# FRA Project Milestone 1

Ashwani Balyan
Batch- G4
balyanashwani@gmail.com

## Table Of Content

# 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.

# About Data

### Data Info

- Total 67 columns
- Total 3586 rows
- Data types for all columns are as per expectations
- Null values present in few columns

### Checking duplicate rows

- No duplicate columns are present

### Checking unique valued columns(IDs)

- Columns 'Co_Code' and 'Co_Name' are unique valued(IDs) columns. Dropping these for further analysis

# 1.1 Outliers Treatment

### Outliers Check

## Boxplots

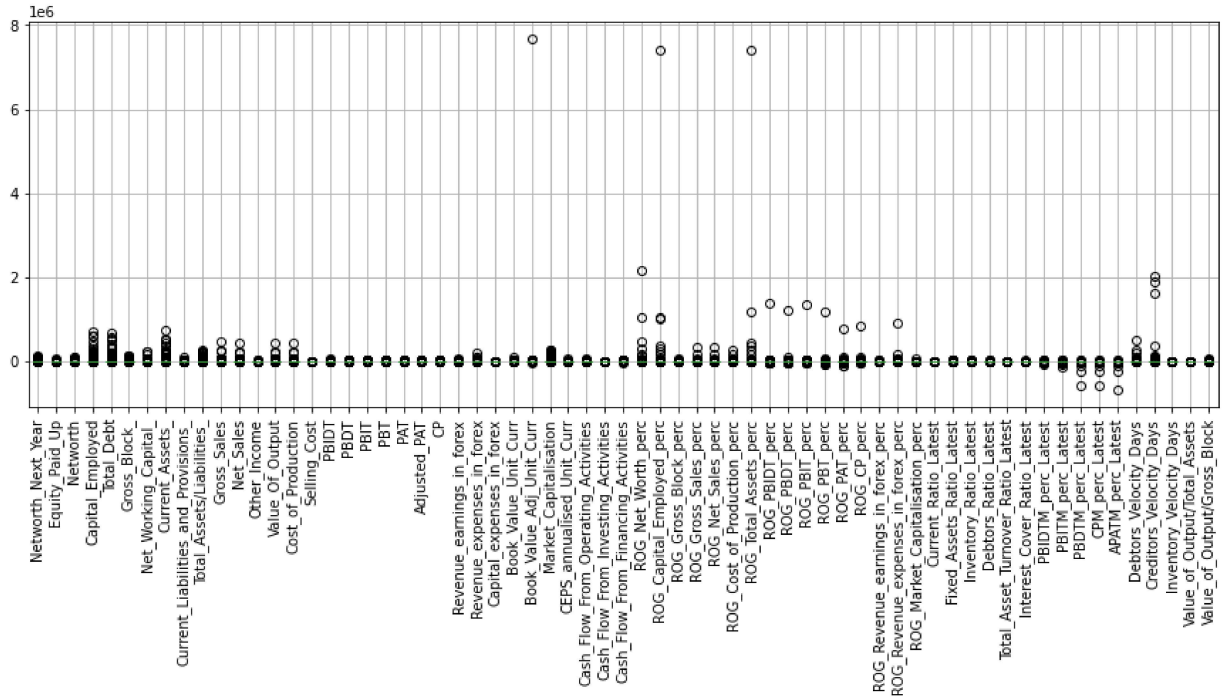


Fig.1

- There are outliers in all the columns

## Treating Outliers

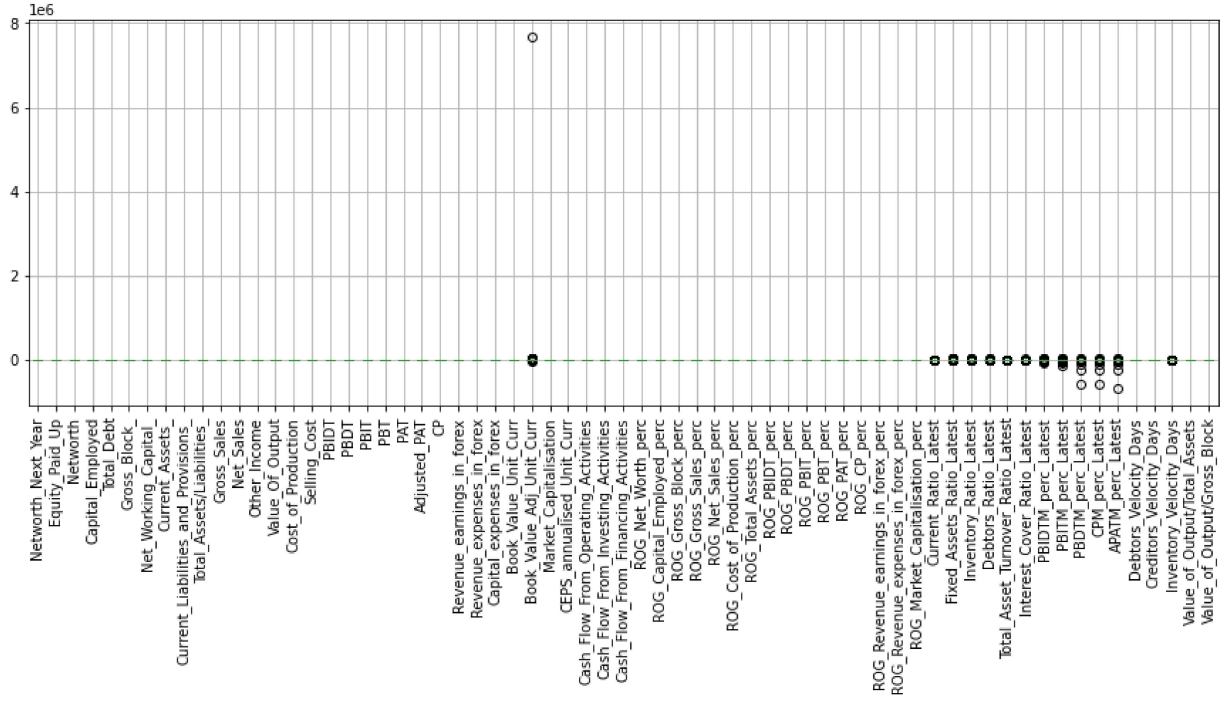


Fig.2

# 1.2 Missing Value Treatment

## Columns-wise check

| | 0 |
|---|---|
| Inventory_Velocity_Days | 103 |
| Book_Value_Adj_Unit_Curr | 4 |
| Interest_Cover_Ratio_Latest | 1 |
| PBITM_perc_Latest | 1 |
| Fixed_Assets_Ratio_Latest | 1 |
| Inventory_Ratio_Latest | 1 |
| Debtors_Ratio_Latest | 1 |
| Total_Asset_Turnover_Ratio_Latest | 1 |
| PBIDTM_perc_Latest | 1 |
| PBDTM_perc_Latest | 1 |
| CPM_perc_Latest | 1 |
| APATM_perc_Latest | 1 |
| Current_Ratio_Latest | 1 |

Table.1

- Most Null values present in the column; Inventory_Velocity_Days

## Rows-wise check

```
2825    11
393      1
277      1
598      1
3001     1
        ..
1213     0
1215     0
1216     0
1217     0
3585     0
```

Table.2

- Row 2825 has most number of null values

### Treating missing values

- Since all the columns having missing values have outliers also, so we will replace null values by there repective medians

# 1.3 Transform Target variable into 0 and 1

Dependent variable - We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

- So, we are replacing the values in column "Networth_Next_Year" by default(1) and no default(0).
- Negative values will be considered as default(1)
- Positive values will be considered as non default (0)
- A column added named "Default"

# 1.4 Univariate & Bivariate analysis with interpretation. (only those variables which were significant in the model building)

- Most significant variables identified after the model building are;
  - Book_Value_Adj_Unit_Curr
  - Current_Ratio_Latest
  - Debtors_Ratio_Latest
  - Interest_Cover_Ratio_Latest
  - PBDTM_perc_Latest
- So, analysing these

