ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

CREACIÓN DE UN MODELO PREDICTIVO BASADO EN INTELIGENCIA ARTIFICIAL PARA LA EVALUACIÓN DE RIESGO CREDITICIO

Profesor
Ing. Manuel Eugenio Morocho

Autores
Adriana Tana Pilacuán
Ana Cristina Viteri Orozco

2024

RESUMEN

En el entorno financiero actual, la evaluación del riesgo crediticio es esencial para las instituciones financieras. Tradicionalmente, se ha basado en modelos estadísticos y datos históricos, pero el avance de la inteligencia artificial (IA), especialmente en aprendizaje automático y análisis de grandes datos, ha mejorado significativamente la precisión y eficiencia de este proceso.

Este estudio se centra en desarrollar un modelo predictivo basado en inteligencia artificial (IA) para evaluar el riesgo crediticio. La IA puede analizar datos en tiempo real, identificar patrones complejos y adaptarse a nueva información, mejorando la precisión de las predicciones y optimizando la toma de decisiones, lo que reduce costos y mitiga riesgos. Se exploran metodologías de aprendizaje automático, tanto supervisadas como no supervisadas, y se abordan los desafíos éticos, así como las estrategias para mejorar la transparencia y la interpretabilidad de las decisiones automatizadas.

Este trabajo busca presentar un modelo innovador y ofrecer una guía práctica para integrar efectivamente la inteligencia artificial en el sector financiero, logrando un equilibrio entre innovación tecnológica y gestión prudente del riesgo crediticio.

El objetivo general de este proyecto es determinar la factibilidad de implementar un modelo basado en inteligencia artificial para la otorgación de créditos.
En este proyecto se han utilizado varios modelos para implementar el scoring crediticio:

a) **Regresión Logística**: Un modelo tradicional utilizado para predecir la probabilidad de que un cliente cumpla con sus obligaciones crediticias.
b) **Random Forest**: Un método que utiliza múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste.

c) **XGBoost**: Un algoritmo de aprendizaje automático basado en árboles de decisión, conocido por su alto rendimiento y eficiencia en tareas de clasificación y regresión.

En resumen, la creación de un modelo predictivo basado en inteligencia artificial para la evaluación de riesgo crediticio puede proporcionar insights valiosos y mejoras significativas en la gestión de riesgos financieros, pero requiere un enfoque integral que abarque desde la construcción y validación del modelo hasta su implementación efectiva y cumplimiento ético y regulatorio.

ABSTRACT

In today's financial environment, credit risk assessment is essential for financial institutions. Traditionally, it has relied on statistical models and historical data, but advances in artificial intelligence (AI), particularly in machine learning and big data analysis, have significantly improved the accuracy and efficiency of this process.

This study focuses on developing a predictive model based on artificial intelligence (AI) to evaluate credit risk. AI can analyze data in real-time, identify complex patterns, and adapt to new information, enhancing prediction accuracy and optimizing decision-making, which reduces costs and mitigates risks. Both supervised and unsupervised machine learning methodologies are explored, and ethical challenges, as well as strategies to improve transparency and interpretability of automated decisions, are addressed.

This work aims to present an innovative model and provide a practical guide for effectively integrating artificial intelligence into the financial sector, achieving a balance between technological innovation and prudent credit risk management.

The overall objective of this project is to determine the feasibility of implementing an AI-based model for credit granting. In this project, several models have been used to implement credit scoring:

Logistic Regression: A traditional model used to predict the probability of a client meeting their credit obligations.

Random Forest: A method that uses multiple decision trees to improve accuracy and avoid overfitting.

XGBoost: A machine learning algorithm based on decision trees, known for its high performance and efficiency in classification and regression tasks.

In summary, creating a predictive model based on artificial intelligence for credit risk assessment can provide valuable insights and significant improvements in financial risk management, but it requires a comprehensive approach that encompasses model development, validation, effective implementation, and ethical and regulatory compliance.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En el dinámico panorama financiero actual, la evaluación precisa del riesgo crediticio es fundamental para las instituciones financieras. Tradicionalmente, este proceso ha dependido en gran medida de modelos estadísticos y análisis de datos históricos. Sin embargo, con el avance de la inteligencia artificial (IA), especialmente en técnicas de aprendizaje automático y procesamiento de grandes volúmenes de datos, se ha abierto un nuevo horizonte para mejorar la precisión y eficiencia en la evaluación de riesgos.

Este estudio se centra en la creación y aplicación de un modelo predictivo basado en inteligencia artificial para la evaluación de riesgo crediticio. La IA ofrece la capacidad única de analizar múltiples fuentes de datos en tiempo real, identificar patrones complejos y adaptarse continuamente a nuevas informaciones. Estas capacidades no solo prometen mejorar la precisión de las predicciones, sino también optimizar el proceso de toma de decisiones crediticias, reduciendo costos operativos y mitigando riesgos inherentes.

En este contexto, exploraremos las metodologías más relevantes de aprendizaje automático aplicadas a la evaluación de riesgo crediticio, incluyendo modelos supervisados y no supervisados. Se discutirán también los desafíos y consideraciones éticas asociadas con la implementación de estos modelos, así como las estrategias para mejorar la transparencia y la interpretabilidad de las decisiones automatizadas.

Finalmente, este trabajo no solo aspira a presentar un modelo innovador, sino también a proporcionar una guía práctica para la integración efectiva de la inteligencia artificial en el sector financiero, asegurando así un equilibrio óptimo entre la innovación tecnológica y la gestión prudente del riesgo crediticio.

REVISIÓN DE LITERATURA

La predicción del riesgo de impago en el sector financiero es crucial, especialmente en el ámbito bancario, donde es fundamental evaluar la solvencia crediticia de los clientes y reducir los riesgos vinculados a los préstamos concedidos.

El machine learning (ML) es una rama de la inteligencia artificial que puede ser usada en múltiples áreas de investigación académica, desarrollo tecnológico y predicción de datos para empresas (Enterprise.nxt Staff, 2021), ya que su proceso consiste en crear un modelo algorítmico que permite ingresar datos históricos, hacer un entrenamiento según el enfoque que se requiera y, a partir de este, encontrar patrones para predecir alguna información futura. Esto genera que las empresas pongan énfasis en el desarrollo del machine learning, pero es importante que se identifique cuáles son los objetivos que se quieren lograr con su implementación (SAS: Analytics, Artificial Intelligence and Data Management, 2017). Según una publicación en Forbes (2021), el 43 % de empresas ha detectado que la implementación de estos algoritmos ha sido más útil de lo que creían inicialmente y el 50 % ha planeado invertir más en estos modelos el presente año, lo cual empieza a demostrar que ponerlos en funcionamiento es de gran importancia en cualquier sector. Un sector que importa y que genera el movimiento en la economía, es el financiero, el cual lleva implementándolo para mejorar la experiencia al usuario a través de soluciones personalizadas o chatbots como destacó Forbes a la BBVA Corporation (2019). Por otro lado, una actividad que es de gran importancia para cualquier organización o entidad financiera es la gestión de riesgos que puede ser dividida según su taxonomía en: riesgo de seguros y demográfico, riesgo de mercado, riesgo de crédito y riesgo operacional (Mashrur et al., 2020). Al que más enfoque se ha prestado es a los riesgos crediticios para poder mitigar las pérdidas ya sea un banco o una fintech dedicada a préstamos, ya que en el último Risk Dashboard Q1 2021 de la European Banking Authority (EBA) se enfatizó que los bancos seguían siendo vulnerables a los movimientos adversos del riesgo crediticio (KPMG Company, 2021). Por este motivo, se percibe un cambio en el paradigma de no solo optar por los modelos estadísticos tradicionales, sino que cada vez procuren ser más

sofisticados como el machine learning para cuantificar y mitigar el riesgo de manera correcta (Mashrur et al., 2020).

Hoy en día, la calificación crediticia es fundamental para ayudar a las instituciones financieras a conocer bien a las empresas para mitigar los riesgos crediticios. Es una indicación del nivel del riesgo de invertir con la corporación y representa la probabilidad de que la corporación pague sus obligaciones financieras a tiempo. Por lo tanto, es de gran importancia modelar el perfil de la corporación para predecir el nivel de calificación crediticia. Sin embargo, este proceso de evaluación suele ser muy costoso y complicado, lo que a menudo lleva meses con muchos expertos involucrados para analizar todo tipo de variables, que reflejan la confiabilidad de una corporación.

La industria bancaria ha desarrollado algunos modelos de riesgo de crédito desde mediados del siglo XX. La calificación de riesgo también es el negocio principal de miles de corporaciones mundiales, incluidas docenas de empresas públicas. Debido al valor altamente práctico, se han desarrollado muchos tipos de modelos de calificación crediticia. Tradicionalmente, los modelos de crédito se proponen por algoritmos de regresión logística con la calificación crediticia temporal, así como información financiera agregada. Hoy en día, los modelos de aprendizaje automático y de aprendizaje profundo han mostrado su poder en una variedad de aplicaciones, incluidos los campos financieros.

Con el tiempo, se han desarrollado varios modelos para mejorar la evaluación del riesgo crediticio, como el modelo de puntuación crediticia basado en regresión logística creado por la Facultad de Economía de Nuevo León. Este modelo se diseñó para analizar la probabilidad de incumplimiento en diferentes segmentos de la cartera de clientes de tarjetas de crédito de una institución mexicana. Los resultados demuestran que el modelo propuesto ofrece una alta capacidad predictiva y estabilidad tanto dentro como fuera del periodo de modelado. Además, la verificación de la monotonicidad asegura que el modelo mantenga un alto nivel de precisión.

El Sistema de Expertos con Aplicaciones propone un nuevo modelo de evaluación de clasificación múltiple de riesgo de crédito personal basado en la teoría de la fusión de información (MIFCA) mediante el uso de seis algoritmos de

aprendizaje automático. El modelo MIFCA puede integrar simultáneamente las ventajas de múltiples clasificadores y reducir la interferencia de la información incierta. Para verificar el modelo MIFCA, el conjunto de datos recopilado de un conjunto de datos real de CommercialBank en China. Los resultados experimentales muestran que el modelo MIFCA tiene dos puntos sobresalientes en varios criterios de evaluación. Una es que tiene mayor precisión para la evaluación de la clasificación múltiple, y la otra es que es adecuado para diversas evaluaciones de riesgos y tiene aplicabilidad universal. Además, los resultados de esta investigación también pueden proporcionar referencias a los bancos y otras instituciones financieras para fortalecer sus capacidades de prevención y control de riesgos, mejorar sus capacidades de identificación de riesgo de crédito y evitar pérdidas financieras.

El Journal Data Science propone un modelo basado en Redes Neuronales de Grafos (GNN) el cual mejora en la precisión de calificación crediticia, superando métodos tradicionales.(J. Wang et al., 2021)

El Hal open Science de Ayoub El Qadi El Haouari, propone un modelo de un sistema de calificación crediticia de inteligencia artificial explicable el cual mejora en la transparencia y explicabilidad de los modelos de calificación crediticia incrementando la confianza de las instituciones financieras en los modelos de AI. En este modelo se incorpora evaluaciones de crédito textual en el modelo de calificación crediticia utilizando técnicas de procesamiento del lenguaje natural de vanguardia (PNL). Los resultados demostraron que los modelos capacitados con datos financieros y evaluaciones de crédito textuales superaron a los que dependían únicamente de los datos financieros. Además, demuestra que el enfoque podría generar efectivamente puntajes de crédito utilizando solo evaluaciones de riesgos textuales, ofreciendo así una solución viable para escenarios en los que las métricas financieras tradicionales no están disponibles o no son insuficientes.(El Qadi El Haouari, n.d.)

El IEEEAccess propone un modelo de ensamble multietapa con algoritmo genético híbrido que mejora en la precisión de predicción, este modelo permite eliminar las características redundantes e irrelevantes en el conjunto de datos y seleccionar clasificadores base de rendimiento bien realizado, con un nuevo

algoritmo genético híbrido que permita obtener el subconjunto de características óptimo y el subconjunto de clasificadores base. Para agregar el poder predictivo de los clasificadores base, se adopta un enfoque de apilamiento para integrar los clasificadores base óptimos en el modelo de conjunto. El modelo propuesto se prueba en tres conjuntos de datos de calificación crediticia desequilibrada estándar, en comparación con enfoques de estado de arte similares, y se evalúa utilizando cuatro indicadores de evaluación bien conocidos. Los resultados experimentales demuestran la efectividad del modelo propuesto y demuestran su superioridad.(Jin et al., 2021)

El China Mobile Information Technology Co., Ltd. Propone un modelo que permite extraer múltiples relaciones de eventos, como relaciones secuenciales, relaciones causales, relaciones inversas y relaciones paralelas. Construye el gráfico lógico de eventos utilizando eventos como nodos y relaciones de eventos múltiples como bordes. Crea un modelo de red convolucional temporal impulsado por el gráfico de lógica de eventos, que integra eventos de noticias y datos de transacciones de bonos de crédito corporativo para pronosticar riesgos de crédito. Los experimentos comparativos y los experimentos de ablación muestran que nuestro enfoque supera a los métodos de referencia, y la relación múltiple tiene un mejor rendimiento en el pronóstico de tendencias que la relación única.(Zhu et al., 2016)

Otro modelo propuesto es una red neuronal de gráficos con conocimiento temporal para predecir el riesgo de crédito del usuario. El problema se modela en varias instantáneas de gráficos ordenadas por el tiempo. Primero un modelo separado para extraer la información estática. En cada instantánea se propone un codificador gráfico a corto plazo para capturar la información temporal y estructural a corto plazo. Además, se propone un LSTM con atención descendida por intervalos para ensamblar la información temporal a largo plazo y los factores estáticos. Se realiza los experimentos utilizando los comportamientos de préstamo para predecir el comportamiento predeterminado del usuario. Los resultados demuestran la efectividad de TEMGNN en comparación con todas las líneas de base. (D. Wang et al., 2021)

Para dar un mejor entendimiento de las referencias literarias, se presenta la Tabla 1 detallando por cada referencia el tipo de dato, la metodología, los resultados obtenidos y las implicaciones generales de los modelos utilizados por los autores.

