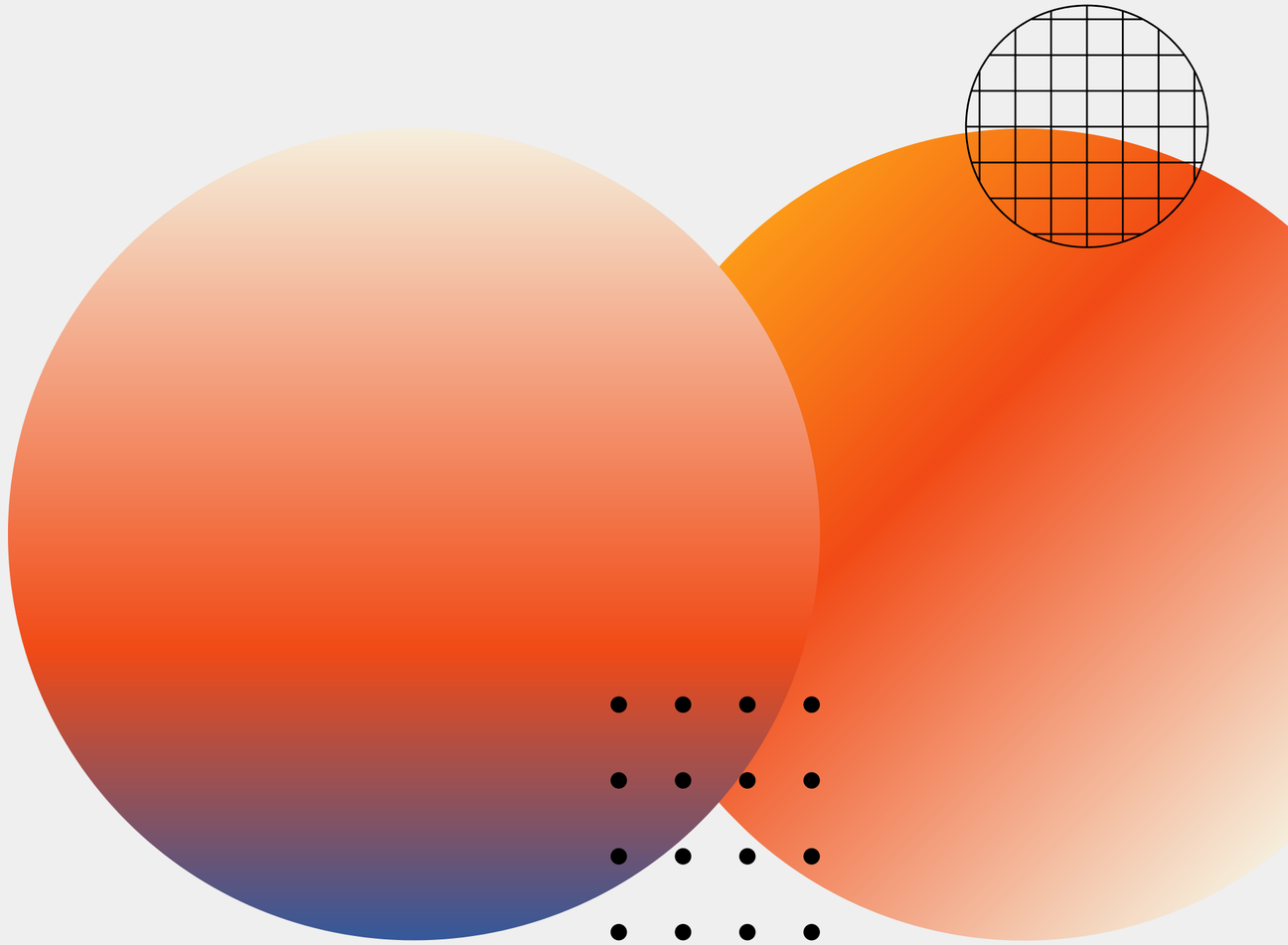
Let's Start

Fraudulent Transaction Analysis



Objectives

- Analyze Customer Transactions through a Bank.
- Create an awareness of internal audit's responsibilities relating to fraud detection and prevention.
- Become alert to fraud red flags and optimize current system.
- Identify basic fraud audit techniques and create a model based on that.

Context

It is important that finance companies are able to recognize fraudulent transactions so that customers are not charged for items that they did not purchase.