### Boxplots

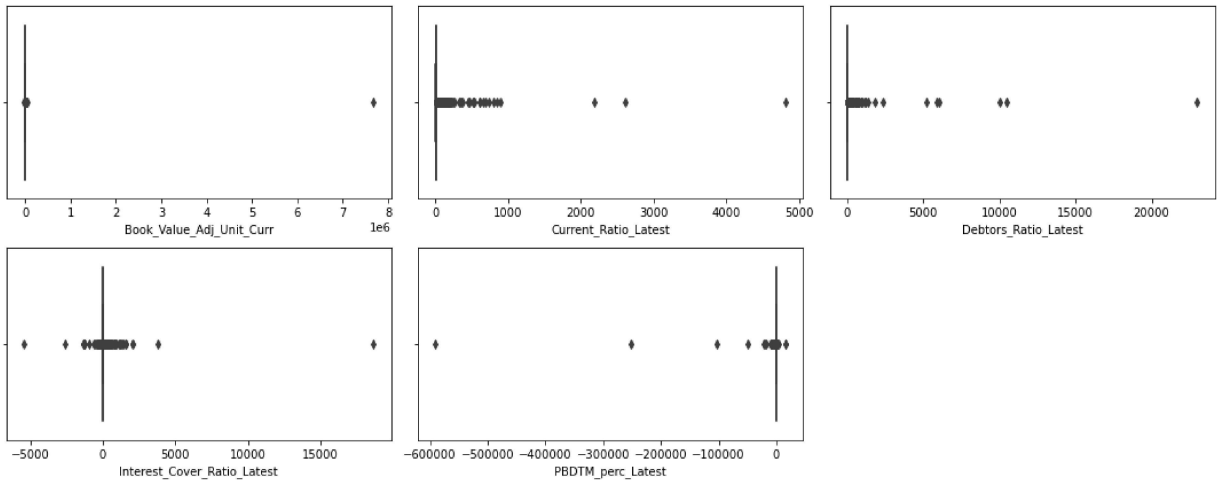


Fig.3

## Distribution plots

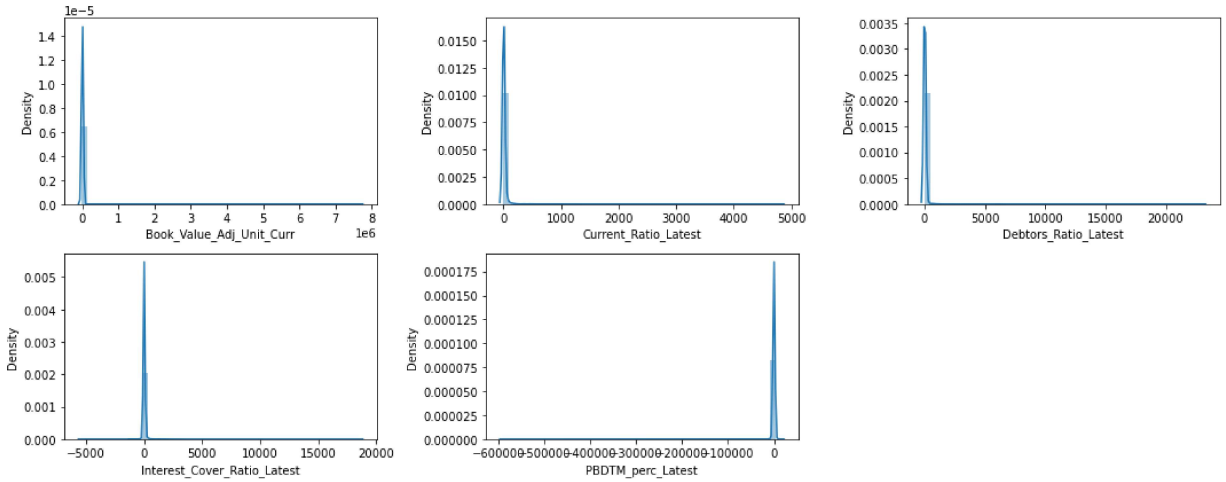


Fig.4

## Countplot for 'Default'

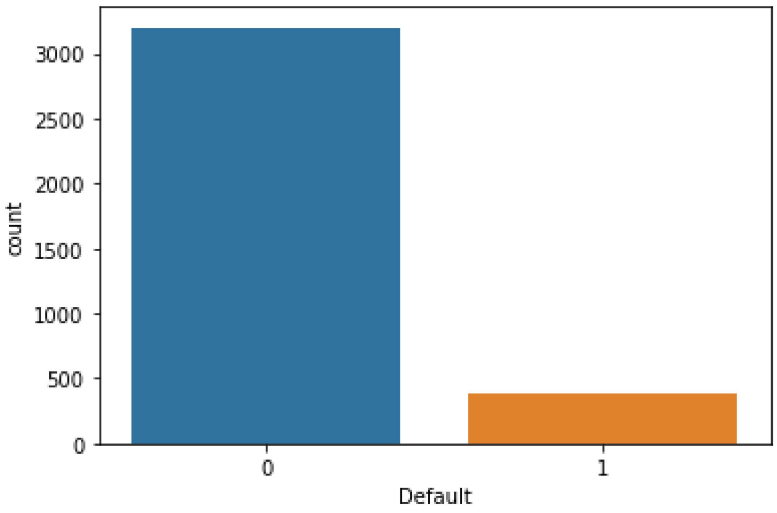


Fig.5

## Plots of most significant variables v/s "Default"

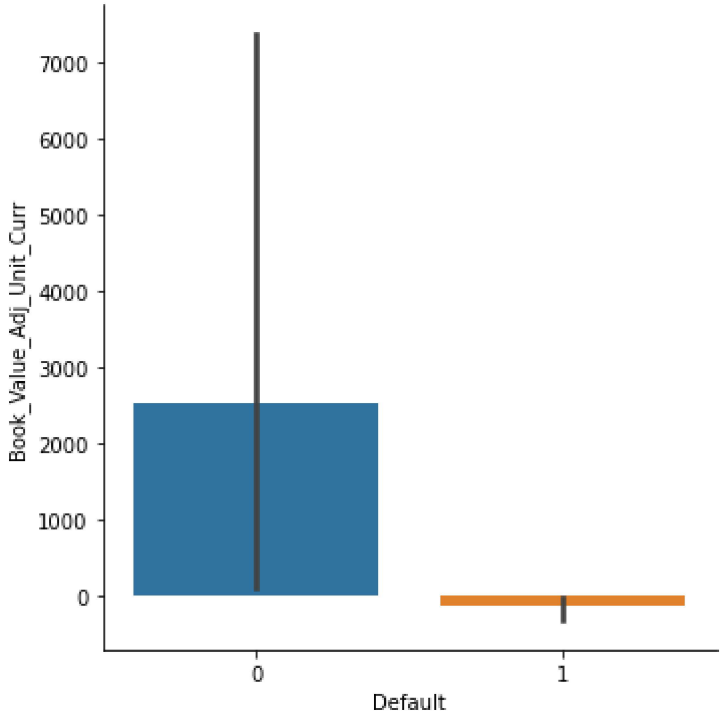


