Capstone Project 3

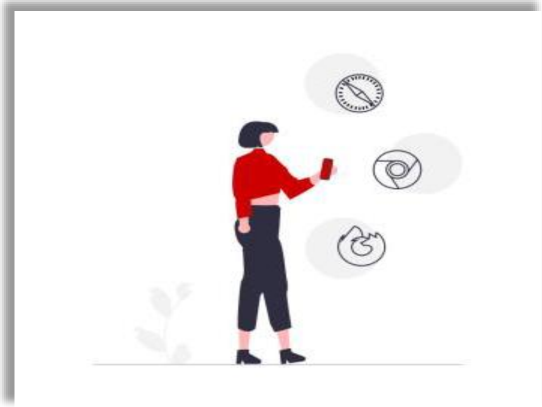
CREDIT CARD DEFAULT PREDICTION

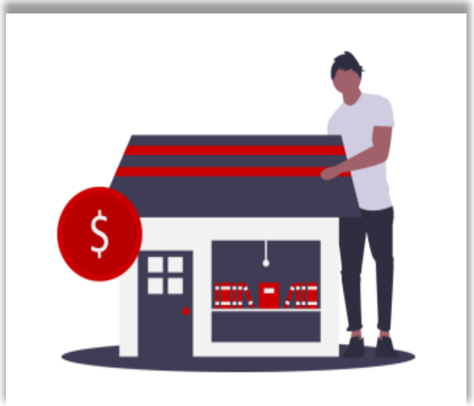
Akash Salmuthe

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 there will be defaulters and non-defaulters. Based on Age, Education, Gender and other features the limit is provided by the issuer.

- 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'

