

Finance Based Model: Financial Fraud Detection

1. Problem Statement: Apply various classifiers to detect fraud on the mobile money transaction dataset.

- 1.1. Logistic Regression
- 1.2. KNN Algorithm
- 1.3. Gaussian Naive Classifier
- 1.4. Decision Tree
- 1.5. Random Forest Classifier

2. Dataset used:

We have download our dataset from from :

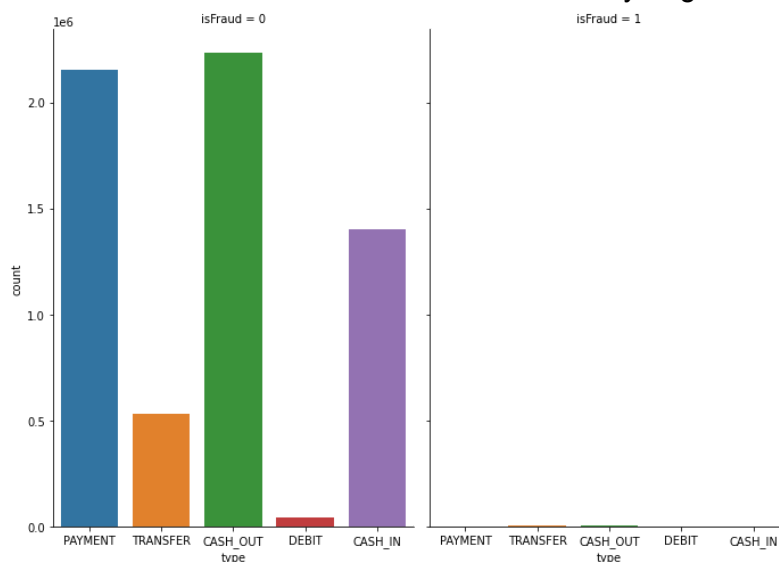
<https://www.kaggle.com/kartik2112/fraud-detection-on-paysim-dataset/data>

It is a mobile money transaction dataset which consists of a total of 63,62,620 datas and 11 columns.

3. Procedures:

Following steps are implemented:

- 3.1. Importing required libraries.
- 3.2. Reading the dataset and exploring its different attributes.
- 3.3. Exploratory Data Analysis (EDA):
 - 3.3.1. Checking for null values.
 - 3.3.2. Various plots to check the different counts of data under binary target arritube (isFraud)



