Proyecto Final
Default of Credit Card Clients

Ing. Omar Manuel Zúniga Ortega[1]
Ing. Salvador Ignacio Gurdián Jarquín[2]
Universidad Americana (UAM)
Managua
Nicaragua

Introducción

La gestión eficaz del riesgo crediticio es un factor crítico para la estabilidad financiera de las instituciones bancarias y emisoras de tarjetas de crédito. En un contexto de creciente morosidad y presiones regulatorias, la capacidad de anticipar posibles incumplimientos de pago resulta fundamental para minimizar pérdidas y optimizar las estrategias de cobranza y otorgamiento de crédito.

El dataset «Default of Credit Card Clients» (ID: 350) del UCI Machine Learning Repository reúne datos reales de 30 000 clientes de tarjetas de crédito en Taiwán, incluyendo variables demográficas, límites de crédito, historial de pagos y montos facturados y pagados durante seis meses consecutivos. Este conjunto de datos constituye un referente académico ampliamente utilizado para la investigación en predicción de riesgo crediticio (Yeh & Lien, 2009).

El presente trabajo se enmarca como proyecto final de la asignatura *Machine Learning para la Analítica de Negocios* del programa de Maestría en Inteligencia de Negocios y Análisis de Datos de la Universidad Americana. El objetivo principal ha sido desarrollar un modelo predictivo supervisado capaz de anticipar el incumplimiento de pago en la próxima facturación, utilizando técnicas de machine learning y aplicando rigurosamente los procesos de exploración, selección de variables, ingeniería de características, ajuste de hiperparámetros y evaluación de modelos.

No obstante, a pesar del exhaustivo pipeline de experimentación, los resultados obtenidos en este estudio no lograron un nivel de recall o sensibilidad plenamente satisfactoria para un entorno real de producción bancaria. Esta limitación, lejos de invalidar el trabajo, aporta un análisis crítico sobre las dificultades prácticas que enfrentan los modelos de machine learning al lidiar con datasets inherentemente desbalanceados, complejidad multivariada y limitaciones en la calidad de los datos históricos. Como contribución académica, se documentan tanto los avances como las oportunidades de mejora, sentando una base sólida para futuras investigaciones que permitan superar las barreras encontradas y evolucionar hacia modelos más robustos y confiables para la predicción temprana de incumplimientos crediticios.

Datos

Para la ejecución del presente proyecto se utilizó el dataset «Default of Credit Card Clients» (ID: 350), disponible públicamente en el UCI Machine Learning Repository. La fuente original corresponde a registros reales anonimizados de clientes de tarjetas de crédito de una institución financiera en Taiwán, recolectados a partir de sistemas OLTP (Online Transaction Processing) y posteriormente preparados como dataset académico (Yeh & Lien, 2009).

La base de datos consta de 30 000 registros y 25 variables, incluyendo características demográficas, historial de pagos mensuales, montos facturados, montos pagados y la variable objetivo correspondiente al incumplimiento de pago al mes siguiente.

La estructura de los datos incluye variables numéricas y categóricas. Entre las primeras se encuentran el límite de crédito otorgado, la edad y los montos facturados/pagados en cada mes (abril a septiembre de 2005). Las variables categóricas incluyen el género (hombre/mujer), nivel educativo (posgrado, universidad, secundaria, otros) y estado civil (casado, soltero, otros). Asimismo, se incluyen seis variables ordinales que representan el estado del historial de pago de los últimos seis meses, codificadas de -2 (pago adelantado) a 8 (retraso de ocho meses).

La variable objetivo Y indica incumplimiento de pago: 1 para clientes que no cumplieron sus obligaciones y 0 para aquellos que sí lo hicieron. La distribución de clases presenta un claro desbalance: aproximadamente 22 % de incumplimientos frente a 78 % de pagos regulares.

Durante la etapa de preprocesamiento, se realizó una limpieza exhaustiva de los datos. Se verificaron y gestionaron valores nulos (inexistentes en este caso), se eliminaron registros duplicados y se redefinieron las etiquetas de las variables para facilitar su interpretación. Se aplicó codificación categórica (*label mapping*) y estandarización de variables numéricas, así como la creación de nuevas variables mediante *feature engineering*, incorporando indicadores como promedio de retraso, tendencia de morosidad, uso relativo del crédito, ratio de pago sobre facturación, escalamiento de mora y disciplina de pago en los meses previos.

