**Capstone Project - 3**

**Credit Card Default Prediction Supervised ML Classification Model**

**Raavi sahu**

# Understanding the concept

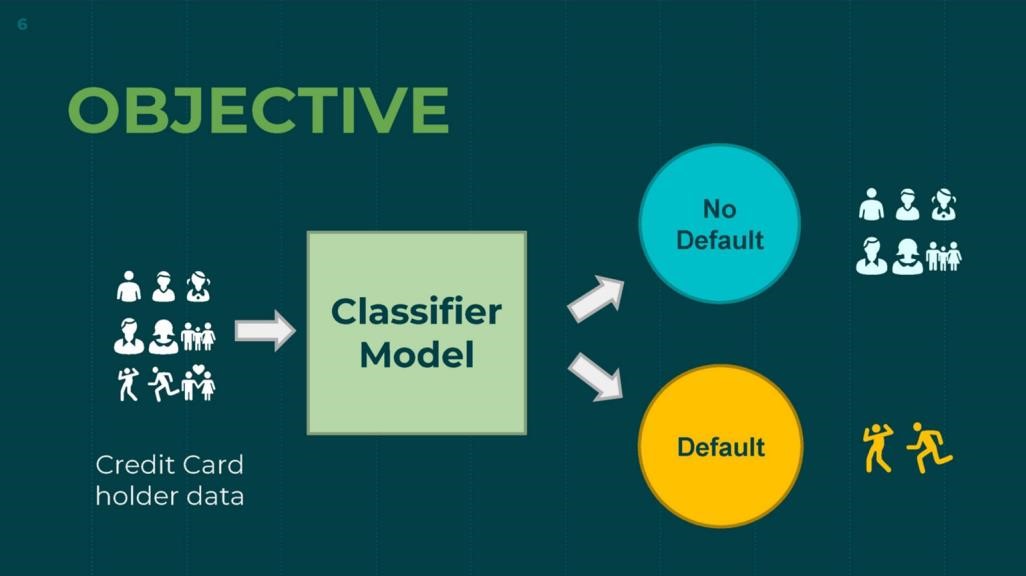
**What is credit card default?**

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 inspection and cleaning Exploring data, checking for outliers

Clean data to get it ready for Analysis

* EDA & Data Pre-processing Checking distributions of variable

Univariate and multivariate analysis

Checking for imbalanced dataset

* Modelling (Implementing Machine Learning Algorithms)

Logistic

KNN

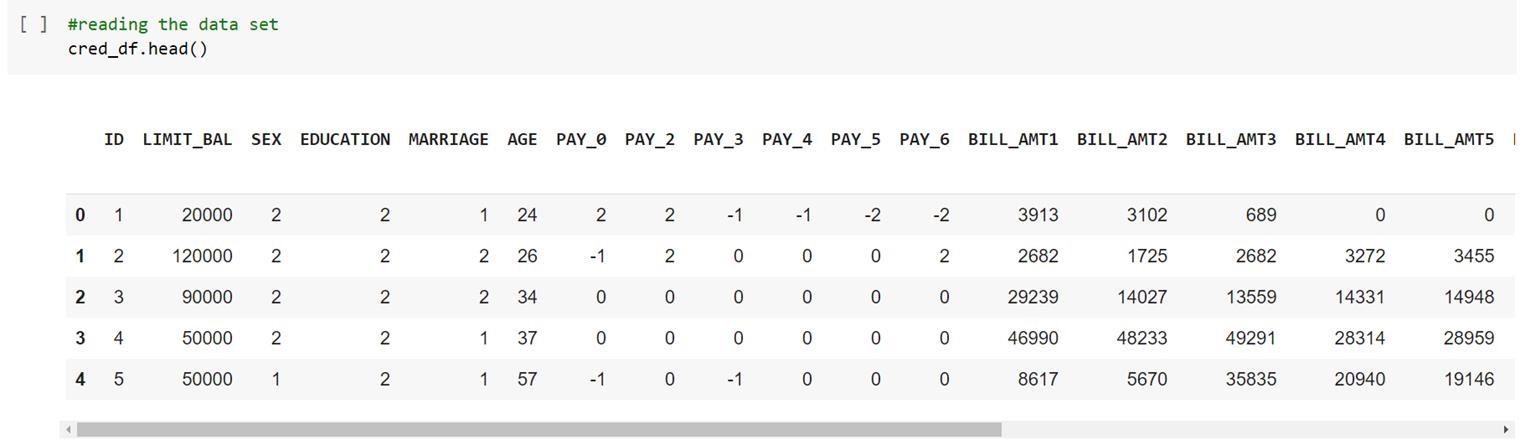
Random Forest +Decision Tree

XGBoost

SVM

## 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.



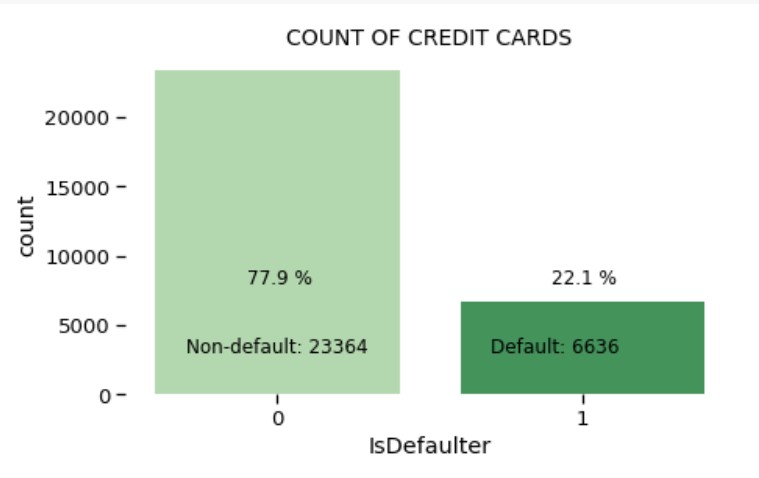
## Feature Analysis - The frequency of defaults

