Revanth Janapriyan

 PGP-DSBA ONLINE

FINANCE AND RISK ANALYTICS

Contents

[**EXECUTIVE SUMMARY** 2](#_Toc138367056)

[**INTRODUCTION** 2](#_Toc138367057)

[**DATA DESCRIPTION** 2](#_Toc138367058)

[**DATA SAMPLE** 4](#_Toc138367059)

[**EXPLORATORY DATA ANALYSIS** 5](#_Toc138367060)

[**MODEL BUILDING** 15](#_Toc138367061)

[**CONCLUSION** 20](#_Toc138367062)

## **EXECUTIVE SUMMARY**

The objective of this project is to create an Indian credit risk(default) model, using the data provided. There are two machine learning models – Logistic Regression and Random Forest classifier models have been used to predict the defaulter in this project.

## **INTRODUCTION**

Though the main objective of the project is to build an Indian credit risk(default) model, various steps have been taken before the same model. First, the given dataset was imported using pandas library and preprocessing steps such as missing value treatment, outlier treatment, descriptive analysis, graphical representations of the features are done. With all these preprocessing a better cleaner data is obtained and is used for training and testing the model built. Logistic Regression and Random Forest Classifier are the two models built and results are evaluated through metrics such as accuracy score, classification report, confusion matrix and roc\_auc curve.

## **DATA DESCRIPTION**

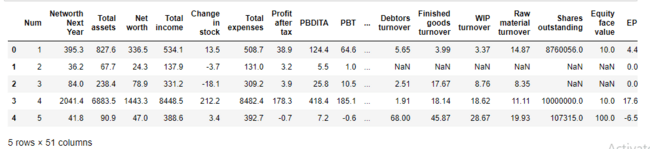
The below table shows the features present in the dataset and its description.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Networth Next Year | Net worth of the customer in next year |
| Total assets | Total assets of customer |
| Net worth | Net worth of the customer of present year |
| Total income | Total income of the customer |
| Change in stock | difference between value of current stock and the value of stock in last trading day |
| Total expenses | Total expense done by customer |
| Profit after tax | Profit after tax deduction |
| PBDITA | Profit before depreciation, income tax and amortization |
| PBT | Profit before tax deduction |
| Cash profit | Total Cash profit |
| PBDITA as % of total income | PBDITA / Total income |
| PBT as % of total income | PBT / Total income |
| PAT as % of total income | PAT / Total income |
| Cash profit as % of total income | Cash Profit / Total income |
| PAT as % of net worth | PAT / Net worth |
| Sales | Sales done by customer |
| Income from financial services | Income from financial services |
| Other income | Income from other sources |
| Total capital | Total capital of the customer |
| Reserves and funds | Total reserves and funds of the customer |
| Deposits (accepted by commercial banks) | All blank values |
| Borrowings | Total amount borrowed by customer |
| Current liabilities & provisions | current liabilities of the customer |
| Deferred tax liability | Future income tax customer will pay because of the current transaction |
| Shareholders funds | Amount of equity in a company, which is belong to shareholder |
| Cumulative retained profits | Total cumulative profit retained by customer |
| Capital employed | Current asset minus current liabilities |
| TOL/TNW | Total liabilities of the customer divided by Total net worth |
| Total term liabilities / tangible net worth | Short + long term liabilities divided by tangible net worth |
| Contingent liabilities / Net worth (%) | Contingent liabilities / Net worth |
| Contingent liabilities | Liabilities because of uncertain events |
| Net fixed assets | purchase price of all fixed assets |
| Investments | Total invested amount |
| Current assets | Assets that are expected to be converted to cash within a year |
| Net working capital | Difference of current liabilities and current assets |
| Quick ratio (times) | Total cash divided by current liabilities |
| Current ratio (times) | Current assets divided by current liabilities |
| Debt to equity ratio (times) | Total liabilities divided by its shareholder equity |
| Cash to current liabilities (times) | Total liquid cash divided by current liabilities |
| Cash to average cost of sales per day | Total cash divided by average cost of the sales |
| Creditors turnover | Net credit purchase divided to average trade creditors |
| Debtors turnover | Net credit sales divided by average accounts receivable |
| Finished goods turnover | Annual sales divided by average inventory |
| WIP turnover | The cost of goods sold for a period divided by the average inventory for that period |
| Raw material turnover | Cost of goods sold is divided by the average inventory for the same period |
| Shares outstanding | Number of issued shares minus the number of share held in the company |
| Equity face value | cost of the equity at the time of issuing |
| EPS | Net income divided by total number of outstanding share |
| Adjusted EPS | Adjusted net earning divided by the weighted average number of common share outstanding on a diluted basis during the plan year |
| Total liabilities | Sum of all type of liabilities |
| PE on BSE | Company current stock price divided by its earning per share |

There are 52 features in the given dataset.

