Proyecto de desarrollo de un MVP para detección de fraudes

Matias Lucero
Analista de BI

Adrian Szklar
Data Analyst

Manuel Ruiz M
Machine Learning Developer

- Librerías de Python que Utilizamos

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Análisis Exploratorio de Datos (EDA)

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.0	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.0	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.0	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.0	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.0	0.00	0	0
5	1	PAYMENT	7817.71	C90045638	53860.00	46042.29	M573487274	0.0	0.00	0	0
6	1	PAYMENT	7107.77	C154988899	183195.00	176087.23	M408069119	0.0	0.00	0	0
7	1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0.0	0.00	0	0
8	1	PAYMENT	4024.36	C1265012928	2671.00	0.00	M1176932104	0.0	0.00	0	0
9	1	DEBIT	5337.77	C712410124	41720.00	36382.23	C195600860	41898.0	40348.79	0	0

Primeras 10 filas, el dataset contiene 6,5 millones de registros

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

- Obtenemos información de nuestro dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  ------          -----
 0   step            int64
 1   type            object
 2   amount          float64
 3   nameOrig        object
 4   oldbalanceOrg   float64
 5   newbalanceOrig  float64
 6   nameDest        object
 7   oldbalanceDest  float64
 8   newbalanceDest  float64
 9   isFraud         int64
 10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 534.0+ MB
```

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Limpieza de Datos

- Detectamos si existen duplicados y valores nulos

- Eliminamos la columna “isFlaggedFraud”

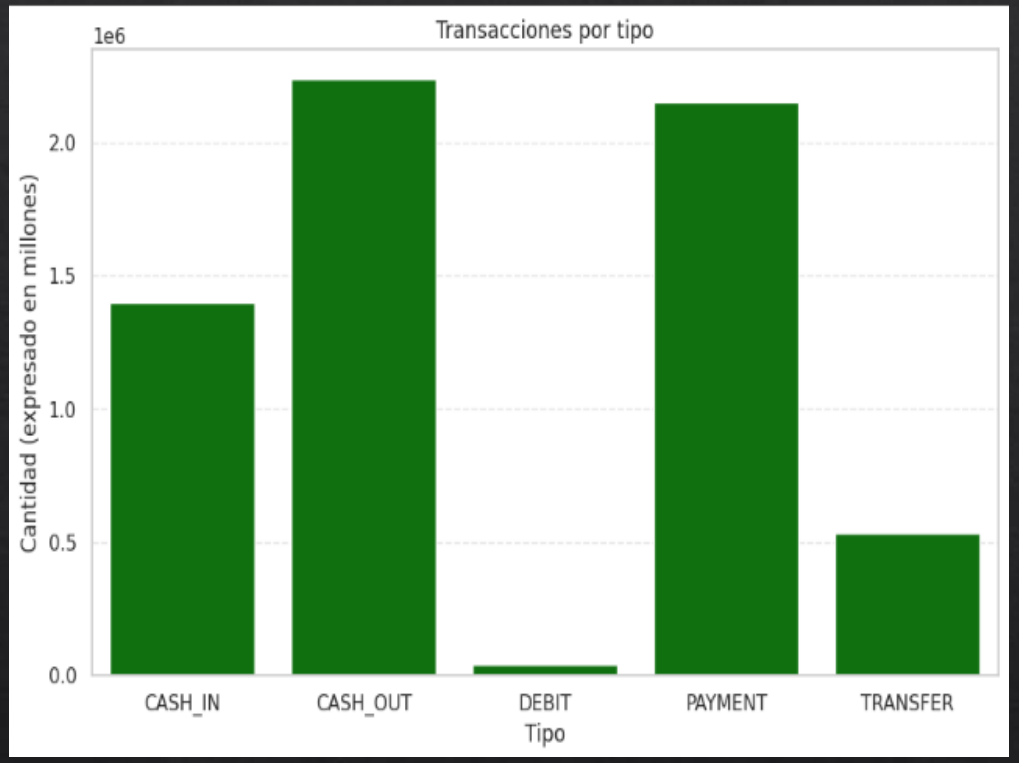
Observación: En este caso el dataframe no presentaba ni datos duplicados y tampoco valores nulos.