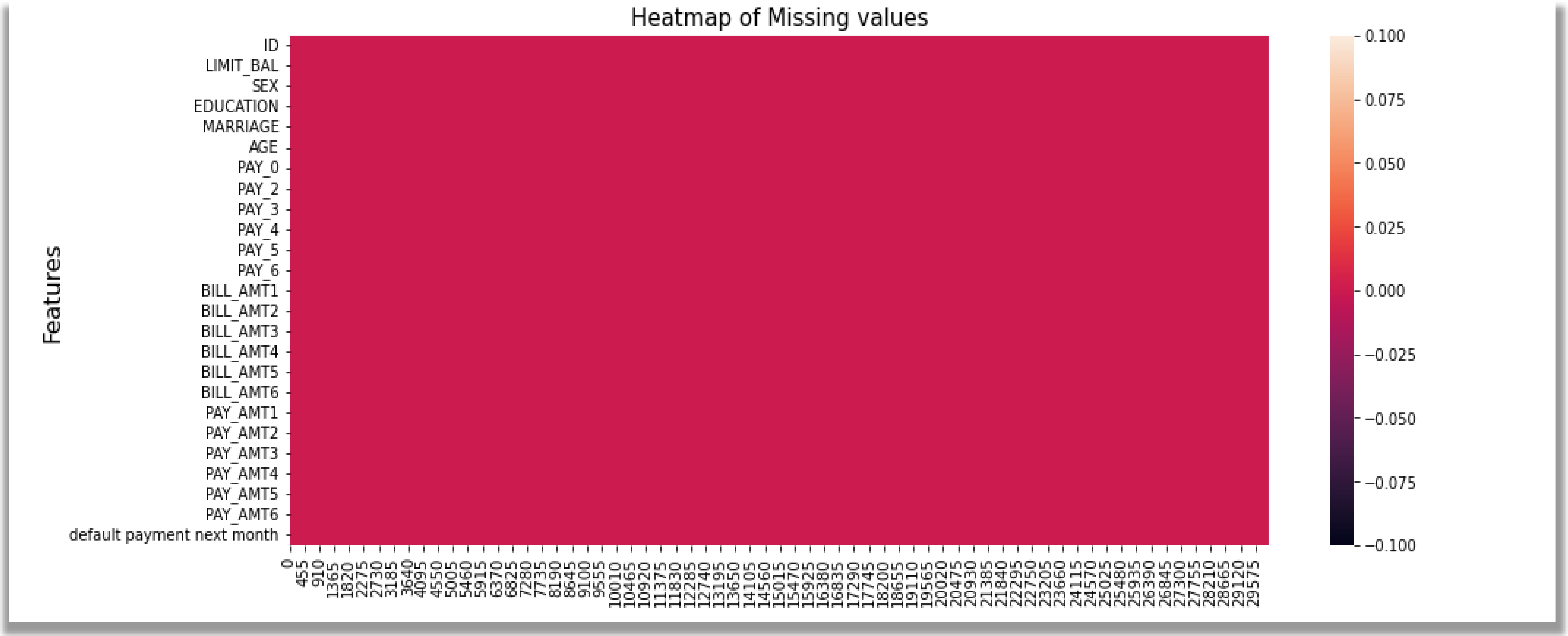
Data Summary

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

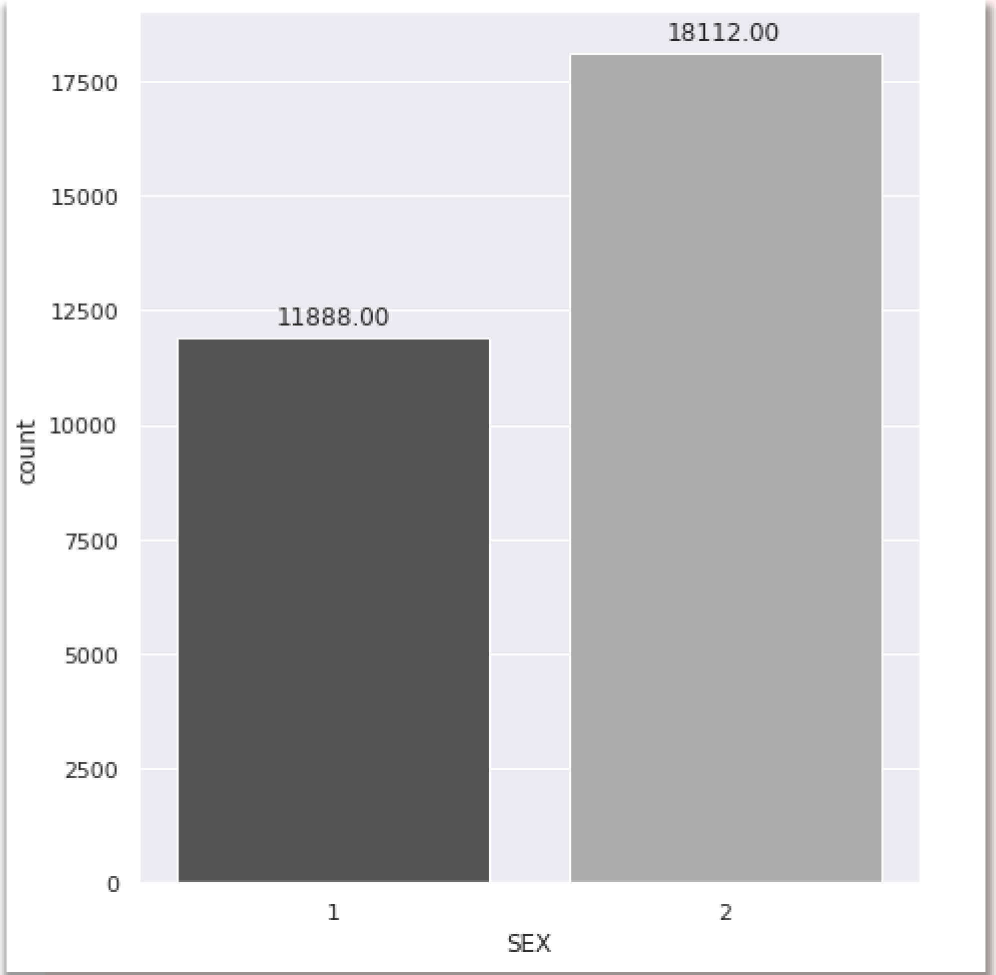


Exploratory Data Analysis

Gender

1 : Male
2 : Female

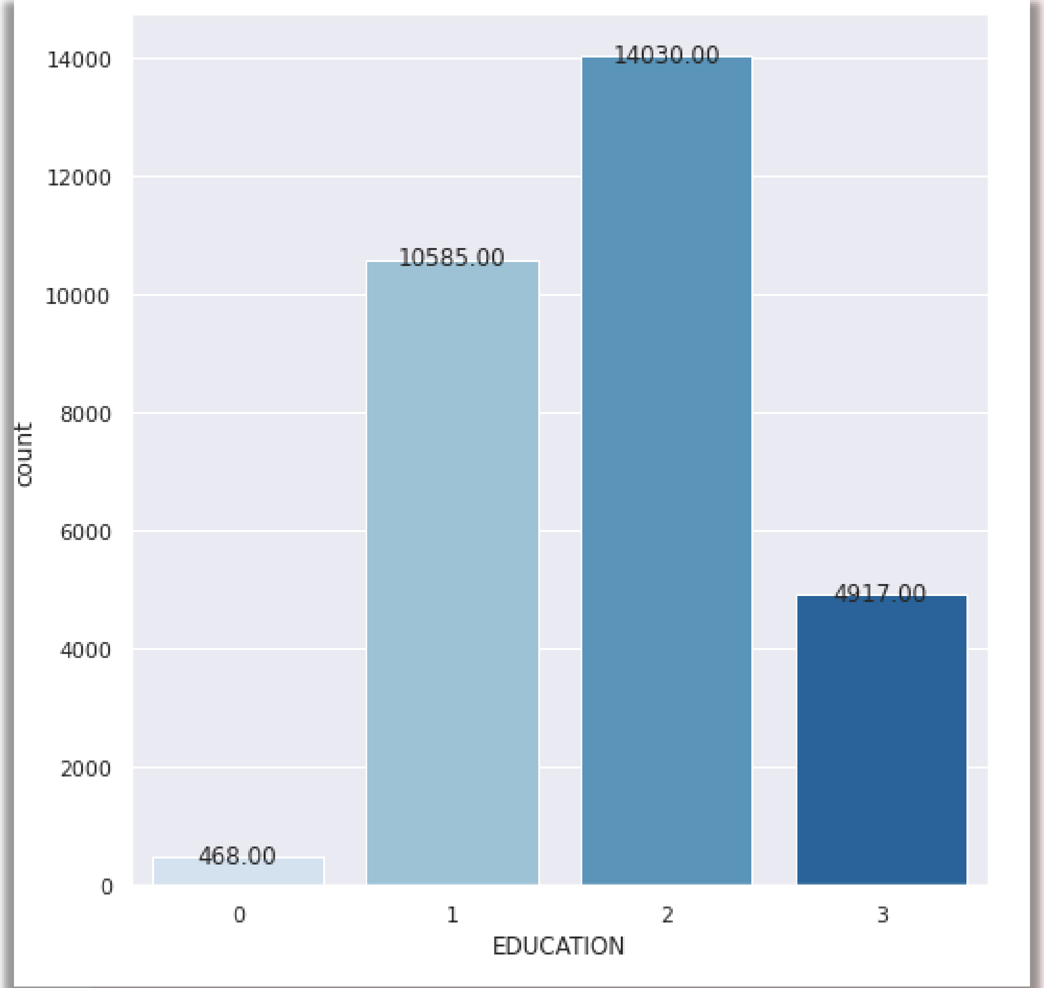
Here we can clearly see that female holds most of credit cards



Education

More number of credit holders are:

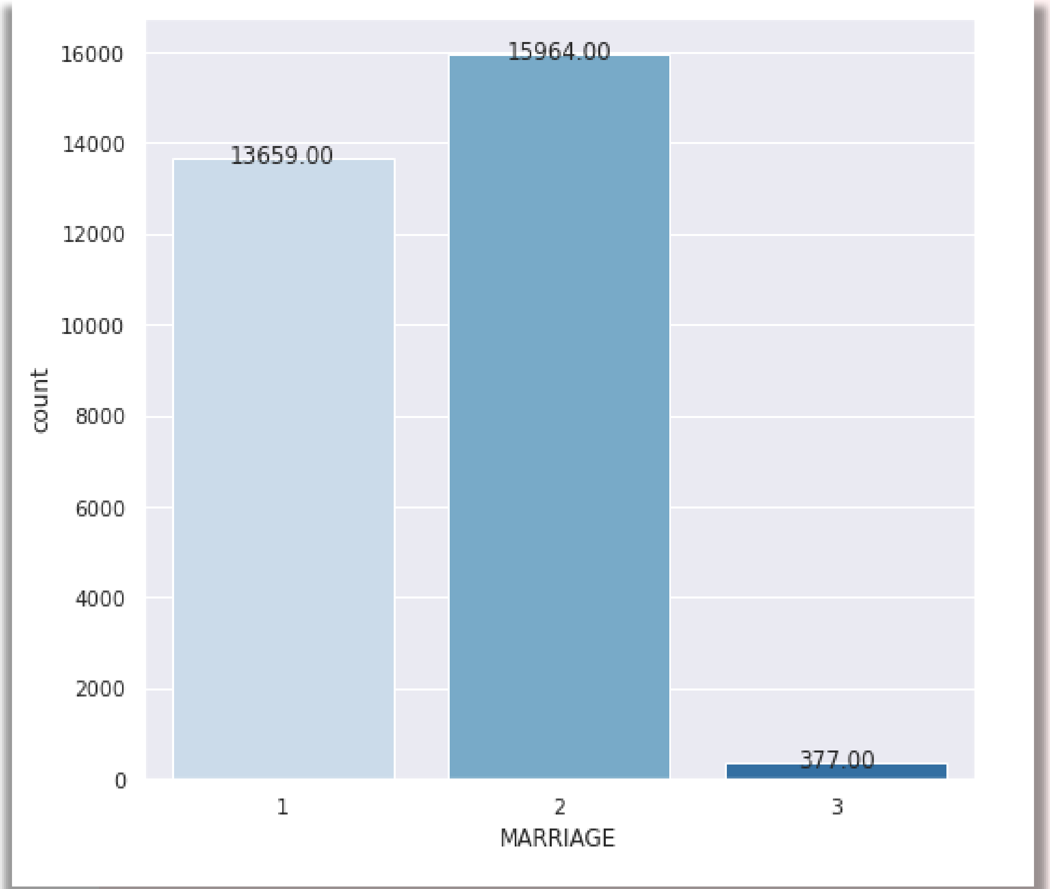
- University students (14030)
- Graduates students (10585)
- High school students (4917)
- Other (468)



