Finance & Risk Analytics - Business Report

Rohan R. Khade

Table of Contents

List of Tables

List of Figures

Chapter 1. FRA Project (Milestone-1)

1.1 Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests 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 (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

1.1.1 Data Dictionary

Table 1 Dataframe: df (with head function)

#	Field Name	Description	New Field Name
1	Co_Code	Company Code	Co_Code
2	Co_Name	Company Name	Co_Name
3	Networth Next Year	Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)	Networth_Next_Year
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders	Equity_Paid_Up
5	Networth	Value of a company as on 2015 - Current Year	Networth
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company	Capital_Employed
7	Total Debt	The sum of money borrowed by the company and is due to be paid	Total_Debt
8	Gross Block	Total value of all of the assets that a company owns	Gross_Block
9	Net Working Capital	The difference between a company's current assets (cash, accounts receivable, inventories of raw	Net_Working_Capital

Back to Top

		materials and finished goods) and its current liabilities (accounts payable).	
10	Current Assets	All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year.	Curr_Assets
11	Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)	Curr_Liab_and_Prov
12	Total Assets/Liabilities	Ratio of total assets to liabailities of the company	Total_Assets_to_Liab
13	Gross Sales	The grand total of sale transactions within the accounting period	Gross_Sales
14	Net Sales	Gross sales minus returns, allowances, and discounts	Net_Sales
15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)	Other_Income
16	Value Of Output	Product of physical output of goods and services produced by company and its market price	Value_Of_Output
17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service	Cost_of_Prod
18	Selling Cost	Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms)	Selling_Cost
19	PBIDT	Profit Before Interest, Depreciation & Taxes	PBIDT
20	PBDT	Profit Before Depreciation and Tax	PBDT
21	PBIT	Profit before interest and taxes	PBIT
22	PBT	Profit before tax	PBT
23	PAT	Profit After Tax	PAT
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit	Adjusted_PAT
26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.	CP
27	Revenue earnings in forex	Revenue earned in foreign currency	Rev_earn_in_forex
28	Revenue expenses in forex	Expenses due to foreign currency transactions	Rev_exp_in_forex
29	Capital expenses in forex	Long term investment in forex	Capital_exp_in_forex
30	Book Value (Unit Curr)	Net asset value	Book_Value_Unit_Curr
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value	Book_Value_Adj_Unit_Curr

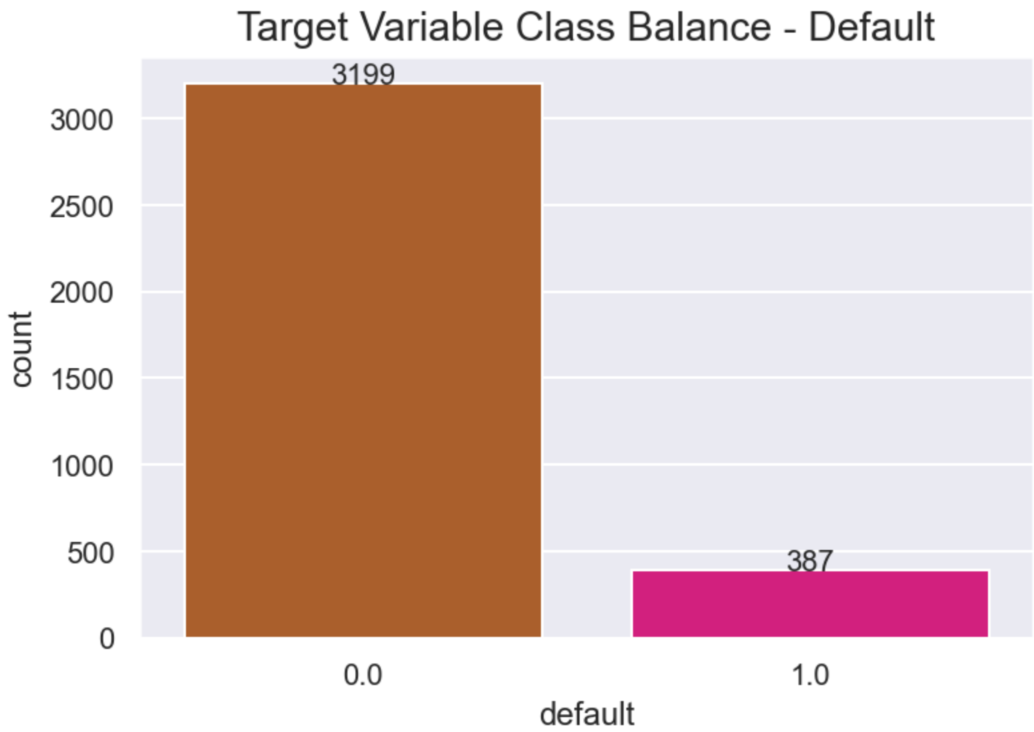
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share	Market_Capitalisation
33	CEPS (annualised) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis	CEPS_annualised_Unit_Curr
34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities	Cash_Flow_From_Opr
35	Cash Flow From Investing Activities	Cash used in the purchase of non-current assets–or long-term assets–that will deliver value in the future	Cash_Flow_From_Inv
36	Cash Flow From Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)	Cash_Flow_From_Fin
37	ROG-Net Worth (%)	Rate of Growth - Networth	ROG_Net_Worth_perc
38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed	ROG_Capital_Employed_perc
39	ROG-Gross Block (%)	Rate of Growth - Gross Block	ROG_Gross_Block_perc
40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales	ROG_Gross_Sales_perc
41	ROG-Net Sales (%)	Rate of Growth - Net Sales	ROG_Net_Sales_perc
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production	ROG_Cost_of_Prod_perc
43	ROG-Total Assets (%)	Rate of Growth - Total Assets	ROG_Total_Assets_perc
44	ROG-PBIDT (%)	Rate of Growth- PBIDT	ROG_PBIDT_perc
45	ROG-PBDT (%)	Rate of Growth- PBDT	ROG_PBDT_perc
46	ROG-PBIT (%)	Rate of Growth- PBIT	ROG_PBIT_perc
47	ROG-PBT (%)	Rate of Growth- PBT	ROG_PBT_perc
48	ROG-PAT (%)	Rate of Growth- PAT	ROG_PAT_perc
49	ROG-CP (%)	Rate of Growth- CP	ROG_CP_perc
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex	ROG_Rev_earn_in_forex_perc
51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex	ROG_Rev_exp_in_forex_perc
52	ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation	ROG_Market_Capitalisation_perc
53	Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year	Curr_Ratio_Latest
54	Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating	Fixed_Assets_Ratio_Latest

