Construcción de scores crediticios utilizando modelos de aprendizaje automático

Artículo presentado como requisito parcial para el curso
Herramientas de Datos 2

Profesor: M.Sc. Luis Alberto Juárez Potoy

Autores:

Joseph Romero Chinchilla - C37006
Sebastián Astúa Morales - C10763
Pablo Alonso Chaves Madrigal - B82150

Universidad de Costa Rica
Departamento de Ciencias Actuariales y Matemáticas Puras
Julio, 2025
Código de GitHub: https://github.com/Q-V3pv/credit-risk-analysis.git

1. Introducción

Los *scores* crediticios constituyen herramientas esenciales para las entidades financieras, ya que permiten evaluar la solvencia de los solicitantes de crédito. Según MyFICO (s.f.), "un puntaje crediticio informa a los prestamistas sobre su solvencia, es decir, la probabilidad de que pague un préstamo según su historial crediticio. Se calcula utilizando la información de sus informes crediticios" (MyFICO, n.d.).

Estos puntajes se construyen a partir de variables como el historial de pagos, la utilización del crédito disponible, la antigüedad del historial crediticio, la diversidad de tipos de crédito y otros factores relevantes. Esta segmentación permite a las instituciones tomar decisiones más informadas sobre la aprobación de solicitudes, el monto otorgado y las condiciones del crédito (Anderson, 2007; Group, 2019).

En los últimos años, el uso de modelos estadísticos y algoritmos de aprendizaje automático ha transformado la forma en que se construyen y aplican los *scores* crediticios. Técnicas como *Random Forest*, *Gradient Boosting* y *XGBoost* han demostrado ser eficaces para predecir el riesgo de impago, especialmente en contextos donde los datos son complejos o desbalanceados (Chang, Sivakulasingam & Wang, 2024; Fernandez Vidal & Barbon, 2019).

Este trabajo se enfoca en el desarrollo de modelos predictivos que estimen la probabilidad de impago de clientes, utilizando variables relevantes y técnicas modernas de aprendizaje automático.

> ¿Qué características están asociadas con una mayor probabilidad de impago, según un modelo actuarial basado en modelos de predicción?

2. Metodología

La base de datos utilizada en este estudio fue obtenida de la plataforma *Kaggle*, específicamente del conjunto denominado "Credit Risk Dataset" (Kaggle Dataset, 2025). Esta base contiene información financiera y crediticia de clientes, adecuada para la construcción y evaluación de modelos de riesgo crediticio.

Este estudio adopta un enfoque cuantitativo, utilizando técnicas de aprendizaje supervisado para modelar la probabilidad de impago de clientes en función de sus características demográfi-

cas, financieras y comportamentales. Se emplearon tres algoritmos ampliamente validados en la literatura: **Random Forest**, **Gradient Boosting** y **XGBoost**, seleccionados por su capacidad para manejar relaciones no lineales, interacciones entre variables y conjuntos de datos desbalanceados (Chang, Sivakulasingam & Wang, 2024; Wang et al., 2025).

La variable `historial_impago` fue excluida del conjunto de entrenamiento para evitar *data leakage*, es decir, la incorporación de información futura que podría sesgar el modelo (Kuhn & Johnson, 2013). El preprocesamiento incluyó codificación *one-hot* para variables categóricas, normalización de variables numéricas y balanceo de clases mediante la técnica `SMOTE` (Chawla et al., 2002), que genera observaciones sintéticas de la clase minoritaria para mejorar la capacidad de generalización del modelo (Petropoulos et al., 2018).

Para el modelo XGBoost se aplicaron pesos personalizados utilizando `compute_sample_weight`, lo que permitió ajustar la penalización por errores en función del desequilibrio de clases y mejorar la sensibilidad del modelo ante casos de impago (Chang, Sivakulasingam & Wang, 2024; Chang, Wang & Luo, 2024).

2.1. Herramientas y tecnologías

Python: Lenguaje de programación ampliamente utilizado en ciencia de datos y análisis predictivo por su sintaxis clara y ecosistema robusto. En este estudio se utilizó para todo el procesamiento, modelado y visualización. Se emplearon las siguientes bibliotecas clave:

- `pandas`: Para manipulación de datos tabulares, limpieza de variables y estructuración de conjuntos de entrenamiento y prueba .
- `scikit-learn`: Para la implementación de algoritmos como `RandomForestClassifier` y `GradientBoostingClassifier`, partición de datos (`train_test_split`), y evaluación del rendimiento (`accuracy_score`, `classification_report`, `confusion_matrix`) .
- `xgboost`: Para la implementación eficiente y optimizada del algoritmo XGBoost, incluyendo ajustes de hiperparámetros y balanceo con `XGBClassifier` .
- `imblearn`: Para el balanceo de clases utilizando `SMOTE` (*Synthetic Minority Oversampling Technique*), lo cual mejora la sensibilidad del modelo frente a clases minoritarias.

