|  |
| --- |
| CP G1 Group Assignment |
| Finance & Risk Analytics – Indian Credit Risk Model |

Amit Sharma

Amitesh Bajpai

Sahil Sachdeva

Kushal Maheshwari

Neeraj Singh

# Project Objective

**The objective of this assignment is to build a model which can identify the right set of business firms who are predicted to default. Basically, build a model predicting their probability of Default.** 3,541 observations of various companies’ financial parameters are given in the dataset. A separate validation dataset is given to test out this model. For both datasets, the default variable is also given. This will be the dependent variable for the model.

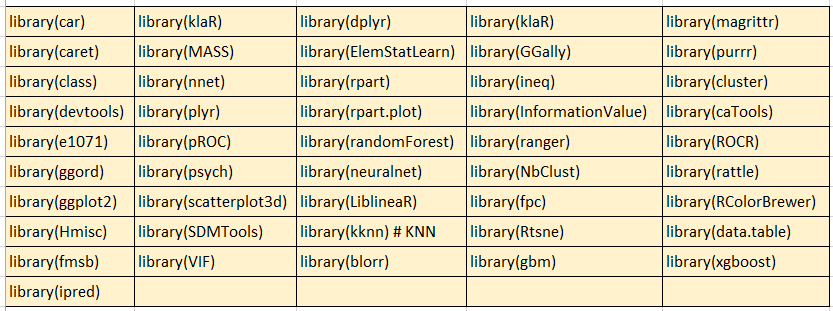
As per the problem statement, this exploratory report will consist of the following:

* + Basic Exploratory Data Analysis of dataset given.
    - Basic Data Summary
    - Outlier Treatment
    - Missing Value Treatment
    - New Variables Creation (for profitability, leverage, liquidity and company's size each)
    - Check for Multi-Collinearity
    - Univariate Analysis
    - Bivariate Analysis
  + Modelling
    - Build Logistic Regression Model on most important variables
    - Analyze coefficient & their signs
  + Model Performance Measures
    - Predict accuracy of model on train and validation datasets
    - Sort the data in descending order based on probability of default and then divide into 10 Deciles based on probability & check how well the model has performed

# Data Preparation and Exploratory Data Analysis

## Libraries

The following set of libraries were installed in R for achieving the project objective.



## Dataset and its variables

Around 3,541 observations of Company’s parameters are given in the dataset. 50 variables are present in the dataset (1 variable - Deposits (accepted by commercial banks) – doesn’t have any values and hence we have omitted it). The following table shows a quick summary of the Category and the type of variables.



There are 21 Ratio variables and 29 Absolute variables in the dataset. Most of the Ratio Variables are taken as the ratio between the absolute variables only.

There are 10 different categories here. 13 variables from Profit, 12 variables from the Size category and 12 variables in Liquidity. These are then followed by Leverage (6 variables) and Profitability (2 variables). Size/Costs, Market Sentiment, Change in Size and Liquidity/Size has single variables each.

The Dependent variable here is “Default” which will be used for the models we run.

## Dataset Cleansing

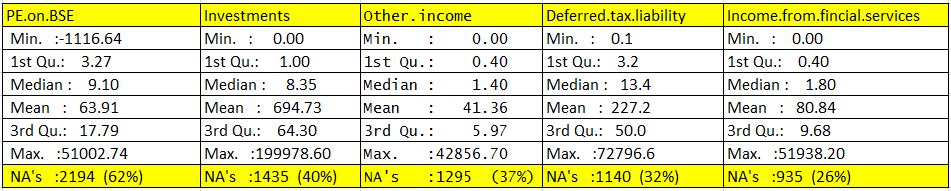
This dataset *(raw-data)* was uploaded into R. It was observed that one variable had all values as blanks. To avoid further issues with these variables, this variable was deleted. While visually exploring excel, it was found that there were two types of missing values i.e. Blanks and “NAs”. All the NAs were replaced by blanks to do the missing value treatments. Also, as discussed before, one variable was completely deleted as all the values were missing in it - Deposits (accepted by commercial banks)

After this, we tried in identifying how many NA values were there.

*> sum(is.na(company)) [1] 14992*

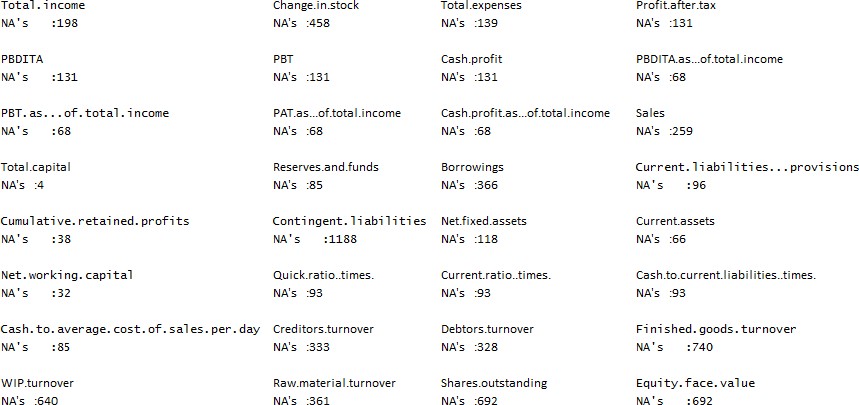
There were around 14,992 missing values reported. All NA values needs to be treated.

