

# Capstone Project

## Credit Card Default Prediction

Individual Project  
Shubham Naik

# Presentation Overview

- Introduction
- Problem statement
- Data Information
- EDA & feature engineering
- Preparing Data for modeling
- Implementing Model
- Model summary
- Conclusion



# Introduction

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Credit card default happens when you have become severely delinquent on your credit card payments. Default is a serious credit card status that affects not only your standing with that credit card issuer but also your credit standing in general and your ability to get approved for other credit-based services.

# 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 various machine learning classification techniques to evaluate which customers will default on their credit card payments

# Data Information

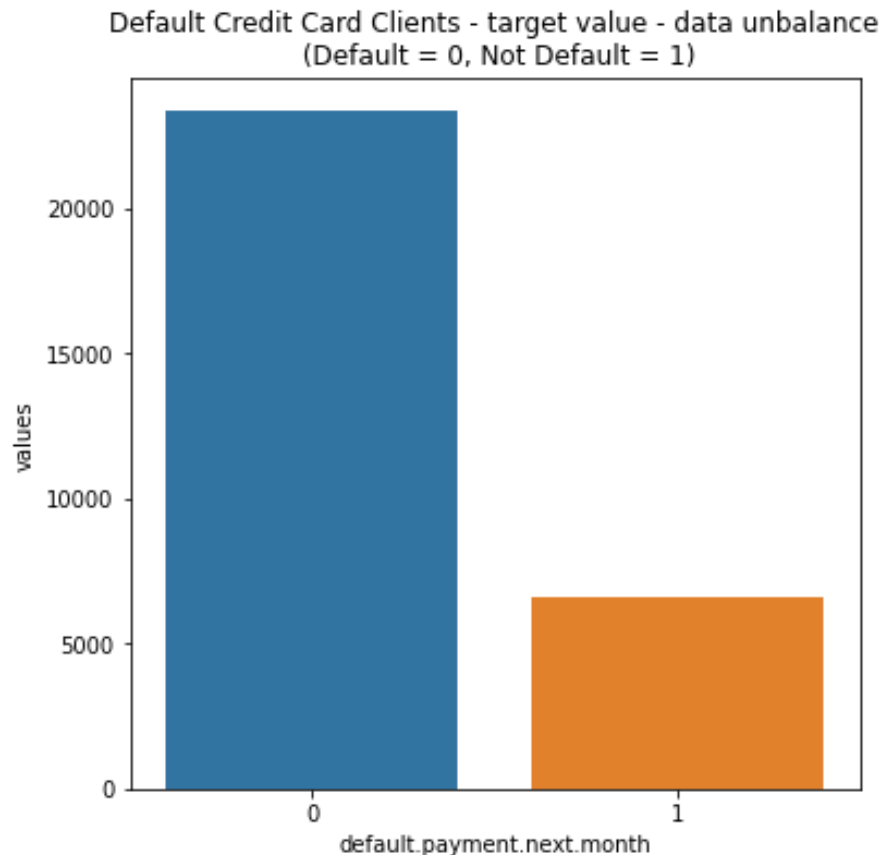
- **LIMIT\_BAL** : Amount of the given credit it includes both the individual consumer credit and his/her family credit.
- **SEX** : Gender (1 = male; 2 = female).
- **EDUCATION** : Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- **MARRIAGE** : Marital status (1 = married; 2 = single; 3 = others).
- **AGE** : Age (year).
- **PAY\_0 - PAY\_6** : History of past payment from April to September 2005
- **BILL\_AMT1 - BILL\_AMT6** : Amount of bill statement from April to September 2005
- **PAY\_AMT1 - PAY\_AMT6** : Amount of previous payment from April to September 2005
- **DEFAULT PAYMENT NEXT MONTH** : default payment (Yes = 1, No = 0), as the response variable

# Basic Exploration

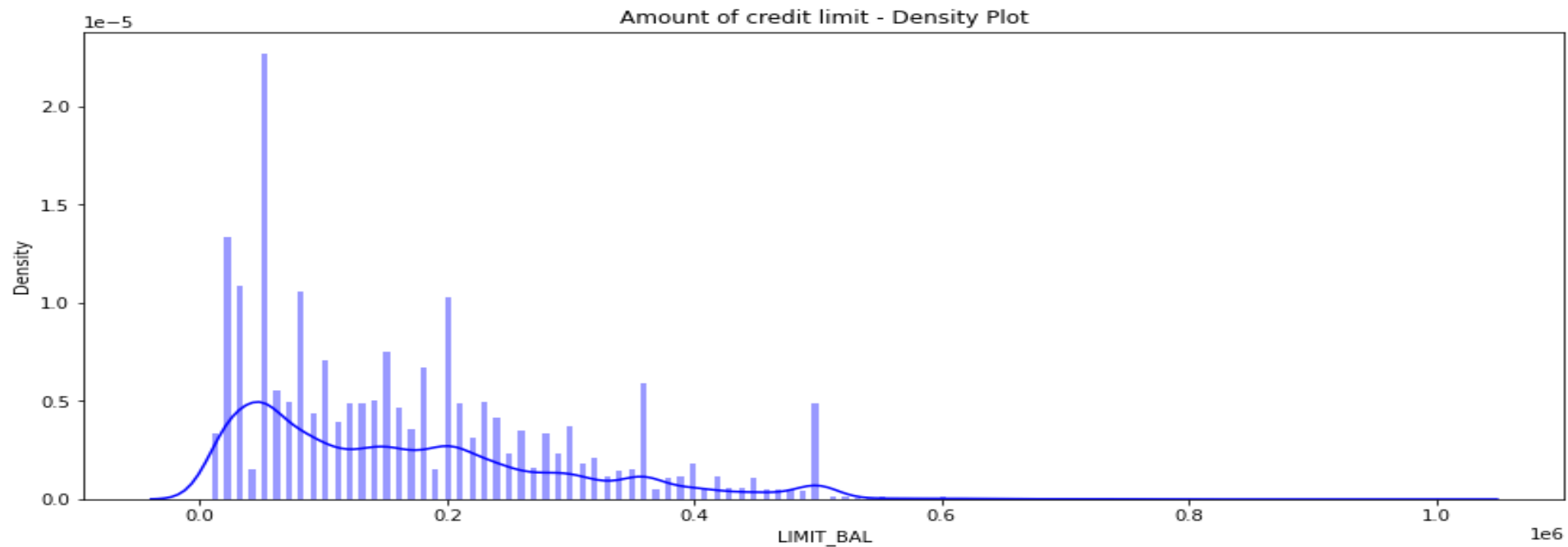
- There are 30,000 distinct credit card clients
- There is no missing data in the entire dataset.
- A number of 6,636 out of 30,000 (or 22%) of clients will default next month
- Education level is mostly graduate school and university.
- Average age is 35.5 years
- The average value for the amount of credit card limit is 167,484.

