# DSC 550 DATA MINING TERM PROJECT

*Credit Card Fraud Detection*
*Bellevue University*

*DSC 550 DATA MINING*
*DIPIKA SHARMA*
*WEEK 10*

## Introduction

### 1. Introduce the problem

*The world we are living in, is undeniably digital. Since covid 19 there is a tremendous increase in the number of users who prefer online shopping as it is convenient and easy. Most of the people use credit cards for online purchase. Credit Card offers lot of the benefits to users and one of the advantages of credit card is - it allows users to buy something even if they do not have money at that time. This feature has also increased the financial fraud. Although cybersecurity is there and plays an important role in providing digital security, but it is not easy to track down the unusual activity.*

### 2. Justify why it is important/useful to solve this problem

*Credit card fraud detection is very important for any financial organization or banks. By identifying the fraud transactions, we are helping consumer so that they will be not charged for any fraud transaction which are not done by them. With the help of this project, we will try to solve the credit card fraud by detecting the fraud transactions. We will create models to classify the fraud and non-fraud transaction and then we will see which model better fit the problem for better result.*

### 3. How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

*The banks or financial organization are the main stakeholders for the credit card fraud detection project. While allocating the credit cards to the consumer they go through the approval process and promise them the easy and secured life where they can use this credit cards for their routine. Most of the consumers who do not have money at time of purchase use credit card and pay installment every month to pay back to bank or financial organization. By detecting the fraud transaction, we are not only making consumers happy but also increasing the trust of the consumer which eventually will help any organization to grow.*

*Using the machine learning model, we can classify non fraud and fraud transaction and use fraud transactions data to answers some of the important questions like - At what time most of the fraud transaction are happening, how much money deducted in fraud transaction, and others.*

## 4. Explain where you obtained your data

*I found the credit card dataset at below location in Kaggle website:*

*https://www.kaggle.com/datasets/jacklizhi/creditcard*

*This data set contained 284807 rows and 31 columns. I understand for security purpose the dataset has column names v1, v2, … v28. We have 28 features with these names. The remaining three features Time, Amount and Class using the original names. Class is the target columns when class labels is 0 then it is non-fraudulent transaction and when it is 1 then it is fraudulent transaction. Below are the different features available in the dataset:*

```
Time      float64
V1        float64
V2        float64
V3        float64
V4        float64
V5        float64
V6        float64
V7        float64
V8        float64
V9        float64
V10       float64
V11       float64
V12       float64
V13       float64
V14       float64
V15       float64
V16       float64
V17       float64
V18       float64
V19       float64
V20       float64
V21       float64
V22       float64
V23       float64
V24       float64
V25       float64
V26       float64
V27       float64
V28       float64
Amount    float64
Class       int64
dtype: object
```

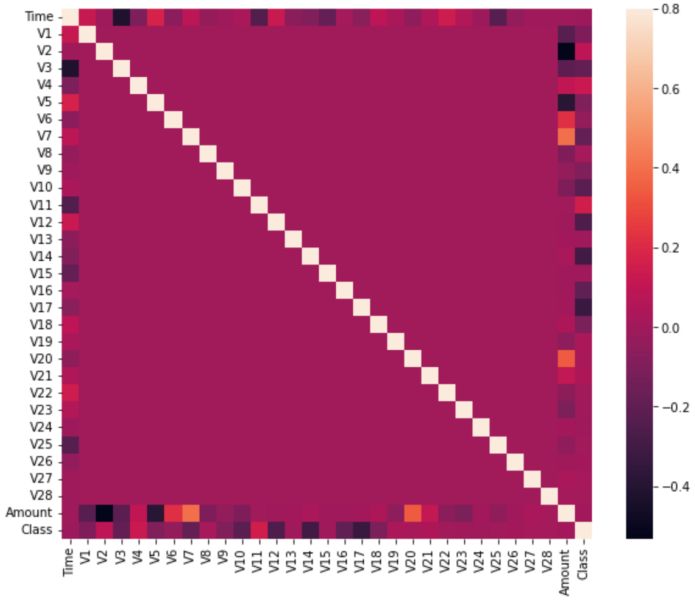
# Organized and detailed summary of Milestones 1-3