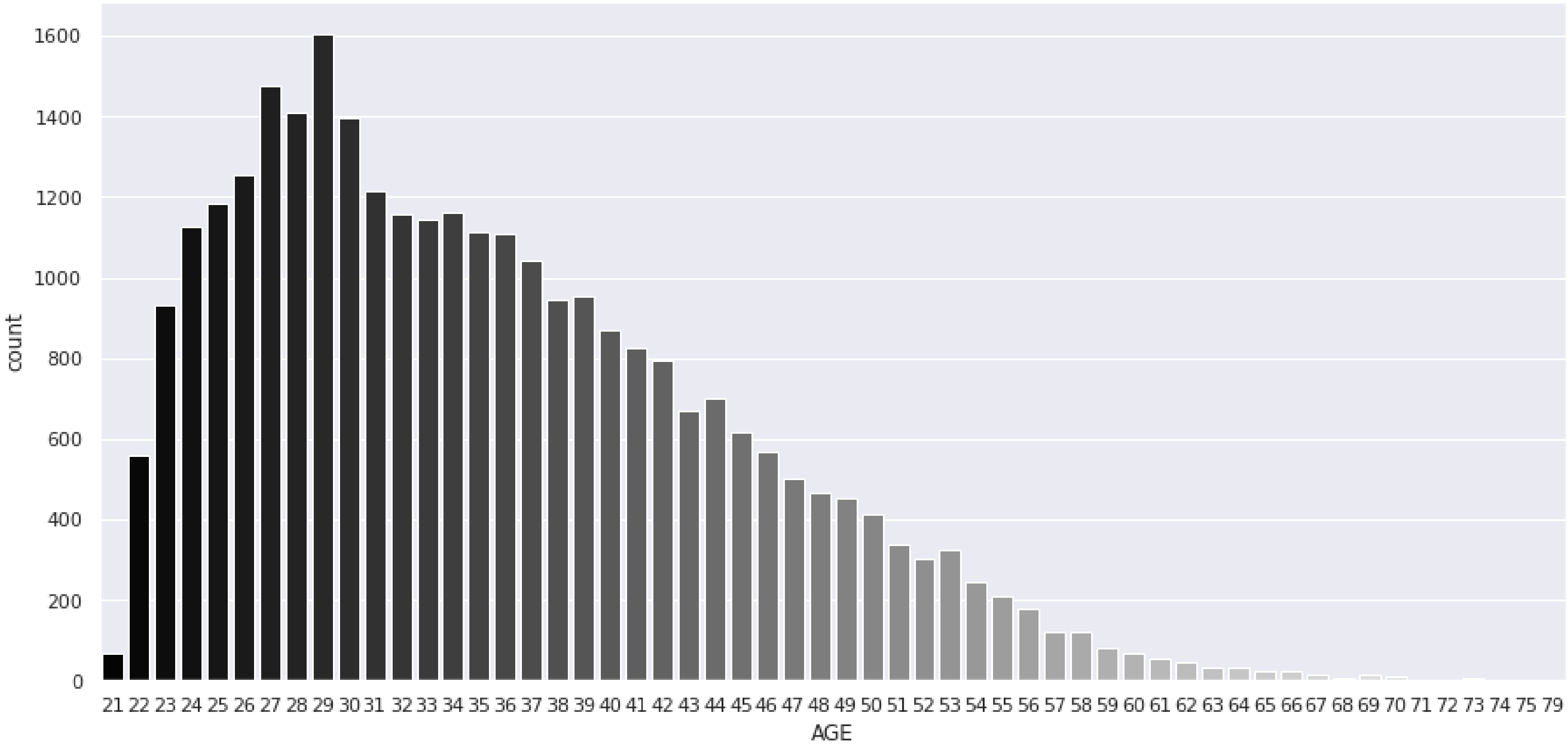
Marriage

Here,

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

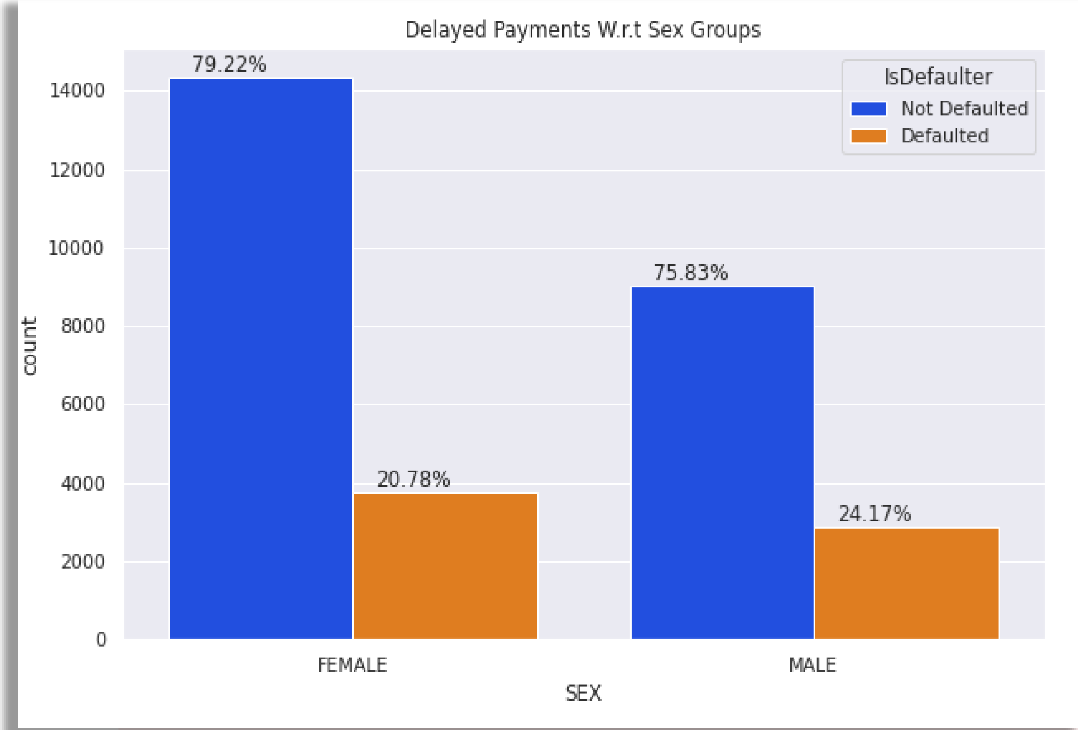


Age Distribution

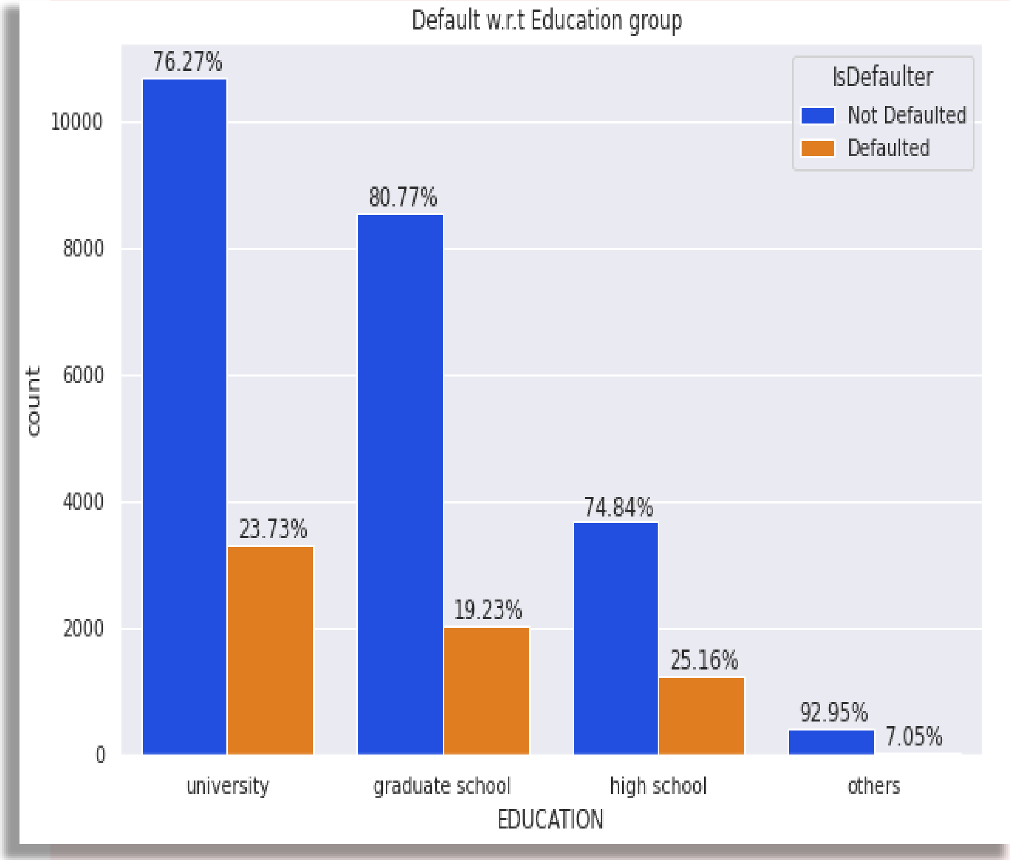


Gender Vs Defaulter

Clearly see that Male has higher default rate

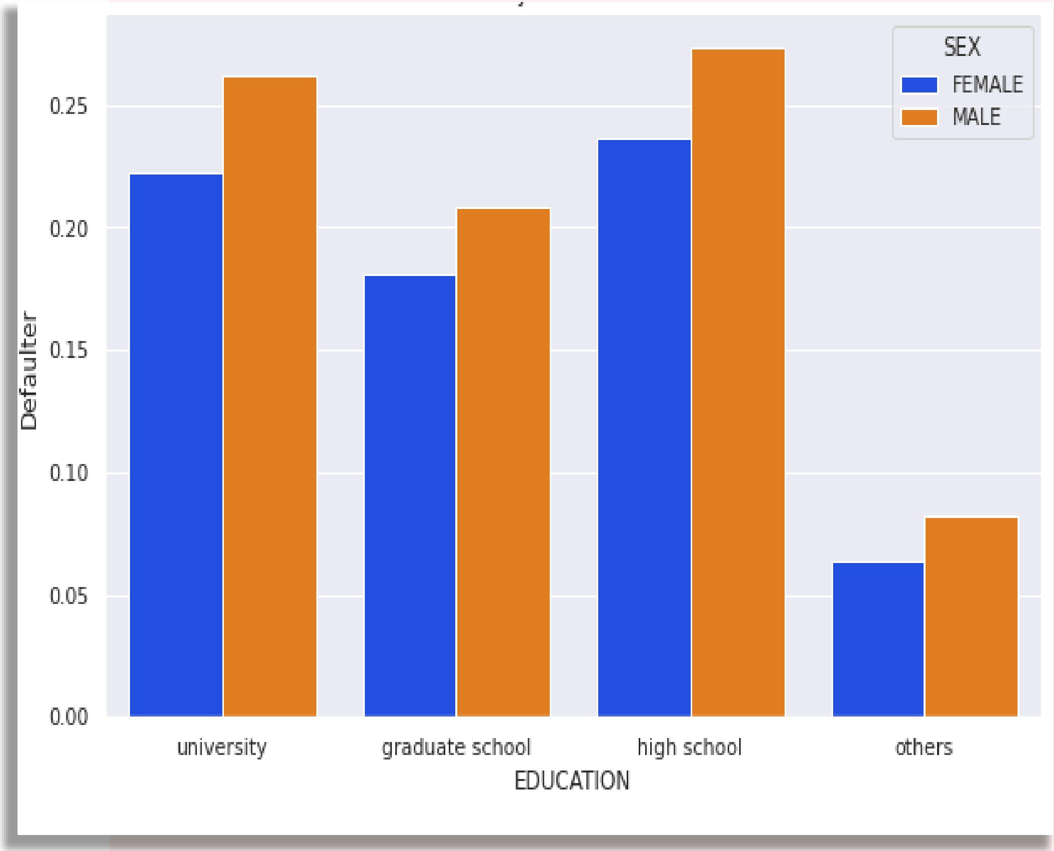


