Finance & Risk Analytics

India Credit Risk

K .VINEET *PATNAIK*

Table of Contents

1.Understanding Data

A dataset is given with 52 variables and 3541 observations in it .The variables are divided across 4 main categories i.e Profitability,leverage,size of the company and liquidity . Our first aim is to read the data and understand it and check the data further for NA's and outliers present in the data i.e cleaning the data for better understanding and further processing .

```
getwd()
setwd("C:/Users/vineet patnaik/Desktop/R language/text files/")
dataset = read_excel('raw-data.xlsx')
library(readxl)
library(rms)
library(DMwR)
names(dataset)     ## all 52 variables names
head(dataset)
str(dataset)        ## category of variables
```

attach(dataset)

summary(dataset)

```
      Num        Networth Next Year   Total assets         Net worth          Total income        Change in stock
 Min.   :   1   Min.   :-74265.6   Min.   :      0.1   Min.   :     0.0   Min.   :      0.0   Min.   :-3029.40
 1st Qu.: 886   1st Qu.:    31.7   1st Qu.:     91.3   1st Qu.:    31.3   1st Qu.:    106.5   1st Qu.:   -1.80
 Median :1773   Median :   116.3   Median :    309.7   Median :   102.3   Median :    444.9   Median :    1.60
 Mean   :1772   Mean   :  1616.3   Mean   :   3443.4   Mean   :  1295.9   Mean   :   4582.8   Mean   :   41.49
 3rd Qu.:2658   3rd Qu.:   456.1   3rd Qu.:   1098.7   3rd Qu.:   377.3   3rd Qu.:   1440.9   3rd Qu.:   18.05
 Max.   :3545   Max.   :805773.4   Max.   :1176509.2   Max.   :613151.6   Max.   :2442828.2   Max.   :14185.50
                                                                          NA's   :198         NA's   :458
 Total expenses      Profit after tax       PBDITA             PBT               Cash profit
 Min.   :     -0.1   Min.   : -3908.30   Min.   :  -440.7   Min.   : -3894.80   Min.   : -2245.70
 1st Qu.:     95.8   1st Qu.:     0.50   1st Qu.:     6.9   1st Qu.:     0.70   1st Qu.:     2.90
 Median :    407.7   Median :     8.80   Median :    35.4   Median :    12.40   Median :    18.85
 Mean   :   4262.9   Mean   :   277.36   Mean   :   578.1   Mean   :   383.81   Mean   :   392.07
 3rd Qu.:   1359.8   3rd Qu.:    52.27   3rd Qu.:   150.2   3rd Qu.:    71.97   3rd Qu.:    93.20
 Max.   :2366035.3   Max.   :119439.10   Max.   :208576.5   Max.   :145292.60   Max.   :176911.80
 NA's   :139         NA's   :131         NA's   :131        NA's   :131         NA's   :131
 PBDITA as % of total income PBT as % of total income PAT as % of total income Cash profit as % of total income
 Min.   :-6400.000           Min.   :-21340.00       Min.   :-21340.00        Min.   :-15020.000
 1st Qu.:    5.000           1st Qu.:     0.55       1st Qu.:     0.35        1st Qu.:     2.020
 Median :    9.660           Median :     3.31       Median :     2.34        Median :     5.640
 Mean   :    4.571           Mean   :   -17.28       Mean   :   -19.20        Mean   :    -8.229
 3rd Qu.:   16.390           3rd Qu.:     8.80       3rd Qu.:     6.34        3rd Qu.:    10.700
 Max.   :  100.000           Max.   :   100.00       Max.   :   150.00        Max.   :   100.000
 NA's   :68                  NA's   :68              NA's   :68               NA's   :68
 PAT as % of net worth      Sales           Income from financial services  Other income         Total capital
 Min.   :-748.72      Min.   :      0.1   Min.   :    0.00                Min.   :    0.00   Min.   :    0.1
 1st Qu.:   0.00      1st Qu.:    112.7   1st Qu.:    0.40                1st Qu.:    0.40   1st Qu.:   13.1
 Median :   7.92      Median :    453.1   Median :    1.80                Median :    1.40   Median :   42.1
 Mean   :  10.27      Mean   :   4549.5   Mean   :   80.84                Mean   :   41.36   Mean   :  216.6
 3rd Qu.:  20.19      3rd Qu.:   1433.5   3rd Qu.:    9.68                3rd Qu.:    5.97   3rd Qu.:  100.3
 Max.   :2466.67      Max.   :2384984.4   Max.   :51938.20                Max.   :42856.70   Max.   :78273.2
                      NA's   :259         NA's   :935                     NA's   :1295       NA's   :4
 Reserves and funds Deposits (accepted by commercial banks)  Borrowings         Current liabilities & provisions
 Min.   : -6525.9   Mode:logical                             Min.   :     0.10   Min.   :     0.1
 1st Qu.:     5.0   NA's:3541                                1st Qu.:    23.95   1st Qu.:    17.8
 Median :    54.8                                            Median :    99.20   Median :    69.4
 Mean   :  1163.8                                            Mean   :  1122.28   Mean   :   940.6
 3rd Qu.:   277.3                                            3rd Qu.:   352.60   3rd Qu.:   261.7
 Max.   :625137.8                                            Max.   :278257.30   Max.   :352240.3
 NA's   :85                                                  NA's   :366         NA's   :96
```

