December 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Financial Risk Analytics

Business Report

Thakur Arun Singh

Contents

[Problem 1: 2](#_Toc89639278)

[Problem 1.1 3](#_Toc89639279)

[Problem 1.2 8](#_Toc89639280)

[Problem 1.3 9](#_Toc89639281)

[Problem 1.4 10](#_Toc89639282)

[Problem 1.5 17](#_Toc89639283)

[Problem 1.6 18](#_Toc89639284)

[Problem 1.7 22](#_Toc89639285)

## Problem 1:

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

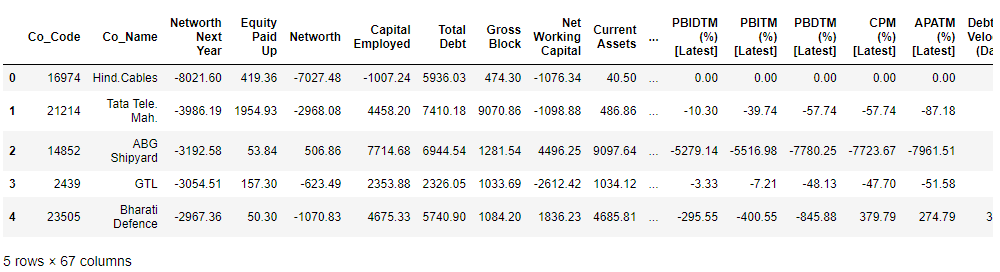
Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

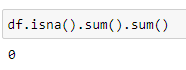
**Exploratory data analysis**

Dataset has 67 variables of which 63 are of float data type, 3 are integer type and 1 is object type.

The head of the dataset is as below:



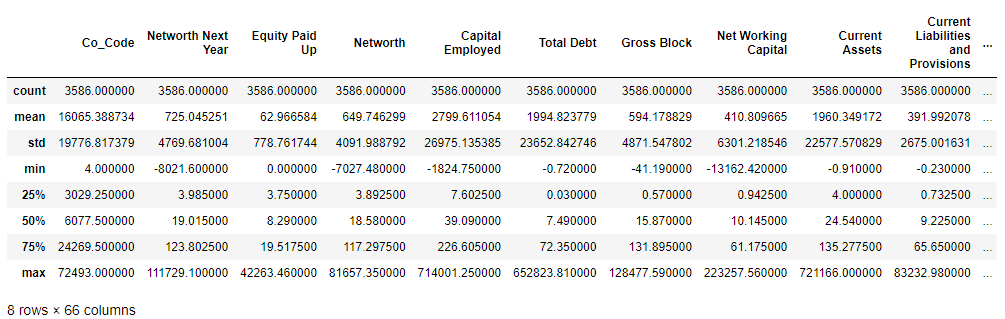
The data has 3586 Rows and 67 Columns. No duplicate data is present in the data set.

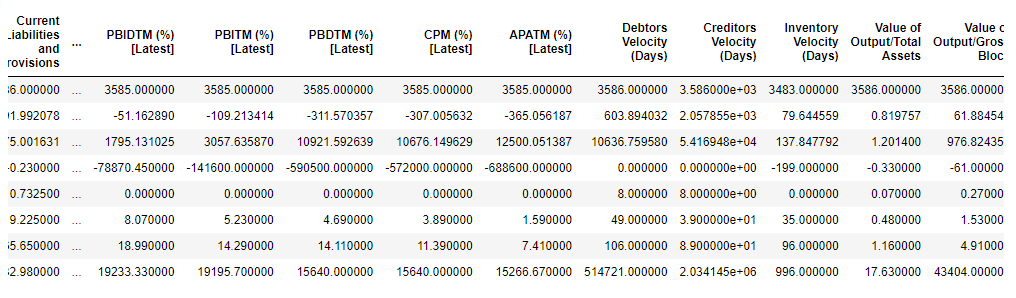


The data has 3586 Rows and 67 Columns. No duplicate data is present in the data set.

We dropped unrequired columns like Co\_Code and Co\_Name since they do not add value to the analysis.

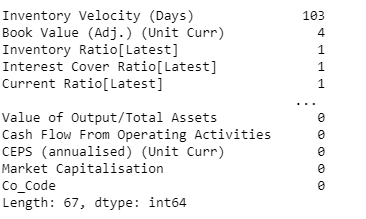
Descriptive statistics / 5 point summary is shown below.





The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile are mentioned in the above tables.

Next we checked for null values.



