# 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 that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

### Data set :

| Co_Code | Co_Name | Networth Next | Equity Paid Up | Networth | Capital Employe | Total Debt | Gross Block | Net Working Ca | Current Assets | Current Liabiliti | Total Assets/Lia |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 16974 | Hind.Cables | -8021.6 | 419.36 | -7,027.48 | -1,007.24 | 5,936.03 | 474.3 | -1,076.34 | 40.5 | 1,116.85 | 109.6 |
| 21214 | Tata Tele. Mah. | -3986.19 | 1,954.93 | -2,968.08 | 4,458.20 | 7,410.18 | 9,070.86 | -1,098.88 | 486.86 | 1,585.74 | 6,043.94 |
| 14852 | ABG Shipyard | -3192.58 | 53.84 | 506.86 | 7,714.68 | 6,944.54 | 1,281.54 | 4,496.25 | 9,097.64 | 4,601.39 | 12,316.07 |
| 2439 | GTL | -3054.51 | 157.3 | -623.49 | 2,353.88 | 2,326.05 | 1,033.69 | -2,612.42 | 1,034.12 | 3,646.54 | 6,000.42 |
| 23505 | Bharati Defence | -2967.36 | 50.3 | -1,070.83 | 4,675.33 | 5,740.90 | 1,084.20 | 1,836.23 | 4,685.81 | 2,849.58 | 7,524.91 |
| 2484 | Usha Ispat | -2519.4 | 179.35 | -2,519.39 | -1,824.75 | 694.64 | 0.02 | -1,843.74 | 0 | 1,843.74 | 18.99 |
| 23633 | Hanung Toys | -2125.05 | 30.82 | -1,031.57 | 1,536.08 | 2,567.65 | 949.98 | 804.82 | 834.86 | 30.04 | 1,566.12 |
| 3226 | K S Oils | -2100.56 | 45.92 | -1,945.45 | 979.13 | 2,664.04 | 920.67 | 263.95 | 705.76 | 441.81 | 1,420.94 |
| 1541 | Quadrant Tele. | -1695.75 | 61.23 | -1,560.94 | -613.79 | 597.82 | 1,700.27 | -1,121.96 | 117.67 | 1,239.63 | 625.85 |
| 2334 | ITI | -1677.18 | 288 | -1,947.85 | 86.35 | 1,220.83 | 1,329.82 | -390.53 | 2,536.78 | 2,927.31 | 3,013.66 |
| 430 | Parasram. Synth | -1403.7 | 66.28 | -1,400.79 | -220.49 | 1,158.34 | 399.8 | -611.84 | 22.29 | 634.12 | 413.62 |
| 4169 | Electrotherm(I) | -1243.33 | 11.48 | -1,025.02 | 2,057.86 | 3,063.30 | 2,141.72 | 565.64 | 868.34 | 302.7 | 2,360.56 |
| 5926 | ICSA (India) | -1138.48 | 9.63 | -1,117.68 | 39.83 | 1,157.04 | 236.5 | -147.75 | 209.05 | 356.8 | 396.63 |
| 3367 | SpiceJet | -1038.86 | 599.45 | -1,085.93 | 602.19 | 1,477.34 | 2,113.77 | -1,456.25 | 548.14 | 2,004.40 | 2,606.59 |
| 2302 | Hind.Organ.Che | -981.21 | 67.27 | -804.16 | -32.81 | 402.03 | 639.79 | -197.77 | 131.31 | 329.07 | 296.26 |
| 4397 | Ricoh India | -949.14 | 39.77 | 168.6 | 885.37 | 701.52 | 144.13 | 766.7 | 1,176.47 | 409.77 | 1,295.14 |
| 24936 | Zylog Systems | -854.42 | 29.5 | -115.42 | 807.85 | 919.71 | 908.28 | 300.83 | 501.39 | 200.56 | 1,008.41 |
| 24619 | Jai Balaji Inds. | -838.28 | 73.78 | -180.81 | 2,799.73 | 2,980.54 | 2,490.95 | 789.42 | 2,181.38 | 1,391.96 | 4,191.69 |
| 6927 | Mackinnon Mac | -834.09 | 0.25 | -833.53 | -7.54 | 825.61 | 1.24 | -8.45 | 6.1 | 14.56 | 7.02 |

We are provided with the above data set of 3586 rows and 67 columns. Of the above columns, One column is integer data type and four columns are of float data type and remaining columns are object data type.

The columns **'Book Value (Adj.) (Unit Curr)', 'Inventory Velocity (Days)','Current Ratio[Latest]',' Fixed Assets Ratio[Latest]','Inventory Ratio[Latest]',' Debtors Ratio[Latest]',' Total Asset Turnover Ratio[Latest]','Interest Cover Ratio[Latest]',' Inventory Velocity (Days)'** in the given dataset have presence of total 118 **Null values**.

```
df.isnull().sum().sum()
```

```
118
```