Content

The datasets contains transactions made by customers of a finance company. This dataset presents transactions that occurred in some days, where we have 8213 frauds out of 63,62,619 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.129% of all transactions.

Data Overview

- step
- type
- amount
- nameOrig
- oldbalanceOrg
- newbalanceOrig
- nameDest
- oldbalanceDest
- newbalanceDest
- isFraud
- isFlaggedFraud

Packages and Libraries Used

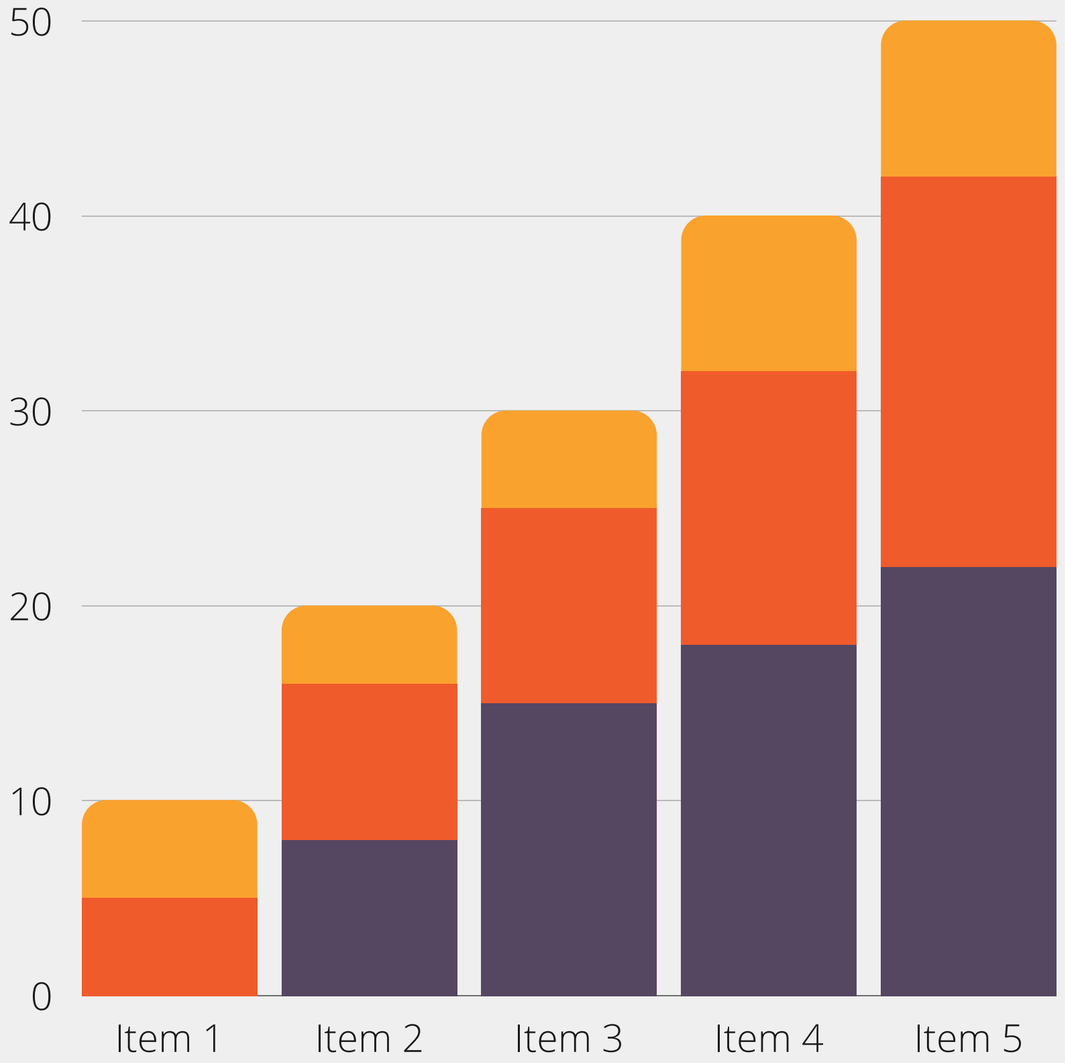
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.tree import DecisionTreeClassifier       # Decision tree algorithm
from sklearn.ensemble import RandomForestClassifier   # Random forest tree algorithm
from sklearn.metrics import average_precision_score
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score, roc_curve, auc, precision_score
import imblearn
from collections import Counter
from sklearn.datasets import make_classification
from imblearn.under_sampling import RandomUnderSampler
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

