

# Capstone Project 3

## CREDIT CARD DEFAULT PREDICTION

Akash Salmuthe

# Content

**Introduction**

**Problem And Objective**

**Data Summary**

**Data Analysis**

**EDA**

**ML Models**

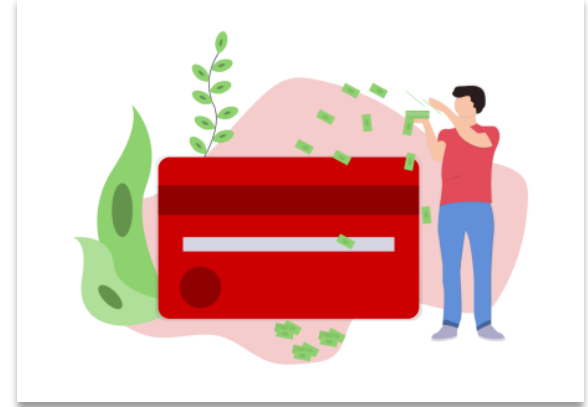
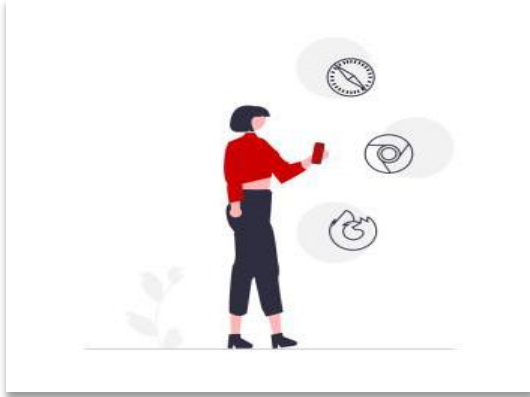
**Metrics**

**Feature Importance**

**Conclusion**

# 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

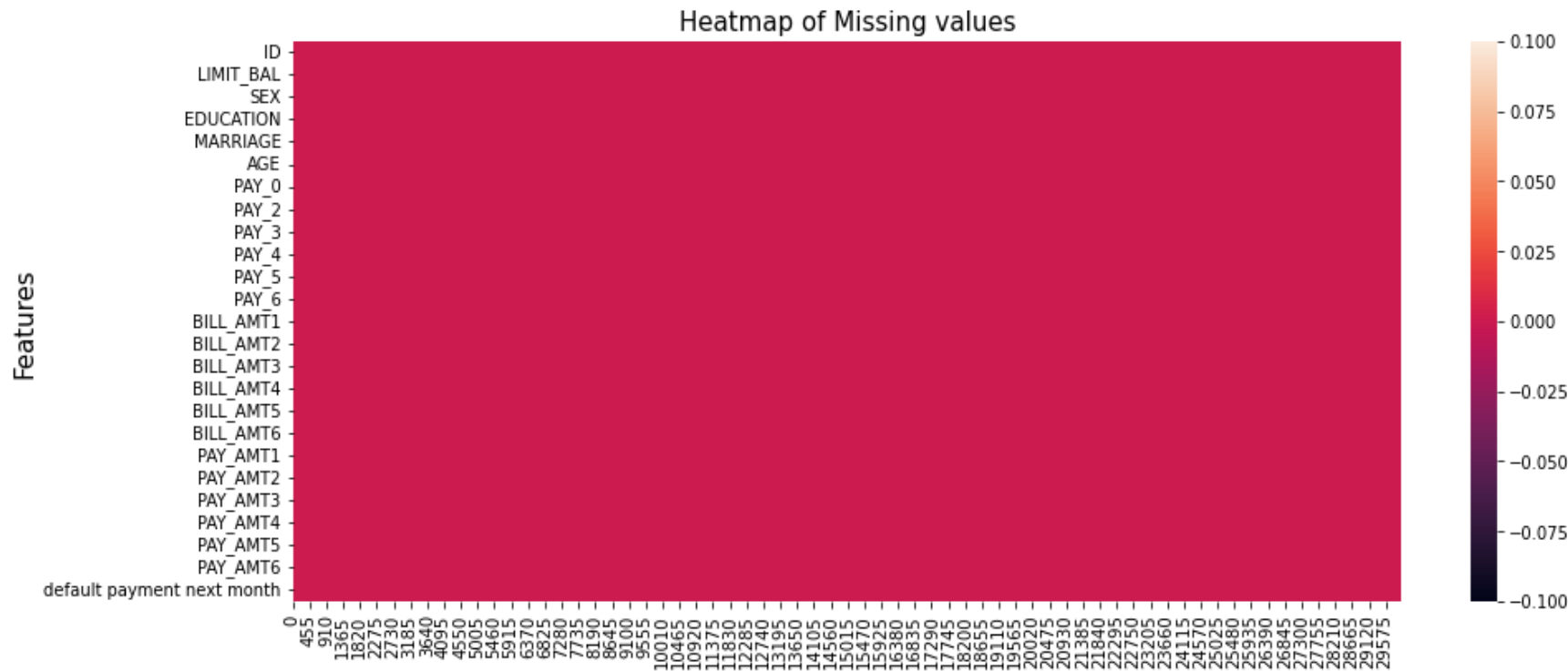
**X12 to X17** - Amount of bill statement from April to September