3.3.3. Data Cleaning

3.3.4. Undersampling

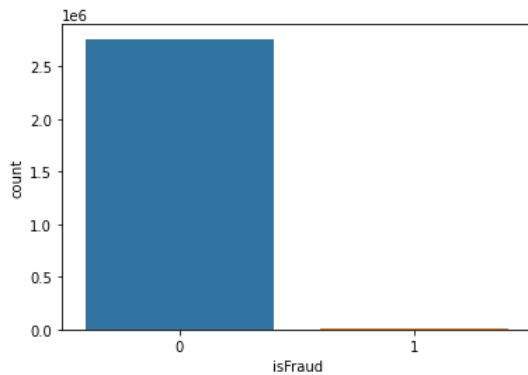


Fig: Data distribution before sampling

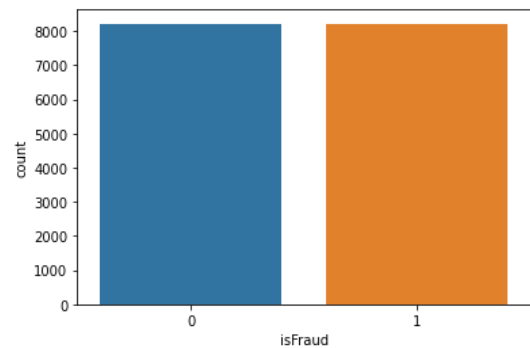
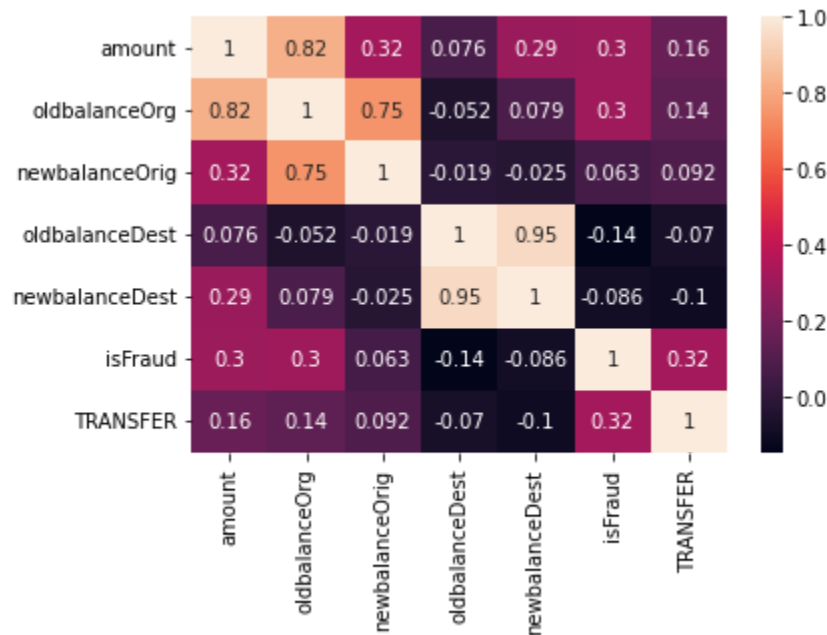


Fig: Data distribution after sampling

3.3.5. One Hot Encoding

Changed categorical data “Transfer and Cashout” as a new binary column “Transfer” where 1 is for type Transfer and 0 for cashout.

3.3.6. Heatmap plot to find the correlation between different attributes

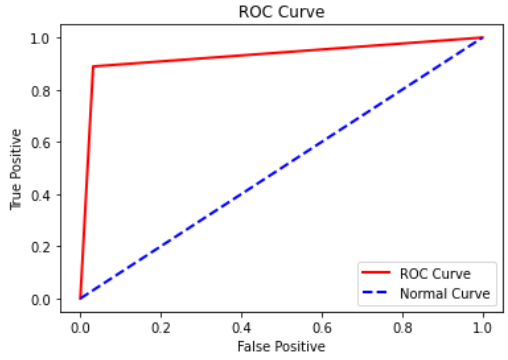
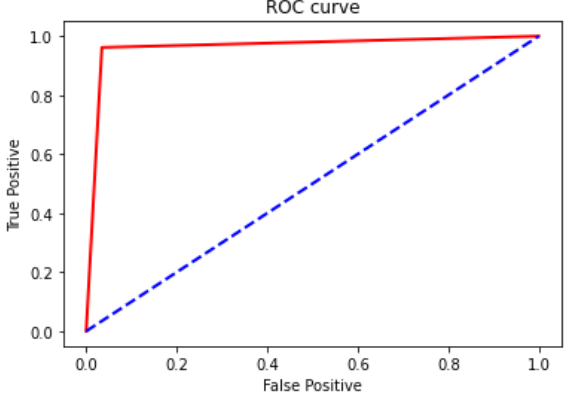
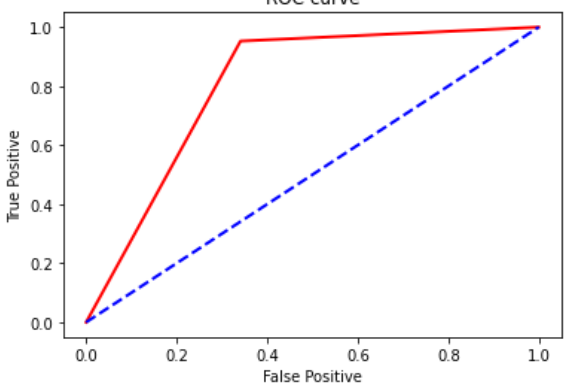