Looking at data, we

get this idea that it is

a case of Imbalanced

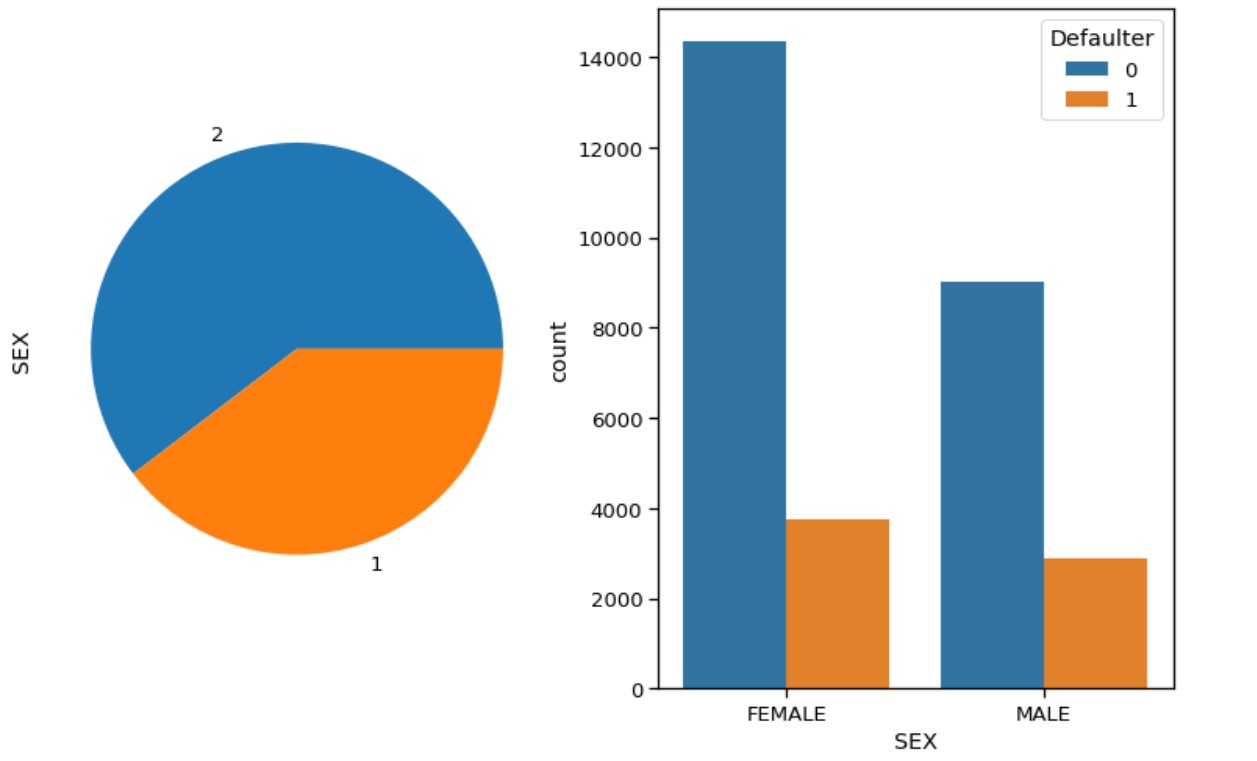
dataset.



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

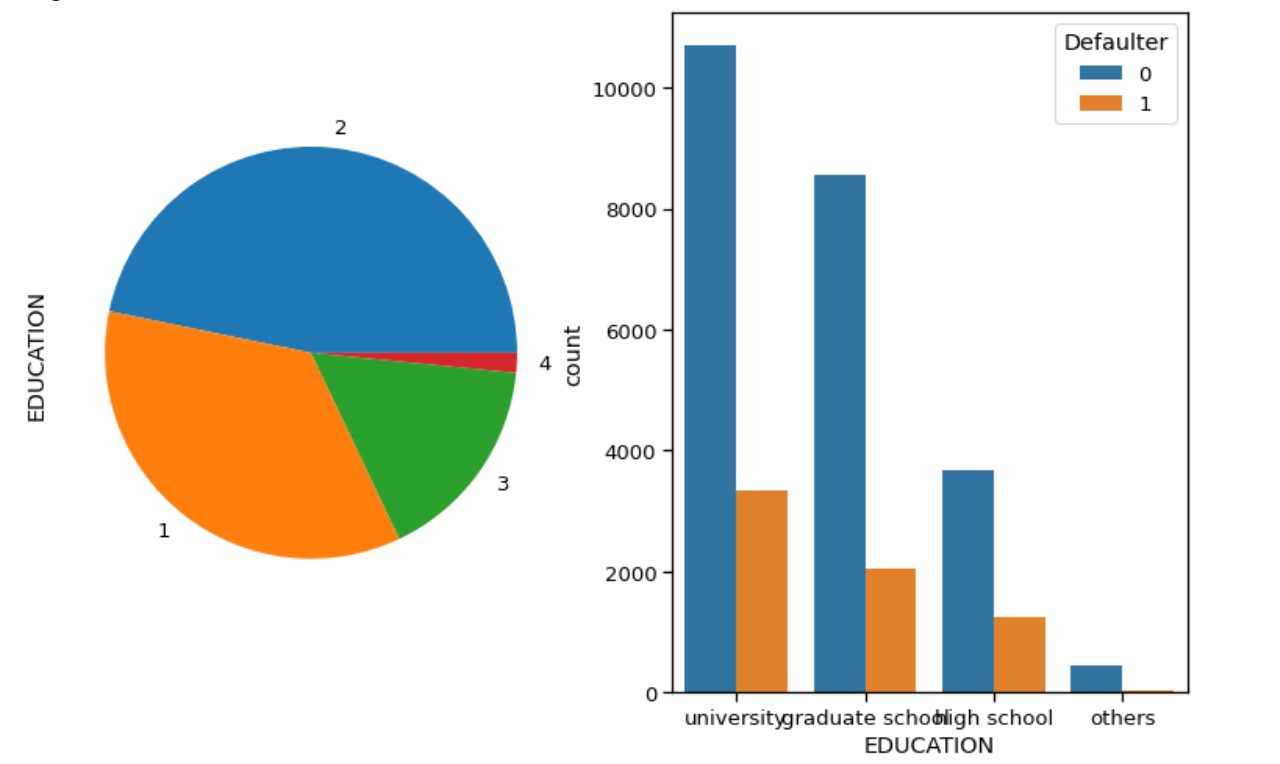
**Non- Default data is 77.9% while Default cases are 22.1% as in dataset**.

