Capstone Project

SUPERVISED ML- CLASSIFICATION

CREDIT CARD DEFAULT PREDICTION

BY
SARANYA N
Email :SNK4411@GMAIL.COM
STUDENT ALMABETTER

POINT FOR DISCUSSION

- PROBLEM STATEMENT
- INDRODUCTION
- DATA SUMMARY
- EDA
- DATA PREPROCESSING
- ML MODELS
- 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. Credit card debt results when a client of a credit card company purchases an item or service through the card system. Debt accumulates and increases via interest and penalties when the consumer does not pay the company for the money they have spent

DATA SUMMARY

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, 0,4,5,6 = others)
MARRIAGE: Marital status (0 = others, 1 = married, 2 = single, 3 = others)
AGE: Age in years
PAY_1: Repayment status in September, 2005 (scale same as above)
PAY_2: Repayment status in August, 2005 (scale same as above)
PAY_3: Repayment status in July, 2005 (scale same as above)
PAY_4: Repayment status in June, 2005 (scale same as above)
PAY_5: Repayment status in May, 2005 (scale same as above)
PAY_6: Repayment status in April, 2005 (scale same as above)

DATA SUMMARY

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5: Amount of bill statement in May, 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_AMT3: Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
default.payment.next.month: Default payment (1=yes, 0=no)

INTRODUCTION

- The given dataset consist of 30000 rows and 25 columns, the columns description is :This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

OBJECTIVE

Objective of our project is to predict which customer might default in upcoming months. let's have a quick look on definition of what actually meant by Credit Card Default.

a small plastic card issued by a bank, building society, etc., allowing the holder to purchase goods or services on credit. We are all aware what is credit card. It is type of payment card in which charges are made against a line of credit instead of the account holder's cash deposits. When someone uses a credit card to make a purchase, that person's account accrues a balance that must be paid off each month.

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.

DATA PREPARATION

```
[ ]  #checking missing value count
     credit_df.isnull().sum()
```

```
ID                          0
LIMIT_BAL                   0
SEX                         0
EDUCATION                   0
MARRIAGE                    0
AGE                         0
PAY_0                       0
PAY_2                       0
PAY_3                       0
PAY_4                       0
PAY_5                       0
PAY_6                       0
BILL_AMT1                   0
BILL_AMT2                   0
BILL_AMT3                   0
BILL_AMT4                   0
BILL_AMT5                   0
BILL_AMT6                   0
PAY_AMT1                    0
PAY_AMT2                    0
PAY_AMT3                    0
PAY_AMT4                    0
PAY_AMT5                    0
PAY_AMT6                    0
default payment next month  0
dtype: int64
```

No null values

No duplicate values

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is done with the help of a package of python by data visualization and their interpretations. In the given datasets, there are 25 feature columns and 30000 rows. Its distribution can be classified based on the following criteria

- Analysis of dependent variable
- Analysis of independent variables

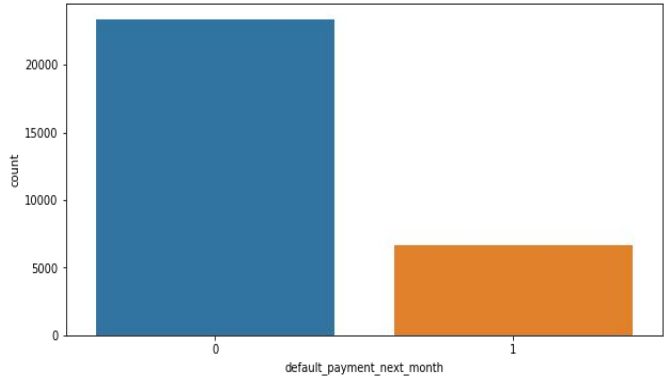
Analysis Of Dependent Variable

We can see class imbalance

Present,

0- not default

1- default



From the graph we can understand that:

About 22% people are expected to default next month and 77.8% are not expected to default.

Defaulters are less than non defaulters in the given dataset . so both the classes are not in proportion so we have to work on that to normalize the dataset in our dataset

Analysis Of Independent Variables

We have few categorical features in dataset

- SEX
- EDUCATION
- MARRIAGE
- AGE
- LIMIT BALANCE
- PAYMENT STATUS
- BILL AMOUNT
- PAYMENT AMOUNT

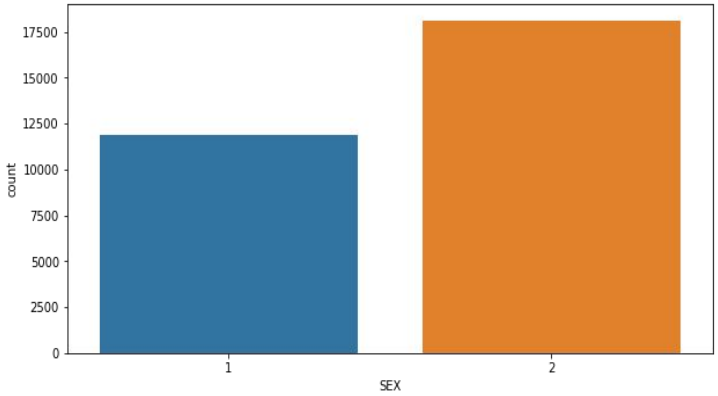
ANALYSIS OF SEX VARIABLE

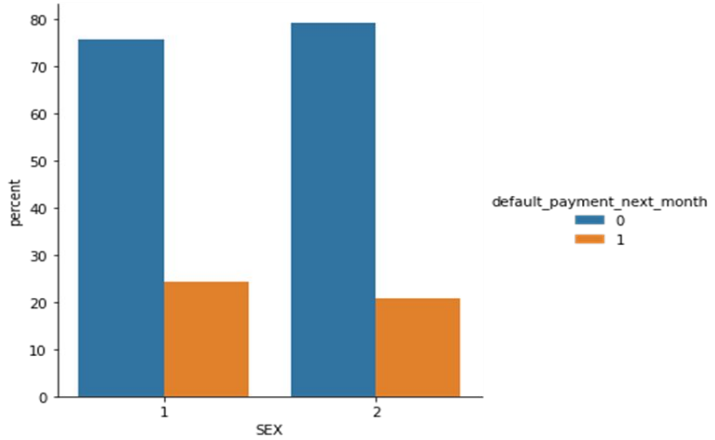
1-MALE
2-FEMALE

In the given dataset, there is a total of 30000 rows of data, of which 18112 are related to females and 11888 data's are belong to male.

