Capstone Project – 3
Supervised ML – Classification
Credit Card Default Prediction

By-
KISHOR SHIVAJI PATIL
Data Science Trainee,
AlmaBetter, Banglore.

Presentation Content -

- Introduction
- Problem Statement
- Data Summary
- Approach Overview
- Exploratory Data Analysis
- Model Overview
- Implementing ML algorithms
- Evaluating Models
- Challenges
- Conclusion

Introduction

With growing technology in this 21st century, many of Customer have become dependent on credit card for daily transactions. This increase in use of credit card has resulted in increase of Credit card frauds. The most common issue in providing these facilities are people not being able to pay the bills. These people are what we call “defaulters”.

Problem Statement

- Predicting whether a customer will default on his/her credit card

OBJECTIVE

Credit Card holder data

Classifier Model

No Default

Default

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

Features:

ID: ID of each client.
LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit.

SEX: Gender (1=male, 2=female).
EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown).
MARRIAGE: Marital status (1=married, 2=single, 3=others).
AGE: Age in years.

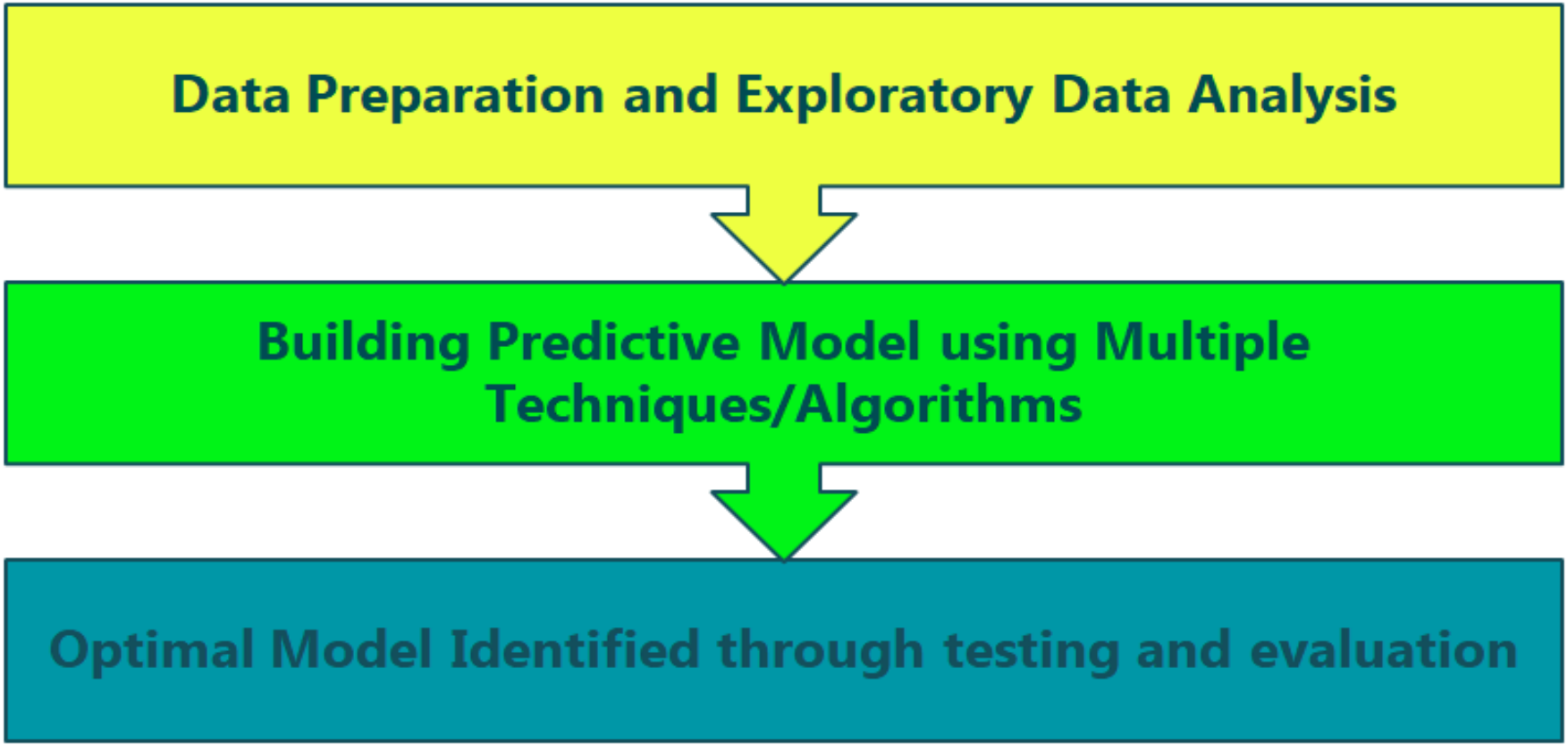
— —

PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2: Repayment status in August, 2005 (scale same as above)
…;
PAY_6: Repayment status in April, 2005 (scale same as above)

