UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Tecnologías Aplicadas a Soluciones de Datos
“Trabajo práctico cuatrimestral”
Detección de fraudes con tarjetas de crédito

Grupo n° 5

Integrantes – Año 2022

Legajo	Nombre	E-Mail
163.724-1	Busso, Alejo Joel	abusso@frba.utn.edu.ar
152.312-0	Gonzalez, Juan Pablo	juanpgonzalez@frba.utn.edu.ar
138.390-5	Costas, Lucas Omar	lcostas@frba.utn.edu.ar
153.643-6	Egüez, Cristian	ceguez@frba.utn.edu.ar

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Índice

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Introducción

En la actualidad existen alrededor de 17 mil millones de tarjetas de crédito emitidas en el mundo y se espera que se incremente a 19 mil millones en el año 2024. Con este incremento constante de usuarios de tarjetas de crédito también han aumentado las estafas y/o intentos de estafa.

Se estima que existen 24 mil millones de datos obtenidos ilegalmente, siendo los más buscados los datos de tarjetas de crédito y/o débito. Existe incluso un mercado para los estafadores que compran (y venden) datos en internet y los utilizan para comprar artículos de lujo o lavar fondos ilícitos.

Actualmente las formas más utilizadas para la extracción de datos fraudulenta son las siguientes:

- **Robo de datos por phishing**
  El phishing es un engaño por el cuál el cibercriminal se hace pasar por una entidad legítima para engañar al usuario y convencerlo de que ingrese sus datos personales o descargue un software malicioso sin darse cuenta.
  El phishing, o pesca en su traducción al español, consiste en utilizar la ingeniería social para solicitar información personal con la excusa de avanzar con la compra de un producto en oferta, sorteo, promociones u otra actividad legítima. Luego utilizarán estos datos para vulnerar al usuario, vulnerar sus activos o suplantarlo en otra operación.
  Esta forma de ataque es muy conocida y alcanzó máximos en el primer trimestre de 2022. Se incrementó mucho en la pandemia de sars-covid-19.

- **Robo de datos con malware**
  La técnica de robo de datos con malware se basa en que el usuario ingrese a un sitio web o descargue una aplicación, el cuál está infectado con un un troyano capaz de robar información o instalar otras amenazas como virus y ransomware.
  Esto les proporciona a los criminales acceso a sus dispositivos permitiéndoles robar sus datos.

- **Robo de datos por web skimming**
  Los ciberdelincuentes instalan malware en páginas de pago de sitios de comercio electrónico legítimos. Estos códigos maliciosos son invisibles para un usuario pero sustraen el detalle de la tarjeta a medida que son ingresados.

- **Filtraciones de datos**
  Esta es una infracción de la seguridad, en la que datos sensibles, protegidos o confidenciales son copiados, transmitidos, vistos, robados o utilizados por una persona no autorizada para hacerlo.

- **Robos de datos en redes wifi públicas**
  Este caso suele darse en lugares donde los usuarios utilizan redes públicas para las conexiones desde un celular o notebook. Los lugares más frecuentes son aeropuertos, hoteles y cafeterías entre otros. Los ciberdelincuentes se conectan a la red y pueden espiar datos de terceros a medida que son ingresados.

Objetivo

El objetivo principal del presente trabajo de investigación es informar sobre la utilidad de la ciencia de datos y aprendizaje automatizado en la identificación de compras fraudulentas en tarjetas de crédito. A su vez, ejemplificaremos este tipo de funcionalidades con el desarrollo de una técnica de detección, desarrollada en lenguaje de programación Python, que predice si una compra es fraudulenta o no en base a datos obtenidos.

Desarrollo

Etapas de transacción

En base a la introducción y a las múltiples maneras que existen y existirán de sustracción de datos sensibles surgió la necesidad por parte de las entidades financieras de tratar de prever la autenticidad de las compras realizadas por los usuarios. Con esto se busca minimizar las operaciones fraudulentas ya que en el año 2020 la cifra mundial de pérdidas por estas transacciones fue de US$31.000 millones. Por lo tanto, en este trabajo de investigación se abordará el tema de la detección de dichas transacciones utilizando aprendizaje automático e inteligencia artificial.

