CAPSTONE PROJECT 3

Credit Card Default Prediction
Supervised ML Classification Model

Team Members

Anas Mustafa

Chetan Rajput

Sarthak Rastogi

Understanding the concept

What is credit card default?

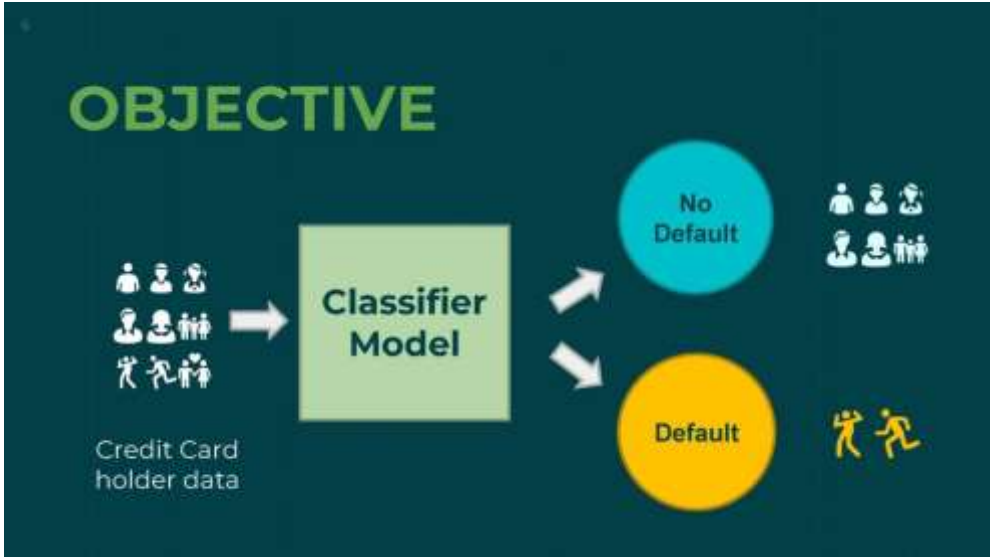
Credit card default happens when you have become severely delinquent on your credit card payments. Default is a serious credit card status that affects not only your standing with that credit card issuer but also your credit standing in general and your ability to get approved for other credit-based services.

Why Do we need to predict Credit card default beforehand ?

The **financial institution** can be capable of preventing the loss. Here, we have used various machine learning classification techniques to carry out Default related analysis.

Content

- Introduction
- Defining Problem Statement
- Data Summary
- Approach Overview
- EDA / Feature analysis
- Modelling Overview
- Model evaluation
- Feature Importance
- Challenges
- Conclusion



Introduction

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Overview and Objective

Overview

In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards. Now we know the most common issue there is in providing these kind of deals are people not being able to pay the bills. These people are what we call "Defaulters".

Objective

The main objective of this project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Data Summary

- X1 - Amount of credit(includes individual as well as family credit)
- X2 - Gender
- X3 - Education
- X4 - Marital Status
- X5 – Age
- X6 to X11 - History of past payments from April to September
- X12 to X17 - Amount of bill statement from April to September
- X18 to X23 - Amount of previous payment from April to September
- Y - Default payment

Approach Overview

- Data inspection and cleaning Exploring data, checking for outliers Clean data to get it ready for Analysis
- EDA & Data Pre-processing Checking distributions of variable Univariate and multivariate analysis Checking for imbalanced dataset
- Modelling

