Fraud Detection in credit cards..

By Alanoud Almutairi, Alnirah Alqahtani

Introduction:

Credit card fraud detection is one of the most important issues for credit card companies to deal with in order to earn trust from its customers. So, we aim to analysis fraud transaction dataset and classify it using python language to allow banks and card credit companies to understand and focus more in solving this fraud.

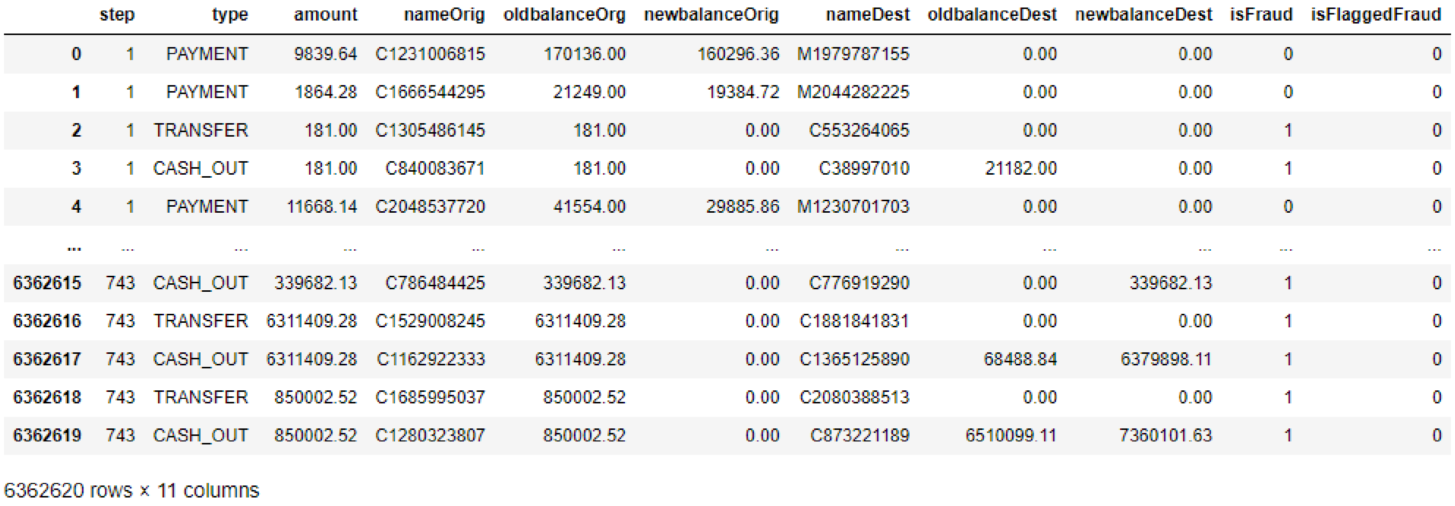
Objectives:

The objective is to classify the data transactions as either legal (isFraud 0) or frauds (isFraud 1) using a machine learning model. Five machine learning models will be trained and tested to determine which will shows the best results:

1. Logistic Regression
2. Decision Trees
3. Random Forest
4. Gradient Boosting
5. XGBoost

Description of the data:

This dataset presents credit card transactions for number of customers , where we have 8213 frauds out of 6354407 transactions, Also we have 11 features in this dataset.



	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.00	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	0	0
...	...	...	...	...	...	...	...	...	...	...	...
6362615	743	CASH_OUT	339682.13	C786484425	339682.13	0.00	C776919290	0.00	339682.13	1	0
6362616	743	TRANSFER	6311409.28	C1529008245	6311409.28	0.00	C1881841831	0.00	0.00	1	0
6362617	743	CASH_OUT	6311409.28	C1162922333	6311409.28	0.00	C1365125890	68488.84	6379898.11	1	0
6362618	743	TRANSFER	850002.52	C1685995037	850002.52	0.00	C2080388513	0.00	0.00	1	0
6362619	743	CASH_OUT	850002.52	C1280323807	850002.52	0.00	C873221189	6510099.11	7360101.63	1	0

6362620 rows × 11 columns

Data Preparation:

-First, we use feature selection to select data of type (int64-float64)

