Capstone Project 3

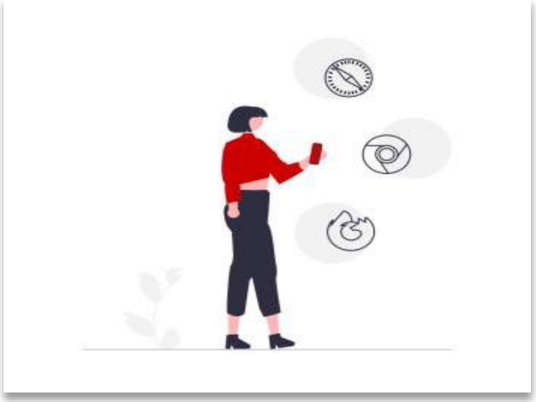
CREDIT CARD DEFAULT PREDICTION

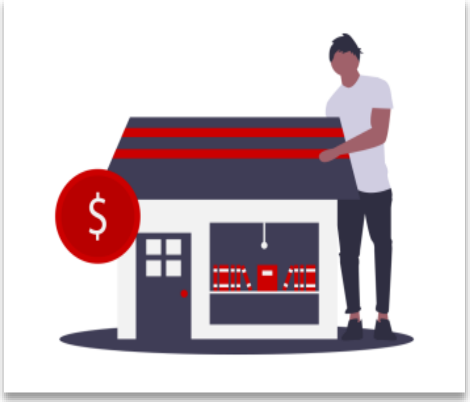
Akash Salmuthe

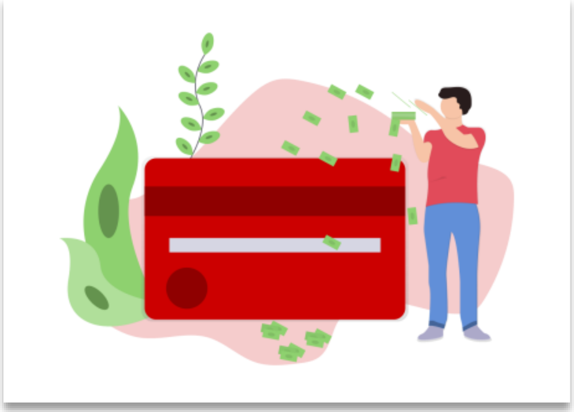
Content

Introduction

How Credit card Works







The credit card is good option until the customer repay on time. But when the customer spends more than his earning limit and unable to pay the loan. The credit default happens.

Problem Statement

- The Taiwan Credit card issuer issues credit limits to the customer and in that there will be defaulters and non-defaulters. Based on the limit the issuer provided, Age, Education, Gender and other features the limit is provided.
- To evaluate which customers will default on their credit card payments.

Data Description

- **Data Set Name :** default of credit card clients.xls

- **Data Set Information:**
  Number of instances: 30,000
  Number of attributes: 25

- **Features:**

'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'