Here we can understand that most of the credit users are females i.e., 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 On Education Variable

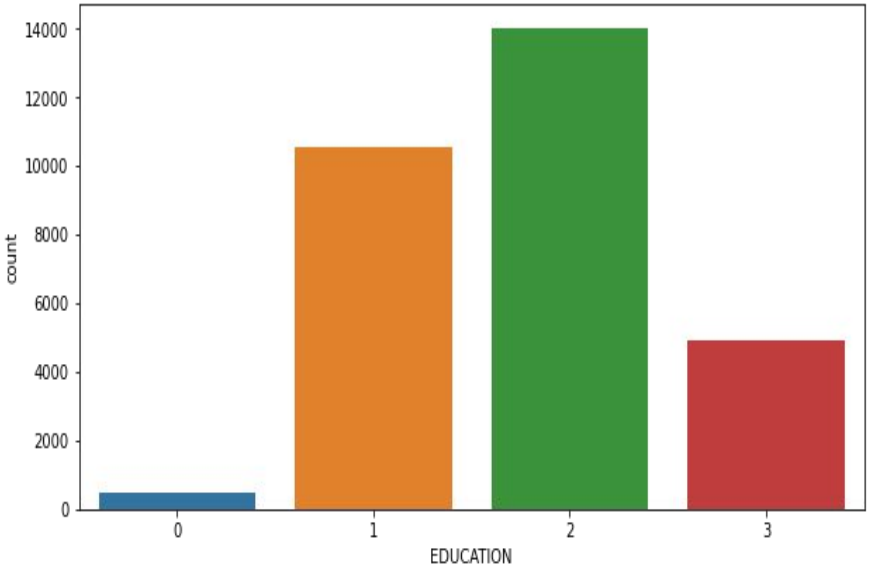
1 = GRADUATE SCHOOL

2 = UNIVERSITY

3 = HIGH SCHOOL

0 = OTHERS

- From the data analysis we can say that More number of credit holders are university students i.e. around 47 percentage followed by Graduates and then High school students

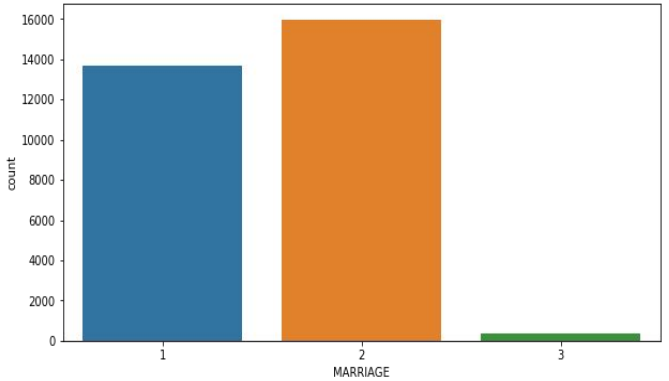


Analysis Of Marriage Variable

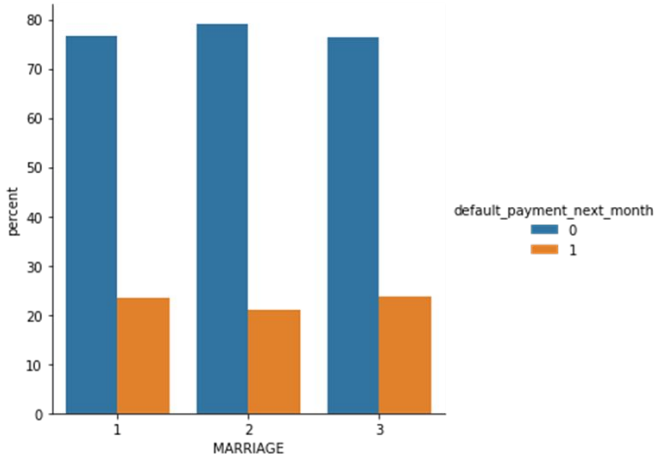
1 = MARRIED

2 = SINGLE

3 = OTHERS

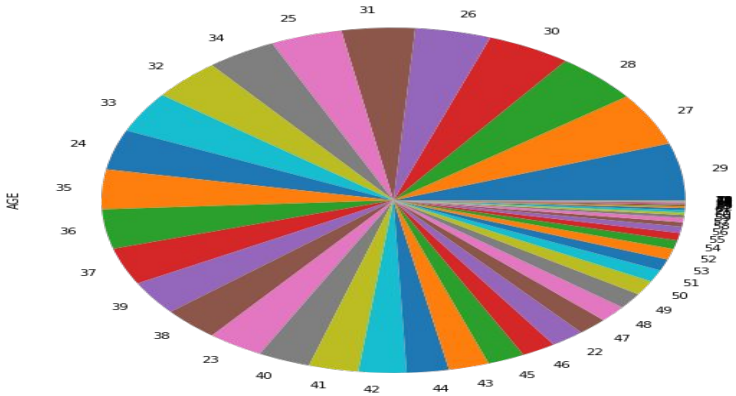


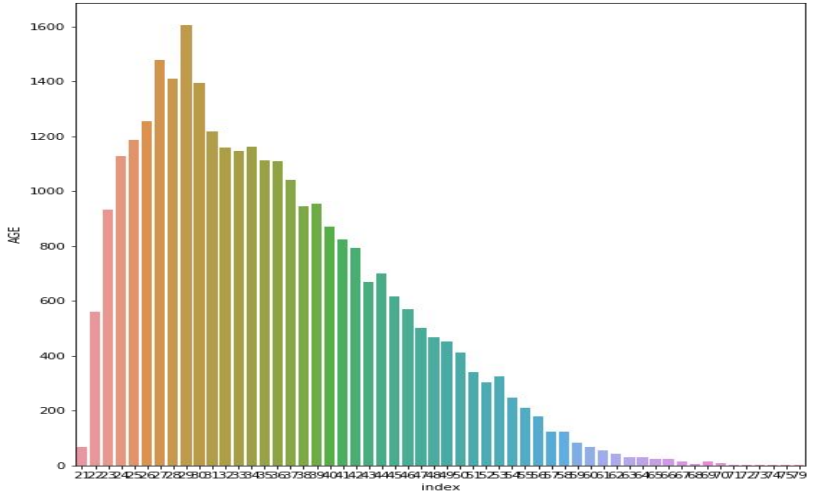
- From the above plot we can understand that most of the credit card users are single
- High defaulter rate when it comes to other



