REPORT

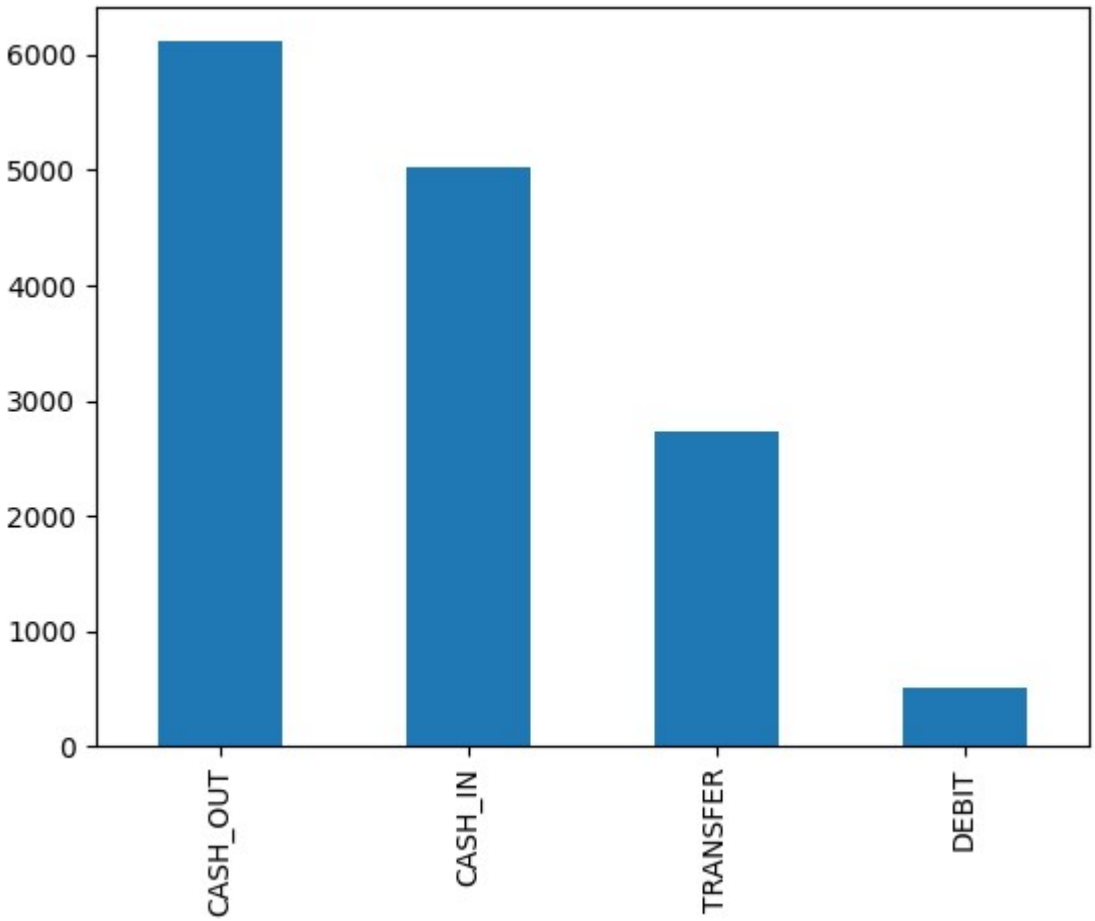
1. Data cleaning including missing values, outliers and multi-collinearity.

- Removed missing values as the data is neither random nor scarce, therefore it is better to remove the data points with missing values instead of replacing them by some measure of central tendency.

- Removed all rows containing ‘Merchant (M)’ entries as the account details of such entries were not present. Therefore, it would not contribute to the model in any way.

- Explored Types of transactions. There are mainly 4 types of transactions, namely

