CAPSTONE PROJECT

Credit Card Default Prediction

TEAM MEMBERS

Rishabh Kumar

Shivansh Yadav

Madhur Awasthi

(Data Science Trainees @ Almabetter)

CREDIT CARD DEFAULT PREDICTION

- Introduction
- Data Pipeline
- Data Summary
- Data Exploration
- Visualization and Analysis
- Feature Engineering
- Handling Class Imbalance
- Classification Modeling and Model selection
- Technologies used

INTRODUCTION

A **credit card** is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's accrued debt (i.e., promise to the card issuer to pay them for the amounts plus the other agreed charges). The card issuer (**usually a bank or credit union**) creates a revolving account and grants a line of **credit to the cardholder**, from which the cardholder can **borrow money** for payment to a merchant or as a **cash advance**. There are two credit card groups: consumer credit cards and business credit cards

A **Taiwan-based credit card issuer** wants to better **predict the likelihood of default** for its customers, as well as identify the **key drivers** that determine this likelihood.

This would inform the issuer's **decisions** on who to **give a credit card** and what **credit limit** to provide. It would also help the issuer have a better understanding of their current and **potential customers**, which would inform their **future strategy**, including their **planning** of offering **targeted credit products** to their customers.

DATA PIPELINE

- **EDA** : In this part, we do some exploratory data analysis (EDA) on the data to see how is the data.
- **Data cleaning** : In this part we have removed or replaced unnecessary data. Since there were some data with null or unwanted values.
- **Data visualization** : In this part we visualize and analyze the data from which we get the trend and relation between features which is useful for prediction.
- **Feature Engineering:** We converted multiple features into single feature and removed unnecessary features.
- **Data modeling** : In this part we trained and predicted data in different classification models for getting final model.
- **Model Selection** : In this part we selected model which gives best result.

THE DATA

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars :it includes both the individual credit and his/her family (supplementary) credit.
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

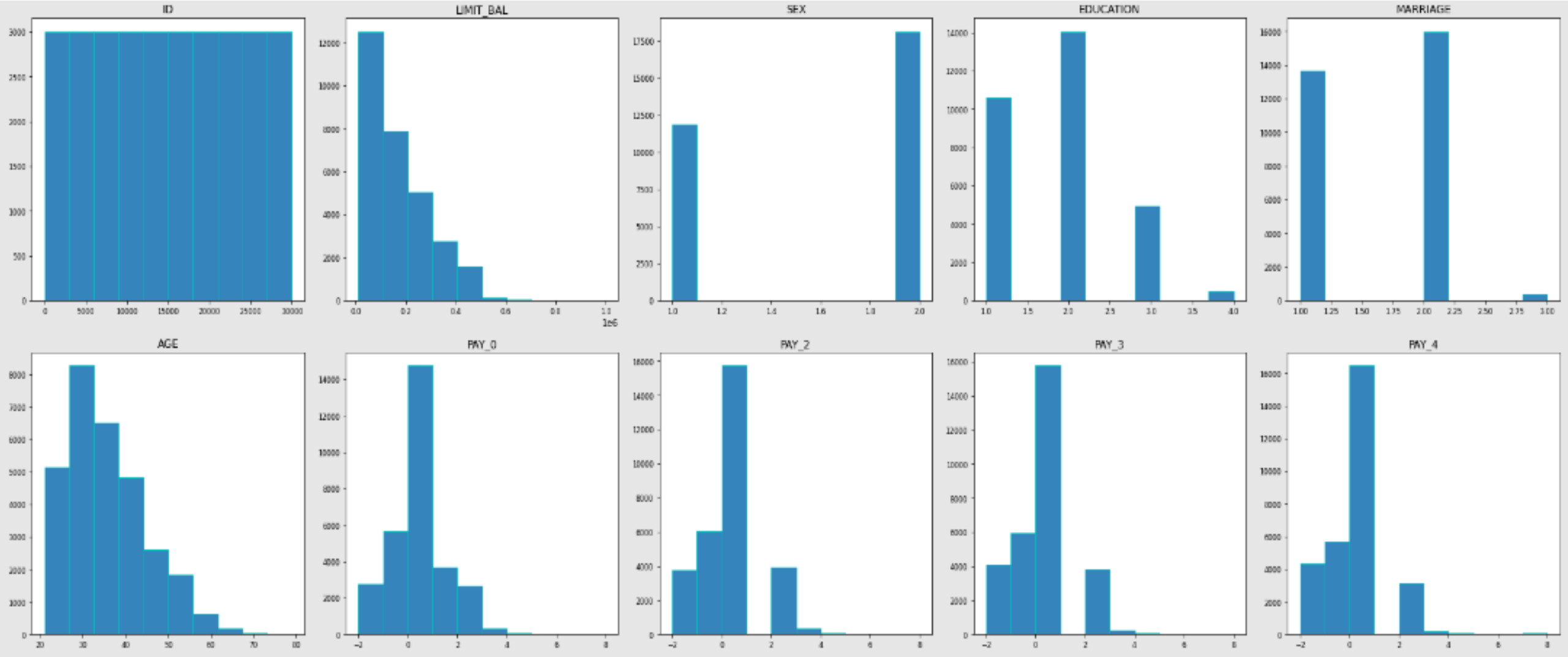
THE DATA

- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)

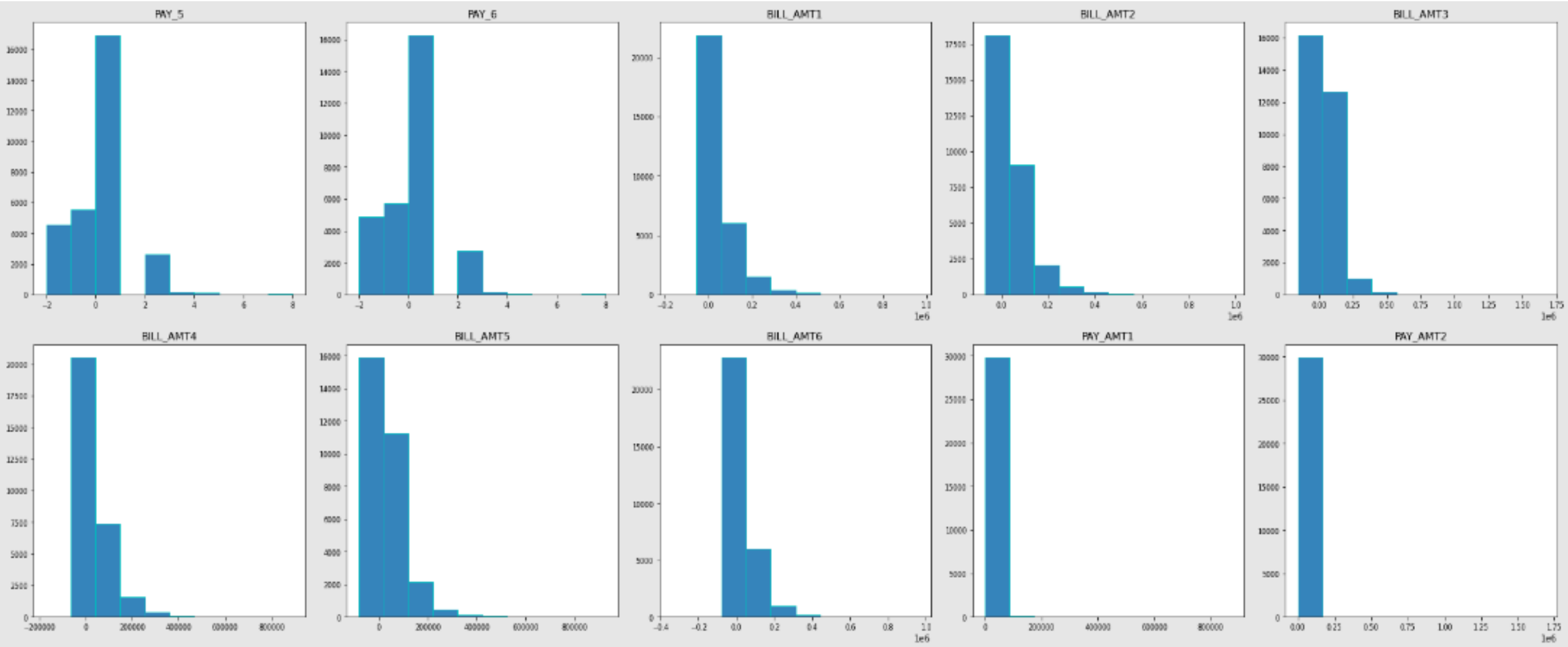
THE DATA

- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of payment in April, 2005 (NT dollar)
- **default payment next month**: The target variable indicating default of payment (1=default, 0=non-default)

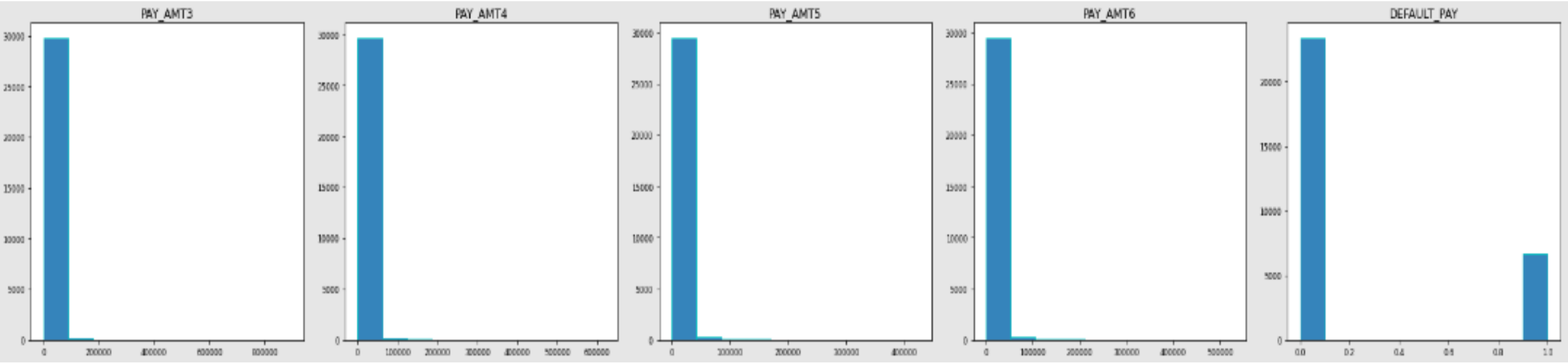
DISTRIBUTION OF DATA FOR CREDIT CARD DEFAULT DATAFRAME