Process of Optimizing Fraud Detection System

1. Data Pre-Processing & Visualization
2. Feature Selection & Hyperparameter Tuning
3. Scaling and Splitting the Data into Test and Train
4. Model Creation and Deployment

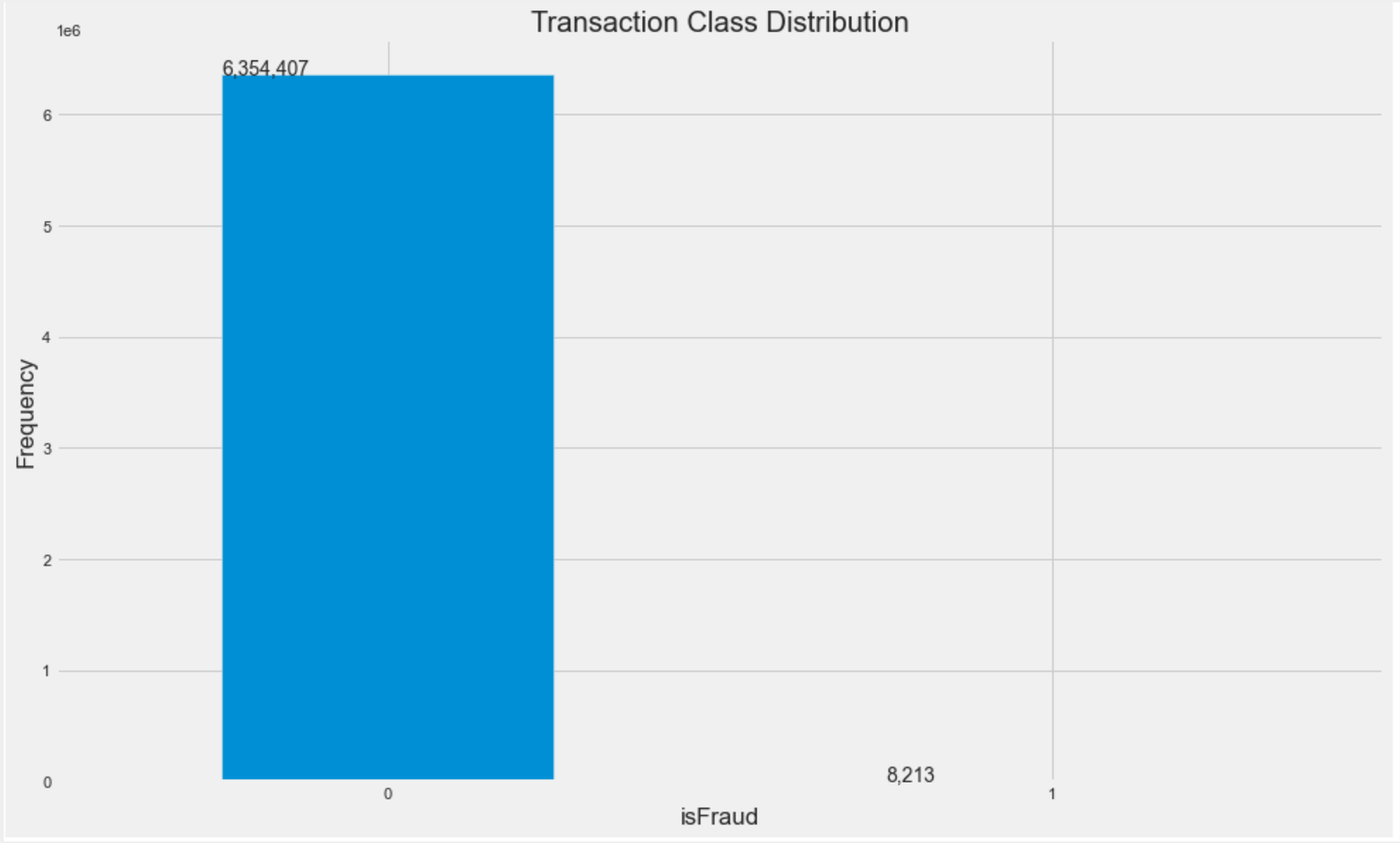
Visualizations

The following are some important graphs from the analysis and pre-processing of the data.



