Capstone Project-3

Credit Card Default Prediction

Submitted by

Shubham Kadu

Data science trainee, Almabetter

Contents:

- Introduction
- Problem Statement and Overview
- Data Summary
- Exploratory data analysis
- Data Preprocessing
- ML-Model Algorithms
- Model Evaluations
- Conclusion

Problem Statement :

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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

Overview

In today's world credit cards have become a lifeline to a lot of people so banks provide us with credit cards. Credit card is a commonly used transaction method in modern society and one of the main business of banks. Credit card fraud is a huge ranging term for theft and fraud committed using or involving at the time of payment by using this card. The purpose may be to purchase goods without paying or to transfer unauthorized funds from an account. Credit card fraud is also an add-on to identity theft. also For banks, it helps the bank to generate interest revenue but at the same time, it raises the liquidity risk and credit risk to the bank. In order to control the cash flow and risk, detecting the customers with default payments next month could play an important role in estimating the potential cash flow and risk management.

Data Information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   ID                         30000 non-null  int64
 1   LIMIT_BAL                  30000 non-null  int64
 2   SEX                        30000 non-null  int64
 3   EDUCATION                  30000 non-null  int64
 4   MARRIAGE                   30000 non-null  int64
 5   AGE                        30000 non-null  int64
 6   PAY_0                      30000 non-null  int64
 7   PAY_2                      30000 non-null  int64
 8   PAY_3                      30000 non-null  int64
 9   PAY_4                      30000 non-null  int64
 10  PAY_5                      30000 non-null  int64
 11  PAY_6                      30000 non-null  int64
 12  BILL_AMT1                  30000 non-null  int64
 13  BILL_AMT2                  30000 non-null  int64
 14  BILL_AMT3                  30000 non-null  int64
 15  BILL_AMT4                  30000 non-null  int64
 16  BILL_AMT5                  30000 non-null  int64
 17  BILL_AMT6                  30000 non-null  int64
 18  PAY_AMT1                   30000 non-null  int64
 19  PAY_AMT2                   30000 non-null  int64
 20  PAY_AMT3                   30000 non-null  int64
 21  PAY_AMT4                   30000 non-null  int64
 22  PAY_AMT5                   30000 non-null  int64
 23  PAY_AMT6                   30000 non-null  int64
 24  default payment next month  30000 non-null  int64
```

Data Summary :

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,4=others, 5=unknown, 6=unknown)

5. MARRIAGE: Marital status (1=married, 2=single, 3=others)

6. AGE: Age in years

7. PAY_0-6: History of past payments from April to September

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

9. PAY_AMT1-6: Amount of previous payment from April to September 2005 (NT dollar)

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

Exploratory Data Analysis (EDA):

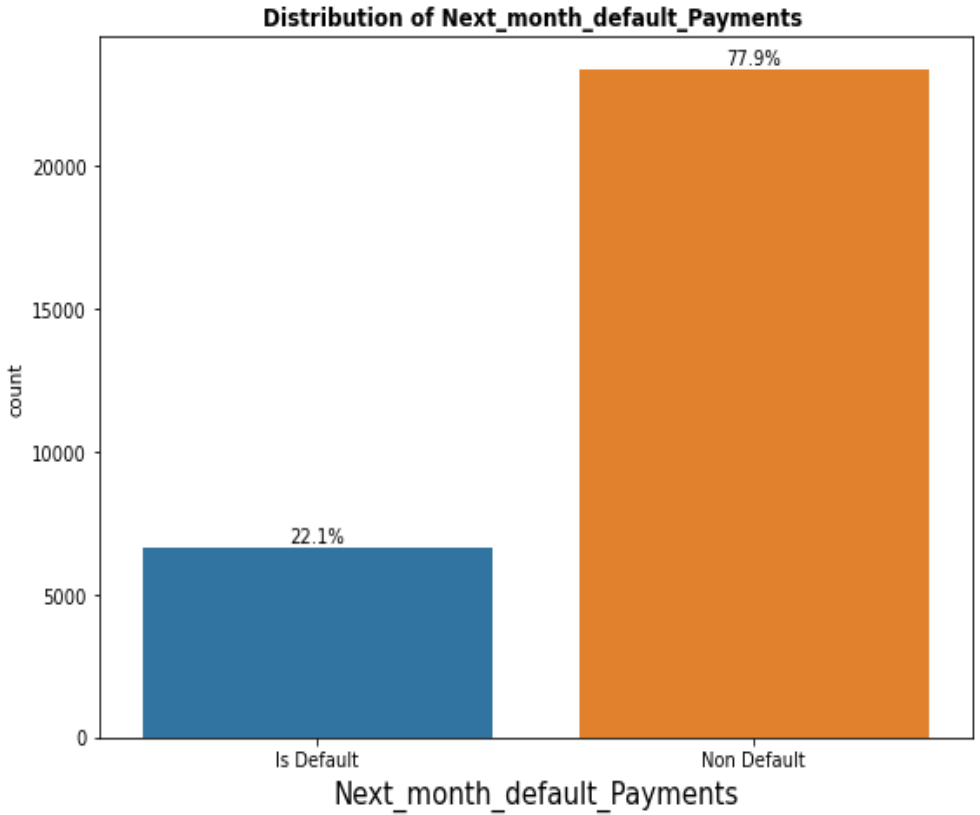
Dependent Variable: Defaulter

As we can see from the graph Here, there is a huge difference between non-defaulter(0) and defaulter(1).

Defaulters are less than the Non-Defaulters.