Para poder comprender de forma efectiva el riesgo que existe en las transacciones con tarjetas de crédito es necesario tener un conocimiento del ciclo de vida de la transacción y sus riesgos asociados en cada etapa dentro del ciclo.

Las etapas de la transacción son las siguientes:

- Autorización
- Liquidación y compensación
- Contracargo

En este trabajo de investigación, se hará foco en la etapa de autorización de la compra, en compras de bajo la modalidad no presencial, en la cual la etapa consta de los siguientes pasos:

Inicialmente, (1) el cliente ingresa los datos de la tarjeta en la página web donde desea comprar, (2) el comercio envía la petición de autorización al adquiriente, (3) el adquiriente del comercio le pide al sello la autorización del emisor de la tarjeta,(4) el sello envía la transacción al emisor para que la autorice, (5) el emisor evalúa la transacción por medio de sus sistemas de seguridad, (6) luego el emisor da una respuesta en línea al sello, (7) el sello envía la respuesta al adquiriente, (8) el adquiriente envía al comercio la respuesta recibida desde el emisor por medio del sello. En este último paso el comercio puede recibir tres respuestas: rechazo, aprobación o aprobación parcial. Este último ítem se refiere a

cuando el emisor no pudo validar la transacción y depende del comercio si continúa con la transacción o no. En caso de que el comercio continúe, sí la misma resulta en un fraude, será responsabilidad de dicho comercio. (9) finalmente en respuesta al comercio, responde al cliente si la compra es aceptada o rechazada. Nuestro algoritmo de detección de compras fraudulentas se centrará en el paso (5). *Figura 1*

Datos

El rendimiento de la detección del fraude en las transacciones de tarjetas de crédito se ve afectado en gran medida por el enfoque de muestreo del conjunto de datos, la selección de variables y las técnicas de detección utilizadas. En este caso, utilizaremos aprendizaje automatizado ya que presenta algunas ventajas como un procesamiento eficiente de datos, reconocimiento de patrones y predicciones precisas utilizando regresiones logísticas, redes neuronales y bosques aleatorios. Es importante destacar que a medida que se alimente con más datos el sistema, el mismo será capaz en consecuencia de aumentar su precisión general. Para este algoritmo, se utilizará un dataset con la ubicación geográfica de las compras además de:

- Tiempo: representa los segundos transcurridos entre la transacción actual y la primera transacción en el conjunto de datos
- Cantidad: representa el valor total de la transacción

Y obtendremos una variable de respuesta, llamada “Clase” con dos valores:

- 1 para casos fraudulentos
- 0 para casos no fraudulentos

Para el entrenamiento del modelo, será necesario contar con la misma cantidad de casos de ambas clases. Es decir, necesitamos un 50% de clase 1 y el restante 50% de clase 0. *Figura 2*

Según el dataset obtenido, la mayoría de las compras fraudulentas tienden a tener valores más bajos. Por otra parte, las no fraudulentas no se distribuyen de manera uniforme.

En términos del tiempo, no existe una diferenciación entre ambas clases, estos valores son consistentes.

La demostración se desarrolló en python versión superior a 3.7 utilizando las bibliotecas de código abierto numpy, pandas y sklearn. Dichas bibliotecas se utilizaron para manejar el dataset, alimentar a la inteligencia artificial y efectuar la predicción.

Entrenamiento

Al tener el conjunto de datos equilibrados, se procede a construir los modelos para el aprendizaje automático, utilizando dos técnicas:

- Bosque aleatorio
- Regresión logística

La ejecución del bosque aleatorio con 15 estimadores arrojó los resultados expuestos en la *Figura 3 i*ndicando una precisión de 86.36%

Por otra parte, la ejecución de la regresión logística arrojó los resultados de la *Figura 4* indicando una precisión del 8.78%.