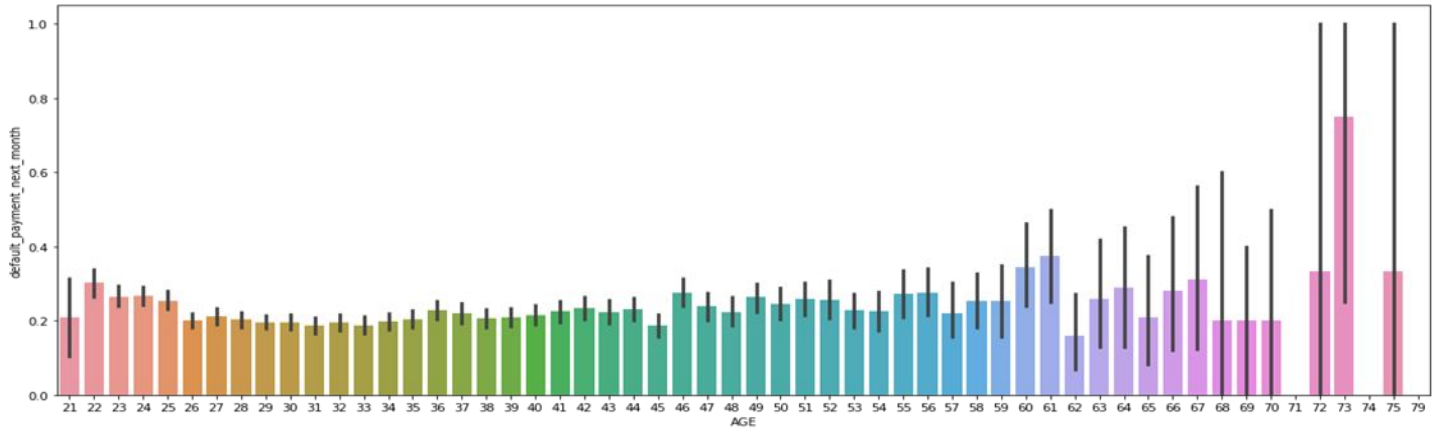
Analysis Of Age Variable

- Plotting graph according to age of the users irrespective of gender





Analysis of Age variable

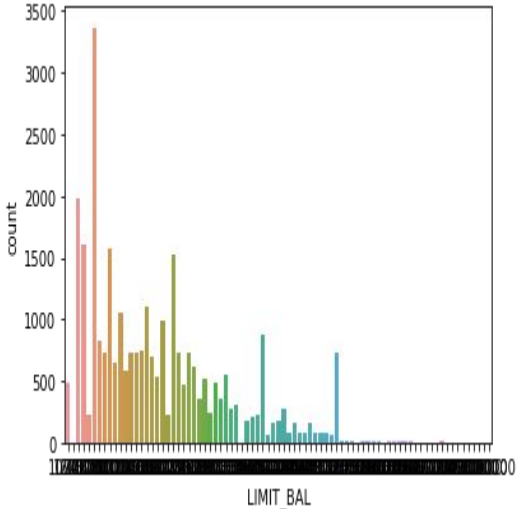


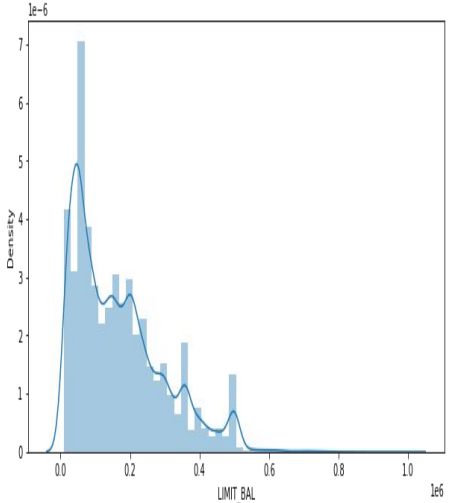
From the above plot analysis we can say that

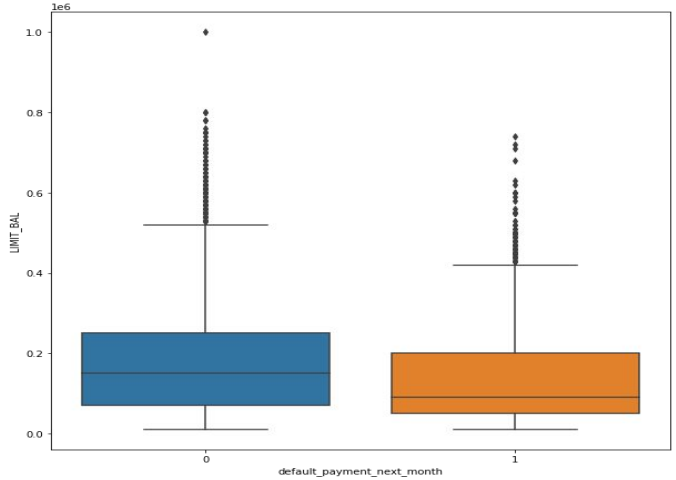
- We can see more number of credit cards holder age are between 26-30 years old.
- Age above 60 years old rarely uses the credit card.

Analysis Of Limit Balance Variable

Limit balance is the range of amount, given to a customer by a lending / financial institution. Also the credit limit is vary from person to person .



