Capstone Project-3
Card Default Prediction

Team Members

Abdul Rahman Talha

Mohd Danish

Sharath Diwakar

Table of contents:-

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 Description

Attributes of dataset:-

1. **ID:** ID of each client
2. **LIMIT_BAL:** Amount of given credit in NT dollars (includes individual and family/supplementary credit
3. **SEX:** Gender (1=Male, 2=Female)
4. **EDUCATION:** (1=Graduate school, 2=University, 3=High school, 0,4,5,6=others
5. **MARRIAGE:** Marital status (1=married, 2=single, 0,3=others)
6. **AGE:** Age in years

Continued….

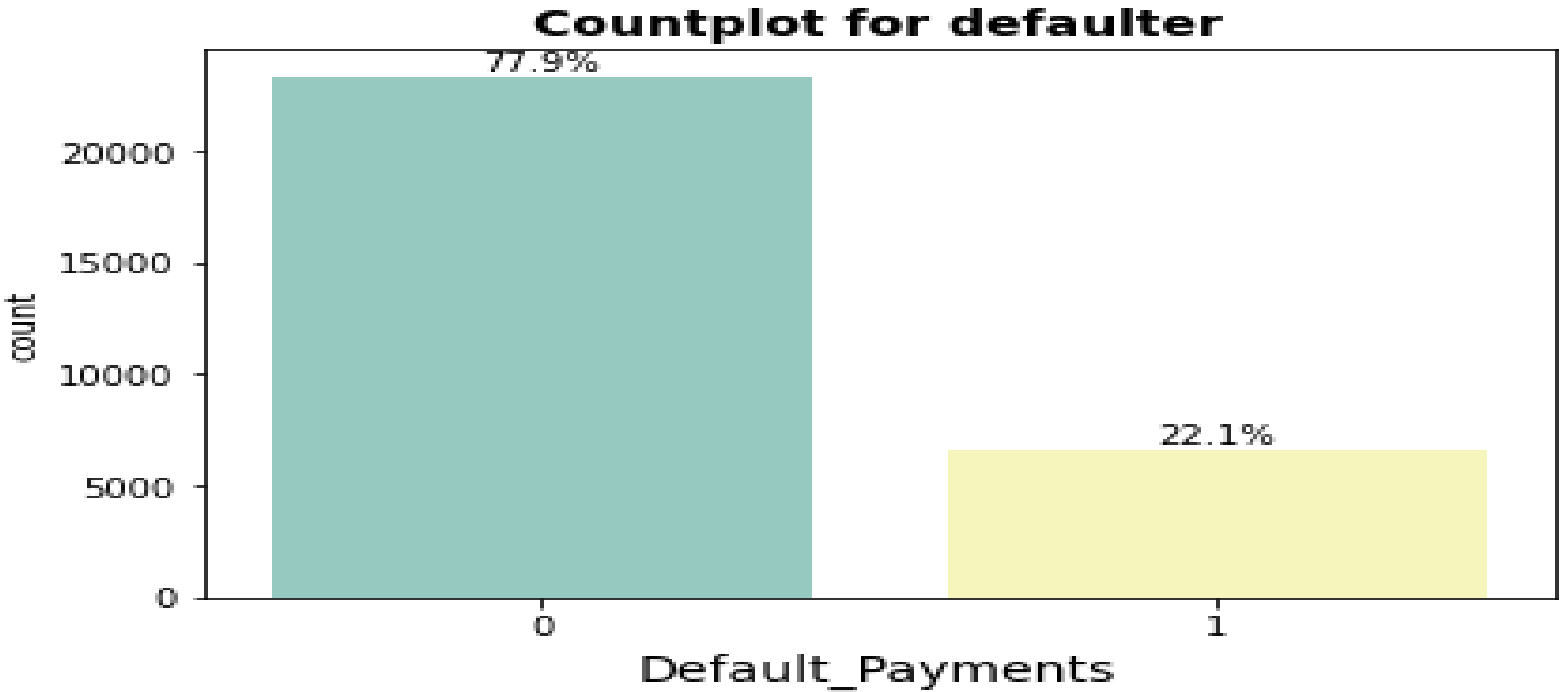
7. **PAY_0-6 :** Repayment status in September 2005 - April, 2005

8. **BILL_AMT1-6:** Amount of bill statement in September- April 2005 (NT dollar)

9. **PAY_AMT1-6:** Amount of previous payment in September- April, 2005 (NT dollar)

10. **default.payment.next.month:** Default payment (1=yes, 0=no)

Exploratory Data Analysis (EDA)



Approximately 78% are Non Defaulters and 22% are Defaulters respectively.

EDA Continued…

SEX

Number of Male credit holder is less than Female.

The ratio of defaulters is High in Male.

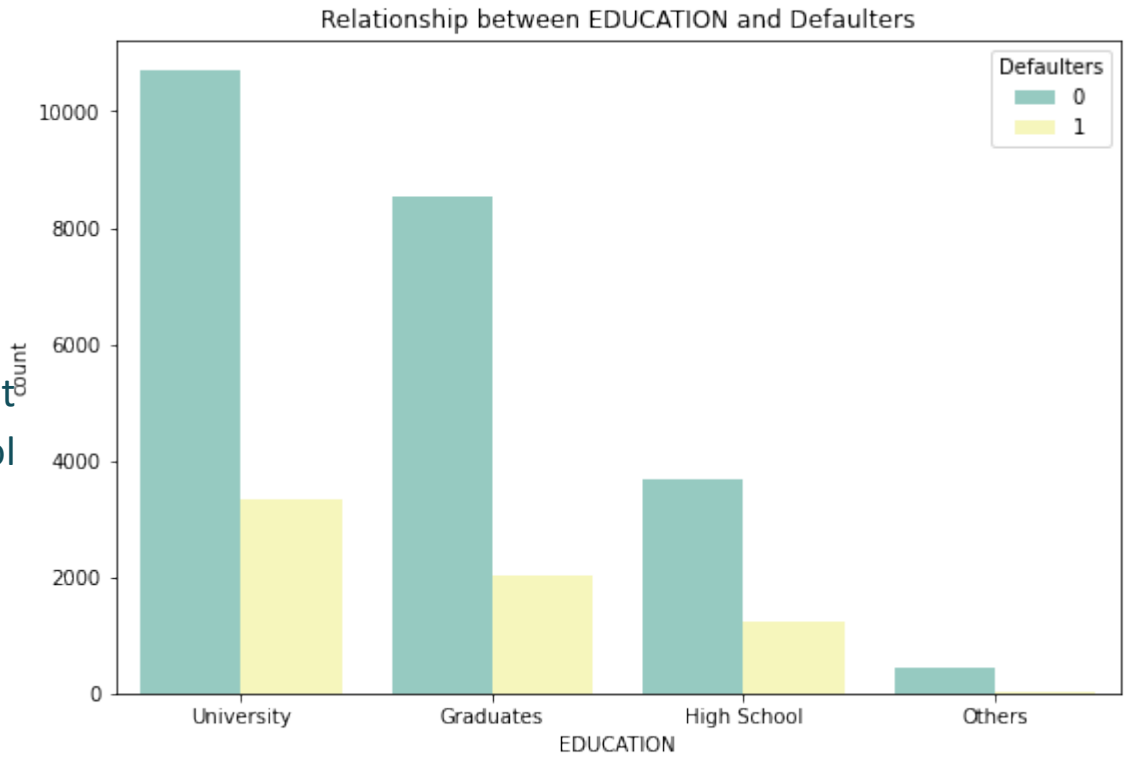
EDA Continued…

EDUCATION

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.



