Finance and Risk Analytics Project

Business Report

Yedupati Venkata Yamini

8th October 2023

Table of Contents

List of figures

List of tables

1. 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 interest 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.

Dependent variable - No need to create any new variable, as the 'Default' variable is already provided in the dataset, which can be considered as the dependent variable.

Test Train Split - Split the data into train and test datasets in the ratio of 67:33 and use a random state of 42 (*random_state=42*). Model building is to be done on the train dataset and model validation is to be done on the test dataset.

1.1. Outlier and missing value treatment

Knowing the dataset:

- There are 2058 rows and 58 columns in the dataset. They belong to various metrics of 2058 companies.

No	Column Name	Description
1	Co_Code	Company Code
2	Co_Name	Company Name

3	_Operating_Expense_Rate	Operating Expense Rate: Operating Expenses/Net Sales. The operating expense ratio (OER) is the cost to operate a piece of property compared to the income the property brings in.
4	_Research_and_development_expense_rate	Research and development expense rate: (Research and Development Expenses)/Net Sales. Research and development (R&D) expenses are direct expenditures relating to a company's efforts to develop, design, and enhance its products, services, technologies, or processes.
5	_Cash_flow_rate	Cash flow rate: Cash Flow from Operating/Current Liabilities. Cash flow is a measure of how much cash a business brought in or spent in total over a period of time.
6	_Interest_bearing_debt_interest_rate	Interest-bearing debt interest rate: Interest-bearing Debt/Equity
7	_Tax_rate_A	Tax rate (A): Effective Tax Rate. Effective tax rate represents the percentage of their taxable income that individuals pay in taxes. For corporations, the effective corporate tax rate is the rate they pay on their pre-tax profits.
8	_Cash_Flow_Per_Share	Cash Flow Per Share. It is the after-tax earnings plus depreciation on a per-share basis that functions as a measure of a firm's financial strength
9	_Per_Share_Net_profit_before_tax_Yuan_	Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share. Pretax income, also known as earnings before tax or pretax earnings, is the net income earned by a business before taxes are subtracted/accounted for.
10	_Realized_Sales_Gross_Profit_Growth_Rate	Realized Sales Gross Profit Growth Rate.
11	_Operating_Profit_Growth_Rate	Operating Profit Growth Rate: Operating Income Growth. It is the rate of increase in operating income over the last year.
12	_Continuous_Net_Profit_	Continuous Net Profit Growth Rate: Net

	Growth_Rate	Income-Excluding Disposal Gain or Loss Growth
13	_Total_Asset_Growth_Rate	Total Asset Growth Rate: Total Asset Growth. It is the rate at which how quickly the company has been growing its Assets
14	_Net_Value_Growth_Rate	Net Value Growth Rate: Total Equity Growth
15	_Total_Asset_Return_Growth_Rate_Ratio	Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
16	_Cash_Reinvestment_perc	Cash Reinvestment %: Cash Reinvestment Ratio. It is the valuation ratio that is used to measure the percentage of annual cash flow that the company invests back into the business as a new investment.
17	_Current_Ratio	Current Ratio. The current ratio describes the relationship between a company's assets and liabilities
18	_Quick_Ratio	Quick Ratio: Acid Test. Acid-test ratio (also known as quick ratio) is a measure of a company's liquidity, which is its ability to pay its short-term obligations using only its most liquid assets.
19	_Interest_Expense_Ratio	Interest Expense Ratio: Interest Expenses/Total Revenue
20	_Total_debt_to_Total_net_worth	Total debt/Total net worth: Total Liability/Equity Ratio
21	_Long_term_fund_suitability_ratio_A	Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
22	_Net_profit_before_tax_to_Paid_in_capital	Net profit before tax/Paid-in capital: Pretax Income/Capital
23	_Total_Asset_Turnover	Total Asset Turnover. Net Sales/Average Total Assets
24	_Accounts_Receivable_Turnover	Accounts Receivable Turnover. The accounts receivable turnover ratio, or receivables turnover, is used in business accounting to quantify how well companies are managing the credit that they extend to their customers by evaluating how long it takes to collect the outstanding debt throughout the accounting period.

25	_Average_Collection_Days	Average Collection Days: Days Receivable Outstanding
26	_Inventory_Turnover_Rate_times	Inventory Turnover Rate (times). The inventory turnover ratio is the number of times a company has sold and replenished its inventory over a specific amount of time. The formula can also be used to calculate the number of days it will take to sell the inventory on hand.
27	_Fixed_Assets_Turnover_Frequency	Fixed Assets Turnover Frequency. Fixed Asset Turnover (FAT) is an efficiency ratio that indicates how well or efficiently a business uses fixed assets to generate sales. This ratio divides net sales by net fixed assets, calculated over an annual period.
28	_Net_Worth_Turnover_Rate_times	Net Worth Turnover Rate (times): Equity Turnover. Equity turnover is a ratio that measures the proportion of a company's sales to its stockholders' equity. The intent of the measurement is to determine the efficiency with which management is using equity to generate revenue.
29	_Operating_profit_per_person	Operating profit per person: Operation Income Per Employee
30	_Allocation_rate_per_person	Allocation rate per person: Fixed Assets Per Employee
31	_Quick_Assets_to_Total_Assets	Quick Assets/Total Assets
32	_Cash_to_Total_Assets	Cash/Total Assets
33	_Quick_Assets_to_Current_Liability	Quick Assets/Current Liability
34	_Cash_to_Current_Liability	Cash/Current Liability
35	_Operating_Funds_to_Liability	Operating Funds to Liability
36	_Inventory_to_Working_Capital	Inventory/Working Capital

37	_Inventory_to_Current_Liability	Inventory/Current Liability
38	_Long_term_Liability_to_Current_Assets	Long-term Liability to Current Assets
39	_Retained_Earnings_to_Total_Assets	Retained Earnings to Total Assets
40	_Total_income_to_Total_expense	Total income/Total expense
41	_Total_expense_to_Assets	Total expense/Assets
42	_Current_Asset_Turnover_Rate	Current Asset Turnover Rate: Current Assets to Sales. The current assets turnover ratio indicates how many times the current assets are turned over in the form of sales within a specific period of time. A higher asset turnover ratio means a better percentage of sales.
43	_Quick_Asset_Turnover_Rate	Quick Asset Turnover Rate: Quick Assets to Sales. The asset turnover ratio measures the efficiency of a company's assets in generating revenue or sales.
44	_Cash_Turnover_Rate	Cash Turnover Rate: Cash to Sales. The cash turnover ratio is an efficiency ratio that reveals the number of times that cash is turned over in an accounting period.
45	_Fixed_Assets_to_Assets	Fixed Assets to Assets. Fixed assets are also known as non-current assets—assets that can't be easily converted into cash.
46	_Cash_Flow_to_Total_Assets	Cash Flow to Total Assets. This ratio indicates the cash a company can generate in relation to its size.
47	_Cash_Flow_to_Liability	Cash Flow to Liability. The amount of money available to run business operations and complete transactions. This is calculated as current assets (cash or near-cash assets, like notes receivable) minus current liabilities (liabilities due during the upcoming accounting period)
48	_CFO_to_Assets	CFO to Assets. Cash flow on total assets is an efficiency ratio that rates cash flows to the company assets without

		being affected by income recognition or income measurements.
49	_Cash_Flow_to_Equity	Cash Flow to Equity. cash flow to equity is a measure of how much cash is available to the equity shareholders of a company after all expenses, reinvestment, and debt are paid.
50	_Current_Liability_to_Current_Assets	Current Liability to Current Assets. Current liabilities are a company's financial commitments that are due and payable within a year, Current assets are projected to be consumed, sold, or converted into cash within a year or within the operational cycle.
51	_Liability_Assets_Flag	Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
52	_Total_assets_to_GNP_price	Total assets to GNP price. Gross National Product (GNP) is the total value of all finished goods and services produced by a country’s citizens in a given financial year, irrespective of their location.
53	_No_credit_Interval	No-credit Interval
54	_Degree_of_Financial_Leverage_DFL	Degree of Financial Leverage (DFL). The degree of financial leverage is a financial ratio that measures the sensitivity in fluctuations of a company's overall profitability to the volatility of its operating income caused by changes in its capital structure.
55	_Interest_Coverage_Ratio_Interest_expense_to_EBIT	Interest Coverage Ratio (Interest expense to EBIT). The interest coverage ratio is a debt and profitability ratio used to determine how easily a company can pay interest on its outstanding debt. The interest coverage ratio is calculated by dividing a company's earnings before interest and taxes (EBIT) by its interest expense during a given period.
56	_Net_Income_Flag	Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
57	_Equity_to_Liability	Equity to Liability Ratio.

58	Default	Whether the Company has Default (Bankrupted) or not? 1 - Defaulted, 0 - Not Defaulted.

Table 1: Data dictionary - Credit Risk dataset

- Viewing the first 5 rows of the dataset(for better view the rows are represented as columns below):

	0	1	2	3	4
Co_Code	16974	21214	14852	2439	23505
Co_Name	Hind.Cables	Tata Tele. Mah.	ABG Shipyard	GTL	Bharati Defence
_Operating_Expense_Rate	8820000000.0	9380000000.0	3800000000.0	6440000000.0	3680000000.0
_Research_and_development_expense_rate	0.0	4230000000.0	815000000.0	0.0	0.0
_Cash_flow_rate	0.462045	0.460116	0.449893	0.462731	0.463117
_Interest_bearing_debt_interest_rate	0.000352	0.000716	0.000496	0.000592	0.000782
_Tax_rate_A	0.001417	0.0	0.0	0.009313	0.400243
_Cash_Flow_Per_Share	0.322558	0.31552	0.299851	0.319834	0.325104
_Per_Share_Net_profit_before_tax_Yuan_	0.194472	0.161633	0.172554	0.174738	0.176546
_Realized_Sales_Gross_Profit_Growth_Rate	0.022074	0.021902	0.022186	0.027638	0.022072
_Operating_Profit_Growth_Rate	0.848021	0.839645	0.848196	0.848391	0.847987
_Continuous_Net_Profit_Growth_Rate	0.21759	0.21736	0.217573	0.217662	0.217589
_Total_Asset_Growth_Rate	7500000000.0	6750000000.0	9680000000.0	7520000000.0	7120000000.0
_Net_Value_Growth_Rate	0.000441	0.000403	0.000452	0.000448	0.000454
_Total_Asset_Return_Growth_Rate_Ratio	0.263902	0.263714	0.264095	0.264766	0.263966
_Cash_Reinvestment_perc	0.369137	0.372676	0.34886	0.379876	0.389609
_Current_Ratio	0.008324	0.006939	0.008669	0.01775	0.008427
_Quick_Ratio	0.000255	0.004787	0.005912	0.001738	0.003967
_Interest_Expense_Ratio	0.631513	0.628055	0.631688	0.632588	0.632682
_Total_debt_to_Total_net_worth	0.026006	0.006812	0.004105	0.007846	0.013671
_Long_term_fund_suitability_ratio_A	0.005767	0.00523	0.005139	0.01196	0.005822
_Net_profit_before_tax_to_Paid_in_capital	0.192859	0.160682	0.171548	0.172159	0.175598
_Total_Asset_Turnover	0.053973	0.056972	0.154423	0.101949	0.163418
_Accounts_Receivable_Turnover	0.014004	0.000306	0.001045	0.005411	0.000814
_Average_Collection_Days	0.000452	0.020645	0.006048	0.001169	0.007776
_Inventory_Turnover_Rate_times	707000000.0	0.000278	0.00017	1340000000.0	0.000134
_Fixed_Assets_Turnover_Frequency	0.000305	8850000000.0	0.000149	0.001827	0.00083

-

_Net_Worth_Turnover_Rate_times	0.029839	0.018387	0.029839	0.028387	0.052258
_Operating_profit_per_person	0.611689	0.386626	0.393263	0.43978	0.392766
_Allocation_rate_per_person	0.139494	0.022805	0.012358	0.009049	0.002069
_Quick_Assets_to_Total_Assets	0.176438	0.40204	0.318921	0.137092	0.739193
_Cash_to_Total_Assets	0.025626	0.004529	0.008242	0.05351	0.082328
_Quick_Assets_to_Current_Liability	0.001509	0.006584	0.00609	0.002437	0.007271
_Cash_to_Current_Liability	0.000676	0.000216	0.000458	0.002793	0.002376
_Operating_Funds_to_Liability	0.342391	0.337476	0.306993	0.3435	0.345796
_Inventory_to_Working_Capital	0.278434	0.277221	0.277473	0.27763	0.277235
_Inventory_to_Current_Liability	0.017945	0.001271	0.007012	0.039872	0.003342
_Long_term_Liability_to_Current_Assets	0.003064	0.004813	0.0	0.004472	0.0
_Retained_Earnings_to_Total_Assets	0.93763	0.926251	0.933155	0.928037	0.934421
_Total_income_to_Total_expense	0.002687	0.002044	0.002324	0.002334	0.00231
_Total_expense_to_Assets	0.007059	0.015441	0.009771	0.013607	0.010493
_Current_Asset_Turnover_Rate	0.000732	0.000301	0.000127	0.000401	0.000208
_Quick_Asset_Turnover_Rate	0.000142	0.000299	941000000.0	5310000000.0	0.000189
_Cash_Turnover_Rate	5470000000.0	882000000.0	679000000.0	6020000000.0	5670000000.0
_Fixed_Assets_to_Assets	0.09427	0.351895	0.463276	0.026433	0.103303
_Cash_Flow_to_Total_Assets	0.632666	0.642967	0.644486	0.656832	0.656549
_Cash_Flow_to_Liability	0.458073	0.459282	0.4597	0.46186	0.461238
_CFO_to_Assets	0.576869	0.551523	0.463045	0.577212	0.594038
_Cash_Flow_to_Equity	0.310901	0.314572	0.314777	0.316974	0.317729
_Current_Liability_to_Current_Assets	0.034913	0.041653	0.03356	0.016527	0.034497
_Liability_Assets_Flag	0	0	0	0	0
_Total_assets_to_GNP_price	0.028801	0.006191	0.001095	0.003749	0.006595
_No_credit_Interval	0.620927	0.622513	0.623749	0.622963	0.624419
_Degree_of_Financial_Leverage_DFL	0.02693	0.026297	0.027276	0.026988	0.027498
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	0.565744	0.560741	0.566744	0.56595	0.567177
_Net_Income_Flag	1	1	1	1	1
_Equity_to_Liability	0.015338	0.029445	0.041718	0.026956	0.0199
Default	0	1	0	0	0

Table 2: First 5 rows of credit risk dataset

- Viewing the basic info of the dataset:
    - <class 'pandas.core.frame.DataFrame'>
    - RangeIndex: 2058 entries, 0 to 2057
    - Data columns (total 58 columns):
    - # Column Non-Null Count Dtype
    - --- ------ -------------- -----
    - 0 Co_Code 2058 non-null int64
    - 1 Co_Name 2058 non-null object
    - 2 _Operating_Expense_Rate 2058 non-null float64
    - 3 _Research_and_development_expense_rate 2058 non-null float64
    - 4 _Cash_flow_rate 2058 non-null float64
    - 5 _Interest_bearing_debt_interest_rate 2058 non-null float64
    - 6 _Tax_rate_A 2058 non-null float64
    - 7 _Cash_Flow_Per_Share 1891 non-null float64
    - 8 _Per_Share_Net_profit_before_tax_Yuan_ 2058 non-null float64
    - 9 _Realized_Sales_Gross_Profit_Growth_Rate 2058 non-null float64
    - 10 _Operating_Profit_Growth_Rate 2058 non-null float64
    - 11 _Continuous_Net_Profit_Growth_Rate 2058 non-null float64
    - 12 _Total_Asset_Growth_Rate 2058 non-null float64
    - 13 _Net_Value_Growth_Rate 2058 non-null float64
    - 14 _Total_Asset_Return_Growth_Rate_Ratio 2058 non-null float64
    - 15 _Cash_Reinvestment_perc 2058 non-null float64
    - 16 _Current_Ratio 2058 non-null float64
    - 17 _Quick_Ratio 2058 non-null float64
    - 18 _Interest_Expense_Ratio 2058 non-null float64
    - 19 _Total_debt_to_Total_net_worth 2037 non-null float64
    - 20 _Long_term_fund_suitability_ratio_A 2058 non-null float64
    - 21 _Net_profit_before_tax_to_Paid_in_capital 2058 non-null float64

- 22 _Total_Asset_Turnover 2058 non-null float64
- 23 _Accounts_Receivable_Turnover 2058 non-null float64
- 24 _Average_Collection_Days 2058 non-null float64
- 25 _Inventory_Turnover_Rate_times 2058 non-null float64
- 26 _Fixed_Assets_Turnover_Frequency 2058 non-null float64
- 27 _Net_Worth_Turnover_Rate_times 2058 non-null float64
- 28 _Operating_profit_per_person 2058 non-null float64
- 29 _Allocation_rate_per_person 2058 non-null float64
- 30 _Quick_Assets_to_Total_Assets 2058 non-null float64
- 31 _Cash_to_Total_Assets 1962 non-null float64
- 32 _Quick_Assets_to_Current_Liability 2058 non-null float64
- 33 _Cash_to_Current_Liability 2058 non-null float64
- 34 _Operating_Funds_to_Liability 2058 non-null float64
- 35 _Inventory_to_Working_Capital 2058 non-null float64
- 36 _Inventory_to_Current_Liability 2058 non-null float64
- 37 _Long_term_Liability_to_Current_Assets 2058 non-null float64
- 38 _Retained_Earnings_to_Total_Assets 2058 non-null float64
- 39 _Total_income_to_Total_expense 2058 non-null float64
- 40 _Total_expense_to_Assets 2058 non-null float64
- 41 _Current_Asset_Turnover_Rate 2058 non-null float64
- 42 _Quick_Asset_Turnover_Rate 2058 non-null float64
- 43 _Cash_Turnover_Rate 2058 non-null float64
- 44 _Fixed_Assets_to_Assets 2058 non-null float64
- 45 _Cash_Flow_to_Total_Assets 2058 non-null float64
- 46 _Cash_Flow_to_Liability 2058 non-null float64
- 47 _CFO_to_Assets 2058 non-null float64
- 48 _Cash_Flow_to_Equity 2058 non-null float64
- 49 _Current_Liability_to_Current_Assets 2044 non-null float64
- 50 _Liability_Assets_Flag 2058 non-null int64

- 51 _Total_assets_to_GNP_price 2058 non-null float64
- 52 _No_credit_Interval 2058 non-null float64
- 53 _Degree_of_Financial_Leverage_DFL 2058 non-null float64
- 54 _Interest_Coverage_Ratio_Interest_expense_to_EBIT 2058 non-null float64
- 55 _Net_Income_Flag 2058 non-null int64
- 56 _Equity_to_Liability 2058 non-null float64
- 57 Default 2058 non-null int64
- dtypes: float64(53), int64(4), object(1)
- memory usage: 932.7+ KB

- There are no duplicate rows in the dataset but there are null values for the below columns in the dataset with their counts:

```
_Cash_Flow_Per_Share                    167
_Total_debt_to_Total_net_worth           21
_Cash_to_Total_Assets                    96
_Current_Liability_to_Current_Assets     14
dtype: int64
```

Table 3: Null values in Credit Risk dataset

- Imputing the null values using KNN Imputer.

- The KNN imputer fills the null values in a particular column based on its nearest predictor variable neighbors in the dataset which have values filled.

- Viewing the information of the dataset after applying KNN imputer

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2058 entries, 0 to 2057

Data columns (total 58 columns):

#	Column	Non-Null Count	Dtype
0	_Operating_Expense_Rate	2058 non-null	float64
1	_Research_and_development_expense_rate	2058 non-null	float64
2	_Cash_flow_rate	2058 non-null	float64
3	_Interest_bearing_debt_interest_rate	2058 non-null	float64
4	_Tax_rate_A	2058 non-null	float64
5	_Cash_Flow_Per_Share	2058 non-null	float64
6	_Per_Share_Net_profit_before_tax_Yuan_	2058 non-null	float64
7	_Realized_Sales_Gross_Profit_Growth_Rate	2058 non-null	float64
8	_Operating_Profit_Growth_Rate	2058 non-null	float64
9	_Continuous_Net_Profit_Growth_Rate	2058 non-null	float64
10	_Total_Asset_Growth_Rate	2058 non-null	float64
11	_Net_Value_Growth_Rate	2058 non-null	float64
12	_Total_Asset_Return_Growth_Rate_Ratio	2058 non-null	float64
13	_Cash_Reinvestment_perc	2058 non-null	float64
14	_Current_Ratio	2058 non-null	float64
15	_Quick_Ratio	2058 non-null	float64
16	_Interest_Expense_Ratio	2058 non-null	float64
17	_Total_debt_to_Total_net_worth	2058 non-null	float64
18	_Long_term_fund_suitability_ratio_A	2058 non-null	float64
19	_Net_profit_before_tax_to_Paid_in_capital	2058 non-null	float64
20	_Total_Asset_Turnover	2058 non-null	float64

21	_Accounts_Receivable_Turnover	2058	non-null	float64
22	_Average_Collection_Days	2058	non-null	float64
23	_Inventory_Turnover_Rate_times	2058	non-null	float64
24	_Fixed_Assets_Turnover_Frequency	2058	non-null	float64
25	_Net_Worth_Turnover_Rate_times	2058	non-null	float64
26	_Operating_profit_per_person	2058	non-null	float64
27	_Allocation_rate_per_person	2058	non-null	float64
28	_Quick_Assets_to_Total_Assets	2058	non-null	float64
29	_Cash_to_Total_Assets	2058	non-null	float64
30	_Quick_Assets_to_Current_Liability	2058	non-null	float64
31	_Cash_to_Current_Liability	2058	non-null	float64
32	_Operating_Funds_to_Liability	2058	non-null	float64
33	_Inventory_to_Working_Capital	2058	non-null	float64
34	_Inventory_to_Current_Liability	2058	non-null	float64
35	_Long_term_Liability_to_Current_Assets	2058	non-null	float64
36	_Retained_Earnings_to_Total_Assets	2058	non-null	float64
37	_Total_income_to_Total_expense	2058	non-null	float64
38	_Total_expense_to_Assets	2058	non-null	float64
39	_Current_Asset_Turnover_Rate	2058	non-null	float64
40	_Quick_Asset_Turnover_Rate	2058	non-null	float64
41	_Cash_Turnover_Rate	2058	non-null	float64
42	_Fixed_Assets_to_Assets	2058	non-null	float64
43	_Cash_Flow_to_Total_Assets	2058	non-null	float64
44	_Cash_Flow_to_Liability	2058	non-null	float64

45	_CFO_to_Assets	2058 non-null	float64
46	_Cash_Flow_to_Equity	2058 non-null	float64
47	_Current_Liability_to_Current_Assets	2058 non-null	float64
48	_Total_assets_to_GNP_price	2058 non-null	float64
49	_No_credit_Interval	2058 non-null	float64
50	_Degree_of_Financial_Leverage_DFL	2058 non-null	float64
51	_Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058 non-null	float64
52	_Equity_to_Liability	2058 non-null	float64
53	Co_Code	2058 non-null	int64
54	Co_Name	2058 non-null	object
55	Default	2058 non-null	int64
56	_Liability_Assets_Flag	2058 non-null	int64
57	_Net_Income_Flag	2058 non-null	int64

- As observed, all the null values have been treated.

Outlier treatment:

- Variables like Co_Code, Co_Name are not non significant for analysis. So they are not considered.

- Net_Income_Flag, Liability_Assets_Flag are categories represented as numerical variables, so they are not necessary for outlier treatment.

- 'Default' is the response variable and it is also not considered in outlier treatment.

- Initial glimpse of all the outliers for the remaining numerical variables in the dataset:

-

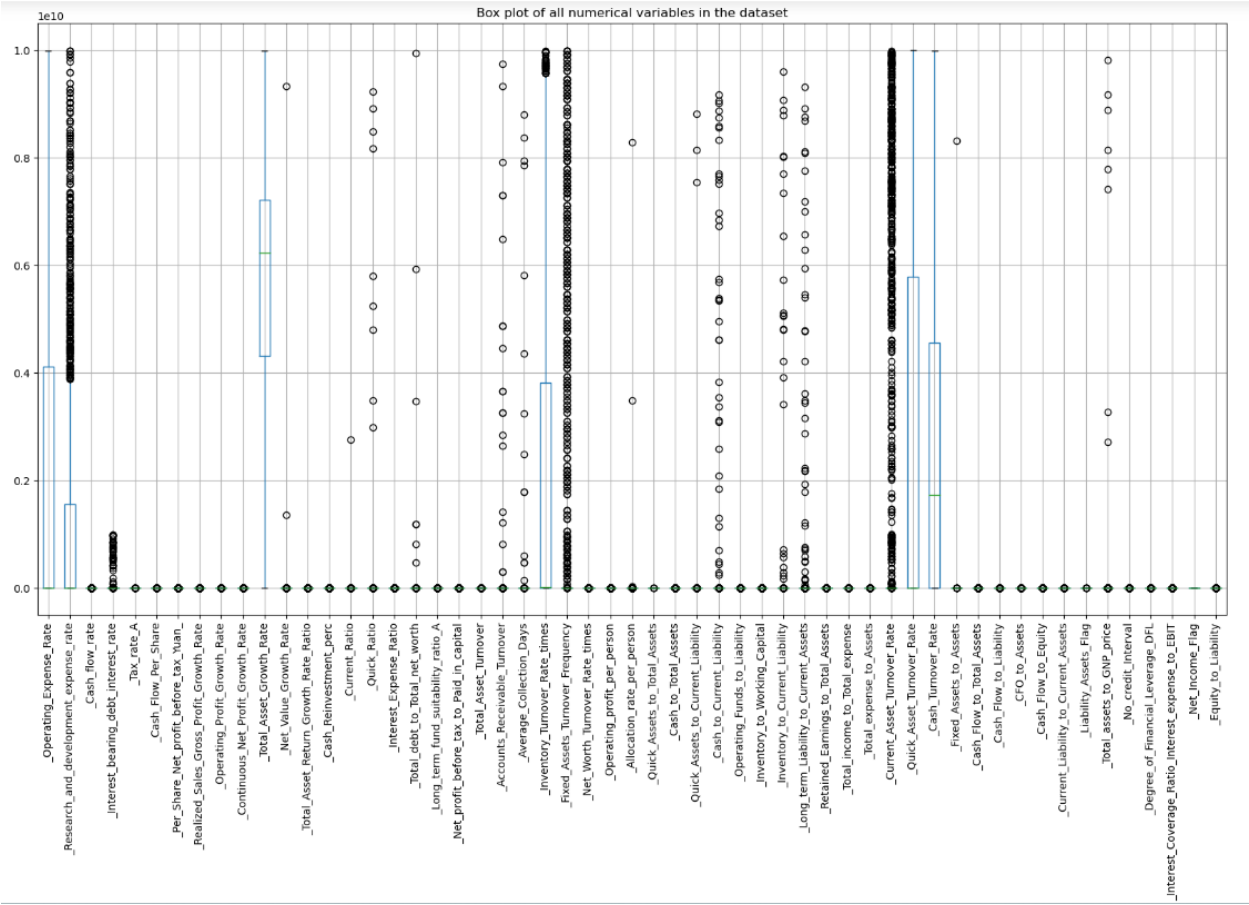


Figure 1: Box plot of numeric variables in CreditRisk dataset

- As we can see there are so many outliers in the dataset.

- Scaling the data using a robust scaler to bring all the numeric variables to a comparable scale.

- Robust scaler scales the data by centering the median and setting the IQR to 1. This can come handy in some cases when outliers are causing a problem.

- There are a total of 10891 outliers from all columns in the dataset and 50 rows consisting of these outliers.

- Applying robust scaling on the dataset with default parameters.

    - Did not remove outliers.

