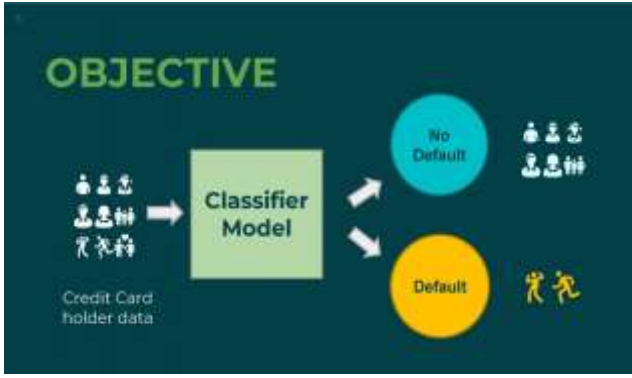
Capstone Project

Credit Card Default Prediction

GOPAL JI GUPTA

Outline

1. Overview & Objective.
2. Data outline.
3. Approach overview
4. Exploratory data analysis
5. Model implementation
6. Model Comparison via evaluation metrics.
7. Conclusion



Overview & Objective

Overview

- Credit card is a commonly used transaction method in modern society and one of the main source of business for banks, as it generates revenue in form of interest but at the same time, it raise the liquidity risk and credit risk to the bank.
- In order to control the cash flow risk, detecting the customers with risk of defaulting is vital.

Objective

This Project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, predicting whether a customer will default on their credit card or not is critically important as we can learn from the past data and use it to be selective in the future.

Data Outline

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

1. **ID**:- Denotes a unique identifier for our clients.
2. **LIMIT_BAL**:- Balance limit on credit card.
3. **SEX**:- Gender of clients 1 for Male And 2 for Female.
4. **EDUCATION**:- Information regarding educational background of clients.
5. **MARRIAGE**:- Marital status of the clients.
6. **PAY_0 –PAY_6** :- Represents the history of past payments.
7. **BILL_AMT1-BILL_AMT6**:- Represents the amount of bill statements for various months.
8. **PAY_AMT1-PAY_AMT6**:- Represents the amount of previous payments like amount paid in September 2005 to April 2005.
9. **default.payment.every.month:-** This variable denotes a client is defaulter or not.

Approach Overview

Understanding and Cleaning

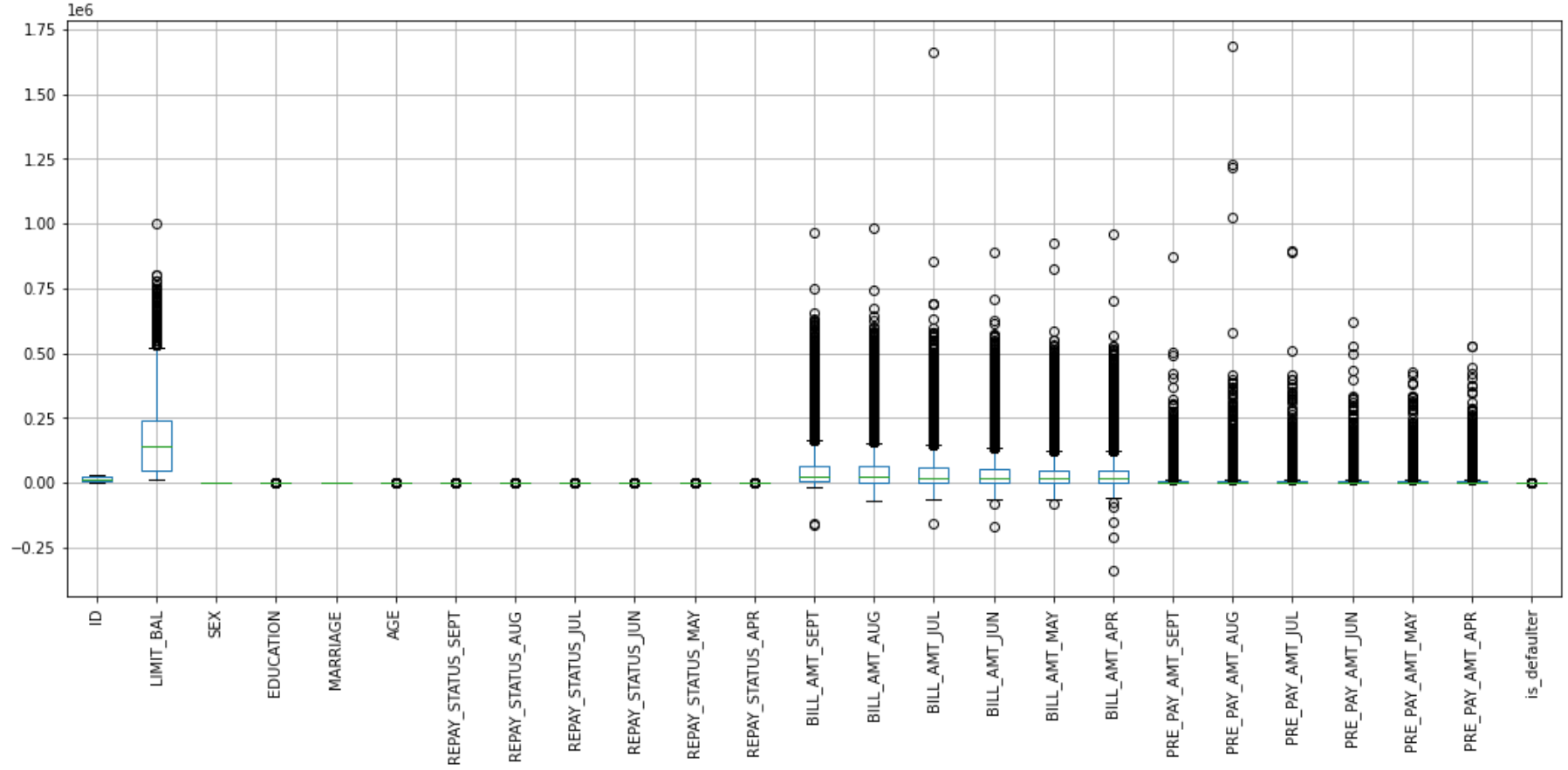
- Find information on documented columns values
- Clean data to get it ready for Analysis

Graphical

- Examining the data with various types of plots.
- Perform some univariate and bivariate analysis.

