Universidad EAFIT

Maestría en Ciencia de Datos y Analítica

Aprendizaje Automático

S2261-0136

Aprendizaje No Supervisado Detección de Transacciones Fraudulentas

---

Autores:

Alejandro Barrientos-Osorio

Luis Miguel Caicedo-Jimenez

Omar Alejandro Henao-Zapata

26 de abril de 2022

I. INTRODUCCIÓN

Las transacciones fraudulentas en la industria bancaria se han convertido parte del paisaje diario que enfrentan los bancos. Estas transacciones fraudulentas son una forma ilegal de usar datos de la tarjeta de crédito sin el conocimiento del titular real de la tarjeta. Normalmente esto sucede cuando una tarjeta de crédito o su información confidencial es robada. Usualmente estas situaciones se resuelven cuando el titular de la tarjeta hace una reclamación a su banco. El banco o la compañía de la tarjeta de crédito realizan una investigación y devuelven el dinero al titular de la tarjeta.

Pero, si podemos detectar las transacciones fraudulentas en tiempo real, podemos tomar las medidas necesarias para detenerlas. Esto evitaría miles de reclamaciones, papeleos y perdida de recursos que terminan en manos ajenas.

Las transacciones fraudulentas son casos atípicos dentro del conjunto de transacciones diarias. Esto debido a que la mayoría de las transacciones realizadas en la industria son legales, por lo que las transacciones fraudulentas son muy pocas dentro del conjunto de transacciones diarias.

Para la detección de atípicos hay métodos de aprendizaje no supervisado que son útiles como:

(i) Isolation forest;

(ii) Local Outlier Factor (LOF);

A continuación se presentan estos algoritmos de aprendizaje no supervisado con un set de datos de transacciones bancarias.

II. METODOLOGÍA

La metodología que se usó fue CRISP-DM, conocida para afrontar problemas de desarrollo de analítica, sin embargo, el paso de implementación no fue llevado a cabo debido a que el propósito del trabajo es una comparación y estudio.

En la figura 1 se puede ver una representación gráfica de la metodología usada.

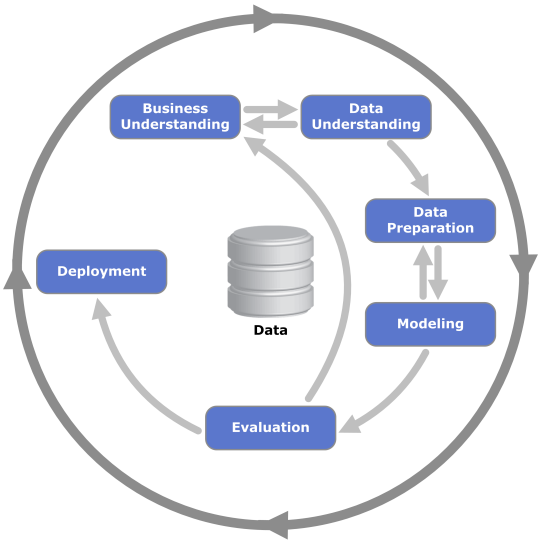

Figura 1: Metodología CRISP-DM

A. Entendimiento del Negocio

La prevención del fraude es la implementación de una estrategia para detectar transacciones fraudulentas o acciones bancarias y evitar que estas acciones causen daños financieros y de reputación al cliente y a la institución financiera (FI). A medida que los canales de banca en línea y móvil se vuelven más populares y las instituciones financieras continúan digitalizándose, una estrategia sólida de prevención de fraude solo será más importante [1].

La prevención del fraude y el cibercrimen están conectados y siempre cambian. A medida que los profesionales de prevención de fraude desarrollan nuevas soluciones de autenticación y detección de fraude, los estafadores se conectan entre sí, monetizan e intercambian información en la Dark Web [2]. Los estafadores de hoy utilizan estrategias sofisticadas y malware para tener éxito en sus actividades fraudulentas. Aunque tecnología de prevención de fraude ha hecho grandes avances y continúa haciéndolo, es importante tener en cuenta las tácticas fraudulentas y comprender cómo prevenir el fraude[3].

En Colombia, Asobancaria no tienen una cifra consolidada de cuánto dinero pierden las personas y las propias entidades cada año por cuenta del accionar de los ciberdelincuentes, pero sí advierten que por cada 100.000 pesos transados en el sistema financiero en general, 4,9 pesos fueron reclamaciones por fraude el año pasado, indicador que fue de 4,3 en el 2019 [4].

