Filed by Aditya Rishi

January 9, 2022, PGP-DSBA-Feb 21

Financial Risk Analytics (FRA), Project for Great Learning

FRA Milestone 1

Tables and Figures

Problem set

Problem description: Executive summary

Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the net worth of the company in the following year (2016) is provided which can be used to drive the labelled field. The aim is 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.

Usage

Default

Format

A dataframe with 3587 observations on 68 variables

Introduction

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

'Credit Default Data Dictionary.xlsx

Data file: Company_Data2015-1.xlsx

Source:-

Great Learning

Data dictionary

	#	Field Name	Description	New Field Name
0	1	Co_Code	Company Code	Co_Code
1	2	Co_Name	Company Name	Co_Name
2	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
3	4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders	Equity_Paid_Up
4	5	Networth	Value of a company as on 2015 - Current Year	Networth
5	6	Capital Employed	Total amount of capital used for the acquisition of profits by a company	Capital_Employed
6	7	Total Debt	The sum of money borrowed by the company and is due to be paid	Total_Debt
7	8	Gross Block	Total value of all of the assets that a company owns	Gross_Block
8	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
9	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
10	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
11	12	Total Assets/Liabilities	Ratio of total assets to liabailities of the company	Total_Assets_to_Liab
12	13	Gross Sales	The grand total of sale transactions within the accounting period	Gross_Sales
13	14	Net Sales	Gross sales minus returns, allowances, and discounts	Net_Sales
14	15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)	Other_Income
15	16	Value Of Output	Product of physical output of goods and services produced by company and its market price	Value_Of_Output
16	17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service	Cost_of_Prod
17	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 ...	Selling_Cost
18	19	PBIDT	Profit Before Interest, Depreciation & Taxes	PBIDT
19	20	PBDT	Profit Before Depreciation and Tax	PBDT
20	21	PBIT	Profit before interest and taxes	PBIT
21	22	PBT	Profit before tax	PBT
22	23	PAT	Profit After Tax	PAT
23	24	Adjusted PAT	Adjusted profit is the best estimate of the true profit	Adjusted_PAT
24	26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.	CP
25	27	Revenue earnings in forex	Revenue earned in foreign currency	Rev_earn_in_forex
26	28	Revenue expenses in forex	Expenses due to foreign currency transactions	Rev_exp_in_forex
27	29	Capital expenses in forex	Long term investment in forex	Capital_exp_in_forex
28	30	Book Value (Unit Curr)	Net asset value	Book_Value_Unit_Curr
29	31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value	Book_Value_Adj_Unit_Curr
30	32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share	Market_Capitalisation
31	33	CEPS (annualised) (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
32	34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities	Cash_Flow_From_Opr
33	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
34	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
35	37	ROG-Net Worth (%)	Rate of Growth - Networth	ROG_Net_Worth_perc
36	38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed	ROG_Capital_Employed_perc
37	39	ROG-Gross Block (%)	Rate of Growth - Gross Block	ROG_Gross_Block_perc
38	40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales	ROG_Gross_Sales_perc

Tab 1

Data dictionary

39	41	ROG-Net Sales (%)	Rate of Growth - Net Sales	ROG_Net_Sales_perc
40	42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production	ROG_Cost_of_Prod_perc
41	43	ROG-Total Assets (%)	Rate of Growth - Total Assets	ROG_Total_Assets_perc
42	44	ROG-PBIDT (%)	Rate of Growth- PBIDT	ROG_PBIDT_perc
43	45	ROG-PBDT (%)	Rate of Growth- PBDT	ROG_PBDT_perc
44	46	ROG-PBIT (%)	Rate of Growth- PBIT	ROG_PBIT_perc
45	47	ROG-PBT (%)	Rate of Growth- PBT	ROG_PBT_perc
46	48	ROG-PAT (%)	Rate of Growth- PAT	ROG_PAT_perc
47	49	ROG-CP (%)	Rate of Growth- CP	ROG_CP_perc
48	50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex	ROG_Rev_earn_in_forex_perc
49	51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex	ROG_Rev_exp_in_forex_perc
50	52	ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation	ROG_Market_Capitalisation_perc
51	53	Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year	Curr_Ratio_Latest
52	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
53	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
54	56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company	Debtors_Ratio_Latest
55	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
56	58	Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt	Interest_Cover_Ratio_Latest
57	59	PBIDTM (%)[Latest]	Profit before Interest Depreciation and Tax Margin	PBIDTM_perc_Latest
58	60	PBITM (%)[Latest]	Profit Before Interest Tax Margin	PBITM_perc_Latest
59	61	PBDTM (%)[Latest]	Profit Before Depreciation Tax Margin	PBDTM_perc_Latest
60	62	CPM (%)[Latest]	Cost per thousand (advertising cost)	CPM_perc_Latest
61	63	APATM (%)[Latest]	After tax profit margin	APATM_perc_Latest
62	64	Debtors Velocity (Days)	Average days required for receiving the payments	Debtors_Vel_Days
63	65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers	Creditors_Vel_Days
64	66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales	Inventory_Vel_Days
65	67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets	Value_of_Output_to_Total_Assets
66	68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block	Value_of_Output_to_Gross_Block

Tab 2

Descriptive stats

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586.00	16065.39	19776.82	4.00	3029.25	6077.50	24269.50	72493.00
Networth_Next_Year	3586.00	725.05	4769.68	-8021.60	3.98	19.02	123.80	111729.10
Equity_Paid_Up	3586.00	62.97	778.76	0.00	3.75	8.29	19.52	42263.46
Networth	3586.00	649.75	4091.99	-7027.48	3.89	18.58	117.30	81657.35
Capital_Employed	3586.00	2799.61	26975.14	-1824.75	7.60	39.09	226.61	714001.25
Total_Debt	3586.00	1994.82	23652.84	-0.72	0.03	7.49	72.35	652823.81
Gross_Block_	3586.00	594.18	4871.55	-41.19	0.57	15.87	131.90	128477.59
Net_Working_Capital_	3586.00	410.81	6301.22	-13162.42	0.94	10.14	61.17	223257.56
Current_Assets_	3586.00	1960.35	22577.57	-0.91	4.00	24.54	135.28	721166.00
Current_Liabilities_and_Provisions_	3586.00	391.99	2675.00	-0.23	0.73	9.23	65.65	83232.98
Total_Assets_to_Liabilities_	3586.00	1778.45	11437.57	-4.51	10.55	52.01	310.54	254737.22
Gross_Sales	3586.00	1123.74	10603.70	-62.59	1.44	31.21	242.25	474182.94
Net_Sales	3586.00	1079.70	9996.57	-62.59	1.44	30.44	234.44	443775.16
Other_Income	3586.00	48.73	426.04	-448.72	0.02	0.45	3.63	14143.40
Value_Of_Output	3586.00	1077.19	9843.88	-119.10	1.41	30.89	235.84	435559.09
Cost_of_Production	3586.00	798.54	9076.70	-22.65	0.94	25.99	189.55	419913.50
Selling_Cost	3586.00	25.55	194.24	0.00	0.00	0.16	3.88	5283.91
PBIDT	3586.00	248.18	1949.59	-4655.14	0.04	2.04	23.52	42059.26
PBDT	3586.00	116.27	956.20	-5874.53	0.00	0.80	12.95	23215.00
PBIT	3586.00	217.66	1850.97	-4812.95	0.00	1.15	16.67	41402.96
PBT	3586.00	85.75	799.93	-6032.34	-0.06	0.31	7.42	16798.00
PAT	3586.00	61.22	620.30	-6032.34	-0.06	0.26	5.54	13383.39
Adjusted_PAT	3586.00	60.06	580.43	-4418.72	-0.09	0.21	5.34	13384.11
CP	3586.00	91.73	780.79	-5874.53	0.00	0.74	10.91	20760.20
Revenue_earnings_in_forex	3586.00	131.17	1150.73	0.00	0.00	0.00	7.20	46158.00
Revenue_expenses_in_forex	3586.00	256.33	4132.34	0.00	0.00	0.00	6.99	193979.73
Capital_expenses_in_forex	3586.00	7.66	111.43	0.00	0.00	0.00	0.00	3722.10
Book_Value_Unit_Curr	3586.00	157.24	1622.66	-3371.57	7.96	21.66	71.67	75790.00
Book_Value_Adj._Unit_Curr	3582.00	2243.15	128283.73	-33715.70	7.06	18.93	60.01	7677600.29
Market_Capitalisation	3586.00	1664.09	12805.17	0.00	0.00	8.37	111.46	260865.08
CEPS_annualised_Unit_Curr	3586.00	36.02	828.42	-1808.00	0.00	1.15	8.77	45438.44
Cash_Flow_From_Operating_Activities	3586.00	65.77	1455.05	-25469.23	-0.31	0.45	12.65	44529.40
Cash_Flow_From_Investing_Activities	3586.00	-60.87	701.97	-23843.45	-5.12	-0.12	0.12	3732.98
Cash_Flow_From_Financing_Activities	3586.00	11.44	1272.26	-38374.04	-5.85	0.00	0.46	28846.00
ROG-Net_Worth_perc	3586.00	1237.62	41041.93	-14485.71	-1.49	1.84	11.36	2144020.00
ROG-Capital_Employed_perc	3586.00	2988.88	126472.87	-8614.63	-3.83	1.38	12.59	7412700.00
ROG-Gross_Block_perc	3586.00	37.55	893.62	-116.12	0.00	0.25	6.72	47400.00
ROG-Gross_Sales_perc	3586.00	242.67	6103.53	-5503.70	-8.08	3.31	21.53	320200.00
ROG-Net_Sales_perc	3586.00	242.59	6103.49	-5503.70	-8.12	3.21	21.57	320200.00
ROG-Cost_of_Production_perc	3586.00	310.49	5573.22	-2130.23	-7.24	4.42	23.12	267150.00
ROG-Total_Assets_perc	3586.00	2793.28	125941.65	-136.13	-3.97	1.48	12.50	7422120.00
ROG-PBIDT_perc	3586.00	375.85	23278.40	-52200.00	-23.36	4.57	47.88	1386200.00
ROG-PBDT_perc	3586.00	336.38	20353.40	-52200.00	-30.60	3.37	52.91	1208700.00

Tab 3

Descriptive stats

ROG-PAT_perc	3586.00	112.23	13480.52	-114500.00	-43.73	0.00	65.35	774200.00
ROG-CP_perc	3586.00	221.09	13980.20	-52200.00	-29.50	4.62	52.91	822400.00
ROG-Revenue_earnings_in_forex_perc	3586.00	37.23	658.67	-100.00	0.00	0.00	0.00	29084.77
ROG-Revenue_expenses_in_forex_perc	3586.00	364.86	15233.64	-100.00	0.00	0.00	0.00	894591.69
ROG-Market_Capitalisation_perc	3586.00	63.68	1047.93	-98.05	0.00	0.00	47.52	61865.26
Current_Ratio[Latest]	3585.00	12.06	108.41	0.00	0.88	1.36	2.77	4813.00
Fixed_Assets_Ratio[Latest]	3585.00	51.54	681.15	0.00	0.27	1.56	4.74	22172.00
Inventory_Ratio[Latest]	3585.00	37.80	458.19	0.00	0.00	3.56	8.94	15472.00
Debtors_Ratio[Latest]	3585.00	33.03	489.56	0.00	0.42	3.82	8.52	22992.67
Total_Asset_Turnover_Ratio[Latest]	3585.00	1.24	2.67	0.00	0.07	0.60	1.55	57.75
Interest_Cover_Ratio[Latest]	3585.00	16.39	351.74	-5450.00	0.00	1.08	3.71	18639.40
PBIDTM_perc[Latest]	3585.00	-51.16	1795.13	-78870.45	0.00	8.07	18.99	19233.33
PBITM_perc[Latest]	3585.00	-109.21	3057.64	-141600.00	0.00	5.23	14.29	19195.70
PBDTM_perc[Latest]	3585.00	-311.57	10921.59	-590500.00	0.00	4.69	14.11	15640.00
CPM_perc[Latest]	3585.00	-307.01	10676.15	-572000.00	0.00	3.89	11.39	15640.00
APATM_perc[Latest]	3585.00	-365.06	12500.05	-688600.00	0.00	1.59	7.41	15266.67
Debtors_Velocity_Days	3586.00	603.89	10636.76	0.00	8.00	49.00	106.00	514721.00
Creditors_Velocity_Days	3586.00	2057.85	54169.48	0.00	8.00	39.00	89.00	2034145.00
Inventory_Velocity_Days	3483.00	79.64	137.85	-199.00	0.00	35.00	96.00	996.00
Value_of_Output_to_Total_Assets	3586.00	0.82	1.20	-0.33	0.07	0.48	1.16	17.63
Value_of_Output_to_Gross_Block	3586.00	61.88	976.82	-61.00	0.27	1.53	4.91	43404.00
default	3586.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00

Checking the dataset (top 5 rows) again after fixing messy names¶

	Co_Code	Co_Name	Networth_Next_Year	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block_	Net_Working_Capital_	Current_Assets_
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81

Checking the number of rows (observations) and columns (variables)

```
The number of rows (observations) is 3586
The number of columns (variables) is 68
```

Tab 4 and 5

Checking datatype of all columns

```
RangeIndex: 3586 entries, 0 to 3585
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_to_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
```

Tab 6

Checking datatype of all columns

```
 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-PBIDT_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_Ratio[Latest]                  3585 non-null   float64
 52  Fixed_Assets_Ratio[Latest]             3585 non-null   float64
 53  Inventory_Ratio[Latest]                3585 non-null   float64
 54  Debtors_Ratio[Latest]                  3585 non-null   float64
 55  Total_Asset_Turnover_Ratio[Latest]     3585 non-null   float64
 56  Interest_Cover_Ratio[Latest]           3585 non-null   float64
 57  PBIDTM_perc[Latest]                    3585 non-null   float64
 58  PBITM_perc[Latest]                     3585 non-null   float64
 59  PBDTM_perc[Latest]                     3585 non-null   float64
 60  CPM_perc[Latest]                       3585 non-null   float64
 61  APATM_perc[Latest]                     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_to_Total_Assets        3586 non-null   float64
 66  Value_of_Output_to_Gross_Block         3586 non-null   float64
 67  default                                3586 non-null   int32
dtypes: float64(63), int32(1), int64(3), object(1)
memory usage: 1.8+ MB
```

Tab 7

Checking for duplicate values

```
The credit risk dataset has 0 duplicate values
```

Check the basic measures of descriptive statistics for the continuous variables once again (taking just 2 decimal places)

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586.00	16065.39	19776.82	4.00	3029.25	6077.50	24269.50	72493.00
Networth_Next_Year	3586.00	725.05	4769.68	-8021.60	3.98	19.02	123.80	111729.10
Equity_Paid_Up	3586.00	62.97	778.76	0.00	3.75	8.29	19.52	42263.46
Networth	3586.00	649.75	4091.99	-7027.48	3.89	18.58	117.30	81657.35
Capital_Employed	3586.00	2799.61	26975.14	-1824.75	7.60	39.09	226.61	714001.25
Total_Debt	3586.00	1994.82	23652.84	-0.72	0.03	7.49	72.35	652823.81
Gross_Block_	3586.00	594.18	4871.55	-41.19	0.57	15.87	131.90	128477.59
Net_Working_Capital_	3586.00	410.81	6301.22	-13162.42	0.94	10.14	61.17	223257.56

Creating a binary target variable using 'Networth_Next_Year',

Checking top 10 rows

	default	Networth_Next_Year
0	1	-8021.60
1	1	-3986.19
2	1	-3192.58
3	1	-3054.51
4	1	-2967.36
5	1	-2519.40
6	1	-2125.05
7	1	-2100.56
8	1	-1695.75
9	1	-1677.18

Tab 8 and 9

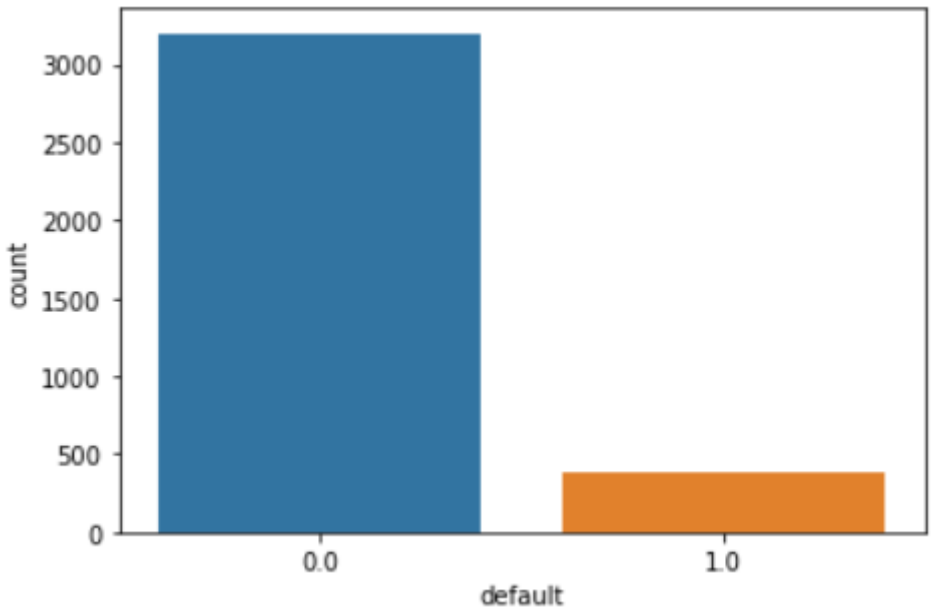
What does variable 'default' look like

```
0    3198
1     388
Name: default, dtype: int64
```

Checking proportion of default

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

Split of default variable



Default is way less than non-default, which is good for the company that provided credit.