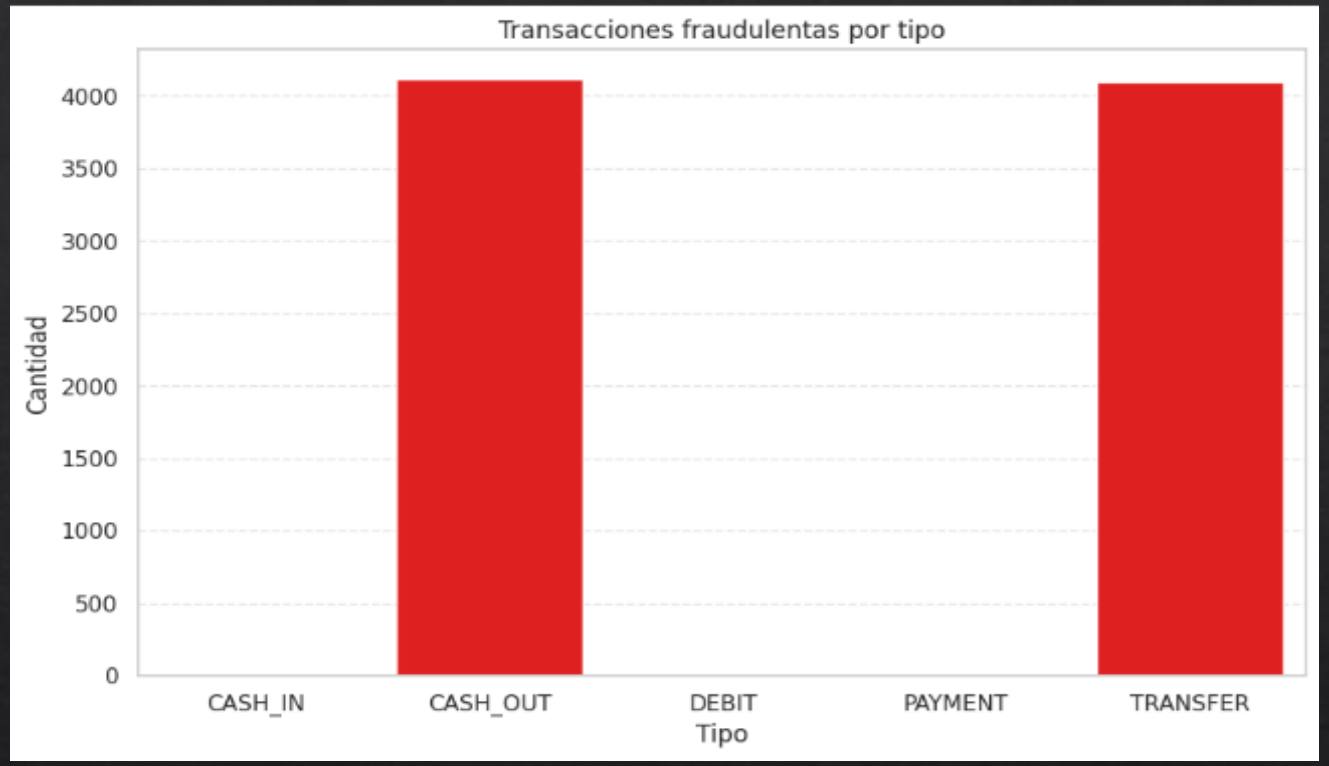
Transformación de Datos

- La columna “step” mostraba las hora mensual cuando se hizo la transacción (la primera hora del mes ***step = 1*** y la última ***step =743***) entonces agregamos columnas para resolver este problema y así poder convertir por ejemplo: step 1 = 2020/01/01 01:00:00.