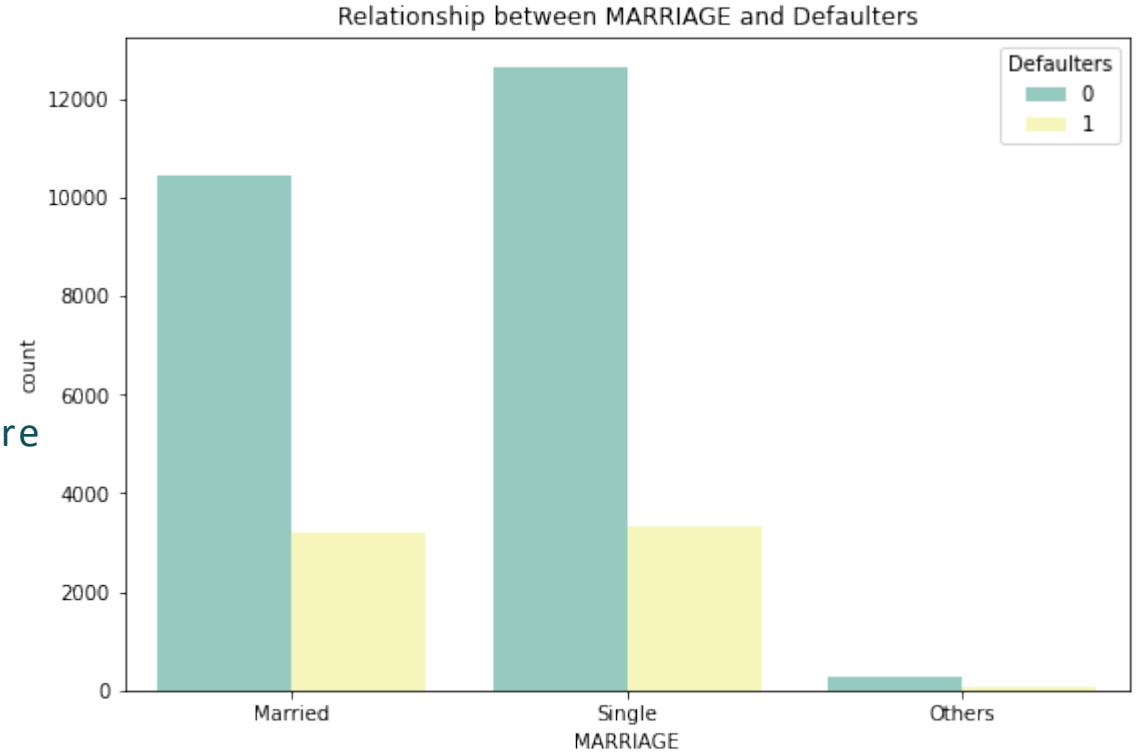
EDA Continued…

MARRIAGE

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.

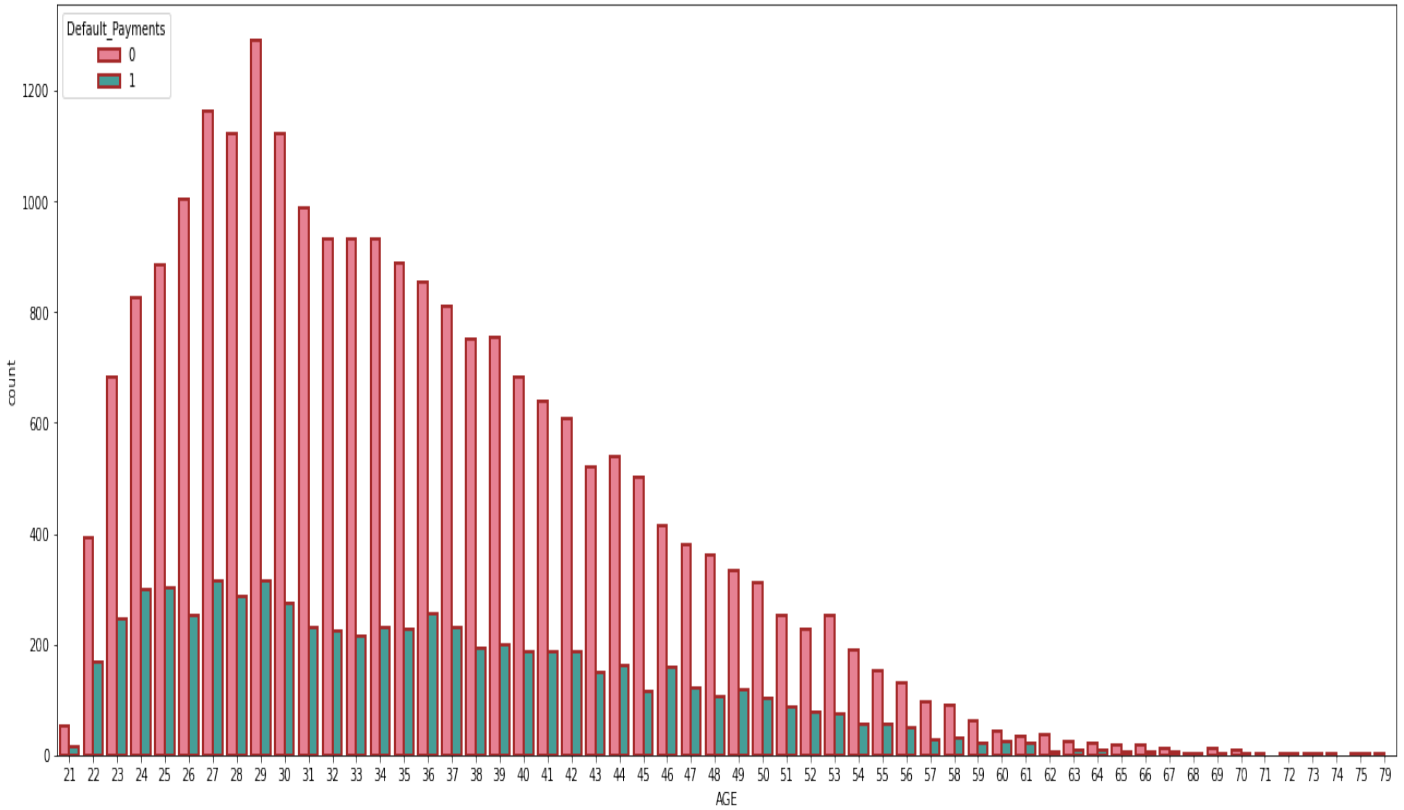


EDA Continued…

AGE

More number of credit card holders age between 26-3.