**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.

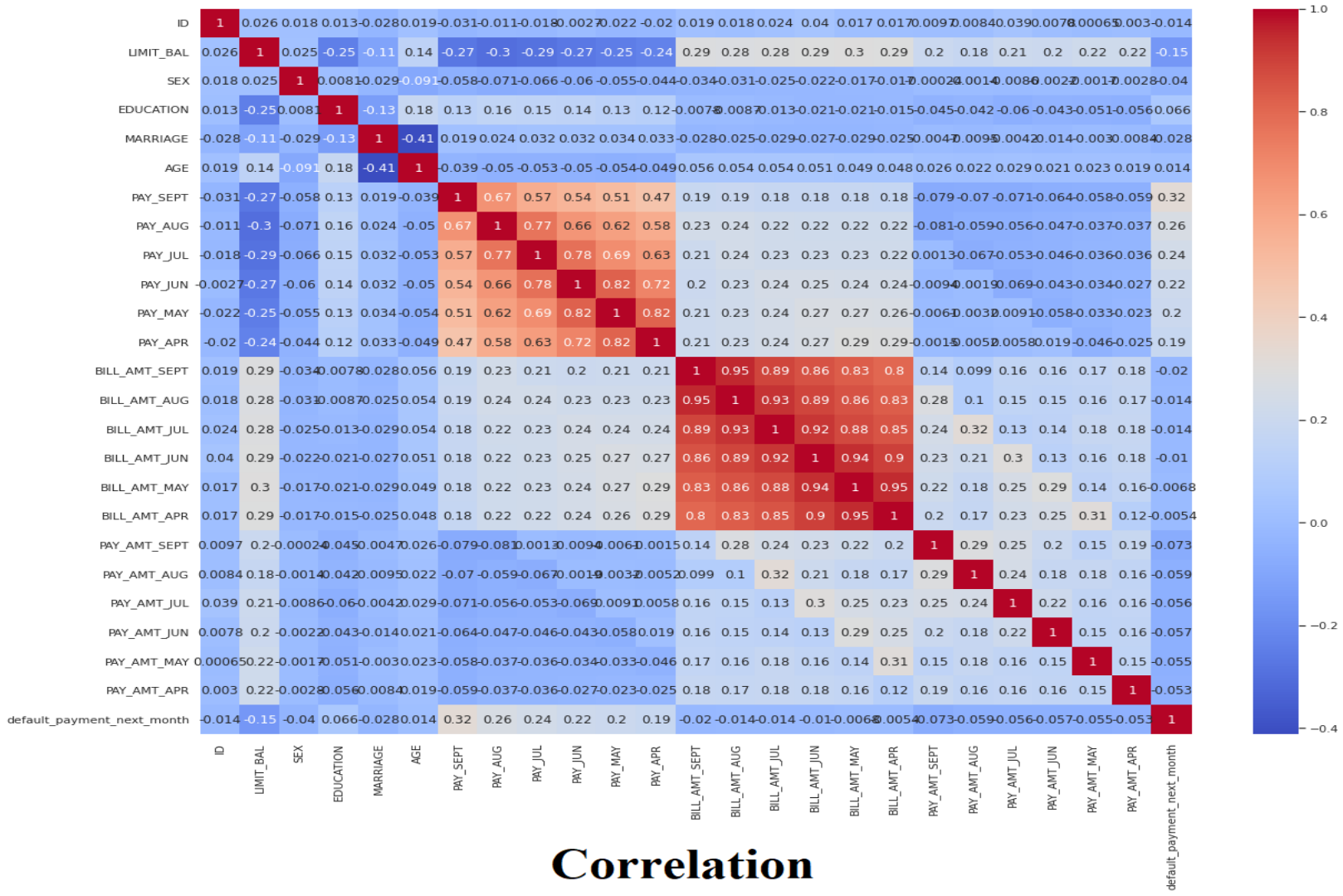
# Missing Value





# Exploratory Data Analysis



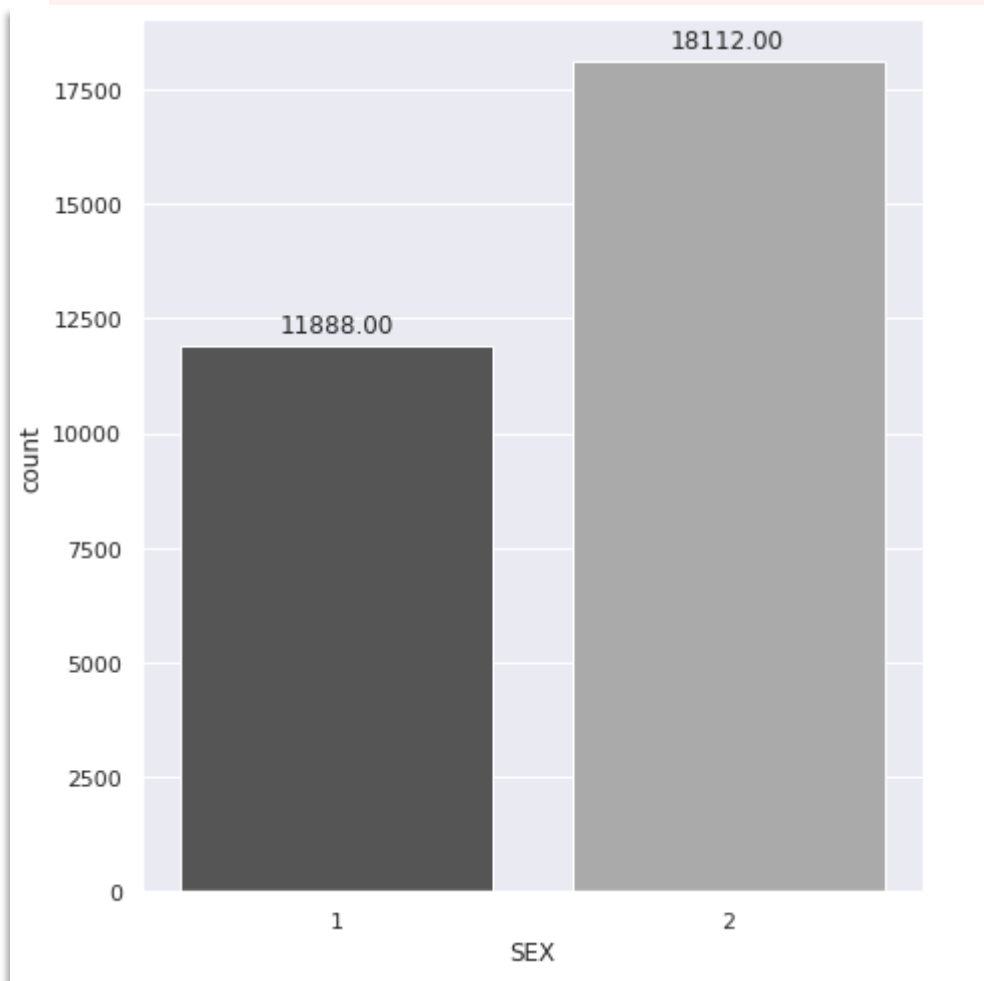


# Gender

**1 : Male**

**2 : Female**

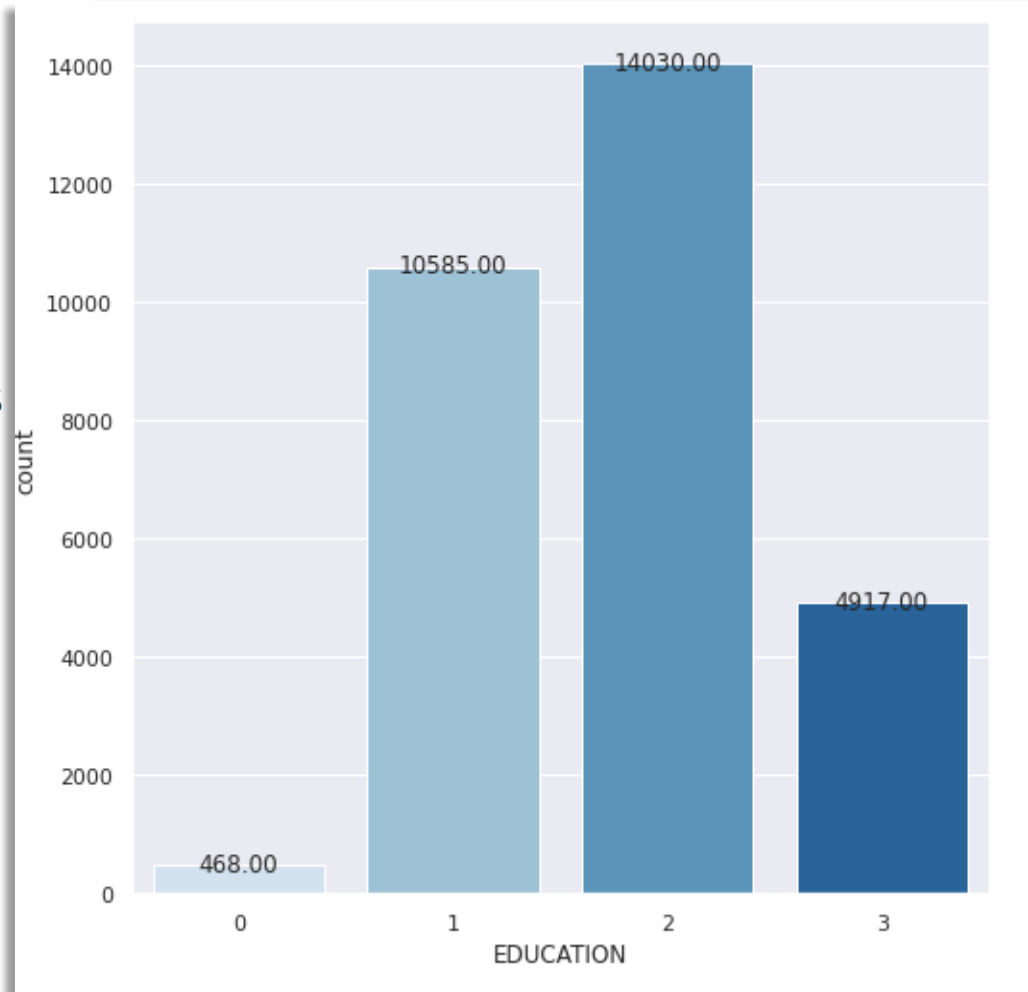
Here we can clearly see that female holds most of credit cards



