Capstone project

Credit Card Default Prediction

BY
SAMEER
THETE

Content

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

Introduction

With growing technology in this 21st century, many of them 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**

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 and Understanding

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

Data Exploration (EDA)

- Examining the data with visualization
- Plotting graphs

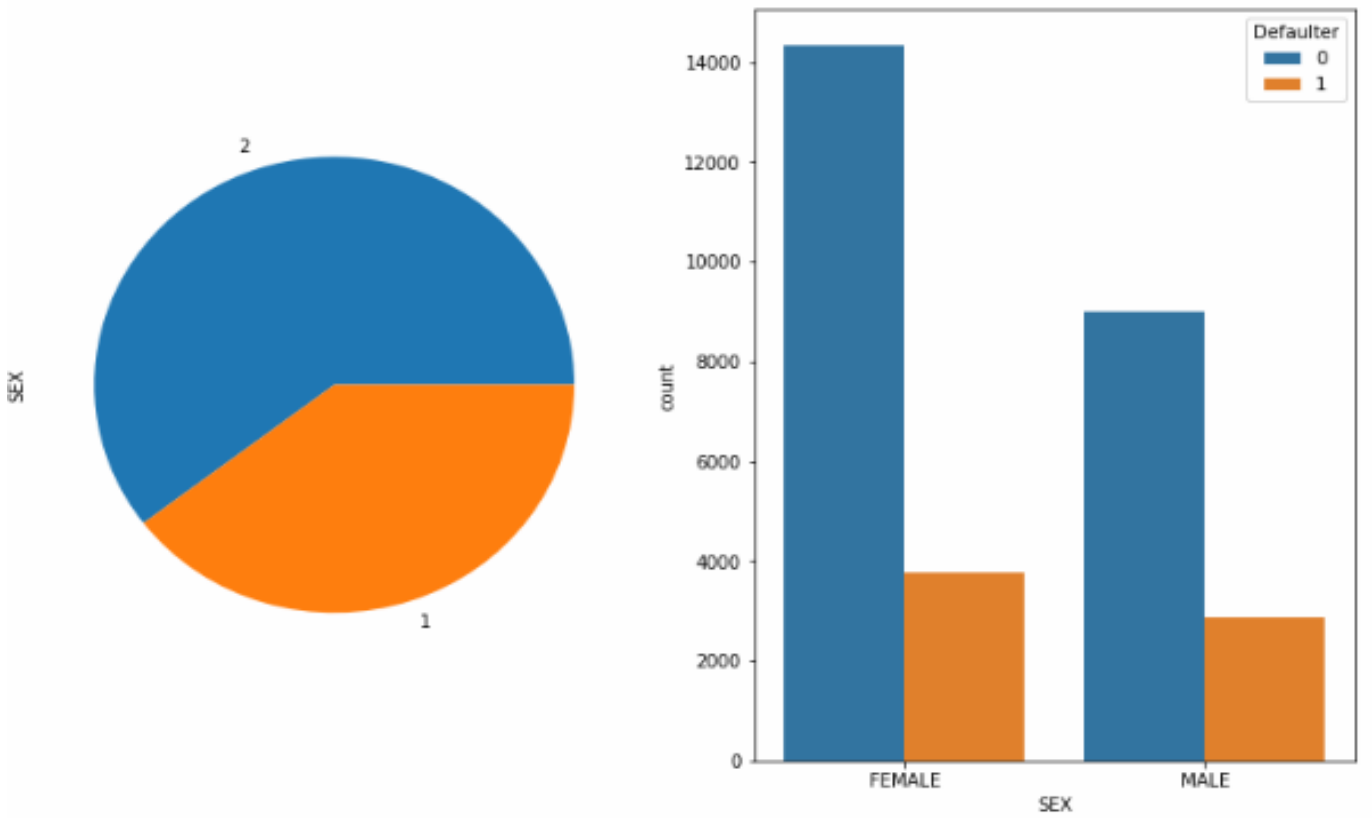
Modeling (Machine Learning)

- Logistic
- SVM
- Random Forest
- XGBoost

Basic Exploration

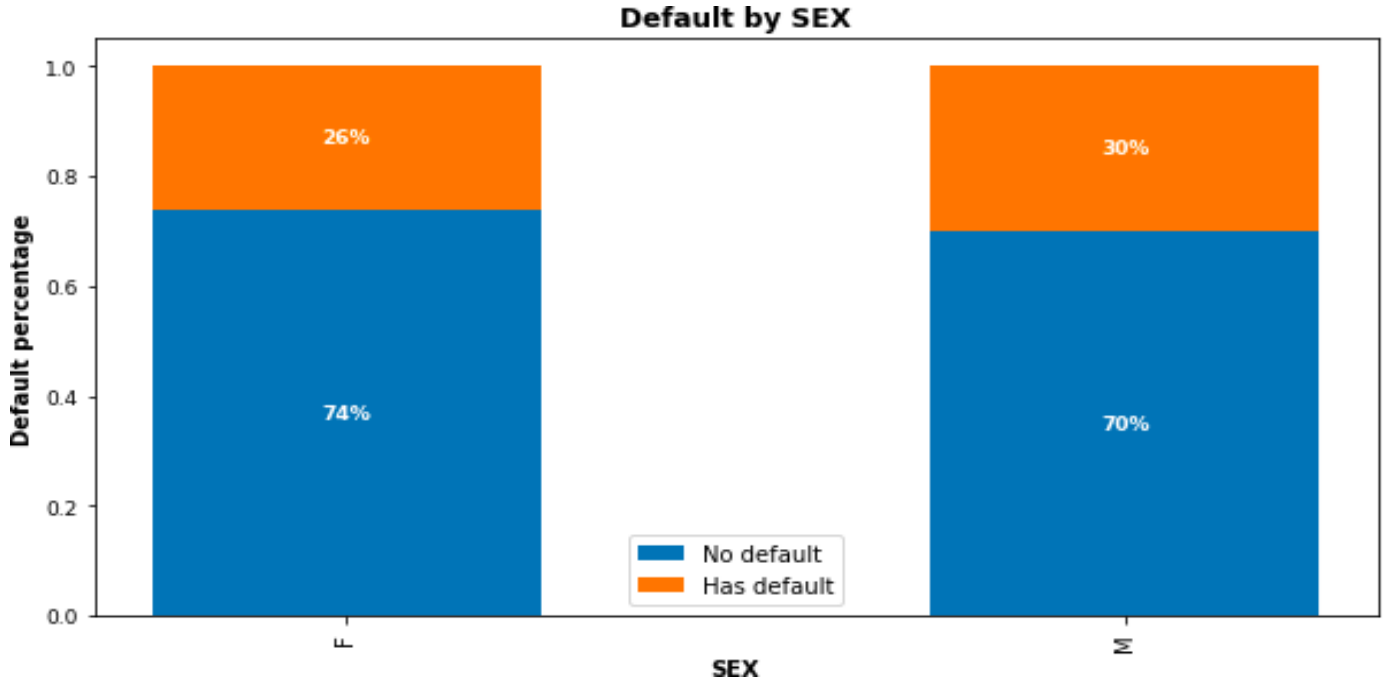
- Dataset for Taiwan.
- Shape of data is 30000 rows and 25 columns
- 6 months payment and bill data available.
- No null data.
- 9 Categorical variables present.
- ID column can be drop