Fig 1

Outlier Analysis

Let's check the number of outliers per column

APATM_perc[Latest]	933
Adjusted_PAT	954
Book_Value_Adj._Unit_Curr	486
Book_Value_Unit_Curr	485
CEPS_annualised_Unit_Curr	602
CP	816
CPM_perc[Latest]	720
Capital_Employed	596
Capital_expenses_in_forex	694
Cash_Flow_From_Financing_Activities	1005
Cash_Flow_From_Investing_Activities	876
Cash_Flow_From_Operating_Activities	801
Co_Code	291
Co_Name	0
Cost_of_Production	560
Creditors_Velocity_Days	391
Current_Assets_	577
Current_Liabilities_and_Provisions_	581
Current_Ratio[Latest]	565
Debtors_Ratio[Latest]	371
Debtors_Velocity_Days	398
Equity_Paid_Up	448
Fixed_Assets_Ratio[Latest]	495
Gross_Block_	540
Gross_Sales	554
Interest_Cover_Ratio[Latest]	725
Inventory_Ratio[Latest]	375
Inventory_Velocity_Days	262
Market_Capitalisation	639
Net_Sales	556
Net_Working_Capital_	625

Logistic regression models are not as much impacted due to the presence of outliers as the linear regression modes because the sigmoid function tapers the outliers. However, the presence of extreme outliers may affect the model's performance for some cases. The presence of the outliers may also affect the prediction probability of the data points. The outliers in the present dataset are **extreme outliers**, which lie beyond

Tab 10

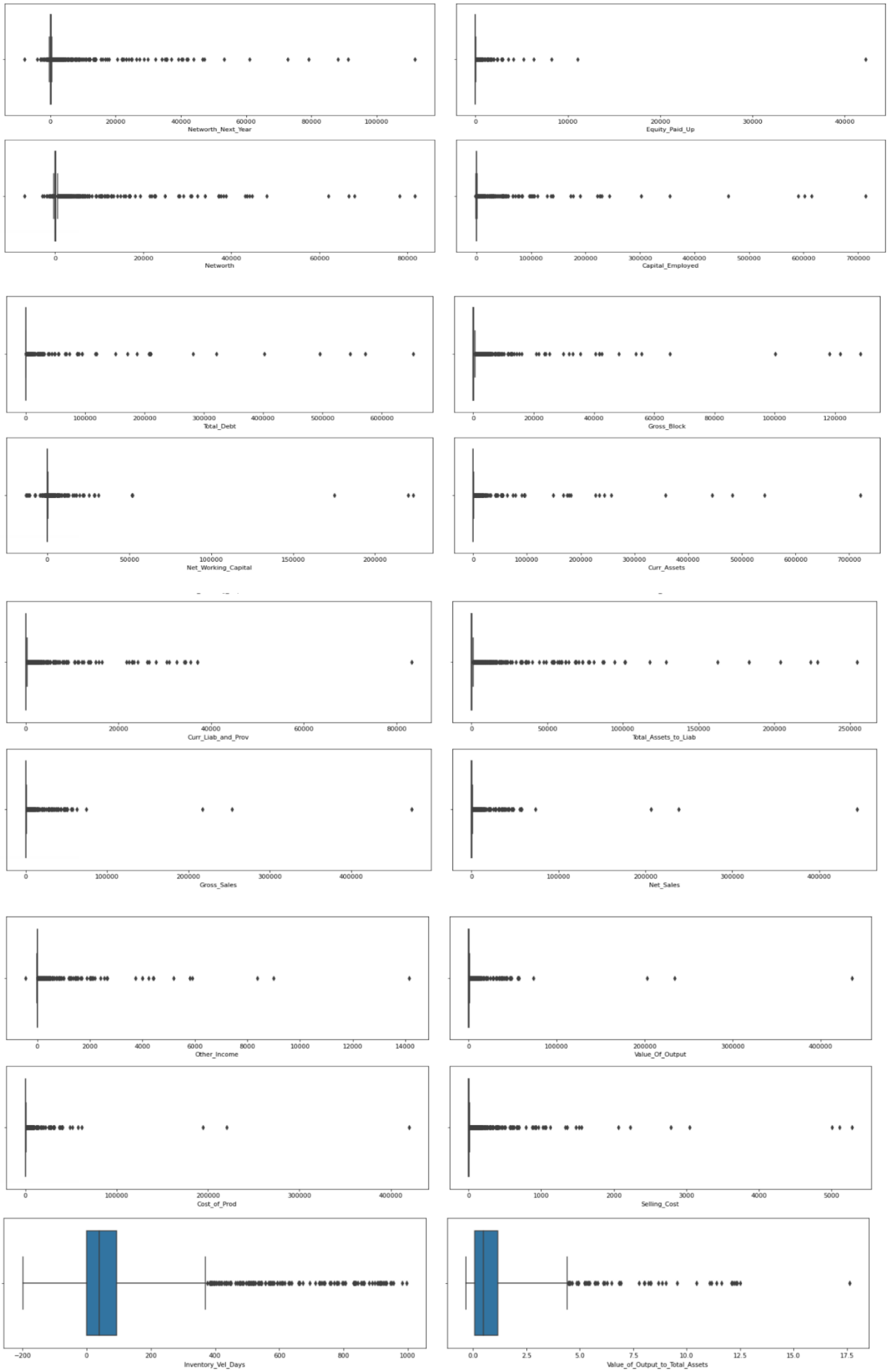

Networth_Next_Year
Equity_Paid_Up
Networth
Capital_Employed
Total_Debt
Gross_Block
Net_Working_Capital
Curr_Assets
Curr_Liab_and_Prov
Total_Assets_to_Liab
Gross_Sales
Net_Sales
Other_Income
Value_Of_Output
Cost_of_Prod
Selling_Cost
Inventory_Vel_Days
Value_of_Output_to_Total_Assets

Fig 2

```
Networth                                650
Networth_Next_Year                      676
Other_Income                            603
PAT                                     959
PBDT                                    815
PBDTM_perc[Latest]                      695
PBIDT                                   671
PBIDTM_perc[Latest]                     595
PBIT                                    720
PBITM_perc[Latest]                      717
PBT                                     941
ROG-CP_perc                             637
ROG-Capital_Employed_perc               572
ROG-Cost_of_Production_perc             675
ROG-Gross_Block_perc                    830
ROG-Gross_Sales_perc                    671
ROG-Market_Capitalisation_perc          497
ROG-Net_Sales_perc                      667
ROG-Net_Worth_perc                      747
ROG-PAT_perc                            598
ROG-PBDT_perc                           628
ROG-PBIDT_perc                          611
ROG-PBIT_perc                           616
ROG-PBT_perc                            611
ROG-Revenue_earnings_in_forex_perc     1317
ROG-Revenue_expenses_in_forex_perc     1615
ROG-Total_Assets_perc                   483
Revenue_earnings_in_forex               738
Revenue_expenses_in_forex               693
Selling_Cost                            605
Total_Asset_Turnover_Ratio[Latest]      201
Total_Assets_to_Liabilities_            574
Total_Debt                              583
Value_Of_Output                         559
Value_of_Output_to_Gross_Block          481
Value_of_Output_to_Total_Assets         150
dtype: int64
```

Total outliers in the dataset

```
Total outliers in the dataset are 42322
```

Tab 11

Size of the dataset

The size of the credit risk dataset is 240262

Percentage of outliers in the dataset

Percentage of outlier values in the dataset is 17.61

What percent of values in each column are outliers

Percentage of data points above the upper limit or Q3 + 1.5*IQR

ROG-Revenue_expenses_in_forex_perc	22.89
Revenue_earnings_in_forex	20.58
Capital_expenses_in_forex	19.35
Revenue_expenses_in_forex	19.33
PAT	18.18
PBT	18.07
ROG-Revenue_earnings_in_forex_perc	17.88
CP	17.85
Market_Capitalisation	17.82
Adjusted_PAT	17.79
PBDT	17.76
Cash_Flow_From_Operating_Activities	17.09
Selling_Cost	16.87
PBIDT	16.84
Networth_Next_Year	16.82
PBIT	16.82
Other_Income	16.79
Capital_Employed	16.48
Networth	16.29
Total_Debt	16.26
Current_Liabilities_and_Provisions_	16.20
Current_Assets_	16.09
Total_Assets_to_Liabilities_	16.01
Current_Ratio[Latest]	15.76
Cost_of_Production	15.62
Value_Of_Output	15.59
Net_Sales	15.50
Gross_Sales	15.45
Gross_Block_	15.06
Net_Working_Capital_	14.92
Interest_Cover_Ratio[Latest]	14.89
Fixed_Assets_Ratio[Latest]	13.80
ROG-Gross_Block_perc	13.69
CEPS_annualised_Unit_Curr	13.58
Value_of_Output_to_Gross_Block	13.33

Tab 12

```
ROG-Market_Capitalisation_perc        12.58
Equity_Paid_Up                        12.49
Book_Value_Unit_Curr                  12.02
ROG-Cost_of_Production_perc           11.77
Book_Value_Adj._Unit_Curr             11.71
APATM_perc[Latest]                    11.49
Debtors_Velocity_Days                 11.10
ROG-Gross_Sales_perc                  10.96
Creditors_Velocity_Days               10.90
ROG-Net_Sales_perc                    10.90
Cash_Flow_From_Financing_Activities   10.74
ROG-Net_Worth_perc                    10.74
PBITM_perc[Latest]                    10.68
Inventory_Ratio[Latest]               10.46
Debtors_Ratio[Latest]                 10.35
ROG-Capital_Employed_perc              9.90
CPM_perc[Latest]                       9.54
ROG-PBIDT_perc                         9.54
PBIDTM_perc[Latest]                    9.48
PBDTM_perc[Latest]                     9.43
ROG-CP_perc                            9.31
ROG-PBIT_perc                          9.15
ROG-PBDT_perc                          9.09
ROG-Total_Assets_perc                  8.95
ROG-PAT_perc                           8.20
Co_Code                                8.11
ROG-PBT_perc                           7.98
Inventory_Velocity_Days                7.28
Cash_Flow_From_Investing_Activities    6.64
Total_Asset_Turnover_Ratio[Latest]     5.61
Value_of_Output_to_Total_Assets        4.18
Co_Name                                0.00
dtype: float64
```

A huge number of outliers such as in the case at hand can neither be dropped not taken into the mandated logistics regression model. Regression tree would have served the purpose better but we'll see what all we can do with the information we have.

Tab 13

Percentage of data points below the lower limit or Q1 - 1.5*IQR

ROG-Revenue_expenses_in_forex_perc	22.14
ROG-Revenue_earnings_in_forex_perc	18.85
Cash_Flow_From_Investing_Activities	17.79
Cash_Flow_From_Financing_Activities	17.29
APATM_perc[Latest]	14.53
CPM_perc[Latest]	10.54
ROG-Net_Worth_perc	10.09
PBDTM_perc[Latest]	9.96
ROG-Gross_Block_perc	9.45
PBITM_perc[Latest]	9.31
ROG-PBT_perc	9.06
Adjusted_PAT	8.81
PAT	8.56
ROG-PAT_perc	8.48
ROG-CP_perc	8.45
ROG-PBDT_perc	8.42
PBT	8.17
ROG-PBIT_perc	8.03
ROG-Gross_Sales_perc	7.75
ROG-Net_Sales_perc	7.70
ROG-PBIDT_perc	7.50
PBIDTM_perc[Latest]	7.11
ROG-Cost_of_Production_perc	7.06
ROG-Capital_Employed_perc	6.05
Interest_Cover_Ratio[Latest]	5.33
Cash_Flow_From_Operating_Activities	5.24
PBDT	4.96
CP	4.91
ROG-Total_Assets_perc	4.52
PBIT	3.26
CEPS_annualised_Unit_Curr	3.21
Net_Working_Capital_	2.51
Networth_Next_Year	2.04
PBIDT	1.87
Networth	1.84

Tab 14

```
Book_Value_Adj._Unit_Curr              1.84
Book_Value_Unit_Curr                   1.51
ROG-Market_Capitalisation_perc         1.28
Capital_Employed                       0.14
Value_of_Output_to_Gross_Block         0.08
Other_Income                           0.03
Inventory_Velocity_Days                0.03
Inventory_Ratio[Latest]                0.00
Gross_Sales                            0.00
Value_Of_Output                        0.00
Total_Debt                             0.00
Total_Assets_to_Liabilities_           0.00
Total_Asset_Turnover_Ratio[Latest]     0.00
Selling_Cost                           0.00
Revenue_expenses_in_forex              0.00
Revenue_earnings_in_forex              0.00
Capital_expenses_in_forex              0.00
Net_Sales                              0.00
Market_Capitalisation                  0.00
Co_Code                                0.00
Co_Name                                0.00
Cost_of_Production                     0.00
Creditors_Velocity_Days                0.00
Current_Assets_                        0.00
Current_Liabilities_and_Provisions_    0.00
Current_Ratio[Latest]                  0.00
Debtors_Ratio[Latest]                  0.00
Debtors_Velocity_Days                  0.00
Equity_Paid_Up                         0.00
Fixed_Assets_Ratio[Latest]             0.00
Gross_Block_                           0.00
Value_of_Output_to_Total_Assets        0.00
dtype: float64
```

The outliers were added to the missing values.

