Capstone Project - 3

Credit Card Default Prediction

Team Members

Kartike

Animesh Chakraborty

Anupam Mishra

❖ Contents

- ❑ Problem Description
- ❑ Objective
- ❑ Approach Overview
- ❑ Modeling Steps
- ❑ Data Description
- ❑ Exploratory Data Analysis
  1) Feature Correlation Graph
  2) Correlation With Default_Payment
  3) Dependent Feature
  4) Independent Features
- ❑ Models performed
- ❑ Model Validation & Selection
- ❑ Feature Importance of XGBoost
- ❑ Overall ROC Curve Analysis
- ❑ Precision – Recall 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

❖ Approach Overview

Data Cleaning → Data Exploration → Predictive Modeling

Understand and Clean

- Find information on undocumented columns values
- Clean data to get it ready for analysis

Graphical and Statistical

- Exam data with visualization
- Verify findings with statistical tests

Machine Learning

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- SVC
- KNN

❖ Modeling Steps

Data Preprocessing

- Feature selection
- Feature engineering
- Train-test data splitting (70%/30%)
- Training data rescaling
- SMOTE oversampling

Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure ROC_AUC on training data

Model Evaluation

- Models testing
- Precision_Recall score
- Compare within the 6 models

❖ Data Description

❑ Dependent variable:

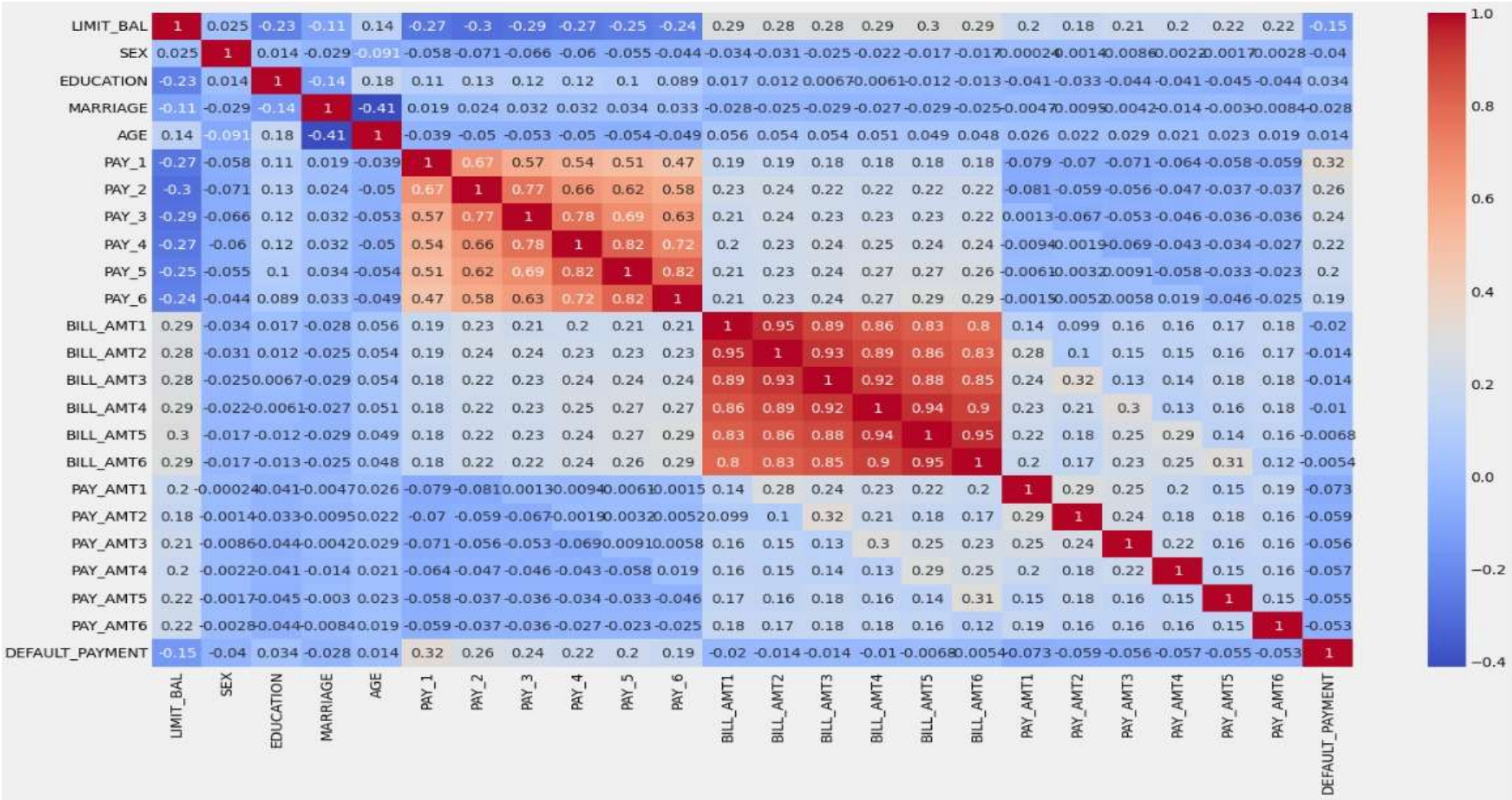
Default Payment: Default payment (1=yes, 0=no)

❑ Independent variables:

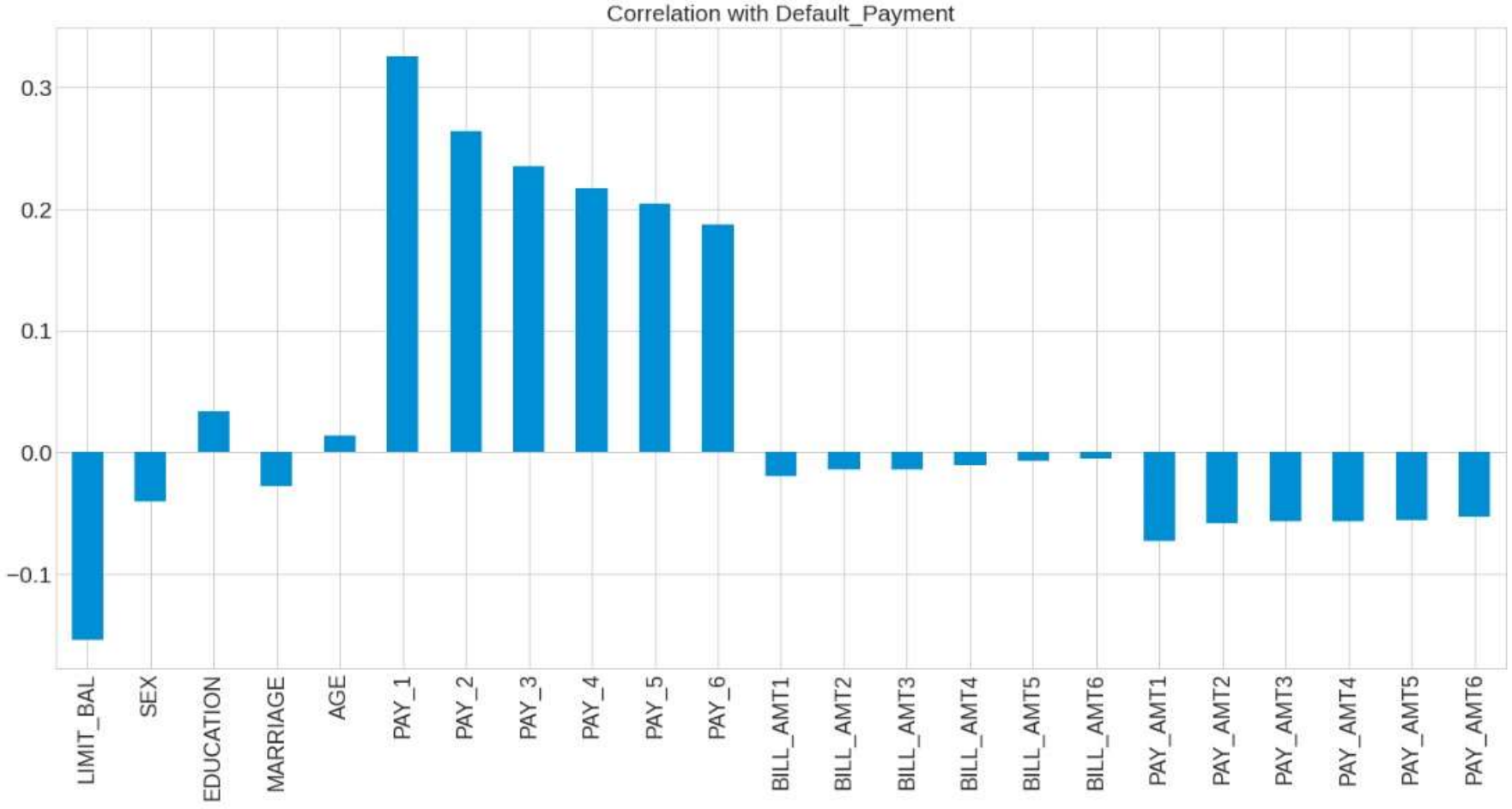
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_1 to PAY_6: 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



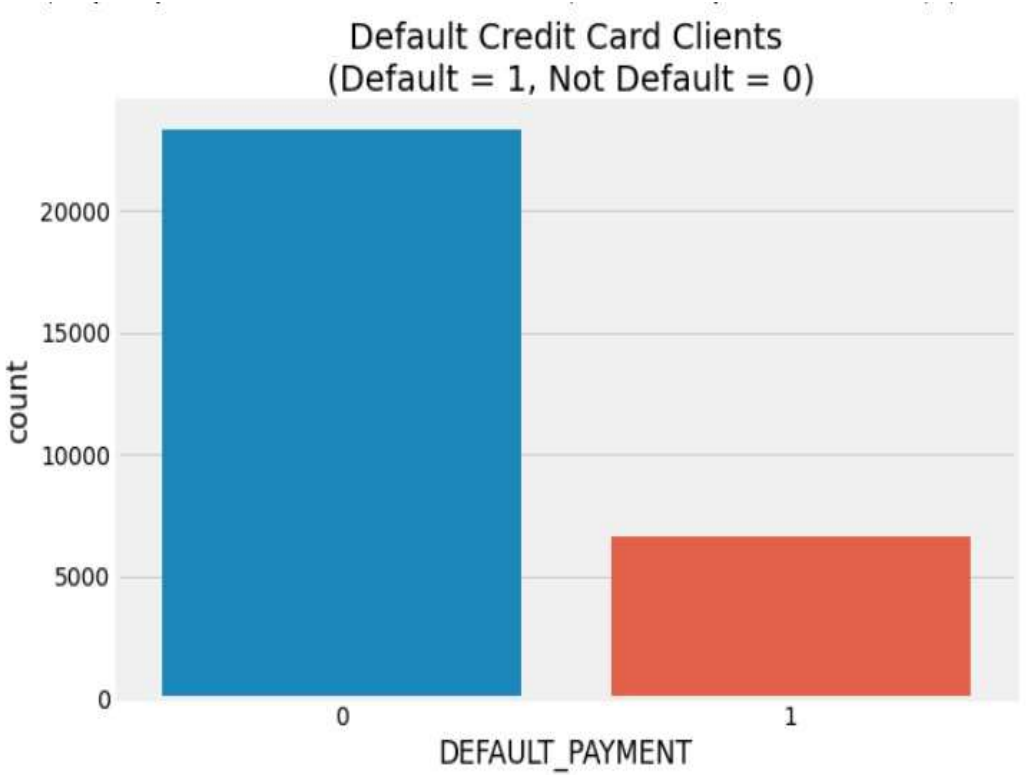
❖ EDA – Correlation With Default_Payment



