ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Desarrollo de un sistema predictivo para la asignación óptima de cupos de crédito en una empresa de nutrición animal

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Andy Evirley Sánchez Astudillo

GUAYAQUIL - ECUADOR

Año: 2025

DEDICATORIA

A Dios, por ser mi guía constante y fuente inagotable de fe y esperanza.

A mi familia, quienes han sido mi pilar fundamental a lo largo de este camino, brindándome amor, apoyo incondicional y la fuerza necesaria para superar cada reto.

A Ingrid Olmedo, por su apoyo, comprensión y compañía en cada etapa de este proyecto. Gracias por ser mi mayor inspiración y por creer en mí incluso en los momentos más difíciles.

A todos aquellos que, de una u otra forma, contribuyeron con palabras de aliento, consejos o simplemente con su presencia, haciéndome sentir acompañado en esta meta.

AGRADECIMIENTOS

Agradezco a Dios por iluminar mi camino y brindarme la fortaleza necesaria para alcanzar este objetivo.

A mi familia, por su amor incondicional, su apoyo permanente y por ser siempre mi mayor motivación.

A Ingrid Olmedo, por su paciencia, comprensión y acompañamiento constante, siendo un pilar fundamental en este proceso.

A mi tutor, Christian Galarza, por su valiosa orientación, exigencia y dedicación, cuyo acompañamiento fue clave para la culminación de este trabajo.

A mi alma máter, la Escuela Superior Politécnica del Litoral, por brindarme la formación académica y los valores que hoy me permiten alcanzar esta meta.

DECLARACIÓN EXPRESA

Yo Andy Evirley Sánchez Astudillo acuerdo y reconozco que: La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores. El o los estudiantes deberán procurar en cualquier caso de cesión de sus derechos patrimoniales incluir una cláusula en la cesión que proteja la vigencia de la licencia aquí concedida a la ESPOL.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, secreto empresarial, derechos patrimoniales de autor sobre software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 18 de septiembre del 2025.

Firmado electrónicamente por:
ANDY EVIRLEY
SANCHEZ ASTUDILLO
Validar únicamente con FirmaEC

Andy Evirley Sánchez Astudillo

Autor

COMITÉ EVALUADOR

Christian Eduardo Galarza Morales
Time Stamping Security Data

Christian Eduardo Galarza

PROFESOR TUTOR

Firmado electrónicamente por:
EDUARDO SEGUNDO CRUZ RAMIREZ
Validar únicamente con FirmaEC

Eduardo Cruz

PROFESOR EVALUADOR

RESUMEN

Este proyecto se centra en el desarrollo de un modelo predictivo que, al emitir una factura, estime el riesgo de impago en las cuentas por cobrar de un negocio. La hipótesis plantea que mediante técnicas de machine learning es posible identificar a clientes con alto riesgo de impago antes de que ocurra, pretendiendo cambiar la gestión de cobranza de reactiva a preventiva. La motivación surge de la necesidad de disminuir los costos y el esfuerzo involucrados en la recuperación de cuentas vencidas, al tiempo que se optimiza el flujo de caja y se mejora la salud financiera.

Se utilizaron datos históricos simulados, inspirados en el sector de nutrición animal, con estructuras y variables similares a las que se gestionan habitualmente en negocios del rubro. El preprocesamiento de los datos incluyó la limpieza de registros, generación de variables adicionales, eliminación de características redundantes tras un análisis de correlación y balanceo de clases mediante SMOTE, asegurando un conjunto de datos robusto y representativo para el entrenamiento de los modelos.

Se compararon cuatro algoritmos: regresión logística, árboles de decisión, random forest y LightGBM; mediante validación temporal y búsqueda en cuadrícula de hiperparámetros. Posteriormente se evaluó su desempeño con métricas como precisión, recall y F1-score, poniendo especial énfasis en la capacidad de detectar verdaderos impagos. El random forest demostró un equilibrio superior entre precisión y generalización, imponiéndose como la opción más robusta para su implementación en producción.

Finalmente, se lanzó un prototipo interactivo con Gradio que, al seleccionar un cliente y ajustarse opcionalmente sus parámetros de pago, calcula en tiempo real la probabilidad de impago de una nueva factura. Esta herramienta se propone como un activo estratégico para una gestión de cartera proactiva, reduciendo el riesgo de morosidad y fortaleciendo la sostenibilidad financiera del negocio.

Palabras Clave: modelo predictivo, morosidad, machine learning, gestión de cartera, riesgo financiero.

ABSTRACT

This project focuses on the development of a predictive model that, at invoice issuing, estimates the risk of non-payment in a company's accounts receivable. The hypothesis posits that by applying machine learning techniques it is possible to identify customers at elevated risk of default before it occurs, shifting collections management from a reactive to a proactive approach. The motivation stems from the need to reduce the costs and effort involved in recovering overdue accounts, while optimizing cash flow and improving the organization's financial health.

Simulated historical data inspired by the animal nutrition sector was used, with structures and variables like those typically managed by companies in the industry. Data preprocessing included record cleaning, generation of additional features, removal of redundant variables through correlation analysis, and class balancing via SMOTE, ensuring a robust and representative dataset for model training.

Logistic regression, decision trees, random forest, and LightGBM algorithms were trained using time-based validation and grid search for hyperparameter tuning. Performance was evaluated with metrics such as precision, recall, and F1-score, with particular emphasis on the ability to detect true defaults. The random forest model demonstrated a superior balance between accuracy and generalization, emerging as the most robust choice for production deployment.

Finally, an interactive Gradio prototype was developed that, upon selecting a customer and optionally adjusting their payment parameters, calculates in real time the probability of default for a new invoice. This tool is proposed as a strategic asset for proactive receivables management, reducing delinquency risk and strengthening the company's financial sustainability.

Keywords: *predictive model, default risk, machine learning, receivables management, financial risk.*

ÍNDICE GENERAL

ABREVIATURAS

ESPOL	Escuela Superior Politécnica del Litoral
AI	Artificial Intelligence (Inteligencia Artificial)
PYME	Pequeña y mediana empresa
RFM	Recencia, frecuencia, poder monetario
AUC-ROC	Area Under the Receiver Operating Characteristic Curve (área bajo la curva ROC)
BI	Business Intelligence (Inteligencia de Negocios)
DAX	Data Analysis Expressions (Expresiones de Análisis de Datos)
F1-Score	F1 Score (Puntuación F1, métrica de rendimiento)
FN	Predicción incorrecta de casos negativos (False Negative)
FP	Predicción incorrecta de casos positivos (False Positive)
KPIs	Key Performance Indicators (Indicadores Clave de Desempeño)
LightGBM	Light Gradient Boosting Machine (Algoritmo eficiente de boosting)
ML	Machine Learning (Aprendizaje Automático)
PCA	Principal Component Analysis (Análisis de Componentes Principales)
SMOTE	Synthetic Minority Over-sampling Technique (Técnica de Sobremuestreo de Minorías Sintéticas)
SQL	Structured Query Language (Lenguaje de Consulta Estructurada)
TN	Predicción correcta de casos negativos (True Negative)
TP	Predicción correcta de casos positivos (True Positive)
XGBoost	Extreme Gradient Boosting (Modelo de optimización de gradiente extremo)

SIMBOLOGÍA

TM	tonelada métrica
mio USD	millones de dólares de los Estados Unidos de América
α	nivel de significancia (usado en pruebas estadísticas)
μ	media o valor promedio
σ	desviación estándar
Σ	suma total de un conjunto de valores
X	variable independiente o característica en el modelo
Y	variable dependiente u objetivo
R^2	coeficiente de determinación (medida de ajuste del modelo)
p-value	valor p (probabilidad asociada a un estadístico)
λ	parámetro de regularización en algunos modelos
θ	parámetro o coeficiente en modelos de regresión
t	estadístico t (usado en pruebas t)
n	tamaño de la muestra o número de observaciones
%	porcentaje
>	mayor que
<	menor que

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE PLANOS

CAPÍTULO 1

1. PLANTEAMIENTO DE LA PROBLEMÁTICA

1.1 Descripción del problema

El negocio, dedicada a la producción de balanceado para nutrición animal, enfrenta desafíos significativos en la gestión de su cartera y la asignación de cupos de crédito a sus clientes. Uno de los problemas clave es la caída continua del precio del camarón, que afecta especialmente a los pequeños productores. Estos productores han luchado por mantener su rentabilidad en un contexto de márgenes de ganancia decrecientes y altas tasas de interés. Como resultado, ha restringido el acceso al crédito para este segmento, lo que ha llevado a una destacada expansión de la cartera vencida y una baja recuperación de los créditos otorgados.

La ausencia de un sistema dinámico para asignar cupos de crédito provoca el bloqueo de la mayoría de las órdenes de venta por incumplir los requisitos crediticios. Esto genera un trabajo operativo adicional, ya que es necesario desbloquear las órdenes, lo que interrumpe la entrega de productos a los clientes y afecta la eficiencia operativa.

1.2 Justificación del problema

Las condiciones macroeconómicas descritas previamente han creado un desafío significativo para los productores camaroneros de pequeña escala. La restricción en el acceso al crédito ha generado una inflación en la cartera y una baja recuperación de los créditos otorgados, lo que afecta negativamente la salud financiera. La falta de un sistema de asignación de cupo para los clientes ha provocado bloqueos en las órdenes de venta y una interrupción en la entrega de productos, impactando en la experiencia del cliente.

La implementación de un modelo predictivo permitirá al negocio mejorar la precisión en la asignación de cupo, identificar patrones de comportamiento de los clientes y evaluar el riesgo crediticio de manera más efectiva. Esto optimizará la gestión de la cartera y aumentará la recuperación de créditos, mejorando la eficiencia operativa al reducir los bloqueos en las órdenes de venta y garantizando una entrega oportuna de productos

1.3 Solución propuesta

Este proyecto se desarrollará y evaluará en Python, usando Google Colab como entorno de prototipado. Las fases clave incluirán la extracción de datos, el análisis exploratorio y el preprocesamiento en Python, el entrenamiento y validación de los modelos en Colab, y el prototipado interactivo con Gradio.

1.3.1 Componentes principales de la solución

- Modelo predictivo de riesgo crediticio: Se desarrollará un modelo predictivo que estime la probabilidad de que un cliente no cumpla con los plazos de pago estipulados. Este modelo se entrenará con datos históricos de clientes, patrones de pago y condiciones de crédito, y se evaluarán algoritmos como regresión logística, árboles de decisión, Random Forest y LightGBM para seleccionar el más adecuado.
- Asignación óptima de cupos de crédito: Con base en las predicciones generadas, se implementará un módulo de asignación de cupos de crédito que ajustará los límites crediticios según el perfil de riesgo y el historial de pago de cada cliente. El negocio podrá ajustar los cupos en tiempo real, maximizando la utilización del límite disponible sin comprometer la recuperación.
- Integración con el ecosistema de Microsoft: Posterior despliegue en Power Platform como opción a producción, aprovechando Power BI como tablero de control y se diseñará un formulario en Power Apps que replique los campos del prototipo en Gradio. Seguidamente, un servicio en Python ejecutará el flujo de preprocesamiento y aplicará el modelo para calcular la probabilidad de impago. Los resultados se mostrarán de forma interactiva, tanto en informes de Power BI como embebidos en la aplicación de Power Apps.
- Ejecutabilidad en tiempo real: El sistema se diseñará para ejecutarse automáticamente al registrar un nuevo pedido. Cuando se procese la orden de venta, se activará el modelo para evaluar el riesgo de morosidad y permitir ajustar las condiciones de crédito antes de aprobar la factura.
- Dashboard interactivo: Se desplegará un dashboard de monitoreo y análisis dentro de Power Platform que permitirá visualizar en tiempo real el comportamiento de la cartera y el desempeño de las ejecuciones del modelo. A

través de esta interfaz, se podrá identificar de inmediato a los clientes con mayor probabilidad de impago o cuyas predicciones hayan caído recientemente en mora, así como filtrar y segmentar por periodos, condiciones de pago o grupos de clientes para detectar tendencias. Además, se incorporarán gráficas de importancia de variables que señalarán los factores más influyentes en cada predicción, y se configurarán alertas automáticas para notificar cuando un cliente supere un umbral de riesgo o cuando el porcentaje de facturas vencidas exceda niveles predefinidos. Con esta solución, los stakeholders contarán con una herramienta proactiva para anticipar problemas, ajustar cupos de crédito y optimizar la gestión de la cartera antes de autorizar nuevos pedidos.

1.3.2 Beneficios Esperados:

- Optimizar de la asignación de crédito al ajustar proactivamente los cupos de crédito con base en predicciones precisas de riesgo, maximizando la utilización de la capacidad crediticia disponible.
- Reducir de los bloqueos de órdenes de venta al permitir el procesamiento de órdenes de cliente sin interrupciones innecesarias, lo que incrementará la eficiencia operativa y minimizará retrasos.
- Mejorar en la recuperación de créditos al identificar clientes con alto riesgo de morosidad, se enfocarán los esfuerzos de cobranza de manera más efectiva, aumentando las tasas de recuperación.
- Escalabilidad y mantenimiento sin intervención manual al integrar el servicio de Python con Power Platform y los sistemas corporativos, de modo que los datos se actualizarán y procesarán automáticamente tras cada transacción, garantizando un flujo escalable y libre de intervención manual.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar un sistema predictivo que estime la probabilidad de impago al emitir una factura en un negocio de nutrición animal para mejorar la gestión de la cartera.

1.4.2 Objetivos Específicos

1. Analizar la información histórica de la cartera utilizando técnicas de análisis exploratorio de datos, buscando comprender el comportamiento de pago de los clientes, identificar patrones de comportamiento y determinar los factores que influyen en la recuperación de créditos.
2. Desarrollar modelos predictivos de riesgo crediticio empleando modelos estadísticos y técnicas de aprendizaje automático para evaluar el riesgo crediticio de los clientes. Se explorarán algoritmos como regresión logística, árboles de decisión, entre otros, para identificar el modelo más adecuado.
3. Evaluar el desempeño de los modelos para estimar la probabilidad de impago de una factura, considerando su perfil de riesgo, capacidad de pago, comportamiento histórico, entre otros; maximizando la utilización de la capacidad crediticia disponible sin comprometer la recuperación de créditos.
4. Proponer un prototipo de inferencia que permita anticipar impagos al emitir facturas.

1.5 Metodología

Este proyecto se fundamentará en el uso de técnicas tradicionales de análisis de datos y machine Learning para desarrollar un sistema predictivo de asignación de cupos de crédito. A continuación, se describen las fases clave del proceso, desde la extracción y análisis de los datos hasta la evaluación del prototipo en Gradio y su futura integración en Power Platform.

1.5.1 Fuentes de Datos

Los datos utilizados en este proyecto corresponden a un conjunto simulado, construido a partir de estructuras y variables comúnmente empleadas en la gestión de cartera, facturación, pagos y cuentas por cobrar en negocios del sector de nutrición animal. Para el análisis y la visualización preliminar, este dataset fue procesado en un entorno de Python utilizando herramientas estándar de ciencia de datos.

1.5.2 Análisis Exploratorio de Datos (EDA)

Etapa fundamental para comprender la estructura y el comportamiento de los datos antes de la construcción de los modelos predictivos. En esta fase, se explorarán variables clave como el historial de pago de los clientes, los días de vencimiento, el saldo adeudado, y otros factores que pueden influir en el riesgo de morosidad. Se realizarán visualizaciones en Python para identificar patrones y relaciones entre las variables, como la distribución de los datos y la presencia de valores atípicos.

Este análisis facilitará la construcción de modelos más precisos al proporcionar una visión clara de los factores que afectan la morosidad de los clientes.

1.5.3 Preprocesamiento de Datos

Se integrará un script Python que aplique limpieza, imputación y codificación sin necesidad de herramientas externas. Este proceso incluirá:

- **Limpieza de datos**: Eliminación de registros duplicados, imputación de valores faltantes y corrección de errores.
- **Codificación de variables categóricas**: Transformación de las variables categóricas en un formato adecuado para los algoritmos de machine Learning, utilizando técnicas como one-hot encoding con scikit-learn.
- **Escalado de variables numéricas**: Se utilizará StandardScaler para homogeneizar rangos y valores numéricos.