Checking the outliers after treatment

```
APATM_perc[Latest]                  0
Adjusted_PAT                        0
Book_Value_Adj._Unit_Curr           0
Book_Value_Unit_Curr                0
CEPS_annualised_Unit_Curr           0
CP                                  0
CPM_perc[Latest]                    0
Capital_Employed                    0
Capital_expenses_in_forex           0
Cash_Flow_From_Financing_Activities 0
Cash_Flow_From_Investing_Activities 0
Cash_Flow_From_Operating_Activities 0
Co_Code                             0
Co_Name                             0 ...
```

Tab 15

Missing Values Analysis

The size of the credit risk dataset is 243848

The credit risk dataset has 118 missing values

Percentage of missing values in the dataset is 0.05

Missing values in each individual column

Co_Code	0
Co_Name	0
Networth_Next_Year	0
Equity_Paid_Up	0
Networth	0
Capital_Employed	0
Total_Debt	0
Gross_Block_	0
Net_Working_Capital_	0
Current_Assets_	0
Current_Liabilities_and_Provisions_	0
Total_Assets_to_Liabilities_	0
Gross_Sales	0
Net_Sales	0
Other_Income	0
Value_Of_Output	0
Cost_of_Production	0
Selling_Cost	0
PBIDT	0

Tab 16

PBDT	0
PBIT	0
PBT	0
PAT	0
Adjusted_PAT	0
CP	0
Revenue_earnings_in_forex	0
Revenue_expenses_in_forex	0
Capital_expenses_in_forex	0
Book_Value_Unit_Curr	0
Book_Value_Adj._Unit_Curr	4
Market_Capitalisation	0
CEPS_annualised_Unit_Curr	0
Cash_Flow_From_Operating_Activities	0
Cash_Flow_From_Investing_Activities	0
Cash_Flow_From_Financing_Activities	0
ROG-Net_Worth_perc	0
ROG-Capital_Employed_perc	0

ROG-Gross_Block_perc	0
ROG-Gross_Sales_perc	0
ROG-Net_Sales_perc	0
ROG-Cost_of_Production_perc	0
ROG-Total_Assets_perc	0
ROG-PBIDT_perc	0
ROG-PBDT_perc	0
ROG-PBIT_perc	0
ROG-PBT_perc	0
ROG-PAT_perc	0
ROG-CP_perc	0
ROG-Revenue_earnings_in_forex_perc	0
ROG-Revenue_expenses_in_forex_perc	0
ROG-Market_Capitalisation_perc	0
Current_Ratio[Latest]	1
Fixed_Assets_Ratio[Latest]	1
Inventory_Ratio[Latest]	1

Tab 17

```
Debtors_Ratio[Latest]                    1
Total_Asset_Turnover_Ratio[Latest]       1
Interest_Cover_Ratio[Latest]             1
PBIDTM_perc[Latest]                      1
PBITM_perc[Latest]                       1
PBDTM_perc[Latest]                       1
CPM_perc[Latest]                         1
APATM_perc[Latest]                       1
Debtors_Velocity_Days                    0
Creditors_Velocity_Days                  0
Inventory_Velocity_Days                103
Value_of_Output_to_Total_Assets          0
Value_of_Output_to_Gross_Block           0
default                                  0
dtype: int64
```

Observation on missing values

Most of the missing values (103) are in the column Inventory_Velocitiy_Days.

Column Book_Value_Adj._Unit_Curr has 4 missing values.

The following 11 columns have 1 missing value each:-

((Current_Ratio[Latest], Fixed_Assets_Ratio[Latest], Inventory_Ratio[Latest], Debtors_Ratio[Latest], Total_Asset_Turnover_Ratio[Latest], Interest_Cover_Ratio[Latest], PBIDTM_perc[Latest], PBITM_perc[Latest], PBDTM_perc[Latest], CPM_perc[Latest], APATM_perc[Latest]))

Missing values in dataset after converting outliers to NAN

```
Co_Code                                  291
Co_Name                                    0
Networth_Next_Year                       676
Equity_Paid_Up                           448
Networth                                 650
Capital_Employed                         596
Total_Debt                               583
Gross_Block_                             540
Net_Working_Capital_                     625
Current_Assets_                          577
Current_Liabilities_and_Provisions_      581
Total_Assets_to_Liabilities_             574
Gross_Sales                              554
Net_Sales                                556
Other_Income                             603
Value_Of_Output                          559
Cost_of_Production                       560
Selling_Cost                             605
PBIDT                                    671
PBDT                                     815
PBIT                                     720
PBT                                      941
PAT                                      959
Adjusted_PAT                             954
CP                                       816
Revenue_earnings_in_forex                738
Revenue_expenses_in_forex                693
Capital_expenses_in_forex                694
Book_Value_Unit_Curr                     485
Book_Value_Adj._Unit_Curr                490
Market_Capitalisation                    639
CEPS_annualised_Unit_Curr                602
Cash_Flow_From_Operating_Activities      801
Cash_Flow_From_Investing_Activities      876
Cash_Flow_From_Financing_Activities     1005
ROG-Net_Worth_perc                       747
ROG-Capital_Employed_perc                572
ROG-Gross_Block_perc                     830
ROG-Gross_Sales_perc                     671
ROG-Net_Sales_perc                       667
ROG-Cost_of_Production_perc              675
ROG-Total_Assets_perc                    483
ROG-PBIDT_perc                           611
ROG-PBDT_perc                            628
ROG-PBIT_perc                            616
ROG-PBT_perc                             611
ROG-PAT_perc                             598
ROG-CP_perc                              637
ROG-Revenue_earnings_in_forex_perc      1317
ROG-Revenue_expenses_in_forex_perc      1615
ROG-Market_Capitalisation_perc           497
Current_Ratio[Latest]                    566
Fixed_Assets_Ratio[Latest]               496
Inventory_Ratio[Latest]                  376
Debtors_Ratio[Latest]                    372
Total_Asset_Turnover_Ratio[Latest]       202
Interest_Cover_Ratio[Latest]             726
PBIDTM_perc[Latest]                      596
PBITM_perc[Latest]                       718
PBDTM_perc[Latest]                       696
CPM_perc[Latest]                         721
APATM_perc[Latest]                       934
Debtors_Velocity_Days                    398
Creditors_Velocity_Days                  391
Inventory_Velocity_Days                  365
Value_of_Output_to_Total_Assets          150
Value_of_Output_to_Gross_Block           481
dtype: int64
```

Tab 18 and 19

Let's check for missing values in the dataset again

```
The size of the credit risk dataset is 240262

The credit risk dataset has 42440 missing values

Percentage of missing values in the dataset is 17.66
```

Why we dropped some columns:

'Co_Code' and 'Co_Name':- These are identifiers, not needed for model building.

Networth_Next_Year:- Features that have been translated directly to get the target can never be used as predictors, because then the model will be 100% accurate. In this case Networth_Next_Year was translated into default.

Shape of the dataset minus the default variable, after dropping three columns.

```
(3586, 64)
```

Distribution of variables (univariate analysis)

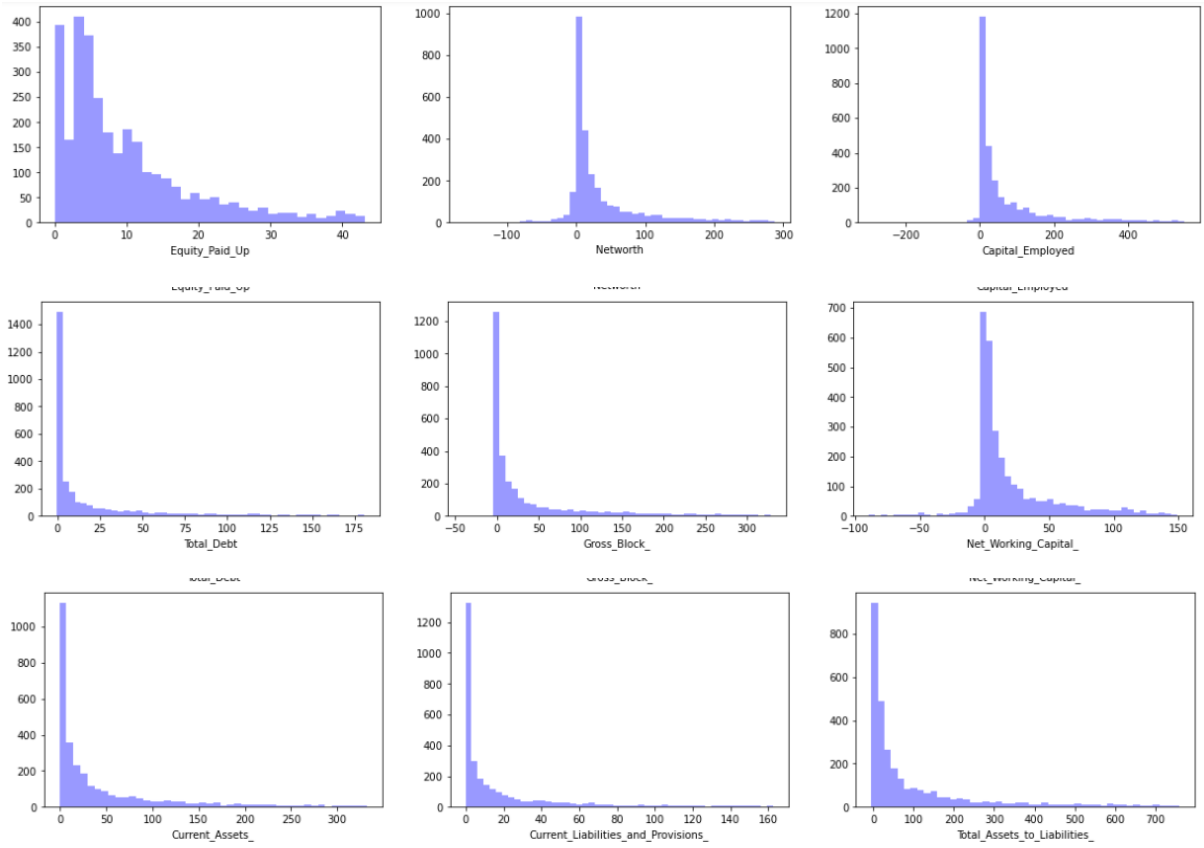
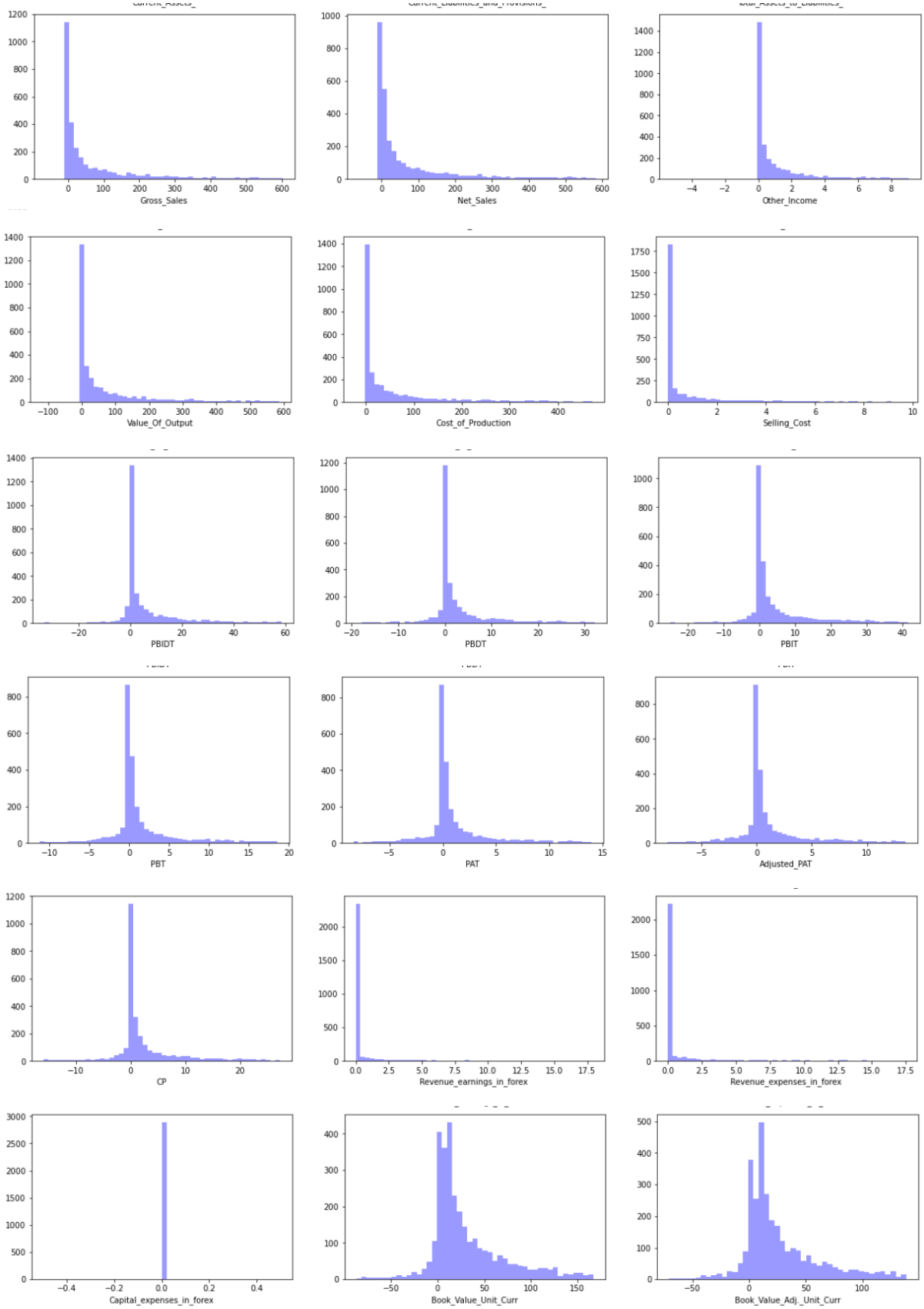
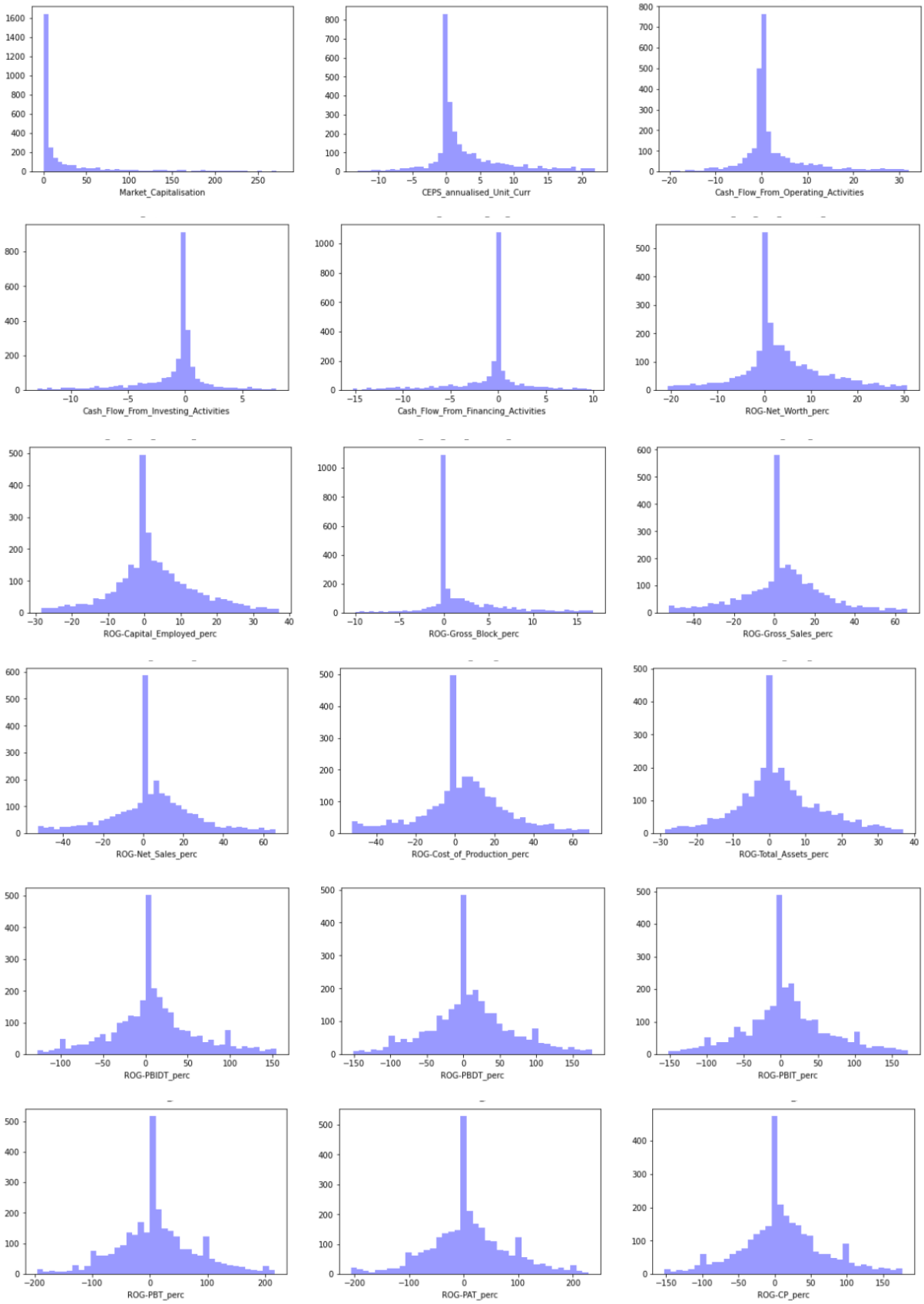


