Capstone Project - 3

Credit Card Default Prediction

Supervised ML Classification Model

Vineeta Singh

Understanding the concept

What is credit card default?

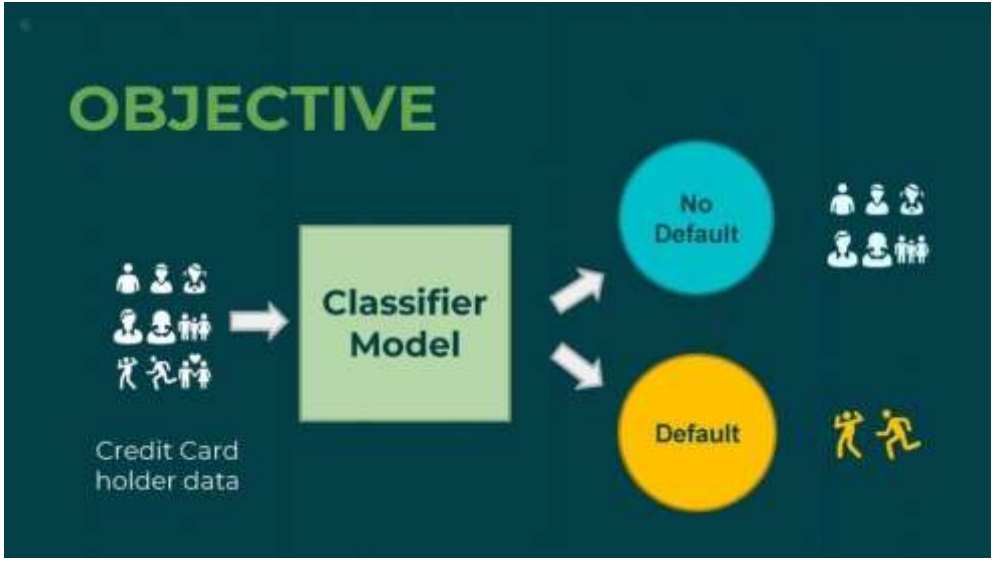
Credit card default happens when you have become severely delinquent on your credit card payments. Default is a serious credit card status that affects not only your standing with that credit card issuer but also your credit standing in general and your ability to get approved for other credit-based services.

Why Do we need to predict Credit card default beforehand ?

The **financial institution** can be capable of preventing the loss. Here, we have used various machine learning classification techniques to carry out Default related analysis.

Content

- Introduction
- Defining Problem Statement
- Data Summary
- Approach Overview
- EDA / Feature analysis
- Modelling Overview
- Feature Importance
- Model evaluation
- Challenges
- Conclusion
- Q&A



Introduction

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

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.

Defining Problem Statement

- Identify the key drivers that determine the likelihood of credit card default.

- Predict the likelihood of credit card default for customers of the Bank.

Data Summary

- X1 - Amount of credit(includes individual as well as family credit)
- X2 - Gender
- X3 - Education
- X4 - Marital Status
- X5 – Age
- X6 to X11 - History of past payments from April to September
- X12 to X17 - Amount of bill statement from April to September
- X18 to X23 - Amount of previous payment from April to September
- Y - Default payment

Approach Overview

Data Cleaning → **Data Exploration** → **Modeling**

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

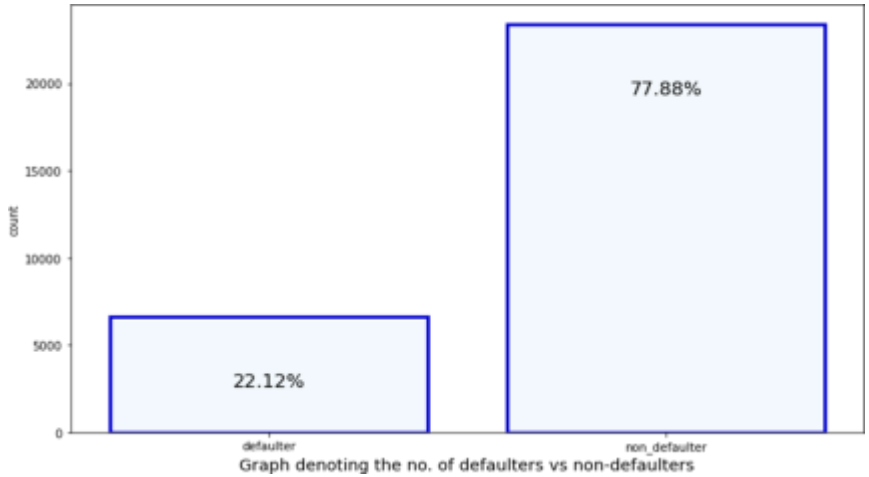
Basic Data Exploration

- Taiwan from April 2005 to September 2005.
- Dataset contains 30000 rows & 25 columns.
- In dataset, 6 months payment and bill data available.
- There are no null or duplicate values.

```
[ ] #reading the data set
    cred_df.head()
```

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5
0	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0
1	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455
2	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948
3	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959
4	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146

Feature Analysis - The frequency of defaults



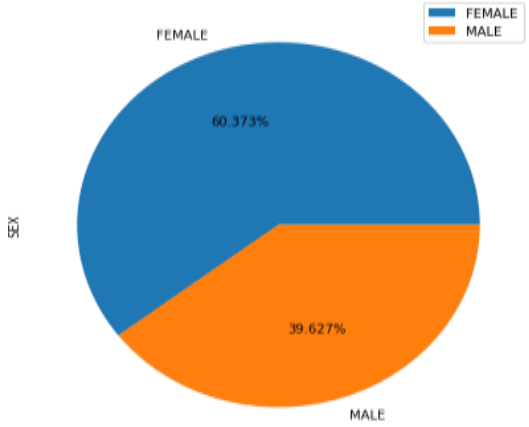
Graph denoting the no. of defaulters vs non-defaulters

