

Prueba de ML Engineer: Detección de Fraude en Transacciones Financieras

Contexto:

Las actividades fraudulentas pueden costar a las empresas un porcentaje significativo de sus ventas anuales, lo que representa un problema económico considerable. Es crucial para las compañías de tarjetas de crédito comprender las características de las transacciones fraudulentas y desarrollar modelos predictivos para identificar y prevenir actividades riesgosas. El control de pagos online es uno de los principales casos de uso para lo que nuestros clientes usan Simetrik, cargando millones de transacciones al día a nuestra aplicación.

Objetivo de la Prueba:

Analizar un dataset de transacciones financieras que ha sido etiquetado para indicar si cada transacción es fraudulenta o no. El objetivo es aplicar técnicas de ML para clasificar adecuadamente si una transacción es o no fraudulenta. En el .zip adjunto se encuentra el dataset de entrenamiento y test, así como la definición de cada columna.

Instrucciones:

Análisis Exploratorio de Datos:

- Realizar un análisis detallado, utilizando gráficos para explorar las relaciones entre características y la etiqueta de fraude.
- Identificar las features más relevantes, realice análisis de correlación y considere transformaciones de datos apropiadas.

Manejo del Desbalance de Datos:

- Note que hay más transacciones no fraudulentas que fraudulentas.
- Evaluar y justificar métodos para balancear el dataset.

Modelado:

- Implementar y comparar al menos cuatro modelos de ML para clasificar las transacciones.
- Discutir la selección de modelos, hiperparámetros y métricas de evaluación.

Entrega:

GitHub que contenga:

1. Un Jupyter notebook que documente claramente cada paso mencionado arriba, desde el análisis hasta la evaluación de modelos.
2. La implementación de una **aplicación sencilla usando FastAPI** que sirva como punto de entrada para realizar inferencias sobre el mejor modelo encontrado.

Nota: Este ejercicio es clave para entender su *enfoque* en el manejo de problemas reales de ML y su habilidad para *comunicar* los resultados obtenidos.