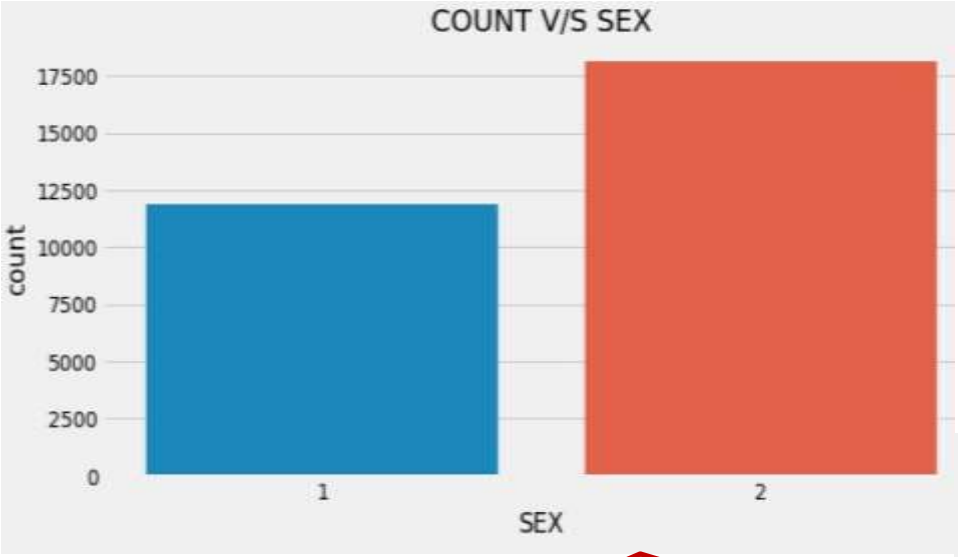
❖ EDA - Define Dependent Feature

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 23364 clients are not defaulters where 6636 clients are defaulters.

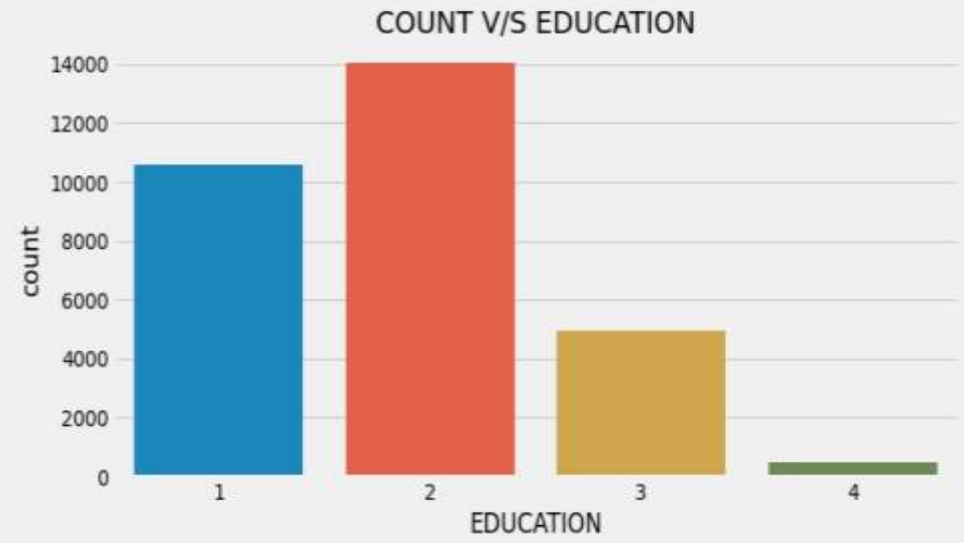


❖ EDA – Gender Count & Education Count

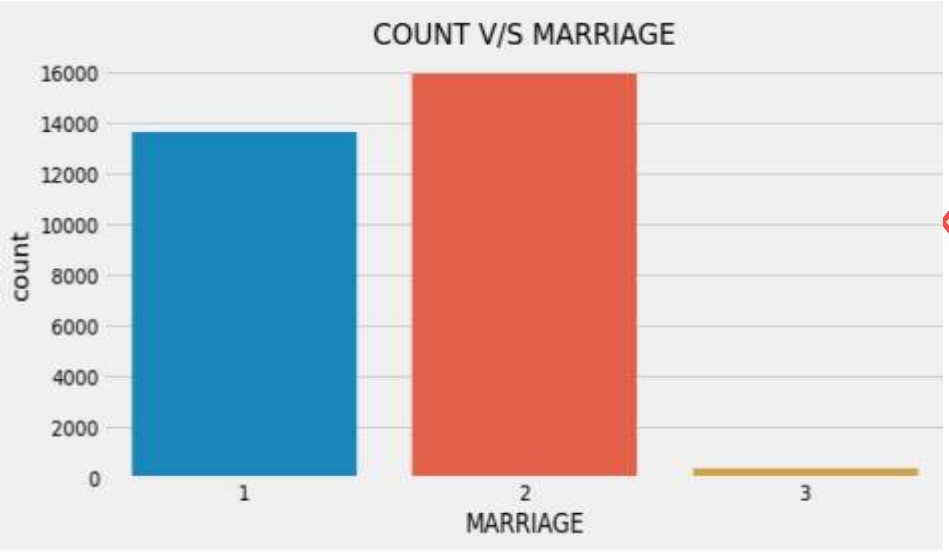


More number of credit card holders are university students (approx. 14030) followed by Graduates (approx. 10585) and then High school students (approx. 4917) and others are (approx. 468).

The Countplot for gender showing that Females are leading in the use of Credit Card. Approx. 18112 females are using the credit cards at the same time males are somewhere lagging in use of credit cards (approx. 11888 credit cards were issued for males.