- Glimpse of the data after applying robust scaling:

	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_Cash_Flow_Per_Sha
0	2.146	-0.000	-0.176	-0.264	-0.165	0.1
1	2.282	2.729	-0.418	0.677	-0.172	-0.5
2	0.925	0.526	-1.700	0.109	-0.172	-2.0
3	1.567	-0.000	-0.090	0.357	-0.129	-0.0
4	0.895	-0.000	-0.041	0.848	1.680	0.4

Table 4: Robust scaled data

- Robust scaling has brought the median to 0 and IQR (75% - 25%) value to 1. Summary below

	count	mean	std	min	25%	50%	75%	max
_Operating_Expense_Rate	2058.0	4.993586e-01	7.913847e-01	-0.000	0.00000	0.0	1.00000	2.428000e+00
_Research_and_development_expense_rate	2058.0	7.797653e-01	1.383591e+00	-0.000	0.00000	0.0	1.00000	6.439000e+00
_Cash_flow_rate	2058.0	2.255476e-01	2.843516e+00	-58.149	-0.41975	0.0	0.58000	6.732200e+01
_Interest_bearing_debt_interest_rate	2058.0	2.875744e+10	2.336358e+11	-1.173	-0.46000	0.0	0.54000	2.557888e+12
_Tax_rate_A	2058.0	3.591220e-01	7.052793e-01	-0.172	-0.17200	0.0	0.82875	4.453000e+00
_Cash_Flow_Per_Share	2058.0	-7.025948e-02	1.424168e+00	-14.576	-0.52725	0.0	0.47300	1.362900e+01
_Per_Share_Net_profit_before_tax_Yuan_	2058.0	6.873081e-02	1.564045e+00	-9.109	-0.46900	0.0	0.53100	3.198400e+01
_Realized_Sales_Gross_Profit_Growth_Rate	2058.0	7.057287e+00	2.316384e+02	-190.190	-0.44500	0.0	0.55500	1.043817e+04
_Operating_Profit_Growth_Rate	2058.0	4.946526e-01	3.259710e+01	-792.774	-0.45925	0.0	0.54100	1.079407e+03
_Continuous_Net_Profit_Growth_Rate	2058.0	-4.478038e+00	1.242927e+02	-4762.575	-0.48100	0.0	0.51875	3.416260e+02
_Total_Asset_Growth_Rate	2058.0	-3.226866e-01	1.002652e+00	-2.143	-0.65725	0.0	0.34300	1.293000e+00
_Net_Value_Growth_Rate	2058.0	9.962095e+10	3.988900e+12	-8.742	-0.36700	0.0	0.63275	1.791045e+14
_Total_Asset_Return_Growth_Rate_Ratio	2058.0	1.474266e-01	4.228144e+00	-21.697	-0.48625	0.0	0.51375	1.656010e+02
_Cash_Reinvestment_perc	2058.0	-1.194227e-01	1.846033e+00	-23.816	-0.55550	0.0	0.44450	4.188200e+01
_Current_Ratio	2058.0	1.925887e+08	8.736821e+09	-1.289	-0.34275	0.0	0.65725	3.963475e+11
_Quick_Ratio	2058.0	4.659567e+09	7.468464e+10	-0.887	-0.39275	0.0	0.60775	1.549546e+12
_Interest_Expense_Ratio	2058.0	4.340675e-01	5.993789e+00	-93.344	-0.16600	0.0	0.83350	1.602040e+02
_Total_debt_to_Total_net_worth	2058.0	1.225726e+09	2.955073e+10	-0.797	-0.36600	0.0	0.63375	1.089656e+12
_Long_term_fund_suitability_ratio_A	2058.0	2.757825e+00	2.780873e+01	-1.108	-0.28300	0.0	0.71650	7.935100e+02
_Net_profit_before_tax_to_Paid_in_capital	2058.0	4.472546e-02	1.411184e+00	-9.394	-0.46875	0.0	0.53175	3.323200e+01
_Total_Asset_Turnover	2058.0	2.366628e-01	9.452819e-01	-0.972	-0.39400	0.0	0.60600	7.662000e+00
_Accounts_Receivable_Turnover	2058.0	3.748176e+10	4.548121e+11	-0.974	-0.30300	0.0	0.69700	8.776064e+12
_Average_Collection_Days	2058.0	5.194524e+09	8.118274e+10	-1.185	-0.47900	0.0	0.52100	1.738233e+12
_Inventory_Turnover_Rate_times	2058.0	5.271467e-01	8.066152e-01	-0.005	-0.00500	0.0	0.99475	2.614000e+00
_Fixed_Assets_Turnover_Frequency	2058.0	1.501951e+11	3.232682e+11	-0.073	-0.04500	-0.0	0.95450	1.218987e+12
_Net_Worth_Turnover_Rate_times	2058.0	4.552575e-01	1.776046e+00	-0.831	-0.34500	0.0	0.65500	4.068900e+01
_Operating_profit_per_person	2058.0	9.037362e-01	5.637322e+00	-41.560	-0.38800	0.0	0.61175	6.363400e+01

Table 5: Summary of robust scaled data

- Outlier analysis after applying robust scaling:

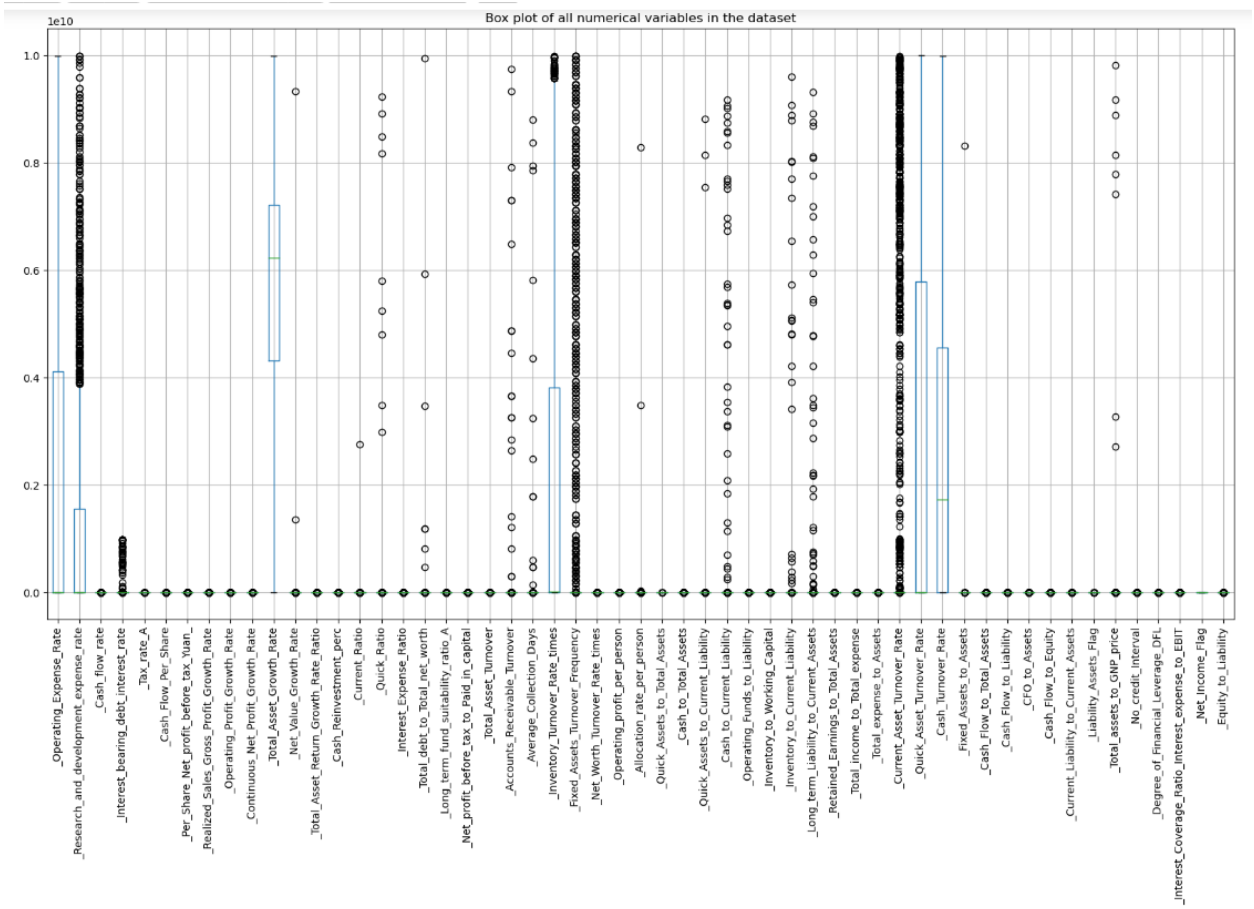


Figure 2: Outliers after applying Robust scaling

- As observed it did not remove outliers.
- So, to treat outliers, capped the values to lower and higher whisker values.
- Below is the outlier plot of the missing value and outlier treated scaled dataset.

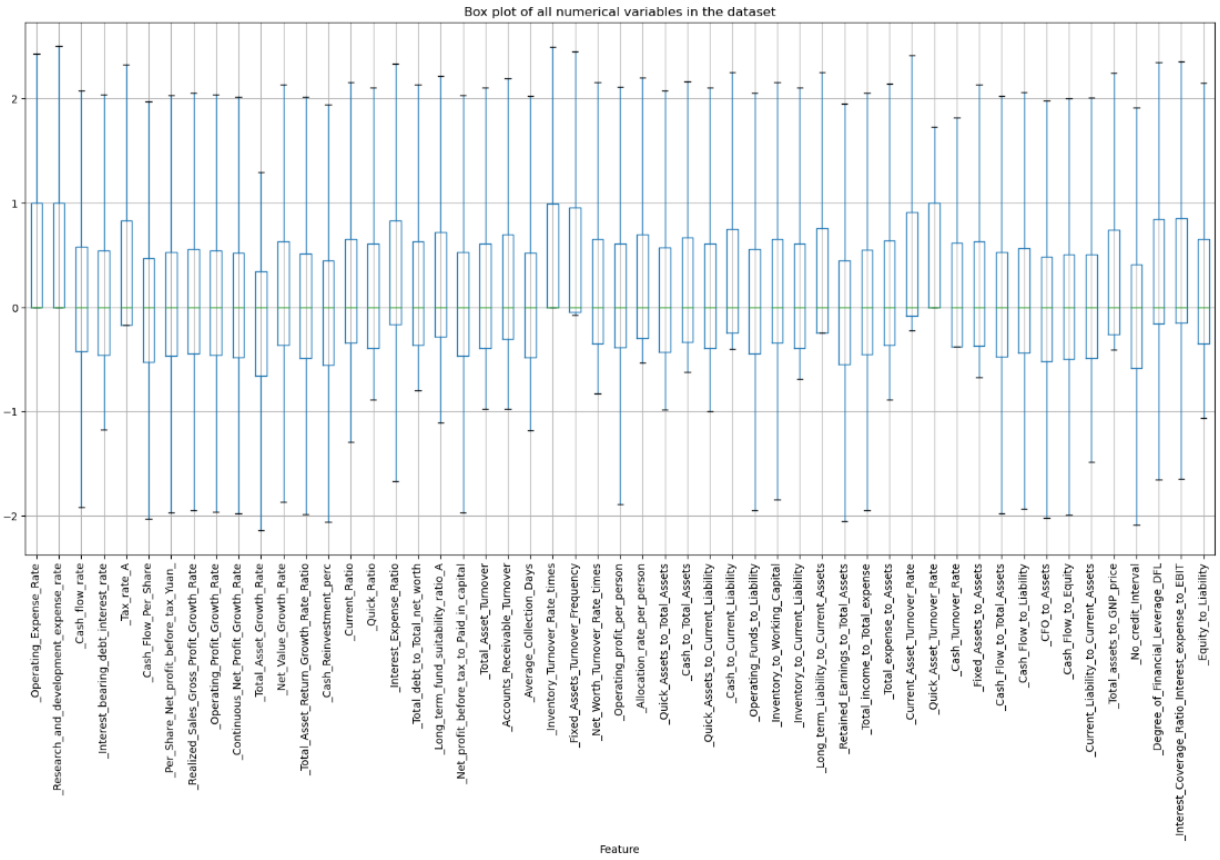


Figure 3: Outlier analysis after outlier treatment

- Finally, the outliers for the numerical variables have been treated and this treated data will be used for model building.

1.2. Univariate and Bivariate Analysis with Interpretation

Univariate analysis:

- **Distribution of defaulters and non defaulters in the dataset:**

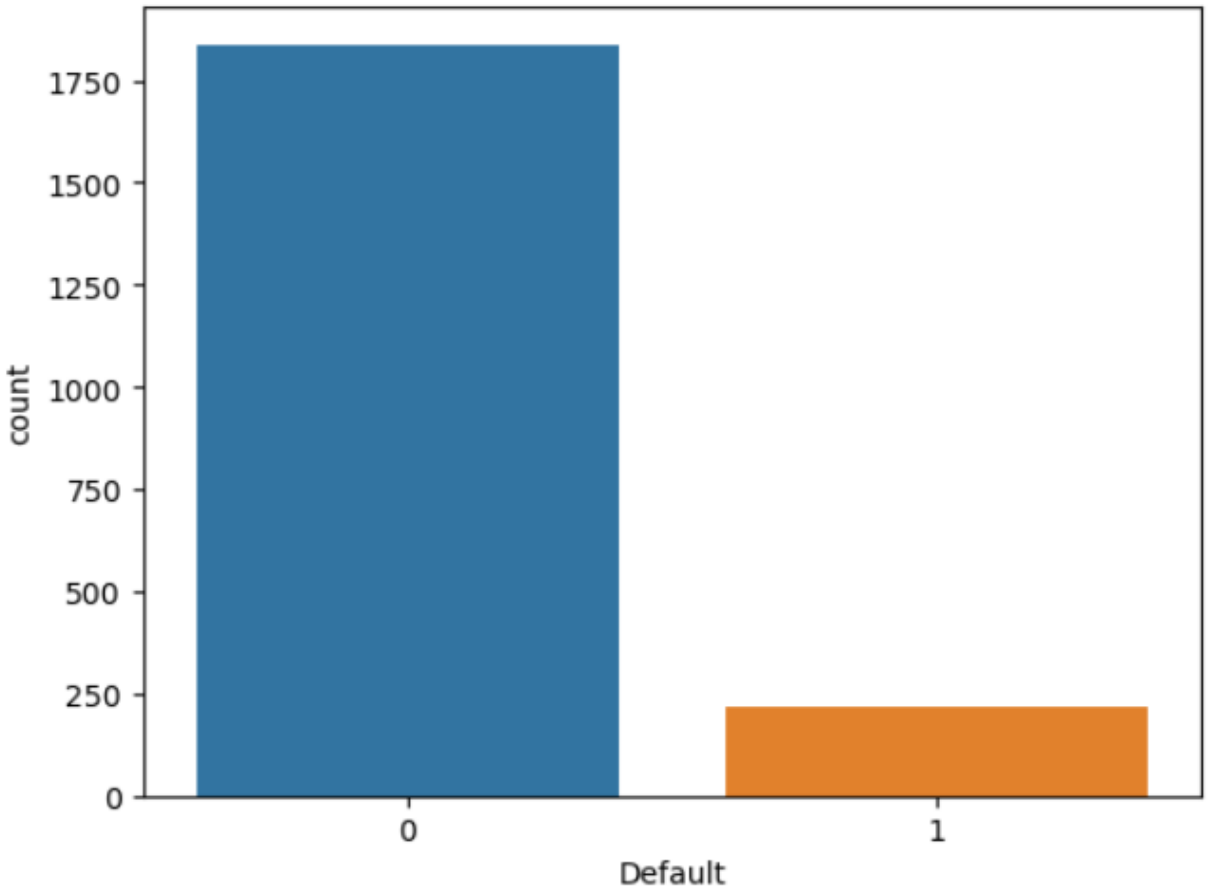
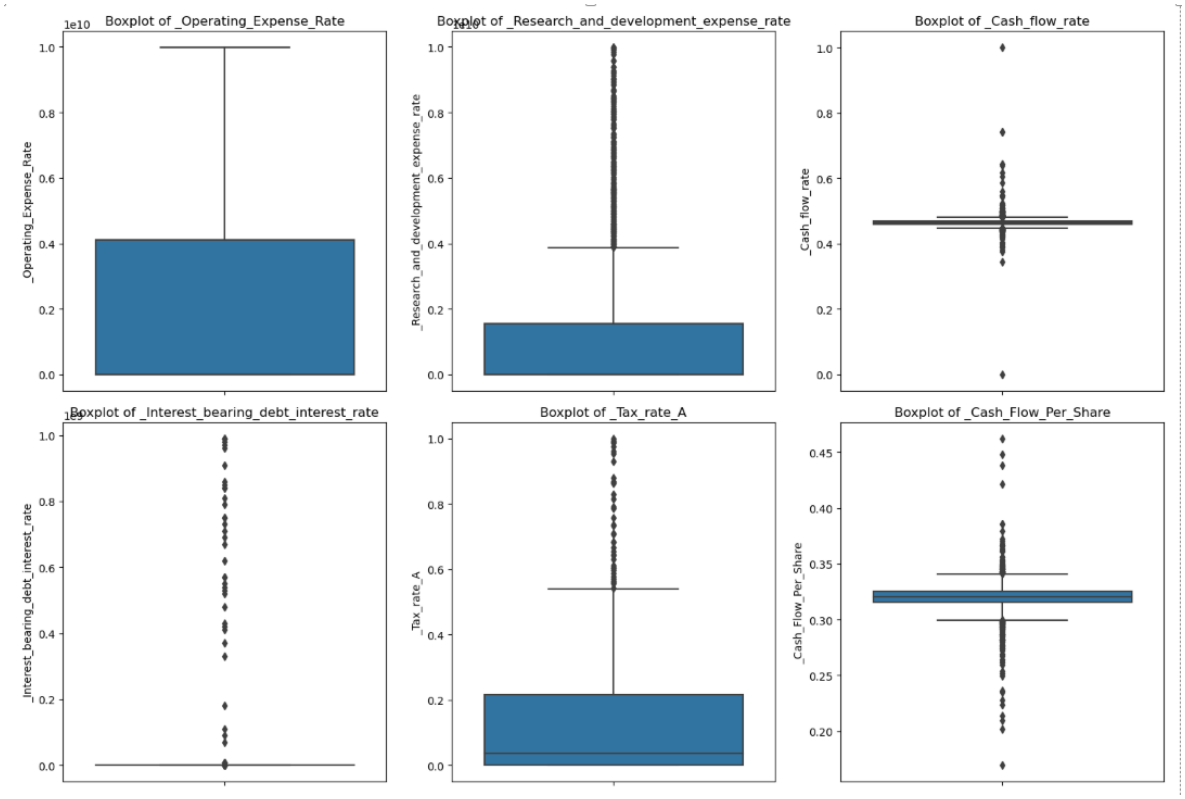
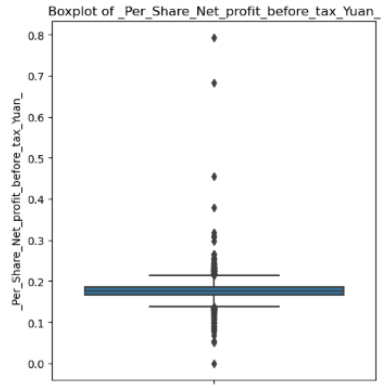


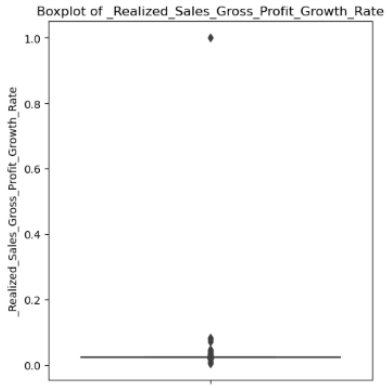
Figure 4: Distribution of defaulters and non defaulters

- Around 220 of the 2058 companies in the dataset have defaulted.

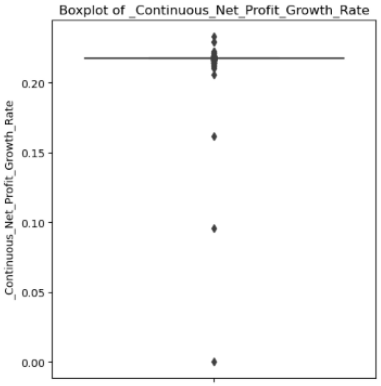
Box plots of all numeric variables:

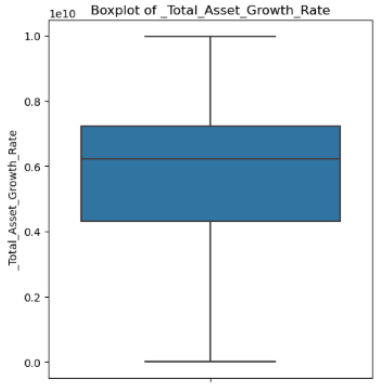


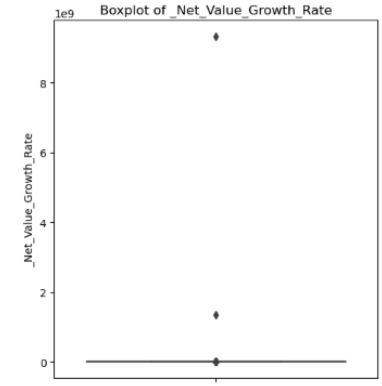

Boxplot of _Per_Share_Net_profit_before_tax_Yuan_
_Per_Share_Net_profit_before_tax_Yuan_

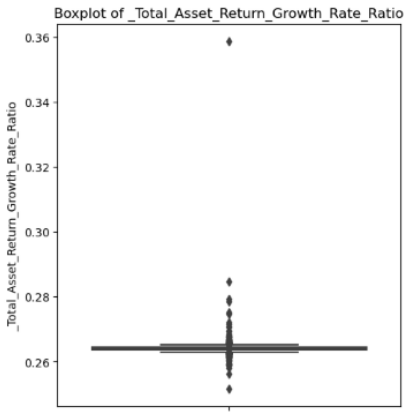

Boxplot of _Realized_Sales_Gross_Profit_Growth_Rate
_Realized_Sales_Gross_Profit_Growth_Rate

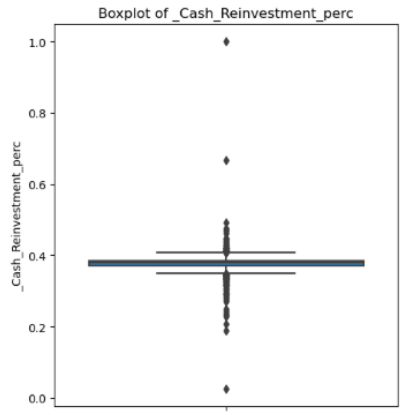

Boxplot of _Operating_Profit_Growth_Rate
_Operating_Profit_Growth_Rate

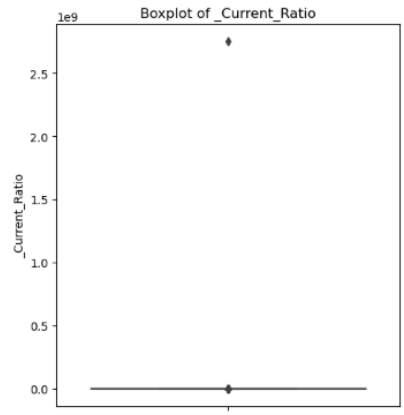

Boxplot of _Continuous_Net_Profit_Growth_Rate
_Continuous_Net_Profit_Growth_Rate

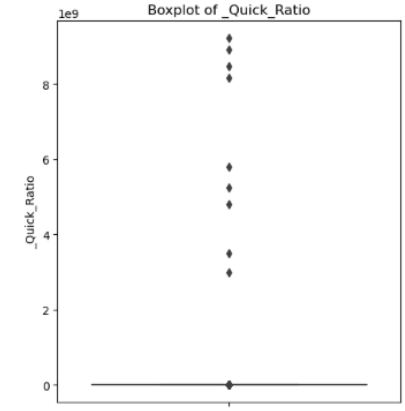

Boxplot of _Total_Asset_Growth_Rate
1e10
_Total_Asset_Growth_Rate

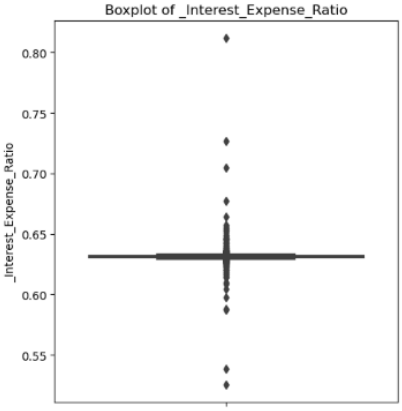

Boxplot of _Net_Value_Growth_Rate
1e9
_Net_Value_Growth_Rate

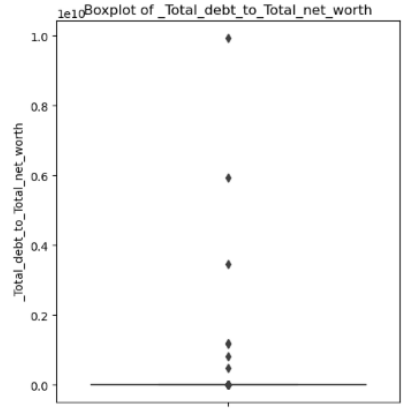

Boxplot of _Total_Asset_Return_Growth_Rate_Ratio
_Total_Asset_Return_Growth_Rate_Ratio

