**FRA Group Assignment**

You are requested to create an India credit risk (default) model, using the data provided in the spreadsheet raw-data.xlsx, and validate it on validation\_data.xlsx. Please use the **logistic regression** framework to develop the credit default model.

**Group Members**

1. Aditya Godse
2. Akshay Surve
3. Amay Saxena

Contents

[**1.** **Data Preparation and Data Cleaning** 2](#_Toc37019760)

[**2. EDA, Outlier Treatment - Univariate and Bivariate Analysis** 11](#_Toc37019761)

[**3.** **Creating new Variables in Ration format to bring the scale of all companies on the same scale** 19](#_Toc37019762)

[**4. Dimension Reduction and Logistic Model development on train dataset** 20](#_Toc37019763)

[**5. Evaluating model output against actual values of train dataset using the validation metrics** 27](#_Toc37019764)

[**6.** **Applying the model on test dataset** 29](#_Toc37019765)

[**7.** **Model Accuracies and Business Insights** 32](#_Toc37019766)

[**Appendix** 33](#_Toc37019767)

# **Data Preparation and Data Cleaning**

library(Amelia)

library(caret)

library(DMwR)

library(tidyverse)

library(gridExtra)

library(car)

library(lmtest)

library(caret)

library(SDMTools)

library(pROC)

library(Hmisc)

library(pscl)

library(ModelMetrics)

setwd("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment FRA")

getwd()

raw\_data <- read.csv("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment FRA/raw-data CSV.csv")

str(raw\_data)

'data.frame': 3541 obs. of 52 variables:

$ Num : int 1 2 3 4 5 6 7 8 9 10 ...

$ Default : int 0 0 0 0 0 0 0 0 0 1 ...

$ Total.assets : num 17512.3 941 232.8 2.7 478.5 ...

$ Net.worth : num 7093.2 351.5 100.6 2.7 107.6 ...

$ Total.income : num 24965 1527 477 NA 1580 ...

$ Change.in.stock : num 235.8 42.7 -5.2 NA -17 ...

$ Total.expenses : num 23658 1455 479 NA 1558 ...

$ Profit.after.tax : num 1543.2 115.2 -6.6 NA 5.5 ...

$ PBDITA : num 2860.2 283 5.8 NA 31 ...

$ PBT : num 2417.2 188.4 -6.6 NA 6.3 ...

$ Cash.profit : num 1872.8 158.6 0.3 NA 11.9 ...

$ PBDITA.as...of.total.income : num 11.46 18.53 1.22 0 1.96 ...

$ PBT.as...of.total.income : num 9.68 12.33 -1.38 0 0.4 ...

$ PAT.as...of.total.income : num 6.18 7.54 -1.38 0 0.35 2.81 0 0.72 8.29 -2.88

$ Cash.profit.as...of.total.income : num 7.5 10.38 0.06 0 0.75 ...

$ PAT.as...of.net.worth : num 23.78 38.08 -6.35 0 5.25 ...

$ Sales : num 24458 1504 476 NA 1575 ...

$ Income.from.financial.services : num 158 4 1.5 NA 3.9 6.4 NA NA 7.3 NA ...

$ Other.income : num 297.2 15.9 0.2 NA 0.9 ...

$ Total.capital : num 423.8 115.5 81.4 0.5 6.2 ...

$ Reserves.and.funds : num 6822.8 257.8 19.2 2.2 161.8 ...

$ Deposits..accepted.by.commercial.banks. : logi NA NA NA NA NA NA ...

$ Borrowings : num 14.9 272.5 35.4 NA 193.1 ...

$ Current.liabilities...provisions : num 9965.9 210 96.8 NA 112.8 ...

$ Deferred.tax.liability : num 284.9 85.2 NA NA 4.6 ...

$ Shareholders.funds : num 7093.2 351.5 100.6 2.7 107.6 ...

$ Cumulative.retained.profits : num 6263.3 247.4 32.4 2.2 82.7 ...

$ Capital.employed : num 7108.1 624 136 2.7 300.7 ...

$ TOL.TNW : num 1.33 1.23 1.44 0 2.83 1.8 0.03 5.17 1.05 3.25

$ Total.term.liabilities...tangible.net.worth: num 0 0.34 0.29 0 1.59 0.37 0.03 0.94 0.3 0.54 ...

$ Contingent.liabilities...Net.worth.... : num 14.8 19.2 45.8 0 34.9 ...

$ Contingent.liabilities : num 1049.7 67.6 46.1 NA 37.6 ...

$ Net.fixed.assets : num 1900.2 286.4 38.7 2.5 94.8 ...

$ Investments : num 1069.6 2.2 4.3 NA 7.4 ...

$ Current.assets : num 13277.5 563.9 167.5 0.2 349.7 ...

$ Net.working.capital : num 3588.5 203.5 59.6 0.2 215.8 ...

$ Quick.ratio..times. : num 1.18 0.95 1.11 NA 1.41 0.48 NA 0.54 0.59 0.39

$ Current.ratio..times. : num 1.37 1.56 1.55 NA 2.54 1.27 NA 1.15 1.58 0.5 .

$ Debt.to.equity.ratio..times. : num 0 0.78 0.35 0 1.79 1.09 0.32 2.31 0.94 3.13 ..

$ Cash.to.current.liabilities..times. : num 0.43 0.06 0.21 NA 0 0.11 NA 0.04 0.19 0 ...

$ Cash.to.average.cost.of.sales.per.day : num 68.21 5.96 17.07 NA 0 ...

$ Creditors.turnover : num 3.62 9.8 5.28 0 13 ...

$ Debtors.turnover : num 3.85 5.7 5.07 0 9.46 ...

$ Finished.goods.turnover : num 200.55 14.21 9.24 NA 12.68 ...

$ WIP.turnover : num 21.78 7.49 0.23 NA 7.9 ...

$ Raw.material.turnover : num 7.71 11.46 NA 0 17.03 ...

$ Shares.outstanding : num 42381675 11550000 8149090 52404 619635 ...

$ Equity.face.value : num 10 10 10 10 10 10 10 NA 10 10 ...

$ EPS : num 35.52 9.97 -0.5 0 7.91 ...

$ Adjusted.EPS : num 7.1 9.97 -0.5 0 7.91 ...

$ Total.liabilities : num 17512.3 941 232.8 2.7 478.5 ...

$ PE.on.BSE : num 27.31 8.17 -5.76 NA NA ...

summary(raw\_data)

Num Default Total.assets Net.worth Total.income

Min. : 1 Min. :0.00000 Min. : 0.1 Min. : 0.0 Min. : 0.0

1st Qu.: 886 1st Qu.:0.00000 1st Qu.: 91.3 1st Qu.: 31.3 1st Qu.: 106.5

Median :1773 Median :0.00000 Median : 309.7 Median : 102.3 Median : 444.9

Mean :1772 Mean :0.06608 Mean : 3443.4 Mean : 1295.9 Mean : 4582.8

3rd Qu.:2658 3rd Qu.:0.00000 3rd Qu.: 1098.7 3rd Qu.: 377.3 3rd Qu.: 1440.9

Max. :3545 Max. :1.00000 Max. :1176509.2 Max. :613151.6 Max. :2442828.2

NA's :198

Change.in.stock Total.expenses Profit.after.tax PBDITA PBT

Min. :-3029.40 Min. : -0.1 Min. : -3908.30 Min. : -440.7 Min. : -3894.80

1st Qu.: -1.80 1st Qu.: 95.8 1st Qu.: 0.50 1st Qu.: 6.9 1st Qu.: 0.70

Median : 1.60 Median : 407.7 Median : 8.80 Median : 35.4 Median : 12.40

Mean : 41.49 Mean : 4262.9 Mean : 277.36 Mean : 578.1 Mean : 383.81

3rd Qu.: 18.05 3rd Qu.: 1359.8 3rd Qu.: 52.27 3rd Qu.: 150.2 3rd Qu.: 71.97

Max. :14185.50 Max. :2366035.3 Max. :119439.10 Max. :208576.5 Max. :145292.60

NA's :458 NA's :139 NA's :131 NA's :131 NA's :131

Cash.profit PBDITA.as...of.total.income PBT.as...of.total.income PAT.as...of.total.income

Min. : -2245.70 Min. :-6400.000 Min. :-21340.00 Min. :-21340.00

1st Qu.: 2.90 1st Qu.: 5.000 1st Qu.: 0.55 1st Qu.: 0.35

Median : 18.85 Median : 9.660 Median : 3.31 Median : 2.34

Mean : 392.07 Mean : 4.571 Mean : -17.28 Mean : -19.20

3rd Qu.: 93.20 3rd Qu.: 16.390 3rd Qu.: 8.80 3rd Qu.: 6.34

Max. :176911.80 Max. : 100.000 Max. : 100.00 Max. : 150.00

NA's :131 NA's :68 NA's :68 NA's :68

Cash.profit.as...of.total.income PAT.as...of.net.worth Sales

Min. :-15020.000 Min. :-748.72 Min. : 0.1

1st Qu.: 2.020 1st Qu.: 0.00 1st Qu.: 112.7

Median : 5.640 Median : 7.92 Median : 453.1

Mean : -8.229 Mean : 10.27 Mean : 4549.5

3rd Qu.: 10.700 3rd Qu.: 20.19 3rd Qu.: 1433.5

Max. : 100.000 Max. :2466.67 Max. :2384984.4

NA's :68 NA's :259

Income.from.financial.services Other.income Total.capital Reserves.and.funds

Min. : 0.00 Min. : 0.00 Min. : 0.1 Min. : -6525.9

1st Qu.: 0.40 1st Qu.: 0.40 1st Qu.: 13.1 1st Qu.: 5.0

Median : 1.80 Median : 1.40 Median : 42.1 Median : 54.8

Mean : 80.84 Mean : 41.36 Mean : 216.6 Mean : 1163.8

3rd Qu.: 9.68 3rd Qu.: 5.97 3rd Qu.: 100.3 3rd Qu.: 277.3

Max. :51938.20 Max. :42856.70 Max. :78273.2 Max. :625137.8

NA's :935 NA's :1295 NA's :4 NA's :85

Deposits..accepted.by.commercial.banks. Borrowings Current.liabilities...provisions

Mode:logical Min. : 0.10 Min. : 0.1

NA's:3541 1st Qu.: 23.95 1st Qu.: 17.8

Median : 99.20 Median : 69.4

Mean : 1122.28 Mean : 940.6

3rd Qu.: 352.60 3rd Qu.: 261.7

Max. :278257.30 Max. :352240.3

NA's :366 NA's :96

Deferred.tax.liability Shareholders.funds Cumulative.retained.profits Capital.employed

Min. : 0.1 Min. : 0.0 Min. : -6534.3 Min. : 0.0

1st Qu.: 3.2 1st Qu.: 32.0 1st Qu.: 1.1 1st Qu.: 60.8

Median : 13.4 Median : 105.6 Median : 37.1 Median : 214.7

Mean : 227.2 Mean : 1322.1 Mean : 890.5 Mean : 2328.3

3rd Qu.: 50.0 3rd Qu.: 393.2 3rd Qu.: 202.3 3rd Qu.: 767.3

Max. :72796.6 Max. :613151.6 Max. :390133.8 Max. :891408.9

NA's :1140 NA's :38

TOL.TNW Total.term.liabilities...tangible.net.worth

Min. :-350.480 Min. :-325.600

1st Qu.: 0.600 1st Qu.: 0.050

Median : 1.430 Median : 0.340

Mean : 3.994 Mean : 1.844

3rd Qu.: 2.830 3rd Qu.: 1.000

Max. : 473.000 Max. : 456.000

Contingent.liabilities...Net.worth.... Contingent.liabilities Net.fixed.assets

Min. : 0.00 Min. : 0.1 Min. : 0.0

1st Qu.: 0.00 1st Qu.: 6.3 1st Qu.: 26.0

Median : 5.33 Median : 38.0 Median : 93.5

Mean : 53.94 Mean : 932.9 Mean : 1189.7

3rd Qu.: 30.76 3rd Qu.: 192.7 3rd Qu.: 344.9

Max. :14704.27 Max. :559506.8 Max. :636604.6

NA's :1188 NA's :118

Investments Current.assets Net.working.capital Quick.ratio..times.