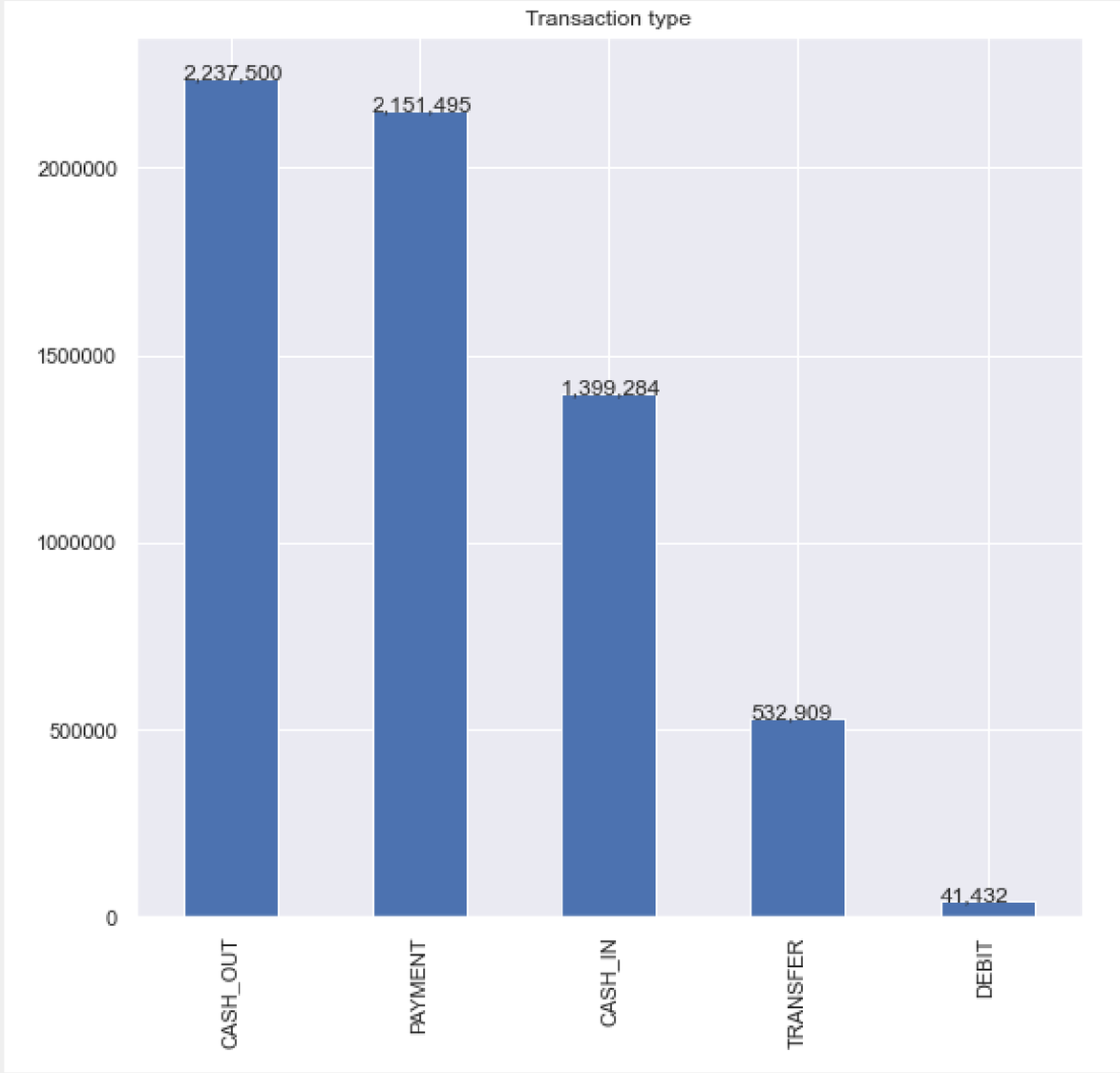
Transaction Class Distribution

Total number of Fraud transactions as per the column "isFraud"



