

Finance and Risk Analytics

Pentapalli Suneel Kumar | [Date]

Contents

List of Figures	2
List of Tables	3
PART A:	4
Data Shape and Dictionary:	4
Data Info:.....	8
Null Check:.....	9
Duplicate Check:	9
Data Description:	9
Outlier Check and Treatment:.....	13
Missing value Treatment:.....	17
Univariate & Bivariate analysis:.....	18
Train Test Split:.....	24
Logistic Regression(From Statsmodel):	25
(Logistic Regression)Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model:	26
Build a Random Forest Model on Train Dataset. Also showcase your model building approach:	27
Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model:	27
Build a LDA Model on Train Dataset. Also showcase your model building approach:	28
Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model:.....	28
Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve): ...	28
Conclusions and Recommendations:.....	31
PART B:.....	31
Problem Statement:.....	31
Exploratory Data Analysis:	32
Data Info:.....	32
Five Point summary:.....	32
Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference:	33
Calculate Returns for all stocks with inference:	34
Calculate Stock Means and Standard Deviation for all stocks with inference:	35
Draw a plot of Stock Means vs Standard Deviation and state your inference:.....	36
Conclusions and Recommendations:.....	36

List of Figures

Fig 1: Credit Risk Data Missing Values	9
Fig 2: Credit Risk Box Plots(Before Outlier Treatment)	14
Fig 3: Credit Risk Box Plots(After Outlier Treatment)	16
Fig 4: Credit Risk(Missing Values Before Treatment)	18
Fig 5: Credit Risk(Box Plots after Outlier and Missing Value Treatment)	18
Fig 6: Credit Risk(Histograms)	20
Fig 7: Credit Risk(Heatmap)	22
Fig 8: Scatter Plot 1	23
Fig 9: Scatter Plot 2	24
Fig 10: Scatter Plot 3	24
Fig 11: Train Test Split(Data Shape)	26
Fig 12: Logit Model Summary	26
Fig 13: Classification Report(Logit)	28
Fig 14: Classification Report(Random Forest)	28
Fig 15: Classification Report(LDA)	29
Fig 16: AUC Score Interpretation	30
Fig 17: ROC Curve Logit Model	30
Fig 18: ROC Curve Random Forest Model	31
Fig 19: ROC Curve LDA Model	31
Fig 20: Data head Stock data	33
Fig 21: Data Info Stock data	33
Fig 22: Data Description Stock data	34
Fig 23: Price vs Time Plot 1	34
Fig 24: Price vs Time Plot 2	35
Fig 25: Logarithmic Returns	36
Fig 26: Stock Means	36
Fig 27: Stock Standard Deviation	37
Fig 28: Stock Mean vs Standard Deviation Plot	37

Fig 29: Best Stocks Top to Bottom 39

List of Tables

Table 1: Credit Rist Data Dictionary 4

Table 2: Credit Rist 5 point summary 10

Table 3: Sum of nulls after Missing value treatment 18

PART 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 Shape and Dictionary:

Given Credit Risk data set has 2058 entries with total 58 features. Data dictionary for each of these features is as follows.

Table 1: Credit Risk 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.

Data Info:

All the features of of int/float data types. Default is out target feature with values 0 and 1. 1 being identified as defaulters and 0 being non defaulters.

Given data set has data related to 1838 non defaulters and 220 defaulters. i.e we only have 10% of records that give information about defaulters. Hence we call this imbalanced data set which needs to be handled using balancing techniques like SMOTE.

Null Check:

From the 58 features, below features have nulls/missing data.

Fig 1: Credit Risk Data Missing Values

Cash_Flow_Per_Share	167
Total_debt_to_Total_net_worth	21
Cash_to_Total_Assets	96
Current_Liability_to_Current_Assets	14

When we convert above null entries in each column to percentages, it will be too low, but still as part of data preparation, we will impute those nulls going further.

Duplicate Check:

No Duplicates were found in the dataset.

Data Description:

Five point summary of all the features is as follows,

From the below summary, it is evident that there are few features with outliers as the mean and median are not in sync.

Table 2: Credit Risk 5 point summary

	count	mean	std	min	25%	50%	75%	max
Co_Code	2058	1.76E+04	2.19E+04	4	3.67E+03	6.24E+03	2.43E+04	7.25E+04
Operating_Expense_Rate	2058	2.05E+09	3.25E+09	0	0.00E+00	0.00E+00	4.11E+09	9.98E+09
Research_and_development_expense_rate	2058	1.21E+09	2.14E+09	0	0.00E+00	0.00E+00	1.55E+09	9.98E+09
Cash_flow_rate	2058	4.70E-01	2.00E-02	0	4.60E-01	4.60E-01	4.70E-01	1.00E+00
Interest_bearing_debt_interest_rate	2058	1.11E+07	9.04E+07	0	0.00E+00	0.00E+00	0.00E+00	9.90E+08
Tax_rate_A	2058	1.10E-01	1.50E-01	0	0.00E+00	4.00E-02	2.20E-01	1.00E+00

Cash_Flow_Per_Share	1891	3.20E-01	2.00E-02	0.17	3.10E-01	3.20E-01	3.30E-01	4.60E-01
Per_Share_Net_profit_before_tax_Yuan_	2058	1.80E-01	3.00E-02	0	1.70E-01	1.80E-01	1.90E-01	7.90E-01
Realized_Sales_Gross_Profit_Growth_Rate	2058	2.00E-02	2.00E-02	0	2.00E-02	2.00E-02	2.00E-02	1.00E+00
Operating_Profit_Growth_Rate	2058	8.50E-01	0.00E+00	0.74	8.50E-01	8.50E-01	8.50E-01	1.00E+00
Continuous_Net_Profit_Growth_Rate	2058	2.20E-01	1.00E-02	0	2.20E-01	2.20E-01	2.20E-01	2.30E-01
Total_Asset_Growth_Rate	2058	5.29E+09	2.91E+09	0	4.32E+09	6.23E+09	7.22E+09	9.98E+09
Net_Value_Growth_Rate	2058	5.19E+06	2.08E+08	0	0.00E+00	0.00E+00	0.00E+00	9.33E+09
Total_Asset_Return_Growth_Rate_Ratio	2058	2.60E-01	0.00E+00	0.25	2.60E-01	2.60E-01	2.60E-01	3.60E-01
Cash_Reinvestment_perc	2058	3.80E-01	3.00E-02	0.03	3.70E-01	3.80E-01	3.90E-01	1.00E+00
Current_Ratio	2058	1.34E+06	6.06E+07	0	1.00E-02	1.00E-02	1.00E-02	2.75E+09
Quick_Ratio	2058	2.78E+07	4.45E+08	0	0.00E+00	1.00E-02	1.00E-02	9.23E+09
Interest_Expense_Ratio	2058	6.30E-01	1.00E-02	0.53	6.30E-01	6.30E-01	6.30E-01	8.10E-01
Total_debt_to_Total_net_worth	2037	1.07E+07	2.70E+08	0	0.00E+00	1.00E-02	1.00E-02	9.94E+09
Long_term_fund_suitability_ratio_A	2058	1.00E-02	3.00E-02	0	1.00E-02	1.00E-02	1.00E-02	1.00E+00
Net_profit_before_tax_to_Paid_in_capital	2058	1.80E-01	3.00E-02	0	1.70E-01	1.70E-01	1.80E-01	7.90E-01
Total_Asset_Turnover	2058	1.30E-01	1.00E-01	0	6.00E-02	1.00E-01	1.70E-01	9.20E-01

Accounts_Receivable_Turnover	2058	4.16E+07	5.05E+08	0	0.00E+00	0.00E+00	0.00E+00	9.74E+09
Average_Collection_Days	2058	2.63E+07	4.11E+08	0	0.00E+00	1.00E-02	1.00E-02	8.80E+09
Inventory_Turnover_Rate_times	2058	2.03E+09	3.08E+09	0	0.00E+00	1.91E+07	3.82E+09	9.99E+09
Fixed_Assets_Turnover_Frequency	2058	1.23E+09	2.65E+09	0	0.00E+00	0.00E+00	1.00E-02	9.99E+09
Net_Worth_Turnover_Rate_times	2058	4.00E-02	4.00E-02	0.01	2.00E-02	3.00E-02	4.00E-02	1.00E+00
Operating_profit_per_person	2058	4.00E-01	5.00E-02	0	3.90E-01	4.00E-01	4.00E-01	1.00E+00
Allocation_rate_per_person	2058	5.73E+06	1.98E+08	0	0.00E+00	1.00E-02	2.00E-02	8.28E+09
Quick_Assets_to_Total_Assets	2058	3.40E-01	2.10E-01	0	1.70E-01	3.10E-01	4.80E-01	9.90E-01
Cash_to_Total_Assets	1962	8.00E-02	1.00E-01	0	2.00E-02	5.00E-02	1.00E-01	9.30E-01
Quick_Assets_to_Current_Liability	2058	1.19E+07	3.12E+08	0	0.00E+00	1.00E-02	1.00E-02	8.82E+09
Cash_to_Current_Liability	2058	9.28E+07	7.85E+08	0	0.00E+00	0.00E+00	1.00E-02	9.17E+09
Operating_Funds_to_Liability	2058	3.50E-01	4.00E-02	0.03	3.40E-01	3.50E-01	3.50E-01	1.00E+00
Inventory_to_Working_Capital	2058	2.80E-01	2.00E-02	0	2.80E-01	2.80E-01	2.80E-01	1.00E+00
Inventory_to_Current_Liability	2058	5.79E+07	6.28E+08	0	0.00E+00	1.00E-02	1.00E-02	9.60E+09
Long_term_Liability_to_Current_Assets	2058	7.34E+07	6.69E+08	0	0.00E+00	0.00E+00	1.00E-02	9.31E+09
Retained_Earnings_to_Total_Assets	2058	9.30E-01	3.00E-02	0	9.30E-01	9.40E-01	9.40E-01	9.70E-01

