# Micro Credit Project


**Project By -**

**Amar Kumar**

# Problem Statement

- A Microfinance institution (MFI) is an organization that includes the provision of financial services such as Group Loans, Agricultural Loans, Individual Business Loan, savings, insurance etc. to low income individuals.

- A Telecom Industry is collaborating with an MFI to provide micro-credit on mobile balances to low income families and poor customers that needs be paid back within the time duration of 5 days.

- Therefore, identification and management of customers paying back the loan within the time duration is very important.

# Objective and Motivation

- Microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.
- Because micro-credit is aimed at the poorest, micro-finance lending technology needs to mimic.
- Default is the inability to repay the loan by either failing to complete the loan as per the loan agreement or neglect to service the loan
- Hence, to improve the selection of customers for the credit, building a machine learning model to make predictions (fraudulent or not) will be the best way to help the organizations in further investment and selection of customers.

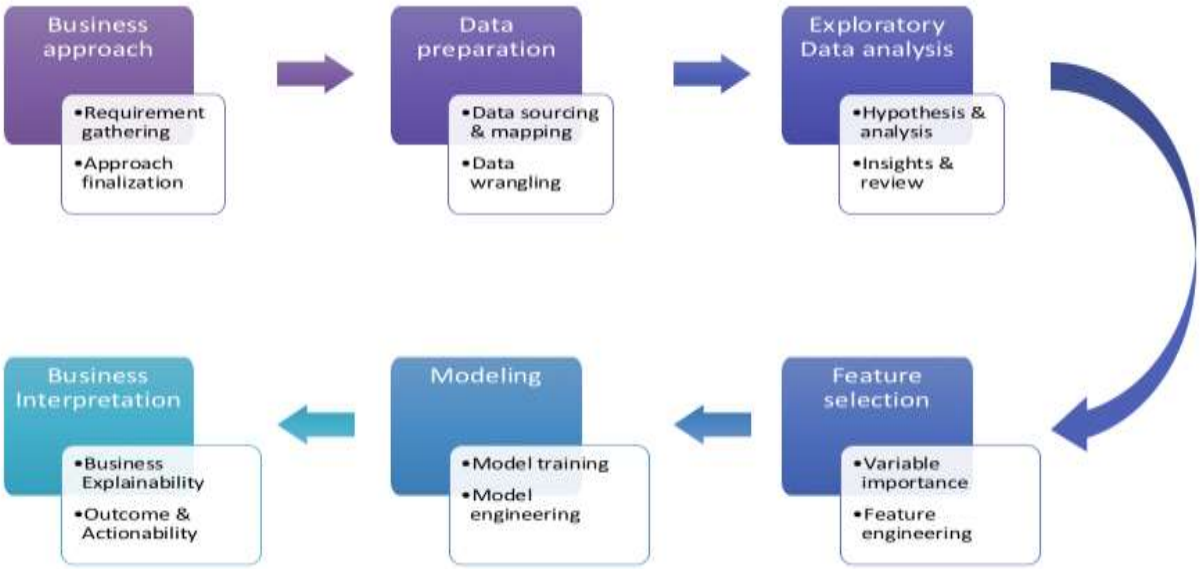
# Research and Solution

- A huge dataset of 209593 records and 37 attributes of customer behaviour is taken and used to analyze the deformities and abnormalities leading to frauds in Microfinance Organizations.
- The dataset is resized and transformed using various pre-processing steps and applying EDA onto the Jupyter Notebook through which the data can be easily accessed.

- We aim at implementing a model by performing various steps –

1. Business Understanding
2. Performing EDA
3. Training 4/5 ML models to test metrics of various models such accuracy, precision, recall and f1 score
4. Hypertuning the models
5. Selecting the best fitted model for data.

- Training is performed by using oversampling techniques to deal with the imbalance nature of the Target Variable.
- The adjustment of high range between values is performed by normalizing the values using Standard Scaler.
- Output is a real number(accuracy) ranging from 0 to 1 indicating strength of a Target Variable.



