Financial Risk Analytics

PROJECT REPORT

HARI HARAN

20h AUG, 2023

Great Learning

CONTENTS

`

Part A :

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.

Dataset: CompData-1.xlsx

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.

PART A: Outlier Treatment.

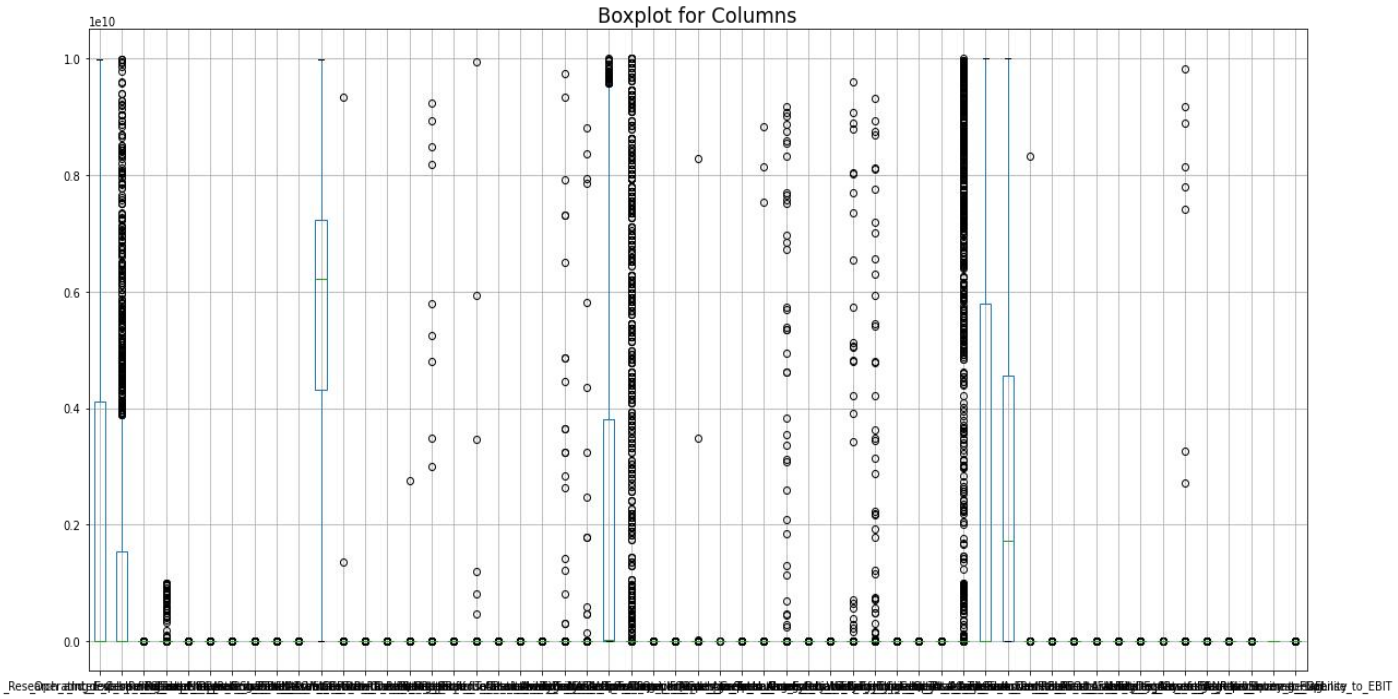


Fig.A.1. Boxplot with Outliers

The above plot shows presence of outliers in most of the features.
For this reason outlier treatment is necessary so that we can have unbiased accuracy during model building.

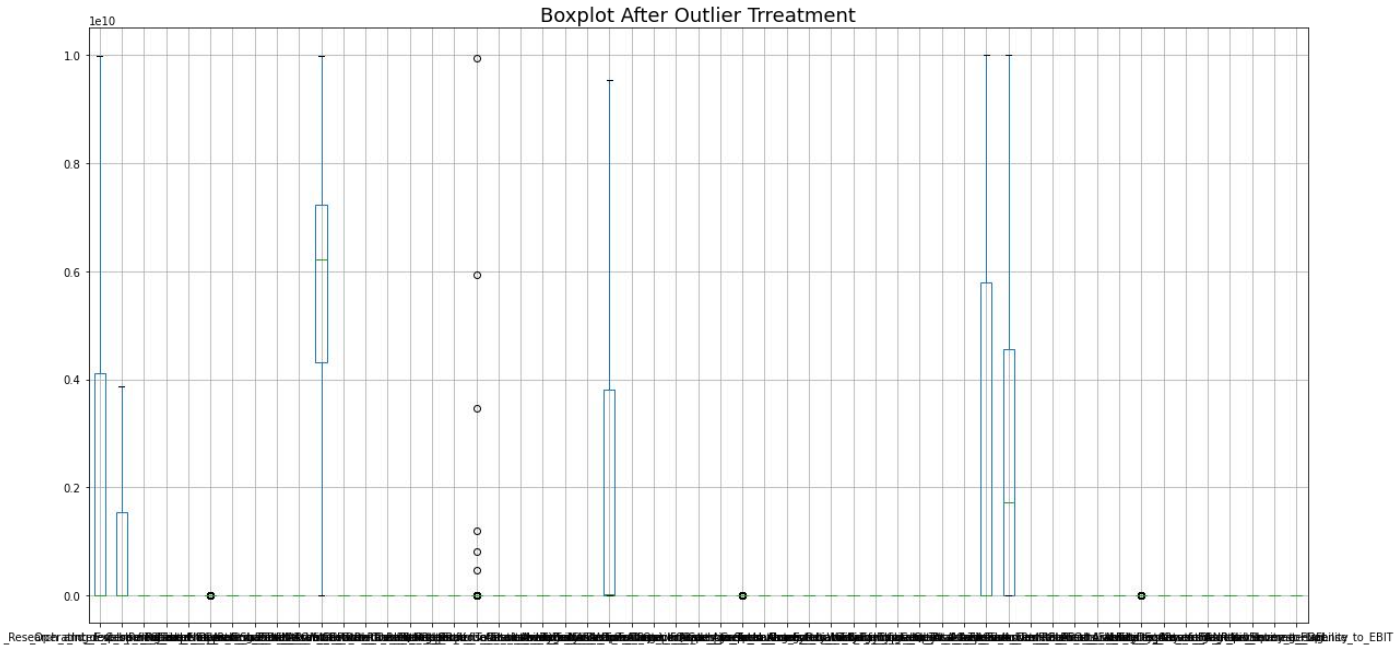


Fig.A.2. Boxplot after outlier treatment

- The above boxplot shows we have treated the outliers within upper and lower limits using IQR.
- Now the current data is almost ready for further analysis.
- But there are missing values is the data which needs to be imputed.
- There are a total of **298 missing values**.

Cash_Flow_Per_Share	67
_Total_debt_to_Total_net_worth	21
_Cash_to_Total_Assets	96
_Current_Liability_to_Current_Assets	14

PART A: Missing Value Treatment.

To treat the missing values we have used **Standard Scaler** to scale the data " **cdf_x** ".

Percentage of missing values= (298 / 119364)*100 = 0.24%, where 119364 is the size of the data.

Although, only 0.24% data seems insignificant value but for further process the these cannot ignored. As it may , statistically speaking, affect the final model building and may have significance.

After scaling is done we have used **KNN Imputer** to impute the missing the values in the data. The imputation of missing data is important so that we can use the data from model building.

The **Fig.A.3.** (Page 6) shows the missing values highlighted with yellow lines in the heatmap plot. While **Fig.A.4.** (Page 6) is the evidence of imputation of missing values.