Further details on missing values is covered under 1.2

### Problem 1.1

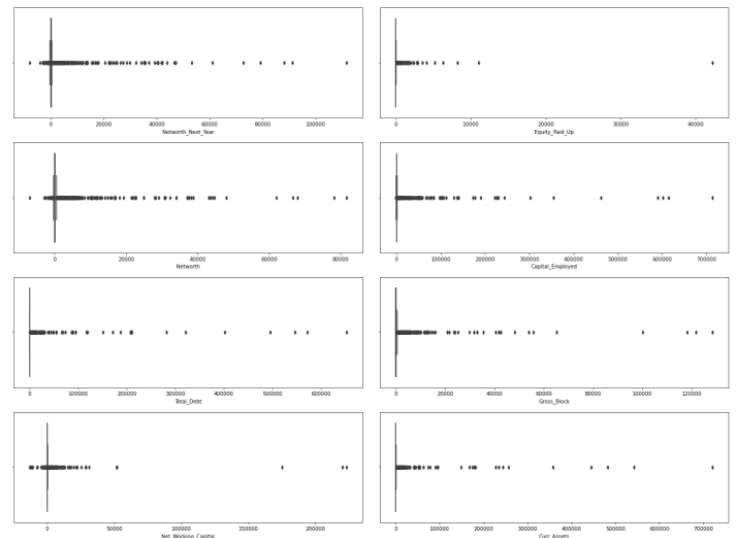
**Outlier Treatment.**

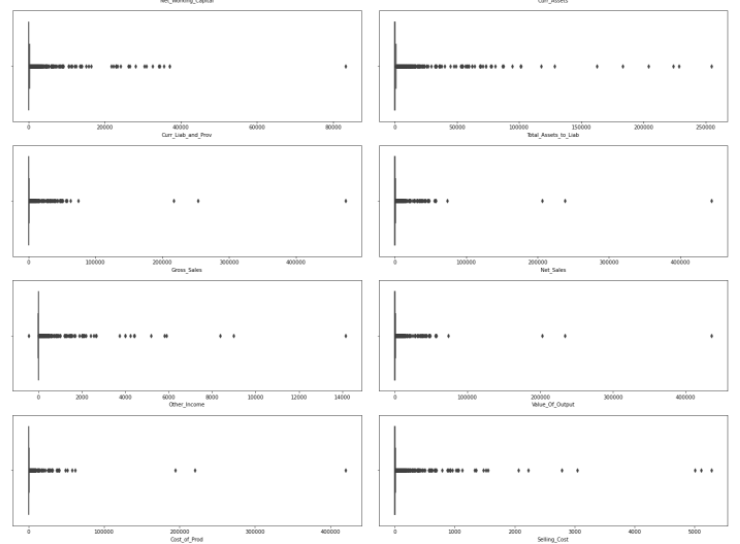
**Resolution:**

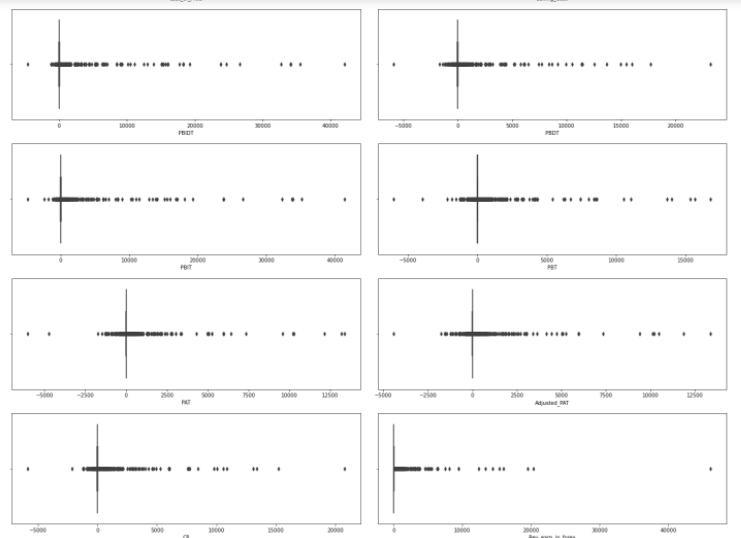
**Describing the data:**

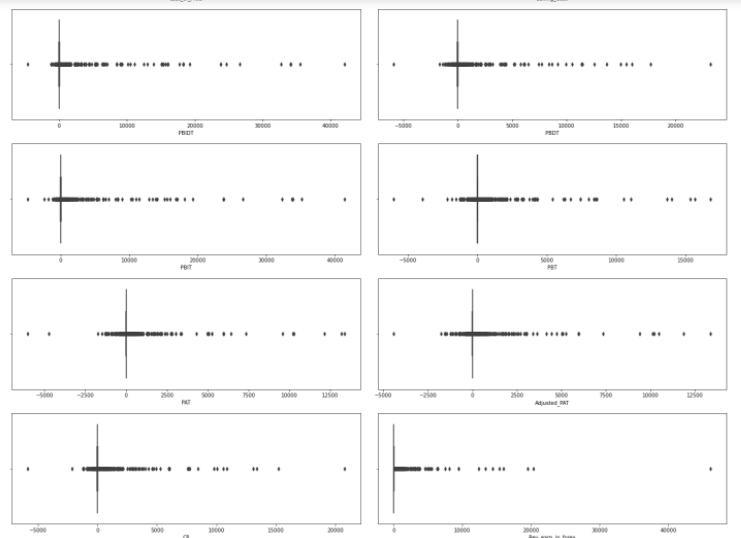
* First we import all the necessary libraries in Python, and then import the data file which is ‘Company\_Data2015-1’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
* Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
* To check the data type of all the columns and also to check the null values, ‘info’ function. Has been used.
* To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,
* Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
* Using the ‘dups’ function we check for the duplicates and there were no duplicate values.
* We also identified the unique values in categorical data.

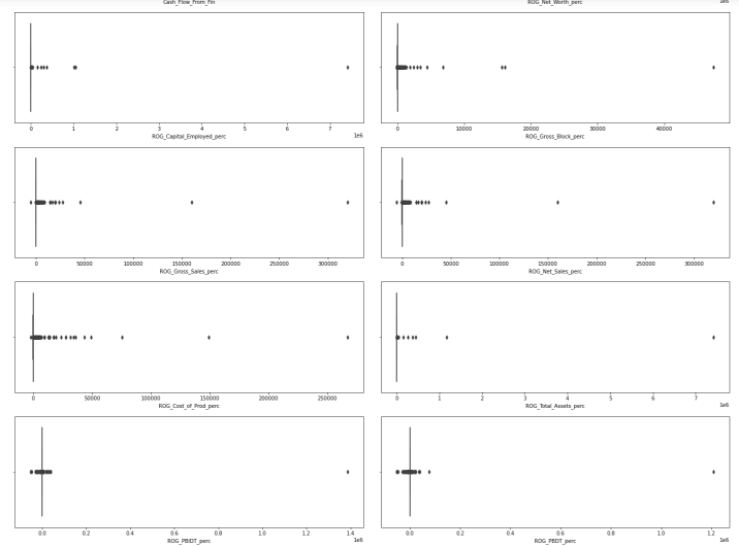
We used 3 times the IQR range as the criteria to determine the outliers. Our analysis gave significant chunk of outliers in the data. Below are boxplots which were plotted to analyze this data.