Total_income_to_Total_expense	2058	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	1.00E-02
Total_expense_to_Assets	2058	3.00E-02	4.00E-02	0	1.00E-02	2.00E-02	4.00E-02	1.00E+00
Current_Asset_Turnover_Rate	2058	1.27E+09	2.84E+09	0	0.00E+00	0.00E+00	0.00E+00	9.99E+09
Quick_Asset_Turnover_Rate	2058	2.57E+09	3.45E+09	0	0.00E+00	0.00E+00	5.79E+09	1.00E+10
Cash_Turnover_Rate	2058	2.65E+09	2.82E+09	0	0.00E+00	1.73E+09	4.55E+09	9.99E+09
Fixed_Assets_to_Assets	2058	4.04E+06	1.83E+08	0	1.00E-01	2.10E-01	4.20E-01	8.32E+09
Cash_Flow_to_Total_Assets	2058	6.40E-01	5.00E-02	0	6.30E-01	6.40E-01	6.50E-01	1.00E+00
Cash_Flow_to_Liability	2058	4.60E-01	3.00E-02	0.03	4.60E-01	4.60E-01	4.60E-01	9.10E-01
CFO_to_Assets	2058	5.80E-01	6.00E-02	0	5.50E-01	5.80E-01	6.10E-01	9.80E-01
Cash_Flow_to_Equity	2058	3.10E-01	1.00E-02	0	3.10E-01	3.10E-01	3.20E-01	5.70E-01
Current_Liability_to_Current_Assets	2044	4.00E-02	5.00E-02	0	2.00E-02	3.00E-02	4.00E-02	1.00E+00
Liability_Assets_Flag	2058	0.00E+00	6.00E-02	0	0.00E+00	0.00E+00	0.00E+00	1.00E+00
Total_assets_to_GNP_price	2058	2.78E+07	4.72E+08	0	0.00E+00	0.00E+00	1.00E-02	9.82E+09
No_credit_Interval	2058	6.20E-01	1.00E-02	0.41	6.20E-01	6.20E-01	6.20E-01	9.60E-01
Degree_of_Financial_Leverage_DFL	2058	3.00E-02	1.00E-02	0.01	3.00E-02	3.00E-02	3.00E-02	4.60E-01
Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058	5.70E-01	1.00E-02	0.17	5.70E-01	5.70E-01	5.70E-01	6.70E-01

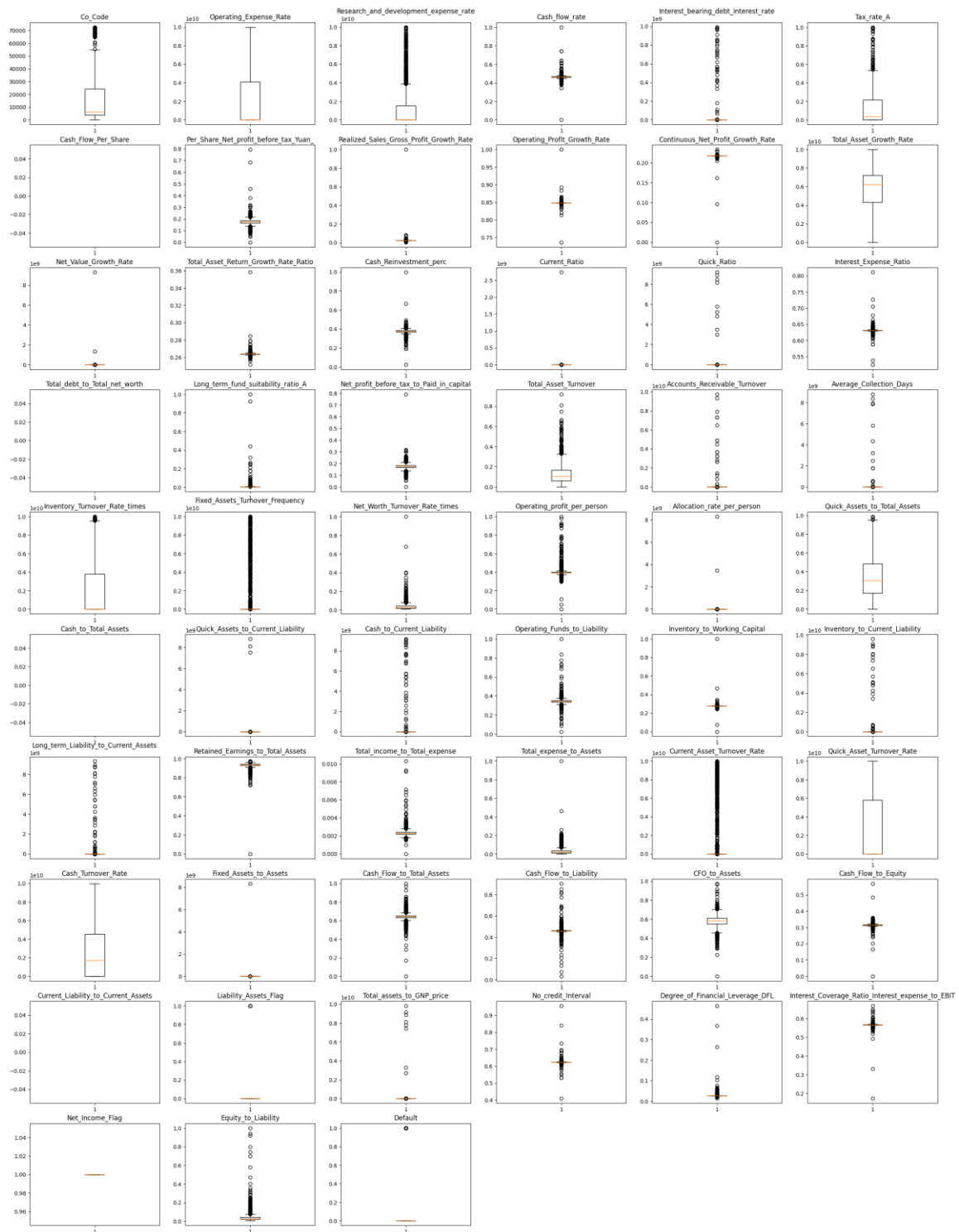
Net_Income_Flag	2058	1.00E+00	0.00E+00	1	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Equity_to_Liability	2058	4.00E-02	6.00E-02	0	2.00E-02	3.00E-02	4.00E-02	1.00E+00
Default	2058	1.10E-01	3.10E-01	0	0.00E+00	0.00E+00	0.00E+00	1.00E+00

Outlier Check and Treatment:

As seen below, most of the features have outliers in them which can be seen as dots.

Outlier treatment is necessary as they will impact our model performance and accuracy.

Fig 2: Credit Risk Box Plots(Before Outlier Treatment)



Here we have treated outliers using inter quartile method.

Inter Quantile Range(IQR)