Min. : 0.00 Min. : 0.1 Min. :-63839.0 Min. : 0.000

1st Qu.: 1.00 1st Qu.: 36.2 1st Qu.: -1.1 1st Qu.: 0.410

Median : 8.35 Median : 145.1 Median : 16.2 Median : 0.670

Mean : 694.73 Mean : 1293.4 Mean : 138.6 Mean : 1.401

3rd Qu.: 64.30 3rd Qu.: 502.2 3rd Qu.: 84.2 3rd Qu.: 1.030

Max. :199978.60 Max. :354815.2 Max. : 85782.8 Max. :341.000

NA's :1435 NA's :66 NA's :32 NA's :93

Current.ratio..times. Debt.to.equity.ratio..times. Cash.to.current.liabilities..times.

Min. : 0.00 Min. : 0.00 Min. : 0.0000

1st Qu.: 0.93 1st Qu.: 0.22 1st Qu.: 0.0200

Median : 1.23 Median : 0.79 Median : 0.0700

Mean : 2.13 Mean : 2.78 Mean : 0.4904

3rd Qu.: 1.71 3rd Qu.: 1.75 3rd Qu.: 0.1900

Max. :505.00 Max. :456.00 Max. :165.0000

NA's :93 NA's :93

Cash.to.average.cost.of.sales.per.day Creditors.turnover Debtors.turnover

Min. : 0.00 Min. : 0.000 Min. : 0.00

1st Qu.: 2.79 1st Qu.: 3.700 1st Qu.: 3.76

Median : 8.03 Median : 6.095 Median : 6.32

Mean : 158.44 Mean : 15.446 Mean : 17.04

3rd Qu.: 21.79 3rd Qu.: 11.490 3rd Qu.: 11.68

Max. :128040.76 Max. :2401.000 Max. :3135.20

NA's :85 NA's :333 NA's :328

Finished.goods.turnover WIP.turnover Raw.material.turnover Shares.outstanding

Min. : -0.09 Min. : -0.18 Min. : -2.00 Min. :-2.147e+09

1st Qu.: 8.20 1st Qu.: 5.10 1st Qu.: 2.99 1st Qu.: 1.316e+06

Median : 17.27 Median : 9.76 Median : 6.40 Median : 4.672e+06

Mean : 87.08 Mean : 27.93 Mean : 19.09 Mean : 2.207e+07

3rd Qu.: 40.35 3rd Qu.: 20.24 3rd Qu.: 11.85 3rd Qu.: 1.065e+07

Max. :17947.60 Max. :5651.40 Max. :21092.00 Max. : 4.130e+09

NA's :740 NA's :640 NA's :361 NA's :692

Equity.face.value EPS Adjusted.EPS Total.liabilities PE.on.BSE

Min. :-999999 Min. :-843181.8 Min. :-843181.8 Min. : 0.1 Min. :-1116.64

1st Qu.: 10 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 91.3 1st Qu.: 3.27

Median : 10 Median : 1.4 Median : 1.2 Median : 309.7 Median : 9.10

Mean : -1334 Mean : -220.3 Mean : -221.5 Mean : 3443.4 Mean : 63.91

3rd Qu.: 10 3rd Qu.: 9.6 3rd Qu.: 7.5 3rd Qu.: 1098.7 3rd Qu.: 17.79

Max. : 100000 Max. : 34522.5 Max. : 34522.5 Max. :1176509.2 Max. :51002.74

NA's :692 NA's :2194

* **number of NA's in all parameters**

missing\_values = lapply(raw\_data,function(x)sum(is.na(x)))

sum(is.na(raw\_data))

[1] 18533

missing\_values

$Num

[1] 0

$Default

[1] 0

$Total.assets

[1] 0

$Net.worth

[1] 0

$Total.income

[1] 198

$Change.in.stock

[1] 458

$Total.expenses

[1] 139

$Profit.after.tax

[1] 131

$PBDITA

[1] 131

$PBT

[1] 131

$Cash.profit

[1] 131

$PBDITA.as...of.total.income

[1] 68

$PBT.as...of.total.income

[1] 68

$PAT.as...of.total.income

[1] 68

$Cash.profit.as...of.total.income

[1] 68

$PAT.as...of.net.worth

[1] 0

$Sales

[1] 259

$Income.from.financial.services

[1] 935

$Other.income

[1] 1295

$Total.capital

[1] 4

$Reserves.and.funds

[1] 85

$Deposits..accepted.by.commercial.banks.

[1] 3541

$Borrowings

[1] 366

$Current.liabilities...provisions

[1] 96

$Deferred.tax.liability

[1] 1140

$Shareholders.funds

[1] 0

$Cumulative.retained.profits

[1] 38

$Capital.employed

[1] 0

$TOL.TNW

[1] 0

$Total.term.liabilities...tangible.net.worth

[1] 0

$Contingent.liabilities...Net.worth....

[1] 0

$Contingent.liabilities

[1] 1188

$Net.fixed.assets

[1] 118

$Investments

[1] 1435

$Current.assets

[1] 66

$Net.working.capital

[1] 32

$Quick.ratio..times.

[1] 93

$Current.ratio..times.

[1] 93

$Debt.to.equity.ratio..times.

[1] 0

$Cash.to.current.liabilities..times.

[1] 93

$Cash.to.average.cost.of.sales.per.day

[1] 85

$Creditors.turnover

[1] 333

$Debtors.turnover

[1] 328

$Finished.goods.turnover

[1] 740

$WIP.turnover

[1] 640

$Raw.material.turnover

[1] 361

$Shares.outstanding

[1] 692

$Equity.face.value

[1] 692

$EPS

[1] 0

$Adjusted.EPS

[1] 0

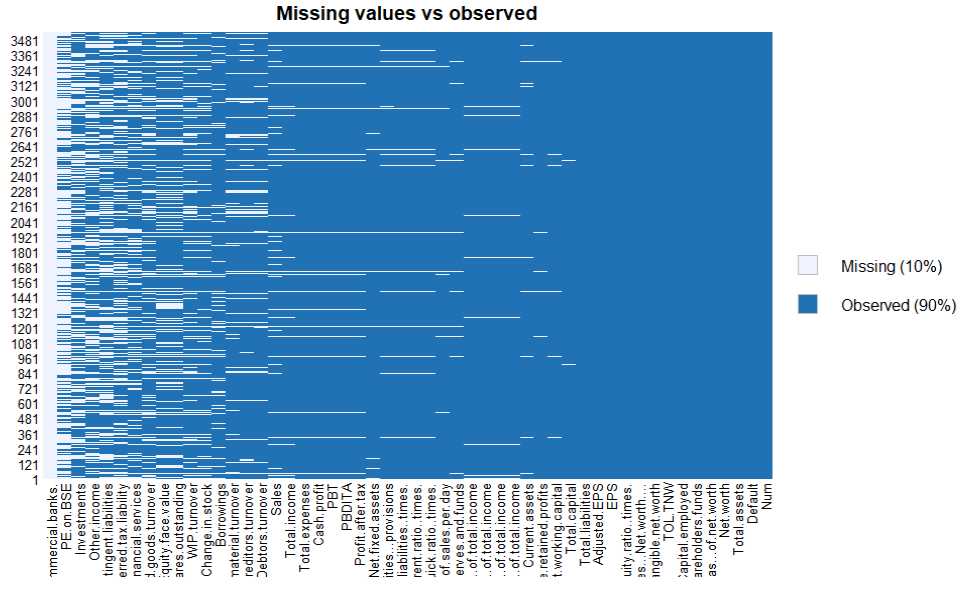
$Total.liabilities

[1] 0

$PE.on.BSE

[1] 2194

missmap(raw\_data, main = "Missing values vs observed")



10 percent values are Missing values if we look from all columns in total. NA values have been Changed with median value for that column as outlier are too broad.

raw\_data\_prepared <- data.frame(

sapply( raw\_data,

function(x) ifelse(is.na(x),

median(x, na.rm = TRUE), x)))

summary(raw\_data\_prepared)