Fig.6

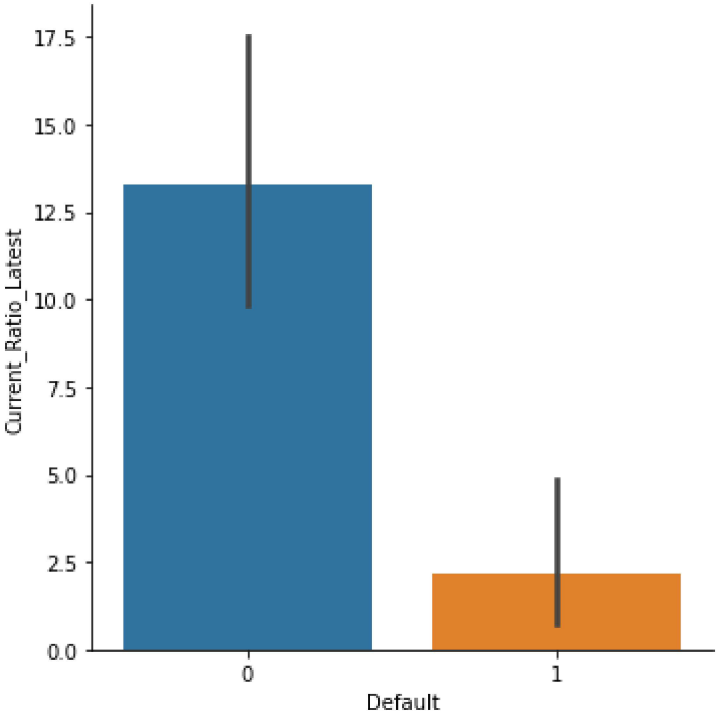
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Current_Ratio_Latest
0
1
Default

Fig.7

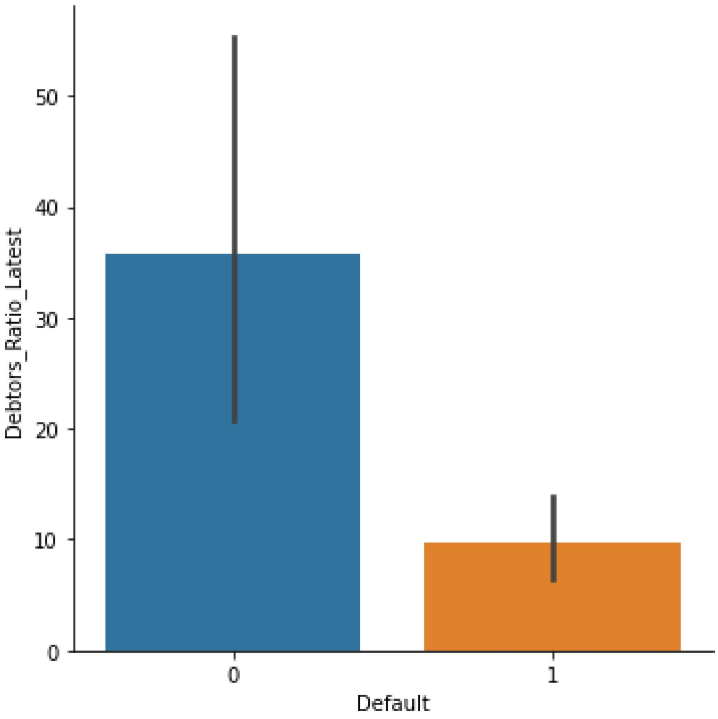
50
40
30
20
10
0
Debtors_Ratio_Latest
0
1
Default