En la Tabla 2 se presenta un resumen de las fuentes primarias y secundarias del presente proyecto.

Tabla 1.Matriz de Referencias literarias

MATRIZ DE REFERENCIAS					
N°	**Referencia**	**Tipos de Datos Utilizados**	**Metodologías para el Análisis de Datos**	**Resultados**	**Implicaciones Gerenciales**
1	Wang, T., Liu, R., & Qi, G. (2022). Multi-classification assessment of bank personal credit risk based on multi-source information fusion	Datos de un banco comercial en China. Datos Estructurados.	Fusión de información, seis algoritmos de machine learning.	Alta precisión en evaluación multiclasificación, aplicabilidad universal	Mejora en la capacidad de identificación de riesgos y prevención de pérdidas financieras
2	Feng, B., Xu, H., Xue, W., & Xue, B. (2020). Every Corporation Owns Its Structure: Corporate Credit Ratings via Graph Neural Networks	Datos de clasificación de crédito de empresas públicas chinas. Datos Semi estrucutrados.	Redes neuronales de grafos (GNN).	Mejora en la precisión de calificación crediticia, superando métodos tradicionales	Optimización de la evaluación de riesgos corporativos y toma de decisiones de inversión
3	El Qadi El Haouari, A. (2023). An Explainable Artificial Intelligence Credit Rating System	Datos de crédito. Datos Estructurados.	AI explicable (XAI).	Mejora en la transparencia y explicabilidad de los modelos de calificación crediticia	Incremento en la confianza de las instituciones financieras en los modelos de AI

4	Jin, Y., Zhang, W., Wu, X., Liu, Y., & Hu, Z. (2021).A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data	Datos desequilibrados de puntuación de crédito Datos Estructurados.	Modelo de ensamble multietapa con algoritmo genético híbrido.	Mejora en la precisión de predicción, reducción del desequilibrio de datos	Mejora en la estabilidad y precisión de modelos de predicción de crédito
5	Noriega, J., Rivera, L., & Herrera, J. (2023).Machine Learning for Credit Risk Prediction: A Systematic Literature Review	Revisión sistemática de estudios sobre ML en predicción de riesgo crediticio Datos Estructurados.	Métodos de aprendizaje automático (Boosted Category, AUC, ACC, F1).	Identificación de modelos y métricas efectivas, limitaciones en representatividad de datos	Guía para desarrolladores de herramientas de gestión de riesgo crediticio
6	Liu, J., Zhang, S., & Fan, H. (2022).A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network	Datos de club de préstamos (2007-2016) Datos Estructurados.	Modelo híbrido de dos etapas con XGBoost y GNN.	Mejora en precisión, F1-score, y G-mean en comparación con benchmarks	Robustez contra ciclos económicos, mejor toma de decisiones crediticias
7	Chen, H., Yang, C., Du, M., & Zhang, Y. (2023). Research on Credit Risk Prediction under Unbalanced Dataset Based on Ensemble Learning	Datos de evaluación de riesgo crediticio Datos Estructurados.	Aprendizaje en conjunto, selección de características, LightGBM.	Mejora en G-mean y AUC, mejor efecto de predicción de impagos	Aumento de la precisión en la clasificación de riesgos y reducción de pérdidas por impagos

8	Addula, S. R., Meduri, K., Nadella, G. S., & Gonaygunta, H. (2024. AI and Blockchain in Finance: Opportunities and Challenges for the Banking Sector)	Datos financieros de bancos. Datos Semi estructurados.	AI y Blockchain.	Mejora en la gestión de riesgos financieros, mayor transparencia y seguridad	Adaptación a nuevas tecnologías, reducción de costos y mejora en la eficiencia operativa
9	Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., & Li, J. (2021). Intelligent financial fraud detection practices in post-pandemic era	Datos financieros y comportamiento de usuarios durante la pandemia. Datos Semi estructurados.	Métodos de detección de fraudes basados en GNN.	Identificación efectiva de fraudes financieros post-pandemia	Mejora en la seguridad financiera y prevención de fraudes en el contexto post-pandemia
10	Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020).Machine learning for financial risk management: A survey	Datos financieros diversos. Datos estrucuturados.	Encuesta sistemática sobre ML en gestión de riesgos.	Identificación de métodos de ML relevantes, tendencias emergentes y direcciones de investigación	Mejora en la toma de decisiones informadas, adaptación a cambios en el mercado financiero
11	Oualid, A., Hansali, A., Balouki, Y., & Moumoun, L. (2022).Application of Machine Learning Techniques for Credit Risk Management: A Survey	Datos de bancos. Datos estrucuturados.	Algoritmos de aprendizaje automático para gestión de riesgos.	Evaluación de técnicas de AI para mejorar la gestión de riesgos crediticios	Optimización de procesos bancarios y mejora en la toma de decisiones de crédito

12	Wang , D., Zhang, Z., Zhou, J., Cui, P., Fang, J., Jia, Q., Fang, Y., & Qi, Y. (2021).A Review on Graph Neural Network Methods in Financial Applications	Datos dinámicos de comportamiento de usuarios. Datos estrucuturados.	GNN temporal para predicción de riesgos crediticios.	Mejor rendimiento en la predicción de riesgos crediticios a lo largo del tiempo	Mejora en la evaluación continua de riesgos y adaptación a cambios dinámicos en el comportamiento de usuarios
13	Bi, W., Xu, B., Sun, X., Wang, Z., Shen, H., Cheng, X. (2022).Company-as-Tribe: Company Financial Risk Assessment on Tribe-Style Graph with Hierarchical Graph Neural Networks	Datos financieros de empresas y noticias financieras. Datos Semi estrucuturado.	Red Neural Jerárquica de Grafos (TH-GNN) en grafos estilo tribu.	Mejora significativa en la evaluación de riesgos financieros comparado con métodos anteriores	Evaluación temprana y precisa de riesgos financieros, reducción de pérdidas financieras
14	Wang, J., Zhang, S., Xiao, Y., Song, R. (2021).A Review on Graph Neural Network Methods in Financial Applications	Datos financieros presentados como grafos. Datos estructurados.	Revisión de modelos de redes neuronales de grafos (GNN).	Resumen de metodologías GNN y aplicaciones en finanzas	Orientación para futuras investigaciones y aplicaciones de GNN en finanzas
15	Zhou, B., Jin, J., Zhou, H., Zhou, X., Shi, L., Ma, J., Zheng, Z. (2023).Forecasting credit default risk with graph attention networks	Historial de crédito, estado de crédito y perfil personal. Datos estructurados.	Red de Atención de Grafos (GAT)	Predicción precisa del riesgo de impago, superando métodos de referencia	Mejora en la gestión de riesgos de crédito, precisión en la predicción de impagos

16	Wang, Y., Wang, J., Shang, J., Chen, Z., Ding, X., Wang, H. (2023).Learning Event Logic Graph Knowledge for Credit Risk Forecast	Noticias financieras estructuradas y series temporales de bonos. Datos semi estructurados.	Red lógica de eventos con red de convolución temporal.	Mejora en la predicción de riesgos de crédito corporativo respecto a métodos de referencia	Identificación temprana de riesgos crediticios, mitigación de impactos financieros de eventos adversos
17	Yang, S., Zhang, Z., Zhou, J., Wang, Y., Sun, W., Zhong, X., Fang, Y., Yu, Q., Qi, Y. (2020).Financial Risk Analysis for SMEs with Graph-based Supply Chain Mining	Relaciones interactivas entre PYMEs en la cadena de suministro. Datos estructurado.	Red neuronal de grafos espacial-temporal para minería de cadena de suministro.	Mejora en el análisis de riesgos financieros de PYMEs, superando problemas de deficiencia de datos	Optimización de la evaluación de riesgos para PYMEs, identificación de relaciones crediticias críticas
18	Leo, M., Sharma, S., Maddulety, K. (2019).Machine learning in banking risk management: A literature review	Datos relacionados con riesgos de crédito, mercado, operativos y de liquidez en bancos. Datos estructurados	Revisión de técnicas de machine learning aplicadas a la gestión de riesgos bancarios.	Identificación de técnicas de ML aplicadas a la gestión de riesgos bancarios, áreas de investigación insuficientemente exploradas	Mejora en la detección, medición, reporte y gestión de riesgos bancarios, identificación de áreas potenciales para futuras investigaciones
19	Reyes Morales, M.A. y Sosa, M. (2022). Modelo de puntuación crediticia para tarjeta de crédito en México: una aproximación logística	Datos de clientes de tarjeta de crédito de una institución mexicana. Datos esctructurados.	Regresión logística.	El modelo de puntuación crediticia tiene un alto nivel de predictibilidad y estabilidad	Proporciona una herramienta confiable y precisa para evaluar la probabilidad de incumplimiento, facilitando la toma de decisiones para el personal bancario

20	Pérez González, G. (s/f).Detección de transacciones fraudulentas en tarjetas de crédito mediante el uso de modelos de Machine Learning	Datos no simulados de transacciones con tarjetas de crédito. Datos estructurados.	Algoritmos de aprendizaje no supervisado (Machine Learning)	Tres modelos lograron un desempeño superior en la detección de anomalías comparado con otros modelos de la literatura	Mejora la detección de transacciones fraudulentas, lo cual puede ayudar a reducir pérdidas financieras y aumentar la seguridad en el uso de tarjetas de crédito

Fuente: Elaboración propia de los autores.

Tabla 2. Matriz de fuentes primarias y secundarias

MATRIZ DE FUENTES PRIMARIAS Y SECUNDARIAS				
Autor(es)	**Año**	**Título**	**Tipo de Documento**	**Fuente**
Pucha Gualoto, O. I.	2022	Desarrollo de un modelo de predicción basado en Algoritmos de Machine Learning para medir el riesgo crediticio.	Tesis	Quito : EPN
Fida, Alex Eduardo.	n.d.	Universidad de San Andrés Escuela de Administración y Negocios Desarrollo de score crediticio a través de redes neuronales y	Tesis	Universidad de San Andrés

		regresión logística.		
Gutierrez Portela, F., Rodríguez Cárdenas, S., Patiño Ospina, L. P., & Hernandez Aros, L.	2023	Estudio de la prevención y detección de fraudes financieros a través de técnicas de aprendizaje automático.	Artículo	CAFI, 6(1), 77–101. https://doi.org/10.23925/cafi.v6i1.58372
Hermitaño Castro, J. A.	2022a	Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática.	Artículo	Interfases, 015, 160–178. https://doi.org/10.26439/interfases2022.n015.5898
González, L. N.	2023	El impacto de la inteligencia artificial en los negocios.	Artículo	Difusiones, 25(25), 153–161. https://doi.org/10.5281/zenodo.10729342
Tadeo Espinoza, F. E., & Coral Ygnacio, M. A.	2023	Modelos para la evaluación de riego crediticio en el ámbito de la tecnología financiera: una revisión.	Artículo	TecnoLógicas, 26(58), e2679. https://doi.org/10.22430/22565337.2679

López Malca, J. C.	2019	Comparación de modelos de aprendizaje de máquina en la predicción del incumplimiento de pago en el sector de las microfinanzas.	Trabajo de investigación	Pontificia Universidad Católica del Perú
García Hernández, J. M., & Torres Moreno, W. N.	2023	Predicción de riesgo de impago en institución financiera usando modelos de machine learning.	Tesis de maestría	Universidad Tecnológica Centroamericana (UNITEC), Tegucigalpa, Honduras

***Fuente*:** Elaboración propia de los autores.

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

La manera en que las organizaciones recopilan y entienden los datos de los clientes ha cambiado drásticamente con la llegada del 'big data', una tecnología exponencial revolucionaria para entender las necesidades del cliente y ofrecer unas soluciones más efectivas y a la medida. Ecuador ha empezado a aplicar el 'big data' y las Instituciones Financieras buscan esta implementación y llegar a la transformación digital.

El big data es la tecnología que facilita la gestión ágil de vastas cantidades de información en constante cambio. Las tecnologías previas, como las bases de datos y estadísticas convencionales, resultan insuficientes frente al enorme volumen de datos generados por los clientes. Esta información necesita ser filtrada y analizada utilizando herramientas sofisticadas que emplean diversos algoritmos, adaptándose así a las necesidades específicas de la empresa.

La Inteligencia Artificial ha llegado para quedarse y tiene un enorme potencial para que las empresas revolucionen sus procesos internos y aborden proyectos de desarrollo global que les permitan implantar sus productos y servicios de manera simultánea en distintos países.

Antes las personas se fidelizaban con un solo banco, invirtiendo su dinero con ellos y realizando operaciones crediticias que permitan cubrir sus necesidades inmediatas. Pero el modelo ha cambiado. Las personas se informan antes de comprar o acceder a alguna oferta del mercado, es por eso por lo que los bancos ya no ofrecen un 'producto' sino soluciones que cada vez son más personalizadas a las necesidades de las personas.

Para comprender las necesidades individuales de manera más personalizada, es fundamental disponer de grandes volúmenes de datos, y aquí es donde entra en juego el big data. Con datos abundantes, verificados, analizados y desglosados, las instituciones financieras pueden tomar decisiones que mejoren significativamente la experiencia de sus clientes.

