GREAT LAKES
INSTITUTE OF MANAGEMENT

Financial Risk Analytics Project

Amit KULKARNI

GREAT LAKES
INSTITUTE OF MANAGEMENT

Table of Contents:

Project Approach.....03

Discovery.....03

1.1: Objective.....03

1.2: The hypothesis.....03

1.3: Potential Data Sources.....03

1.1.3: Applications list.....03

2: Data – Basic Exploratory Analysis.....04

2.1: Dimensions of the dataset provided.....04

2.2: Structure of the data 04

2.3: Missing data.....05

2.4: Density plot of the data.....06

3: Outliers.....07

3.1: Quantiles.....07

3.2: Exploratory data and density plots post the capping and flooring of the data.....08

3.3: Boxplot for the variables.....09

3.4: Imputation of the missing data.....09

4: New Variable creations:.....10

4.1: Ratios.....10

4.2: Multicollinearity and exclusions of variables.....11

5: Univariate Analysis and multivariate Analysis.....11

6: Modelling.....13

6.1: Modelling Technique.....13

6.2: Testing the validation data.....13

6.3: Comparison.....14

1: Project Approach

A typical developmental Lifecycle can be adopted for this project which consists the following

1. Discovery

1.1: Objective

Varity of financial institutions advance finances to companies. To provide the finances financial institutions study variety of parameters like the net worth, profits (after taxes, before taxes etc.) revenue, financial ratios, current ratios etc. Even after the study there are chances that the companies may default. This is called financial credit risk. The uncertainties associated with recovering the advanced money is called the financial risks.

Using the given data set, develop a predictive model using the logistic regression model on the train and test data

1.2: The hypothesis: All the provided variables do not have impact on the dependent variables.

Alternative Hypothesis: All the provided variables impact the dependent variables.

1.3: Potential Data Sources: The data source referred in this project are 2 .csv files which has the requisite information for the modelling purpose.

1.4: Applications list:

i. R
ii. MS Office

2: Data – Basic Exploratory Analysis

2.1: Dimensions of the dataset provided

No of Rows	No of columns
3541	50

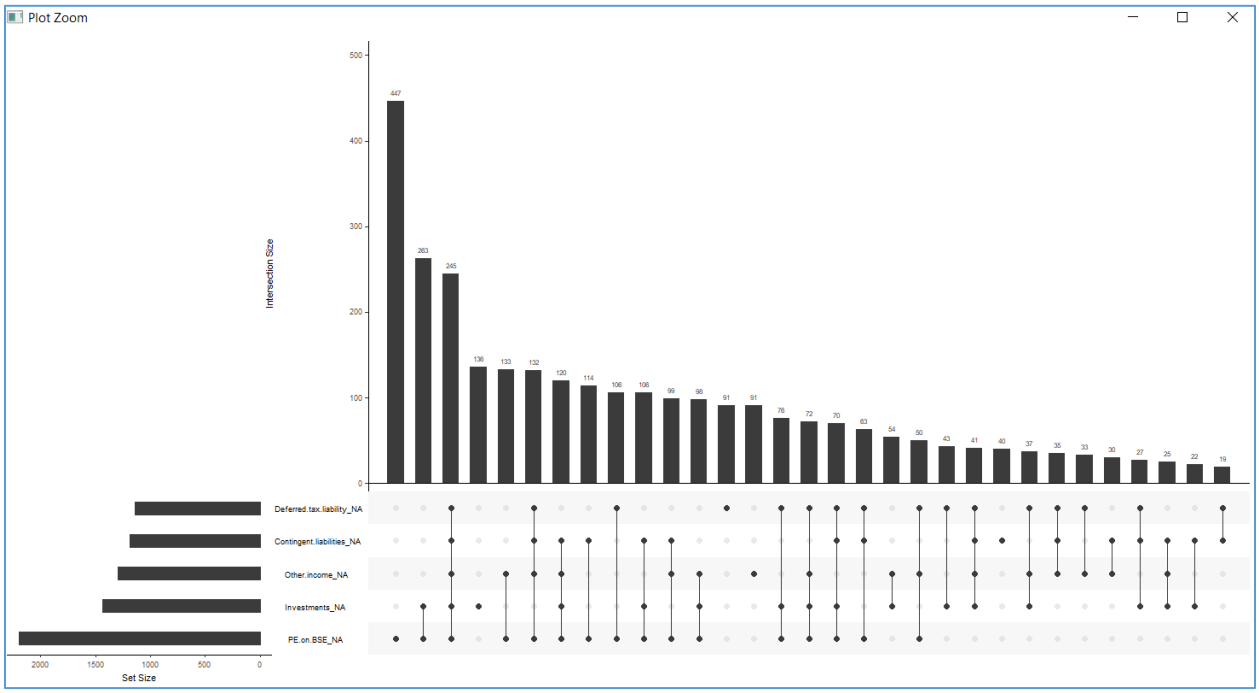
2.2: Structure of the data

data.frame': 3541 obs. of 50 variables:	
Variables	**Defintion**
$ Networth.Next.Year	num 8890.6 394.3 92.2 2.7 109 ...
$ Total.assets	num 17512.3 941 232.8 2.7 478.5 ...
$ Net.worth	num 7093.2 351.5 100.6 2.7 107.6 ...
$ Total.income	num 24965 1527 477 NA 1580 ...
$ Change.in.stock	num 235.8 42.7 -5.2 NA -17 ...
$ Total.expenses	num 23658 1455 479 NA 1558 ...
$ Profit.after.tax	num 1543.2 115.2 -6.6 NA 5.5 ...
$ PBDITA	num 2860.2 283 5.8 NA 31 ...
$ PBT	num 2417.2 188.4 -6.6 NA 6.3 ...
$ Cash.profit	num 1872.8 158.6 0.3 NA 11.9 ...
$ PBDITA.as...of.total.income	num 11.46 18.53 1.22 0 1.96 ...
$ PBT.as...of.total.income	num 9.68 12.33 -1.38 0 0.4 ...
$ PAT.as...of.total.income	num 6.18 7.54 -1.38 0 0.35 2.81 0 0.72 8.29
$ Cash.profit.as...of.total.income	num 7.5 10.38 0.06 0 0.75 ...
$ PAT.as...of.net.worth	num 23.78 38.08 -6.35 0 5.25 ...
$ Sales	num 24458 1504 476 NA 1575 ...
$ Income.from.financial.services	num 158 4 1.5 NA 3.9 6.4 NA NA 7.3 NA ...
$ Other.income	num 297.2 15.9 0.2 NA 0.9 ...
$ Total.capital	num 423.8 115.5 81.4 0.5 6.2 ...
$ Reserves.and.funds	num 6822.8 257.8 19.2 2.2 161.8 ...
$ Borrowings	num 14.9 272.5 35.4 NA 193.1 ...
$ Current.liabilities...provisions	num 9965.9 210 96.8 NA 112.8 ...
$ Deferred.tax.liability	num 284.9 85.2 NA NA 4.6 ...
$ Shareholders.funds	num 7093.2 351.5 100.6 2.7 107.6 ...
$ Cumulative.retained.profits	num 6263.3 247.4 32.4 2.2 82.7 ...
$ Capital.employed	num 7108.1 624 136 2.7 300.7 ...
$ TOL.TNW	num 1.33 1.23 1.44 0 2.83 1.8 0.03 5.17 1.05
$ Total.term.liabilities...tangible.net.worth	num 0 0.34 0.29 0 1.59 0.37 0.03 0.94 0.3 0.
$ Contingent.liabilities...Net.worth....	num 14.8 19.2 45.8 0 34.9 ...
$ Contingent.liabilities	num 1049.7 67.6 46.1 NA 37.6 ...
$ Net.fixed.assets	num 1900.2 286.4 38.7 2.5 94.8 ...
$ Investments	num 1069.6 2.2 4.3 NA 7.4 ...
$ Current.assets	num 13277.5 563.9 167.5 0.2 349.7 ...
$ Net.working.capital	num 3588.5 203.5 59.6 0.2 215.8 ...
$ Quick.ratio..times.	num 1.18 0.95 1.11 NA 1.41 0.48 NA 0.54 0.59
$ Current.ratio..times.	num 1.37 1.56 1.55 NA 2.54 1.27 NA 1.15 1.58
$ Debt.to.equity.ratio..times.	num 0 0.78 0.35 0 1.79 1.09 0.32 2.31 0.94 3
$ Cash.to.current.liabilities..times.	num 0.43 0.06 0.21 NA 0 0.11 NA 0.04 0.19 0
$ Cash.to.average.cost.of.sales.per.day	num 68.21 5.96 17.07 NA 0 ...
$ Creditors.turnover	num 3.62 9.8 5.28 0 13 ...
$ Debtors.turnover	num 3.85 5.7 5.07 0 9.46 ...
$ Finished.goods.turnover	num 200.55 14.21 9.24 NA 12.68 ...
$ WIP.turnover	num 21.78 7.49 0.23 NA 7.9 ...
$ Raw.material.turnover	num 7.71 11.46 NA 0 17.03 ...
$ Shares.outstanding	num 42381675 11550000 8149090 52404 619635 .
$ Equity.face.value	num 10 10 10 10 10 10 10 NA 10 10 ...
$ EPS	num 35.52 9.97 -0.5 0 7.91 ...
$ Adjusted.EPS	num 7.1 9.97 -0.5 0 7.91 ...
$ Total.liabilities	num 17512.3 941 232.8 2.7 478.5 ...
$ PE.on.BSE	num 27.31 8.17 -5.76 NA NA ...

