SAYAN MONDAL

GREAT LEARNING

FRA PROJECT

**Contents**

|  |
| --- |
| 1.Problem Statement |
| 2.Data Types, Duplicate rows, Data description |
| 3.Outlier treatment |
| 4.Missing value treatment |
| 5.Univariate analysis |
| 6.Bivariate analysis |
| 7.Logistic Regression |
| 8.LDA (Linear Discriminant Analysis) |
| 9.Random Forest Classifier |
| 10.Insights |
| 11.Recommendation |
|  |
|  |
|  |
|  |

**Tables**

|  |
| --- |
| 1.Description of data |
| 2.Outliers |
| 3.Check for missing values |
| 4.Missing values |
| 5.VIF < 5 |
| 6.Model Summary (Logistic Regression) |
| 7.Score for Logistic Regression |
| 8.Score for LDA |
| 9.Score for Random Forest Classifier |
|  |

**Figures**

|  |
| --- |
| 1.Outliers check |
| 2.Boxplot of Net worth Next Year |
| 3.Boxplot of Total Assets |
| 4.Boxplot of Total Liabilities |
| 5.Boxplot of Shareholder’s Funds |
| 6.Barplot of PAT to Net worth |
| 7.Barplot of Net Working Capital |
| 8.Barplot of Contingent Liabilities/Net worth (%) |
| 9.Heat map of Correlation |
|  |

**Problem Statement-**

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

**Data Types-**

There are 51 variables and 4256 observations in the dataset. There is one variable ‘Num’ which is integer data type and other variables are all float. Since ‘Num’ denotes only indexing of different companies, it is dropped from the dataset. So, there are 50 variables in total now.

**Duplicate Rows-**

Companies having same financial record will not add any value while building model for credit risk default. In addition there are only 2 duplicated rows. So, we just drop the duplicated rows keeping the first row.

**Data Description-**

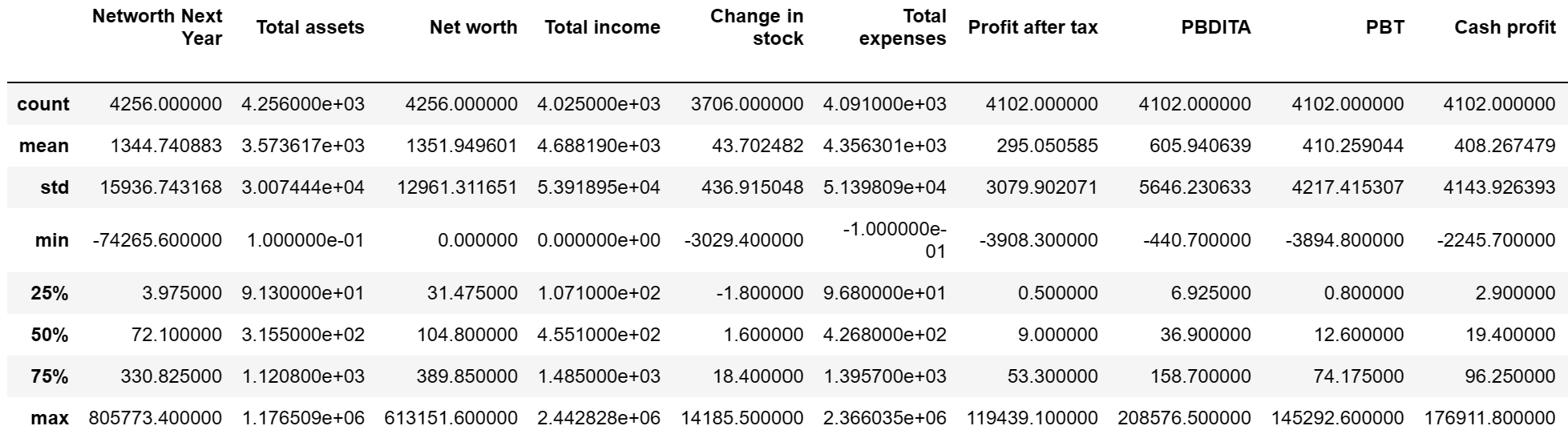


Table-1

Observing the dataset, we can simply have an overview that there are outliers present in the data. The data description of a subset of the dataframe is given.

