FINANCE AND RISK ANALYSIS

PROJECT REPORT

by: PRADEEP PAL

TABLE OF CONTENTS

Credit Risk Problem

- Outlier Treatment 10
- Missing Value Treatment 19
- Univariate & Bivariate analysis with proper interpretation. 22
- Train Test Split 31
- 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 32
- Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model 37
- Build a Random Forest Model on Train Dataset. Also showcase your model building approach 40
- Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model 43

- Build a LDA Model on Train Dataset. Also showcase your model building approach 46
- Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model 49
- Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve) 52
- Conclusions and Recommendations 57

Market Risk Problem

- Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference 31
- Calculate Returns for all stocks with inference 68
- Calculate Stock Means and Standard Deviation for all stocks with inference 72
- Draw a plot of Stock Means vs Standard Deviation and state your inference 76
- Conclusions and Recommendations 78

LIST OF TABLES

Table 1.1: Top rows and columns of Dataset	10
Table 1.2: Shape of dataset	10
Table 1.3: Information about Data types of dataset	11-12
Table 1.4: Statistical summary of dataset	13-14
Table 1.5: Columns having missing values present	14
Table 1.6: Duplicated row's sum	15
Table 1.7: Number of Defaulters	15
Table 1.8: Missing values in dataset	19
Table 1.9: Standardising the data	20
Table 1.10: Data after imputation	21
Table 1.11: Count of null values after imputation	21
Table 1.12: Shape of train and test data	31
Table 1.13: Top 5 rows of train data	31
Table 1.14: Top 5 rows of test data	31
Table 1.15: VIF values of variables	32-33
Table 1.16: Logistic regression summary table	33
Table 1.17: Characteristic table for RF train data	42
Table 1.18: Characteristic table for RF test data	43
Table 1.19: Characteristic table for LDA of train data	46
Table 1.19: Characteristic table for LDA of train data	48
Table 1.20: Characteristic table for LDA of test data	49

Table 2.1: Top five rows of the data 61
Table 2.2: Shape the data 61
Table 2.3: Datatype information about data 62
Table 2.4: Sum of duplicated rows in data 62
Table 2.5: Statistical summary of numeric columns 63
Table 2.6: Data frame with change in stock price 68
Table 2.7: Average and Volatility of stocks 72

LIST OF FIGURES

Figure 1.1: Outliers present in columns 16
Figure 1.2: Box plot after outlier treatment 18
Figure 1.3: Univariate analysis of significant features 22-25
Figure 1.4: Correlation heat map 26
Figure 1.5: Pair Plot 28
Figure 1.6: Continuous variables vs Default Boxplot 29-30
Figure 1.7: CM for 0.5 threshold on train data 35
Figure 1.8: CM for 0.106 threshold on train data 36
Figure 1.9: CM for 0.5 threshold on test data 37
Figure 1.10: CM for 0.106 threshold on test data 38
Figure 1.11: ROC curve for test data 39
Figure 1.12: Confusion matrix of RF for train data 41
Figure 1.13: Confusion matrix of RF for test data 44

Figure 1.14: ROC curve of RF for test data 44
Figure 1.15: Confusion matrix for LDA of train data 47
Figure 1.16: Confusion matrix for LDA of test data 50
Figure 1.17: ROC curve for LDA of test data 50
Figure 1.18: ROC curve for Logistic Regression of test data 52
Figure 1.19: ROC curve for Random Forest of test data 53
Figure 1.20: ROC curve for LDA of test data 54
Figure 2.1: Stock price graph for Infosys 64
Figure 2.2: Stock price graph for Idea_Vodafone 66
Figure 2.3: Stock Mean vs Standard deviation 76

A: 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.

Data Dictionary

Column Name	Description
Co_Code	Company Code
Co_Name	Company Name
_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.
_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.
_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.
Interest_bearing_debt_interest_rate	Interest-bearing debt interest rate: Interest-bearing Debt/Equity
_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.
_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
_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.
_Realized_Sales_Gross_Profit_Growth_Rate	Realized Sales Gross Profit Growth Rate.
_Operating_Profit_Growth_Rate	Operating Profit Growth Rate: Operating Income Growth. It is the rate of increase in operating income over the last year.
_Continuous_Net_Profit_Growth_Rate	Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth
_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
Net_Value_Growth_Rate	Net Value Growth Rate: Total Equity Growth
Total_Asset_Return_Growth_Rate_Ratio	Total Asset Return Growth Rate Ratio: Return on Total Asset Growth
_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.
_Current_Ratio	Current Ratio. The current ratio describes the relationship between a company's assets and liabilities
_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.
_Interest_Expense_Ratio	Interest Expense Ratio: Interest Expenses/Total Revenue
_Total_debt_to_Total_net_worth	Total debt/Total net worth: Total Liability/Equity Ratio
_Long_term_fund_suitability_ratio_A	Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets
_Net_profit_before_tax_to_Paid_in_capital	Net profit before tax/Paid-in capital: Pretax Income/Capital
_Total_Asset_Turnover	Total Asset Turnover. Net Sales/Average Total Assets
_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.
_Average_Collection_Days	Average Collection Days: Days Receivable Outstanding
_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.
_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.
_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.
_Operating_profit_per_person	Operating profit per person: Operation Income Per Employee
_Allocation_rate_per_person	Allocation rate per person: Fixed Assets Per Employee
_Quick_Assets_to_Total_Assets	Quick Assets/Total Assets
_Cash_to_Total_Assets	Cash/Total Assets
_Quick_Assets_to_Current_Liability	Quick Assets/Current Liability

Data Dictionary

_Cash_to_Current_Liability	Cash/Current Liability
_Operating_Funds_to_Liability	Operating Funds to Liability
_Inventory_to_Working_Capital	Inventory/Working Capital
_Inventory_to_Current_Liability	Inventory/Current Liability
_Long_term_Liability_to_Current_Assets	Long-term Liability to Current Assets
_Retained_Earnings_to_Total_Assets	Retained Earnings to Total Assets
_Total_income_to_Total_expense	Total income/Total expense
_Total_expense_to_Assets	Total expense/Assets
_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.
_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.
_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.
_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.
_Cash_Flow_to_Total_Assets	Cash Flow to Total Assets. This ratio indicates the cash a company can generate in relation to its size.
_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)
_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.
_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.
_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.
_Liability_Assets_Flag	Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise
_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.
_No_credit_Interval	No-credit Interval
_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.
_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.
_Net_Income_Flag	Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise
_Equity_to_Liability	Equity to Liability Ratio.
Default	Whether the Company has Default (Bankrupted) or not? 1 - Defaulted, 0 - Not Defaulted.

1. Perform Outlier Treatment

To perform outlier treatment, we first need to know details about the data set. Below is glimpse of data set.

	0	1	2	3	4
Co_Code	16974	21214	14852	2439	23505
Co_Name	Hind.Cables	Tata Tele. Mah.	ABG Shipyard	GTL	Bharati Defence
_Operating_Expense_Rate	8820000000.00	9380000000.00	3800000000.00	6440000000.00	3680000000.00
_Research_and_development_expense_rate	0.00	4230000000.00	815000000.00	0.00	0.00
_Cash_flow_rate	0.46	0.46	0.45	0.46	0.46
_Interest_bearing_debt_interest_rate	0.00	0.00	0.00	0.00	0.00
_Tax_rate_A	0.00	0.00	0.00	0.01	0.40
_Cash_Flow_Per_Share	0.32	0.32	0.30	0.32	0.33

Table 1.1: Top rows and columns of Dataset

Shape of data set is given below

```
The number of rows (observations) is 2058
The number of columns (variables) is 58
```

Table 1.2: Shape of dataset

Let's now examine the datatype information of the variables.

```
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)

Table 1.3: Information about Data types of dataset

The dataset comprises of 57 variables with numeric datatype and one variable with object datatype.

Now, let's examine the 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

Table 1.4: Statistical summary of dataset

The table above illustrates considerable disparity in the values of various columns.

Missing values in data set

_Cash_Flow_Per_Share	167
_Total_debt_to_Total_net_worth	21
_Cash_to_Total_Assets	96
_Current_Liability_to_Current_Assets	14

Table 1.5: Columns having missing values present

Four variables within the dataset contain missing values, accounting for 0.25% of the overall data.

There are no duplicated rows.

```
The sum of duplicate rows :0
```

Table 1.6: Duplicated row's sum

Now checking the target column 'Default'.
Defaulters = 220
Non-defaulters = 1838
This implies that 11% of companies have defaulted.

```
Number of defaulters:
1:Defaulted
0:Not defaulted
0    1838
1     220
```

```
Number of defaulters ratio:
1:Defaulted
0:Not defaulted

0   0.89
1   0.11
```

Table 1.7: Number of Defaulters

Let's examine the outliers that exist in the various columns.

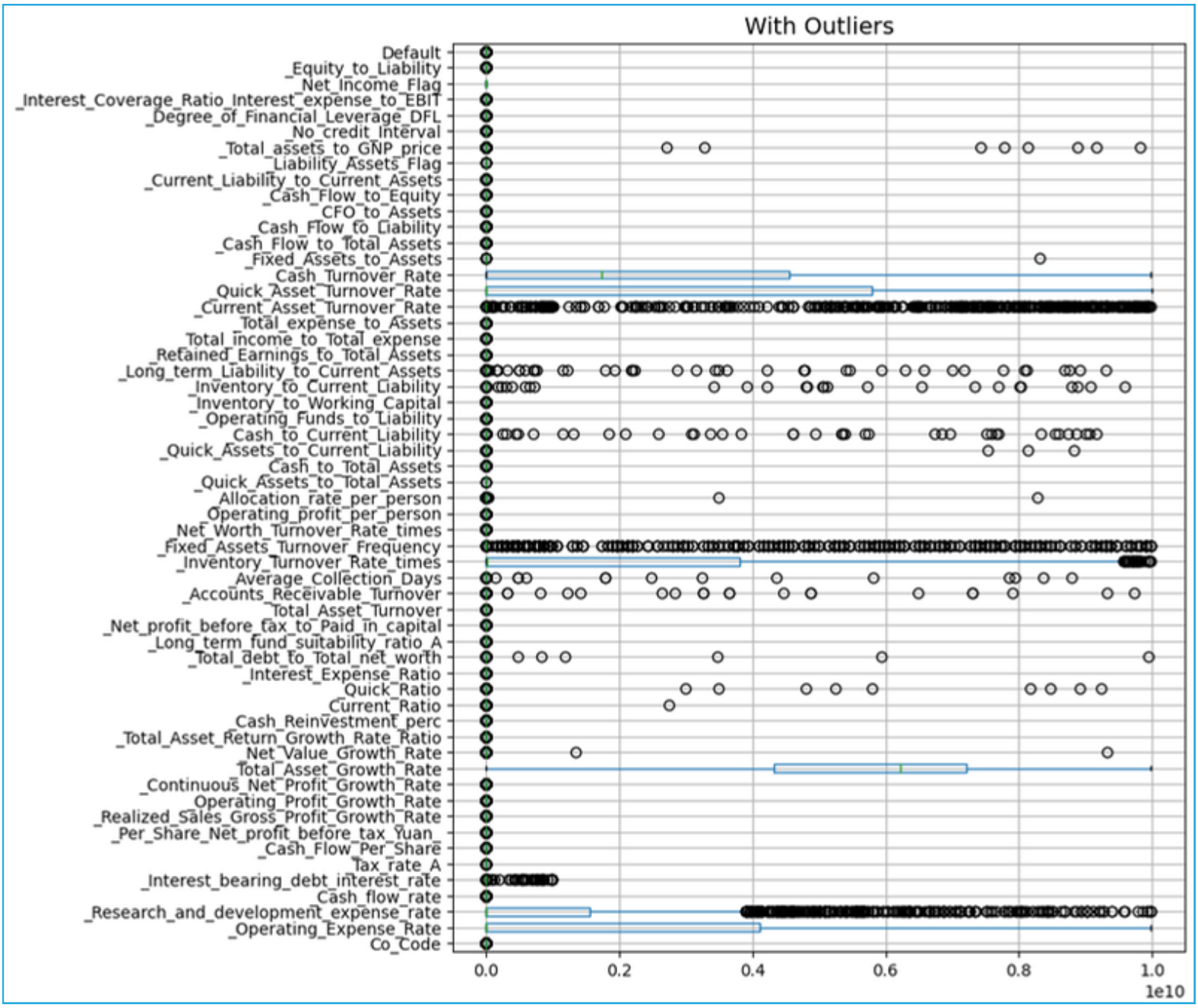


Figure 1.1: Outliers present in columns

Outlier above are identified using Inter-Quartile Rule (IQR). The IQR method defines outliers as data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR, where Q1 is the first quartile (25 percentile) and Q3 is the third quartile (75 percentile).

In order to handle outliers, we are converting the outliers to NaN values. Later, we will impute values to these NaN values. This approach offers multiple benefits, as opposed to merely dropping the outliers.

1. Converting outliers to NaN allows you to keep the integrity of your original dataset.
2. After converting outliers to NaN, we can choose from various imputation techniques to fill in the missing values, depending on the nature of your data.
3. Converting outliers to NaN and imputing values can reduce the influence of outliers on model training, making the model more robust and less sensitive to extreme values.

After outliers treatment.

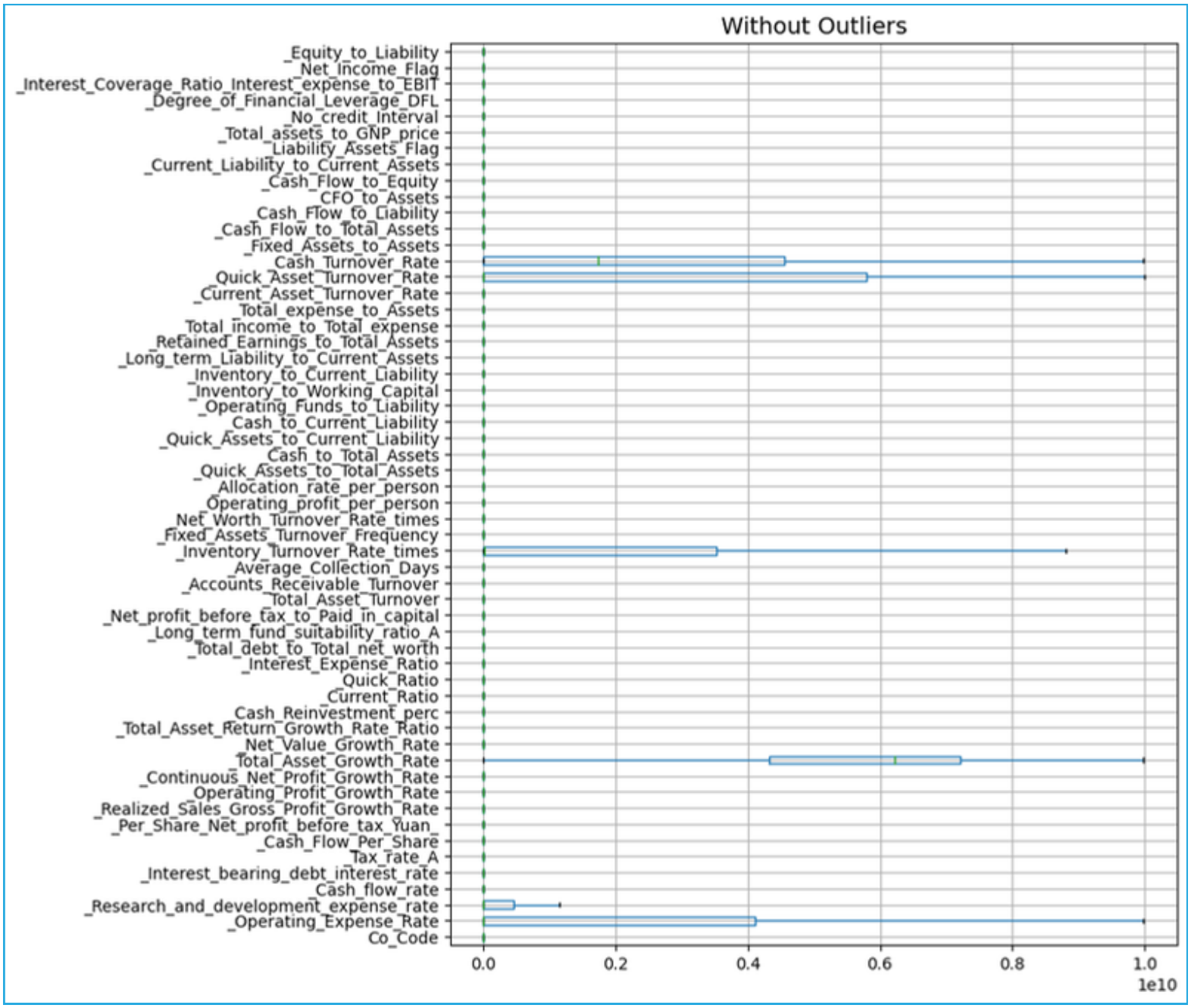


Figure 1.2: Box plot after outlier treatment

2. Missing Value Treatment

Lets find out how many missing/null values present in dataset

```
Total number of Null values present: 11155
```

```
Percentage of Null values present: 9.35%
```

Table 1.8: Missing values in dataset

To address missing or null values, we will utilize imputation techniques.

Some of the popular imputation techniques are below:

- Mean, Median, or Mode Imputation
- Linear Interpolation
- Polynomial Interpolation
- K-Nearest Neighbors (KNN) Imputation

Out of the techniques mentioned above, we opted to use the K-Nearest Neighbor (KNN) imputer for treating missing or null values in our model. This is due to its unique ability to preserve the data structure, handle mixed data types, and capture local relationships, all of which are critical factors in our data analysis.

To ensure the accuracy of our inputs, we must standardize the data. Therefore, we will be utilizing the StandardScaler method.

_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A
2.08	-0.52	-0.31	-0.38	-0.83
2.25	NaN	-0.64	1.07	-0.84
0.54	0.33	-2.41	0.19	-0.84
1.35	-0.52	-0.19	0.57	-0.76
0.50	-0.52	-0.12	1.33	2.47

Table 1.9: Standardising the data

Here are some rows after imputation of data. We have taken value for n_neighbours as 25.

_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A
2.08	-0.52	-0.31	-0.38	-0.83
2.25	-0.52	-0.64	1.07	-0.84
0.54	0.33	-2.41	0.19	-0.84
1.35	-0.52	-0.19	0.57	-0.76
0.50	-0.52	-0.12	1.33	2.47

Table 1.10: Data after imputation

Now lets check the count of missing/null values.

