Detección de Fraude en Transacciones con Tarjeta de Crédito

Aplicación de técnicas de aprendizaje de máquina en un problema de clasificación binaria con datos desbalanceados

Sara Palacios Duque, Manuela Granda Muñoz, Sebastián Restrepo Betancur, Simón Cuartas Rendón

UNIVERSIDAD NACIONAL DE COLOMBIA

Introducción

El fraude con tarjetas de crédito es un fenómeno creciente en sistemas financieros digitales, donde uno de los mayores retos es que las clases referentes al fraude están en desventaja frente a los datos de transacciones legítimas. Por lo cual, se plantea la implementación y comparación de diferentes modelos, integrando técnicas de reducción de dimensionalidad y estrategias de balanceo de clases para mejorar la capacidad predictiva frente a eventos poco frecuentes.

El conjunto de datos contiene 284,807 transacciones, de las cuales 492 corresponden a fraudes. El total de 30 variables incluye transformaciones aplicadas mediante análisis de componentes principales (PCA), implementadas para preservar la confidencialidad de la información sensible de los clientes.

Método

Esta problemática se abordó como un problema de clasificación binaria con un desbalance marcado entre clases. Para abordar este desafío, se desarrolló una metodología estructurada en tres etapas: preparación de datos, selección de variables y entrenamiento/evaluación de modelos bajo validación cruzada y optimización de hiperparámetros.

Los datos se particionaron en conjuntos de entrenamiento y validación, con muestreo estratificado para mantener la proporción de clases. Posteriormente, se aplicó SMOTE sobre el conjunto de entrenamiento para corregir el desbalance de clases, generando observaciones sintéticas de la clase minoritaria. Posterior, se estandarizaron los datos y se aplicó PCA para reducir la dimensionalidad reteniendo el 95 % de la varianza explicada.

Por otra parte, con el fin de garantizar reproducibilidad y escalabilidad, se construyeron pipelines, integrando selección de variables y ajuste del modelo utilizando validación cruzada estratificada de 3 pliegues y métrica de evaluación F1-score.

Se compararon siete clasificadores, tanto lineales como no lineales, paramétricos y de ensamblado, cada modelo fue entrenado sobre el conjunto balanceado y validado sobre el conjunto original para evaluar en condiciones reales. Las configuraciones se seleccionaron mediante búsqueda aleatoria de hiperparámetros.

Modelo	Tipo	Configuraciones
Regresión logística	Lineal	Penalización L2, selección entre liblinear y lbfgs, C ∈ [0.001–100]
Árbol de decisión	No lineal	Ajuste de max_depth y min_samples_split
Random Forest	Ensamblado	Combinación de árboles, ajuste de n_estimators y profundidad
HistGradientBoosting	Ensamblado	Boosting basado en histogramas, learning_rate, max_iter, l2
LightGBM	Ensamblado	Optimizado para rendimiento, max_depth, n_estimators, lr
XGBoost	Ensamblado	Árboles optimizados con regularización, ajuste de lr, depth
MLPClassifier	Red neuronal	Red multicapa, funciones de activación ReLU, ajuste de alpha y tamaño de capas ocultas

Tabla 1. Configuración de los modelos

Resultados

Los modelos evaluados demostraron un buen desempeño en el conjunto de entrenamiento, con métricas cercanas a la unidad en AUC, exactitud, coeficiente kappa de Cohen y F1-score.

Este comportamiento indica un ajuste óptimo sobre datos balanceados con los datos sintéticos. En el conjunto de validación, el desempeño general fue bueno en términos de AUC y exactitud. Sin embargo, las métricas más sensibles al desbalance, como el F1-score y el coeficiente kappa, mostraron caídas importantes. Esto evidencia que, aunque los modelos identifican correctamente la mayoría de las transacciones legítimas, su capacidad para detectar fraudes reales presenta limitaciones.

Por otra parte, el modelo que mostró mayor estabilidad y rendimiento global fue el de Bosques Aleatorios, con AUC y exactitud de 0.94 y 1.00, respectivamente, y un F1-score de 0.49 en validación, el más alto entre todos los modelos evaluados.

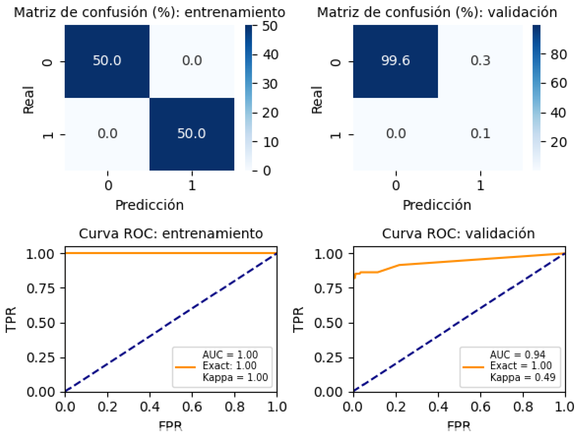

Figura 1. Desempeño del modelo Bosque Aleatorio

Otros modelos como LightGBM, XGBoost, MLP y HistGradientBoosting también alcanzaron valores altos de AUC, superiores a 0.90, pero mostraron F1-scores por debajo de 0.20, lo cual indica una mayor dificultad para generalizar en cuando hay clases minoritarias.

La regresión logística y el árbol de decisión, presentaron el menor F1 y Kappa en validación, lo que limita su aplicabilidad práctica en este tipo de problemas.

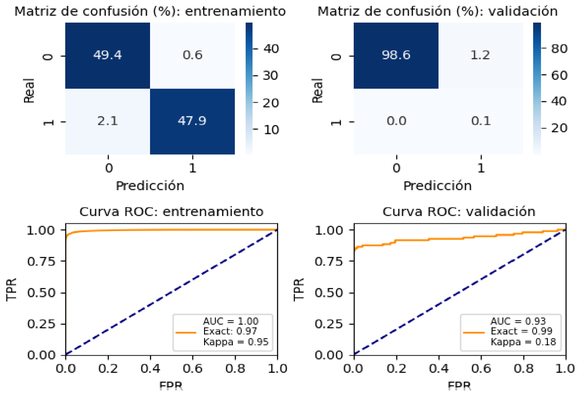

Figura 2. Desempeño del modelo LightGBM

Conclusión

Si bien todos los modelos lograron buenas métricas en entrenamiento, solo los algoritmos de ensamble, mostraron una mejor capacidad de generalización en validación y estabilidad frente a métricas sensibles al desbalance. Estos resultados indican que, en este tipo de problemas, la elección del modelo debe priorizar el balance entre clases y que las técnicas de preprocesamiento y evaluación son fundamentales para desarrollar soluciones robustas.

Referencias

1. Kaggle: Credit Card Fraud Detection Dataset
2. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.