**Outlier treatment-**

The data set contains outliers. The variables having outliers are shown below.

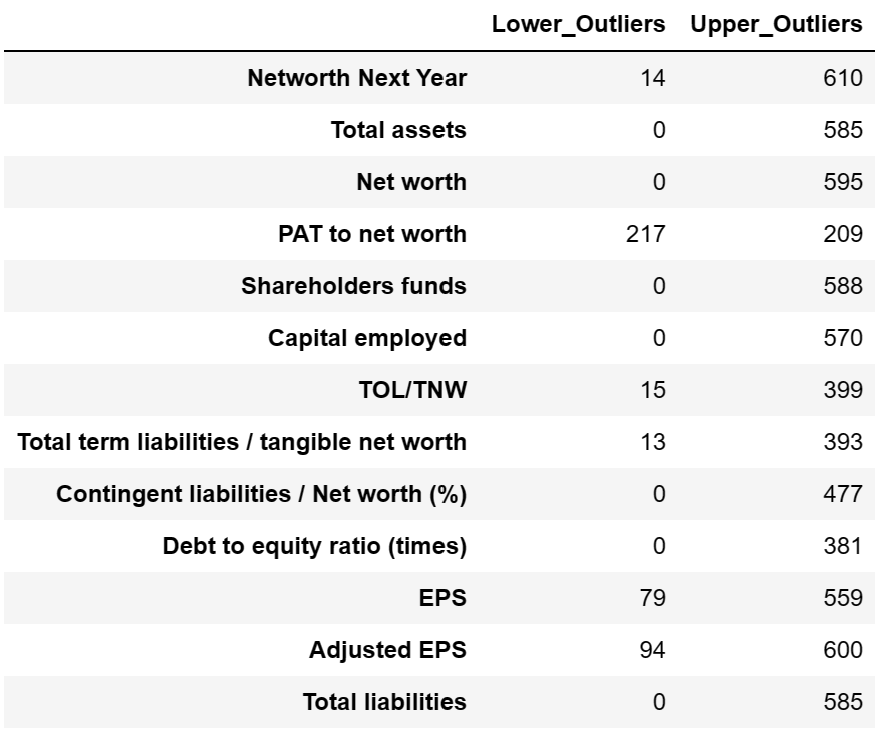


Table-2

Now to have an idea how the outliers are distributed we do a boxplot for some of the variables having outliers.