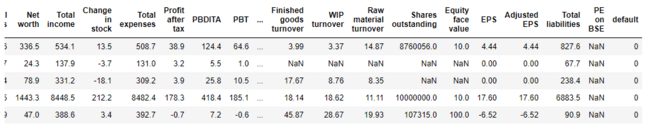
## DATA SAMPLE

The below table shows the first few rows of the given dataset.



**Table 1.1**

Table 1.2 below is the new dataset with ‘default’ feature added to it.



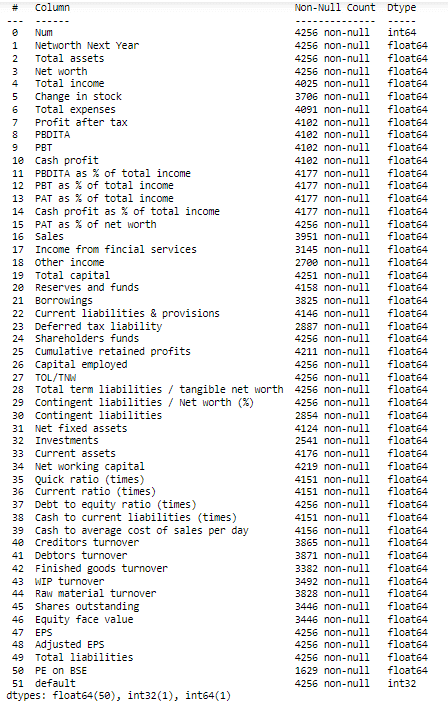
**Table 1.2**

Now with addition to the new variable ‘default’ to the dataset, there are total of 52 independent variables and one dependent variable.

## **EXPLORATORY DATA ANALYSIS**

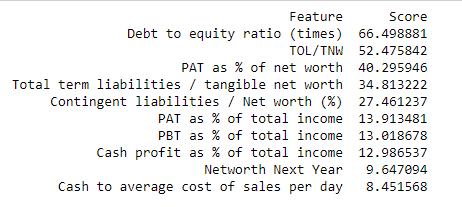
Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.

The below shows all the variables along with its null count and data type.



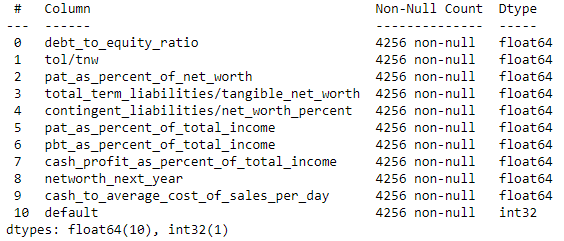
**Table 1.3**

As the first step of data exploration feature engineering is used to find 10 best features for the given dataset.



**Fig 1.1**

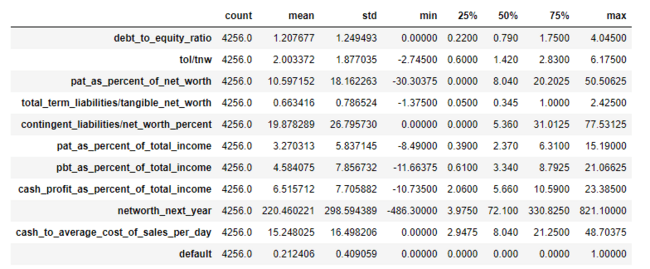
The selected 10 features and its data types are given below.



**Table 1.4**

There are 10 float variables and 1 integer variable is the feature selected dataset.

Let’s look into the descriptive stats of the new dataset.