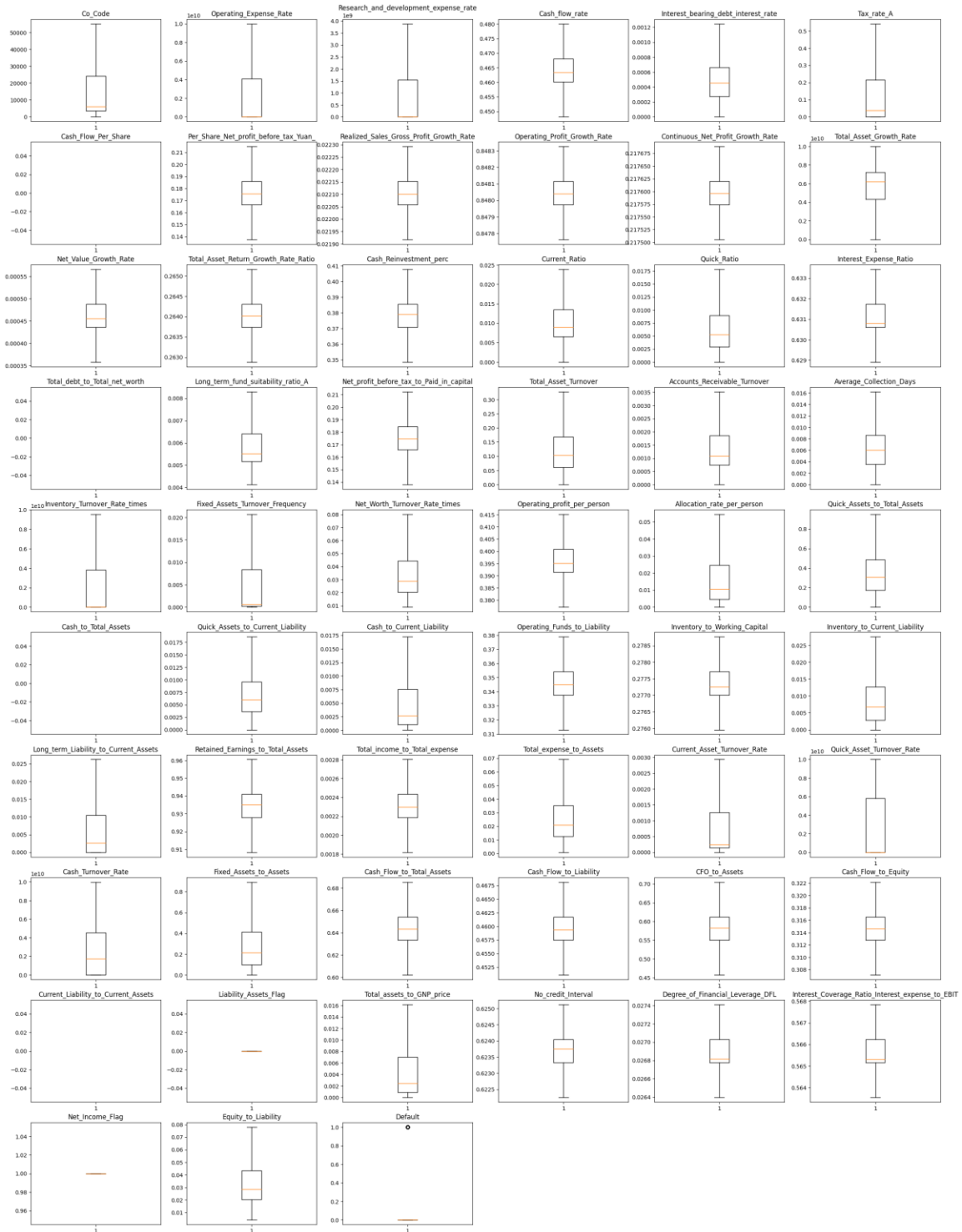
Criteria: data points that lie 1.5 times of IQR above Q_3 and below Q_1 are outliers. This shows in detail about outlier treatment in Python.

Steps:

- Sort the dataset in ascending order
- calculate the 1st and 3rd quartiles(Q_1 , Q_3)
- compute $IQR=Q_3-Q_1$
- compute lower bound = $(Q_1-1.5*IQR)$, upper bound = $(Q_3+1.5*IQR)$
- loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

Below is the box plot matrix for all the features after outlier treatment,

Fig 3: Credit Rist Box Plots(After Outlier Treatment)



Thus, we have removed in all the features except our target feature i.e., Default

Also we can see few features being shown as empty in above matrix, that is due to null values in those features. This will be fixed after removing null values.

Missing value Treatment:

We have imputed the nulls in below features with their respective means.

Fig 4: Credit Risk(Missing Values Before Treatment)

Cash_Flow_Per_Share	167
Total_debt_to_Total_net_worth	21
Cash_to_Total_Assets	96
Current_Liability_to_Current_Assets	14

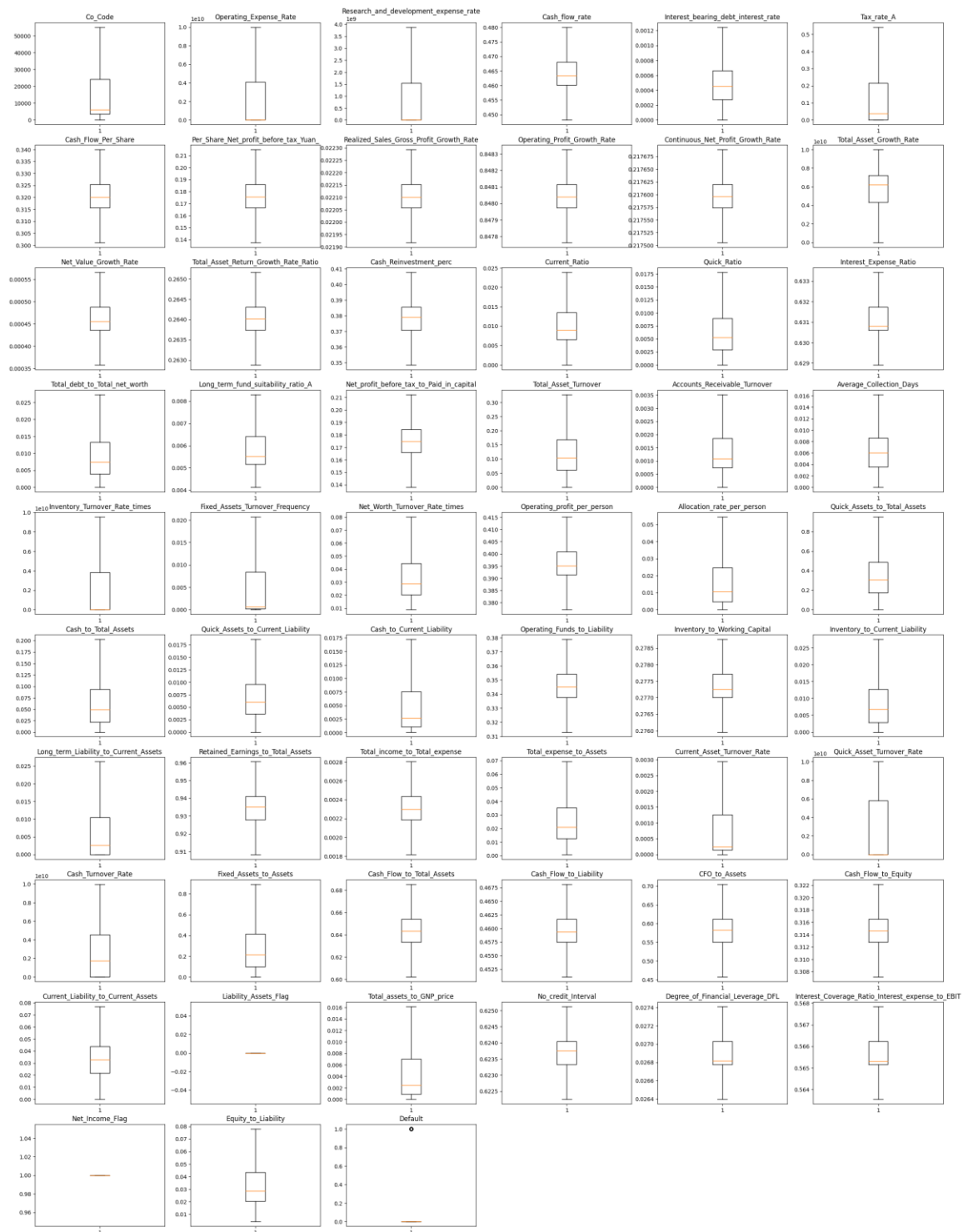
Here is the sum of nulls in above features after removing nulls.

Table 3: Sum of nulls after Missing value treatment

Cash_Flow_Per_Share	0
Total_debt_to_Total_net_worth	0
Cash_to_Total_Assets	0
Current_Liability_to_Current_Assets	0

Now as we have removed nulls, we can do outlier treatment on these columns, after treating outliers in these columns, box plots are as follows,

Fig 5: Credit Risk(Box Plots after Outlier and Missing Value Treatment)