Solo en los canales digitales, ese indicador pasó, en el mismo periodo, de 2,7 a 3,5 pesos por

cada 100.000 pesos transados, precisó Gómez, quien señala que esos datos son positivos pues indican que las pérdidas no son muy elevadas teniendo en cuenta el volumen de operaciones y de recursos que se transan hoy por los canales digitales [4]. En el 2020 fueron cerca de COP$ 3.500 billones de pesos solo por internet, mientras que por la llamada banca móvil (celulares y dispositivos electrónicos) fueron más de COP$ 179,3 billones, según la Superfinanciera [4].

La entidad advierte que en el 2020 se evidenció un incremento del 10,7 % de fraude en canales digitales [4].

Y si bien Colombia ocupa el tercer lugar dentro del top 5 de fraudes y es el sexto de Latinoamérica en materia de ataques cibernéticos detectados, según el más reciente Informe Global de Fraude e Identidad 2021, elaborado por Datacrédito Experian, este no es un flagelo exclusivo del país [5].

Por ello, la tendencia emergente en la detección y prevención del fraude en este momento se centra en el aprendizaje automático. El aprendizaje automático es el uso de inteligencia artificial para mejorar un sistema sin estar específicamente programado para realizar estas mejoras. En el contexto de la prevención del fraude, hay dos tipos de aprendizaje automático: aprendizaje automático no supervisado y supervisado [3].

En este reporte se utilizaron dos métodos de aprendizaje no supervisado para la detección de transacciones fraudulentas: (i) Isolation forest [6]; y (ii) LOF [7].

B. Entendimiento de los datos

Los datos de transacciones bancarias estudiados en este reporte tienen 15 variables financieras las cuales no se especifican que son para mantener la confidencialidad de los clientes (Cuadro I). 14 de estas variables son números reales mientras que una es entero. El conteo de los datos no nulos muestra como en 10 variables hay entre 10 mil y 25 mil datos nulos.

La ultima variable *Class* es la clasificación de cada una de las transacciones. "0"para las transacciones legales y "1"para la transacciones fraudulentas. Revisando que tan balanceado está el set de datos, se obtiene que hay 139747 transacciones legales y 253 transacciones fraudulentas (Figura 2). Estos resultados indican que el set de datos no esta balanceado. Situación normal para transaccciones bancarias.

Ahora obteniendo los estadísticos básicos del set de datos se evidencia la multidimensionalidad de los datos (Figura 3).

N°	Column	Non-Null Count	Dtype
0	Timestamp	140000 non-null	float64
1	Value	140000 non-null	float64
2	C1	116232 non-null	float64
3	C2	129731 non-null	float64
4	C3	129693 non-null	float64
5	C4	140000 non-null	float64
6	C5	129678 non-null	float64
7	C6	116529 non-null	float64
8	C7	125595 non-null	float64
9	C8	129645 non-null	float64
10	C9	140000 non-null	float64
11	C10	129891 non-null	float64
12	C11	125695 non-null	float64
13	C12	125833 non-null	float64
14	Class	140000 non-null	int64

Cuadro I: Descripción general de los datos de transacciones bancarias extraídos de `https://www.projectpro.io/`

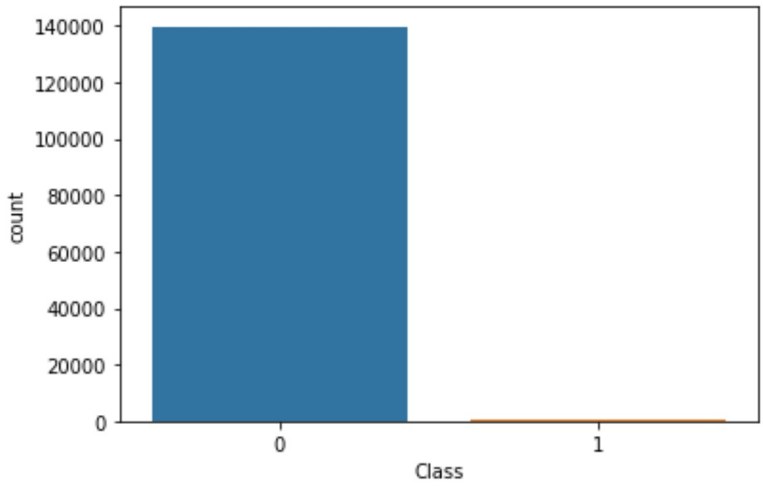

Figura 2: Clasificación en transacciones legales (0) y fraudulentas (1)