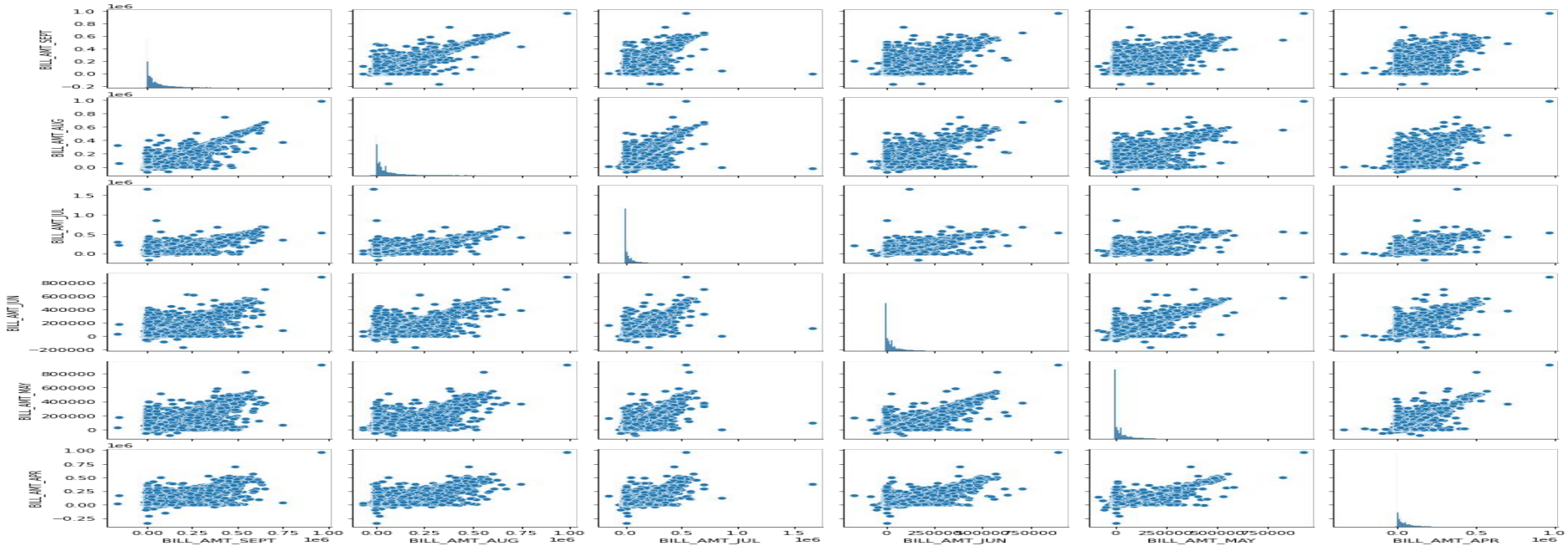




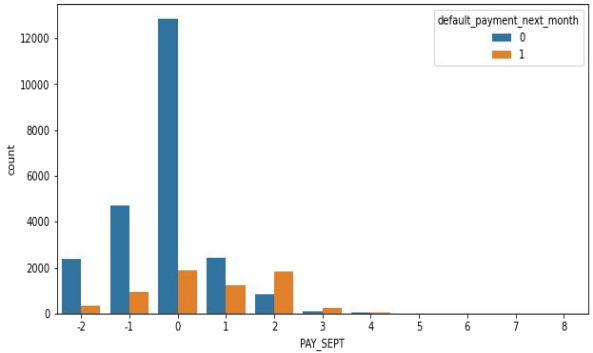
From the above plots we can say that maximum amount of given credit in NT dollars is 50,000 .

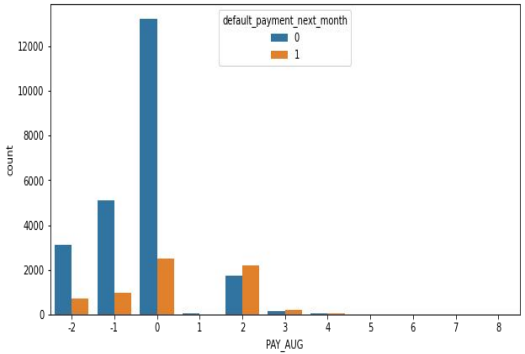
Analysis of bill amount

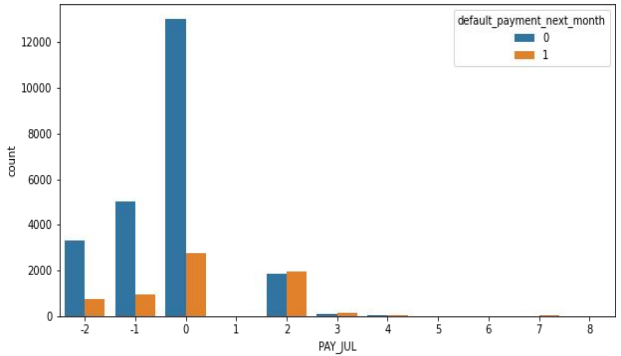
- Bill amount is the amount, due to be paid on the due date.

