INTRODUCCIÓN A LA CIENCIA DE DATOS

Tarea Final

AÑO 2024

GRUPO 14:

Fernando Sellanes Fajardo
Lucas Sellanes Fajardo

1. Presentación del Dataset

La identificación de usuarios fraudulentos es uno de los tantos problemas a los que se pueden enfrentar las empresas. Gracias a los avances tecnológicos en el área de la ciencia de datos es posible, a través de diversos modelos, enfrentar esta problemática de una forma más rápida y eficiente. Para este trabajo seleccionamos una base de datos libre con el fin de explicar las etapas del proceso de análisis, así como las dificultades de cada una de ellas.

1.1. Dataset

Para el desarrollo de esta tarea, se eligió un data set de la plataforma Kaggle, el cual consta de 16 variables de información sobre transacciones web y sus compradores. Descripción de algunas variables:

- ***Transaction ID:*** identificador único de la transacción
- ***Customer ID:*** identificador único para cada comprador
- ***Transaction Amount:*** el valor total de la transacción
- ***Payment Method:*** método con el cual se hizo el pago (*credit card*, *PayPal*, etc.)
- ***Product Category:*** categoría del producto involucrado en la transacción
- ***Customer Age:*** edad del comprador
- ***Customer Location:*** ubicación del comprador
- ***Device Used:*** tipo de dispositivo en el cual se llevó a cabo la transacción
- ***IP Address:*** dirección IP del dispositivo
- ***Account Age Days:*** días que pasaron entre que se creó el usuario y se realizó la transacción
- ***Is Fraudulent:*** *True* si la transacción fue fraudulenta *False* si es legítima

Esta es una base de datos sintética que simula transacciones con patrones reales y escenarios de fraude. No es el caso de esta base de datos, pero las usadas para analizar fraude pueden contener información sensible y personal, como nombres, números de identificación (cédula, pasaporte, etc.), detalles bancarios y más. Para ello es importante tener buenos protocolos de almacenamiento y transferencia de datos, así como también técnicas adecuadas de anonimización para evitar la reidintificación de los individuos.

1.2. Preprocesamiento de datos

Esta etapa tiene como fin limpiar y transformar los datos previo al análisis. La identificación de datos faltantes y/o de doble interpretación, el entendimiento de las variables y la visualización de los datos son algunos de los pasos que se dan en esta primera sección del análisis.

1.2.1. Datos faltantes y de doble interpretación

En este caso un posible dato faltante en la base de datos podría ser la edad del comprador (*Costumer Age*), el cual puede tratarse de un dato importante para el modelo. Las opciones para resolverlo podrían ser directamente eliminar estos ID con datos faltantes, pero para

no perder mucha información se puede optar por utilizar la edad promedio de los clientes sabiendo que podría existir un sesgo en el modelo en caso de que falte mucha información.

Por otro lado, en esta base de datos se encuentra la variable de la ciudad del comprador (*Customer Location*), la cual puede tener una doble interpretación si se presenta una ciudad que se repite en diferentes países. Si bien es una variable que puede confundir al modelo en caso de repetirse, es una variable de importancia para chequear que el IP utilizado efectivamente coincide con esta dirección o se trata de una persona que utiliza un VPN para cambiar el IP, esto podría indicar actividad fraudulenta.

1.2.2. Visualizaciones

En una primer instancia es posible realizar diferentes visualizaciones a modo de análisis exploratorio que caractericen la base de datos y permitan una mejor comprensión. En este sentido la realización de histogramas de variables como *Costumer Age*, *Transaction Amount* o *Payment Method* permiten ver como estas variables están distribuidas. Asimismo, la visualización de variables en conjunto es de gran ayuda para poder identificar correlaciones, como por ejemplo *Costumer Age* vs *Device used*.

Otra alternativa puede ser hacer un análisis exploratorio utilizando técnicas de aprendizaje no supervisado, para lo que hay que dejar de lado la variable *Is Fraudulent* y buscar patrones o anomalías en el conjunto de datos. Teniendo en consideración la cantidad de variables, es importante buscar reducir la dimensión de la base de datos utilizando métodos como Análisis de Componentes Principales (PCA, por sus siglas en inglés). Con el fin de obtener representaciones en un plano o 3 dimensiones, y observar si existe alguna relación que sobresalga entre las demás. Si las primeras componentes tienen suficiente varianza como par explicar las relaciones entre si, se podrían graficar las proyección de las variables de interés en las componentes principales y observar si estas se correlacionan.

1.2.3. Proceso previo a la utilización del modelo

Dado que los modelos predictivos de aprendizaje automático trabajan con variables numéricas es importante transformar este dataset en una matriz numérica capaz de ser leída por el modelo.

De esta forma, se pueden identificar los diferentes ID y variables con valores numéricos, que luego deben ser estandarizados para el modelo.

A su vez, se separa el dataset en *train/test* de manera estratificada ya que en el contexto del problema a resolver el desbalance de datos es importante, las transacciones legítimas son mucho mas frecuentes que las frauduentas.

2. Modelo predictivo

Debido a la naturaleza del problema, el desbalance de datos tiene importancia al momento de elegir un algoritmo, por lo que se optó por el de *Random Forest*. Es especialmente útil debido a su capacidad de manejar datos desequilibrados y una gran robustez contra el sobre ajuste.

El algoritmo se basa en la construcción de múltiples árboles de decisión durante el entrenamiento. Utilizando una técnica llamada “bagging” crea subconjuntos de datos de entrenamiento mediante muestreo con reemplazo. Esto significa que algunos datos pueden

repetirse en diferentes subconjuntos, mientras que otros pueden no aparecer. Para cada subconjunto de datos, se construye un árbol de decisión. Para hacer una predicción de clasificación, se pasan los datos de entrada a través de cada uno de los árboles, emitiendo un "voto" y la clase final se determina por mayoría.

Ventajas:

- **Robustez:** es menos propenso al sobreajuste en comparación a un solo árbol de decisión.
- **Capacidad de manejo de datos:** puede manejar grandes conjuntos de datos con alta dimensionalidad y alto desequilibrio.
- **Importancia de características:** proporciona una medida de la importancia de cada caracterísitca en la predicción, lo que puede ayudar a identificar patrones y tendencias relevantes.

Desventajas:

- **Interpretación:** puede ser difícil interpretar el modelo en comparación con un solo árbol de decisión.
- **Tiempo de cálculo:** la construcción de muchos árboles de decisión es computacionalmente costosa.

2.1. Métricas de evaluación

Una vez entrenado el modelo, este debe pasar por un proceso de evaluación utilizando métricas que representen su comportamiento. Así es posible utilizar métricas como las siguientes: *accuracy*, *precision* y *recall*.

Sin embargo, al tratarse de un problema de identificación y etiquetado de clientes como fraudulentos, para una empresa es muy importante lograr un correcto etiquetado disminuyendo al mínimo los casos Falsos Positivos, de modo de no afectar buenos clientes. La métrica de evaluación del modelo que destaca sobre las demás debe ser el *recall*.

3. Comentarios finales

A partir de esta esquematización de un proyecto de análisis de datos es posible tener una herramienta que identifique a clientes fraudulentos de una empresa.

En específico, el modelo puede identificar cuales características comparten los usuarios fraudulentos y asociar una probabilidad a que estos efectivamente lo sean.

Los patrones de fraude pueden cambiar con el tiempo, por lo que es importante ajustar el modelo regularmente. Monitoreando las métricas antes mencionadas se puede detectar este desvío.