  Logistic Regression

  Decision Tree

  Random Forest

  XGBoost

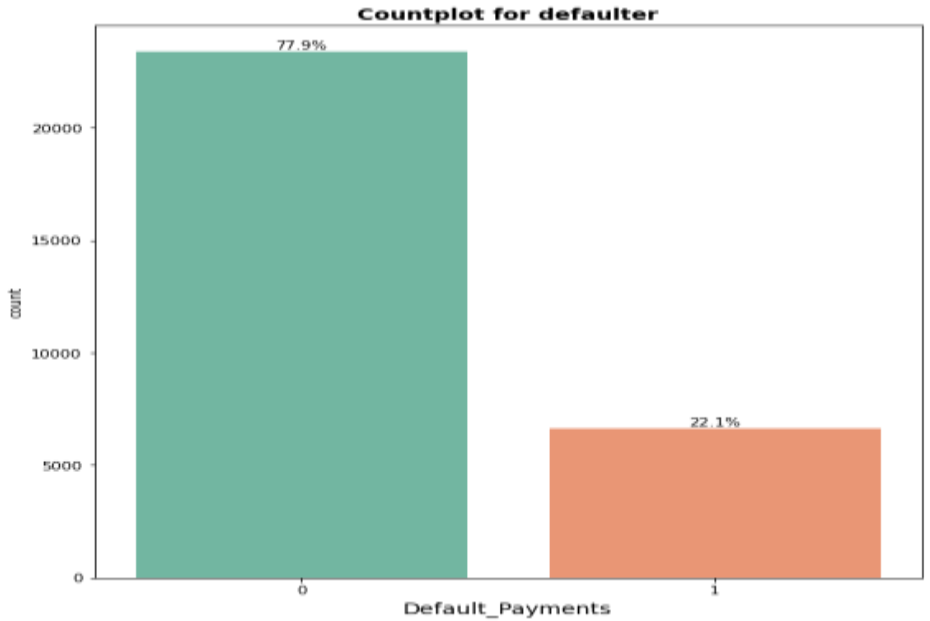
Basic Data Exploration

- Taiwan from April 2005 to September 2005.
- Dataset contains 30000 rows & 25 columns.
- In dataset, 6 months payment and bill data available.
- There are no null or duplicate values.

```
[ ] #reading the data set
    cred_df.head()
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5
0	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0
1	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455
2	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948
3	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959
4	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146

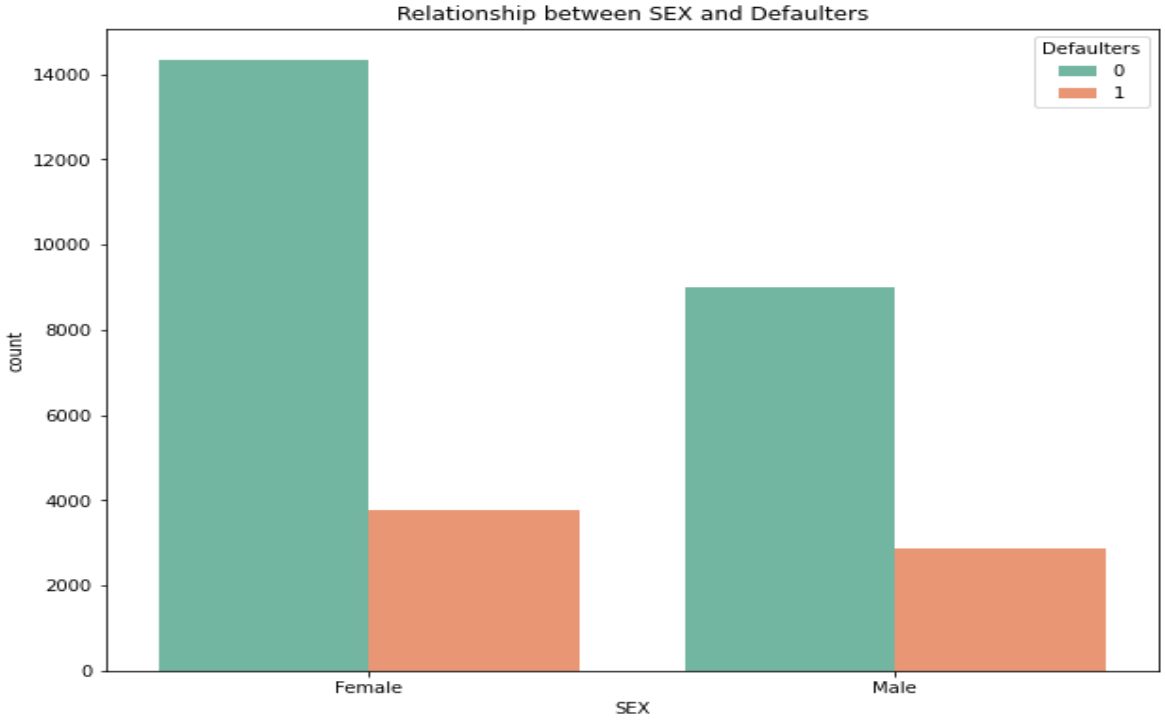
Exploratory Data Analysis



Looking at data, we will need to check if it is a case of Imbalanced dataset.

- Count of Non-default is a lot higher than default value.
- Non- Default data is 77.9% while Default cases are 22.1% as in dataset.

EDA- Gender Wise Defaulter Prediction



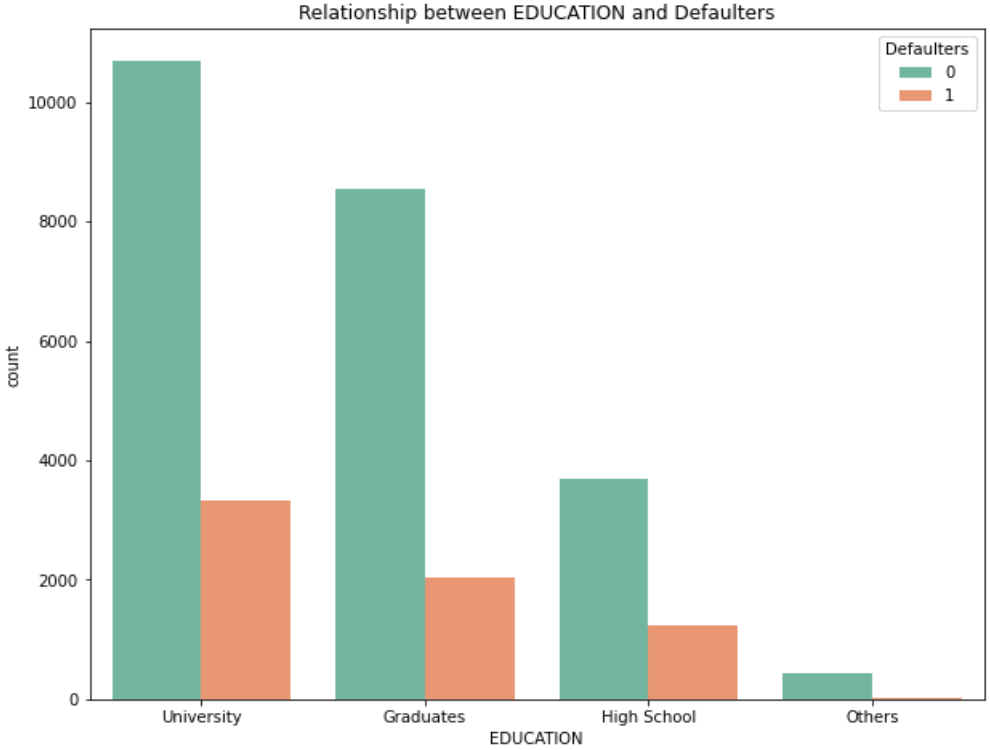
- Number of Male credit holder is less than Female.
- The ratio of defaulters is High in Male.

EDA - Education wise defaulter prediction

- More number of credit holders are university students followed by Graduates and then High school students.