55	Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company	Inventory_Ratio_Latest
56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company	Debtors_Ratio_Latest
57	Total Asset Turnover Ratio[Latest]	The value of a company's revenues relative to the value of its assets	Total_Asset_Turnover_Ratio_Latest
58	Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt	Interest_Cover_Ratio_Latest
59	PBIDTM (%)[Latest]	Profit before Interest Depreciation and Tax Margin	PBIDTM_perc_Latest
60	PBITM (%)[Latest]	Profit Before Interest Tax Margin	PBITM_perc_Latest
61	PBDTM (%)[Latest]	Profit Before Depreciation Tax Margin	PBDTM_perc_Latest
62	CPM (%)[Latest]	Cost per thousand (advertising cost)	CPM_perc_Latest
63	APATM (%)[Latest]	After tax profit margin	APATM_perc_Latest
64	Debtors Velocity (Days)	Average days required for receiving the payments	Debtors_Vel_Days
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers	Creditors_Vel_Days
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales	Inventory_Vel_Days
67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets	Value_of_Output_to_Total_Assets
68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block	Value_of_Output_to_Gross_Block

1.1.2 Project Details

- Total Number of Companies (observations) = 3586
- Total Number of Variables = 67 (1 target and 66 predictors)
- **Target Variable:**
  - We have created a target variable - **'default'**
  - Where, if Net-worth next year is zero or positive, then default = 0
  - If Net-worth next year is negative, then default = 1

Figure 1. Class Balance of Target Variable



- Number of Duplicates = 0
- Missing Value Treatment:
    - Less than 1% missing values present
    - We impute these missing values by using KNN Imputer (n_neighbors=10)
- Zero Values:
    - Large amount of zero values present (total = 15.1 %)
    - We drop columns with more than 30% of zero values (9 columns)
    - We found that 164 out of total 387 defaulting companies had more than 5 zero values in their rows
        - We conclude, more the missing or zero values, higher is the probability of default
    - For the rest of columns: We convert zeros to Missing NaN values
    - Impute all these missing values using KNN Imputer (n_neighbors=10)
- Outlier Treatment:
    - IQR and Z-Score methods - used separately to identify and treat outliers
    - Different Logistic Regression models fitted and tested using both
    - Z-Score outlier treatment was found to give better results on Test Data
- Scaling - We use Z-score Standard scaling
- Multi-Collinearity:
    - Many variables in the data are extracts of each other
    - Hence, there is a high correlation between many of them
    - This causes multi-collinearity and can harm a model's interpretability
    - Also, these columns don't add any more value to predictions by regression

  - Variance Inflation Factor method is used to check and drop columns causing Multi-Collinearity
  - Recursively, one-by-one, columns with VIF > 5 are dropped
- Feature Engineering:
  - We start with large number of 66 predictor variables
  - There are various methods employed to extract the best features
  - Methods and Steps taken for all modelling:
    - Drop unique identifiers which add no value to predictions - Company Code and name: 64 variables remaining
    - Drop variables with zeros > 30% (9 cols dropped): 55 variables remaining
    - Drop variables one-by-one with VIF > 5: 27 variables remaining after IQR outliers: 23 varaibles remaining after zscore outliers
  - Also, for some models, we test by dropping insignificant variables for prediction (variables with p-values > 0.05) at 95% confidence
  - For Model #7 - we use Recursive Feature Elimination (RFE) technique to select 15 best features for modelling
- We choose Model #8 as the best model for deployment -
  - This has the best combination of Recall and Precision for default = 1
  - This model:
    - Outlier treatment: Z-score with values capped to ±3 std dev RFE: with top 15 features
    - Oversampling method: SMOTE with 50-50 balance of 0 & 1 Choosing Optimum Threshold = 0.5
  - Metrics for default = 1:
    - Recall = 95%, Precision = 78%, Accuracy = 96%, f1-score = 86%