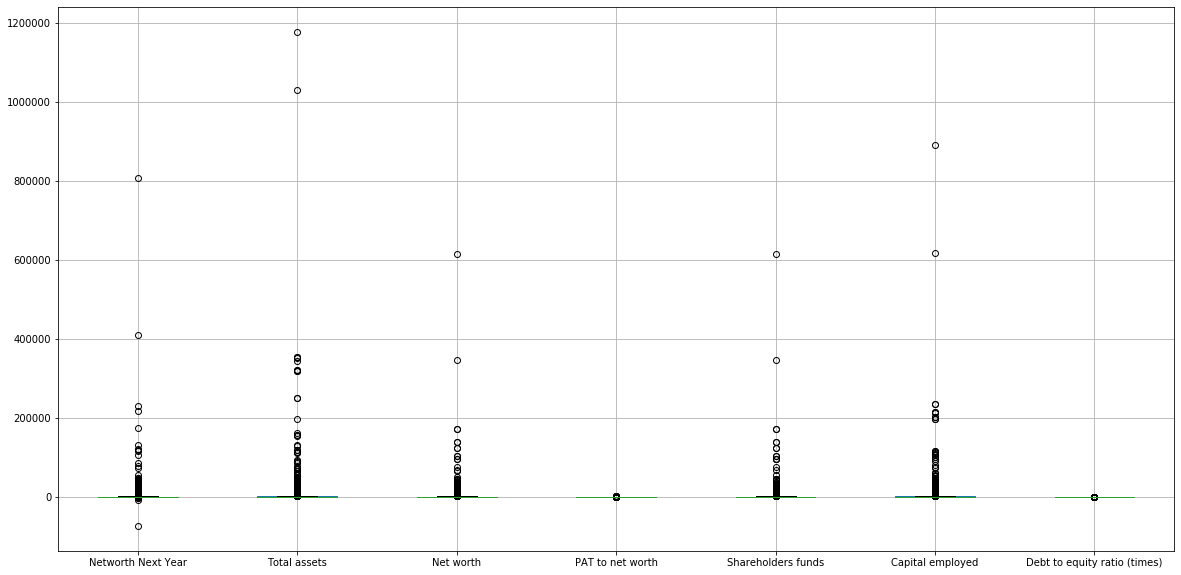


Fig-1

Here we can see the upper range and lower range for each boxplot is very close to zero. While the values of outliers are very high in comparison to upper range and lower range. So, if we replace the outliers by upper range and lower range respectively the data can be misleading. Instead, we decide to cap the upper outliers by 90th percentile