Capstone Project
Credit Card Default Prediction

Individual Project
Uthaman A

Content

- Introduction
- Problem Statement
- Data Summary
- Approach Overview
- Exploratory Data Analysis
- Modelling Overview
- Feature Importance
- Challenges
- Conclusion

Introduction

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'.

Problem Statement

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

Data Summary

- X1 – Amount of credit(includes individual as well as family credit)
- X2 – Genser
- 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 Cleaning

Data Exploration

Modeling

Understanding and Cleaning

- Find information on documented columns values
- Clean data to get it ready for Analysis

Graphical

- Examining the data with visualization

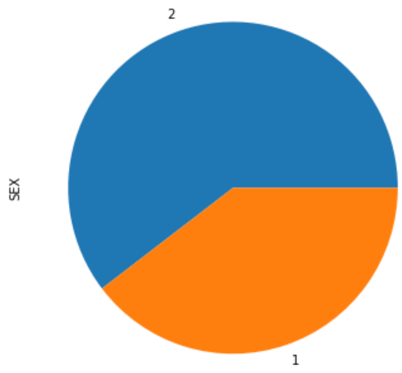
Machine Learning

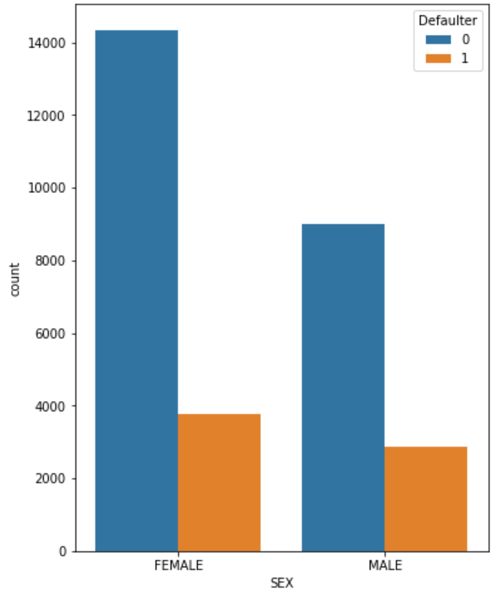
- Logistic
- SVM
- Random Forest
- XGBoost

Basic Exploration

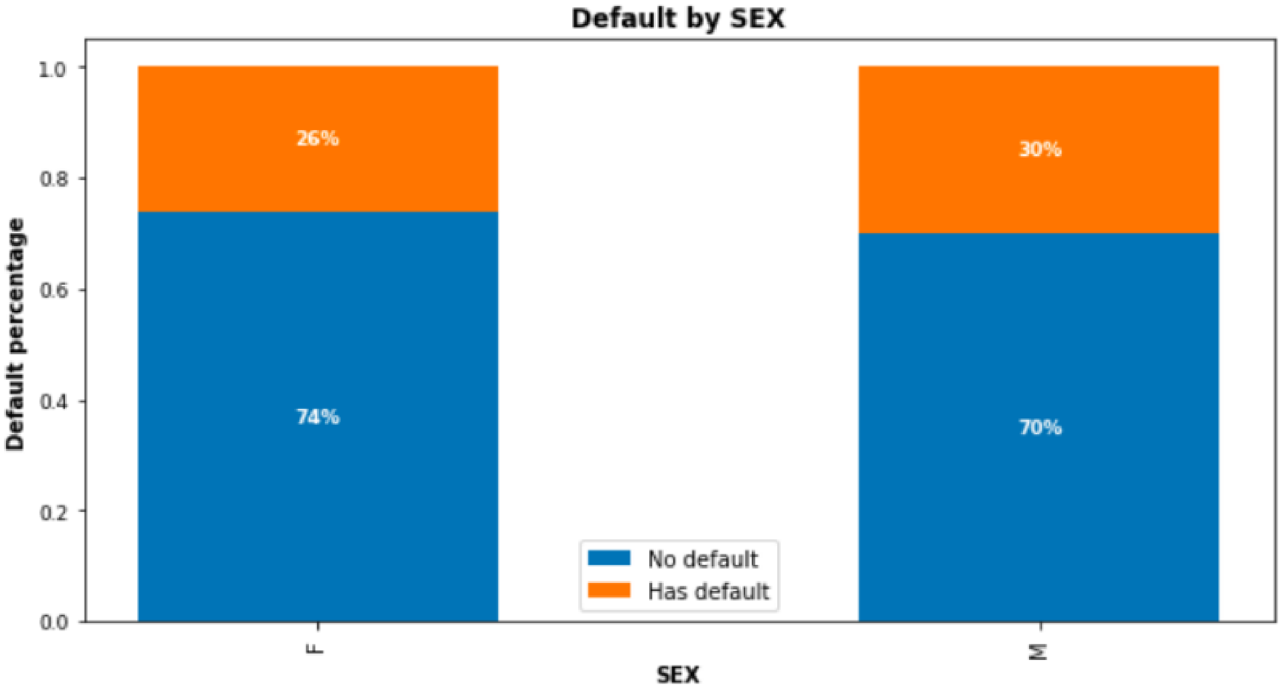
- Data for Taiwan.
- Data for 30000 customers.
- 6 Months payment and bill data available.
- No null data.
- 9 Categorical variables present