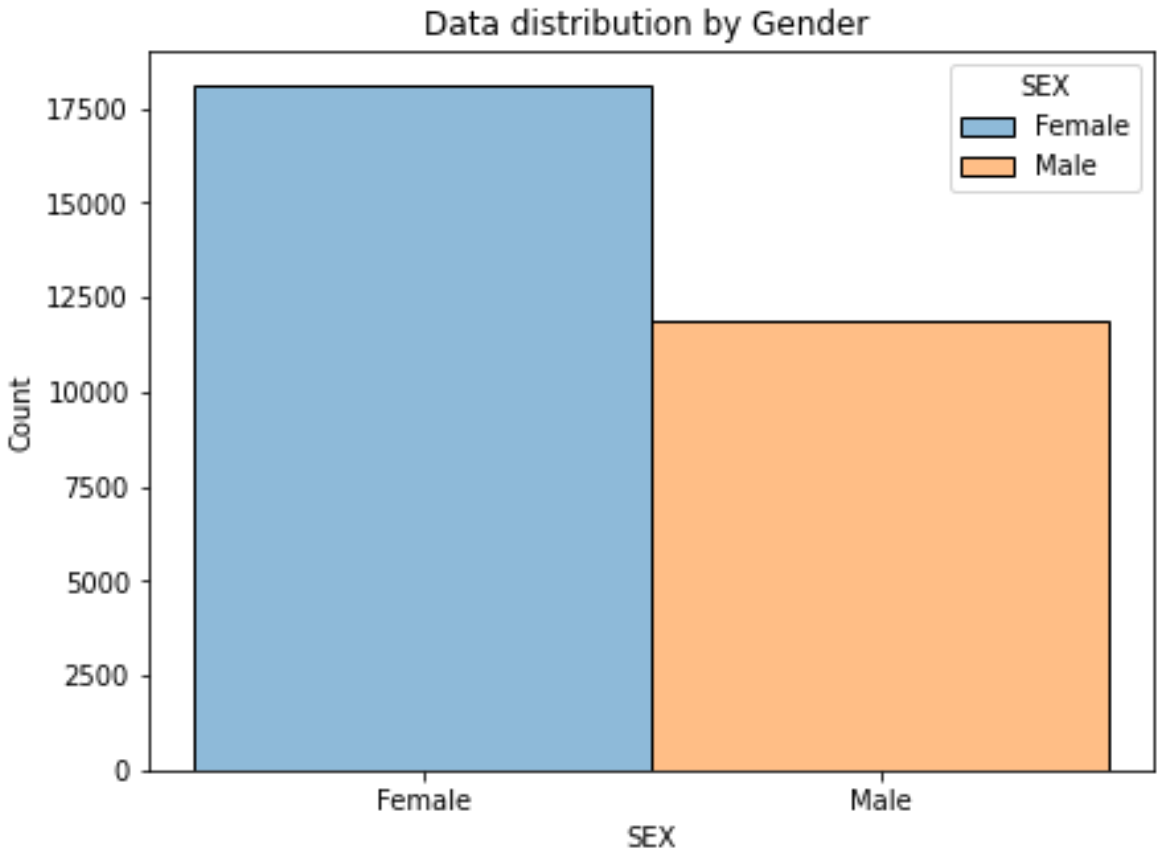
DISTRIBUTION OF DATA FOR CREDIT CARD DEFAULT DATAFRAME

