Plus Ti

Convenio PLUS TI – Universidad del Valle 2025

Trabajo teórico - práctico 1:

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

Objetivo

El objetivo de este trabajo es explorar y evaluar distintas estrategias para mejorar la detección de fraudes en un conjunto de datos proporcionado. Actualmente, el modelo presenta una alta tasa de falsos positivos por cada fraude identificado. Para abordar este problema, se propone probar diversas métricas personalizadas en la función de evaluación (feval) de LightGBM, con el fin de identificar cuál logra reducir los falsos positivos sin afectar negativamente la capacidad de detección. Adicionalmente, se busca explorar funciones de evaluación que permitan optimizar aspectos específicos del modelo, como la precisión, la cobertura o la priorización de alertas relevantes.

1. Diseñar y evaluar funciones de evaluación personalizadas enfocadas en reducir la cantidad de falsos positivos por cada fraude detectado. Este será el punto de partida común para todos los estudiantes.

2. Asignar a cada alumno un objetivo específico de optimización, el cual servirá como enfoque principal para desarrollar y ajustar sus modelos.

3. Cada alumno deberá proponer al menos tres funciones de evaluación distintas alineadas con el objetivo asignado, implementarlas y compararlas entre sí para analizar cuál se desempeña mejor.

4. Seleccionar la función de evaluación más efectiva entre las propuestas, justificando la elección con base en los resultados obtenidos.

Plus Ti

Descripción del Dataset

Este es un dataset simulado de transacciones con tarjeta de crédito que contiene transacciones legítimas y fraudulentas desde el 1 de enero de 2019 hasta el 31 de diciembre de 2020. Cubre tarjetas de crédito de 1000 clientes que realizan transacciones con un conjunto de 800 comercios. Cuenta con 23 variables originales. Se usará como test de evaluación el mes de diciembre 2020.

Metodología

1. EDA

- Cargar el dataset en un entorno de trabajo como Python con Pandas y NumPy.
- Realizar un análisis de distribución y balanceo de clases.

2. Ingeniería de variables

- Crear nuevas variables que aporten información relevante al modelo, como contadores, acumuladores, condiciones.

3. Implementación del modelo base

- Entrenar un modelo inicial con LightGBM utilizando métricas tradicionales como AUC-ROC y F1-score.
- Evaluar el rendimiento inicial en términos de fraude detectado y falsos positivos.

4. Definición de métricas personalizadas

- Implementar diferentes funciones feval para LightGBM, por ejemplo:
- Penalización de falsos positivos.
- Métricas balanceadas que favorezcan una menor tasa de falsos positivos sin reducir drásticamente la detección de fraude.
- La métrica que evaluamos es la ratio de falsos positivos definido como: **ratio falsos positivos = (TP + FP) / TP.**

5. **Evaluación de resultados**

- Comparar el rendimiento de las distintas métricas feval usando el último trimestre del dataset como testing. El comparativo debería ser la ratio de falsos positivos buscando tener un 90% de detección.
- Determinar cuál estrategia logra el mejor balance entre falsos positivos y detección de fraude.

info@plusti.com 706 257 - 6555

1921 Whittlesey Road Suite 500 Columbus, Georgia 31904

6. **Optimización con objetivos diferenciados por alumno**

- A cada alumno se le asignará un objetivo de optimización específico, el cual puede estar relacionado con distintos aspectos del desempeño del modelo. En función del objetivo asignado, cada alumno deberá diseñar al menos tres funciones de evaluación personalizadas.

- Ajustar parámetros del modelo y/o pesos de clases según sea necesario para cumplir con el objetivo.

- Evaluar y comparar el rendimiento de cada función diseñada, seleccionando la más efectiva y justificando su elección con evidencia cuantitativa obtenida del modelo

Recomendaciones

- Documentar cada paso del proceso en el notebook.
- Utilizar visualizaciones para respaldar el análisis.
- Realizar optimización de hiperparámetros
- Explicar el razonamiento detrás de cada decisión tomada.

Entrega

El trabajo deberá entregarse en forma de un informe técnico de no más de cuatro páginas, anexando los notebooks utilizados para la implementación y los resultados de los experimentos. El informe debe incluir:

- Resumen

- Metodología
- Descripción de la implementación práctica
- Conclusiones
- Análisis de los resultados de la evaluación, con énfasis en el comparativo de estrategias

Recursos Adicionales

Se puede optar por utilizar Google Colab para facilitar el uso de recursos computacionales, incluyendo GPU si es necesario. Se recomienda el uso de bibliotecas de Python como Scikit-learn, lgbm para la implementación de los modelos.

Link descarga Dataset y Script Feature Engineering:

https://drive.google.com/drive/folders/1KLbqDPMAIB58zMEeNkPSnQzm21TbTGwN?usp=sharing

Enrique Coloch | Coordinador Data Science
ecoloch@plusti.com
Plus Technologies & Innovations
Por un Mundo Financiero Más Seguro
ecoloch@plusti.com
www.plusti.com

Headquarters Columbus, GA, USA
Regional Office Guatemala, Central America
Office +502 2383 1616
LinkedIn®