Some of the variables are taken for understanding ,we can clearly see there are NA’s in the dataset and also extreme values present in the data ,so we have to remove outliers and NA’s from the data

```
### remove NA'S

d = as.data.frame(dataset)
d[is.na(d)]=0
d
any(is.na(d))   #[1] FALSE
sum(is.na(d))  #[1] 0
```

To remove outliers a function quantile is used with an intervalof(0.05,0.95)

```
`Networth Next Year`= quantile(d$`Networth Next Year`,probs = c(0.005,0.95))
`Total assets`= quantile(d$`Total assets`,probs = c(0.005,0.95))
`Net worth`= quantile(d$`Net worth`,probs = c(0.001,0,95))
```

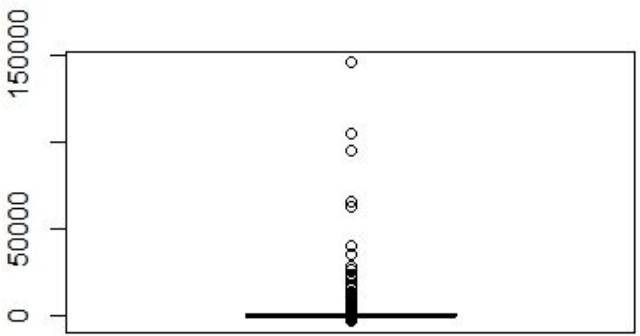
```
`Total income` = quantile(d$`Total income`,probs = c(0.05,0.95))
`Change in stock`= quantile(d$`Change in stock`,probs = c(0.05,0.95))
`Total expenses`= quantile(d$`Total expenses`,probs = c(0.05,0.95))
`Profit after tax`= quantile(d$`Profit after tax`,probs = c(0.05,0.95))
PBDITA = quantile(d$PBDITA ,probs = c(0.05,0.95))                    `Profit
after tax` = quantile(d$`Total expenses`,probs = c(0.05,0.95))
```

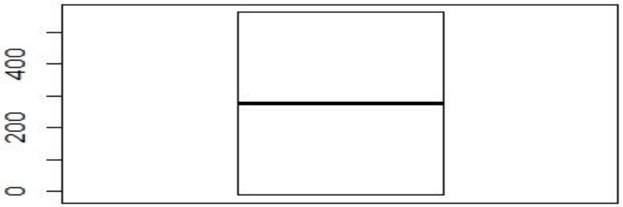
```
> summary(`Networth Next Year`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -222.2   774.5  1771.1  1771.1  2767.8  3764.4
> summary(`Total income`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    2218    4435    4435    6653    8870
> summary(`Change in stock`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -12.90   26.05   65.00   65.00  103.95  142.90
> summary(`Total expenses`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    2147    4293    4293    6440    8587
> summary(`Profit after tax`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -13.3   130.6   274.6   274.6   418.5   562.4
```

Similarly the summaries are noted and the extremities or outliers are quiet balanced for all the variables present in the dataset .

boxplot(`Networth Next Year`)
boxplot(`Change in stock`)
boxplot(`Profit after tax`)
boxplot(`Total assets`)

Different boxplot are taken to check if there were outliers before and after.





Also there are two variables with similar data in them so as a process of data cleaning we can remove one of the variable i.e either ‘Total assets’ or ‘Total liabilities’ and also there is an entire column ‘Deposits ‘which has no data ’.so we can directly remove this in the excel sheet or in R by selecting the column number to remove and create a new dataset but that may not be good idea as that might be data loss or can be affected while validating the data so make sure that you remove the column in the validation_dataset

```
dataset = subset( dataset, select = -c(3))
```

2. Selecting variables

Here Net worth next year variable is taken as a dependent variable and the remaining variables are taken as independent variable which show how big the company is or leverage of the company or liquidity of the company and the net profits measurements and some of them are ratios variables

So from the dependent variable we can conclude that a negative value indicates that the company will default and vice versa .

Usually the ratios are selected for better understanding of companies with different sizes and working .