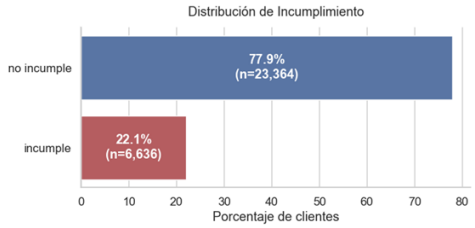

Figura 1

Distribución de incumplimiento.
Fuente: *Elaboración propia.*

Las variables con alta correlación ($\rho \geq 0{,}8$) fueron eliminadas para reducir la multicolinealidad y mejorar la estabilidad del modelo.

El conjunto de datos final, luego de la depuración y creación de variables, consta de 19 variables predictoras más la variable objetivo. Para la construcción de los modelos, el dataset se particionó mediante *hold-out*, asignando el 70 % a entrenamiento (*train set*) y el 30 % a validación (*test set*), garantizando el equilibrio de clases mediante *stratification* en la división. Esta partición permitió contar con muestras independientes para ajustar los modelos y evaluar su capacidad predictiva sin sesgos.

Metodología

El proceso metodológico seguido en este estudio respondió a un enfoque sistemático orientado a maximizar la capacidad predictiva del modelo para detectar clientes en riesgo de incumplimiento de pago.

Selección de algoritmos

Se exploraron múltiples algoritmos de clasificación supervisada ampliamente reconocidos en la literatura de machine learning y en aplicaciones financieras. Los modelos evaluados fueron:

- **CatBoost Classifier**: un algoritmo de boosting que maneja eficientemente variables categóricas y reduce el sesgo de predicción (Prokhorenkova et al., 2018).
- **LightGBM Classifier**: un algoritmo de boosting que utiliza técnicas como GOSS y EFB para mejorar la eficiencia y precisión en grandes volúmenes de datos (Ke et al., 2017).
- **Balanced Random Forest Classifier**: una variante del Random Forest que equilibra las clases mediante submuestreo de la clase mayoritaria en cada árbol (Lemaitre et al., 2017).
- **Multi-Layer Perceptron (MLP)**: una red neuronal feedforward que aprende funciones no lineales para tareas de clasificación (Pedregosa et al., 2011).

Estos modelos fueron seleccionados por su capacidad para manejar datos estructurados, su robustez frente a *outliers* y su versatilidad para trabajar con datasets desbalanceados. Se emplearon tres enfoques de ajuste de hiperparámetros:

- *RandomizedSearchCV*: búsqueda aleatoria de combinaciones de hiperparámetros.
- *BayesSearchCV*: optimización bayesiana para encontrar combinaciones óptimas.
- *GridSearchCV*: búsqueda exhaustiva en una cuadrícula de hiperparámetros.

La métrica de selección utilizada para optimizar cada modelo fue el *recall* sobre la clase positiva (incumplimiento), por ser la métrica crítica en gestión de riesgo crediticio.

Feature Engineering

Se llevó a cabo una profunda fase de ingeniería de características con el fin de enriquecer la señal predictiva del dataset original. Se desarrollaron variables derivadas agrupadas en cuatro bloques:

- **Señales de mora**: promedio de retraso mensual, número de meses en mora, banderas de escalamiento de morosidad y severidad.
- **Capacidad de crédito**: indicadores como uso promedio y máximo del límite de crédito y su tendencia temporal.
- **Disciplina de pago**: relaciones entre montos pagados y montos facturados, coeficiente de variación y tendencia de ratios de pago.
- **Demografía**: variables dummificadas correspondientes a nivel educativo, género y estado civil.

Se aplicaron técnicas de reducción de multicolinealidad (Spearman $\rho \geq 0{,}8$) eliminando variables redundantes, priorizando aquellas con mayor capacidad discriminativa (AUC univariante superior).

Arquitectura del modelo

Se diseñó un pipeline estandarizado bajo la librería `scikit-learn`, incorporando las siguientes etapas:

- **Preprocesamiento**: imputación de valores faltantes,

estandarización de variables numéricas y codificación *one-hot* para variables categóricas.

- **Balanceo de clases**: aplicación de SMOTE (*Synthetic Minority Over-sampling Technique*) para mitigar el fuerte desbalance de clases, excepto para Balanced Random Forest que lo gestiona internamente (Chawla et al., 2002).

- **Modelo base**: inclusión del clasificador correspondiente con hiperparámetros definidos a través de las técnicas de búsqueda mencionadas.

- **Calibración de probabilidad**: ajuste final mediante `CalibratedClassifierCV` para obtener probabilidades de salida más fiables (Niculescu-Mizil & Caruana, 2005).

- **Optimización de umbral**: proceso de optimización de umbral sobre la curva Precision–Recall, maximizando el *recall* sujeto a una restricción mínima de *precision* $\geq 0{,}50$.

