Business Report

Finance and Risk Analytics

Prepared by Dhruv Dosad

Great Learning

Table of Contents

Table of Contents	2
List of Figures	3
List of Tables	4
PART A: Companies Credit Risk	**6-55**
Outlier Treatment	13
Missing Value Treatment	16
Univariate & Bivariate analysis with proper interpretation.	24
Train Test Split	34
Build Logistic Regression Model (using statsmodels library) on most important variables on train dataset and choose the optimum cut-off. Also showcase your model building approach	34
Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model	38
Build a Random Forest Model on Train Dataset. Also showcase your model building approach	45
Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model	47
Build a LDA Model on Train Dataset. Also showcase your model building approach	49
Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model	50
Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve)	53
Conclusions and Recommendations	54
PART B: Market Risk Analysis	**56-63**
Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference	57
Calculate Returns for all stocks with inference	58
Calculate Stock Means and Standard Deviation for all stocks with inference	60
Draw a plot of Stock Means vs Standard Deviation and state your inference	60
Conclusions and Recommendations	63

List Of Tables

Table	Page No
Table 1.1 Sample of the dataset	10
Table 1.2 Information of the dataset	11
Table 1.3 Descriptive Statistics of the dataset	12
Table 1.4 Outliers in Each variable	14
Table 1.5 Sample of scaled variables	18
Table 1.6 Variables with highest VIF (>5.0)	19
Table 1.7 Variables with highest VIF - Top 5	19
Table 1.8 Variables with highest VIF - Top 5	20
Table 1.9 Variables with highest VIF - Top 5	20
Table 1.10 Variables with highest VIF - Top 5	20
Table 1.11 Variables with highest VIF - Top 5	20
Table 1.12 Variables with highest VIF - Top 5	20
Table 1.13 Variables with highest VIF - Top 5	20
Table 1.14 Variables with highest VIF - Top 5	20
Table 1.15 Variables with highest VIF - Top 5	21
Table 1.16 Variables with highest VIF - Top 5	21
Table 1.17 Variables with highest VIF - Top 5	21
Table 1.18 Final Variables after removing VIF > 5	22
Table 1.19 RFE Rank of all Variables	23
Table 1.20 Selected Variables with Rank 1	24
Table 1.21 Descriptive Statistics of the Variable	25
Table 1.22 Descriptive Statistics of the Variable	25
Table 1.23 Descriptive Statistics of the Variable	25
Table 1.24 Descriptive Statistics of the Variable	26
Table 1.25 Descriptive Statistics of the Variable	26
Table 1.26 Descriptive Statistics of the Variable	26
Table 1.27 Descriptive Statistics of the Variable	27
Table 1.28 Descriptive Statistics of the Variable	27
Table 1.29 Descriptive Statistics of the Variable	27
Table 1.30 Descriptive Statistics of the Variable	28
Table 1.31 Descriptive Statistics of the Variable	28
Table 1.32 Descriptive Statistics of the Variable	28
Table 1.33 Descriptive Statistics of the Variable	29
Table 1.34 Descriptive Statistics of the Variable	29
Table 1.35 Logistic Regression Model Summary	36
Table 1.36 Logistic Regression Model Summary	37
Table 1.37 Logistic Regression Model Summary	37
Table 1.38 Logistic Regression Model Summary	37
Table 1.39 Logistic Regression Classification report	38
Table 1.40 Logistic Regression Conclusion	39
Table 1.41 Logistic Regression Optimal Cutoff Classification report	40
Table 1.42 Logistic Regression Conclusion	41
Table 1.43 Logistic Regression SMOTE Model Summary	42

Table 1.44 Logistic Regression SMOTE Model Summary	42
Table 1.45 Logistic Regression SMOTE Classification report	43
Table 1.46 Logistic Regression Conclusion	44
Table 1.47 Logistic Regression Optimal Cutoff Classification report	44
Table 1.48 Logistic Regression Conclusion	45
Table 1.49 Logistic Regression Metric Comparison	46
Table 1.50 Random Forest Classification report	47
Table 1.51 Random Forest Conclusion	48
Table 1.52 LDA Classification report	50
Table 1.53 LDA Conclusion	51
Table 2.1 Dataset Information	56
Table 2.2 Sample of the Dataset	57
Table 2.3 Descriptive Statistics	57
Table 2.4 Sample of the Stock Returns	59
Table 2.5 Sample of the Stock Returns	59
Table 2.6 Stock Mean and Volatility	60

List Of Figures

Figure	Page No
fig 1.1 Outliers in Each variable	15
fig 1.2 Outliers after Treatment	16
fig 1.3 Missing Value heatmap	17
fig 1.4 Missing Values After Treatment	18
fig 1.5 Multicollinearity between Variables	24
fig 1.6 Target Variable Overview	25
fig 1.7 Univariate Analysis of Research_and_development_expense_rate	25
fig 1.8 Univariate Analysis of Interest_bearing_debt_interest_rate	25
fig 1.9 Univariate Analysis of Net_Value_Growth_Rate	26
fig 1.10 Univariate Analysis of Cash_Reinvestment_perc	26
fig 1.11 Univariate Analysis of Total_debt_to_Total_net_worth	26
fig 1.12 Univariate Analysis of Total_Asset_Turnover	27
fig 1.13 Univariate Analysis of Accounts_Receivable_Turnover	27
fig 1.14 Univariate Analysis of Fixed_Assets_Turnover_Frequency	27
fig 1.15 Univariate Analysis of Operating_profit_per_person	28
fig 1.16 Univariate Analysis of Allocation_rate_per_person	28
fig 1.17 Univariate Analysis of Retained_Earnings_to_Total_Assets	28
fig 1.18 Univariate Analysis of Total_income_to_Total_expense	29
fig 1.19 Univariate Analysis of Total_expense_to_Assets	29
fig 1.20 Univariate Analysis of Equity_to_Liability	29
fig 1.21 Pairplot	30
fig 1.22 Correlation Heatmap	30
fig 1.23 Default against Research and Development Expense Rate	31
fig 1.24 Default against Interest-bearing Debt Interest Rate:	31
fig 1.25 Default against Net Value Growth Rate	31

fig 1.26 Default against Cash Reinvestment Percentage	31
fig 1.27 Default against Total Debt to Total Net Worth	32
fig 1.28 Default against Total Asset Turnover	32
fig 1.29 Default against Accounts Receivable Turnover	32
fig 1.30 Default against Fixed Assets Turnover Frequency	32
fig 1.31 Default against Operating Profit per Person	33
fig 1.32 Default against Retained Earnings to Total Assets	33
fig 1.33 Default against Retained Earnings to Total Assets	33
fig 1.34 Default against Total Income to Total Expense	33
fig 1.35 Default against Total Expense to Assets	34
fig 1.36 Default against Equity to Liability	34
fig 1.37 Logistic Regression Explanation	35
fig 1.38 Confusion Matrix	35
fig 1.39 ROC - AUC	36
fig 1.40 Logistic Regression Confusion Matrix	38
fig 1.41 Logistic Regression ROC curve	39
fig 1.42 Logistic Regression Optimal Cutoff Confusion Matrix	40
fig 1.43 Logistic Regression Optimal Cutoff ROC curve	41
fig 1.44 Logistic Regression SMOTE Confusion Matrix	43
fig 1.45 Logistic Regression SMOTE ROC curve	43
fig 1.46 Logistic Regression SMOTE Optimal Cutoff Confusion Matrix	44
fig 1.47 Logistic Regression SMOTE Optimal ROC curve	45
fig 1.48 Random Forrest Explanation	46
fig 1.49 RandomForest Confusion Matrix	47
fig 1.50 RandomForest ROC curve	48
fig 1.51 RandomForest Feature Importance	48
fig 1.52 LDA Explaination	49
fig 1.53 LDA Confusion Matrix	50
fig 1.54 LDA ROC curve	51
fig 1.55 LDA Feature Importance	51
fig 1.56 All Model Comparison Metrics	52
fig 1.57 All Model Comparison Confusion Matrix	53
fig 1.58 All Model Comparison ROC curve	54
Fig 2.1 Stock Price vs Time Graph for Shree Cement	58
Fig 2.2 Stock Price vs Time Graph for Idea-Vodafone	58
Fig 2.3 Stock Returns & Price Trend	59
Fig 2.4 Stock Means vs Standard Deviation	61
Fig 2.5 Stock Means vs Standard Deviation with Stock Names	61

PART A: Company Credit Risk

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.

Data Dictionary:

Sl. 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_Growth_Rate	Continuous Net Profit Growth Rate: Net 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.

About the data

Sample of data

The sameple is Transposed for the visibility on a single view & exported in table.

index	0	1	2	3	4
Co_Code	16974	21214	14852	2439	23505
Co_Name	Hind.Cables	Tata Tele. Mah.	ABG Shipyard	GTL	Bharati Defence
Default	0	1	0	0	0
_Accounts_Receivable_Turnover	0.014003835	0.000306372	0.001044911	0.005411218	0.00081361
_Allocation_rate_per_person	0.13949366	0.022804742	0.012358345	0.009048551	0.002069073
_Average_Collection_Days	0.000451718	0.020644957	0.006048237	0.001169001	0.007775695
_CFO_to_Assets	0.576869319	0.55152318	0.463045085	0.577212195	0.594037614
_Cash_Flow_Per_Share	0.322557827	0.315519559	0.299851454	0.319834477	0.325104336
_Cash_Flow_to_Equity	0.310900603	0.314572393	0.314776836	0.316974404	0.317729099
_Cash_Flow_to_Liability	0.458072673	0.459282055	0.459700257	0.461860245	0.461237791
_Cash_Flow_to_Total_Assets	0.632665988	0.642967042	0.644485668	0.656831595	0.656549332
_Cash_Reinvestment_perc	0.369136768	0.37267613	0.348859953	0.379875936	0.38960918
_Cash_Turnover_Rate	5470000000	882000000	679000000	6020000000	5670000000
_Cash_flow_rate	0.462045075	0.460115917	0.449893058	0.462731184	0.463117016
_Cash_to_Current_Liability	0.000676239	0.000216355	0.00045779	0.002792591	0.002375847
_Cash_to_Total_Assets	0.025625562	0.004529484	0.008241543	0.053510183	0.082328432
_Continuous_Net_Profit_Growth_Rate	0.217590395	0.217359818	0.217573146	0.21766244	0.217589117
_Current_Asset_Turnover_Rate	0.000731942	0.000300807	0.000127161	0.000401223	0.000207952
_Current_Liability_to_Current_Assets	0.034913465	0.041652748	0.033560129	0.016527162	0.034497326
_Current_Ratio	0.008324467	0.006938512	0.008668928	0.017749711	0.008427431
_Degree_of_Financial_Leverage_DFL	0.0269302	0.026297476	0.027276372	0.026987988	0.027497703
_Equity_to_Liability	0.015337823	0.029444909	0.041718317	0.02695599	0.019900264
_Fixed_Assets_Turnover_Frequency	0.000304869	8850000000	0.00014903	0.001826942	0.000829878
_Fixed_Assets_to_Assets	0.094269738	0.351894542	0.463276072	0.0264327	0.103302506
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	0.565744293	0.560740561	0.566743977	0.565949567	0.567177433
_Interest_Expense_Ratio	0.631512924	0.628055438	0.631687922	0.632587567	0.632682429
_Interest_bearing_debt_interest_rate	0.000352035	0.000716072	0.00049605	0.000592059	0.000782078
_Inventory_Turnover_Rate_times	707000000	0.000278202	0.000169864	1340000000	0.000134133
_Inventory_to_Current_Liability	0.017944511	0.00127057	0.007011756	0.039872089	0.003341549
_Inventory_to_Working_Capital	0.278433848	0.277221033	0.277473023	0.277630169	0.277235295
_Liability_Assets_Flag	0	0	0	0	0
_Long_term_Liability_to_Current_Assets	0.00306419	0.004813246	0	0.004471604	0
_Long_term_fund_suitability_ratio_A	0.005766944	0.005230461	0.005139082	0.011960026	0.005822005
_Net_Income_Flag	1	1	1	1	1
_Net_Value_Growth_Rate	0.000441218	0.000403288	0.000451574	0.000448181	0.000454494
_Net_Worth_Turnover_Rate_times	0.02983871	0.018387097	0.02983871	0.028387097	0.052258065
_Net_profit_before_tax_to_Paid_in_capital	0.192858759	0.160681698	0.171548149	0.172158962	0.175597617
_No_credit_Interval	0.620927233	0.622513412	0.62374948	0.622962701	0.624418719
_Operating_Expense_Rate	8820000000	9380000000	3800000000	6440000000	3680000000

_Operating_Funds_to_Liability	0.342391123	0.337475735	0.306992828	0.343499741	0.345795846
_Operating_Profit_Growth_Rate	0.848021065	0.839645289	0.848196489	0.848390958	0.847986749
_Operating_profit_per_person	0.611689017	0.386626264	0.393263382	0.439779743	0.39276588
_Per_Share_Net_profit_before_tax_Yuan_	0.194471643	0.161632899	0.172554041	0.174738269	0.176545906
_Quick_Asset_Turnover_Rate	0.000142041	0.000298559	941000000	5310000000	0.000189202
_Quick_Assets_to_Current_Liability	0.001509387	0.006584496	0.006090393	0.002437173	0.007270916
_Quick_Assets_to_Total_Assets	0.176437639	0.402040088	0.318920592	0.137092399	0.739193107
_Quick_Ratio	0.000254527	0.004787352	0.005912136	0.001738019	0.003967001
_Realized_Sales_Gross_Profit_Growth_Rate	0.022074133	0.021902022	0.022186493	0.027637697	0.022072095
_Research_and_development_expense_rate	0	4230000000	815000000	0	0
_Retained_Earnings_to_Total_Assets	0.937630313	0.92625076	0.93315478	0.928037417	0.934421037
_Tax_rate_A	0.001417147	0	0	0.009312683	0.40024294
_Total_Asset_Growth_Rate	7500000000	6750000000	9680000000	7520000000	7120000000
_Total_Asset_Return_Growth_Rate_Ratio	0.263901905	0.263713857	0.264094614	0.26476599	0.2639664
_Total_Asset_Turnover	0.053973013	0.056971514	0.154422789	0.101949025	0.163418291
_Total_assets_to_GNP_price	0.028800595	0.006190921	0.001094543	0.003748761	0.006594729
_Total_debt_to_Total_net_worth	0.026006073	0.006812026	0.004104542	0.007846141	0.013670561
_Total_expense_to_Assets	0.007059242	0.015440696	0.009770697	0.013607109	0.01049254
_Total_income_to_Total_expense	0.002687011	0.002043933	0.002324166	0.002333806	0.002310273

Table 1.1 Sample of the dataset

- Messy column names have been fixed by adding uniformity throughout in the names of the columns (variables).

Information about the data

- This dataset contains information on **2,058 companies** and having **58 columns.**
- **Data Type**: It's mainly made up of **numerical data**. Only 1 Object datatype as Company Name.
- **Missing Values**: Some columns have missing values. For example, the "Cash_Flow_Per_Share" column has 1891 non-null values indicating missing. We'll have deeper look at this.
- **Target Column: "Default"** which indicates whether a company defaulted (1) or not (0).
- Overall, the dataset contains information about the financial performance and health of different companies.

```
<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
```

Table 1.2 Information of the dataset

Statistical Summary of the Dataset:

	count	mean	std	min	25%	50%	75%	max
Co_Code	2058.00	17572.11	21892.89	4.00	3674.00	6240.00	24280.75	72493.00
Operating_Expense_Rate	2058.00	2052388835.76	3252623690.29	0.00	0.00	0.00	4110000000.00	9980000000.00
Research_and_development_expense_rate	2058.00	1208634256.56	2144568158.08	0.00	0.00	0.00	1550000000.00	9980000000.00
Cash_flow_rate	2058.00	0.47	0.02	0.00	0.46	0.46	0.47	1.00
Interest_bearing_debt_interest_rate	2058.00	11130223.52	90425949.04	0.00	0.00	0.00	0.00	990000000.00
Tax_rate_A	2058.00	0.11	0.15	0.00	0.00	0.04	0.22	1.00
Cash_Flow_Per_Share	1891.00	0.32	0.02	0.17	0.31	0.32	0.33	0.46
Per_Share_Net_profit_before_tax_Yuan_	2058.00	0.18	0.03	0.00	0.17	0.18	0.19	0.79
Realized_Sales_Gross_Profit_Growth_Rate	2058.00	0.02	0.02	0.00	0.02	0.02	0.02	1.00
Operating_Profit_Growth_Rate	2058.00	0.85	0.00	0.74	0.85	0.85	0.85	1.00
Continuous_Net_Profit_Growth_Rate	2058.00	0.22	0.01	0.00	0.22	0.22	0.22	0.23
Total_Asset_Growth_Rate	2058.00	5287663257.05	2912614769.58	0.00	4315000000.00	6225000000.00	7220000000.00	9980000000.00
Net_Value_Growth_Rate	2058.00	5189504.37	207791797.86	0.00	0.00	0.00	0.00	9330000000.00
Total_Asset_Return_Growth_Rate_Ratio	2058.00	0.26	0.00	0.25	0.26	0.26	0.26	0.36
Cash_Reinvestment_perc	2058.00	0.38	0.03	0.03	0.37	0.38	0.39	1.00
Current_Ratio	2058.00	1336248.80	60619173.20	0.00	0.01	0.01	0.01	2750000000.00
Quick_Ratio	2058.00	27755102.05	444865390.47	0.00	0.00	0.01	0.01	9230000000.00
Interest_Expense_Ratio	2058.00	0.63	0.01	0.53	0.63	0.63	0.63	0.81
Total_debt_to_Total_net_worth	2037.00	10714285.73	269696017.59	0.00	0.00	0.01	0.01	9940000000.00
Long_term_fund_suitability_ratio_A	2058.00	0.01	0.03	0.00	0.01	0.01	0.01	1.00
Net_profit_before_tax_to_Paid_in_capital	2058.00	0.18	0.03	0.00	0.17	0.17	0.18	0.79
Total_Asset_Turnover	2058.00	0.13	0.10	0.00	0.06	0.10	0.17	0.92
Accounts_Receivable_Turnover	2058.00	41598639.46	504767266.59	0.00	0.00	0.00	0.00	9740000000.00
Average_Collection_Days	2058.00	26297862.01	410996733.83	0.00	0.00	0.01	0.01	8800000000.00
Inventory_Turnover_Rate_times	2058.00	2030227259.48	3077250265.27	0.00	0.00	19100000.00	3815000000.00	9990000000.00
Fixed_Assets_Turnover_Frequency	2058.00	1230897959.18	2649288936.44	0.00	0.00	0.00	0.01	9990000000.00
Net_Worth_Turnover_Rate_times	2058.00	0.04	0.04	0.01	0.02	0.03	0.04	1.00
Operating_profit_per_person	2058.00	0.40	0.05	0.00	0.39	0.40	0.40	1.00
Allocation_rate_per_person	2058.00	5725558.82	197949961.06	0.00	0.00	0.01	0.02	8280000000.00
Quick_Assets_to_Total_Assets	2058.00	0.34	0.21	0.00	0.17	0.31	0.48	0.99
Cash_to_Total_Assets	1962.00	0.08	0.10	0.00	0.02	0.05	0.10	0.93
Quick_Assets_to_Current_Liability	2058.00	11904761.91	312292270.93	0.00	0.00	0.01	0.01	8820000000.00
Cash_to_Current_Liability	2058.00	92825072.90	785189881.95	0.00	0.00	0.00	0.01	9170000000.00
Operating_Funds_to_Liability	2058.00	0.35	0.04	0.03	0.34	0.35	0.35	1.00
Inventory_to_Working_Capital	2058.00	0.28	0.02	0.00	0.28	0.28	0.28	1.00
Inventory_to_Current_Liability	2058.00	57863459.68	627879536.23	0.00	0.00	0.01	0.01	9600000000.00
Long_term_Liability_to_Current_Assets	2058.00	73401069.01	669352618.01	0.00	0.00	0.00	0.01	9310000000.00
Retained_Earnings_to_Total_Assets	2058.00	0.93	0.03	0.00	0.93	0.94	0.94	0.97
Total_income_to_Total_expense	2058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Total_expense_to_Assets	2058.00	0.03	0.04	0.00	0.01	0.02	0.04	1.00
Current_Asset_Turnover_Rate	2058.00	1273303377.07	2839740987.63	0.00	0.00	0.00	0.00	9990000000.00
Quick_Asset_Turnover_Rate	2058.00	2571767687.08	3453544121.67	0.00	0.00	0.00	5790000000.00	10000000000.00
Cash_Turnover_Rate	2058.00	2653695544.22	2821244732.19	0.00	0.00	1730000000.00	4550000000.00	9990000000.00
Fixed_Assets_to_Assets	2058.00	4042760.23	183400553.09	0.00	0.10	0.21	0.42	8320000000.00
Cash_Flow_to_Total_Assets	2058.00	0.64	0.05	0.00	0.63	0.64	0.65	1.00
Cash_Flow_to_Liability	2058.00	0.46	0.03	0.03	0.46	0.46	0.46	0.91
CFO_to_Assets	2058.00	0.58	0.06	0.00	0.55	0.58	0.61	0.98
Cash_Flow_to_Equity	2058.00	0.31	0.01	0.00	0.31	0.31	0.32	0.57
Current_Liability_to_Current_Assets	2044.00	0.04	0.05	0.00	0.02	0.03	0.04	1.00
Liability_Assets_Flag	2058.00	0.00	0.06	0.00	0.00	0.00	0.00	1.00
Total_assets_to_GNP_price	2058.00	27793974.74	471771444.55	0.00	0.00	0.00	0.01	9820000000.00
No_credit_Interval	2058.00	0.62	0.01	0.41	0.62	0.62	0.62	0.96
Degree_of_Financial_Leverage_DFL	2058.00	0.03	0.01	0.01	0.03	0.03	0.03	0.46
Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058.00	0.57	0.01	0.17	0.57	0.57	0.57	0.67
Net_Income_Flag	2058.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Equity_to_Liability	2058.00	0.04	0.06	0.00	0.02	0.03	0.04	1.00
Default	2058.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00

Table 1.3 Descriptive Statistics of the dataset

Descriptive Statistics

- **Company information:** The total count of companies in the dataset are 2058.
- **Operating Expenses:** The operating expense rates vary significantly, with values ranging from 0 to approximately 9.98 billion. This column exhibits a very high degree of variability, including extreme outliers towards right.
- **Research and Development Expenses:** Similar to operating expenses, research and development expenses show a wide range, spanning from 0 to approximately 9.98 billion. This column also contains significant variability and potential outliers.
- **Cash Flow and Profitability:** The dataset includes cash flow-related metrics such as cash flow rate and cash flow per share, with values around 0.47 and 0.32, respectively. Per share net profit before tax averages at around 0.18. These metrics appear to have a relatively narrower distribution compared to some other columns.
- **Tax and Financial Ratios:** The tax rate (A) ranges from 0 to 1.0, with a mean of approximately 0.11. Several financial ratios, such as realized sales gross profit growth rate and operating profit growth rate, show moderate variability with means around 0.02 and 0.85, respectively. Other financial ratios also demonstrate variability within their respective ranges.

Duplicate records:

- There are no duplicate records in the dataset.