**Descriptive statistics for the dataset:**

| | Networth Next Year | Equity Paid Up | Networth | Capital Employed | Total Debt | Gross Block | Net Working Capital | Current Assets | Current Liabilities and Provisions | Total Assets/Liabilities | ... | PBIDTM (%) [Latest] | PBITM (%) [Latest] | PBDTM (% [Latest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | 3586.00 | ... | 3585.00 | 3585.00 | 3585.0 |
| **mean** | 725.05 | 62.97 | 649.75 | 2799.61 | 1994.82 | 594.18 | 410.81 | 1960.35 | 391.99 | 1778.45 | ... | -51.16 | -109.21 | -311.5 |
| **std** | 4769.68 | 778.76 | 4091.99 | 26975.14 | 23652.84 | 4871.55 | 6301.22 | 22577.57 | 2675.00 | 11437.57 | ... | 1795.13 | 3057.64 | 10921.5 |
| **min** | -8021.60 | 0.00 | -7027.48 | -1824.75 | -0.72 | -41.19 | -13162.42 | -0.91 | -0.23 | -4.51 | ... | -78870.45 | -141600.00 | -590500.0 |
| **25%** | 3.98 | 3.75 | 3.89 | 7.60 | 0.03 | 0.57 | 0.94 | 4.00 | 0.73 | 10.55 | ... | 0.00 | 0.00 | 0.0 |
| **50%** | 19.02 | 8.29 | 18.58 | 39.09 | 7.49 | 15.87 | 10.14 | 24.54 | 9.23 | 52.01 | ... | 8.07 | 5.23 | 4.6 |
| **75%** | 123.80 | 19.52 | 117.30 | 226.61 | 72.35 | 131.90 | 61.17 | 135.28 | 65.65 | 310.54 | ... | 18.99 | 14.29 | 14.1 |
| **max** | 111729.10 | 42263.46 | 81657.35 | 714001.25 | 652823.81 | 128477.59 | 223257.56 | 721166.00 | 83232.98 | 254737.22 | ... | 19233.33 | 19195.70 | 15640.0 |

8 rows × 65 columns

We can see the descriptive statistics of the dataset from above table.

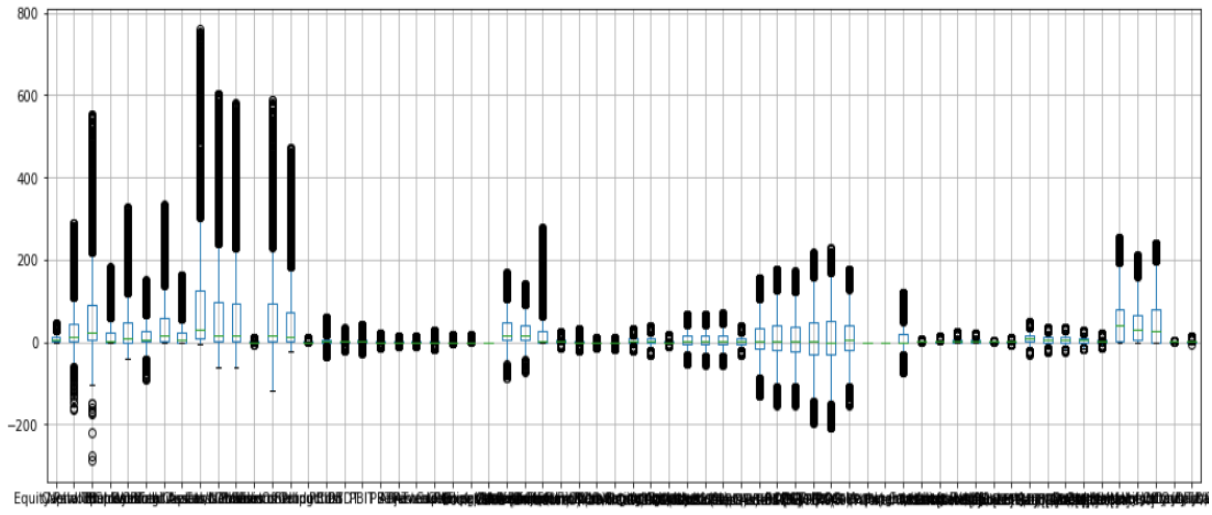
The given dataset has no duplicate rows

### 1.1) Outlier Treatment

Almost all the variables in the given dataset are having outliers. There is a Total of **41355** outlier values in the given dataset.

```
Equity Paid Up                    448
Networth                          650
Capital Employed                  596
Total Debt                        583
Gross Block                       540
                                 ...
Debtors Velocity (Days)           398
Creditors Velocity (Days)         391
Inventory Velocity (Days)         262
Value of Output/Total Assets      150
Value of Output/Gross Block       481
Length: 64, dtype: int64
```

BoxPlot for all variables of the dataset:



So, as part of the outlier treatment, we are replacing all the outliers values, which are above **UL = Q3 + 1.5*IQR** and below **LL = Q1 - 1.5*IQR** with null values.

