Capstone Project

Credit Card Default Prediction

By
NIDHI PANDEY

AlmaBetter

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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

Content

- Problem Statement
- Data Pipeline
- Data Description
- Handling Null values
- Exploratory Data Analysis
- Hypothesis Testing
- Models Performed
- ROC AUC for all the models
- Model Validation & Selection
- Evaluation Matrix of All the models
- Challenges
- Conclusion

Data Pipeline

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.

- Data Processing: In this part we went through each attributes and encoded the categorical features.

- Model Creation: Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.

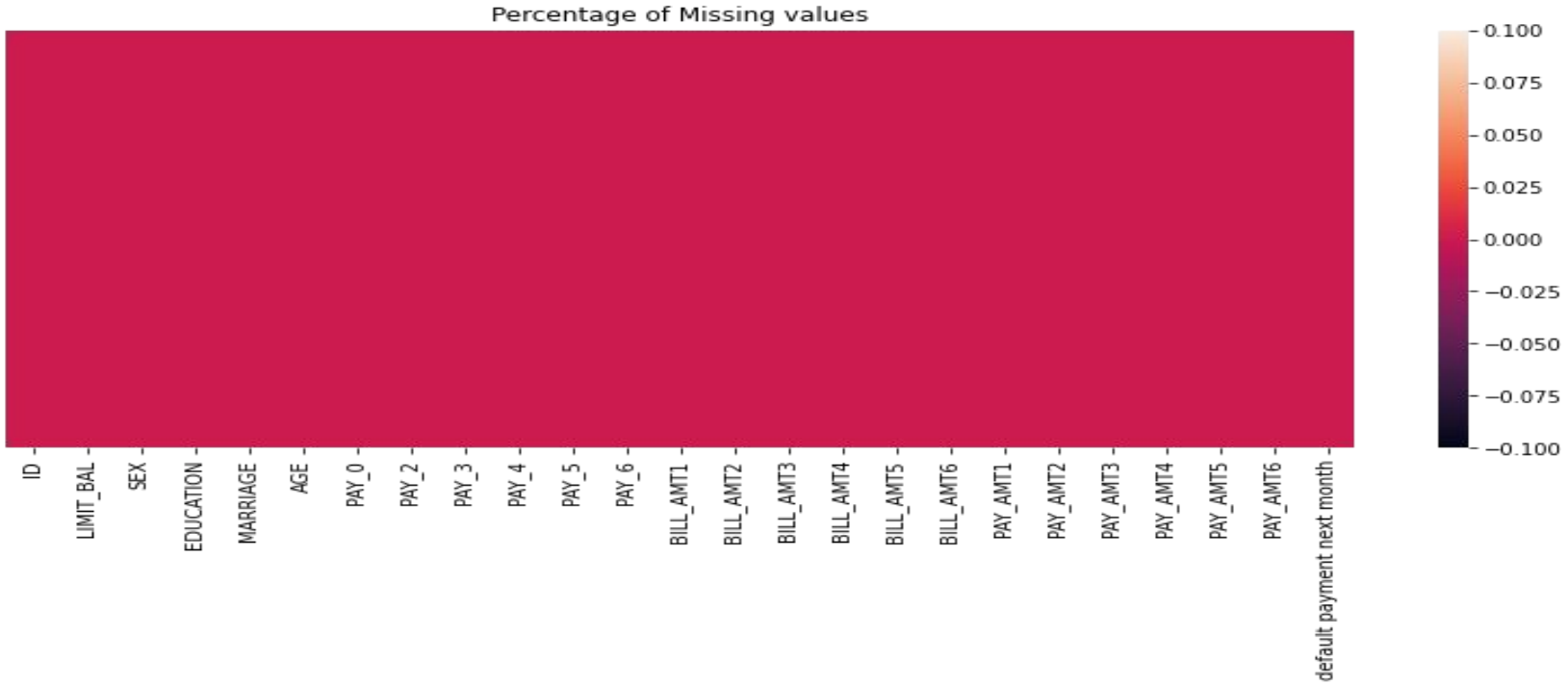
Data Description

We have considered 23 variables as explanatory variables:

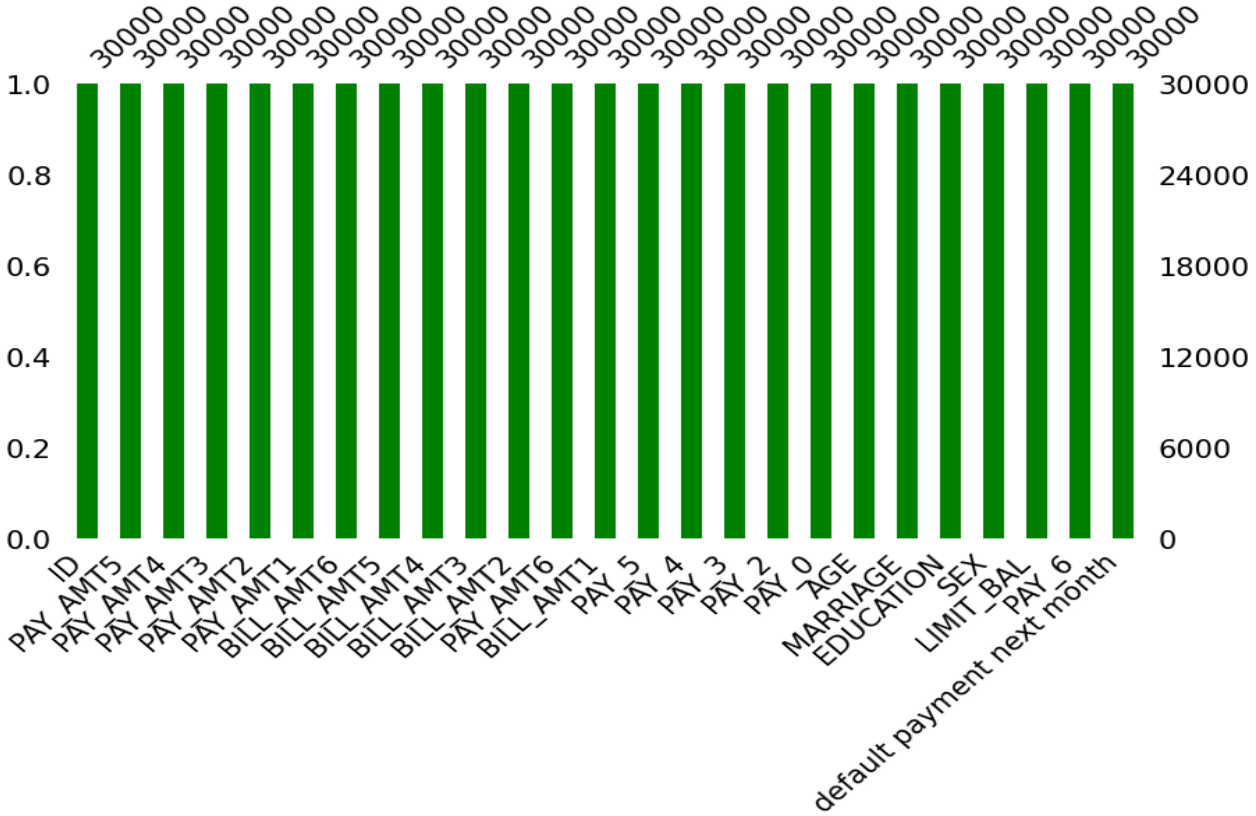
- X1: Amount of the given credit -it includes both the individual consumer credit and his/her family credit
- X2: Gender
- X3: Education
- X4: Marital status
- X5: Age (year)
- X6 -X11: History of past payment (from April to September, 2005)
- X12-X17: Amount of bill statement (from April to September, 2005)
- X18-X23: Amount of previous payment (from April to September, 2005)

Handling Missing / Null / Duplicate Values

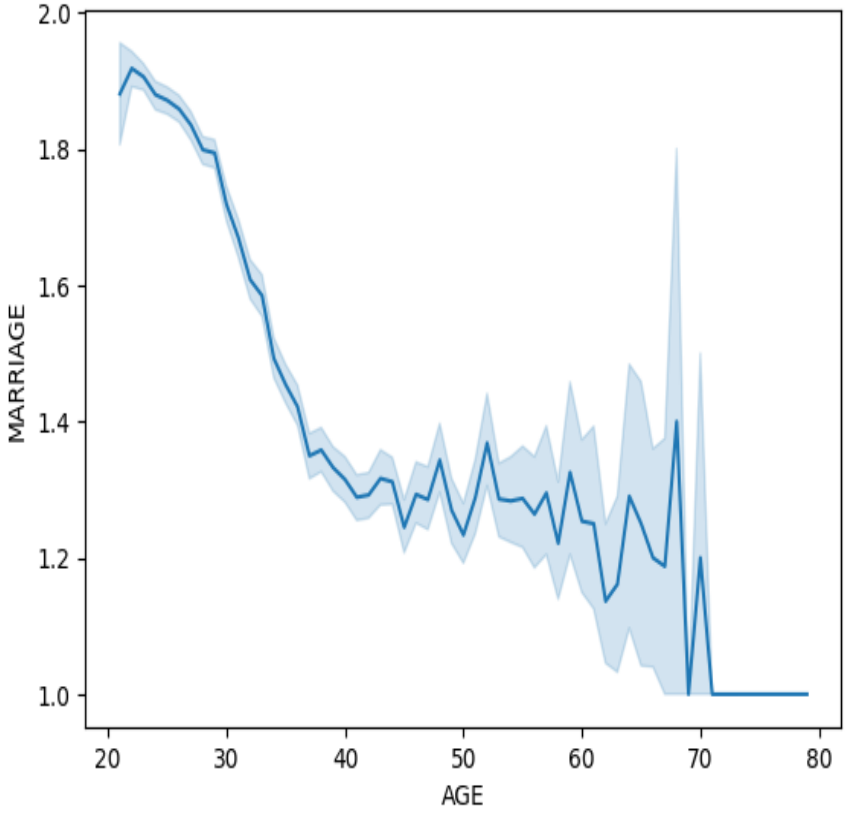
No missing or null values or duplicates are found in our dataset.