Outlier Treatment

Number of Observations Beyond Upper & Lower Limit for Each Column:

Operating_Expense_Rate	0
Research_and_development_expense_rate	264
Cash_flow_rate	206
Interest_bearing_debt_interest_rate	94
Tax_rate_A	42
Cash_Flow_Per_Share	146
Per_Share_Net_profit_before_tax_Yuan_	186
Realized_Sales_Gross_Profit_Growth_Rate	283
Operating_Profit_Growth_Rate	317
Continuous_Net_Profit_Growth_Rate	340
Total_Asset_Growth_Rate	0
Net_Value_Growth_Rate	304
Total_Asset_Return_Growth_Rate_Ratio	226
Cash_Reinvestment_perc	220
Current_Ratio	193
Quick_Ratio	190
Interest_Expense_Ratio	328
Total_debt_to_Total_net_worth	105
Long_term_fund_suitability_ratio_A	234
Net_profit_before_tax_to_Paid_in_capital	173
Total_Asset_Turnover	101
Accounts_Receivable_Turnover	281
Average_Collection_Days	77
Inventory_Turnover_Rate_times	29
Fixed_Assets_Turnover_Frequency	501
Net_Worth_Turnover_Rate_times	165
Operating_profit_per_person	357
Allocation_rate_per_person	200
Quick_Assets_to_Total_Assets	4
Cash_to_Total_Assets	163
Quick_Assets_to_Current_Liability	185
Cash_to_Current_Liability	253
Operating_Funds_to_Liability	219
Inventory_to_Working_Capital	247
Inventory_to_Current_Liability	129
Long_term_Liability_to_Current_Assets	213
Retained_Earnings_to_Total_Assets	208
Total_income_to_Total_expense	136
Total_expense_to_Assets	168
Current_Asset_Turnover_Rate	464
Quick_Asset_Turnover_Rate	0
Cash_Turnover_Rate	0
Fixed_Assets_to_Assets	10
Cash_Flow_to_Total_Assets	317
Cash_Flow_to_Liability	407
CFO_to_Assets	110
Cash_Flow_to_Equity	306
Current_Liability_to_Current_Assets	121
Liability_Assets_Flag	7
Total_assets_to_GNP_price	235
No_credit_Interval	396
Degree_of_Financial_Leverage_DFL	438
Interest_Coverage_Ratio_Interest_expense_to_EBIT	376
Net_Income_Flag	0
Equity_to_Liability	190
dtype: int64	

Table 1.4 Outliers in Each variable

Outliers Visualization

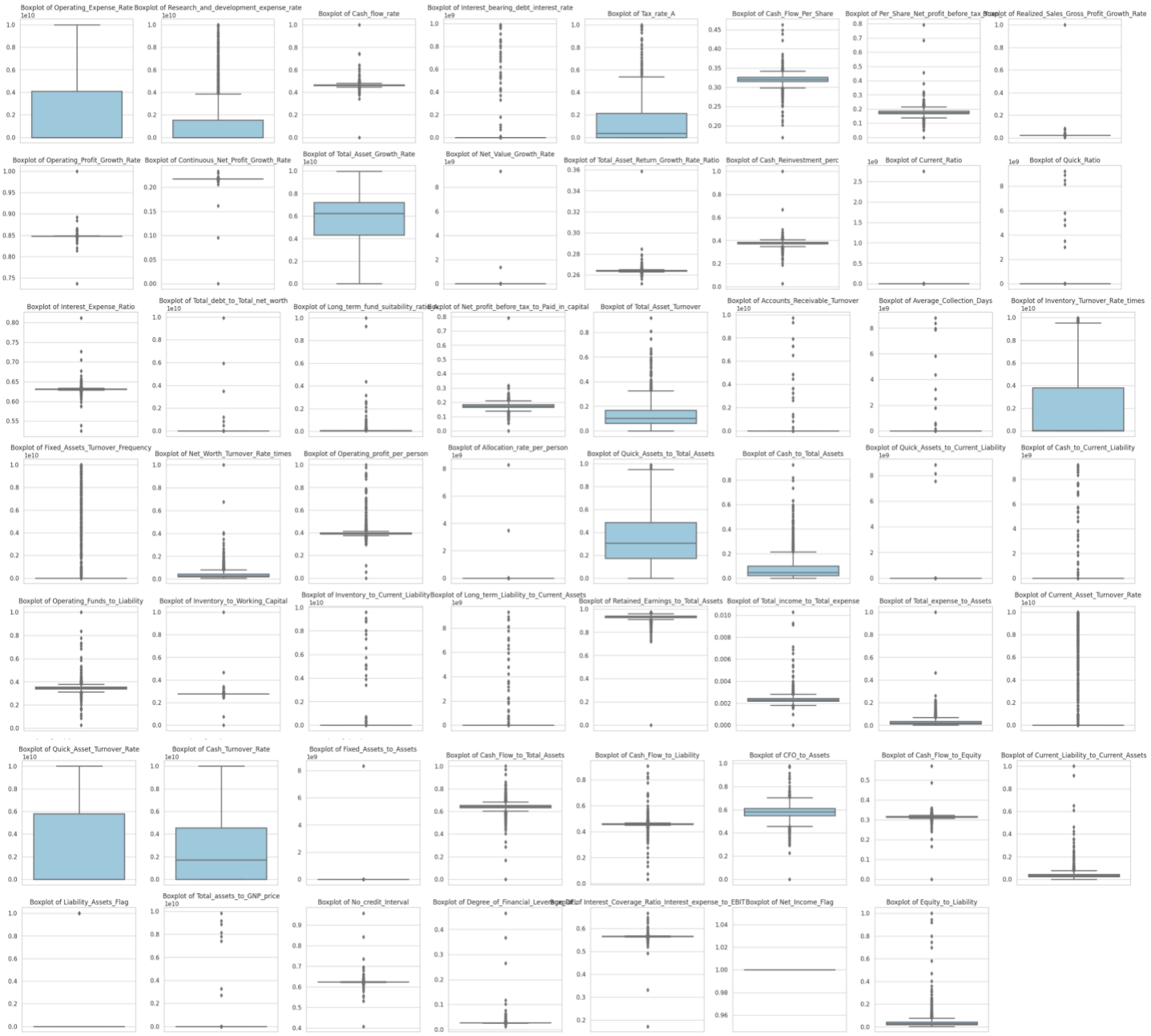


fig 1.1 Outliers in Each variable

Treatment Approach:

- The Outliers are Treated by capping them on **Upper Limit** and **Lower limit** for each column.

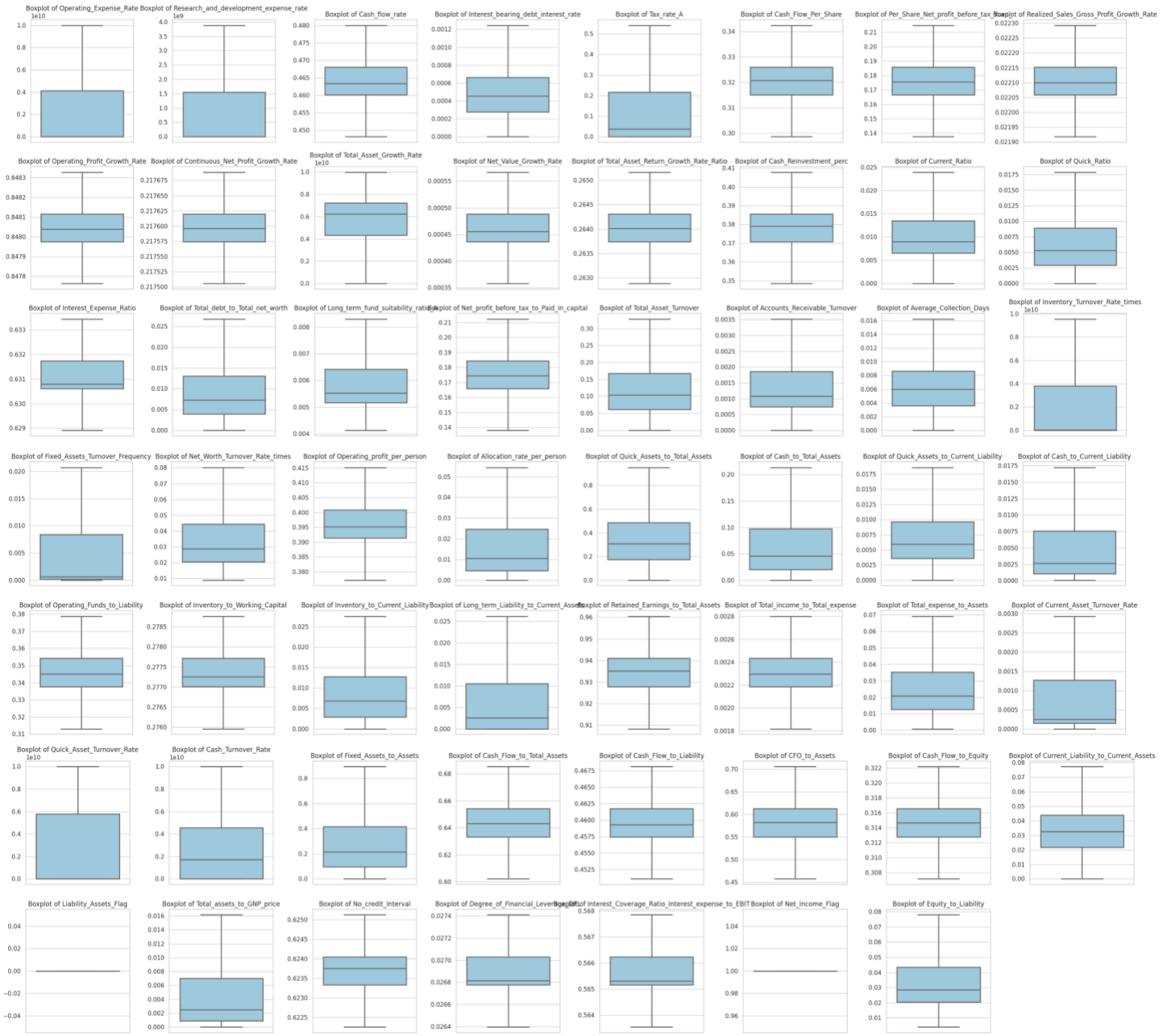


fig 1.2 Outliers after Treatment

Missing Values

- **Overall**: The total number of missing values are **298**, which is **0.25%** of total data.

Column Wise:

- **Cash_Flow_Per_Share** has **167** missing values, which is **8.11%** of the column.
- **Total_debt_to_Total_net_worth** has **21** missing values, which is **1.02%** of the column.
- **Cash_to_Total_Assets** has **96** missing values, which is **4.66%** of the column.
- **Current_Liability_to_Current_Assets** has **14** missing values, which is **0.68%** of the column.

Missing Value visualization:-

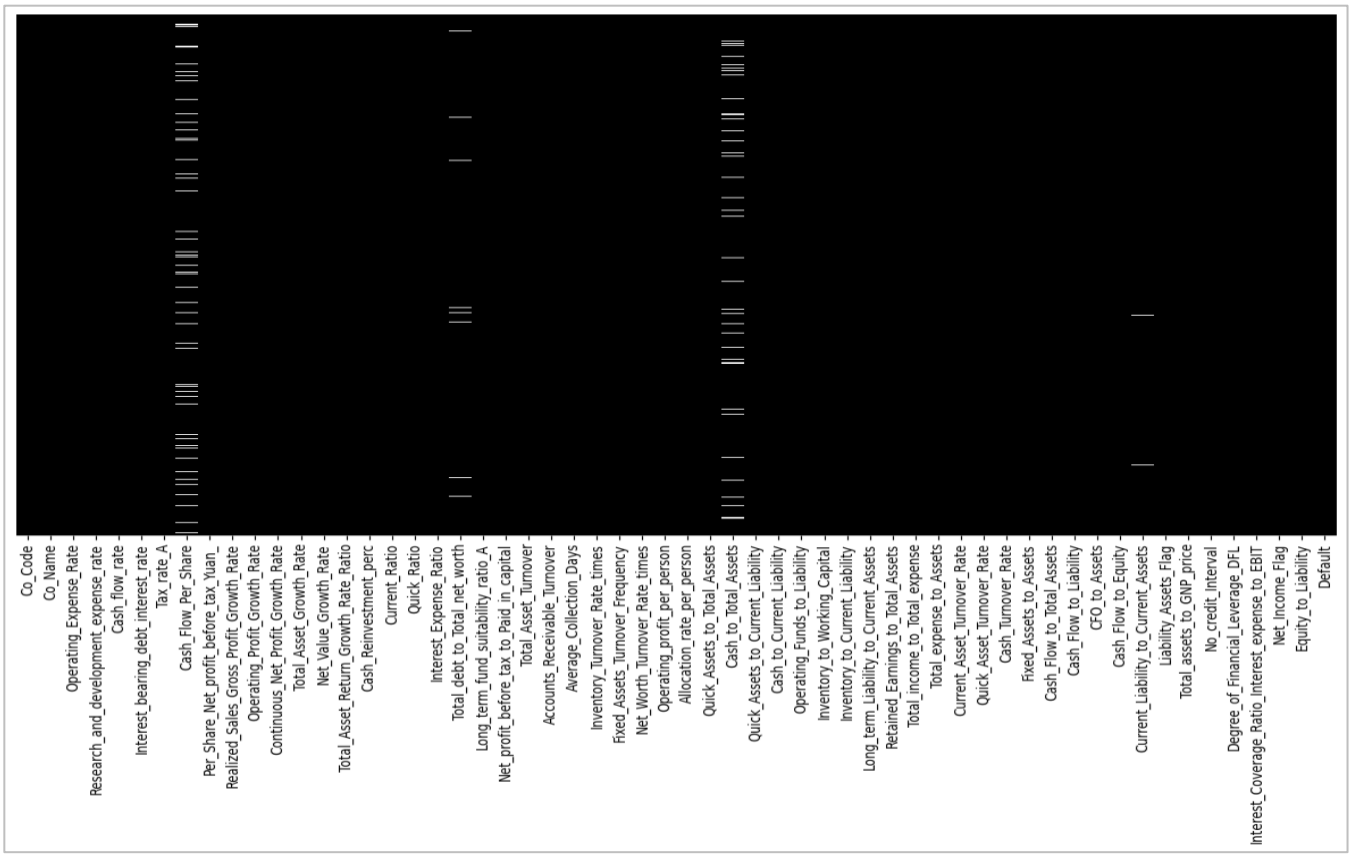


fig 1.3 Missing Value heatmap

Missing Value Treatment:

The missing values are Treated using KNN imputer post outlier Treatment.

- Scaling was done on the predictors to make them on same scale
- KNN imputer imputed the missing values using the mean value from the 'n' nearest neighbours found in the training set (here, n = 5).
- There are no missing values remaining after imputation using KNN

```
Missing Values in the dataset after treatment : 0
```

fig 1.4 Missing Values After Treatment

Univariate & Bivariate analysis with proper interpretation.

Before we do Univariate & Bivariate analysis, Let's do feature selection for the most important features.

Feature Scaling

The features are scaled as per the standard scaler(Z score) technique as the standardisation helps in easy & correct interpretation of the coefficients of Logistic Regression & LDA. A snippet of the dataset after scaling is

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share	Per_Share_Net_profit_before_tax_Yuan_	Realized_Sa
0	2.08	-0.65	-0.32	-0.45	-0.81	0.22	1.00	
1	2.25	2.04	-0.58	0.78	-0.82	-0.49	-0.82	
2	0.54	-0.08	-1.95	0.04	-0.82	-2.07	-0.21	
3	1.35	-0.65	-0.23	0.36	-0.75	-0.06	-0.09	
4	0.50	-0.65	-0.17	1.01	2.15	0.48	0.01	

Table 1.5 Sample of scaled variables

Z-score is a technique of scaling that tells us the number of standard deviations a data is away from the mean. After Z-score scaling the features have a mean = 0 and standard deviation = 1

Multicollinearity

fig 1.5 Multicollinearity between Variables

- Multicollinearity is evident as many features exhibit high correlations with each other, potentially complicating the interpretation of individual feature impacts on the target variable and the stability of model coefficients
- Features such as , such as "Realized_Sales_Gross_Profit_Growth_Rate" and "Operating_Profit_Growth_Rate," exibits strong positive correlation, indicating that they tend to move together positively in financial performance.
- Some pairs of features, like "Total_assets_to_GNP_price" and "Total_expense_to_Assets," show negative correlations, implying an inverse relationship between them.
- **Variables with a high multicollinearity will be removed, i.e., variables with a VIF > 5**

Eliminating variables that are highly correlated with one another using VIF:

- A **variance inflation factor (VIF)** provides a measure of multicollinearity among the independent variables in a multiple regression model.
- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

VIF > 5 is not suitable as it is mostly compensated by other IVs. Hence we use VIF to clean the data of redundant variables. Variables VIF>5 are shown as below without any elimination yet:

	variables	VIF
6	Per_Share_Net_profit_before_tax_Yuan_	100.74
19	Net_profit_before_tax_to_Paid_in_capital	99.88
43	Cash_Flow_to_Total_Assets	44.51
45	CFO_to_Assets	30.62
32	Operating_Funds_to_Liability	21.86
30	Quick_Assets_to_Current_Liability	19.85
44	Cash_Flow_to_Liability	17.85
2	Cash_flow_rate	16.53
25	Net_Worth_Turnover_Rate_times	15.51
46	Cash_Flow_to_Equity	15.15
14	Current_Ratio	14.78
20	Total_Asset_Turnover	14.51
13	Cash_Reinvestment_perc	13.92
15	Quick_Ratio	12.59
5	Cash_Flow_Per_Share	9.70
47	Current_Liability_to_Current_Assets	8.37
54	Equity_to_Liability	6.35
28	Quick_Assets_to_Total_Assets	6.19
52	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.10
36	Retained_Earnings_to_Total_Assets	5.27
37	Total_income_to_Total_expense	5.25
17	Total_debt_to_Total_net_worth	5.02

Table 1.6 Variables with highest VIF (>5.0)

Let's eliminate the variables one by one. We will iterate through the values one by one. Starting the first iteration by dropping variable with highest VIF & show updated VIF values (Top 5)

Iteration 1: Highest VIF is 100.74 for **Per_Share_Net_profit_before_tax_Yuan_**, dropping it.

```
Updated VIF values:
                                    variables   VIF
42                    Cash_Flow_to_Total_Assets 44.50
44                                 CFO_to_Assets 30.33
31                Operating_Funds_to_Liability 21.83
29              Quick_Assets_to_Current_Liability 19.84
43                    Cash_Flow_to_Liability 17.85
```

Table 1.7 Variables with highest VIF - Top 5

Iteration 2: Highest VIF is 44.5 for **Cash_Flow_to_Total_Assets**, dropping it.

Updated VIF values:

	variables	VIF
43	CFO_to_Assets	30.14
31	Operating_Funds_to_Liability	21.80
29	Quick_Assets_to_Current_Liability	19.81
2	Cash_flow_rate	16.50
24	Net_Worth_Turnover_Rate_times	15.51

Table 1.8 Variables with highest VIF - Top 5

Iteration 3: Highest VIF is 30.14 for **CFO_to_Assets**, dropping it.

Updated VIF values:

	variables	VIF
29	Quick_Assets_to_Current_Liability	19.81
31	Operating_Funds_to_Liability	18.87
2	Cash_flow_rate	16.12
24	Net_Worth_Turnover_Rate_times	15.49
13	Current_Ratio	14.76

Table 1.9 Variables with highest VIF - Top 5

Iteration 4: Highest VIF is 19.81 for **Quick_Assets_to_Current_Liability**, dropping it.

Updated VIF values:

	variables	VIF
30	Operating_Funds_to_Liability	18.71
2	Cash_flow_rate	15.90
24	Net_Worth_Turnover_Rate_times	15.47
19	Total_Asset_Turnover	14.40
13	Current_Ratio	13.06

Table 1.10 Variables with highest VIF - Top 5

Iteration 5: Highest VIF is 18.71 for **Operating_Funds_to_Liability**, dropping it.

Updated VIF values:

	variables	VIF
24	Net_Worth_Turnover_Rate_times	15.31
19	Total_Asset_Turnover	14.25
13	Current_Ratio	12.91
42	Current_Liability_to_Current_Assets	8.26
5	Cash_Flow_Per_Share	8.24

Table 1.11 Variables with highest VIF - Top 5

Iteration 6: Highest VIF is 15.31 for **Net_Worth_Turnover_Rate_times**, dropping it.

Updated VIF values:

	variables	VIF
13	Current_Ratio	12.87
41	Current_Liability_to_Current_Assets	8.26
5	Cash_Flow_Per_Share	8.24
18	Net_profit_before_tax_to_Paid_in_capital	8.11
12	Cash_Reinvestment_perc	7.80

Table 1.12 Variables with highest VIF - Top 5

Iteration 7: Highest VIF is 12.87 for **Current_Ratio**, dropping it.

Updated VIF values:

	variables	VIF
5	Cash_Flow_Per_Share	8.23
17	Net_profit_before_tax_to_Paid_in_capital	8.10
12	Cash_Reinvestment_perc	7.80
13	Quick_Ratio	6.32
45	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.08

Table 1.13 Variables with highest VIF - Top 5

Iteration 8: Highest VIF is 8.23 for **Cash_Flow_Per_Share**, dropping it.

Updated VIF values:

	variables	VIF
16	Net_profit_before_tax_to_Paid_in_capital	7.86
12	Quick_Ratio	6.28
44	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.07
24	Quick_Assets_to_Total_Assets	5.45
39	Current_Liability_to_Current_Assets	5.19

Table 1.14 Variables with highest VIF - Top 5

Iteration 9: Highest VIF is 7.86 for **Net_profit_before_tax_to_Paid_in_capital**, dropping it.

Updated VIF values:

	variables	VIF
12	Quick_Ratio	6.27
43	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.04
23	Quick_Assets_to_Total_Assets	5.45
38	Current_Liability_to_Current_Assets	5.18
37	Cash_Flow_to_Equity	5.11

Table 1.15 Variables with highest VIF - Top 5

Iteration 10: Highest VIF is 6.27 for **Quick_Ratio**, dropping it.

Updated VIF values:

	variables	VIF
42	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.04
36	Cash_Flow_to_Equity	5.11
35	Cash_Flow_to_Liability	4.92
22	Quick_Assets_to_Total_Assets	4.92
34	Fixed_Assets_to_Assets	4.61

Table 1.16 Variables with highest VIF - Top 5

Iteration 11: Highest VIF is 6.04 for **Interest_Coverage_Ratio_Interest_expense_to_EBIT**, dropping it.

Updated VIF values:

	variables	VIF
36	Cash_Flow_to_Equity	5.10
35	Cash_Flow_to_Liability	4.92
22	Quick_Assets_to_Total_Assets	4.91
34	Fixed_Assets_to_Assets	4.61
29	Total_income_to_Total_expense	4.33

Table 1.17 Variables with highest VIF - Top 5

Iteration 12: Highest VIF is 5.1 for **Cash_Flow_to_Equity,** dropping it.

Final VIF values:

	variables	VIF
22	Quick_Assets_to_Total_Assets	4.90
34	Fixed_Assets_to_Assets	4.61
29	Total_income_to_Total_expense	4.33
36	Current_Liability_to_Current_Assets	4.14
42	Equity_to_Liability	4.14
24	Cash_to_Current_Liability	3.82
23	Cash_to_Total_Assets	3.80
28	Retained_Earnings_to_Total_Assets	3.67
6	Operating_Profit_Growth_Rate	3.61
2	Cash_flow_rate	3.54
15	Total_Asset_Turnover	3.48
7	Continuous_Net_Profit_Growth_Rate	3.43
13	Total_debt_to_Total_net_worth	3.21
10	Total_Asset_Return_Growth_Rate_Ratio	3.00
20	Operating_profit_per_person	2.94
14	Long_term_fund_suitability_ratio_A	2.85
5	Realized_Sales_Gross_Profit_Growth_Rate	2.74
21	Allocation_rate_per_person	2.67
16	Accounts_Receivable_Turnover	2.60
11	Cash_Reinvestment_perc	2.58
17	Average_Collection_Days	2.52
40	Degree_of_Financial_Leverage_DFL	2.40
12	Interest_Expense_Ratio	2.39
26	Inventory_to_Current_Liability	2.37
9	Net_Value_Growth_Rate	2.36
30	Total_expense_to_Assets	2.25
19	Fixed_Assets_Turnover_Frequency	1.90
27	Long_term_Liability_to_Current_Assets	1.73
39	No_credit_Interval	1.70
38	Total_assets_to_GNP_price	1.70
31	Current_Asset_Turnover_Rate	1.63
25	Inventory_to_Working_Capital	1.52
4	Tax_rate_A	1.49
32	Quick_Asset_Turnover_Rate	1.41
35	Cash_Flow_to_Liability	1.41
0	Operating_Expense_Rate	1.31
18	Inventory_Turnover_Rate_times	1.22
1	Research_and_development_expense_rate	1.19
8	Total_Asset_Growth_Rate	1.15
3	Interest_bearing_debt_interest_rate	1.11
33	Cash_Turnover_Rate	1.10
37	Liability_Assets_Flag	NaN
41	Net_Income_Flag	NaN

Table 1.18 Final Variables after removing VIF > 5

We can see that we do not have any variable now having VIF>5. We have dropped 12 Variables & left with 43.

We see there are 2 variables **Liability_Assets_Flag', 'Net_Income_Flag'** having only one type of values which are not significant, hence, we chose to drop them as well.

After dropping all columns with **VIF > 5 & 2 constant value variables,** we are left with **41 Variables.**

Recursive Feature Elimination:

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

- Once the feature importance has been determined, it then removes the less important features one by one in each iteration.
- We use the **Logistic Regression algorithm** of sci-kit learn to determine the most important features.
- Since there are still a lot of variables (41) remaining in the dataset, we use recursive feature elimination to select the most important features. As per the thumb rule, we select 1/3rd of the features (14) that will be critical in the model development.