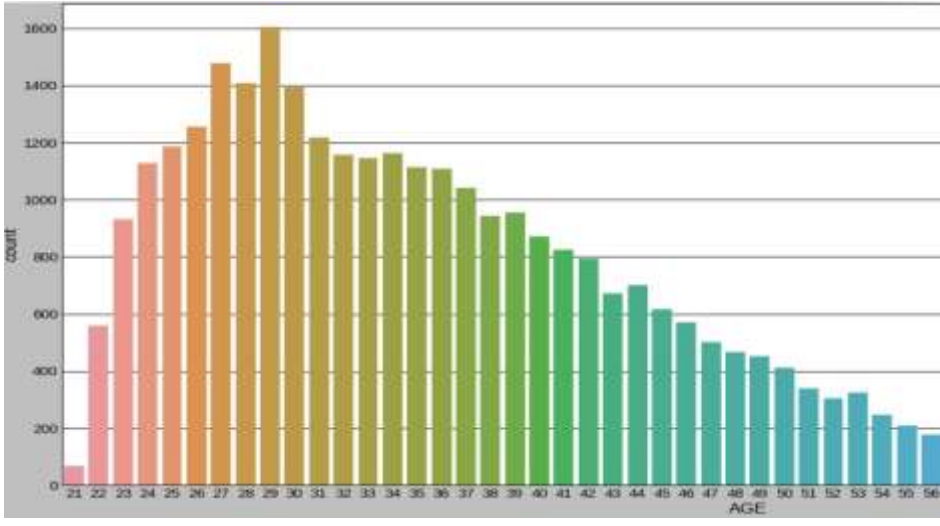


❖ EDA – Marriage Count & Age Count

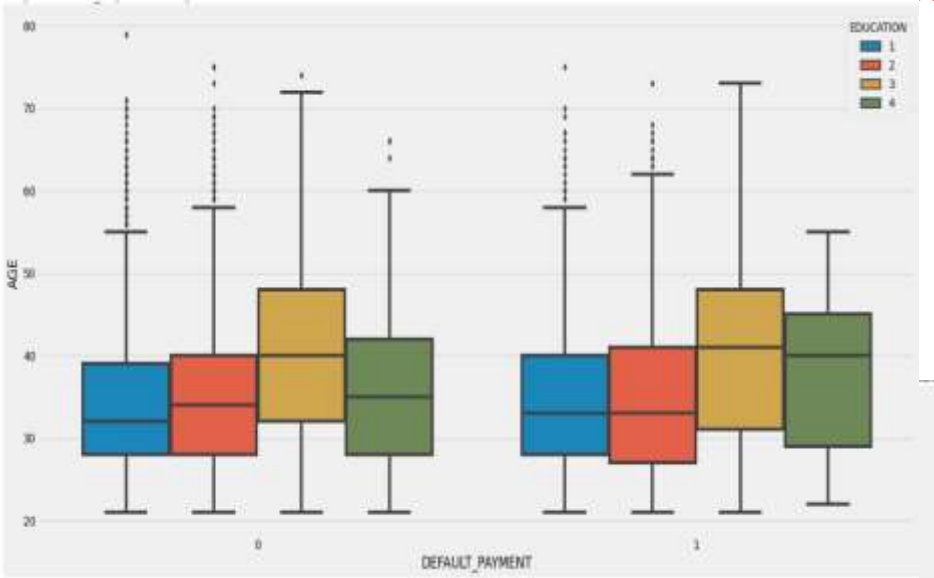


As we can see that, large amount of credit card holders are singles. There are 15964 clients who are unmarried and using credit cards, also 13659 credit cards users are married which are quite near to the unmarried clients and other which are 377 very less as compare to others.

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.