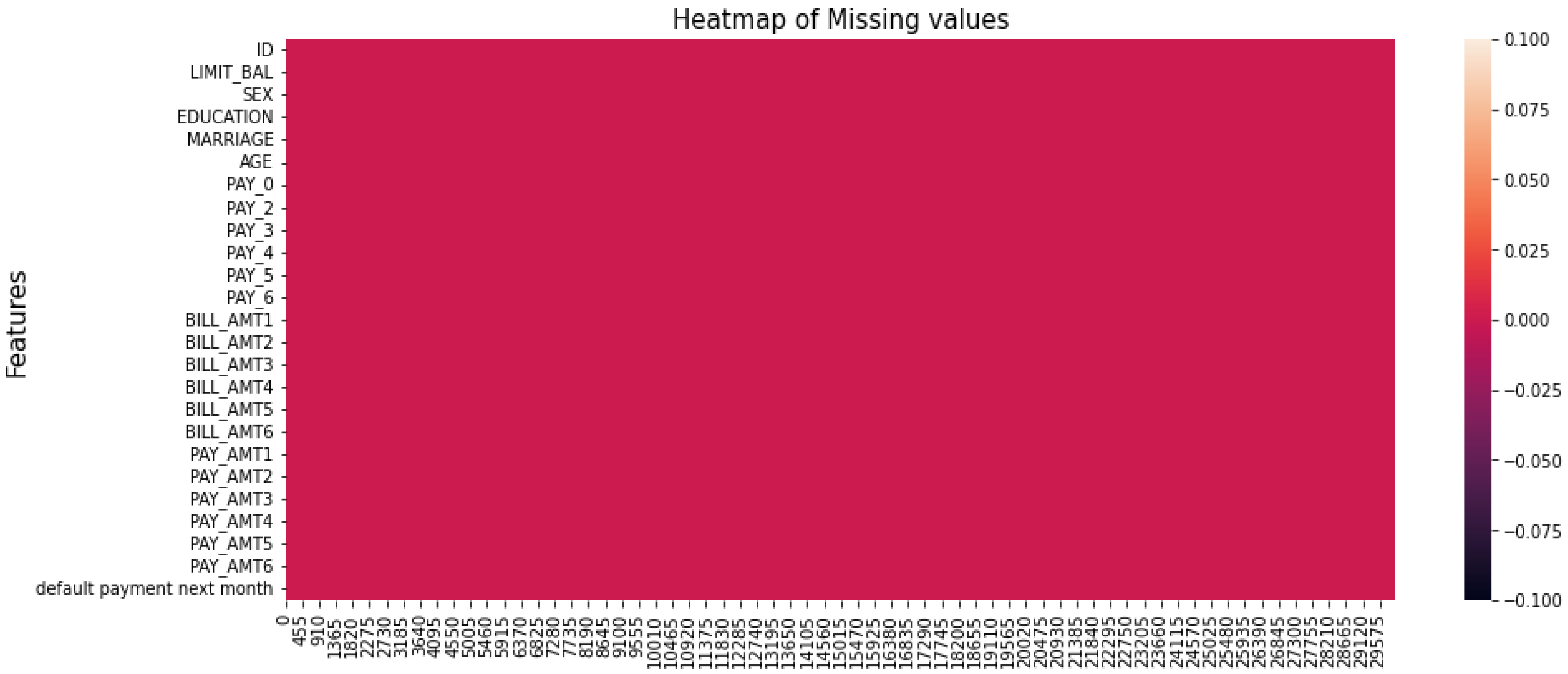
Data Summary

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

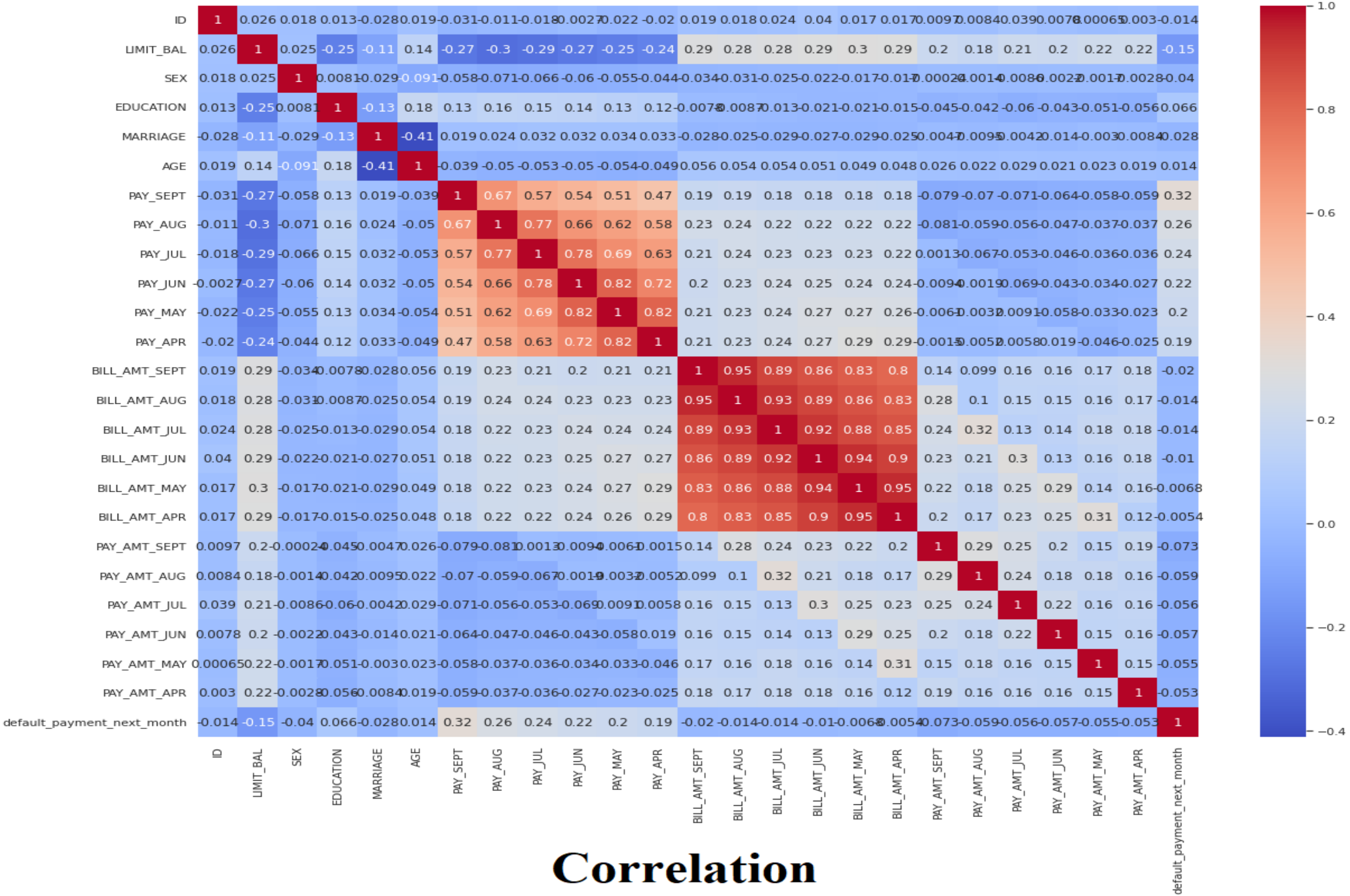
Data Cleaning

- Converting the column names to proper names
- Renaming column PAY_0 to PAY_1 and default.payment.next.month as DEFAULT
- There is no missing data in the entire dataset.
- Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns.

Missing Value



Exploratory Data Analysis

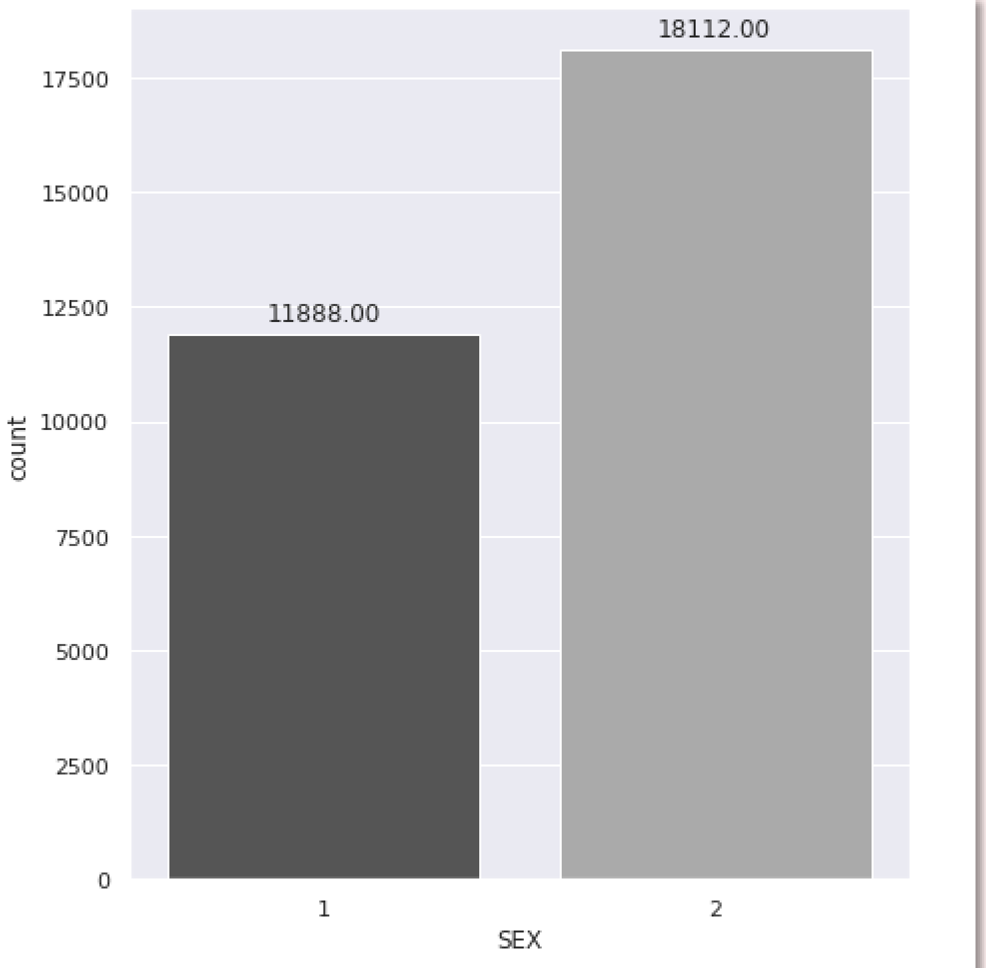


Correlation

Gender

1 : Male
2 : Female

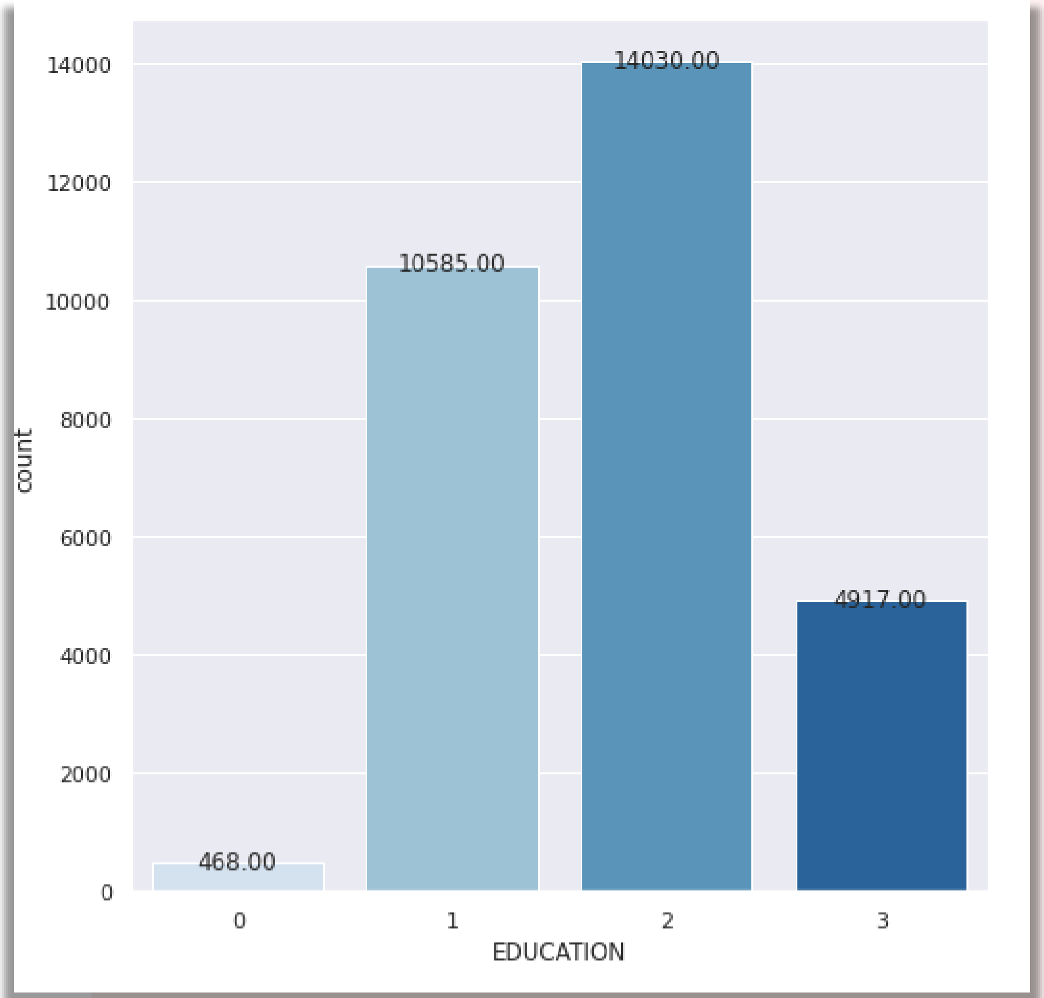
Here we can clearly see that female holds most of credit cards