2.3: Missing data

There are **14992** NAs in the given data set

Intersection of the data to check the NA values



To treat the missing values outliers and imputing the missing data below approach is followed.

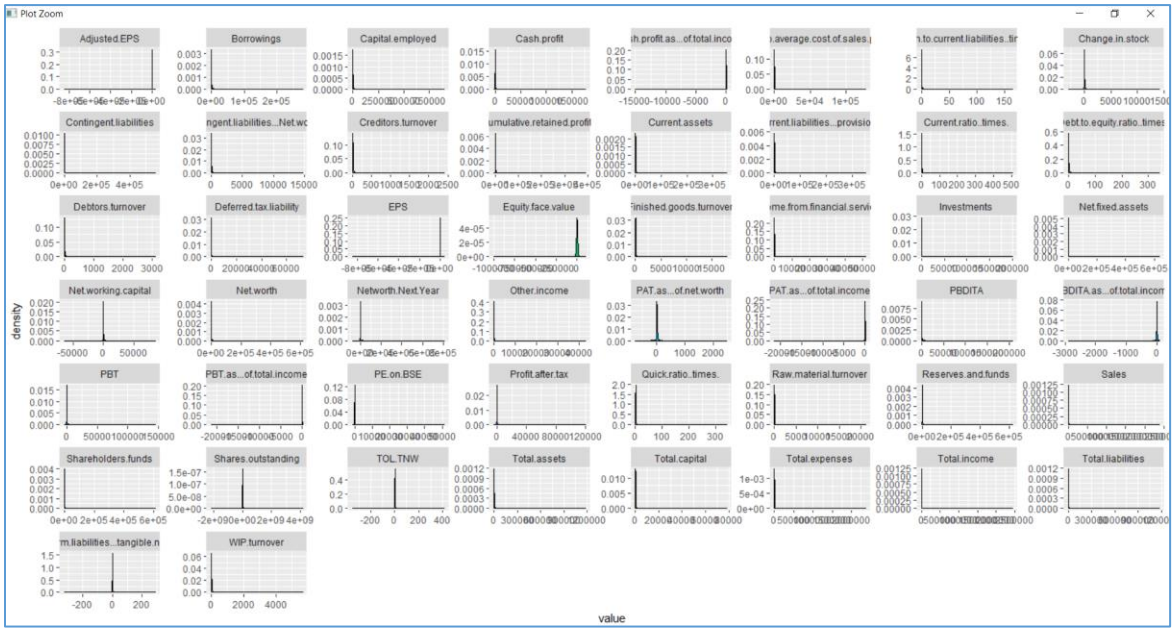
Library upset and ggplot2 is used to treat and impute the missing values.

We now know that there are 14992 missing values across the dataset. With various intersections of missing data across variables we can exclude the intersections and visualize the density plots and boxplots.

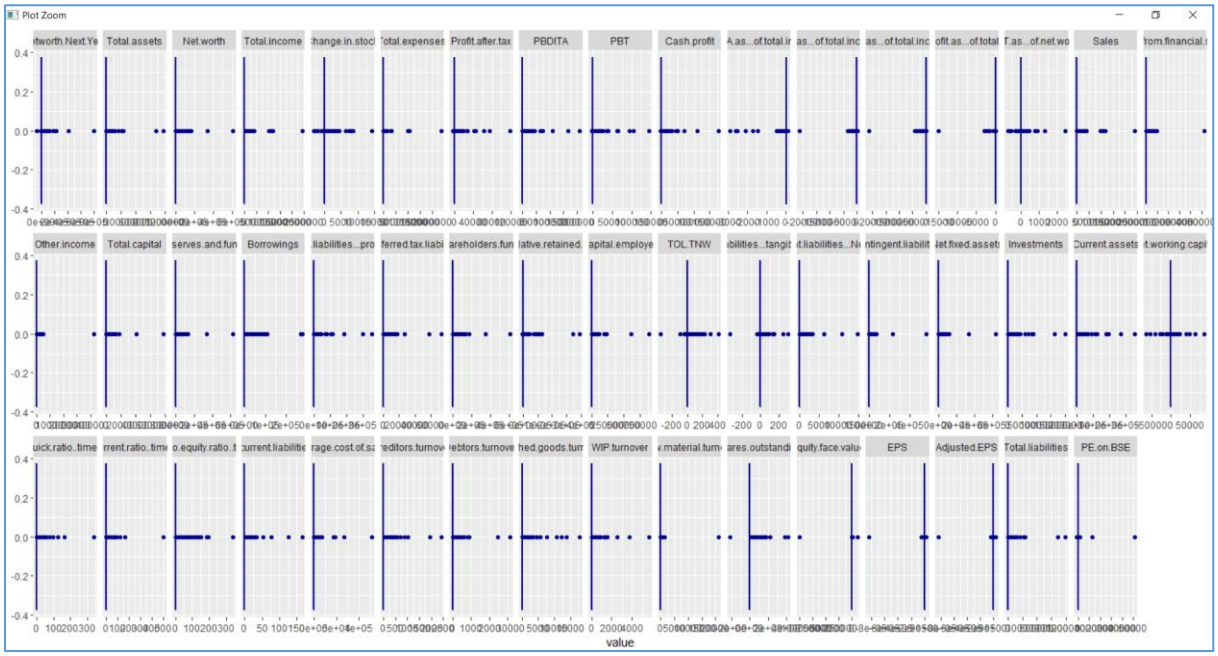
```
data1<-data[!(is.na(data$Other.income))|
            !(is.na(data$Deferred.tax.liability))|
            !(is.na(data$Contingent.liabilities))|
            !(is.na(data$Investments))|
            !(is.na(data$PE.on.BSE)),]
```