Machine Learning

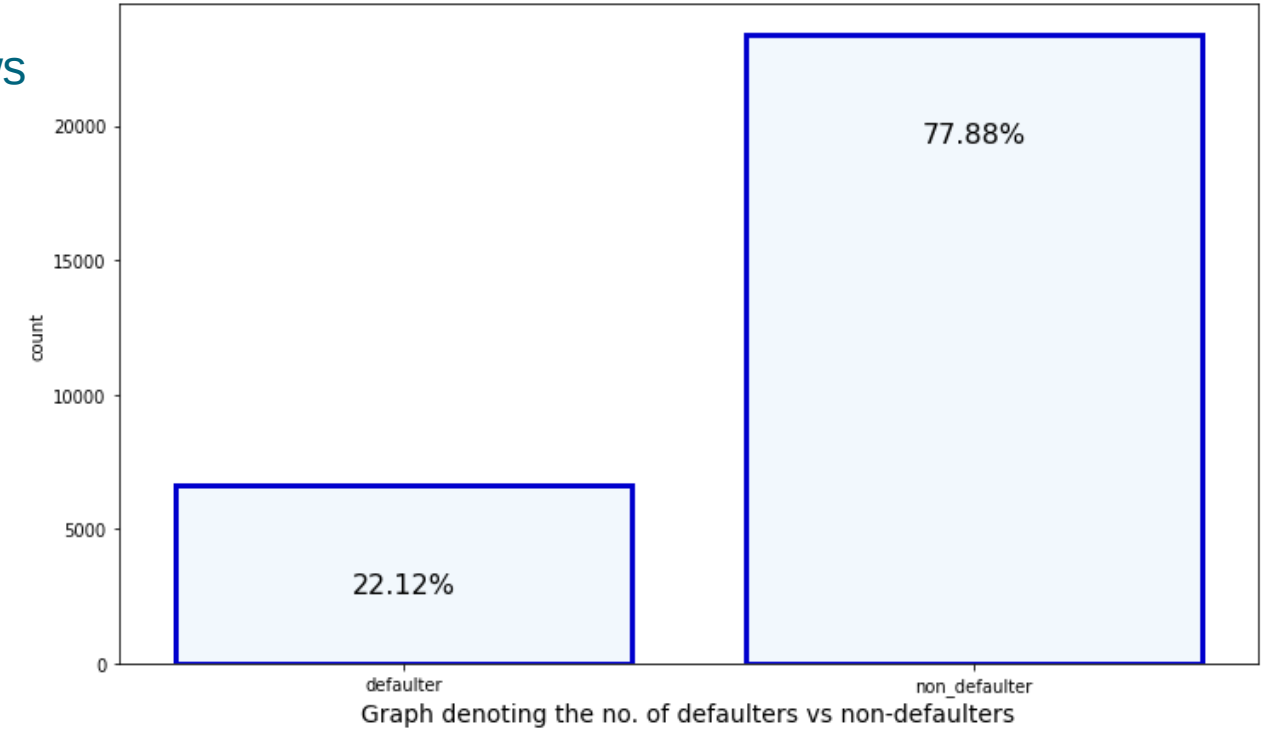
- Logistic
- SVM
- Random Forest
- KNN



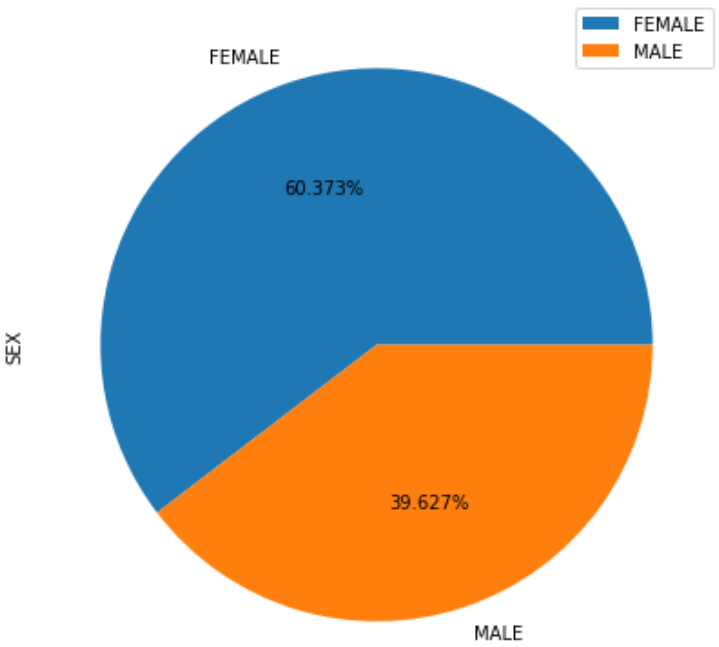
Detecting outliers via a boxplot for all our features.