Fig 3



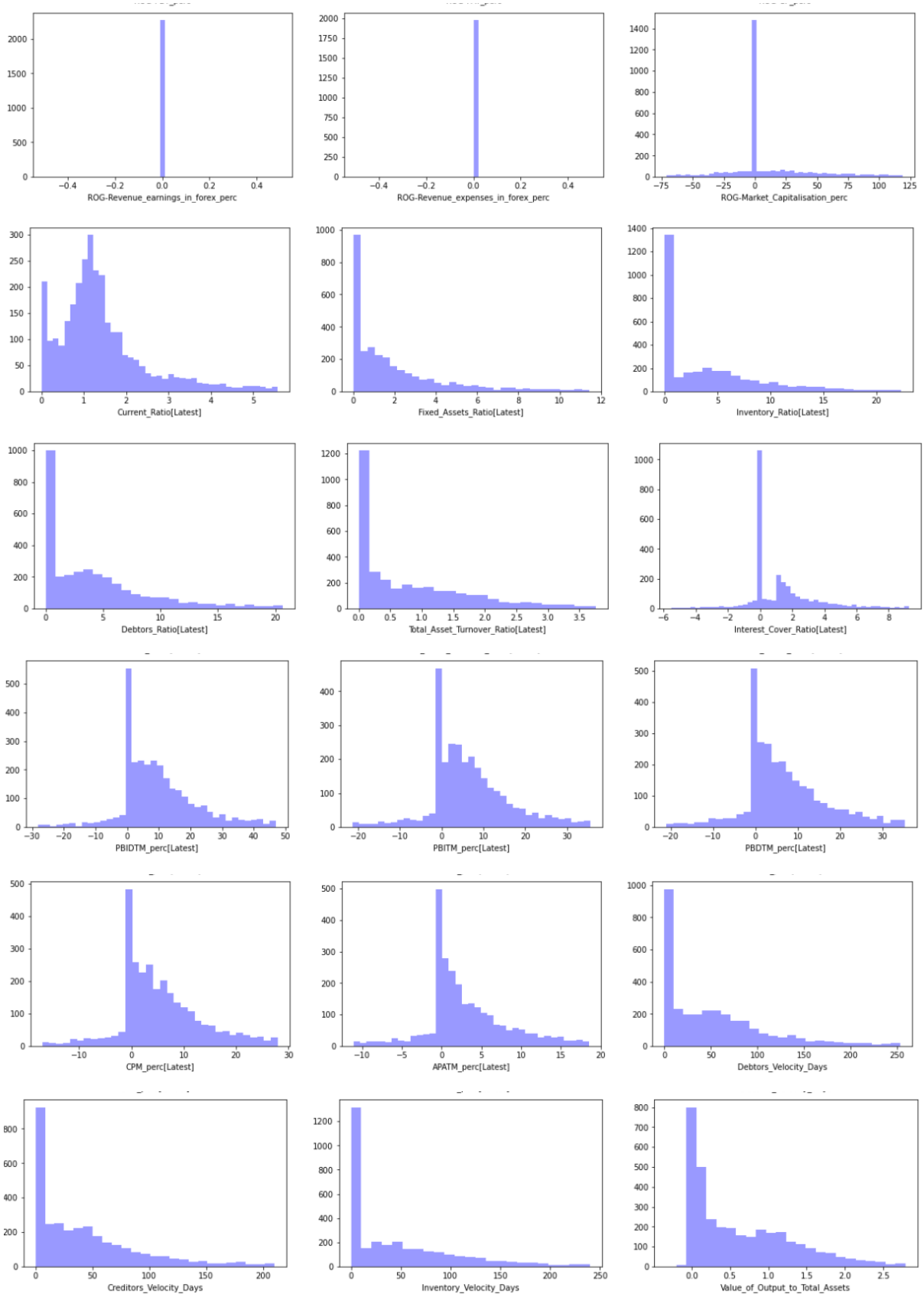
Most of the data is right skewed. The tallest bars are for non-defaulters. Income, gross sales, Net sales, other income etc. have non-defaulters as the highest frequency.

Fig 4



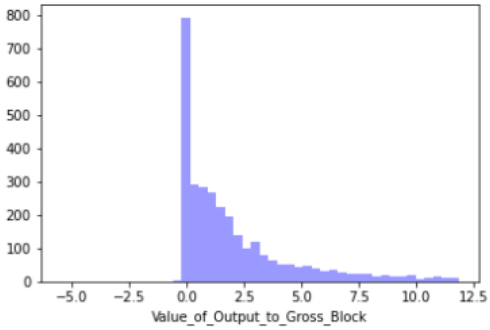
Except market capitalisation, others variables in the page are almost normally distributed. We will have to normalize our data and transform it a bit to bring extreme values more together, but as long as the response is linear it should be fine.

Fig 5



Top three variables in this page, if turned out to be useful, will require log transformation. A skewness chart is also give along, in the descending order.

Fig 6



Checking for skewness

	Skewness
Revenue_earnings_in_forex	3.61
Revenue_expenses_in_forex	3.10
Market_Capitalisation	2.52
Selling_Cost	2.41
Total_Debt	2.25
Other_Income	2.24
Current_Liabilities_and_Provisions_	2.22
Total_Assets_to_Liabilities_	2.14
Gross_Sales	2.14
Net_Sales	2.14
Value_Of_Output	2.13
Gross_Block_	2.11
Cost_of_Production	2.08
Current_Assets_	2.07
Capital_Employed	2.06
Networth	1.84
PBIDT	1.82
Fixed_Assets_Ratio[Latest]	1.71
Value_of_Output_to_Gross_Block	1.68
PBDT	1.65
PBIT	1.59

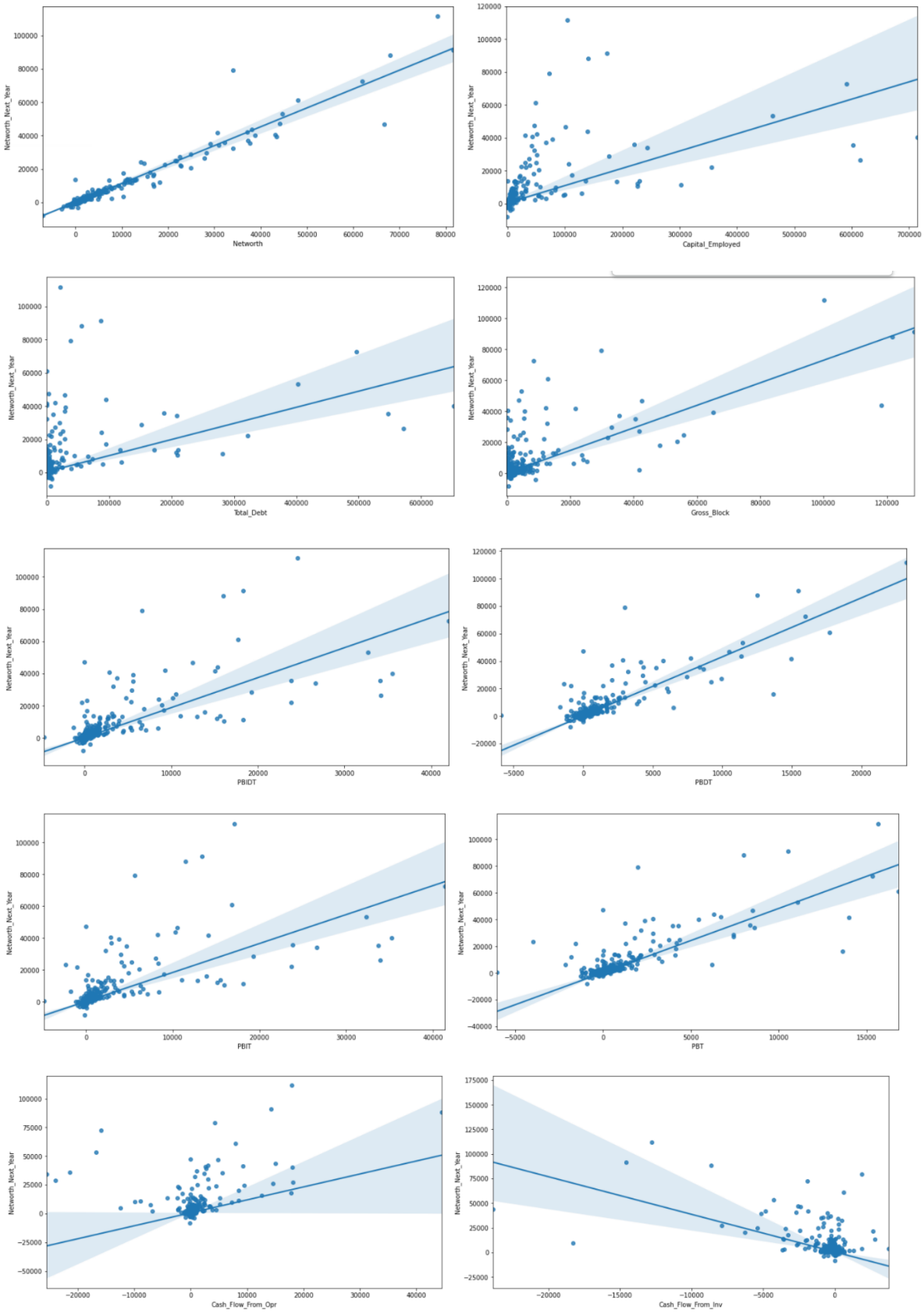
Tab 20

	Skewness
Adjusted_PAT	1.50
Equity_Paid_Up	1.49
PAT	1.48
Net_Working_Capital_	1.46
Cash_Flow_From_Operating_Activities	1.44
Current_Ratio[Latest]	1.41
Inventory_Ratio[Latest]	1.33
Debtors_Ratio[Latest]	1.31
Creditors_Velocity_Days	1.30
Debtors_Velocity_Days	1.26
Inventory_Velocity_Days	1.22
CEPS_annualised_Unit_Curr	1.19
Book_Value_Adj._Unit_Curr	1.15
Book_Value_Unit_Curr	1.14
Total_Asset_Turnover_Ratio[Latest]	1.13
ROG-Gross_Block_perc	1.11
ROG-Market_Capitalisation_perc	0.97
Value_of_Output_to_Total_Assets	0.96
Interest_Cover_Ratio[Latest]	0.96
PBIDTM_perc[Latest]	0.59
APATM_perc[Latest]	0.58
PBDTM_perc[Latest]	0.55
CPM_perc[Latest]	0.53
PBITM_perc[Latest]	0.48

	Skewness
ROG-Net_Worth_perc	0.39
ROG-Total_Assets_perc	0.29
ROG-Capital_Employed_perc	0.28
ROG-PBIDT_perc	0.24
ROG-PBDT_perc	0.19
ROG-PBT_perc	0.18
ROG-PBIT_perc	0.16
ROG-CP_perc	0.15
ROG-PAT_perc	0.11
ROG-Gross_Sales_perc	0.07
ROG-Net_Sales_perc	0.06
ROG-Revenue_expenses_in_forex_perc	0.00
ROG-Revenue_earnings_in_forex_perc	0.00
Capital_expenses_in_forex	0.00
ROG-Cost_of_Production_perc	-0.03
Cash_Flow_From_Financing_Activities	-1.07
Cash_Flow_From_Investing_Activities	-1.29

Tab 21

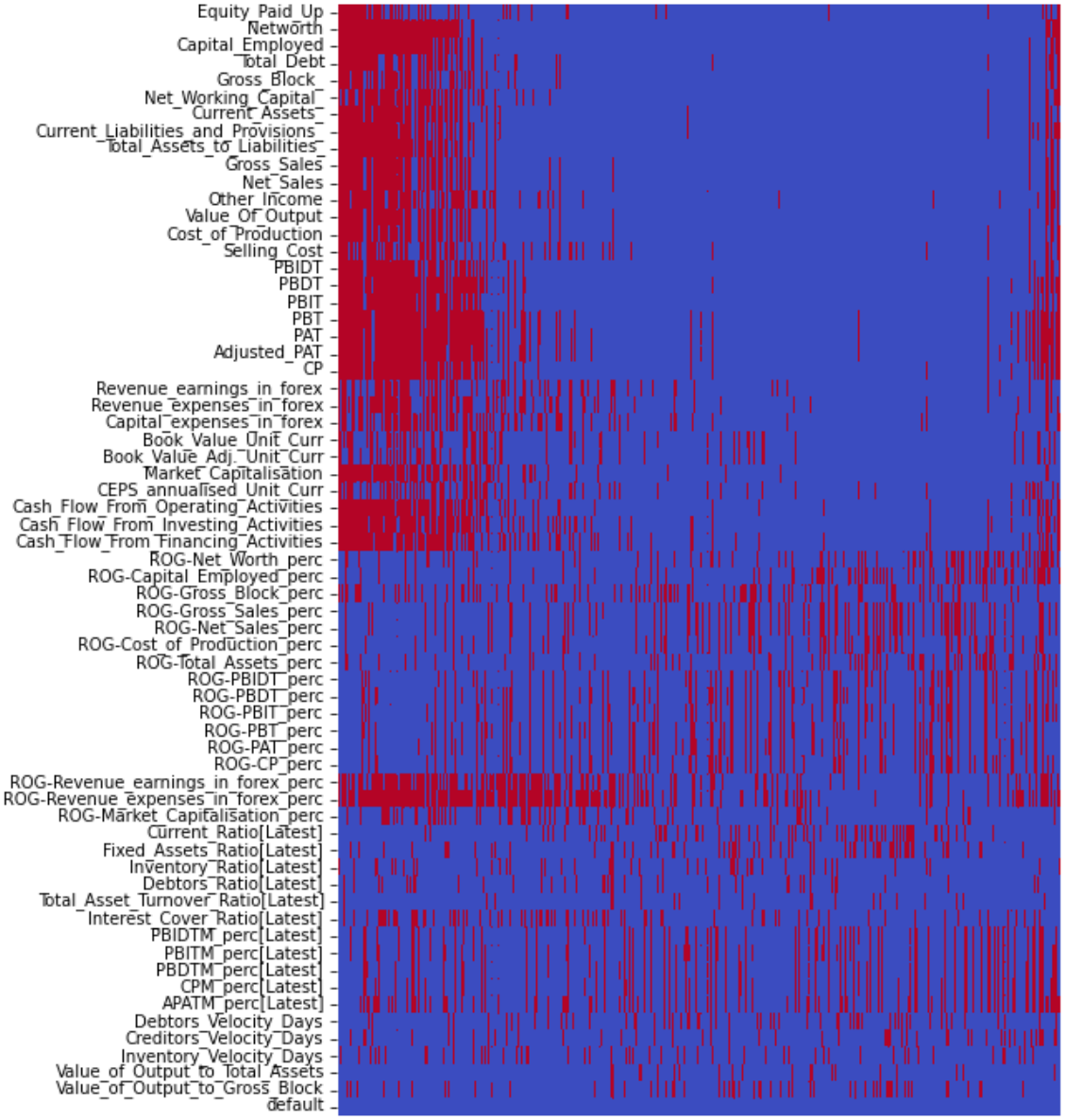
Studying Networth_Next_Year, since it will define our dependent variable



Cash flow from investing activities has a negative relation with Networth_Next_Year, while cash flow from operating activities has a positive relation.