1.5.4 Desarrollo de Modelos Predictivos

Se desarrollarán modelos predictivos que permitan estimar el riesgo de morosidad de los clientes y la probabilidad de que paguen sus facturas en el plazo estipulado. Para ello, se utilizarán algoritmos de machine Learning tradicionales como:

- **Regresión Logística**: Este modelo es ideal para la clasificación binaria, permitiendo predecir si una factura será pagada o no, en función de los patrones históricos de pago de los clientes. La regresión logística ayudará a identificar los factores más influyentes en el comportamiento crediticio de los clientes.
- **Árboles de Decisión**: Este algoritmo facilita la segmentación de los clientes en categorías de riesgo y es particularmente útil para interpretar de manera sencilla los resultados obtenidos, además de permitir una fácil implementación dentro de los sistemas de gestión de créditos existentes.

1.5.5 Validación Cruzada y Ajuste de Hiperparámetros

Para garantizar la precisión y robustez de los modelos, se aplicará **validación cruzada**, lo que permitirá evaluar su desempeño en diferentes subconjuntos del dataset. Adicionalmente, se utilizarán técnicas de ajuste de hiperparámetros como **búsqueda en cuadrícula (grid search)** para optimizar los parámetros clave de los modelos, como la profundidad de los árboles en los árboles de decisión o los coeficientes de regularización en la regresión logística.

1.5.6 Evaluación del Rendimiento del Modelo

El rendimiento de los modelos será evaluado mediante el uso de métricas adecuadas para problemas de clasificación, tales como:

- **Precisión**: Mide la proporción de predicciones correctas.
- **Recall**: Evalúa la capacidad del modelo para identificar correctamente las facturas que no serán pagadas.
- **F1-score**: Combina precisión y recall en una sola métrica para ofrecer una evaluación equilibrada del rendimiento del modelo.
- **Área bajo la curva ROC (AUC-ROC)**: Mide la capacidad del modelo para diferenciar entre las clases de pago y no pago, ofreciendo una visión integral de su desempeño.

1.5.7 Monitoreo y Mantenimiento del Modelo

El modelo será monitoreado continuamente para asegurar que siga siendo preciso y relevante. Se establecerán indicadores clave de rendimiento (KPIs) para evaluar el impacto de las predicciones en la gestión de la cartera, y se realizarán ajustes al modelo según sea necesario. Además, se implementarán actualizaciones periódicas de los datos y ajustes en los hiperparámetros para asegurar que el modelo mantenga un alto nivel de precisión a lo largo del tiempo.

1.5.8 Próximos pasos para producción

Se construirá un prototipo de inferencia en Gradio dentro del mismo notebook de Colab. Este prototipo hace:

- **Carga de datos de cliente/factura:** Se conecta a una base con información histórica de clientes a la fecha actual.

- **Preprocesamiento:** aplica el pipeline entrenado (imputación, escalado, one-hot encoding) exactamente igual que en la fase de entrenamiento.
- **Inferencia:** carga el archivo pickle con el modelo, genera las probabilidades de impago y el indicador binario.
- **Interfaz Gradio:** muestra en tiempo real la probabilidad y el flag de impago, y permite editar cualquier campo para probar distintos escenarios.

Para actualizar el modelo, se reemplaza el archivo pickle con los nuevos pesos; el prototipo de Gradio cargará automáticamente la nueva versión en la próxima ejecución.

1.6 Resultados esperados

El resultado final será un informe interactivo implementado en Power Platform, integrado con las herramientas del ecosistema de Microsoft. El sistema predictivo estará diseñado para realizar predicciones sobre la asignación de cupos de crédito y el riesgo de morosidad de los clientes, lo cual se actualizará de manera automática cada vez que se carguen nuevos datos de la cartera en el sistema.

Un informe en Power BI permitirá a las partes interesadas acceder a análisis en tiempo real sobre el comportamiento de pago de los clientes, tomando en cuenta la información histórica disponible. La actualización del dataset activará el modelo predictivo, que calculará la probabilidad de pago o no pago de las facturas pendientes. Con esta funcionalidad, el negocio podrá tomar decisiones más informadas al momento de otorgar créditos, ajustando los cupos de crédito según el riesgo crediticio de cada cliente.

Adicionalmente, el modelo será programado para ejecutarse automáticamente en tiempo real cuando un cliente realice un nuevo pedido. En ese momento, el sistema activará el modelo predictivo que evaluará si el pedido será pagado dentro del plazo establecido o si existe un riesgo de incumplimiento. Esta predicción brindará a los stakeholders la posibilidad de actuar de forma proactiva para mitigar riesgos, permitiendo ajustar las condiciones de crédito en tiempo real.

El código del modelo predictivo se ejecutará dentro de la infraestructura de Power BI, utilizando Power Query para procesar y transformar los datos, y DAX (Data Analysis

Expressions) para generar las predicciones en base a los datos históricos. El informe será diseñado para actualizarse automáticamente a partir de la fuente de datos simulada, reproduciendo la integración que podría aplicarse en un entorno corporativo real.

Al finalizar el proyecto, se proveerá un dashboard interactivo en Power Platform que ofrecerá una visión integral del estado de la cartera, incluyendo las predicciones del modelo sobre el riesgo crediticio.

1.7 Conjunto de datos

Cada fila del conjunto de datos corresponde a una factura emitida durante los últimos tres años. Para cada factura se derivaron variables de comportamiento de pago, características del cliente al momento de emisión y métricas agregadas de su historial crediticio. La variable objetivo es binaria e indica si la factura incurrió en impago (1) o no (0), considerando un umbral de 30 días de retraso.
A continuación, se detallan las 20 variables finales que alimentan el modelo.

Tabla 1. Variables del conjunto de datos

Variable	Descripción	Tipo de dato
invoice_amount	Monto de la factura en moneda local.	flotante
term_days	Plazo pactado de pago (días entre emisión y vencimiento).	entero
num_prev_invoices	Número de facturas previas emitidas al mismo cliente antes de la factura actual.	entero
avg_prev_days_late	Promedio de días de retraso en facturas previas del cliente.	flotante
num_prev_late_invoices	Conteo de facturas previas que se pagaron con al menos 1 día de retraso.	entero
num_prev_paid_31_60d	Conteo de facturas previas pagadas con retraso entre 31 y 60 días.	entero
num_prev_paid_61_90d	Conteo de facturas previas pagadas con retraso entre 61 y 90 días.	entero
avg_prev_invoice_amount	Promedio de monto de facturas previas del cliente.	flotante
stddev_prev_invoice_amount	Desviación estándar del monto de facturas previas.	flotante
min_prev_invoice_amount	Monto mínimo entre las facturas previas.	flotante
max_prev_invoice_amount	Monto máximo entre las facturas previas.	flotante
days_since_last_invoice	Días transcurridos entre la emisión de la última factura previa y la factura actual.	entero

ratio_invoice_to_avg	Relación entre el monto de la factura actual y el promedio de las facturas previas.	flotante
pct_prev_paid_on_time	Porcentaje de facturas previas pagadas puntualmente (0 días de retraso).	flotante
days_since_last_pay	Días transcurridos desde la última fecha de pago hasta la emisión de la factura actual.	entero
avg_hist_days_late	Promedio histórico de días de retraso (todas las facturas previas).	flotante
payment_terms_code	Código de condición de pago del cliente (codificada mediante one-hot encoding).	categórica codificada
payment_method_code	Código de método de pago del cliente (codificada mediante one-hot encoding).	categórica codificada
customer_group_code	Código de grupo al que pertenece el cliente (codificada mediante one-hot encoding).	categórica codificada
state	Estado o región del cliente (codificada mediante one-hot encoding).	categórica codificada
label_default	Variable objetivo: 1 si la factura incurre en impago (retraso ≥ 30 días), 0 en caso contrario.	binaria

Después del preprocesamiento, el conjunto de entrenamiento quedó compuesto por 67 variables de entrada: 15 características numéricas estandarizadas y 52 variables binarias generadas mediante one-hot encoding de las columnas categóricas. Este juego de variables finales, libre de valores faltantes y preparado según el pipeline de transformación, es el que utiliza el modelo Random Forest para estimar la probabilidad de impago.

CAPÍTULO 2

2. ESTADO DEL ARTE

2.1 Estatus actual del sector camaronero y su acceso a crédito

El sector camaronero es uno de los pilares fundamentales de la economía ecuatoriana, representando una parte significativa del Producto Interno Bruto (PIB) y de las exportaciones del país. En los últimos años, este sector ha enfrentado diversos desafíos, como la fluctuación de los precios internacionales, la creciente competencia global y el aumento de costos de materia prima, agravado por el conflicto bélico entre Ucrania y Rusia. Según CFN (2023), aunque el sector exportó 14% más en volumen en comparación al año anterior, los ingresos en facturación se redujeron un 5%, lo que refleja las dificultades generadas por la baja de precios internacionales.

A nivel gubernamental, han surgido varias iniciativas en busca de mitigar la falta de liquidez en el sector, mediante líneas de crédito específicas gestionadas por BanEcuador y la Corporación Financiera Nacional (CFN). Estas han ofrecido montos de hasta 200 mil dólares para asociaciones camaroneras, dependiendo de su grado de afectación y necesidades productivas. A estas iniciativas, se suman los esfuerzos de la banca privada, con banco Produbanco a la cabeza, el cual desde 2022 ha destinado más de 27 millones de dólares en financiamiento al sector camaronero, con enfoque en la adopción de tecnologías ecoeficientes, prácticas sostenibles y automatización. (Produbanco, 2022).

A pesar de estos esfuerzos, el sector sigue enfrentando desafíos importantes. En 2023, la participación el sector camaronero en el PIB alcanzó un 2.08%, lo que representa un crecimiento en comparación con años anteriores, consolidando aún más su relevancia económica en el país.

Año	VAB Acuicultura y pesca de camarón (Millones USD de 2007)	PIB Total (Millones USD de 2007)	Participación PIB
2019 p	933.85	71,879.22	1.30%
2020 p	997.25	66,281.55	1.50%
2021 prev	1,158.67	69,088.74	1.68%
2022 prev	1,295.14	71,125.24	1.82%
2023 prev	1,499.11	72,164.26	2.08%

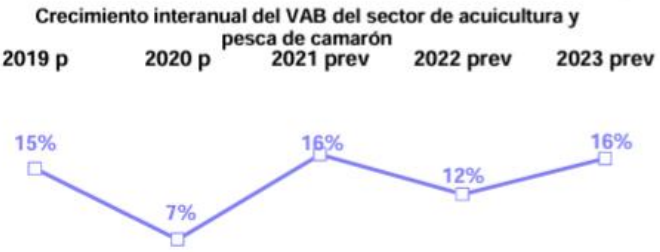


Ilustración 1. Ficha Sectorial Camarón. (2024). PIB sector acuicultura.

Fuente: CFN

Los productores han sido forzados a adaptarse al contexto económico. Este ajuste ha sido especialmente complicado para los pequeños y medianos productores, que se enfrentan a mayores dificultades para mantenerse rentables y competitivos. Comparando los datos de 2022 y 2024 que la CFN publicó en su Ficha Sectorial Camarón, se evidencia un claro proceso de concentración en el sector, en el que las empresas más grandes han absorbido o desplazado a los productores más pequeños. Por ejemplo, en 2020 existían 1.301 empresas dedicadas a la explotación de criaderos de camarón, de las cuales el 87% eran micro y pequeñas empresas, empleando a más de 6000 personas en conjunto. Sin embargo, para 2023, este número se redujo a 893 empresas. Mientras, las microempresas pasaron de emplear a 2,332 personas a tan solo 1,476. A su vez, las empresas grandes incrementaron su participación en el total de empleados y la producción, demostrando una tendencia hacia la concentración de la capacidad productiva en manos de las compañías con un mayor nivel de capital.

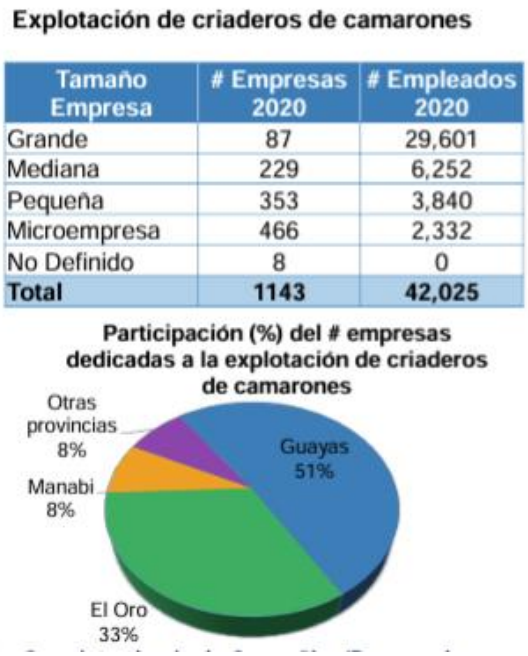


Ilustración 2. Ficha Sectorial Camarón. (2022). Empresas del sector.

Fuente: CFN

Tamaño de Empresa	Explotación de criaderos de camarones	
	# Empresas	# Empleados
Grande	79	14,094
Mediana	163	4,917
Microempresa	439	1,476
Pequeña	212	1,385
Total	**893**	**21,872**

Ilustración 3. Ficha Sectorial Camarón. (2024). Empresas del sector.

Fuente: CFN

Este fenómeno de concentración de mercado en manos de las grandes empresas ha generado una mayor estabilidad para los actores más poderosos, mientras que los pequeños productores, ante la pérdida de rentabilidad, se ven obligados a vender sus hectáreas productivas, lo que contribuye a la desaparición de estas empresas.

En este contexto, es crucial que las instituciones y empresas camaroneras adapten sus procesos de evaluación de crédito. Los métodos tradicionales, aunque útiles en su momento, ya no son suficientes para capturar la complejidad del sector. En este sentido, se han comenzado a implementar técnicas avanzadas de gestión de crédito basadas en

machine Learning y modelos probabilísticos, que permiten una evaluación más precisa del riesgo crediticio.

En los últimos años han proliferado implementaciones en Python para la evaluación de riesgo crediticio en la literatura. Por ejemplo, (Fernández & González, 2021) muestran el uso de scikit-learn para Random Forest en PYMEs financieras, mientras que (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) introdujeron SMOTE como técnica de sobremuestreo para clases desbalanceadas. Además, (Taleb, Ruiz, & Sanchez, 2022) documentan prototipos interactivos con Gradio para validación temprana de modelos en entornos de negocio.

2.2 El desafío del riesgo en el sector camaronero

El acceso al crédito en la acuicultura camaronera enfrenta retos específicos derivados de la estacionalidad de la producción, los ciclos de cobro prolongados y la heterogeneidad de sus productores (Smith, 2020). Los pequeños y medianos criadores, con bajos volúmenes y escaso historial financiero, exhiben perfiles de riesgo que los scorecards tradicionales no diferencian adecuadamente, elevando la cartera vencida y los bloqueos de órdenes (Fernández & González, 2021). Además, la variabilidad de los precios internacionales y las condiciones climáticas introduce fluctuaciones significativas en la liquidez, lo cual requiere de modelos de evaluación que incorporen series temporales y métricas de comportamiento de pago más dinámicas (Ramírez L. , 2022). En respuesta, la literatura reciente aboga por enfoques de machine learning que mejoran la calibración y la capacidad predictiva en entornos de datos desbalanceados y variables. (Taleb, Ruiz, & Sanchez, 2022)

En el contexto de la segmentación de clientes, el enfoque RFM (Recency, Frequency, Monetary) se ha consolidado como una herramienta fundamental para cuantificar el valor y el comportamiento de cada cliente a partir de tres dimensiones clave: Recency, que mide el tiempo transcurrido desde la última transacción; Frequency, que evalúa la cantidad de interacciones o compras realizadas en un periodo determinado; y Monetary, que refleja el importe económico asociado a esas transacciones (Hughes, 2002). Al combinar estas métricas, es posible identificar segmentos de clientes. Por ejemplo, aquellos con alta frecuencia y recencia baja, que suelen presentar mayor lealtad y valor

potencial, frente a clientes con baja frecuencia y recencia alta, que representan un mayor riesgo de abandono o impago (Pfeifer, 2000).