Profitability - Shareholders Equity ,Gross profit margin most used variables
Liquidity - Current Assets/Current Liabilities ,Quick Ratio
Leverage - Debt to Equity Ratio,Total Assets/Total Equity
Size - common size ratio (equity/asset),asset size, cash flows .

3. Checking for Multicollinearity

Let’s perform multicollinearity for different variables of four different categories so the different variables taken in each of the category since most of these are ratios

So we will use a linear model to understand which variables have the highest importance with dependent variable i.e net worth next year

```
## multicollinearity

mydata = data.frame(d[,-1])      # removing the num variable
mydata
str(mydata)
summary(lm(`Networth Next Year`~. ,mydata))
```

Some of the important variables taken from each of the category to further do the logistic regression .here the variables needed for logistic regression are already there with us since most of the ratios will be taken for each of the different categories as it gives a better explanation regarding the variance in companies w.r.t to size or cash flows and since linear model(lm) gives importance towards normal variables than ratios we consider ratios and check multicollinearity for them

Size : Total Assets, networth ,Total Income ,change in stocks , Equity/Asset

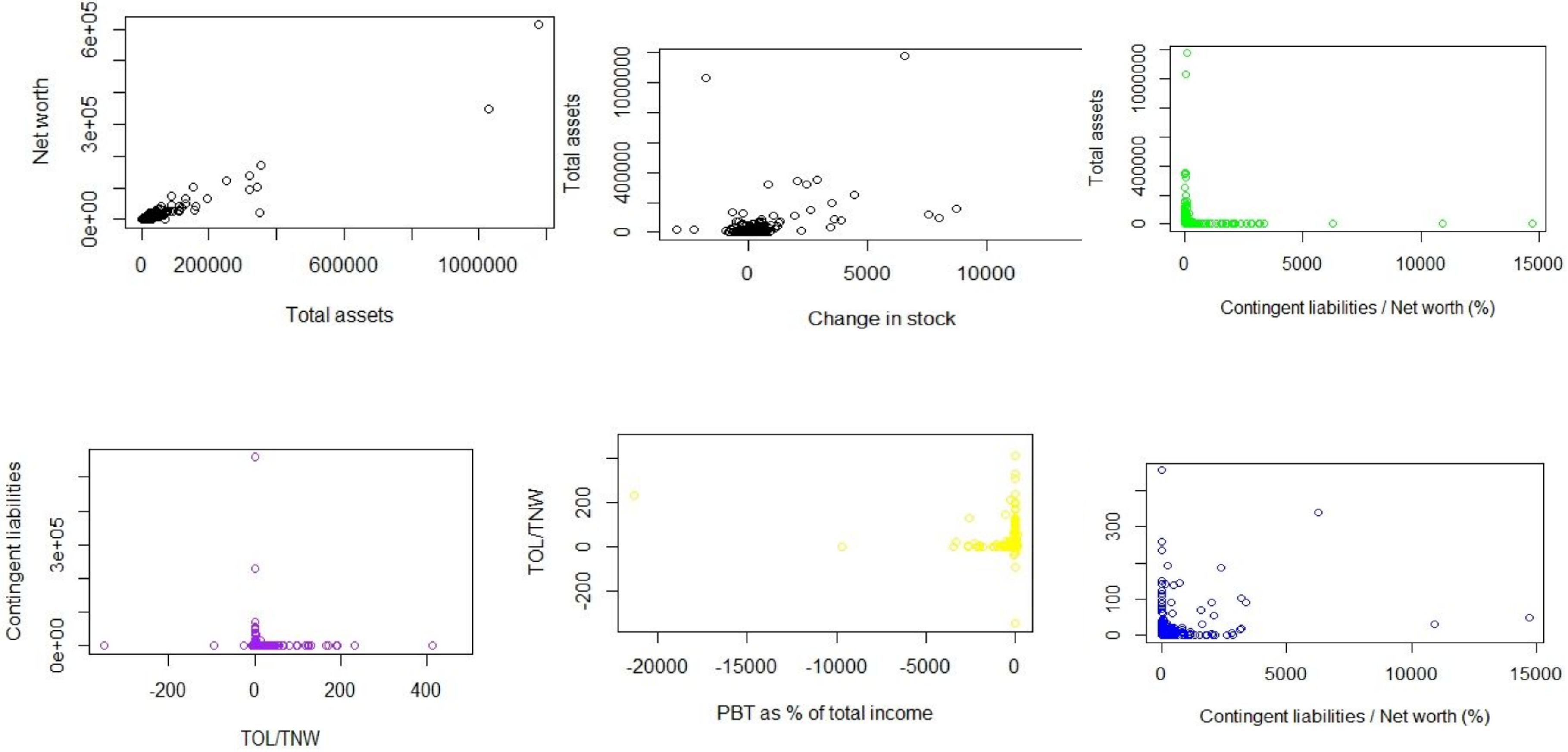
Leverage : TOL/TNW, Contingent Liabilities/Net worth, Current Ratio

Liquidity : Debt to equity ratio ,Quick ratio ,Cash to current liabilities ,Cash to Average cost of sales per day

Profitability : PBDITA as % of total income,PBT as of % of total Income, cash profit as % of total Income ,PAT as of % of total Income,PAT as % of net worth.