# Education

**More number of credit holders are:**

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



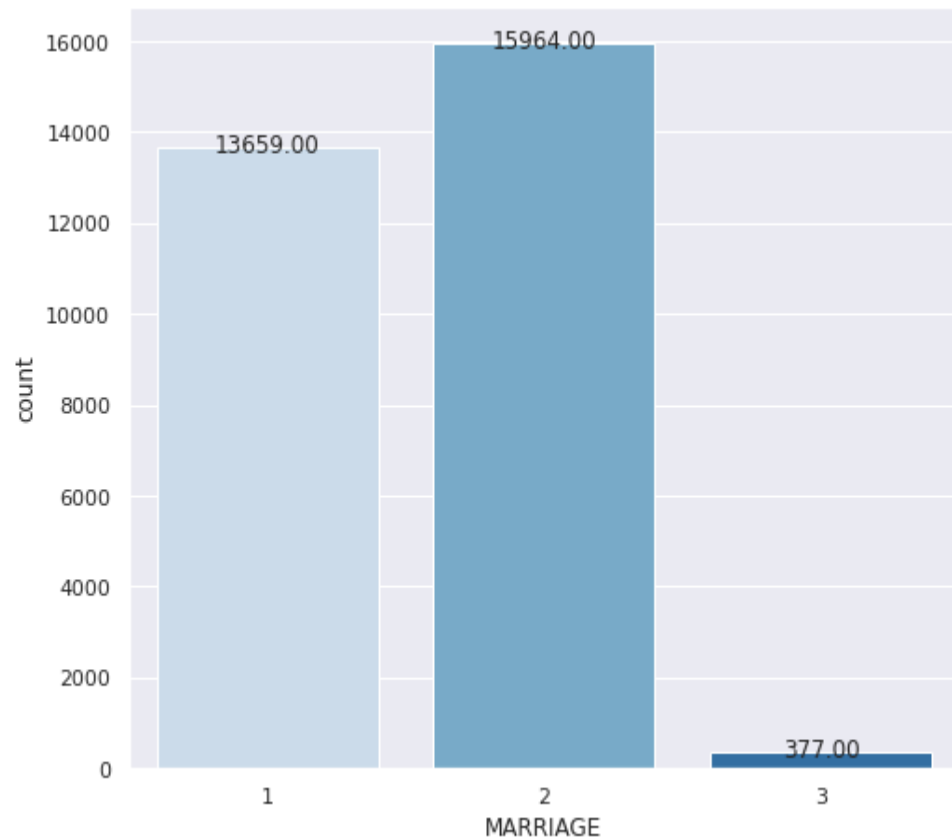
# Marriage

Here,

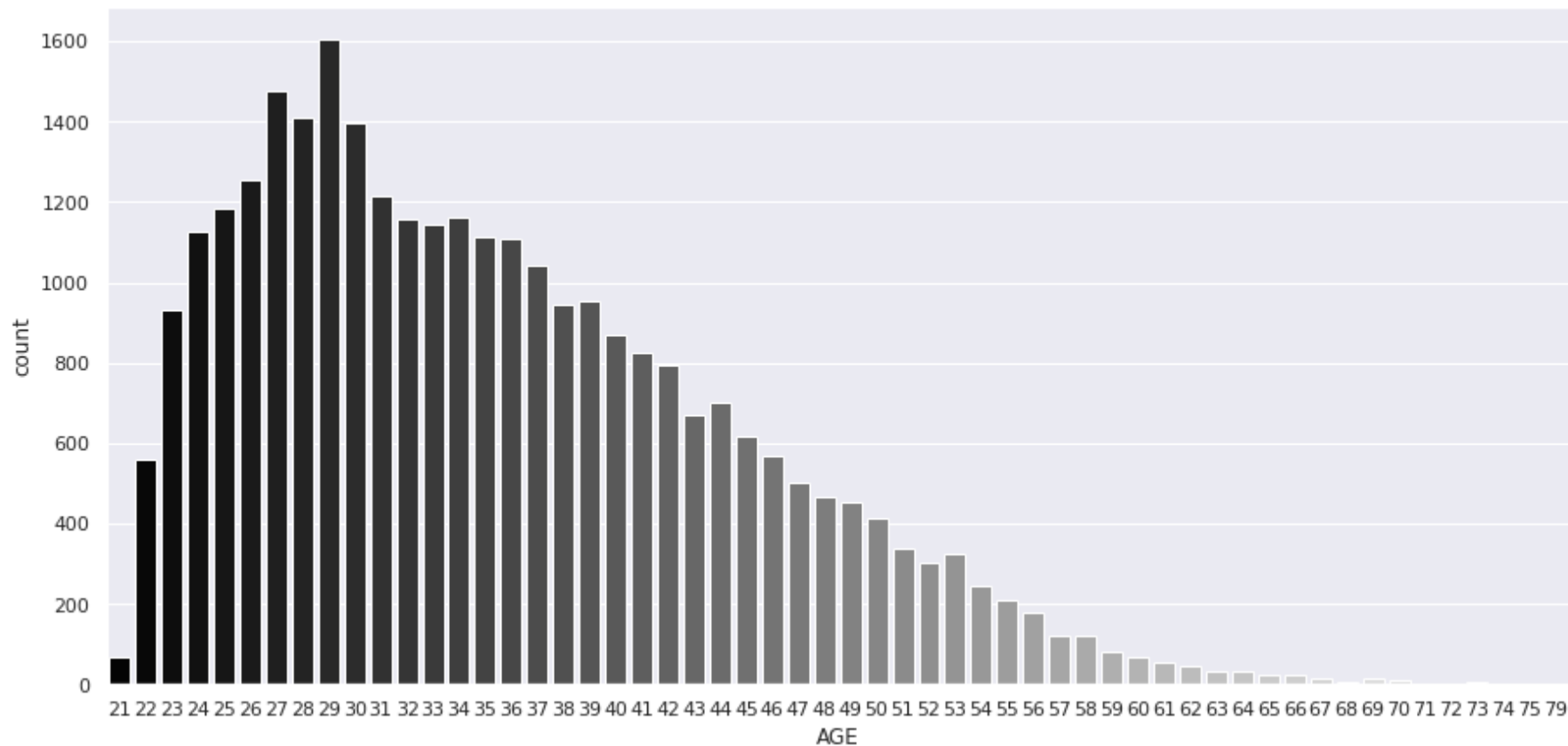
1 : Married - 13659

2 : Unmarried – 15964

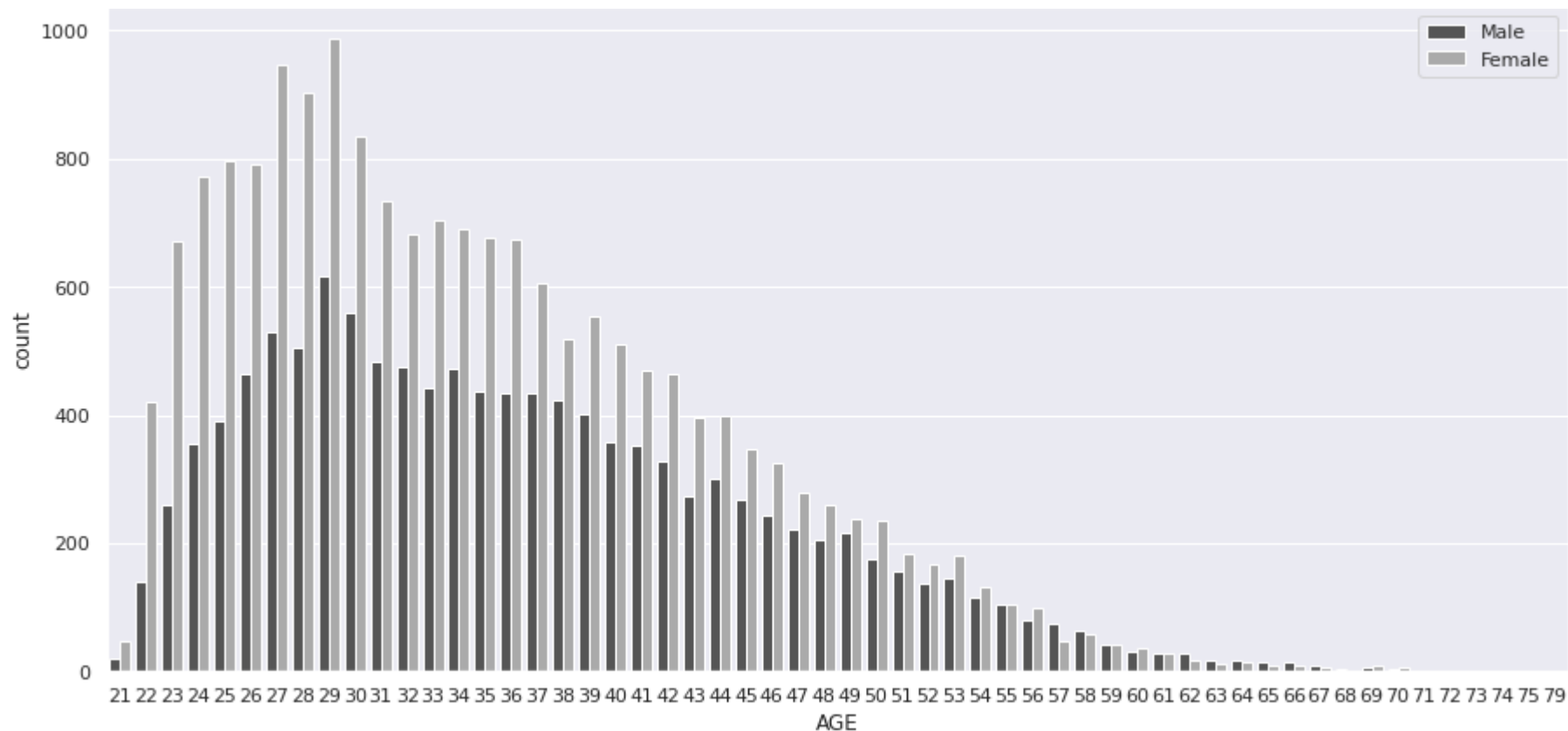
3 : Others – 377 (54 + 323)



