**Abstract:**

Bank of recent years played a significant role in the development of the nation. The bank offers a few things that are directly dependent on any nation's general economic and financial condition. Banking efficiency leads to the business, growth in the industry, economic growth, and support for the common man with savings, improving financial security. This analysis has the function of forecasting the inability to pay the bank loan. The study found more than 10 million Bank of Taiwan records. Analysis of the logistic regression hits the relation between the class variable and the set of independent variables. The primary analysis produces exploratory views of data correctly. Further, this paper used ML algorithms to get predictions with accuracy to detect the default users based on transactional data.

This model will help commercial banks,financial organizations, loan institutes, and other decision-makers to predict the loan defaulter earlier.

**Introduction:**

Bank of current years performed a great position in the improvement of the nation. The financial institution affords a few matters that are immediately based on any nation's widespread financial and economic condition. Banking effectivity leads to the business, increase in the industry, financial growth, and assist for the frequent man with savings, enhancing economic security. This evaluation has the feature of forecasting the incapacity to pay the financial institution loan. The learn about discovered greater than 10 million Bank of Taiwan records. Analysis of the logistic regression hits the relation between the type variable and the set of impartial variables. The most important evaluation produces exploratory views of information correctly. Further, this paper used ML algorithms to get predictions with accuracy to discover the default customers based totally on transactional data.  
This mannequin will assist industrial banks,financial organizations, mortgage institutes, and different decision-makers to predict the mortgage defaulter earlier.

**Problem definition:**

Bank Loan has been one of the fastest-growing financial services banks in recent years. However, with the increasing number of bank loan users, banks face an ever-increasing rate of bank loan decline. This program is offered primarily to a person or company of higher value than another. Under this scheme, a small amount can be provided as a cash transfer or electronic transfer to the debtor when they can be in demand. Few of them have not returned a set amount in time, so sometimes they do not. This situation creates a problem for the bank. Then with the help of historical data, the need to predict bank loan errors can be determined. As such, machine learning may offer options for addressing the current issue and handling credit risk.

**Data Description:**

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.

**Attribute Information:**

* 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 years
* 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\_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)
* 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).

**Pre-processing**

1. Finding Missing value

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data. Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present. This dataset having zero missing values. Hence no pre-processing is required.

1. Looking at Duplicate entries

Many would probably not immediately realize the importance of finding and removing duplicate data in company records. Others might consider the process a total waste of time. However, it is important to realize that duplicate data can create chaos that might, eventually, cost your business a considerable amount of money.

Total 35 entries are duplicate in given dataset. Hence these 35 rows are removed from dataset.

**Exploratory Data Analysis**

The primary objective of exploratory data analysis in order to perform exploratory data analysis is to uncover the underlying structure. The structure of the various data sets determines the trends, patterns, and relationships among them. A business cannot conclude or draw assumptions from a huge quantity of data and rather requires taking an exhaustive look at the data set through an analytical lens. Therefore, performing an Exploratory Data Analysis allows data scientists to detect errors, debunk assumptions, and much more to ultimately select an appropriate predictive model. Following EDA have been implemented.

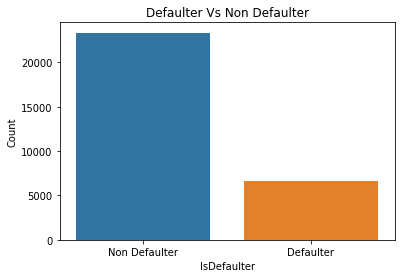


Figure 1 Defaulter vs Non defaulter

The comparison for the count of defaulter and non-defaulter can be observed from above graph. The non-defaulters are greater in number.

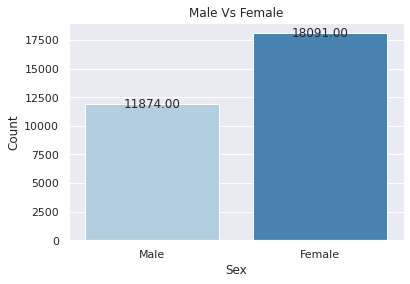


Figure 2 Credit card count vs Male/Female

Female users are having large count as compared to male.