Gender Distribution



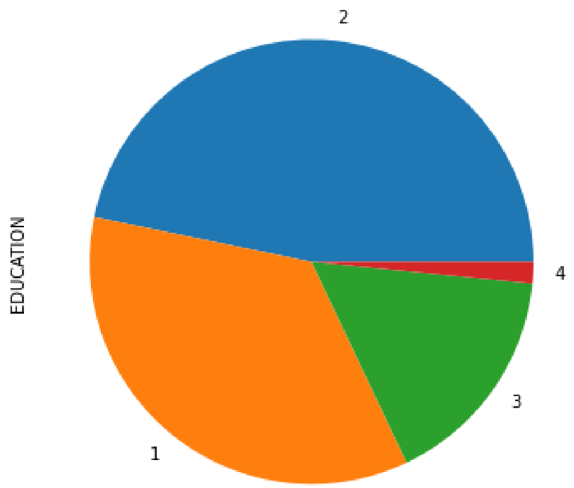


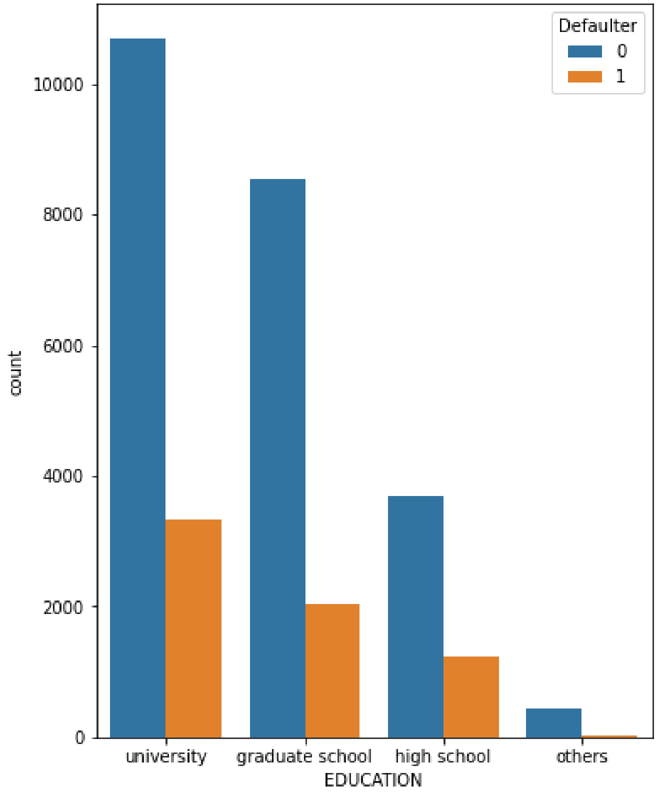
Gender wise defaulters



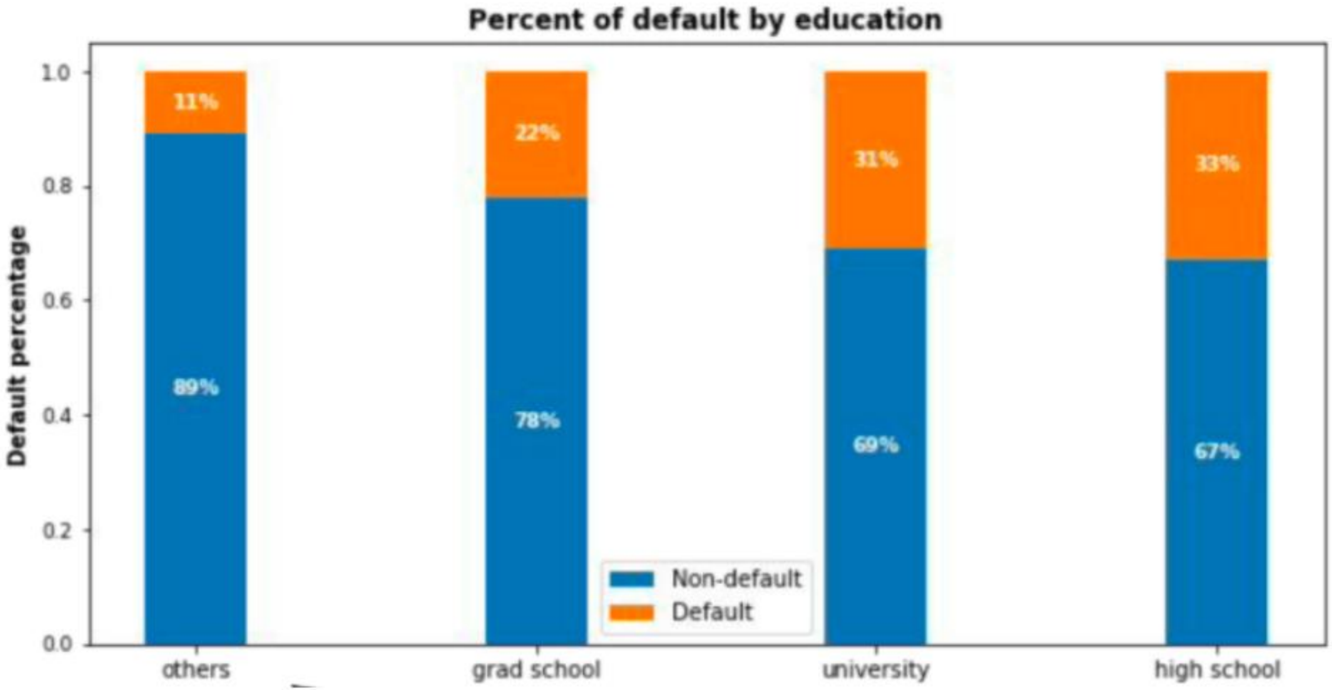
30% of Males and **26%** of Females are defaulters