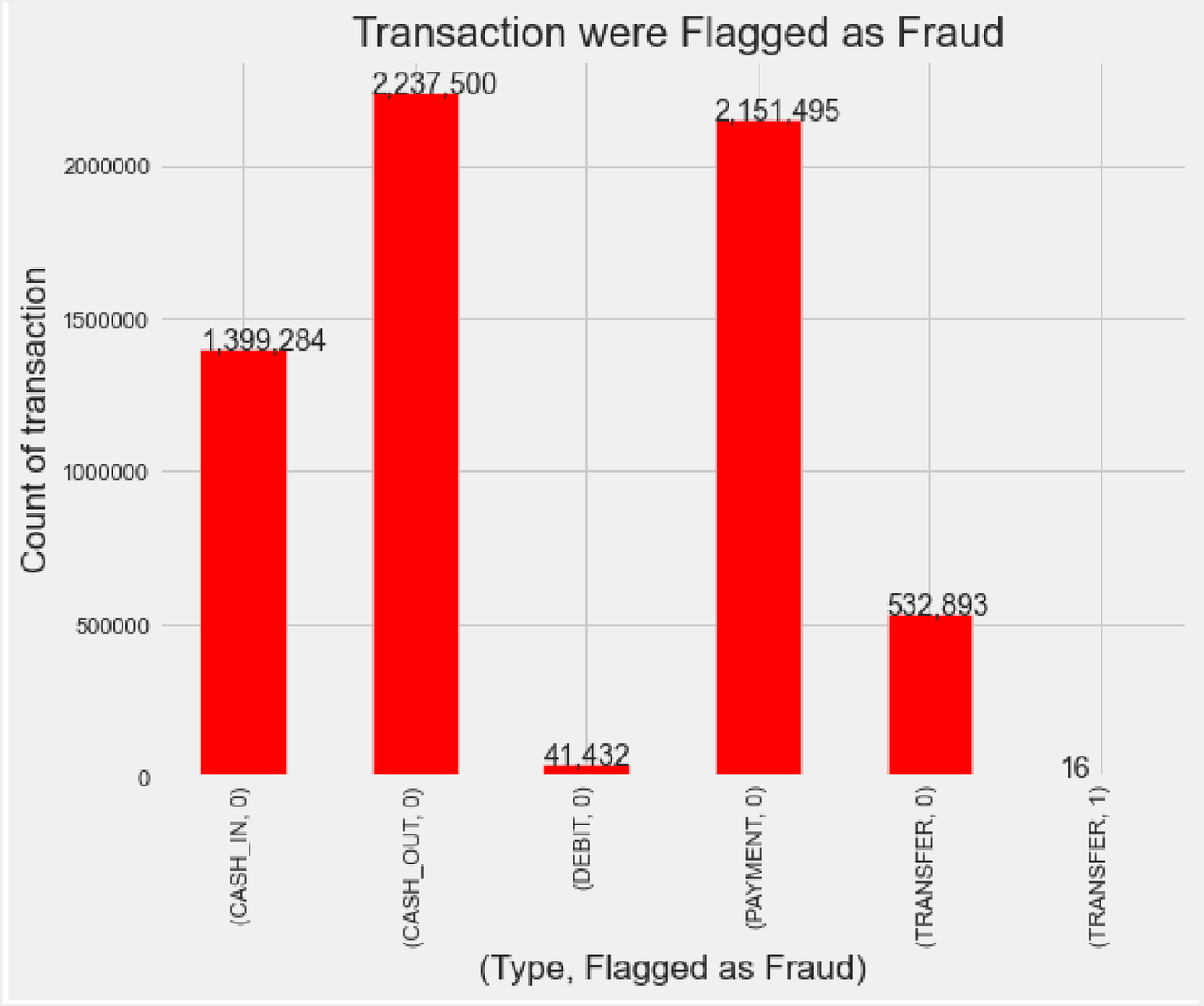
Transaction Type

Visualizing the frequency of transaction type

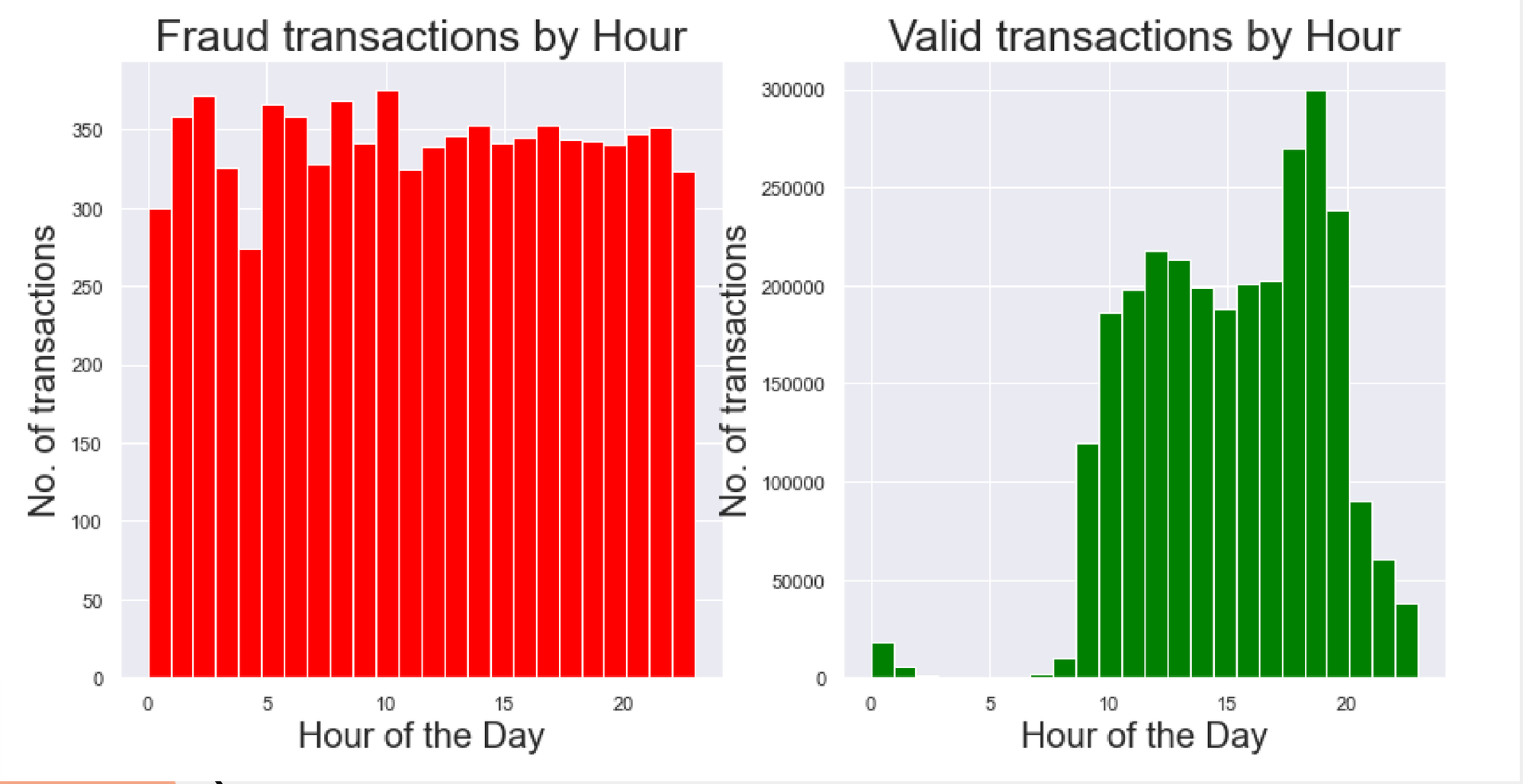


Transaction Flagged as Fraud

Visualizing number of fraudulent transaction which were flagged fraudulent



