Credit Risk(default) Modelling – Logistic Regression

By Aparajita Mukherjee

Table of Content

1. Project Introduction

Business owners use ratio analysis to determine the financial well-being of their companies. Ratio analysis provides an objective measure of the financial effectiveness of its marketing strategies. Ratio analysis is also used by banks and financial institutions to determine the credit worthiness of companies before loans are approved.

The objective of this predictive modelling is:

- Gauge a company's capability to pay back their debt/borrowings.

2. Required Packages

Library	**Description**
library(DataExplorer)	Data Visualisation
library(readxl)	To import a file with .xlsx extension.
library(Hmisc)	The goal of 'readr' is to provide a fast and friendly way to read rectangular data (like 'csv', 'tsv', and 'fwf').
library(naniar)	Functionality to create pretty word clouds, visualize differences and similarity between documents, and avoid over-plotting in scatter plots with text.
library(nFactors)	Principal Component Analysis
library(psych)	Used for rotation while performing variable reduction.
library(VIM)	Recursive partitioning for classification, regression and survival trees.
library(plotly)	Plot random forest model.
library(DMwR)	Classification and regression based on a forest of trees using random inputs.
library(car)	Provides a number of user-level functions to work with "grid" graphics.
Library(ggplot2)	For Visualization
library(ROCR)	
library(ineq)	
library(InformationValue)	
Library(viridis)	Colour palette

3. Data Dictionary

Variable Name	**Discerption**	**Category**
Net worth Next Year	Net worth of the customer in next year	Dependent Variable
Total assets	Total assets of customer	Size
Net worth	Net worth of the customer of present year	Size
Total income	Total income of the customer	Size

Change in stock	difference between value of current stock and the value of stock in last trading day	Change in Size
Total expenses	Total expense done by customer	Size/Costs
Profit after tax	Profit after tax deduction	Profit
PBDITA	Profit before depreciation, income tax and amortization	Profit
PBT	Profit before tax deduction	Profit
Cash profit	Total Cash profit	Profit
PBDITA as % of total income	PBDITA / Total income	Profit
PBT as % of total income	PBT / Total income	Profit
PAT as % of total income	PAT / Total income	Profit
Cash profit as % of total income	Cash Profit / Total income	Profit
PAT as % of net worth	PAT / Net worth	Profit
Sales	Sales done by customer	Size
Income from financial services	Income from financial services	Profit
Other income	Income from other sources	Profit
Total capital	Total capital of the customer	Size
Reserves and funds	Total reserves and funds of the customer	Profit
Deposits (accepted by commercial banks)	All blank values	Profit/Size
Borrowings	Total amount borrowed by customer	Leverage
Current liabilities & provisions	current liabilities of the customer	Liquidity
Deferred tax liability	Future income tax customer will pay because of the current transaction	Liquidity
Shareholders funds	Amount of equity in a company, which is belong to shareholder	Size
Cumulative retained profits	Total cumulative profit retained by customer	Profit
Capital employed	Current asset minus current liabilities	Size
TOL/TNW	Total liabilities of the customer divided by Total net worth	Leverage
Total term liabilities / tangible net worth	Short + long term liabilities divided by tangible net worth	Leverage
Contingent liabilities / Net worth (%)	Contingent liabilities / Net worth	Leverage
Contingent liabilities	Liabilities because of uncertain events	Liquidity
Net fixed assets	purchase price of all fixed assets	Size
Investments	Total invested amount	Size
Current assets	Assets that are expected to be converted to cash within a year	Size
Net working capital	Difference of current liabilities and current assets	Liquidity/Size
Quick ratio (times)	Total cash divided by current liabilities	Liquidity
Current ratio (times)	Current assets divided by current liabilities	Leverage
Debt to equity ratio (times)	Total liabilities divided by its shareholder equity	Liquidity
Cash to current liabilities (times)	Total liquid cash divided by current liabilities	Liquidity
Cash to average cost of sales per day	Total cash divided by average cost of the sales	Liquidity
Creditors turnover	Net credit purchase divided to average trade creditors	Liquidity

Debtors turnover	Net credit sales divided by average accounts receivable	Liquidity
Finished goods turnover	Annual sales divided by average inventory	Liquidity
WIP turnover	The cost of goods sold for a period divided by the average inventory for that period	Liquidity
Raw material turnover	Cost of goods sold is divided by the average inventory for the same period	Liquidity
Shares outstanding	Number of issued shares minus the number of share held in the company	Size
Equity face value	cost of the equity at the time of issuing	Size
EPS	Net income divided by total number of outstanding share	Profit
Adjusted EPS	Adjusted net earning divided by the weighted average number of common share outstanding on a diluted basis during the plan year	Profit
Total liabilities	Sum of all type of liabilities	Leverage
PE on BSE	Company current stock price divided by its earning per share	Market Segment

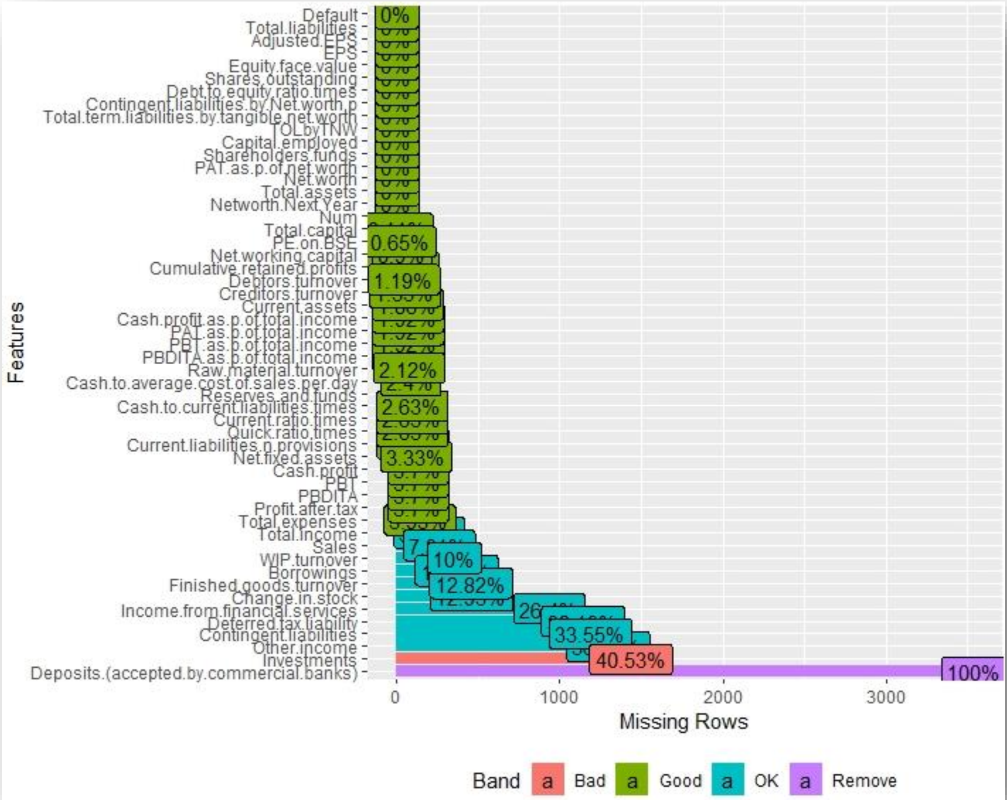
4. Basic EDA(Exploratory Data Analysis)

4.1 Add "Default Variable

```
> names(train)
 [1] "Num"                                "Networth.Next.Year"
 [3] "Total.assets"                       "Net.worth"
 [5] "Total.income"                       "PBDITA.as.p.of.total.income"
 [7] "PBT.as.p.of.total.income"           "PAT.as.p.of.total.income"
 [9] "Cash.profit.as.p.of.total.income"   "PAT.as.p.of.net.worth"
[11] "TOLbyTNW"                           "Total.term.liabilities.by.tangible.net.worth"
[13] "Contingent.liabilities.by.Net.worth.p" "Net.working.capital"
[15] "Debt.to.equity.ratio.times"         "Cash.to.average.cost.of.sales.per.day"
[17] "Creditors.turnover"                 "Debtors.turnover"
[19] "EPS"                                "PE.on.BSE"
[21] "Default"
```

A "Default variable has to be added to the train dataset using the Networth Next Year variable. But, the same variable will not be included while building the regression model in order to avoid high multicollinearity.

4.2 Missing Values



Missing values in the dataset need to be treated before proceeding to outlier treatment.
Replace data having "NA" as value with "0". Replace missing values of total income using the formula. The remaining missing value that is now an insignificant value will be omitted from the dataset.