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

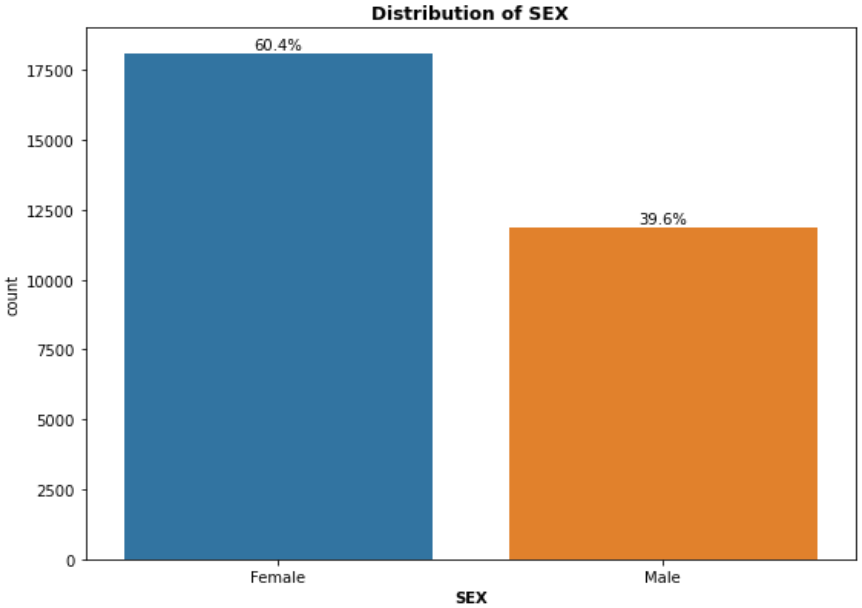
Both classes are not in proportion and we have an imbalanced dataset. We need to normalize the data in the next step.



Univariate Analysis

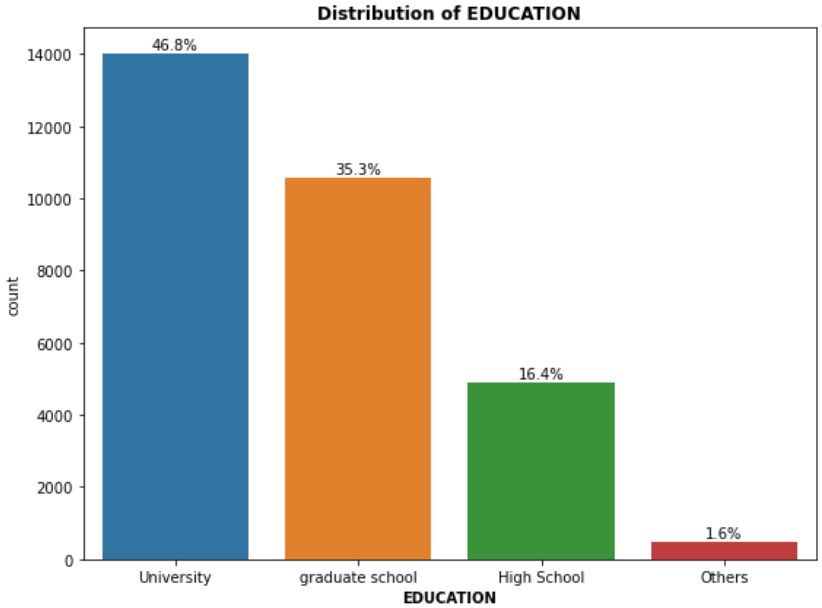
SEX

- From the below graph, we can see that the Number of Male credit holders is less than females.
- Approximately 40% are male and
- 60% are Female.



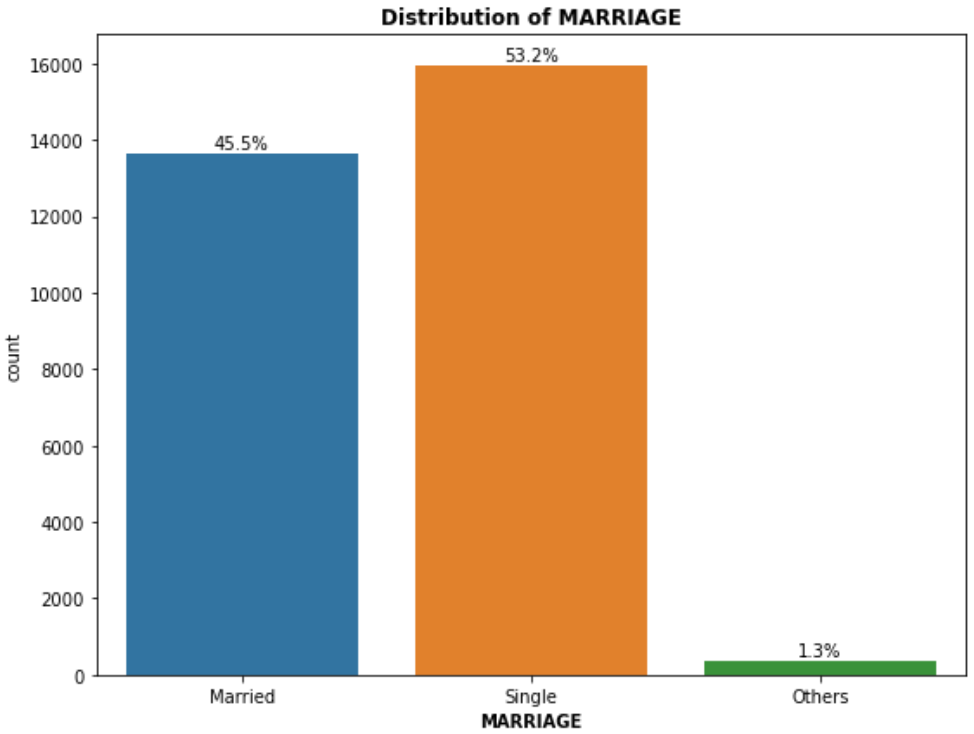
EDUCATION