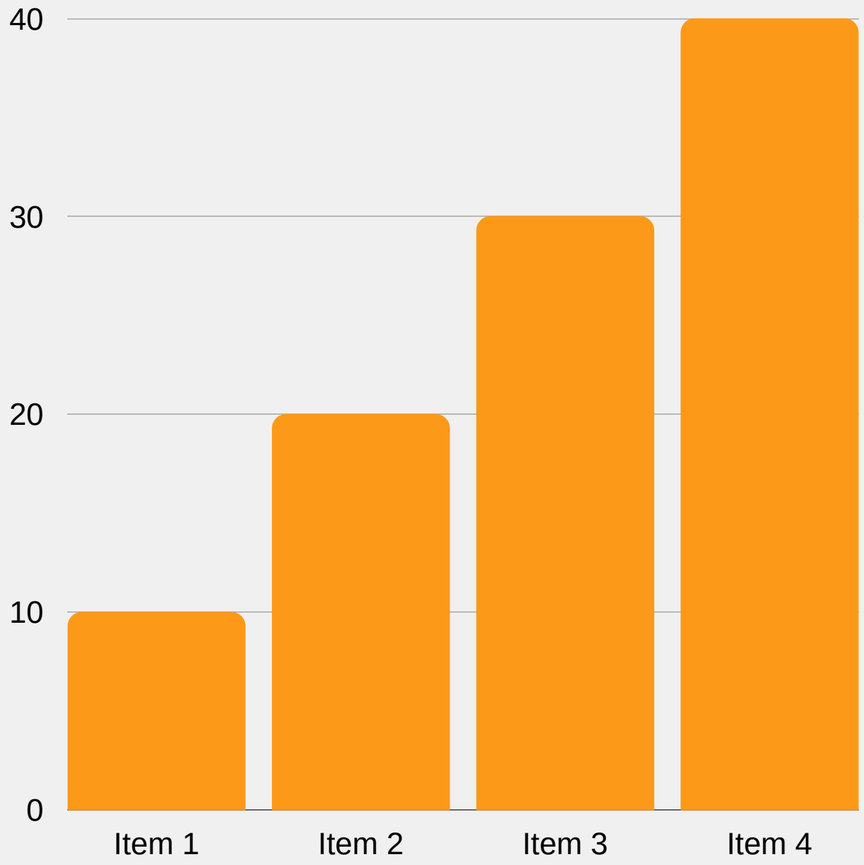
Fraud vs Valid Transactions in a Day



MODEL SELECTION

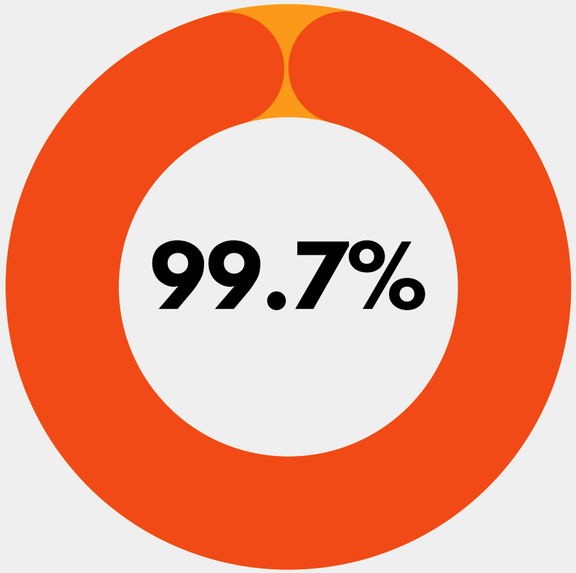
CRITERIAS

- Generalizes data well
- Interpretable
- Good fit on the test data, i.e. no over fitting and under fitting observed
