

# **Business Report**

## **Project – FRA Project(Milestone-1)**

### **Predicting Credit Risk for Company data**

**Created by Amit Jain**

## Contents

List of Figure .....	2
Credit Risk Dataset : Introduction .....	3
Problem statement : .....	3
Read the data as Dataframe in python and analyze the data. ....	4
Fixing messy column names (containing spaces) for ease of use : .....	5
Data dictionary : .....	6
Create dependent variable: .....	9
Data types of all variables: .....	10
Dropping unnecessary columns:.....	11
NULL Checks: .....	12
Treat missing values: .....	14
Outlier detection: .....	14
Outlier Treatment: .....	15
Univariate analysis:.....	16
Distinct Value Counts of each field:.....	16
Correlation heatmap .....	18
Model building using stats model:.....	19
Stats Model definitions: .....	19
Partitioning the data into train and test: .....	20
Start with the model .....	20

## List of Figure

Figure 1 HeatMap .....	13
Figure 2 BoxPlot for all data elements .....	14
Figure 3 ScatterPlot.....	18

## Credit Risk Dataset : Introduction

This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment. Also, the purpose of this exercise is to execute Stats Model supervised learning techniques on the given data, combine all predictions and find out the model with best prediction or accuracy. In supervised learning techniques, there are clearly defined X and Y variables. Supervised Learning is used to predict either a continuous response (as in regression) or a categorical response (as in classification). These are machine learning models for combining predictions from multiple separate models. Both regression and classification can be done using Ensemble Learning. Combining all the individual predictions can be done using either voting or averaging.

### 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.

#### **Dataset for Problem 1: Company\_Data2015-1.xlsx**

To understand the problem, for Credit risk checking company has given sample of 3586 Company records collected data in the Company\_Data2015-1.xlsx , which have financial information of the companies for previous year 3025 and we need to predict Company status based on next Year Networth.

**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.

#### **Assumption:**

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Step of understanding the data:

Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

## Read the data as Dataframe in python and analyze the data.

This is how the data look like:

	Co_Code	Co_Name	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	Gross Sales	Net Sales	Other Income
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	1116.85	109.60	0.00	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	1585.74	6043.94	2892.73	2892.73	41.00
2	14852	ABG Shipyards	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	4601.39	12316.07	392.13	392.13	1.00
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	3646.54	6000.42	1354.39	1354.39	22.00
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	2849.58	7524.91	38.72	38.72	1.00

Shape of the data:

The number of rows (observations) is 3586

The number of columns (variables) is 67

Insights:

1. There are only 3586 Rows in sample, data , which are Companies previous year financial statements
2. We have given total 67 different fields for the data, so good amount of observation points.
3. Column names of the data:

```
'Co_Code', 'Co_Name', '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 ',, 'Gross Sales',  
'Net Sales', 'Other Income', 'Value Of Output',, 'Cost of Production', 'Selling  
Cost', 'PBIDT', 'PBDT', 'PBIT', 'PBT',, 'PAT', 'Adjusted PAT', 'CP', 'Revenue  
earnings in forex',, 'Revenue expenses in forex', 'Capital expenses in forex',, 'Book  
Value (Unit Curr)', 'Book Value (Adj.) (Unit Curr)',, 'Market Capitalisation', 'CEPS  
(annualised) (Unit Curr)',, 'Cash Flow From Operating Activities',, 'Cash Flow From  
Investing Activities',, 'Cash Flow From Financing Activities', 'ROG-Net Worth (%)',,  
'ROG-Capital Employed (%)', 'ROG-Gross Block (%)',, 'ROG-Gross Sales (%)', 'ROG-Net  
Sales (%)',, 'ROG-Cost of Production (%)', 'ROG-Total Assets (%)', 'ROG-PBIDT (%)',,  
'ROG-PBDT (%)', 'ROG-PBIT (%)', 'ROG-PBT (%)', 'ROG-PAT (%)',, 'ROG-CP (%)', 'ROG-  
Revenue earnings in forex (%)',, 'ROG-Revenue expenses in forex (%)', 'ROG-Market  
Capitalisation (%)',, 'Current Ratio[Latest]', 'Fixed Assets Ratio[Latest]',,  
'Inventory Ratio[Latest]', 'Debtors Ratio[Latest]',, 'Total Asset Turnover  
Ratio[Latest]', 'Interest Cover Ratio[Latest]',, 'PBIDTM (%) [Latest]', 'PBITM  
(%) [Latest]', 'PBDTM (%) [Latest]',, 'CPM (%) [Latest]', 'APATM (%) [Latest]', 'Debtors  
Velocity (Days)',, 'Creditors Velocity (Days)', 'Inventory Velocity (Days)',, 'Value  
of Output/Total Assets', 'Value of Output/Gross Block'
```

## Fixing messy column names (containing spaces) for ease of use :

We can also observe that above listed column names have multiple Special characters in it , example [ , ( , ) , %, - white space etc.

In order to start analyzing them in our Python tool, we need to clear this clutter and modify Column names only.

After fixing column names, this is how the Column names look like:

```
'Co_Code', 'Co_Name', 'Networth_Next_Year', 'Equity_Paid_Up', 'Networth', 'Capital_Employed', 'Total_Debt',  
'Gross_Block', 'Net_Working_Capital', 'Current_Assets', 'Current_Liabilities_and_Provisions',  
'Total_Assets_by_Liabilities', 'Gross_Sales', 'Net_Sales', 'Other_Income', 'Value_Of_Output',  
'Cost_of_Production', 'Selling_Cost', 'PBDT', 'PBDT', 'PBIT', 'PBT', 'PAT', 'Adjusted_PAT', 'CP',  
'Revenue_earnings_in_forex', 'Revenue_expenses_in_forex', 'Capital_expenses_in_forex',  
'Book_Value_Unit_Curr', 'Book_Value_Adj_Unit_Curr', 'Market_Capitalisation', 'CEPS_annualised_Unit_Curr',  
'Cash_Flow_From_Operating_Activities', 'Cash_Flow_From_Investing_Activities',  
'Cash_Flow_From_Financing_Activities', 'ROG_Net_Worth_perc', 'ROG_Capital_Employed_perc',  
'ROG_Gross_Block_perc', 'ROG_Gross_Sales_perc', 'ROG_Net_Sales_perc', 'ROG_Cost_of_Production_perc',  
'ROG_Total_Assets_perc', 'ROG_PBDT_perc', 'ROG_PBDT_perc', 'ROG_PBIT_perc', 'ROG_PBT_perc',  
'ROG_PAT_perc', 'ROG_CP_perc', 'ROG_Revenue_earnings_in_forex_perc',  
'ROG_Revenue_expenses_in_forex_perc', 'ROG_Market_Capitalisation_perc', 'Current_RatioLatest',  
'Fixed_Assets_RatioLatest', 'Inventory_RatioLatest', 'Debtors_RatioLatest',  
'Total_Asset_Turnover_RatioLatest', 'Interest_Cover_RatioLatest', 'PBDTM_perclatest', 'PBITM_perclatest',  
'PBDTM_perclatest', 'CPM_perclatest', 'APATM_perclatest', 'Debtors_Velocity_Days',  
'Creditors_Velocity_Days', 'Inventory_Velocity_Days', 'Value_of_Output_by_Total_Assets',  
'Value_of_Output_by_Gross_Block'
```

## Data dictionary :

#	Field Name	Description	New Field Name
1	Co_Code	Company Code	Co_Code
2	Co_Name	Company Name	Co_Name
3	Networth Next Year	Value of a company as on 2016 – Next Year(difference between the value of total assets and total liabilities)	Networth_Next_Year
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders	Equity_Paid_Up
5	Networth	Value of a company as on 2015 – Current Year	Networth
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company	Capital_Employed
7	Total Debt	The sum of money borrowed by the company and is due to be paid	Total_Debt
8	Gross Block	Total value of all of the assets that a company owns	Gross_Block
9	Net Working Capital	The difference between a company's current assets (cash, accounts receivable, inventories of raw materials and finished goods) and its current liabilities (accounts payable).	Net_Working_Capital
10	Current Assets	All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year.	Curr_Assets
11	Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)	Curr_Liab_and_Prov
12	Total Assets/Liabilities	Ratio of total assets to liabilities of the company	Total_Assets_to_Liab
13	Gross Sales	The grand total of sale transactions within the accounting period	Gross_Sales
14	Net Sales	Gross sales minus returns, allowances, and discounts	Net_Sales
15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)	Other_Income
16	Value Of Output	Product of physical output of goods and services produced by company and its market price	Value_Of_Output
17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service	Cost_of_Prod
18	Selling Cost	Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms)	Selling_Cost
19	PBIDT	Profit Before Interest, Depreciation & Taxes	PBIDT
20	PBDT	Profit Before Depreciation and Tax	PBDT
21	PBIT	Profit before interest and taxes	PBIT