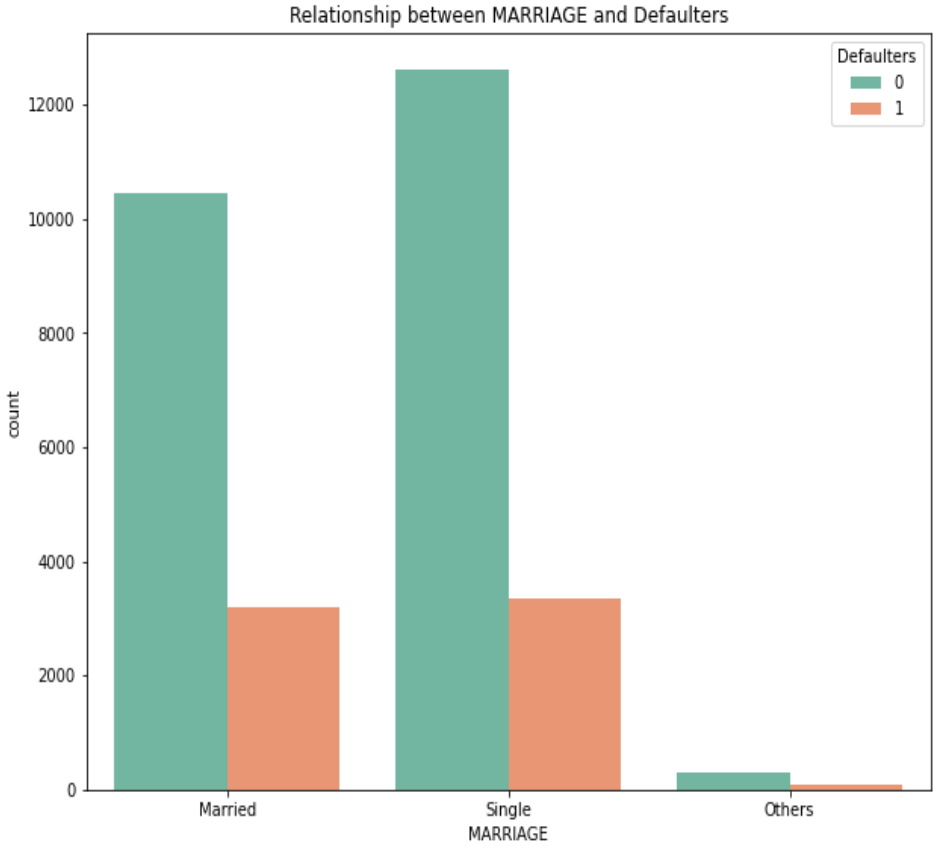
- University students have higher default payment than graduates and high school people.

- From university 11%are default, from graduate 7% are default, and from high school 4% are default.

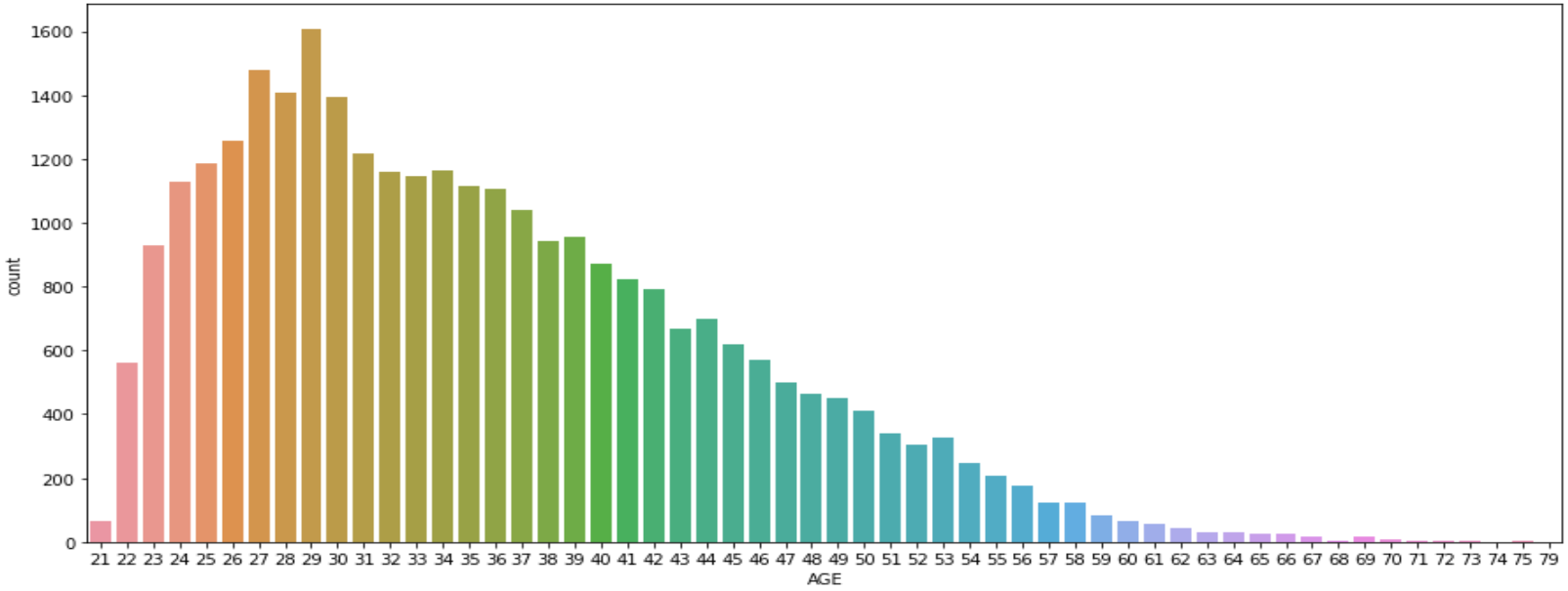


EDA – Marital status wise defaulter prediction

- From graph, we can say that more number of credit cards holders are Single as compared to Married and others.
- Here it seems that married are more likely to default than single.
- From single 11% are default and from married approx 11% are defaulter.

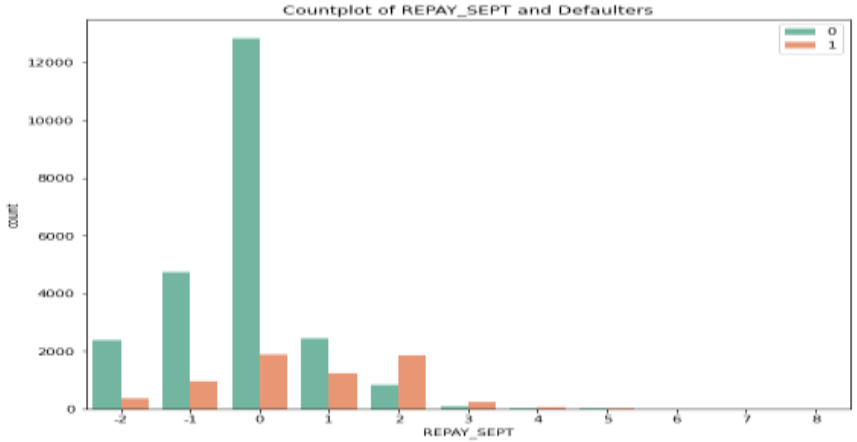


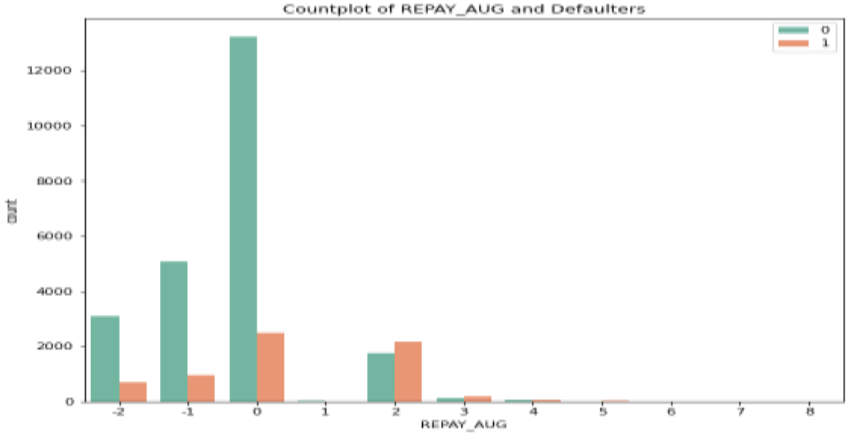
EDA– AGE wise Defaulter Prediction

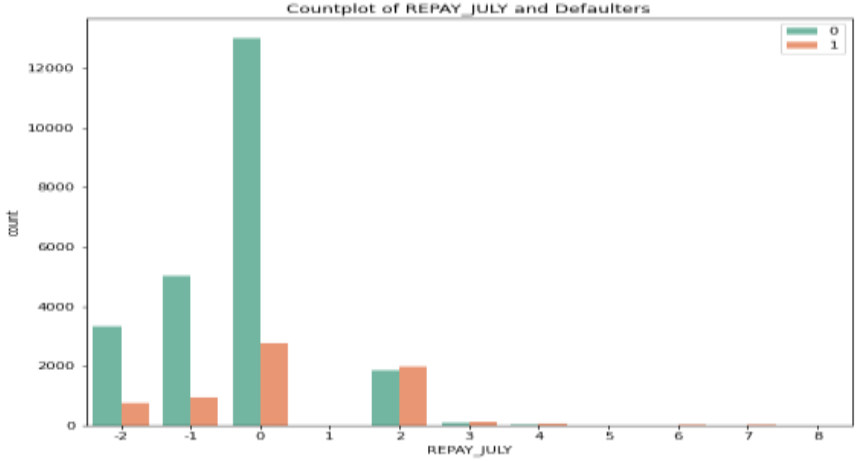


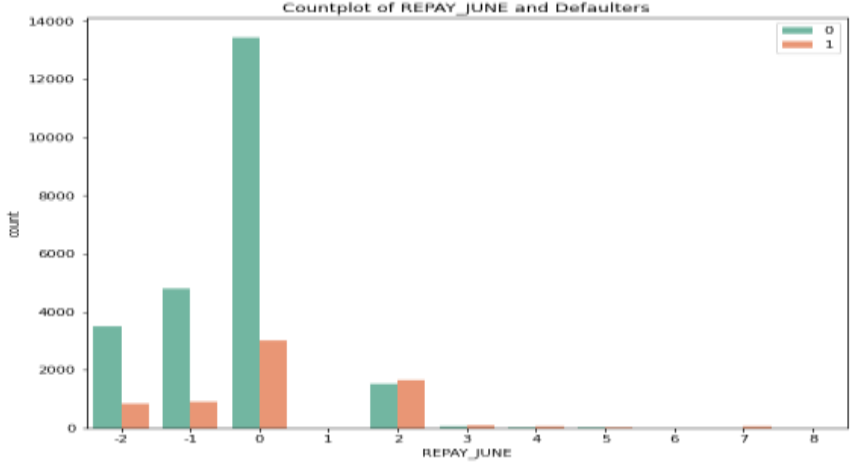
- More number of credit card holders age between 26-30.
- Also more number of Defaulters are between 27-29 years.

REPAYMENT STATUS









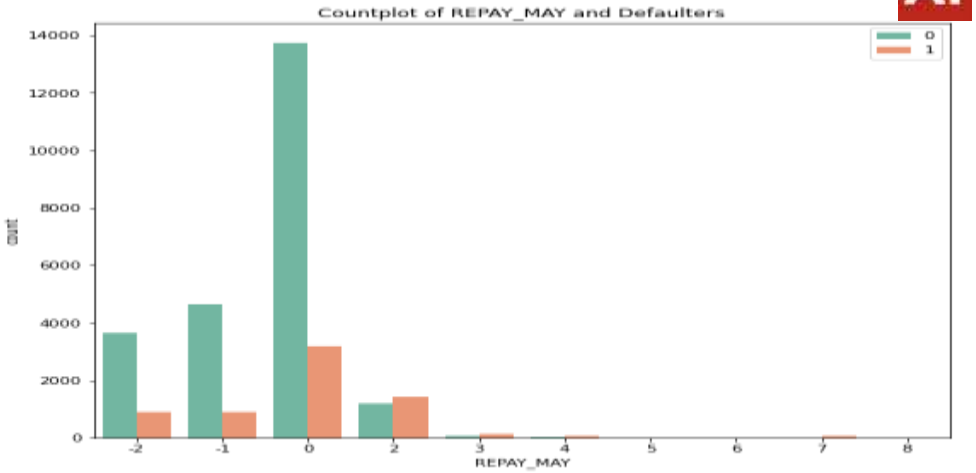
REPAYMENT STATUS

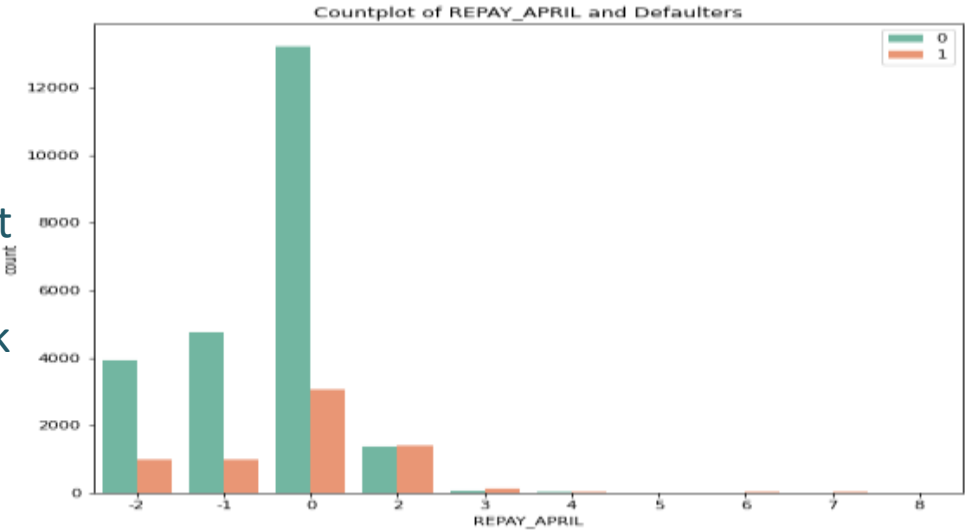
Repayment of different months.

- Blue – Non Defaulters
- Pink – Defaulter

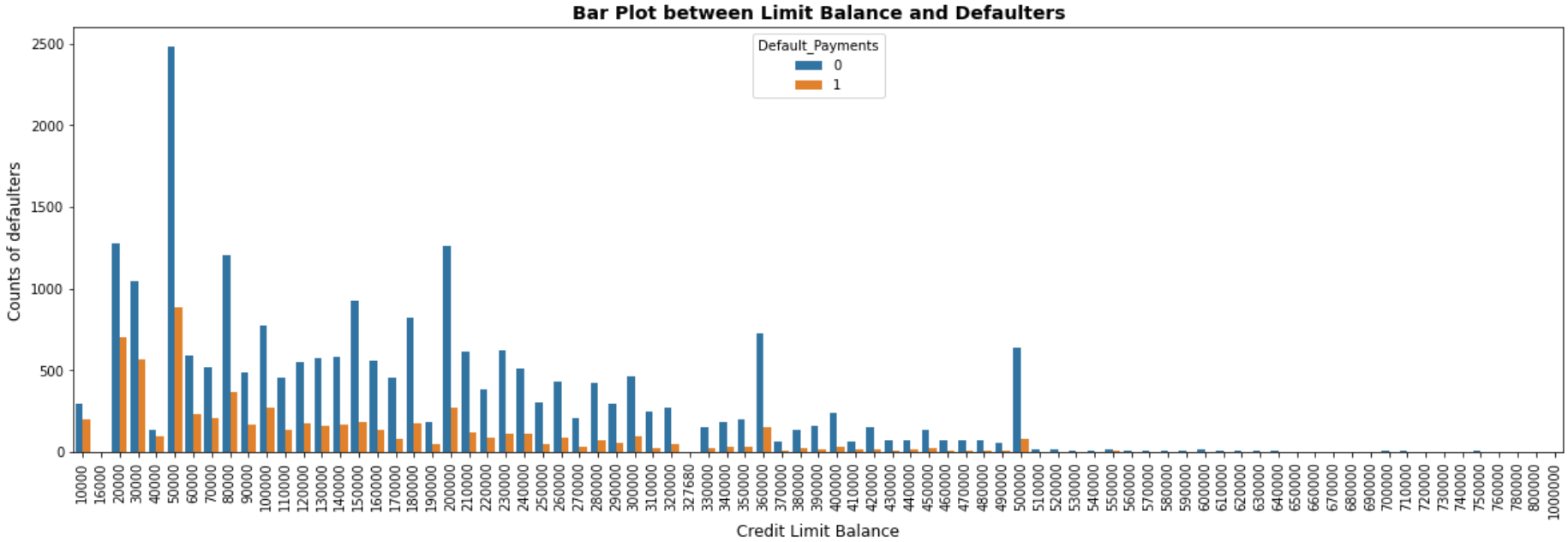
- 2 = No consumption
- 1 = paid in full
  0 = use of revolving credit (paid minimum only)
  1 = payment delay for one month
  2 – 8 (for consecutive months).