Fig 7

Visual inspection of missing values in the dataset



The missing values, marked in the red, are just too many to drop or carry into the logistic regression model that's ordered to be made. Wherever the missing values make up more than 30% of the column, that feature will not be taken into the analysis.

Fig 8

Inspecting total missing values by each row

The objective is to capture as much of the data as possible if the complete picture is unavailable.

0	19
1	34
2	43
3	36
4	35
5	9
6	37
7	34
8	24
9	27
10	20
11	34
12	21
13	35
14	27
15	35
16	34
17	31
18	12

...

Many rows have a huge volume of missing information. This data given to us is not ready for logistic regression

0	3198
1	388

0	1319
1	141

388 are the total defaulters out of 3586 observations available with us in the dataset, while the proportion of default in the data section fit enough for analysis is 141. If we consider availability of features for deciding the observations to be considered, we will end up losing more than 64% of the actual defaulters. Only 36% will be captured.

Percentage of missing values in each column

ROG-Revenue_expenses_in_forex_perc	0.45
ROG-Revenue_earnings_in_forex_perc	0.37
Cash_Flow_From_Financing_Activities	0.28
PAT	0.27
Adjusted_PAT	0.27
PBT	0.26
APATM_perc[Latest]	0.26
Cash_Flow_From_Investing_Activities	0.24
ROG-Gross_Block_perc	0.23
CP	0.23
PBDT	0.23
Cash_Flow_From_Operating_Activities	0.22
ROG-Net_Worth_perc	0.21
Revenue_earnings_in_forex	0.21
Interest_Cover_Ratio[Latest]	0.20
CPM_perc[Latest]	0.20
PBIT	0.20

PBITM_perc[Latest]	0.20
PBDTM_perc[Latest]	0.19
Capital_expenses_in_forex	0.19
Revenue_expenses_in_forex	0.19
ROG-Cost_of_Production_perc	0.19
PBIDT	0.19
ROG-Gross_Sales_perc	0.19
ROG-Net_Sales_perc	0.19
Networth	0.18
Market_Capitalisation	0.18
ROG-CP_perc	0.18
ROG-PBDT_perc	0.18
Net_Working_Capital_	0.17
ROG-PBIT_perc	0.17
ROG-PBIDT_perc	0.17
ROG-PBT_perc	0.17
Selling_Cost	0.17
Other_Income	0.17
CEPS_annualised_Unit_Curr	0.17
ROG-PAT_perc	0.17
Capital Employed	0.17

Tab 22 and 23

Percentage of missing values in each column

PBIDTM_perc[Latest]	0.17
Total_Debt	0.16
Current_Liabilities_and_Provisions_	0.16
Current_Assets_	0.16
Total_Assets_to_Liabilities_	0.16
ROG-Capital_Employed_perc	0.16
Current_Ratio[Latest]	0.16
Cost_of_Production	0.16
Value_Of_Output	0.16
Net_Sales	0.16
Gross_Sales	0.15
Gross_Block_	0.15
ROG-Market_Capitalisation_perc	0.14
Fixed_Assets_Ratio[Latest]	0.14
Book_Value_Adj._Unit_Curr	0.14
Book_Value_Unit_Curr	0.14
ROG-Total_Assets_perc	0.13
Value_of_Output_to_Gross_Block	0.13
Equity_Paid_Up	0.12
Debtors_Velocity_Days	0.11
Creditors_Velocity_Days	0.11
Inventory_Ratio[Latest]	0.10
Debtors_Ratio[Latest]	0.10
Inventory_Velocity_Days	0.10
Total_Asset_Turnover_Ratio[Latest]	0.06
Value_of_Output_to_Total_Assets	0.04
default	0.00
dtype: float64	

Dropping columns with more than 30% missing values

ROG-Revenue_expenses_in_forex_perc	0.45
ROG-Revenue_earnings_in_forex_perc	0.37

Shape of the dataset after dropping these two columns

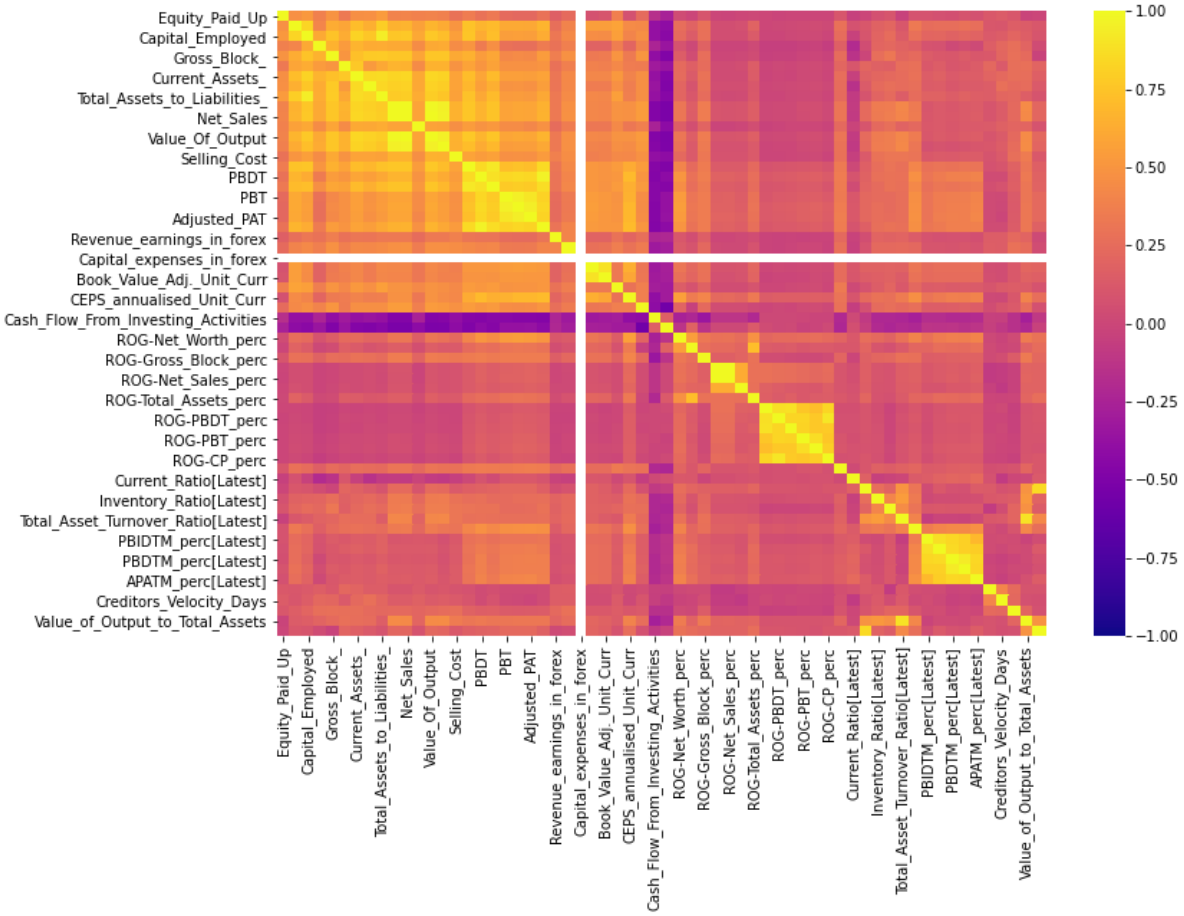
(3586, 63)

Predictors and response variable were segregated before the predictors were scaled using the Standard Scalar. And then the remaining missing values were imputed using the KNN Imputer.

About KNN Imputer

KNN Imputation is for completing missing values using k-Nearest Neighbors, an ideal method for us since this data set has quite a few extreme outliers that can be misclassified so easily, otherwise. Each sample's missing values are imputed using the mean value from n-neighbors (nearest neighbors) found in the training set. Two samples are close if the features that neither is missing are close. KNN looks for the closest match before it clubs the data point with anything.

Inspect possible correlations between independent variables



A good smattering of yellow and gold on the heat map indicates that quite a few variables have a high correlation with multiple variables in this dataset, and now to end this problem of multicollinearity, we can go to either PCA (principal component analysis), which is not explainable, or we can go to the method of VIF, which is length. Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis. Multicollinearity inflated 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. However, we have another option, which is to use recursive feature elimination or RFE. We will use this method to select roundabout 15 top features out of 60-odd.

For applying logistic regression, we first convert the data to X and Y or predictors and response and then we split the data into train and test, in a ratio of 67:33 and use random_state =42. Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset.

After split

```
Shape of train data: (2402, 62)
Shape of test data: (1184, 62)
Percentage of train data: 67
Percentage of test data: 33
```

Fig 9

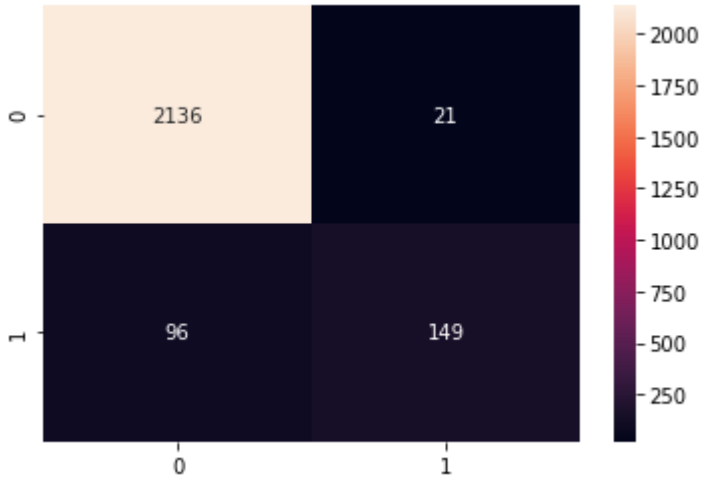
Selector ranking

```
array([38,  1,  1,  4,  1, 11, 10,  1,  1, 20, 40, 17,  1,  1, 24,  1,  2,
        1, 13, 15, 16,  3, 46, 21, 48,  1,  1, 18, 47, 45, 14, 32,  1,  1,
       39, 34, 35,  9, 23, 43, 42, 30, 41, 29, 44, 36,  1, 12, 28, 33,  7,
        1, 27,  5, 25, 26,  6, 31, 19, 22,  8, 37])
```

Features suggested by Logistic regression with RFE

	Feature	Rank
1	Networth	1
2	Capital_Employed	1
4	Gross_Block_	1
7	Current_Liabilities_and_Provisions_	1
8	Total_Assets_to_Liabilities_	1
12	Value_Of_Output	1
13	Cost_of_Production	1
15	PBIDT	1
17	PBIT	1
25	Book_Value_Unit_Curr	1
26	Book_Value_Adj._Unit_Curr	1
32	ROG-Net_Worth_perc	1
33	ROG-Capital_Employed_perc	1
46	Current_Ratio[Latest]	1
51	Interest_Cover_Ratio[Latest]	1

Validating the model on train and test set



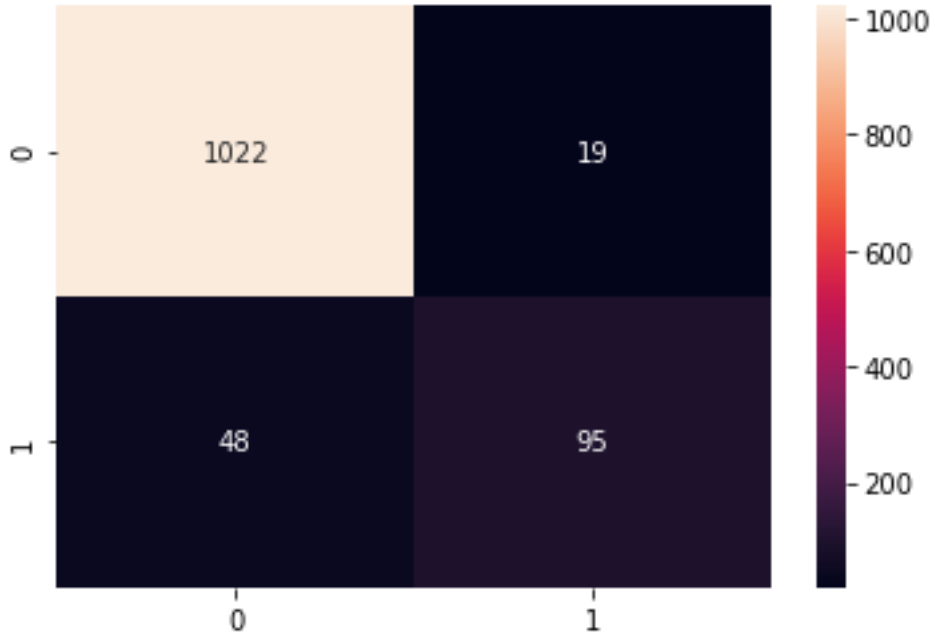
Confusion matrix (train)

The model calls 149 defaults truly and predicts 2136 non-defaults also accurately. But it also calls 96 defaults as non-default because it is an imbalanced dataset, which is why the model take majority of the cases to be non-default. It also predicts 21 non-defaults to be defaults, which doesn't hurt the lending company, even though the 96 misclassified defaults will. The experiment should be tried with SMOTE first to solve the problem of balance.

Tab 25 and Fig 10

Classification report (train): Logistic regression with RFE

	precision	recall	f1-score	support
0.0	0.96	0.99	0.97	2157
1.0	0.88	0.61	0.72	245
accuracy			0.95	2402
macro avg	0.92	0.80	0.85	2402
weighted avg	0.95	0.95	0.95	2402



Confusion matrix (test)

The model failed to predict 48 defaults but called 95 right. When it calls default, it is right 66% of the time.

Classification report (test): Logistic regression with RFE

	precision	recall	f1-score	support
0.0	0.96	0.98	0.97	1041
1.0	0.83	0.66	0.74	143
accuracy			0.94	1184
macro avg	0.89	0.82	0.85	1184
weighted avg	0.94	0.94	0.94	1184

The model performs slightly better in test, and although the accuracy drops, the recall level increases. The data is not overfitting but we'll look for a candidate with a more robust performance.

