CAPSTONE PROJECT - III

Credit card default prediction

by

Nakshith D N

Contents

Introduction

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can be proven by industry business reports and statistical data. So, assessing, detecting and managing default risk is the key factor in generating revenue and reducing loss for the banking and credit card industry.

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

Data description

This dataset has 25 columns with 30000 numerical values.

The features of the dataset are

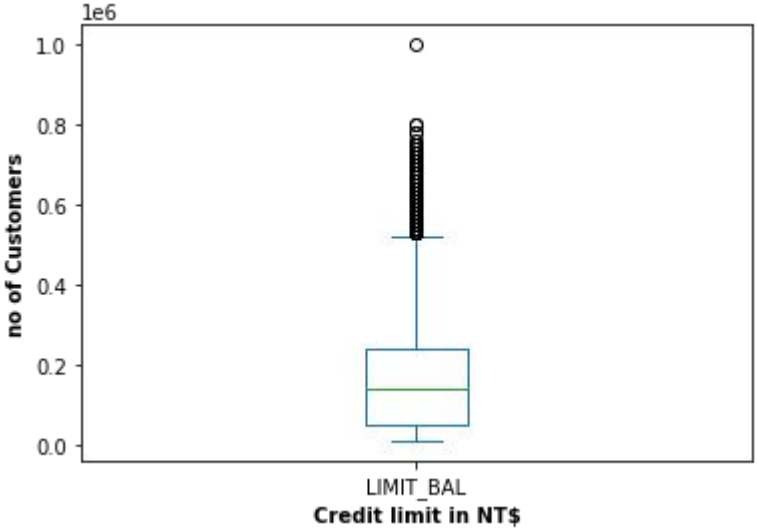
‘ID’
'LIMIT_BAL'
'SEX', 'EDUCATION', 'MARRIAGE', 'AGE'
'PAY_0', 'PAY_2','PAY_3', 'PAY_4', 'PAY_5', 'PAY_6'
'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6'
'PAY_AMT1','PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6',
'default payment next month'

Data cleaning

There are no null values,

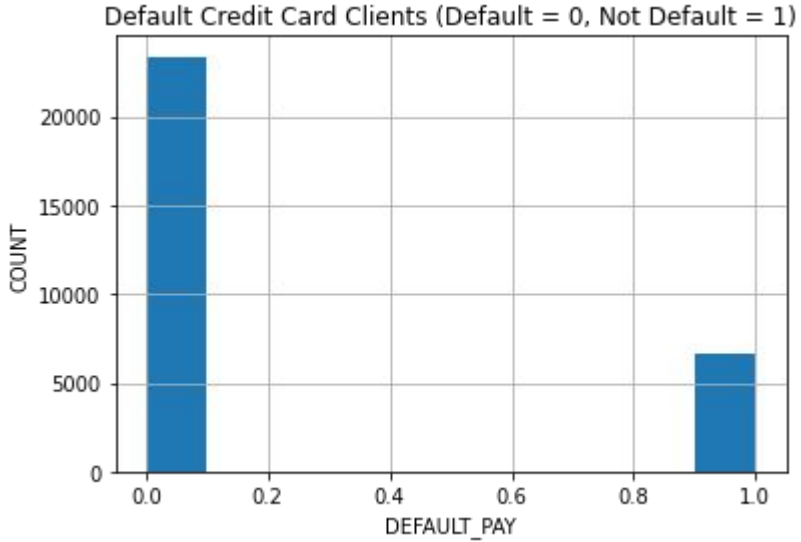
But the observations from other columns indicate that the outlier point in the figure around 90000 was just a customer who has an excellent payment history.

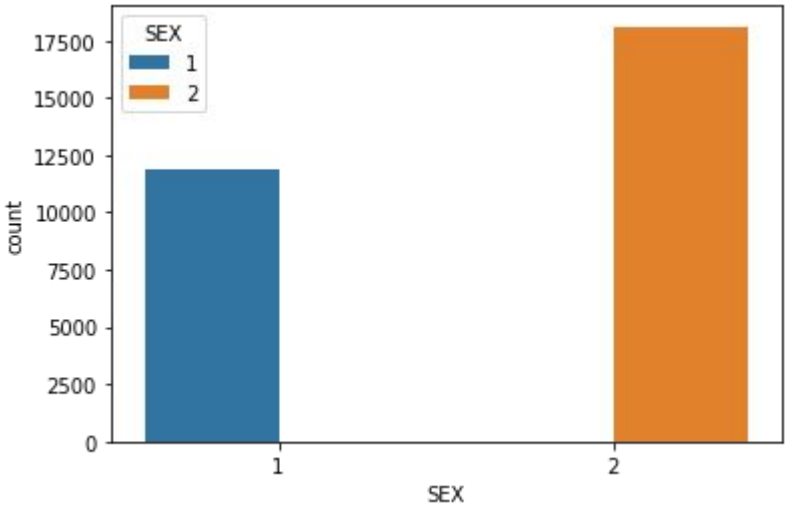
Therefore it should be considered as a valid data