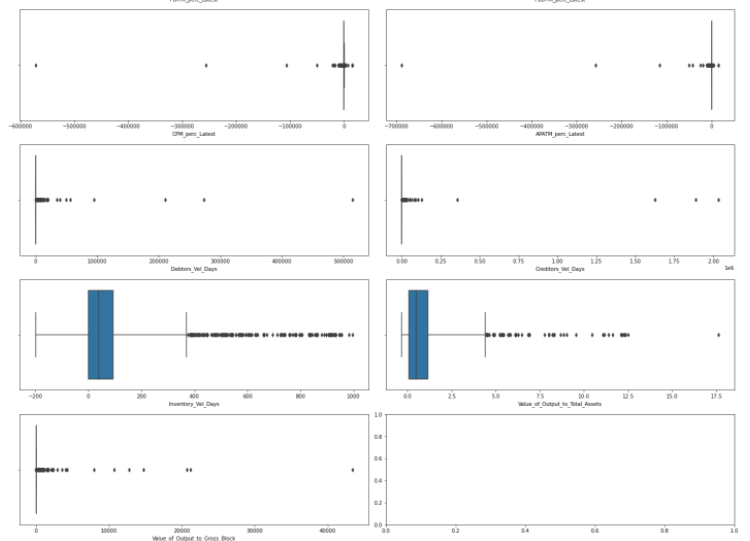








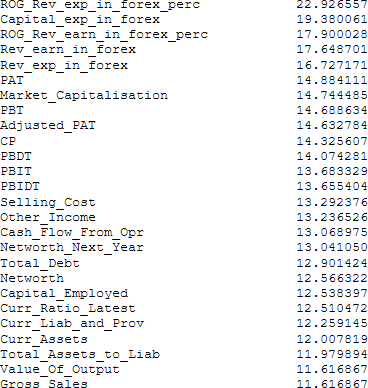




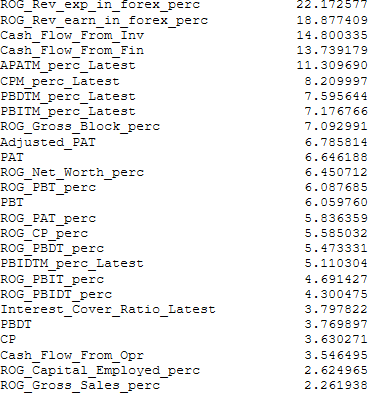
**O U T L I E R T R E A T M E N T**

Significant number of outliers were present for almost all the variables. We captured the actual percentage of data which was above and below the third and first quintiles respectively.

**Data above third quintile.**



**Data below first quintile.**



Since the number of outliers are too large in number to be treated, as treated such large number of records would mean changing the essence of the data. Also given the fact that this is a financial data and the outliers might very well reflect the information which is genuine in nature. Since there is data captured for small, medium as well as large companies.

Hence we decided against treating the outliers in this data set.

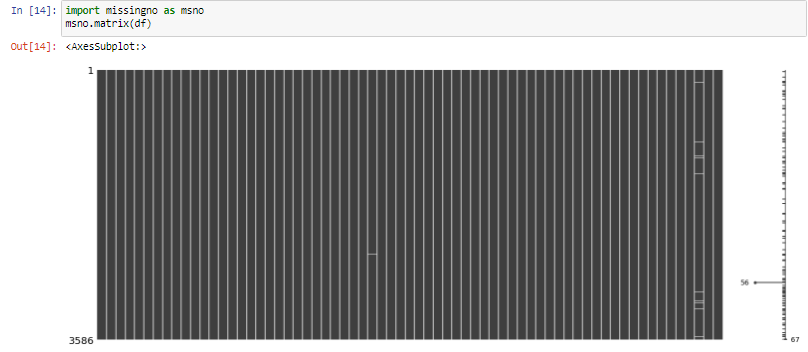
### Problem 1.2

Missing Value Treatment

**Resolution:**

Given the size of the data set i.e. 3586 rows, there were not many missing values to start with. There were a total of 118 missing records observed in the entire data.

Snapshot from missingno library has been published below for reference.

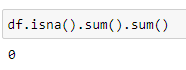


Null values were present in many columns, however significant number was present in "Inventory\_Vel\_Days" column. This is the one which we treated.

Records with missing value in "Inventory\_Vel\_Days" column were imputed with the average value.

After this imputation, there were another 15 rows with missing data, however this number was too small to warrant any additional efforts. Hence we dropped these rows the purpose of the analysis.

No more missing values were present after treatment.



### Problem 1.3

Transform Target variable into 0 and 1

**Resolution:**

A new dependent variable named "Default" was created based on the criteria given in the project notes.

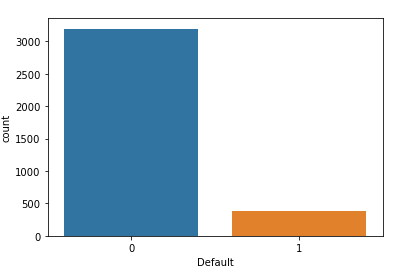
**Criteria -**

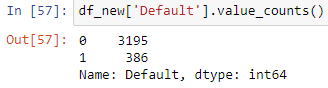
1 - If the Net Worth Next Year is negative for the company 0 - If the Net Worth Next Year is positive for the company

Making use of np.where function to achieve this.



After generating the dependent column, we checked for the split of data based on this dependent variable. Below is a bar plot showing the same.



