UNIVERSIDAD DEL VALLE DE GUATEMALA

Security Data Science

Sección 10

Catedrático: Jorge Yass

Informe Final

Proyecto Final

Oscar Esteban Donis Martínez 21610

Resumen

El proyecto tiene como objetivo identificar transacciones fraudulentas que forman parte de ataques coordinados mediante el uso del algoritmo LightGBM, un modelo de gradient boosting basado en árboles de decisión. Se implementó un preprocesamiento exhaustivo que incluyó la limpieza de datos (uniformización de variables categóricas y conversión de fechas a formato Unix) y la ingeniería de características, creando variables como time_diff_seconds, hour_window, trans_per_hour, unusual_distance, velocity_km_h, amt_month_ratio, high_amt_first_time, unique_cards_per_hour, entre otras, para capturar patrones de comportamiento anómalo. La variable objetivo is_coordinated_attack se definió combinando ocho criterios de sospecha en cuatro patrones complejos, identificando transacciones fraudulentas que forman parte de campañas orquestadas. Se utilizaron funciones de evaluación personalizadas (Unique Card, Merchant Frequency, Distance Anomaly) para priorizar la detección de patrones específicos, como el uso excesivo de tarjetas únicas, frecuencias de compra inusuales o distancias geográficas anómalas. Los modelos fueron evaluados mediante métricas como la matriz de confusión, curva ROC-AUC, precisión, accuracy y F1-Score. Los resultados muestran que el modelo Unique Cards obtuvo el mejor rendimiento, con una mejora significativa en la reducción de falsos positivos (50% menos que el modelo base Basic LGBM), aunque la precisión general (9.2%) se vio afectada por el desbalance de clases y los falsos positivos. La curva ROC-AUC indicó que los modelos con funciones especializadas (Unique Cards y Merchant Frequency) tienen mayor certeza en la clasificación que el modelo base, mientras que Distance Anomaly mostró el peor desempeño, con un 13% menos de accuracy y un 62% más de falsos positivos.

Metodología

Objetivo

Nuestro objetivo para este proyecto es poder identificar las transacciones que se realicen como parte de un ataque coordinado de fraude de tarjetas. Para ello debemos poder identificar los factores que hacen que un fraude pueda ser parte de una estafa mucho más grande.

Factores

Algunos de los factores para identificar ataques coordinados de estafas de tarjetas si el vendedor tiene, en una ventana de tiempo muy pequeña, una cantidad inusual de tarjetas de crédito únicas utilizadas para realizar compras a un solo vendedor. Diferencia en la distancia de transacciones seguidas, también si el usuario gasta por primera vez en un vendedor en concreto y si la cantidad está muy por encima de lo que gasta la persona usualmente. También debemos de tomar en cuenta si existe una variación muy alta en la cantidad gastada durante una ventana de tiempo, comparado con lo que usualmente compra el usuario.

Preprocesamiento y Transformación de Datos

Definición de la Variable Objetivo para Ataques Coordinados:

Para determinar si un ataque es coordinado, se establecen ocho criterios basados en anomalías como: el uso excesivo de tarjetas en un mismo comerciante, transacciones físicamente imposibles, compras iniciales con montos muy altos, distancias o frecuencias extremas, montos con muy baja varianza, e intervalos muy cortos entre transacciones. Luego se combinan en cuatro patrones complejos, y una transacción se marca como parte de un ataque coordinado si, además de ser fraudulenta, coincide con al menos uno de estos patrones.

Descripción de la Implementación Práctica

Funciones de Evaluación Personalizada

Unique Card

Esta función evalúa un modelo penalizando su incapacidad para detectar transacciones con patrones típicos de ataques coordinados, como el uso de muchas tarjetas únicas en poco tiempo o montos de transacción con muy baja varianza. Combina el error general de clasificación con un "recall coordinado" para priorizar la identificación de estos patrones específicos

Merchant Frequency

Esta función guía al modelo para identificar transacciones sospechosas de ser parte de un ataque coordinado, como la primera transacción con un comerciante, una primera transacción de alto monto o una frecuencia de compra diaria anómala. Combina la precisión general del modelo con un "recall de frecuencia" para priorizar la detección de estos patrones específicos, dando más peso a este último.

Distance Anomaly

Esta función prioriza la identificación de transacciones geográficamente sospechosas, como aquellas con distancias inusuales o velocidades imposiblemente altas entre transacciones. Combina métricas generales de precisión y recall con un F1-score específico para anomalías de distancia, dando mayor peso a este último, para evaluar el rendimiento del modelo.