Education Vs Defaulter



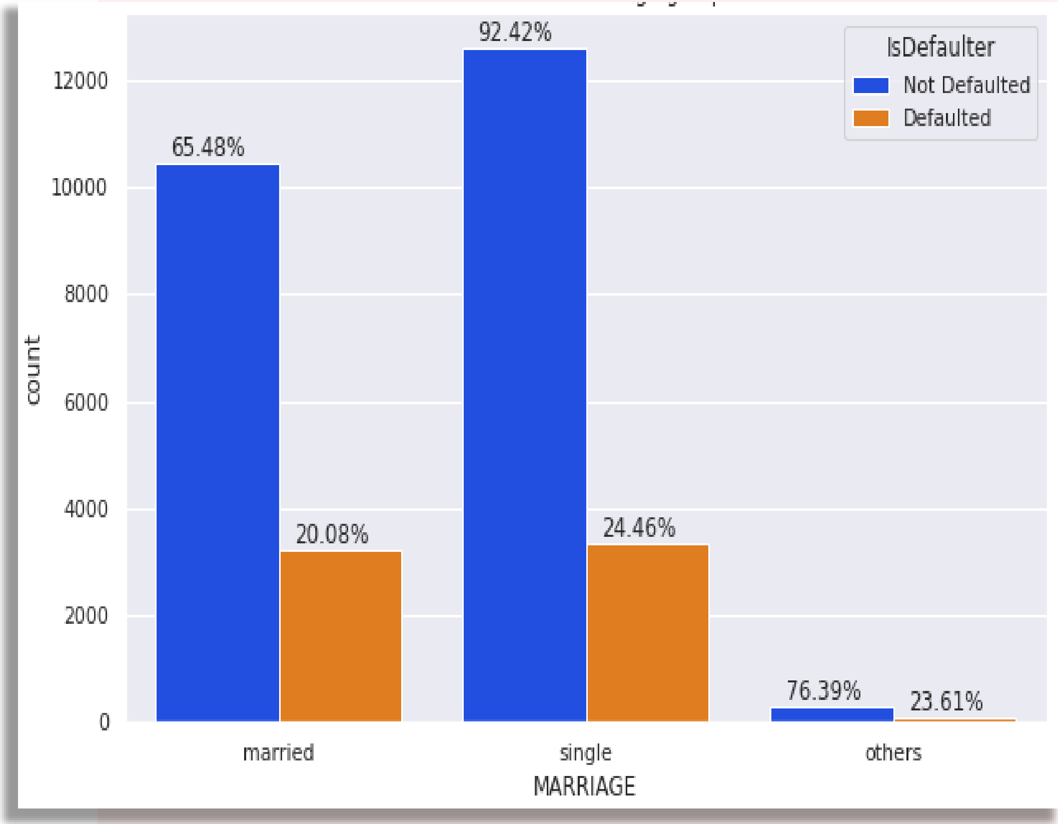
Default by Education and Gender

Male has higher tendency towards default in each educational group



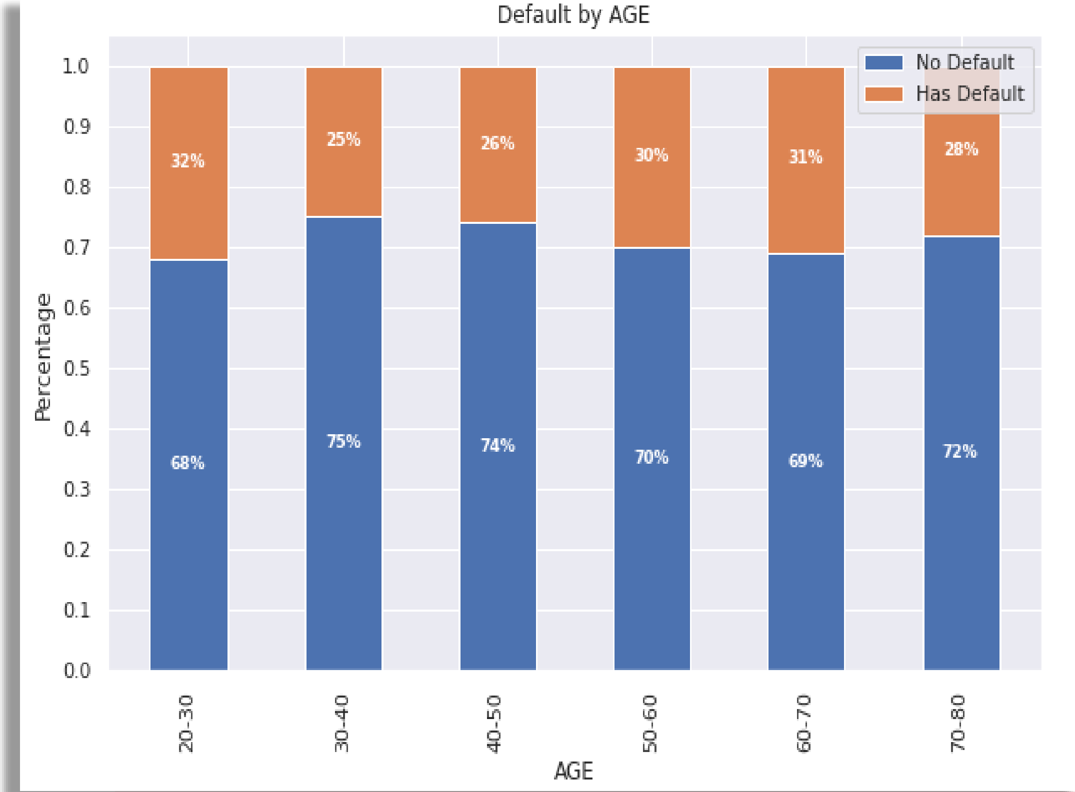
Marriage Vs Defaulter

High defaulter rate when it comes to Singles and Others