Gender Distribution

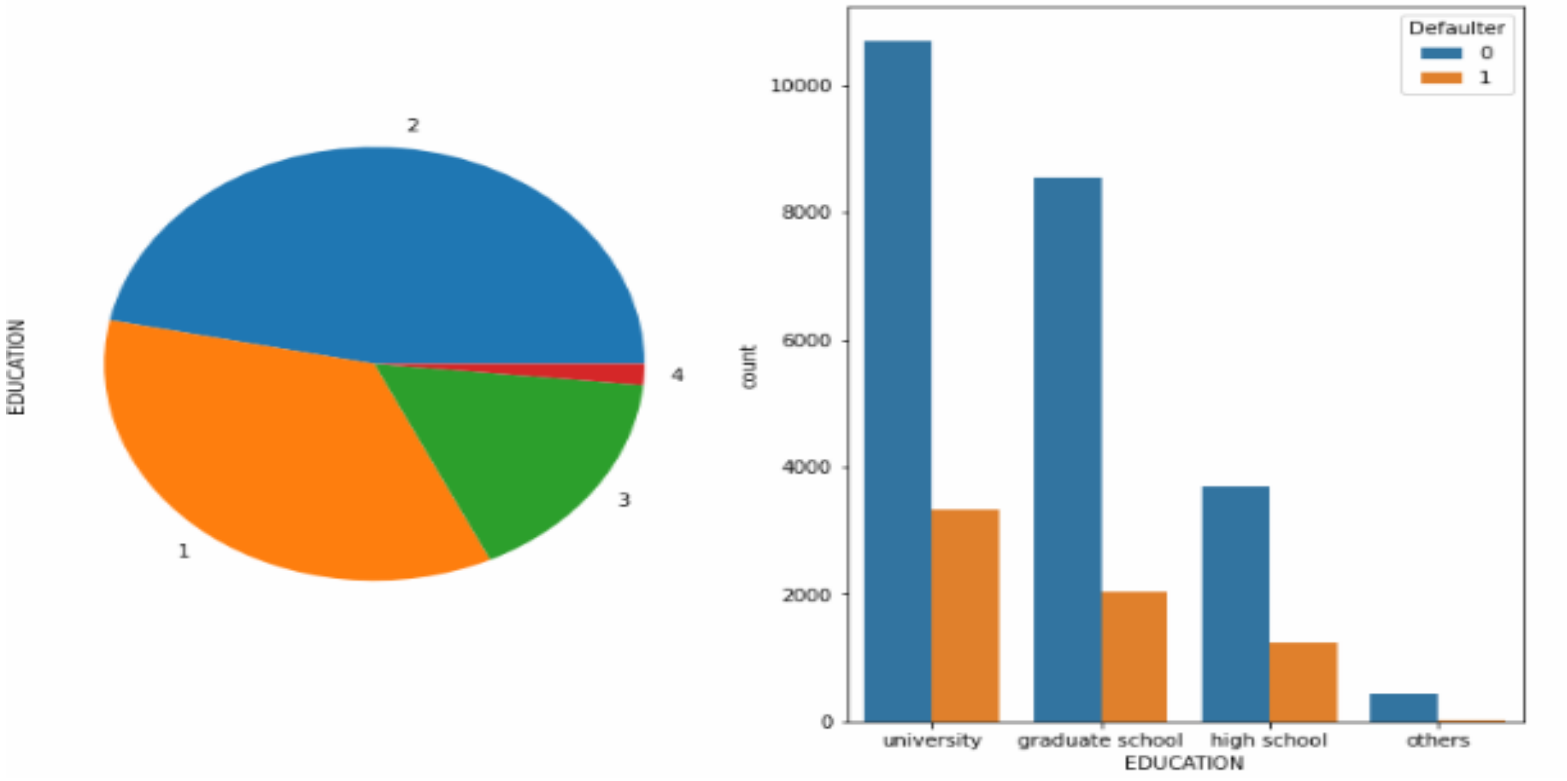


Gender wise defaulters

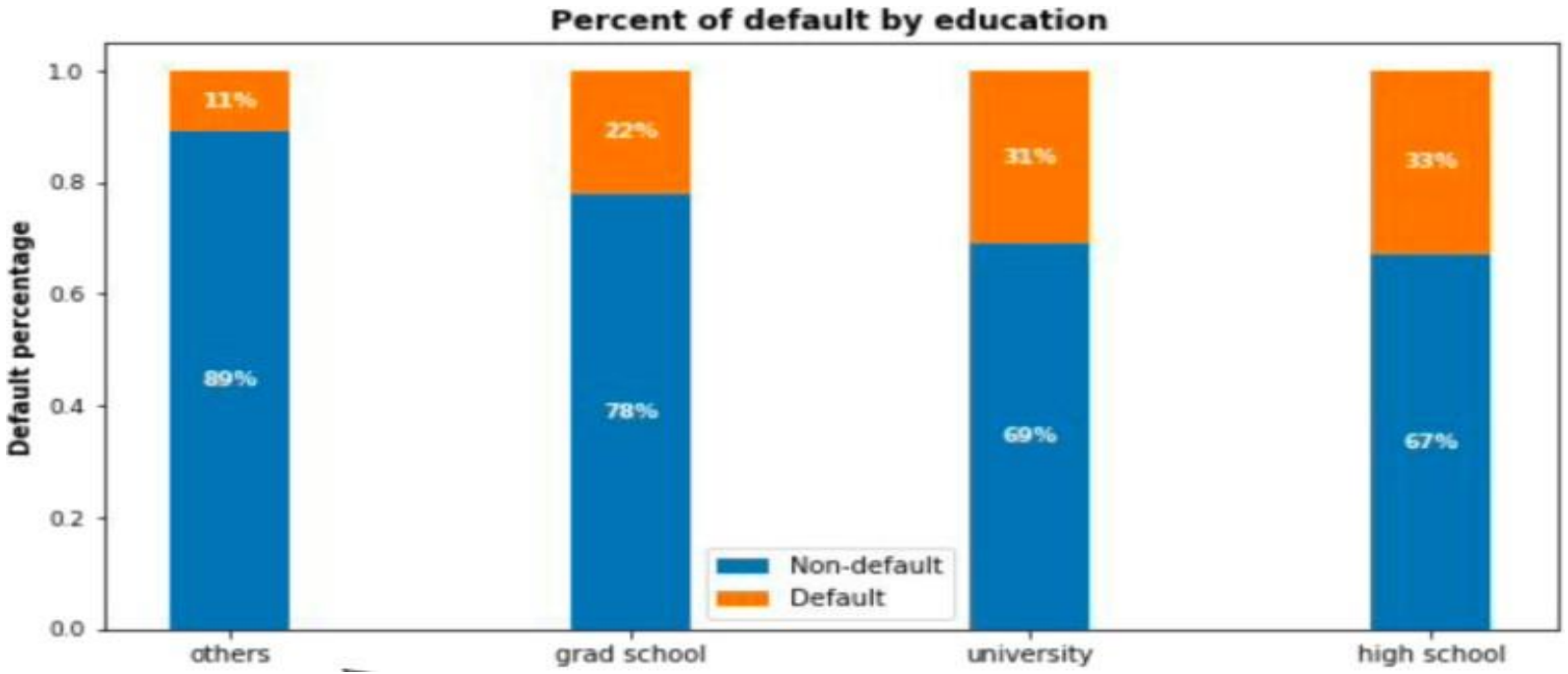
Observation- 30%of Males and 26%of Females are defaulters



