Credit Risk Prediction

-Project

Vompolu Sai Tanuj

G1 – PGPBABI(June 2019)

Table of Contents

1) Introduction:

The project given to us is of **Credit Risk Estimation to be done by the bank** on various **companies** before giving them loans.

The word **Credit** in the context of the project refers to the **loan** that the bank provides to the company when required. The bank, after providing the loan, expects the customer (**i.e.** company) to pay back the **loan amount** along with the **interest levied on it.** Any company which is **not able to pay back** the **loan** to the bank **adhering to the contractual terms** is considered a **loss for the bank.**

So the term **credit risk** refers to the **possibility of loss** that occurs when the bank gives a **loan** to a company which **fails to pay back the loan to the bank based adhering to the contractual terms.**

Credit Risk can be averted by **analysing the potential of the customer** to **pay back the loan.** This risk is assessed based on the **5 Cs which are as follows:**

- **Credit History**
- **Capacity to repay**
- **Capital**
- **Conditions of the loan**
- **Collateral**

While the last two factors are under the control of the bank, the **first three factors** are all dependent on the **financial stability of the company,** these must be assessed and decided by the bank based on the **data got from the financial portfolios and statements of the company.**

After assessing the **financial stability of the company,** then it is important for the company to decide on three things,

I. ***Whether the company can be provided with a loan or not.***
II. ***If the company can be provided with a loan, to what amount the credit can be given?***
III. ***What must be the collaterals and conditions for the said company given it is provided with a loan?***

a. Project Introduction:

We are requested to create an **India credit risk (default) model**, using the data provided in the spreadsheet **training.xlsx,** and validate it on **test.xlsx.** We are to use the **logistic regression framework** to develop the **credit default model while trying to reduce the loss incurred by the bank**

The **Raw and Validation** datasets contain data based on the following variables

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

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

2) Exploratory Data Analysis:

The dataset must be imported from the **excel file** named **Raw.xlsx.** This dataset must be imported into the **R session using the function readxl()**. The dataset is imported under the name "**training".** The dataset can be viewed using the **View()** function.

```
> ### EDA #####
> training = read_xlsx("training.xlsx")
> View(training)
```

a. Initial Exploration:

The initial exploration of the dataset can be done using the following functions:

- **class() –** This function helps in telling us the **format of the dataset.**
- **str() –** This function helps in giving the **basic structure of the dataset.**
- **head() –** This function helps in displaying the **top rows of the dataset.**