Education

More number of credit holders are:

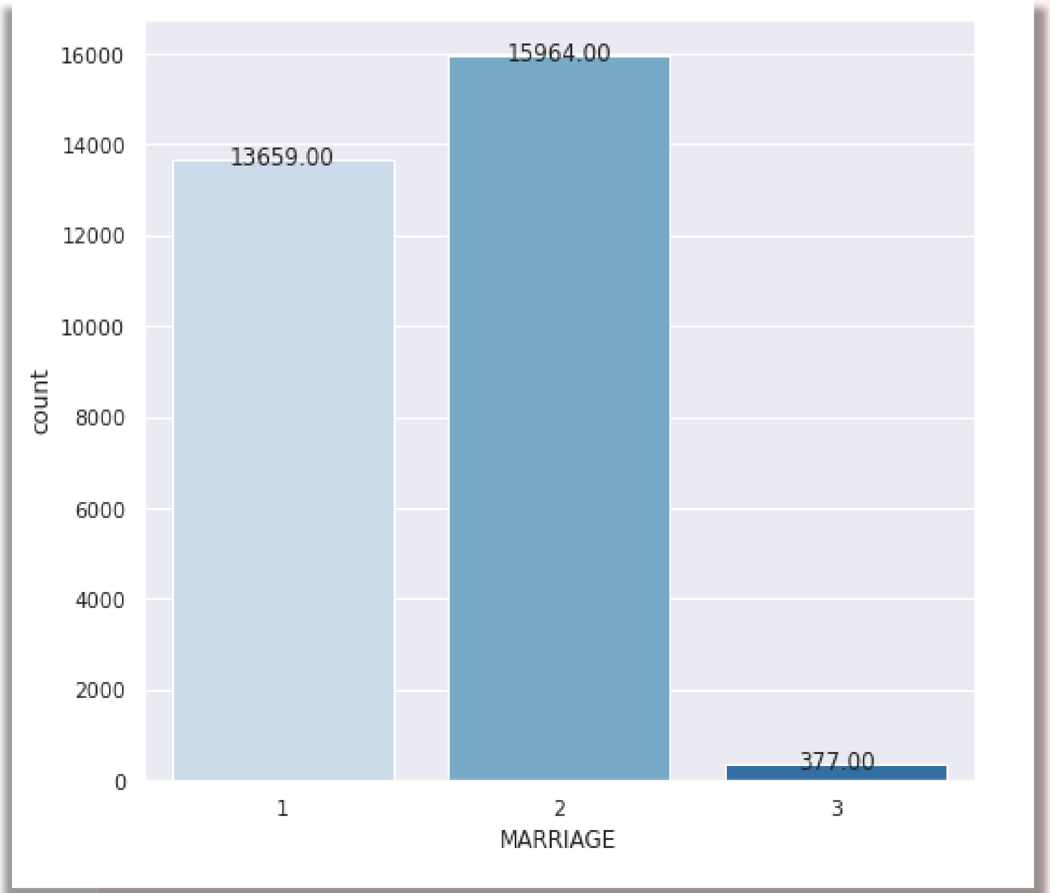
- University students (14030)
- Graduates students (10585)
- High school students (4917)
- Other (468)