4.3 Convert to correct Data Type

The "str" function shows that there are variables that are in the incorrect data type format. Therefore:

- Creditors Turnover, Debtors Turnover and PE on BSE is converted to "numeric" datatype.
- Default variable is converted to "factor" datatype.

```
$ Investments                                   : num  NA NA 0.7 NA NA NA 1
$ Current.assets                                : num  560 407 148 536 472
$ Net.working.capital                           : num  134.2 123.6 -97.1 99
$ Quick.ratio.(times)                           : num  0.92 0.48 0.32 0.51
$ Current.ratio.(times)                         : num  1.31 1.39 0.6 1.23 1
$ Debt.to.equity.ratio.(times)                  : num  0.64 1.61 0.15 2.6 0
$ Cash.to.current.liabilities.(times)           : num  0.09 0.03 0.04 0.08
$ Cash.to.average.cost.of.sales.per.day         : num  7.56 3.88 4.63 3.71
$ Creditors.turnover                            : chr  "5.94" "10.59" "2.35
$ Debtors.turnover                              : chr  "5.74" "6.03" "9.6"
$ Finished.goods.turnover                       : chr  "25.11" "28.96" "8.2
$ WIP.turnover                                  : chr  "20.010000000000002"
$ Raw.material.turnover                         : chr  "17.579999999999998"
$ Shares.outstanding                            : chr  "4800000" "11400000"
$ Equity.face.value                             : chr  "10" "10" "100" "10"
$ EPS                                           : num  18.6 1.65 -90.39 -7.
$ Adjusted.EPS                                  : num  18.6 1.65 -90.39 -7.
$ Total.liabilities                             : num  971 675 532 858 823
$ PE.on.BSE                                     : chr  "NA" "NA" "-15.5" "-
```

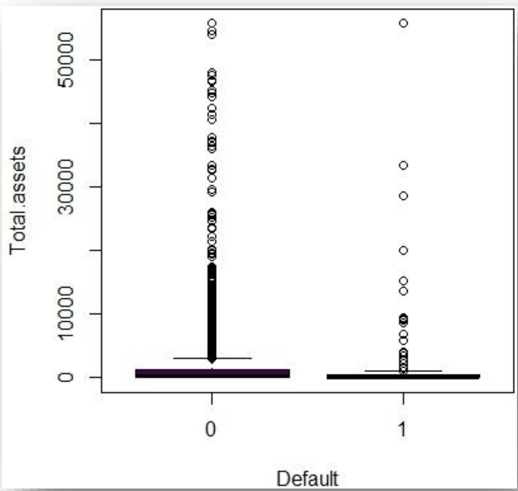
```
> str(train)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':        3268 obs. of  21 variables
 $ Num                                          : num  1 2 3 5 6 8 9 11 12
 $ Networth.Next.Year                           : num  8890.6 394.3 92.2 10
 $ Total.assets                                 : num  17512 941 233 478 24
 $ Net.worth                                    : num  7093 352 101 108 676
 $ Total.income                                 : num  24965 1527 477 1580
 $ PBDITA.as.p.of.total.income                  : num  11.46 18.53 1.22 1.9
 $ PBT.as.p.of.total.income                     : num  9.68 12.33 -1.38 0.4
 $ PAT.as.p.of.total.income                     : num  6.18 7.54 -1.38 0.35
 $ Cash.profit.as.p.of.total.income             : num  7.5 10.38 0.06 0.75
 $ PAT.as.p.of.net.worth                        : num  23.78 38.08 -6.35 5.
 $ TOLbyTNW                                     : num  1.33 1.23 1.44 2.83
 $ Total.term.liabilities.by.tangible.net.worth: num  0 0.34 0.29 1.59 0.3
 $ Contingent.liabilities.by.Net.worth.p        : num  14.8 19.2 45.8 34.9
 $ Net.working.capital                          : num  3588.5 203.5 59.6 21
 $ Debt.to.equity.ratio.times                   : num  0 0.78 0.35 1.79 1.0
 $ Cash.to.average.cost.of.sales.per.day        : num  68.21 5.96 17.07 0 1
 $ Creditors.turnover                           : num  3.62 9.8 5.28 13 6.5
 $ Debtors.turnover                             : num  3.85 5.7 5.07 9.46 2
 $ EPS                                          : num  35.52 9.97 -0.5 7.91
 $ PE.on.BSE                                    : num  27.31 8.17 -5.76 0 0
 $ Default                                      : Factor w/ 2 levels "0","1
 - attr(*, "na.action")= 'omit' Named int  4 7 10 37 101 106 107 120 143 1
  ..- attr(*, "names")= chr  "4" "7" "10" "37" ...
```

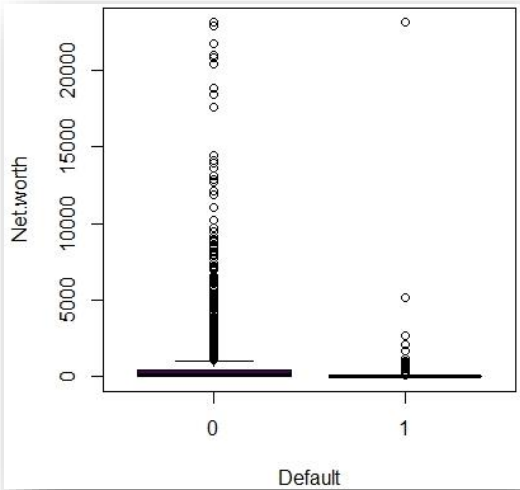
4.4 Outlier Treatment

	n	nmiss	outlier_flag	mean	stdev	min	p1.1%	p99.99%	max	UC	LC
Num	3268	0	0	1761.755508	1018.76292	1.00	37.3400	3506.3300	3544.00	4818.04428	-1294.53327
Networth.Next.Year	3268	0	1	1738.280753	18167.60123	-74265.60	-89.2060	26139.3680	805773.40	56241.08444	-52764.52293
Total.assets	3268	0	1	3692.257283	32209.78305	0.30	5.0340	55820.2200	1176509.20	100321.60642	-92937.09186
Net.worth	3268	0	1	1392.479070	13929.80762	0.10	1.3000	23127.1110	613151.60	43181.90192	-40396.94378
Total.income	3268	0	1	4689.174939	56312.36208	0.10	0.6000	44433.5240	2442828.20	173626.26118	-164247.91131
PBDITA.as.p.of.total.income	3268	0	1	6.435398	101.51295	-2900.00	-61.2596	76.2977	100.00	310.97426	-298.10347
PBT.as.p.of.total.income	3268	0	1	-14.377870	402.28385	-21340.00	-214.6497	47.2500	96.00	1192.47369	-1221.22943
PAT.as.p.of.total.income	3268	0	1	-16.542653	407.78941	-21340.00	-211.8646	38.9231	96.00	1206.82558	-1239.91088
Cash.profit.as.p.of.total.income	3268	0	1	-6.342225	291.31066	-15020.00	-127.0264	51.1793	100.00	867.58975	-880.27420
PAT.as.p.of.net.worth	3268	0	1	11.317570	66.61432	-748.72	-133.6732	97.8288	2466.67	211.16053	-188.52539
TOLbyTNW	3268	0	1	3.542249	15.85412	-350.48	0.0000	44.1112	411.27	51.10460	-44.02010
Total.term.liabilities.by.tangible.net.worth	3268	0	1	1.401031	10.67645	-325.60	0.0000	19.7320	292.02	33.43039	-30.62833
Contingent.liabilities.by.Net.worth.p	3268	0	1	51.435428	338.72123	0.00	0.0000	545.2563	14704.27	1067.59912	-964.72826
Net.working.capital	3268	0	1	146.929590	3047.69386	-63839.00	-1780.2320	3985.1770	85782.80	9290.01117	-8996.15199
Debt.to.equity.ratio.times	3268	0	1	2.340168	10.45242	0.00	0.0000	27.7198	341.18	33.69742	-29.01708
Cash.to.average.cost.of.sales.per.day	3268	0	1	125.233880	2692.15493	0.00	0.0000	716.4800	128040.76	8201.69866	-7951.23090
Creditors.turnover	3268	0	1	15.049183	67.61527	0.00	0.0000	135.0888	2401.00	217.89500	-187.79663
Debtors.turnover	3268	0	1	15.716968	62.89787	0.00	0.0000	198.9774	2473.04	204.41057	-172.97663
EPS	3268	0	1	-226.643501	14829.41281	-843181.82	-62.6006	1141.9311	34522.53	44261.59494	-44714.88194
PE.on.BSE	3268	0	1	27.198739	920.49329	-435.84	-37.1095	162.9211	51002.74	2788.67860	-2734.28112

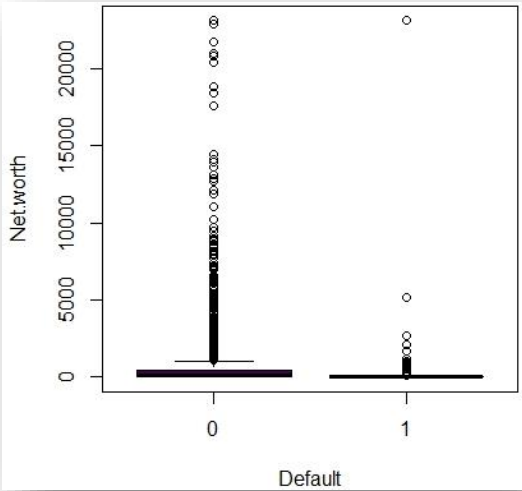
The existence of outliers can indicate individuals or groups that have behaviour very different from the most of the individuals of the dataset. We will remove outliers to improve the accuracy of estimators, by capping the lower limit of the dataset at 0.01 quartile and the upper limit of the dataset at 0.99 quartile.