2.4: Basic Density plots for all variables of the data

Below graphs depicts the distribution of the data for all the variables. Skewness and tails are shown below. All the data show a normal distribution with long tails on either sides



Additionally boxplots of each variable has been summarized in the below image. There are outliers for all the variables which can be treated



3: Outlier Treatment

3.1: Quantiles

We can check the minimum and maximum values of the columns at 1% and 99% quantile

```
aa<-quantile(p.data$Total.assets, probs = c(0.01, 0.99),na.rm = TRUE)
aa
summary(p.data$Total.assets)
```

If the minimum and maximum values are seen with the 1st and 3rd quartile values a 118.10 and 1249.70, it can observed that the minimum and maximum values are 0.50 and 1,176,509.20 which is significantly away from the 1st and 3rd quartile values and thus it becomes necessary to have a capping and flooring for the data.

A loop formula can be used to cap and floor the outlier values in the dataset.

Below is the result of the capping and flooring

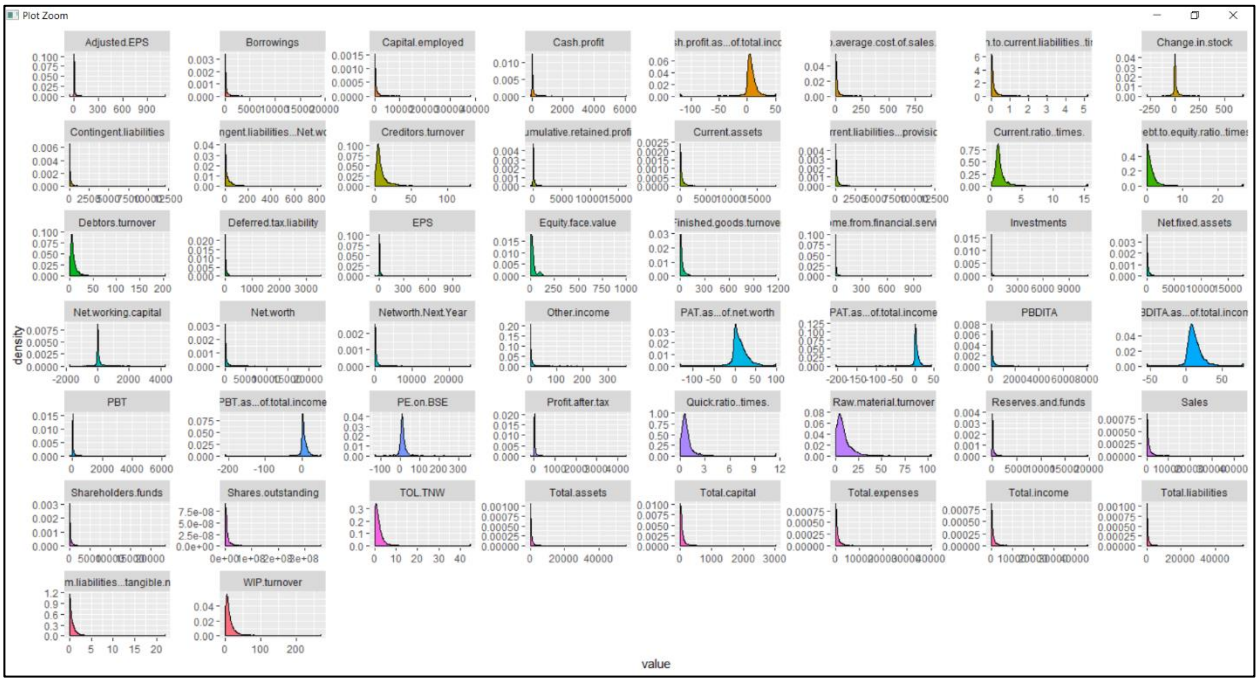
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.500	118.100	359.200	3694.400	1249.700	1176509.200