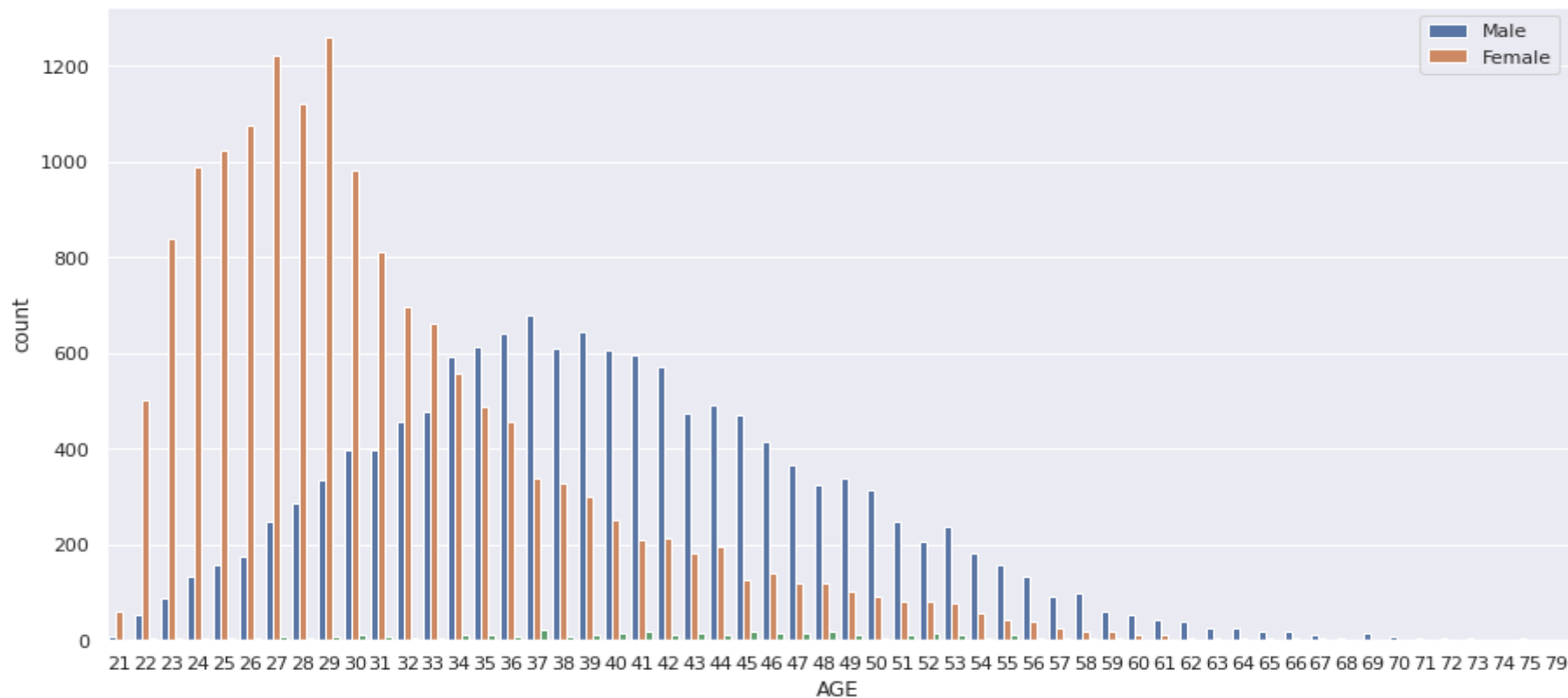
# Age Vs Card Holders



# Age vs Gender



# Age vs Marriage



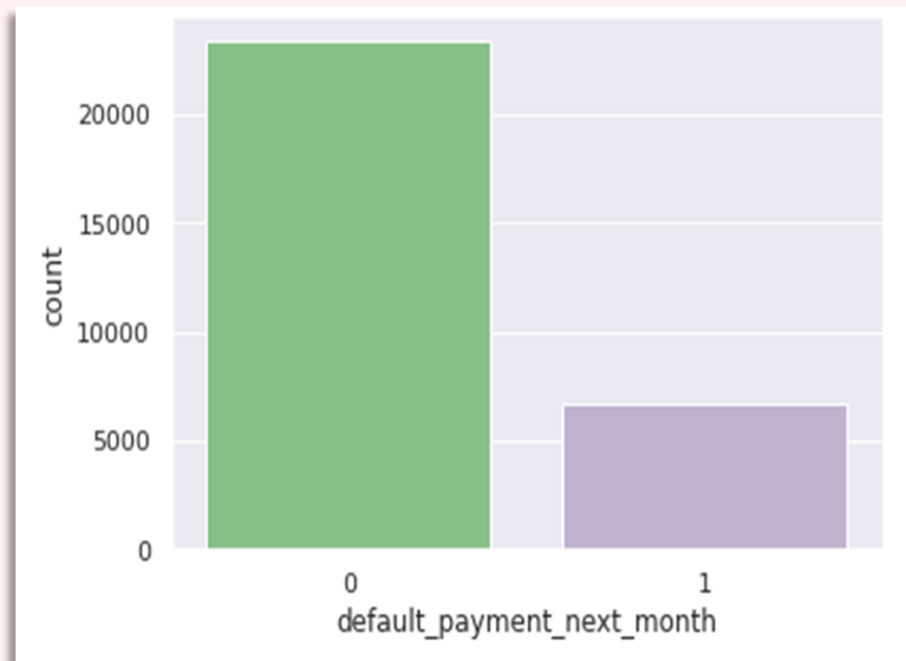


# Default Next Month

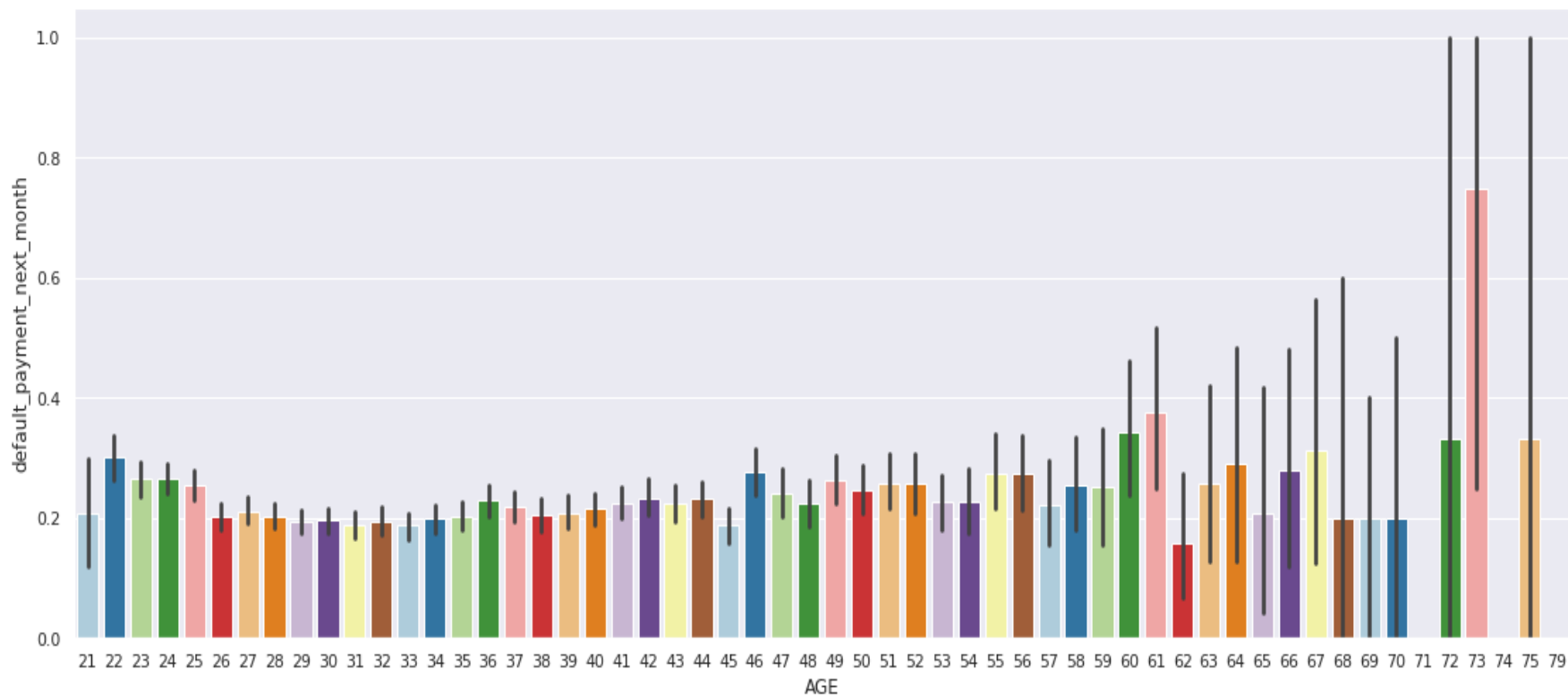
**Here,**

0 : Not Default

