Presentacion Proyecto Final Data Science

Análisis de Detección de Fraude para una Empresa Financiera

Equipo de Trabajo: - Ertola Urtubay, Galo
- Gasaro, Emiliano

Tutor: Surijón, Ernesto
Profesor: Magaña Fuentes, Miguel Angel

Objetivos

Elegimos este Dataset porque nos interesó la idea de poder colaborar en un proyecto para analizar posibles casos de Fraude. Además los registros, más de 6 millones, con sus variables, son una buena fuente de información para poder manipularla y trabajarla, representando al mismo tiempo un gran desafio a la hora de trabajar. También es llamativo poder tener resultados positivos que ayuden a evitar que las personas pierdan su dinero o este sea robado.

- **R**ealizar un análisis EDA que nos permita identificar los datos más relevantes del dataset para alimentar de forma eficiente el modelo de ML a utilizar.
- **U**na vez realizado el EDA, alimentaremos el modelo con los datos modificados para comenzar a obtener resultados y usar estos como una retroalimentación para generar modelos más robustos para la detección de transacciones fraudulentas.
- **C**on los resultados obtenidos, se ideará un plan de acción para prevenir transacciones fraudulentas y así mejorar la seguridad de la empresa financiera.

Link Kaggle: https://www.kaggle.com/datasets/miznaaroob/fraudulent-transactions-data

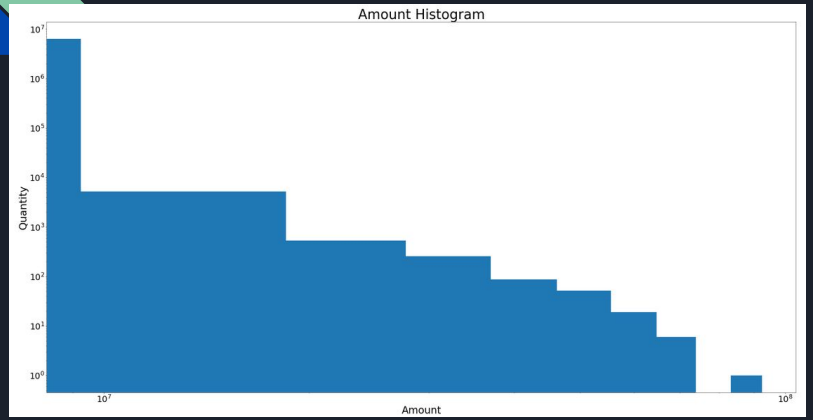
EDA

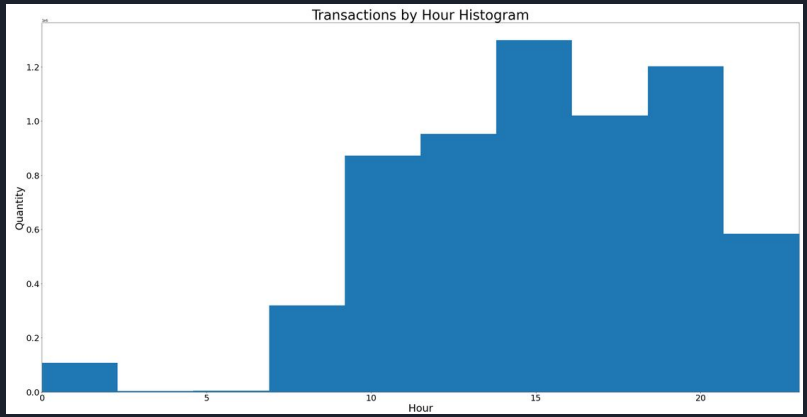
Luego de una primera exploración de datos se decidió que las columnas “isFlaggedFraud”, “nameOrig” y “nameDest” serían eliminadas de la tabla ya que la información que brindaban no eran de ayuda para alimentar a los distintos modelos que se van a analizar en el presente trabajo. Se agruparon la variable step por hora del día y se renombró la columna de type por hour.

