Capstone Project

Credit Card Default Prediction

Presented by
JAGAT PAL

Content

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 - 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 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

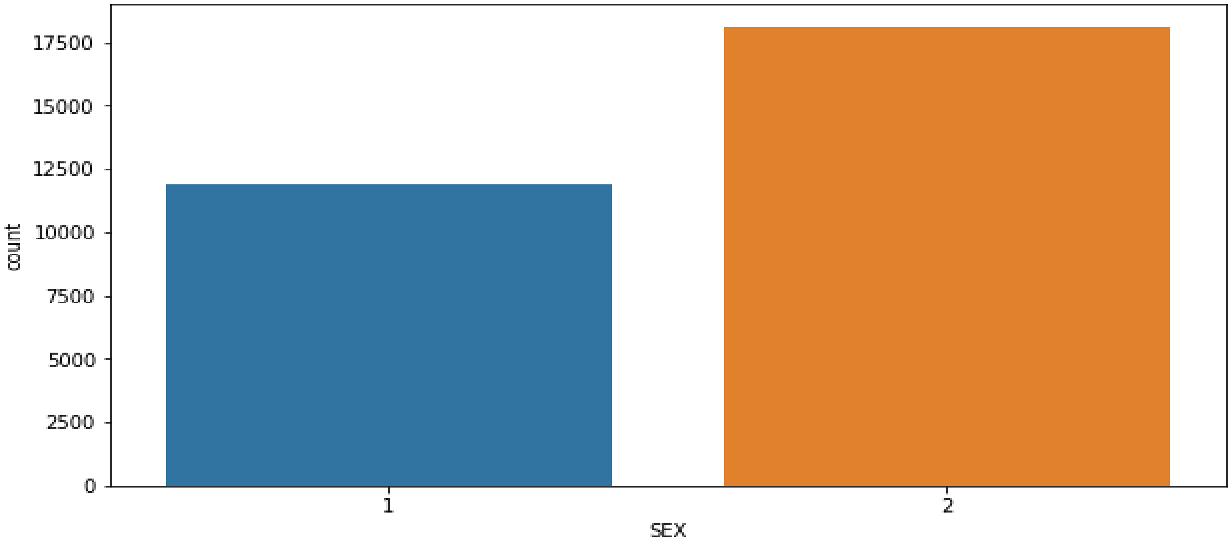
- Logistic
- SVM
- Random Forest
- XGBoost

Basic Exploration

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

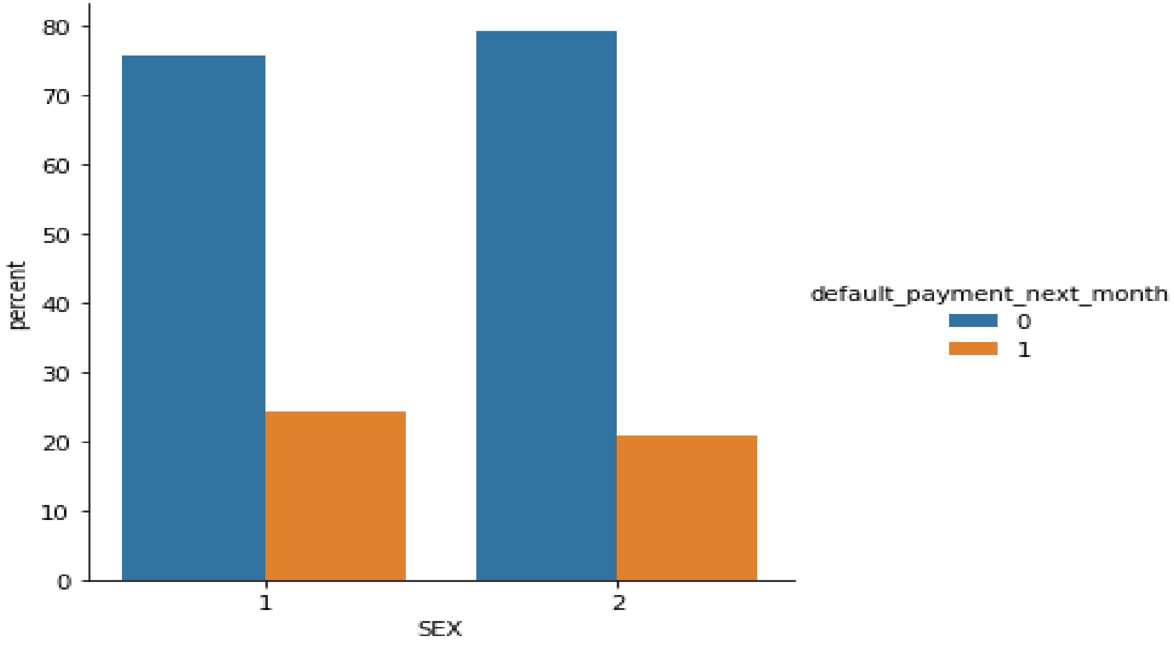
Gender Distribution

From the above data analysis we can say that
1 - Male
2 - Female
Number of Male credit holder is less than Female.



