Capstone Project-3

Credit Card Default Prediction

Team Members

Adityasingh Thakur
Meenakshi
Tushar R. Wagh

Let's Predict!

1. Overview & Objective
2. Data Pipeline
3. Data Summary
4. EDA (Exploratory Data Analysis)
5. Feature Engineering
6. Handling Imbalance data
7. Building model
    a) Logistic Regression
    b) Random Forest Classifier with CV
    c) Support Vector Classifier
    d) K-Neighbor Classifier with CV
    e) XG Boosting CV
    f) Decision Tree Classifier
8. Model Evaluation
9. Conclusion

Overview & Objective

Overview

Credit card is a commonly used transaction method in modern society and one of the main business of banks. For banks, it helps the bank to generate interest revenue but at the same time, it raise the liquidity risk and credit risk to the bank.

Objective

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.

Data Pipeline

Data Preprocessing:- At this stage,

- We check for duplicate values and missing values and treat them if any.
- Detecting the outliers and removed it.
- We check the data-type of the features present in our dataset transform them if necessary.
- Rename the column for better understanding.

Exploratory Data Analysis (EDA):- At this stage, we conduct an EDA on the selected features in order to better understand their spread , pattern and relationship with the other features. It gives us an intuition as to what is going on in the dataset.

Handling Imbalance Dataset:- We have used SMOTE(**Synthetic Minority Oversampling Technique**) technique which can handle imbalance dataset.

Model Building:- At this stage, we apply various models to understand which one will give us the best result.

Data Summary

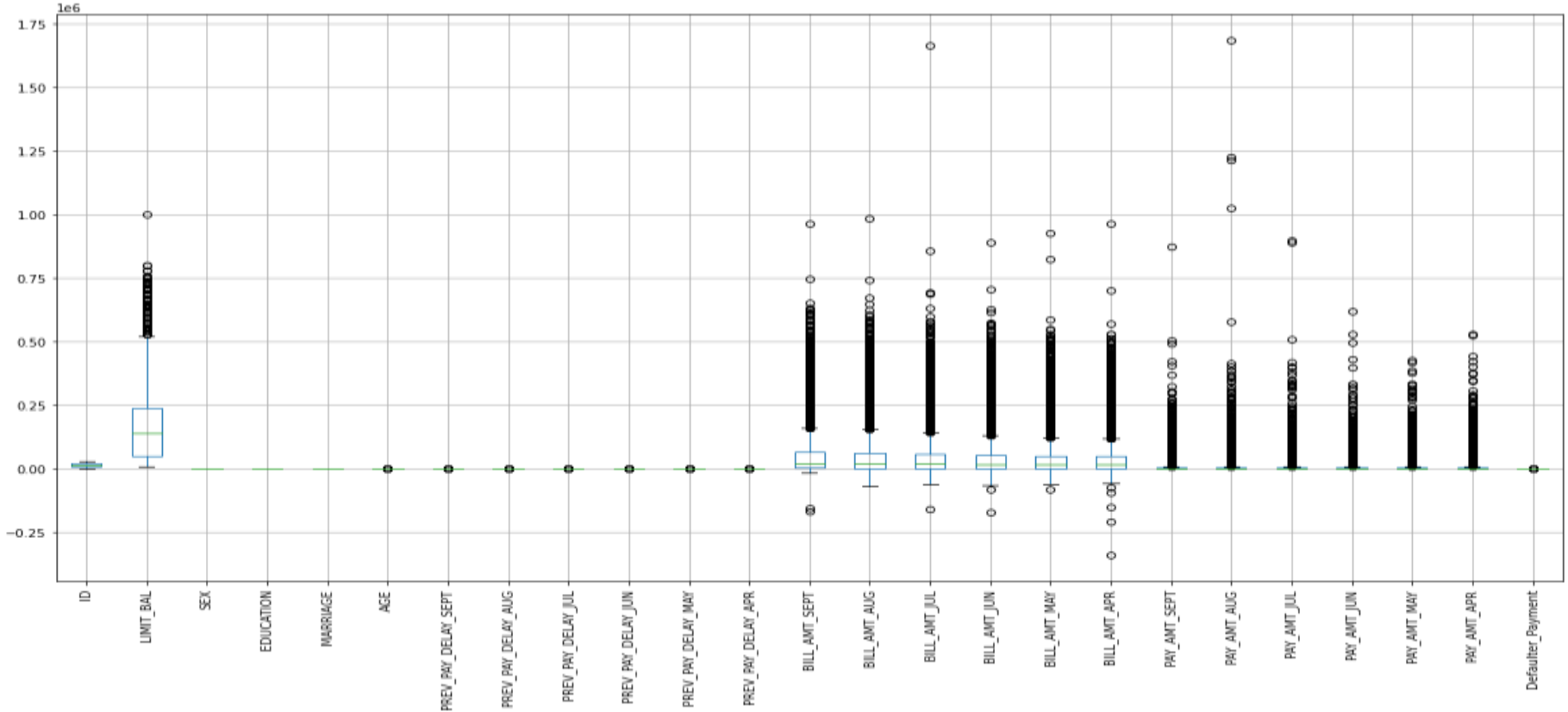
We have credit card default prediction dataset. It has following features (Columns):

1 **ID:-** Denotes a special unique number for clients in our dataset.
2 **LIMIT_BAL:-** It includes both the individual consumer credit and his/her family credit.
3 **GENDER:-** Includes 1 for Male And 2 for Female.
4 **EDUCATION:-** Includes graduate school , university , high school and others clients in our dataset.
5 **MARRIAGE:-** Includes Married , Single and others clients in our dataset.
6 **PAY_0 – PAY_6 :-** Represents the history of past payments.
7 **BILL_AMT1-BILL_AMT6:-** Represents the amount of bill statements from September 2005 to April 2005.
8 **PAY_AMT1-PAY_AMT6:-** Represents the amount of previous payments like amount paid in September 2005 to April 2005.

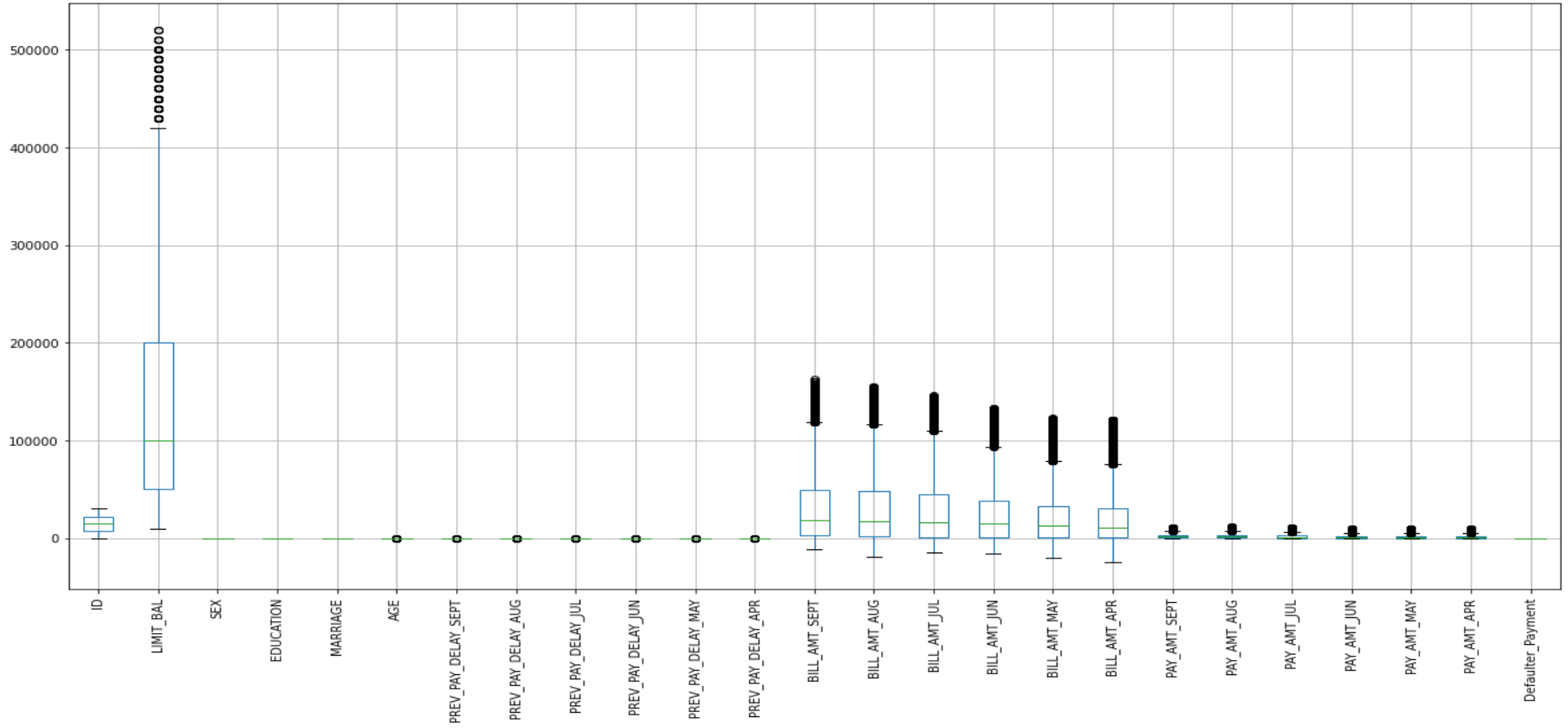
1.

Note: 'default payment next month' will be our Dependent variable & Others will be Independent variable.