Also more number of Defaulters are between 27-29 years.



Repayment Status

Repayment of different months.

- Red – Non Defaulters
- Blue – Defaulter

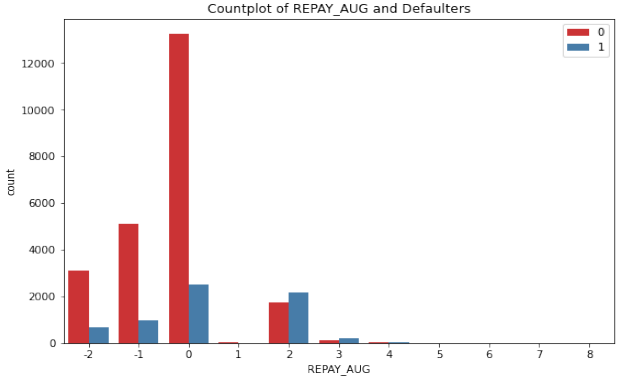
-2 = No consumption

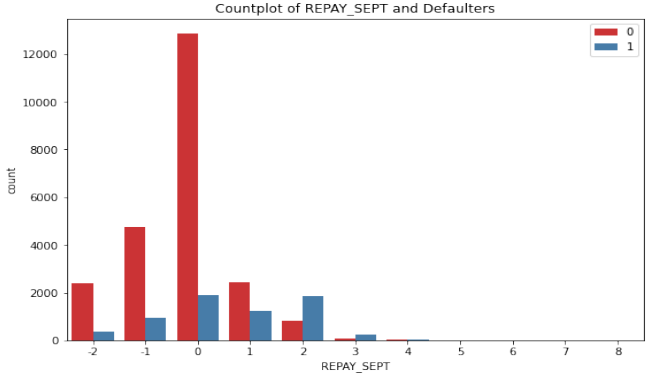
-1 = paid in full

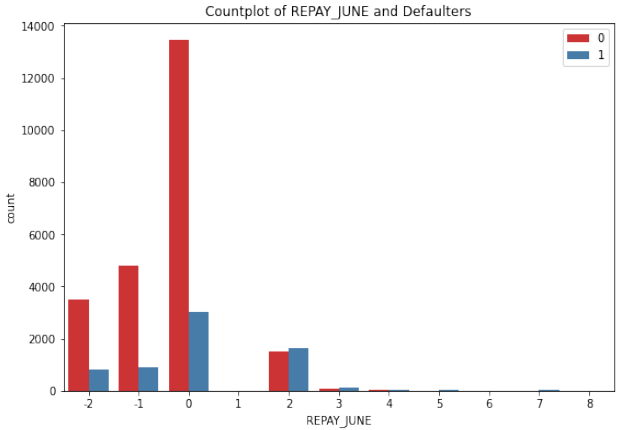
0 = use of revolving credit (paid minimum only)

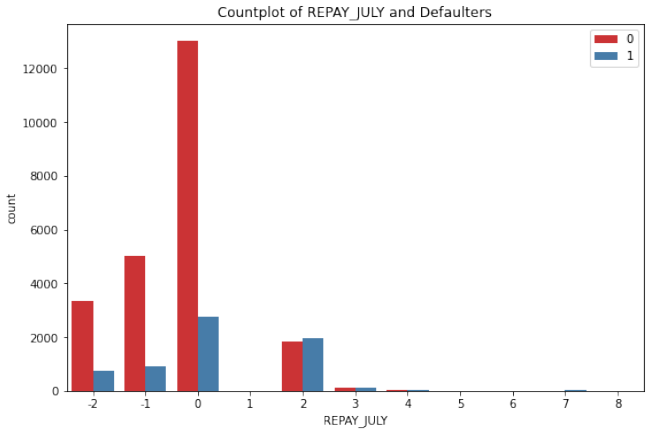
1 = payment delay for one month

2 – 8 (for consecutive months)