La métrica de selección utilizada para optimizar cada modelo fue el *recall* sobre la clase positiva (incumplimiento), por ser la métrica crítica en gestión de riesgo crediticio.

Resultados

Métricas cuantitativas

Los modelos entrenados para predecir el incumplimiento crediticio fueron evaluados mediante métricas estándar: precisión, *accuracy*, *recall*, F1-score y AUC–ROC, así como mediante matrices de confusión.

Cuadro 1

Resumen de métricas cuantitativas por modelo y método

Modelo y Método	Accuracy	Precisión	Recall	F1-Score	ROC AUC
BalRF (Randomized)	0.779	0.5	0.561	0.529	0.771
LightGBM (Randomized)	0.779	0.5	0.556	0.527	0.765
LightGBM (Grid)	0.779	0.5	0.553	0.525	0.762
LightGBM (Bayes)	0.780	0.503	0.551	0.526	0.763
CatBoost (Bayes)	0.779	0.5	0.550	0.524	0.752
MLP (Bayes)	0.779	0.5	0.547	0.522	0.764
MLP (Grid)	0.779	0.5	0.546	0.522	0.766
CatBoost (Grid)	0.779	0.5	0.541	0.520	0.755
BalRF (Bayes)	0.779	0.5	0.540	0.519	0.760
BalRF (Grid)	0.779	0.5	0.535	0.517	0.760
MLP (Randomized)	0.779	0.5	0.518	0.509	0.756
CatBoost (Randomized)	0.779	0.5	0.142	0.221	0.671

Se evidencia que el modelo **Balanced Random Forest (BalRF)** ajustado mediante *RandomizedSearchCV* logró el mayor *recall* (0.561), sugiriendo un mejor desempeño para identificar correctamente a los clientes en riesgo, criterio prioritario en contextos financieros.

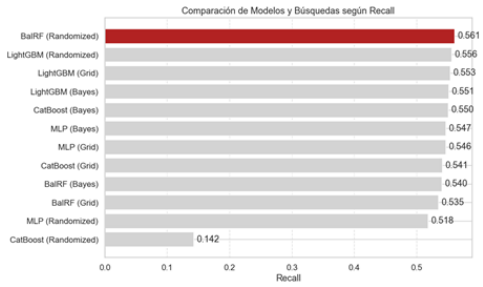

Figura 2

Comparación de Modelos y Búsquedas según Recall.
Fuente: *Elaboración propia.*

Análisis Cualitativo

La evaluación cualitativa mostró que:

- **Balanced Random Forest (BalRF)** presentó la mejor capacidad para captar verdaderos positivos, siendo especialmente eficiente para gestionar clases desbalanceadas.

- **LightGBM** y **CatBoost** exhibieron buena precisión pero baja sensibilidad al emplear umbrales predeterminados (0.5). El ajuste fino del umbral mostró una mejora notable en *recall*, manteniendo una precisión aceptable (0.50).

- El modelo **MLP** mostró dificultades de convergencia, limitando su potencial para una precisión elevada.

Ejemplos de Clasificación y Análisis de Errores

Mediante la matriz de confusión, el modelo BalRF (Randomized) identificó correctamente 1 117 de 1 991 casos positivos. Sin embargo, mostró una tasa considerable de falsos positivos (1 117 casos). Esto implica que aunque el modelo es capaz de detectar correctamente muchos casos de incumplimiento, también genera un número importante de falsas alarmas, lo que podría resultar en costos operativos adicionales para las entidades financieras.

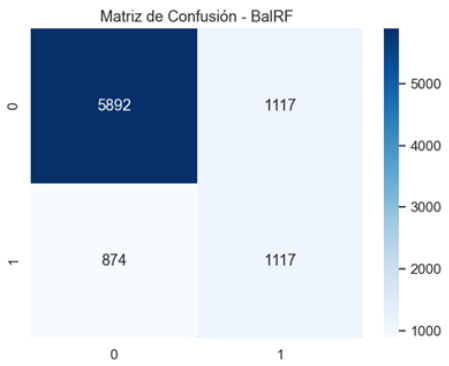

Figura 3

*Matriz de confusión del modelo **Balanced Random Forest (BalRF)** ajustado mediante RandomizedSearchCV.*
***Fuente:** Elaboración propia.*

Interpretación

Los resultados revelan la complejidad inherente a la predicción del incumplimiento crediticio. Aunque ninguno de los modelos alcanza un desempeño perfecto, el **Balanced Random Forest** demostró ser el más apto para un escenario bancario donde minimizar falsos negativos (clientes que incumplen y no son detectados) es crítico. El ajuste del umbral de decisión según la curva Precision–Recall fue clave para optimizar los resultados.