Num Default Total.assets Net.worth

Min. : 1 Min. :0.00000 Min. : 0.1 Min. : 0.0

1st Qu.: 886 1st Qu.:0.00000 1st Qu.: 91.3 1st Qu.: 31.3

Median :1773 Median :0.00000 Median : 309.7 Median : 102.3

Mean :1772 Mean :0.06608 Mean : 3443.4 Mean : 1295.9

3rd Qu.:2658 3rd Qu.:0.00000 3rd Qu.: 1098.7 3rd Qu.: 377.3

Max. :3545 Max. :1.00000 Max. :1176509.2 Max. :613151.6

Total.income Change.in.stock Total.expenses

Min. : 0.0 Min. :-3029.40 Min. : -0.1

1st Qu.: 121.2 1st Qu.: -0.70 1st Qu.: 104.1

Median : 444.9 Median : 1.60 Median : 407.7

Mean : 4351.4 Mean : 36.33 Mean : 4111.6

3rd Qu.: 1340.3 3rd Qu.: 13.40 3rd Qu.: 1284.6

Max. :2442828.2 Max. :14185.50 Max. :2366035.3

Profit.after.tax PBDITA PBT

Min. : -3908.3 Min. : -440.7 Min. : -3894.8

1st Qu.: 0.6 1st Qu.: 7.3 1st Qu.: 0.9

Median : 8.8 Median : 35.4 Median : 12.4

Mean : 267.4 Mean : 558.0 Mean : 370.1

3rd Qu.: 48.1 3rd Qu.: 139.1 3rd Qu.: 67.5

Max. :119439.1 Max. :208576.5 Max. :145292.6

Cash.profit PBDITA.as...of.total.income PBT.as...of.total.income

Min. : -2245.70 Min. :-6400.000 Min. :-21340.00

1st Qu.: 3.10 1st Qu.: 5.070 1st Qu.: 0.60

Median : 18.85 Median : 9.660 Median : 3.31

Mean : 378.26 Mean : 4.668 Mean : -16.88

3rd Qu.: 86.80 3rd Qu.: 16.150 3rd Qu.: 8.63

Max. :176911.80 Max. : 100.000 Max. : 100.00

PAT.as...of.total.income Cash.profit.as...of.total.income

Min. :-21340.00 Min. :-15020.000

1st Qu.: 0.39 1st Qu.: 2.090

Median : 2.34 Median : 5.640

Mean : -18.79 Mean : -7.963

3rd Qu.: 6.25 3rd Qu.: 10.560

Max. : 150.00 Max. : 100.000

PAT.as...of.net.worth Sales Income.from.financial.services

Min. :-748.72 Min. : 0.1 Min. : 0.00

1st Qu.: 0.00 1st Qu.: 133.3 1st Qu.: 0.70

Median : 7.92 Median : 453.1 Median : 1.80

Mean : 10.27 Mean : 4249.9 Mean : 59.97

3rd Qu.: 20.19 3rd Qu.: 1314.7 3rd Qu.: 5.40

Max. :2466.67 Max. :2384984.4 Max. :51938.20

Other.income Total.capital Reserves.and.funds

Min. : 0.00 Min. : 0.1 Min. : -6525.9

1st Qu.: 0.80 1st Qu.: 13.1 1st Qu.: 5.8

Median : 1.40 Median : 42.1 Median : 54.8

Mean : 26.74 Mean : 216.4 Mean : 1137.2

3rd Qu.: 2.50 3rd Qu.: 100.3 3rd Qu.: 263.2

Max. :42856.70 Max. :78273.2 Max. :625137.8

Deposits..accepted.by.commercial.banks. Borrowings

Min. : NA Min. : 0.1

1st Qu.: NA 1st Qu.: 29.7

Median : NA Median : 99.2

Mean :NaN Mean : 1016.5

3rd Qu.: NA 3rd Qu.: 296.0

Max. : NA Max. :278257.3

NA's :3541

Current.liabilities...provisions Deferred.tax.liability Shareholders.funds

Min. : 0.1 Min. : 0.1 Min. : 0.0

1st Qu.: 18.7 1st Qu.: 6.7 1st Qu.: 32.0

Median : 69.4 Median : 13.4 Median : 105.6

Mean : 917.0 Mean : 158.4 Mean : 1322.1

3rd Qu.: 249.1 3rd Qu.: 26.9 3rd Qu.: 393.2

Max. :352240.3 Max. :72796.6 Max. :613151.6

Cumulative.retained.profits Capital.employed TOL.TNW

Min. : -6534.3 Min. : 0.0 Min. :-350.480

1st Qu.: 1.3 1st Qu.: 60.8 1st Qu.: 0.600

Median : 37.1 Median : 214.7 Median : 1.430

Mean : 881.4 Mean : 2328.3 Mean : 3.994

3rd Qu.: 199.4 3rd Qu.: 767.3 3rd Qu.: 2.830

Max. :390133.8 Max. :891408.9 Max. : 473.000

Total.term.liabilities...tangible.net.worth

Min. :-325.600

1st Qu.: 0.050

Median : 0.340

Mean : 1.844

3rd Qu.: 1.000

Max. : 456.000

Contingent.liabilities...Net.worth.... Contingent.liabilities

Min. : 0.00 Min. : 0.1

1st Qu.: 0.00 1st Qu.: 16.0

Median : 5.33 Median : 38.0

Mean : 53.94 Mean : 632.7

3rd Qu.: 30.76 3rd Qu.: 84.2

Max. :14704.27 Max. :559506.8

Net.fixed.assets Investments Current.assets

Min. : 0.0 Min. : 0.00 Min. : 0.1

1st Qu.: 27.3 1st Qu.: 4.60 1st Qu.: 37.2

Median : 93.5 Median : 8.35 Median : 145.1

Mean : 1153.1 Mean : 416.58 Mean : 1272.0

3rd Qu.: 328.8 3rd Qu.: 16.10 3rd Qu.: 485.9

Max. :636604.6 Max. :199978.60 Max. :354815.2

Net.working.capital Quick.ratio..times. Current.ratio..times.

Min. :-63839.0 Min. : 0.000 Min. : 0.000

1st Qu.: -1.0 1st Qu.: 0.420 1st Qu.: 0.940

Median : 16.2 Median : 0.670 Median : 1.230

Mean : 137.5 Mean : 1.382 Mean : 2.107

3rd Qu.: 81.6 3rd Qu.: 1.020 3rd Qu.: 1.690

Max. : 85782.8 Max. :341.000 Max. :505.000

Debt.to.equity.ratio..times. Cash.to.current.liabilities..times.

Min. : 0.00 Min. : 0.0000

1st Qu.: 0.22 1st Qu.: 0.0200

Median : 0.79 Median : 0.0700

Mean : 2.78 Mean : 0.4794

3rd Qu.: 1.75 3rd Qu.: 0.1900

Max. :456.00 Max. :165.0000

Cash.to.average.cost.of.sales.per.day Creditors.turnover Debtors.turnover

Min. : 0.00 Min. : 0.000 Min. : 0.00

1st Qu.: 2.89 1st Qu.: 3.940 1st Qu.: 4.01

Median : 8.03 Median : 6.095 Median : 6.32

Mean : 154.83 Mean : 14.566 Mean : 16.05

3rd Qu.: 21.15 3rd Qu.: 10.550 3rd Qu.: 10.82

Max. :128040.76 Max. :2401.000 Max. :3135.20

Finished.goods.turnover WIP.turnover Raw.material.turnover

Min. : -0.09 Min. : -0.18 Min. : -2.00

1st Qu.: 10.21 1st Qu.: 5.93 1st Qu.: 3.41

Median : 17.27 Median : 9.76 Median : 6.40

Mean : 72.49 Mean : 24.65 Mean : 17.80

3rd Qu.: 30.72 3rd Qu.: 16.94 3rd Qu.: 10.92

Max. :17947.60 Max. :5651.40 Max. :21092.00

Shares.outstanding Equity.face.value EPS

Min. :-2.147e+09 Min. :-999999 Min. :-843181.8

1st Qu.: 2.210e+06 1st Qu.: 10 1st Qu.: 0.0

Median : 4.672e+06 Median : 10 Median : 1.4

Mean : 1.867e+07 Mean : -1071 Mean : -220.3

3rd Qu.: 8.320e+06 3rd Qu.: 10 3rd Qu.: 9.6

Max. : 4.130e+09 Max. : 100000 Max. : 34522.5

Adjusted.EPS Total.liabilities PE.on.BSE

Min. :-843181.8 Min. : 0.1 Min. :-1116.64

1st Qu.: 0.0 1st Qu.: 91.3 1st Qu.: 9.10

Median : 1.2 Median : 309.7 Median : 9.10

Mean : -221.5 Mean : 3443.4 Mean : 29.95

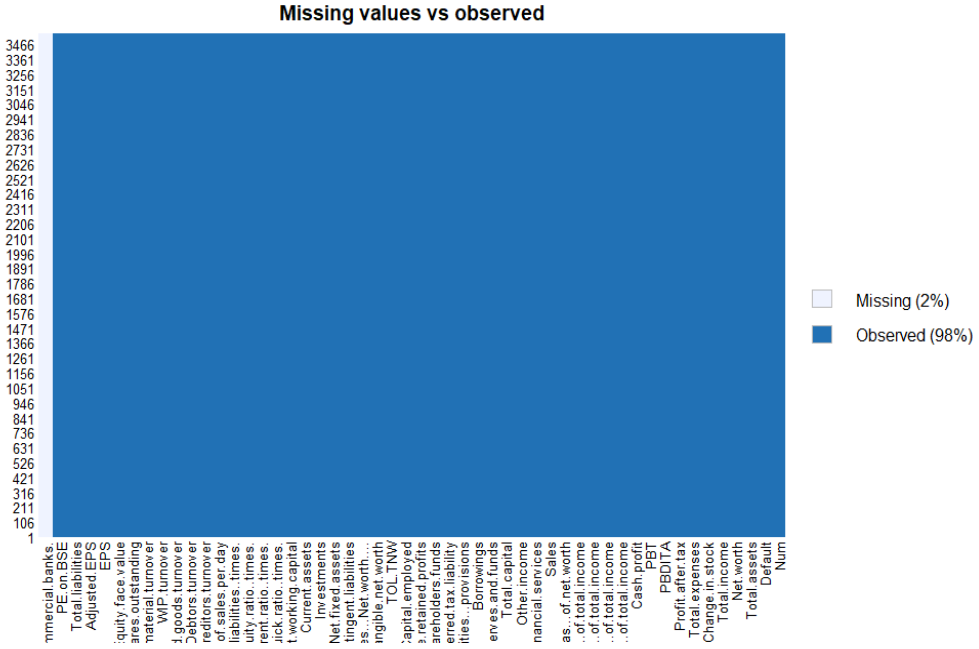
3rd Qu.: 7.5 3rd Qu.: 1098.7 3rd Qu.: 9.10

Max. : 34522.5 Max. :1176509.2 Max. :51002.74

str(raw\_data\_prepared)

missmap(raw\_data\_prepared, main = "Missing values vs observed")

2 percent values are Missing values if we look from all columns in total after missing data treatment.



**Removing the only column which has all values as NA, ie ‘Deposits..accepted.by.commercial.banks’**

raw\_data\_prepared<- raw\_data\_prepared[,-22]

colnames(raw\_data\_prepared)

[1] "Num"

[2] "Default"

[3] "Total.assets"

[4] "Net.worth"

[5] "Total.income"

[6] "Change.in.stock"

[7] "Total.expenses"

[8] "Profit.after.tax"

[9] "PBDITA"

[10] "PBT"

[11] "Cash.profit"

[12] "PBDITA.as...of.total.income"

[13] "PBT.as...of.total.income"

[14] "PAT.as...of.total.income"

[15] "Cash.profit.as...of.total.income"

[16] "PAT.as...of.net.worth"

[17] "Sales"

[18] "Income.from.financial.services"

[19] "Other.income"

[20] "Total.capital"

[21] "Reserves.and.funds"

[22] "Deferred.tax.liability"

[23] "Shareholders.funds"

[24] "Cumulative.retained.profits"

[25] "Capital.employed"

[26] "TOL.TNW"

[27] "Total.term.liabilities...tangible.net.worth"

[28] "Contingent.liabilities...Net.worth...."

[29] "Contingent.liabilities"

[30] "Net.fixed.assets"

[31] "Investments"

[32] "Current.assets"

[33] "Net.working.capital"

[34] "Quick.ratio..times."

[35] "Current.ratio..times."

[36] "Debt.to.equity.ratio..times."

[37] "Cash.to.current.liabilities..times."