## 1. EDA: include any visuals you think are important to your project

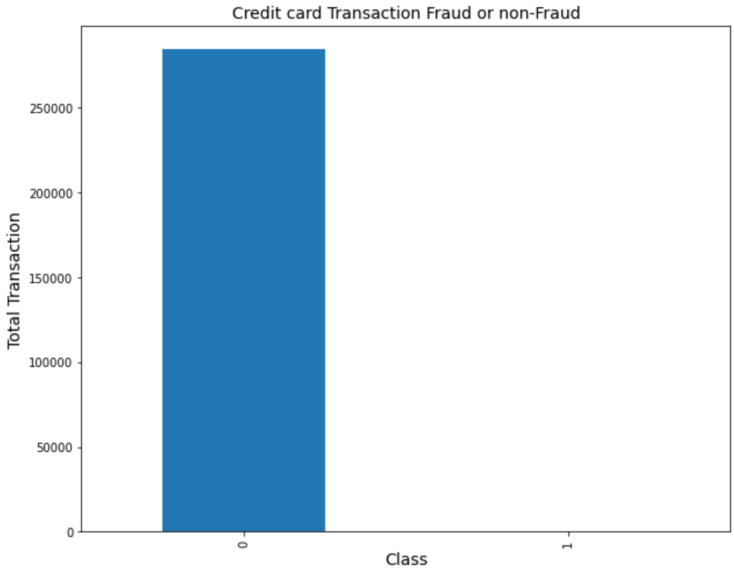
*While working on the credit card data set, I have following observation:*

- *The dataset has 284807 rows and 31 columns.*

- *Using the describe function I have noticed that there is no missing values or Null values in the dataset.*
- *Using the heatmap below we can clearly say that v7 and v20 features are positively correlated to Amount. Class and Amount is negatively correlated to v3. Also, we can see that V2 and V5 are negatively correlated to Amount. The features V1, V2, … and V28 are weakly correlated to each other. Class is negatively correlated to v14, v17 and v12.*
  *The features V1, V2, … and V28 are weakly correlated to each other, and they do not have much dependency on others. But we do find that Class and Amount variable are positively and negatively correlated to some of the features like V2, V3, V5, V7, V14, V17 and V12.*

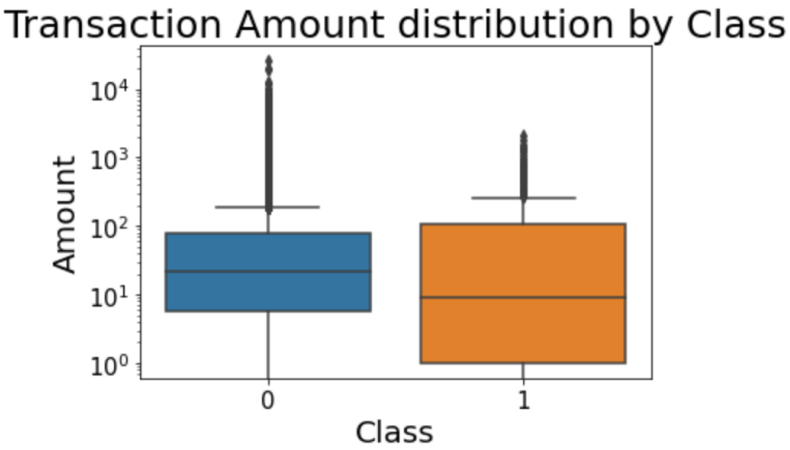



- *The bar graph is used to represent the categorical feature of the dataset with rectangular bars. Used the bar graph to understand the feature "Class" of our dataset. Also, we know class labels as 0 for non-fraudulent transactions and 1 for fraudulent transactions.*