```
Total number of Null values after imputation: 0
```

Table 1.11: Count of null values after imputation

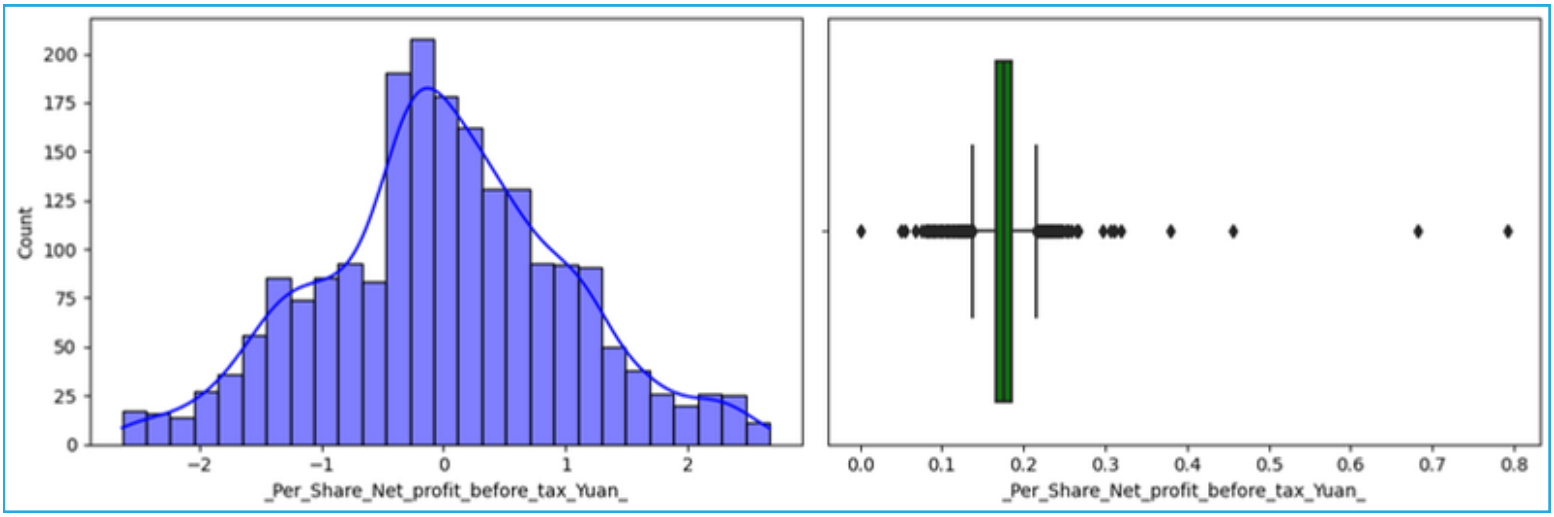
Now we can use this data for further analysis.

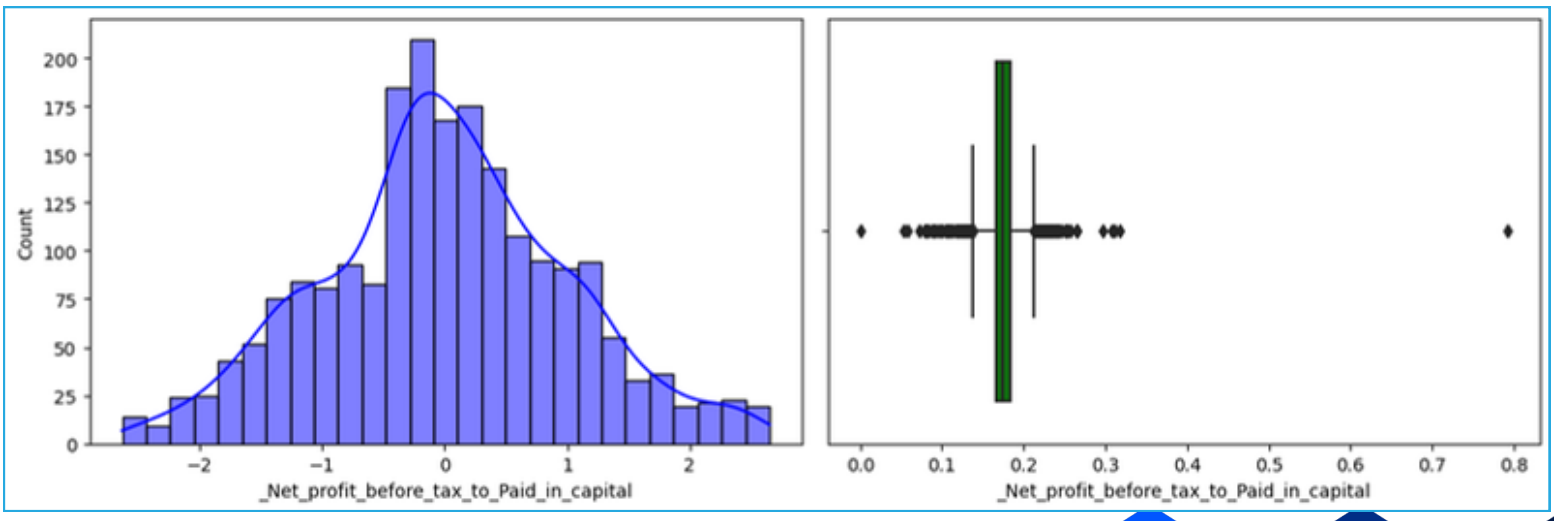
3. Univariate and Bivariate analysis

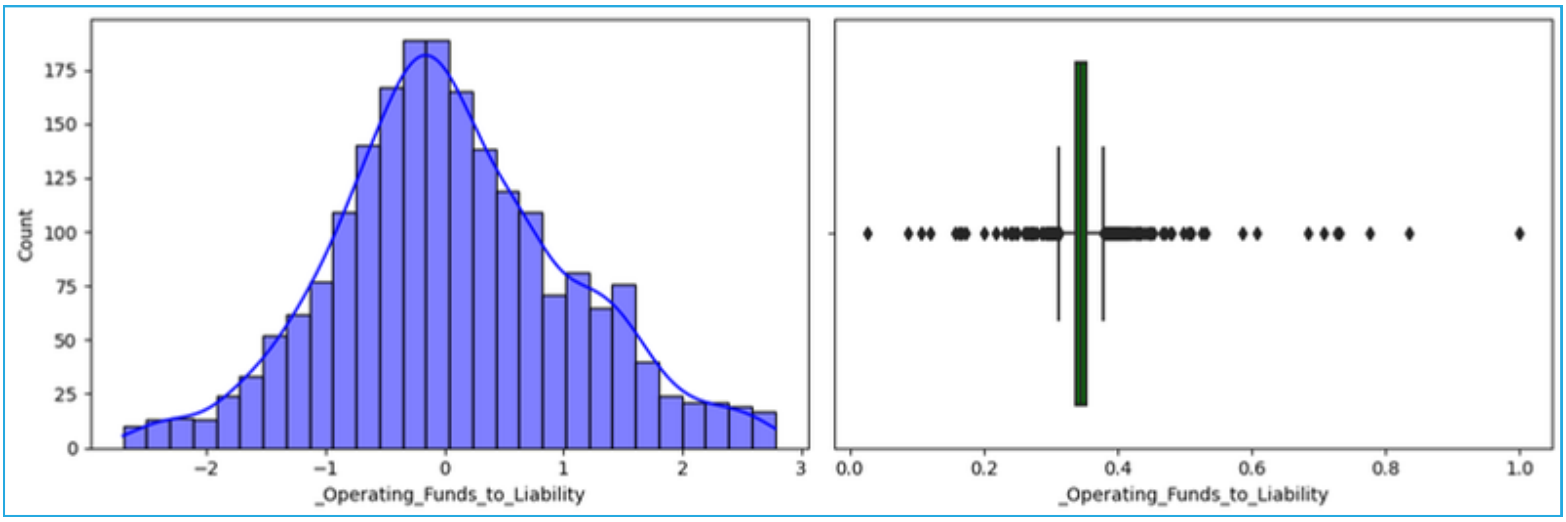
Univariate Analysis

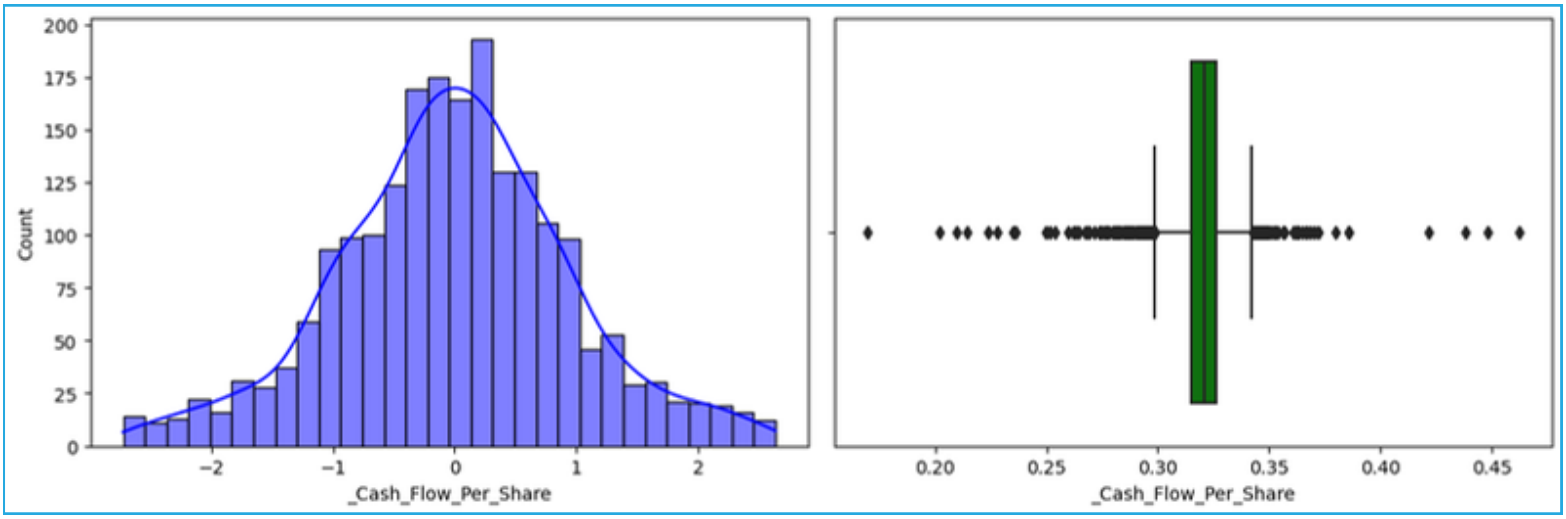
Univariate analysis is a statistical method used to analyze and describe the distribution, central tendency, and variability of a single variable or a single feature in a dataset.

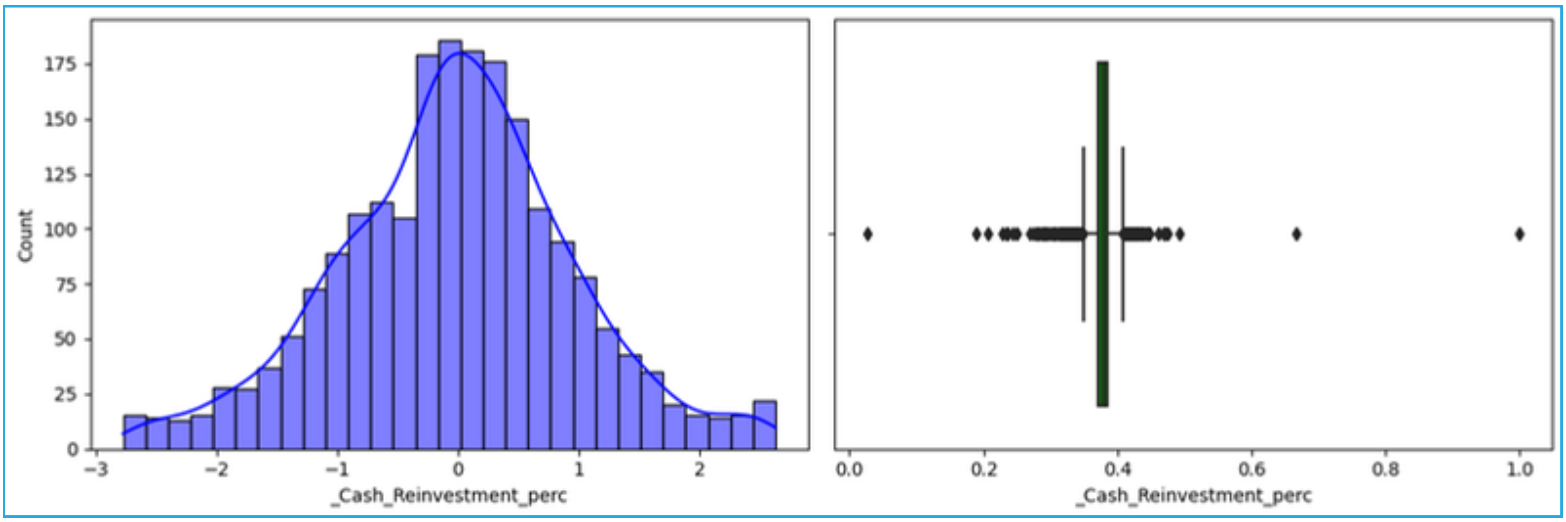
Below is univariate analysis for few significant parameters

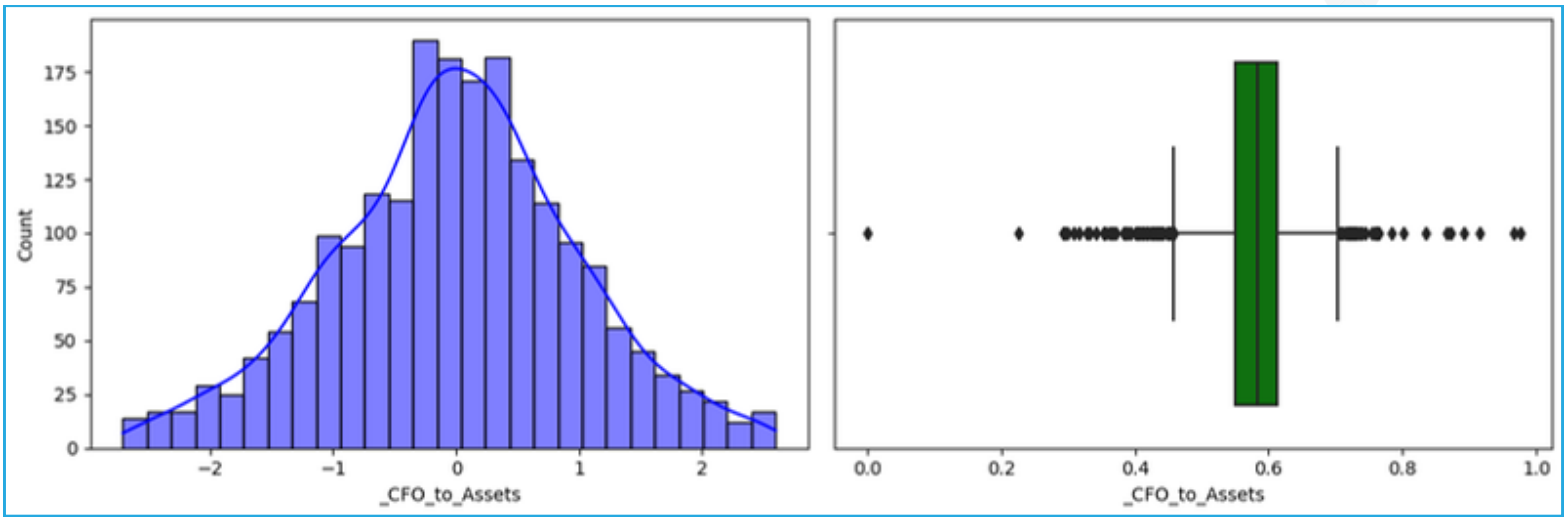


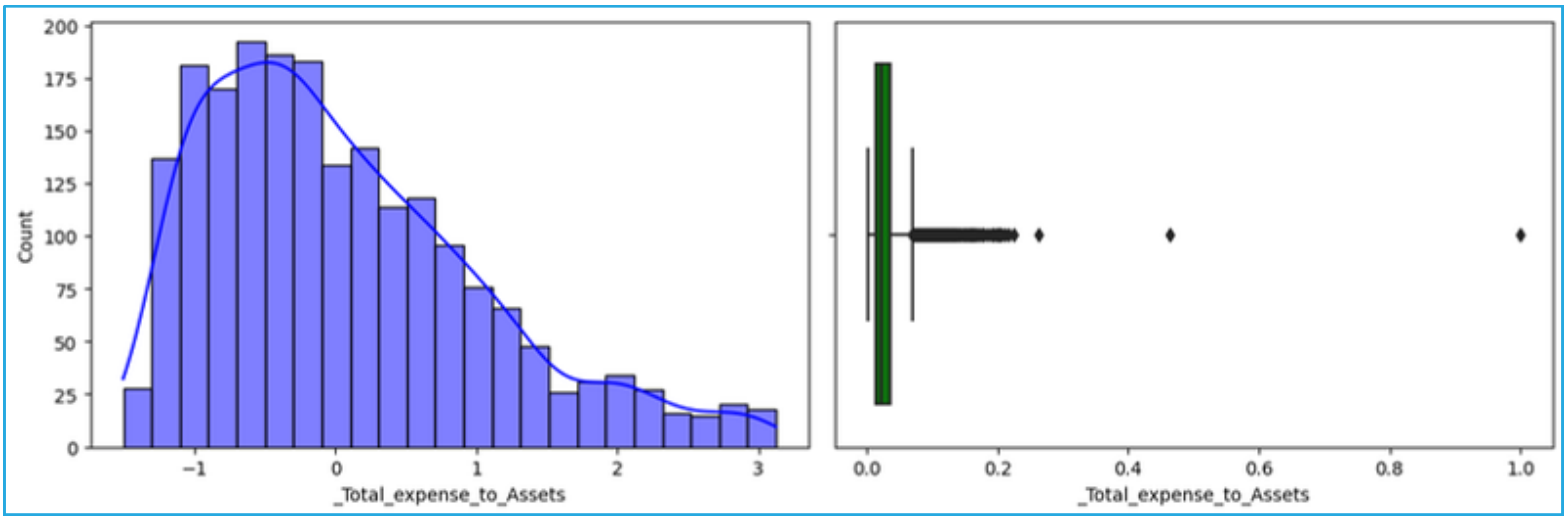


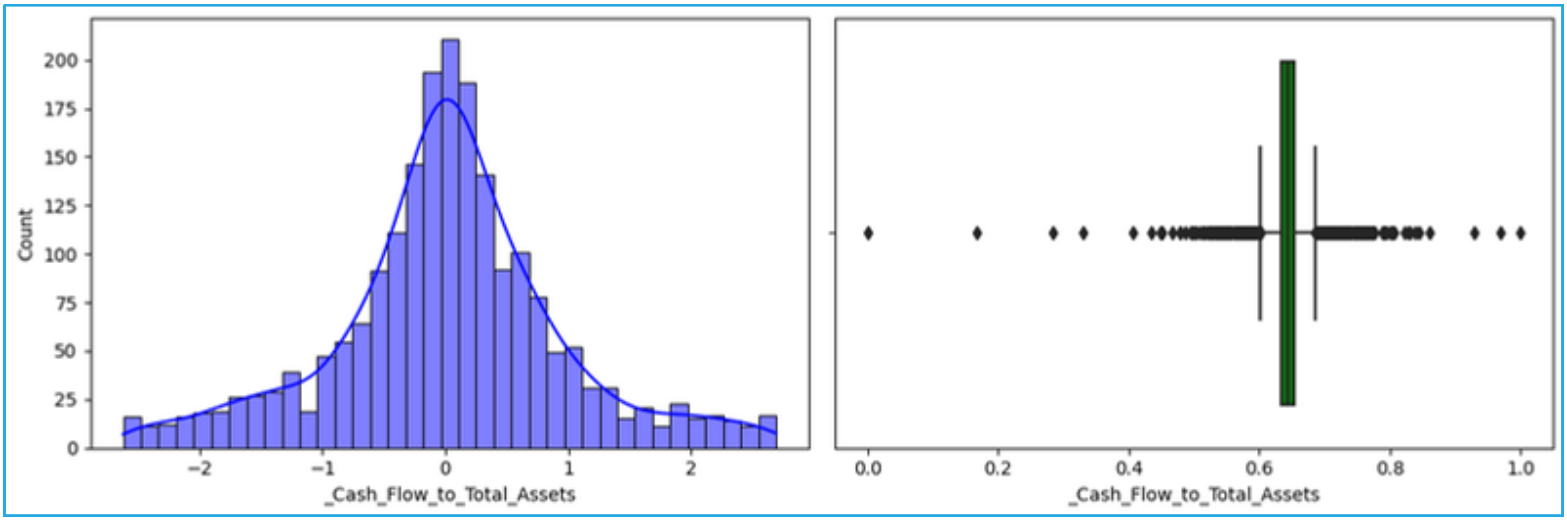

Count
_Operating_Funds_to_Liability
_Operating_Funds_to_Liability


Count
_Cash_Flow_Per_Share
_Cash_Flow_Per_Share


Count
_Cash_Reinvestment_perc
_Cash_Reinvestment_perc


Count
_CFO_to_Assets
_CFO_to_Assets


Count
_Total_expense_to_Assets
_Total_expense_to_Assets


Count
_Cash_Flow_to_Total_Assets
_Cash_Flow_to_Total_Assets

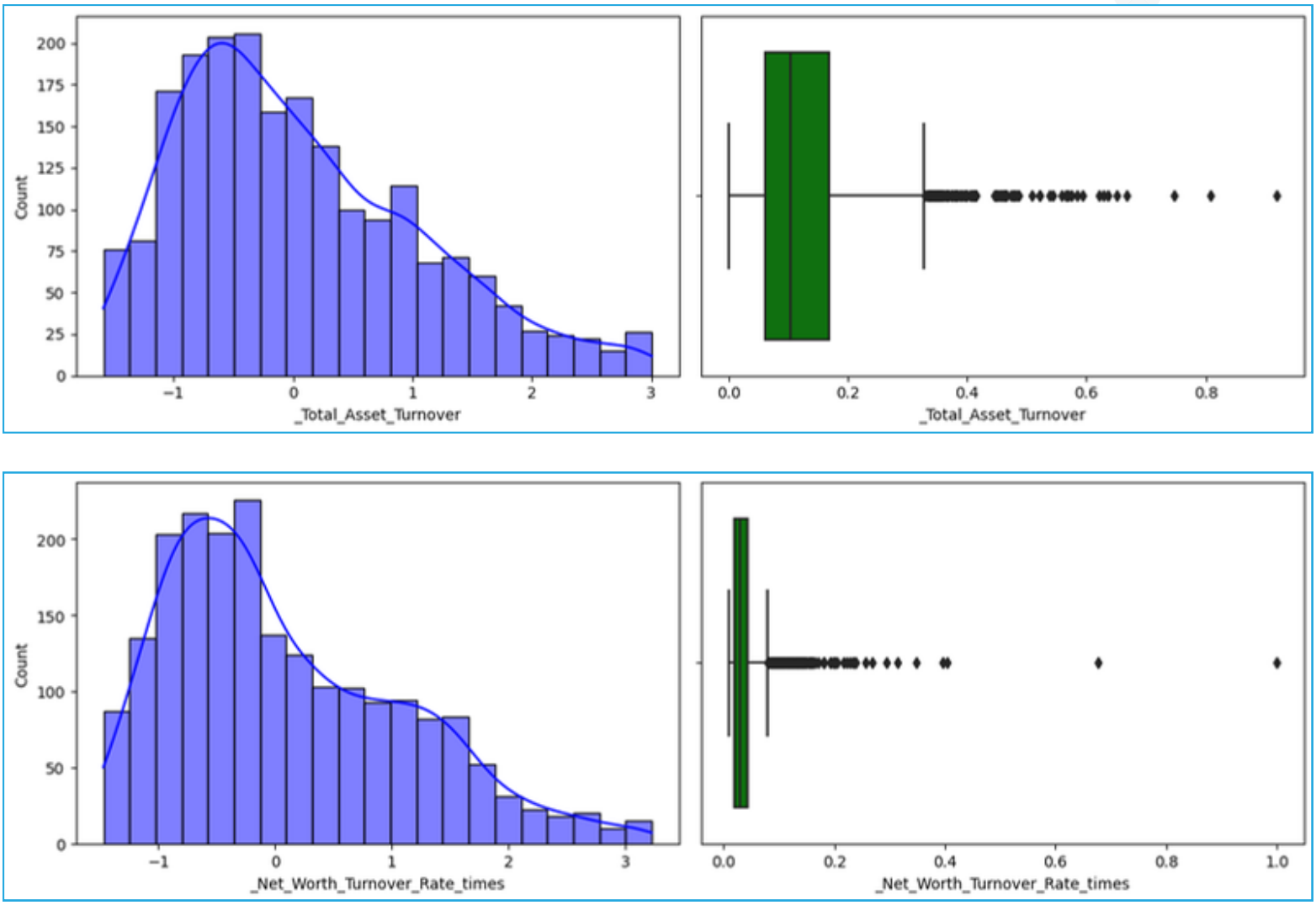


Figure 1.3: Univariate analysis of significant features

From the above graphs we can see than most of variable are normally distributed. Whereas variables like ‘Total_expense_to_Assets’, 'Total_Asset_Turnover' and '_Net_Worth_Turnover_Rate_times' are left skewed.

Bivariate Analysis