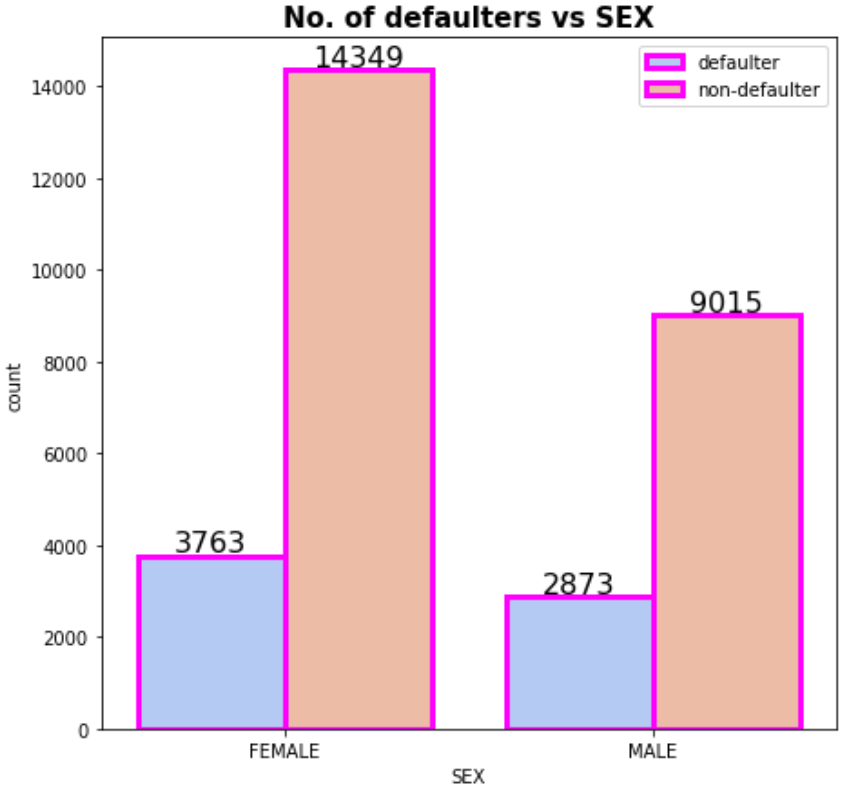
EDA : Visualizing our dependent variable.

- The adjacent graph shows us the distribution of our dependent variable.
- We can clearly see from the adjacent graph that around 22 percent of credit card users are defaulters.

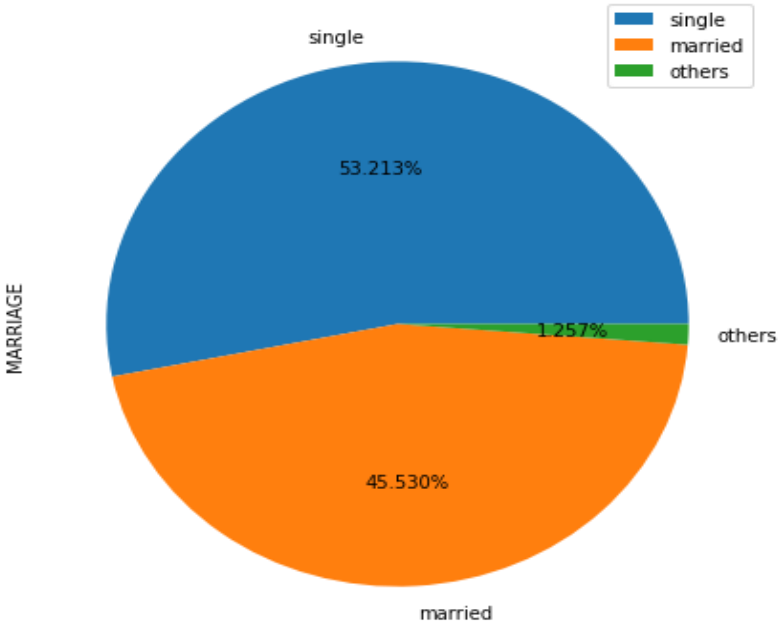


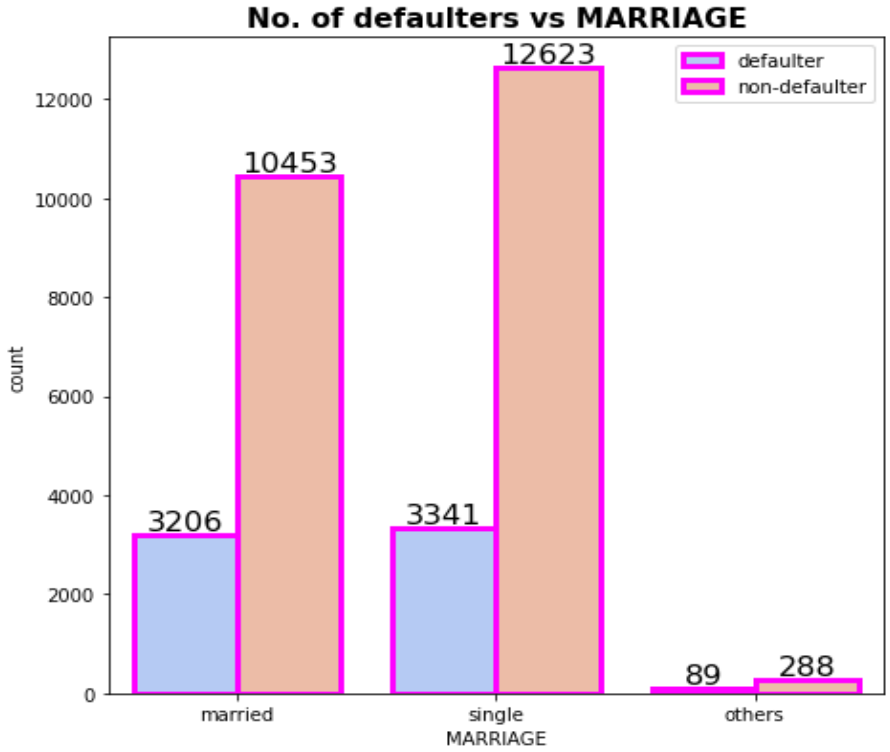
Graph denoting the no. of defaulters vs non-defaulters