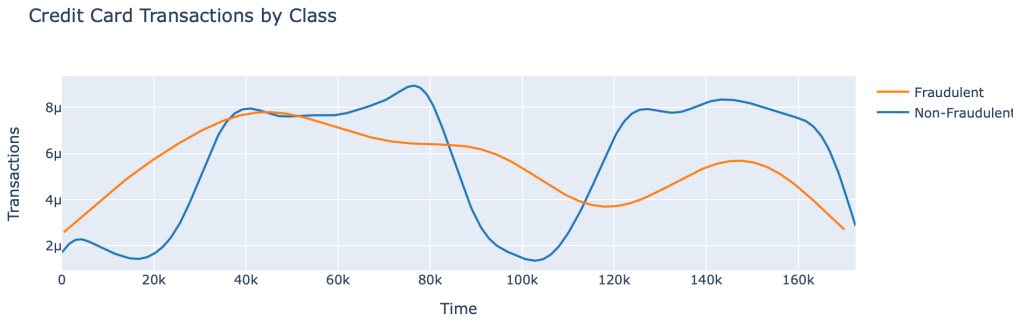

*Used bar graph analysis to find out the number of fraud transaction against the total number of transactions. We saw that the fraud transaction is very less compared to valid non fraud transactions. This also raise a question if the dataset is best fit or not as the training set is seems to be very small.*
*The bar graph show instability in the credit card dataset.*

- *Boxplot helps us to identify the outliers and gave us better understanding of outliers that whether it is required to remove outliers from the data or not. Outliers are considered to be any value coming outside of the 1.5 times of the Inter-Quartile range. But with credit card data set where the fraud transaction is only 17% compared to the total transactions it would be not a good idea to remove all the outliers of the data. So instead of removing all the outlier's data points we will be focusing only on the extreme outliers.*



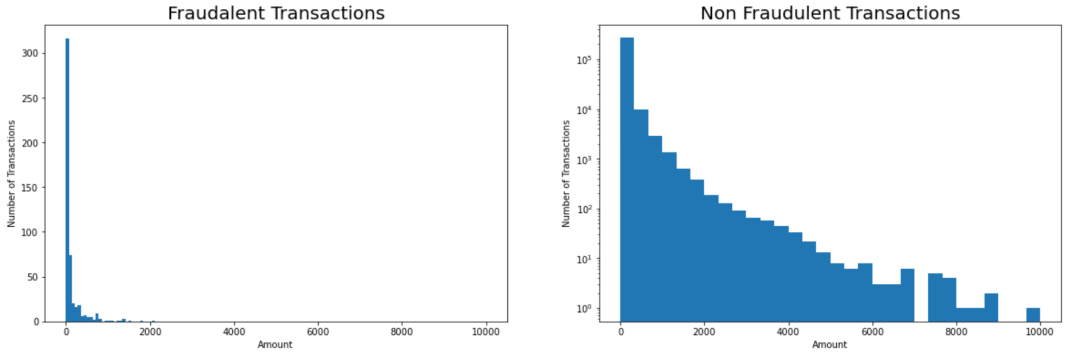
*With help of boxplot will be finding out the outliers in the data set for amount variable.*
*Looking at the box plot we can Cleary see that we do not have outliers for fraud transactions after 3000 but outliers in non-fraud transaction is quite visible and present even after 10000. Let's consider any amount beyond to 10000 as extreme outliers and we will be creating the dataset after removing the extreme outliers.*

- *Used the density plot to see pattern in fraud and valid non fraud transactions.*




*I observed that fraud transactions are quite consistent compared to non-fraud transactions.*

- *Next, I used histogram to find out if there is any unbalanced in amount data when comes to fraud transaction and non-fraud transaction.*




*Each above graph is right skewed, and the Amount withdraws with Fraudulent transactions is very less compared to non-Fraudulent transactions. This also indicate that with respect to Class variable, the Credit card data is unbalanced.*

*I also noticed that zero amount transaction happened for both fraud and non-fraud transactions. although the number of transactions with no amount is very less but it indicates the presence of some data errors in the dataset.*

## 2. Data preparation

*Data Preparation is an important step in data analysis as this process helps us to clean the data and prepare the dataset for evaluation phase that is building the model.*