- `matplotlib` y `seaborn`: Para visualización de resultados y métricas, incluyendo gráficos de matrices de confusión, distribución de errores y relevancia de variables (**hunter2007matplotl waskom2021seaborn**).
- `sklearn.utils.class_weight`: Para aplicar pesos a las muestras según el desbalance de clases mediante la función `compute_sample_weight`, especialmente útil en el modelo XGBoost.

GitHub: Plataforma de control de versiones utilizada para almacenar y gestionar el código fuente del proyecto, promoviendo la colaboración, documentación y auditoría del flujo de .

2.2. Descripción de los modelos

Random Forest. Algoritmo de ensamblado que construye múltiples árboles de decisión entrenados sobre subconjuntos aleatorios del conjunto de datos y de las variables. La predicción final se obtiene por agregación (votación mayoritaria) (Chang, Wang & Luo, 2024; Freire López, 2020).

Gradient Boosting. Algoritmo secuencial que construye árboles de decisión de forma iterativa, corrigiendo los errores de los modelos anteriores. (Clark & Lee, 2025; Petropoulos et al., 2018).

XGBoost. Implementación optimizada de Gradient Boosting que incorpora regularización L1 y L2, manejo automático de valores faltantes, y paralelización del entrenamiento. Es especialmente eficaz en contextos financieros por su capacidad para manejar grandes volúmenes de datos y clases desbalanceadas (Chang, Wang & Luo, 2024; Wang et al., 2025).

3. Resultados

En esta sección se presentan las métricas de desempeño obtenidas por los modelos evaluados sobre el conjunto de prueba. El modelo **XGBoost** evidenció el mejor rendimiento general en todas las métricas, superando a **Random Forest** y **Gradient Boosting**. Estos resultados son consistentes con estudios previos que destacan su capacidad para manejar datos desbalanceados y capturar relaciones no lineales complejas (Chang, Sivakulasingam & Wang, 2024; Hernes et al., 2023).

Cuadro 1: Métricas de desempeño por modelo

Métrica	**Random Forest**	**Gradient Boosting**	**XGBoost**
Accuracy	0.5949	0.6214	**0.6373**
Precision	0.5772	0.5838	**0.6031**
Recall	0.5949	0.6214	**0.6373**
F1-Score	0.5835	0.5871	**0.6019**

Las métricas reflejan una mejora consistente en la capacidad predictiva del modelo XGBoost, especialmente en cuanto a *recall* y *precision*, lo cual resulta crucial en contextos financieros donde la identificación temprana de posibles impagos puede mitigar pérdidas económicas considerables (Chang, Wang & Luo, 2024; Wang et al., 2025).

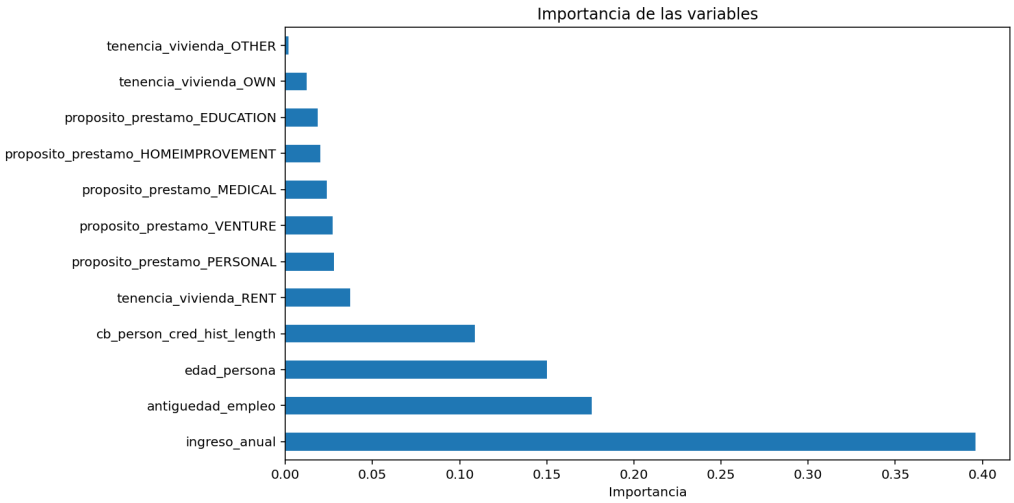

Figura 1: Importancia de variables según XGBoost.

La Figura 1 presenta la importancia relativa de las variables predictoras según el modelo XGBoost, calculada con el criterio de ganancia. Se observa que variables como el nivel de ingresos, el historial crediticio y la antigüedad laboral son las que más contribuyen a la predicción de impago (FasterCapital, 2025; Hernes et al., 2023).