22	PBT	Profit before tax	PBT
23	PAT	Profit After Tax	PAT
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit	Adjusted_PAT
26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.	CP
27	Revenue earnings in forex	Revenue earned in foreign currency	Rev_earn_in_forex
28	Revenue expenses in forex	Expenses due to foreign currency transactions	Rev_exp_in_forex
29	Capital expenses in forex	Long term investment in forex	Capital_exp_in_forex
30	Book Value (Unit Curr)	Net asset value	Book_Value_Unit_Curr
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value	Book_Value_Adj_Unit_Curr
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share	Market_Capitalisation
33	CEPS (7nnualized) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis	CEPS_annualised_Unit_Curr
34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities	Cash_Flow_From_Opr
35	Cash Flow From Investing Activities	Cash used in the purchase of non-current assets—or long-term assets– that will deliver value in the future	Cash_Flow_From_Inv
36	Cash Flow From Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)	Cash_Flow_From_Fin
37	ROG-Net Worth (%)	Rate of Growth – Networth	ROG_Net_Worth_perc
38	ROG-Capital Employed (%)	Rate of Growth – Capital Employed	ROG_Capital_Employed_perc
39	ROG-Gross Block (%)	Rate of Growth – Gross Block	ROG_Gross_Block_perc
40	ROG-Gross Sales (%)	Rate of Growth – Gross Sales	ROG_Gross_Sales_perc
41	ROG-Net Sales (%)	Rate of Growth – Net Sales	ROG_Net_Sales_perc
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production	ROG_Cost_of_Prod_perc
43	ROG-Total Assets (%)	Rate of Growth – Total Assets	ROG_Total_Assets_perc
44	ROG-PBIDT (%)	Rate of Growth- PBIDT	ROG_PBIDT_perc
45	ROG-PBDT (%)	Rate of Growth- PBDT	ROG_PBDT_perc
46	ROG-PBIT (%)	Rate of Growth- PBIT	ROG_PBIT_perc

47	ROG-PBT (%)	Rate of Growth- PBT	ROG_PBT_perc
48	ROG-PAT (%)	Rate of Growth- PAT	ROG_PAT_perc
49	ROG-CP (%)	Rate of Growth- CP	ROG_CP_perc
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex	ROG_Rev_earn_in_forex_perc
51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex	ROG_Rev_exp_in_forex_perc
52	ROG-Market Capitalisation (%)	Rate of Growth – Market Capitalisation	ROG_Market_Capitalisation_perc
53	Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year	Curr_Ratio_Latest
54	Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating	Fixed_Assets_Ratio_Latest
55	Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company	Inventory_Ratio_Latest
56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company	Debtors_Ratio_Latest
57	Total Asset Turnover Ratio[Latest]	The value of a company's revenues relative to the value of its assets	Total_Asset_Turnover_Ratio_Latest
58	Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt	Interest_Cover_Ratio_Latest
59	PBIDTM (%) [Latest]	Profit before Interest Depreciation and Tax Margin	PBIDTM_perc_Latest
60	PBITM (%) [Latest]	Profit Before Interest Tax Margin	PBITM_perc_Latest
61	PBDTM (%) [Latest]	Profit Before Depreciation Tax Margin	PBDTM_perc_Latest
62	CPM (%) [Latest]	Cost per thousand (advertising cost)	CPM_perc_Latest
63	APATM (%) [Latest]	After tax profit margin	APATM_perc_Latest
64	Debtors Velocity (Days)	Average days required for receiving the payments	Debtors_Vel_Days
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers	Creditors_Vel_Days
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales	Inventory_Vel_Days
67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets	Value_of_Output_to_Total_Assets
68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block	Value_of_Output_to_Gross_Block



## Description of the data elements:

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586.0	16065.388734	19776.817379	4.00	3029.2500	6077.500	24269.5000	72493.00
Networth_Next_Year	3586.0	725.045251	4769.681004	-8021.60	3.9850	19.015	123.8025	111729.10
Equity_Paid_Up	3586.0	62.966584	778.761744	0.00	3.7500	8.290	19.5175	42263.46
Networth	3586.0	649.746299	4091.988792	-7027.48	3.8925	18.580	117.2975	81657.35
Capital_Employed	3586.0	2799.611054	26975.135385	-1824.75	7.6025	39.090	226.6050	714001.25
...	...	...	...	...	...	...	...	...
Debtors_Velocity_Days	3586.0	603.894032	10636.759580	0.00	8.0000	49.000	106.0000	514721.00
Creditors_Velocity_Days	3586.0	2057.854992	54169.479197	0.00	8.0000	39.000	89.0000	2034145.00
Inventory_Velocity_Days	3483.0	79.644559	137.847792	-199.00	0.0000	35.000	96.0000	996.00
Value_of_Output_by_Total_Assets	3586.0	0.819757	1.201400	-0.33	0.0700	0.480	1.1600	17.63
Value_of_Output_by_Gross_Block	3586.0	61.884548	976.824352	-61.00	0.2700	1.530	4.9100	43404.00

66 rows x 8 columns

## Insights:

1. We can clearly see that data have different range of Min, max and Median.
2. We can Scale the data, based on our what prediction method we are using.

## Create dependent variable:

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.

Lets check for the proportion of the data:

```
1    3198
1     388
```

Now lets check for the summary of the “default” target data :

```
count    3586.000000
mean      0.108199
std       0.310674
min       0.000000
25%       0.000000
50%       0.000000
75%       0.000000
max       1.000000
Name: default, dtype: float64
```

## Insights:

1. Target variable is 1, which denotes the Default companies
2. Value 0 denotes, non default company
3. Default companies data is very less as compare to Non-default company.
4. Ratio of the target variable is 10.8%

## Data types of all variables:

Data columns (total 68 columns):

#	Column	Non-Null Count	Dtype
0	Co_Code	3586 non-null	int64
1	Co_Name	3586 non-null	object
2	Networth_Next_Year	3586 non-null	float64
3	Equity_Paid_Up	3586 non-null	float64
4	Networth	3586 non-null	float64
5	Capital_Employed	3586 non-null	float64
6	Total_Debt	3586 non-null	float64
7	Gross_Block	3586 non-null	float64
8	Net_Working_Capital	3586 non-null	float64
9	Current_Assets	3586 non-null	float64
10	Current_Liabilities_and_Provisions	3586 non-null	float64
11	Total_Assets_by_Liabilities	3586 non-null	float64
12	Gross_Sales	3586 non-null	float64
13	Net_Sales	3586 non-null	float64
14	Other_Income	3586 non-null	float64
15	Value_Of_Output	3586 non-null	float64
16	Cost_of_Production	3586 non-null	float64
17	Selling_Cost	3586 non-null	float64
18	PBIDT	3586 non-null	float64
19	PBDT	3586 non-null	float64
20	PBIT	3586 non-null	float64
21	PBT	3586 non-null	float64
22	PAT	3586 non-null	float64
23	Adjusted_PAT	3586 non-null	float64
24	CP	3586 non-null	float64
25	Revenue_earnings_in_forex	3586 non-null	float64
26	Revenue_expenses_in_forex	3586 non-null	float64
27	Capital_expenses_in_forex	3586 non-null	float64
28	Book_Value_Unit_Curr	3586 non-null	float64
29	Book_Value_Adj_Unit_Curr	3582 non-null	float64
30	Market_Capitalisation	3586 non-null	float64
31	CEPS_annualised_Unit_Curr	3586 non-null	float64
32	Cash_Flow_From_Operating_Activities	3586 non-null	float64
33	Cash_Flow_From_Investing_Activities	3586 non-null	float64
34	Cash_Flow_From_Financing_Activities	3586 non-null	float64
35	ROG_Net_Worth_perc	3586 non-null	float64
36	ROG_Capital_Employed_perc	3586 non-null	float64
37	ROG_Gross_Block_perc	3586 non-null	float64

38	ROG_Gross_Sales_perc	3586	non-null	float64
39	ROG_Net_Sales_perc	3586	non-null	float64
40	ROG_Cost_of_Production_perc	3586	non-null	float64
41	ROG_Total_Assets_perc	3586	non-null	float64
42	ROG_PBITD_perc	3586	non-null	float64
43	ROG_PBDT_perc	3586	non-null	float64
44	ROG_PBIT_perc	3586	non-null	float64
45	ROG_PBT_perc	3586	non-null	float64
46	ROG_PAT_perc	3586	non-null	float64
47	ROG_CP_perc	3586	non-null	float64
48	ROG_Revenue_earnings_in_forex_perc	3586	non-null	float64
49	ROG_Revenue_expenses_in_forex_perc	3586	non-null	float64
50	ROG_Market_Capitalisation_perc	3586	non-null	float64
51	Current_RatioLatest	3585	non-null	float64
52	Fixed_Assets_RatioLatest	3585	non-null	float64
53	Inventory_RatioLatest	3585	non-null	float64
54	Debtors_RatioLatest	3585	non-null	float64
55	Total_Asset_Turnover_RatioLatest	3585	non-null	float64
56	Interest_Cover_RatioLatest	3585	non-null	float64
57	PBITM_percLatest	3585	non-null	float64
58	PBITM_percLatest	3585	non-null	float64
59	PBDTM_percLatest	3585	non-null	float64
60	CPM_percLatest	3585	non-null	float64
61	APATM_percLatest	3585	non-null	float64
62	Debtors_Velocity_Days	3586	non-null	int64
63	Creditors_Velocity_Days	3586	non-null	int64
64	Inventory_Velocity_Days	3483	non-null	float64
65	Value_of_Output_by_Total_Assets	3586	non-null	float64
66	Value_of_Output_by_Gross_Block	3586	non-null	float64
67	default	3586	non-null	int32

dtypes: float64(63), int32(1), int64(3), object(1)

#### Insights:

1. All of the data is in Numeric Format, except Co\_Code and Co\_Name, and these 2 fields are not required , so looks good to me.