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	hour	day
0	2020-01-01 01:00:00	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	False	1	1
1	2020-01-01 01:00:00	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	False	1	1
2	2020-01-01 01:00:00	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	True	1	1
3	2020-01-01 01:00:00	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.00	0.00	True	1	1
4	2020-01-01 01:00:00	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	False	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
6362615	2020-01-31 23:00:00	CASH_OUT	339682.13	C786484425	339682.13	0.00	C776919290	0.00	339682.13	True	23	3
6362616	2020-01-31 23:00:00	TRANSFER	6311409.28	C1529008245	6311409.28	0.00	C1881841831	0.00	0.00	True	23	3
6362617	2020-01-31 23:00:00	CASH_OUT	6311409.28	C1162922333	6311409.28	0.00	C1365125890	68488.84	6379898.11	True	23	3

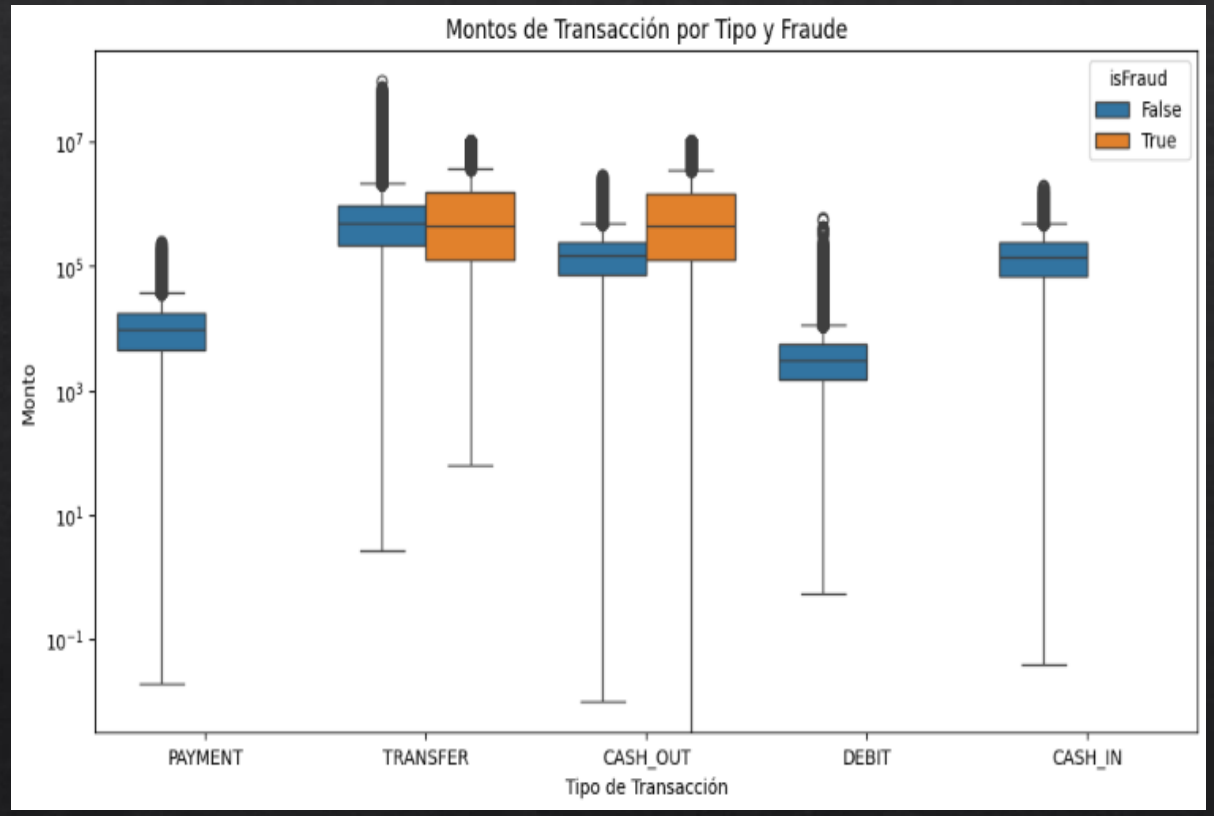
• Visualización de Datos

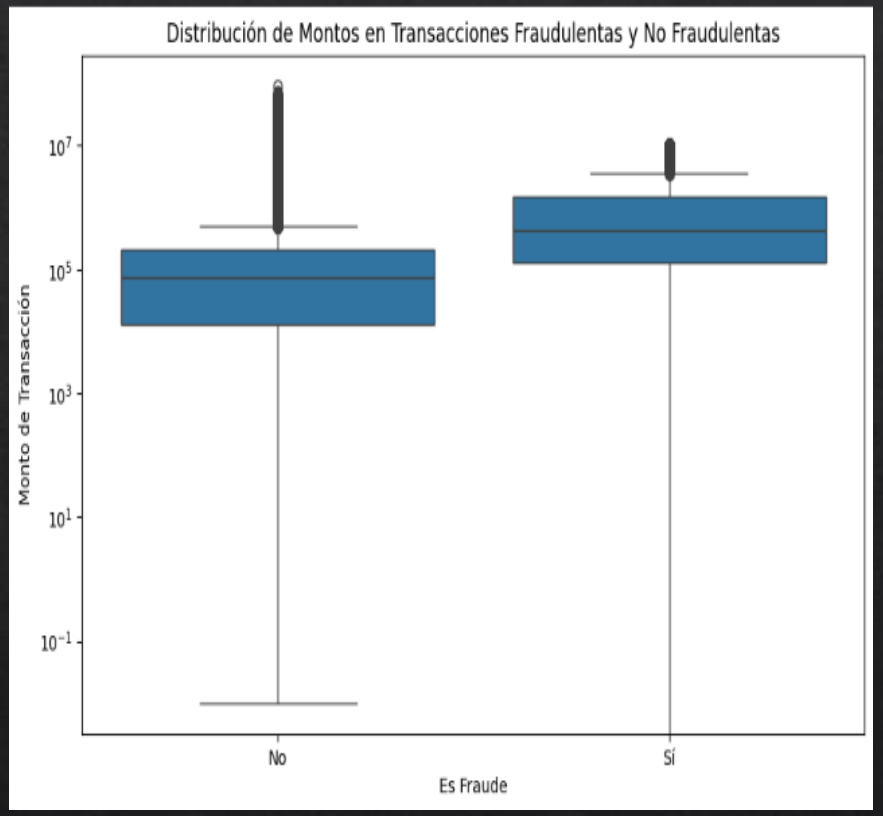




Matias Lucero , Adrian Szklar, Manuel Ruiz M.

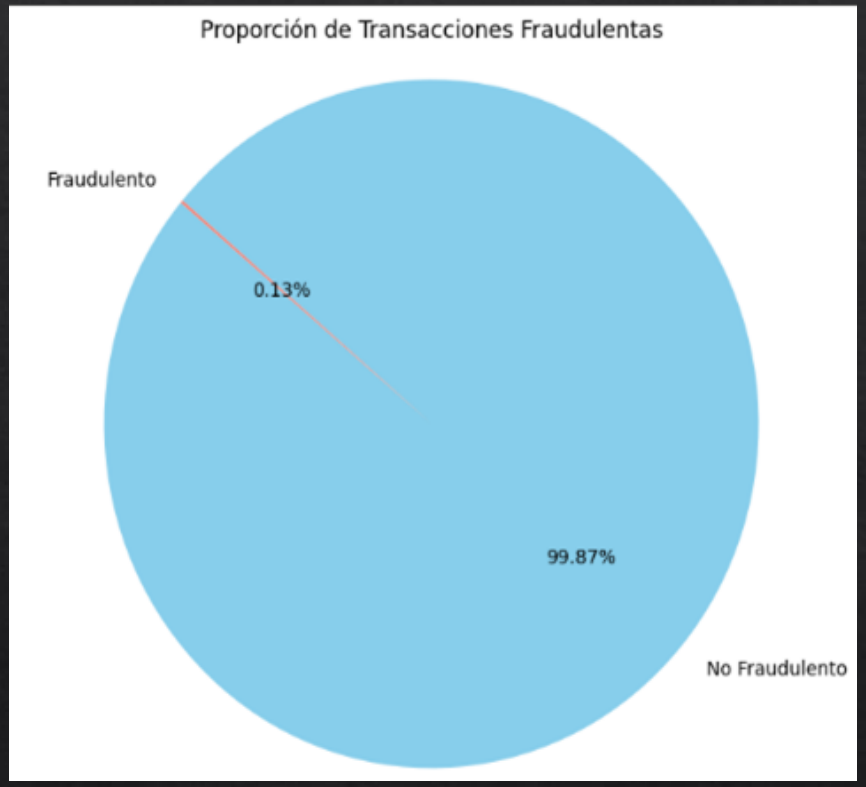
• Visualización de Datos

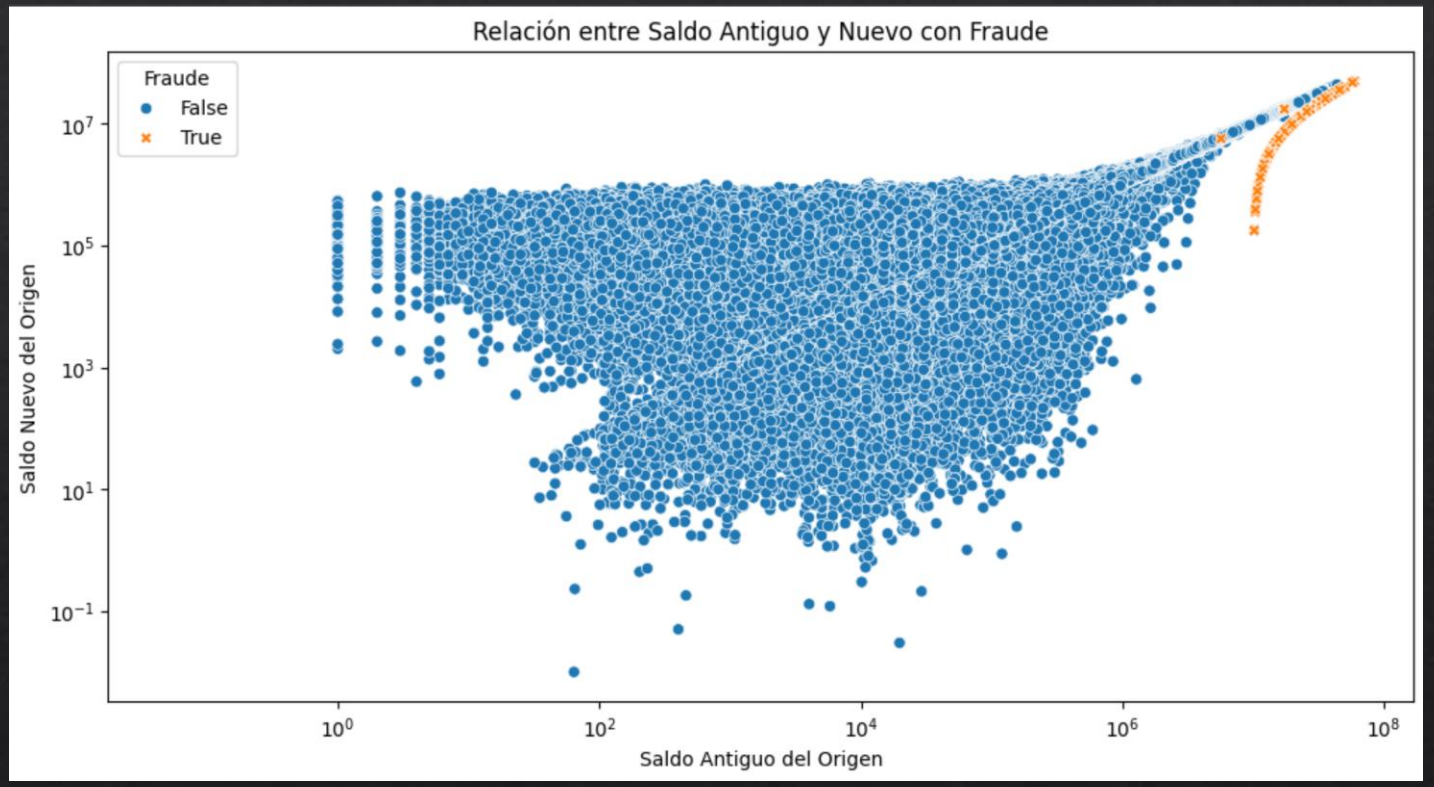




Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Visualización de Datos





Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Resumen y Análisis Estadístico de los Datos

```
     step           type               amount          nameOrig          oldbalanceOrg      newbalanceOrig
Min.   :  1.0   Length:6362620     Min.   :       0   Length:6362620     Min.   :       0   Min.   :       0
1st Qu.:156.0   Class :character   1st Qu.:   13390   Class :character   1st Qu.:       0   1st Qu.:       0
Median :239.0   Mode  :character   Median :   74872   Mode  :character   Median :   14208   Median :       0
Mean   :243.4                      Mean   :  179862                      Mean   :  833883   Mean   :  855114
3rd Qu.:335.0                      3rd Qu.:  208721                      3rd Qu.:  107315   3rd Qu.:  144258
Max.   :743.0                      Max.   :92445517                      Max.   :59585040   Max.   :49585040
  nameDest         oldbalanceDest     newbalanceDest         isFraud
Length:6362620     Min.   :        0   Min.   :        0   Min.   :0.000000
Class :character   1st Qu.:        0   1st Qu.:        0   1st Qu.:0.000000
Mode  :character   Median :   132706   Median :   214661   Median :0.000000
                   Mean   :  1100702   Mean   :  1224996   Mean   :0.001291
                   3rd Qu.:   943037   3rd Qu.:  1111909   3rd Qu.:0.000000