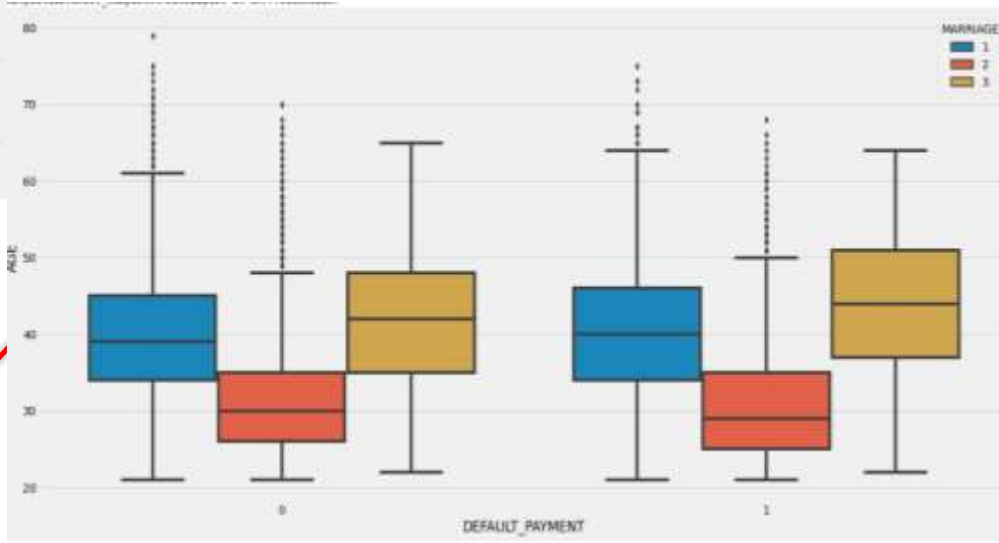


❖ EDA – Default Payment vs Age With Education & Marriage

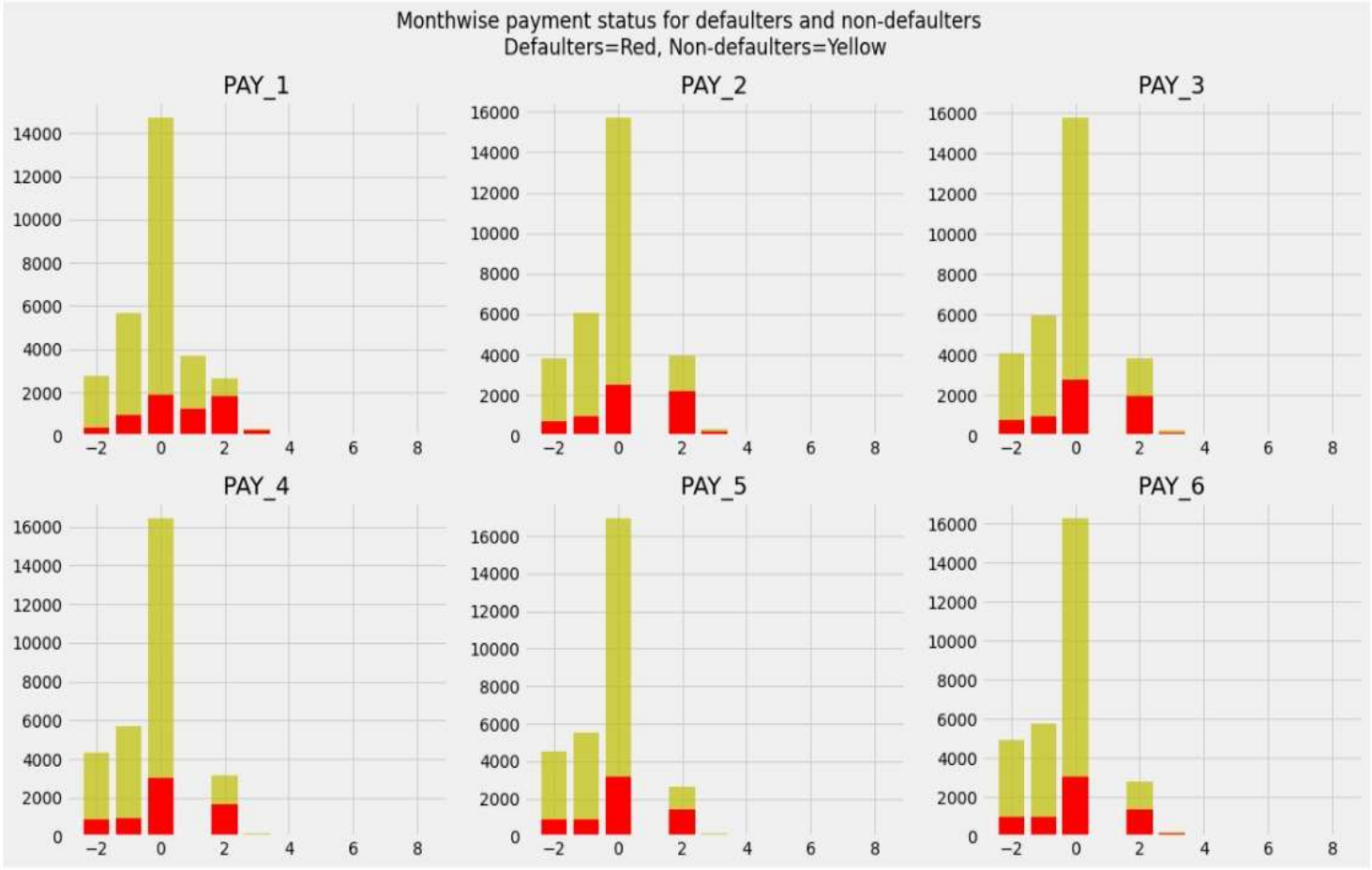


From this Boxplot, we can conclude that more number of defaulters are high school students followed by university students.