and floor the lower outliers by 10th percentile. This will not remove all the outliers but a significant amount of them can be removed.

**Missing value treatment-**

Almost every variable in the data set have missing values. The table showing number of missing values for each variable is given below.

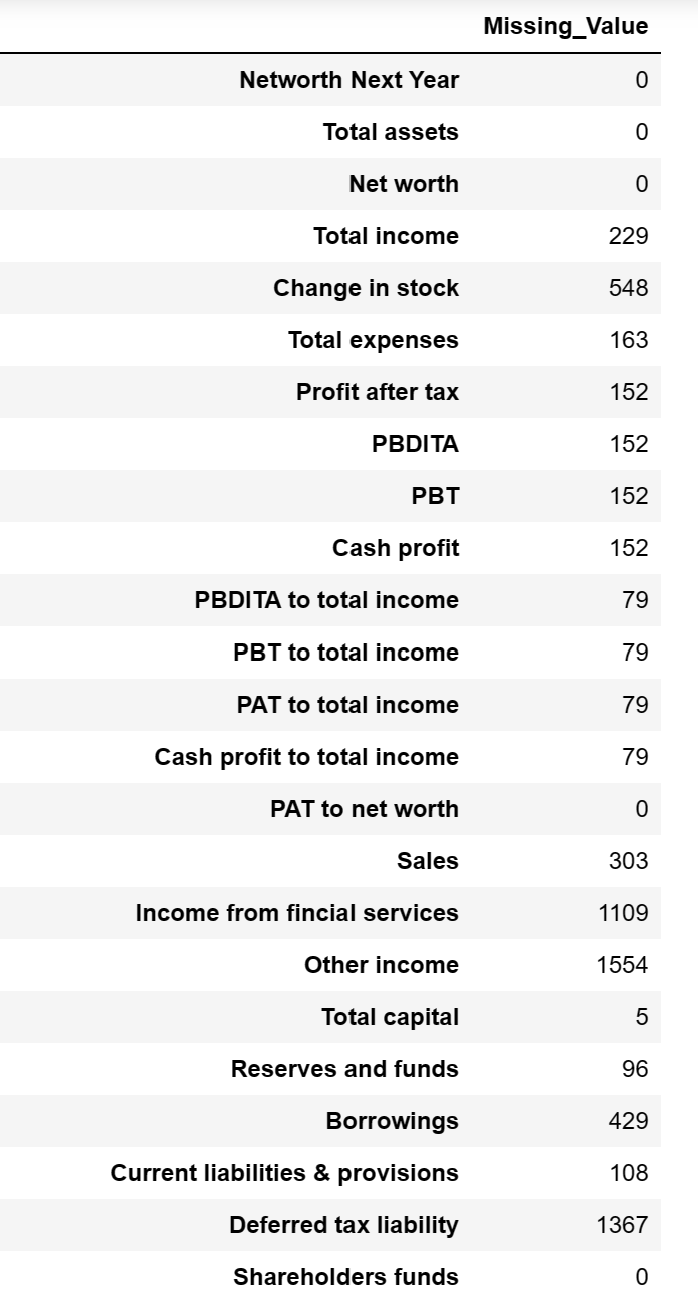
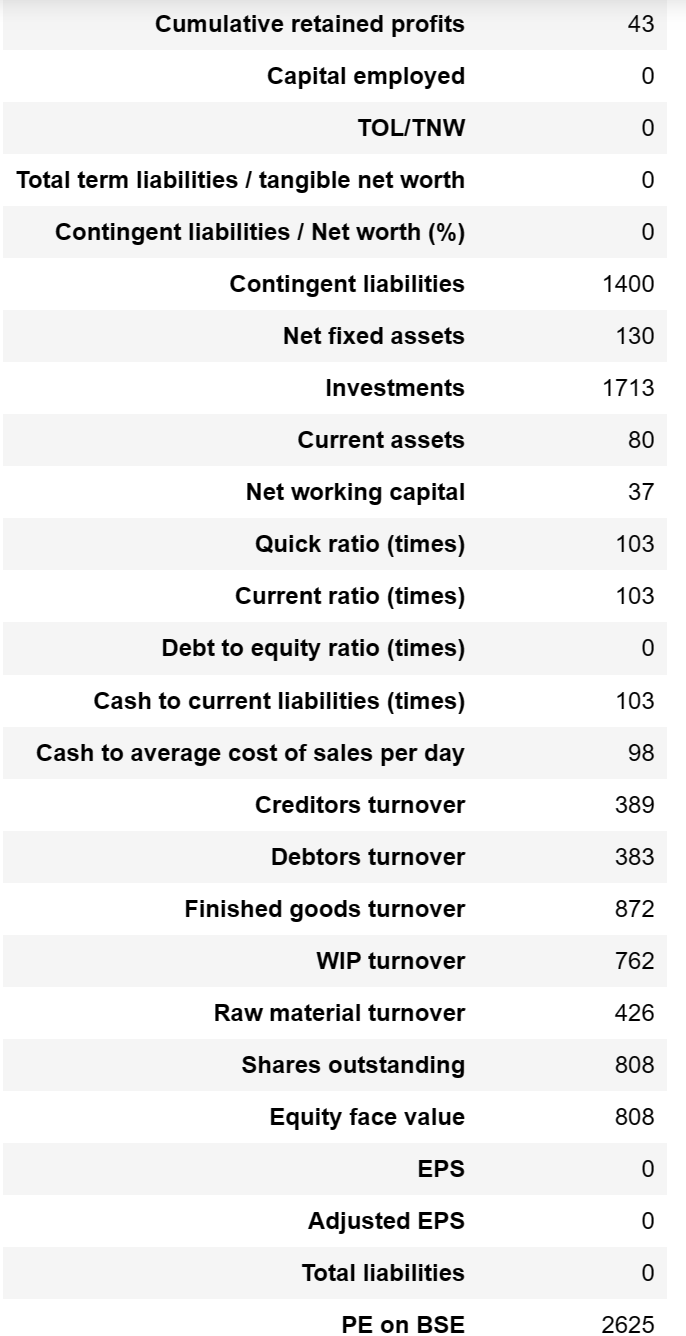
 