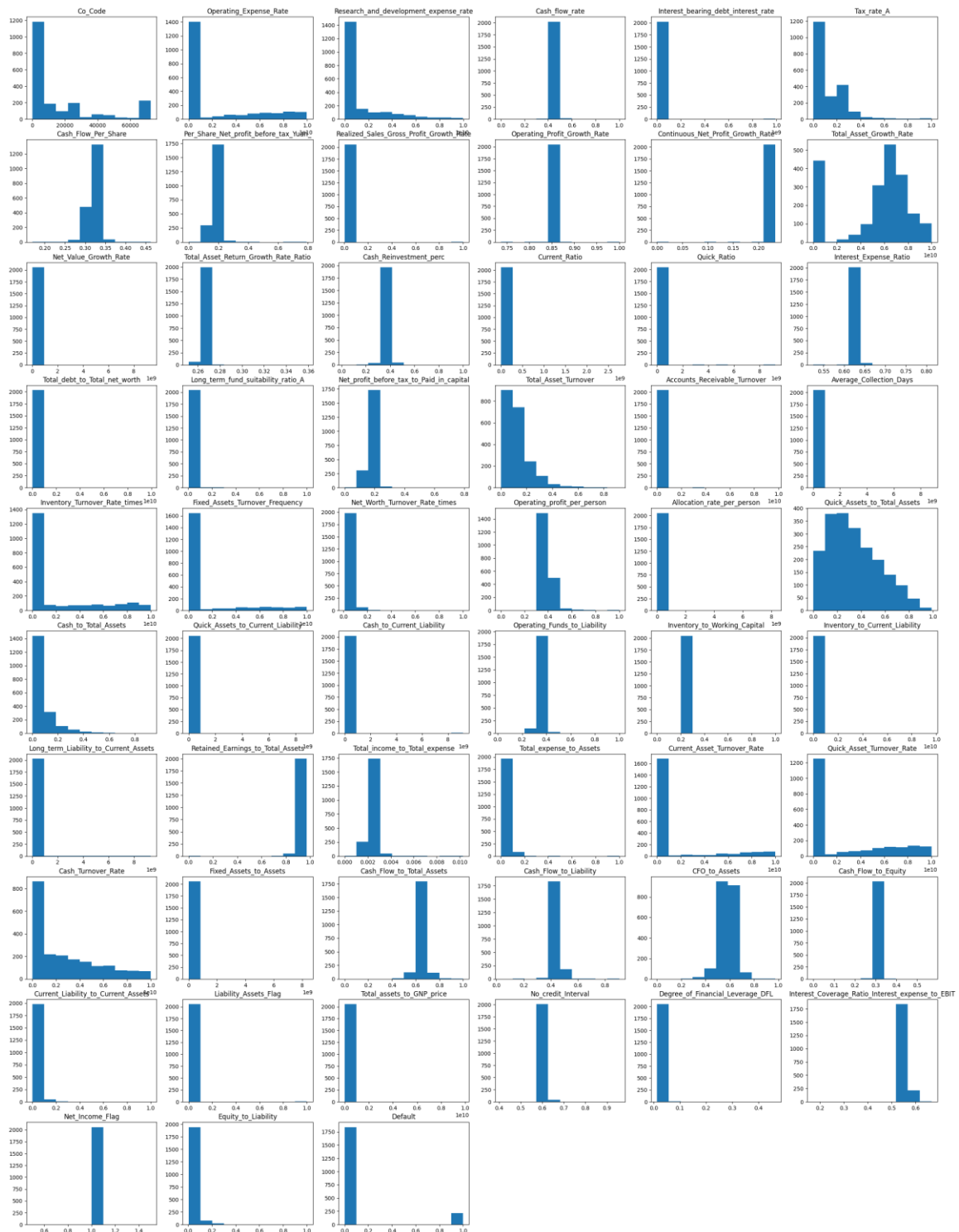
Univariate & Bivariate analysis:

Apart from box plots which we discussed above, we can also visualize distributions using Histplot.

Histograms of all the features is as follows,

We can see that none of the feature follow any distribution pattern. Most of them have the data skewed towards right, with exceptions in Total_Asset_Growth_Rate, Cash_Flow_Rate, Operating_Profit_Growth_Rate, Continuous_Net_Profit_Growth_Rate, Cash_Flow_To_Total_Assets, Cash_Flow_to_Equity and Net_Income_Flag.

Fig 6: Credit Risk(Histograms)



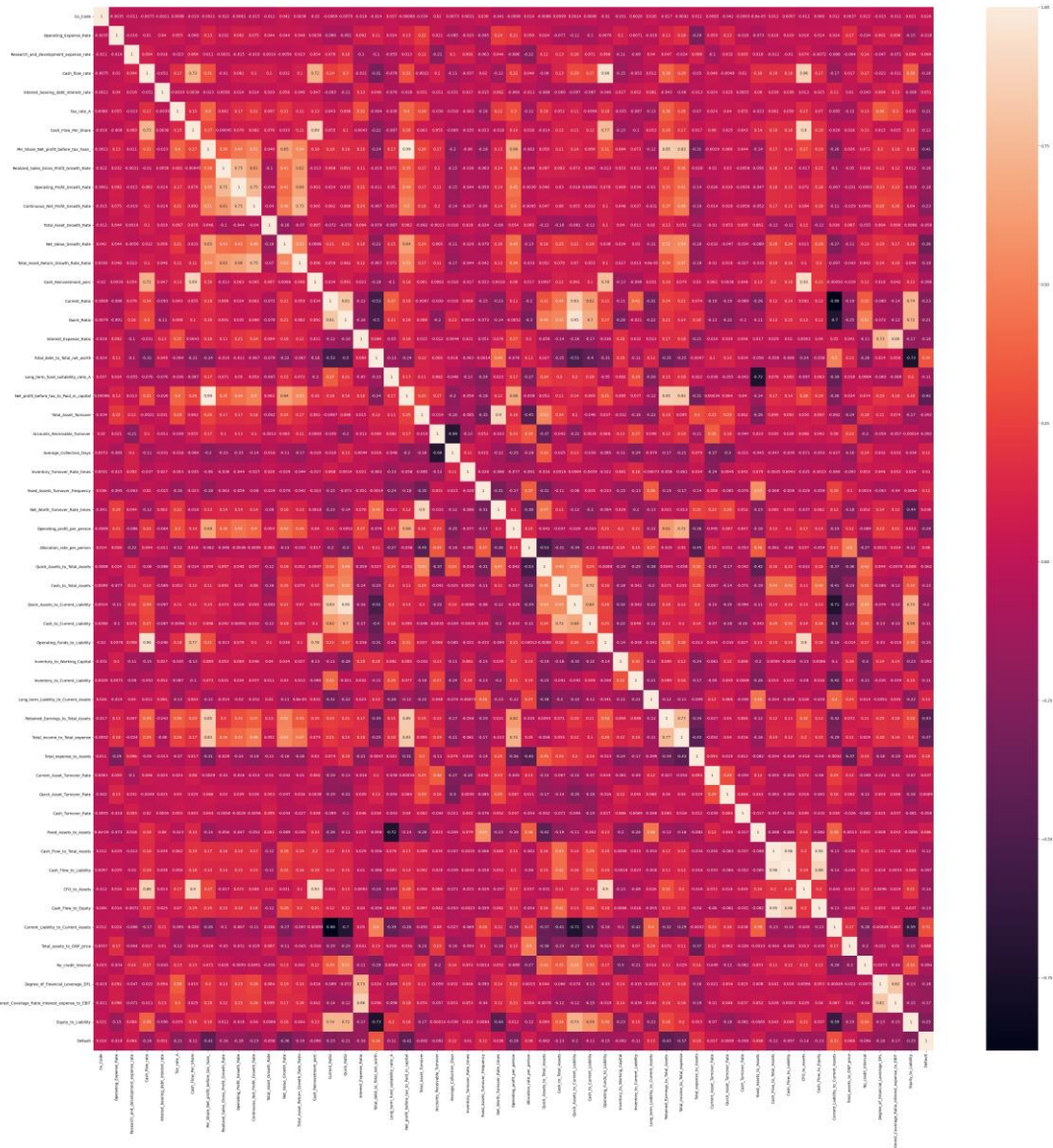
Heatmap for all the features is as follows,

Heatmap is nothing but the visual representation of correlation matrix.

The value of the correlation coefficient can take any values from -1 to 1.

If the value is 0, there is no correlation between the two variables. This means that the variables changes in a random manner with respect to each other.

Fig 7: Credit Risk(Heatmap)



In the above chart, the more darker the block is, the more negatively correlated are those respective features and vice-versa.

For instance, (Cash_Reinvestment_perc - Cash_Flow_Per_Share) and (Operating_Funds_to_Liability - Cash_flow_rate) have high correlation coefficients with values 0.89 and 0.96 respectively.

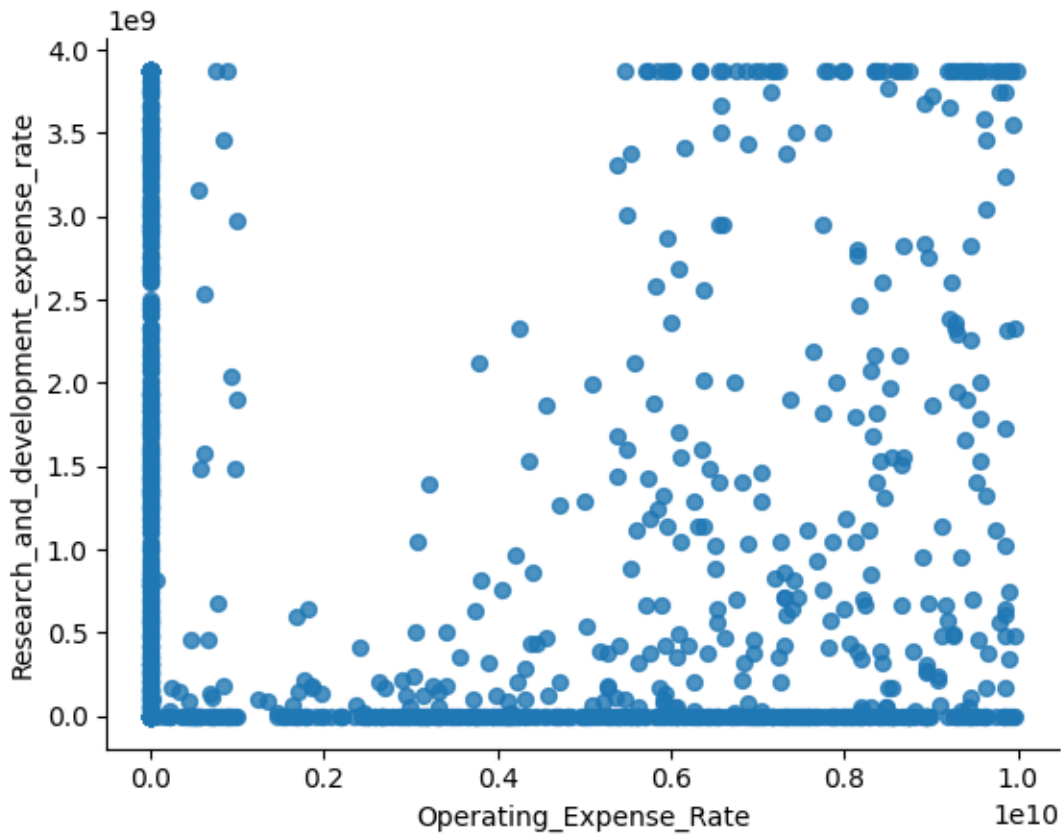
Similarly, $(\text{Current_Ratio} - \text{Current_Liability_to_Current_Assets})$ and

$(\text{Current_Liability_to_Current_Assets} - \text{Quick_Assets_to_Current_Liability})$ are highly negatively correlated with coefficients of -0.88 and -0.71 respectively.