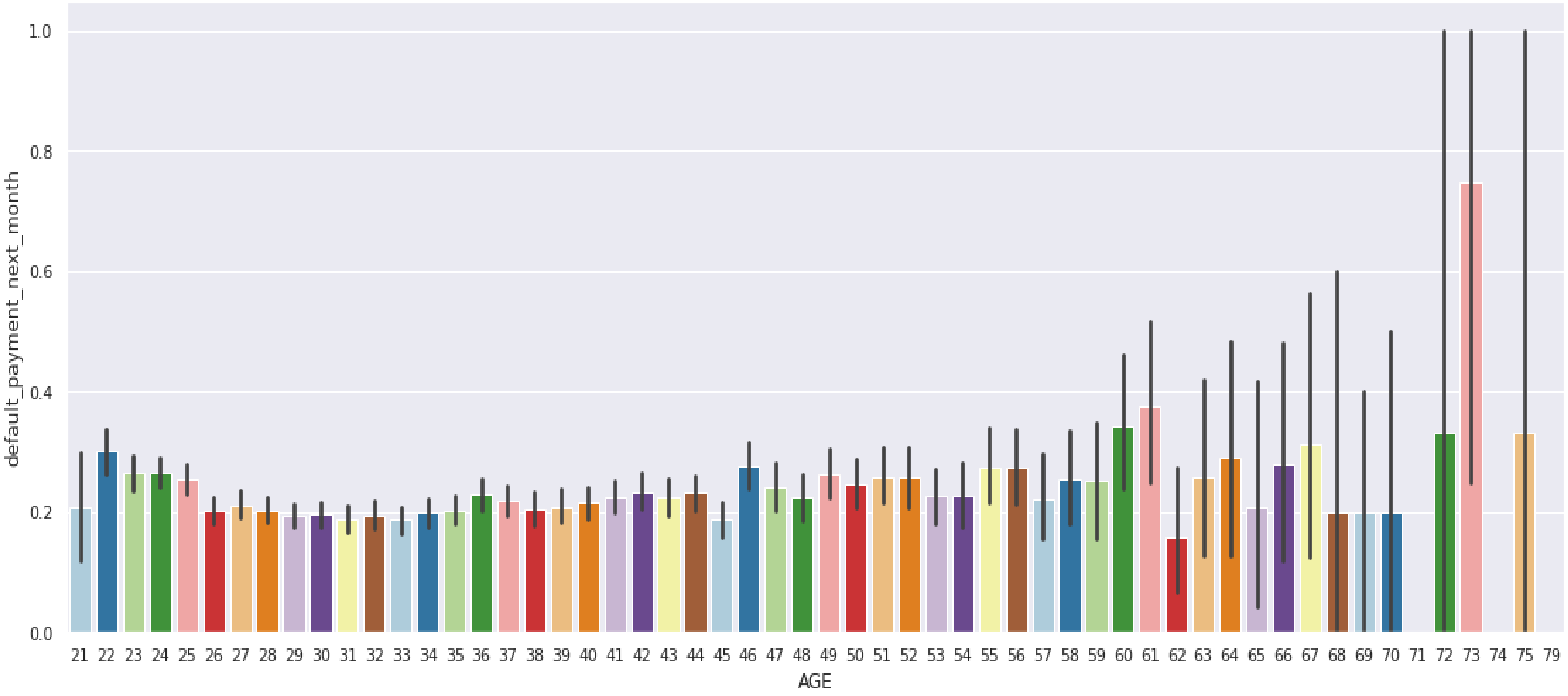


Age Vs Defaulter

High defaulter rate in thirties and after fifties

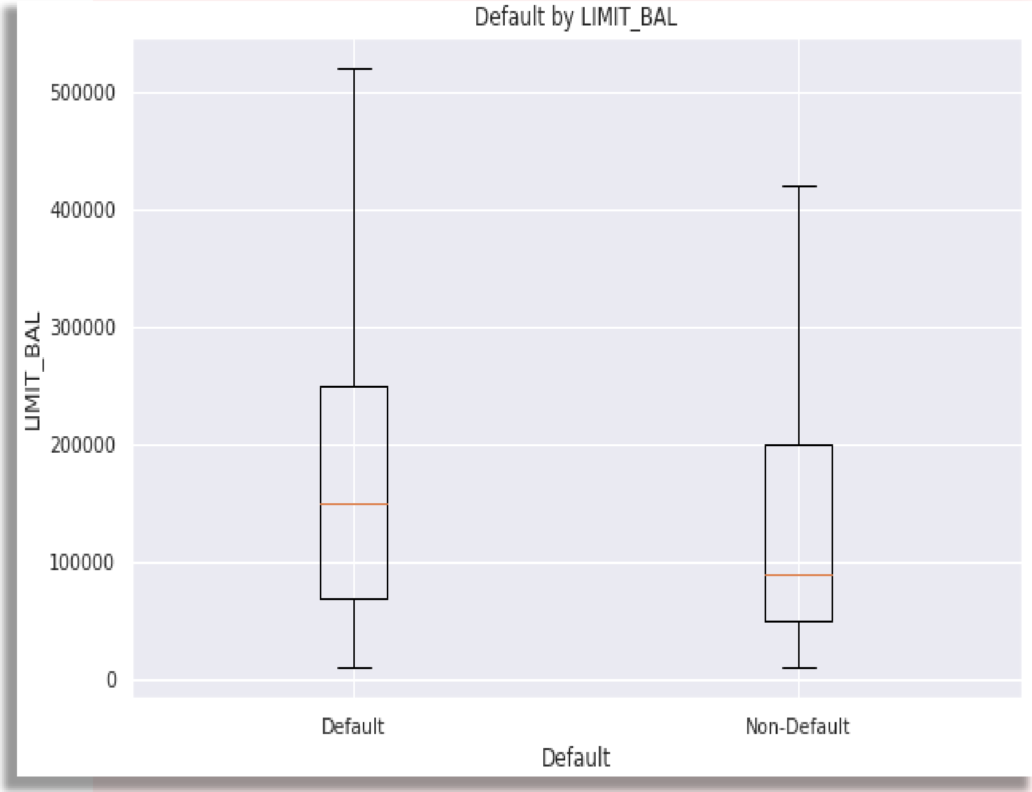


Age Vs Defaulter



Limit Balance Vs Defaulter

Higher the Limit Balance
Lower the default rate



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)