Fig.8

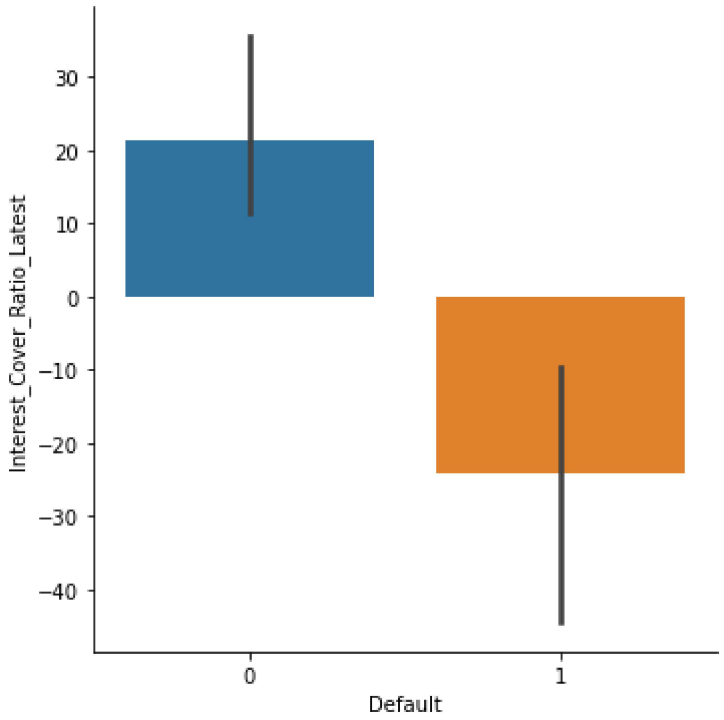


Fig.9

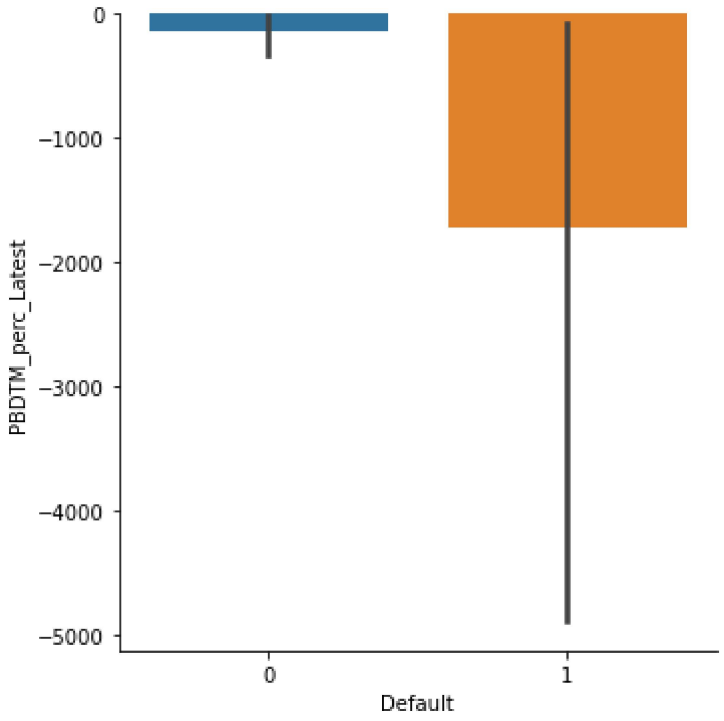


Fig.10

**Heatmap ( most significant variables and "default")**

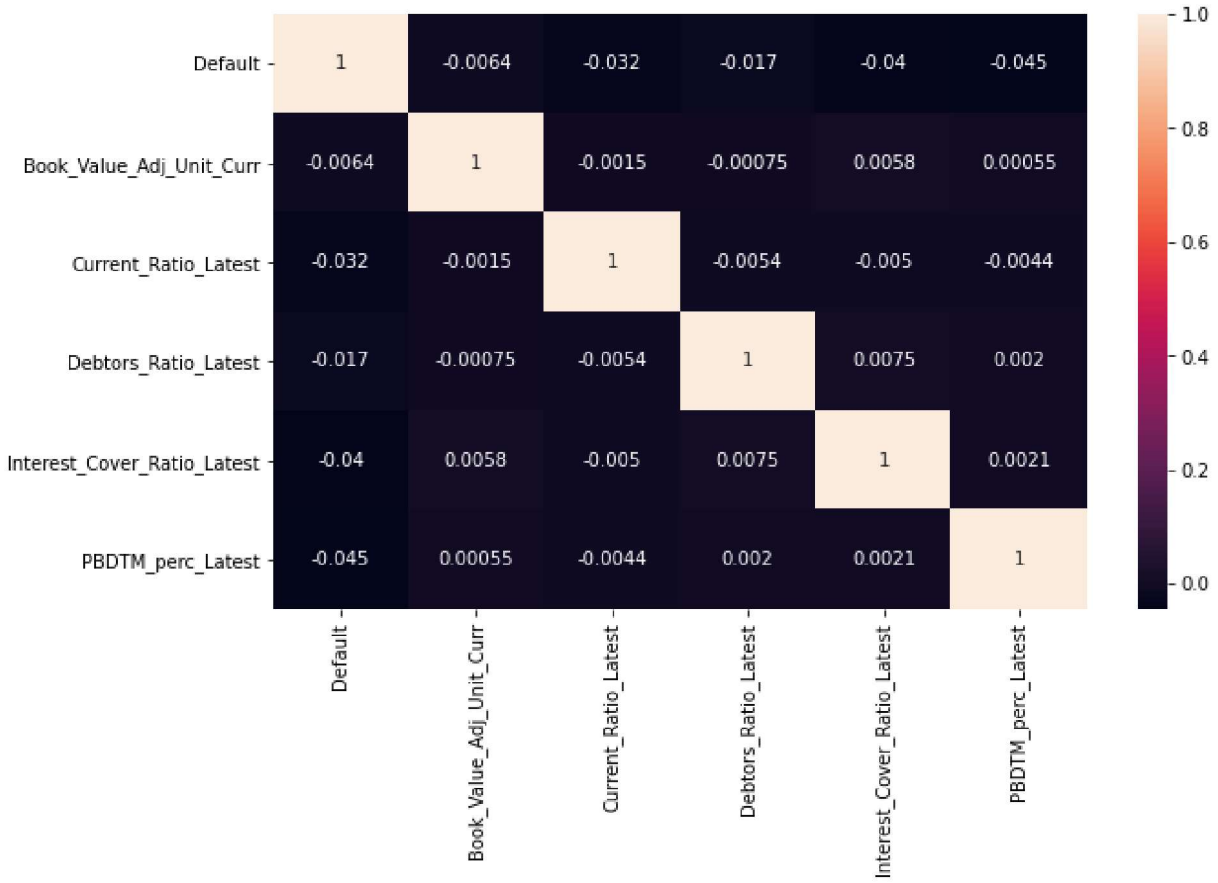


Fig.11

## Interpretations from the graphs of most significant variables and 'Default' variable;

- Boxplot- All the variables are having outliers.
- Distribution plots- All the variables are concentrated at value zero.
- For 'Default' v/s "Book_Value_Adj_Unit_Curr','Current_Ratio_Latest' and 'Debtors_Ratio_Latest', the non-defaulters are high as compared to defaulters.
- For 'Default' v/s "Book_Value_Adj_Unit_Curr' and 'PBDTM_perc_Latest', the relation is in negative direction.
- Heatmap- All the variables are showing no correlation among each other and with 'Default".