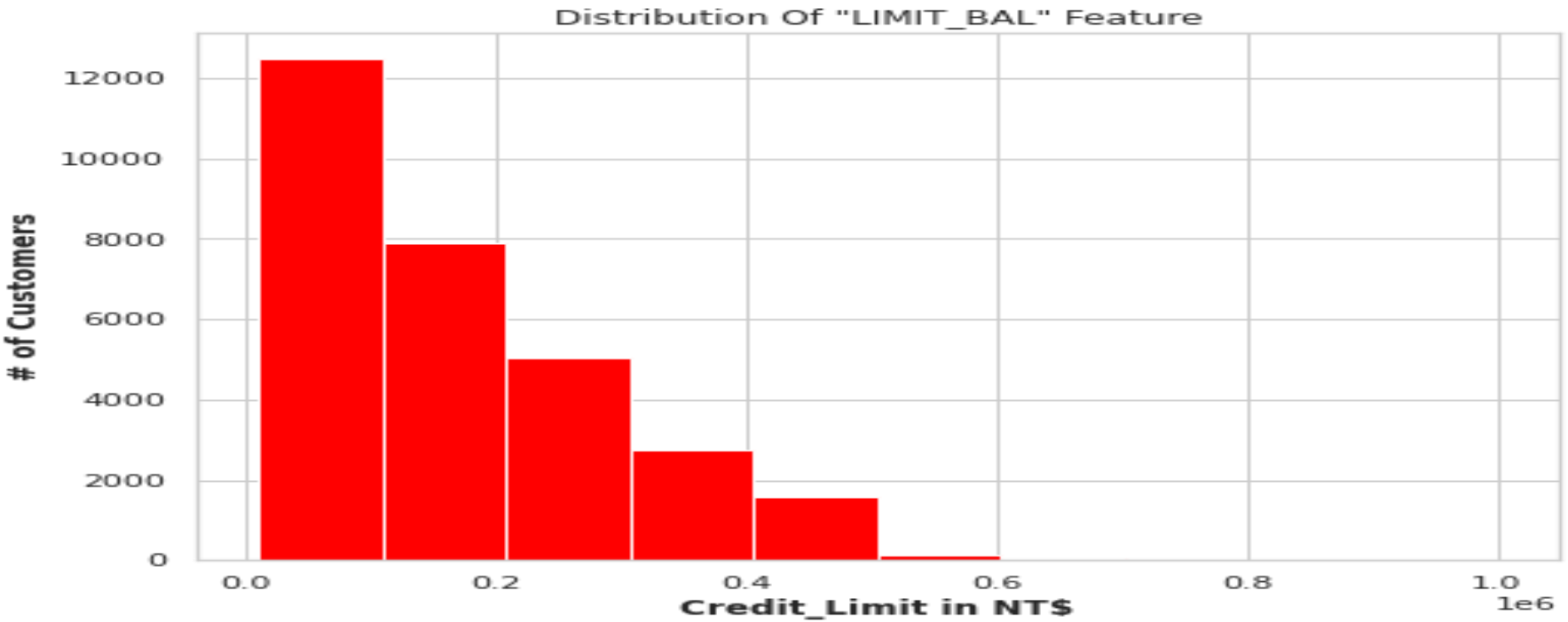
Box plot for whole dataframe to detect outliers



Box plot after removing outliers from the dataset



Distribution Of ‘LIMIT_BAL’Feature

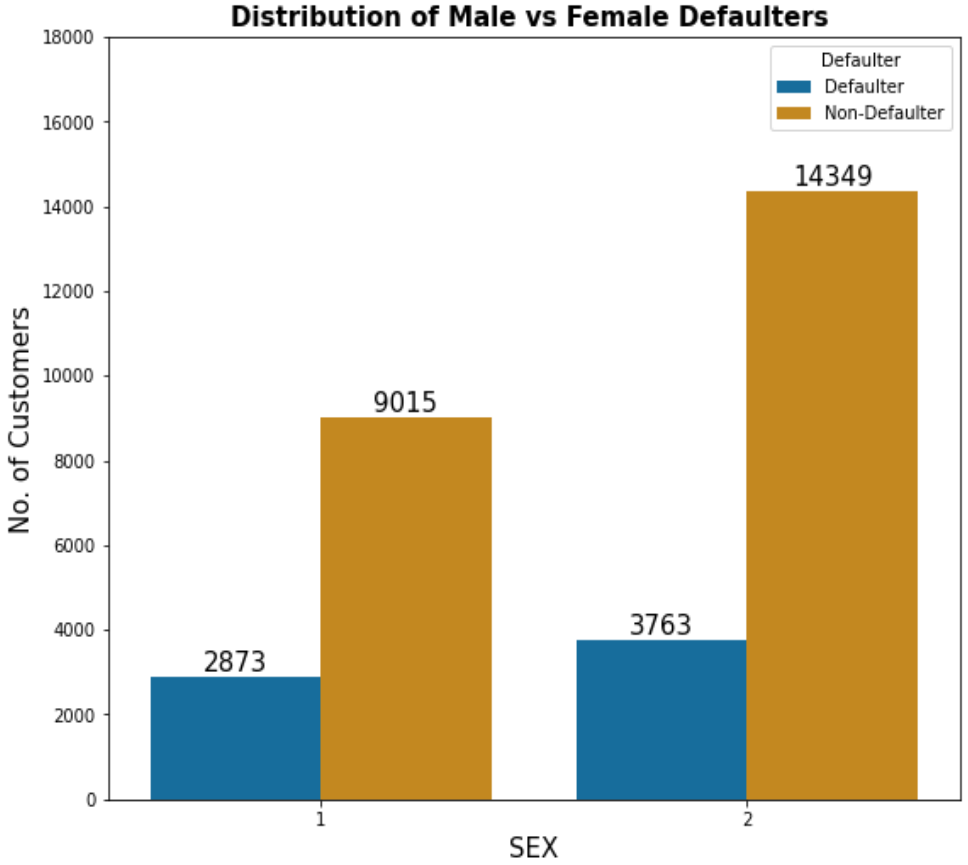


- We can see that the above distribution is a positively skewed distribution i.e. mean != median.
- We can also say that in 0.1 to 1.1 range there are more customers and as the credit limit increases the number of customers tends to decrease.

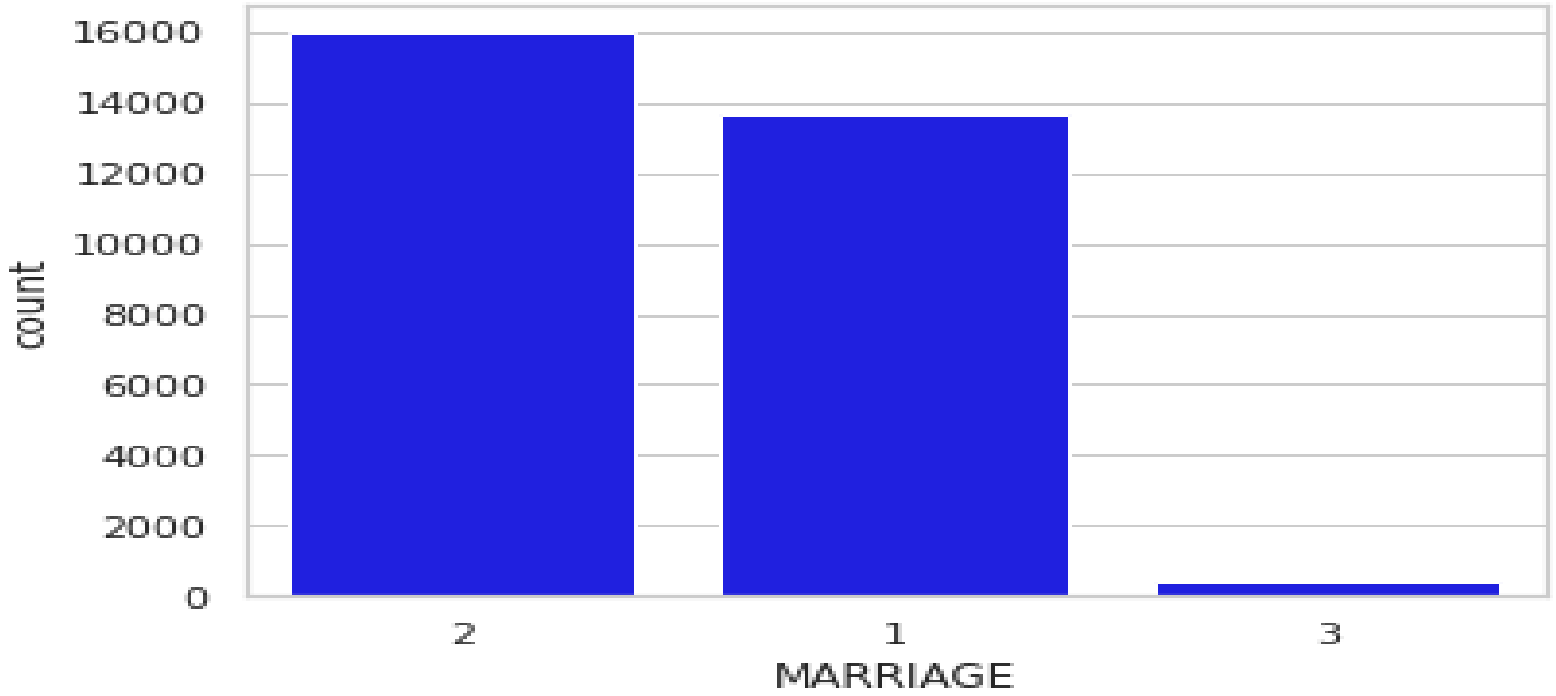
EDA(Exploratory Data Analysis)

The graph conclude that :

- Number of female defaulters are more than male defaulters
- Around 9015 are male non-defaulters and 14349 are female non-defaulters
- Male defaulters are less



Countplot For The Feature ‘MARRIAGE’