Features(Contd.):

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
…;
BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
— —
PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
…;
PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
— —
Default payment next month: Default payment (1=yes, 0=no)

Approach:

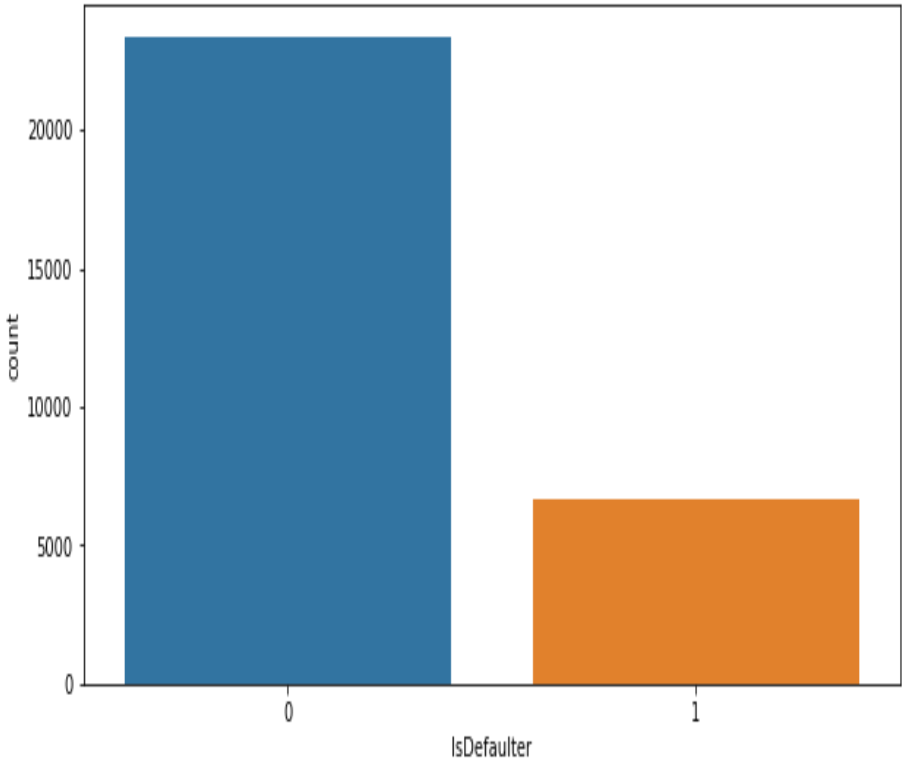

Data Preparation and Exploratory Data Analysis
Building Predictive Model using Multiple Techniques/Algorithms
Optimal Model Identified through testing and evaluation

Basic Exploration

- Dataset for Taiwan.
- Shape of data is 30000 rows and 25 columns
- 6 months payment and bill data available.
- No null data.
- 9 Categorical variables present.
- ID column can be drop

ANALYSIS OF DEPENDENT VARIABLE

As we can see from a graph that both classes are not in proportion and we have imbalanced dataset. We need to do normalize the data in next step.



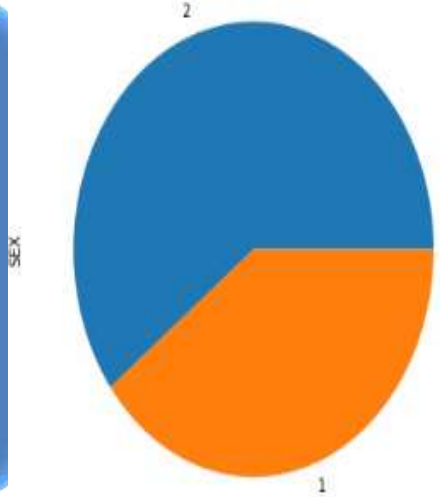
0 - Not Default

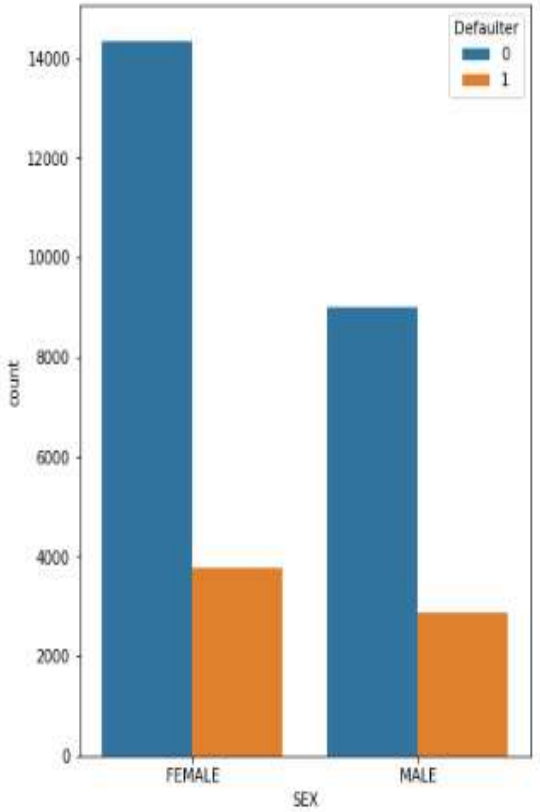
1 – Default

Defaulters are less than the Non Defaulters in the given dataset.