- From the below graph, we can see that More credit holders are university students followed by Graduates and then High school students



MARRIAGE

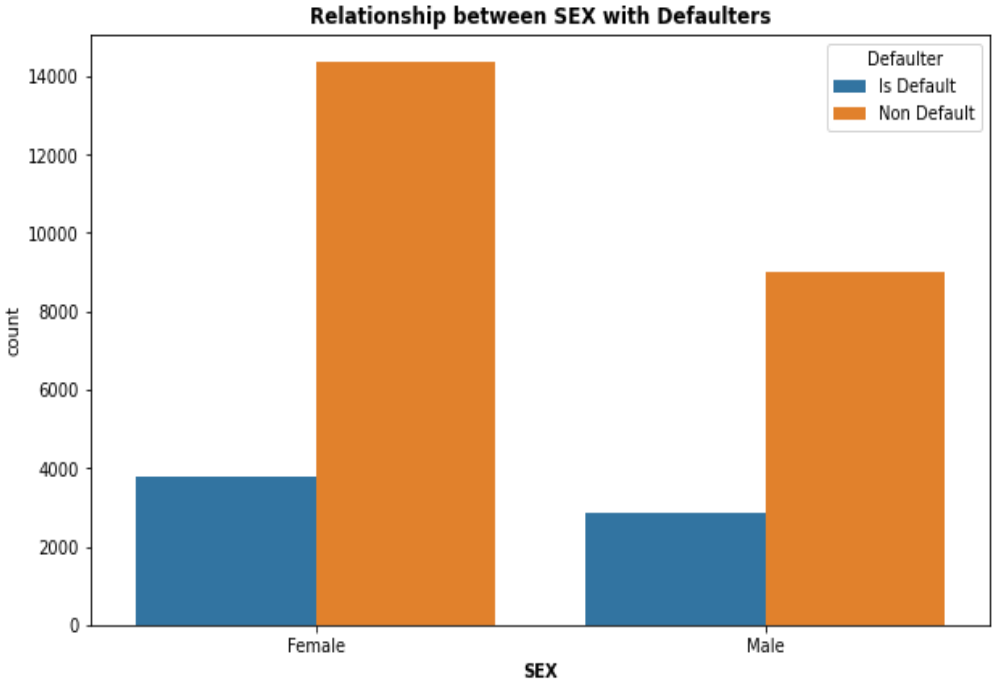
- From the above data analysis, we can say that More number of credit cards holder are Single as compared to Married and others.



Bivariate Analysis

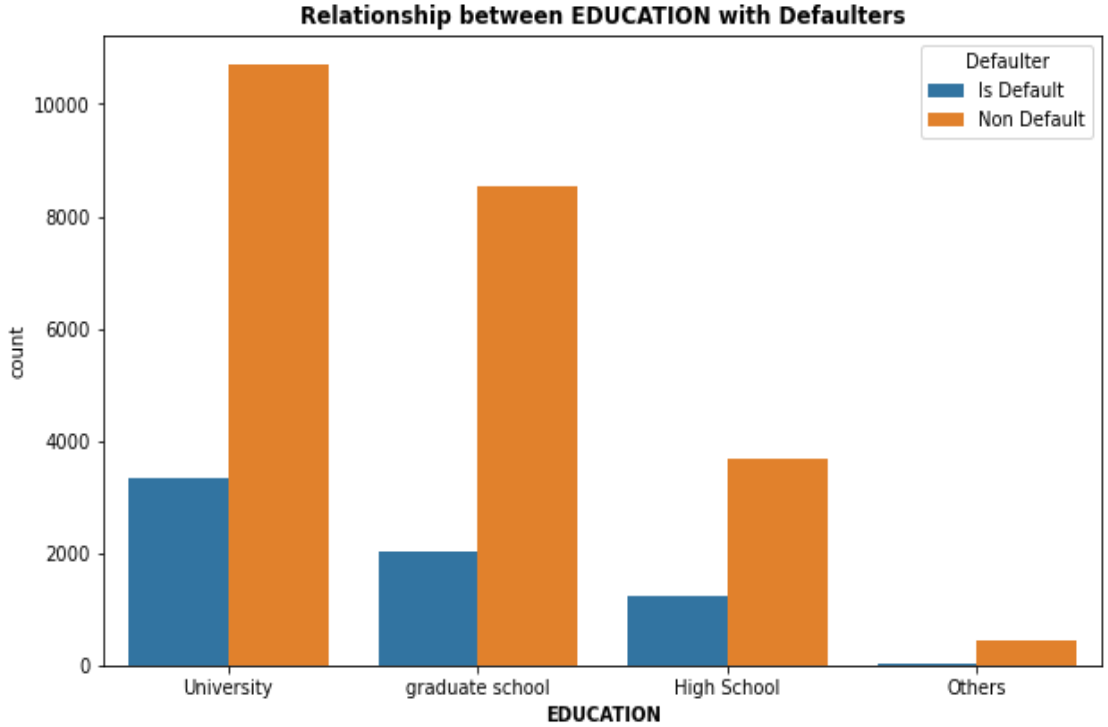
SEX vs DEFAULTER

It is evident from the above graph that the number of defaulters has a high proportion of females.