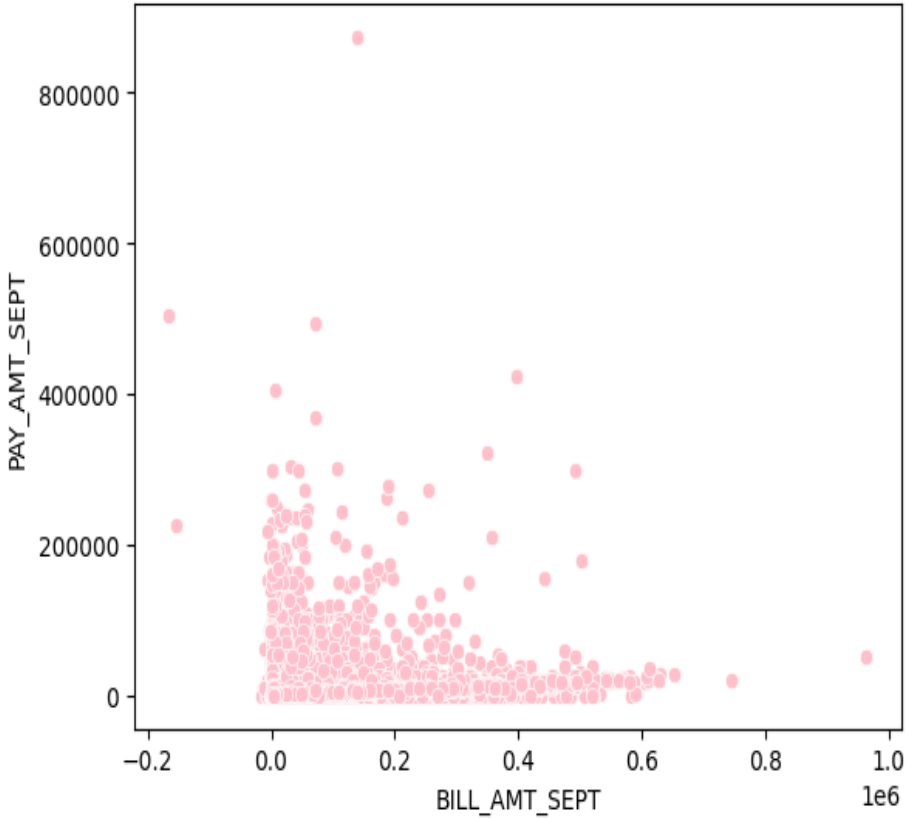


Handling Missing / Null / Duplicate Values Contd.

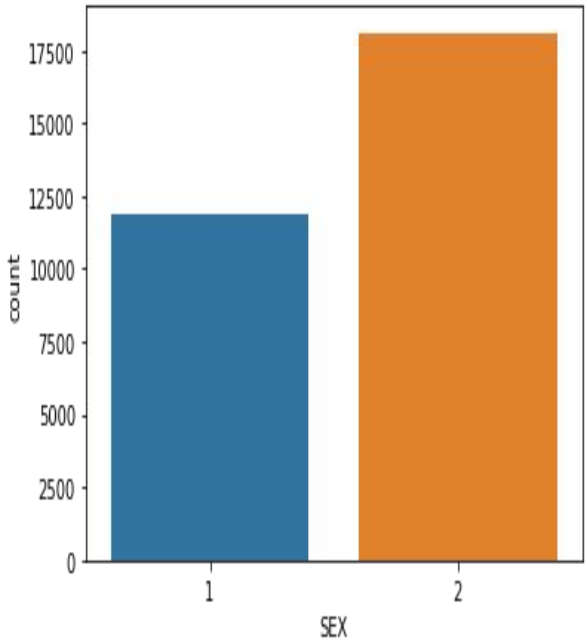


EDA- Dependent Variables

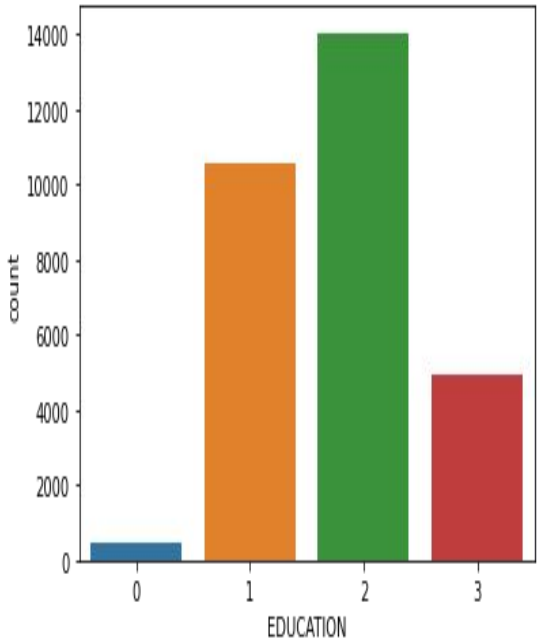




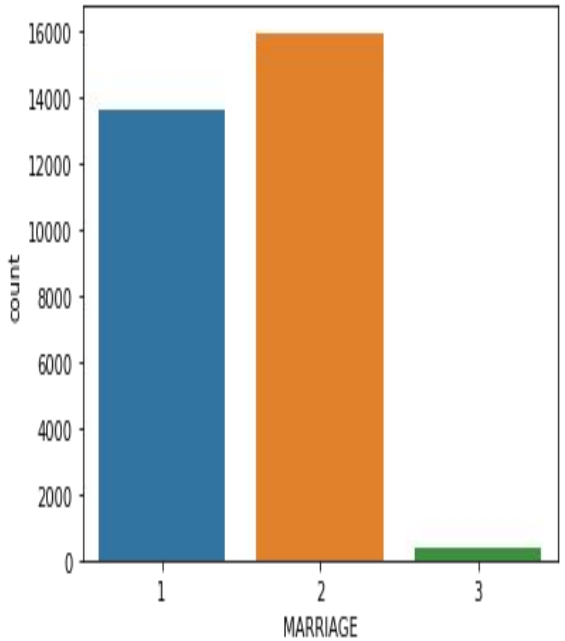
EDA- Independent Variables




```
1 : 'MALE'
2 : 'FEMALE'
```

