ML Project on

FRAUD TRANSACTION'S

- Ramdas B Yamgar

Steps

1. Import Libraries¶
2. Data Preprocessing
3. Exploratory Data Analysis
4. Feature engineering
5. Lebel Encoding
6. Balancing Dependent Variable
7. Train Test Split¶
8. Scaling values
9. Model training : Logistic Regression
10. Improving accuracy

Business Context

- This case requires trainees to develop a model for predicting fraudulent transactions for a financial company and use insights from the model to develop an actionable plan. Data for the case is available in CSV format having 6362620 rows and 10 columns.
- Candidates can use whatever method they wish to develop their machine learning model. Following usual model development procedures, the model would be estimated on the calibration data and tested on the validation data. This case requires both statistical analysis and creativity/judgment. We recommend you spend time on both fine-tuning and interpreting the results of your machine learning model.

Assumptions

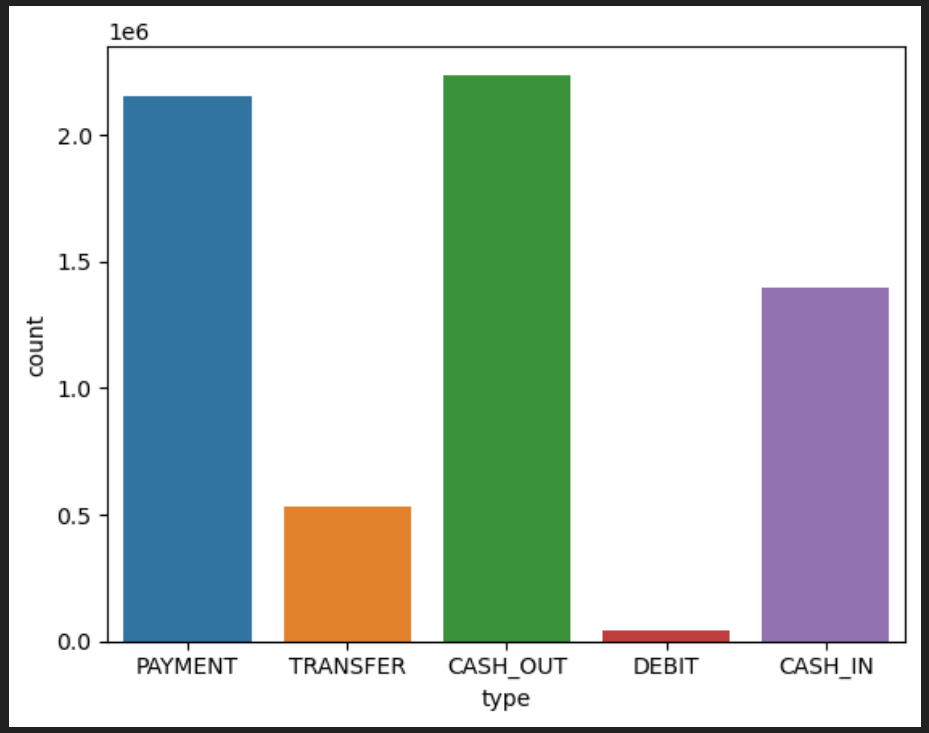
- No unusual occurences between 2015 to 2017 will have impact o the data used.
- The information is still current and can be used to analyze hotel possible plans in an efficient manner
- There are no unanticipated negatives to the hotel employing any advised techniques.
- The hotels are not currently using any of the suggested solutions.

Research Questions

- 1. Data cleaning including missing values, outliers and multi-collinearity.
- 2. Describe your fraud detection model in elaboration.
- 3. How did you select variables to be included in the model?
- 4. Demonstrate the performance of the model by using best set of tools.
- 5. What are the key factors that predict fraudulent customer?
- 6. Do these factors make sense? If yes, How? If not, How not?
- 7. What kind of prevention should be adopted while company update its infrastructure?
- 8. Assuming these actions have been implemented, how would you determine if they work