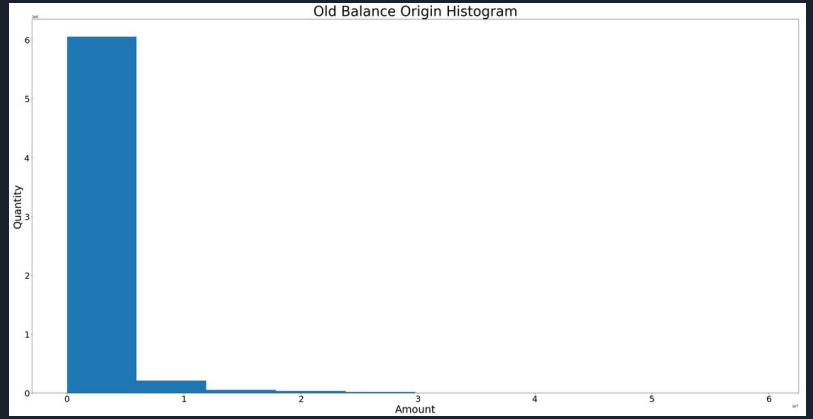
	CASH_IN	CASH_OUT	DEBIT	PAYMENT	TRANSFER	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
0	0	0	0	1	0	1	9839.64	170136.00	160296.36	0.00	0.00	0
1	0	0	0	1	0	1	1864.28	21249.00	19384.72	0.00	0.00	0
2	0	0	0	0	1	1	181.00	181.00	0.00	0.00	0.00	1
3	0	1	0	0	0	1	181.00	181.00	0.00	21182.00	0.00	1
4	0	0	0	1	0	1	11668.14	41554.00	29885.86	0.00	0.00	0
...	...	...	...	...	...	...	...	...	...	...	...	...
6362615	0	1	0	0	0	743	339682.13	339682.13	0.00	0.00	339682.13	1
6362616	0	0	0	0	1	743	6311409.28	6311409.28	0.00	0.00	0.00	1
6362617	0	1	0	0	0	743	6311409.28	6311409.28	0.00	68488.84	6379898.11	1
6362618	0	0	0	0	1	743	850002.52	850002.52	0.00	0.00	0.00	1
6362619	0	1	0	0	0	743	850002.52	850002.52	0.00	6510099.11	7360101.63	1

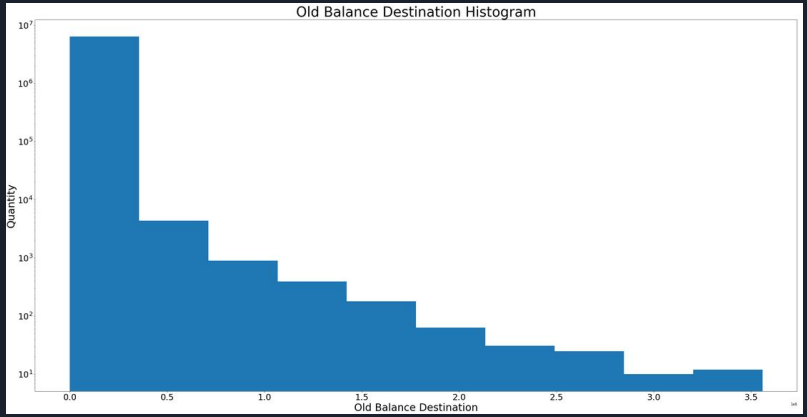
6362620 rows × 12 columns

Distribución de las Variables Numéricas









Análisis de la Data Fraudulenta

Casos totales de Fraude en el Dataset

8213

Este total será nuestro Target a trabajar en los modelos y en los análisis.