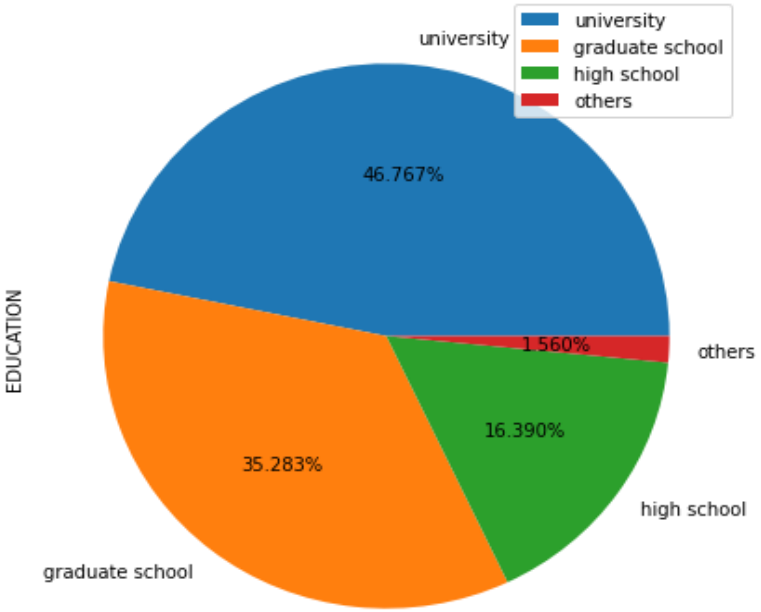


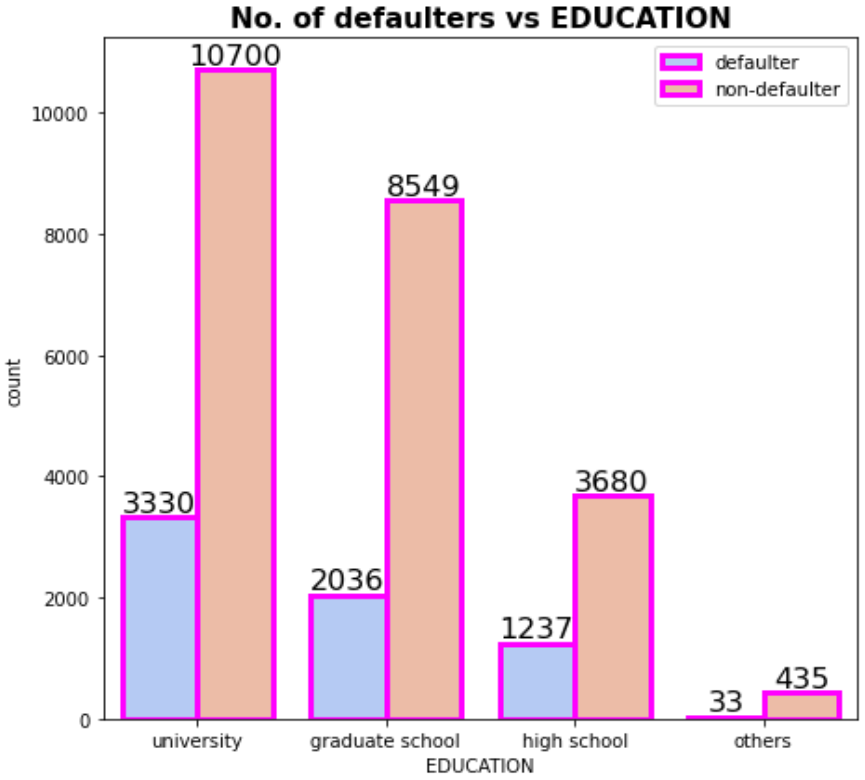
- ✓ We can clearly see from the graph above that most of the credit card holders are female and most of our defaulters are female as well. This is normal as most of purchases are made by women.
- ✓ In fact around 60 percent of our users are female.



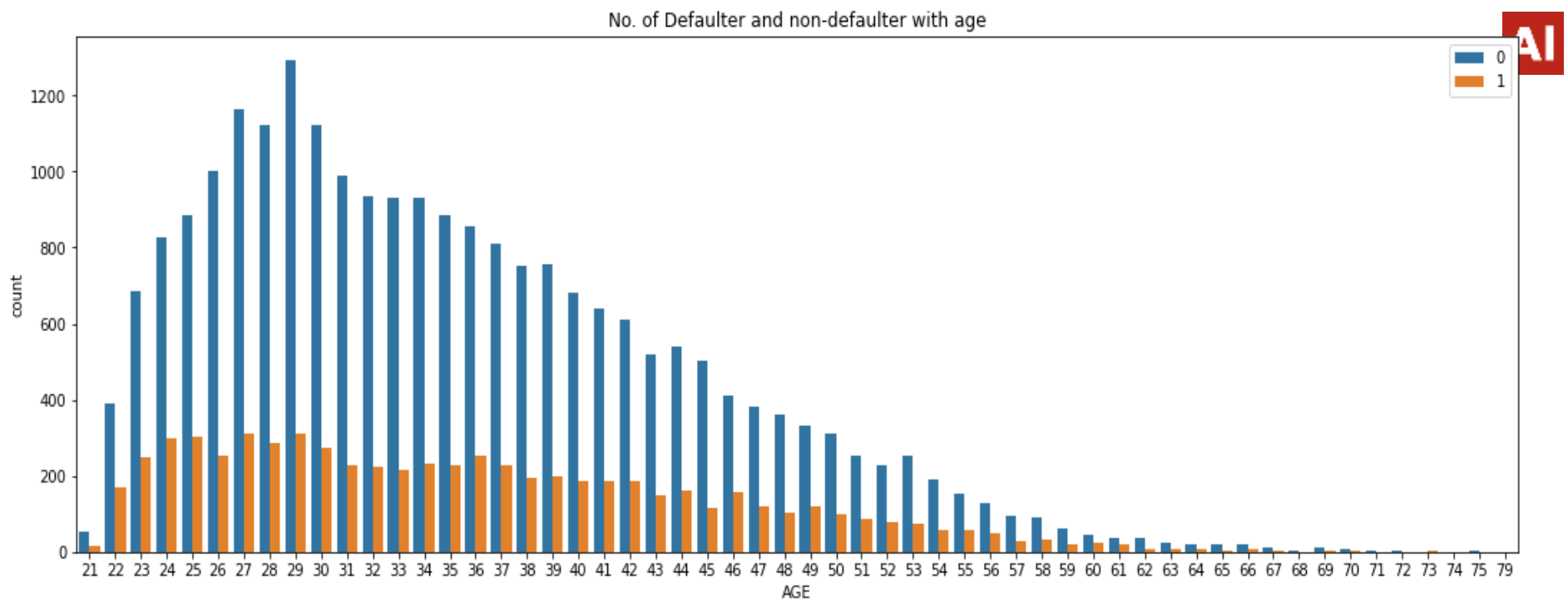


- It can be clearly seen from the above graphs that there are more single credit card holders than married people.
- Around 53 percent of our credit card holders are single.