**Table 1.5**

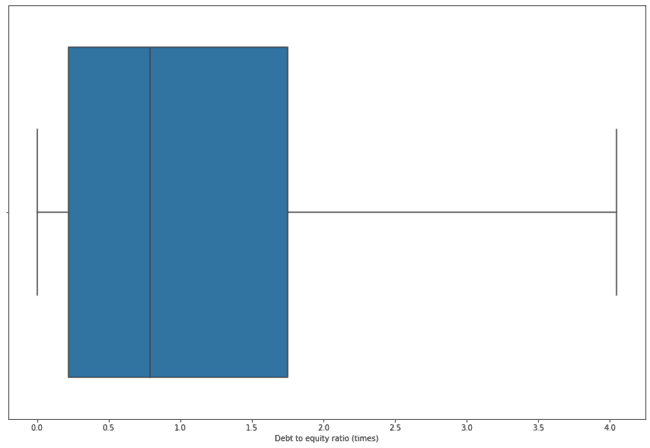
With the help of the descriptive statistics, we can see various measure such as mean, Q1, Q2, Q3, min, max and standard deviation of each feature. Below are few of the insights from above table:

* The average net worth of the companies is 220 million dollars while the least is -10.7 million dollars.
* The highest debt to equity ratio is 4.045 and the average is 1.2.

Visualizing variables:

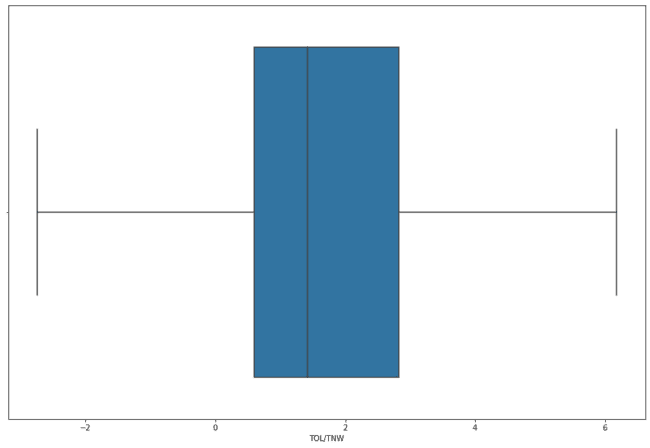
Let’s see the univariate analysis of each variable using boxplot. (**Note:** All the variables have been treated with outlier treatment and are visualized).

The figure 1.2 shows boxplot of debt-to-equity ratio.



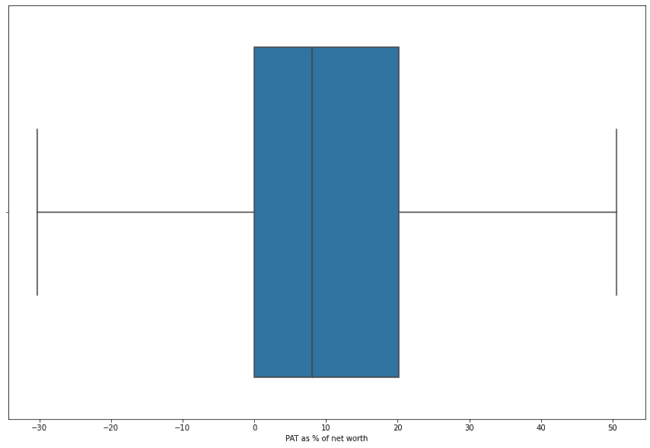
**Fig 1.2**

The figure below shows the distribution of TOL/TNW variable in the forma of boxplot.



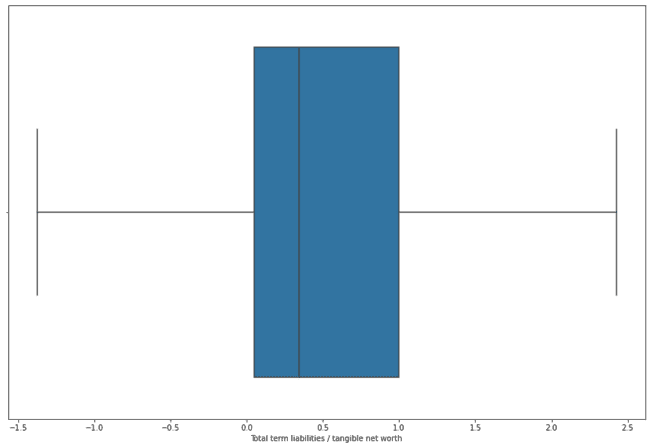
**Fig 1.3**

The fig 1.4 shows the PAT as percentage of net worth. The box plot is slightly right skewed here.