Scatter Plot:

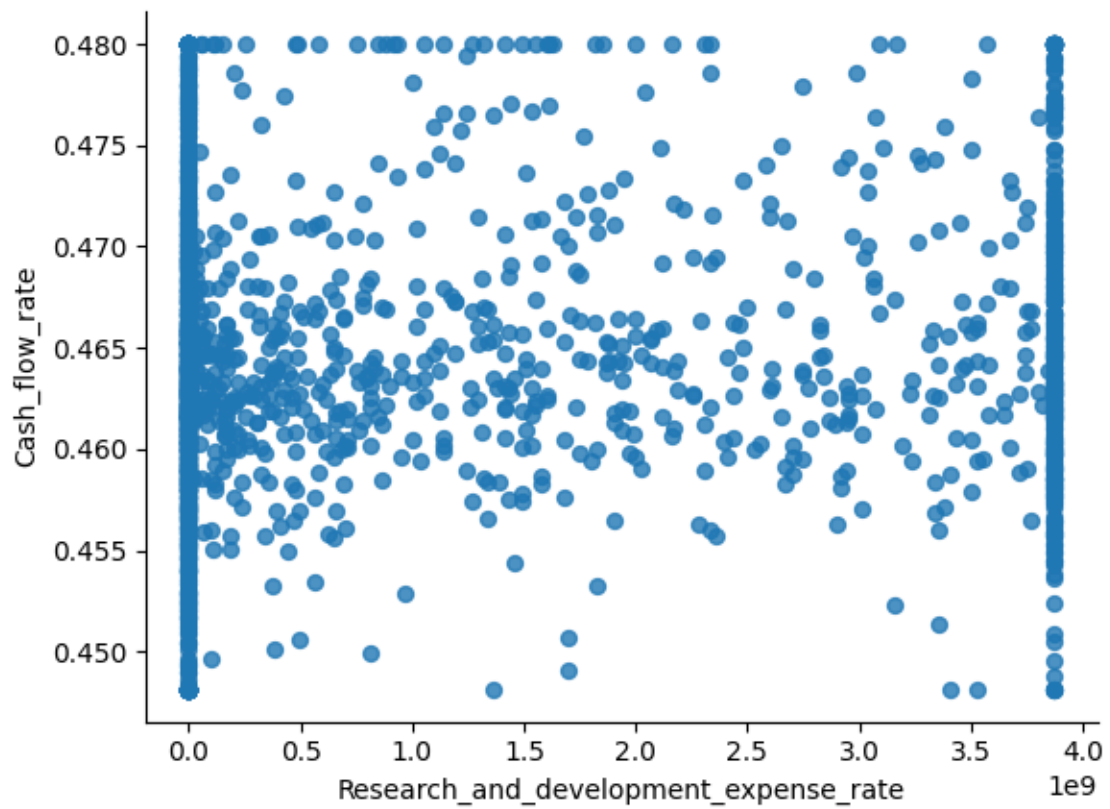
Scatter Plot between $\text{Research_and_development_expense_rate}$ and $\text{Operating_Expense_Rate}$ shows more values being concentrated near 0 for one feature and values being spread across in other feature.

Fig 8: Scatter Plot 1



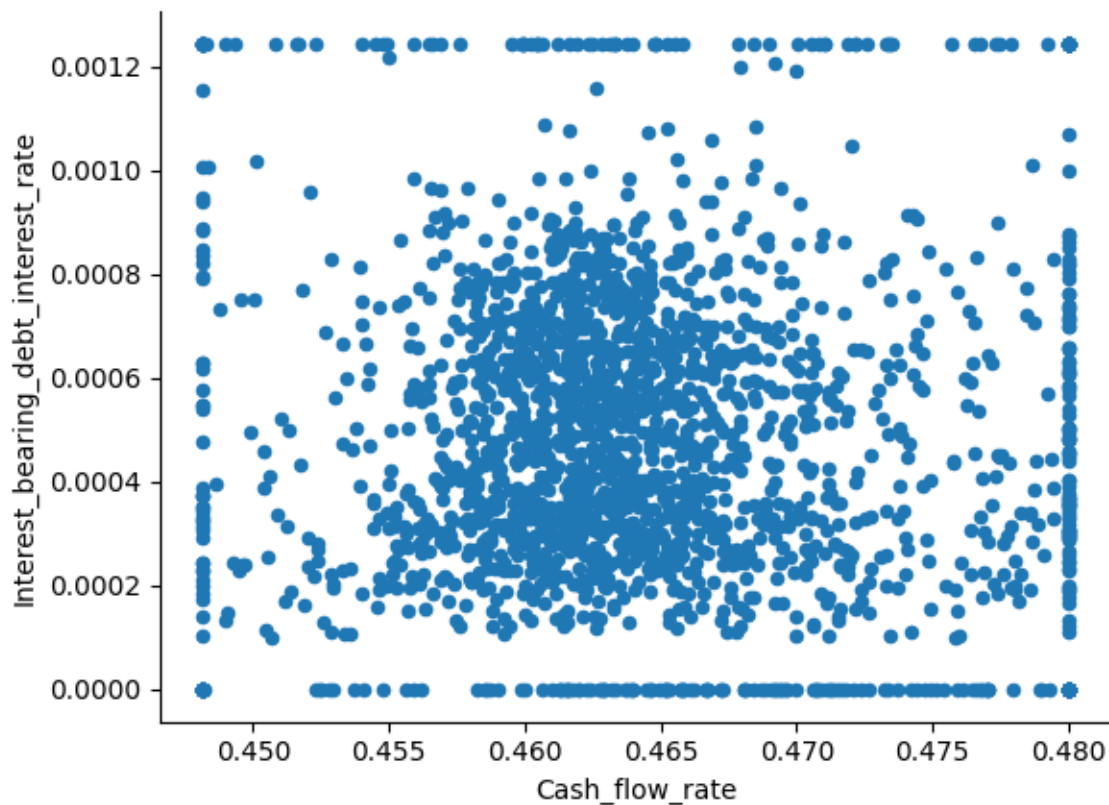
Similarly, scatter plot between Cash_flow_rate and $\text{Research_and_development_expense_rate}$ shows Cash_flow_rate is slightly high near lower values of $\text{Research_and_development_expense_rate}$.

Fig 9: Scatter Plot 2



Scatter plot between Interest_bearing_debt_interest_rate and Cash_flow_rate shows more values are being concentrated towards center.

Fig 10: Scatter Plot 3



Train Test Split:

As we have total 2058 rows with 58 features, now we divide features into independent and dependent sets.

We have split original dataset into X(independent variables) and Y(dependent variables).

FYI, we have removed column “Co_Name” from the list as it doesn’t add any value to the prediction.

Now “X” will have all features except “Co_Name” and “Default”.

“Y” contains only “Default”

After splitting data to independent and dependent sets, now we split data into train and test sets in 67:33 ratio.

Also, as already mentioned, the target variable data is not balanced as it has only 10% of data related to defaulters,

Hence we apply SMOTE oversampling technique here to balance data,

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Below are the counts of train and test data sets after applying SMOTE oversampling

Fig 11: Train Test Split(Data Shape)

```
Shape of training predictor dataset(x_train): (2462, 56)
Shape of test predictor dataset(x_test): (1214, 56)
Shape of training target dataset(y_train): (2462,)
Shape of test target dataset(y_test): (1214,)
```

Logistic Regression(From Statsmodel):

Now that we have 56 independent variable, all of them may not be useful when it comes to model building and prediction. Hence we have used RFE(Recursive Feature Elimination) technique for our feature selection.

Feature selection refers to techniques that select a subset of the most relevant features (columns) for a dataset. Fewer features can allow machine learning algorithms to run more efficiently (less space or time complexity) and be more effective. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance.