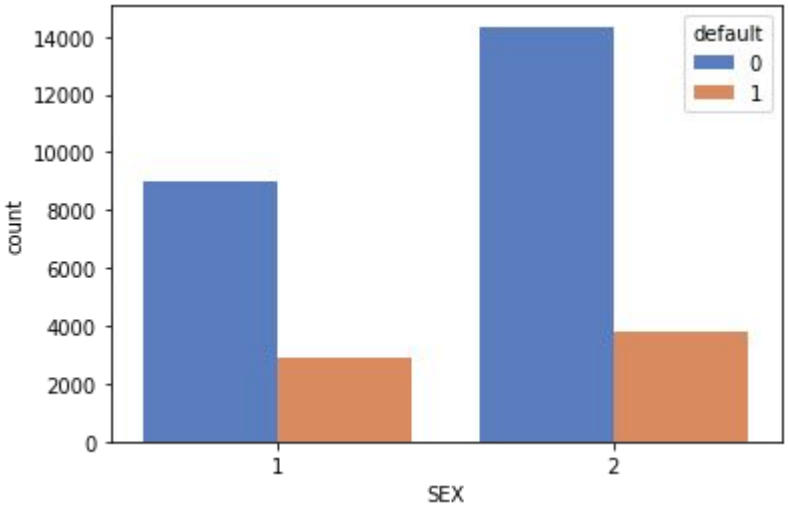


Data visualization and Analysis

Percentage of Defaulters are smaller than the Non Defaulters in the given dataset



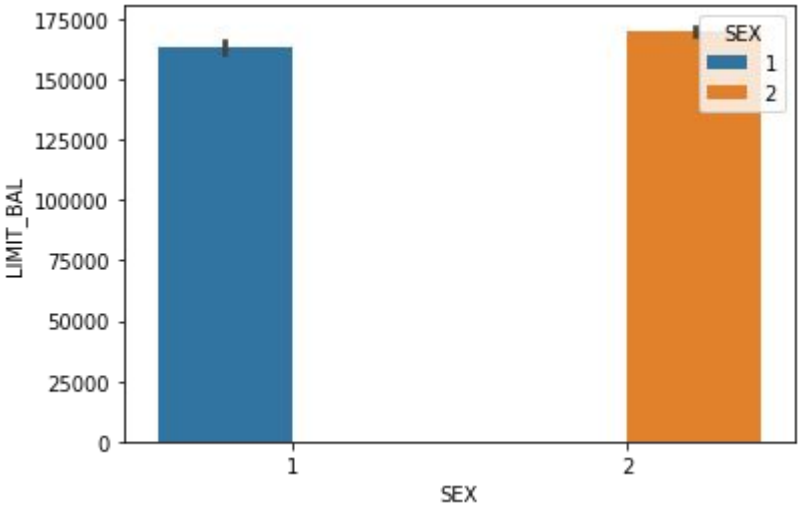


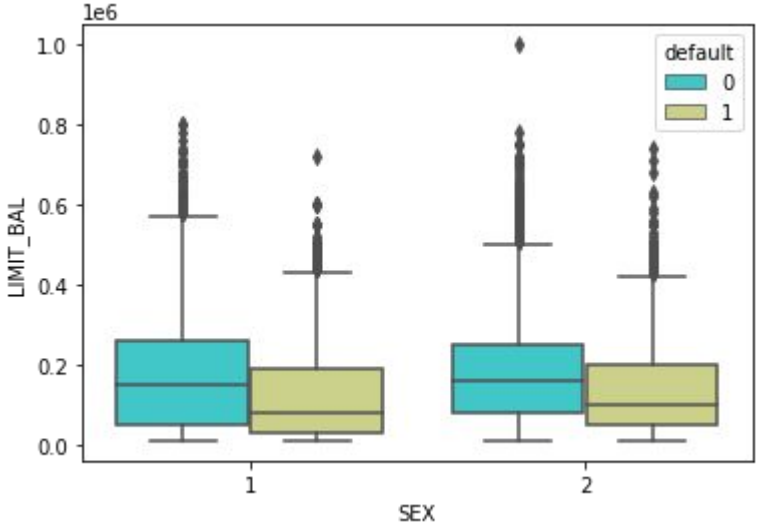


SEX column's distribution. 1: male; 2: female

- Female count is higher compared to men.
- It is evident from the above figure that females have overall less default payments wrt males

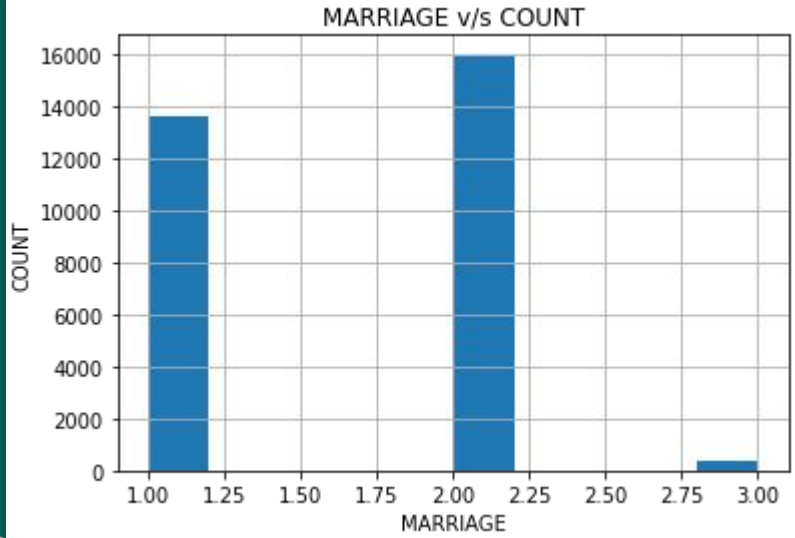
Non-Defaults have a higher proportion of Females