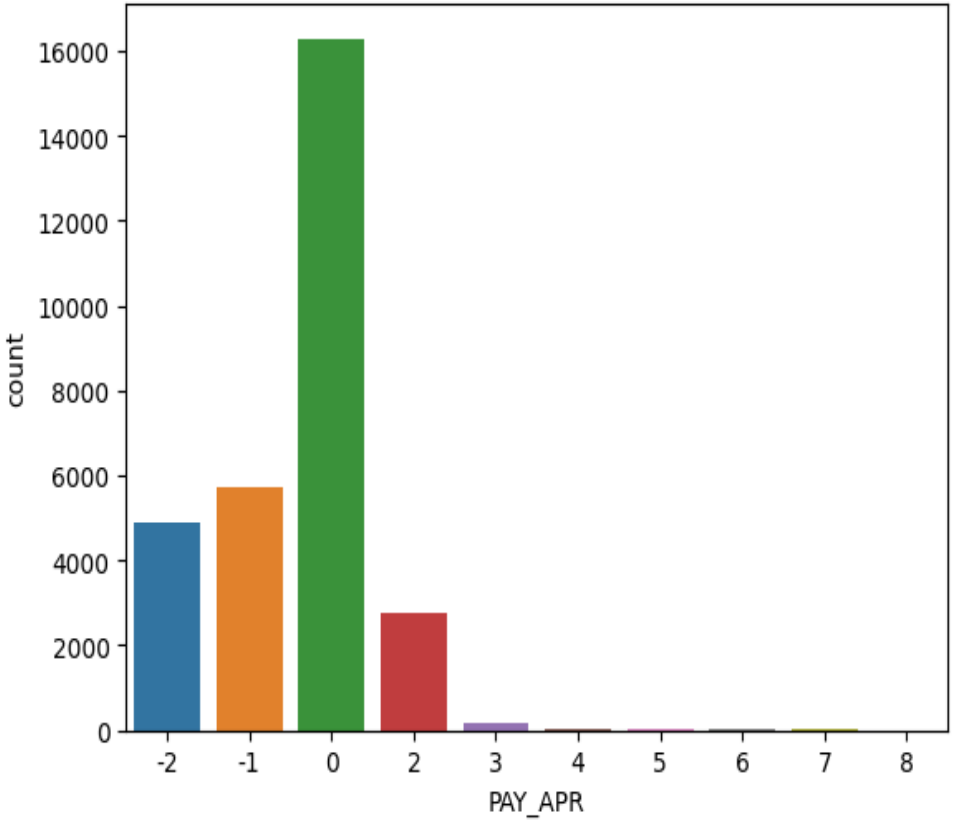


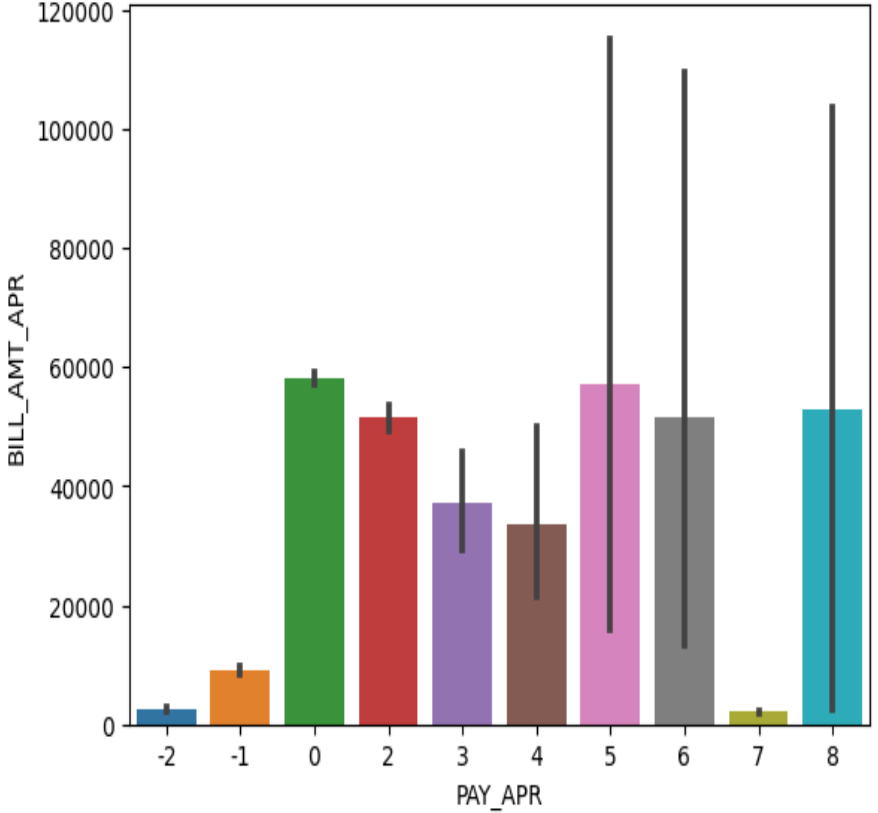

```
0 : 'graduate school'
1 : 'university'
2 : 'high school'
3 : 'others'
```



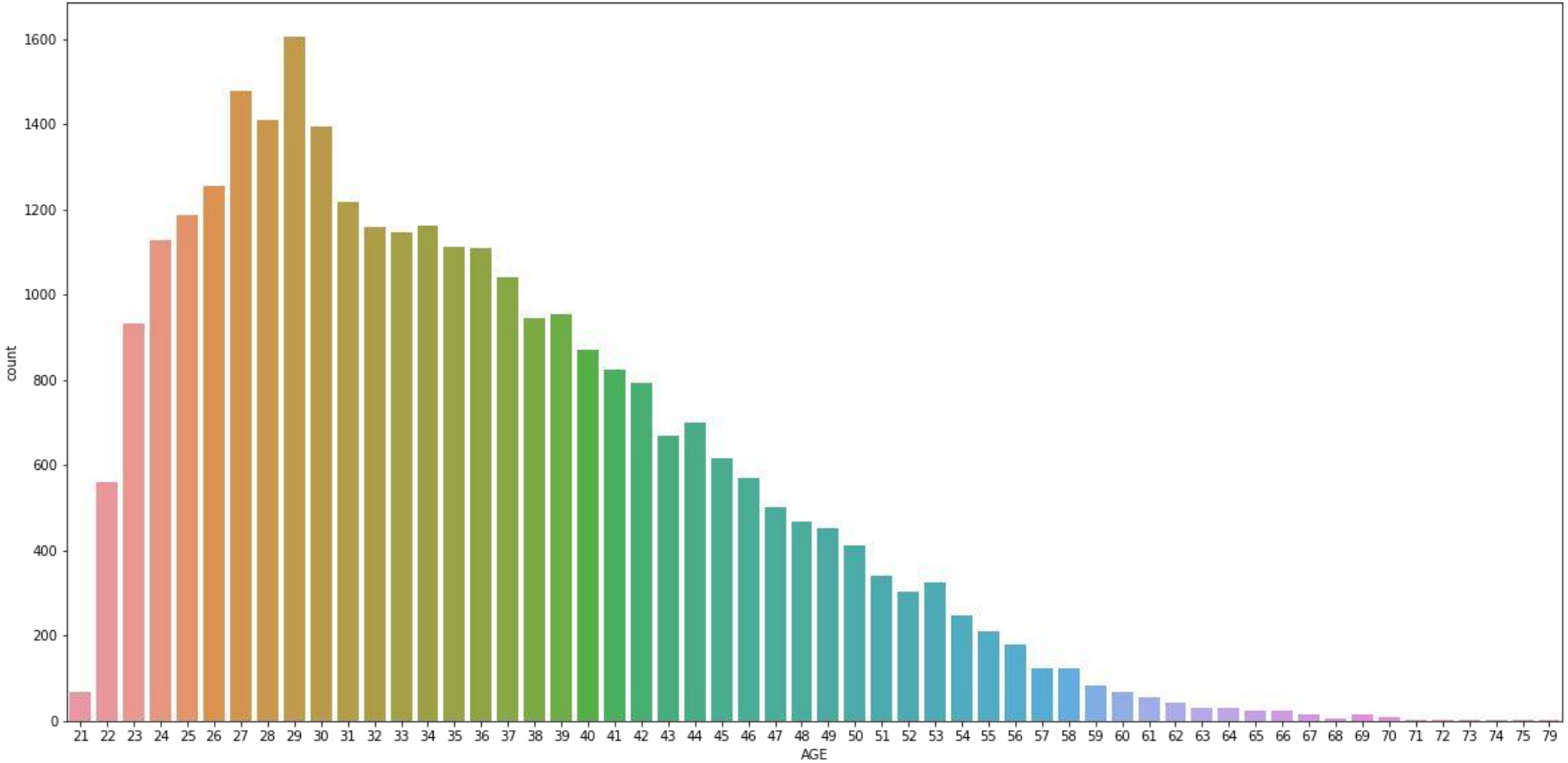

```
1 : 'married'
2 : 'single'
3 : 'others'
```

Visualization between BILL_AMT_APR and PAY_APR

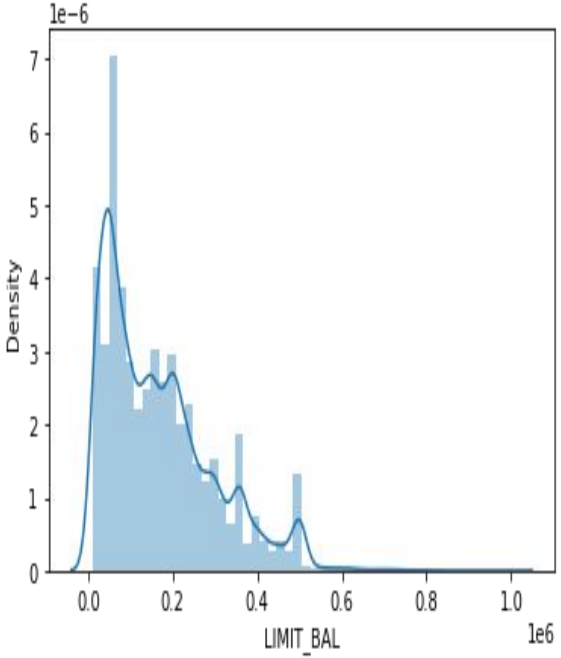


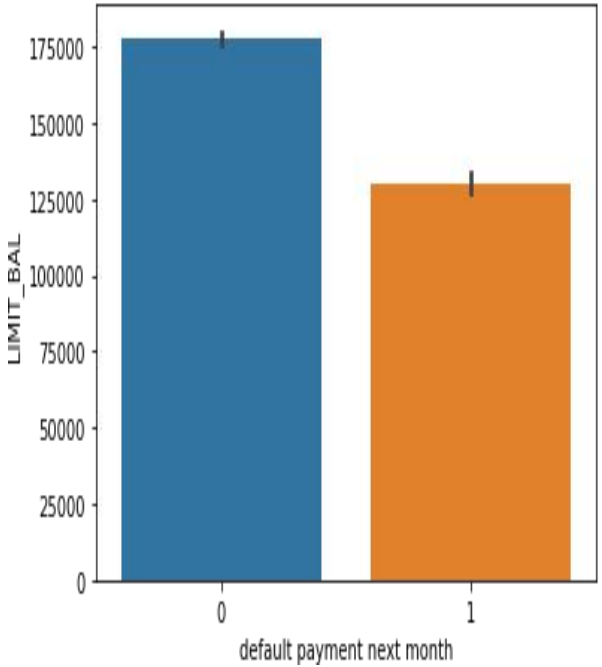


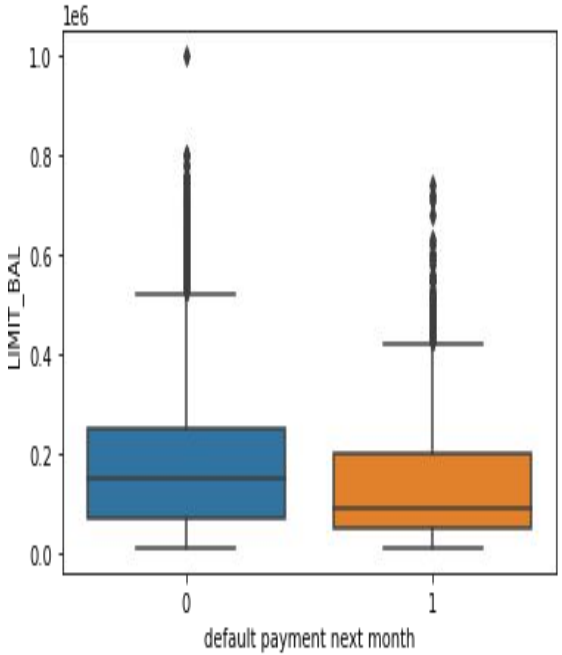
EDA- Independent Variables contd.