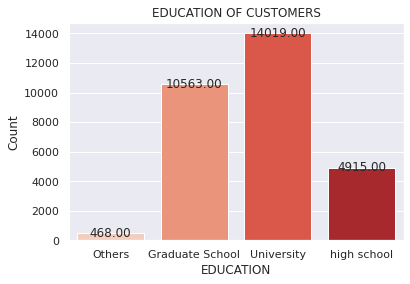


Figure 3 Education level vs card count

There are four levels of education. The count of credit card users increases with increase in level of education.

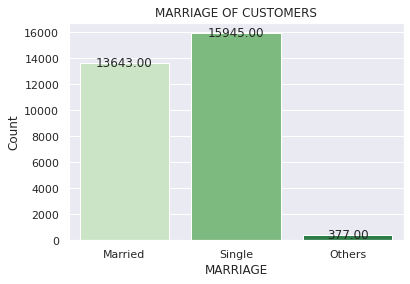


Figure 4 Marriage of customers vs card count

The marital status may affect the use of count of credit card users. The count for Single persons is more compared to married and other people.

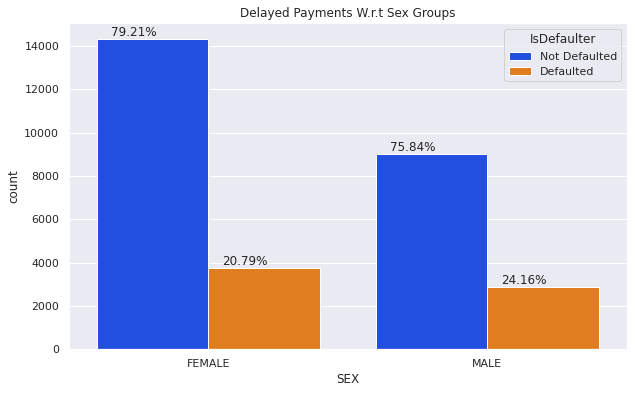


Figure 5 Delayed payment vs sex groups

Delayed payments can be compared on basis of gender. The female count is more than males.

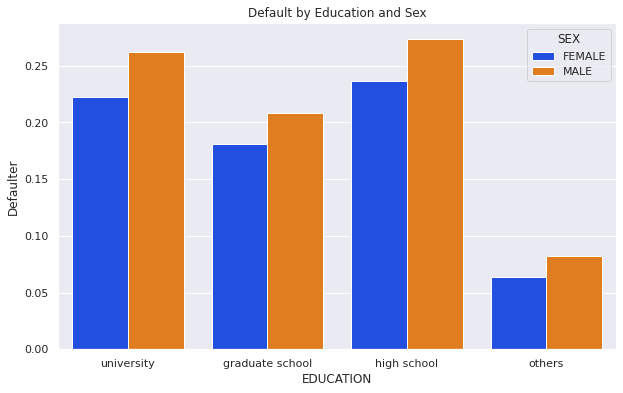


Figure 6 Default by education and sex

The above graph shows comparison between education and sex with defaulter category.

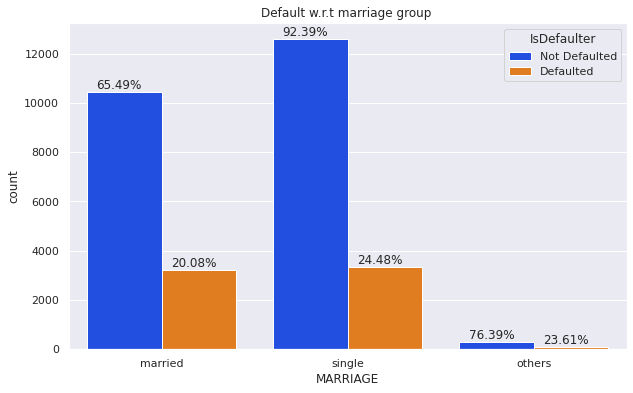


Figure 7 Default vs marriage group

The comparison of marital status vs defaulter or non-defaulter.