After this method, we have come up with 20 best features, which are as follows

```
['Co_Code', 'Operating_Expense_Rate', 'Research_and_development_expense_rate', 'Cash_flow_rate',
'Tax_rate_A', 'Operating_Profit_Growth_Rate', 'Total_Asset_Growth_Rate', 'Cash_Reinvestment_perc',
'Interest_Expense_Ratio', 'Inventory_Turnover_Rate_times', 'Quick_Assets_to_Total_Assets',
'Retained_Earnings_to_Total_Assets', 'Quick_Asset_Turnover_Rate', 'Cash_Turnover_Rate',
'Fixed_Assets_to_Assets', 'Cash_Flow_to_Total_Assets', 'Cash_Flow_to_Liability', 'No_credit_Interval',
'Interest_Coverage_Ratio_Interest_expense_to_EBIT', 'Net_Income_Flag']
```

Upon fitting the model on train data using above features, these are the coefficients we get from Logit Model,

Fig 12: Logit Model Summary

Logit Regression Results						
=====						
Dep. Variable:	Default	No. Observations:	2462			
Model:	Logit	Df Residuals:	2441			
Method:	MLE	Df Model:	20			
Date:	Sun, 07 Jan 2024	Pseudo R-squ.:	0.4188			
Time:	07:23:27	Log-Likelihood:	-991.86			
converged:	True	LL-Null:	-1706.5			
Covariance Type:	nonrobust	LLR p-value:	5.678e-291			
=====						
	coef	std err	z	P> z	[0.025	0.975]
Intercept	789.9144	1.91e+11	4.13e-09	1.000	-3.75e+11	3.75e+11
Co_Code	7.817e-06	3.39e-06	2.308	0.021	1.18e-06	1.45e-05
Operating_Expense_Rate	7.296e-11	1.8e-11	4.054	0.000	3.77e-11	1.08e-10
Research_and_development_expense_rate	2.591e-10	4.29e-11	6.045	0.000	1.75e-10	3.43e-10
Cash_flow_rate	-46.6140	15.781	-2.954	0.003	-77.545	-15.683
Tax_rate_A	-0.8162	0.567	-1.438	0.150	-1.928	0.296
Operating_Profit_Growth_Rate	-1568.3665	58.169	-26.962	0.000	-1682.376	-1454.357
Total_Asset_Growth_Rate	1.207e-11	2.28e-11	0.531	0.596	-3.25e-11	5.67e-11
Cash_Reinvestment_perc	7.1796	6.465	1.110	0.267	-5.492	19.851
Interest_Expense_Ratio	-3.3149	86.836	-0.038	0.970	-173.509	166.880
Inventory_Turnover_Rate_times	-3.616e-11	1.88e-11	-1.920	0.055	-7.31e-11	7.47e-13
Quick_Assets_to_Total_Assets	-1.6412	0.422	-3.886	0.000	-2.469	-0.813
Retained_Earnings_to_Total_Assets	-134.1512	6.937	-19.339	0.000	-147.747	-120.555
Quick_Asset_Turnover_Rate	-1.725e-11	1.73e-11	-0.997	0.319	-5.11e-11	1.66e-11
Cash_Turnover_Rate	-1.083e-10	2.33e-11	-4.653	0.000	-1.54e-10	-6.27e-11
Fixed_Assets_to_Assets	0.7209	0.338	2.134	0.033	0.059	1.383
Cash_Flow_to_Total_Assets	-32.2611	11.967	-2.696	0.007	-55.716	-8.807
Cash_Flow_to_Liability	129.6713	60.340	2.149	0.032	11.408	247.935
No_credit_Interval	-272.2113	83.078	-3.277	0.001	-435.042	-109.380
Interest_Coverage_Ratio_Interest_expense_to_EBIT	46.6122	81.807	0.570	0.569	-113.727	206.951
Net_Income_Flag	789.9149	1.91e+11	4.13e-09	1.000	-3.75e+11	3.75e+11
=====						

Interpretations:

Pseudo R-squ. : a substitute for the R-squared value in Least Squares linear regression. It is the ratio of the log-likelihood of the null model to that of the full model. This value ranges from 0 to 1. 0 being the poor model and 1 being the best model. In our case, we got this value around 0.41 which is pretty decent.

LLR p-value: Null hypothesis for this model says that the model is no significant in making predictions and alternate hypotheses says the model is significant and efficient in predicting our dependent variable. If p value is less than 0.05 then we can say this model is significant and useful. In our case its less than significant value 0.05 and hence we can reject null hypothesis.

(Logistic Regression)Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model:

Below is the Classification report of Logistic Regression model on test data,

1. Precision: Percentage of correct positive predictions relative to total positive predictions.
2. Recall: Percentage of correct positive predictions relative to total actual positives.
3. F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

F1 Score: $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

Using these three metrics, we can understand how well a given classification model is able to predict the outcomes for some response variable.

Fig 13: Classification Report(Logit)

	precision	recall	f1-score	support
0	0.82	0.85	0.84	607
1	0.84	0.82	0.83	607
accuracy			0.83	1214
macro avg	0.83	0.83	0.83	1214
weighted avg	0.83	0.83	0.83	1214

In this particular case we are more interested in predicting defaulters i.e., 1's. Hence recall would be correct parameter to rely on.

Hence our Logistic Regression model can predict 83% of total defaulters.

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

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the `max_samples` parameter if `bootstrap=True` (default), otherwise the whole dataset is used to build each tree.

We can use `GridSearchCV` algorithm to find the best possible parameters to be used for our train data in Random Forest model.

Upon Passing Random forest model to `GridSearchCV` algorithm with our train data, below are the best parameters suggested, Hence the model has been built on the same hyper parameters.

```
{'max_depth': 7, 'min_samples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 50}
```

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

Classification report for our test data using Random Forest Model is as follows,

We can see recall value has been improved a lot when compared to Logistic regression model. Recall in this case can be interpreted as, Our model can predict 93% of total defaulters.

Fig 14: Classification Report(Random Forest)

	precision	recall	f1-score	support
0	0.93	0.91	0.92	607
1	0.91	0.93	0.92	607
accuracy			0.92	1214
macro avg	0.92	0.92	0.92	1214
weighted avg	0.92	0.92	0.92	1214

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

Linear Discriminant Analysis (LDA) is a supervised learning algorithm used for classification tasks in machine learning. It is a technique used to find a linear combination of features that best separates the classes in a dataset.

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

Classification report of LDA model on test data is as follows, here based on Recall value, this model can predict 89% of our defaulters. Hence we can say this models predictability is between Logistic Regression and Random Forest Model.

Fig 15: Classification Report(LDA)

	precision	recall	f1-score	support
0	0.89	0.87	0.88	607
1	0.87	0.89	0.88	607
accuracy			0.88	1214
macro avg	0.88	0.88	0.88	1214
weighted avg	0.88	0.88	0.88	1214

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

Now that we have built classification models using Logistic Regression, Random Forest and Linear Discriminant Analysis, let compare performance of each model using AUC Score and ROC Curve.

AUC is a common abbreviation for Area Under the Receiver Operating Characteristic Curve (ROC AUC). It's a metric used to assess the performance of classification machine learning models.

The ROC is a graph which maps the relationship between true positive rate (TPR) and the false positive rate (FPR), showing the TPR that we can expect to receive for a given trade-off with FPR. The AUC score is the area under this ROC curve, meaning that the resulting score represents in broad terms the model's ability to predict classes correctly.

AUC score is interpreted as the probability that the model will assign a larger probability to a random positive observation than a random negative observation. More simplistically, AUC score can be interpreted as the model's ability to accurately classify classes on a scale from 0 to 1, where 1 is best and 0.5 is as good as random choice.

Fig 16: AUC Score Interpretation

AUC score	Interpretation
>0.8	Very good performance
0.7-0.8	Good performance
0.5-0.7	OK performance
0.5	As good as random choice

So lets see the AUC Score and ROC Curves of all our models on train and test data sets,

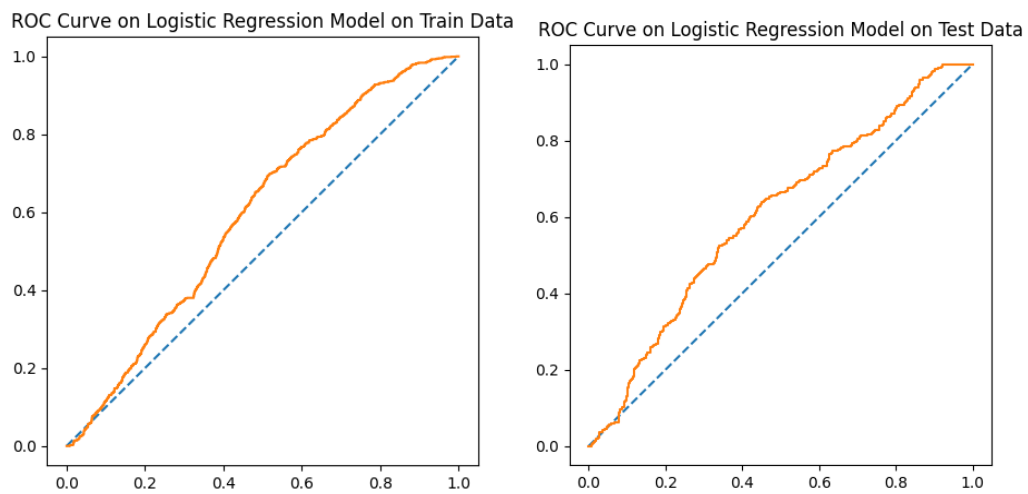
Logistic Regression:

AUC Score on Train data – 0.596

AUC Score on Test data – 0.6

ROC Curve –

Fig 17: ROC Curve Logit Model



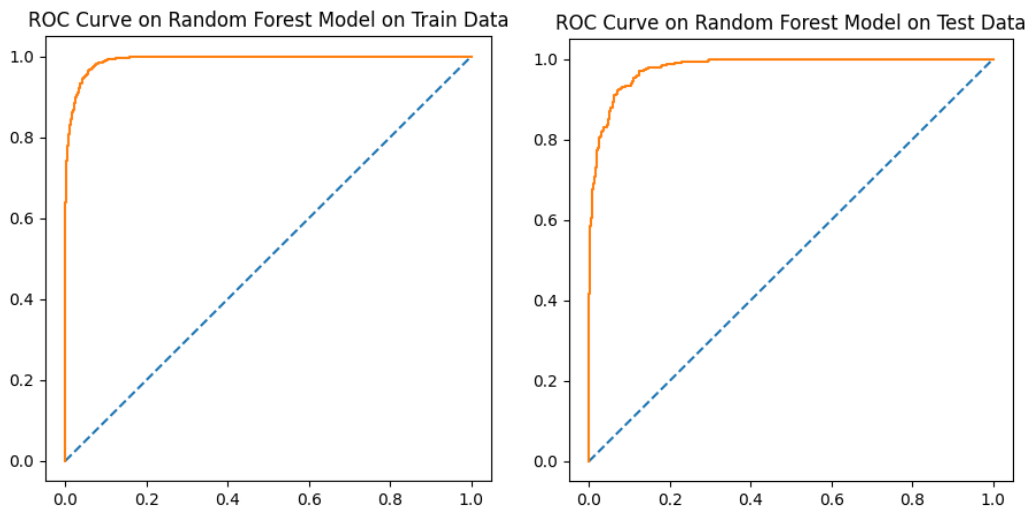
Random Forest:

AUC Score on Train data – 0.992

AUC Score on Test data – 0.98

ROC Curve –

Fig 18: ROC Curve Random Forest Model



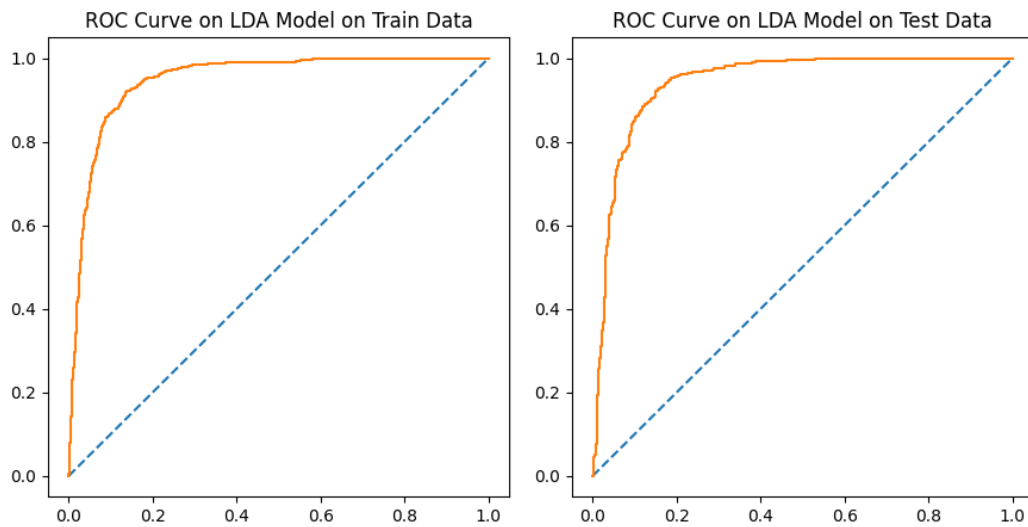
Linear Discriminant Analysis(LDA):

AUC Score on Train data – 0.948

AUC Score on Test data – 0.943

ROC Curve –

Fig 19: ROC Curve LDA Model



Conclusions and Recommendations:

Based on prediction capabilities of all the above models on our data, random forest clearly is the best model to use which can predict 93% of total defaulters.

It is recommended that the companies that are predicted as defaulters should be given credit only after assessing company's future plans and also we might consider different interest rates for these companies.

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.

Exploratory Data Analysis:

Sample Data:

Fig 20: Data head Stock data

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

Data Info:

Given dataset has total 314 entries with no duplicates and 11 columns or features. Non of the feature has null or missing values.

Fig 21: Data Info Stock data

```
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   IndianHotel         314 non-null   int64
 3   Mahindra&Mahindra   314 non-null   int64
 4   AxisBank            314 non-null   int64
 5   SAIL                314 non-null   int64
 6   ShreeCement         314 non-null   int64
 7   SunPharma           314 non-null   int64
 8   JindalSteel         314 non-null   int64
 9   IdeaVodafone        314 non-null   int64
10   JetAirways          314 non-null   int64
dtypes: int64(10), object(1)
```

Five Point summary:

Given data is related to stock prices of 10 companies across 6 years with weekly frequency. There is no much skewness in the data as mean and median values are almost close. ShreeCement is the company with highest average stock price of 14806.41 and IdeaVodafone is the company with least average price of around 53.71

Fig 22: Data Description Stock data

	count	mean	std	min	25%	50%	75%	max
Infosys	314.0	511.34	135.95	234.0	424.00	466.5	630.75	810.0
IndianHotel	314.0	114.56	22.51	64.0	96.00	115.0	134.00	157.0
Mahindra&Mahindra	314.0	636.68	102.88	284.0	572.00	625.0	678.00	956.0
AxisBank	314.0	540.74	115.84	263.0	470.50	528.0	605.25	808.0
SAIL	314.0	59.10	15.81	21.0	47.00	57.0	71.75	104.0
ShreeCement	314.0	14806.41	4288.28	5543.0	10952.25	16018.5	17773.25	24806.0
SunPharma	314.0	633.47	171.86	338.0	478.50	614.0	785.00	1089.0
JindalSteel	314.0	147.63	65.88	53.0	88.25	142.5	182.75	338.0
IdeaVodafone	314.0	53.71	31.25	3.0	25.25	53.0	82.00	117.0
JetAirways	314.0	372.66	202.26	14.0	243.25	376.0	534.00	871.0

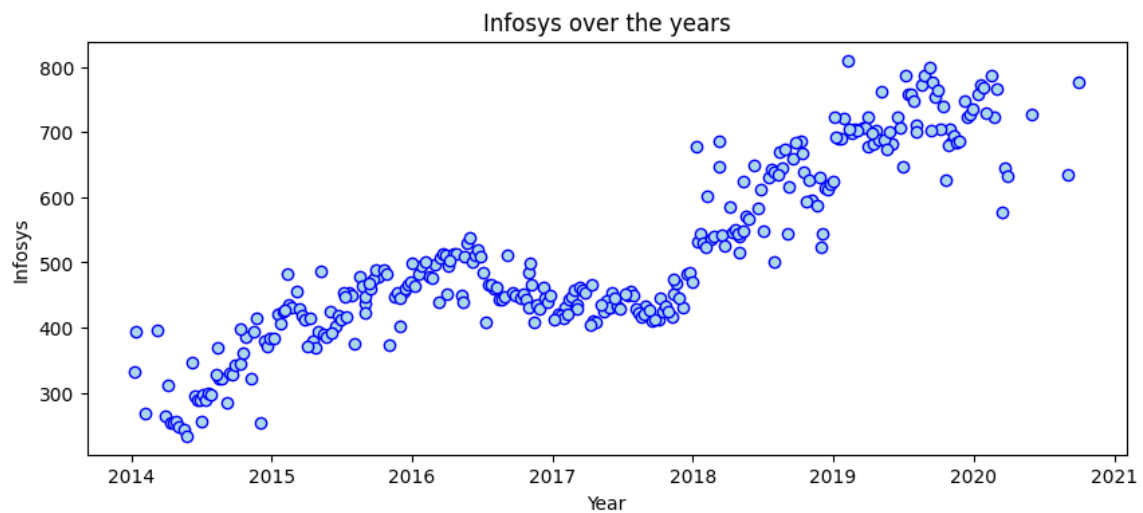
Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference:

Lets plot Stock Price vs Time graphs for Infosys and IdeaVodafone,

Infosys(Stock Price vs Time):

We can see the stock price has seen a decent upward trend across years and this stock seems pretty bullish.