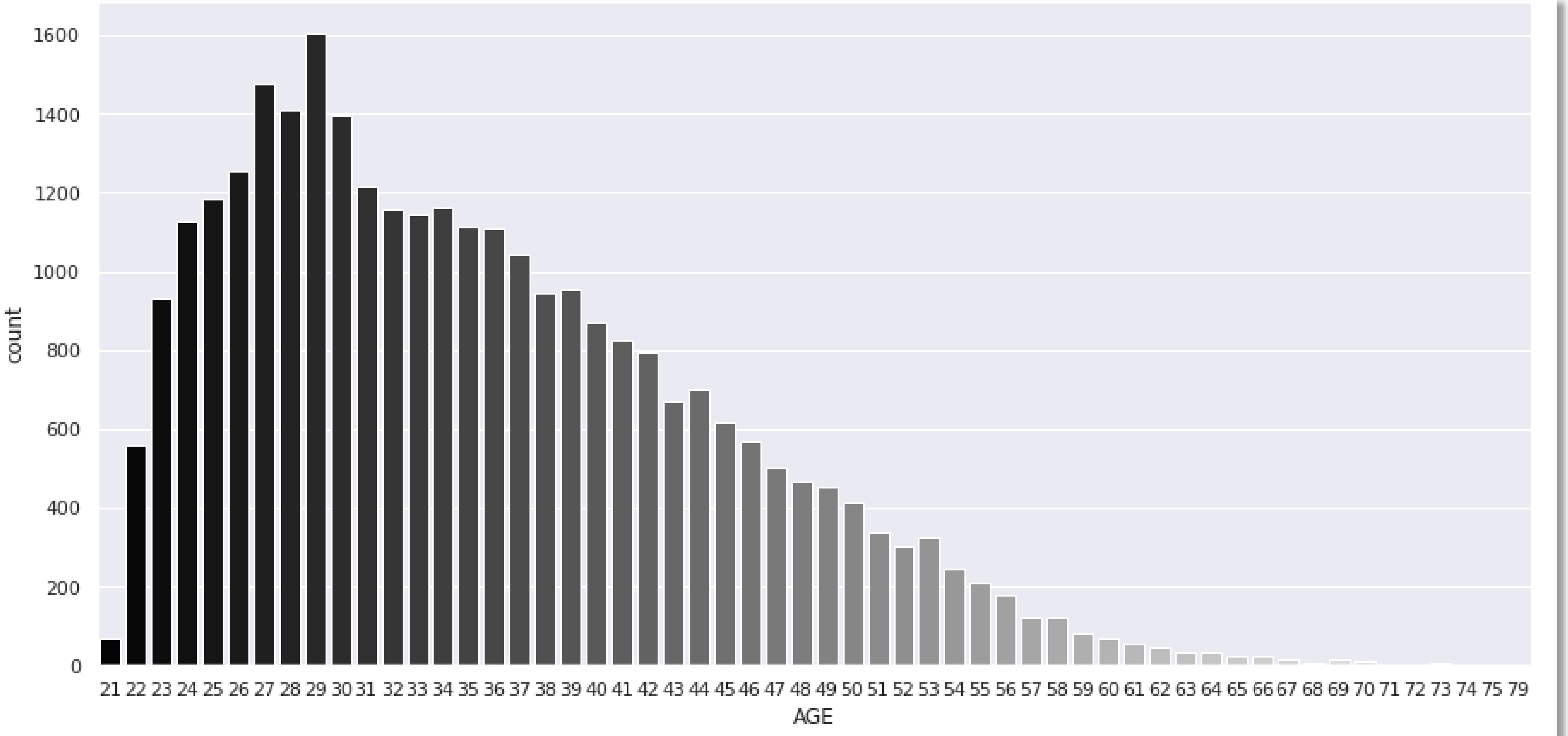
Marriage

Here,

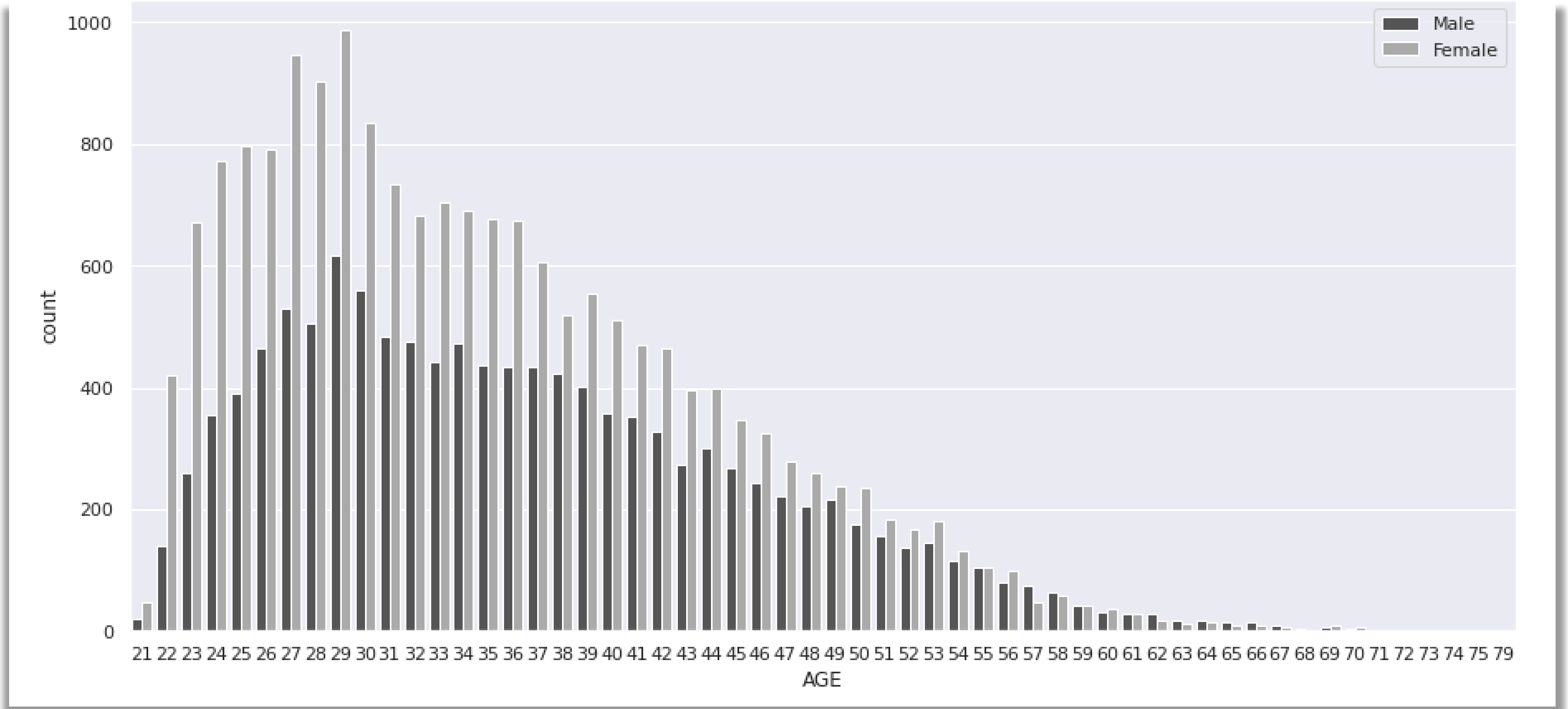
1 : Married - 13659
2 : Unmarried – 15964
3 : Others – 377 (54 + 323)