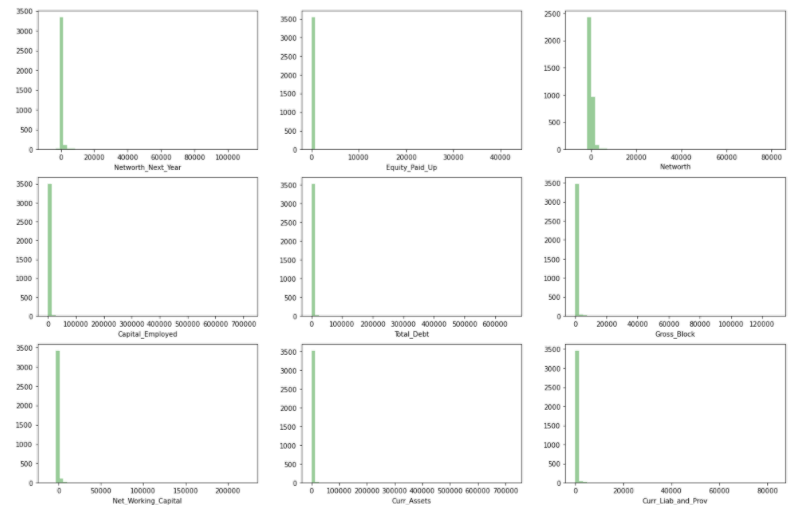


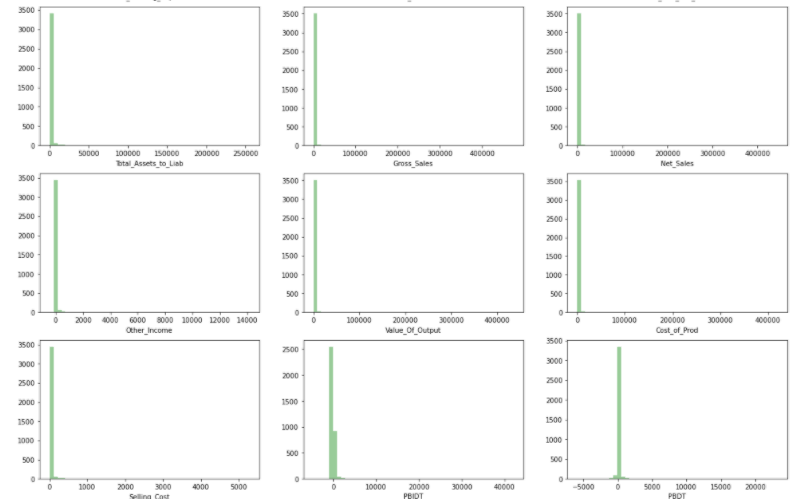
### Problem 1.4

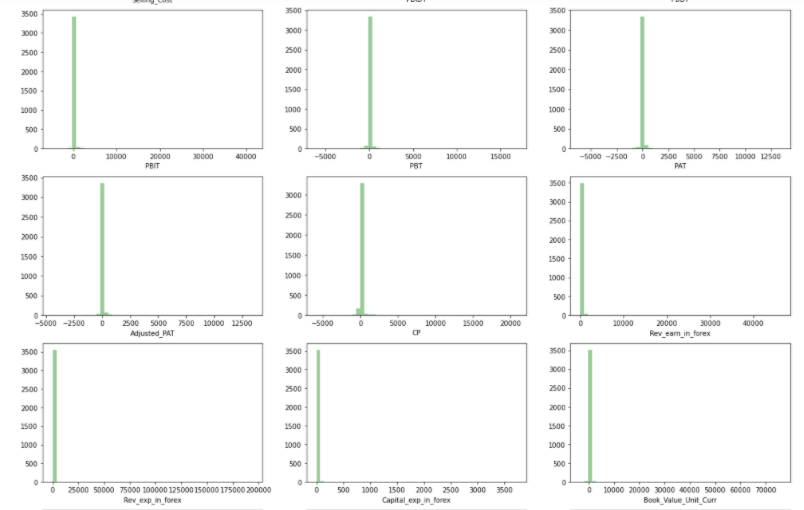
Univariate & Bivariate analysis with proper interpretation.

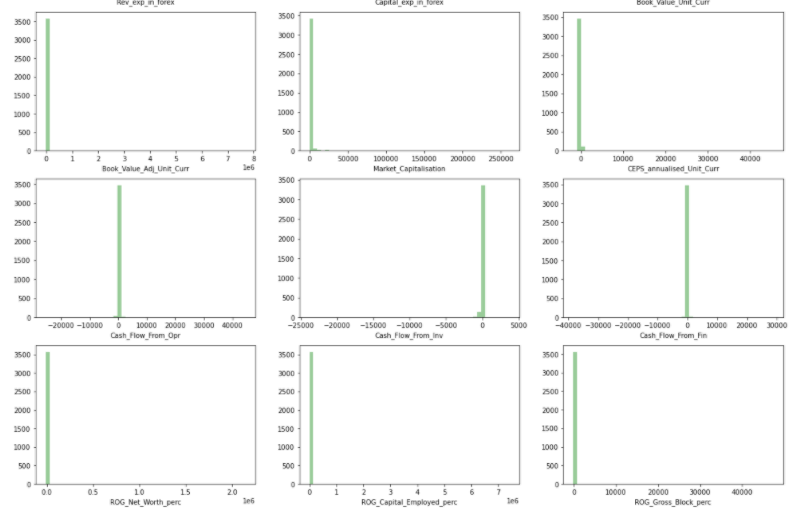
**Resolution:**

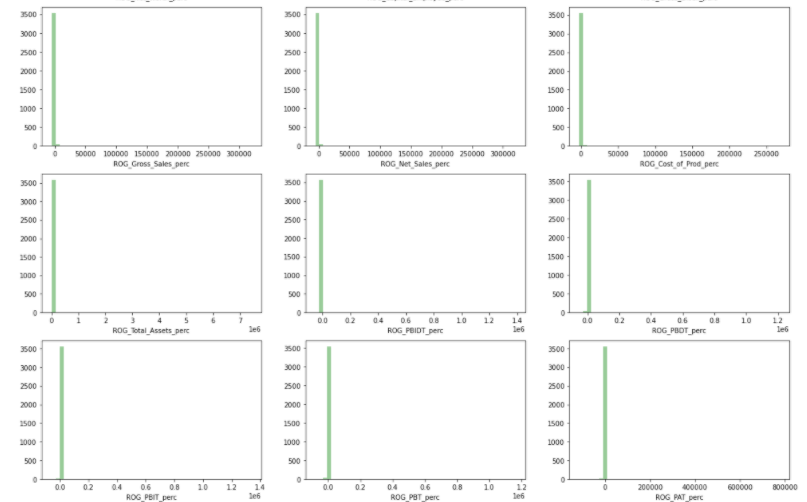
Distplot were plotted for all the variables to analyze the distribution of all the variables.