Table-3

Since the data contains outlier, we replace the missing values by median value.

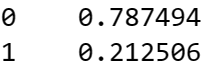
Target variable-

We create a target variable by defining default as,

* If ‘Networth Next Year’ is negative then 1.
* If ‘Networth Next Year’ is positive then 0.

We name the target variable as ‘Default’.

The distribution of default values,



The data is imbalanced with respect to the default values. We can use some sampling techniques to balance the data.

Median values of different variables according to default and non-default values.



Table-4

There is significant difference in median values of many variables for default and non-default.

**Univariate Analysis-**

We perform univariate analysis to observe how the variables vary for those companies who defaulted and those who did not.

1.Boxplot-

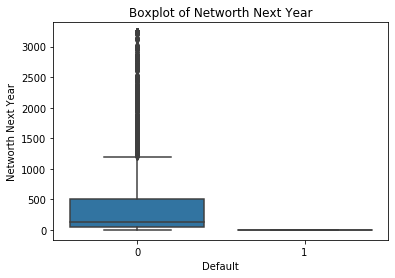


Fig-2

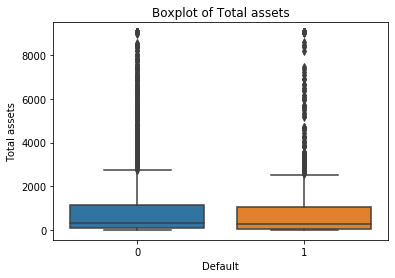


Fig-3

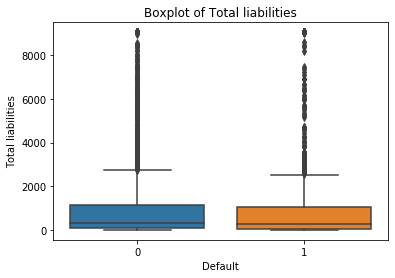


Fig-4

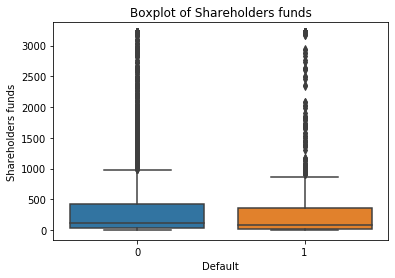


Fig-5

We can see the median value of the variable ‘Networth Next year’ is significantly different for companies who have defaulted from companies who have not defaulted. Though we can not make any differences for the variables ‘Total assets’, ‘Total liabilities’ and ‘Shareholders funds’ from their boxplots.

2.Barplot-

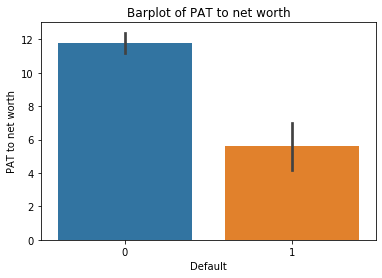


Fig-6

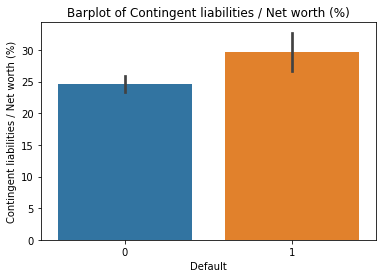


Fig-7

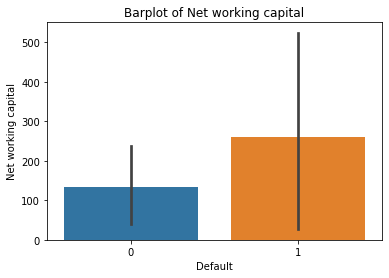


Fig-8

We observe from these barplots that the mean value of the variables shown above are well separated for defaulters and non-defaulters.

Here we see the mean value of the variable ‘Net Working Capital’ is significantly high for defaulters than non-defaulters. This indicates working inefficiency of those companies.

**Bivariate Analysis-**

We make a heatmap of correlation of the given data. It gives an idea about how the independent variables related with each other and also with the dependent variable.