- We can clearly see from the above graphs that most credit card holders are highly educated people around 46 percent of them are university educated and around 35 percent are graduate school educated.



- We can clearly see from the above graph that as age increases, the no. of credit cards are low i.e. we don't see many elderly people that have credit cards.
- Also we can see that most no. of credit cards are held by people in 23 to 35 age bracket.

Correlation Heatmap

- The correlation matrix helps us visualize the correlation between our numerical variables. We can see in the heatmap that there is some correlation in our bill amount features.

Model Implementation

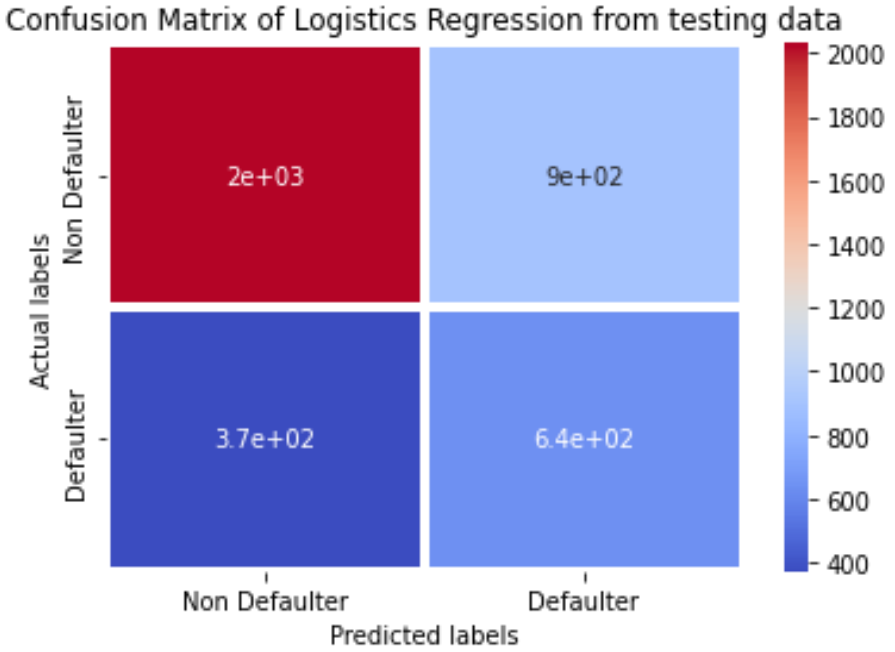
Based on the linear relationship between the dependent and independent variables present in our data, we implemented following models on our data.

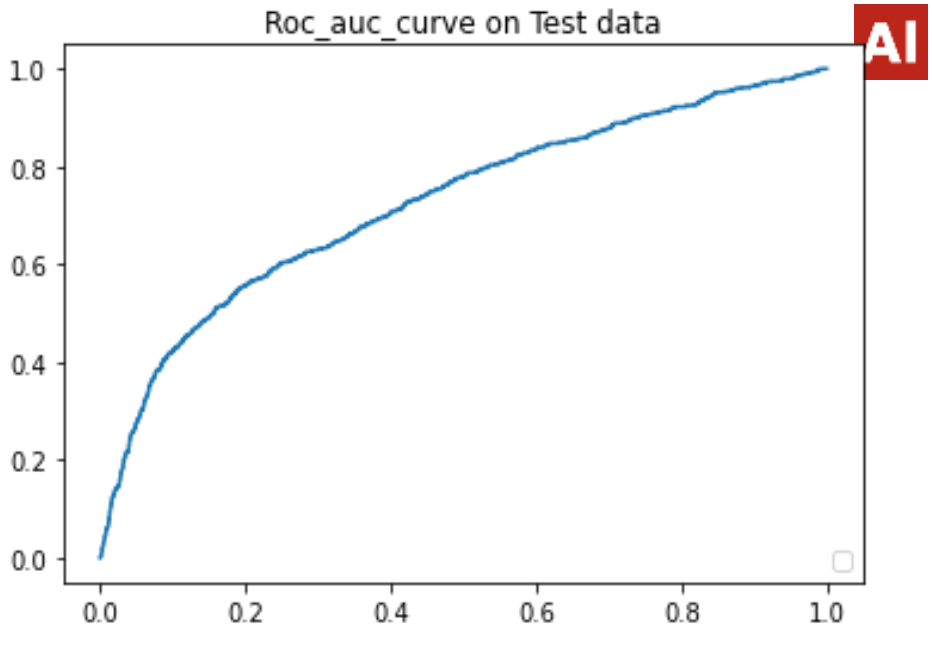
- Logistic Regression
- Random Forest Classification
- K Nearest Neighbor Classification
- Support Vector Classifier

We fit these models on training data, learn the model parameters and then make predictions on test dataset. Then we check the performance of these models using various evaluation metrics such as :-