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

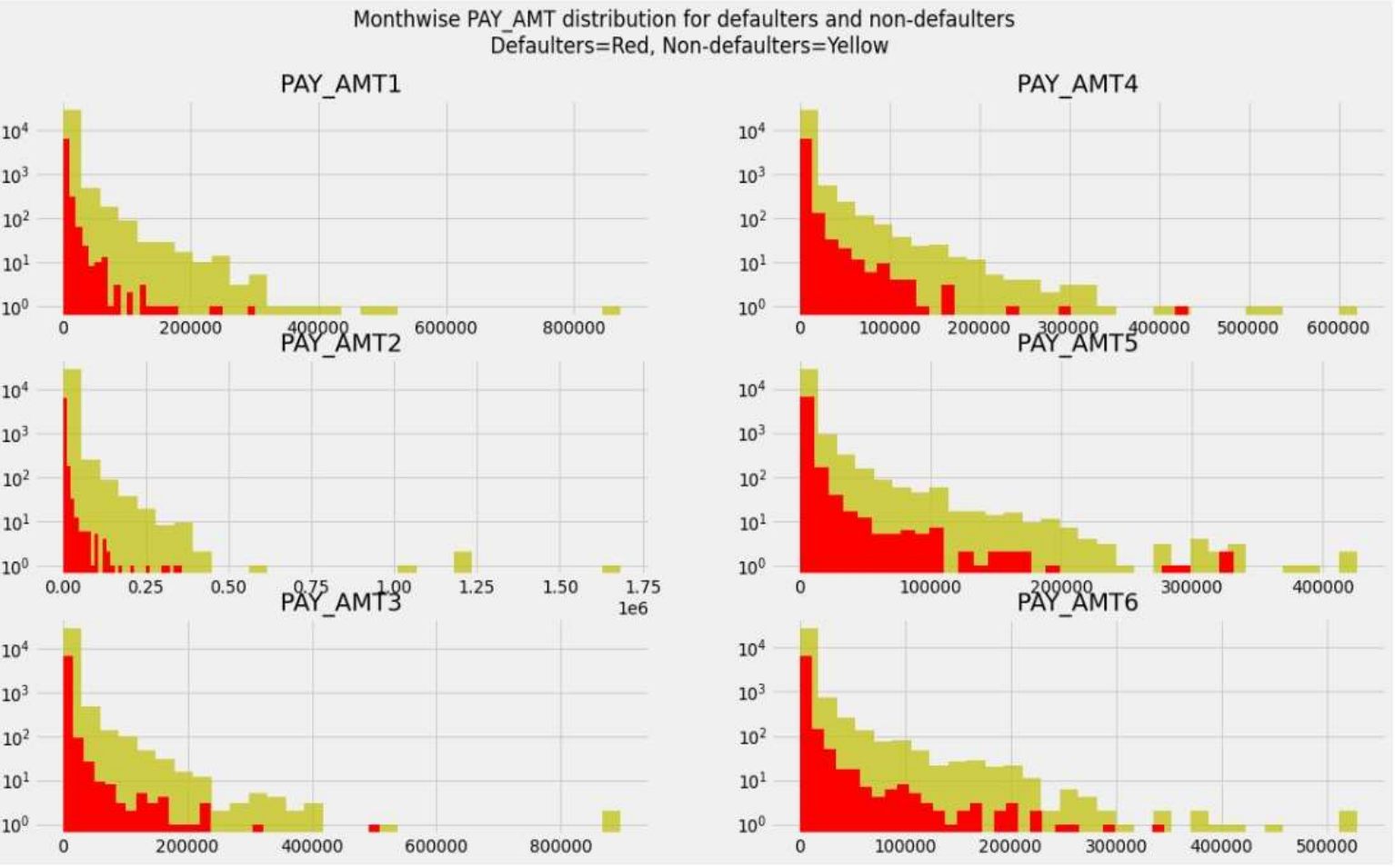


❖ EDA – Monthwise Payment Status



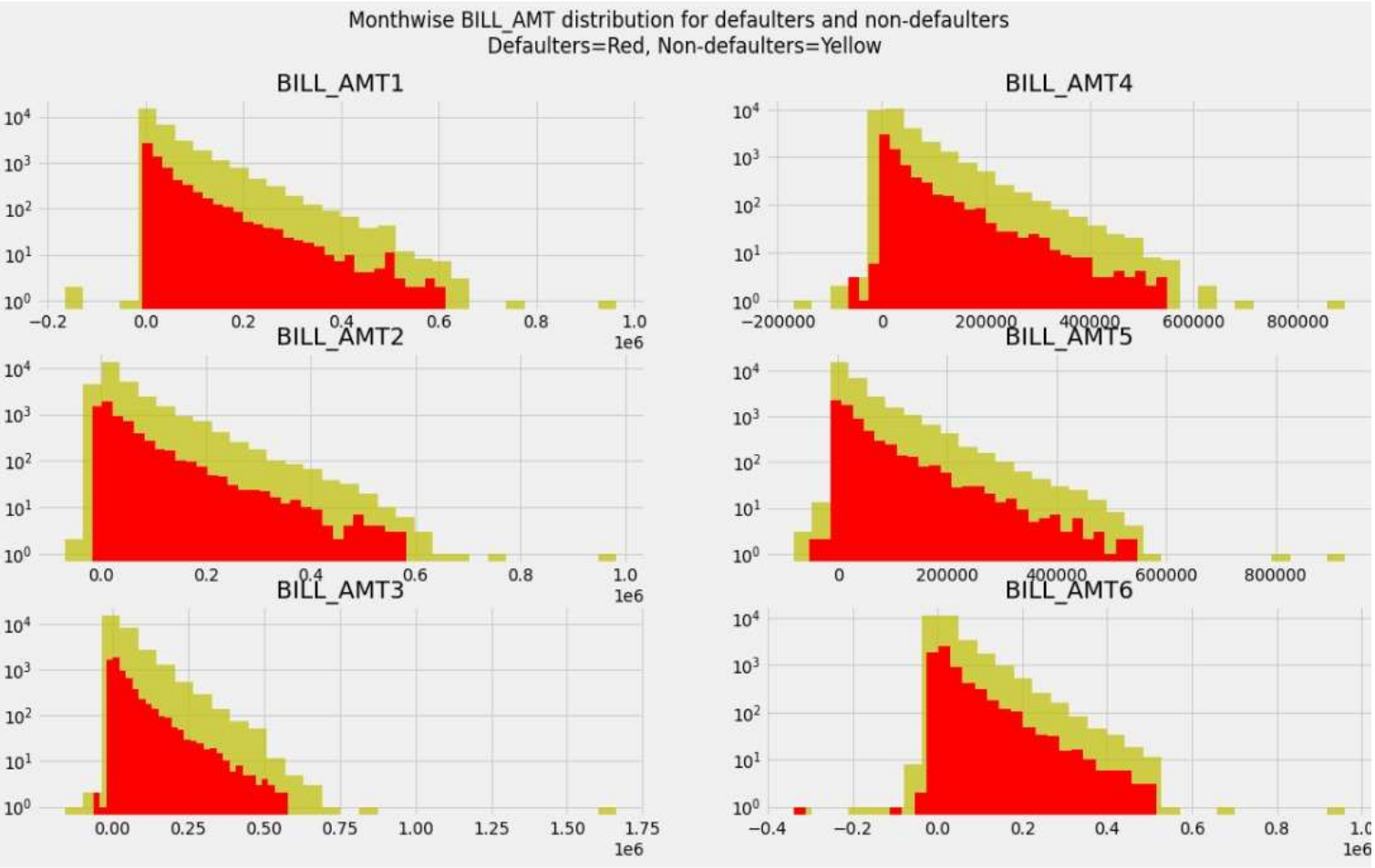
Highest number of defaulters are in May & June followed by April, July & August where lowest number of defaulters are from month of September.

❖ EDA – Monthwise Payment Distribution



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

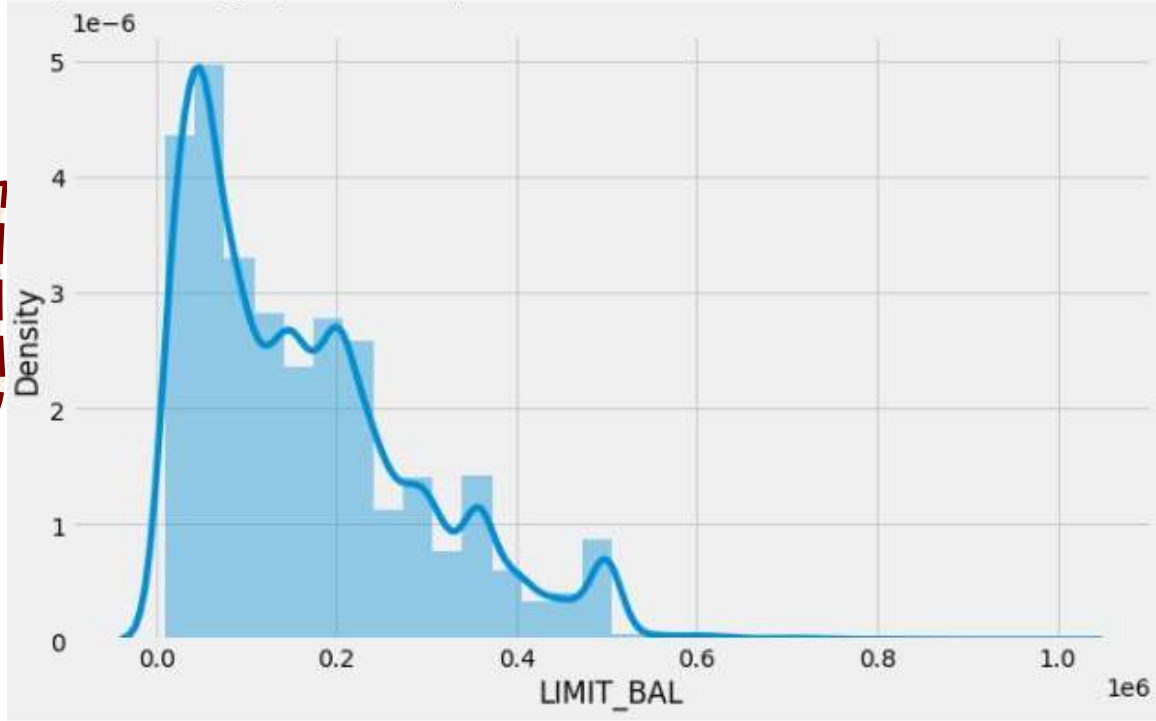
❖ EDA – Monthwise Bill Amount Distribution



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 – 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.