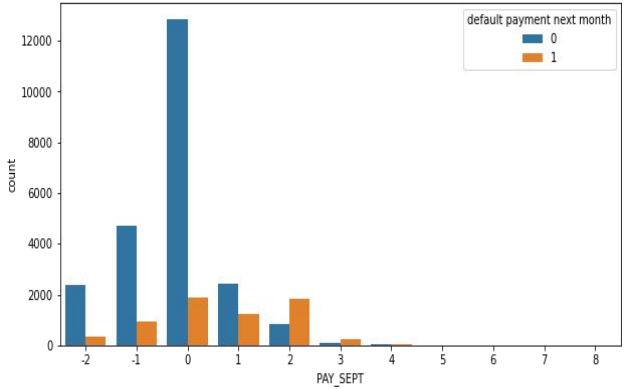
Limit Balance

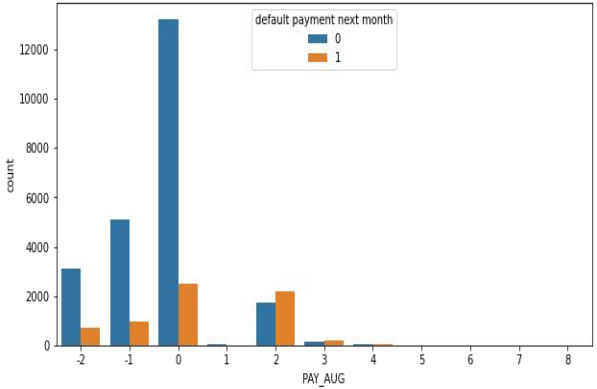


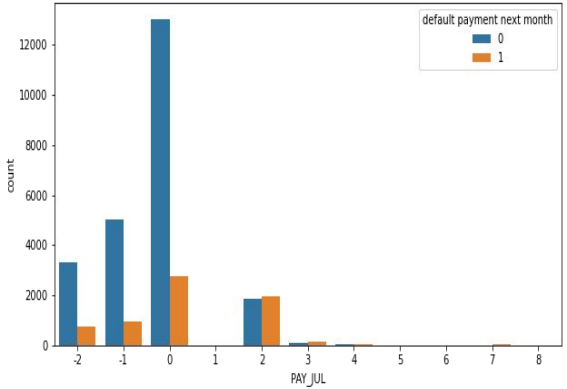


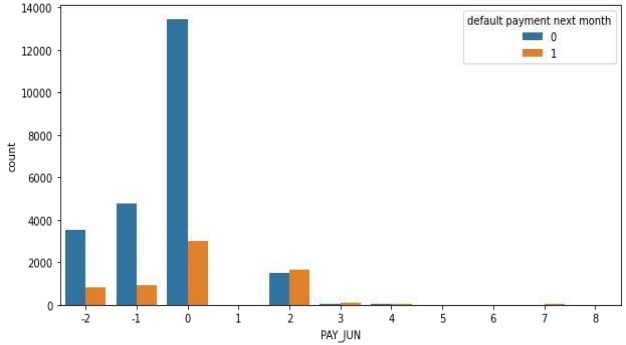


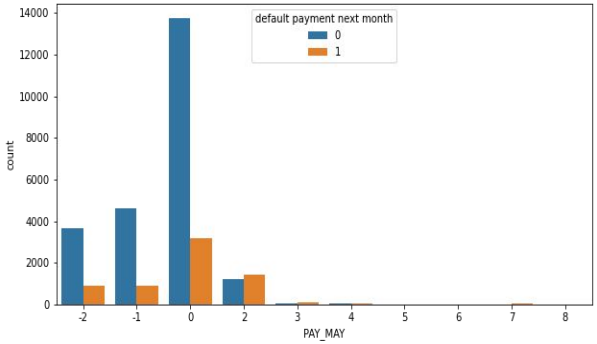
History / Previous Payments

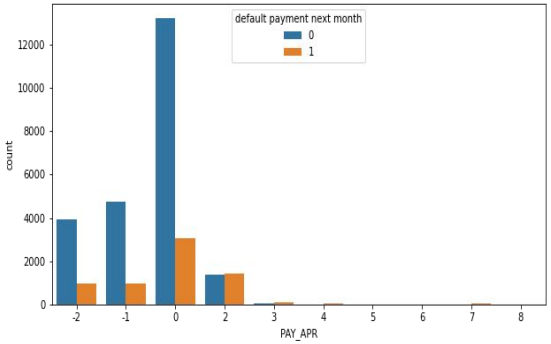