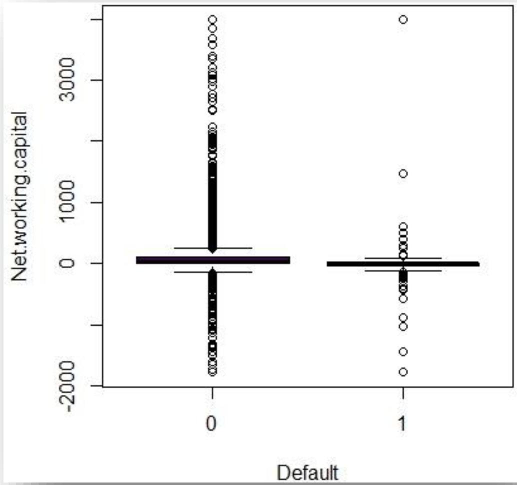
4.5 Univariate Analysis



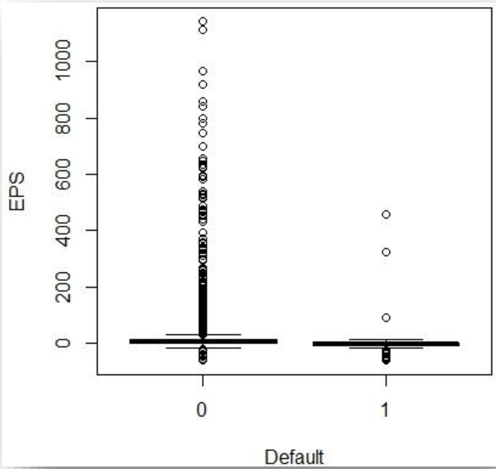


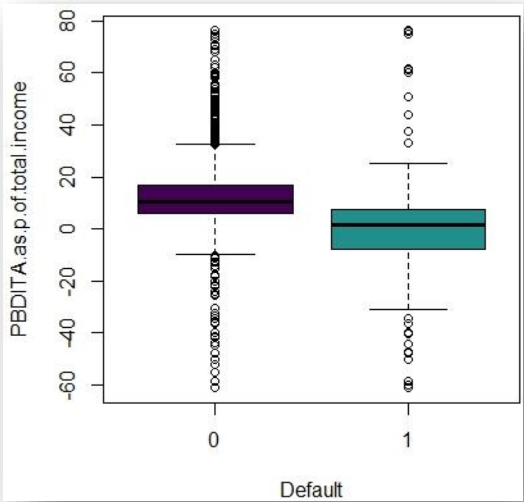
The variables are categorized into indicating the size of the company. The larger the size of a firm, the less likely it is to default.



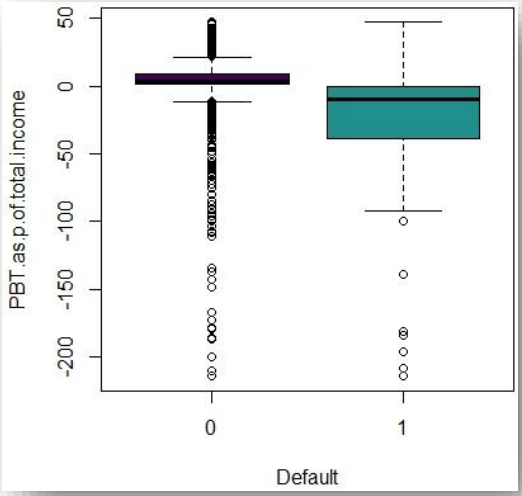


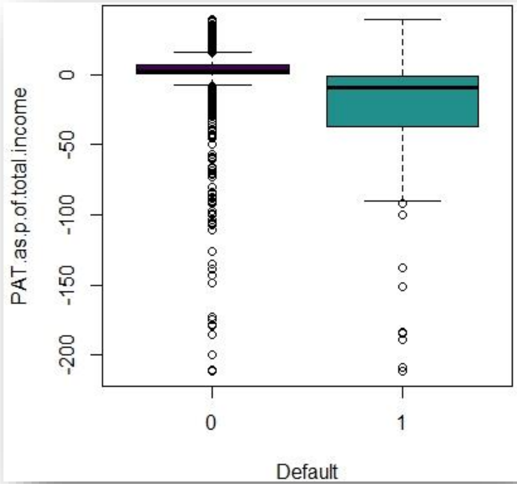
The companies having more assets/backing are less likely to default as compared to companies with lesser assets.



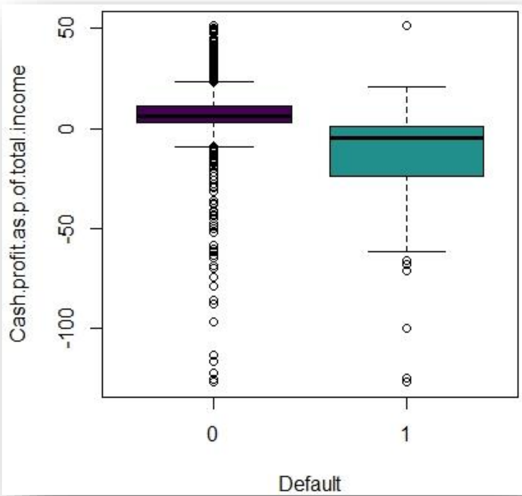


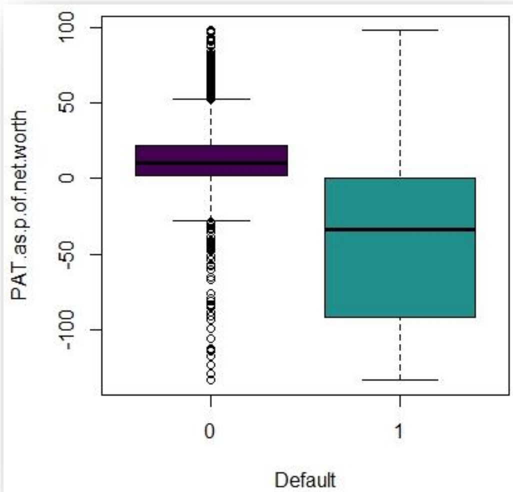
Higher the Profit percentage, lesser are the company's likelihood to default.

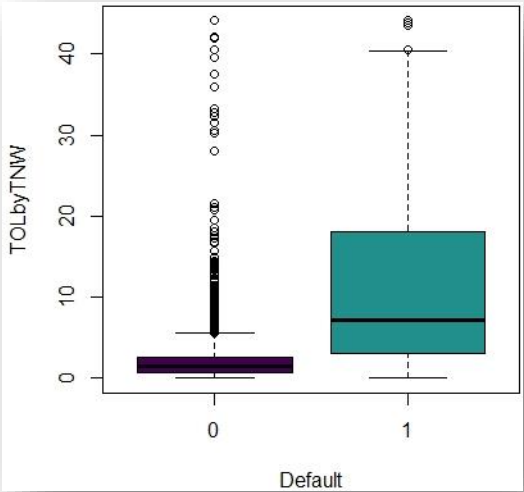


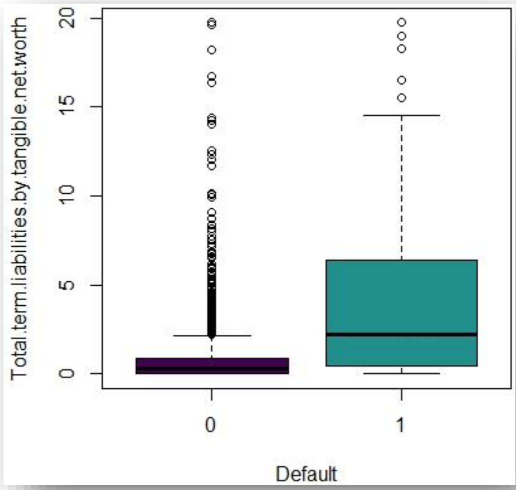


The ratios(PBDITA.as.p.of.total.income, PBT.as.p.of.total.income, PAT.as.p.of.total.income, Cash.profit.as.p.of.total.income, PAT.as.p.of.net.worth, EPS) fall into the category "Profitability". The five plots indicate that, higher the profit ratios of a company, the less likely it is to default on its credit.