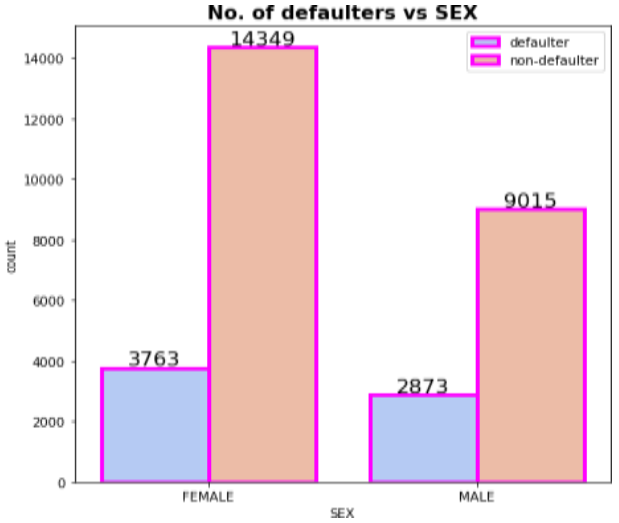
Looking at data, we get this idea that it is a case of Imbalanced dataset.

Count of Non-default is a lot higher than default value.

Non- Default data is 77.9% while Default cases are 22.1% as in dataset.

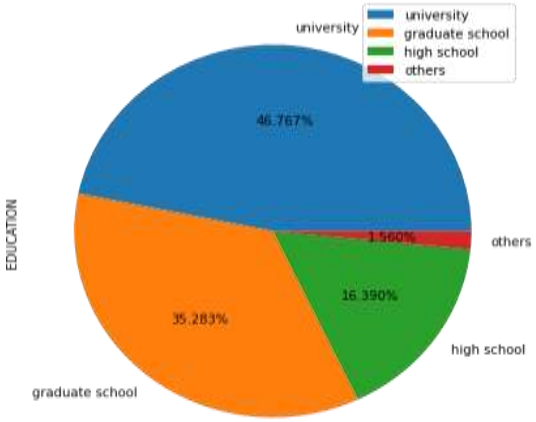
Feature Analysis - Gender wise defaulter prediction

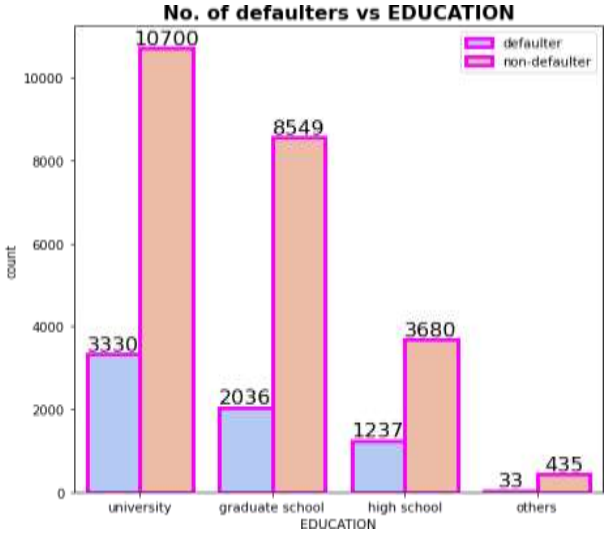




- ✓ We can clearly see from the graph above that most of the credit card holders are female and most of our defaulters are female as well. This is normal as most of purchases are made by women.
- ✓ In fact around 60 percent of our users are female.

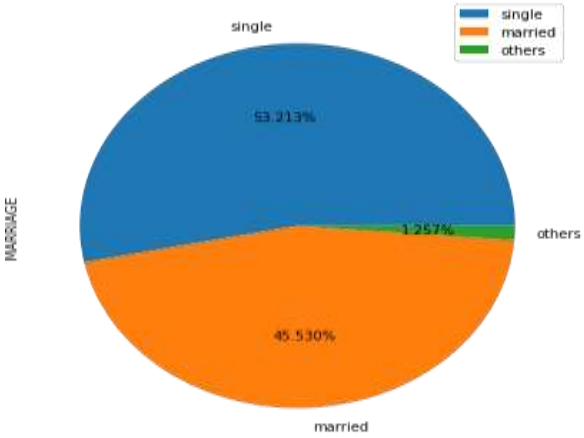
Feature Analysis - Education wise defaulter prediction

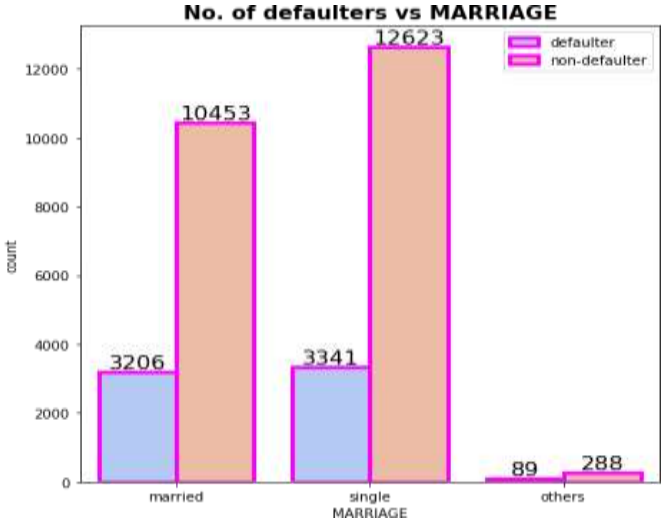




We can clearly see from the above graphs that most credit card holders are highly educated people around 46 percent of them are university educated and around 35 percent are graduate school educated.