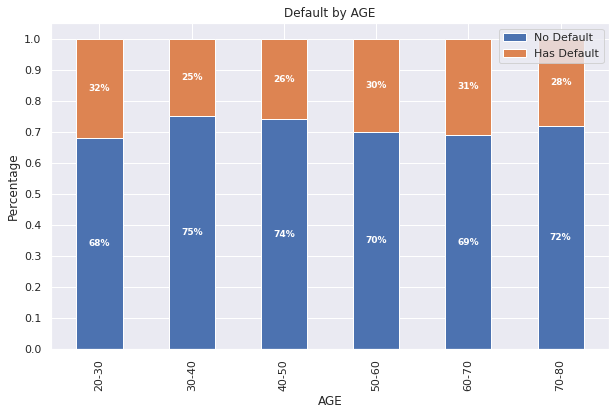


Figure 8 Defaults by Age

The count of defaulter and non-defaulter can be studied from different age group. The age group is divided into different bins and then analysed.

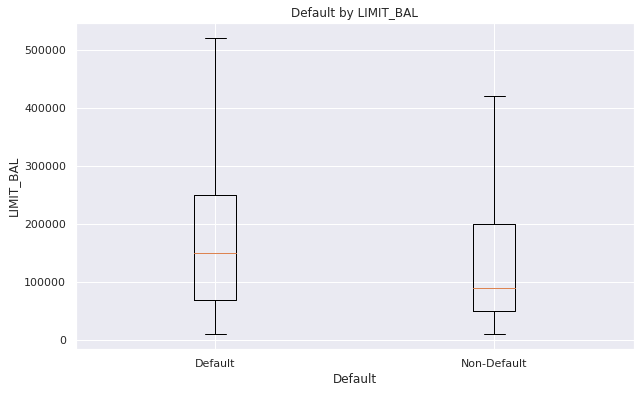


Figure 9 Defaults by Limit balance

The box plot is used to study the comparison for defaulter count vs credit limit. Customers with high credit limits tend to have higher 'no-default' rate.

Figure 10, The limit amount vs density graph shows rightly skewed. For the further processing this graph need to be adjusted.