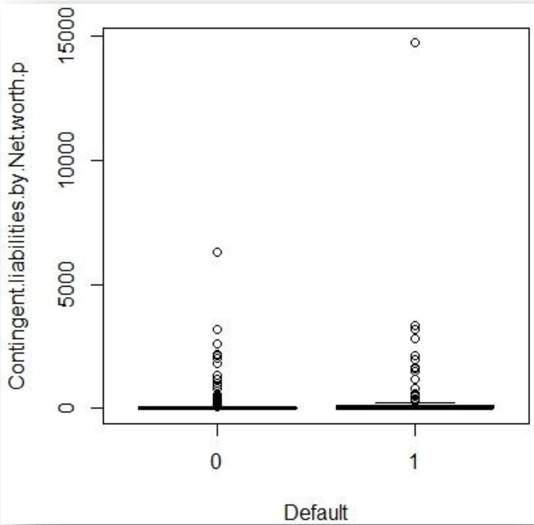


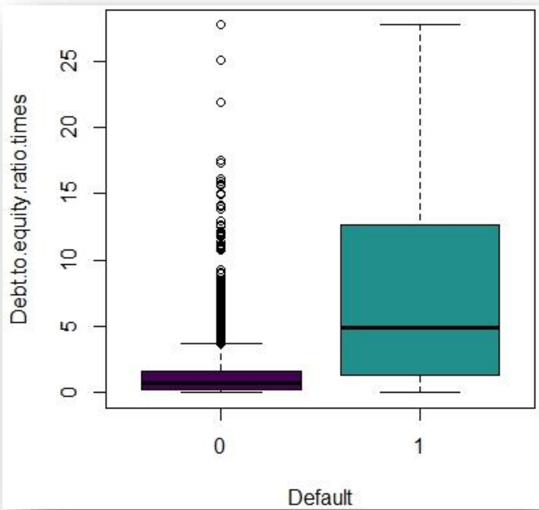


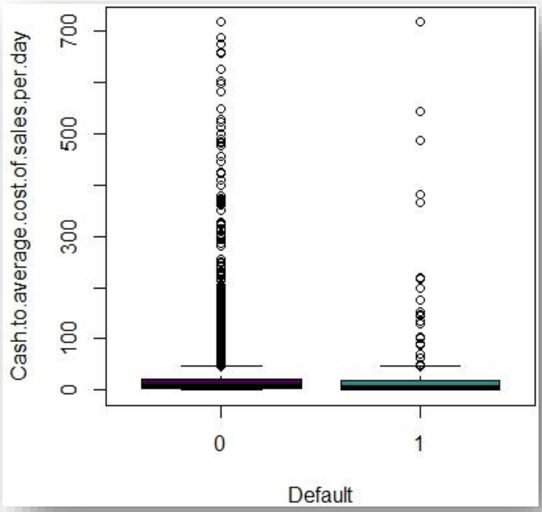


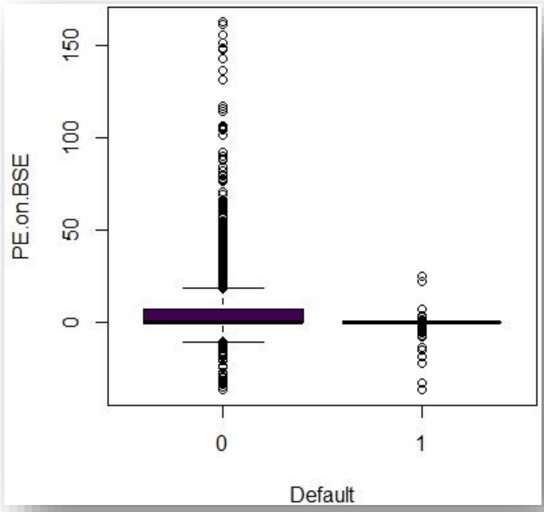


The ratios(TOLbyTNW, Total.term.liabilities.by.tangible.net.worth, Contingent.liabilities.by.Net.worth, Debt.to.equity.ratio.times) fall under the category of Leverage. It refers to the amount of debt a firm uses to finance assets. "Highly leveraged," means that the item has more debt than equity. Therefore, companies with lower leverage ratio will have a lesser propensity to default.



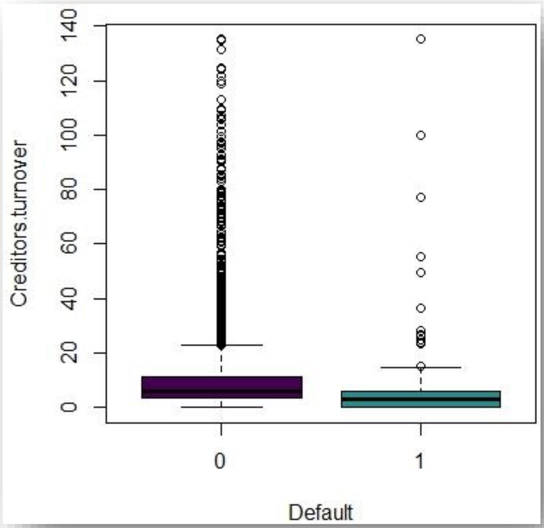


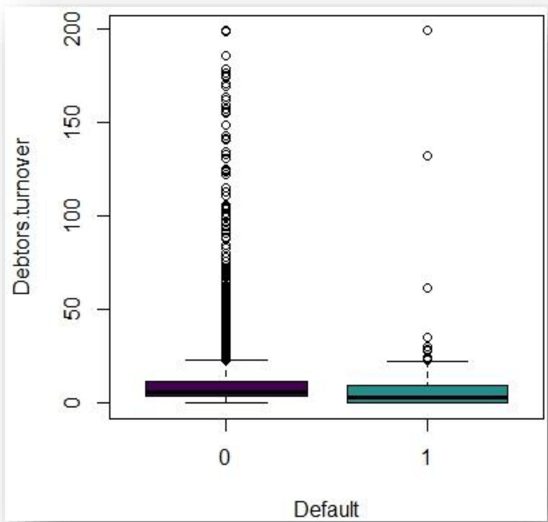




The higher the ratio for a company, the less likely it is to default on its credit.

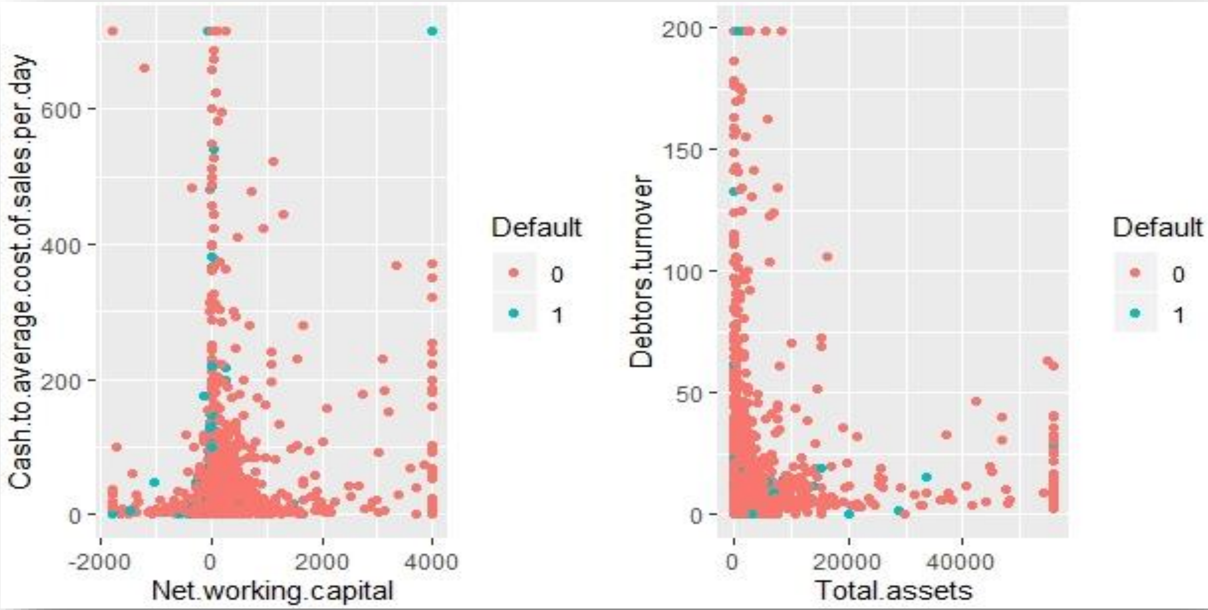
The Price-Earning Ratio indicates the valuation of a company. Therefore, higher the valuation, lesser is the propensity to default.



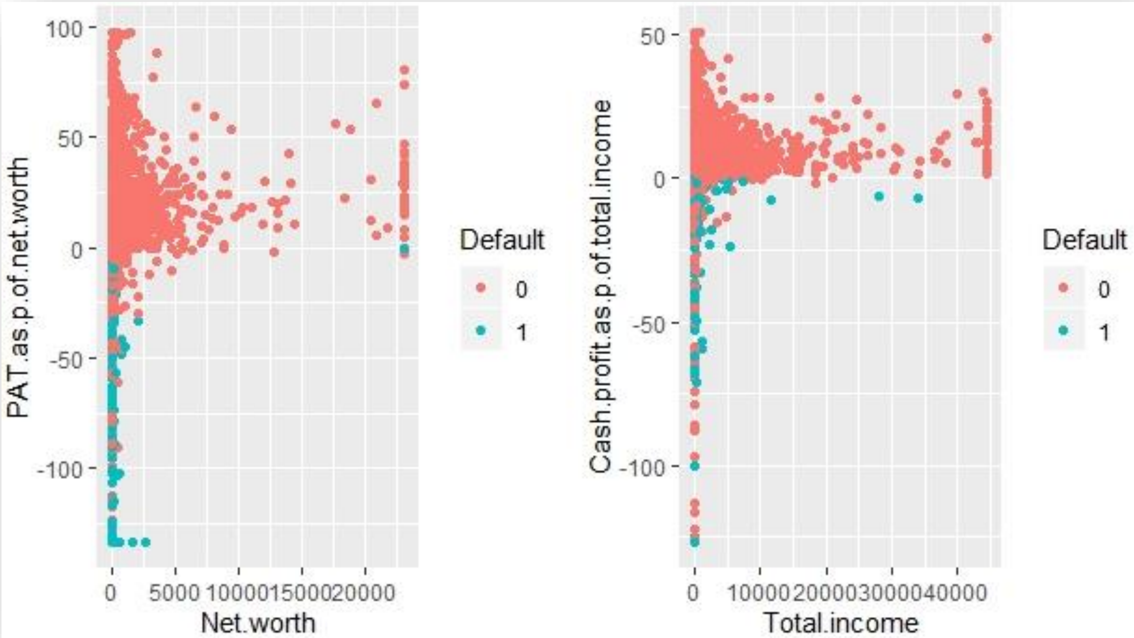


The companies having higher Liquidity ratio means that a company is capable of paying off its suppliers/effectively collect it's receivables. This more capable a company is, the less likely it is to default.