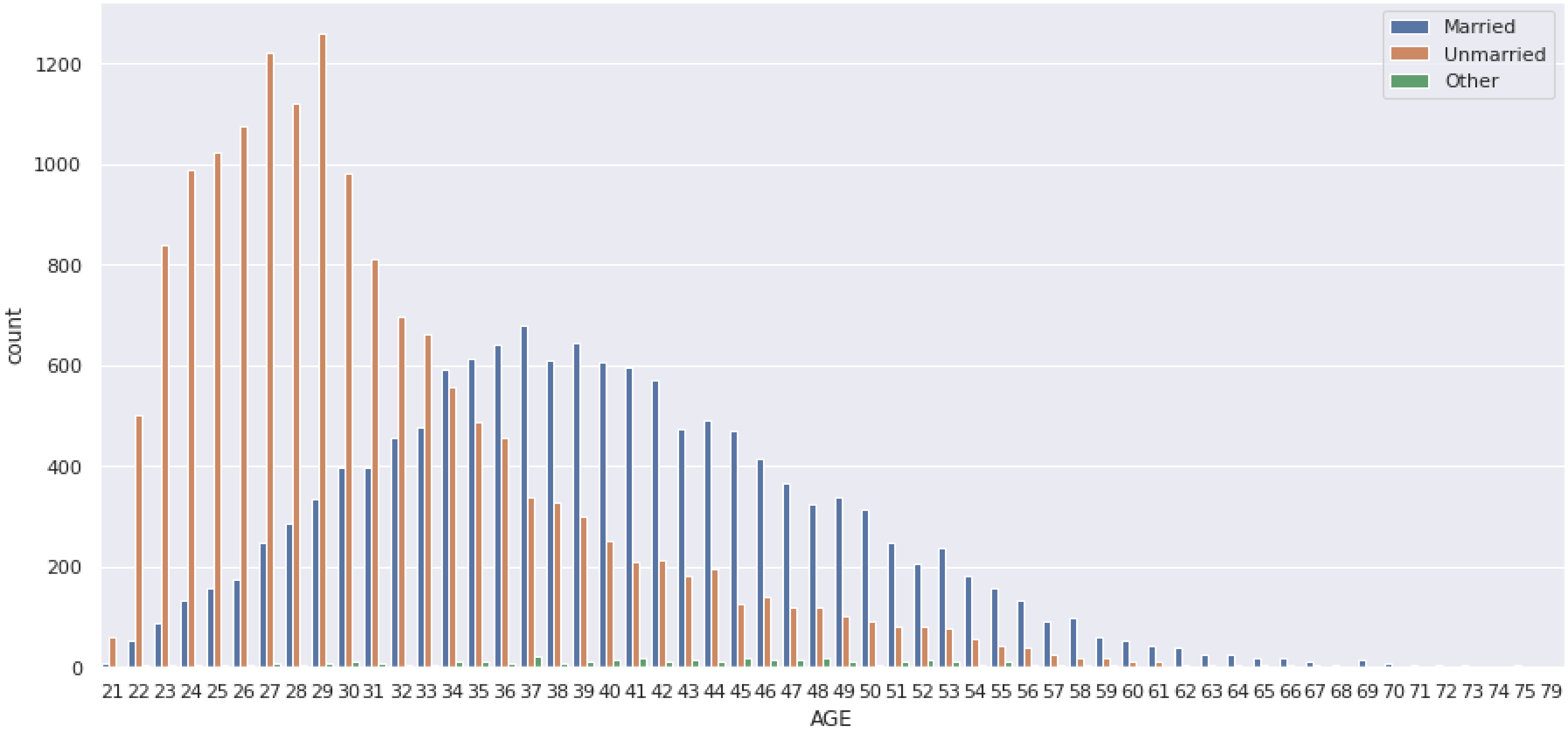
Age Distribution



Age by Gender

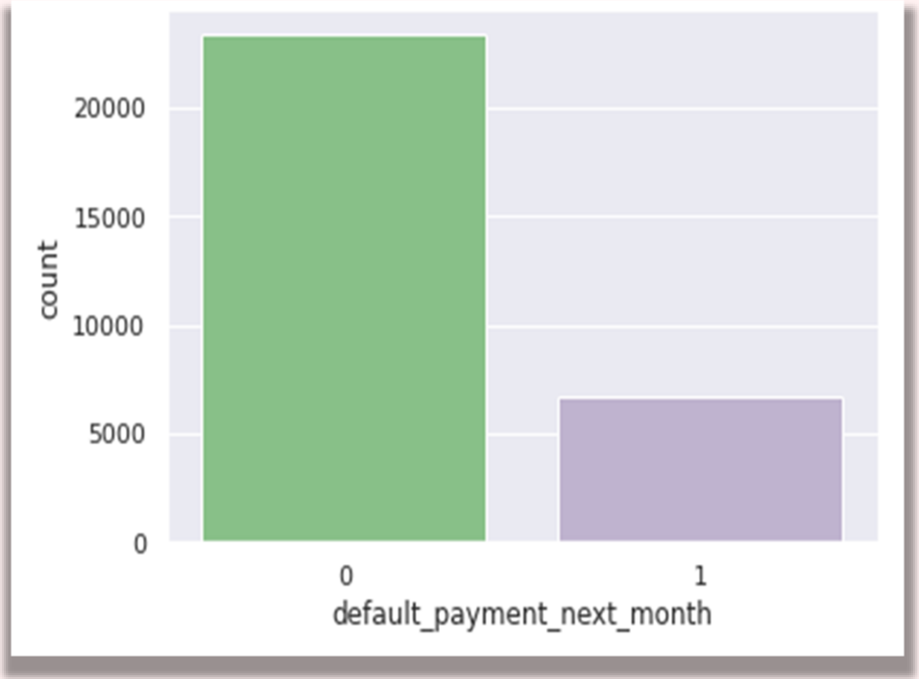


Age vs Marriage

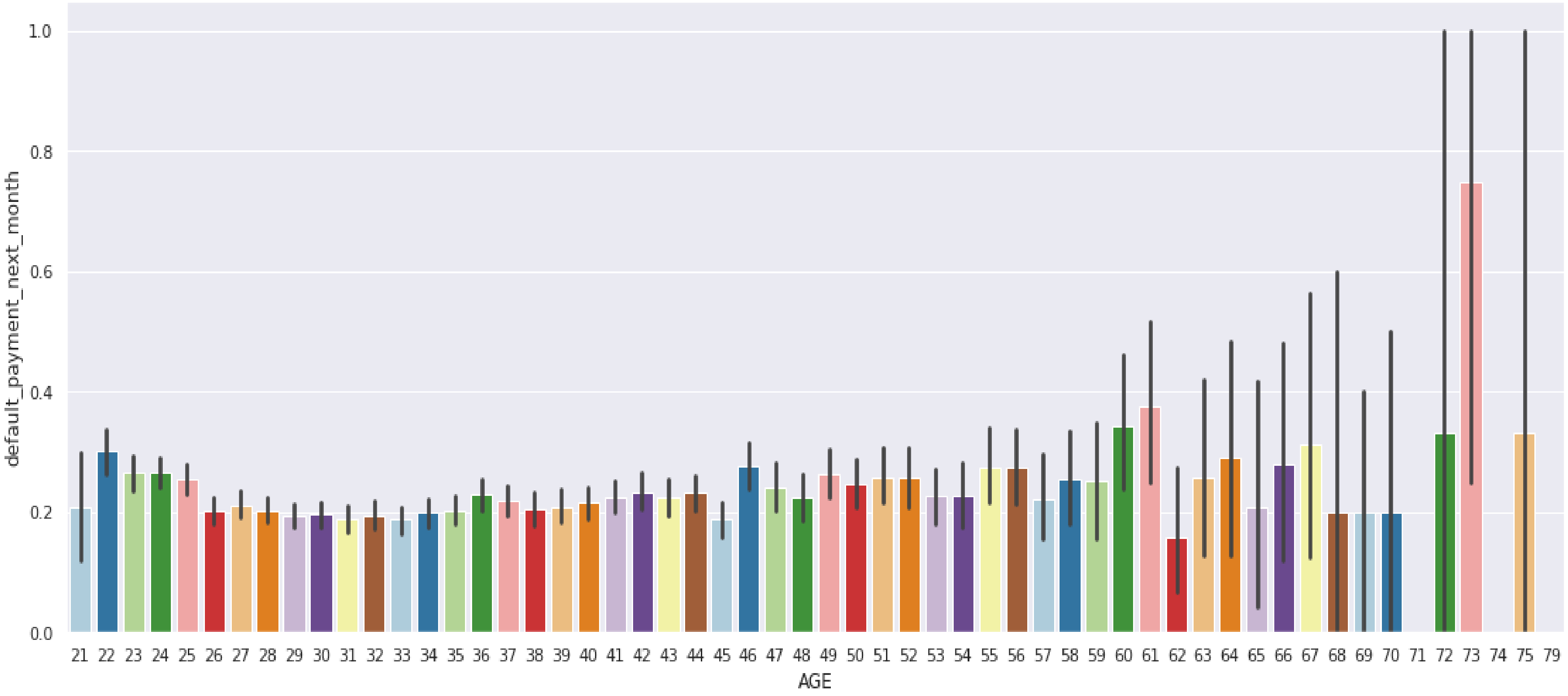


Default Next Month

Here,
0 : Not Default
1 : Default

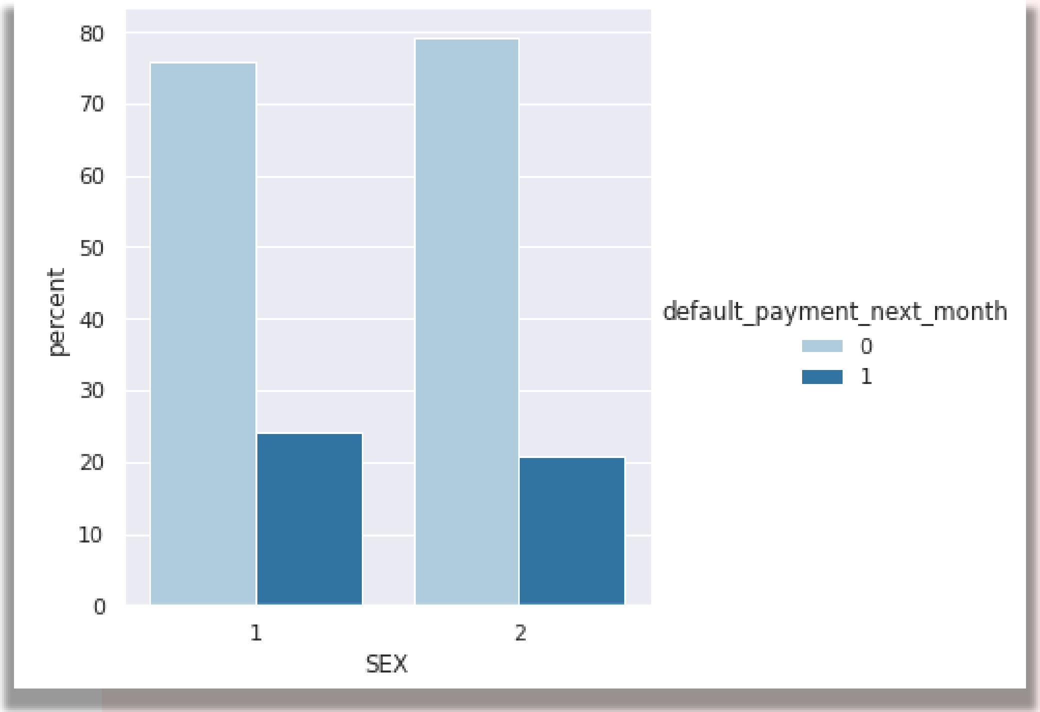


Age Vs Defaulter



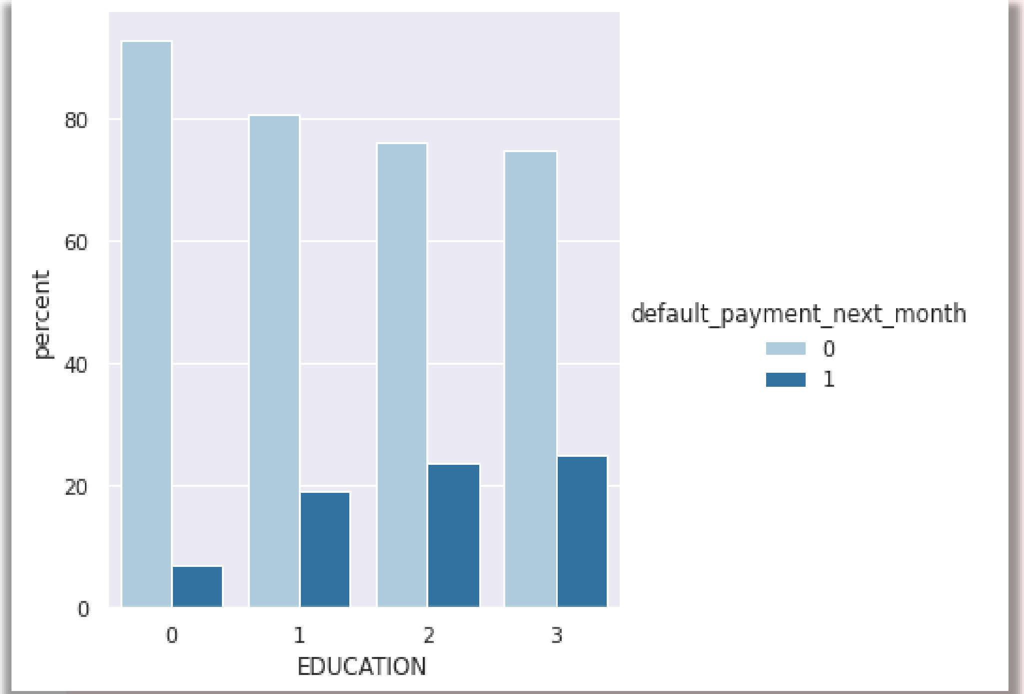
Gender Vs Defaulter

Clearly see that Male has higher default rate



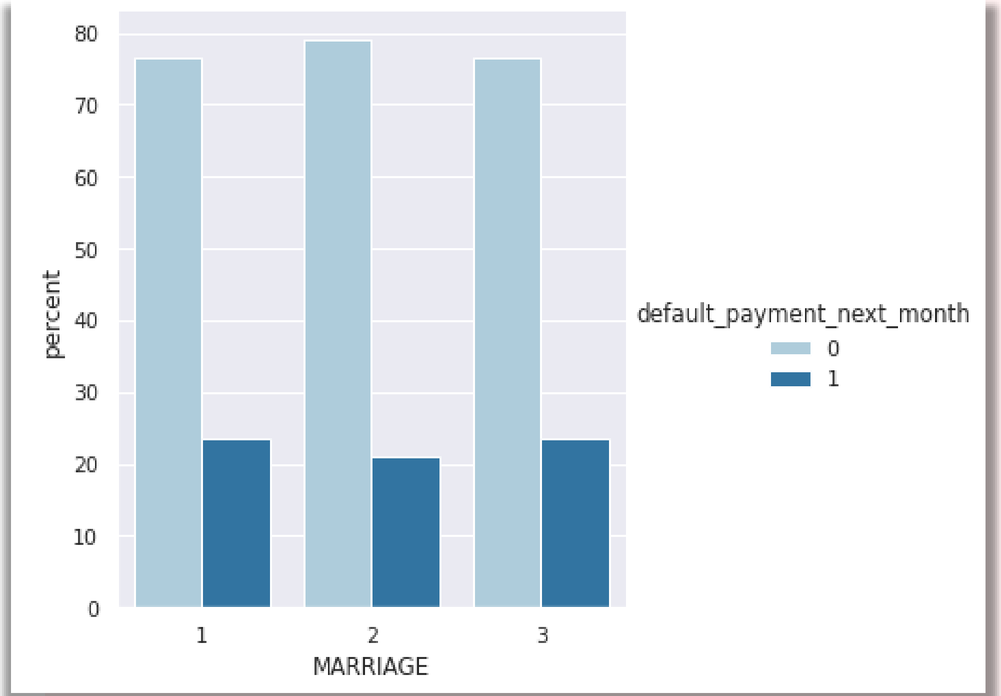
Education Vs Defaulter

Higher the education lower default risk



Marriage Vs Defaulter

High defaulter rate when it comes to others



EDA Summary

Credit card Holder

1. As per gender **Female** Holds (18112) Cards while male has (11888)