### Feature Analysis - Gender wise defaulter prediction



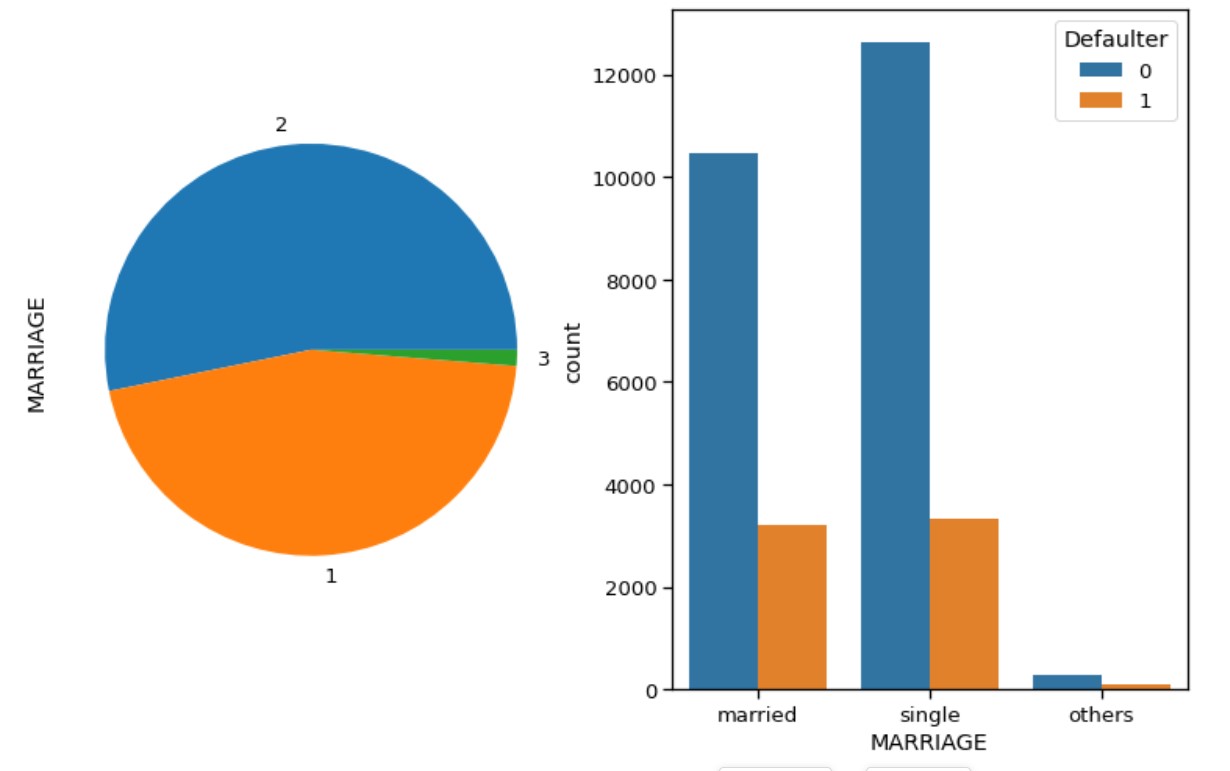
● **Larger percentage of females than males in default payment category.**