Como consecuencia, entre ambos modelos se opta por seleccionar el de bosque aleatorio ya que es más preciso. Una vez obtenido el modelo con dicha precisión ya está disponible para ser utilizado en un entorno productivo el cuál se encargará de predecir el origen de las compras.

Conclusiones

Se destaca que es sumamente importante para una institución financiera, la utilización de herramientas que permitan predecir el fraude en las compras. Esto es debido a su utilidad y capacidad para evitar grandes pérdidas. Si bien el aprendizaje automatizado suele ser eficiente es importante destacar que pueden existir falsos positivos los cuales generan situaciones incómodas y engorrosas para los clientes que desean adquirir legítimamente un bien y/o servicio. De todas formas, al existir el autoaprendizaje, a medida que el modelo sea más utilizado se reducirán los falsos positivos y/o falsos negativos.

Referencias

Calvo Pérez, Ismael (2021). Algoritmos de aprendizaje automático para detección de fraudes con tarjetas de crédito: Análisis y comparativa. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. de Sistemas Informáticos (UPM), Madrid.

RENJITH MADHAVAN(2018), Algoritmo de aprendizaje automático para detección de fraudes con tarjeta de credito. https://www.kaggle.com/code/renjithmadhavan/credit-card-fraud-detection-using-python/notebook