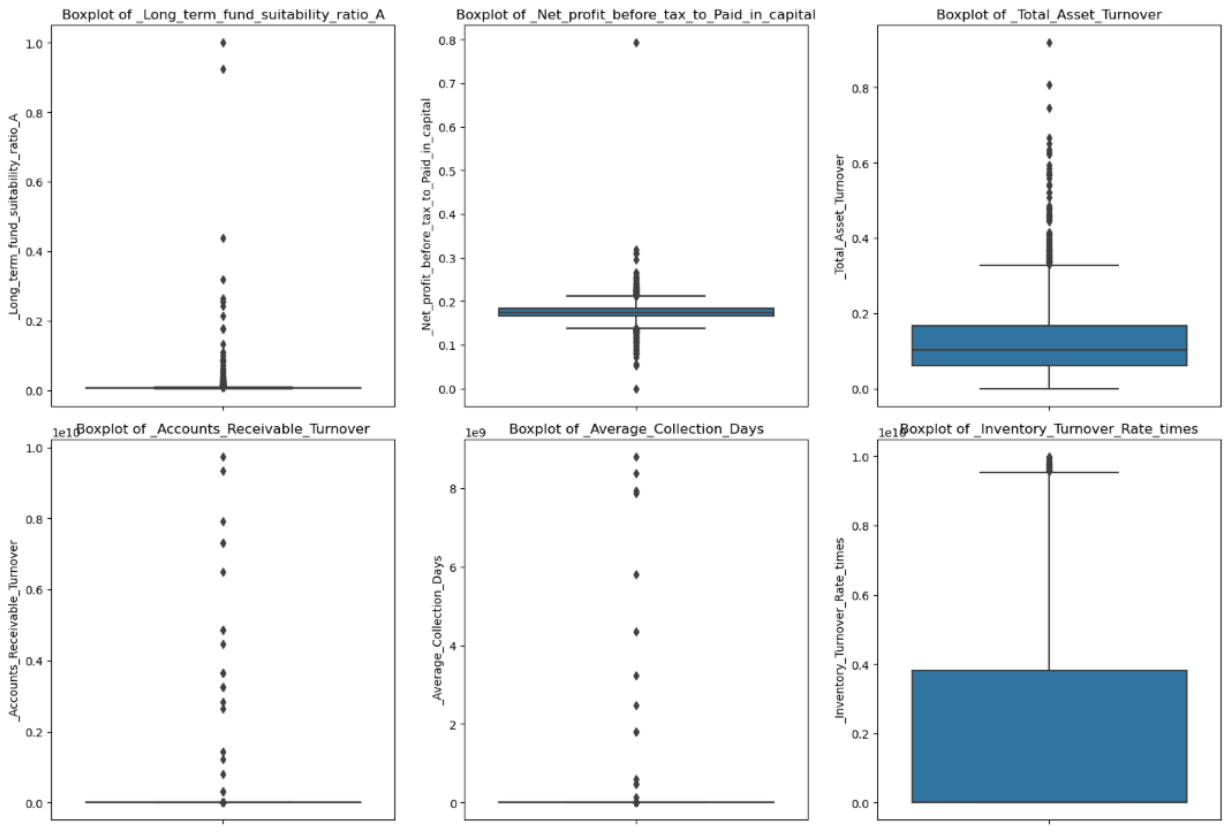

Boxplot of _Cash_Reinvestment_perc
_Cash_Reinvestment_perc

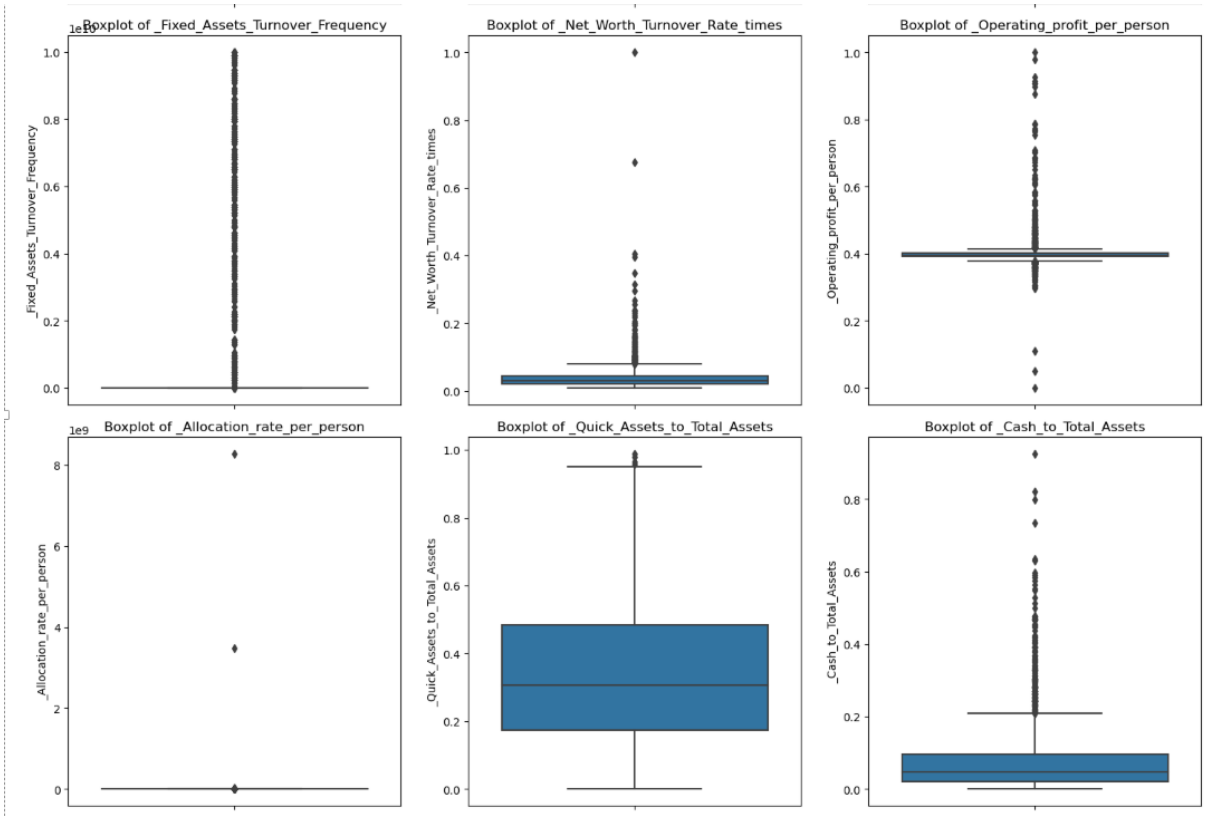

Boxplot of _Current_Ratio
1e9
_Current_Ratio

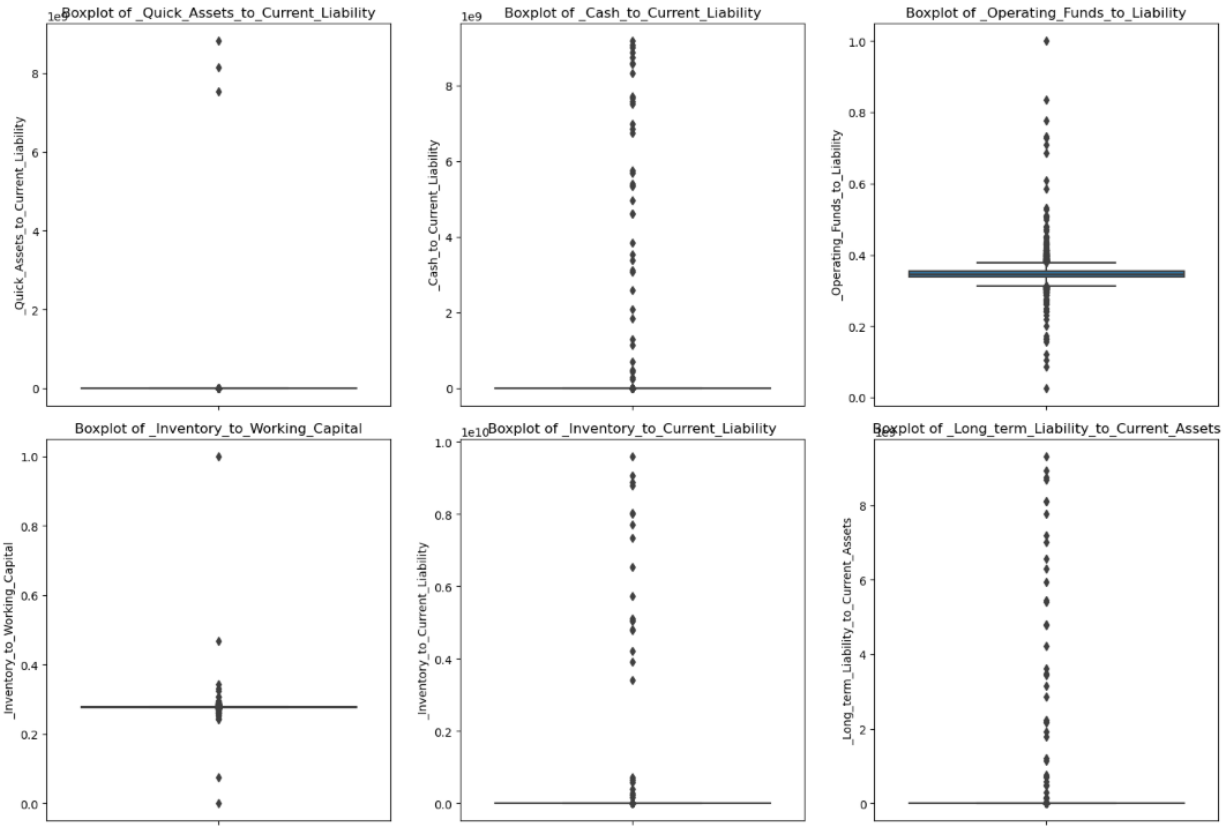

Boxplot of _Quick_Ratio
1e9
_Quick_Ratio

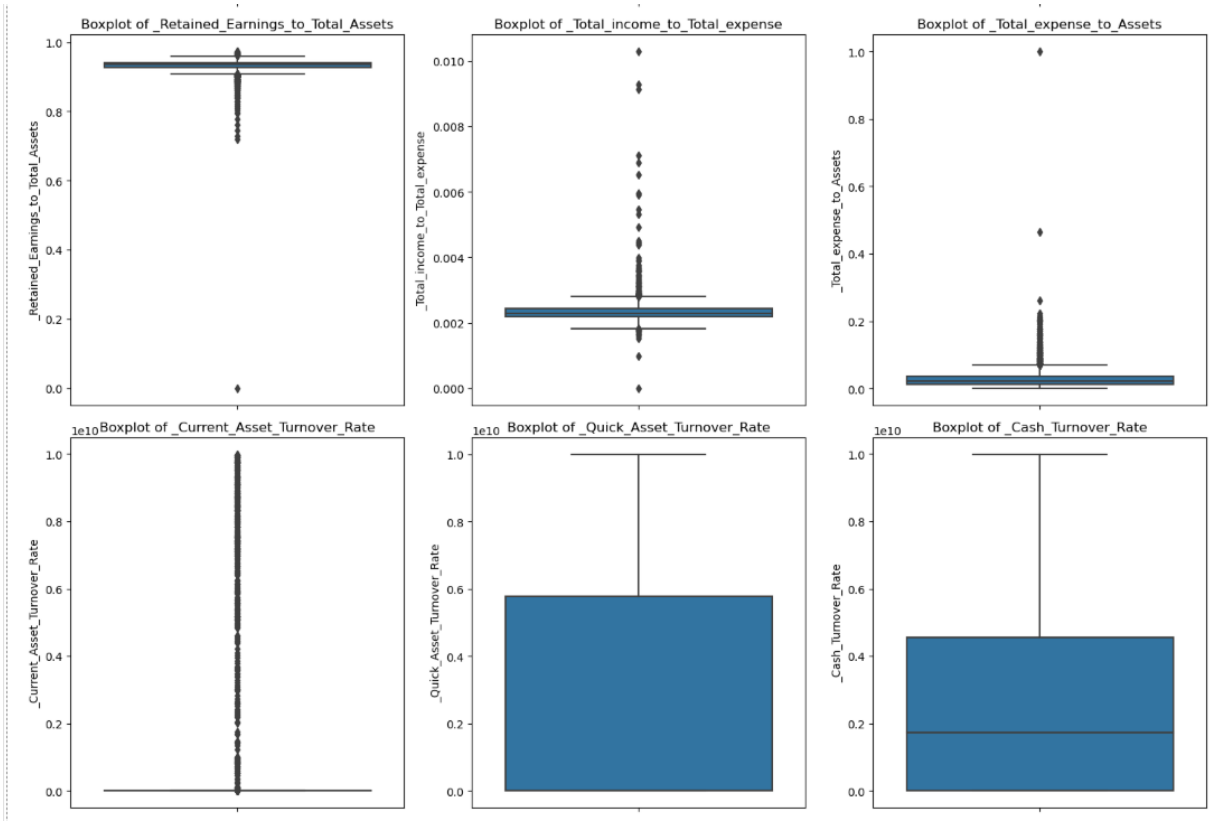

Boxplot of _Interest_Expense_Ratio
_Interest_Expense_Ratio

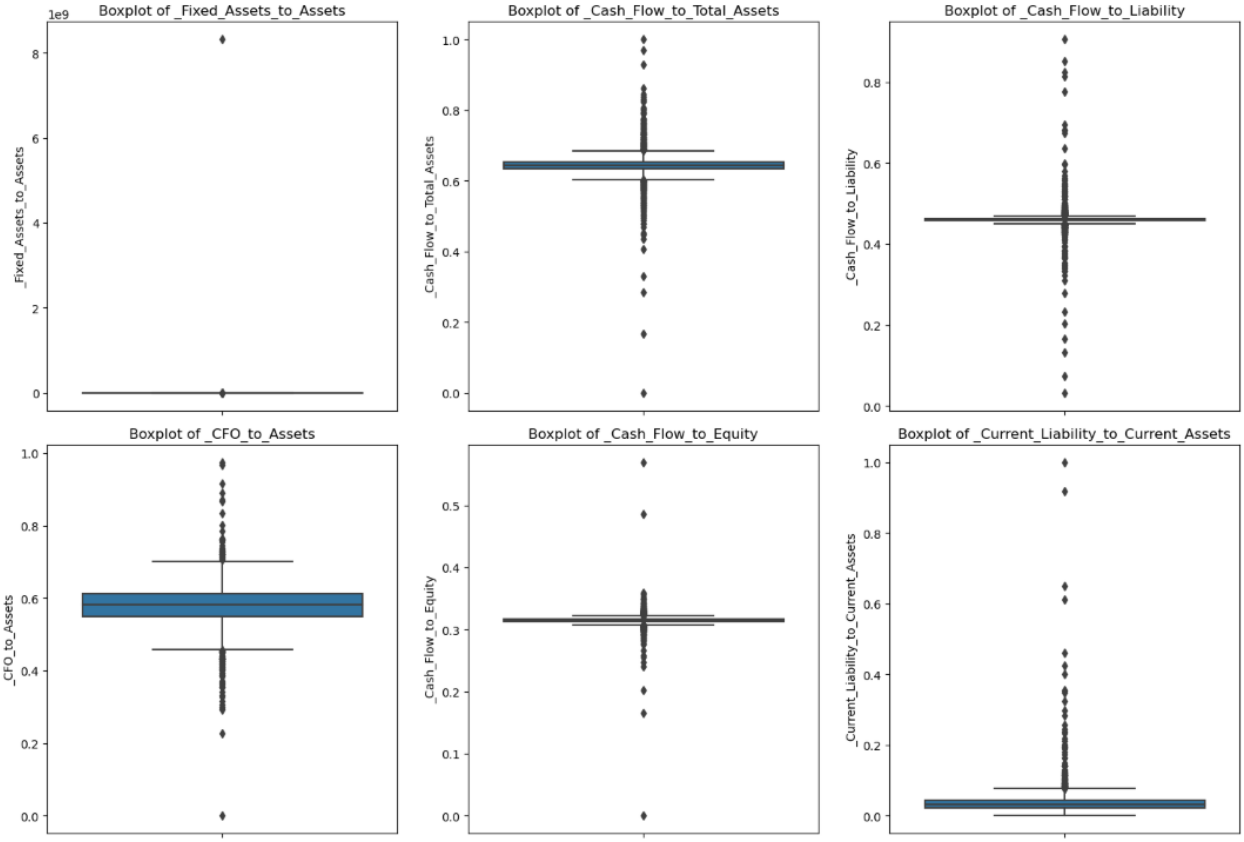

Boxplot of _Total_debt_to_Total_net_worth
1e10
_Total_debt_to_Total_net_worth


Boxplot of _Long_term_fund_suitability_ratio_A
_Long_term_fund_suitability_ratio_A
Boxplot of _Net_profit_before_tax_to_Paid_in_capital
_Net_profit_before_tax_to_Paid_in_capital
Boxplot of _Total_Asset_Turnover
_Total_Asset_Turnover
Boxplot of _Accounts_Receivable_Turnover
1e10
_Accounts_Receivable_Turnover
Boxplot of _Average_Collection_Days
1e9
_Average_Collection_Days
Boxplot of _Inventory_Turnover_Rate_times
1e10
_Inventory_Turnover_Rate_times


Boxplot of _Fixed_Assets_Turnover_Frequency
1e10
_Fixed_Assets_Turnover_Frequency
Boxplot of _Net_Worth_Turnover_Rate_times
_Net_Worth_Turnover_Rate_times
Boxplot of _Operating_profit_per_person
_Operating_profit_per_person
Boxplot of _Allocation_rate_per_person
1e9
_Allocation_rate_per_person
Boxplot of _Quick_Assets_to_Total_Assets
_Quick_Assets_to_Total_Assets
Boxplot of _Cash_to_Total_Assets
_Cash_to_Total_Assets


Boxplot of _Quick_Assets_to_Current_Liability
_Quick_Assets_to_Current_Liability
Boxplot of _Cash_to_Current_Liability
_Cash_to_Current_Liability
Boxplot of _Operating_Funds_to_Liability
_Operating_Funds_to_Liability
Boxplot of _Inventory_to_Working_Capital
_Inventory_to_Working_Capital
Boxplot of _Inventory_to_Current_Liability
_Inventory_to_Current_Liability
Boxplot of _Long_term_Liability_to_Current_Assets
_Long_term_Liability_to_Current_Assets


Boxplot of _Retained_Earnings_to_Total_Assets
_Retained_Earnings_to_Total_Assets
Boxplot of _Total_income_to_Total_expense
_Total_income_to_Total_expense
Boxplot of _Total_expense_to_Assets
_Total_expense_to_Assets
Boxplot of _Current_Asset_Turnover_Rate
_Current_Asset_Turnover_Rate
Boxplot of _Quick_Asset_Turnover_Rate
_Quick_Asset_Turnover_Rate
Boxplot of _Cash_Turnover_Rate
_Cash_Turnover_Rate


Boxplot of _Fixed_Assets_to_Assets
1e9
_Fixed_Assets_to_Assets
Boxplot of _Cash_Flow_to_Total_Assets
_Cash_Flow_to_Total_Assets
Boxplot of _Cash_Flow_to_Liability
_Cash_Flow_to_Liability
Boxplot of _CFO_to_Assets
_CFO_to_Assets
Boxplot of _Cash_Flow_to_Equity
_Cash_Flow_to_Equity
Boxplot of _Current_Liability_to_Current_Assets
_Current_Liability_to_Current_Assets

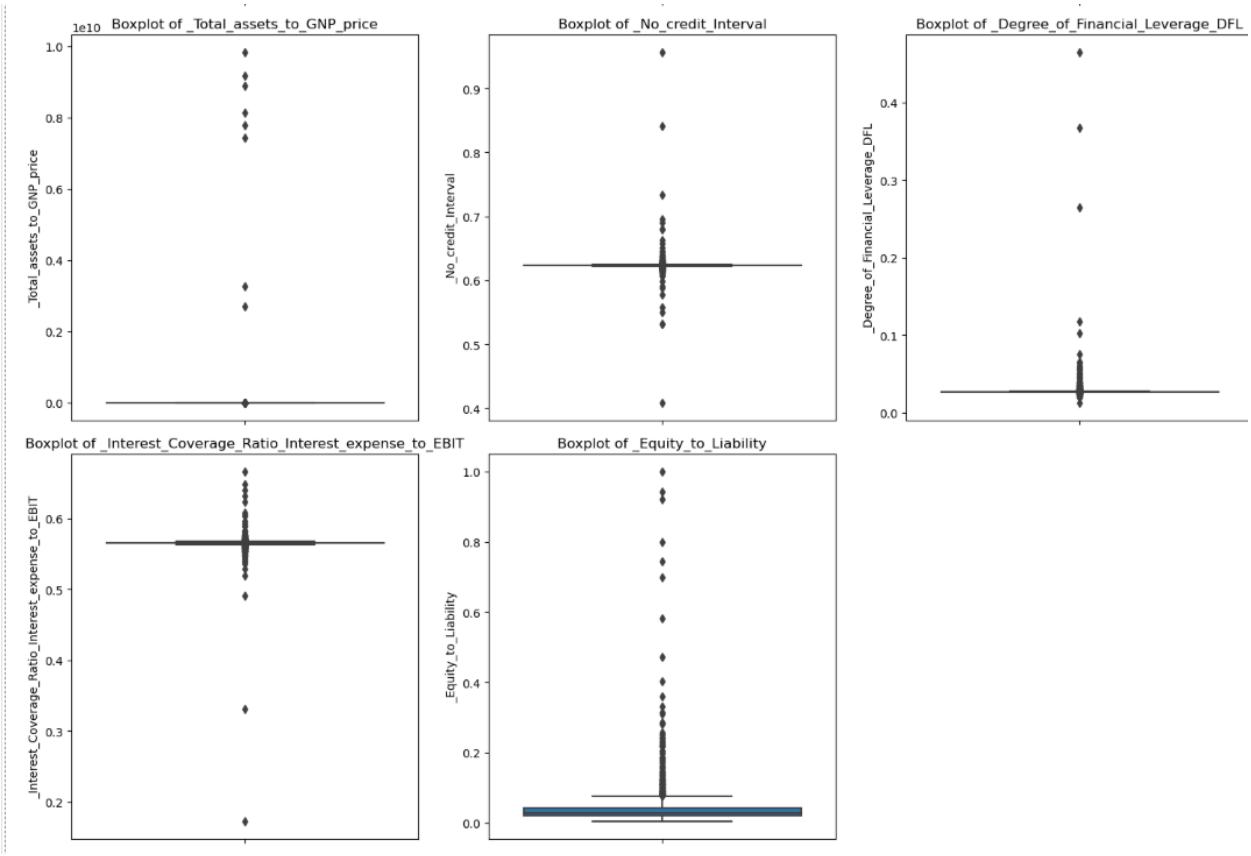
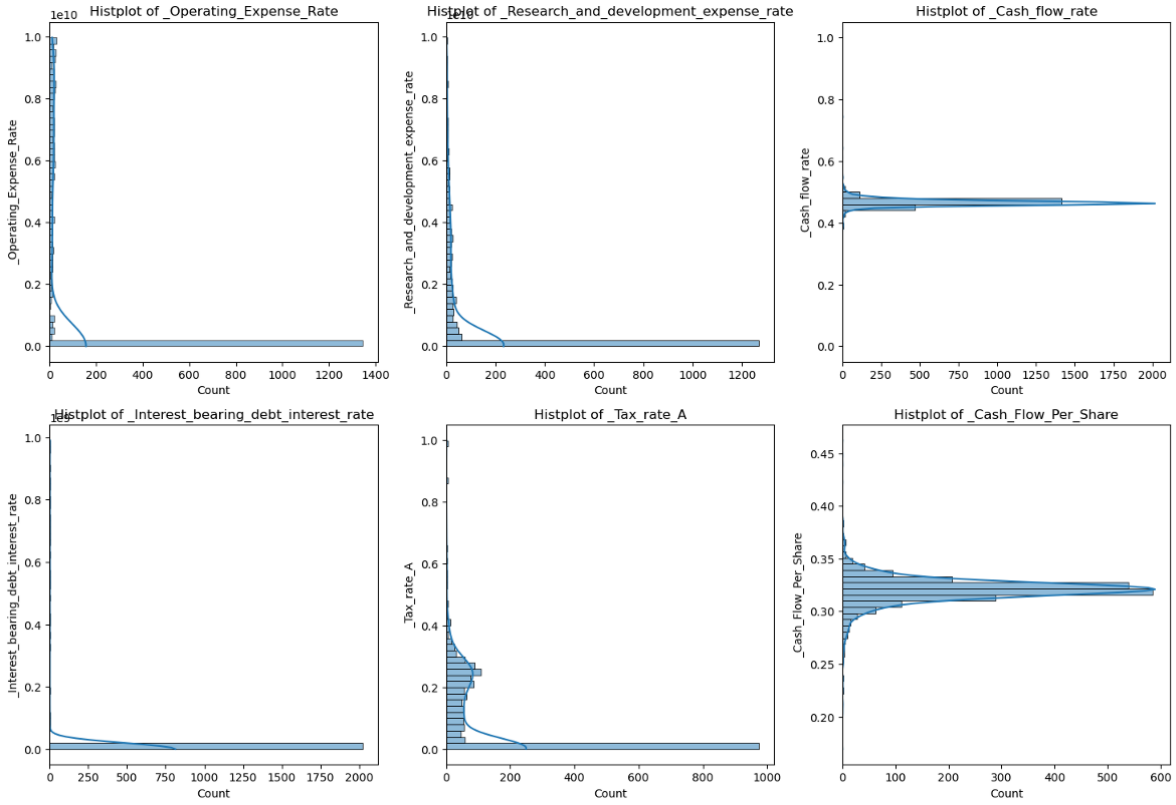
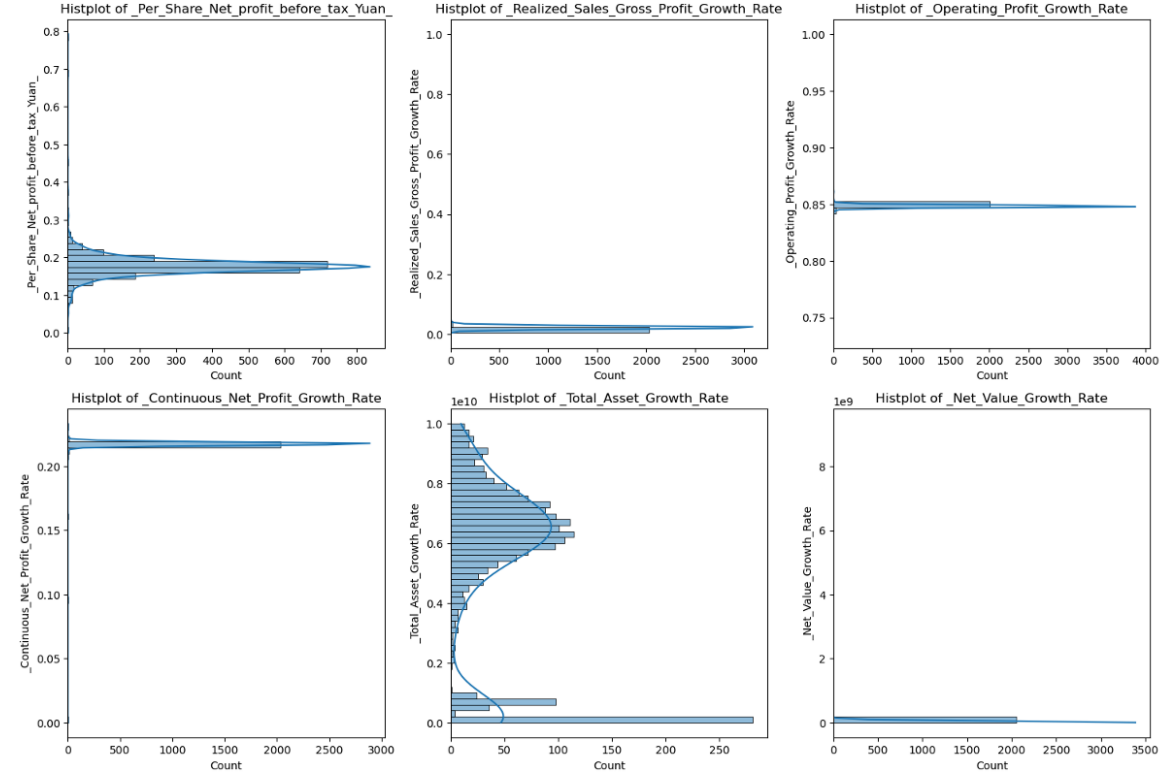


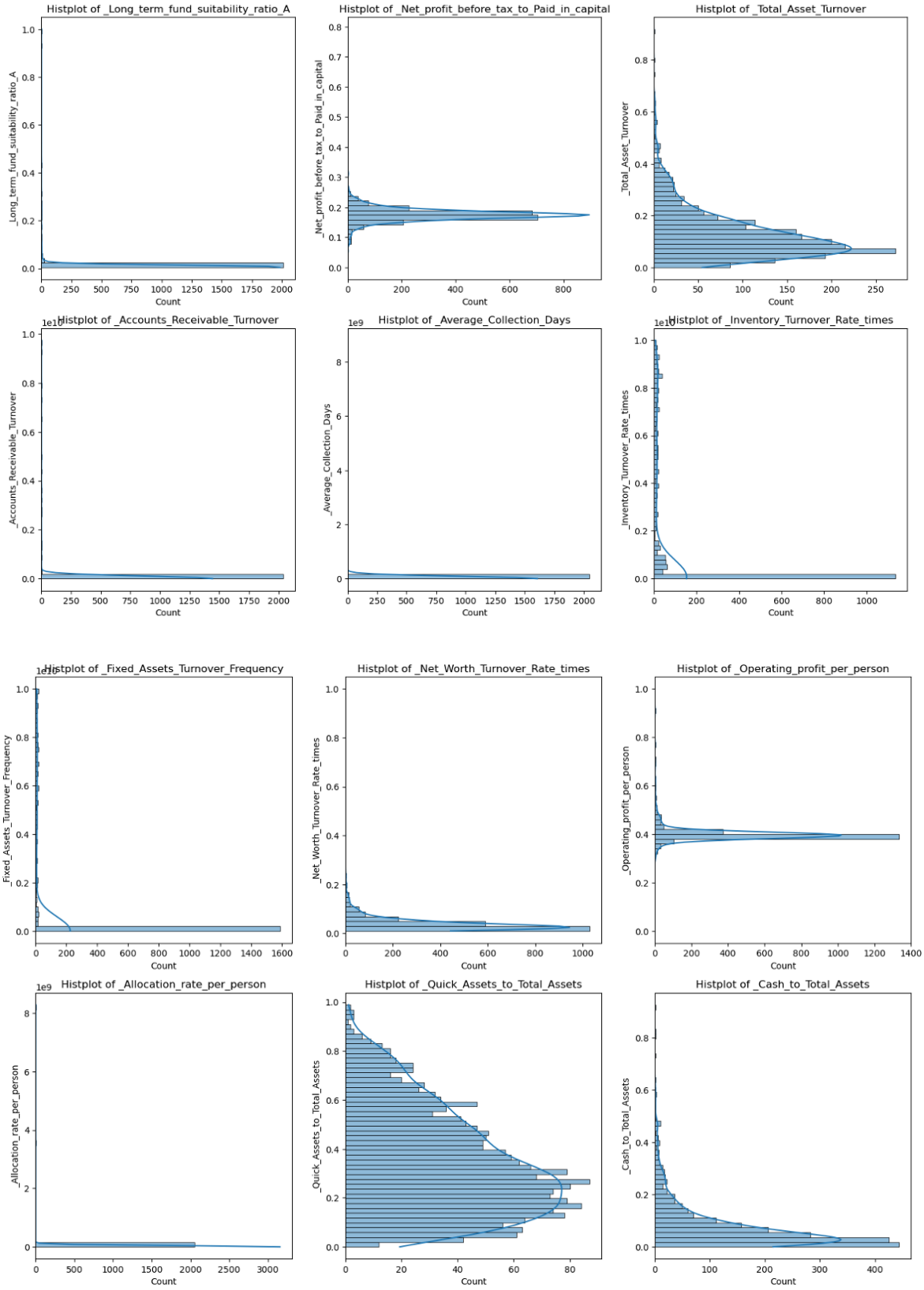
Figure 5: Box plots of all numeric variables

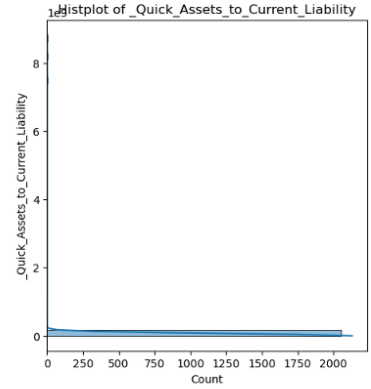
Histograms of all numeric variables:

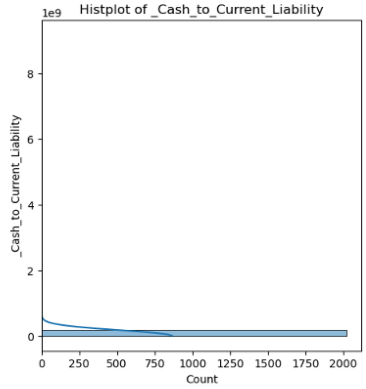


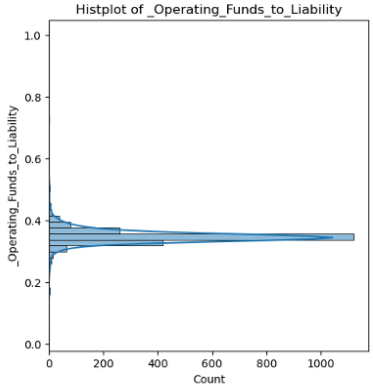


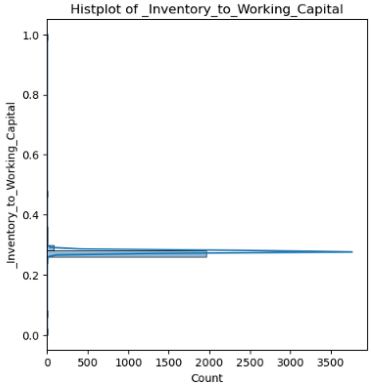
E

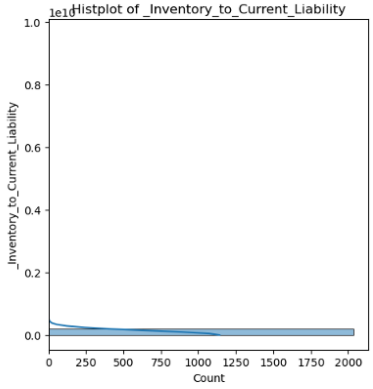

Histplot of _Long_term_fund_suitability_ratio_A
_Long_term_fund_suitability_ratio_A
Count
Histplot of _Net_profit_before_tax_to_Paid_in_capital
_Net_profit_before_tax_to_Paid_in_capital
Count
Histplot of _Total_Asset_Turnover
_Total_Asset_Turnover
Count
Histplot of _Accounts_Receivable_Turnover
_Accounts_Receivable_Turnover
Count
Histplot of _Average_Collection_Days
_Average_Collection_Days
Count
Histplot of _Inventory_Turnover_Rate_times
_Inventory_Turnover_Rate_times
Count
Histplot of _Fixed_Assets_Turnover_Frequency
_Fixed_Assets_Turnover_Frequency
Count
Histplot of _Net_Worth_Turnover_Rate_times
_Net_Worth_Turnover_Rate_times
Count
Histplot of _Operating_profit_per_person
_Operating_profit_per_person
Count
Histplot of _Allocation_rate_per_person
_Allocation_rate_per_person
Count
Histplot of _Quick_Assets_to_Total_Assets
_Quick_Assets_to_Total_Assets
Count
Histplot of _Cash_to_Total_Assets
_Cash_to_Total_Assets
Count