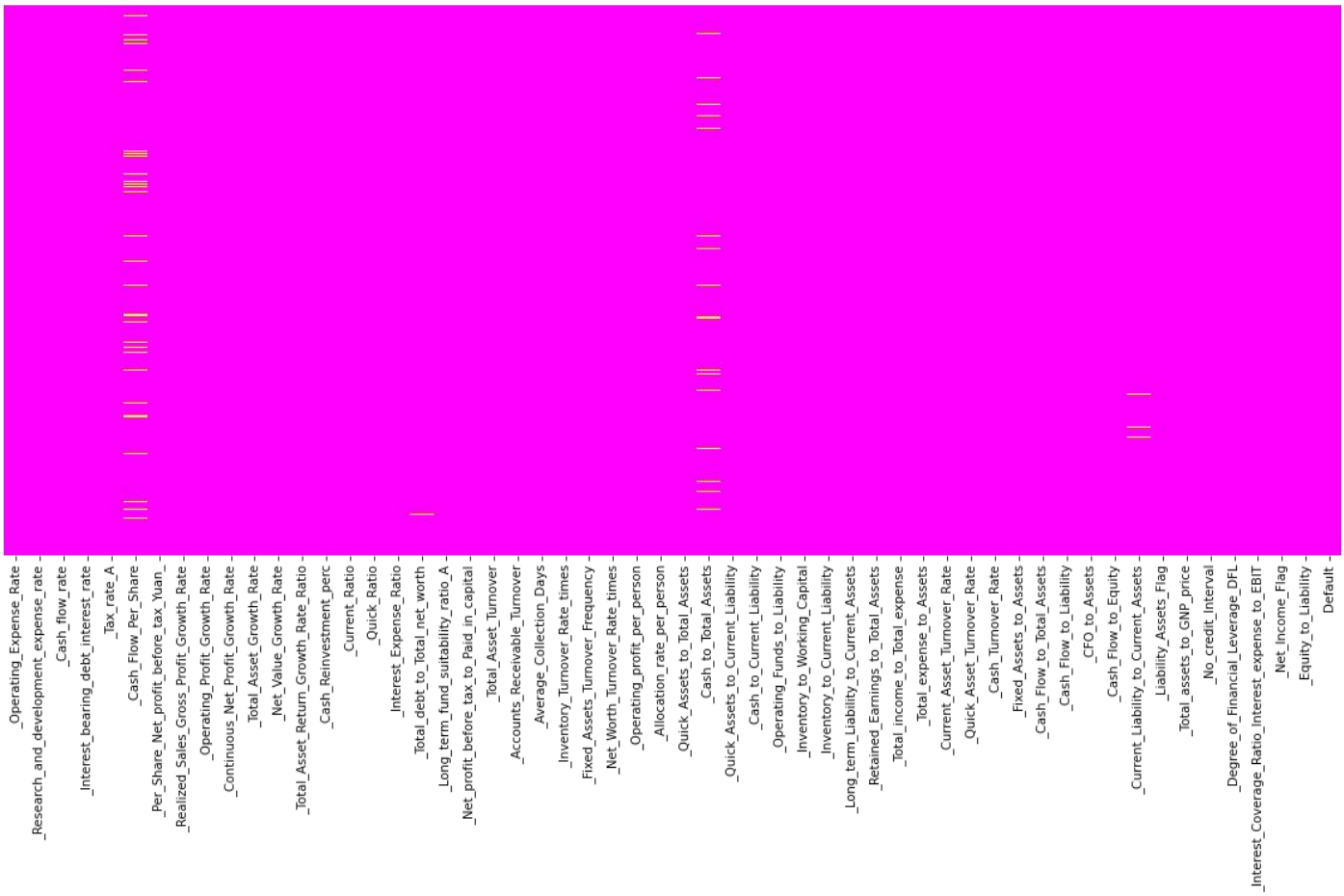


Fig.A.3. Heat Map with with missing values

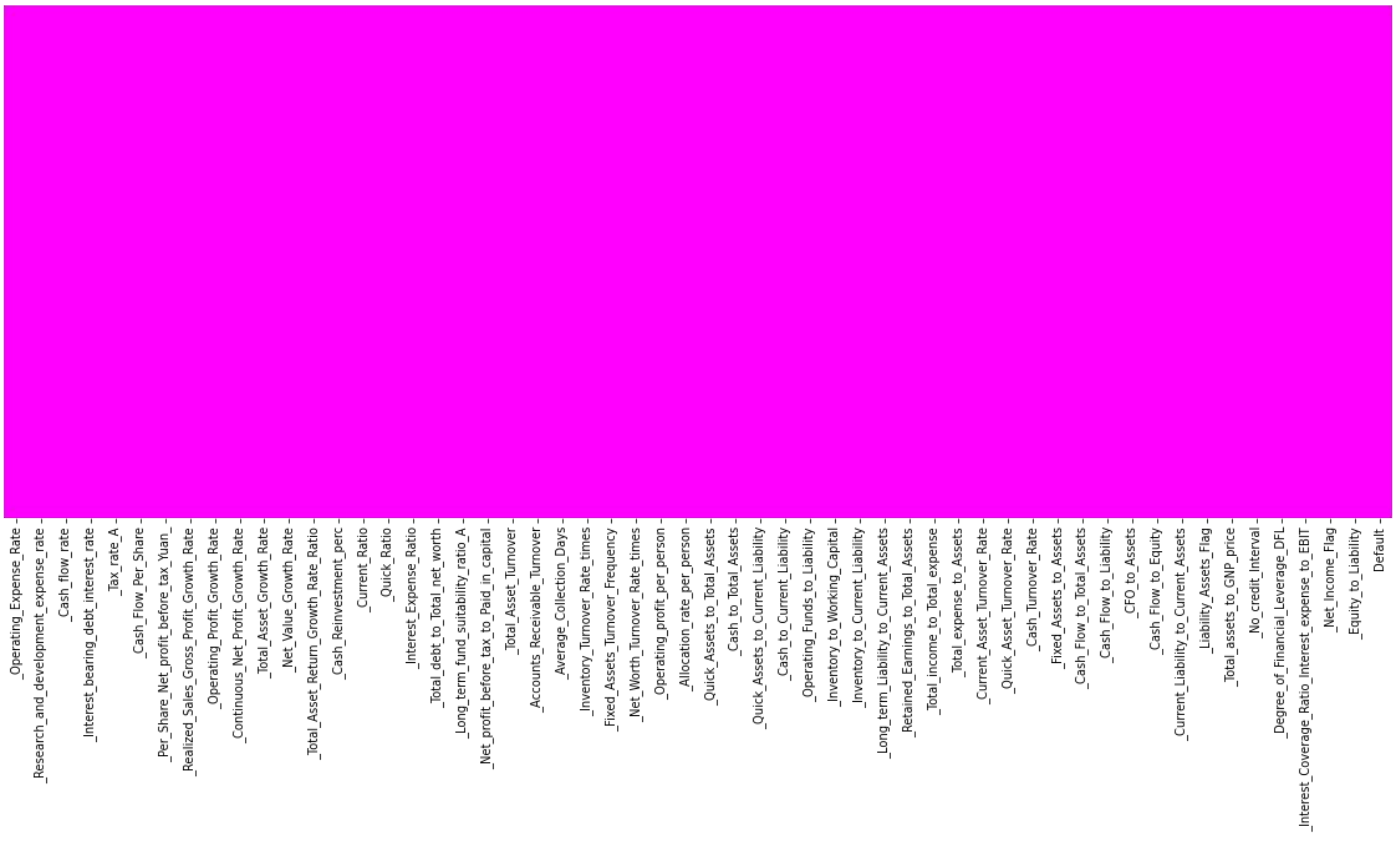


Fig.A.4. Heat Map with after imputing the missing values

PART A: Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building).

*Please not that the below Univariate analysis and Bivariate analysis is based on the final model for the variables whose **VIF< 5** and **P-value <0.05**.*