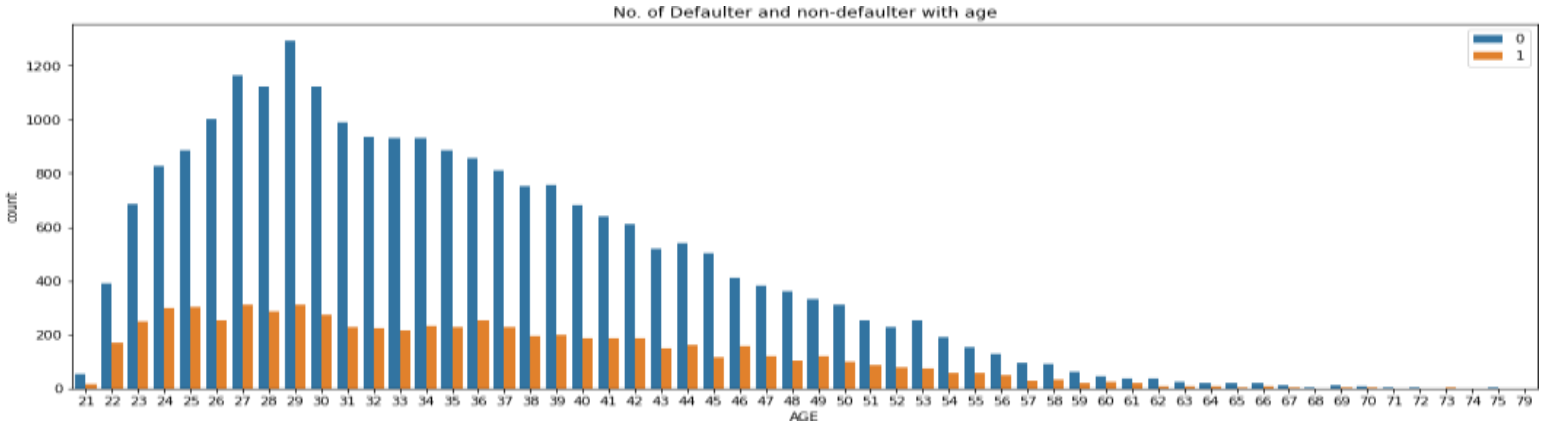
Feature Analysis – Marital status wise defaulter prediction





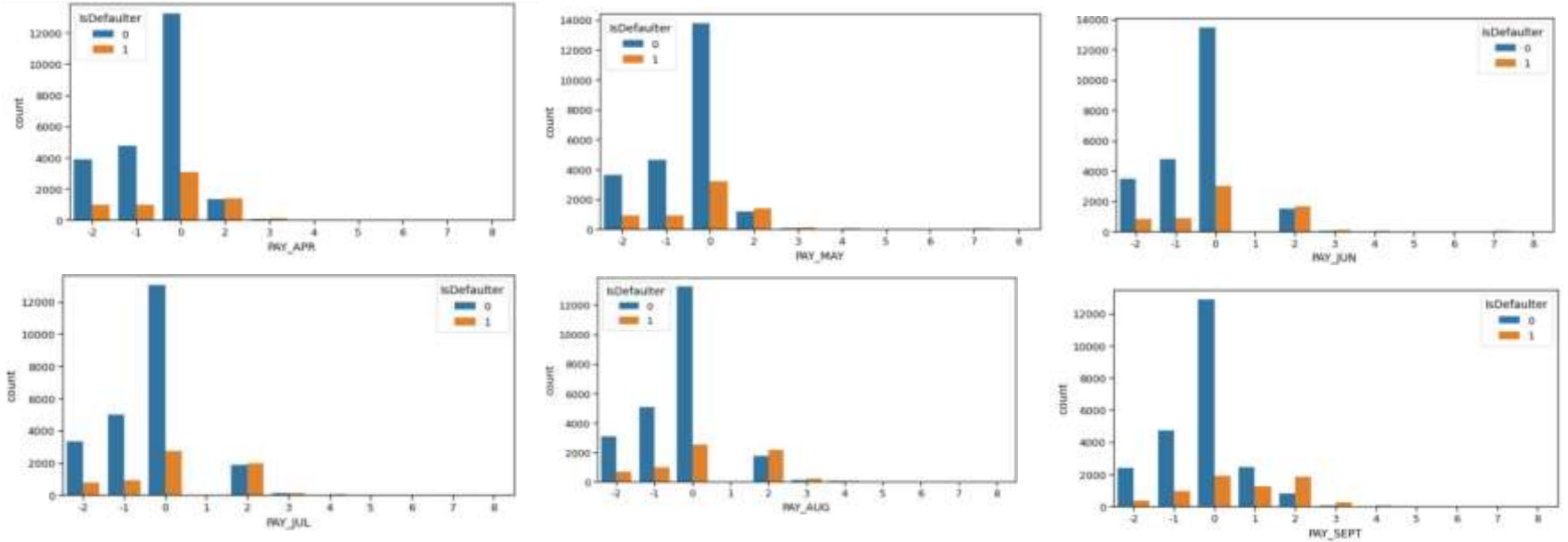
- It can be clearly seen from the above graphs that there are more single credit card holders than married people.
- Around 53 percent of our credit card holders are single.