#### Dropping unnecessary columns:

Drop the fields CO\_Code and CO\_name , since these are not required for our model. Also Drop column Networth\_Next\_Year, because we used this field to build the Dependent Field "default"

## NULL Checks:

Equity_Paid_Up	0
Networth	0
Capital_Employed	0
Total_Debt	0
Gross_Block	0
...	
Creditors_Velocity_Days	0
Inventory_Velocity_Days	103
Value_of_Output_by_Total_Assets	0
Value_of_Output_by_Gross_Block	0
default	0

## Insights:

1. We checked that data have NULL values in all of these positional columns:  
(array([26, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 61], dtype=int64),)
2. Total NULL values in data set are : 103
3. Let's check how many percentage values are missing in each columns.  
if there are more than 30% of the values are missing, we will drop them

Inventory_Velocity_Days	0.028723
Book_Value_Adj_Unit_Curr	0.001115
Total_Asset_Turnover_RatioLatest	0.000279
CPM_percLatest	0.000279
PBDTM_percLatest	0.000279
...	
ROG_Net_Worth_perc	0.000000
Networth	0.000000
PBIDT	0.000000
Capital_Employed	0.000000
default	0.000000

No Column have values more than 30%, so we are good in this case, and no need to drop any field because of Missing values. Let's visually inspect the missing values in our data

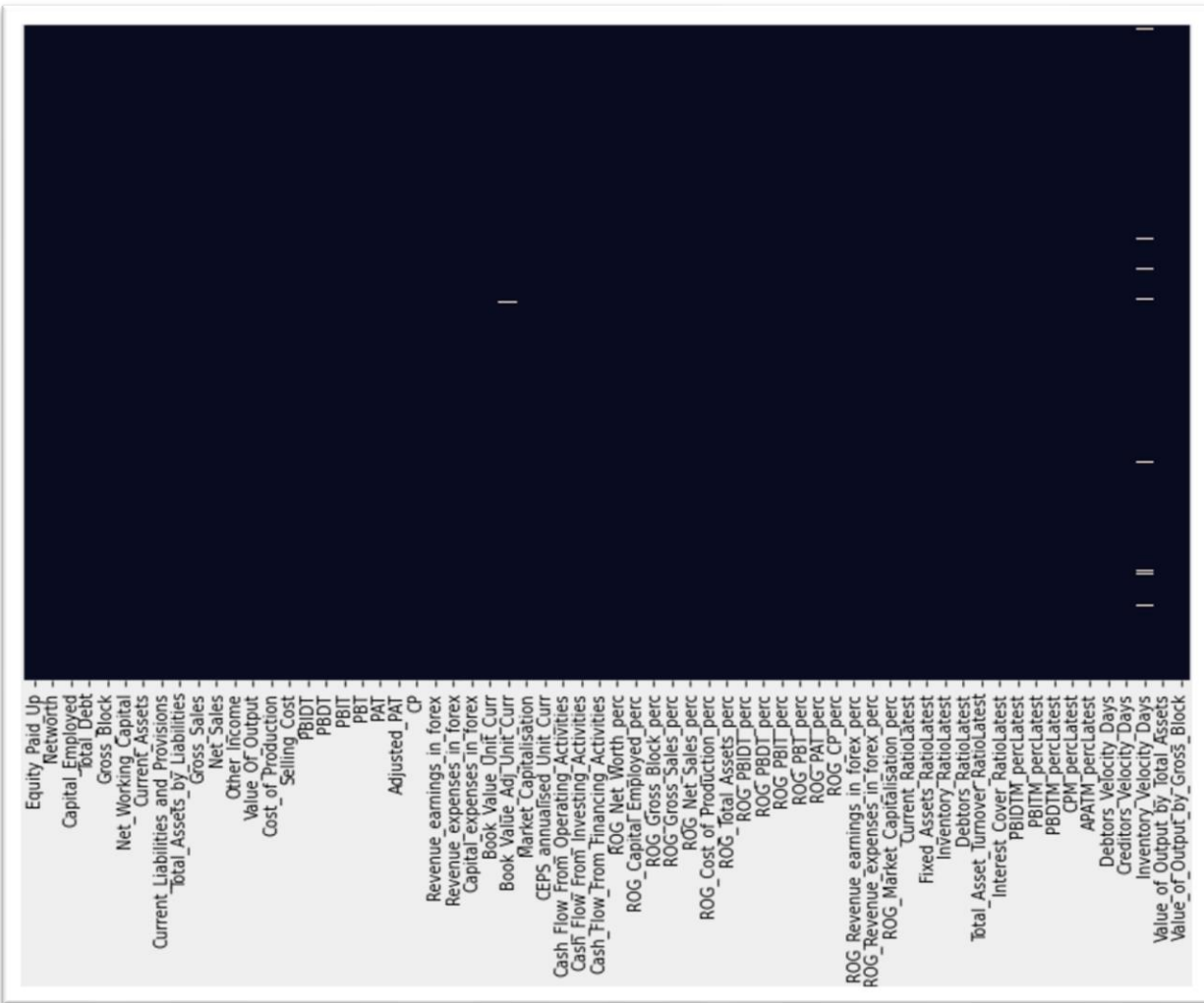


Figure 1 HeatMap

#### Insights:

1. We can see that data have NULL values in a few columns only, and that is spotted in white dashes in the data . This data is too less. And we can impute them
2. We can see that maximum percentage of missing are 2.8% in Column **Inventory\_Velocity\_Days**

## Treat missing values:

There are many ways to Treat Missing values.

1. Either Replace them with Median
2. Impute them with any imputer

We will use both methods and analyze them separately . First of all we will impute missing values with Median value of that column:

Replace NULL with Median and this is how the data look like :

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

0

## Outlier detection:

Plot all data n box plot:

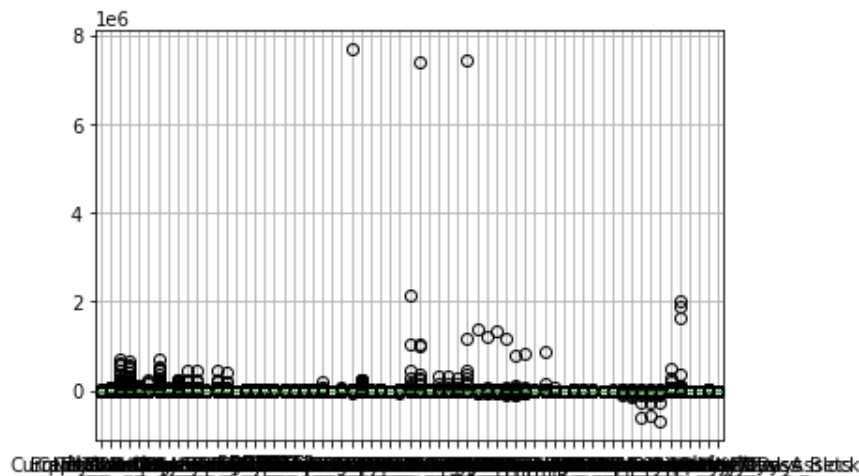


Figure 2 BoxPlot for all data elements

Insights:

1. Also individual box plot can not be draw in this document, as there are 65 box plot to draw, which wont fit into page.
2. But we can clearly see some bubbles in the above graph , which is depicting outliers in the data

## Outlier Treatment:

We have created a function for checking upper and lower limit of the data and we will use that function for calculating, total no of records in each field, which are beyond this range.