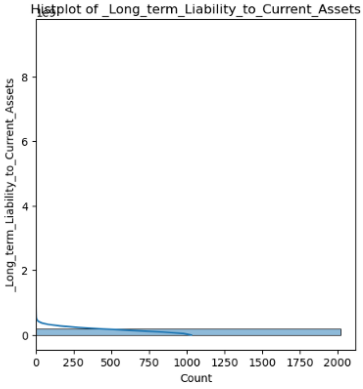

Histplot of _Quick_Assets_to_Current_Liability
1e9
_Quick_Assets_to_Current_Liability
Count

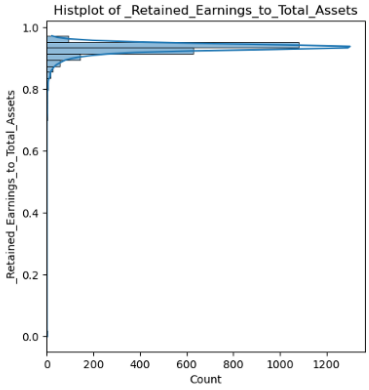

Histplot of _Cash_to_Current_Liability
1e9
_Cash_to_Current_Liability
Count

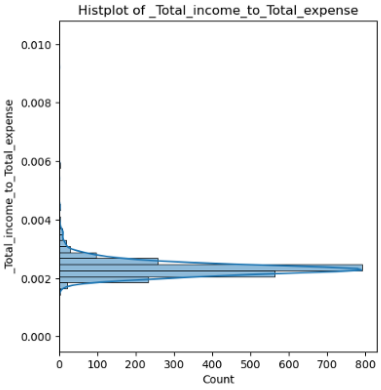

Histplot of _Operating_Funds_to_Liability
_Operating_Funds_to_Liability
Count

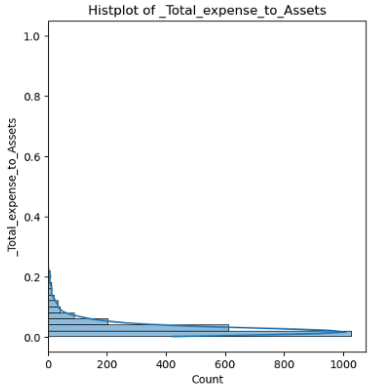

Histplot of _Inventory_to_Working_Capital
_Inventory_to_Working_Capital
Count

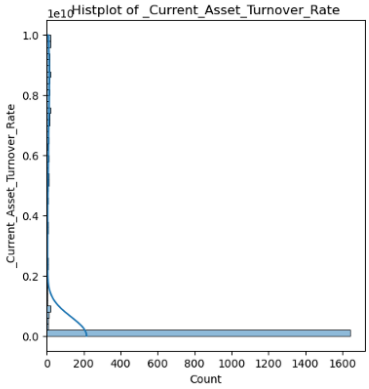

Histplot of _Inventory_to_Current_Liability
1e10
_Inventory_to_Current_Liability
Count

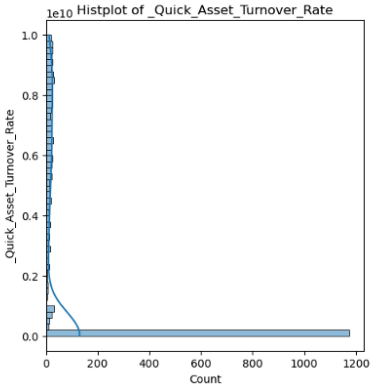

Histplot of _Long_term_Liability_to_Current_Assets
1e9
_Long_term_Liability_to_Current_Assets
Count

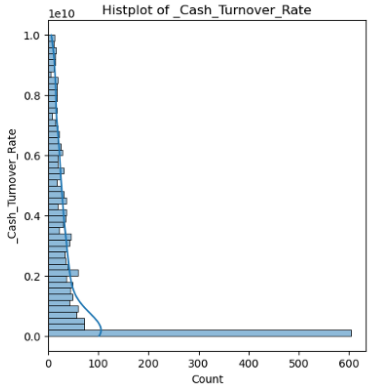

Histplot of _Retained_Earnings_to_Total_Assets
_Retained_Earnings_to_Total_Assets
Count

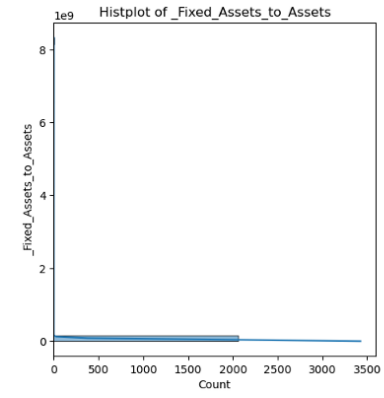

Histplot of _Total_income_to_Total_expense
_Total_income_to_Total_expense
Count

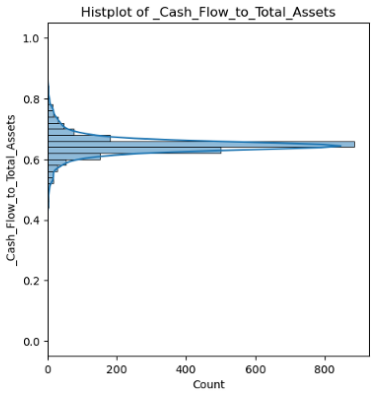

Histplot of _Total_expense_to_Assets
_Total_expense_to_Assets
Count

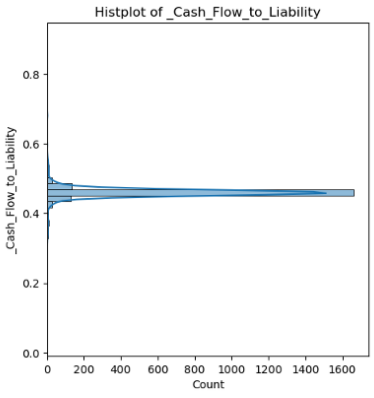

Histplot of _Current_Asset_Turnover_Rate
1e10
_Current_Asset_Turnover_Rate
Count

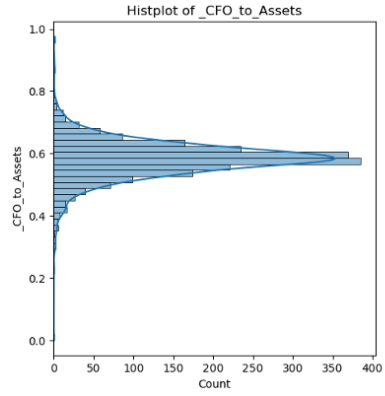

Histplot of _Quick_Asset_Turnover_Rate
1e10
_Quick_Asset_Turnover_Rate
Count

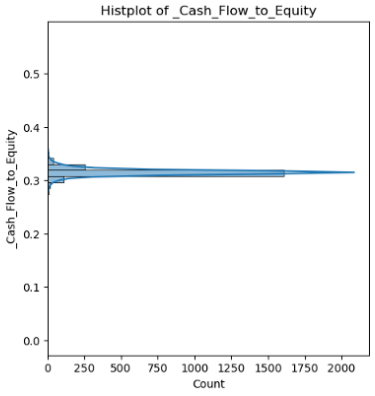

Histplot of _Cash_Turnover_Rate
1e10
_Cash_Turnover_Rate
Count

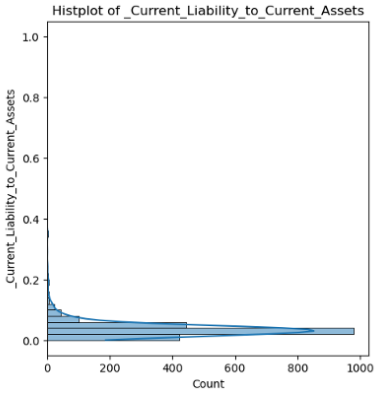

Histplot of _Fixed_Assets_to_Assets
1e9
_Fixed_Assets_to_Assets
Count

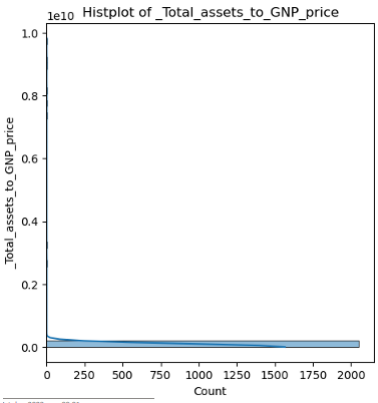

Histplot of _Cash_Flow_to_Total_Assets
_Cash_Flow_to_Total_Assets
Count

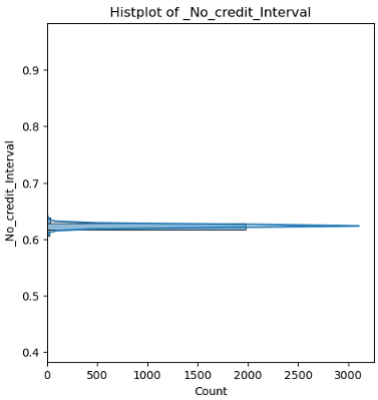

Histplot of _Cash_Flow_to_Liability
_Cash_Flow_to_Liability
Count

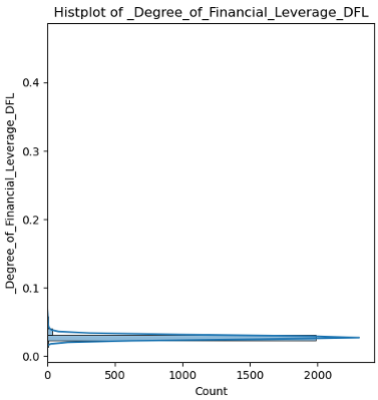

Histplot of _CFO_to_Assets
_CFO_to_Assets
Count

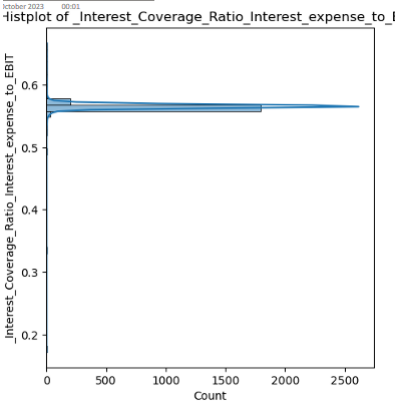

Histplot of _Cash_Flow_to_Equity
_Cash_Flow_to_Equity
Count


Histplot of _Current_Liability_to_Current_Assets
_Current_Liability_to_Current_Assets
Count


1e10
Histplot of _Total_assets_to_GNP_price
_Total_assets_to_GNP_price
Count


Histplot of _No_credit_Interval
_No_credit_Interval
Count


Histplot of _Degree_of_Financial_Leverage_DFL
_Degree_of_Financial_Leverage_DFL
Count


Histplot of _Interest_Coverage_Ratio_Interest_expense_to_EBIT
_Interest_Coverage_Ratio_Interest_expense_to_EBIT
Count

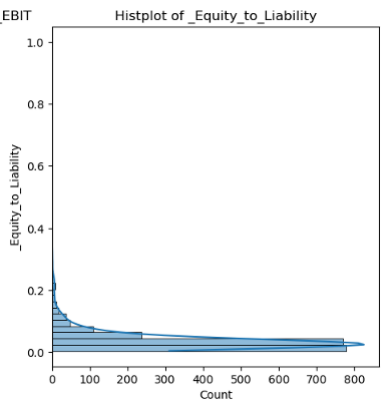

Histplot of _Equity_to_Liability
_Equity_to_Liability
Count

Figure 6: Histograms of all numeric variables

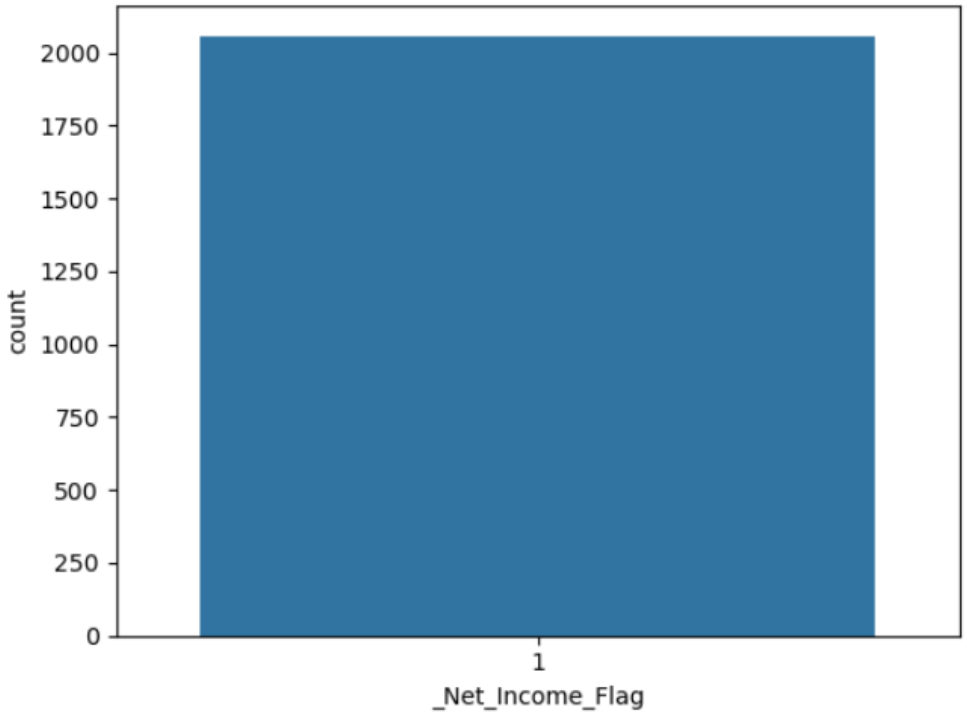


Figure 7: Distribution of Net_Income_Flag

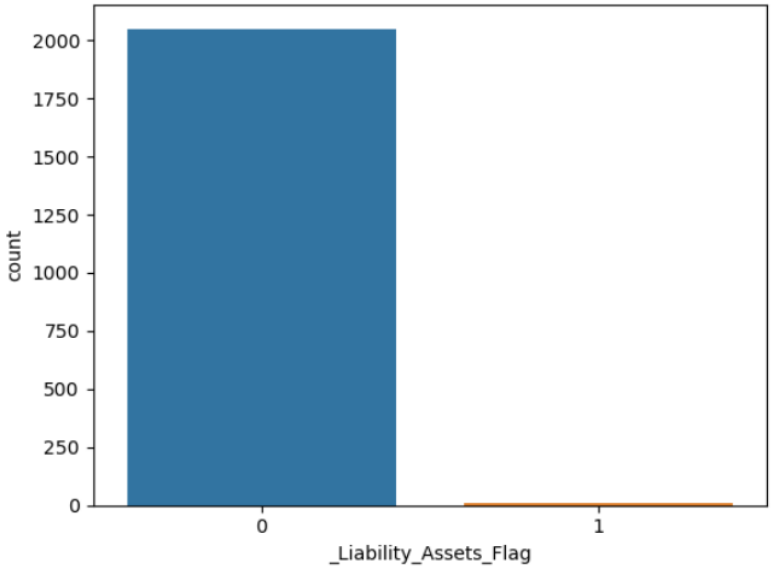


Figure 8: Distribution of Liability_Assets_Flag

Insights from univariate analysis:

- As observed in outlier treatment, there are huge outliers for many predictor variables in the dataset.

- Some companies spend considerable amount on R&D but they all have been recorded as outliers. 75% of the companies have R&D expense rate below 0.2

- At least 75% of companies do not have debt interest rates.

- Some companies show negative signs of few important revenue indicators like Operating_Proft_Growth_Rate, Net_Proft_Growth_Rate, Net_Value_Growth_Rate, Total_debt_to_total_net_worth, LongtermLiability_to_current_assets

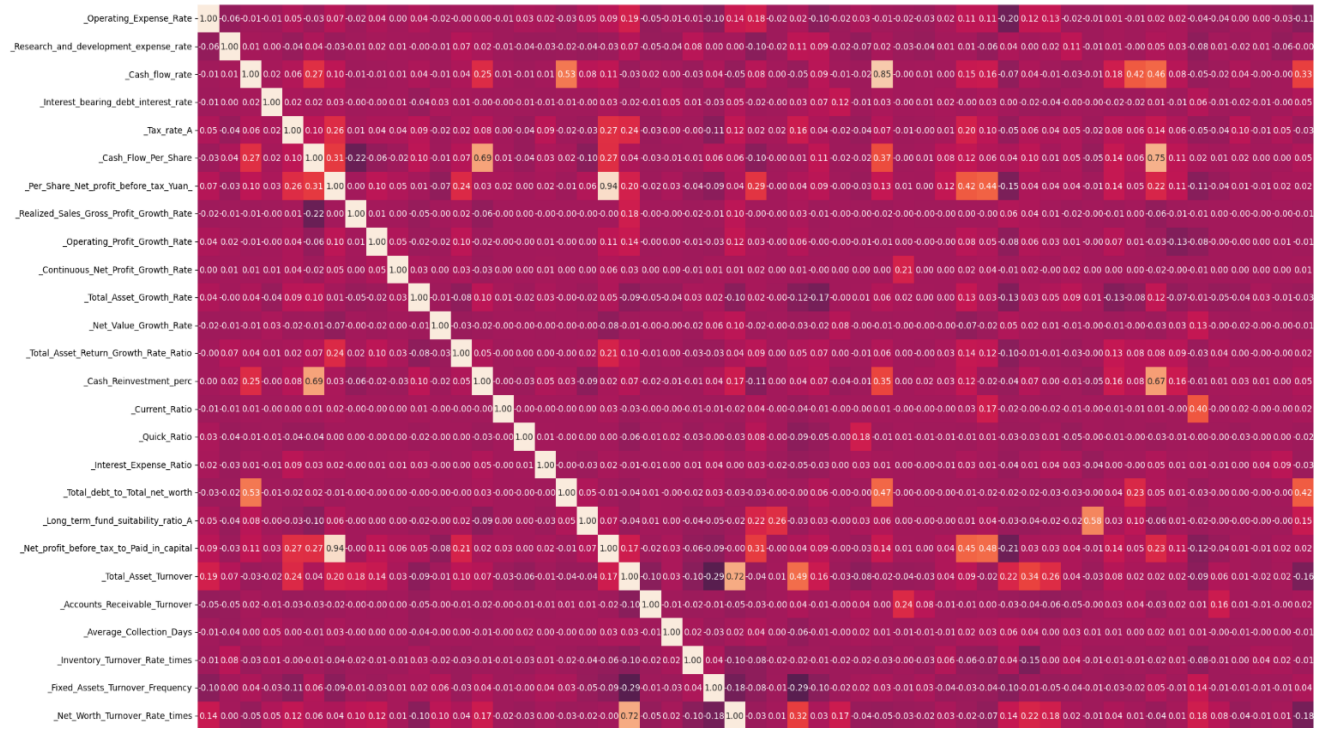
- Income_to_expense and Equity_to_liability ratios for most of the companies is either 0 or on the positive side.

- Variables like Cash_Flow_Rate, Cash_Flow_Per_Share, Per_Share_Net_Profit_before_tax_Yuan, Cash_investment_perc, Interest_Expense_Ratio, Net_Profit_tax_to_Paid_in_Capital, Operating_profit_per_person, Total_income_to_total_expense, Cash_flow_to_total_assets, CFO_to_assets.

- All companies have Net_Income_Flag to 1, so it can be removed for model building.

Bivariate Analysis:

Correlation plot:



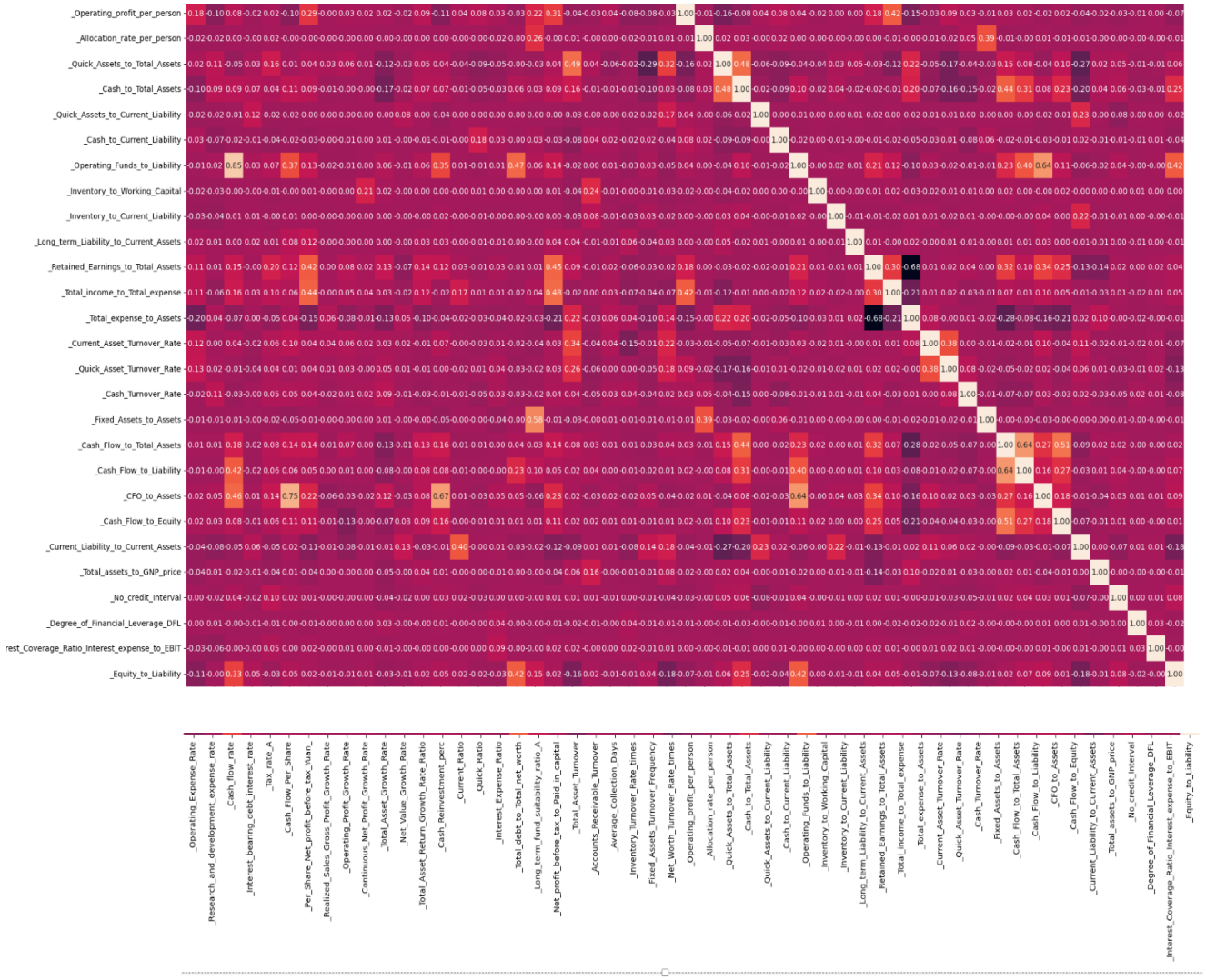
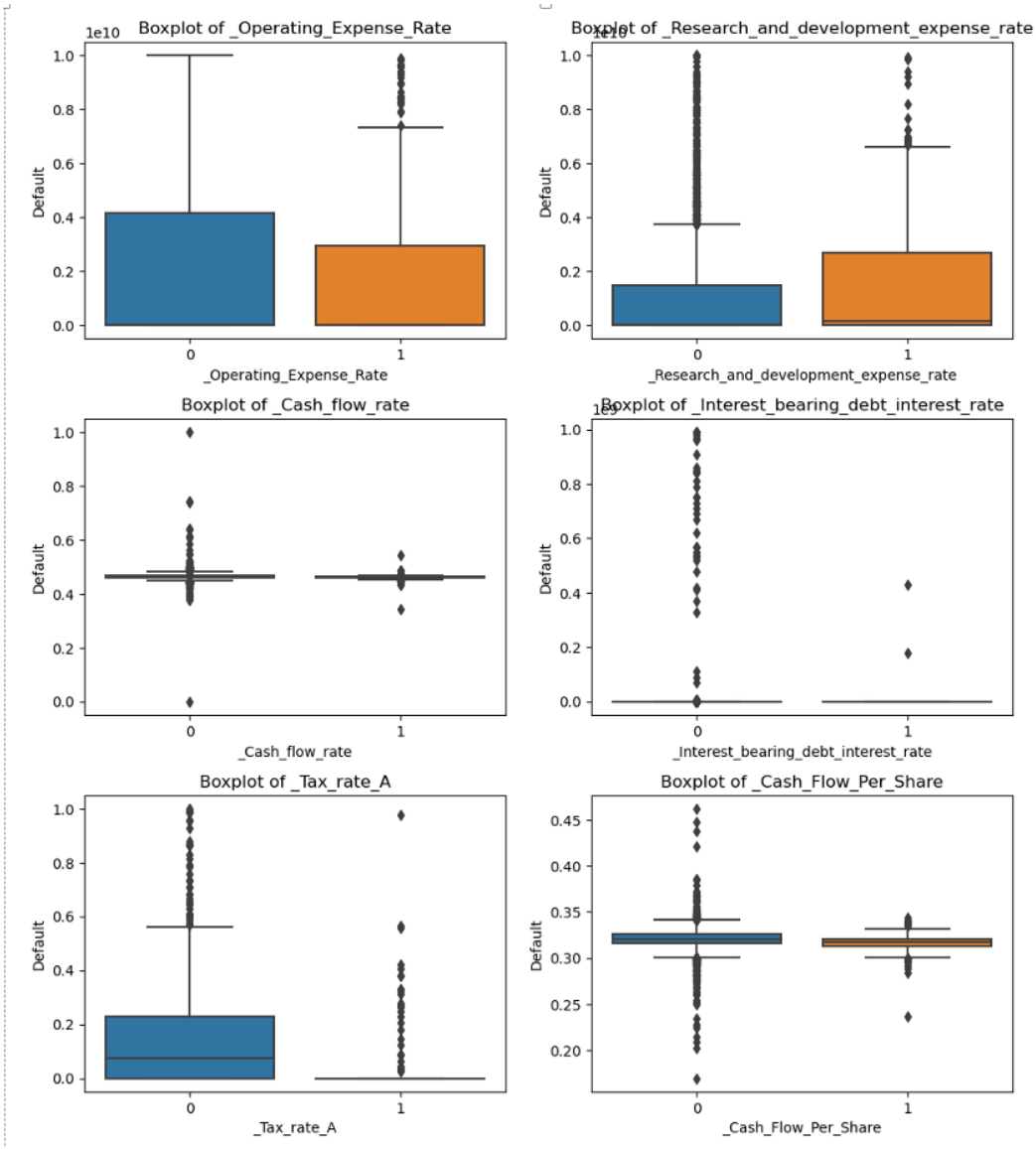
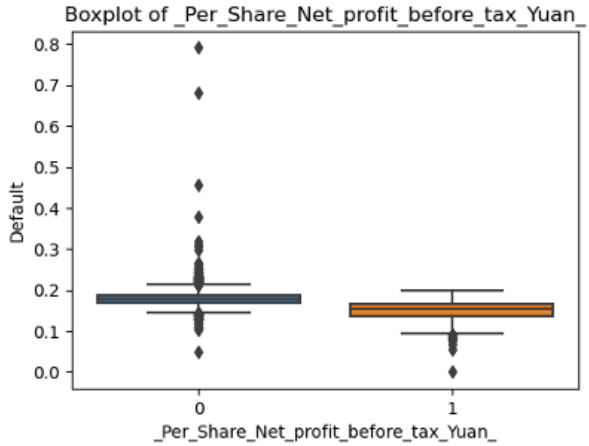
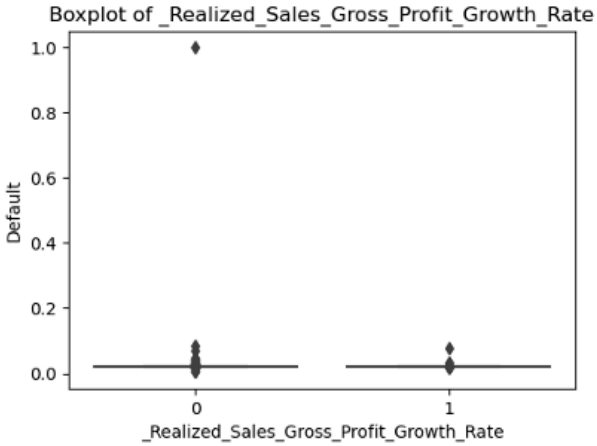


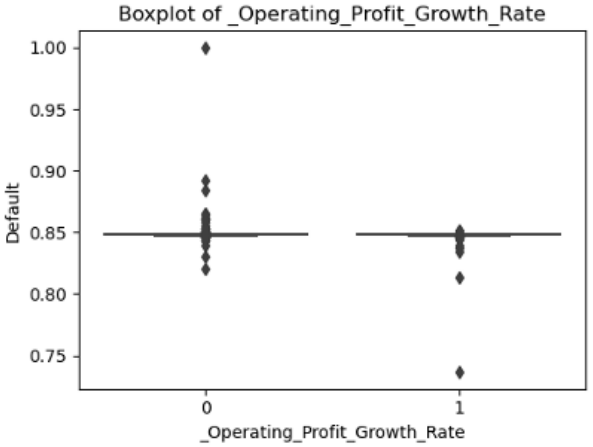
Figure 9: Correlation plot - Bivariate analysis