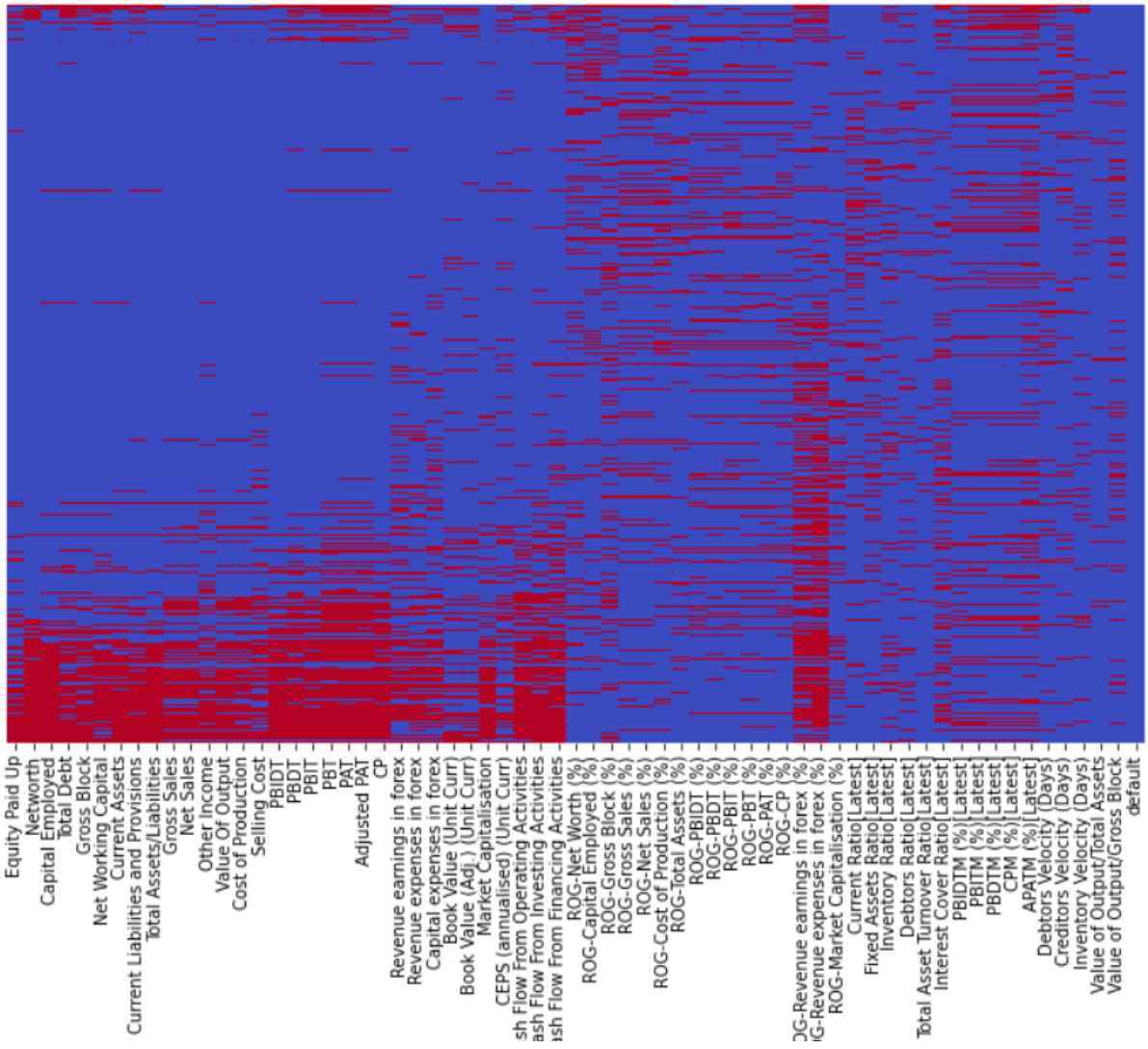
```
Company_X[((Company_X > UL) | (Company_X < LL))]= np.nan
```

Later, we are imputing null values with the help of KNN Imputer,which imputes the values by comparing with row wise nearest variables.

**1.2) Missing value Treatment.**

The given dataset have presence of total 118 **Null values** in columns **'Book Value (Adj.) (Unit Curr)', 'Inventory Velocity (Days)','Current Ratio[Latest]',' Fixed Assets Ratio[Latest]','Inventory Ratio[Latest]',' Debtors Ratio[Latest]',' Total Asset Turnover Ratio[Latest]','Interest Cover Ratio[Latest]',' Inventory Velocity (Days)'**

As we are converting all outlier values into null values. The total count of null values has become **41473**




All the redcolor highlighted portion shows the null values present in the respective columns.

```
ROG-Revenue expenses in forex (%)      0.45
ROG-Revenue earnings in forex (%)      0.37
Cash Flow From Financing Activities    0.28
PAT                                    0.27
Adjusted PAT                           0.27
                                       ...
Debtors Ratio[Latest]                  0.10
Inventory Velocity (Days)              0.10
Total Asset Turnover Ratio[Latest]     0.06
Value of Output/Total Assets           0.04
default                                0.00
Length: 65, dtype: float64
```

Variable '**ROG-Revenue expenses in forex (%)'** is having **highest** percentage(**45%**) of null values of the total values in that column.
These null values are treated with the help of KNN Imputer,which imputes the values by comparing with row wise nearest variables.

```
imputer = KNNImputer(n_neighbors=10)
```

The final result after imputing all the null values :

```
Equity Paid Up                  0
Networth                        0
Capital Employed                0
Total Debt                      0
Gross Block                     0
                               ..
Creditors Velocity (Days)       0
Inventory Velocity (Days)       0
Value of Output/Total Assets    0
Value of Output/Gross Block     0
default                         0
Length: 63, dtype: int64
```

### 1.3) Transform Target variable into 0 and 1

We are creating a Target variable which takes the value of 1 when '**Networth Next Year'** is negative & 0 when net worth is positive.