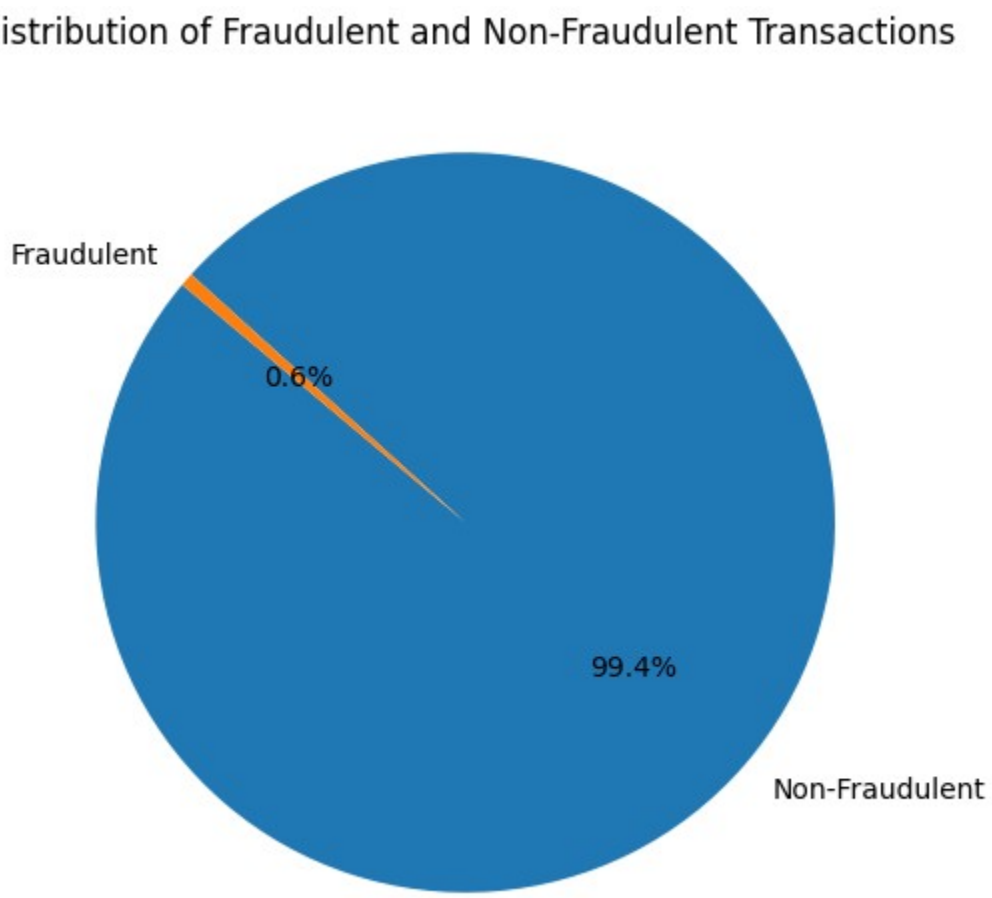
1. **Transfer:** This category represents transfers of funds between accounts. Transfers can occur between a customer's own accounts or between accounts belonging to different customers.
2. **Cash out:** This category represents transactions involving the withdrawal or "cashing out" of funds from an account.
3. **Debit:** This category represents debit card transactions, where funds are debited or subtracted from the account for purchases or payments.
4. **Cash in:** This category represents transactions where money is being deposited or "cashed in" to an account.



The major difference between "Cash In" and "Transfer" is that "Cash In" indicates that a customer is adding funds to their account, such as depositing cash or checks while "Transfer" often used for moving money between accounts within the same bank

The most amount of transactions are of "Cash Out" type followed by "Cash In" which indicates that there is a higher amount of cash outflow as compared to cash inflow.

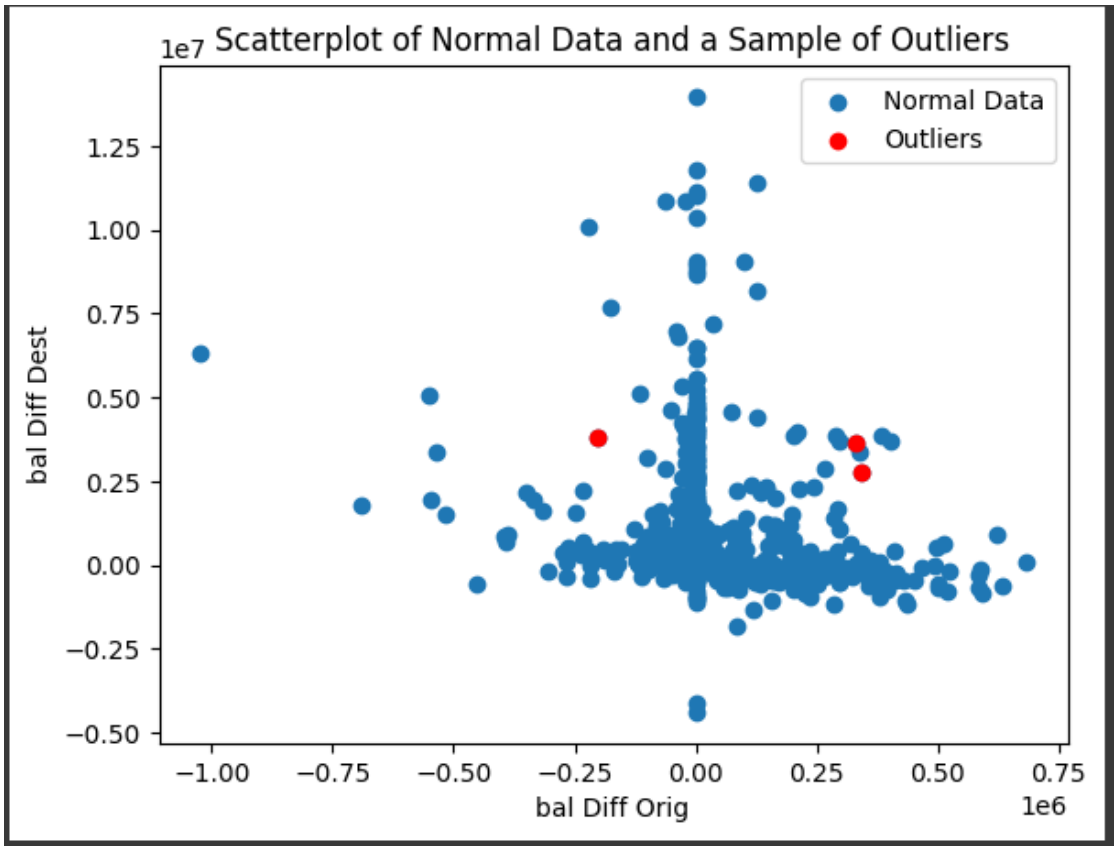
- Explore the proportions of fraudulent and non-fraudulent transactions and visualized using Pie chart