#### Feature Analysis - Education wise defaulter prediction



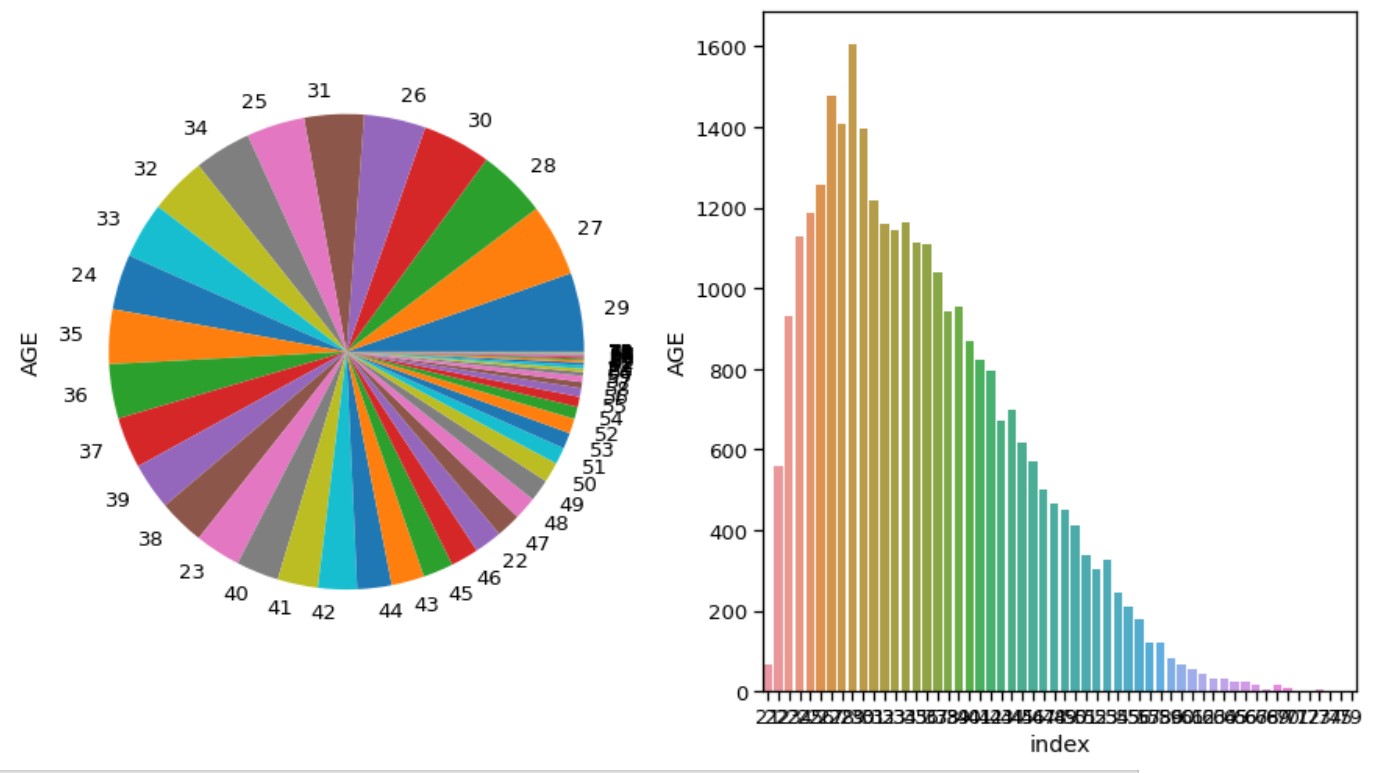
**Number of defaulters have a higher possibility that the person is educated (graduated from school or university).**

##### Feature Analysis – Marital status wise defaulter prediction



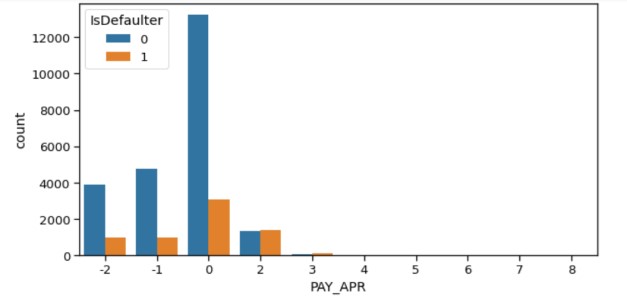
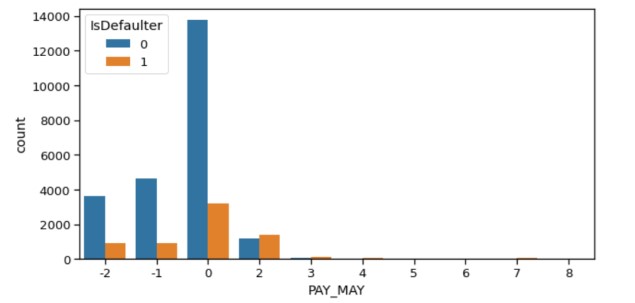
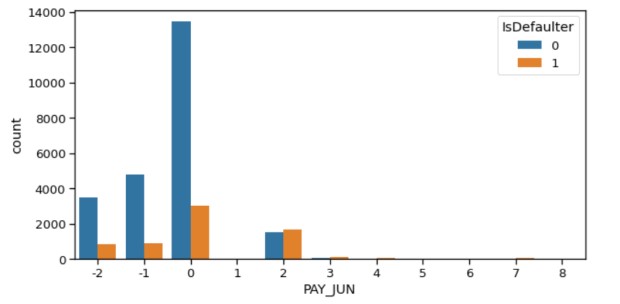
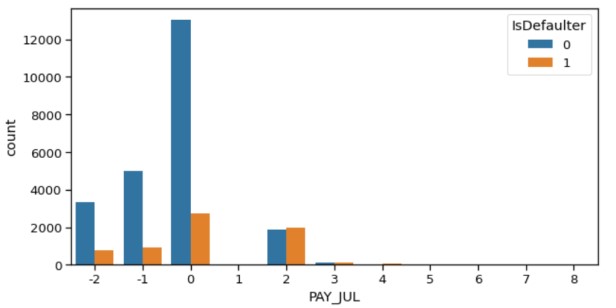
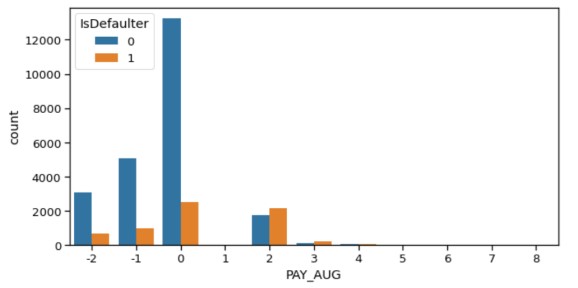
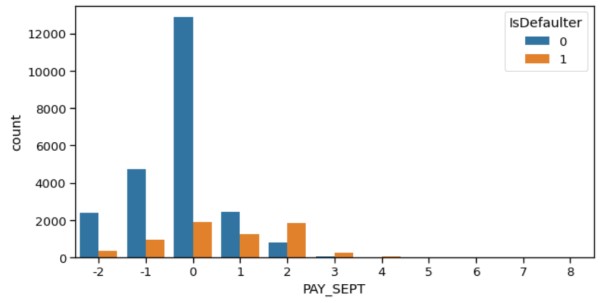
**Individuals having single status have higher percentage of default than married.**