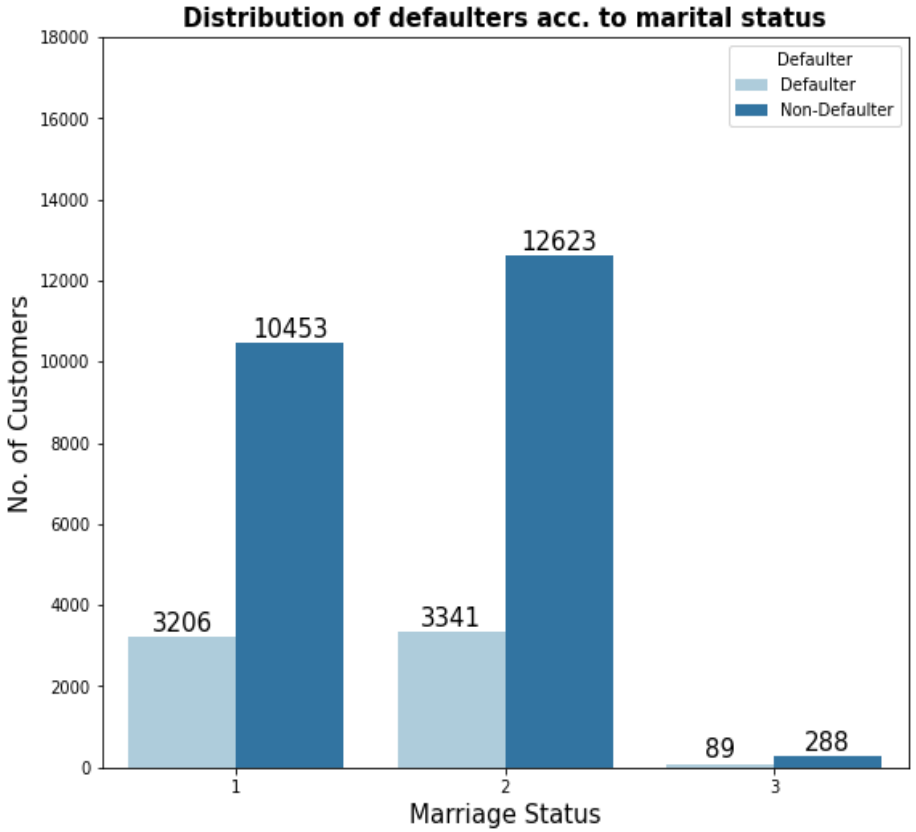


- We can easily say that there are more number of single customers as compared to married customers in our dataset.

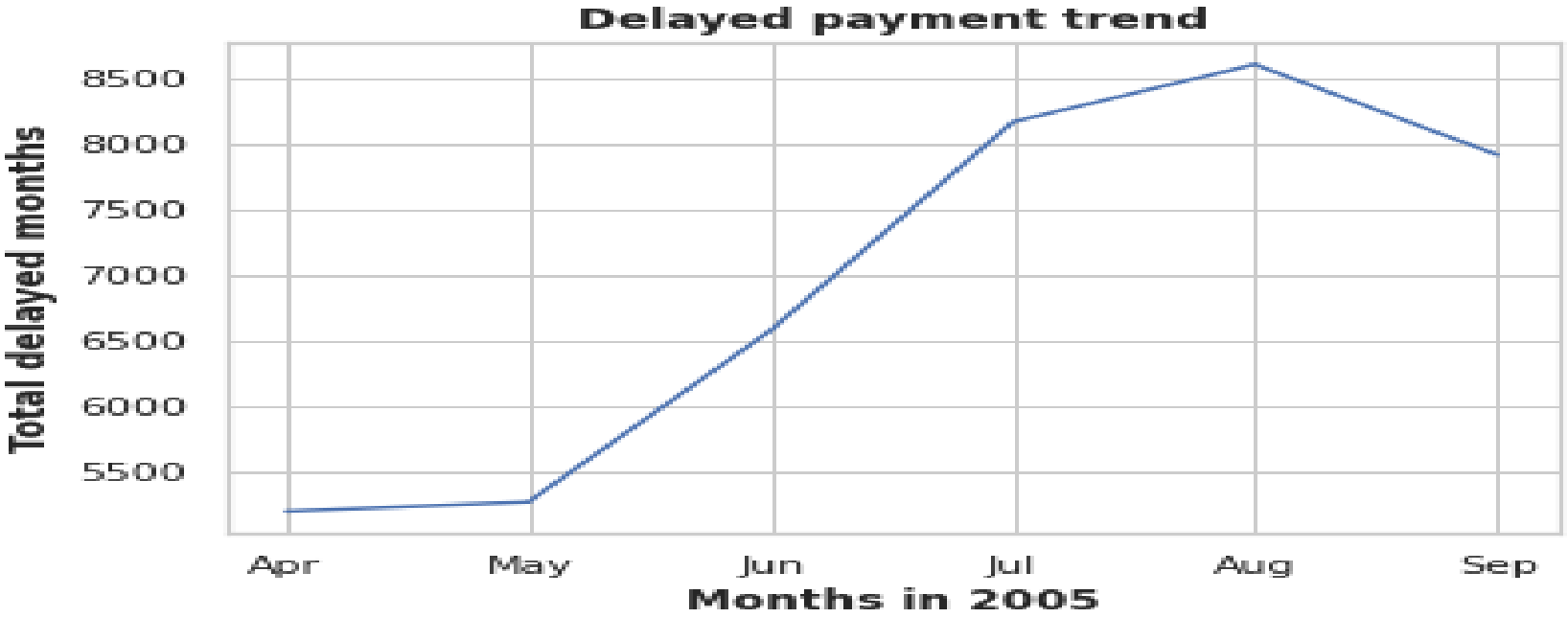
Defaulters acc. to marital status

From the graph we conclude that:

- “1” represent “Married”
- “2” represent “Single”
- “3” represent “Others”
- Most of the defaulters and non-defaulters customers are “Single”

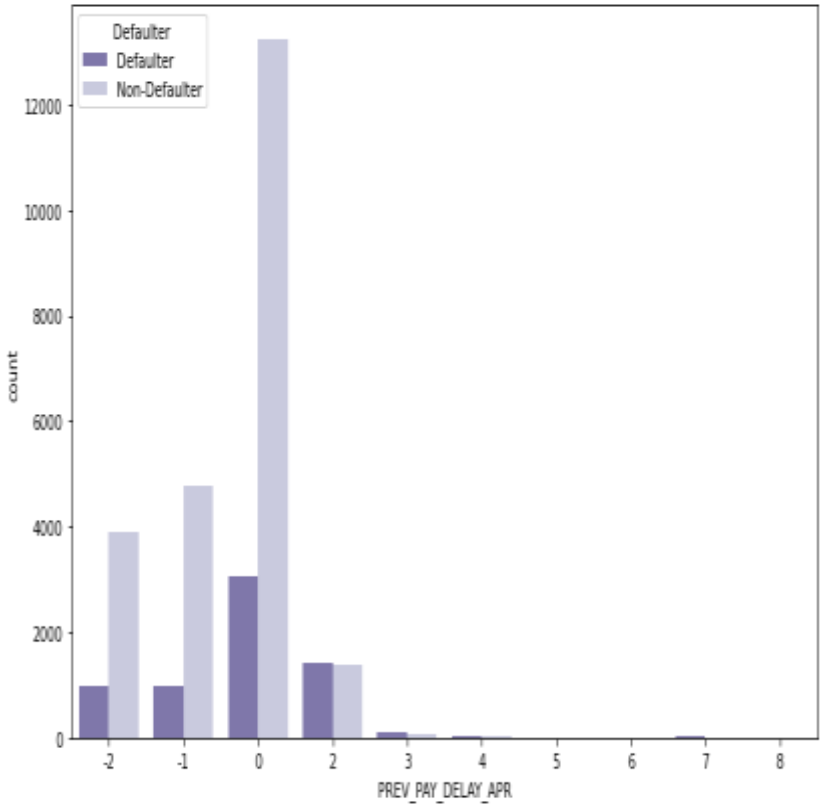


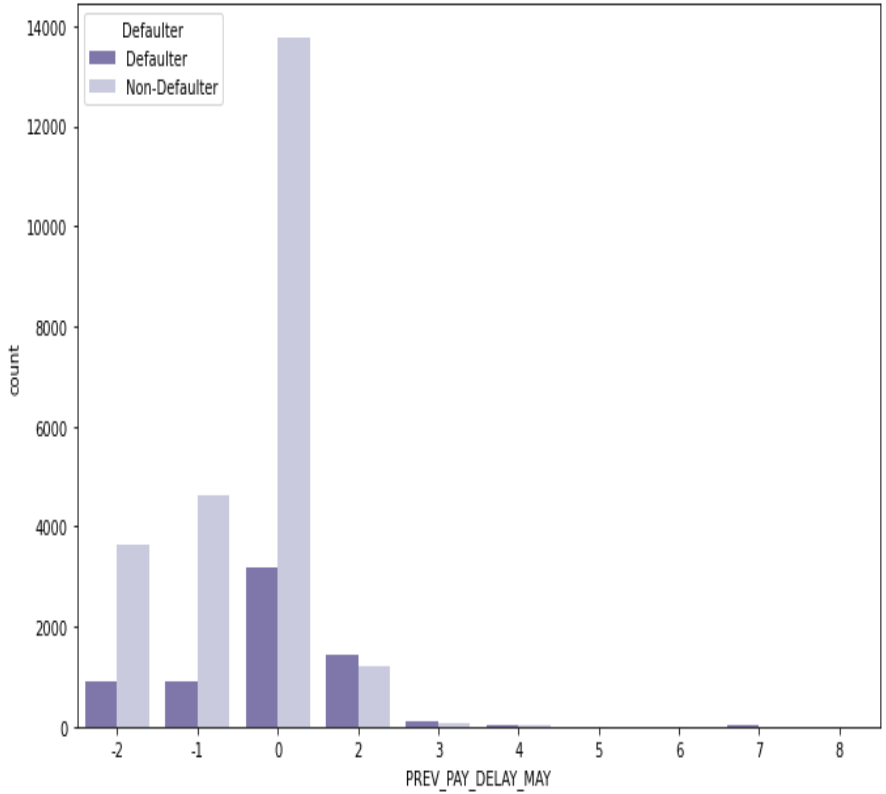
Line Chart To Show The Trend From April 2005 to September 2005



