MODELO DE CLASIFICACIÓN DE RIESGO CREDITICIO

Autor: Sebastian Cruz Castro
junio 2025

OBJETIVO DEL PROYECTO

El objetivo es construir un modelo de machine learning que clasifique a los clientes como buenos o riesgosos con base en información interna y externa, maximizando la métrica AUC (Area Under Curve). Esta clasificación servirá para optimizar la toma de decisiones crediticias.

DESCRIPCIÓN DEL DATASET

- Dataset principal: 14,454 registros con variables relacionadas a la solicitud y otorgamiento del crédito.
- Dataset externo: historial crediticio con variables como pagos atrasados, balances y cuentas activas.
- Variable objetivo: target (1 = cliente riesgoso, 0 = cliente confiable).

ANÁLISIS EXPLORATORIO Y LIMPIEZA DE DATOS

- Se identificó un **desbalance** en la variable objetivo: **81%** buenos, **19%** riesgosos.
- Se revisaron valores nulos y tipos de datos para decidir estrategias de limpieza.

El desbalance impacta la elección del modelo y la métrica. La limpieza asegura la calidad de entrada para el algoritmo.

- Imputación: variables numéricas con 0, fechas conservadas o transformadas.
- Conversión de columnas datetime a variables numéricas: días desde evento.
- Se descartaron columnas con tipos no compatibles con LightGBM (como datetime64).

SELECCIÓN DEL MODELO

Se eligió **LightGBM** por:

- Ser rápido y eficiente con grandes volúmenes.
- Tener soporte para desbalanceo (class_weight='balanced').
- Ser altamente interpretable a través de importancia de variables.

Otros modelos como Random Forest o XGBoost mostraron rendimiento similar o inferior. LightGBM fue el mejor balance entre rendimiento y velocidad.

RESULTADOS FINALES

- AUC alcanzado: 0.6043 (mejor modelo).
- Recall > 0.99 para umbrales bajos.
- Matriz de confusión, curva ROC, y reporte de clasificación disponibles como evidencia.

Se priorizó una alta tasa de recall para detectar clientes riesgosos a costa de un menor precision. Esto es coherente con el objetivo del reto.

CONCLUSIONES

- El modelo LightGBM optimizado cumple con los requisitos del challenge, ya que logra una separación razonable entre clientes buenos y riesgosos (AUC > 0.60) en un contexto altamente desbalanceado.
- Se identificó una inconsistencia estructural en los datos: muchas observaciones no tienen historial crediticio, lo que limita el poder predictivo de ciertas variables.
- A pesar de esta limitación, el modelo logró detectar patrones relevantes gracias al feature engineering, permitiendo extraer valor incluso de clientes sin historial.