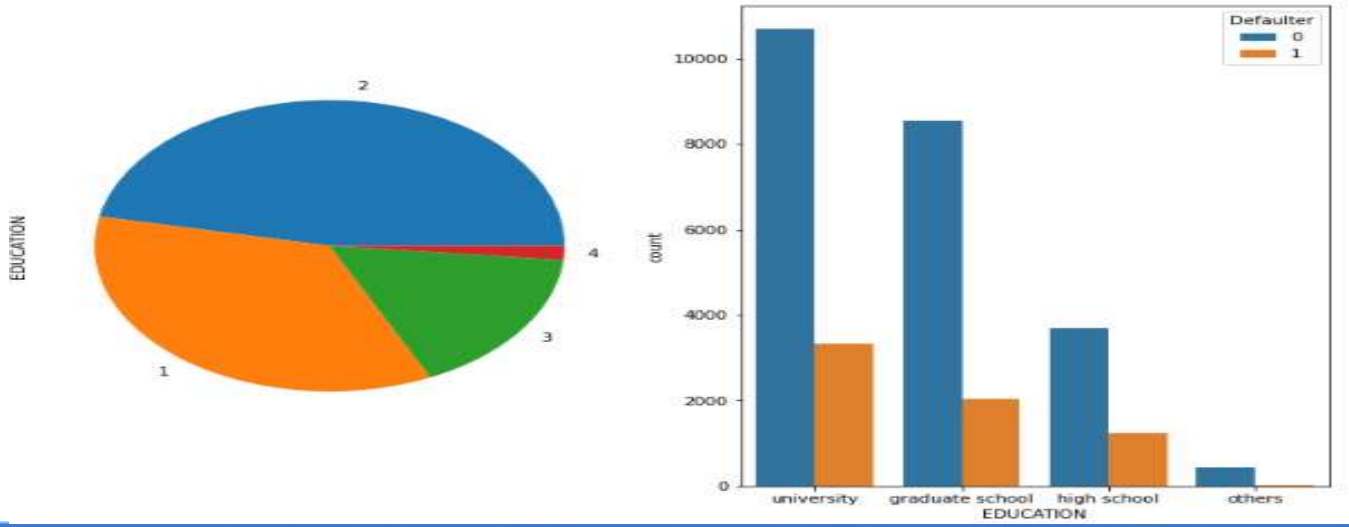
ANALYSIS OF SEX VARIABLE

- 1 - Male
- 2 - Female
- Number of Male credit holder is less than Female.
- It is evident from the graph that the number of defaulter have high proportion of males





ANALYSIS OF EDUCATION VARIABLE

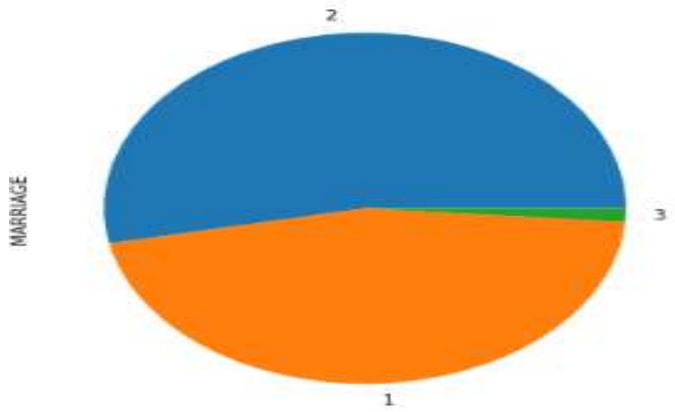


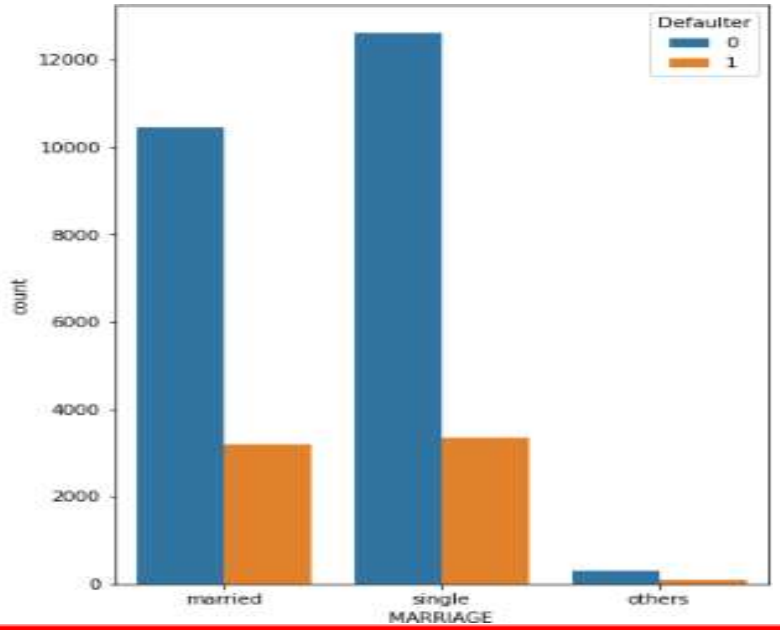
❖1=graduate school, 2=university, 3=high school, 4=others