4. Univariate and Bivariate analysis

```
cor(`Total assets`,`Net worth`) # 0.95
plot(`Total assets`,`Net worth`)
cor(`Change in stock`,`Total assets`) # 0.231
plot(`Change in stock`,`Total assets`)
plot(`TOL/TNW`,`Total assets`)
cor(`Contingent liabilities / Net worth (%)`,`Total assets`) # 0.0018
plot(`Contingent liabilities / Net worth (%)`,`Total assets`)
cor(`PBT as % of total income`,`TOL/TNW`) # 0.375
plot(`PBT as % of total income`,`TOL/TNW`)
cor(`Contingent liabilities / Net worth (%)`,`Debt to equity ratio (times)`) # 0.254
plot(`Contingent liabilities / Net worth (%)`,`Debt to equity ratio (times)`)
```



Net worth
0e+00
3e+05
6e+05
0
200000
600000
1000000
Total assets
Total assets
0
400000
1000000
0
5000
10000
Change in stock
Total assets
0
400000
1000000
0
5000
10000
15000
Contingent liabilities / Net worth (%)
Contingent liabilities
0e+00
3e+05
-200
0
200
400
TOL/TNW
TOL/TNW
-200
0
200
-20000
-15000
-10000
-5000
0
PBT as % of total income
0
100
300
0
5000
10000
15000
Contingent liabilities / Net worth (%)

summary(`Total assets`)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1	91.3	309.7	3443.4	1098.7	8543.7

summary('Total Capital')

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	13.0	42.1	216.4	100.3	8732.6

summary(`TOL/TNW`)

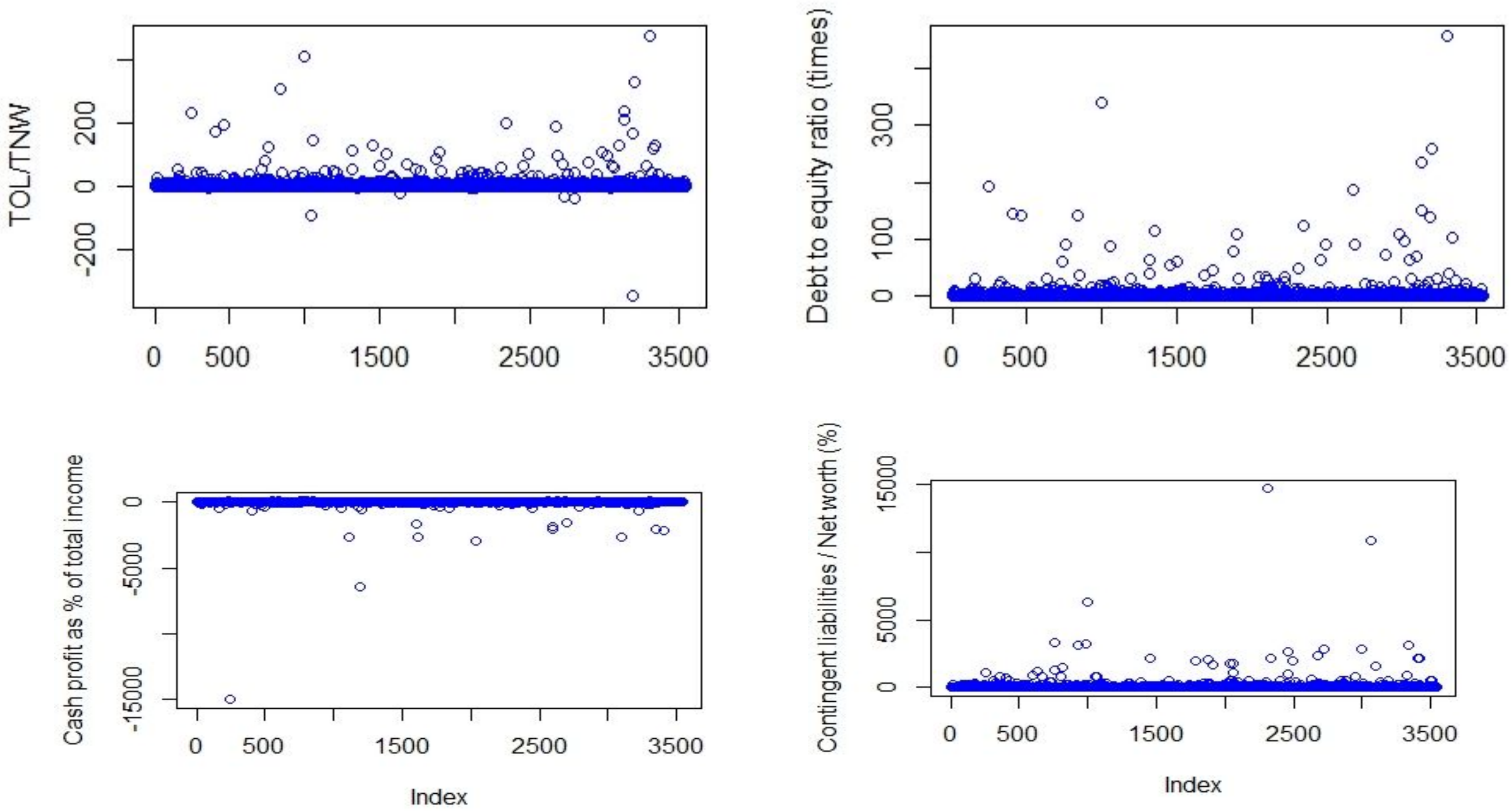
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-350.480	0.600	1.430	3.994	2.830	473.000