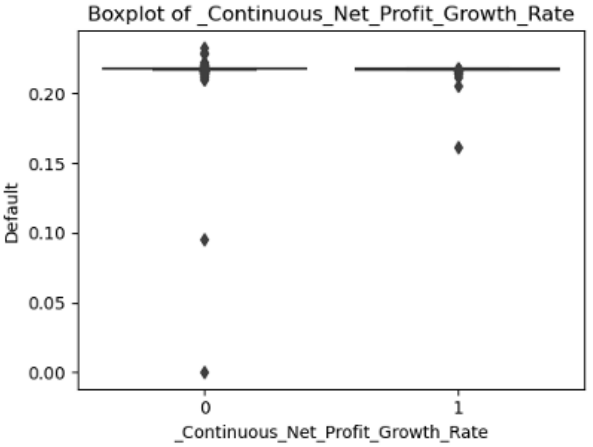
Box plots of all numeric variables for default vs non default

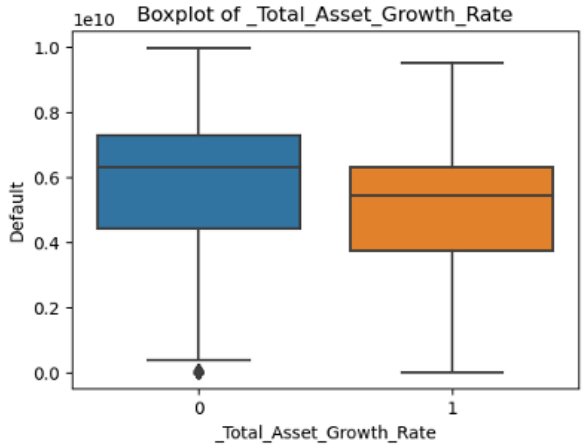

Boxplot of _Operating_Expense_Rate
1e10
Default
_Operating_Expense_Rate
Boxplot of _Research_and_development_expense_rate
Default
_Research_and_development_expense_rate
Boxplot of _Cash_flow_rate
Default
_Cash_flow_rate
Boxplot of _Interest_bearing_debt_interest_rate
Default
_Interest_bearing_debt_interest_rate
Boxplot of _Tax_rate_A
Default
_Tax_rate_A
Boxplot of _Cash_Flow_Per_Share
Default
_Cash_Flow_Per_Share

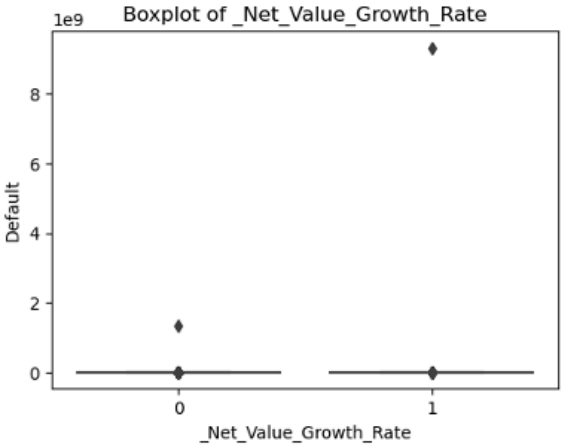

Boxplot of _Per_Share_Net_profit_before_tax_Yuan_
Default
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
_Per_Share_Net_profit_before_tax_Yuan_

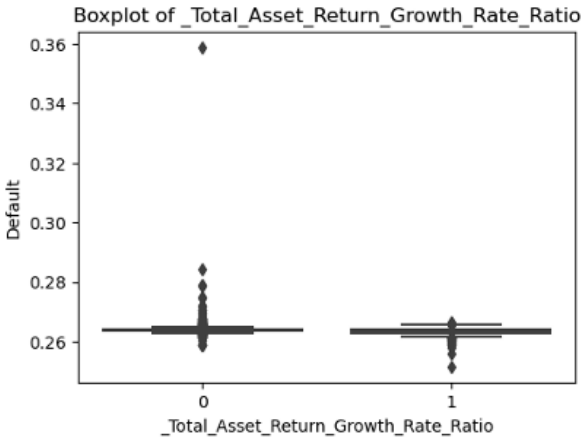

Boxplot of _Realized_Sales_Gross_Profit_Growth_Rate
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Realized_Sales_Gross_Profit_Growth_Rate

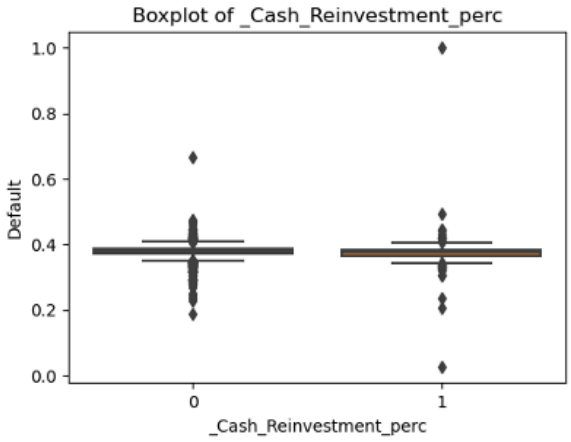

Boxplot of _Operating_Profit_Growth_Rate
Default
1.00
0.95
0.90
0.85
0.80
0.75
0
1
_Operating_Profit_Growth_Rate

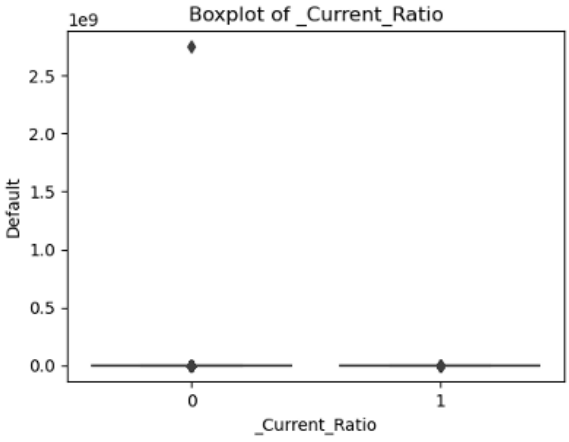

Boxplot of _Continuous_Net_Profit_Growth_Rate
Default
0.20
0.15
0.10
0.05
0.00
0
1
_Continuous_Net_Profit_Growth_Rate

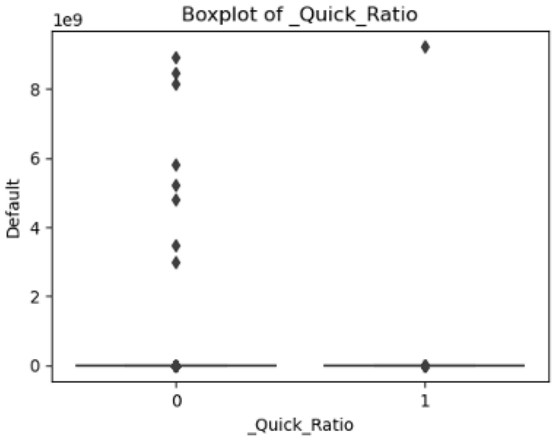

1e10
Boxplot of _Total_Asset_Growth_Rate
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Total_Asset_Growth_Rate

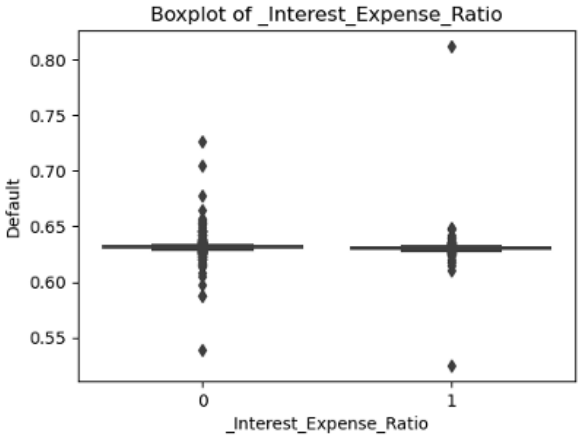

1e9
Boxplot of _Net_Value_Growth_Rate
Default
8
6
4
2
0
0
1
_Net_Value_Growth_Rate

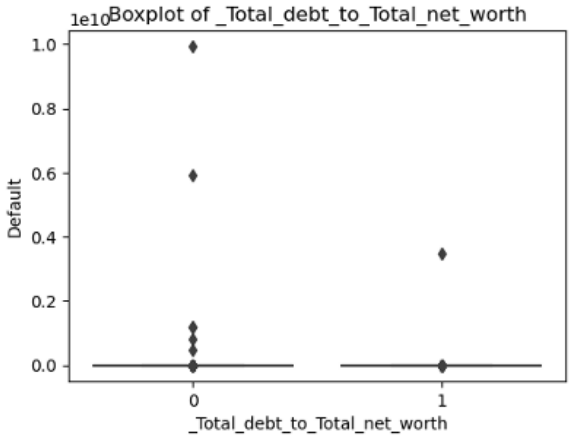

Boxplot of _Total_Asset_Return_Growth_Rate_Ratio
Default
0.36
0.34
0.32
0.30
0.28
0.26
0
1
_Total_Asset_Return_Growth_Rate_Ratio

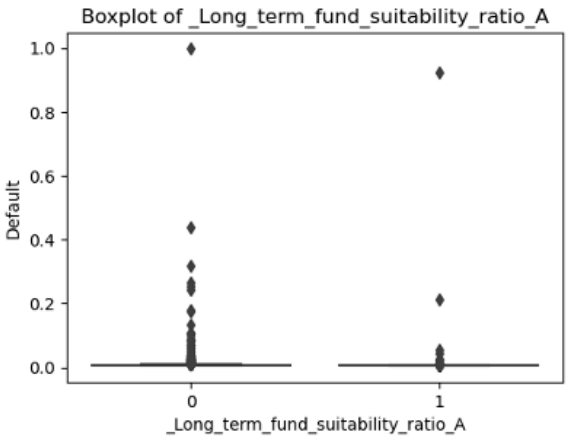

Boxplot of _Cash_Reinvestment_perc
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Cash_Reinvestment_perc

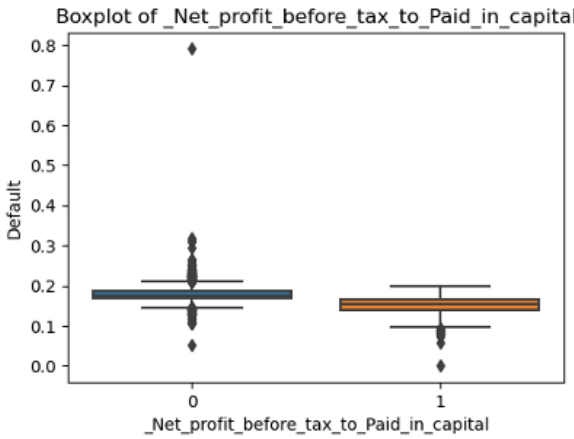

Boxplot of _Current_Ratio
1e9
Default
2.5
2.0
1.5
1.0
0.5
0.0
0
1
_Current_Ratio

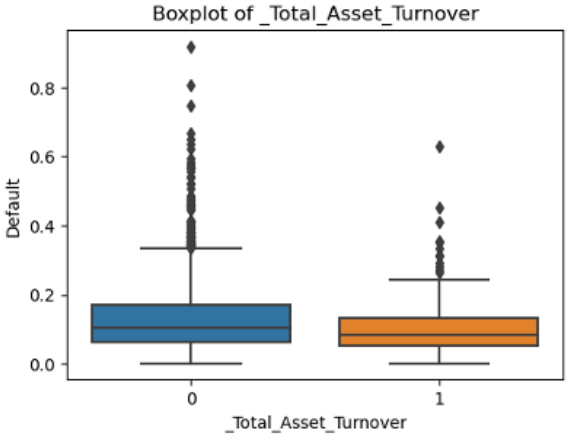

Boxplot of _Quick_Ratio
1e9
Default
8
6
4
2
0
0
1
_Quick_Ratio

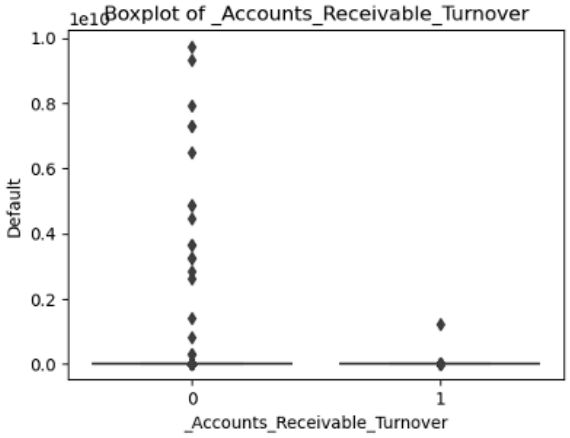

Boxplot of _Interest_Expense_Ratio
Default
0.80
0.75
0.70
0.65
0.60
0.55
0
1
_Interest_Expense_Ratio

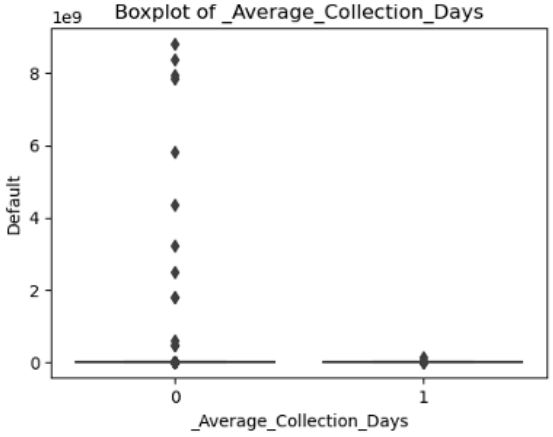

Boxplot of _Total_debt_to_Total_net_worth
1e10
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Total_debt_to_Total_net_worth

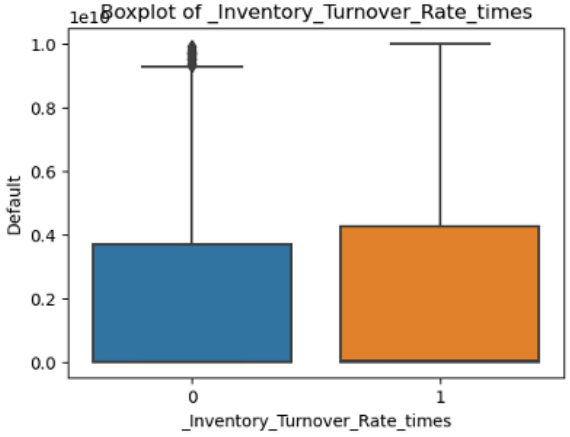

Boxplot of _Long_term_fund_suitability_ratio_A
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Long_term_fund_suitability_ratio_A

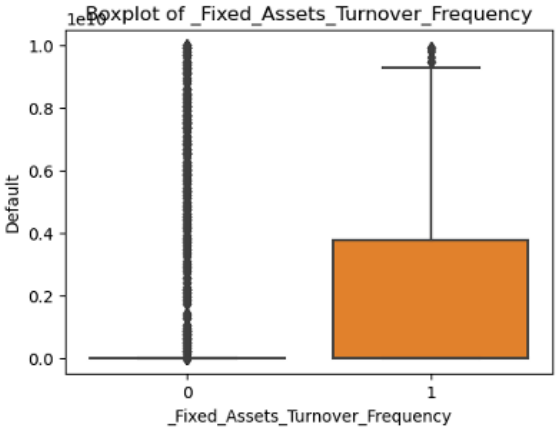

Boxplot of _Net_profit_before_tax_to_Paid_in_capital
Default
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
_Net_profit_before_tax_to_Paid_in_capital

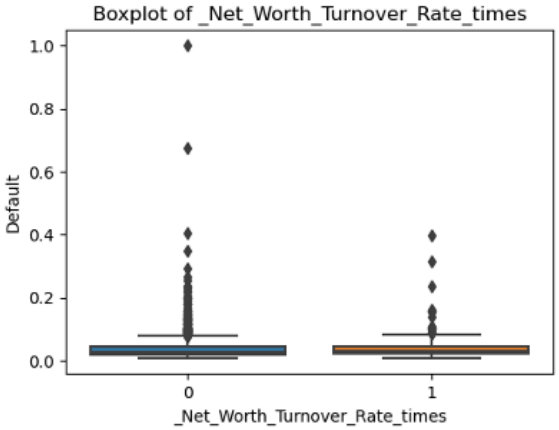

Boxplot of _Total_Asset_Turnover
Default
0.8
0.6
0.4
0.2
0.0
0
1
_Total_Asset_Turnover

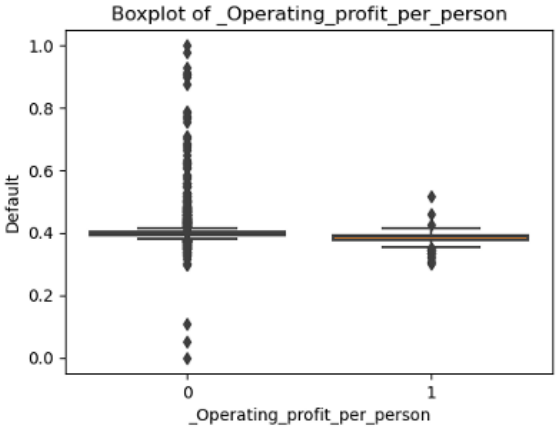

Boxplot of _Accounts_Receivable_Turnover
1e10
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Accounts_Receivable_Turnover

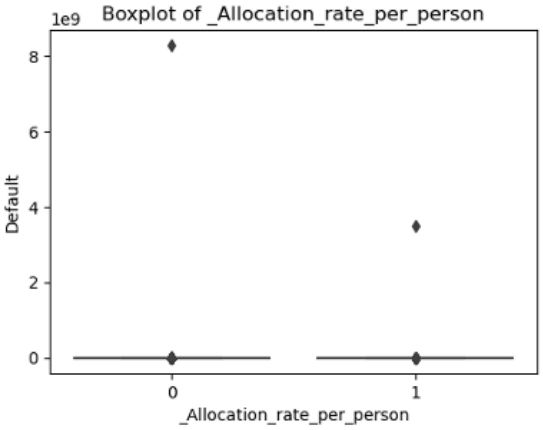

Boxplot of _Average_Collection_Days
1e9
Default
8
6
4
2
0
0
1
_Average_Collection_Days

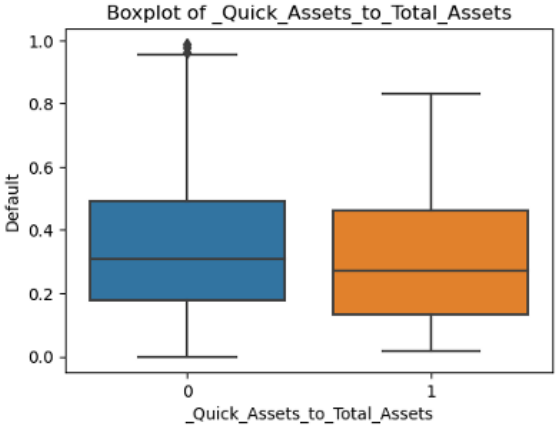

Boxplot of _Inventory_Turnover_Rate_times
1e10
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Inventory_Turnover_Rate_times

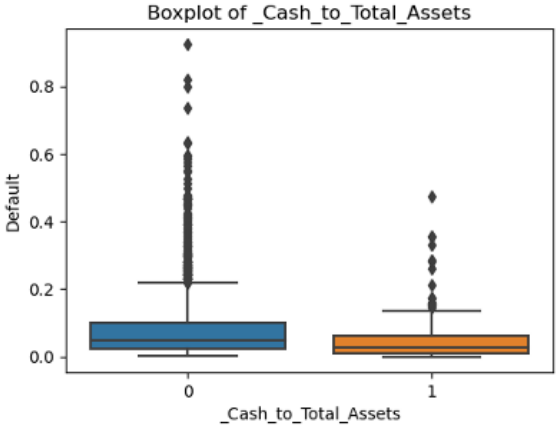

Boxplot of _Fixed_Assets_Turnover_Frequency
1e10
Default
_Fixed_Assets_Turnover_Frequency

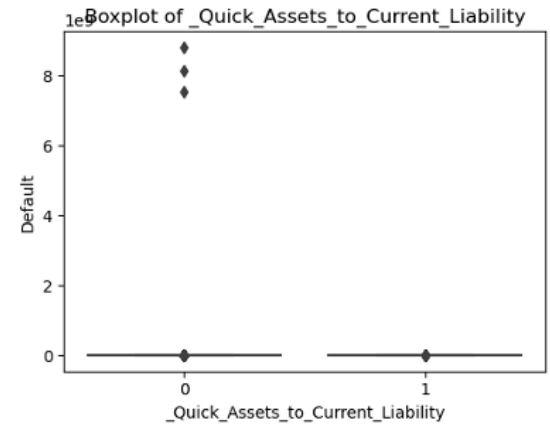

Boxplot of _Net_Worth_Turnover_Rate_times
Default
_Net_Worth_Turnover_Rate_times

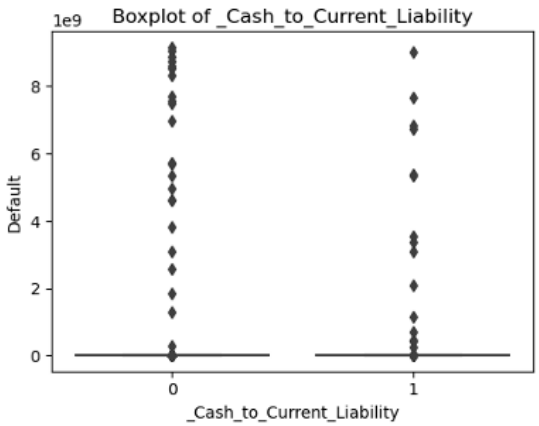

Boxplot of _Operating_profit_per_person
Default
_Operating_profit_per_person

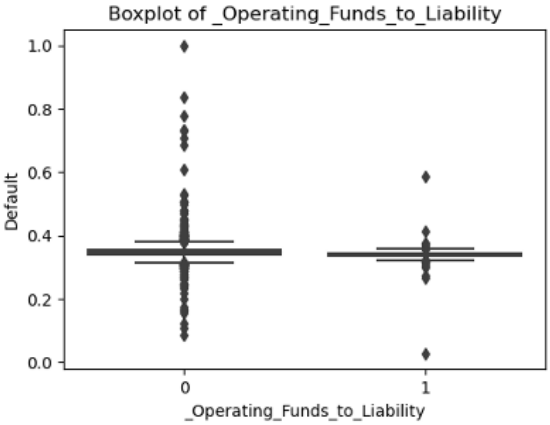

Boxplot of _Allocation_rate_per_person
1e9
Default
_Allocation_rate_per_person

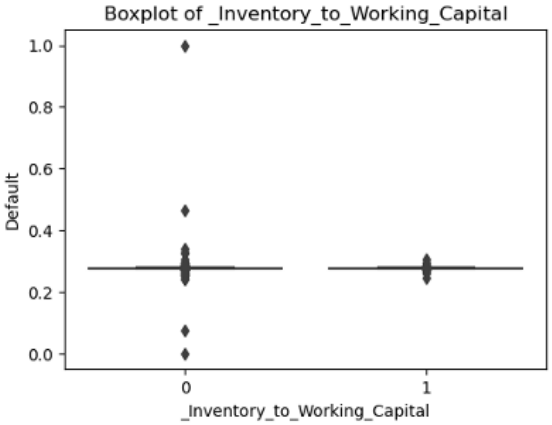

Boxplot of _Quick_Assets_to_Total_Assets
Default
_Quick_Assets_to_Total_Assets

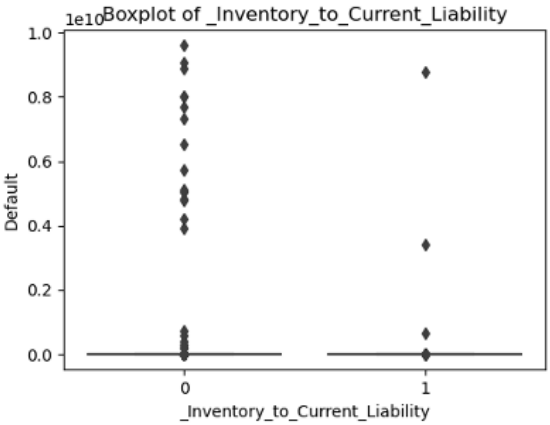

Boxplot of _Cash_to_Total_Assets
Default
_Cash_to_Total_Assets

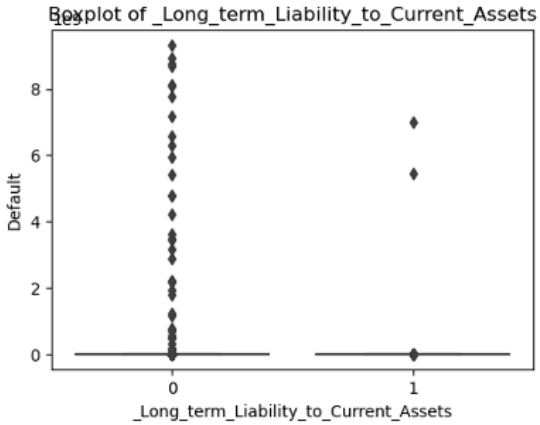

Boxplot of _Quick_Assets_to_Current_Liability
1e9
Default
_Quick_Assets_to_Current_Liability

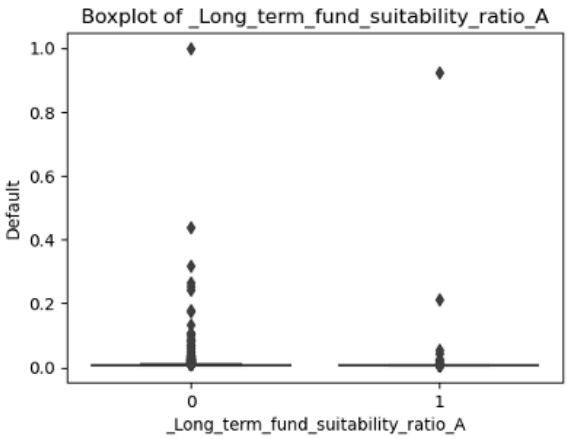

Boxplot of _Cash_to_Current_Liability
1e9
Default
_Cash_to_Current_Liability

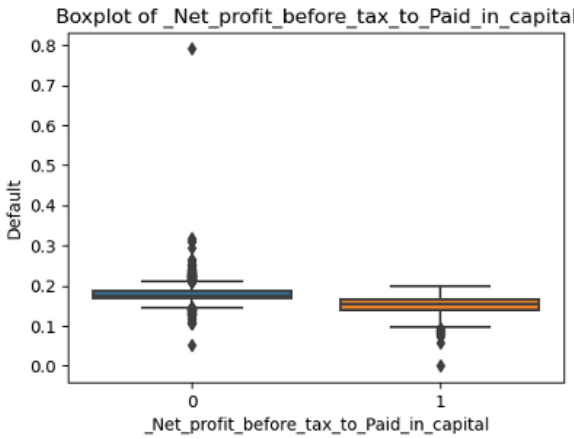

Boxplot of _Operating_Funds_to_Liability
Default
_Operating_Funds_to_Liability

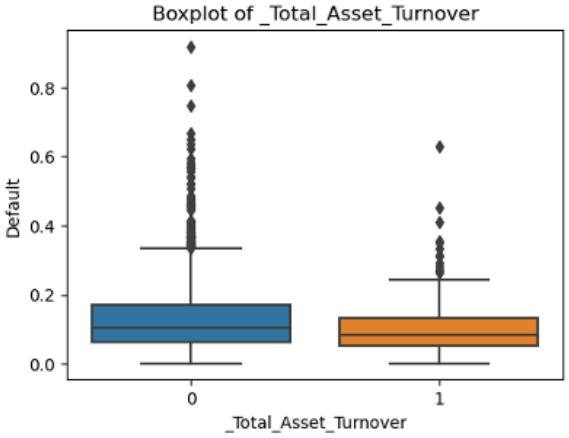

Boxplot of _Inventory_to_Working_Capital
Default
_Inventory_to_Working_Capital

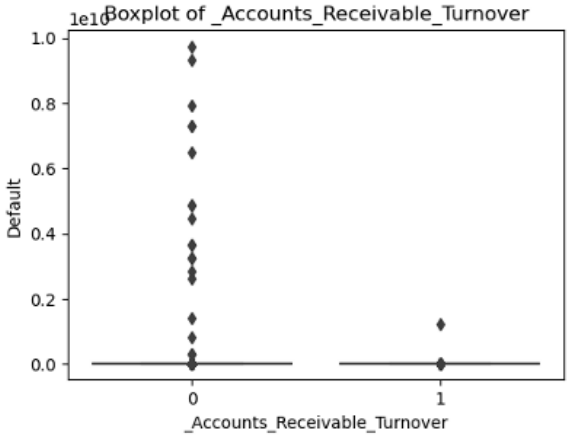

Boxplot of _Inventory_to_Current_Liability
1e10
Default
_Inventory_to_Current_Liability

