Detección de anomalías en e-commerce utilizando aprendizaje automático y profundo

FERNANDEZ DEL VALLE PANIAGUA Santiago

Universidad Autónoma de Nuevo León

santiago.fernandezp@uanl.edu.mx

1. Introduccion

El tema de la detección de anomalías no es un tema nuevo, sin embargo su desarrollo en el área digital aun está lejos de terminar. El objetivo principal es poder identificar comportamientos atípicos en el comportamiento del sujeto analizado. La tarea de localizar datos anómalos puede ayudar en la seguridad contra ataques cibernéticos, detección de fraudes, forja de seguros o inclusive diagnósticos médicos. Según Market Analysis Report (2023), el 2022 se invirtió un monto de 4.3 mil millones de dólares para el desarrollo de herramientas que permitan localizar anomalías. En el presente trabajo se pretende analizar un año de transacciones de un E-commerce real de tamaño pequeño. Se pretende analizar y comparar diversos tipos de algoritmos de aprendizaje automático no supervisados con el objetivo de explorar los puntos fuertes y débiles de los métodos utilizados. De igual manera se discute cómo evaluar la presición, ya que los datos no cuentan con una etiqueta que nos permita saber si son anómalos o no.

2. Datos y Métodos

2.1 Data

Los datos utilizados son las transacciones de un E-Commerce local al cuál se denominara como el comercio X. Se cuenta con 9462 transacciones únicas. La fuente cuenta con una totalidad de 78 parámetros algunos de los cuales son nombre, estatus de la transaccion, el día de pago, si se completó la entrega del pedido, cuándo se completó, los artículos ordenados, el precio, la dirección de entrega, datos de la persona a quien se entrega y de quien realiza el pedido, si se canceló el pedido y la fecha de cancelación.

2.2 Metodología

Se realizó un análisis principal sobre la base de datos con el fin de realizar limpieza de datos categóricos, como errores de ortografía; eliminar aquellas características que no contribuyeran información para el análisis y crear nuevas características de importancia. Se utilizaron las distintas fechas proporcionadas en la fuente para crear los campos de tiempo entre un evento y otro. Se utilizaron los SKU de los artículos ordenados para crear un parámetro que enlistara los artículos ordenados en la transacción.

2.3 Métodos utilizados

- K-Prototypes (39 variables)
- K-Prototypes (42 variables)
- K-means
- DBScan
- Isolation Forest

3. Resultados

3.1 K-Prototypes

Cluster analysis (CA) is an unsupervised classification method of a set of objects in groups. The HCA is useful for variables or observations, categorical data, or continuous variables. Principal component analysis (PCA) is another form of unsupervised classification method applying the rotation of covariance matrix to reduce the dimension of the dataset by retaining most of the information in the original set (Chen et al., 2007; Fitzpatrick et al., 2007; Irawan et al., 2009; King et al., 2014; Ritzi et al., 1993; Seyhan et al., 1985).

3.2 K-Prototype sin pre-procesamiento

Para este caso se utilizaron los datos sin procesamiento alguno. El único cambio realizado fue la exclusión de características irrelevantes. Podemos ver en la figura 1 que muy pocas entradas se marcan como anomalías. Analizando más a detalle, los datos marcados como anómalos consisten de transacciones cuyo precio es mucho más elevado a los precios de todas las otras transacciones.

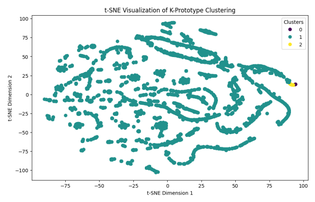
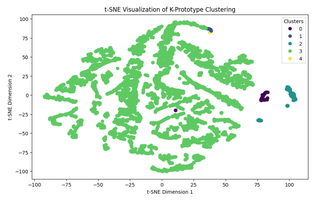


Figura 1: *Clasificación usando k-prototypes sin características creadas.*

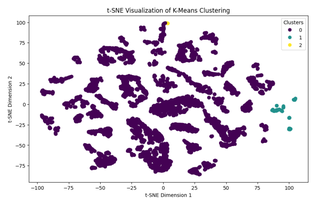
3.3 K-Prototype con variables de tiempos entre accciones.