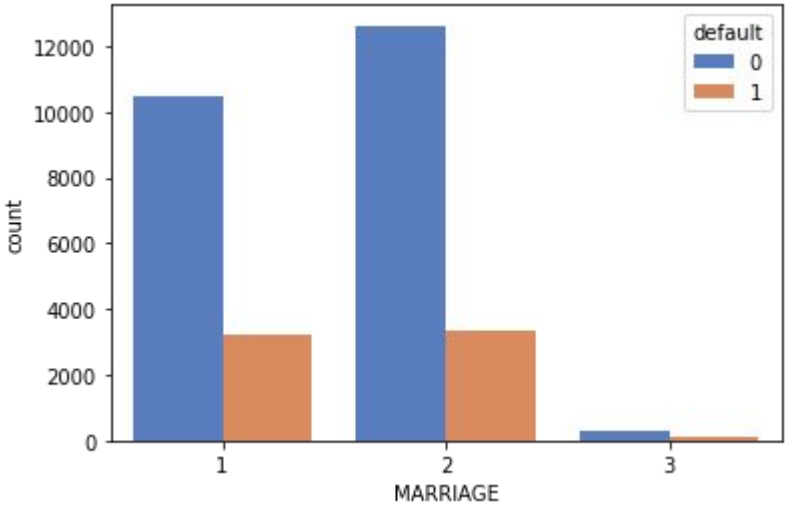




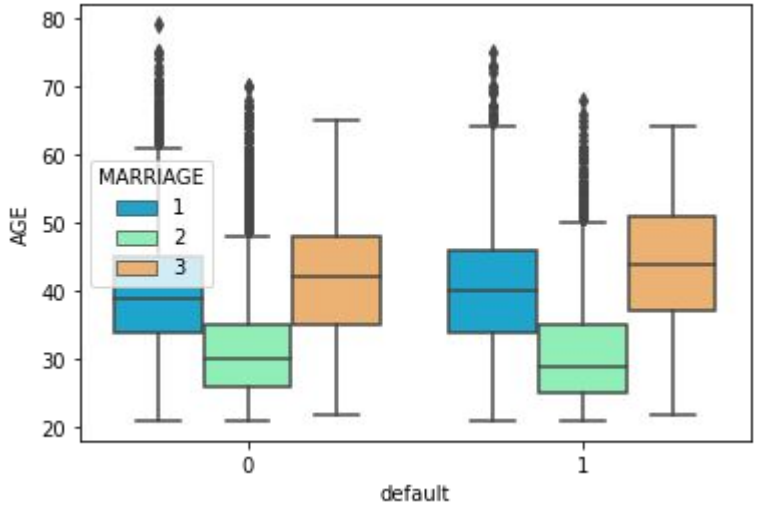
SEX column's distribution. 1: male; 2: female

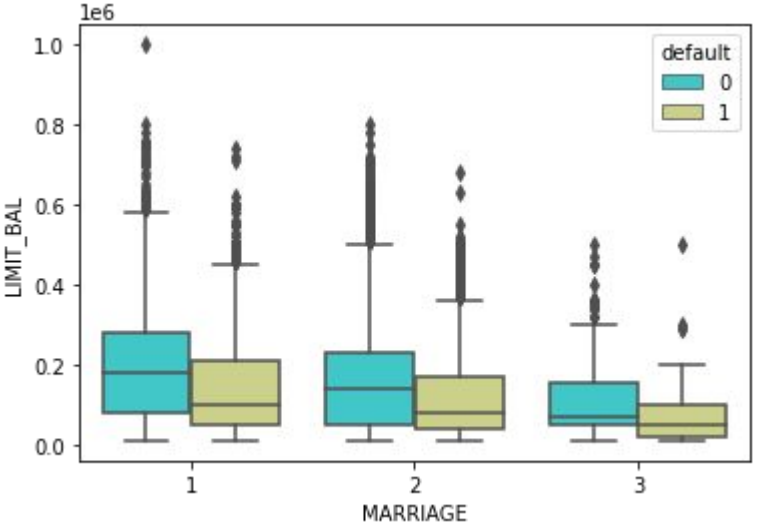
- Amount of given credit in NT dollars sexwise
- Amount of given credit in NT dollars sexwise with both default and no default





- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- More number of credit holders are single
- From the above plot it is clear that those people who have marital status single have less default payment wrt married status people