- Accuracy, Precision and Recall
- F1 score and roc-auc score
- Confusion Matrix

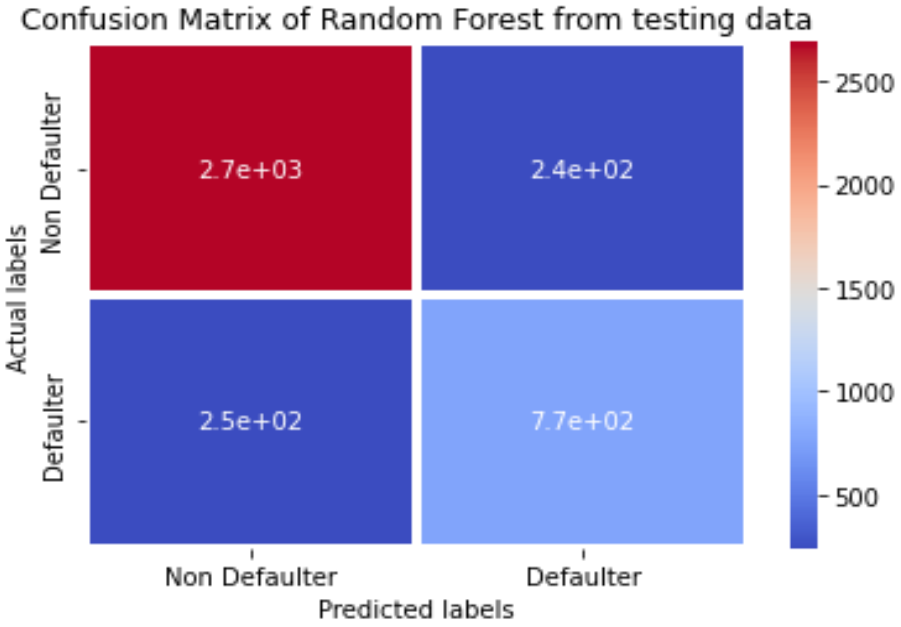
Finally, we select the best performing model based on these metrics.

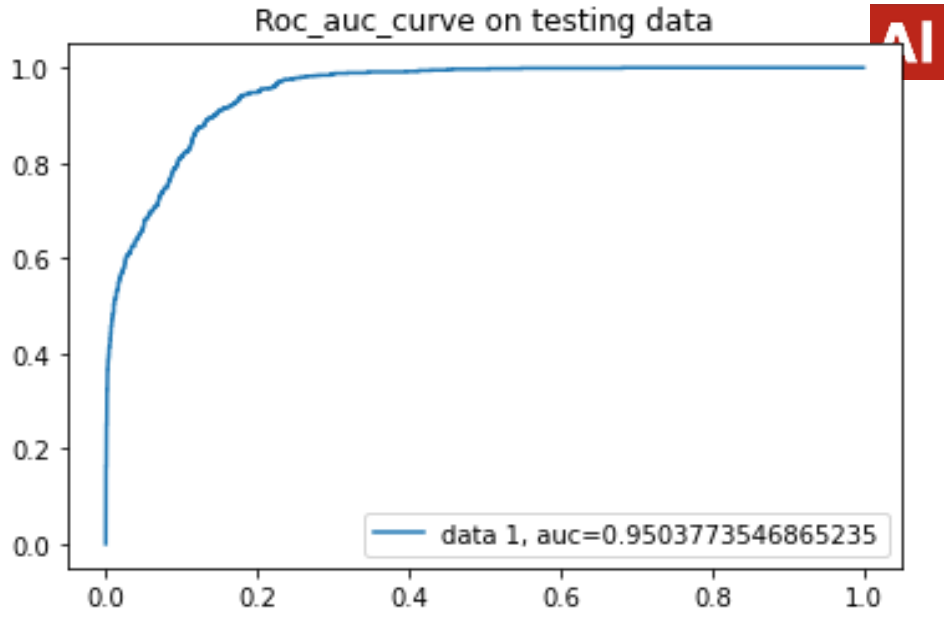




Logistic Regression Implementation

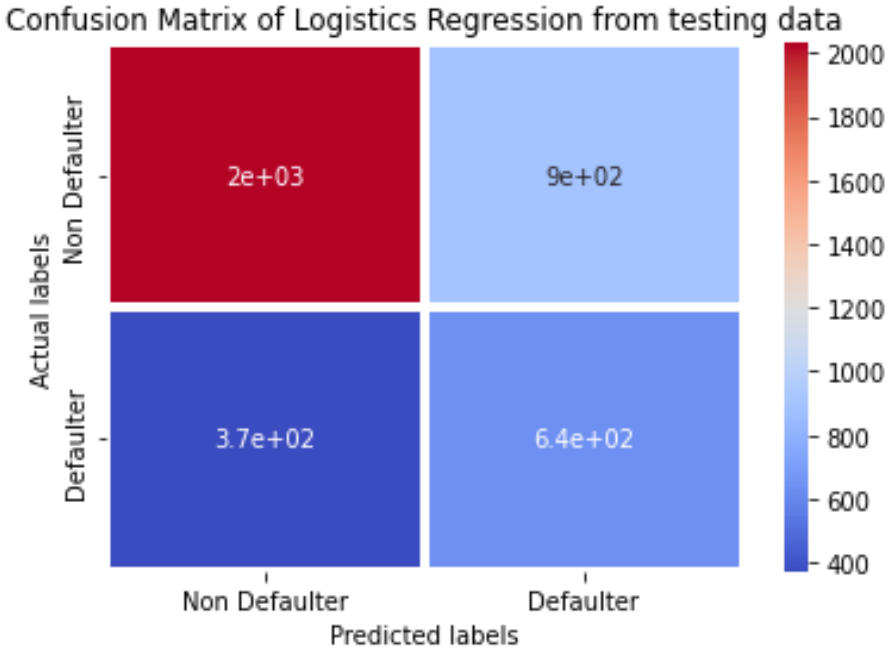
- We can see the plotted Roc-Auc curve of logistic regression.
- The logistic regression makes predictions with accuracy of 0.677, precision 0.416, recall of 0.633 and roc auc score of 0.663 & The F1 score is 0.502