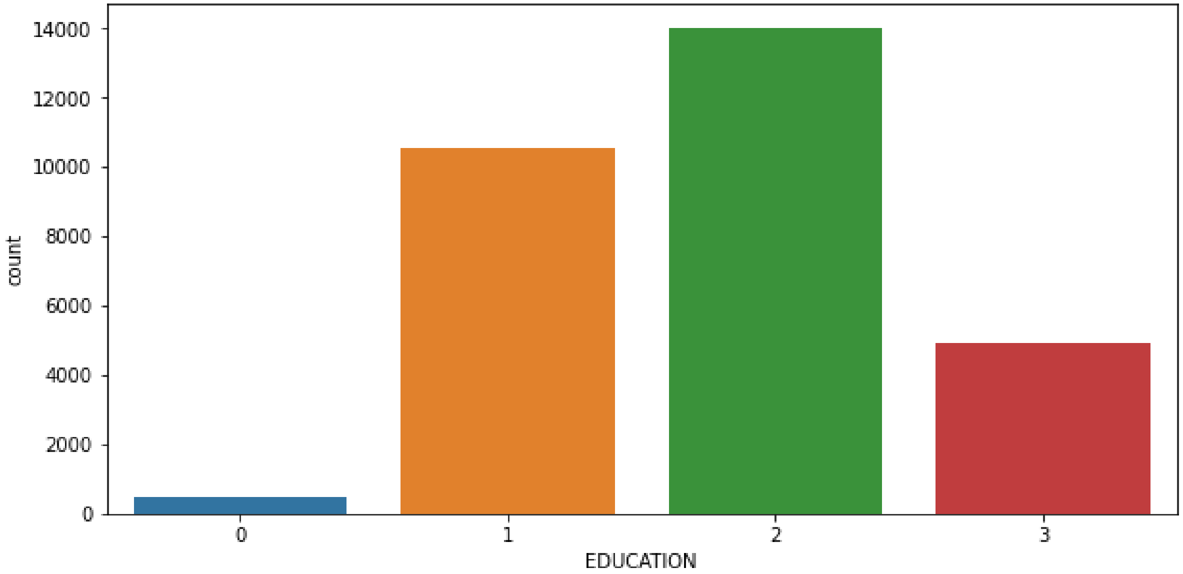
Gender wise defaulters

It is evident from the above graph that the number of defaulter have high proportion of males.



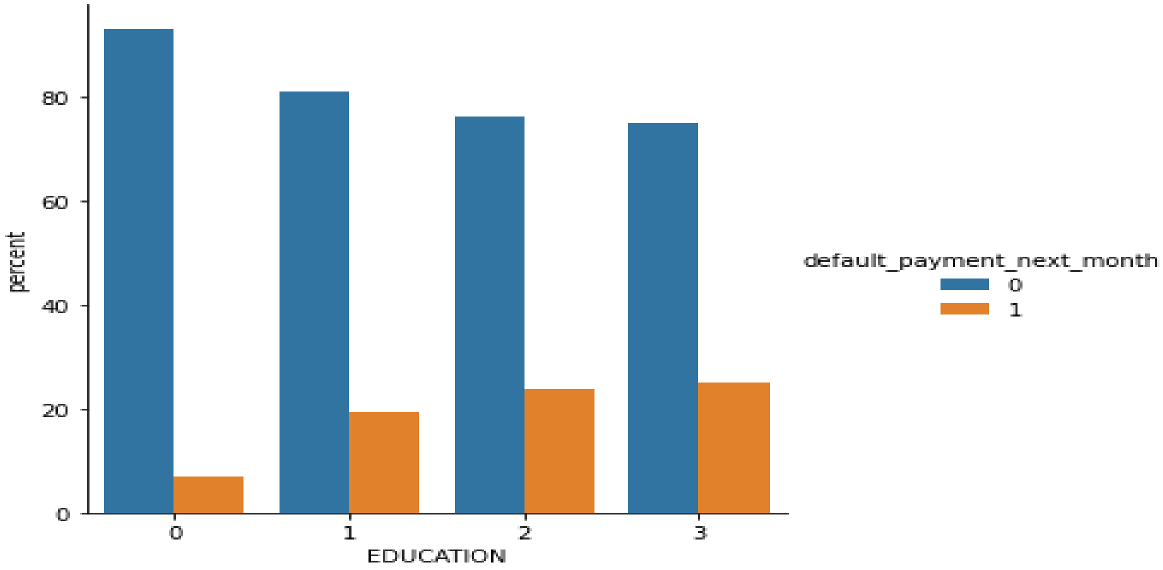
Education Distribution

1 = graduate school; 2 = university; 3 = high school; 0 = others
From the above data analysis we can say that, ***More number of credit holders are university students followed by Graduates and then High school students.***