After careful observation we found that most credit card payment did by the customer on time. we see the distribution of the plot peak at 0 value ,which means on the x scale no delay in the payment of card





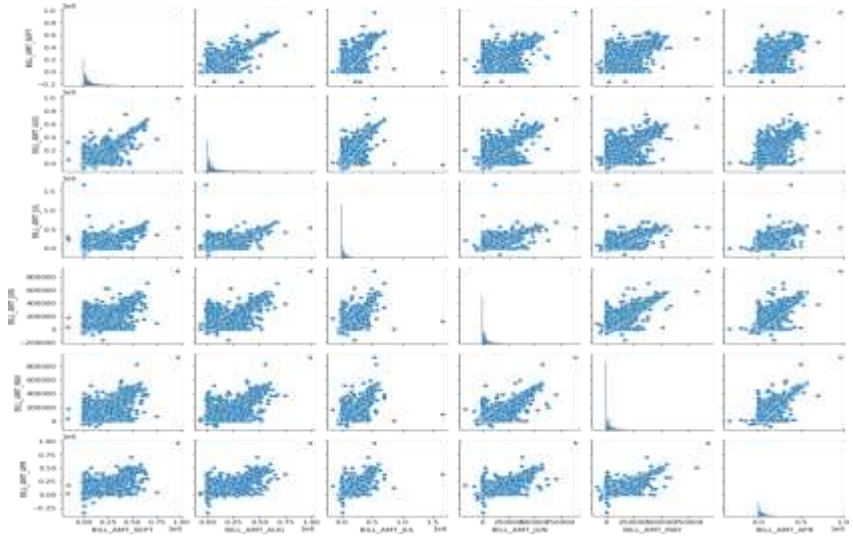
LIMIT BALANCE



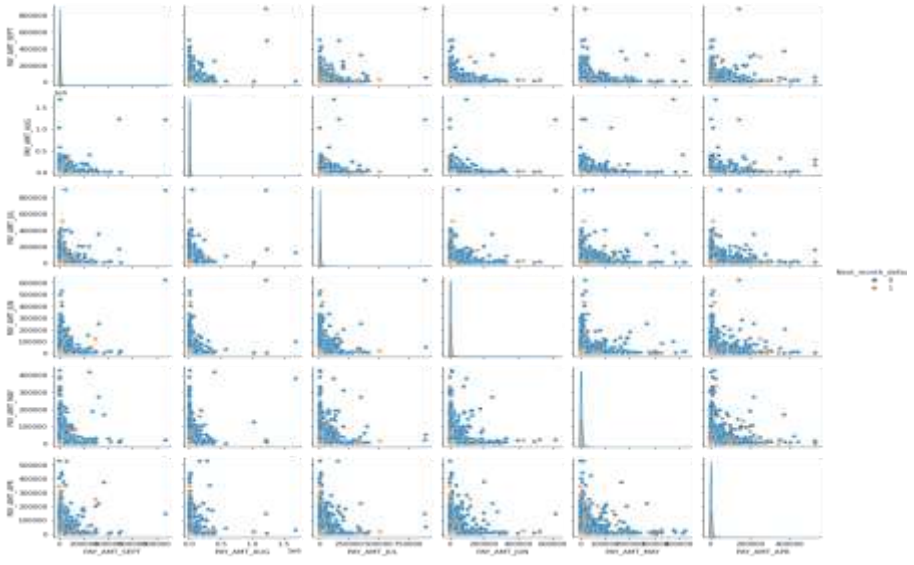
- Majority of defaulters are those who have credit limit between 20k to 300k.
- After Credit limit 500k, numbers of defaulters are almost negligible.

PAIR PLOTS

PAIR PLOT OF PAY

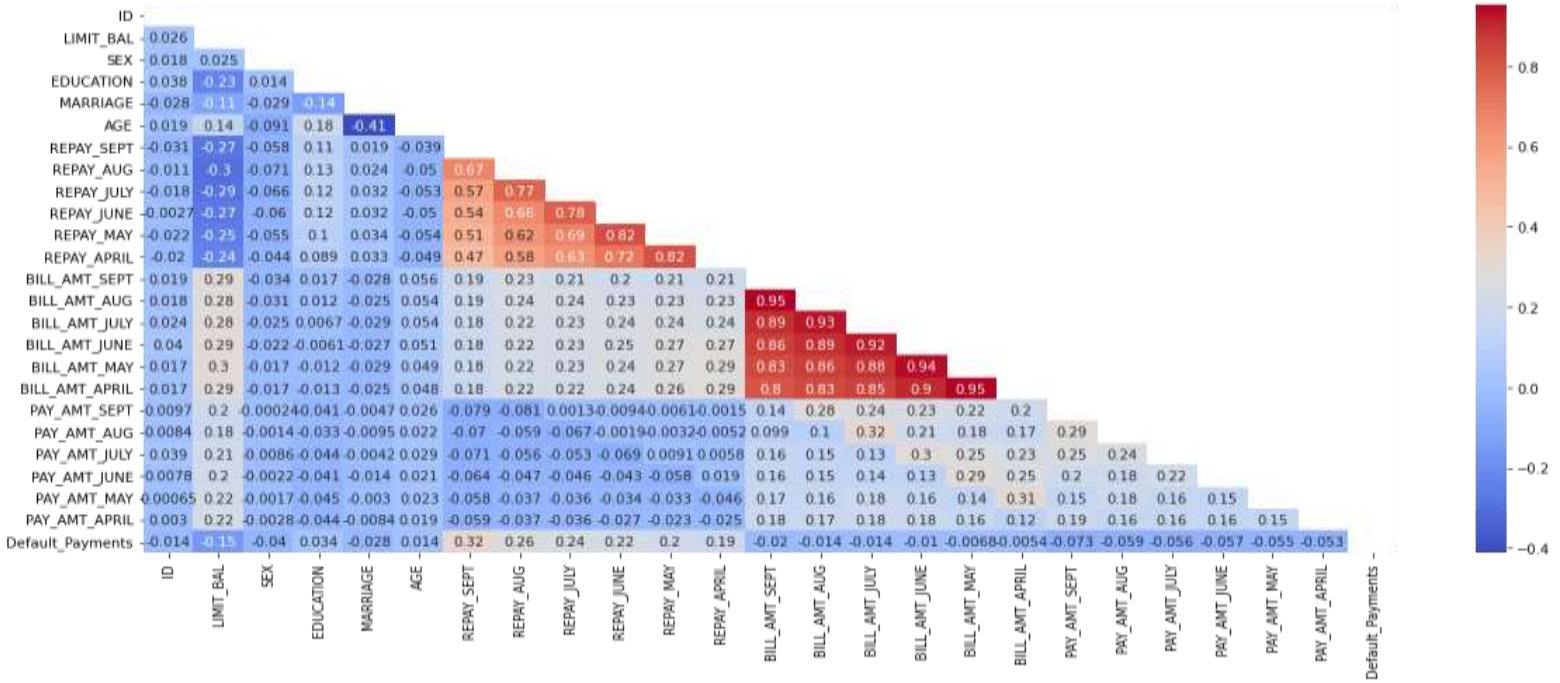


PAIR PLOT OF BILL



- The 1st pair plot shows the distribution of bill amount statements for each month explicitly for defaulters and non-defaulters.
- The 2nd pair plot shows the distribution of payment statements for each month explicitly for defaulters and non-defaulters

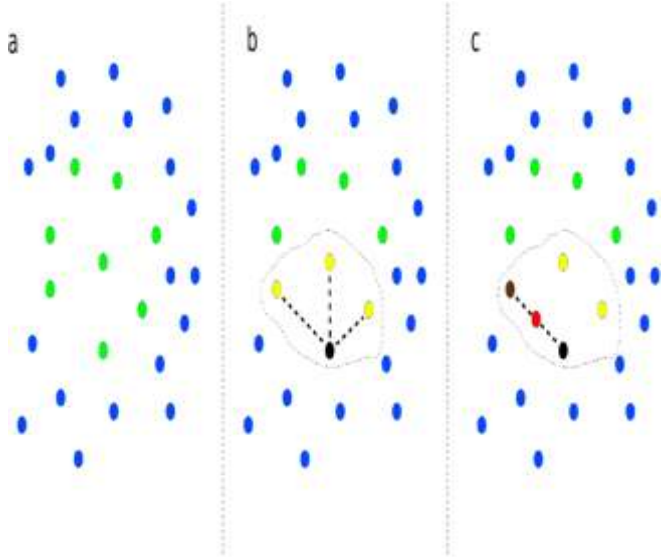
Correlation between parameters



We can see that no correlation between the features so there is no need to remove or drop some features.

SMOTE(Synthetic Minority Oversampling Technique)

- Our dataset is imbalanced which can lead to Biasness While Building Model.
- For Balancing We Use SMOTE. SMOTE is an algorithm that performs data augmentation by creating synthetic data points based on the original data points.