❖From the above left side plot we can say that more number of credit holders are university students followed by Graduates and then High school students.

❖From the right side plot we can see that High School Students have higher proportion of Defaulters followed by University, Graduates & Others

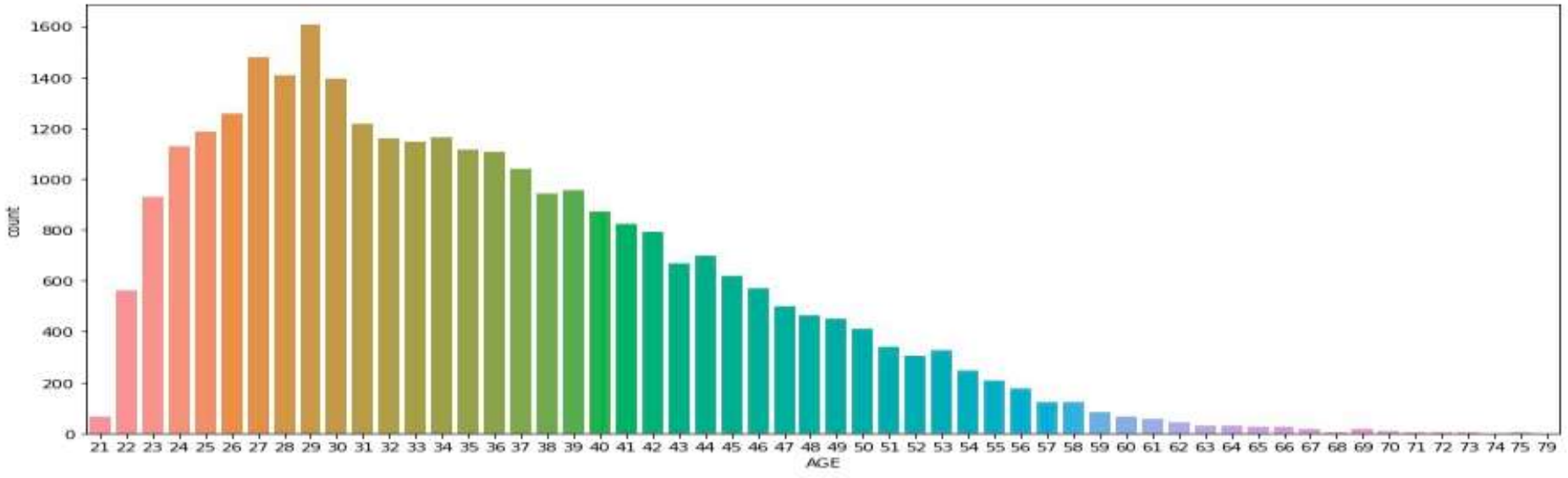
ANALYSIS OF MARRIAGE VARIABLE





- 1 – married, 2 – single, 3 – others.
- More number of credit cards holder are Single.
- As We can see in Right side Graph Others have Higher proportion of Defaulters & There is no Significant Co-relation of Defaulter Risk & Maritial Status.

ANALYSIS OF AGE VARIABLE



❖We can see more number of credit cards holder age are between 26-30 years old.

❖ Age above 60 years old rarely uses the credit card.

Modeling Overview

Supervised learning/Binary Classification

Imbalance data with 78% non-defaulters and 22% defaulters

Models Used:

- Logistic Regression
- Decision Trees
- Random Forest
- SVC
- XGBoost

Modeling Steps

AI

Data Preprocessing

- Feature selection
- Feature engineering
- Train test data split(75%-25%)
- SMOTE (Synthetic Minority Oversampling Technique)

Data Fitting and Tuning

- Start with default model parameters
- Hyperparameter tuning
- Measure AUC- ROC on training data

