Fraud Detection — Memo Técnico

BAF NeurIPS 2022 · XGBoost + Optuna + SHAP

Eduardo Rdgz-Á

2026-05-28

Contexto

Este documento registra las decisiones de diseño del modelo de detección de fraude entrenado sobre el dataset BAF (NeurIPS 2022), compuesto por 1 millón de transacciones bancarias con 30 features cuyos valores fueron transformados para preservar privacidad. La clase positiva (fraude) representa aproximadamente 1.1% de las observaciones, lo que determina las decisiones centrales del proyecto: métrica de optimización, tratamiento del desbalance y criterio de threshold.

Decisiones de diseño

Tratamiento del desbalance de clases

El dataset presenta una proporción aproximada de 98.9:1.1 entre transacciones legítimas y fraudulentas — aproximadamente 90 casos legítimos por cada fraude. Las alternativas evaluadas fueron SMOTE (oversampling sintético), undersampling aleatorio y ajuste de pesos mediante `scale_pos_weight`.

Se eligió `scale_pos_weight = 90` por tres razones: no modifica la distribución real de los datos, es nativo en XGBoost sin pasos adicionales en el pipeline, y preserva toda la información de la clase mayoritaria.

Tratamiento de ausencias

El dataset no presenta valores nulos. Sin embargo, varias columnas contienen valores `-1` que, según el paper, son la convención del dataset para indicar dato no disponible — no valores faltantes en el sentido estadístico. La columna más afectada es `prev_address_months_count` con 71.3% de registros en esa condición.

No se aplicó imputación. XGBoost interpreta estos valores nativamente como una categoría implícita de "ausente", y en contexto de fraude la ausencia de un dato puede ser señal por sí misma.

Encoding de variables categóricas

El dataset contiene variables categóricas nominales. La alternativa evaluada fue `OneHotEncoder`, que genera una columna binaria por categoría.

Se eligió `OrdinalEncoder` con `unknown_value=-1` para categorías no vistas en producción. Para modelos basados en árboles como XGBoost el encoding ordinal es suficiente — los splits no asumen orden entre categorías, y evita la expansión dimensional que genera OneHot.

Métrica de optimización

La métrica natural en clasificación binaria es AUC-ROC. Sin embargo, con una clase positiva del 1%, un clasificador que predice siempre "legítima" obtiene AUC-ROC 0.50 mientras que AUC-PR 0.01 — el baseline aleatorio. AUC-PR no involucra los verdaderos negativos en ninguno de sus ejes, por lo que el desbalance no distorsiona el score.

Se eligió **AUC-PR** como métrica de optimización en Optuna y como referencia principal de evaluación. AUC-ROC se reporta como métrica secundaria para comparabilidad con la literatura.

Búsqueda de hiperparámetros — Optuna

Se realizó una búsqueda bayesiana con TPE (Tree-structured Parzen Estimator) sobre 50 trials, maximizando AUC-PR en validación cruzada estratificada (5 folds). El espacio de búsqueda:

Hiperparámetro	Rango
`n_estimators`	[200, 800]
`max_depth`	[3, 8]
`learning_rate`	[0.001, 0.3] log
`subsample`	[0.5, 1.0]
`colsample_bytree`	[0.5, 1.0]
`reg_alpha`	[0.0001, 10.0] log
`reg_lambda`	[0.0001, 10.0] log

Mejor trial: `n_estimators=581`, `max_depth=3`, `learning_rate=0.065`, `subsample=0.729`, `colsample_bytree=0.543`, `reg_alpha=0.002`, `reg_lambda=3.022`. Mejor AUC-PR en CV: **0.1778**. La ganancia sobre el baseline se reporta en la sección de Resultados, evaluada sobre el conjunto de prueba.

Selección de threshold — modelo final (post Optuna)

El threshold por defecto (0.5) no es óptimo bajo desbalance severo. El análisis se realiza sobre las probabilidades del modelo final (XGBoost + Optuna, 50 trials) — no del baseline.