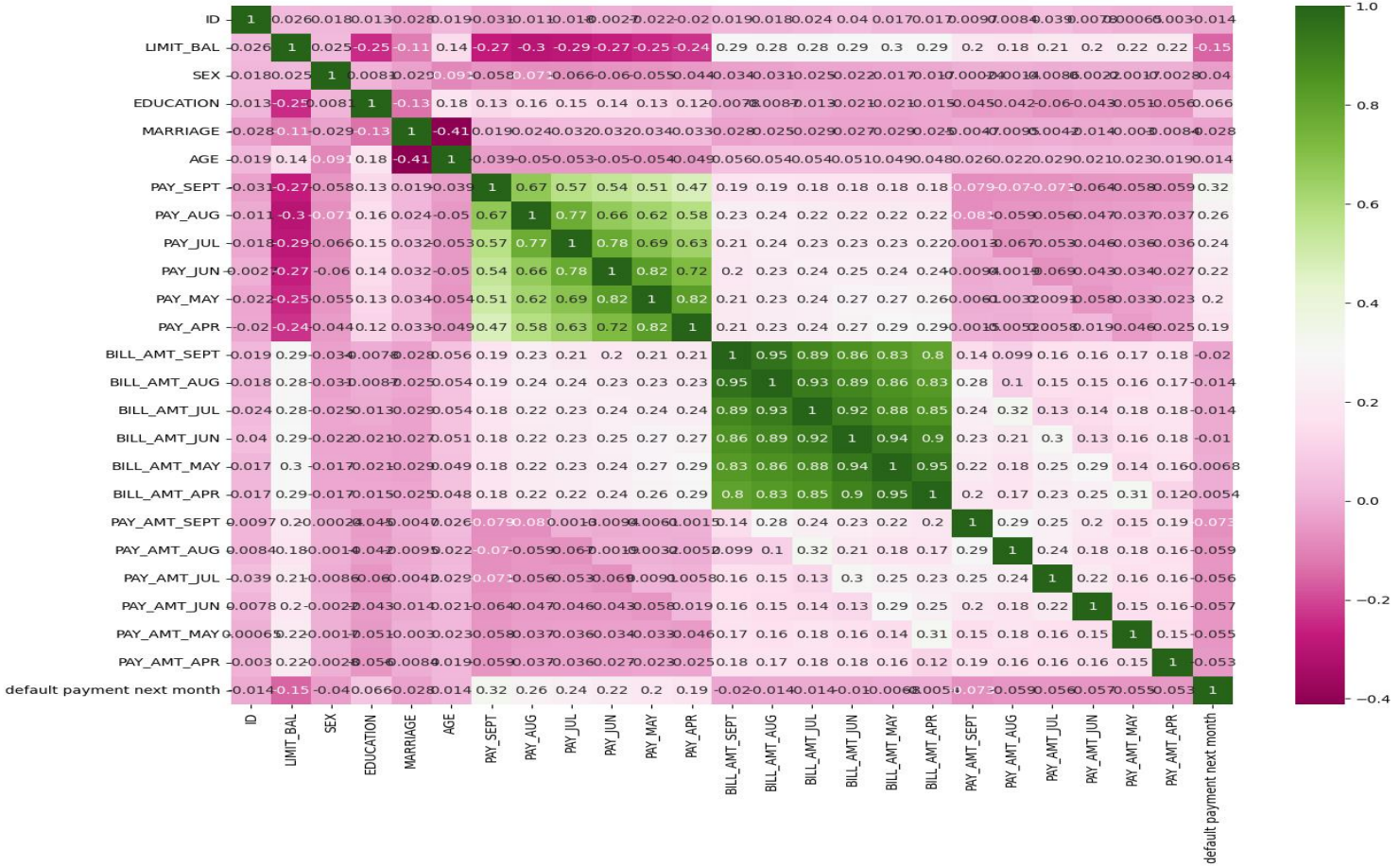




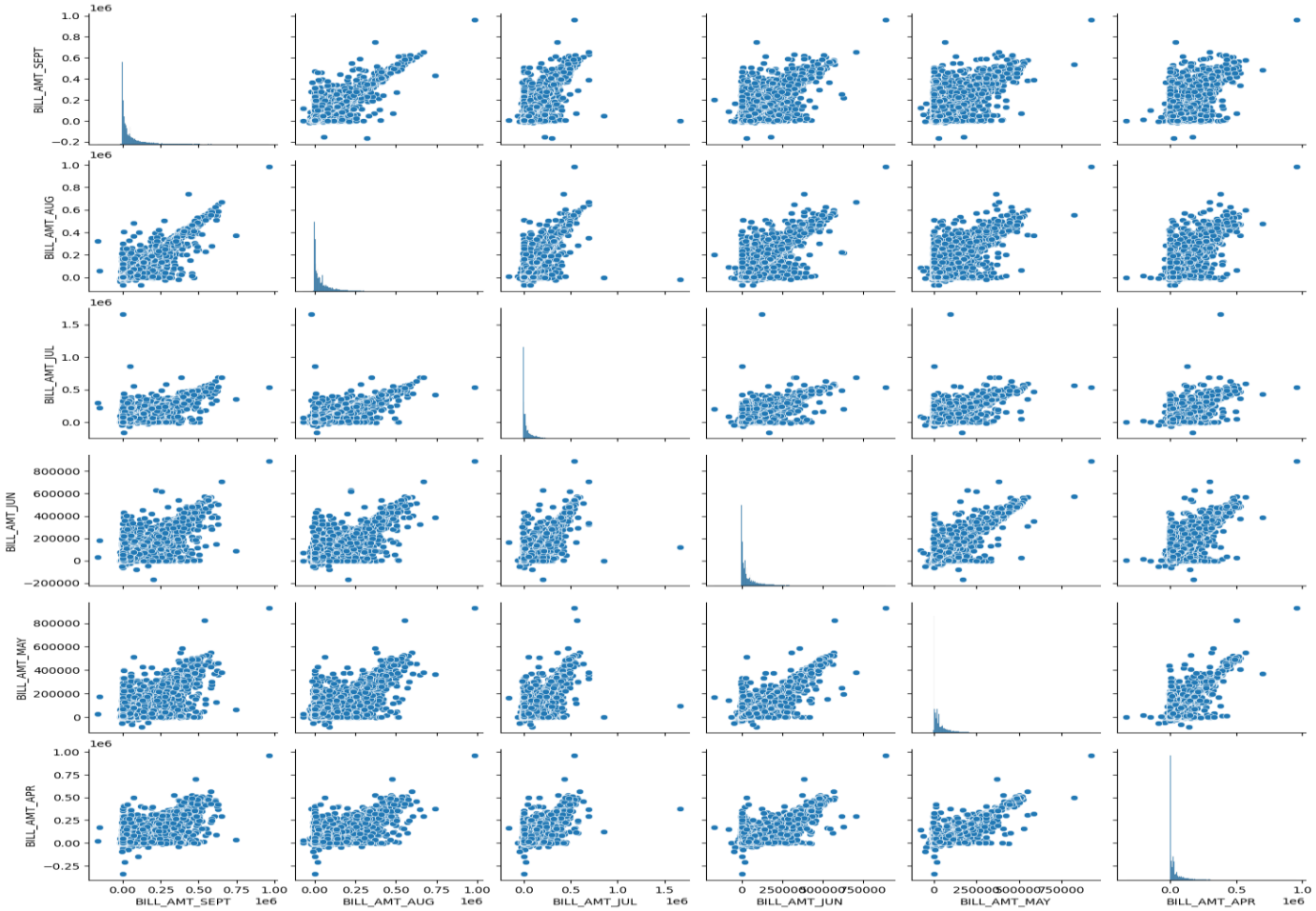


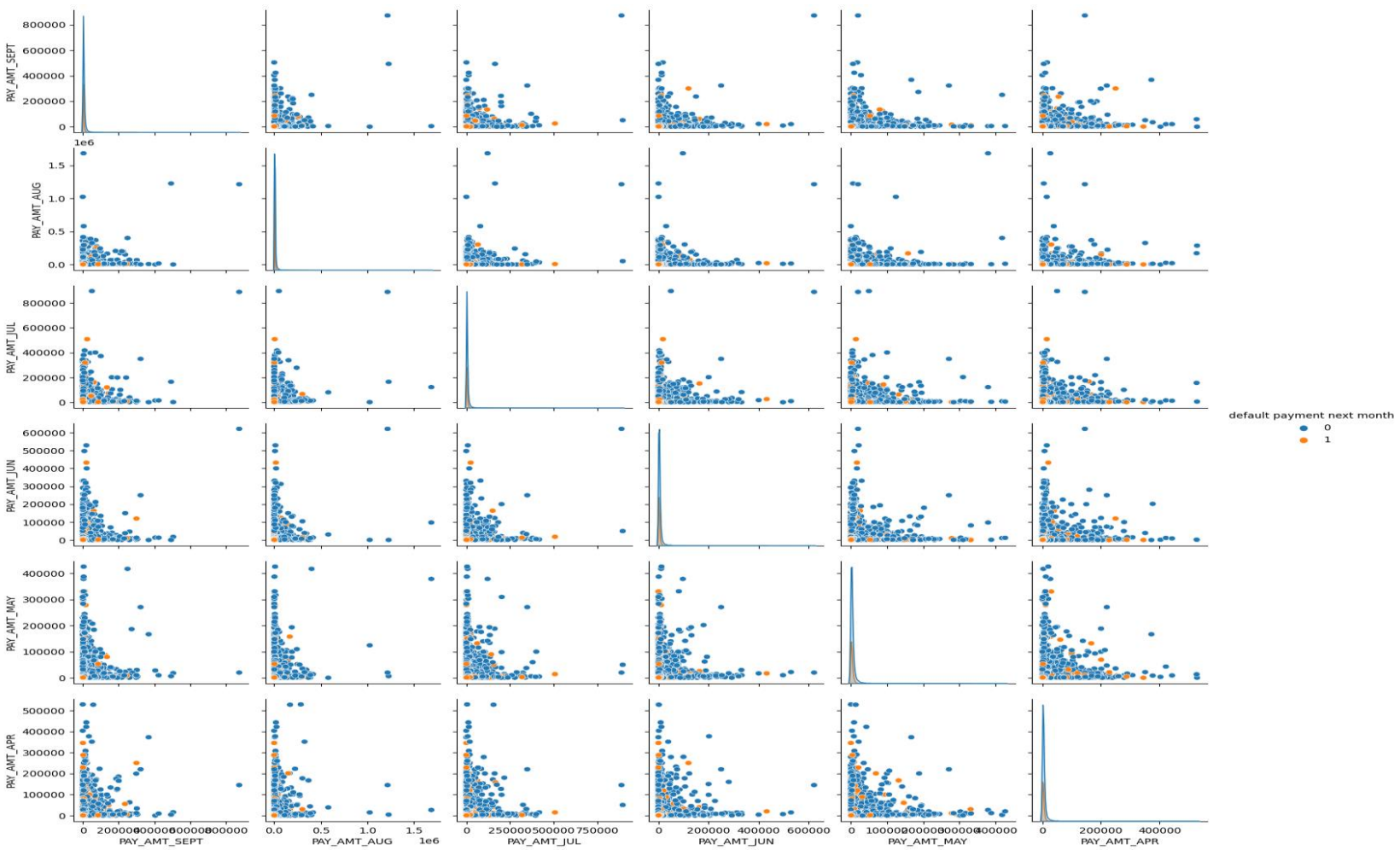


Checking Correlations



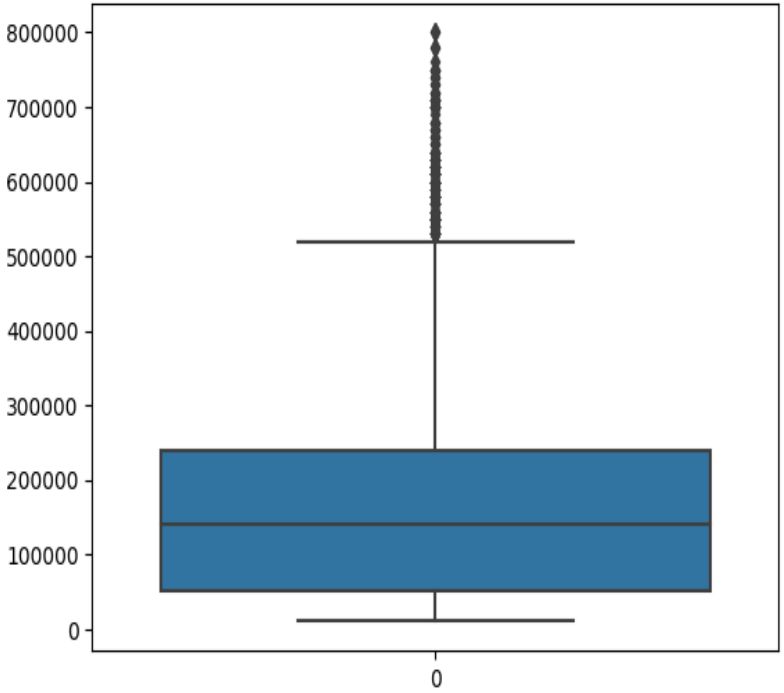
PAIR PLOT

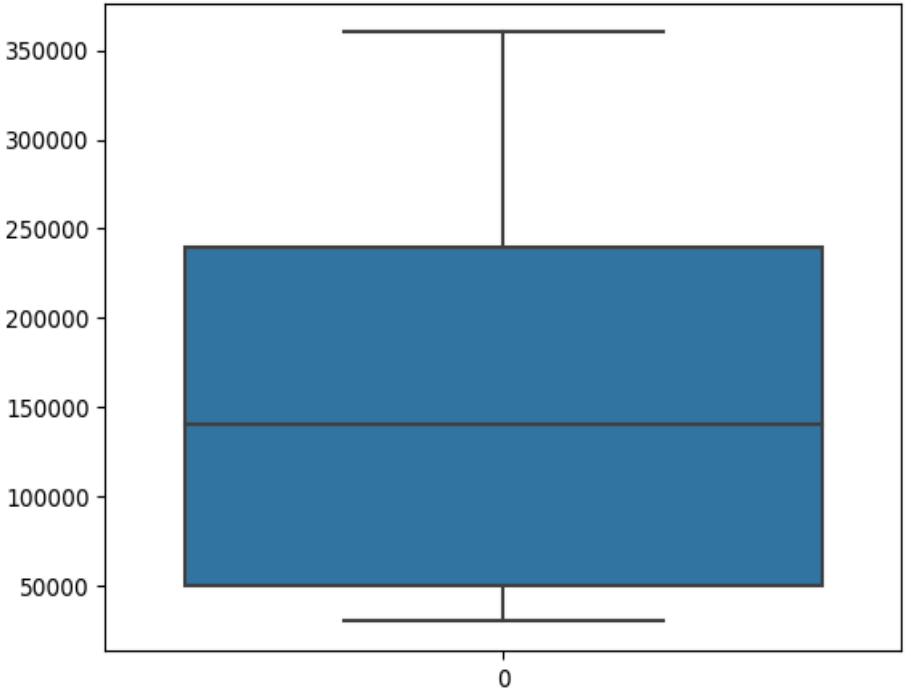

BILL_AMT_SEPT
BILL_AMT_AUG
BILL_AMT_JUL
BILL_AMT_JUN
BILL_AMT_MAY
BILL_AMT_APR
1e6


PAY_AMT_SEPT