Education Distribution



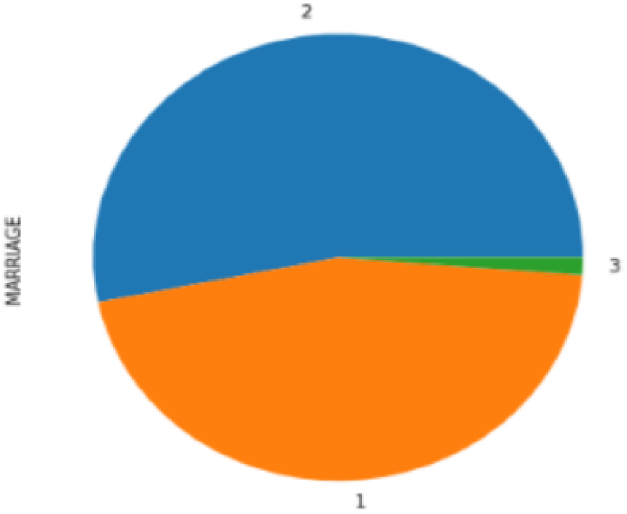


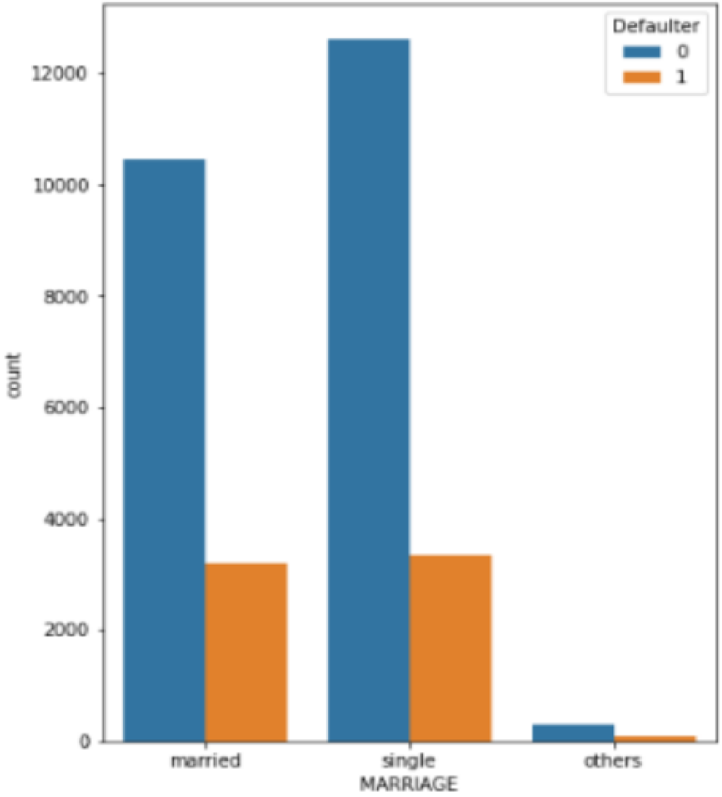
Education wise defaulters



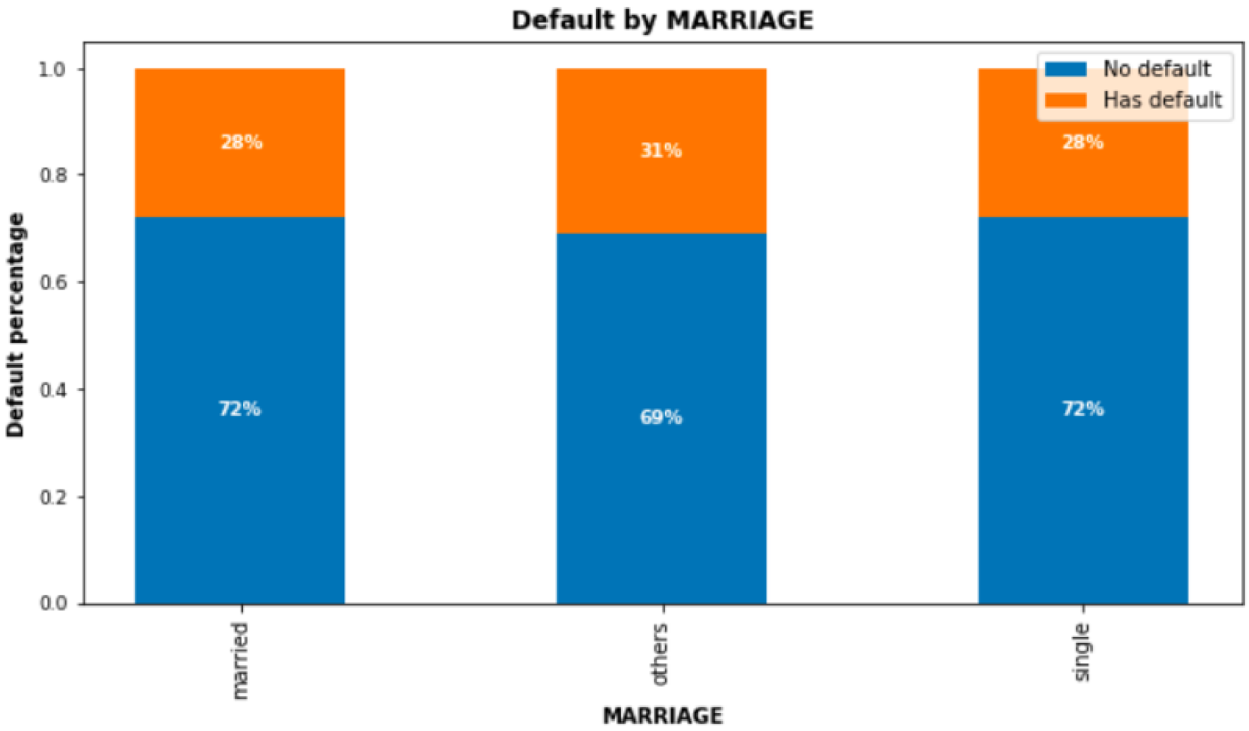
Higher Education Level, lower Default Risk