List the total no of values in each column, which are above upper limit or less than lower limit of that column.

```
Equity_Paid_Up          905
Networth                1342
Capital_Employed        1220
Total_Debt              1247
Gross_Block             1236
...
Debtors_Velocity_Days   738
Creditors_Velocity_Days 770
Inventory_Velocity_Days 719
Value_of_Output_by_Total_Assets 633
Value_of_Output_by_Gross_Block 984
Length: 64, dtype: int64
```

## Insights:

1. There are total of 76984 records, which have values beyond upper and lower range of the value in that field.
2. There are many ways we can Treat Outliers ,
  - a. Either we can replace Higher values to Upper Limit or very less value to Lower limit of outliers
  - b. Or we can impute them to K Nearest neighbor Method.
  - c. In our Case, we dont know Company Segmentations, and Companies can have different levels of Revenue, Profits and all can be Valid at same time. So we should Treat them with K nearest neighbor method , and Treat them same as we did for Missing values.
  - d. Still I will be using both method for KNN as well as replacing higher values with Upper/lower values

First of all, we would replace the outlier data with upper and lower value.

Again , since data fields are too large, we can display the Box plot in the Business report.

## Univariate analysis:

### Distinct Value Counts of each field:

Equity\_Paid\_Up:1586  
Networth:2334  
Capital\_Employed:2512  
Total\_Debt:1557  
Gross\_Block:1877  
Net\_Working\_Capital:2078  
Current\_Assets:2199  
Current\_Liabilities\_and\_Provisions:1698  
Total\_Assets\_by\_Liabilities:2565  
Gross\_Sales:2081  
Net\_Sales:2079  
Other\_Income:545  
Value\_Of\_Output:2097  
Cost\_of\_Production:1975  
Selling\_Cost:531  
PBIDT:1418  
PBDT:1179  
PBIT:1327  
PBT:969  
PAT:885  
Adjusted\_PAT:883  
CP:1129  
Revenue\_earnings\_in\_forex:431  
Revenue\_expenses\_in\_forex:528  
Capital\_expenses\_in\_forex:1  
Book\_Value\_Unit\_Curr:2540  
Book\_Value\_Adj\_Unit\_Curr:2483  
Market\_Capitalisation:1469  
CEPS\_annualised\_Unit\_Curr:1315  
Cash\_Flow\_From\_Operating\_Activities:1340  
Cash\_Flow\_From\_Investing\_Activities:922  
Cash\_Flow\_From\_Financing\_Activities:989  
ROG\_Net\_Worth\_perc:1725  
ROG\_Capital\_Employed\_perc:1972  
ROG\_Gross\_Block\_perc:1103  
ROG\_Gross\_Sales\_perc:2084  
ROG\_Net\_Sales\_perc:2085  
ROG\_Cost\_of\_Production\_perc:2069  
ROG\_Total\_Assets\_perc:2107  
ROG\_PBIDT\_perc:2208  
ROG\_PBDT\_perc:2201  
ROG\_PBIT\_perc:2216  
ROG\_PBT\_perc:2150  
ROG\_PAT\_perc:2097  
ROG\_CP\_perc:2175  
ROG\_Revenue\_earnings\_in\_forex\_perc:1  
ROG\_Revenue\_expenses\_in\_forex\_perc:1  
ROG\_Market\_Capitalisation\_perc:1600  
Current\_RatioLatest:455



```

Fixed_Assets_RatioLatest:741
Inventory_RatioLatest:1134
Debtors_RatioLatest:1161
Total_Asset_Turnover_RatioLatest:352
Interest_Cover_RatioLatest:776
PBIDTM_percLatest:1835
PBITM_percLatest:1654
PBDTM_percLatest:1707
CPM_percLatest:1551
APATM_percLatest:1268
Debtors_Velocity_Days:241
Creditors_Velocity_Days:203
Inventory_Velocity_Days:225
Value_of_Output_by_Total_Assets:283
Value_of_Output_by_Gross_Block:764
default:2

```

insights:

1. These fields have distinct count of value as only 1

```

Column name is Capital_expenses_in_forex and its unique value count is : 1
Column name is ROG_Revenue_earnings_in_forex_perc and its unique value count is : 1
Column name is ROG_Revenue_expenses_in_forex_perc and its unique value count is : 1

```

2. Since these columns have only 1 value in it, we can not use it for our Predictions. So we should Drop these fields, as they are not going to contribute anything in Target variable prediction.
3. Individual Box plot and distribution of the data . Please refer Notebook for the diagram.
4. Data Skewness :

Equity_Paid_Up	1.141900
Networth	0.903328
Capital_Employed	1.137131
Total_Debt	1.197970
Gross_Block	1.228742
	...
Creditors_Velocity_Days	1.143302
Inventory_Velocity_Days	1.206446
Value_of_Output_by_Total_Assets	1.110709
Value_of_Output_by_Gross_Block	1.183618
default	2.523672

5. Data is not normally distributed and it is Right skewed and left skewed in both directions for many fields.
6. Checking proportion of default

0	0.891801
1	0.108199

# Correlation heatmap

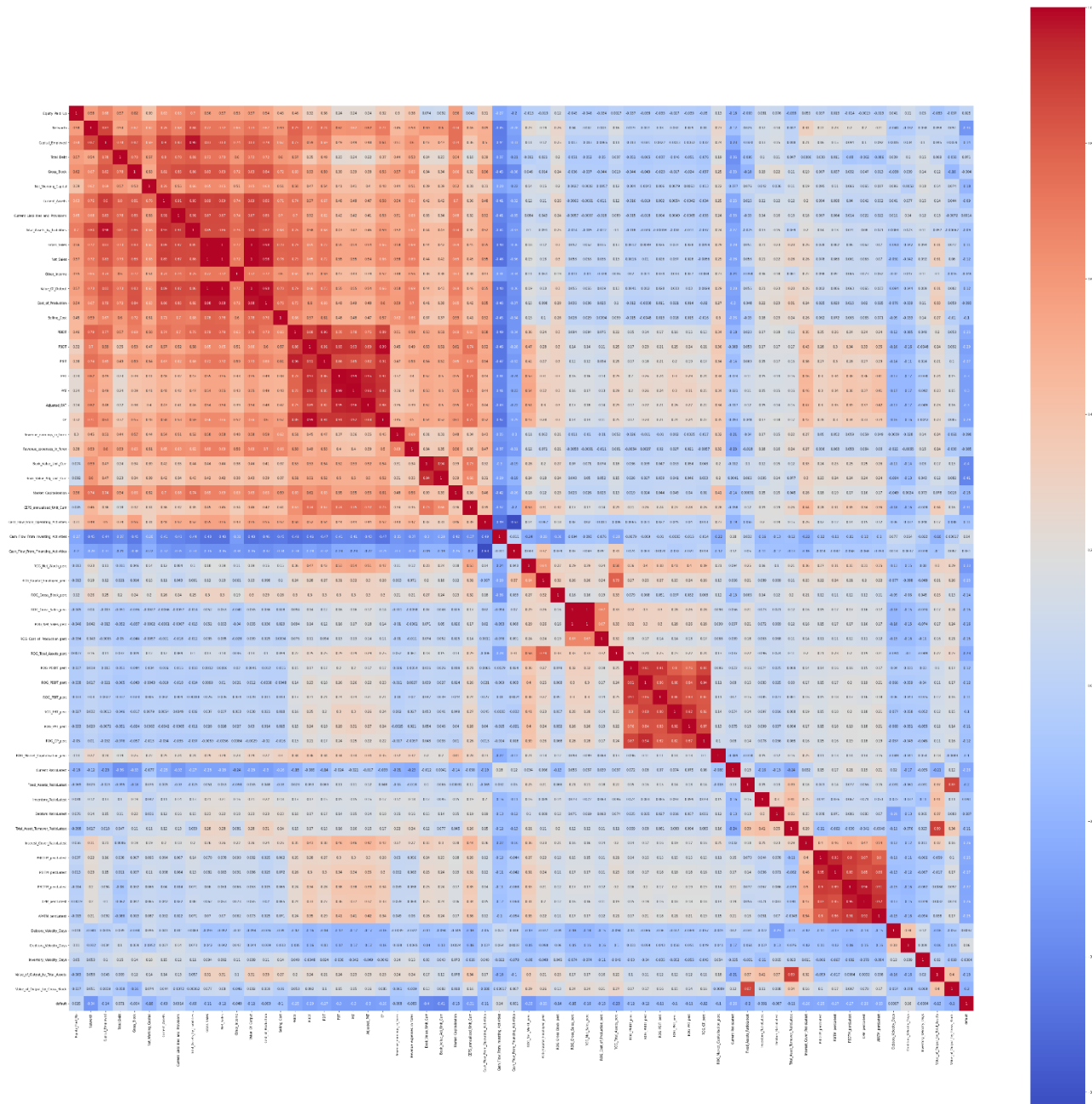


Figure 3 ScatterPlot

## Insights:

1. We don't need to look at the Score of co-relation for each column, we can just look at the color coding and Dark Red and Dark Blue showing Strong Positive and negative co-relations, among fields. We can clearly see that there are many fields, which can be avoided for analyzing and building prediction models, because they have co-relation between them,

## Model building using stats model:

### Stats Model definitions:

We will use statsmodels module to implement Ordinary Least Squares(OLS) method of linear regression for predicting “Default “ Companies based on the data provided for us.