Model Evaluation

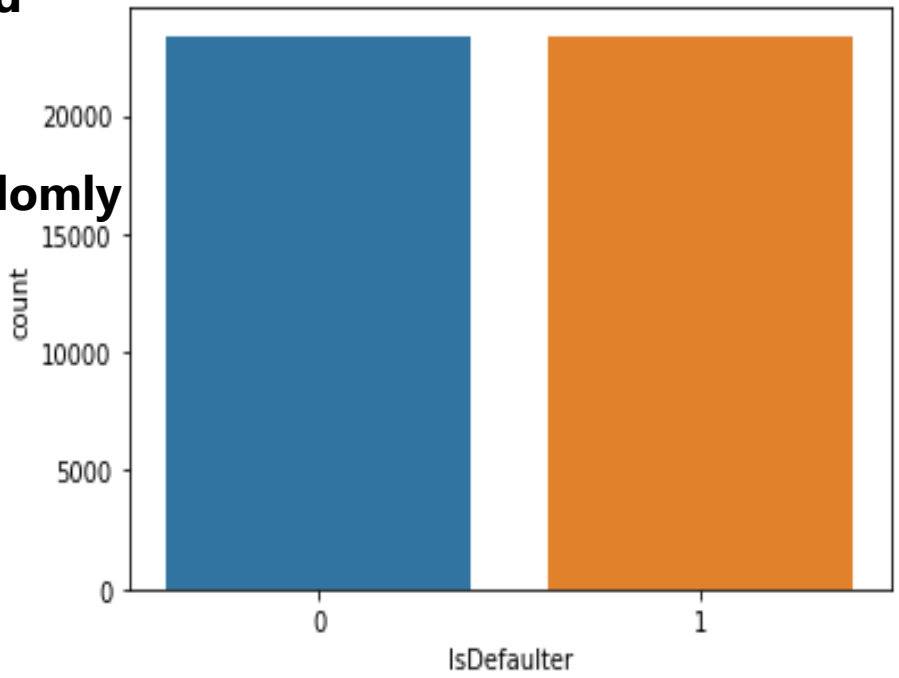
- Model testing
- Precision_Recall Score
- Compare with the other models

SMOTE -

SMOTE (Synthetic Minority Oversampling Technique) – Oversampling is one of the most commonly used oversampling method 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



ONE HOT ENCODING & LABEL ENCODING –

- One hot encoding is a process by which categorical variables are converted into a form 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'.
- label encoding for 'SEX'.