summary(`Contingent liabilities / Net worth (%)`)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	5.33	53.94	30.76	147

```
summary(d$`Cash profit`)
Min.  1st Qu.   Median    Mean  3rd Qu.    Max.
-245.7      1.9     16.7    377.6     86.8     987.9

summary(`PBT as % of total income`)
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-3.10      0.55      3.31    -17.28      8.80      100.00

summary(`Debt to equity ratio (times)`)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    0.22    0.79      2.78     1.75     456.00

summary(d$`Cash to current liabilities (times)`)
Min.  1st Qu.      Median      Mean     3rd Qu.   Max.
0.0000   0.0200   0.0700    0.4775    0.1900   165.0000
```

```
> summary(`Networth Next Year`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -222.2   774.5  1771.1  1771.1  2767.8  3764.4 
> summary(`Total income`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    2218    4435    4435    6653    8870 
> summary(`Change in stock`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -12.90   26.05   65.00   65.00  103.95  142.90 
> summary(`Total expenses`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    2147    4293    4293    6440    8587 
> summary(`Profit after tax`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  -13.3   130.6   274.6   274.6   418.5   562.4 
```



TOL/TNW
200
0
-200
0
500
1500
2500
3500
Debt to equity ratio (times)
300
100
0
0
500
1500
2500
3500
Cash profit as % of total income
0
-5000
-15000
0
500
1500
2500
3500
Index
Contingent liabilities / Net worth (%)
15000
5000
0
0
500
1500
2500
3500
Index

5. Analysing variables & signs

First of all let us assume a variable ‘default’ from the dependent variable Net worth next year

Default = ifelse(`Networth Next Year`>0,0,1)
summary(as.factor(Default)) #0 1
3298 243

Profitability
summary(`PBT as % of total income`)
summary(`PBT as % of total income`[Default==0]) # a typical good company makes a profit of 3.54 out of 100 units
summary(`PBT as % of total income`[Default==1]) # a typical bad company makes a loss of 5.09 out of 100 units

using logistic regression

```
model1 = glm(as.factor(Default)~`PBT as % of total income`,family = binomial)
model1
summary(glm(as.factor(Default)~`PBT as % of total income`,family = binomial))
```

`PBT as % of total income` -0.0011237 0.0002471 -4.549 5.4e-06 ***

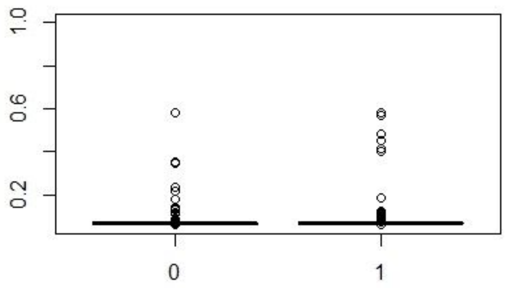
```
summary(model1$fitted.values)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05987	0.06592	0.06629	0.06862	0.06649	1.00000

These are percentages for min ,median and max value

of a company to default