Monto totales de Fraude en el Dataset

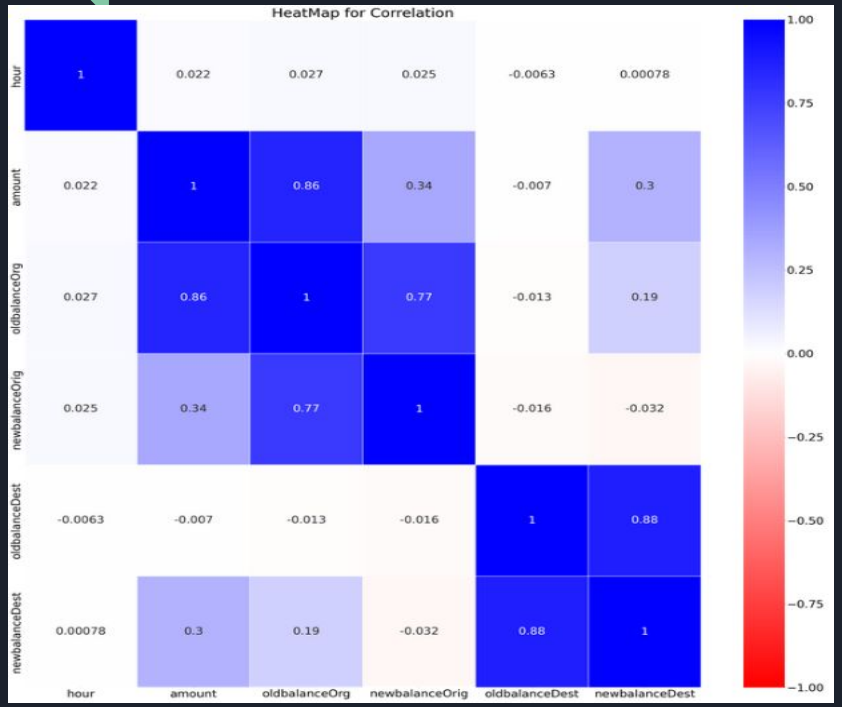
12.056.415.427,84 ₹

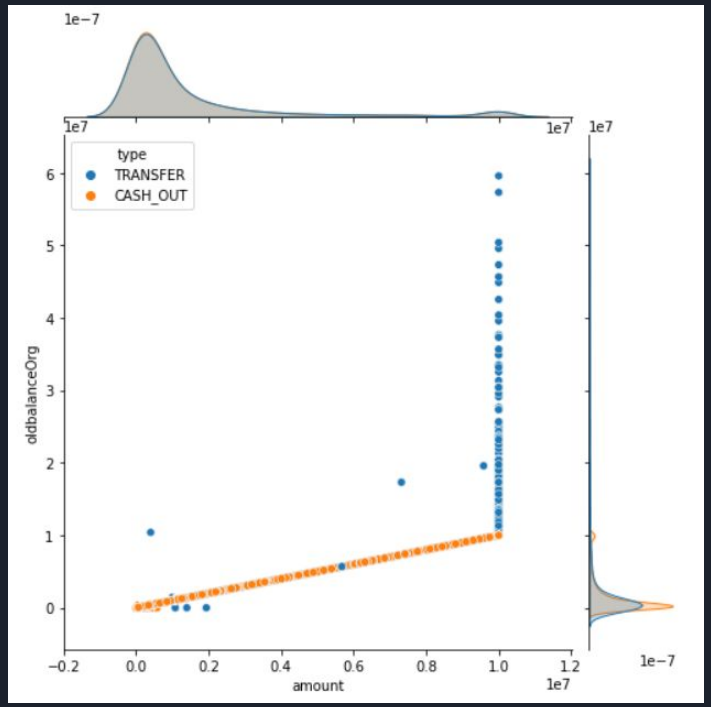
Monto Promedio de Transacción Fraudulenta