1 : Default

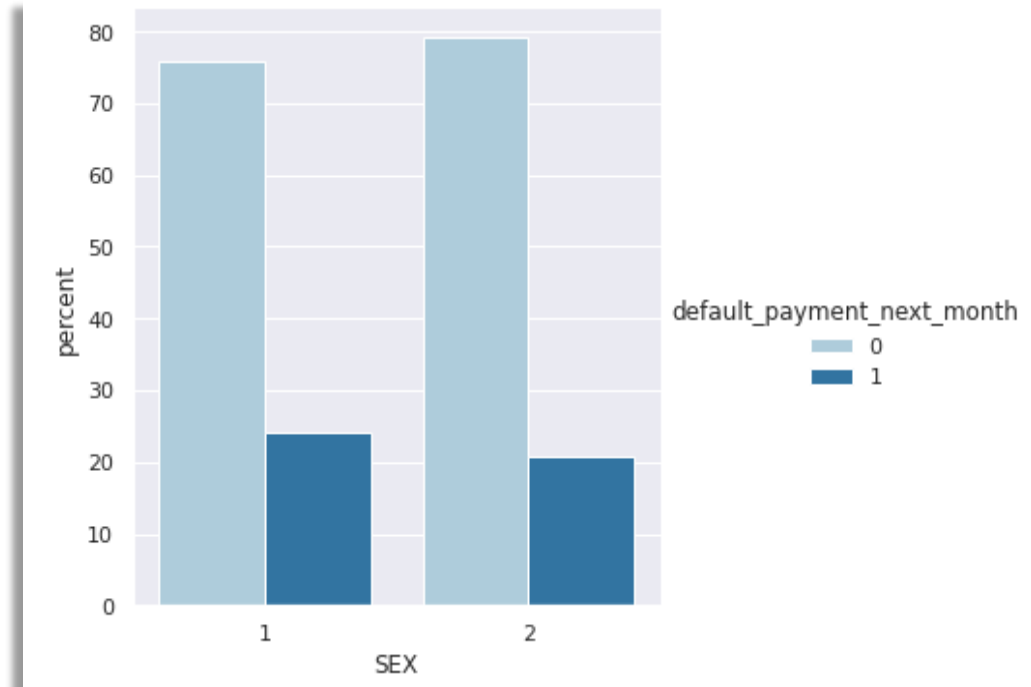


# Age Vs Defaulter



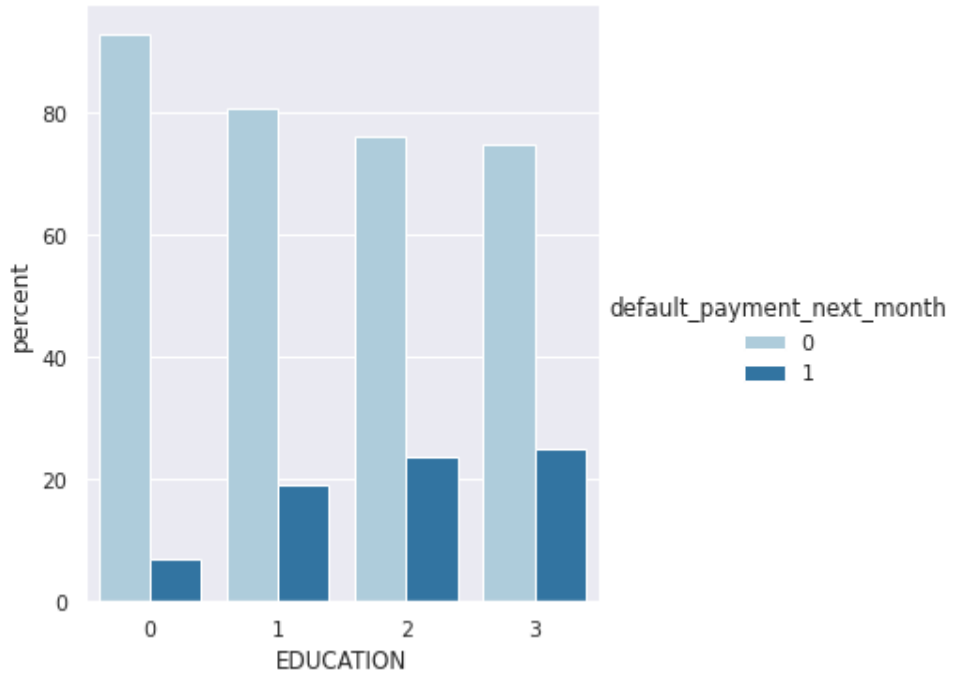
# Gender Vs Defaulter

Clearly see that Male has higher default rate



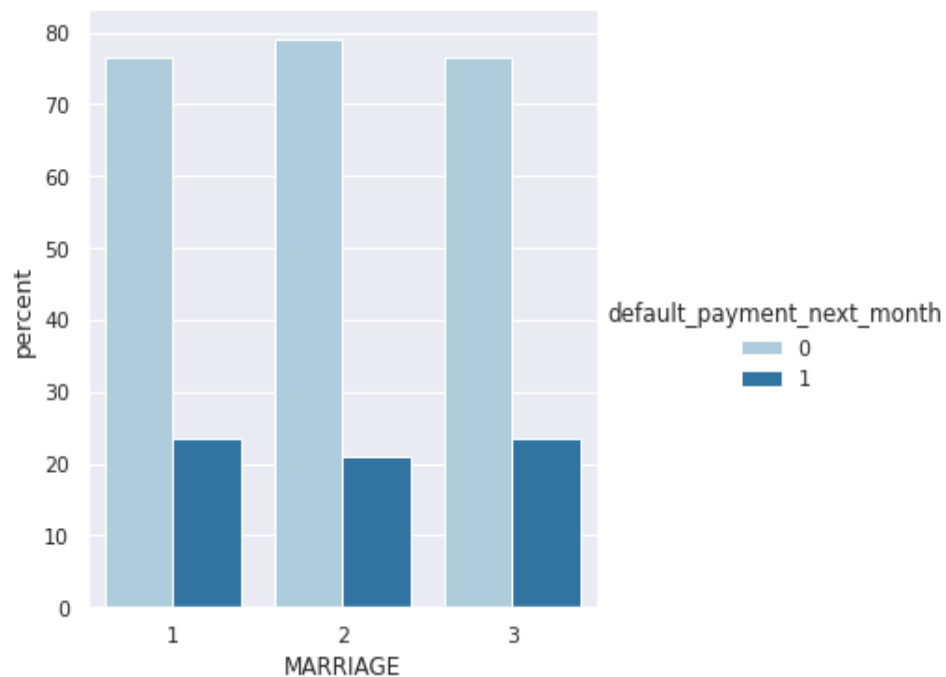