- **tail() –** This function helps in displaying the **bottom rows of the dataset.**
- **colnames() –** This function helps in displaying the **column names of the dataset.**
- **summary() –** This function helps in giving a **summary of all the variables present in the dataset.**
- **dim() –** This function gives us the dimensions of the dataset.

```
> ### EDA #####
> training = read_xlsx("training.xlsx")
> ### Initial Exploration ####
> class(training)
[1] "tbl_df"     "tbl"        "data.frame"
> training = as.data.frame(training)
```

```
> head(training)
  Num Networth Next Year Total assets Net worth Total income Change in stock Total expenses Profit after tax PBDITA    PBT
1   1             8890.6      17512.3    7093.2      24965.2           235.8        23657.8           1543.2 2860.2 2417.2
2   2              394.3        941.0     351.5       1527.4            42.7         1454.9            115.2  283.0  188.4
3   3               92.2        232.8     100.6        477.3            -5.2          478.7             -6.6    5.8   -6.6
4   4                2.7          2.7       2.7           NA              NA             NA               NA     NA     NA
5   5              109.0        478.5     107.6       1580.5           -17.0         1558.0              5.5   31.0    6.3
6   6              688.6       2434.4     675.8       2648.6            62.3         2636.4             74.5  200.1   74.5
  Cash profit PBDITA as % of total income PBT as % of total income PAT as % of total income Cash profit as % of total income
1      1872.8                      11.46                     9.68                      6.18                              7.50
2       158.6                      18.53                    12.33                      7.54                             10.38
3         0.3                       1.22                    -1.38                     -1.38                              0.06
4          NA                       0.00                     0.00                      0.00                              0.00
5        11.9                       1.96                     0.40                      0.35                              0.75
6       146.9                       7.55                     2.81                      2.81                              5.55
  PAT as % of net worth   Sales Income from financial services Other income Total capital Reserves and funds
1                 23.78 24458.0                            158.0        297.2         423.8             6822.8
2                 38.08  1504.3                              4.0         15.9         115.5              257.8
3                 -6.35   475.6                              1.5          0.2          81.4               19.2
4                  0.00      NA                               NA           NA           0.5                2.2
5                  5.25  1575.1                              3.9          0.9           6.2              161.8
6                 21.78  2639.5                              6.4          0.2          33.8              972.0
  Deposits (accepted by commercial banks) Borrowings Current liabilities & provisions Deferred tax liability Shareholders funds
1                                      NA       14.9                           9965.9                 284.9             7093.2
2                                      NA      272.5                            210.0                  85.2              351.5
3                                      NA       35.4                             96.8                    NA              100.6
4                                      NA         NA                               NA                    NA                2.7
5                                      NA      193.1                            112.8                   4.6              107.6
6                                      NA      717.1                            555.9                  54.4              698.2
  Cumulative retained profits Capital employed TOL/TNW Total term liabilities / tangible net worth
1                      6263.3            7108.1    1.33                                       0.00
2                       247.4             624.0    1.23                                       0.34
3                        32.4             136.0    1.44                                       0.29
4                         2.2               2.7    0.00                                       0.00
5                        82.7             300.7    2.83                                       1.59
6                       317.7            1415.3    1.80                                       0.37
```

```
  Contingent liabilities / Net worth (%) Contingent liabilities Net fixed assets Investments Current assets Net working capital
1                                  14.80                1049.7           1900.2      1069.6        13277.5               3588.5
2                                  19.23                  67.6            286.4         2.2          563.9                203.5
3                                  45.83                  46.1             38.7         4.3          167.5                 59.6
4                                   0.00                    NA              2.5          NA            0.2                  0.2
5                                  34.94                  37.6             94.8         7.4          349.7                215.8
6                                  36.28                 245.2            864.9        22.7         1296.2                278.5
  Quick ratio (times) Current ratio (times) Debt to equity ratio (times) Cash to current liabilities (times)
1                1.18                   1.37                       0.00                               0.43
2                0.95                   1.56                       0.78                               0.06
3                1.11                   1.55                       0.35                               0.21
4                  NA                     NA                       0.00                                 NA
5                1.41                   2.54                       1.79                               0.00
6                0.48                   1.27                       1.09                               0.11
  Cash to average cost of sales per day Creditors turnover   Debtors turnover Finished goods turnover       WIP turnover
1                                 68.21               3.62               3.85                  200.55              21.78
2                                  5.96 9.8000000000000007                5.7                   14.21               7.49
3                                 17.07               5.28               5.07                    9.24               0.23
4                                    NA                  0                  0                   <NA>               <NA>
5                                  0.00                 13 9.4600000000000009                   12.68                7.9
6                                 15.78                6.5              21.13                   10.14 8.3800000000000008
  Raw material turnover Shares outstanding Equity face value   EPS Adjusted EPS Total liabilities PE on BSE
1                  7.71           42381675                10 35.52         7.10            17512.3     27.31
2                 11.46           11550000                10  9.97         9.97              941.0      8.17
3                  <NA>            8149090                10 -0.50        -0.50              232.8     -5.76
4                     0              52404                10  0.00         0.00                2.7        NA
5                 17.03             619635                10  7.91         7.91              478.5        NA
6                  4.74            1141718                10 30.57        15.28             2434.4        NA
> tail(training)
      Num Networth Next Year Total assets Net worth Total income Change in stock Total expenses Profit after tax PBDITA   PBT
3536 3540               1.2          17.8       1.2         15.5            -1.2           14.2             0.1    1.8   0.2
3537 3541             226.4         450.5     172.3        565.0            30.5          581.1            14.4   76.7  41.1
3538 3542              89.4          97.6      82.0         75.8            -4.0           66.5             5.3   11.1   6.2
3539 3543             246.2         902.9     209.1       1005.1             5.6          966.5            44.2  120.3  70.0
3540 3544             146.9         177.0     137.2        371.0             3.9          348.9            26.0   50.5  40.8
3541 3545              -0.2           0.6       0.3           NA              NA           17.4           -17.4  -17.4 -17.4

      Cash profit PBDITA as % of total income PBT as % of total income PAT as % of total income Cash profit as % of total income
3536          0.5                       11.61                    1.29                     0.65                            3.23
3537         48.4                       13.58                    7.27                     2.55                            8.57
3538          9.2                       14.64                    8.18                     6.99                           12.14
3539         62.6                       11.97                    6.96                     4.40                            6.23
3540         33.6                       13.61                   11.00                     7.01                            9.06
3541        -17.4                          NA                      NA                       NA                              NA
      PAT as % of net worth Sales Income from financial services Other income Total capital Reserves and funds
3536                  8.70  14.3                              NA          1.2          1.0               0.2
3537                  8.71 564.5                             0.5           NA         89.0              85.5
3538                  6.68  73.9                             1.7           NA         38.6              48.4
3539                 22.77 995.9                             2.6          0.3         30.0             179.1
3540                 20.30 365.8                             3.3          1.6         50.9              86.3
3541               -193.33    NA                              NA           NA         28.3             -28.0
      Deposits (accepted by commercial banks) Borrowings Current liabilities & provisions Deferred tax liability
3536                                      NA       14.5                              2.1                     NA
3537                                      NA      190.2                             42.5                   36.8
3538                                      NA        3.0                              7.6                     NA
3539                                      NA      305.0                            363.4                   25.4
3540                                      NA        1.3                             21.1                   17.4
3541                                      NA         NA                              0.3                     NA
      Shareholders funds Cumulative retained profits Capital employed TOL/TNW Total term liabilities / tangible net worth
3536                1.2                         0.2              15.7   13.83                                      4.83
3537              172.3                        76.8             362.5    1.30                                      0.72
3538               87.0                        36.6              90.0    0.12                                      0.02
3539              209.1                       179.1             514.1    2.45                                      0.68
3540              137.2                        77.1             138.5    0.10                                      0.01
3541                0.3                       -28.0               0.3    1.00                                      0.00
```

```
     Contingent liabilities / Net worth (%) Contingent liabilities Net fixed assets Investments Current assets
3536                                   0.00                    NA             5.7         0.1           6.4
3537                                   0.00                    NA           227.0          NA         187.0
3538                                   5.12                   4.2            21.9         6.8          55.8
3539                                  93.45                 195.4           217.7        17.5         477.5
3540                                   6.20                   8.5            73.5          NA          80.8
3541                                   0.00                    NA              NA          NA           0.6
     Net working capital Quick ratio (times) Current ratio (times) Debt to equity ratio (times)
3536                -4.4                 0.46                  0.59                      12.08
3537                78.3                 0.41                  1.71                       1.10
3538                47.2                 4.58                  6.49                       0.10
3539               -49.5                 0.59                  0.91                       1.46
3540                59.7                 2.83                  3.83                       0.01
3541                 0.3                 2.00                  2.00                       0.00
     Cash to current liabilities (times) Cash to average cost of sales per day Creditors turnover Debtors turnover
3536                                0.07                                20.71                5.81             3.67
3537                                0.07                                 5.67               15.65            20.64
3538                                3.88                               177.71               10.07            14.21
3539                                0.05                                11.05                3.96             3.76
3540                                1.35                                29.93                  25            13.75
3541                                2.00                              2190.00                   0               0
     Finished goods turnover WIP turnover Raw material turnover Shares outstanding Equity face value   EPS Adjusted EPS
3536                    8.33          7.52                10.92                 NA                NA  0.00         0.00
3537                    8.66          5.14                19.47           14904213                10  0.97         0.97
3538                    5.13          4.17                 4.83            3362800                10  1.61         1.61
3539                   33.03         11.68                 4.63            3000000                10 13.10        13.10
3540                      49         47.03   17.420000000000002            4422346                10  6.06         6.06
3541                    <NA>          <NA>                    0            5220000                10 -0.02        -0.02
     Total liabilities          PE on BSE
3536              17.8                 NA
3537             450.5                 NA
3538              97.6 2.4900000000000002
3539             902.9              12.62
3540             177.0               4.07
3541               0.6                 NA
```

```
> summary(training)
      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       PBDITA as % of total income
 Min.   :     -0.1   Min.   : -3908.30   Min.   :  -440.7   Min.   : -3894.80   Min.   : -2245.70   Min.   :-6400.000
 1st Qu.:     95.8   1st Qu.:     0.50   1st Qu.:     6.9   1st Qu.:     0.70   1st Qu.:     2.90   1st Qu.:    5.000
 Median :    407.7   Median :     8.80   Median :    35.4   Median :    12.40   Median :    18.85   Median :    9.660
 Mean   :   4262.9   Mean   :   277.36   Mean   :   578.1   Mean   :   383.81   Mean   :   392.07   Mean   :    4.571
 3rd Qu.:   1359.8   3rd Qu.:    52.27   3rd Qu.:   150.2   3rd Qu.:    71.97   3rd Qu.:    93.20   3rd Qu.:   16.390
 Max.   :2366035.3   Max.   :119439.10   Max.   :208576.5   Max.   :145292.60   Max.   :176911.80   Max.   :  100.000
 NA's   :139         NA's   :131         NA's   :131        NA's   :131         NA's   :131         NA's   :68
 PBT as % of total income PAT as % of total income Cash profit as % of total income PAT as % of net worth      Sales
 Min.   :-21340.00        Min.   :-21340.00        Min.   :-15020.000               Min.   :-748.72        Min.   :     0.1
 1st Qu.:      0.55       1st Qu.:      0.35       1st Qu.:      2.020              1st Qu.:   0.00        1st Qu.:   112.7
 Median :      3.31       Median :      2.34       Median :      5.640              Median :   7.92        Median :   453.1
 Mean   :    -17.28       Mean   :    -19.20       Mean   :     -8.229              Mean   :  10.27        Mean   :  4549.5
 3rd Qu.:      8.80       3rd Qu.:      6.34       3rd Qu.:     10.700              3rd Qu.:  20.19        3rd Qu.:  1433.5
 Max.   :    100.00       Max.   :    150.00       Max.   :    100.000              Max.   :2466.67        Max.   :2384984.4
 NA's   :68               NA's   :68               NA's   :68                                             NA's   :259
 Income from financial services  Other income      Total capital      Reserves and funds Deposits (accepted by commercial banks)
 Min.   :     0.00               Min.   :    0.00   Min.   :     0.1   Min.   : -6525.9   Mode:logical
 1st Qu.:     0.40               1st Qu.:    0.40   1st Qu.:    13.1   1st Qu.:     5.0   NA's:3541
 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
```

```
   Borrowings       Current liabilities & provisions Deferred tax liability Shareholders funds Cumulative retained profits
 Min.   :     0.10   Min.   :     0.1                 Min.   :    0.1        Min.   :     0.0   Min.   : -6534.3
 1st Qu.:    23.95   1st Qu.:    17.8                 1st Qu.:    3.2        1st Qu.:    32.0   1st Qu.:     1.1
 Median :    99.20   Median :    69.4                 Median :   13.4        Median :   105.6   Median :    37.1
 Mean   :  1122.28   Mean   :   940.6                 Mean   :  227.2        Mean   :  1322.1   Mean   :   890.5
 3rd Qu.:   352.60   3rd Qu.:   261.7                 3rd Qu.:   50.0        3rd Qu.:   393.2   3rd Qu.:   202.3
 Max.   :278257.30   Max.   :352240.3                 Max.   :72796.6        Max.   :613151.6   Max.   :390133.8
 NA's   :366         NA's   :96                       NA's   :1140                              NA's   :38
 Capital employed      TOL/TNW          Total term liabilities / tangible net worth Contingent liabilities / Net worth (%)
 Min.   :     0.0   Min.   :-350.480   Min.   :-325.600                            Min.   :    0.00
 1st Qu.:    60.8   1st Qu.:   0.600   1st Qu.:   0.050                            1st Qu.:    0.00
 Median :   214.7   Median :   1.430   Median :   0.340                            Median :    5.33
 Mean   :  2328.3   Mean   :   3.994   Mean   :   1.844                            Mean   :   53.94
 3rd Qu.:   767.3   3rd Qu.:   2.830   3rd Qu.:   1.000                            3rd Qu.:   30.76
 Max.   :891408.9   Max.   : 473.000   Max.   : 456.000                            Max.   :14704.27

 Contingent liabilities Net fixed assets    Investments        Current assets     Net working capital Quick ratio (times)
 Min.   :     0.1       Min.   :     0.0   Min.   :     0.00   Min.   :     0.1   Min.   :-63839.0    Min.   : 0.000
 1st Qu.:     6.3       1st Qu.:    26.0   1st Qu.:     1.00   1st Qu.:    36.2   1st Qu.:    -1.1    1st Qu.: 0.410
 Median :    38.0       Median :    93.5   Median :     8.35   Median :   145.1   Median :    16.2    Median : 0.670
 Mean   :   932.9       Mean   :  1189.7   Mean   :   694.73   Mean   :  1293.4   Mean   :   138.6    Mean   : 1.401
 3rd Qu.:   192.7       3rd Qu.:   344.9   3rd Qu.:    64.30   3rd Qu.:   502.2   3rd Qu.:    84.2    3rd Qu.: 1.030
 Max.   :559506.8       Max.   :636604.6   Max.   :199978.60   Max.   :354815.2   Max.   : 85782.8    Max.   :341.000
 NA's   :1188           NA's   :118        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) Cash to average cost of sales per day
 Min.   :  0.00        Min.   :  0.00               Min.   :  0.0000                    Min.   :     0.00
 1st Qu.:  0.93        1st Qu.:  0.22               1st Qu.:  0.0200                    1st Qu.:     2.79
 Median :  1.23        Median :  0.79               Median :  0.0700                    Median :     8.03
 Mean   :  2.13        Mean   :  2.78               Mean   :  0.4904                    Mean   :   158.44
 3rd Qu.:  1.71        3rd Qu.:  1.75               3rd Qu.:  0.1900                    3rd Qu.:    21.79
 Max.   :505.00        Max.   :456.00               Max.   :165.0000                    Max.   :128040.76
 NA's   :93                                         NA's   :93                          NA's   :85
 Creditors turnover Debtors turnover   Finished goods turnover WIP turnover       Raw material turnover Shares outstanding
 Length:3541        Length:3541        Length:3541             Length:3541        Length:3541           Length:3541
 Class :character   Class :character   Class :character        Class :character   Class :character      Class :character
 Mode  :character   Mode  :character   Mode  :character        Mode  :character   Mode  :character      Mode  :character

 Equity face value       EPS              Adjusted EPS       Total liabilities    PE on BSE
 Length:3541        Min.   :-843181.8   Min.   :-843181.8   Min.   :      0.1   Length:3541
 Class :character   1st Qu.:      0.0   1st Qu.:      0.0   1st Qu.:     91.3   Class :character
 Mode  :character   Median :      1.4   Median :      1.2   Median :    309.7   Mode  :character
                    Mean   :   -220.3   Mean   :   -221.5   Mean   :   3443.4
                    3rd Qu.:      9.6   3rd Qu.:      7.5   3rd Qu.:   1098.7
                    Max.   :  34522.5   Max.   :  34522.5   Max.   :1176509.2
```

```
> colnames(training)
 [1] "Num"                                      "Networth Next Year"
 [3] "Total assets"                             "Net worth"
 [5] "Total income"                             "Change in stock"
 [7] "Total expenses"                           "Profit after tax"
 [9] "PBDITA"                                   "PBT"
[11] "Cash profit"                              "PBDITA as % of total income"
[13] "PBT as % of total income"                 "PAT as % of total income"
[15] "Cash profit as % of total income"         "PAT as % of net worth"
[17] "Sales"                                    "Income from financial services"
[19] "Other income"                             "Total capital"
[21] "Reserves and funds"                       "Deposits (accepted by commercial banks)"
[23] "Borrowings"                               "Current liabilities & provisions"
[25] "Deferred tax liability"                   "Shareholders funds"
[27] "Cumulative retained profits"              "Capital employed"
[29] "TOL/TNW"                                  "Total term liabilities / tangible net worth"
[31] "Contingent liabilities / Net worth (%)"   "Contingent liabilities"
[33] "Net fixed assets"                         "Investments"
[35] "Current assets"                           "Net working capital"
[37] "Quick ratio (times)"                      "Current ratio (times)"
[39] "Debt to equity ratio (times)"             "Cash to current liabilities (times)"
[41] "Cash to average cost of sales per day"    "Creditors turnover"
[43] "Debtors turnover"                         "Finished goods turnover"
[45] "WIP turnover"                             "Raw material turnover"
[47] "Shares outstanding"                       "Equity face value"
[49] "EPS"                                      "Adjusted EPS"
[51] "Total liabilities"                        "PE on BSE"
>
```

```
> dim(training)
[1] 3541   52
```