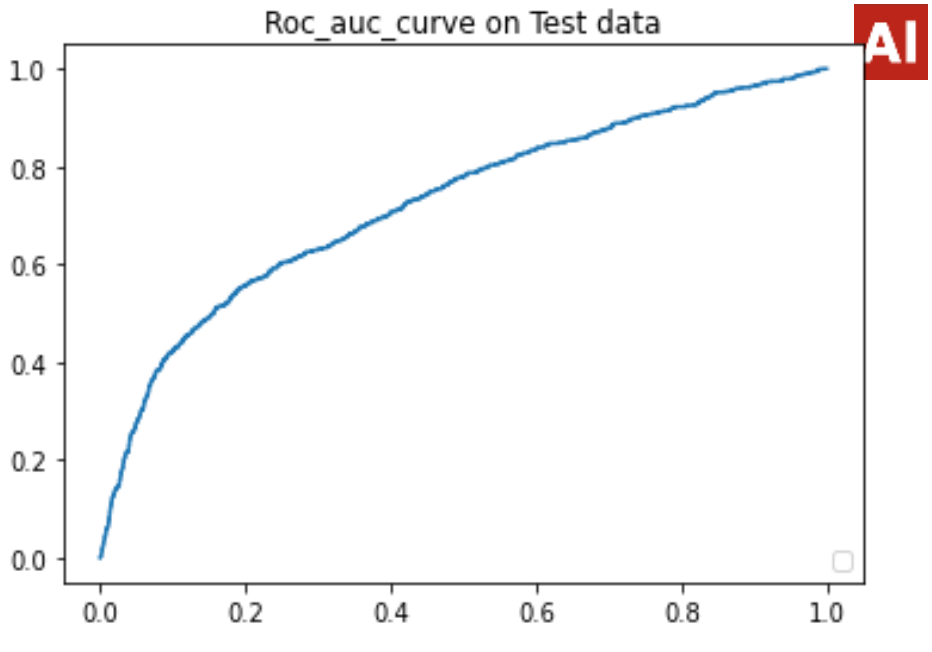




Random Forest Implementation

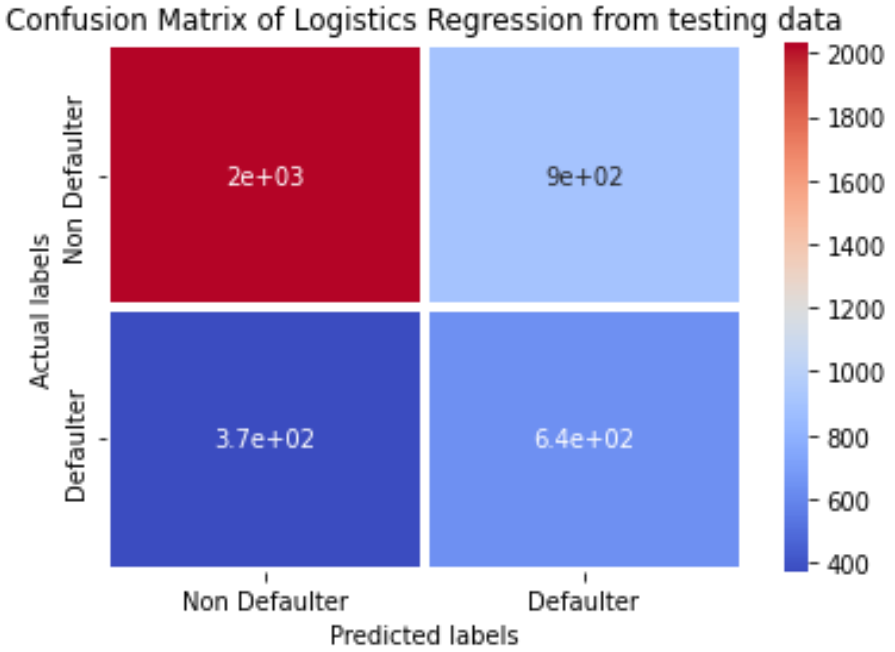
- We can see the plotted Roc-Auc curve of Random Forest classification.
- The random forest model makes predictions with accuracy of 0.876, precision 0.760, recall of 0.757 and roc auc score of 0.837 & The F1 score is 0.759