# EDA

Distribution of target classes is highly imbalanced, non-defaults far outnumber defaults. This is common in these datasets since most people pay credit cards on time



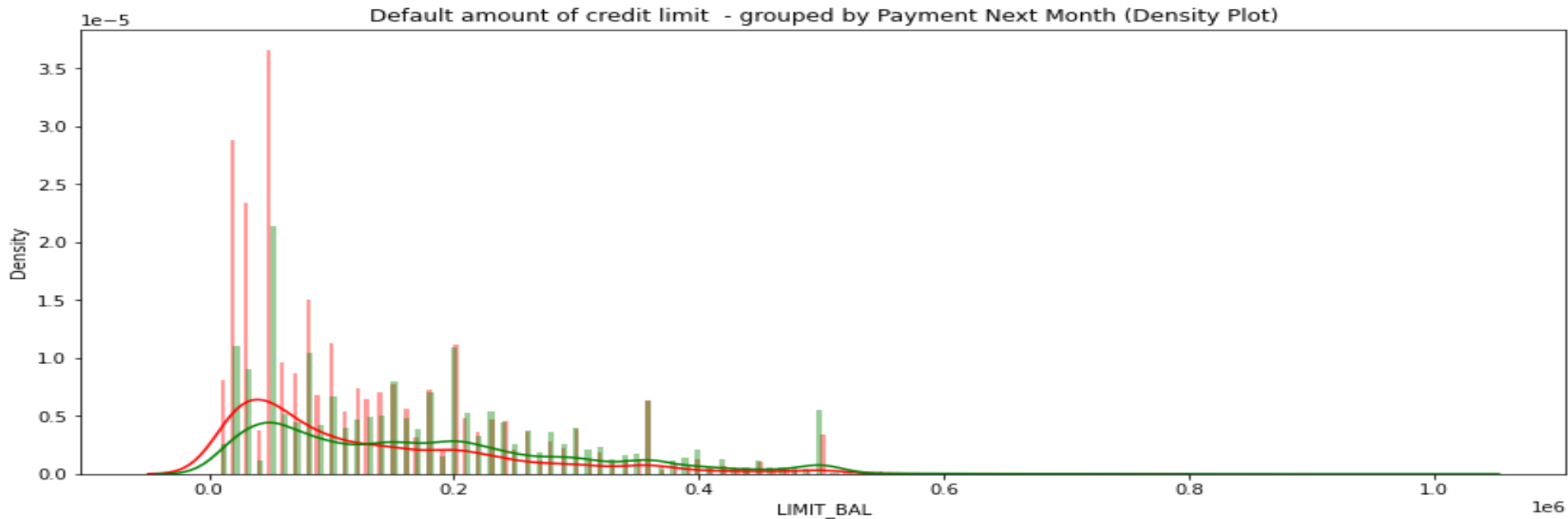
Distribution of credit limit amounts. The largest credit limit amount is \$50k