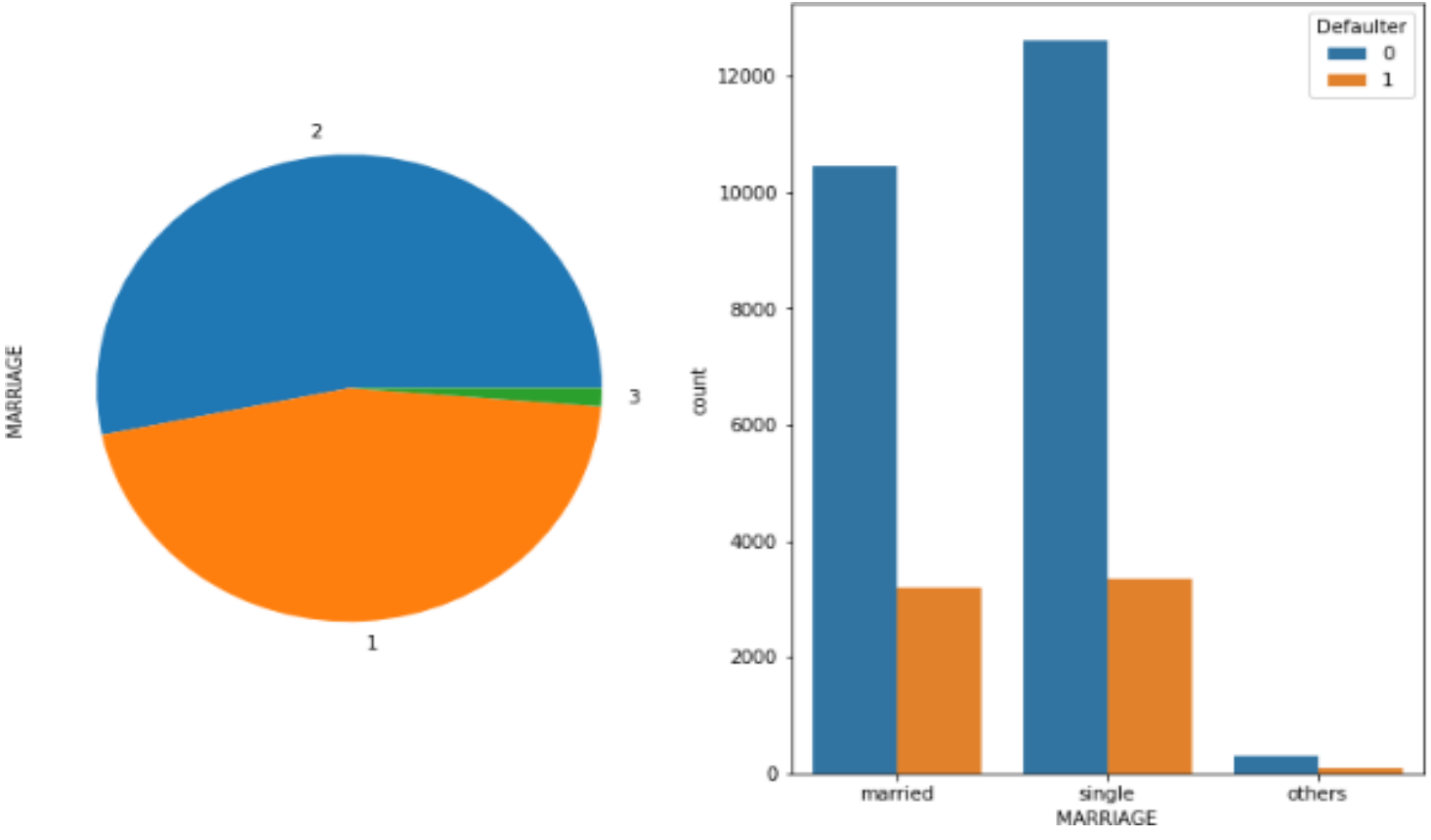
Education Distribution



Education wise defaulters

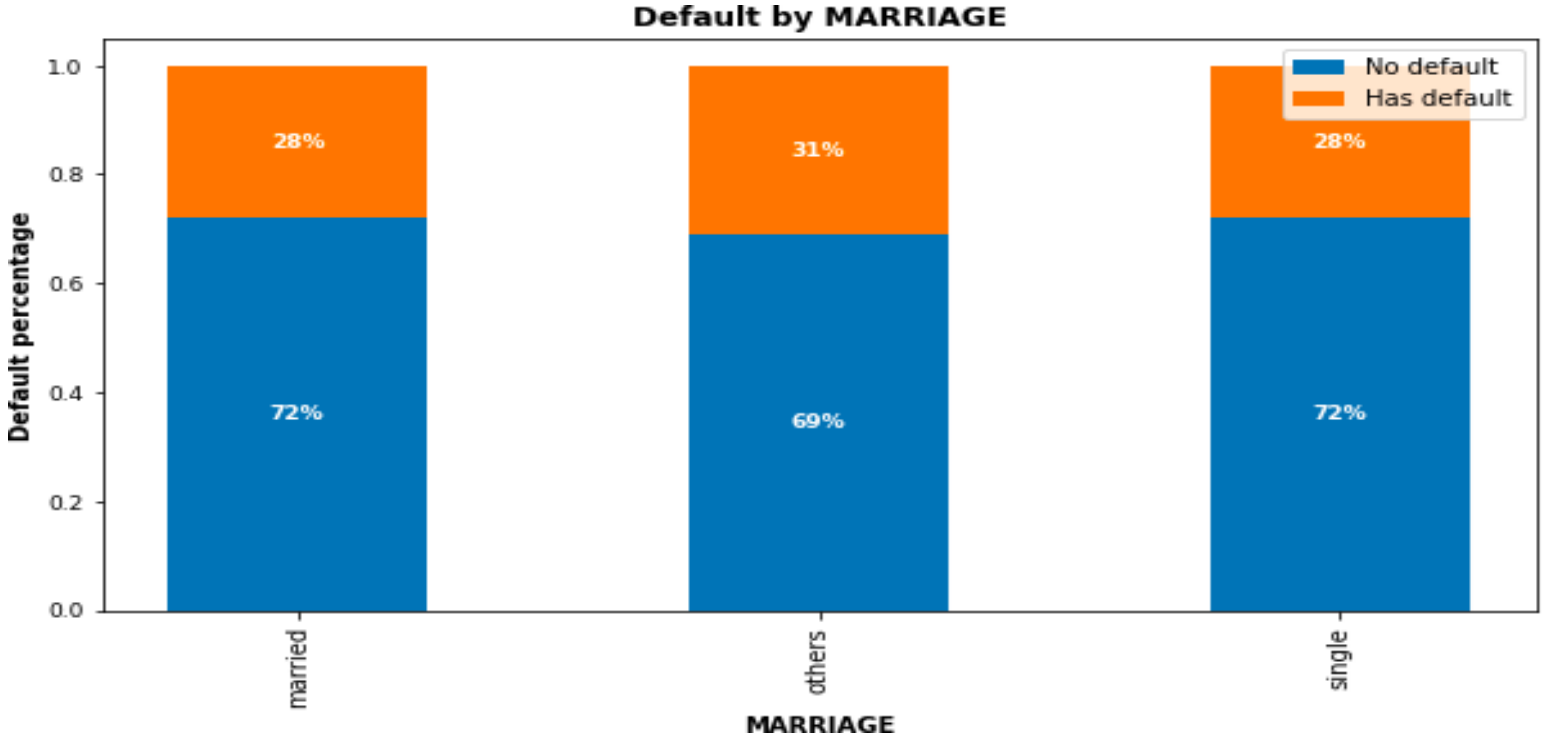


Marital Distributions