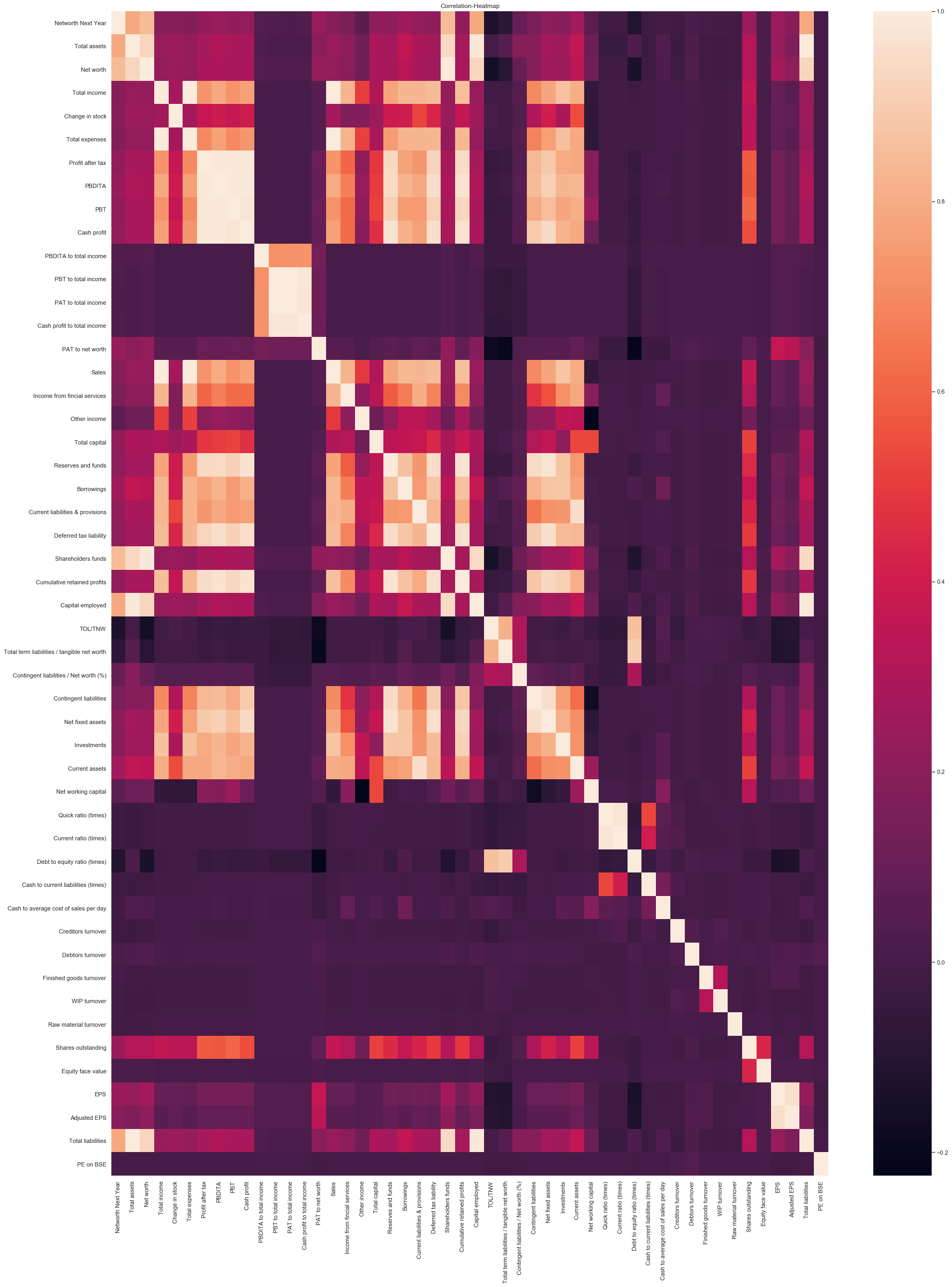


Fig-9

We observe that the there is high correlation among some independent variables. For example ‘Profit After Tax’ and ‘Investments’ is highly correlated with each other (~0.8). Also ‘Total Income’ and ‘Sales’ is highly correlated (~1.0). Also, there is instance of high correlation within ‘Total assets’ and ‘Total liabilities’. So, we can conclude that there is multicollinearity in the data.

We also observe that ‘Current assets’, ‘Investments’, ‘Capital employed’, ‘Shareholders fund’, ‘Networth’, ‘Total assets’ etc. are significantly correlated with the dependent variable ‘Networth Next Year’.

Train and Test split-

We split the data into (70:30) ratio using train-test split function with random state=1.

No.of records in the Train set is:2977

No.of records in the Test set is:1277

**Models:**

**Logistic Regression:**

1. At first, we handle the problem of multi-collinearity in the data. We use VIF criterion for this purpose. We exclude the variables from the model where VIF > 5. The variables appropriate for model building are given below.