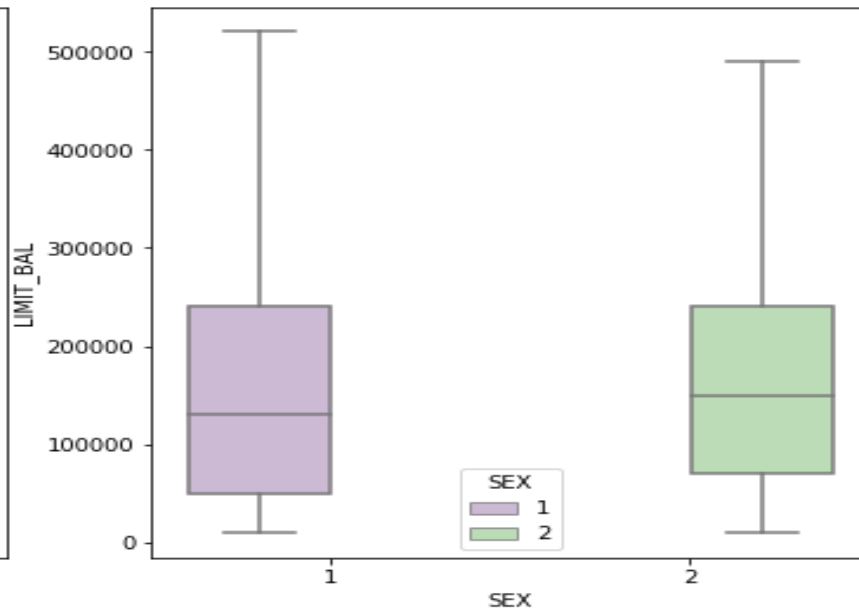
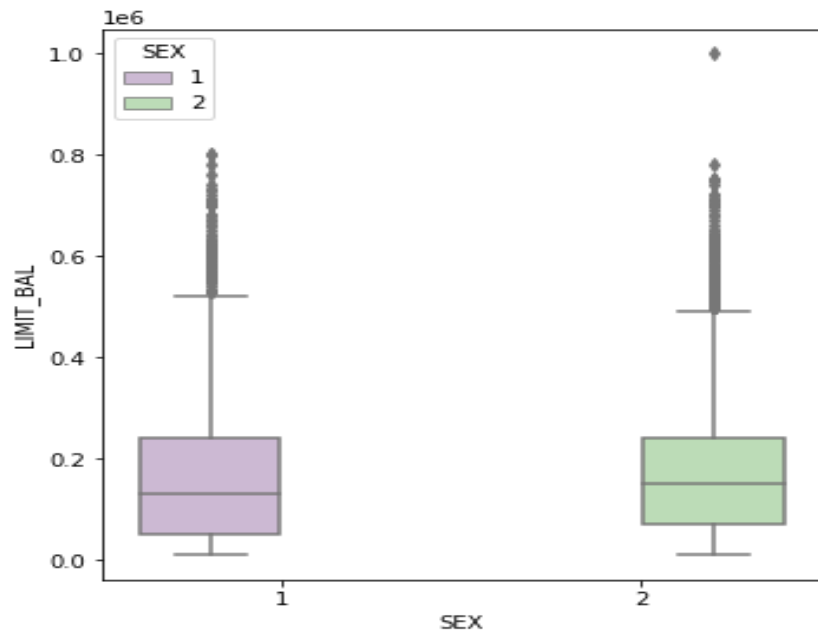


# EDA

Most of defaults are for credit limits 0-100,000 (and density for this interval is larger for defaults than for non-defaults).

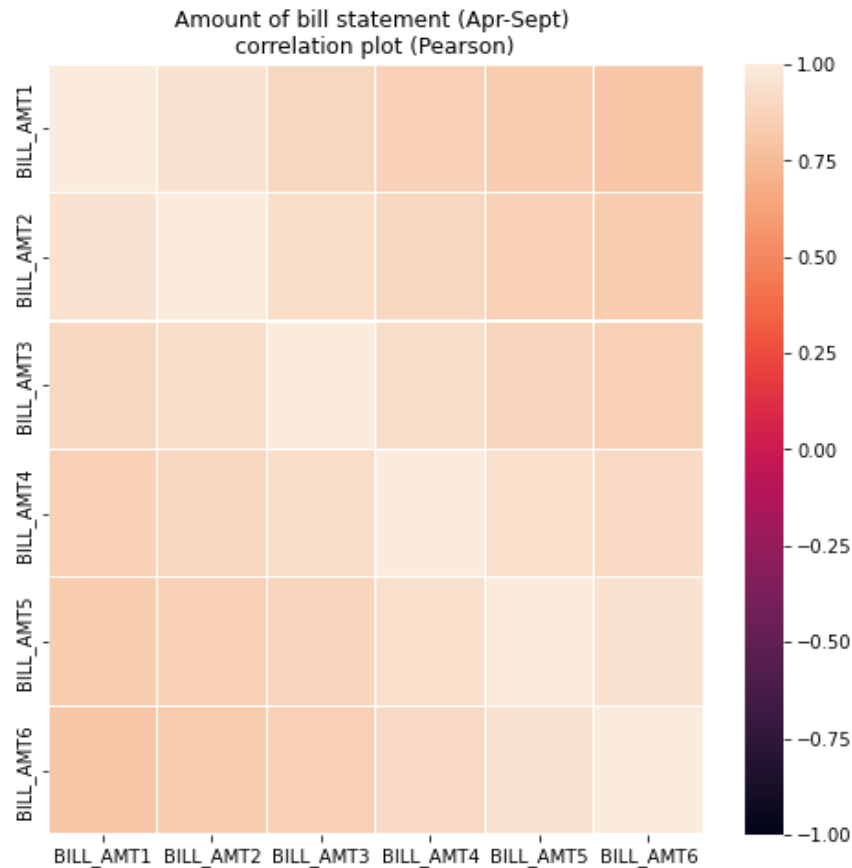


Credit Limit by Sex. The data is evenly distributed amongst males and females.