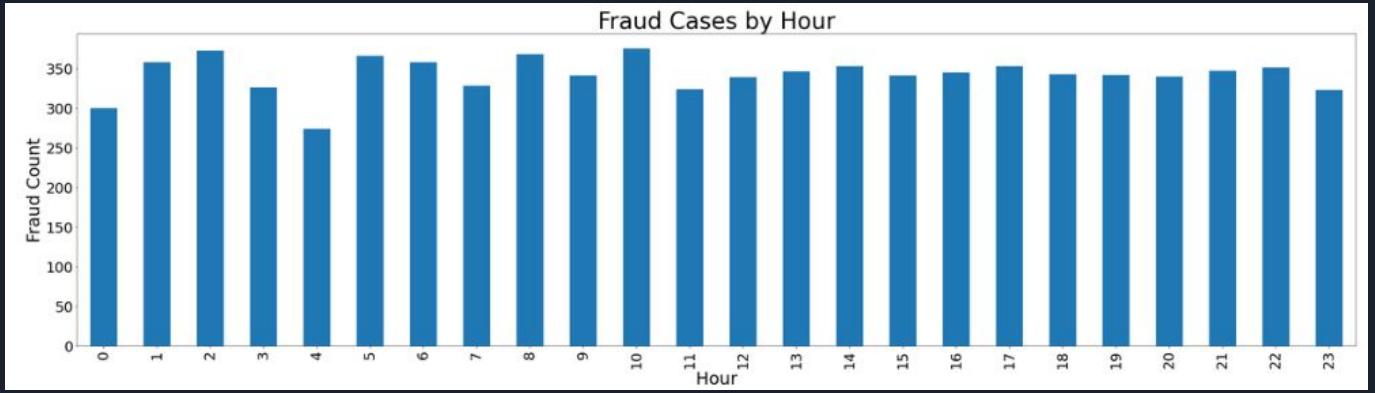
1.467.967,30 ₹

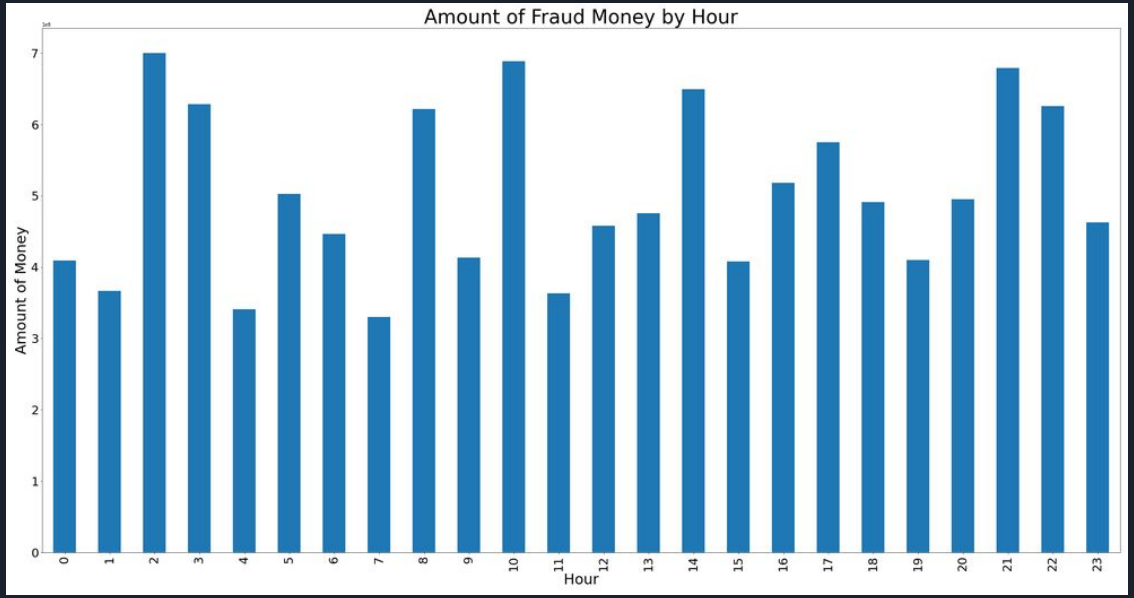
Una ₹ a dólar es igual a: 0,012Usd$

Análisis de la Data Fraudulenta

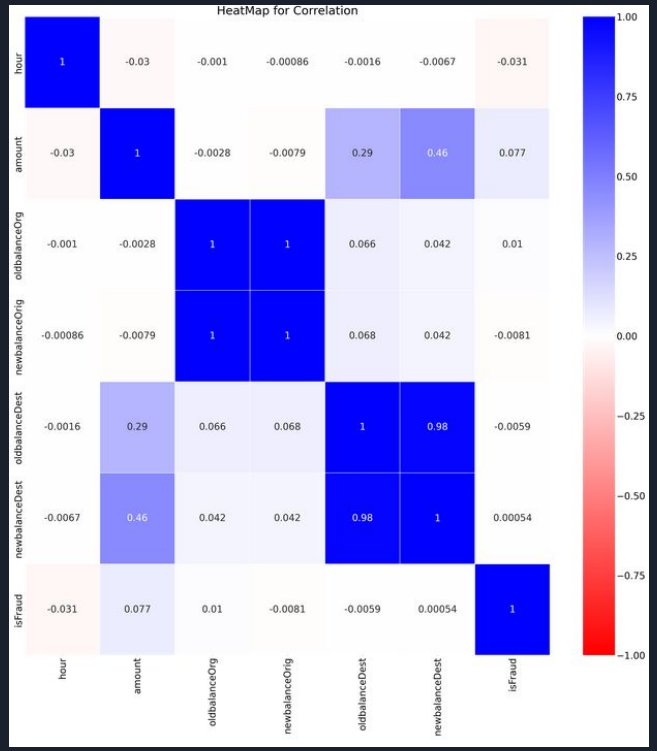


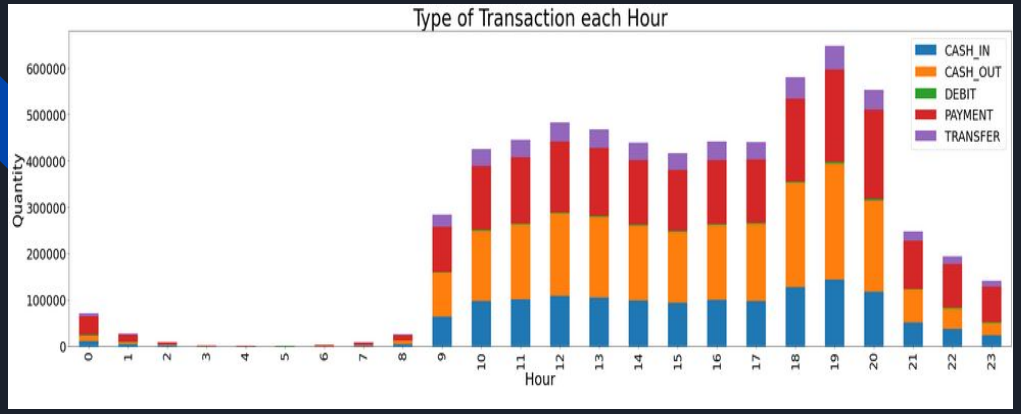


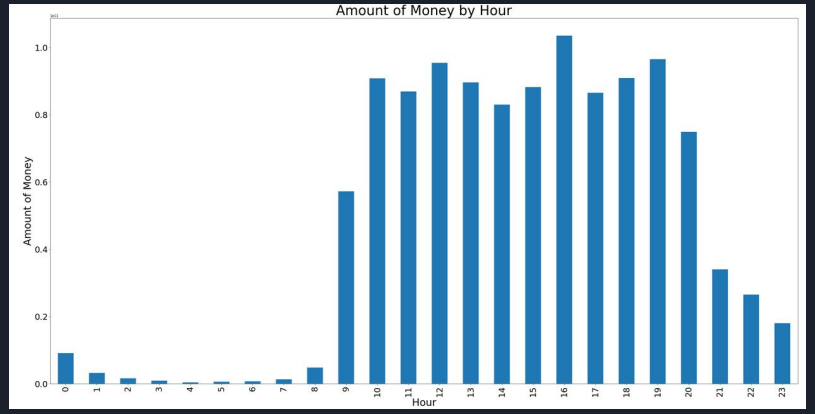

Fraud Cases by Hour
Fraud Count
0
50
100
150
200
250
300
350
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