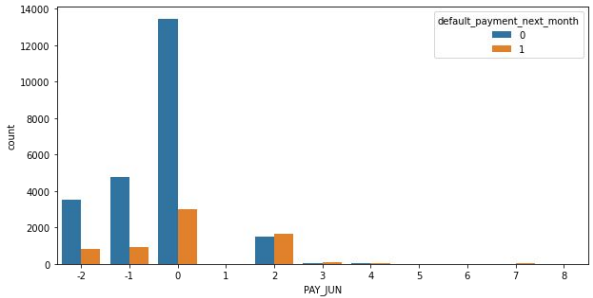


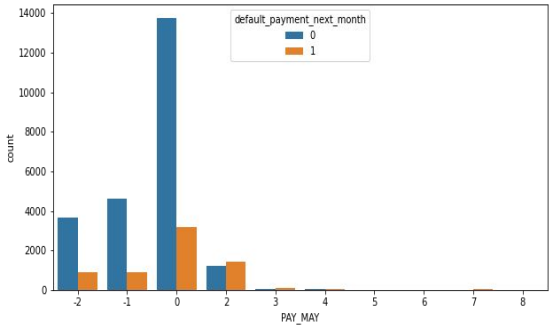
Analysis Of Payment Status

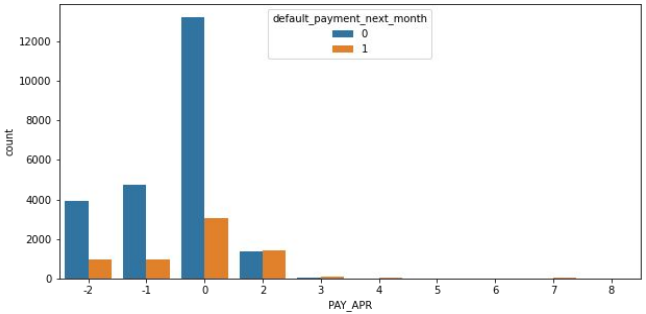






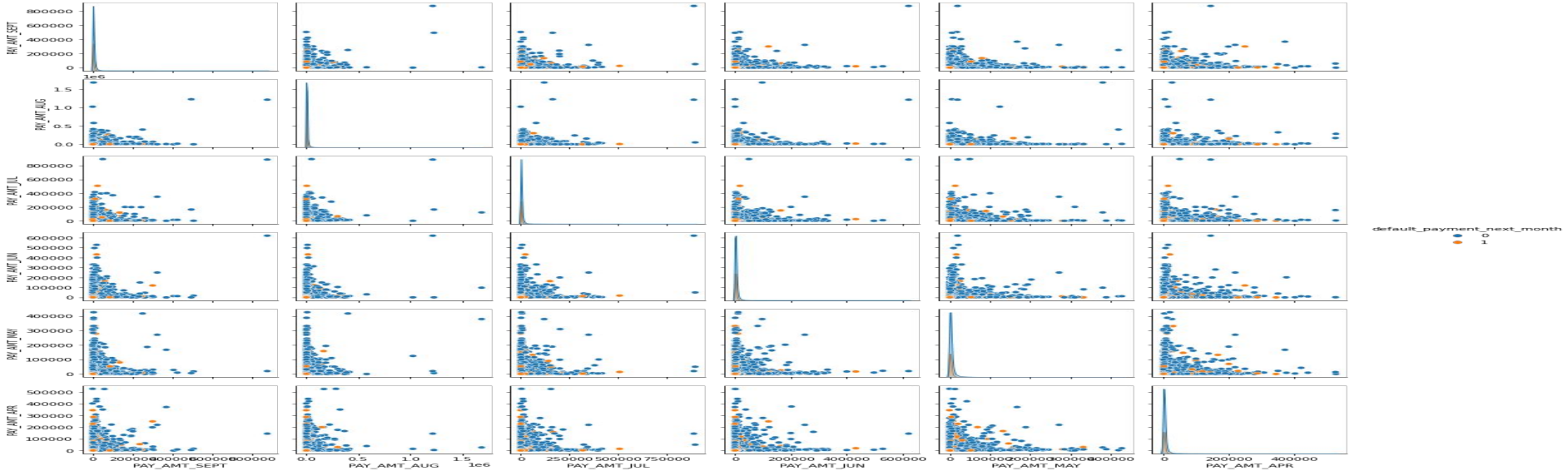






Payment status is the status of payment, which is due to be paid on the due date and paid by the customer.

Analysis Of Paid Amount Status



Amount payment is the amount of money paid by the customer on the due date.

INSIGHTS- EDA

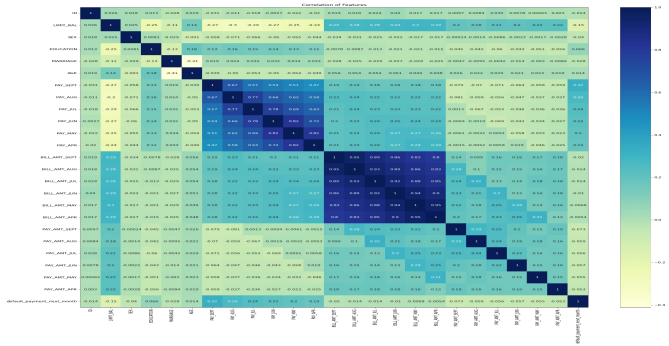
- About 22% people are expected to default next month and 77.8% are not expected to default.
- Defaulters are less than non defaulters in the given dataset . so both the classes are not in proportion so we have to work on that to normalize the dataset in our dataset.
- most of the credit users are females i.e., Number of Male credit holder is less than Female.
- More number of credit holders are university students i.e. around 47 percentage followed by Graduates and then High school students.
- most of the credit card users are single. High defaulter rate when it comes to other
- 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.
- From the above plots we can say that maximum amount of given credit in NT dollars is 50,000 .

Correlation Relationship

- Correlation is a mutual relationship or connection between two or more features.

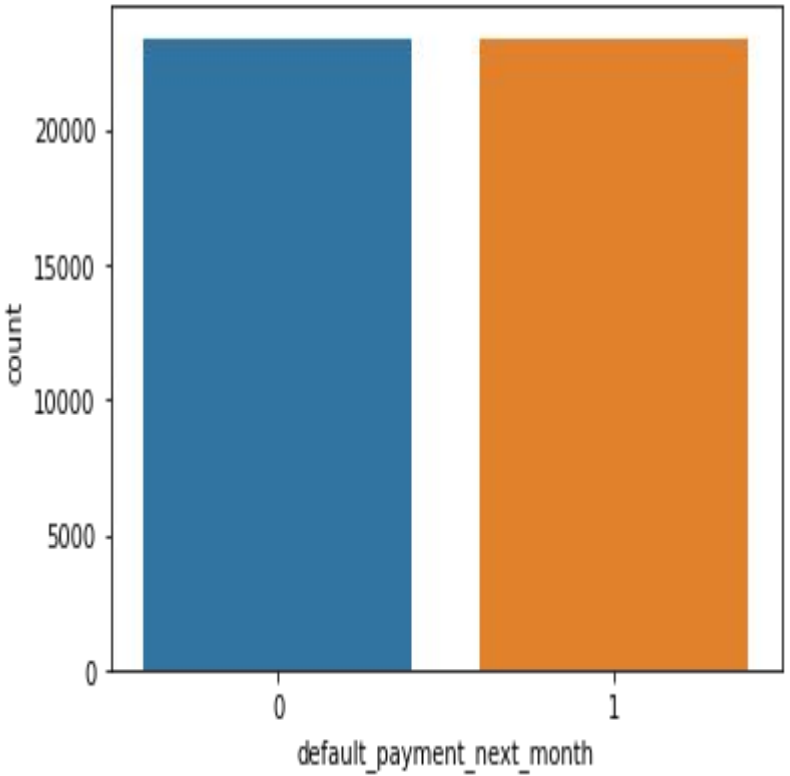
- From this correlation heat map there we can see that some of the values are negatively correlated . like age , but we can't delete age because it is an important feature for further prediction

- The column ID also have some negative values it is not that much important for analysis , so we can remove it



SMOTE

- SMOTE (synthetic minority oversampling technique) 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.
- SMOTE synthesis new minority instances between existing minority
- instances. It generates the virtual training records by linear
- interpolation for the minority class.
-
- The Original dataset had 30000 data points but after resampled dataset using SMOTE the data points increased to 46728.
-
- 0: 23364, 1: 23364 we can see that now both class are balance



One Hot Encoding

- Creating dummy variables for categorical variables using one hot encoding.
- A one hot encoding allows the representation of categorical data to be more expressive.
- Many machine learning algorithms cannot work with categorical data directly.
- This is required for both input and output variables that are categorical
- 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_Payment_Value Dues EDUCATION_graduate school EDUCATION_high school EDUCATION_others EDUCATION_university MARRIAGE_married MARRIAGE_others MARRIAGE_single ... 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