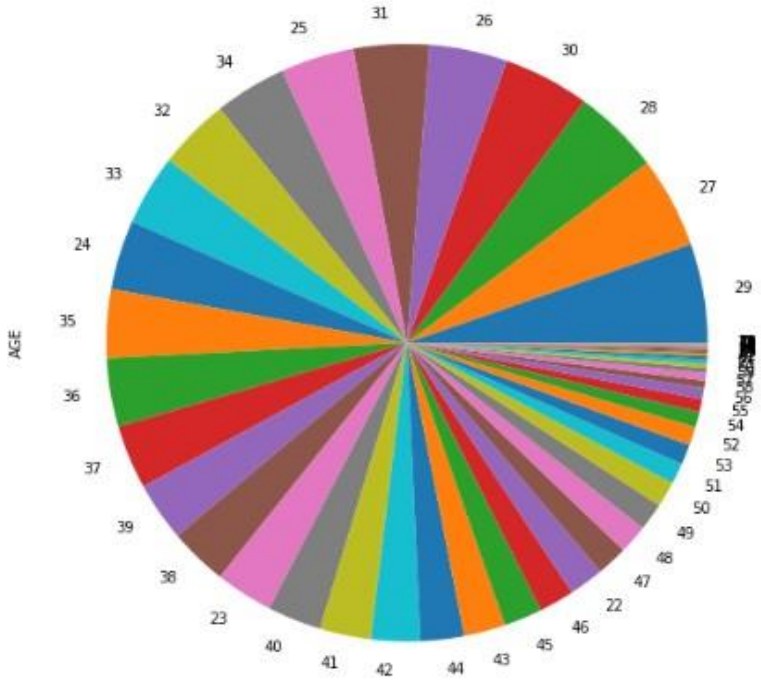


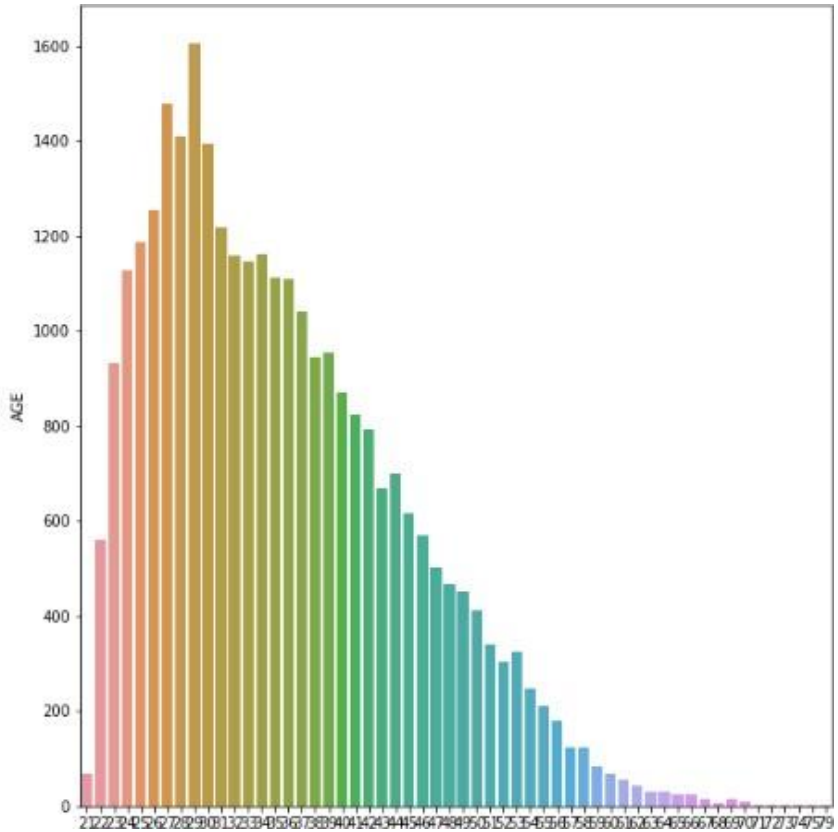
Marital Status

There is nearly no Significant correlation of default risk and marital status