- The advantage of SMOTE is that you are not generating duplicates, but rather creating synthetic data points that are slightly different from the original data points.



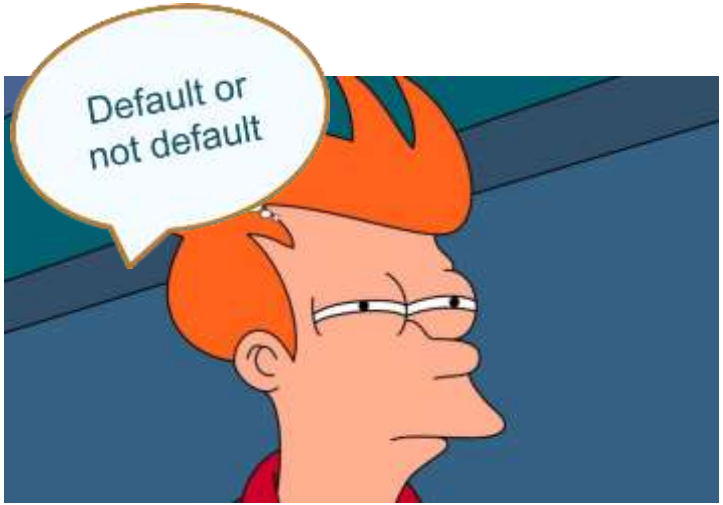
Modeling Overview

This is Classification problem statement.

There is Imbalance data with 78% non-defaulters and 22% defaulters.

We have applied following models :

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost

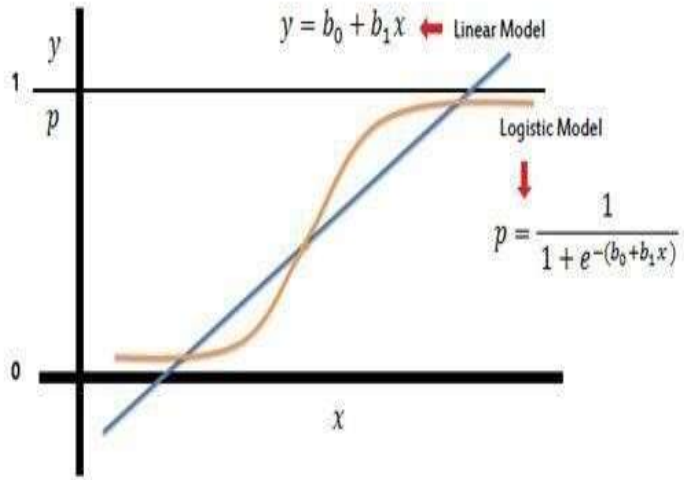


Logistic Modelling

LOGISTIC REGRESSION

- Logistic Regression is a Machine Learning algorithm and is basically used for binary classifications like yes-no, true-false, male-Female, etc.

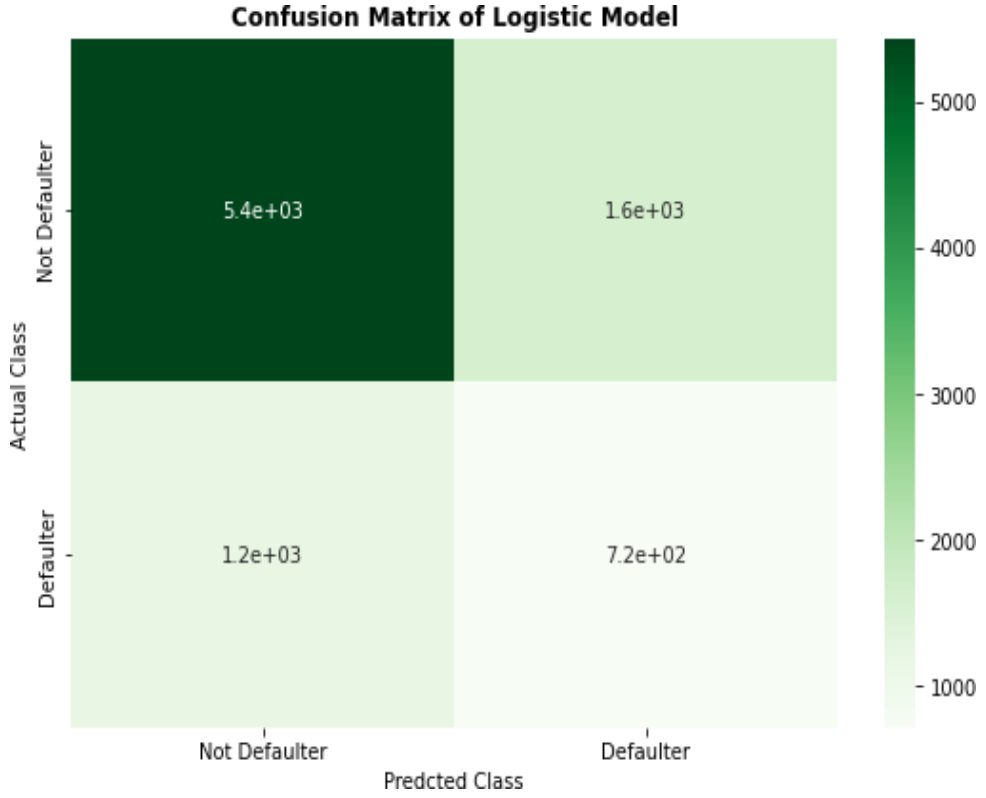
- It take the linear combination and apply a sigmoid function (logit). The Sigmoid curve gives value between 0 & 1.



Best Parameters {'C': 0.2, 'penalty': 'l2'}

Confusion Matrix of Logistic Model

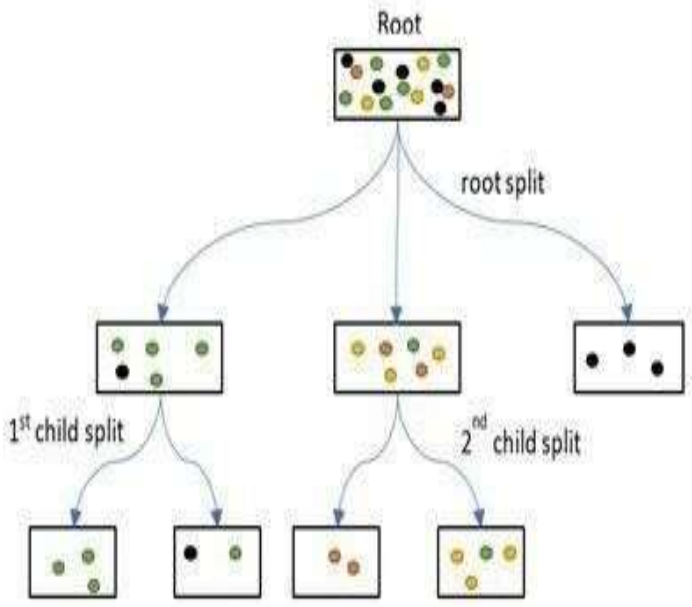
- Precision score of logistic model: 0.3064
- Recall score of logistic model: 0.3701
- F1 score of logistic model: 0.3352
- ROC AUC score of logistic model: 0.5699



Decision Tree

DECISION TREE CLASSIFIER

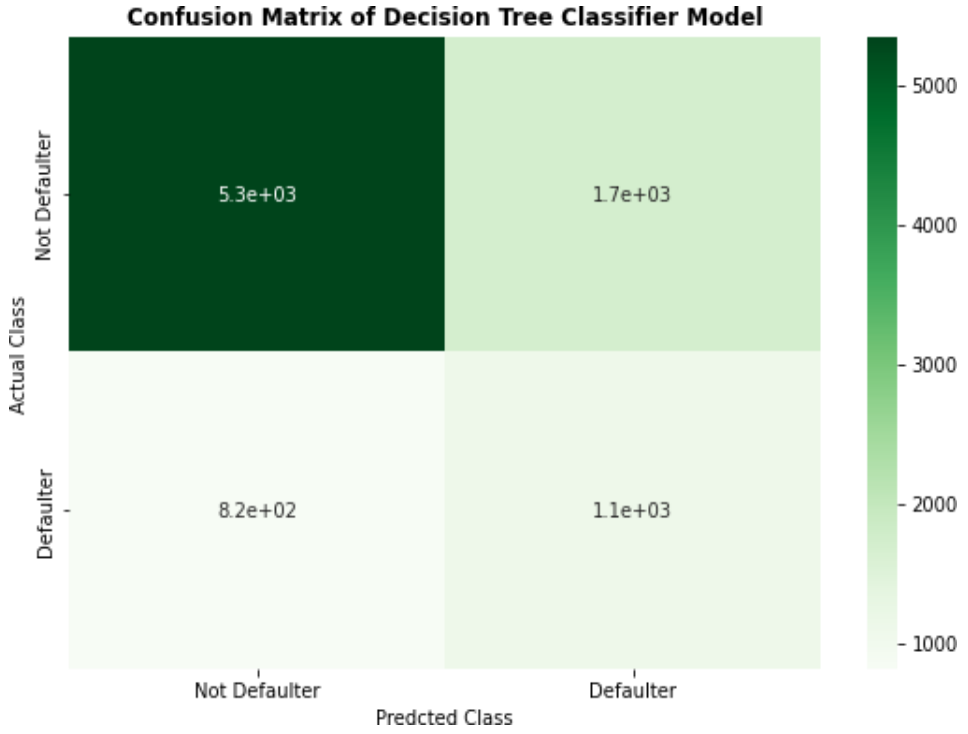
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- The objective of Decision tree algorithm is to find the relationship between the target column and the independent variables and Express it as a tree structure.