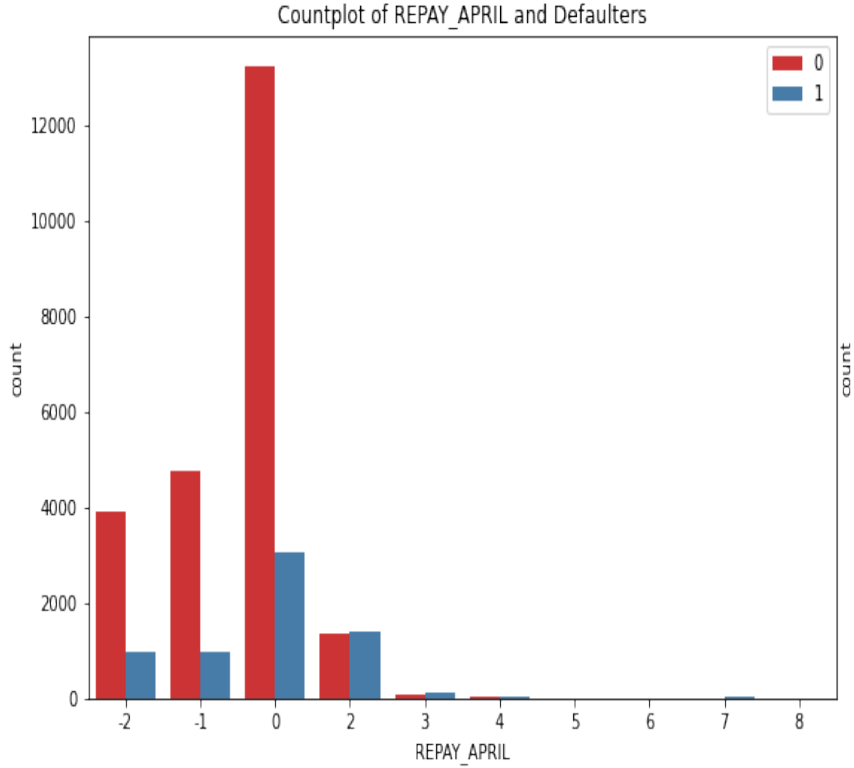


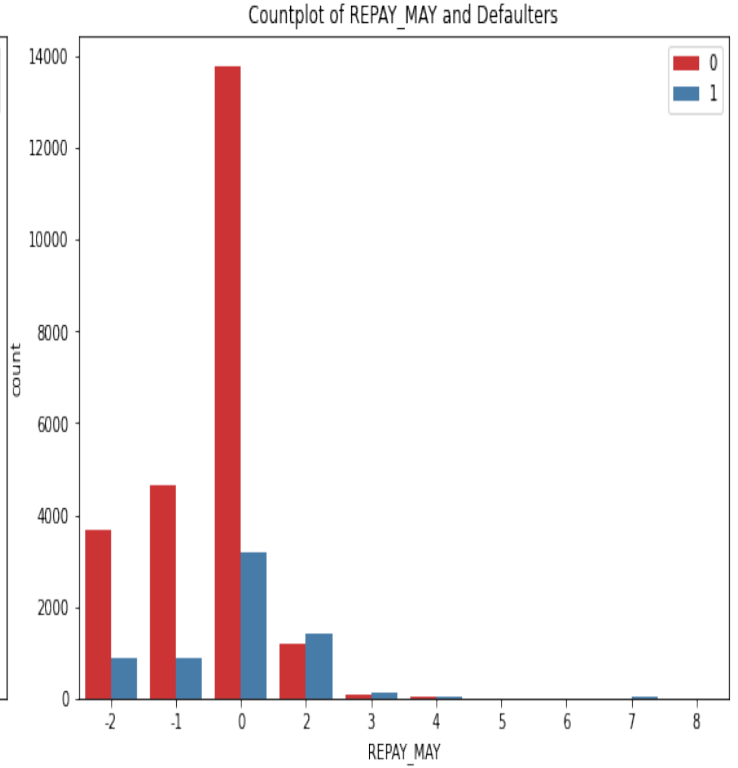




Repayment Cont.

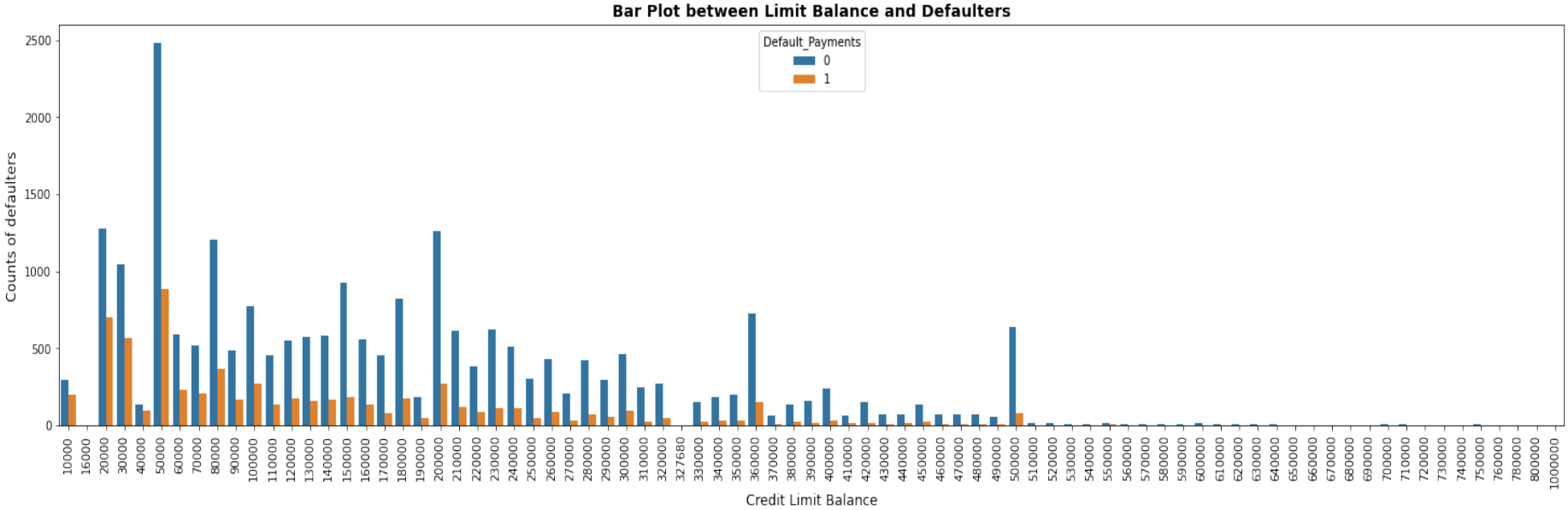
Contd.