2.2.1 Técnicas tradicionales en la gestión del crédito y cuentas por cobrar

Históricamente, las entidades financieras han confiado en scorecards estadísticos y programación lineal para evaluar el riesgo crediticio de sus clientes.

1. **Scorecards de crédito:** Asignan a cada solicitante un puntaje basado en una combinación lineal de variables (antigüedad, historial de pagos, ratio deuda/ingresos, etc.). Este score, usualmente entre 300 y 850, permite clasificar rápidamente a grandes volúmenes de clientes, pero depende de la calibración de la función logit y puede subestimar la variabilidad de poblaciones con datos desbalanceados (Shewell, 2024). Estos scorecards ofrecen una evaluación rápida y estandarizada de grandes volúmenes de clientes, lo que es esencial en sectores con alta rotación de ventas, como el retail. Además, pueden ajustarse con base en indicadores adicionales como el comportamiento reciente del cliente, lo que mejora su precisión.
2. **Programación lineal:** Se utiliza en entornos más complejos para optimizar la gestión de cuentas por cobrar, balanceando las decisiones de crédito con las restricciones operativas del negocio. Formula un problema de optimización que maximiza ingresos sujetos a restricciones operativas (presupuestos, capacidad de producción). Aunque útil para determinar límites de crédito agregados, su esquema estático no incorpora información de comportamiento de pago en tiempo real ni ciclos estacionales de la acuicultura (Czopelk, 2013). Un modelo de programación lineal podría plantearse tal como se muestra en la ilustración 4.

1) optimise (maximise) the objective function:

$$P = c_1 \cdot x_1 + c_2 \cdot x_2 + \dots c_n \cdot x_n \quad (1)$$

2) for the following side restrictions:

$$\left.\begin{array}{l} a_{11} \cdot x_1 + a_{12} \cdot x_2 + \dots a_{1n} \cdot x_n \le b_1 \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 + \dots a_{2n} \cdot x_n \le b_2 \\ \cdot \\ \cdot \\ \cdot \\ a_{m1} \cdot x_1 + a_{m2} \cdot x_2 + \dots a_{mn} \cdot x_n \le b_m \end{array}\right\} \quad (2)$$

3) with boundaries:

$$x_j \ge 0 \quad (3)$$

4) assuming:

$$m \le n \quad (4)$$

Ilustración 4. Planteamiento de problema usando programación lineal

En la ilustración 4 muestra un esquema genérico de un problema de programación lineal, donde la función objetivo P, la cual es igual a una combinación lineal de variables o cantidades de interés, está sujeta a restricciones, las cuales vienen dadas por el problema. El objetivo es hallar la combinación de valores de x dentro de la región de solución al problema de optimización que maximiza los ingresos (dadas las condiciones de operación). Este tipo de modelos ha sido utilizado en sectores como la minería para gestionar cuentas por cobrar y mantener un balance óptimo entre las ventas a crédito y la capacidad del negocio para cubrir sus operaciones (Czopelk, 2013).

2.2.2 Soluciones avanzadas en la gestión del crédito y cuentas por cobrar.

Las estrategias tradicionales usadas en la gestión de cuentas por cobrar no logran capturar la complejidad, estacionalidad y variabilidad que existe en el sector camaronero. Por tanto, con la llegada de herramientas de machine learning y big data, las instituciones han incorporado modelos no lineales y técnicas de sobremuestreo para mejorar la predicción de impagos y capturar relaciones complejas entre variables. (Brown, 2019).

1. **Regresión logística:** Sigue siendo un referente para la evaluación de riesgo crediticio debido a su simplicidad y su capacidad de interpretación en entornos altamente regulados como el financiero. Nos permite predecir la probabilidad de

ocurrencia de un evento, como el impago de un cliente en una deuda. A diferencia de modelos más complejos, como los basados en árboles de decisión o redes neuronales, la regresión logística se destaca por ofrecer transparencia en los procesos de toma de decisiones y una mayor facilidad de explicación. En la industria del riesgo crediticio, esta técnica permite generar un "score" basado en la probabilidad de incumplimiento de pago por parte de los clientes. La fórmula básica emplea la función logística para convertir la salida de una combinación lineal de variables en un valor que oscila entre 0 y 1, lo que se puede interpretar como la probabilidad de morosidad.

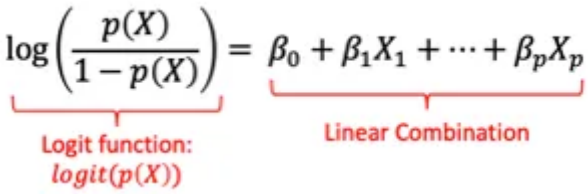


Ilustración 5. Fórmula de la regresión logística

Además, uno de los beneficios más importantes de la regresión logística es su capacidad de asegurar relaciones monotónicas entre las variables predictoras y la variable objetivo, lo que garantiza que el aumento en el riesgo crediticio siempre corresponda con cambios en las características financieras del cliente. Una relación monotónica implica, que a medida que el riesgo del cliente incrementa, la variable predictora debe aumentar o disminuir de forma estricta (Shewell, Medium, 2024).

2. **Árboles de decisión y Random Forest:** son una de las técnicas más utilizadas en la clasificación y predicción dentro de la gestión del riesgo crediticio. Este enfoque destaca por su capacidad para generar reglas comprensibles que pueden aplicarse directamente a los procesos de toma de decisiones. Un árbol de decisión divide los datos en subconjuntos más homogéneos a través de nodos sucesivos, utilizando criterios como la ganancia de información o el índice Gini para evaluar qué características proporcionan el mejor punto de división. En el caso del riesgo

crediticio, las variables clave pueden incluir características demográficas de los clientes, historial de pago y montos adeudados. Esta técnica permite crear reglas claras para la segmentación de deudas, facilitando la identificación temprana de los casos con mayor probabilidad de ser recuperados y aquellos con menor viabilidad, lo que mejora tanto la precisión del modelo como la eficiencia operativa de las empresas de cobranza.

Uno de los beneficios clave del uso de árboles de decisión es su capacidad para transformar los datos de historial de crédito en reglas específicas de acción. En un estudio de Jankowski & Palinski (2024), el árbol de decisión permitió generar 16 reglas estables que ayudan a decidir si se debe continuar con la cobranza amigable o proceder a la vía judicial. Este enfoque permite, por ejemplo, abandonar ciertos casos en las primeras etapas cuando la probabilidad de recuperación es baja, o bien aplicar diferentes estrategias de cobro según las características del deudor, como su historial crediticio o la cantidad adeudada. La capacidad de los árboles de decisión para interpretar los patrones de pago y otras variables relevantes los convierte en una herramienta fundamental en la optimización del proceso de cobranza.

También se tienen las máquinas de vectores de soporte (SVM), una técnica de machine Learning utilizada para la clasificación en diferentes contextos, así como en la evaluación de riesgo crediticio. Esta técnica es particularmente útil para problemas de clasificación con muestras pequeñas, como sucede frecuentemente en los sectores industriales donde los datos disponibles son limitados. En el caso del análisis de crédito, SVM busca encontrar un hiperplano óptimo que maximice la separación entre los clientes solventes y los morosos, basándose en múltiples características financieras. El poder de SVM radica en su capacidad para manejar datos no lineales, mediante el uso de funciones kernel que permiten transformar los problemas en un espacio de mayor dimensión, donde las clases se vuelven más fáciles de separar.

Sum (2016) menciona el uso de SVM para evaluar el riesgo crediticio de las cuentas por cobrar, demostrando una alta precisión predictiva en comparación con otros modelos, como la regresión logística. Esto se debe a la capacidad del modelo para adaptarse

mejor a los conjuntos de datos pequeños, asegurando un buen ajuste incluso en entornos complejos y variables. Un aspecto clave del modelo es la aplicación de técnicas como el análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos, mejorando así la eficiencia y precisión del proceso de clasificación especialmente en casos donde el número de variables es grande.

Aunque SVM y PCA ofrecen ventajas teóricas para reducción de dimensionalidad y separación de clases en espacios de alta dimensión, en este proyecto hemos optado por Random Forest como modelo final. Su capacidad de proveer métricas claras de importancia de variable facilita la interpretabilidad de cada predicción, y su robustez frente a datos desequilibrados y heterogéneos lo hace especialmente adecuado para implementaciones en Google Colab. De esta manera, combinamos una alta performance predictiva con explicabilidad, clave para la adopción de la solución por parte de los stakeholders.

En cuanto a la segmentación de clientes y la predicción de riesgo, los algoritmos de k-means también han sido implementados para agrupar clientes en clústeres según su comportamiento de pago. Este método divide los clientes en grupos homogéneos, lo que permite a las empresas personalizar las políticas de crédito según el perfil de riesgo de cada grupo. El algoritmo k-means busca minimizar la distancia entre los puntos de cada clúster y su centroide, logrando así una segmentación más precisa.

En los últimos dos años se han generalizado en el sector finanzas y agroindustria las implementaciones en Python usando scikit-learn, XGBoost (López, 2024), junto con técnicas de sobremuestreo (SMOTE) (Ramírez J. &., 2023) para mitigar el desbalance de impagos. Asimismo, el uso de prototipos web ligeros ha acelerado la validación con usuarios de negocio. (Vega, 2022)

Finalmente, el Gradient Boosting y XGBoost son técnicas avanzadas de machine learning que ha sido ampliamente aplicadas en la predicción del comportamiento de pago debido a su capacidad para mejorar la precisión de los modelos. Gradient Boosting se basa en la combinación secuencial de predicciones realizadas por múltiples modelos

débiles, típicamente árboles de decisión. A lo largo de varias iteraciones, el algoritmo ajusta el peso de los modelos en función de los errores de las iteraciones anteriores, optimizando de esta manera el rendimiento predictivo general. Este proceso tiene como objetivo reducir gradualmente los errores de predicción y mejorar la precisión del modelo final.

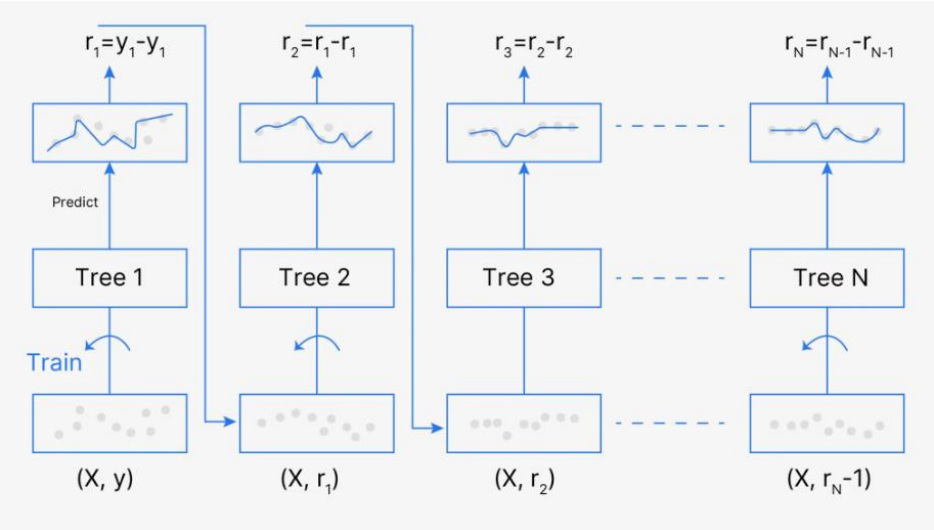


Ilustración 6. Diagrama Gradient Boosting.

En la práctica, Gradient Boosting construye un modelo inicial a partir de los datos originales y luego entrena modelos sucesivos, donde cada uno trata de corregir los errores residuales del modelo anterior. Como se muestra en la imagen, el modelo comienza con un conjunto de datos (X, y), donde y representa las etiquetas o valores reales, y genera un primer árbol de decisión. Posteriormente, se calcula el error residual $r_1=y-y_1$, que es la diferencia entre los valores predichos y_1 y los valores reales y. Este residual se utiliza para entrenar el segundo árbol de decisión. Este proceso se repite hasta que se alcanza el número de iteraciones definido, con cada árbol sucesivo ajustándose a los errores residuales del árbol anterior.

Gradient Boosting ha mostrado un desempeño excepcional en la predicción del comportamiento de pago de usuarios, alcanzando precisiones de hasta el 95% en ciertos conjuntos de datos de comportamiento de pago. Esta alta precisión se debe en gran medida a la capacidad del algoritmo de capturar las variaciones complejas que no pueden ser identificadas por modelos más simples como la regresión logística o incluso los árboles de decisión individuales (Vora, Choudhary, Kumar, & Kadam, 2024).

Por otro lado, XGBoost (eXtreme Gradient Boosting) es una implementación mejorada de Gradient Boosting que ha demostrado ser altamente eficiente y precisa. XGBoost optimiza el proceso de Gradient Boosting mediante técnicas como la regularización para evitar el sobreajuste, el uso de múltiples núcleos de CPU para acelerar el entrenamiento, y la implementación de algoritmos de búsqueda más avanzados que permiten una mayor optimización de los parámetros. XGBoost ha ganado popularidad debido a su capacidad para manejar grandes volúmenes de datos y su flexibilidad para ajustarse a distintos problemas de clasificación y regresión Se basa en la combinación secuencial de predicciones realizadas por múltiples modelos débiles, típicamente árboles de decisión. A lo largo de varias iteraciones, el algoritmo ajusta el peso de los modelos en función de los errores de las iteraciones anteriores, optimizando de esta manera el rendimiento predictivo general. Este proceso tiene como objetivo reducir gradualmente los errores de predicción y mejorar la precisión del modelo final. (Chent, 2016).

2.2.3 Métricas de evaluación

Para evaluar el desempeño de los modelos de clasificación desarrollados en este proyecto, se han utilizado métricas ampliamente aceptadas en problemas de clasificación binaria. Estas métricas, derivadas de la matriz de confusión, permiten analizar tanto la precisión general del modelo como su capacidad para minimizar errores críticos.

La matriz de confusión es una herramienta fundamental para evaluar modelos de clasificación, ya que proporciona una representación tabular de las predicciones del modelo frente a los valores reales. En un problema de clasificación binaria como el presente, donde se predice si una factura será morosa (clase positiva) o pagada (clase negativa), la matriz de confusión incluye las siguientes categorías:

- **Verdaderos Positivos (TP)**: Facturas correctamente predichas como morosas.
- **Falsos Positivos (FP):** Facturas predichas como morosas, pero que en realidad fueron pagadas. Este tipo de error implica una pérdida de oportunidades de venta, ya que clientes solventes son clasificados erróneamente como riesgosos.
- **Verdaderos Negativos (TN):** Facturas correctamente predichas como pagadas.

- **Falsos Negativos (FN):** Facturas predichas como pagadas, pero que en realidad resultaron morosas. Este tipo de error tiene consecuencias financieras significativas, como el aumento de la cartera vencida y riesgos de incobrabilidad.

A partir de la matriz de confusión, se calculan las siguientes métricas clave:

- **Precisión (Accuracy):** Mide la proporción total de predicciones correctas, calculada como:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

  Si bien es útil como medida general, su relevancia puede ser limitada en conjuntos de datos desbalanceados, como en este proyecto.

- **Recall (Sensibilidad o Tasa de Verdaderos Positivos):** Evalúa la capacidad del modelo para identificar correctamente las facturas morosas (clase positiva):

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

  En este contexto, el recall es especialmente crítico, ya que un bajo recall aumenta el riesgo de no identificar facturas en mora, lo que puede resultar en pérdidas financieras.

- **Precisión:** Indica la proporción de facturas predichas como morosas que realmente lo son:

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

  Es útil para minimizar falsos positivos, asegurando que no se clasifiquen erróneamente clientes solventes como riesgosos.

- F1 Score: Combina precisión y recall en una sola métrica:

$$F1 - Score = 2 \times \frac{Precision\ X\ Recall}{Precision + Recall}$$

  Es particularmente valiosa en problemas con clases desbalanceadas, ya que proporciona una evaluación equilibrada del modelo.

- **Curva Precisión-Recall.** Para problemas con clases muy desbalanceadas, como el de detección de impagos, la curva precisión-recall (PR) aporta una visión más informativa que la ROC. En ella se observa cómo varía la precisión al cambiar el recall según el umbral de decisión, permitiendo elegir el punto óptimo para maximizar la detección de impagos sin disparar demasiadas falsas alarmas.