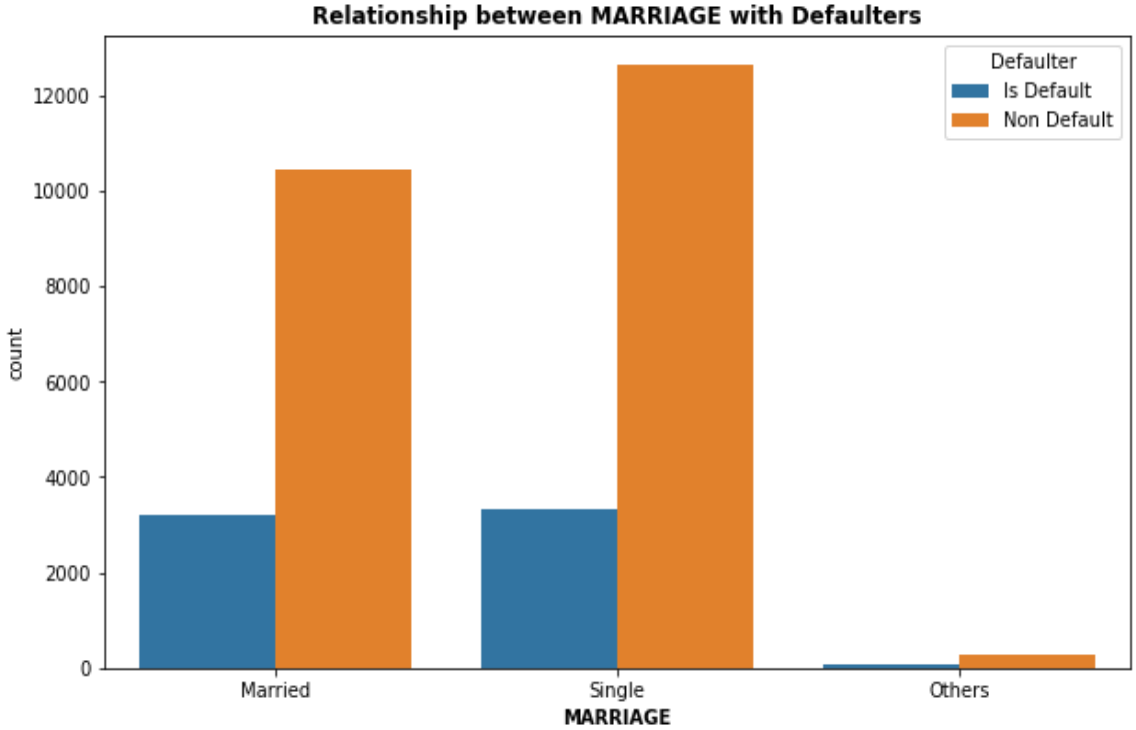
EDUCATION vs DEFAULTER

From the above graph, it is clear that those people who are university students have higher default payments w.r.to graduates and high school people.



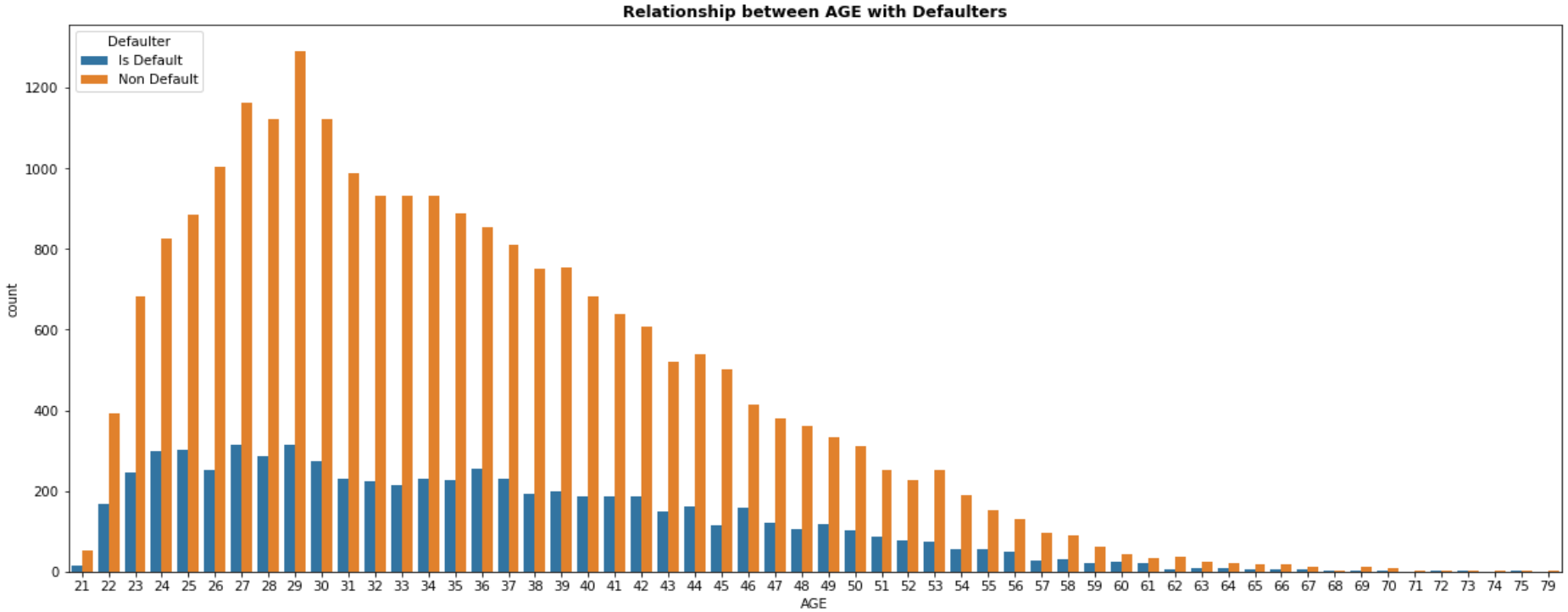
MARRIAGE vs DEFAULTER

From the above graph, Here it seems that married, or single is most likely to default.