Bivariate analysis is a statistical analysis technique that focuses on examining the relationship or association between two variables in a dataset.

Below is Correlation heat map for few significant features

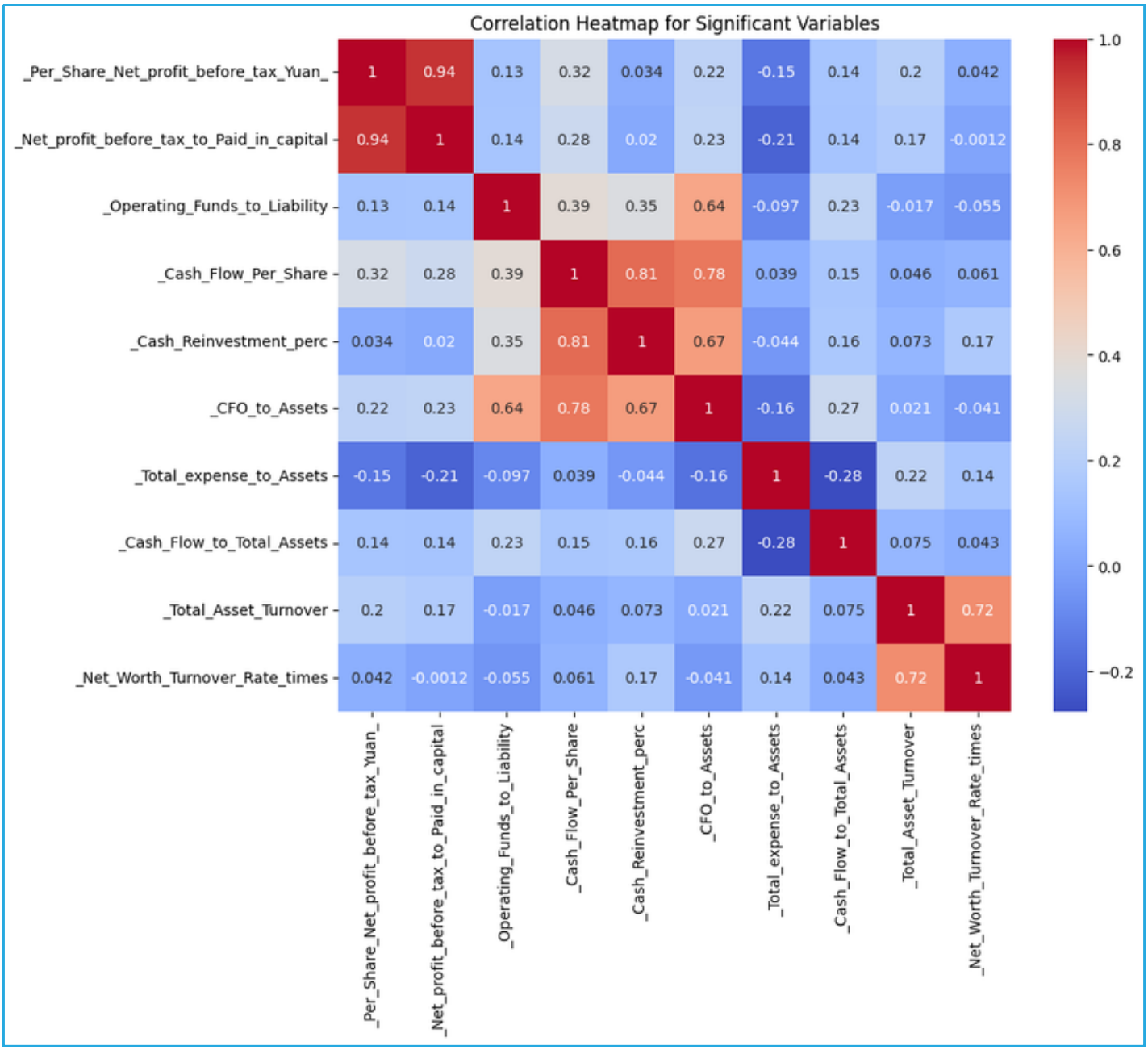


Figure 1.4: Correlation heat map

From the above graph here are few observations:

Strong positive correlations:

- Net_profit_before_tax_to_Paid_in_capital and Per_Share_Net_profit_before_tax_Yuan
- CFO_to_Assets and Operating_Funds_to_Liability
- Cash_Flow_Per_Share and Cash_Reinvestment_perc
- Cash_Flow_Per_Share and Operating_Funds_to_Liability

Negative correlations:

- Total_expense_to_Assets and Cash_Flow_to_Total_Assets
- CFO_to_Assets and Total_expense_to_Assets
- Total_expense_to_Assets and Net_profit_before_tax_to_Paid_in_capital
- Per_Share_Net_profit_before_tax_Yuan and Total_expense_to_Assets

These correlation we can visualise using a pairplot

Now lets look at Pairplot for significant features

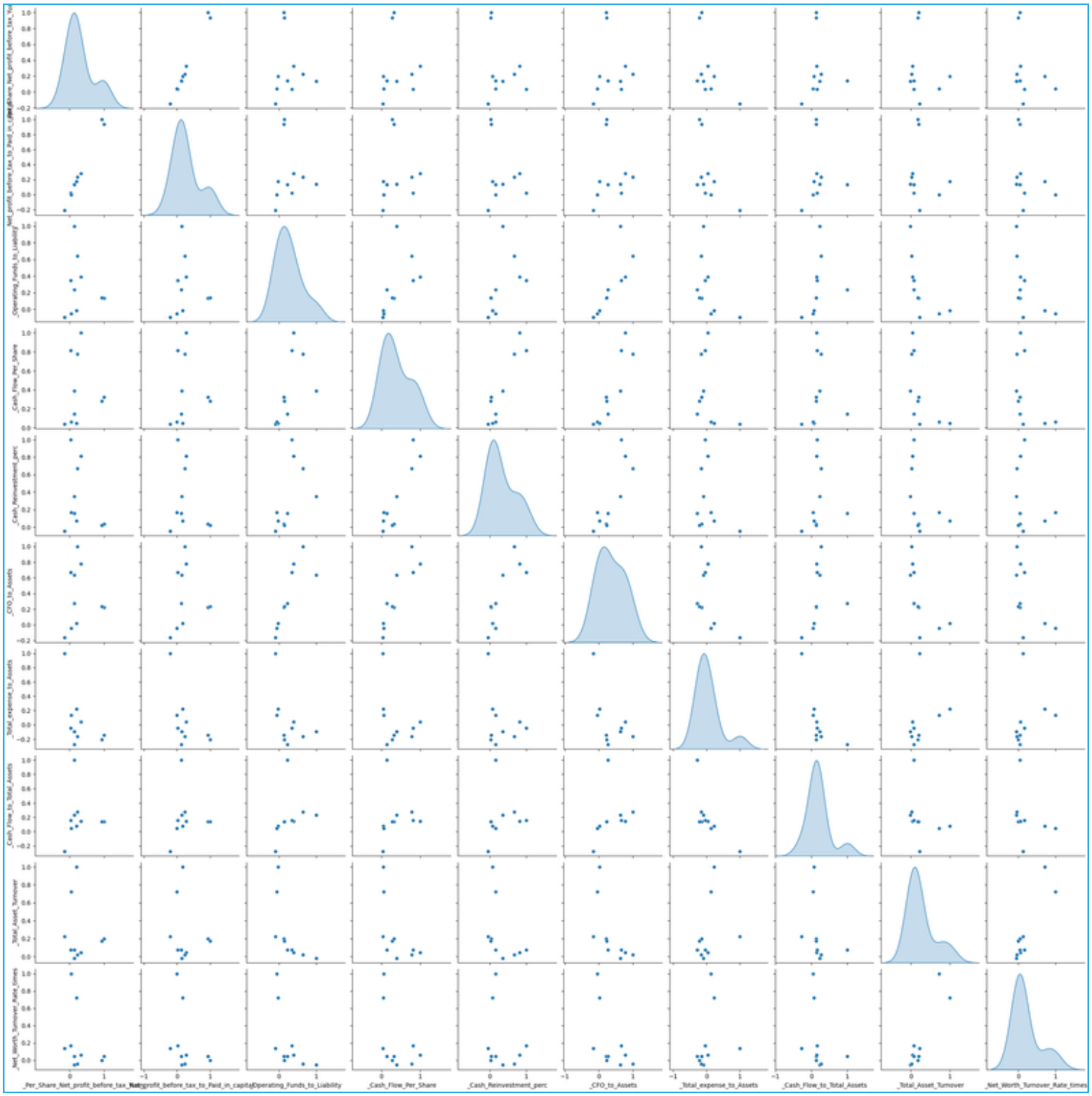
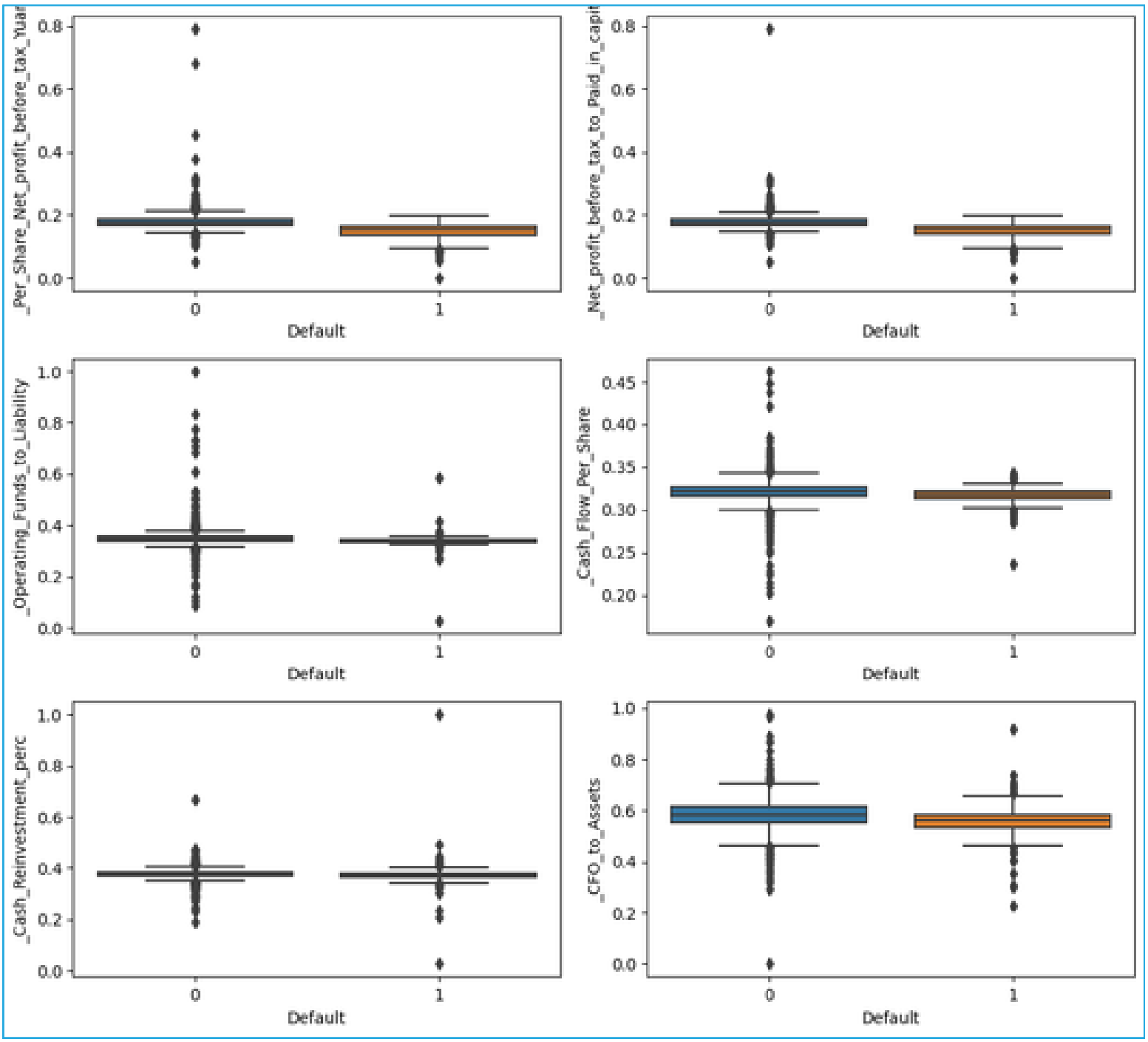


Figure 1.5: Pair Plot

Boxplot for continuous variables vs Default



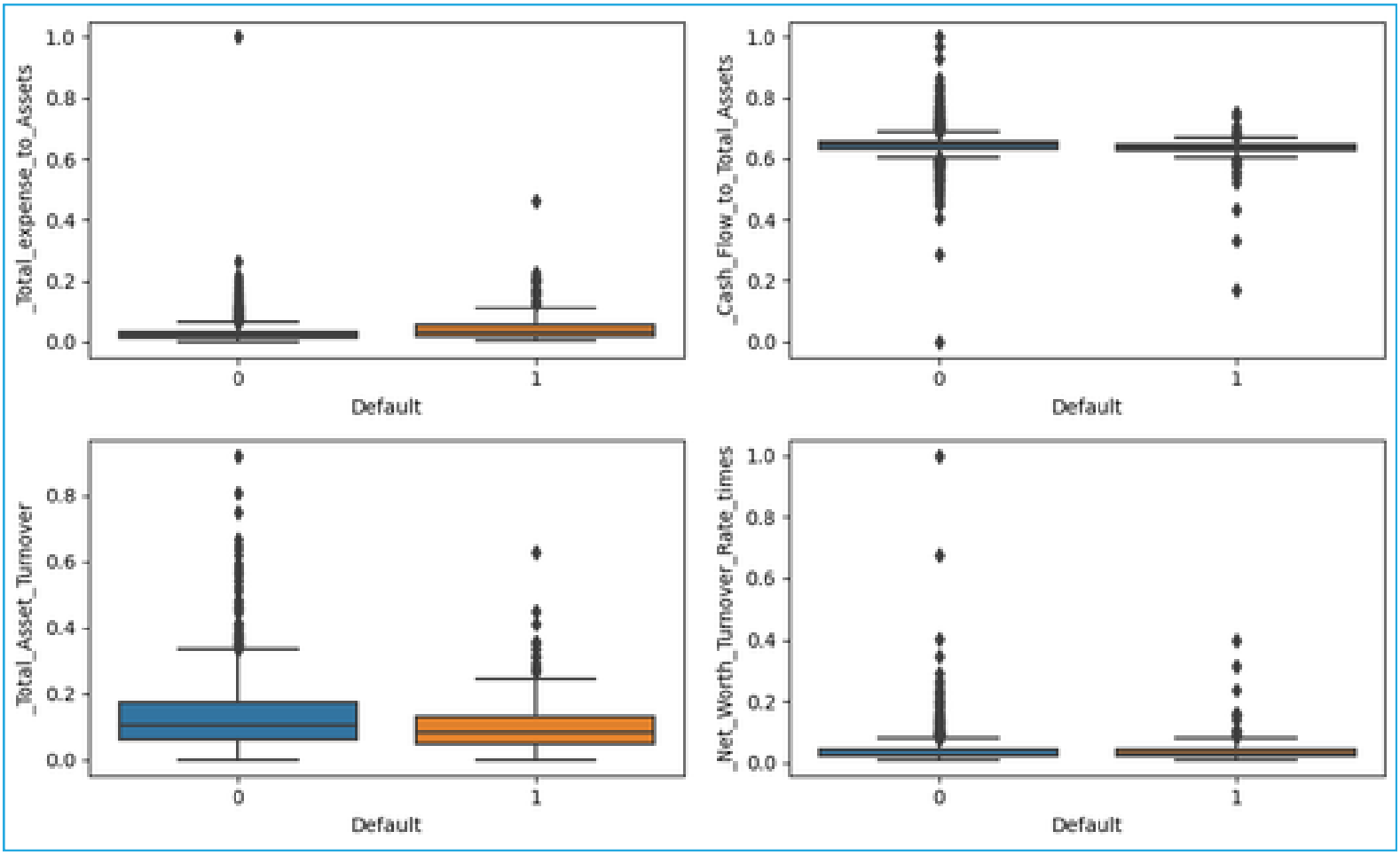


Figure 1.6: Continuous variables vs Default Boxplot

Based on the graphs above, the probability of default remains relatively constant across different variables, with the exception of a few outliers.

4. Create Train Test Split

We have divided the data into a 67:33 ratio of train to test data