## Missing value treatment

Missing value treatment is an important step in this assignment. There are several packages in R which deals this with nicely. First, we will have to understand which of the variables have the most missing value. Having very large number of missing values might not make much sense in pushing into the model. A cut off percentage of 20% was kept for the selection variables i.e. If you have more than 20% of the values as missing values, then that variable is to be avoided in the model. The variables which were more than 20% of the missing values are as shown below:

All these high percentage shares of missing values were avoided from the Dataset. The remaining set of variables were adequately treated for the missing values.

The variables and their corresponding missing values are as follows:



All the above Missing values were to be treated in the dataset. Now, in R there are various type of missing value treatment packages.

### Imputation with Mean / Mode / Median

Replacing the missing values with the mean statistical measures is a crude way of treating missing values. Depending on the context, like if the variation is low or if the variable has low leverage over the response, such a rough approximation is acceptable and could give fine results. But in our case as there are several variables and most of them are quite an important measure, we are not going ahead with this measure.

### “knn” Imputation

kNN imputation uses k-Nearest Neighbours approach to impute missing values. For every observation to be imputed, it identifies ‘k’ closes observation based on the Euclidean distance and computes the weighted average (weight based on distance) of these ‘k’ observations. This is using the “DMwR” library.

*library(DMwR)*

*company.knn<- knnImputation(company[,*

*!names(company) %in% "medv"])*

### “mice” Treatment of missing values

“mice” short for Multivariate Imputation by chained equations is an R package that provides advanced features for missing value treatment. It uses a slightly uncommon way of implementing the imputation in 2-steps, using mice() to build the model and complete() to generate the completed data. The mice(df) function produces multiple complete copies of df, each with different imputations of the missing data. The [complete()](http://www.inside-r.org/packages/cran/mice/docs/complete) function returns one or several of these data sets, with the default being the first.

*library(mice)*

*miceMod <- mice(company[, !names(company) %in% "medv"], method="rf") # perform mice imputation, based on random forests.*

*miceOutput <- complete(miceMod) # generate the completed data. anyNA(miceOutput)*

*company.mice = miceOutput*

In this assignment, for treating the missing values knn imputation and mice were both tried. After a thorough research it was found that using the mice package predicts the missing values better and hence the mice package was used.

## Outlier Treatment

Almost all the variables are skewed very highly. Also, it indicates a presence of outliers on the very extreme side. This makes the distribution of the data very unorderly and uneven. There is a very high scope of proper outlier treatment and feature re-engineering for this dataset.

For better plots, we have tried doing a sample outlier treatment for the variables wherever there is the highest gap from the 95th to the 100th Percentile. For example, if the gap is the highest between 95th percentile and 96th Percentile – all the values will be capped at 95th percentile and similarly on the lower extreme side if the gap is the highest between the 4th and the 5th percentile. All the values will be capped at 5th percentile. Different variables were treated differently.

An outlier will be capped if the values is below its first quartile – 1.5IQR or above third quartile

+ 1.5IQR.

The following is the code snippet used for treating the outliers:

*library(outliers) outlier\_capping <- function(x){*

*qnt <- quantile(x, probs=c(.25, .75), na.rm = T) caps <- quantile(x, probs=c(.05, .95), na.rm = T) H <- 1.5 \* IQR(x, na.rm = T)*

*x[x < (qnt[1] - H)] <- caps[1]*

*x[x > (qnt[2] + H)] <- caps[2] return(x)*

*}*

And once the outlier capping function is defined, the capping is done according to the following code: *company.mice.2.ot$Net.worth=outlier\_capping(company.mice.2.ot$Net.worth) company.mice.2.ot$Total.income=outlier\_capping(company.mice.2.ot$Total.income) company.mice.2.ot$Change.in.stock=outlier\_capping(company.mice.2.ot$Change.in.stock) company.mice.2.ot$Total.expenses=outlier\_capping(company.mice.2.ot$Total.expenses) company.mice.2.ot$Profit.after.tax=outlier\_capping(company.mice.2.ot$Profit.after.tax) company.mice.2.ot$PBDITA=outlier\_capping(company.mice.2.ot$PBDITA) company.mice.2.ot$PBT=outlier\_capping(company.mice.2.ot$PBT)*

Usually in financial projects, it is not advised to do outlier treatment, as they say that these outliers are actual values. But for the model to be a good predictive one and an accurate one, outliers are necessary. Other outlier treatment methods are like Imputation with Mean/Mode/Median and Prediction based (the outliers can be replaced with missing values (NA) and then can be predicted by considering them as a response variable). The Imputation method might be very basic and cannot be applied here as this is a financial project. The Prediction based method might too beyond according to the scope of this project. Hence the best method used here was the capping approach by Quantiles.