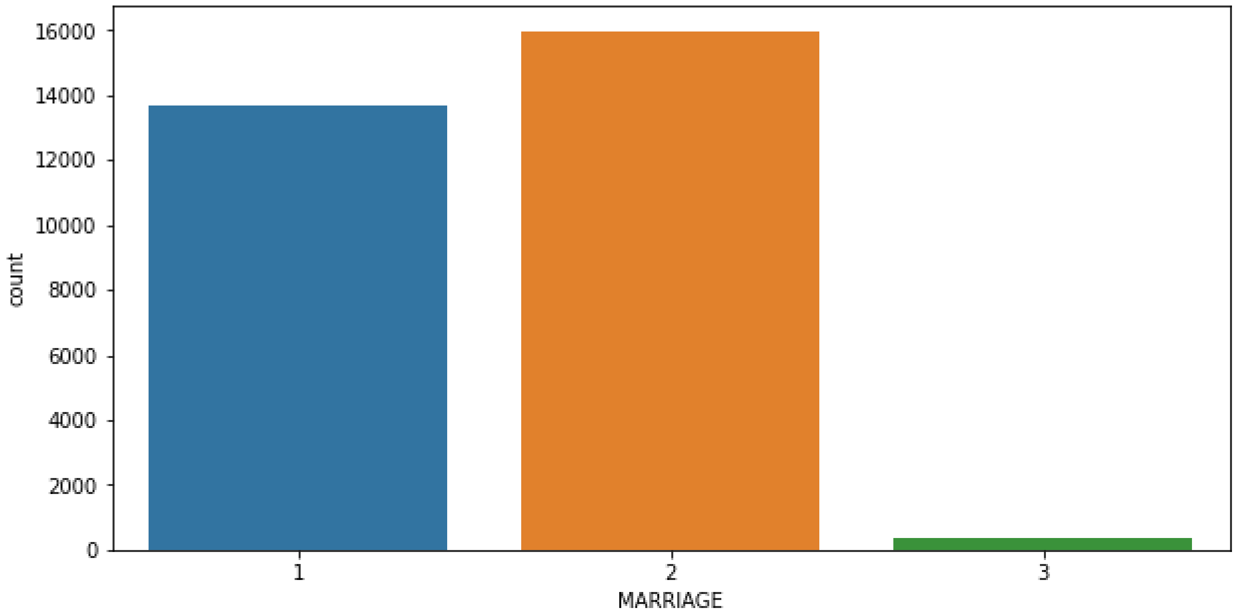
Education wise defaulters

From the above plot it is clear that those people who are other students have higher default payment wrt graduates and university people