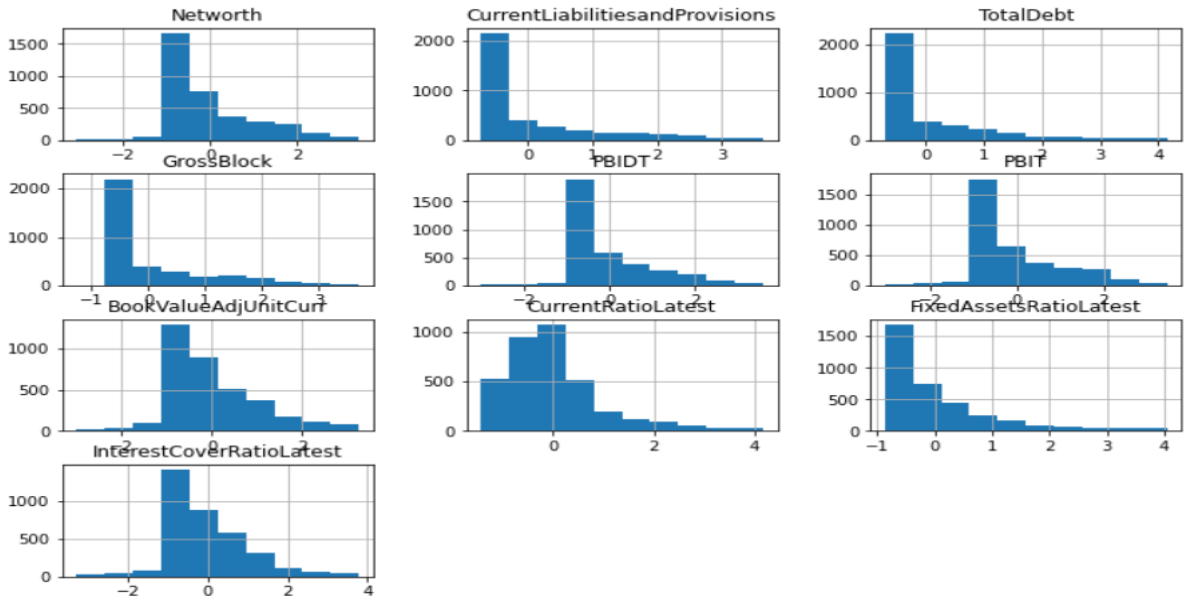
```
df['default'] = np.where((df['Networth Next Year'] > 0), 0, 1)
```

In the target variable we are having 89% of values as '0' and 11% of values as '1'

```
0   0.89
1   0.11
Name: default, dtype: float64
```

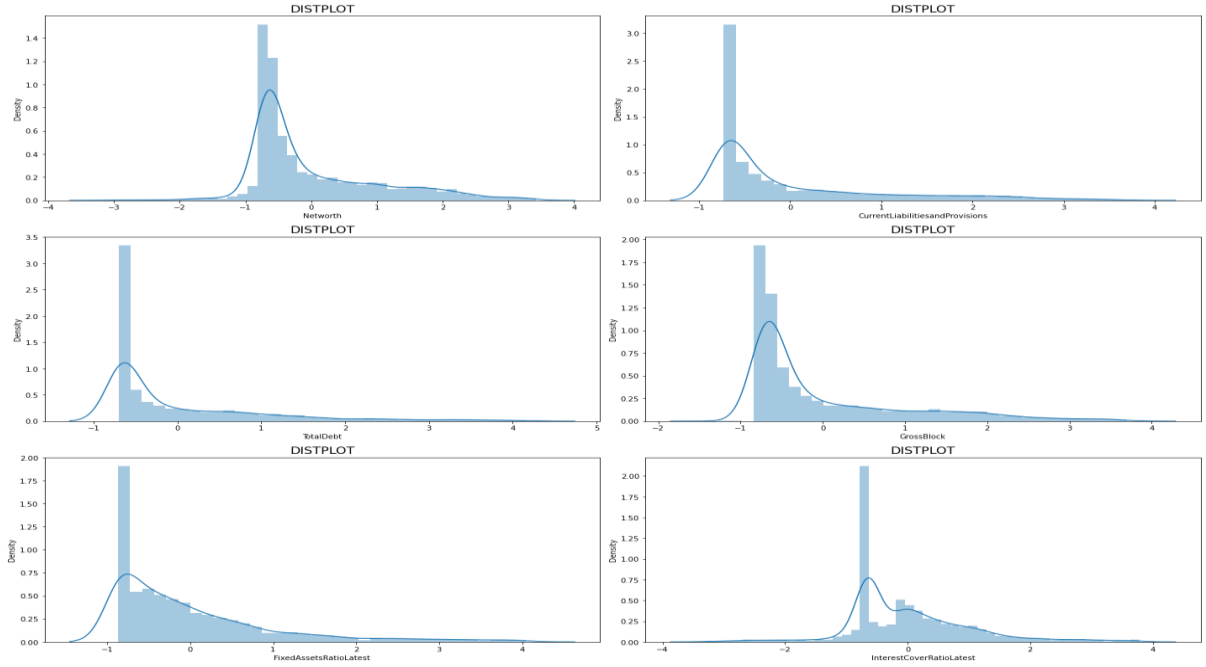
**1.4) Univariate (4 marks) & Bivariate ( 6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)**

After checking variable importance and multicollinearity, we have found the following variables as significant in the model building
**'Networth','CurrentLiabilitiesandProvisions','TotalDebt','GrossBlock','PBIDT','PBIT','BookValueAdjUnitCurr','CurrentRatioLatest','FixedAssetsRatioLatest','InterestCoverRatioLatest'**