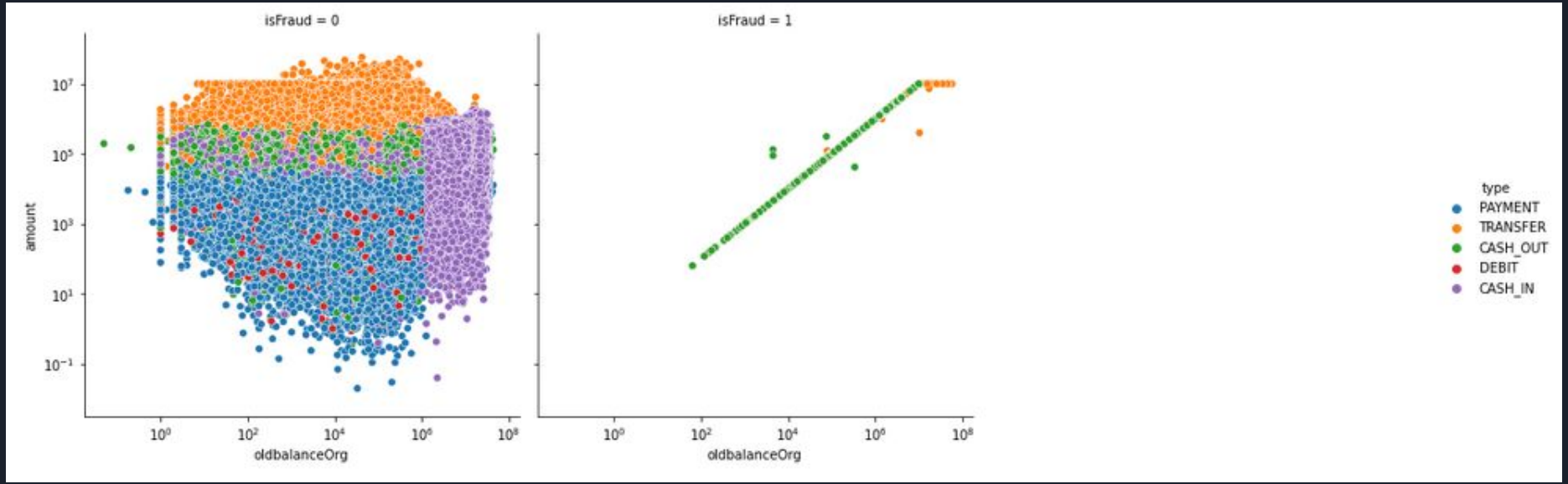

Amount of Fraud Money by Hour
Amount of Money
0
1
2
3
4
5
6
7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

Análisis de la Data General




Type of Transaction each Hour
Quantity
Hour
CASH_IN
CASH_OUT
DEBIT
PAYMENT
TRANSFER

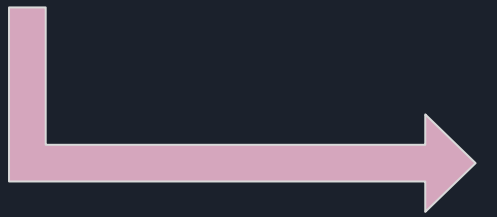

Amount of Money by Hour
Amount of Money
Hour

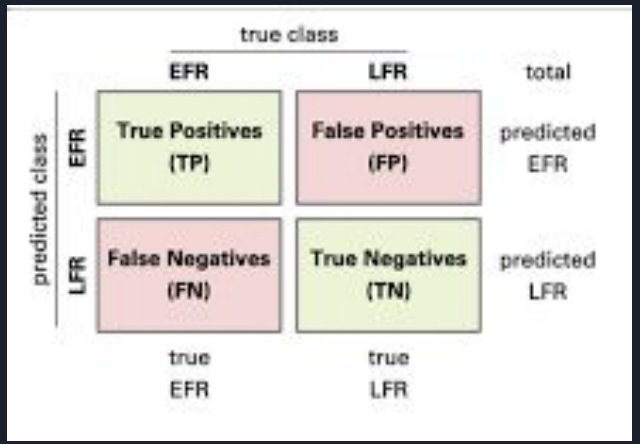

isFraud = 0
isFraud = 1
amount
oldbalanceOrg
oldbalanceOrg
type
PAYMENT
TRANSFER
CASH_OUT
DEBIT
CASH_IN

Métrica Considerada: Recall

Dado que nuestro objetivo es detectar casos de fraude, la métrica considerada en el desarrollo de este trabajo es la de Recall. Esta métrica lo que hace es darle un peso alto a la cantidad de falsos negativos. Si uno piensa detenidamente, se encuentra que en el caso de fraude tener un falso negativo es equivalente a decir que un caso de fraude no fue localizado, lo que se traduce en una pérdida económica para la entidad financiera, es por esto que a la hora de entrenar nuestros modelos vamos a buscar maximizar el valor de la métrica de recall (en el caso ideal tendríamos que el recall es de uno ya que no habría falsos negativos).