# Education Vs Defaulter

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

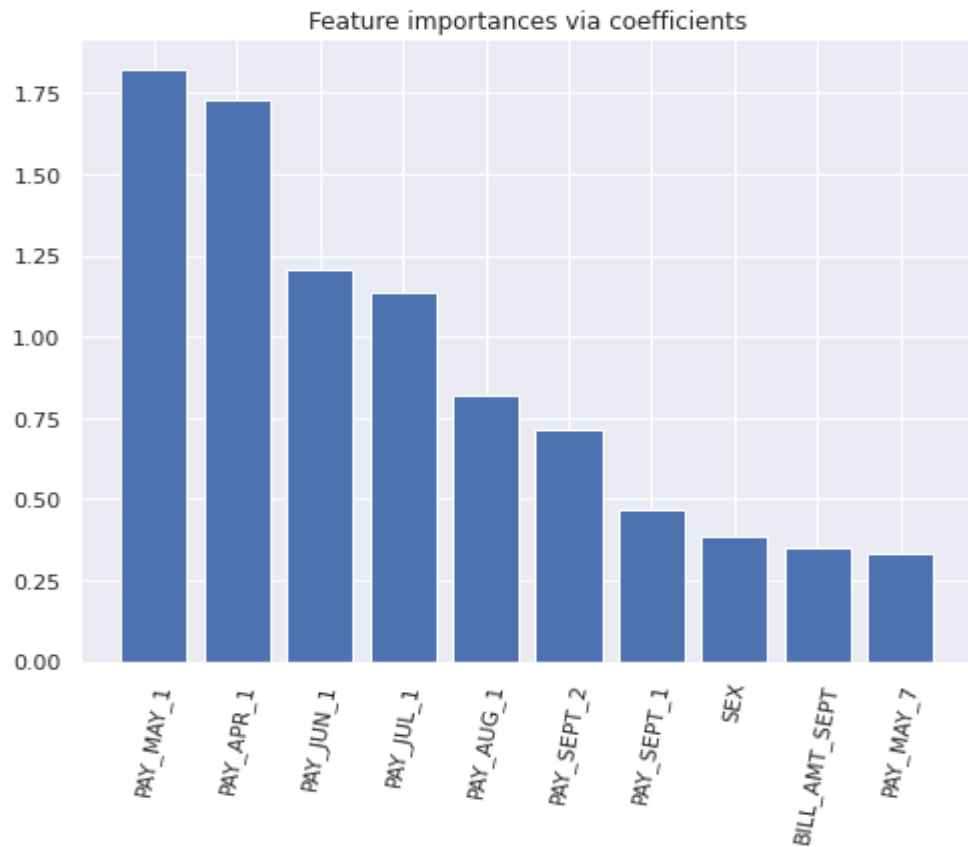
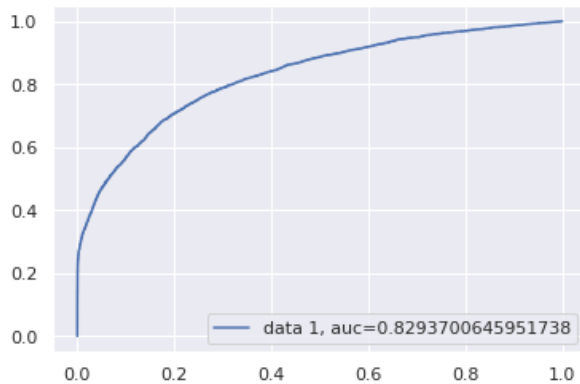
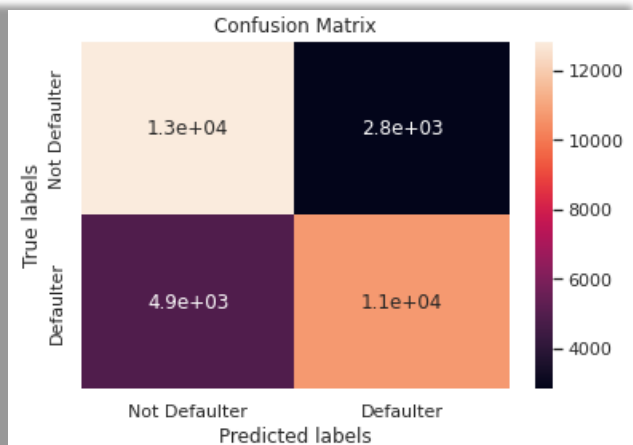
# Modeling