Inferences:

- We can see that the **dataset** has been imported in the form of **“tbl_df”** which is acronym for **Table Dataframe.**
- The dataframe has been arranged according to the **increasing order of the variable Num.**
- We can see that most of the variables in the dataframe contain **missing values.**
- There are many variables which are actually **numeric** but are shown as **character vectors.**
- The column names don’t have discrepancies and can be used as it is.
- The **dataframe** contains **52 Variables and 3541 Observations.**
- The **dataframe doesn’t** contain **response variable** and hence must be created.

b.Clearing the incosistencies:

- The dataset must converted to **dataframe** using the function **as.dataframe()**.
- The **"Deposits (accepted by commercial banks)"** is to be removed from the dataframe
- A **response variable** must be created using the **variable “Networth Next Year”**. If the **“Networth Next Year”** is greater than **zero,** then the **company** is

said to **not have defaulted** and if the value less than **zero,** the **company** is said to **have defaulted.** This can be done using the function **ifelse().**

```
> ### Creation of the response variable ####
> training$default = ifelse(training$`Networth Next Year`>0,"1","0")
> training$default = as.factor(training$default)
> table(training$default)

   0    1
 243 3298
```

- Changing the **character vectors** to **numeric vectors** using the function **as.numeric()**

```
> ### changing to numeric ####
> training$`Creditors turnover` = as.numeric(training$`Creditors turnover`)
> #### Clearing the Inconsistencies ####
> ### Removal of Variable with all NAs ####
> training = training[,-c(22)]
> ### changing to numeric ####
> training$`Creditors turnover` = as.numeric(training$`Creditors turnover`)
> training$`Finished goods turnover` = as.numeric(training$`Finished goods turnover`)
Warning message:
NAs introduced by coercion
> training$`WIP turnover` = as.numeric(training$`WIP turnover`)
Warning message:
NAs introduced by coercion
> training$`Shares outstanding` = as.numeric(training$`Shares outstanding`)
Warning message:
NAs introduced by coercion
> training$`Equity face value` = as.numeric(training$`Equity face value`)
Warning message:
NAs introduced by coercion
> training$`PE on BSE` = as.numeric(training$`PE on BSE`)
Warning message:
NAs introduced by coercion
> training$`Debtors turnover` = as.numeric(training$`Debtors turnover`)
Warning message:
NAs introduced by coercion
> training$`Raw material turnover` = as.numeric(training$`Raw material turnover`)
Warning message:
NAs introduced by coercion
```

c. Missing Value Treatment:

The **missing values** can be termed as the values that are unknown to the analyst when he gets the data. These values must be dealt with in a proper way so as to not disturb the structure of the dataset. These kind of values also cause hindrances to the **model building process.**

```
> sum(is.na(training))
[1] 14992
```

We can see that the number of **NAs** are **very high** compared to the total number of observations in the dataframe. Hence, these values cannot be removed.

Therefore, these **NAs** can be treated by imputing them to the **median** of the particular **column** in which the **Missing Values** exist. We can do this with the help of **for loop** which **does median imputation** for all the **missing values.**

```
> training2 = training
> for(i in c(1:51)){
+   if(sum(is.na(training2[,i])) > 0){
+     training2[,i][is.na(training2[,i])] = median(training2[,i],na.rm = TRUE)
+   }
+ }
> sum(is.na(training2))
[1] 0
```

d. Outlier Treatment:

An **outlier** can be defined as those values which are at an abnormal distance from the other values in a **sample distribution.** These **outliers** can disturb the **distribution of the sample** and can hinder the **performance** of the **model.**

Lower Outliers are the values which are **less than 1st quartile - 1.5*IQR(Inter Quartile Range) .**

Upper Outliers are the values which are **greater than 3rd quartile + 1.5*IQR(Inter Quartile Range).**

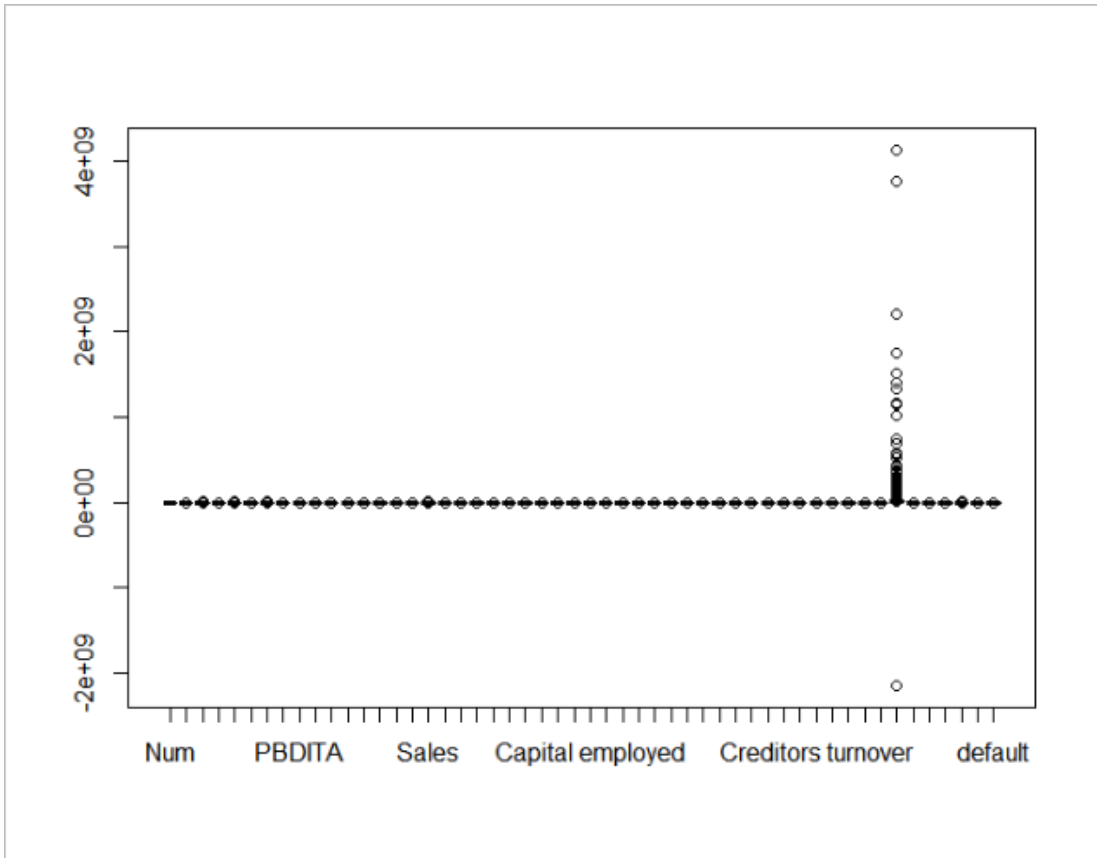
The **outliers** are more in number for our dataset. **Hence** omitting them is not an option.

Therefore to deal with them, we use the method called **capping** where we **impute the outlier values** to the **thresholds** based on whether they are **upper outliers** or **lower outliers.**

Lower threshold can be set at ***1st quartile - 1.5*IQR(Inter Quartile Range)*** while the **upper threshold** can be set at **3rd quartile + 1.5*IQR(Inter Quartile Range).**