El criterio de selección depende del costo relativo entre FN y FP: un fraude no detectado genera pérdida directa (Error Tipo II — FN); una falsa alarma genera fricción operativa (Error Tipo I — FP). El threshold es una decisión de negocio, no del modelo.

Threshold	Precision	Recall	F1	FN	FP
0.50	0.0509	0.8005	0.0957	440	32,921
0.40	0.0408	0.8622	0.0779	304	44,724
0.30	0.0321	0.9089	0.0620	201	60,435
0.20	0.0244	0.9438	0.0476	124	83,107
0.10	0.0175	0.9787	0.0344	47	121,169
0.05	0.0141	0.9964	0.0277	8	154,024

Resultados

Métricas evaluadas sobre el conjunto de prueba (200,000 observaciones, threshold = 0.5).

Métrica	Baseline	Final (Optuna)	Δ
AUC-ROC	0.8444	0.8991	+6.5%
AUC-PR	0.1238	0.1800	+45.4%

AUC-PR es la métrica de optimización — la ganancia relevante es de +45.4%. AUC-ROC se reporta como referencia secundaria.

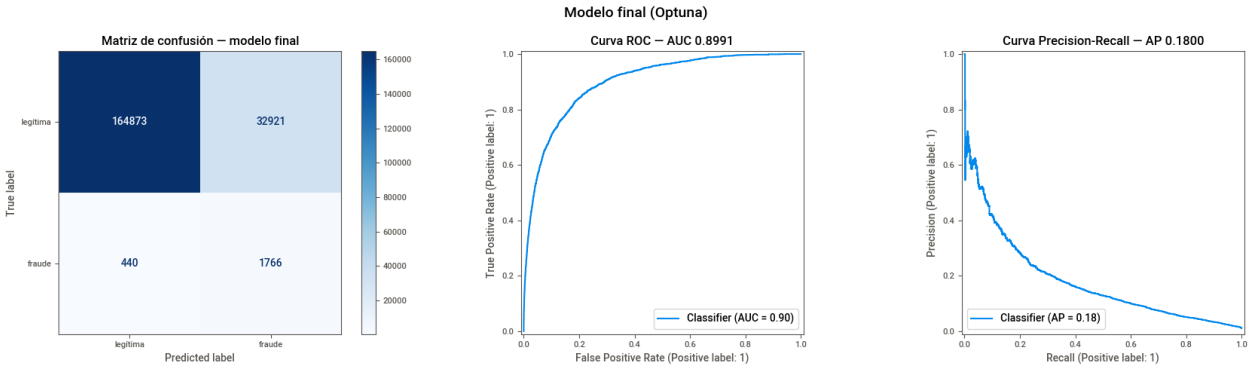

Figure 1: Curvas de evaluación — modelo final (Optuna)

Interpretabilidad — Importancia global (SHAP)

Se usó SHAP (SHapley Additive exPlanations) con `TreeExplainer` sobre una muestra aleatoria de 2,000 observaciones del test set.

Cada punto es una observación. El color indica el valor de la feature: rojo = valor alto, azul = valor bajo. La posición horizontal indica la dirección del impacto sobre el score de fraude.

Las features con señal direccional clara son `housing_status` y `device_os` (SHAP medio = 0.51 cada una): valores bajos de `housing_status` empujan hacia fraude; valores altos de `device_os` también. `phone_home_valid` (+0.41), `keep_alive_session` (+0.35) y `has_other_cards` (+0.34) siguen el mismo patrón.

`name_email_similarity` y `current_address_months_count` presentan señal mixta: sus distribuciones se extienden en ambas direcciones, lo que indica que el impacto depende del valor específico de la observación. La suma acumulada de las 21 features restantes contribuye con SHAP medio = 2.54 — mayor que cualquier feature individual, lo que sugiere que el modelo distribuye información predictiva de forma difusa sobre el espacio de features.