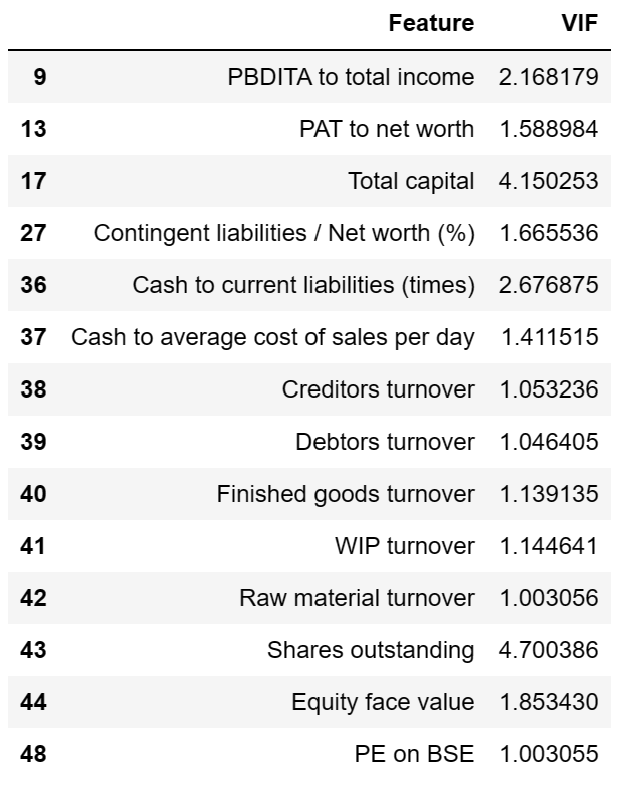


Table-5

2. Then we build a model for logistic regression with variables which have VIF <= 5. The Null Hypothesis for this model is there is no relationship between dependent variable and the independent variable. So, if p-value < 0.05 we keep the variable otherwise we remove the variable from our model. We keep doing this till p-value for all variables is less than 0.05.

The formula used for the model is given by,

f\_1='Default ~ PBDITA\_to\_total\_income + PAT\_to\_net\_worth + Total\_capital + Contingent\_liabilities\_by\_Net\_worth\_in\_Percentage + Cash\_to\_current\_liabilities\_in\_times + Cash\_to\_average\_cost\_of\_sales\_per\_day + Creditors\_turnover + Debtors\_turnover + Finished\_goods\_turnover + WIP\_turnover + Raw\_material\_turnover + Shares\_outstanding + Equity\_face\_value + PE\_on\_BSE'.