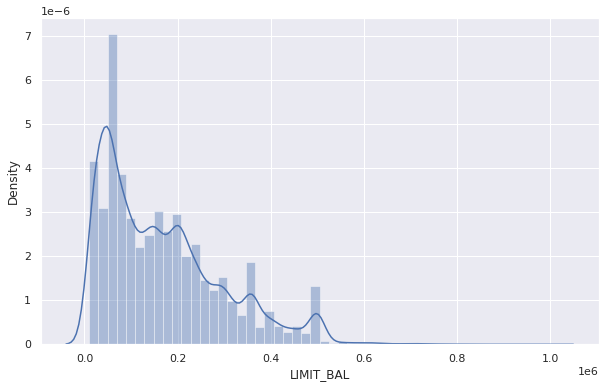


Figure 10 Density vs Limit balance

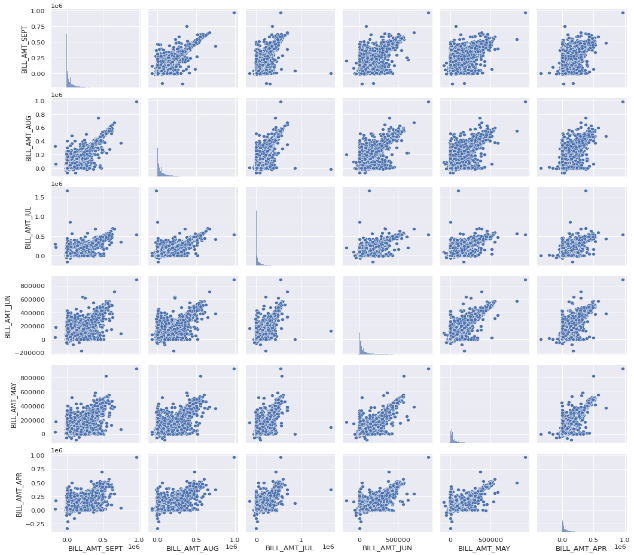


Figure 11 Month wise analysis of balance limit

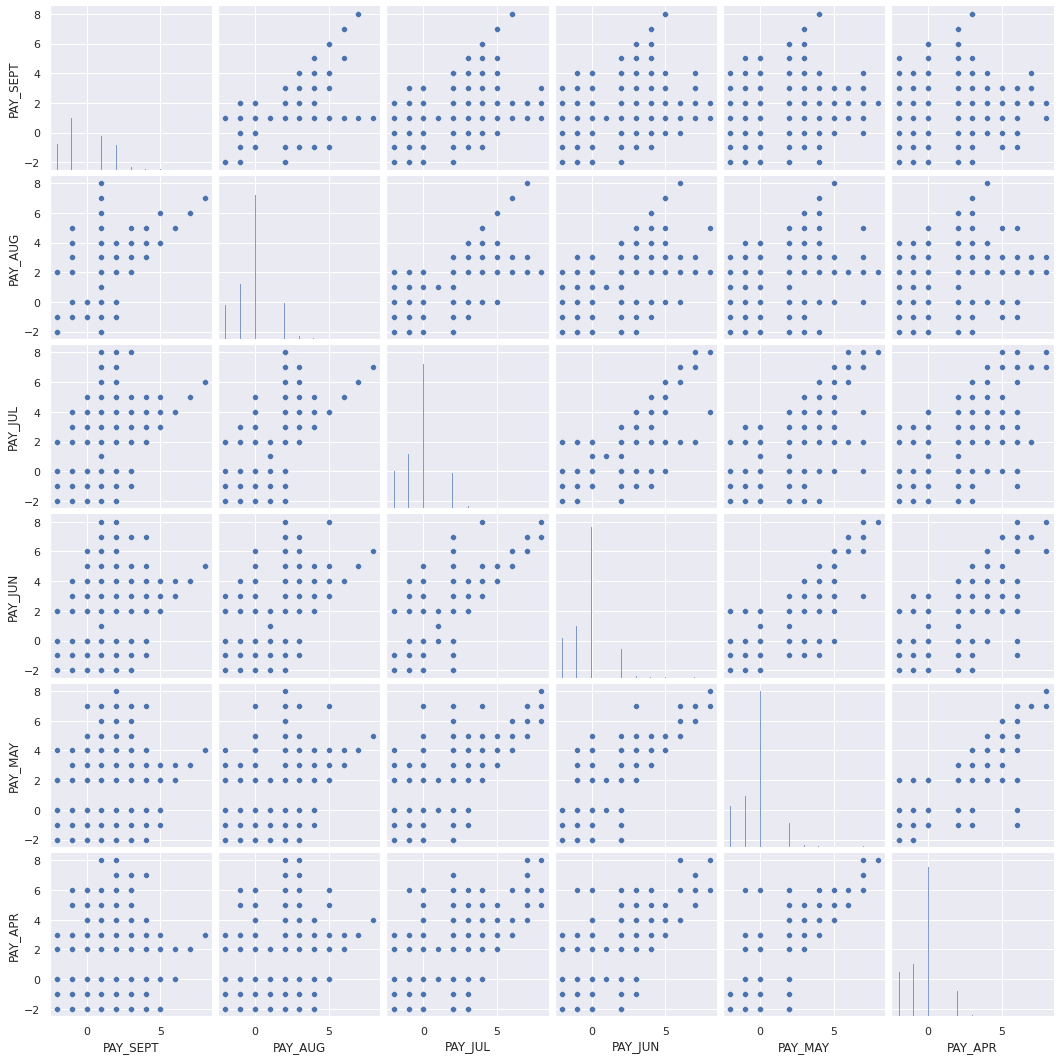


Figure 12 Month wise status of payment

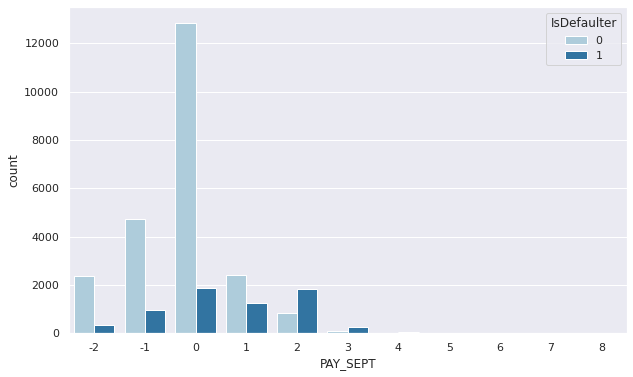


Figure 13 Counter plot for month vs due payment

Above figure shows the counter plot for every month and this will give analysis of due payment.

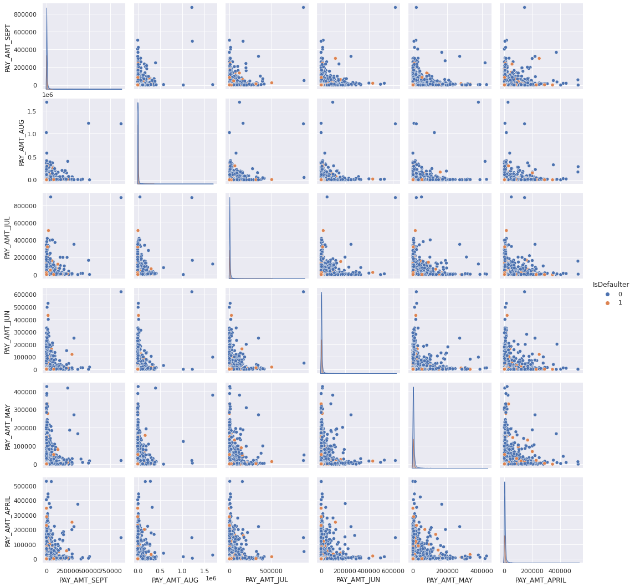


Figure 14 Already paid amounts

The above plot shows month wise analysis of already paid customers count.