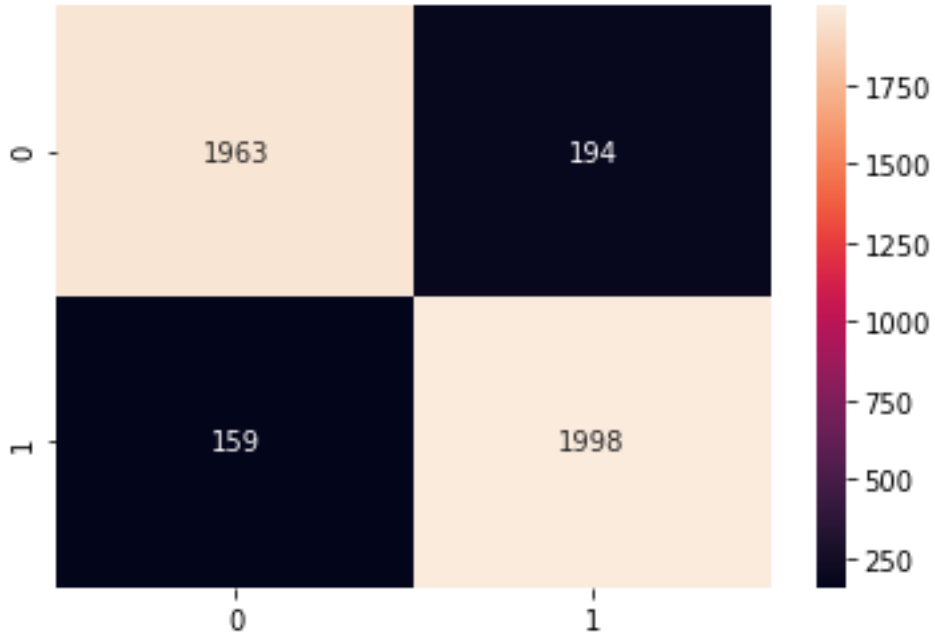
Tab 26

Logistic regression with RFE, on SMOTE

Classification report (train)

	precision	recall	f1-score	support
0.0	0.93	0.91	0.92	2157
1.0	0.91	0.93	0.92	2157
accuracy			0.92	4314
macro avg	0.92	0.92	0.92	4314
weighted avg	0.92	0.92	0.92	4314

Confusion matrix (train)

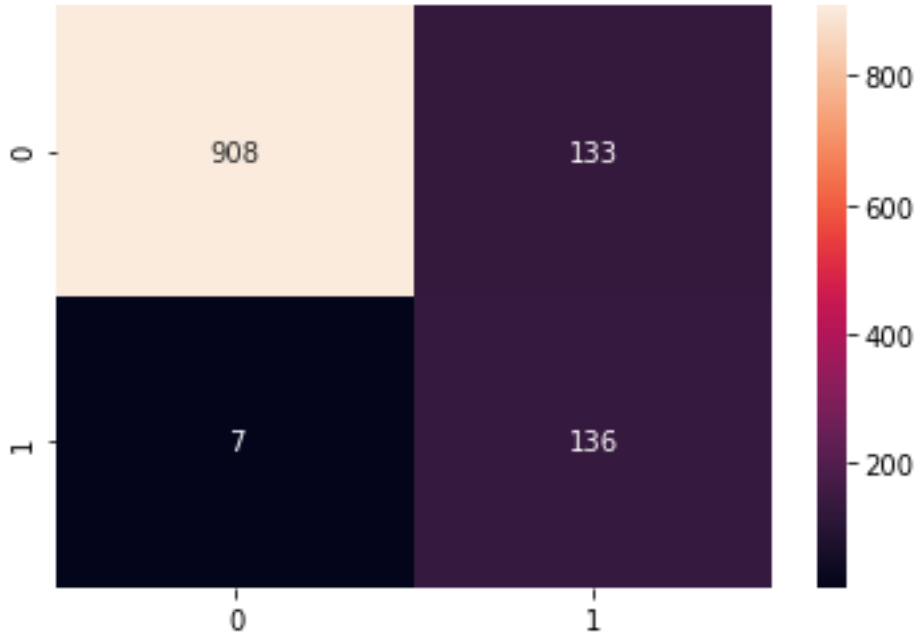


Precision, recall, and accuracy improve greatly with RFE and SMOTE. The model is able to call right in case of default 93% of the time in training.

Classification report (test)

	precision	recall	f1-score	support
0.0	0.99	0.87	0.93	1041
1.0	0.51	0.95	0.66	143
accuracy			0.88	1184
macro avg	0.75	0.91	0.79	1184
weighted avg	0.93	0.88	0.90	1184

Confusion matrix (test)



The model performs well even when it comes to the test dataset. Precision drops, but that's a trade-off, as the recall and accuracy are still very good. Now the model is calling default right 95% of the time. The misclassification of non-default has increased now but the creditor company will feel happy that misclassification of default has reduced a great deal. Only 7 defaults were misclassified in the test stage.

Finally, we are able to achieve a descent recall value without overfitting. Considering the opportunities such as outliers, missing values and correlated features this is a fairly good model. It can be improved if we get better quality data where the features explaining the default are not missing to this extent. Of course, we can try other techniques which are not sensitive towards missing values and outliers.

Tab 28

Model building using statsmodels library

Model Building using Logistic Regression for 'Probability at default'¶

For further analysis, we use Statsmodels Logit Modeling with backward elimination. After each model is built, the variable that has a p-value of >0.05 will be dropped as those coefficients are unreliable. Otherwise, the null hypothesis is that the feature is not significant in predicting default. If p is low, null will go. If p is high, null will fly.

The equation of the Logistic Regression by which we predict the corresponding probabilities and then go on predict a discrete target variable is

$$y = \frac{1}{1+e^{-z}}$$

$$\text{Note: } z = \beta_0 + \sum_{i=1}^{n}(\beta_i X_1)$$

Feature Selection

Since there are too many columns, we need to determine the columns which are related and eliminate them if possible. We will use VIF to determine the collinearity and eliminate using a threshold of 5.

```
dropping 'Net_Sales' at index: 10
dropping 'ROG-Gross_Sales_perc' at index: 34
dropping 'Value_Of_Output' at index: 11
dropping 'PAT' at index: 17
dropping 'PBDT' at index: 14
dropping 'Gross_Sales' at index: 9
dropping 'Total_Assets_to_Liabilities_' at index: 8
dropping 'PBDTM_perc[Latest]' at index: 47
dropping 'ROG-PBDT_perc' at index: 33
dropping 'Current_Assets_' at index: 6
dropping 'PBIT' at index: 11
dropping 'Adjusted_PAT' at index: 12
dropping 'CP' at index: 12
dropping 'ROG-PBIDT_perc' at index: 28
dropping 'Capital_Employed' at index: 2
dropping 'PBITM_perc[Latest]' at index: 39
dropping 'Value_of_Output_to_Total_Assets' at index: 44
dropping 'Cost_of_Production' at index: 7
dropping 'Book_Value_Unit_Curr' at index: 13
dropping 'ROG-PBT_perc' at index: 26
dropping 'CPM_perc[Latest]' at index: 36
Remaining variables:
Index(['Equity_Paid_Up', 'Networth', 'Total_Debt', 'Gross_Block_',
       'Net_Working_Capital_', 'Current_Liabilities_and_Provisions_',
       'Other_Income', 'Selling_Cost', 'PBIDT', 'PBT',
       'Revenue_earnings_in_forex', 'Revenue_expenses_in_forex',
       'Capital_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-PAT_perc', 'ROG-CP_perc',
       'ROG-Market_Capitalisation_perc', 'Current_Ratio[Latest]',
       'Fixed_Assets_Ratio[Latest]', 'Inventory_Ratio[Latest]',
       'Debtors_Ratio[Latest]', 'Total_Asset_Turnover_Ratio[Latest]',
       'Interest_Cover_Ratio[Latest]', 'PBIDTM_perc[Latest]',
       'APATM_perc[Latest]', 'Debtors_Velocity_Days',
       'Creditors_Velocity_Days', 'Inventory_Velocity_Days',
       'Value_of_Output_to_Gross_Block'],
      dtype='object')
```

Predictors

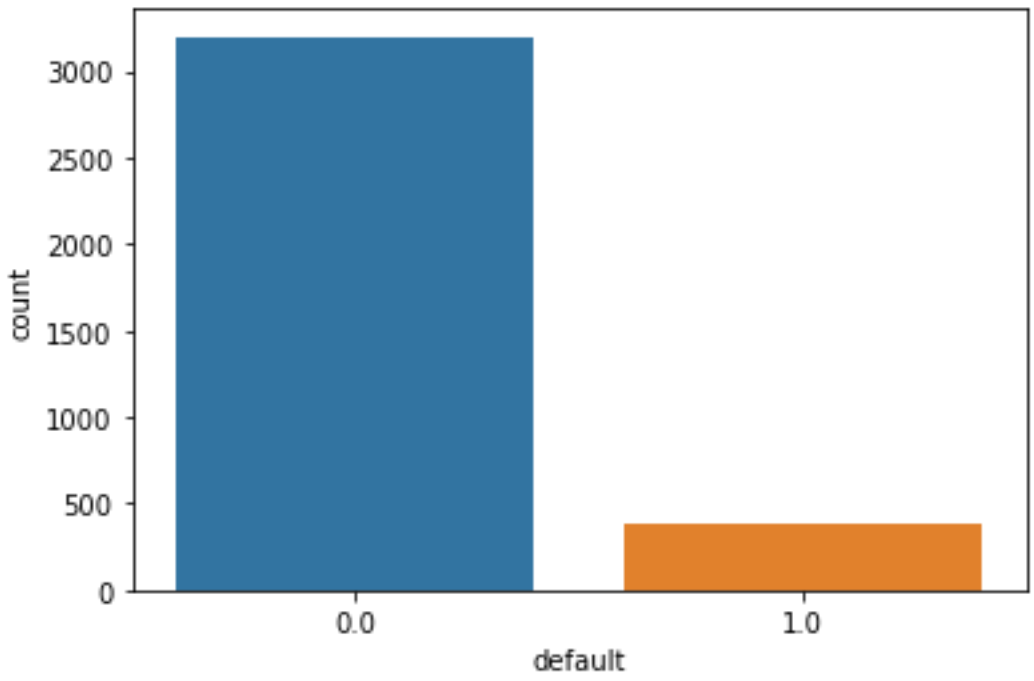
	Equity_Paid_Up	Networth	Total_Debt	Gross_Block_	Net_Working_Capital_	Current_Liabilities_and_Provisions_	Other_Income	Selling_Cost	PBIDT	PBT
0	-0.04	-0.37	0.06	-0.37	-0.22	-0.34	3.44	-0.51	-0.24	-0.91
1	1.73	-0.21	1.02	1.13	0.68	1.01	0.41	-0.01	0.24	0.01
2	0.88	2.01	-0.03	2.13	0.12	1.85	1.14	0.60	1.57	0.04
3	0.13	0.47	0.42	0.66	0.90	0.15	-0.32	1.18	0.19	-0.52
4	1.32	1.27	0.91	0.81	0.28	0.22	0.75	0.49	0.89	0.03
5	0.52	-0.73	-0.44	-0.60	-0.78	-0.37	-0.58	-0.51	-0.48	-0.33
6	2.28	-0.26	0.73	0.32	-0.49	0.29	-0.25	0.33	-0.57	-0.07

After concatenating predictors and response

Creditors_Velocity_Days	Inventory_Velocity_Days	Value_of_Output_to_Gross_Block	default
-0.95	-0.05	-0.81	1.00
1.21	-0.81	-0.71	1.00
0.38	-0.85	-0.91	1.00
0.40	-0.81	-0.06	1.00
1.10	-0.85	-0.79	1.00
-0.95	-0.85	-0.81	1.00
-0.69	0.65	-0.81	1.00
-0.01	-0.07	-0.82	1.00
2.15	-0.81	-0.69	1.00

...