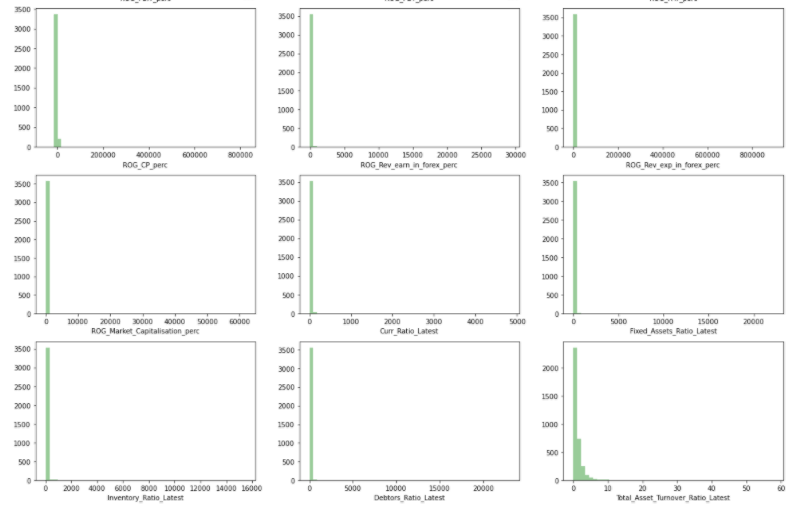


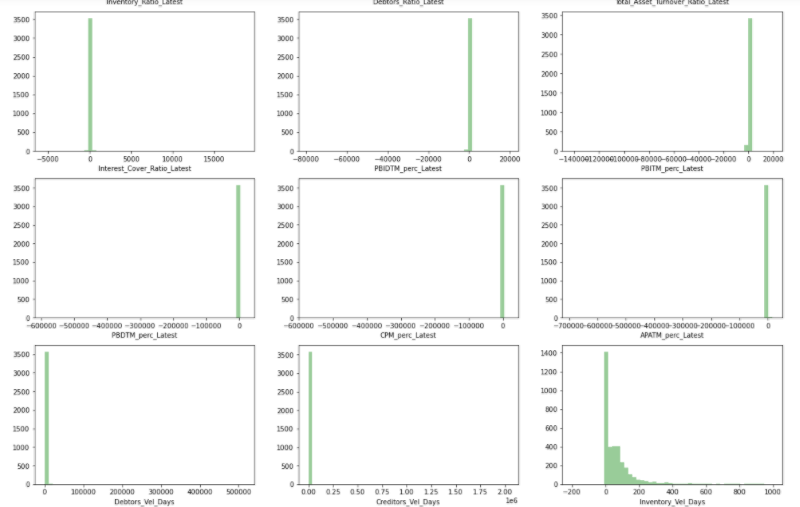


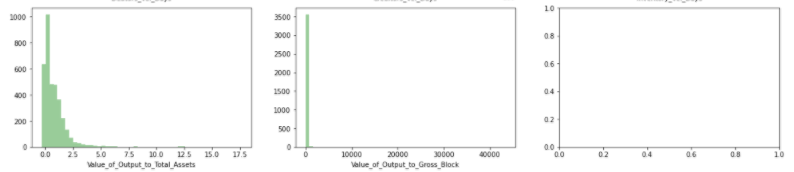






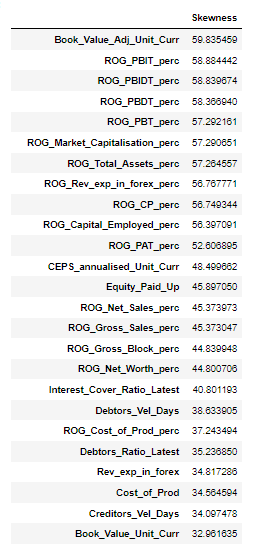
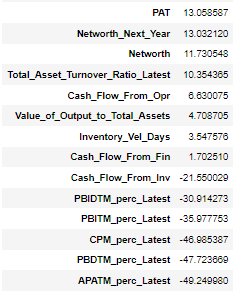






None of the variables show perfect normal distribution. Few of the variables have skewness in data. There are no duplicate values.

Skewness was observed in almost all teh variables. Most of the variables were right skewed while a few were also found to be left skewed.

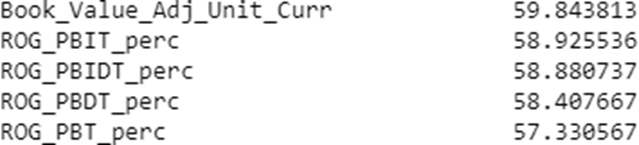
Univariate Analysis

Data is highly skewed and most of the data is found to be right skewed.

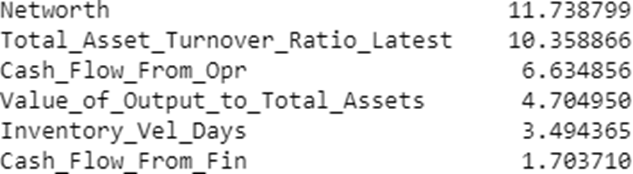
A total of 61 variables were found having tails to the right and hence were right skewed.

There were a total of 6 variables which were found to be left skewed i.e. they had a longer tail on the left hand side of the distribution.

The top 5 variables that have the highest skew are:



The top variables that have the least skew are (in decreasing order):

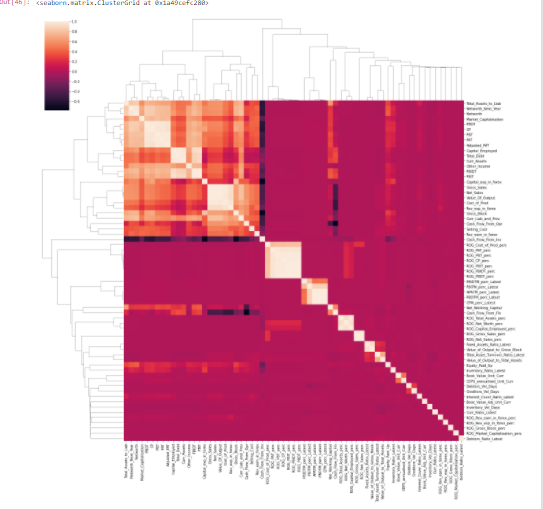


**M U L T I V A R I A T E A N A L Y S I S**

We also performed multi Variate analysis on the data to see if there are any correlation that are observed within the data. Correlations function was used and seaborn cluster map was used to plot the correlations and to make better sense of the data.

We observed that net worth and net worth next year were highly correlated. Apart from this, we also found various Rate of Growth variables were highly correlated.

This analysis tells us that there is a problem of collinearity with this data set. Heat map has been plotted on the next page.