## New ratio variables creation

Based on the existing set of variables, it is proposed to create new set of ratio variables to get the modelling more accurate. As per the problem statement it is expected to create one set of variables from each of the following domains:

* Profitability
* Leverage
* Liquidity
* Company Size

#### Profitability

**Profitability Ratio**  This is the ratio between the profit and sales. This will give an idea on how much profit the company has in terms of sales. The ratio = (Profit after Tax) / (Sales)

**Profitability in terms of Assets**  This is the ratio between Profit Before Tax and Total Assets. This is one of the main parameters in the Altman Z Score. The corresponding ratio in the India Z score model is Profit After Tax and Depreciation and Total.Assets. This might not be required separately as it will just increase the multi collinearity.

#### Leverage

**Total Equity**  This will help us to identify the total shareholder equity value. This is attained by the ratio of (Total.liabilities) and (Debt.to.equity.ratio..times.)

**Equity Multiplier**  The equity multiplier is a financial leverage ratio that measures the amount of a firm's assets that are financed by its shareholders by comparing total assets with total shareholder's equity. In other words, the equity multiplier shows the %age of assets that are financed or owed by the shareholders. The ratio between (Total.assets) and (TotalEquity).

**Borrowing Ratio**  This is the ratio between (Total.Borrowings) and (Total.assets). This is one of the main parameters in the India Z Score Model.

#### Liquidity

**Liquidity Ratio**  Liquidity ratios are measurements used to examine the ability of an organization to pay off its short-term obligations. This is attained by the ratio of (Net.working.capital) and (Total.assets).

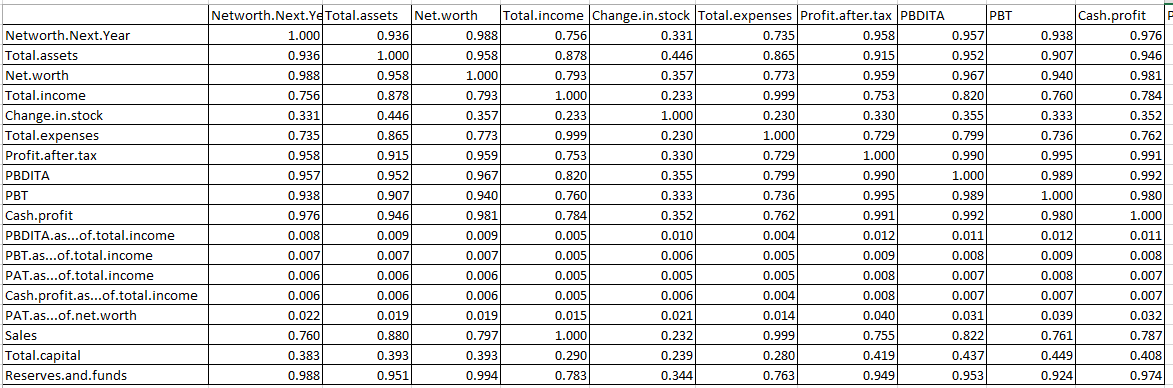
**Asset Turn Over Ratio**  This is the Ratio between (Sales) and (Total.assets). One of the main parameters in the Altman Z Score

#### Company Size

**Company Size** This is to determine on how the company is doing in terms of its Net worth. This is attained by the Ratio of (Net.worth) and (Total.assets).

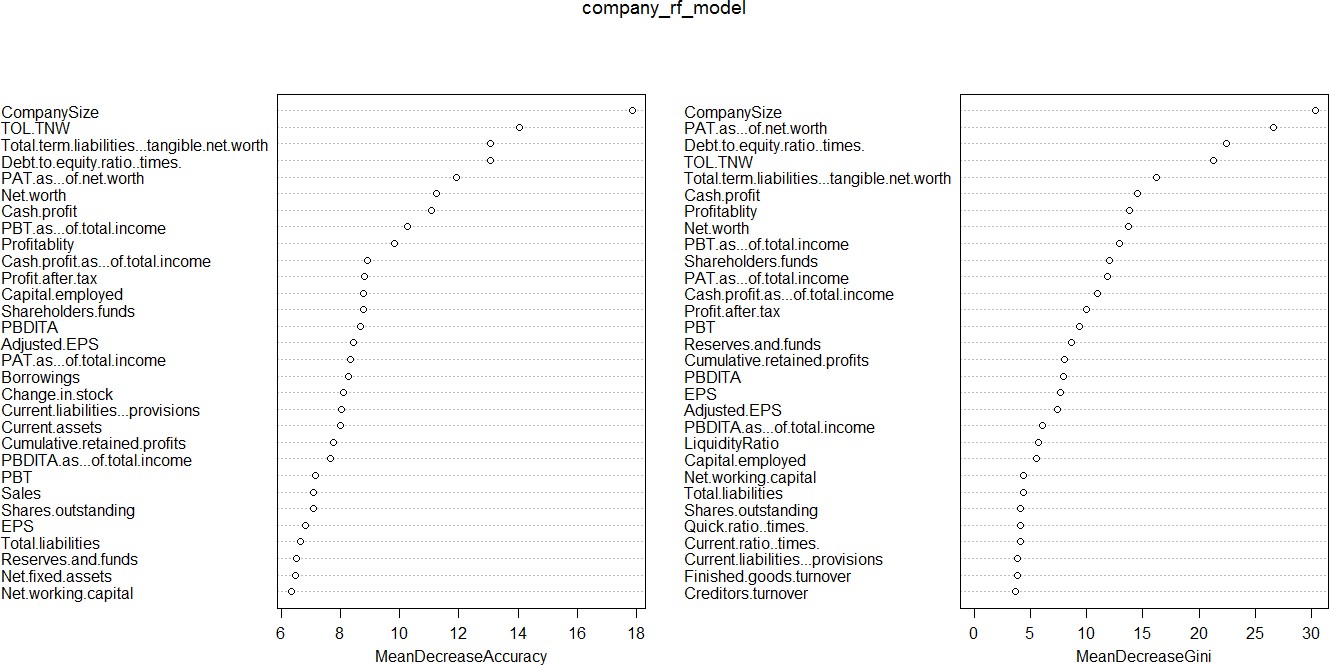
## Checking for Correlation and Multi Collinearity/Dimension Reduction