                   Max.   :356015889   Max.   :356179279   Max.   :1.000000
```

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Resumen y Análisis Estadístico de los Datos

Basado en las estadísticas proporcionadas, podemos realizar el siguiente análisis inicial del dataset:

Tamaño del Dataset:

- El dataset contiene 6,362,620 filas (observaciones) y 8 columnas (variables).

Tipos de Datos:

- Las columnas step, amount, nameOrig, oldbalanceOrg, newbalanceOrg, nameDest, oldbalanceDest, y newbalanceDest son de tipo numérico, (aunque nameOrig y nameDest parecen contener valores de texto que podrían representar nombres u otras identificaciones).
- La columna isFraud es de tipo binario (0 representa transacciones no fraudulentas y 1 representa transacciones fraudulentas).

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Resumen y Análisis Estadístico de los Datos

Distribución de los Datos:

- Existen valores mínimos y máximos para todas las columnas numéricas.
- Se proporcionan cuartiles (percentiles 25, 50 y 75) para todas las columnas numéricas, lo que permite observar la distribución de los datos. Por ejemplo, el primer cuartil para la columna amount es 156, lo que significa que el 25% de los valores son menores a este valor.
- La presencia de valores máximos muy altos en algunas columnas como amount, oldbalanceOrg, newbalanceOrg, oldbalanceDest, y newbalanceDest podría indicar la existencia de outliers (valores atípicos).

Conclusiones Preliminares:

- El dataset parece ser adecuado para el desarrollo de un modelo de detección de fraudes, ya que contiene información sobre transacciones (monto, origen, destino, saldos) y una etiqueta que indica si una transacción es fraudulenta o no.

Matias Lucero , Adrian Szklar,Manuel Ruiz M.

Machine Learning

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Introducción al Machine Learning

Uno de los principales riesgos a los que están sometidas las entidades financieras son los ataques de fraudes electrónicos. Billones de dólares en pérdidas son absorbidas cada año por las entidades financieras debido a transacciones fraudulentas.

Se plantea un modelo que considera los principales retos en el diseño de un sistema de detección de fraudes, El dataset con el cual planteamos el modelo fue obtenido de Kaggle ,ante la falta de información pública acerca del tema, es un dataset fuertemente desbalanceado ,como se expone en el EDA, dicha cuestión la abordamos en la configuración de Hiperparametros del modelo de Machine Learning, para el cual elegimos por su versatilidad el algoritmo de Random Forest.

