Capstone Project

Credit Card Default Prediction

Team Members

Gaurav Gade

Hitesh Verma

Anand Gend

❖ Contents

- ❑ Problem Description
- ❑ Objective
- ❑ Concept Of Credit Card
- ❑ What To Expect When You Are Unable To Clear Credit Card Dues
- ❑ Data Pipeline
- ❑ Data Description
- ❑ Exploratory Data Analysis
  - 1) Feature Correlation Graph
  - 2) Correlation With Default
  - 3) Dependent Feature
  - 4) Independent Features
- ❑ Models performed
- ❑ Model Validation & Selection
- ❑ Feature Importance
- ❑ Overall ROC Curve Analysis
- ❑ Conclusion
- ❑ Challenges

❖ Credit Card Default Prediction

❑ Problem Description

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.

OBJECTIVE

Credit Card holder data

Classifier Model

No Default

Default

❖ Concept of Credit Card

AI

A credit card is a type of payment card in which charges are made against a line of credit instead of the account holder's cash deposits. Although failure to pay off the credit card on time could result in interest charges and late fees, credit cards can also help users build a positive credit history. According to the Diner's Club, the idea of the credit card came to **Frank McNamara** in 1949 while he was having dinner at a restaurant in New York City. When it was time to pay the bill, McNamara realized he had forgotten his wallet.

Advantages

1.Building credit history.
2.A quick source of funds in an “absolute” emergency
3.No accrued interest if bill is paid on time and in full each month
4.Zero liability as consumers is not responsible for fraudulent charges when reported promptly.
5.Consumer protection ($50.00) if fraudulent charges are reported promptly in case the card is stolen or lost.

Disadvantages

1.Established credit-worthiness needed before getting a credit card
2.Encouraging impulsive and unnecessary “wanted” purchases
3.High-interest rates if not paid in full by the due date
4.Annual fees for some credit cards – can become expensive over the years
5.Fee charged for late payments
6.Negative effect on credit history and credit score in case of improper usage

WHAT TO EXPECT WHEN YOU ARE UNABLE TO CLEAR CREDIT CARD DUES

Post Due Date

Reminders & Notifications

Continued Reminders, Notifications, & Phone Calls

Post 30 Days

Post 90 Days

- Loan Recovery Agents
- Arbitration & Conciliation
- Police Complaint

- Account classified as Recovery Pool
- Lawsuit Filed
- Debt sold to collection Agency

Post 191 Days

Note: Once default is declared on your credit card account, you can also be blacklisted with credit rating agencies like CIBIL and Experian making it difficult for you to get loans and credit cards in future.

❖Data Pipeline

❑ Exploratory Data Analysis

After loading the dataset we performed EDA by comparing our target variable that is Default Payment with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

❑ Null values Treatment

After the data is loaded , The missing data is checked using isna() or isnull() function and the identical rows are checked by using duplicated() function. The output depicted that there is no missing values and identical rows in our dataset. So our dataset does not contain any missing values and duplicate rows.

❑ Feature Engineering

To make the data tenable for understanding and further analysis , the data set was analysed for identifiable statistical trends and patterns. After preliminary analysis, the following steps were undertaken to transform the data into a systematically workable dataset:

1) We combined all the features of monthwise payment status into a new feature named as Payment Value.
2) Then, we combined all the features of monthwise bill amount into a new feature named as Dues.
3) Reducing the values of Education feature by replacing and maintaining the four categories.
4) Same operation perform with Marriage feature and maintaining three types of values.

❑ Encoding of Categorical features

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

Categorical variables- (Education and Marriage) both features were converted into numerical depictions to fit our Model to Predict Credit Card Defaulters.

Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

Feature Selection

In these steps we used algorithms like Decision Tree, Random Forest, XGBoost to check the results of each feature i.e., which feature is more important compared to our model and which is of less importance.

Among all the features Limit Balance, Payment and Bill amount feature are the most important in model prediction. While almost all features are the least important, we found that they help enhance the performance of the models.

❖ Data Description

❑ Dependent variable:

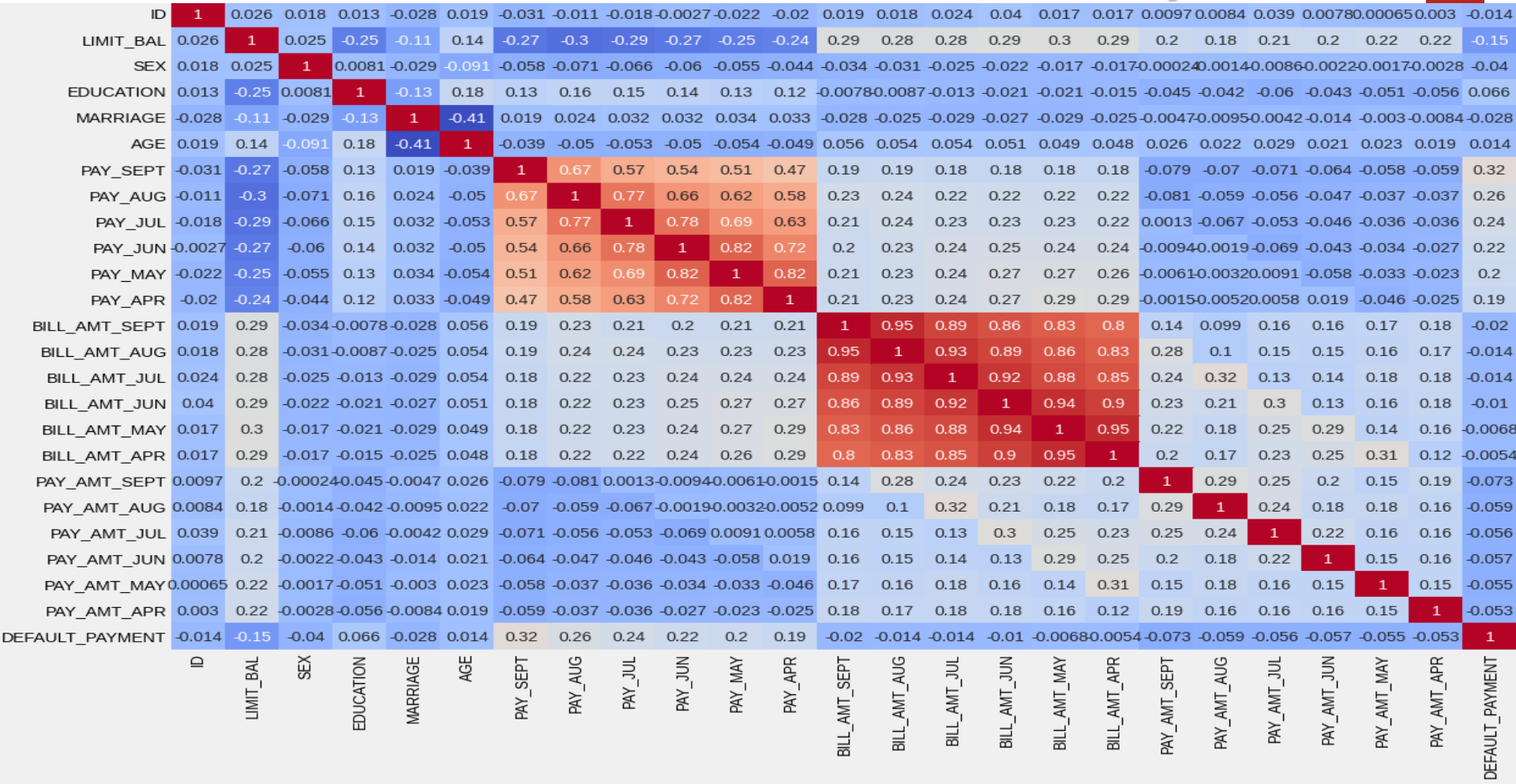
Default Payment: Default payment (1=yes, 0=no)

❑ Independent variables:

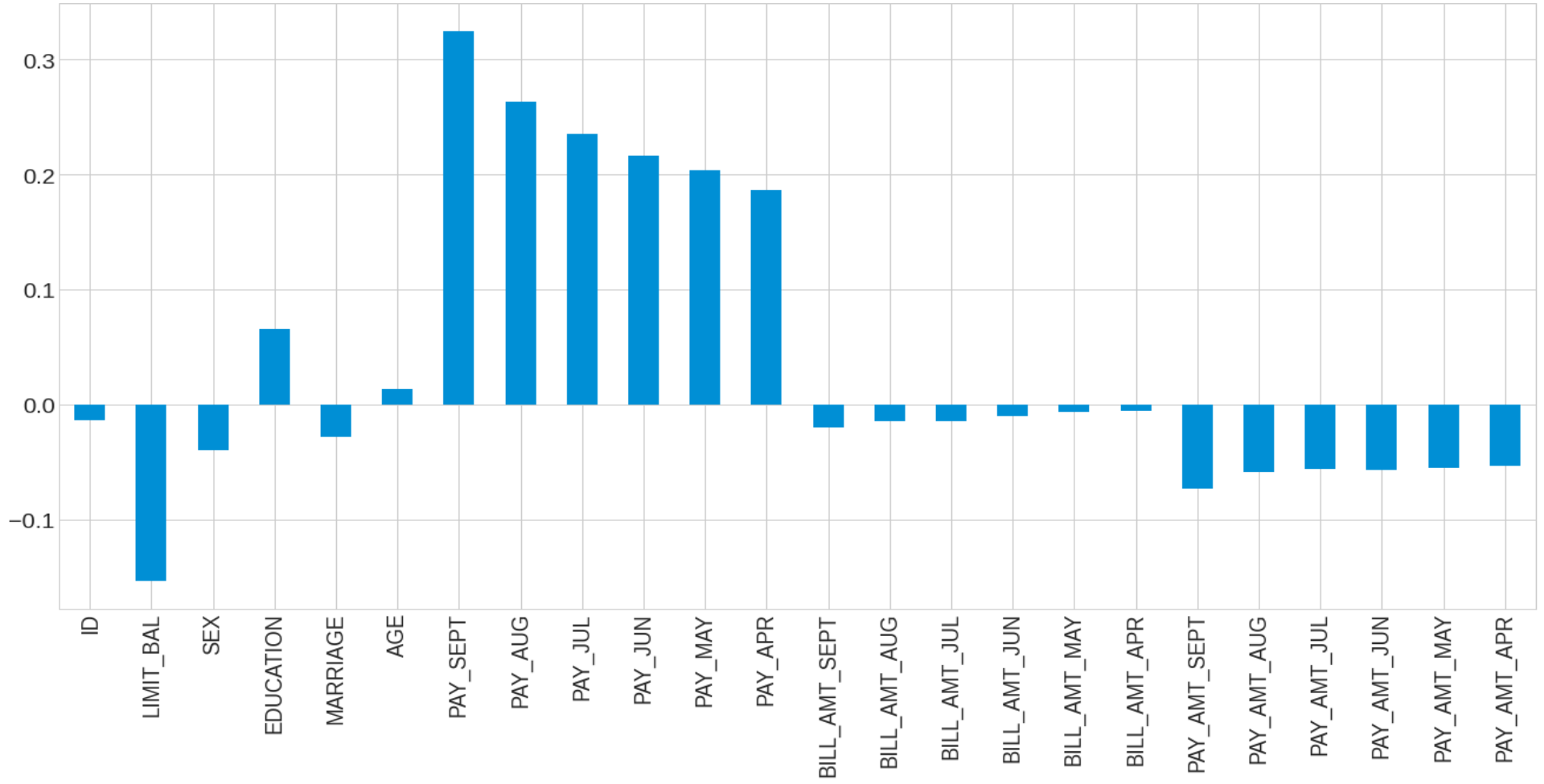
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0 to PAY_6: Repayment status in April to September, 2005
- BILL_AMT1 to BILL_AMT6 : Amount of bill statement in April to September, 2005 (NT dollar)
- PAY_AMT1 to PAY_AMT6 : Amount of previous payment in April to September, 2005 (NT dollar)

(Scale for last three index : 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)

❖ EDA - Feature Correlation Graph



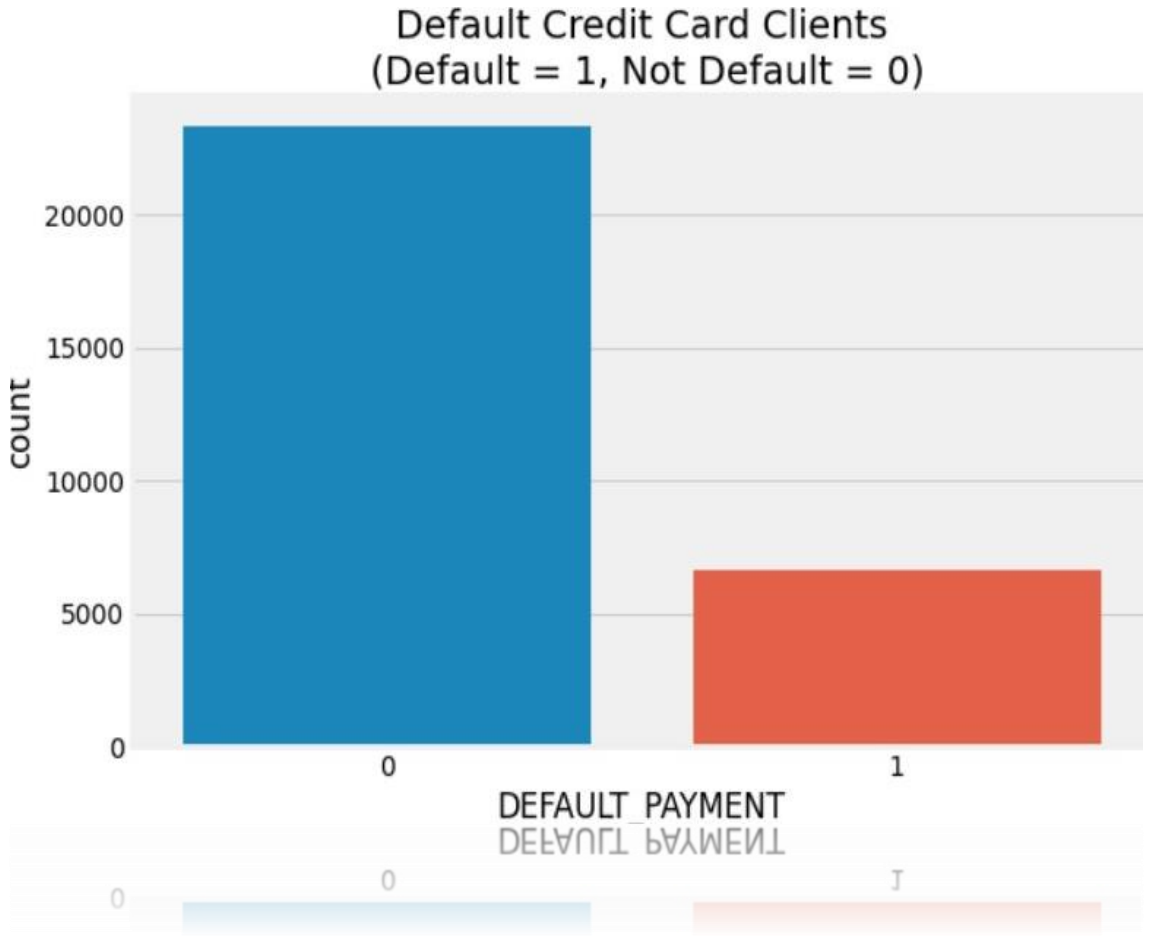
❖ EDA – Correlation With Default



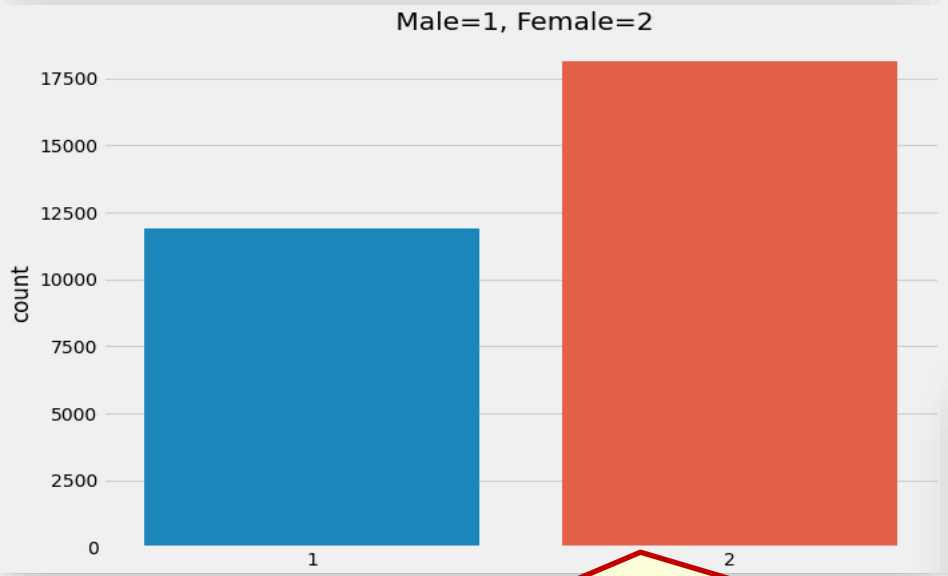
❖ EDA - Define Dependent Feature

Since Default Payment is our dependent feature and we have a Barplot for count of default payment.

So, from this Barplot we can conclude that Defaulters are less as compared to Non-defaulter. If we go for numbers, there are near about 23400 clients are not defaulters where 6600 clients are defaulters.