❖ Model's Performed

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

❖ Correct Imbalanced Classes

- Fit every model with SMOTE oversampling for comparison.
- Random Forest has highest ROC_AUC score out of all model.

	Model	ROC_AUC Score
0	Logistic Regression	0.791440
1	Decision Tree	0.742458
2	Random Forest	0.903192
3	XGBoost	0.858813
4	SVC	0.832468
5	KNN	0.822064

❖ Model Comparisons

- Compare the models to each other.
- SVC Model has overall best precision, recall and F1 score.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.723375	0.720152	0.730632	0.725354
1	Decision Tree	0.600185	0.660201	0.412755	0.507945
2	Random Forest	0.649832	0.890335	0.341704	0.493865
3	XGBoost	0.648976	0.623113	0.753888	0.682291
4	SVC	0.776589	0.790412	0.752746	0.771120
5	KNN	0.757757	0.754186	0.764731	0.759422

❖ Model Validation And Selection

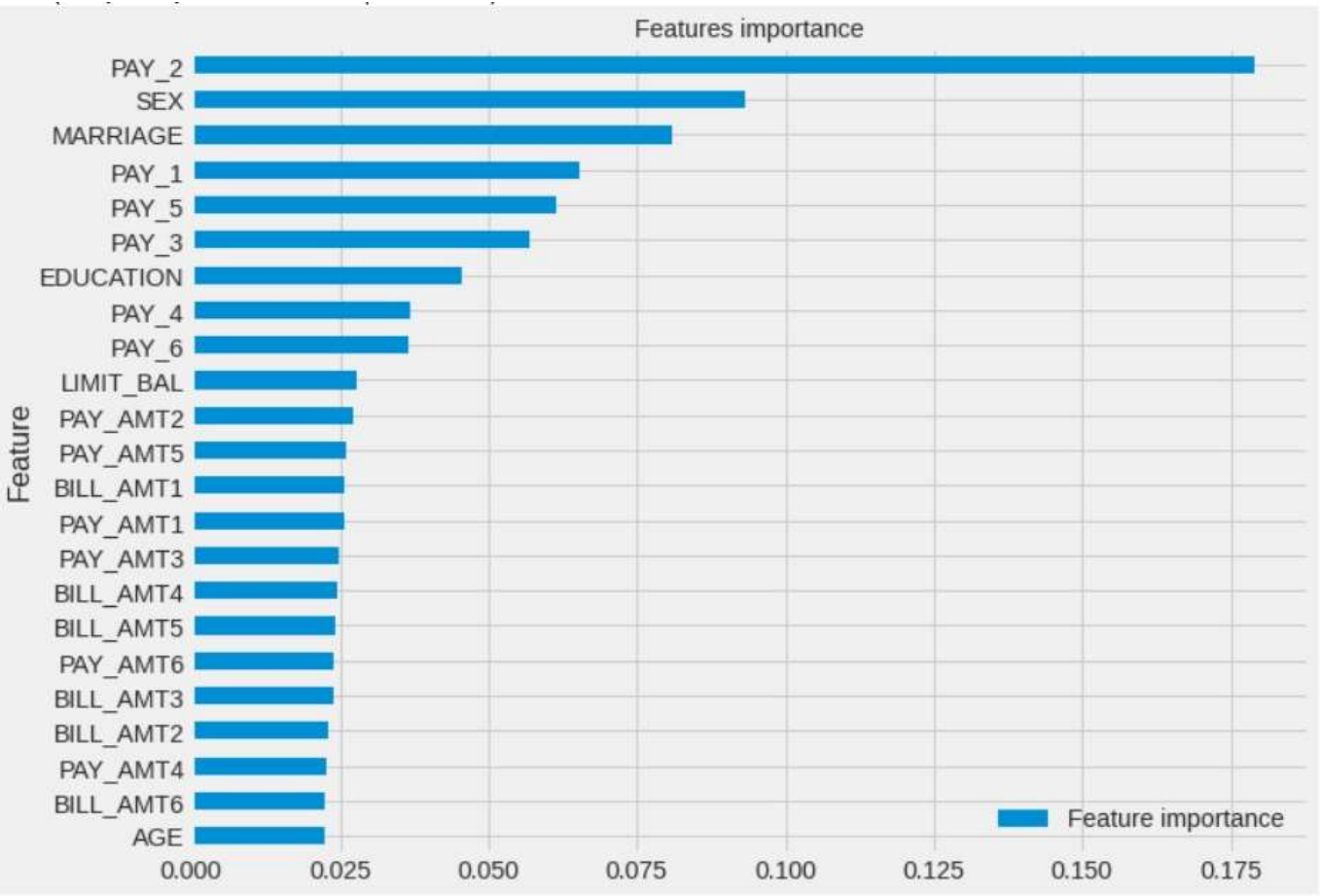
❑ **Observation 1:**

As seen in the previous slide, Decision Tree, Random Forest and XGBoost are not giving better results as compared to other models.

❑ **Observation 2:**

Support vector classifier & KNN have best performed near about close to each other in terms of Accuracy and other scores. We can use either SVC or KNN model for the Prediction of Credit Card Defaulters.