Best parameters : {'max_depth': 2, 'min_samples_leaf': 2, 'min_samples_split': 2}

Confusion Matrix of Decision tree Classifier

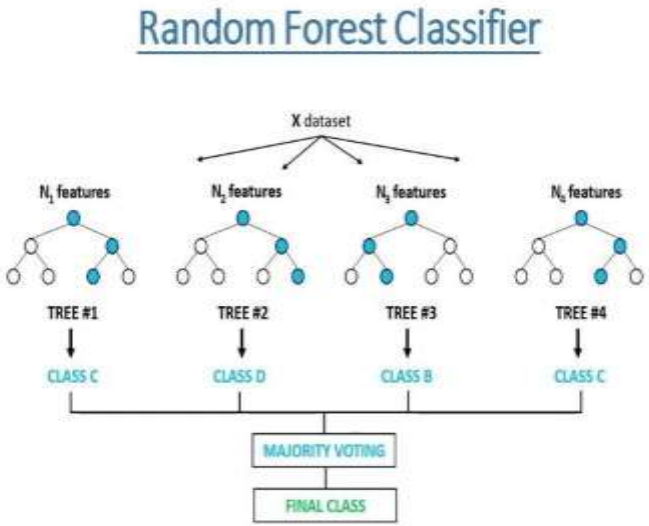
- Recall score of Decision Tree Classifier: 0.5773
- Precision score of Decision Tree Classifier: 0.5773
- F1 score of Decision Tree Classifier: 0.4691
- 6 ROC-AUC score of Decision Tree Classifier: 0.667200



RANDOM FOREST CLASSIFIER

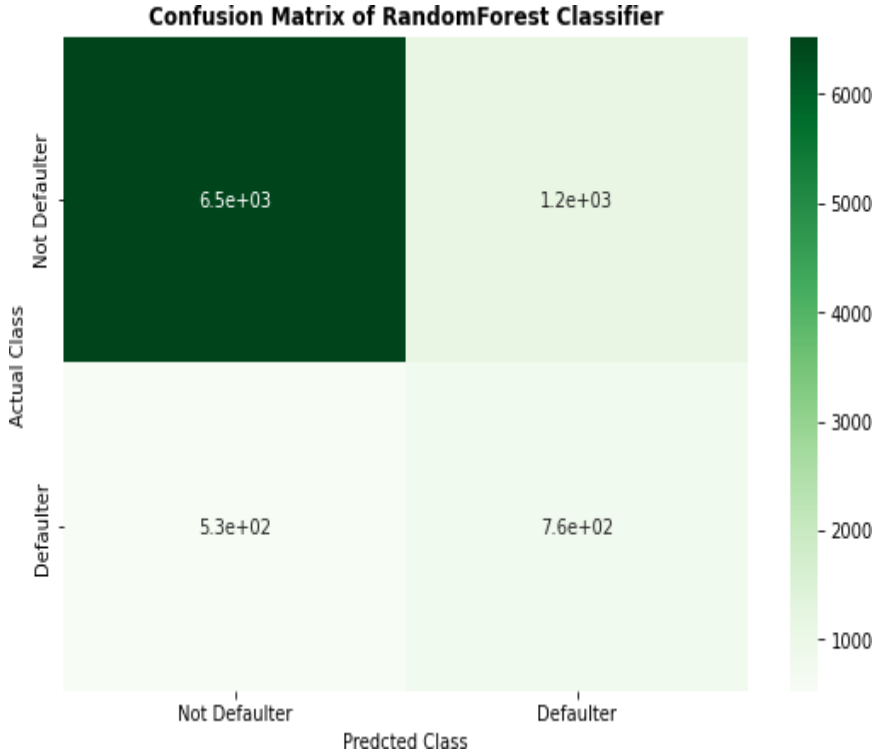
Random Forest Classifier

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



Confusion Matrix of Random Forest Classifier:-

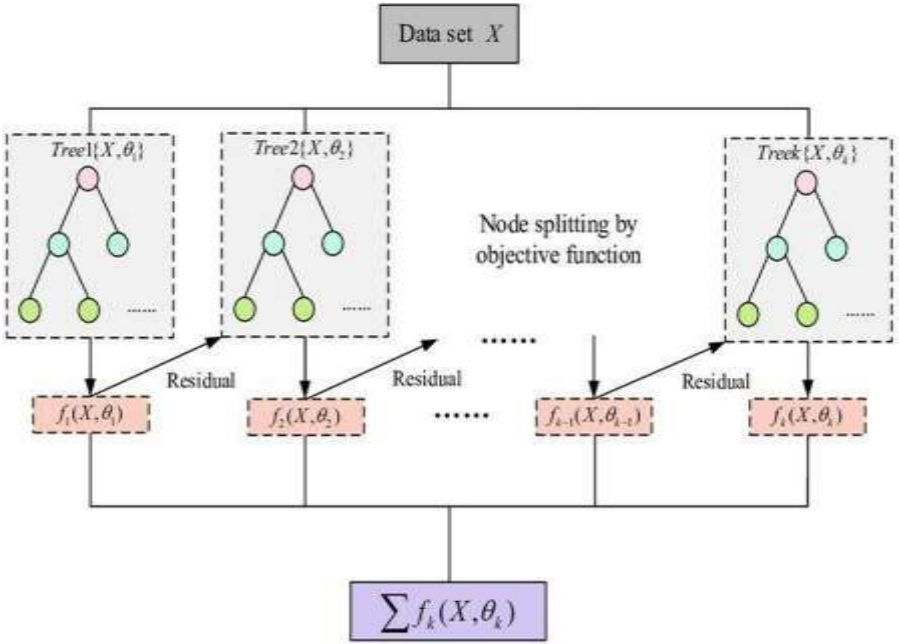
- Precision Score of Random Forest with Hyperparameter Tuning: 0.39381
- Recall Score of Random Forest with Hyperparameter Tuning: 0.59087
- F1 Score of Random Forest with Hyperparameter Tuning: 0.47262
- ROC AUC Score of Random Forest with Hyperparameter Tuning: 0.71914



XGBOOST CLASSIFIER

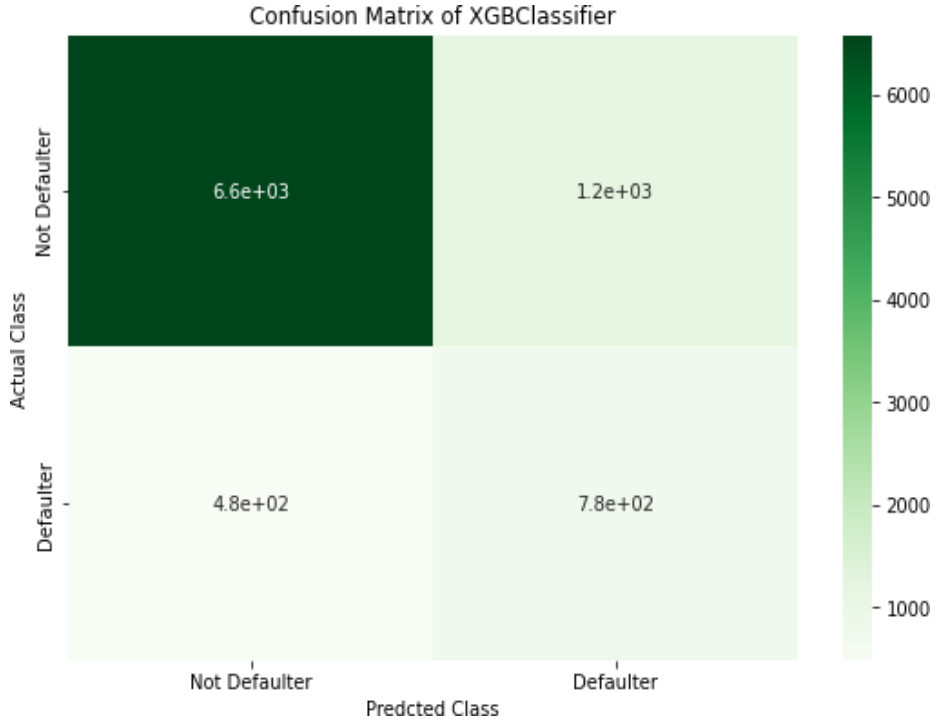
XG BOOST CLASSIFIER

- Stands for:– Extreme Gradient Boosting.
- XGBoost is a powerful iterative learning algorithm based on gradient boosting.
- Regularization to avoid over fitting
- Tree pruning using depth-first approach
- It is generally used for very large dataset



Confusion Matrix of XG Boost Classifier

- Precision Score of Random Forest with Hyperparameter Tuning: 0.400515