**Feature Engineering**

Imbalanced Dataset of Target Variable

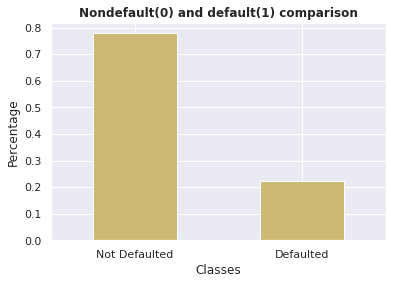


Figure 15 Percentage of Non default vs defaulters

The percentage graph analysis non defaulters vs defaulters is shown in figure 15.

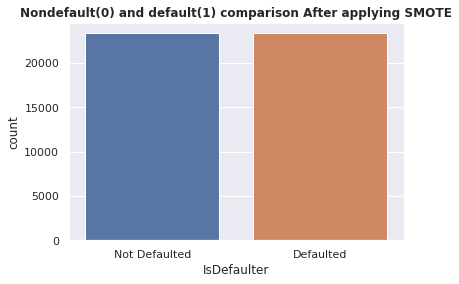


Figure 16 Comparing both defaulters and non-defaulters after applying SMOT

After applying SMOT technique to make both entries count equal. This will help well-functioning of classifier.

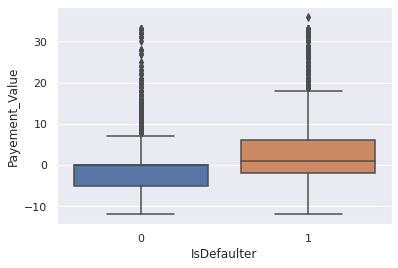


Figure 17 Payment value vs Is Defaulter

This figure explains the new feature generated.