```
plot(as.factor(model1$y),model1$fitted.values)
```



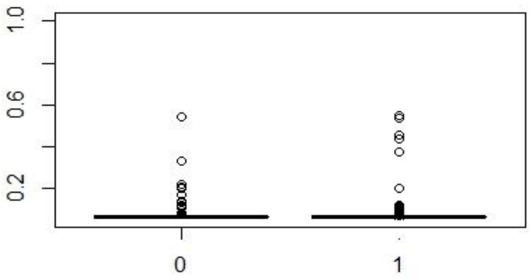
Profitability

summary(`PAT as % of total income`)

summary(`PAT as % of total income`[Default==0]) # a typical good company makes a profit of 2.570 out of 100 units summary(`PAT as % of total income`[Default==1]) # a typical bad company makes a loss of 4.57 out of 100 units

using logistic regression

```
Call:  glm(formula = as.factor(Default) ~ `PAT as % of total income`,
    family = binomial)

Coefficients:
              (Intercept)  `PAT as % of total income`
                -2.643137                   -0.001074

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.6431366  0.0676064 -39.096  < 2e-16 ***
`PAT as % of total income` -0.0010744  0.0002407  -4.463 8.07e-06 ***
---

> summary(model1$fitted.values)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05709 0.06600 0.06626 0.06862 0.06640 1.00000
> plot(as.factor(model1$y),model1$fitted.values) #original default values vs predicted default values
```



From analysing variables we can understand that -ve median in the summary of the variable means there is a loss and +ve median means there is a profit or increase in profit since it a Profitability variable the standardised value is made between default and non-default companies

Similarly there will be a -ve flow of cash(going out of company) or +ve flow of cash(coming to company) of the median in terms of liquidity between standardised default and non-default companies .e.g debt to equity ratio

```
> summary(`Debt to equity ratio (times)`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.22    0.79    2.78    1.75  456.00 
> summary(`Debt to equity ratio (times)`[Default==0]) #good company has less debt so less ratio i.e 0.740
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.200   0.740   1.601   1.590 341.180 
> summary(`Debt to equity ratio (times)`[Default==1]) #bad company has more debt so more ratio i.e 4.56
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    1.10    4.56   18.78   12.85  456.00 
```

Similarly there will be a difference in the median of the size of the assets ,income ,net worth or total expenses

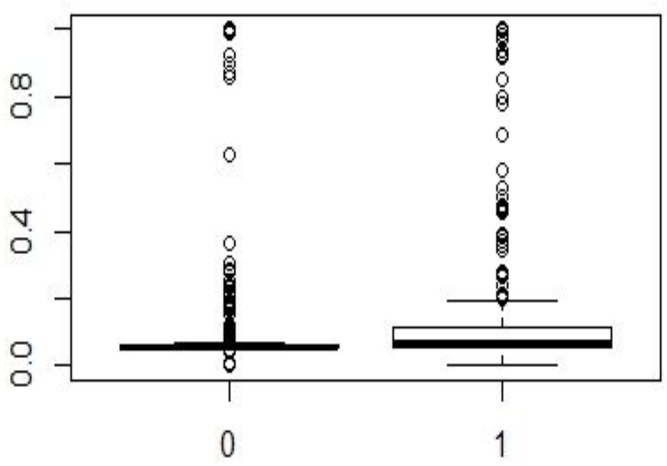
summary(`Total assets`[Default==0]) # a good company has 332.6 units of assets

summary(`Total assets`[Default==1]) # a bad company has 102.6 units of assets

But this cannot entirely say whether a company is good or bad so we take ratios

For leverage we consider a variable ratio TOL/TNW which is the ratio of total liabilities to total net worth . So less the liabilities better is the company

```
> ## Leverage
> summary(`TOL/TNW`)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-350.480    0.600    1.430    3.994    2.830  473.000 
> summary(`TOL/TNW`[Default==0])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-94.580   0.570   1.340   2.553   2.540 411.270 
> summary(`TOL/TNW`[Default==1])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-350.480    2.125    6.920   23.552   19.530  473.000 
> model1 = glm(as.factor(Default)~`TOL/TNW`,family = binomial)
> model1

Call:  glm(formula = as.factor(Default) ~ `TOL/TNW`, family = binomial)

Coefficients:
(Intercept)    `TOL/TNW`  
   -2.86894      0.04256  

Degrees of Freedom: 3540 Total (i.e. Null);  3539 Residual
Null Deviance:      1771 
Residual Deviance: 1620        AIC: 1624
> summary(model1$fitted.values)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.05502 0.05689 0.06862 0.06017 1.00000 
> plot(as.factor(model1$y),model1$fitted.values) #original default values vs predicted default values
```



For a better company or a standardised good company the ratio is around 1.4 but for the standardised bad company the ratio is around 6.920 which is high liability rate on the company .

6. Applying Logistic regression analysis

```
Default = ifelse(`Networth Next Year`>0,0,1)
summary(as.factor(Default))                      # 0      1
.                                                 3298    243

model2 = glm(as.factor(Default)~`PBT as % of total income`+`Debt to equity ratio (times)`+`Quick ratio (times)`+`Contingent liabilities / Net worth (%)`,family = binomial)
model2
```