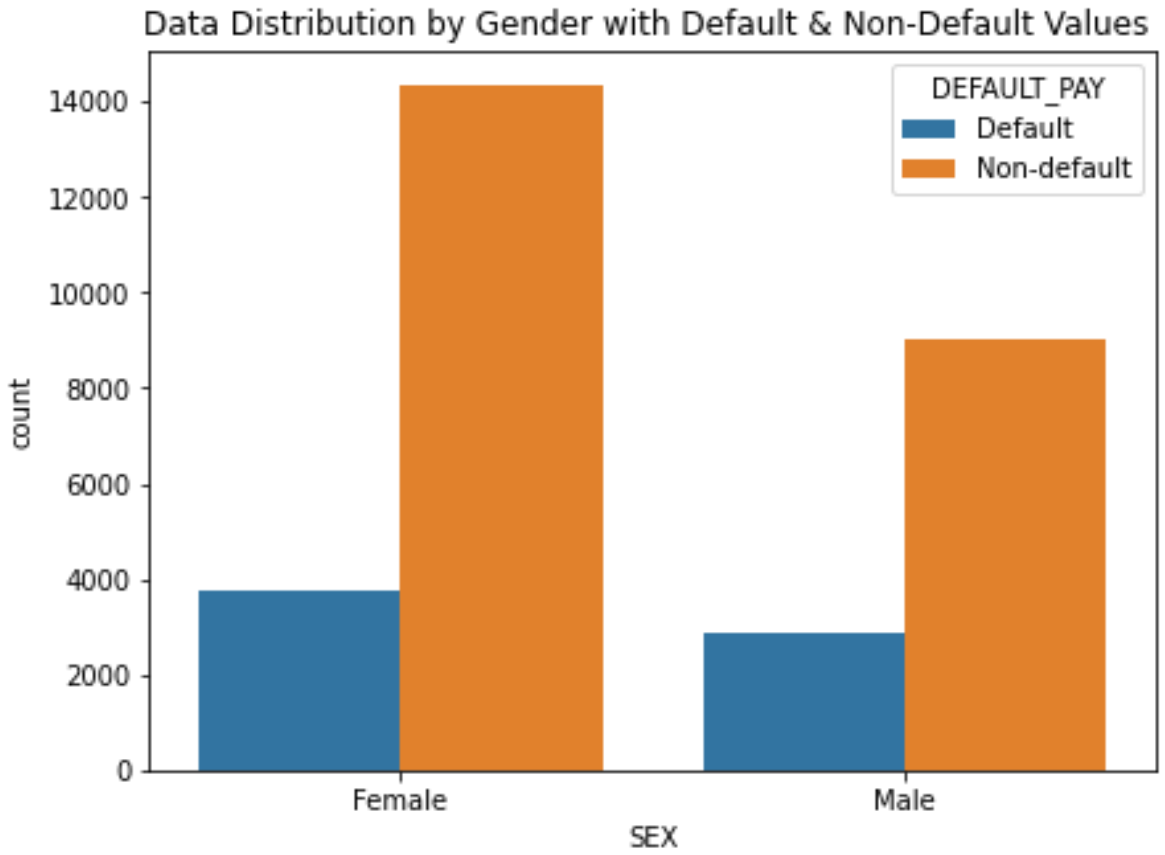


DISTRIBUTION OF DATA FOR CREDIT CARD DEFAULT DATAFRAME

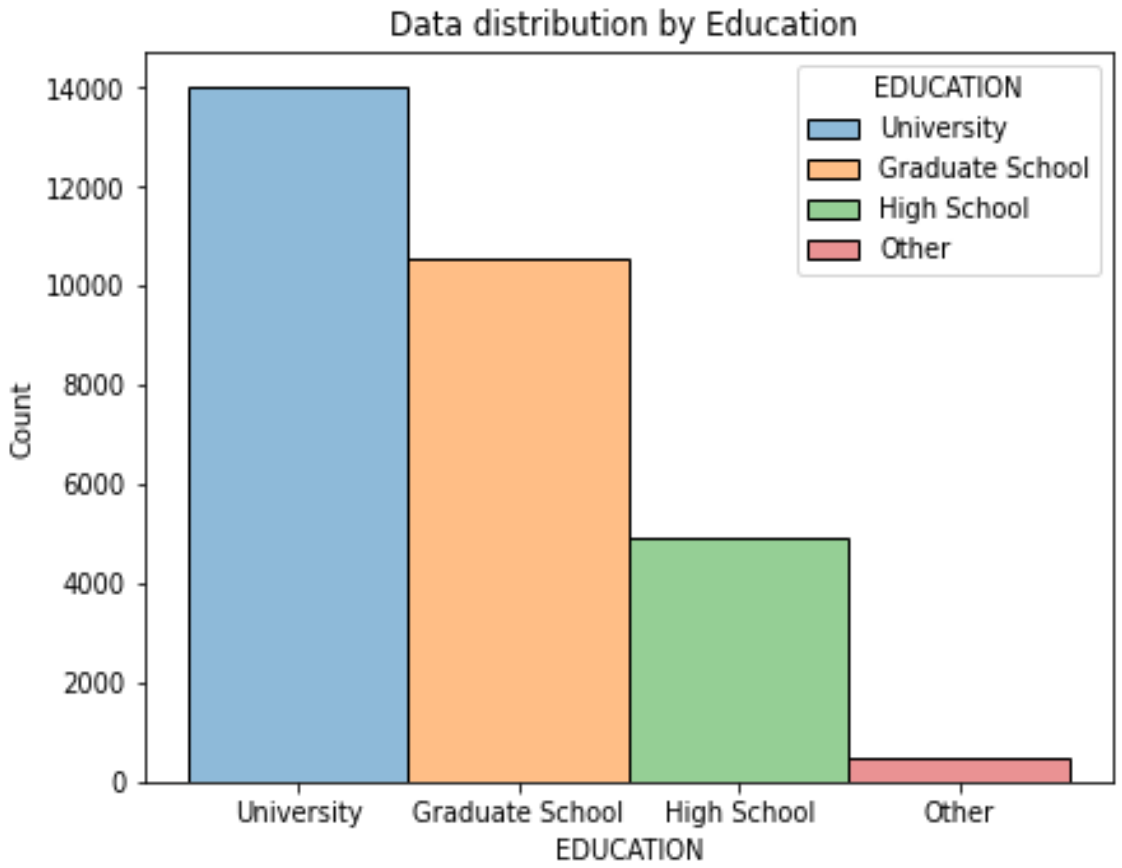


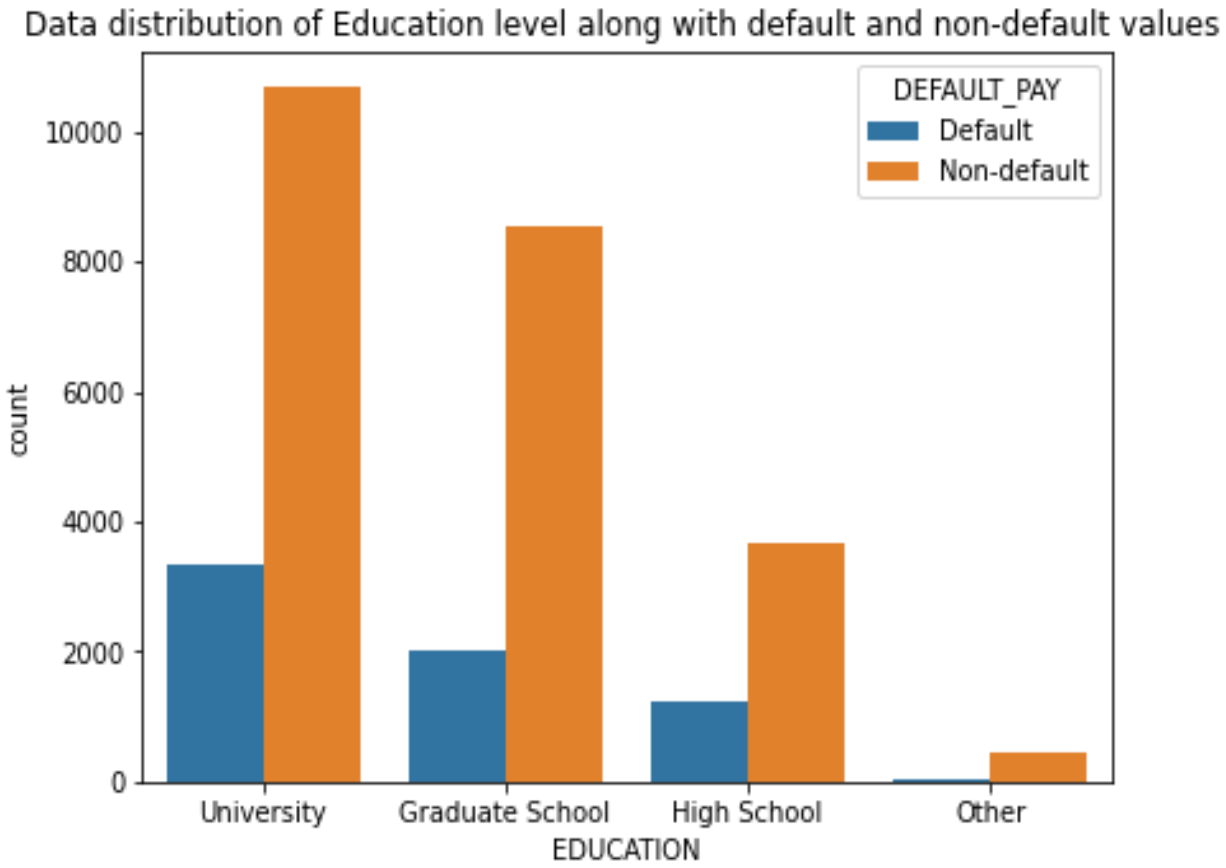
Visualisation & Analysis



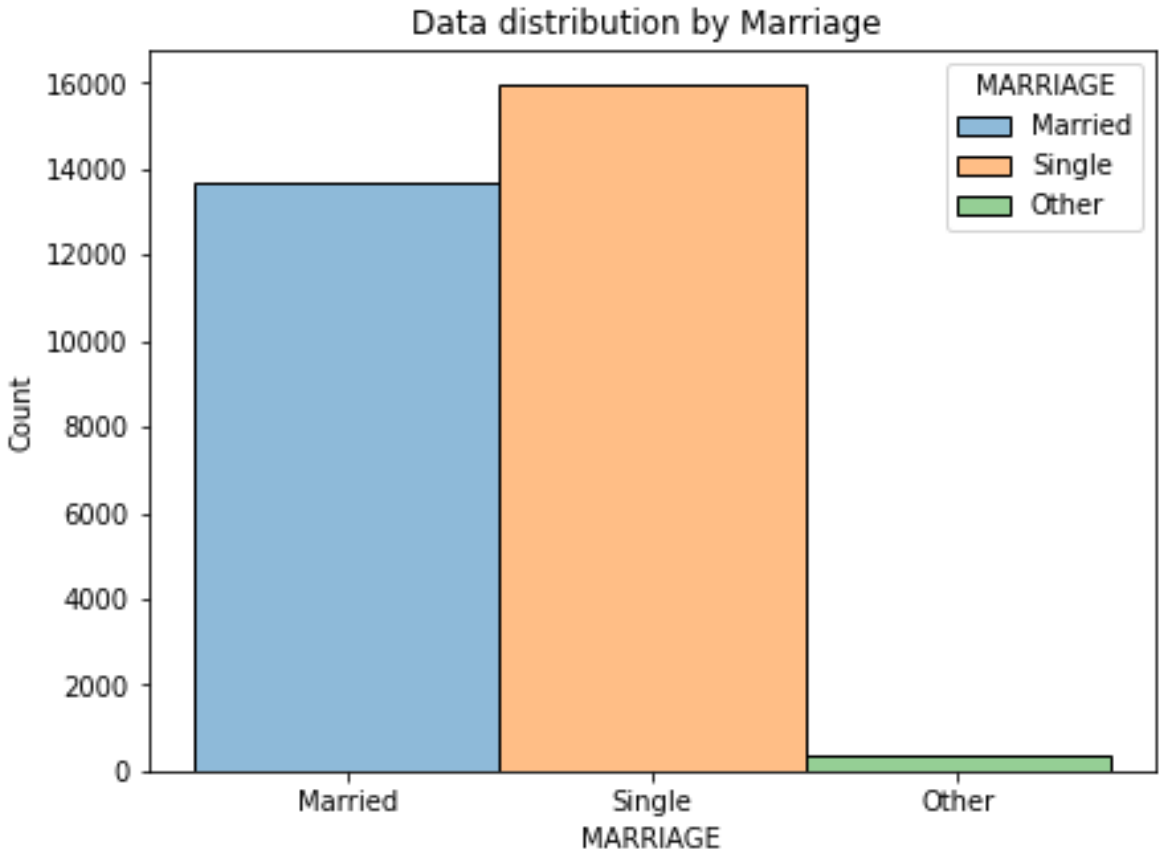


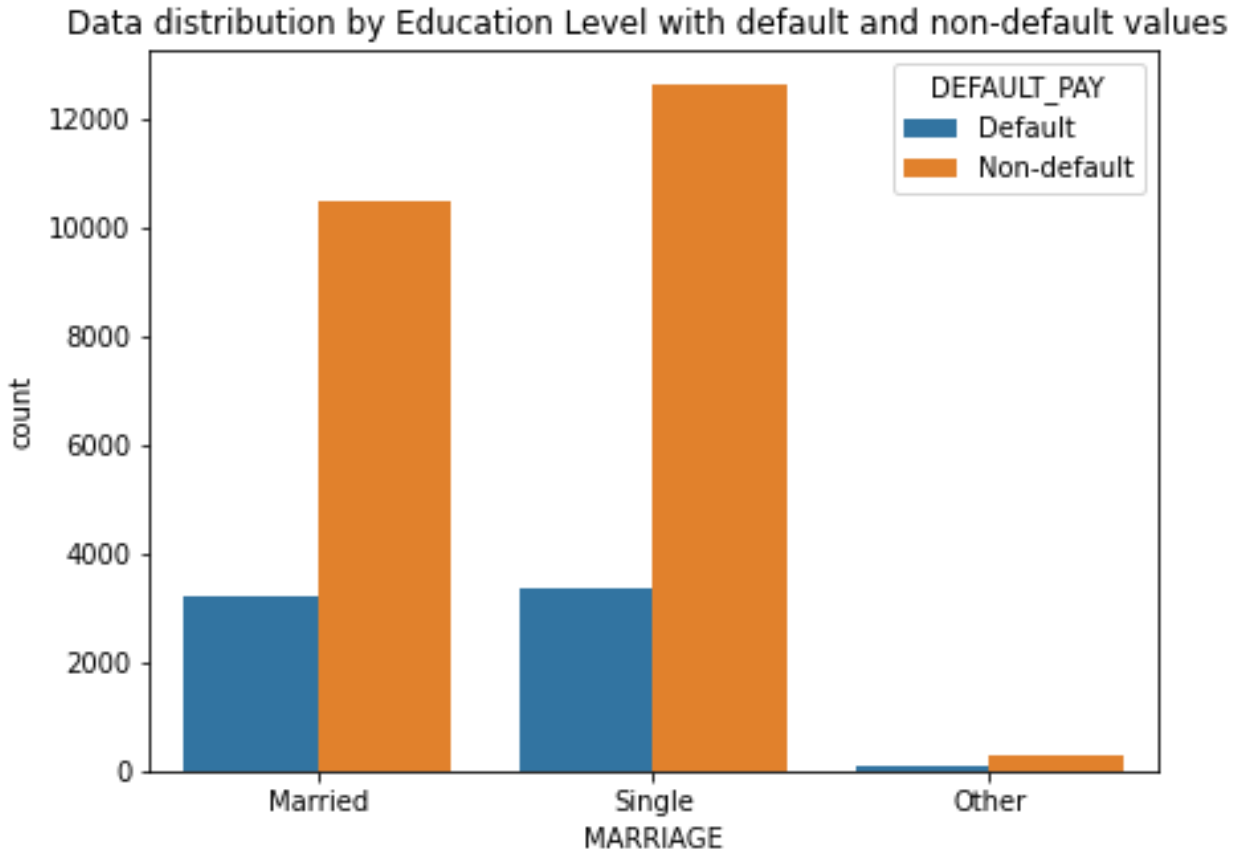
Visualisation & Analysis



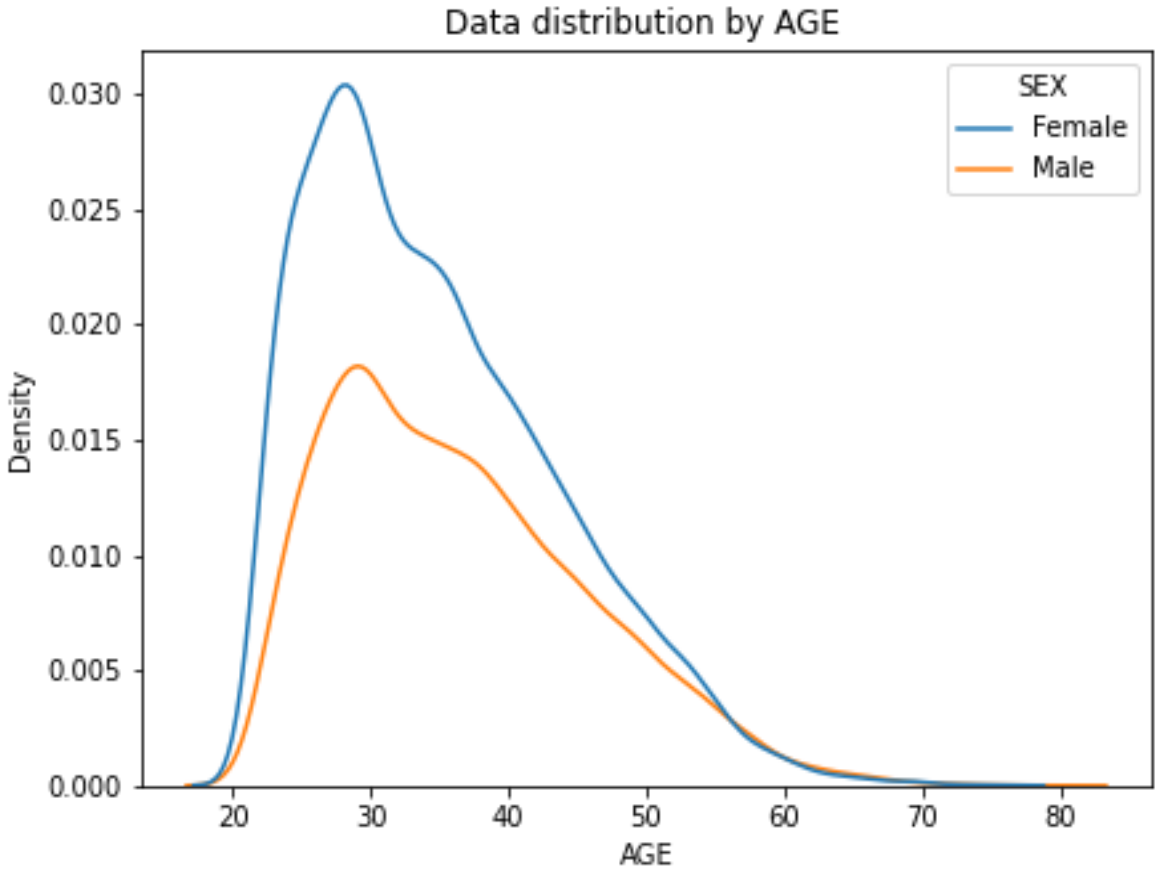


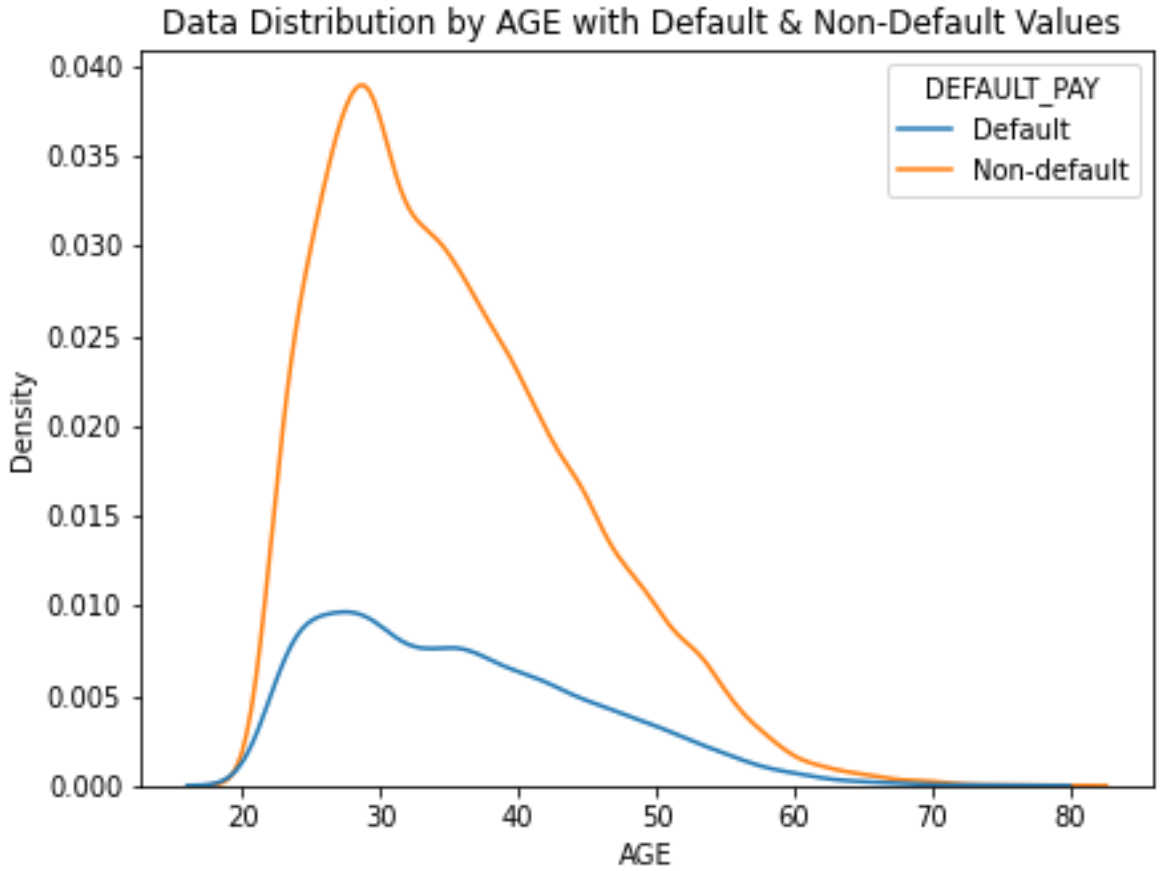