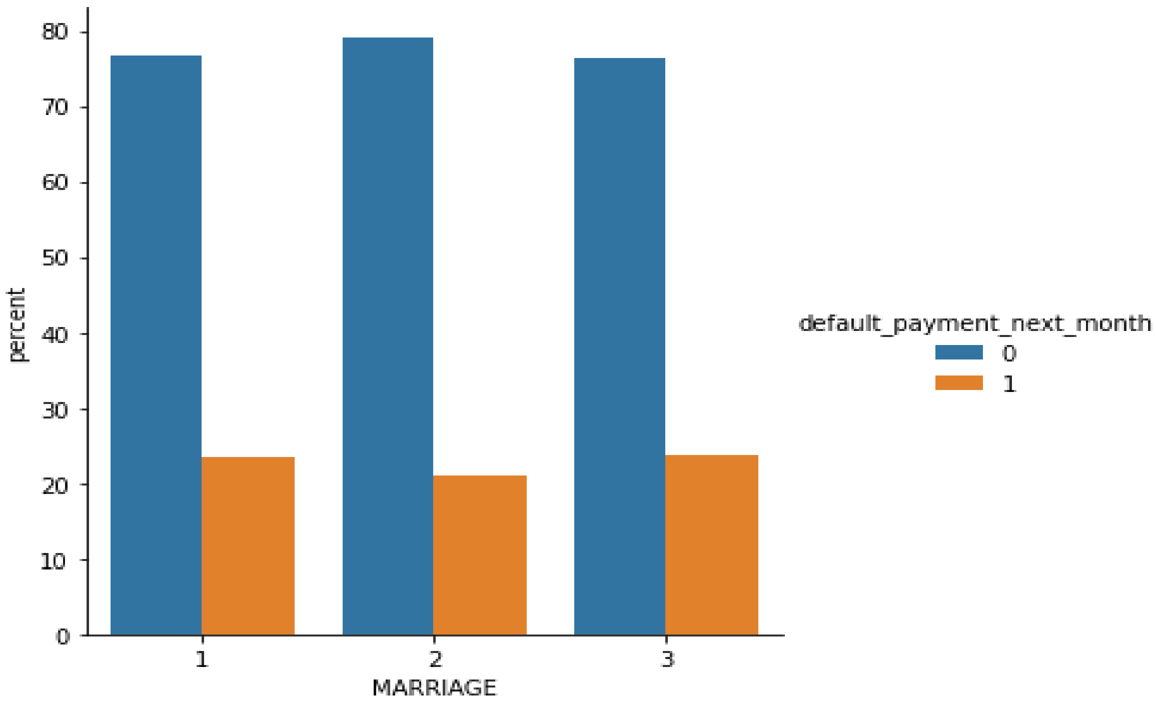
Marital Distributions

From the above data analysis we can say that
1 - married
2 - single
3 - others
***More number of credit cards holder are Single*.**