	Feature	Rank
0	Operating_Expense_Rate	10
1	Research_and_development_expense_rate	1
2	Cash_flow_rate	27
3	Interest_bearing_debt_interest_rate	1
4	Tax_rate_A	13
5	Realized_Sales_Gross_Profit_Growth_Rate	14
6	Operating_Profit_Growth_Rate	15
7	Continuous_Net_Profit_Growth_Rate	4
8	Total_Asset_Growth_Rate	9
9	Net_Value_Growth_Rate	1
10	Total_Asset_Return_Growth_Rate_Ratio	3
11	Cash_Reinvestment_perc	1
12	Interest_Expense_Ratio	24
13	Total_debt_to_Total_net_worth	1
14	Long_term_fund_suitability_ratio_A	7
15	Total_Asset_Turnover	1
16	Accounts_Receivable_Turnover	1
17	Average_Collection_Days	20
18	Inventory_Turnover_Rate_times	17
19	Fixed_Assets_Turnover_Frequency	1
20	Operating_profit_per_person	1
21	Allocation_rate_per_person	1
22	Quick_Assets_to_Total_Assets	11
23	Cash_to_Total_Assets	25
24	Cash_to_Current_Liability	12
25	Inventory_to_Working_Capital	19
26	Inventory_to_Current_Liability	18
27	Long_term_Liability_to_Current_Assets	2
28	Retained_Earnings_to_Total_Assets	1
29	Total_income_to_Total_expense	1
30	Total_expense_to_Assets	1
31	Current_Asset_Turnover_Rate	22
32	Quick_Asset_Turnover_Rate	28
33	Cash_Turnover_Rate	5
34	Fixed_Assets_to_Assets	23
35	Cash_Flow_to_Liability	8
36	Current_Liability_to_Current_Assets	6
37	Total_assets_to_GNP_price	16
38	No_credit_Interval	26
39	Degree_of_Financial_Leverage_DFL	21
40	Equity_to_Liability	1

Table 1.19 RFE Rank of all Variables

The ranks of the **most important features are 1 and** the rest follow it. Below given is the list of the **14 most important features** that will be used in model building

	Feature	Rank
1	Research_and_development_expense_rate	1
3	Interest_bearing_debt_interest_rate	1
9	Net_Value_Growth_Rate	1
11	Cash_Reinvestment_perc	1
13	Total_debt_to_Total_net_worth	1
15	Total_Asset_Turnover	1
16	Accounts_Receivable_Turnover	1
19	Fixed_Assets_Turnover_Frequency	1
20	Operating_profit_per_person	1
21	Allocation_rate_per_person	1
28	Retained_Earnings_to_Total_Assets	1
29	Total_income_to_Total_expense	1
30	Total_expense_to_Assets	1
40	Equity_to_Liability	1

Table 1.20 Selected Variables with Rank 1

Univariate Analysis

We will now take a look at only the most important variables for univariate & bivariate analysis.

Note: The Univariate & bivariate Analysis is done on the original data, not the treated

Target Variable:

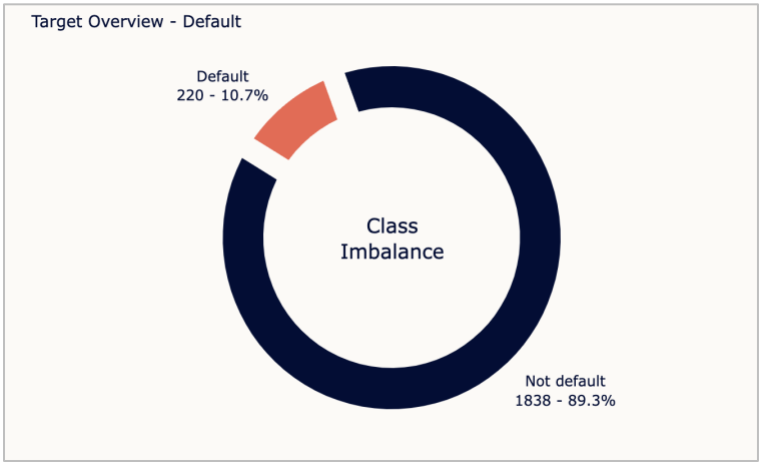


fig 1.6 Target Variable Overview

- The default class of the "Default " variable is imbalanced as observed earlier as well with 10.7% data in the underrepresented class.
- We will try to oversample the default class using SMOTE to see if it improves our model performance.

Distribution of each Variable

Research_and_development_expense_rate

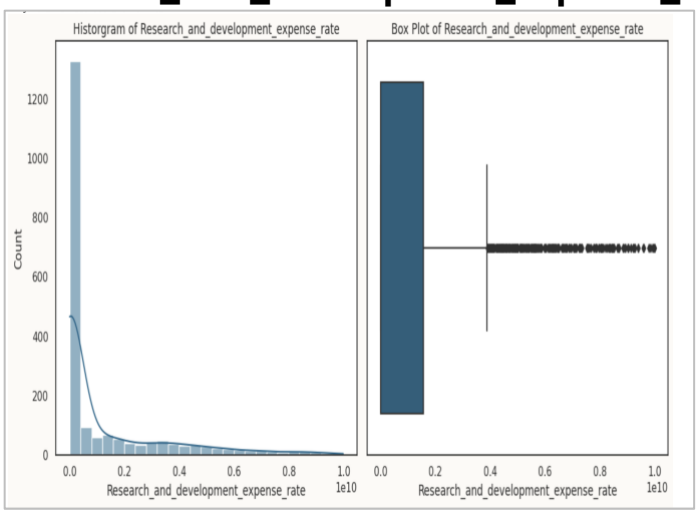


fig 1.7 Univariate Analysis of Research_and_development_expense_rate

```
Description of Research_and_development_expense_rate
------------------------------------------------------------
count          2058.00
mean     1208634256.56
std      2144568158.08
min               0.00
25%               0.00
50%               0.00
75%      1550000000.00
max      9980000000.00
Name: Research_and_development_expense_rate, dtype: float64
Skewness: 1.99
Kurtosis: 3.30
```

Table 1.21 Descriptive Statistics of the Variable

- The mean value is 1.21Bil. but it has a wide range from 0 to 9.98Bil, indicating a high variation in this feature.
- The data is positively skewed (1.99) and has positive kurtosis 3.30), suggesting that it is not normally distributed and has a heavy right tail.

Interest_bearing_debt_interest_rate

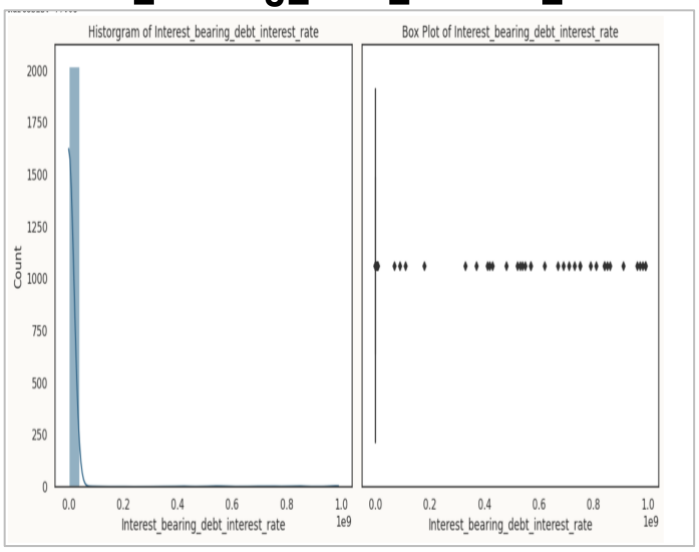


fig 1.8 Univariate Analysis of Interest_bearing_debt_interest_rate

```
Description of Interest_bearing_debt_interest_rate
------------------------------------------------------------
count        2058.00
mean     11130223.52
std      90425949.04
min              0.00
25%              0.00
50%              0.00
75%              0.00
max      990000000.00
Name: Interest_bearing_debt_interest_rate, dtype: float64
Skewness: 8.67
Kurtosis: 77.03
```

Table 1.22 Descriptive Statistics of the Variable

- The mean value is 11130223.52, but it has a wide range from 0 to 990000000.00, indicating significant variation.
- This feature is highly positively skewed (Skewness: 8.67) and has extremely high kurtosis (Kurtosis: 77.03), indicating a non-normal distribution with a long right tail.

Net_Value_Growth_Rate

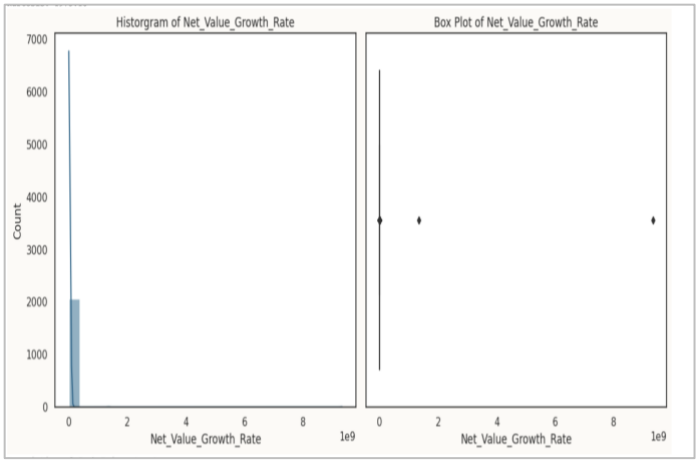


fig 1.9 Univariate Analysis of Net_Value_Growth_Rate

```
Description of Net_Value_Growth_Rate
------------------------------------------------
count          2058.00
mean        5189504.37
std       207791797.86
min               0.00
25%               0.00
50%               0.00
75%               0.00
max      9330000000.00
Name: Net_Value_Growth_Rate, dtype: float64
Skewness: 44.11
Kurtosis: 1975.18
```

Table 1.23 Descriptive Statistics of the Variable

- The mean value is 5189504.37, a wide range from 0 to 9.33B, indicating a large variation.
- This feature is extremely positively skewed (Skewness: 44.11) and has very high kurtosis (Kurtosis: 1975.18), indicating a highly non-normal distribution with a heavy right tail.

Cash_Reinvestment_perc

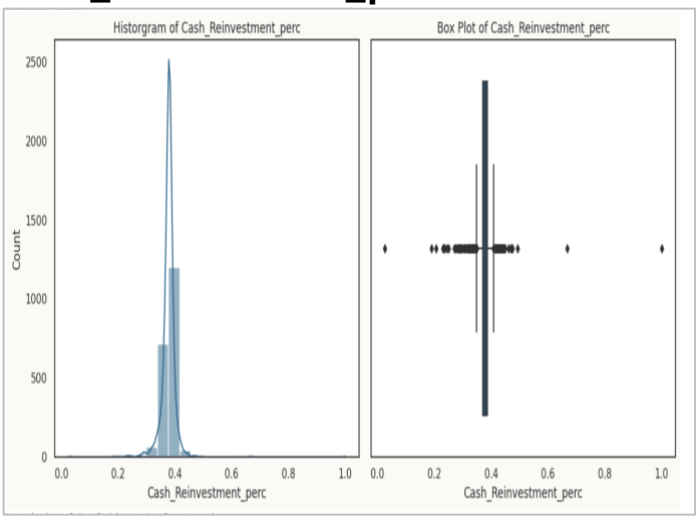


fig 1.10 Univariate Analysis of Cash_Reinvestment_perc

```
Description of Cash_Reinvestment_perc
-----------------------------------------------
count    2058.00
mean        0.38
std         0.03
min         0.03
25%         0.37
50%         0.38
75%         0.39
max         1.00
Name: Cash_Reinvestment_perc, dtype: float64
Skewness: 4.42
Kurtosis: 152.75
```

Table 1.24 Descriptive Statistics of the Variable

- The data has a relatively low standard deviation (0.03) compared to the mean (0.38), suggesting low variability.
- The skewness (Skewness: 4.42) and kurtosis (Kurtosis: 152.75) values indicate that this feature is positively skewed and has a heavy right tail.

Total_debt_to_Total_net_worth

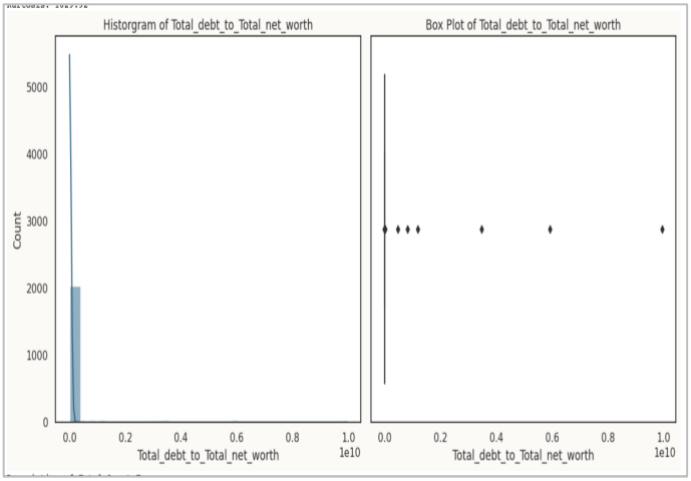


fig 1.11 Univariate Analysis of Total_debt_to_Total_net_worth

```
Description of Total_debt_to_Total_net_worth
-----------------------------------------------------
count          2037.00
mean       10714285.73
std       269696017.59
min               0.00
25%               0.00
50%               0.01
75%               0.01
max      9940000000.00
Name: Total_debt_to_Total_net_worth, dtype: float64
Skewness: 30.83
Kurtosis: 1029.92
```

Table 1.25 Descriptive Statistics of the Variable

- The mean value is 10714285.73, with a wide range from 0 to 9940000000.00.
- The feature is highly positively skewed (Skewness: 30.83) and has extremely high kurtosis (Kurtosis: 1029.92), indicating a non-normal distribution with a heavy right tail.

Total_Asset_Turnover

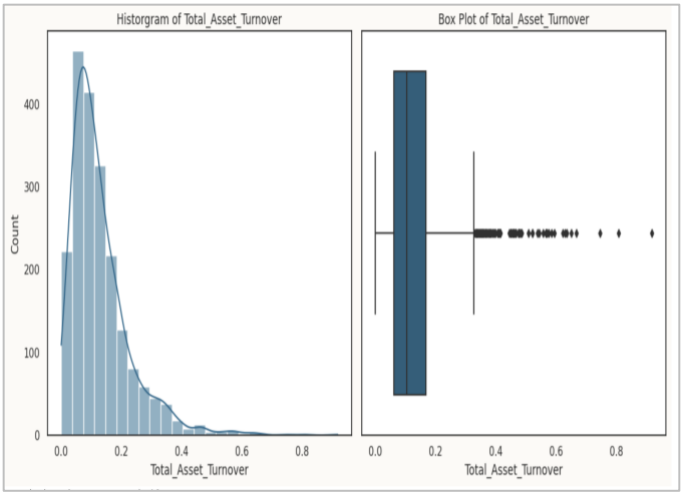


fig 1.12 Univariate Analysis of Total_Asset_Turnover

```
Description of Total_Asset_Turnover
-----------------------------------------------
count    2058.00
mean        0.13
std         0.10
min         0.00
25%         0.06
50%         0.10
75%         0.17
max         0.92
Name: Total_Asset_Turnover, dtype: float64
Skewness: 2.04
Kurtosis: 6.79
```

Table 1.26 Descriptive Statistics of the Variable

- The mean value is 0.13, indicating the average asset turnover rate.
- The data is moderately positively skewed (Skewness: 2.04) and has moderate kurtosis (Kurtosis: 6.79), suggesting a somewhat non-normal distribution.

Accounts_Receivable_Turnover

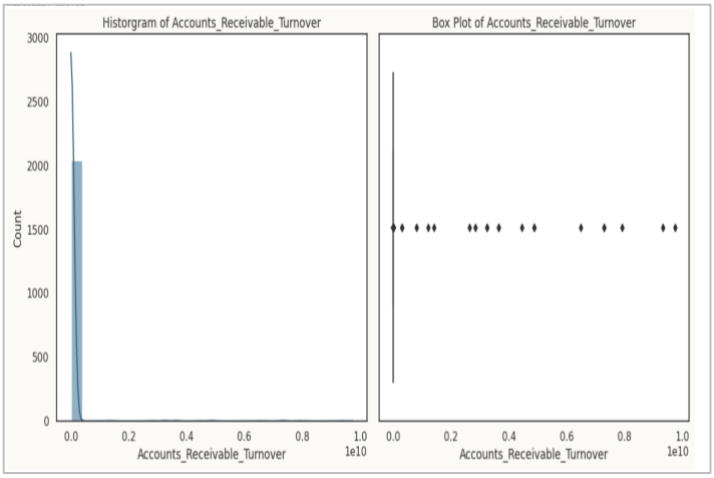


fig 1.13 Univariate Analysis of Accounts_Receivable_Turnover

```
Description of Accounts_Receivable_Turnover
-------------------------------------------------------------
count          2058.00
mean       41598639.46
std       504767266.59
min               0.00
25%               0.00
50%               0.00
75%               0.00
max      9740000000.00
Name: Accounts_Receivable_Turnover, dtype: float64
Skewness: 14.19
Kurtosis: 218.82
```

Table 1.27 Descriptive Statistics of the Variable

- The mean value is 41598639.46, but it has a wide range from 0 to 9740000000.00.
- This feature is highly positively skewed (Skewness: 14.19) and has extremely high kurtosis (Kurtosis: 218.82), indicating a non-normal distribution with a heavy right tail.

Fixed_Assets_Turnover_Frequency

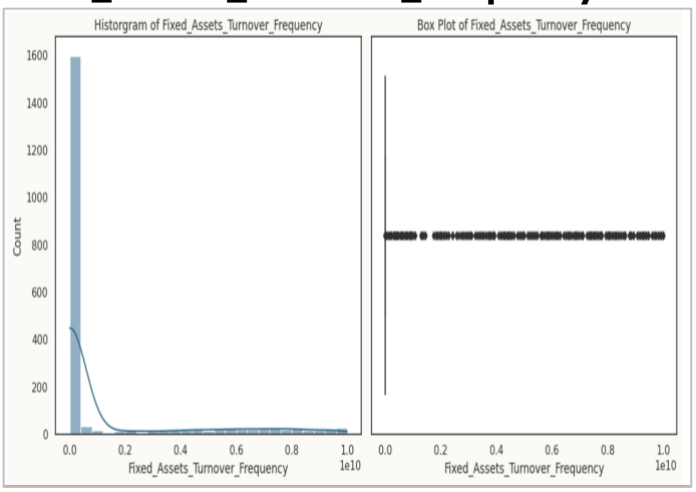


fig 1.14 Univariate Analysis of Fixed_Assets_Turnover_Frequency

```
Description of Fixed_Assets_Turnover_Frequency
-------------------------------------------------------------
count          2058.00
mean     1230897959.18
std      2649288936.44
min               0.00
25%               0.00
50%               0.00
75%               0.01
max      9990000000.00
Name: Fixed_Assets_Turnover_Frequency, dtype: float64
Skewness: 2.01
Kurtosis: 2.61
```

Table 1.28 Descriptive Statistics of the Variable

- The mean value is 1230897959.18, but it has a wide range from 0 to 9990000000.00.
- The data is moderately positively skewed (Skewness: 2.01) and has moderate kurtosis (Kurtosis: 2.61), suggesting a somewhat non-normal distribution.

Operating_profit_per_person

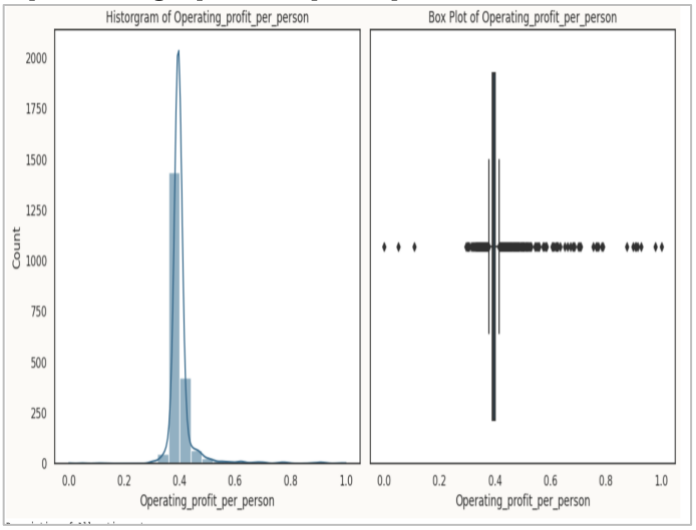


fig 1.15 Univariate Analysis of Operating_profit_per_person

```
Description of Operating_profit_per_person
-------------------------------------------------------------
count   2058.00
mean       0.40
std        0.05
min        0.00
25%        0.39
50%        0.40
75%        0.40
max        1.00
Name: Operating_profit_per_person, dtype: float64
Skewness: 5.34
Kurtosis: 48.27
```

Table 1.29 Descriptive Statistics of the Variable

- The data has a relatively low standard deviation (0.05) compared to the mean (0.40), suggesting low variability.
- This feature is highly positively skewed (Skewness: 5.34) and has high kurtosis (Kurtosis: 48.27), indicating a non-normal distribution with a heavy right tail.

Allocation_rate_per_person

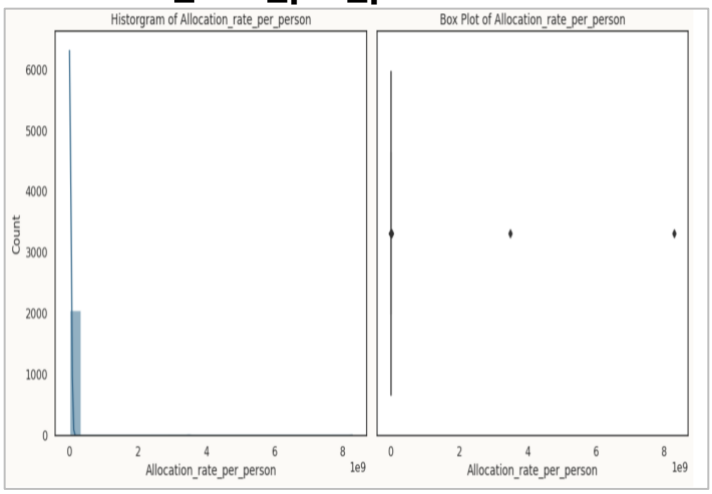


fig 1.16 Univariate Analysis of Allocation_rate_per_person

```
Description of Allocation_rate_per_person
-----------------------------------------------------------
count          2058.00
mean        5725558.82
std       197949961.06
min               0.00
25%               0.00
50%               0.01
75%               0.02
max      8280000000.00
Name: Allocation_rate_per_person, dtype: float64
Skewness: 38.17
Kurtosis: 1531.70
```

Table 1.30 Descriptive Statistics of the Variable

- The mean value is 5725558.82, with a wide range from 0 to 8280000000.00.
- The feature is highly positively skewed (Skewness: 38.17) and has extremely high kurtosis (Kurtosis: 1531.70), indicating a non-normal distribution with a heavy right tail.

Retained_Earnings_to_Total_Assets

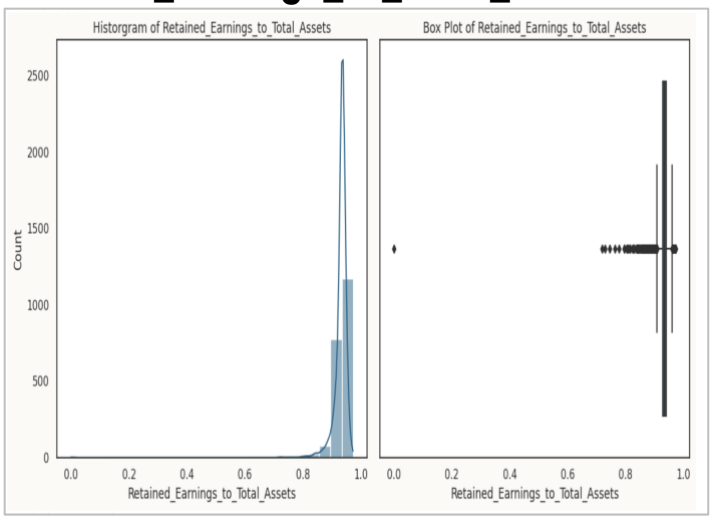


fig 1.17 Univariate Analysis of Retained_Earnings_to_Total_Assets

```
Description of Retained_Earnings_to_Total_Assets
-----------------------------------------------------------
count   2058.00
mean       0.93
std        0.03
min        0.00
25%        0.93
50%        0.94
75%        0.94
max        0.97
Name: Retained_Earnings_to_Total_Assets, dtype: float64
Skewness: -16.14
Kurtosis: 468.82
```

Table 1.31 Descriptive Statistics of the Variable

- The mean value is 0.93, suggesting a high proportion of retained earnings to total assets on average.
- The feature is highly negatively skewed (Skewness: -16.14) and has high kurtosis (Kurtosis: 468.82), indicating a non-normal distribution with a heavy left tail.