PAY_AMT_AUG
PAY_AMT_JUL
PAY_AMT_JUN
PAY_AMT_MAY
PAY_AMT_APR
1e6
default payment next month
0
1

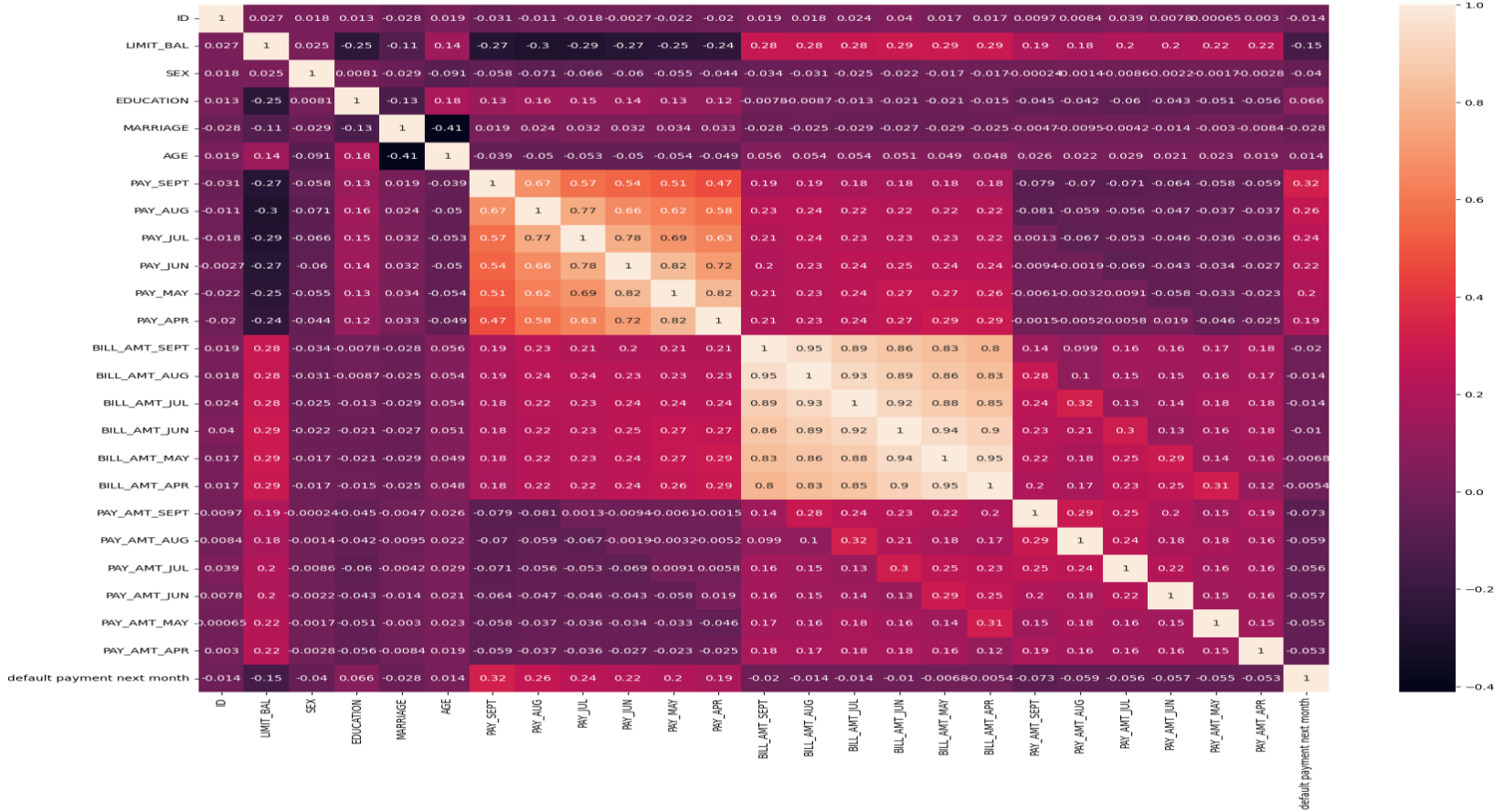
HANDLING OUTLIERS

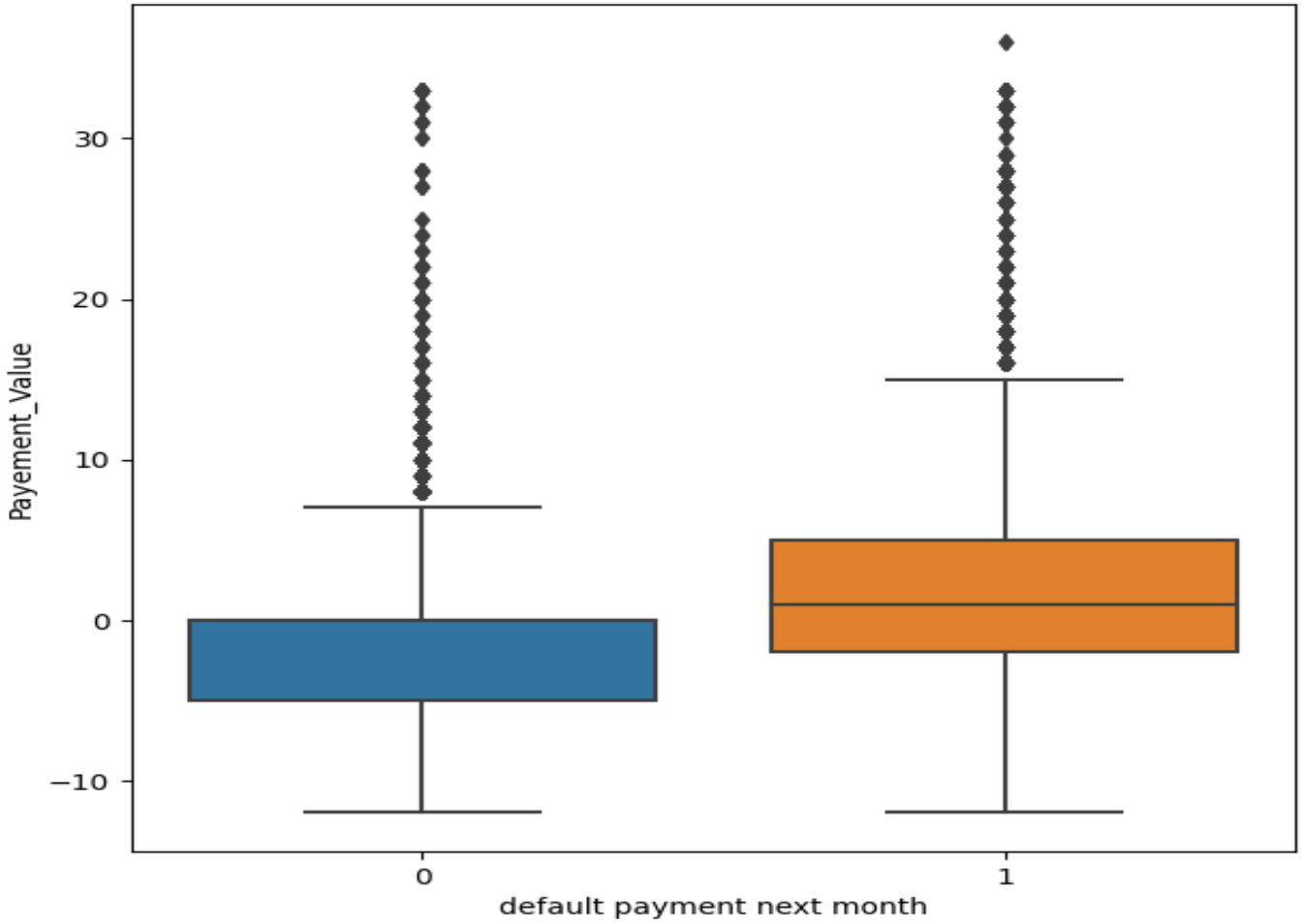
I have used Removing/deleting the outliers, Replacing them with mean/median, Quantile based flooring and capping.