Model Building

Model training is the process of fitting data into the required model after completing the data preprocessing part.
In this project , training 4 models , which are,

- LOGISTIC REGRESSION
- RANDOM FOREST
- SVM
- XGBOOST

Logistic Regression

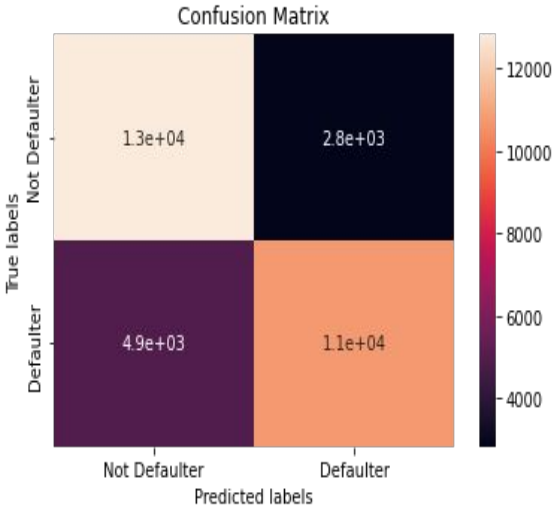
- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.
- It is used for predicting the categorical dependent variable using a given set of independent variables.
- From this regression model we get the results as below:

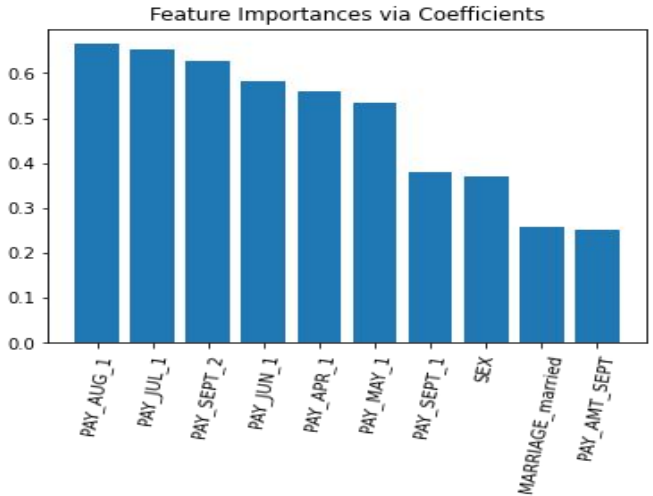
- **The accuracy on train data is 0.7525473536269843**
- **The accuracy on test data is 0.7536476233707282**
- ***The accuracy on test data is 0.7536476233707282***
- ***The precision on test data is 0.6894941634241245***
- ***The recall on test data is 0.790953727123939 9***
- ***The f1 on test data is 0.736747280160765***
- ***The roc_score on test data is 0.7578906566654182***

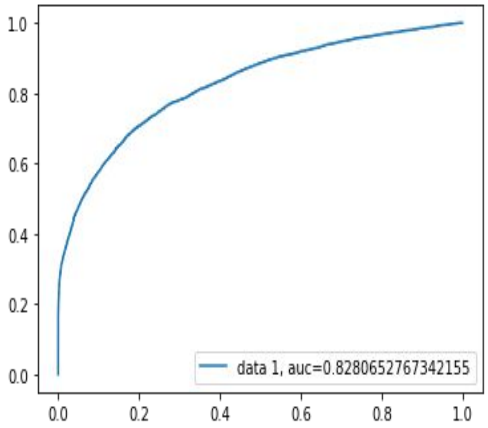
CONFUSION MATRIX

FEATURE IMPORTANCES

ROC AUC CURVE





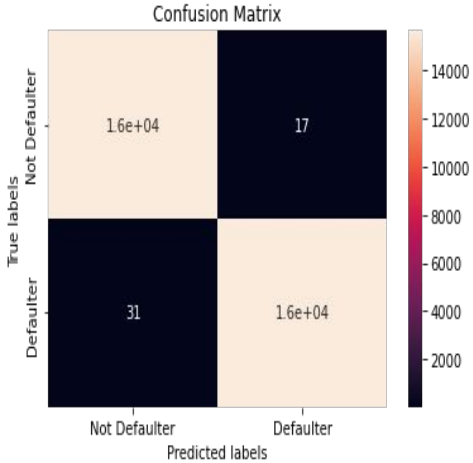


Random Forest

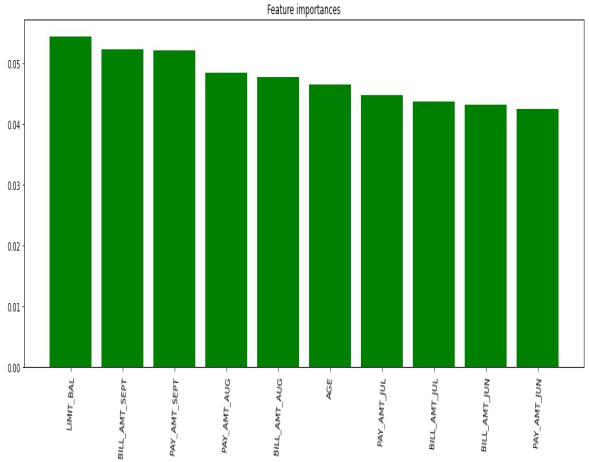
- A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

- *The accuracy on train data is 0.9993292234963427*
- *The accuracy on test data is 0.8377537124700084*
- *The accuracy on test data is 0.8377537124700084*
- *The precision on test data is 0.8075226977950714*
- *The recall on test data is 0.8594699061292104*
- *The f1 on test data is 0.8326869065133075*
- *The roc_score on test data is 0.8389926270281615*