Total_income_to_Total_expense

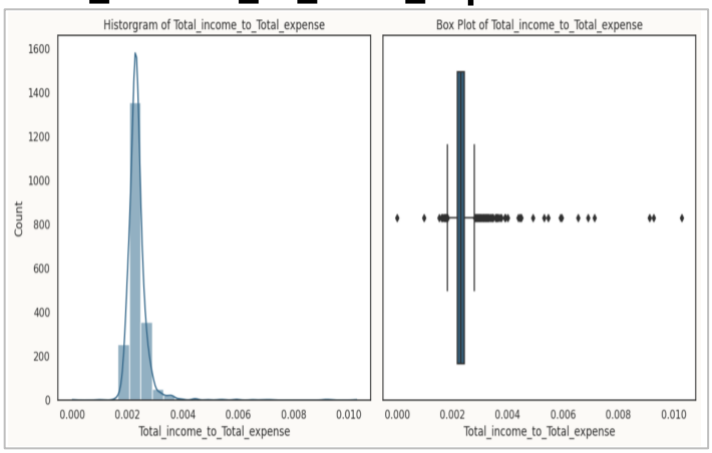


fig 1.18 Univariate Analysis of Total_income_to_Total_expense

```
Description of Total_income_to_Total_expense
-----------------------------------------------------------
count   2058.00
mean       0.00
std        0.00
min        0.00
25%        0.00
50%        0.00
75%        0.00
max        0.01
Name: Total_income_to_Total_expense, dtype: float64
Skewness: 8.02
Kurtosis: 105.17
```

Table 1.32 Descriptive Statistics of the Variable

- Mean of 0.0, indicating that total income is very close to total expenses on average.
- This feature is highly positively skewed (Skewness: 8.02) and has high kurtosis(105.17), indicating a non-normal distribution with a heavy right tail.

Total_expense_to_Assets

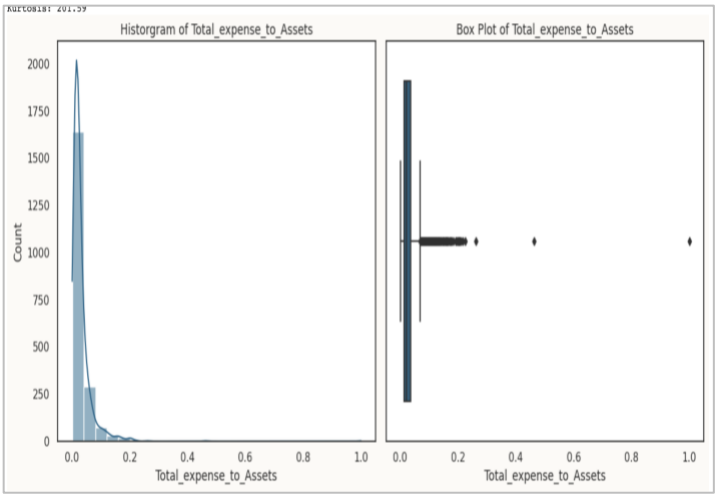


fig 1.19 Univariate Analysis of Total_expense_to_Assets

```
Description of Total_expense_to_Assets
------------------------------------------------
count   2058.00
mean       0.03
std        0.04
min        0.00
25%        0.01
50%        0.02
75%        0.04
max        1.00
Name: Total_expense_to_Assets, dtype: float64
Skewness: 9.75
Kurtosis: 201.59
```

Table 1.33 Descriptive Statistics of the Variable

- The data has a mean of 0.03, suggesting that total expenses are a small proportion of total assets on average.
- This feature is highly positively skewed (Skewness: 9.75) and has high kurtosis (Kurtosis: 201.59), indicating a non-normal distribution with a heavy right tail.

Equity_to_Liability

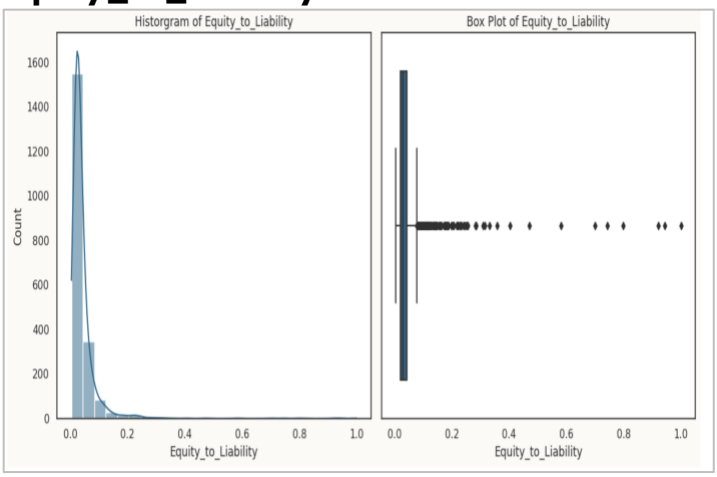


fig 1.20 Univariate Analysis of Equity_to_Liability

```
Description of Equity_to_Liability
------------------------------------------------
count   2058.00
mean       0.04
std        0.06
min        0.00
25%        0.02
50%        0.03
75%        0.04
max        1.00
Name: Equity_to_Liability, dtype: float64
Skewness: 9.14
Kurtosis: 115.45
```

Table 1.34 Descriptive Statistics of the Variable

- The mean value is 0.04, suggesting that equity is a small proportion of liabilities on average.
- This feature is highly positively skewed (Skewness: 9.14) and has high kurtosis (Kurtosis: 115.45), indicating a non-normal distribution with a heavy right tail.

Bivariate Analysis

In Bivariate Analysis, we will check how the features are related with each other, correlation between them.

Additionally we will explore how the important features are related to the target “Default”

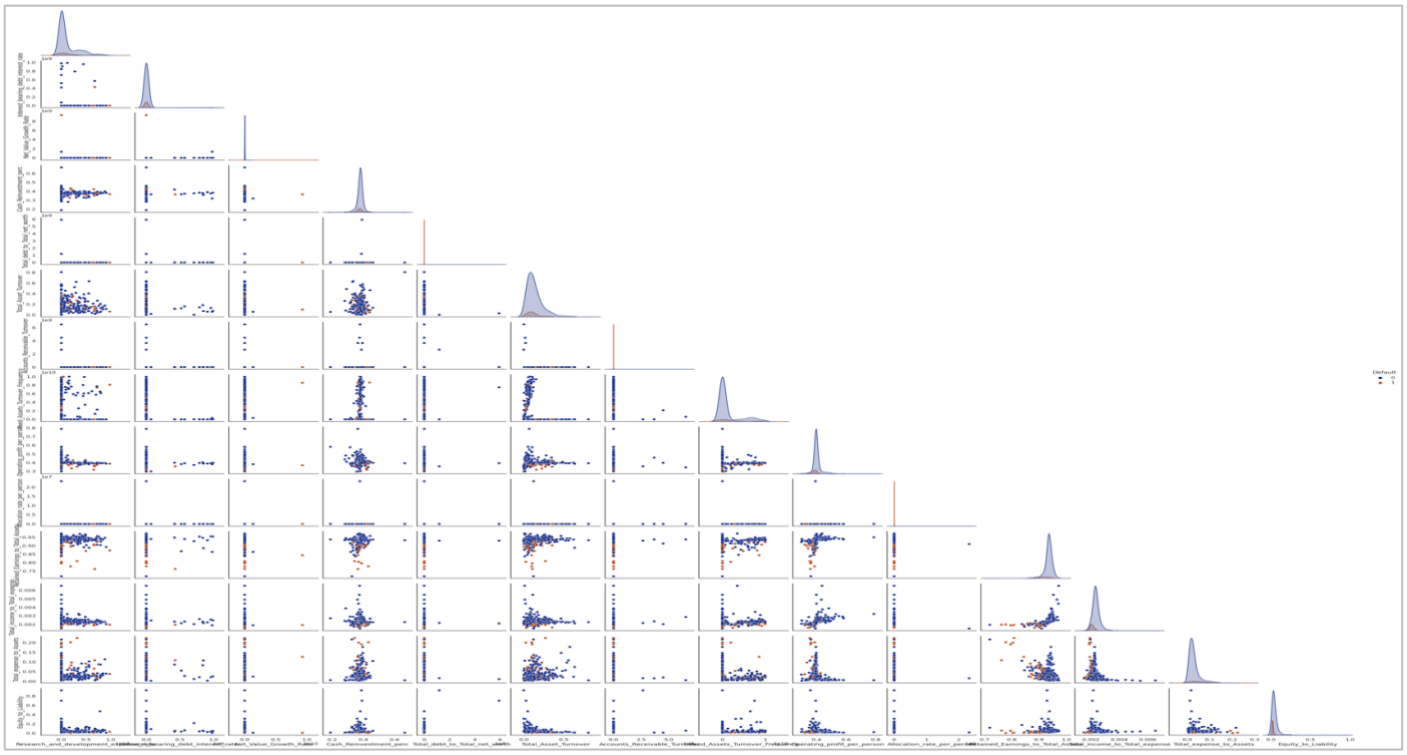


fig 1.21 Pairplot

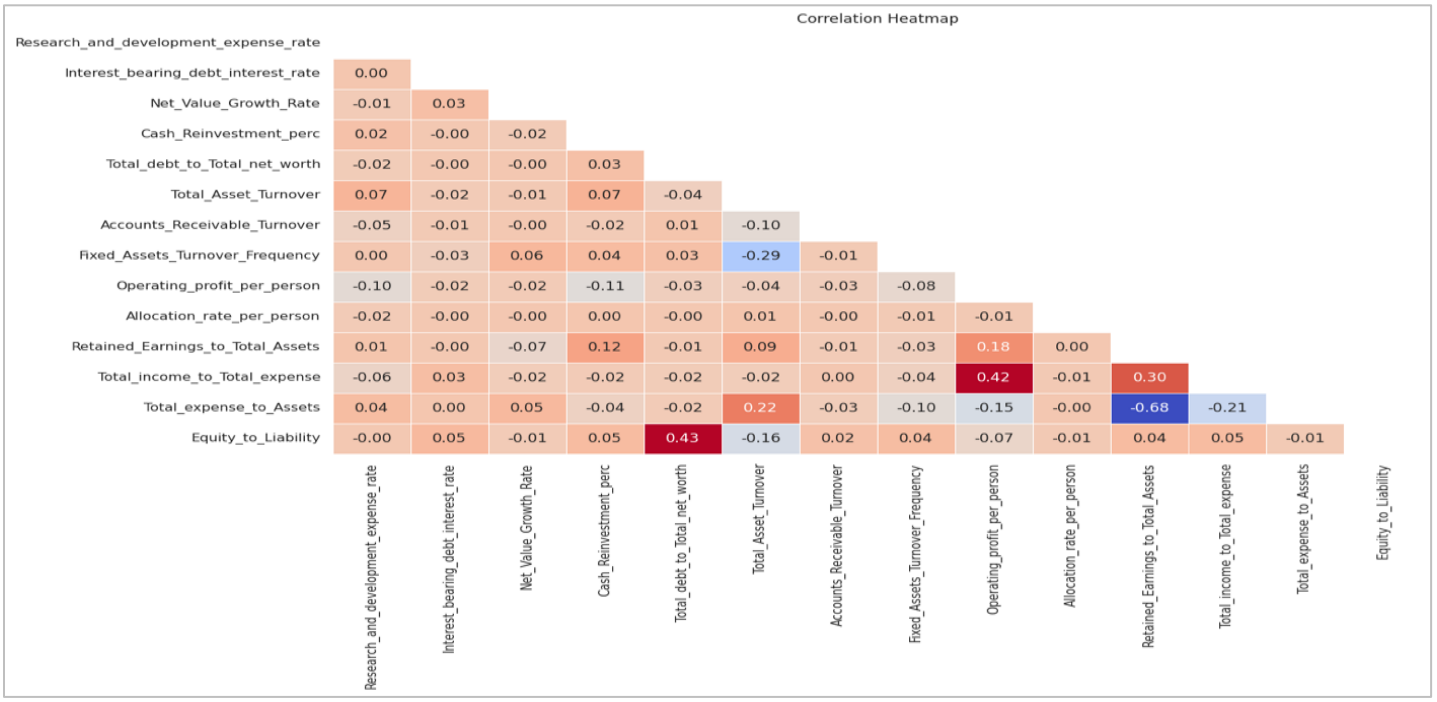


fig 1.22 Correlation Heatmap

- **Operating Profit per Person** and **Total Income to Total Expense** show **strong positive correlations**, indicating a close relationship between these variables.
- **Total Debt to Total Net Worth** has a moderate **positive correlation** with **Equity to Liability**, suggesting a strong financial structure.
- **Retained Earnings to Total Assets** has a **strong positive correlation** with **Total Income to Total Expense**, indicating healthy financial management.
- **Several variables show weak or no significant correlations**, indicating their **independence from other factors** in the dataset.

Default against most important variables:

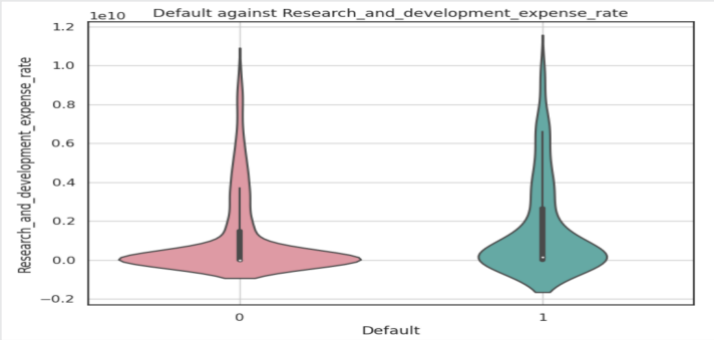


fig 1.23 Default against Research and Development Expense Rate

- The 'Default' 1 class tends to have higher research and development expenses compared to 'Default' 0.

- There is a wide range of variation in research and development expenses within both classes.

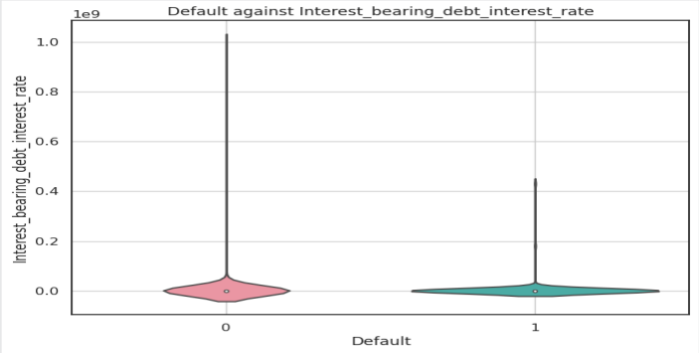


fig 1.24 Default against Interest-bearing Debt Interest Rate:

- 'Default' 0 has a wider distribution of interest rates, while 'Default' 1 has relatively lower interest rates.
- Most observations in both classes have low or zero interest rates.

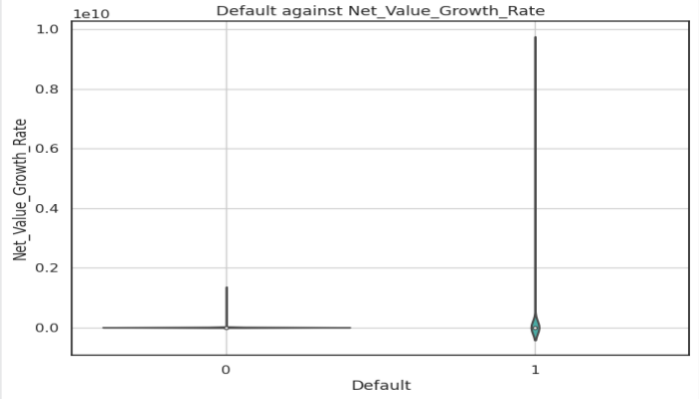


fig 1.25 Default against Net Value Growth Rate

- The growth rate is significantly higher for 'Default' 1 compared to 'Default' 0.
- There is a wide range of variation in growth rates within both classes.

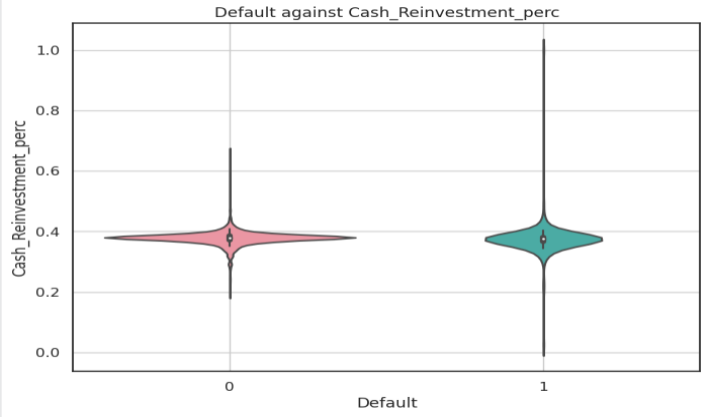


fig 1.26 Default against Cash Reinvestment Percentage

- Both classes have similar mean cash reinvestment percentages, but 'Default' 1 has a wider distribution.

- 'Default' 1 has some extreme values with very low or very high reinvestment percentages.

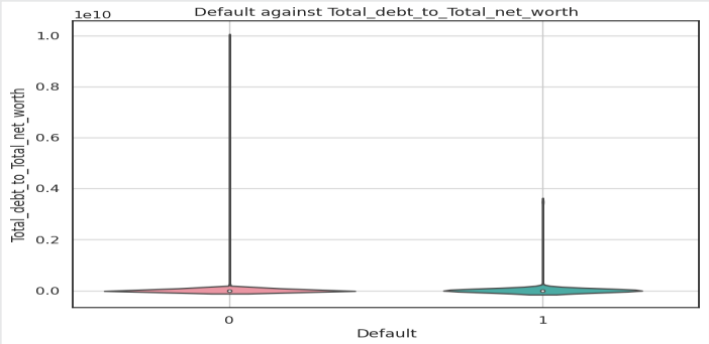


fig 1.27 Default against Total Debt to Total Net Worth

- 'Default' 1 has a slightly higher mean total debt-to-net-worth ratio compared to 'Default' 0.
- Both classes have a significant number of observations with low ratios.

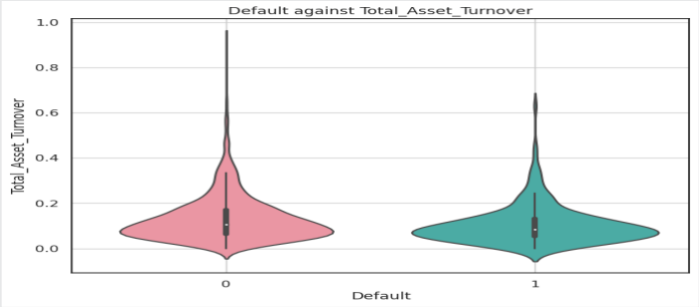


fig 1.28 Default against Total Asset Turnover

- 'Default' 0 tends to have a higher total asset turnover rate compared to 'Default' 1.

- There is a wide distribution of turnover rates in both classes.

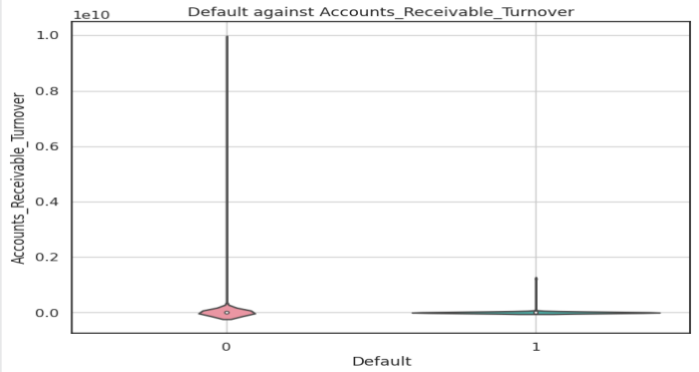


fig 1.29 Default against Accounts Receivable Turnover

- 'Default' 1 has a lower mean accounts receivable turnover compared to 'Default' 0.

- Both classes have a majority of observations with zero or low turnover.

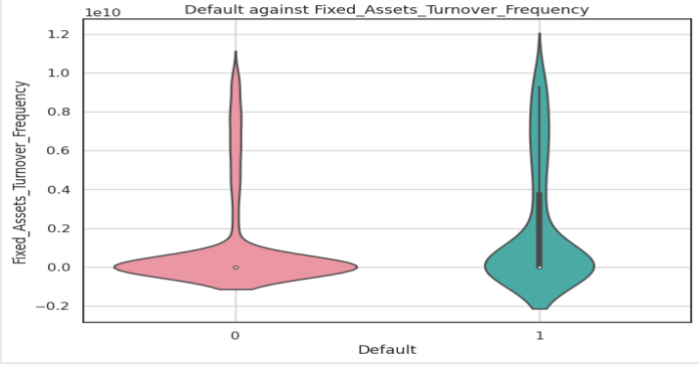


fig 1.30 Default against Fixed Assets Turnover Frequency

- 'Default' 1 has a wider distribution of turnover frequency, with some extreme values.

- 'Default' 0 has a narrower distribution with lower frequencies.

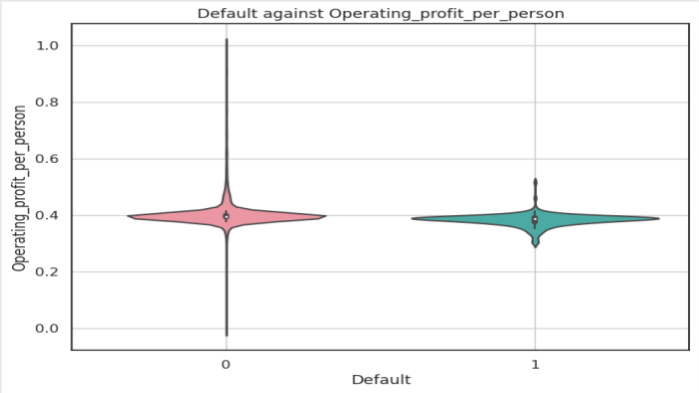


fig 1.31 Default against Operating Profit per Person

- 'Default' 0 has a slightly higher mean operating profit per person compared to 'Default' 1.

- Both classes have observations with relatively high profit per person.

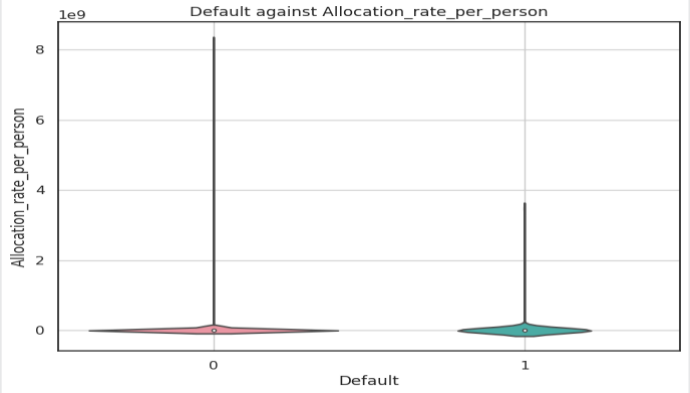


fig 1.32 Default against Retained Earnings to Total Assets

- 'Default' 1 has a higher mean allocation rate per person compared to 'Default' 0.

- Both classes have observations with a wide range of allocation rates

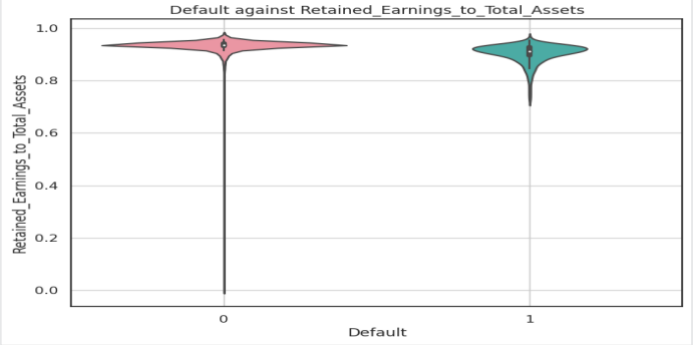


fig 1.33 Default against Retained Earnings to Total Assets

- 'Default' 0 has a slightly higher mean ratio of retained earnings to total assets compared to 'Default' 1.