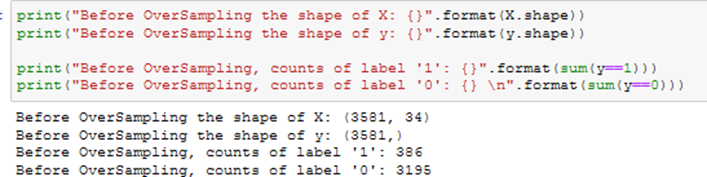


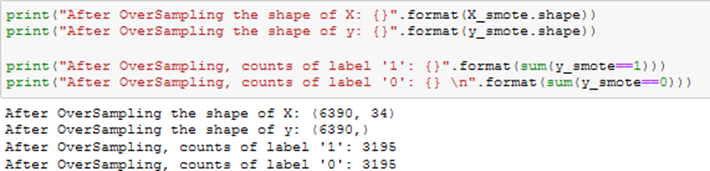
### Problem 1.5

Train Test Split

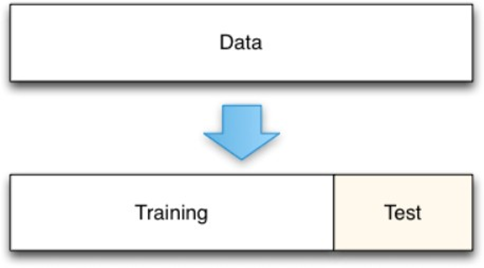
**Resolution:**

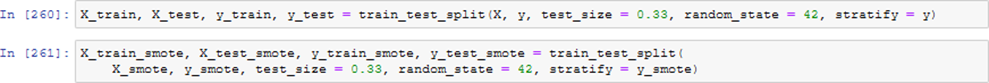
Since there was a great imbalance in the data set, we also created a parallel data set with SMOTE and evaluated the performance on smote as well as non smote data.





After this data was split into train and testing set, using the stratify = y argument, keeping the ratio of Default variable more or less similar in training as well as testing set.





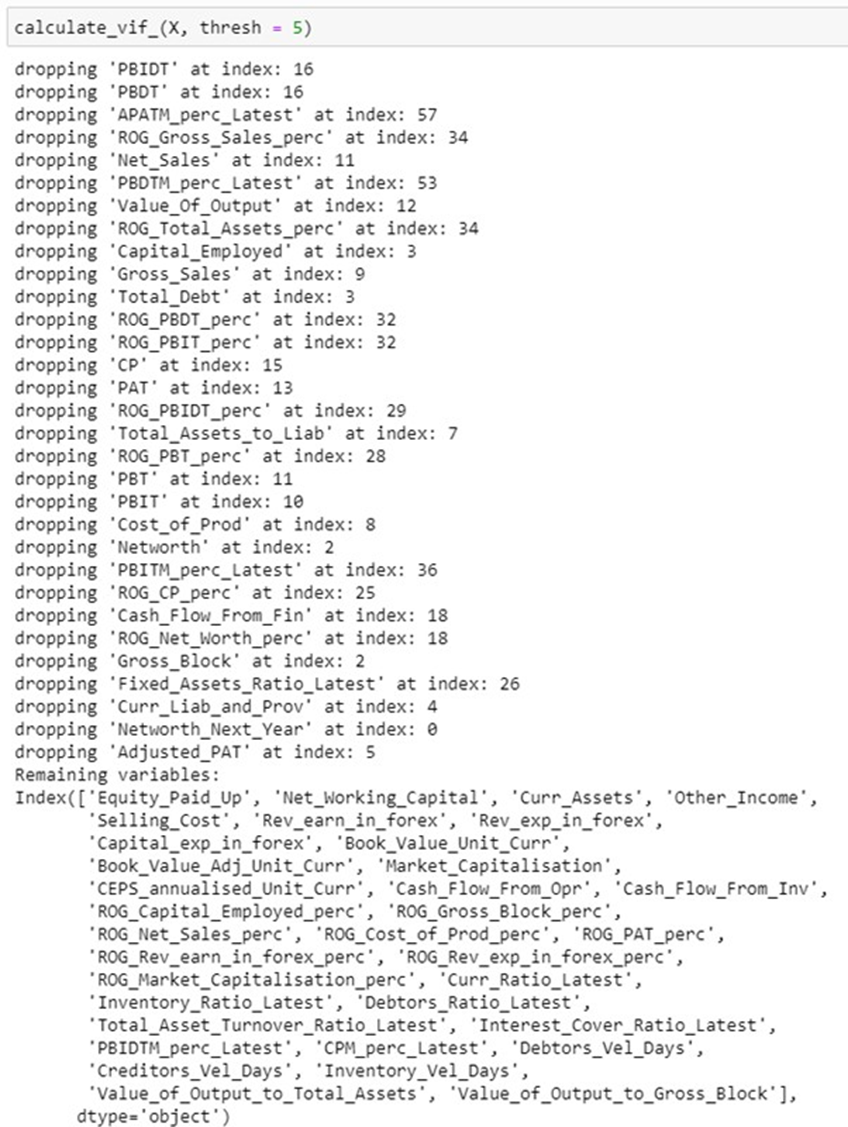
Data was split in the 67:33 ratio as per project notes using sklearn's train\_test\_split function. Also seed value of 42 was used

### Problem 1.6

Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach

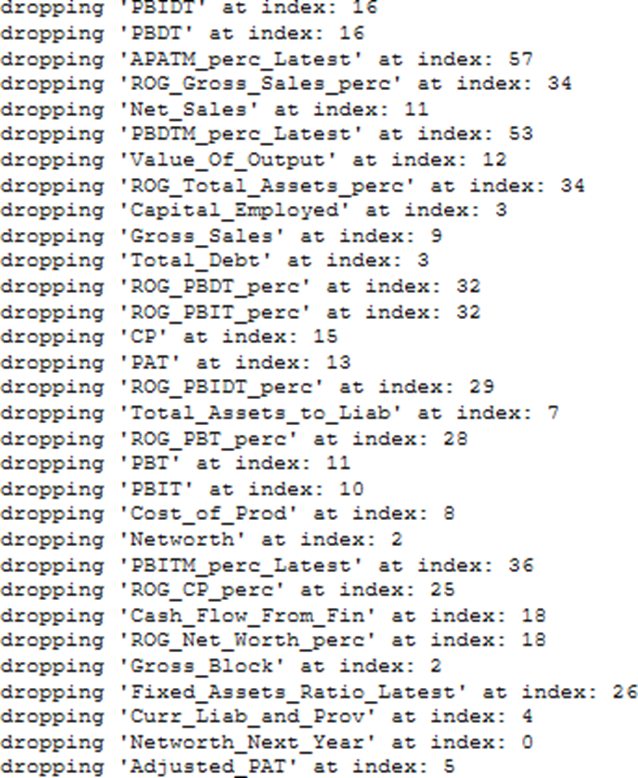
**Resolution:**

Prior to building the logistic regression model, we had to work on feature selection since there were too many columns to start with and we decided to eliminate a few of the columns using the Variation Inflation Factor i.e. VIF