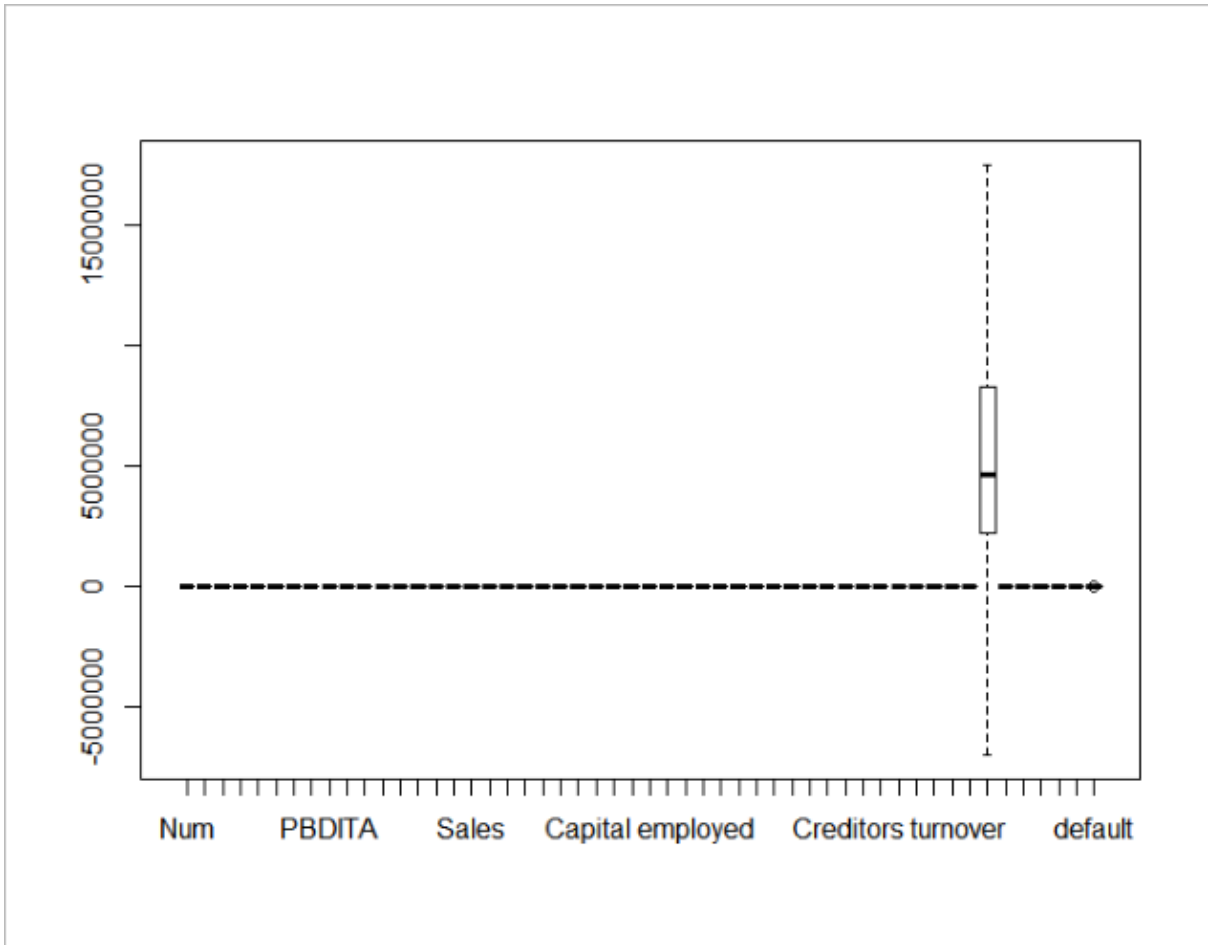
This process of capping can be done by the usage of **custom functions** to find the thresholds and then the actual **imputation** can be done by the means of **for loop.**

<u>Before Outlier Removal:</u>



```
### Treating Outliers ####
training2 = as.data.frame(training2)
boxplot(training2)
outlier2 = function(i)
{}

outcol = function(j){}

for (i in c(1:51)){}

for(g in c(1:51)){}
boxplot(training2)
```

After Outlier Removal:

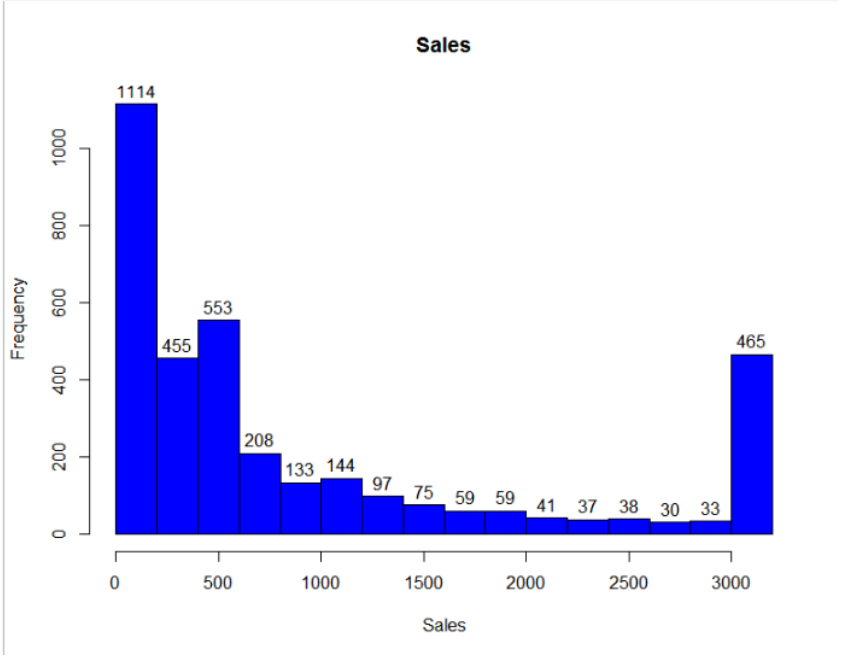


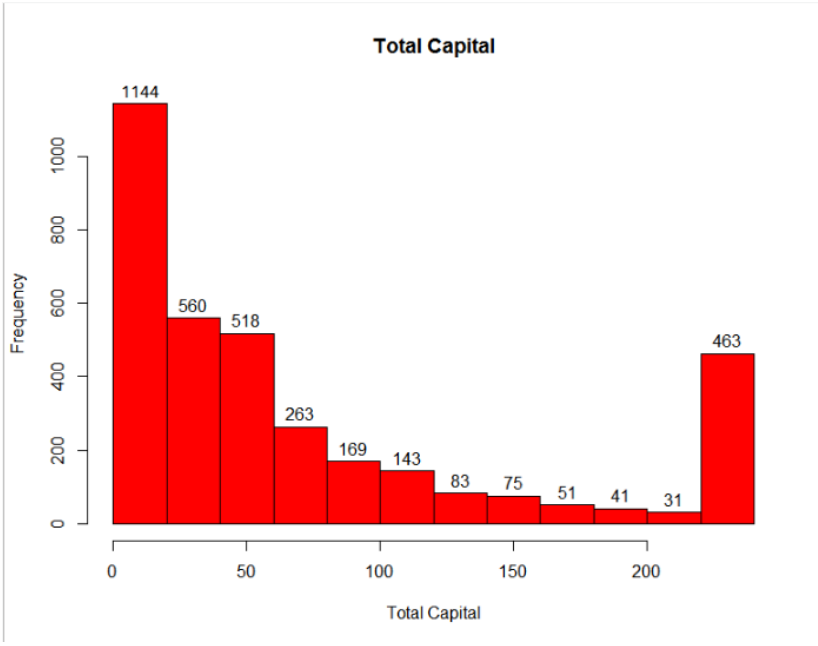
a. Univariate and Bivariate Analysis:

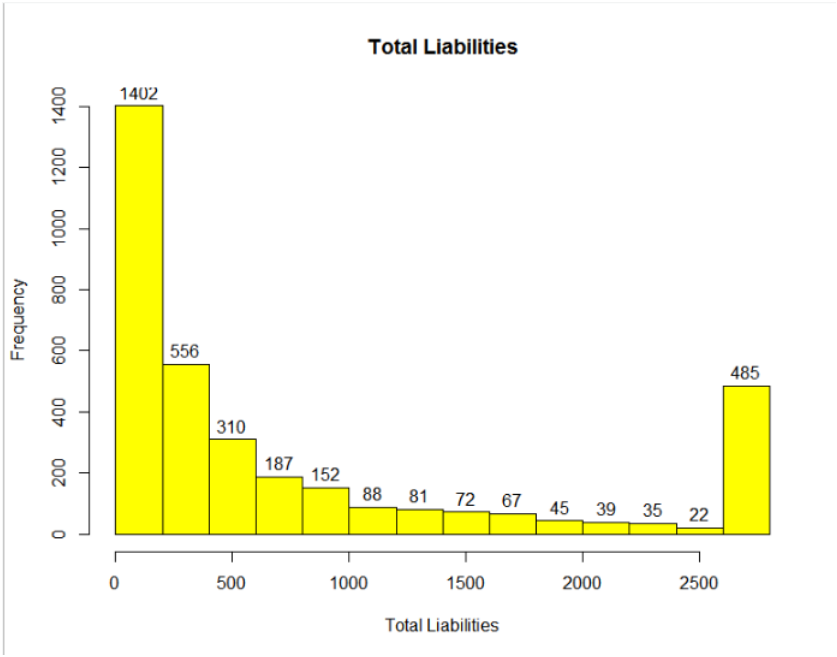
Univariate Analysis:

The **univariate analysis** in our context can be carried out on the important variables such as **Sales, Total Capital and Total Liabilities.** These variables are important because they give us a basic **idea** on **major**

determinants of **cash flow** in the **financial statements** of the companies. It also gives us an idea on **variation** in the **customer base** of the bank.