Introduction :

A linear regression model establishes the relation between a dependent variable(y) and at least one independent variable(x) as:

$$\hat{y} = b_1x + b_0$$

In OLS method, we have to choose the values of  $b_1$  and  $b_0$  such that, the total sum of squares of the difference between the calculated and observed values of y, is minimized.

Formula for OLS:

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_1x_i - b_0)^2 = \sum_{i=1}^n (\epsilon_i)^2 = \min$$

Where,

$\hat{y}_i$ = predicted value for the ith observation

$y_i$ = actual value for the ith observation

$\epsilon_i$ = error/residual for the ith observation

n = total number of observations

To get the values of  $b_0$  and  $b_1$  which minimize S, we can take a partial derivative for each coefficient and equate it to zero.

### Approach :

First of all we define the variables **x** and **y**. In the example below, the variables are read from a csv file using pandas.

Next, We need to add the constant  $b_0$  to the equation using the **add\_constant()** method.

The OLS() function of the statsmodels.api module is used to perform OLS regression. It returns an OLS object.

Then fit() method is called on this object for fitting the regression line to the data.

The summary() method is used to obtain a table which gives an extensive description about the regression results

## Partitioning the data into train and test:

We divided the data into train and Test data set. With given parameters:

```
test_size = 0.33, random_state = 42, stratify = y
```

**Test Size specify**, what should be the size of train and test data set, in our case, train set will have 66 % of data and test will have about 33% of data elements

**Random state** given as 42.

Check the shape of the data :

Train size: (2402, 61)

Test Size : (1184, 61)

**Why stratify = y?**

Please note, because this data is highly imbalanced and could possibly result into different proportions in the y variable between train and test set.

## Start with the model

This is how the final data looks like :

```
Company_imputed.head()
```

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Current_Assets	Current_Liabilities_and_Provisions	Total_A
0	43.16875	-166.215	-320.90125	180.83	328.8825	-89.40625	40.50000	163.02625	
1	43.16875	-166.215	555.10875	180.83	328.8825	-89.40625	332.19375	163.02625	
2	43.16875	287.405	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	
3	43.16875	-166.215	555.10875	180.83	328.8825	-89.40625	332.19375	163.02625	
4	43.16875	-166.215	555.10875	180.83	328.8825	151.52375	332.19375	163.02625	

Once the data is split into 4 parts : X\_train, X\_test, y\_train, y\_test

We need to Concatenate train data (X\_train, y\_train) with name as TRAIN and test data (X\_test, y\_test) as TEST separately, because Stats model work on Complete data, it does need Separation of independent variables into 2 form of train and test.

After concatenation of data, we have got these list of fields in Train data , which are all fields and all rows of independent variables:

```
Index(['Equity_Paid_Up', 'Networth', 'Capital_Employed', 'Total_Debt',
      'Gross_Block', 'Net_Working_Capital', 'Current_Assets',
      'Current_Liabilities_and_Provisions', 'Total_Assets_by_Liabilities',
      'Gross_Sales', 'Net_Sales', 'Other_Income', 'Value_Of_Output',
      'Cost_of_Production', 'Selling_Cost', 'PBIDT', 'PBDT', 'PBIT', 'PBT',
      'PAT', 'Adjusted_PAT', 'CP', 'Revenue_earnings_in_forex',
      'Revenue_expenses_in_forex', 'Book_Value_Unit_Curr',
      'Book_Value_Adj_Unit_Curr', 'Market_Capitalisation',
      'CEPS_annualised_Unit_Curr', 'Cash_Flow_From_Operating_Activities',
      'Cash_Flow_From_Investing_Activities',
      'Cash_Flow_From_Financing_Activities', 'ROG_Net_Worth_perc',
      'ROG_Capital_Employed_perc', 'ROG_Gross_Block_perc',
      'ROG_Gross_Sales_perc', 'ROG_Net_Sales_perc',
      'ROG_Cost_of_Production_perc', 'ROG_Total_Assets_perc',
      'ROG_PBIDT_perc', 'ROG_PBDT_perc', 'ROG_PBIT_perc', 'ROG_PBT_perc',
      'ROG_PAT_perc', 'ROG_CP_perc', 'ROG_Market_Capitalisation_perc',
      'Current_RatioLatest', 'Fixed_Assets_RatioLatest',
      'Inventory_RatioLatest', 'Debtors_RatioLatest',
      'Total_Asset_Turnover_RatioLatest', 'Interest_Cover_RatioLatest',
      'PBIDTM_percLatest', 'PBITM_percLatest', 'PBDTM_percLatest',
      'CPM_percLatest', 'APATM_percLatest', 'Debtors_Velocity_Days',
      'Creditors_Velocity_Days', 'Inventory_Velocity_Days',
      'Value_of_Output_by_Total_Assets', 'Value_of_Output_by_Gross_Block',
      'default'],
```

And similar fields will be part of Test data as well.

## Model 1 (With all columns ):

We have used all of the fields listed above and build our first model.

And this is the final result for our first model:

Logit Regression Results			
<b>Dep. Variable:</b>	default	<b>No. Observations:</b>	2402
<b>Model:</b>	Logit	<b>Df Residuals:</b>	2340
<b>Method:</b>	MLE	<b>Df Model:</b>	61
<b>Date:</b>	Tue, 15 Nov 2022	<b>Pseudo R-squ.:</b>	0.6745
<b>Time:</b>	18:07:50	<b>Log-Likelihood:</b>	-268.02
<b>converged:</b>	True	<b>LL-Null:</b>	-823.47

Covariance Type:	nonrobust	LLR p-value: 1.198e-192					
		coef	std err	z	P> z	[0.025	0.975]
	Intercept	-0.1885	0.258	-0.732	0.464	-0.693	0.316
	Equity_Paid_Up	-0.0079	0.014	-0.585	0.559	-0.035	0.019
	Networth	-0.0056	0.005	-1.215	0.224	-0.015	0.003
	Capital_Employed	-0.0124	0.010	-1.294	0.196	-0.031	0.006
	Total_Debt	0.0213	0.007	3.008	0.003	0.007	0.035
	Gross_Block	0.0012	0.004	0.288	0.773	-0.007	0.009
	Net_Working_Capital	-0.0040	0.010	-0.400	0.689	-0.023	0.015
	Current_Assets	0.0053	0.008	0.640	0.522	-0.011	0.022
	Current_Liabilities_and_Provisions	0.0010	0.013	0.081	0.935	-0.024	0.026
	Total_Assets_by_Liabilities	0.0065	0.008	0.795	0.427	-0.010	0.023
	Gross_Sales	-0.0114	0.010	-1.113	0.266	-0.031	0.009
	Net_Sales	0.0203	0.022	0.939	0.348	-0.022	0.063
	Other_Income	-0.0076	0.079	-0.097	0.923	-0.162	0.147
	Value_Of_Output	-0.0188	0.014	-1.348	0.178	-0.046	0.009
	Cost_of_Production	0.0065	0.013	0.510	0.610	-0.018	0.031
	Selling_Cost	-0.0350	0.103	-0.340	0.734	-0.237	0.167
	PBIDT	-0.0177	0.040	-0.438	0.661	-0.097	0.062
	PBDT	-0.1551	0.145	-1.069	0.285	-0.440	0.129
	PBIT	0.0184	0.050	0.368	0.713	-0.080	0.117
	PBT	0.0480	0.212	0.226	0.821	-0.368	0.464
	PAT	-0.0803	0.247	-0.325	0.745	-0.565	0.404
	Adjusted_PAT	0.0038	0.077	0.050	0.960	-0.147	0.155
	CP	0.1585	0.154	1.028	0.304	-0.144	0.461
	Revenue_earnings_in_forex	-0.0277	0.037	-0.754	0.451	-0.100	0.044
	Revenue_expenses_in_forex	0.0533	0.038	1.414	0.157	-0.021	0.127
	Book_Value_Unit_Curr	-0.0215	0.035	-0.622	0.534	-0.089	0.046
	Book_Value_Adj_Unit_Curr	-0.0684	0.036	-1.878	0.060	-0.140	0.003
	Market_Capitalisation	-0.0043	0.004	-1.205	0.228	-0.011	0.003