# Scope of Project

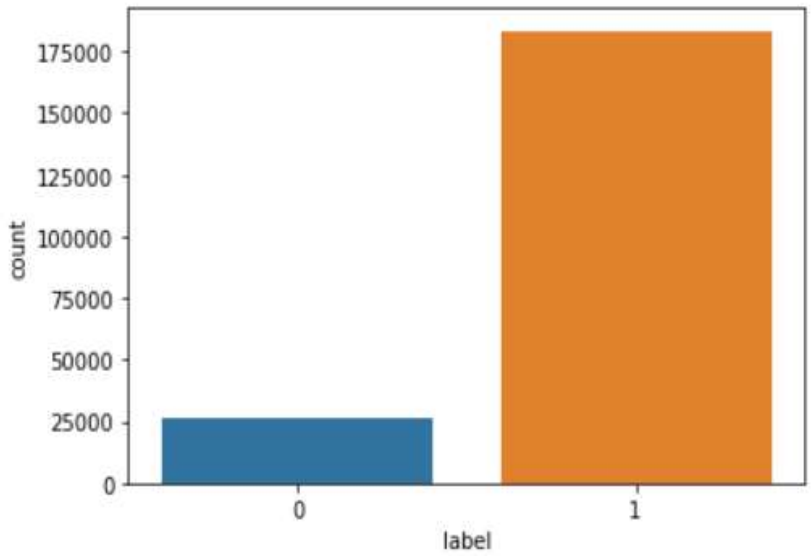
- The project aims at training a Machine Learning model to predict or identify whether the consumer is fraudulent or not i.e. paying the loan amount within the time duration or not using the dataset that consists of the behavioural patterns/details of the customer.

- The project is strictly limited to binary class labels 0 & 1 wherein 0 indicates that the loan has not been paid (fraudulent) and 1 indicates that the loan has been paid (not-fraudulent).

- The model works for fraud detection of credit-loans but can be further improved to work on more customers and different attributes.

# Target Variable

```
1    183431
0     26162
Name: label, dtype: int64
```




- Class “label” is the Target Variable.
- The nature of the Target Variable is imbalanced wherein the Label ‘1’ has approximately 87.5% records, while, Label ‘0’ has approximately 12.5% records which is 183431 & 26162 respectively.

# Brief Description of the Features

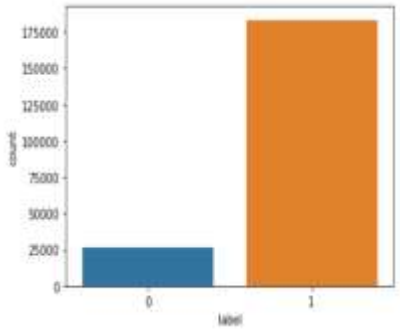
## Categorical

- label
- pcircle
- Pyear
- medianamnt_loans90
- maxamnt_loans90
- medianamnt_loans30
- pmonth

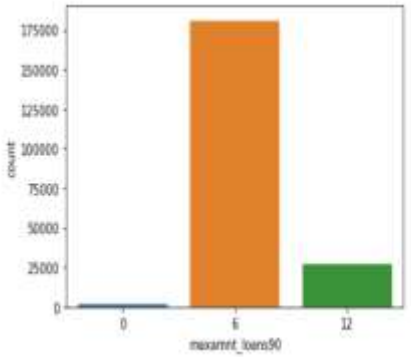
## Continuous

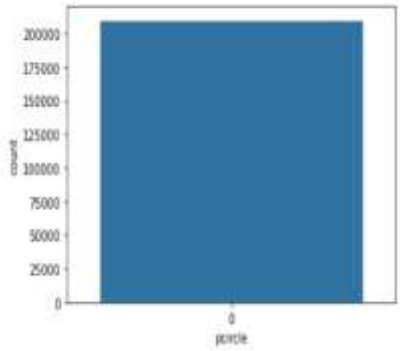
- aon
- daily_decr30
- daily_decr90
- rental30
- rental90
- last_rech_date_ma
- last_rech_date_da
- last_rech_amt_ma
- cnt_ma_rech30
- fr_ma_rech30
- sumamnt_ma_rech30
- medianamnt_ma_rech30
- medianmarechprebal30
- cnt_ma_rech90
- fr_ma_rech90
- sumamnt_ma_rech90
- medianamnt_ma_rech90
- medianmarechprebal90
- cnt_da_rech30
- fr_da_rech30
- cnt_da_rech90
- fr_da_rech90
- cnt_loans30
- amnt_loans30
- maxamnt_loans30
- cnt_loans90
- amnt_loans90
- payback30
- payback90
- pdate