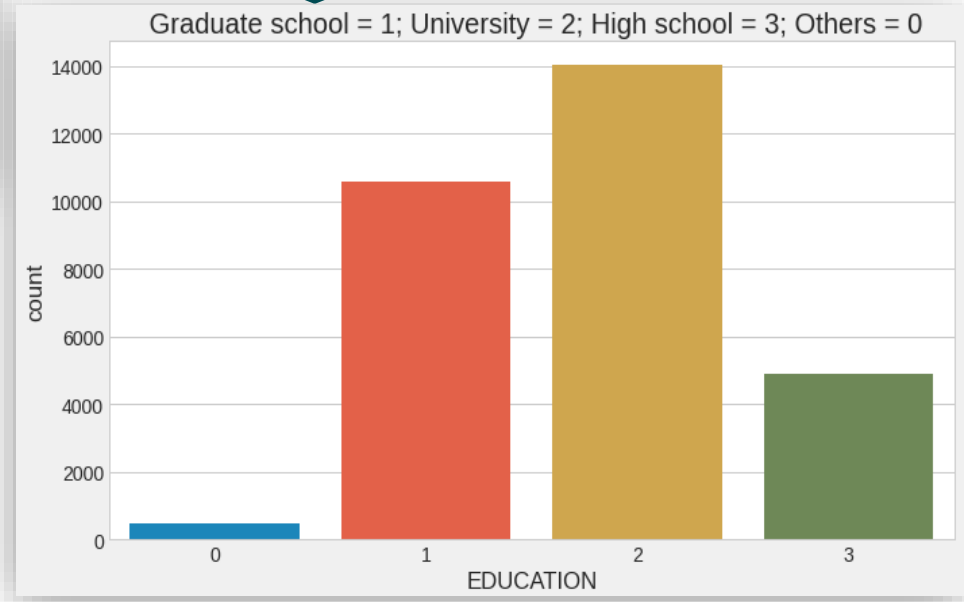


❖ EDA – Gender Count & Education Count

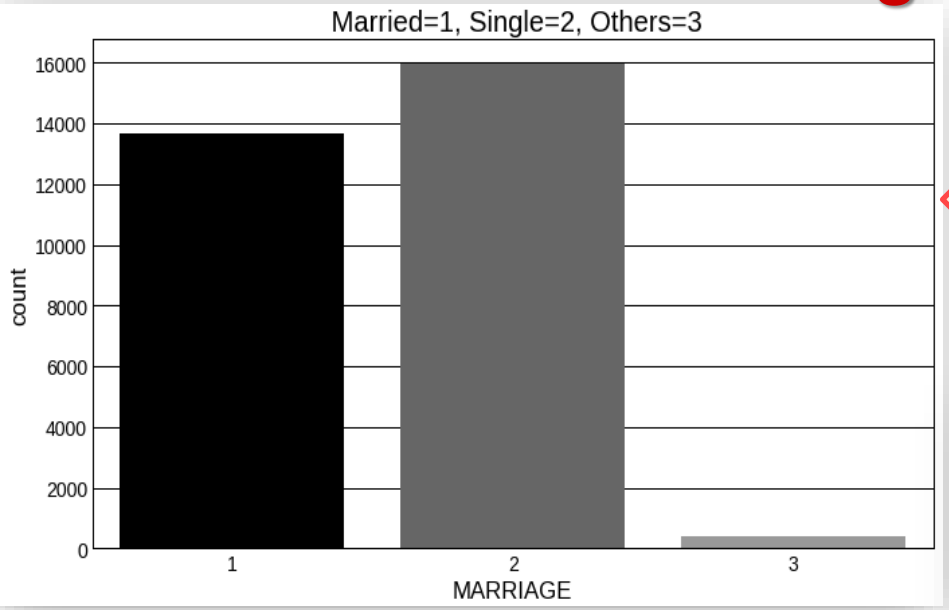


The Countplot for gender showing that Females are leading in the use of Credit Card. Approx. 18100 females are using the credit cards at the same time males are somewhere lagging in use of credit cards (approx. 11900 credit cards were issued for men.

More number of credit card holders are university students (approx. 14000) followed by Graduates (approx. 10500) and then High school students (approx. 4900).

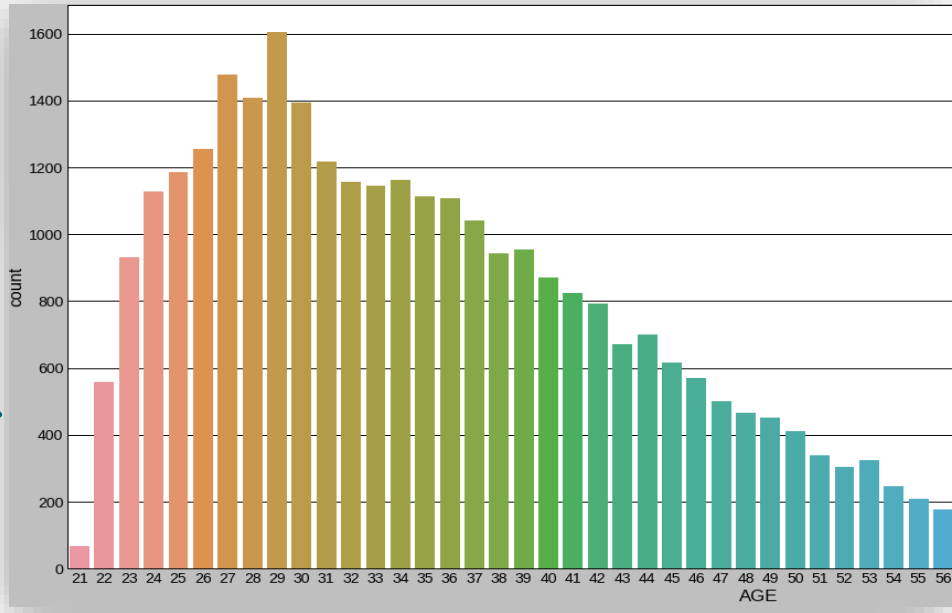


❖ EDA – Marriage Count & Age Count

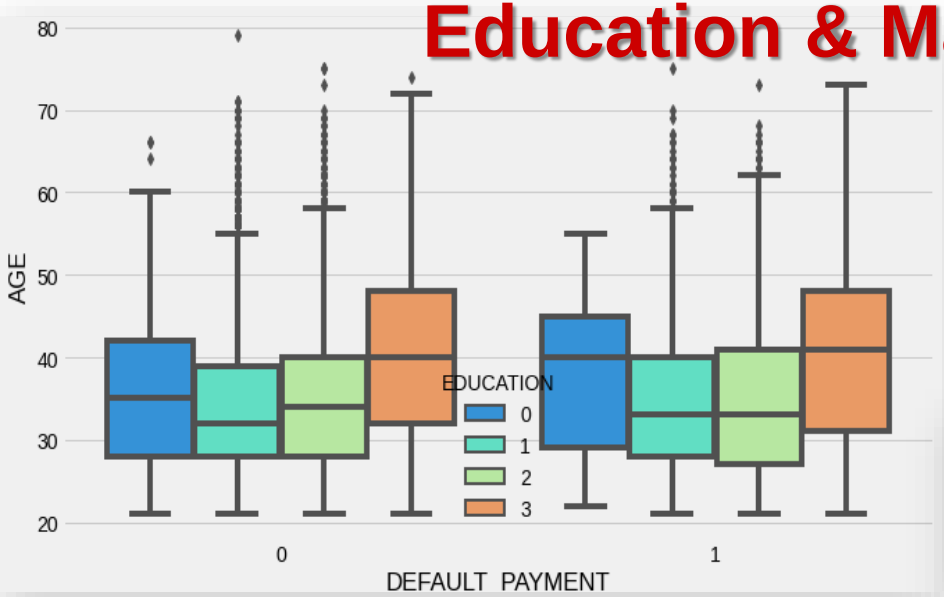


As we can see that, large amount of credit card holders are singles. There are 16000 clients who are unmarried and using credit cards, also 13700 credit cards users are married which are quite near to the unmarried clients.

This Countplot represents that the people in the age group of 24-37 are extensively using the credit cards. And the people from age group of less than 24 and greater than 37 are comparatively less.