Inferences:

- We can see that **frequency** of the companies whose **sales** are in the range of **0-500** is very high as compared to the **companies** with **higher sales.** This indicates that the data contains more number of companies **whose sales figures** are **less or even meagre** from the whole dataset. This is an indication of the data containing more number of **non-defaulters** since most of them would not be needing a **huge loans** to **run** their operations and are able to pay off their loans.
- As we can see that higher number of companies have **total capital less than 50** indicating huge number of **small companies** present in the dataset. This is an inclination towards the fact that these small companies would require **small amount loans** and

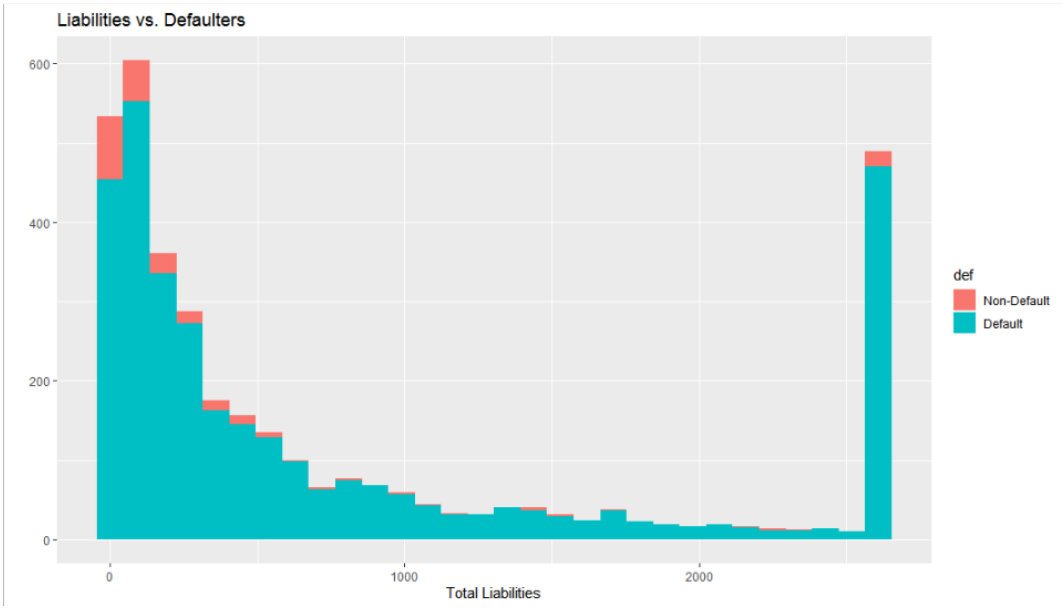
hence are **less prone to defaulting** than the **other large companies.**

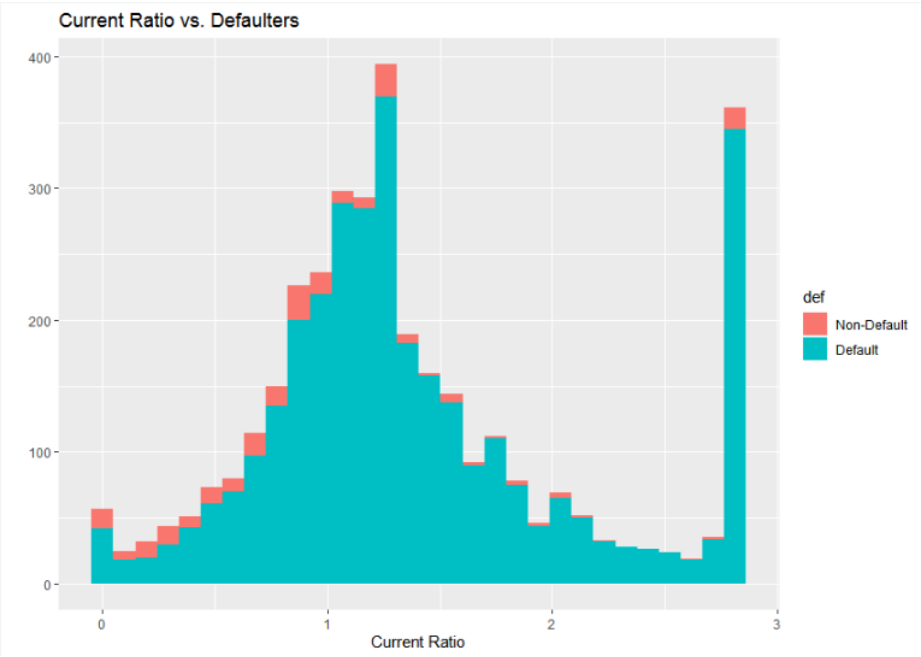
- The number of companies with **higher amount of liabilities** are **very less** in number indicating that the **number of defaulters** that are likely to arise from this dataset would be very less or meagre **compared** to the **number of non-defaulters.**

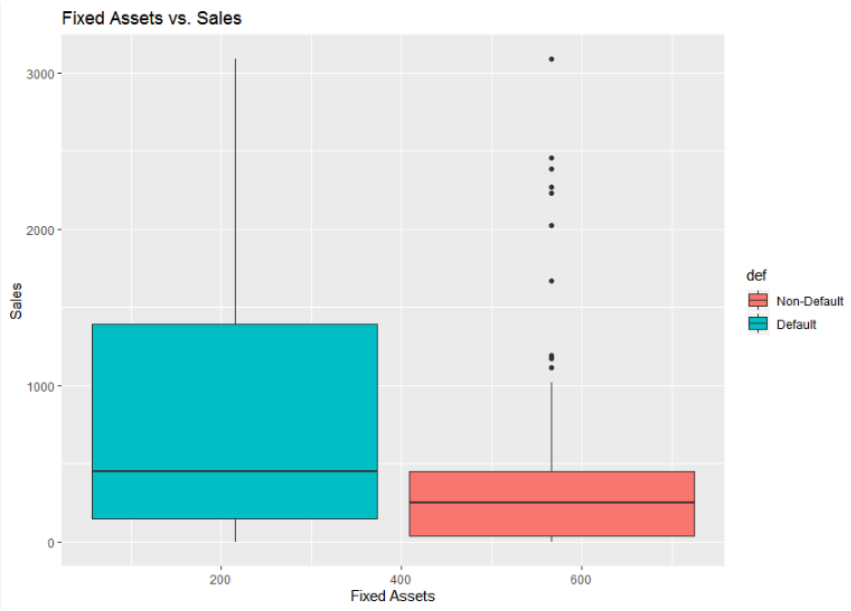
Bivariate Analysis:

The **bivariate Analysis** in our context can be done using the **independent variables** such as **liabilities, Current Ratio** and **Fixed Assets** along with the **response variable "default"**.

```
> ## Bi-Variate Analysis ####
> def = training2$default
> levels(def)
[1] "0" "1"
> levels(def) = c("Non-Default","Default")
```







Inferences:

- There are more number of defaulters have **lower liabilities** than most of the other customers. The **lower the liabilities** the company has, the **lesser the chance** the company has to default on a loan.

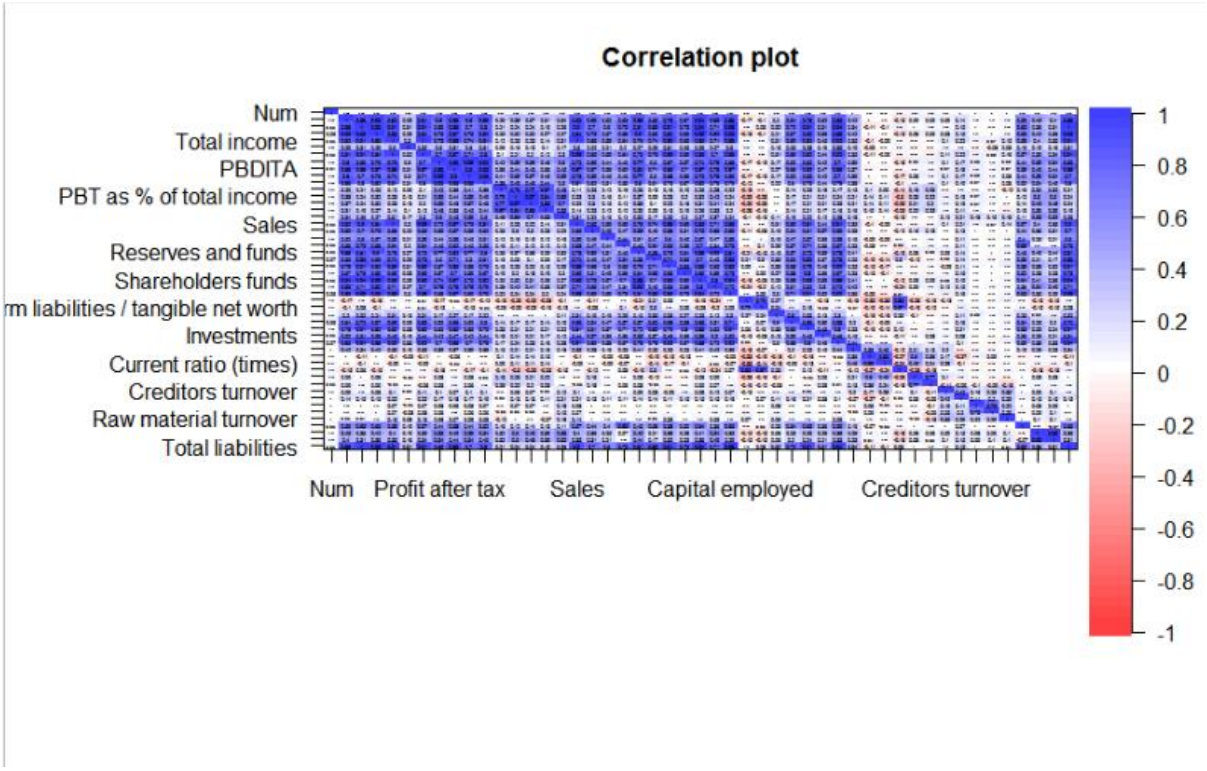
- The **frequency** of **defaulters** and **non-defaulters** seems to be **highest** when the **Current Ratio** of the **companies** is **1** or little **more than 1.** This **indicates** that the companies that have **Current Assets more than or equal to** that of **Current Liabilities** are the ones which most probably will **not default on loan.**
- The **companies** with **more fixed assets** and **less sales** are more likely to **default on their loan** because these companies take **more amount of loan** while building up their **initial capital** but are unable to get back that money due to **less sales** and **fail to pay back** the **loan** let alone **fall** into **debt crunch.**

b. Checking for Multicollinearity:

Multicollinearity is the situation in which one or more **independent variables** are **linearly related** with each other. This situation **degrades the performance of the model** because all the variables in the **model** must always be **truly independent** of each other.
Checking for **multicollinearity** can be done in **two ways:**

I. Correlation:

We can say that if more number of **variables** have **high correlation** between them, we can infer that the **multicollinearity** exists.




```
> ### Finding Multicollinearity ####
> ## Correlation Plot ####
> corr = training2[,-c(52)]
> corr = corr[,-c(51,47)]
> cor.plot(corr,numbers = TRUE)
> cor.m = cor(corr)
> cor.m = as.data.frame(cor.m)
> value = cor.m[cor.m[]>0.90]
> length(value)
[1] 129
```