Falsos Negativos





Modelos

Se optó por hacer un análisis con cinco distintos modelos, donde se decidió tomar como modelo de referencia a la regresión logística, ya que es de los algoritmos más baratos en términos computacionales a la hora de realizar un problema de clasificación. Adicionalmente se consideraron al KNN (K-Nearest-Neighbours), un random forest, un MLPClassifier (Multi-layer Precepton) y al XGBoost.

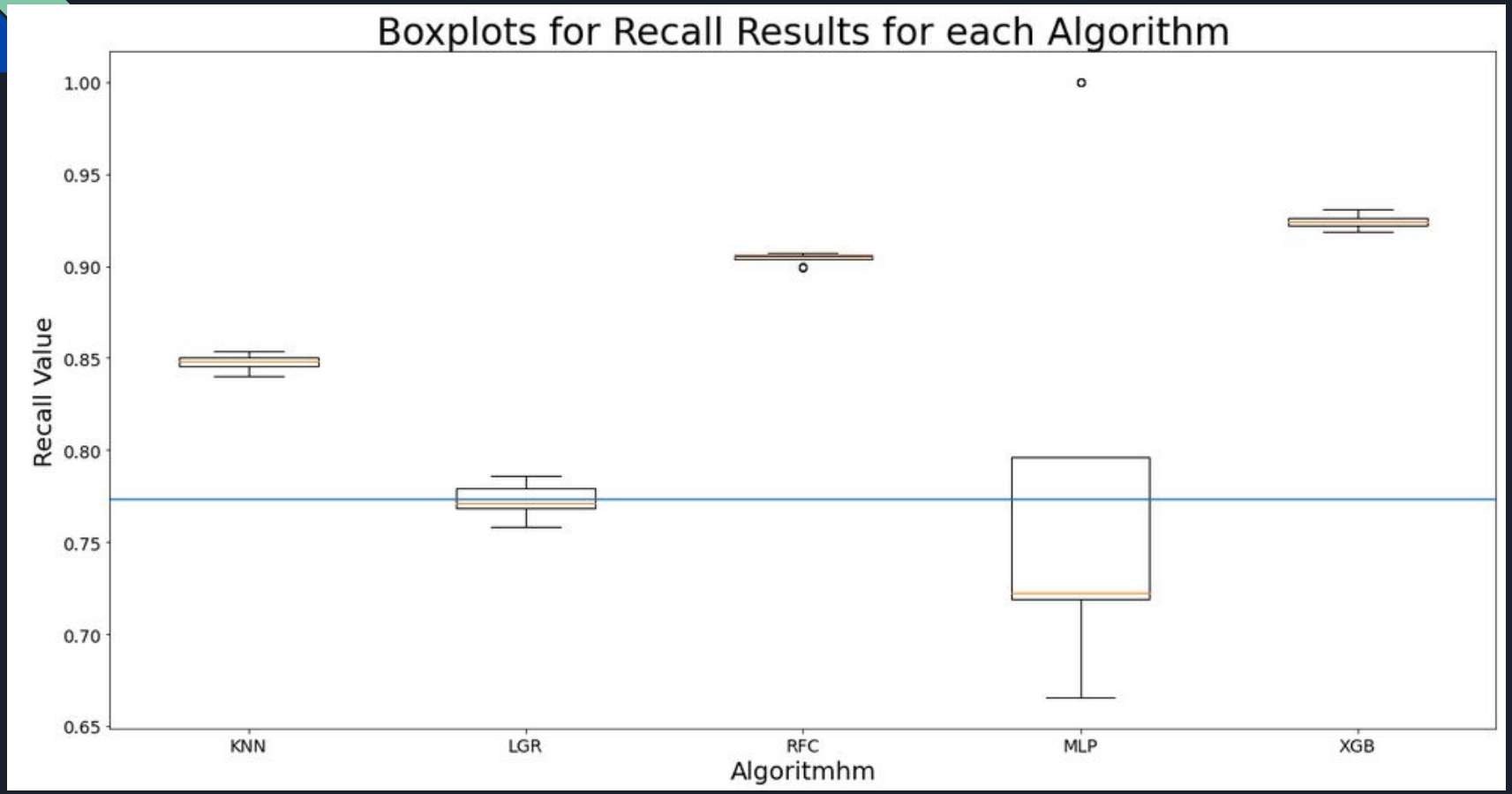
A la hora de entrenar y dada la característica desbalanceada del dataset se optó por hacer uso del Stratified K Fold. Los resultados obtenidos para las métricas de Recall y Precision* son los siguientes:

	Precision	Recall
KNN	0.867367	0.847663
LGR	0.450612	0.772929
RFC	0.979722	0.904252
MPL	0.382761	0.778849
XGB	0.972932	0.924120

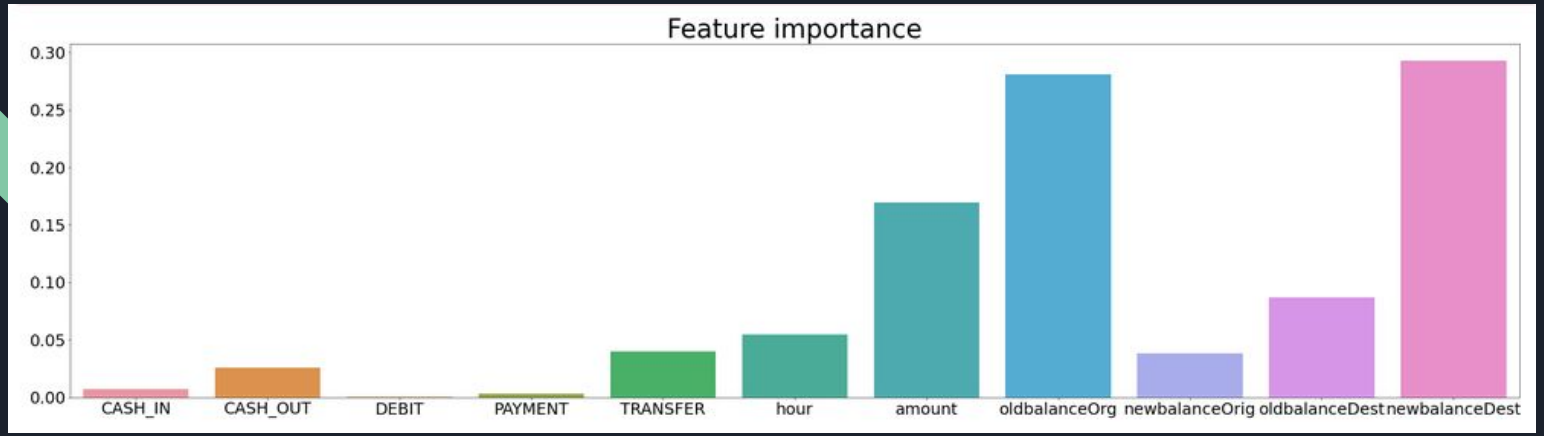
Precision*

*Se la calculó como una métrica adicional de referencia ya que la misma da peso a la cantidad de falsos positivos, lo que equivale a decir la cantidad de transacciones tildadas de fraude que en realidad no lo son.

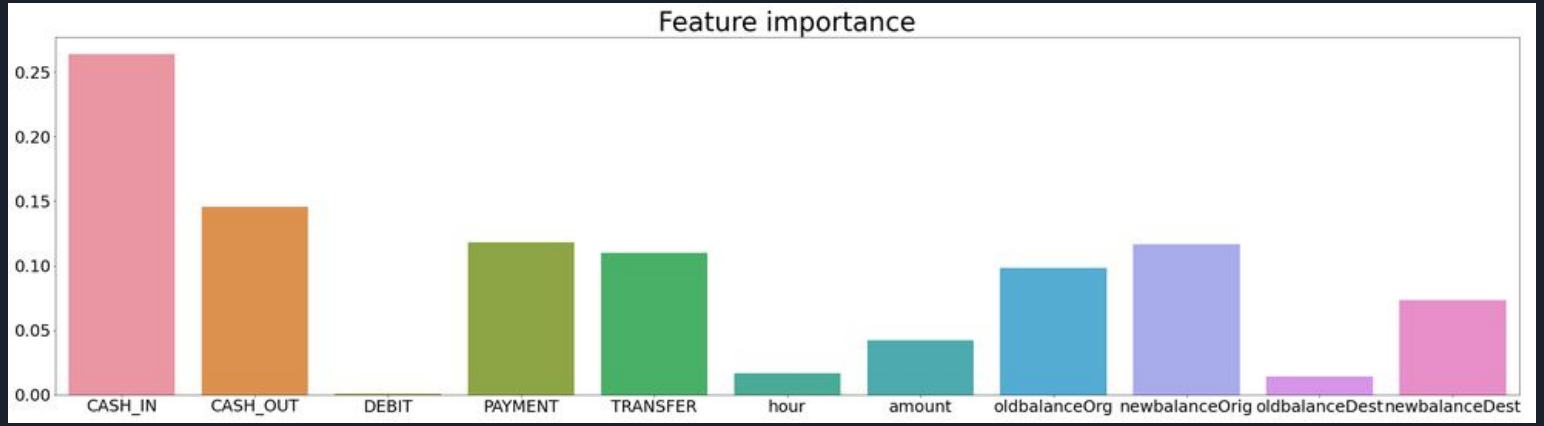
Resultados del Recall para cada algoritmo