****

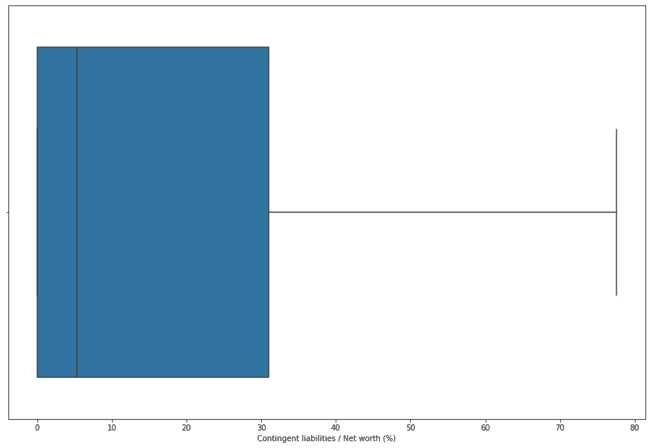
**Fig 1.4**

The distribution of total term liabilities/Tangible net worth is shown in the boxplot below.

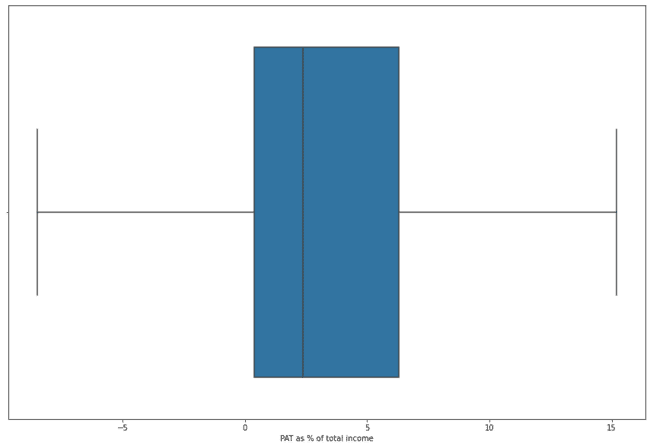


**Fig 1.5**

Next figure shows boxplot of contingent liabilities/Net worth in percentage. The boxplot shows that the variable is heavily right skewed.

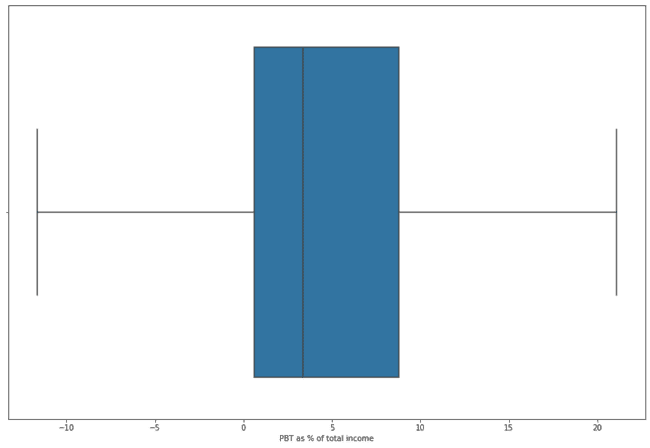


**Fig 1.6**



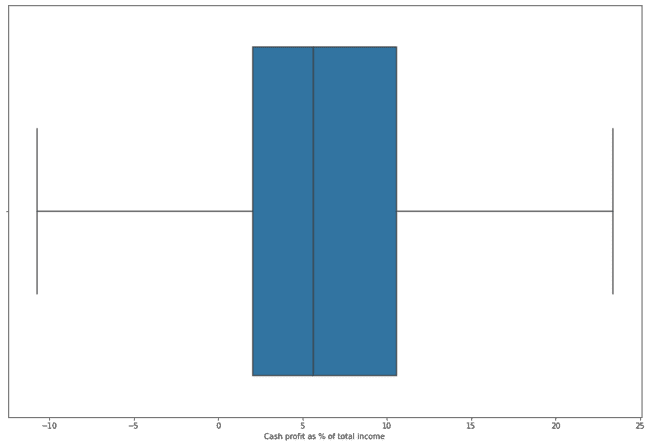
**Fig 1.7**

The figures 1.7 and 1.8 box plots of PAT and PBT as percentage of total income. We can clearly see the outlier treated boxplots. (Refer fig 1.7 and 1.8).

