Capstone Project – 03
Supervised ML – Classification
Credit Card Default Prediction

Presented By :- Abhishek Kirar

Presentation Outline

- Problem Statement
- Introduction
- Data Summary
- Methodology
- Exploratory Data Analysis
- Data Processing
- Implementing ML algorithms
- Challenges
- Conclusion

Problem Statement

- 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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

Introduction

- **The aim of this study is to exploit some supervised machine learning algorithms to identify the key drivers that determine the likelihood of credit card default, underlining the mathematical aspects behind the methods used. Credit card default happens when you have become severely delinquent on your credit card payments. In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, the overused credit card for consumption and accumulated heavy credit and debts**

- **The goal is to build an automated model for both identifying the key factors, and predicting a credit card default based on the information about the client and historical transactions. The general concepts of the supervised machine learning paradigm are later reported, together with a detailed explanation of all techniques and algorithms used to build the models. In particular, Logistic Regression, Random Forest and Support Vector Machines algorithms have been applied.**


Exploring
the Dataset
Sales by day of the week
Total earned
Number of sales
Wednesday
Thursday
Stats by rating earning
60 %
70 %

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

Features:

- **ID: ID of each client**
- **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, 5=unknown, 6=unknown)**
- **MARRIAGE: Marital status (1=married, 2=single, 3=others)**
- **AGE: Age in year**
- **PAY_0: Repayment status in September, 2005 (-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)**
- **PAY_2: Repayment status in August, 2005 (scale same as above)**
- **PAY_6: Repayment status in April, 2005 (scale same as above)**

Features(Contd.):

- **BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)**
- **BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)**

  …;
- **BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)**

  — —
- **PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)**
- **PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)**

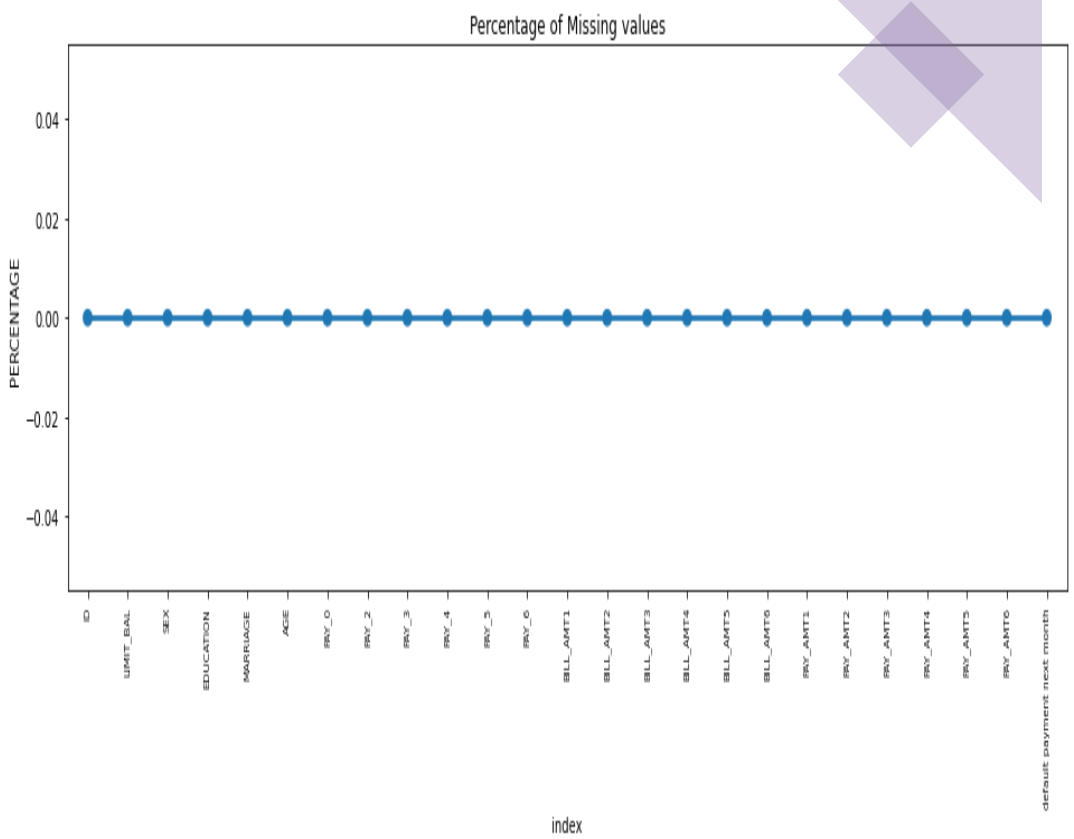
  …;
- **PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)**

  — —
- **Default payment next month: Default payment (1=yes, 0=no)**

Missing value :

- Plot the graph to check whether th ere are any missing value present.

- As we can see above there are no missing value presents thankfully.