Marital Distributions



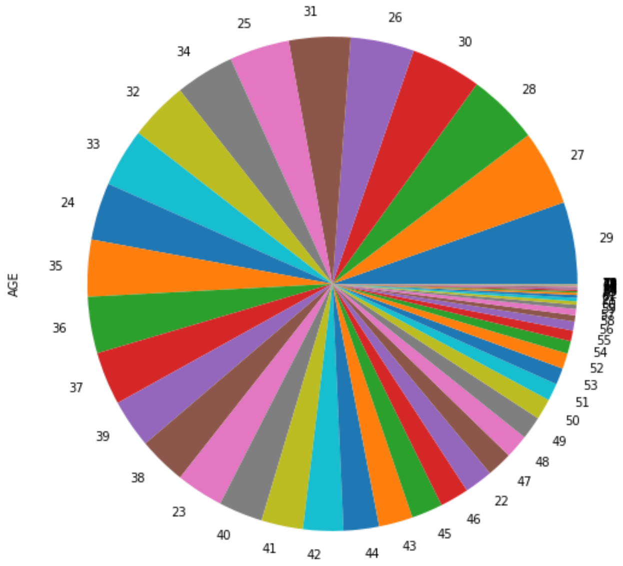


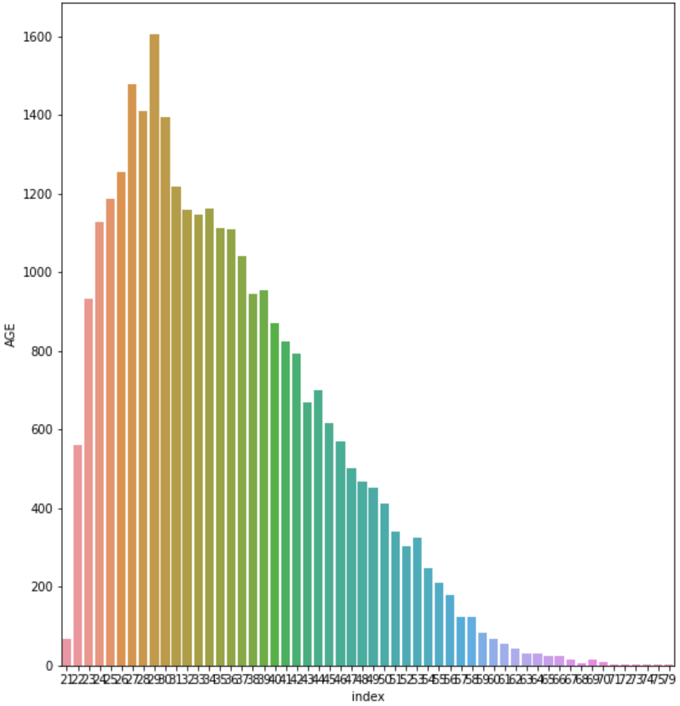
Marital Status



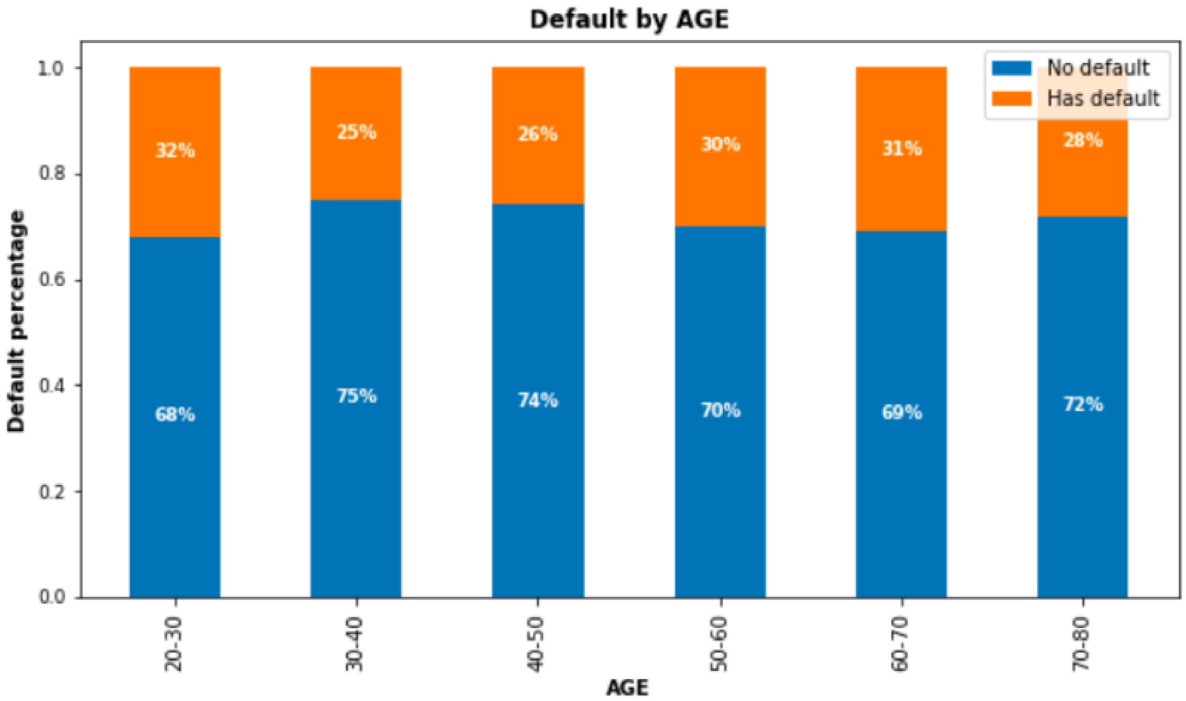
No
Significant
Correlation of
Default risk
And marital
status

