Universidad del Valle de Guatemala

SECURITY DATA SCIENCE

Sección 10

Convenio PLUS TI – Universidad del Valle 2025

“Métricas custom para reducción de falsos positivos en clasificación binaria - fraude”

Fabián Juárez - 21440

Guatemala, 02 de Junio del 2025

Resumen

Este proyecto aborda el desafío de mejorar la detección de fraudes en un entorno altamente desbalanceado, con especial énfasis en comercios con baja frecuencia de transacciones. El modelo inicial presentaba una alta tasa de falsos positivos, lo que comprometía su utilidad práctica. Para resolverlo, se diseñaron métricas de evaluación personalizadas para LightGBM, orientadas a maximizar la detección de fraude sin sacrificar precisión.

Se exploraron distintas estrategias enfocadas en priorizar los casos de baja frecuencia y se identificó la más efectiva tras evaluar el rendimiento sobre un conjunto de pruebas. Los datos utilizados reflejan un fuerte desbalance de clases. La métrica personalizada low_volume_recall logró una precisión del 91% y un recall del 67%, con solo 1.1 falsos positivos por cada fraude detectado.

Metodología

La metodología implementada siguió las siguientes etapas:

Se usó Python con pandas y LightGBM. El dataset fue dividido en entrenamiento (80%) y prueba (20%, correspondiente a diciembre 2020). Los comercios de baja frecuencia se definieron como aquellos en el percentil 16.5 en el dataset de transacciones mensuales (este valor puede extraerse del notebook si se aplicó un umbral).

1. **Análisis exploratorio de datos (EDA):** inspección de la distribución de clases, comportamiento por comercio, y frecuencia de transacciones.

2. **Ingeniería de variables:** se generaron nuevas variables agregadas por comercio y cliente, como conteos, promedios de montos y ratios temporales.

3. **Modelo base:** se entrenó un modelo Light GBM inicial con métricas estándar (AUC, F1-score).

4. **Diseño de métricas personalizadas:** se implementaron funciones feval en LightGBM, enfocadas en:

    - Minimizar la razón de falsos positivos (fp_ratio)

    - Optimizar el recall en comercios con baja frecuencia (low_volume_recall)

    - Maximizar el F1-score en esos comercios (low_volume_f1)

5. **Evaluación en test:** se utilizó el mes de diciembre de 2020 como conjunto de evaluación, comparando precisión, recall, F1 y razón de FP.

Descripción de la implementación práctica

Las métricas personalizadas se integraron mediante la interfaz feval de LightGBM. Se empleó validación cruzada estratificada para garantizar la representación proporcional de fraudes en cada fold. Se implementaron tres funciones de evaluación personalizadas:

- **fp_ratio:** penaliza falsos positivos directamente.
- **low_volume_recall:** favorece el recall en comercios con baja actividad.
- **low_volume_f1:** balancea precisión y recall en ese mismo segmento.

Para el test final, se calculó una matriz de confusión y se compararon los scores para las métricas principales.

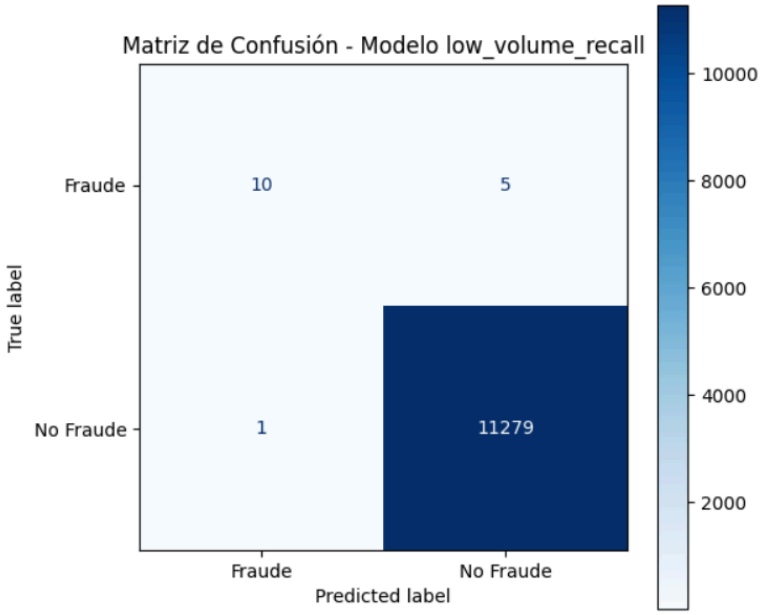
Análisis de los resultados

La siguiente tabla resume el rendimiento final sobre el conjunto de test:

Métrica personalizada	Precisión	Recall	F1-score	FP Ratio
`low_volume_recall`	0.91	0.67	0.77	1.1
`low_volume_f1`	0.83	0.67	0.74	1.2
`fp_ratio`	0.00	0.00	0.00	0.0

- **low_volume_recall** fue la estrategia más efectiva: logró detectar el 67% de los fraudes reales con una precisión del 91% y solo 1.1 falsos positivos por cada fraude detectado.
- La **matriz de confusión** confirma este buen balance, mostrando 10 fraudes correctamente detectados, 5 no detectados, y solo 1 falso positivo (ver Figura 1).
- La función fp_ratio, si bien logró eliminar falsos positivos, lo hizo a costa de no detectar ningún fraude, lo que resulta inaceptable en un sistema práctico.

Figura 1 - Matriz de Confusión (low_volume_recall)



Conclusiones

- Se logró el objetivo de optimizar la detección de fraude en comercios con baja frecuencia, mediante el diseño de métricas personalizadas.

- La métrica low_volume_recall fue la más efectiva, alcanzando un buen trade-off entre recall y precisión, y resultando útil en entornos donde es clave no perder fraudes.

- El experimento confirma que adaptar la función de evaluación al dominio específico mejora significativamente el rendimiento del modelo, en comparación con métricas genéricas.

- Como trabajo futuro, se recomienda explorar técnicas complementarias como el balanceo de clases (por ejemplo, SMOTE o submuestreo) y el uso de modelos alternativos como XGBoost o redes neuronales, así como la integración de la métrica personalizada en un pipeline de producción para monitoreo en tiempo real.

- Este enfoque demuestra el valor de adaptar las métricas de evaluación al dominio, especialmente en sistemas sensibles como la detección de fraude, donde los falsos positivos tienen costos operativos reales.