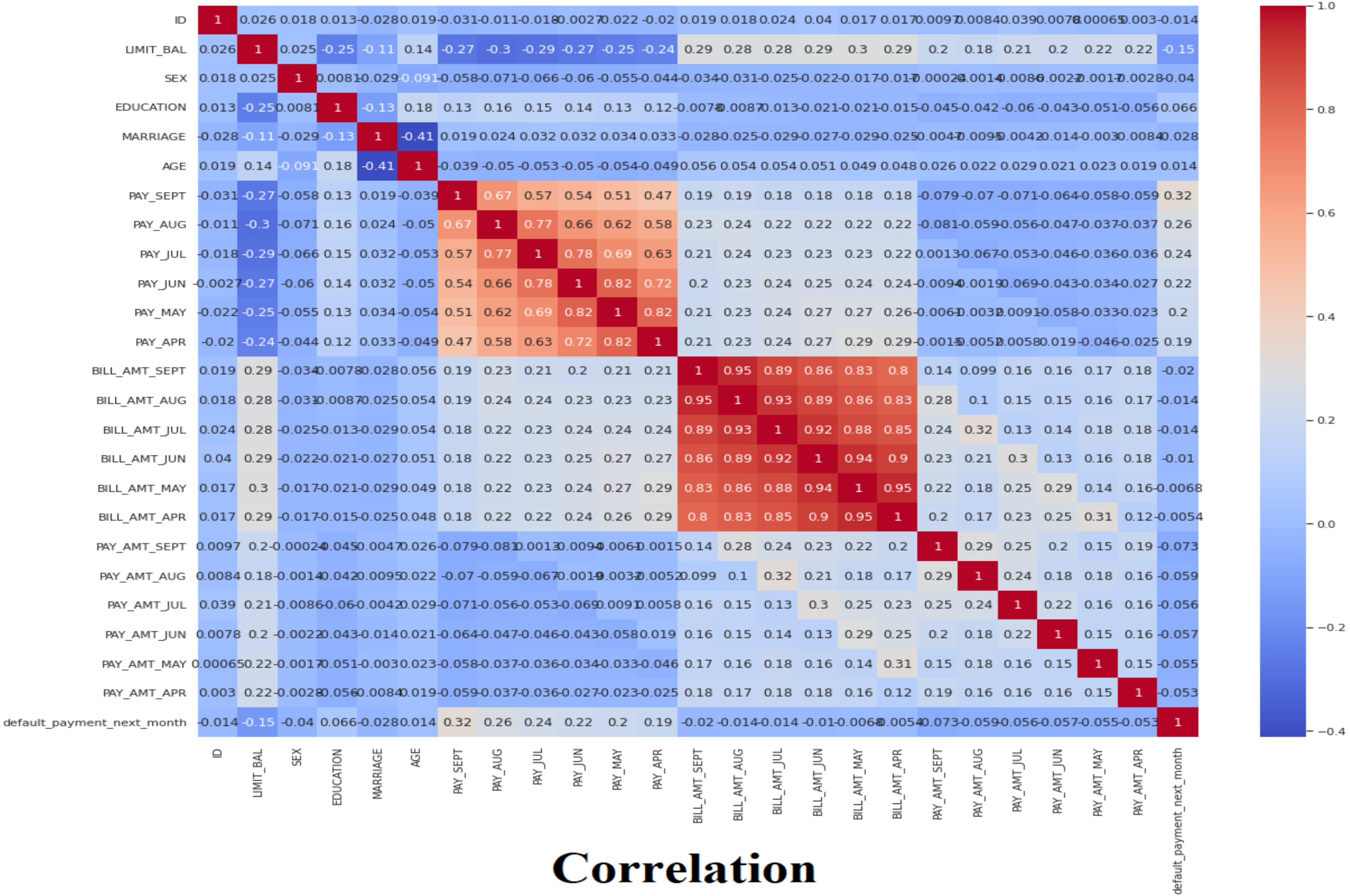
Defaulter

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



Correlation

SMOTE : (Synthetic Minority Oversampling Technique)

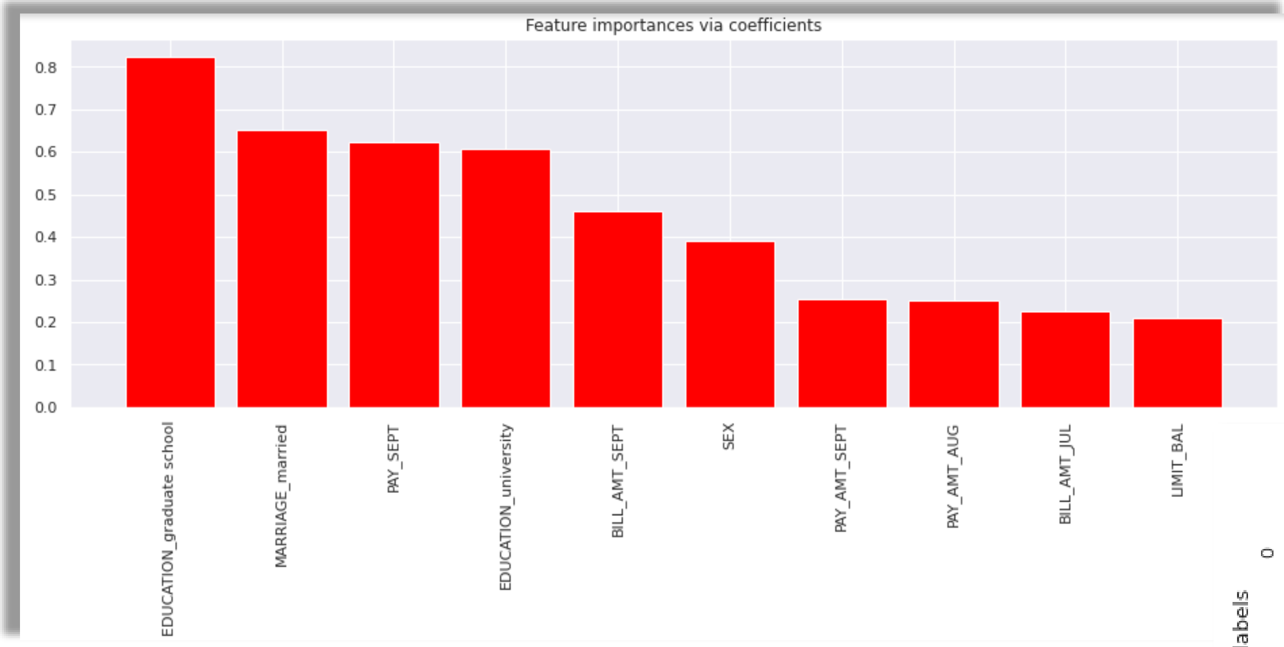
- SMOTE is an **oversampling** technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling

- Original dataset shape 30000
- Resampled dataset shape 46728

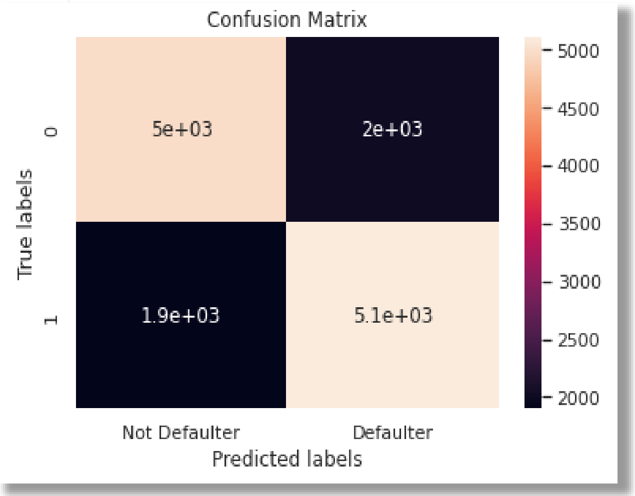
Feature Engineering

1. IsDefaulter
2. Label encoding: Gender
3. One hot encoding: Education and Marriage
4. Separating Independent and Dependent variables
5. Rescaling values using StandardScaler
6. Train test split