As we can see, there are **many combination** of **variables** where there is **significant correlation.**

II. Eigen values:

We can find out the **Eigen values** of all the **variables** and check if any of those values are close to **zero.** If so, we can say that **multicollinearity** exists.

```
> ## Eigen Values ####
> eigen = eigen(cor(corr))
> sum(eigen$values[]<0.001)
[1] 1
```

We can see that **one** of the values is very close to **zero.**

Inferences:

From the above two results, it is proved that **multicollinearity exists** in the **dataset.** Multicollinearity can be treated by dropping out the variables **during the model creation** which are the main reason behind the **multicollinearity.**

3) Model Building:

For this dataset, we can use the **logistic regression model** to build a **prediction model. Logistic regression** uses the **logit function** of **probability** of **an event occurring** or **not occurring** to make **predictions.** This **technique** is useful here because we can set the **right threshold** for the probabilities which will help us make predictions in the way which will be **favourable** for the **management.** And since the **dataset** is very imbalanced with **Non-defaulters** in **higher numbers** than **defaulters, logistic regression** will help us build better models.

a. Building the Regression Model:

The **regression model** must be built in **multiple iterations** in order to remove the multicollinearity that is present in the **dataset.** At each iteration, we must remove those variables which are not significant for the model and whose **p-values** are **more than 0.05.** This will help us give a model that will make **good predictions.**

First Iteration:

In this **iteration,** we make a model with all the variables except **Networth next year** because it is the variable that was used to create the **response** variable will result in **mulitcollinearity.**

```
Call:
glm(formula = default ~ . - `Networth Next Year`, family = "binomial",
    data = training3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2045   0.0242   0.1073   0.2390   2.9447

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             2.797e+00  4.202e-01   6.657 2.80e-11 ***
Num                                    -2.262e-04  8.900e-05  -2.541  0.01105 *
`Total assets`                          3.326e-04  1.266e-03   0.263  0.79271
`Net worth`                             1.232e-03  2.350e-03   0.524  0.59995
`Total income`                         -1.275e-03  7.316e-04  -1.743  0.08127 .
`Change in stock`                      -1.680e-02  8.648e-03  -1.943  0.05202 .
`Total expenses`                        7.627e-04  8.019e-04   0.951  0.34156
`Profit after tax`                     -2.094e-02  1.426e-02  -1.468  0.14201
PBDITA                                  3.262e-03  3.215e-03   1.015  0.31030
PBT                                     4.341e-03  1.107e-02   0.392  0.69495
`Cash profit`                           8.515e-03  4.385e-03   1.942  0.05214 .
`PBDITA as % of total income`           1.616e-02  1.538e-02   1.051  0.29336
`PBT as % of total income`              1.693e-02  6.200e-02   0.273  0.78481
`PAT as % of total income`              2.358e-02  8.103e-02   0.291  0.77103
`Cash profit as % of total income`      4.075e-03  2.479e-02   0.164  0.86943
`PAT as % of net worth`                 3.254e-02  7.466e-03   4.358 1.31e-05 ***
Sales                                   3.314e-04  6.185e-04   0.536  0.59209
`Income from financial services`       -5.572e-02  4.376e-02  -1.273  0.20287
`Other income`                         -1.650e-01  8.256e-02  -1.998  0.04569 *
`Total capital`                         7.027e-03  3.214e-03   2.186  0.02880 *
`Reserves and funds`                    3.965e-04  1.079e-03   0.367  0.71328
Borrowings                             -9.790e-04  1.586e-03  -0.617  0.53711
`Current liabilities & provisions`     -6.800e-04  2.195e-03  -0.310  0.75667
`Deferred tax liability`               -6.344e-03  9.659e-03  -0.657  0.51131
`Shareholders funds`                   -3.571e-03  2.440e-03  -1.464  0.14325
`Cumulative retained profits`           7.211e-03  1.700e-03   4.242 2.21e-05 ***
`Capital employed`                      1.703e-03  1.642e-03   1.037  0.29956
`TOL/TNW`                              -2.334e-01  7.896e-02  -2.956  0.00311 **
```

```
`Total term liabilities / tangible net worth`  2.617e-01  1.805e-01   1.450  0.14707
`Contingent liabilities / Net worth (%)`      -5.660e-03  3.625e-03  -1.561  0.11845
`Contingent liabilities`                       5.269e-03  2.705e-03   1.948  0.05144 .
`Net fixed assets`                            -1.202e-03  1.156e-03  -1.040  0.29835
Investments                                   -1.353e-03  1.256e-02  -0.108  0.91420
`Current assets`                               8.913e-04  1.119e-03   0.796  0.42593
`Net working capital`                         -2.588e-03  2.208e-03  -1.172  0.24123
`Quick ratio (times)`                          2.042e-01  3.943e-01   0.518  0.60451
`Current ratio (times)`                        5.482e-01  2.566e-01   2.136  0.03268 *
`Debt to equity ratio (times)`                -4.624e-01  1.169e-01  -3.954 7.68e-05 ***
`Cash to current liabilities (times)`         -3.189e+00  1.095e+00  -2.912  0.00359 **
`Cash to average cost of sales per day`        9.105e-03  7.814e-03   1.165  0.24390
`Creditors turnover`                           2.350e-02  2.247e-02   1.046  0.29563
`Debtors turnover`                             1.233e-02  1.972e-02   0.625  0.53171
`Finished goods turnover`                      5.459e-03  9.190e-03   0.594  0.55247
`WIP turnover`                                 8.242e-03  1.763e-02   0.467  0.64016
`Raw material turnover`                        3.317e-02  1.810e-02   1.833  0.06680 .
`Shares outstanding`                          -2.303e-08  3.334e-08  -0.691  0.48973
`Adjusted EPS`                                 1.055e-01  2.382e-02   4.432 9.35e-06 ***
---
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:  923.3  on 3494  degrees of freedom
AIC: 1017.3

Number of Fisher Scoring iterations: 9
```

We can see that **only 10** variables out of **50** variables are **significant** at a **confidence interval of 95%.**

Second Iteration:

In this iteration, we use all the variables which were significant in the **first iteration** and build a model with them.

```
Call:
glm(formula = default ~ Num + `PAT as % of net worth` + `TOL/TNW` +
    `Total term liabilities / tangible net worth` + `Total capital` +
    `Other income` + `Cumulative retained profits` + `TOL/TNW` +
    `Current ratio (times)` + `Debt to equity ratio (times)` +
    `Cash to current liabilities (times)` + `Adjusted EPS`, family = "binomial",
    data = training3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1360   0.0305   0.1312   0.2631   2.3112

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  3.064e+00  3.066e-01   9.992  < 2e-16 ***
Num                                         -2.306e-04  8.475e-05  -2.721 0.006513 **
`PAT as % of net worth`                      4.502e-02  5.664e-03   7.949 1.88e-15 ***
`TOL/TNW`                                   -2.201e-01  6.730e-02  -3.271 0.001073 **
`Total term liabilities / tangible net worth` 3.243e-01  1.651e-01   1.964 0.049517 *
`Total capital`                              8.254e-03  1.914e-03   4.312 1.62e-05 ***
`Other income`                              -5.112e-02  7.052e-02  -0.725 0.468519
`Cumulative retained profits`                7.240e-03  1.331e-03   5.439 5.36e-08 ***
`Current ratio (times)`                      6.615e-01  1.481e-01   4.466 7.95e-06 ***
`Debt to equity ratio (times)`              -4.217e-01  1.070e-01  -3.940 8.13e-05 ***
`Cash to current liabilities (times)`       -2.391e+00  7.144e-01  -3.347 0.000818 ***
`Adjusted EPS`                               1.006e-01  2.080e-02   4.836 1.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1770.97  on 3540  degrees of freedom
Residual deviance:  990.31  on 3529  degrees of freedom
AIC: 1014.3

Number of Fisher Scoring iterations: 8

> |
```

We can see almost **all the variables** are **significant** in this model. We also see that the model is **better than the** previous model **due to the better AIC value.**

b. Creation of new variables:

The **financial status of the company** can be summarized by the aspect of **financial statements** known as **Ratio Analysis.** This is an important tool in **management accounting** which helps understand the company from **different views.** The major areas where Ratio Analysis can be used are as follows:

I. **Size of the company**
II. **Liquidity**
III. **Leverage**
IV. **Profitability**

We can calculate the **ratios** for these criteria with the **existing variables** and **create new variables** with these ratios that can be used in our model for **better predictions**.

1) **Size of the company**

These ratios help in **comparison** of **two or more companies** on which of them is larger. This comparison can be of **either volume of Sales, Income, profits** or **Equity held with the company.**

One such ratio that can be calculated is as follows:

$$\boldsymbol{Fixed\ Assets\ Ratio} = \frac{\boldsymbol{Fixed\ Assets}}{\boldsymbol{Total\ Assets}} * \mathbf{100}$$

This ratio explains how much **percentage** of the **total assets** present at the company are **Fixed Assets.** A company **having** more number of **fixed assets** such as **machinery, buildings, etc.** is generally a **bigger company.**

2) Liquidity

Liquidity in financial terms can be defined as the **flow of cash during day to day** operations that occur in the company. This **ratios** help in identifying on the extent of **cash flow** in the company.

$$\boldsymbol{Working\ Capital\ Ratio} = \frac{\boldsymbol{Working\ Capital}}{\boldsymbol{Working\ Capital + Total\ Assets}}$$

This ratio represents the **contribution of Working Capital (One of Current Assets)** to the **total Assets.**

3) Leverage

Leverage refers to the extent of **company's capital which** is under **obligations.**

$$\boldsymbol{Assets\ to\ Equity\ Ratio} = \frac{\boldsymbol{Total\ Assets}}{\boldsymbol{Total\ Assets + Total\ Equity}}$$

This ratio represents on how much of **the total equity** of the company is **comprised** of **Assets** spent on the company.

