Primera entrega de proyecto

POR:

Santiago Escobar Casas

MATERIA:

Inteligencia artificial para las ciencias e ingenierías

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERIA
MEDELLIN 2022

1. Planteamiento del problema

Actualmente, en pleno siglo XXI muchas de las compras realizadas en el mundo se hacen mediane tarjetas de crédito lo cual conlleva beneficios y facilidades tanto para el comprador como para el vendedor. Sin embargo, existen casos en donde se realizan transacciones fraudulentas en la cuales son cargados artículos o montos a usuarios sin haber sido ellos quienes realizaran la compra. Para evitar esto es posible generar modelos, utilizando conceptos de IA, los cuales permitan reconocer si una transacción es fraudulenta o no.

2. Data set

El conjunto de datos a utilizar (llamado Credit Card Fraud Detection) es tomado de Kaggle y contiene 284,807 transacciones realizadas (de las cuales 492 fueron fraudulentas) a lo largo de dos días en septiembre de 2013 por tarjetahabientes europeos.
Consta de 31 columnas

- **Time**: indica la diferencia (en segundos) entre dicha transacción y la primera del dataset)
- **V1** hasta **V28**: variables numéricas resultado de una transformación de PCA realizada a los datos. Por motivos de confidencialidad no es dado ningún contexto ni referencia individual adicional de ninguna de estas variables.
- **Amount**: cantidad de dinero por la que fue realizada la transacción (no se especifica moneda)
- **Class**: indica con 1 las transacciones fraudulentas y con 0 lo demás. Esta es la variable objetivo a predecir

Nota: Por la gran cantidad de muestras y el costo computacional que conlleva su procesamiento, es posible que solo se tome una parte del total para realizar el proyecto.

3. Métricas

Debido a que las clases del dataset no están balanceadas lo mas razonable es utilizar métricas apropiadas para esta clase de situaciones tal y como lo es “F1-score” la cual combina “Precision” (de todo aquello que ha sido clasificado como positivo, cuanto lo es realmente) y “Recall” (de todo lo que es realmente positivo, cuanto fue correctamente clasificado).

$$Precision = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Positives}$$

$$Recall = \frac{\#\ of\ True\ Positives}{\#\ of\ True\ Positives + \#\ of\ False\ Negatives}$$

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Nota: Otras posibles métricas para problemas desbalanceados y que se tendrán como secundarias (en el sentido de que podrán ser obtenidas y analizadas, pero F1 será la que determine el desempeño del modelo) son la matriz de confusión y las anteriormente mencionadas precision y recall.

4. Desempeño

Para este caso es de vital importancia tener en cuenta el desbalance de clase, en este mismo sentido el objetivo es doble. Principalmente, detectar la mayor cantidad de transacciones fraudulentas sobre todo en transacciones con montos altos para así evitar inconvenientes para los tarjetahabientes y perdidas para los vendedores o los mismos bancos. Y de manera secundaria reducir los falsos positivos (teniendo el fraude como clase positiva) ya que esto podría bloquear y hacer infructuoso el manejo de la tarjeta por los usuarios.

5. Bibliografía

- *Credit card fraud detection*. (s. f.). Recuperado 5 de julio de 2022, de https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- *Credit fraud || dealing with imbalanced datasets*. (s. f.). Recuperado 5 de julio de 2022, de https://kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets
- Korstanje, J. (2021, agosto 31). *The F1 score*. Medium. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6