[38] "Cash.to.average.cost.of.sales.per.day"

[39] "Creditors.turnover"

[40] "Debtors.turnover"

[41] "Finished.goods.turnover"

[42] "WIP.turnover"

[43] "Raw.material.turnover"

[44] "Shares.outstanding"

[45] "Equity.face.value"

[46] "EPS"

[47] "Adjusted.EPS"

[48] "Total.liabilities"

[49] "PE.on.BSE"

dim(raw\_data\_prepared)

[1] 3541 49

**Default Rate for the data set,**

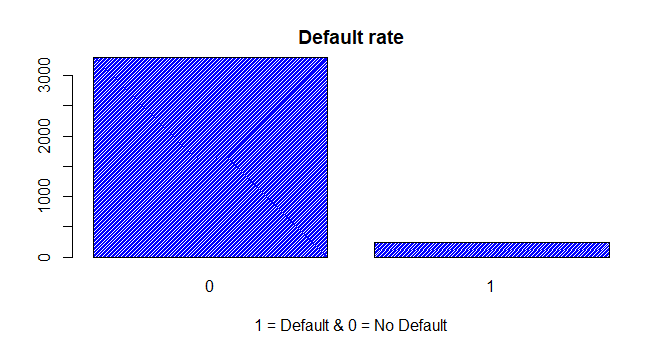
default\_rate = (sum(raw\_data\_prepared$Default)/(nrow(raw\_data\_prepared)))\*100

paste("Default Rate for the dataset is ", default\_rate, "%")

|  |
| --- |
| "Default Rate for the dataset is 6.60830273933917 %"  **Convert "Default" variable as categorical variable,**  raw\_data\_prepared$Default <- as.factor(raw\_data\_prepared$Default) |
|  |
| |  | | --- | |  | |

# **2. EDA, Outlier Treatment - Univariate and Bivariate Analysis**

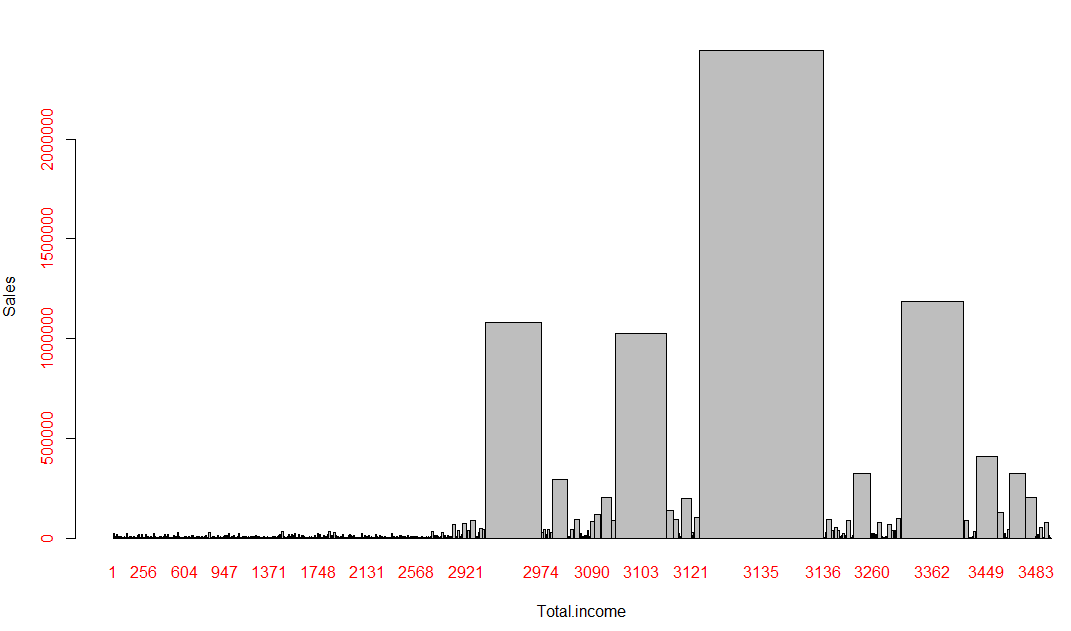
Plotting bar plot companies which default,



It can be seen from the graph that most companies are not defaulters.

barplot(raw\_data\_prepared$Total.income,raw\_data\_prepared$Sales, xlab = "Total.income", ylab = "Sales", axisnames = TRUE,

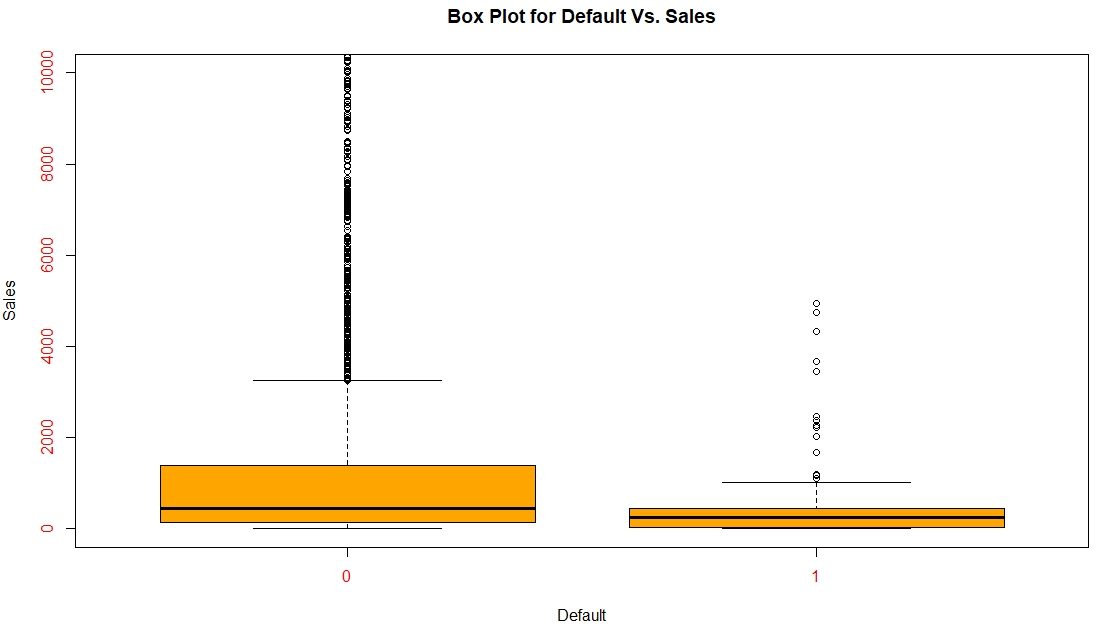
names.arg = raw\_data\_prepared$Num,col.axis = "Red")



We can see that, companies having more sales have high total income. Also, Company number 3135 has higher value for both parameters which indicate it could be one of the outlier.

p1 <-plot(raw\_data\_prepared$Default,raw\_data\_prepared$Sales, xlab = "Default", ylab = "Sales",

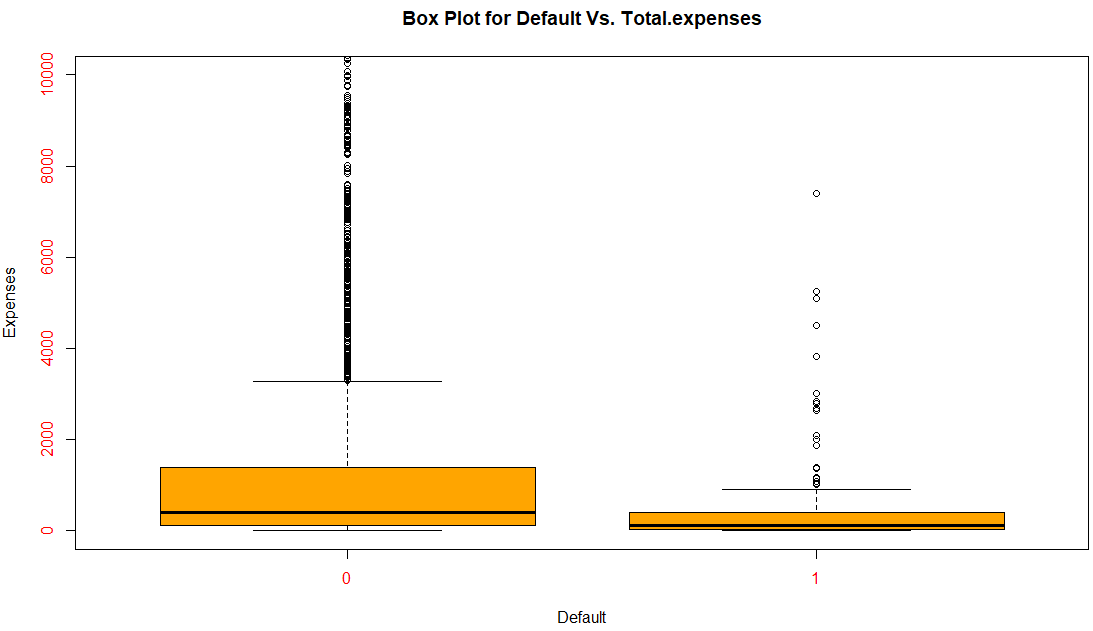
col.axis = "Red",col = 'orange', ylim = c(0,10000), main = "Box Plot for Default Vs. Sales")



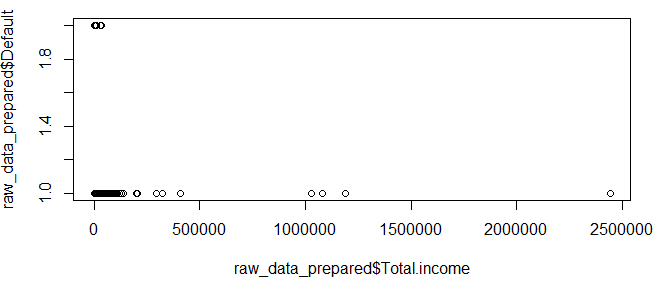
We can see companies which default have sales lesser than companies which don't Default.

p2 <-plot(raw\_data\_prepared$Default,raw\_data\_prepared$Total.expenses, xlab = "Default", ylab = "Expenses",

col.axis = "Red", col = 'orange', ylim = c(0,10000), main = "Box Plot for Default Vs. Total.expenses")



plot(raw\_data\_prepared$Total.income,raw\_data\_prepared$Default)



Above plot indicates companies who will probably default.

