondientes al Modelo de Contactabilidad.

Organización de la carpeta resultados

Carpeta	Descripción
Estrategia	Resultados obtenidos por el Modelo de Priorización de Gestión de Cobranza.
TargetMora	Resultados experimentales del Modelo Predictivo de Mora.
TargetPago	Resultados experimentales del Modelo Predictivo de Pago.
TargetRegularizacion	Resultados experimentales del Modelo Predictivo de Regularización.
TargetContacto	Resultados experimentales del Modelo de Contactabilidad.

Repositorio Digital

El repositorio digital que contiene los archivos descritos en este anexo forma parte de los materiales complementarios entregados junto con la presente tesis y se encuentra disponible mediante el siguiente enlace.

https://momaservices-my.sharepoint.com/:f:/p/jairo_morales/IgBkOoXQ_1S4Rp_o_ojue7fkAYcirpzlPsgilLv1f7DBXq4?e=8UyrzX

Clave de acceso: A7!jM#4oJ9dUiDE2