```
The number of rows (observations) in train data is 1378
The number of columns (variables) is 54
```

```
The number of rows (observations) in test data is 680
The number of columns (variables) is 54
```

Table 1.12: Shape of train and test data

Below are top few rows of train and test data

_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_Cash_Flow_Per_Share	_Per_Share_Net_profit_before_tax_Yuan_
-0.15	-0.26	-0.84	0.07	-0.02
1.58	-0.54	-0.84	-0.32	-0.34
1.49	-1.02	1.18	0.94	0.39
0.03	1.10	0.19	0.40	1.41
0.63	-0.23	-0.22	0.23	-0.26

Table 1.13: Top 5 rows of train data

_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_Cash_Flow_Per_Share	_Per_Share_Net_profit_before_tax_Yuan_
-1.27	0.45	1.31	-1.12	1.44
2.20	1.53	-0.63	1.14	1.24
-0.48	-0.91	-0.84	-0.45	-0.80
0.81	0.43	0.88	0.54	0.40
0.09	-0.17	-0.84	0.11	-0.71

Table 1.14: Top 5 rows of test data

5. 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

For building Logistic Regression model we are utilizing statsmodel library. Before building the model we need to reduce multicollinearity.
Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to problems in estimating the coefficients accurately and interpreting the model.

We have used Variance Inflation Factor (VIF) to quantify multicollinearity and dropped columns having VIF >10.

	variables	VIF
	_Current_Ratio	8.65
	_Operating_Funds_to_Liability	8.42
	_Quick_Ratio	7.88
	_Cash_flow_rate	7.78

variables	VIF
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	7.32
_Total_Asset_Turnover	6.85
_Cash_Flow_to_Total_Assets	6.61
_Net_Worth_Turnover_Rate_times	6.24
_Per_Share_Net_profit_before_tax_Yuan_	6.16
_Current_Liability_to_Current_Assets	5.64
_Quick_Assets_to_Total_Assets	5.32
_Degree_of_Financial_Leverage_DFL	5.04
_Interest_Expense_Ratio	4.98
_Equity_to_Liability	4.80
_Cash_Flow_Per_Share	4.53
_Total_income_to_Total_expense	4.53
_Cash_Flow_to_Equity	4.36
_Cash_Reinvestment_perc	4.24
_Retained_Earnings_to_Total_Assets	4.09
_Total_debt_to_Total_net_worth	4.01
_Cash_to_Current_Liability	3.75
_Cash_Flow_to_Liability	3.57
_Inventory_to_Current_Liability	3.49
_Fixed_Assets_to_Assets	3.43
_Average_Collection_Days	3.14
_Cash_to_Total_Assets	3.08
_Accounts_Receivable_Turnover	2.67

variables	VIF
_Operating_Profit_Growth_Rate	2.61
_Operating_profit_per_person	2.57
_Net_Value_Growth_Rate	2.40
_Continuous_Net_Profit_Growth_Rate	2.38
_Long_term_fund_suitability_ratio_A	2.25
_Current_Asset_Turnover_Rate	2.21
_Total_Asset_Return_Growth_Rate_Ratio	2.18
_Allocation_rate_per_person	2.13
_Realized_Sales_Gross_Profit_Growth_Rate	2.13
_Inventory_to_Working_Capital	2.02
_Total_expense_to_Assets	2.01
_No_credit_Interval	1.95
_Tax_rate_A	1.70
_Long_term_Liability_to_Current_Assets	1.60
_Total_assets_to_GNP_price	1.51
_Quick_Asset_Turnover_Rate	1.42
_Fixed_Assets_Turnover_Frequency	1.39
_Operating_Expense_Rate	1.30
_Interest_bearing_debt_interest_rate	1.29
_Inventory_Turnover_Rate_times	1.26
_Research_and_development_expense_rate	1.18
_Total_Asset_Growth_Rate	1.10
_Cash_Turnover_Rate	1.08

Table 1.15: VIF values of variables

After calculating VIF, now we will fit our model to logit function and create summary table. We have dropped variables with p-value less than 0.05 as they re not significant and optimised table is given below.

Logit Regression Results

Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1362
Method:	MLE	Df Model:	15
Date:	Fri, 06 Oct 2023	Pseudo R-squ.:	0.4368
Time:	09:20:13	Log-Likelihood:	-270.61
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	4.976e-80

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9733	0.258	-15.410	0.000	-4.479	-3.468
_Research_and_development_expense_rate	0.3612	0.121	2.978	0.003	0.123	0.599
_Interest_bearing_debt_interest_rate	0.3495	0.133	2.626	0.009	0.089	0.610
_Tax_rate_A	-0.4318	0.173	-2.496	0.013	-0.771	-0.093
_Cash_Flow_Per_Share	-0.3577	0.127	-2.826	0.005	-0.606	-0.110
_Quick_Ratio	-1.8076	0.353	-5.128	0.000	-2.499	-1.117
_Accounts_Receivable_Turnover	-0.6295	0.148	-4.267	0.000	-0.919	-0.340
_Operating_profit_per_person	0.5165	0.177	2.924	0.003	0.170	0.863