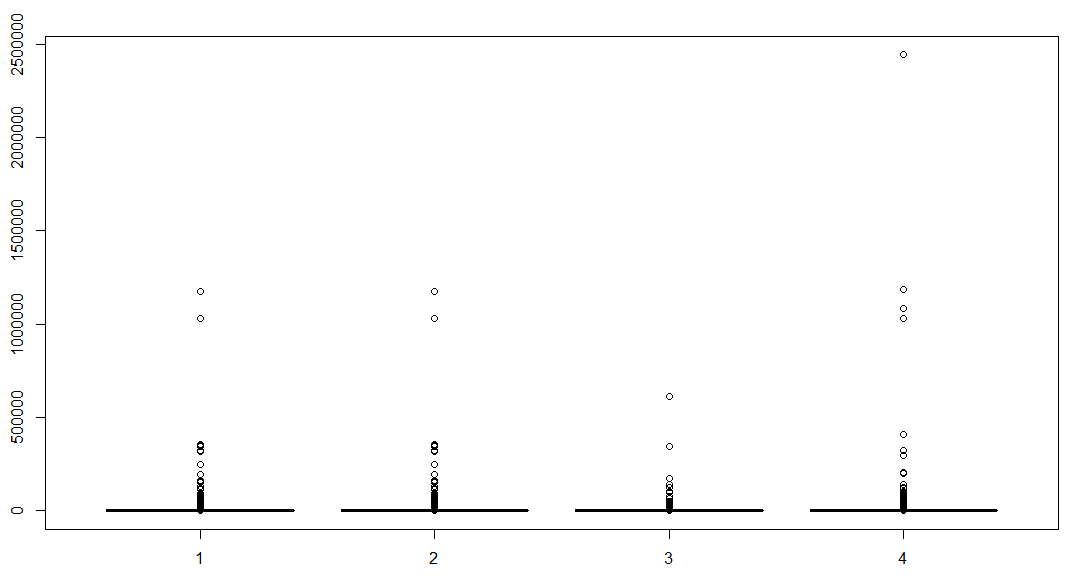
ggplot2::aes(raw\_data\_prepared$Total.expenses,raw\_data\_prepared$Cash.profit)

Aesthetic mapping:

\* `x` -> `raw\_data\_prepared$Total.expenses`

\* `y` -> `raw\_data\_prepared$Cash.profit`

boxplot(raw\_data\_prepared$Total.liabilities,raw\_data\_prepared$Total.assets,raw\_data\_prepared$Net.worth,raw\_data\_prepared$Total.income, col = "red")



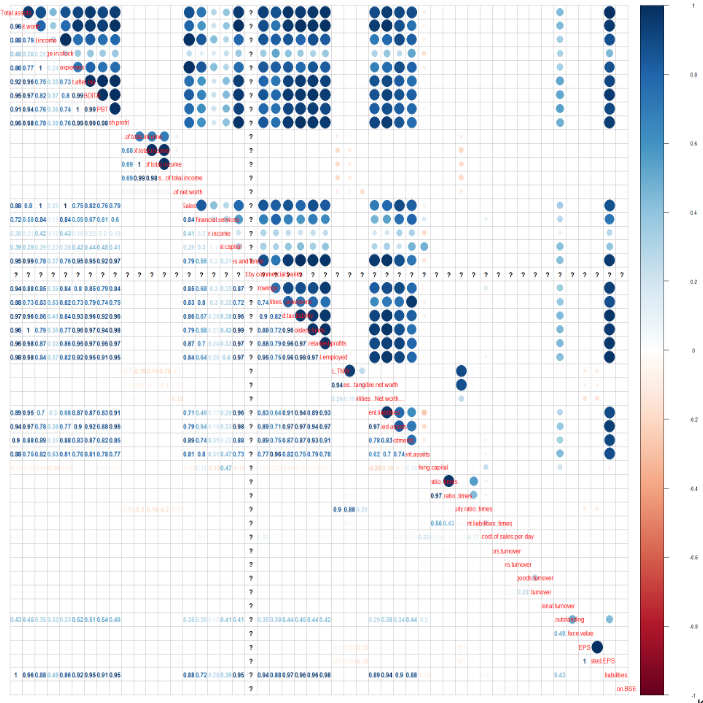
**Check for Multicollinearity - Plot the graph based on Multi-collinearity**

library(corrplot)

library(Hmisc)

numericdata<-raw\_data\_prepared[,c(-1,-2)]

print(cor(numericdata),digits = 3)

****

**Outlier Treatment Using capping for 95% and 5%,**

capOutlier <- 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)

}

raw\_data\_prepared$Total.income=capOutlier(raw\_data\_prepared$Total.income)

raw\_data\_prepared$Total.assets=capOutlier(raw\_data\_prepared$Total.assets)

raw\_data\_prepared$Total.expenses=capOutlier(raw\_data\_prepared$Total.expenses)

raw\_data\_prepared$Net.worth=capOutlier(raw\_data\_prepared$Net.worth)

raw\_data\_prepared$Change.in.stock=capOutlier(raw\_data\_prepared$Change.in.stock)

raw\_data\_prepared$Profit.after.tax=capOutlier(raw\_data\_prepared$Profit.after.tax)

raw\_data\_prepared$PBDITA=capOutlier(raw\_data\_prepared$PBDITA)

raw\_data\_prepared$PBT=capOutlier(raw\_data\_prepared$PBT)

raw\_data\_prepared$Cash.profit=capOutlier(raw\_data\_prepared$Cash.profit)

raw\_data\_prepared$PBT.as...of.total.income=capOutlier(raw\_data\_prepared$Cash.profit)

raw\_data\_prepared$PBDITA.as...of.total.income=capOutlier(raw\_data\_prepared$PBDITA.as...of.total.income)

raw\_data\_prepared$PAT.as...of.total.income=capOutlier(raw\_data\_prepared$PAT.as...of.total.income)

raw\_data\_prepared$Cash.profit.as...of.total.income=capOutlier(raw\_data\_prepared$Cash.profit.as...of.total.income)

raw\_data\_prepared$PAT.as...of.net.worth=capOutlier(raw\_data\_prepared$PAT.as...of.net.worth)

raw\_data\_prepared$Sales=capOutlier(raw\_data\_prepared$Sales)

raw\_data\_prepared$Income.from.financial.services=capOutlier(raw\_data\_prepared$Income.from.financial.services)

raw\_data\_prepared$Other.income=capOutlier(raw\_data\_prepared$Other.income)

raw\_data\_prepared$Total.capital=capOutlier(raw\_data\_prepared$Total.capital)

raw\_data\_prepared$Reserves.and.funds=capOutlier(raw\_data\_prepared$Reserves.and.funds)

raw\_data\_prepared$Borrowings=capOutlier(raw\_data\_prepared$Borrowings)

raw\_data\_prepared$Current.liabilities...provisions=capOutlier(raw\_data\_prepared$Current.liabilities...provisions)

raw\_data\_prepared$Deferred.tax.liability=capOutlier(raw\_data\_prepared$Deferred.tax.liability)

raw\_data\_prepared$Shareholders.funds=capOutlier(raw\_data\_prepared$Shareholders.funds)

raw\_data\_prepared$Cumulative.retained.profits=capOutlier(raw\_data\_prepared$Cumulative.retained.profits)

raw\_data\_prepared$Capital.employed=capOutlier(raw\_data\_prepared$Capital.employed)

raw\_data\_prepared$TOL.TNW=capOutlier(raw\_data\_prepared$TOL.TNW)

raw\_data\_prepared$Total.term.liabilities...tangible.net.worth=capOutlier(raw\_data\_prepared$Total.term.liabilities...tangible.net.worth)

raw\_data\_prepared$Contingent.liabilities...Net.worth....=capOutlier(raw\_data\_prepared$Contingent.liabilities...Net.worth....)

raw\_data\_prepared$Contingent.liabilities=capOutlier(raw\_data\_prepared$Contingent.liabilities)

raw\_data\_prepared$Net.fixed.assets=capOutlier(raw\_data\_prepared$Net.fixed.assets)

raw\_data\_prepared$Investments=capOutlier(raw\_data\_prepared$Investments)

raw\_data\_prepared$Current.assets=capOutlier(raw\_data\_prepared$Current.assets)

raw\_data\_prepared$Net.working.capital=capOutlier(raw\_data\_prepared$Net.working.capital)

#raw\_data\_prepared$Quick.ratio..times.=capOutlier(raw\_data\_prepared$Quick.rao..times.)

#raw\_data\_prepared$Current.ratio..times.=capOutlier(raw\_data\_prepared$Current.rtiatio..times.)

raw\_data\_prepared$Debt.to.equity.ratio..times.=capOutlier(raw\_data\_prepared$Debt.to.equity.ratio..times.)

raw\_data\_prepared$Cash.to.average.cost.of.sales.per.day=capOutlier(raw\_data\_prepared$Cash.to.average.cost.of.sales.per.day)

raw\_data\_prepared$Cash.to.current.liabilities..times.=capOutlier(raw\_data\_prepared$Cash.to.current.liabilities..times.)

raw\_data\_prepared$Creditors.turnover=capOutlier(raw\_data\_prepared$Creditors.turnover)

raw\_data\_prepared$Debtors.turnover=capOutlier(raw\_data\_prepared$Debtors.turnover)

raw\_data\_prepared$Finished.goods.turnover=capOutlier(raw\_data\_prepared$Finished.goods.turnover)

raw\_data\_prepared$WIP.turnover=capOutlier(raw\_data\_prepared$WIP.turnover)

raw\_data\_prepared$Raw.material.turnover=capOutlier(raw\_data\_prepared$Raw.material.turnover)

raw\_data\_prepared$Shares.outstanding=capOutlier(raw\_data\_prepared$Shares.outstanding)

raw\_data\_prepared$Equity.face.value=capOutlier(raw\_data\_prepared$Equity.face.value)

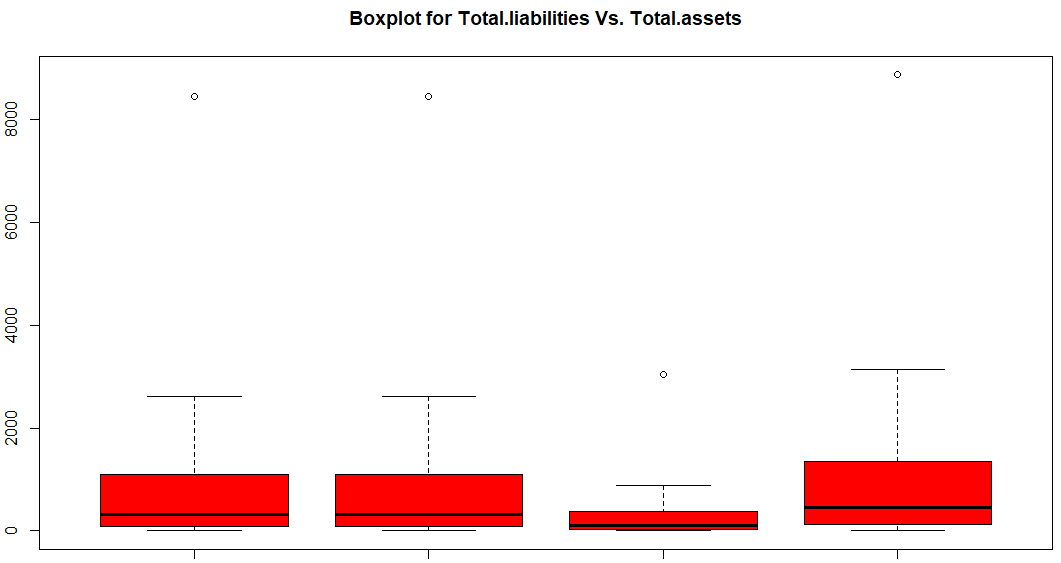
raw\_data\_prepared$EPS=capOutlier(raw\_data\_prepared$EPS)

raw\_data\_prepared$Adjusted.EPS=capOutlier(raw\_data\_prepared$Adjusted.EPS)

raw\_data\_prepared$Total.liabilities=capOutlier(raw\_data\_prepared$Total.liabilities)

raw\_data\_prepared$PE.on.BSE=capOutlier(raw\_data\_prepared$PE.on.BSE)

boxplot(raw\_data\_prepared$Total.liabilities,raw\_data\_prepared$Total.assets,raw\_data\_prepared$Net.worth,raw\_data\_prepared$Total.income, col = "red", main = "Boxplot for Total.liabilities Vs. Total.assets")



write.csv(raw\_data\_prepared,"raw\_data\_prepared\_updated.csv")

# **Creating new Variables in Ration format to bring the scale of all companies on the same scale**

New variables created on some critical variables as per business knowledge and after seeing Outlier behaviour, they have been created as ratios to scale down the number data for companies.