Para pasar de ser una idea a una realidad, el uso del big data implica el cambio de tres pilares en las organizaciones:

La infraestructura, donde la inversión en tecnología es clave para tener las herramientas necesarias para filtrar la información.
El fortalecimiento y capacitación del equipo humano aumentando sus destrezas.
La información, donde se forma un ecosistema informacional para determinar procesos y fuente de información.
El enfoque estratégico de las Instituciones Financieras hacia la analítica de datos desempeña un papel fundamental en su crecimiento. La capacidad de tomar decisiones informadas basadas en el análisis de datos ha contribuido significativamente a su posición destacada en el sector financiero.

Ante el desafío presentado por la pandemia en 2020, los bancos tuvieron que buscar alternativas y soluciones de pago antes las operaciones crediticias de los individuos y de las empresas en el contexto post-COVID. La iniciativa refleja la adaptabilidad y la respuesta proactiva del banco ante las cambiantes circunstancias, buscando optimizar la experiencia del cliente en un entorno financiero transformado por la pandemia.

La necesidad de financiamiento y refinanciamiento surge a medida que las personas, afectadas por el aislamiento social, buscan opciones seguras y eficientes para realizar operaciones financieras. En este contexto, las entidades bancarias buscan transformar por completo su proceso de calificación de créditos, implementando un modelo que analizará los datos históricos financieros y no financieros para predecir la probabilidad de incumplimiento de pago, permitiendo a las instituciones financieras tomar decisiones más informadas sobre las aprobaciones de créditos.

Para las personas que tienen un negocio o una empresa es imperativo reducir el número de impagos y mantener un historial crediticio bueno que les permita acceder a fuentes de financiamiento para capital de trabajo. Por lo cual es

importante para las entidades bancarias conocer los motivos generales y específicos que pueden llevar a incumplimientos de pagos de sus obligaciones.

Conforme lo determina la Norma para las Instituciones Financieras sobre la administración del Riesgo de Crédito indica en sus artículos lo siguiente:(X-de Gestión Y Administración De Riesgos, n.d.)

ARTICULO 6.- Las instituciones controladas deberán contar con un sistema para monitorear los niveles del riesgo de crédito en forma permanente a través de las diferentes metodologías adoptadas por cada entidad para cada modalidad de crédito (comercial, consumo, vivienda y microcrédito), dentro de las cuales se determinarán los principios y criterios generales para la evaluación del riesgo de crédito.

ARTICULO 7.- Las metodologías implantadas deben considerar la combinación de criterios cuantitativos y cualitativos, de acuerdo con la experiencia y las políticas estratégicas de la entidad; deben permitir monitorear y controlar la exposición crediticia de los diferentes portafolios. Esta metodología debe ser evaluada periódicamente a fin de garantizar la idoneidad de esta, al igual que la relevancia de las variables utilizadas.

La administración del portafolio de crédito incluye las siguientes etapas fundamentales: el otorgamiento que incluye las fases de evaluación, aprobación, instrumentación y desembolso; seguimiento; recuperación; y, control, para lo cual es necesario que las entidades establezcan:
7.1 Criterios, metodologías y sistemas internos de evaluación crediticia para la selección y otorgamiento de los créditos, que se ajusten al perfil de riesgo de la entidad, los que deben ser consistentes con la naturaleza, tamaño y complejidad de las operaciones de la institución controlada; y, estar basados en el análisis de los estados financieros, flujos de caja del proyecto, calidad de la gerencia, entre otros, para los clientes de los que se dispone de suficiente información financiera (créditos comerciales); y, en sistemas de evaluación crediticia, por

ejemplo: “credit scoring” para créditos a la microempresa y a las personas naturales (créditos de consumo y créditos para la vivienda);

Un sistema de seguimiento y control del riesgo de crédito de los diferentes portafolios, lo que implica un proceso continuo de calificación de los sujetos y operaciones coherente con el proceso de otorgamiento, que incluya un esquema para realizar el seguimiento del nivel de riesgo de cada sujeto y operación, sin perjuicio de lo dispuesto en el capítulo II “Calificación de activos de riesgo y constitución de provisiones por parte de las instituciones controladas por la Superintendencia de Bancos y Seguros”, título IX. Adicionalmente, el control del riesgo incorpora la adopción de medidas para mitigar los riesgos, cuando se identifican debilidades potenciales o reales en un cliente, tales como: reducción o transferencia de exposición, nuevas garantías, entre otras;

Metodologías y técnicas analíticas basadas en el comportamiento histórico de los portafolios de inversión y de las operaciones de crédito y contingentes, que permitan determinar la pérdida esperada sobre la base de la probabilidad de incumplimiento, el nivel de exposición y la severidad de la pérdida. Para el cálculo de estos componentes se deberá disponer de una base de datos mínima de tres (3) años inmediatos anteriores, que deberá contener elementos suficientes para el cálculo de los aspectos señalados en este numeral; y,

Un sistema de información basado en reportes objetivos, con información suficiente para satisfacer las necesidades de la institución, apoyar los procesos de toma de decisiones de la administración del riesgo de crédito y asegurar una revisión oportuna de las posiciones de riesgo y de las excepciones.

La información debe ser permanente, oportuna y consistente; y, ser distribuida a los niveles administrativos correspondientes para asegurar que se tomen acciones correctivas.

PLANTEAMIENTO DEL PROBLEMA

La necesidad de financiamiento y refinanciamiento surge a medida que las personas, afectadas por el aislamiento social, buscan opciones seguras y eficientes para realizar operaciones financieras. En este contexto, las entidades bancarias buscan transformar por completo su proceso de calificación de créditos, implementando un modelo que analizará los datos históricos financieros y no financieros para predecir la probabilidad de incumplimiento de pago, permitiendo a las instituciones financieras tomar decisiones más informadas sobre las aprobaciones de créditos.

Se ofrecerá un plan de implementación de tecnologías Big Data correctamente asociado a los procesos de negocio, haciendo uso de una adecuada selección de herramientas, para permitir a la empresa procesar los datos derivados del negocio y posteriormente generar información que soporte nuevos niveles de análisis para mejorar el conocimiento empresarial, lo cual a su vez generará grandes beneficios y mayor productividad dentro de los grupos de trabajo.

OBJETIVO GENERAL

- Determinar la factibilidad de implementar un modelo basado en inteligencia artificial para la otorgación de créditos.

OBJETIVOS ESPECÍFICOS

- Desarrollar el modelo de score crediticio basado en inteligencia artificial utilizando la herramienta de Power BI y Tableau.

- Analizar los resultados obtenidos del modelo de inteligencia artificial y determinar su aplicabilidad.

- Implementar un análisis macroeconómico de las instituciones financieras sobre el comportamiento de la mora durante los últimos 2 años.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

La razón para adoptar un enfoque analítico del problema se fundamenta en la viabilidad y utilidad que ofrece el análisis de datos mediante técnicas estadísticas y de aprendizaje automático para gestionar información compleja.

En el análisis de credit scoring, se pueden utilizar varios modelos estadísticos y de aprendizaje automático. Algunos de los modelos más comúnmente utilizados incluyen:

a) **Regresión Logística**: Es uno de los modelos más tradicionales para el credit scoring. Se utiliza para predecir la probabilidad de que un cliente cumpla con sus obligaciones de crédito.

b) **Random Forest:** Es una extensión de los árboles de decisión que utiliza múltiples árboles para mejorar la precisión y evitar el sobreajuste.

c) **Modelo XGBoost:** es un algoritmo de aprendizaje automático basado en árboles de decisión. Se ha destacado por su rendimiento y eficiencia en una amplia variedad de tareas de clasificación y regresión.

La metodología utilizada en el desarrollo de este trabajo está basada en la regresión logística y en el modelo XGBoost, explicaremos porque para este desarrollo consideramos importante los mismos:

a) REGRESIÓN LOGÍSTICA:

“La regresión logística es una técnica fundamental en el campo de la inteligencia artificial y el aprendizaje automático (AI/ML). Los modelos de ML son programas de software que se pueden entrenar para realizar tareas complejas de procesamiento de datos sin intervención humana. Los modelos de ML creados mediante regresión logística ayudan a las organizaciones a obtener información procesable a partir de sus datos empresariales. Esta información se puede utilizar para análisis predictivos con el fin de reducir costos operativos, aumentar

la eficiencia y escalar más rápidamente. Por ejemplo, las empresas pueden descubrir patrones que mejoran la retención de empleados o que conducen a un diseño de productos más rentable.

A continuación, enumeramos algunos beneficios del uso de la regresión logística en comparación con otras técnicas de ML.

1. Simplicidad

Los modelos de regresión logística son matemáticamente menos complejos que otros métodos de ML. Por lo tanto, puede implementarlos incluso si nadie de su equipo tiene una profunda experiencia en ML.

2. Velocidad

Los modelos de regresión logística pueden procesar grandes volúmenes de datos a alta velocidad porque requieren menos capacidad computacional, como memoria y potencia de procesamiento. Esto los hace ideales para que las organizaciones que están empezando con proyectos de ML obtengan ganancias rápidas.

3. Flexibilidad

Puede usar la regresión logística para encontrar respuestas a preguntas que tienen dos o más resultados finitos. También puede usarlo para preprocesar datos. Por ejemplo, puede ordenar los datos con un amplio rango de valores, como las transacciones bancarias, en un rango de valores más pequeño y finito mediante la regresión logística. A continuación, puede procesar este conjunto de datos más pequeño mediante el uso de otras técnicas de ML para obtener un análisis más preciso.

4. Visibilidad

El análisis de regresión logística ofrece a los desarrolladores una mayor visibilidad de los procesos de software internos que otras técnicas de análisis de

datos. La solución de problemas y la corrección de errores también son más fáciles porque los cálculos son menos complejos.

El análisis de regresión lineal funciona de la siguiente manera:

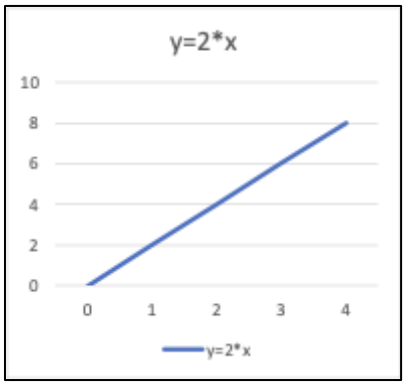
- Identificar la pregunta para obtener resultados concretos.
- Recopilar datos históricos.
- Entrenar el modelo de análisis de regresión.
- Realizar predicciones para valores desconocidos.

Para entender cómo funciona el modelo de regresión logística vamos a entender las ecuaciones y las variables.

5. Ecuaciones

En matemáticas, las ecuaciones dan la relación entre dos variables: x e y. Puede usar estas ecuaciones, o funciones, para trazar un gráfico a lo largo de los ejes x e y poniendo diferentes valores de x e y. Por ejemplo, si traza el gráfico para la función y = 2*x, obtendrá una línea recta como se muestra en el Grafico 1, por lo tanto, esta función también se denomina función lineal.

Ilustración 1. Ecuación del modelo de regresión logística



Fuente: *Elaboración de los autores*

6. Variables

En estadística, las variables son los factores o atributos de datos cuyos valores varían. Para cualquier análisis, ciertas variables son independientes o explicativas. Estos atributos son la causa de un resultado. Otras variables son variables dependientes o de respuesta; sus valores dependen de las variables independientes. En general, la regresión logística explora cómo las variables independientes afectan a una variable dependiente al observar los valores de datos históricos de ambas variables.

En nuestro ejemplo anterior, x se denomina variable independiente, variable predictora o variable explicativa porque tiene un valor conocido. Y se denomina variable dependiente, variable de resultado o variable de respuesta porque se desconoce su valor.

7. Función de regresión logística

La regresión logística es un modelo estadístico que utiliza la función logística, o función logit, en matemáticas como la ecuación entre x ey. La función logit mapea y como una función sigmoidea de x.

Ilustración 2. Función de regresión logística

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

Fuente: *Elaboración de los autores*

8. Análisis de regresión logística con múltiples variables independientes

En muchos casos, múltiples variables explicativas afectan al valor de la variable dependiente. Para modelar dichos conjuntos de datos de entrada, las fórmulas de regresión logística asumen una relación lineal entre las diferentes variables independientes. Puede modificar la función sigmoidea y calcular la variable de salida final como:

$y = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n)$

El símbolo β representa el coeficiente de regresión. El modelo logit puede calcular de forma inversa estos valores de coeficientes cuando se le proporciona un conjunto de datos experimentales suficientemente grande con valores conocidos de variables dependientes e independientes.

9. Registrar probabilidades

El modelo logit también puede determinar la relación entre el éxito y el fracaso o registrar las probabilidades. Por ejemplo, si estaba jugando al póker con sus amigos y ganó cuatro partidos de 10, sus probabilidades de ganar son cuatro sextos, o cuatro de seis, que es la relación entre su éxito y su fracaso. La probabilidad de ganar, por otro lado, es de cuatro sobre 10.

Matemáticamente, sus probabilidades en términos de probabilidad son p/(1 - p), y sus registros de probabilidades son (p/[1 - p]). Se puede representar la función logística como registro de probabilidades como se muestra a continuación:

$$\text{Logit Function} = \log\left(\frac{p}{1-p}\right)$$

Hay tres enfoques para el análisis de regresión logística basados en los resultados de la variable dependiente.

10. Regresión logística binaria

La regresión logística binaria funciona bien para problemas de clasificación binaria que solo tienen dos resultados posibles. La variable dependiente solo puede tener dos valores, como sí y no o 0 y 1.