```
Coefficients:
                          (Intercept)              `PBT as % of total income`
                           -3.0164559                              -0.0021931
         `Debt to equity ratio (times)`                `Quick ratio (times)`
                            0.0643330                              -0.0207570
`Contingent liabilities / Net worth (%)`
                            0.0004998

Degrees of Freedom: 3391 Total (i.e. Null);  3387 Residual
  (149 observations deleted due to missingness)
Null Deviance:      1586
Residual Deviance: 1391        AIC: 1401
```

The values form a product with the corresponding variable to form a linear equation with the dependent variable .

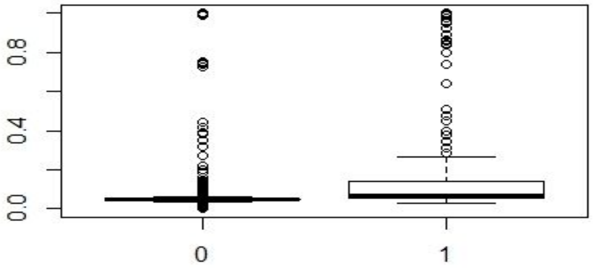
Here a positive coefficient means that for the independent variable if the value is increasing then the dependent variable is directly proportional relation to the variable with the positive coefficient and inversely proportional relation to that of variable with negative coefficient and the p-value show the significance of the variable and there will be a linear equation formed with these variables w.r.t the dependent variable summary(glm(as.factor(Default)~`PBT as % of total income`+`Debt to equity ratio (times)`+`Quick ratio (times)`+ `Contingent liabilities / Net worth (%)` ,family = binomial))

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)                             -2.9011364  0.0781275 -37.133  < 2e-16 ***
`PBT as % of total income`              -0.0010221  0.0002581  -3.959 7.52e-05 ***
`Debt to equity ratio (times)`           0.0621338  0.0086398   7.192 6.41e-13 ***
`Quick ratio (times)`                   -0.0225218  0.0151349  -1.488   0.1367    
`Contingent liabilities / Net worth (%)`  0.0005784  0.0002455   2.356   0.0185 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1771.0  on 3540  degrees of freedom
Residual deviance: 1572.2  on 3536  degrees of freedom
AIC: 1582.2

Number of Fisher Scoring iterations: 6
```

Clearly the asterisks are marked less the p-value better is the variable to consider and the z-values show the standard deviation from the mean variable and std error shows the error rate from the estimate value

The values form a product with the corresponding variable to form a linear equation with the dependent variable

plot(as.factor(model2$y),model2$fitted.values)



model2$fitted.values

1	2	3	5	6	8	9	10	11	12	13
0.04501147	0.04727134	0.04783187	0.05146664	0.05000528	0.05305527	0.04860490	0.05812294	0.04517121	0.04946342	0.04399592
14	15	16	17	18	19	20	21	22	23	24
0.08798070	0.04774794	0.05173840	0.05011203	0.04830510	0.08409875	0.04690387	0.04996089	0.05834946	0.05004554	0.04675816
25	26	27	28	29	30	31	32	33	34	35
0.04548626	0.04644499	0.04591600	0.04511457	0.04663294	0.05212011	0.04628319	0.04717657	0.04611073	0.05067986	0.04440304
36	37	38	39	40	41	42	43	44	45	46
0.04853564	0.03696091	0.04596696	0.04754090	0.04588612	0.05168242	0.04552111	0.04789189	0.06728980	0.04694825	0.04357588
47	48	49	50	51	52	53	54	55	56	57
0.04585427	0.04630154	0.05104869	0.04768944	0.05669168	0.04799013	0.05499753	0.04510922	0.02987300	0.04631177	0.05168885
58	59	60	61	62	63	64	65	66	67	68
0.04602147	0.04749382	0.06386089	0.04488760	0.04591726	0.04777518	0.04608583	0.04983439	0.04666484	0.04871869	0.04500252
69	70	71	72	73	74	75	76	77	78	79
0.06705694	0.05522873	0.04719169	0.04863307	0.04813929	0.06215610	0.04604684	0.04916850	0.04635824	0.04670816	0.04700751
80	81	82	83	84	85	86	87	88	89	90
0.04714067	0.04527513	0.05067608	0.04768699	0.05021358	0.04772674	0.04857285	0.05686365	0.04596436	0.05458839	0.04834810
91	92	93	94	95	96	97	98	99	100	101
0.06288992	0.04588294	0.05163560	0.05272784	0.04486910	0.05753172	0.05393681	0.04579043	0.05121834	0.05063096	0.03997524
102	103	104	105	107	108	109	110	111	112	113
0.05021296	0.04784001	0.05727608	0.04782546	0.05347163	0.08133772	0.04837708	0.04506156	0.04604084	0.04781188	0.04628945
114	115	116	117	118	119	121	122	123	124	125
0.04522895	0.04787372	0.05148213	0.04110900	0.05041290	0.04854097	0.05606552	0.05000269	0.04835481	0.04583354	0.04732837
126	127	128	129	130	131	132	133	134	135	136
0.05053522	0.04781379	0.04636101	0.04604531	0.04993572	0.04191395	0.04518747	0.05119558	0.04784264	0.10962455	0.04838304
137	138	139	140	141	142	144	145	146	147	148
0.04693816	0.05371087	0.04644412	0.08004022	0.05098252	0.04297588	0.05655974	0.05721340	0.10602218	0.04670207	0.04623545
150	151	152	153	154	155	156	157	158	159	160
0.04832008	0.04994404	0.04790520	0.04663236	0.04708991	0.28385835	0.04664685	0.05081631	0.05841423	0.06146124	0.04860043
161	162	163	164	165	166	167	168	169	170	171
0.04698844	0.05038631	0.14409213	0.10816588	0.05097276	0.04677470	0.04495013	0.05096825	0.04900611	0.06307578	0.04474654
172	173	174	175	176	177	178	179	180	181	182
0.03168700	0.04545606	0.04840585	0.04903562	0.04657935	0.04867524	0.04626089	0.04917554	0.05031447	0.05488969	0.07825507

```
summary(model2$fitted.values)
Min.       1st Qu.    Median     Mean       3rd Qu.    Max.
0.0000413 0.0463585 0.0487013 0.0625000 0.0524144 1.00000
```