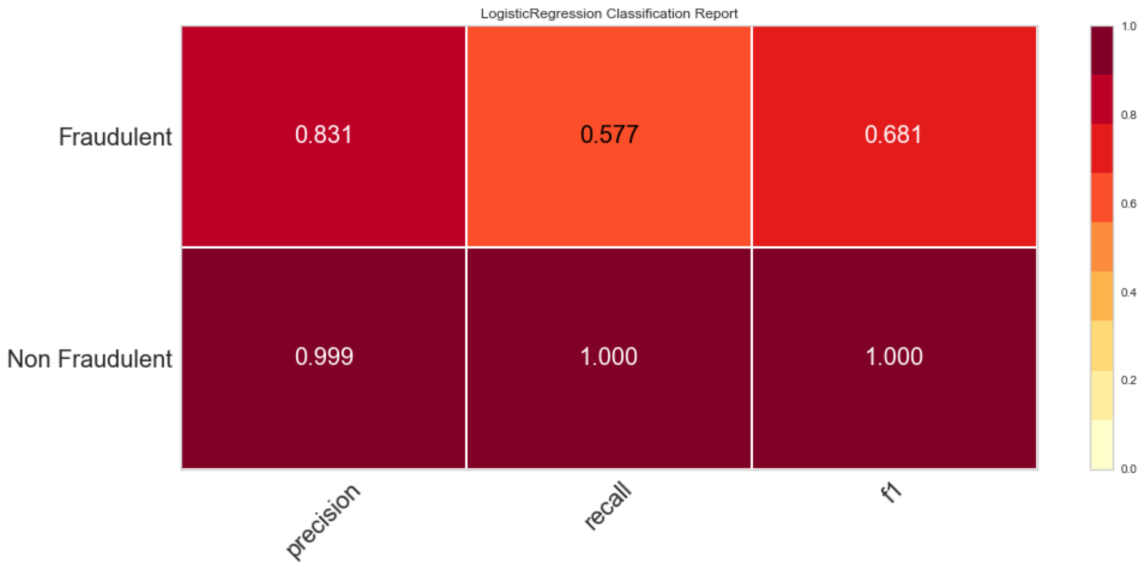
*As part of data preparation, we have done the following in the credit card dataset.*

- *Drop the irrelevant feature from the dataset.*
  *I drop the "Time" feature from the dataset as this column will not help in building the model.*
- *Taking care of NULL or Missing values.*
  *It is important to deal with any missing or null values in the credit card dataset before building the model.*
  *Used the info() and isnull functions I have noticed that there is no missing or null values in the dataset.*
- *Transform the dataset to remove outliers and scale them using statistics that are robust to outliers.*
- *Added the categorical feature in our dataset using the Class column, as we know in Class feature all 0 values represent non-Fraudulent transactions and 1 represent the Fraudulent transactions.*
- *After preparing the dataset, I have split the credit card dataset using Pareto Principle Split into 80% training and 20% testing.*

## 3. Model building and evaluation

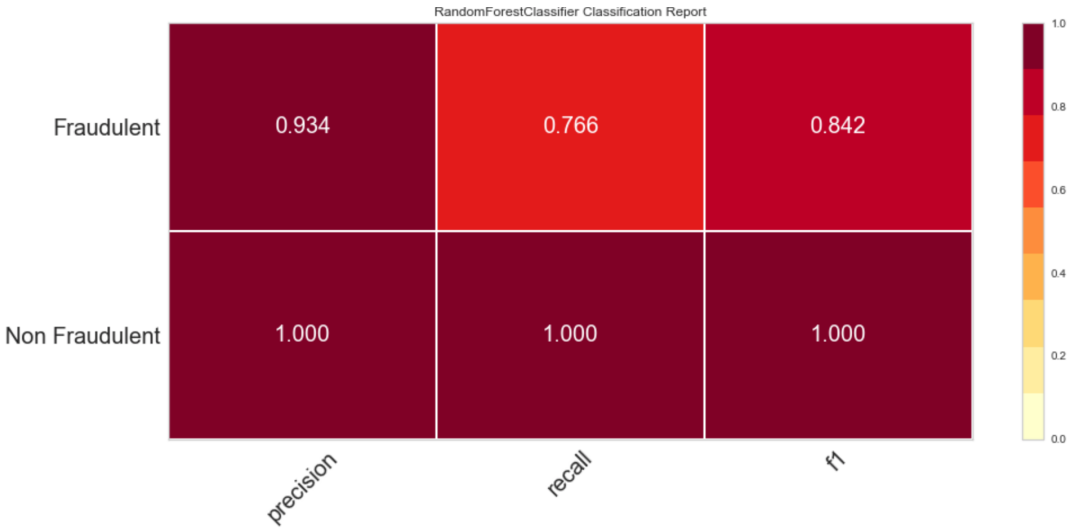
*Credit card fraud detection problem is binary classification problem as we have to detect only 2 class transactions - fraud or non-fraud. A variety of algorithm can be used to solve this problem but let's train Logistic Regression model and Random Forest Classifier on our dataset and see which algorithm is better fit for our problem.*