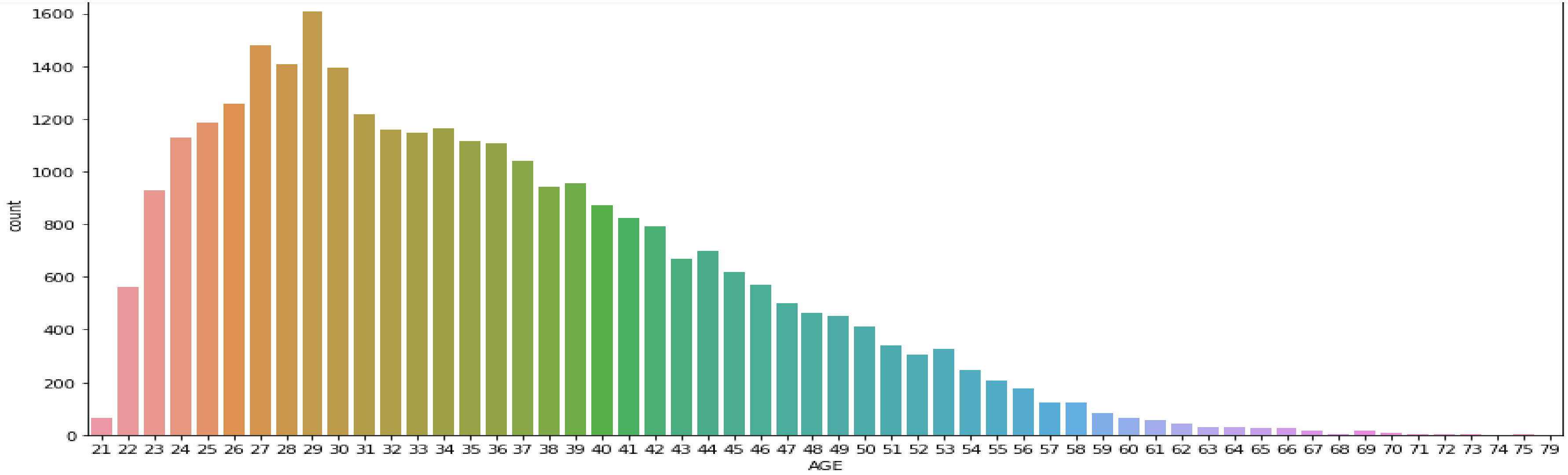


Marital Status

High defaulter rate when it comes to others



Age Distribution

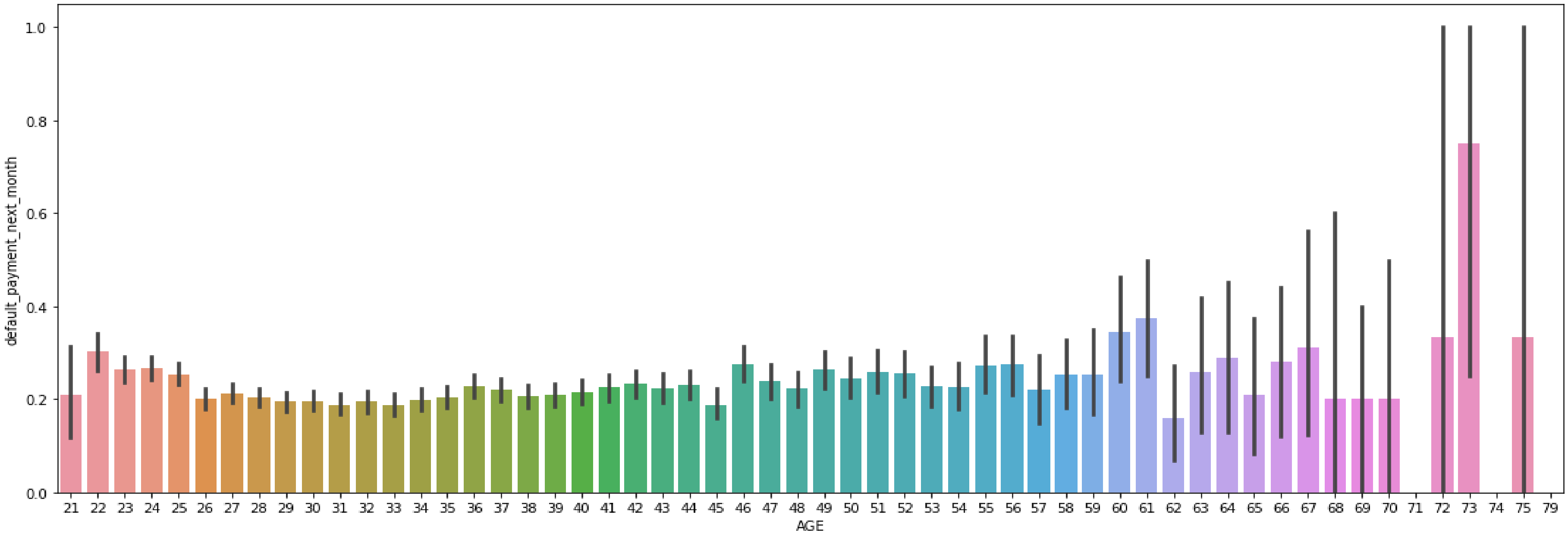


From the above data analysis we can say that

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.

Age wise defaulters



Slightly higher defaulter rate in 60's.

Modeling 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
- Naive Bayes

Modeling Steps

Data Preprocessing

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

Data Fitting and Tuning

- Start with default model parameters
- Hyperparameter tuning
- Measure RUC-AOC on training data

Model Evaluation

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

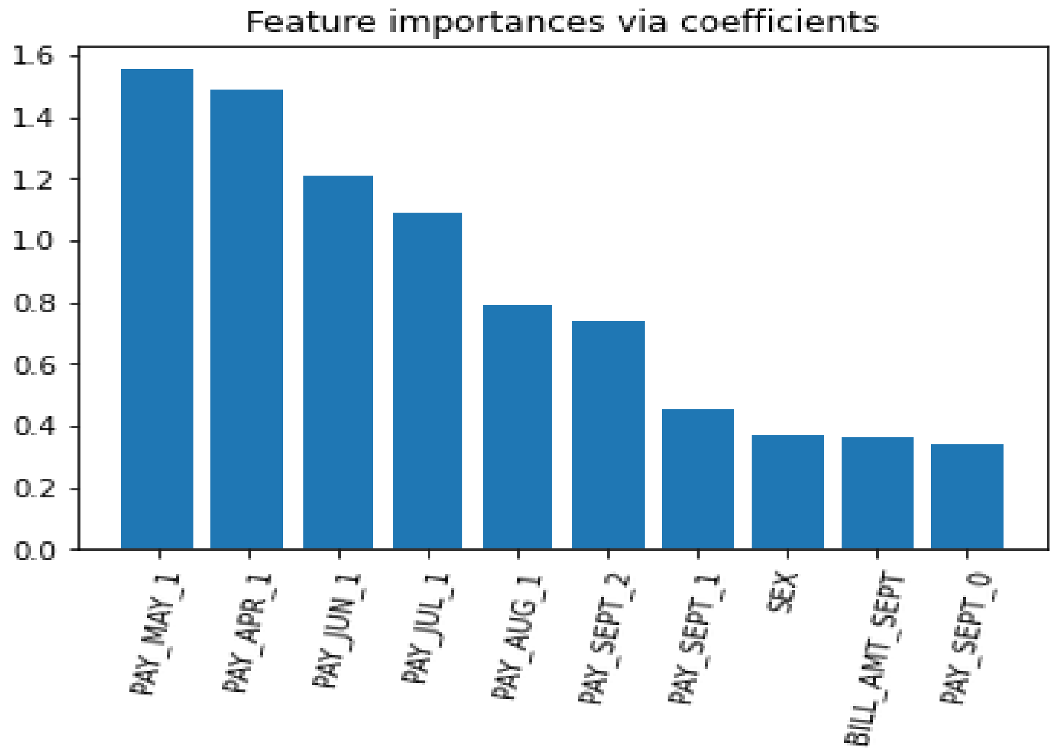
Logistic Modelling

Parameters :