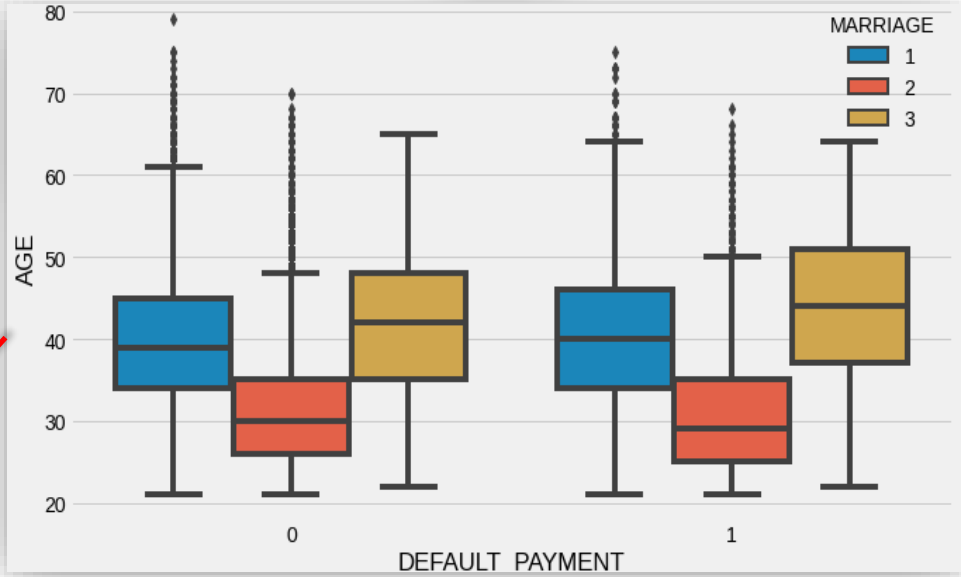


❖ EDA – Default Payment vs Age With Education & Marriage

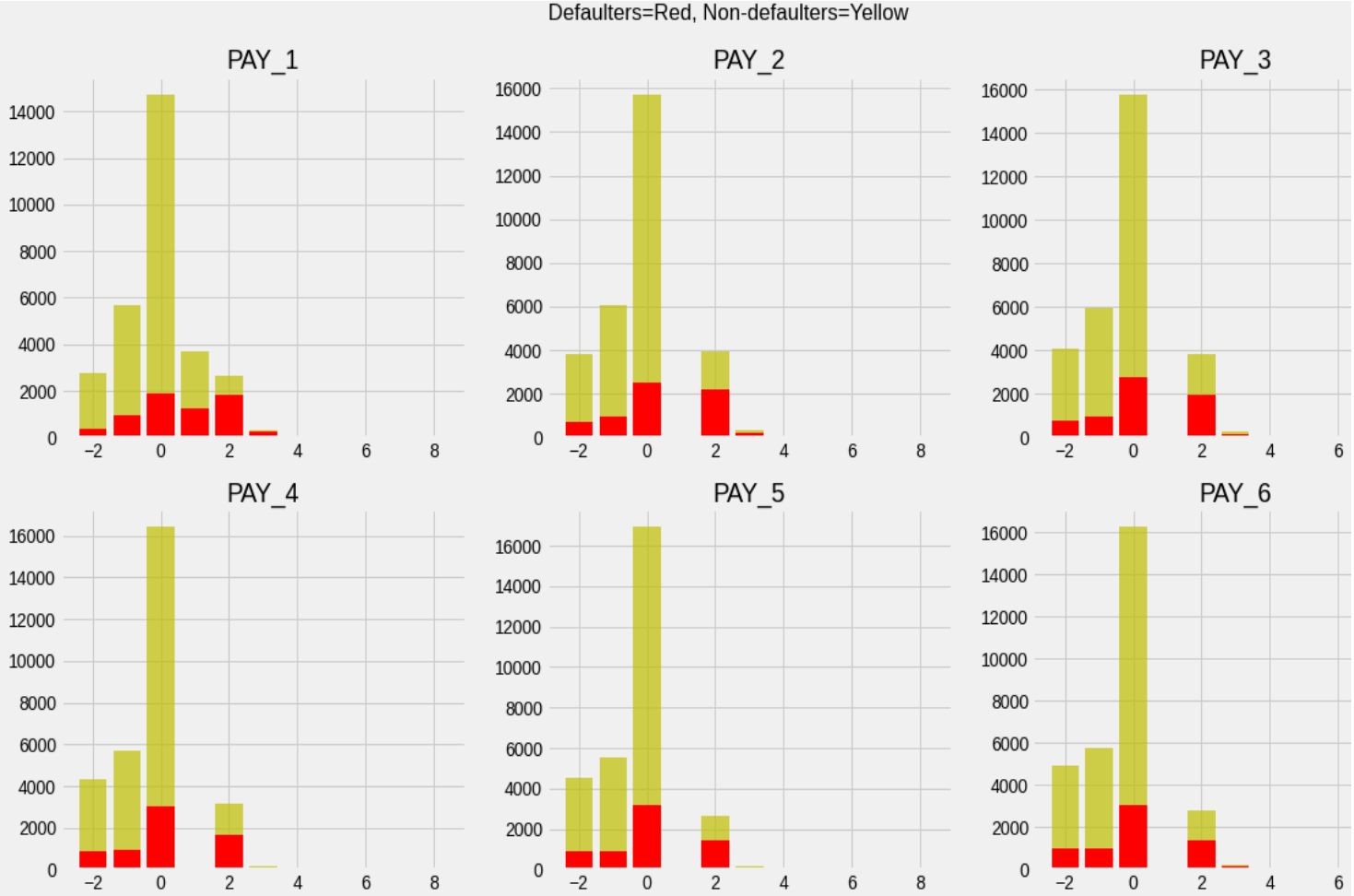


From this Boxplot, we can conclude that more number of defaulters are high school students followed by university students.

From this Boxplot, we can say that large number of defaulters are married as compared to unmarried clients.