RFC


Feature importance
0.30
0.25
0.20
0.15
0.10
0.05
0.00
CASH_IN
CASH_OUT
DEBIT
PAYMENT
TRANSFER
hour
amount
oldbalanceOrg
newbalanceOrig
oldbalanceDest
newbalanceDest

XGB


Feature importance
0.25
0.20
0.15
0.10
0.05
0.00
CASH_IN
CASH_OUT
DEBIT
PAYMENT
TRANSFER
hour
amount
oldbalanceOrg
newbalanceOrig
oldbalanceDest
newbalanceDest

Optimización

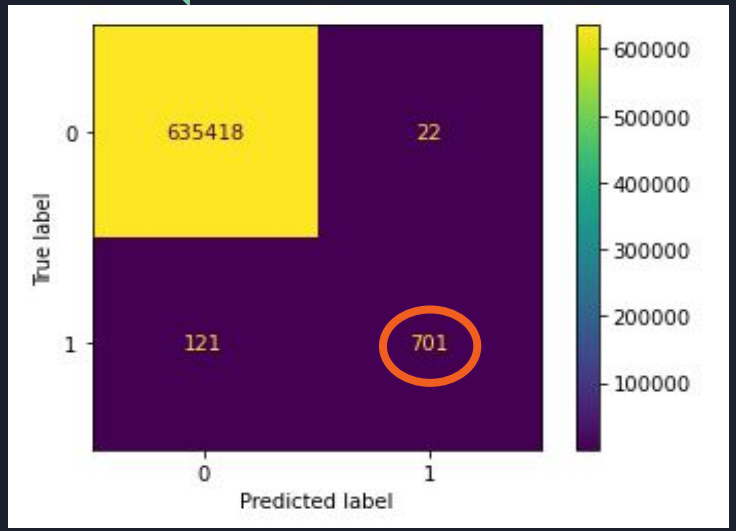
Se decidió utilizar la librería Optuna para buscar una manera de optimizar los resultados obtenidos para la métrica de recall con los distintos algoritmos. Esto se logra por medio de una optimización de los hiperparámetros que definen a cada modelo.

Para los algoritmos de KNN y LGR no hubo mejoras sustanciales tampoco en el recall. En el caso del algoritmo de Random Forest y del XGBoost los nuevos valores de recall obtenidos con la optimización de los hiperparámetros fueron los siguientes:

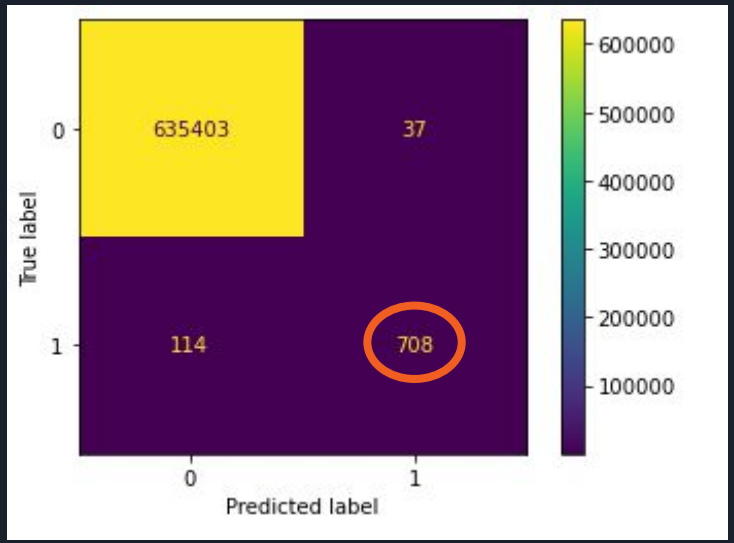
	Recall mean en las 3 iteraciones para el algoritmo RFC optimizado	Precision mean en las 3 iteraciones para el algoritmo RFC optimizado
Random Forest	**0.9075070518095113**	**0.9997384724640035**

Resultados Optimización

Antes



Después



Se detectaron 7 casos más de fraude, lo que equivale a una tengo una mejora del 5%, o 10 millones de rupias

Líneas De Trabajo a Futuro

Algunas de las líneas de los desarrollos que se pueden hacer a futuro son:

- Entrenar modelos de ML más robustos como redes neuronales profundas para obtener resultados con un mejor recall.
- Realizar un model ensamble, utilizando por ejemplo la técnica de stacking para aprovechar las ventajas de los distintos algoritmos y generar un super modelo que arroje mejores resultados.
- Presentar un plan de acción concreto para que la empresa financiera pueda comenzar a reducir los montos fraudulentos.

GRACIAS