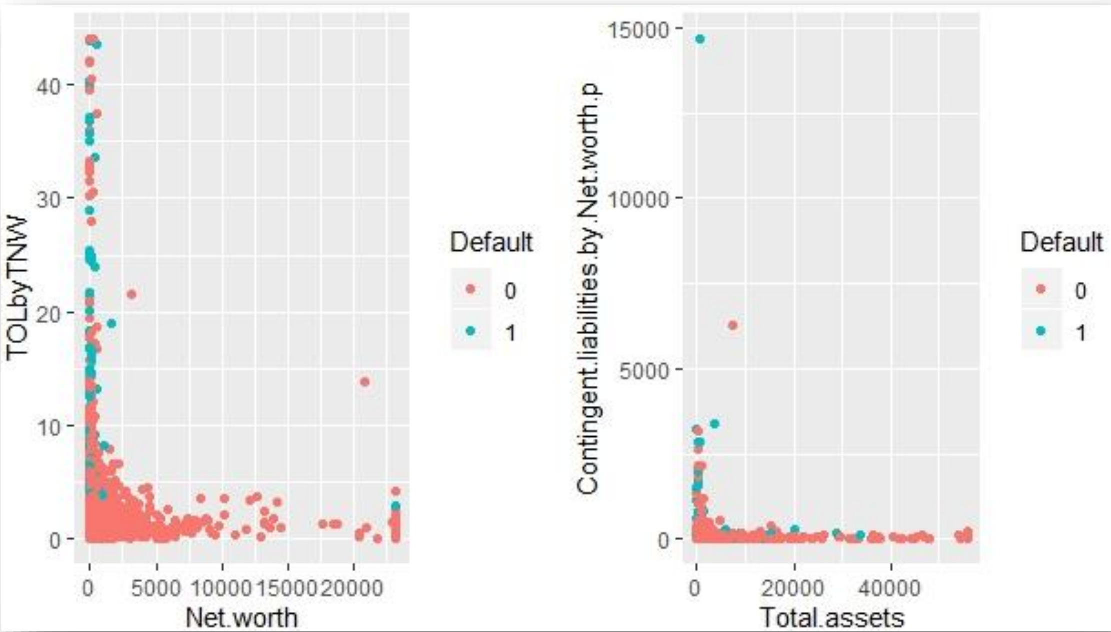
4.6 Bivariate Analysis



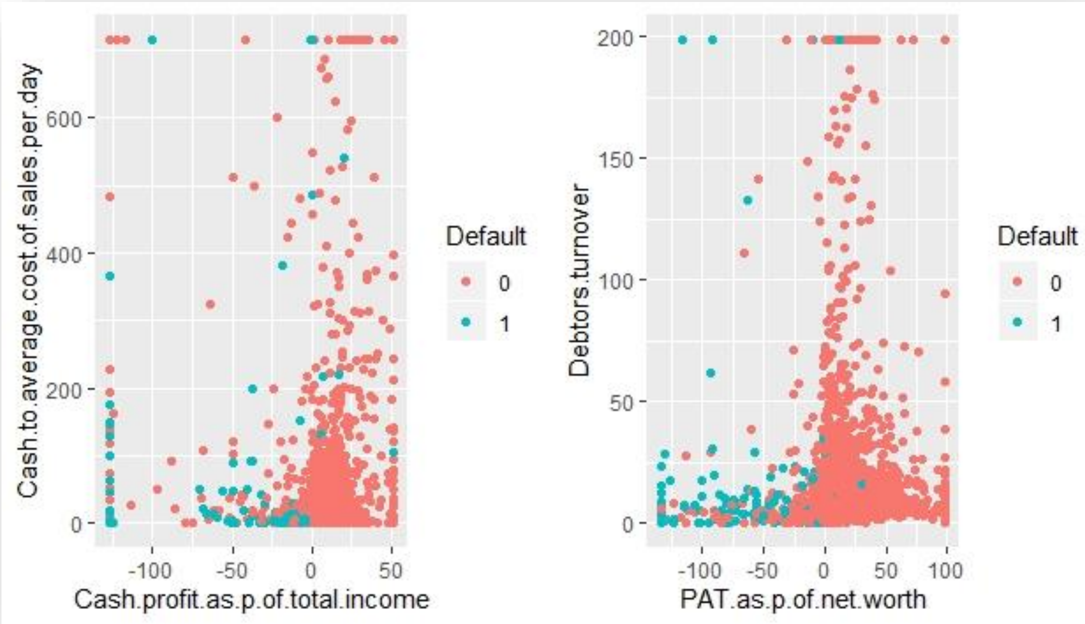
The above plots are a comparison between a "Size" variable and a variable from the "Liquidity" category of company ratios. It indicates that companies with smaller size and lesser the liquidity ratio have a higher tendency to default.



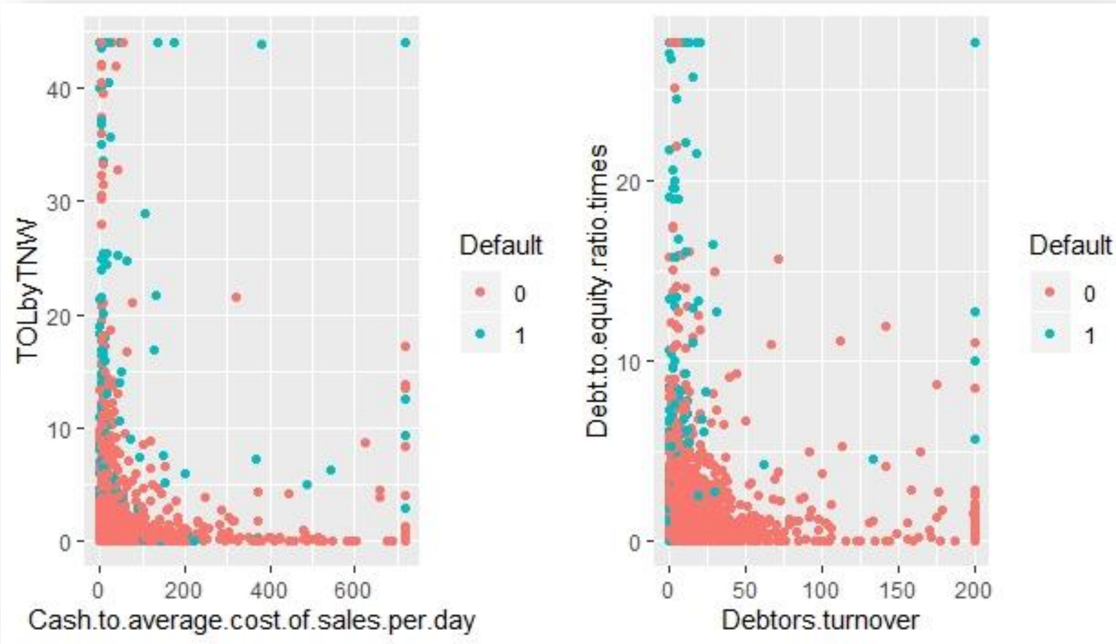
The plot is a comparison between the Size and Profitability of companies. It clearly shows that smaller the size and lesser the profitability of a company, the more likely they are to default.



The plot is a comparison between the Size and Leverage variables of companies. It clearly indicates that smaller the size and higher the leverage strategy of a company, they are more inclined to default.



The plot is a comparison between the Profitability and Liquidity variables of companies. Lesser the profits and liquidity ratio of a company, more the companies are likely to default.



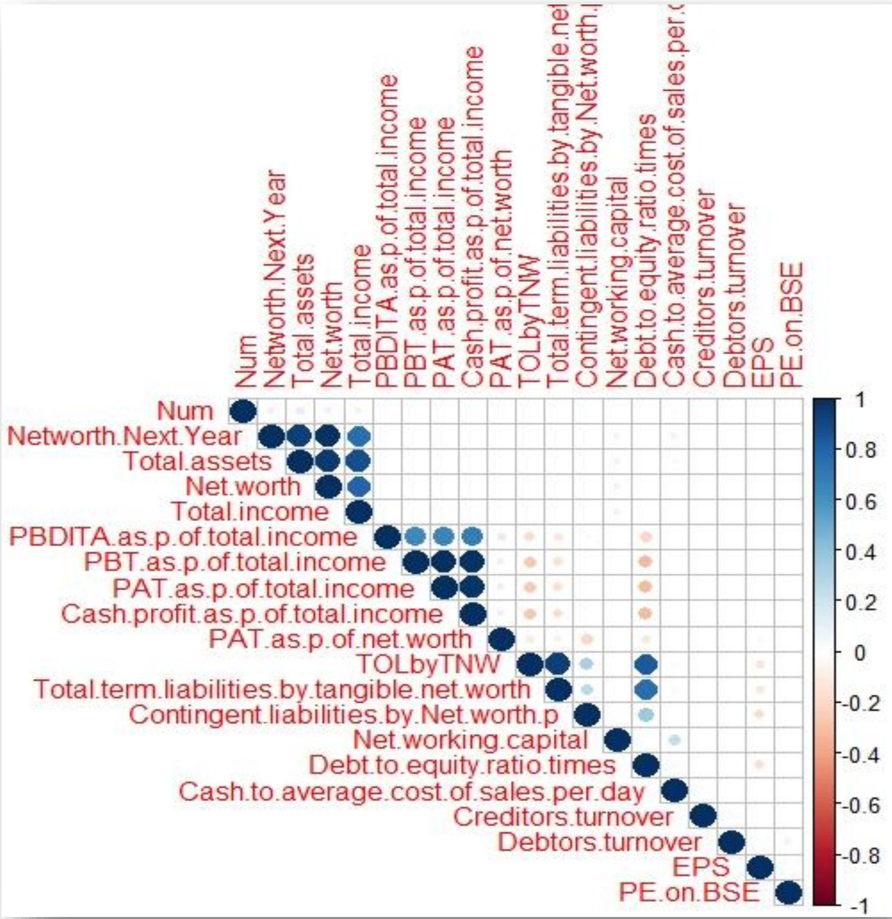
The plot is a comparison between the Leverage and Liquidity variables of companies. It clearly indicates that a high leverage ratio will directly indicate towards companies that can default.