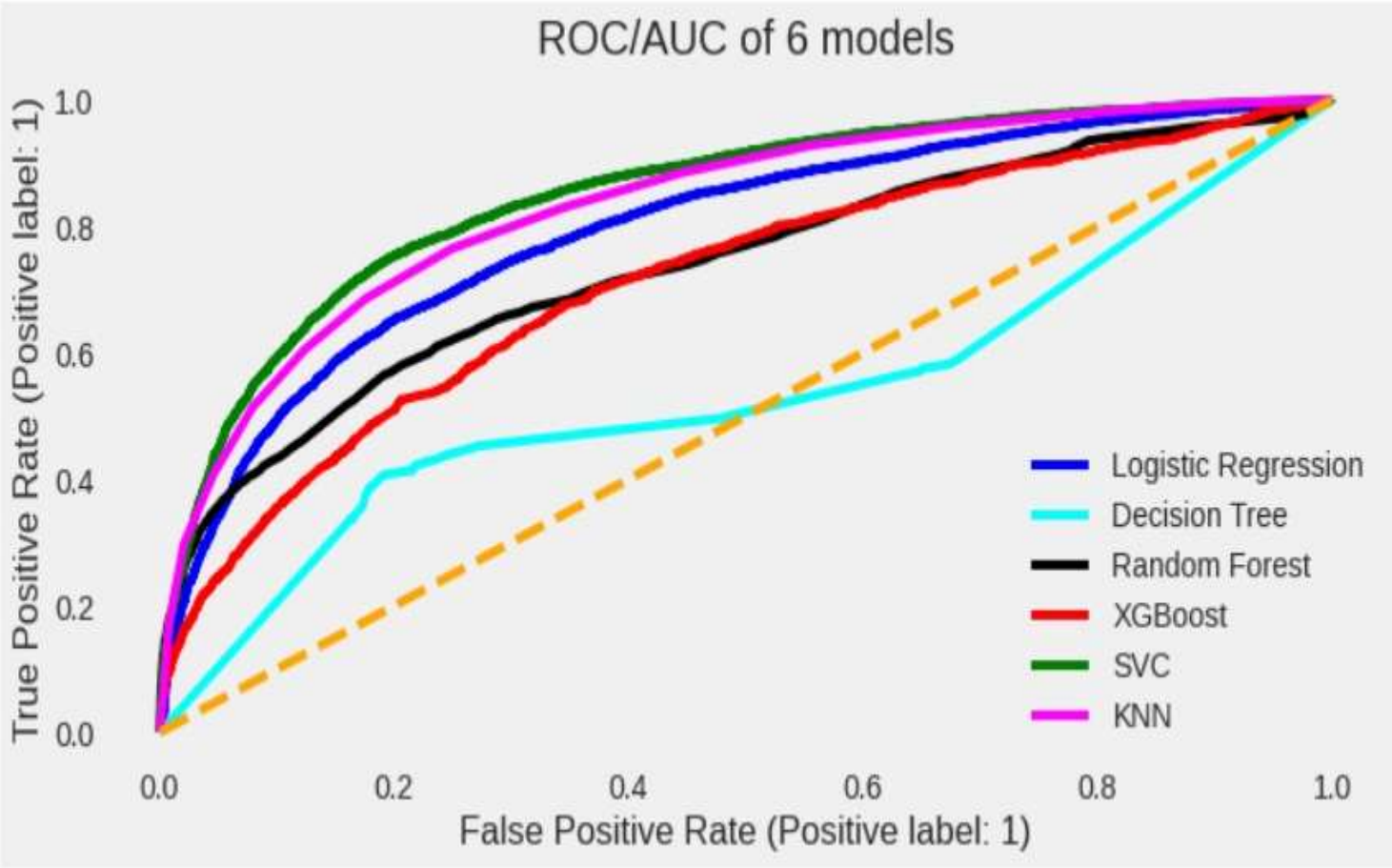
❖ Feature Importance For XGBoost



XGBoost Classifier feature importances plot.

- ★ PAY_2: the month prior to current month's payment status.
- ★ SEX: gender (M/F)
- ★ MARRIAGE: marital status
- ★ PAY_1: most recent month's payment status.

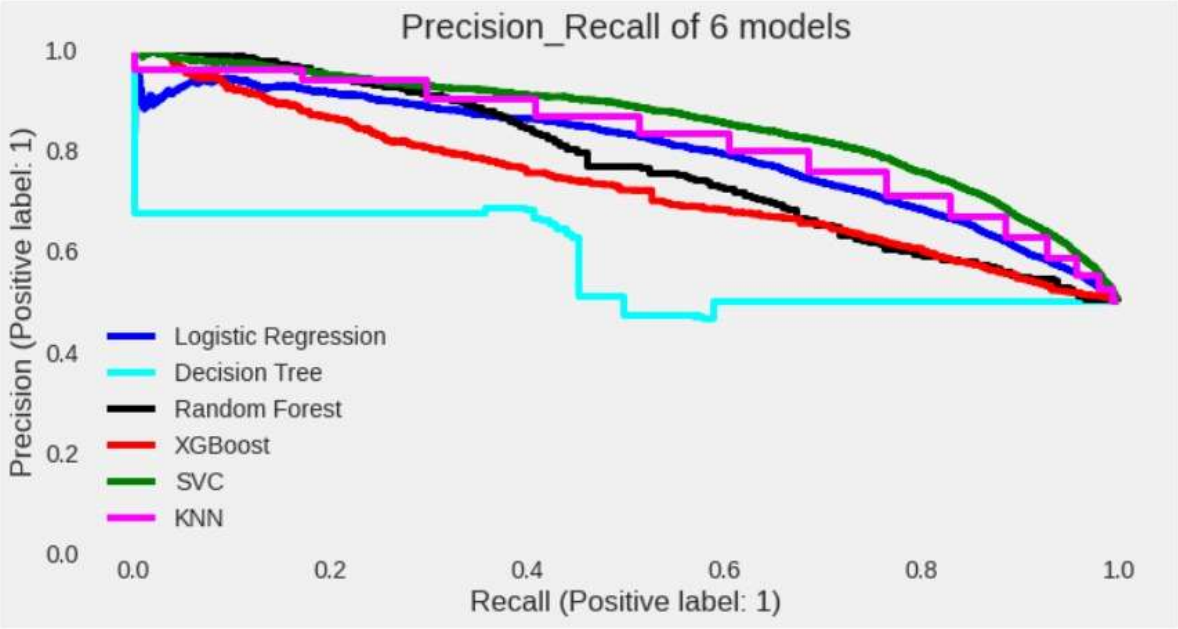
❖ Overall ROC Curve Analysis



Support Vector Classifier (SVC) & KNN have almost equal and highest value for ROC_AUC compared to other models, otherwise Decision Tree whereas gives the lowest ROC_AUC value.

❖ Precision - Recall Analysis

- Compare within 6 models.
- SVC (green line) has the best precision_recall score.



Terminology:

★ Recall: how many 1s are being identified?

★ Precision: Among all the 1s that are flagged, how many are truly 1s?

★ Precision and recall trade-off: high recall will cause low precision

❖ Conclusion

- After performing the various model we the get the best accuracy from the SVC and KNN.
- Decision Tree is the least accurate as compared to other models performed.
- SVC(Support Vector Classifier) 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!