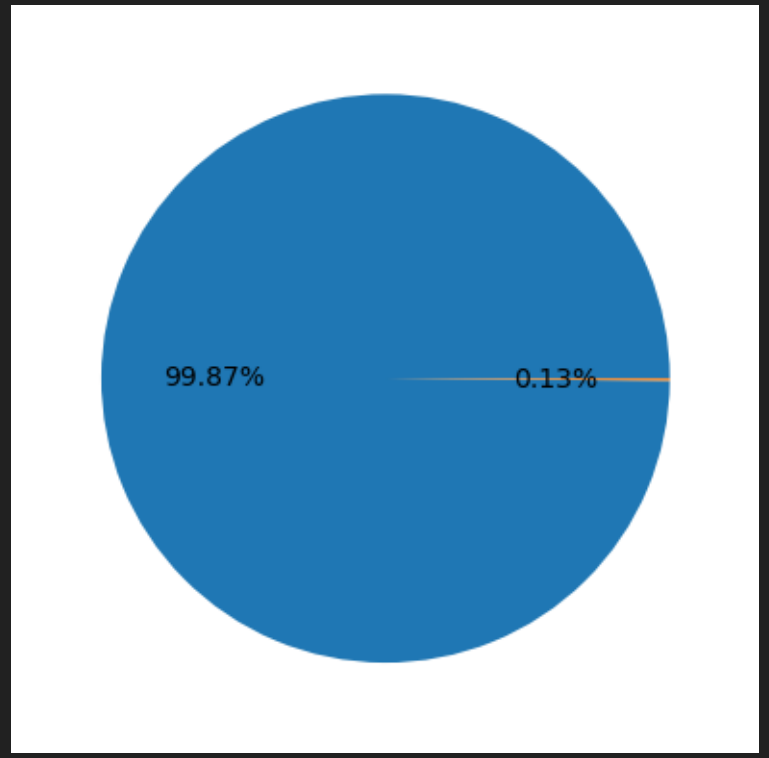
Analysis and Finding

- Graph shows ‘type’ of transaction
- ‘Payment’ and ‘Cash_Out’ has maximum no of transaction occur as compare to ‘Transfer’ and ‘Cash_In’.
- ‘Debit’ has least no of transactions

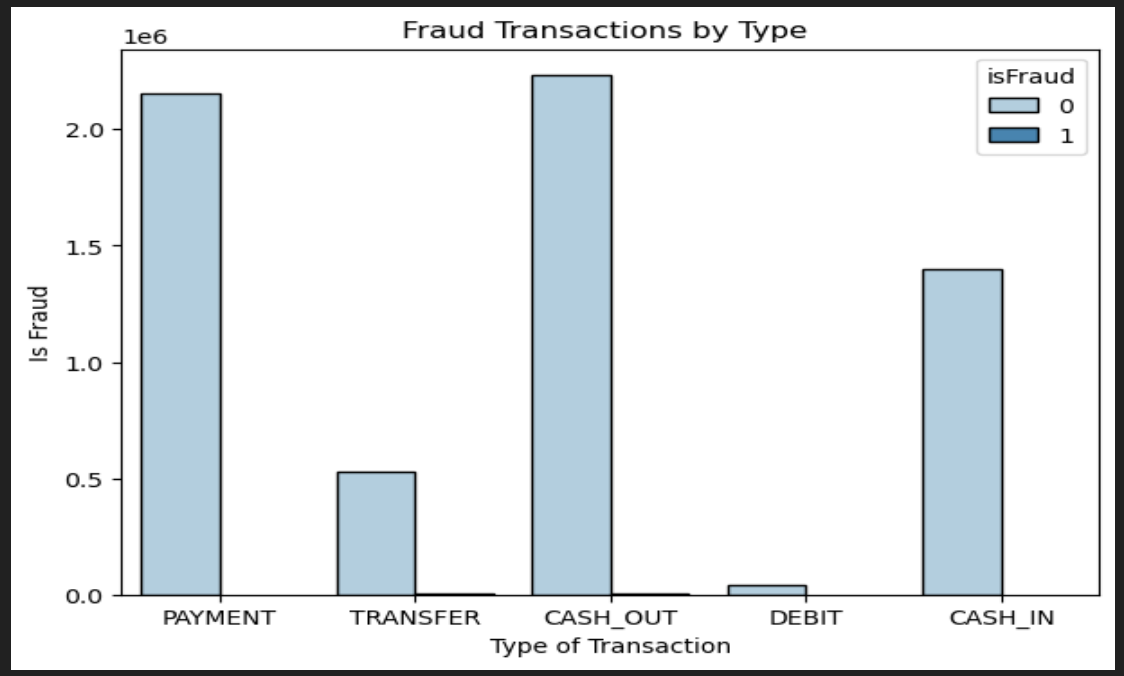


Percentage of Fraud transactions.