Análisis de Resultados

Matrices de Confusión

Se puede observar que al comparar las matrices de confusión de todos los modelos, tenemos que el modelo **Basic LGBM** pudo clasificar todas las transacciones parte de ataques coordinados, sin embargo, podemos ver que clasificó una cantidad muy grande de falsos positivos (alrededor del 75% de las transacciones que no son parte de ataques coordinados). Tenemos que los otros dos mejores fueron el **Unique Cards** y el **Merchant Frequency**, teniendo una caída considerable de falsos positivos comparado con el **Basic LGBM**. Siendo el mejor de los dos el **Unique Cards**, con una mejora del 50% en los falsos positivos, mientras mantiene la misma precisión en los ataque coordinados. Y, por último, tenemos el modelo **Distance Anomaly**, el cual tuvo el peor rendimiento clasificando de los cuatro, teniendo una acertividad 13% menor al de los modelos más cercanos y un incremento del 62% en los falsos positivos con el modelo más cercano.

Curva ROC-AUC

Se puede observar que el modelo **Basic LGBM**, se encuentra muy cerca de la pendiente, lo cuál podría indicar que su método de clasificación, por más efectivo que sea, tiende a generar muchos de los resultados sin mucha certeza, resultando así en una vasta cantidad de falsos positivos. Se puede ver que los modelos con funciones de evaluación especializadas cuentan con una mejor trayectoría, mostrándonos que sus predicciones no son al azar y que si poseen cierta certeza a la hora de clasificar. Se puede ver que el mejor modelo es el de **Unique Cards**, con una diferencia del 5% comparado con el siguiente mejor modelo, que es el **Merchant Frequency.** Se puede ver que los resultados de los modelos con función de evaluación se encuentran con una diferencia de 4.07% entre cada uno de los modelos, lo cual señala que los modelos si están prediciendo con conocimientos adquiridos y no con predicciones aleatorias.

Precisión - Accuracy - F1_Score

Al visualizar las métricas de precisión, se puede ver que el mejor modelo es el de **Unique Cards**, con una diferencia del 52% con el modelo siguiente mejor modelo. Sin embargo, se ve claramente que este valor está muy por debajo del valor que esperamos, apenas llegando al 9.2%, lo cual se podría explicar por la gran cantidad de falsos positivos que obtuvimos, que son muchos más que nuestros verdaderos positivos. Aunque la cantidad de falsos positivos es mucho menor comparada con la cantidad de transacciones en el dataset de test, siendo solo el 0.1% de todos los negativos. Podemos ver que en el recall, el modelo sin ninguna función de evaluación es perfecto, sin embargo, esto es porque el Recall no toma en cuenta los falsos positivos, lo cual se ve reflejado tanto en Precisión como en el F1-Score. Luego tenemos un empate entre **Unique Cards** y **Merchant Frequency**, los cuales se pueden ver que obtuvieron los mismos resultados en los verdaderos positivos, y por último tenemos el modelo **Distance Anomaly**, el cual solo logró acertarle a 13 de los 16 valores. En el F1-Score, se puede observar que la métrica es bastante similar a la de precisión, que similar a esta, se ve el peso que tienen los falsos positivos sobre el

rendimiento de los modelos. Aunque la diferencia de los falsos positivos no es tan significativa, al comparar la cantidad de transacciones que son parte de un ataque coordinado contra las que no lo son; aún así podemos ver que el tener esos falsos positivos afecta mucho las métricas que toman en cuenta el rendimiento overall de nuestro modelo.

Conclusiones

El desarrollo del modelo basado en LightGBM demostró ser efectivo para identificar transacciones fraudulentas dentro de ataques coordinados, destacando la importancia de la ingeniería de características y las funciones de evaluación personalizadas para capturar patrones específicos de fraude. El modelo Unique Cards se destacó como el más robusto, logrando un equilibrio entre la detección de ataques coordinados y la reducción de falsos positivos, aunque la precisión general se vio limitada por el desbalance de clases y la gran cantidad de falsos positivos en comparación con los verdaderos positivos. La curva ROC-AUC confirmó que los modelos con funciones especializadas superan al modelo base en certeza predictiva, lo que resalta la utilidad de métricas personalizadas para priorizar patrones de ataque específicos. Sin embargo, el modelo Distance Anomaly mostró un rendimiento inferior, sugiriendo que las anomalías geográficas por sí solas no son suficientes para detectar ataques coordinados en este contexto. Para futuras mejoras, se recomienda optimizar los hiperparámetros de LightGBM, explorar técnicas de balanceo de clases (como sobremuestreo o submuestreo) y considerar la integración de más patrones de ataque para aumentar la precisión y reducir falsos positivos, mejorando así la aplicabilidad práctica del modelo en la detección de fraudes coordinados.