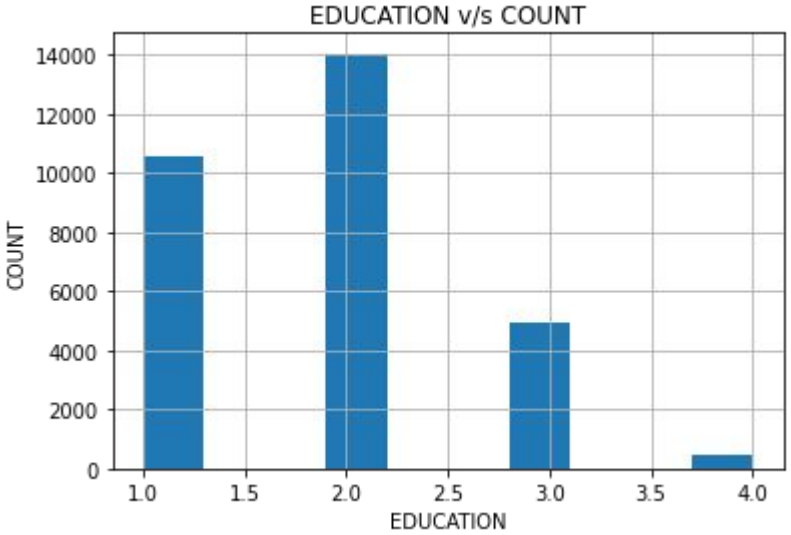


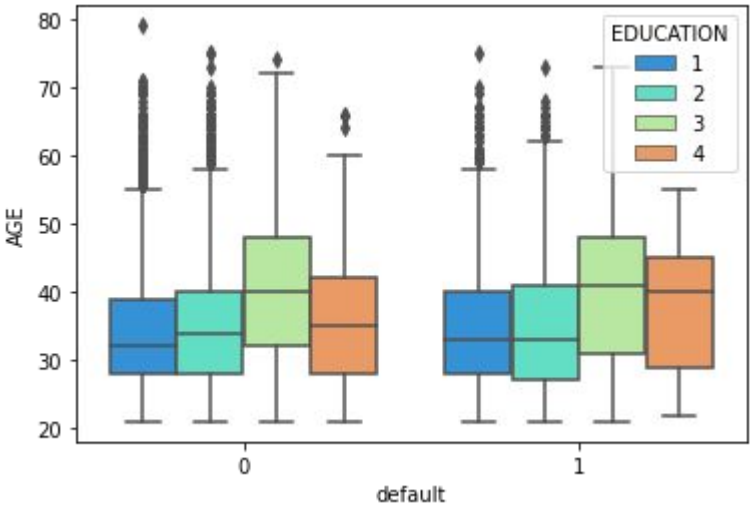
- MARRIAGE: Marital status (1=married, 2=single, 3=others)

The above plot shows the age group of marriage feature with default and non default,

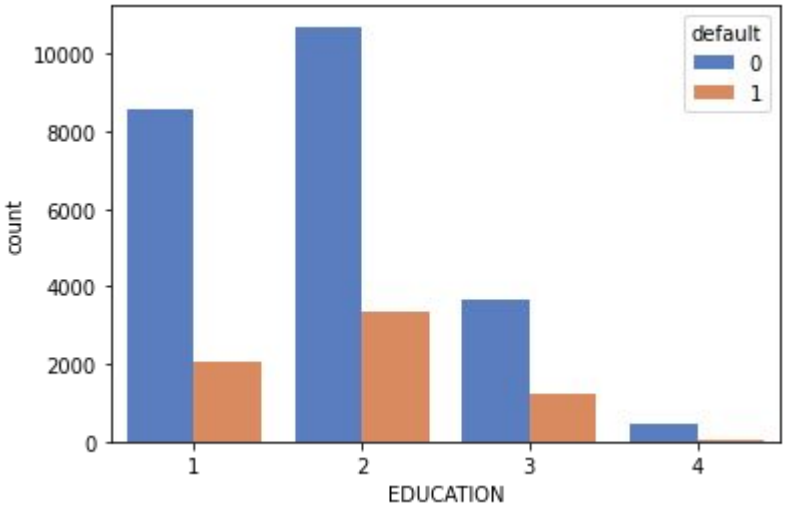
And the age group of married around 40 years old have high defaults compared to single people,

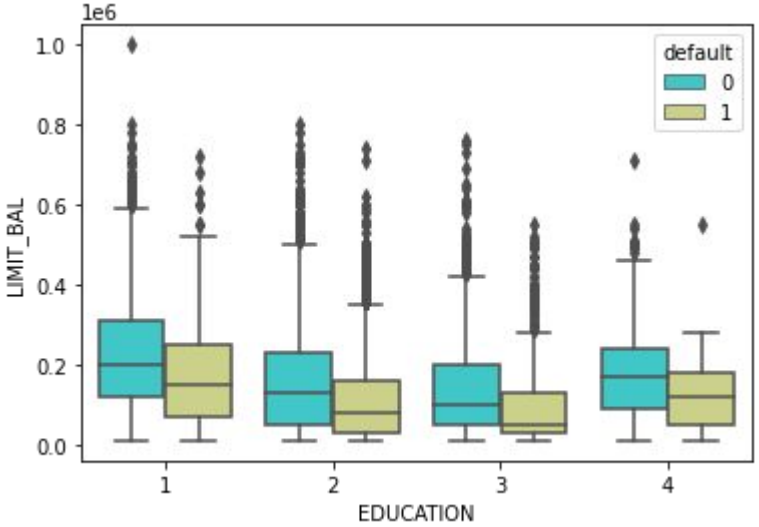
And also checking for default and no default w.r.t marriage groups and limit balance





- Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- More number of credit holders are university students followed by Graduates and then High school students
- And we have are checking for default and no default w.r.t age wise and with educational background
- Graduate and university students of age around 35 years have defaults whereas high school graduates of age 40 has more default compared to all.

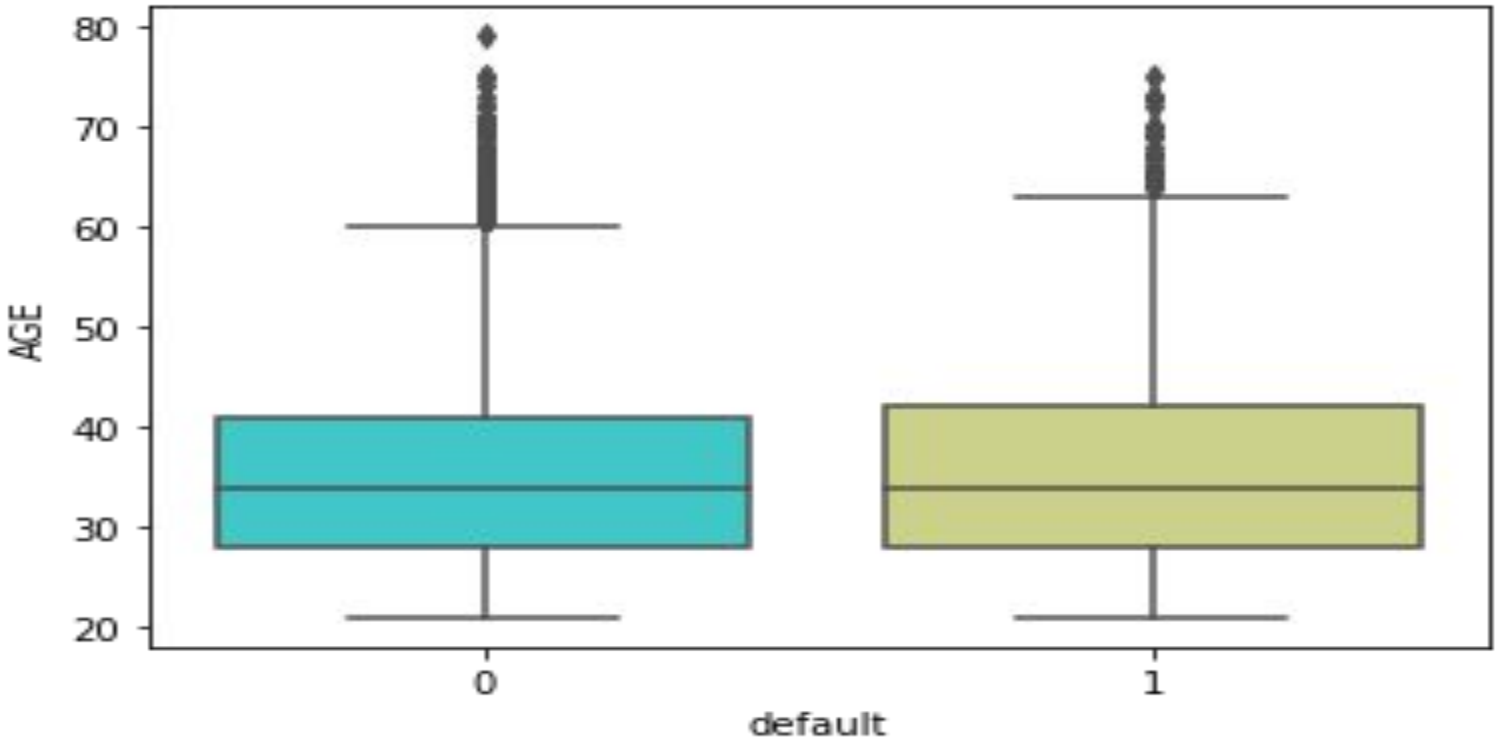




- Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

From the above plot it is clear that those people who are university students have less default payment wrt graduates and high school people.

Here we are also checking for default and no default w.r.t limit balance education wise.