EDA Continued…

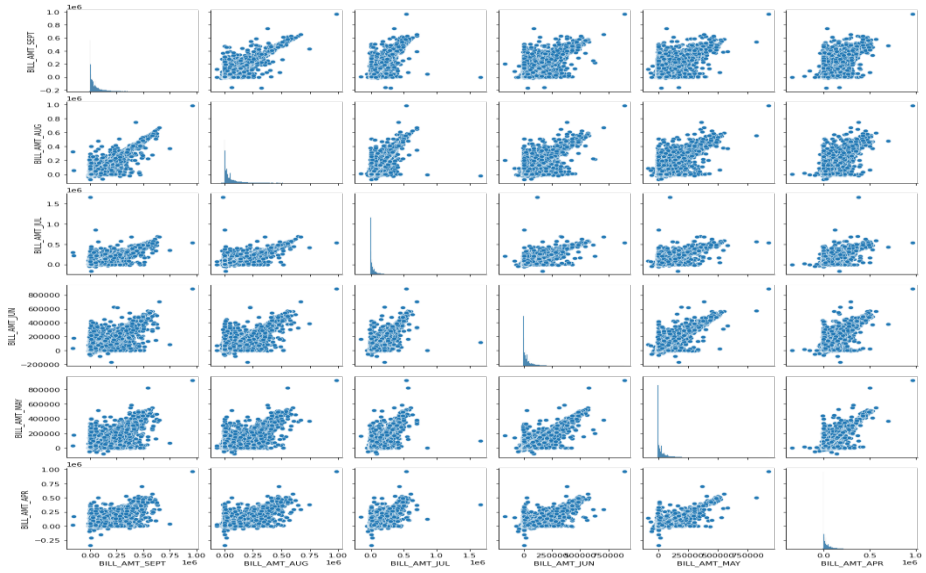
LIMIT BALANCE



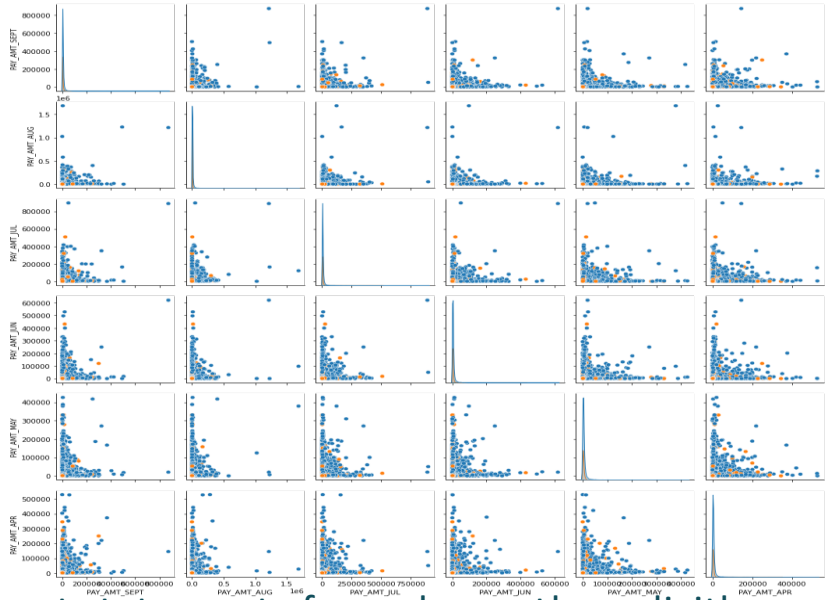
- Majority of defaulters are those who have credit limit between 20k to 300k
- After Cred limit 500k, numbers of defaulters are almost negligible

EDA Continued…

Pairplot of bill



Pairplot of pay



The 1st pairplot shows the distribution of bill amount statements for each month explicitly for defaulters and non-defaulters.

The 2nd pairplot shows the distribution of payment statements for each month explicitly for defaulters and non-defaulters.

FEATURE ENGINEERING

AI

ONE HOT ENCODING:

One-hot encoding is a technique which is used to convert or transform a categorical feature having string labels into K numerical features in such a manner that the value of one out of K (one-of-K) features is 1 and the value of rest (K-1) features is 0. It is also called as dummy encoding as the features created as part of these techniques are dummy features which don’t represent any real world features. Rather they are created for encoding the different values of categorical feature using dummy numerical features.

The primary need for using one-hot encoding technique is to transform or convert the categorical features into numerical features such that machine learning libraries can use the values to train the model .

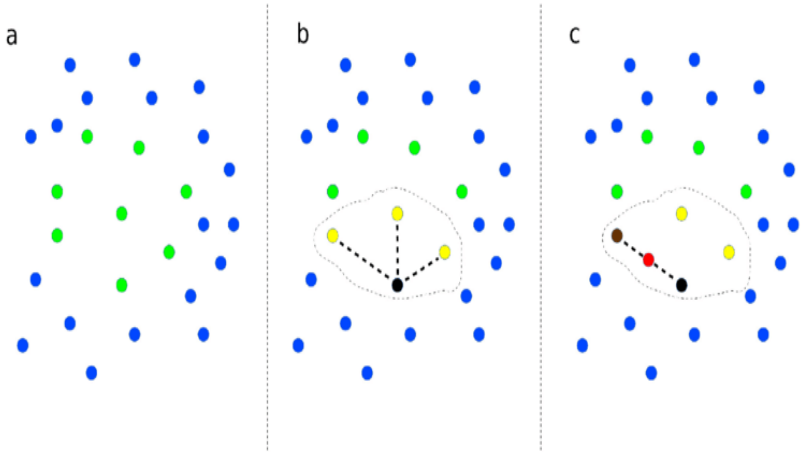
Here we perform one hot encoding on 'EDUCATION', 'MARRIAGE', and 'SEX’.

Synthetic Minority Oversampling Technique(SMOTE):

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.

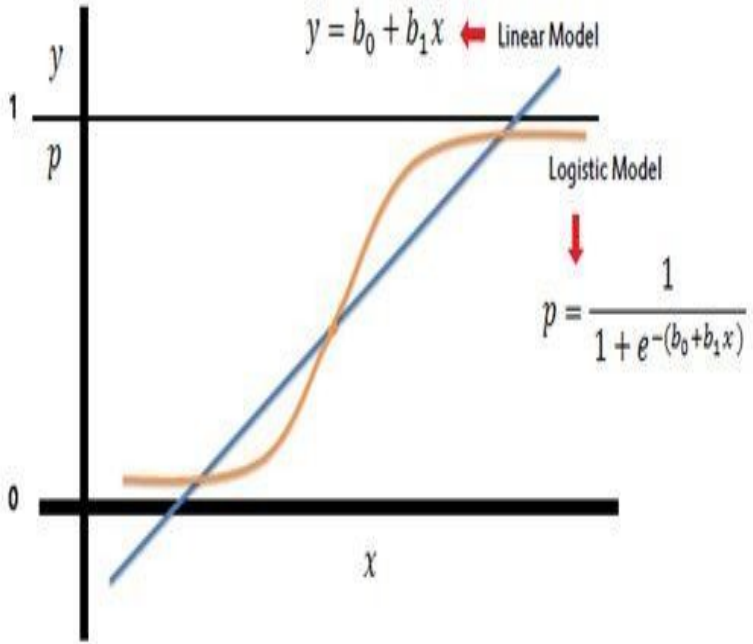


MODEL BUILDING:

LOGISTIC REGRESSION

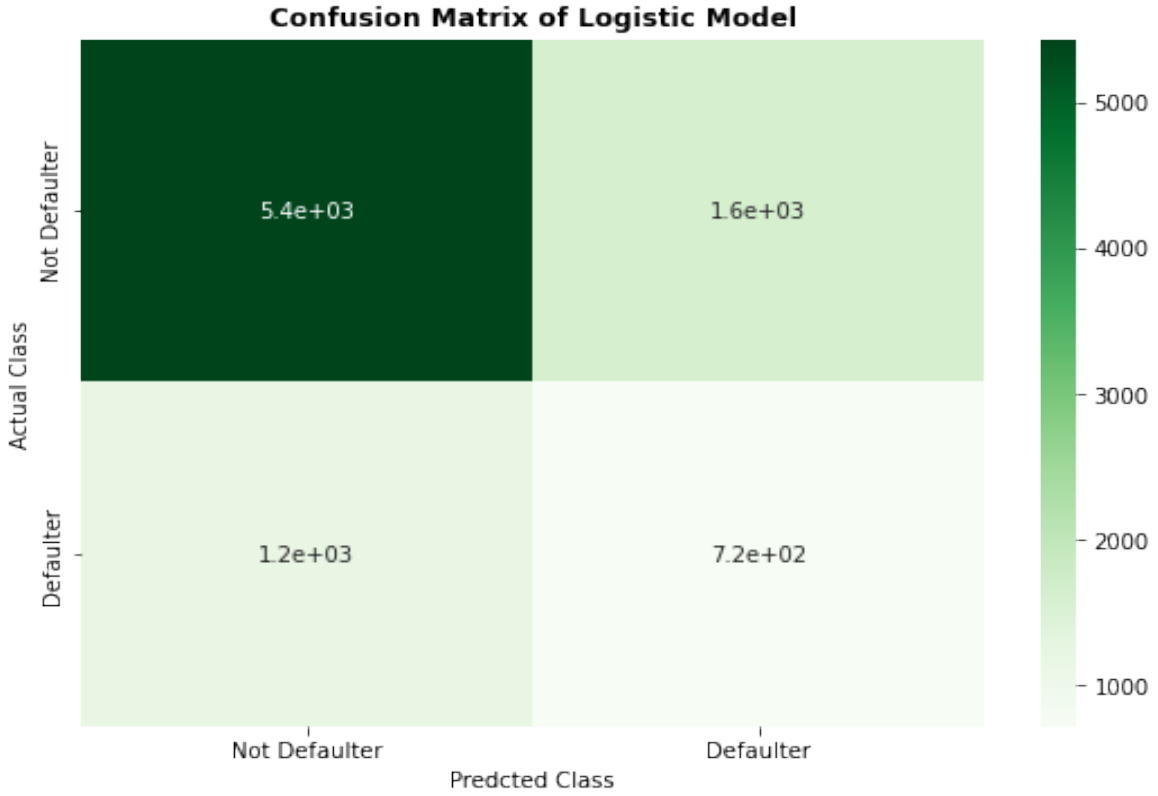
Logistic Regression is a Machine Learning algorithm and is basically used for binary classifications like yes-no, true-false, male-Female, etc.

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



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

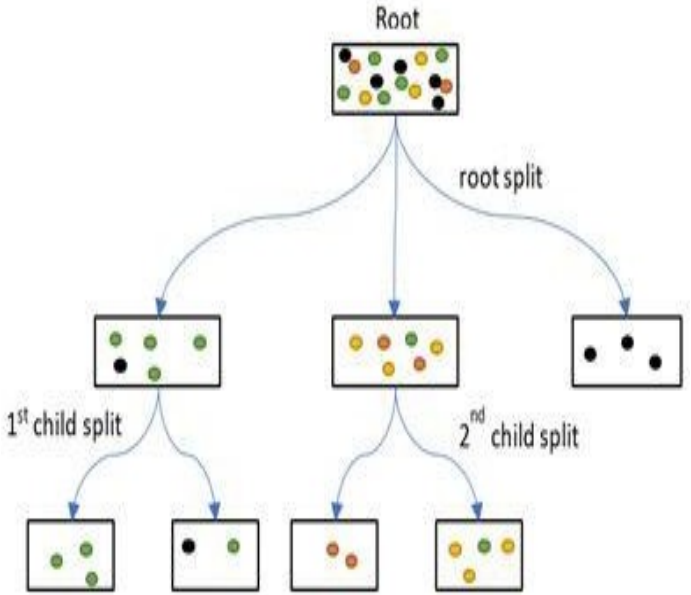


MODEL BUILDING (continued):

DECISION TREE CLASSIFIER

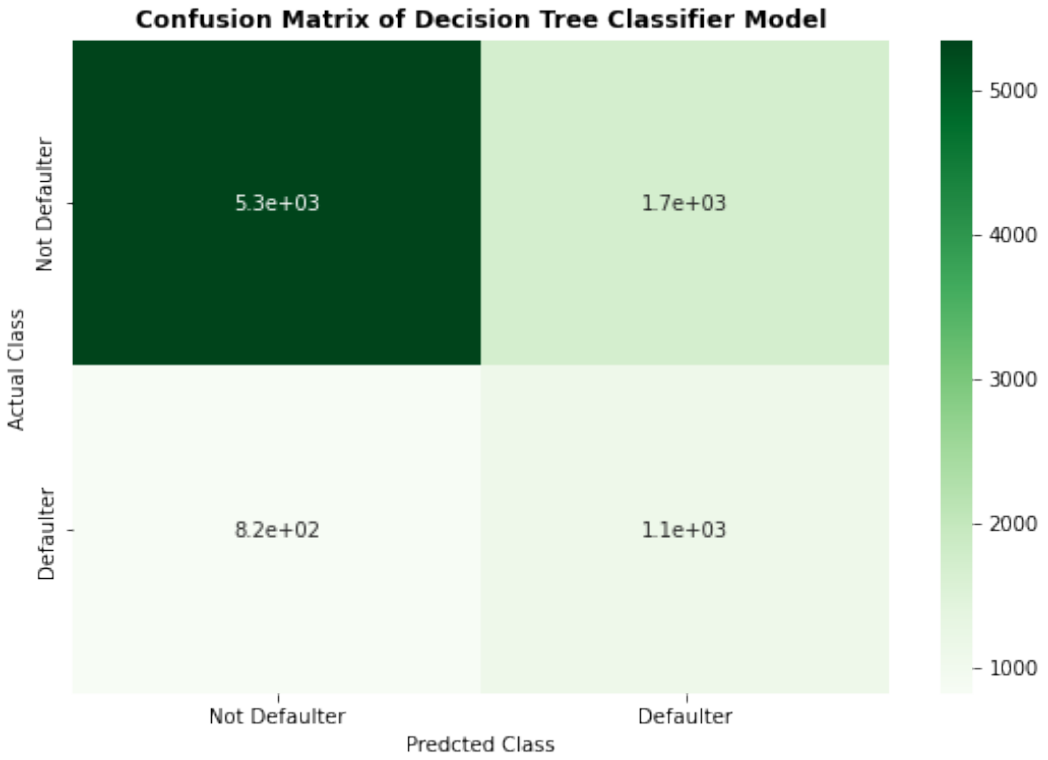
It is agraphical representation for getting all the possible solutions to a problem/decision based on given conditions.

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.