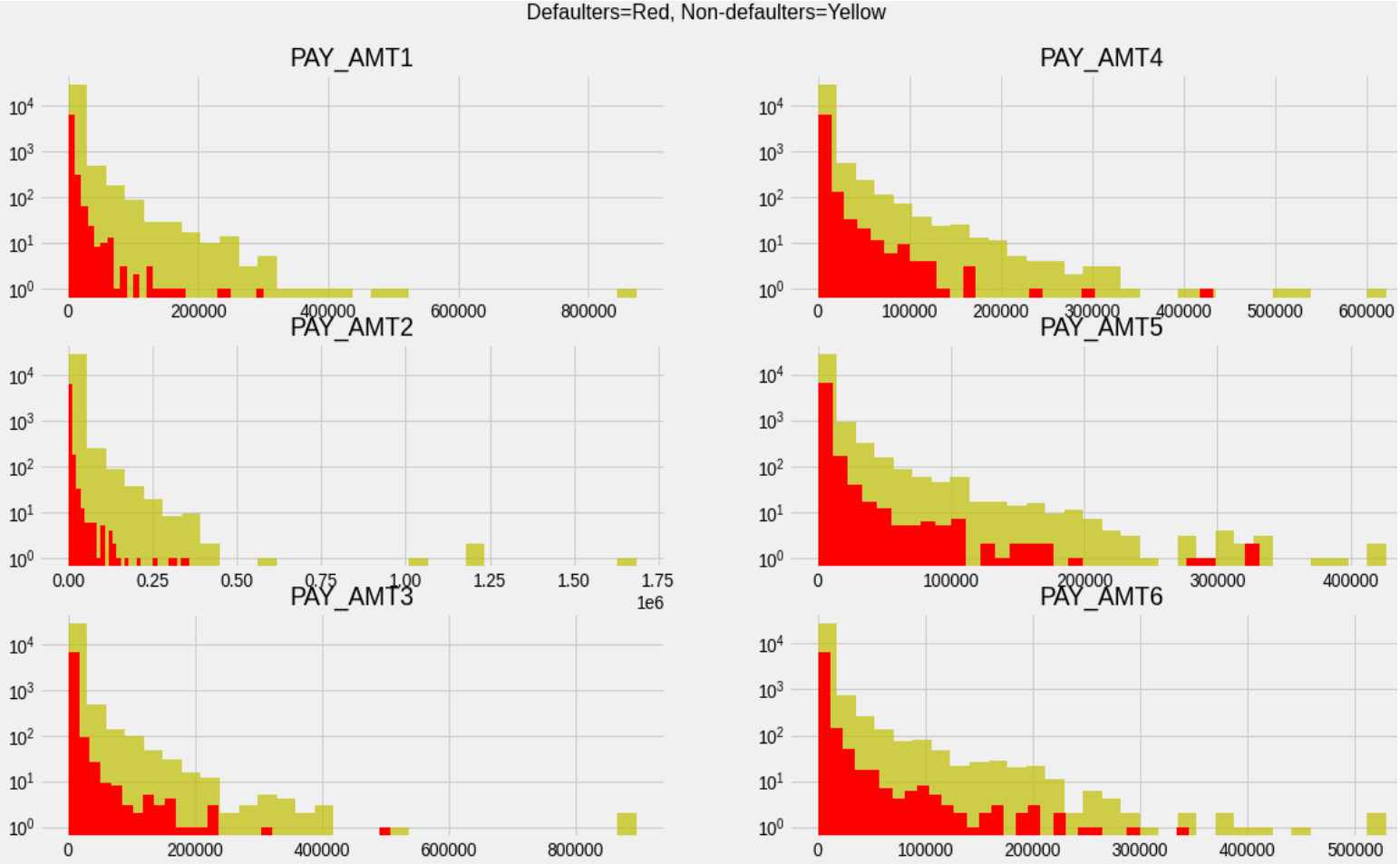


❖ EDA – Monthwise Payment Status



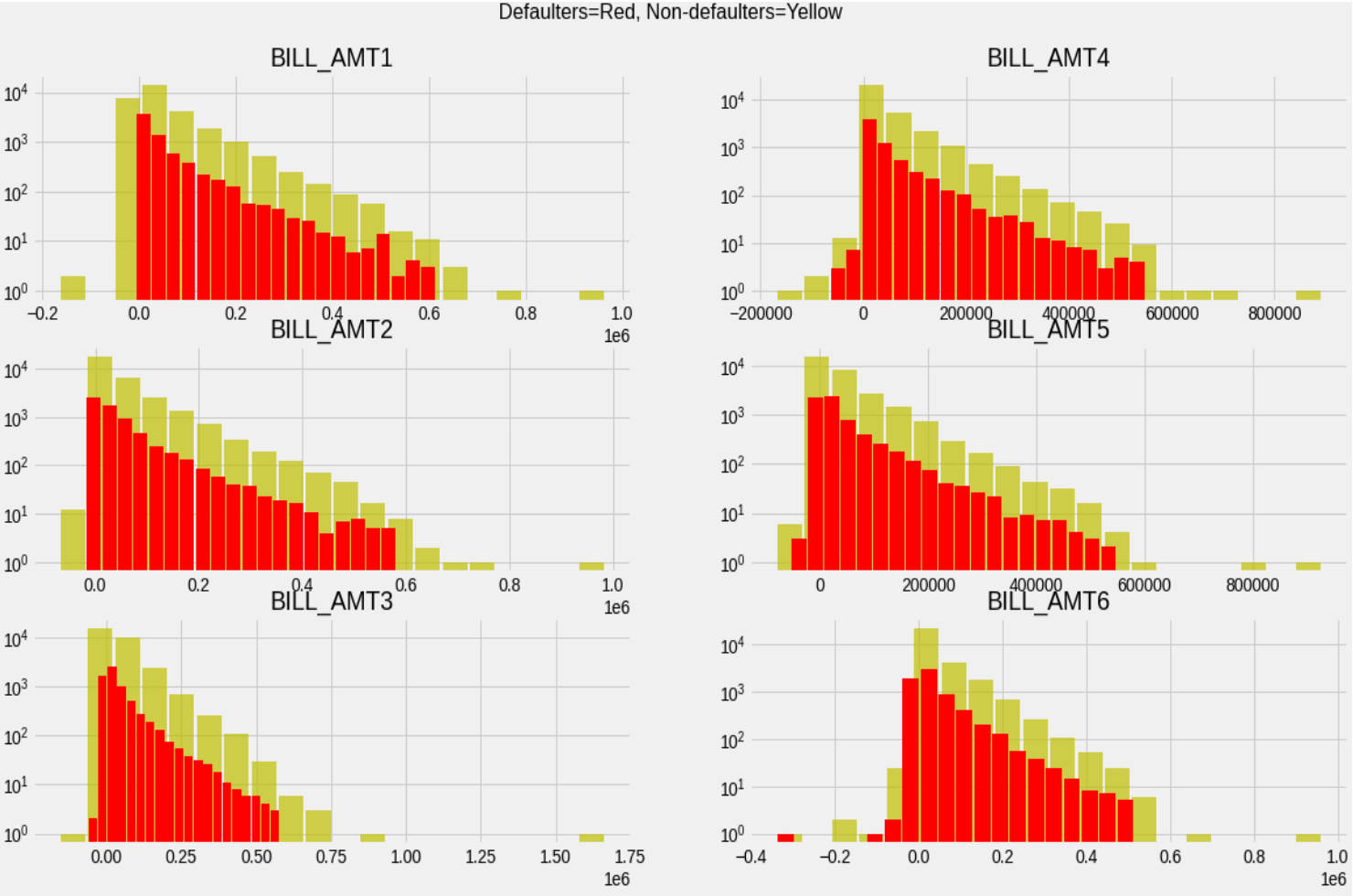
Highest number of defaulters are in May & June followed by April, July & August where lowest number of defaulters are from month of September.

❖ EDA – Monthwise Payment Distribution



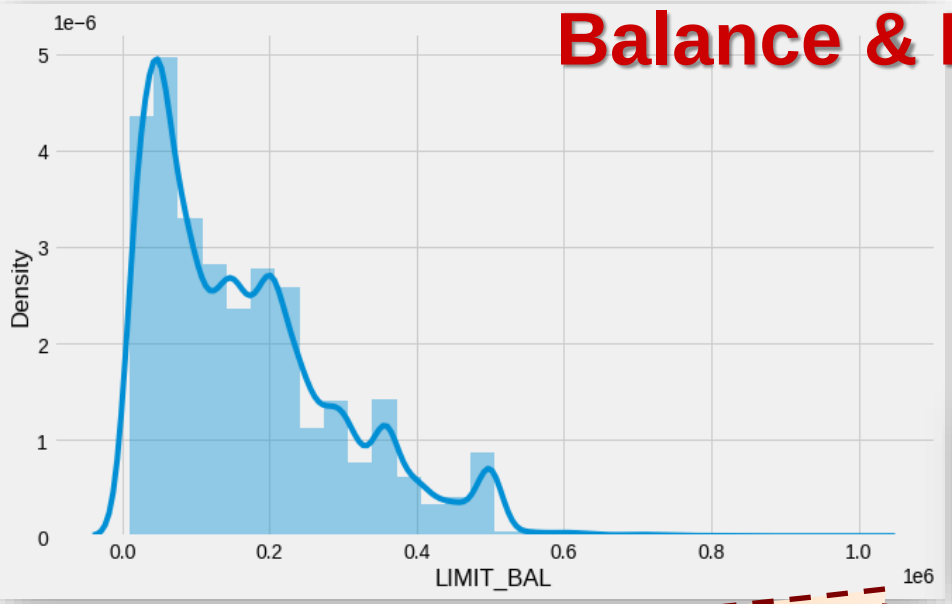
Highest payment distribution was done in April, May & June since lowest payment distribution was done in month of August.

❖ EDA – Monthwise Bill Amount Distribution



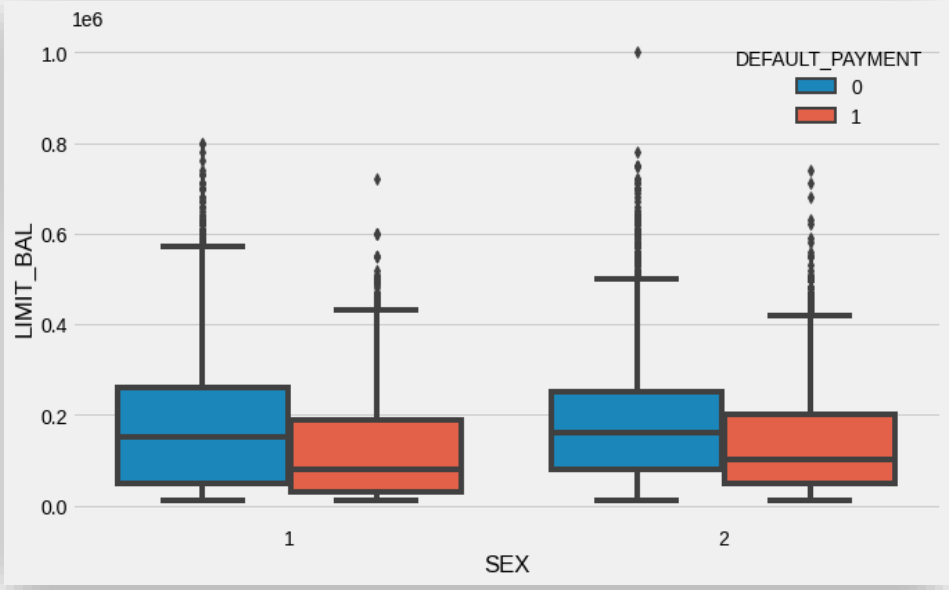
Highest bill amount was generated in May followed by August, and June. If we consider only for defaulters then the highest bill amount was generated in the month of May.

❖ EDA – Default Payment with Limit Balance & Distribution

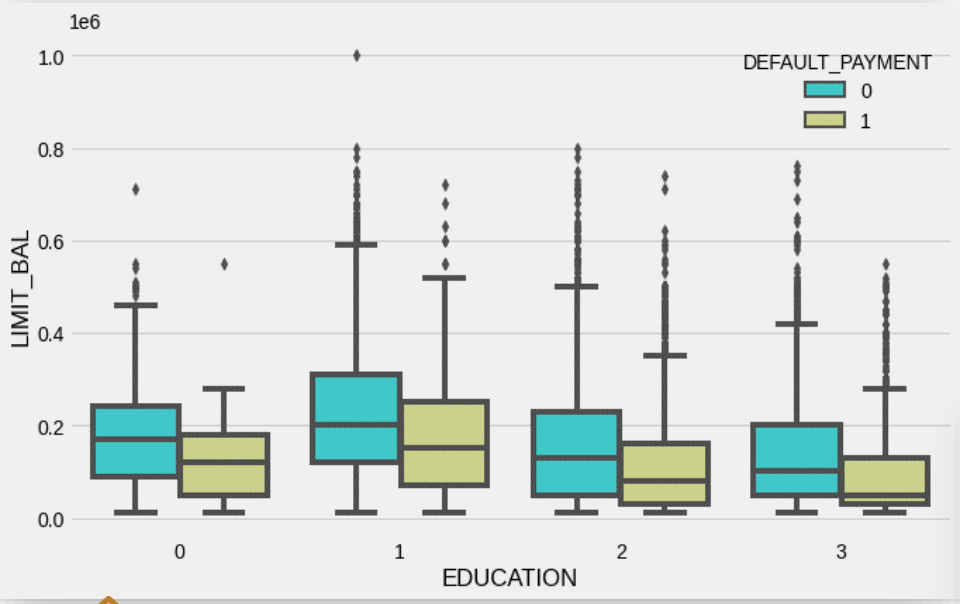


Here, we plotted distplot with KDE so we can understand distribution for Limit Balance. And it is observed that the plot is positively skewed.