Feature Analysis – AGE wise Defaulter Prediction



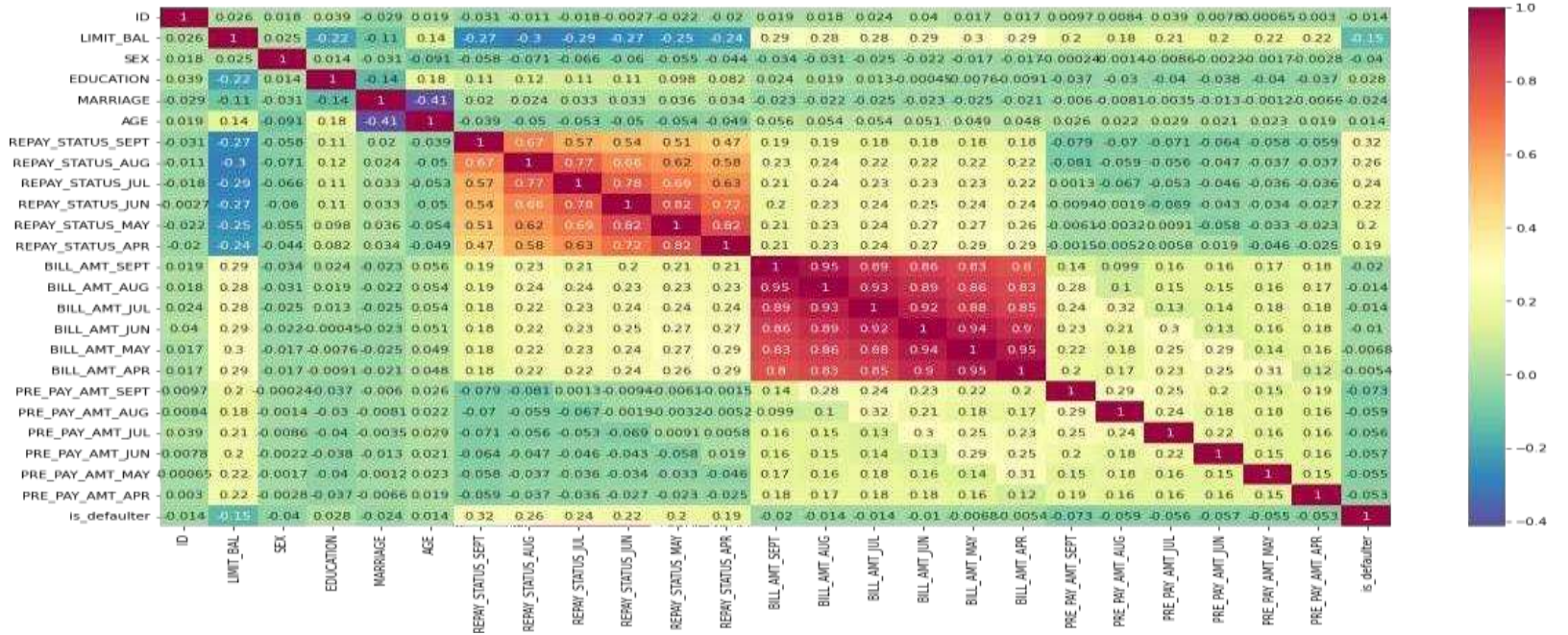
- We can clearly see from the above graph that as age increases, the no. of credit cards are low i.e. we don't see many elderly people that have credit cards.
- Also we can see that most no. of credit cards are held by people in 23 to 35 age bracket.

Observation on payment history



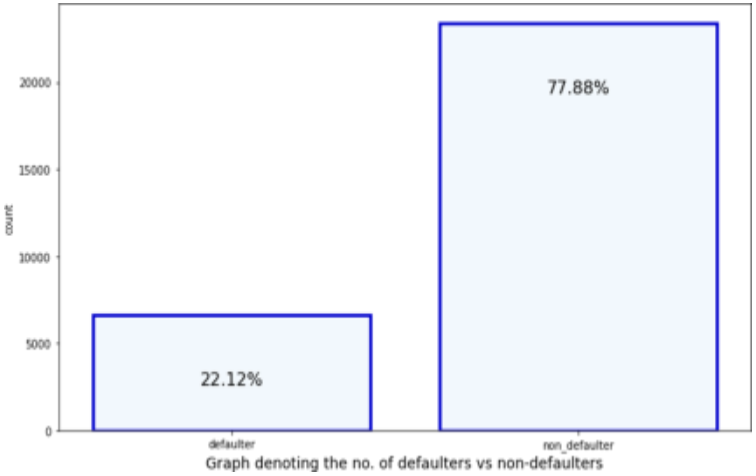
After careful observation, we found that most credit card payment by the customers were on time. We see the distribution of the plot peak at 0 value, which means that on the x scale no delay in the payment of card.

Feature Analysis - Correlation Heatmap

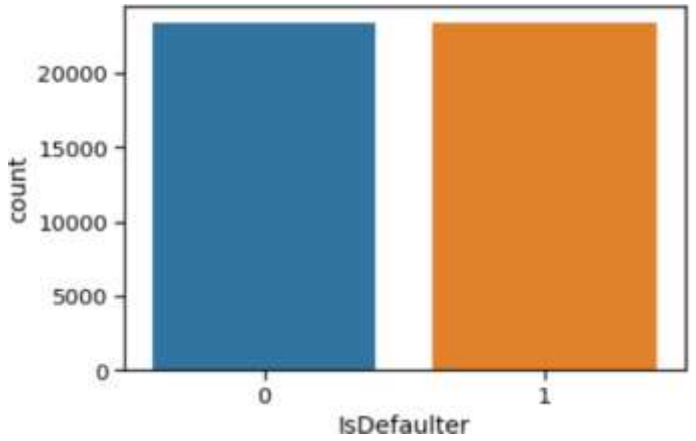


The correlation matrix helps us visualize the correlation between our numerical variables. We can see in the heatmap that there is some correlation in our bill amount features.