- Both classes have a majority of observations with high ratios.

fig 1.34 Default against Total Income to Total Expense

- Both classes have very low mean ratios of total income to total expense, indicating low profitability.

- The majority of observations in both classes have low ratios.

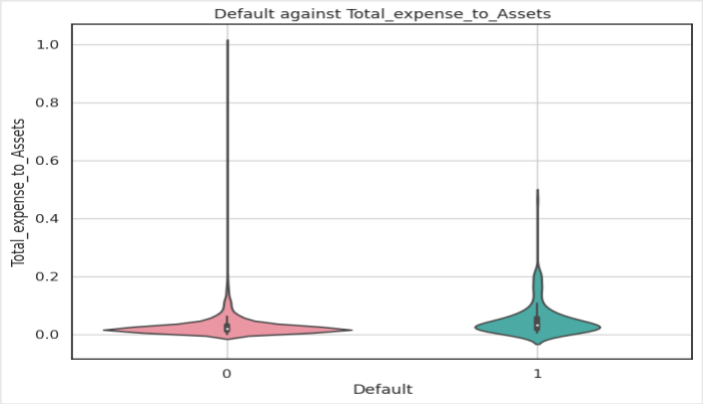


fig 1.35 Default against Total Expense to Assets

- 'Default' 1 has a higher mean ratio of total expenses to assets compared to 'Default' 0.

- Both classes have observations with a wide range of expense ratios.

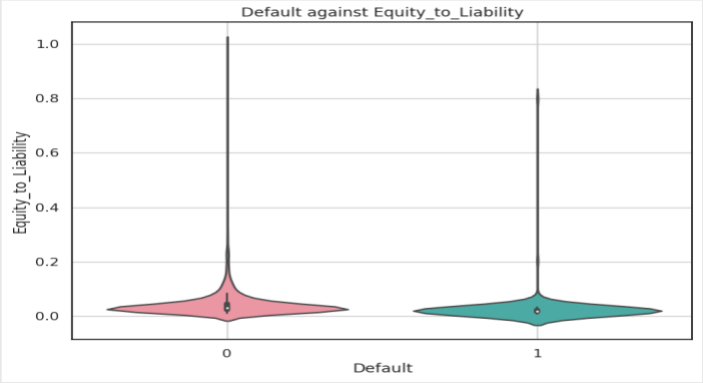


fig 1.36 Default against Equity to Liability

- 'Default' 0 has a higher mean equity-to-liability ratio compared to 'Default' 1.

- Both classes have observations with a wide range of equity ratios.

Train Test Split

We Split the data into train and test datasets in the ratio of 67:33 and used a random state of 42 (random_state=42).

Train set has 1378 records whereas the test set has 680 records.

However, for the logistic regression model using Statsmodels we do not need to segregate the X & the Y separately. Hence, for the purpose of Logistic Regression we concatenate the X & Y for both train & test set.

Build Logistic Regression Model (using statsmodels library) on most important variables on train dataset and choose the optimum cut-off. Also showcase your model building approach.

Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model

Logistic Regression Model (using Statsmodels library)

Logistic Regression is a supervised learning for classification. Logistic regression is the type of regression analysis used to find the probability of a certain event occurring. It is the best suited type of regression for cases where we have a categorical dependent variable which can take only discrete values.

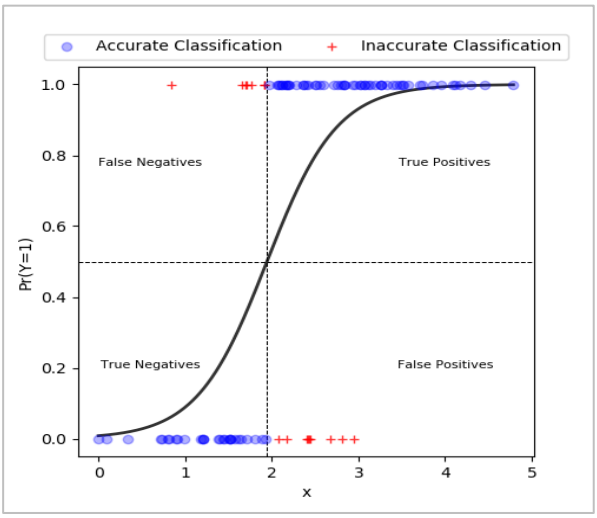


fig 1.37 Logistic Regression Explanation

We will build a logistic regression model using the statsmodels library. Statsmodels is a Python module that provides various functions for estimating different statistical models and performing statistical tests

First, we define the set of dependent(y) and independent(X) variables. Statsmodels provides a Logit() function for performing logistic regression. The Logit() function accepts y and X as parameters and returns the Logit object. The model is then fitted to the data.

Performance Metrices.
Since the dataset is skewed and type – II error is costly, i.e., **actually the company defaults when it is predicted as not going to default, we give priority to the recall score** of the default class over the accuracy to adjudge the model performance as incorrect prediction of default will lead to incorrect assessment of the company and a credit line may be
extended to the defaulting company which will lead to huge losses to the business.

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

fig 1.38 Confusion Matrix

Confusion Matrix:

- **Actual 0, Predicted 0 (TN - True Negative)**: This represents cases where the actual class is 0 (negative) and the model correctly predicts it as 0.
- **Actual 1, Predicted 1 (TP - True Positive)**: This represents cases where the actual class is 1 (positive) and the model correctly predicts it as 1.
- **Actual 0, Predicted 1 (FP - False Positive)**: This represents cases where the actual class is 0 (negative) but the model incorrectly predicts it as 1 (positive). It's also known as a Type I error or a false alarm.
- **Actual 1, Predicted 0 (FN - False Negative)**: This represents cases where the actual class is 1 (positive) but the model incorrectly predicts it as 0 (negative). It's also known as a Type II error or a missed detection.
- **Recall (Sensitivity)**: This is all about how well the model can spot the actual credit defaults. If it has high recall, it's good at catching most of the real defaults. That's a plus because it means fewer

cases of missing out on potential defaulters (those false negatives), and that's something lenders want to avoid because it can lead to losses.

- **Precision**: Precision looks at the proportion of cases the model predicts as defaults that actually turn out to be true credit defaults. If precision is a bit lower, it means there might be more cases where the model predicts defaults that don't really happen (false positives). This cautious approach helps minimize the risk of lending to people who might not actually default.

The ROC (Receiver Operating Characteristic) curve evaluates a classification model's performance by plotting the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) as the classification distinguishes positive and negative instances.

AUC (Area Under the Curve) summarizes overall performance, with higher values indicating better model performance threshold varies. It helps visualize how well the model

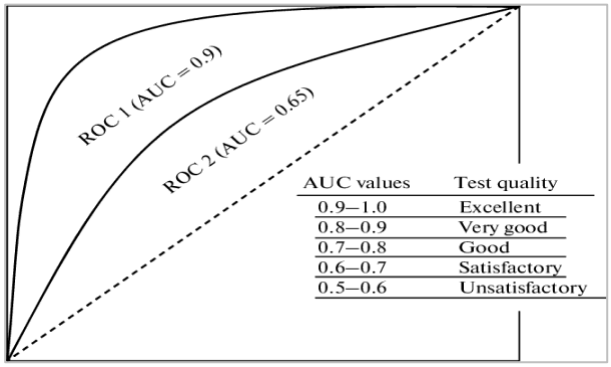


fig 1.39 ROC-AUC

Model Building

Model 1 :

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1363
Method:	MLE	Df Model:	14
Date:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.4321
Time:	04:38:36	Log-Likelihood:	-272.84
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	7.772e-80

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8741	0.247	-15.658	0.000	-4.359	-3.389
Research_and_development_expense_rate	0.3853	0.112	3.432	0.001	0.165	0.605
Interest_bearing_debt_interest_rate	0.4901	0.139	3.516	0.000	0.217	0.763
Net_Value_Growth_Rate	-0.2870	0.161	-1.778	0.075	-0.603	0.029
Cash_Reinvestment_perc	-0.3396	0.110	-3.075	0.002	-0.556	-0.123
Total_debt_to_Total_net_worth	0.4346	0.176	2.472	0.013	0.090	0.779
Total_Asset_Turnover	-0.1997	0.160	-1.249	0.212	-0.513	0.114
Accounts_Receivable_Turnover	-0.5475	0.138	-3.960	0.000	-0.818	-0.277
Fixed_Assets_Turnover_Frequency	0.1590	0.121	1.311	0.190	-0.079	0.397
Operating_profit_per_person	0.5307	0.192	2.769	0.006	0.155	0.906
Allocation_rate_per_person	0.5723	0.159	3.603	0.000	0.261	0.884
Retained_Earnings_to_Total_Assets	-0.7195	0.216	-3.332	0.001	-1.143	-0.296
Total_income_to_Total_expense	-0.9353	0.289	-3.231	0.001	-1.503	-0.368
Total_expense_to_Assets	0.3807	0.157	2.420	0.016	0.072	0.689
Equity_to_Liability	-0.8258	0.311	-2.655	0.008	-1.435	-0.216

Table 1.35 Logistic Regression Model Summary

We Observe that there is presence of **Insignificant predictors like "Total_Asset_Turnover", "Fixed_Assets_Turnover_Frequency" etc having p-value > 0.05 need to be removed**. We will remove all features one by one, till we have any features with p-value>0.05

Model 2: Removed ' **Total_Asset_Turnover'** since p-value is 0.212 > 0.05

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1364
Method:	MLE	Df Model:	13
Date:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.4305
Time:	04:38:36	Log-Likelihood:	-273.63
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	2.921e-80

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8697	0.248	-15.632	0.000	-4.355	-3.384
Research_and_development_expense_rate	0.3722	0.112	3.335	0.001	0.153	0.591
Interest_bearing_debt_interest_rate	0.4698	0.138	3.407	0.001	0.200	0.740
Net_Value_Growth_Rate	-0.3024	0.162	-1.871	0.061	-0.619	0.014
Cash_Reinvestment_perc	-0.3403	0.110	-3.088	0.002	-0.556	-0.124
Total_debt_to_Total_net_worth	0.3898	0.171	2.277	0.023	0.054	0.725
Accounts_Receivable_Turnover	-0.5643	0.138	-4.076	0.000	-0.836	-0.293
Fixed_Assets_Turnover_Frequency	0.1767	0.121	1.465	0.143	-0.060	0.413
Operating_profit_per_person	0.5373	0.192	2.794	0.005	0.160	0.914
Allocation_rate_per_person	0.6424	0.150	4.296	0.000	0.349	0.936
Retained_Earnings_to_Total_Assets	-0.7786	0.211	-3.697	0.000	-1.191	-0.366
Total_income_to_Total_expense	-0.9501	0.290	-3.271	0.001	-1.519	-0.381
Total_expense_to_Assets	0.3341	0.152	2.191	0.028	0.035	0.633
Equity_to_Liability	-0.8070	0.310	-2.602	0.009	-1.415	-0.199

Table 1.36 Logistic Regression Model Summary

Model 3: Removed ' **Fixed_Assets_Turnover_Frequency'** since p-value is 0.143 > 0.05

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1365
Method:	MLE	Df Model:	12
Date:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.4282
Time:	04:38:36	Log-Likelihood:	-274.70
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	1.380e-80

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.8886	0.248	-15.676	0.000	-4.375	-3.402
Research_and_development_expense_rate	0.3822	0.111	3.434	0.001	0.164	0.600
Interest_bearing_debt_interest_rate	0.4706	0.138	3.420	0.001	0.201	0.740
Net_Value_Growth_Rate	-0.2690	0.160	-1.677	0.094	-0.583	0.045
Cash_Reinvestment_perc	-0.3349	0.109	-3.059	0.002	-0.549	-0.120
Total_debt_to_Total_net_worth	0.3894	0.171	2.284	0.022	0.055	0.724
Accounts_Receivable_Turnover	-0.5685	0.138	-4.126	0.000	-0.839	-0.298
Operating_profit_per_person	0.5279	0.190	2.778	0.005	0.155	0.900
Allocation_rate_per_person	0.7229	0.139	5.196	0.000	0.450	0.996
Retained_Earnings_to_Total_Assets	-0.7993	0.211	-3.781	0.000	-1.214	-0.385
Total_income_to_Total_expense	-1.0241	0.286	-3.579	0.000	-1.585	-0.463
Total_expense_to_Assets	0.3042	0.151	2.013	0.044	0.008	0.600
Equity_to_Liability	-0.7933	0.309	-2.565	0.010	-1.399	-0.187

Table 1.37 Logistic Regression Model Summary

Model 4: Removed ' **Net_Value_Growth_Rate'** since p-value is 0.094 > 0.05

Iterations 9

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1366
Method:	MLE	Df Model:	11
Date:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.4251
Time:	04:38:36	Log-Likelihood:	-276.21
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	9.561e-81

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9483	0.250	-15.777	0.000	-4.439	-3.458
Research_and_development_expense_rate	0.3894	0.110	3.526	0.000	0.173	0.606
Interest_bearing_debt_interest_rate	0.4397	0.135	3.262	0.001	0.176	0.704
Cash_Reinvestment_perc	-0.3172	0.108	-2.948	0.003	-0.528	-0.106
Total_debt_to_Total_net_worth	0.4057	0.169	2.400	0.016	0.074	0.737
Accounts_Receivable_Turnover	-0.5761	0.137	-4.194	0.000	-0.845	-0.307
Operating_profit_per_person	0.5205	0.191	2.727	0.006	0.146	0.895
Allocation_rate_per_person	0.7605	0.138	5.498	0.000	0.489	1.032
Retained_Earnings_to_Total_Assets	-0.8377	0.207	-4.050	0.000	-1.243	-0.432
Total_income_to_Total_expense	-1.1833	0.273	-4.328	0.000	-1.719	-0.647
Total_expense_to_Assets	0.3511	0.146	2.401	0.016	0.065	0.638
Equity_to_Liability	-0.8667	0.308	-2.814	0.005	-1.470	-0.263

Table 1.38 Logistic Regression Model Summary

After removing non-significant features, the **model 4** is the Optimized model with **12 features.**

Logit Score Equation: = -3.9483 + 0.3894 * Research_and_development_expense_rate
+ 0.4397 * Interest_bearing_debt_interest_rate - 0.3172 * Cash_Reinvestment_perc
+ 0.4057 * Total_debt_to_Total_net_worth - 0.5761 * Accounts_Receivable_Turnover
+ 0.5205 * Operating_profit_per_person + 0.7605 * Allocation_rate_per_person
- 0.8377 * Retained_Earnings_to_Total_Assets
- 1.1833 * Total_income_to_Total_expense
+ 0.3511 * Total_expense_to_Assets - 0.8667 * Equity_to_Liability

The logit score equation assesses financial risk, with positive coefficients increasing Default likelihood, and negative coefficients decreasing it. For example, keeping others constant for each one-unit increase in the "Research and Development Expense Rate," the log-odds of Default increase by 0.3894 units. Conversely, for each one-unit increase in the " Total_income_to_Total_expense ," the log-odds of Default decrease by 1.1833 units.

Model Performance Evaluation

Classification Report on Train & Test data:-

Train data Classification Report

	precision	recall	f1-score	support
0.0	0.94	0.98	0.96	1225
1.0	0.71	0.46	0.56	153
accuracy			0.92	1378
macro avg	0.82	0.72	0.76	1378
weighted avg	0.91	0.92	0.91	1378

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.94	0.96	0.95	613
1.0	0.56	0.46	0.51	67
accuracy			0.91	680
macro avg	0.75	0.71	0.73	680
weighted avg	0.91	0.91	0.91	680

Table 1.39 Logistic Regression Classification report

Confusion Matrix:-

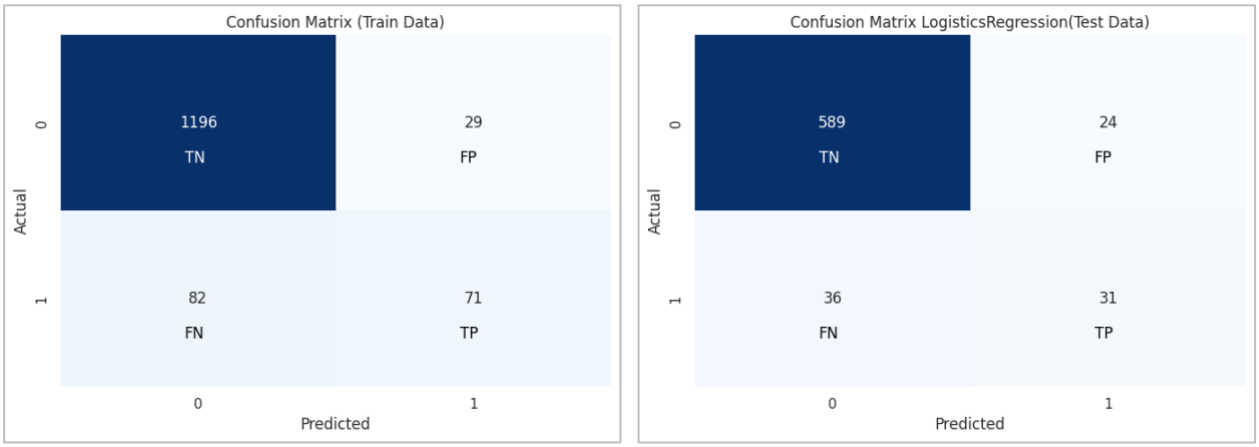


fig 1.40 Logistic Regression Confusion Matrix

ROC Curve

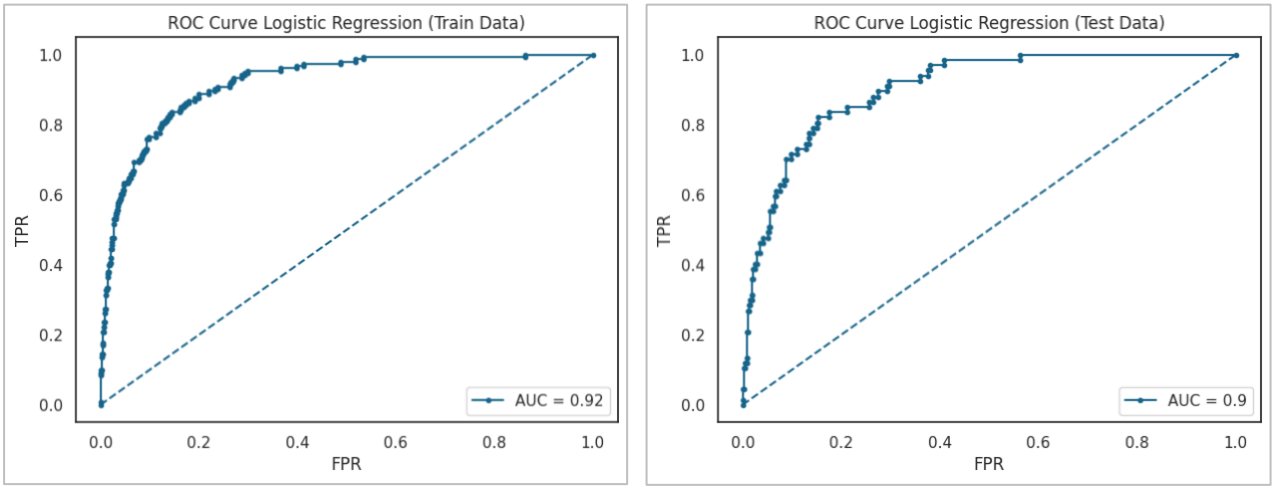


fig 1.41 Logistic Regression ROC curve

Logistic Regression Conclusion

Metrics	Train	Test
True Negative	1196	589
True Positive	71	31
False Negative	82	36
False Positive	29	24
AUC	92%	90%
Accuracy	92%	91%
Precision	71%	56%
Recall	46%	46%
f1-score	56%	51%

Table 1.40 Logistic Regression Conclusion

Insights

The Logistic Regression model demonstrates good predictive capabilities for identifying default and non-default cases. When analyzing its performance on both the training and test datasets, several key findings emerge:

- **True Positives and True Negatives:** The model successfully identifies more true positives (default cases) in the test dataset, while maintaining a high accuracy in predicting true negatives (non-default cases) in the training dataset.
- **False Positives and False Negatives:** The model exhibits a lower rate of false positives (incorrectly predicted default cases) on the test dataset, indicating its ability to avoid erroneous classifications. Additionally, the test data shows fewer false negatives (missed default cases) compared to the training data.
- **AUC (Area Under the ROC Curve):** Although the AUC is slightly higher on the training data, both datasets display high AUC values, suggesting excellent discrimination between classes. This indicates the model's robustness in distinguishing between default and non-default cases.

- **Accuracy:** The model maintains a remarkably high level of accuracy on both the training and test datasets. Despite a slight drop in accuracy on the test data, the model's overall accuracy remains impressive.
- **Precision and Recall:** Precision, which measures the ability to avoid false positives, demonstrates a minor decrease on the test data but remains relatively high. Recall, which measures the ability to capture true positives, remains consistent between the two datasets.
- **F1-Score:** The F1-score, which balances precision and recall, experiences a slight decrease on the test data but remains at a satisfactory level.

Since, Recall is important in Financial Risks, the recall from above model was very less. We'll find optimum cut-off to improve Recall

Adjusting Optimal Cutoff

We adjust the threshold to find the optimum cut-off value that will improve the recall of the model. The optimum cut-off value comes out to be 0.128.
Let's see the evaluation Metrices:

Classification Report on Train & Test data:-

Train data Classification Report

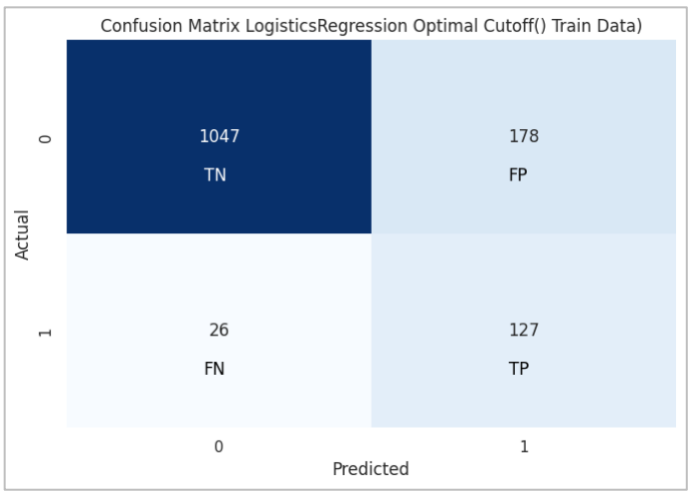
	precision	recall	f1-score	support
0.0	0.98	0.85	0.91	1225
1.0	0.42	0.83	0.55	153
accuracy			0.85	1378
macro avg	0.70	0.84	0.73	1378
weighted avg	0.91	0.85	0.87	1378

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.97	0.84	0.90	613
1.0	0.34	0.79	0.48	67
accuracy			0.83	680
macro avg	0.66	0.81	0.69	680
weighted avg	0.91	0.83	0.86	680

Table 1.41 Logistic Regression Optimal Cutoff Classification report

Confusion Matrix:-



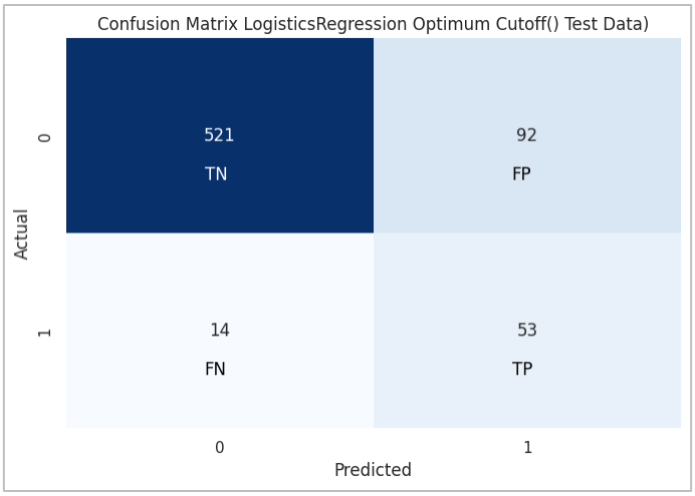


fig 1.42 Logistic Regression Optimal Cutoff Confusion Matrix

ROC Curve

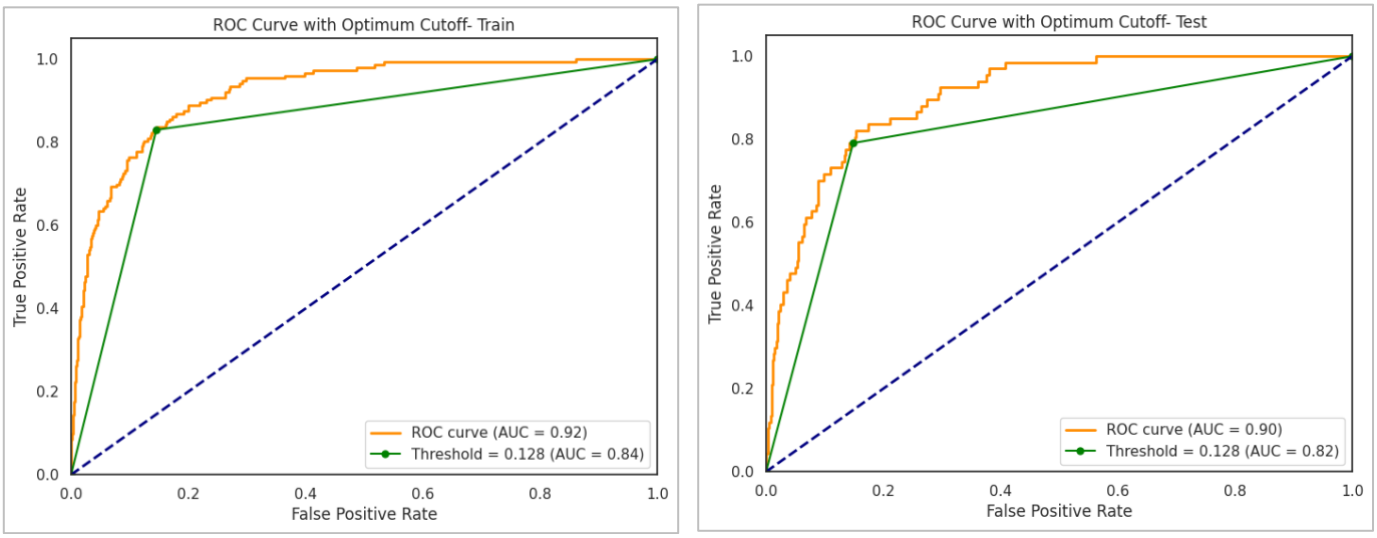


fig 1.43 Logistic Regression Optimal Cutoff ROC curve

Logistic Regression Conclusion

Metrics	Train	Test
True Negative	1047	521
True Positive	127	53
False Negative	27	14
False Positive	178	92
AUC	84%	82%
Accuracy	85%	83%
Precision	42%	34%
Recall	83%	79%
f1-score	55%	48%

Table 1.42 Logistic Regression Conclusion

Insights

After adjusting the **optimum cutoff to 0.128**, we can see **significantly improved Recall score.**

- The Precision, Recall, F1 score & Accuracy of training data for the model is in line with the testing data.
- The **recall score** for the **default class has increased** but the accuracy of the model has slightly decreased. Since we need better recall here, the model performs well in doing so.
- The Precision dropped drastically In this case along with slight decrease in AUC and F1.