	Timestamp	Value	C1	C2	C3	C4	C5
count	140000.000000	140000.000000	116232.000000	129731.000000	129693.000000	140000.000000	129678.000000
mean	105334.592698	71.102883	-13.647954	-15.797094	-16.141105	0.000037	-15.540173
std	52763.641695	212.359700	224.599903	72.631165	74.585111	0.031540	72.120519
min	0.000000	0.000000	-2000.000000	-500.000000	-500.000000	-0.212540	-500.000000
25%	60107.500002	4.400000	-6.419730	-0.262951	-9.288367	-0.017731	-0.205542
50%	94276.111110	17.584000	-1.250014	-0.019520	-1.739613	-0.000140	0.000903
75%	154845.833375	61.522000	5.063801	0.220734	6.436411	0.017776	0.206203

Figura 3: Estadísticos básicos de las primeras 7 variables del set de datos estudiado

C. Preparación de los datos

Para poder procesar estos datos, se requiere que ninguna de las variables tenga datos nulos.

Para ello, se procede a calcular la mediana de cada variables y a reemplazar cada dato nulo con su mediana. Se utiliza la mediana porque es un estadístico mas robusto que la media (o promedio).

De esta manera se obtiene un set de datos sin valores nulos. Luego se revisan las posibles correlaciones que haya entre las variables y determinar si se puede reducir la dimensionalidad descartando una variable que dependa de las otras.

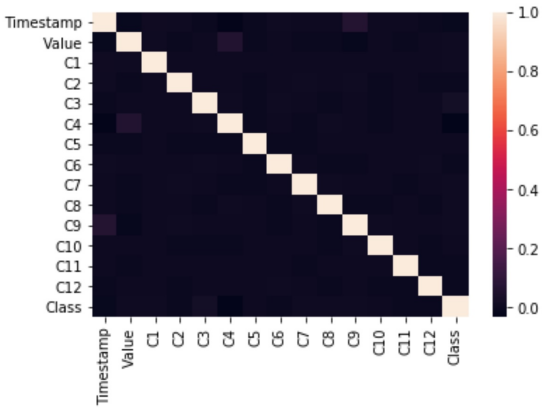

Figura 4: Correlacion entre las variables del data set estudiado

El resultado obtenido de las correlaciones entre las 15 variables estudiadas, es que son independientes. Tiene una correlación que tiende a cero. Por ello se toma todo el data set completo y no se puede descartar ninguna variable.

D. Modelado

(i) **Isolation forest:** Isolation forest (o bosque de aislamiento) es similar a Random Forest (o bosque aleatorio) que se crean con base en árboles de decisión. Este algoritmo hace parte del aprendizaje no supervisado porque no utiliza etiquetas predefinidas.

Isolation forest se construye con base en las anomalías (o valores atípicos; pocos y diferentes) de un data set. En un bosque de aislamiento, los datos submuestreados aleatoriamente se procesan en una estructura de árbol basada en características seleccionadas aleatoriamente.

Al adentrarnos en el árbol de decisión, es menos probable que la submuestra estudiada tenga una anomalía. Esto debido a que cada vez se requieren mas cortes para aislar las anomalías. Si una anomalía es identificada después de varias ramificaciones, es porque era un inlier (dato atípico interior difícil de aislar). Si las ramas para identificar un atípico son cortas indican anomalías fáciles de detectar (o outliers). Estas son relativamente mas fáciles de separarlas de otras (Figura 5).

Para la modelación de Isolation forest el parámetro mas importante que define la cantidad de atípicos que hay en el data set es el nivel de contaminación. Sin embargo, del nivel de contaminación solo se puede aproximar al valor del sector estudiado. Para el caso de las transacciones bancarias es $< 1\,\%$. Es muy posible que ese valor de atípicos asociado al set de datos estudiado sea muy diferente.

El nivel de contaminación (C) se calcula de la siguiente manera:

$C = T_{valida}/T_{fraudulenta}$.

donde, T_{valida} son el numero total de transacciones validas; $T_{fraudulenta}$ son el numero total de transacciones fraudulentas.

El resultado obtenido para la contaminación es igual a $0.18\,\%$.

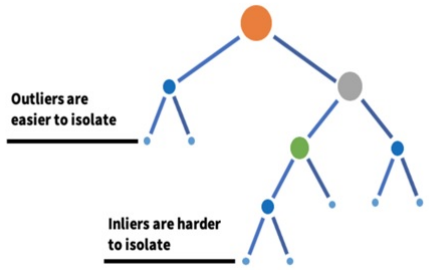

Figura 5: Isolation Forest: Forma operativa

(ii) **LOF:** Local Outlier factor (LOF) es un algoritmo de aprendizaje automático no supervisado que identifica valores atípicos con respecto a los vecindarios locales, en lugar de utilizar la distribución asociada a los datos.