MODEL BUILDING

LOGISTIC REGRESSION

RANDOM FOREST

SVC

XGBOOST

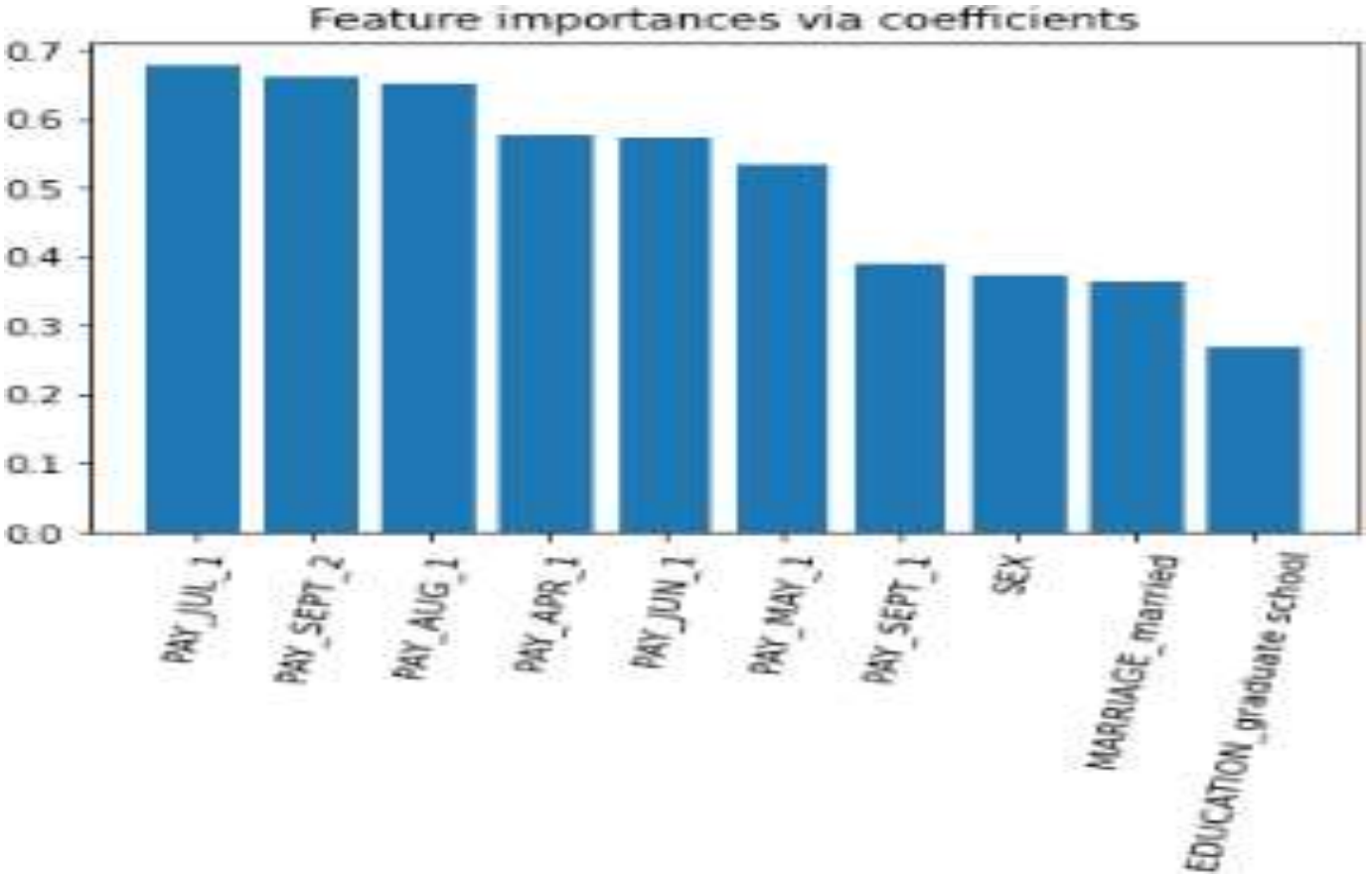
Logistic Regression

Parameters :C = 0.01 , Penalty = L2

Results:

- The accuracy on test data is 0.752867659647320 7
- The precision on test data is 0.6884095189179935
- The recall on test data is 0.7902908805031447
- The f1 on test data is 0.7358404245585141
- The roc_score on test data is 0.7571411939670714

Logistic feature Importances



SVC Modelling

Parameters :- C = 10 , Kernel = ‘rbf’(Radial Basis Function)

Results:-

- The accuracy on test data is 0.7528676596473207
- The precision on test data is 0.6884095189179935
- The recall on test data is 0.7902908805031447
- The f1 on test data is 0.7358404245585141
- The roc_score on test data is 0.7571411939670714