**Calculations has been done in Excel and then data is read again with new variables.**

**Formulae used:**

Net worth/ Total assets = NW\_new,

Total income / Total assets= TI\_new,

Total expenses / Total assets = TE\_new,

Profit after tax / Total assets = PAT\_new,

PBT / Total assets = PBT\_new,

Sales / Total assets = Sales\_new.

Current liabilities & provisions / Total assets= CLP\_new,

Capital employed / Total assets = CE\_new,

Net fixed assets / Total assets = NFA\_new,

Investments / Total assets = Investments\_new,

Total liabilities / Total assets = TL\_new

Shares.outstanding / Total assets = SO\_new

raw\_data\_prepared\_updated\_data<-read.csv("raw\_data\_prepared\_updated.csv")

raw\_data\_prepared\_updated\_data$Default <- as.factor(raw\_data\_prepared\_updated\_data$Default)

# **4. Dimension Reduction and Logistic Model development on train dataset**

**a) Applying Logistic Regression and Interpret the Regression model output**

Build the initial Logistic Regression Model taking all independent variables and new variables into consideration.

logit <- glm(raw\_data\_prepared\_updated\_data$Default~ .,family = 'binomial',data = raw\_data\_prepared\_updated\_data)

logit

Call: glm(formula = raw\_data\_prepared\_updated\_data$Default ~ ., family = "binomial",

data = raw\_data\_prepared\_updated\_data)

Coefficients:

(Intercept)

-3.010e+00

X

-1.489e-01

Num

1.489e-01

Total.assets

1.314e-04

Net.worth

-7.672e-04

Total.income

1.878e-03

Change.in.stock

6.840e-03

Total.expenses

-1.174e-03

Profit.after.tax

1.146e-02

PBDITA

-2.542e-03

PBT

-4.046e-03

Cash.profit

-3.108e-03

PBDITA.as...of.total.income

-1.864e-02

PBT.as...of.total.income

NA

PAT.as...of.total.income

-1.103e-02

Cash.profit.as...of.total.income

-3.223e-02

PAT.as...of.net.worth

-3.631e-02

Sales

-7.757e-04

Income.from.financial.services

-4.101e-03

Other.income

9.815e-03

Total.capital

-2.720e-03

Reserves.and.funds

-5.567e-04

Borrowings

-5.848e-04

Current.liabilities...provisions

7.453e-04

Deferred.tax.liability

1.071e-03

Shareholders.funds

1.051e-04

Cumulative.retained.profits

-8.687e-03

Capital.employed

5.203e-04

TOL.TNW

1.320e-01

Total.term.liabilities...tangible.net.worth

-9.052e-02

Contingent.liabilities...Net.worth....

3.353e-03

Contingent.liabilities

-7.838e-04

Net.fixed.assets

1.921e-04

Investments

-4.145e-05

Current.assets

-4.204e-04

Net.working.capital

-7.705e-05

Quick.ratio..times.

-3.897e-03

Current.ratio..times.

1.860e-03

Debt.to.equity.ratio..times.

2.506e-01

Cash.to.current.liabilities..times.

4.604e-01

Cash.to.average.cost.of.sales.per.day

1.250e-03

Creditors.turnover

-1.674e-02

Debtors.turnover

4.300e-03

Finished.goods.turnover

-1.136e-03

WIP.turnover

-3.750e-03

Raw.material.turnover

-2.305e-02

Shares.outstanding

-1.256e-09

Equity.face.value

1.220e-03

EPS

-2.616e-02

Adjusted.EPS

-1.659e-03

Total.liabilities

NA

PE.on.BSE

-2.535e-02

Degrees of Freedom: 3540 Total (i.e. Null); 3491 Residual

Null Deviance: 1724

Residual Deviance: 888 AIC: 988

summary(logit)

Call:

glm(formula = raw\_data\_prepared\_updated\_data$Default ~ ., family = "binomial",

data = raw\_data\_prepared\_updated\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4938 -0.2359 -0.1115 -0.0117 3.5185

Coefficients: (2 not defined because of singularities)

Estimate Std. Error z value

(Intercept) -3.010e+00 3.594e-01 -8.376

X -1.489e-01 1.535e-01 -0.970

Num 1.489e-01 1.533e-01 0.972

Total.assets 1.314e-04 2.002e-04 0.656

Net.worth -7.672e-04 5.785e-04 -1.326

Total.income 1.878e-03 5.305e-04 3.540

Change.in.stock 6.840e-03 2.976e-03 2.298

Total.expenses -1.174e-03 4.536e-04 -2.588

Profit.after.tax 1.146e-02 5.535e-03 2.070

PBDITA -2.542e-03 1.114e-03 -2.282

PBT -4.046e-03 4.339e-03 -0.933

Cash.profit -3.108e-03 2.227e-03 -1.396

PBDITA.as...of.total.income -1.864e-02 1.563e-02 -1.193

PBT.as...of.total.income NA NA NA

PAT.as...of.total.income -1.103e-02 1.565e-02 -0.705

Cash.profit.as...of.total.income -3.223e-02 2.345e-02 -1.375

PAT.as...of.net.worth -3.631e-02 7.228e-03 -5.024

Sales -7.757e-04 3.244e-04 -2.391

Income.from.financial.services -4.101e-03 8.018e-03 -0.511

Other.income 9.815e-03 1.210e-02 0.811

Total.capital -2.720e-03 1.446e-03 -1.881

Reserves.and.funds -5.567e-04 4.765e-04 -1.168

Borrowings -5.848e-04 3.346e-04 -1.748

Current.liabilities...provisions 7.453e-04 5.284e-04 1.411

Deferred.tax.liability 1.071e-03 2.495e-03 0.429

Shareholders.funds 1.051e-04 5.759e-04 0.182

Cumulative.retained.profits -8.687e-03 1.945e-03 -4.467

Capital.employed 5.203e-04 3.097e-04 1.680

TOL.TNW 1.320e-01 4.204e-02 3.140

Total.term.liabilities...tangible.net.worth -9.052e-02 1.134e-01 -0.798

Contingent.liabilities...Net.worth.... 3.353e-03 1.976e-03 1.696

Contingent.liabilities -7.838e-04 5.075e-04 -1.544

Net.fixed.assets 1.921e-04 3.397e-04 0.566

Investments -4.145e-05 7.971e-04 -0.052

Current.assets -4.204e-04 4.158e-04 -1.011

Net.working.capital -7.705e-05 1.087e-03 -0.071

Quick.ratio..times. -3.897e-03 4.022e-02 -0.097

Current.ratio..times. 1.860e-03 3.166e-02 0.059

Debt.to.equity.ratio..times. 2.506e-01 8.052e-02 3.113

Cash.to.current.liabilities..times. 4.604e-01 3.046e-01 1.512

Cash.to.average.cost.of.sales.per.day 1.250e-03 1.783e-03 0.701

Creditors.turnover -1.674e-02 1.175e-02 -1.425

Debtors.turnover 4.300e-03 9.456e-03 0.455

Finished.goods.turnover -1.136e-03 3.359e-03 -0.338

WIP.turnover -3.750e-03 8.357e-03 -0.449

Raw.material.turnover -2.305e-02 1.383e-02 -1.667

Shares.outstanding -1.256e-09 1.067e-08 -0.118

Equity.face.value 1.220e-03 4.210e-03 0.290

EPS -2.616e-02 3.026e-02 -0.864

Adjusted.EPS -1.659e-03 2.695e-02 -0.062

Total.liabilities NA NA NA

PE.on.BSE -2.535e-02 1.391e-02 -1.822

Pr(>|z|)

(Intercept) < 2e-16 \*\*\*

X 0.33211

Num 0.33124

Total.assets 0.51171

Net.worth 0.18479

Total.income 0.00040 \*\*\*

Change.in.stock 0.02156 \*

Total.expenses 0.00965 \*\*

Profit.after.tax 0.03847 \*

PBDITA 0.02250 \*

PBT 0.35104

Cash.profit 0.16279

PBDITA.as...of.total.income 0.23290

PBT.as...of.total.income NA

PAT.as...of.total.income 0.48072

Cash.profit.as...of.total.income 0.16928

PAT.as...of.net.worth 5.07e-07 \*\*\*

Sales 0.01679 \*

Income.from.financial.services 0.60904

Other.income 0.41740

Total.capital 0.05996 .

Reserves.and.funds 0.24266

Borrowings 0.08049 .

Current.liabilities...provisions 0.15836

Deferred.tax.liability 0.66763

Shareholders.funds 0.85523

Cumulative.retained.profits 7.92e-06 \*\*\*

Capital.employed 0.09294 .

TOL.TNW 0.00169 \*\*

Total.term.liabilities...tangible.net.worth 0.42473

Contingent.liabilities...Net.worth.... 0.08980 .

Contingent.liabilities 0.12252

Net.fixed.assets 0.57170

Investments 0.95853

Current.assets 0.31199

Net.working.capital 0.94350

Quick.ratio..times. 0.92280

Current.ratio..times. 0.95314

Debt.to.equity.ratio..times. 0.00185 \*\*

Cash.to.current.liabilities..times. 0.13060

Cash.to.average.cost.of.sales.per.day 0.48315

Creditors.turnover 0.15411

Debtors.turnover 0.64928

Finished.goods.turnover 0.73520

WIP.turnover 0.65357

Raw.material.turnover 0.09547 .

Shares.outstanding 0.90629

Equity.face.value 0.77199

EPS 0.38733

Adjusted.EPS 0.95092

Total.liabilities NA

PE.on.BSE 0.06840 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1723.67 on 3540 degrees of freedom

Residual deviance: 888.04 on 3491 degrees of freedom

AIC: 988.04

Number of Fisher Scoring iterations: 11

Above Run will helps us to reduce the variables to be considered significant for model building and thus help in variable reduction.

**b) Applying Logistic Regression with relevant variables and Interpret the Regression model output.**

logit2<-glm(raw\_data\_prepared\_updated\_data$Default~ Total.income + Change.in.stock +Total.expenses +PAT.as...of.net.worth +Sales +Total.capital+Borrowings+Cumulative.retained.profits +Capital.employed+Debt.to.equity.ratio..times +Cash.to.current.liabilities..times.+PE.on.BSE+NW\_new+TE\_new+Sales\_new+NFA\_new + SO\_new ,family = 'binomial',data = raw\_data\_prepared\_updated\_data)

logit2

Call: glm(formula = raw\_data\_prepared\_updated\_data$Default ~ Total.income +

Change.in.stock + Total.expenses + PAT.as...of.net.worth +

Sales + Total.capital + Borrowings + Cumulative.retained.profits +

Capital.employed + Debt.to.equity.ratio..times. + Cash.to.current.liabilities..times. +

PE.on.BSE, family = "binomial", data = raw\_data\_prepared\_updated\_data)

Coefficients:

(Intercept) Total.income

-2.8403189 0.0009991

Change.in.stock Total.expenses

0.0050296 -0.0005795

PAT.as...of.net.worth Sales

-0.0606055 -0.0004867

Total.capital Borrowings

-0.0032200 -0.0001751

Cumulative.retained.profits Capital.employed

-0.0103270 0.0001761

Debt.to.equity.ratio..times. Cash.to.current.liabilities..times.