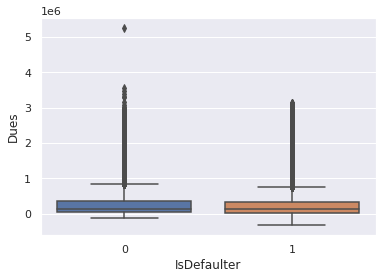


Figure 18 Due feature vs Is Counter

The Due feature is generated with balanced payments of customers. The label encoding and one hot encoding part is performed under feature engineering for this dataset.

**Correlation Heatmap**

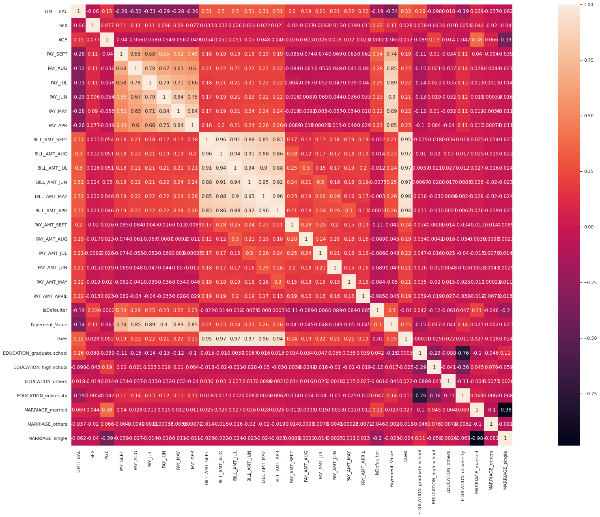
****

Figure 19 Correlation Heatmap

This heatmap graph gives correlation values between two feature vectors.

The dependent and independent variables are separated. After that rescaling values to standard values.

**Model Training and Hyper Parameter Tuning**

Train-Test-Split method have been implemented for applying classification algorithm.

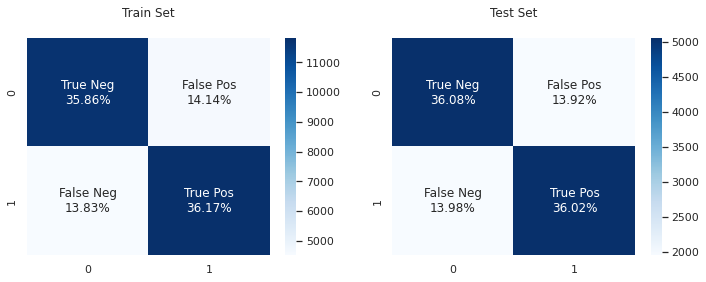


Figure 20 Logistic Regression model performance

The figure 20 gives the performance of the logistic regression model.

Hyperparameter tuning of Logistic Regression

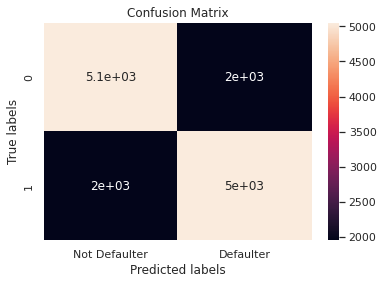


Figure 21 Confusion matrix

The confusion matrix gives details of true positive rate, classification accuracy, and precision, recall. A confusion matrix is the summary of prediction results. A confusion matrix prints the correct and incorrect values in number count. It helps us for a good Data Visualization It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being

Feature Importance of Logistic Regression

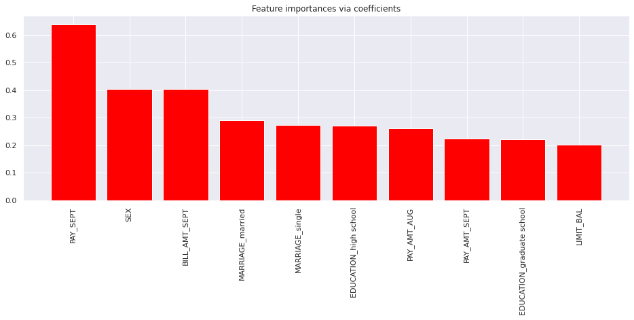


Figure 22 Feature Importance

All the features are not that much important. Hence these features are removed from future processing.