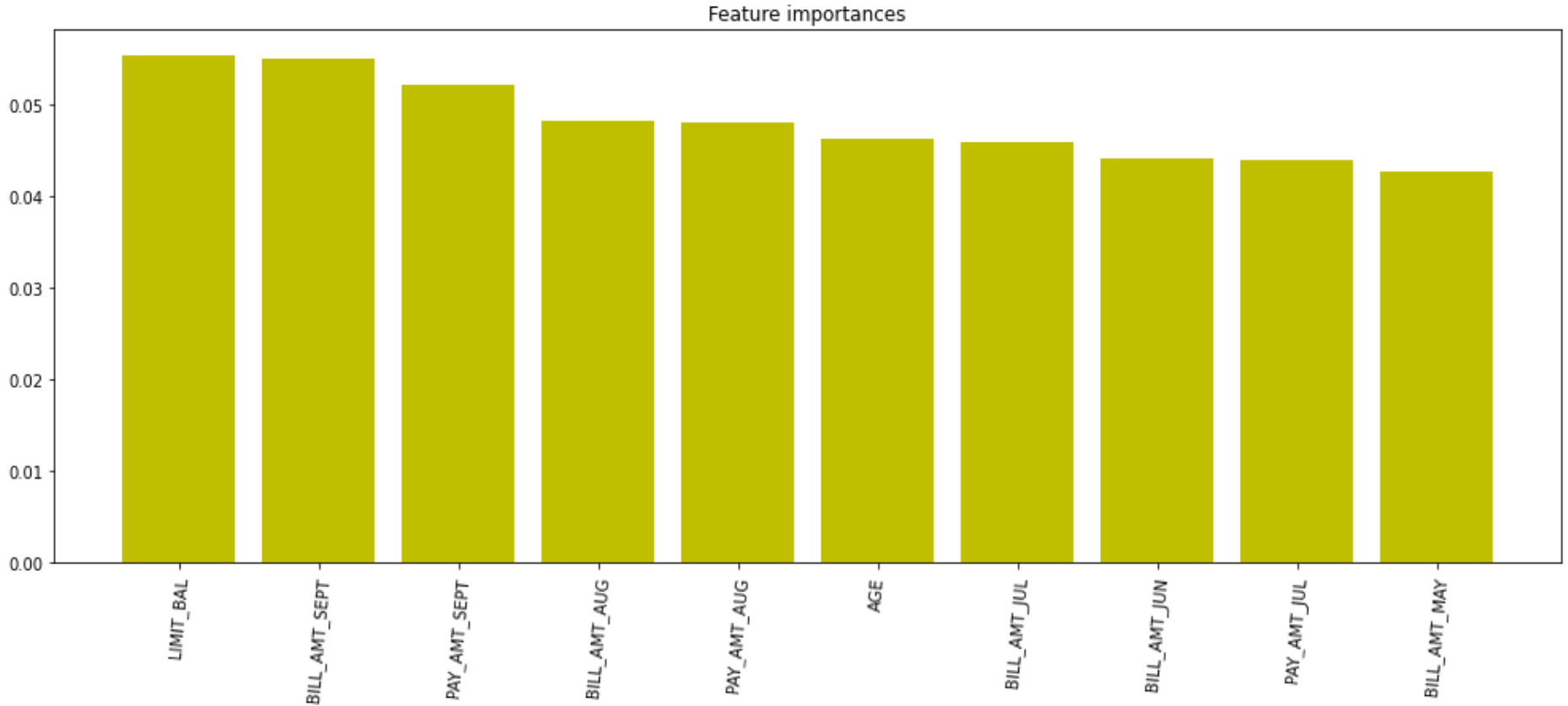
Random Forest

Parameters :-max_depth= 30 , n_estimators= 150

Results :-

- The accuracy on test data is 0.8421503167265879
- The precision on test data is 0.8121896935456258
- The recall on test data is 0.863959205973411
- The f1 on test data is 0.8372749735262972
- The roc_score on test data is 0.8433832534470512