(Saito, 2015) demostró que, cuando la clase positiva es rara, la curva PR es más adecuada que la curva ROC tradicional para comparar clasificadores.

2.3 Pipeline de extracción, modelado y prototipado

En el ámbito de la gestión de riesgos crediticios, diversos autores subrayan la importancia de construir flujos de procesamiento de datos reproducibles y modulables que integren limpieza, generación de características y escalado previo al modelado. (Pedregosa, 2011) presenta en la biblioteca scikit-learn la clase Pipeline como estándar de facto para encadenar transformaciones y modelos de forma segura y mantenible; mientras que (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) demostraron la eficacia de SMOTE para corregir desequilibrios de clase en problemas de clasificación raros, como la predicción de impagos. Aplicaciones prácticas en el sector financiero, como las descritas por (Fernández & González, 2021), confirman que estas técnicas, combinadas con prácticas de serialización de modelos (joblib), facilitan la escalabilidad y el mantenimiento continuo de soluciones predictivas basadas en machine learning.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

3.1 Obtención y Validación de Datos

Los datos empleados en este estudio corresponden a un conjunto simulado, diseñado a partir de variables y estructuras que reflejan prácticas habituales de facturación y gestión de cartera en el sector de nutrición animal. Se asumió que estos datos habían pasado por procesos estándar de validación y control de calidad, tal como ocurre en la industria, garantizando así su consistencia y representatividad.

Para construir la base de datos del modelo predictivo, se replicó el proceso habitual de integración de información de facturación y pagos, aplicando técnicas comunes de consulta y transformación de datos, como el uso de Common Table Expressions (CTEs) en SQL, para la combinación de distintas fuentes y la generación de métricas relevantes.

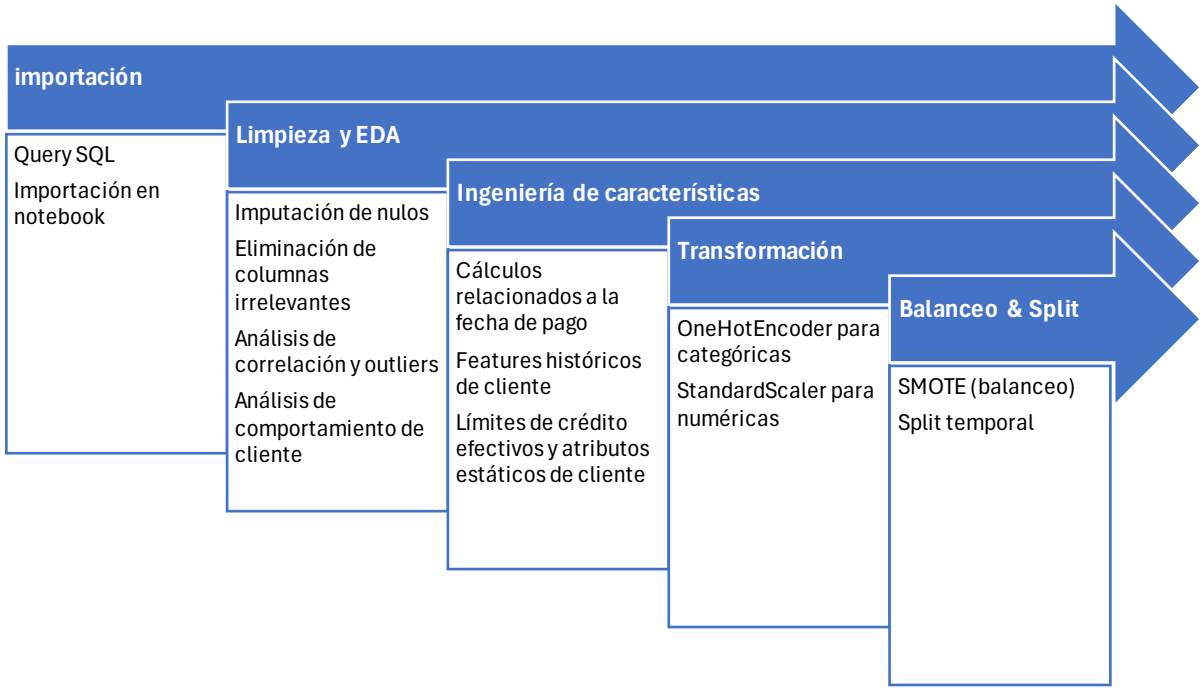
Todo el procesamiento y análisis posterior se realizó utilizando herramientas de ciencia de datos ampliamente aceptadas, como Python y Power BI, simulando un flujo típico de trabajo en el sector.

3.2 Preprocesamiento de los Datos

3.2.1 Pipeline de preprocesamiento de datos

El preprocesamiento de los datos fue crucial para garantizar la calidad y la reproductibilidad del análisis predictivo. El conjunto de datos inicialmente constaba de 186 172 registros y 28 columnas. Este dataset incluye todas las facturas emitidas, enriquecidas con su historial completo de pagos y atributos generales de cada cliente (métricas RFM, límites de crédito y demás máster data), obtenidos mediante la unificación de las tablas de facturación y pagos en la consulta SQL. El pipeline utilizado se resume en la siguiente tabla:

Tabla 2. Pipeline de preprocesamiento de datos.



Para garantizar la coherencia y reproducibilidad de los resultados, el preprocesamiento se articula en cinco fases claramente definidas. Cada fase se implementa en el entorno de Colab tras extraer el dataset ya consolidado por SQL.

Consolidación y carga del dataset: El primer paso consistió en simular el proceso de integración de información relevante para la gestión de cartera, utilizando técnicas estándar de consulta SQL, como las Common Table Expressions (CTEs). Esto permitió combinar datos de facturación, pagos, historial crediticio y datos maestros de clientes, así como calcular métricas como RFM y plazos de crédito efectivos. El dataset resultante fue consolidado en un único archivo para su análisis posterior en el entorno de ciencia de datos. De esta manera, se garantizó que todo el procesamiento subsecuente partiera de una base de datos homogénea y estructurada.

Exploración y diagnóstico inicial: Se inspecciona la estructura y calidad del dataset, cuantificando valores faltantes, revisando tipos de variable y detectando valores atípicos mediante análisis de correlación y diagramas de caja. Este diagnóstico orienta las decisiones de limpieza y transformación posteriores.

Limpieza e imputación:

- **Imputación de nulos:** todas las métricas históricas y montos faltantes se rellenan con cero; la variable que mide la recencia de la última factura sin historial recibe un valor sintético de –1 para diferenciar los casos sin observaciones previas.
- **Eliminación de variables irrelevantes:** se descartan las columnas intermedias que podrían filtrar información directa sobre el estado de pago.

Asimismo, se eliminan aquellas variables con alta correlación identificadas en el EDA, reduciendo redundancias y mejorando la estabilidad del modelo.

Ingeniería y transformación de características: Se construye un pipeline modular que aplica:

- Codificación one-hot a todas las variables categóricas, manejando categorías nuevas o desconocidas de manera segura.
- Estandarización de las variables numéricas (plazos, demoras, sumas y promedios históricos, límites de crédito), asegurando media cero y varianza unitaria.

Este enfoque permite encapsular todas las transformaciones en un único objeto reutilizable en la fase de entrenamiento y en futuros entornos de producción.

Balanceo y partición temporal: Para corregir el desbalance entre facturas pagadas e impagas, se aplica SMOTE únicamente sobre el subconjunto de entrenamiento. A continuación, se realiza un corte temporal: todas las facturas anteriores al 1 de enero de 2023 conforman el conjunto de entrenamiento, y las posteriores se destinan al conjunto de prueba. Este esquema respeta la secuencia cronológica de los datos y reproduce fielmente el escenario real de predicción.

3.2.2 Resultado del preprocesamiento

Tras completar el proceso de preprocesamiento, el conjunto de datos resultante contiene 186,172 registros y 17 columnas, distribuidas de la siguiente manera:

- invoice_amount (float64): monto total de la factura.
- term_days (int64): plazo pactado en días entre emisión y vencimiento.

- num_prev_invoices (int64): número total de facturas previas del cliente.
- avg_prev_days_late (float64): demora promedio de pagas anteriores, en días.
- num_prev_late_invoices (int64): conteo de facturas previas con atraso.
- num_prev_paid_31_60d (int64): facturas previas con retraso entre 31 y 60 días.
- avg_prev_invoice_amount (float64): monto promedio de facturas previas.
- stddev_prev_invoice_amount (float64): desviación estándar del monto de facturas previas.
- min_prev_invoice_amount (float64): monto mínimo de facturas previas.
- max_prev_invoice_amount (float64): monto máximo de facturas previas.
- days_since_last_invoice (int64): días transcurridos desde la factura anterior (–1 si no existe).
- ratio_invoice_to_avg (float64): relación entre el monto actual y el promedio histórico.
- pct_prev_paid_on_time (float64): porcentaje de facturas previas pagadas puntualmente.
- payment_terms_code (object): código del término de pago del cliente.
- payment_method_code (object): código del método de pago.
- customer_group_code (object): código del grupo de cliente.
- state (object): indicador de zona geográfica del cliente.
- days_since_last_pay (int64): días desde el último pago registrado (–1 si no existe).

Nota: en este punto las variables categóricas aún están en tipo objeto; su codificación one-hot se realiza posteriormente dentro del pipeline de entrenamiento.

Este dataset es el resultado de un riguroso análisis de correlación, que permitió eliminar variables redundantes o altamente correlacionadas para mitigar el riesgo de sobreajuste. Una vez definidas las variables resultantes, se procede a abordar un problema típico en los datasets utilizados para la predicción de morosidad: el desbalance de clases. En este caso, el número de facturas impagas y pagadas tienen la proporción mostrada en la figura a continuación:

Tabla 3. Proporción de la variable objetivo en el dataset.

1	10,4%
0	89,6%

Para corregir el desbalance de la variable objetivo en el entrenamiento, se aplicó SMOTE al subconjunto anterior al 1 de enero de 2023. La distribución resultante en el conjunto de entrenamiento quedó en un 50 % de facturas impagas y un 50 % de facturas pagadas, garantizando que el modelo aprenda de ambas clases sin sesgo.

Con estas 18 variables preprocesadas y balanceadas, el dataset está preparado para la fase de entrenamiento, asegurando la consistencia y representatividad necesarias para obtener modelos robustos y generalizables.

3.3 Análisis exploratorio de datos

Se realizó visualizaciones para entender mejor la distribución de las variables y las relaciones entre ellas. Este análisis exploratorio permitió identificar patrones y validar la consistencia de los datos. Se evidencia que en el dataset la cantidad de facturas impagas es mayor que la cantidad de facturas pagadas a la fecha de corte.

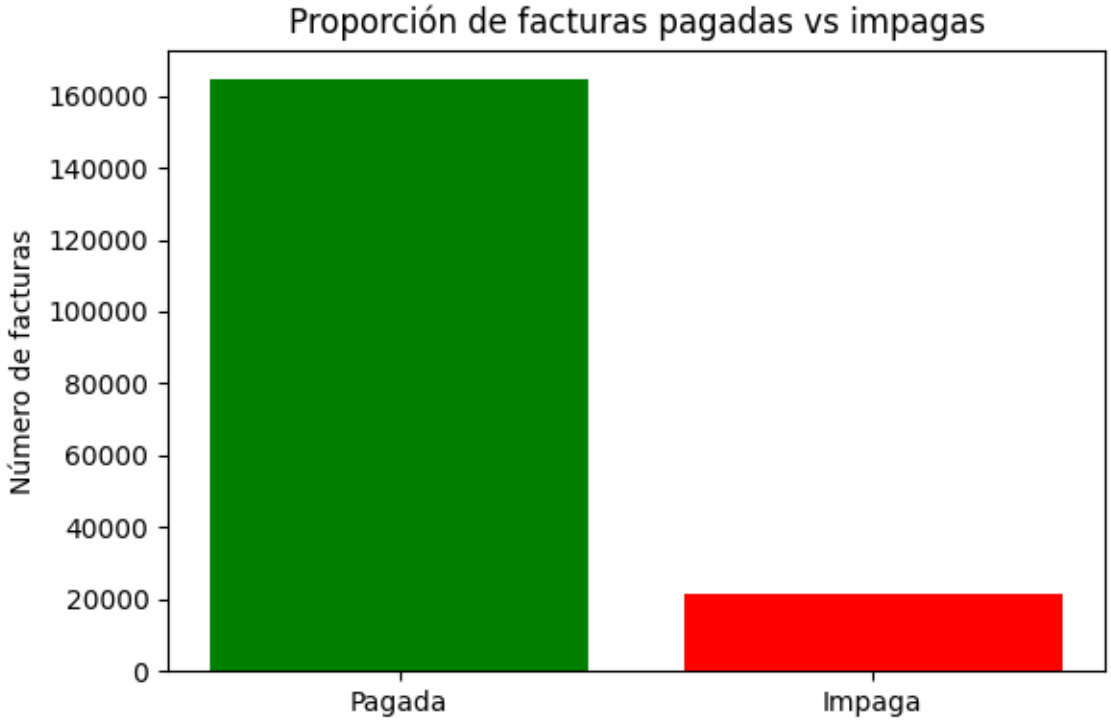


Ilustración 7. Proporción de facturas pagadas vs impagas.

Esto sugiere la necesidad posterior de realizar un balance de clases, de tal manera que durante el entrenamiento no exista un sesgo hacia alguna clase en específico.

Posteriormente se generó un diagrama de dispersión Recency–Frequency coloreado por Monetary en escala logarítmica, con ejes en escala lineal (Ilustración 8). Este tipo de visualización permite evaluar simultáneamente las tres dimensiones del modelo RFM en el conjunto de facturas:

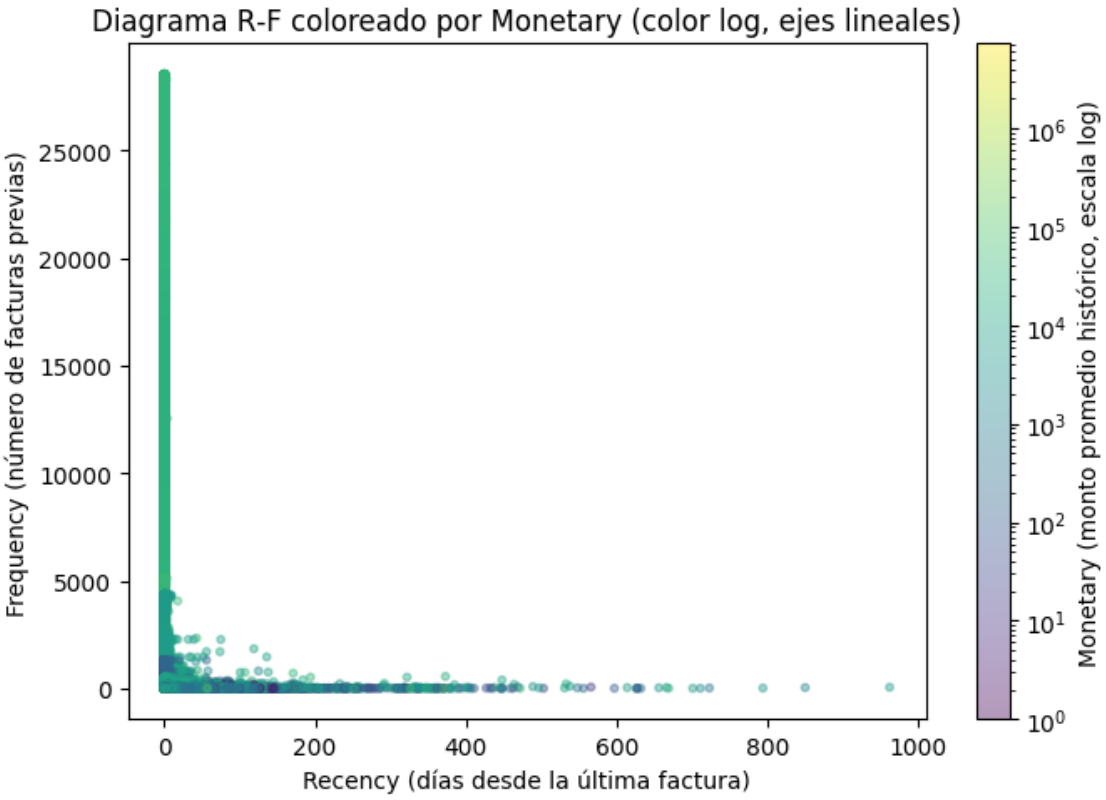


Ilustración 8. Diagrama RFM

La concentración de puntos en la esquina inferior-izquierda (Recency próxima a cero y Frequency baja) revela que la mayoría de los clientes ha realizado pocas facturas y lo ha hecho muy recientemente, lo cual ofrece escasa capacidad discriminatoria por sí solo. En contraste, los clientes de bajo riesgo, que son aquellos con Recency baja (< 30 días) y Frequency alta (> 20 facturas) se distinguen por tonos más claros, indicando montos históricos elevados y, por ende, una probabilidad de impago mínimamente baja. Por otro lado, los clientes de mayor riesgo aparecen en la zona de Recency alta (> 90 días) y Frequency baja (< 3 facturas), con colores oscuros que reflejan bajo Monetary; este patrón sugiere interacciones esporádicas y de bajo valor económico, lo que los posiciona como prioritarios en las estrategias de cobranza.