Age Distribution





Age wise defaulters



30 to 50:
Lower Risk

<30 and >50
Risk Increases

Modelling Overview

- Supervised learning/Binary Classification
- Imbalance data with 78% non-defaulters and 22% defaulters

Models Used:

- Logistic Regression
- Knn
- Decision Trees
- Random Forest
- SVM
- XGBoost
- Naïve Bayes

Modelling Steps

Data Preprocessing

- Feature Selection
- Feature engineering
- Train test data split(80%-20)
- SMOTE oversampling

Data Fitting and Tuning

- Start with default model parameter
- Hyperparameter tuning
- Measure Ruc_AOC on training data

Model Evaluation

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

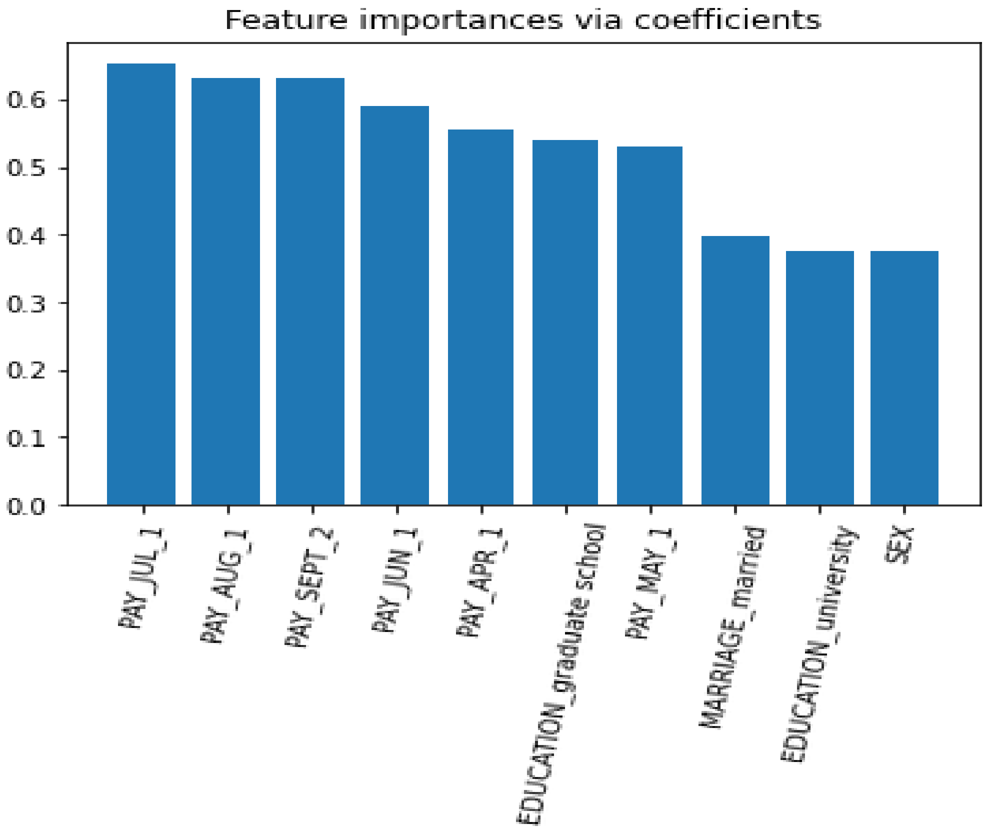
Logistic Modelling

Parameters :

- C = 0.01
- Penalty = L2

```
The accuracy on test data is  0.7498865183840218
The precision on test data is  0.6862516212710765
The recall on test data is  0.7862981126467529
The f1 on test data is  0.732876237966182
The roc_score on test data is  0.75399811292715
```

Logistic feature importance



SVM Modelling

Parameters :

- C = 10
- Kernel= 'rbf'

```
The accuracy on test data is  0.7786135788859347
The precision on test data is  0.7175097276264591
The recall on test data is  0.8173758865248227
The f1 on test data is  0.7641939494405305
The roc_score on test data is  0.7828356377036455
```