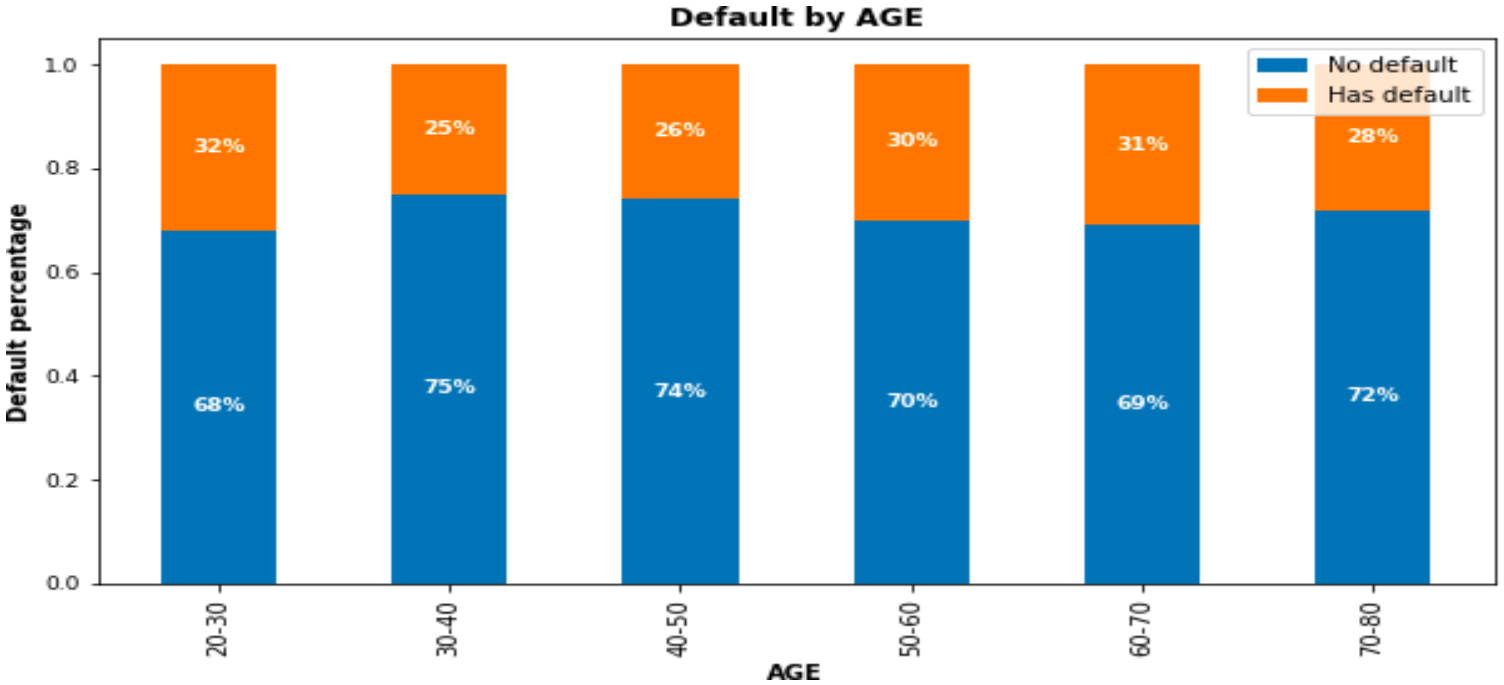
Age Distribution





Age wise defaulters

Observation-(30 to 50:Lowest Risk) & (<30 and >50:Risk Increases)