<b>CEPS_annualised_Unit_Curr</b>	-0.0612	0.050	-1.218	0.223	-0.160	0.037
<b>Cash_Flow_From_Operating_Activities</b>	0.0100	0.027	0.375	0.707	-0.042	0.062
<b>Cash_Flow_From_Investing_Activities</b>	-0.0427	0.051	-0.834	0.404	-0.143	0.058
<b>Cash_Flow_From_Financing_Activities</b>	0.0023	0.043	0.053	0.958	-0.083	0.087
<b>ROG_Net_Worth_perc</b>	-0.0236	0.012	-1.899	0.058	-0.048	0.001
<b>ROG_Capital_Employed_perc</b>	0.0237	0.011	2.224	0.026	0.003	0.045
<b>ROG_Gross_Block_perc</b>	-0.0316	0.021	-1.476	0.140	-0.074	0.010
<b>ROG_Gross_Sales_perc</b>	0.0896	0.120	0.748	0.454	-0.145	0.325
<b>ROG_Net_Sales_perc</b>	-0.0914	0.119	-0.766	0.444	-0.325	0.143
<b>ROG_Cost_of_Production_perc</b>	-0.0060	0.004	-1.432	0.152	-0.014	0.002
<b>ROG_Total_Assets_perc</b>	-0.0256	0.010	-2.519	0.012	-0.046	-0.006
<b>ROG_PBDT_perc</b>	-0.0058	0.006	-1.016	0.310	-0.017	0.005
<b>ROG_PBDT_perc</b>	0.0064	0.006	1.148	0.251	-0.005	0.017
<b>ROG_PBIT_perc</b>	0.0049	0.005	0.954	0.340	-0.005	0.015
<b>ROG_PBT_perc</b>	-0.0021	0.005	-0.435	0.663	-0.012	0.008
<b>ROG_PAT_perc</b>	0.0014	0.004	0.351	0.726	-0.006	0.009
<b>ROG_CP_perc</b>	-0.0042	0.004	-0.928	0.353	-0.013	0.005
<b>ROG_Market_Capitalisation_perc</b>	-0.0024	0.003	-0.799	0.424	-0.008	0.003
<b>Current_RatioLatest</b>	-0.5445	0.094	-5.763	0.000	-0.730	-0.359
<b>Fixed_Assets_RatioLatest</b>	-0.0015	0.112	-0.014	0.989	-0.221	0.218
<b>Inventory_RatioLatest</b>	-0.0678	0.026	-2.561	0.010	-0.120	-0.016
<b>Debtors_RatioLatest</b>	-0.0396	0.026	-1.523	0.128	-0.091	0.011
<b>Total_Asset_Turnover_RatioLatest</b>	-0.0678	0.207	-0.328	0.743	-0.473	0.337
<b>Interest_Cover_RatioLatest</b>	-0.1067	0.053	-2.023	0.043	-0.210	-0.003
<b>PBDTM_percLatest</b>	0.0133	0.032	0.419	0.675	-0.049	0.076
<b>PBITM_percLatest</b>	-0.0783	0.040	-1.943	0.052	-0.157	0.001
<b>PBDTM_percLatest</b>	0.0095	0.054	0.177	0.859	-0.096	0.115
<b>CPM_percLatest</b>	-0.0490	0.068	-0.723	0.470	-0.182	0.084
<b>APATM_percLatest</b>	0.1252	0.069	1.802	0.072	-0.011	0.261
<b>Debtors_Velocity_Days</b>	-0.0042	0.002	-2.771	0.006	-0.007	-0.001

<b>Creditors_Velocity_Days</b>	0.0011	0.002	0.676	0.499	-0.002	0.004
<b>Inventory_Velocity_Days</b>	0.0015	0.002	0.827	0.408	-0.002	0.005
<b>Value_of_Output_by_Total_Assets</b>	0.7860	0.358	2.196	0.028	0.085	1.487
<b>Value_of_Output_by_Gross_Block</b>	-0.0875	0.111	-0.790	0.430	-0.305	0.130

Insights:

1. We can clearly see that results from our 1<sup>st</sup> Model have all 61 columns,,after removing unnecessary fields of Co\_num, next\_year\_networth etc.
2. There are many columns with more than 0.05 P value, which denotes those fields are not required for predicting the Target field.
3. As per initial Heatmap, we already know, that data have multi collinearity in the data, so we don't need all of the columns for predicting target values

## Removing multicollinearity using VIF:

### What is VIF:

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

In ordinary least square (OLS) regression analysis, multicollinearity exists when two or more of the independent variables demonstrate a linear relationship between them. VIF can be calculated by the formula below:

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{\text{Tolerance}}$$

Where **Ri2** represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones. The reciprocal of VIF is known as **tolerance**. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.

### Interpreting the Variance Inflation Factor

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.



Exactly how large a VIF has to be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above.

### Summary:

- Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
- Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.
- VIF measures the number of inflated variances caused by multicollinearity.
- Eliminate fields with Greater than 5 , which means there is highly correlated. We should do it one by one for each field.

### Model 2 with VIF threshold 4:

To Notice, since we have lot of independent fields, we should go and check for each single columns and check VIF multiple times.

Rather we have used a Loop function, its checking VIF in Iterations and remove one column in one iteration and calculate VIF again. This loop will run until we get all the fields with lesser value of VIF than threshold limit.

For this testing we will keep VIF threshold value as 5.

Keep VIF threshold value as 5:

After executing Loop function, we received 34 columns and about half of the fields are eliminated, which had very high co-relation in it.

Remained fields are:

```
Index(['Total_Debt', 'Net_Working_Capital', 'Other_Income', 'Selling_Cost',
      'Adjusted_PAT', 'Revenue_earnings_in_forex',
      'Revenue_expenses_in_forex', 'Book_Value_Adj_Unit_Curr',
      'Market_Capitalisation', 'CEPS_annualised_Unit_Curr',
      'Cash_Flow_From_Operating_Activities',
      'Cash_Flow_From_Investing_Activities',
      'Cash_Flow_From_Financing_Activities', 'ROG_Net_Worth_perc',
      'ROG_Capital_Employed_perc', 'ROG_Gross_Block_perc',
      'ROG_Net_Sales_perc', 'ROG_Cost_of_Production_perc',
      'ROG_Total_Assets_perc', 'ROG_PBIT_perc', 'ROG_CP_perc',
      'ROG_Market_Capitalisation_perc', 'Current_RatioLatest',
      'Inventory_RatioLatest', 'Debtors_RatioLatest',
      'Total_Asset_Turnover_RatioLatest', 'Interest_Cover_RatioLatest',
      'PBITM_percLatest', 'CPM_percLatest', 'Debtors_Velocity_Days',
      'Creditors_Velocity_Days', 'Inventory_Velocity_Days',
      'Value_of_Output_by_Gross_Block', 'default'],
```

We ran the stats Model on above listed field, and this is the Stats summary for same:

<b>Dep. Variable:</b>	default	<b>No. Observations:</b>	2402				
<b>Model:</b>	Logit	<b>Df Residuals:</b>	2368				
<b>Method:</b>	MLE	<b>Df Model:</b>	33				
<b>Date:</b>	Tue, 15 Nov 2022	<b>Pseudo R-squ.:</b>	0.6546				
<b>Time:</b>	18:08:40	<b>Log-Likelihood:</b>	-284.41				
<b>converged:</b>	True	<b>LL-Null:</b>	-823.47				
<b>Covariance Type:</b>	nonrobust	<b>LLR p-value:</b>	3.364e-205				
		<b>coef</b>	<b>std err</b>	<b>z</b>	<b>P&gt; z </b>	<b>[0.025</b>	<b>0.975]</b>
<b>Intercept</b>		-0.2422	0.226	-1.072	0.284	-0.685	0.201
<b>Total_Debt</b>		0.0157	0.004	3.831	0.000	0.008	0.024
<b>Net_Working_Capital</b>		-0.0067	0.004	-1.643	0.100	-0.015	0.001
<b>Other_Income</b>		-0.0227	0.066	-0.341	0.733	-0.153	0.108
<b>Selling_Cost</b>		-0.0587	0.081	-0.725	0.469	-0.218	0.100
<b>Adjusted_PAT</b>		-0.0126	0.049	-0.259	0.796	-0.108	0.083
<b>Revenue_earnings_in_forex</b>		-0.0358	0.033	-1.084	0.278	-0.101	0.029
<b>Revenue_expenses_in_forex</b>		0.0480	0.034	1.433	0.152	-0.018	0.114
<b>Book_Value_Adj_Unit_Curr</b>		-0.1004	0.010	-10.195	0.000	-0.120	-0.081
<b>Market_Capitalisation</b>		-0.0081	0.003	-3.015	0.003	-0.013	-0.003
<b>CEPS_annualised_Unit_Curr</b>		-0.0670	0.037	-1.820	0.069	-0.139	0.005
<b>Cash_Flow_From_Operating_Activities</b>		0.0037	0.022	0.169	0.866	-0.039	0.047
<b>Cash_Flow_From_Investing_Activities</b>		-0.0143	0.041	-0.346	0.730	-0.095	0.067
<b>Cash_Flow_From_Financing_Activities</b>		0.0071	0.038	0.186	0.853	-0.067	0.082
<b>ROG_Net_Worth_perc</b>		-0.0267	0.012	-2.317	0.021	-0.049	-0.004
<b>ROG_Capital_Employed_perc</b>		0.0220	0.010	2.230	0.026	0.003	0.041