Adicionalmente, se construyó un mapa de calor de recency × frequency con el promedio de monetary en cada celda (Ilustración 9), que permite identificar de forma agregada los segmentos de mayor valor económico. En este mapa, el cuartil de recency medio (R2) combinado con el tercer cuartil de frequency (F3) registra el mayor promedio de monetary (≈ 45 000), lo que sugiere que los clientes que compran con periodicidad moderada y frecuencia relativamente alta constituyen el núcleo de facturación más rentable. Asimismo, los segmentos R1–F2 y R1–F3 también muestran valores elevados, confirmando que una alta frecuencia de facturación, incluso con recencia mínima, incrementa sustancialmente el valor histórico. Este análisis refuerza la estrategia de focalizar esfuerzos de crédito y fidelización en aquellos clientes que, aun sin ser los más recientes, mantienen un patrón de compras recurrente y de alto importe.

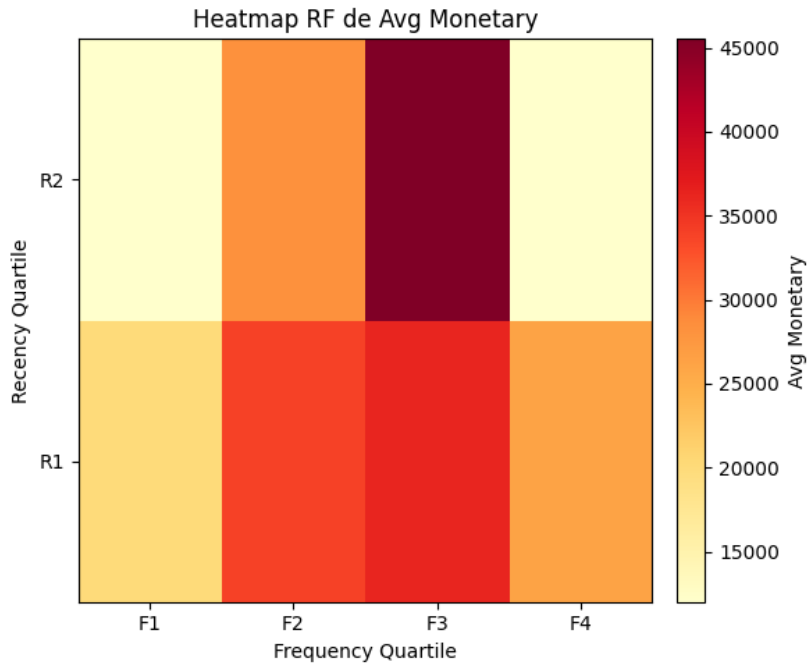


Ilustración 9. Mapa de Calor RFM

3.3.1 Análisis de correlación

Se realizó un análisis exhaustivo de la correlación entre variables. Las características con alta correlación se eliminaron para evitar redundancias y reducir el riesgo de sobreajuste. Se consideró como alta correlación todos aquellos valores muy cercanos a

1 y -1. Por ejemplo, variables relacionadas con rangos de vencimiento que estaban altamente correlacionadas con saldo y otras métricas clave fueron descartadas, siendo las siguientes variables descartadas en base a este análisis:

- 'num_prev_paid_1_30d',
- 'sum_prev_days_late'
- ,'stddev_prev_invoice_amount',
- 'num_prev_invoices'
- 'max_prev_invoice_amount'

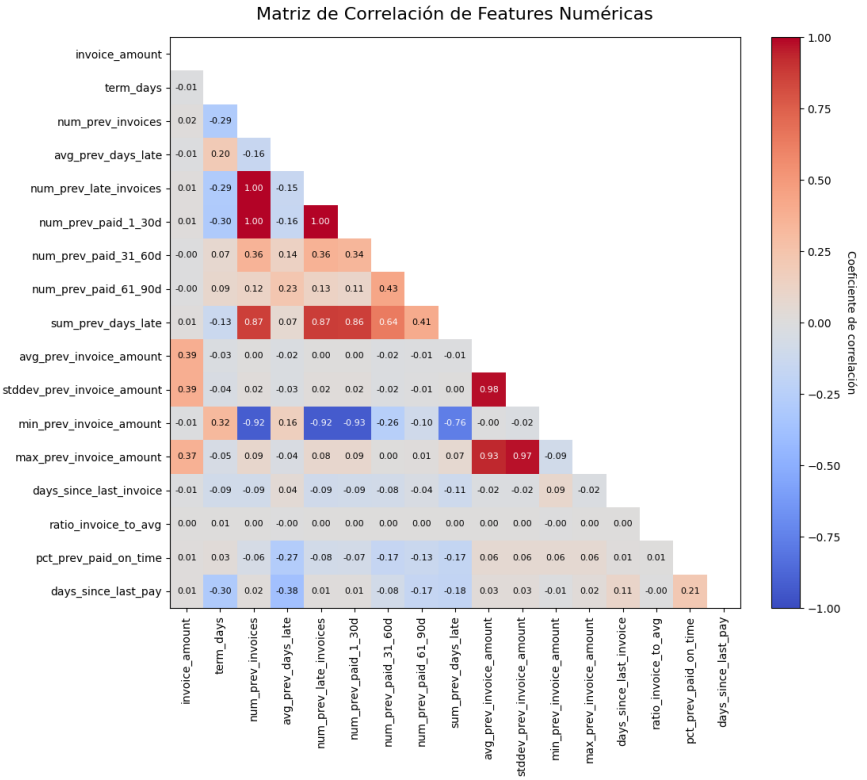


Ilustración 10. Matriz de correlaciones.

3.3.2 Comportamiento de pago de clientes

Para comprender mejor qué tan centrada está la morosidad en unos pocos días de retraso y quiénes concentran la mayor parte de los pagos tardíos, se evaluaron varias métricas de Pareto y se sintetizaron las siguientes tres visualizaciones clave:

1. **Pareto de días de retraso (deciles):** Se agruparon todos los retrasos distintos en deciles, calculando el porcentaje acumulado de facturas tardías por día de

demora. Este análisis muestra que el 80 % de los retrasos se concentra en apenas 26 días de demora, y que el 90 % lo cubre hasta los 44 días.

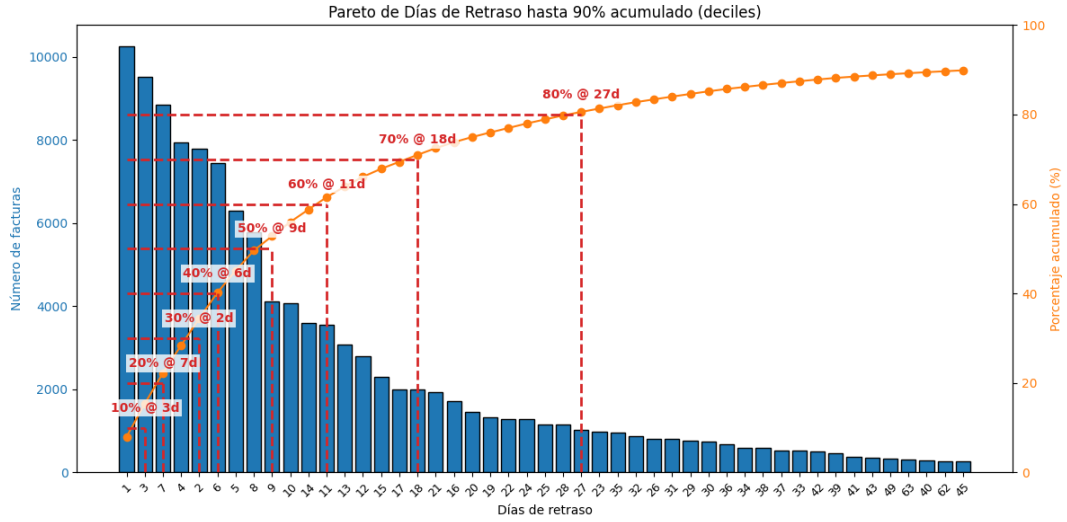


Ilustración 11. Pareto días de retraso

2. **Curva de envejecimiento hasta 60 días:** Considerando que el 80 % de los retrasos ocurre antes de 26 días, se trazó la curva acumulada de morosidad hasta 60 días. Se destacan líneas punteadas en 50 % (9 días), 70 % (17 días), 80 % (26 días) y 90 % (49 días), lo que permite ver cómo a partir del día 26 la curva empieza a aplanarse significativamente.

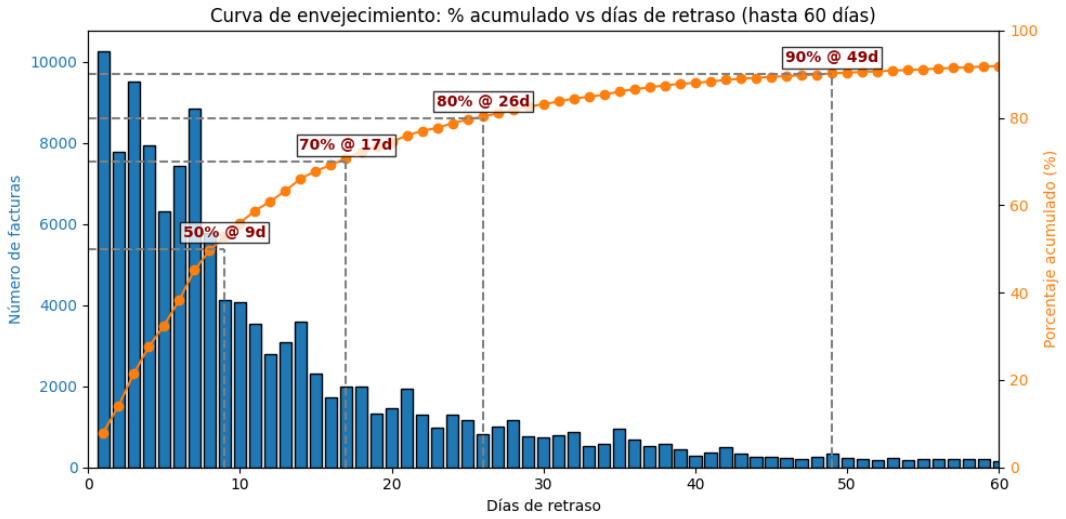


Ilustración 12. Curva de envejecimiento

3. **Gráfico de sectores por rango de demora:** Se agrupó el total de facturas en tres categorías.
    - 0 días (pagadas a tiempo): 30.9 %
    - 1–30 días de retraso: 57.5 %
    - Más de 30 días de retraso: 11.7 %

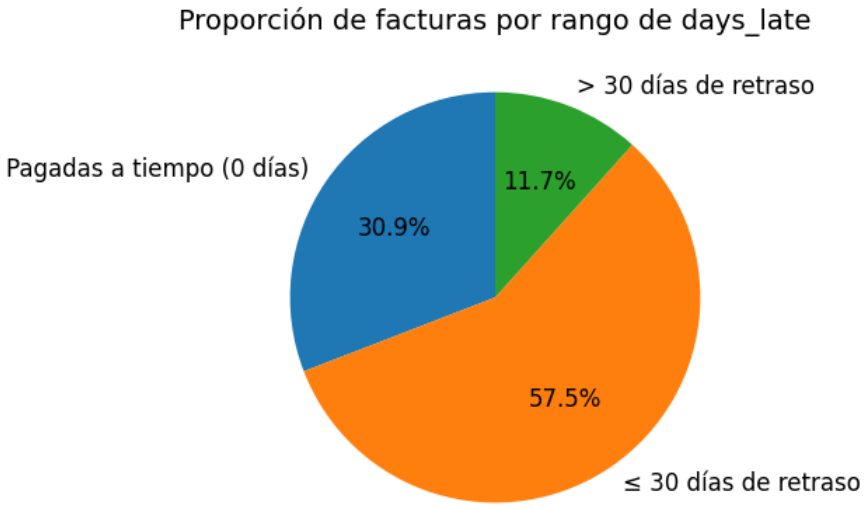


Ilustración 13. Gráfico de sectores por rango de demora

Con base en estos hallazgos y buscando un equilibrio entre cobertura de casos y acción temprana, se estableció un umbral operativo de 30 días para la variable objetivo. Así, la etiqueta se define como:

label_default = 1 si days_late > 30 días
0 en caso contrario

Este criterio permite capturar el 88.4 % de los retrasos (sumando los que superan 30 días) y pone el foco en aquellos casos con más probabilidad de convertirse en incobrables, optimizando los recursos de cobranza.

3.4 Preparación para el entrenamiento

Antes de ajustar los algoritmos predictivos, se consolidó toda la lógica de transformación, balanceo y partición de datos en un pipeline unificado, de modo que el proceso resulte íntegro, reproducible y compatible con producción.

Guardado automático de modelos

Para facilitar la experimentación y el despliegue, se implementó una función para el guardado de modelos, que serializa cada estimador entrenado en un archivo con marca de tiempo. Esta utilidad crea automáticamente un directorio de salida, convirtiendo el objeto entrenado en un archivo formato pickle y almacenándolo en la misma.

Pipeline de preprocesamiento y entrenamiento

Toda la transformación de atributos, el balanceo de clases y el ajuste del modelo se agrupan en un solo pipeline, que consta de tres etapas:

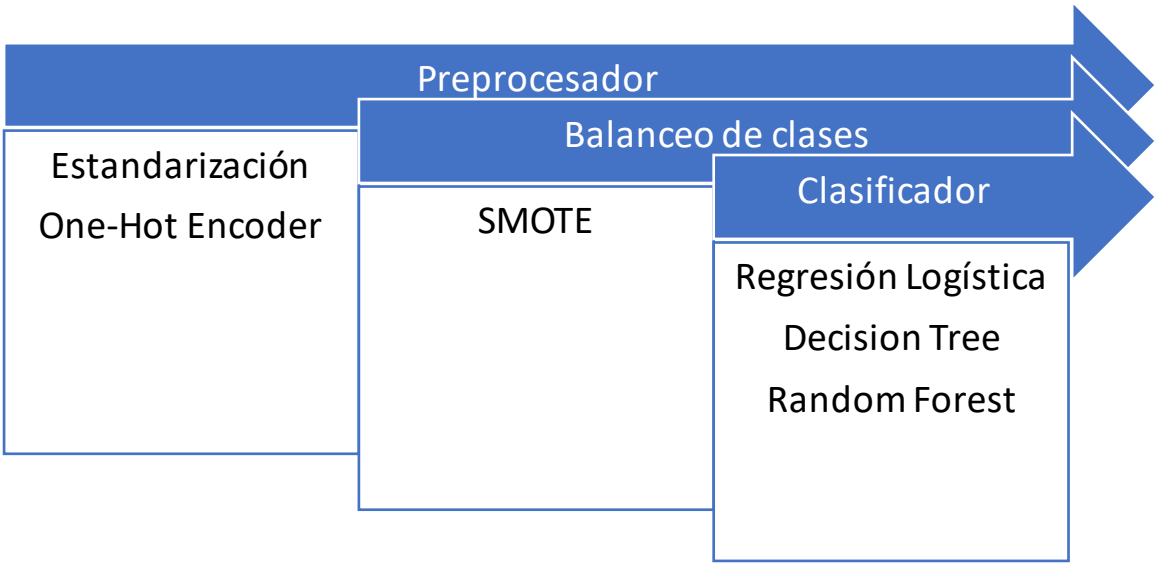


Ilustración 14. Pipeline de Preprocesamiento y entrenamiento

Preprocesador (ColumnTransformer)

- **StandardScaler:** estandariza las variables numéricas para media cero y varianza unitaria, asegurando que ninguna domine por su escala.
- **OneHotEncoder:** codifica las 4 variables categóricas en 52 dummies. Se incorpora el argumento handle_unknown='ignore' para prevenir errores si en producción aparece una categoría nueva.