LOF se basa en un concepto de densidad local, que es calculada con base en las distancias de un punto A a sus k vecinos más cercanos. Al comparar la densidad del punto A con las densidades de sus vecinos, se pueden identificar regiones de densidad similar y puntos que tienen una densidad sustancialmente menor que sus vecinos. Estos últimos, con densidad menor, se consideran valores atípicos.

La densidad local se estima utilizando la "distancia de accesibilidad"(Figura 6). Esta se define como el máximo entre de dos distancias: (i) la distancia entre dos puntos A y B; y (ii) la k-distancia(A) (distancia del objeto A a su k-ésimo vecino más cercano). Para los puntos dentro de una aglomeración (o cluster) se considera la distancia k. Mientras que para los puntos fuera del cluster, se considera la distancia entre puntos.

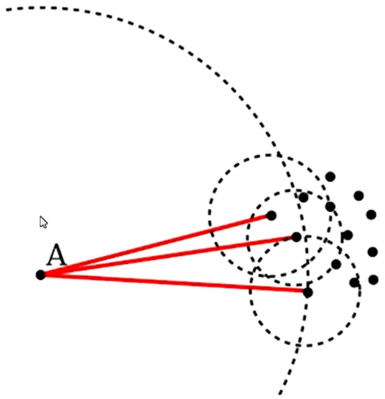

Figura 6: LOF: distancia de accesibilidad

Aquí se calculan las distancias de accesibilidad a todos los k vecinos más cercanos de un punto para determinar la densidad de accesibilidad local (LRD en ingles) de ese punto. LRD es una medida de la densidad de k-puntos más cercanos alrededor de un punto que se calcula dividiendo 1 sobre la suma de todas las distancias de accesibilidad de todos los k-puntos vecinos más cercanos. Por lo tanto, cuanto más cerca están los puntos, la distancia es menor y la densidad es mayor.

El cálculo de LOF se realiza tomando la relación entre el promedio de los LRDs de k número de vecinos de un punto y el LRD de ese punto.

Analisis del valor de LOF:

(i) Si LOF < 1, entonces el punto está dentro del grupo de densidad: Esto sucede si la densidad de los vecinos es menor que la densidad del punto;

(ii) Si LOF $== 1$, entonces el punto es muy similar a sus vecinos: Esto sucede cuando la densidad de los vecinos y el punto en evaluación son casi iguales;

(iii) Si LOF > 1 , valor atípico: Esto sucede si la densidad de los vecinos es mayor que la densidad del punto.

E. Evaluación y Análisis

1. IF

El modelo se inicializa con con 500 estimadores por defecto y una contaminación definida. Esto permite decirle al modelo los parámetros a tener en cuenta en la detección de los outliers. Además, se alimenta la totalidad del dataset al modelo.

Luego de correr el modelo con estos parámetros, es posible observar una asimetría negativa (Figura 7) en los puntajes otorgados a cada una de las observaciones, con lo cual se puede confirmar que el algoritmo es capaz de identificar una lista de datos atípicos.

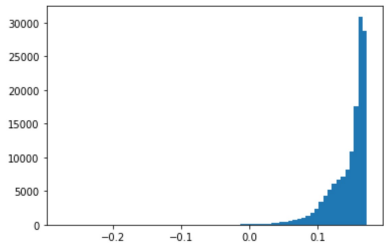

Figura 7: Asimetría a negativa.

Al obtener los datos anómalos, se observa que el algoritmo identificó 280 valores atípicos, cuando el total del dataset son 253. Esto se explica dada la naturaleza no supervisada del algoritmo, lo que ocasiona que puedan obtenerse falsos negativos (Figura 8).

Con el fin de mejorar el desempeño del modelo, se aplica una división entre set de entrenamiento y otro de testeo. Se ajusta el modelo con los mismos parámetros, pero en este caso aplicado únicamente a los datos de entrenamiento. En este caso, se obtiene un total de 257 anomalías, con lo cual la precisión se acerca mucho más al número real encontrado en el dataset.

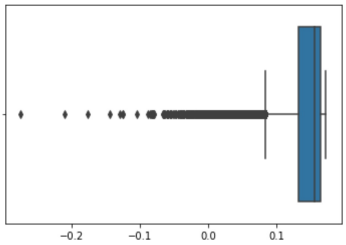

Figura 8: Datos atípicos a las izquierda.