SMOTE(Synthetic Minority Oversampling Technique)



Graph denoting the no. of defaulters vs non-defaulters



Before SMOTE

After SMOTE

Data Pre-processing

- Feature engineering
- Feature selection
- Train test data split (67%-33%)
- SMOTE oversampling(Synthetic Minority Oversampling Technique)
- Data Fitting and Tuning
- Start with default model parameters
- Hyperparameter tuning
- Measure AUC- ROC after training data
- Model Evaluation
- Model testing
- Precision Recall Score
- Compare with the other models

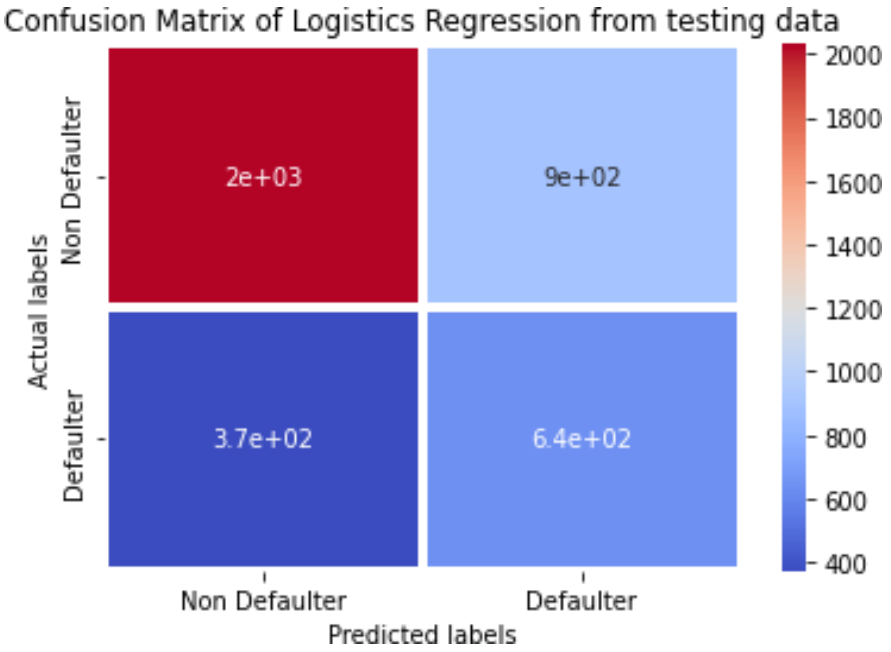
Modeling Overview

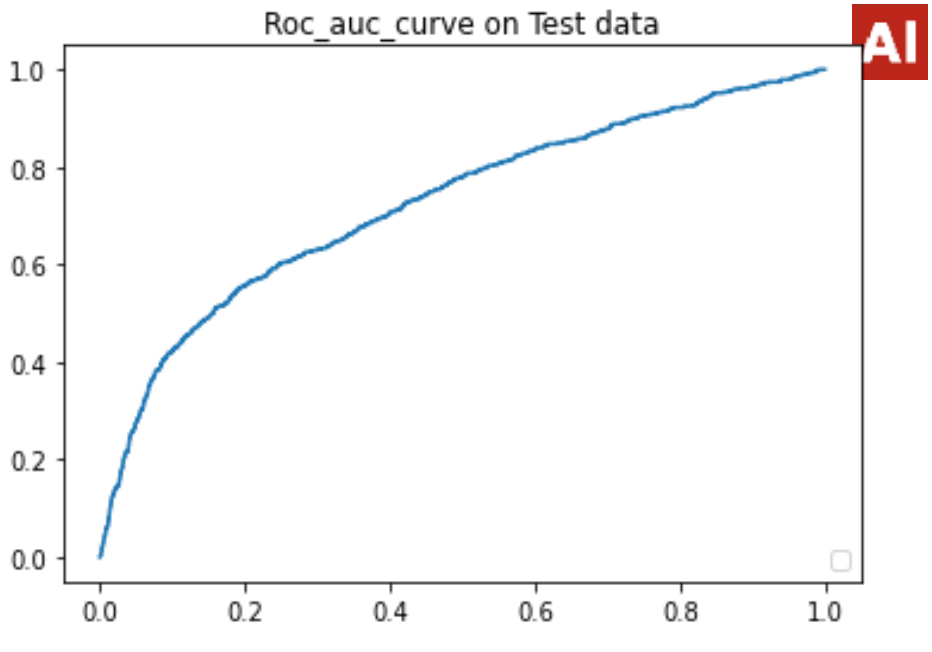
This is Classification problem statement.

There is Imbalance data with 78% non-defaulters and 22% defaulters.