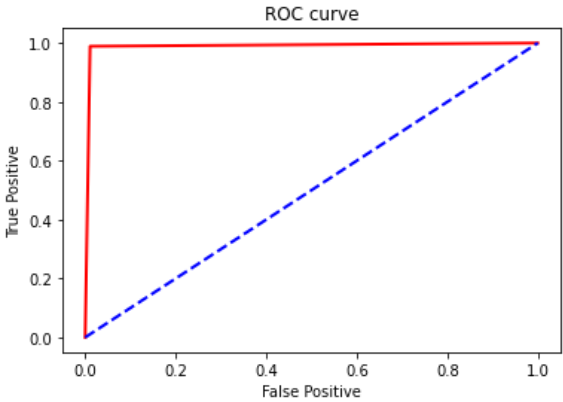
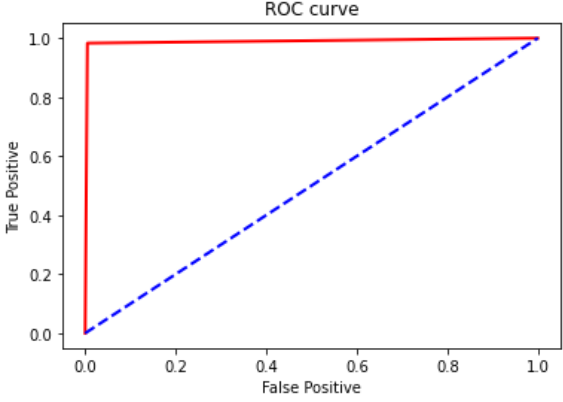
A value close to 1, shows that the columns are highly correlated and value 0 indicates there is no correlation between the corresponding columns.

4. Divided the data frame into X(containing feature columns) and Y (Target Column “isFraud”).
5. Train Test Split (Splitted the data in the ratio 70% (train) and 30% (test)).
6. Used different classifiers to train the model and checked their respective accuracy over both train and test data.

4. Code : Jupyter notebook has been attached in the file containing the codes

5. Observation

Algorithms Used	Accuracy	ROC Curve
Logistic Regression	92.87%	 <p>The ROC Curve for Logistic Regression shows a red solid line (ROC Curve) that rises sharply from (0,0) to approximately (0.05, 0.9) and then gradually approaches the top-right corner. A blue dashed line (Normal Curve) represents a random classifier. The y-axis is labeled 'True Positive' and the x-axis is labeled 'False Positive', both ranging from 0.0 to 1.0.</p>
KNN Algorithm	95.34%	 <p>The ROC curve for the KNN Algorithm shows a red solid line (ROC curve) that rises very steeply from (0,0) to approximately (0.05, 0.95) and then remains nearly horizontal at a True Positive rate of 1.0. A blue dashed line (Normal Curve) represents a random classifier. The y-axis is labeled 'True Positive' and the x-axis is labeled 'False Positive', both ranging from 0.0 to 1.0.</p>
Gaussian Naive Classifier	80.58%	 <p>The ROC curve for the Gaussian Naive Classifier shows a red solid line (ROC curve) that rises from (0,0) to approximately (0.35, 0.95) and then remains nearly horizontal at a True Positive rate of 1.0. A blue dashed line (Normal Curve) represents a random classifier. The y-axis is labeled 'True Positive' and the x-axis is labeled 'False Positive', both ranging from 0.0 to 1.0.</p>

Decision Tree	98.904%	 <p>ROC curve</p> <p>The plot shows True Positive Rate on the y-axis and False Positive Rate on the x-axis, both ranging from 0.0 to 1.0. A red solid line represents the classifier's performance, which is nearly vertical at x=0 and horizontal at y=1. A blue dashed line represents the random classifier's performance, running diagonally from (0,0) to (1,1).</p>
Random Forest Classifier	98.906%	 <p>ROC curve</p> <p>The plot shows True Positive Rate on the y-axis and False Positive Rate on the x-axis, both ranging from 0.0 to 1.0. A red solid line represents the classifier's performance, which is nearly vertical at x=0 and horizontal at y=1. A blue dashed line represents the random classifier's performance, running diagonally from (0,0) to (1,1).</p>

6. Conclusion:

Based on our observation, we found that “Decision Tree” and “Random Forest Classifier” give the maximum accuracy. So we can use these two classifiers to train our model and use that model for further fraud detection.