0.3552957 0.2468643

PE.on.BSE

-0.0321213

Degrees of Freedom: 3540 Total (i.e. Null); 3528 Residual

Null Deviance: 1724

Residual Deviance: 991.2 AIC: 1017

summary(logit2)

Call:

glm(formula = raw\_data\_prepared\_updated\_data$Default ~ Total.income +

Change.in.stock + Total.expenses + PAT.as...of.net.worth +

Sales + Total.capital + Borrowings + Cumulative.retained.profits +

Capital.employed + Debt.to.equity.ratio..times. + Cash.to.current.liabilities..times. +

PE.on.BSE, family = "binomial", data = raw\_data\_prepared\_updated\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4012 -0.2826 -0.1475 -0.0239 4.0181

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.8403189 0.1773734 -16.013 < 2e-16

Total.income 0.0009991 0.0003498 2.856 0.004293

Change.in.stock 0.0050296 0.0025565 1.967 0.049142

Total.expenses -0.0005795 0.0002939 -1.972 0.048662

PAT.as...of.net.worth -0.0606055 0.0057138 -10.607 < 2e-16

Sales -0.0004867 0.0002549 -1.909 0.056199

Total.capital -0.0032200 0.0009377 -3.434 0.000595

Borrowings -0.0001751 0.0002356 -0.743 0.457333

Cumulative.retained.profits -0.0103270 0.0017380 -5.942 2.82e-09

Capital.employed 0.0001761 0.0001554 1.133 0.257222

Debt.to.equity.ratio..times. 0.3552957 0.0338452 10.498 < 2e-16

Cash.to.current.liabilities..times. 0.2468643 0.2329264 1.060 0.289218

PE.on.BSE -0.0321213 0.0125356 -2.562 0.010395

(Intercept) \*\*\*

Total.income \*\*

Change.in.stock \*

Total.expenses \*

PAT.as...of.net.worth \*\*\*

Sales .

Total.capital \*\*\*

Borrowings

Cumulative.retained.profits \*\*\*

Capital.employed

Debt.to.equity.ratio..times. \*\*\*

Cash.to.current.liabilities..times.

PE.on.BSE \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1723.67 on 3540 degrees of freedom

Residual deviance: 991.17 on 3528 degrees of freedom

AIC: 1017.2

Number of Fisher Scoring iterations: 10

**c) Applying Logistic Regression again with relevant variables and Interpret the Regression model output.**

logit3<-glm(raw\_data\_prepared\_updated\_data$Default~ PAT.as...of.net.worth

+Total.capital+Cumulative.retained.profits +Debt.to.equity.ratio..times.

+Cash.to.current.liabilities..times.+PE.on.BSE+NW\_new+TE\_new+Sales\_new+NFA\_new + SO\_new

,family = 'binomial',data = raw\_data\_prepared\_updated\_data)

logit3

Call: glm(formula = raw\_data\_prepared\_updated\_data$Default ~ PAT.as...of.net.worth +

Total.capital + Cumulative.retained.profits + Debt.to.equity.ratio..times. +

Cash.to.current.liabilities..times. + PE.on.BSE, family = "binomial",

data = raw\_data\_prepared\_updated\_data)

Coefficients:

(Intercept) PAT.as...of.net.worth

-2.846389 -0.058870

Total.capital Cumulative.retained.profits

-0.002588 -0.009758

Debt.to.equity.ratio..times. Cash.to.current.liabilities..times.

0.353294 0.266926

PE.on.BSE

-0.031190

Degrees of Freedom: 3540 Total (i.e. Null); 3534 Residual

Null Deviance: 1724

Residual Deviance: 1004 AIC: 1018

summary(logit3)

Call:

glm(formula = raw\_data\_prepared\_updated\_data$Default ~ PAT.as...of.net.worth +

Total.capital + Cumulative.retained.profits + Debt.to.equity.ratio..times. +

Cash.to.current.liabilities..times. + PE.on.BSE, family = "binomial",

data = raw\_data\_prepared\_updated\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5294 -0.2862 -0.1534 -0.0268 4.0500

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.8463895 0.1756288 -16.207 < 2e-16

PAT.as...of.net.worth -0.0588695 0.0055930 -10.526 < 2e-16

Total.capital -0.0025881 0.0006073 -4.262 2.03e-05

Cumulative.retained.profits -0.0097580 0.0016399 -5.950 2.68e-09

Debt.to.equity.ratio..times. 0.3532941 0.0324990 10.871 < 2e-16

Cash.to.current.liabilities..times. 0.2669256 0.2280035 1.171 0.2417

PE.on.BSE -0.0311898 0.0124076 -2.514 0.0119

(Intercept) \*\*\*

PAT.as...of.net.worth \*\*\*

Total.capital \*\*\*

Cumulative.retained.profits \*\*\*

Debt.to.equity.ratio..times. \*\*\*

Cash.to.current.liabilities..times.

PE.on.BSE \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1723.7 on 3540 degrees of freedom

Residual deviance: 1003.9 on 3534 degrees of freedom

AIC: 1017.9

Number of Fisher Scoring iterations: 10

**From Below output we can see that the most significant variables for our model are:**

PAT.as...of.net.worth

Total.capital

Cumulative.retained.profits and others

Debt.to.equity.ratio..times.#Cash.to.current.liabilities..times.

PE.on.BSE

TE\_new

NFA\_new

SO\_new

NW\_new

Sales\_new

**Checking for Variation Inflation Factor,**

x<-vif(logit3)

x

PAT.as...of.net.worth Total.capital

1.136645 1.361098

Cumulative.retained.profits Debt.to.equity.ratio..times.

1.400794 1.109343

Cash.to.current.liabilities..times. PE.on.BSE

1.040443 1.053278

**Likelihood Ratio Test (LR test),**

lrtest(logit3)

Likelihood ratio test

Model 1: raw\_data\_prepared\_updated\_data$Default ~ PAT.as...of.net.worth +

Total.capital + Cumulative.retained.profits + Debt.to.equity.ratio..times. +

Cash.to.current.liabilities..times. + PE.on.BSE

Model 2: raw\_data\_prepared\_updated\_data$Default ~ 1

#Df LogLik Df Chisq Pr(>Chisq)

1 7 -501.97

2 1 -861.83 -6 719.72 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Looking for McFadden value,**

pR2(logit3)

odds<-exp(coef(logit3))

fitting null model for pseudo-r2

llh llhNull G2 McFadden r2ML

-501.9725200 -861.8334541 719.7218683 0.4175528 0.1839289

r2CU

0.4772490

# **5. Evaluating model output against actual values of train dataset using the validation metrics**

Model Performance Measures (Train)

* **Checking the prediction accuracy on the train data**

predict.logit<-predict.glm(logit3, newdata= raw\_data\_prepared\_updated\_data, type="response")

summary(predict.logit)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000000 0.001799 0.016728 0.066083 0.046884 0.976373

table.logit<- confusionMatrix(raw\_data\_prepared\_updated\_data$Default,predict.logit,cutoff = 0.9)

table.logit

* **ROC curve for train data,**

accuracy.logit<-

roc.logit<-roc(raw\_data\_prepared\_updated\_data$Default,predict.logit)

roc.logit

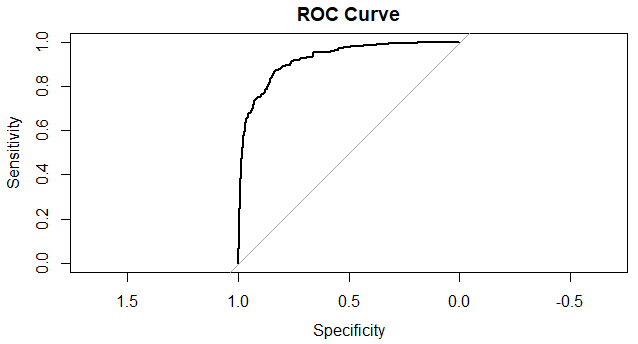
plot(roc.logit,main="ROC Curve")

Call:

roc.default(response = raw\_data\_prepared\_updated\_data$Default, predictor = predict.logit)

Data: predict.logit in 3307 controls (raw\_data\_prepared\_updated\_data$Default 0) < 234 cases (raw\_data\_prepared\_updated\_data$Default 1).

Area under the curve: 0.9231



From the output it is seen that the area under the curve (AUC) is 92.31% percent which is >70% and hence is good fit indicator.

* **KS score**

library(ROCR)

ks1<-prediction(predict.logit,raw\_data\_prepared\_updated\_data$Default)

perf\_m3\_train<-performance(ks1,"tpr","fpr")

ks\_stats<-round(max(attr(perf\_m3\_train,'y.values')[[1]]-attr(perf\_m3\_train,'x.values')[[1]]),4)\*100

ks\_stats

[1] 70.58

Since the KS stats is 70.58% which is greater than 20% indicates **good separation between the cumulative good rate & bad rate**.

* **Gini values for logistics regression**

library(ineq)

gini\_stats=round(ineq(predict.logit,type="Gini"),4)\*100

gini\_stats

[1] 77.64

The **gini value** for the model is **77.64** which is a **fair fit**.

# **Applying the model on test dataset**

Importing validation (test) dataset,

validation\_data <- read.csv("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment FRA/validation\_data\_updated.csv", header=FALSE)

str(validation\_data)

summary(validation\_data)

'data.frame': 716 obs. of 56 variables:

$ V1 : Factor w/ 716 levels "1","1000","101",..: 716 120 652 368 67 78 689 534 247 232 ...

$ V2 : Factor w/ 3 levels "0","1","Default...1": 3 1 1 2 1 1 1 1 1 1 ...

$ V3 : Factor w/ 682 levels "0.1","0.7","1",..: 682 675 569 483 646 637 142 640 328 626 ...

$ V4 : Factor w/ 637 levels "0.1","0.2","0.3",..: 637 288 233 64 218 346 598 553 34 422 ...

$ V5 : Factor w/ 658 levels "0","0.1","0.2",..: 658 236 596 487 374 25 207 543 504 618 ...

$ V6 : Factor w/ 421 levels "-0.1","-0.2",..: 421 201 166 64 6 284 5 163 133 280 ...

$ V7 : Factor w/ 665 levels "0","0.1","0.2",..: 665 237 604 508 387 28 225 518 514 619 ...

$ V8 : Factor w/ 490 levels "-0.1","-0.2",..: 490 115 212 58 51 236 97 98 24 329 ...

$ V9 : Factor w/ 567 levels "-0.1","-0.2",..: 567 290 93 28 481 554 66 362 101 510 ...

$ V10: Factor w/ 505 levels "-0.1","-0.2",..: 505 180 323 63 266 295 302 96 86 398 ...

$ V11: Factor w/ 519 levels "-0.1","-0.2",..: 519 201 404 44 338 447 297 347 502 419 ...

$ V12: Factor w/ 625 levels "-0.17","-0.82",..: 625 152 180 37 69 582 472 520 276 582 ...