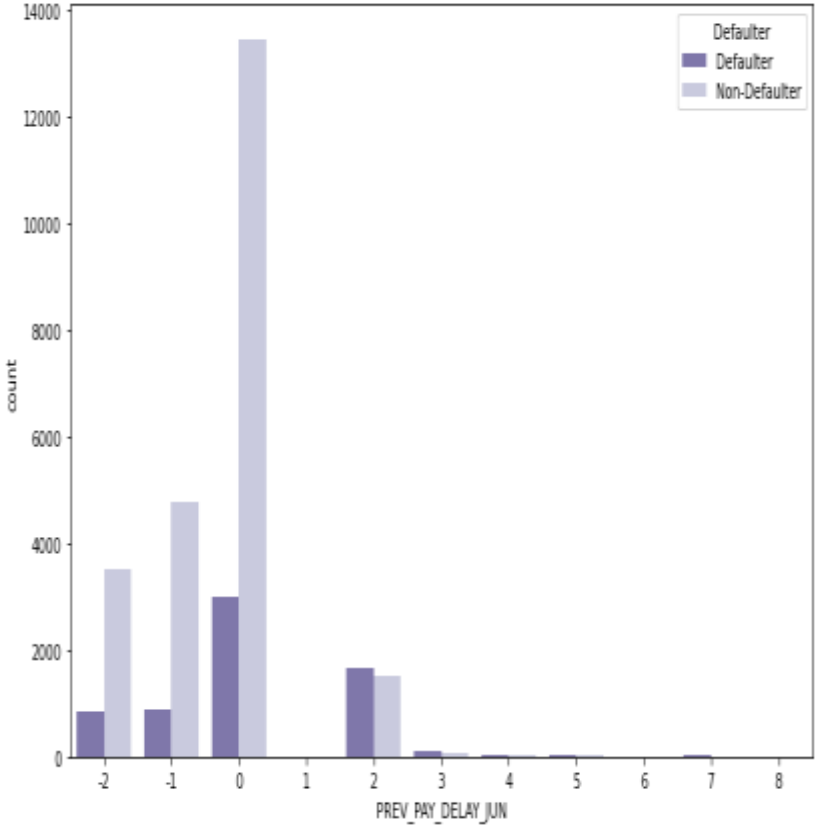
- There was a huge jump from May 2005 (PAY_5) to July 2005 (PAY_3) when delayed payment increased significantly, then it peaked at August 2005 (PAY_2) , After August 2005 (PAY_2) it tends to lost its peakness in September 2005 (PAY_1).

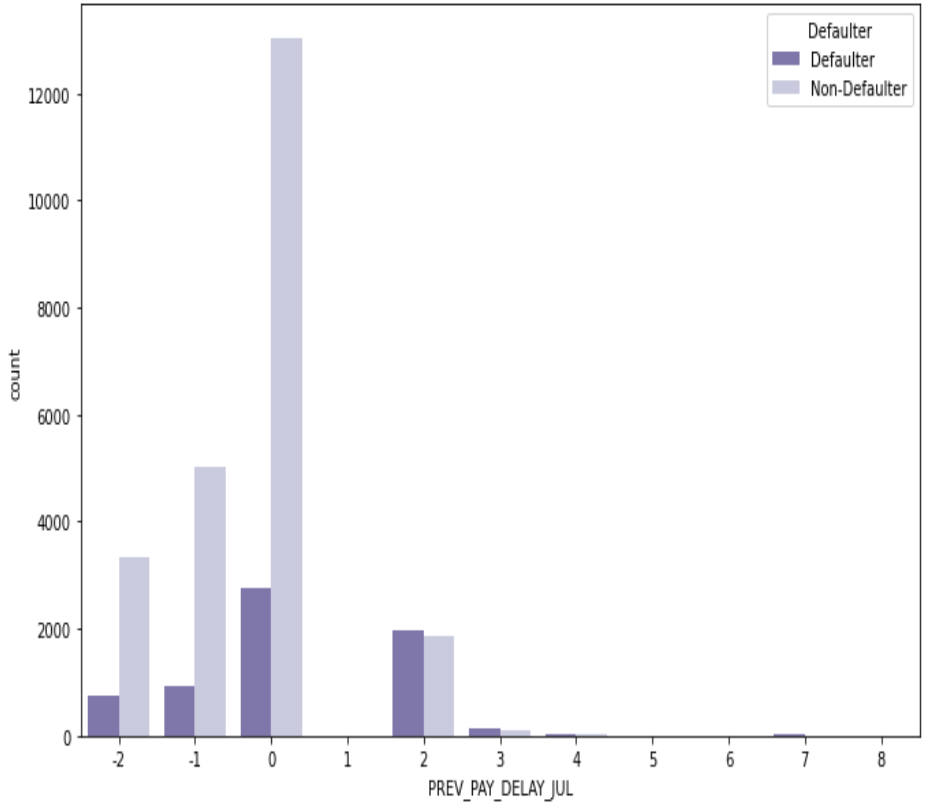
Total no. of months the defaulters have delayed their payment in the previous months



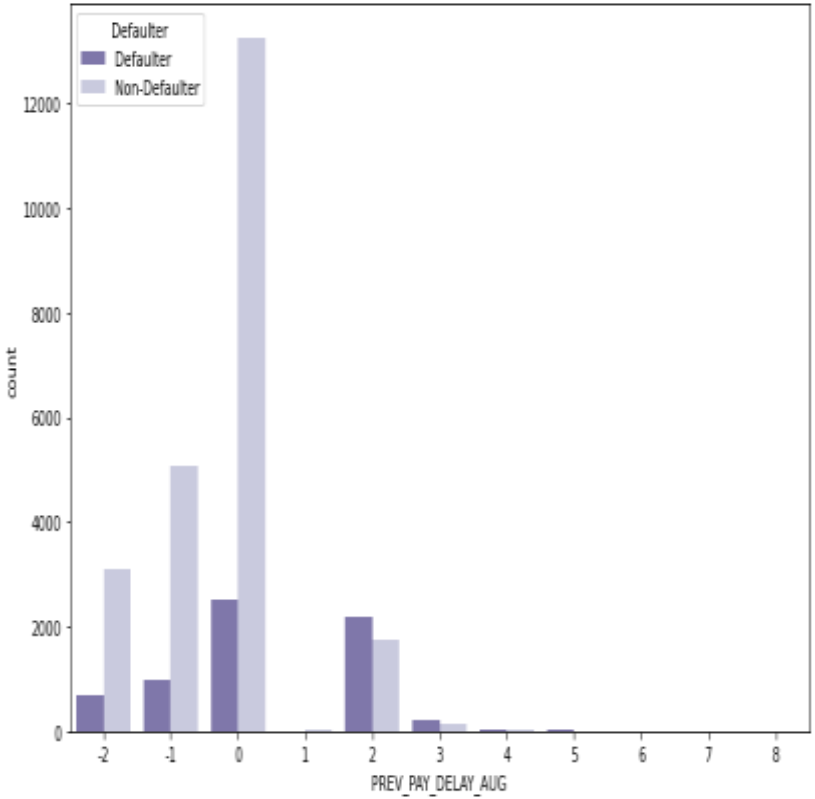


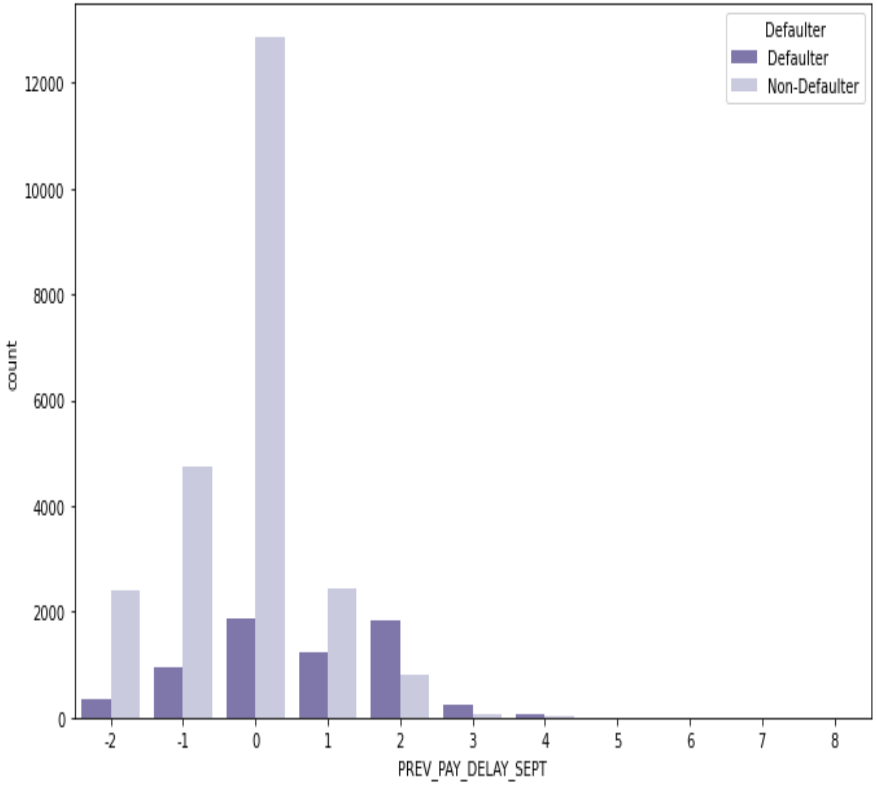
Total no. of months the defaulters have delayed their payment in the previous months