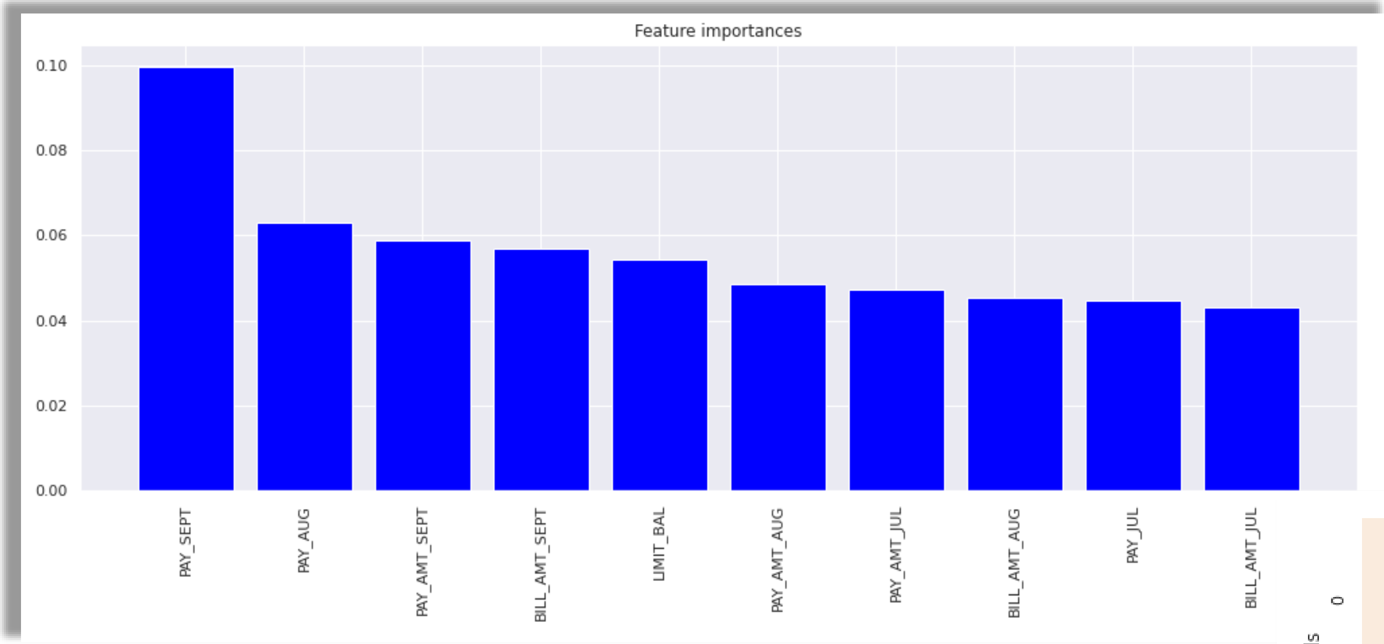
Logistic Regression



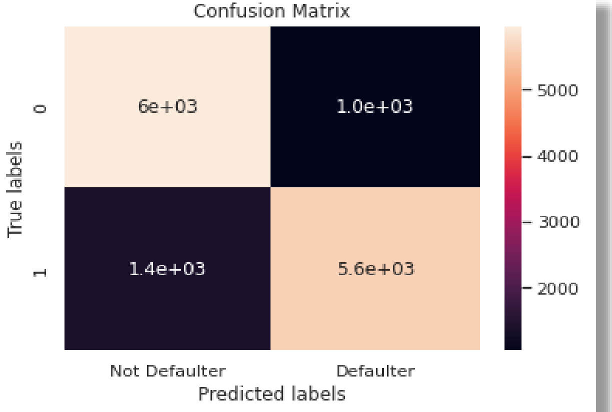
[[4991 2019]
[1897 5112]]



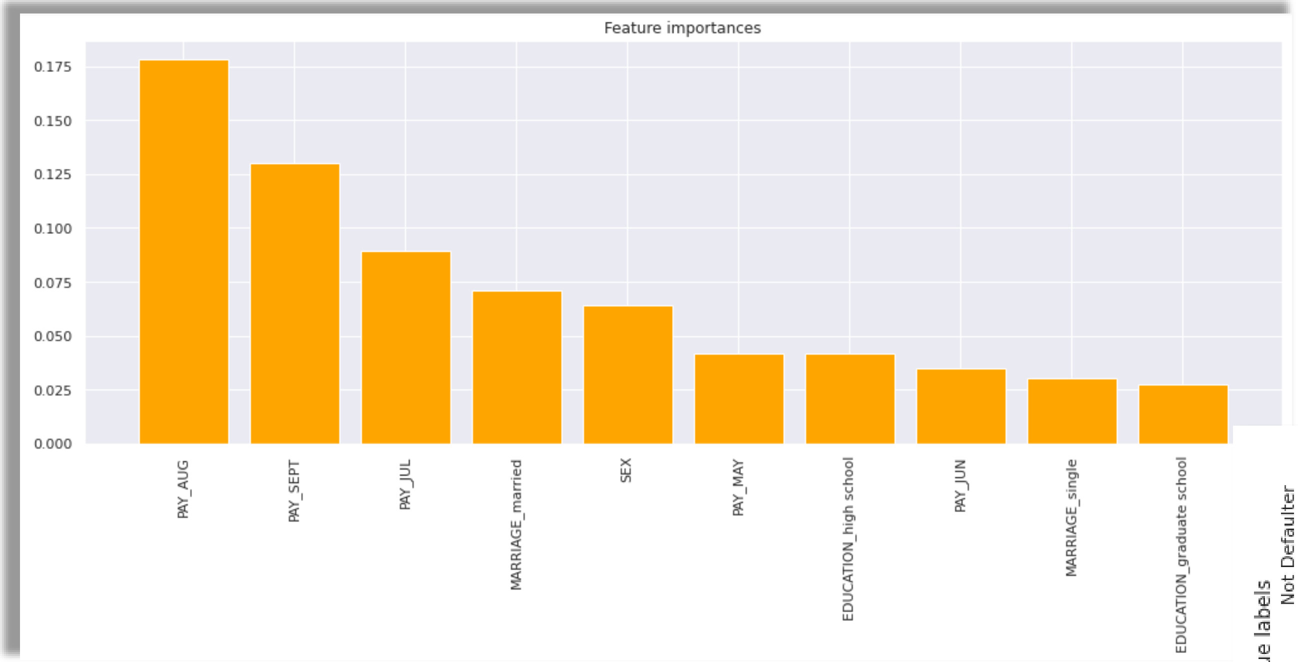
Random Forest

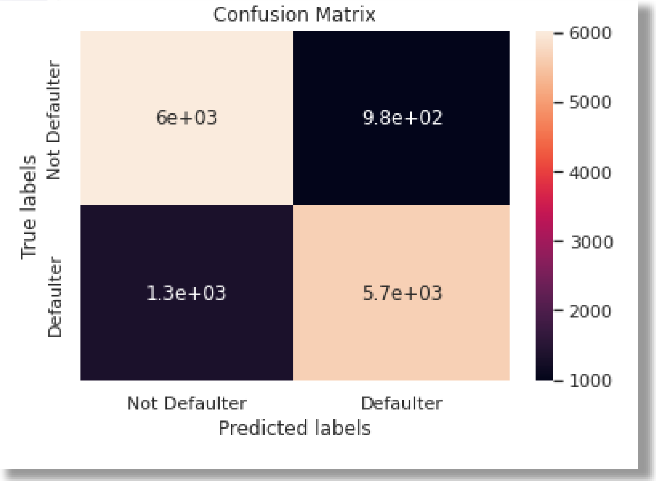


- [[5960 1050]
- [1407 5602]]