Balanceo de clases (SMOTE)

Se aplica SMOTE únicamente sobre el conjunto de entrenamiento para generar instancias sintéticas de la clase minoritaria (facturas impagas), equilibrando la proporción de ambas clases sin sobreajuste.

Clasificador

Se utiliza para fijar el algoritmo a entrenar. Iniciamos con Regresión Logística con regularización por defecto y alta tolerancia de iteraciones (max_iter=10000). Este mismo paso del pipeline permite sustituir fácilmente el algoritmo por Random Forest, LightGBM u otros.

Al encapsular estos tres componentes en un solo objeto, basta con llamar a *pipeline.fit(X_train, y_train)* y luego a *pipeline.predict(X_test)* para ejecutar toda la secuencia de forma idéntica y libre de fugas de datos.

Split temporal de datos

Dado el carácter secuencial de las facturas y la necesidad de replicar un escenario real de predicción futura, la partición entre entrenamiento y prueba se realizó mediante un split temporal con fecha de corte el 1 de enero de 2023:

Tabla 4. Split temporal de datos

Fecha de emisión	Tipo de Split
Facturas anteriores a la fecha de corte	entrenamiento
Facturas emitidas a partir de la fecha de corte	prueba

Esta estrategia preserva la dependencia cronológica y evita aprender de datos futuros, ofreciendo una evaluación honesta de la capacidad de generalización del modelo en producción.

3.5 Entrenamiento de modelos

La selección del modelo predictivo adecuado fue una etapa crucial en este proyecto, ya que la meta era desarrollar una herramienta capaz de predecir con precisión el riesgo de

morosidad en las facturas de clientes. Dado el enfoque de clasificación binaria del problema, se exploraron cuatro modelos de aprendizaje supervisado ampliamente reconocidos: Regresión Logística, Árboles de Decisión, Random Forest, y LightGBM.

Se evaluaron cuatro algoritmos de clasificación binaria ampliamente utilizados en problemas de riesgo crediticio: regresión logística, árboles de decisión, Random Forest y LightGBM. Cada modelo fue entrenado sobre el conjunto de datos preprocesado y balanceado, utilizando validación cruzada temporal y ajuste de hiperparámetros mediante GridSearchCV. La regresión logística sirvió como línea base, destacando por su interpretabilidad, aunque con limitaciones para capturar relaciones no lineales. Los árboles de decisión ofrecieron reglas claras de segmentación, pero mostraron tendencia al sobreajuste. Random Forest demostró un equilibrio superior entre precisión y generalización, mientras que LightGBM alcanzó el mayor AUC, aunque con menor interpretabilidad y sensibilidad a la variable cliente.

Los resultados de cada modelo se resumen en las Tablas 5 a 8, donde se presentan sus métricas clave (precisión, recall, F1-score, AUC) y los hiperparámetros óptimos encontrados. Esta comparación permitió seleccionar el modelo más robusto para la implementación final.

Tabla 5. Resultados del modelo regresión logística

Mejores hiperparámetros:
{'clf__C': 0.01,
'clf__penalty': 'l2',
'clf__solver': 'saga'}
Mejor CV AUC: 0.8363720089614975
Test ROC AUC: 0.8072708838906533
Evaluación en el conjunto de prueba:

	precision	**recall**	**f1-score**	**support**
0	0.96	0.71	0.82	128644
1	0.27	0.76	0.40	17540
accuracy			0.72	146184
macro avg	0.61	0.74	0.61	146184
weighted avg	0.87	0.72	0.77	146184x

Tabla 6. Resultados del modelo de árboles de decisión

Mejores hiperparámetros DT:

```
       {'clf__max_depth': 5,
       'clf__min_samples_leaf': 5,
       'clf__min_samples_split': 2}
```

Mejor CV AUC: 0.7621395281496264

Test ROC AUC: 0.7962361608394368

Evaluación en el conjunto de prueba para árboles de decisión:

	precision	recall	f1-score	support
0	0.96	0.68	0.80	128644
1	0.25	0.78	0.38	17540
accuracy			0.70	146184
macro avg	0.60	0.73	0.59	146184
weighted avg	0.87	0.70	0.75	146184

Tabla 7. Resultados del modelo random forest

Mejores hiperparámetros RF:

```
       {'clf__max_depth': 10,
       'clf__min_samples_leaf': 3,
       'clf__min_samples_split': 2,
       'clf__n_estimators': 300}
```

Mejor CV AUC: 0.8377832849030582

Test ROC AUC: 0.858666615367019

Evaluación en el conjunto de prueba para Random Forest:

	precision	recall	f1-score	support
0	0.96	0.79	0.87	128644
1	0.33	0.77	0.47	17540
accuracy			0.79	146184
macro avg	0.65	0.78	0.67	146184
weighted avg	0.89	0.79	0.82	146184

Tabla 8. Resultados del modelo Light GBM

Mejores hiperparámetros Light GBM:

```
       {'clf__learning_rate': 0.05,
       'clf__max_depth': 10,
       'clf__n_estimators': 50}
```

Mejor CV AUC: 0.8629370223860588

Evaluación en el conjunto de prueba para LightGBM:

	precision	recall	f1-score	support
0	0.97	0.76	0.85	128644
1	0.31	0.81	0.45	17540
accuracy			0.76	146184
macro avg	0.64	0.78	0.65	146184
weighted avg	0.89	0.76	0.80	146184

3.5.1 Validación y Ajuste de Hiperparámetros

Para garantizar la fiabilidad de los modelos y respetar la dependencia cronológica de las facturas, se implementó una validación cruzada con TimeSeriesSplit (n_splits=5), de modo que el conjunto de entrenamiento se divide en cinco particiones temporales, entrenando en cada una con datos previos y validando con bloques posteriores. Sobre este esquema, se ejecutó una búsqueda en cuadrícula (GridSearchCV) optimizando la métrica ROC AUC. El grid incluyó parámetros clave como la profundidad máxima de los árboles, el número de estimadores y la tasa de aprendizaje, y se aprovechó n_jobs=-1 para paralelizar el proceso. De este modo, se identificaron las combinaciones de hiperparámetros que maximizan la capacidad discriminativa del modelo sin sobreajuste.

3.6 Evaluación del rendimiento de los modelos

3.6.1 Métricas de evaluación

La evaluación de los modelos se realizó utilizando múltiples métricas que permiten tener una visión integral de su rendimiento. En términos generales, los algoritmos de Random Forest y LightGBM demostraron un mejor desempeño en la predicción del riesgo de incobrabilidad, superando tanto los modelos de regresión logística y árboles de decisión. Las métricas clave consideradas fueron el área bajo la curva AUC-ROC, que mide la capacidad de los modelos para discriminar entre facturas cobrables y morosas, así como el Recall (sensibilidad), la Precisión y la precisión promedio, especialmente relevantes en problemas con clases desbalanceadas. El modelo LightGBM alcanzó un AUC de 0.865, seguido muy de cerca por Random Forest (AUC = 0.859), evidenciando una excelente capacidad predictiva. El rendimiento de la Regresión Logística (AUC = 0.807) y el Árbol de Decisión (AUC = 0.796) fue menor, mostrando limitaciones para capturar la complejidad del problema.

A nivel de métricas de clasificación, la matriz de confusión permitió analizar en detalle los aciertos y errores de cada modelo, distinguiendo entre verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Así, se pudo evidenciar que los modelos más robustos, como Random Forest y LightGBM, lograron un mejor equilibrio entre la identificación correcta de facturas morosas y la minimización de falsos positivos.

3.6.2 Comparación de modelos

Matrices de confusión:

Muestran el desempeño de cada modelo en términos de predicciones correctas e incorrectas. Random Forest y LightGBM destacan por una mayor cantidad de verdaderos positivos (identificación correcta de facturas morosas) y una menor proporción de falsos negativos, lo que se traduce en menor riesgo de dejar pasar facturas que finalmente no se cobrarán. Por su parte, los modelos basados en Regresión Logística y Árbol de Decisión presentan una mayor cantidad de errores, especialmente en la clasificación de facturas morosas.

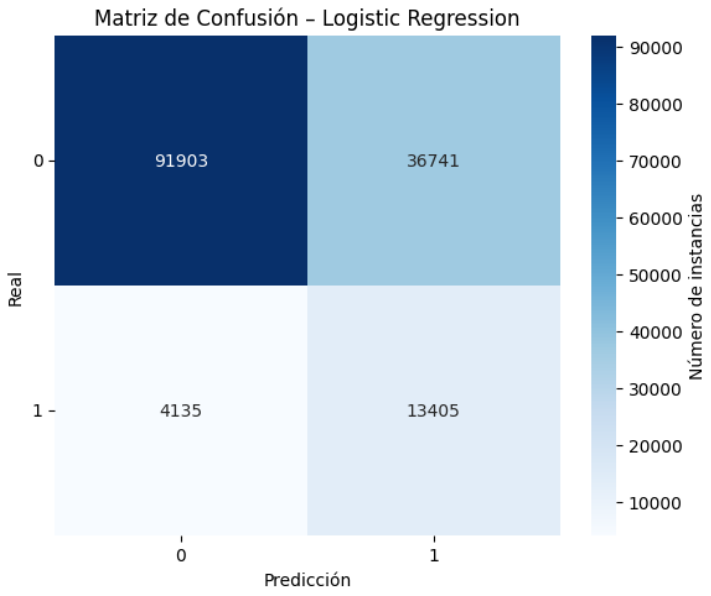


Ilustración 15. Matriz de confusión modelo regresión logística.

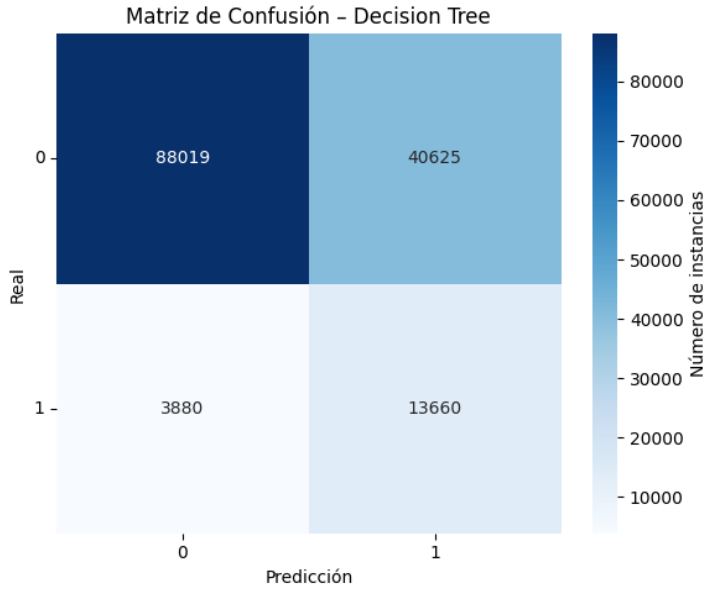


Ilustración 16. Matriz de confusión modelo árboles de decisión.

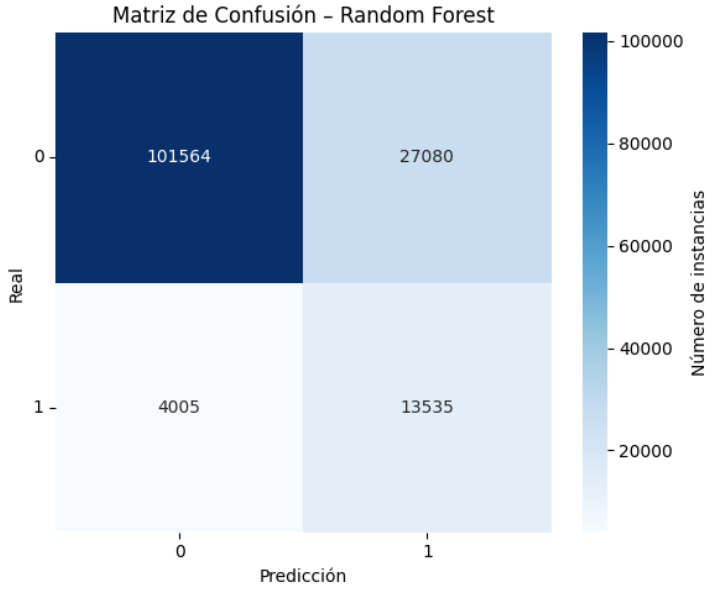


Ilustración 17. Matriz de confusión modelo random forest.

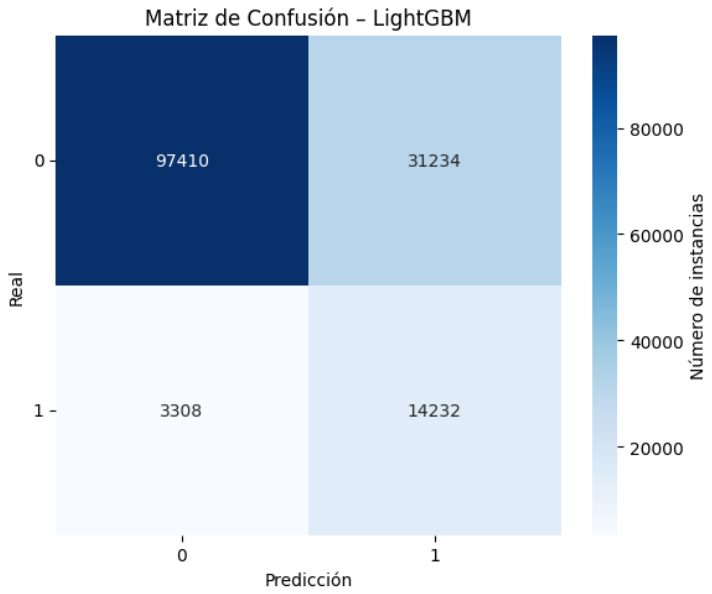


Ilustración 18. Matriz de confusión modelo LightGBM

Curvas ROC:

En la ilustración 19, podemos observar que cuanto más cerca está la curva de la esquina superior izquierda, mejor es el modelo, ya que logra un alto nivel de aciertos identificando morosos y comete pocos errores señalando como morosos a quienes realmente pagan. Los resultados evidencian que LightGBM y Random Forest sobresalen frente a los demás, alcanzando los valores de área bajo la curva más altos, lo que indica una mayor capacidad discriminativa. Es decir, estos modelos son mucho más efectivos diferenciando entre clientes de alto y bajo riesgo de impago, reduciendo tanto los falsos positivos como los falsos negativos en comparación con los modelos más simples.

En conclusión, una mayor área bajo la curva ROC refuerza que estos algoritmos proporcionan predicciones más confiables y robustas, favoreciendo la toma de decisiones en la gestión del riesgo crediticio.