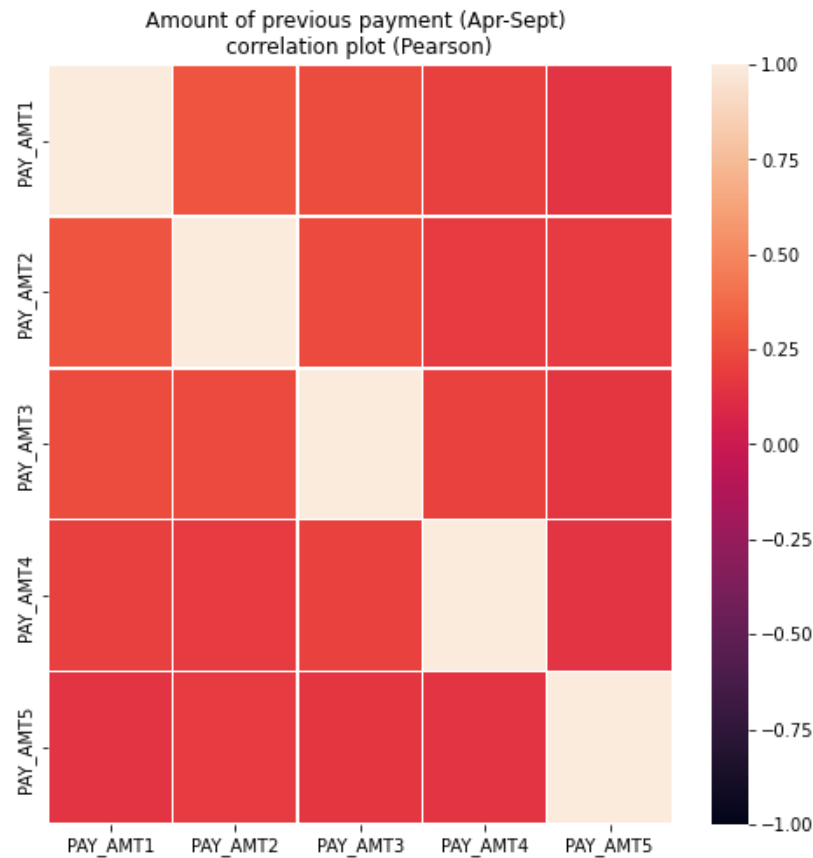
# EDA

- Amount of bill statement.  
Correlation is decreasing with  
distance between months.  
Lowest correlations are between  
Sept-April.



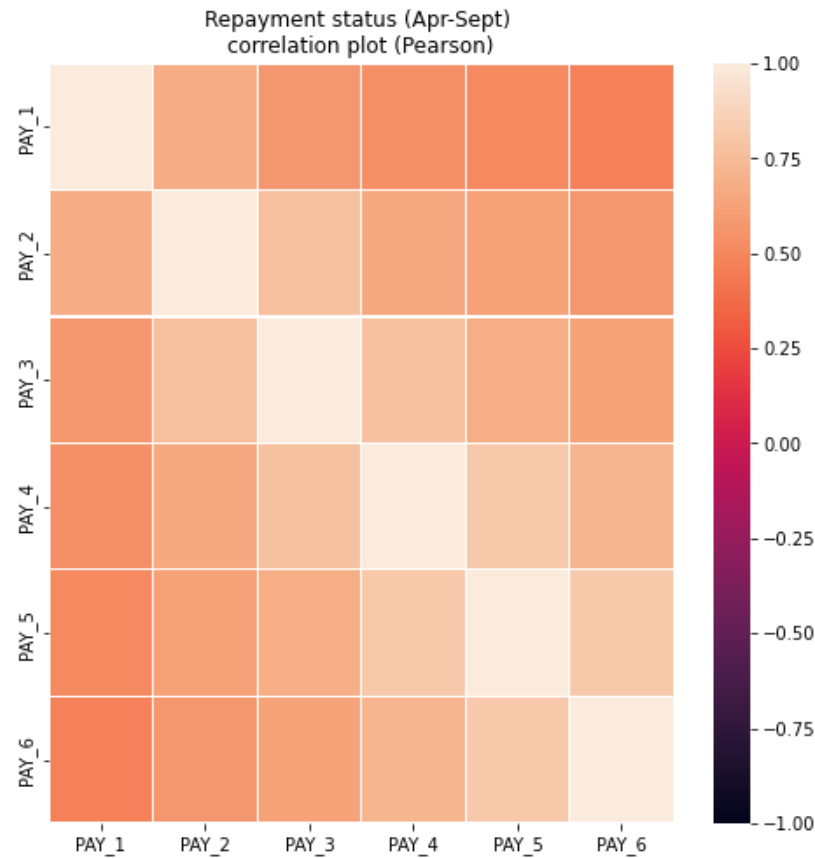
# EDA

- Amount of previous payment.  
There are no correlations  
between amounts of previous  
payments for April-Sept 2005.