- *Logistic Regression Model*
  *Using Logistic Regression Model, we can predict whether the transaction is fraud or not.*




*F1 score is 68%, the higher the F1 score means higher the accuracy of the model and it is considered to be better model. 68% is not as good and not bad either. Let's try Random Forest classifier to achieve better accuracy of the model.*

- *Random Forest Model*
  *One of the popular algorithms in machine learning is Random Forest. It is a collection of decision tree called "forest" which trained by combining different improved models. Random Forest classifier give more precise and reliable result as it is created using several decision trees.*




*I have noticed that the Random Forest Classifier is running for longer time because Data-set values are large in number but the F1 score is 84.2 % which is better score compared to Logistic Regression Model.*

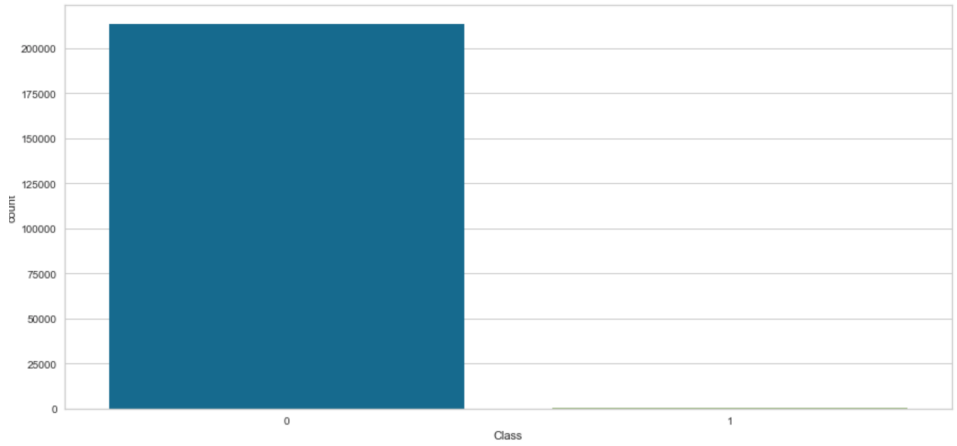
- *Class-Imbalance issue*

  *The above analysis clearly state that Random Forest classifier works better than the Logistic Regression model. Also, while preparing the dataset for building model, we also observe that the existing dataset class is highly imbalance with 99% of non-fraud transactions and only 0.17% of fraud transactions.*

  *Hence it is really important to address the class-imbalance issue before training the model otherwise most of the positive prediction will come from non-fraud transactions.*