4.7 Data Selection

```
> dim(train)
[1] 3268   21
```

```
> names(train)
 [1] "Num"                                    "Networth.Next.Year"
 [3] "Total.assets"                           "Net.worth"
 [5] "Total.income"                           "PBDITA.as.p.of.total.income"
 [7] "PBT.as.p.of.total.income"               "PAT.as.p.of.total.income"
 [9] "Cash.profit.as.p.of.total.income"       "PAT.as.p.of.net.worth"
[11] "TOLbyTNW"                               "Total.term.liabilities.by.tangible.net.worth"
[13] "Contingent.liabilities.by.Net.worth.p"  "Net.working.capital"
[15] "Debt.to.equity.ratio.times"             "Cash.to.average.cost.of.sales.per.day"
[17] "Creditors.turnover"                     "Debtors.turnover"
[19] "EPS"                                    "PE.on.BSE"
[21] "Default"
```

4.8 Correlation Check



Variables that fall under different specific categories such as profitability, size, liquidity or leverage have a strong positive relation amongst themselves.

This problem of multicollinearity needs to be resolved before building a regression model.

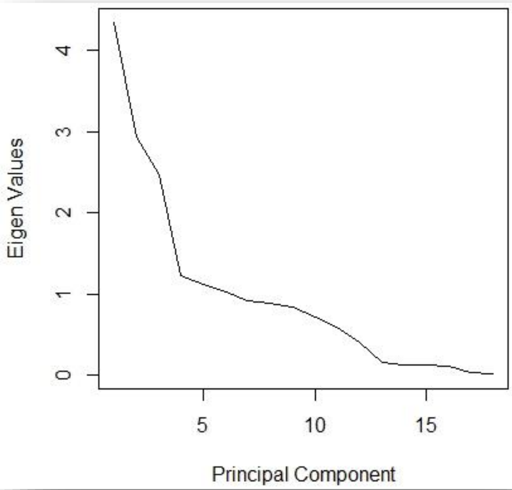
5. Variable Reduction

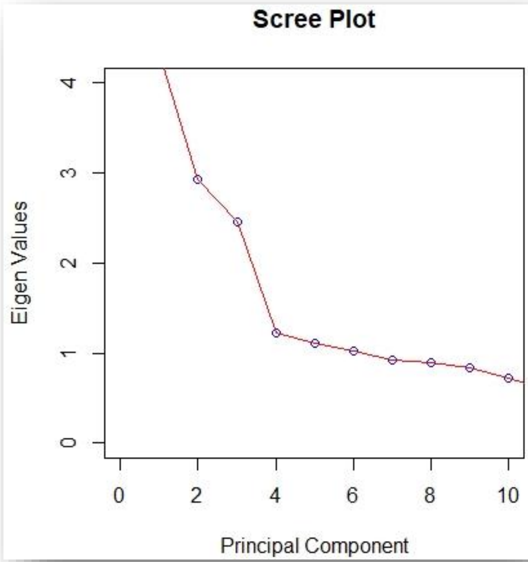
Principal Component Analysis (PCA) is a useful technique that allows you to better visualize the variation present in a dataset with many variables. It is helpful in the case of "wide" datasets. Therefore, we will use PCA for variables reduction.

5.1 EigenValue

Eigenvalues gives the amount of variation explained by each Principal Component(PC).

```
> eigenvalues <- ev$values
> #Eigen value comoutation
> ev <- eigen(cor(PCA.data))
> eigenvalues <- ev$values
> eigenvectors <- ev$vectors
> eigenvalues
 [1] 4.34245440 2.93278609 2.46229652 1.22802623 1.10981370 1.02481029 0.91816691 0.88851383
[12] 0.40493508 0.14862567 0.12728519 0.11708744 0.10472186 0.03678529 0.00741545
> plot(eigenvalues, type = "lines", xlab = "Principal Component", ylab = "Eigen Values")
```





In this case, we use the elbow method, which leads towards selection of “4” as the appropriate number of components for model building.

In order to get a better understanding, a scree plot was used.

5.2 PCA(Principal Component Analysis)

The analytical procedure of factor analysis involves creating uncorrelated (orthogonal) combinations of the initial independent variables. The purpose of the analysis is to reduce a mass of variables to a reasonable number of elements that the analyst can understand and explain.

```
> unrotate <- principal(PCA.data, nfactors = 4, rotate = "none")
> unrotate
Principal Components Analysis
Call: principal(r = PCA.data, nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                                           PC1   PC2   PC3   PC4    h2    u2 com
Total.assets                              0.34  0.90 -0.15  0.03 0.946 0.054 1.3
Net.worth                                 0.37  0.87 -0.19  0.03 0.934 0.066 1.5
Total.income                              0.35  0.86 -0.15  0.10 0.896 0.104 1.4
PBDITA.as.p.of.total.income               0.67 -0.05  0.42 -0.26 0.696 0.304 2.0
PBT.as.p.of.total.income                  0.81 -0.14  0.48  0.05 0.909 0.091 1.7
PAT.as.p.of.total.income                  0.79 -0.15  0.48  0.07 0.888 0.112 1.7
Cash.profit.as.p.of.total.income          0.81 -0.14  0.45 -0.07 0.888 0.112 1.7
PAT.as.p.of.net.worth                     0.67 -0.12  0.02  0.03 0.462 0.538 1.1
TOLbyTNW                                 -0.57  0.32  0.66 -0.04 0.874 0.126 2.4
Total.term.liabilities.by.tangible.net.worth -0.55  0.33  0.68 -0.03 0.877 0.123 2.4
Contingent.liabilities.by.Net.worth.p    -0.22  0.17  0.39  0.07 0.231 0.769 2.1
Net.working.capital                       0.24  0.43 -0.12 -0.18 0.286 0.714 2.2
Debt.to.equity.ratio.times               -0.58  0.32  0.67 -0.01 0.888 0.112 2.4
Cash.to.average.cost.of.sales.per.day     0.04  0.09 -0.09 -0.50 0.268 0.732 1.1
Creditors.turnover                        0.07 -0.06  0.01  0.62 0.387 0.613 1.0
Debtors.turnover                          0.07  0.04  0.09  0.68 0.480 0.520 1.1
EPS                                       0.18  0.05 -0.01  0.01 0.034 0.966 1.2
PE.on.BSE                                 0.13  0.07 -0.01  0.02 0.021 0.979 1.6

                       PC1  PC2  PC3  PC4
SS loadings           4.34 2.93 2.46 1.23
Proportion Var        0.24 0.16 0.14 0.07
Cumulative Var        0.24 0.40 0.54 0.61
Proportion Explained  0.40 0.27 0.22 0.11
Cumulative Proportion 0.40 0.66 0.89 1.00

Mean item complexity =  1.7
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  3448.93  with prob <  0

Fit based upon off diagonal values = 0.95
```

The above PCA has rotation as “none”

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of values of possibly M correlated variables into a set of K uncorrelated variables called principal components.

Varimax changes the coordinates that maximize the sum of square loadings. It is used to make an attempt to clarify the relationship among factors.

```
> rotate <- principal(PCA.data, nfactors = 4, rotate = "varimax") #orthogonal rotation will make the factors independent
> rotate
Principal Components Analysis
Call: principal(r = PCA.data, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                                          RC1   RC2   RC3   RC4    h2    u2 com
Total.assets                             0.03  0.97  0.05 -0.01 0.946 0.054 1.0
Net.worth                                0.04  0.97 -0.01 -0.01 0.934 0.066 1.0
Total.income                             0.04  0.94  0.02  0.07 0.896 0.104 1.0
PBDITA.as.p.of.total.income              0.80  0.08  0.00 -0.21 0.696 0.304 1.2
PBT.as.p.of.total.income                 0.94  0.05 -0.05  0.11 0.909 0.091 1.0
PAT.as.p.of.total.income                 0.93  0.04 -0.05  0.13 0.888 0.112 1.1
Cash.profit.as.p.of.total.income         0.94  0.05 -0.07  0.00 0.888 0.112 1.0
PAT.as.p.of.net.worth                    0.57  0.12 -0.35  0.06 0.462 0.538 1.8
TOLbyTNW                                -0.14 -0.05  0.92 -0.04 0.874 0.126 1.1
Total.term.liabilities.by.tangible.net.worth -0.12 -0.03  0.93 -0.02 0.877 0.123 1.0
Contingent.liabilities.by.Net.worth.p    0.00  0.00  0.47  0.08 0.231 0.769 1.1
Net.working.capital                      0.07  0.49 -0.05 -0.20 0.286 0.714 1.4
Debt.to.equity.ratio.times              -0.14 -0.05  0.93  0.00 0.888 0.112 1.1
Cash.to.average.cost.of.sales.per.day    0.00  0.10 -0.05 -0.51 0.268 0.732 1.1
Creditors.turnover                       0.03 -0.01 -0.05  0.62 0.387 0.613 1.0
Debtors.turnover                         0.05  0.07  0.04  0.69 0.480 0.520 1.0
EPS                                      0.12  0.11 -0.08  0.02 0.034 0.966 2.7
PE.on.BSE                                0.08  0.11 -0.04  0.02 0.021 0.979 2.3

                       RC1  RC2  RC3  RC4
SS loadings           3.70 3.08 2.95 1.24
Proportion Var        0.21 0.17 0.16 0.07
Cumulative Var        0.21 0.38 0.54 0.61
Proportion Explained  0.34 0.28 0.27 0.11
Cumulative Proportion 0.34 0.62 0.89 1.00

Mean item complexity =  1.3
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  3448.93  with prob <  0

Fit based upon off diagonal values = 0.95
```