Visualisation & analysis



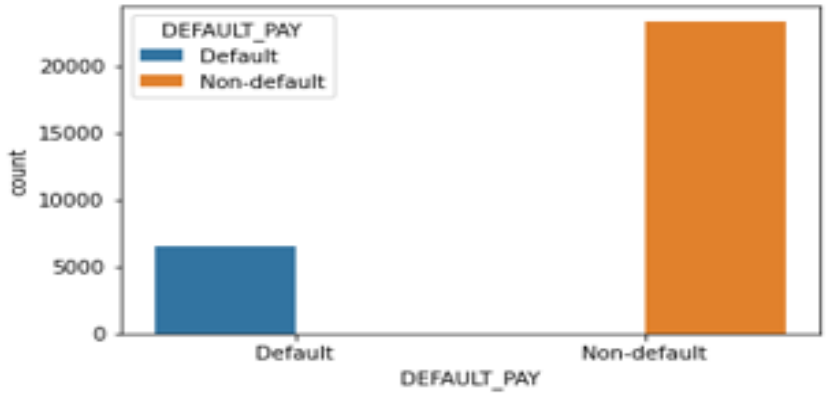


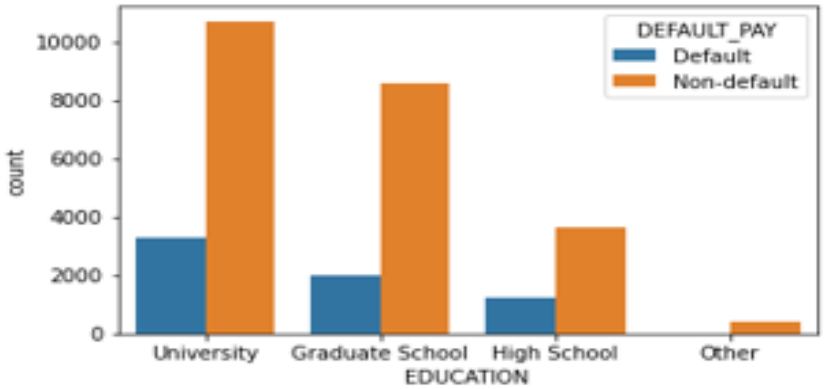
Visualisation & analysis

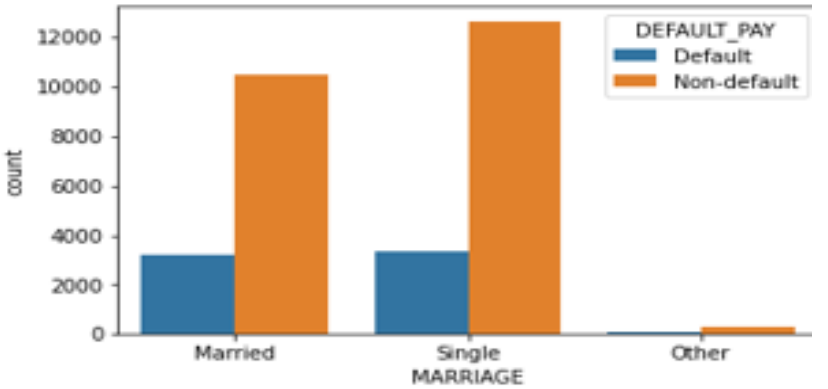


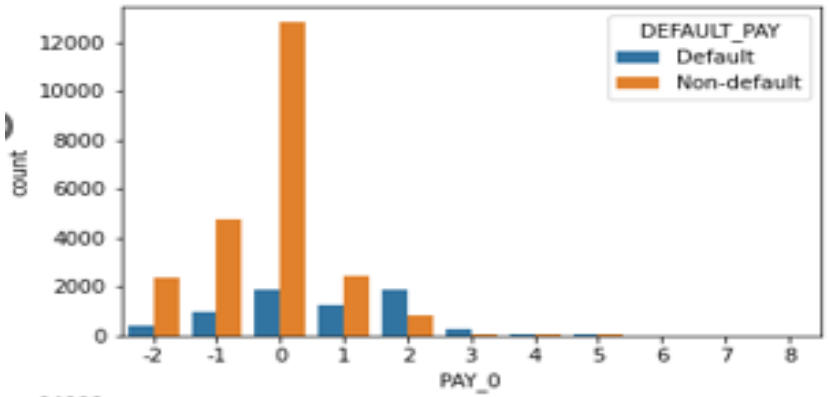


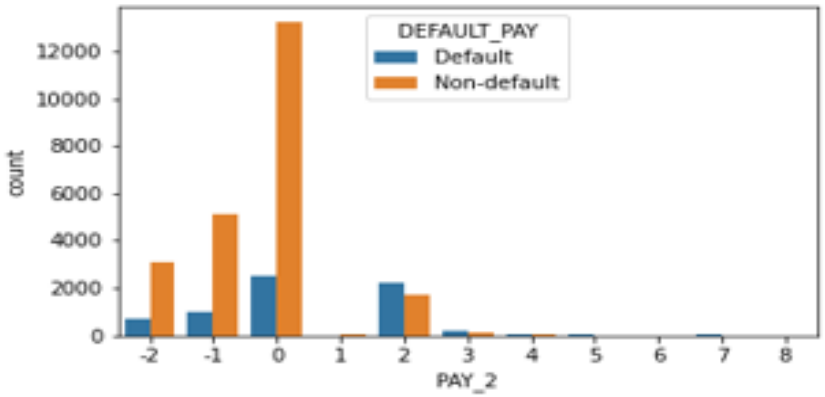
DATAPOINT DISTRIBUTION OF DEFAULT & NON DEFAULT VS SOME FEATURES

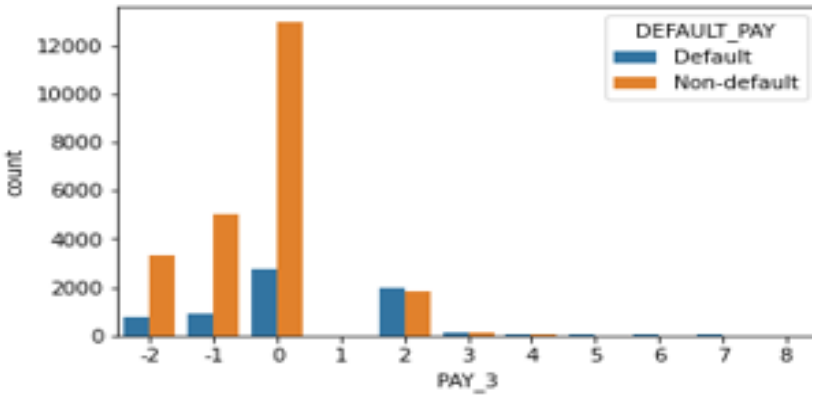


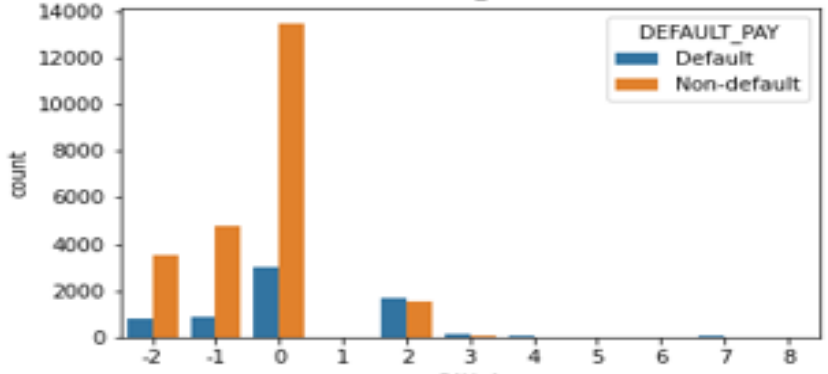