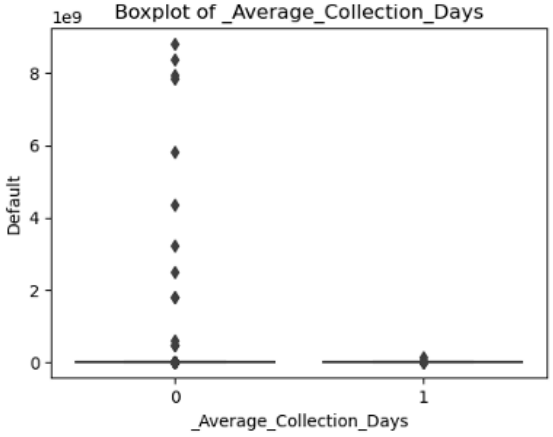

Boxplot of _Long_term_Liability_to_Current_Assets
Default
_Long_term_Liability_to_Current_Assets

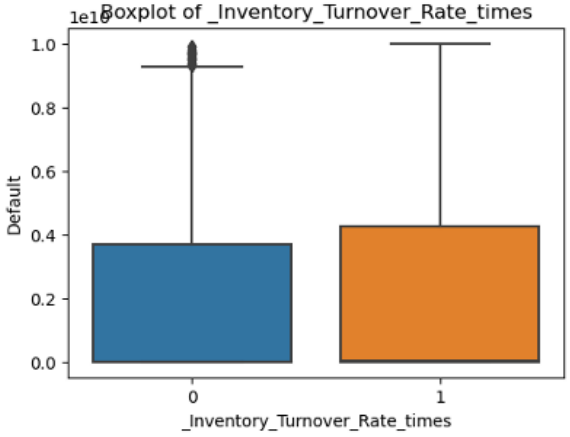

Boxplot of _Long_term_fund_suitability_ratio_A
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Long_term_fund_suitability_ratio_A


Boxplot of _Net_profit_before_tax_to_Paid_in_capital
Default
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
_Net_profit_before_tax_to_Paid_in_capital


Boxplot of _Total_Asset_Turnover
Default
0.8
0.6
0.4
0.2
0.0
0
1
_Total_Asset_Turnover


Boxplot of _Accounts_Receivable_Turnover
1e10
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Accounts_Receivable_Turnover


Boxplot of _Average_Collection_Days
1e9
Default
8
6
4
2
0
0
1
_Average_Collection_Days


Boxplot of _Inventory_Turnover_Rate_times
1e10
Default
1.0
0.8
0.6
0.4
0.2
0.0
0
1
_Inventory_Turnover_Rate_times

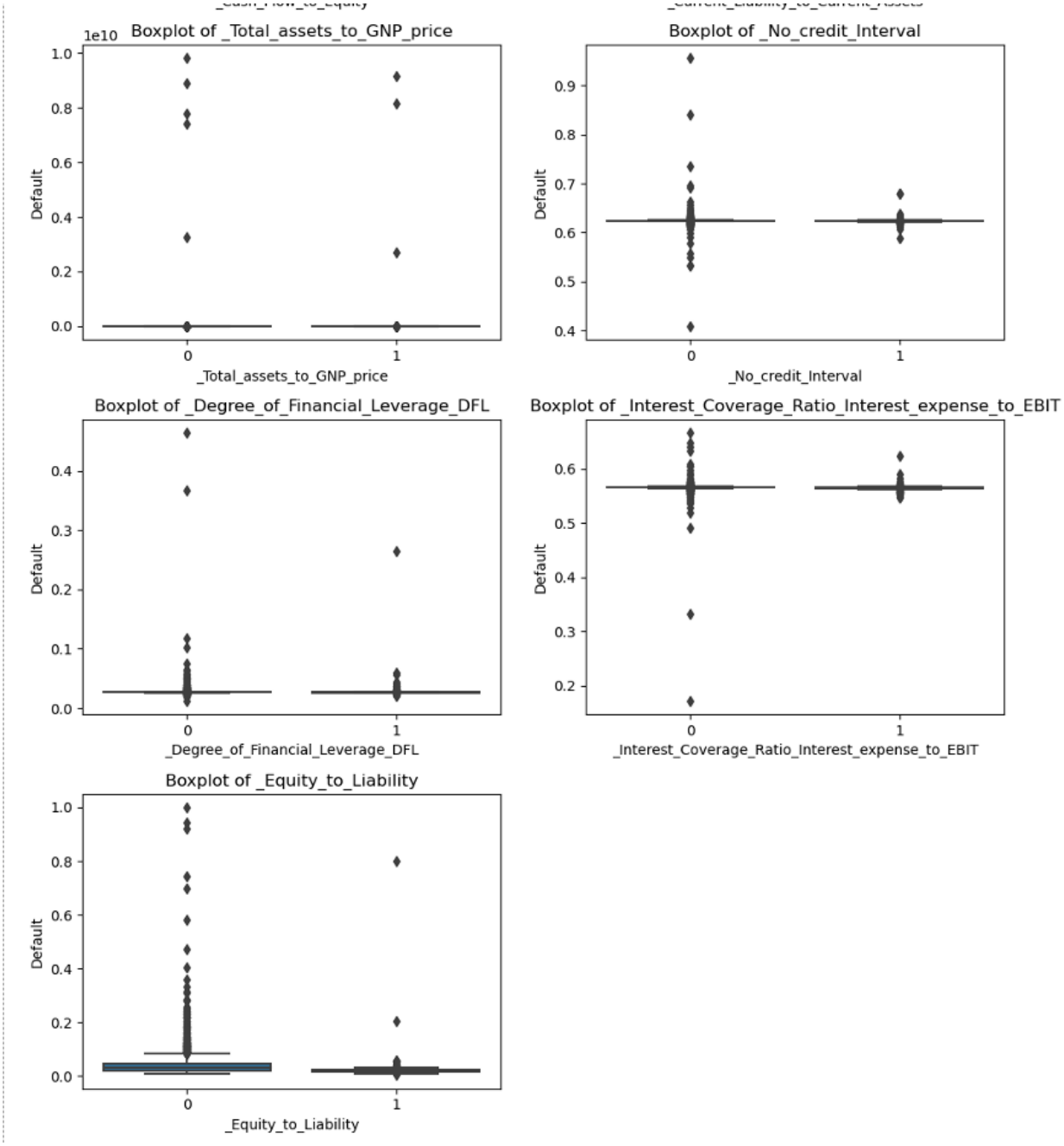


Figure 10: Box plots of all numeric variables for defaulters vs non defaulters

- **Insights from bivariate analysis:**
  - Below are the variables that exhibit high correlations with each other
    - Cash_flow_rate -- Cash_to_Current_liability

    - Per_share_net_proft_before_tax_yuan --net_profit_before_tax_to_paid_in_capital
    - cash_flow_per_share -- _cash_flow_to_liability
    - cash_reinvestment_perc -- _cash_flow_to_liability
    - net_worth_turnover_rate_times -- net_profit_before_tax_to_paid_in_capital
    - operating_funds_to_liability -- cash_flow_rate
    - total_expense_to_assets -- total_income_to_total_expense

- Below are the insights from box plots hued by the 'Default' response variable. Companies who have defaulted in the past have below statistics

    - They have a high expense rate, less cash flow rate, less cash flow per share, less per share net profit before tax, less operating profit growth rate, net profit before tax to paid in capital, high fixed assets turnover frequency, less operating profit per person.

- Removed Net_Income_Flag as it is same for all the entries
- Applied Recursive feature elimination using logistic regression to find out the most important variables as the variables are observed to be correlated with each other.
    - Parameters:
        - Estimator: base logistic regression algorithm
        - N_features: 30 (to find top 30 features)
        - Step = 1

- Top 30 important features observed after applying recursive feature elimination

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 31 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Researchanddevelopmentexpenserate               2058 non-null   float64
 1   Cashflowrate                                    2058 non-null   float64
 2   TaxrateA                                        2058 non-null   float64
 3   CashFlowPerShare                                2058 non-null   float64
 4   PerShareNetprofitbeforetaxYuan                  2058 non-null   float64
 5   TotalAssetGrowthRate                            2058 non-null   float64
 6   TotalAssetReturnGrowthRateRatio                 2058 non-null   float64
 7   CashReinvestmentperc                            2058 non-null   float64
 8   InterestExpenseRatio                            2058 non-null   float64
 9   LongtermfundsuitabilityratioA                   2058 non-null   float64
 10  NetprofitbeforetaxtoPaidincapital               2058 non-null   float64
 11  TotalAssetTurnover                              2058 non-null   float64
 12  NetWorthTurnoverRatetimes                       2058 non-null   float64
 13  Operatingprofitperperson                        2058 non-null   float64
 14  QuickAssetstoTotalAssets                        2058 non-null   float64
 15  CashtoTotalAssets                               2058 non-null   float64
 16  OperatingFundstoLiability                       2058 non-null   float64
 17  RetainedEarningstoTotalAssets                   2058 non-null   float64
 18  TotalincometoTotalexpense                       2058 non-null   float64
 19  TotalexpensetoAssets                            2058 non-null   float64
 20  QuickAssetTurnoverRate                          2058 non-null   float64
 21  CashTurnoverRate                                2058 non-null   float64
 22  CashFlowtoTotalAssets                           2058 non-null   float64
 23  CashFlowtoLiability                             2058 non-null   float64
 24  CFOtoAssets                                     2058 non-null   float64
 25  CashFlowtoEquity                                2058 non-null   float64
 26  CurrentLiabilitytoCurrentAssets                 2058 non-null   float64
 27  InterestCoverageRatioInterestexpensetoEBIT      2058 non-null   float64
 28  EquitytoLiability                               2058 non-null   float64
 29  LiabilityAssetsFlag                             2058 non-null   int64
 30  Default                                         2058 non-null   int64
dtypes: float64(29), int64(2)
memory usage: 498.5 KB
```

Table 6: Information of top 30 features

- We will proceed with modeling using the above variables.

1.3. Train test Split

The data has been split into training and test data set

- Train test split has been done with ratio 67 : 33
- Random state - 42
- Viewing the first rows of predictor and response variables of training set

	Researchanddevelopmentexpenserate	Cashflowrate	TaxrateA	CashFlowPerShare	PerShareNetprofitbeforetaxYuan	TotalAssetGrowthRate	TotalAssetReturn(
2011	-0.000	0.017	-0.037	0.053	-0.029	-0.187	
697	0.175	0.003	0.279	0.012	0.022	-0.061	
160	-0.000	-0.004	-0.037	-0.020	-0.005	0.046	
1273	0.288	-0.001	-0.037	0.000	-0.011	-0.075	
541	0.066	-0.003	-0.037	-0.015	-0.014	0.052	

Table 7: First rows of training data - independent variables

```
2011    0
697     0
160     0
1273    0
541     0
Name: Default, dtype: int64
```

Table 8: First rows of training data - dependent variable

- Viewing the first rows of predictor and response variables of testing data set

	Researchanddevelopmentexpenserate	Interestbearingdebtinterestrate	TotaldebttoTotalnetworth	AccountsReceivableTurnover	Operatingprofitperperson	Alloc
1298	-0.000	0.274	0.359	0.375	-0.042000	
591	0.735	0.974	-0.377	0.435	2.111375	
1318	0.000	-0.605	-0.483	-0.384	-1.113000	
1067	-0.000	0.264	-0.265	-0.243	1.998000	
29	0.614	-0.129	-0.263	-0.327	-0.438000	

Table 9: First rows of test data - independent variables

```
]: 974     0
   134     0
   1267    0
   464     0
   579     0
   Name: Default, dtype: int64
```

Table 10: First 5 rows of test data - dependent variable

1.4. Logistic Regression(using StatsModels library) - Choosing the most important variables and optimum cut-off

- **Built a logistic regression model considering the below formula:**

f_1 = 'Default ~ Researchanddevelopmentexpenserate + Cashflowrate + TaxrateA + CashFlowPerShare + PerShareNetprofitbeforetaxYuan + TotalAssetGrowthRate + TotalAssetReturnGrowthRateRatio + CashReinvestmentperc + InterestExpenseRatio + LongtermfundsuitabilityratioA + NetprofitbeforetaxtoPaidincapital + TotalAssetTurnover + NetWorthTurnoverRatetimes + Operatingprofitperperson + QuickAssetstoTotalAssets + CashtoTotalAssets + OperatingFundstoLiability + RetainedEarningstoTotalAssets + TotalincometoTotalexpense + TotalexpensetoAssets + QuickAssetTurnoverRate + CashTurnoverRate + CashFlowtoTotalAssets + CashFlowtoLiability + CFOtoAssets + CashFlowtoEquity + CurrentLiabilitytoCurrentAssets + InterestCoverageRatioInterestexpensetoEBIT + EquitytoLiability + LiabilityAssetsFlag'

Below is the summary of the obtained model:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2058
Model:	Logit	**Df Residuals:**	2027
Method:	MLE	**Df Model:**	30
Date:	Sat, 07 Oct 2023	**Pseudo R-squ.:**	0.4105
Time:	17:11:24	**Log-Likelihood:**	-412.49
converged:	True	**LL-Null:**	-699.69
Covariance Type:	nonrobust	**LLR p-value:**	5.833e-102

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.6272	0.234	-15.505	0.000	-4.086	-3.169
Researchanddevelopmentexpenserate	2.6828	0.659	4.072	0.000	1.392	3.974
Cashflowrate	-22.4333	61.323	-0.366	0.714	-142.625	97.758
TaxrateA	-0.2456	0.965	-0.255	0.799	-2.136	1.645
CashFlowPerShare	0.9410	7.435	0.127	0.899	-13.631	15.513
PerShareNetprofitbeforetaxYuan	57.4881	41.563	1.383	0.167	-23.973	138.949
TotalAssetGrowthRate	0.0511	0.375	0.136	0.892	-0.684	0.786
TotalAssetReturnGrowthRateRatio	0.6078	20.094	0.030	0.976	-38.776	39.992
CashReinvestmentperc	9.1889	23.106	0.398	0.691	-36.098	54.476
InterestExpenseRatio	-31.7902	35.486	-0.896	0.370	-101.342	37.761
LongtermfundsuitabilityratioA	215.3754	112.777	1.910	0.056	-5.663	436.414
NetprofitbeforetaxtoPaidincapital	-99.1715	44.942	-2.207	0.027	-187.256	-11.087
TotalAssetTurnover	-8.1760	3.166	-2.583	0.010	-14.381	-1.971
NetWorthTurnoverRatetimes	17.3917	12.821	1.357	0.175	-7.736	42.520
Operatingprofitperperson	19.0362	14.390	1.323	0.186	-9.169	47.241
QuickAssetstoTotalAssets	-0.0493	0.659	-0.075	0.940	-1.341	1.242
CashtoTotalAssets	-2.1034	2.213	-0.950	0.342	-6.442	2.235
OperatingFundstoLiability	54.2515	39.307	1.380	0.168	-22.788	131.291

RetainedEarningstoTotalAssets	-46.4079	15.514	-2.991	0.003	-76.815	-16.001
TotalincometoTotalexpense	-16.8597	12.184	-1.384	0.166	-40.740	7.020
TotalexpensetoAssets	3.3557	7.120	0.471	0.637	-10.599	17.311
QuickAssetTurnoverRate	-0.0068	0.303	-0.023	0.982	-0.602	0.588
CashTurnoverRate	-1.2189	0.386	-3.155	0.002	-1.976	-0.462
CashFlowtoTotalAssets	-13.2116	29.728	-0.444	0.657	-71.477	45.053
CashFlowtoLiability	24.4629	101.797	0.240	0.810	-175.055	223.981
CFOtoAssets	-13.9089	10.645	-1.307	0.191	-34.773	6.955
CashFlowtoEquity	9.1861	44.998	0.204	0.838	-79.009	97.381
CurrentLiabilitytoCurrentAssets	9.0419	7.272	1.243	0.214	-5.211	23.295
InterestCoverageRatioInterestexpensetoEBIT	99.3480	65.288	1.522	0.128	-28.614	227.310
EquitytoLiability	-63.8313	12.510	-5.103	0.000	-88.350	-39.313
LiabilityAssetsFlag	0.6552	1.376	0.476	0.634	-2.042	3.353

Figure 11: Logistic regression Model 1 - Summary

- The column named ‘coef’ in the above summary represents a linear equation that is a result of multiplication of coefficients and the corresponding predictor variables.
- Interpreting significant coefficients:
    - Null hypothesis: Predictor variable is not significant
    - Alternate hypothesis: predictor variable is significant (P > |z|) gives the p-value for each predictor variable to check the null hypothesis.
    - If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

- As an example, consider CFOtoAssets, p-value for the variable is -2.991, probability that the coefficient being -13.9089 is 0.191 which is greater than 0.05, this indicates that the null hypothesis that the variable is not significant cannot be ruled out.

- It indicates that this variable is insignificant.

- In this case we shall find the correlated variable by using Variance inflation Method.

- Variance inflation method considers one independent variable and estimates how effectively it can be predicted from other independent variables, if a variable is highly correlated with other variables in the data, then the VIF score of that variable is obtained higher.

- The VIF scores of the independent variables are as below:

	variables	VIF
18	QuickAssetstoCurrentLiability	19.614185
6	CurrentRatio	14.856060
7	QuickRatio	12.988497
20	OperatingFundstoLiability	7.912648
10	NetprofitbeforetaxtoPaidincapital	7.268975
26	CFOtoAssets	6.900905
27	CurrentLiabilitytoCurrentAssets	6.049353
28	EquitytoLiability	5.736218
22	RetainedEarningstoTotalAssets	4.934115
23	TotalincometoTotalexpense	4.717501
17	QuickAssetstoTotalAssets	4.246279
21	InventorytoCurrentLiability	3.242692
8	TotaldebttoTotalnetworth	3.115746
5	TotalAssetReturnGrowthRateRatio	2.823870
11	TotalAssetTurnover	2.808594
19	CashtoCurrentLiability	2.757314
12	AccountsReceivableTurnover	2.693366
15	Operatingprofitperperson	2.682514
3	ContinuousNetProfitGrowthRate	2.517664
16	Allocationrateperperson	2.367380
13	AverageCollectionDays	2.273547
4	NetValueGrowthRate	2.270418
24	TotalexpensetoAssets	2.209165
14	FixedAssetsTurnoverFrequency	1.868417
9	LongtermfundsuitabilityratioA	1.698698
0	OperatingExpenseRate	1.625370
1	Researchanddevelopmentexpenserate	1.414828

25	CashTurnoverRate	1.142578
2	Interestbearingdebtinterestrate	1.107938
29	LiabilityAssetsFlag	1.059592

Table 11: Variance inflation factor scores for independent variables

- Removed the variables that have VIF score of greater than 5 - 'QuickAssetstoCurrentLiability','CurrentRatio', 'QuickRatio', 'OperatingFundstoLiability', 'NetprofitbeforetaxtoPaidincapital', 'CFOtoAssets', 'CurrentLiabilitytoCurrentAssets', 'EquitytoLiability'

Second Logistic Regression model with the remaining variables:

- f_2 = 'Default ~ OperatingExpenseRate + Researchanddevelopmentexpenserate + Interestbearingdebtinterestrate + ContinuousNetProfitGrowthRate + NetValueGrowthRate + TotalAssetReturnGrowthRateRatio + TotaldebttoTotalnetworth + LongtermfundsuitabilityratioA + TotalAssetTurnover + AccountsReceivableTurnover + AverageCollectionDays + FixedAssetsTurnoverFrequency + Operatingprofitperperson + Allocationrateperperson + QuickAssetstoTotalAssets + CashtoCurrentLiability + InventorytoCurrentLiability + RetainedEarningstoTotalAssets + TotalincometoTotalexpense + TotalexpensetoAssets + CashTurnoverRate + LiabilityAssetsFlag'
- Applying logistic regression from statsmodel from the above with all default parameters
- Below is the summary obtained:
  - Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1355
Method:	MLE	**Df Model:**	22
Date:	Sun, 08 Oct 2023	**Pseudo R-squ.:**	0.4278
Time:	15:15:51	**Log-Likelihood:**	-274.90
converged:	False	**LL-Null:**	-480.46
Covariance Type:	nonrobust	**LLR p-value:**	2.074e-73

-

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.4480	0.332	-13.385	0.000	-5.099	-3.797
OperatingExpenseRate	0.1390	0.163	0.853	0.394	-0.180	0.458
Researchanddevelopmentexpenserate	0.4756	0.125	3.820	0.000	0.232	0.720
Interestbearingdebtinterestrate	0.5855	0.186	3.140	0.002	0.220	0.951
ContinuousNetProfitGrowthRate	-0.2112	0.172	-1.230	0.219	-0.548	0.125
NetValueGrowthRate	-0.3093	0.171	-1.812	0.070	-0.644	0.025
TotalAssetReturnGrowthRateRatio	0.2665	0.184	1.448	0.147	-0.094	0.627
TotaldebttoTotalnetworth	1.0825	0.163	6.656	0.000	0.764	1.401
LongtermfundsuitabilityratioA	0.2946	0.170	1.736	0.082	-0.038	0.627
TotalAssetTurnover	-0.2178	0.257	-0.849	0.396	-0.721	0.285
AccountsReceivableTurnover	-0.5605	0.197	-2.851	0.004	-0.946	-0.175
AverageCollectionDays	0.1139	0.203	0.561	0.575	-0.284	0.512
FixedAssetsTurnoverFrequency	0.1335	0.120	1.110	0.267	-0.102	0.369
Operatingprofitperperson	0.5110	0.185	2.760	0.006	0.148	0.874
Allocationrateperperson	0.6752	0.193	3.496	0.000	0.297	1.054
QuickAssetstoTotalAssets	-0.2812	0.285	-0.986	0.324	-0.840	0.278
CashtoCurrentLiability	-0.0880	0.165	-0.533	0.594	-0.412	0.236
InventorytoCurrentLiability	-0.1203	0.203	-0.593	0.553	-0.518	0.277
RetainedEarningstoTotalAssets	-0.7064	0.231	-3.056	0.002	-1.160	-0.253
TotalincometoTotalexpense	-1.0214	0.336	-3.039	0.002	-1.680	-0.363
TotalexpensetoAssets	0.5128	0.199	2.571	0.010	0.122	0.904
CashTurnoverRate	-0.3433	0.206	-1.668	0.095	-0.747	0.060
LiabilityAssetsFlag	9.3155	81.194	0.115	0.909	-149.821	168.452

Figure 12: Logistic Regression Model 2 - Summary

- From the above summary, we observe that there still exist some insignificant variables (variables for which z-value is greater than 0.05 at 95% confidence interval.)
- The variables removed from the above summary for the next model are

- 'OperatingExpenseRate','ContinuousNetProfitGrowthRate','TotalAssetReturnGrowthRateRatio','TotalAssetTurnover','AverageCollectionDays','FixedAssetsTurnoverFrequency','QuickAssetstoTotalAssets','CashtoCurrentLiability','InventorytoCurrentLiability','CashTurnoverRate','LiabilityAssetsFlag'

- Third logistic regression model with the remaining variables has been applied with default parameters.

- Summary obtained:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1366
Method:	MLE	**Df Model:**	11
Date:	Sun, 08 Oct 2023	**Pseudo R-squ.:**	0.4135
Time:	15:15:51	**Log-Likelihood:**	-281.77
converged:	True	**LL-Null:**	-480.46
Covariance Type:	nonrobust	**LLR p-value:**	2.196e-78

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.4008	0.296	-14.866	0.000	-4.981	-3.821
Researchanddevelopmentexpenserate	0.4402	0.119	3.708	0.000	0.207	0.673
Interestbearingdebtinterestrate	0.5974	0.176	3.397	0.001	0.253	0.942
NetValueGrowthRate	-0.2803	0.155	-1.807	0.071	-0.584	0.024
TotaldebttoTotalnetworth	1.0143	0.146	6.969	0.000	0.729	1.300
LongtermfundsuitabilityratioA	0.1651	0.153	1.079	0.280	-0.135	0.465
AccountsReceivableTurnover	-0.6229	0.151	-4.115	0.000	-0.920	-0.326
Operatingprofitperperson	0.5393	0.176	3.065	0.002	0.194	0.884
Allocationrateperperson	0.8339	0.164	5.075	0.000	0.512	1.156
RetainedEarningstoTotalAssets	-0.8461	0.216	-3.923	0.000	-1.269	-0.423
TotalincometoTotalexpense	-1.1653	0.311	-3.744	0.000	-1.775	-0.555
TotalexpensetoAssets	0.3218	0.176	1.827	0.068	-0.023	0.667

Figure 13: Logistic Regression Model 3 - Summary

- From the above summary, we observed there still are some variables which are insignificant according to their P|z| scores.
- Eliminating the variables that have P|z| > 0.05 as the hypothesis test suggests that these are the contributions made by these variables for predicting the response are insignificant.
- Variables removed are: 'NetValueGrowthRate', 'LongtermfundsuitabilityratioA', 'TotalexpensetoAssets'

Fourth and final logistic regression model on the remaining independent variables are:

- Equation:
    - f_4 = 'Default ~ Researchanddevelopmentexpenserate + Interestbearingdebtinterestrate + TotaldebttoTotalnetworth + AccountsReceivableTurnover + Operatingprofitperperson + Allocationrateperperson + RetainedEarningstoTotalAssets + TotalincometoTotalexpense'

- Logistic regression model has been applied on the above variables with the default parameters.

- Below is the summary obtained:

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1369
Method:	MLE	**Df Model:**	8
Date:	Sun, 08 Oct 2023	**Pseudo R-squ.:**	0.4032
Time:	15:15:51	**Log-Likelihood:**	-286.75
converged:	True	**LL-Null:**	-480.46
Covariance Type:	nonrobust	**LLR p-value:**	9.207e-79

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.2780	0.262	-16.338	0.000	-4.791	-3.765
Researchanddevelopmentexpenserate	0.4415	0.115	3.834	0.000	0.216	0.667
Interestbearingdebtinterestrate	0.5103	0.171	2.983	0.003	0.175	0.846
TotaldebttoTotalnetworth	1.0294	0.137	7.536	0.000	0.762	1.297
AccountsReceivableTurnover	-0.5767	0.148	-3.896	0.000	-0.867	-0.287
Operatingprofitperperson	0.4725	0.176	2.683	0.007	0.127	0.818
Allocationrateperperson	0.6653	0.135	4.930	0.000	0.401	0.930
RetainedEarningstoTotalAssets	-1.0470	0.193	-5.434	0.000	-1.425	-0.669
TotalincometoTotalexpense	-1.2461	0.284	-4.384	0.000	-1.803	-0.689

Figure 14: Logistic Regression Model 4 - Summary

Interpretation of the results

- The resultant logistic regression equation obtained is
    - Default = -4.7280 + 0.4415 * Researchanddevelopmentexpensetate + 0.5103 * interestbearingdebtinterestrate + 1.0294 * TotaldebttoTotalnetworth - 0.5767 * AccountsReceivableTurnover + 0.4725 * Operatingprofitperperson + 0.6653 * Allocationrateperperson - 1.0470 * RetainedEarningstoTotalAssets - 1.2461 * TotalincometoTotalexpense

- According to the z-scores of the variables from the above summary all the above variables can be declared as significant in prediction.

- Validating the above model on the training dataset with chosen threshold as 0.5

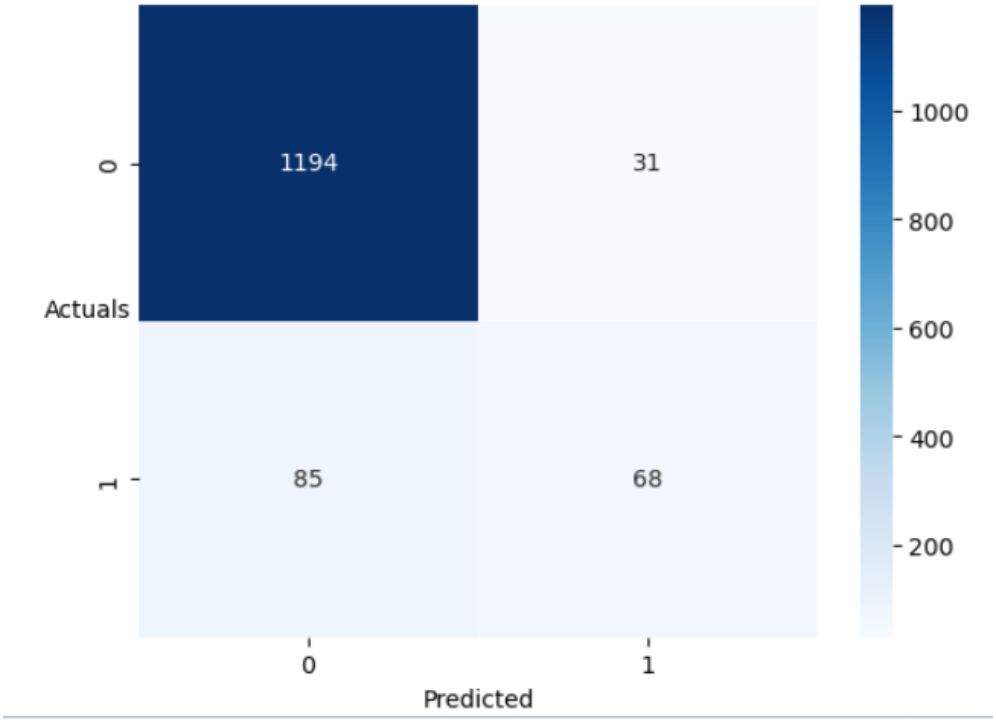


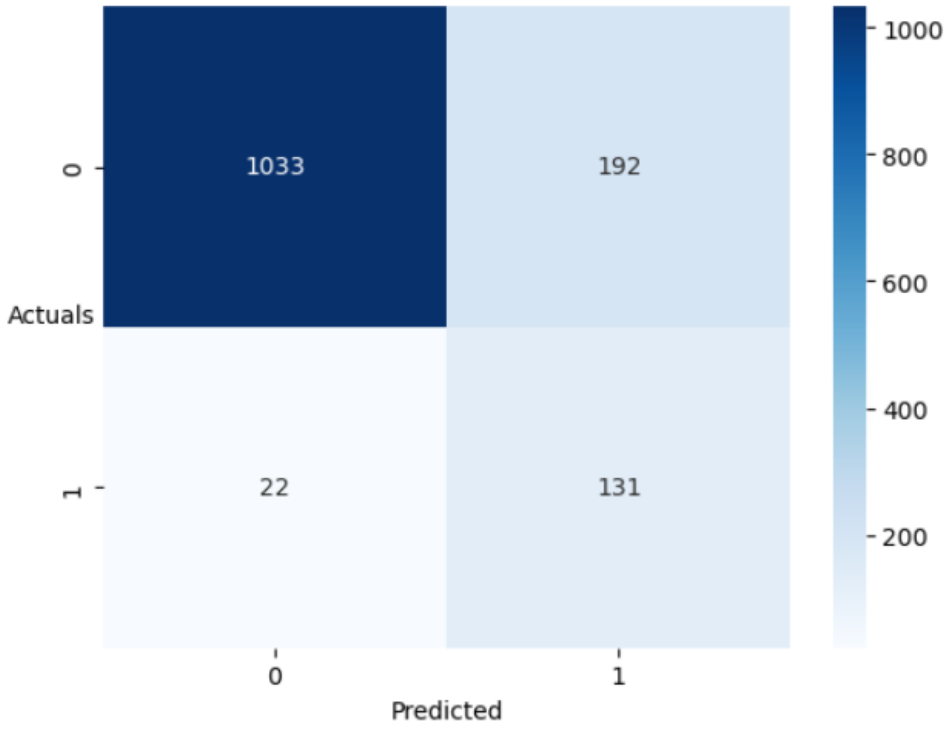
Figure 15: Classification matrix of Model 4 on training data (Cutoff - 0.5)

- As visible, there are 116 variables not correctly predicted and 1262 variables are correctly predicted to be defaults or not defaults by the model.
- The choice of the cutoff value depends on the business context and the dataset considered. Choosing 0.5 is only a general practice.
- Choosing optimum cutoff for the problem:
    - We use the ROC curve to determine optimum cutoff. We find false positive rates, true positive rates and thresholds from the roc curve obtained on training data.
    - The largest difference between false positive rate and true positive rate is considered and the threshold at the combination of false positive and true positive rates is chosen as the optimal threshold.

- The optimal threshold obtained in our case is 0.124.

1.5. Logistic Regression Model - Validation of the model on test dataset, Performance Metrics & Interpretation

- Validating the model with the above obtained cutoff on training dataset
- Classification matrix for training dataset:



-

Figure 16: Classification matrix on training data - minimal cutoff

- Classification report for training dataset:

	precision	recall	f1-score	support
0	0.979	0.843	0.906	1225
1	0.406	0.856	0.550	153
accuracy			0.845	1378
macro avg	0.692	0.850	0.728	1378
weighted avg	0.915	0.845	0.867	1378

-

Figure 17: Classification report on training data - minimal cutoff

- Interpretation:

- The resultant logistic regression model has predicted 1164 records correctly.

- 192 records have been detected as default when they don't actually default

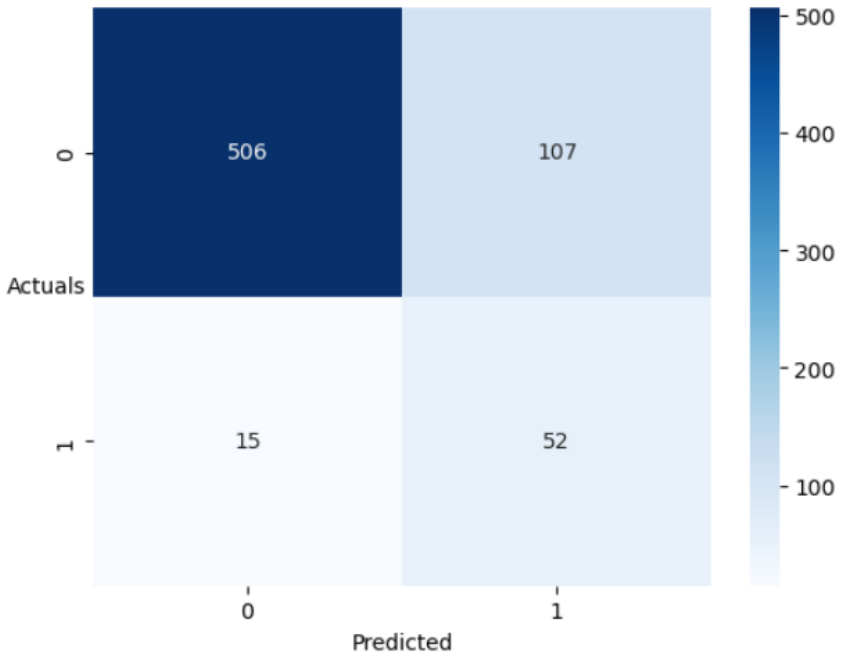
- 22 records have been recorded as non-default when they actually default.

- Recall for a model is calculated as TruePositives/ TruePositives + FalseNegatives. It is 0.86 on the training dataset.

- For this problem statement we consider recall as our important metric as the lender does not want to miss out on the defaulters because of the model predicting them as non-default.

- Having high recall is a sign of a good model as it indicates that the number of FalseNegatives are low.

- Classification matrix for test data:



-

Figure 18: Classification matrix on test data - minimal cutoff

- Classification report for test data:

	precision	recall	f1-score	support
0	0.971	0.825	0.892	613
1	0.327	0.776	0.460	67
accuracy			0.821	680
macro avg	0.649	0.801	0.676	680
weighted avg	0.908	0.821	0.850	680

-

Figure 19: Classification report on test data - minimal cutoff

- Interpretation:

- The resultant logistic regression model has predicted 1164 records correctly.

- 107 records have been detected as default when they don't actually default

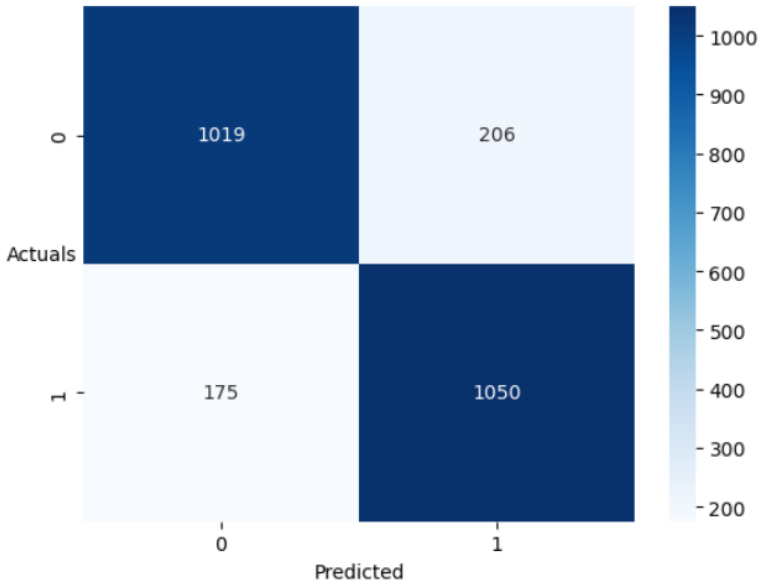
- 15 records have been recorded as non-default when they actually default.

- Recall for a model is calculated as TruePositives/ TruePositives + FalseNegatives. It is 0.78 on the training dataset.

- Recall values on training and test dataset are pretty close. This indicates that the model has not overfitted.

Logistic regression model with SMOTE:

- In the dataset the distribution of defaulters and non defaulters are far away from each other, the response variable is imbalanced in terms of the number of records of each type.
- So to lessen biasing of the model, we apply SMOTE which will populate synthetic data and make the counts of the categories of the response variable comparable.
- This helps in better model performance and can help improve precision and recall metrics.
- Applying SMOTE on the dataset with random state = 42
- Classification matrix on training dataset:



-

Figure 20: Classification matrix on training data - SMOTE treated Logit model

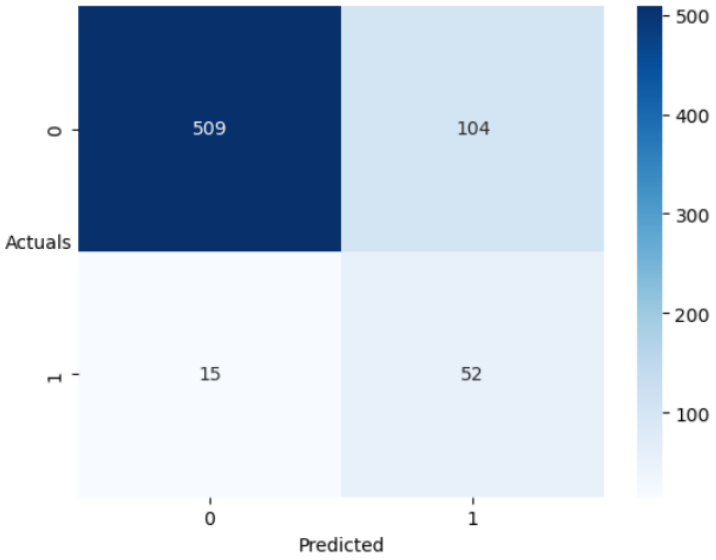
- Classification report on training dataset:

	precision	recall	f1-score	support
0	0.85	0.83	0.84	1225
1	0.84	0.86	0.85	1225
accuracy			0.84	2450
macro avg	0.84	0.84	0.84	2450
weighted avg	0.84	0.84	0.84	2450

-

Figure 21: Classification report on training data - SMOTE treated Logit model

- Classification matrix on test dataset:



-

Figure 22: Classification matrix on test data - SMOTE treated Logit model

- Classification report on test dataset

	precision	recall	f1-score	support
0	0.97	0.83	0.90	613
1	0.33	0.78	0.47	67
accuracy			0.82	680
macro avg	0.65	0.80	0.68	680
weighted avg	0.91	0.82	0.85	680

-

Figure 23: Classification report on test data - SMOTE treated Logit model

- Interpretation:
  - The precision values for the training dataset has improved but the precision on the test data set is the same.
  - The recall values are not changed much with both models either SMOTE applied or not.
  - We will consider the model without SMOTE applied.

1.6. Random Forest Model

- A random forest model has been applied on the training data considering the following variables
- 'Researchanddevelopmentexpenserate', 'Interestbearingdebtinterestrate', 'TotaldebttoTotalnetworth', 'AccountsReceivableTurnover', 'Operatingprofitperperson', 'Allocationrateperperson', 'RetainedEarningstoTotalAssets', 'TotalincometoTotalexpense'
- Basic information of the dataset considered for building the random forest model:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 8 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Researchanddevelopmentexpenserate  2058 non-null   float64
 1   Interestbearingdebtinterestrate    2058 non-null   float64
 2   TotaldebttoTotalnetworth           2058 non-null   float64
 3   AccountsReceivableTurnover         2058 non-null   float64
 4   Operatingprofitperperson           2058 non-null   float64
 5   Allocationrateperperson            2058 non-null   float64
 6   RetainedEarningstoTotalAssets      2058 non-null   float64
 7   TotalincometoTotalexpense          2058 non-null   float64
dtypes: float64(8)
memory usage: 128.8 KB
```