- Better prediction accuracy based on AUC

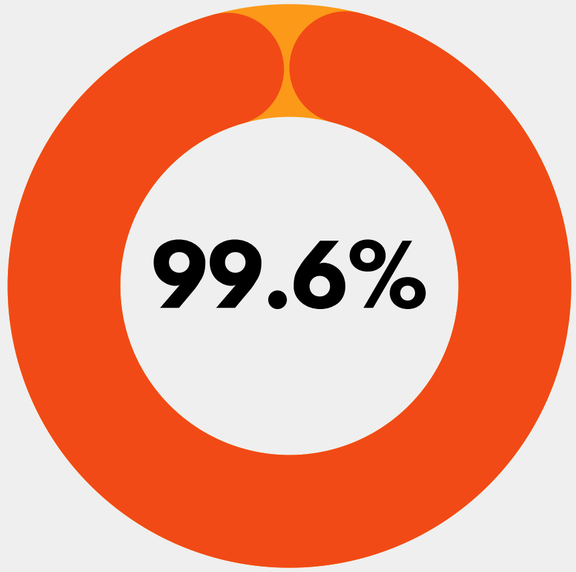


Model Performace

Accuracy percentage of prediction done by two machine learning models
The model accuracy is determined using AUC method i.e. Area Under the Curve.