Based on Model_32 the significant variables are :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 9 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Default                               2058 non-null   int64  
 1   _Total_income_to_Total_expense        2058 non-null   float64
 2   _Cash_Reinvestment_perc               2058 non-null   float64
 3   _Retained_Earnings_to_Total_Assets    2058 non-null   float64
 4   _Allocation_rate_per_person           2058 non-null   float64
 5   _Accounts_Receivable_Turnover         2058 non-null   float64
 6   _Total_expense_to_Assets              2058 non-null   float64
 7   _Research_and_development_expense_rate 2058 non-null   float64
 8   _Interest_bearing_debt_interest_rate  2058 non-null   float64
dtypes: float64(8), int64(1)
memory usage: 144.8 KB
```

Table.A.1. Data Info - All Significant Variables.

	count	mean	std	min	25 %	50 %	75%	max
Default	2058.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00
_Total_income_to_Total_expense	2058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
_Cash_Reinvestment_perc	2058.00	0.38	0.03	0.03	0.37	0.38	0.39	1.00
_Retained_Earnings_to_Total_Assets	2058.00	0.93	0.03	0.00	0.93	0.94	0.94	0.97
_Allocation_rate_per_person	2058.00	5725558.82	197949961.06	0.00	0.00	0.01	0.02	8280000000.00
_Accounts_Receivable_Turnover	2058.00	41598639.46	504767266.59	0.00	0.00	0.00	0.00	9740000000.00
_Total_expense_to_Assets	2058.00	0.03	0.04	0.00	0.01	0.02	0.04	1.00
_Research_and_development_expense_rate	2058.00	1208634256.56	2144568158.08	0.00	0.00	0.00	1550000000.00	9980000000.00
_Interest_bearing_debt_interest_rate	2058.00	11130223.52	90425949.04	0.00	0.00	0.00	0.00	990000000.00

Table.A.2. Data Description - All Significant Variables.

Univariate Analysis,

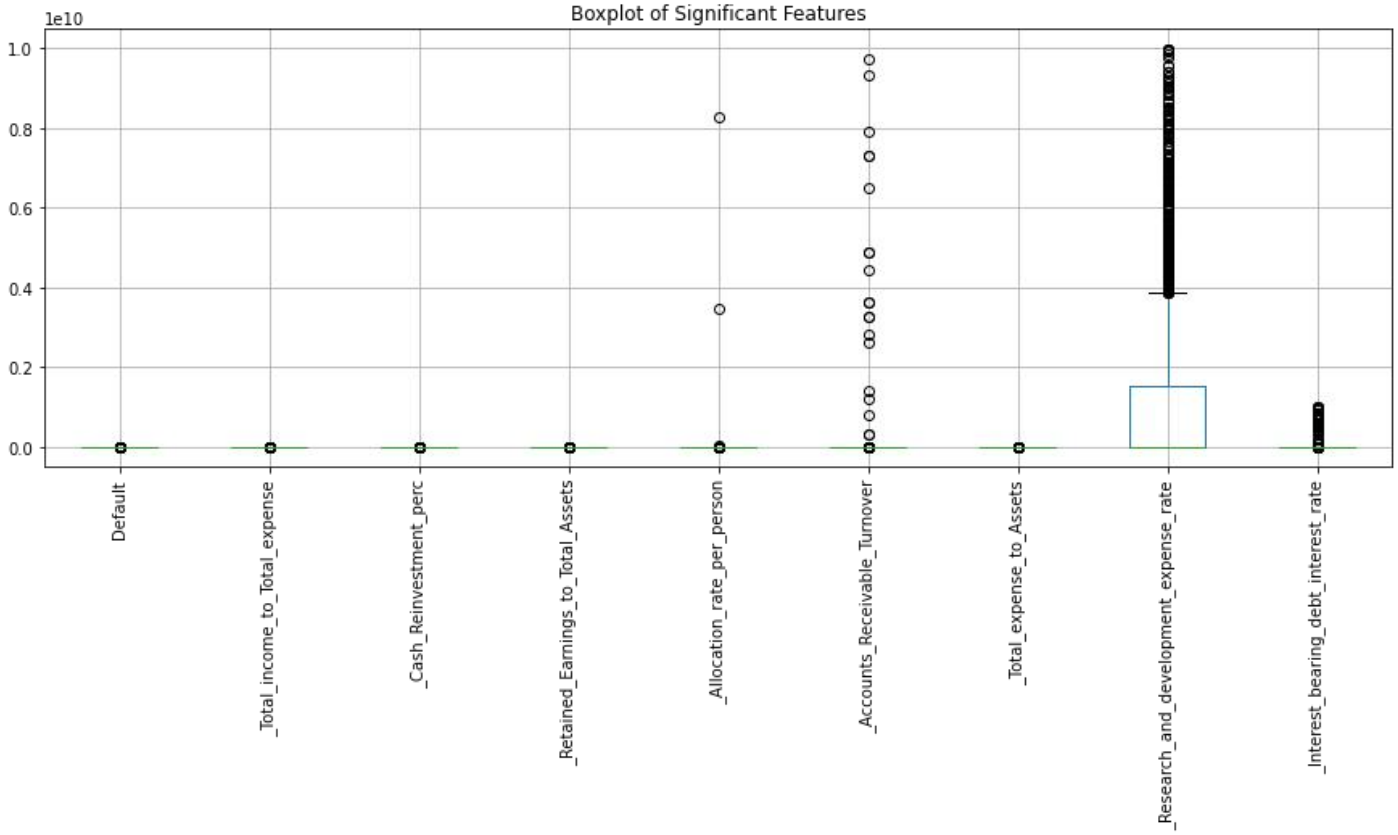


Fig.A.5. Boxplot of Significant Variables.

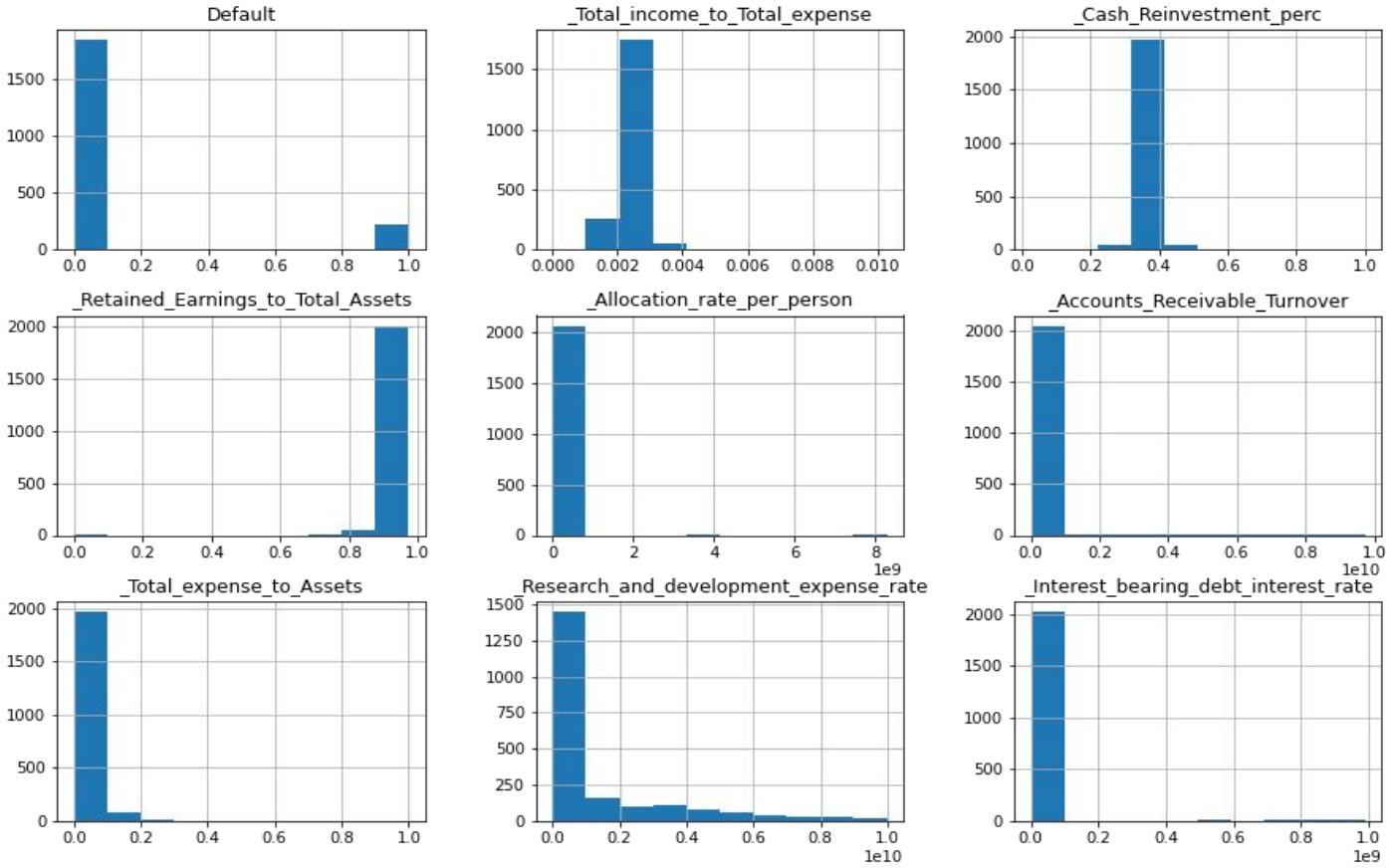


Fig.A.6. Histogram plot of Significant Variables.

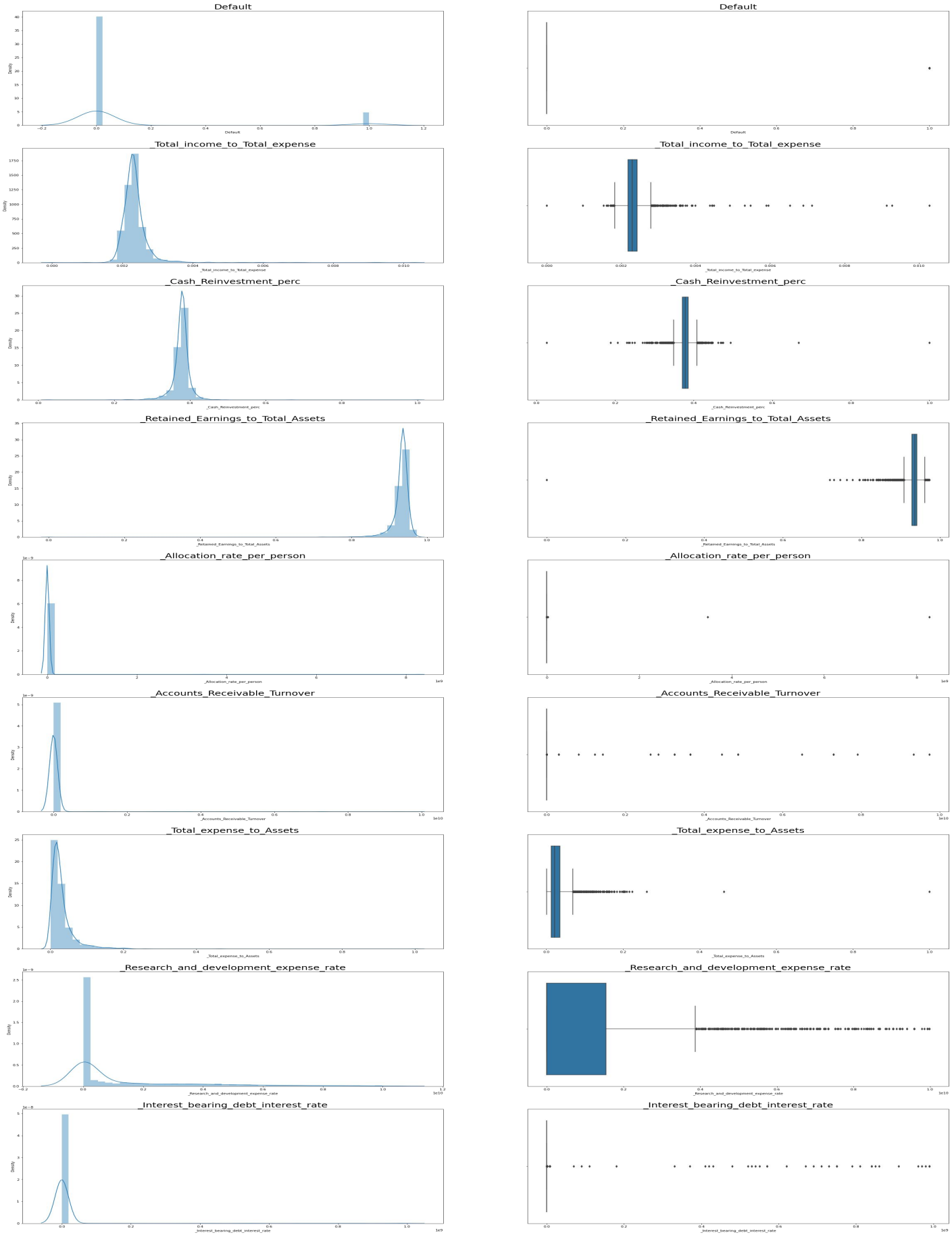
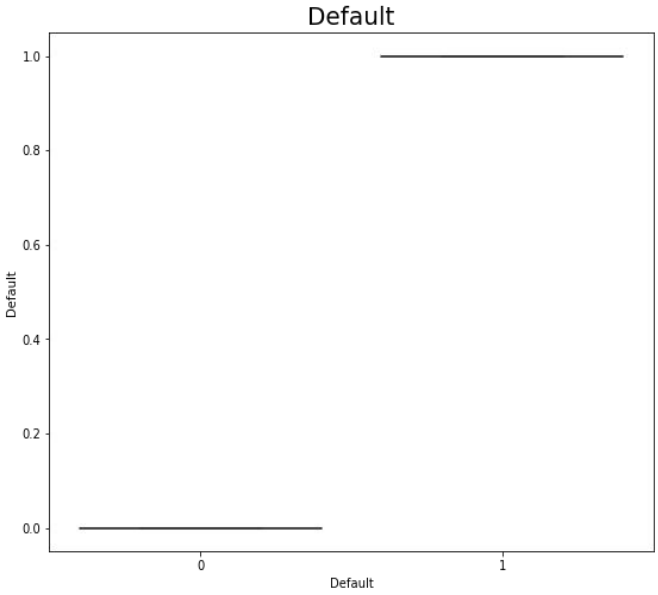
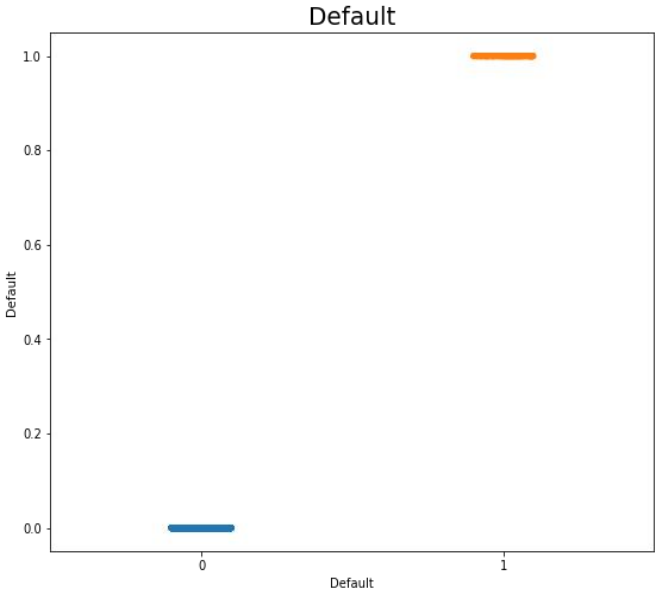