Matías Lucero , Adrián Szklar, Manuel Ruiz M.

Ingeniería de Características

Optuna, Optimización de Hiperparametros

Construir Modelo

Predicciones

Métricas del Modelo

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Ingeniería de Características

Un Bosque Aleatorio (Random Forest) es un algoritmo de aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y la robustez de las predicciones.
Se basa en la idea de que un conjunto de árboles diversificados puede ofrecer mejores resultados que un solo árbol.

Cuando enviamos datos a cualquier modelo de aprendizaje automático (ML), debemos hacerlo en el formato adecuado, ya que los algoritmos solo entienden números.

En este enfoque, a cada etiqueta se le asigna un número entero único según el orden alfabético. implementamos esto usando la biblioteca Scikit-learn.

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Optuna

Optuna es una biblioteca de Python para la optimización de hiperparámetros, Permite automatizar la búsqueda de la mejor configuración de un modelo de aprendizaje automático, evaluando diferentes valores de los hiperparámetros y seleccionando la combinación que optimiza un criterio específico.

Entre las ventajas de usar Optuna podemos mencionar brevemente:

- Mejora del rendimiento del modelo.
- Ahorro de tiempo.
- Eficiencia.
- Escalabilidad.
- Facilidad de uso.
- Independiente de la plataforma.

Ingeniería de Características

Label Encoding :

```
le = LabelEncoder()
df_fraude_sin_nulos['type'] =
le.fit_transform(df_fraude_sin_nulos['type'])
df_fraude_sin_nulos['nameOrig'] =
le.fit_transform(df_fraude_sin_nulos['nameOrig'])
df_fraude_sin_nulos['nameDest'] =
le.fit_transform(df_fraude_sin_nulos['nameDest'])
```

Optuna, optimización de Hiperparametros

• Best Parameters :

```
{'criterion': 'entropy', 'n_estimators': 23,
'max_depth': 18, 'min_samples_split': 7,
 'min_samples_leaf': 8, 'max_features': 16}
```

Construir Modelo

```
model =
RandomForestClassifier(n_estimators=23,max_depth=18,
                       min_samples_split=7,
                       min_samples_leaf=8,random_state=42,
                       criterion='entropy',class_weight = 'balanced',
                       max_features = 16, n_jobs=-1)
model.fit(X_train, y_train)
```

Métricas del Modelo

Clases	No Fraude	Fraude
Precisión =	99.98	92,76
F1-score =	99,99	89,14
Recall =	99,99	85,79

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Predicciones :

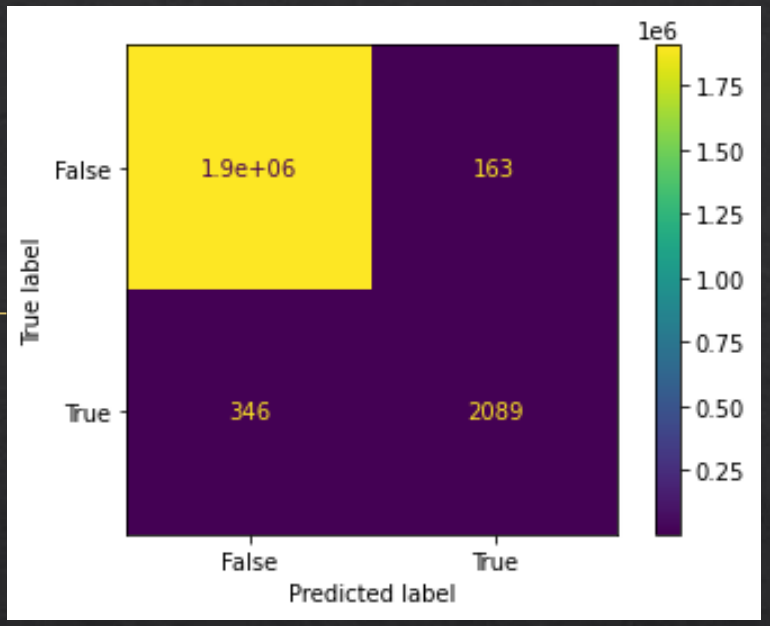
Umbral de clasificación en Random Forest

En un modelo de Random Forest, cada árbol individual genera una probabilidad de que una instancia pertenezca a una clase específica. El **umbral de clasificación** es un valor que se utiliza para convertir estas probabilidades en predicciones de clase.