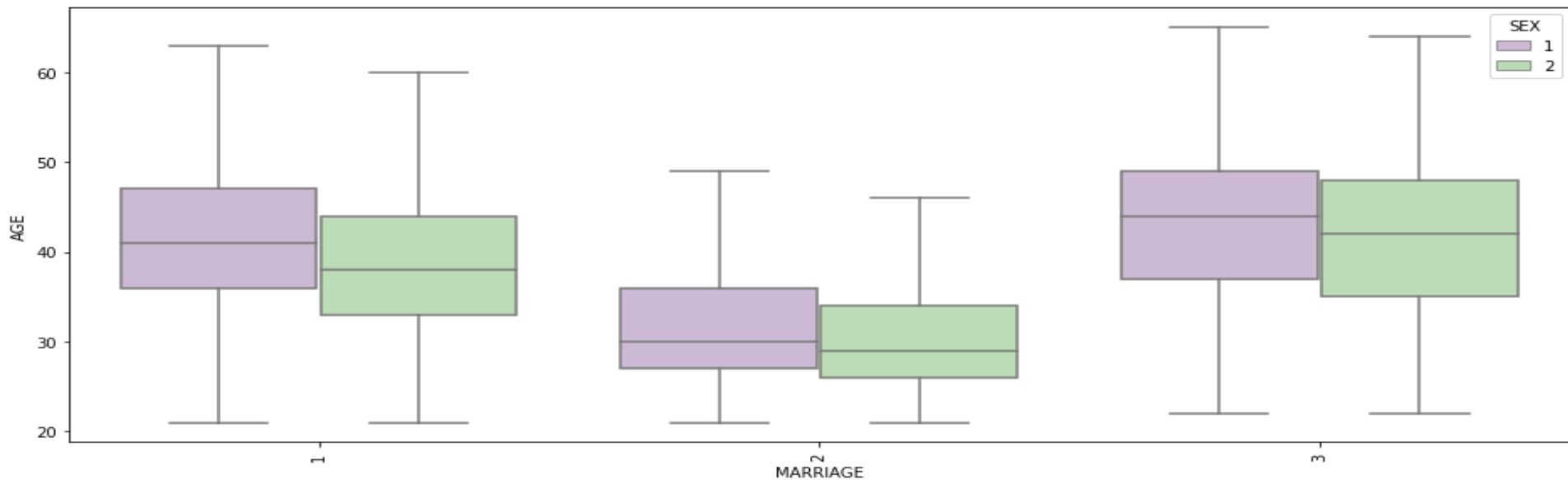
# EDA

- Payment status. Correlation is decreasing with distance between months. Lowest correlations are between Sept-April.



Marriage, age, and sex. The dataset mostly contains couples in their mid-30s to mid-40s and single people in their mid-20s to early-30s.

Marriage status meaning is:  
1 : married  
2 : single  
3 : others



Boxplots with age distribution grouped by education and marriage

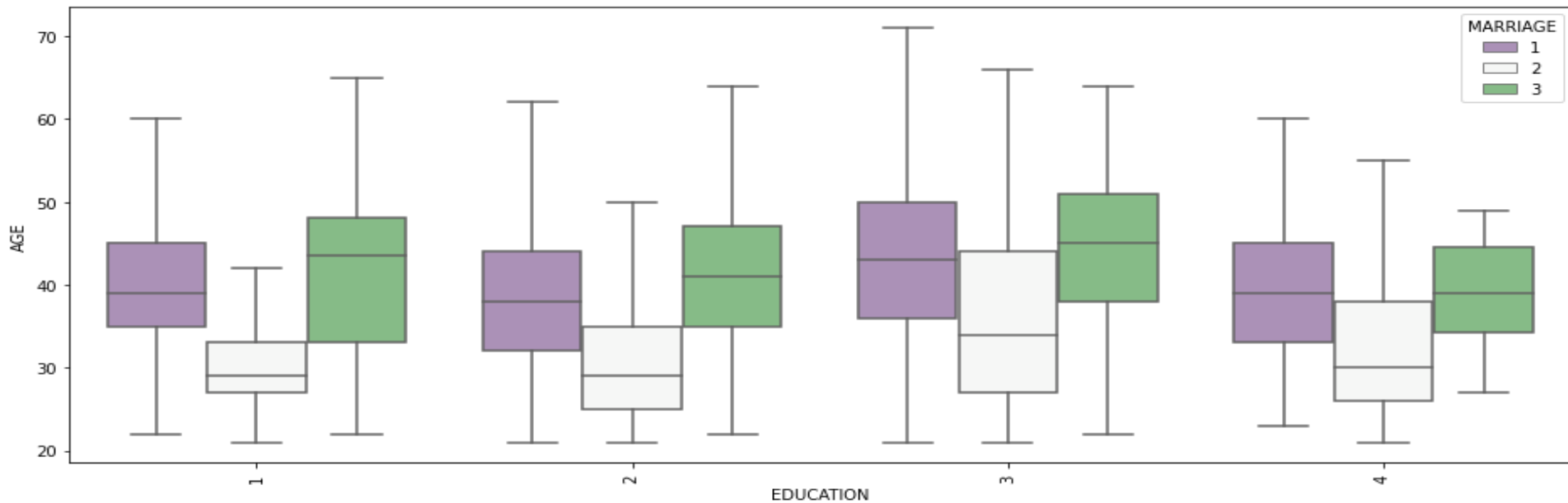
Education status meaning is:

1 : graduate school

2 : university

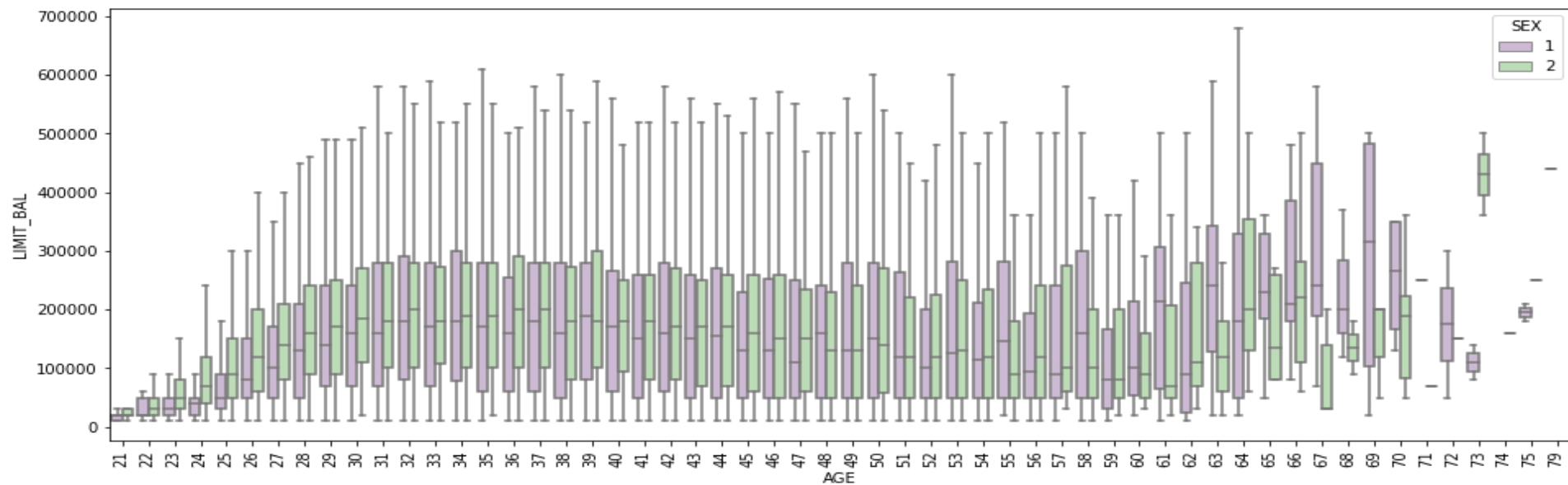
3 : high school

4 : others



## Boxplots with credit amount limit distribution grouped by age and sex.

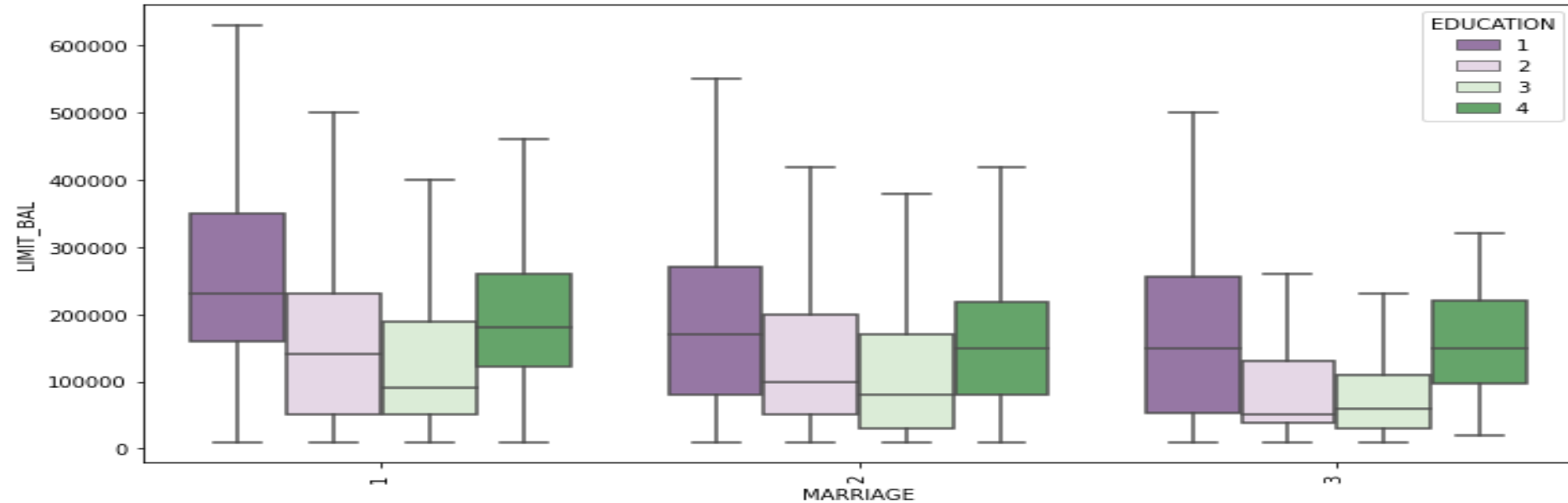
Mean values are generally smaller for males than for females, with few exceptions, for example at age 39, 48, until approximately 60, where mean values for males are generally larger than for females.





Boxplots with credit amount limit distribution grouped by marriage status and education level.

Marriage status meaning is:  
1 : married  
2 : single  
3 : others



# Modeling Steps

## Data Preprocessing

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

## Data Fitting & Tuning

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

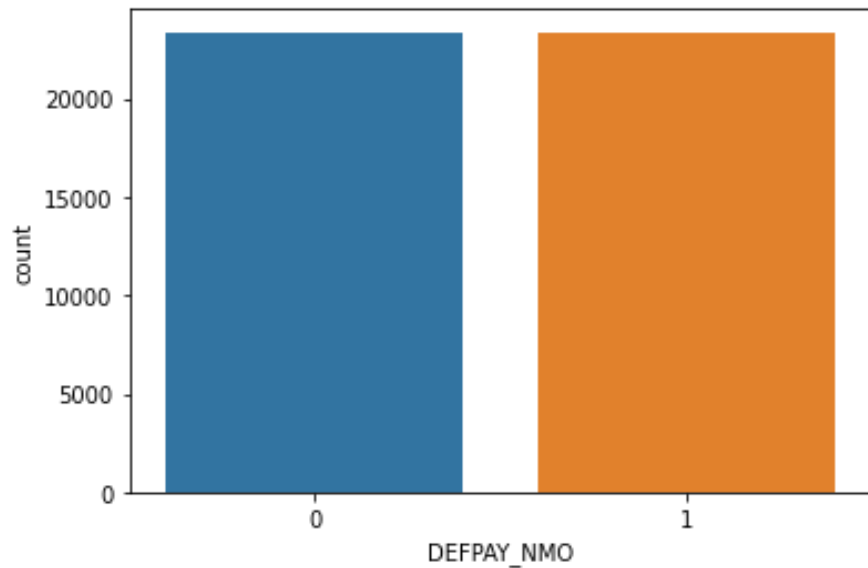
## Model Evaluation

- Model testing
- Precision\_Recall Score
- Compare with the other models

# Preparing Data for modeling

As we have seen earlier that we have imbalanced dataset. So to remediate Imbalance we are using SMOTE(Synthetic Minority Oversampling Technique)