Aunque la función logística calcula un rango de valores entre 0 y 1, el modelo de regresión binaria redondea la respuesta a los valores más cercanos. Por lo general, las respuestas por debajo de 0,5 se redondean a 0 y las respuestas por encima de 0,5 se redondean a 1, de modo que la función logística devuelve un resultado binario.

11. MODELO XGBOOST:

XGBoost (eXtreme Gradient Boosting) es un algoritmo de aprendizaje automático basado en árboles de decisión. Se ha destacado por su rendimiento y eficiencia en una amplia variedad de tareas de clasificación y regresión. Aquí explico sus características principales:

1. Gradient Boosting: XGBoost es una implementación optimizada del algoritmo de Gradient Boosting. Gradient Boosting construye modelos secuenciales, donde cada nuevo modelo intenta corregir los errores de los modelos anteriores. Los modelos se combinan para producir un predictor más fuerte.

2. Eficiencia Computacional: XGBoost está diseñado para ser altamente eficiente en términos de tiempo de entrenamiento y uso de recursos computacionales. Utiliza técnicas como la paralelización de cálculos, la poda de árboles para evitar el sobreajuste y la utilización de memoria fuera del núcleo (out-of-core) para manejar grandes conjuntos de datos que no caben en la memoria RAM.

3. Regularización: Una característica clave de XGBoost es la regularización, que ayuda a prevenir el sobreajuste. La regularización en XGBoost incluye términos que penalizan la complejidad del modelo, lo que mejora su capacidad para generalizar a datos no vistos.

4. Flexibilidad: XGBoost puede manejar tanto tareas de clasificación (por ejemplo, predicción de categorías) como de regresión (por ejemplo,

predicción de valores continuos). Además, soporta la personalización de funciones de pérdida y de evaluación, lo que permite adaptar el modelo a necesidades específicas.

5. Manejo de Valores Faltantes: XGBoost tiene la capacidad de manejar datos con valores faltantes de manera eficiente, determinando automáticamente las mejores rutas de los árboles de decisión en presencia de estos valores.

6. Peso de las Observaciones: Permite asignar pesos diferentes a las observaciones, lo que es útil cuando se trabaja con datos desequilibrados o con diferentes niveles de importancia en las observaciones.

7. XGBoost ha ganado popularidad en competiciones de ciencia de datos y se utiliza ampliamente en la industria debido a su alto rendimiento y flexibilidad para abordar diversos problemas de machine learning.

8. A continuación, presentamos las fases implementadas en la metodología utilizada:

RECOLECCIÓN DE DATOS

Para la aplicación de los modelos de machine learning, hemos recolectado una base de datos de la página web de Kaggle. Kaggle: Your Machine Learning and Data Science Community que tiene por nombre original: Home Credit Default Risk. (KAGGLE, 2018).

Esta base de datos contiene datos financieros y no financieros de los clientes que han solicitado préstamos a la entidad financiera Home Credit. Como se presenta en la Tabla 3, la base de datos consta de las siguientes tablas:

Tabla 3. Base de Datos

N°	Tablas	Información	Columnas	Datos	Volumen
1.-	**Application_Data**	Variable objetivo: Condición de Pago (binaria).	122 columnas que constan en el Anexo 1.	105 columnas con datos floats.	307511 filas
		Información respecto al préstamo, solicitud de préstamo y al tiempo de solicitud.		17 columnas con datos objects	
2.-	**POS_CASH_balance**	Saldos mensuales de préstamos previos de clientes de Home Credit.	8 columnas que constan en el Anexo 2.	7 columnas con datos inter o float.	10001358 filas
		Datos de comportamiento del préstamo.		1 columna con dato categórico.	

Fuente: Elaboración propia de los autores.

La primera tabla Application_Data, nos proporciona datos no financieros de los solicitantes de préstamos tales como: Género, si posee o no vehículos e inmuebles, cantidad de hijos, etc y datos financieros tales como: monto solicitado, precio de los bienes a adquirir con el préstamo, monto de la anualidad a pagar, monto de los ingresos anuales, etc. Se presenta la Tabla 4 con los nombres de las diez primeras variables y los datos de las primeras y últimas filas:

Tabla 4. Datos de la Tabla Applicaation Data

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5
...	...	...	...	...	...	...	...	...	...	...
307506	456251	0	Cash loans	M	N	N	0	157500.0	254700.0	27558.0
307507	456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	12001.5
307508	456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	29979.0
307509	456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	20205.0
307510	456255	0	Cash loans	F	N	N	0	157500.0	675000.0	49117.5

Fuente: Generado con Matplotblit a partir del script de Pyhton.

La segunda tabla POS_CASH_balance, nos proporciona datos no financieros de los solicitantes de préstamos tales como: el número de meses en cartera, cantidad de cuotas pagadas, cantidad de cuotas a pagar, estado del contrato,

días de retraso en el pago y un indicador de incumplimiento. Se presenta la Tabla 5 con los nombres de las diez primeras variables y los datos de las primeras y últimas filas:

Tabla 5. POS_CASH_balance

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRACT_STATUS	SK_DPD	SK_DPD_DEF
0	1803195	182943	-31	48.0	45.0	Active	0	0
1	1715348	367990	-33	36.0	35.0	Active	0	0
2	1784872	397406	-32	12.0	9.0	Active	0	0
3	1903291	269225	-35	48.0	42.0	Active	0	0
4	2341044	334279	-35	36.0	35.0	Active	0	0
...	...	...	...	...	...	...	...	...
10001353	2448283	226558	-20	6.0	0.0	Active	843	0
10001354	1717234	141565	-19	12.0	0.0	Active	602	0
10001355	1283126	315695	-21	10.0	0.0	Active	609	0
10001356	1082516	450255	-22	12.0	0.0	Active	614	0
10001357	1259607	174278	-52	16.0	0.0	Completed	0	0

Fuente: Generado con Matplotblit a partir del script de Pyhton.

LIMPIEZA, PRE-PROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS

Para la fase de limpieza, pre-procesamiento y/o transformación de datos se utiliza el lenguaje de programación Pyhton Software License.

En la fase de limpieza y transformación de las bases de datos aplicamos las acciones que se detallan en la Tabla 6:

Tabla 6. Acciones de limpieza, pre-procesamiento y transformación de datos

N	Acción	Códigos Python	Resultado	Antes	Después
1	Recodificación de variables del inglés al español.	df.rename({'variables';' traducción ' }, axis=1, inplace=True)	Variables traducidas	SK_ID_CURR MONTHS_BALANCE CNT_INSTALMENT CNT_INSTALMENT_FUTURE NAME_CONTRACT_STATUS SK_DPD SK_DPD_DEF	ID_CLIENTE MES_SALDO CANT_CUOTAS_PAGADAS CANT_CUOTAS_FUTURAS ESTADO_CONTRATO DIAS_RETRASO INDICADOR_INCUMPLIMIENTO
2	Conversión de columnas días a columnas fechas	df['variable'] = df_original['variable'].apply(pd.Timestamp('1900-01-01'))	4 columnas convertidas a datos con tipo Fecha.	DIAS_NACIMIENTO DIAS_EMPLEADO DIAS_REGISTRO DIAS_PUBLICACION_ID	FECHA_NACIMIENTO FECHA_EMPLEADO FECHA_REGISTRO FECHA_PUBLICACION_ID :1856-02-04 00:07:06.224753664

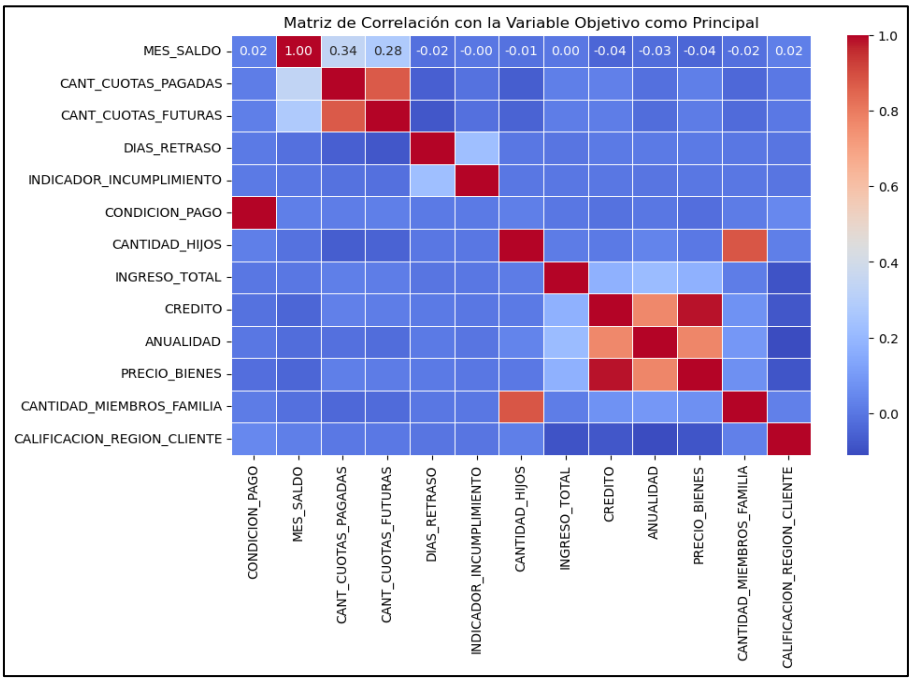
3	Identificación de valores nulos.	valores_nulos =df.isnull().sum() y barplot = sns.barplot(x='nulos', y='variable', data=df_nulos, palette='husl') df_originalinput['TIPO_SUITE'] = df_originalinput['variable'].replace([0, None, np.nan], 'Sin_respuesta') df_originalinput[columnas_a_reemplazar] = df_originalinput[columnas_a_reemplazar].fillna(0)	Por el gran volumen de datos se realiza un gráfico de barras con seaborn.	En la tabla Application Data se identifica 77 variables con datos nulos. Se identifica que estas variables tienen más del 50% de datos nulos. En la tabla POS_CASH_balance se identifican dos variables con 52158 datos nulos.	En la tabla Aplication data, al ser un gran volumen de datos se reemplaza con 0 en las variables numéricas y se reemplaza con el texto "Sin respuesta" en las variables categóricas. En la tabla POS_CASH_balance se eliminan los valores nulos.
4	Identificación de valores no nulos.	sin_valores_nulos = valores_nulos[valores_nulos == 0] y sns.barplot(x='Cantidad de Valores Nulos', y='Variable', data=sin_nulos_df, palette='viridis')	Por el gran volumen de datos se realiza un gráfico de barras con seaborn.	Se identifica 45 variables sin datos nulos. Se identifica que estas variables tienen 0% de datos nulos.	No aplica.
5	Selección de variables	df_seleccionado = df_originalinput[variables_seleccionadas]	Se eligen 20 variables de la tabla Application Data. Se eligen 8 columnas de la base POS_CASH_balance.	Tabla 1.- 122 columnas. Tabla 2. 8 columnas.	Tabla 1.- 20 columnas. Tabla 2. 8 columnas.
6	Unión de tablas	df5_merged = pd.merge(df_seleccionado4limpio, df_seleccionado, on='ID_CLIENTE', how='inner')	Una sola tabla.	Tabla 1.- 20 columnas. Tabla 2. 8 columnas.	Tabla 3. 27 columnas y 8521412 filas. Variable común es el código del cliente.
7	Identicación de datos atípicos en las variables numéricas de la tabla 3.	sns.histplot(df_seleccionado3[column], kde=True, color='cyan', bins=30) plt.axvline(df_seleccionado3[column].quantile(0.01), color='firebrick', linestyle='--', linewidth=2, label='Umbral Inferior (1%)') plt.axvline(df_seleccionado3[column].quantile(0.99), color='firebrick', linestyle='--', linewidth=2, label='Umbral Superior (99%)')	Se grafica histogramas con umbrales del 1% y del 99% para identificar valores atípicos.	Se identifica que en las 13 variables numéricas no presentan datos atípicos significativos, pues al ser una tabla con 8521412 filas, los valores atípicos que están fuera del umbral no generan cambios. Sin embargo, la variable objetivo presenta un gran desbalance de clases.	No se eliminan los datos atípicos.

Fuente: Elaboración propia de los autores.

VISUALIZACION DE LAS VARIABLES.

a) Para la visualización de las variables numéricas se genera la siguiente matriz de correlación como se presenta en el Gráfico 1:

Ilustración 3. Matriz de correlación de variables numéricas

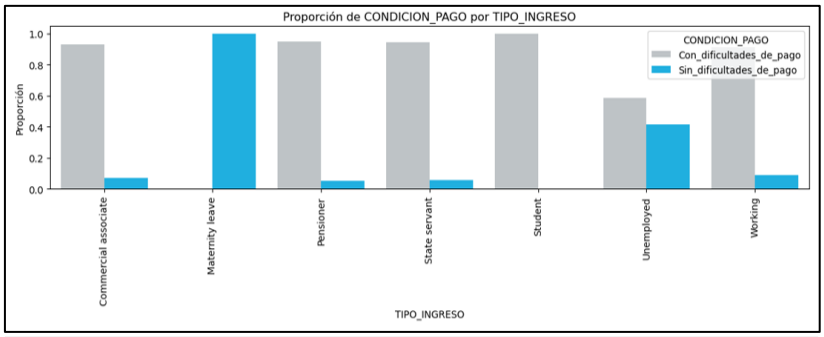


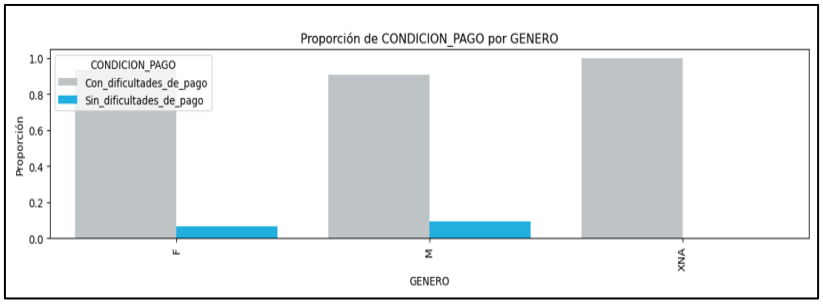
Fuente: Gráfico generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

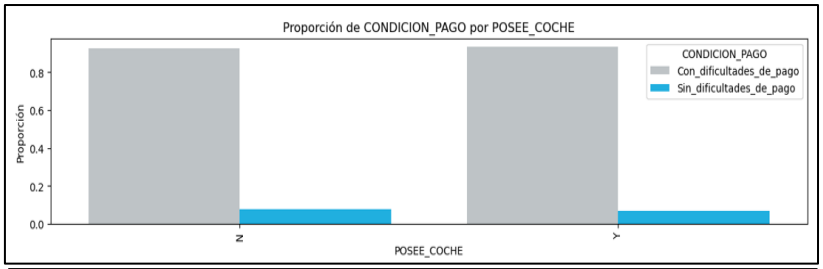
La variable dependiente CONDICION DE PAGO tiene una correlación de 0.34 con la variable cantidad de cuotas pagadas, 0.28 con cantidad de cuotas futuras, así como también con el monto del crédito, con la anualidad, con el precio de los bienes, con la cantidad de hijos y miembros de familia.