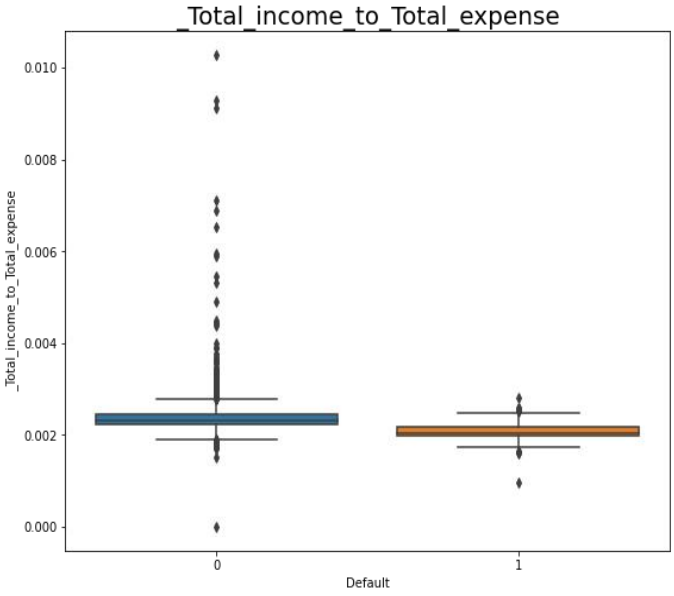


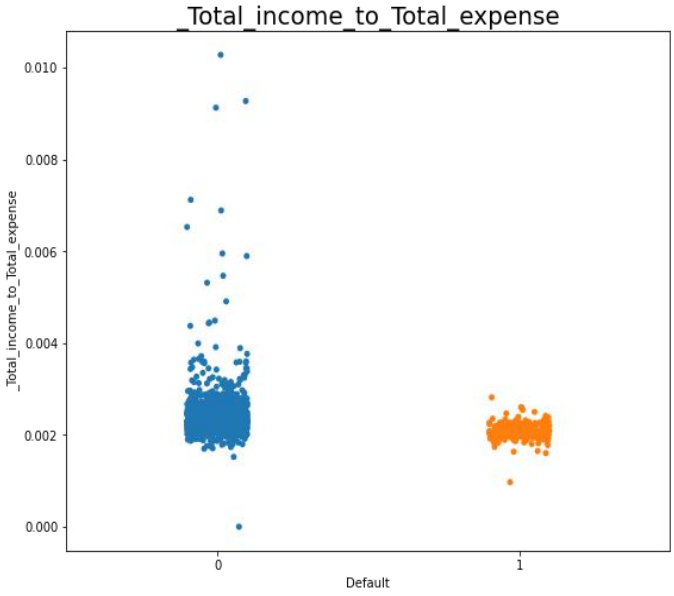
Fig.A.7. Distribution and Boxplot plot of Significant Variables.

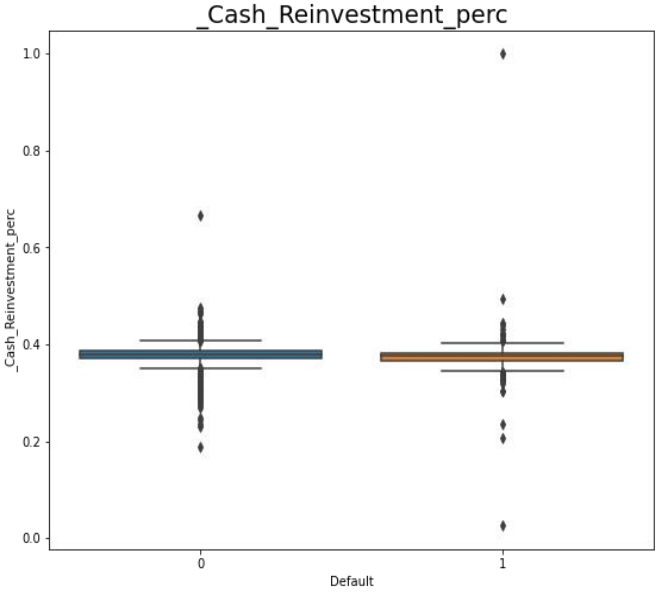
Bivariate Analysis,

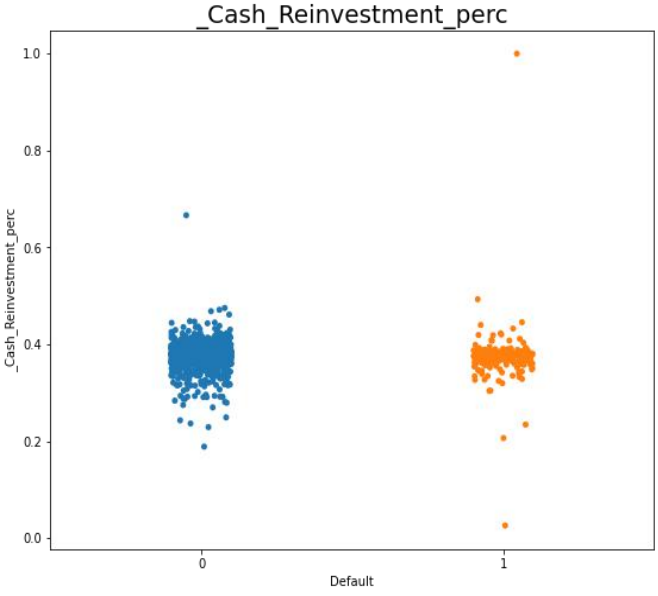


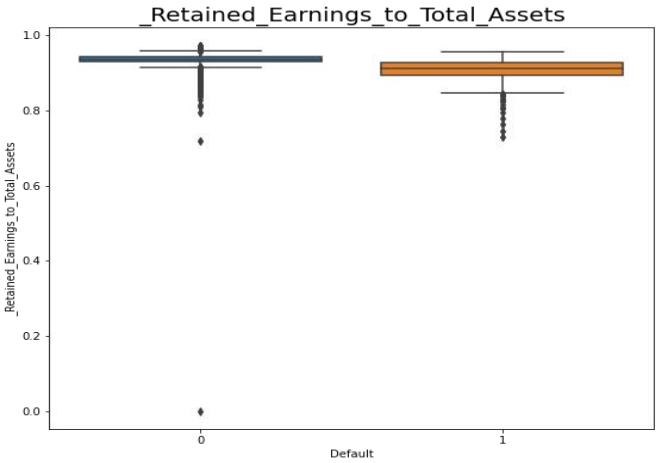


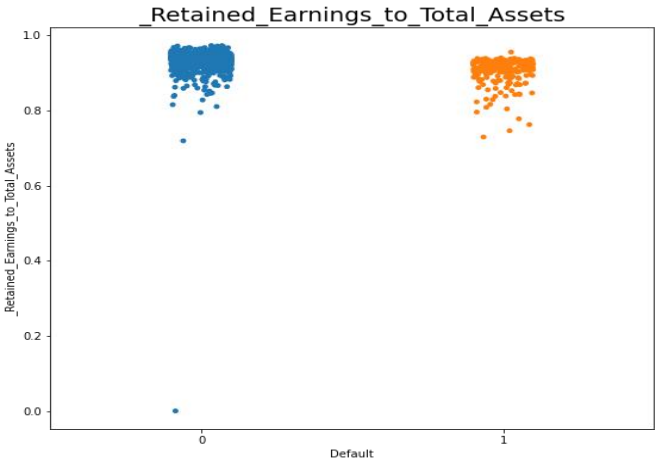


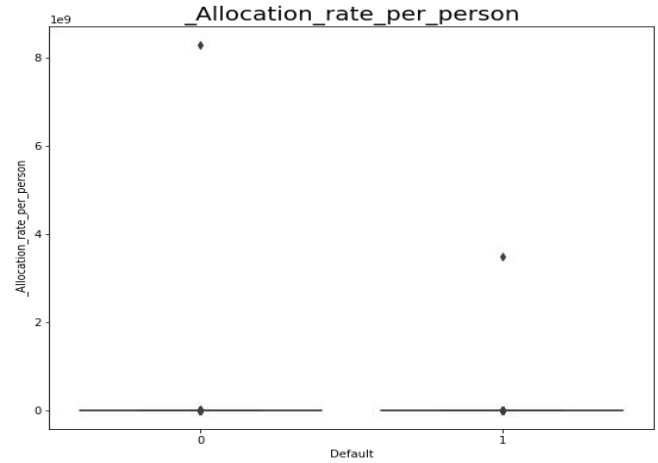


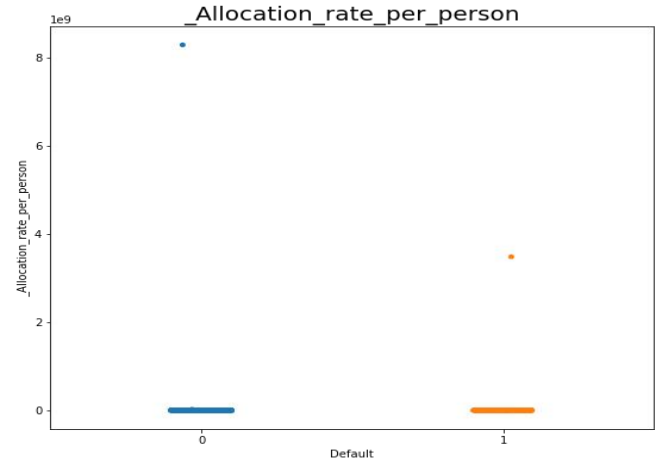


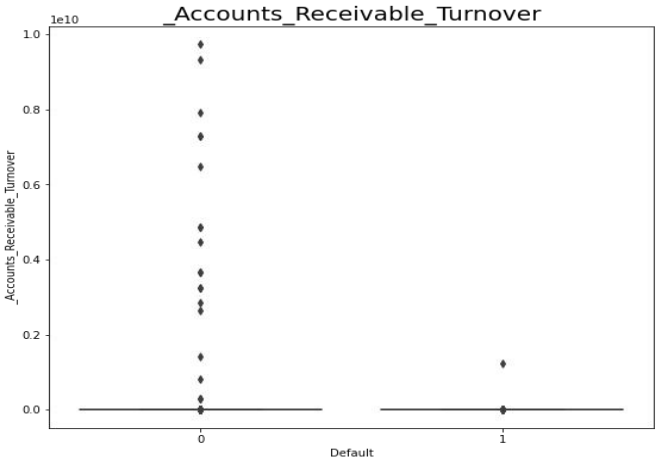


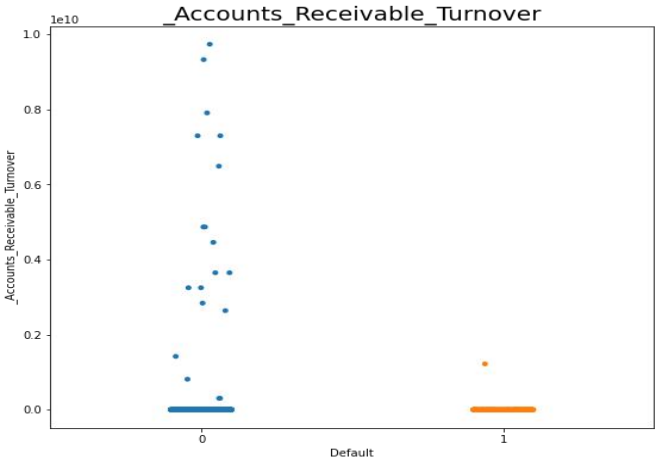

_Retained_Earnings_to_Total_Assets
_Retained_Earnings_to_Total_Assets
Default

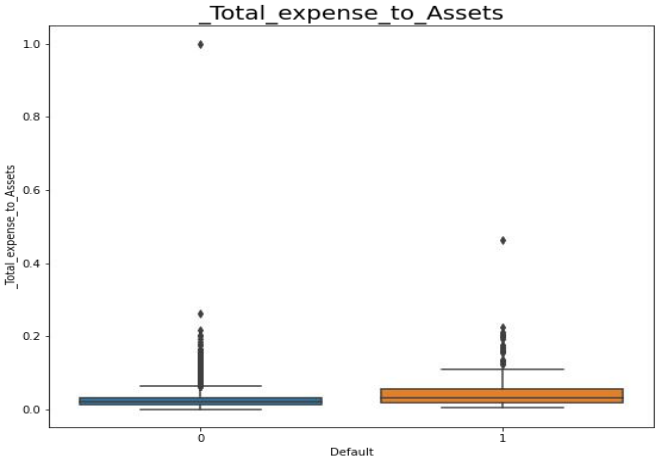

_Retained_Earnings_to_Total_Assets
_Retained_Earnings_to_Total_Assets
Default

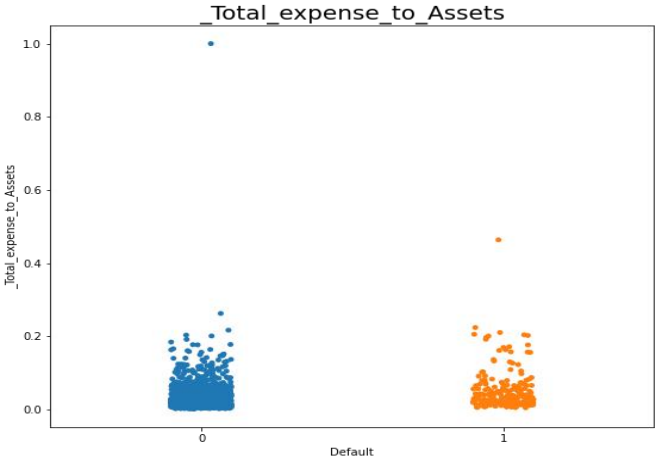

_Allocation_rate_per_person
1e9
_Allocation_rate_per_person
Default


_Allocation_rate_per_person
1e9
_Allocation_rate_per_person
Default


_Accounts_Receivable_Turnover
1e10
_Accounts_Receivable_Turnover
Default


_Accounts_Receivable_Turnover
1e10
_Accounts_Receivable_Turnover
Default


_Total_expense_to_Assets
_Total_expense_to_Assets
Default


_Total_expense_to_Assets
_Total_expense_to_Assets
Default

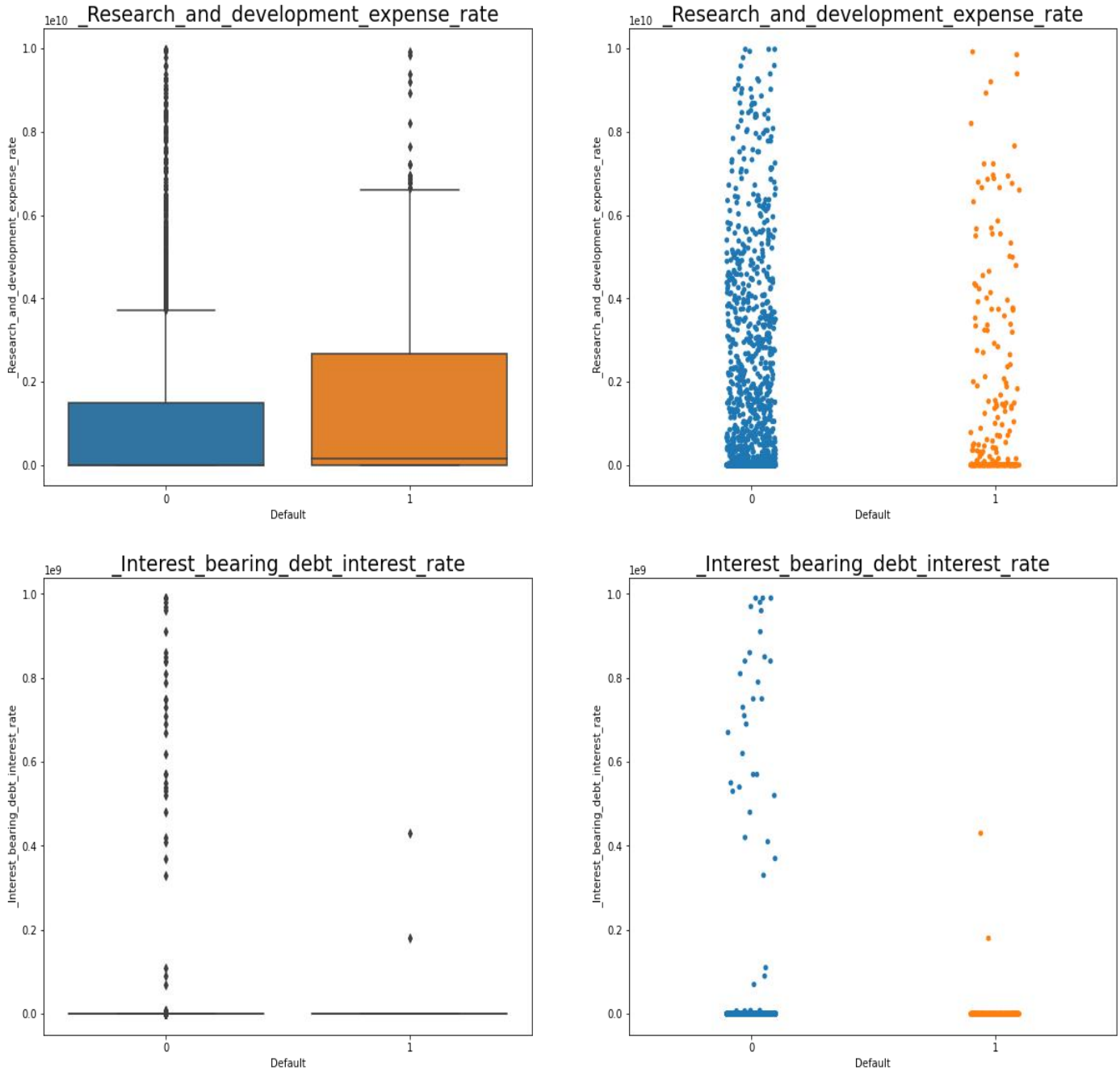


Fig.A.8. Boxplot and Strip Plot of Significant Variables w.r.t. "Default".

Inferences :

- The original data set contains 89% non-defaulters and 11% defaulters.
- The above bivariate and univariate analysis shows the presence of outliers through box plot.
- The distribution shows almost normal distribution of the data achieved through scaling and imputation of the data.
- The bivariate analysis shows that the non-defaulters are less likely to default unless there is some unpredictable circumstances which may cause financial liability.
- On the other hand, the boxplot and strip plot shows that there are high chances of defaulters may continue to default so investment in such companies are risky.
- The non-defaulters have higher earning w.r.t. to defaulters as they have less obligation to debts.

PART A: Train Test Split:

The train and test data is split into 67:33