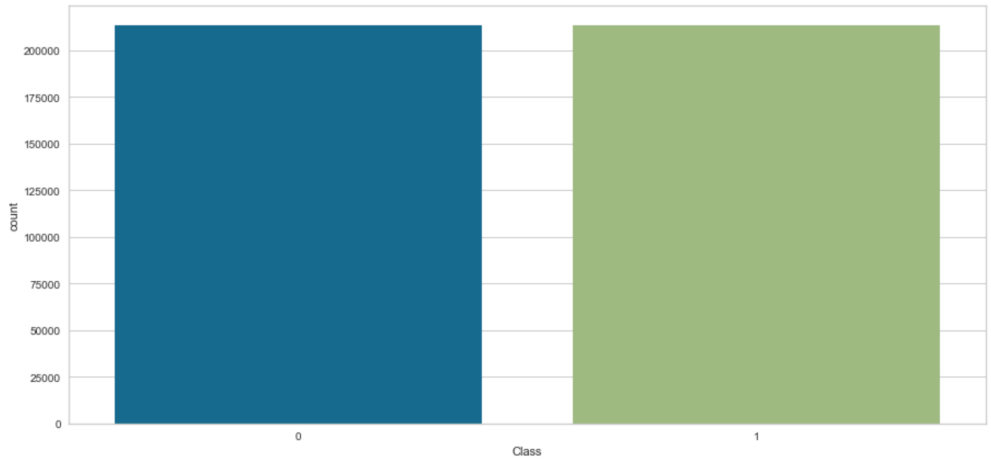
  *Various technique can be used to solve the imbalance issue. We will be using the SMOTE - Synthetic Minority Oversampling Technique for data augmentation of the minority class of fraud transactions.*

  *Let's plot the class to see how the values are distributed.*




  *We can see from above plot that the data is not balanced, and we have only few Fraudulent transaction compared to non-Fraudulent transactions in our dataset.*

  *Let's see how the class values distributed after using the SMOTE oversampling technique on our training dataset.*



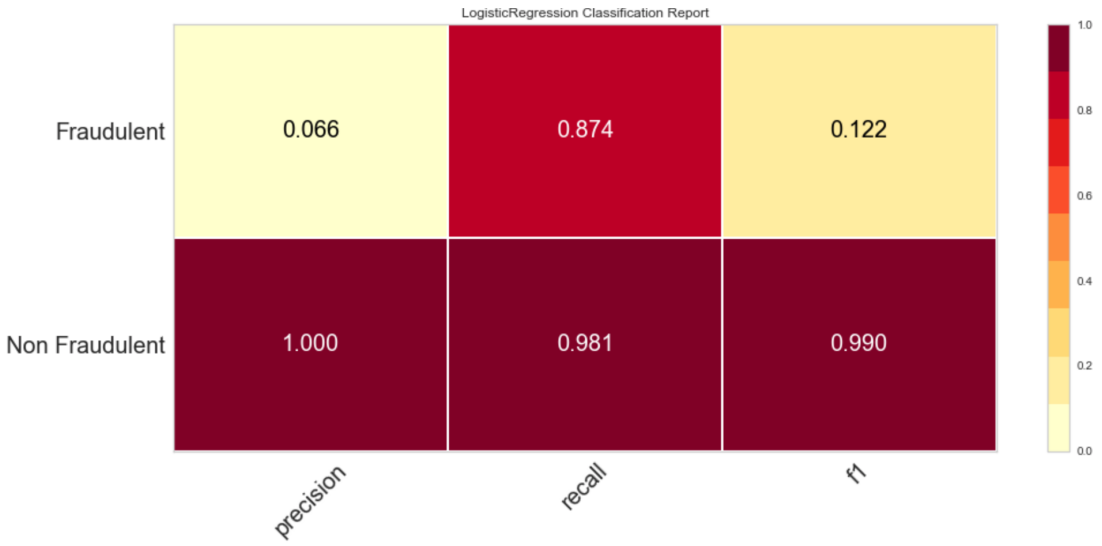
*The above plot shows random increase in fraudulent transactions in training set.*