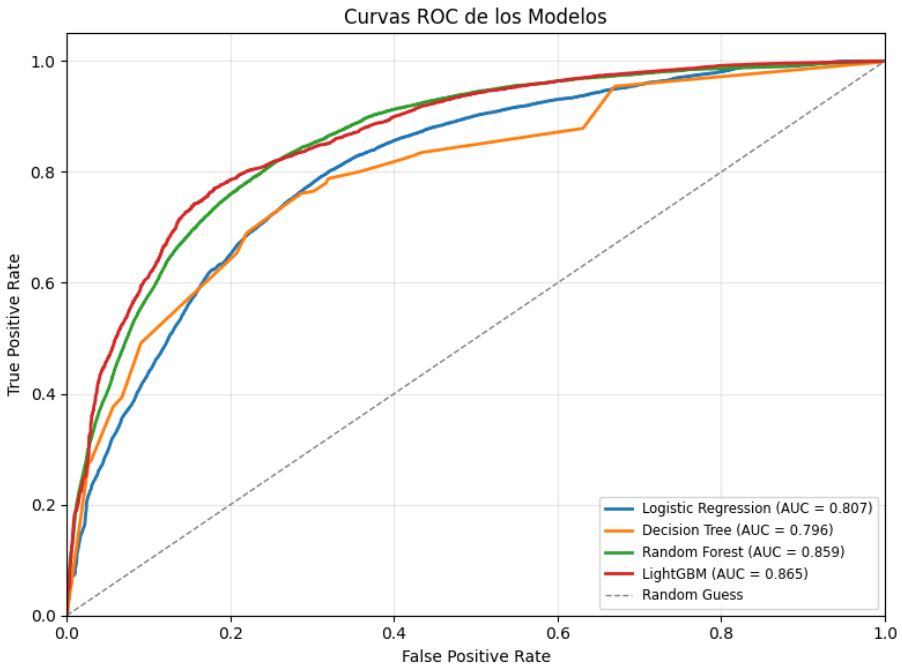


Ilustración 19. Curva ROC

Curvas Precision-Recall:

Son particularmente útiles cuando el problema tiene muchas más facturas pagadas que impagas, es decir, cuando hay desbalance de clases. En este gráfico, la precisión mide qué porcentaje de las facturas que el modelo predijo como “morosas” realmente lo son, y el recall indica qué porcentaje de todas las facturas morosas reales logra identificar el modelo.

Al analizar la curva, observamos que LightGBM y Random Forest tienen líneas que se mantienen más arriba a lo largo de la gráfica, lo que significa que, en todos los niveles de recall (sensibilidad), sus predicciones son más precisas que las de los otros modelos. El área promedio bajo la curva es también más alta para estos dos modelos, lo cual indica que tienen un mejor equilibrio entre encontrar la mayor cantidad de morosos posible (alto recall) sin sacrificar la calidad de las predicciones (alta precisión).

En resumen, esto significa que, al usar cualquiera de estos modelos, se podrá identificar a la mayoría de los clientes de alto riesgo, cometiendo menos errores al señalar como morosos a clientes que en realidad sí van a pagar.

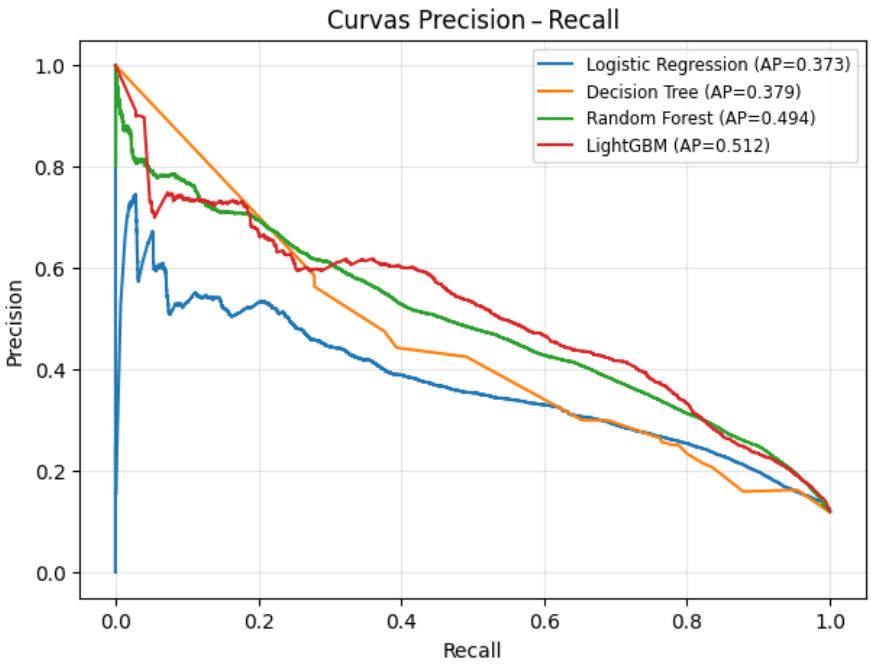


Ilustración 20. Curva Precision-Recall

Curvas de calibración

Permiten evaluar qué tan bien las probabilidades de impago estimadas por cada modelo reflejan la realidad observada. En otras palabras, muestran si las predicciones de probabilidad generadas por los modelos corresponden con la frecuencia real de impagos en los datos de prueba.

Idealmente, los puntos deberían alinearse lo más cerca posible de la diagonal, lo que indicaría que, por ejemplo, de todas las facturas a las que el modelo asigna una probabilidad de impago del 40%, aproximadamente el 40% realmente termina en mora.

En la comparación, se observa que LightGBM y Random Forest presentan curvas más próximas a la diagonal, lo que implica que sus probabilidades predichas son más confiables y representan adecuadamente el riesgo real. Por otro lado, los modelos de regresión logística y árboles de decisión muestran mayor dispersión, lo que sugiere que tienden a subestimar o sobrestimar la probabilidad real de impago.

En resumen, los modelos mejor calibrados ofrecen mayor utilidad práctica, ya que permiten interpretar y actuar sobre sus probabilidades de forma directa y con confianza en los procesos de evaluación y toma de decisiones crediticias.

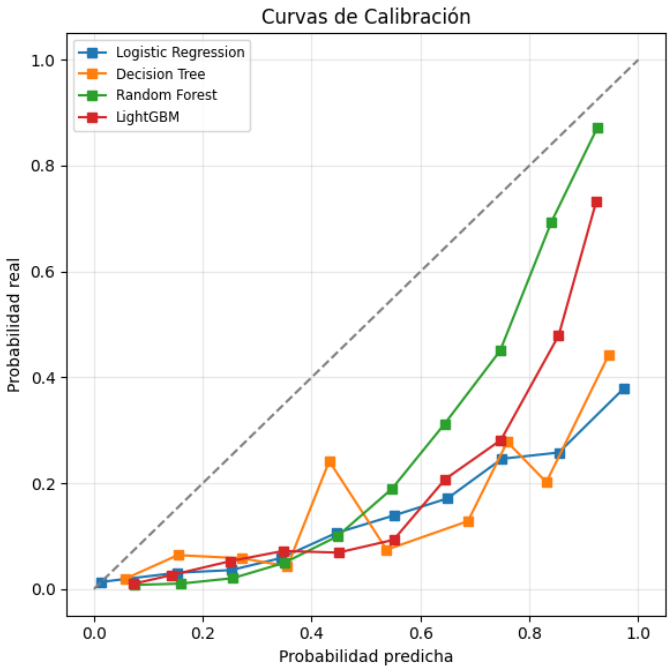


Ilustración 21. Curvas de calibración.

Curvas de ganancia

Muestran el porcentaje acumulado de facturas morosas que cada modelo logra identificar a medida que se aumenta la proporción de población ordenada según el score de riesgo. En la práctica, estas curvas permiten visualizar la capacidad del modelo para priorizar correctamente a los clientes con mayor probabilidad de incurrir en mora.

Una curva de ganancia ideal se aproxima rápidamente al 100% acumulado de morosos utilizando el menor porcentaje posible de la población, lo que se traduce en un mayor poder de discriminación y eficiencia en el proceso de selección y gestión de riesgo crediticio.

Al comparar los modelos, tanto LightGBM como Random Forest presentan curvas que crecen de forma más rápida y alcanzan altos niveles de recuperación de morosos en los primeros segmentos de la población, superando significativamente la línea base aleatoria y también a los modelos de regresión logística y árboles de decisión. Esto confirma que estos modelos no solo logran distinguir mejor a los clientes con mayor riesgo, sino que también permiten priorizar acciones preventivas de manera más eficiente.

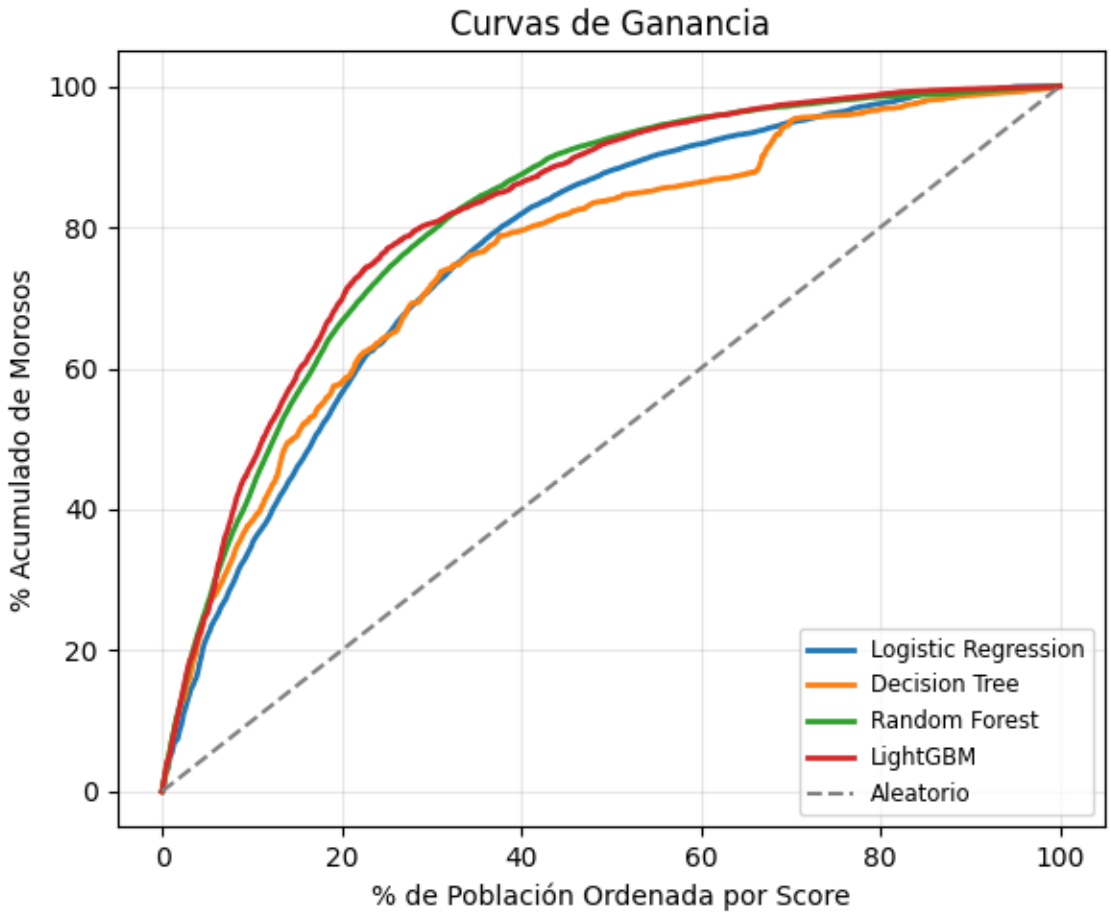


Ilustración 22. Curvas de ganancia

Importancia de características

El análisis de importancia de características revela que las variables relacionadas con el historial de pago y comportamiento previo de los clientes son las que más peso tienen en la predicción de incobrabilidad. Destacan especialmente el promedio de días de atraso en pagos (*avg_prev_days_late*) y la cantidad de facturas pagadas entre 31 y 60 días (*num_prev_paid_31_60d*), seguidas por el tiempo transcurrido desde el último pago (*days_since_last_pay*) y otras variables asociadas al historial de facturación y puntualidad en los pagos. Este resultado es consistente con la lógica de negocio: los clientes con mayores atrasos y menor regularidad en los pagos anteriores presentan un mayor riesgo de caer en mora. Por otro lado, variables como el monto de la factura actual (*invoice_amount*) tienen menor influencia en la predicción. En conjunto, estos hallazgos confirman que el modelo está capturando patrones relevantes y alineados con la realidad operativa para anticipar el riesgo crediticio.

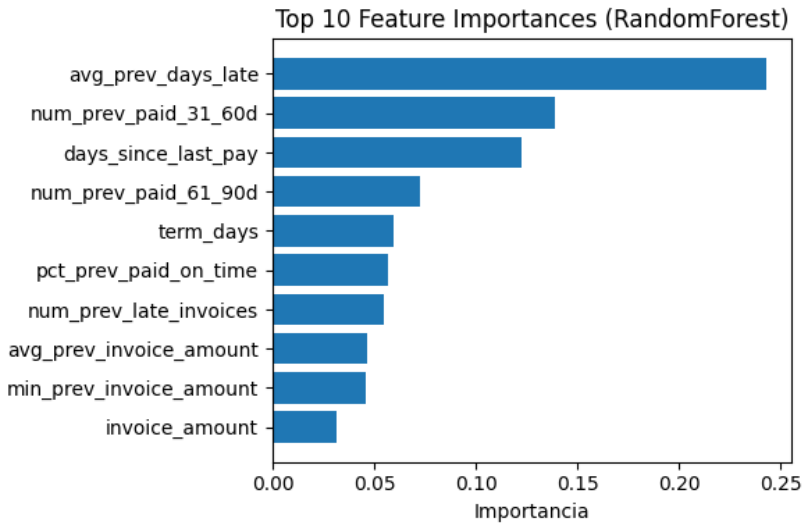


Ilustración 23. Importancia de características

3.6.3 Selección de modelo

Tras una evaluación exhaustiva utilizando múltiples métricas y visualizaciones comparativas, el modelo seleccionado para el prototipo fue Random Forest. Esta decisión se fundamenta en varios factores observados durante el análisis.

En primer lugar, Random Forest mostró un desempeño competitivo en todas las métricas clave: obtuvo un alto valor de AUC en la curva ROC, mantuvo buena precisión y recall incluso frente al desbalance de clases, y presentó una curva de calibración aceptable, lo que sugiere que las probabilidades predichas pueden interpretarse de manera confiable en términos de riesgo real.

Si bien LightGBM demostró un rendimiento ligeramente superior en algunas métricas, se identificó que este modelo tendía a asignar un peso excesivo a la variable de identificación de cliente, lo que puede introducir sesgos y limitar la generalización en escenarios de validación real o con nuevos datos. Por el contrario, Random Forest mostró un balance más adecuado en la importancia de las variables y una mayor estabilidad en las predicciones.

Adicionalmente, Random Forest destaca por su robustez ante el ruido y la variabilidad de los datos, además de ofrecer una interpretación clara de los factores que influyen en el resultado de la predicción, lo cual resulta especialmente valioso para el análisis y la

toma de decisiones. Por estas razones, se eligió Random Forest como el modelo base para la implementación y despliegue del sistema de predicción de incobrabilidad.

3.7 Implementación del Modelo en prototipo

Una vez que se ha seleccionado el modelo predictivo más adecuado, se procede a su implementación en Gradio.

3.7.1 Diseño del Dashboard

El prototipo desarrollado se presenta como un dashboard interactivo, diseñado específicamente para facilitar la interacción con los stakeholders. Este formulario digital permite seleccionar un cliente desde una lista desplegable y visualizar automáticamente su información histórica relevante, como número de facturas emitidas, comportamiento de pago, promedios de atraso, entre otros indicadores fundamentales para la evaluación crediticia.

Además de la información histórica, el usuario puede simular nuevas condiciones de crédito; por ejemplo, ajustando el monto de la factura o los días de vencimiento, y obtener al instante una predicción de incobrabilidad para el escenario planteado. El dash board está construido para ser intuitivo, guiando al usuario en la introducción de los datos mediante campos interactivos y controles desplegables que validan las opciones permitidas. Si bien la funcionalidad central es la predicción de impago, el diseño del prototipo permite, en futuras iteraciones, la integración de visualizaciones adicionales como gráficas de riesgo por segmento de cliente, evolución temporal del score de crédito, o alertas visuales sobre cambios en los patrones de pago.

El resultado principal entregado por el formulario es una probabilidad estimada de impago, junto con la clasificación automática de la factura como potencialmente pagada o incobrable, facilitando la toma de decisiones informadas y ágiles.

3.7.2 Integración del modelo predictivo

El dashboard está completamente integrado con el modelo predictivo seleccionado durante la etapa de entrenamiento y evaluación. Esta integración permite que los datos introducidos por el usuario en el formulario sean transformados internamente en el formato requerido por el modelo, incluyendo la codificación de variables categóricas y la validación de los valores ingresados.