Let's find out if we can further improve the performance using Oversampling.

Oversampling underrepresented class using SMOTE

Synthetic Minority Over-sampling Technique a method used in machine learning to address class imbalance in datasets, particularly in scenarios where one class is underrepresented compared to others. Class imbalance can lead to biased models, and SMOTE helps balance the class distribution.

Sampling Approach: We chose sampling_strategy as= .75 which means minority class will be upsampled to have 75% as many samples as the majority class

Model 5 :

Logit Regression Results

ep. Variable:	Default	No. Observations:	2143
odel:	Logit	Df Residuals:	2131
ethod:	MLE	Df Model:	11
ate:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.5262
ime:	07:11:14	Log-Likelihood:	-693.27
onverged:	True	LL-Null:	-1463.3
ovariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
ntercept	-2.6458	0.145	-18.222	0.000	-2.930	-2.361
esearch_and_development_expense_rate	0.4944	0.073	6.813	0.000	0.352	0.637
nterest_bearing_debt_interest_rate	0.5442	0.092	5.942	0.000	0.365	0.724
ash_Reinvestment_perc	-0.3631	0.074	-4.924	0.000	-0.508	-0.219
otal_debt_to_Total_net_worth	0.4081	0.107	3.824	0.000	0.199	0.617
ccounts_Receivable_Turnover	-0.7779	0.093	-8.321	0.000	-0.961	-0.595
perating_profit_per_person	0.5554	0.127	4.377	0.000	0.307	0.804
llocation_rate_per_person	0.8263	0.088	9.420	0.000	0.654	0.998
etained_Earnings_to_Total_Assets	-1.0218	0.140	-7.296	0.000	-1.296	-0.747
otal_income_to_Total_expense	-1.3575	0.180	-7.560	0.000	-1.709	-1.006
otal_expense_to_Assets	0.1883	0.100	1.877	0.061	-0.008	0.385
quity_to_Liability	-1.1014	0.173	-6.376	0.000	-1.440	-0.763

Table 1.43 Logistic Regression SMOTE Model Summary

Model 6 : All features are not significant features, i.e. p-value<0.05. We have Total_expense_to_Assets having p-value > 0.05 i.e. **0.061. Let's drop that**.

Logit Regression Results

Dep. Variable:	Default	No. Observations:	2143
Model:	Logit	Df Residuals:	2132
Method:	MLE	Df Model:	10
Date:	Sat, 07 Oct 2023	Pseudo R-squ.:	0.5251
Time:	07:13:54	Log-Likelihood:	-695.01
converged:	True	LL-Null:	-1463.3
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.6617	0.145	-18.305	0.000	-2.947	-2.377
Research_and_development_expense_rate	0.5083	0.073	7.009	0.000	0.366	0.650
Interest_bearing_debt_interest_rate	0.5433	0.092	5.914	0.000	0.363	0.723
Cash_Reinvestment_perc	-0.3551	0.073	-4.854	0.000	-0.498	-0.212
Total_debt_to_Total_net_worth	0.3943	0.106	3.722	0.000	0.187	0.602
Accounts_Receivable_Turnover	-0.7441	0.091	-8.154	0.000	-0.923	-0.565
Operating_profit_per_person	0.5381	0.126	4.282	0.000	0.292	0.784
Allocation_rate_per_person	0.7477	0.076	9.807	0.000	0.598	0.897
Retained_Earnings_to_Total_Assets	-1.1394	0.127	-8.981	0.000	-1.388	-0.891
Total_income_to_Total_expense	-1.3234	0.178	-7.451	0.000	-1.672	-0.975
Equity_to_Liability	-1.1075	0.173	-6.394	0.000	-1.447	-0.768

Table 1.44 Logistic Regression SMOTE Model Summary

After removing non-significant features, the **model 4** is the Optimized model with **11 features.**

Logit Score Equation: = -2.6617 + 0.5083 * Research_and_development_expense_rate
+ 0.5433 * Interest_bearing_debt_interest_rate - 0.3551 * Cash_Reinvestment_perc
+ 0.3943 * Total_debt_to_Total_net_worth - 0.7441 * Accounts_Receivable_Turnover
+ 0.5381 * Operating_profit_per_person + 0.7477 * Allocation_rate_per_person
- 1.1394 * Retained_Earnings_to_Total_Assets
- 1.3234 * Total_income_to_Total_expense - 1.1075 * Equity_to_Liability

The logit score equation assesses financial risk, with positive coefficients increasing Default likelihood, and negative coefficients decreasing it. For example, keeping others constant for each one-unit increase in the "Research and Development Expense Rate," the log-odds of Default increase by 0.5083 units. Conversely, for each one-unit increase in the " Accounts_Receivable_Turnover," the log-odds of Default decrease by 0.7441units.

Model Performance Evaluation

Classification Report on Train & Test data:-

Train data Classification Report

	precision	recall	f1-score	support
0.0	0.88	0.87	0.87	1225
1.0	0.83	0.84	0.83	918
accuracy			0.86	2143
macro avg	0.85	0.86	0.85	2143
weighted avg	0.86	0.86	0.86	2143

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.97	0.86	0.91	613
1.0	0.37	0.76	0.50	67
accuracy			0.85	680
macro avg	0.67	0.81	0.70	680
weighted avg	0.91	0.85	0.87	680

Table 1.45 Logistic Regression SMOTE Classification report

Confusion Matrix:-

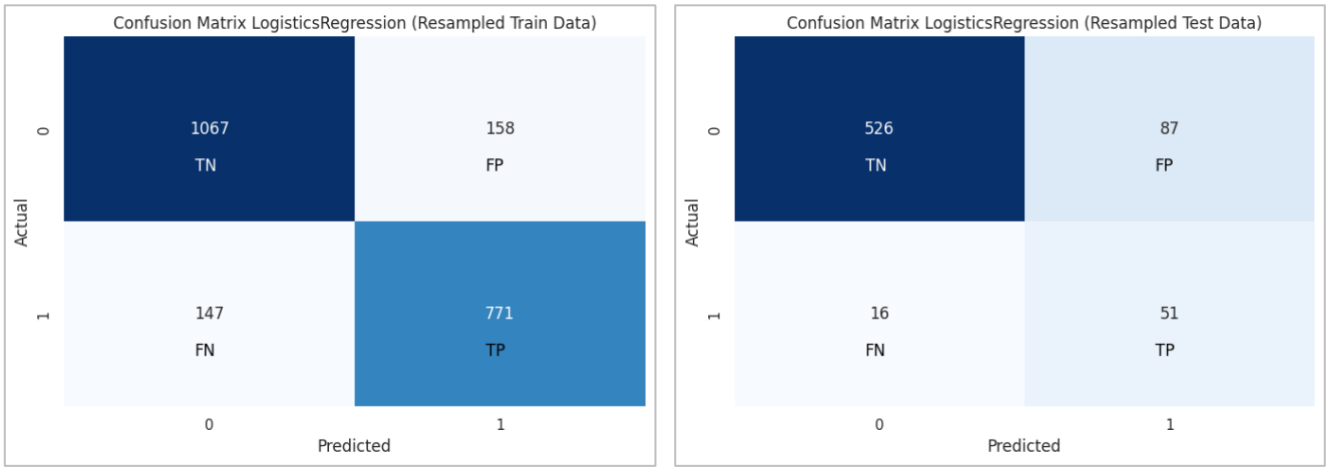


fig 1.44 Logistic Regression SMOTE Confusion Matrix

ROC Curve

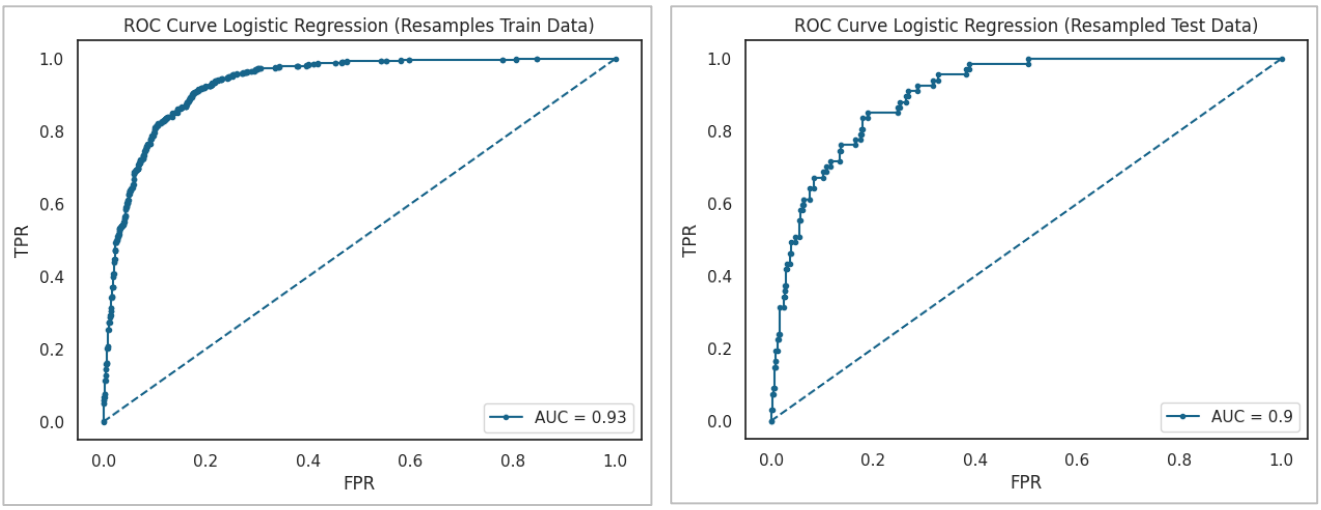


fig 1.45 Logistic Regression SMOTE ROC curve

Logistic Regression Conclusion

Metrics	Train	Test
True Negative	1067	526
True Positive	771	51
False Negative	147	16
False Positive	158	87
AUC	93%	90%
Accuracy	86%	85%
Precision	83%	37%
Recall	84%	76%
f1-score	83%	50%

Table 1.46 Logistic Regression Conclusion

Insights

- Applying SMOTE has led to a reduction in false negatives, improved recall, and a more balanced F1-score, indicating a better ability to detect actual default cases as compared to the Logistics Model without optimized cut off.
- There is a slight decrease in precision, and the AUC has also decreased, suggesting a trade-off between different aspects of model performance.
- However, recall is slightly lower than that of the optimized cut-off model. Let's use optimized cut off here as well to see if any improvement.

Adjusting Optimum Cutoff

We adjust the threshold to find the optimum cut-off value that will improve the recall of the model. The **optimum cut-off value comes out to be 0.367**. Let's see the evaluation Metrices:

Classification Report on Train & Test data:-

Train data Classification Report

	precision	recall	f1-score	support
0.0	0.92	0.82	0.87	1225
1.0	0.79	0.91	0.85	918
accuracy			0.86	2143
macro avg	0.86	0.87	0.86	2143
weighted avg	0.87	0.86	0.86	2143

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.98	0.82	0.89	613
1.0	0.33	0.82	0.47	67
accuracy			0.82	680
macro avg	0.66	0.82	0.68	680
weighted avg	0.91	0.82	0.85	680

Table 1.47 Logistic Regression Optimal Cutoff Classification report

Confusion Matrix:-

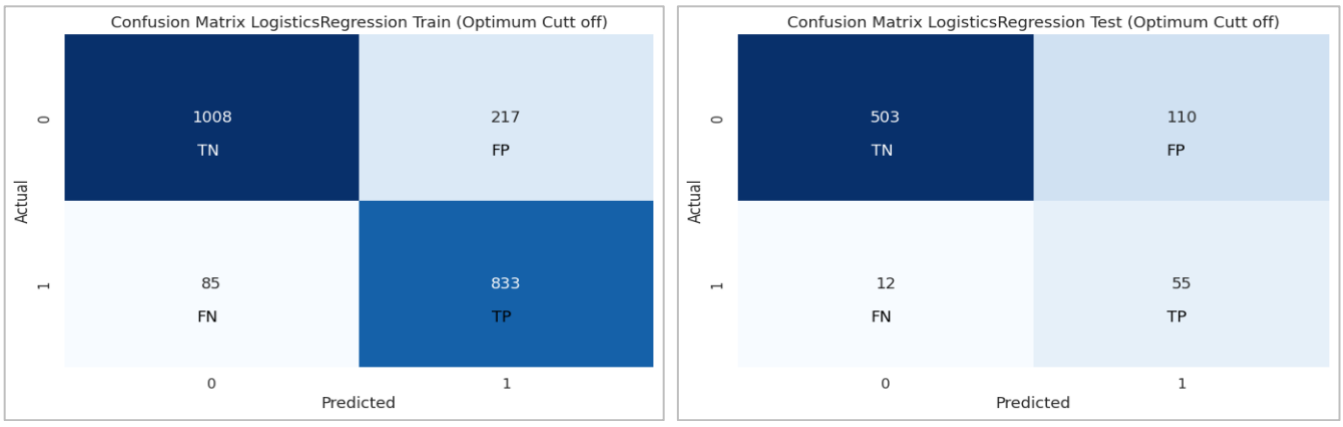


fig 1.46 Logistic Regression SMOTE Optimal Cutoff Confusion Matrix

ROC Curve

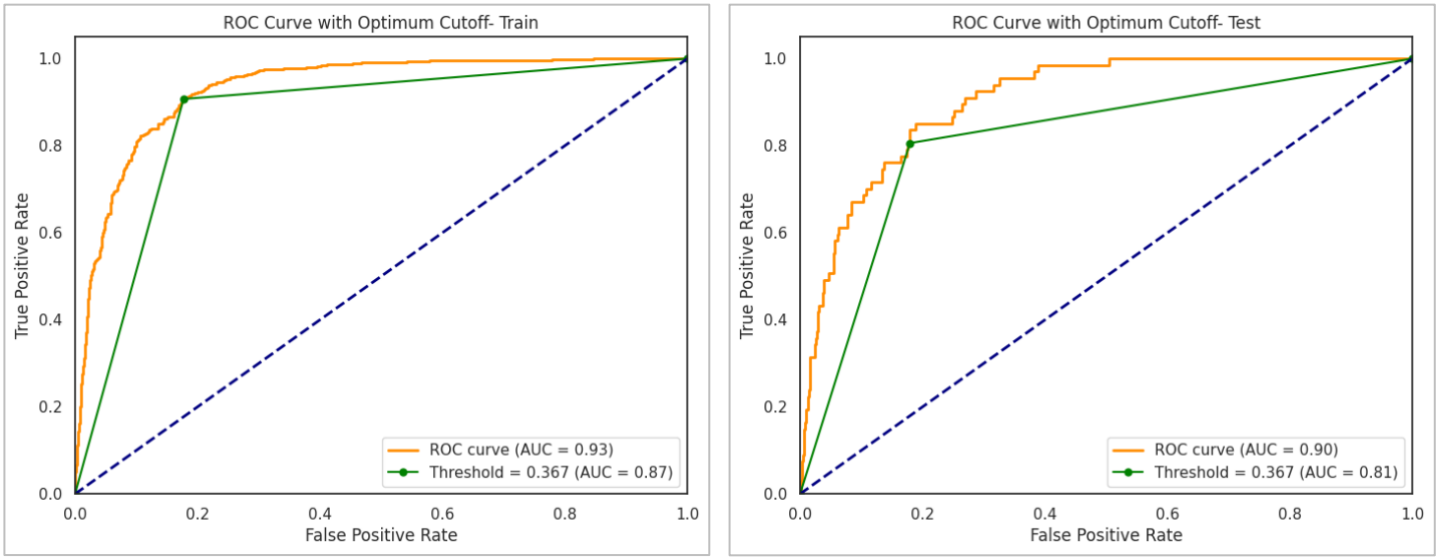


fig 1.47 Logistic Regression SMOTE Optimal ROC curve

Logistic Regression Conclusion

Metrics	Train	Test
True Negative	1008	503
True Positive	833	55
False Negative	85	12
False Positive	217	110
AUC	87%	81%
Accuracy	86%	82%
Precision	79%	33%
Recall	91%	82%
f1-score	85%	47%

Table 1.48 Logistic Regression Conclusion

Insights

- SMOTE with optimum cut off provides excellent recall, while with default it offers a balanced trade-off between precision and recall. This model achieves a high recall of 91% in training and 82% in test, indicating excellent detection of actual default cases.

- Precision has decreased to 33% on test data due to the optimized threshold, leading to a trade-off between precision and recall.

- The F1-score and accuracy has slightly reduced. The F1 score reflecting a good balance between precision and recall.

Model Selection for Financial Credit Default Prediction:

Metrix	Logit		Logit Optimum Cut off 0.128		Logit SMOTE		Logit SMOTE Optimum Cut off 0.367	
	Train Data	Test Data	Train Data	Test Data	Train Data	Test Data	Train Data	Test Data
True Negative	1196	589	1047	521	1067	526	1008	503
True Positive	71	31	127	53	771	51	833	55
False Negative	82	36	27	14	147	16	85	12
False Positive	29	24	178	92	158	87	217	110
AUC	92%	90%	84%	82%	92%	90%	87%	81%
Accuracy	92%	91%	85%	83%	86%	85%	86%	82%
Precision	71%	56%	42%	34%	83%	37%	79%	33%
Recall	46%	46%	83%	79%	84%	76%	91%	82%
f1	56%	51%	55%	48%	83%	50%	85%	47%

Table 1.49 Logistic Regression Metric Comparison

The choice of the model depends on the specific goals and constraints of the credit default prediction task. Since the main objective here is to maximize the detection of actual default cases while tolerating some false positives, **SMOTE with optimized threshold** would be the **preferred choice due to its high recall on test (82%)** among all the models evaluated so far.

Build a Random Forest Model on Train Dataset. Also showcase your model building approach. Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model

Random Forest is a powerful ensemble learning technique, often regarded as an extension of decision trees. It operates through the following steps:-

1.Random Sampling: Random Forest selects random subsets of data from the given dataset.

2.Decision Trees: For each of these data subsets, it constructs a decision tree, making predictions on each subset.

3.Voting: Random Forest then aggregates the predictions from all decision trees through a voting mechanism.

4.Majority Rule: The prediction result with the highest number of votes becomes the final prediction.

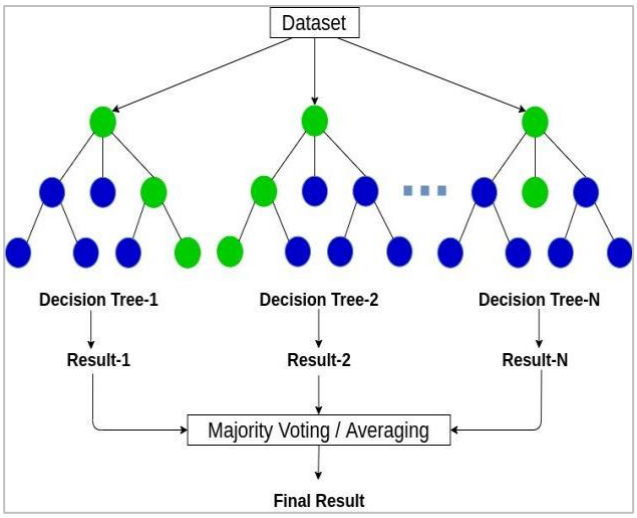


fig 1.48 Random Forrest Explanation

Approach:-

We build a random forest classifier & search for the best fit model, using Grid Search Cross Validation Technique with the following values:

Max depth: 5, 6, 7 => Depth of each decision tree

Max features: 3, 4 => based on ✓ (No. of Independent Variables)

Min samples leaf: 40, 50, 60 => based on 1 - 3 % of the records

Min samples split: 120, 150, 180 => based on 3 times min sample split
No. of estimators: 101, 301, 501 => No. of Decision Trees
Random State : 42

We obtain the best fit model with the parameters:

{'max_depth': 4, 'max_features': 2, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 101, random_state : 42}

The best fit model parameters are then used for creating a random forest classifier model & predicting the train & test labels based on the independent variable values.