- Handled Multicollinearity by combining the two high features with high Variance Inflation Rate using feature engineering

balanceDiffOrig	balanceDiffDest
-181.00	0.00
-181.00	-21182.00
-5337.77	-1549.21
-4465.00	147137.12
-15325.00	46430.44
...	...
0.00	78635.42
0.00	7270.37
0.00	-71114.64
0.00	89346.61
0.00	138651.85

- Detected outliers and visualized using scatter plot



2. Describe your fraud detection model in elaboration

- I have treated the problem as a supervised machine learning problem using classification with the target variable being a binary categorical variable, namely:

1: Is Fraud,
0: Not Fraud

- The problem can also be identified as unsupervised anomaly detection problem using Isolation forest algorithm and performance metrics as ROC curve

- In my model, I've performed Data Cleaning, handling missing values, handling multicollinearity and detecting outliers.

- Then, I've introduced new features using feature engineering techniques and created 3 new features:

1. balanceDiffOrig: Change in newbalanceOrig and oldbalanceOrg
2. balanceDiffDest : Change in newbalanceDest and oldbalanceDest
3. timeOfDay(hourly): Created using the "step" feature to find the hourly day time

- I have encoded the categorical data present in the dataset using sklearn's LabelEncoder module

- Since the values of various features differed in scale, I used sklearn's Normalizer module to normalize the data into same range

- Splitting the dataset into training and validation(test) sets to check the performance of the model on unseen data

- Applied 3 different algorithms to train the data and compare the results. These are:

1. Logistic Regression
2. Random Forest
3. Support Vector Machines

- After training the model, I've measured the performance of the model using accuracy score and classification report of all three algorithms, which performs extraordinary well with accuracy of 100% in the unseen data

3. How did you select variables to be included in the model?

- The selection process of the features was rather straight forward. First, the bank account details are absolutely critical for this model so the 'oldbalanceOrig', 'oldbalanceDest', 'newbalanceOrig' and 'newbalanceDest' and the 'amount' features are picked

- 'step' variable maps the unit of time in real world, therefore it tells us in what unit of time are the fraudulent transactions occurring the most. It can help us determine the time period for occurrence of majority of fraudulent transactions.

- 'type' variable explains the type of transaction taking place. Different types of transactions may be more susceptible to fraud than others

- 'nameOrig' and 'nameDest' are string values that are used for identifying an individual's transaction. This bit of information can not be used to understand the data and find patterns. Therefore, these variables are dropped.

- Since the problem is treated as a supervised classification problem, the 'isFraud' variable is considered as target variable and is used to make associations with the independent variables

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

The performance of the model is measured using the “accuracy score” and “classification report” with an overwhelming performance of 100% on the unseen data

5. What are the key factors that predict fraudulent customer?

1. **Amount of the Transaction:** High value transactions could be particularly at risk. Fraudsters often aim to maximize their gain from each fraudulent act.

2. **Type of Transaction:** Different types of transactions may be more susceptible to fraud than others. For instance, 'transfer' and 'cash out' types might be more often associated with fraud as they involve moving money to other accounts.

3. **Old/New Balance of Origin/Destination:** Large changes in account balance can indicate fraudulent activity. For instance, if an account's balance drops significantly following a transaction, that could be a sign of fraud. It might also be suspicious if the account balance remains unchanged even after a transaction.

4. **Step (or hour of the day):** Fraudulent transactions can also depend on the hour of the day. There might be certain times of the day when fraudulent activities are more likely to take place.

7. What kind of prevention should be adopted while company update its infrastructure?

Some of the prevention techniques that can be adapted by a company can be:

1. **Data Encryption:** Sensitive data should be encrypted both at rest and in transit. This ensure that even in event of a data breach, the data remains unintelligible and useless to attackers.

2. **Authentication:** Tech giants such as Google, incorporate strong authentication and authorization policies such as two-factor authentication (2FA) or multi-factor authentication (MFA)

3. **Fraud Detection Systems:** Building machine learning-based fraud detection systems. Real-time fraud detection can help identify and prevent fraudulent transactions as they occur.

8. Assuming these actions have been implemented, how would you determine if they work?

- Continuous Monitoring and logging activities across the network can help identify any unusual behavior and assess whether the preventative measures are functioning as expected.

- Building bigger neural network with more hidden layers to improve the likelihood of predicting fraudulent behavior.