As there are several variables, it might not really make sense in doing a univariate and bivariate analysis for all the variables. It would be good first to check the correlation between each of the variables and then remove those variables which are highly correlated. The correlation matrix cannot be shown below due to the high number of variables and it is 51x51 matrix. A part of the matrix is shown below:



This matrix is further analyzed in Excel and all those variables which has a correlation of greater than 0.95 will be grouped together.

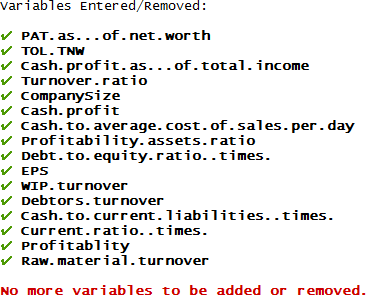
Also, the Decision Tree and Random Forest method was tried for identifying the important variables. The Mean Decrease Accuracy and Mean Decrease Gini plot would give a good idea of the significant variables.



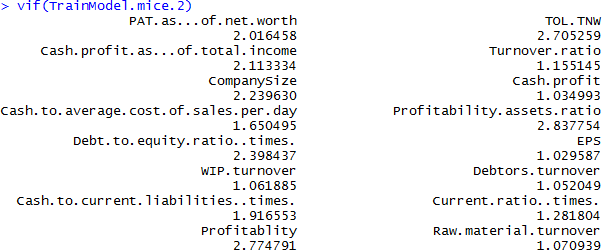
Along with this and the removal of VIFs we will be identifying one final set of variables for the Univariate and Bivariate Analysis.

As seen in most of the Credit Scoring methods, we can observe that there is only the significance of Ratios which is relevant for all the scoring methods. Hence here also we will be taking mainly the ratios and generating the model. Only these variables will be used for the Univariate and Bivariate Analysis.

Now, when we ran Logistic Model using BLR it selects us the adequate variables (Step Wise Selection) for the prediction and gives us the following output.



The VIF was checked across these variables and it was found that no high VIF has been observed. All the higher VIF variables has been eliminated and we have reached at the perfect selection of the variables.



With the above table we have ensured that all the high VIF values has been removed and we are good to go with the model.

For better understanding and to avoid clutter, the Univariate and Bivariate analysis is done based on the above variables only. We can understand that most of the variables obtained are ratio variables which makes sense as other standard credit scores like Altman Z and India Z also has ratio variables as their inputs.

## Univariate Analysis

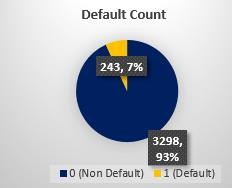
Though we are not left with 16 variables, we will be concerned with only some of the variables as per the problem statement - as most of the variables are not significant in predicting the Default. We will start off with a quick statistical summary for all the variables:

The default Variable is the only categorical variable here with summary as follows: (0) – 3,298; (1) – 243