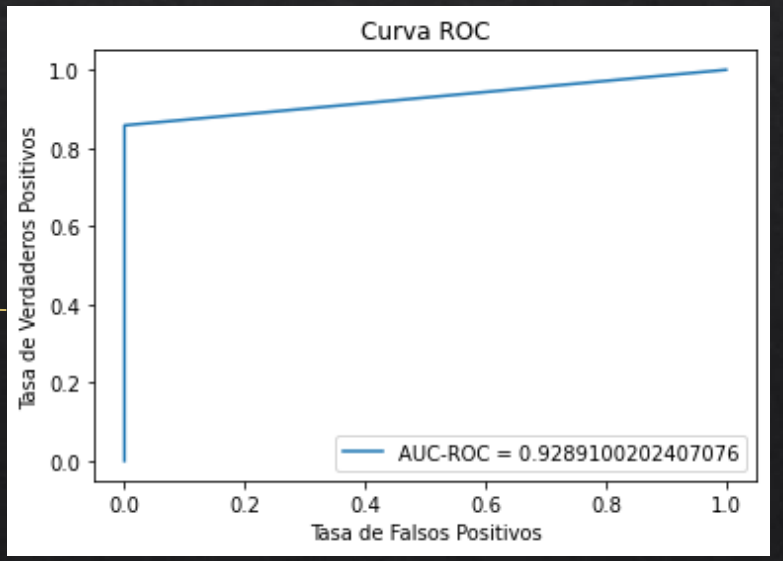
Lo que estamos buscando es dónde fijar la probabilidad de nuestro modelo, el denominado “umbral de probabilidad” (probability threshold) para encontrar el punto medio más óptimo, que nos permita clasificar con una determinada certeza cuales de nuestras transacciones son o no fraude.

Para hacer esto utilizamos **Optuna** ,lo cual nos permite optimizar una función de acuerdo a sus valores.

Matriz de Confusión



Curva Roc



Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Interpretación de las métricas específicas para el desbalanceo.

Precisión por clase:

• **La precisión para la clase "legítima"** es del 99.98%. Esto significa que el 99.98% de las transacciones que el modelo clasificó como legítimas eran realmente legítimas.

• **La precisión para la clase "fraude"** es del 92.76%. Esto significa que el 92.76% de las transacciones que el modelo clasificó como fraude eran realmente fraude.

Recall por clase:

• **La Recall para la clase "legítima"** es del 99.99%. Esto significa que el modelo identificó correctamente el 99.99% de las transacciones legítimas.
• **La Recall para la clase "fraude"** es del 85.79%. Esto significa que el modelo identificó correctamente el 85.79% de los casos de fraude.

F1-score por clase:

• **El F1-score para la clase "legítima"** es del 99.99%. Esto indica que el modelo tiene un buen rendimiento para identificar las transacciones legítimas.
• **El F1-score para la clase "fraude"** es del 89.14%. Esto indica que el modelo tiene un buen rendimiento para identificar los casos de fraude, aunque un poco menor que para la clase "legítima".

Matias Lucero , Adrian Szklar, Manuel Ruiz M.

Conclusión :

- El modelo tiene un **alto rendimiento** para identificar tanto las transacciones legítimas como los casos de fraude.
- El rendimiento es ligeramente mejor para la clase "legítima" que para la clase "fraude", lo cual es esperable en un contexto de desbalanceo.

Enlaces de Interés :

https://optuna.readthedocs.io/en/stable/index.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://github.com/No-Country/c16-93-ft-data-bi

Matias Lucero , Adrian Szklar, Manuel Ruiz M.