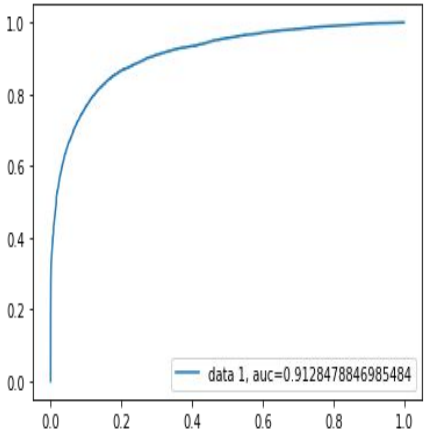
CONFUSION MATRIX



FEATURE IMPORTANCES



ROC AUC CURVE

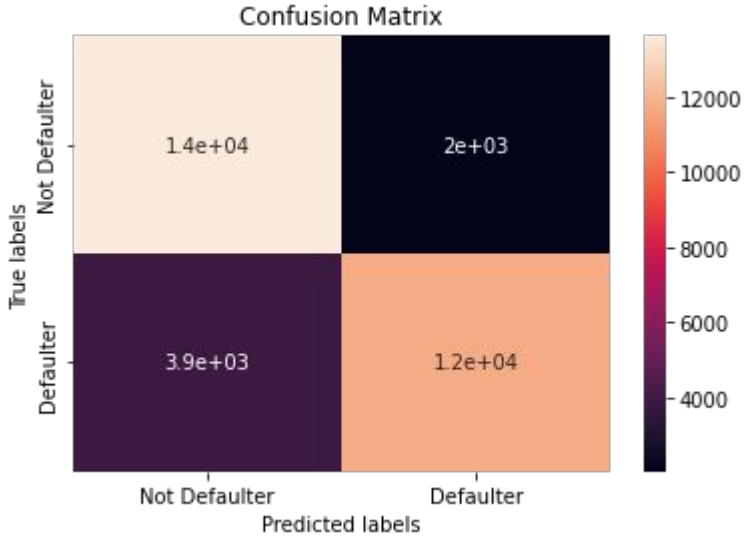


Support Vector Machine

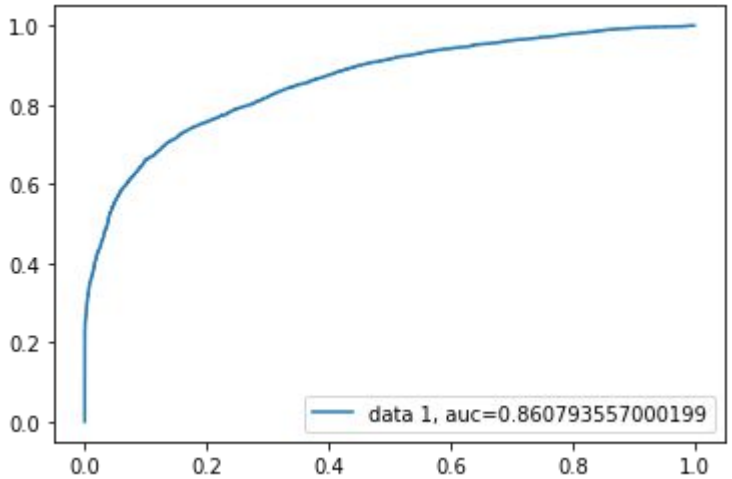
- The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function. The kernel method can not be changed in linear SVC, because it is based on the kernel linear method

- ***The accuracy on train data is 0.7525473536269843***
- ***The accuracy on test data is 0.7536476233707282***
- ***The accuracy on test data is 0.7847091628299072***
- ***The precision on test data is 0.72905317769131***
- ***The recall on test data is 0.8203444249854057***
- ***The f1 on test data is 0.7720093393764592***
- ***The roc_score on test data is 0.7882793428463031***

CONFUSION MATRIX



ROC AUC CURVE

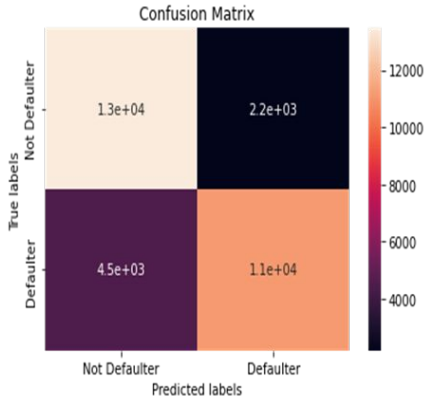


XG BOOST

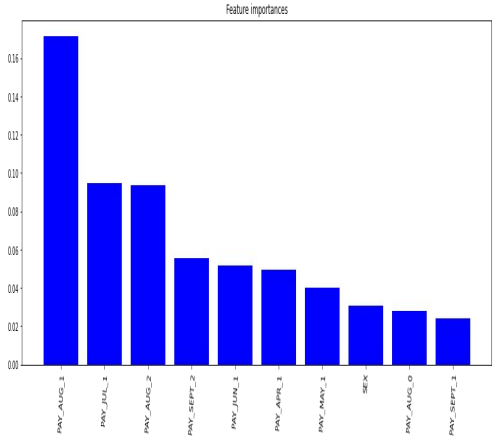
- **XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.**

- ***The accuracy on train data is 0.7859264701185038***
- ***The accuracy on test data is 0.7764087932040724***
- ***The accuracy on test data is 0.7764087932040724***
- ***The precision on test data is 0.7083009079118029***
- ***The recall on test data is 0.81996996996997***
- ***The f1 on test data is 0.7600556715379263***
- ***The roc_score on train data is 0.7816320572370111***