- *Logistic Regression Model after SMOTE*

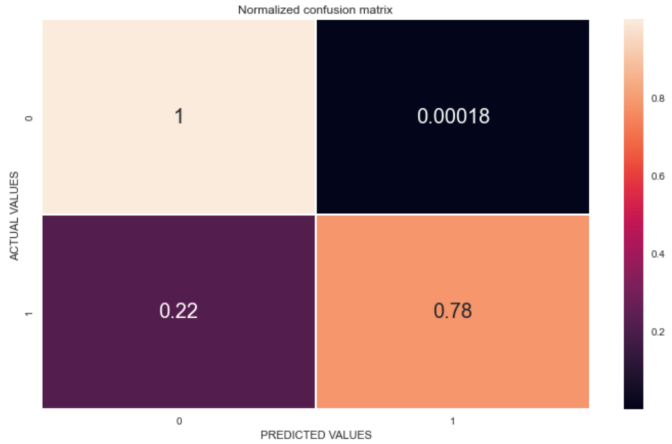
  *Although the accuracy of the model is 98% but the F1 score is very low as .12 for Logistic Regression model. The model performed not well because of the large number of samples.*

```
F1 Score is : 0.1220125786163522
Logistic Regression Model Accuracy is : 0.9803938091626639
Area under curve score for Logistic Regression Model is:0.9272170005174183
```

*Even after balancing the dataset, the logistic regression model does not seem to be a better fit for predicting the Fraudulent transaction in dataset.*

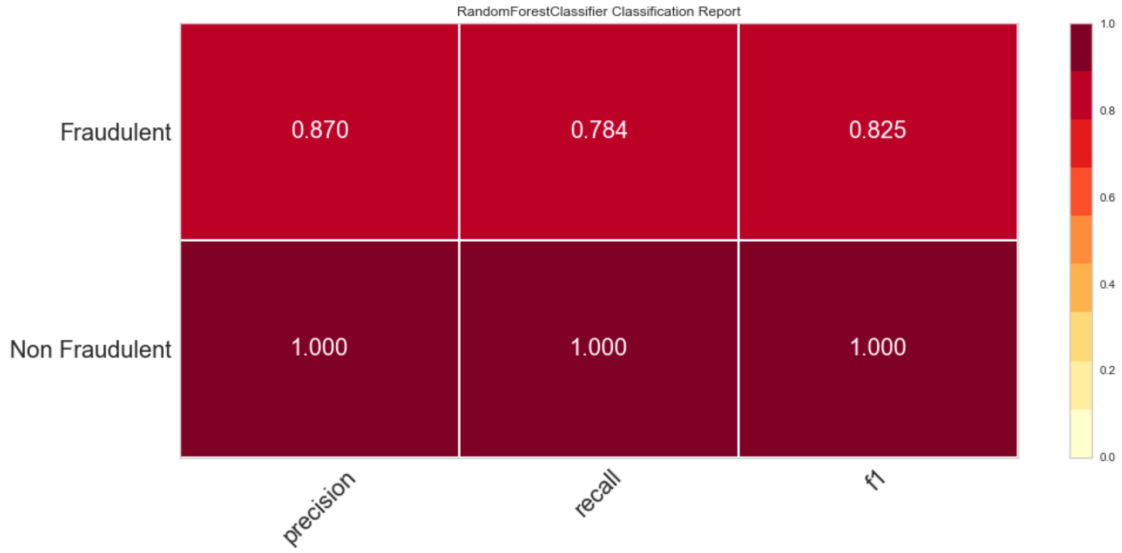



- *Random Forest Model after SMOTE*
  *Looking at the below Random Forest classifier confusion matrix we can see the model correctly classified .78 of the non-Fraudulent transaction.*



*We can see below the Random Forest Model show 99.9% of model accuracy with F1 score of 82 %. It shows model can correctly predict non fraudulent and fraudulent transactions.*

```
F1 Score: 0.8246445497630331
RandomForest Classifier Model Accuracy is : 0.9994803516755147
Area under curve score for RandomForest Classifier Model is:0.8918004597837487
```