```
cdf_x = cdf.drop(['Default','Co_Code','Co_Name'], axis = 1)

cdf_y = cdf['Default']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =0.33, random_state=42)
```

PART A: 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.

Logistic Regression Model (using statsmodels library) ,

We are using **VIF (Variance Inflation Factor)**which is a measure to detect multicollinearity in multiple regression models. It tells us how an independent variable is a linear combination of other independent variables. It can also quantify how much the variance of a regression coefficient has inflated due to its correlations with other predictors indicating multicollinearity.

Higher the VIF value higher is the multicollinearity. So, for the present case ,i.e., for Financial Risk Analytics (FRA), we have to tried to find out the significant which **VIF** less than 5 and also **P-value** less than 0.05.

So for the steps involved we have one-by-one checked the VIF from model_1 to model_32 to find the variables which has VIF < 5. Higher multicollinearity gives biased results in model building so to avoid biased results we choose the optimum mode,i.e., **model_32** .

For VIF for all columns (without dropping any columns) in descending order,

	variables	VIF
6	_Per_Share_Net_profit_before_tax_Yuan_	98.99
19	_Net_profit_before_tax_to_Paid_in_capital	98.75
43	_Cash_Flow_to_Total_Assets	44.45
45	_CFO_to_Assets	28.30
32	_Operating_Funds_to_Liability	21.22
30	_Quick_Assets_to_Current_Liability	19.90
44	_Cash_Flow_to_Liability	17.86
2	_Cash_flow_rate	16.60
46	_Cash_Flow_to_Equity	15.13
15	_Quick_Ratio	12.45
13	_Cash_Reinvestment_perc	12.32

14	_Current_Ratio	11.00
20	_Total_Asset_Turnover	10.96
25	_Net_Worth_Turnover_Rate_times	10.57
28	_Quick_Assets_to_Total_Assets	6.37
52	_Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.06
54	_Equity_to_Liability	5.35
36	_Retained_Earnings_to_Total_Assets	5.24
37	_Total_income_to_Total_expense	5.22
42	_Fixed_Assets_to_Assets	4.84
16	_Interest_Expense_Ratio	4.58
8	_Operating_Profit_Growth_Rate	3.66
31	_Cash_to_Current_Liability	3.59
9	_Continuous_Net_Profit_Growth_Rate	3.46
51	_Degree_of_Financial_Leverage_DFL	3.34
34	_Inventory_to_Current_Liability	3.24
12	_Total_Asset_Return_Growth_Rate_Ratio	3.23
29	_Cash_to_Total_Assets	3.10
26	_Operating_profit_per_person	3.06
5	_Cash_Flow_Per_Share	3.03
18	_Long_term_fund_suitability_ratio_A	2.96
7	_Realized_Sales_Gross_Profit_Growth_Rate	2.81
27	_Allocation_rate_per_person	2.74
11	_Net_Value_Growth_Rate	2.64
21	_Accounts_Receivable_Turnover	2.63
22	_Average_Collection_Days	2.61
38	_Total_expense_to_Assets	2.27
24	_Fixed_Assets_Turnover_Frequency	1.93
35	_Long_term_Liability_to_Current_Assets	1.78
50	_No_credit_Interval	1.75
49	_Total_assets_to_GNP_price	1.73
39	_Current_Asset_Turnover_Rate	1.65
47	_Current_Liability_to_Current_Assets	1.61
4	_Tax_rate_A	1.60
33	_Inventory_to_Working_Capital	1.45

40	_Quick_Asset_Turnover_Rate	1.41
0	_Operating_Expense_Rate	1.33
23	_Inventory_Turnover_Rate_times	1.22
1	_Research_and_development_expense_rate	1.20
10	_Total_Asset_Growth_Rate	1.18
3	_Interest_bearing_debt_interest_rate	1.12
41	_Cash_Turnover_Rate	1.11
17	_Total_debt_to_Total_net_worth	1.07
48	_Liability_Assets_Flag	NaN
53	_Net_Income_Flag	NaN

Table.A.3. VIF with all the columns

After dropping all columns with VIF > 5,

	variables	VIF
35	_Fixed_Assets_to_Assets	4.43
30	_Total_income_to_Total_expense	4.25
13	_Quick_Ratio	4.17
43	_Equity_to_Liability	3.89
12	_Cash_Reinvestment_perc	3.88
7	_Operating_Profit_Growth_Rate	3.61
29	_Retained_Earnings_to_Total_Assets	3.58
2	_Cash_flow_rate	3.47
8	_Continuous_Net_Profit_Growth_Rate	3.44
25	_Cash_to_Current_Liability	3.35
11	_Total_Asset_Return_Growth_Rate_Ratio	3.00
22	_Operating_profit_per_person	2.93
16	_Long_term_fund_suitability_ratio_A	2.85
21	_Net_Worth_Turnover_Rate_times	2.80
6	_Realized_Sales_Gross_Profit_Growth_Rate	2.75
5	_Cash_Flow_Per_Share	2.73
23	_Allocation_rate_per_person	2.65
24	_Cash_to_Total_Assets	2.61
17	_Accounts_Receivable_Turnover	2.58
41	_Degree_of_Financial_Leverage_DFL	2.40

14	_Interest_Expense_Ratio	2.38
18	_Average_Collection_Days	2.37
10	_Net_Value_Growth_Rate	2.37
31	_Total_expense_to_Assets	2.17
20	_Fixed_Assets_Turnover_Frequency	1.90
27	_Inventory_to_Current_Liability	1.70
39	_Total_assets_to_GNP_price	1.69
28	_Long_term_Liability_to_Current_Assets	1.64
40	_No_credit_Interval	1.59
32	_Current_Asset_Turnover_Rate	1.54
4	_Tax_rate_A	1.48
37	_Current_Liability_to_Current_Assets	1.46
26	_Inventory_to_Working_Capital	1.43
36	_Cash_Flow_to_Liability	1.37
33	_Quick_Asset_Turnover_Rate	1.37
0	_Operating_Expense_Rate	1.30
19	_Inventory_Turnover_Rate_times	1.21
1	_Research_and_development_expense_rate	1.17
9	_Total_Asset_Growth_Rate	1.15
3	_Interest_bearing_debt_interest_rate	1.10
34	_Cash_Turnover_Rate	1.10
15	_Total_debt_to_Total_net_worth	1.05
38	_Liability_Assets_Flag	NaN
42	_Net_Income_Flag	NaN

Table.A.4. VIF < 5 - after dropping columns

variables	count	mean	std	min	25%	50%	75%	max
_Operating_Expense_Rate	2058.00	0.00	1.00	-0.63	-0.63	-0.63	0.63	2.44
_Research_and_development_expense_rate	2058.00	0.00	1.00	-0.65	-0.65	-0.65	0.43	2.04
_Cash_flow_rate	2058.00	0.00	1.00	-2.18	-0.58	-0.13	0.49	2.09
_Interest_bearing_debt_interest_rate	2058.00	0.00	1.00	-1.63	-0.70	-0.10	0.60	2.56
_Tax_rate_A	2058.00	-0.00	1.00	-0.82	-0.82	-0.54	0.78	3.19
_Cash_Flow_Per_Share	2058.00	-0.00	0.97	-9.84	-0.32	0.05	0.37	9.30
_Realized_Sales_Gross_Profit_Growth_Rate	2058.00	-0.00	1.00	-2.05	-0.55	-0.10	0.45	1.96

_Operating_Profit_Growth_Rate	2058.00	-0.00	1.00	-1.98	-0.50	-0.05	0.48	1.96
_Continuous_Net_Profit_Growth_Rate	2058.00	0.00	1.00	-1.91	-0.45	0.01	0.52	1.98
_Total_Asset_Growth_Rate	2058.00	0.00	1.00	-1.82	-0.33	0.32	0.66	1.61
_Net_Value_Growth_Rate	2058.00	-0.00	1.00	-1.97	-0.51	-0.15	0.47	1.94
_Total_Asset_Return_Growth_Rate_Ratio	2058.00	-0.00	1.00	-2.09	-0.53	-0.02	0.51	2.07
_Cash_Reinvestment_perc	2058.00	-0.00	1.00	-2.09	-0.52	0.07	0.53	2.10
_Quick_Ratio	2058.00	-0.00	1.00	-1.31	-0.73	-0.28	0.43	2.18
_Interest_Expense_Ratio	2058.00	-0.00	1.00	-1.84	-0.43	-0.27	0.51	1.93
_Total_debt_to_Total_net_worth	2058.00	-0.00	1.00	-0.04	-0.04	-0.04	-0.04	36.83
_Long_term_fund_suitability_ratio_A	2058.00	0.00	1.00	-1.68	-0.74	-0.41	0.41	2.12
_Accounts_Receivable_Turnover	2058.00	-0.00	1.00	-1.46	-0.72	-0.39	0.37	2.02
_Average_Collection_Days	2058.00	-0.00	1.00	-1.63	-0.72	-0.11	0.56	2.49
_Inventory_Turnover_Rate_times	2058.00	-0.00	1.00	-0.66	-0.66	-0.65	0.58	2.45
_Fixed_Assets_Turnover_Frequency	2058.00	-0.00	1.00	-0.66	-0.64	-0.59	0.31	1.74
_Net_Worth_Turnover_Rate_times	2058.00	-0.00	1.00	-1.33	-0.74	-0.32	0.47	2.29
_Operating_profit_per_person	2058.00	-0.00	1.00	-1.91	-0.48	-0.11	0.47	1.89
_Allocation_rate_per_person	2058.00	0.00	1.00	-1.03	-0.75	-0.40	0.43	2.19
_Cash_to_Total_Assets	2058.00	-0.00	0.99	-0.81	-0.60	-0.35	0.17	8.57
_Cash_to_Current_Liability	2058.00	-0.00	1.00	-0.91	-0.73	-0.45	0.40	2.11
_Inventory_to_Working_Capital	2058.00	0.00	1.00	-2.19	-0.59	-0.22	0.47	2.07
_Inventory_to_Current_Liability	2058.00	-0.00	1.00	-1.14	-0.77	-0.28	0.47	2.34
_Long_term_Liability_to_Current_Assets	2058.00	-0.00	1.00	-0.78	-0.78	-0.48	0.42	2.22
_Retained_Earnings_to_Total_Assets	2058.00	-0.00	1.00	-2.01	-0.44	0.14	0.61	2.18
_Total_income_to_Total_expense	2058.00	-0.00	1.00	-2.25	-0.59	-0.10	0.51	2.17
_Total_expense_to_Assets	2058.00	0.00	1.00	-1.37	-0.74	-0.31	0.45	2.23
_Current_Asset_Turnover_Rate	2058.00	-0.00	1.00	-0.80	-0.67	-0.58	0.32	1.79