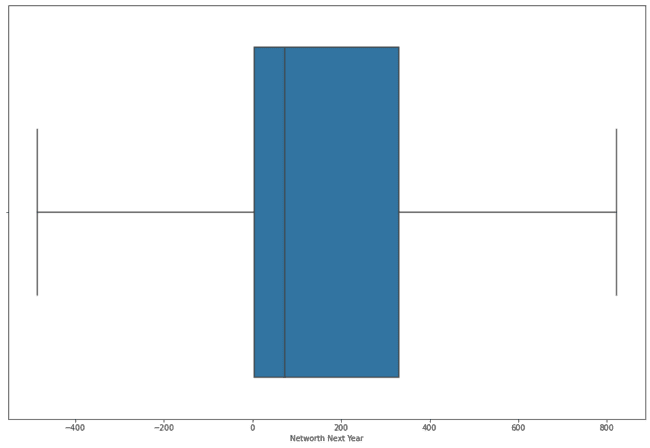
****

**Fig 1.8**

The cash profit as % of total income box plot is shown in the figure 1.9. It is almost normally distributed box plot.

****

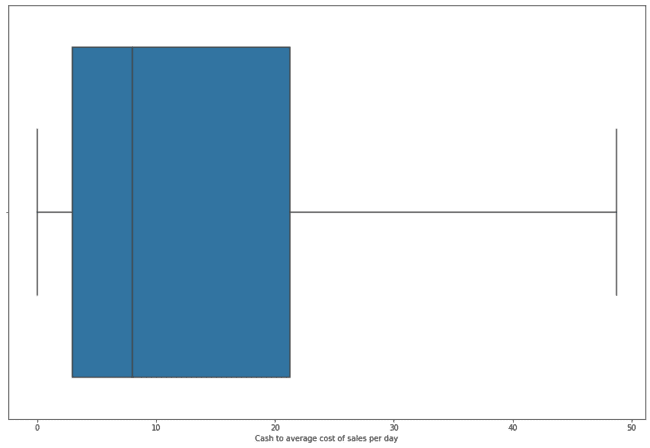
**Fig 1.9**

****

**Fig 1.10**

The above figure shows the next worth of companies next year and it is heavily right skewed.

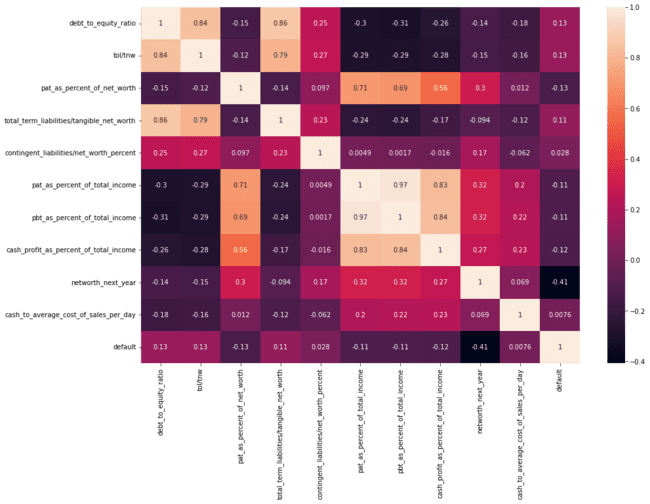
The box plot of cash to average cost of sales per day is shown below. (Refer fig 1.12)



**Fig 1.12**

Heatmap

The heatmap of the top 10 features selected.