Below coding can be used for capping all the variables in a loop

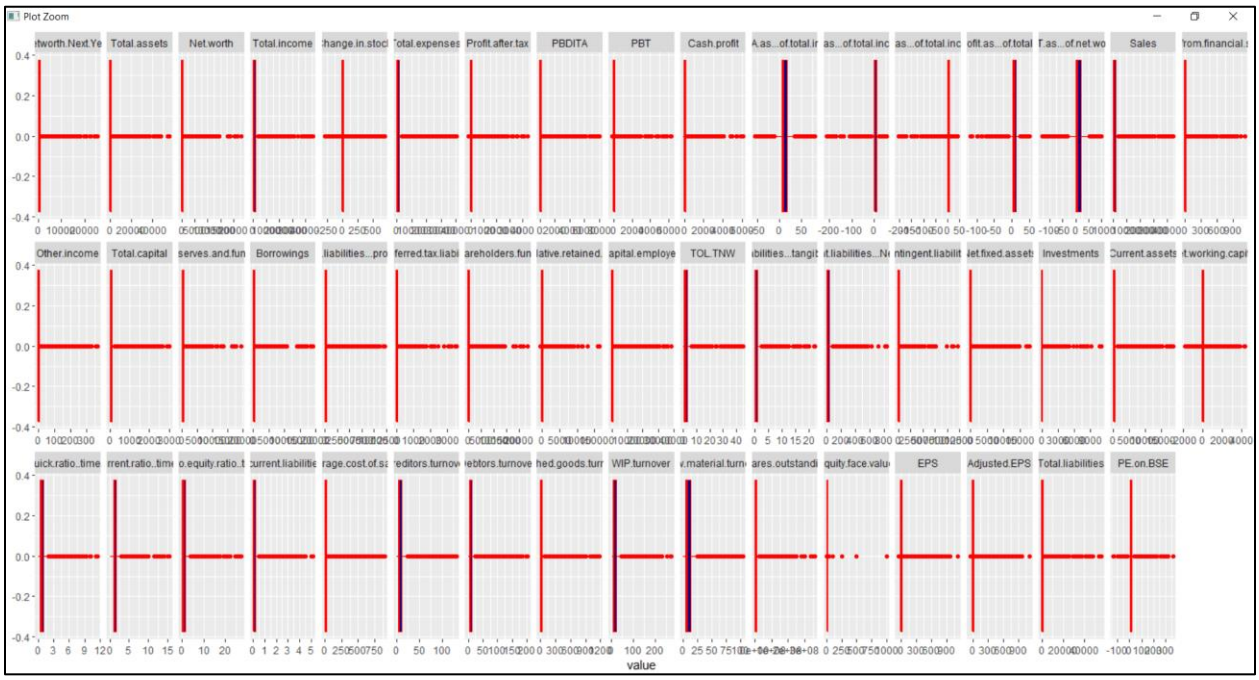
```
p.range<- NA
for(i in 1:(ncol(p.data))){
  Statistic <- data.frame(
    "column" = colnames(p.data[i]),
    "min value" = min(p.data[[i]], na.rm = TRUE),
    "1st Percentile" = quantile(p.data[[i]],probs = c(0.01), na.rm = TRUE),
    "max value" = max(p.data[[i]], na.rm = TRUE),
    "99th Percentile" = quantile(p.data[[i]],probs = c(0.99), na.rm = TRUE))
  p.range<-rbind(p.range, Statistic)
}
p.range <- data.table :: data.table(p.range)
p.range
```