_Allocation_rate_per_person	0.4171	0.141	2.953	0.003	0.140	0.694
_Quick_Assets_to_Total_Assets	0.4772	0.193	2.477	0.013	0.100	0.855
_Cash_to_Current_Liability	0.6739	0.210	3.210	0.001	0.262	1.085
_Inventory_to_Working_Capital	-0.3872	0.103	-3.761	0.000	-0.589	-0.185
_Total_income_to_Total_expense	-1.2733	0.223	-5.705	0.000	-1.711	-0.836
_Total_expense_to_Assets	0.3933	0.129	3.047	0.002	0.140	0.646
_Cash_Flow_to_Liability	-0.2950	0.142	-2.080	0.038	-0.573	-0.017
_Equity_to_Liability	-1.0321	0.253	-4.078	0.000	-1.528	-0.536

Table 1.16: Logistic regression summary table

In the summary table we can see now we are down to 15 variables. Positive coefficient shows the direct relation with probability of default and vice versa

To check the precision and recall on train data, we have taken a threshold of 0.5 in the start and later we will modify it to find optimum threshold

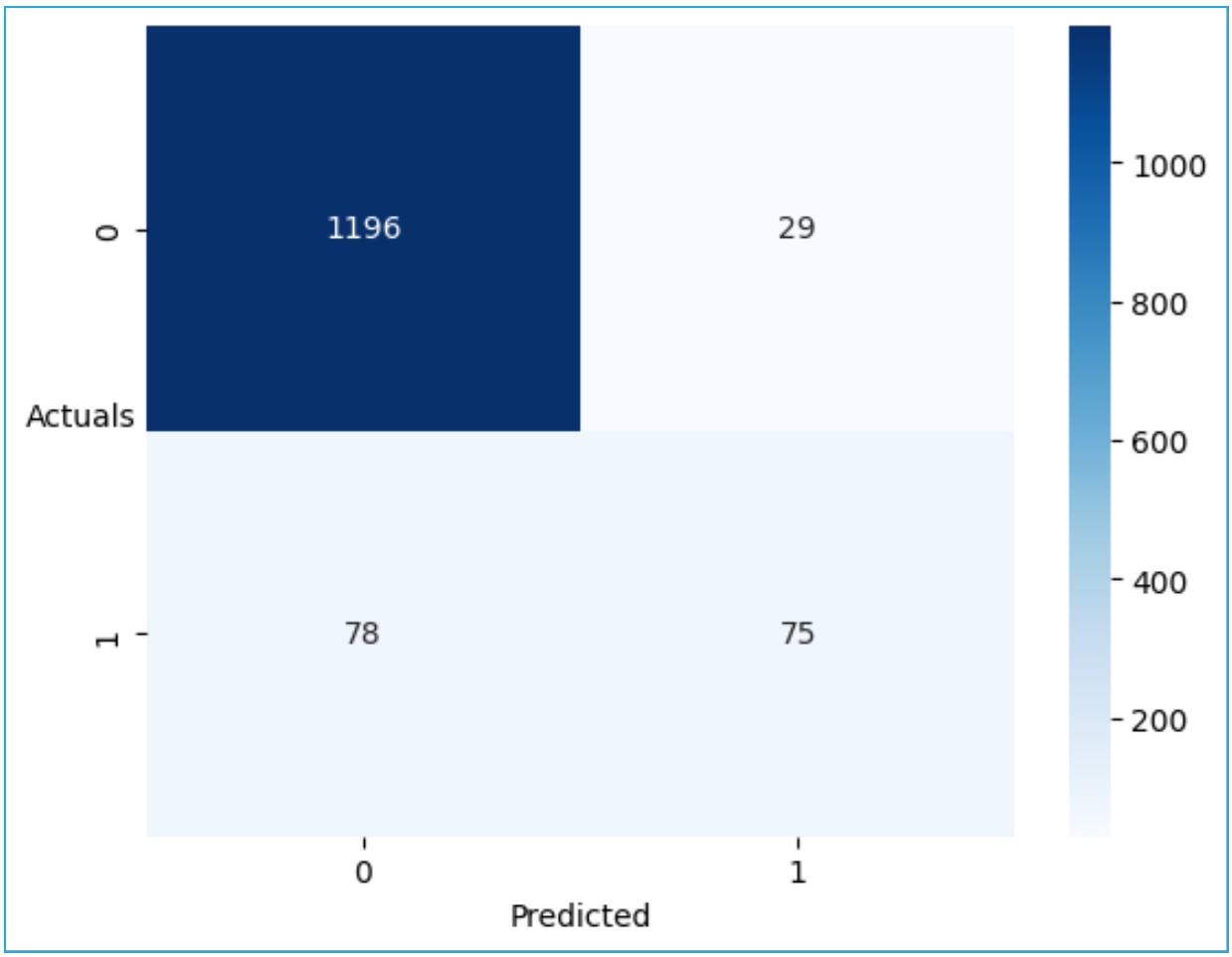


Figure 1.7: Confusion matrix for 0.5 threshold on train data

Recall = 49% Precision = 72%

So the current threshold does not have good Recall and Precision

We are using roc_curve library to find the optimum threshold which will ensure the maximum difference between true positive rate and false positive rate

New threshold came out to be 0.106 and the updated confusion matrix using this threshold is below.

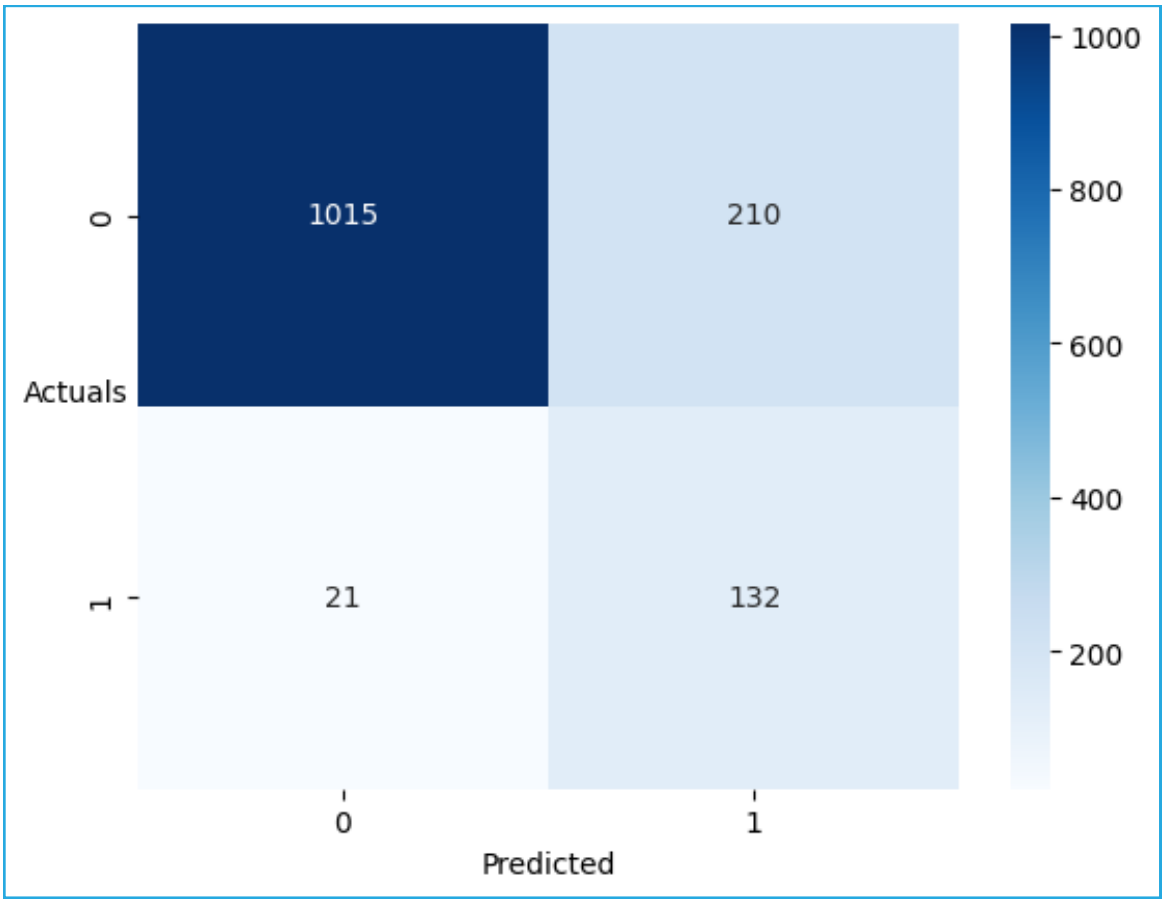


Figure 1.8: Confusion matrix for 0.106 threshold on train data

Recall = 0.86 Precision = 0.39

So we have managed to increase our Recall to 0.86 but in return precision has suffered.

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

Now validating it on test data.

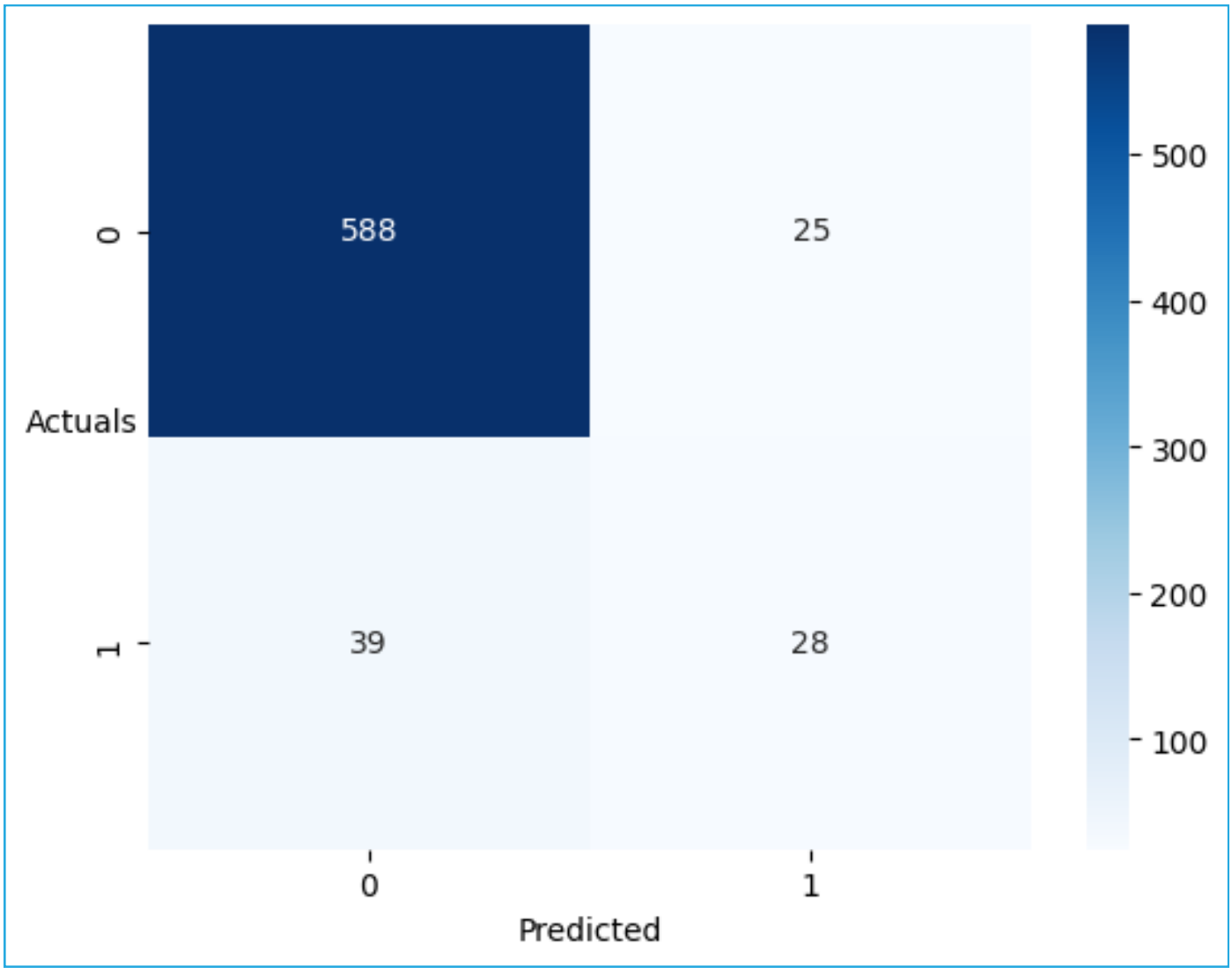


Figure 1.9: Confusion matrix for 0.5 threshold on test data

Recall = 0.42 Precision = 0.53

So for test data also we are having low recall for 0.5 threshold value. now we will apply optimum threshold value of 0.106

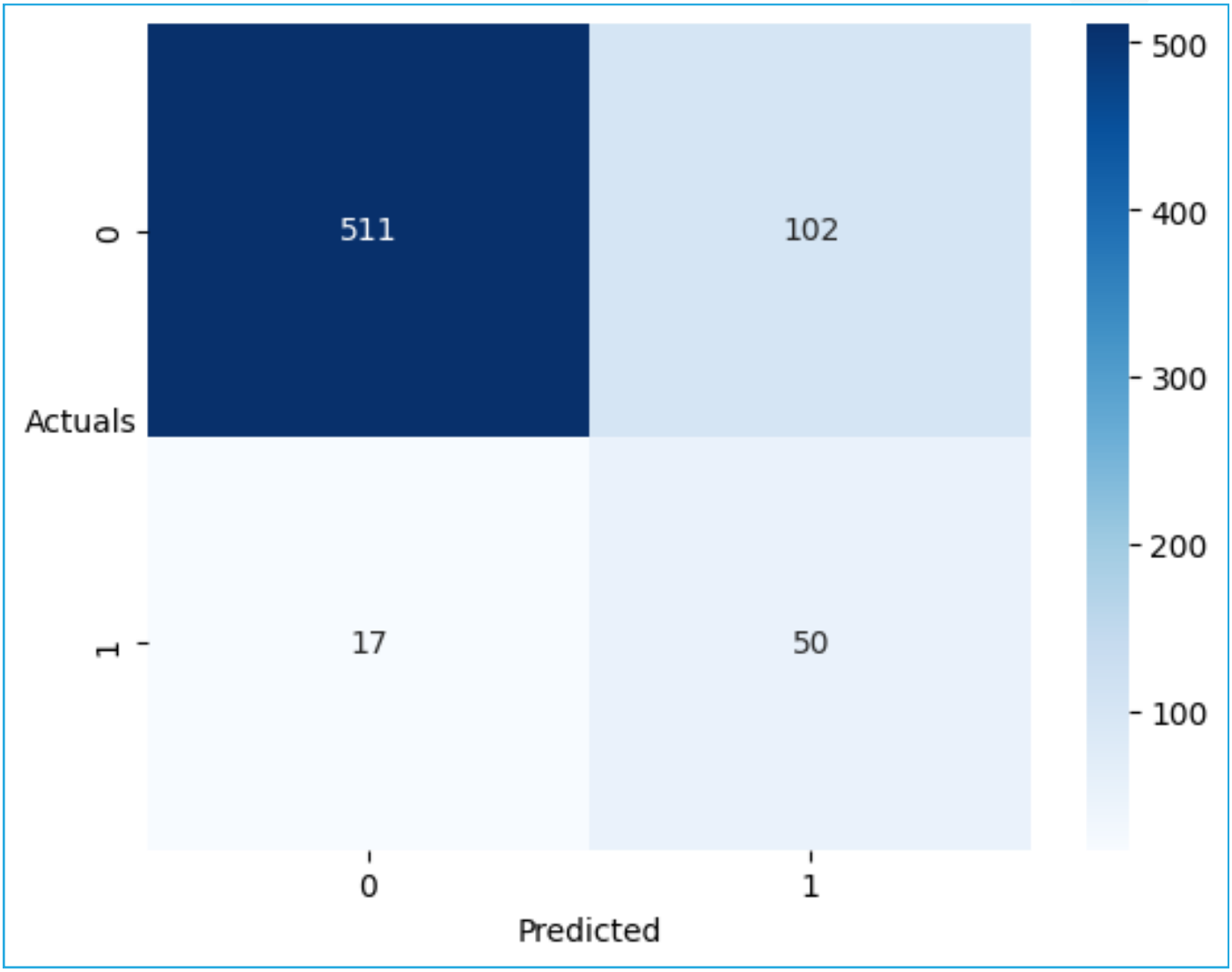


Figure 1.10: Confusion matrix for 0.106 threshold on test data

Recall = 0.75 Precision = 0.33

So for test data also we are having good recall score with optimum threshold value.

Now looking at the ROC curve

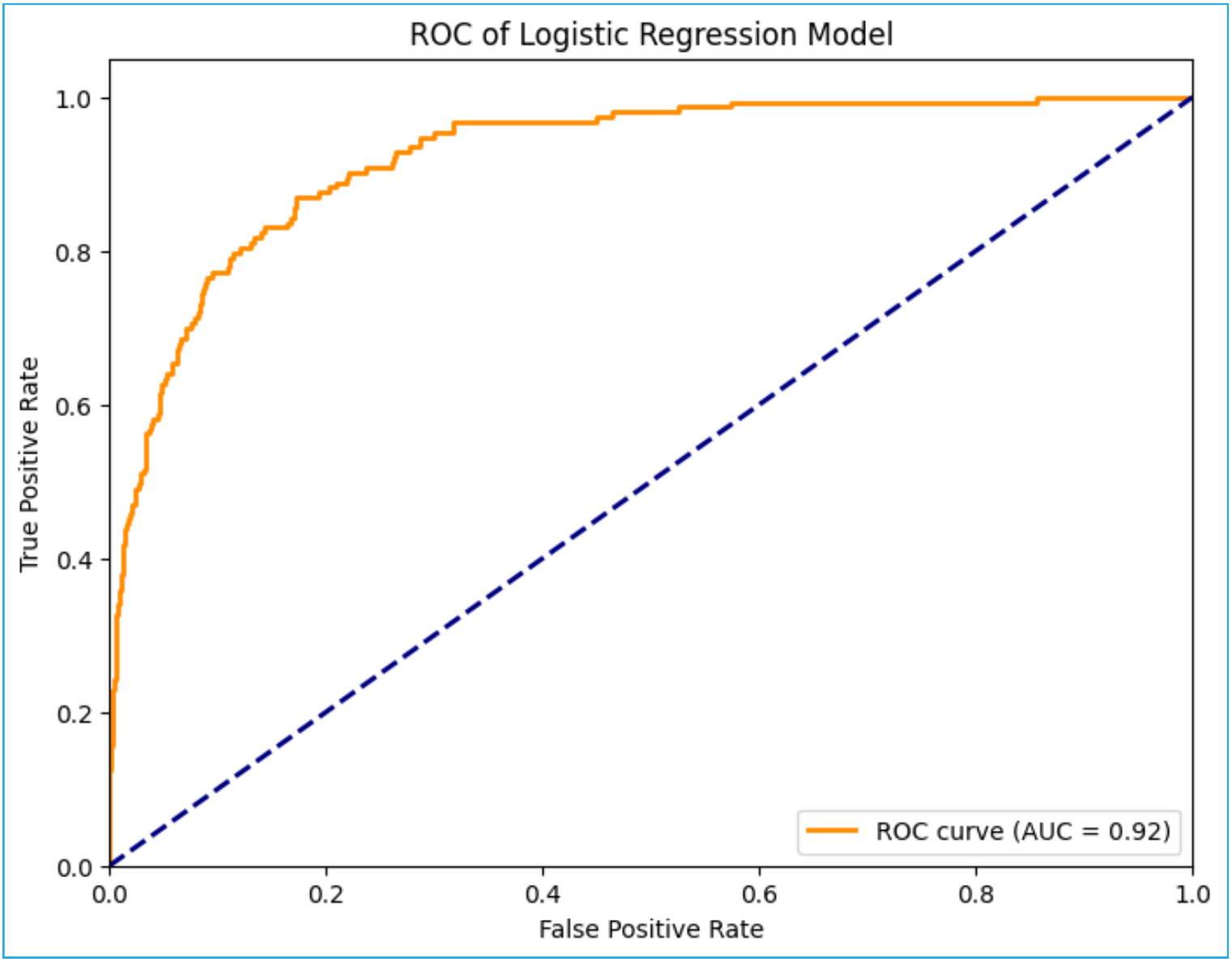


Figure 1.11: ROC curve for test data

Since the ROC curve is closer to the top-left corner of the plot, the model's performance is good. This corresponds to higher sensitivity and lower false positive rate.

7. Build a Random Forest Model on Train Dataset. Also showcase your model building approach

We have build Random forest model using RandomForestClassifier and used GridSearchCV to select the best parameters.

- n_estimators is the number of decision trees (estimators) to be included in the random forest.
- Increasing n_estimators generally improves the performance of the random forest by reducing overfitting and making predictions more robust.
- max_depth is the maximum depth or levels that an individual decision tree in the random forest can grow.
- Limiting the depth of each tree can help prevent overfitting and improve generalization.
- min_samples_split is the minimum number of samples required to split an internal node during the construction of a decision tree.
- It controls how finely the decision tree can partition the feature space.

For our model we have used
'max_depth'= 5,
'min_samples_leaf'= 10,
'min_samples_split'= 15,
'n_estimators'= 50

Now lets check the confusion matrix.

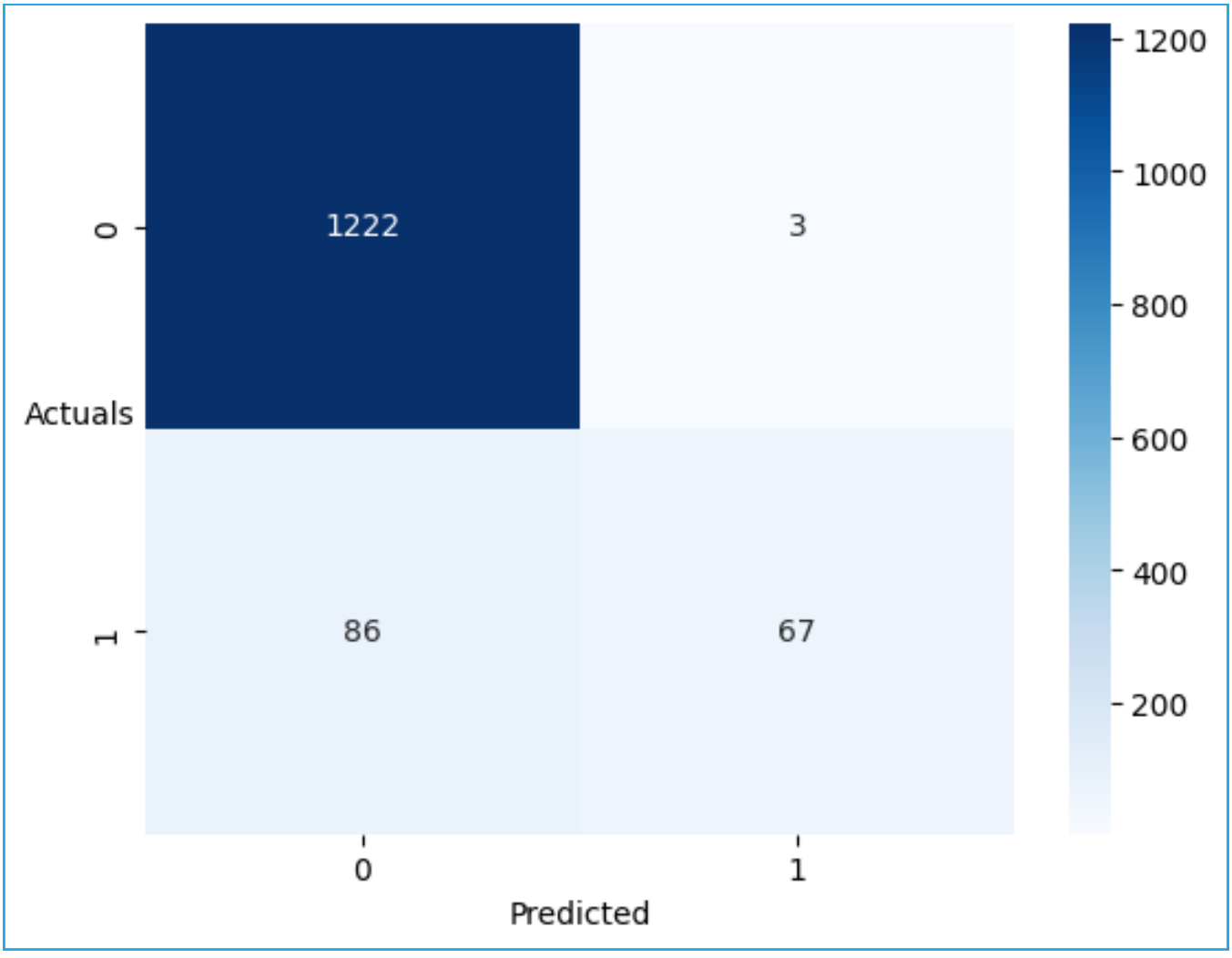


Figure 1.12: Confusion matrix of RF for train data

	precision	recall	f1-score	support
0.0	0.93	1.00	0.96	1225
1.0	0.96	0.44	0.60	153
accuracy			0.94	1378
macro avg	0.95	0.72	0.78	1378
weighted avg	0.94	0.94	0.92	1378

Table 1.17: Characteristic table for RF train data

From here we can see that train data has 0.44 Recall and 0.96 Precision whereas Accuracy is 0.94

8. Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model

Now lets check model characterstics for test data

	precision	recall	f1-score	support
0.0	0.93	0.98	0.95	613
1.0	0.63	0.33	0.43	67
accuracy			0.91	680
macro avg	0.78	0.65	0.69	680
weighted avg	0.90	0.91	0.90	680