Confusion Matrix of Decision tree Classifier

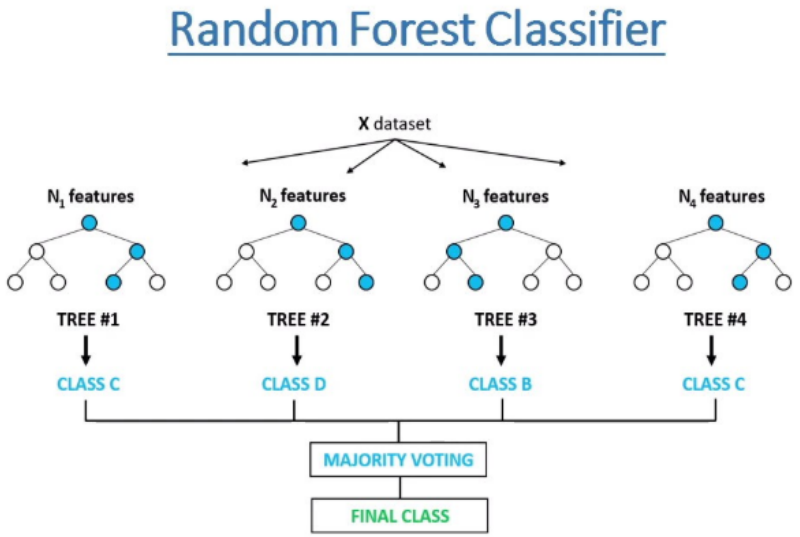
- Recall score of Decision Tree Classifier: 0.5773
- Precision score of Decision Tree Classifier: 0.5773
- F1 score of Decision Tree Classifier: 0.4691
- 6 ROC-AUC score of Decision Tree Classifier: 0.667200



MODEL BUILDING (continued):

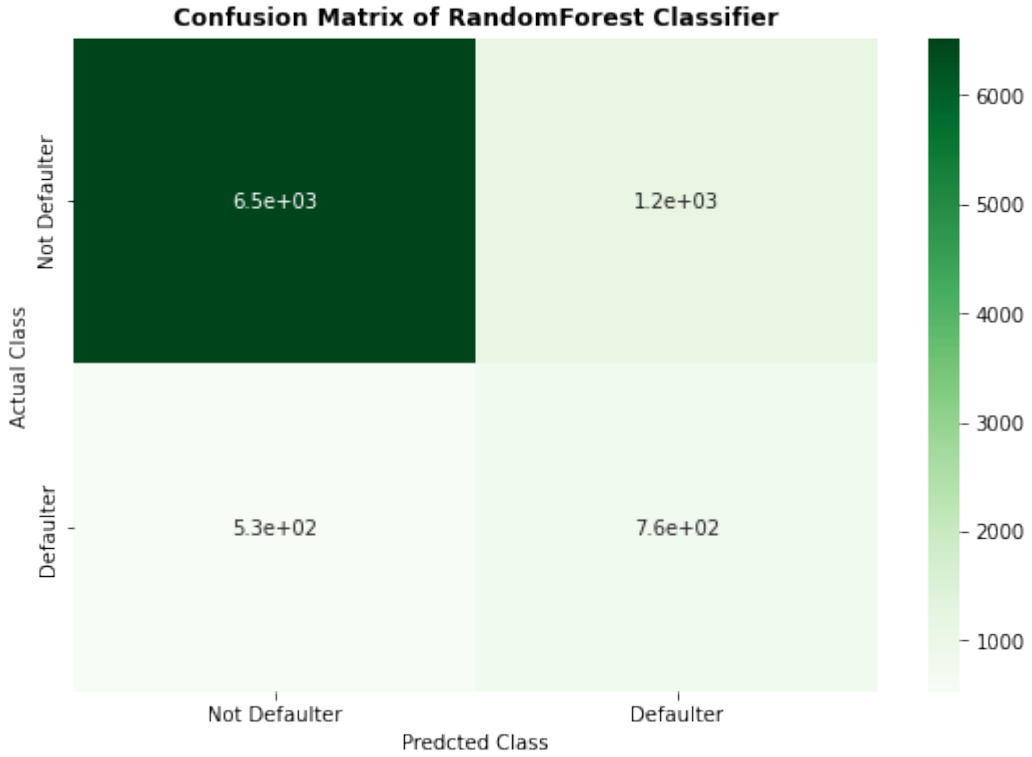
Random Forest Classifier

The random forest is a classification algorith
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



MODEL BUILDING (continued):

XG BOOST CLASSIFIER

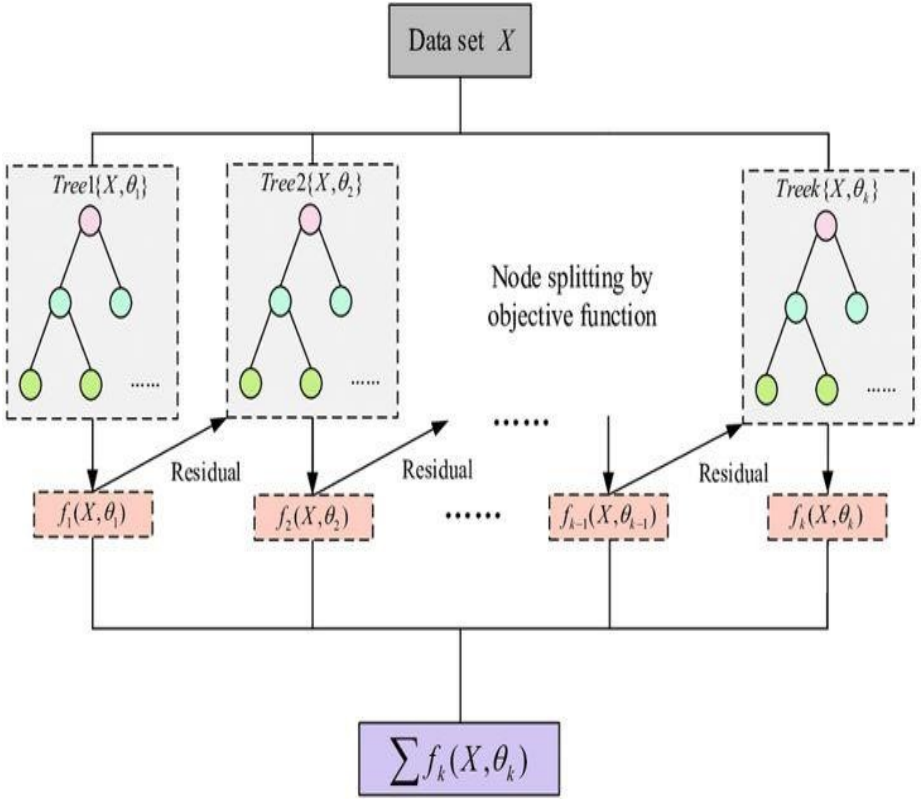
- Stands for:– e**X**treme **G**radient **B**oosting.

- XGBoost is a powerful iterative learning algorithm based on gradient boosting.