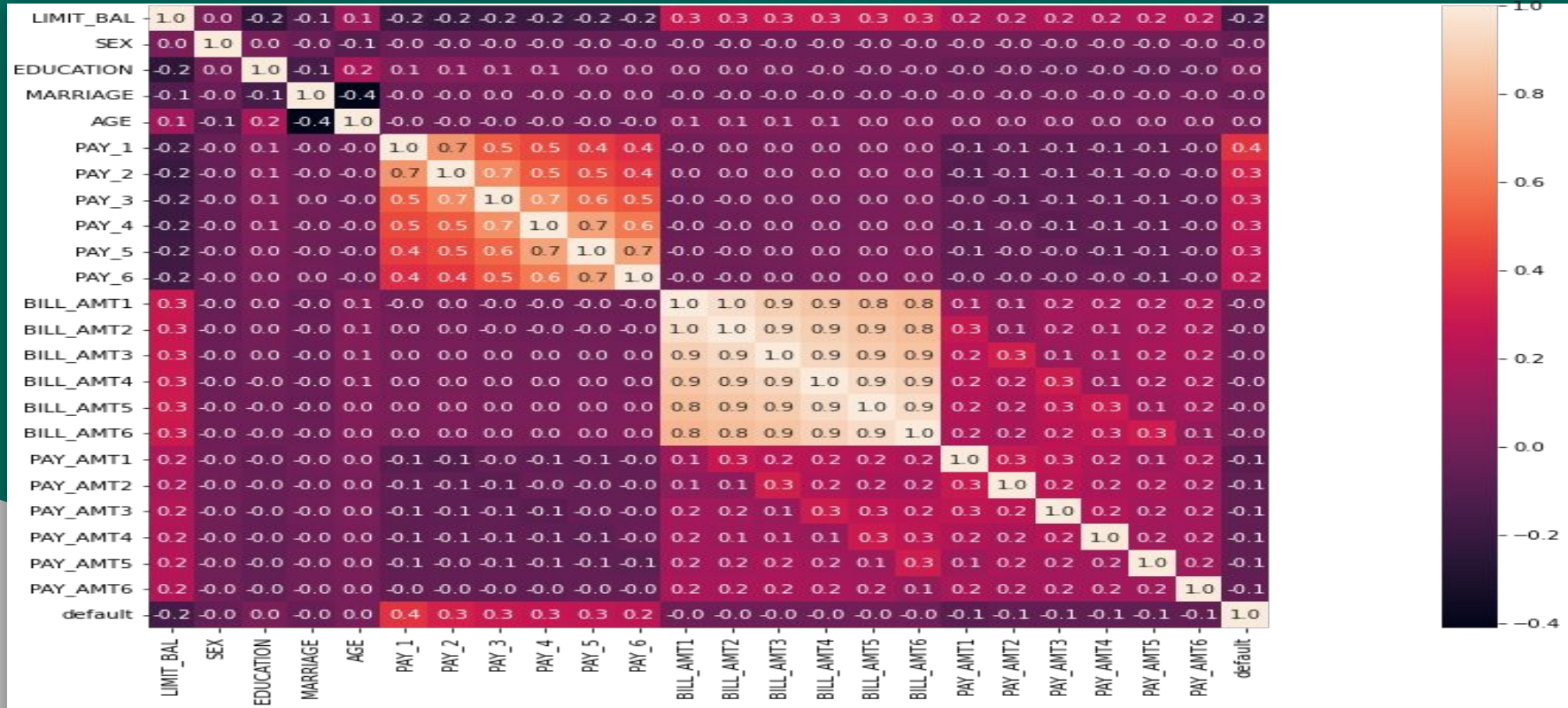


This plot shows the age group of people with default and no default,

And we can see that people around the age group of 35 tend to default.

EDA Summary

- Demographic factors that impact default risk are:
  - Education: Higher education is associated with lower default risk.
  - Age: Customers aged 30-50 have the lowest default risk.
  - Sex: Females have lower default risk than males in this dataset.
  - Credit limit: Higher credit limit is associated with lower default risk.



Correlation matrix - So it looks like the PAY_1, PAY_X variables are the strongest predictors of default, followed by the LIMIT_BAL and PAY_AMT_X variables.

Model building

The major steps involved in model building are

- Data preprocessing - Feature selection
- Feature engineering
- Train test data split
- Training data rescaling