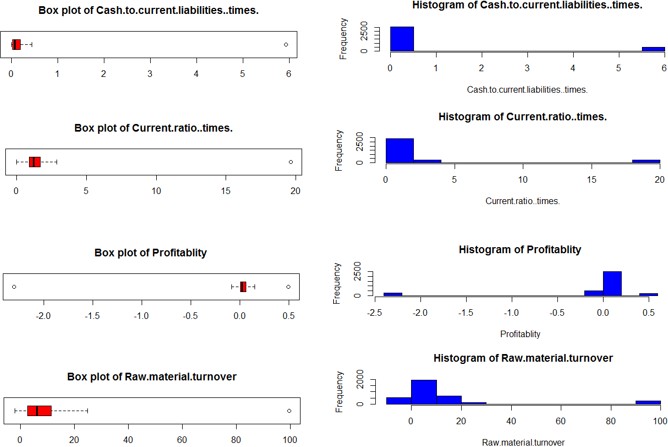
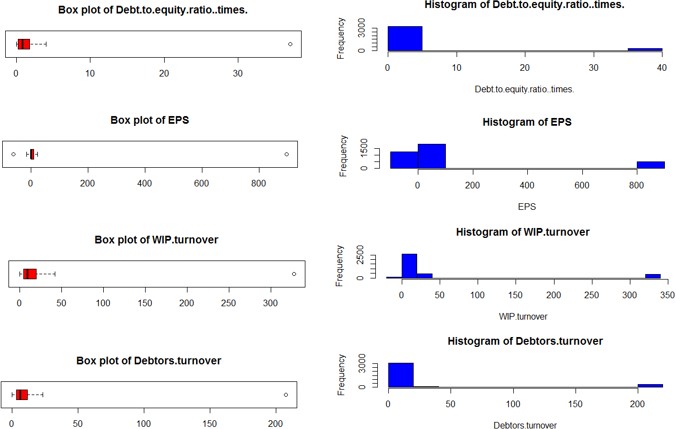
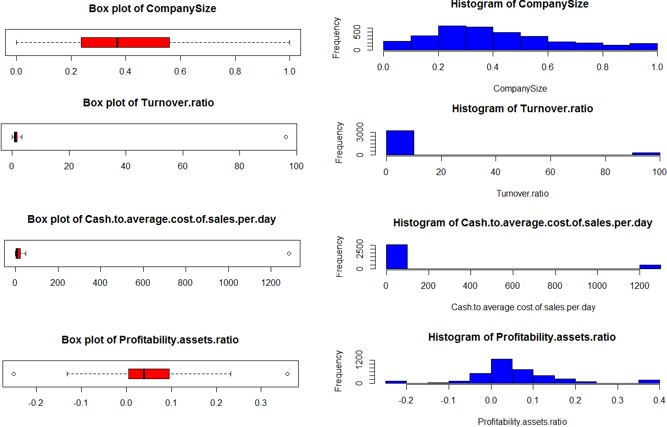
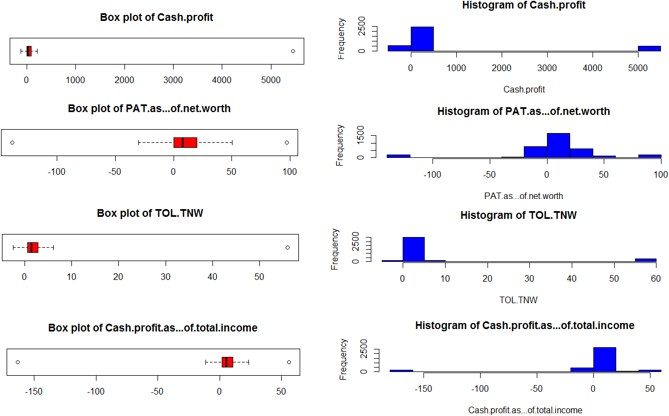
For the other continuous variables, following is the statistical summary:



Following shows the Pie Chart of the Dependent Variable:



Following shows the box plot and histograms for the continuous variables.

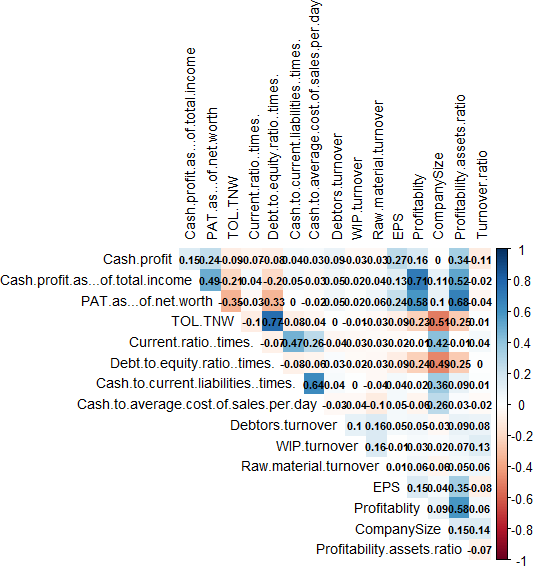
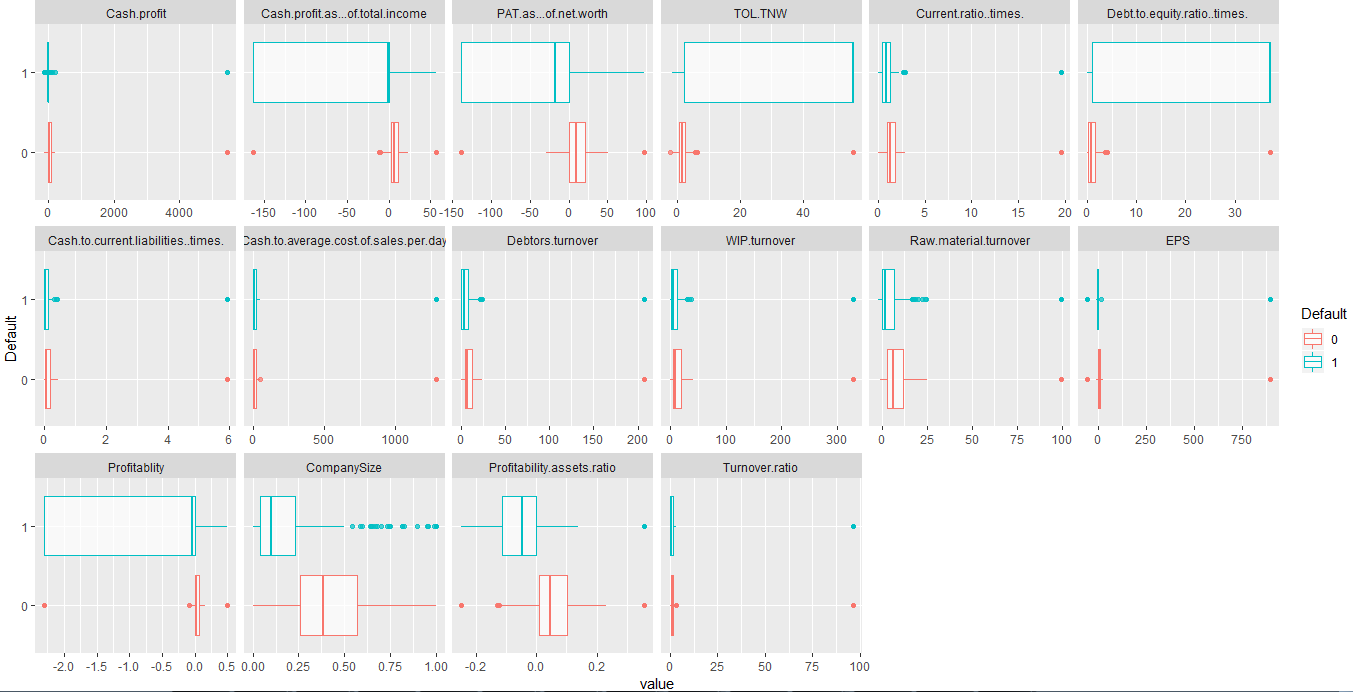