Table 1.18: Characteristic table for RF test data

From the test data Recall is 0.33 and 0.63 Precision whereas Accuracy is 0.91. So, clearly there is slight dip in results

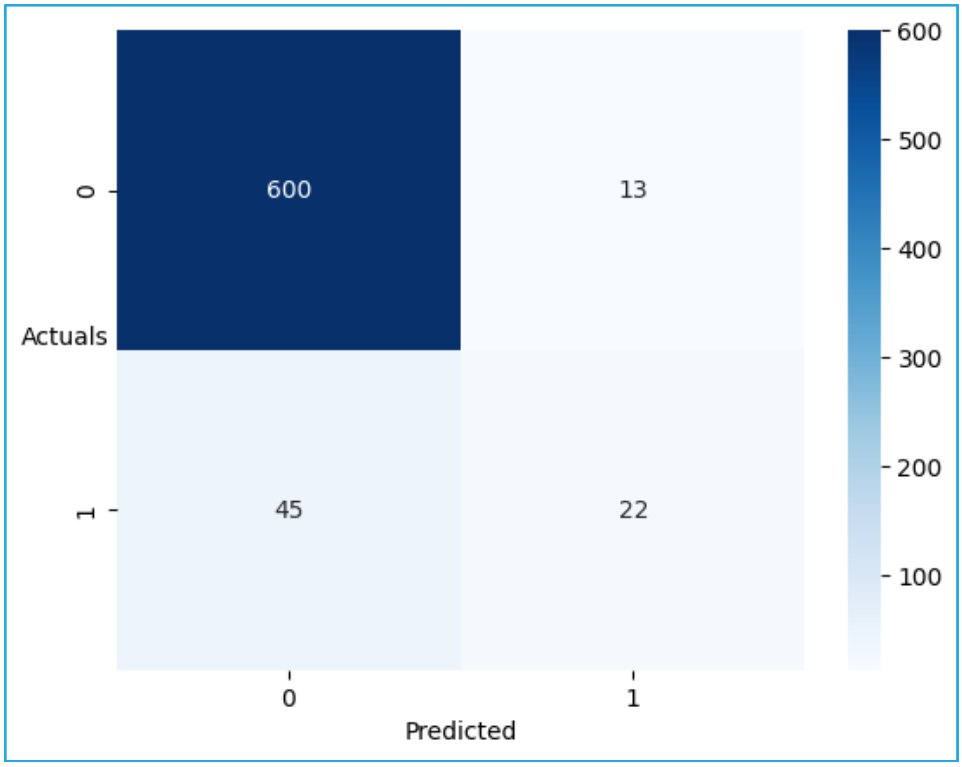


Figure 1.13: Confusion matrix of RF for test data

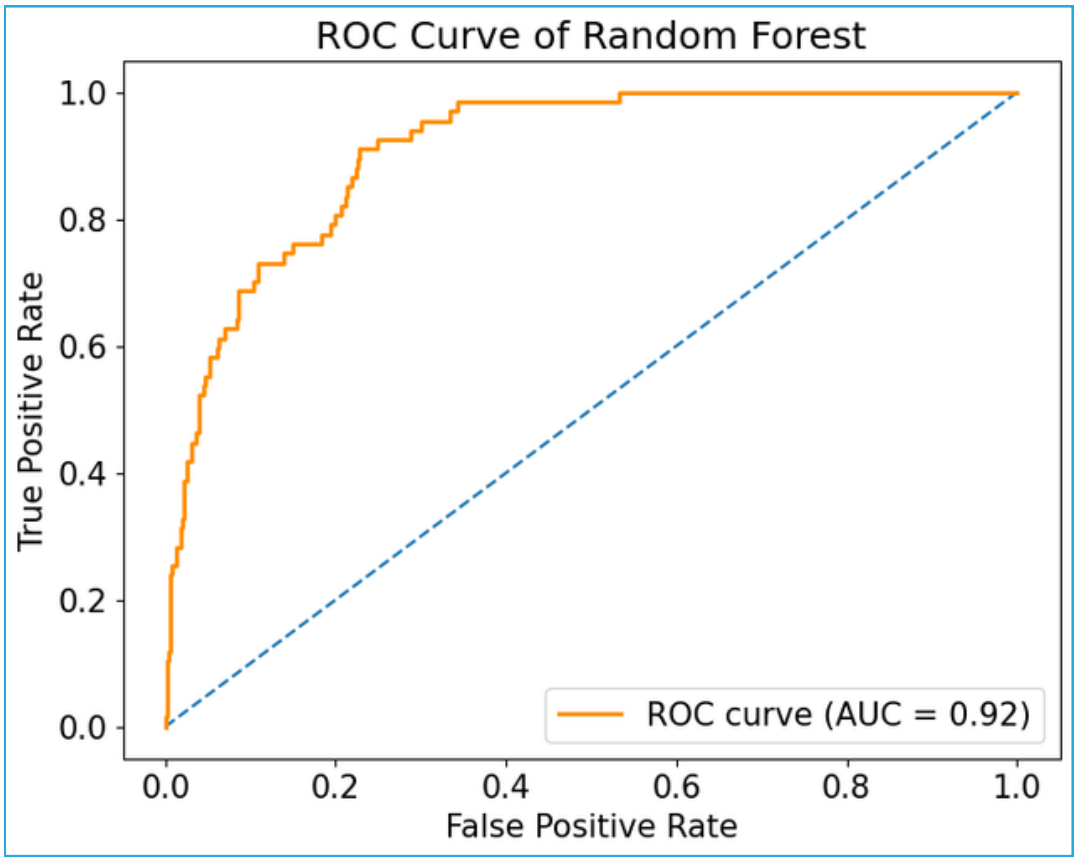


Figure 1.14: ROC curve of RF for test data

Interpretations from model:

- True Positives (TP): 22
- False Positives (FP): 13
- True Negatives (TN): 600
- False Negatives (FN): 45

- The model correctly predicted 600 instances as negative (true negatives). These are cases where the model correctly identified non-events or negative outcomes.
- The model is not as effective at correctly identifying positive instances (low TP).
- Minimizing false negatives (Defaulters identified as non defaulter) is a priority, even if it means accepting some false positives .
- The trade-off here is that we might have lower precision (higher false positive rate), but we are capturing a larger proportion of the actual positive cases (higher Recall).

9. Build a LDA Model on Train Dataset. Also showcase your model building approach

Linear Discriminant Analysis (LDA) is a statistical and machine learning technique used for dimensionality reduction and supervised classification. It is particularly useful when dealing with multi-class classification problems

We have used LinearDiscriminantAnalysis function from sklearn library to build our model.

	precision	recall	f1-score	support
0.0	0.95	0.95	0.95	1225
1.0	0.60	0.56	0.58	153
accuracy			0.91	1378
macro avg	0.77	0.76	0.77	1378
weighted avg	0.91	0.91	0.91	1378

Table 1.19: Characteristic table for LDA of train data

When we look at result from LDA model of train data without setting any threshold value, we see a low value of 0.56 for Recall. So now we will try to find a optimum threshold for our model and re-run the model.

After using roc_curve function we get a threshold value of 0.065.

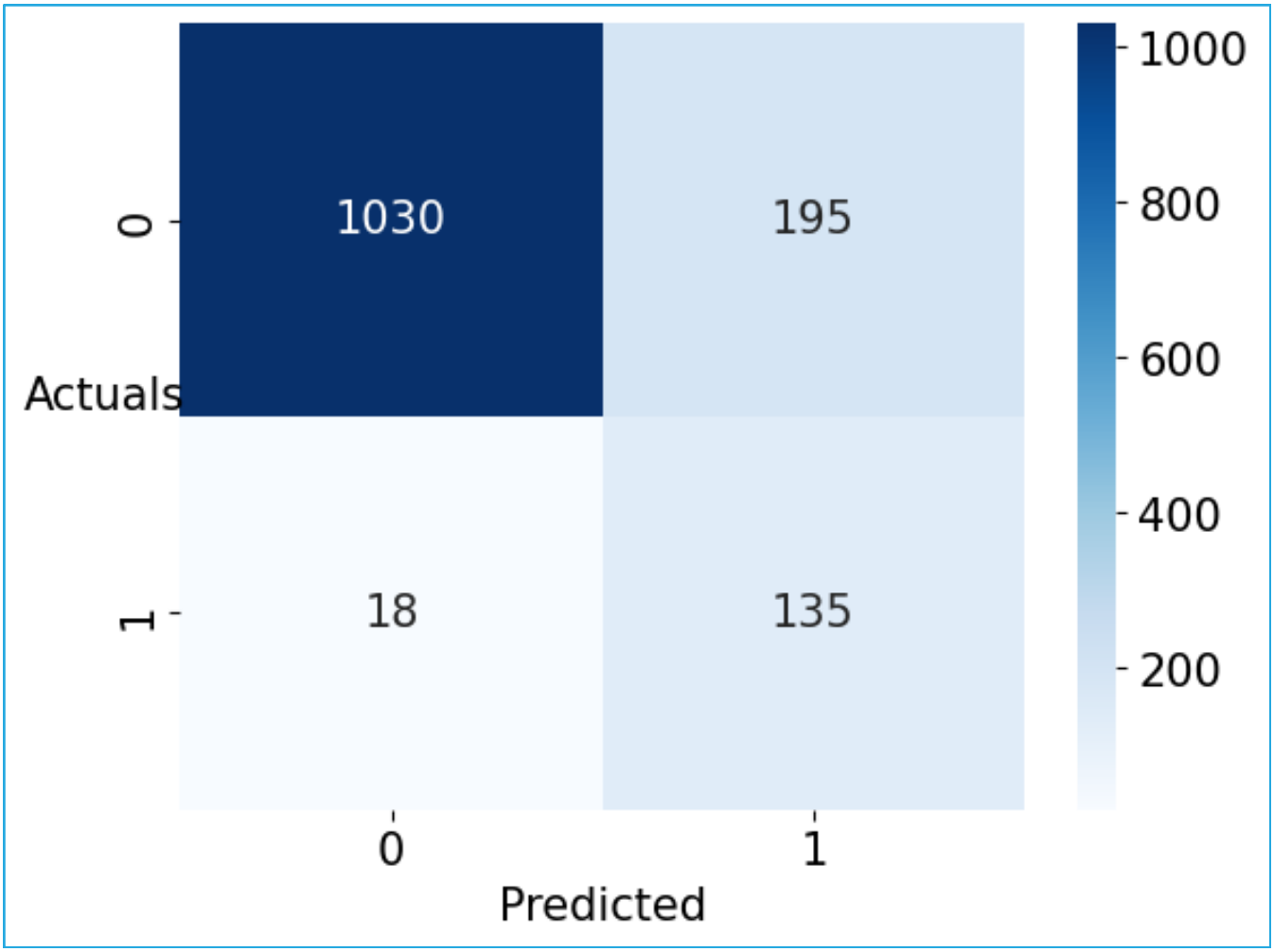


Figure 1.15: Confusion matrix for LDA of train data

	precision	recall	f1-score	support
0.0	0.98	0.84	0.91	1225
1.0	0.41	0.88	0.56	153
accuracy			0.85	1378
macro avg	0.70	0.86	0.73	1378
weighted avg	0.92	0.85	0.87	1378

Table 1.19: Characteristic table for LDA of train data

Now we can observe that recall has significantly increased and model is able to predict true positives more accurately.

For the train data Recall is 0.88 and 0.41 Precision whereas Accuracy is 0.85.

10. Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model

Now lets check model characterstics for test data

	precision	recall	f1-score	support
0.0	0.98	0.83	0.90	613
1.0	0.34	0.81	0.48	67
accuracy			0.83	680
macro avg	0.66	0.82	0.69	680
weighted avg	0.91	0.83	0.86	680

Table 1.20: Characteristic table for LDA of test data

For the test data Recall is 0.81 and 0.34 Precision whereas Accuracy is 0.83 So, clearly there is slight dip in results

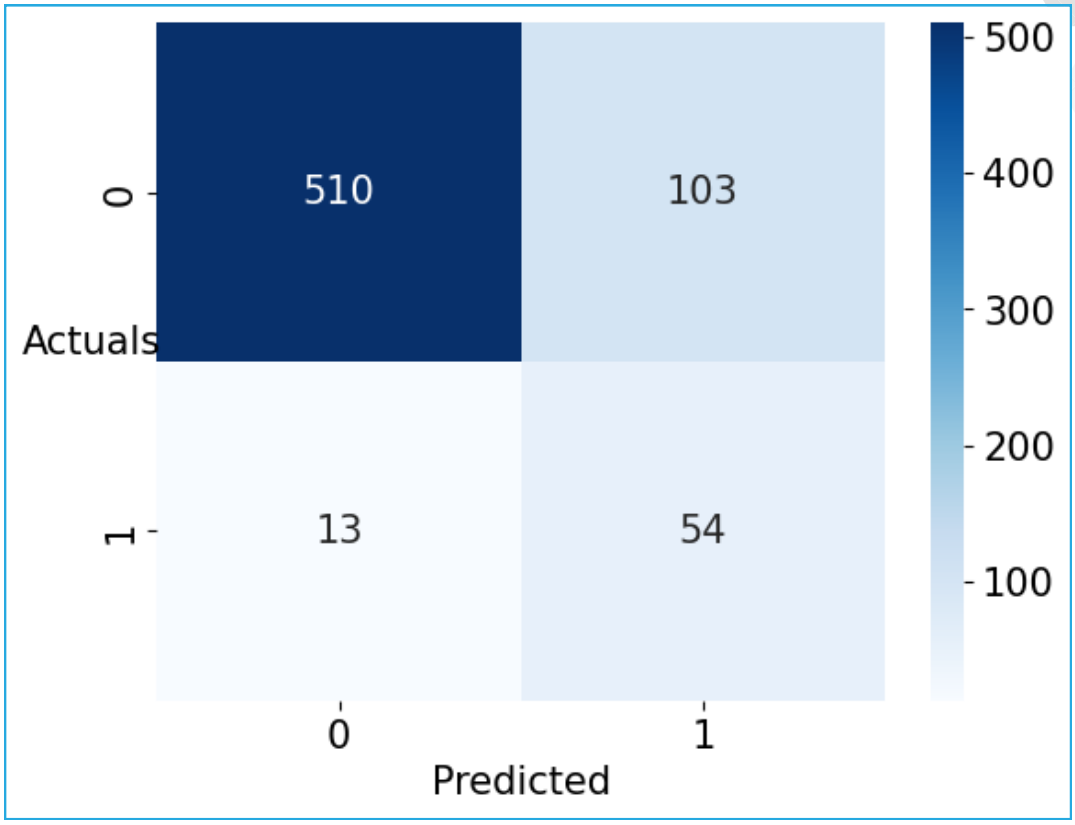


Figure 1.16: Confusion matrix for LDA of test data

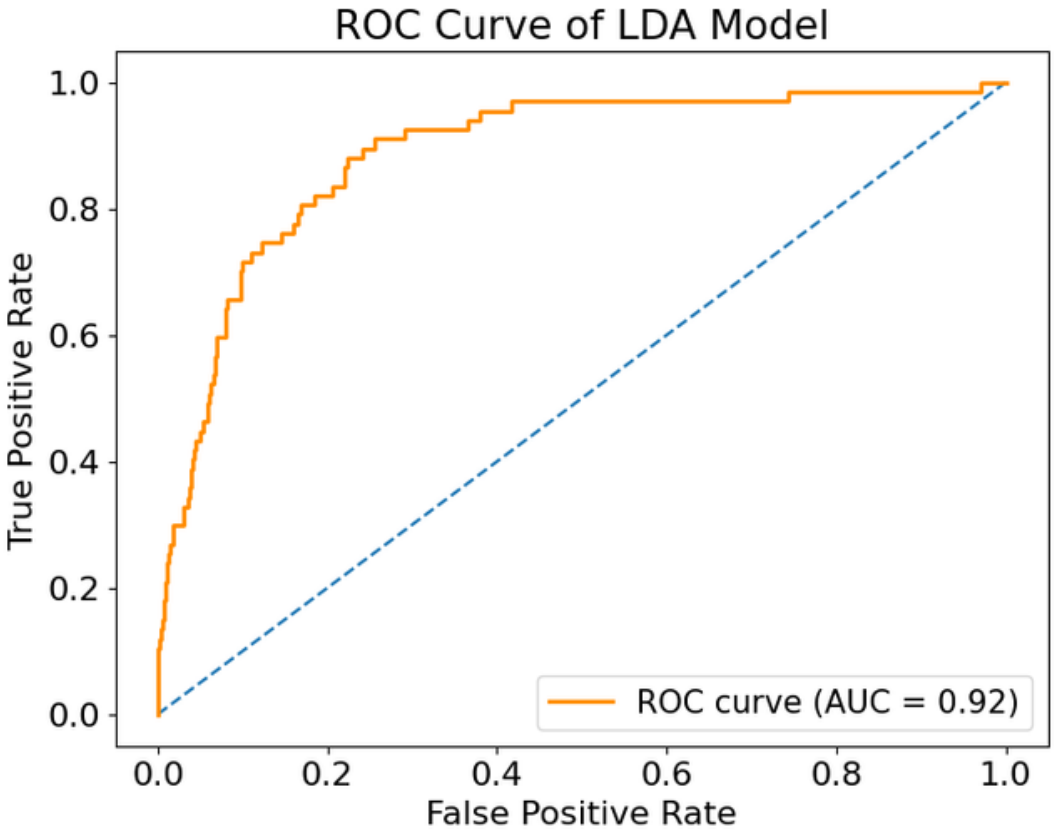


Figure 1.17: ROC curve for LDA of test data

Interpretations from model:

- True Positives (TP): 54
- False Positives (FP): 103
- True Negatives (TN): 510
- False Negatives (FN): 13
- The model correctly predicted 510 instances as negative (true negatives). These are cases where the model correctly identified non-events or negative outcomes.
- One of our primary objectives is to minimize false negatives. In this case, model incorrectly predicted 13 positive instances as negative.
- Model incorrectly predicted 103 negative instances as positive. Minimizing false positives is important, but in a Recall-focused scenario, it's acceptable to have more false positives.

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

Let's compare all three models.

Logistic Regression

True Positives (TP): 50 False Positives (FP): 102

True Negatives (TN): 511 False Negatives (FN): 17

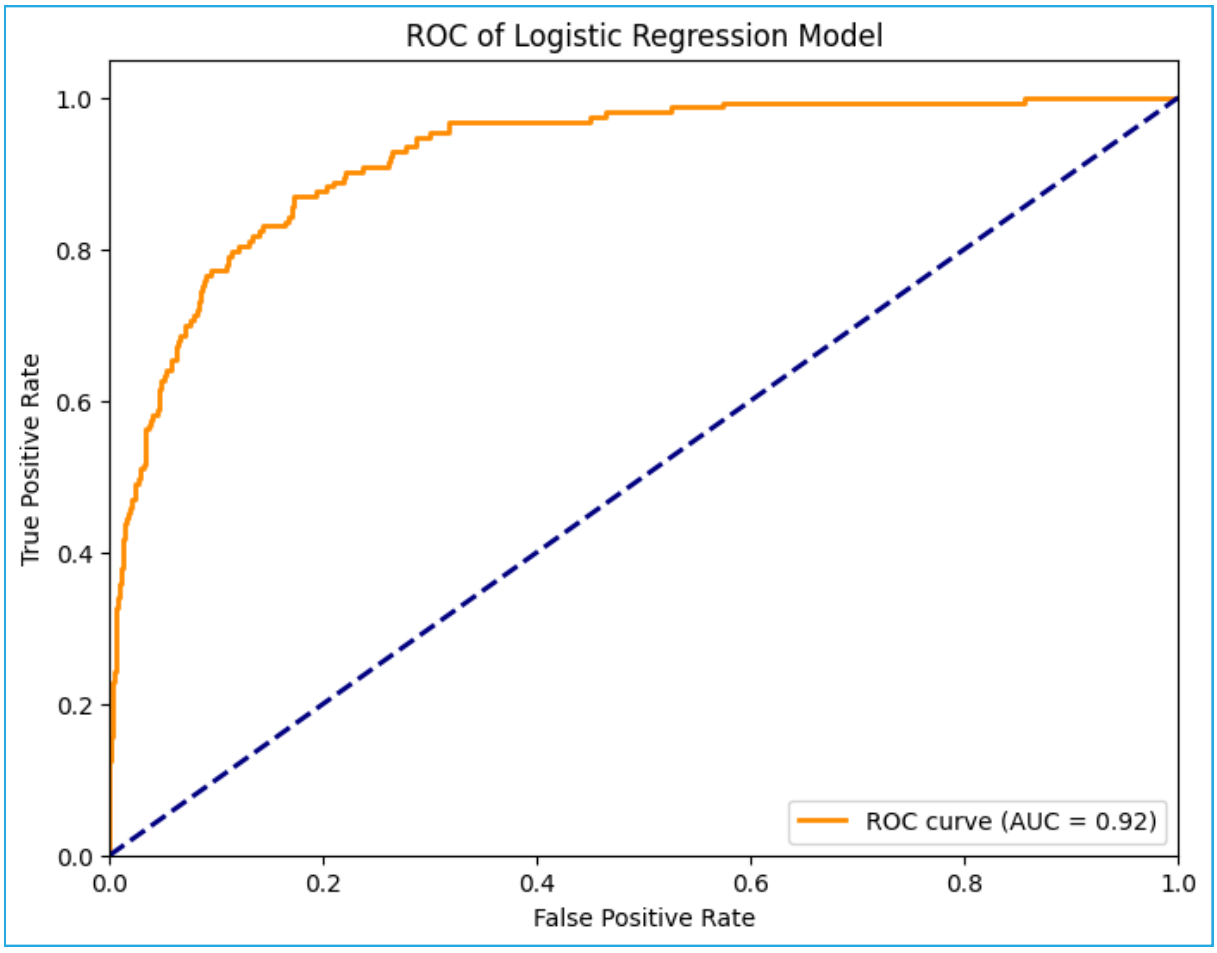


Figure 1.18: ROC curve for Logistic Regression of test data

Random Forest

True Positives (TP): 22