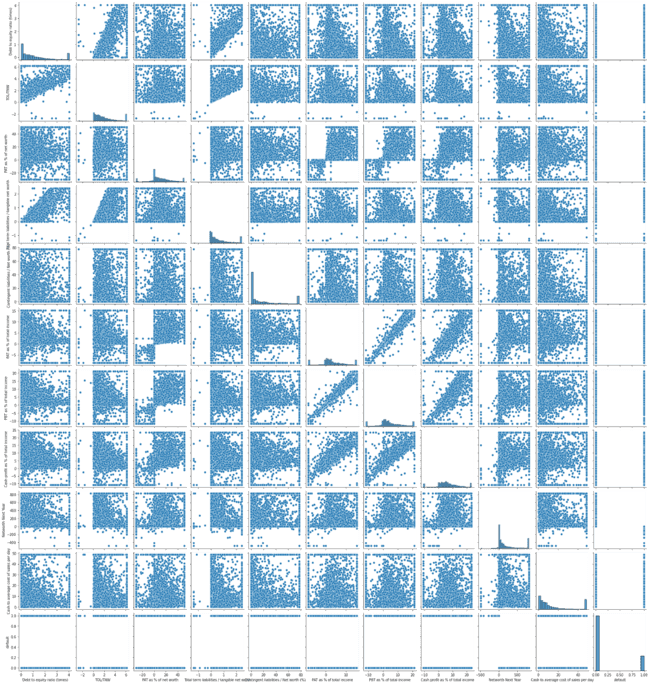


**Fig 1.13**

* Pbt as percent of total income and pat as percent of total income has the highest correlation among all the variables.
* Total term liabilities/Tangible net worth and debt to equity ratio is the second highest correlation.
* Total term liabilities/Tangible net worth and cash to average cost of sales per day has the lowest correlation.

Pair plot

The below figure shows the pair plot of the features. (Refer fig 1.14).



**Fig 1.14**

## Model building

With all data preprocessing such as data cleaning, feature engineering done, now let’s build machine learning models to predict the defaulters.

Two models – Logistic Regression model and Random Forest Classifier model has been built with cleaned data.

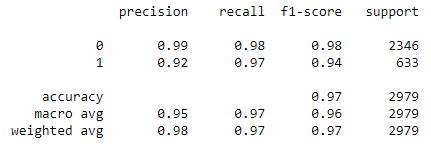
Both the models have trained and tested with training and testing data respectively. Both of these models are also evaluated using metrics such as accuracy score, confusion matrix, roc\_auc score.

Logistic Regression model

Let’s see the Accuracy score, classification report, confusion matrix, ROC\_AUC score and ROC curve of the Logistic Regression model on both train and test data.

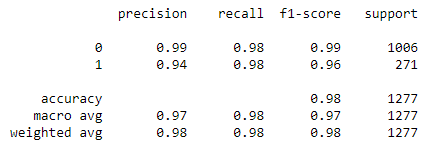
**Classification Report**

**Below figure shows the classification report of the logistic model on training dataset.**



**Fig 1.15**

* The model accuracy is quite good as 97% with training data.
* Though the model has good precision and f1-score, recall is best when compared with other aspects of the classification report.



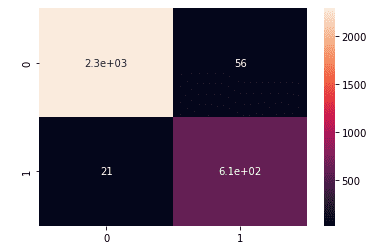
**Fig 1.16**

The above figure shows the classification report of the logistic regression model on the test dataset.

* Here, just like the model’s performance on training dataset the recall is good.
* The accuracy score of the model has also improved.

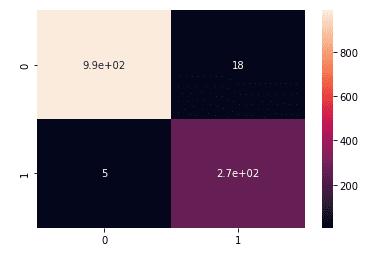
**Confusion Matrix**

**The figure below shows the confusion matrix of the logistic regression model on the training dataset. Here Type I error is higher than the Type II error.**

****

**Fig 1.17**

The figure below is the confusion matrix of logistic model on the test dataset.