### Observations from Univariate Analysis

* + - * **Default** : This is the dependent variable which we will be using. Almost 93% of the Customers are non-defaulters while the remaining 7% are Defaulters.
      * **CashProfit**: The Cash Profits are in the range of -120 to 1000. There are a few outliers, and these are capped at 5000.
      * **PAT.as...of.net.worth**: The PAT.as...of.net.worth are in the range of 0 to 20. There are a few outliers, and these are capped at -138 and 97.
      * **TOL.TNW**: The TOL.TNW are in the range of 0 to 3. There are a few outliers, and these are capped at -2 and 55.
      * **Cash.profit.as...of.total.income**: The Cash.profit.as...of.total.income are in the range of 1 to 10.5. There are a few outliers, and these are capped at -163 and 56.
      * **CompanySize**: The CompanySize is in the range of 0 to 1. The Median Value is at 0.35
      * **Turnover.ratio**: The Turnover.ratio is in the range of 0 to 11. There are a few outliers, and these are capped at 0 and 96.
      * **Cash.to.average.cost.of.sales.per.day**: This is in the range of 0 to 1.75. There are a few outliers, and these are capped at 0 and 37.
      * **Debt.to.equity.ratio..times.**This is in the range of 0 to 21. There are a few outliers, and these are capped at 0 and 1284.
      * **Cash.to.current.liabilities..times.**: This is in the range of 0 to 21. There are a few outliers, and these are capped at 0 and 1284.
      * **Debtors.turnover**: This is in the range of 0 to 11. There are a few outliers, and these are capped at 0 and 207.
      * **WIP.turnover**: This is in the range of 0 to 20. There are a few outliers, and these are capped at -0.18 and 328.
      * **Raw.material.turnover**: This is in the range of 2 to 12. There are a few outliers, and these are capped at -2 and 100.
      * **EPS**: This is in the range of 0 to 10. There are a few outliers, and these are capped at - 60 and 896.
      * **Profitablity**: This is in the range of 0 to 0.06. There are a few outliers, and these are capped at 0 and 0.5.

## Bivariate and Multivariate Analysis

This Section will contain the Bivariate and Multivariate analysis of this dataset. Following shows the box plot and histograms for the continuous variables.



### Observations from Bivariate Analysis and Multi Variate Analysis

* + - * **CashProfit**: The Median of Cash Profit is on the lower side for the Defaulters v/s Non-Defaulters.
      * **Cash.profit.as...of.total.income**: Most of the values lies in the negative region for the defaulters and the median is also lesser than 0 for the defaulters. For the Non defaulters – most of the values are in the positive side and median is also positive.
      * **PAT.as...of.net.worth**: Most of the values lies in the negative region for the defaulters and the median is also lesser than 0 for the defaulters. For the Non defaulters – most of the values are in the positive side and median is also positive.
      * **TOL.TNW**: This is the ratio between the Liabilities and the Net Worth. For the defaulters as expected the Liabilities are more than the Net worth. Hence the values range above 5 and most of them have it that way. For the Non-Defaulters the median value is less than 5.
      * **Current ratio (times)**: The Defaulters Median value is lesser than the Non-Defaulters. This implies the current assets is lower than their current liabilities for the Defaulters.
      * **Debt to equity ratio (times)**: The Defaulters values are much higher than the non-defaulters. This implies that the liabilities are much greater than the shareholders equity values.
      * **Debtors turnover:** Net credit sales divided by average accounts receivable. The values are much lower than the non-defaulters. This implies that the sales are low as compared to the accounts receivables.
      * **Raw material turnover:** The values are much lower than that of the non-defaulters. This implies that the cost of goods sold is much lesser than the average inventory.
      * **Profitability**: The value is much lower or rather in the negative region for defaulters as compared to the non-defaulters.
      * **Company Size**: The defaulters’ company’s net worth is much lower than their total assets. The median ratio is below 0.1 which is alarming. For the Non defaulters the median ratio is around 0.35
      * **Profitability Assets Ratio**: The value is much lower or rather in the negative region for defaulters as compared to the non-defaulters.
      * **Correlation:** From the Correlation table also a lot of insights can be figured out. The TOL-TNW and Debt to Equity Ratio is correlated while Cash Profit as of Income and Profitability are correlated. Both have the coefficients higher than 0.6
      * TOL-TNW and Debt to Equity ratio are negatively correlated with Company Size.