	column	min.value	X1st.Percentile	max.value	X99th.Percentile
1:	<NA>	NA	NA	NA	NA
2:	Networth.Next.Year	-7.426560e+04	-95.1950	8.057734e+05	2.579228e+04
3:	Total.assets	5.000000e-01	4.5950	1.176509e+06	5.612527e+04
4:	Net.worth	1.000000e-01	1.3000	6.131516e+05	2.297124e+04
5:	Total.income	0.000000e+00	0.6000	2.442828e+06	4.567120e+04
6:	Change.in.stock	-3.029400e+03	-283.0520	1.418550e+04	6.856860e+02
7:	Total.expenses	-1.000000e-01	0.3000	2.366035e+06	4.030095e+04
8:	Profit.after.tax	-3.908300e+03	-192.7600	1.194391e+05	4.419584e+03
9:	PBDITA	-4.407000e+02	-24.1680	2.085765e+05	8.049580e+03
10:	PBT	-3.894800e+03	-209.5920	1.452926e+05	6.265964e+03
11:	Cash.profit	-2.245700e+03	-117.4400	1.769118e+05	6.100020e+03
12:	PBDITA.as...of.total.income	-2.900000e+03	-55.4085	1.000000e+02	7.959330e+01
13:	PBT.as...of.total.income	-2.134000e+04	-209.6263	1.000000e+02	5.204680e+01
14:	PAT.as...of.total.income	-2.134000e+04	-204.5594	1.500000e+02	4.350500e+01
15:	Cash.profit.as...of.total.income	-1.502000e+04	-119.6433	1.000000e+02	5.121710e+01
16:	PAT.as...of.net.worth	-7.487200e+02	-131.0310	2.466670e+03	9.885800e+01
17:	Sales	1.000000e-01	0.8600	2.384984e+06	4.420142e+04
18:	Income.from.financial.services	0.000000e+00	0.1000	5.193820e+04	1.107292e+03
19:	Other.income	0.000000e+00	0.1000	4.285670e+04	3.680800e+02
20:	Total.capital	1.000000e-01	0.5000	7.827320e+04	3.028585e+03
21:	Reserves.and.funds	-6.525900e+03	-233.1500	6.251378e+05	1.952580e+04
22:	Borrowings	1.000000e-01	0.3000	2.782573e+05	1.917242e+04
23:	Current.liabilities...provisions	1.000000e-01	0.2000	3.522403e+05	1.216246e+04
24:	Deferred.tax.liability	1.000000e-01	0.1000	7.279660e+04	3.536700e+03
25:	Shareholders.funds	1.000000e-01	1.4000	6.131516e+05	2.351967e+04
26:	Cumulative.retained.profits	-6.534300e+03	-515.7300	3.901338e+05	1.638021e+04
27:	Capital.employed	2.000000e-01	2.8950	8.914089e+05	3.826386e+04
28:	TOL.TNW	-3.504800e+02	0.0000	4.112700e+02	4.502600e+01
29:	Total.term.liabilities...tangible.net.worth	-3.256000e+02	0.0000	2.920200e+02	2.199050e+01
30:	Contingent.liabilities...Net.worth....	0.000000e+00	0.0000	1.470427e+04	8.330960e+02
31:	Contingent.liabilities	1.000000e-01	0.1000	5.595068e+05	1.215832e+04
32:	Net.fixed.assets	0.000000e+00	0.4000	6.366046e+05	1.773118e+04
33:	Investments	0.000000e+00	0.1000	1.999786e+05	1.105417e+04
34:	Current.assets	1.000000e-01	0.4680	3.548152e+05	1.872664e+04
35:	Net.working.capital	-6.383900e+04	-1783.1240	8.578280e+04	4.297456e+03
36:	Quick.ratio..times.	0.000000e+00	0.0100	3.410000e+02	1.155940e+01
37:	Current.ratio..times.	0.000000e+00	0.0400	5.050000e+02	1.543970e+01
38:	Debt.to.equity.ratio..times.	0.000000e+00	0.0000	3.411800e+02	2.714300e+01

3.2: Exploratory data and density plots post the capping and flooring of the data.



3.3: Boxplot for the variables



3.4: Imputation of the missing data

Using the below codes we can impute the values for the data set

```
p.newdata<-p.data %>%
  transmute_all(funs(squish(.,quantile(.,c(0.01,0.99), na.rm = TRUE))))
summary(p.newdata)
# rechecking the boxplot after capping and flooring
p.stack<-melt(p.newdata)
ggplot(data=p.stack, aes(y=value))+
  geom_boxplot(color= "red", fill="dark blue", aplha=0.3)+
  facet_wrap(~variable, scales = "free_x", nrow = 3)+
  coord_flip()
# Rechecking the distribution post the capping and flooring
p.newdata%>%
  gather(metric, value)%>%
  ggplot(aes(value, fill = metric))+
  geom_density(show.legend = FALSE)+
  facet_wrap(~metric, scales = "free")
# imputing missing values
p.newdata <- p.newdata%>%
mutate_if(is.numeric, zoo:: na.aggregate)
summary(p.newdata)
View(p.newdata)
View(data1)
dim(p.newdata)
colnames(p.newdata)[49]<-"Networth.Next.Year"
View(p.newdata)
#adding feature variable to dataset
p.newdata$default <- ifelse(p.newdata$Networth.Next.Year<=0, 1, 0)
```