Methodology

Data preparation and exploratory analysis

Building predictive model using multiple techniques & algorithm

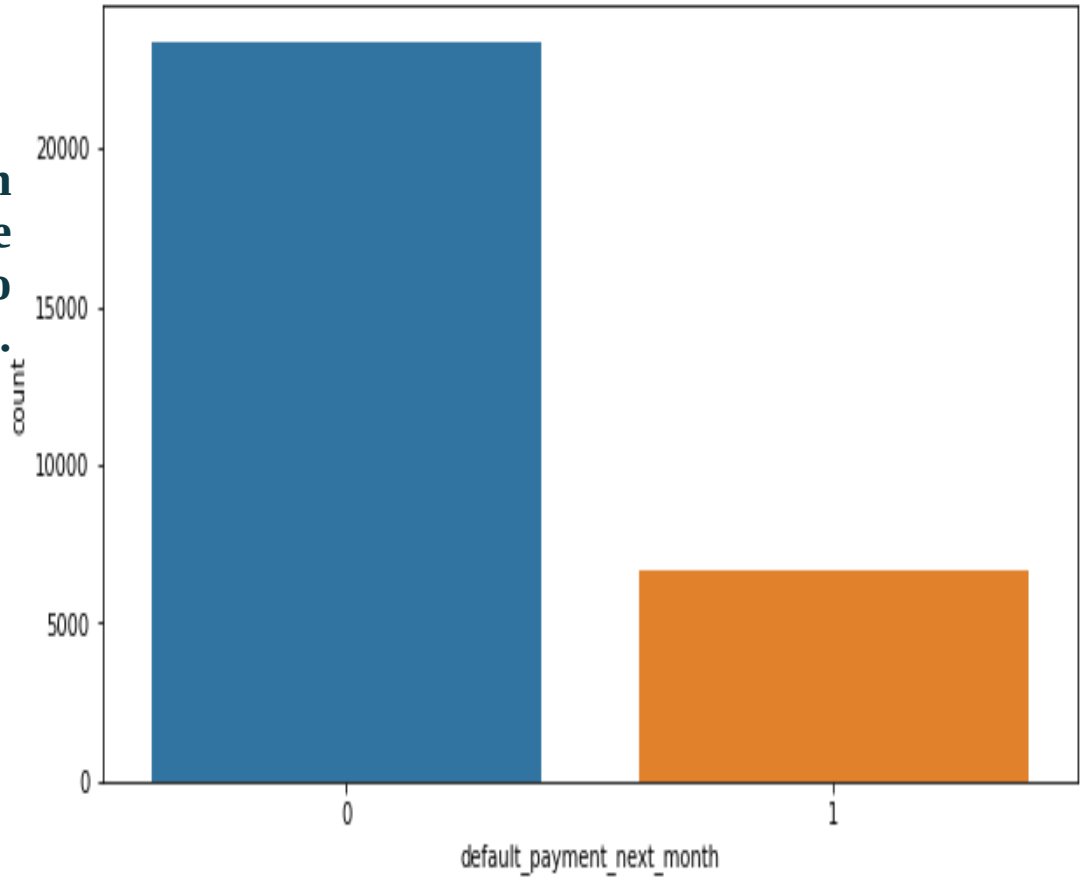
Optimal Model identified through testing and evaluation

EDA AND DATA PROCESSING

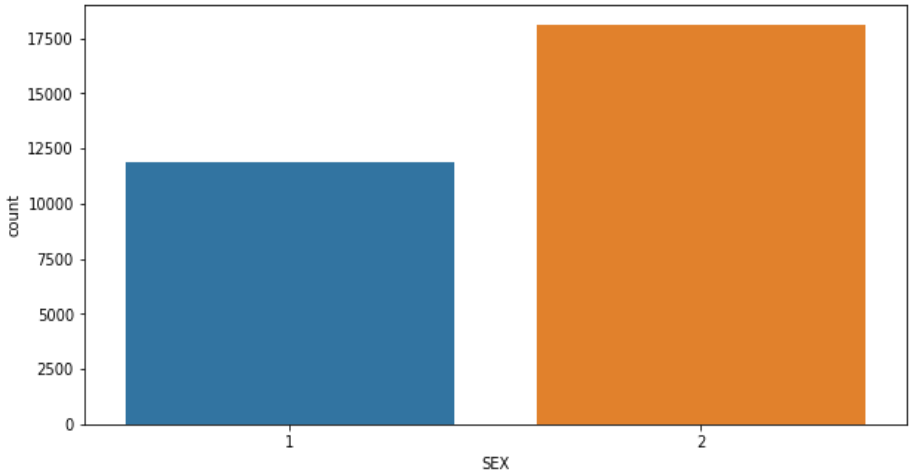
ANALYSIS OF DEPENDENT VARIABLE

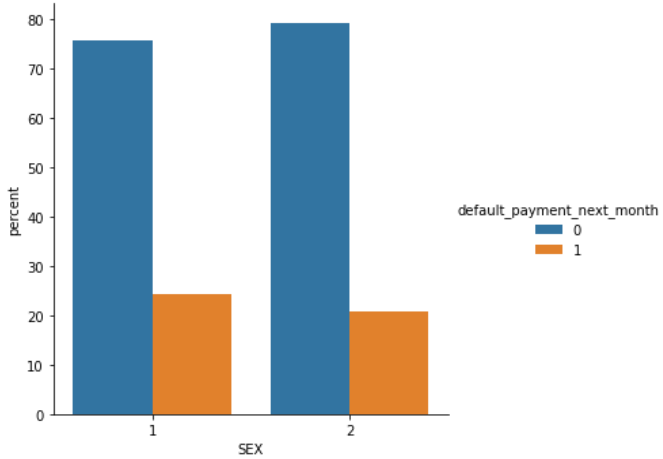
- As we can see from graph that both classes are not in proportion and we have imbalanced dataset. We need to do normalize the data in next step.

- 0 - Not Default
- 1 – Default
- Defaulters are less than the Non Defaulters in the given dataset.



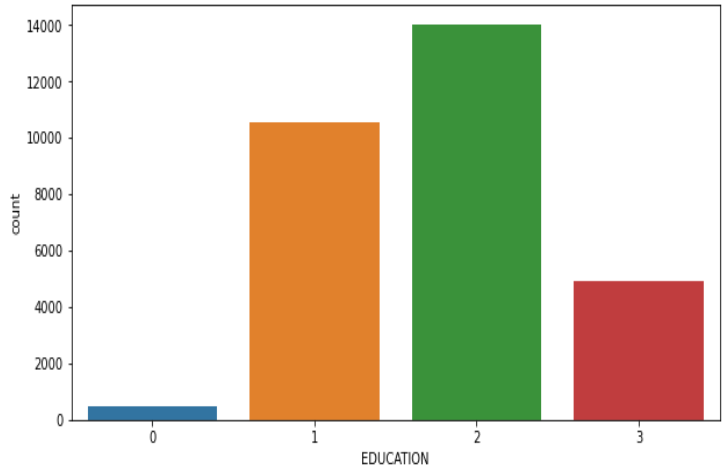
ANALYSIS OF SEX VARIABLE

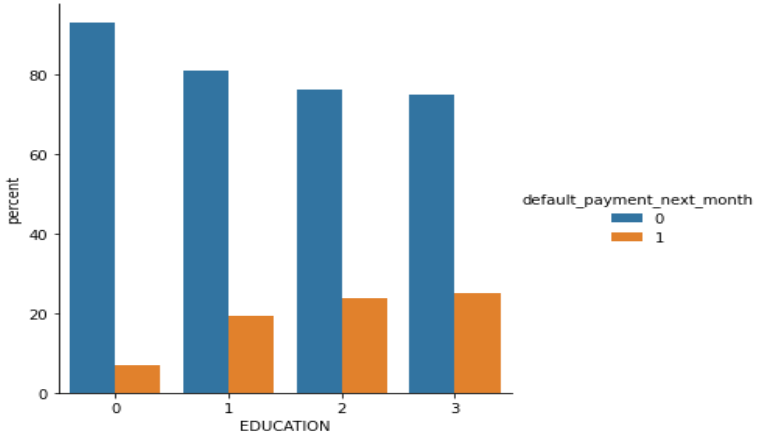




- **1 - Male**
- **2 - Female**
- **Number of Male credit holder is less than Female.**
- **It is evident from the above graph that the number of defaulter have high proportion of males**