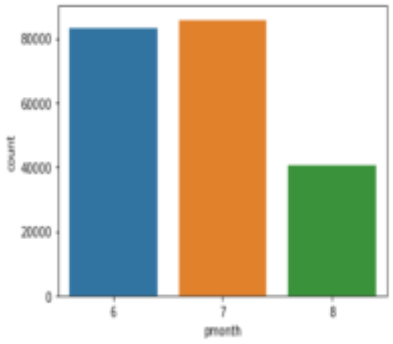
# Dashboard of Categorical Features

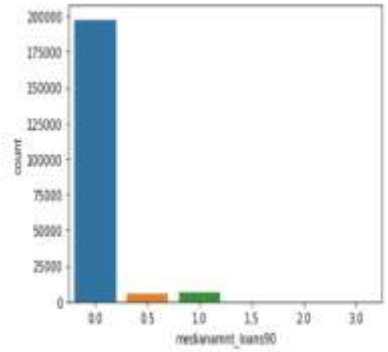


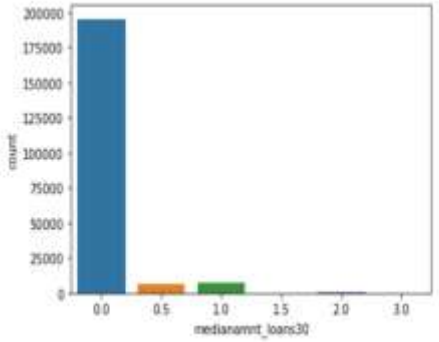




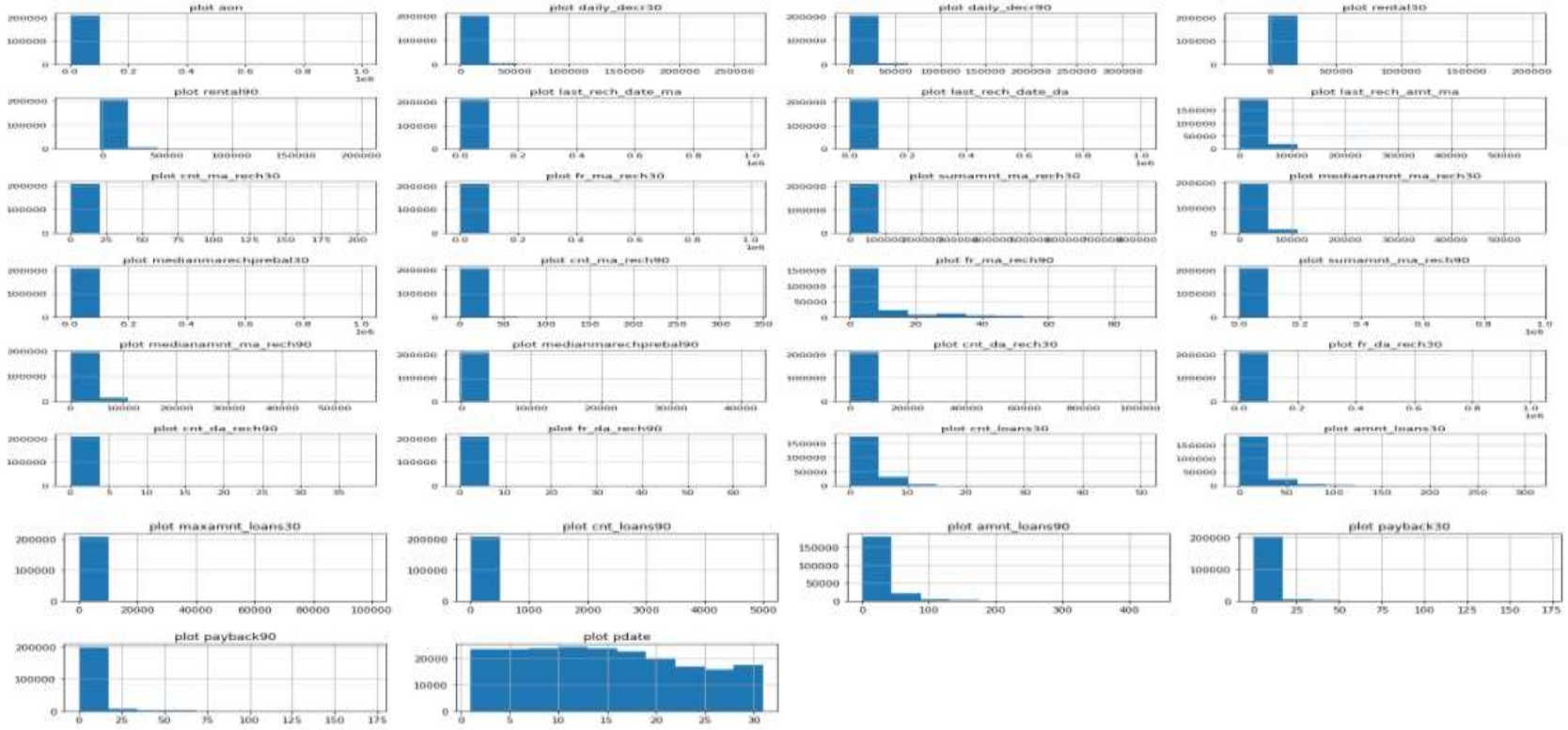



# Dashboard of Continuous Features



# Visualization Observations

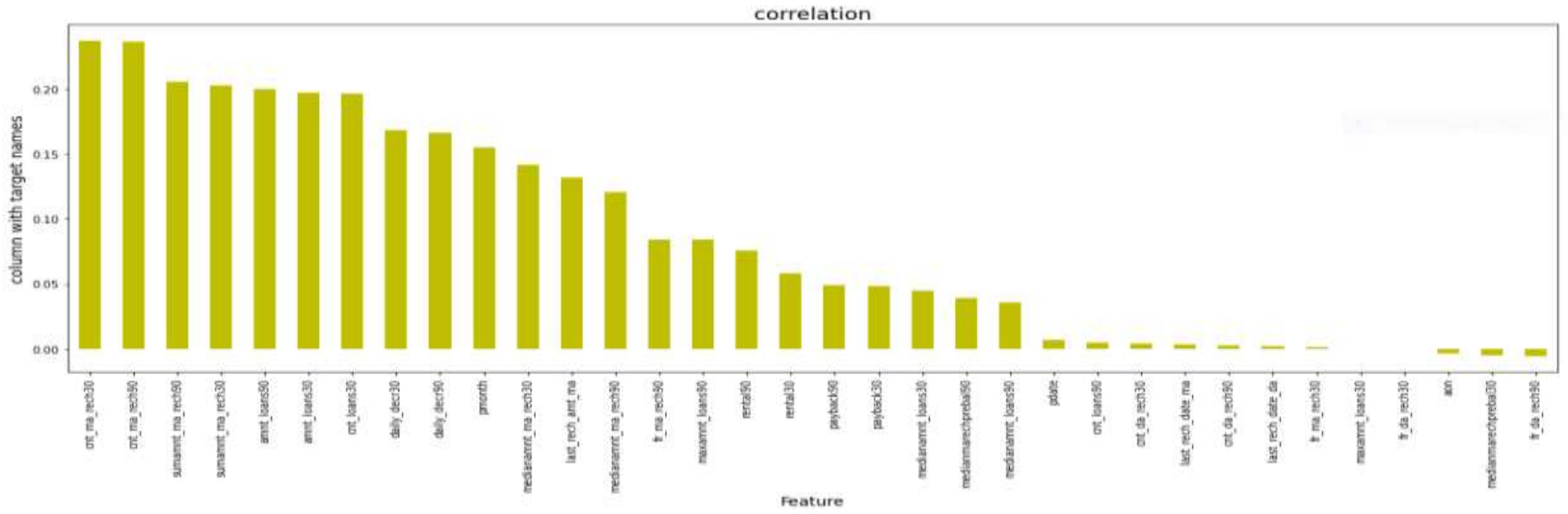
## Categorical

- Since the data of variables “pyear” and “pcircle” belongs to only one category i.e. “2016” and “UPW’ respectively which does not serve as an important information for analyzing data as it makes no difference to the dataset.
- Hence, the columns were deleted from the data because there is no information gain.

## Continuous

- Looking at the data of the continuous variables, huge skewness in all the variables can be observed. The skewness of data is resolved using the technique of “Power Transform”.
- Features are seen to have maximum records at value 0.
- **Assumption -** Here, an assumption was made of having too many outliers in the dataset.