False Positives (FP): 13

True Negatives (TN): 600

False Negatives (FN): 45

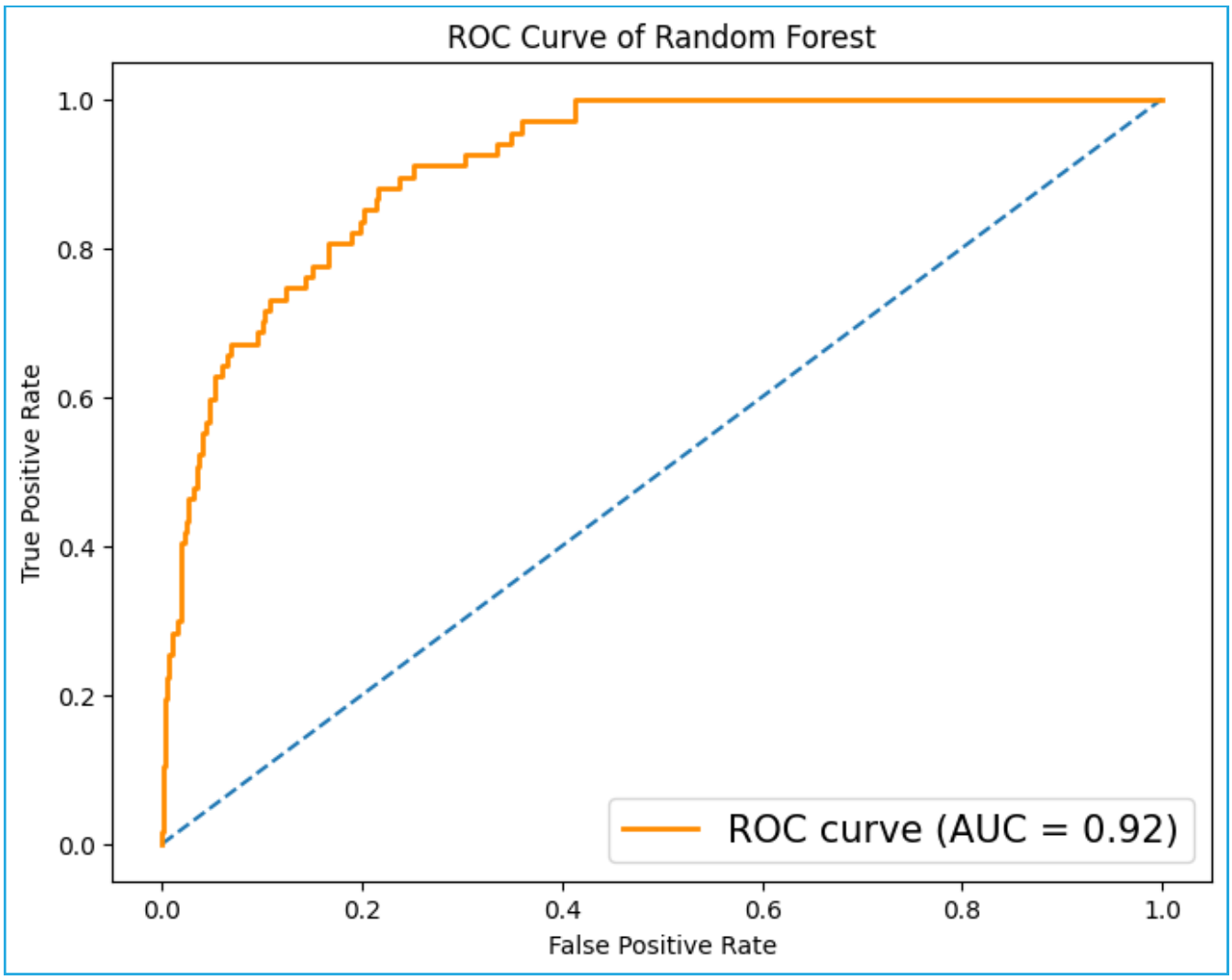


Figure 1.19: ROC curve for Random Forest of test data

Linear Discriminant Analysis

True Positives (TP): 54 False Positives (FP): 103

True Negatives (TN): 510 False Negatives (FN): 13

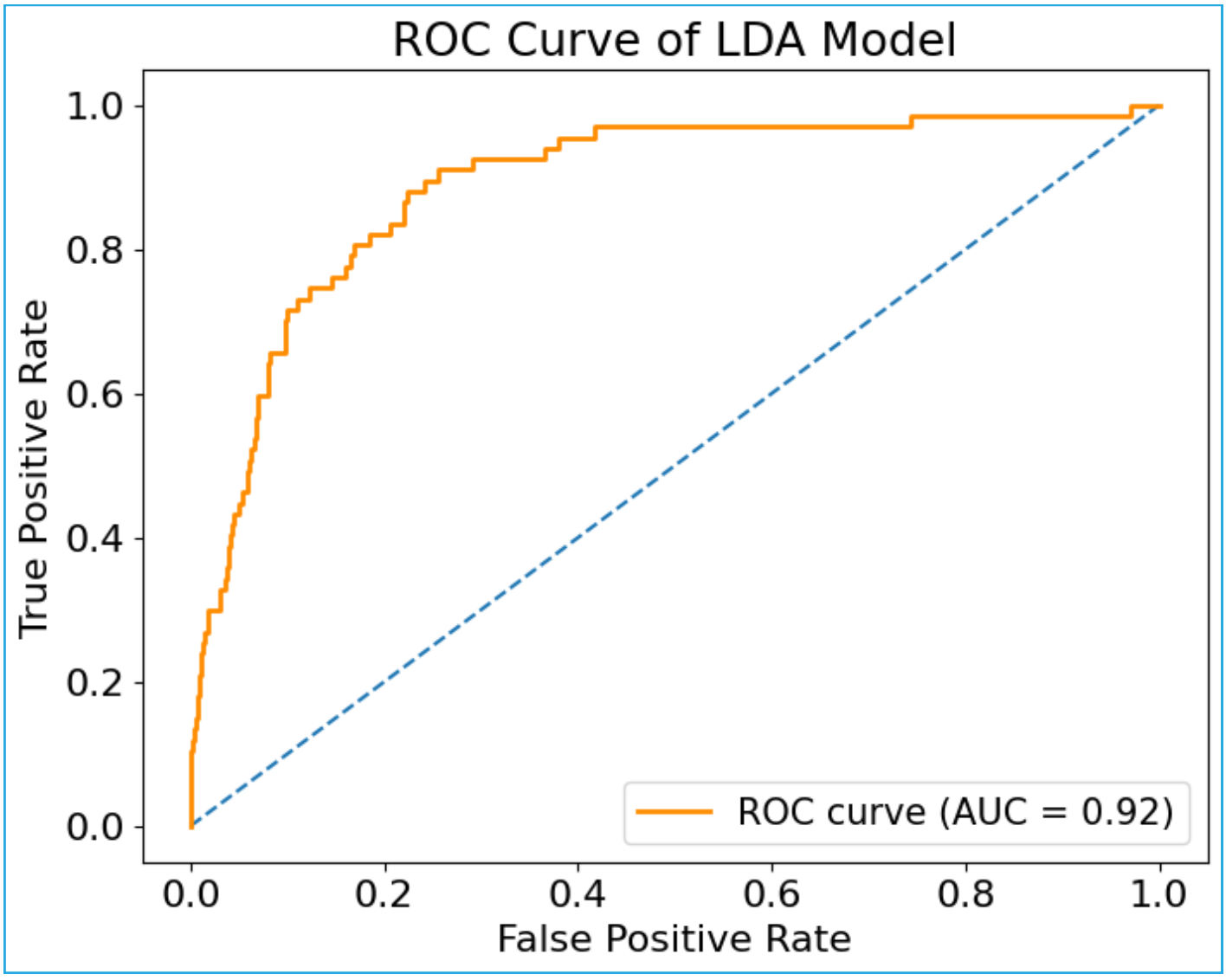


Figure 1.20: ROC curve for LDA of test data

Comparison and Interpretation:

Logistic Regression:

- TP: 50
- FN: 17
- This model has a moderate Recall, capturing a reasonable number of positive instances while keeping the number of false negatives relatively low.
- The false positive rate (FP) is also moderate, indicating that some negative instances were classified as positive.

Random Forest:

- TP: 22
- FN: 45
- This model has a lower Recall compared to Logistic Regression and LDA.
- It correctly identifies fewer positive instances but has fewer false positives.

LDA:

- TP: 54
- FN: 13
- LDA outperforms both Logistic Regression and Random Forest in terms of Recall, as it captures the highest number of positive instances while keeping false negatives relatively low.
- However, it comes at the cost of a higher false positive rate.

In a Default-focused scenario, we prioritize models that maximize Recall (capturing positive instances) while being mindful of false positives. Based on the confusion matrices, the LDA model appears to perform the best in this regard, followed by Logistic Regression, and then Random Forest.

12. Conclusions and Recommendations

Based on the comparison of the Logistic Regression, Random Forest, and Linear Discriminant Analysis (LDA) models we can draw conclusions and make recommendations:

Conclusions:

- LDA Outperforms for Recall: In the context of maximizing Recall (minimizing false negatives), the LDA model consistently outperforms both Logistic Regression and Random Forest. It captures the highest number of positive instances while maintaining a relatively low false negative rate.
- Logistic Regression Offers a Balanced Trade-off: Logistic Regression provides a reasonable balance between Recall and precision. It captures a moderate number of positive instances while keeping false negatives in check. This makes it a good choice when you want to prioritize Recall without excessively increasing false positives.

- Random Forest Lags in Recall: The Random Forest model has a lower Recall compared to the other two models. It correctly identifies fewer positive instances but also has fewer false positives. This model may be suitable when we want to control false positives at the expense of capturing all positive instances.

Recommendations:

- Choose LDA for High Recall: If the primary objective is to maximize Recall, the Linear Discriminant Analysis (LDA) model should be the preferred choice.
- Consider Logistic Regression for Balance: Logistic Regression strikes a balance between Recall and precision.
- Evaluate Random Forest for Specific Needs: While Random Forest may not perform as well in terms of Recall, it should be considered in scenarios where controlling false positives is critical.

- Threshold Tuning: For all models, consider tuning the decision threshold based on the specific needs of your application. Adjusting the threshold can help strike the right balance between Recall and precision.
- Collect More Data: In situations where Recall is crucial, collecting more data, especially positive instances, can help improve model performance across all algorithms.

B: PROBLEM STATEMENT

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

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

Before drawing the stock price graph, let's explore the data first.
Here are top five rows of the data set

	Date	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement
0	31-03-2014	264	69	455	263	68	5543
1	07-04-2014	257	68	458	276	70	5728
2	14-04-2014	254	68	454	270	68	5649
3	21-04-2014	253	68	488	283	68	5692
4	28-04-2014	256	65	482	282	63	5582

Table 2.1: Top five rows of the data

Now looking at the shape of data, dataset has 314 rows and 11 columns.