Classification Report on Train & Test data:-

Train data Classification Report

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	1225
1.0	0.84	0.40	0.54	153
accuracy			0.92	1378
macro avg	0.88	0.69	0.75	1378
weighted avg	0.92	0.92	0.91	1378

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	613
1.0	0.70	0.28	0.40	67
accuracy			0.92	680
macro avg	0.82	0.64	0.68	680
weighted avg	0.90	0.92	0.90	680

Table 1.50 Random Forest Classification report

Confusion Matrix:-

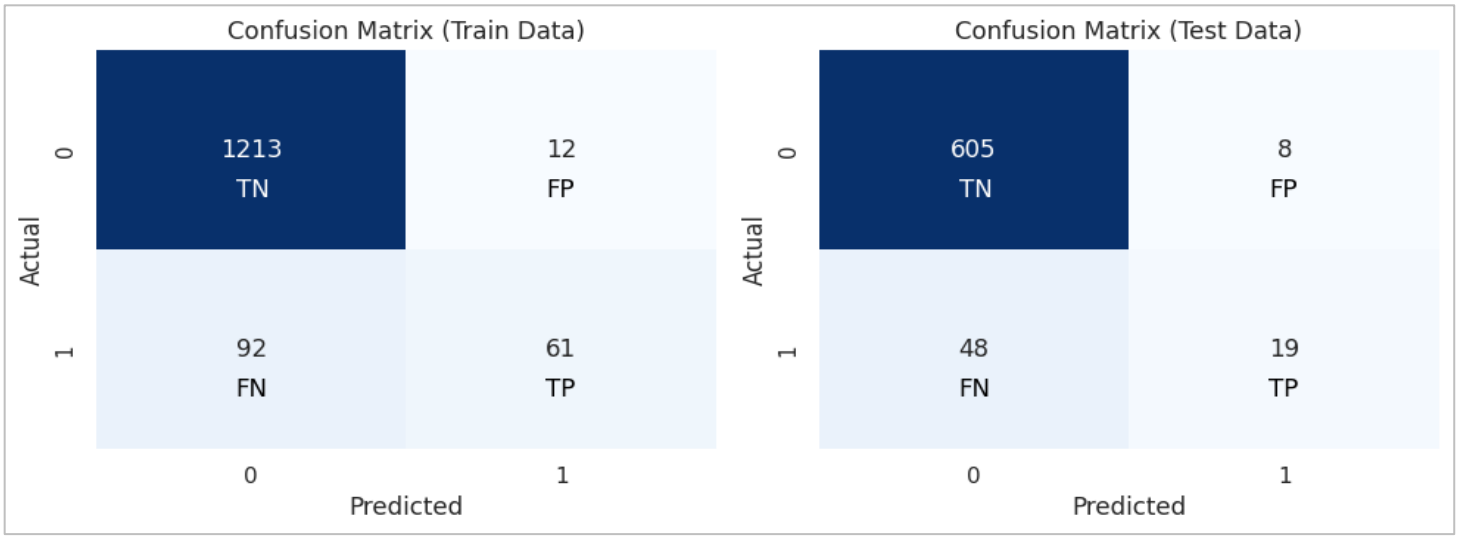


fig 1.49 RandomForest Confusion Matrix

ROC Curve

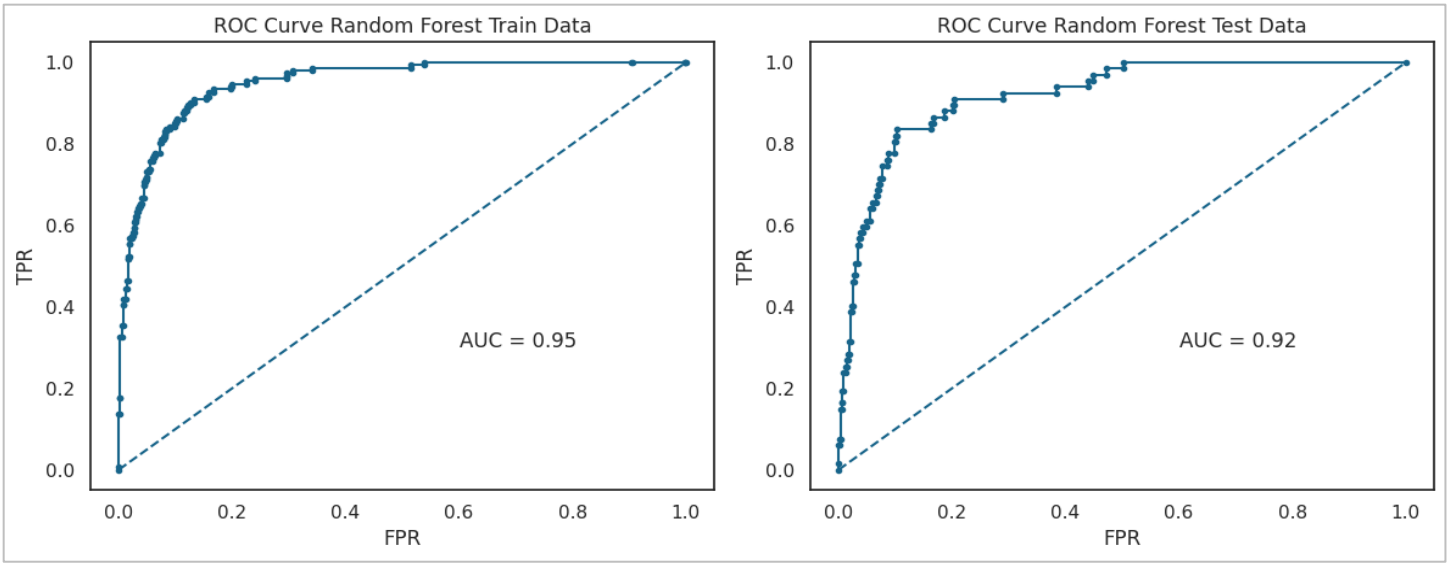


fig 1.50 RandomForest ROC curve

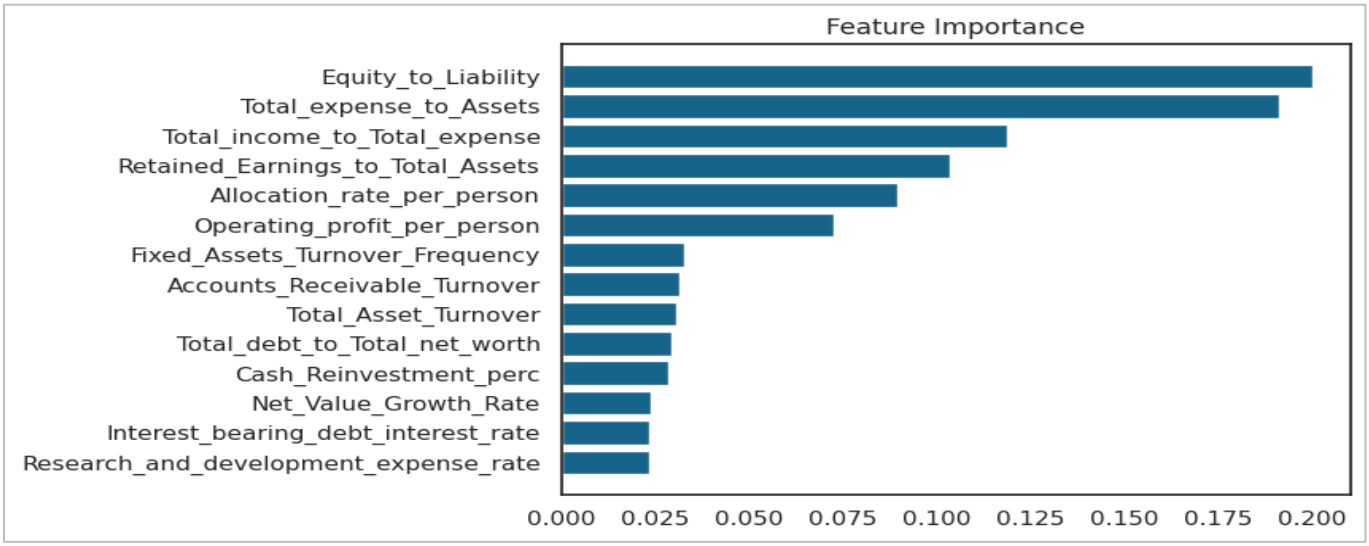


fig 1.51 RandomForest Feature Importance

Random Forest Conclusion

Metrics	Train	Test
True Negative	1213	605
True Positive	61	19
False Negative	92	48
False Positive	12	8
AUC	95%	92%
Accuracy	92%	92%
Precision	84%	70%
Recall	40%	28%
f1-score	54%	40%

Table 1.51 Random Forest Conclusion

- The model shows consistent Precision, Recall, F1-score, and Accuracy scores between the training and testing data. However, it **struggles to accurately predict defaulting companies in both datasets**. The model shows higher AUC and Accuracy as compared to previous models.

- Notably, the model performs relatively **better in predicting the majority class** but **significantly worse in predicting the minority class.**

- Given its weak performance in predicting defaults but strong performance in identifying non-defaulting companies, we can **utilize this model primarily for classifying the non-default category**.

- Key features for classification include **Equity_to_Liability**, **Total_expense_to_Assests, Total_Income_to_Total_Expense.**

Build a LDA Model on Train Dataset. Also showcase your model building approach. Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model

Linear Discriminant Analysis predicts the class in dependent variable using a linear combination of independent variables.

LDA projects the features in higher-dimensional space onto a lower dimensional space. LDA then searches for a linear combination of independent variables (line, plane, or hyperplane) that best separates the classes of the dependent variable.

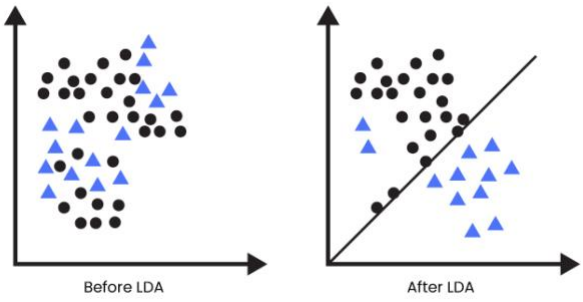


fig 1.52 LDA Explaination

Assumptions of LDA:

1. Each feature in the dataset is a gaussian distribution
2. Each feature has the same variance, the value of each feature varies around the mean with the same amount on average.
3. Each feature is assumed to be randomly sampled.
4. Lack of multicollinearity in independent features. Increase in correlations between independent features and the power of prediction decreases.

Approach:-

We use Grid Search Cross Validation Technique to obtain the best fit LDA model
We began the search for the best fit model with the following values:
Solver: svd, lsqr, eigen
Tolerance (learning rate): 0.001, 0.0001, 0.00001

We obtain the best fit model with the parameters:
Solver = lsqr
Tolerance = 0.1

The best fit model parameters are then used for creating a LDA model & predicting the train & test labels based on the independent variables

Classification Report on Train & Test data:-

Train data Classification Report

	precision	recall	f1-score	support
0.0	0.94	0.96	0.95	1225
1.0	0.64	0.52	0.57	153
accuracy			0.91	1378
macro avg	0.79	0.74	0.76	1378
weighted avg	0.91	0.91	0.91	1378

Test data Classification Report

	precision	recall	f1-score	support
0.0	0.95	0.96	0.95	613
1.0	0.59	0.49	0.54	67
accuracy			0.92	680
macro avg	0.77	0.73	0.75	680
weighted avg	0.91	0.92	0.91	680

Table 1.52 LDA Classification report

Confusion Matrix:-

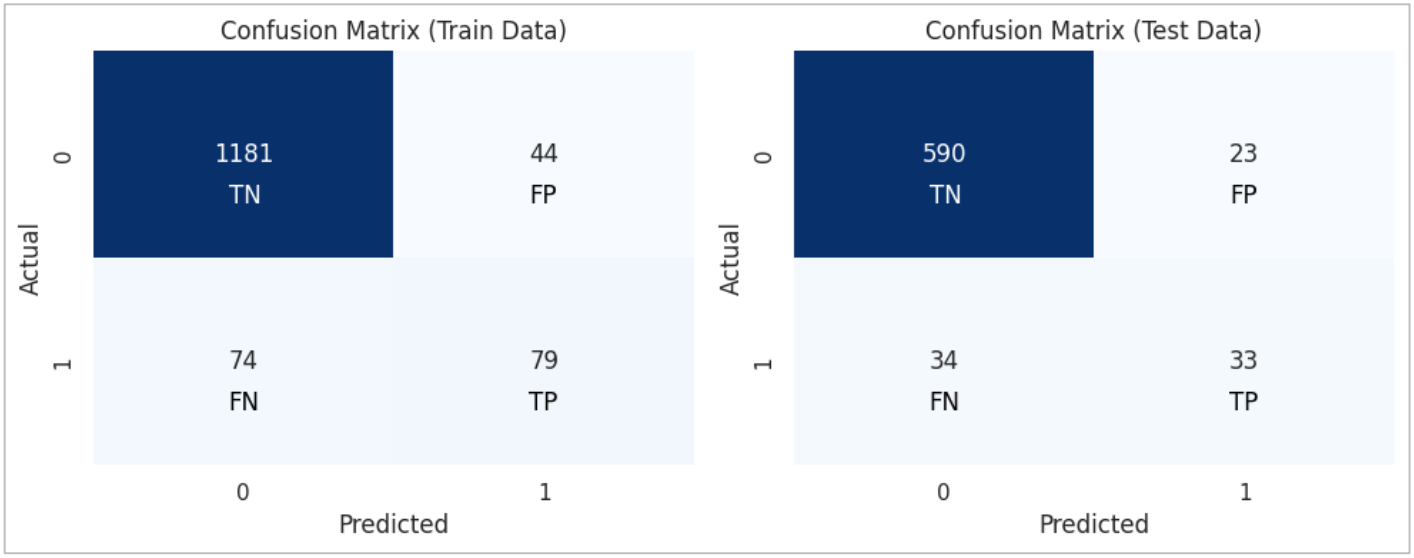


fig 1.53 LDA Confusion Matrix

ROC Curve

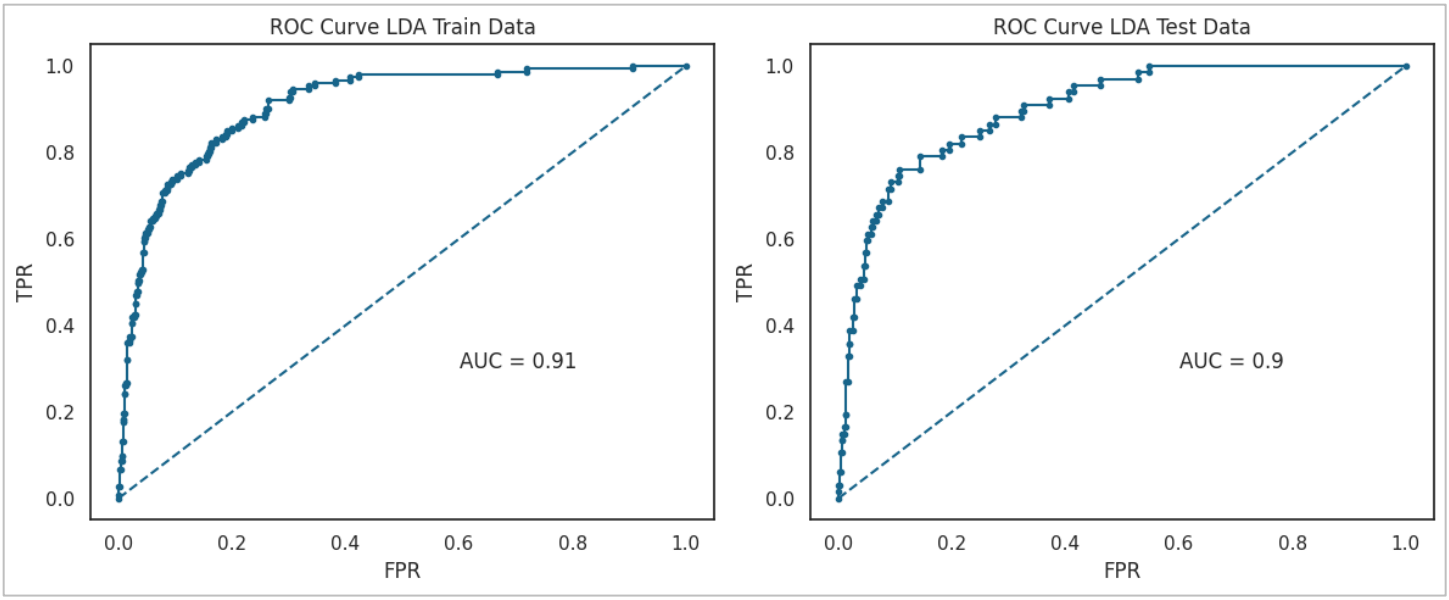


fig 1.54 LDA ROC curve

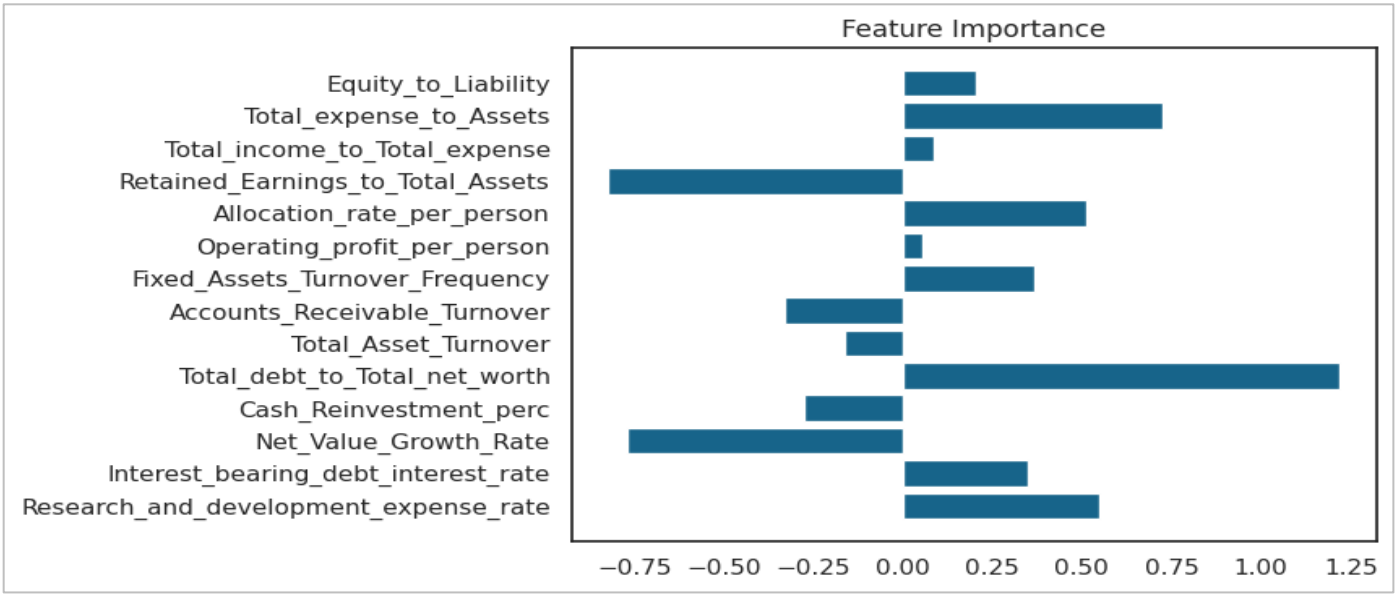


fig 1.55 LDA Feature Importance

LDA Conclusion

Metrics	Train	Test
True Negative	1181	590
True Positive	79	33
False Negative	74	34
False Positive	44	23
AUC	91%	90%
Accuracy	91%	92%
Precision	64%	59%
Recall	52%	49%
f1-score	57%	54%

Table 1.53 LDA Conclusion

The Precision, Recall, F1-score, and Accuracy of the training data closely match those of the testing data. However, the model struggles to effectively identify defaulting companies in both datasets.

- Notably, the model performs relatively better in predicting the majority class (non-default) but exhibits poorer performance in predicting the minority class (default).
- The model performed slightly better than Random forest on test data. However, still not a good model to predict for minority class.
- Given its limited ability to predict defaults but relatively strong performance in identifying non-defaulting companies, it's advisable to employ this model primarily for classifying the non-default category.
- Key metrics such as Precision, Recall, F1-score, and Accuracy are reasonably consistent between training and testing data, suggesting that the model's performance generalizes well to unseen data.
- Key features for classification include **Total_Debit_to_Total_Net_worth, Net_Value_Growth_Rate, Retained_Earnings_to_Total_Earning**

Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve)

Metric Comparison

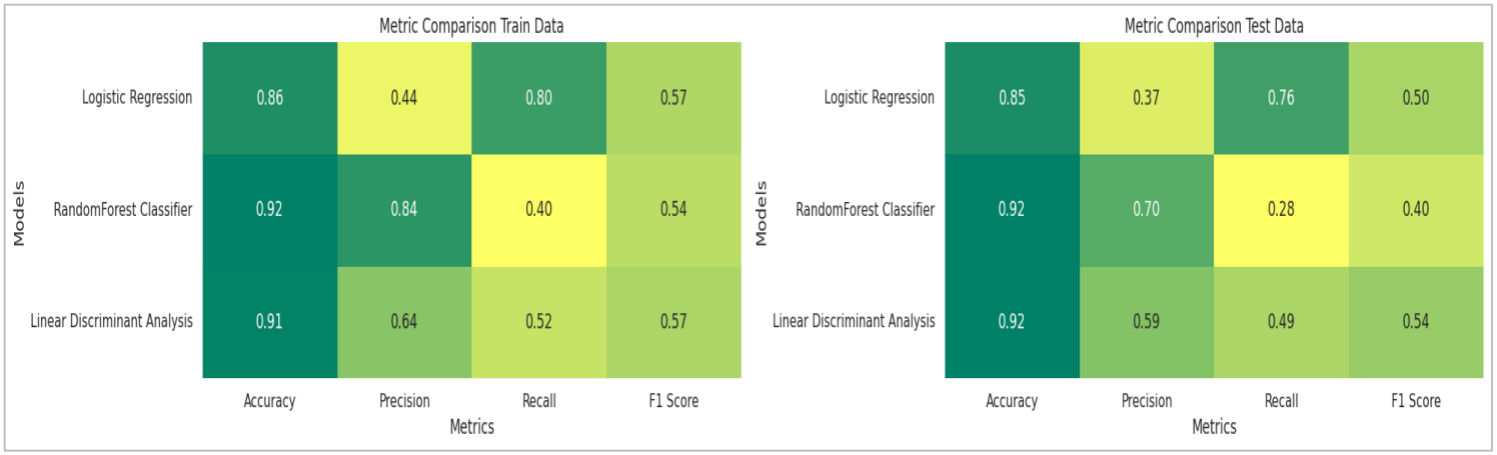


fig 1.56 All Model Comparison Metrics

- **Logistic Regression with SMOTE and Optimum Cut-off 0.367** has the **highest and recall** both in test & train data, indicating that it performs well in identifying Default 1 correctly and has a balanced trade-off between precision and recall.
- RandomForest has the highest accuracy in the test data, but its precision and recall are lower, indicating that it has a higher number of false positives and false negatives. it's advisable to employ this model primarily for classifying the non-default category.
- LDA performs consistently well in both train and test data with balanced precision and recall. However, is not upto the mark in identifying the defaults.

Confusion Matrix on Train and Test Data

(Note: Upper part is Train, lower one is Test)

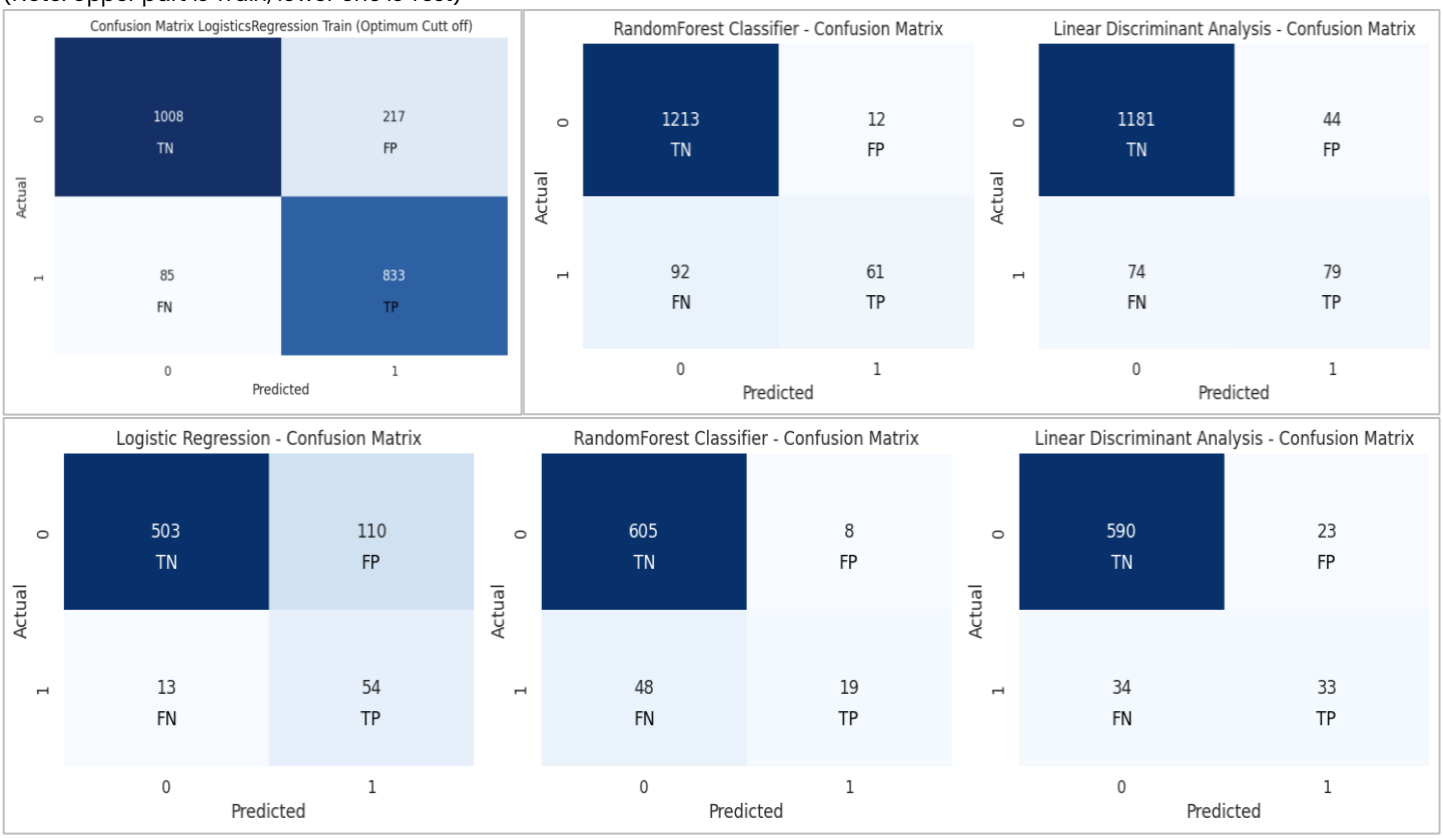


fig 1.57 All Model Comparison Confusion Matrix

- Among the three models, **Logistic Regression with SMOTE and Optimum Cut-off 0.367** performs the best in correctly **identifying Default 1 in both the train and test data, with 833 and 55 true positives**, respectively.
- RandomForest performs better in the train data compared to the test data, indicating that it may be overfitting the training data.
- LDA also performs better in the train data compared to the test data, but it is the second-best model in terms of identifying Default 1 in both datasets.

