Segunda entrega

Integrantes

Diego Alejandro Poveda Alzate
Pedro Pablo Saldarriaga Jaramillo
Jhonier Andrés Córdoba Asprilla

Tutor

Raul Ramos Pollán

Modelos y Simulación de Sistemas 1 / Introducción a la Inteligencia Artificial

Universidad de Antioquia
Medellín, Antioquia
2023

1. Planteamiento del problema

Las tarjetas de crédito son uno de los medios de pago más utilizados en todo el mundo. Con el aumento del comercio electrónico, las transacciones con estas tarjetas en línea también han aumentado significativamente. Esto, lamentablemente, también ha llevado a un aumento en las actividades fraudulentas relacionadas con las tarjetas. Los bancos y las empresas emisoras de tarjetas están constantemente buscando formas de detectar y prevenir transacciones fraudulentas para minimizar las pérdidas financieras y mantener la confianza de sus clientes.

2. Dataset a utilizar

El dataset a utilizar proviene de una competencia de Kaggle, compuesto por un conjunto de archivos **.csv** (comma separated values), los cuales nos dan la información requerida.

Este dataset fue generado mediante un modelo de deep learning que previamente fue entrenada en el problema de Detección de Fraudes con Tarjetas de Crédito (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

El enlace a la competencia es https://www.kaggle.com/competitions/playground-series-s3e4

Columnas del dataset

- **Id**: un identificador único para cada transacción, utilizado para indexar los datos y no tiene poder predictivo.
- **Tiempo**: el tiempo transcurrido en segundos entre esta transacción y la primera transacción en el dataset. No es la marca de tiempo real de la transacción sino más bien una medida de tiempo relativa.
- **V1 a V28:** son el resultado de una transformación (específicamente Análisis de Componentes Principales, o PCA) aplicada a los datos originales sin procesar. PCA es una técnica de reducción de dimensionalidad que puede transformar un conjunto de variables correlacionadas en un conjunto de variables no correlacionadas llamadas componentes principales. La idea principal detrás del PCA es capturar la mayor cantidad de variabilidad en los datos con la menor cantidad posible de componentes principales. La naturaleza exacta y el significado de estas variables no se proporcionan por razones de confidencialidad. Sin embargo, son características numéricas derivadas de los datos originales sin procesar sobre la transacción.
- **Amount (Monto):** El monto de compra de la transacción. Esta es una característica crucial, ya que ciertas cantidades pueden ser indicativas de actividades fraudulentas, según el contexto.
- **Class (Clase):** Variable respuesta que tiene el valor de 1 en caso de que la transacción sea fraudulenta, y 0 en otro caso.

Columnas simuladas

Además de las columnas que tiene el dataset, se simularán otras columnas, como por ejemplo

- **Medio:** Columna categórica, con valores como “compra en línea”, “punto de venta” o “retiro ATM”.
- **AmountRange:** Columna categórica que indica el rango en el que se encuentra la transacción según su **Amount**(alto, medio o bajo).
- **TipoTarjeta:** Columna categórica que indica el tipo de tarjeta que se usó en la transacción.

3. Métricas

La competencia recomienda el cálculo de la métrica AUC-ROC pero, debido a que este cálculo es muy complicado a nuestro parecer, decidimos hacer el análisis de otras métricas como, por ejemplo:

- **Accuracy**: Calcula la proporción de predicciones de fraude correctas. Si solo el 0.1% de las transacciones son fraudulentas, un modelo que predice que todas las transacciones son legítimas aún tendría una precisión del 99.9%.

- **Precision:** Indica qué proporción de las transacciones que el modelo predice como fraudulentas lo son realmente. Dada la gravedad del fraude, es importante tener una precisión alta para evitar falsas alarmas.

- **Recall:** Mide qué proporción de las transacciones fraudulentas reales fueron correctamente detectadas por el modelo. En el contexto de detección de fraude, un recall alto es crucial porque no queremos pasar por alto eventos fraudulentos

- **F1-Score:** Dado que tanto la precisión como el recall son importantes en la detección de fraudes, el F1-Score ayuda a encontrar un equilibrio entre ambos.

4. Desempeño

Con este modelo en ambiente productivo el objetivo es que logre analizar de forma correcta las transacciones de los clientes de cierta organización y que catalogue con buena precisión las operaciones que considere como fraudulentas, además, el número de las transacciones etiquetadas como falsos positivos (fraude que en realidad no lo sean) sea el más bajo posible. Para esto se tomarán en cuenta algunos criterios como por ejemplo:

- **Alto Recall (Sensibilidad):** Es esencial que el modelo sea capaz de identificar la mayoría, si no todas, de las transacciones fraudulentas. Un falso negativo, donde una transacción fraudulenta no se detecta, puede tener consecuencias financieras y de reputación para la empresa.

- **Precision Aceptable:** Aunque queremos un recall alto, la precisión también es importante. Un gran número de falsos positivos, donde las transacciones legítimas se marcan como fraudulentas, puede llevar a una mala experiencia del cliente y a un aumento de los costos operativos, ya que cada alerta podría necesitar revisión manual.

- **F1-Score:** Debido a que tanto el recall como la precisión son importantes, el F1-score puede ser una métrica útil para equilibrar ambos.

5. Exploración y simulación de datos

Al hacer la revisión del dataset que nos brinda Kaggle para esta competencia, y de acuerdos a las indicaciones dadas por el profesor a la hora de proponer el uso del mismo para el proyecto, se evidencia que no se tiene el número de columnas categóricas necesarias, y tampoco se tiene el porcentaje mínimo de datos faltantes requeridos para el proyecto.

Por esto mismo, se decidió simular los requisitos faltantes, es decir, agregar al menos 3 columnas categóricas y simular la falta de información.

Las columnas categóricas que se agregaron mediante la simulación son:

- Rango de monto pagado
- Medio en el cual se realizó la transacción
- Tipo de tarjeta usada para la transacción

Además de esto, en estas columnas simuladas se eliminaron un porcentaje de datos, para así cumplir con el requisito de valores faltantes. Se decidió hacerlo en estas columnas ya que, en un caso de la vida real, es posible que alguno de estos valores no se conozca.