Modeling Overview

Supervised learning/Binary Classification

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

Models Used:

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

Modeling Steps

Data Preprocessing

- Feature selection
- Feature engineering
- Train test data split(75%-25%)
- SMOTE oversampling(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

Logistic Modelling

Parameters :C = 0.01 , Penalty = L2

Results:

The accuracy on test data is 0.7483307652799178

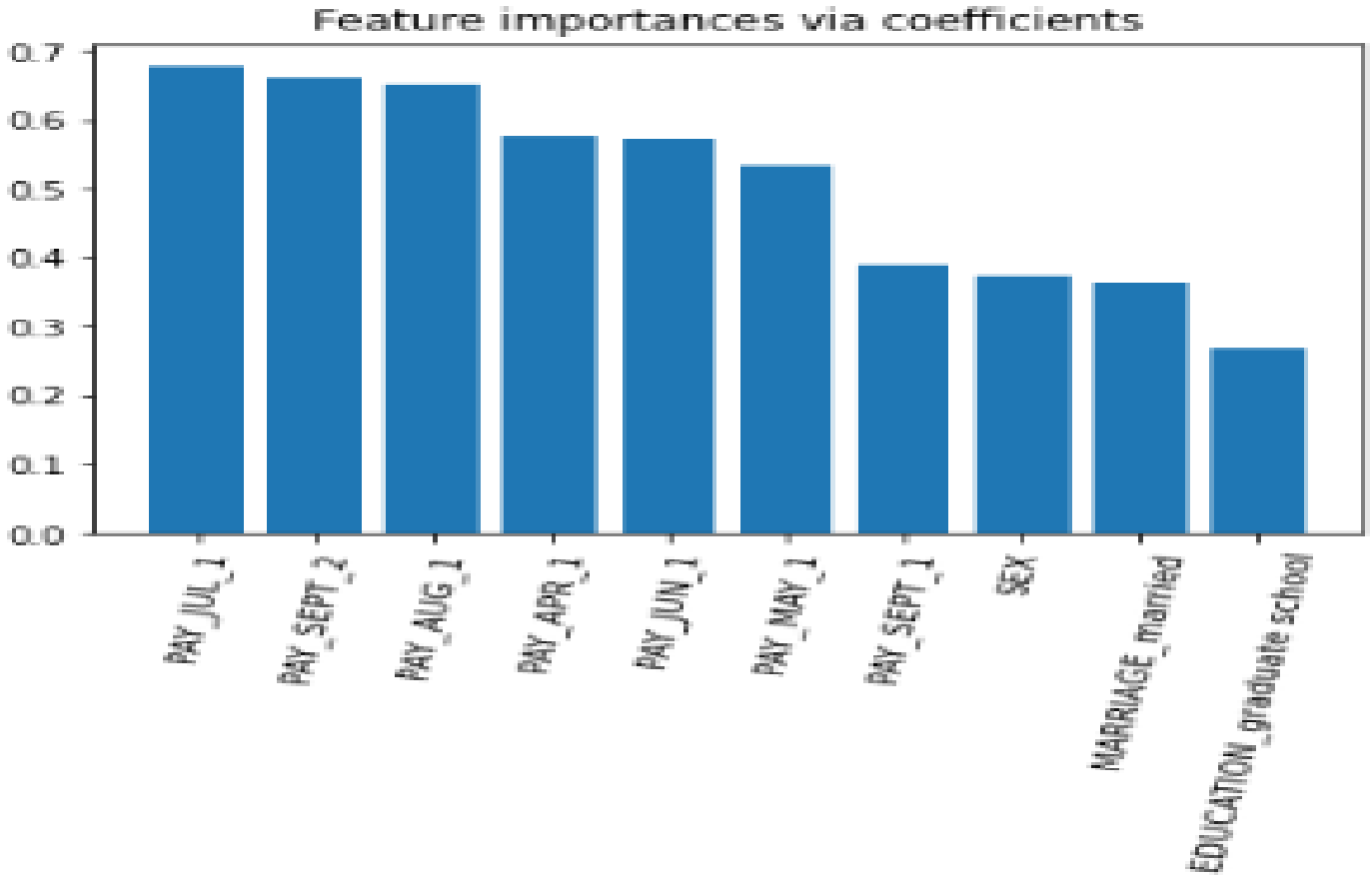
The precision on test data is 0.6831022085259374