- Regularisation to avoid overfitting

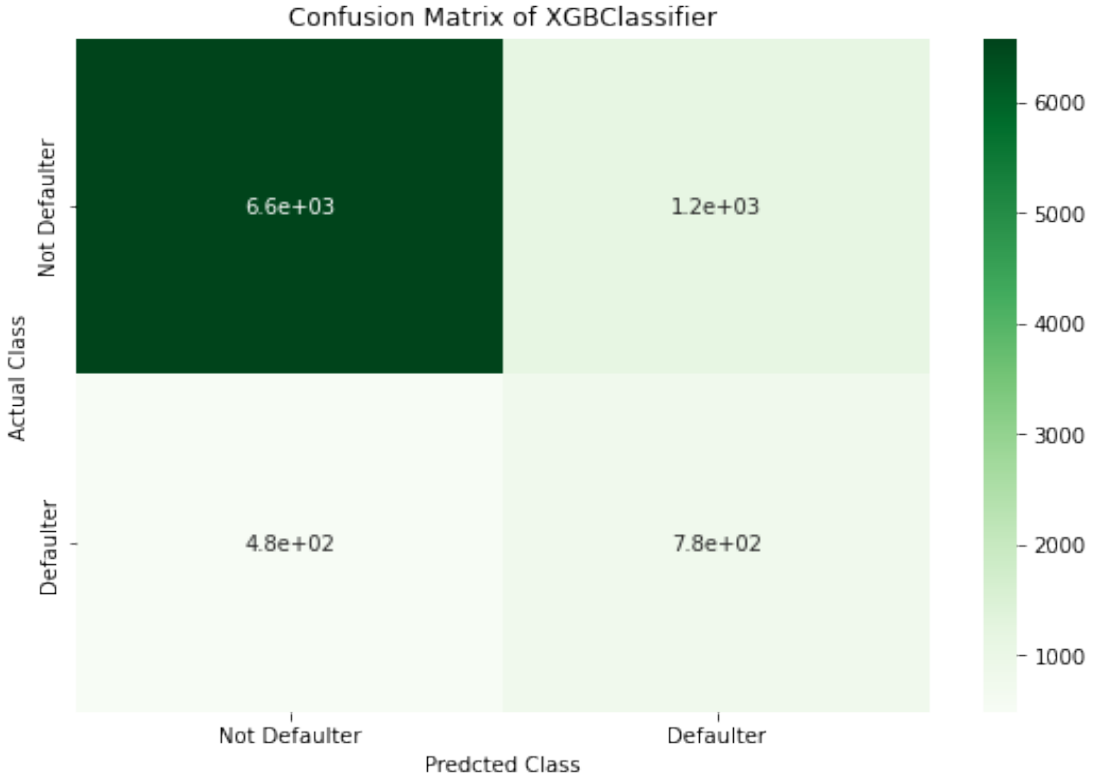
- Tree pruning using depth-first approach

- 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

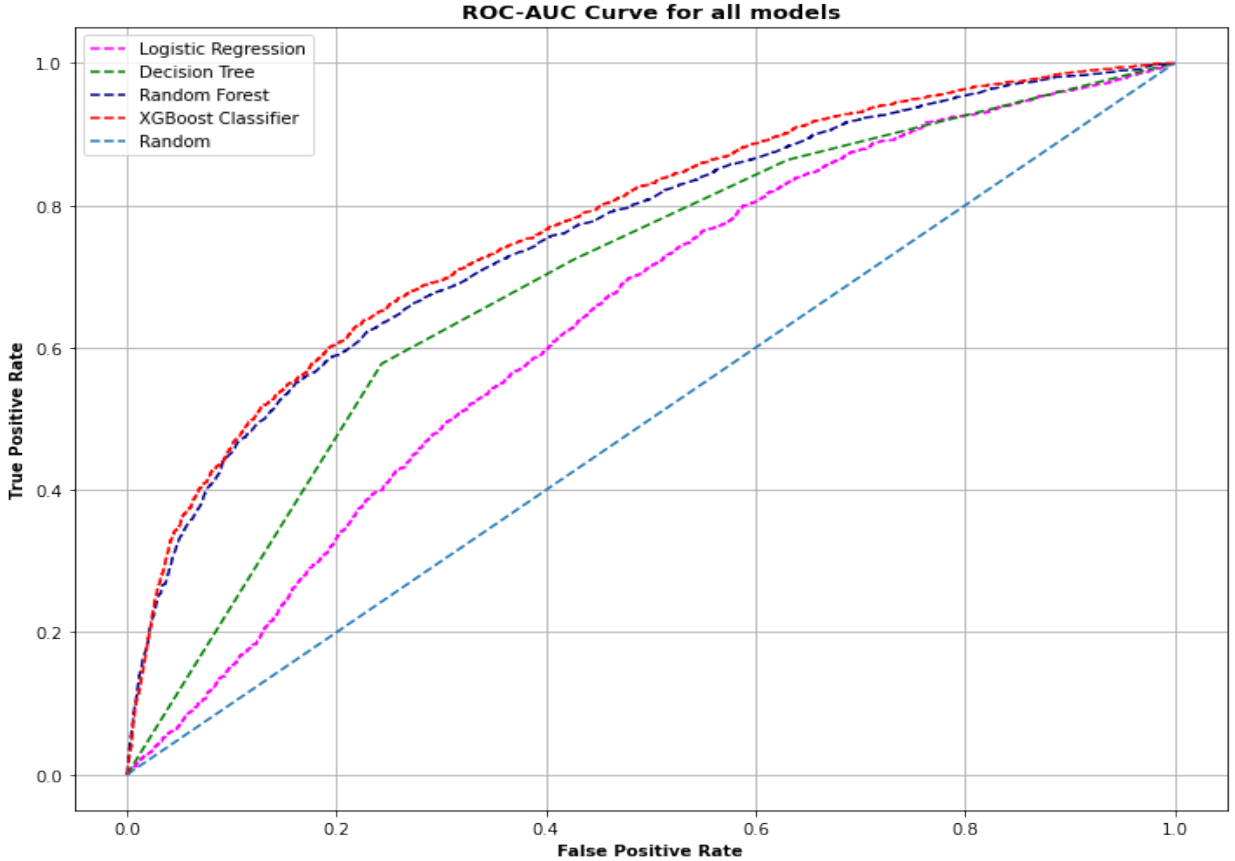


MODEL EVALUATION:

	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.



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.

Continued…

• Using a Logistic Regression classifier, we can predict with 65.97% accuracy, whether a customer is likely to default next month.

• With Decision Tree classifier having precision 57.77%, we can predict with accuracy of 71.83%, whether customer is likely to default next month.

• Using Random Forest, we can predict with accuracy of 81.13%, whether customer will be defaulter in next month.

• 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 metrices(precision, F1-score and ROC-AUC score).
On behalf of these metrices 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.

THANK YOU!!