- **C = 1000**
- **Penalty = L2**

```
The accuracy on test data is  0.7499513650217237
The precision on test data is  0.6862516212710765
The recall on test data is  0.7864149821640903
The f1 on test data is  0.7329269981991965
The roc_score on test data is  0.7540725549265177
```

Logistic feature importances



SVC Modelling

Parameters

C =10

Kernel =‘rbf’

```
The accuracy on test data is  0.7794565851760586
The precision on test data is  0.7167315175097276
The recall on test data is  0.8195165356666172
The f1 on test data is  0.7646855324154155
The roc_score on test data is  0.7839228218780194
```

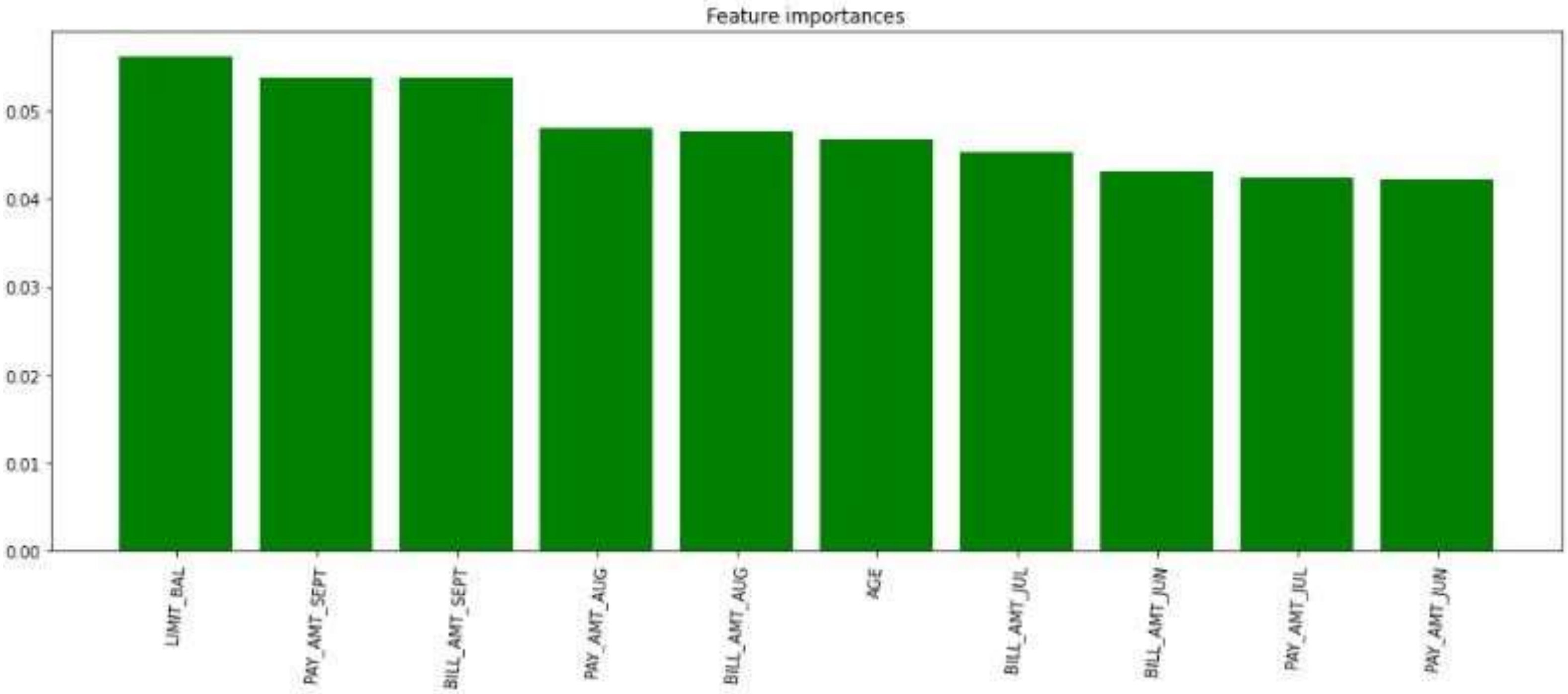
Random Forest Metrics

Parameters :