## Feature Analysis – AGE wise Defaulter Prediction



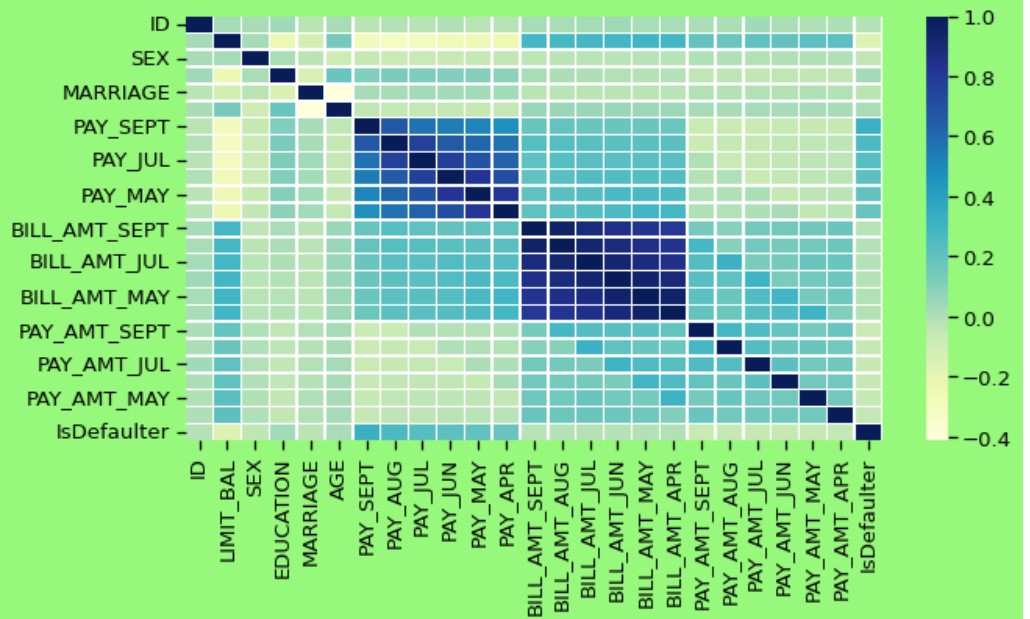
● **Age peaks around 28-29 years in default payment category. But there is no major relation between age and defaulter prediction.**

## Observation on payment history



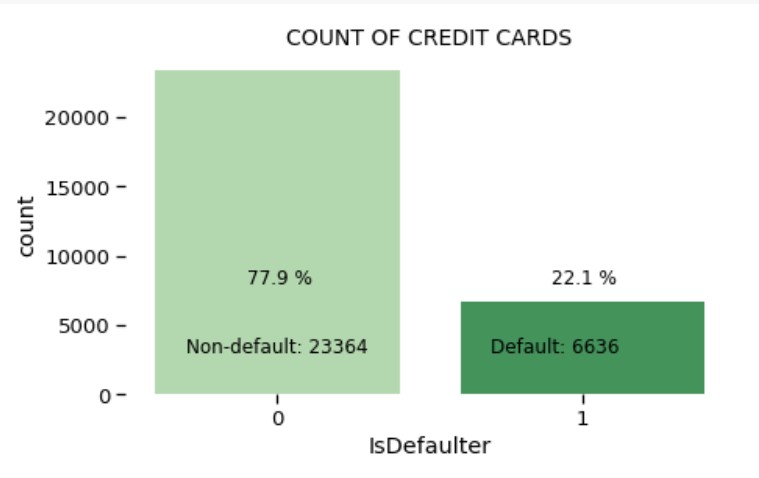
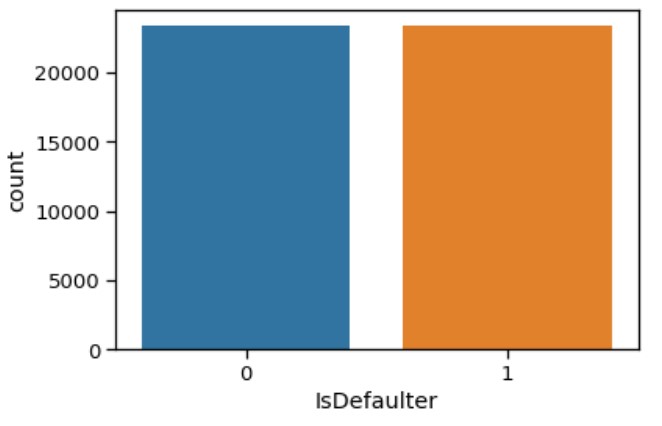
**After careful observation, we found that most credit card payment by the customers were on time. We see the distribution of the plot peak at 0 value, which means that on the x scale no delay in the payment of card.**