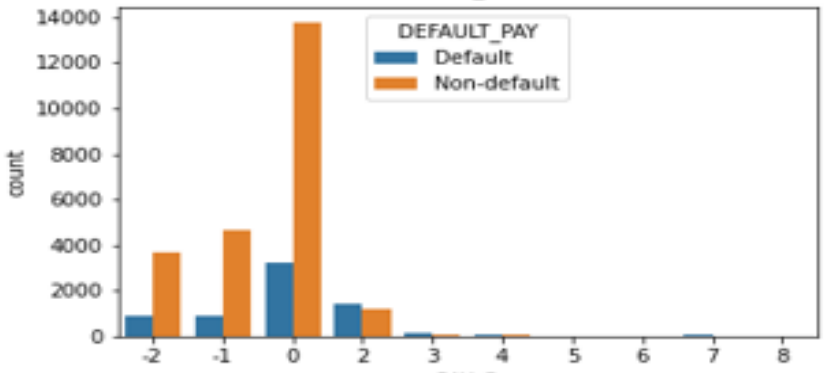


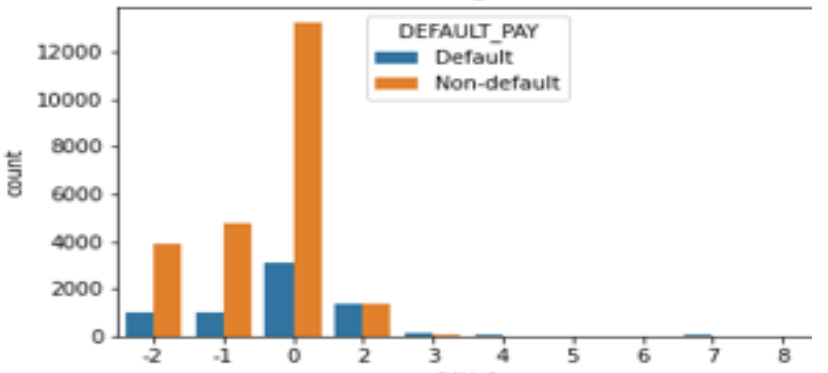












FEATURE ENGINEERING

- We have bill amount and pay amount from April to September so let us make columns having remaining payment (unpaid amount) for each month.
- We are creating list of months pending as :

['april_pending','may_pending','june_pending','july_pending',
'august_pending','september_pending']