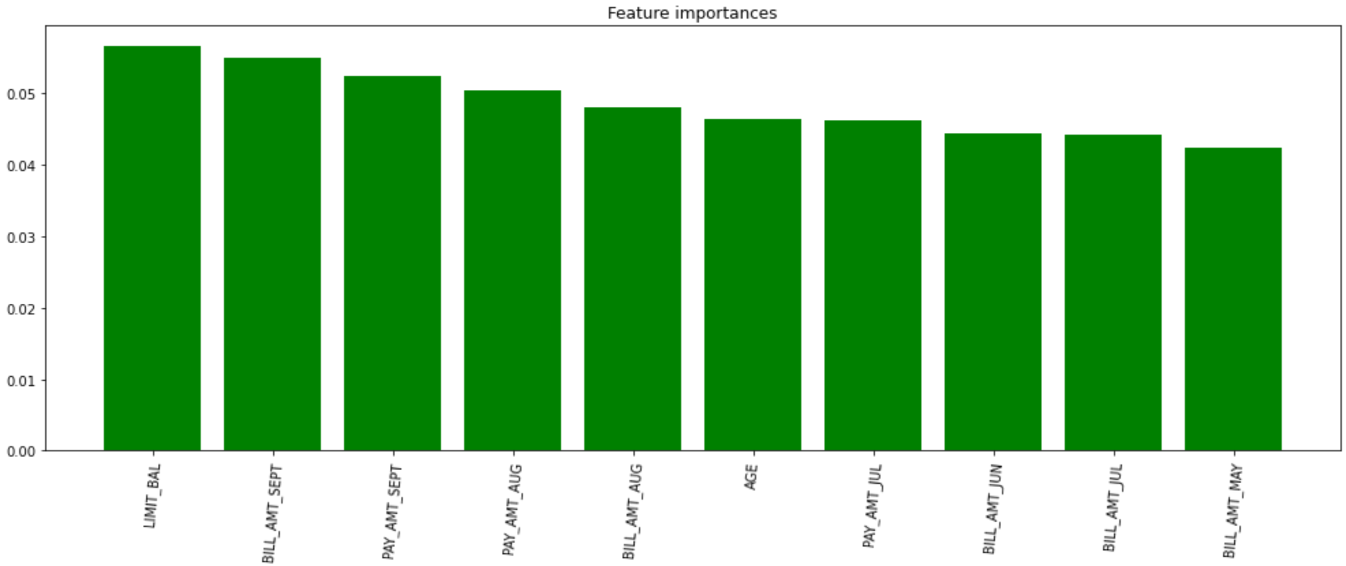
Random Forest Metrics

Parameters :

- **Max_depth=30**
- **N_estimators=150**

```
The accuracy on test data is  0.8349004604111276
The precision on test data is  0.804928664072633
The recall on test data is  0.8562362030905077
The f1 on test data is  0.8297900788875517
The roc_score on test data is  0.8361078238014633
```