Split of default variable




```
0.00    3198
1.00     388
```

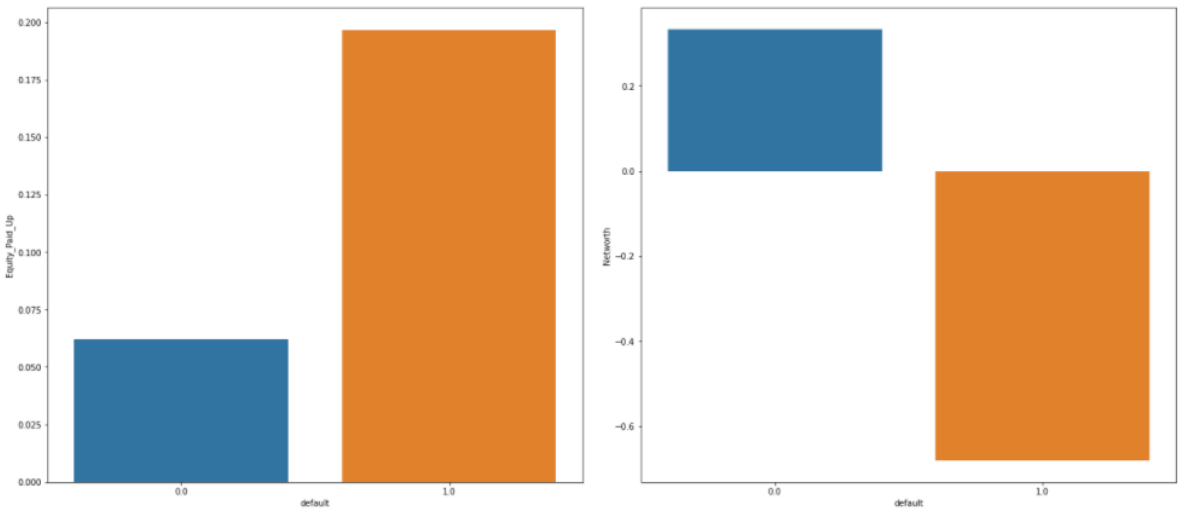
Tab 30-31, Fig 11

Dataset grouped by the variable default

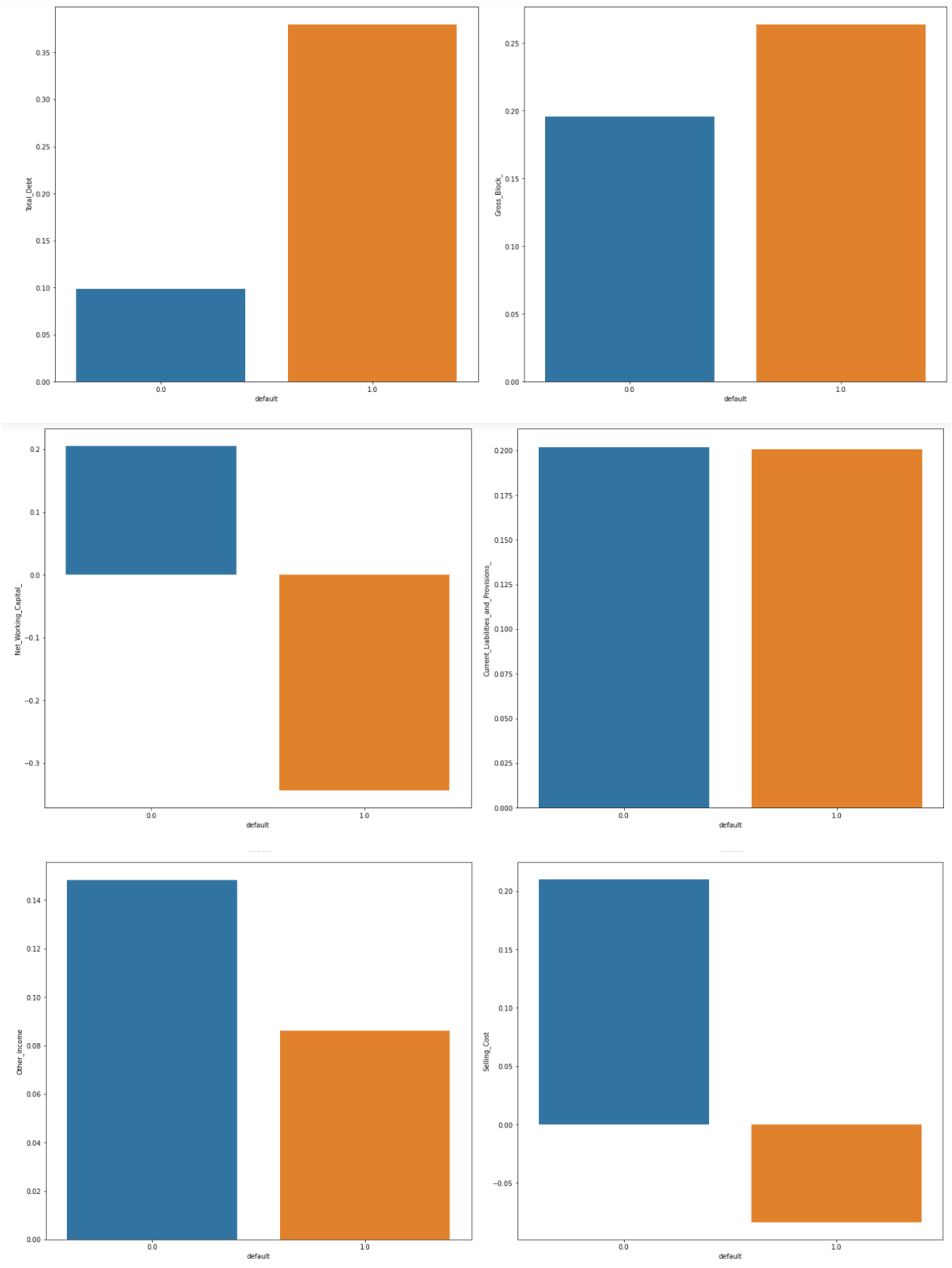
default	0.00	1.00
Equity_Paid_Up	198.90	76.30
Networth	1069.29	-263.94
Total_Debt	316.61	147.36
Gross_Block_	625.86	102.39
Net_Working_Capital_	655.16	-133.40
Current_Liabilities_and_Provisions_	645.81	77.90
Other_Income	473.90	33.38
Selling_Cost	671.45	-32.50
PBIDT	1202.71	-175.91
PBT	1196.65	-206.74
Revenue_earnings_in_forex	420.02	3.06
Revenue_expenses_in_forex	608.12	-2.27
Capital_expenses_in_forex	0.00	0.00
Book_Value_Adj._Unit_Curr	683.09	-382.41
Market_Capitalisation	569.78	-72.86
CEPS_annualised_Unit_Curr	793.74	-256.34
Cash_Flow_From_Operating_Activities	612.11	-53.29
Cash_Flow_From_Investing_Activities	-666.92	84.43
Cash_Flow_From_Financing_Activities	-652.70	-26.34
ROG-Net_Worth_perc	211.16	-268.25

default	0.00	1.00
ROG-Capital_Employed_perc	222.95	-195.55
ROG-Gross_Block_perc	270.66	-127.86
ROG-Net_Sales_perc	112.03	-136.76
ROG-Cost_of_Production_perc	150.79	-153.48
ROG-Total_Assets_perc	245.96	-204.05
ROG-PBIT_perc	85.30	-55.01
ROG-PAT_perc	67.64	-90.55
ROG-CP_perc	67.84	-81.70
ROG-Market_Capitalisation_perc	246.92	-89.98
Current_Ratio[Latest]	421.70	-304.12
Fixed_Assets_Ratio[Latest]	316.44	-199.77
Inventory_Ratio[Latest]	195.96	-100.39
Debtors_Ratio[Latest]	174.41	-92.09
Total_Asset_Turnover_Ratio[Latest]	335.61	-141.91
Interest_Cover_Ratio[Latest]	501.96	-246.66
PBIDTM_perc[Latest]	277.87	-221.43
APATM_perc[Latest]	199.78	-212.12
Debtors_Velocity_Days	11.34	-33.47
Creditors_Velocity_Days	-131.41	56.51
Inventory_Velocity_Days	8.08	-34.64
Value_of_Output_to_Gross_Block	323.90	-197.80

Variables against default

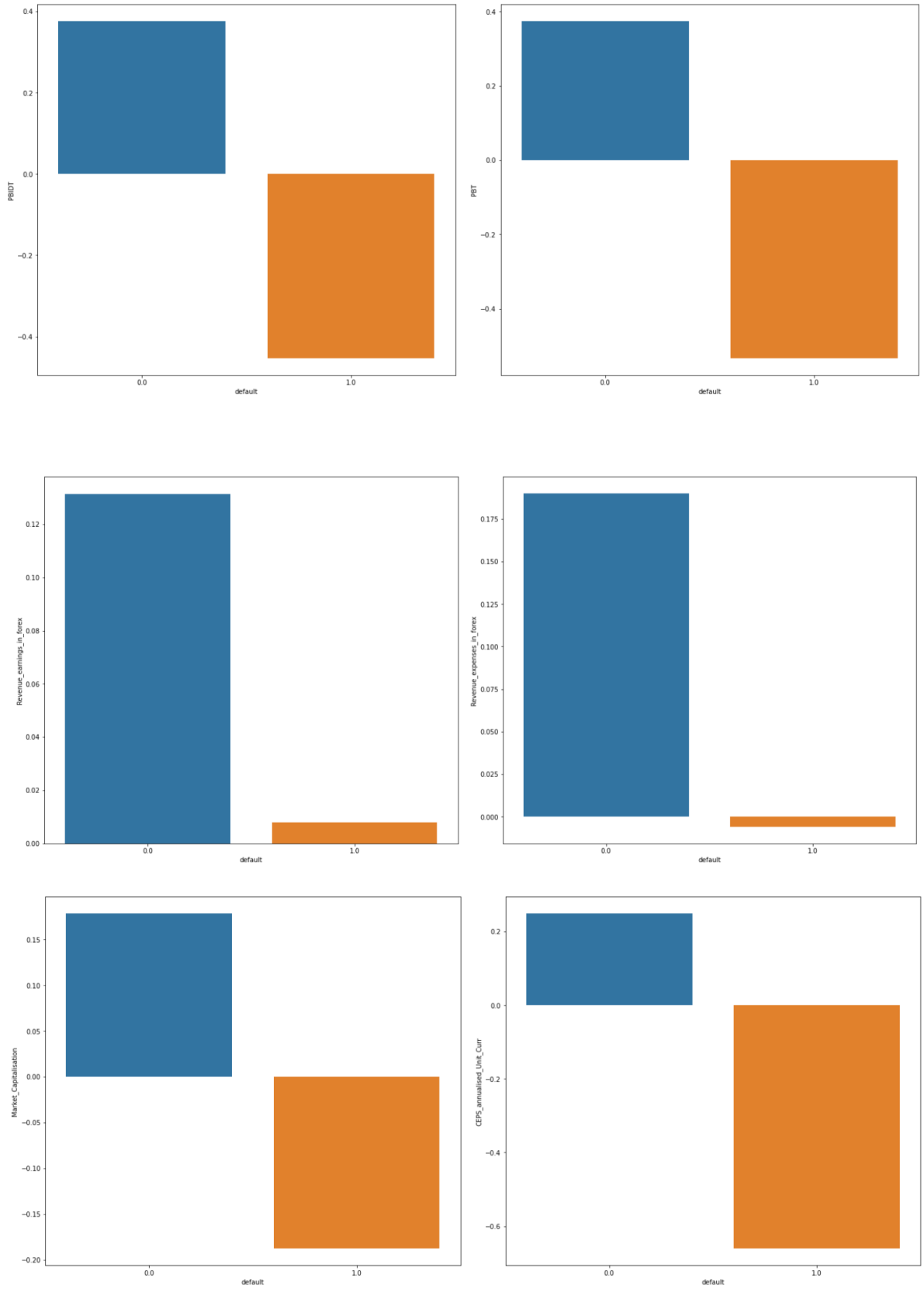


Tab 32, Fig 12



Companies with high total debt tend to default a lot. Companies that have other income tend to stay non-defaulters overwhelmingly. Companies with current liabilities have equal proportion of default and non-default.

Fig 13



Companies that have revenue expenses and earnings in forex tend to default very less.

Fig 14

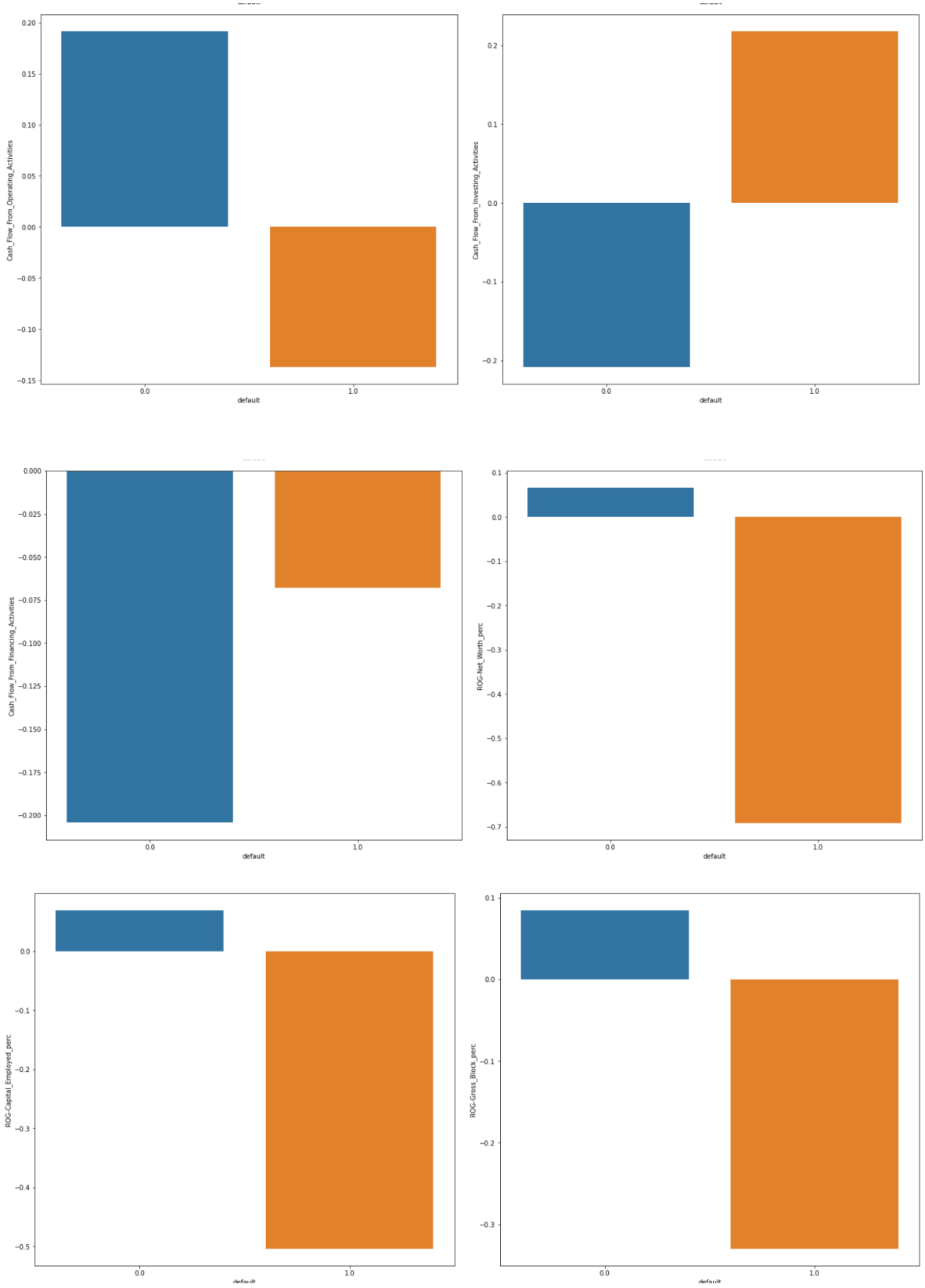

Cash_Flow_From_Operating_Activities
Cash_Flow_From_Investing_Activities
Cash_Flow_From_Financing_Activities
ROG-Net_Worth_perc
ROG-Capital_Employed_perc
ROG-Gross_Block_perc
default
0.0
1.0

Fig 15

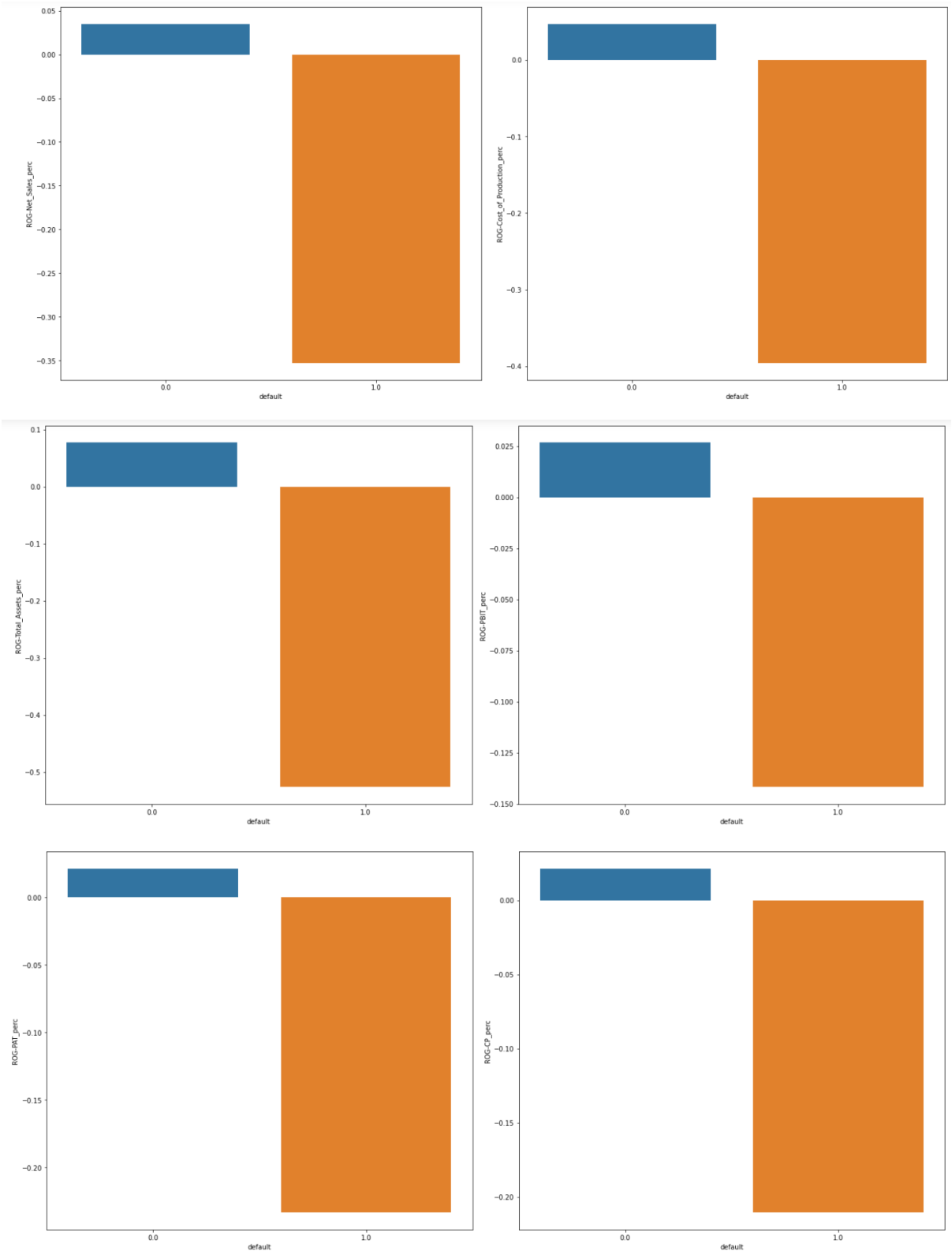

ROG-Net_Sales_perc
ROG-Cost_of_Production_perc
ROG-Total_Assets_perc
ROG-PBIT_perc
ROG-PAT_perc
ROG-CP_perc
default