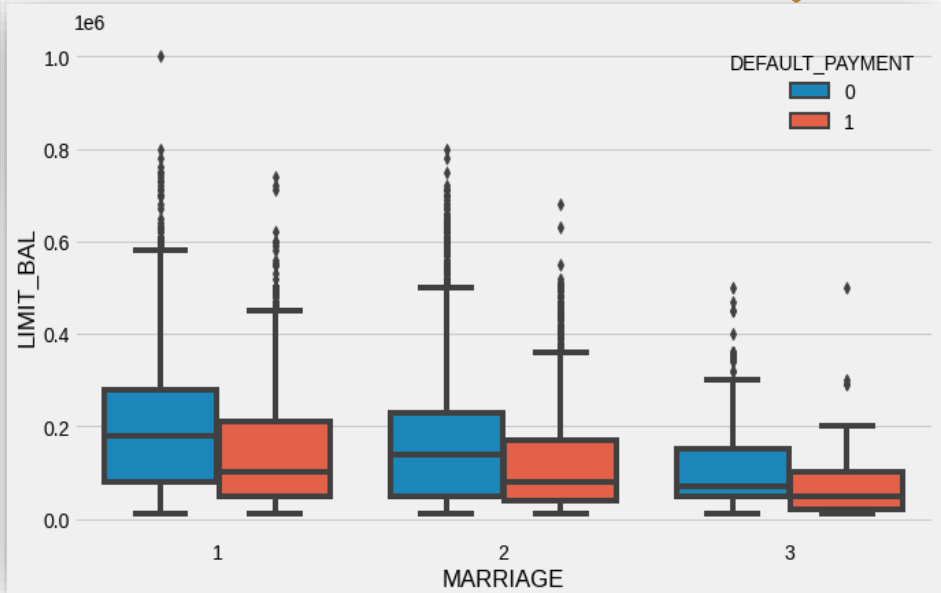
Proportion of defaulter clients is less in both Male & Female as compared to non-defaulters clients.



EDA – Limit Balance vs Education & Marriage



Limit Balance amount is quite high for married clients followed by unmarried clients. Proportion of defaulter clients is also quite high in case of married people as compared to unmarried.

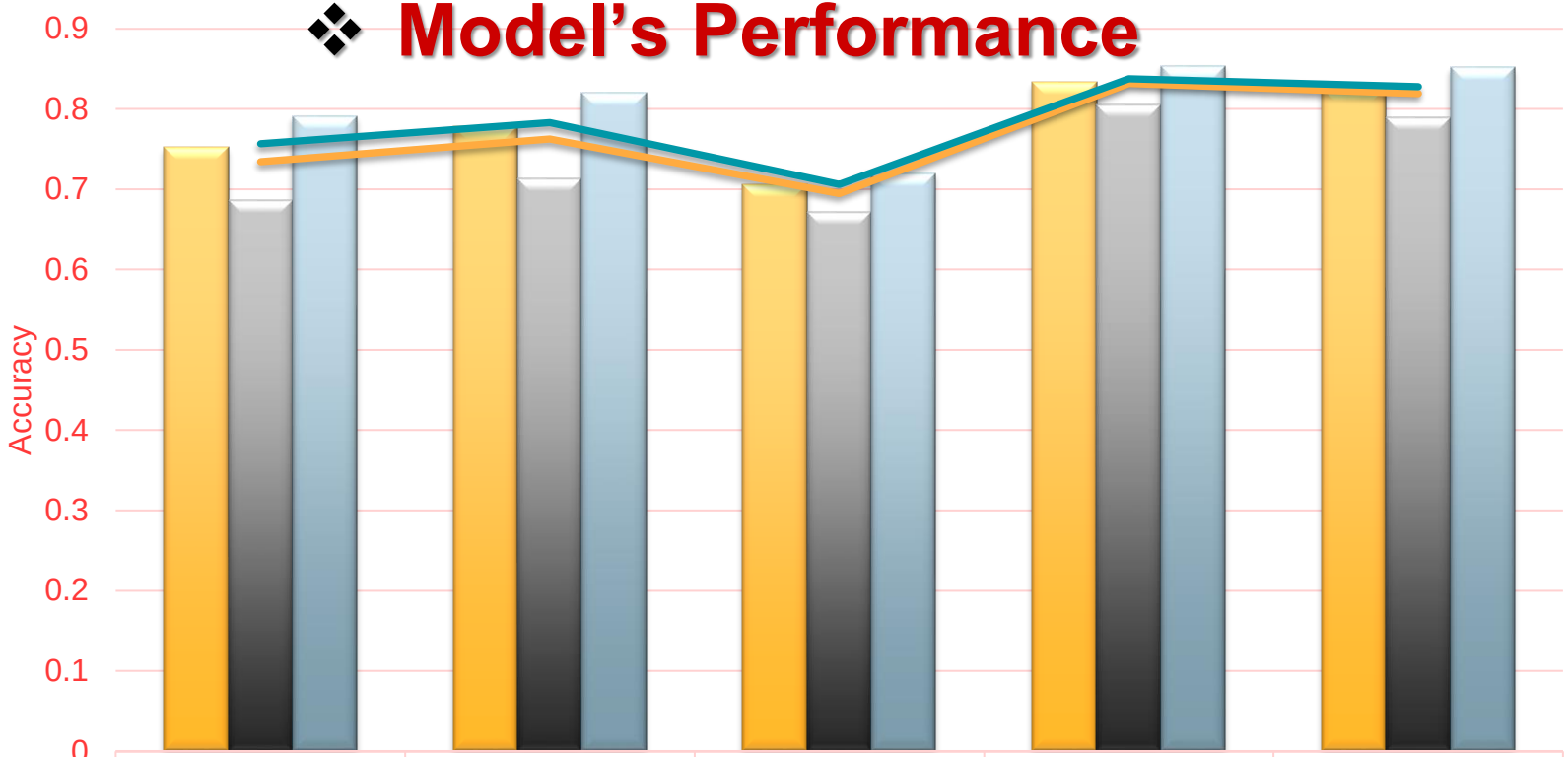


Defaulters are less in each category of Education where Limit Balance amount is quite high in case of Graduate students and less in case of High School students.

❖ Model's Performed

- ✓ Logistic Regression
- ✓ Support Vector Classifier
- ✓ Decision Tree
- ✓ Random Forest
- ✓ XGBoost

❖ Model's Performance



	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	XGBoost
Accuracy	0.7522	0.7781	0.7051	0.8335	0.8261
Precision	0.6856	0.7128	0.6718	0.8049	0.7889
Recall	0.7909	0.8198	0.7195	0.8537	0.8521
f1 score	0.7345	0.7625	0.6948	0.8318	0.8193
roc score	0.7567	0.7829	0.7059	0.8375	0.8278

❖ Model Validation And Selection

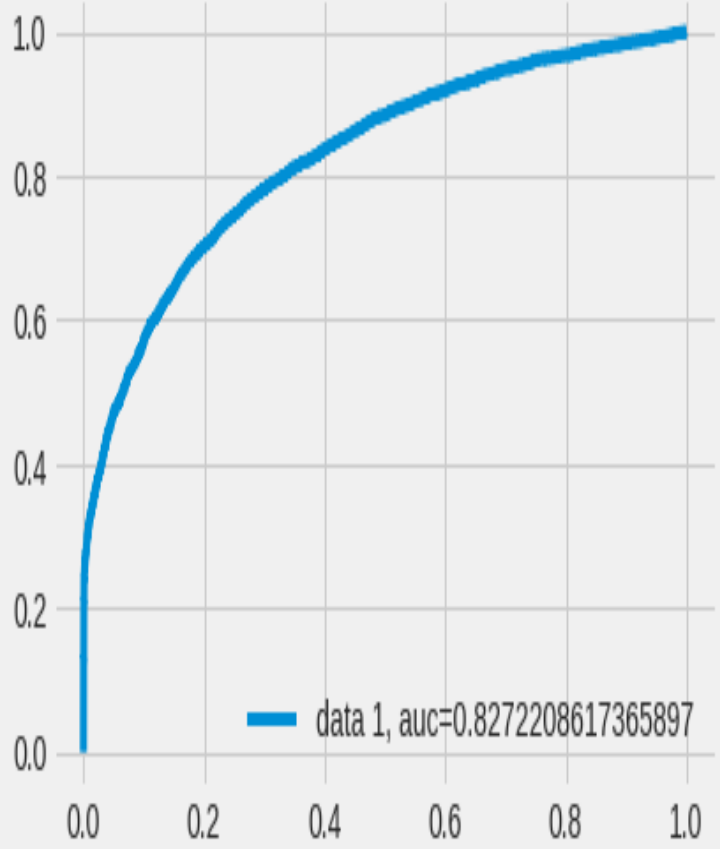
❑ Observation 1:

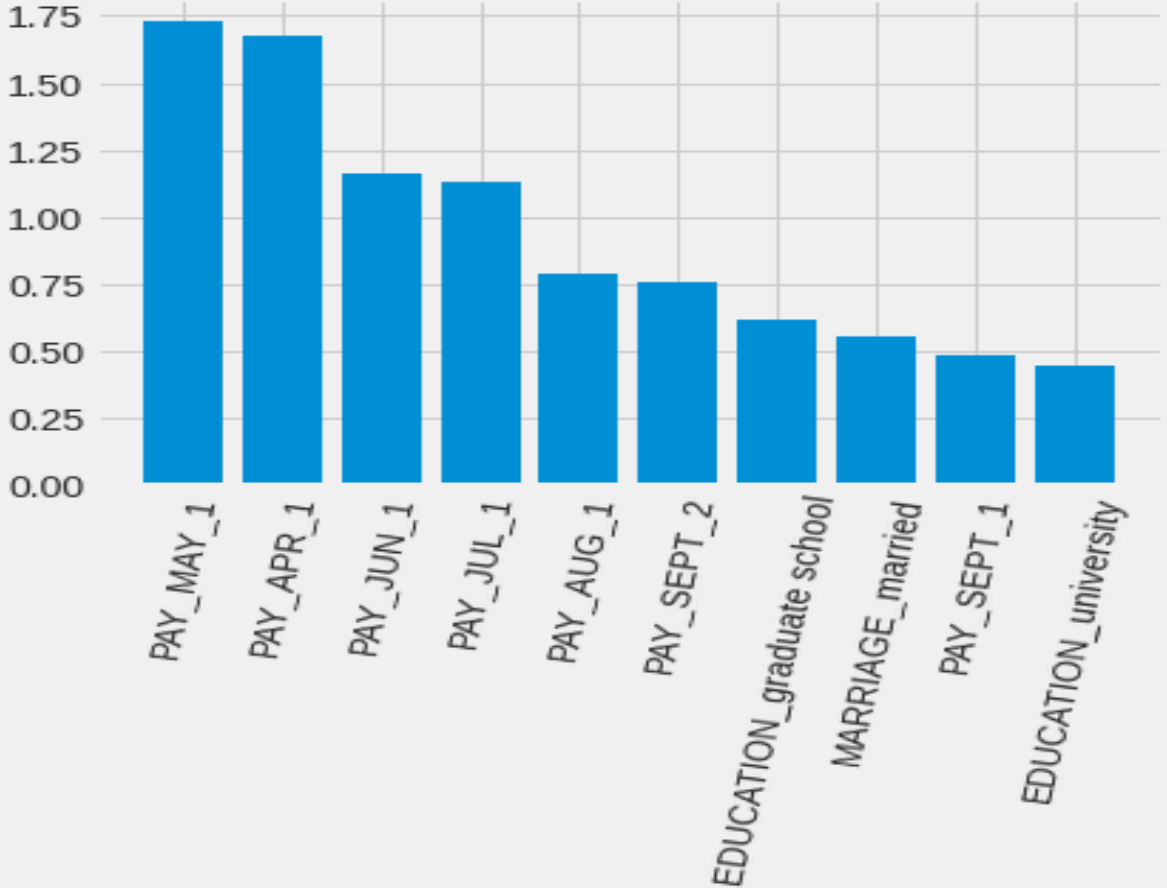
As seen in the previous slide, Logistic Regression, Support Vector Classifier and Decision Tree are not giving greater results.

❑ Observation 2:

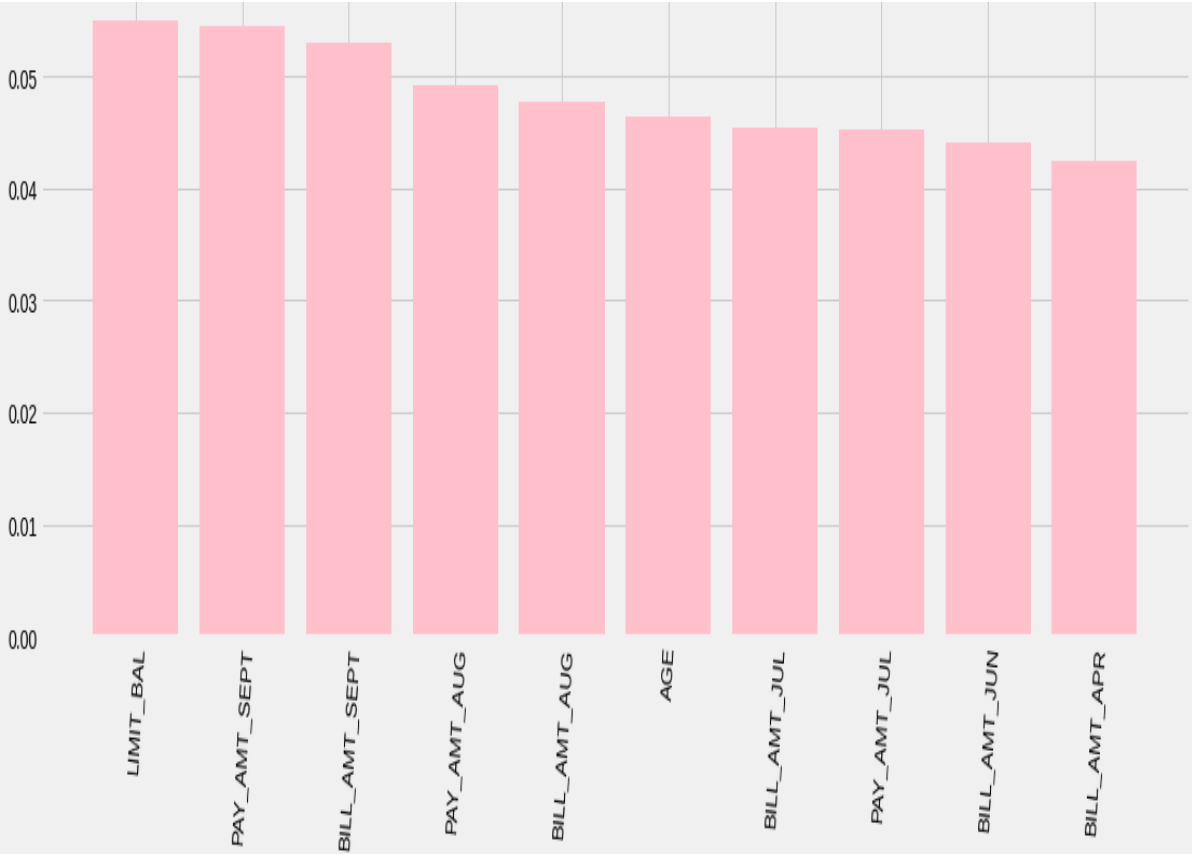
Random forest & XGBoost have best performed near about close to each other in terms of Accuracy and other scores. We can use either Random Forest or Gradient Boosting model for the Prediction of Credit Card Defaulters.