De igual manera se utiliza el algoritmo de k-prototypes al construir nuevas variables numéricas con la información de los tiempos en las transacciones. Se obtienen los resultados mostrados en la figura 2. En este los grupos anómalos contienen transacciones que fueron canceladas, anuladas, no pagadas y a aquellas que se les hizo una debolución de dinero.



3.4 K-means

Se convierten todas las variables categóricas a numéricas utilizando embeding de frecuencias y se utiliza el algoritmo de k-means, los resultados se muestran en la figura 3. En este los grupos anómalos contienen transacciones que fueron canceladas, anuladas, no pagadas y a aquellas que se les hizo una debolución de dinero.



3.5 DBScan

Se utiliza un método similar al de codo para encontrar el número óptimo de clusters para encontrar el hiperparámetro adecuado. Se obtienen los resultados mostrados en la figura 5. En este caso los grupos anómalos contienen transacciones que fueron canceladas, anuladas, no pagadas y a aquellas que se les hizo una debolución de dinero. Sin embargo, también encuentra un grupo de transacciones que no tienen información del usuario que hace la compra ni de quien la recibe.

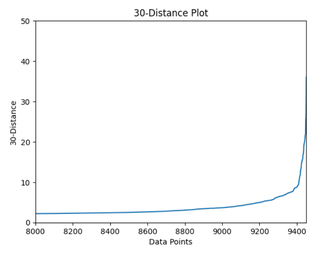


Figura 4: *distance plot.*

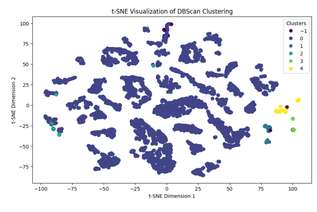


Figura 5: *DBScan.*

3.6 Isolation Forest.

Para este caso se utilizan 150 estimadores y una contaminación del 0.05. Se obtienen los resultados mostrados en la figura 6. En este los grupos anómalos contienen transacciones que fueron canceladas, anuladas, no pagadas y a aquellas que se les hizo una debolución de dinero. Sin embargo este método resultó ser el que más datos regulares incluye en el agrupamiento de datos atípicos.

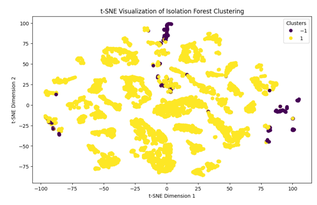


Figura 6: *Isolation Forest.*

4. Concluciones

Podemos concluir que los métodos utilizados hasta ahora son eficientes en encontrar datos anómalos. Inclusive el primer k-prototypes sin procesamiento alguno encontró que un cliente manejaba la mayor parte de las ganancias para esta pequeña empresa. Al procesar un poco los datos se observó que todos los métodos utilizados tenían la misma tendencia de predicción de transacciones atípicas. Se puede ver la potencia del uso de DBScan, con afinar los hiperparámetros este método ha sido aquel que más información de transacciónes atípicas ha logrado distinguir. Isolation Forest logró distinguir parte de los datos anómalos, sin embargo la limitación principal de este algoritmo es que solo clasifica en 2 grupos e intenta excluir todo dato que ve como anómalo y es probable que haya requerido bajar la dimensionalidad de la fuente para tener un mejor resultado.

5. References

1. Market Analysis Report, 2023, https://www.grandviewresearch.com/inc analysis/anomaly-detection-market-report

2. Global Industry Analysts, Inc, 2022, https://www.prnewswire.com/news releases/valued-to-be-8-6-billion-by-2026–anomaly-detection-slated-for-robust-growth-worldwide-301507777.html

3. AWS, 2023, “What is anomaly detection?”, https://aws.amazon.com/wh is/anomaly-detection/

4. NVIDIA 2023, “Applications of AI for Anomaly Detection”, https://www.nvidia.com/en-us/training/instructor-led-workshops/anomaly-detection/