This show a minimum value of a company defaulting is 0 and the maximum probability that the company is going to default is 1 and in the first quartile there is a 4.6% chance of defaulting and the median value is 4.8% over 75% comapanies there is a 5.2% chance it is going to default according to the summary of our logistic model as per the fitted models. These predicted values will differ from the good and bad companies .

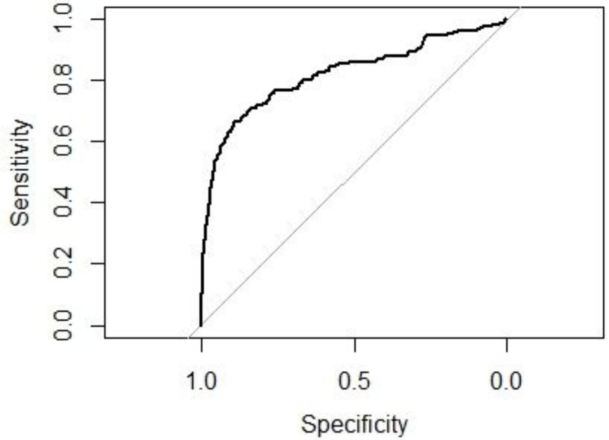
7. Predicticting Accuracy

```
Default2 = ifelse(model2$fitted.values>0.0585,1,0)
```

We can set the fitted value over a wide range w.r.t predicting more companies without a default or predict less companies which have actually defaulted or a balanced over predicted and actual defaulters and non-defaulters.We have considered a value so here we have set a fitted value of around 0.0585

table(model2$y,Default2)

```
> table(model2$y,Default2)
   Default2
       0    1
  0 2923  257
  1   75  137
> 
```



Sensitivity : 2923/(2923+257) = 92% accurate

Specificity : 137/(137+75) = 65 % accurate

Overall accuracy : (2923+137)/(2923+137+75+257) = 90.2 % accurate

library(pROC)
plot.roc(model2$y,model2$fitted.values)

We create a test and training data from the raw_data and compare it to the validation dataset for accuracy and make sure that the argument length doesn't differ

We divide the test and train data

```
summary(as.factor(data$`Default - 1`))
0   1
661  54
ind = sample(2,nrow(d),replace = T,prob = c(0.8,0.2))
train1 = d[ind==1,]
test1 = d[ind==2,]
test.predicted = predict(model3,newdata = 'Default-1',type = 'response')

table(test.predicted,default3)
```

	Default3	
	0	1
0	509	137
1	11	36

Sensitivity : 509/646 = 78.8 % Specificity : 36/47 = 76.7 % Overall accuracy : 79.35 %

8.Sorting the data in deciles

sort(Default3,decreasing = FALSE)

order(Default3)

quantile(Default3,decreasing = FALSE,prob = seq(0,1,length = 11))

```
> quantile(Default3,decreasing = FALSE,prob = seq(0,1,length = 11))
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100% 
   0    0    0    0    0    0    0    0    1    1    1 
> 
```

The model has a accuracy rate of around 80 %