### Feature Analysis - Correlation between parameters



**We can see that no correlation between the features. So, no need to remove or drop some features.**

#### SMOTE(Synthetic Minority Oversampling Technique)



● **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 :

Default or

Not\_defa

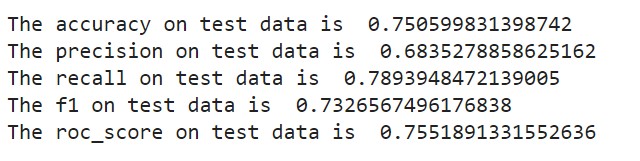
ult



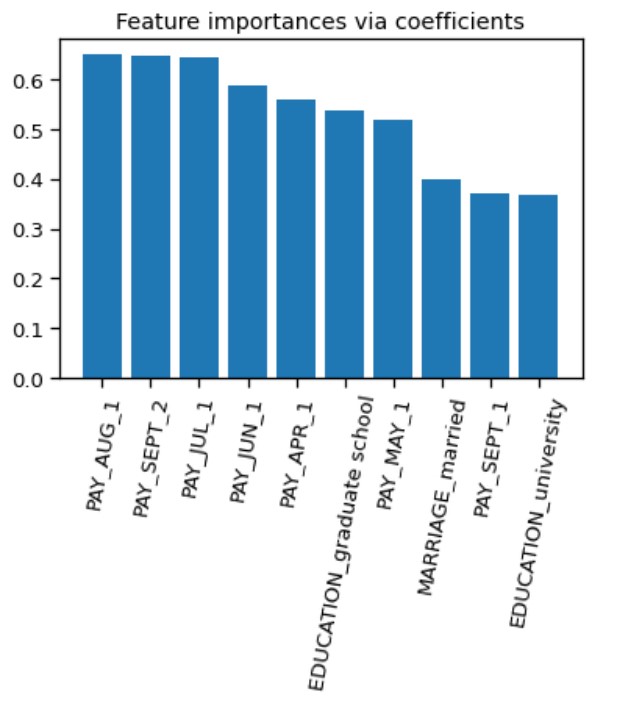
* Logistic Regression
* KNN
* Decision Tree
* Random Forest
* XGBoost
* SVM

## Logistic Modelling

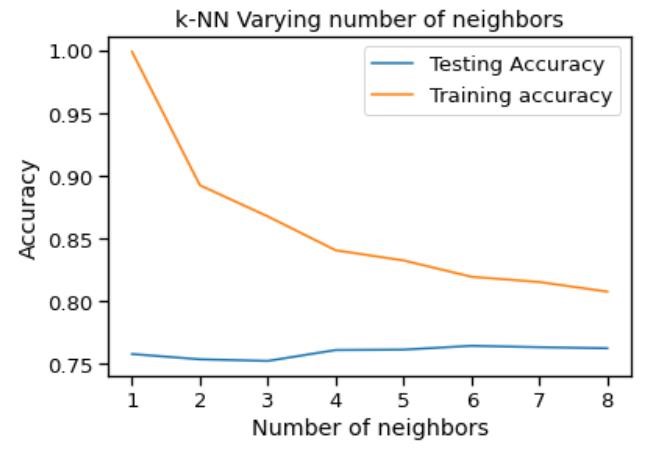
**Best parameters : C= 0.01, penalty = L2**



**Features importance according to logistic regression**

● **Payment-month is important feature with respect to different month’s defaulter rate.**

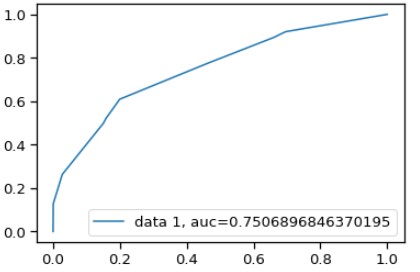
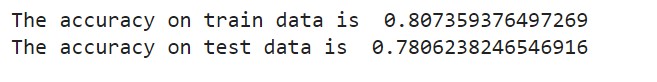
## Implementing K-NN



We found that k=7 is the best nearest neighbour because the test and train accuracy meet near at this point so calculating model performance and we get AUC-ROC score around 83% after running KNN model for the data set.

## Decision Tree

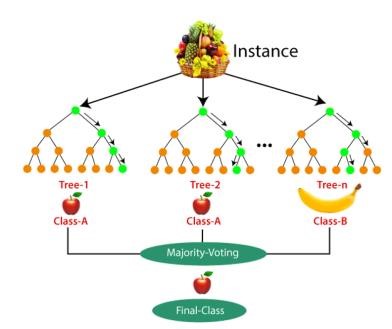
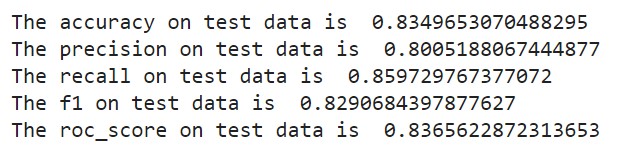
Best parameters : max\_depth = 20, min\_samples\_split = 0.1



The decision tree AUC-ROC score is 75% which not satisfactory comparison to other model score.