_Quick_Asset_Turnover_Rate	2058.00	-0.00	1.00	-0.74	-0.74	-0.74	0.93	2.15
_Cash_Turnover_Rate	2058.00	-0.00	1.00	-0.94	-0.94	-0.33	0.67	2.60
_Fixed_Assets_to_Assets	2058.00	0.00	1.00	-1.26	-0.81	-0.27	0.66	2.86
_Cash_Flow_to_Liability	2058.00	0.00	1.00	-1.82	-0.45	-0.05	0.47	1.84
_Current_Liability_to_Current_Assets	2058.00	-0.00	1.00	-0.82	-0.37	-0.14	0.10	20.03
_Liability_Assets_Flag	2058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
_Total_assets_to_GNP_price	2058.00	-0.00	1.00	-0.93	-0.76	-0.46	0.41	2.15
_No_credit_Interval	2058.00	0.00	1.00	-1.81	-0.41	0.14	0.52	1.92

_Degree_of_Financial_Leverage_DFL	2058.00	-0.00	1.00	-1.75	-0.42	-0.28	0.47	1.81
_Net_Income_Flag	2058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
_Equity_to_Liability	2058.00	-0.00	1.00	-1.60	-0.75	-0.34	0.44	2.22

Table.A.5. Data Description for Columns with VIF < 5

Applying Logistic Regression with Statsmodel library,

Model_1 :

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	680
Model:	Logit	**Df Residuals:**	637
Method:	MLE	**Df Model:**	42
Date:	Sat, 19 Aug 2023	**Pseudo R-squ.:**	0.5066
Time:	23:19:57	**Log-Likelihood:**	-107.98
converged:	False	**LL-Null:**	-218.85
Covariance Type:	nonrobust	**LLR p-value:**	2.795e-26

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-5.1161	2.53e+05	-2.02e-05	1.000	-4.96e+05	4.96e+05
_Operating_Expense_Rate	0.0596	0.226	0.263	0.792	-0.384	0.503
_Research_and_development_expense_rate	0.3432	0.206	1.668	0.095	-0.060	0.746
_Cash_flow_rate	-0.2540	0.463	-0.549	0.583	-1.161	0.653
_Interest_bearing_debt_interest_rate	-0.0537	0.232	-0.232	0.817	-0.508	0.401
_Tax_rate_A	-0.1383	0.255	-0.542	0.588	-0.639	0.362
_Cash_Flow_Per_Share	-0.4443	0.577	-0.771	0.441	-1.574	0.686
_Realized_Sales_Gross_Profit_Growth_Rate	0.0020	0.253	0.008	0.994	-0.494	0.498
_Operating_Profit_Growth_Rate	0.3062	0.301	1.016	0.310	-0.285	0.897
_Continuous_Net_Profit_Growth_Rate	-0.8373	0.330	-2.539	0.011	-1.484	-0.191
_Total_Asset_Growth_Rate	0.2774	0.256	1.086	0.278	-0.223	0.778
_Net_Value_Growth_Rate	-0.4935	0.241	-2.045	0.041	-0.967	-0.020
_Total_Asset_Return_Growth_Rate_Ratio	0.2773	0.286	0.969	0.333	-0.284	0.838
_Cash_Reinvestment_perc	0.5836	0.437	1.335	0.182	-0.273	1.440
_Quick_Ratio	0.1873	0.531	0.353	0.724	-0.853	1.228
_Interest_Expense_Ratio	-0.0144	0.243	-0.059	0.953	-0.491	0.462

_Total_debt_to_Total_net_worth	-5.7337	6.36e+06	-9.01e-07	1.000	-1.25e+07	1.25e+07
_Long_term_fund_suitability_ratio_A	0.6785	0.348	1.950	0.051	-0.003	1.360
_Accounts_Receivable_Turnover	-0.7843	0.378	-2.073	0.038	-1.526	-0.043
_Average_Collection_Days	-0.0225	0.354	-0.063	0.949	-0.716	0.671
_Inventory_Turnover_Rate_times	-0.0385	0.217	-0.178	0.859	-0.463	0.386
_Fixed_Assets_Turnover_Frequency	0.1122	0.270	0.416	0.677	-0.416	0.641
_Net_Worth_Turnover_Rate_times	-0.3323	0.316	-1.051	0.293	-0.952	0.287
_Operating_profit_per_person	-0.2634	0.344	-0.765	0.444	-0.939	0.412
_Allocation_rate_per_person	0.0022	0.318	0.007	0.995	-0.620	0.625
_Cash_to_Total_Assets	0.2891	0.438	0.660	0.509	-0.570	1.148
_Cash_to_Current_Liability	0.0289	0.355	0.082	0.935	-0.666	0.724
_Inventory_to_Working_Capital	0.0242	0.183	0.132	0.895	-0.334	0.383
_Inventory_to_Current_Liability	0.0705	0.280	0.252	0.801	-0.478	0.619
_Long_term_Liability_to_Current_Assets	0.2211	0.224	0.988	0.323	-0.218	0.660
_Retained_Earnings_to_Total_Assets	-1.0912	0.395	-2.762	0.006	-1.866	-0.317
_Total_income_to_Total_expense	-0.0222	0.560	-0.040	0.968	-1.120	1.076
_Total_expense_to_Assets	0.3321	0.283	1.172	0.241	-0.223	0.887
_Current_Asset_Turnover_Rate	0.1330	0.225	0.591	0.555	-0.308	0.574
_Quick_Asset_Turnover_Rate	0.1896	0.228	0.832	0.405	-0.257	0.636
_Cash_Turnover_Rate	-0.9423	0.286	-3.293	0.001	-1.503	-0.381
_Fixed_Assets_to_Assets	0.3733	0.423	0.882	0.378	-0.456	1.203
_Cash_Flow_to_Liability	0.0145	0.288	0.050	0.960	-0.550	0.579
_Current_Liability_to_Current_Assets	0.0392	0.232	0.169	0.866	-0.415	0.494
_Total_assets_to_GNP_price	0.3622	0.252	1.435	0.151	-0.133	0.857
_No_credit_Interval	0.0908	0.220	0.413	0.679	-0.340	0.522
_Degree_of_Financial_Leverage_DFL	0.3312	0.264	1.256	0.209	-0.186	0.848
_Equity_to_Liability	-2.6631	0.711	-3.747	0.000	-4.056	-1.270

Table.A.6. Model_1

Please note there are lots of variables which have P-value > 0.05. So we have optimized the model up to 32nd iteration to obtain the significant variables which have P-value < 0.05.

Model_32 :

After dropping the insignificant columns, we get

Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1366
Method:	MLE	**Df Model:**	11
Date:	Sun, 20 Aug 2023	**Pseudo R-squ.:**	0.4307
Time:	00:31:11	**Log-Likelihood:**	-273.52
converged:	True	**LL-Null:**	-480.46
Covariance Type:	nonrobust	**LLR p-value:**	6.904e-82

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.1868	0.266	-15.722	0.000	-4.709	-3.665
_Total_income_to_Total_expense	-1.0671	0.271	-3.935	0.000	-1.599	-0.536
_Quick_Ratio	-0.7482	0.240	-3.116	0.002	-1.219	-0.278
_Equity_to_Liability	-1.0776	0.267	-4.033	0.000	-1.601	-0.554
_Cash_Reinvestment_perc	-0.3557	0.109	-3.267	0.001	-0.569	-0.142
_Retained_Earnings_to_Total_Assets	-0.8801	0.205	-4.298	0.000	-1.281	-0.479
_Operating_profit_per_person	0.4480	0.188	2.377	0.017	0.079	0.817
_Allocation_rate_per_person	0.7054	0.138	5.108	0.000	0.435	0.976
_Accounts_Receivable_Turnover	-0.6219	0.139	-4.482	0.000	-0.894	-0.350
_Total_expense_to_Assets	0.4029	0.149	2.708	0.007	0.111	0.695
_Research_and_development_expense_rate	0.3895	0.111	3.520	0.000	0.173	0.606
_Interest_bearing_debt_interest_rate	0.3878	0.142	2.739	0.006	0.110	0.665

Table.A.7. Model_32 with *P < 0.05*

So now we have obtained **P-value < 0.05 at model_32 with VIF< 5**, optimised for the regression model.

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

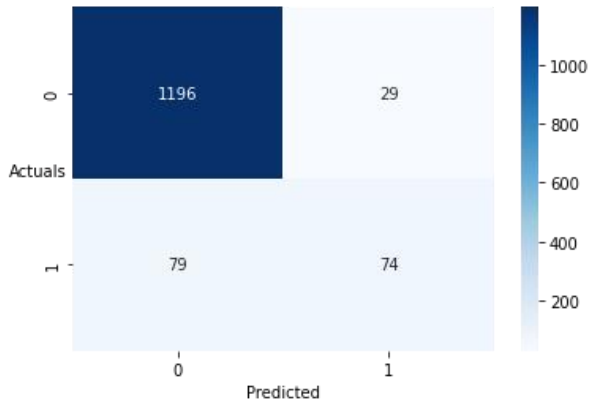


Fig.A.9. Confusion Matrix- Logistic regression for Model_32

	precision	recall	f1-score	support
0.0	0.938	0.976	0.957	1225
1.0	0.718	0.484	0.578	153
accuracy			0.922	1378
macro avg	0.828	0.730	0.767	1378
weighted avg	0.914	0.922	0.915	1378

Precision = 71%
Recall = 48%

Validating train and test data,

For FRA, better recall value is important than precison value.
Now optimising threshold using ROC curve, we obtain optimal_threshold = 0.10709
We will use the above threshold value to validate our train and test data.

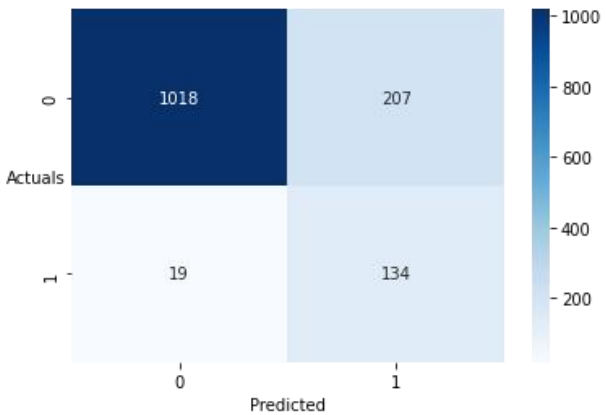


Fig.A.10. Confusion Matrix- Logistic regression for train data

	precision	recall	f1-score	support
0.0	0.982	0.831	0.900	1225
1.0	0.393	0.876	0.543	153
accuracy			0.836	1378
macro avg	0.687	0.853	0.721	1378
weighted avg	0.916	0.836	0.860	1378

Precison = 39%
Recall = 87%

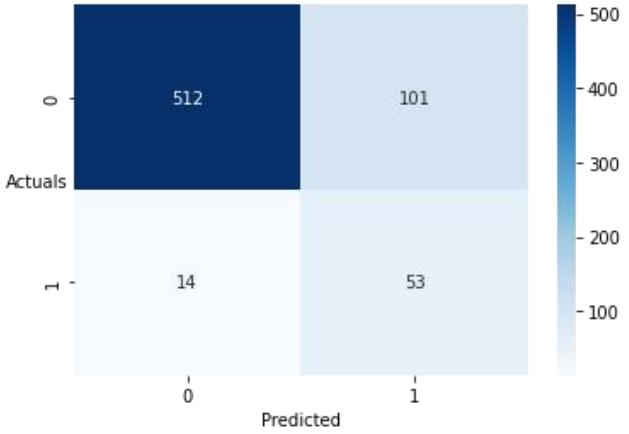


Fig.A.11. Confusion Matrix- Logistic regression for test data

	precision	recall	f1-score	support
0.0	0.973	0.835	0.899	613
1.0	0.344	0.791	0.480	67
accuracy			0.831	680
macro avg	0.659	0.813	0.689	680
weighted avg	0.911	0.831	0.858	680

Precison = 34%
Recall = 79%

Now we have obtained a better recall value after optimising the threshold value for train data,i.e., 48% to 87% which is good. In financial credit risk analysis, a high recall value and a lower precision value is acceptable.

In both train and test data we have obtained recall around 80% which shows that there is no over-fitting in t he model. Hence, statistically speaking it's a good model.

The ROC-AUC Score for model_32 on the test set is 0.90.

PART A: Build a Random Forest Model on Train Dataset. Also showcase your m odel building approach.

Random Forest Model :

Applying grid-search we obtain,