ANALYSIS OF EDUCATION VARIABLE

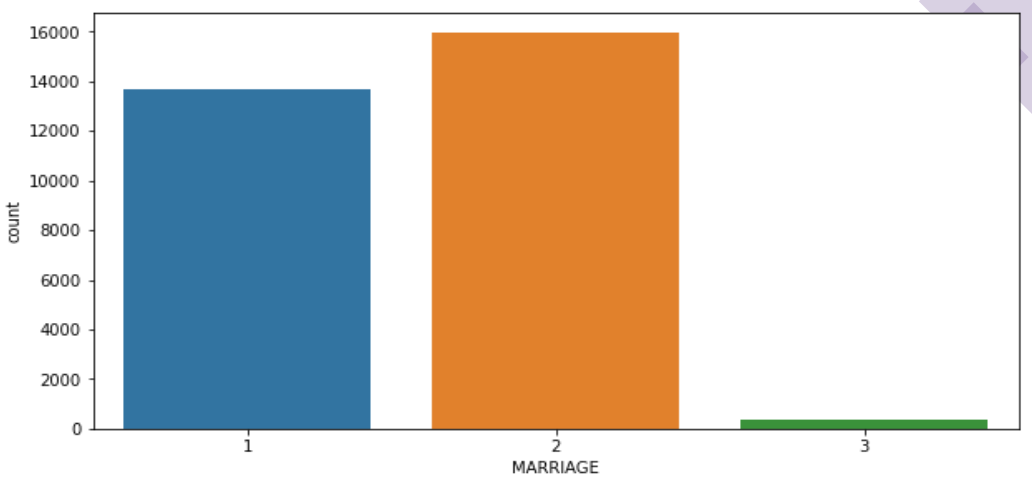


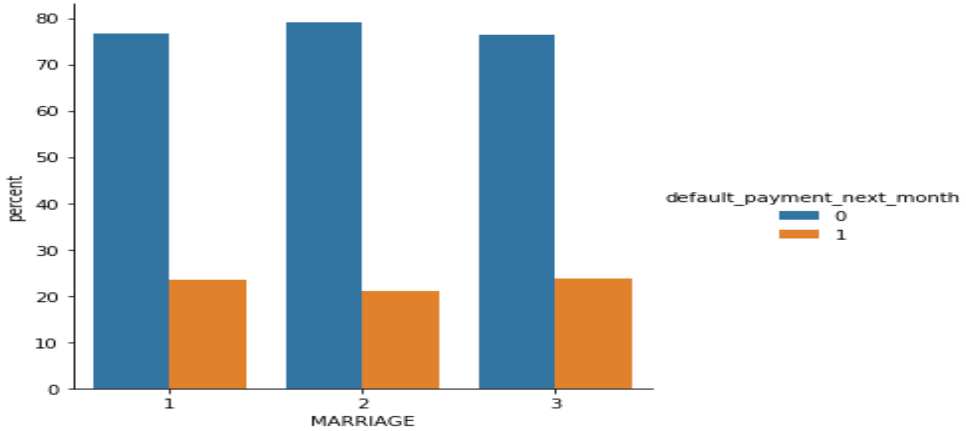


- 1=graduate school, 2=university, 3=high school, 0=others
- From the above left side plot we can say that more number of credit holders are university students followed by Graduates and then High school students.
- From the right side plot it is clear that those people who are other students have higher default payment w.r.t. graduates and university people.

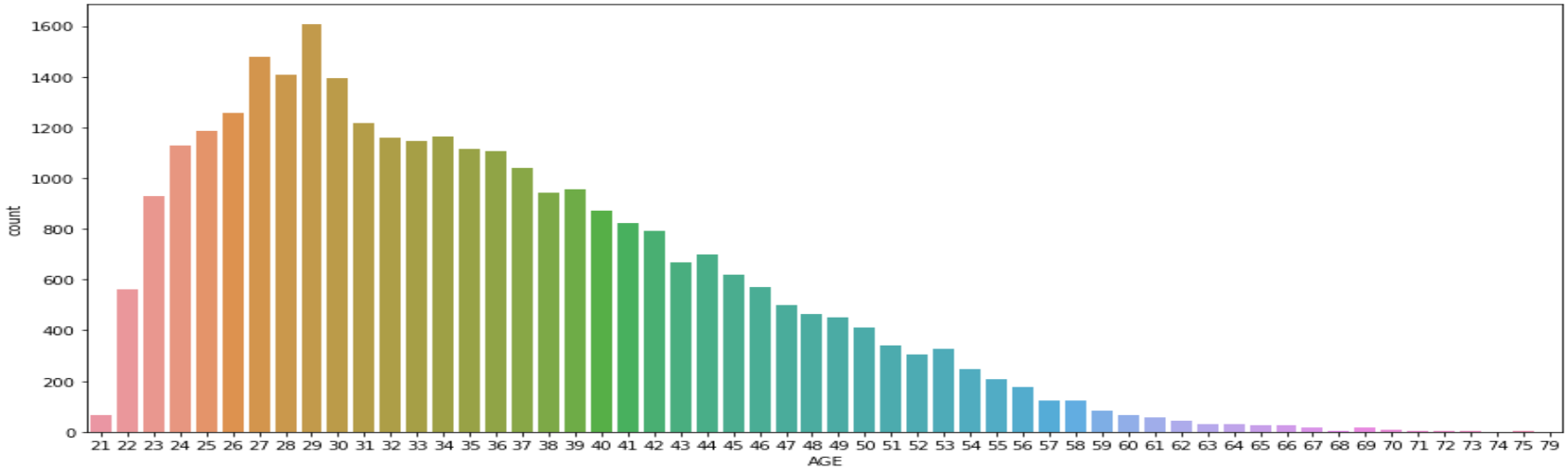
ANALYSIS OF MARRIAGE VARIABLE

- 1 – married, 2 – single, 3 - others
- From the above data analysis we can say that
- More number of credit cards holder are Single.
- High defaulter rate when it comes to other