where ,

april_pending = Bill amount of April - Pay amount of April

may_pending = Bill amount of May - Pay amount of May

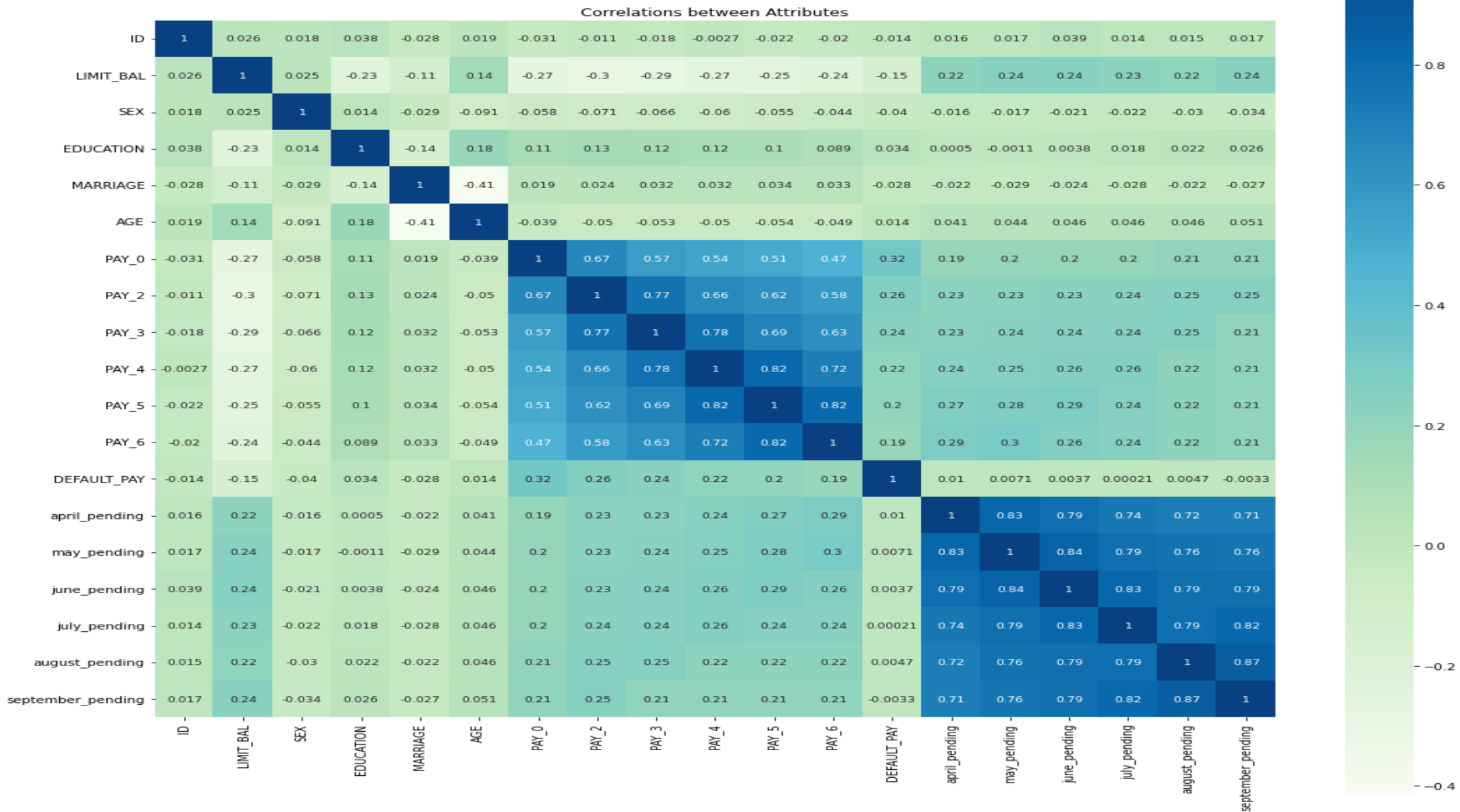
june_pending = Bill amount of June - Pay amount of June

july_pending = Bill amount of July - Pay amount of July

august_pending = Bill amount of August - Pay amount of August

september_pending = Bill amount of September - Pay amount of September

CORRELATION AFTER FEATURE ENGINEERING



HANDLING CLASS IMBALANCE

- As value count of our target variable (DEFAULT_PAY) is:

  0 - 23364

  1 - 6636

  where , 0→Non default

  1→Default

- We can see there is class Imbalance, 0 count is way more than 1 count. It could lead to bias in prediction.
- So we are using SMOTE() oversampling for balancing classes.
- After balancing :

  Original dataset shape - 30000

  Resampled dataset shape - 46728

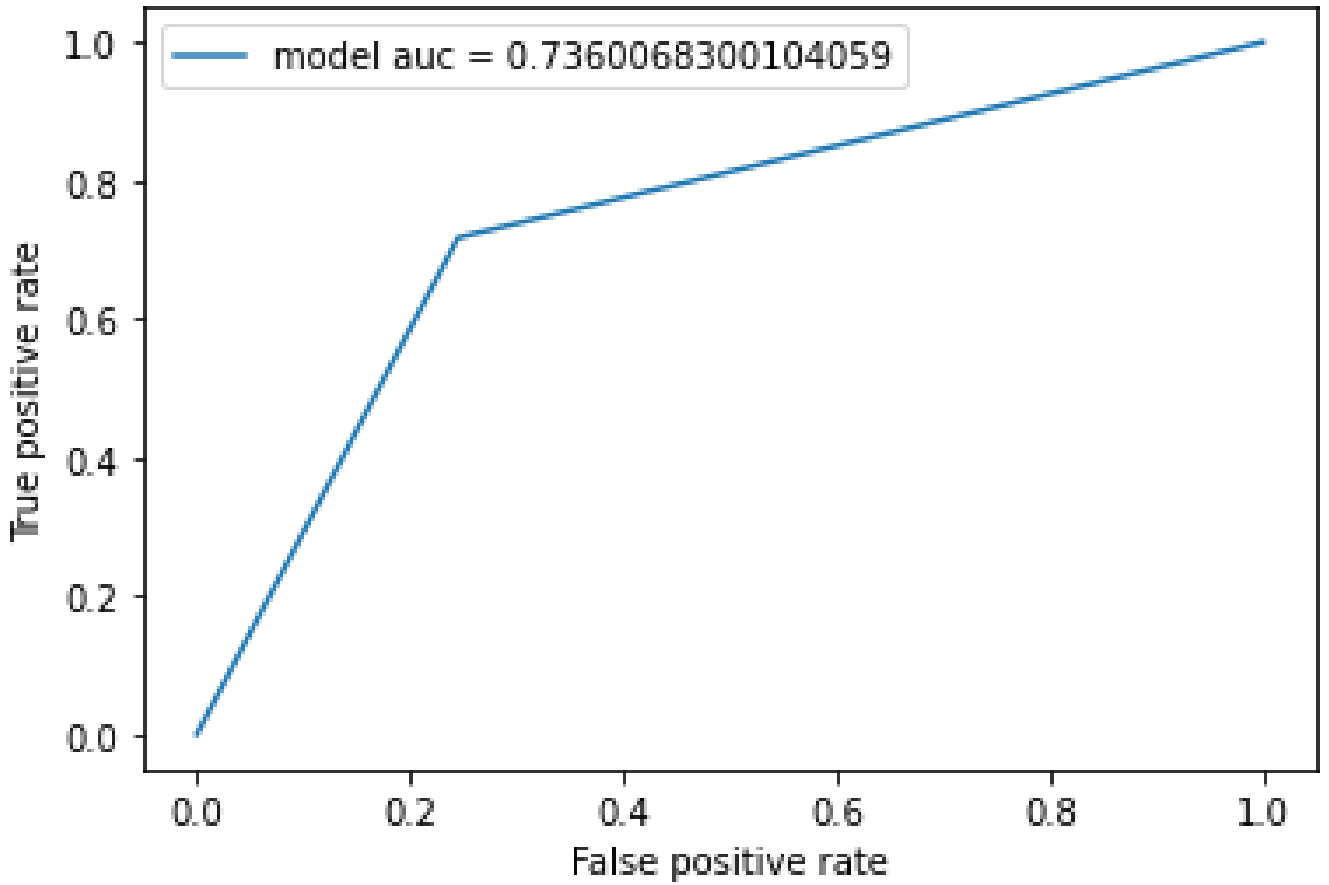
We used some models and find which model is best fitting with data.

Models we used are given below.

- KNN Classifier
- XGBoost
- Logistic Regression
- SVM(Support Vector Machine)
- Naive Bayes Classifier

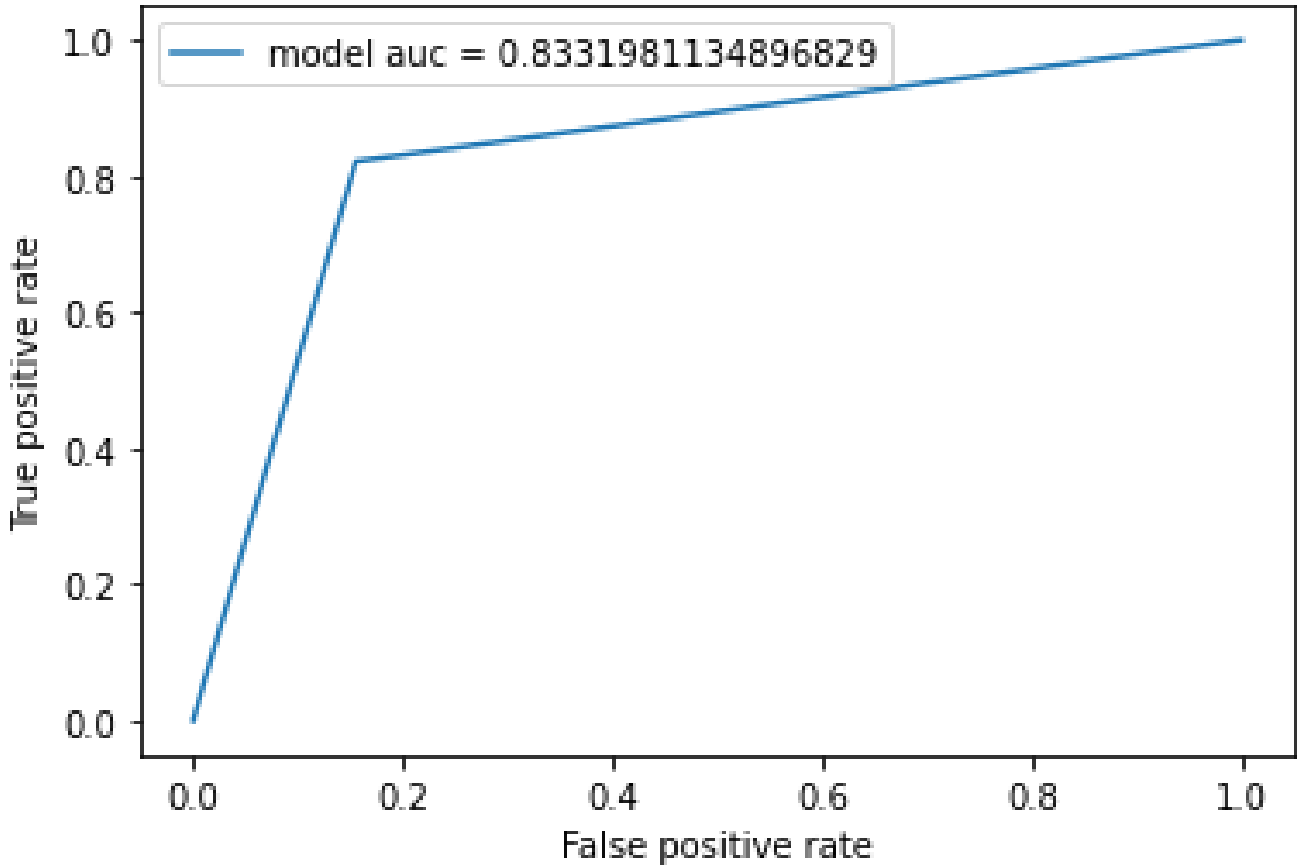
KNN CLASSIFIER

- Accuracy for train dataset:- 73.29%
- Accuracy for test dataset:- 73.58%
- ROCAUC score:- 0.7360068300104