Random Forest



Decision Tree

Key features – Top 3

errorbalanceOrg

error in the final balance of the account initiating the transaction

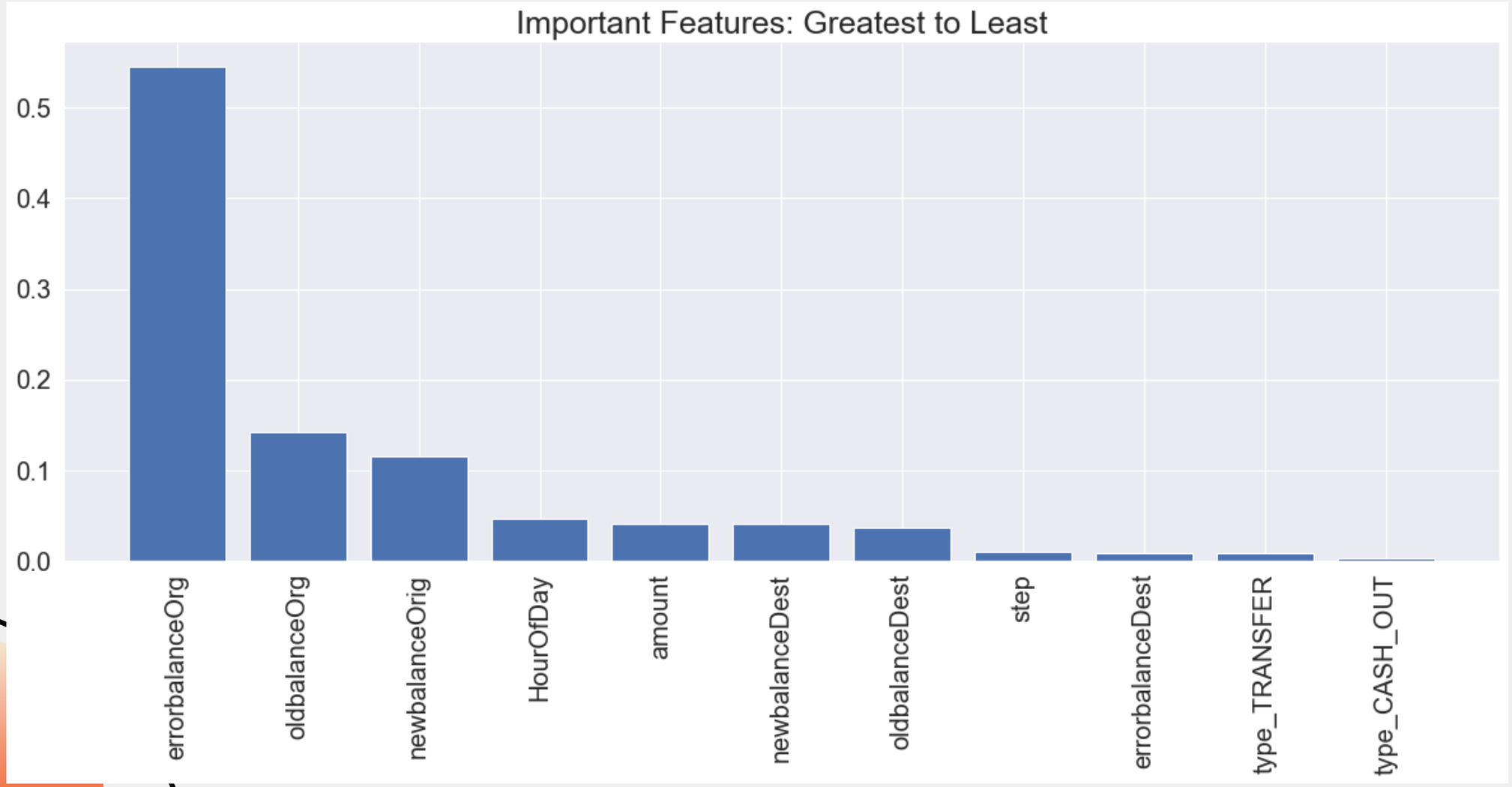
oldbalanceOrg

balance in the account initiating the transaction before the transaction

newbalanceOrg

balance in the account initiating the transaction after the transaction

Key Features - All



Conclusion

- The dataset is huge with over million data points, and the ratio of fraud to valid data is heavily skewed towads valid data

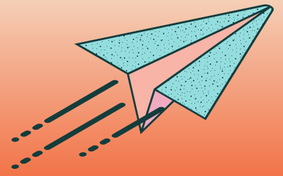
- Feature engineering and creation of two new features namely 'errorbalance' and 'HourofDay' yielded fruitful results.

- Random Forest Classifier is the best model in the given situation as it is fairly accurate in predicting both fraud and valid data, and has the heigest AUC .

Strategy for Production and Maitainance

GOAL

Track fraudulent transaction



A/B Tests Before Implementing

01 **Quantify Baseline Performance**

02 **Check for Improvement over current system**

Continue Optimizing Model

01 **Collect more data**

More data will help us creating more functional model

02 **Change Matrics**

Customize model training to gain higher accuracy with test dataset

End

Thank you

Do you have any questions?

Reach out to me at shahooda637@gmail.com

Click here for full Jupyter Notebook Link