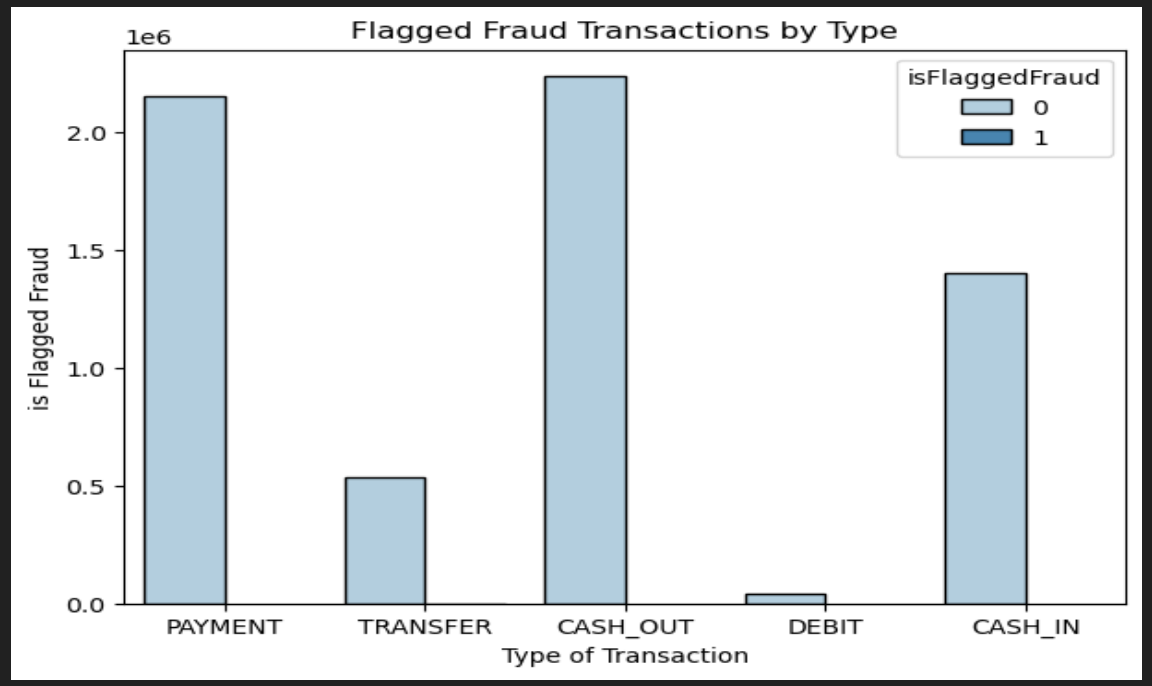
- In our data only 0.13 % transactions are fraud and 99.87% are not frauded.



Count of Fraud and Flagged Fraud



There are total 8213 Fraud Transaction
4116 are in 'Cash_Out and 4097 are in ''Transfer' type of transaction.