Variable Names	RC1	RC2	RC3	RC4
Total.assets	0.03	0.97	0.05	0.01
Net.worth	0.04	0.97	0.01	0.01
Total.income	0.04	0.94	0.02	0.07
PBDITA.as.p.of.total.income	0.8	0.08	0	0.21
PBT.as.p.of.total.income	0.94	0.05	0.05	0.11
PAT.as.p.of.total.income	0.93	0.04	0.05	0.13
Cash.profit.as.p.of.total.income	0.94	0.05	0.07	0
PAT.as.p.of.net.worth	0.57	0.12	0.35	0.06
TOLbyTNW	0.14	0.05	0.92	0.04
Total.term.liabilities.by.tangible.net.worth	0.12	0.03	0.93	0.02
Contingent.liabilities.by.Net.worth.p	0	0	0.47	0.08
Net.working.capital	0.07	0.49	0.05	0.2
Debt.to.equity.ratio.times	0.14	0.05	0.93	0
Cash.to.average.cost.of.sales.per.day	0	0.1	0.05	0.51
Creditors.turnover	0.03	0.01	0.05	0.62
Debtors.turnover	0.05	0.07	0.04	0.69
EPS	0.12	0.11	0.08	0.02
PE.on.BSE	0.08	0.11	0.04	0.02

The above variables have been categorized into the groups “Profit”, “Liquidity”, “Leverage” and “Size” according to their absolute scores as is given below:

Profitability	Size	Leverage	Liquidity
PBDITA.as.p.of.total.income	Net.working.capital	TOLbyTNW	Cash.to.average.cost.of.sales.per.day
PBT.as.p.of.total.income	PE.on.BSE	Total.term.liabilities.by.tangible.net.worth	Creditors.turnover
PAT.as.p.of.total.income	Total.assets	Contingent.liabilities.by.Net.worth.p	Debtors.turnover
Cash.profit.as.p.of.total.income	Net.worth	Debt.to.equity.ratio.times	
PAT.as.p.of.net.worth	Total.income		
EPS			

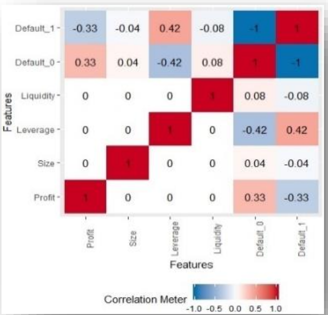
5.3 Creating New Dataset & Renaming Variables

The above variables after renaming have been added to a data frame along with the dependent variable("Default").

```
> #Translate PCA into regression
> newdf <- rotate$scores
> ndata <- as.data.frame(newdf)
> regPCA <- cbind(mydata$Default, mydata$Num, ndata)
> names(regPCA) <- c("Default", "Num", "Profit", "Liquidity", "Leverage", "Size")
> names(regPCA)
[1] "Default"   "Num"       "Profit"    "Liquidity" "Leverage"  "Size"
```

```
> summary(regPCA[,c(-2)])
 Default       Profit              Liquidity            Leverage               Size
 0:3065   Min.   :-7.37787   Min.   :-1.10156   Min.   :-0.54748   Min.   :-1.57069
 1: 191   1st Qu.:-0.14513   1st Qu.:-0.41196   1st Qu.:-0.33794   1st Qu.:-0.37782
          Median : 0.05319   Median :-0.27070   Median :-0.27566   Median :-0.24296
          Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
          3rd Qu.: 0.32724   3rd Qu.:-0.02227   3rd Qu.:-0.08499   3rd Qu.:-0.01204
          Max.   : 2.52944   Max.   : 7.48911   Max.   : 8.36751   Max.   :14.65345
```

The independent variables in the new dataset is devoid of any multicollinearity amongst themselves, but show a correlation with the dependent variable.



6. Predictive Modelling

Logistic Regression is the most used technique to predict the probability of company defaulting on a credit.

1.1 SMOTE

The SMOTE function handles unbalanced classification problems. It oversamples the rare event by using bootstrapping and k-nearest neighbour to synthetically create additional observations of that event.

1.2 Logistic Regression

```
> summary(model2)

Call:
glm(formula = train$Default ~ ., family = "binomial", data = train)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
 -8.49     0.00     0.00     0.00     8.49

Coefficients:
                                          Estimate Std. Error    z value Pr(>|z|)
(Intercept)                             -5.028e+14  2.917e+06 -172379809   <2e-16 ***
Num                                      8.170e+10  1.174e+03   69608486   <2e-16 ***
Networth.Next.Year                      -3.779e+11  1.650e+03 -229019459   <2e-16 ***
Total.assets                             2.881e+10  6.940e+02   41508674   <2e-16 ***
Net.worth                                3.242e+11  2.142e+03  151336878   <2e-16 ***
Total.income                            -5.655e+10  5.562e+02 -101668063   <2e-16 ***
PBDITA.as.p.of.total.income             -4.088e+11  1.299e+05   -3146670   <2e-16 ***
PBT.as.p.of.total.income                 2.434e+12  3.286e+05    7408383   <2e-16 ***
PAT.as.p.of.total.income                -4.398e+12  3.343e+05  -13155915   <2e-16 ***
Cash.profit.as.p.of.total.income         1.075e+13  1.500e+05   71634754   <2e-16 ***
PAT.as.p.of.net.worth                   -5.262e+12  4.976e+04 -105747142   <2e-16 ***
TOLbyTNW                                 3.382e+13  4.340e+05   77921730   <2e-16 ***
Total.term.liabilities.by.tangible.net.worth -8.877e+13  1.071e+06  -82864405   <2e-16 ***
Contingent.liabilities.by.Net.worth.p    3.296e+10  3.682e+03    8952877   <2e-16 ***
Net.working.capital                     -1.480e+11  2.257e+03  -65586001   <2e-16 ***
Debt.to.equity.ratio.times               2.189e+13  7.747e+05   28253035   <2e-16 ***
Cash.to.average.cost.of.sales.per.day   -8.009e+11  1.338e+04  -59846573   <2e-16 ***
Creditors.turnover                      -1.533e+13  6.066e+04 -252656256   <2e-16 ***
Debtors.turnover                        -2.389e+11  4.528e+04   -5275956   <2e-16 ***
EPS                                     -4.138e+11  8.811e+03  -46963022   <2e-16 ***
PE.on.BSE                               -6.792e+12  5.385e+04 -126119175   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  1466.5  on 3267  degrees of freedom
Residual deviance: 12182.8  on 3247  degrees of freedom
AIC: 12225

Number of Fisher Scoring iterations: 25
```

```
> summary(model1)

Call:
glm(formula = SMOTE.train$Default ~ ., family = "binomial", data = SMOTE.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3323  -0.7058   0.0098   0.5050   2.1880

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6495     0.1018  -6.377 1.80e-10 ***
Profit       -1.2851     0.1293  -9.936  < 2e-16 ***
Size         -0.4609     0.1503  -3.066  0.00217 **
Leverage      1.9047     0.1743  10.928  < 2e-16 ***
Liquidity    -0.5792     0.1069  -5.419 5.99e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.91  on 964  degrees of freedom
Residual deviance:  724.23  on 960  degrees of freedom
AIC: 734.23

Number of Fisher Scoring iterations: 7
```

The estimate of the (Intercept) is unrelated to the number of predictors. The value of the coefficients help determine the magnitude of effect a variable has. The four predictor variables (aka features) are:

- Profitability of a Company: The variable has a negative sign meaning that it has a negative relation with the dependent variable i.e. the higher the profitability of a company, the less likely they are to default.

- Size of Company: The negative sign denotes a negative relation with the Default variable. The larger the size of the company, the less probable it is to default in its payments.

- Leveraging power of Company: A positive sign means that all else being equal, a company higher liquidity is more likely to have not churned. The higher the leverage means that a company has more debts and is more inclined to default in compassion to companies with a lower leverage ratio.

- Liquidity of Company: The liquidity variable has a negative relation to the dependent variable. If the liquidity ratio of a company goes up, the company is more probable to default as compared to companies with a lower liquidity ratio.

*All the variables in the model are “significant.

Fisher scoring iterations uses an iterative approach (the Newton-Raphson algorithm by default) that looks for the best model. The algorithm stops when it doesn't perceive that moving again would yield much additional improvement. This line tells you how many iterations there were before the process stopped and output the results.

Akaike's An Information Criterion provides a method for assessing the quality of your model through comparison of related models. It's based on the Deviance, but penalizes you for making the model more complicated. Much like adjusted R-squared, it's intent is to prevent you from including irrelevant predictors.

Logistic Regression was performed on the original dataset(train) and the refined dataset after applying variable reduction and SMOTE on the original dataset.

The model with a lower AIC is always preferred, therefore the model1 will be selected for further analysing the accuracy of the model.