**Random Forest Classifier**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

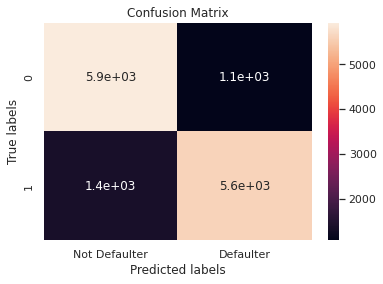


Figure 23 Confusion Matrix of random forest

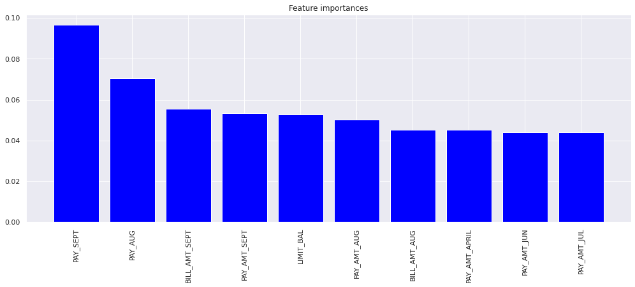


Figure 24 Feature importance for RF

**XGBoost**

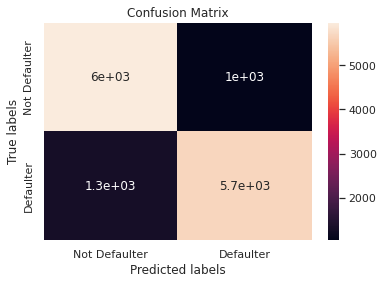


Figure 25 Confusion matrix for XG Boost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

It’s vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.

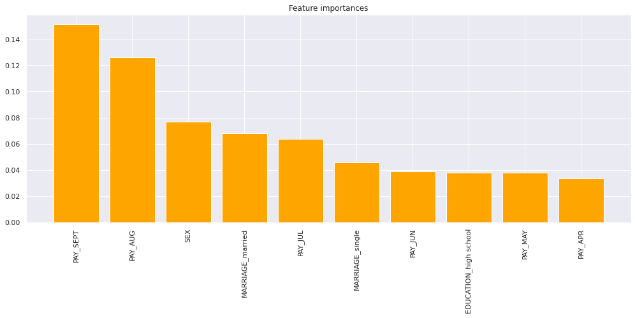
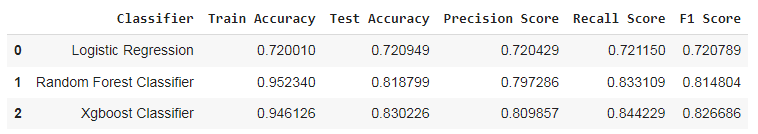


Figure 26 Feature Selection for XGBoost

**Evaluating the models**



Comparison of classifiers train and Test dataset.

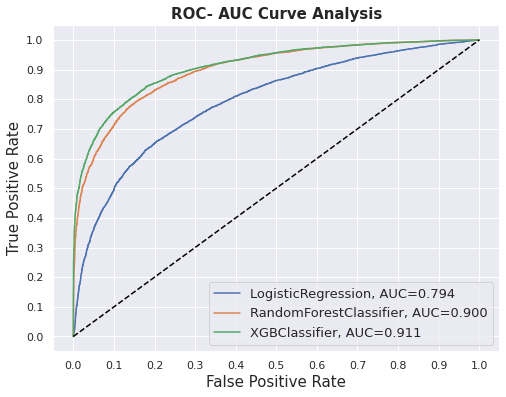


Figure 27 ROC curve for classifier

**Conclusion**

To quit stage via correctly constructing a mannequin to predict whether or not the purchaser will default his / her payment. We have carried out characteristic engineering, characteristic selection, hyperparameter tuning to forestall overfitting and for lowering error. The recall is the measure of our mannequin efficiently figuring out True Positives. Thus, for all the Customers who default, recall tells us how many we efficaciously recognized as default. As we had regarded recall, XGBoost is our nice mannequin as we can see the roc- aoc curve is maximum.