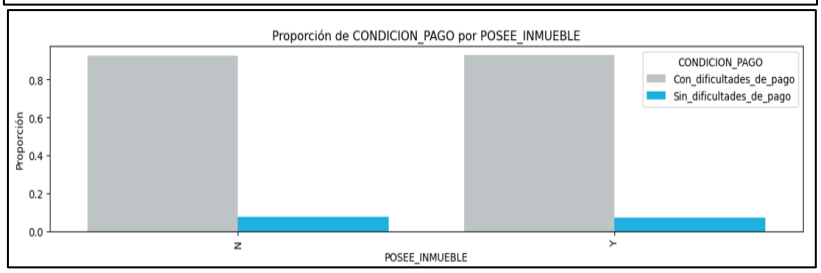
a) Para la visualización de las variables categóricas utilizamos gráficos de barras generados en Python por cada variable, tal como se visualiza en el Gráfico 4. que se muestra a continuación:

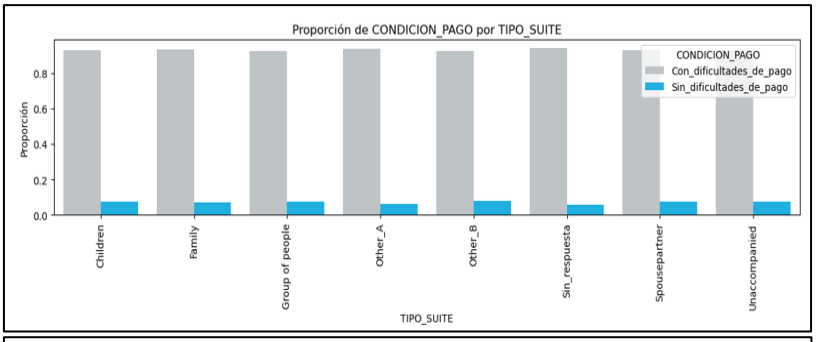
Ilustración 4. Gráfico de barras de las variables categóricas

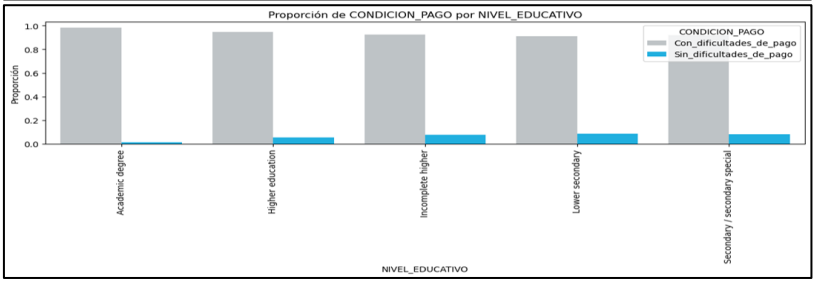

Proporción de CONDICION_PAGO por TIPO_INGRESO
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
TIPO_INGRESO

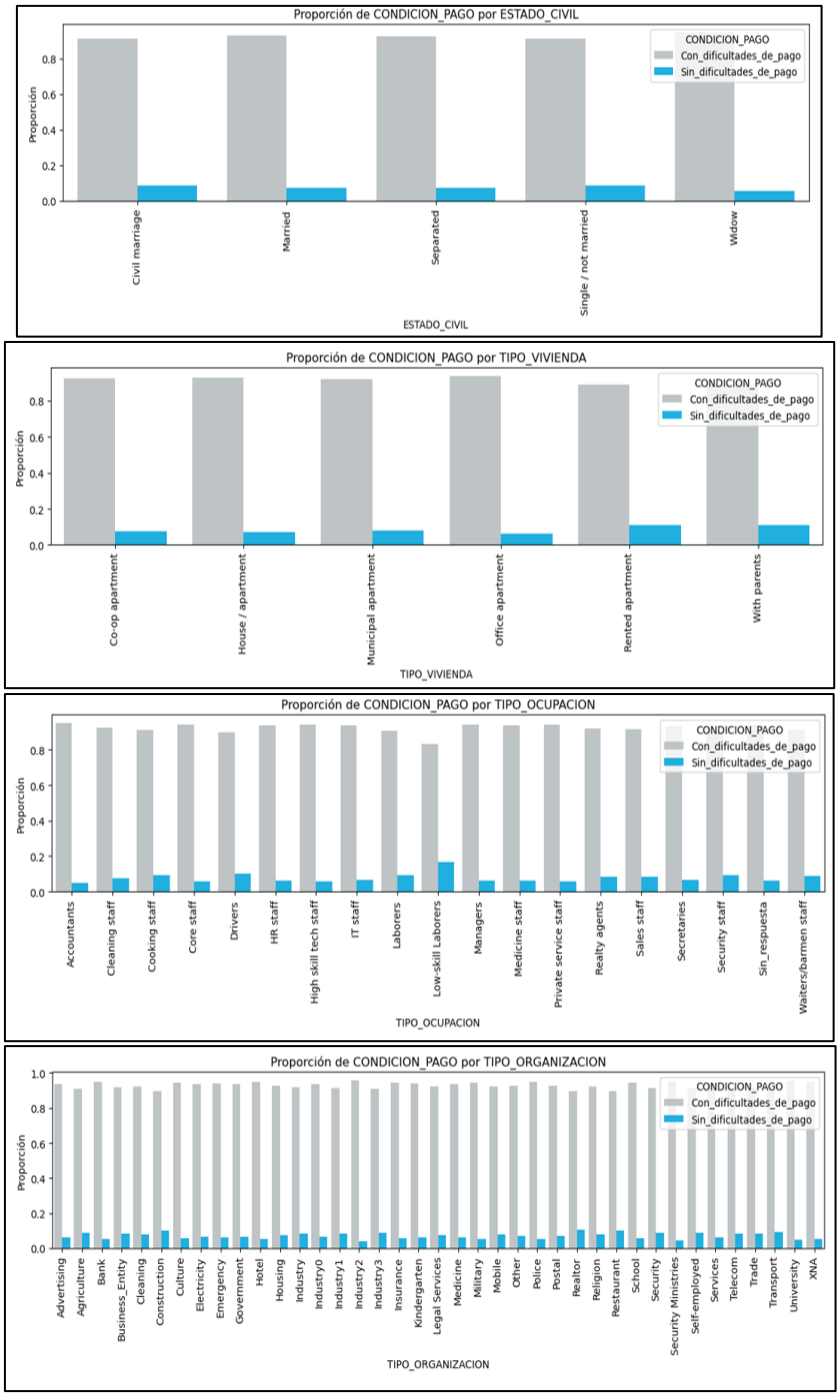

Proporción de CONDICION_PAGO por GENERO
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
F
M
XNA
GENERO


Proporción de CONDICION_PAGO por POSEE_COCHE
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
0.8
0.6
0.4
0.2
0.0
N
Y
POSEE_COCHE


Proporción de CONDICION_PAGO por POSEE_INMUEBLE
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
0.8
0.6
0.4
0.2
0.0
N
Y
POSEE_INMUEBLE


Proporción de CONDICION_PAGO por TIPO_SUITE
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
0.8
0.6
0.4
0.2
0.0
Children
Family
Group of people
Other_A
Other_B
Sin_respuesta
Spousepartner
Unaccompanied
TIPO_SUITE


Proporción de CONDICION_PAGO por NIVEL_EDUCATIVO
CONDICION_PAGO
Con_dificultades_de_pago
Sin_dificultades_de_pago
Proporción
1.0
0.8
0.6
0.4
0.2
0.0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NIVEL_EDUCATIVO



Fuente: Gráficos generados con Matplotlib a partir de la base de datos de las variables seleccionadas.

Según los gráficos de barras podemos identificar que los clientes sin dificultades de pago son en menor proporción que los que tienen dificultades de pago. Los clientes con dificultades de pago presentan en mayor proporción en todas las categorías de las variables.

SELECCIÓN DEL MODELO ESTADISTICO.

Para analizar nuestra base de datos seleccionamos dos modelos estadísticos:

1.- REGRESION LOGISTICA

Se elige aplicar el modelo de regresión logística porque permite clasificar a los solicitantes de crédito en: tiene dificultades de pago o no tiene dificultades de pago, es decir nuestra variable objetivo es binaria donde el objetivo es predecir una categoría basada en variables predictoras.

Se define la variable dependiente y las variables independientes:

Como se mencionó anteriormente se eligen 27 columnas y 8521412 filas donde:

Variable Dependiente:

CONDICION_PAGO. Esta variable binaria y numérica presenta dos categorías:
1: cliente sin dificultades de pago y;

0: clientes con dificultades de pago.

Variables Independientes:

En la Tabla 7 se presenta el vocabulario de las 27 variables independientes seleccionadas.

Tabla 7. Variables Independientes seleccionadas

N	Nombre	Descripción	Tipo
1	ID_CLIENTE	Identificador único del cliente	object
2	TIPO_CONTRATO	Tipo de contrato de crédito (por ejemplo, Cash loans, Revolving loans)	object
3	GENERO	Género del cliente (M: masculino, F: femenino)	object
4	POSEE_COCHE	Indicador de si el cliente posee un vehículo (Y: sí, N: no)	object
5	POSEE_INMUEBLE	Indicador de si el cliente posee una propiedad inmobiliaria (Y: sí, N: no)	object
6	CANTIDAD_HIJOS	Número de hijos del cliente	float64
7	INGRESO_TOTAL	Ingreso total anual del cliente	float64
8	CREDITO	Monto del crédito solicitado	float64

	9	ANUALIDAD	Monto de la anualidad del crédito	float64	
	10	PRECIO_BIENES	Precio de los bienes para los cuales se está solicitando el crédito	float64	
	11	TIPO_SUITE	Con quién vive el cliente (por ejemplo, Unaccompanied, Family, Spouse, etc.)	object	
	12	TIPO_INGRESO	Tipo de ingreso del cliente (por ejemplo, Working, State servant, Pensioner, etc.)	object	
	13	NIVEL_EDUCATIVO	Nivel de educación del cliente (por ejemplo, Higher education, Secondary education, etc.)	object	
	14	ESTADO_CIVIL	Estado civil del cliente (por ejemplo, Married, Single, Widow, etc.)	object	
	15	TIPO_VIVIENDA	Tipo de vivienda del cliente (por ejemplo, House, Rented apartment, etc.)	object	
	16	TIPO_OCUPACION	Ocupación del cliente (por ejemplo, Laborers, Core staff, Managers, etc.)	object	
	17	CANTIDAD_MIEMBROS_FAMILIA	Número de miembros de la familia del cliente	float64	
	18	CALIFICACION_REGION_CLIENTE	Calificación de la región del cliente (1: peor, 2: promedio, 3: mejor)	float64	
	19	TIPO_ORGANIZACION	Tipo de organización donde trabaja el cliente	object	
	20	ID_CLIENTE_PREV	Identificador único del cliente previo	inter64	
	21	ID_CLIENTE	Identificador único del cliente	inter64	
	22	MES_SALDO	Meses desde la solicitud	inter64	
	23	CANT_CUOTAS_PAGADAS	Número de pagos realizados	inter6	
	24	CANT_CUOTAS_FUTURAS	Número de pagos futuros pendientes	inter64	
	25	ESTADO_CONTRATO	Estado del contrato	inter64	
	26	DIAS_RETRASO	Días de retraso en el pago	float64	
	27	INDICADOR_INCUMPLIMIENTO	Días de retraso en el pago con definición	float64	

***Fuente*:** Elaboración propia de los autores.

Aplicando el modelo de Regresión Logística en la base de prueba obtuvimos el siguiente Reporte de Clasificación que se visualiza en la Tabla 8:

Tabla 8. Reporte de Clasificación – Modelo Regresión Logística

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.93	1.00	0.96	2368239
1	0.67	0.00	0.00	188185
accuracy			0.93	2556424
macro avg	0.80	0.50	0.48	2556424
weighted avg	0.91	0.93	0.89	2556424

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

El modelo tiene un alto rendimiento en la clase 0, con una precisión y recall perfectos para esta clase.

Sin embargo, tiene un desempeño muy deficiente en la clase 1, con un recall de 0.00, lo que indica que el modelo no está identificando correctamente las instancias de esta clase.