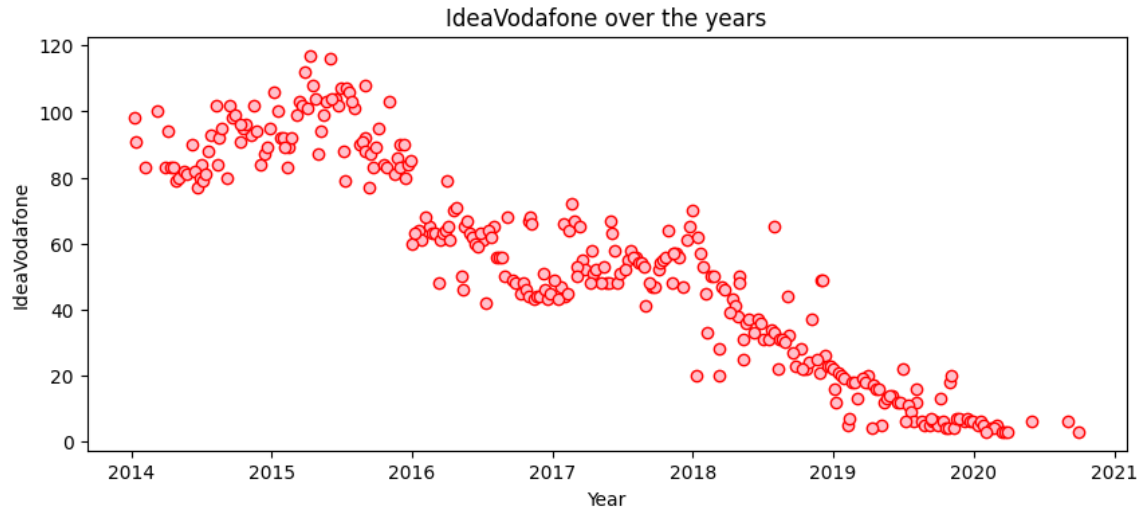
Fig 23: Price vs Time Plot 1



IdeaVodafone(Stock Price vs Time):

The price of this Stock has seen a continuous downward trend from the past 6 years and it seems pretty bearish.

Fig 24: Price vs Time Plot 2



Calculate Returns for all stocks with inference:

Steps for calculating returns from prices:

- Take logarithms
- Take differences

The logarithmic returns of all the stocks each week is as follows(Showing only first 5 and last 5 records)

Fig 25: Logarithmic Returns

	Infosys	IndianHotel	Mahindra&Mahindra	AxisBank	SAIL	ShreeCement	SunPharma	JindalSteel	IdeaVodafone	JetAirways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846
...
309	0.009649	-0.110348	0.030305	-0.057580	-0.087011	0.023688	0.072383	-0.053346	-0.287682	-0.127833
310	-0.139625	-0.051293	-0.093819	-0.145324	-0.095310	-0.081183	-0.043319	-0.187816	0.693147	-0.200671
311	-0.094207	-0.236389	-0.285343	-0.284757	-0.105361	-0.119709	-0.050745	-0.141830	-0.693147	-0.117783
312	0.109856	-0.182322	-0.091269	-0.173019	-0.251314	-0.067732	-0.076851	-0.165324	0.000000	-0.133531
313	-0.017228	0.000000	-0.031198	0.051432	0.090972	-0.006816	0.040585	-0.081917	0.000000	0.000000

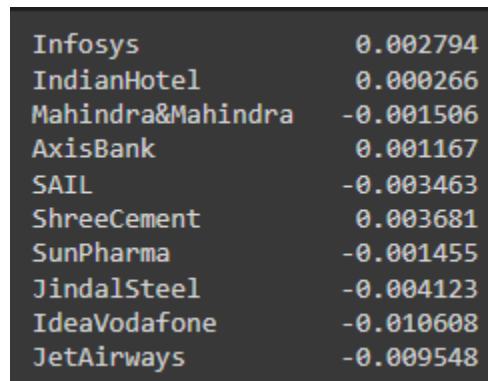
This can be interpreted as the logarithmic change of stock price of a week compared to its previous week. If it is positive then it means the stock has seen a price increase from previous and if its negative then it indicates

the stock price has been reduced since last week. Higher the value, higher is the change of price compared to previous week.

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

Stock means here indicate the average return on logarithmic scale over six years,

Fig 26: Stock Means

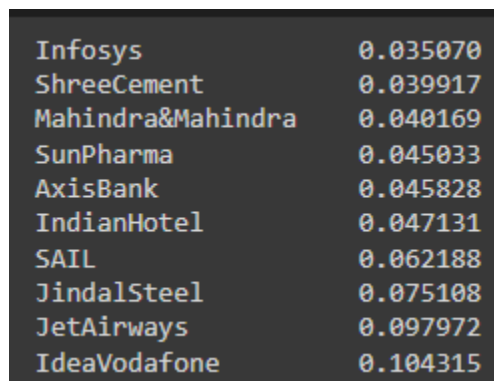


Infosys	0.002794
IndianHotel	0.000266
Mahindra&Mahindra	-0.001506
AxisBank	0.001167
SAIL	-0.003463
ShreeCement	0.003681
SunPharma	-0.001455
JindalSteel	-0.004123
IdeaVodafone	-0.010608
JetAirways	-0.009548

As per the data, only Infosys, IndianHotel, AxisBank and ShreeCement are the companies that gave profits when stock in these companies are held for six years. IdeaVodafone Seems to be the lowest performing stock from the given list followed by JetAirways and JindalSteel.

Stock Standard Deviation indicates the volatility/risk associated with each stock. In long term,

Fig 27: Stock Standard Deviation



Infosys	0.035070
ShreeCement	0.039917
Mahindra&Mahindra	0.040169
SunPharma	0.045033
AxisBank	0.045828
IndianHotel	0.047131
SAIL	0.062188
JindalSteel	0.075108
JetAirways	0.097972
IdeaVodafone	0.104315

Above is the volatility of stocks in ascending order, Infosys is the stock with least stock followed by ShreeCement and Mahindra&Mahindra.

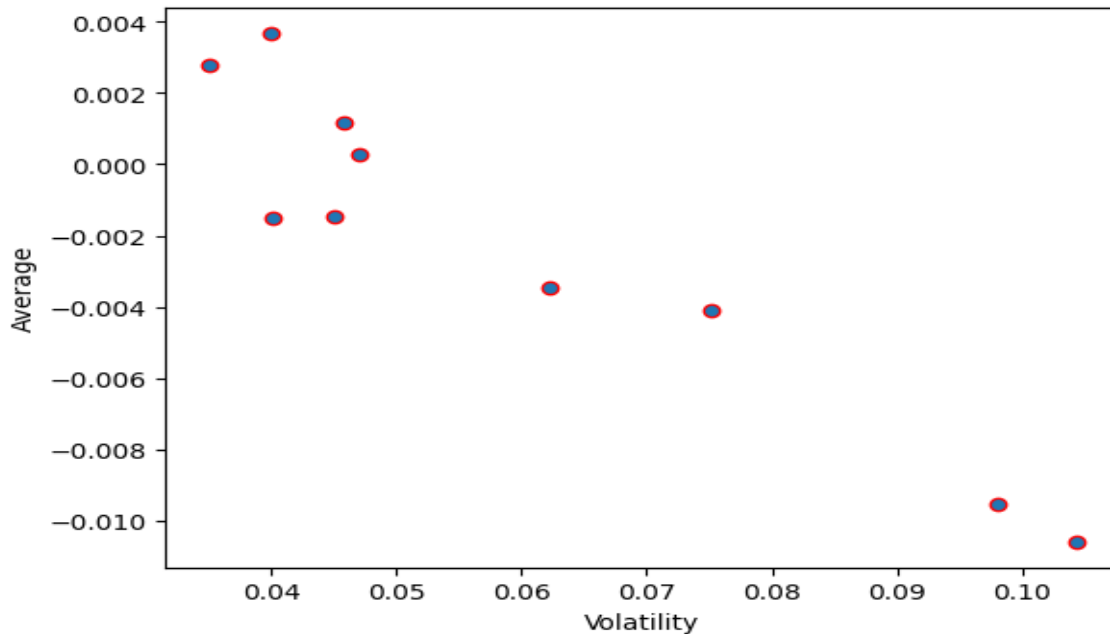
IdeaVodafone is the stock with high volatility followed by JetAirways and JindalSteel.

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

Below is the graph between Stock Means(Returns) and Standard Deviation(Volatility) for all the stocks.

As per the graph, the average returns are increasing with decrease in volatility. Stocks with high average returns are the ones with least volatility. Hence these stocks are stable and high performing stocks.

Fig 28: Stock Mean vs Standard Deviation Plot



Conclusions and Recommendations:

Best stocks are the ones with higher average and less volatility, now if we sort the stocks by higher average and less volatility, we could pick the best stocks from the top to bottom. The top ones are the best performing stocks compared to bottom ones.

As per this analogy, if we have to pick top 5 stocks, then ShreeCement, Infosys, AxisBank, IndianHotel and SunPharms would be the best choice.

Hence when a portfolio to be built from the below 10 stocks, then pick the top 4 stocks and avoid the rest. Provided, there are no developments in the bottom stocks.

Fig 29: Best Stocks Top to Bottom

	Average	Volatility
ShreeCement	0.003681	0.039917
Infosys	0.002794	0.035070
AxisBank	0.001167	0.045828
IndianHotel	0.000266	0.047131
SunPharma	-0.001455	0.045033
Mahindra&Mahindra	-0.001506	0.040169
SAIL	-0.003463	0.062188
JindalSteel	-0.004123	0.075108
JetAirways	-0.009548	0.097972
IdeaVodafone	-0.010608	0.104315