- Recall Score of Random Forest with Hyperparameter Tuning: 0.617647
- F1 Score of Random Forest with Hyperparameter Tuning: 0.485928
- ROC AUC Score of Random Forest with Hyperparameter Tuning: 0.7337137

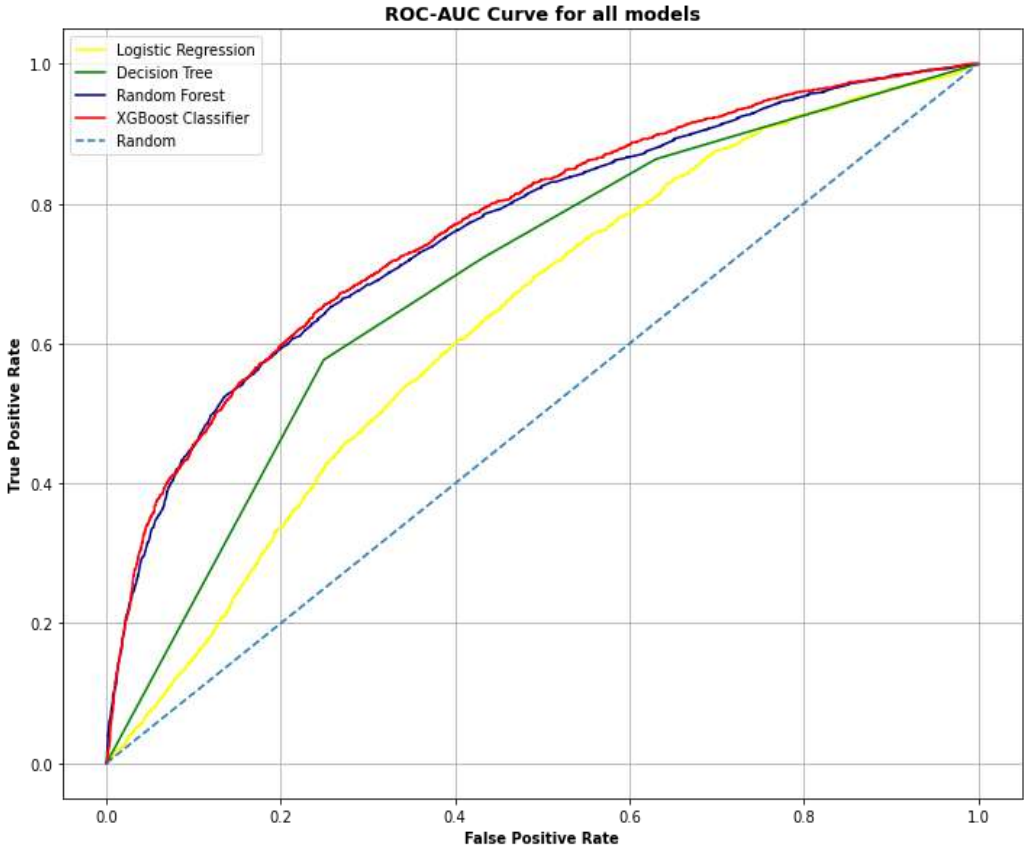


MODEL EVALUTION

	Classification Models	Training Accuracy	Testing Accuracy	Precision Score	Recall Score	F1 Score	ROC-AUC Score
0	Logistic Regression	0.614665	0.659778	0.309572	0.470103	0.373312	0.591001
1	Decision Tree Classifier	0.745952	0.718333	0.577320	0.577320	0.469110	0.667201
2	Random Forest	0.999816	0.811333	0.396392	0.593364	0.475278	0.720683
3	XGBoost Classifier	0.870461	0.817333	0.400515	0.617647	0.485929	0.733714

ROC CURVE

Receiver Operating Characteristic (ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.



Challenges

- Reading the dataset and understanding the problem statement.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Dealing with Imbalanced Dataset
- Feature engineering
- Feature selection - Making sure we don’t miss any important feature.
- Careful tuning of hyperparameters as it affects accuracy.
- Computation time was a big challenge for us.

Conclusion