All the values are now imputed and the data is ready for a model building

4: New Variable creations

4.1: Below ratios need to be created

- Profitability
- Leverage
- Liquidity
- Profitability ratio

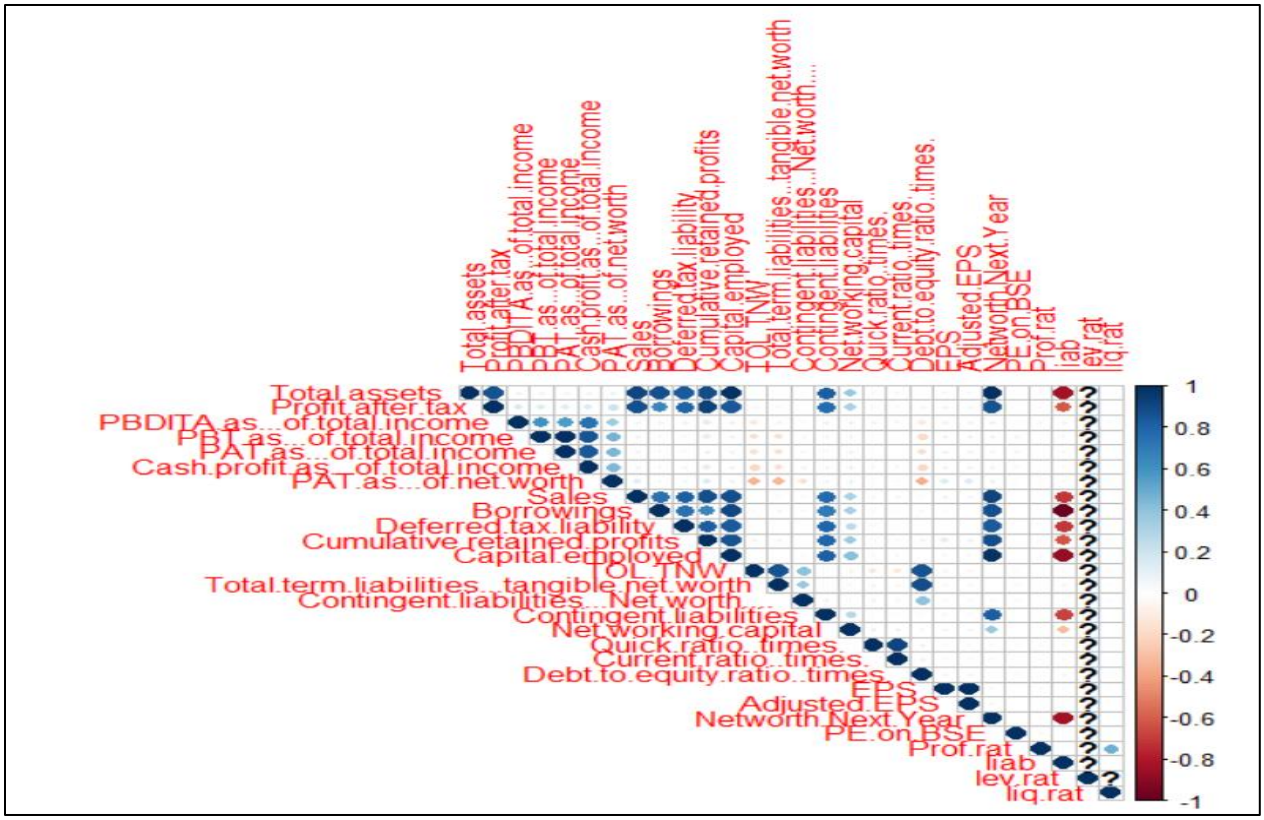
Basis the given the dataset information, below ratio can be calculated

```
#creating new ratios
p.newdata$Prof.rat<- p.newdata$Total.income/p.newdata$Total.assets
p.newdata$liab<- p.newdata$Total.capital- p.newdata$Borrowings
p.newdata$lev.rat<- p.newdata$Borrowings/p.newdata$liab
p.newdata$liq.rat<- p.newdata$Current.assets/p.newdata$Current.liabilities...provisions
```

4.2: Multicollinearity and exclusions of variables

A sample Multicollinearity between variables has been done and below variables will be dropped

	default	Networth.Next.Year	Total.assets	Net.worth	Total.income	Change.in.stock	Total.expenses	Profit.after.tax	PBDITA	PBT
default	1.0000	-0.0767	-0.0233	-0.0495	-0.0484	-0.0337	-0.0453	-0.0725	-0.0541	-0.0699
Networth.Next.Year	-0.0767	1.0000	0.9349	0.9733	0.8914	0.4644	0.8731	0.9238	0.9398	0.9252
Total.assets	-0.0233	0.9349	1.0000	0.9505	0.9165	0.4827	0.9066	0.8659	0.9365	0.8707
Net.worth	-0.0495	0.9733	0.9505	1.0000	0.8846	0.4621	0.8678	0.9035	0.9438	0.9094
Total.income	-0.0484	0.8914	0.9165	0.8846	1.0000	0.4997	0.9963	0.8786	0.9295	0.8869
Change.in.stock	-0.0337	0.4644	0.4827	0.4621	0.4997	1.0000	0.5125	0.4323	0.4775	0.4413
Total.expenses	-0.0453	0.8731	0.9066	0.8678	0.9963	0.5125	1.0000	0.8504	0.9110	0.8599
Profit.after.tax	-0.0725	0.9238	0.8659	0.9035	0.8786	0.4323	0.8504	1.0000	0.9619	0.9963
PBDITA	-0.0541	0.9398	0.9365	0.9438	0.9295	0.4775	0.9110	0.9619	1.0000	0.9642



5: Univariate Analysis and multivariate Analysis

Summary of Univariate Analysis is provided below

All the cells colored in green are the significant variables that tend to impact the dependent variable of default which is the dependent variable.