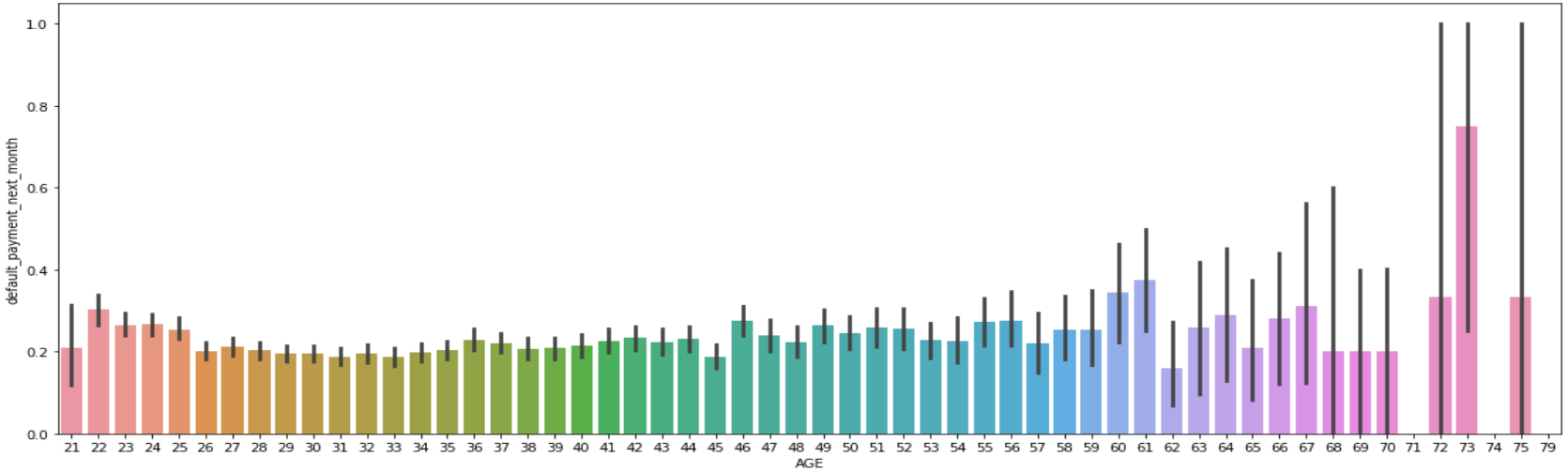


ANALYSIS OF AGE VARIABLE



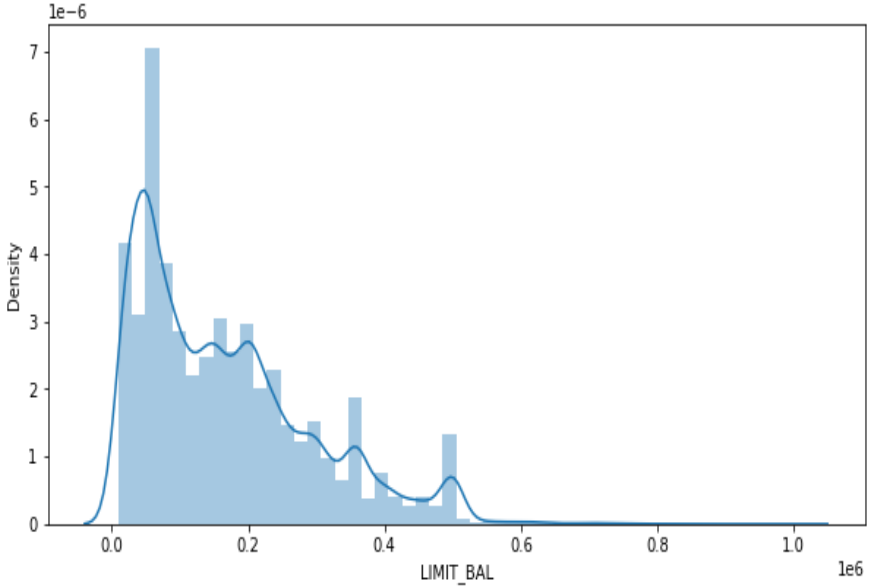
- **From the above count plot analysis we can say that**
- **We can see more number of credit cards holder age are between 26-30 years old.**
- **Age above 60 years old rarely uses the credit card.**

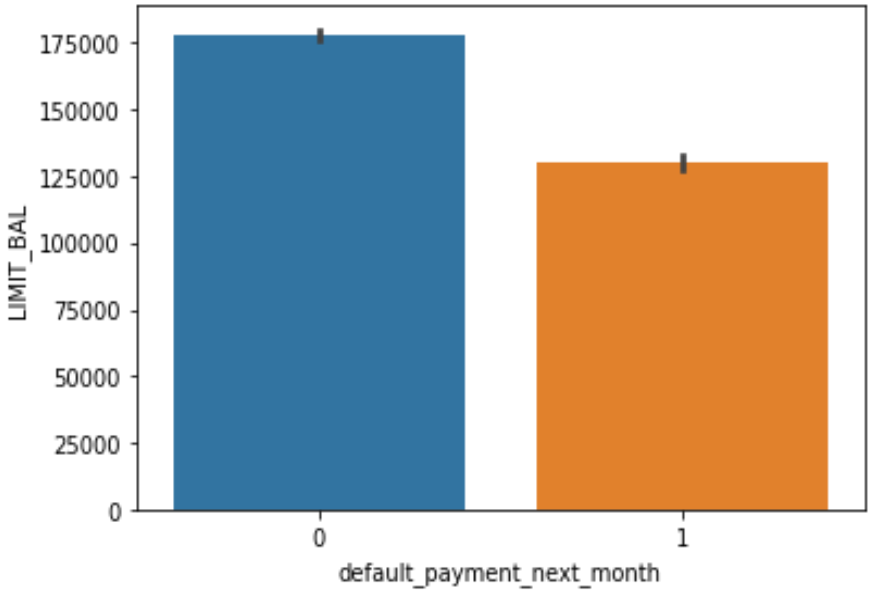
ANALYSIS OF AGE VARIABLE



- **From the above bar plot which shows the relationship between age and defaulter, we can say that those who default are 60 years and older, that may be they don't use their card frequently.**