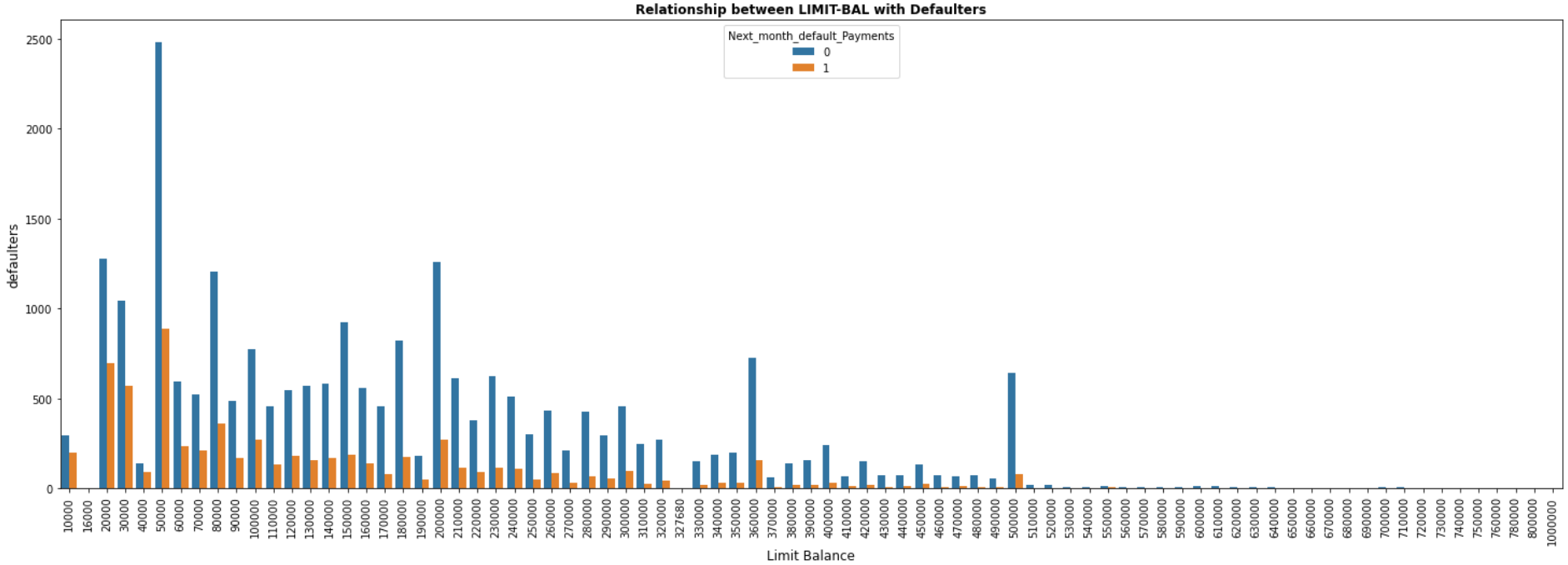
AGE

More number of credit card holders aged between 26-32 years and 29 years age is the highest uses of credit cards. Age above 60 years old rarely uses a credit card. Also, more Defaulters are between 27-29 years.



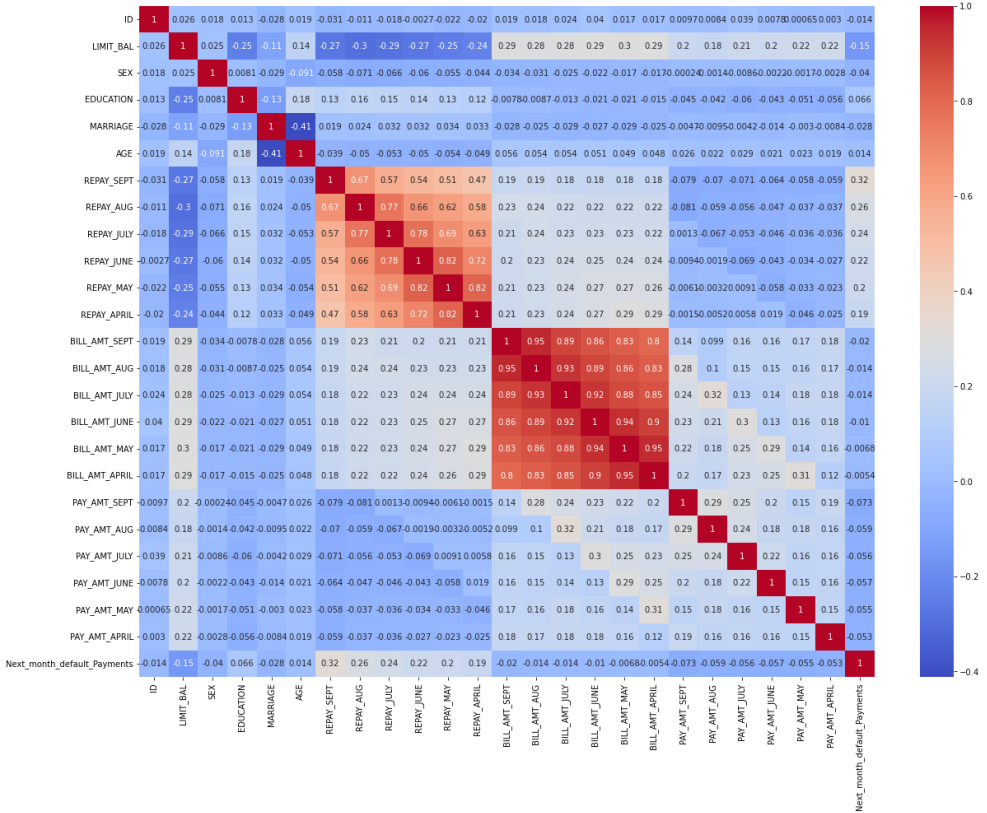
LIMIT BALANCE

From the below plot, we clear that the majority of the defaulters are those who have a credit limit balance between 20,000 to 3,00,000 After the credit limit of 5,00,000, the number of defaulters is almost negligible.



Correlation Analysis

- The heatmap shows some high correlations between variables and also visualizes how each parameter's correlation with respect to every other parameter.