<b>ROG_Gross_Block_perc</b>	-0.0308	0.020	-1.512	0.131	-0.071	0.009
<b>ROG_Net_Sales_perc</b>	-0.0020	0.004	-0.474	0.636	-0.010	0.006
<b>ROG_Cost_of_Production_perc</b>	-0.0062	0.004	-1.551	0.121	-0.014	0.002
<b>ROG_Total_Assets_perc</b>	-0.0216	0.010	-2.231	0.026	-0.041	-0.003
<b>ROG_PBIT_perc</b>	0.0028	0.002	1.188	0.235	-0.002	0.008
<b>ROG_CP_perc</b>	-0.0016	0.002	-0.716	0.474	-0.006	0.003
<b>ROG_Market_Capitalisation_perc</b>	-0.0008	0.003	-0.290	0.772	-0.006	0.005
<b>Current_RatioLatest</b>	-0.5418	0.091	-5.931	0.000	-0.721	-0.363
<b>Inventory_RatioLatest</b>	-0.0537	0.023	-2.382	0.017	-0.098	-0.010
<b>Debtors_RatioLatest</b>	-0.0274	0.023	-1.186	0.236	-0.073	0.018
<b>Total_Asset_Turnover_RatioLatest</b>	0.2183	0.132	1.648	0.099	-0.041	0.478
<b>Interest_Cover_RatioLatest</b>	-0.0728	0.047	-1.547	0.122	-0.165	0.019
<b>PBITM_percLatest</b>	-0.0307	0.015	-2.040	0.041	-0.060	-0.001
<b>CPM_percLatest</b>	0.0044	0.017	0.251	0.802	-0.030	0.038
<b>Debtors_Velocity_Days</b>	-0.0042	0.001	-2.909	0.004	-0.007	-0.001
<b>Creditors_Velocity_Days</b>	0.0011	0.002	0.735	0.462	-0.002	0.004
<b>Inventory_Velocity_Days</b>	0.0011	0.002	0.689	0.491	-0.002	0.004
<b>Value_of_Output_by_Gross_Block</b>	-0.0515	0.044	-1.157	0.247	-0.139	0.036

Still we see lot of Columns have P value more than 0.05, so lets run VIF function one more time, and eliminate fields , which have VIF more than 4

### Model 3 with VIF threshold 4:

After adjusting threshold limit to 4 , we were successfully able to eliminate 3 more fields. Which means we have now below listed fields for our Third model:

```
Index(['Total_Debt', 'Net_Working_Capital', 'Other_Income', 'Adjusted_PAT',
      'Revenue_earnings_in_forex', 'Revenue_expenses_in_forex',
      'Book_Value_Adj_Unit_Curr', 'Market_Capitalisation',
      'Cash_Flow_From_Operating_Activities',
      'Cash_Flow_From_Investing_Activities',
      'Cash_Flow_From_Financing_Activities', 'ROG_Net_Worth_perc',
      'ROG_Capital_Employed_perc', 'ROG_Gross_Block_perc',
      'ROG_Net_Sales_perc', 'ROG_Cost_of_Production_perc',
      'ROG_Total_Assets_perc', 'ROG_PBIT_perc', 'ROG_CP_perc',
      'ROG_Market_Capitalisation_perc', 'Current_RatioLatest',
      'Inventory_RatioLatest', 'Debtors_RatioLatest',
      'Total_Asset_Turnover_RatioLatest', 'Interest_Cover_RatioLatest',
      'CPM_percLatest', 'Debtors_Velocity_Days', 'Creditors_Velocity_Days',
      'Inventory_Velocity_Days', 'Value_of_Output_by_Gross_Block', 'default'])
```

Again, run the stats model on top of above listed columns.

Dep. Variable:	default	No. Observations:	2402				
Model:	Logit	Df Residuals:	2371				
Method:	MLE	Df Model:	30				
Date:	Tue, 15 Nov 2022	Pseudo R-squ.:	0.6501				
Time:	18:08:42	Log-Likelihood:	-288.13				
converged:	True	LL-Null:	-823.47				
Covariance Type:	nonrobust	LLR p-value:	5.990e-206				
		coef	std err	z	P> z	[0.025	0.975]
Intercept		-0.2233	0.225	-0.993	0.321	-0.664	0.218
Total_Debt		0.0143	0.004	3.595	0.000	0.006	0.022
Net_Working_Capital		-0.0064	0.004	-1.687	0.092	-0.014	0.001
Other_Income		-0.0546	0.065	-0.843	0.399	-0.182	0.072
Adjusted_PAT		-0.0448	0.044	-1.021	0.307	-0.131	0.041
Revenue_earnings_in_forex		-0.0454	0.031	-1.457	0.145	-0.106	0.016

Revenue_expenses_in_forex	0.0477	0.033	1.463	0.143	-0.016	0.112
Book_Value_Adj_Unit_Curr	-0.1006	0.010	-10.439	0.000	-0.120	-0.082
Market_Capitalisation	-0.0079	0.003	-3.007	0.003	-0.013	-0.003
Cash_Flow_From_Operating_Activities	-0.0026	0.022	-0.119	0.905	-0.045	0.040
Cash_Flow_From_Investing_Activities	-0.0153	0.039	-0.388	0.698	-0.092	0.062
Cash_Flow_From_Financing_Activities	0.0049	0.037	0.132	0.895	-0.068	0.078
ROG_Net_Worth_perc	-0.0284	0.011	-2.531	0.011	-0.050	-0.006
ROG_Capital_Employed_perc	0.0197	0.010	2.033	0.042	0.001	0.039
ROG_Gross_Block_perc	-0.0265	0.020	-1.310	0.190	-0.066	0.013
ROG_Net_Sales_perc	-0.0022	0.004	-0.537	0.591	-0.010	0.006
ROG_Cost_of_Production_perc	-0.0060	0.004	-1.519	0.129	-0.014	0.002
ROG_Total_Assets_perc	-0.0229	0.010	-2.372	0.018	-0.042	-0.004
ROG_PBIT_perc	0.0025	0.002	1.049	0.294	-0.002	0.007
ROG_CP_perc	-0.0017	0.002	-0.769	0.442	-0.006	0.003
ROG_Market_Capitalisation_perc	-0.0008	0.003	-0.304	0.761	-0.006	0.005
Current_RatioLatest	-0.5498	0.091	-6.054	0.000	-0.728	-0.372
Inventory_RatioLatest	-0.0503	0.023	-2.224	0.026	-0.095	-0.006
Debtors_RatioLatest	-0.0281	0.023	-1.224	0.221	-0.073	0.017
Total_Asset_Turnover_RatioLatest	0.1895	0.130	1.458	0.145	-0.065	0.444
Interest_Cover_RatioLatest	-0.0896	0.047	-1.902	0.057	-0.182	0.003
CPM_percLatest	-0.0253	0.011	-2.345	0.019	-0.046	-0.004
Debtors_Velocity_Days	-0.0037	0.001	-2.616	0.009	-0.006	-0.001
Creditors_Velocity_Days	0.0010	0.001	0.689	0.491	-0.002	0.004
Inventory_Velocity_Days	0.0011	0.002	0.659	0.510	-0.002	0.004
Value_of_Output_by_Gross_Block	-0.0611	0.044	-1.392	0.164	-0.147	0.025

Check for VIF for all the fields in this model :

	<b>variables</b>	<b>VIF</b>
<b>8</b>	Cash_Flow_From_Operating_Activities	3.804148
<b>7</b>	Market_Capitalisation	3.769023
<b>12</b>	ROG_Capital_Employed_perc	3.671089
<b>0</b>	Total_Debt	3.620201
<b>2</b>	Other_Income	3.483271
<b>5</b>	Revenue_expenses_in_forex	3.418506
<b>18</b>	ROG_CP_perc	3.262816
<b>17</b>	ROG_PBIT_perc	3.251599
<b>23</b>	Total_Asset_Turnover_RatioLatest	3.135009
<b>16</b>	ROG_Total_Assets_perc	3.108078
<b>3</b>	Adjusted_PAT	3.033204
<b>1</b>	Net_Working_Capital	2.902452
<b>4</b>	Revenue_earnings_in_forex	2.872264
<b>10</b>	Cash_Flow_From_Financing_Activities	2.863477
<b>11</b>	ROG_Net_Worth_perc	2.719630
<b>9</b>	Cash_Flow_From_Investing_Activities	2.503964
<b>6</b>	Book_Value_Adj_Unit_Curr	2.460811
<b>22</b>	Debtors_RatioLatest	2.346833
<b>26</b>	Debtors_Velocity_Days	2.345360
<b>29</b>	Value_of_Output_by_Gross_Block	2.316941
<b>24</b>	Interest_Cover_RatioLatest	2.250516
<b>21</b>	Inventory_RatioLatest	2.221555
<b>27</b>	Creditors_Velocity_Days	2.180764

	variables	VIF
20	Current_RatioLatest	2.175016
14	ROG_Net_Sales_perc	2.081818
15	ROG_Cost_of_Production_perc	1.995917
28	Inventory_Velocity_Days	1.860591
25	CPM_perLatest	1.744019
19	ROG_Market_Capitalisation_perc	1.621405
13	ROG_Gross_Block_perc	1.477094
30	default	1.397951

We can see that all of the fields have less than 4 VIF value for this.