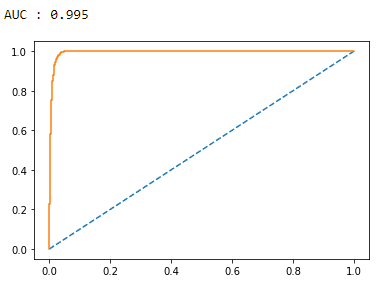


**Fig 1.18**

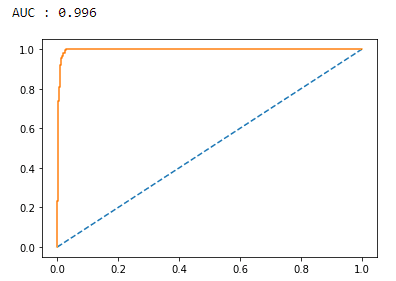
The confusion matrix above shows that the model performed well on the test dataset.

**Auc and Roc**

**The figure below shows the roc\_auc curve and auc score of the logistic regression model on training dataset.**



**Fig 1.19**



**Fig 1.20**

The above figure shows the auc score and the roc\_auc curve of the logistic regression model on the test dataset.

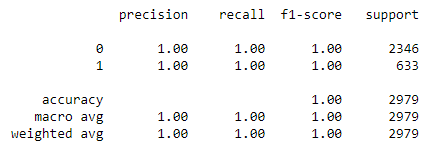
We can conclude that the logistic regression model’s auc score and roc\_auc curve for training and testing dataset is almost identical.

Random Forest Classifier

The other model build for predicting the defaulters is Random Forest classifier.

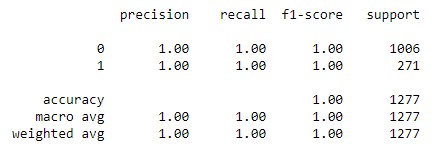
Just like the Logistic Regression model, the Accuracy score, classification report, confusion matrix, ROC\_AUC score and ROC curve of the Random Forest classifier on both train and test data are also evaluated to choose the best model.

**Classification Report**

****

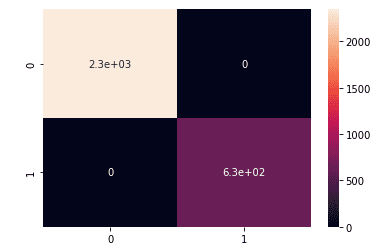
**Fig 1.21**

**Both the classification reports of Random Forest Model on Training and testing data are identical.**

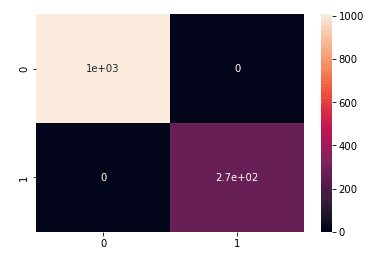


**Fig 1.22**

**Confusion Matrix**



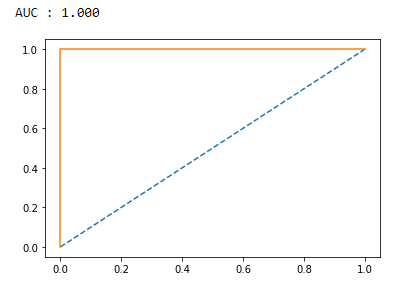
**Fig 1.23**

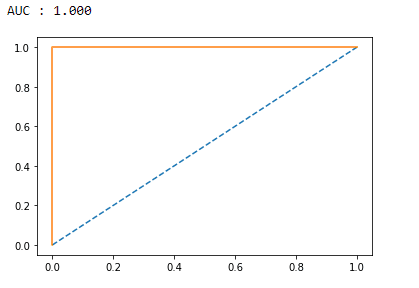


**Fig 1.24**

Similar to the classification report the confusion matrix of the Random Forest classifier on training and testing is exact same in terms of Type I and Type II error.

**Auc and Roc**



  
***Test Train***

**Fig 1.25**

Like the other two performance metrics the auc score and roc\_auc curve of the Random Forest Classifier on the training and test dataset is same.

## Conclusion

Looking at the performance of Logistic Regression model and Random Forest Classifier model can conclude the following:

* The Random Forest Classifier has better performance than the Logistic Regression model.
* Though the performance of the Random Forest Classifier is better it is too early and lacks evidence to prove it as the best model. Since, amount of data given is comparatively lesser in terms of training and testing logistic regression model.
* With this, we can clearly choose the logistic regression model as best choice for the current scenario.