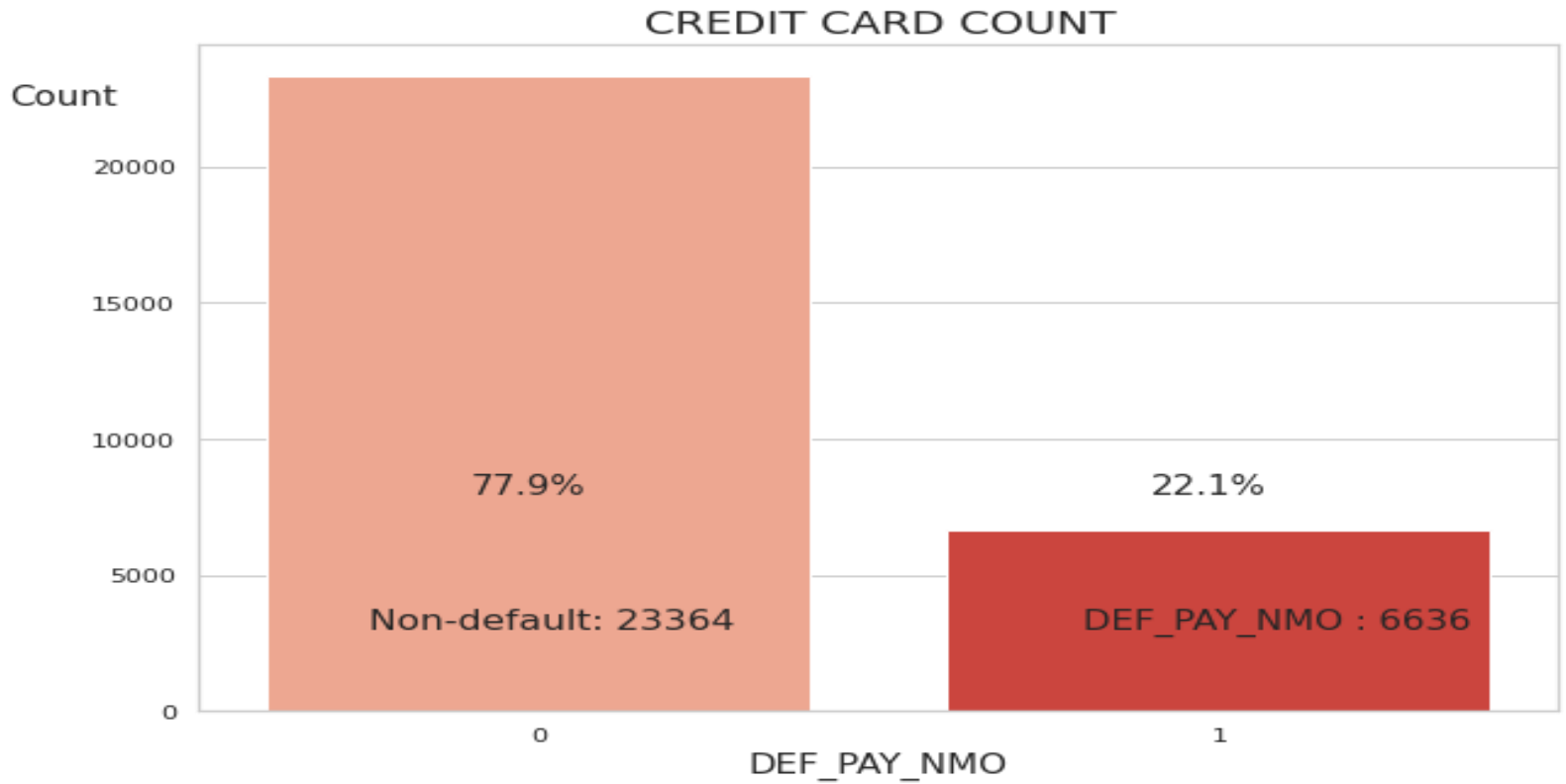


Total no. of months the defaulters have delayed their payment in the previous months



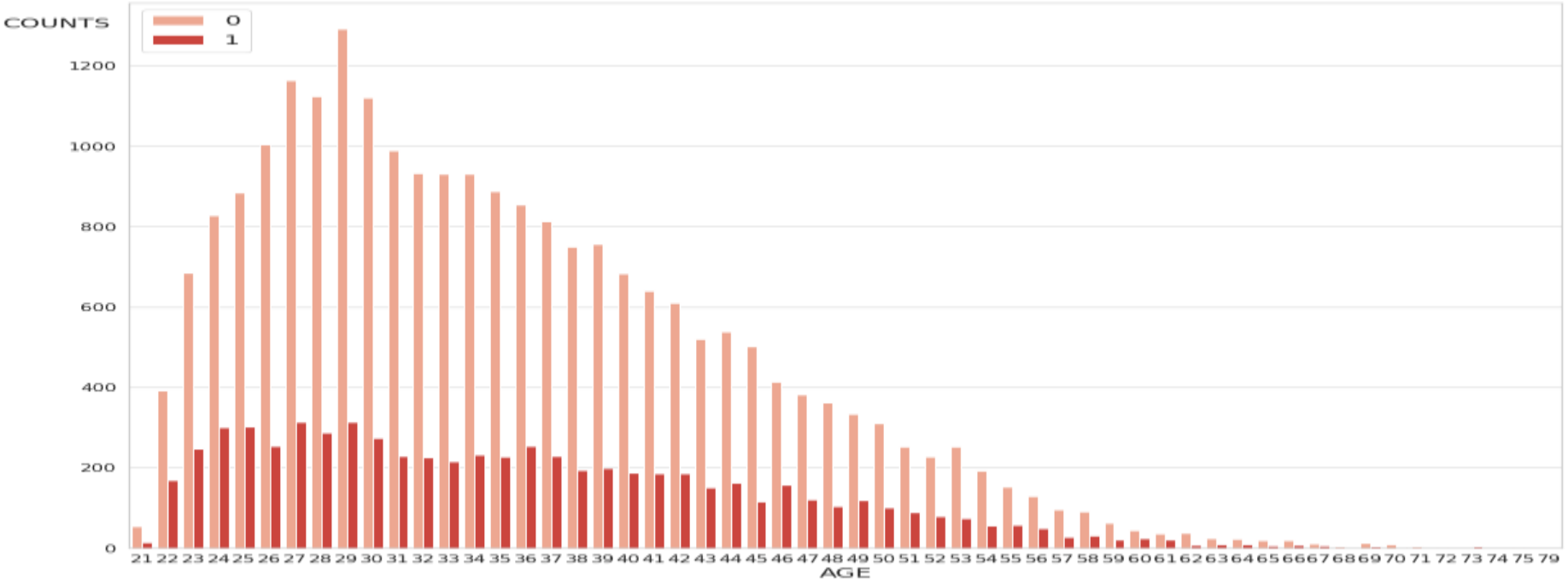


Frequency of Defaulters and Non-Defaulters



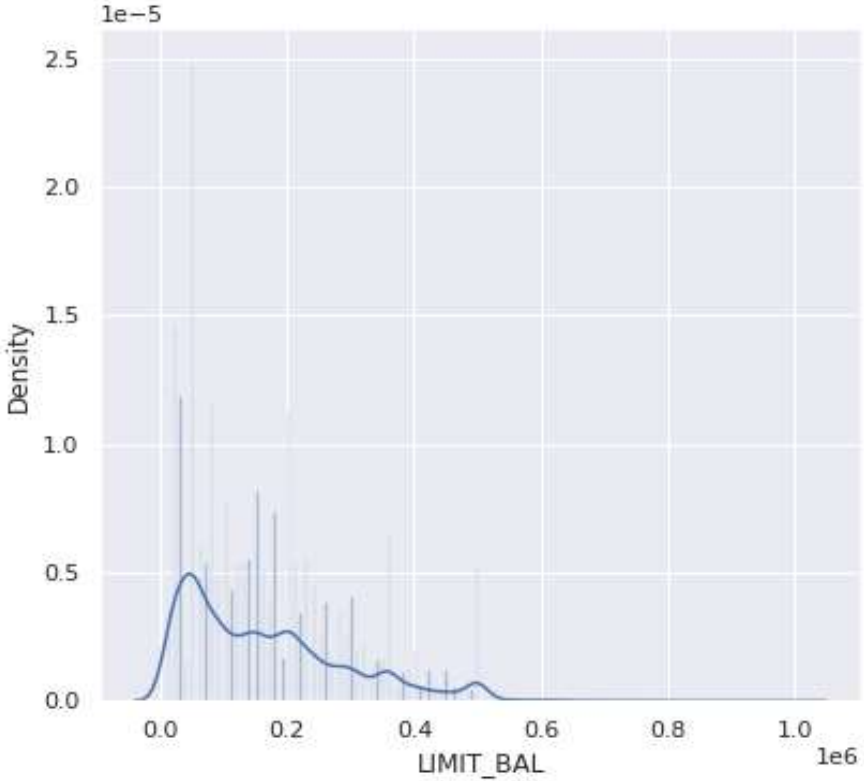
From the above count plot ,we can see that there are more number of Non-Defaulters compared to defaulters in our dataset.

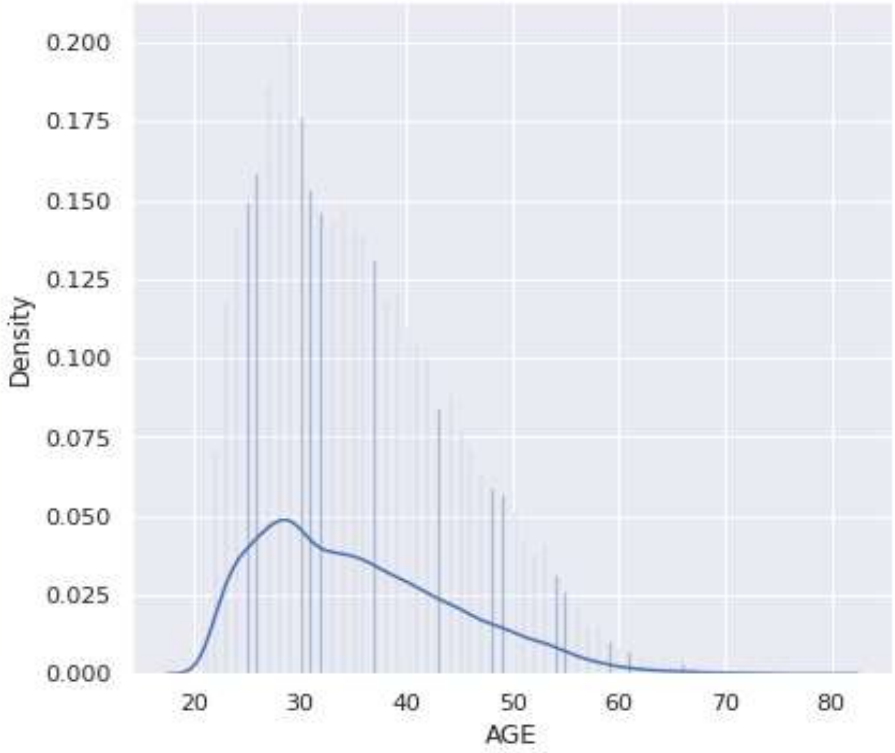
Count plot for The Numerical Feature ‘AGE’



- More no. of credit card holders are age between 26-32 years
- 29 years is the highest age for credit card holder
- Age above 60 years are old customers who rarely uses the credit card.
- Also, more number of Defaulters are between 27-29 years.