## Implementing Random Forest Classifier

**Best Parameters : max\_depth = 30, n\_estimators = 200**



Random forest having approx. 92% AUC

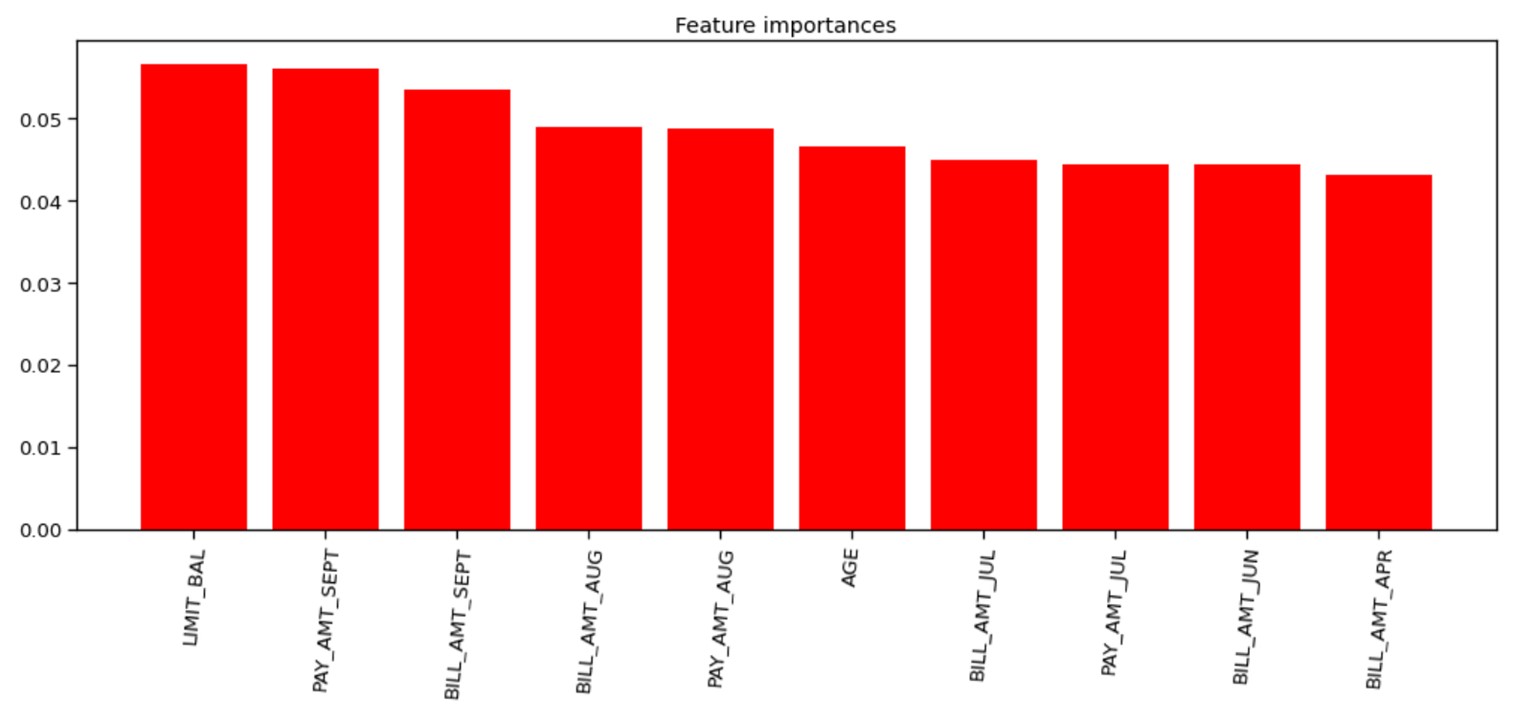
-

ROC score

which is high then the all model score or the best

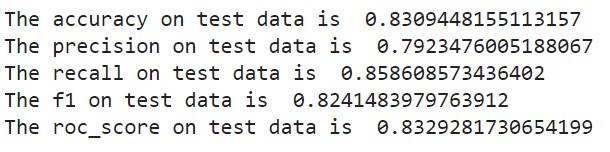
recall approx. 86%

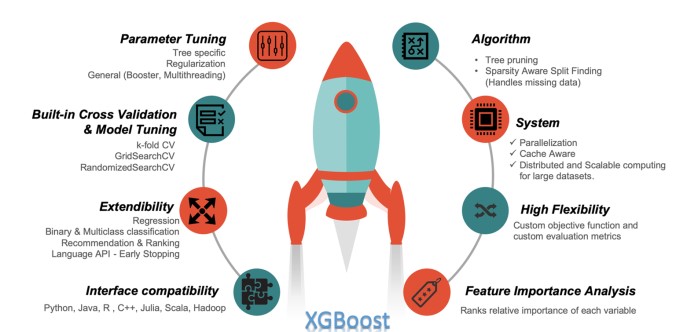
## Feature Importance according to Random forest