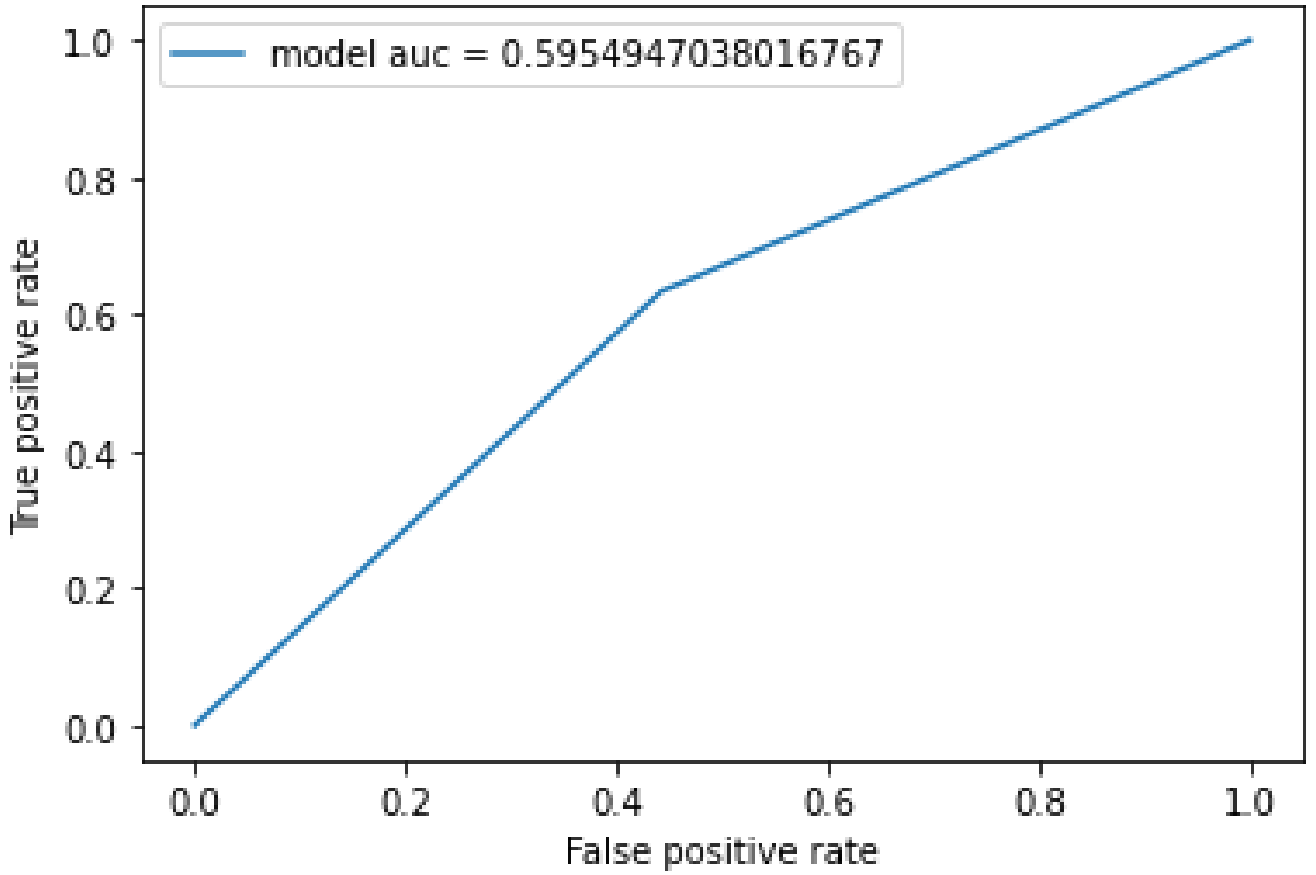
XGBOOST

- Accuracy for train dataset:- 99.95%
- Accuracy for test dataset:- 83.31%
- ROCAUC score:- 0.8331981134896



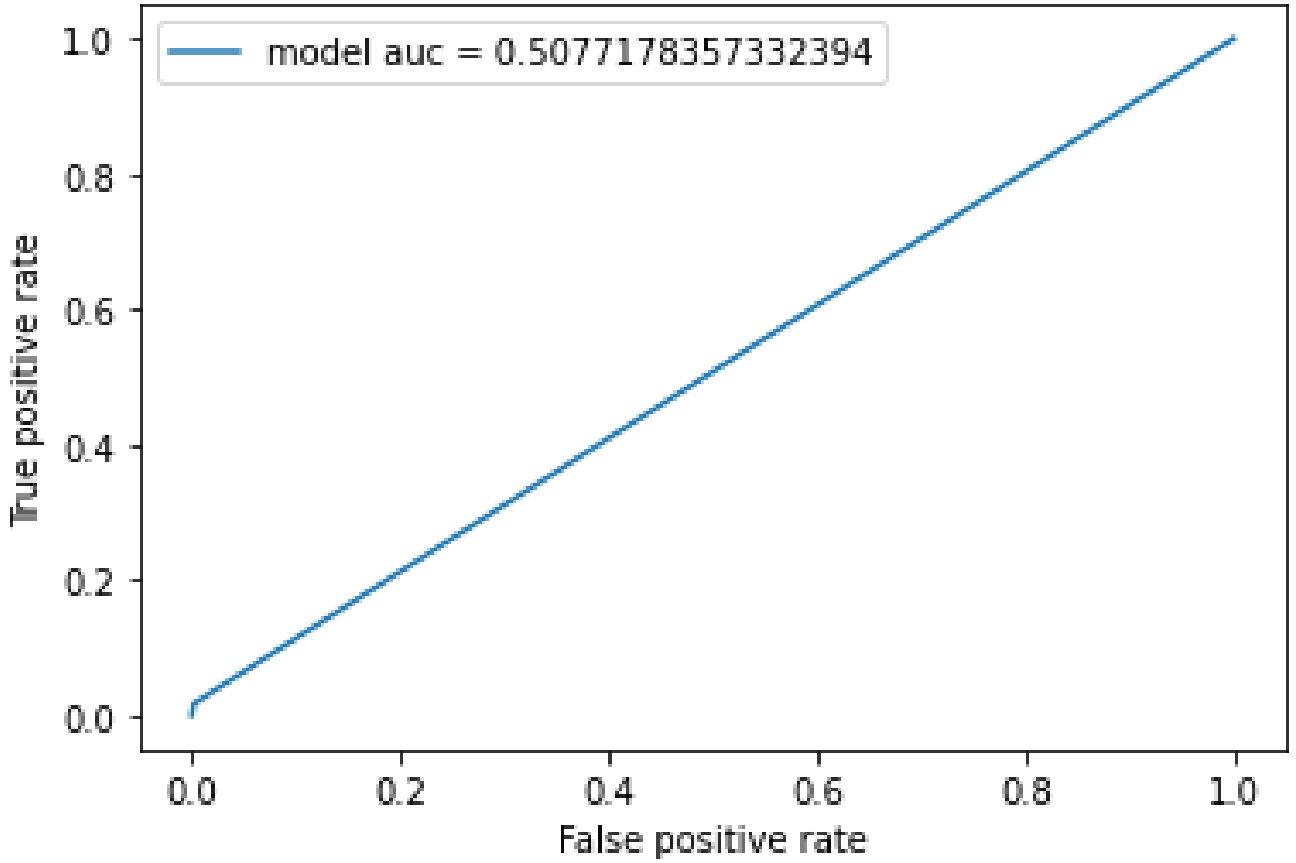
LOGISTIC REGRESSION

- Accuracy for train dataset:- 59.01%
- Accuracy for test dataset:- 59.59%
- ROCAUC score:- 0.6055282909