## Heatmap for complete dataset

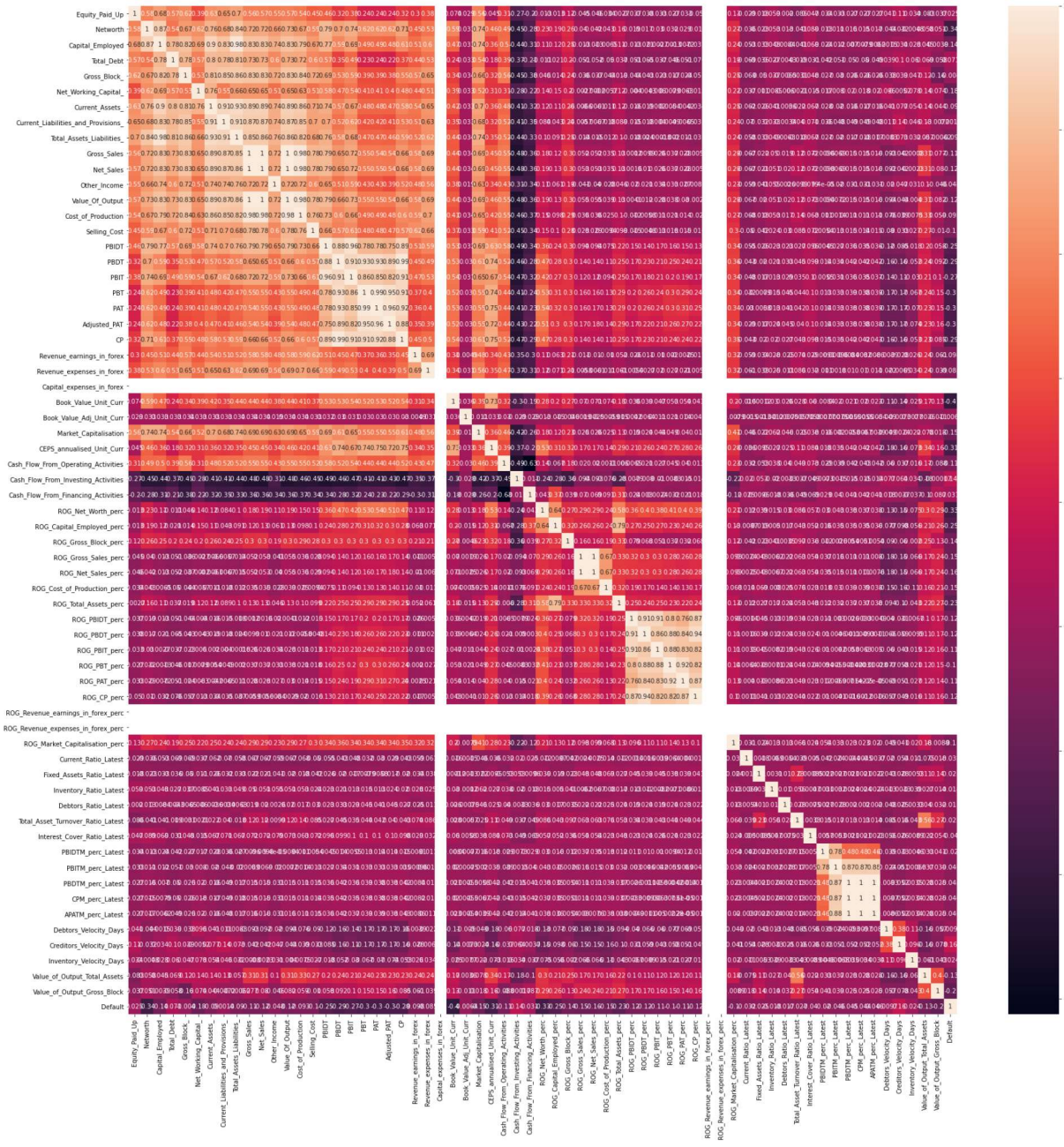

Fig.12

- There is a perfect correlation between following variables:
  - 'Gross_Sales', 'Net_Sales' and 'Value_Of_Output'
  - 'ROG_Gross_Sales_perc'and 'ROG_Net_Sales_perc'
  - 'PBDTM_perc_Latest','CPM_perc_Latest' and 'APATM_perc_Latest So, dropping the columns for further analysis ; 'Net_Sales', 'Value_Of_Output','ROG_Net_Sales_perc','CPM_perc_Latest' and 'APATM_perc_Latest

# 1.5 Train Test Split

- Since, there is a perfect correlation between following variables:

    - 'Gross_Sales', 'Net_Sales' and 'Value_Of_Output'
    - 'ROG_Gross_Sales_perc'and 'ROG_Net_Sales_perc'
    - 'PBDTM_perc_Latest','CPM_perc_Latest' and 'APATM_perc_Latest
    - 'ROG_Revenue_expenses_in_forex_perc',
    - 'ROG_Revenue_earnings_in_forex_perc'
    - 'Capital_expenses_in_forex'
- So, dropping the columns for further analysis ; 'Net_Sales', 'Value_Of_Output','ROG_Net_Sales_perc','CPM_perc_Latest' and 'APATM_perc_Latest

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

## After trying seven different models, by dropping the variables using Variance Inflation Factor and p-values, The Model8 is finalised

## Model8

### Summary of Model8

Logit Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Default | **No. Observations:** | 2402 |
| **Model:** | Logit | **Df Residuals:** | 2396 |
| **Method:** | MLE | **Df Model:** | 5 |