ANALYSIS OF LIMIT BALANCE VARIABLE

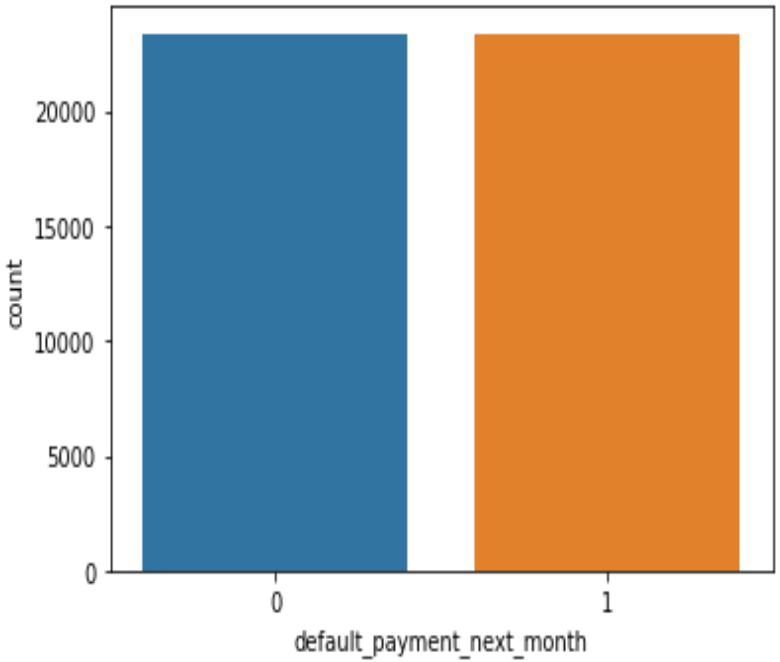




- **From the above plots analysis we can say that maximum amount of given credit in NT dollars is 50,000 followed by 30,000 and 20,000.**

SMOTE

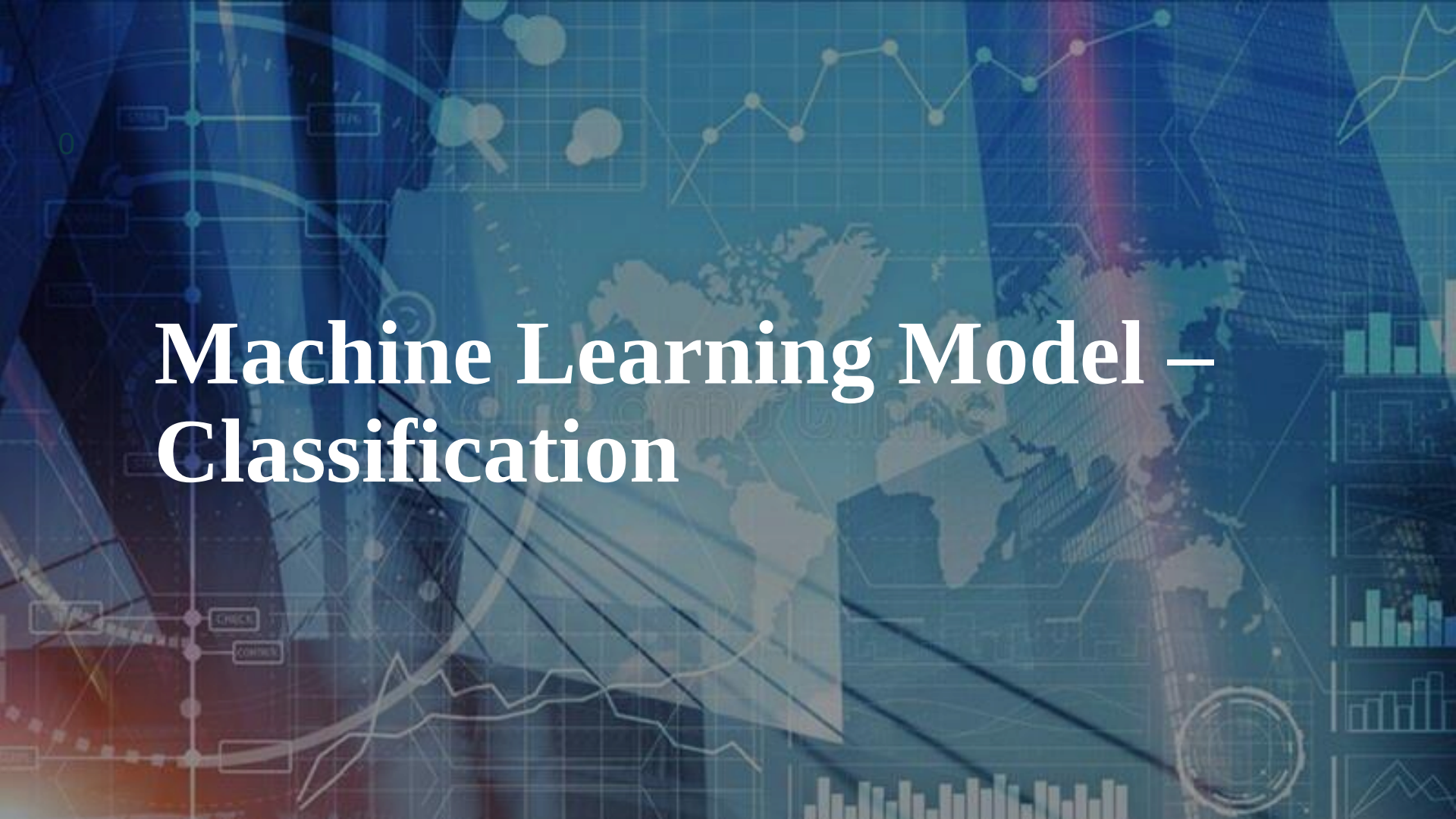
- **SMOTE (Synthetic Minority Oversampling Technique)- Oversampling is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.**
- **After performing SMOTE operation we get this balance dataset**