Plot a continuous variable





Correlation Between The Variables

We can see that “AGE” and “MARRIAGE” column are negatively correlated with each other

Model Selection

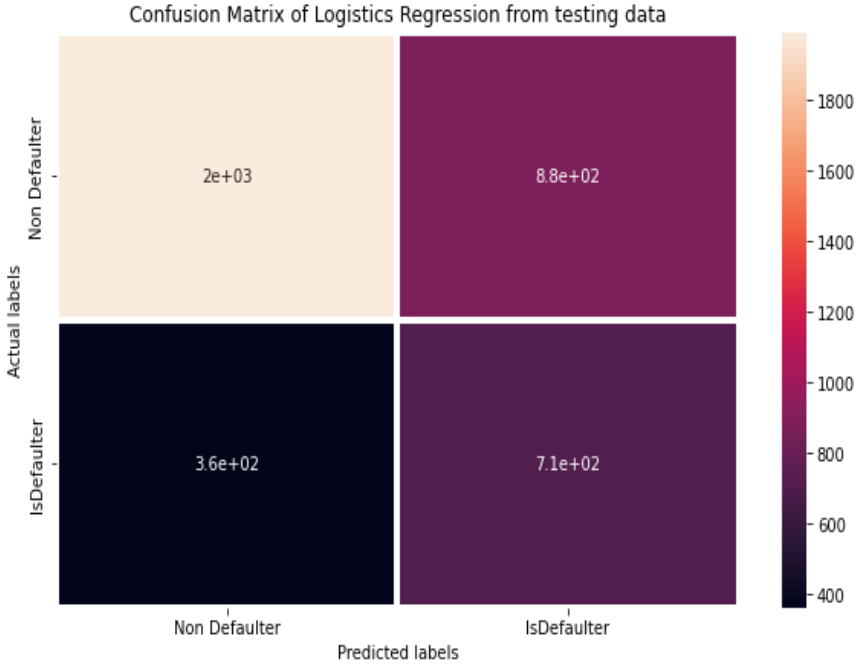
We passed the data into different models like:-

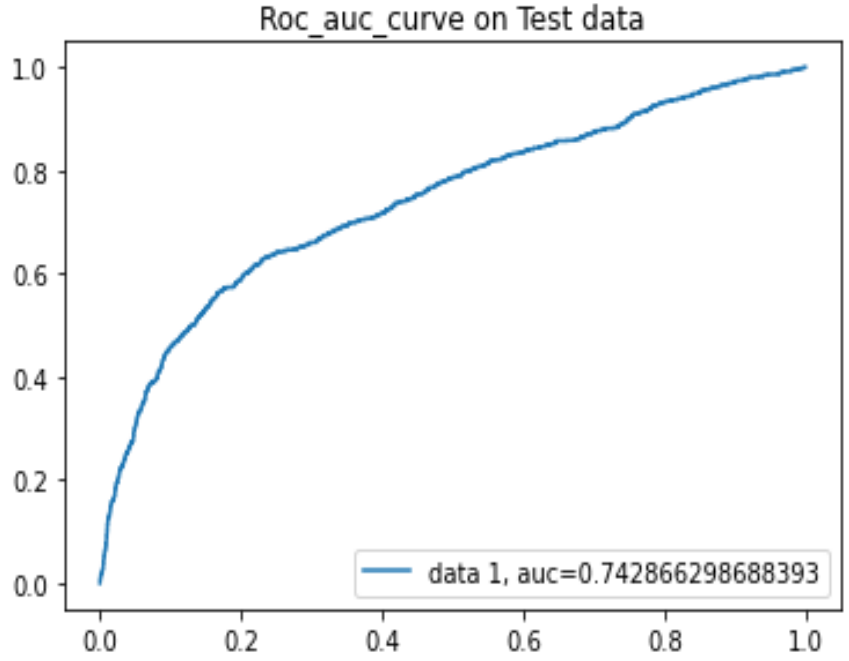
- Logistic Regression
- Random Forest Classifier with CV
- Support Vector Classifier
- K-Neighbor Classifier with CV
- XG Boosting CV
- Decision Tree Classifier

We checked the performance of the each model across various parameters:-

- **Accuracy Score:-** finding the difference between actual and predicted value.
- **Precision:-** is a good metric to use when the costs of false positive(FP) is high.
- **Recall:-** is a good metric to use when the cost associated with false negative(FN) is high.
- **F1-Score:-** is a weighted average of precision and recall.
- **ROC-AUC**
  - **ROC curve (receiver operating characteristic curve):-** is a graph showing the performance of a classification model at all classification thresholds.
  - **AUC stands for "Area under the ROC Curve."** That is, AUC measures the entire two-dimensional area underneath the entire ROC curve

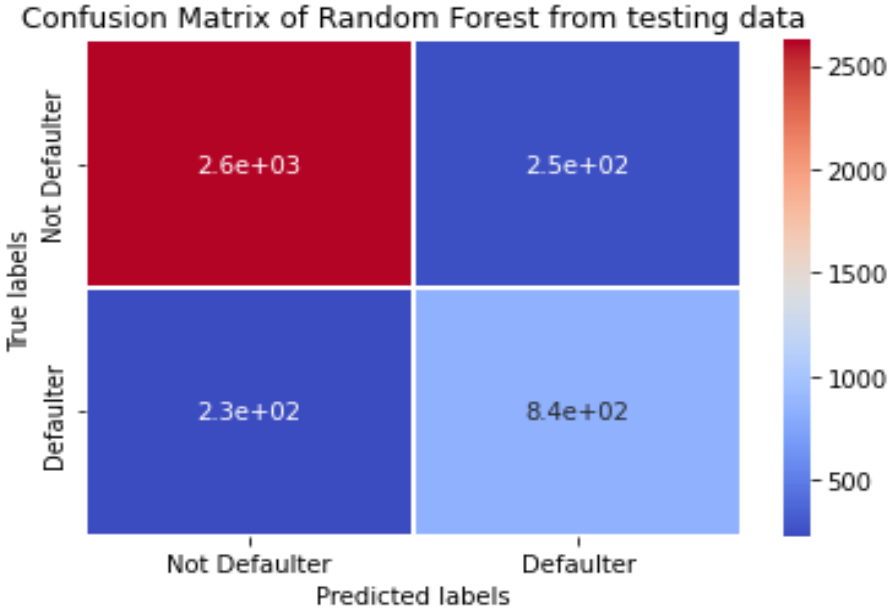
Logistic Regression

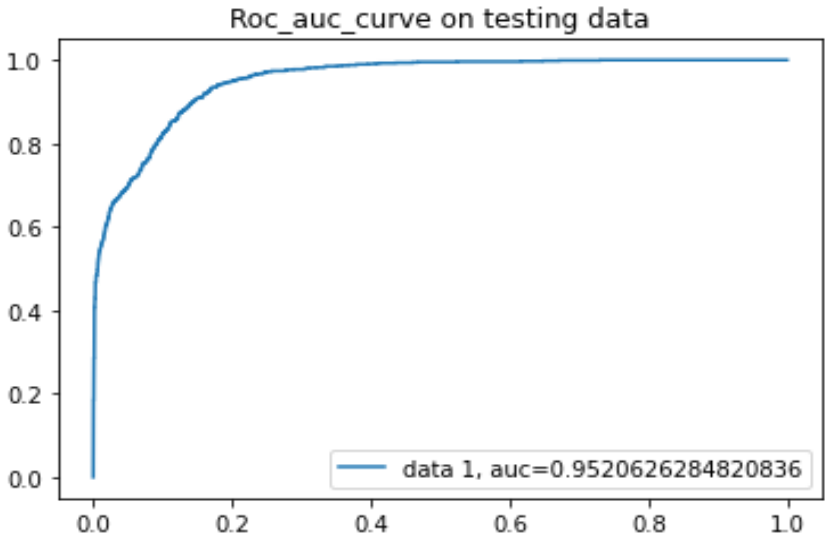




- We have implemented logistic regression and we are getting accuracy score is approx 68%
- Precision score approx is 68% and f1_score is 53%
- roc_auc approx is 67% and recall score is approx 67%
- Here, 1 denote the class for defaulters in AUC curve having value of 0.741