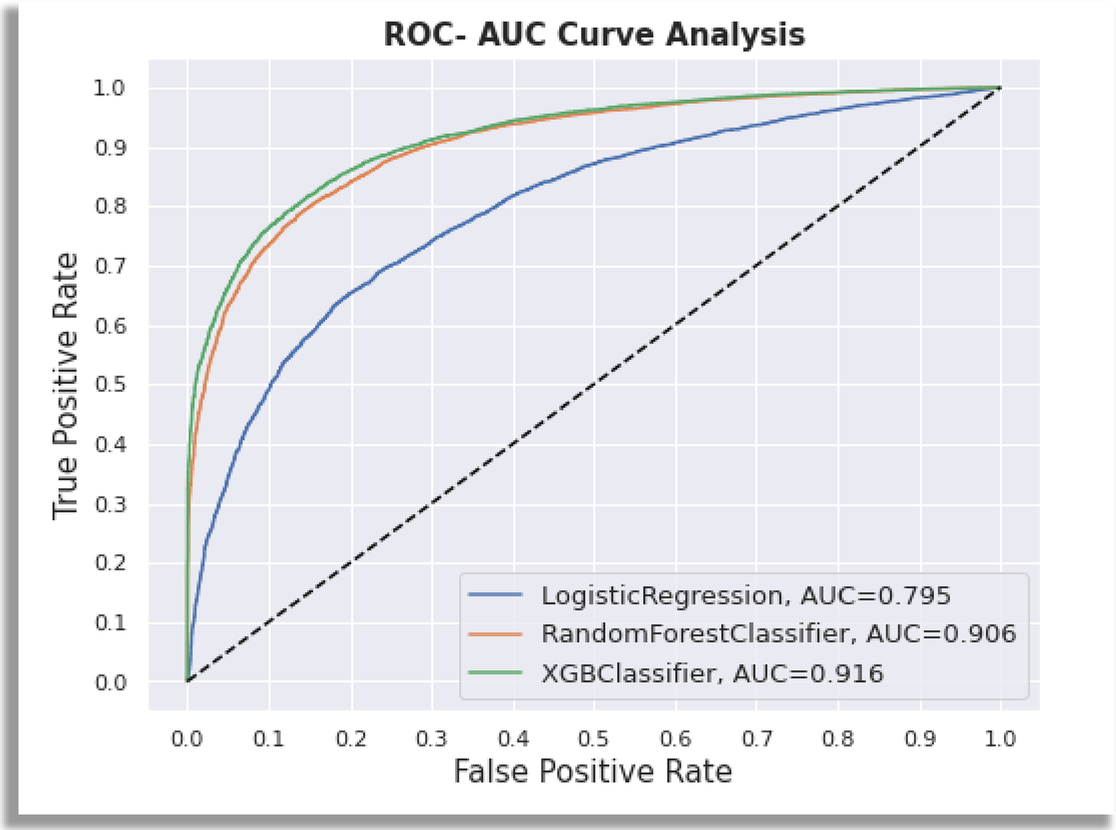
XGBoost

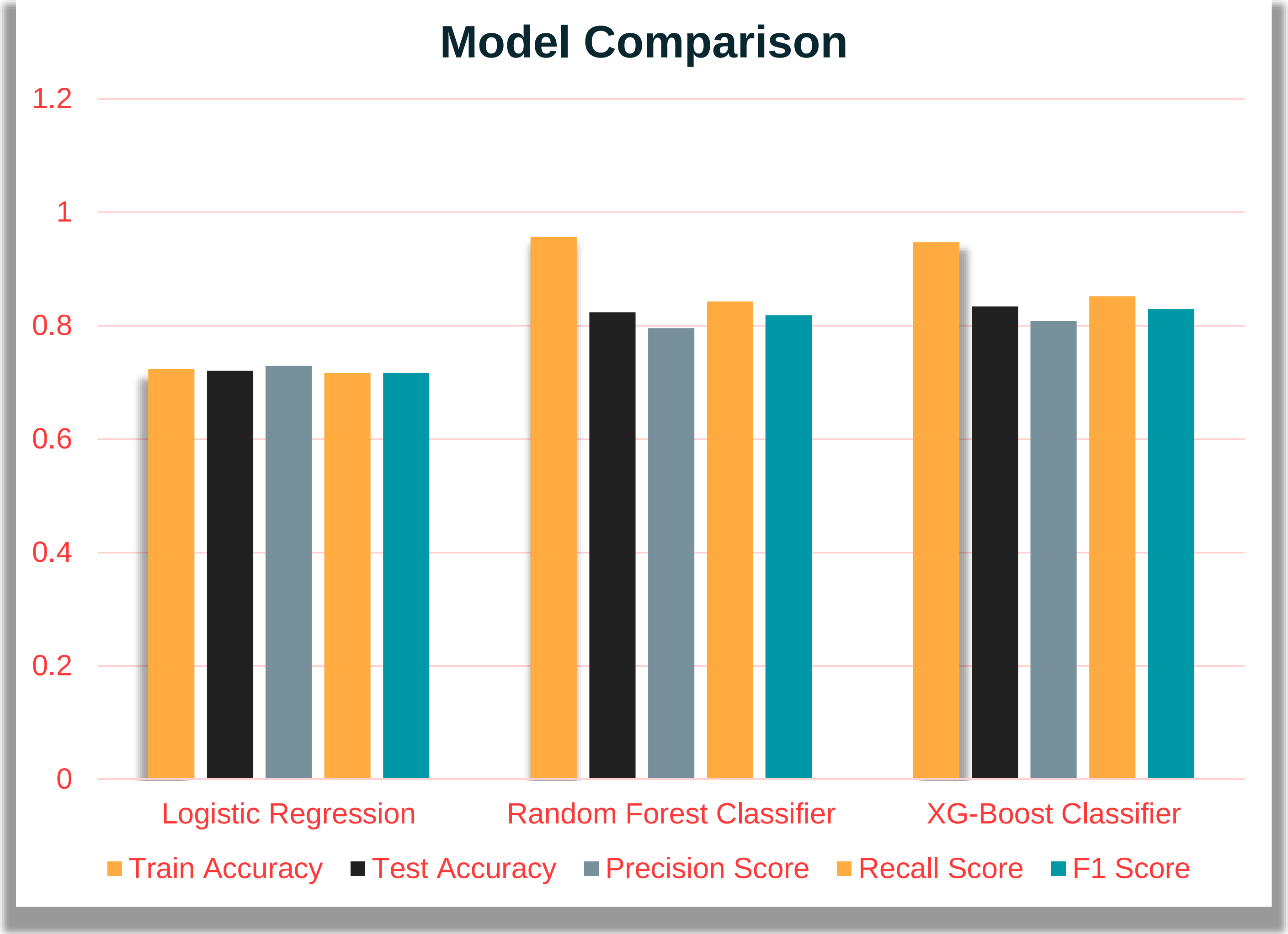




[[6028 982]
[1346 5663]]

ROC Curve


ROC- AUC Curve Analysis
True Positive Rate
False Positive Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
LogisticRegression, AUC=0.795
RandomForestClassifier, AUC=0.906
XGBClassifier, AUC=0.916



Model Comparison
1.2
1
0.8
0.6
0.4
0.2
0
Logistic Regression
Random Forest Classifier
XG-Boost Classifier
Train Accuracy
Test Accuracy
Precision Score
Recall Score
F1 Score

Conclusion

- Data categorical variables had minority classes which were added to their closest majority class

- We have built predictive model for credit card agency to predict if a person would default on his/her payment of credit card.

- We have performed feature engineering, feature selection, hyperparameter tuning to prevent overfitting and decrease error rate in the model.

- Since the business nature of credit card default prediction requires model to have a high recall. Therefore we selected XGBoostas our best model.

Thank You