Limitaciones y trabajo futuro

Datos sintéticos. BAF NeurIPS 2022 es un dataset sintético generado a partir de datos reales. El modelo puede no capturar patrones de fraude que emergen en producción real — deriva de distribución, fraude coordinado, y nuevos vectores de ataque no representados en el dataset.

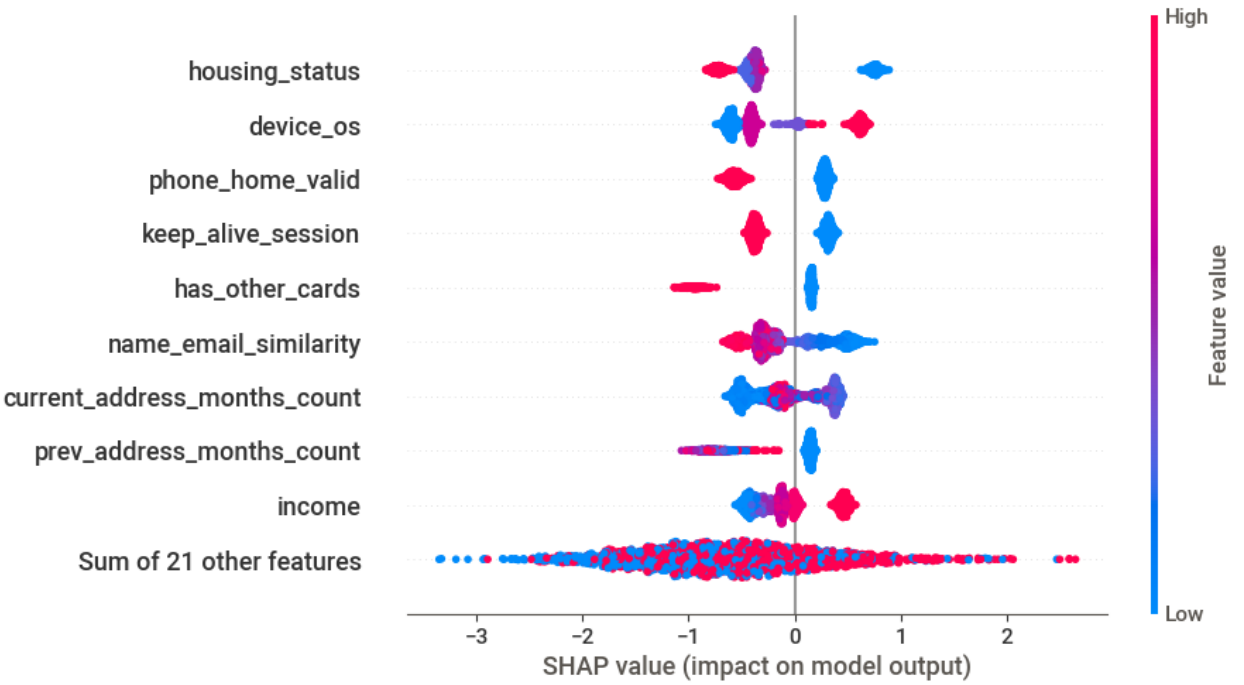

Figure 2: SHAP beeswarm — importancia global de features

Features con valores transformados. Los valores de las 30 features están transformados para preservar privacidad. Esto limita la interpretabilidad operacional: es posible identificar qué features importan estadísticamente, pero no traducirlas directamente a reglas de negocio accionables sin el mapeo original.

Threshold estático. El threshold guardado (0.5) es una referencia neutral. En producción, el threshold óptimo depende del costo relativo FN/FP del negocio — un parámetro que varía por institución y contexto regulatorio.

Trabajo futuro. Reentrenamiento sobre datos reales con drift monitoring, optimización de threshold bajo función de costo explícita, y exploración de modelos secuenciales (RNN/Transformer) para capturar patrones temporales entre solicitudes del mismo dispositivo.