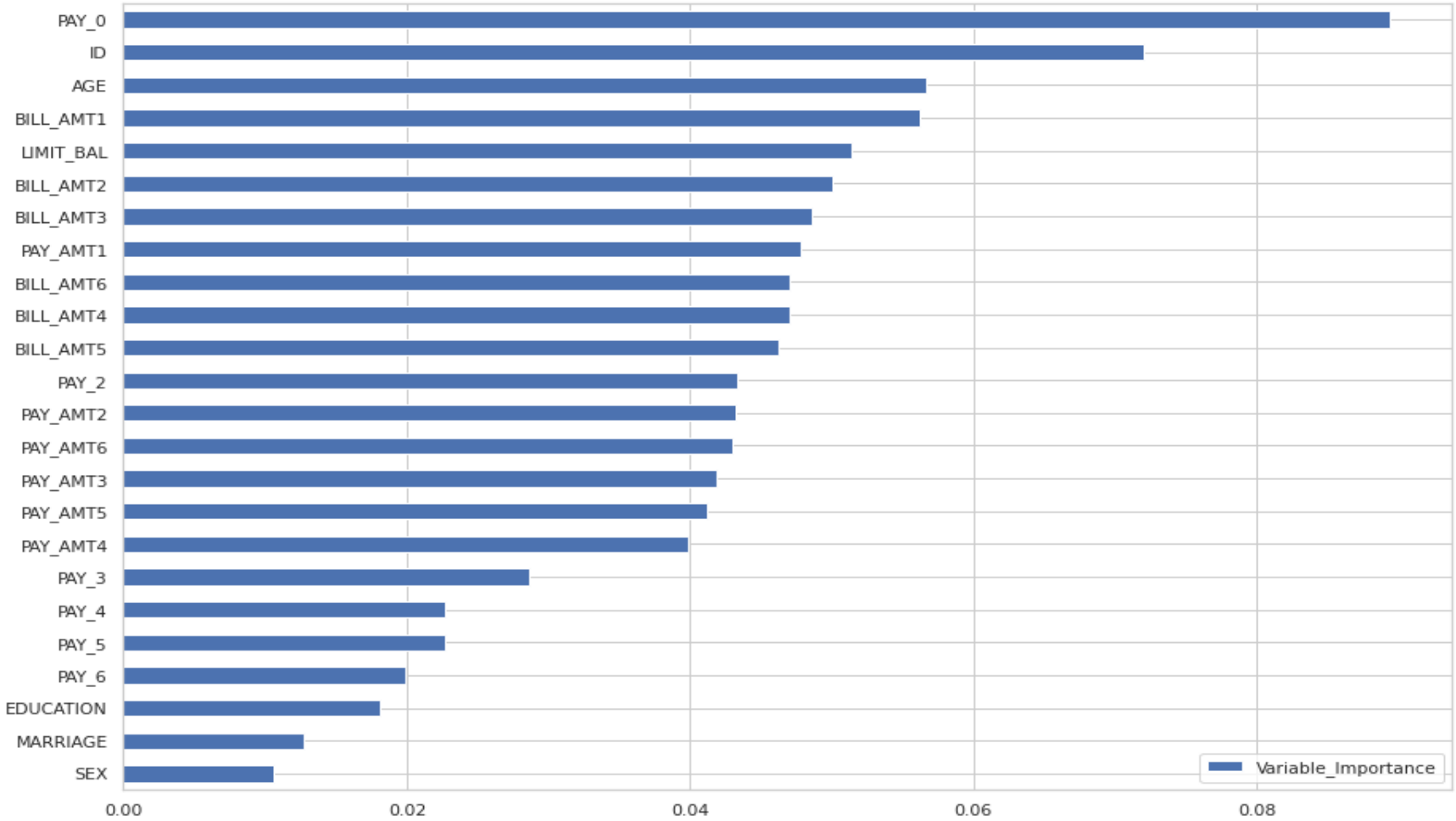
Random Forest Classifier CV



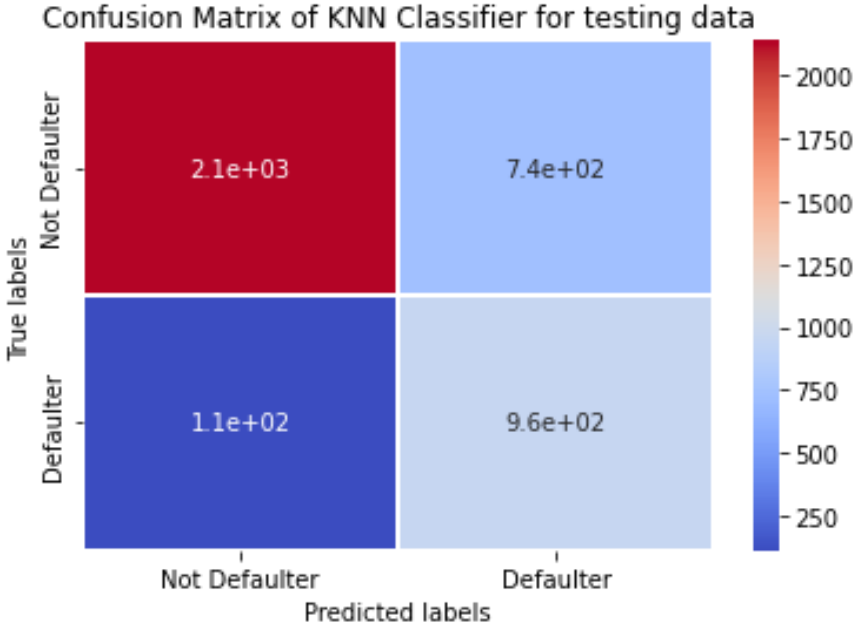


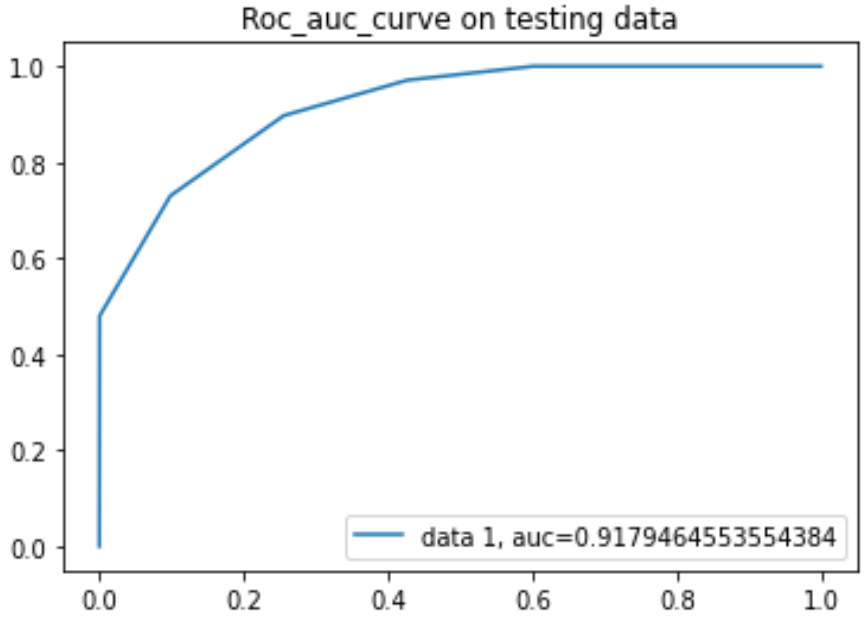
- We have implemented random forest and we are getting accuracy score is approx 88%.
- Precision score approx is 77% and f1_score is 79%
- roc_auc approx is 86% and recall score is approx 81%
- Here, 1 denote the class for defaulters in AUC curve having value of 0.95

Feature Selection Described By Random Forest



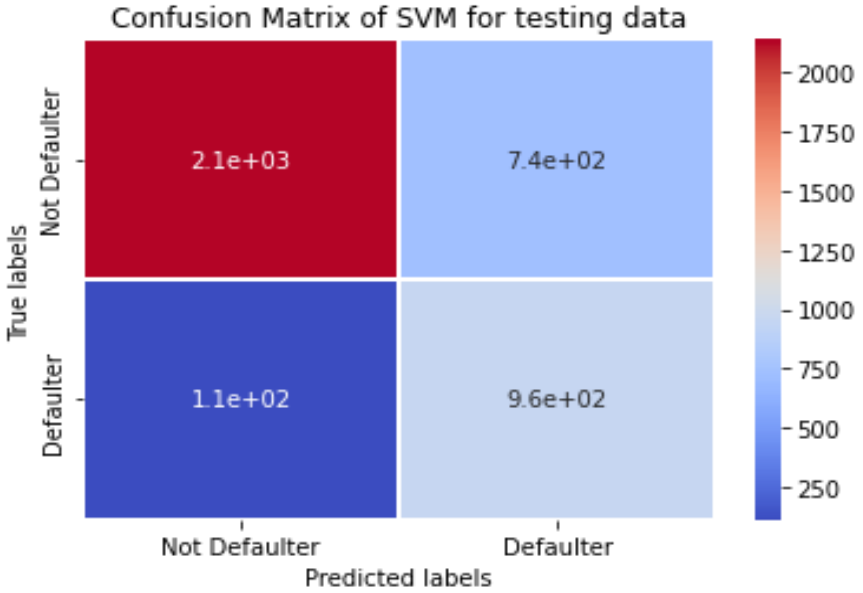
K-Neighbor Classifier CV

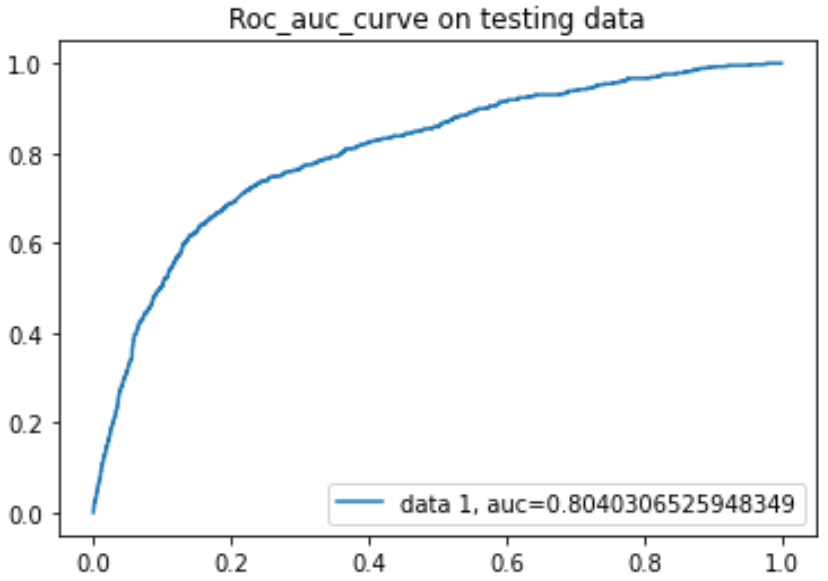




- We have implemented K-Neighbor Classifier and we are getting accuracy score is approx 78%.
- Precision score approx is 57% and f1_score is 79%
- roc_auc approx is 82% and recall score is approx 90%
- Here, 1 denote the class for defaulters in AUC curve having value of 0.917