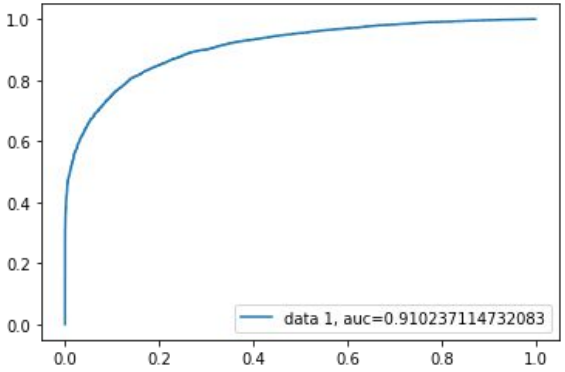
CONFUSION MATRIX



FEATURE IMPORTANCES



ROC AUC CURVE

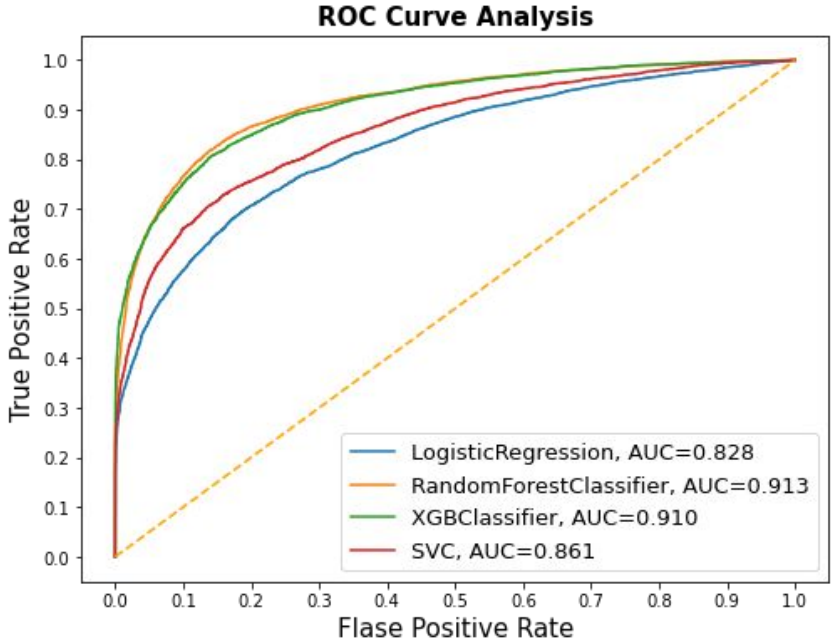


ROC AND AUC OF ALL MODELS

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:

True Positive Rate
False Positive Rate

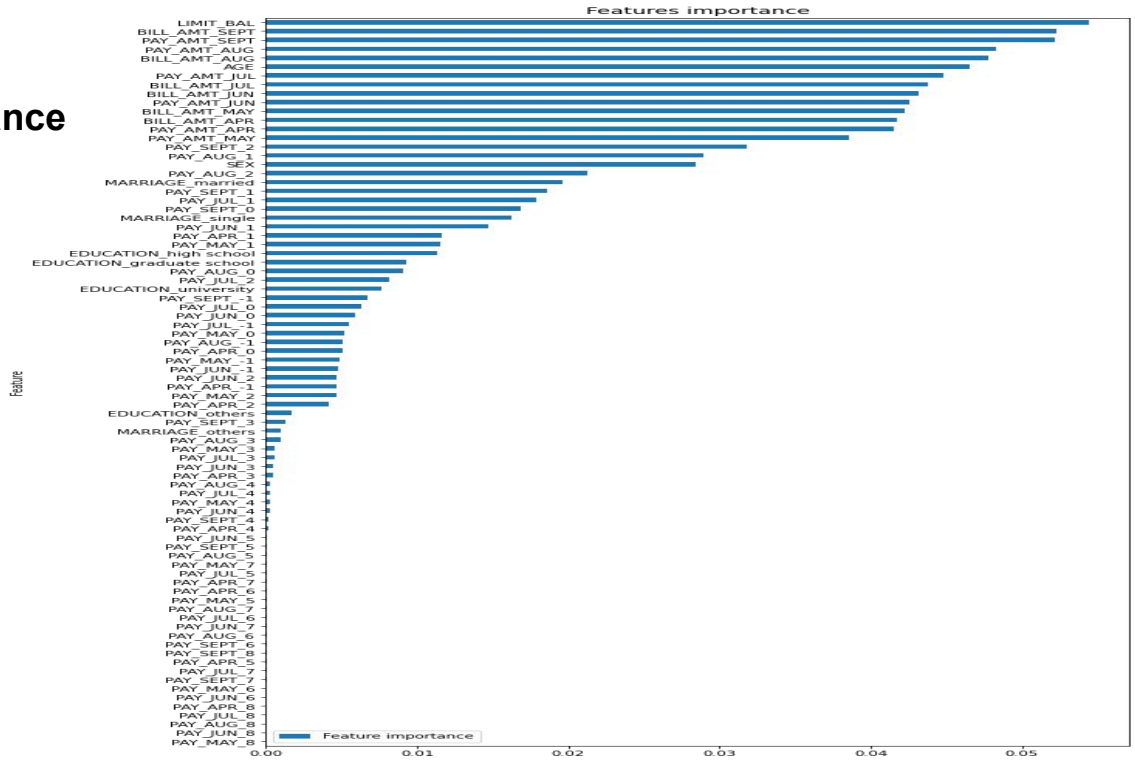
AUC stands for "Area under the ROC Curve. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example



FEATURE IMPORTANCE

Credit limit has given highest importance comparatively

LIMT_BAL","BILL_AMT_SEPT" AND "PAY_AMT_SEPT" are the most recent 2 months' payment status and they are the strongest predictors of future payment default risk.



MODEL EVALUATION AND SELECTION

- Model evaluation and model selection are important parts of a data science project.
- Hyperparameter tuning in the model building helps to improve model performance and can select the best among them. The same methodology has been followed in this case.

	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.75	0.75	0.69	0.79	0.74
1	SVC	0.81	0.78	0.73	0.82	0.77
2	Random Forest CLf	1.00	0.84	0.81	0.86	0.83
3	Xgboost Clf	0.91	0.83	0.79	0.86	0.82

CHALLENGES

- Understanding the columns.
- Understanding the data because 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

- We have predicted the defaulters using multiple models in this project. We have used Logistic regression, Random forest, XGBoost, SVM. We have also used GridSearchCV to tune hyper parameters.

- 1. In conclusion , we can conclude that boost is the best model for the present problem.

- 2. We have predicted the defaulters using multiple models in this project. We have used Logistical regression, Decision Tree, XGBoost, SVM. We have also used GridSearchCV to tune hyperparameters. Logistic Regression, xgboost,svm, Random Forest algorithms were implemented. The important metric to compare all the algorithms in this case is ‘Recall’. As the company can’t afford to predict False negative i.e. predict defaulter as a non defaulter. Since, company is one, who will give to money to the customers, if, for any reason giving money to defaulter is gaining more risk to getting the investment back. Hence, here identifying false negative is important.

- 3. We have also seen the class imbalance so we did SMOTE to handle imbalance.
- 4. We did train test split and stratify the target variable.
- 5. We conclude that out of all models XGBoost performed well with ROC AUC score of 0.832.
- 6. The best accuracy is obtained for the Random forest and XGBoost classifier.
- 7. XGBoost model solves the problem with high accuracy than others. It has the precision, recall ,F!-score and ROC Score scores of 79% , 85% , 82% and 83% respectively, which is the highest among other models.
- 8. The next best models are the Random Forest classifier , it has the precision, recall ,F!-score and ROC Score scores of 80% , 85% , 83% and 83% respectively


Thank
you!!