| **Date:** | Sun, 05 Dec 2021 | **Pseudo R-squ.:** | 0.5657 |
| **Time:** | 19:45:35 | **Log-Likelihood:** | -343.70 |
| **converged:** | True | **LL-Null:** | -791.34 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 2.778e-191 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -0.4054 | 0.140 | -2.901 | 0.004 | -0.679 | -0.132 |
| **Book_Value_Adj_Unit_Curr** | -0.1354 | 0.011 | -12.621 | 0.000 | -0.156 | -0.114 |
| **Current_Ratio_Latest** | -0.4051 | 0.077 | -5.230 | 0.000 | -0.557 | -0.253 |
| **Debtors_Ratio_Latest** | -0.0062 | 0.003 | -1.987 | 0.047 | -0.012 | -8.41e-05 |
| **Interest_Cover_Ratio_Latest** | -0.0020 | 0.001 | -2.805 | 0.005 | -0.003 | -0.001 |
| **PBDTM_perc_Latest** | -7.193e-05 | 1.99e-05 | -3.617 | 0.000 | -0.000 | -3.3e-05 |

Table.3

**All the variables are now having p value less than 0.05, so we may finalise this model i.e Model8**

# 1.7 Validation of Model on Test Dataset and the performance matrices with interpretation from the model

**Validating on the train set**

**confusion matrix for train set**

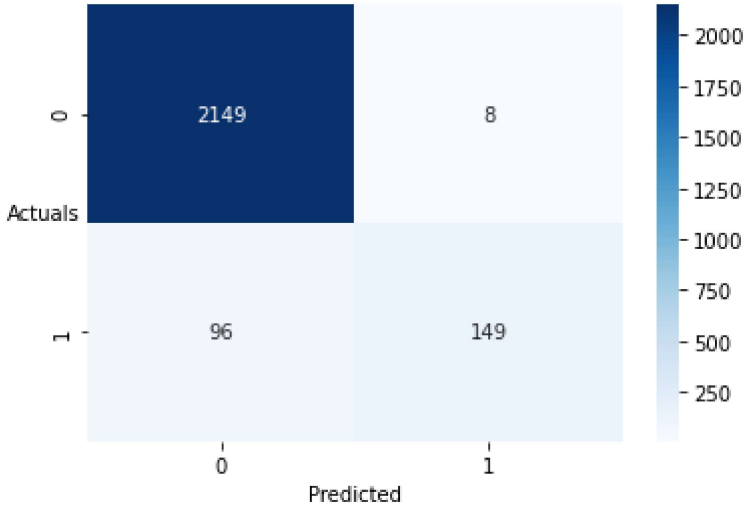


Fig.13

**Classification report for train set**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.957 | 0.996 | 0.976 | 2157 |
| 1 | 0.949 | 0.608 | 0.741 | 245 |
| accuracy | | | 0.957 | 2402 |
| macro avg | 0.953 | 0.802 | 0.859 | 2402 |
| weighted avg | 0.956 | 0.957 | 0.952 | 2402 |