We have applied following models :

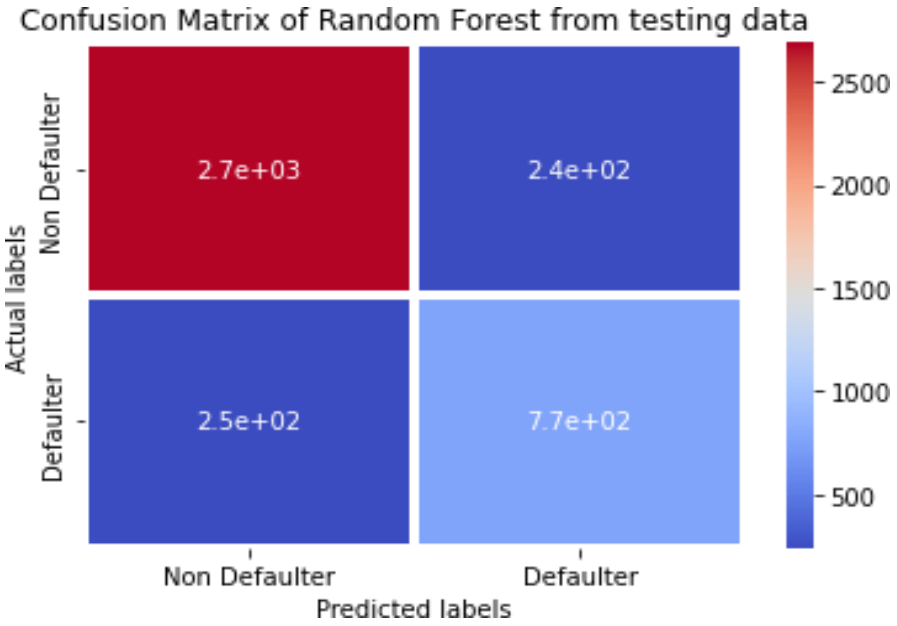
- Logistic Regression
- KNN
- Random Forest Classification
- SVM

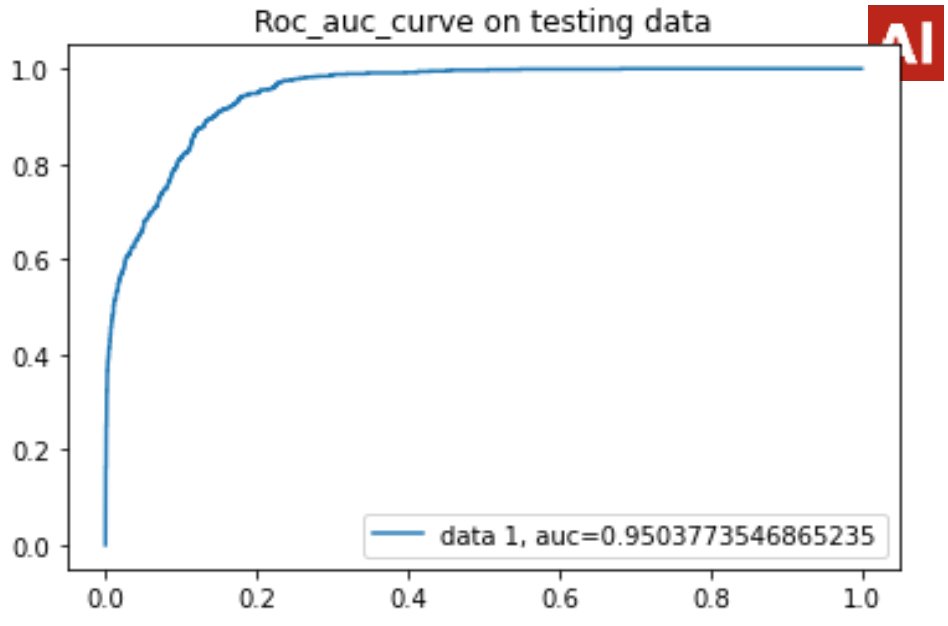




Logistic Regression Implementation

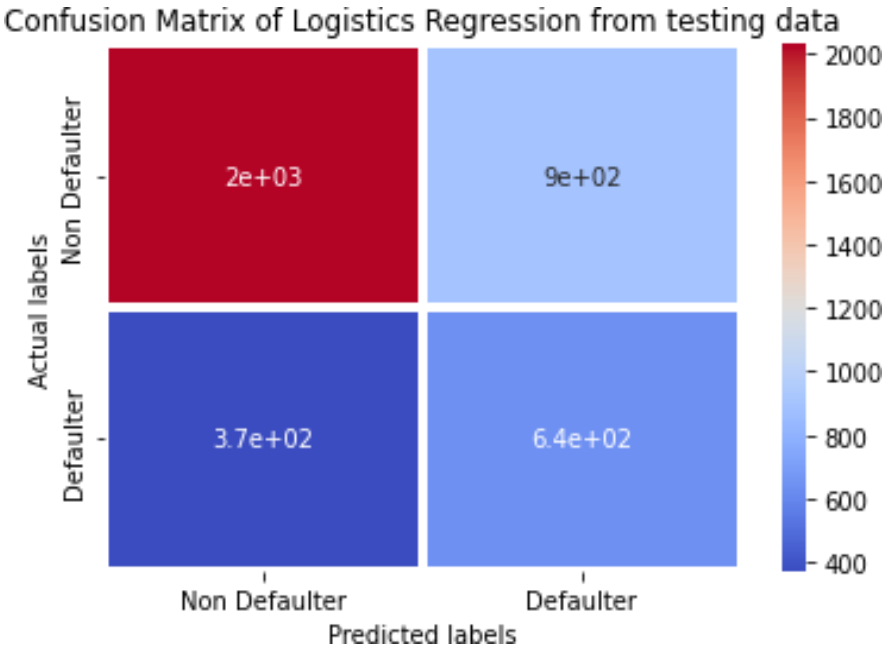
- We can see the plotted Roc-Auc curve of logistic regression.
- The logistic regression makes predictions with accuracy of 0.677, precision 0.416, recall of 0.633 and roc auc score of 0.663 & The F1 score is 0.502