### ***Validating the model on train set***

Confusion Matrix for train data:

```
[[2112   30]
 [  74  186]]
```

Calculate Recall and other matrix:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2142
1	0.86	0.72	0.78	260
accuracy			0.96	2402
macro avg	0.91	0.85	0.88	2402
weighted avg	0.95	0.96	0.95	2402

## Validating the model on test set

Confusion Matrix for train data:

```
[[1035   21]
 [   38   90]]
```

Calculate Recall and other matrix:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	1056
1	0.81	0.70	0.75	128
accuracy			0.95	1184
macro avg	0.89	0.84	0.86	1184
weighted avg	0.95	0.95	0.95	1184

Insights:

1. We are more interested in RECALL value as it denotes, how much “default” value or Targeted value we are predicting correctly.
2. Recall value for train data and test data are 78% and 75% in sequence, which is not bad
3. Still data is imbalanced, because our targeted values are very less in Sample data.

## Model 4 - Balance data using SMOTE and threshold VIF as 4:

We will be using SMOTE functionality to generate data for target values, and we will upscaling the data for this case.

Steps:

1. Again, we will perform Smote on full sample data , so total fields are 62
2. Upscale data using imblearn.over\_sampling library
3. We also run VIF function to check best features, which can be used for Stats model, after executing VIF calculation function, we found 31 fields and we kept threshold limit as 4 for eliminating features with High co-relation.
4. We generated the Stats model technique and following is the result for Stats model:

<b>Dep. Variable:</b>	default	<b>No. Observations:</b>	3748
<b>Model:</b>	Logit	<b>Df Residuals:</b>	3717
<b>Method:</b>	MLE	<b>Df Model:</b>	30
<b>Date:</b>	Tue, 15 Nov 2022	<b>Pseudo R-squ.:</b>	0.7452
<b>Time:</b>	18:08:45	<b>Log-Likelihood:</b>	-652.27



<b>converged:</b>	True	<b>LL-Null:</b> -2559.5					
<b>Covariance Type:</b>	nonrobust	<b>LLR p-value:</b> 0.000					
		<b>coef</b>	<b>std err</b>	<b>z</b>	<b>P&gt; z </b>	<b>[0.025</b>	<b>0.975]</b>
	<b>Intercept</b>	1.6931	0.162	10.462	0.000	1.376	2.010
	<b>Total_Debt</b>	0.0240	0.003	7.305	0.000	0.018	0.030
	<b>Net_Working_Capital</b>	-0.0053	0.003	-1.782	0.075	-0.011	0.001
	<b>Other_Income</b>	-0.1006	0.050	-2.028	0.043	-0.198	-0.003
	<b>Adjusted_PAT</b>	-0.0177	0.029	-0.602	0.547	-0.075	0.040
	<b>Revenue_earnings_in_forex</b>	-0.0699	0.024	-2.966	0.003	-0.116	-0.024
	<b>Revenue_expenses_in_forex</b>	0.0389	0.024	1.609	0.108	-0.008	0.086
	<b>Book_Value_Adj_Unit_Curr</b>	-0.1220	0.007	-17.710	0.000	-0.135	-0.108
	<b>Market_Capitalisation</b>	-0.0097	0.002	-4.788	0.000	-0.014	-0.006
	<b>Cash_Flow_From_Operating_Activities</b>	-0.0262	0.016	-1.620	0.105	-0.058	0.006
	<b>Cash_Flow_From_Investing_Activities</b>	-0.0400	0.031	-1.305	0.192	-0.100	0.020
	<b>Cash_Flow_From_Financing_Activities</b>	-0.0079	0.028	-0.283	0.777	-0.062	0.047
	<b>ROG_Net_Worth_perc</b>	-0.0340	0.008	-4.402	0.000	-0.049	-0.019
	<b>ROG_Capital_Employed_perc</b>	0.0231	0.007	3.354	0.001	0.010	0.037
	<b>ROG_Gross_Block_perc</b>	-0.0330	0.015	-2.201	0.028	-0.062	-0.004
	<b>ROG_Net_Sales_perc</b>	-0.0058	0.003	-1.793	0.073	-0.012	0.001
	<b>ROG_Cost_of_Production_perc</b>	-0.0070	0.003	-2.447	0.014	-0.013	-0.001
	<b>ROG_Total_Assets_perc</b>	-0.0314	0.007	-4.631	0.000	-0.045	-0.018
	<b>ROG_PBIT_perc</b>	0.0043	0.002	2.548	0.011	0.001	0.008
	<b>ROG_CP_perc</b>	-0.0021	0.002	-1.300	0.194	-0.005	0.001
	<b>ROG_Market_Capitalisation_perc</b>	-0.0027	0.002	-1.358	0.174	-0.007	0.001
	<b>Current_RatioLatest</b>	-0.6289	0.056	-11.244	0.000	-0.739	-0.519
	<b>Inventory_RatioLatest</b>	-0.0543	0.014	-3.784	0.000	-0.082	-0.026
	<b>Debtors_RatioLatest</b>	-0.0576	0.016	-3.514	0.000	-0.090	-0.025
	<b>Total_Asset_Turnover_RatioLatest</b>	0.3287	0.089	3.707	0.000	0.155	0.503
	<b>Interest_Cover_RatioLatest</b>	-0.0856	0.034	-2.515	0.012	-0.152	-0.019
	<b>CPM_percLatest</b>	-0.0507	0.008	-6.185	0.000	-0.067	-0.035

<b>Debtors_Velocity_Days</b>	-0.0059	0.001	-5.662	0.000	-0.008	-0.004
<b>Creditors_Velocity_Days</b>	0.0010	0.001	0.909	0.363	-0.001	0.003
<b>Inventory_Velocity_Days</b>	0.0005	0.001	0.414	0.679	-0.002	0.003
<b>Value_of_Output_by_Gross_Block</b>	-0.0945	0.028	-3.384	0.001	-0.149	-0.040

### ***Validating the model on train set***

Confusion Matrix for train data:

```
[[2009  133]
 [  84 1522]]
```

Calculate Recall and other matrix:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	2142
1	0.92	0.95	0.93	1606
accuracy			0.94	3748
macro avg	0.94	0.94	0.94	3748
weighted avg	0.94	0.94	0.94	3748

### ***Validating the model on test set***

Confusion Matrix for train data:

```
[[972  84]
 [ 11 117]]
```

Calculate Recall and other matrix:

	precision	recall	f1-score	support
0	0.99	0.92	0.95	1056
1	0.58	0.91	0.71	128
accuracy			0.92	1184
macro avg	0.79	0.92	0.83	1184
weighted avg	0.94	0.92	0.93	1184

Insights:

1. We are more interested in RECALL value as it denotes, how much “default” value or Targeted value we are predicting correctly.
2. Recall value for train data and test data are 95% and 91% in sequence, which is very good.
3. Accuracy on Train and Test data is also very good which is more than 92% with both train and test data.
4. Precision is Bad after SMOTE is implemented, because data was imbalanced, and we had to upscale data using SMOTE.

## Conclusion and Recommendations:

1. Sample data had 3586 rows and 67 columns; total number of column values were not bad.
2. We have not given Company types, like Mid-Size, Small Size or Large capital companies.
3. We have not given revenue of the companies , so that we can segregate companies based on their revenue
4. We have not given industry type of companies like, IT, Manufacturing, Retail , Pharma, etc . We consider them all of same type and not taken any action based on what type company it is
5. Data have about 100+ NULL values in sample data, which we imputed by taking median.
6. Data was very large number of outliers , and for our modelling purpose, we treated outliers, using bringing them back to Normal Upper and Lower limits
7. Data was very imbalanced , and we upscale data using SMOTE feature
8. In the Final model , we have used 31 columns for deciding “DEFAULT” nature of companies, and there were total 67 features, were given.
9. There are other method also available for classification model, but for this exercise purpose, we have used only stats model .
10. NULL Values and Outlier treatment:

Following methods were used in this exercise for treating NULL and outlier values:

- a. Impute NULL with Median and Impute Outlier by bringing them back to Normal range. This Method is used for this project submission, because it has the best RECALL and Accuracy score.
- b. I have also tested NULL values and Outlier treatment with KNN method of data imputations, but recall values and precision are very less in this case , so I switch back to above point a only . Following attached is notebook, where I have tested using KNN model as well, but RECALL values are very less. This is just for reference purpose.:



**Amit\_Jain\_Milestone-1\_KNN\_Treatment.**

This is a Notebook file, in this Notebook, I have tested KNN treatment and calculated RECALL after KNN treatment of Missing values and Outliers.