Table.4

## Choosing the optimal threshold

**Validating with revised threshold of 0.218**

**confusion matrix for train set with revised threshold of 0.218**

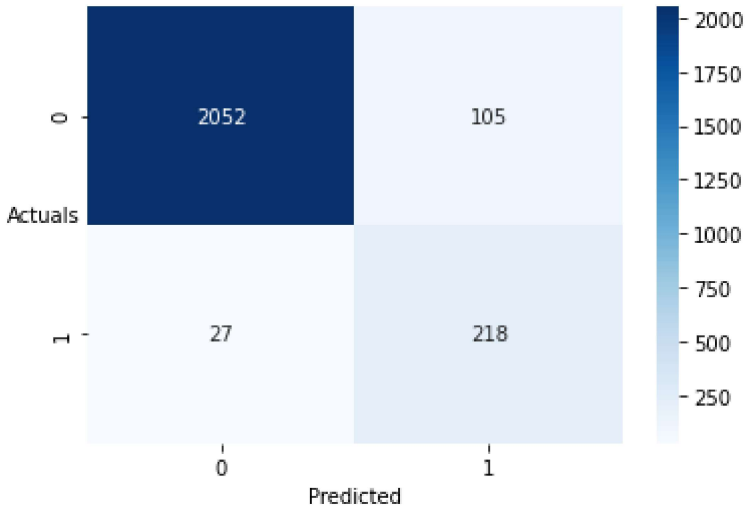


Fig.14

**classification report for train set with revised threshold of 0.218**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.987 | 0.951 | 0.969 | 2157 |
| 1 | 0.675 | 0.890 | 0.768 | 245 |
| accuracy | | | 0.945 | 2402 |
| macro avg | 0.831 | 0.921 | 0.868 | 2402 |
| weighted avg | 0.955 | 0.945 | 0.948 | 2402 |

Table.5

**Validating on the test set**

**confusion matrix for test set (threshold=0.218)**

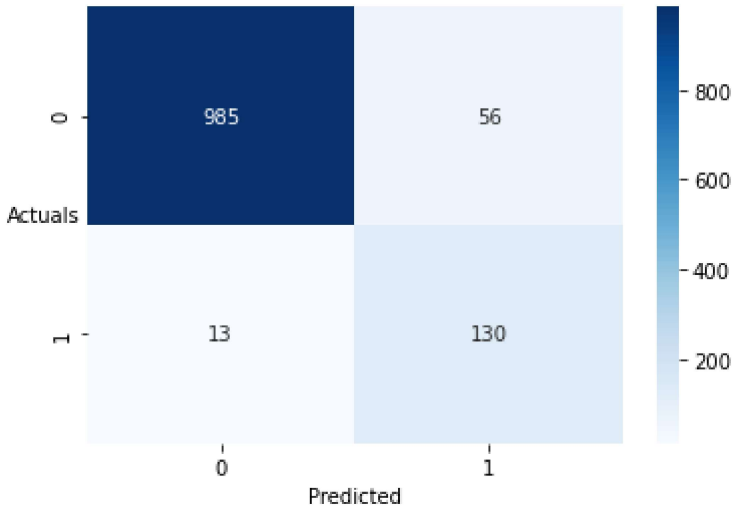


Fig.15

**classification report for test set matrix (threshold=0.218)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.987 | 0.946 | 0.966 | 1041 |
| 1 | 0.699 | 0.909 | 0.790 | 143 |
| accuracy | | | 0.942 | 1184 |
| macro avg | 0.843 | 0.928 | 0.878 | 1184 |
| weighted avg | 0.952 | 0.942 | 0.945 | 1184 |

Table.6

## Interpretations;

- The Recall value for train test is 0.890 and for test set is 0.909
- The Precision value for train test is 0.675 and for test set is 0.699
- Most significant features in data set are;
  - Book_Value_Adj_Unit_Curr
  - Current_Ratio_Latest
  - Debtors_Ratio_Latest
  - Interest_Cover_Ratio_Latest
  - PBDTM_perc_Latest
- All the above variables are in a negative relationship with "Default" i.e an increment in these variables may negatively affect the "Default".
- The increment in above variables may increase the Credit Risk.
- the variable "Current_Ratio_Latest" is showing to have most negative affect and thus an in crease in this will increase the Risk most significantly.

**Thanks.....**