Model summary-

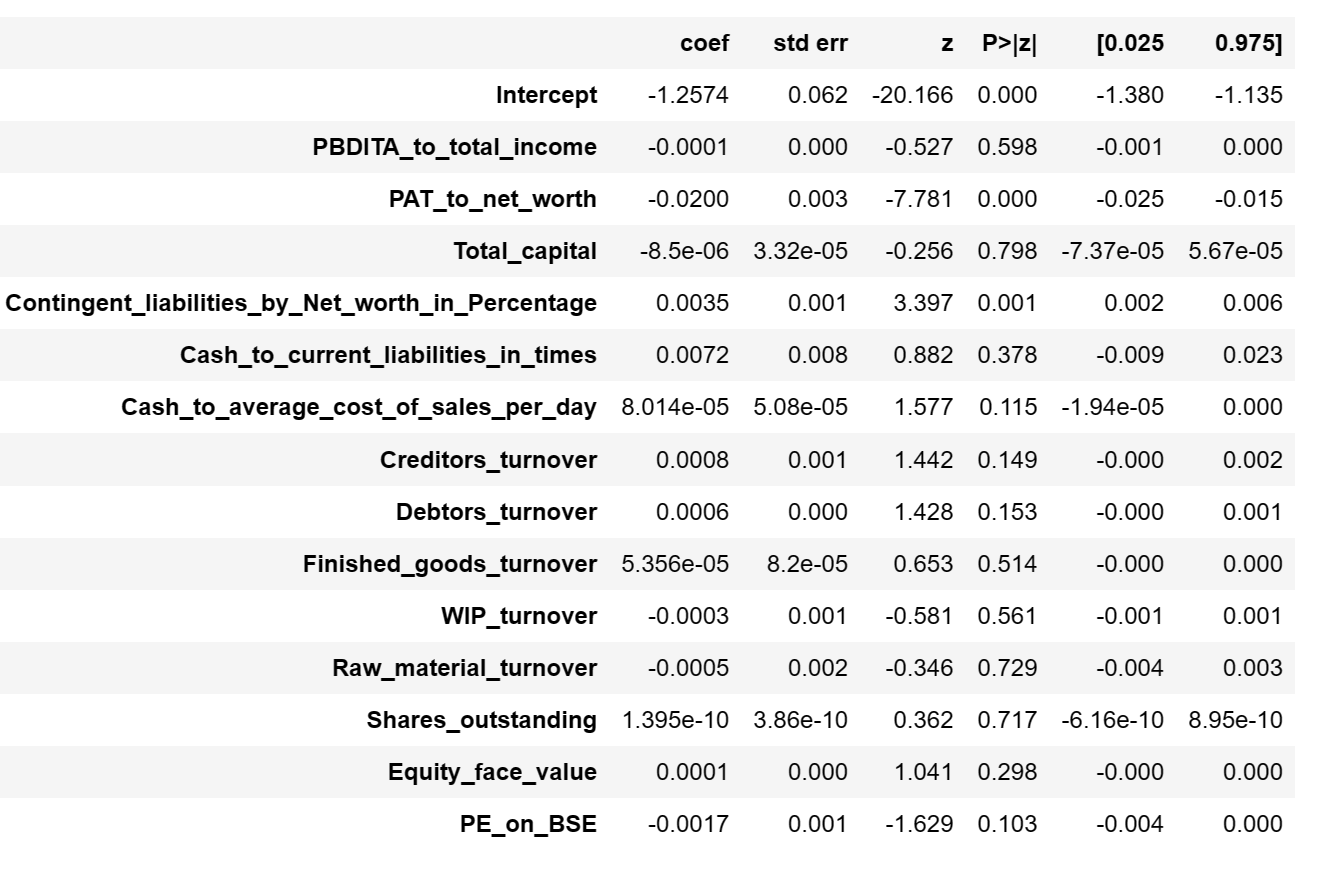


Table-6

Number of variables with p-value > 0.05 = 12.

We keep on removing variables one by one for which p-value > 0.05 and finally have the model,

f\_2='Default ~ PAT\_to\_net\_worth + Contingent\_liabilities\_by\_Net\_worth\_in\_Percentage'.

Optimum cut off-

We find that optimal threshold is 0.22 i.e.

If probability of default is > 0.22 we consider it in the default class.

If probability of default < 0.22 we consider it in the non-default class.

Performance metric-

There needs to be a balance between the following 2 opposing factors:

False Positives -This means the model predicts the company is a defaulter, but in reality, is not a defaulter. This scenario represents a Lost Opportunity for the investor, since he would not have invested in the company, thinking it was a defaulter.

False negatives -This is when the model predicts the company is not a defaulter, but in reality, actually defaults. This is a big loss to the investor, and therefore, needs to be minimized.

Thus, the most important metric would be recall and then it would be precision.

Scores-

|  |  |  |
| --- | --- | --- |
|  | Recall | Precision |
| Train Data | 0.509 | 0.268 |
| Test Data | 0.487 | 0.254 |
|  |  |  |

Table-7

So overall, this is not a great model as the recall value is not very high and precision is very low. But given the fact we could work with only two variables in this model while other variables were not adding any value in the prediction it is a decent model for the given data set.

**LDA (Linear Discriminant Analysis):**

Scores-

|  |  |  |
| --- | --- | --- |
|  | Recall | Precision |
| Train Data | 0.38 | 0.34 |
| Test Data | 0.33 | 0.32 |

Table-8

It performs very poorly compared to Logistic Regression to predict defaulting companies.

**Random Forest Classifier:**

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Grid Search

We perform a grid search on some randomly chosen parameters and thus find the best parameters for modelling.

Max\_depth: [3, 5, 7]

Min\_samples\_leaf: [5, 10, 15]

Min\_samples\_split: [15, 30, 45]

N\_estimators: [25, 50]

Scores-

|  |  |  |
| --- | --- | --- |
|  | Recall | Precision |
| Train Data | 0.14 | 0.76 |
| Test Data | 0.13 | 0.57 |

Table-9

This model also performs very poorly. If consider only the precision value, the model over fitting.

So, the best model for prediction of defaulting companies is Logistic Regression.

**Insights:**

* The most important features for prediction whether a company will default or not are

‘PAT\_to\_net\_worth’ and ‘Contingent\_liabilities\_by\_Net\_worth\_in\_Percentage'.

* Companies not defaulting have almost twice mean PAT\_to\_net\_worth value in comparison to companies who are defaulting.
* The mean percentage of Contingent liabilities by net worth of companies not defaulting is significantly less than those are defaulting.

**Recommendations:**

* We can avoid providing big loans to companies who are prone to default.
* We can charge high rate of interest for loans from those companies who have a high chance to default.
* The companies who are predicted as non-defaulter can be pushed to take big loans.