Figure 24: Basic info of resultant dataset

- Applied random forest with grid search algorithm in order to obtain best parameters to be applied to the model.
- Model parameters considered for grid search:
  - {
  - 'max_depth': [3, 5, 7],
  - 'min_samples_leaf': [5, 10, 15],
  - 'min_samples_split': [15, 30, 45],
  - 'n_estimators': [25, 50]
  - }

- Best parameters obtained after applying grid search algorithm are:

{'max_depth': 7, // maximum depth of the tree to be constructed

'min_samples_leaf': 5, // minimum samples to be remaining in leaf node

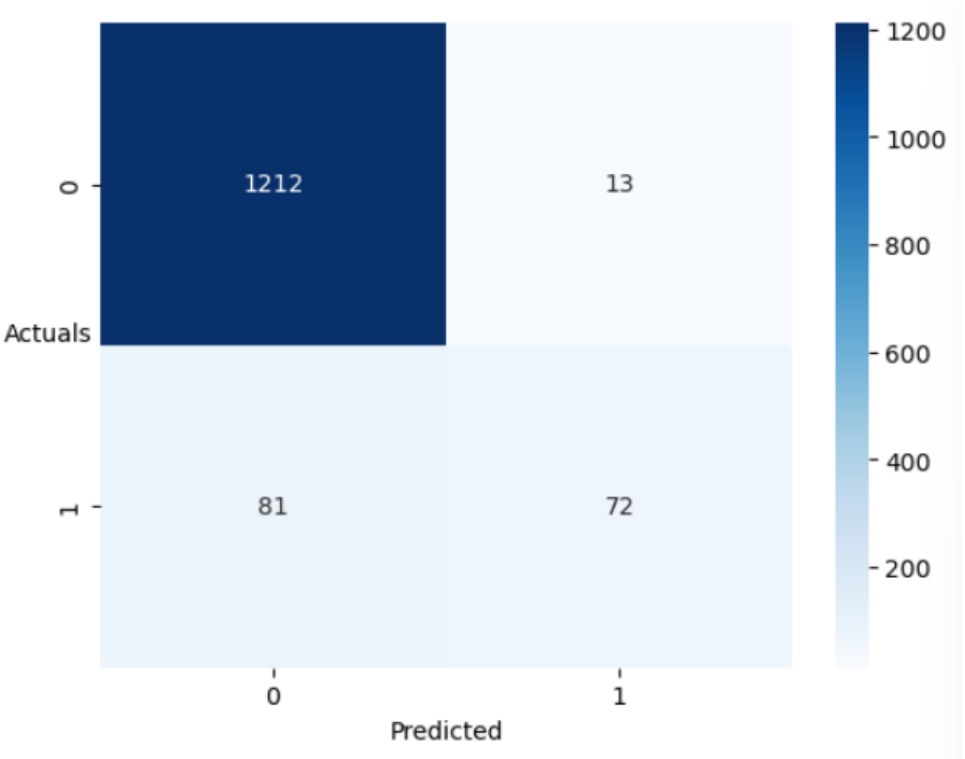
'min_samples_split': 15, // minimum samples to be present to be able to split the node

'n_estimators': 25} // number of trees to construct

- Applying a random forest model with the above parameters.

1.7. Random Forest Model - Validation of the model on test dataset, Performance metrics and Interpretation

- Classification matrix on the training dataset



-

Figure 25: Classification matrix on training data - Random Forest model

- Classification matrix on test dataset

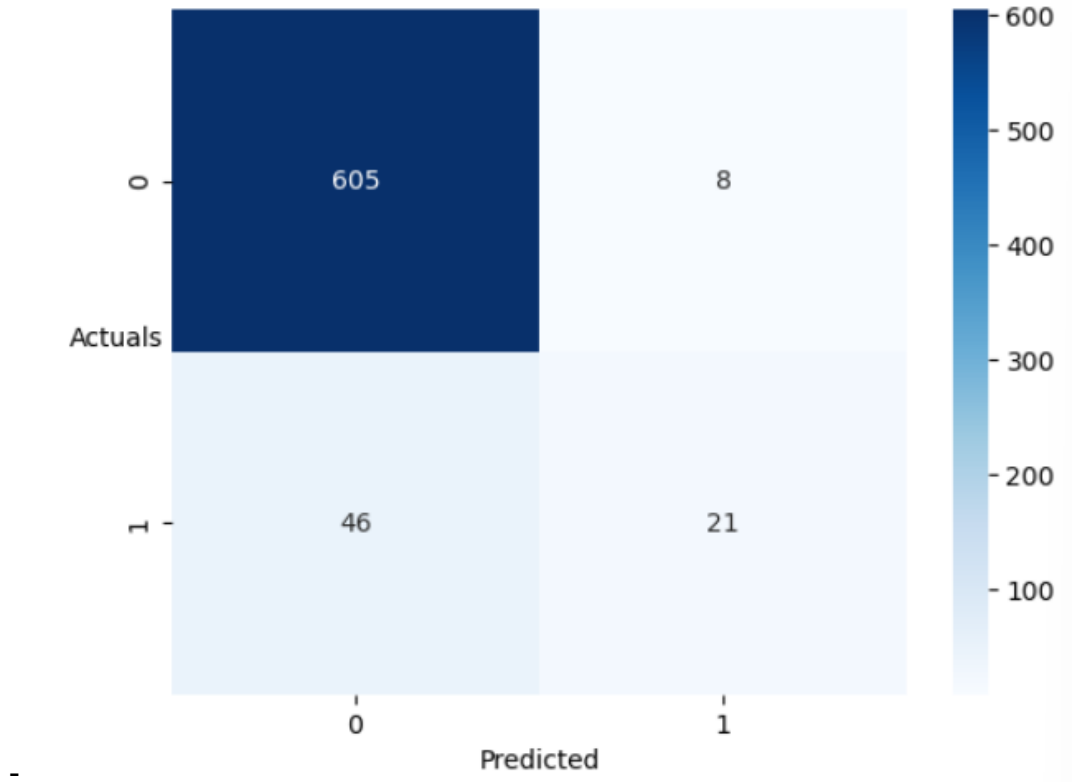


Figure 26: Classification matrix on test data - Random Forest model

- Classification report on training dataset:

	precision	recall	f1-score	support
0	0.94	0.99	0.96	1225
1	0.85	0.47	0.61	153
accuracy			0.93	1378
macro avg	0.89	0.73	0.78	1378
weighted avg	0.93	0.93	0.92	1378

Figure 27: Classification report on training data - Random Forest model

- Classification report on test dataset:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	613
1	0.72	0.31	0.44	67
accuracy			0.92	680
macro avg	0.83	0.65	0.70	680
weighted avg	0.91	0.92	0.91	680

Figure 28: Classification report on test data - Random Forest model

- Interpretation:
    - The random forest model built on the training dataset and test data set show accurate predictions in terms of false positives.
    - Precision of the model for both training and test dataset is high whereas the recall metric has been recorded low.
    - In the context of this business problem, it is important for us to reduce the false negatives i.e., we prioritize high recall metric.
    - Considering the above criteria, logistic regression model performed better than random forest model.

1.8. Linear Discriminant Analysis Model

- A linear discriminant analysis model has been applied on the training data considering the following variables
- 'Researchanddevelopmentexpenserate', 'Interestbearingdebtinterestrate', 'TotaldebttoTotalnetworth', 'AccountsReceivableTurnover', 'Operatingprofitperperson', 'Allocationrateperperson', 'RetainedEarningstoTotalAssets', 'TotalincometoTotalexpense'
- Basic information of the dataset considered for building the model:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 8 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Researchanddevelopmentexpenserate  2058 non-null   float64
 1   Interestbearingdebtinterestrate    2058 non-null   float64
 2   TotaldebttoTotalnetworth           2058 non-null   float64
 3   AccountsReceivableTurnover         2058 non-null   float64
 4   Operatingprofitperperson           2058 non-null   float64
 5   Allocationrateperperson            2058 non-null   float64
 6   RetainedEarningstoTotalAssets      2058 non-null   float64
 7   TotalincometoTotalexpense          2058 non-null   float64
dtypes: float64(8)
memory usage: 128.8 KB
```

-

Figure 29: Basic info of resultant dataset

- Applied LDA model on the above dataset with default parameters
  - solver=**'svd',**
  - shrinkage=**None,**
  - priors=**None,**
  - n_components=**None,**
  - store_covariance=**False,**
  - tol=**0.0001,**
  - covariance_estimator=**None,**

- The coefficients obtained by the matrix are array([[0.5797559 , 0.24854823, 1.26895049, -0.26042433, -0.12009656, 0.58589859, -1.15990773, -0.39523498]])

- The equation obtained by the LDA model is: **0.5797559 * Researchanddevelopmentexpenserate + 0.24854823 * Interestbearingdebtinterestrate + 1.26895049 * TotaldebttoTotalnetworth -0.26042433 * AccountsReceivableTurnover -0.12009656 * Operatingprofitperperson + 0.58589859 * Allocationrateperperson -1.15990773 * RetainedEarningstoTotalAssets -0.39523498 * TotalincometoTotalexpense**

- The metrics like total debt, interest bearing rate and research and development expense rate contributing to company being default whereas the remaining variables contribute to the company being classified as non default.

1.9. Linear Discriminant Analysis Model - Validation of the model on test dataset, Performance metrics and Interpretation

- Classification matrix on training dataset:

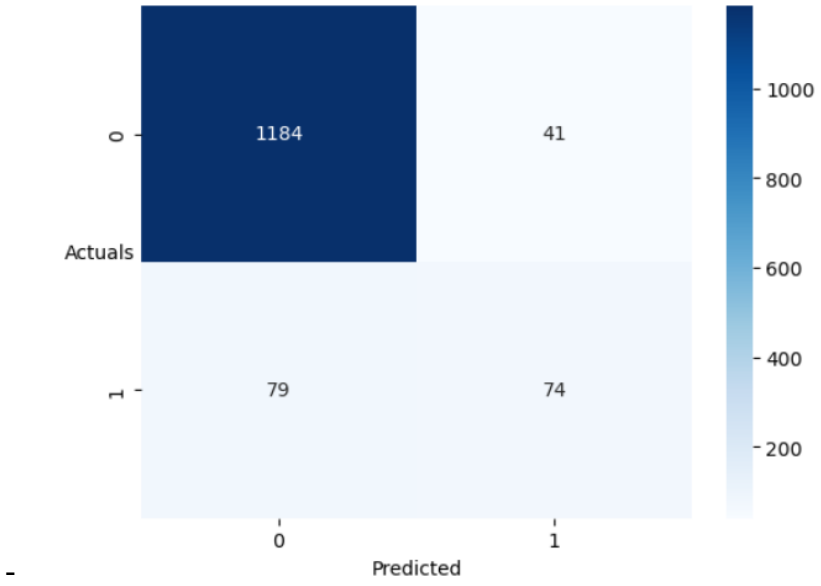


Figure 30: Classification matrix on training data - LDA model

- Classification matrix on test dataset:

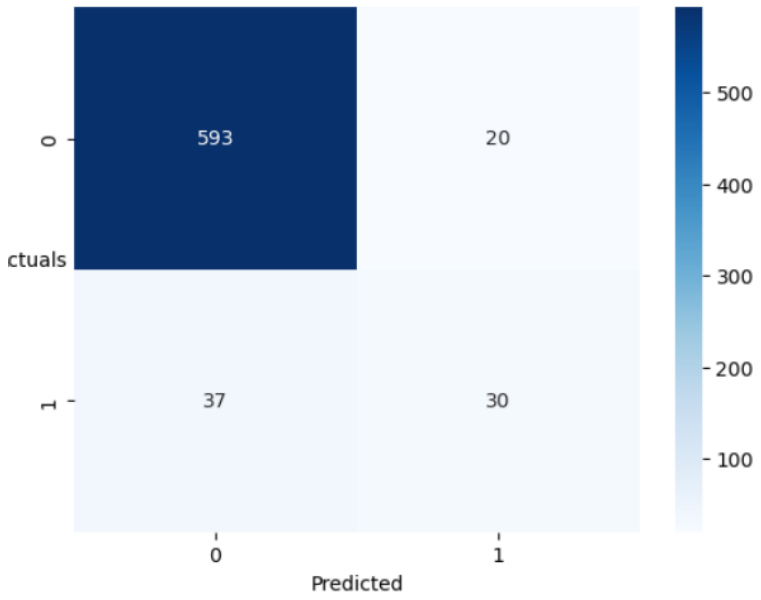


Figure 31: Classification matrix on test data - LDA model

- Classification report on training dataset:

	precision	recall	f1-score	support
0	0.937	0.967	0.952	1225
1	0.643	0.484	0.552	153
accuracy			0.913	1378
macro avg	0.790	0.725	0.752	1378
weighted avg	0.905	0.913	0.907	1378

Figure 32: Classification report on training data - LDA model

- Classification report on test dataset:

	precision	recall	f1-score	support
0	0.941	0.967	0.954	613
1	0.600	0.448	0.513	67
accuracy			0.916	680
macro avg	0.771	0.708	0.733	680
weighted avg	0.908	0.916	0.911	680

-

Figure 33: Classification report on test data - LDA model

- Interpretation:

- The above metrics are calculated with a threshold of 0.5.

- LDA model has improved on precision and balanced recall well as well compared to logistic regression and random forest models.

- Since our goal is to maximize recall, we can say that logistic regression models perform better than LDA model in this case.

1.10. Comparison of Logistic Regression, Random Forest and Linear Discriminant Analysis Models

ModelName/Metrics on test data	Accuracy	Precision	Recall	AUC score
Logistic Regression	**0.821**	**0.327**	**0.776**	**0.800**
Random Forest	**0.92**	**0.72**	**0.31**	**0.65**
Linear Discriminant Analysis	**0.916**	**0.6**	**0.448**	**0.90**

Table 12: Performance metrics of Logit, Random Forest, LDA models

AUC ROC curves for the three models:

Logistic Regression model - Training data:

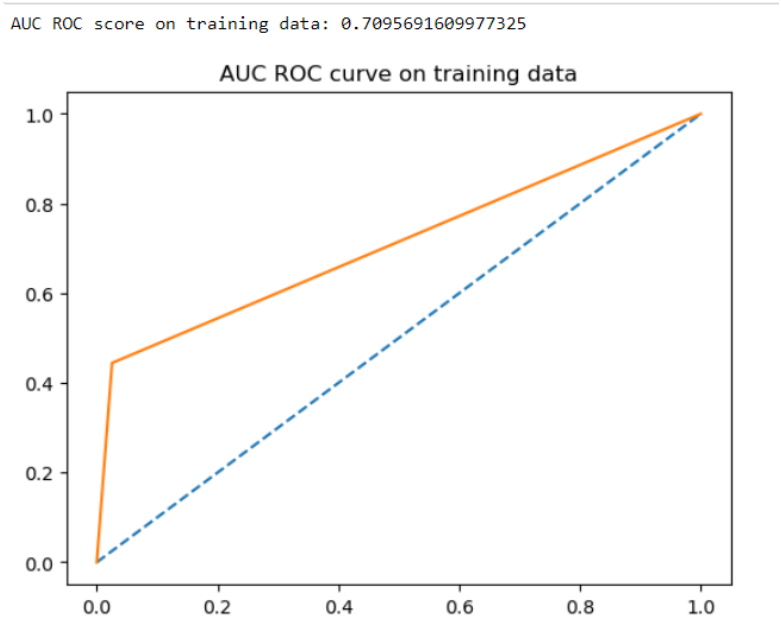


Figure 34: AUC ROC curve - Training data - Logit Model

Test data:

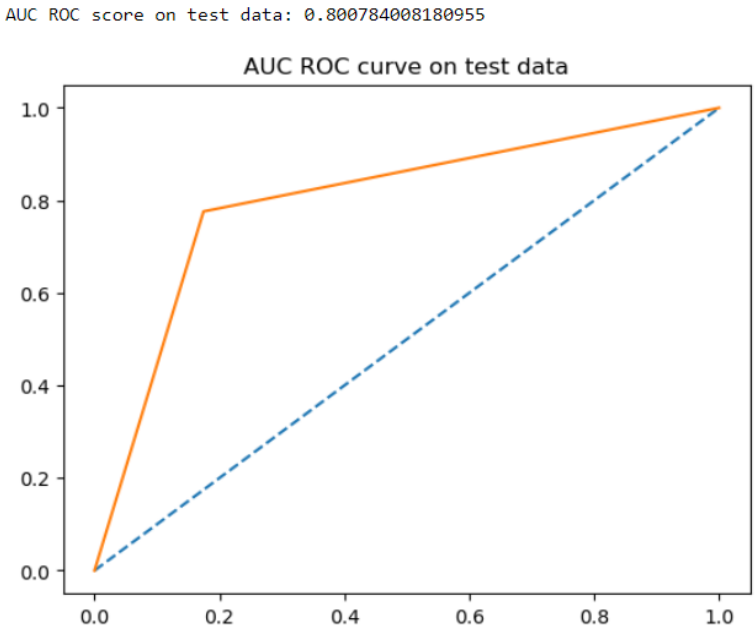


Figure 35: AUC ROC curve - Test data - Logit Model

Random Forest model:

Training data:

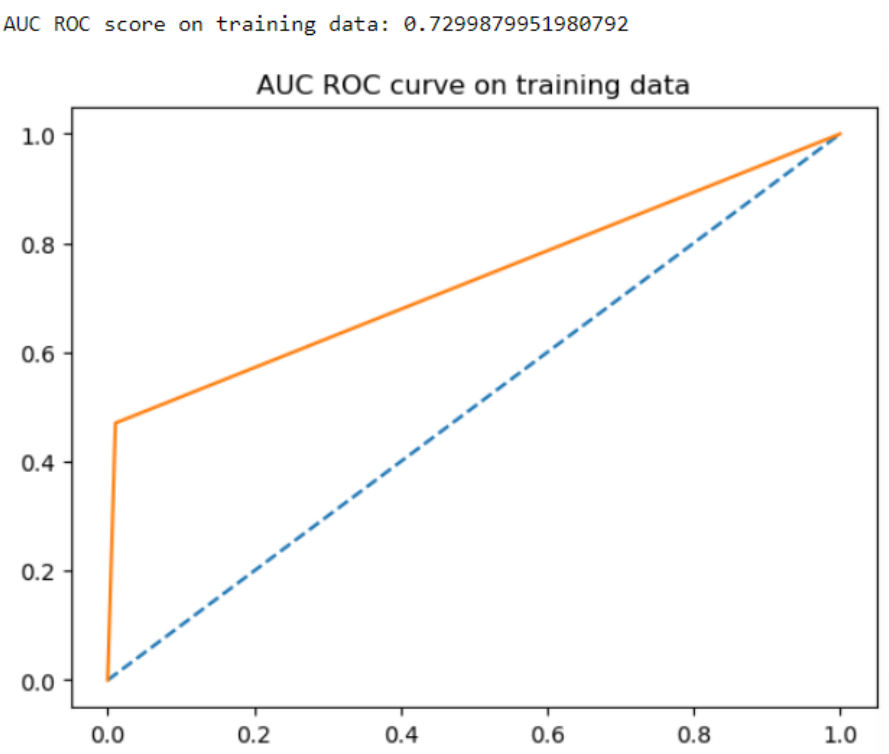


Figure 36: AUC ROC curve - Training data - Random Forest Model

Test data:

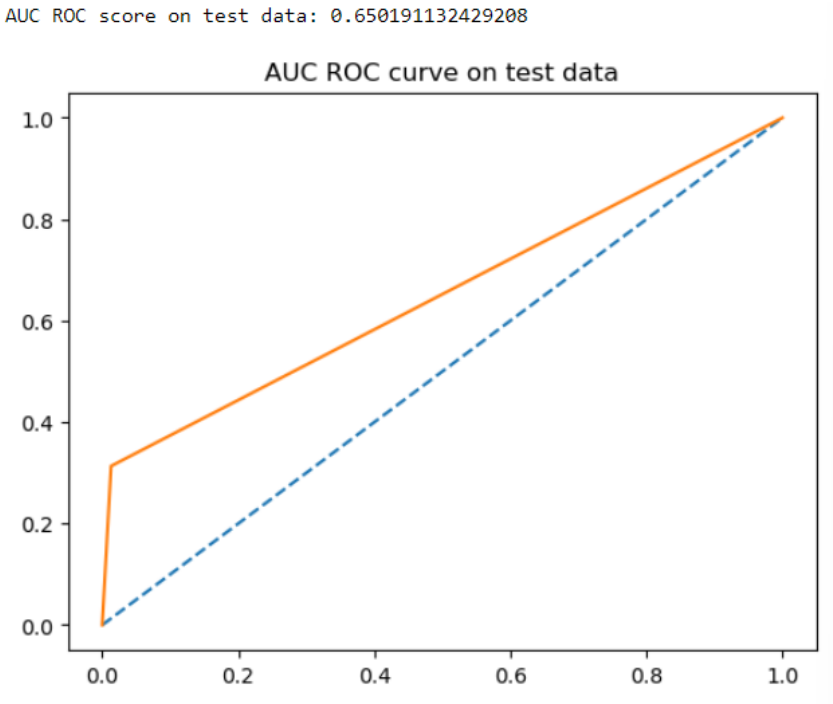


Figure 37: AUC ROC curve - Test data - Random Forest Model

Linear Discriminant Analysis model:

Training data:

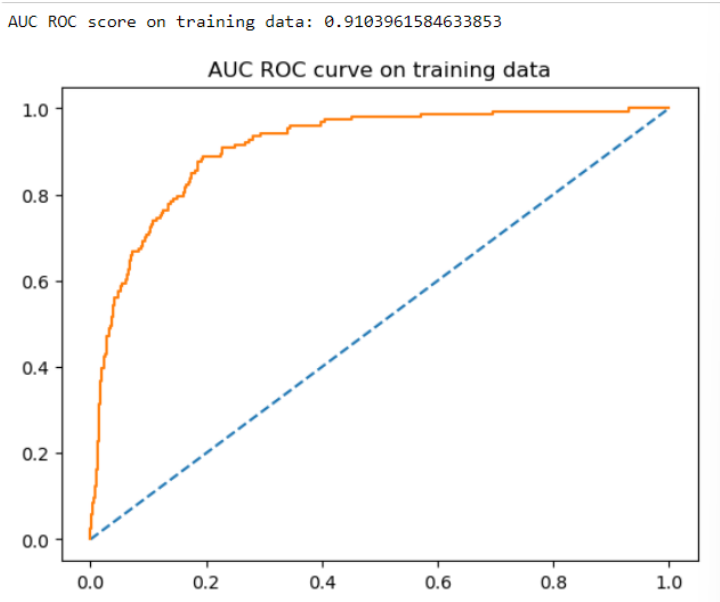


Figure 38: AUC ROC curve - Training data - LDA Model

Test data:

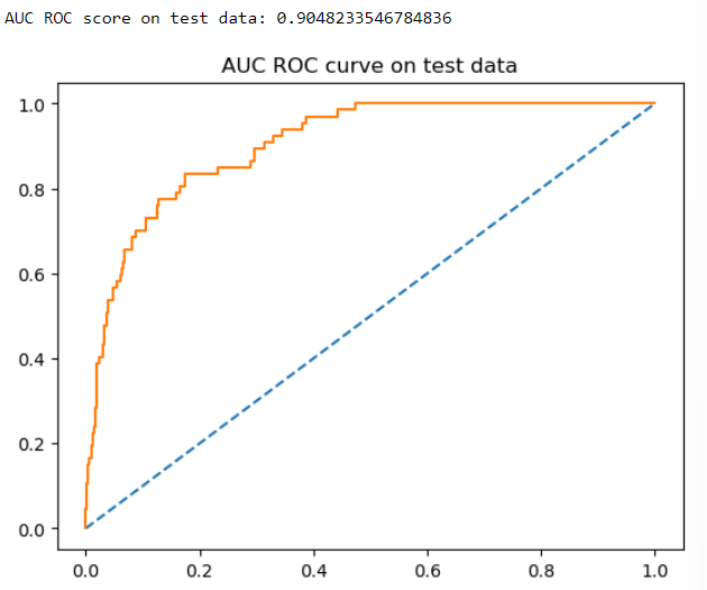


Figure 39: AUC ROC curve - Test data - LDA Model

Choosing the best model:

- Accuracy of the Linear Discriminant Analysis is the highest.
- Precision given by the Random Forest model is the highest.
- Recall provided by the Logistic Regression model is the highest.
- AUC ROC score by Linear Discriminant model is highest.
- Though, comparatively Linear Discriminant Analysis performs better among the three models, we need to consider Recall as our most important metric in choosing the best model.
- This is because banks want to know most of the companies which are going to be default which means we need the least false negatives.
- A metric of high recall provides us the strength of correct predictions when false negatives are low.
- Hence, we choose Logistic Regression model with the formula as our best model

- 'Default ~ Researchanddevelopmentexpenserate + Interestbearingdebtinterestrate + TotaldebttoTotalnetworth + AccountsReceivableTurnover + Operatingprofitperperson + Allocationrateperperson + RetainedEarningstoTotalAssets + TotalincometoTotalexpense'

1.11. Conclusions and Recommendations

- Below are the conclusions and recommendations from the logistic regression model chosen:
    - **Equation obtained by the most optimum model:**
        - **Default = -4.7280 + 0.4415 * Researchanddevelopmentexpensetate + 0.5103 * interestbearingdebtinterestrate + 1.0294 * TotaldebttoTotalnetworth - 0.5767 * AccountsReceivableTurnover + 0.4725 * Operatingprofitperperson + 0.6653 * Allocationrateperperson - 1.0470 * RetainedEarningstoTotalAssets - 1.2461 * TotalincometoTotalexpense**

- A unit change in

- Researchanddevelopmentexpensetate is going to increase the probability of default by 0.4 considering all other variables not to affect.

- Similar effect is shown by variables like interestbearingdebtinterestrate , TotaldebttoTotalnetworth , Operatingprofitperperson, Allocationrateperperson

- In contrast, AccountsReceivableTurnover is going to increase the probability of default by 0.4 considering all other variables not to affect.

- Similar effect is shown by variables like RetainedEarningstoTotalAssets , TotalincometoTotalexpense

- The above conclusions indicate that for a company to default it's net revenue or net profit are on a lower side and its expenses are on a higher side.

- Similarly, the probability of a company defaulting is less if the company has positive income over expense, more ways or means of obtaining cash or profits, more retained earnings and more income.

2. Market Risk Analytics - Problem Statement

The dataset contains 6 years of information(weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights. You are expected to do the Market Risk Analysis using Python.

2.1.Stock Price Graph(Stock Price vs time) for 2 stocks - Inference

Knowing the dataset:

- The dataset has 314 rows and 11 columns
- <u>First and last rows of the dataset:</u>

	Date	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
0	31-03-2014	264	69	455	263	68	5543	555	298	83	278
1	07-04-2014	257	68	458	276	70	5728	610	279	84	303
2	14-04-2014	254	68	454	270	68	5649	607	279	83	280
3	21-04-2014	253	68	488	283	68	5692	604	274	83	282
4	28-04-2014	256	65	482	282	63	5582	611	238	79	243

Table 13: First 5 rows of Stocks dataset

	Date	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
309	02-03-2020	729	120	469	658	33	23110	401	146	3	22
310	09-03-2020	634	114	427	569	30	21308	384	121	6	18
311	16-03-2020	577	90	321	428	27	18904	365	105	3	16
312	23-03-2020	644	75	293	360	21	17666	338	89	3	14
313	30-03-2020	633	75	284	379	23	17546	352	82	3	14

Table 14: Last 5 rows of Stocks dataset

- The data contains stock prices of 10 different stocks (visible in the column names) starting from 31st March 2014 to 20th March 2020 i.e 6 years of data.
- <u>Basic information of the dataset:</u>

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 314 entries, 0 to 313
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Date                 314 non-null    object
 1   Infosys              314 non-null    int64
 2   Indian Hotel         314 non-null    int64
 3   Mahindra & Mahindra  314 non-null    int64
 4   Axis Bank            314 non-null    int64
 5   SAIL                 314 non-null    int64
 6   Shree Cement         314 non-null    int64
 7   Sun Pharma           314 non-null    int64
 8   Jindal Steel         314 non-null    int64
 9   Idea Vodafone        314 non-null    int64
 10  Jet Airways          314 non-null    int64
dtypes: int64(10), object(1)
memory usage: 27.1+ KB
```