2. **University** students has (14030), **Graduates** students has (10585), **High school** students has (4917) and Other (468) Credit cards

3. **Married** :13659, **Unmarried**: 15964 Others – 377 (54 + 323)

Defaulter

1. Male have higher default rate

2. Higher Education level, lower default risk

3. **Age:** Default rate is slightly higher in **60's**

4. High defaulter rate when it comes to others

ML Models

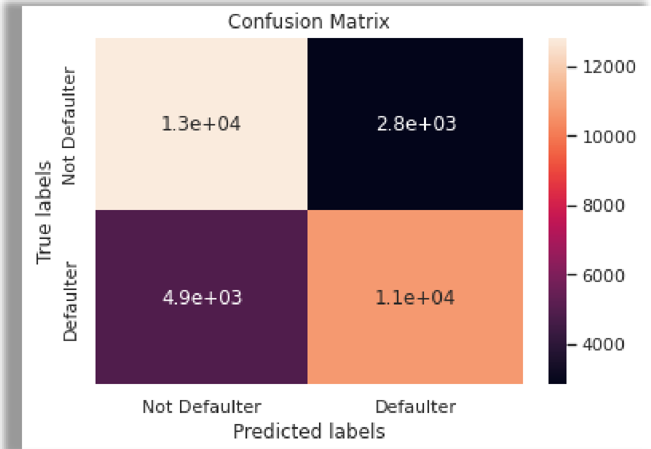
SMOTE : (Synthetic Minority Oversampling Technique)

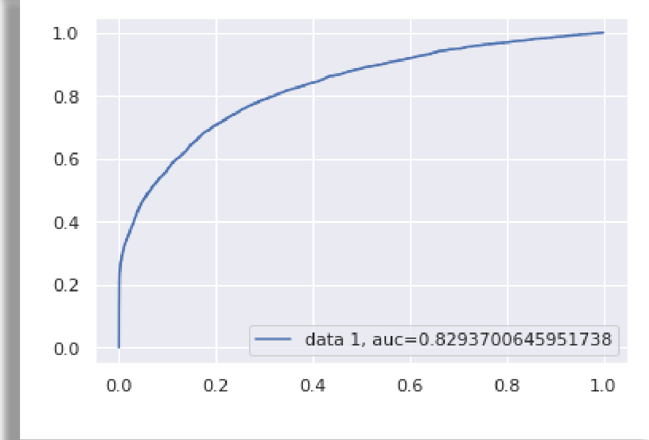
- In heatmap, Highly correlated items "PAY_SEPT","BILL_AMT_SEPT","PAY_AMT_SEP removed.

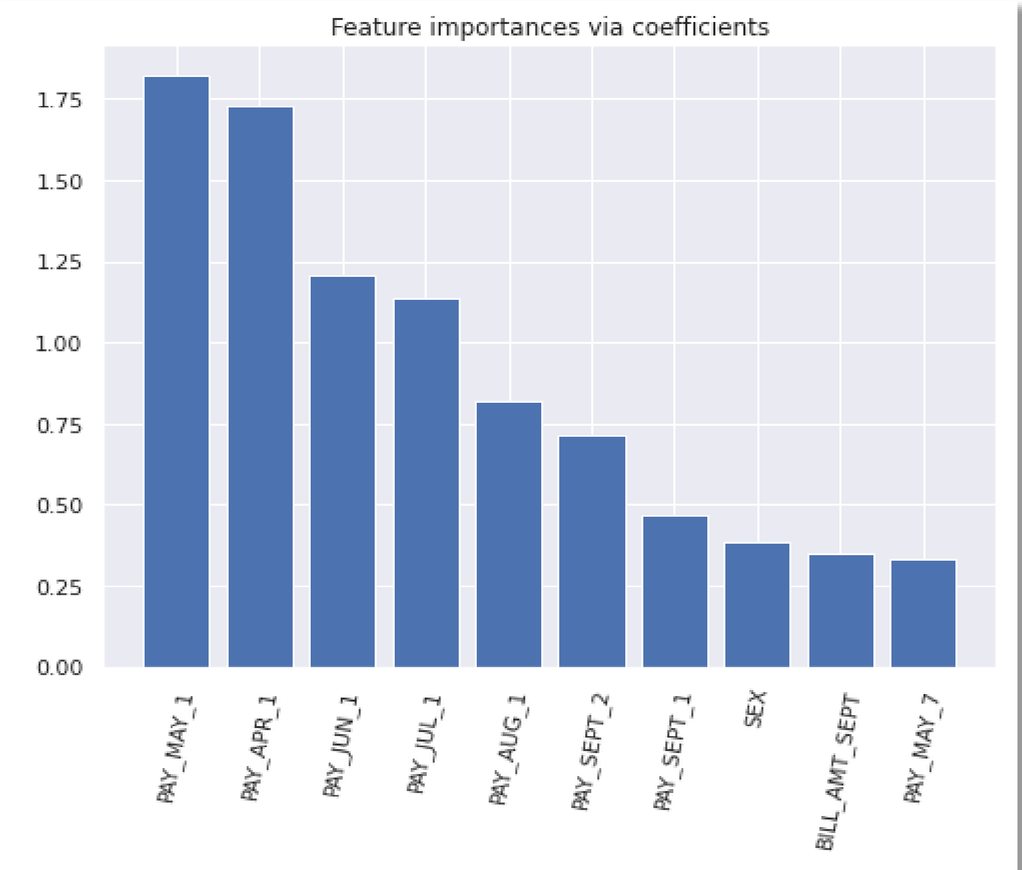
- After dataset is imbalanced dataset so we need to do the balance using SMOTE