4) Profitability

Profitability means the **extent** to which the company is **acquiring profits** when compared to its **capital and sales.**

$$\textbf{Return on Assets} = \frac{\textbf{Profit before Tax}}{\textbf{Total Assets}} * 100$$

This ratio asses the **percentage of profit** that is got back after **company's investments** in the **Assets.**

Now we can create **new variables** using these **ratios** and **build the model** again **using the** significant **variables.**

```
Call:
glm(formula = default ~ `PAT as % of net worth` + `Cumulative retained profits` +
    `Current ratio (times)` + `Debt to equity ratio (times)` +
    `Cash to current liabilities (times)` + `Adjusted EPS` +
    Fixed.by.total + returnonassets + asset.equity + networkingcapitalratio,
    family = "binomial", data = training4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2043   0.0315   0.1245   0.2430   2.3674

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        6.909e+00  8.795e-01   7.856 3.96e-15 ***
`PAT as % of net worth`            4.333e-02  5.619e-03   7.711 1.25e-14 ***
`Cumulative retained profits`      5.535e-03  1.061e-03   5.218 1.81e-07 ***
`Current ratio (times)`            6.812e-01  1.564e-01   4.354 1.33e-05 ***
`Debt to equity ratio (times)`    -2.191e-01  8.787e-02  -2.494 0.012644 *
`Cash to current liabilities (times)` -2.454e+00  7.600e-01  -3.229 0.001242 **
`Adjusted EPS`                     1.079e-01  2.171e-02   4.972 6.64e-07 ***
Fixed.by.total                    -1.800e-04  5.158e-05  -3.490 0.000483 ***
returnonassets                     9.437e-04  3.906e-04   2.416 0.015681 *
asset.equity                      -6.157e+00  1.171e+00  -5.257 1.47e-07 ***
networkingcapitalratio             5.993e-04  1.545e-04   3.879 0.000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1759.39  on 3532  degrees of freedom
Residual deviance:  944.91  on 3522  degrees of freedom
  (8 observations deleted due to missingness)
AIC: 966.91

Number of Fisher Scoring iterations: 8

>
```

We can see that after the **addition** of the **new variables,** we were able to build a **better model** than the model created in the **second iteration** due to the lower **AIC value.**

c. Analysis of the Coefficients:

We must perform an **analysis** on the **coefficients** of the variables present in the final model so that we can better understand on how the model works and whether it is inclined towards predicting **non-defaulters** or **defaulters.** The **Variables** are as follows:

- **PAT as % of Net Worth –** The **coefficient** is **positive** meaning the **higher** the **profit**, it is more likely that company is **non-defaulter**
- **Cumulative Retained Profits –** The **positive coefficient** refers to the all the profits retained by the company after all the pay-outs suggesting the company would **not-default.**
- **Current Ratio** – The **positive coefficient** tells us how much **current assets** such as cash is available to pay back the loan at a given time and **not be a defaulter.**
- **Debt to Equity Ratio –** The **negative coefficient** tells us that **lower** the **debt** of the company, **higher the chance of the company to not default.**
- **Cash to current liabilities –** We can see that it has **negative coefficient** and its value **is more than 1** meaning that decrease in the **current liabilities** will enable the company to **not to default** on the **loan.**
- **Adjusted EPS –** This variable has **positive coefficient** meaning that **increase in earnings per shares** will lead

to **increase** the chances of the company **not defaulting.**

- **Expenses by Capital –** This variable has **negative coefficient** meaning that **decrease** in percentage of **Expenses** out of the **company's capital** will give the company a chance to **not default** on the **loan.**
- **Fixed by Total Assets –** The **negative coefficient** suggests that the **decrease** in the **amount** of **fixed assets** will lead to **increase** in **probability** of the company **not defaulting.**
- **Return on Assets** – The **positive coefficient** tells us that **more** the number of **profits** earned on the **assets,** there is a **better chance** of the **company not defaulting.**
- **Asset on Equity –** The **negative coefficient** indicates that **less** the company spends on the **assets** from the **equity,** there is **higher probability** of the company **not defaulting** on the loan.
- **Working capital Ratio –** The **positive coefficient** suggests that **increase in working capital's contribution** in the **total capital** gives a **chance** for the company to **not default** on their **loans.**

Coefficients of all these variables suggest that **regression model** we built will give us the **probability of non-defaulters.**

4) Model Performance:

Along with the **raw data**, we are also given a **validation dataset** to **check** the **performance** on how well the **model** is **performing.**

The **validation dataset**, called **testing.xlsx,** must be **imported** into the **R session** using the function called **read_xlsx()** and can be named as **“testing”.**

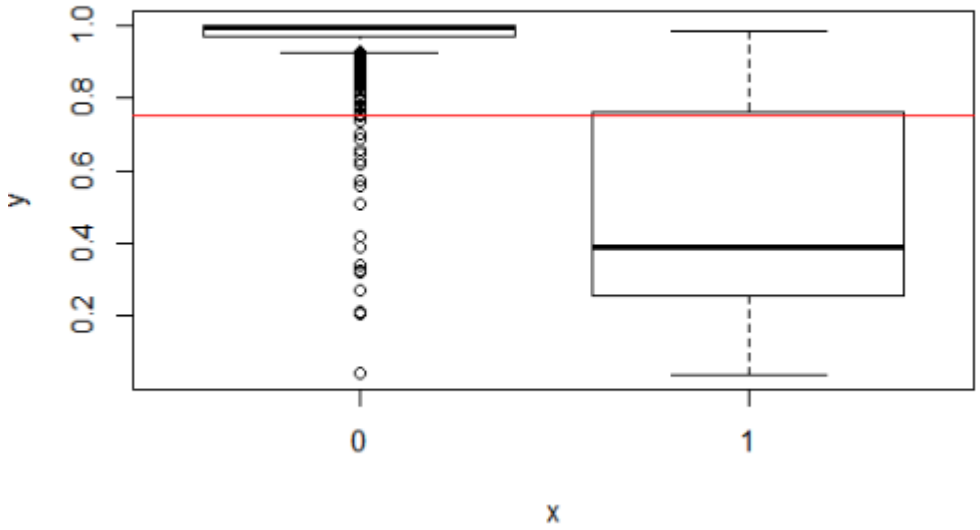
a. Making predictions

Before we can make **predictions** on the **validation dataset,** we must **clear** the dataset of the **NAs** and **outliers** in similar way as done for the **raw data.** Since the **validation data** already has a **response variable,** there is **no necessity** of **creation** of **new variables.** It must also be noted that **the new variables** created using the **ratios** must also be created in this **dataset** also.

After making the dataset ready for **predictions,** we can use the **predict()** function to get our **predictions** with the **final model.**

```
> ### Prediction and Comparision ####
> validpredicitonO = predict(fit3,newdata = testing,type = "response")
> validpredicitonO
        1         2         3         4         5         6         7         8         9
0.99975889 0.98696373 0.54636466 0.92935397 0.99807925 0.99306086 0.99045883 0.95693971 0.99979499
       10        11        12        13        14        15        16        17        18
0.77187837 0.97543176 0.99926715 0.98500094 0.99968443 0.98365738 0.98024048 0.99981724 0.99527776
       19        20        21        22        23        24        25        26        27
0.86379141 0.82315946 0.92606350 0.30149275 0.30037561 0.99079878 0.04456767 0.99416261 0.99994822
       28        29        30        31        32        33        34        35        36
0.99998008 0.96497214 0.96933320 0.83613781 0.99998021 0.99521128 0.96872427 0.98701450 0.64996893
       37        38        39        40        41        42        43        44        45
0.99063038 0.99975570 0.65840421 0.50279404 0.55912209 0.93419940 0.94264796 0.99978770 0.98363419
```

It must be noted that since the **model** has been created in such a way that it **predicts 0,** the **probabilities** in the above results are that of **company <u>not defaulting.</u>**



In the above graph between the **probability of non-default, and response variable** in the **validation dataset,** we can see that **putting** a **threshold** at **0.75** will give us a right **distinction** of the **defaulters** and **non-defaulters.** *Thus we can say that companies with* ***probabilities greater than 0.75*** *are* ***non-defaulters*** *while* ***lesser than that*** *are* ***defaulters.***
Using this **threshold,** we can convert the **probabilities into proper prediction responses.**

b. Analysis on the performance of Model:

To test the performance of **any Classification model,** we have various measures which are as follows:

1) **Confusion Matrix:** This is an important **model performance measure** which consists of a **2 X 2 matrix** of the **rightly predicted and wrongly predicted values.**

```
> caret::confusionMatrix(testing$`Default - 1`,validprediciton,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 639  22
         1  13  40

               Accuracy : 0.951
                 95% CI : (0.9325, 0.9656)
    No Information Rate : 0.9132
    P-Value [Acc > NIR] : 7.717e-05

                  Kappa : 0.6692

 Mcnemar's Test P-Value : 0.1763

            Sensitivity : 0.64516
            Specificity : 0.98006
         Pos Pred Value : 0.75472
         Neg Pred Value : 0.96672
             Prevalence : 0.08683
         Detection Rate : 0.05602
   Detection Prevalence : 0.07423
      Balanced Accuracy : 0.81261

       'Positive' Class : 1
```

We can see that the model has an **accuracy** of **95%** and **balanced accuracy** of **81%** which showcases that the model built is a **good model.** The two types of **wrongly predicted values** that might lead to **losses** for the bank are as follows:

- The first kind of **error** occurs when we give **loan** to a company because we predicted it as **non-defaulter**

but that company **defaults** on the **loan.** These type of customers are the main reason for the **huge losses** incurred by the bank. But in our model, we are able to predict more than **50%** of the **defaulters** correctly and thereby **reducing** the **major part** of the **losses.**