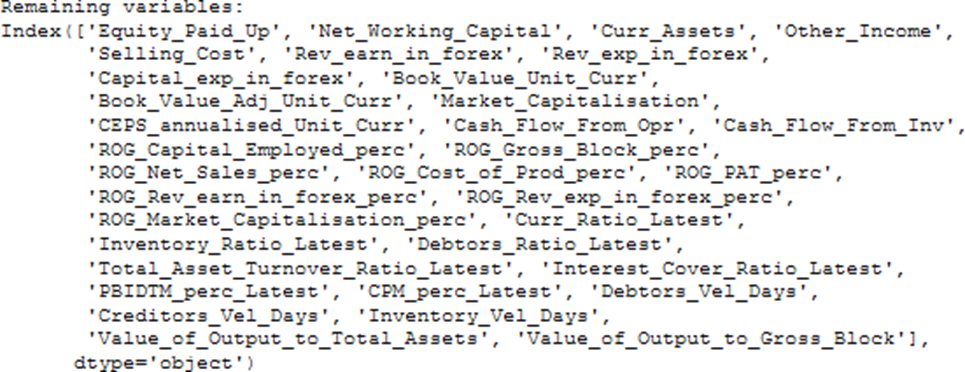


**LOGISTIC REGRESSION**

A number of variables were dropped as part of this VIF calculation. These were as below.



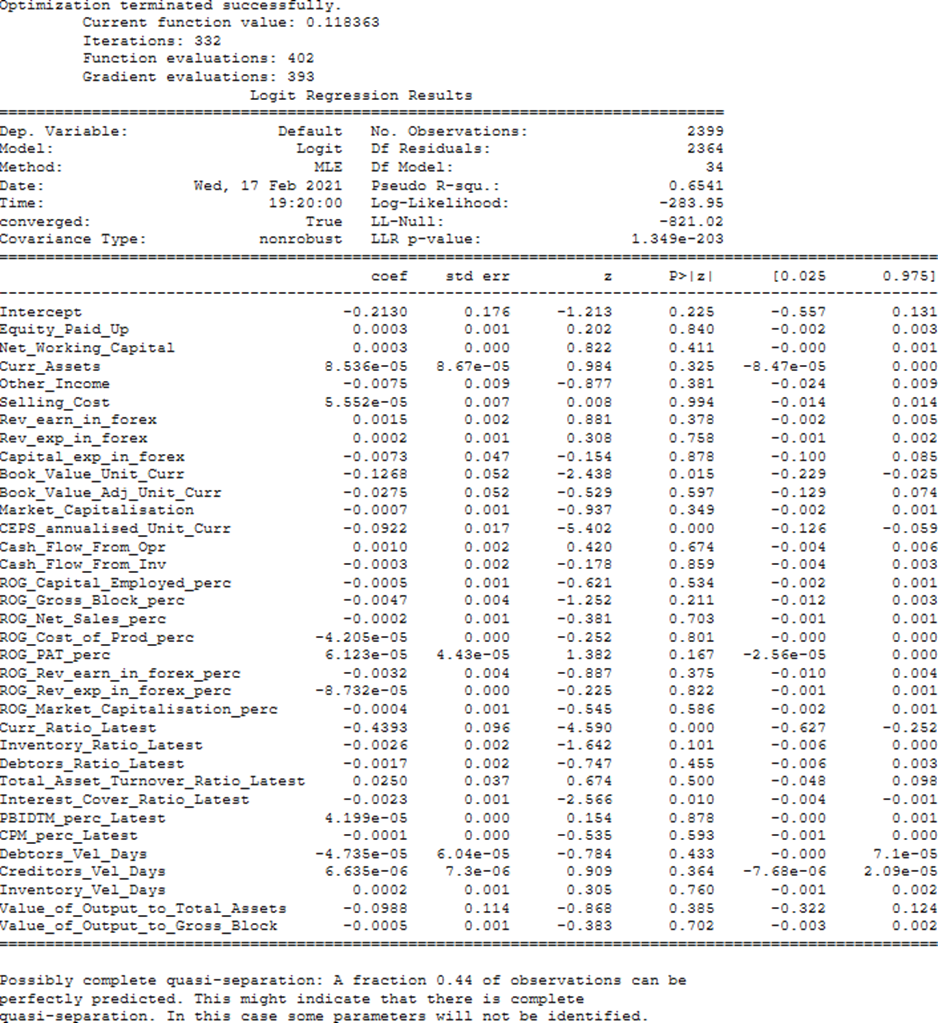
A total of 34 variables were retained after this exercise. These were as below.



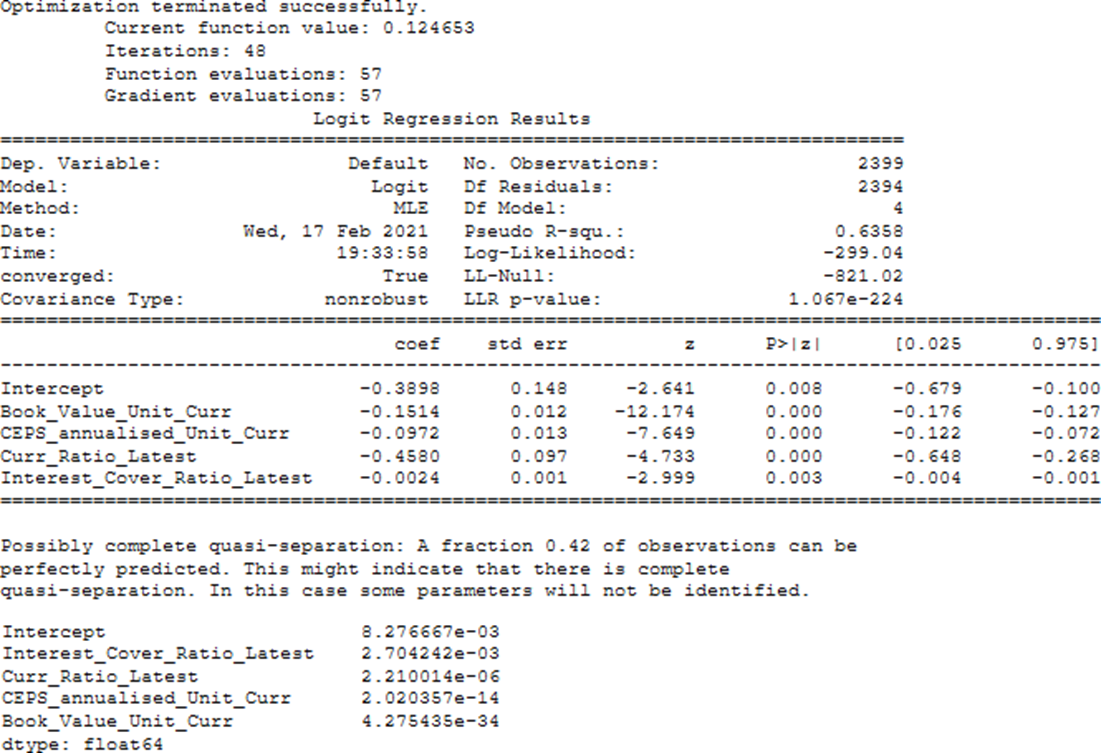
Backward elimination method was used for model tuning, we started wtih all 34 variables and built a model, evaluated the p-values at the end of it. Then we removed the variable with highest p-value and then re-ran the model. This process was re-iterated multiple times until we had all variables whose p-value was less than 0.05.