Fig 16

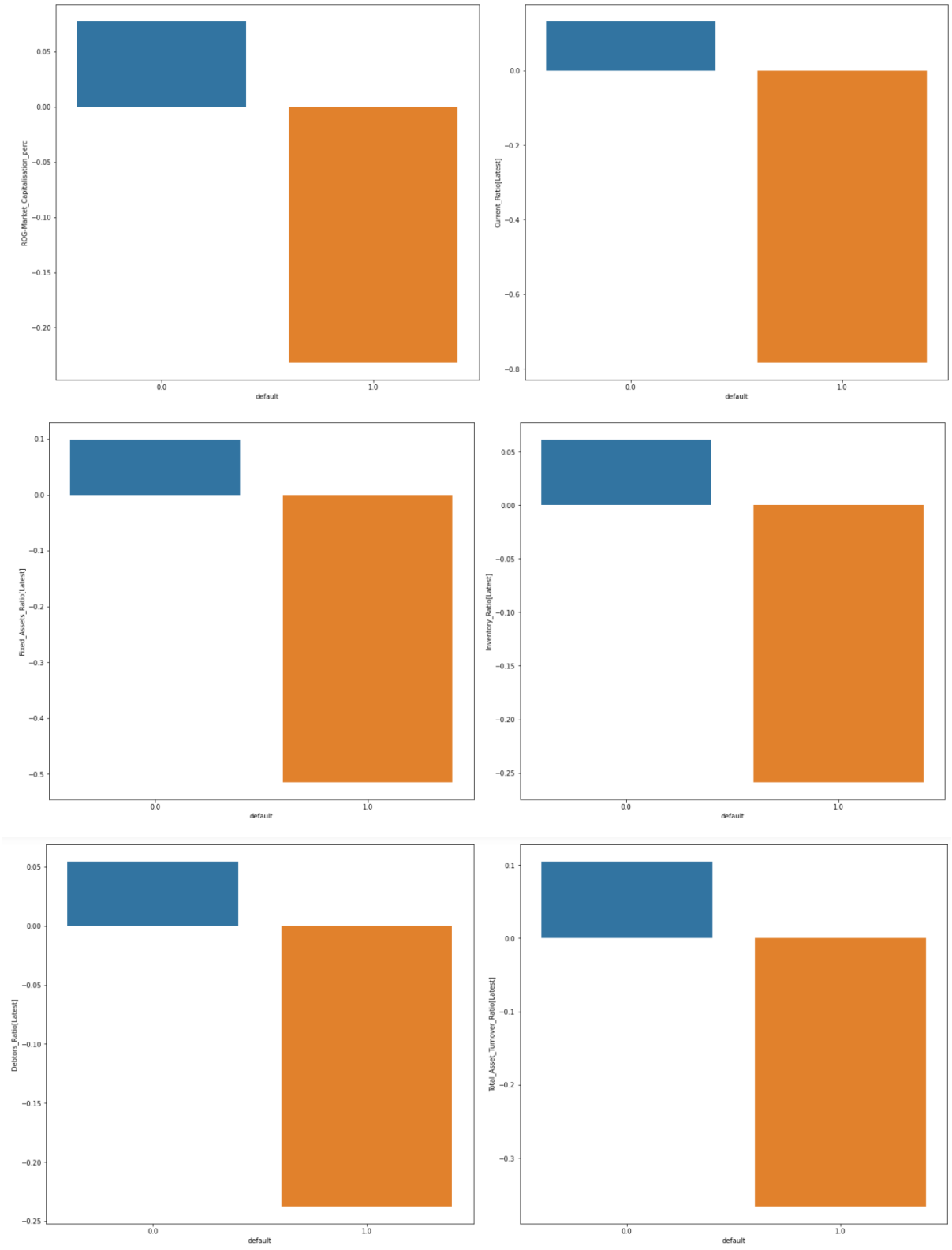
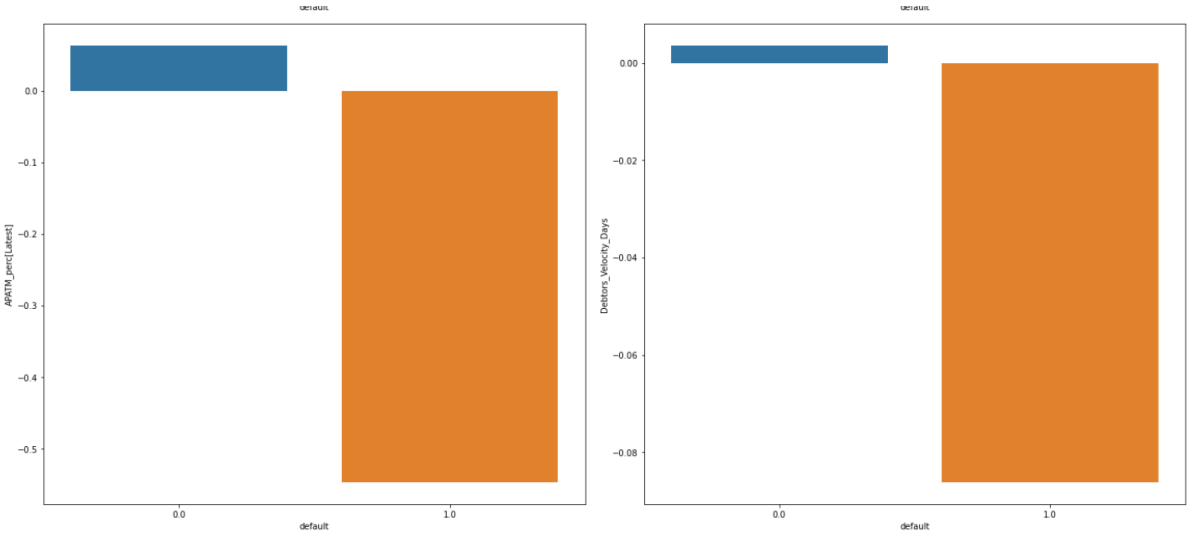

ROG-Market_Capitalisation_perc
Current_Ratio[Latest]
Fixed_Assets_Ratio[Latest]
Inventory_Ratio[Latest]
Debtors_Ratio[Latest]
Total_Asset_Turnover_Ratio[Latest]
default
0.0
1.0

Fig 17



The credit risk data set is quite imbalanced, so one needs to apply SMOTE before evaluating all the models on both SMOTE as well as non-SMOTE dataset for comparison.

```
Before OverSampling the shape of predictors: (3586, 41)
Before OverSampling the shape of response: (3586,)
Before OverSampling, counts of label '1': 388
Before OverSampling, counts of label '0': 3198
```

```
After OverSampling the shape of predictors: (6396, 62)
After OverSampling the shape of response: (6396,)
After OverSampling, counts of label '1': 3198
After OverSampling, counts of label '0': 3198
```

Data is now balanced:-

```
1.00    0.50
0.00    0.50
Name: default, dtype: float64
```

Fig 18

Fitting the logistic regression model on imbalanced data

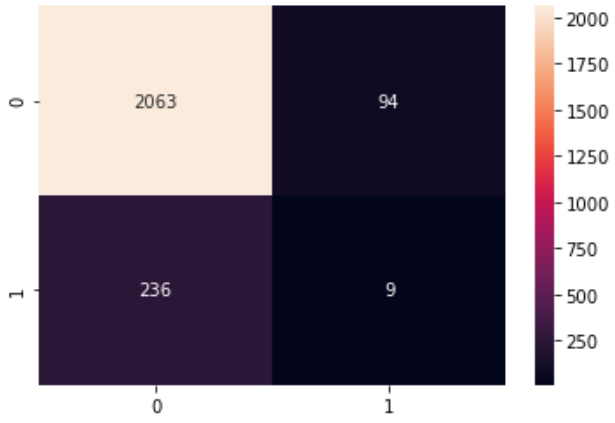
Model_1: Checking the parameters

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2394
Method:	MLE	**Df Model:**	7
Date:	Sun, 09 Jan 2022	**Pseudo R-squ.:**	0.3461
Time:	05:51:29	**Log-Likelihood:**	-538.49
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	7.167e-119

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.1823	0.147	-21.670	0.000	-3.470	-2.894
Equity_Paid_Up	0.5689	0.108	5.285	0.000	0.358	0.780
Networth	-3.8834	0.329	-11.787	0.000	-4.529	-3.238
Total_Debt	0.4673	0.170	2.752	0.006	0.135	0.800
Gross_Block_	0.5600	0.177	3.161	0.002	0.213	0.907
Net_Working_Capital_	-0.3988	0.158	-2.530	0.011	-0.708	-0.090
Current_Liabilities_and_Provisions_	0.2351	0.171	1.371	0.170	-0.101	0.571
Other_Income	0.3079	0.132	2.334	0.020	0.049	0.566

Validating the model on train set



p-value is low for all the variables except Current_Liabilities_and_Provisions_, so this feature will be dropped from the future model building.

Default is greatly misclassified, 236 cases of it on top of that, while just 9 defaults are called correctly in train.

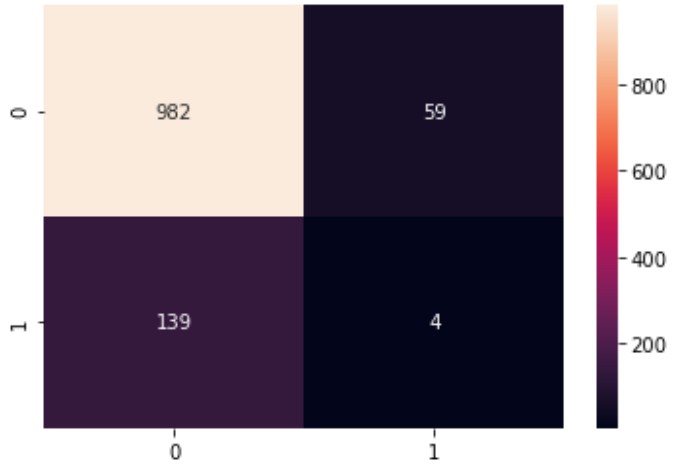
The bias for non-default is because the dataset is imbalanced.

Tab 33

Classification report (train)

	precision	recall	f1-score	support
0.0	0.90	0.96	0.93	2157
1.0	0.09	0.04	0.05	245
accuracy			0.86	2402
macro avg	0.49	0.50	0.49	2402
weighted avg	0.81	0.86	0.84	2402

Validating the model on test set



Recall is important to the credit company, since it would want maximum defaults to be called right, which is where this model fails even when it comes to the test dataset.

The next step will be to balance the dataset to eliminate the model's bias for the non-default class.

Classification report (test)

	precision	recall	f1-score	support
0.0	0.88	0.94	0.91	1041
1.0	0.06	0.03	0.04	143
accuracy			0.83	1184
macro avg	0.47	0.49	0.47	1184
weighted avg	0.78	0.83	0.80	1184

Model is not overfitting but the recall is really poor.

Tab 34

Fitting the logistic regression model on balanced data

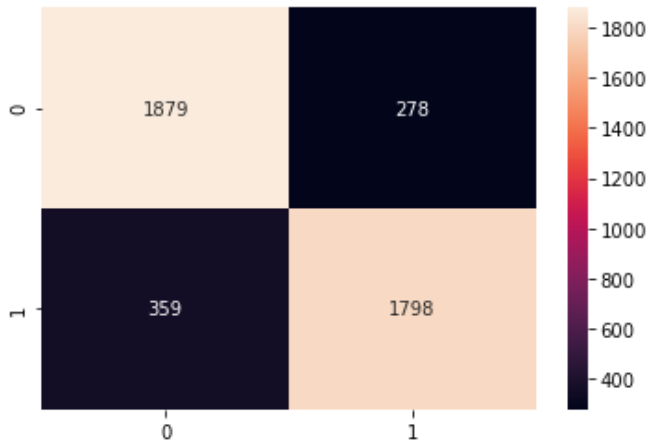
Logit Regression Results

Dep. Variable:	default	**No. Observations:**	4314
Model:	Logit	**Df Residuals:**	4306
Method:	MLE	**Df Model:**	7
Date:	Sun, 09 Jan 2022	**Pseudo R-squ.:**	0.3612
Time:	06:05:29	**Log-Likelihood:**	-1910.1
converged:	True	**LL-Null:**	-2990.2
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.9292	0.055	-17.049	0.000	-1.036	-0.822
Equity_Paid_Up	0.7415	0.064	11.592	0.000	0.616	0.867
Networth	-3.5765	0.154	-23.219	0.000	-3.878	-3.275
Total_Debt	0.3860	0.098	3.937	0.000	0.194	0.578
Gross_Block_	0.9027	0.107	8.443	0.000	0.693	1.112
Net_Working_Capital_	-0.7301	0.091	-7.993	0.000	-0.909	-0.551
Current_Liabilities_and_Provisions_	0.6546	0.108	6.073	0.000	0.443	0.866
Other_Income	0.2363	0.077	3.068	0.002	0.085	0.387

p-value is low for all the top features suggested earlier by the RFE model. We reject the null hyphothesis that these aren't good indicators of default

Validating on resampled train set



The model, while in training, has called 1798 defaults correctly. When it calls so, the prediction is right 83% of the time.

The recall rate is very healthy. Also, 359 defaults and 278 non-defaults are still misclassified but it's a trade-off.

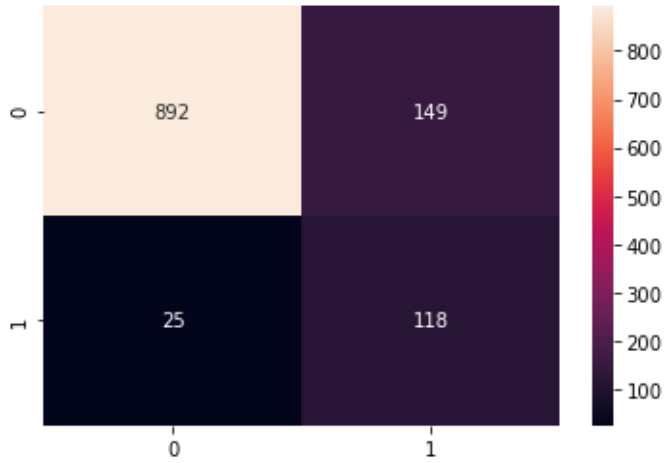
The model is not overfitting and the test accuracy is also a good 85%.

Tab 35

Classification report (train)

	precision	recall	f1-score	support
0.0	0.84	0.87	0.86	2157
1.0	0.87	0.83	0.85	2157
accuracy			0.85	4314
macro avg	0.85	0.85	0.85	4314
weighted avg	0.85	0.85	0.85	4314

Validating on test set



Classification report (test)

	precision	recall	f1-score	support
0.0	0.97	0.86	0.91	1041
1.0	0.44	0.83	0.58	143
accuracy			0.85	1184
macro avg	0.71	0.84	0.74	1184
weighted avg	0.91	0.85	0.87	1184

Conclusion

We can see that we get better recall value after balancing the data but precision is now a problem. This trade-off between recall and precision can be approached by adjusting the threshold. At present, we are using a threshold of 0.5.