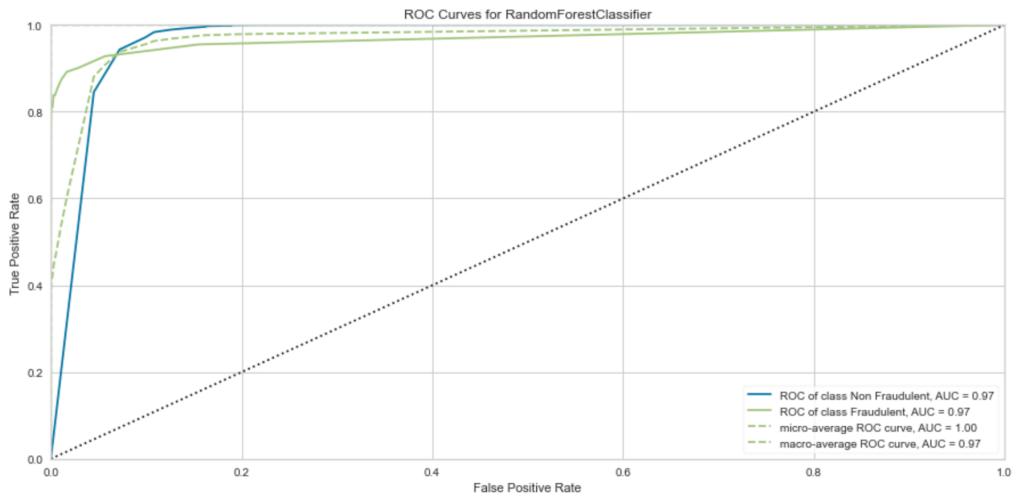



*The above is to plot the F1, precision, and recall and see model performed good with 82.5 % of F1 score.*

## *Conclusion*

### *1. What does the analysis/model building tell you?*

- *The Logistic Regression model 98% accuracy and Random Forest classifier model shows 99.9% accuracy.*
- *Both models performed really well but the F1 score of Logistic regression model is very less as .12 whereas for Random Forest model the F1 score is .86.*
- *The trained Random Forest model performed much better on the test data set compared to Logistic Regression model.*
- *Random Forest model shows 96% of the Area Under the ROC curve value with 99.9% accuracy and .86 F1 score meaning the model build above can correctly classify Non-Fraudulent and Fraudulent Transactions.*




- *ROC curve above state that the Random Forest model has strong predictive power and able to predict non fraudulent and fraudulent transactions.*

- *The classification report and the ROC curve suggest that the Credit card fraud detection model built using a Random Forest classifier shows maximum accuracy of 99.9% and AUC of 96%. The accuracy and AUC results suggest that the model performed well and best suited for our project. This model can be efficiently used in identifying the credit card fraud transactions and will help financial organization and consumers from any fraud activities.*

## 2. Is this model ready to be deployed?

*The built Random Forest classifier model is performing good and shows good accuracy in identifying the fraudulent transactions while maintaining a low false positive rate. But there are some other features that I think we need to explore before deploying this model in production. The dataset we used for our project is missing information like Retailer, transaction type – Online, pin number, or chip.*

*I think it will be good idea to study our dataset with some other dataset also which has above mentioned features. It would be interested to see if the transaction happened from same retailer. Or is the transaction happened using pin number or credit card or is it an online transaction.*

## 3. What are your recommendations?

*To make this Random Forest model ready to deploy, I think we can look for another dataset that can be merge with the existing dataset and we can analyze Retailer and transaction type information of the fraudulent transactions. Also, if get dataset including location information like where the fraudulent transaction happened, is it nearby to retailer or consumer. Are location changes when another fraudulent transaction happened on the same credit card.*

*Information like above will help us to understand the data and we will be able to build robust model to identify the fraudulent transactions.*

## 4. What are some of the potential challenges or additional opportunities that still need to be explored?

*The major challenge I faced was when I was preparing the data before building the model. As we know that the dataset has only few .17% fraudulent transaction compared to 99.8% non-fraudulent transaction. To address the imbalance issue, I have performed SMOTE and was able to build model which efficiently identify fraudulent and non-fraudulent transactions.*

*But I would say the built model can be further improved if we can have more fraudulent transaction to identify the patterns.*

*Also, if we can identify at what time or duration most of the fraudulent transaction happening, Retailer with most fraudulent cases, and Transaction type. All these information will help us to build the robust model to identify fraudulent transaction.*

*We have Time feature in current dataset which we can explore more and try to predict at what time the fraudulent transaction will happen. Adding this significant feature will improve the existing model.*