## SMOTE : (Synthetic Minority Oversampling Technique)

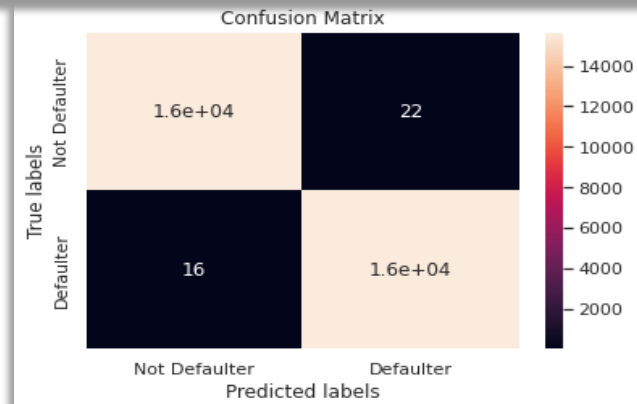
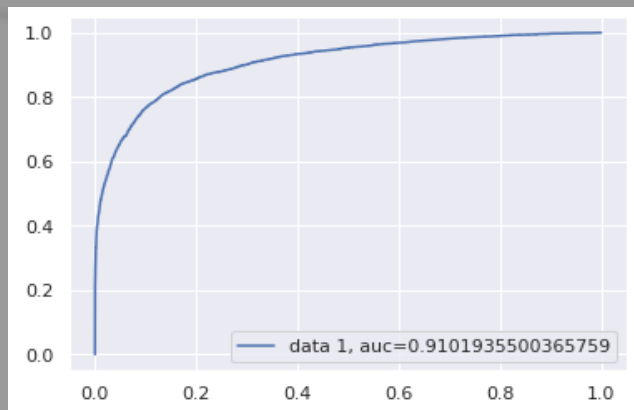
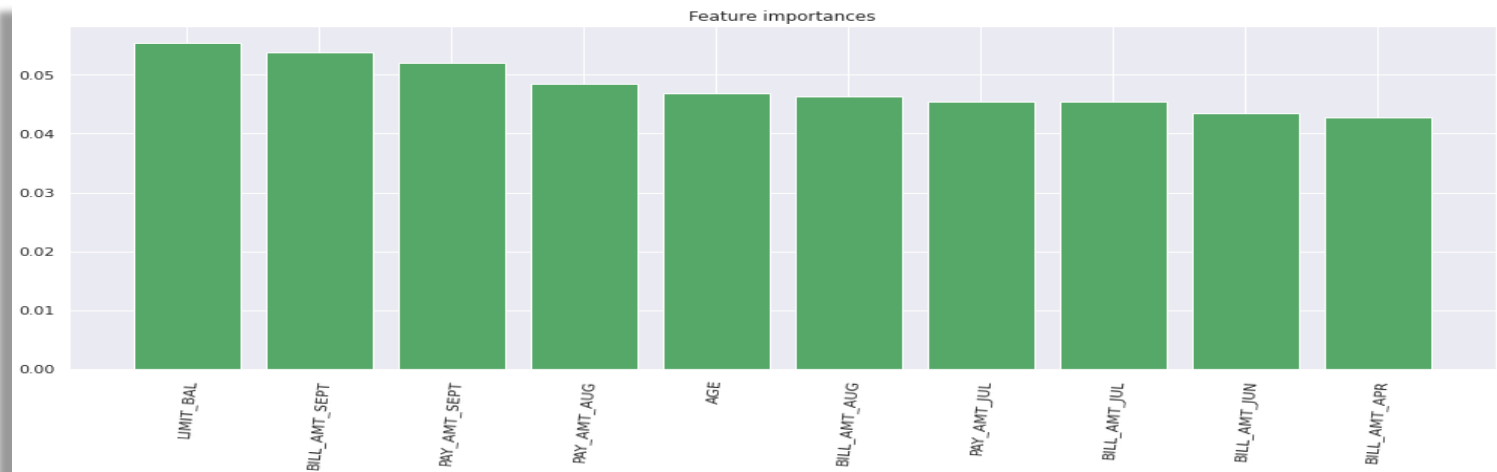
- In heatmap, Highly correlated items "PAY\_SEPT", "BILL\_AMT\_SEPT", "PAY\_AMT\_SEPT" removed.
- After dataset is imbalanced dataset so we need to do the balance using SMOTE
- Original dataset shape Counter ({0: 18691, 1: 5309})
  - Resample dataset shape Counter ({1: 23364, 0: 23364})
  - Counter ({0: 23364, 1: 23364})



# Logistic Regression

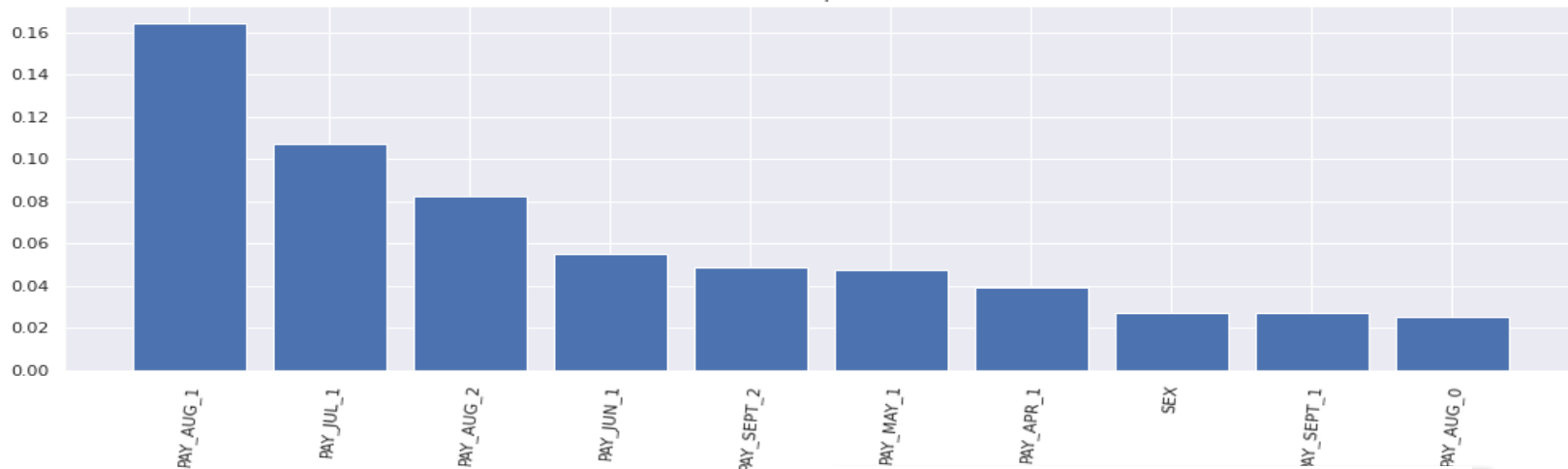


# Random Forest

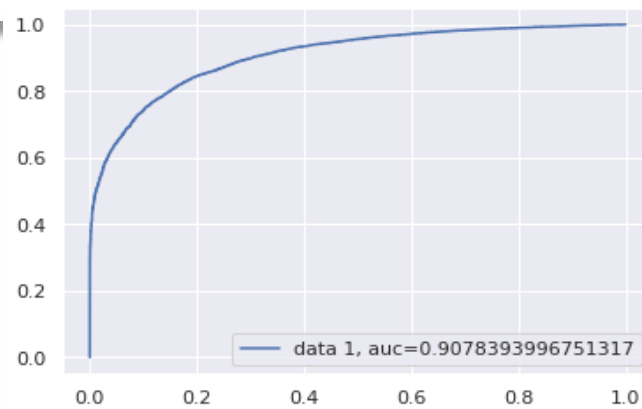
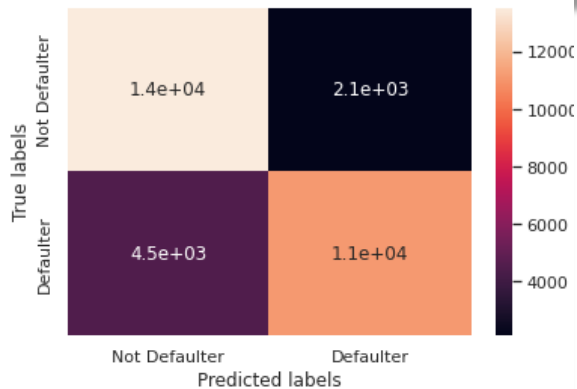