La exactitud global del 93% puede ser engañosa en situaciones de clases desbalanceadas, ya que el modelo puede estar sesgado hacia la clase mayoritaria (en este caso, la clase 0).

Para evitar el desbalanceo de clases se aplica el algoritmo SMOTE (Synthetic Minority Over-sampling Technique). Es una técnica de sobremuestreo utilizada para abordar el problema de desequilibrio de clases en conjuntos de datos y evitar el sesgo hacia la clase mayoritaria en el modelo. SMOTE crea nuevas instancias sintéticas de la clase minoritaria interpolando entre la instancia original y sus datos más cercanos.

El objetivo es mejorar las métricas de rendimiento en la clase minoritaria, como el recall y la precisión.

En este caso la clase minoritaria corresponde a la categoría de los clientes que no tienen dificultades de pago. Aunque la prioridad es predecir a los clientes que tienen dificultades de pago, no es menos importante los clientes que cumplen con el pago de sus préstamos y aplicar beneficios en sus transacciones.

A continuación, los resultados aplicando el logaritmo SMOTE:

Tabla 9. Reporte de Clasificación – Modelo Regresión Logística-SMOTE

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.95	0.59	0.73	2368239
1	0.10	0.59	0.18	188185
accuracy			0.59	2556424
macro avg	0.53	0.59	0.45	2556424
weighted avg	0.89	0.59	0.69	2556424

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

Como se puede visualizar en la Tabla 9. y realizando una comparación con los resultados anteriores, se observa que los resultados han cambiado significativamente, a continuación, analizamos en las métricas de precisión y recall:

- Precisión de la Clase 0: Aumentó de 0.93 a 0.95, lo que indica una mejora en la capacidad del modelo para identificar correctamente las instancias de la clase 0.
- Recall de la Clase 0: Disminuyó de 1.00 a 0.59, lo que sugiere que, aunque la precisión mejoró, el modelo ahora está identificando correctamente una menor proporción de instancias de la clase 0.
- Precisión de la Clase 1: Disminuyó de 0.67 a 0.10, lo que indica que el modelo ahora tiene una peor capacidad para identificar correctamente las instancias de la clase 1.
- Recall de la Clase 1: Aumentó de 0.00 a 0.59, lo que muestra que el modelo ahora está identificando una mayor proporción de instancias de la clase 1.

SMOTE ha ayudado a mejorar el recall de la clase minoritaria (1), pero a costa de una menor precisión en esa clase.

El modelo tiene una baja accuracy global después de aplicar SMOTE, lo que sugiere que el balance entre precisión y recall no ha sido óptimo.

Es crucial considerar si el modelo necesita un equilibrio entre precisión y recall o si se debe priorizar una métrica específica según el contexto del problema.

2.- RANDOM FOREST-XGBOOTS.

Elegimos aplicar el modelo de Random ForestXGboots porque nuestra base de datos es desbalanceada y este modelo permite:

a) Ajustar. - el parámetro “sclae_pos_weight” permite evitar un sesgo hacia la clase minoritaria. Esto significa que puede predecir con alta precisión la clase de los clientes que no tienen problemas de pagos.

b) En contestos desbalanceados, la precisión global(“accurancy) puede ser engañosa. Métricas como la precisión, recall, F1-score y AUC-ROC ofrecen una visión más completa del desempeño del modelo.

Con las mismas variables dependientes e independientes se aplica el modelo en la base de prueba obteniendo los resultados de la Tabla 10:

Tabla 10. Reporte de Clasificación – Modelo Random Forest – XGBoots

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.93	1.00	0.96	7894940
1	0.95	0.06	0.12	626472
accuracy			0.93	8521412
macro avg	0.94	0.53	0.54	8521412
weighted avg	0.93	0.93	0.90	8521412

AUC: **0.7790324455253456**

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

El modelo XGBoost muestra un buen desempeño en la clasificación de la clase mayoritaria (0) pero tiene un rendimiento muy pobre en la clase minoritaria (1). Esto es indicativo de un problema de desbalanceo de clases. A pesar de una alta accuracy y buenas métricas para la clase mayoritaria, el bajo recall y F1-score

para la clase 1 indican que el modelo está fallando en detectar adecuadamente las instancias de la clase minoritaria. Esto sugiere que podría ser beneficioso aplicar técnicas adicionales para manejar el desbalanceo de clases.
Un AUC de 0.779 sugiere una capacidad moderada del modelo para distinguir entre las dos clases. Aunque el modelo tiene un desempeño razonable en términos de clasificación general, hay margen para mejorar en la detección de la clase minoritaria.

Como en el modelo anterior, para evitar el desbalanceo de clases se aplica hiperparámetros que permitan ajustar antes de entrenar un modelo de machine learning.
A continuación de los Hiperparámetros utilizados (IBM, 2024):

- El número de iteraciones de impulso, denominadas nrounds. Nombre en XGBoost: n_estimators.
- La tasa de aprendizaje (también conocida como eta), que es un hiperparámetro que escala la contribución de cada árbol en el conjunto. Nombre en XGBoost: learning_rate.
- La profundidad máxima de un árbol, que es un parámetro de poda diseñado para controlar la profundidad total del árbol. Nombre en XGBoost: max_depth.
- Ajuste del Peso de Clase: Este parámetro multiplica el peso de las instancias de la clase positiva durante el entrenamiento. Por lo general, se usa cuando el número de instancias de la clase positiva es significativamente menor que el de la clase negativa. Nombre en XGBoost :scale_pos_weight.

En la base de entrenamiento se aplican los hiperparámetros obtenidos de la Tabla 11:

Tabla 11. Hiperparámetros del Modelo Random Forest – XGBoots

xgb.XGBClassifier	Valor encontrado
scale_pos_weight	1
n_estimators	100
max_depth	10
learning_rate	0,2

Fuente: Elaboración propia de los autores.

Aplicando los hiperparámetros se obtienen los siguientes resultados:

Tabla 12. Reporte de Clasificación – Modelo Random Forest – XGBoots - Hperparámetros

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.94	1.00	0.97	5526701
1	1.00	0.21	0.35	438287
Accuracy			0.94	5964988
macro avg	0.97	0.61	0.66	5964988
weighted avg	0.95	0.94	0.92	5964988

AUC: **0.9228975505917811**

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

Como se puede visualizar en la Tabla 12 y comparando los dos conjuntos de resultados nos da una visión clara sobre cómo diferentes configuraciones del modelo afectan el rendimiento:

Mejora en AUC: El AUC ha mejorado considerablemente en el modelo actual (0.923 frente a 0.779), lo que sugiere que el modelo actual tiene una mejor capacidad para distinguir entre las clases.

Mejora en Métricas Globales: La exactitud, precisión, recall y F1-Score en el promedio macro y ponderado también han mejorado, indicando un mejor rendimiento general del modelo.

Desempeño en la Clase Minoritaria: Aunque aún hay desafíos con la clase minoritaria, el modelo actual muestra mejoras en el recall y el F1-Score para esa clase en comparación con el modelo anterior.

En resumen, el modelo con los hiperparámetros muestra un mejor equilibrio en el rendimiento general y una mayor capacidad para distinguir entre las clases, aunque sigue enfrentando desafíos con la clase minoritaria.

RESULTADOS

Para proponer una solución efectiva a la transformación del proceso de calificación de créditos en entidades bancarias mediante la implementación de un modelo predictivo, se podría estructurarla de la siguiente manera:

Propuesta de Solución: Implementación de un Modelo Predictivo para la Calificación de Créditos

En respuesta a la necesidad de transformar el proceso de calificación de créditos, proponemos la implementación de un modelo predictivo avanzado basado en inteligencia artificial. Este modelo utilizará tanto datos históricos financieros como no financieros para predecir con mayor precisión la probabilidad de incumplimiento de pago por parte de los solicitantes de crédito.

1. **Recopilación y Preprocesamiento de Datos:**

- Datos Financieros: Incluirán información sobre ingresos, activos, deudas actuales, historial crediticio, entre otros.

- Datos No Financieros: Pueden abarcar variables como la estabilidad laboral, comportamiento de pago de servicios públicos, historial de empleo, y cualquier otro dato relevante que pueda proporcionar una visión más completa del perfil del solicitante.

2. **Selección y Construcción del Modelo Predictivo:**

- Selección de la Metodología: Consideraremos diversas técnicas de aprendizaje automático, como regresión logística para problemas de clasificación binaria.

- Entrenamiento del Modelo: Utilizaremos datos históricos etiquetados para entrenar y ajustar el modelo, optimizando los hiperparámetros para maximizar la precisión y minimizar el sesgo y la varianza.

3. **Validación y Evaluación del Modelo:**

- Validación Cruzada: Para asegurar la robustez y la generalización del modelo, aplicaremos técnicas de validación cruzada.
- Métricas de Evaluación: Utilizaremos métricas como precisión, recall, y el área bajo la curva ROC (AUC-ROC) para evaluar el rendimiento del modelo en la predicción del riesgo crediticio.

4. **Implementación y Monitoreo Continuo:**

- Integración en el Proceso de Toma de Decisiones: El modelo será integrado en el sistema existente de evaluación de créditos, proporcionando una puntuación de riesgo automatizada y basada en datos.
- Monitoreo y Actualización: Estableceremos un proceso de monitoreo continuo para recalibrar y actualizar el modelo a medida que se recopilen nuevos datos y se mejore la precisión predictiva.

5. **Consideraciones Éticas y Transparencia:**

- Interpretabilidad del Modelo: Aseguraremos que las decisiones del modelo sean interpretables, explicando cómo se llega a cada puntuación de riesgo.

- Ética en el Uso de Datos: Implementaremos medidas para garantizar la privacidad y seguridad de los datos de los clientes, cumpliendo con las regulaciones pertinentes (por ejemplo, GDPR).

6. **Beneficios Esperados:**

- Mejora significativa en la precisión de la evaluación del riesgo crediticio.
- Reducción de los costos operativos asociados con la evaluación manual de créditos.
- Mayor eficiencia en el proceso de toma de decisiones crediticias, permitiendo una respuesta más rápida a las solicitudes de crédito.

Esta propuesta no solo representa una oportunidad para mejorar la eficiencia y precisión en la evaluación de riesgo crediticio, sino que también posiciona a nuestra entidad bancaria en la vanguardia de la innovación tecnológica en el sector financiero. La implementación de este modelo predictivo no solo optimizará nuestros procesos internos, sino que también mejorará la experiencia del cliente al facilitar decisiones de crédito más informadas y rápidas.

Esta propuesta proporciona un marco claro para la implementación de un modelo predictivo avanzado, abordando tanto los aspectos técnicos como las consideraciones éticas y de cumplimiento. Puedes ajustar los detalles según las especificaciones y recursos disponibles en tu contexto particular.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

A continuación, se presentan los mejores resultados de los dos modelos aplicados:

REGRESION LOGISTICA:

Tabla 13. Reporte de Clasificación del modelo Regresión Logística final

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.95	0.59	0.73	2368239
1	0.10	0.59	0.18	188185
Accuracy			0.59	2556424
macro avg	0.53	0.59	0.45	2556424
weighted avg	0.89	0.59	0.69	2556424

ROC AUC Score:
0.6244444774099355

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

Tabla 14. Matriz de confusión del modelo Regresión Logística final

MATRIZ DE CONFUSION	
1402805	965434
77465	110720

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

A continuación, se presenta la interpretación de los resultados de la Tabla 13 y en la Tabla 14.

Precisión (Precision): La alta precisión para la clase 0 (0.94) y la clase 1 (1.00) indica que el modelo es confiable cuando realiza una predicción, siendo más preciso en la clase minoritaria. Sin embargo, esto no significa que el modelo sea equilibrado en cuanto a la identificación de ambas clases.

Recuperación (Recall): La recuperación perfecta para la clase 0 (1.00) muestra que el modelo identifica correctamente todos los verdaderos casos de la clase mayoritaria. En contraste, la baja recuperación para la clase 1 (0.21) sugiere que el modelo tiene problemas para detectar todos los casos verdaderos de la clase minoritaria, mostrando un desempeño deficiente en esta área.

F1-Score: El F1-Score alto para la clase 0 (0.97) refleja un buen equilibrio entre precisión y recuperación para la clase mayoritaria. En cambio, el F1-Score bajo para la clase 1 (0.35) indica un desequilibrio significativo en la capacidad del modelo para manejar la clase minoritaria, mostrando un desempeño limitado.

Precisión Global (Accuracy): 0.94
Aunque la precisión global es alta (0.94), esta métrica puede ser engañosa debido al desequilibrio en el número de casos entre las clases. La alta precisión global se debe principalmente al buen desempeño en la clase mayoritaria, sin reflejar adecuadamente el desempeño en la clase minoritaria.

Un AUC de 0.923 indica que el modelo tiene una buena capacidad para distinguir entre las dos clases en general. Sin embargo, el desempeño desigual en la identificación de la clase minoritaria sugiere que hay margen para mejorar el modelo, especialmente en la detección de la clase minoritaria

RANDOM FOREST – XGBOOTS

Tabla 15. Reporte de Clasificación del modelo Random Forest-XGBoots final

VARIABLE DEPENDIENTE	precision	recall	f1-score	support
0	0.94	1.00	0.97	5526701
1	1.00	0.21	0.35	438287
accuracy			0.94	5964988
macro avg	0.97	0.61	0.66	5964988
weighted avg	0.95	0.94	0.92	5964988

AUC: **0.9228975505917811**

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

Tabla 16. Matriz de confusión del modelo Random Forest-XGBoots final

MATRIZ DE CONFUSION	
5526365	336
344253	94034

Fuente: Generado con Matplotlib a partir de la base de datos de las variables seleccionadas.