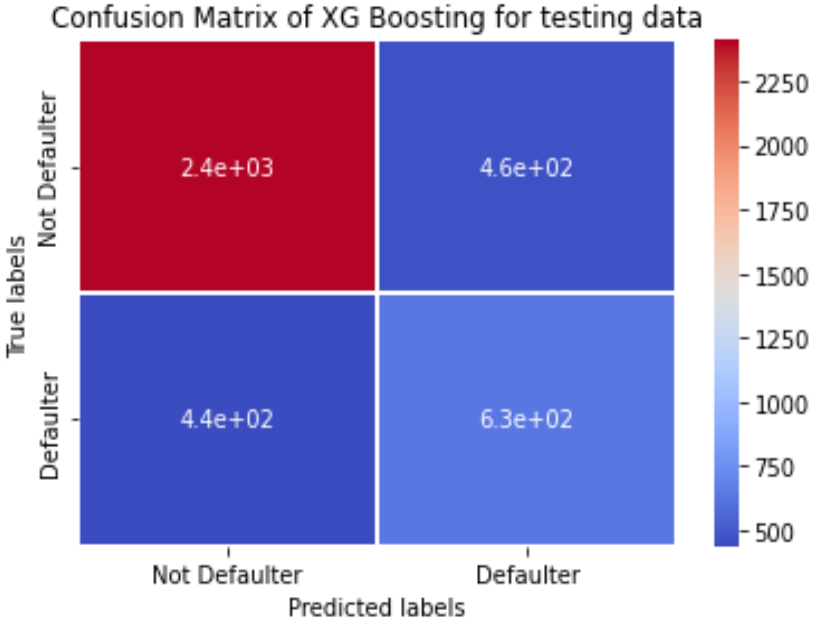
Support Vector Classifier

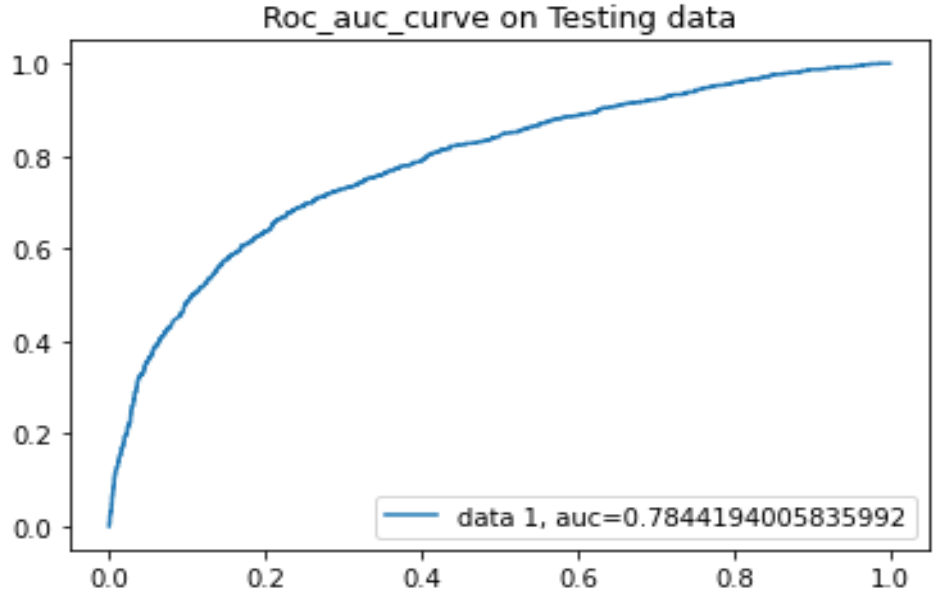




- We have implemented Support Vector Classifier and we are getting accuracy score is approx 77%.
- Precision score approx is 57% and f1_score is 62%
- roc_auc approx is 82% and recall score is approx 68%.
- Here, 1 denote the class for defaulters in AUC curve having value of 0.804

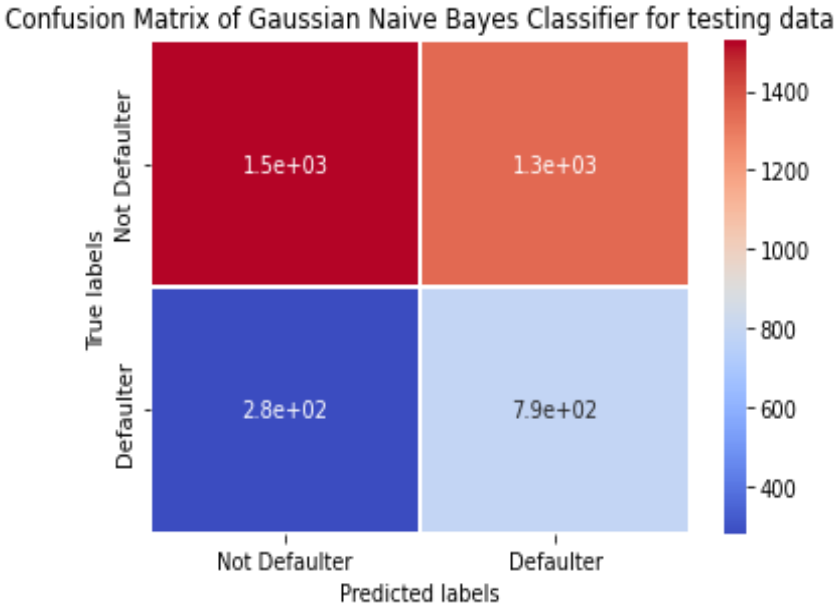
XG Boosting CV

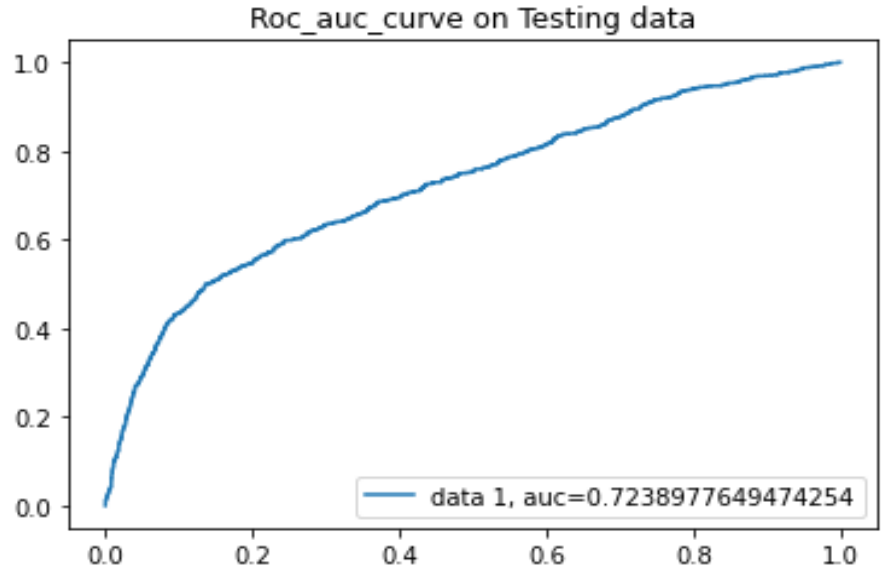




- We have implemented XG Boosting and we are getting accuracy score is approx 76%.
- Precision score approx is 56% and f1_score is 57%
- roc_auc approx is 71% and recall score is approx 59%
- Here, 1 denote the class for defaulters in AUC curve having value of 0.784

Decision Tree Classifier





- We have implemented Decision Tree Classifier and we are getting accuracy score is approx 60%.
- Precision score approx is 37% and f1_score is 49%
- roc_auc approx is 63% and recall score is approx 74%
- Here, 1 denote the class for defaulters in AUC curve having value of 0.723

Evaluation Metrics Table

Model_Name	Accuracy_Score	Precision_Score	Recall_Score	F1_Score	RUC_AUC_Score
Logistics regression	0.6891309855586522	0.4504472681067343	0.6619981325863679	0.5361058601134215	0.6806165906325441
Random Forest	0.8839625031669622	0.7768744354110207	0.8029878618113913	0.7897153351698806	0.8585523453702297
K-Neighbor Classifier	0.7899670635926019	0.5715976331360947	0.9019607843137255	0.6997464686707714	0.8251111292917724
Support Vector Classifier	0.7770458576133773	0.5777054515866559	0.6629318394024276	0.6173913043478262	0.7412364343048299
XG Boosting	0.7727387889536357	0.5820754716981132	0.5760971055088702	0.5790708587517597	0.7110318629074254
Decision Tree Classifier	0.740815809475551	0.5187207488299532	0.6209150326797386	0.5652358691032725	0.7031904787877135

Conclusion

• There is neither null nor duplicate values in our dataset.

• We rename the column for better understanding.

• We check the distribution of defaulter vs. non defaulter and we see that around 78% are non defaulter and 22% are defaulter.

•We have found the proportion of defaulters with respect to Marriage, Education, Sex feature and we found that :

- Most of the defaulters are Female
- Most of the defaulters are from university
- Marital status is Single and single defaulters are more

•We have use the boxplot to detect the outliers and we see that there are so many outliers in the data so we apply IQR(Inter Quartile Range) which is one of the technique to remove outliers.

• We plot heat map to see the correlation between variables and we see that AGE and MARRIAGE columns are negatively correlated.

Conclusion

- **Random Forest Classifier** performs best among all models.
- **Logistic Regression** is not giving good precision score.
- Top 3 models are **Random Forest , K-Neighbor Classifier and Support Vector Classifier** that gives best Precision, Recall ,ROC_AUC and F1 score.

Thank you ....