# XGBoost

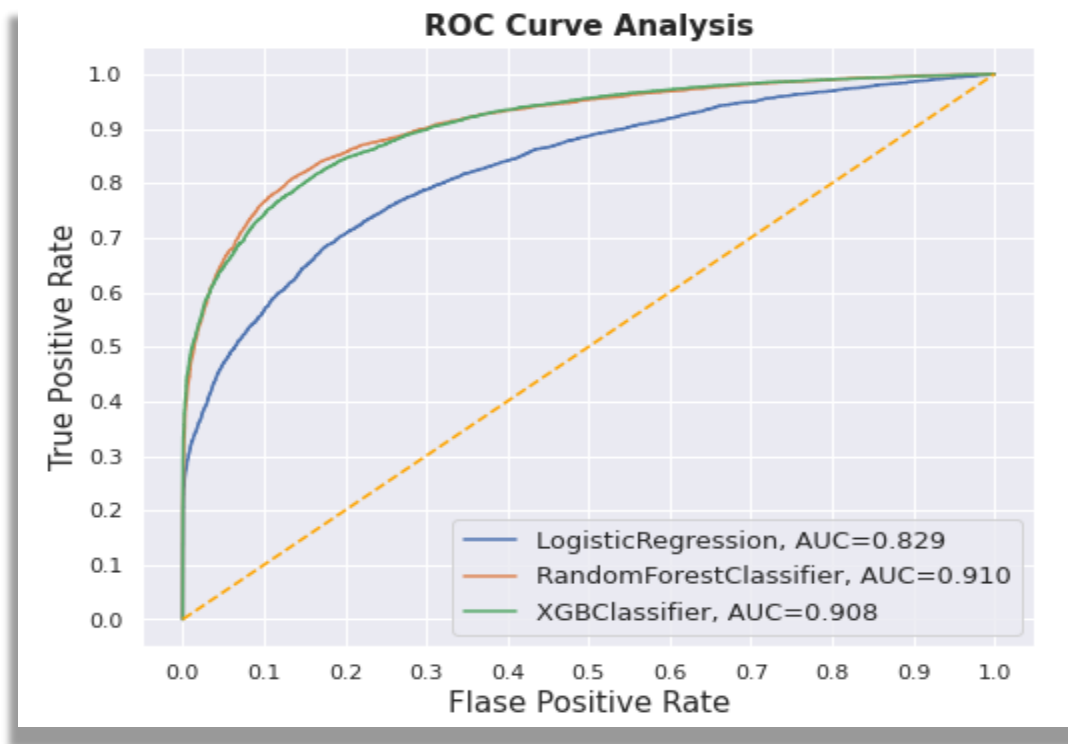
Feature importances



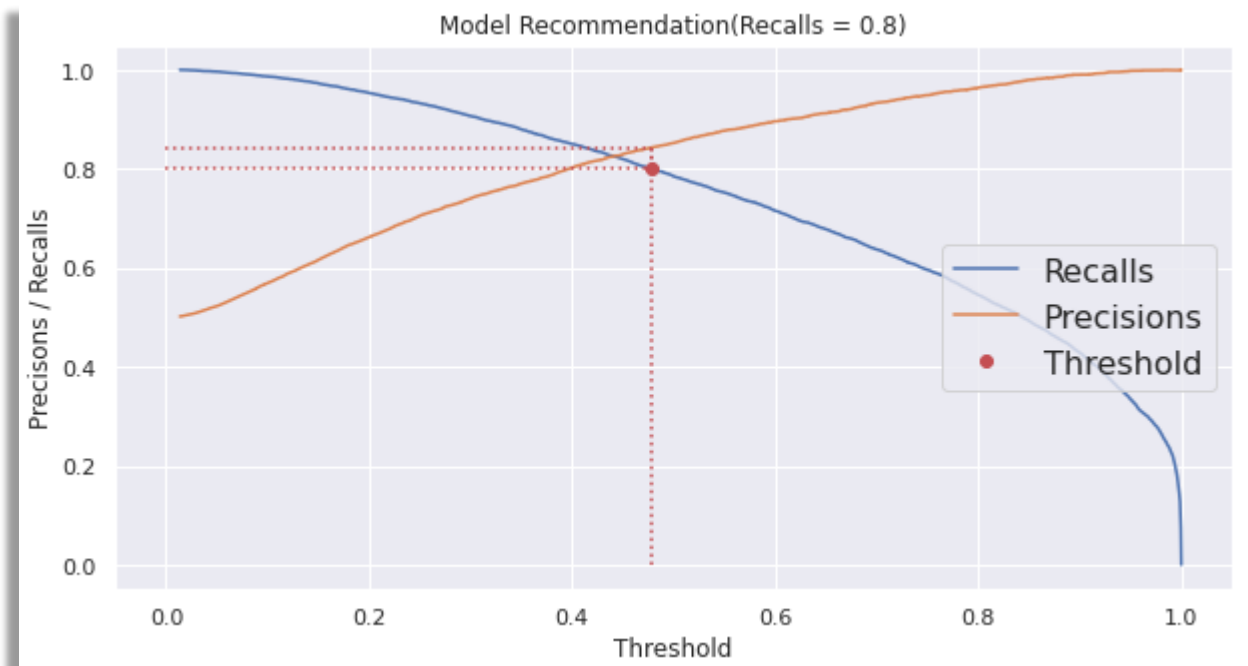
Confusion Matrix



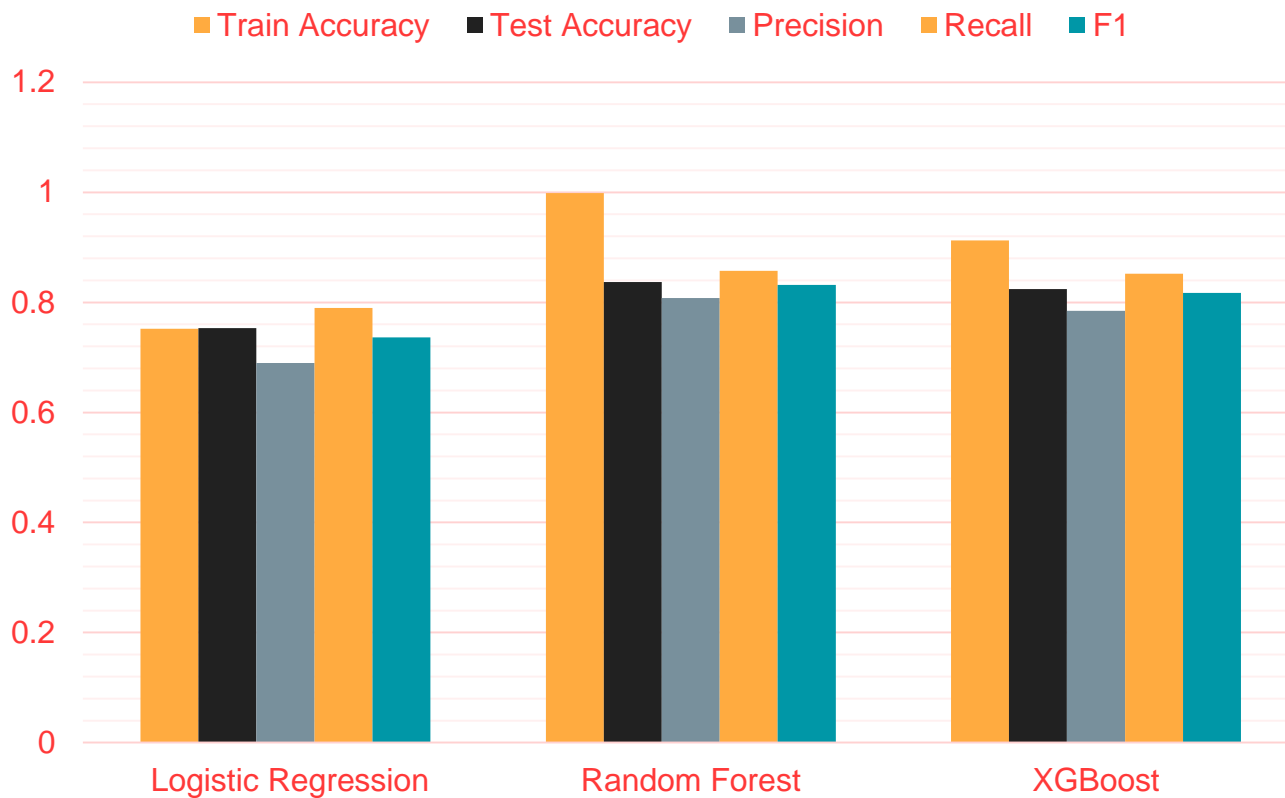
# ROC Curve



# Threshold



## Model Performed



# 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