Random Forest feature importance



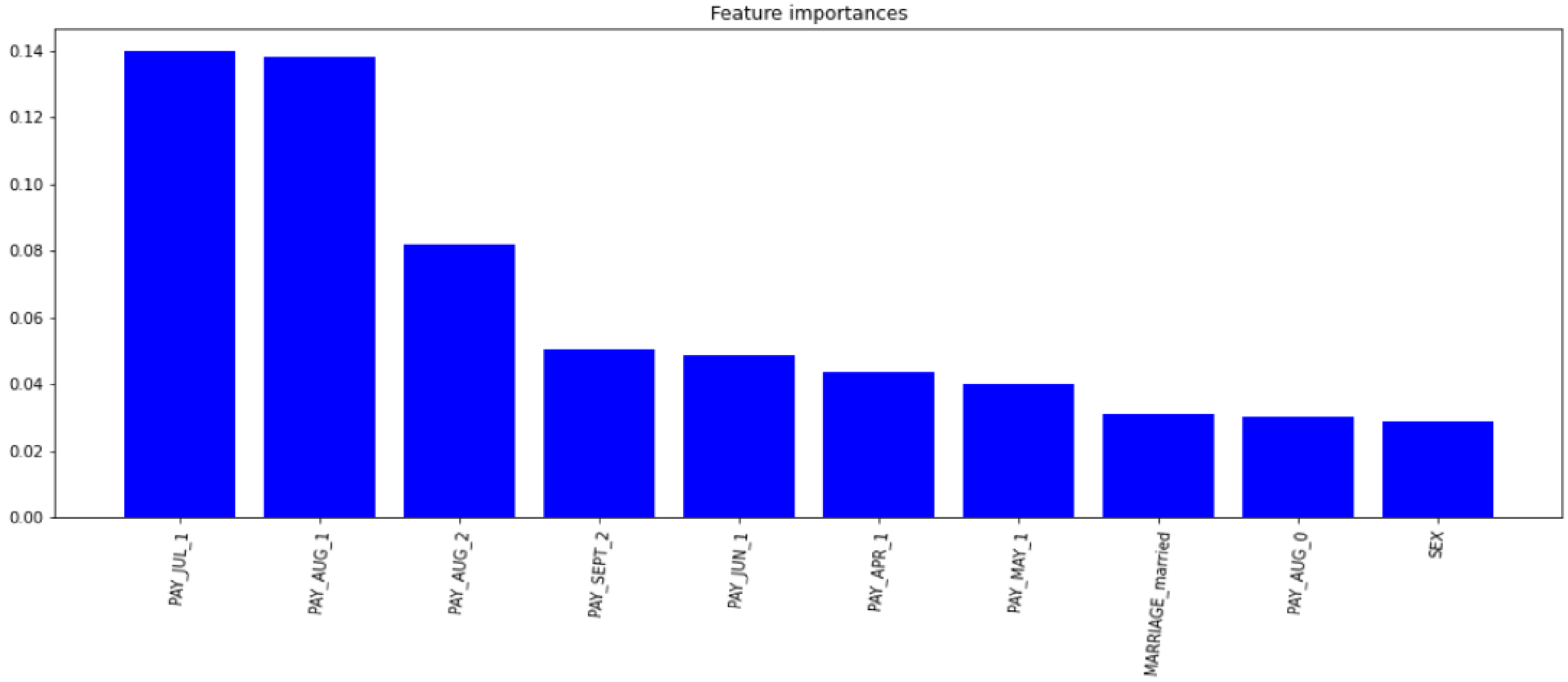
XGBoost Modelling

Parameters:

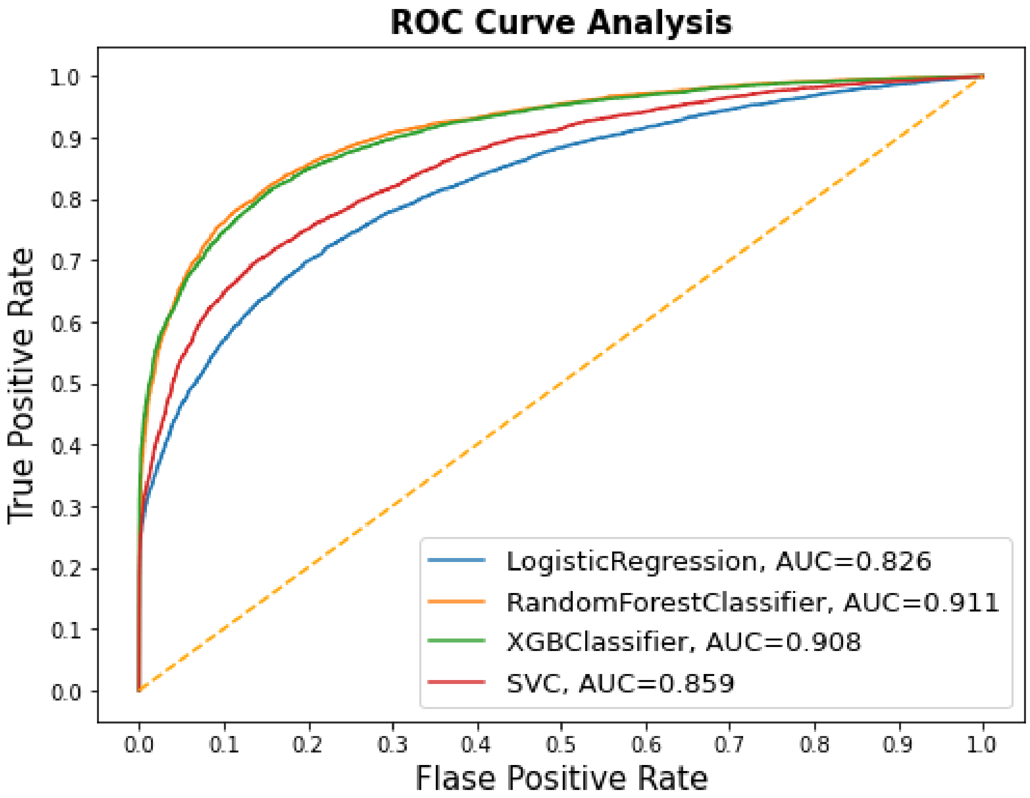
- **Max_depth=15**
- **Min_child_weight=8**

```
The accuracy on test data is  0.7727773814927696
The precision on test data is  0.694163424124513
The recall on test data is  0.823638042474607
The f1 on test data is  0.7533783783783784
The roc_score on train data is  0.779688571836878
```

X Gradient Boosting feature importance



AUC-ROC curve comparison



Challenges

- Understanding the columns.
- Feature engineering
- Getting a higher accuracy on the models

Conclusion

- XGBoost provided us the best results giving us a recall of 85 percent(meaning out of 100 defaulters 85 will be correctly caught by XGBoost)
- Random Forest also had good score as well but leads to overfit the data.
- Logistic regression being the least accurate with recall of 79.

Thank You