A continuación, se presenta la interpretación de los resultados de la Tabla 15 y en la Tabla 16:

Precisión (Precision): La alta precisión para la clase 0 (0.94) y para la clase 1 (1.00) muestra que, cuando el modelo predice la clase, es muy confiable, aunque el modelo tiene dificultad para identificar correctamente la clase minoritaria (1).

Recuperación (Recall): La recuperación de la clase 0 es perfecta (1.00), pero la recuperación de la clase 1 es baja (0.21), lo que indica que el modelo tiene problemas para identificar todos los verdaderos casos de la clase minoritaria.

F1-Score: El F1-Score alto para la clase 0 y bajo para la clase 1 refleja un desempeño desigual entre las dos clases. La clase 0 tiene un buen equilibrio entre precisión y recuperación, mientras que la clase 1 no logra un buen equilibrio.

Accuracy (Precisión Global): La alta precisión global (0.94) se ve afectada por la desproporción en el número de casos entre las clases, y puede ser engañosa si se considera solo esta métrica.

ROC AUC Score: Con un AUC de 0.923, el modelo tiene una buena capacidad para distinguir entre las dos clases, a pesar de que muestra un desempeño desigual en la identificación de la clase minoritaria

En base a la actualidad del país, donde ha incrementado los clientes que no cancelan sus préstamos, y que en nuestra base representa a la clase mayoritaria

y tomando en cuenta que nuestro objetivo es predecir a los clientes que tienen dificultades de pagar las cuotas de los préstamos los dos modelos nos presentan resultados favorables, sin embargo el modelo de Random Forest XGBoots presenta mejores métricas de predicción, lo que permitirá a la institución financiera tener un score de crédito innovador utilizando las técnicas de machine learning.

PROPUESTA DE INNOVACIÓN Y COMPETITIVIDAD:

En base a los resultados proponemos las siguientes estrategias de innovación y competitividad para la institución financiera:

1.- Innovación en análisis de datos:

Incorporar datos externos, como información de redes sociales o datos económicos más claros, para obtener una visión más completa del cliente, su entorno y su historial crediticio en otras instituciones financieras.

Utilizar plataformas de big data para analizar grandes volúmenes de datos y extraer más información valiosa sobre los patrones de riesgo.

2.- Innovación en experiencia del cliente:

Aplicaciones móviles: Desarrollar aplicaciones móviles que permitan a los clientes gestionar sus préstamos, hacer pagos y recibir alertas sobre su situación de riesgo.

Interfaces de usuario mejoradas: Mejorar interfaces de usuario en la plataforma de préstamos para facilitar el acceso a la información y mejorar la interacción con el cliente.

3.- Capacitación y Cultura organizacional:

Entrenamiento en análisis de datos capacitando al personal en el uso de herramientas de análisis de datos y en la interpretación de resultados del modelo.

Desarrollo de habilidades en gestión de riesgos de los empleados.

Crear un entorno que fomente la innovación y la experimentación en la gestión del riesgo y el desarrollo de productos.

CONCLUSIONES

Al concluir la creación de un modelo predictivo basado en inteligencia artificial para la evaluación de riesgo crediticio, se pueden obtener varias conclusiones importantes:

1. **Capacidad predictiva:** Se puede evaluar la capacidad del modelo para predecir con precisión el riesgo de incumplimiento crediticio. Esto se mide generalmente mediante métricas como la precisión, el área bajo la curva ROC (AUC-ROC), y la sensibilidad/especificidad, entre otras.
2. **Importancia de las variables:** A través del modelo, se puede identificar qué variables tienen mayor impacto en la predicción del riesgo crediticio. Esto proporciona insights valiosos sobre qué aspectos financieros y de comportamiento son más relevantes para evaluar la solvencia de los clientes.
3. **Optimización continua:** Es crucial reconocer que el modelo predictivo no es estático. Requiere ajustes y optimizaciones periódicas para mantener su precisión y relevancia a lo largo del tiempo, especialmente ante cambios en los datos o en las condiciones económicas.
4. **Interpretación y explicabilidad:** La capacidad de explicar cómo el modelo toma decisiones es crucial, especialmente en entornos regulados o donde la transparencia es necesaria. Métodos como la importancia de características (feature importance) y la interpretación de modelos pueden ayudar a entender cómo se llega a una determinada predicción.
5. **Implementación y aceptación:** La implementación exitosa del modelo en un entorno de producción es una conclusión clave. Debe integrarse adecuadamente en los sistemas y procesos existentes de la institución financiera, asegurando que los usuarios finales (analistas, gestores de riesgos, etc.) confíen en las predicciones y las utilicen en la toma de decisiones.
6. **Impacto en la eficiencia y gestión de riesgos:** Un buen modelo predictivo puede mejorar significativamente la eficiencia operativa al automatizar procesos de evaluación de riesgo, así como ayudar a

gestionar de manera más efectiva la cartera de créditos al identificar riesgos potenciales de manera temprana.

7. **Ética y cumplimiento**: Es fundamental concluir que el uso de inteligencia artificial en la evaluación de riesgo crediticio debe cumplir con estándares éticos y regulatorios. Esto incluye el manejo adecuado de datos personales, la no discriminación y el cumplimiento de normativas como GDPR u otras leyes de protección de datos.

En resumen, la creación de un modelo predictivo basado en inteligencia artificial para la evaluación de riesgo crediticio puede proporcionar insights valiosos y mejoras significativas en la gestión de riesgos financieros, pero requiere un enfoque integral que abarque desde la construcción y validación del modelo hasta su implementación efectiva y cumplimiento ético y regulatorio.

RECOMENDACIONES

Crear un modelo predictivo basado en inteligencia artificial para la evaluación de riesgo crediticio es un proceso complejo que requiere atención meticulosa a varios aspectos clave. Aquí se presentan algunas recomendaciones importantes:

1. **Definir claramente el objetivo**: Es fundamental establecer qué se quiere predecir (por ejemplo, la probabilidad de incumplimiento de un crédito) y cómo el modelo ayudará a la toma de decisiones en la evaluación de riesgo crediticio.
2. **Recopilar datos relevantes y de alta calidad:** La calidad de los datos es crucial. Hay que asegurarse de tener acceso a datos históricos precisos y completos que sean relevantes para el modelo (por ejemplo, historiales de crédito, comportamiento de pagos, ingresos, etc.).
3. **Realizar un preprocesamiento exhaustivo de los datos**: Antes de construir el modelo, es necesario realizar tareas como limpieza de datos, manejo de valores atípicos, manejo de datos faltantes y normalización de variables para asegurar que los datos estén en condiciones óptimas para el modelado.
4. **Seleccionar el modelo adecuado:** Existen diversas técnicas de inteligencia artificial que se pueden emplear (como regresión logística, árboles de decisión, redes neuronales, etc.). La elección del modelo dependerá de factores como la naturaleza de los datos, la interpretabilidad requerida y la precisión deseada.
5. **Validar el modelo**: Es fundamental evaluar la capacidad predictiva del modelo utilizando técnicas de validación cruzada y pruebas en datos independientes. Esto ayuda a garantizar que el modelo sea robusto y generalice bien a nuevos datos.
6. **Interpretar los resultados**: Asegurarse de entender cómo el modelo toma decisiones y qué variables son más influyentes en las predicciones de riesgo crediticio. Esto es crucial para justificar las decisiones basadas en el modelo.

7. **Implementar y monitorear el modelo:** Después de desarrollar el modelo, implementarlo en un entorno de producción y establecer un mecanismo para monitorear su desempeño continuamente. Los modelos de inteligencia artificial pueden necesitar ajustes periódicos debido a cambios en los datos o condiciones del mercado.
8. **Garantizar la transparencia y la ética:** Garantizar de que el modelo cumpla con las regulaciones y normativas y que se apliquen principios éticos en el uso de datos y en la toma de decisiones basadas en el modelo.

BIBLIOGRAFIA

1. Hermitaño Castro, Juler Anderson. «Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática». *Interfases*, n.º 015 (29 de julio de 2022): 160-78. https://doi.org/10.26439/interfases2022.n015.5898.
2. Feng, Bojing, Haonan Xu, Wenfang Xue, y bindang Xue. «Every Corporation Owns Its Structure: Corporate Credit Ratings vía Graph Neural Networks», noviembre de 2020. http://arxiv.org/abs/2012.01933.
3. Wang, Tianhui, Renjing Liu, y Guohua Qi. «Multi-classification assessment of bank personal credit risk based on multi-source information fusion». *Expert Systems with Applications* 191 (abril de 2022). https://doi.org/10.1016/j.eswa.2021.116236.
4. Wang, Jianian, Sheng Zhang, Yanghua Xiao, y Rui Song. «A Review on Graph Neural Network Methods in Financial Applications», 26 de noviembre de 2021. http://arxiv.org/abs/2111.15367.
5. El Qadi El Haouari, Ayoub. «An Explainable Artificial Intelligence Credit Rating System», s. f. https://theses.hal.science/tel-04472510.
6. Jin, Yilun, Wenyu Zhang, Xin Wu, Yanan Liu, y Zeqian Hu. «A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data». *IEEE Access* 9 (2021): 143593-607. https://doi.org/10.1109/ACCESS.2021.3120086.
7. Wang, Yifan, Jibin Wang, Jing Shang, Zhuo Chen, Xuelian Ding, y Heyuan Wang. «Learning Event Logic Graph Knowledge for Credit Risk Forecast». En *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, 2023-July:345-50. Knowledge Systems Institute Graduate School, 2023. https://doi.org/10.18293/SEKE2023-222.
8. Wang, Daixin, Zhiqiang Zhang, Jun Zhou, Peng Cui, Jingli Fang, Quanhui Jia, Yanming Fang, y Yuan Qi. «Temporal-Aware Graph Neural Network for Credit Risk Prediction», 2021. https://epubs.siam.org/terms-privacy.
9. Superintendencia de Bancos y Seguros, Normas Generales para las Instituciones del Sistema Financiero, Capítulo II de la Administración del Riesgo Crediticio, resolución No JB-2003-602 de 9 de diciembre del 2003, páginas 578-579
10. ¿Qué es la regresión logística? - Explicación del modelo de regresión logística - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/logistic-regression/

11. GARCIA HERNANDEZ, JOSE MANUEL, y WALTHER NAHUN TORRES MORENO. «PREDICCIÓN DE RIESGO DE IMPAGO EN INSTITUCIÓN FINANCIERA USANDO MODELOS DE MACHINE LEARNING». Tesis para AL TÍTULO DE MÁSTER EN ANALÍTICA DE NEGOCIOS, 2023.

12. Reyes Morales, Marco Antonio, y Miriam Sosa. «Modelo de puntuación crediticia para tarjeta de crédito en México: una aproximación logística». *Ensayos Revista de Economía* 41, n.º 1 (18 de mayo de 2022): 17-52. https://doi.org/10.29105/ensayos41.1-2.

13. BBVA;(s.f); El big data: la nueva forma de entender a los clientes; https://www.bbva.com/es/innovacion/big-data-la-nueva-forma-de-entender-a-los clientes/

14. Baoss Analytics Everywhere;(s.f); 10 ejemplos de usos reales de Big Data Analytics; p.1;https://www.baoss.es/10-ejemplos-usos-reales-big-data/.

15. Indeed.com/orientación-profesional/desarrollo-profesional;(s.f); Importancia de la tecnología para empresas;https://www.indeed.com/orientacion-profesional/desarrollo profesional/tecnologia-empresas.

16. González, Y. M., & Rodríguez, A. I. (2021). Alfabetización en datos: Diseño de un nuevo escenario formativo para el contexto universitario. Revista Ibero Americana de Ciência da Informação, 14(1), Article https://www.researchgate.net/publication/348620598_Alfabetizacion_en_datos_Diseno_de_un_nuevo_escenario_formativo_para_el_contexto_universitario.

17. Lurillo, M. (2018). Logical Data Warehouse la nueva generación: Es hora de democratizar los datos | LinkedIn. https://www.linkedin.com/pulse/logical-datawarehouse-la-nueva-generaci%C3%B3n-es-hora-de-micheleiurillo/?originalSubdomain=es.

18. Gonzáles, E. (2012). Validación de la Teoría Unificada de Aceptación y Uso de la Tecnología UTAUT en castellano en el ámbito de las consultas externas de la Red de Salud Mental de Bizkaia.

19. https://openaccess.uoc.edu/bitstream/10609/19284/6/arzaTFM0213memoria.pdf.

20. Dirección de Educación en Línea. (2023, 15 febrero). Marco TOE Tendencias Tecnológicas en Big Data – Semana https://www.youtube.com/watch?v=9p8qJ0I9w6c. 2 [Vídeo]. YouTube.

21. • Andrés, M. B. (2022). Modelos de negocio basados en datos, publicidad programática,

22. inteligencia artificial y regulación: Algunas reflexiones. IDP. Revista de Internet, Derecho y Política, 36, 1–13.https://raco.cat/index.php/IDP/article/view/n36-barrio andres/497508.