- max_depth=30
- n_estimators=100

```
The accuracy on test data is  0.8350301536865313
The precision on test data is  0.8058365758754864
The recall on test data is  0.8557851239669422
The f1 on test data is  0.8300601202404809
The roc_score on test data is  0.8361758605988369
```

Random Forest feature importances



XGBoost Modelling

Parameters :

- max_depth=7
- Learning rate =0.05

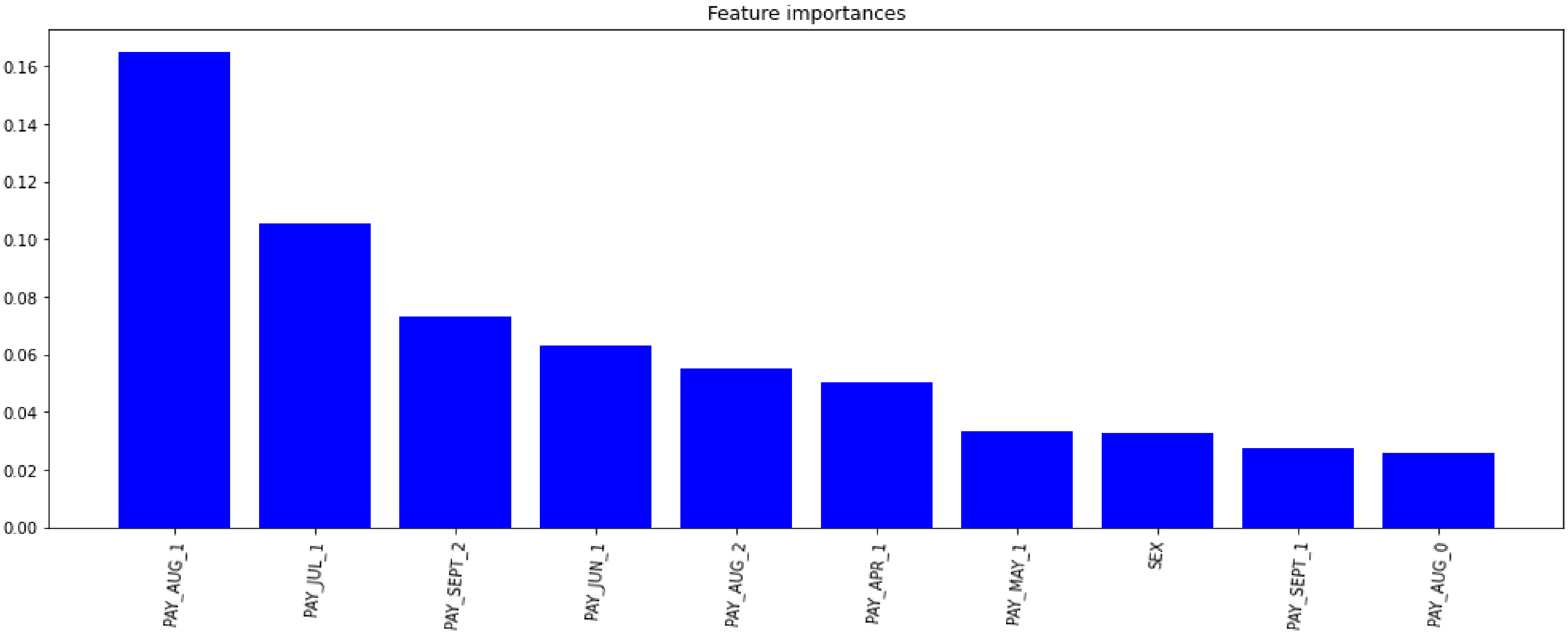
```
The accuracy on test data is  0.7764736398417742
The precision on test data is  0.7035019455252919
The recall on test data is  0.8236902050113896
The f1 on test data is  0.758867366211963
The roc_score on train data is  0.782487927313301
```

Hyperparameter Tuning

Max_depth = 10
Min_child_weight = 1

```
The accuracy on test data is  0.829193956293366
The precision on test data is  0.788197146562905
The recall on test data is  0.858575868889516
The f1 on test data is  0.821882607519610
The roc_score on train data is  0.8314214595556
```