ONE HOT ENCODING

- **One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.**
- **Here we perform one hot encoding on 'EDUCATION','MARRIAGE','PAY_SEPT',**
- **'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR‘.**
- **and label encoding for ‘SEX’**
- **After this we get these features in our dataset:**

(['LIMIT_BAL', 'SEX', 'AGE', 'BILL_AMT_SEPT', 'BILL_AMT_AUG','BILL_AMT_JUL', 'BILL_AMT_JUN', 'BILL_AMT_MAY', 'BILL_AMT_APR','PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUL', 'PAY_AMT_JUN','PAY_AMT_MAY', 'PAY_AMT_APR', 'default_payment_next_month','total_Payement_Value', 'Dues', 'EDUCATION_graduate school','EDUCATION_high school', 'EDUCATION_others', 'EDUCATION_university','MARRIAGE_married', 'MARRIAGE_others', 'MARRIAGE_single', 'PAY_SEPT_-1','PAY_SEPT_0', 'PAY_SEPT_1', 'PAY_SEPT_2', 'PAY_SEPT_3', 'PAY_SEPT_4','PAY_SEPT_5', 'PAY_SEPT_6', 'PAY_SEPT_7', 'PAY_SEPT_8', 'PAY_AUG_-1','PAY_AUG_0', 'PAY_AUG_1', 'PAY_AUG_2', 'PAY_AUG_3', 'PAY_AUG_4','PAY_AUG_5', 'PAY_AUG_6', 'PAY_AUG_7', 'PAY_AUG_8', 'PAY_JUL_-1','PAY_JUL_0', 'PAY_JUL_1', 'PAY_JUL_2', 'PAY_JUL_3', 'PAY_JUL_4','PAY_JUL_5', 'PAY_JUL_6', 'PAY_JUL_7', 'PAY_JUL_8', 'PAY_JUN_-1','PAY_JUN_0', 'PAY_JUN_1', 'PAY_JUN_2', 'PAY_JUN_3', 'PAY_JUN_4','PAY_JUN_5', 'PAY_JUN_6', 'PAY_JUN_7', 'PAY_JUN_8', 'PAY_MAY_-1', 'PAY_MAY_0', 'PAY_MAY_1', 'PAY_MAY_2', 'PAY_MAY_3', 'PAY_MAY_4','PAY_MAY_5', 'PAY_MAY_6', 'PAY_MAY_7', 'PAY_MAY_8', 'PAY_APR_-1','PAY_APR_0', 'PAY_APR_1', 'PAY_APR_2', 'PAY_APR_3', 'PAY_APR_4','PAY_APR_5', 'PAY_APR_6', 'PAY_APR_7', 'PAY_APR_8']



Machine Learning Model – Classification

MODEL BUILDING

- **LOGISTIC REGRESSION**
- **RANDOM FOREST**
- **SVC**
- **XGBOOST**

LOGISTIC REGRESSION

Parameter - {'C': 10, 'penalty': 'l2'}

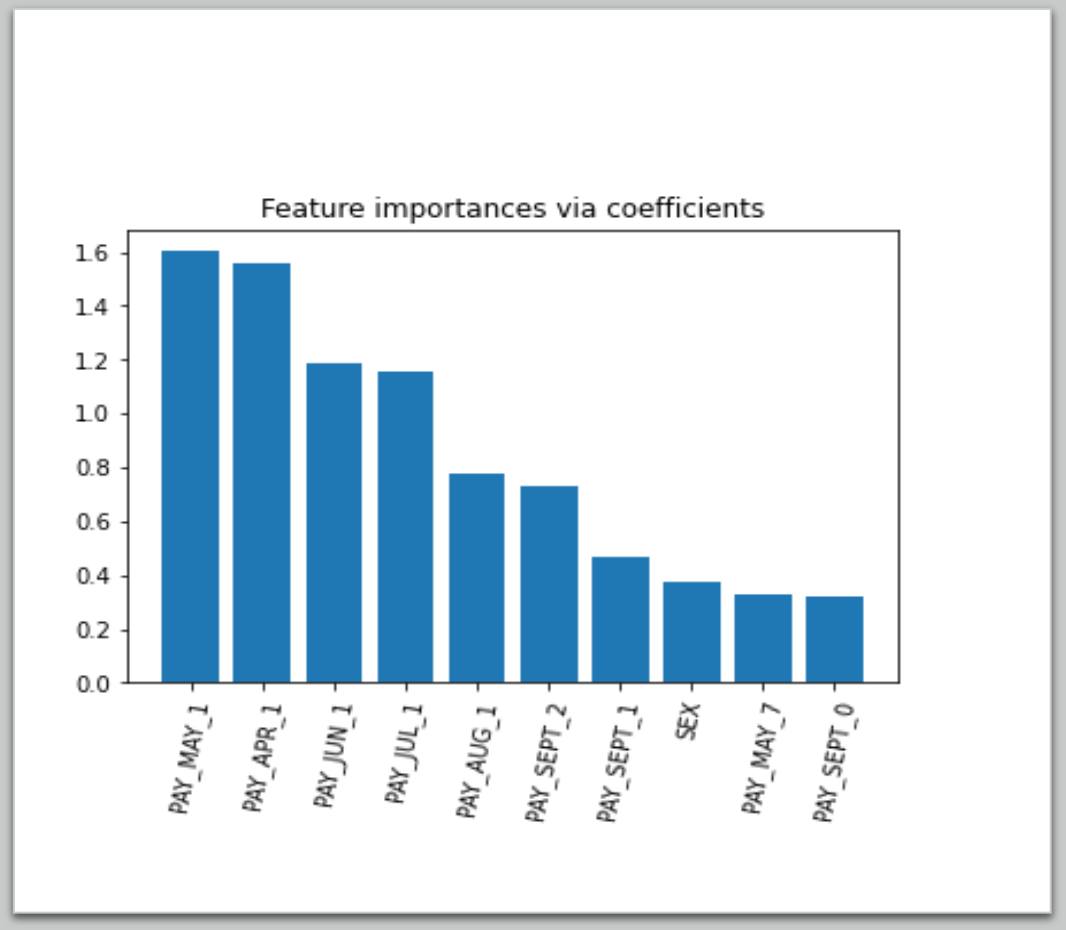
From this regression model we get the results as below:

- The accuracy on test data is 0.7483301990791777
- The precision on test data is 0.6782101167315175
- The recall on test data is 0.7888067581837381
- The f1 score on test data is 0.7293395634284121
- The roc_score on test data is 0.7533092025677562

We have implemented logistic regression and we are getting f1-score approx 73%. As we have imbalanced dataset, F1-score is better parameter. Let's go ahead with other models and see if they can yield better result.