Limitaciones

- El desbalance significativo en el conjunto de datos afectó negativamente la capacidad predictiva de algunos modelos.
- Calidad y cantidad limitadas de características disponibles restringieron la potencialidad de capturar patrones predictivos más robustos.
- La calibración de probabilidades, aunque mejoró los resultados, introdujo complejidad adicional en el pipeline de evaluación.

Comparaciones entre modelos generados

Al comparar los modelos mediante diferentes estrategias de optimización hiperparamétrica, se observó que:

- **RandomizedSearchCV** proporcionó los mejores resultados globales, especialmente en términos de recall.
- **BayesSearchCV** ofreció una búsqueda eficiente, pero resultados ligeramente inferiores en recall respecto a RandomizedSearchCV.
- **GridSearchCV** mostró desempeño competitivo, pero con recall ligeramente inferior, sugiriendo una menor capacidad para explorar eficientemente el espacio de parámetros en comparación con los otros métodos.

En definitiva, el **Balanced Random Forest** optimizado mediante **RandomizedSearchCV** demostró ser la mejor combinación, alcanzando un recall máximo de 0.561, constituyendo un resultado óptimo en términos de gestión de riesgo crediticio.

Conclusiones

Evaluación crítica de los resultados

A pesar de implementar un pipeline exhaustivo que incluyó preprocesamiento, ingeniería de características, balanceo de clases, calibración de probabilidades y optimización de hiperparámetros, los modelos desarrollados no alcanzaron un nivel de *recall* o sensibilidad plenamente satisfactoria para su aplicación en un entorno real de producción bancaria. El modelo que mejor desempeño mostró, el **Balanced Random Forest** optimizado mediante *RandomizedSearchCV*, logró un *recall* del 56.1 %, lo que indica que aún se presentan desafíos significativos en la detección precisa de clientes en riesgo de incumplimiento.

Esta limitación pone de manifiesto las dificultades inherentes al uso de modelos de machine learning en contextos con datasets desbalanceados, complejidad multivariada y restricciones en la calidad y cantidad de los datos históricos disponibles. Además, se observó que técnicas como SMOTE, aunque útiles, pueden introducir problemas de calibración en los modelos, afectando la confiabilidad de las probabilidades predichas.

Contribución académica

A pesar de los resultados subóptimos, este estudio aporta una documentación detallada del proceso de modelado y de las dificultades encontradas, ofreciendo una base sólida para futuras investigaciones. Se destacan las siguientes contribuciones:

- Identificación de las limitaciones de los modelos tradicionales de machine learning en la predicción de incumplimiento crediticio con datos desbalanceados.
- Evaluación de diversas técnicas de balanceo de clases y su impacto en la calibración y desempeño de los modelos.
- Análisis de la necesidad de incorporar fuentes de datos adicionales y técnicas avanzadas para mejorar la capacidad predictiva.

Recomendaciones para futuros trabajos

Para superar las limitaciones identificadas y avanzar hacia modelos más robustos y confiables, se proponen las siguientes líneas de acción:

- **Enriquecimiento de datos:** Incorporar fuentes de datos adicionales (transaccionales en tiempo real, macroeconómicos, redes sociales) que aporten variables predictivas no presentes en el dataset original.
- **Técnicas avanzadas de balanceo:** Explorar métodos como ADASYN o submuestreo informados para mejorar la representación de la clase minoritaria sin comprometer la calibración.
- **Modelos de ensamble y deep learning:** Investigar enfoques de stacking, blending y arquitecturas de deep learning que puedan capturar relaciones no lineales y patrones complejos.
- **Optimización de umbrales y costos:** Implementar aprendizaje sensible al costo y ajuste dinámico de umbrales para minimizar falsos negativos en la gestión del riesgo crediticio.
- **Evaluación continua y validación cruzada:** Adoptar validaciones cruzadas más robustas y evaluaciones continuas para adaptar los modelos a cambios en los patrones de los datos.

En resumen, aunque los resultados obtenidos en este estudio no alcanzaron los niveles deseados de sensibilidad, proporcionan una comprensión valiosa de los desafíos actuales y delinean un camino claro para futuras investigaciones que busquen mejorar la predicción de incumplimiento crediticio mediante técnicas de machine learning.

Referencias

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[2] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.

[3] Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5. https://jmlr.org/papers/volume18/16-365/16-365.pdf

[4] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625–632. https://doi.org/10.1145/1102351.1102430

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[6] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020