- Above heatmap It seems that there is some negatively correlated feature like age and marriage. and ID is unimportant and has no role in prediction so we will remove it.



ONE HOT ENCODING

One hot encoding is a process by which categorical variables are converted into the form of a numerical variable that could be provided to ML algorithms to do a better job in prediction.

Here we perform one hot encoding on 'EDUCATION','MARRIAGE','PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR' and label encoding for 'SEX'

SMOTE

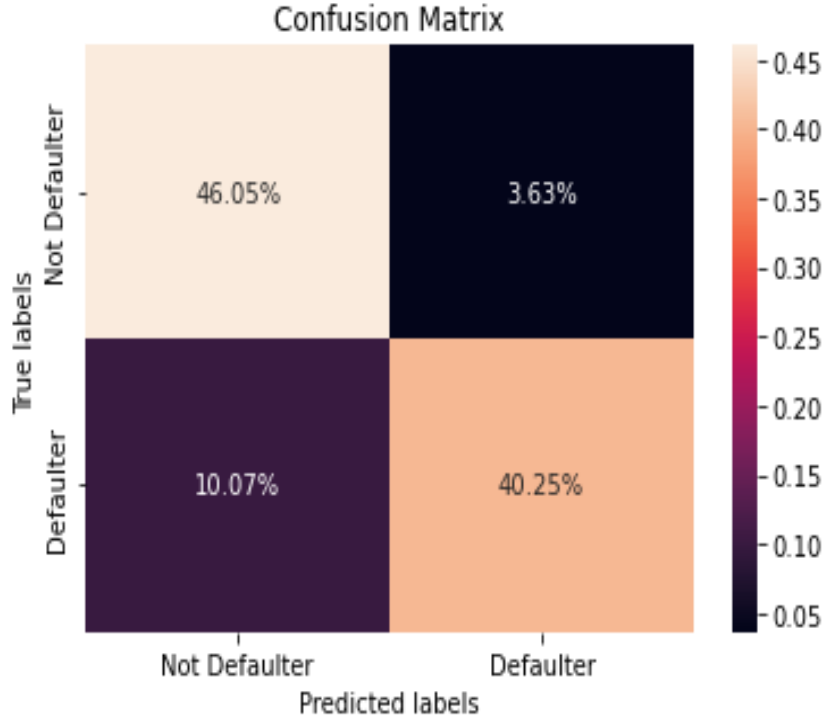
Our dataset is imbalanced which can lead to Biasness While Building the Model. For Balancing We Use SMOTE.

SMOTE (Synthetic Minority Oversampling Technique) – Oversampling is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. After performing SMOTE operation we get this balance dataset

Machine Learning Model

❖ 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 takes the linear combination and applies a sigmoid function (logit). The Sigmoid curve gives a value between 0&1.

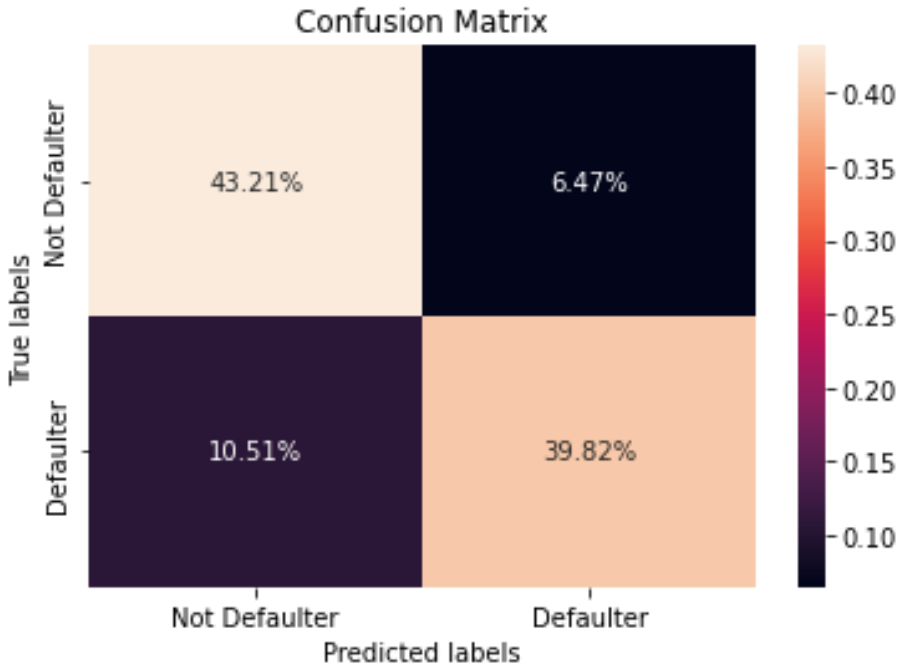


Evaluation Metrics:

Training accuracy = 0.86597
Testing Accuracy = 0.8644
Precision Score = 0.8042
Recall Score = 0.9160
F1_Score = 0.8565
ROC_AUC score = 0.8697