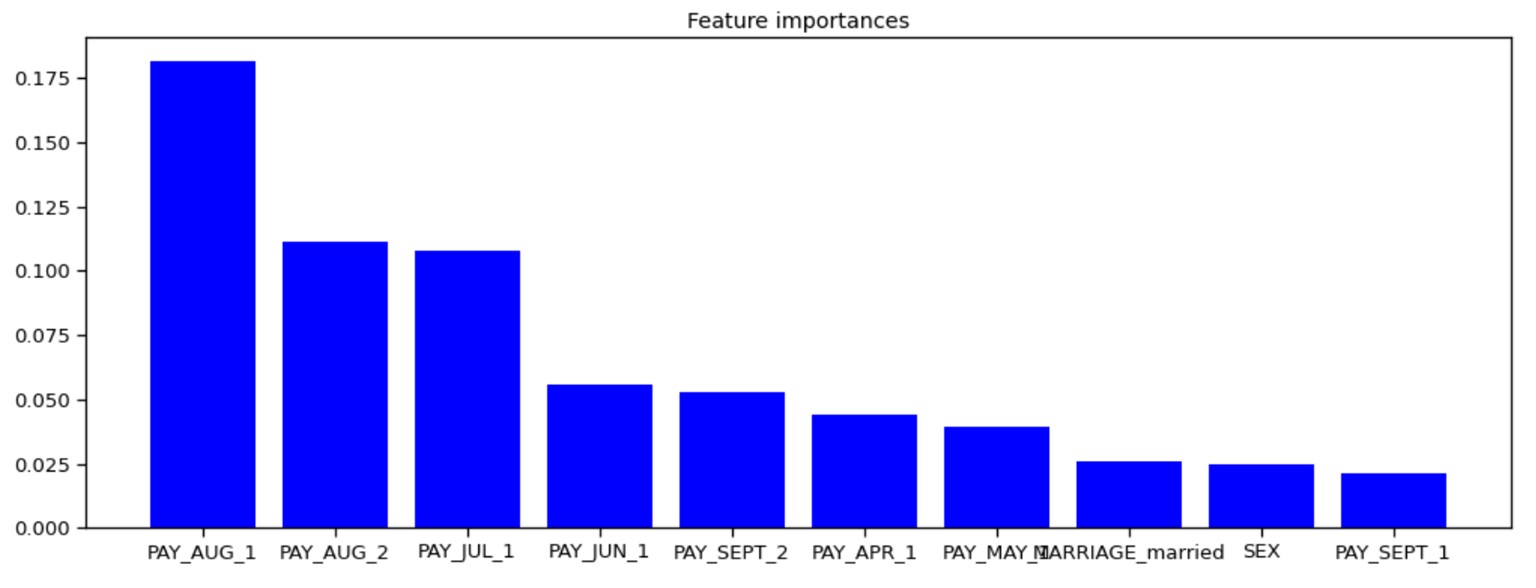
Credit limit is the major reason of defaulter as per Rf

## Implement the XGBoost





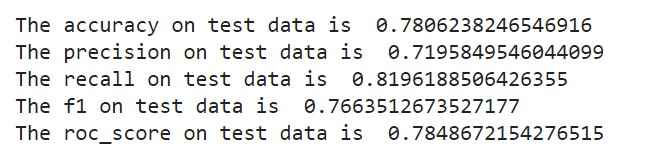
## Feature importance according to XGBoost



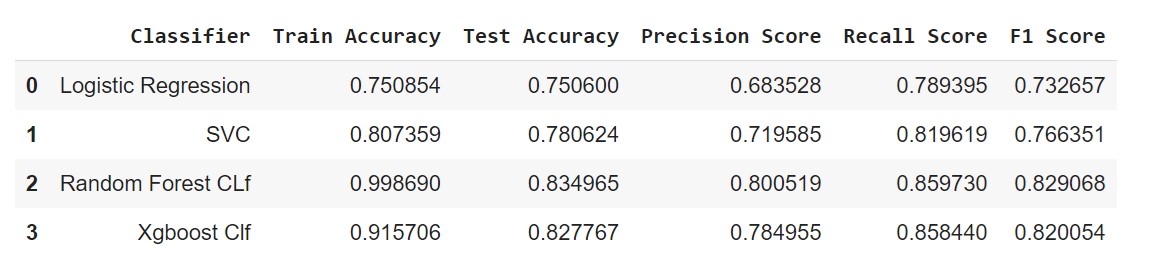
**Pay\_Aug\_1 is the most important feature here.**

## SVM (support vector machine)

**Best parameters : C = 10, kernel = rbf**

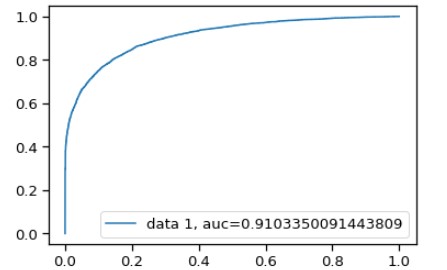
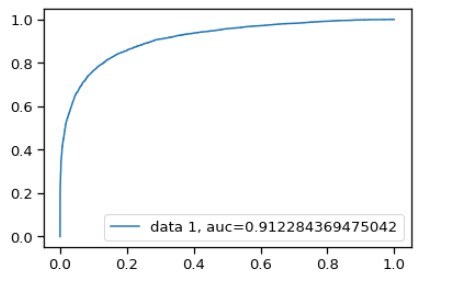


# Model Evaluation



**Random Forest is the winner here with highest recall score 0.85.**

# Model Evaluation



● **Random Forest score XGBoost**

## Observations on basis of AUC – ROC score

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | Observations\_1 : We started with Logistic  Classifier and then with XGBoost , for which we obtained an AUC score of 0.81 and 0.91, respectively, when predicting the target for the test set. We followed with an KNN model, with a lower AUC score (0.83) for the prediction of the test set target value. We then followed with a SVM, with the AUC score after training iterations 0.85. | | |  | | --- | | Observations\_2 : We then experimented with an **Random Forest** model. In this case, we used the validation set for validation of the training model. The best AUC score obtained was **0.912.** | |

# 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.

## Conclusion while training model

* Random Forest provided us the best results giving us a recall of 86% (It means out of 100 defaulters 86 will be correctly caught by Random Forest ).
* XGBoost also had good score as well with recall of 85.8%.
* Logistic regression being the least accurate with a recall of nearly 78%.