# 4 Logistic Regression

The Logistic Regression is done using the “glm” function in R.

## Methodology – without Smote

As there are almost 51 variables, it might not be suitable to check the Wald Test and Significance of all the variables. We will run the Logistic Regression model using all the variables and use the option  *Maxit = 100.* This is to make sure that the Logistic Regression Algorithm converges. As discussed before the main variables will be the ratios which is being used with some of the absolute variables in place.

The code and the variables which are used as below:

*TrainModel.mice.2=glm(Default~ PAT.as...of.net.worth+TOL.TNW+ Cash.profit.as...of.total.income+Turnover.ratio+ CompanySize+Cash.profit+Cash.to.average.cost.of.sales.per.day+ Profitability.assets.ratio+Debt.to.equity.ratio..times.+ EPS+WIP.turnover+Debtors.turnover+ Cash.to.current.liabilities..times.+Current.ratio..times.+ Profitablity+Raw.material.turnover,*

*data=company.mice.2.ot,family = binomial,control = list(maxit = 1000)) summary(TrainModel.mice.2)*

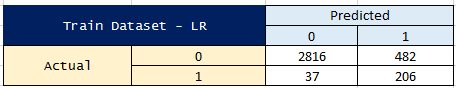
*vif(TrainModel.mice.2)*

The model is run, and the values are predicted. If the predictive score is higher than 0.5 than the Predictive class is taken as 1.

At first the model is run without using Smote. If need arises Smote will be done and data will be predicted accordingly.

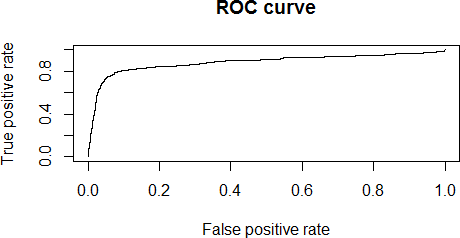
## Results and Model Validation (without Smote)

### Train Dataset

The Logistic model first is checked with the train model itself. Checking the Confusion Matrix on the train model, we get the following:

-

The Accuracy Rate is at 85%. The Sensitivity is at 85% while the Specificity is also at 85%. This implies that we have been mostly identifying the Defaulters and Non-Defaulters 85% of the time correctly. This is a decent model to go ahead with.



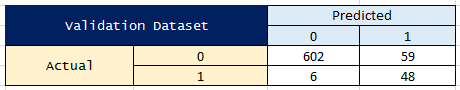
The figure above shows the area under the curve which is close to 1. This implies the model is a good one. Larger the area under the curve (AUC) the best the model is.

The AUC obtained is **0.8829. This is a fine AUC.**

The model is further checked for Gini and KS.

The Gini value is at 0.73 and the KS value is at 0.71. Both the values suggest that this is a good model. Due to this specific reason, SMOTE is ruled out and we go ahead with the Validation data.

### Validation Dataset

The logistic regression then is checked on the validation model. Checking the Confusion Matrix on the test model, we get the following:

The Accuracy Rate is at 91%. The Sensitivity is at 89% while the Specificity is at 91%. This implies that its fairly a good prediction.

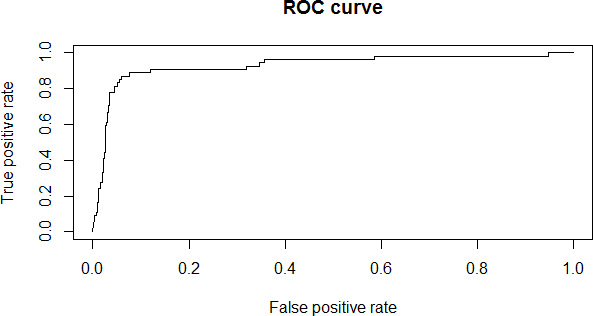


Figure 10 ROC Curve for the validation dataset

The figure above shows the area under the curve which is close to 1. This implies the model is a good one. Larger the area under the curve (AUC) the best the model is.

The AUC value is at 0.93 which indicates a very good prediction. The Model was validated using the KS Stats also.

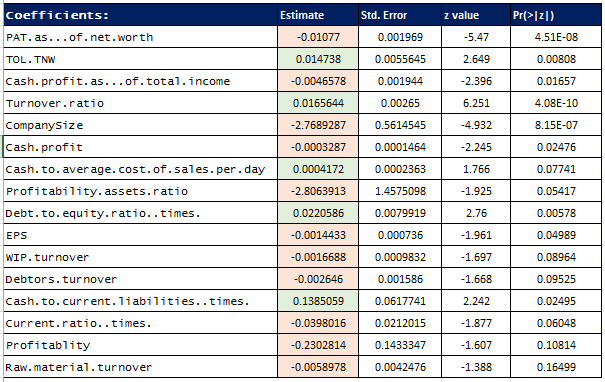


The output (KS = 0.81) shows that the model performance is very good based on the KS stats. The Model was validated using GINI too. How well the data performs on random predictions.