❖ 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 the Decision tree algorithm is to find the relationship between the target column and the independent variables and Express it as a tree structure

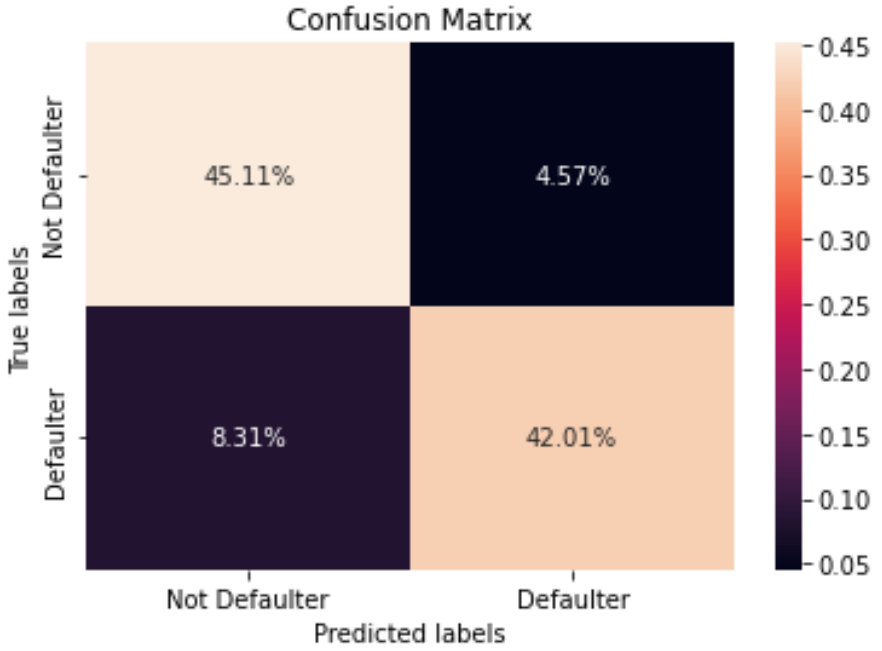


Evaluation Metrics :

Training accuracy = 0.8503
Testing accuracy = 0.8317
Precision score = 0.8600
Recall score = 0.7706
F1_score = 0.8217
ROC-AUC score = 0.83054

❖ Random Forest Classifier

- The random forest is a classification algorithm consisting of many decision 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

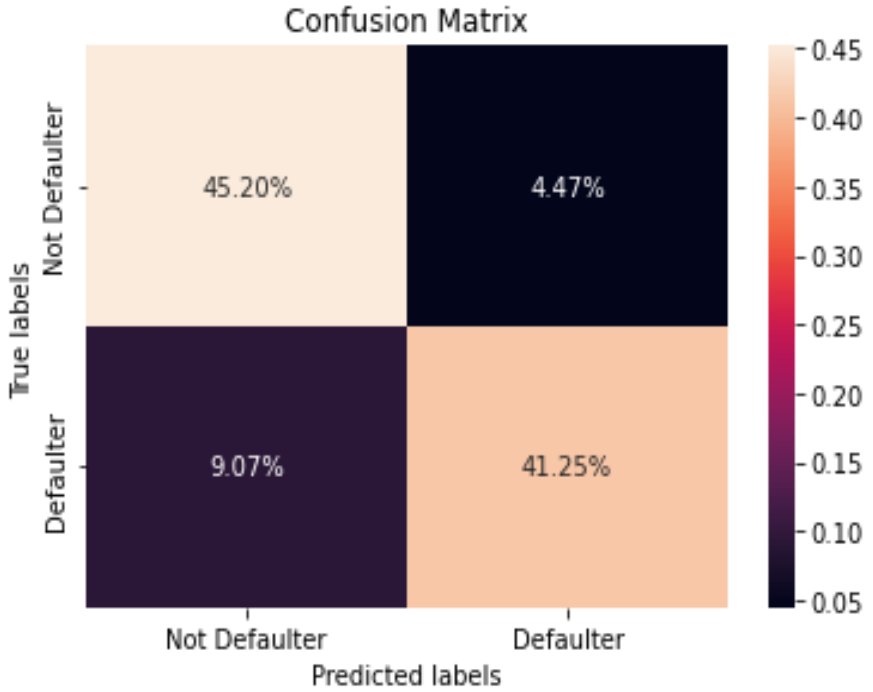


Evaluation Metrics :

Training accuracy = 0.9995
Testing Accuracy = 0.8747
Precision score = 0.8395
Recall score = 0.9046
F1_score = 0.8708
ROC_AUC score = 0.8765

❖ XG-BOOST

- XGBoost is a powerful iterative learning algorithm based on gradient boosting.
- Regularizations to avoid overfitting Tree pruning using a depth-first approach.
- It is generally used for very large dataset

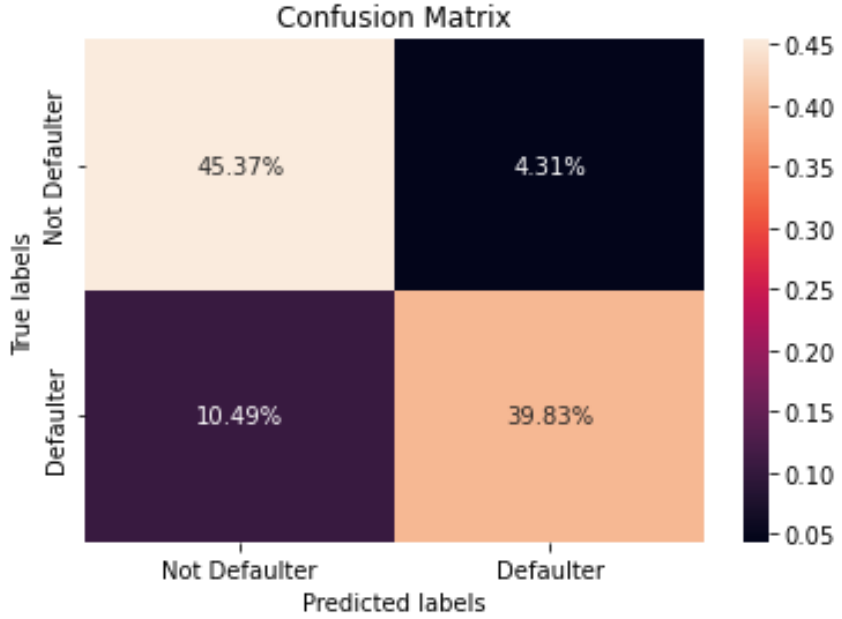


Evaluation Metrics :

Training accuracy = 0.9156
Testing accuracy = 0.8656
Precision score = 0.8186
Recall score = 0.9052
F1_score = 0.8597
ROC_AUC score = 0.8689

❖ KNN-Classifier

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows.