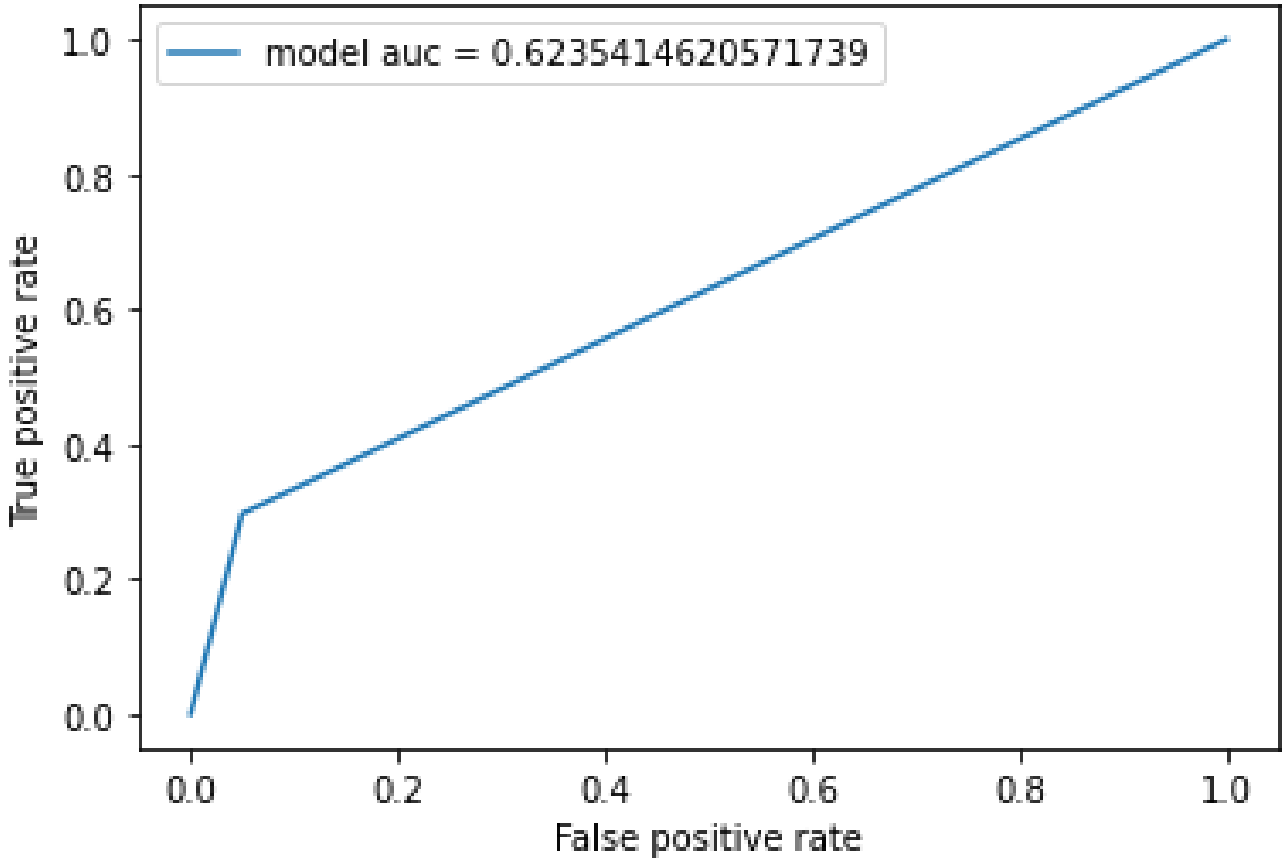
SUPPORT VECTOR MACHINE

- Accuracy for train dataset:- 100.00%
- Accuracy for test dataset:- 50.29%
- ROCAUC score:- 0.5077178357332

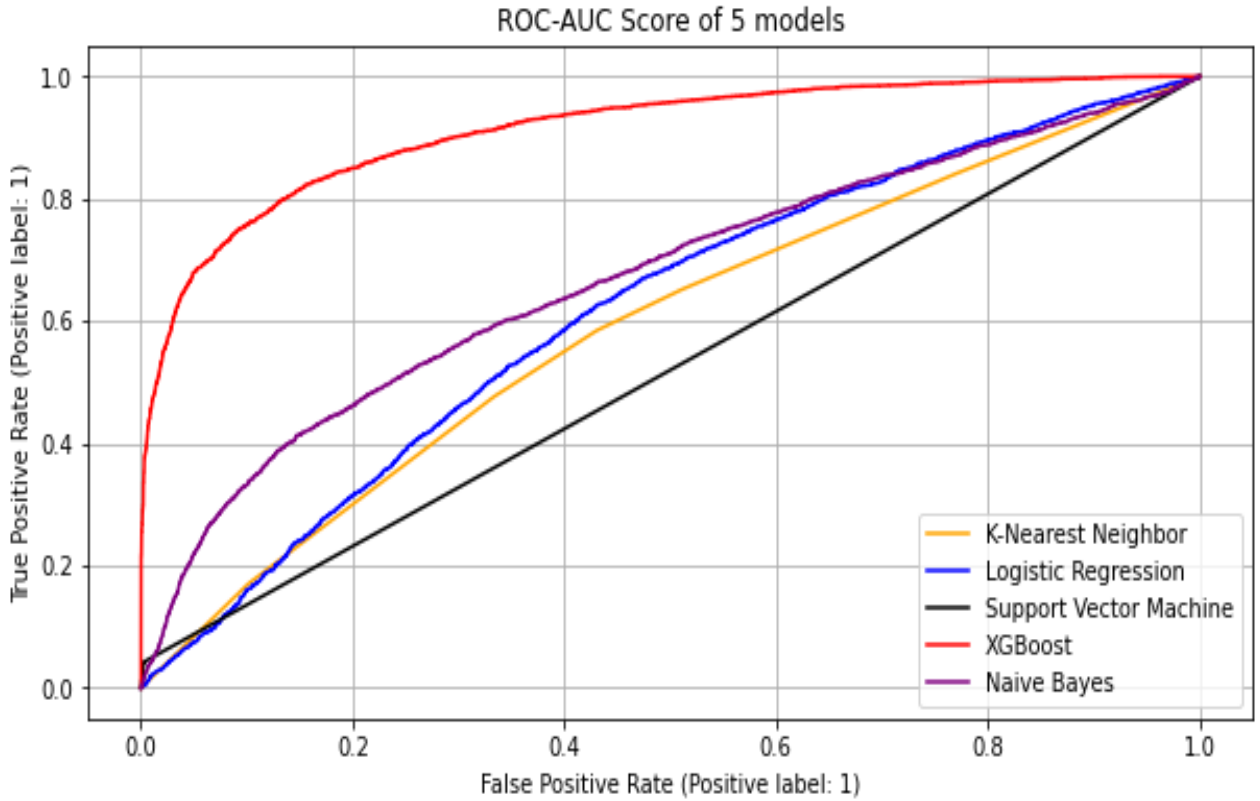


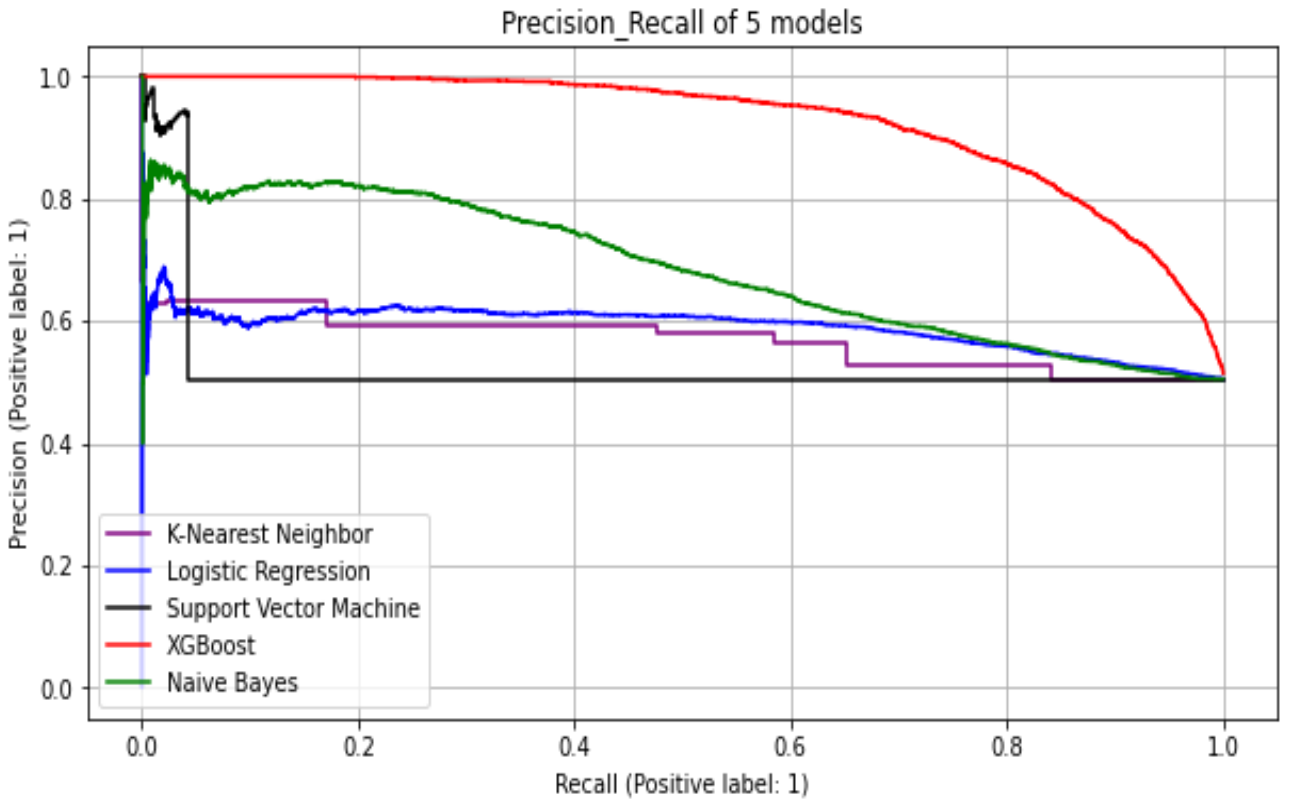
NAIVE BAYES CLASSIFIER

- Accuracy for train dataset:- 80.92%
- Accuracy for test dataset:- 79.98%
- ROCAUC score:- 0.62354146



MODEL COMPARISION





CONCLUSION

Model Name	Accuracy	Roc Auc	Precision	Recall	F1
KNN	73.29%	0.736	0.749	0.717	0.733
XGBoost	99.95%	0.833	0.844	0.822	0.833
Logistic Regression	59.01%	0.595	0.593	0.634	0.613
SVM	100%	0.507	0.91	0.017	0.034
Naïve Bayes	80.92%	0.623	0.641	0.296	0.405

- After applying all models here we can see that XGboost performed better as compared to other models as its accuracy and roc auc score and all other metrics are good.
- So we are considering XGboost Model for future predictions.

TECHNOLOGY USED

- Python(Programming Language)
- Libraries used:1) Pandas
  2) Numpy
  3) Matplotlib
  4) Seaborn
  5) Sklearn

THANK YOU