Variables with p-values less than 0.05 were dropped since their coefficients are unreliable and might very well be just a statistical coincidence.

**First Model**



It is evident from the image that the variable Selling\_Cost has a p-value of 0.993767. Since this is higher than 0.05 and the highest of all the variables, we will drop this variable in subsequent models. This process of dropping variables based on p-values and modeling continued until a model where all the p-values were relevant was achieved. The iterative process got stopped at Model30 which has 4 independent variables and each of them were relevant.



P-values of all the variables are less than 0.05 and thus all the coefficients are relevant. Book\_Value\_Unit\_Curr has the highest coefficient and Interest\_Cover\_Ratio\_Latest the least of all. This model will be used to validate the test dataset.

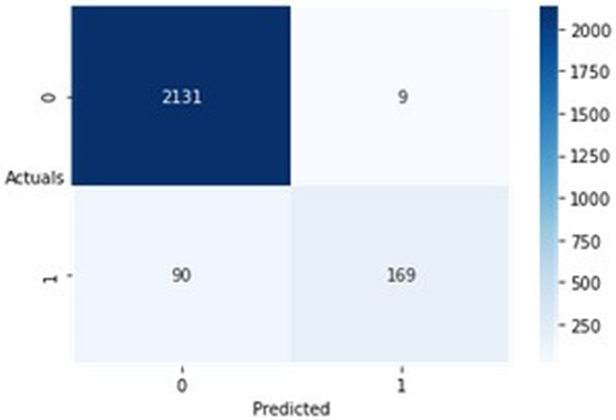
Evaluation on SMOTE set did not yield any better results. Hence we stuck to the original data set.

### Problem 1.7

Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model

**Resolution:**

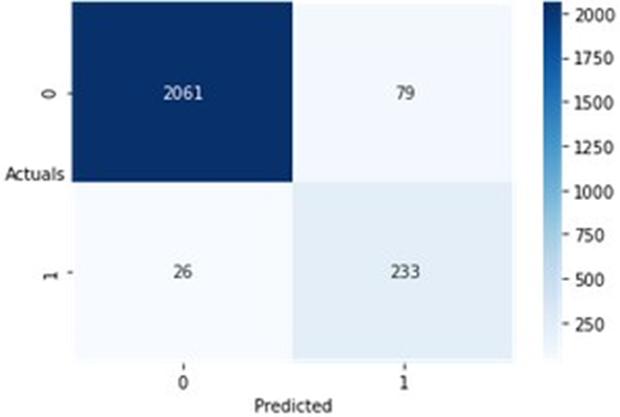
With default probability threshold of 0.5, the confusion matrix for the train set is as follows:

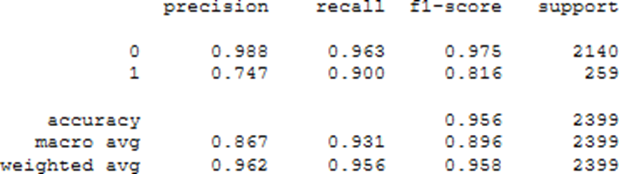


Correctly predicted = 2131 incorrectly predicted records = 169

This was pretty good result on its own, however to further improve the on the results. We decided to look for the optimum threshold.

After evaluating using the optimal threshold. Below was the new classification matrix.

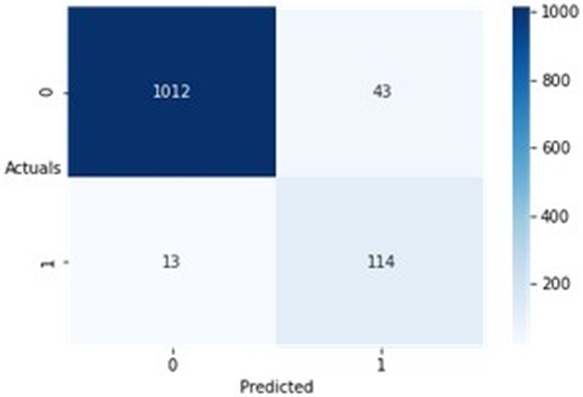


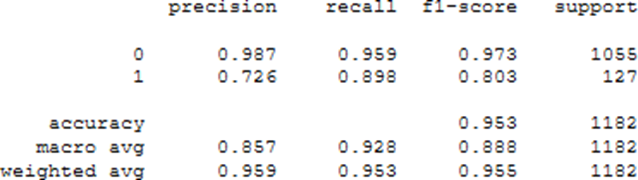


Accuracy of over 95.6% was achieved while recall, precision and f1 score were also very high at 96.3,98.8% and 97.5% respectively.

We also evaluated the test data set for the same model which was built after the above mentioned re-iterative process.

Below are statistics for the test model.





Accuracy of 95.3% and very high recall, precision and f1 score of 95.9% ,98.7% and 97.3% respectively were also observed on the test set. This clearly indicates that the model which has been built is highly efficient and has been able to capture the correct variable for prediction.

Tt has been proven to work on train as well as test data.

The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*