```
{'max_depth': 5,
 'min_samples_leaf': 5,
 'min_samples_split': 45,
 'n_estimators': 25}
```

For train data,

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	1225
1.0	0.86	0.42	0.56	153
accuracy			0.93	1378
macro avg	0.90	0.71	0.76	1378
weighted avg	0.92	0.93	0.92	1378

Precison = 86%
Recall = 42%

For test data,

	precision	recall	f1-score	support
0.0	0.93	0.98	0.96	613
1.0	0.68	0.31	0.43	67
accuracy			0.92	680
macro avg	0.80	0.65	0.69	680
weighted avg	0.90	0.92	0.90	680

Precison = 68%
Recall = 31%

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

For train data,

	precision	recall	f1-score	support
0.0	0.93	0.99	0.96	1225
1.0	0.86	0.42	0.56	153
accuracy			0.93	1378
macro avg	0.90	0.71	0.76	1378
weighted avg	0.92	0.93	0.92	1378

Precison = 86%
Recall = 42%

For test data,

	precision	recall	f1-score	support
0.0	0.93	0.98	0.96	613
1.0	0.68	0.31	0.43	67
accuracy			0.92	680
macro avg	0.80	0.65	0.69	680
weighted avg	0.90	0.92	0.90	680

Precison = 68%
Recall = 31%

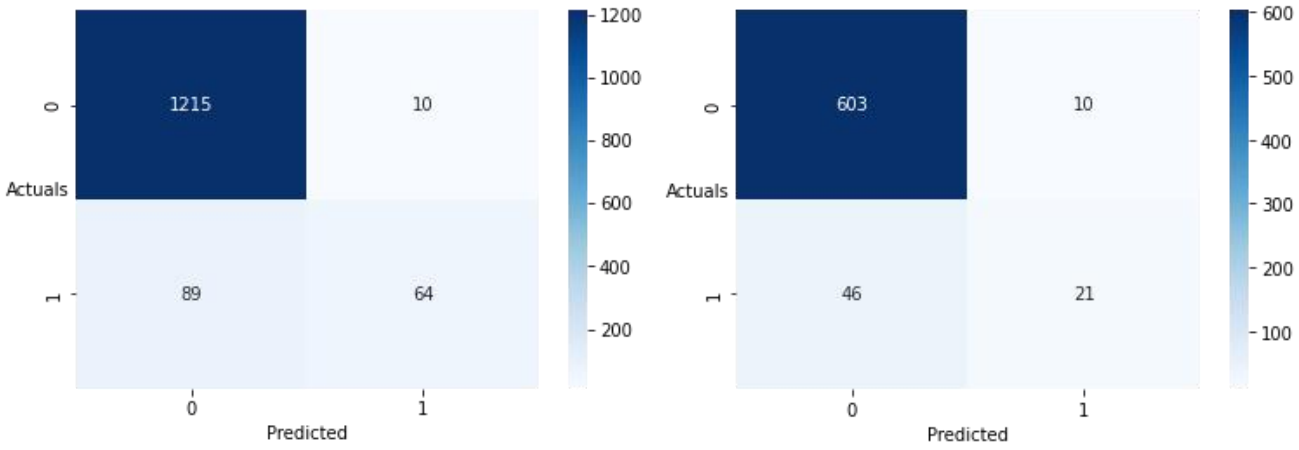


Fig.A.12. Confusion Matrix- Random Forest (Train and Test)

The recall for Random Forest remains the same . The model looks to be underfitting even after optimization.

PART A: Build a LDA Model on Train Dataset. Also showcase your model building approach.

For train data,

	precision	recall	f1-score	support
0.0	0.94	0.96	0.95	1225
1.0	0.61	0.52	0.56	153
accuracy			0.91	1378
macro avg	0.78	0.74	0.76	1378
weighted avg	0.90	0.91	0.91	1378

Precison = 61%
Recall =52%

For test data,

	precision	recall	f1-score	support
0.0	0.96	0.94	0.95	613
1.0	0.53	0.63	0.57	67
accuracy			0.91	680
macro avg	0.74	0.78	0.76	680
weighted avg	0.92	0.91	0.91	680

Precison = 53%
Recall =63%

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

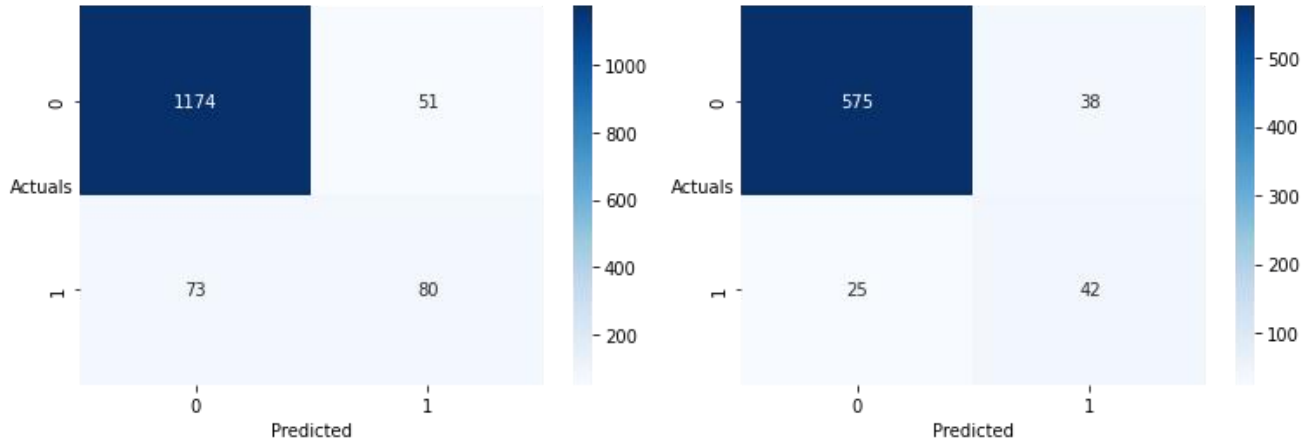


Fig.A.13. Confusion Matrix- LDA (Train and Test)

The recall for LDA model remains the same . The model looks to be underfitting even after optimization as there are no changes in the recall values.

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

For Logistic Regression Model,

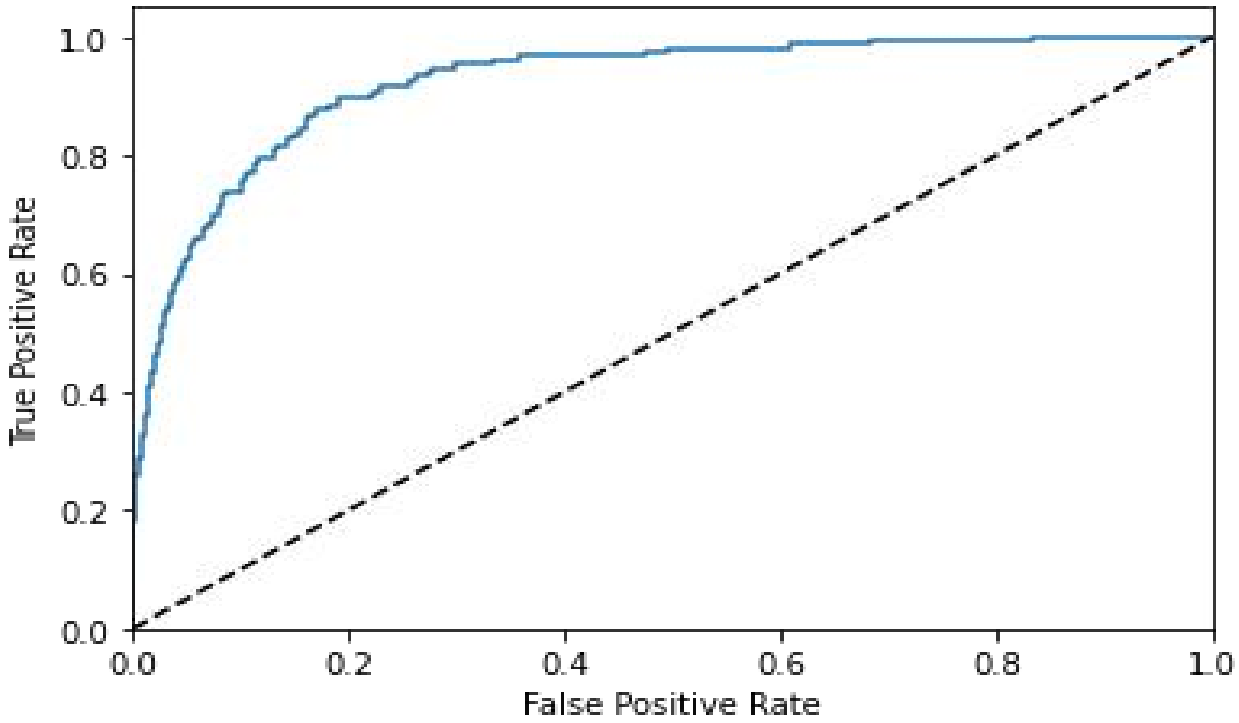


Fig.A.14. AUC-ROC curve- Logistic Regression

For Random Forest Model,

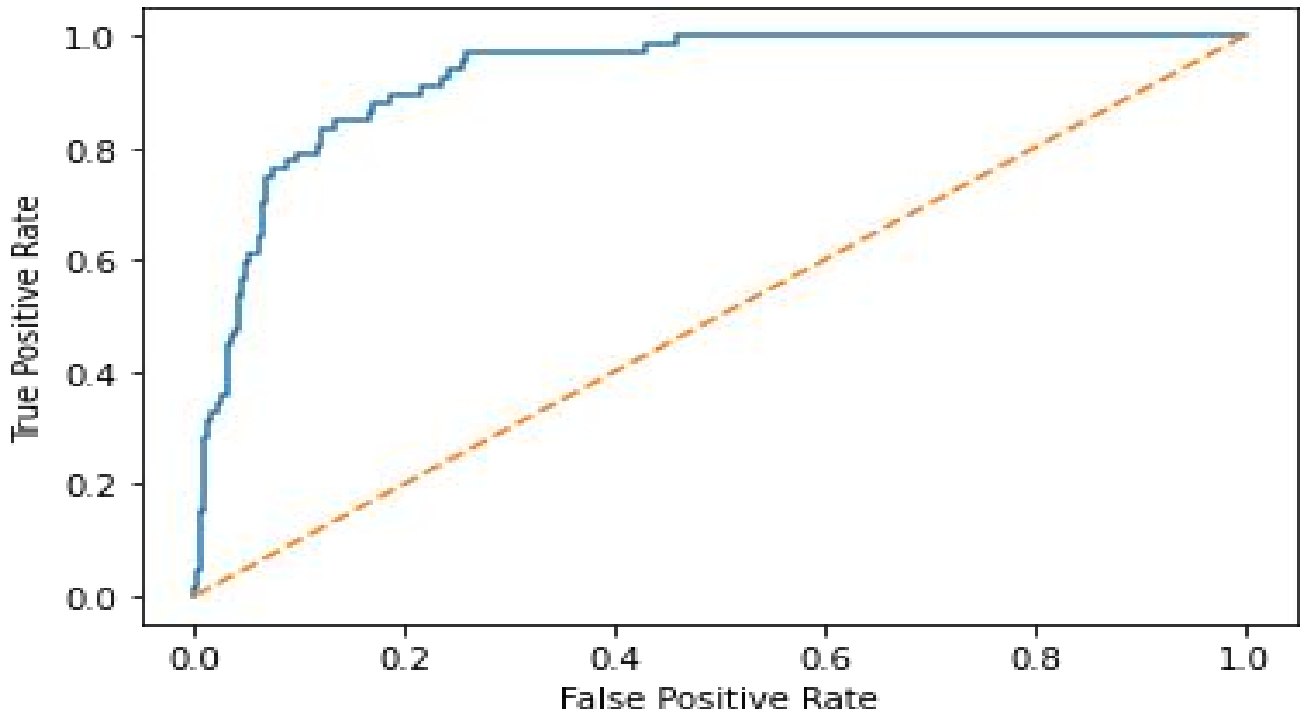


Fig.A.15. AUC-ROC curve- Random Forest

For LDA Model,

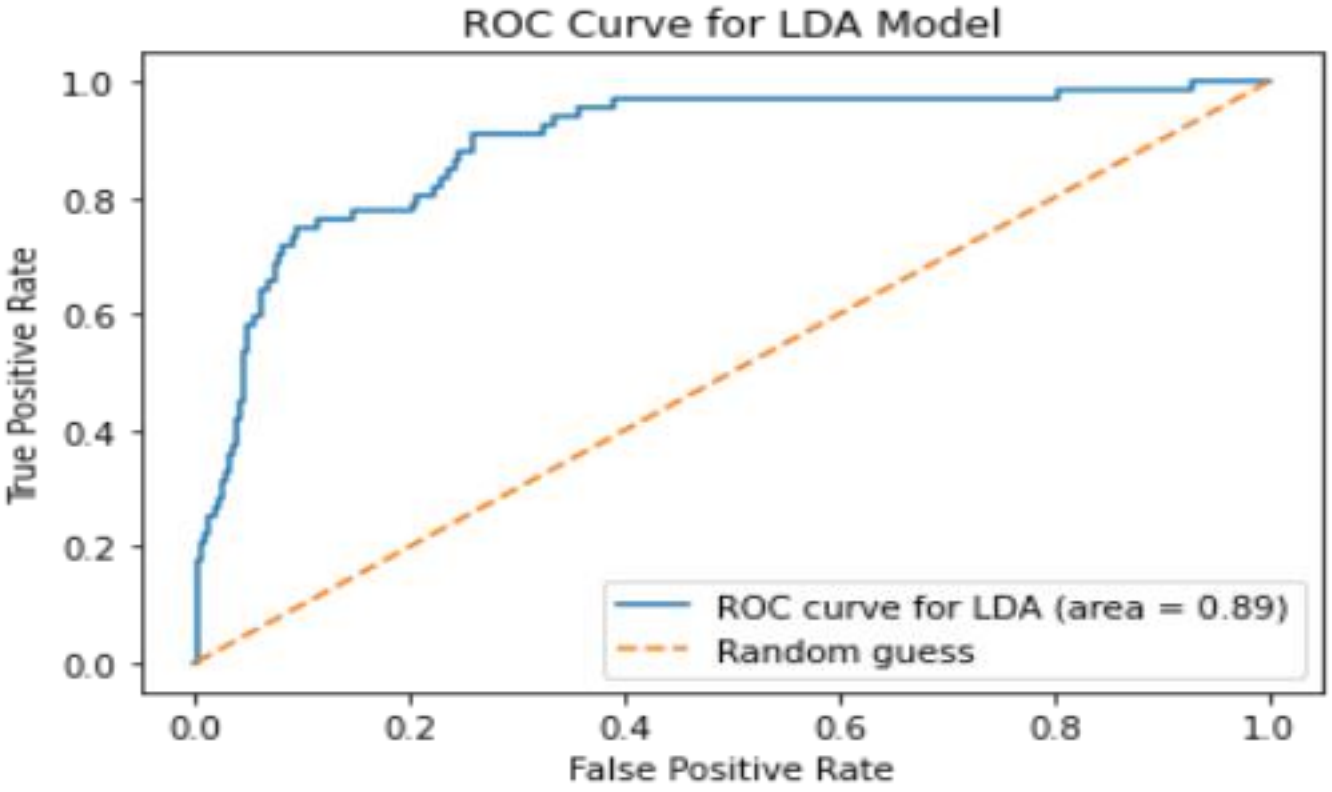


Fig.A.16. AUC-ROC curve- LDA

Model comparisons after threshold optimization:

Log-Reg Model		
	Precison	Recall
Train Data	39 %	87%
Test Data	34%	79 %
ROC score	0.90	
Random Forest Model		
Train Data	86%	42%
Test Data	68%	31%
ROC score	0.93	
LDA Model		
Train Data	61%	52%
Test Data	53%	63%
ROC score	0.89	

Table.A.8. Model Comparison

PART A: Conclusions and Recommendations.

- The best optimised model is Logistic regression model with highest recall value almost 80% in both train and test data. There is overfitting or underfitting in the model.
- The ROC curve plot of Log-Reg model has best curve
- The precison of RF model has the highest precision value of 86% in train data while LDA model test data has higher recall value. Both the model has underfitting, hence not recommended.
- For FRA, recall value is given priority compared to precision value. Hence, Log-Reg Model is recommended.