Figure 40: Basic information of Stocks dataset

- The dataset does not contain any null values and duplicated rows.
- Summary of the stock prices of the stocks:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Date	314	314	31-03-2014	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Infosys	314.0	NaN	NaN	NaN	511.340764	135.952051	234.0	424.0	466.5	630.75	810.0
Indian Hotel	314.0	NaN	NaN	NaN	114.56051	22.509732	64.0	96.0	115.0	134.0	157.0
Mahindra & Mahindra	314.0	NaN	NaN	NaN	636.678344	102.879975	284.0	572.0	625.0	678.0	956.0
Axis Bank	314.0	NaN	NaN	NaN	540.742038	115.835569	263.0	470.5	528.0	605.25	808.0
SAIL	314.0	NaN	NaN	NaN	59.095541	15.810493	21.0	47.0	57.0	71.75	104.0
Shree Cement	314.0	NaN	NaN	NaN	14806.410828	4288.275085	5543.0	10952.25	16018.5	17773.25	24806.0
Sun Pharma	314.0	NaN	NaN	NaN	633.468153	171.855893	338.0	478.5	614.0	785.0	1089.0
Jindal Steel	314.0	NaN	NaN	NaN	147.627389	65.879195	53.0	88.25	142.5	182.75	338.0
Idea Vodafone	314.0	NaN	NaN	NaN	53.713376	31.248985	3.0	25.25	53.0	82.0	117.0
Jet Airways	314.0	NaN	NaN	NaN	372.659236	202.262668	14.0	243.25	376.0	534.0	871.0

Table 15: Summary of Stocks dataset

- Stock prices of Shree Cement are the highest whereas SAIL are the lowest.
- Data type of the ‘Date’ column is converted to datetime.

Stock price trend for Infosys:

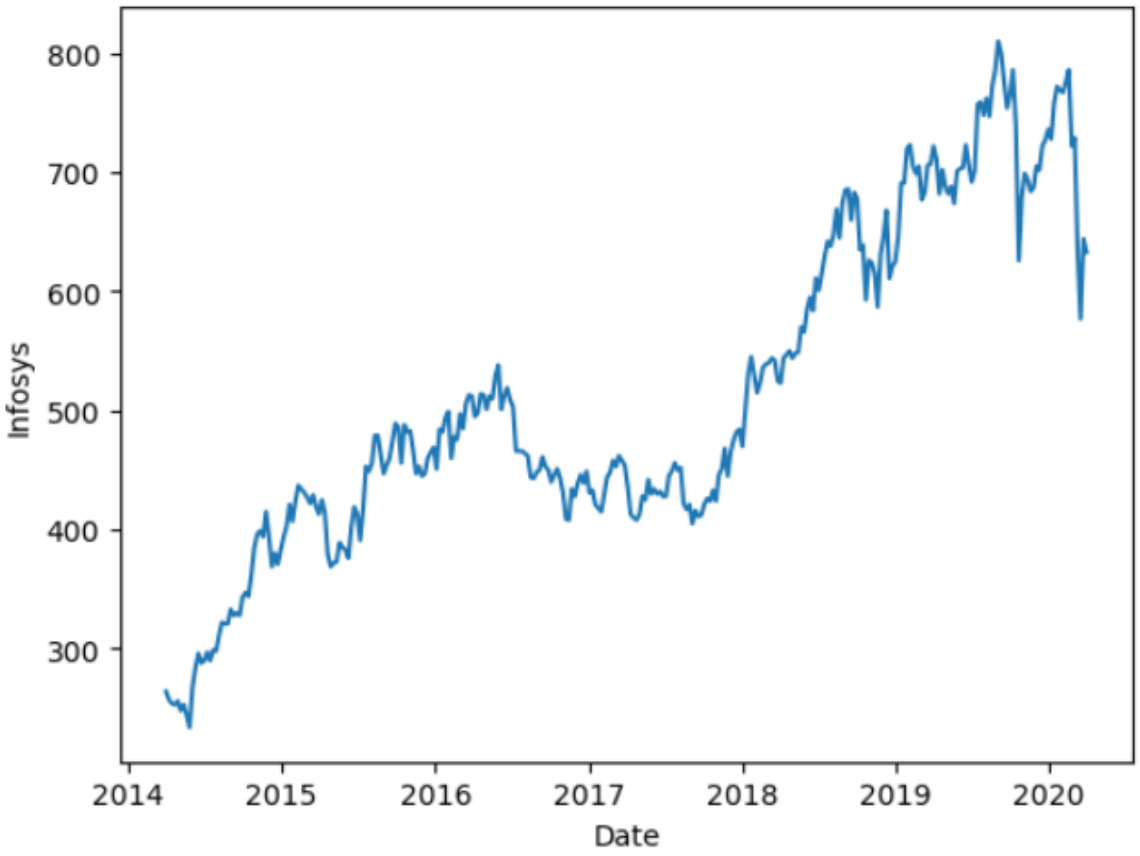


Figure 41: Stock price trend of Infosys

- Stock prices of Infosys belonging to the IT industry have been increasing since 2104 but have seen a dip in the years 2107 - 2018.
- But once the company has recovered from the dip, stock prices have soared high and have been recording their highest.
- Starting stock price was around Rs.234 whereas it is Rs.810 right now. Stock price has increased 4 times over 6 years for the company.

Stock price trend for Shree Cement:

Figure 42: Stock price trend of Shree Cement

- This is the stock that has been recording the highest stock prices among the others in the dataset.
- The stock prices of Shree Cement belonging to the Cement industry have been seeing a few ups and downs during some years like 2016 and 2019 but the price has steadily been increasing.
- The stock prices recorded in March 2014 were around Rs.5543 but in the year 2020 it has soared to Rs.24800 which is almost 5 times the lowest price it has recorded during 2014.

2.2.Calculation of Returns for all stocks - Inference

- Logarithmic difference of the stocks has been calculated.
- Viewing few of differenced data set showing stock returns

	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846

Table 16: Table showing stock returns

Summary of the differences in stock returns:

	count	mean	std	min	25%	50%	75%	max
Infosys	313.0	0.002794	0.035070	-0.167300	-0.014514	0.004376	0.024553	0.135666
Indian_Hotel	313.0	0.000266	0.047131	-0.236389	-0.023530	0.000000	0.027909	0.199333
Mahindra_&_Mahindra	313.0	-0.001506	0.040169	-0.285343	-0.020884	0.001526	0.019894	0.089407
Axis_Bank	313.0	0.001167	0.045828	-0.284757	-0.022473	0.001614	0.028522	0.127461
SAIL	313.0	-0.003463	0.062188	-0.251314	-0.040822	0.000000	0.032790	0.309005
Shree_Cement	313.0	0.003681	0.039917	-0.129215	-0.019546	0.003173	0.029873	0.152329
Sun_Pharma	313.0	-0.001455	0.045033	-0.179855	-0.020699	0.001530	0.023257	0.166604
Jindal_Steel	313.0	-0.004123	0.075108	-0.283768	-0.049700	0.000000	0.037179	0.243978
Idea_Vodafone	313.0	-0.010608	0.104315	-0.693147	-0.045120	0.000000	0.024391	0.693147
Jet_Airways	313.0	-0.009548	0.097972	-0.458575	-0.052644	-0.005780	0.036368	0.300249

Table 17: Summary of stock differences

Inference:

- Logarithmic differences for each column have been calculated considering the previous column.
- All stock prices are recorded in different price scales, so a logarithmic difference has been taken to record the ups and downs stock prices are going through compared to the price in the previous day.
- Returns Idea vodafone stocks decline by 1% on an average.
- Returns on Mahindra & Mahindra, SAIL, Sun_Pharma and Jindal_Steel, Idea Vodafone and Jet Airways are declining by round 01.% - 0.9% on an average over the years.

- Companies like Shree_Cement, Infosys, Axis_Bank have shown positive returns on an average.
- Surprisingly, stocks for JetAirways and SAIL have increased by 30% in the past.

2.3. Calculation of Stock Means and Standard Deviation for all stocks - Inference

Mean of all stock prices

	Company	StockPrices_Average
0	Infosys	511.340764
1	Indian_Hotel	114.560510
2	Mahindra_&_Mahindra	636.678344
3	Axis_Bank	540.742038
4	SAIL	59.095541
5	Shree_Cement	14806.410828
6	Sun_Pharma	633.468153
7	Jindal_Steel	147.627389
8	Idea_Vodafone	53.713376
9	Jet_Airways	372.659236

Mean of returns on all stocks

	Company	StockReturns_Average
0	Infosys	0.002794
1	Indian_Hotel	0.000266
2	Mahindra_&_Mahindra	-0.001506
3	Axis_Bank	0.001167
4	SAIL	-0.003463
5	Shree_Cement	0.003681
6	Sun_Pharma	-0.001455
7	Jindal_Steel	-0.004123
8	Idea_Vodafone	-0.010608
9	Jet_Airways	-0.009548

Figure 43: Average of stock prices and stock returns

Inference:

- Average stock prices in the dataset range from Rs.53 to Rs.14806.
- Highest average stock price is possessed by Shree_Cement whereas least stock price is possessed by Idea_Vodafone.
- The table towards the right indicates the average returns on all stocks per day.
- Positive values indicate that returns on stocks increase on an average and negative values indicate that returns on stocks decrease on average.

- Infosys, Indian_Hotel, Axis_Bank, Shree_Cement show positive signs of stock returns whereas Mahindra & Mahindra, SAIL, Sun_Pharma, Jindal_Steel, Idea_Vodafone, Jet_Airways show negative signs of stock returns.

Standard deviation of all stocks:

Standard deviation of stock prices

	Company	StockPrices_StandardDeviations
0	Infosys	135.952051
1	Indian_Hotel	22.509732
2	Mahindra_&_Mahindra	102.879975
3	Axis_Bank	115.835569
4	SAIL	15.810493
5	Shree_Cement	4288.275085
6	Sun_Pharma	171.855893
7	Jindal_Steel	65.879195
8	Idea_Vodafone	31.248985
9	Jet_Airways	202.262668

Standard deviation of stock returns

	Company	StockReturns_StandardDeviations
0	Infosys	0.035070
1	Indian_Hotel	0.047131
2	Mahindra_&_Mahindra	0.040169
3	Axis_Bank	0.045828
4	SAIL	0.062188
5	Shree_Cement	0.039917
6	Sun_Pharma	0.045033
7	Jindal_Steel	0.075108
8	Idea_Vodafone	0.104315
9	Jet_Airways	0.097972

Figure 44: Standard deviations of stock prices and stock returns

Inference:

- A stock price or return of stock having high values of standard deviation indicates that the stock price is more volatile and is more subject to change.
- This low stability can cause risk while attempting to invest in the stocks and are not a safer bet.
- Companies like Infosys, Jet_Airways and IdeaVodafone have shown more volatility in terms of average stock prices in the past.
- Idea_Vodafone is again the company that showed the highest volatility in terms of stock returns.
- Least volatility in terms of stock returns are shown by Sun_Pharma, Shree_Cement and Infosys.

2.4.Plot of Stock Means vs Stock Standard Deviations - Inference

- Average of all stock prices of all stocks - 1787.6296178343946
- Standard deviation of all stock prices of all stocks - 1327.2139000396146
- A vertical and horizontal line is drawn with these overall values as shown below.

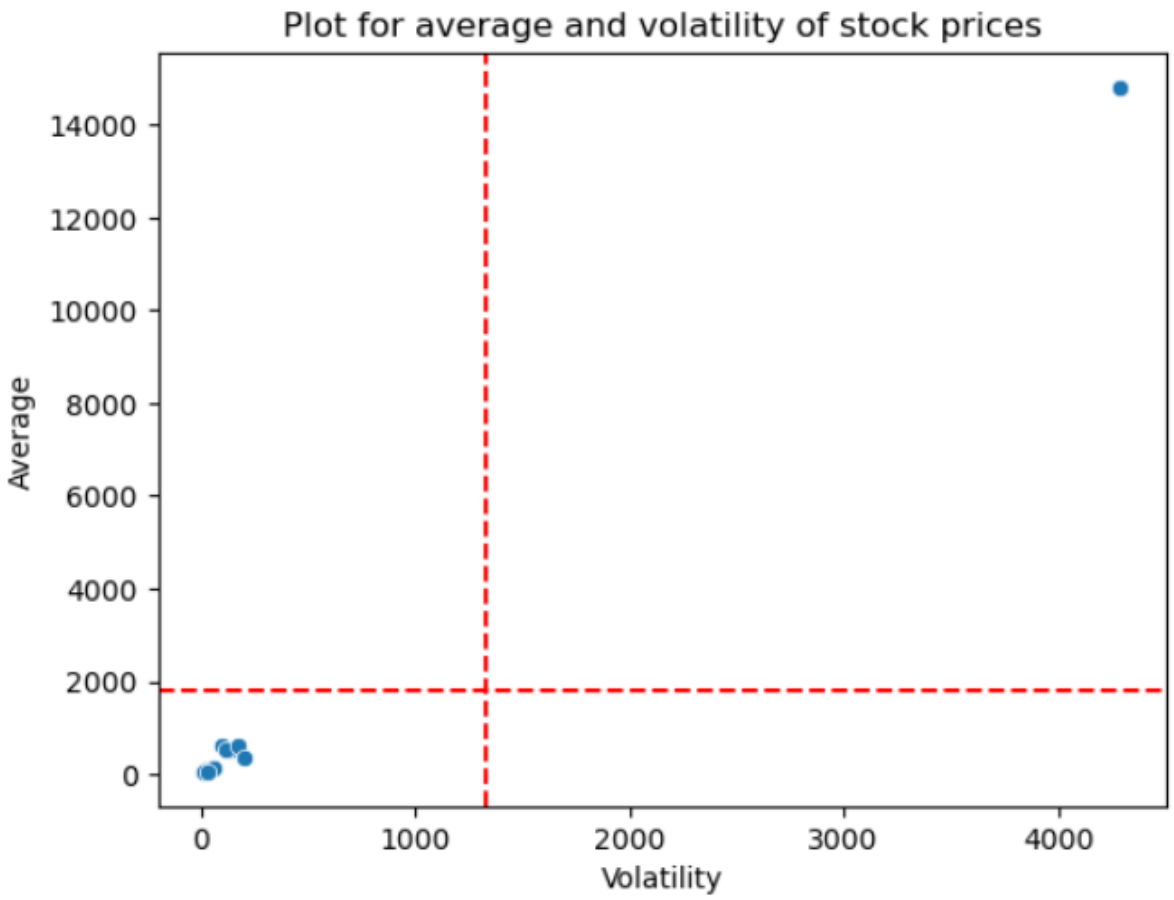


Figure 45: Plot for average and volatility of stock prices

- Average of stock returns of all stocks: -0.0022793178008050996
- Standard deviation on stock returns of all stocks: 0.02503780557315118
- A vertical and horizontal line is drawn with these overall values as shown below.

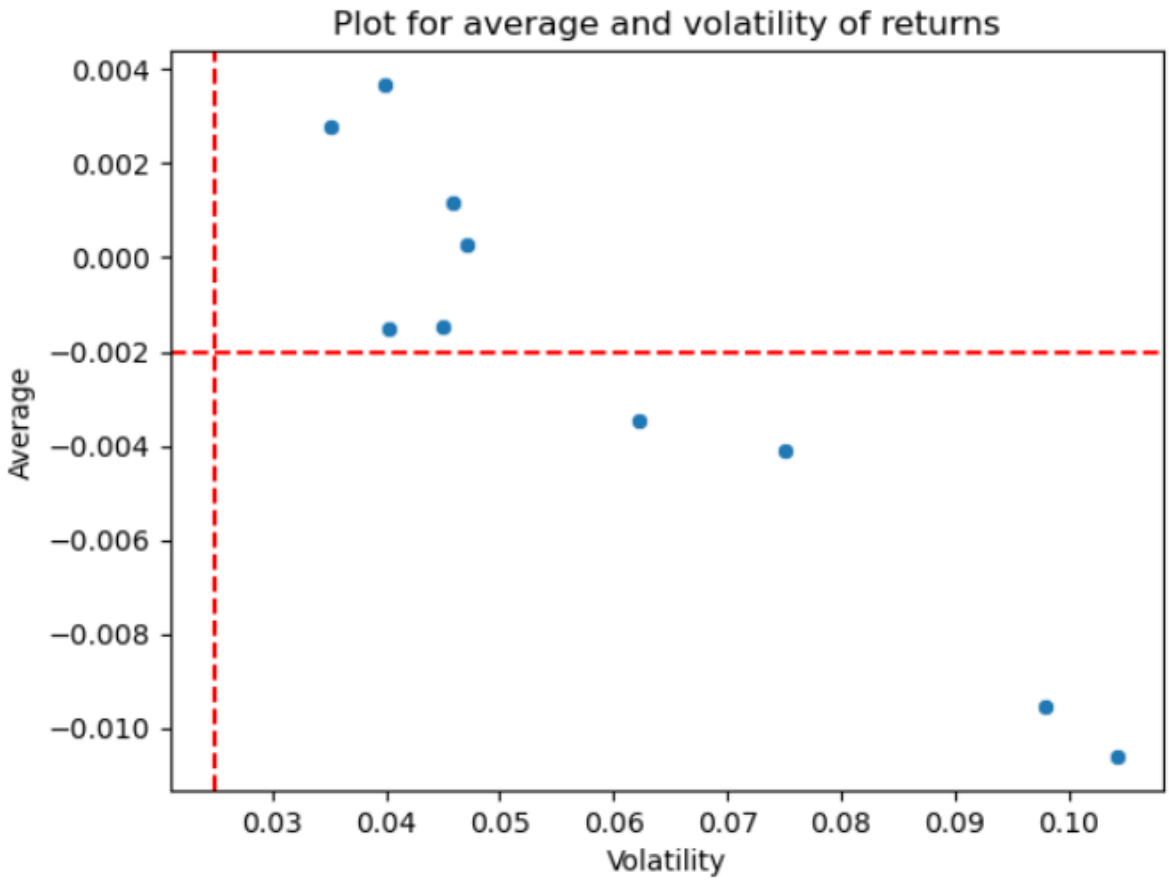


Figure 46: Plot for average and volatility of stock returns

Inference:

- The first graph represents the trend of average and standard deviations of all stock prices compared to the overall average and standard deviation.
- Since Shree_Cement values are higher than all other stocks, the bar was set very high.
- Considering the stocks excluding Shree_Cement, most of them show similar stock prices and similar standard deviations.
- The second graph represents the trend of average and standard deviations of all stock returns compared to the overall average and standard deviation of returns.
- This graph shows that 6 stocks perform better in returns than the overall average and standard deviation.

- 4 stocks underperform. The stocks performing well in terms of returns are Infosys, Shree_Cement, Indian_hotel, Axis_Bank, Mahindra & Mahindra, Sun_Pharma.

2.5.Conclusions and Recommendations

- As described in the above points, Companies like Shree_Cement, Infosys, Axis_Bank have shown positive returns on an average.
- Companies like Infosys, Jet_Airways and IdeaVodafone have shown more volatility in terms of average stock prices in the past.
- The stocks performing well in terms of returns are Infosys, Shree_Cement, Indian_hotel, Axis_Bank, Mahindra & Mahindra, Sun_Pharma.
- To invest in stocks here is the order based on the price trends - Shree_Cement (if investing in the stock for that stock price is affordable), Infosys, Indian_Hotel, Axis_Bank, Mahindra & Mahindra, Sun_Pharma.
- Companies not recommended to invest in stocks are SAIL, Jindal Steel, Jet Airways, Idea_Vodafone.

—---------- END OF REPORT —----------