23. IBM. (15 de MAYO de 2024). Desarrollador de IBM. Obtenido de Implementación de XGBoost en Python: https://developer.ibm.com/tutorials/awb-implement-xgboost-in-python/

24. KAGGLE. (2018). KAGGLE . Obtenido de Home Credit Default Risk: https://www.kaggle.com/competitions/home-credit-default-risk/data

ANEXOS

Tabla 17. TABLA DE APPLICATION_DATA

N	Variable	Definición	Tipo	Selección
0	SK_ID_CURR	Identificador único del cliente	object	SI
1	TARGET	Variable objetivo (1: cliente con dificultades de pago, 0: cliente sin dificultades de pago)	float64	SI
2	NAME_CONTRACT_TYPE	Tipo de contrato de crédito (por ejemplo, Cash loans, Revolving loans)	object	SI
3	CODE_GENDER	Género del cliente (M: masculino, F: femenino)	object	SI
4	FLAG_OWN_CAR	Indicador de si el cliente posee un vehículo (Y: sí, N: no)	object	SI
5	FLAG_OWN_REALTY	Indicador de si el cliente posee una propiedad inmobiliaria (Y: sí, N: no)	object	SI
6	CNT_CHILDREN	Número de hijos del cliente	float64	SI
7	AMT_INCOME_TOTAL	Ingreso total anual del cliente	float64	SI
8	AMT_CREDIT	Monto del crédito solicitado	float64	SI
9	AMT_ANNUITY	Monto de la anualidad del crédito	float64	SI
10	AMT_GOODS_PRICE	Precio de los bienes para los cuales se está solicitando el crédito	float64	SI
11	NAME_TYPE_SUITE	Con quién vive el cliente (por ejemplo, Unaccompanied, Family, Spouse, etc.)	object	SI
12	NAME_INCOME_TYPE	Tipo de ingreso del cliente (por ejemplo, Working, State servant, Pensioner, etc.)	object	SI
13	NAME_EDUCATION_TYPE	Nivel de educación del cliente (por ejemplo, Higher education, Secondary education, etc.)	object	SI

14	NAME_FAMILY_STATUS	Estado civil del cliente (por ejemplo, Married, Single, Widow, etc.)	object	SI
15	NAME_HOUSING_TYPE	Tipo de vivienda del cliente (por ejemplo, House, Rented apartment, etc.)	object	SI
16	REGION_POPULATION_RELATIVE	Población relativa de la región donde vive el cliente	float64	NO
17	DAYS_BIRTH	Edad del cliente en días (negativo: número de días desde el nacimiento)	float64	NO
18	DAYS_EMPLOYED	Días de empleo del cliente (negativo: número de días desde que empezó a trabajar)	float64	NO
19	DAYS_REGISTRATION	Días desde el registro del cliente (negativo: número de días desde el registro)	float64	NO
20	DAYS_ID_PUBLISH	Días desde que se publicó la identificación del cliente (negativo: número de días desde la publicación)	float64	NO
21	OWN_CAR_AGE	Edad del coche del cliente (en años)	float64	SI
22	FLAG_MOBIL	Indicador de si el cliente tiene un teléfono móvil (1: sí, 0: no)	float64	NO
23	FLAG_EMP_PHONE	Indicador de si el cliente tiene un teléfono de trabajo (1: sí, 0: no)	float64	NO
24	FLAG_WORK_PHONE	Indicador de si el cliente tiene un teléfono fijo de trabajo (1: sí, 0: no)	float64	NO
25	FLAG_CONT_MOBILE	Indicador de si el cliente tiene un teléfono móvil continuo (1: sí, 0: no)	float64	NO
26	FLAG_PHONE	Indicador de si el cliente tiene un teléfono fijo (1: sí, 0: no)	float64	NO
27	FLAG_EMAIL	Indicador de si el cliente tiene correo electrónico (1: sí, 0: no)	float64	NO
28	OCCUPATION_TYPE	Ocupación del cliente (por ejemplo, Laborers, Core staff, Managers, etc.)	object	SI
29	CNT_FAM_MEMBERS	Número de miembros de la familia del cliente	float64	SI
30	REGION_RATING_CLIENT	Calificación de la región del cliente (1: peor, 2: promedio, 3: mejor)	float64	NO

31	REGION_RATING_CLIENT_W_CITY	Calificación de la región del cliente con la ciudad (1: peor, 2: promedio, 3: mejor)	float64	NO
32	WEEKDAY_APPR_PROCESS_START	Día de la semana en que se inició el proceso de solicitud	object	NO
33	HOUR_APPR_PROCESS_START	Hora en que se inició el proceso de solicitud	float64	NO
34	REG_REGION_NOT_LIVE_REGION	Indicador de si la región de registro es diferente a la región de residencia (1: sí, 0: no)	float64	NO
35	REG_REGION_NOT_WORK_REGION	Indicador de si la región de registro es diferente a la región de trabajo (1: sí, 0: no)	float64	NO
36	LIVE_REGION_NOT_WORK_REGION	Indicador de si la región de residencia es diferente a la región de trabajo (1: sí, 0: no)	float64	NO
37	REG_CITY_NOT_LIVE_CITY	Indicador de si la ciudad de registro es diferente a la ciudad de residencia (1: sí, 0: no)	float64	NO
38	REG_CITY_NOT_WORK_CITY	Indicador de si la ciudad de registro es diferente a la ciudad de trabajo (1: sí, 0: no)	float64	NO
39	LIVE_CITY_NOT_WORK_CITY	Indicador de si la ciudad de residencia es diferente a la ciudad de trabajo (1: sí, 0: no)	float64	NO
40	ORGANIZATION_TYPE	Tipo de organización donde trabaja el cliente	object	SI
41	EXT_SOURCE_1	Fuente externa 1 (valor numérico)	float64	NO
42	EXT_SOURCE_2	Fuente externa 2 (valor numérico)	float64	NO
43	EXT_SOURCE_3	Fuente externa 3 (valor numérico)	float64	NO
44	APARTMENTS_AVG	Promedio de apartamentos	float64	NO
45	BASEMENTAREA_AVG	Promedio de área del sótano	float64	NO
46	YEARS_BEGINEXPLUATATION_AVG	Promedio de años desde el inicio de la explotación	float64	NO
47	YEARS_BUILD_AVG	Promedio de años desde la construcción	float64	NO

48	COMMONAREA_AVG	Promedio de área común	float64	NO
49	ELEVATORS_AVG	Promedio de elevadores	float64	NO
50	ENTRANCES_AVG	Promedio de entradas	float64	NO
51	FLOORSMAX_AVG	Promedio de pisos máximos	float64	NO
52	FLOORSMIN_AVG	Promedio de pisos mínimos	float64	NO
53	LANDAREA_AVG	Promedio de área de terreno	float64	NO
54	LIVINGAPARTMENTS_AVG	Promedio de apartamentos habitables	float64	NO
55	LIVINGAREA_AVG	Promedio de área habitable	float64	NO
56	NONLIVINGAPARTMENTS_AVG	Promedio de apartamentos no habitables	float64	NO
57	NONLIVINGAREA_AVG	Promedio de área no habitable	float64	NO
58	APARTMENTS_MODE	Modo de apartamentos	float64	NO
59	BASEMENTAREA_MODE	Modo de área del sótano	float64	NO
60	YEARS_BEGINEXPLUATATION_MODE	Modo de años desde el inicio de la explotación	float64	NO
61	YEARS_BUILD_MODE	Modo de años desde la construcción	float64	NO
62	COMMONAREA_MODE	Modo de área común	float64	NO
63	ELEVATORS_MODE	Modo de elevadores	float64	NO
64	ENTRANCES_MODE	Modo de entradas	float64	NO
65	FLOORSMAX_MODE	Modo de pisos máximos	float64	NO
66	FLOORSMIN_MODE	Modo de pisos mínimos	float64	NO
67	LANDAREA_MODE	Modo de área de terreno	float64	NO

68	LIVINGAPARTMENTS_MODE	Modo de apartamentos habitables	float64	NO
69	LIVINGAREA_MODE	Modo de área habitable	float64	NO
70	NONLIVINGAPARTMENTS_MODE	Modo de apartamentos no habitables	float64	NO
71	NONLIVINGAREA_MODE	Modo de área no habitable	float64	NO
72	APARTMENTS_MEDI	Mediana de apartamentos	float64	NO
73	BASEMENTAREA_MEDI	Mediana de área del sótano	float64	NO
74	YEARS_BEGINEXPLUATATION_MEDI	Mediana de años desde el inicio de la explotación	float64	NO
75	YEARS_BUILD_MEDI	Mediana de años desde la construcción	float64	NO
76	COMMONAREA_MEDI	Mediana de área común	float64	NO
77	ELEVATORS_MEDI	Mediana de elevadores	float64	NO
78	ENTRANCES_MEDI	Mediana de entradas	float64	NO
79	FLOORSMAX_MEDI	Mediana de pisos máximos	float64	NO
80	FLOORSMIN_MEDI	Mediana de pisos mínimos	float64	NO
81	LANDAREA_MEDI	Mediana de área de terreno	float64	NO
82	LIVINGAPARTMENTS_MEDI	Mediana de apartamentos habitables	float64	NO
83	LIVINGAREA_MEDI	Mediana de área habitable	float64	NO
84	NONLIVINGAPARTMENTS_MEDI	Mediana de apartamentos no habitables	float64	NO
85	NONLIVINGAREA_MEDI	Mediana de área no habitable	float64	NO
86	FONDKAPREMONT_MODE	Modo de fondos para reparaciones	object	NO

87	HOUSETYPE_MODE	Modo de tipo de casa	object	NO
88	TOTALAREA_MODE	Modo de área total	float64	NO
89	WALLSMATERIAL_MODE	Modo de material de las paredes	object	NO
90	EMERGENCYSTATE_MODE	Modo de estado de emergencia	object	NO
91	OBS_30_CNT_SOCIAL_CIRCLE	Número de observaciones del círculo social en 30 días	float64	NO
92	DEF_30_CNT_SOCIAL_CIRCLE	Número de incumplimientos del círculo social en 30 días	float64	NO
93	OBS_60_CNT_SOCIAL_CIRCLE	Número de observaciones del círculo social en 60 días	float64	NO
94	DEF_60_CNT_SOCIAL_CIRCLE	Número de incumplimientos del círculo social en 60 días	float64	NO
95	DAYS_LAST_PHONE_CHANGE	Días desde el último cambio de teléfono (negativo: número de días desde el cambio)	float64	NO
96	FLAG_DOCUMENT_2	Indicador de documento 2 (1: sí, 0: no)	float64	NO
97	FLAG_DOCUMENT_3	Indicador de documento 3 (1: sí, 0: no)	float64	NO
98	FLAG_DOCUMENT_4	Indicador de documento 4 (1: sí, 0: no)	float64	NO
99	FLAG_DOCUMENT_5	Indicador de documento 5 (1: sí, 0: no)	float64	NO
100	FLAG_DOCUMENT_6	Indicador de documento 6 (1: sí, 0: no)	float64	NO
101	FLAG_DOCUMENT_7	Indicador de documento 7 (1: sí, 0: no)	float64	NO
102	FLAG_DOCUMENT_8	Indicador de documento 8 (1: sí, 0: no)	float64	NO
103	FLAG_DOCUMENT_9	Indicador de documento 9 (1: sí, 0: no)	float64	NO
104	FLAG_DOCUMENT_10	Indicador de documento 10 (1: sí, 0: no)	float64	NO
105	FLAG_DOCUMENT_11	Indicador de documento 11 (1: sí, 0: no)	float64	NO
106	FLAG_DOCUMENT_12	Indicador de documento 12 (1: sí, 0: no)	float64	NO

107	FLAG_DOCUMENT_13	Indicador de documento 13 (1: sí, 0: no)	float64	NO
108	FLAG_DOCUMENT_14	Indicador de documento 14 (1: sí, 0: no)	float64	NO
109	FLAG_DOCUMENT_15	Indicador de documento 15 (1: sí, 0: no)	float64	NO
110	FLAG_DOCUMENT_16	Indicador de documento 16 (1: sí, 0: no)	float64	NO
111	FLAG_DOCUMENT_17	Indicador de documento 17 (1: sí, 0: no)	float64	NO
112	FLAG_DOCUMENT_18	Indicador de documento 18 (1: sí, 0: no)	float64	NO
113	FLAG_DOCUMENT_19	Indicador de documento 19 (1: sí, 0: no)	float64	NO
114	FLAG_DOCUMENT_20	Indicador de documento 20 (1: sí, 0: no)	float64	NO
115	FLAG_DOCUMENT_21	Indicador de documento 21 (1: sí, 0: no)	float64	NO
116	AMT_REQ_CREDIT_BUREAU_HOUR	Cantidad de solicitudes a la oficina de crédito en la última hora	float64	NO
117	AMT_REQ_CREDIT_BUREAU_DAY	Cantidad de solicitudes a la oficina de crédito en el último día	float64	NO
118	AMT_REQ_CREDIT_BUREAU_WEEK	Cantidad de solicitudes a la oficina de crédito en la última semana	float64	NO
119	AMT_REQ_CREDIT_BUREAU_MON	Cantidad de solicitudes a la oficina de crédito en el último mes	float64	NO
120	AMT_REQ_CREDIT_BUREAU_QRT	Cantidad de solicitudes a la oficina de crédito en el último trimestre	float64	NO
121	AMT_REQ_CREDIT_BUREAU_YEAR	Cantidad de solicitudes a la oficina de crédito en el último año	float64	NO

Tabla 18. POS_CASH_balance

N°	Variable	Descripción	Tipo de Dato	Selección
1	SK_ID_PREV	Identificador único de la solicitud previa	inter64	SI
2	SK_ID_CURR	Identificador único del cliente	inter64	SI
3	MONTHS_BALANCE	Meses desde la solicitud	inter64	SI
4	CNT_INSTALMENT	Número de pagos realizados	floar64	SI
5	CNT_INSTALMENT_FUTURE	Número de pagos futuros pendientes	floar64	SI
6	NAME_CONTRACT_STATUS	Estado del contrato	Categórico	SI
7	SK_DPD	Días de retraso en el pago	inter64	SI
8	SK_DPD_DEF	Días de retraso en el pago con definición	inter64	SI

Fuente: Elaboración de los autores