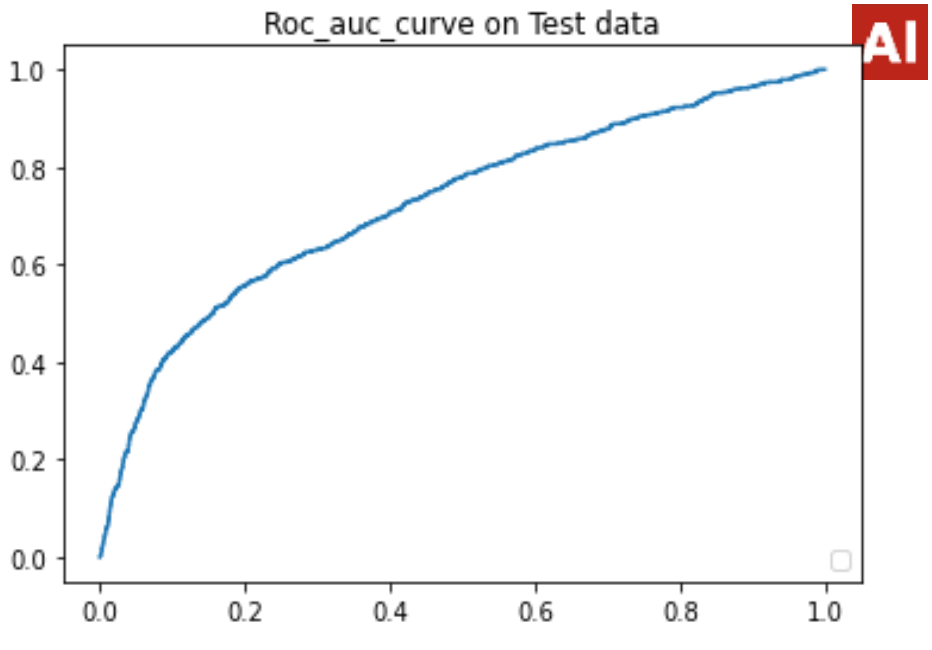




Random Forest Implementation

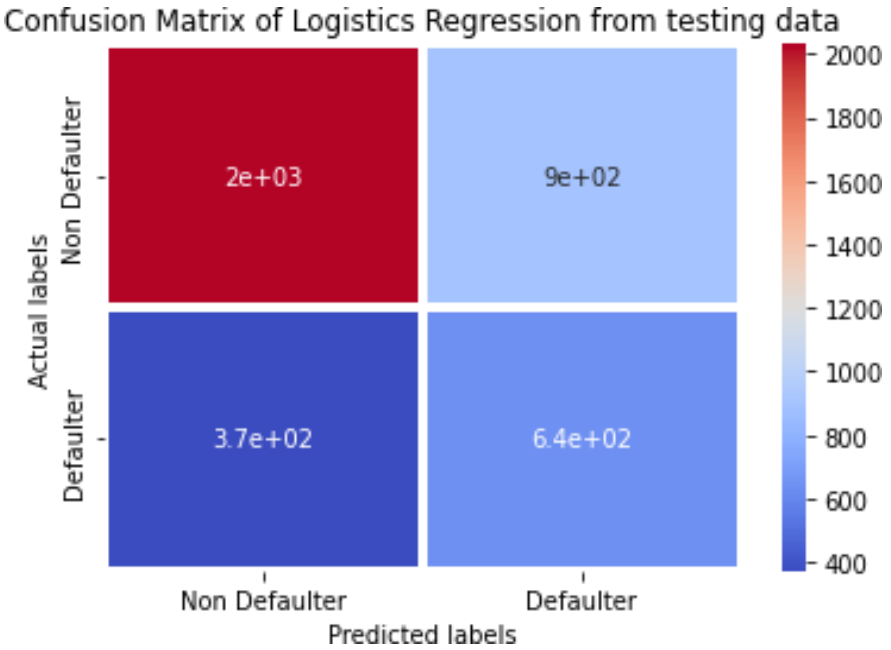
- We can see the plotted Roc-Auc curve of Random Forest classification.
- The random forest model makes predictions with accuracy of 0.876, precision 0.760, recall of 0.757 and roc auc score of 0.837 & The F1 score is 0.759