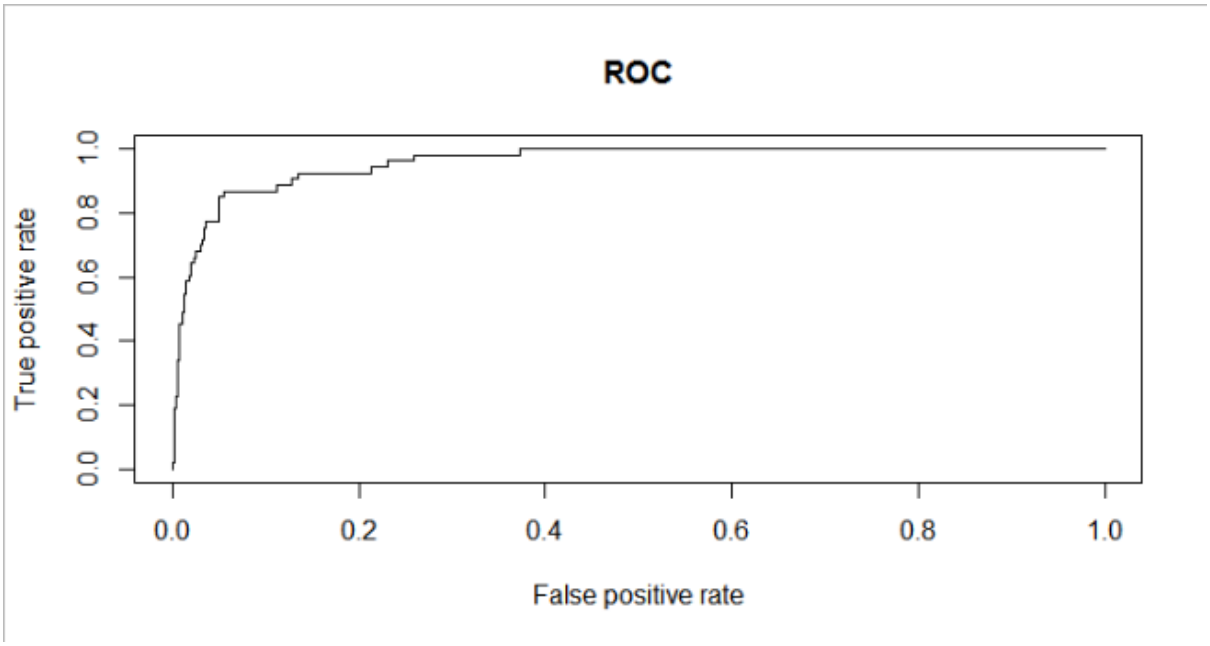
- The **second kind** of **error** occurs when **the bank** predicts a **non-defaulter** as a **defaulter** and doesn't provide him a loan. In the due process, the bank incurs losses due to the number of potential customers it has lost by not providing them loan. This error doesn't lead to as much loss as the **first error** but still must be avoided to some extent. The model was able to predict **93%** correctly **as non-defaulters.**

2) **Concordance Ratio:** In this method, the **values of Response variables and Probabilities are taken as pairs** and then tested if the **probabilities** predicted actually hold true. And then the number of pairs are counted with respect to total number of **pairs.**

```
> Concordance(actuals = x,predictedScores = y)
$Concordance
[1] 0.9588674

$Discordance
[1] 0.04113265

$Tied
[1] -6.938894e-18

$Pairs
[1] 35033
```

The value of **concordance ratio is 95%** meaning that **probabilities** correctly match the **responses** for **95%** of the **pairs.**

3) **ROC & AUC:** Both of these measures help in determining the separation of the different Categories present in the **Target and Prediction Variables.** AUC = Area Under the Curve, ROC = Receiver Operating Characteristics.




```
> #### AUC ####
> auc = performance(pred.obj,"auc")
> auc = as.numeric(auc@y.values)
> print(auc)
[1] 0.9588674
```

We can see that **ROC** value is around **85%** while the **AUC** is **95%** meaning that most of the **True Positives and False Negatives** were identified correctly by the **model.**

4) **Gini:** The **Gini Coefficient** be determined to test the **purity** of the classes divided in the **Target variable using the prediction model.**

```
> ### KS VALUE ###
> print(max(perf@y.values[[1]] - perf@x.values[[1]]))
[1] 0.8119487
```

The **Gini Coefficient** is **81%** meaning that the model is able to remove the **impurities** from the **two sets of classes** while making predictions.

5) **KS Value:** The **KS** (Kolmogorov-Smirnov) value is the **highest separation of classes** that has been achieved by the prediction model.

```
> #### GINI COEFFICIENT ####
> gini1 = ineq(validprediciton.prob1,"gini")
> print(gini1)
[1] 0.8216498
```

The **KS Value** is **82%** indicating that the model is able to **separate** the two classes correctly and give a proper **distinction** while making predictions.

Performance Measure	Value
Accuracy	***95%***
Concordance Ratio	***95%***
ROC	***85%***
AUC	***95%***
Gini	***81%***
KS Value	***82%***
Positive pred. value	***81%***

c. Splitting the dataset into deciles:

We can split the **validation data** into **deciles** depending on the **probability of default.** It arranges the companies into groups based on their **probability of default.** This process of division of the dataset will help management give an idea on **how much risk** they are taking while providing **loan** to a company. The management can decide on which **deciles** the loan must be given to and if given, how much is to be given to each **deciles.** The **deciles** can be created using the **cut()** function and the new dataset with **deciles** can be exported to an **excel file** using **read_excel()** function.

	Fixed.by.total	returnonassets	asset.equity	networkingcapitalratio	probabilityofdefault	predictedvalue	decile
146	6.857118e+01	-8.645722	0.9209494	1409.5	0.961	1	10
25	7.373505e+01	-4.001840	0.9487235	435.8	0.954	1	10
462	8.447059e+01	-3.294118	0.9953162	86.0	0.918	1	10
148	6.424242e+01	43.181818	0.9969789	34.0	0.894	1	9
250	4.145883e+01	-5.227292	0.9143575	2330.6	0.894	1	9
175	4.130435e+02	-10.434783	0.9704641	24.0	0.841	1	9
196	5.757132e+01	-6.291149	0.9125501	137.7	0.826	1	9
495	4.711538e+01	-6.346154	0.9942639	53.0	0.820	1	9
372	2.675097e+01	-2.285992	0.9799809	206.6	0.785	1	8
139	4.515581e+01	-0.509915	0.9135611	177.5	0.782	1	8
683	7.994723e+01	-2.902375	0.9844156	38.9	0.782	1	8
309	1.324825e+01	-1.488909	0.9903701	330.1	0.779	1	8
516	4.902231e+01	-9.880325	0.9186792	1233.5	0.776	1	8
529	7.189073e-02	-15.312725	0.8859873	140.1	0.774	1	8
56	2.726350e+01	-9.611971	0.9385861	395.3	0.755	1	8
313	9.500000e+04	14250.000000	0.5000000	1.1	0.754	1	8
626	3.658537e+01	-2.439024	0.9647059	9.2	0.751	1	8

Like in the above manner, the whole dataset has been **divided** into **deciles.**

The **first decile** contains **3 companies** whose **probability of default** ranges from **0.90-1.00**

The **second decile** contains **5 companies** whose **probability of default** ranges from **0.80-0.90.**

In this way, each **decile** contains certain number of **companies** based on their **probabilities of default** giving management a **clearer** way to take a decision.

5) Conclusion:

The **objective** of this **project** was to create a **logistic regression model** which would assess the **credit risk** and help bank take decision on two things, **whom to provide the loan** and **if provided, how much loan amount is to be sanctioned.** Using the given **raw data,** we were able to create a **logistic regression model** all while including new **ratios** and introducing them in our model for **better predictions.** These **ratios** also helped us better understand **financial statuses** of the companies. The **logistic regression model** was then tested using various performance measures on the **validation dataset** provided to us. Then using the **probability of default** got from the **predictions** made by the **model,** we divided the whole dataset into **deciles.**

The **dataset** with **deciles** can be used by management in **two ways:**

I. The management can decide on whether to **provide a loan** to **a particular decile** by placing a **risk threshold** which would determine the **amount of risk** the bank is willing to take.
II. The management can also decide on a **limit** on the **amount of loan** to be **sanctioned** based on the **decile** the company is present in. **The** companies in the **higher deciles** can be given **lower amounts** can be sanctioned since they have **high probabilities** of **default** and **higher amounts** can be sanctioned to **lower deciles** since they have **low probability of default.**

Based on the analysis done in the project, the following suggestions can be given to the **management:**

- While choosing a dataset for **model creation,** the dataset thus chosen must be **balanced** as the **ratio** of **defaulters** are **only 7%** while the **non-defaulters** are **93%.**
- If the management is inclined **towards** correctly **predicting the defaulters** over the **non-defaulters,** the dataset must either be **balanced** or at the most be inclined towards **defaulters** so that all our predictions will be inclined towards more number of **defaulters.**

- The **dataset** with least number of **missing values** must be favoured because these values can cause **hindrances** to model building.
- The **dataset** must be free of any **outliers** so that it will **reduce** the **deviations** that occur in the model while making **predictions.**
- The dataset had a **very significant multicollinearity** which calls for **careful selection** of variables for the dataset that do not have **correlation** between them.
- The number of **variables** could have been reduced as few of them **didn't** contribute to the **model building.** The **significance** of the variables could have been improved over the number of **variables.**
- The **risk threshold** must be decided carefully depending on the overall **financial statuses** of the **customer base.**
- The **loan amount** to be provided to **the deciles** must be solely dependent on the **financial status** of the bank and how much amount the bank is willing to risk.

Credit Risk is a prevailing problem in the **economy** of **India.** It is a regular problem dealt by every bank in **India.** Therefore the **proper analysis** of the **financial status** of the enemies is necessary for the bank so that **banks** sustain **financially.** Every decision taken by the bank affects the financial status of the bank**.**