- Original dataset shape 30000
- Resampled dataset shape 46728



# Applying Model

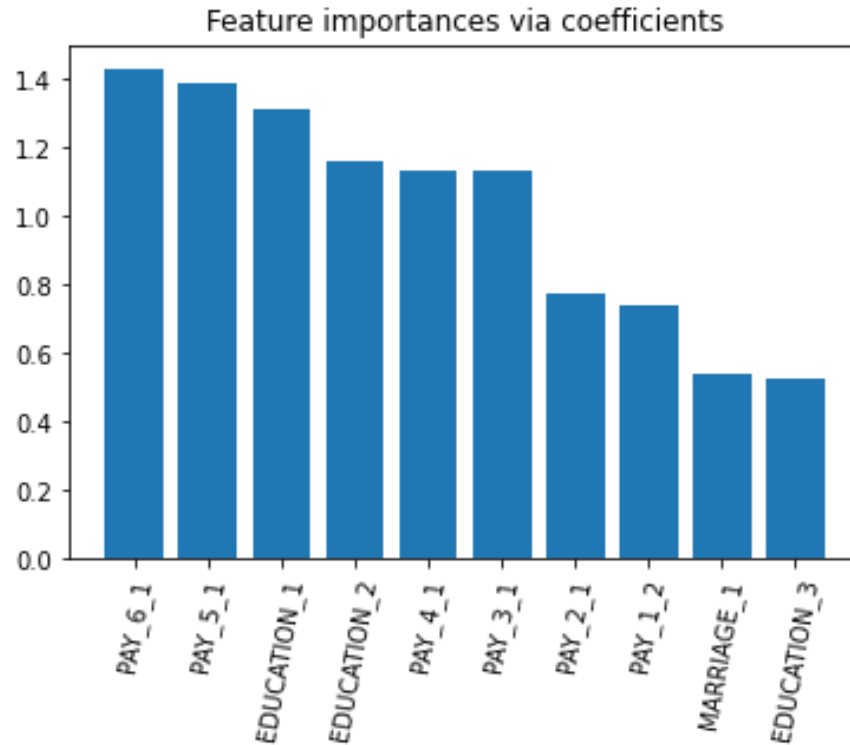
- Logistic Modelling

Parameters :

- $C = 0.01$
- Penalty = L2

```
The accuracy on test data is 0.7522647835080961
The precision on test data is 0.6908260807533172
The recall on test data is 0.7875731945348081
The f1 on test data is 0.7360340503154215
The roc_score on test data is 0.7561293745511038
```

# Logistic feature importance



# Applying Model

- SVM Modelling

Parameters :

- $C = 10$
- Kernel = 'rbf'

```
The accuracy on test data is 0.778086882088594
The precision on test data is 0.7113710943073192
The recall on test data is 0.820875864339809
The f1 on test data is 0.7622105021783995
The roc_score on test data is 0.783125156839508
```