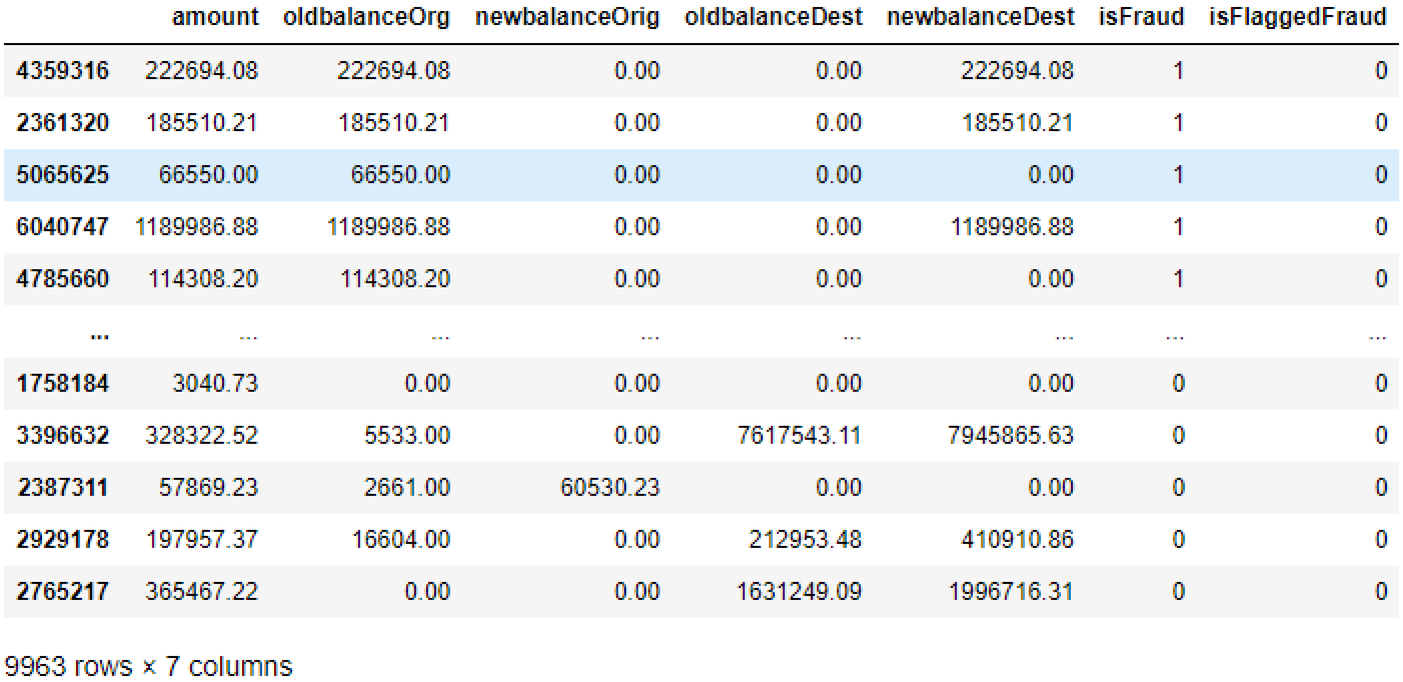
-Second, select balanced dataset from the data that contains 14000 record from the main data.

-Third use selection feature to select data from type (int64 ,float64)

-Forth, drop column 'step' to have a full numeric dataset.

-Fifth, Check and drop duplicated and null values.

The final dataset chosen:

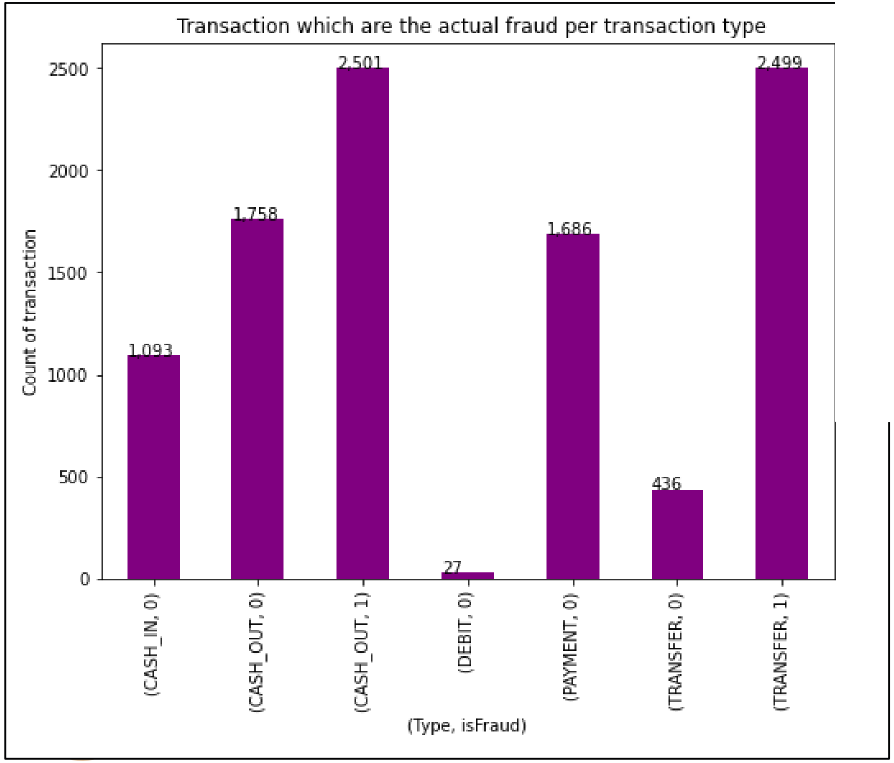


	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
4359316	222694.08	222694.08	0.00	0.00	222694.08	1	0
2361320	185510.21	185510.21	0.00	0.00	185510.21	1	0
5065625	66550.00	66550.00	0.00	0.00	0.00	1	0
6040747	1189986.88	1189986.88	0.00	0.00	1189986.88	1	0
4785660	114308.20	114308.20	0.00	0.00	0.00	1	0
...	...	...	...	...	...	...	...
1758184	3040.73	0.00	0.00	0.00	0.00	0	0
3396632	328322.52	5533.00	0.00	7617543.11	7945865.63	0	0
2387311	57869.23	2661.00	60530.23	0.00	0.00	0	0
2929178	197957.37	16604.00	0.00	212953.48	410910.86	0	0
2765217	365467.22	0.00	0.00	1631249.09	1996716.31	0	0

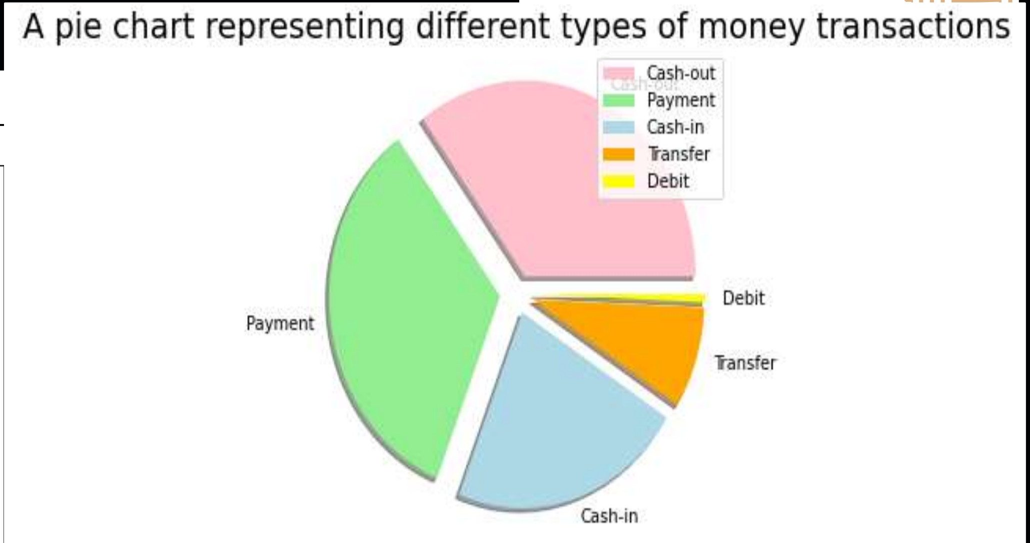
9963 rows × 7 columns

Result:

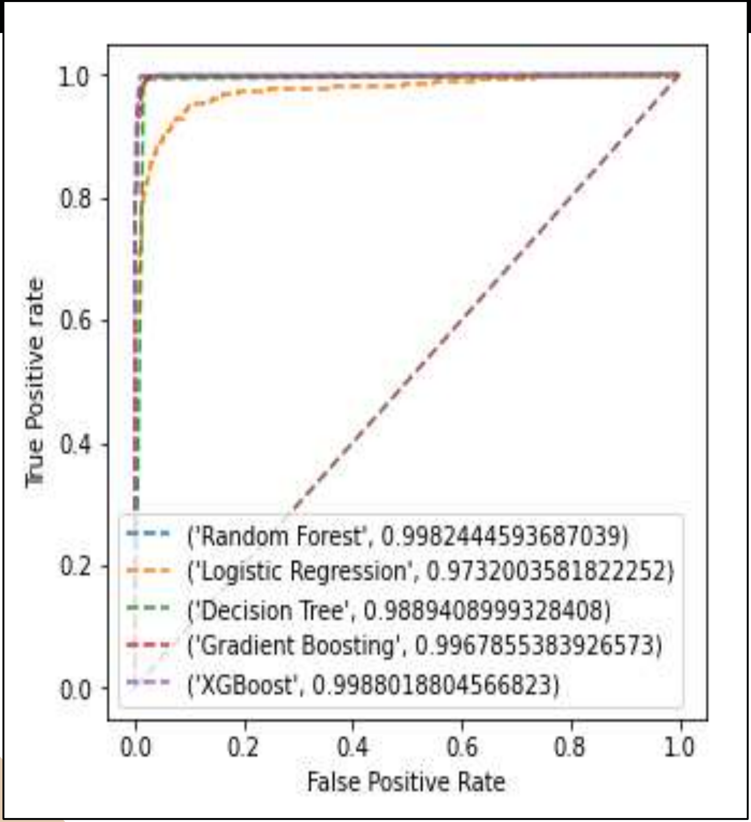
Model	F1_score	Accuracy	AUC
Logistic Regression	84.606613454960	86.450317832050	0.9774251175261
Random Forest	98.904018598472	98.895951823352	0.9986897246474
Decision Tree	99.202657807308	99.197055877152	0.9919581374524
Gradient Boosting	98.378020523005	98.360655737704	0.9954743675845
XGBoost	99.169711059448	99.163599866175	0.9986254757107

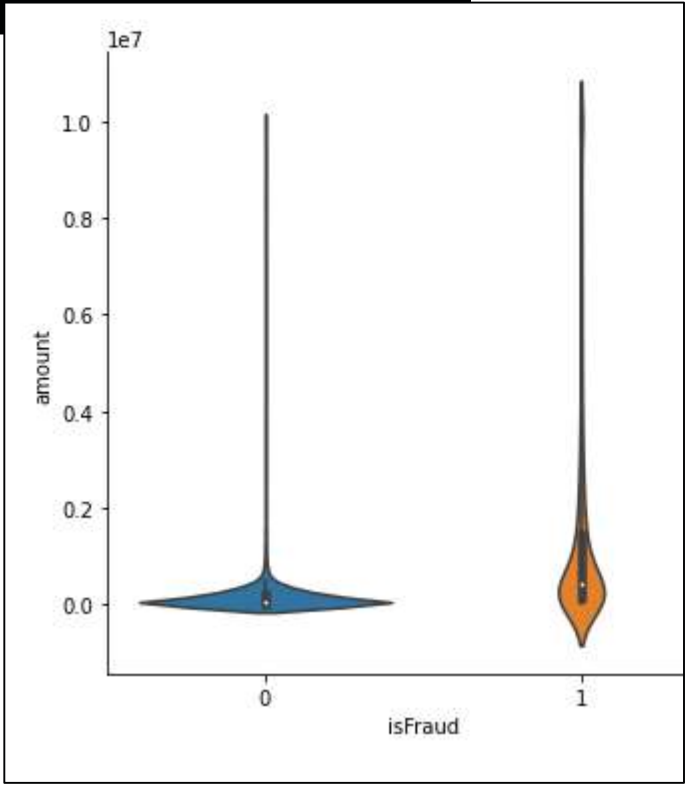


Transaction which are the actual fraud per transaction type



A pie chart representing different types of money transactions


1.0
0.8
0.6
0.4
0.2
0.0
True Positive rate
('Random Forest', 0.9982444593687039)
('Logistic Regression', 0.9732003581822252)
('Decision Tree', 0.9889408999328408)
('Gradient Boosting', 0.9967855383926573)
('XGBoost', 0.9988018804566823)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


1e7
1.0
0.8
0.6
0.4
0.2
0.0
amount
0
1
isFraud

Project limitations:

- Date and day is missing in the dataset, Also if we could have the locations for every transaction record would be more efficient to detect this fraud transactions.

Tools:

- **Technologies :** Python, Jupyter Notebook
- **Libraires :** Pandas, Numpy , Seaborn , Sklearn.

Conclusion:

Based in our analysis for the data you can notice how huge is fraud transactions was, So we suggest to cards companies to increase the systems security and focus more on that side of dangerous frauds transaction..

Thanks..