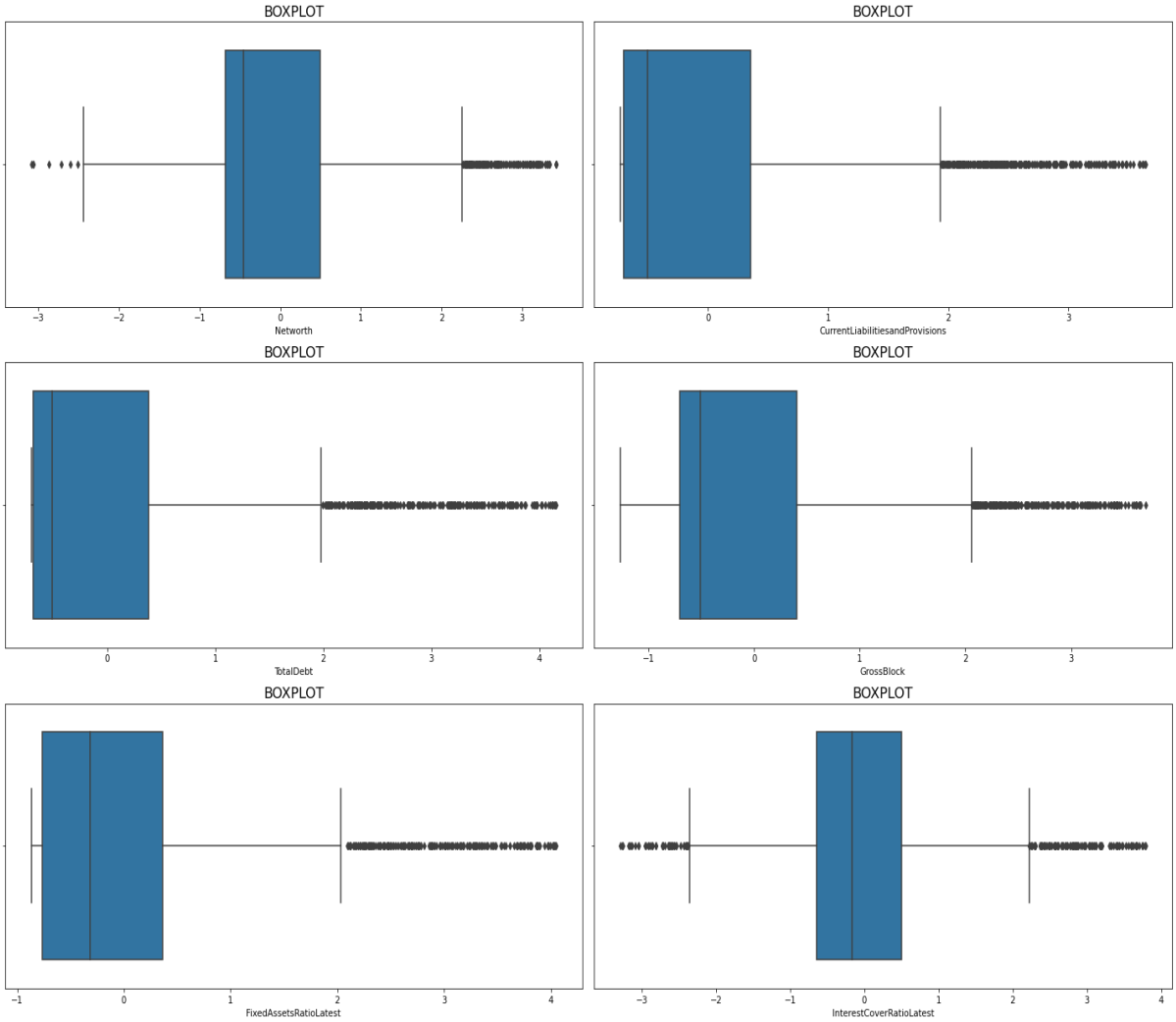



All the above variables are skewed and are not symmetrical. Variable**'TotalDebt'** is having **highest positive** skewness of value **1.86** and Variable **'InterestCoverRatioLatest'** has **lowest positive** skewness of **0.84.**