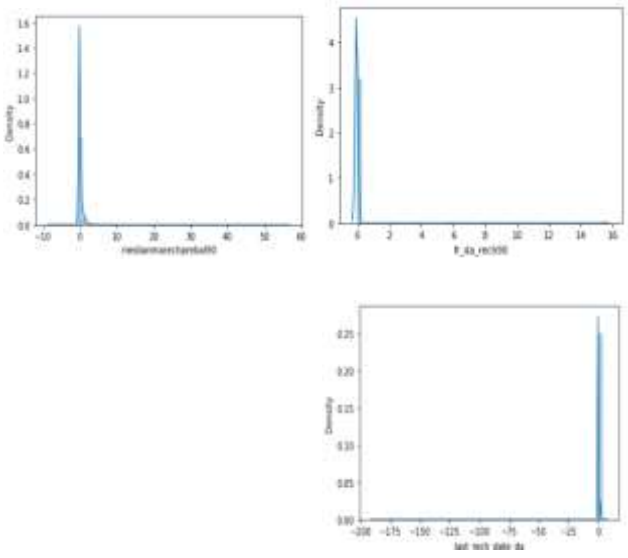
# Correlation of Features with the Target Variable




- No correlation of attributes 'fr_da_rech30' and 'maxamnt_loans30' is found with the Target Variable 'label'.
- This data has no information gain in building an ML model so the features are deleted from the dataset.

# Data Cleaning

| | Correlation with Target | Column Name | Normalised | Outliers |
|---|---|---|---|---|
| **0** | 0.001711 | last_rech_date_da | No | Alot |
| **1** | -0.005418 | fr_da_rech90 | No | Alot |
| **2** | 0.039300 | medianmarechprebal90 | No | Alot |



- Features in the above table contained enormous amount of garbage values consisting of high skewness, no good correlation with target attribute and immense outliers. Therefore, the data was cleaned from the dataset.
- The skewness of these features is seen from the graphs shown in the right, there is no normal distribution of data.

# Techniques Used

**Label Encoder** - The technique is used in ML to convert values in object data type into integer format. In this case, attribute “pcircle” is converted from object data to integer data.

**Power Transform -** Statistics used to transform the skewness in data into a normal distribution curve. This helps to reduce biasness and uncertainties in data for better model training.

**Standard Scaling -** One of the scaling techniques used for standardization of data wherein the values are centered around by removing the mean and scaling data to unit variance. In this case, the value ranges are too high within the dataset is, hence standardization plays a vital role in data normalization.

**Oversampling -** Oversampling is a type of sampling technique used to handle the imbalance nature of the target variable. It duplicates examples from the minority class and adds them to the training dataset.

# Algorithms Used

**Logistic Regression-** Logistic Regression is a supervised learning algorithm. It is a predictive regression analysis which is conducted when there are binary values in the target variable.

**Gaussian NB -** Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. It is a simple classification technique, but has high functionality.

**Decision Tree Classifier -** Visualisation view of the Decision Tree is in the form of graphs. Decision tree divides the main data set into a subset of trees that consists of choices and results. Node of each tree depicts a choice and the edges depicts the decision.

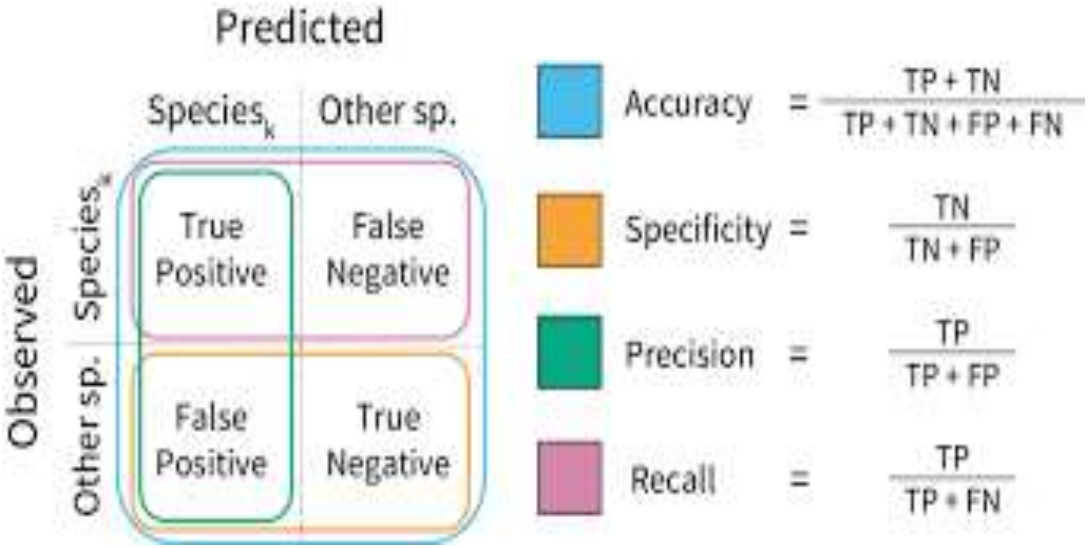
**Kneighbors Classifier -** KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