2. LOF

Para la implementación, se define una vecindad de tamaño 30, y el mismo porcentaje de contaminación utilizado en el método de IF. En este caso, se tiene que el número total de anomalías detectadas fue 252, demostrando un mejor desempeño que el modelo anterior (Figura 9).

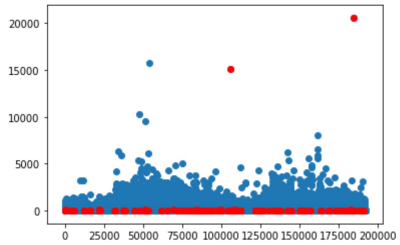

Figura 9: Datos atípicos identificados con LOF (color rojo)

3. Análisis de Sensibilidad

A continuación, se realiza una comparación entre los resultados obtenidos con los parámetros anteriores, y aquellos obtenidos al variar los parámetros, con el fin de determinar el impacto que tiene cada una de las especificaciones en el desempeño del modelo.

Para esto, se ejecuta el algoritmo LOF con un tamaño de vecindad igual a 20 y el cálculo automático de la contaminación, la cual tiene en cuenta un peso de -1.5 para la estimación de las anomalías. Adicional a esto, se determina el umbral para la selección de datos atípicos a ser el 2 %. De esta manera, los valores atípicos son aquellos cuyo score sea menor que este umbral. Así, se tiene un total de 2800 datos atípicos (Figura 10). Por tanto, se concluye que el porcentaje de contaminación es un parámetro crítico para el modelo, y debe realizarse una

optimización de hiperparámetros para determinar la que mejor seleccione las anomalías.

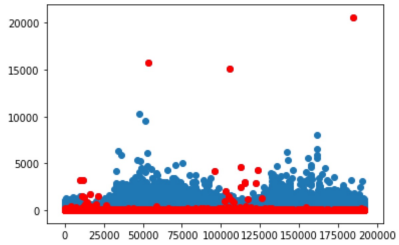

Figura 10: Aumento de datos atípicos debido al umbral seleccionado

III. CONCLUSIONES

- Isolation Forest y LOF son algoritmos muy útiles para detección de atípicos en un set de datos donde las variables son independientes.
- LOF es mejor detectando el numero de atípicos mas cercano a la realidad debido a que se basa en distancias entre los vecinos mas cercanos y la densidad local.
- Ambos algoritmos se pueden trabajar utilizando un score de atípicos. Sin embargo este score de atípicos es influenciado por el % de contaminación. Si se define un error de contaminación alejado de la realidad del sector estudiado, lo mas probable es muchos datos sean seleccionados como atípicos que realmente no lo son. Es sumamente importante conocer muy bien el sector estudiado para definir el % de cotaminación lo mas cercano posible a la realidad.

[1] A. Shabbir, M. Shabir, A. R. Javed, C. Chakraborty, and M. Rizwan, "Suspicious transaction detection in banking cyber–physical systems," *Computers and Electrical Engineering*, vol. 97, no. June 2021, p. 107596, 2022. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2021.107596

[2] A. Kumar, R. D. Gopal, R. Shankar, and K. H. Tan, "Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering," *Decision Support Systems*, vol. 155, no. June 2021, p. 113728, 2022. [Online]. Available: https://doi.org/10.1016/j.dss.2021.113728

[3] J. Domashova and O. Zabelina, "Detection of fraudulent transactions using SAS Viya machine learning algorithms," *Procedia Computer Science*, vol. 190, no. 2020, pp. 204–209, 2021. [Online]. Available: https://doi.org/10.1016/j.procs.2021.06.025

[4] "Cada hora se presentan 5 denuncias por fraude financiero," *El TIEMPO*, 2021. [Online]. Available: https://www.eltiempo.com/economia/sector-financiero/fraudes-financieros-en-aumento-con-la-pandemia-614010

[5] T. T. C. Nguyen, T. H. P. Nguyen, T. B. T. Nguyen, S. K. Selvarajan, and A. Baskaran, "The impact of opportunity factors on fraudulent behavior in the Vietnamese stock market," *Journal of Asian Economics*, vol. 79, no. November 2021, p. 101451, 2022. [Online]. Available: https://doi.org/10.1016/j.asieco.2022.101451

[6] "sklearn.ensemble.IsolationForest," *Scikitlearn*, 2022. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

[7] "sklearn.neighbors.LocalOutlierFactor," *Scikitlearn*, 2022. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html