FEATURE IMPORTANCES

- **From the feature importance graph we can say that the most important feature that make an impact on dependant variable are PAY_JUL_1, PAY_MAY_1, PAY_APR_1.**



RANDOM FOREST

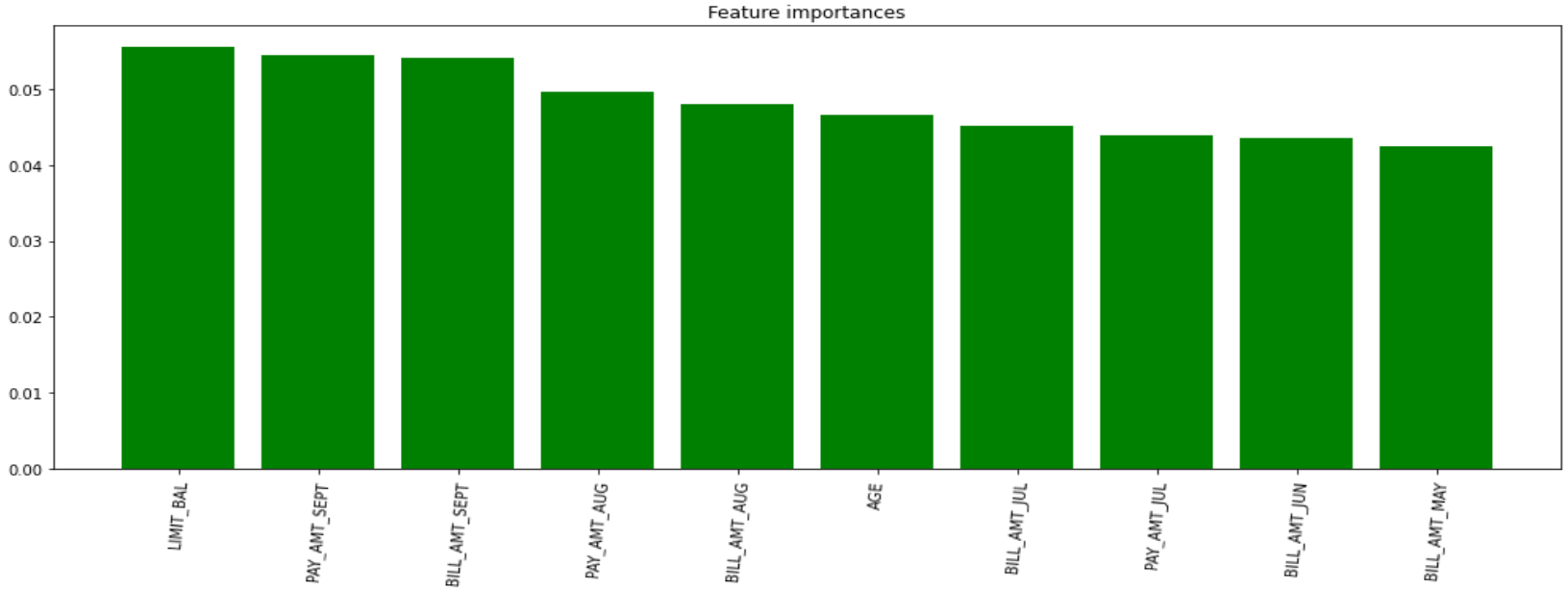
- **PARAMETERS :** {'max_depth': 30, 'n_estimators': 200}

From the regression model we get the results as below:

- The accuracy on test data is 0.8334738343816873
- The precision on test data is 0.799610894941634
- The recall on test data is 0.8576794657762938
- The f1 on test data is 0.8276278695126863
- The roc_score on test data is 0.8350100231833006

- After GridSearchCV we getting f1-sore approx 82%. As we have imbalanced dataset, F1- score is better parameter. Let's go ahead with other models and see if they can yield better result.

FEATURE IMPORTANCES



- **from the above feature importance graph we can say that the most important feature that make an impact on dependent variable are LIMIT_BAL, PAY_AMT_SEPT**

SUPPORT VECTOR CLASSIFIER (SVC)

- PARAMETERS -{'C': 10, 'kernel': 'rbf'}

From the regression model we get the results as below:

- The accuracy on test data is 0.776538486479476
- The precision on test data is 0.7045395590142672
- The recall on test data is 0.823030303030303
- The f1 on test data is 0.7591893780573027
- The roc_score on test data is 0.7823914693929431

