

# **Uso de modelos de clasificación para la detección de fraude de tarjetas bancarias**

**Fernando Poblete**

## **RESUMEN**

Este proyecto desarrolla modelos de Machine Learning para detectar transacciones fraudulentas con tarjetas bancarias, utilizando datos sintéticos generados mediante simulación MABS (Lopez-Rojas & Axelsson, 2014), basada en datos reales agregados de España durante los años 2012-2013.

El proyecto está dividido en dos secciones. La primera sección contiene el análisis de las variables incluidas en la base de datos. Se analiza la densidad de las variables numéricas y el conteo de las variables categóricas, para transacciones fraudulentas y legítimas.

Se discute el razonamiento para eliminar registros de la base de datos, que no afecten al entrenamiento de los modelos de clasificación.

La segunda sección contiene la creación y comparación de modelos de clasificación: CatBoost, MLPClassifier y SVC.

Se explica la división estratificada de datos (importante para modelos de clasificación con clases desbalanceadas), el preprocesamiento correspondiente, la creación de datos sintéticos de la clase minoritaria (en nuestro caso, las transacciones fraudulentas) con SMOTE (Synthetic Minority Oversampling Technique) y finalmente la evaluación de los modelos con la matriz de confusión y la curva PR (precisión-recall).

De la comparación de los modelos, se obtiene como sugerencia práctica, se recomienda el uso de CatBoost para problemas de clasificación con características similares, de grupos desbalanceados, y variables mayoritariamente categóricas. Considerando su facilidad de preprocesamiento, ajustes del modelo y tiempo para obtener resultados.

Finalmente, se discute la contrapartida (trade-off) que existe entre eliminar ambos tipos de errores de clasificación.

## **INTRODUCCIÓN**

Este proyecto desarrolla modelos de Machine Learning para detectar transacciones fraudulentas con tarjetas bancarias, utilizando datos sintéticos generados mediante simulación MABS (Lopez-Rojas & Axelsson, 2014), basados en datos reales agregados de España (2012-2013).

### **Problema de investigación:**

Los datasets de fraude presentan un desbalance extremo (solo 1.21% de transacciones fraudulentas), lo que dificulta la identificación precisa de patrones anómalos. Además, existen sesgos en variables clave como categorías de comercios.

### **Objetivos:**

1. Evaluar el rendimiento comparativo de tres algoritmos (CatBoost, MLP y SVC)
2. Identificar variables predictivas críticas (monto, categoría, edad)
3. Optimizar el *recall* para minimizar falsos negativos (fraudes no detectados)

## **METODOLOGÍA**

### **Datos y preprocesamiento:**

- Dataset sintético (594,643 transacciones, 10 variables incluyendo monto, categoría y fraude binario).
- Preprocesamiento:
  - Eliminación de zipcode (redundante).
  - Reducción de sobremuestreo en es\_transportation.
  - Creación de variable temporal step\_mod7.

### **Herramientas y técnicas:**

- **Algoritmos comparados:** CatBoost, MLP (Multilayer Perceptron), SVC (Support Vector Classifier).
- **Manejo de desbalance:** SMOTE al 30% para transacciones fraudulentas.
- **Validación:** División estratificada 80/20, evaluación con matriz de confusión y curvas Precision-Recall.
- Entorno Python (pandas, scikit-learn, CatBoost).

## **RESULTADOS**

### **Análisis exploratorio (EDA):**

- **Monto:** Transacciones fraudulentas mostraron mayor dispersión (media: \$530 vs. \$31) y distribución de cola larga.
- **Categorías:**
  - es\_leisure: Mayor tasa de fraude (94.99%).
  - es\_travel: Segunda mayor tasa (79.40%).
  - es\_sportsandtoys: Tercera tasa (49.53%) pero mayor volumen absoluto (27.53% del total).
- **Edad:** Rango 26-35 años concentró 32.56% de fraudes.
- **Género:** Mujeres presentaron tasa ligeramente mayor (1.47% vs. 0.91%).

### **Rendimiento de modelos:**

<b>Modelo</b>	<b>Recall (Fraude)</b>	<b>AUC-PR</b>	<b>Precisión (Fraude)</b>
<b>CatBoost</b>	92.43%	0.9032	60.42%
<b>SVC</b>	88.47%	0.8681	66.49%
<b>MLP</b>	85.49%	0.8915	72.54%

### Detalles adicionales:

- CatBoost detectó 1,331 fraudes verdaderos (solo 109 falsos negativos).
- Curvas Precision-Recall confirmaron superioridad de CatBoost (AUC-PR: 0.9032).

## DISCUSIÓN

CatBoost demostró mejor rendimiento debido a:

1. Manejo nativo de variables categóricas (sin preprocesamiento complejo).
2. Optimización para datos desbalanceados (peso 2:1 para fraudes).
3. Velocidad de entrenamiento.

### Implicaciones clave:

- Alto *recall* es crítico: **Un falso negativo cuesta 10x más que un falso positivo** en contextos bancarios.
- Variables como categoría y monto fueron más predictivas que género o edad.

### Limitaciones:

Datos sintéticos pueden subestimar complejidad de fraudes colaborativos.

## CONCLUSIÓN

### Hallazgos principales:

1. **CatBoost es el modelo óptimo** por su alto recall (92.43%) y eficiencia con variables categóricas.
2. **Categorías sensibles:** *es\_sportsandtoys* y *es\_travel* requieren monitoreo prioritario.
3. **Técnicas esenciales:** SMOTE y pesos de clases son indispensables para desbalance extremo.

### Recomendaciones prácticas:

- Implementar CatBoost con umbrales ajustables según riesgo operativo.
- Extender metodología a fraudes en comercio electrónico.

**Impacto:** Marco reproducible para reducir pérdidas financieras y mejorar seguridad bancaria.

## REFERENCIA

Lopez-Rojas, Edgar Alonso & Axelsson, Stefan. (2014). BankSim: A Bank Payment Simulation for Fraud Detection Research. 26th European Modeling and Simulation Symposium, EMSS 2014