Anexo

Figura 1

Figura 2

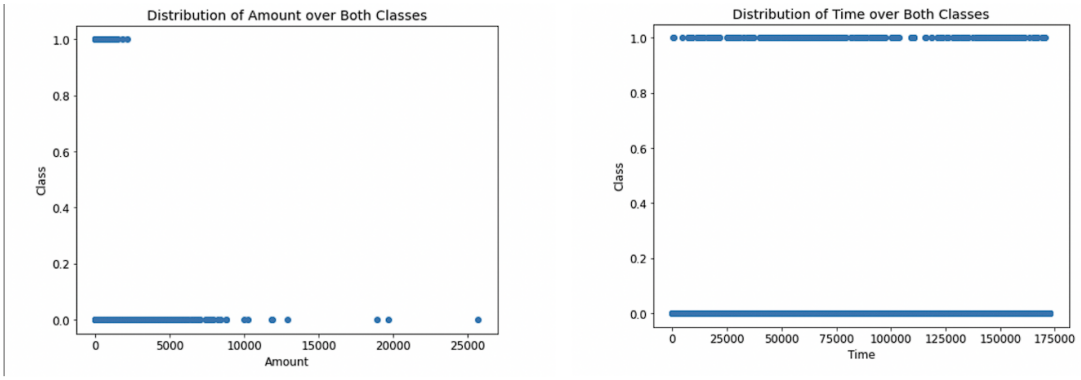
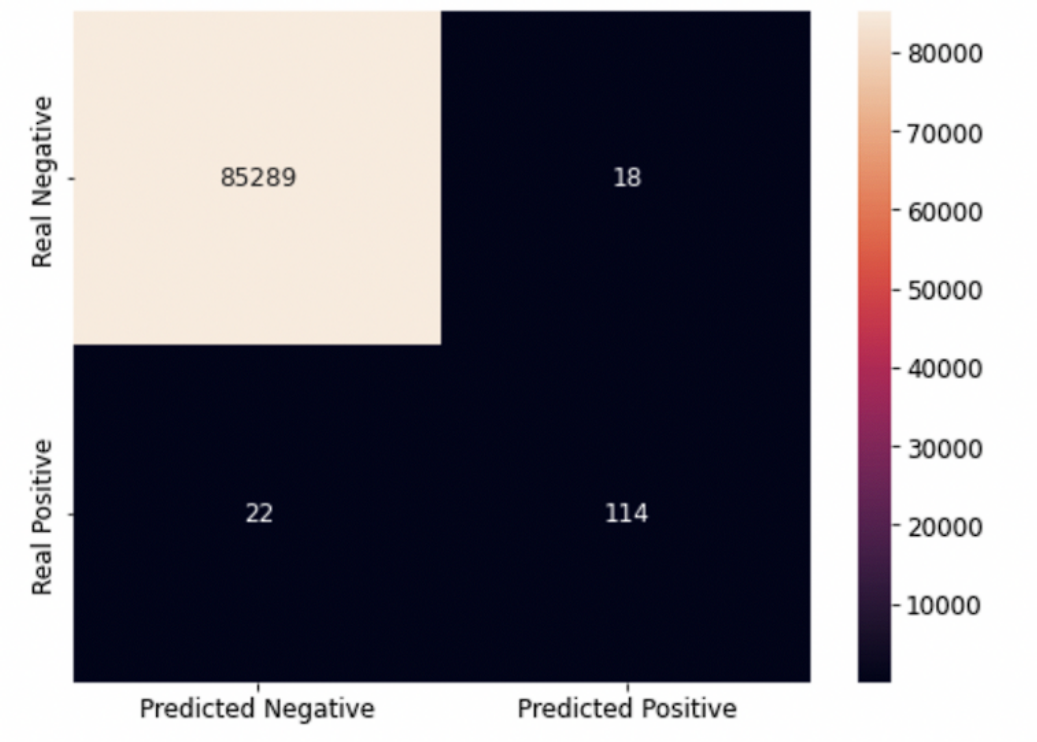
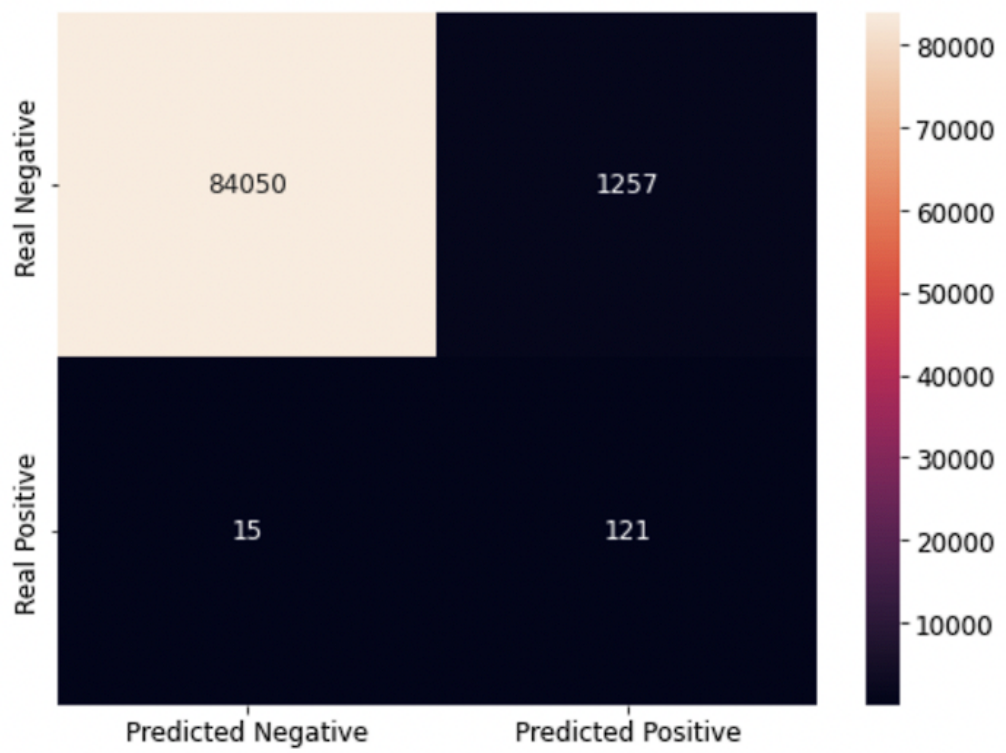


Figura 3



Precision: 86.36 %
Recall: 83.82 %
F1: 85.07 %

Figura 4




```
Precision:  8.78 %
Recall:  88.97 %
F1:  15.98 %
```