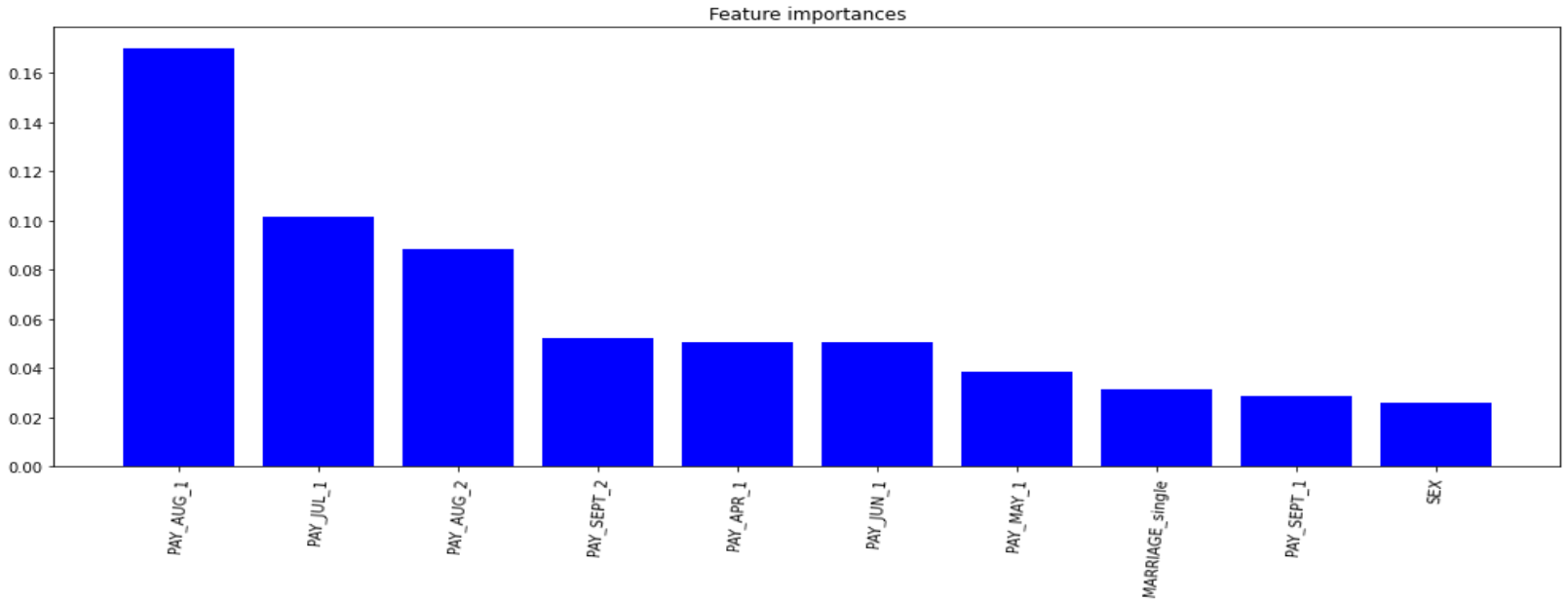
XGBOOST

- PARAMETERS :**{'max_depth': 15'min_child_weight': 8}**

- **From the regression model we get the results as below**

- The accuracy on test data is 0.8313338953375268
- The precision on test data is 0.7929961089494163
- The recall on test data is 0.8588284871470713
- The f1 on test data is 0.8246004450738418
- The roc_score on train data is 0.8332928270473974

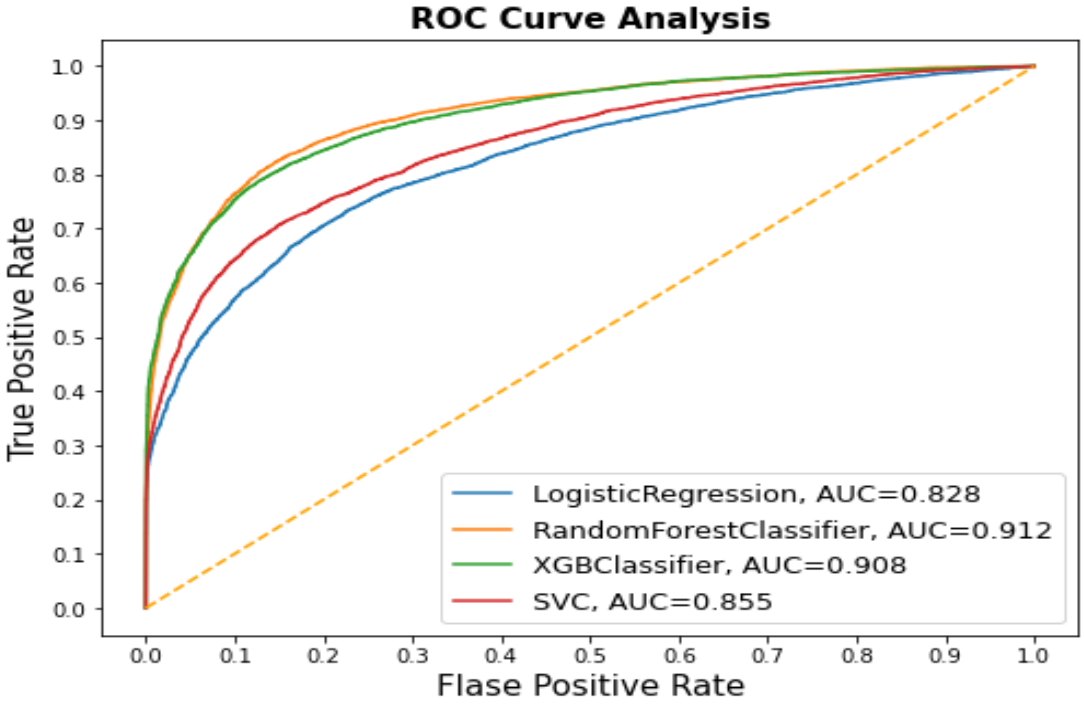
FEATURE IMPORTANCES



- **From the above feature importance graph we can say that the most important feature that make an impact on dependent variable are PAY_AUG_1**

AUC-ROC CURVE COMPARISON

- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
- This curve plots two parameters:
- 1. True Positive Rate
- 2. False Positive Rate
- AUC stands for "Area under the ROC Curve“. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve.



EVALUATING THE MODELS

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.751461	0.749108	0.682879	0.787113	0.731301
1	SVC	0.807711	0.777900	0.714008	0.818587	0.762729
2	Random Forest CLf	0.998595	0.832696	0.800649	0.855460	0.827147
3	Xgboost Clf	0.914620	0.831334	0.792996	0.858828	0.824600

- **From the above table we can find that Random forest classifier perform best among all these models.**

CHALLENGES FACED

- **The data was huge and was to be handled keeping in mind that we do not miss anything which is even of a little relevance.**
- **Computation time.**
- **Getting a higher accuracy on the models.**
- **Carefully handling feature imbalanced data.**
- **Tuning of hyper parameters carefully.**
- **Feature engineering**

CONCLUSION

- **Random Forest model has the highest recall, if the business cares recall the most, then this model is the best candidate. Since Random Forest and XGBoost model has slight difference between their Recall value, F1-Score, Precision Score.**
- **Random Forest model gave the highest accuracy of 83.79% on test dataset. Repayment in the month of September tended to be the most important feature for our machine learning model.**
- **The best accuracy is obtained from the Random Forest and XGBoost Classifier models.**
- **In general, all models have comparable accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading.**
- **From the table in the previous slide we can see that Random Forest Classifier having Recall, F1-score, and ROC Score values equals 85.85%, 83%, and 91% and XGBoost Classifier having Recall, F1-score, and ROC Score values equals 85.69%, 82%, and 90%.**
- **We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis.**

Thank You