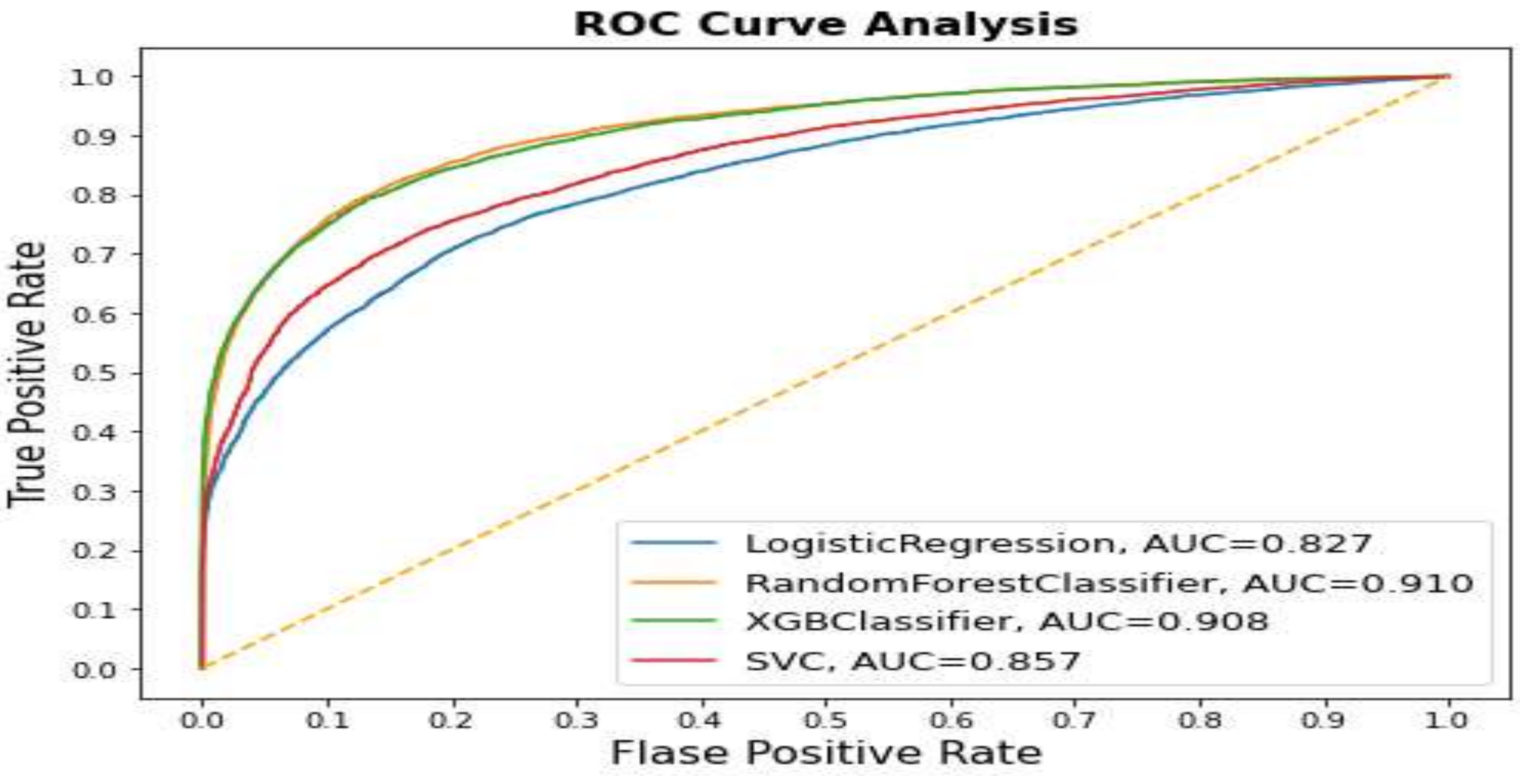
Hyperparameter Tuning feature importance



Evaluating the models

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.753601	0.752091	0.687808	0.789254	0.735047
1	SVC	0.810713	0.779457	0.716732	0.819517	0.764686
2	Random Forest CLf	0.998722	0.832112	0.800389	0.854591	0.826602
3	Xgboost Clf	0.912448	0.829194	0.788197	0.858576	0.821883

Plotting ROC AUC for all the models



Challenges

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

Conclusion

- Data categorical variables had minority classes which were added to their closest majority class
- There were not huge gap but female clients tended to default the most.
- Labels of the data were imbalanced and had a significant difference.
- Gradient boost gave the highest accuracy of 82% on test dataset.
- Repayment in the month of september tended to be the most important feature for our machine learning model.
- The best **accuracy** is obtained for the **Random forest** and **XGBoost classifier.**
- In general, all models have comparable accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading.
- From above table we can see that **XGBoost Classifier** having **Recall**, **F1-score**, and **ROC Score** values equals 82%, 77%, and 86% and **Random forest Classifier** having **Recall**, **F1-score**, and **ROC Score** values equals 81%, 75%, and 84%.

Thank You