- Original dataset shape Counter ({0: 18691, 1: 5309})
- Resample dataset shape Counter ({1: 23364, 0: 23364})
- Counter ({0: 23364, 1: 23364})

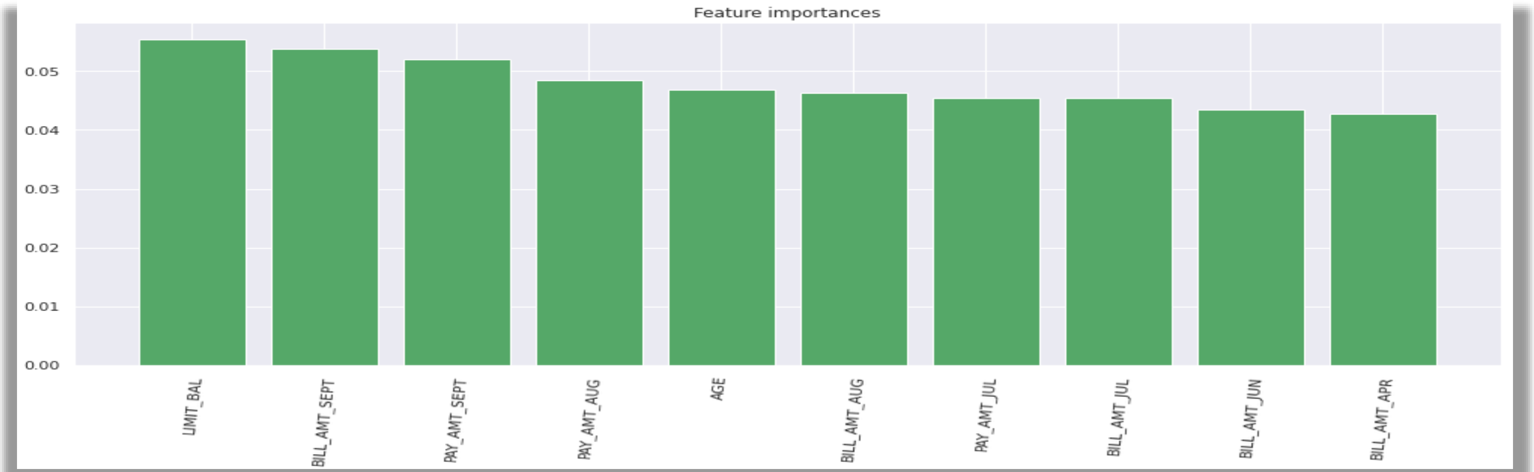
Logistic Regression

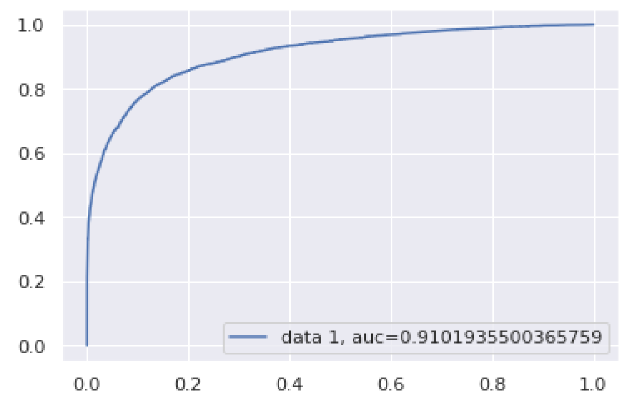


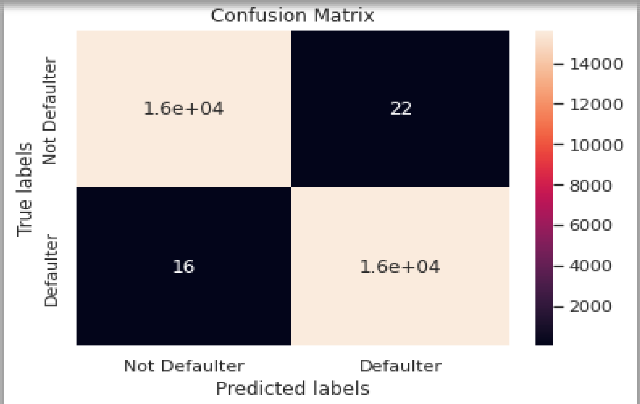




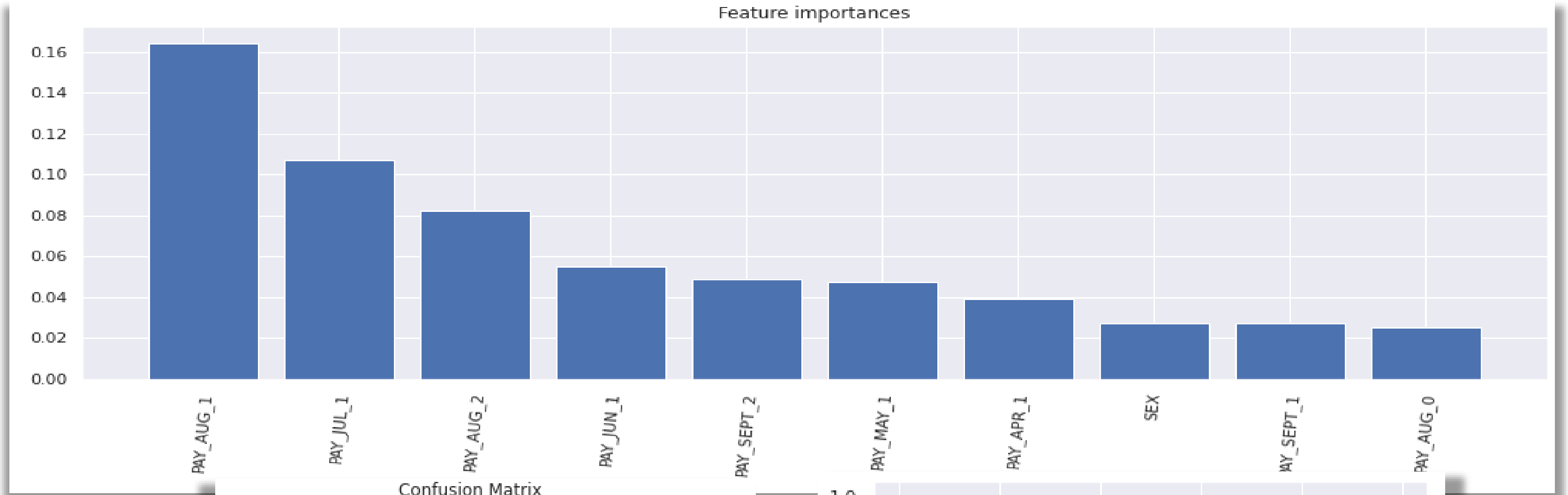
Random Forest

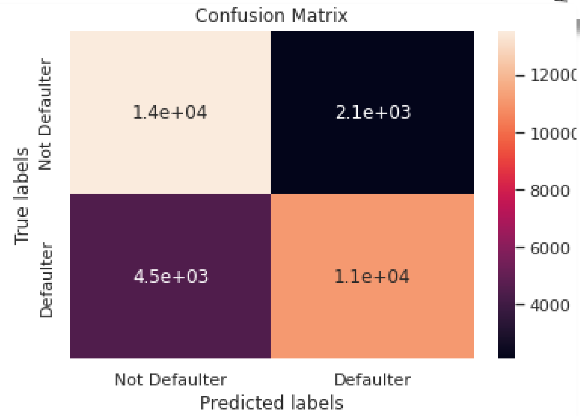


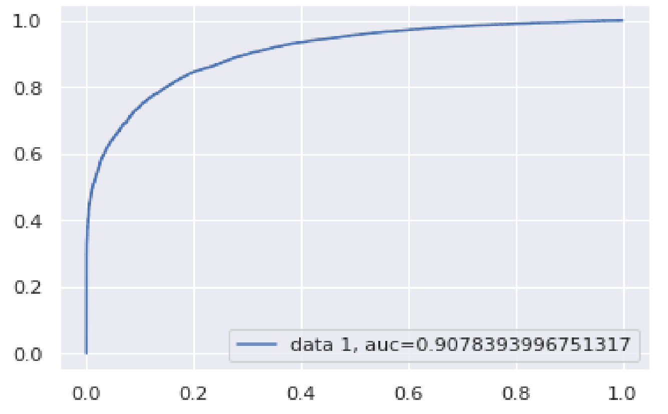




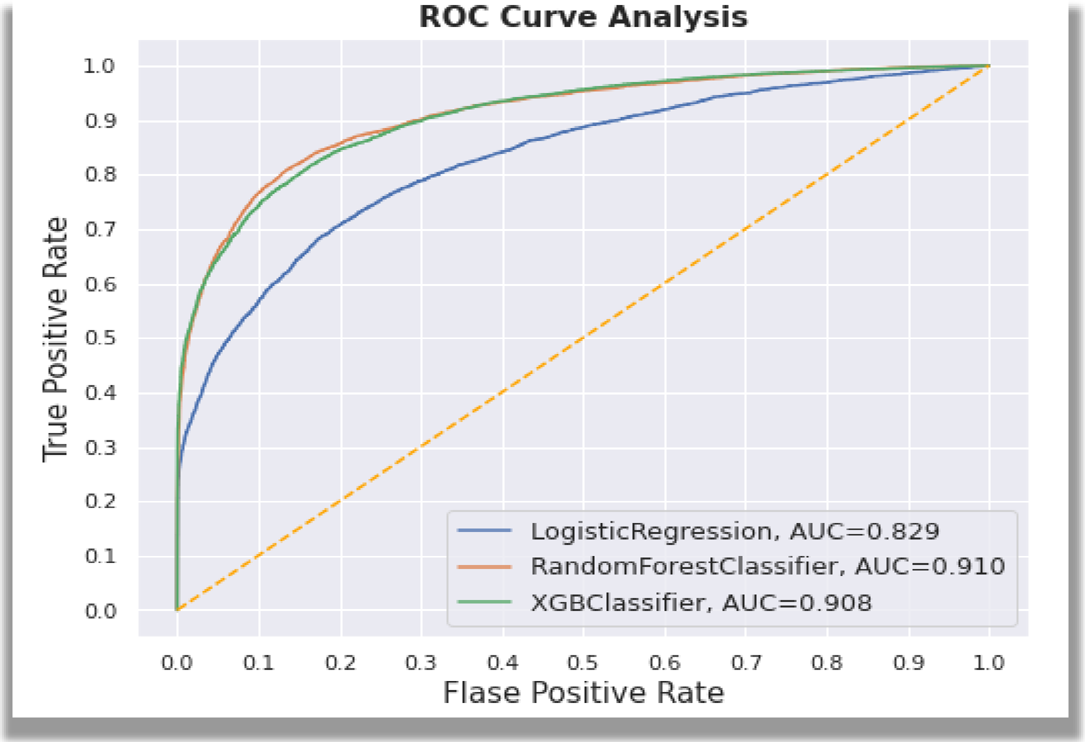
XGBoost


Feature importances
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
PAY_AUG_1
PAY_JUL_1
PAY_AUG_2
PAY_JUN_1
PAY_SEPT_2
PAY_MAY_1
PAY_APR_1
SEX
AY_SEPT_1
PAY_AUG_0

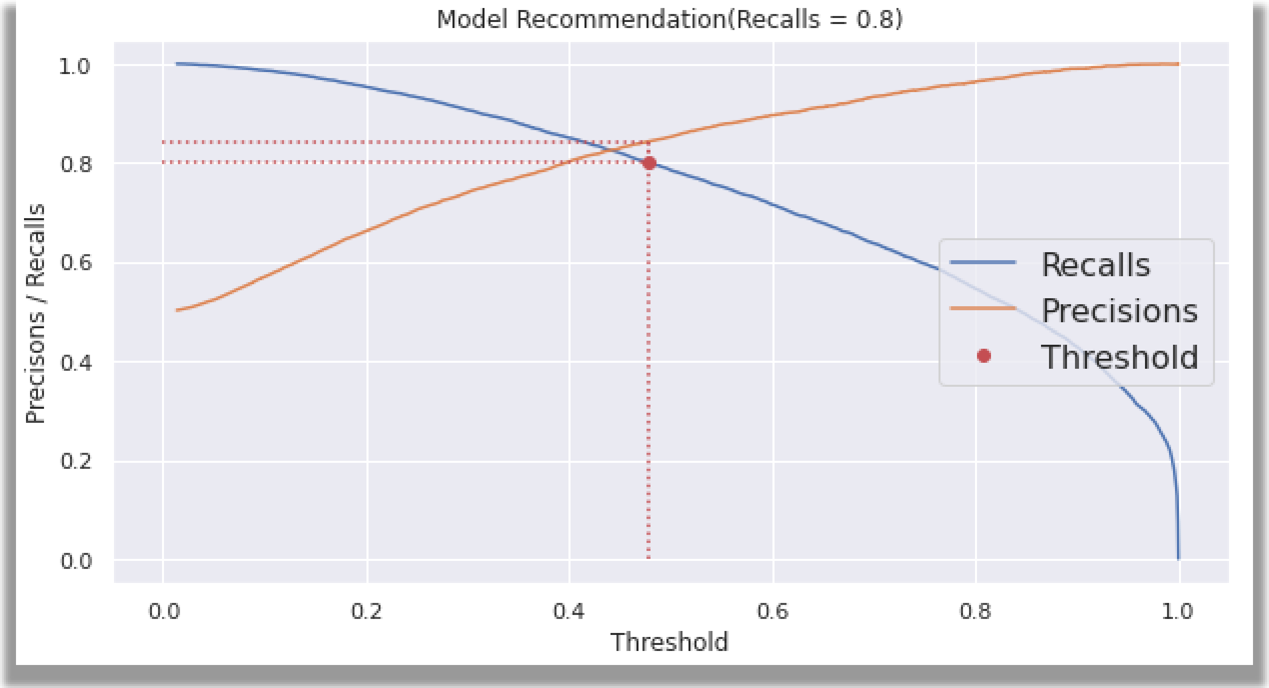

Confusion Matrix
True labels
Not Defaulter
Defaulter
1.4e+04
2.1e+03
4.5e+03
1.1e+04
12000
10000
8000
6000
4000
Not Defaulter
Defaulter
Predicted labels


1.0
0.8
0.6
0.4
0.2
0.0
data 1, auc=0.9078393996751317
0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve


ROC Curve Analysis
True Positive Rate
Flase Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
LogisticRegression, AUC=0.829
RandomForestClassifier, AUC=0.910
XGBClassifier, AUC=0.908

Threshold



Model Recommendation(Recalls = 0.8)
1.0
0.8
0.6
0.4
0.2
0.0
Precisons / Recalls
0.0
0.2
0.4
0.6
0.8
1.0
Threshold
Recalls
Precisions
Threshold


Model Performed
Train Accuracy
Test Accuracy
Precision
Recall
F1
1.2
1
0.8
0.6
0.4
0.2
0
Logistic Regression
Random Forest
XGBoost

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.
- XGBoost Classifier gave the highest accuracy of 83% on test dataset.
- The best **accuracy** is obtained for the **Random forest** and **XGBoost classifier**
- **XGBoost Classifier** giving us the best **Recall, F1-score**, and **ROC Score** among other algorithms
- We can conclude that these two algorithms are the best to predict whether the credit card is default or not