The Gini Score at 0.78 shows very high which implies it has performed well on Gini too.

## Analyzing Coefficients

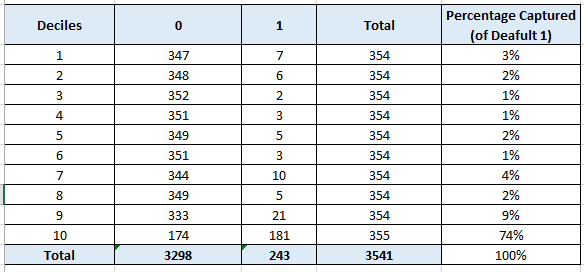
The below table shows the list of coefficients obtained on the model Their relative weights also can be seen by the estimates.

We can understand that the most of them are negative coefficients. That is the lower that values of these the higher the chances of default. This can be mainly seen on the following variables – Profitability.assets.ratio, CompanySize, Profitability, Current.ratio..times and PAT.as...of.net.worth. Hence the higher the decrease in here, the higher the chances of the default.

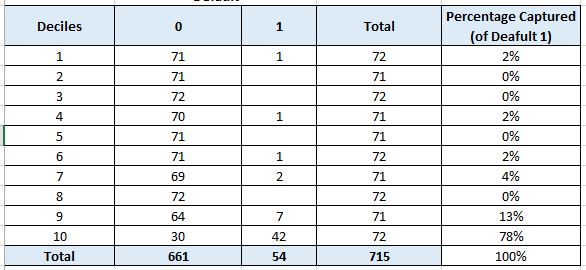
On the positive side, if the variables TOL.TNW, Turnover.ratio, Debt.to.equity.ratio..times. Cash.to.current.liabilities..times are higher there are high chances of default.

## Probability of Default and Deciles

The Probability of Default were calculated, and the Deciles were obtained for each of it. The following table shows the distribution of the Deciles and the Prediction Accuracy:



Decile Analysis on the Train Dataset



Decile Analysis on the Validation Dataset

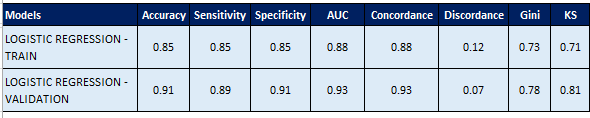
For the Validation Dataset 78% + 13% = 91% of the defaulters are captured in the last 2 deciles which shows the goodness of the model. While for the Train Dataset this at 85%. Both are very good measures for the model. Hence the model can be said to be a very good model. In all the remaining deciles there is relatively very less values on the Default. This implies only in the top deciles most of the Defaulters lie and hence the loans can be accordingly given. For example, seeing the validation dataset – if we plan to rule out the loans for 10th Decile – we are sure that 60% of the companies will default. The remaining 40% (30 companies) of the companies must go unlucky here. But that is just 5% of the overall companies. For running the business, we are still working out with the remaining 95% of the companies.

Similarly, in the train dataset – 174 non defaulters are there in the 10th decile. This is just 5% of the companies where we will be losing business with. We will be able to confidently give loans and do business with all the other deciles. This is the methodology which the financial firms use in doing business with.

# Model Performance Summary and Insights/Conclusions

The following are the observations and conclusions:

* + Around 51 variables were there. New ratios were introduced. Finally, by checking the significance it was observed that the ratios were more significant than the absolute values.
  + In the final equation too, most of the variables were ratios. This can be understood from all the standard credit scores such as Altman Z and India Z score where all are ratios.
  + Logistic Regression was used for building the predictive models. The accuracy has increased by 6% for the Validation Dataset (v/s Train Dataset). The accuracy of the train dataset is at 85% while for the test dataset is at 91%.
  + The Sensitivity for the validation dataset is also higher means better predicting at 89% (v/s 85% for the train dataset)
  + The KS value is at 81% for the validation model. This implies that it can separate out the positives and negatives very clearly.
  + The Gini Value ranges from 0.78 for the validation dataset while 0.71 for the train dataset. Ideally for a good model it should be 0.6. This implies that the model is good.
  + The Decile summary based on the probability of default also shows very good prediction of the model. For the Validation Dataset 78% + 13% = 91% of the defaulters are captured in the last 2 deciles which shows the goodness of the model. While for the Train Dataset this at 85%. Both are very good measures for the model. Hence the model can be said to be a very good model.
  + On Analyzing the Coefficients:
    - We can understand that the most of them are negative coefficients. That is the lower that values of these the higher the chances of default. This can be mainly seen on the following variables – Profitability.assets.ratio, CompanySize, Profitability, Current.ratio..times and PAT.as...of.net.worth. Hence the higher the decrease in here, the higher the chances of the default.
    - On the positive side, if the variables TOL.TNW, Turnover.ratio, Debt.to.equity.ratio..times. Cash.to.current.liabilities..times are higher there are high chances of default.
  + The Overall conclusion table is as shown below:



Summary on Model Validation Parameters – Logistic Regression Method

# Appendix – Source Code

# 