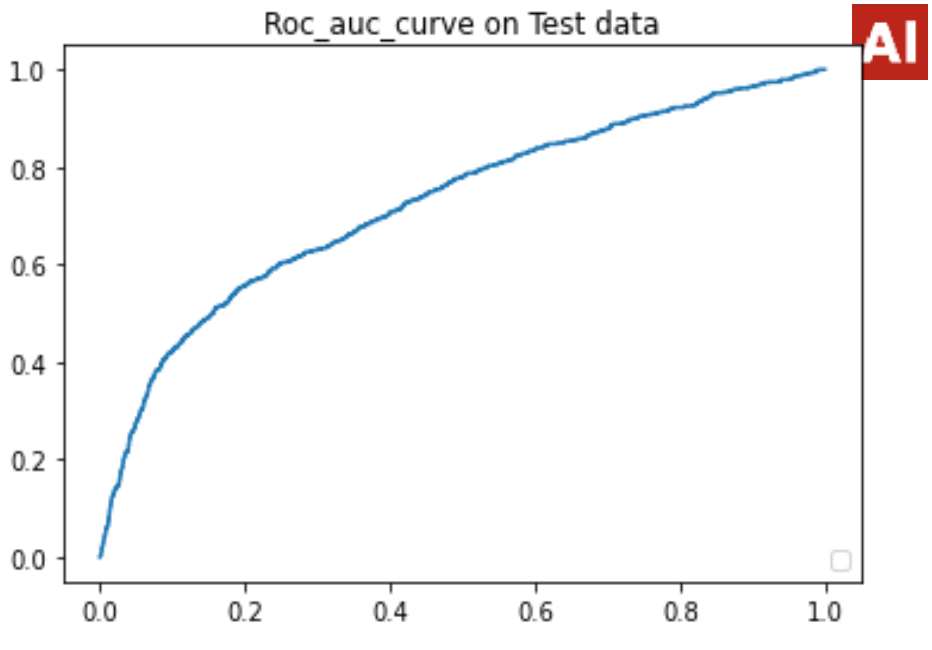




K Neighbors Classifiers Implementation

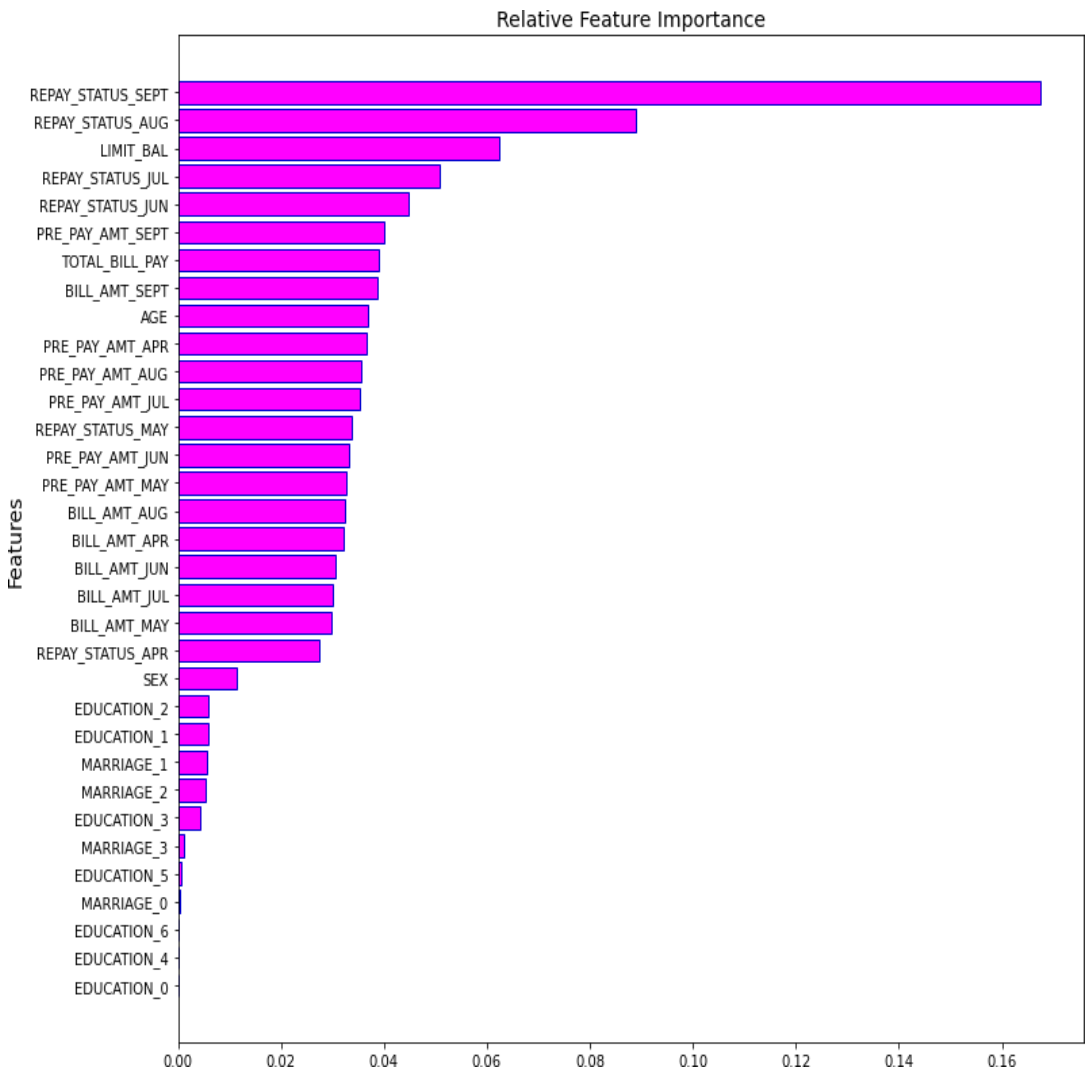
- We can see the plotted Roc-Auc curve of K neighbors classifier.
- The K neighbors classifier makes predictions with accuracy of 0.859, precision 0.680, recall of 0.855 and roc auc score of 0.858 & The F1 score is 0.758





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

Challenges

- Reading the dataset and understanding the problem statement.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Dealing with Imbalanced Dataset
- Feature engineering
- Feature selection - Making sure we don’t miss any important feature.
- Careful tuning of hyperparameters as it affects accuracy.
- Computation time was a big challenge for us.

Conclusion

Descriptive Analytics

In conclusion, the data exploration of credit card default dataset shows：

- No Missing values were found.
- Larger percentage of females than males in default payment category.
- Percentage of customers having graduate/Uni school degree is higher in default payment category.
- Individuals having single status have higher percentage of default than married, age peaks around 28-29 years in default payment category.
- As per the data collection, the number of defaulters lies in the range of 22% usually across all predictors.

Q & A