1.3 Variable Importance

To assess the relative importance' of individual predictors in the model, we can also look at the absolute value of the t-statistic for each model parameter. This technique is utilized by the varImp function in the caret package for general and generalized linear models.

```
> vif(model1)
  Profit      Size  Leverage Liquidity
1.241523  1.056264  1.103593  1.214411
```

VIF(Variable Inflation Factor) a simple approach to identify collinearityamong independent variables. Collinearity, or excessive correlation among explanatory variables, can complicate or prevent the identification of an optimal set of explanatory variables for a statistical model.

```
> imp <- as.data.frame(varImp(model2))
> imp <- data.frame(names = rownames(imp),overall = imp$Overall)
> imp[order(imp$overall, decreasing = T),]
                                         names   overall
17                          Creditors.turnover 252656256
2                           Networth.Next.Year 229019459
4                                    Net.worth 151336878
20                                   PE.on.BSE 126119175
10                       PAT.as.p.of.net.worth 105747142
5                                 Total.income 101668063
12 Total.term.liabilities.by.tangible.net.worth  82864405
11                                    TOLbyTNW  77921730
9                Cash.profit.as.p.of.total.income  71634754
1                                          Num  69608486
14                           Net.working.capital  65586001
16        Cash.to.average.cost.of.sales.per.day  59846573
19                                         EPS  46963022
3                                 Total.assets  41508674
15                     Debt.to.equity.ratio.times  28253035
8                      PAT.as.p.of.total.income  13155915
13          Contingent.liabilities.by.Net.worth.p   8952877
7                      PBT.as.p.of.total.income   7408383
18                              Debtors.turnover   5275956
6                   PBDITA.as.p.of.total.income   3146670
```

```
> varImp(model1)
                        Overall
SMOTE.data$Profit      9.935992
SMOTE.data$Leverage   10.927903
SMOTE.data$Liquidity   5.419042
SMOTE.data$Size        3.066385
```

VarImp can be used to compute variable importance measures. Besides the standard version, a conditional version is available, that adjusts for correlations between predictor variables.

1.4 Model Validation

In order to cross-check the accuracy of the dataset by gauging if it can predict the same way as the train dataset, a validation/unseen dataset is used.

1.4.1 EDA

The same EDA steps are performed on the validation dataset, as in on the test dataset, i.e.

- Import the data
- Replace the "NA" values in the dataset with "0".
- Missing value Treatment
- Outlier Treatment
- PCA

1.4.2 Validate the model

Confusion Matrix

```
> accuracy.test <- table(regPCA.test$Default, LR.pred.test)
> accuracy.test
   LR.pred.test
      0   1
  0 581  43
  1   6  38
> #Accuracy
> (accuracy.test[1,1]+accuracy.test[2,2])/nrow(regPCA.test)
[1] 0.9266467
```

Accuracy

```
> ##CONFUSION MATRIX
> LR_CM.test = table(regPCA.test$Default, LR.pred.test>0.01)
> LR_CM.test

    FALSE TRUE
  0   581   43
  1     6   38
> ##ERROR RATE
> (LR_CM.test[1,2]+LR_CM.test[2,1])/nrow(regPCA.test) #1,2; 2,1 refers to the placements
[1] 0.07335329
> ##ACCURACY
> (LR_CM.test[1,1]+LR_CM.test[2,2])/nrow(regPCA.test)
[1] 0.9266467
```

Sensitivity

```
> #SENSITIVITY
> LR_CM.test[1,1]/sum(LR_CM.test[1,1], LR_CM.test[1,2])
[1] 0.9310897
```

Specificity

```
> #SPECIFICITY
> LR_CM.test[2,2]/sum(LR_CM.test[2,1], LR_CM.test[2,2])
[1] 0.8636364
```

KS

```
> ##KS
> max(test.ROC.plot@y.values[[1]]-test.ROC.plot@x.values[[1]])
[1] 0.8030303
```

Gini Coefficient

```
> ##GINI COEFFICIENT
> ineq(LR.pred.test, "gini")
[1] 0.8787425
```

Area Under the ROC Curve

```
> ##AUC
> test.AUC=performance(test.ROC,"auc")
> slot(test.AUC, "y.values")
[[1]]
[1] 0.9362617
```

7. Model Performance Measure

1.5 Accuracy, Sensitivity, Specificity, ROC, AUC, KS, Gini

Confusion Matrix used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

```
> ##CONFUSION MATRIX
> LR_CM = table(SMOTE.train$Default, LR.pred>0.01)
> LR_CM

    FALSE  TRUE
  0   333    53
  1   111   468
```

Error Rate is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0.

```
> ##ERROR RATE
> (LR_CM[1,2]+LR_CM[2,1])/nrow(SMOTE.train) #1,2; 2,1 refers to the placements
[1] 0.1699482
```

Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset.

```
> ##ACCURACY
> (LR_CM[1,1]+LR_CM[2,2])/nrow(SMOTE.train)
[1] 0.8300518
```

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. The best sensitivity is 1.0, whereas the worst is 0.0.

```
> #SENSITIVITY
> LR_CM[1,1]/sum(LR_CM[1,1], LR_CM[1,2])
[1] 0.8834197
```

Specificity (True negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives. The best specificity is 1.0, whereas the worst is 0.0.

```
> #SPECIFICITY
> LR_CM[2,2]/sum(LR_CM[2,1], LR_CM[2,2])
[1] 0.8013817
```

KS(Kolmogorov Smirnov Chart) measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions.

The model's capability to segregate between the defaulters and not defaulters is 69%.

```
> ##KS
> max(LR.ROC.plot@y.values[[1]]-LR.ROC.plot@x.values[[1]])
[1] 0.6856649
```

Gini Coefficient is nothing but ratio between area between the ROC curve and the diagonal line & the area of the above triangle.

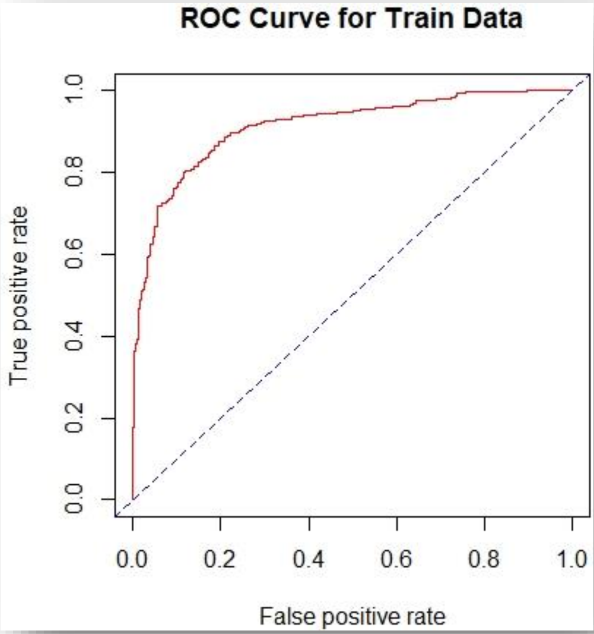
Gini = 2*AUC – 1

```
> ##GINI COEFFICIENT
> ineq(LR.pred, "gini")
[1] 0.4601036
```

Area Under the ROC Curve

ROC curves are a nice way to see how any predictive model can distinguish between the true positives and negatives. The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as false positive rate and sensitivity is also known as True Positive rate.

The AUC is 91%, which implies that it is an excellent model for the given dataset.




```
> ##AUC
> LR.AUC=performance(LR.ROC,"auc")
> slot(LR.AUC, "y.values")
[[1]]
[1] 0.9136979
```

1.6 Deciling

```
> #SMOTE.train <- SMOTE.train[,c(-6,-7,-8)]
> final <- data.frame(SMOTE.train, LR.prob)
> final$LR.prob <- round(final$LR.prob, 2)
> L.F <- arrange(final, desc(LR.prob))
> L.F$decile <- with(L.F, cut_number(LR.prob, 10, labels = 10:1))
> train.score <- L.F %>% group_by(decile)
> train.score
# A tibble: 965 x 7
# Groups:   decile [10]
   Default Profit     Size Leverage Liquidity LR.prob decile
   <fct>    <dbl>    <dbl>    <dbl>     <dbl>   <dbl> <fct>
 1 1        0.889 -0.214     14.5       2.46     24.6 1
 2 1       -6.66  -0.0182     6.77     -0.432    21.1 1
 3 1       -6.68  -0.0216     6.70     -0.474    21.0 1
 4 1       -6.65  -0.0174     6.69     -0.445    20.9 1
 5 1       -6.79  -0.00330    5.98     -1.29     20.2 1
 6 1       -6.79  -0.0501     6.09     -0.820    20.2 1
 7 1       -6.80  -0.00290    5.95     -1.31     20.2 1
 8 1       -6.79  -0.0477     6.09     -0.845    20.2 1
 9 1       -6.80  -0.0417     6.07     -0.907    20.2 1
10 1       -5.91  -0.0594     6.24     -1.89     19.9 1
# ... with 955 more rows
```

The dataset has been divided into 10 groups/deciles on the basis of the descending values of the predicted probability of the model.

8. Insights

Model Performance Measures						
Dataset	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**KS**	**Gini**
Train Dataset	0.8341969	0.8834197	0.8013817	0.9136979	0.6856649	0.4725389
Test Dataset	0.9266467	0.9310897	0.8636364	0.9362617	0.8030303	0.8787425

From the above result it is clear that the model indicates that it is a good fit for the given dataset. Since, the model performs well on an unseen dataset.

9. Source of Data

- Great Learning Mentored Learning Session and Recorded Sessions
- Google
- R Blogger
- stats.stackexchange.com
- uc-r.github.io
- https://tabvizexplorer.com/
- www.rpubs.com
- www.datacamp.com
- https://cran.r-project.org
- https://machinelearningmastery.com
- https://www.analyticsvidhya.com/