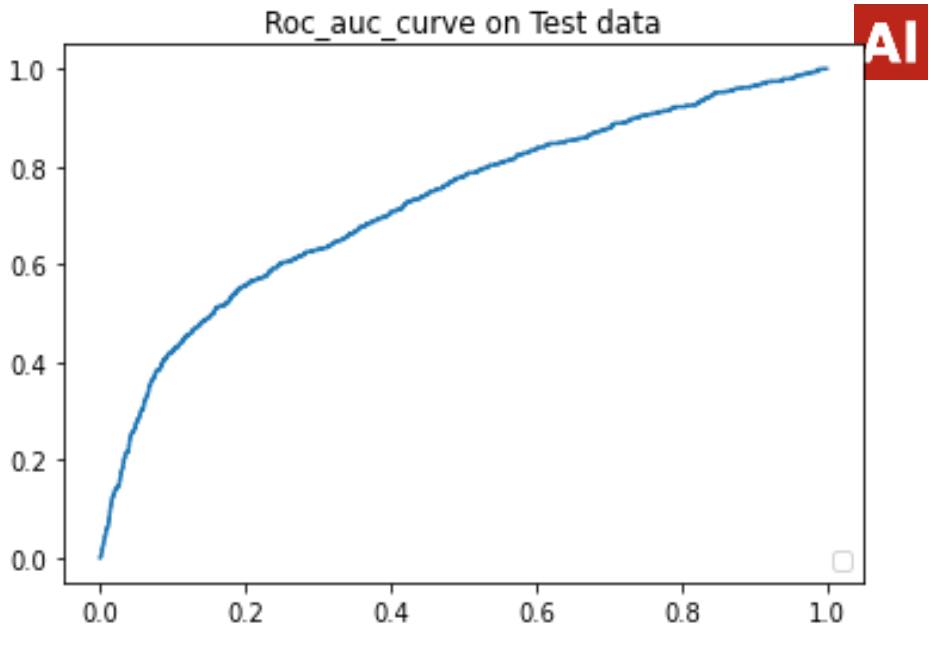




K Neighbors Classifiers Implementation

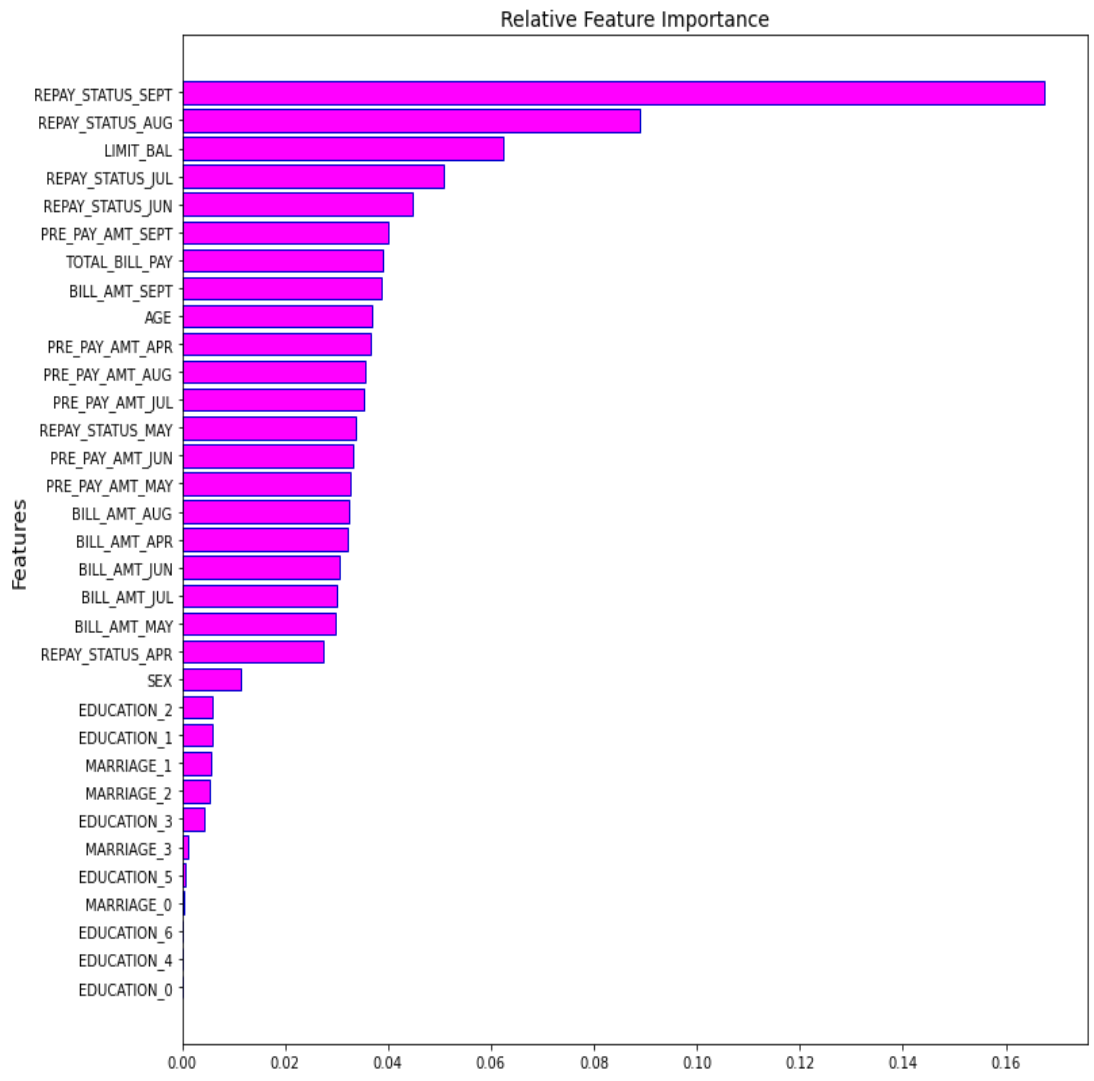
- We can see the plotted Roc-Auc curve of K neighbors classifier.
- The K neighbors classifier makes predictions with accuracy of 0.859, precision 0.680, recall of 0.855 and roc auc score of 0.858 & The F1 score is 0.758





Support Vector Classifier Implementation

- We can see the plotted Roc-Auc curve of support vector classifier.
- The support vector classifier makes predictions with accuracy of 0.766, precision 0.539, recall of 0.635 and roc auc score of 0.723 & The F1 score is 0.583.



- In the adjacent graph, we can clearly see the feature importances for making predictions.
- We can clearly see that repayment status for Sep and Aug are most important features.
- The features are ranked from top to bottom in descending order of their importance.
- This feature importance feature provides us with some insight into why the result of a prediction falls into one label class or the other.

Evaluation Metrics:

	Logistic Regression	Random Forest Classifier	K Neighbours Classifier	Support Vector Classifier
ACCURACY	0.677477	0.876362	0.85964	0.766658
PRECISION	0.416451	0.760633	0.680784	0.539298
RECALL	0.633498	0.757635	0.855172	0.635468
F1 SCORE	0.50254	0.759131	0.758079	0.583446
ROC-AUC SCORE	0.663099	0.837549	0.85818	0.723771

Conclusions Drawn

- We implemented several models on our dataset in order to be able to predict the dependent variable and found that **K nearest classifier and Random Forest Classifier are the best performing models** as they score well on all evaluation metrics**.**
- We saw that logistic regression is the least well performing model.
- Recall for K nearest classifier is around 85% so this means that this model will correctly predict defaulters around 85% of the time.
- Most of the credit card users are Female and thus have higher number of defaults than male. This can be understood as on average, women make more purchases than men.
- Most of the credit card users are highly educated.
- The number of credit card users goes down with increase in age as elderly people have less consumption and may not be able to use credit cards or their purchases may by done by family members.
- With our models making predictions with such high accuracy even on unseen test data, we can confidently deploy this model for further predictive tasks using future real data.

THANK YOU