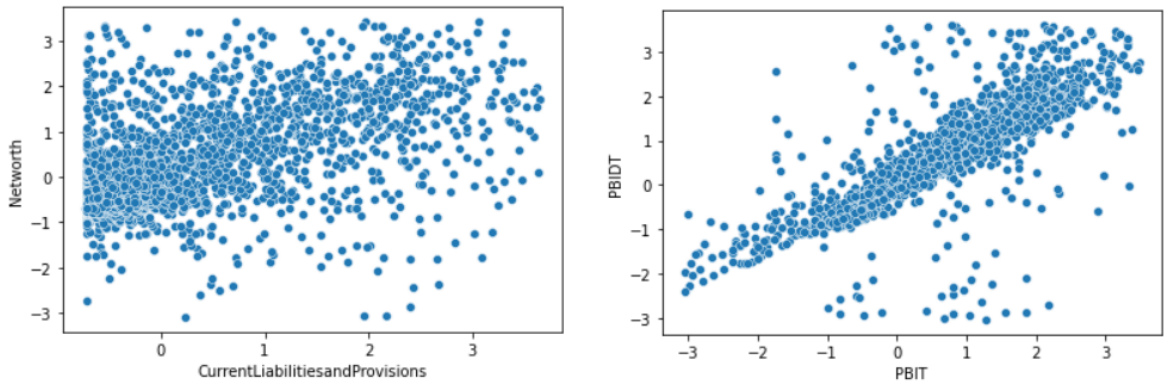
**Dist Plot:**



**Box Plot:**

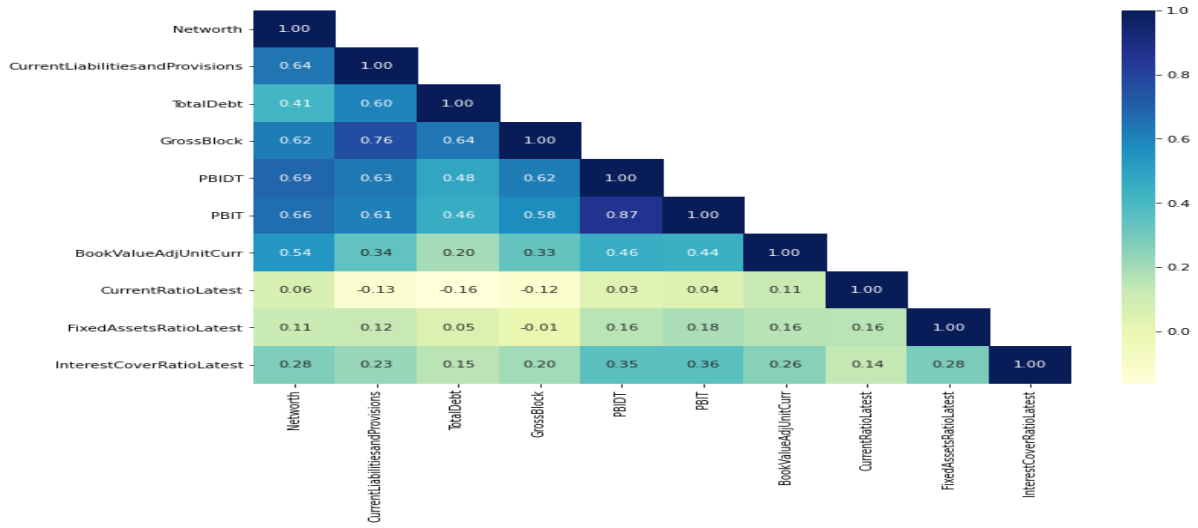



From above plot, we can see that all the variables are having outliers.



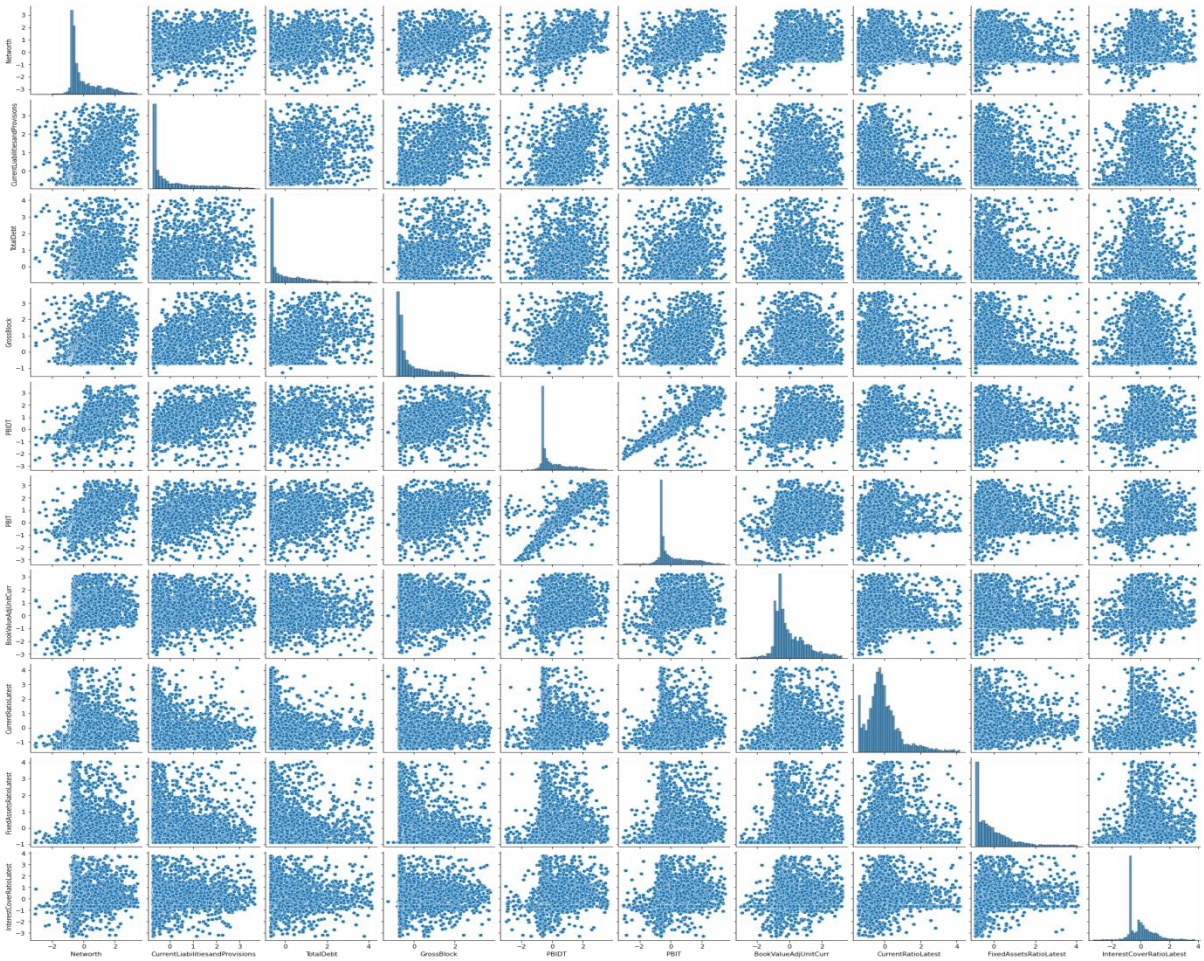

We can see there is high linear correlation between variables ‘Networth’ & CurrentLiabilitiesandProvisions and also between variables ‘PBIDT’ & ‘PBIT’.

**Heat Map:**




From the above map , we can see that many columns are co-related to each other and There is highest positive correlation(0.87) between the variables **'PBIT'** and **'PBIDT'** and lowest negative correlation(-0.16) between the variables **'CurrentRatioLatest'** and **'GrossBlock'**

**PairPlot:**



In the above plot scatter diagrams are plotted for all the columns in the dataset. From the visual representation , we can understand the degree of correlation between any two columns of the given dataset.

We cannot to infer clear relationship between most of the different variables but there is notable relationship between certain variables.

Variables **'PBIT' and 'PBIDT'** shows positive linear correlation with Variable **'Networth'**and Variables **'PBIT' and 'PBIDT'** shows very positive linear correlation between them.

