**Name: Vishwajeet Jagtap**

**Rollno: 230340325079**

**Objective**:

The main objective is to create a predictive model that can accurately classify fradulant transactions

**Steps followed :**

1. Importing Data, Inspecting the Dataframe

2. Data Preparation

3. EDA (univariate analysis, outlier detection, checking data imbalance, checking correlation)

4. Applying Encoding to convert categorical into

5. Balancing the dataset using hybrid sampling

6. Test-Train Split

7. Feature Scaling

8. Building Model

9.Checking Accuracy scores for different models

**Detailed Solution (Implementation done in Python):**

• We are using Fraud.csv dataset for this case study.

• Classification Model and evaluating final model’s performance.

• Looking at the dataset we observed that following this to be done:

**Numeric column Analysis:**

1)We have done statistical analysis of dataset.

2)We observed the correlation between independent variables and according to that dropped columns

3)While observing data we got to know that there is a column which is isFlaggedFraud which is not flagged correctly actually so we dropped that column as it was not giving correct information

4)Also we found that transfer and cashout has more fraud cases

**Categorical Column Analysis:**

1)There are categorical columns Type which is converted into categorical column.

2)There are also column nameOrig and nameDest which needs to be dropped as it is not giving any info. So we are dropping those columns.

3) There is need for conversion for type as it is categorical column so we will apply onehot encoding.

**Balancing:**

During our analysis we found that our data is imbalanced since our target column 0 value has more weightage as compared to 1. So if we look at ratio it is imbalanced.

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations.

So we need to balance it

Method :

To solve the problem of sample imbalance in medical dataset, we propose a hybrid sampling algorithm combining synthetic minority over-sampling technique (SMOTE) .

Procedure:

1)After that we go for feature scaling. In feature scaling we have used standard scaler. Generally we

use feature scaling when our data is containing lot of variances and it is necessary for distance based algorithms like KNN.

2) After that we go for model building

3)These are the some of the algorithms which are suitable and we tried for this dataset

Models :

1)Logistic Regression

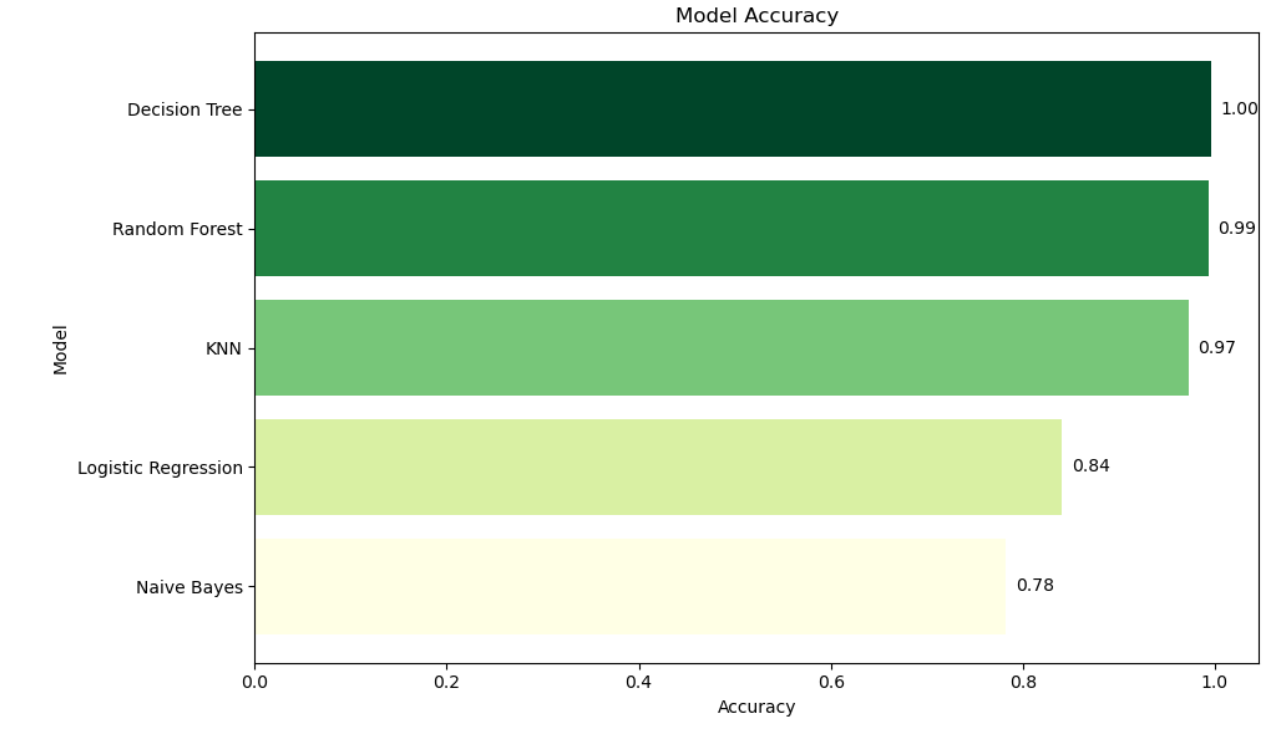
2)Decision Tree

3)Random forest

4)Naïve Bayes

5)KNN

These are the following accuracy results we have achieved:



From above graph we can conclude that Decision Tree,Random Forest,KNN are overfit to models and Naïve Bayes has avg accuracy so we will choose Logistic Regression model. Because Neither it is overfitting nor underfit. So we can go for Logistic Regression model.

These are the following Results and classification Reports of the models:

Logistic Regression:

A screenshot of a graph

Description automatically generated

Decision Tree:

A screenshot of a computer program

Description automatically generated

Random Forest:

A screenshot of a report

Description automatically generated

Naïve Bayes:

A screenshot of a computer screen

Description automatically generated

KNN:

A screenshot of a graph

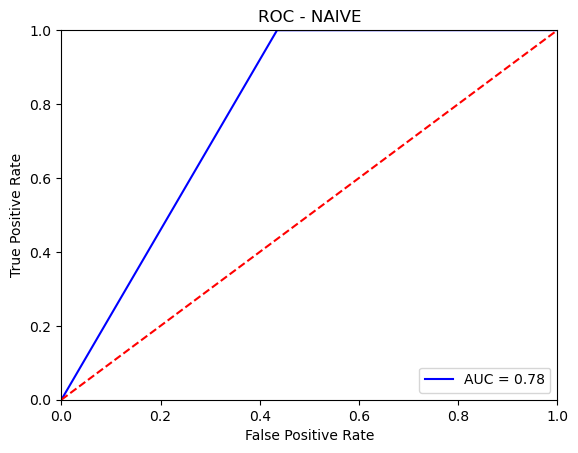
Description automatically generated

Roc curve:

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

A blue line with red lines

Description automatically generated



A graph with a red line

Description automatically generated

A graph with a red line

Description automatically generated

A graph with a red line

Description automatically generated

Conclusion:

As our data is imbalanced we tried to balanced it but when we applied balanced dataset to some of the algorithms. They are prone to overfitting. We are getting logistic and naïve bayes as a generalized model. The choice of the final model should consider not only on its performance metrics but also other factors such as interpretability, computational efficiency, and model complexity.