Cada vez que se ingresan nuevos valores en el formulario y se solicita una predicción, el sistema prepara estos datos de entrada y los envía al modelo entrenado, el cual retorna tanto la clasificación (pago/incobrable) como la probabilidad asociada de impago. Este proceso se realiza de manera transparente para el usuario, asegurando que la predicción esté basada en la misma lógica y preprocesamiento utilizados durante la fase de entrenamiento y validación del modelo.

De esta manera, el prototipo funciona como una interfaz amigable que encapsula toda la complejidad técnica del modelo, aportando capacidades avanzadas de machine learning sin requerir conocimientos técnicos especializados.

3.7.3 Actualización y monitoreo

En esta primera fase de implementación, el uso del modelo predictivo será manual o ad hoc, ejecutándose principalmente en situaciones de análisis crediticio, como la liberación de pedidos antes de su facturación. Este enfoque flexible facilita la adopción gradual de la herramienta dentro del flujo operativo, permitiendo que los stakeholders utilicen el modelo como apoyo en la toma de decisiones, sin requerir una integración inmediata y completa con los sistemas transaccionales.

Se evaluará la posibilidad de aumentar la frecuencia de ejecución, evolucionando hacia un proceso más automatizado, según la adopción y utilidad percibida del modelo. En etapas futuras, el modelo podría ejecutarse de manera periódica o en tiempo real, en función nuevos eventos, como la creación de nuevos pedidos o cambios en la situación crediticia de los clientes.

Es fundamental establecer desde el inicio un proceso de monitoreo continuo del rendimiento del modelo, permitiendo detectar cualquier degradación en la precisión o cambios en los patrones de los datos. De esta manera, se asegura la vigencia y utilidad del sistema predictivo a lo largo del tiempo, ajustándolo y mejorándolo cuando sea necesario.

3.7.4 Indicadores Clave de Desempeño (KPIs)

El desempeño del modelo predictivo será evaluado mediante una serie de indicadores clave de desempeño (KPIs), enfocados tanto en su capacidad técnica como en su impacto operativo:

- **Recall (clase “no paga”):** Métrica prioritaria en este estudio, ya que mide la capacidad del modelo para identificar todos los casos de morosidad. Su optimización es esencial para minimizar el riesgo de dejar pasar clientes morosos y maximizar la efectividad de la gestión de cobranzas.
- **Precisión del modelo:** Proporción de predicciones correctas, fundamental para validar que el modelo identifica patrones de morosidad de manera confiable.
- **Tasa de falsos positivos:** Porcentaje de facturas erróneamente clasificadas como morosas, indicador crítico para evitar rechazos innecesarios de crédito.
- **Tasa de falsos negativos:** Porcentaje de facturas morosas no detectadas por el modelo, cuya minimización es clave para prevenir riesgos financieros.
- **Impacto en la recuperación de cartera:** Métrica orientada a resultados, que evalúa si la utilización del modelo mejora la gestión de cobranzas y reduce la morosidad global. Se utilizarán medidas financieras, como los días de recuperación crediticia y el promedio de cartera vencida mensual.

Estos indicadores serán monitoreados y revisados periódicamente. En caso de observarse deterioro en el rendimiento de alguno de ellos, se tomarán medidas correctivas, como ajustes en el modelo o su reentrenamiento.

3.7.5 Reajuste y Mantenimiento

El modelo predictivo debe ser entendido como un sistema dinámico, sujeto a reajustes periódicos en función de los cambios en los datos y las condiciones económicas. Esto implica volver a entrenar el modelo con información más reciente, revisar y ajustar hiperparámetros, y asegurar que los supuestos originales se mantengan vigentes.

El mantenimiento incluye la revisión regular de resultados y predicciones (por ejemplo, trimestral o semestralmente), la actualización de los conjuntos de datos y la verificación de la estabilidad del modelo frente a cambios en la distribución de las variables de entrada. Si se identifican desviaciones significativas o pérdida de precisión, se procederá con las actualizaciones y ajustes necesarios para restaurar la efectividad del sistema.

3.7.6 Escalabilidad y Mejoras

El prototipo está diseñado para ser escalable y adaptable a futuras necesidades. El sistema puede incorporar nuevas fuentes de información, manejar volúmenes crecientes

de datos, y adoptar algoritmos más avanzados en caso de que los requerimientos lo justifiquen.

Las siguientes etapas de mejora pueden incluir la integración de nuevos modelos de machine learning, la implementación de técnicas de ensamble, o la automatización completa del flujo de análisis crediticio. El objetivo es evolucionar hacia un sistema altamente automatizado y proactivo, capaz de reaccionar a cambios en tiempo real y contribuir de manera estratégica a la gestión de riesgos.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En este estudio, se implementaron y evaluaron cuatro modelos de clasificación para predecir el comportamiento de pagos de los clientes, utilizando un conjunto de datos enriquecido con características históricas de facturación y pagos. Los modelos evaluados incluyeron Regresión Logística, Árbol de Decisión, Random Forest y LightGBM (un método avanzado de Gradient Boosting). Cada modelo fue ajustado y evaluado mediante técnicas como la validación cruzada y la búsqueda de hiperparámetros, y su desempeño se midió con métricas clave como precisión, recall, F1-score y matrices de confusión.

4.1 Conclusiones

Los resultados obtenidos en la evaluación de modelos permiten identificar un claro liderazgo por parte de Random Forest, el cual, aunque no obtuvo el AUC más alto (0.8587), sí logró el mejor equilibrio general entre las métricas clave. Su recall y F1-score de 0.79, junto con una precisión de 0.33, lo posicionan como el modelo más robusto y transparente, lo que justifica su selección como modelo final.

LightGBM, por su parte, alcanzó el AUC más alto (0.8629) y el mayor recall (0.81), lo que indica una mayor capacidad para detectar facturas morosas. Sin embargo, su menor interpretabilidad y una ligera caída en el F1-score (0.76) frente a Random Forest, lo hacen menos adecuado en contextos donde la transparencia es prioritaria.

El Árbol de Decisión mostró un rendimiento más limitado, con un F1-score de apenas 0.38 y una precisión de 0.25, lo que refleja su tendencia al sobreajuste y su menor capacidad para generalizar. A pesar de su simplicidad, estos factores reducen su aplicabilidad práctica.

Finalmente, la Regresión Logística se comportó como un modelo de referencia útil, con métricas moderadas (AUC: 0.8072, F1-score: 0.4), pero con limitaciones claras para capturar patrones complejos, lo que restringe su uso en escenarios más exigentes.

Estos hallazgos, resumidos en la Tabla 9, respaldan la elección de Random Forest como la mejor alternativa para abordar el problema de predicción de incobrabilidad.

Tabla 9. Resultado de modelos evaluados

Modelo	AUC Test	Recall	F1-Score	Precisión	Observaciones
Random Forest	0.8587	0.79	0.79	0.33	Mejor equilibrio global. Modelo seleccionado por robustez y transparencia.
LightGBM	0.8629	0.81	0.76	0.31	Ligeramente superior en recall, pero menos interpretable.
Árbol de Decisión	0.7962	0.78	0.38	0.25	Sencillo, pero con menor capacidad predictiva y tendencia al sobreajuste.
Regresión Logística	0.8072	0.76	0.4	0.27	Línea base, limitada en patrones complejos.

El desarrollo e implementación de modelos de machine learning ha aportado una herramienta robusta y adaptada al contexto empresarial para la gestión del riesgo de impago, considerando específicamente el criterio de facturas impagas por más de 30 días. El modelo Random Forest fue seleccionado como la solución óptima, gracias a su excelente rendimiento, estabilidad, capacidad de generalización y facilidad de interpretación. Este modelo permitirá optimizar la toma de decisiones crediticias, reducir el riesgo financiero y mejorar la eficiencia en la recuperación de cartera. En definitiva, la solución desarrollada no solo optimiza la gestión de cuentas por cobrar, sino que además ha sido diseñada para su futura integración con Power Platform, asegurando una transición fluida hacia la operación diaria y el aprovechamiento de las herramientas corporativas ya implementadas.

4.2 Recomendaciones

Con base en los resultados obtenidos y el modelo final seleccionado (Random Forest), se proponen las siguientes recomendaciones para optimizar el modelo desarrollado y mejorar su aplicabilidad en la gestión de cuentas por cobrar:

1. **Definir y actualizar el horizonte de la variable objetivo:** Se recomienda revisar periódicamente el criterio de clasificación de impago, actualmente definido como facturas impagas por más de 30 días. Este horizonte puede ser ajustado en función de las políticas internas y contextos económicos, o bien diferenciarse para distintos segmentos de clientes según riesgo o perfil de pago.

2. **Simulación de límites y condiciones de crédito:** Utilizar el modelo Random Forest no solo para predecir impago, sino para simular diferentes escenarios, ajustando montos de factura, plazos y condiciones de crédito. Permitiendo optimizar las condiciones ofrecidas, anticipar riesgos y proponer límites de crédito adaptados a cada cliente.
3. **Estrategias para nuevos clientes sin historial:** Para clientes nuevos, sin suficiente historial de facturación, se recomienda desarrollar metodologías de segmentación, clustering, por ejemplo, que permitan asignarles un perfil de riesgo estimado usando información agregada de grupos similares. Adicionalmente, enriquecer el modelo con variables externas (como indicadores financieros o de mercado) puede aumentar la precisión en este segmento.
4. **Implementar monitoreo y recalibración continua del modelo:** Establecer procesos para monitorizar el desempeño del modelo en producción, revisando periódicamente métricas clave (AUC, recall, precisión, falsos positivos/negativos). Ante variaciones significativas en los datos de entrada o cambios en los patrones de pago, será necesario recalibrar o reentrenar el modelo con datos más recientes.
5. **Fomentar la colaboración:** Involucrar activamente a los stakeholders en la interpretación de los resultados del modelo y en la toma de decisiones relacionadas con crédito y cobranza. La retroalimentación de usuarios expertos es esencial para ajustar reglas, mejorar la calidad de los datos y asegurar que la herramienta aporte valor en la operación diaria.
6. **Uso estratégico del prototipo:** Aprovechar el prototipo interactivo desarrollado para facilitar la adopción del modelo y validar su utilidad antes de considerar una integración completa con sistemas transaccionales. De este modo, se facilita la adopción y el ajuste progresivo a las necesidades reales.

4.3 Próximos pasos

En adelante, se sugiere tomar las siguientes acciones para explotar y desarrollar continuamente el proyecto:

- **Escalabilidad y automatización:** Planificar la automatización gradual de la herramienta, integrando el modelo predictivo con los sistemas de gestión de

corporativos para que la evaluación de riesgo se ejecute en tiempo real cada vez que se registre un nuevo pedido o cambio relevante en la situación crediticia de un cliente.

- **Ampliar análisis a nuevos algoritmos y variables:** Explorar en el futuro otros algoritmos avanzados (como XGBoost) o el desarrollo de modelos híbridos/ensamble para comparar resultados. Además, analizar la inclusión de nuevas fuentes de datos, como información sectorial, macroeconómica o comportamental.
- **Validación dinámica y seguimiento de desempeño:** Implementar estrategias de validación continua, como ventanas móviles de tiempo, para evaluar y ajustar el modelo ante cambios de contexto. Crear dashboards automáticos para monitorear los KPIs críticos, alertando tempranamente sobre posibles degradaciones de rendimiento.
- **Análisis por segmentos y personalización:** Realizar análisis de desempeño por segmentos de clientes, productos o regiones, y personalizar los modelos o los umbrales de decisión en función de los perfiles detectados.
- **Estrategias preventivas y comunicación proactiva:** Utilizar las predicciones del modelo para definir estrategias preventivas de cobranza, priorizando acciones sobre clientes con mayor riesgo y ajustando la política comercial y de seguimiento según el perfil de riesgo identificado.
- **Implementación en el ecosistema de Microsoft:** Como parte de la estrategia de despliegue a producción, se recomienda migrar el prototipo a la Power Platform. Esto permitirá utilizar Power BI para el monitoreo y análisis de resultados, y Power Apps para la captura y consulta interactiva de predicciones. El flujo completo se orquestará mediante un servicio en Python que ejecute el preprocesamiento y la predicción, integrándose de manera transparente con otras herramientas utilizadas. Esta integración garantizará facilidad de adopción, escalabilidad y seguridad, aprovechando la infraestructura tecnológica existente.

Estas recomendaciones y próximos pasos buscan asegurar la sostenibilidad y mejora continua del sistema predictivo implementado, maximizando el valor aportado por la analítica avanzada en la gestión del crédito y la cartera.

BIBLIOGRAFÍA

Brown, I. M. (2019). Machine learning for credit risk: Evidence from global banks. *Journal of Banking and Finance*, 107. doi:105285

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321-357.

Chent, T. &. (2016). XGBoost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:10.1145/2939672.2939785

Czopelk, K.-Z. B. (2013). Optimisation of receivables management in a mine, using linear programming. *Archives of Mining Sciences*.

El Universo. (2023, Septiembre 25). El Universo. *Precios a la baja golpean al sector camaronero que hasta agosto percibió -5 % en facturación, aunque produce y exporta más*, p. n.a. Retrieved from Precios a la baja golpean al sector camaronero que hasta agosto percibió -5 % en facturación, aunque produce y exporta más: https://www.eluniverso.com/noticias/economia/precios-a-la-baja-golpean-al-sector-camaronero-que-hasta-agosto-percibio-5-en-facturacion-aunque-produce-y-exporta-mas-nota/

Fernández, M., & González, P. (2021). Apliación de Random Forest en evaluación de crédito para PYMEs. *Revista de Finanzas Aplicadas*, 45-60.

Hand, D. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*.

Hughes, A. M. (2002). *Strategic Database Marketing* (3 ed.). McGraw-Hill.

Jankowski, R., & Palinski, A. (2024, Mayo). Debt Collection Model for Mass Receivables Based on Decision Rules—A Path to Efficiency and Sustainability. *MDPI*, 3-7. doi:10.3390/su16145885

López, A. &. (2024). Integración de scikit-learn, XGBoost y SMOTE en procesos de cobranza automatizada. *Revista de Machine Learning Aplicado*, 23-37.

Pedregosa, F. V. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Pfeifer, P. E. (2000). Modeling customer relationships as Markov chains. *Journal of Interactive Marketing*, 43-51.

Produbanco. (2022, mayo 31). *Produbanco y su contribución al Sector Camaronero*. Retrieved from Banca Responsable: https://www.produbanco.com.ec/Noticias/articulos/sector-camaronero/

Ramírez, J. &. (2023). Aplicaciones de Python y técnicas de sobremuestreo en la gestión de impagos del sector agrícola. *Revista Iberoamericana de Finanzas y Agroindustria*, 98-112.

Ramírez, L. (2022). Estacionalidad y financiamiento en acuicultura: un análisis de riesgo. *Revista Acuícola Latinoamericana*, 112-130.

Saito, T. &. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3). doi:10.1371/journal.pone.0118432

Shewell, K. (2024, Marzo 19). *Medium.* Retrieved from Logistic Regression in the Credit Risk Industry: https://2os.medium.com/logistic-regression-in-the-credit-risk-industry-6eb27bc2784c

Shewell, K. (2024). Monotonic relationships in credit scoring: Theory and practice. *Journal of Credit Risk, 20*(1), 1-25.

Smith, J. &. (2020). Dinámicas de liquidez en camaroneras ecuatorianas. *Economía y Desarrollo*, 77-89.

Sum, H. (2016). Credit Risk Assessment of Receivable Accounts in Industry Chain based on SVM. *Proceedings of Science*, 2-5. Retrieved from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://pos.sissa.it/300/020/pdf

Taleb, A., Ruiz, C., & Sanchez, J. (2022). Rapid prototyping of credit-risk models with Gradio. *Proceedings of PyData Latam.* Ciudad de México: PyData Latam.

Vega, M. H. (2022). Validación rápida de modelos de riesgo crediticio con interfaces web ligeras. *Journal of Financial Technology*, 45-59.

Vora, D., Choudhary, R., Kumar, A., & Kadam, P. (2024). Forecasting User Payment Behavior Using Machine Learning. *Communications in Computer and Information Science*. doi:0.1007/978-3-031-55486-5_6