```
The number of rows (observations) is 314
The number of columns (variables) is 11
```

Table 2.2: Shape the data

Data set is having 10 numeric and one categorical column. Data set is also not having any null values or duplicated rows.

```
RangeIndex: 314 entries, 0 to 313
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Date                  314 non-null    object
 1   Infosys               314 non-null    int64
 2   Indian_Hotel          314 non-null    int64
 3   Mahindra_&_Mahindra   314 non-null    int64
 4   Axis_Bank             314 non-null    int64
 5   SAIL                  314 non-null    int64
 6   Shree_Cement          314 non-null    int64
 7   Sun_Pharma            314 non-null    int64
 8   Jindal_Steel          314 non-null    int64
 9   Idea_Vodafone         314 non-null    int64
 10  Jet_Airways           314 non-null    int64
```

Table 2.3: Datatype information about data

```
The sum of duplicate rows :0
```

Table 2.4: Sum of duplicated rows in data

Here is the statistical summary of numeric columns

	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_&_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 2.5: Statistical summary of numeric columns

Since this is a stock price problem, so we are not concerned with the price of stocks rather we are looking at maximizing profits.

Now look at the stock price graphs of 'Infosys' and 'Idea_Vodaphone'.

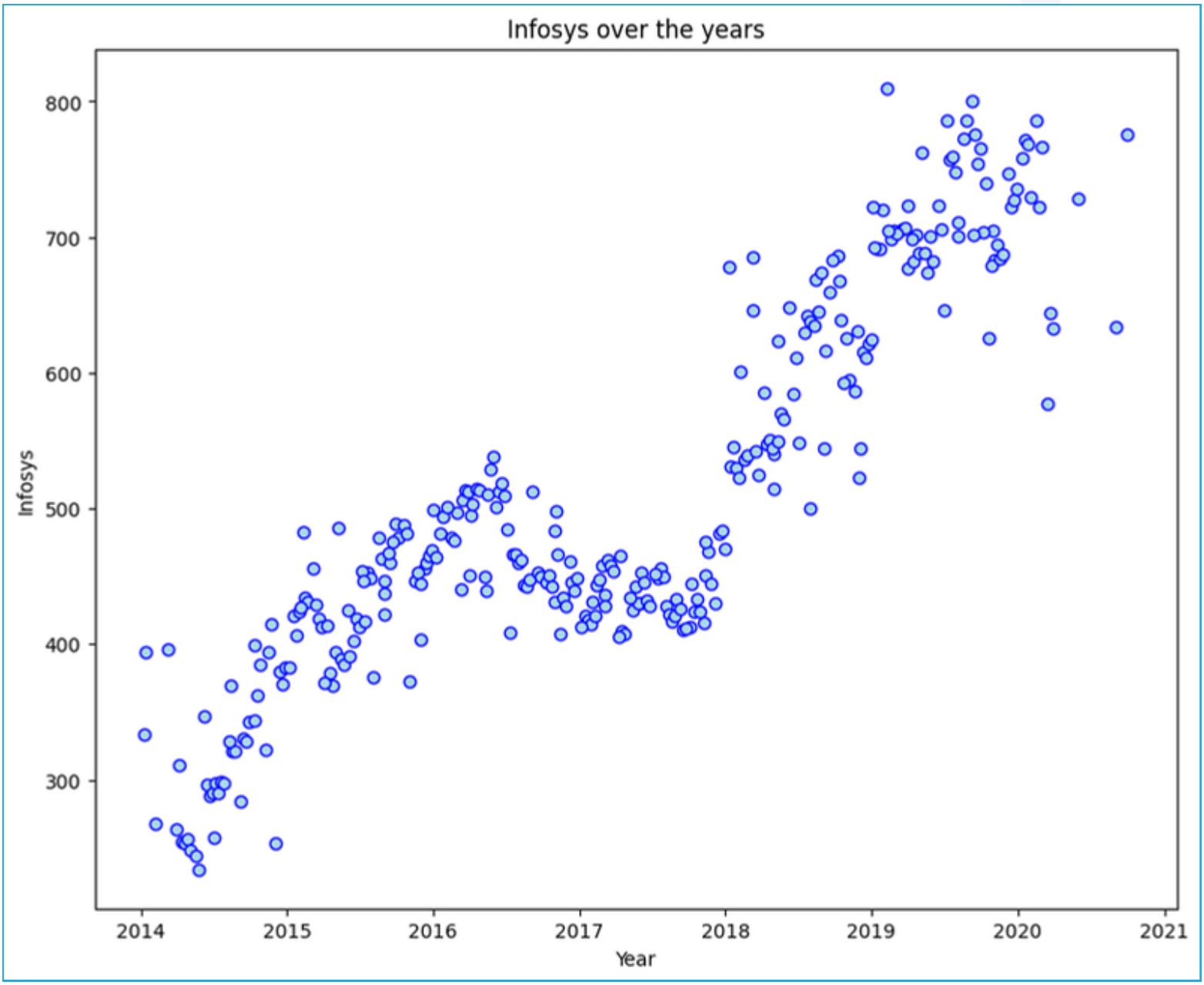


Figure 2.1: Stock price graph for Infosys

Inferences:

- Volatility in the Short Term: In the short term, from March 2014 to early 2015, Infosys stock prices displayed some volatility.
- Steady Growth Phase: From mid-2015 to mid-2016, there appears to be a relatively steady and consistent growth trend in Infosys stock prices.
- Periods of Consolidation: In 2017 and 2018, there are instances of consolidation, where stock prices remained relatively stable or showed limited growth.
- Sharp Decline in Early 2020: In early 2020, Infosys experienced a significant decline in stock prices.
- Overall Positive Trend: Despite short-term fluctuations and the impact of external events like the pandemic, there is an overall positive trend in Infosys stock prices over the analyzed period.

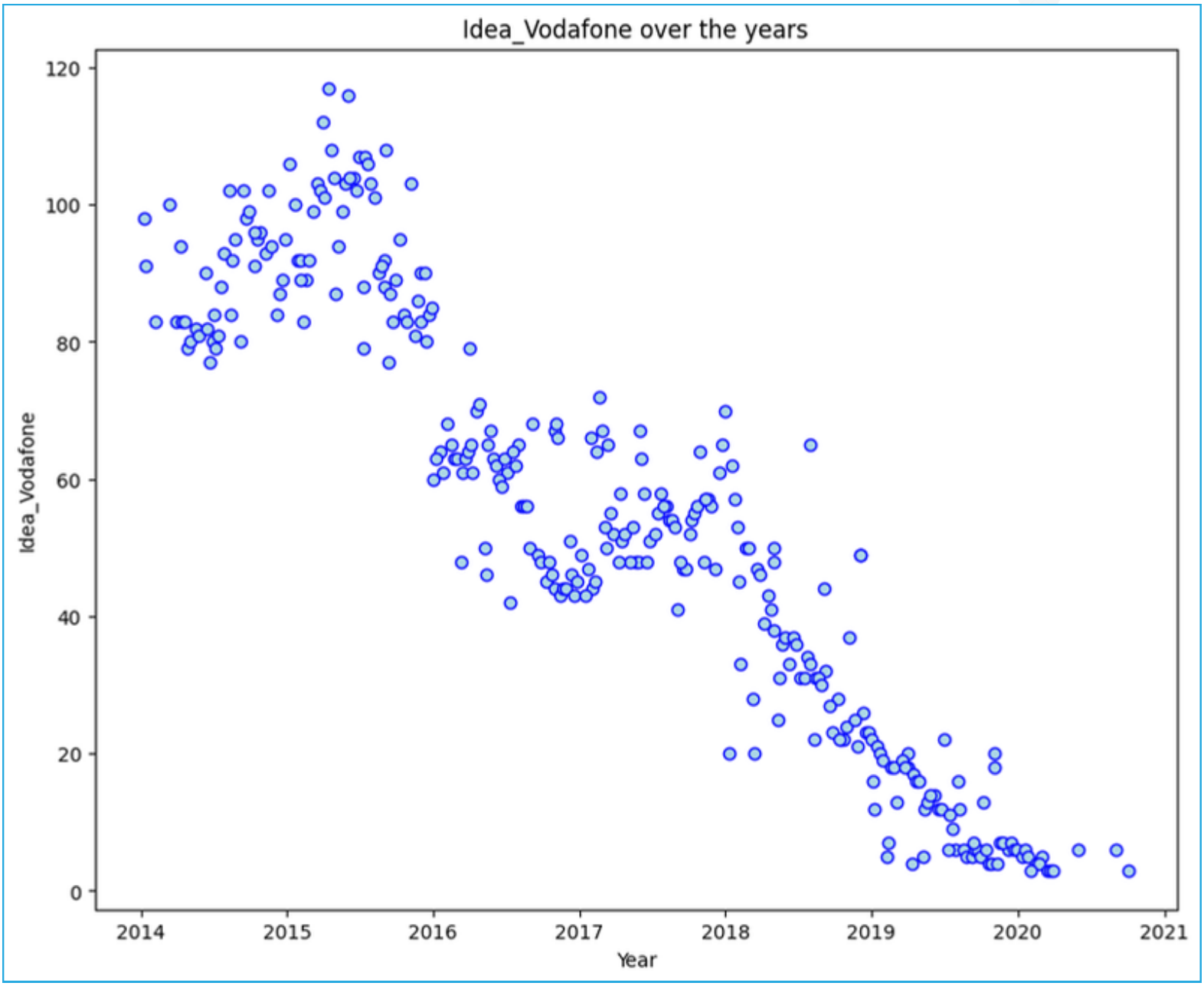


Figure 2.2: Stock price graph for Idea_Vodafone

Inferences:

- Relatively Stable Period (2014-2015): In the early years (2014-2015), Idea Vodafone's stock prices remained relatively stable, with minor fluctuations.
- Sharp Decline (Late 2015-2016): Starting in late 2015 and continuing into 2017, the stock experienced a sharp decline.
- Volatility and Decline (2017-2018): Throughout 2017 and into 2018, the stock displayed high volatility and a continued downward trend.
- Steep Decline (2018-2020): The stock's value plummeted significantly in late 2018 and continued to decline into 2020.
- Overall Positive Trend: The stock price of Idea Vodafone experienced a prolonged period of decline and instability during the analyzed timeframe.

2. Calculate Returns for all stocks with inference

To calculate the weekly returns from all the stocks, we need to calculate the percentage change in stock prices from the previous week.

In other words, we need to calculate the following: *(Current week's closing price - Previous week's closing price) / Previous week's closing price * 100*

Below is the top 5 rows of new data frame showing the percent change in stock prices compared to previous price

Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma
NaN	NaN	NaN	NaN	NaN	NaN	NaN
-2.65	-1.45	0.66	4.94	2.94	3.34	9.91
-1.17	0.00	-0.87	-2.17	-2.86	-1.38	-0.49
-0.39	0.00	7.49	4.81	0.00	0.76	-0.49
1.19	-4.41	-1.23	-0.35	-7.35	-1.93	1.16

Table 2.6: Data frame with change in stock price

Here is the inference from data:

Infosys:

- Initial Price (31-03-2014): 264
- Current Price (14-01-2019): 691
- Return = [(691 - 264) / 264] * 100 ≈ 161.74%

Indian Hotel:

- Initial Price (31-03-2014): 69
- Current Price (14-01-2019): 135
- Return = [(135 - 69) / 69] * 100 ≈ 95.65%

Mahindra & Mahindra:

- Initial Price (31-03-2014): 455
- Current Price (14-01-2019): 712
- Return = [(712 - 455) / 455] * 100 ≈ 56.92%

Axis Bank:

- Initial Price (31-03-2014): 263
- Current Price (14-01-2019): 650
- Return = [(650 - 263) / 263] * 100 ≈ 147.89%

SAIL:

- Initial Price (31-03-2014): 68
- Current Price (14-01-2019): 52
- Return = [(52 - 68) / 68] * 100 ≈ -23.53%

Shree Cement:

- Initial Price (31-03-2014): 5543
- Current Price (14-01-2019): 15749
- Return = [(15749 - 5543) / 5543] * 100 ≈ 184.53%

Sun Pharma:

- Initial Price (31-03-2014): 555
- Current Price (14-01-2019): 439
- Return = [(439 - 555) / 555] * 100 ≈ -20.90%

Jindal Steel:

- Initial Price (31-03-2014): 298
- Current Price (14-01-2019): 161
- Return = [(161 - 298) / 298] * 100 ≈ -46.31%

Idea Vodafone:

- Initial Price (31-03-2014): 83
- Current Price (14-01-2019): 36
- Return = [(36 - 83) / 83] * 100 ≈ -56.63%

Jet Airways:

- Initial Price (31-03-2014): 278
- Current Price (14-01-2019): 239
- Return = [(239 - 278) / 278] * 100 ≈ -14.03%

So stocks like Infosys, Axis bank and Shree Cement has provided exceptionally high returns.
Stocks like Indian Hotel and Mahindra & Mahindra has given good positive return
while stocks like SAIL, Sun Pharma, Jindal Steel, Idea Vodaphone and Jet Airways have given negative returns

3. Calculate Stock Means and Standard Deviation for all stocks with inference

Now we will calculate Stock Means and Stock Standard Deviation to know the average return and volatility of stock.

	Average	Volatility
Infosys	0.34	3.49
Indian_Hotel	0.14	4.69
Mahindra_&_Mahindra	-0.07	3.90
Axis_Bank	0.22	4.50
SAIL	-0.15	6.29
Shree_Cement	0.45	4.02
Sun_Pharma	-0.05	4.46
Jindal_Steel	-0.13	7.51
Idea_Vodafone	-0.51	11.01
Jet_Airways	-0.48	9.65

Table 2.7: Average and Volatility of stocks

Inferences based on this data:

- Infosys (Average: 0.34%, Volatility: 3.49%): Infosys has shown a relatively stable average weekly price change with moderate volatility. This suggests that the stock has been experiencing consistent, albeit modest, price fluctuations.
- Indian Hotel (Average: 0.14%, Volatility: 4.69%): Indian Hotel has a lower average weekly price change compared to Infosys but with higher volatility. This indicates that while the price changes are relatively smaller on average, the stock experiences more significant price swings.
- Mahindra & Mahindra (Average: -0.07%, Volatility: 3.90%): Mahindra & Mahindra has a negative average weekly price change, suggesting a downward trend on average. However, the volatility is moderate, indicating that while the trend is negative, it is not extremely volatile.

- Axis Bank (Average: 0.22%, Volatility: 4.50%): Axis Bank displays a positive average weekly price change with moderate volatility. This suggests that the stock has been on an upward trend on average, with relatively consistent but moderate price fluctuations.
- SAIL (Average: -0.15%, Volatility: 6.29%): SAIL has a negative average weekly price change with relatively high volatility. This indicates that the stock experiences significant price swings, and the overall trend is bearish.
- Shree Cement (Average: 0.45%, Volatility: 4.02%): Shree Cement has the highest average weekly price change among the listed stocks, indicating strong positive momentum. The volatility is moderate, suggesting that the stock is experiencing significant price increases but with some fluctuations.

- Sun Pharma (Average: -0.05%, Volatility: 4.46%): Sun Pharma has a slightly negative average weekly price change with moderate volatility. This implies a relatively stable but slightly bearish trend in the stock.
- Jindal Steel (Average: -0.13%, Volatility: 7.51%): Jindal Steel has a negative average weekly price change with high volatility. The stock experiences substantial price swings, and the overall trend is bearish.
- Idea Vodafone (Average: -0.51%, Volatility: 11.01%): Idea Vodafone has the most negative average weekly price change and the highest volatility among the listed stocks. This indicates a strong bearish trend with significant price fluctuations, making it a highly volatile stock.
- Jet Airways (Average: -0.48%, Volatility: 9.65%): Jet Airways also exhibits a strong negative average weekly price change and high volatility. The stock is in a bearish trend with substantial price swings.

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

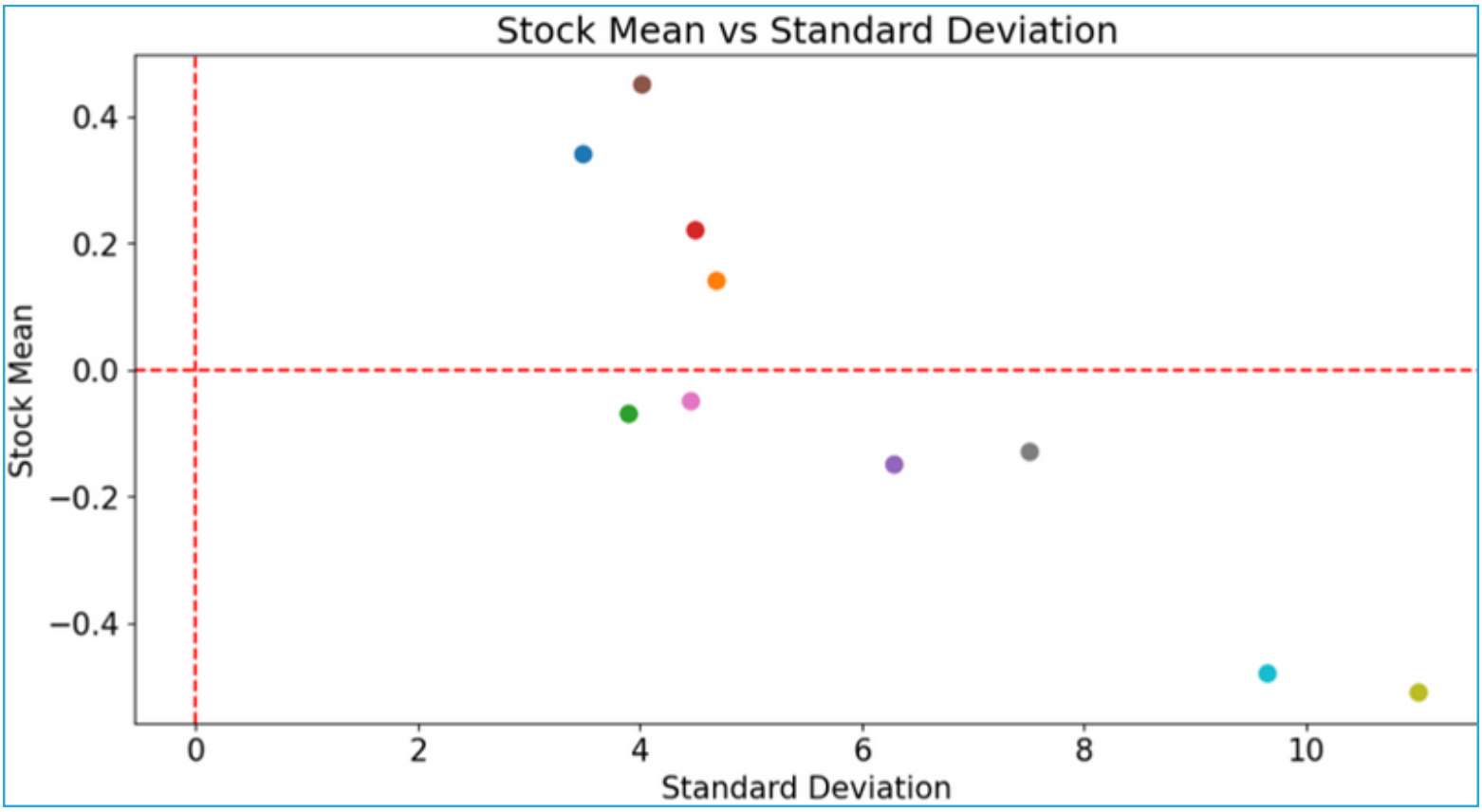


Figure 2.3: Stock Mean vs Standard deviation

- The plot shows the relationship between Stock Mean and Standard deviation for different stocks.
- Stocks are represented by different colors in the plot (shown in the legend).
- It appears that there is a general trend where stocks with lower Stock Mean tend to have higher Standard deviation.
- "Idea_Vodafone" and "Jet_Airways" stand out with the lowest Stock Mean and relatively high Standard deviation, indicating significant negative performance.
- "Shree_Cement" has a relatively high Stock Mean and moderate Standard deviation, suggesting it performed well with lower Standard deviation.

5. Conclusions and Recommendations

Conclusions:

- Stock Performance: Shree Cement stands out as having both a high average return and relatively low volatility, making it an attractive option for investors seeking a balance between returns and risk
- High Volatility Stocks: Stocks like Jindal Steel, Idea Vodafone, and Jet Airways exhibit high volatility, indicating a higher level of risk. Investors should exercise caution when considering these stocks.
- Diversification: Diversifying a portfolio by including stocks with different risk-return profiles, such as combining Shree Cement with lower-performing but less volatile stocks, may help manage overall portfolio risk.
- Stock Selection Strategy: Investors with a higher risk appetite might consider high-volatility stocks for potentially higher returns, but they should also be prepared for greater fluctuations in the stock's value.

Recommendations:

- Diversify Your Portfolio: Consider a diversified portfolio that includes a mix of stocks with varying levels of risk and return. This can help spread risk and potentially provide a more stable overall return.
- Risk Management: If you are risk-averse, prioritize stocks with lower volatility like Shree Cement or Infosys. However, if you are comfortable with risk, you can explore opportunities in high-volatility stocks but be prepared for greater market fluctuations.
- Research and Monitoring: Continuously monitor the performance of the stocks in your portfolio. Stay informed about market trends, news, and company developments that may impact stock prices.
- Long-Term vs. Short-Term: Consider your investment horizon. Stocks with high volatility may be suitable for short-term trading, while less volatile stocks might be better for long-term investments.