The recall on test data is 0.7855877141169522

The f1 on test data is 0.7307692307692308

The roc_score on test data is 0.7526302950412113

Logistic feature Importances



SVM Modelling

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

Results:-

The accuracy on test data is 0.7483307652799178

The precision on test data is 0.6831022085259374

The recall on test data is 0.7855877141169522

The f1 on test data is 0.7307692307692308

The roc_score on test data is 0.7526302950412113

Random Forest Metrics

Parameters :-max_depth=30 , n_estimators=150

Results:-

The accuracy on test data is 0.842663927409690 1

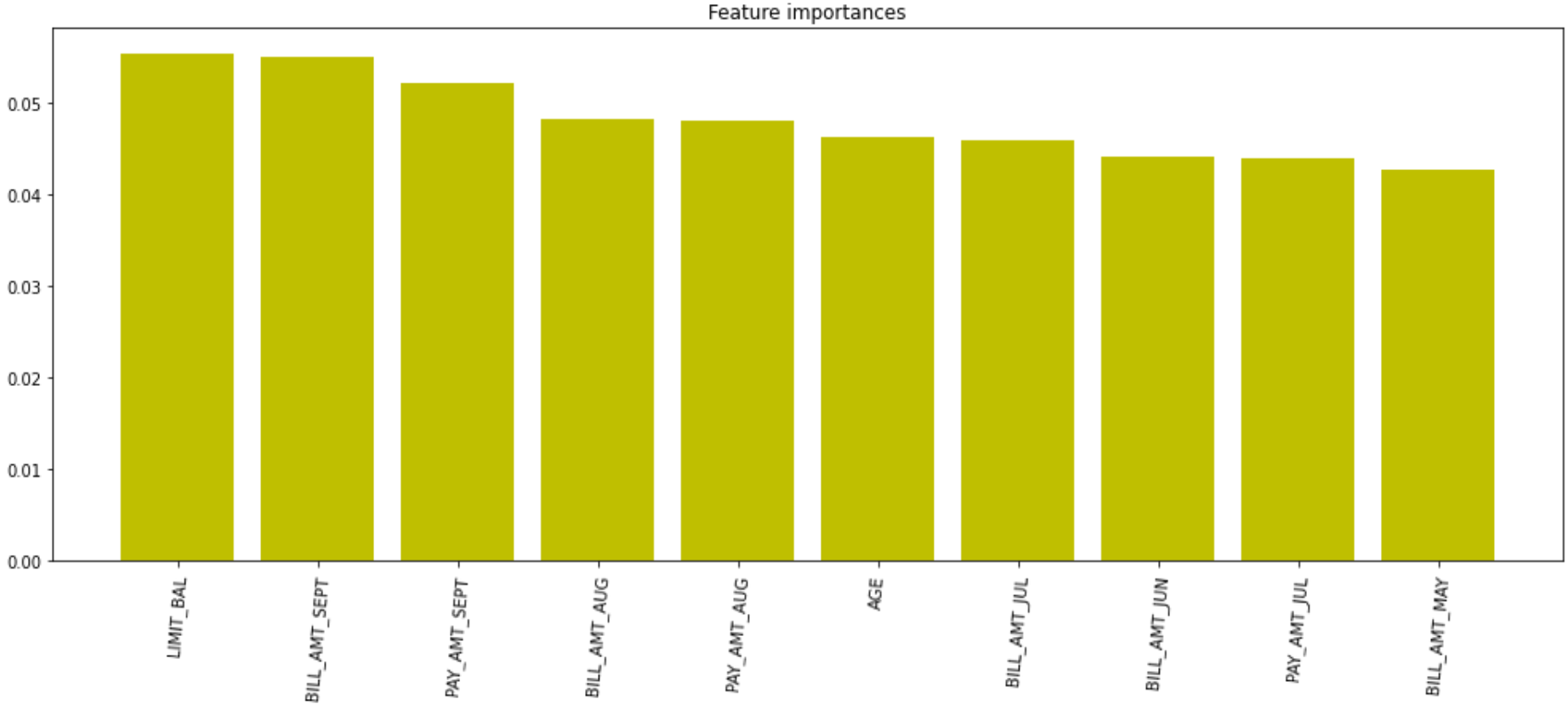
The precision on test data is 0.8145865434001027

The recall on test data is 0.863050970433520 7

The f1 on test data is 0.838118724678527 4

The roc_score on test data is 0.8437478875510123

Random Forest feature Importances



XGBoost Modelling

Parameters : max_depth= 15 , min_child_weight= 8

Results:-

The accuracy on test data is 0.8320493066255779

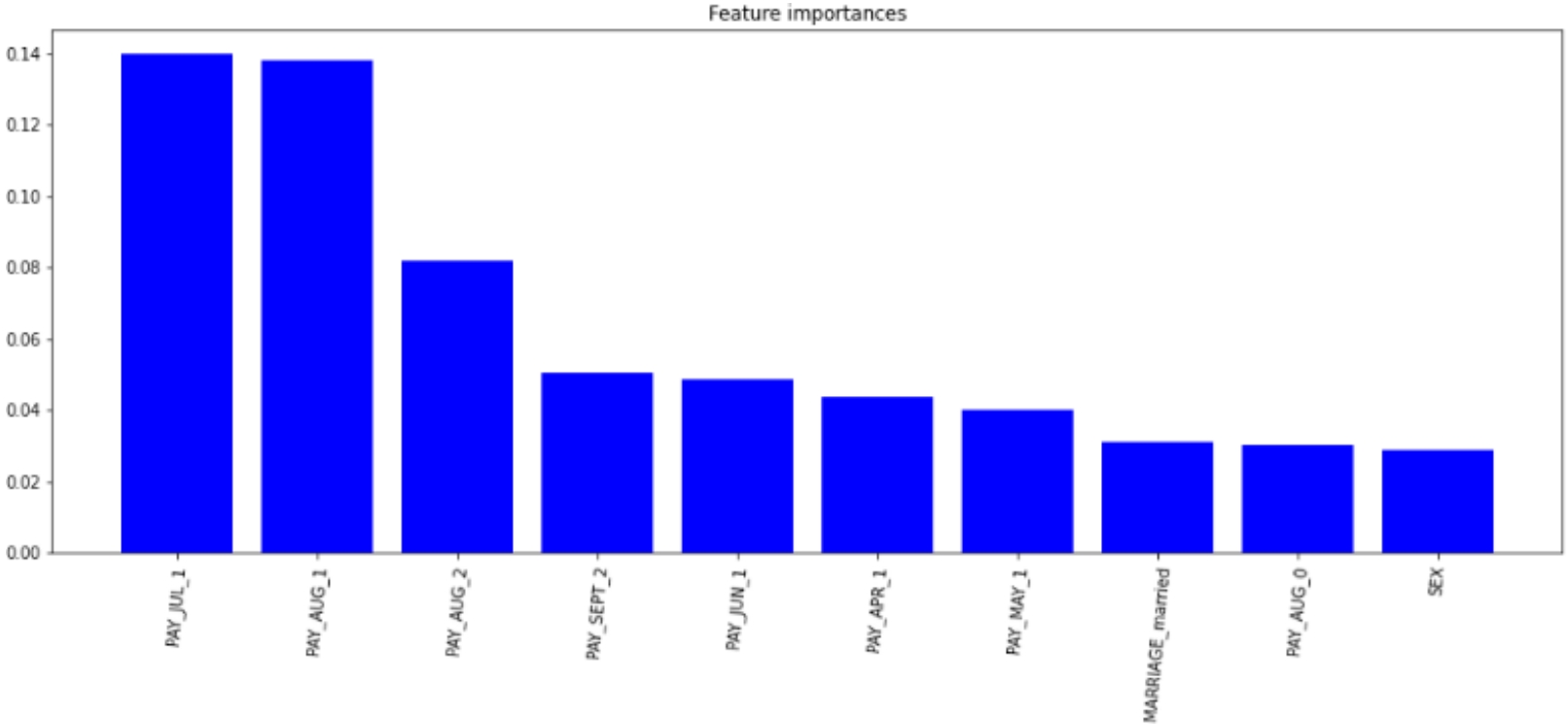
The precision on test data is 0.7890772128060264