### 1.5) Train Test Split

We have Split the data into Train and Test dataset in a ratio of 67:33.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 42)
```

```
X_train.head()
```

| | Networth | TotalDebt | GrossBlock | CurrentLiabilitiesandProvisions | CostofProduction | PBIDT | PBIT | BookValueAdjUnitCurr | CurrentRatioLatest |
|---|---|---|---|---|---|---|---|---|---|
| **662** | -0.70 | -0.68 | -0.70 | -0.72 | -0.72 | -0.63 | -0.61 | -0.62 | -0.37 |
| **1373** | -0.60 | -0.68 | -0.70 | -0.71 | -0.72 | -0.61 | -0.59 | 0.50 | 0.35 |
| **3268** | 0.26 | -0.53 | -0.65 | -0.35 | -0.58 | -0.89 | -0.70 | 0.31 | -0.66 |
| **3246** | 1.88 | 0.57 | 1.84 | 2.17 | 0.21 | 2.43 | 1.90 | 2.11 | -0.18 |
| **1456** | -0.62 | -0.50 | -0.61 | -0.70 | -0.21 | -0.48 | -0.44 | -0.59 | 0.39 |

```
X_test.head()
```

| | Networth | TotalDebt | GrossBlock | CurrentLiabilitiesandProvisions | CostofProduction | PBIDT | PBIT | BookValueAdjUnitCurr | CurrentRatioLatest |
|---|---|---|---|---|---|---|---|---|---|
| **3163** | 1.65 | 0.47 | 0.60 | 0.92 | 1.55 | 1.69 | 1.42 | 0.04 | -0.32 |
| **3133** | 1.89 | 0.83 | 0.20 | 3.23 | 0.83 | 1.20 | 0.88 | 0.36 | -0.69 |
| **937** | -0.68 | -0.68 | -0.71 | -0.72 | -0.72 | -0.69 | -0.68 | -0.87 | -0.06 |
| **196** | -0.89 | -0.28 | -0.36 | -0.40 | -0.56 | -0.58 | -0.64 | -1.69 | -1.16 |
| **2852** | 2.07 | 0.75 | 3.11 | 0.28 | 1.68 | 0.58 | -1.74 | -0.67 | 0.14 |

```
y_train.head()
```

```
662     0
1373    0
3268    0
3246    0
1456    0
Name: default, dtype: int64
```

```
y_test.head()
```

```
3163    0
3133    0
937     0
196     1
2852    0
Name: default, dtype: int64
```

**1.6) Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach.**

Logit Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | default | **No. Observations:** | 2402 |
| **Model:** | Logit | **Df Residuals:** | 2391 |
| **Method:** | MLE | **Df Model:** | 10 |
| **Date:** | Sun, 21 Mar 2021 | **Pseudo R-squ.:** | 0.4938 |
| **Time:** | 17:12:07 | **Log-Likelihood:** | -400.58 |
| **converged:** | True | **LL-Null:** | -791.34 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.930e-161 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -4.3826 | 0.212 | -20.635 | 0.000 | -4.799 | -3.966 |
| **Networth** | -1.2002 | 0.221 | -5.427 | 0.000 | -1.634 | -0.767 |
| **CurrentLiabilitiesandProvisions** | 0.9109 | 0.171 | 5.320 | 0.000 | 0.575 | 1.246 |
| **TotalDebt** | 0.2938 | 0.143 | 2.049 | 0.040 | 0.013 | 0.575 |
| **GrossBlock** | 0.6037 | 0.198 | 3.054 | 0.002 | 0.216 | 0.991 |
| **PBIDT** | -1.2122 | 0.231 | -5.256 | 0.000 | -1.664 | -0.760 |
| **PBIT** | 0.4690 | 0.193 | 2.424 | 0.015 | 0.090 | 0.848 |
| **BookValueAdjUnitCurr** | -2.0044 | 0.216 | -9.264 | 0.000 | -2.428 | -1.580 |
| **CurrentRatioLatest** | -1.0395 | 0.148 | -7.004 | 0.000 | -1.330 | -0.749 |
| **FixedAssetsRatioLatest** | -0.4637 | 0.171 | -2.719 | 0.007 | -0.798 | -0.129 |
| **InterestCoverRatioLatest** | -0.5656 | 0.134 | -4.217 | 0.000 | -0.829 | -0.303 |

From the table, Of the available variables we have got the following variables as the most significant (p<0.05) in predicting the defaulters.

```
f_3 = 'default ~ Networth + CurrentLiabilitiesandProvisions + TotalDebt + GrossBlock + PBIDT + PBIT + BookValueAdjUnitCurr + Curr
```

```
model_3 = SM.logit(formula = f_3, data=Default_train).fit()
```

```
Optimization terminated successfully.
         Current function value: 0.166769
         Iterations 9
```

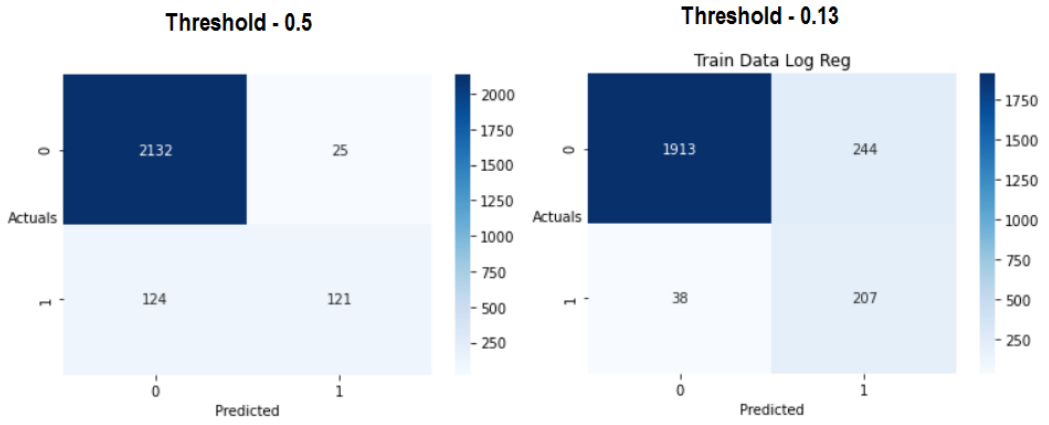
**'Networth','CurrentLiabilitiesandProvisions','TotalDebt','GrossBlock','PBIDT','PBIT','BookValueAdjUnitCurr','CurrentRatioLatest','FixedAssetsRatioLatest','InterestCoverRatioLatest'**

With the default threshold of 0.5, we are getting a recall of the 0.49. So, lets check for optimal threshold which helps better in separating ‘0’s and ‘1’s in the default.

Optimum threshold:

```
optimal_idx = np.argmax(tpr - fpr)
optimal_threshold = thresholds[optimal_idx]
optimal_threshold
```

```
0.12810242710313494
```




```
121/(124+121)
```

**Recall -** `0.49387755102040815`

```
207/(207+38)
```

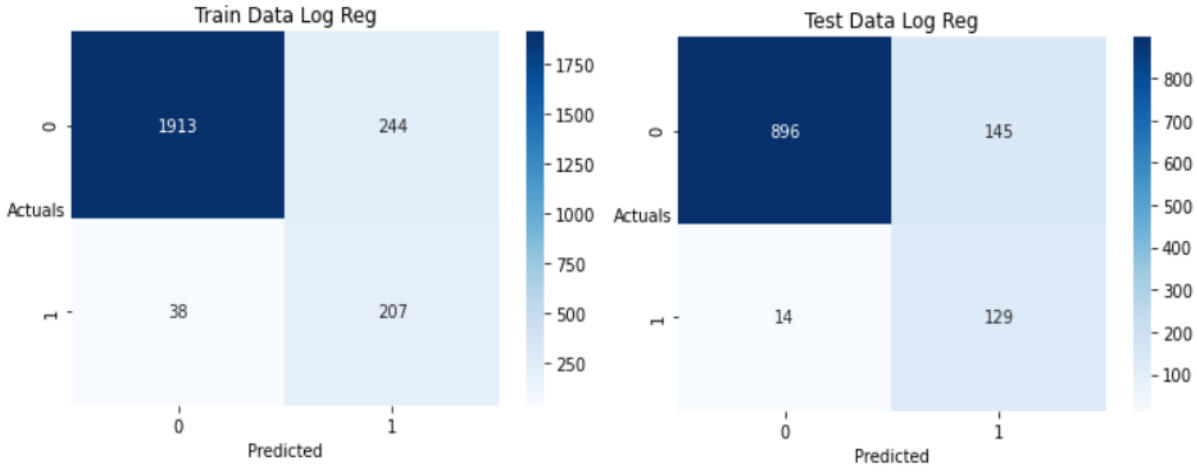
**Recall -** `0.8448979591836735`

Therefore with the threshold of 0.13 , we are able to achieve a recall of 0.844.

### 1.7) Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model.

Comparing the performance metrics on the test and train data of the model,

**Confusion matrix :**



**Classification Report:**

**Train data :**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.981 | 0.887 | 0.931 | 2157 |
| 1 | 0.459 | 0.845 | 0.595 | 245 |
| accuracy | | | 0.883 | 2402 |
| macro avg | 0.720 | 0.866 | 0.763 | 2402 |
| weighted avg | 0.927 | 0.883 | 0.897 | 2402 |

**Test data:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.985 | 0.861 | 0.919 | 1041 |
| 1 | 0.471 | 0.902 | 0.619 | 143 |
| accuracy | | | 0.866 | 1184 |
| macro avg | 0.728 | 0.881 | 0.769 | 1184 |
| weighted avg | 0.923 | 0.866 | 0.882 | 1184 |

The Model has got high recall and accuracy on both the train data and the test data. We know that logistic regression are the most widely used statistical methods for analyzing categorical outcome variable. It is assumed that logistic regression is the more flexible and more robust method in case of violations of the assumptions also logistic regression is preferred when the dependent variable is dichotomous.Therefore we can use logistic regression model in predicting whether a company will go for the default or not.

Due to the importance of understanding and managing the risks in volatile business domains, it is required to find an effective aid in making decisions. The results from model shows that the above algorithm is a promising opportunity in predicting whether a company will go for the default or not through the cause and effect relationship between the independent and dependent variables of the given dataset.

The above model will be helpful in predicting the dependent variables through the independent variables by assigning the probability of company going for the default to the every predictor variable to give the best predictive/dependent variable.

As per predictions of the model, we have got the following coefficients for the independent variables of the given dataset.

The coefficient of the different attributes of the given dataset are:

The coefficient for Networth is -1.2002
The coefficient for CurrentLiabilitiesandProvisions is 0.9109

The coefficient for TotalDebt is 0.2938
The coefficient for GrossBlock is 0.6037
The coefficient for PBIDT is -1.2122
The coefficient for PBIT is 0.469
The coefficient for BookValueAdjUnitCurr is -2.0044
The coefficient for CurrentRatioLatest is -1.039
The coefficient for FixedAssetsRatioLatest is -0.4637
The coefficient for InterestCoverRatioLatest is -0.5656

'BookValueAdjUnitCurr' is the most important feature among all the features of the dataset.

We must look about company based on the feature importance to get the better results in predicting whether an company will go for the default or not.

So, The Overall analysis of given dataset definitely helped to get insights that would help in predicting about the company.