- The RF model has highest ROC score compared to Log-Reg and LDA model.

PART 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.

Dataset: Market Risk Dataset

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

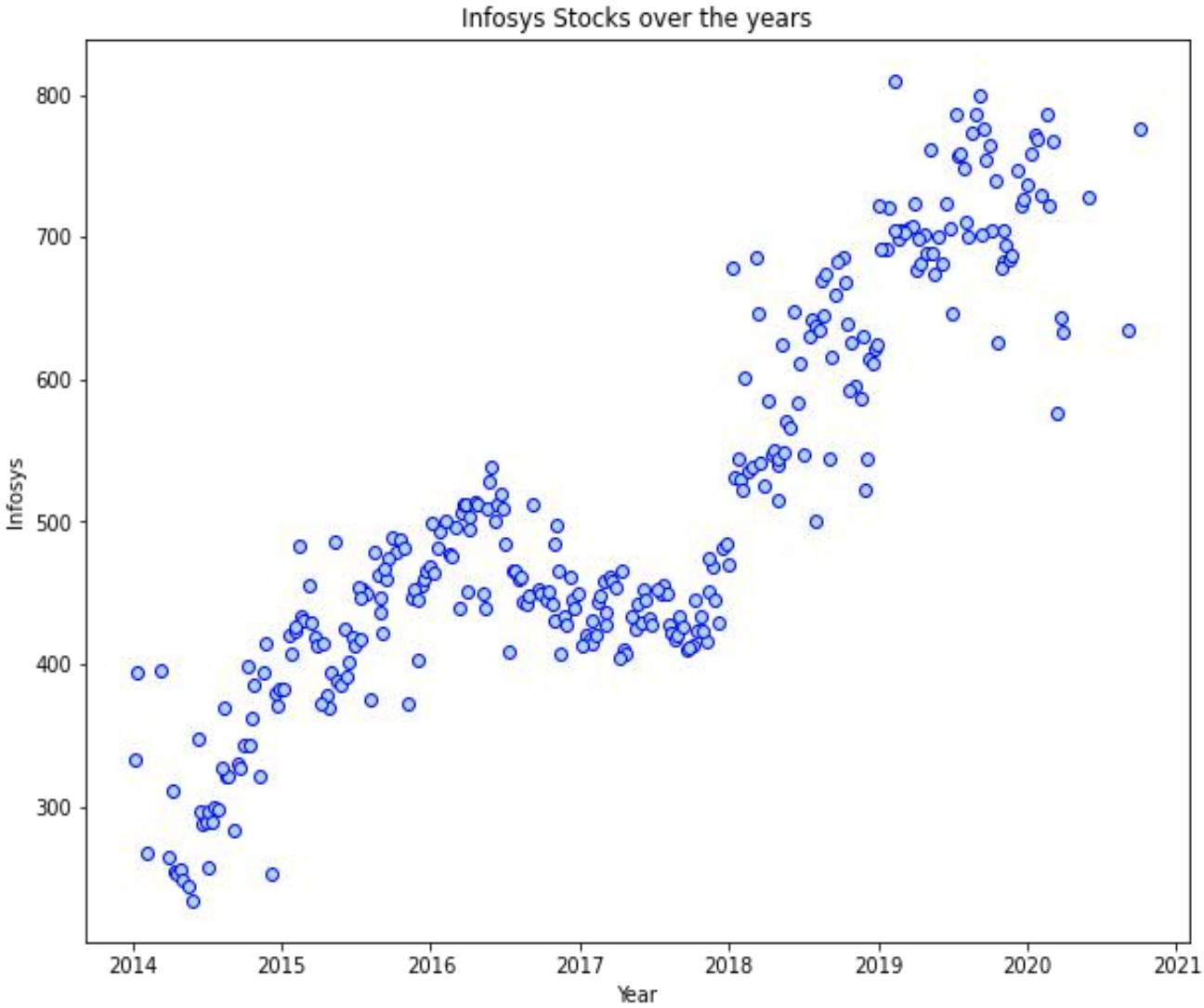


Fig.B.1. Scatter Plot for Infosys stocks- Stock Price vs Time

- The stock price of Infosys has improved over a period of 6 years with a slight decline during 2018.
- Currently the price trend seems to be growing towards 2021.

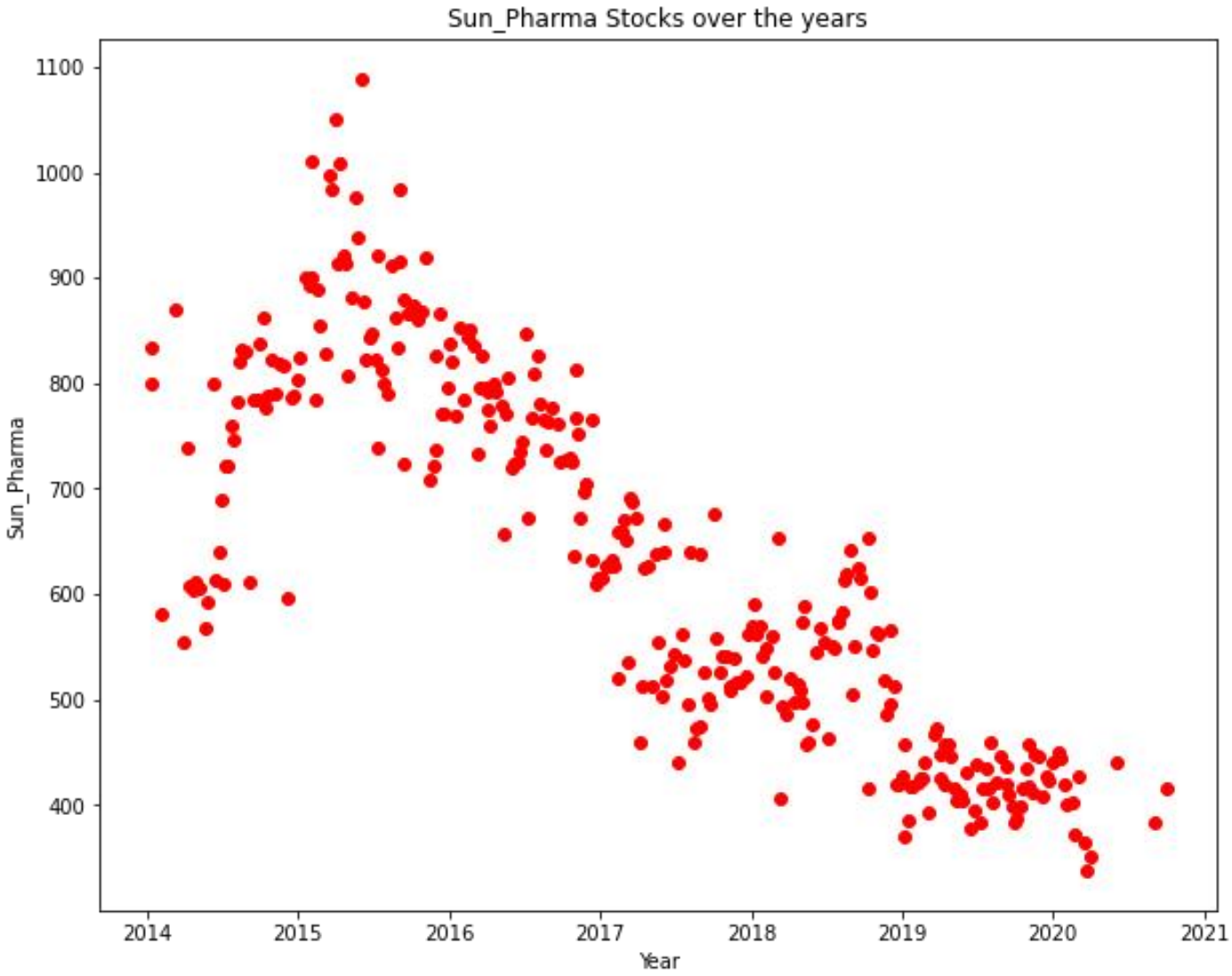


Fig.B.2. Scatter Plot for Sun Pharma stocks- Stock Price vs Time

- The stock price of Sun Pharma has improved over a period of 1 year with with decline after middle of the year 2015.
- Currently the price trend looks to be declining towards the year 2021.
- There was a slight increase in stock in the middle of the year 2018.

PART B: Calculate Returns for all stocks with inference.

Analyzing returns

Steps for calculating returns from prices:

- Take logarithms
- Take differences

Creating a data frame using the above method,

Sl. No.	Infosys	Indian Hotel	Mahindra & Mahindra	Axis Bank	SAIL	Shree Cement	Sun Pharma	Jindal Steel	Idea Vodafone	Jet Airways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.03	-0.01	0.01	0.05	0.03	0.03	0.09	-0.07	0.01	0.09
2	-0.01	0.00	-0.01	-0.02	-0.03	-0.01	-0.00	0.00	-0.01	-0.08
3	-0.00	0.00	0.07	0.05	0.00	0.01	-0.00	-0.02	0.00	0.01
4	0.01	-0.05	-0.01	-0.00	-0.08	-0.02	0.01	-0.14	-0.05	-0.15

Table.B.1. Data Head - stock_returns

PART B: Calculate Stock Means and Standard Deviation for all stocks with inference

Calculating Stock Means :

Infosys	0.00
Indian_Hotel	0.00
Mahindra_&_Mahindra	-0.00
Axis_Bank	0.00
SAIL	-0.00
Shree_Cement	0.00
Sun_Pharma	-0.00
Jindal_Steel	-0.00
Idea_Vodafone	-0.01
Jet_Airways	-0.01

Table.B.2. Stock Means - stock_returns

Stock Means: Average returns that the stock is making on a week to week basis.

- The above results show there are no average returns for the Stock Means for of the stocks.
- Although there are no stock returns Infosys, Indian Hotel, Axis Bank and Shree Cement stock has not declined.
- While the other stocks show negative value.

Calculating stock Standard Deviations :

Infosys	0.04
Indian_Hotel	0.05
Mahindra_&_Mahindra	0.04
Axis_Bank	0.05
SAIL	0.06
Shree_Cement	0.04
Sun_Pharma	0.05
Jindal_Steel	0.08
Idea_Vodafone	0.10
Jet_Airways	0.10

Table.B.3. Stock Standard Deviations - stock_returns

Stock Standard Deviation : It is the measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile stock will become.

- The above table shows that Jindal_Steel, Idea_Vodafone and Jet_airways are the most volatile stock compared to others.

Stocks	Average	Volatility
Infosys	0.00	0.04
Indian_Hotel	0.00	0.05
Mahindra_&_Mahindra	-0.00	0.04
Axis_Bank	0.00	0.05
SAIL	-0.00	0.06
Shree_Cement	0.00	0.04
Sun_Pharma	-0.00	0.05
Jindal_Steel	-0.00	0.08
Idea_Vodafone	-0.01	0.10
Jet_Airways	-0.01	0.10

Table.B.4. Stock Average and Volatility.

PART B: Draw a plot of Stock Means vs Standard Deviation and state your inference.

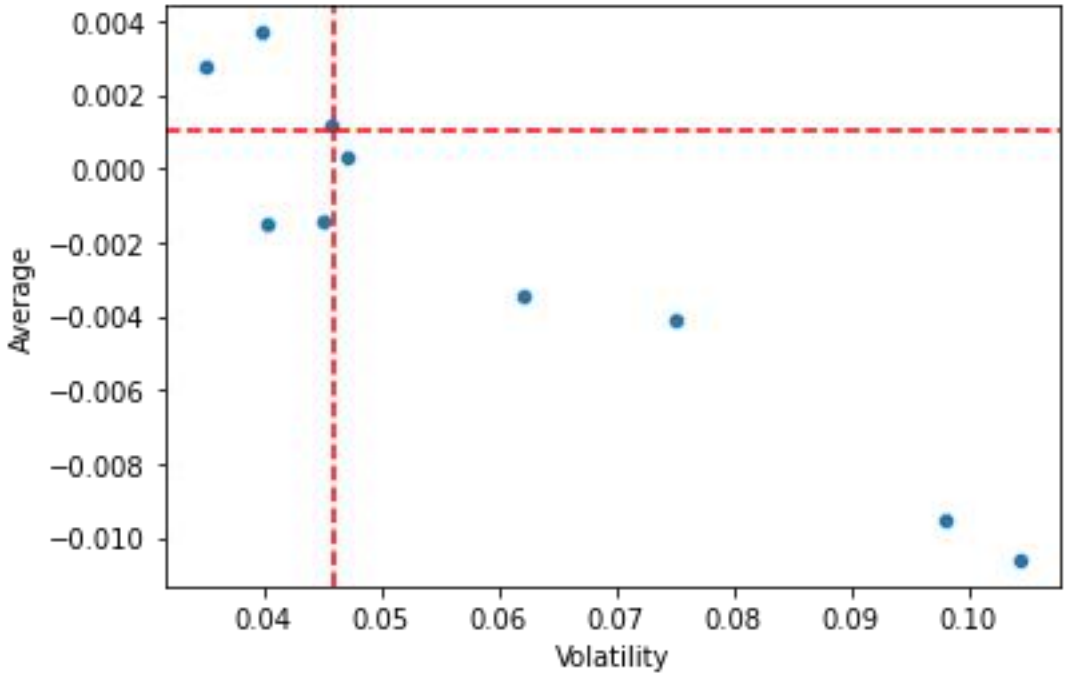


Fig.B.3. Plot for Stock Means vs Standard Deviation

PART B: Conclusions and Recommendations

➢