The recall on test data is 0.8632702753324593

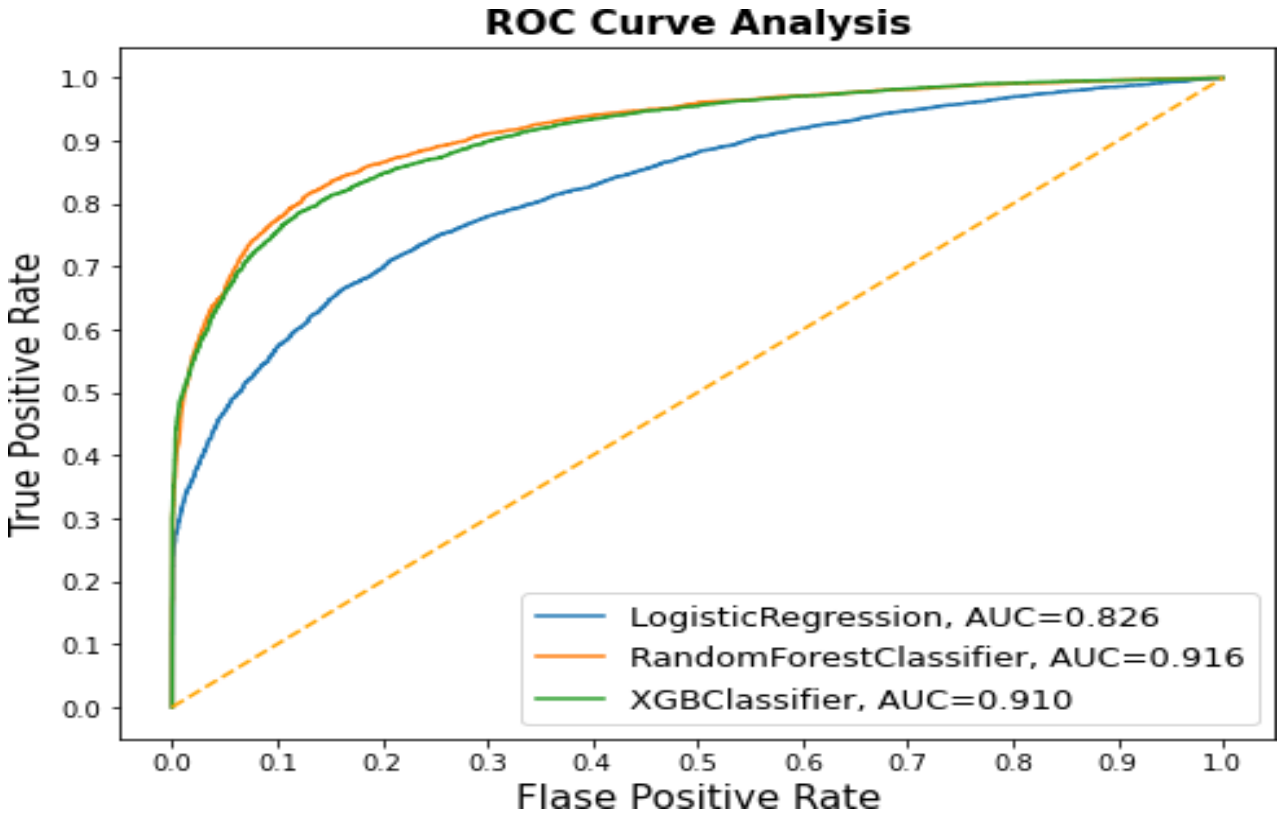
The f1 on test data is 0.8245080500894455

The roc_score on train data is 0.8345202078223072

X Gradient Boosting feature Importances



AUC-ROC curve Comparision



Challenges

- 9 Categorical variables present
- Huge dataset
- Understanding the columns .
- Feature engineering.
- Getting a higher accuracy on the models .

Conclusion

- XGBoost provided us the best results giving us a recall of 86 percent(meaning out of 100 defaulters 86 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 a recall of nearly 79.

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.753267	0.748331	0.683102	0.785588	0.730769
1	SVC	0.809851	0.781207	0.722957	0.818262	0.767663
2	Random Forest CLf	0.998060	0.842664	0.814587	0.863051	0.838119
3	Xgboost Clf	0.910746	0.832049	0.789077	0.863270	0.824508

Thank You