Though the collinearity was done there can be still some variables that may be existing that are co-related between themselves which can cause noise to the data.

variables	Estimate	Std. Error	t value	Pr(>|t|)	Column1
(Intercept)	0.04117	0.00657	6.27	0.000000	***
Total.assets	0.00000	0.00000	1.13	0.257094	
Profit.after.tax	0.00006	0.00002	3.64	0.000274	***
PBDITA.as...of.total.income	0.00167	0.00034	4.84	0.000001	***
PBT.as...of.total.income	0.00032	0.00091	0.35	0.725871	
PAT.as...of.total.income	-0.00064	0.00093	-0.69	0.492703	
Cash.profit.as...of.total.income	-0.00219	0.00044	-5.02	0.000001	***
PAT.as...of.net.worth	-0.00259	0.00014	-18.56	0.000000	***
Sales	0.00000	0.00000	-2.10	0.035839	*
Borrowings	0.00002	0.00001	1.74	0.081501	.
Deferred.tax.liability	0.00001	0.00002	0.38	0.700678	
Cumulative.retained.profits	-0.00001	0.00001	-1.51	0.132206	
Capital.employed	0.00000	0.00001	-0.67	0.501196	
TOL.TNW	0.00736	0.00122	6.03	0.000000	***
Total.term.liabilities...tangible.net.worth	-0.01690	0.00284	-5.96	0.000000	***
Contingent.liabilities...Net.worth....	0.00005	0.00003	1.45	0.146876	
Contingent.liabilities	-0.00001	0.00001	-1.67	0.095266	.
Net.working.capital	0.00000	0.00001	-0.23	0.821642	
Quick.ratio..times.	0.00896	0.00552	1.62	0.104789	
Current.ratio..times.	-0.00990	0.00418	-2.37	0.017791	*
Debt.to.equity.ratio..times.	0.02162	0.00221	9.77	0.000000	***
EPS	0.00041	0.00032	1.26	0.208414	
Adjusted.EPS	-0.00037	0.00032	-1.14	0.253012	
PE.on.BSE	-0.00022	0.00010	-2.27	0.023327	*
Prof.rat	0.00016	0.00012	1.38	0.167586	
liab	0.00003	0.00001	2.15	0.032055	*
liq.rat	-0.00001	0.00002	-0.44	0.659322	

Overall adjusted R-square is at ~16.5% which is very low.

6: Modelling

6.1: Logistical Regression supervised learning technique to be used since we have binary data as a predictor

```
#modelling using the logistic regression
library(SDMTools)
library(pROC)
library(Hmisc)
lrdata= read.csv("rd.csv")
lrm1 = glm (final.data$default~ .,data=final.data, family= binomial)
summary(lrm1)
pred.rd<-predict(lrm1, newdata=final.data, type="response")
table(final.data$default, pred.rd>0.5)
(3077+109)/nrow(na.omit(final.data))
```

The AIC for the model is 830.74

Confusion matrix of the model

Confusion Matrix		
	FALSE	TRUE
0	3077	31
1	109	79

Model Accuracy as below

```
> (3077+109)/nrow(na.omit(final.data))
[1] 0.9666262
```

Accuracy for the true positives is

42.02%

6.2: Testing the validation data.

All the steps mentioned above are used for the validation data and a final model is developed with the below matrices of the

For the validation data

The AIC is 173.63

Confusion matrix

Confusion Matrix		
	FALSE	TRUE
0	625	5
1	15	29

The accuracy of the model is

Overall accuracy of the model is

97.03%

Accuracy for the True Positive is

65.90%

6.3: Comparison of the development and test data

Comparing the overall model for train and test, it occurs that the overall accuracies of the model varies very minutely but for the true positives it varies close to 20% making the model a little ineffective.

The overall the AICs for train is 830.74 and for test data it is 173.63 which is a huge margin. AIC for the train suggests the model to be a bit weak and the AIC for test data shows the model to be strong.

However there is scope to improve the model further by eliminating multi-collinearity if any which could help in reducing the noise in the data and could make the train and test models behave in the same manner to make it a stronger model.