FEATURE MANIPULATION AND SELECTION

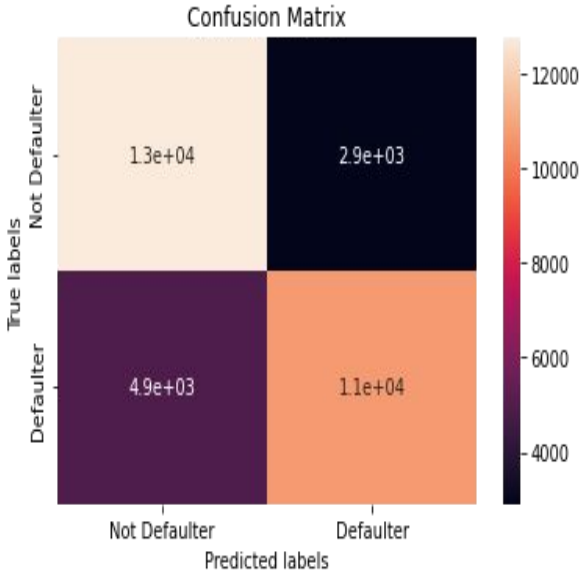


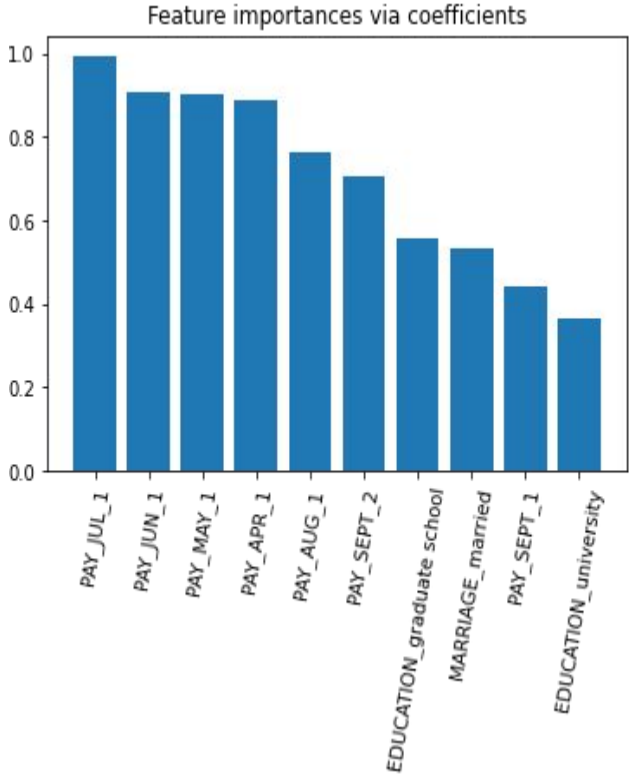

30
20
10
0
−10
Payement_Value
0
1
default payment next month

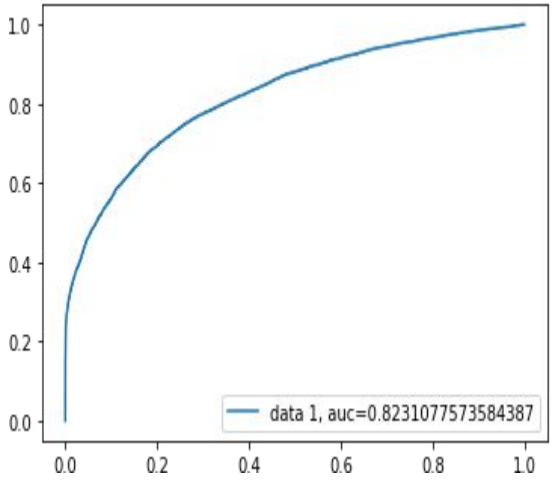
Model’s Performed

- Logistic Regression
- Decision tree
- Random forest
- eXtreme Gradient Boost

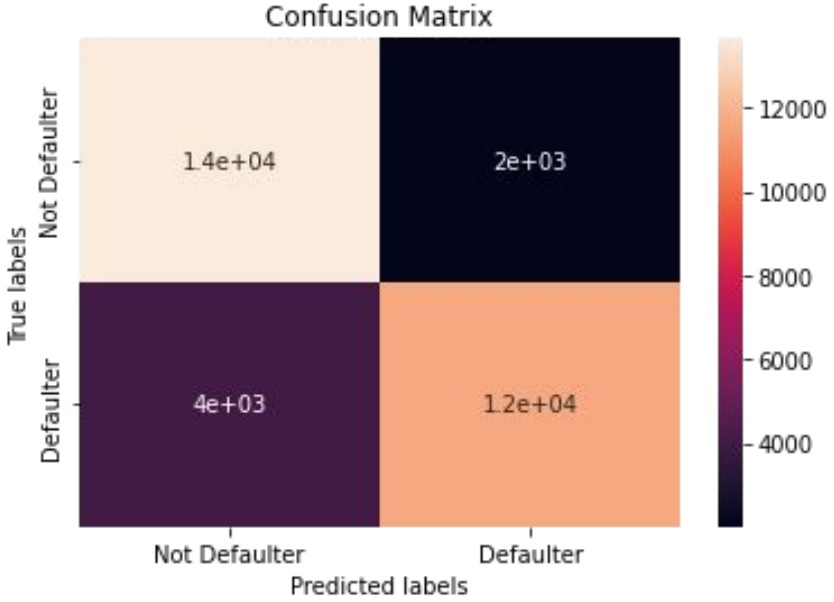
Logistic Regression

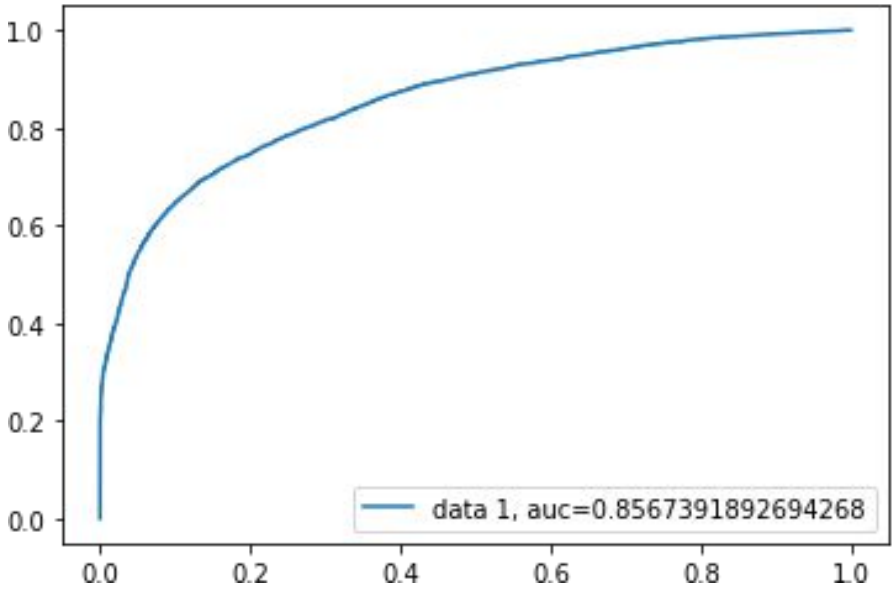




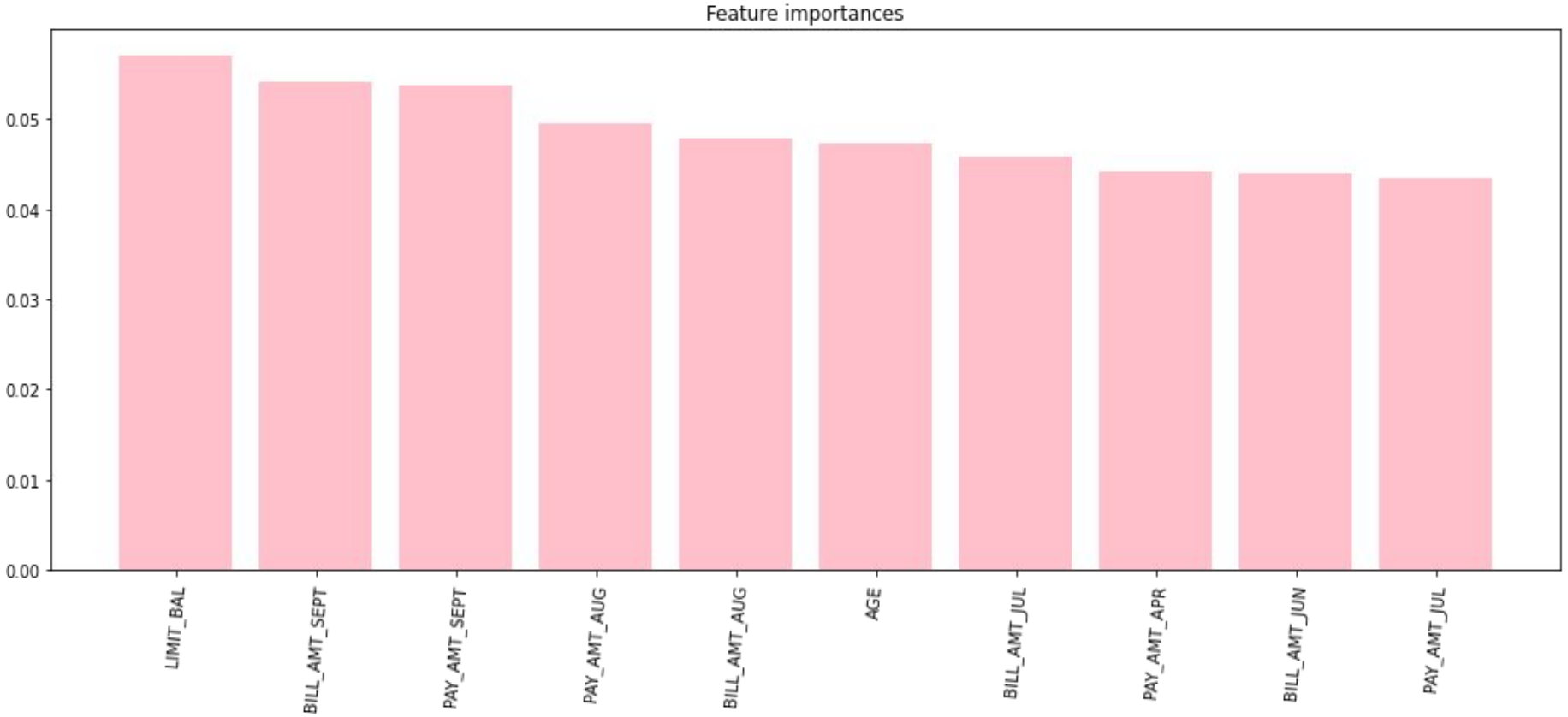


SVC

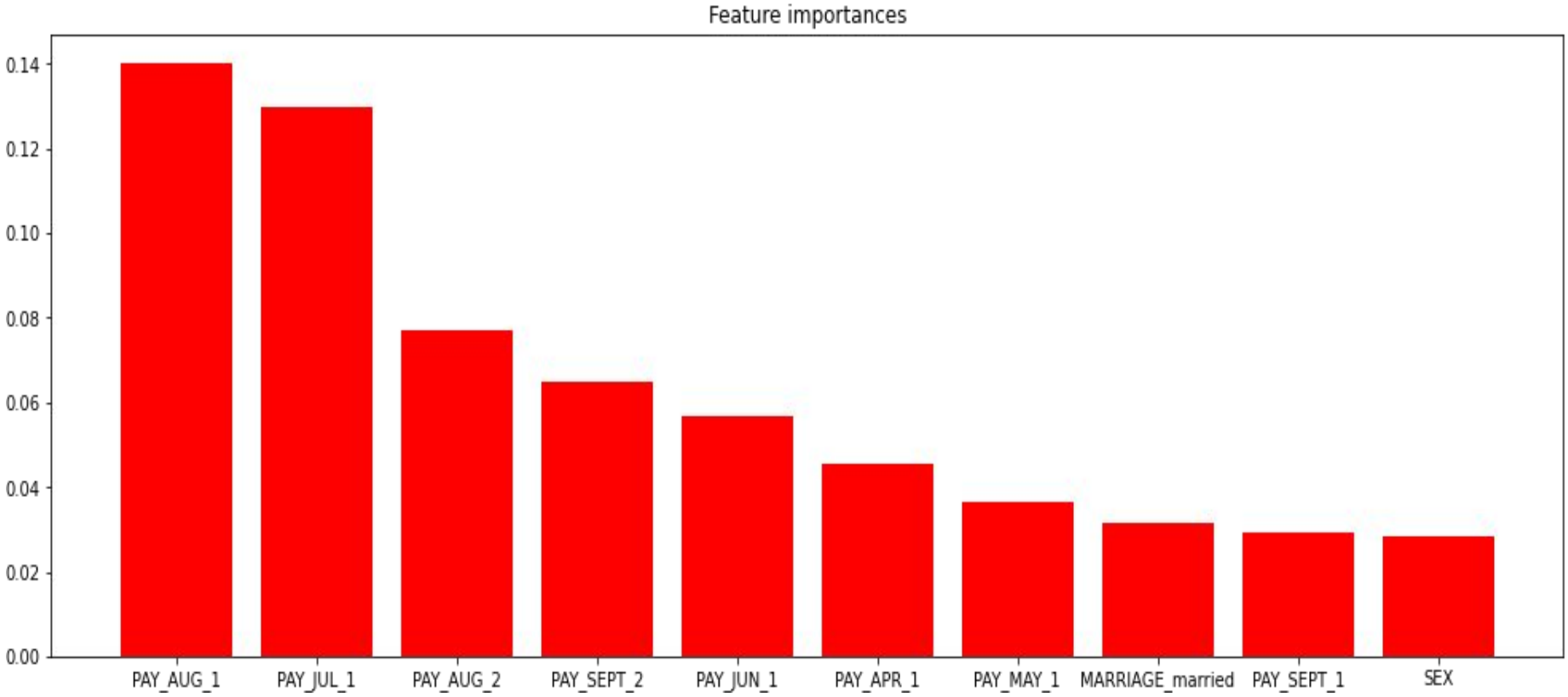

Confusion Matrix
1.4e+04
2e+03
4e+03
1.2e+04
Not Defaulter
Defaulter
True labels
Not Defaulter
Defaulter
Predicted labels
12000
10000
8000
6000
4000


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
data 1, auc=0.8567391892694268

RandomForest

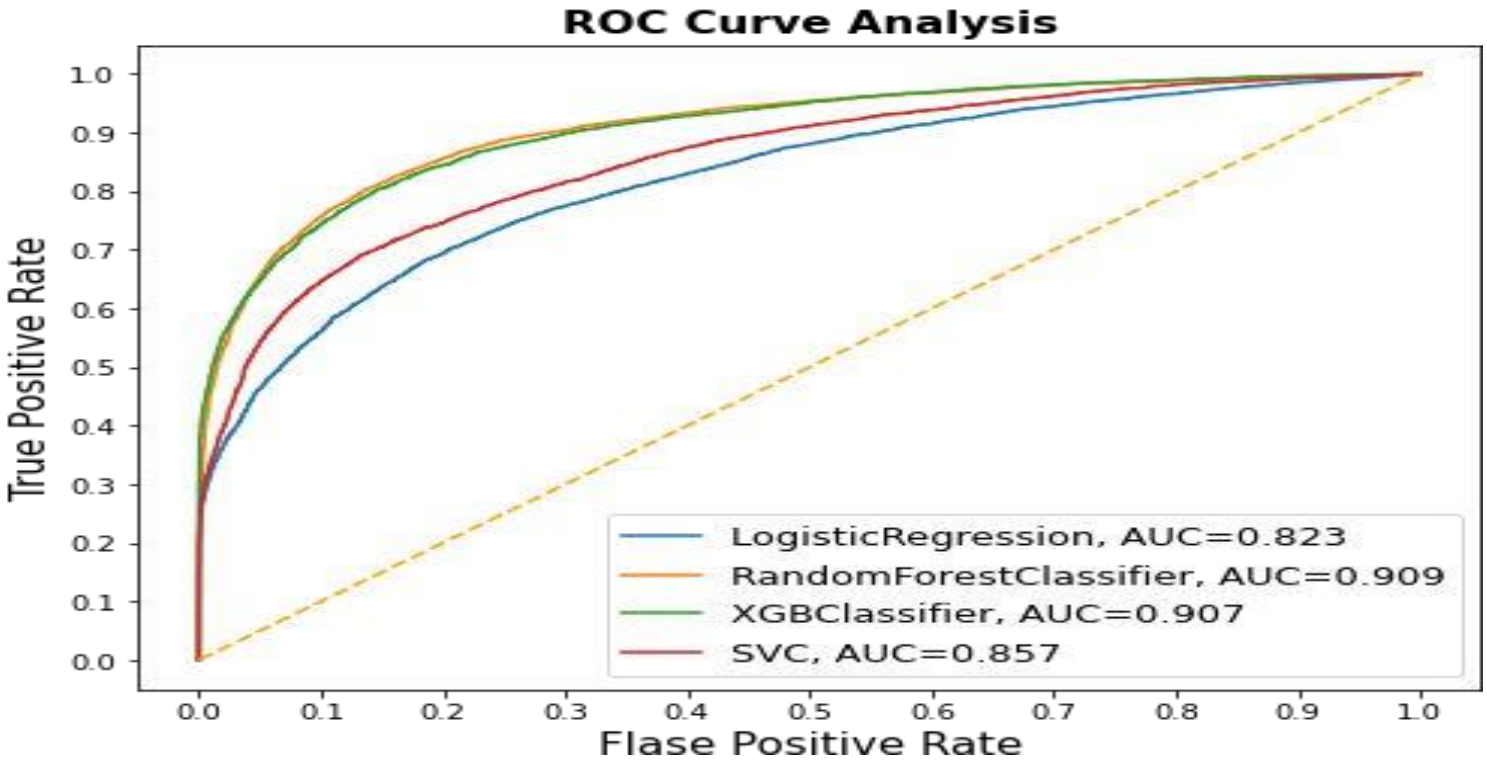


XGBoost feature Importance



ROC AUC for all the Models

ROC - AUC curve is **a performance measurement for the classification problems at various threshold settings**.



Model Validation & Selection (continued)

- **Observation 1:** As seen in the ROC AUC plot, Logistic Regression is not giving great results.

- **Observation 2:** Support Vector Classifier and Decision tree performed equally good.

- **Observation 3:** We are getting the best results from Random forest and XGBoost Classifier.

Challenges

- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.

- As dataset was quite big enough which led more computation time.

- After performing the various model we the get the best accuracy from the Random forest and XGBoost classifier.
- Logistic Regression is the least accurate as compared to other models performed.
- XGBoost has the best precision and the recall balance.
- Higher recall can be achieved if low precision is acceptable.
- We can deploy the model and can be served as an aid to human decision.
- Model can be improved with more data and computational resources

THANK YOU