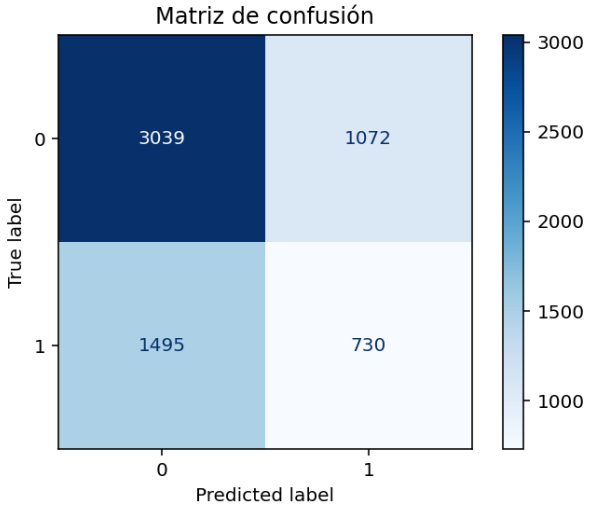

Figura 2: Matriz de confusión del modelo XGBoost.

La Figura 2 ilustra la matriz de confusión generada por el modelo XGBoost. De un total de predicciones, se observa que 3039 observaciones fueron clasificadas correctamente como no impago (verdaderos negativos), y 730 como impago (verdaderos positivos). Sin embargo, se identificaron 1495 falsos negativos clientes con probabilidad de impago mal clasificados como solventes, lo cual es especialmente relevante desde la perspectiva del riesgo financiero.

Podemos ver que los resultados obtenidos evidencian que el modelo XGBoost logró un desempeño sólido y superior respecto a los demás algoritmos evaluados. En particular, mostró un balance favorable entre precisión y sensibilidad. No obstante, el análisis de la matriz de confusión (Figura 2) permite identificar oportunidades adicionales de mejora. Aunque la cantidad de verdaderos positivos es considerable, también se observa la presencia de falsos negativos, es decir, casos de impago que no fueron detectados por el modelo.

En conjunto, los resultados reflejan un modelo con capacidad predictiva , útil como herramienta de apoyo a la toma de decisiones financieras, y con margen para seguir perfeccionándose en escenarios reales de implementación.

4. Conclusiones y recomendaciones

El presente estudio demostró que la aplicación de modelos de aprendizaje automático como *XGBoost*, *Random Forest* y *Gradient Boosting* permite construir sistemas predictivos eficaces para la evaluación del riesgo crediticio. Entre ellos, el modelo **XGBoost** mostró un rendimiento

superior.

Por otra parte los resultados confirman que variables como el nivel de ingresos, historial crediticio y antigüedad laboral tienen un peso significativo en la predicción de impago.

Referencias

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

Chang, V., Sivakulasingam, S., & Wang, H. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks*, *12*(11), 174. https://doi.org/10.3390/risks12110174

Chang, V., Wang, H., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning. *Risks*, *12*(11), 174. https://doi.org/10.3390/risks12110174

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Clark, B., & Lee, F. (2025, abril). What is Gradient Boosting? https://www.ibm.com/think/topics/gradient-boosting

FasterCapital. (2025). Credit Risk XGBoost: A Scalable Technique for Credit Risk Forecasting. https://fastercapital.com/content/Credit-Risk-XGBoost--A-Scalable-Technique-for-Credit-Risk-Forecasting.html

Fernandez Vidal, M., & Barbon, F. (2019). *Credit Scoring in Financial Inclusion* (inf. téc.). CGAP / World Bank. https://www.cgap.org/sites/default/files/publications/2019_07_Technical_Guide_CreditScore.pdf

Freire López, J. (2020). *Modelo de clasificación de riesgo crediticio utilizando Random Forest* [Tesis de pregrado]. Universidad Internacional SEK. https://repositorio.uisek.edu.ec/handle/123456789/4256

Group, W. B. (2019). *Credit Scoring Approaches Guidelines* (inf. téc.). World Bank. https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPR
pdf

Hernes, M., Adaszyński, J., & Tutak, P. (2023). Credit Risk Modeling Using Interpreted XGBoost. *European Management Studies*, *21*(3), 46-70. https://doi.org/10.7172/2956-7602.101.3

Kaggle Dataset. (2025). Credit Risk Dataset [Accedido: junio 2025].

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

MyFICO. (n.d.). ¿Qué es un puntaje de crédito? [Consultado en julio de 2025]. https://www.myfico.com/credit-education/credit-scores

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). *A Robust Machine Learning Approach for Credit Risk Analysis of Large Loan Level Datasets* (inf. téc.). Bank of Greece. https://www.bis.org/ifc/publ/ifcb49_49.pdf

Wang, W., Zuo, X., & Han, D. (2025). Predict Credit Risk with XGBoost. *Neural Computing and Applications*, *37*, 10333-10350.