Evaluation Metrics :

Training accuracy = 0.8724
Testing accuracy = 0.8548
Precision score = 0.8041
Recall score = 0.8968
F1_score = 0.8479
ROC_AUC score = 0.8586

❖ Over all Evaluation matrix

Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score	ROC_AUC Score
Logistic Regression	0.865972	0.864406	0.804253	0.916043	0.856515	0.869790
Decision Tree	0.850385	0.831723	0.880059	0.770619	0.821711	0.832118
Random Forest	0.999553	0.874716	0.839562	0.904610	0.870873	0.876566
XG-BOOST	0.915610	0.865638	0.818686	0.905244	0.859792	0.868902
KNN-Classifier	0.872457	0.854873	0.804124	0.896809	0.847941	0.858602

❖ ROC Curve Comparison

A ROC curve (Receiver operating characteristic) is a graph showing the performance of a classification model at all classification thresholds.

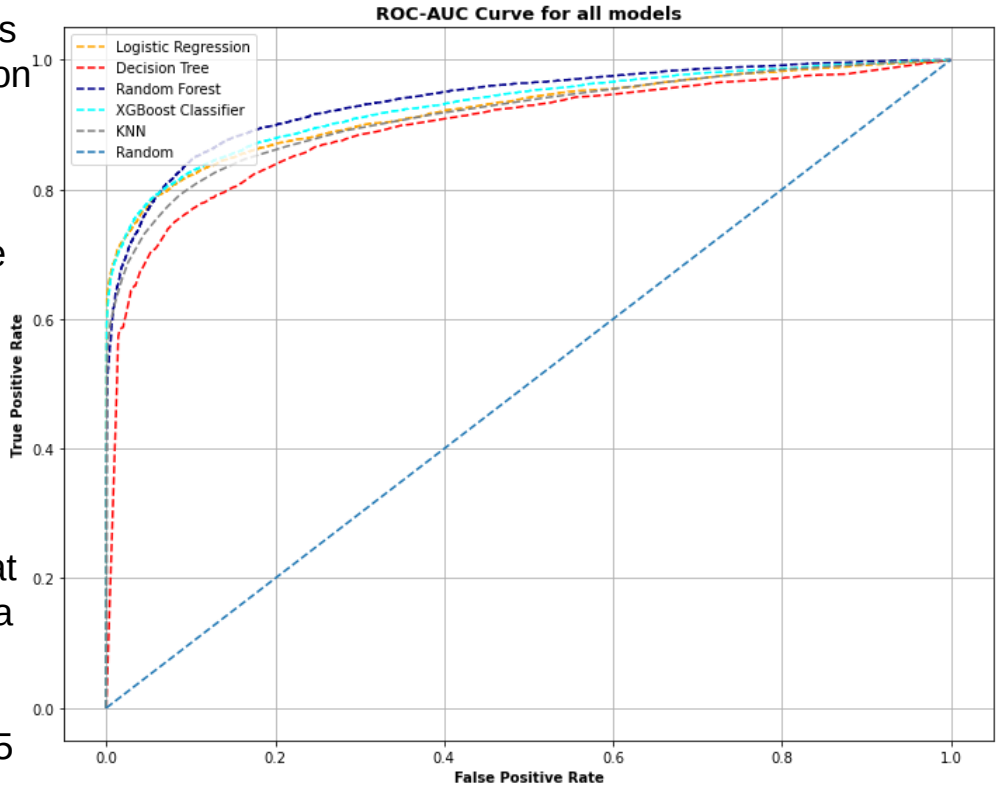
It summarizes the model's performance by evaluating the trade-offs between the true positive rate (sensitivity) and the false positive rate(1-specificity).

This curve plots two parameters:

1. True Positive Rate
2. False Positive Rate

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve.

For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about the success rate



❑ Conclusion:

1)first We started with data inspection, viewed the data distribution

2)By Visualization we have checked the distribution of defaulters vs non-defaulters and we see around 78% are non-defaulters and 22% are defaulters.

3)the distribution of sex, Education, and Marriage with respect to the defaulter. and we found in Sex more defaulter is Female, in Education, more number of the defaulters is a university students and in Marriage more number of the defaulters by single.

4)After that we built a model(Logistic Regression, Decision Tree, Random forest, XGBoot classifier, and KNN), and all of them in, the best accuracy has obtained from the Random Forest Classifier.

5)Using a Logistic Regression classifier, we can predict with 86.38% accuracy, whether a customer is likely to default next month or not. Using the Decision Tree classifier, we can predict with 82% accuracy whether a customer is likely to default next month or not. Using Random Forest, we can predict with 87% accuracy whether a customer will be a defaulter in the next month or not. Using XGBoost Classifier, we can predict with 86.64% accuracy whether a customer will be a defaulter in the next month or not. And By applying KNN Classifier with 85% accuracy whether a customer will be a defaulter in the next month or not.

6)From the Above evaluation table Logistic regression model has the highest recall, if the business cares about recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. but Since the Random Forest classifier has also a higher Recall score.so I would recommend Random Forest.

7) From the above evaluation table we can also see that the Random forest Classifier having Recall, F1-score, and ROC Score values equals 90.46%, 87.08%, and 87.65% resp. and XGBoost Classifier having Recall, F1-score, and ROC Score values equals 90.29%, 85.95%, and 86.89% resp.

8)From the models that are applied to the dataset, We can conclude that these two Random Forest and XGBoost are giving the best evaluation metrics(Recall, F1-score, and ROC-AUC score) and with the help of these two models we are the best to predict whether the credit card is the default or not default according to our analysis.

Thank you!