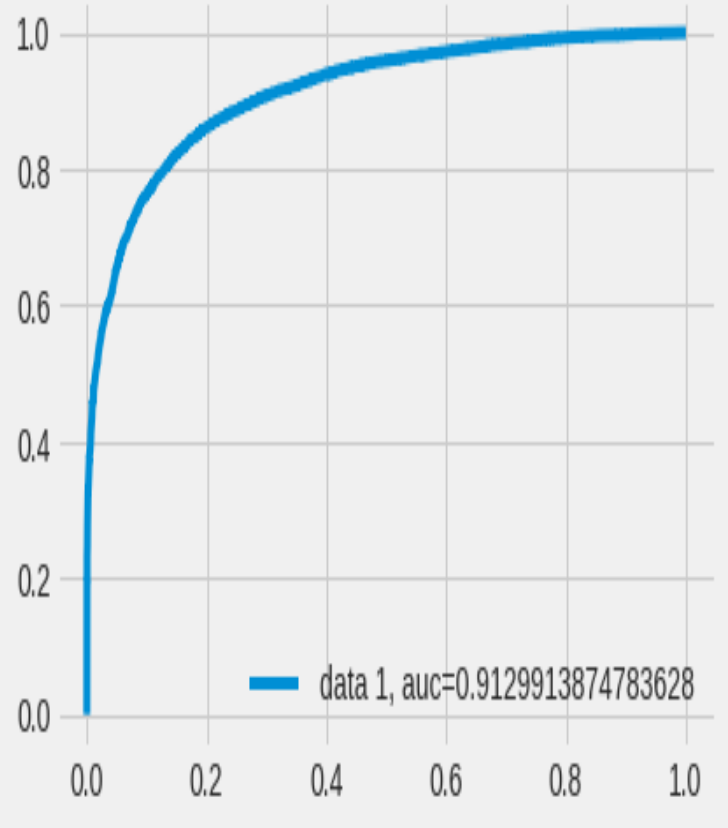
❖ Feature Importance & ROC Curve For Logistic Regression



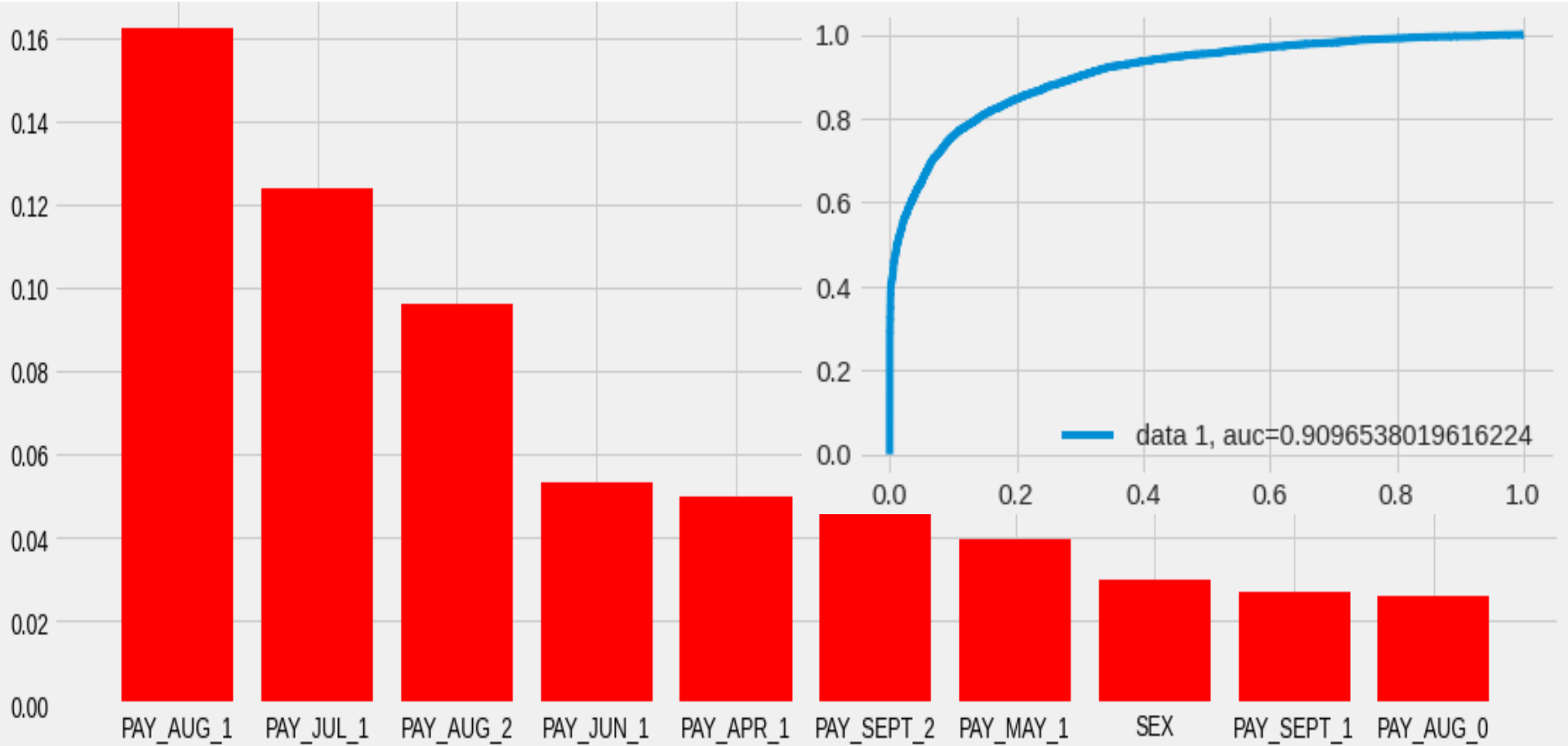


❖Feature Importance & ROC Curve For Random Forest

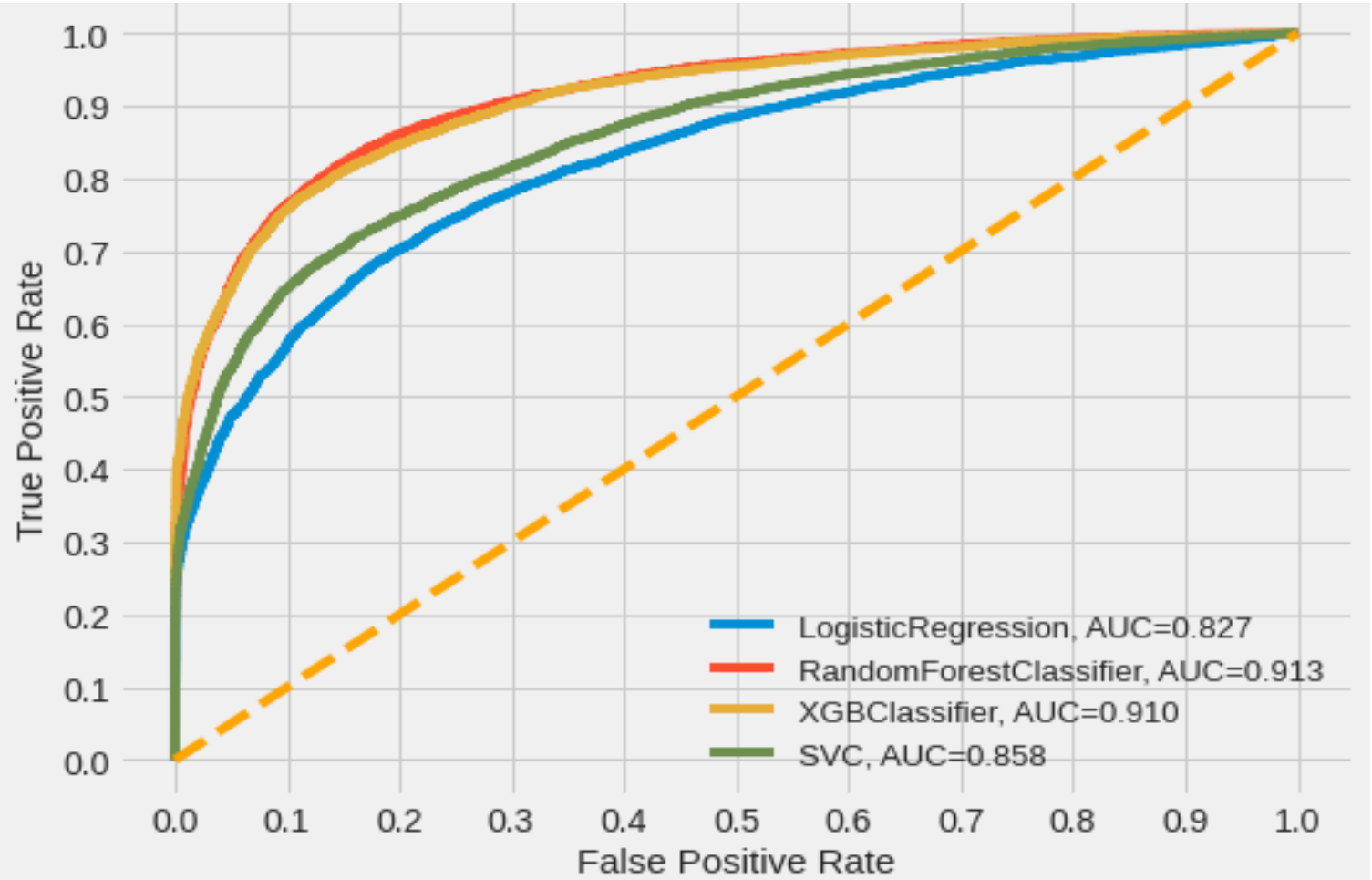




❖ Feature Importance & ROC Curve For XGBoost



❖ Overall ROC Curve Analysis



Random Forest Classifier & XGBoost Classifier have almost equal and highest value for ROC Curve compare to other models, Where Logistic Regression gives the lowest AUC value.

❖ Conclusion

- After performing the various model we the get the best accuracy form 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.

❖ Challenges

- A huge amount of data needed to be deal 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.
- Handling the numerical and categorical data to build high accuracy model.

Thank You!