❖ Start with default model parameters

❖ Measure roc_auc on training data

❖ Model testing and prediction

❖ Precision - recall scores

Logistic regression model

Model -Logistic Regression

Accuracy - 0.814167

Precision - 0.678457

Recall - 0.315632

F1 Score - 0.430832

ROC - 0.636371

Random forest classification

Model - Random tree Classifier

Accuracy - 0.815167

Precision - 0.661473

Recall - 0.349289

F1 Score - 0.457171

ROC - 0.649017

XGboost classifier

Model - XGBOOST Classifier

Accuracy - 0.819667

Precision - 0.686131

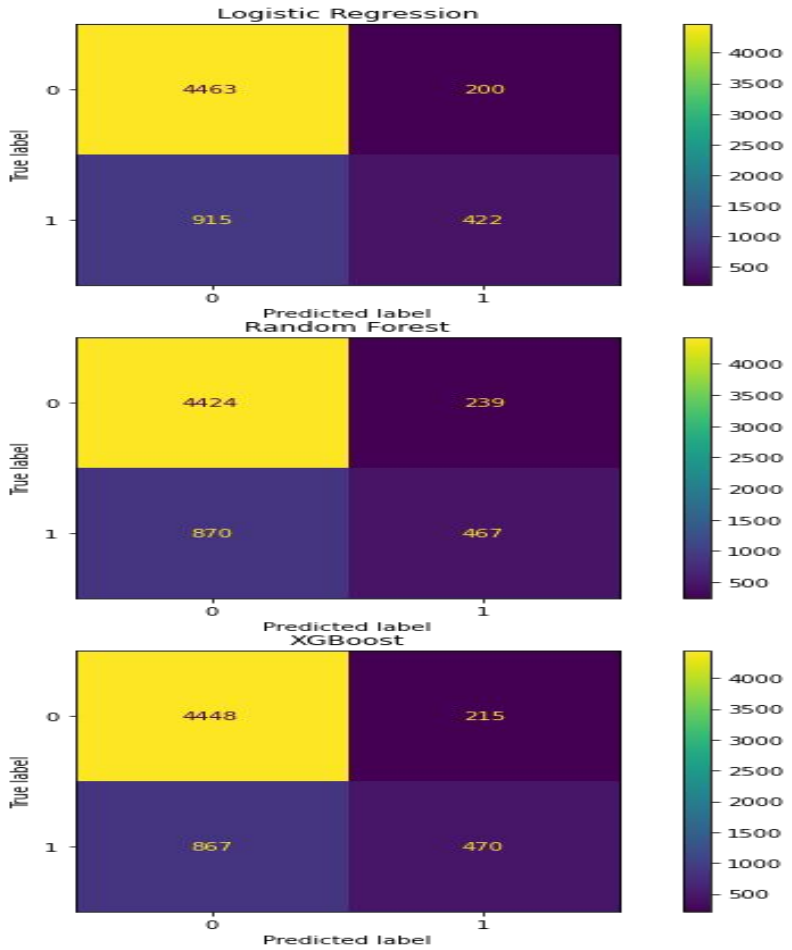
Recall - 0.351533

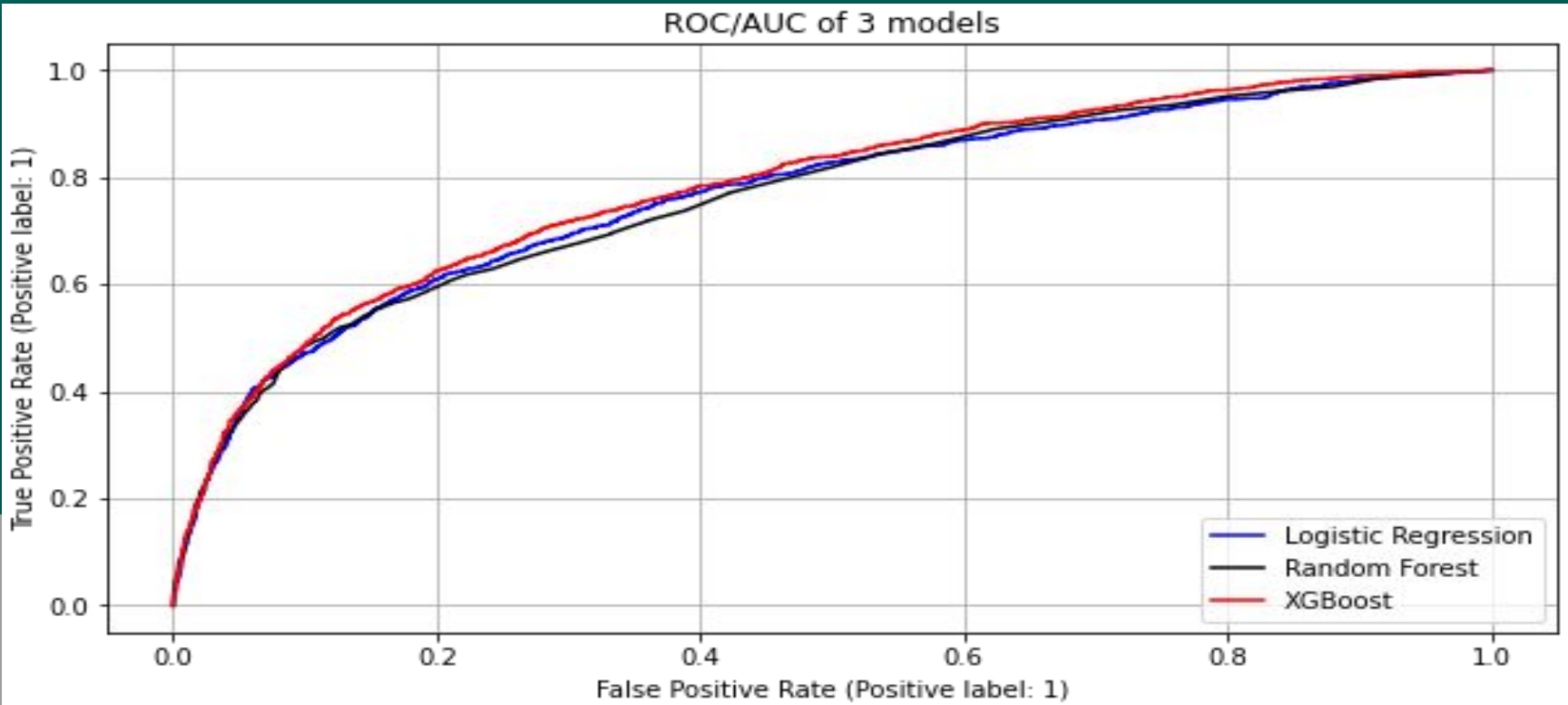
F1 Score - 0.464886

ROC - 0.652713

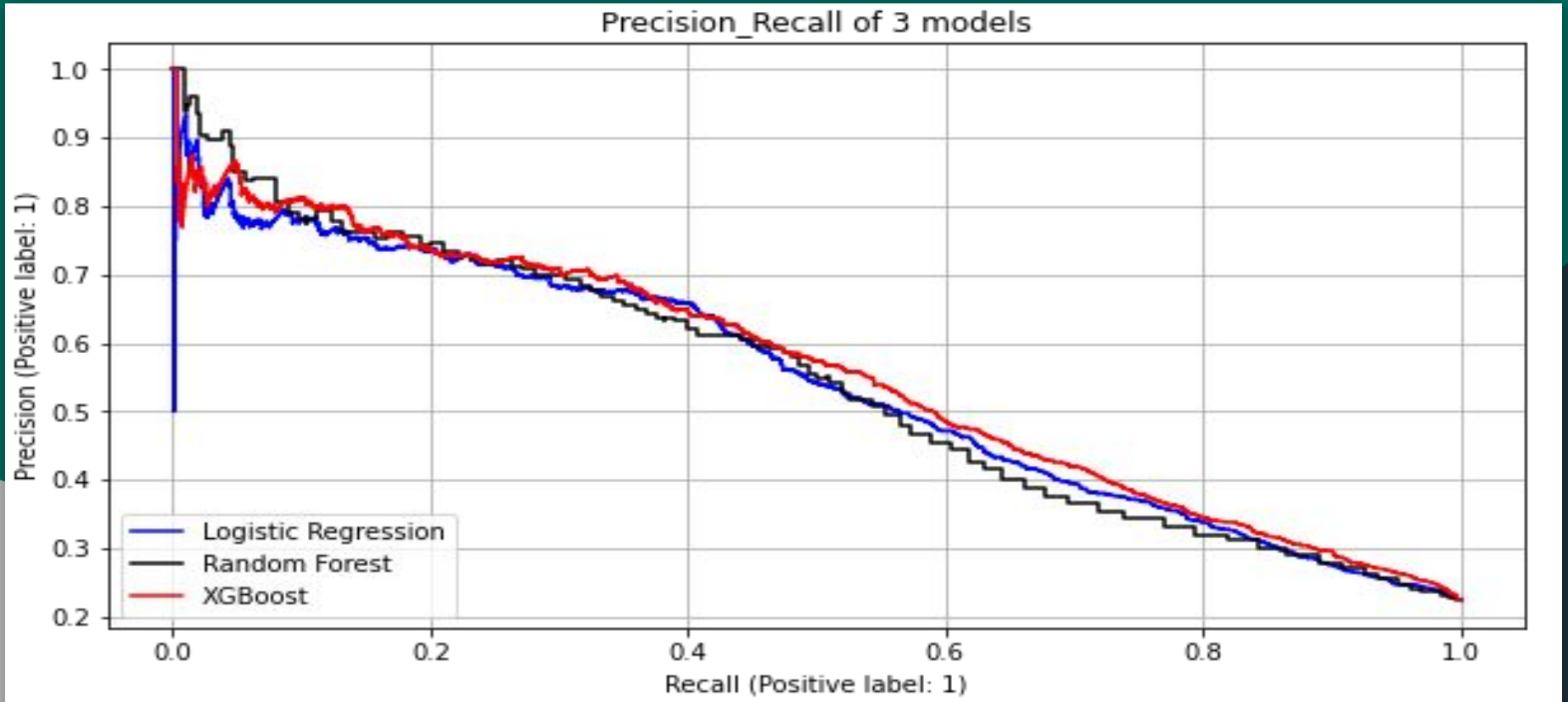
Model comparisons

This plot shows the confusion matrix for all three models.





This plot shows the roc - auc curve for the three models.



The precision_recall curve for all three models have been drawn

Conclusion

1)Using a Logistic Regression classifier, we can predict with 81.4% accuracy, whether a customer is likely to default next month.

2)Using a random forest classifier, we can predict with 81.5% accuracy, whether a customer is likely to default next month.

3)Using a XG_boost classifier, we can predict with 81.9% accuracy, whether a customer is likely to default next month.

Conclusion

- ★ Recent two payments and credit limit are the strongest default predictors
- ★ Dormant customers can also have default risk
- ★ Further modulation of parameters would increase the accuracy of the model
- ★ Model can be improved with more data and computational resources

Thank you