1. Distribution of defaulter vs. non defaulter - around 78% are non defaulter and 22% are defaulter. Also we check for Marriage, Education, Sex with respect to defaulter and we found in marriage more number of defaulter are Male, in Education more no. of defaulter are University Students & in Marriage more no. of defaulter are Married.
2. After that we build the Four models Logistic Regression, Decision Tree, Default XGBoost Classifier & Random Forest . The best accuracy is obtained from the Default XGBoost Classifier
3. Using a Logistic Regression classifier, we can predict with 65.97% accuracy, whether a customer is likely to default next month.

5. With Decision Tree classifier having precision 57.77%, we can predict with accuracy of 71.83%, whether customer is likely to default next month.
6. Using Random Forest, we can predict with accuracy of 81.13%, whether customer will be defaulter in next month.
7. XG Boost Classifier with recall 61.77%, accuracy of 81.73%, we can predict whether customer is likely to default next month.

From the models that are applied on the dataset, *XG Boost and Random Forest* are giving the best evaluation matrices (precision, F1-score and ROC-AUC score).

- On behalf of these matrices we can predict whether customers would be defaulter or not in the next month.
- From the **ROC-AUC** curve, Random Forest and XG Boost classifier are more able to distinguish between positive and negative class.

T H A N K YOU!!