# Applying Model

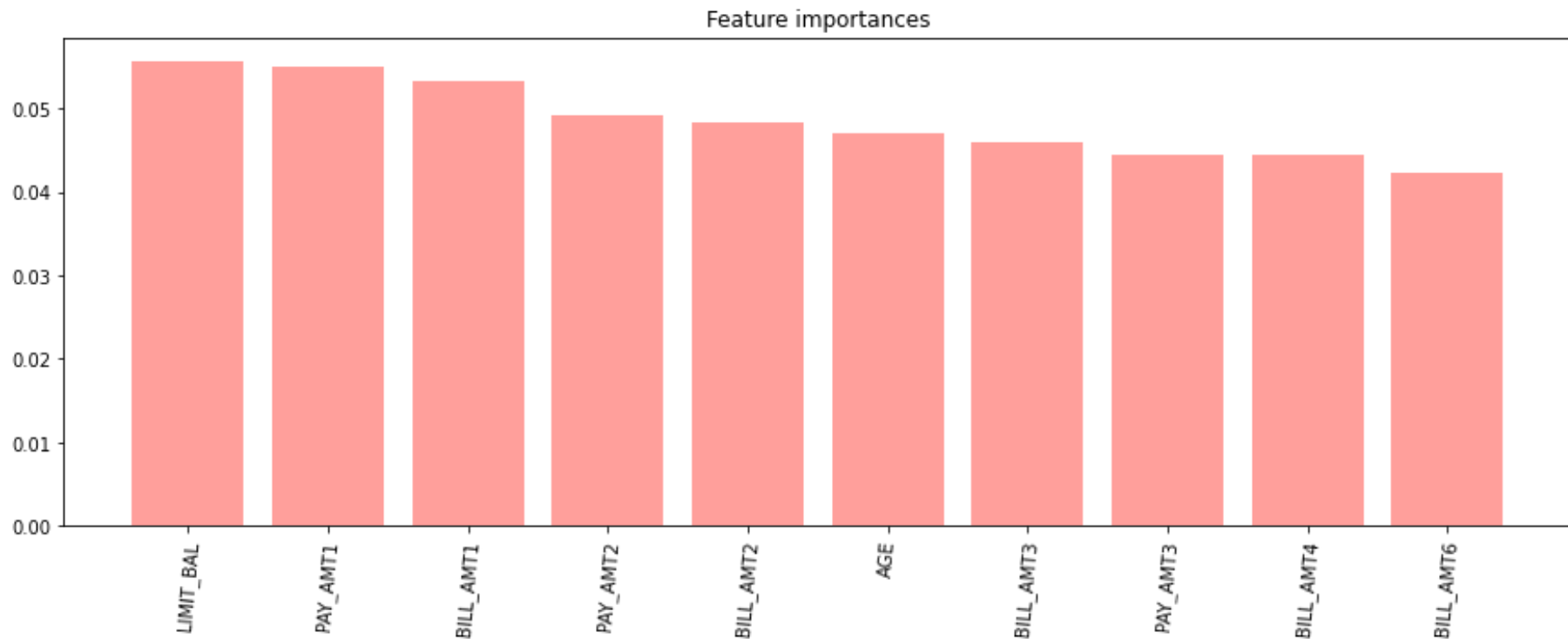
- Random Forest Metrics

Parameters :

- max\_depth=30
- n\_estimators=150

```
The accuracy on test data is  0.8354376203723518
The precision on test data is  0.8058210871736339
The recall on test data is    0.8565362450712769
The f1 on test data is       0.8304050577078586
The roc_score on test data is 0.8366182908858067
```

# Random Forest feature importance





# Applying Model

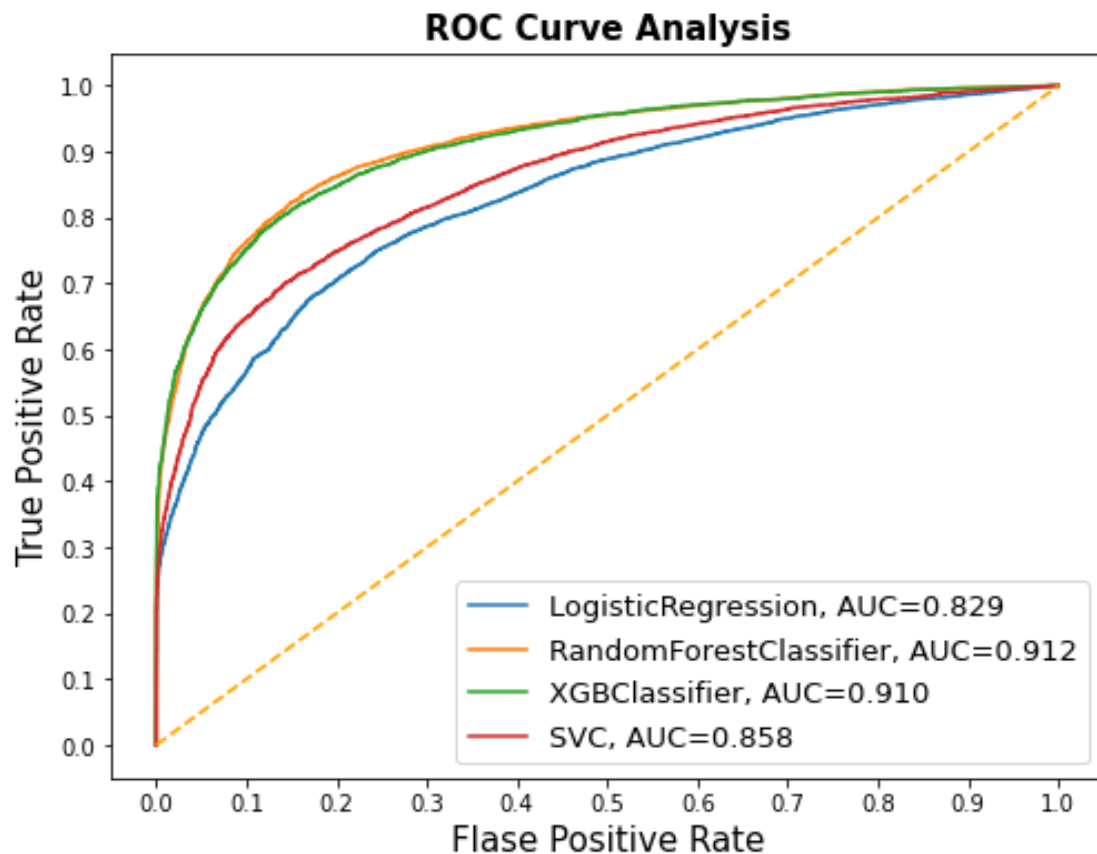
- XGBoost Modelling

Parameters :

- max\_depth= 10
- min\_child\_weight= 6

```
The accuracy on test data is  0.8305157286539696
The precision on test data is  0.7934084748180911
The recall on test data is    0.8569887501926337
The f1 on test data is       0.8239739220625277
The roc_score on train data is 0.8323456367165027
```

# AUC-ROC curve comparison



# 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 a recall of 78.

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 score
0	Logistic Regression	0.752484	0.752265	0.690826	0.787573	0.736034
1	SVC	0.807912	0.778087	0.711371	0.820876	0.762211
2	Random Forest CLf	0.998563	0.835438	0.805821	0.856536	0.830405
3	Xgboost Clf	0.917301	0.830516	0.793408	0.856989	0.823974

Thank You