ROC Curve

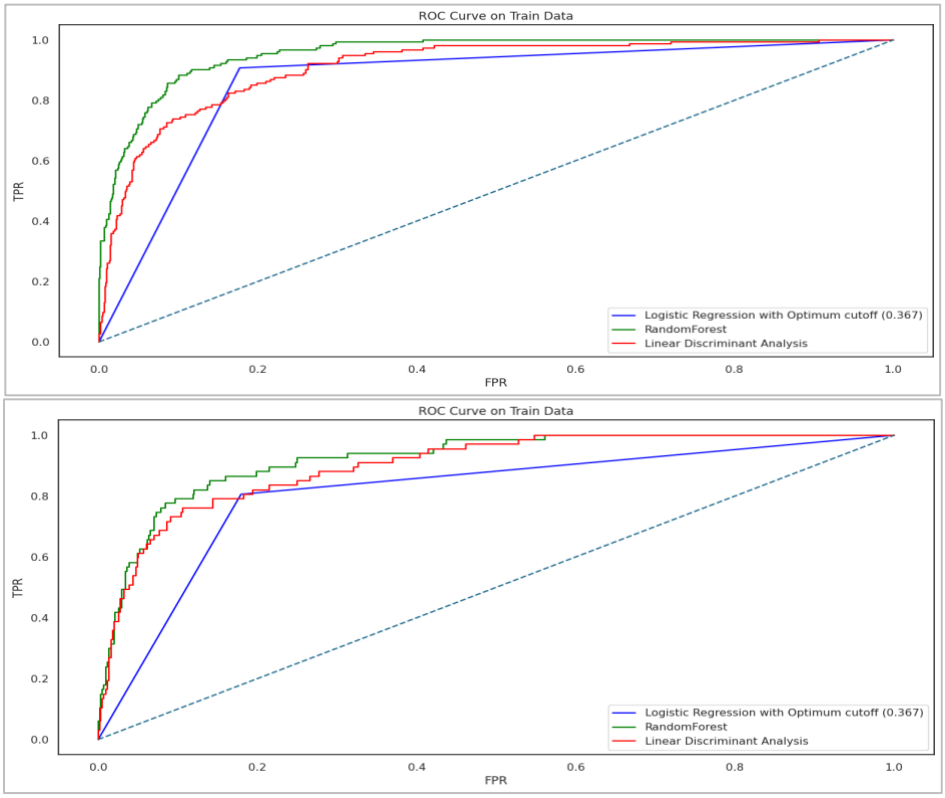


fig 1.58 All Model Comparison ROC curve

- RandomForest has the highest AUC in both train and test data, indicating that it has the best overall discriminatory power for distinguishing between the positive and negative classes.
- Logistic Regression with SMOTE and Optimum Cut-off 0.367 has a lower AUC but still performs reasonably well in distinguishing between the classes.
- LDA has a similar AUC to Logistic Regression with SMOTE but slightly lower performance in test data.

Conclusions and Recommendations

- All the models have approximately similar performance metrics. The models do not perform exceptionally well in predicting which companies will default for both train & test data.

- The models predicts better for the majority class and has a pretty inferior performance for the minority class.
- Since, the performance of the model is too low for the default class but it has a good performance when it comes to predicting the not going to default class. Hence, we can use this model to predict the not going to default class instead of predicting the default class.

- **Logistic Regression with SMOTE and Optimal Cut-off 0.367** stands out as the most effective model for correctly identifying Default 1. It consistently achieves the highest number of true positives in both the training and test datasets, indicating its superior ability to detect actual defaults.

- **RandomForest** demonstrates competitive performance, particularly in terms of AUC, accuracy, and precision. While it excels in true negative predictions, it lags slightly behind Logistic Regression with SMOTE in true positive identifications

- **Linear Discriminant Analysis (LDA),** while still performing reasonably well, falls short of the other two models in several metrics, including precision, recall, and F1-score.

Recommendations

- Since, the performance of the models are low for the defaulting class but it has a good performance when it comes to predicting the non-defaulting class. Hence, we can use this model to predict the non-defaulting class instead of predicting the defaulting class.

- **Prioritize Logistic Regression with SMOTE & Optimal cutoff:** Given its performance in identifying Default 1 with the highest number of true positives, Logistic Regression with SMOTE and Optimum Cut-off 0.367 should be the preferred model for predicting defaults. It offers the best balance between precision and recall.

- **Consider RandomForest:** RandomForest can serve as an alternative when diversity in modeling is needed. It excels in true negative predictions and maintains competitive performance in identifying Default 1. Given its strong performance in identifying non-defaulting companies, we can utilize this model primarily for classifying the non-default category.

- **Data Quality :** There is presence of missing values & outliers in the dataset which was dealt with imputation & capping respectively, which may have affected the model performance. Hence, a better-quality data may have significantly improved the model performance. There also exists high multicollinearity among the variables, meaning they are explaining the same thing. More features need to be added that explain different aspects of accounting as well.

PART B: Market Risk

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.

The Objective is to explore a dataset containing 314 entries, including 10 stock names and a date variable.

About the Dataset

```
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_and_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
```

Table 2.1 Dataset Information

- The dataset consists of 11 columns.
- There are a total of 314 non-null entries in each of the stock price columns, indicating no missing data in these columns.
- All stock price columns contain integer values.
- There are no duplicate values found.

Sample of dataset

'First 5 samples:'

	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

'Last 5 samples:'

	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 2.2 Sample of the Dataset

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Infosys	314.0	511.340764	135.952051	234.0	424.00	466.5	630.75	810.0
Indian_Hotel	314.0	114.560510	22.509732	64.0	96.00	115.0	134.00	157.0
Mahindra_and_Mahindra	314.0	636.678344	102.879975	284.0	572.00	625.0	678.00	956.0
Axis_Bank	314.0	540.742038	115.835569	263.0	470.50	528.0	605.25	808.0
SAIL	314.0	59.095541	15.810493	21.0	47.00	57.0	71.75	104.0
Shree_Cement	314.0	14806.410828	4288.275085	5543.0	10952.25	16018.5	17773.25	24806.0
Sun_Pharma	314.0	633.468153	171.855893	338.0	478.50	614.0	785.00	1089.0
Jindal_Steel	314.0	147.627389	65.879195	53.0	88.25	142.5	182.75	338.0
Idea_Vodafone	314.0	53.713376	31.248985	3.0	25.25	53.0	82.00	117.0
Jet_Airways	314.0	372.659236	202.262668	14.0	243.25	376.0	534.00	871.0

Table 2.3 Descriptive Statistics

- Shree Cement and Jet Airways have the widest price ranges, indicating significant price fluctuations over the years.
- Sun Pharma, Axis Bank, and Infosys also show substantial price volatility.
- Indian Hotel and Jindal Steel have relatively stable price movements compared to other stocks.
- Stocks like Idea_Vodafone and SAIL have moderate volatility with a relatively narrow price range.
- Shree Cement has the highest average stock price, while Idea-Vodafone has the lowest average stock price among the analyzed stocks.

Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference

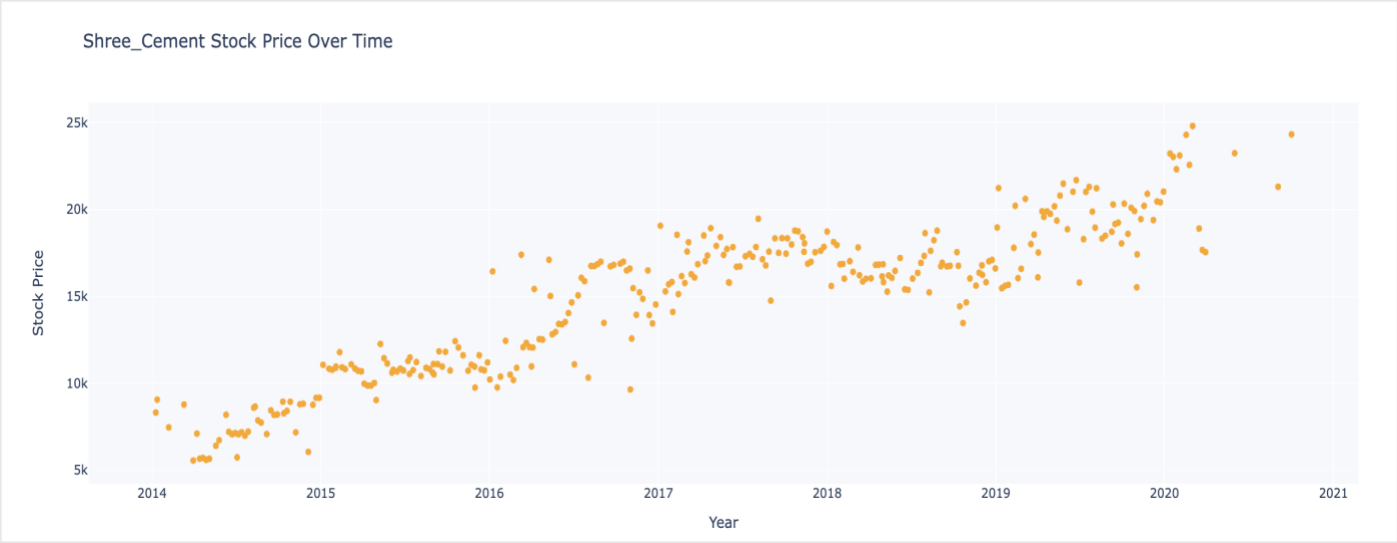


Fig 2.1 Stock Price vs Time Graph for Shree Cement

- The Prices of Shree Cement has shown an increasing Trend over the past 6 years
- Lowest Price recorded in the 6 years was 5543 on 31-03-2014
- Highest Price of Shree_Cement was 24806 on 03-02-2020
- This stock has the highest average stock price of 14806.41

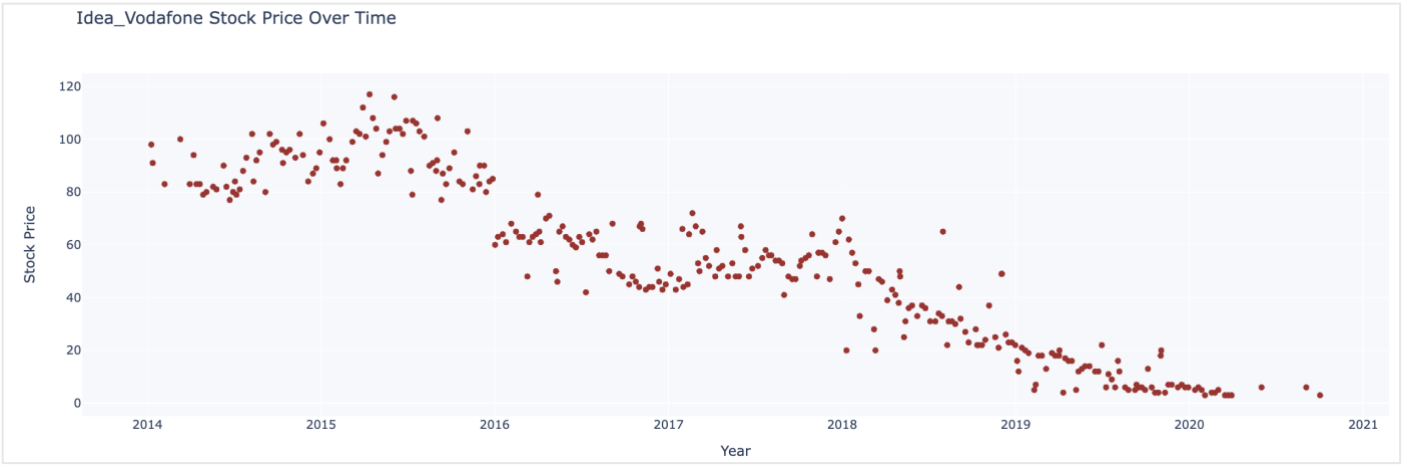


Fig 2.2 Stock Price vs Time Graph for Idea-Vodafone

- The Prices of Idea-Vodafone has shown an decreasing Trend over the past 6 years
- Lowest Price recorded in the 6 years was 3 on 10-02-2020
- Highest Price of Shree_Cement was 117 on 13-04-2015
- Idea_Vodafone has the lowest average stock price of 53.71 among the analyzed stocks

Calculate Returns for all stocks with inference

To calculate the returns form price, we have taken Logarithms and their differences. Returns is the difference between two consecutive week prices for the stock

Note: We have dropped the date column, hence the updated dataset with Stock Returns contains only 10 columns with Stock Names as header and 314 rows.

The negative value refers to decrease in price compared to previous week and vice versa.

	Infosys	Indian_Hotel	Mahindra_and_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 2.4 Sample of the Stock Returns

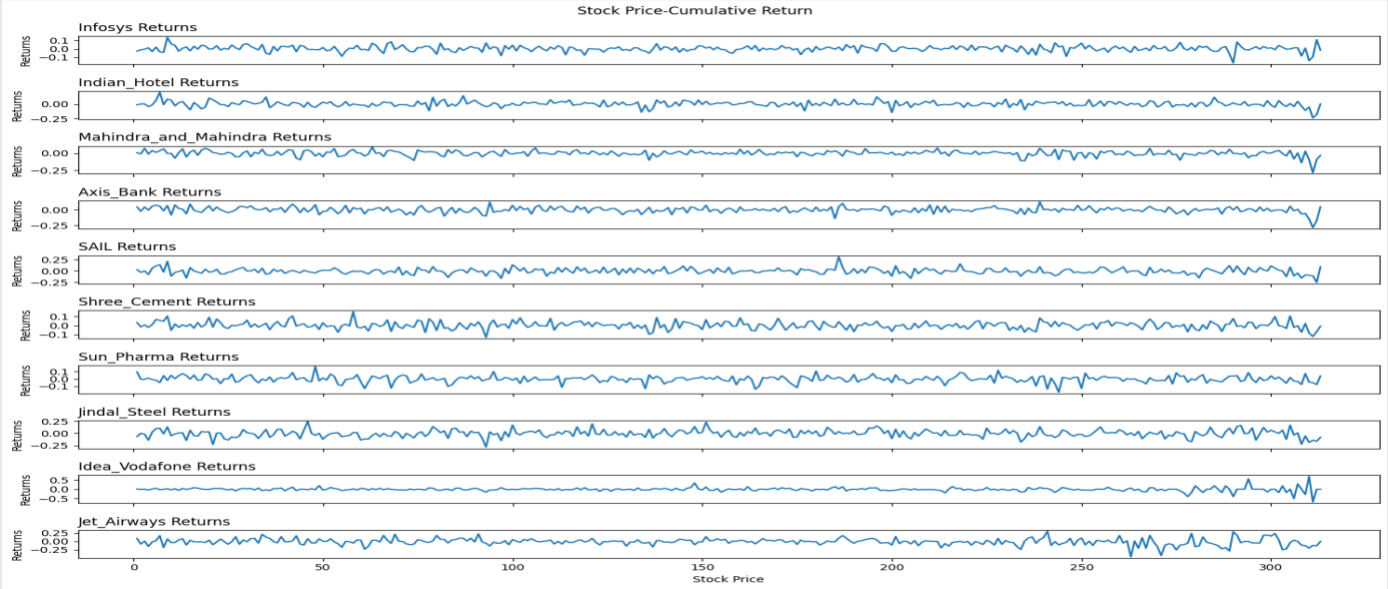


Fig 2.3 Stock Returns & Price Trend

Descriptive Statistics of Stock Returns

Note: The prices are not rounded off due to low values in decimals..

	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_and_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 2.5 Sample of the Stock Returns

- The above stock returns covers the weekly returns of different stocks over a period of time.
- Each stock displays varying levels of average returns, volatility, and ranges of returns.

- Stocks like **Shree Cement and Axis Bank** have relatively **favourable average returns**.
- **Idea Vodafone** and **Jet Airways** might have faced significant challenges, resulting in **substantial negative returns.**

Calculate Stock Means and Standard Deviation for all stocks with inference

Stock Means: Stock Means represent the average returns a stock generates on a week-to-week basis
Stock Standard Deviation: serves as a gauge of volatility. In essence, the higher the deviation of a stock's returns from its average return, the greater its level of volatility

	Average	Volatility
Shree_Cement	0.003681	0.039917
Infosys	0.002794	0.035070
Axis_Bank	0.001167	0.045828
Indian_Hotel	0.000266	0.047131
Sun_Pharma	-0.001455	0.045033
Mahindra_and_Mahindra	-0.001506	0.040169
SAIL	-0.003463	0.062188
Jindal_Steel	-0.004123	0.075108
Jet_Airways	-0.009548	0.097972
Idea_Vodafone	-0.010608	0.104315

Table 2.6 Stock Mean and Volatility

- **Shree Cement** has the **highest average stock price**, while **Idea-Vodafone has the lowest average stock price** among the analyzed stocks.
- **Idea_Vodafonehas** the **highest volatility** while **Infosys has the lowest**

Draw a plot of Stock Means vs Standard Deviation and state your inference

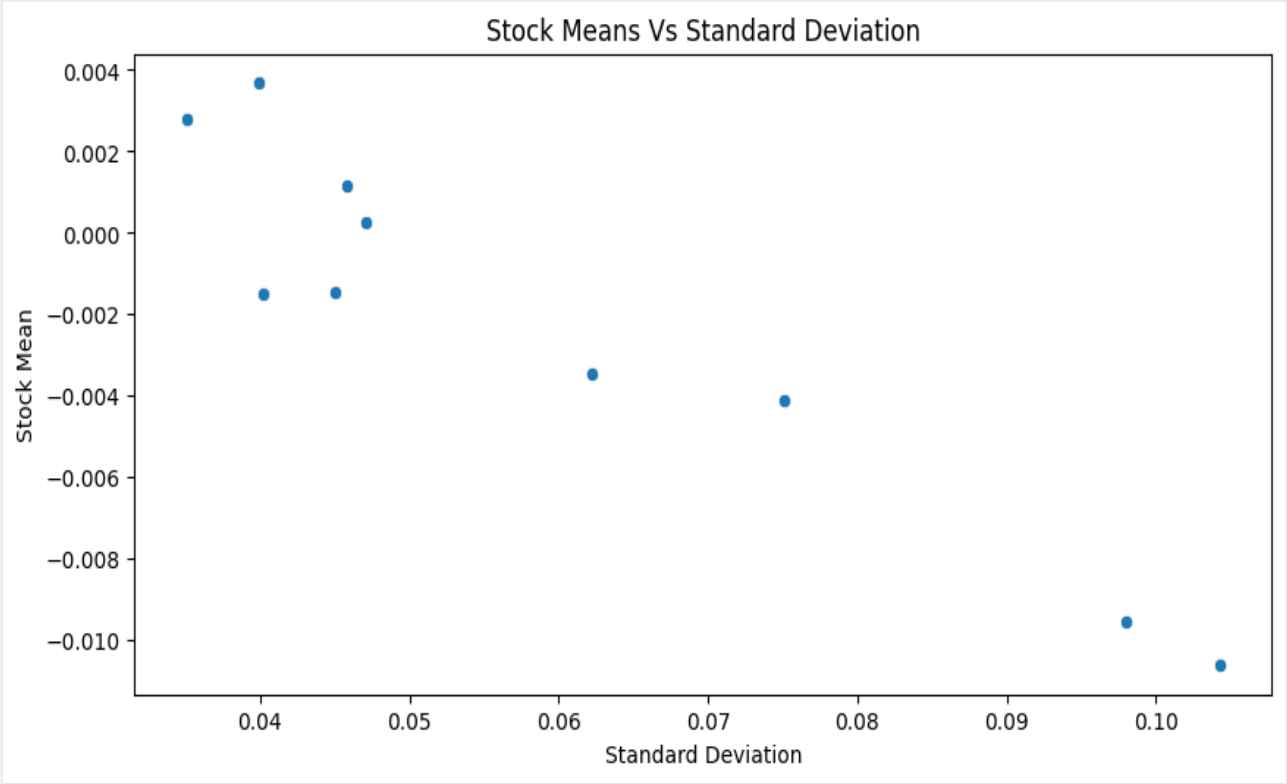


Fig 2.4 Stock Means vs Standard Deviation

Labelling each of the datapoints to identify the stocks

Fig 2.5 Stock Means vs Standard Deviation with Stock Names

Idea Vodafone: This stock has the **highest negative average return (-1.06%)** among the listed stocks. It also exhibits the **highest volatility (10.43%),** indicating **substantial price fluctuations**. The negative average suggests a **consistent downward trend.**

Jet Airways: While having a **negative average return (-0.95%),** Jet Airways shows **slightly lower volatility (9.80%) compared to Idea Vodafone**. This implies that although the stock is declining, it **experiences relatively smaller price swings**.

Jindal Steel: Jindal Steel also has a **negative average return (-0.41%)** but with **higher volatility (7.51%).** This suggests that the **stock has fluctuations**, but the downward trend is not as pronounced as Idea Vodafone or Jet Airways.

SAIL: Steel Authority of India Limited (SAIL) has a **negative average return (-0.35%)** and moderate **volatility (6.22%).** While it **exhibits some price fluctuations**, they are relatively milder compared to highly volatile stocks.

Indian Hotel: Indian Hotel demonstrates a slightly **positive average return (0.03%)** with **moderate volatility (4.71%)**. This indicates that the stock has **relatively stable performance**, with minor price swings.

Axis Bank: Axis Bank showcases a **positive average return (0.12%)** and relatively **moderate volatility (4.58%).** This suggests that the stock has a **positive trend with limited price fluctuations**.

Sun Pharma: Sun Pharma has a **negative average return (-0.15%)** with **moderate volatility (4.50%).** It indicates a **declining trend with relatively stable price movements**.

Mahindra and Mahindra: This stock has a **negative average return (-0.15%)** and **with relatively low (4.02%) than previous stocks.** It suggests that the stock **experiences price fluctuations, but the overall trend is negative**.

Shree Cement: Shree Cement has a **positive average return (0.37%) with relatively low volatility (3.99%) than previous stocks.** It shows a **positive trend with relatively stable price movements**.

Infosys: Infosys has a positive **average return (0.28%) and the lowest volatility (3.51%)** among the listed stocks. This indicates a **consistent positive trend with minimal price fluctuations**.

Conclusion:

The stocks with **negative average returns** (**Idea Vodafone, Jet Airways, Jindal Steel, SAIL, Mahindra and Mahindra**) generally have **higher volatility**, suggesting a **downward trend with significant price swings.**

On the other hand, **stocks with positive average returns** (**Axis Bank, Indian Hotel, Shree Cement, Infosys**) tend to have **lower volatility**, indicating **more stable and positive trends**.

Recommendation

Among the listed stocks, **Infosys and Shree Cement** exhibit the **most favourable combination of high average returns and low volatility**. Hence, stocks with **higher returns relative to their risk level are considered better among the available options.**

For investors seeking a **balance between returns and risk**, we recommend considering stocks from the mid-range performers, such as **Axis Bank, Indian Hotel, Sun Pharma, and Mahindra and Mahindra**. These stocks exhibit **moderate returns with relatively stable price movements.**

On the contrary, it's essential to exercise caution with stocks like **Jet Airways and Idea Vodafone**, which have the **highest losses and significant price fluctuations**. These stocks **may not be suitable for conservative investors**.

Diversifying investments across various stocks can also help spread risk and optimize returns.