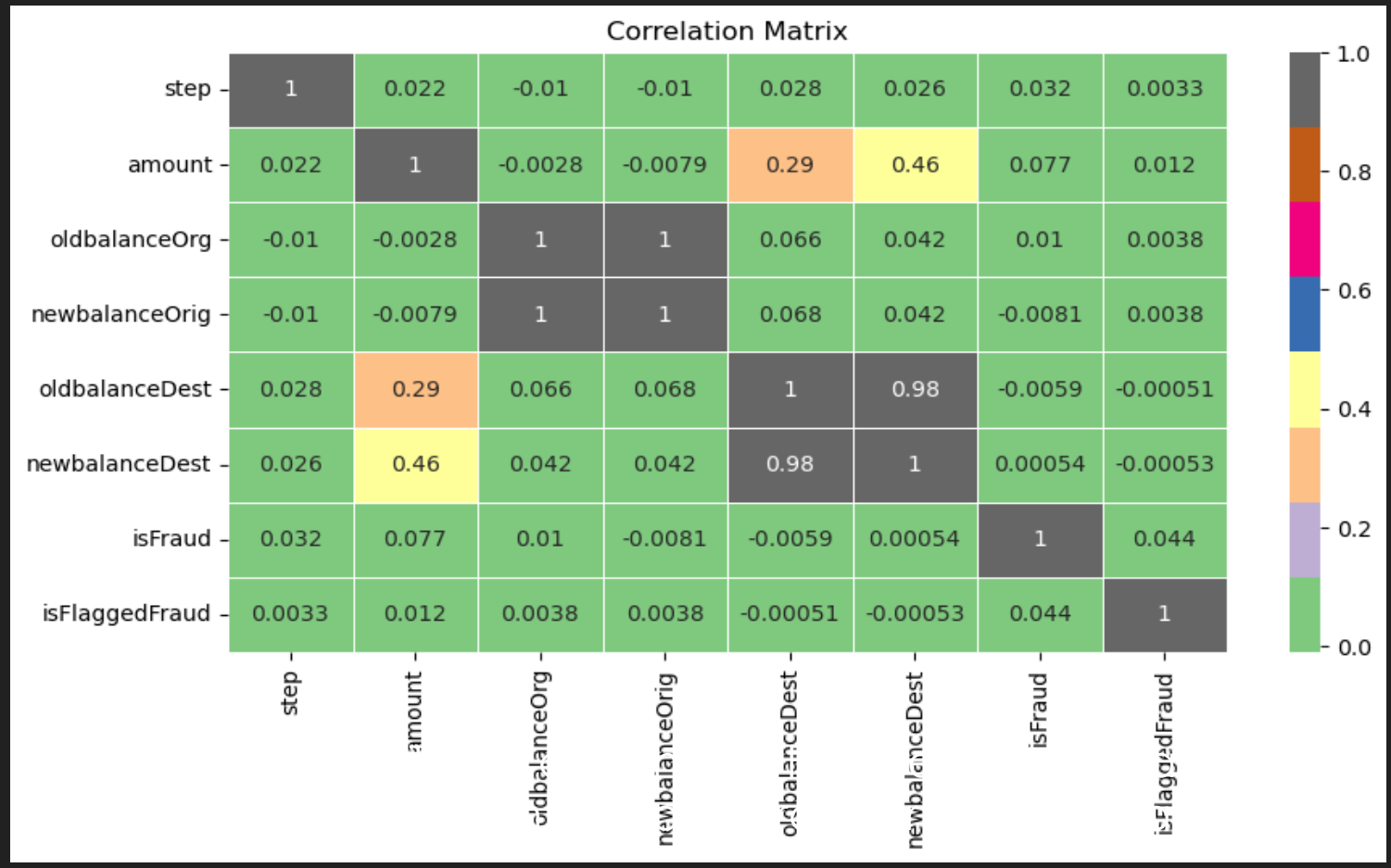


There are only 16 Flagged Fraud Transaction
Which is also in 'Transfer' type of transaction.

Correlation

Graph show, there are two features which shows multicollinearity.
Feature:

1. oldbalanceOrg with newbalanceOrg.
2. oldbalanceDest with newbalanceDest



Task

1. Data cleaning including missing values, outliers and multi-collinearity.
   - In our Data , Not any single null value and also Not any duplicate row.
   - We check Multicollinearity by using Corr.plot which shows Multicollinearity ‘oldbalancedOrg’ with ‘newbalancedOrg’ and ‘oldbalancedDist’ with ‘newbalancedDist’

2. Describe your fraud detection model in elaboration.
   -We use Logistic regression algorithm for model building. Because,
   1. It avoid Multicollinearity
   2. The chances of overfitting is less as compare to other algoritms

3. How did you select variables to be included in the model?
   - First we go with all features including multicollinearity features but over model tends to overfit So, We drop that features.

Tasks

4. Demonstrate the performance of the model by using best set of tools.

Before Drop Multicolinear Features

```
Confusion Matrix :
[[1197339   73425]
 [ 118449 1152550]]

Accuracy    : 92.45114513036818
Precision   : 94.01088929219601
TPR/ Recall : 90.68063782898335
FPR         : 8.351853568621752e-05
F1_Ratio    : 92.31573897044983
```

After Drop Multicolinear Features

```
Confusion Matrix :
[[1197339   73425]
 [ 118449 1152550]]

Accuracy    : 99.99984283204088
Precision   : 0.0
TPR/ Recall : 0.0
FPR         : nan
F1_Ratio    : 0.0
```

Tasks

5. What are the key factors that predict fraudulent customer?
- When Flagged Fraud is happen then Fraud is also happen.
- Only in 'Payment ' and 'Cash_Out' type of transaction fraud happen.

6. Do these factors make sense? If yes, How? If not, How not?
-Yes, Because When we have to make transaction type of 'Transfer' and 'Cash_Out' that time chances of fraud happen is high as compare to Cash_In.

7. What kind of prevention should be adopted while company update its infrastructure?
- Take more care when transaction type is 'Transfer' and 'Cash_Out'
- Do more verification large amount of transaction(i.e >100000).

8. Assuming these actions have been implemented, how would you determine if they work
- First we balanced data (dependent variable) but our accuracy goes down and model become over fit, So We avoid balancing tech. i.e We Assume that our data is balanced.