Random Forest feature Importances



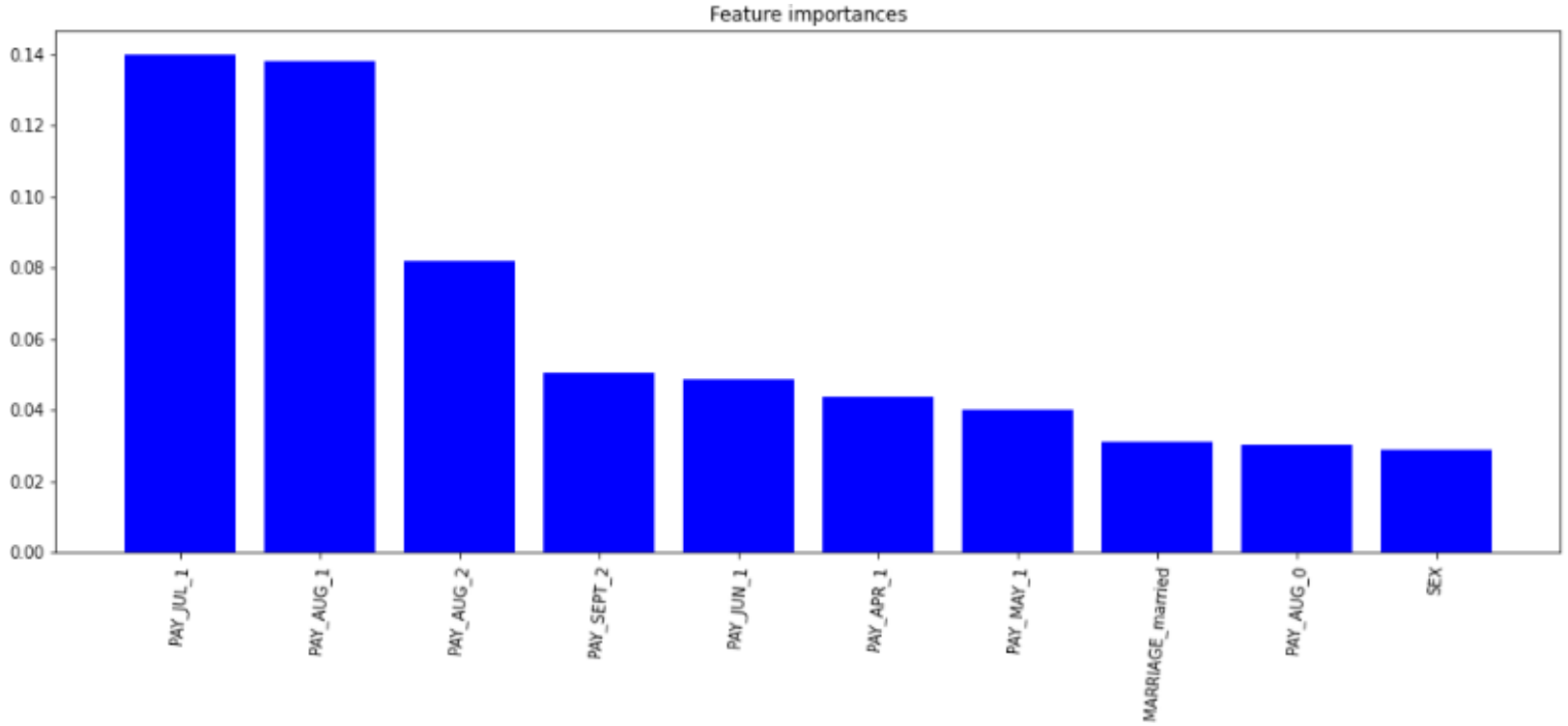
XGBoost Modelling

Parameters : max_depth= 15 , min_child_weight= 8

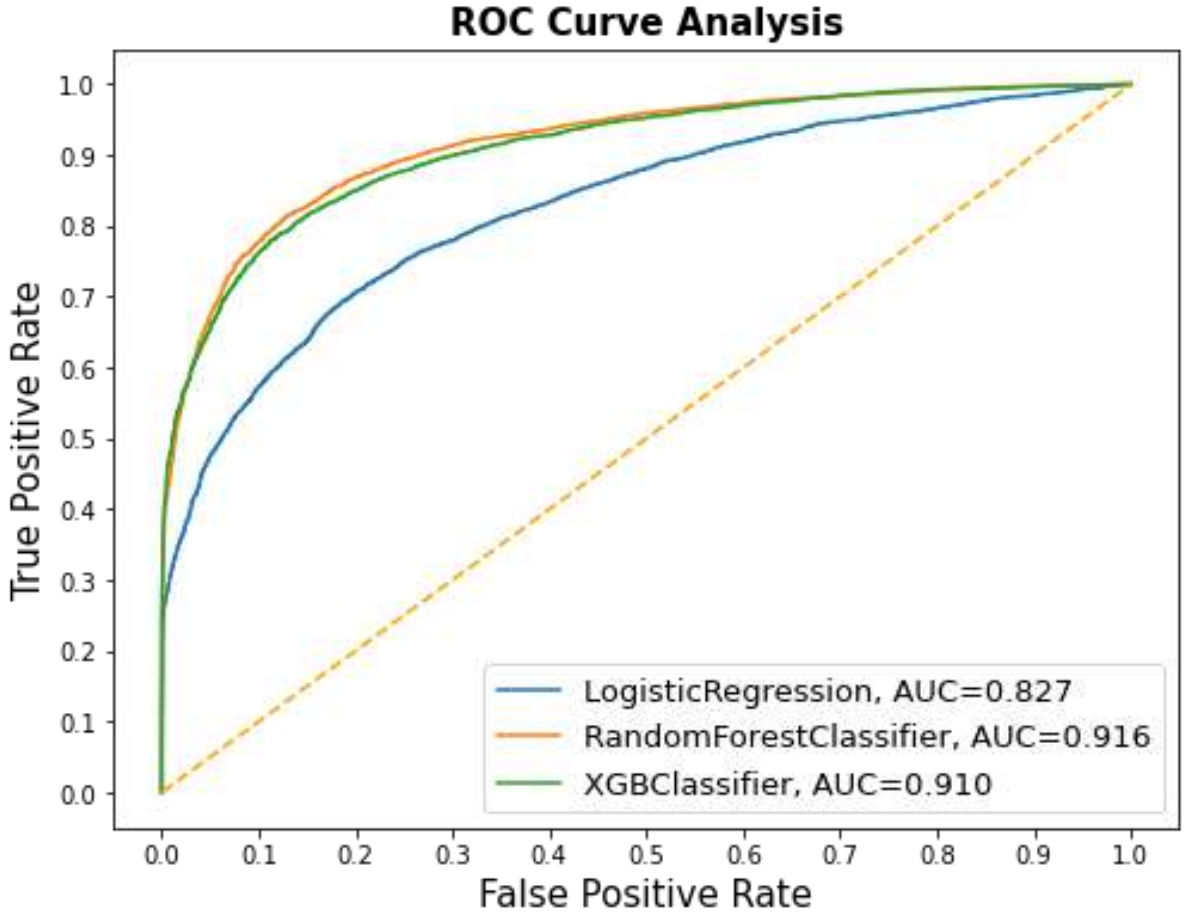
Results :-

- The accuracy on test data is 0.8329909262112652
- The precision on test data is 0.7880499914398219
- The recall on test data is 0.8658765989465764
- The f1 on test data is 0.8251322039974904
- The roc_score on train data is 0.8357029868751105

XGBoosting Feature Importances



AUC-ROC curve Comparision



CHALLENGES FACED

- The data was huge and was to be handled keeping in mind that we do not miss anything which is even of a little relevance.
- Computation time.
- Getting a higher accuracy on the models.
- Carefully handling feature imbalanced data.
- Tuning of hyperparameters carefully.

EVALUATING THE MODELS

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.754009	0.752868	0.68841	0.790291	0.735840
1	SVC	0.754009	0.752868	0.68841	0.790291	0.735840
2	Random Forest CLf	0.998431	0.842150	0.81219	0.863959	0.837275
3	Xgboost Clf	0.909062	0.832991	0.78805	0.865877	0.825132

From the above table we can find that XGBoost classifier perform best among those models

Conclusion

- XGBoost model has the highest recall, if the business cares recall the most, then this model is the best candidate.
- If the precision Score is the most important metric, then Random Forest is the ideal model.
- Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest.
- There were not huge gap but female clients tended to default the most.
- Labels of the data were imbalanced and had a significant difference.
- The best accuracy is obtained for the Random forest and XGBoost classifier.
- XGBoost Provided Us the Best Results by Giving us Recall of 86% (Means Out of 100 Defaulters 86 Will be Correctly Caught by XGBoost.)

Thank You