**Random Forest Classifier -** Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier.

# Evaluating Classifiers

**Outcome:**

1. **Accuracy**
2. **Confusion Matrix**
3. **Precision**
4. **Recall**
5. **F1 Score**
6. **AUC ROC Curve**



# Best Selected Model

## Random Forest Classifier

- Comparing all algorithms, the best model selected is the Random Forest Classifier as the precision, recall and f1 score along with the accuracy is observed to be the best in Random Forest than any other classifiers used. Random Forest is applied on the over sampled data and then hypertuning it to receive the best of it.

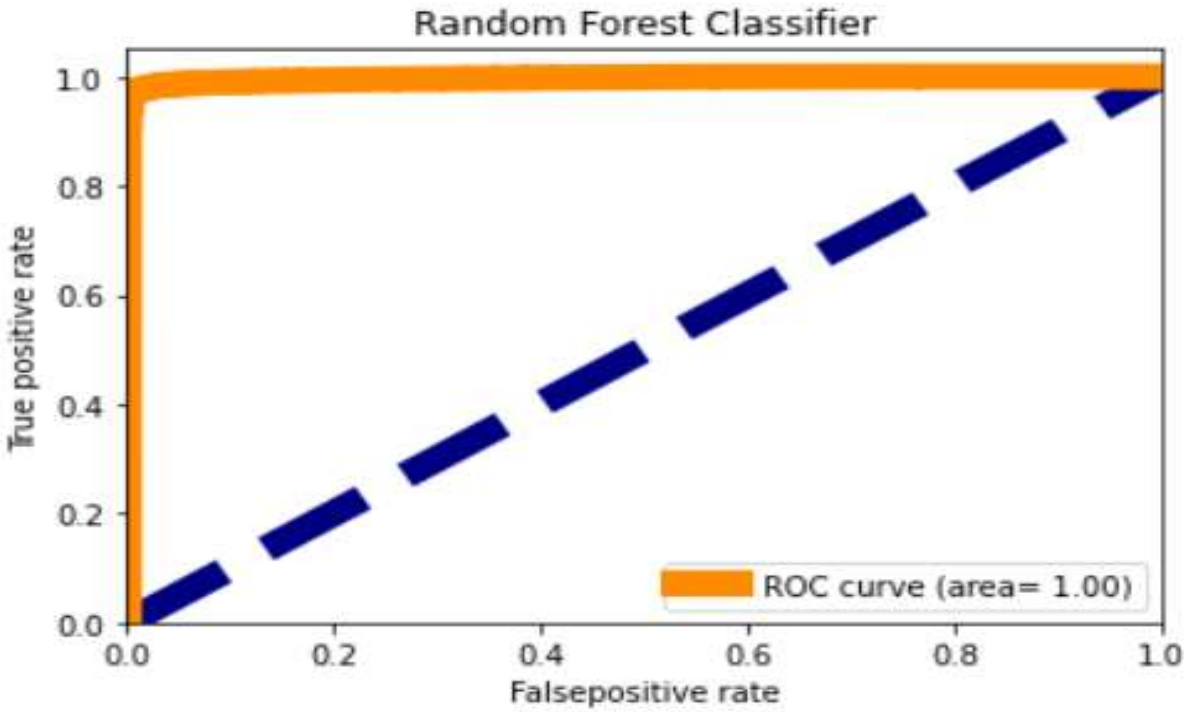
- The evaluation scores of Random Forest Classifier for the given data is :

```
[[ 3501  2148]
 [ 1478 38984]]
0.9213636659365444
              precision    recall  f1-score   support

           0       0.70      0.62      0.66      5649
           1       0.95      0.96      0.96     40462

    accuracy                           0.92     46111
   macro avg       0.83      0.79      0.81     46111
weighted avg       0.92      0.92      0.92     46111
```

# AUC ROC Curve




The **Area Under the Curve** (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.


Thank
you!