$ V13: Factor w/ 583 levels "-0.07","-0.09",..: 583 516 442 95 134 349 184 282 101 500 ...

$ V14: Factor w/ 555 levels "-0.03","-0.06",..: 555 416 314 99 12 296 131 262 6 408 ...

$ V15: Factor w/ 580 levels "-0.03","-0.21",..: 580 540 469 61 94 466 131 463 119 488 ...

$ V16: Factor w/ 598 levels "-0.1","-0.14",..: 598 481 163 60 111 528 111 135 47 576 ...

$ V17: Factor w/ 651 levels "0.1","0.2","0.4",..: 651 228 588 470 367 17 210 319 436 605 ...

$ V18: Factor w/ 204 levels "0.1","0.2","0.3",..: 204 73 8 198 10 7 5 107 39 34 ...

$ V19: Factor w/ 169 levels "0","0.1","0.2",..: 169 18 3 57 16 59 2 18 18 2 ...

$ V20: Factor w/ 529 levels "0.1","0.2","0.4",..: 529 361 56 358 383 264 187 469 489 514 ...

$ V21: Factor w/ 636 levels "-0.1","-0.3",..: 636 430 629 295 212 371 503 422 333 448 ...

$ V22: Factor w/ 578 levels "0.2","0.3","0.5",..: 578 157 310 147 423 134 255 280 80 29 ...

$ V23: Factor w/ 601 levels "0.1","0.2","0.3",..: 601 332 23 255 540 303 416 5 534 348 ...

$ V24: Factor w/ 331 levels "0.1","0.2","0.3",..: 331 173 131 69 258 49 110 19 210 117 ...

$ V25: Factor w/ 636 levels "0.1","0.2","0.3",..: 636 292 234 66 216 349 596 552 36 422 ...

$ V26: Factor w/ 628 levels "-0.1","-0.2",..: 628 321 620 543 245 384 582 14 142 462 ...

$ V27: Factor w/ 669 levels "0.1","0.2","0.3",..: 669 438 491 101 577 471 56 589 254 436 ...

$ V28: Factor w/ 382 levels "-2.41","-350.48",..: 382 164 205 159 255 100 85 12 175 42 ...

$ V29: Factor w/ 246 levels "-0.89","-1.25",..: 246 33 77 15 84 16 17 11 82 6 ...

$ V30: Factor w/ 472 levels "0","0.09","0.1",..: 472 95 379 75 27 263 1 182 1 14 ...

$ V31: Factor w/ 390 levels "0.1","0.2","0.3",..: 390 215 42 49 13 24 245 15 245 16 ...

$ V32: Factor w/ 629 levels "0.1","0.2","0.4",..: 629 324 179 272 261 168 140 279 378 579 ...

$ V33: Factor w/ 264 levels "0","0.1","0.2",..: 264 233 233 8 233 233 233 72 95 233 ...

$ V34: Factor w/ 636 levels "0.1","0.2","0.3",..: 636 495 408 124 483 451 115 143 196 413 ...

$ V35: Factor w/ 614 levels "-0.1","-0.2",..: 614 243 228 166 613 557 560 273 541 537 ...

$ V36: Factor w/ 219 levels "0","0.01","0.02",..: 219 92 48 32 51 58 97 171 52 88 ...

$ V37: Factor w/ 277 levels "0","0.02","0.03",..: 277 109 117 40 101 97 161 174 134 97 ...

$ V38: Factor w/ 308 levels "0","0.01","0.02",..: 308 64 151 16 213 46 33 6 116 1 ...

$ V39: Factor w/ 116 levels "0","0.01","0.02",..: 116 10 4 5 9 9 1 93 4 35 ...

$ V40: Factor w/ 589 levels "0","0.1","0.13",..: 589 526 344 398 338 108 20 166 247 465 ...

$ V41: Factor w/ 536 levels "0","0.01","0.02",..: 536 397 51 161 426 379 426 28 105 167 ...

$ V42: Factor w/ 548 levels "0","0.01","0.06",..: 548 380 398 536 445 326 445 532 76 309 ...

$ V43: Factor w/ 551 levels "-0.09","0","0.07",..: 551 251 278 503 158 430 158 158 158 25 ...

$ V44: Factor w/ 547 levels "0","0.03","0.1",..: 547 220 184 446 50 450 50 50 162 434 ...

$ V45: Factor w/ 506 levels "0","0.01","0.06",..: 506 135 181 263 393 260 393 1 330 368 ...

$ V46: Factor w/ 557 levels "1.00E+05","1.00E+06",..: 557 368 44 364 392 252 93 540 380 545 ...

$ V47: Factor w/ 11 levels "1","10","100",..: 11 2 2 3 2 2 3 2 2 2 ...

$ V48: Factor w/ 523 levels "-0.01","-0.02",..: 523 256 172 101 90 436 460 112 102 402 ...

$ V49: Factor w/ 505 levels "-0.01","-0.02",..: 505 256 176 97 86 420 443 108 98 384 ...

$ V50: Factor w/ 682 levels "0.1","0.7","1",..: 682 675 569 483 646 637 142 640 328 626 ...

$ V51: Factor w/ 275 levels "-0.16","-0.4",..: 275 249 249 15 1 249 249 249 249 249 ...

$ V52: Factor w/ 716 levels "0","0.001230769",..: 716 616 389 347 696 419 391 27 564 335 ...

$ V53: Factor w/ 712 levels "0.000718907",..: 712 381 331 555 343 229 86 372 105 87 ...

$ V54: Factor w/ 712 levels "1.505882353",..: 712 472 159 669 535 400 666 24 135 47 ...

$ V55: Factor w/ 714 levels "0.001996279",..: 714 233 272 166 176 409 517 678 303 540 ...

$ V56: Factor w/ 715 levels "0.001133144",..: 715 604 363 299 687 379 362 13 458 306 ...

Replacing all NA values with median value for each column,

for (i in which(sapply(validation\_data, is.numeric))) {

validation\_data[is.na(validation\_data[, i]), i] <- median(validation\_data[, i], na.rm = TRUE)

}

Removing column as it contains no data,

validation\_data<- validation\_data[,-22]

Creating updated validation dataset after removing NA and outlier treatment,

validation\_data\_updated<-read.csv("validation\_data\_updated.csv")

validation\_data\_updated$Default...1 <- as.factor(validation\_data\_updated$Default...1)

**Model Performance Measures (Test)**

Confusion matrix interpretation for all models

* **Checking the prediction accuracy on the test data**

predict.logit2<-predict.glm(logit3, newdata= validation\_data\_updated, type="response")

summary(predict.logit2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000000 0.0006597 0.0165007 0.1229222 0.0529173 1.0000000

* **Preparing Confusion matrix**

table.logit2<- confusionMatrix(validation\_data\_updated$Default...1,predict.logit2,cutoff = 0.5)

table.logit2

[,1] [,2]

[1,] 0 628

[2,] 0 33

* **ROC curve for test data**

accuracy.logit2<-

roc.logit2<-roc(validation\_data\_updated$Default...1,predict.logit2)

roc.logit2

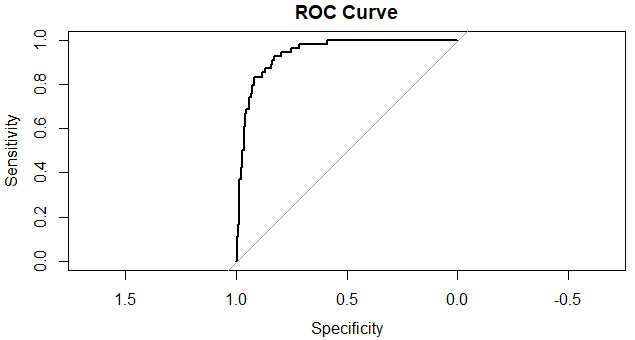
Call:

roc.default(response = validation\_data\_updated$Default...1, predictor = predict.logit2)

Data: predict.logit2 in 661 controls (validation\_data\_updated$Default...1 0) < 54 cases (validation\_data\_updated$Default...1 1).

Area under the curve: 0.9407

plot(roc.logit2,main="ROC Curve")



* **KS score**

library(ROCR)

ks1<-prediction(predict.logit,validation\_data\_updated$Default...1)

perf\_m3\_train<-performance(ks1,"tpr","fpr")

ks\_stats<-round(max(attr(perf\_m3\_train,'y.values')[[1]]-attr(perf\_m3\_train,'x.values')[[1]]),4)\*100

ks\_stats

[1] 70.58

* **Gini values for logistics regression**

library(ineq)

gini\_stats=round(ineq(predict.logit2,type="Gini"),4)\*100

gini\_stats

[1] 81.57

The gini value for the model is 81.57 which is a fair fit.

# **Model Accuracies and Business Insights**

Tweaking the model for better accuracy and use the final model to predict the final test data values and shared the business insights inferred from the model

|  |  |  |
| --- | --- | --- |
|  | **Raw Data** | **Validation Data** |
| **Accuracy** | 0.9048 | 0.8944 |
| **Sensitivity** | 0.9386 | 0.9023 |
| **Specificity** | 0.6833 | 0.7828 |
| **AUC** | 0.9231 | 0.9407 |
| **KS** | 0.7058 | 0.7058 |
| **Gini** | 0.7764 | 0.8157 |

After comparing the accuracy of the AUC, KS Score and Gini value it can be seen AUC gives better performance in terms of accuracy in our Logistic Regression Model.

Most prominent being

* PAT.as...of.net.worth
* Total.capital
* Cumulative.retained.profits and others
* Debt.to.equity.ratio..times.#Cash.to.current.liabilities..times.
* PE.on.BSE
* TE\_new
* NFA\_new
* SO\_new
* NW\_new
* Sales\_new

Cash to current liabilities have a positive coefficient, so it says they have more cash and less current liabilities, so companies having this more will default more as they are only focusing towards current liabilities.

Debt to equity ratio also have a positive coefficient , so we can say that companies having this parameter as positive have  for Debt and less equity , so they might not be able to make money through equity and will default.

* This means that an increase in these variables will be associated with a decreased probability of company Default.
* Better the ratios, lesser chances of company default but having said that if coefficient is negative only for like debt ratio.
* This shows greater PBT, lesser liabilities increase chances of company is not going to default as profit is more and taxation will not hamper loan returns.
* Total capital is also negative linked, so any company having more capital can pay back loans.
* Bigger companies have less chance of defaulting has they have more capital, more profit and more income.
* Cumulative retain profits is inversely linked, so if you have more overall profit, companies can pay back EOY.
* Better management of inventory i.e. converting inventory into final product quickly, getting your receivables quickly and commanding payable with enough time period helps.
* Companies with no leverage and with larger reserves have less chances of Default.
* Debt to equity ratio also have a positive coefficient , so we can say that companies having this parameter as positive have more Debt and less equity , so they might not be able to make money through equity and will default.
* Cash to current liabilities have a positive coefficient, so it says they have more cash and less current liabilities, so companies having this more will default more as they are not focusing towards current liabilities and will lose in current market in longer run as they will stack up liabilities.

# **Appendix**



Excel sheets include new variables created with ratios formed.