1.2 Outlier Treatment

- Outlier treatment is necessary for any regression model
- In Regression, outliers pull the regression line towards itself thereby affecting its slope. This distorts the reality and leads to faulty predictions
- We employ 2 types of Outlier detection and treatments in this case study:
  - Inter-Quartile Range (IQR) Treatment
  - Z-score treatment
- We show box plots of 15 variables before and after Outlier treatment. We scale these variables for better comparison
- These 15 vars are finally chosen as the best predictors for Logistic Regression
- IQR Treatment:
  - Q1 = 25th percentile, Q3 = 75th percentile
  - IQR = Q3 - Q1
  - Outlier = any value which lies beyond 1.5 times of IQR from Q1 and Q3 on either side
  - We cap all outliers to this upper or lower level

Figure 2. Boxplot prior to Outlier Treatment

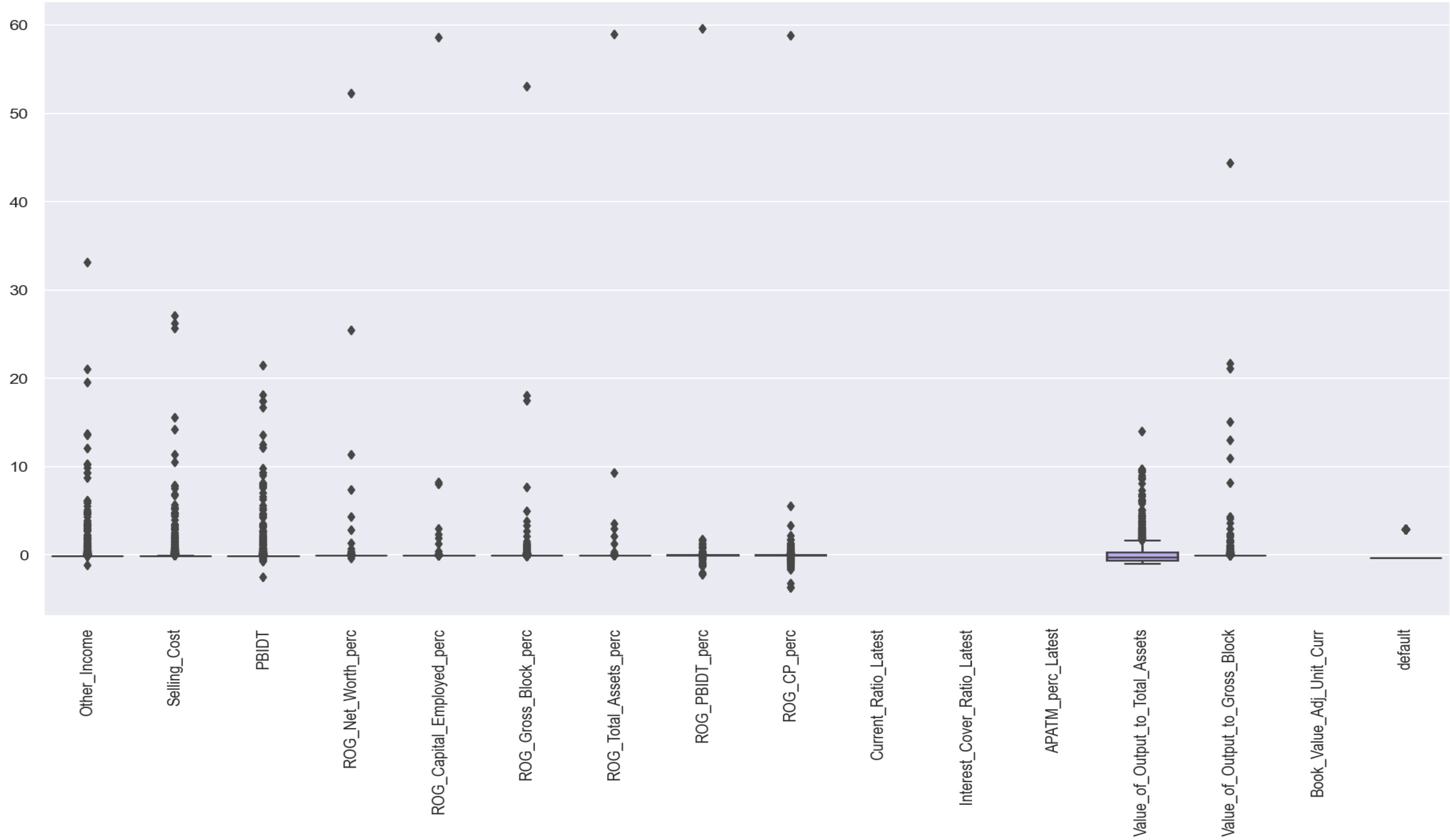
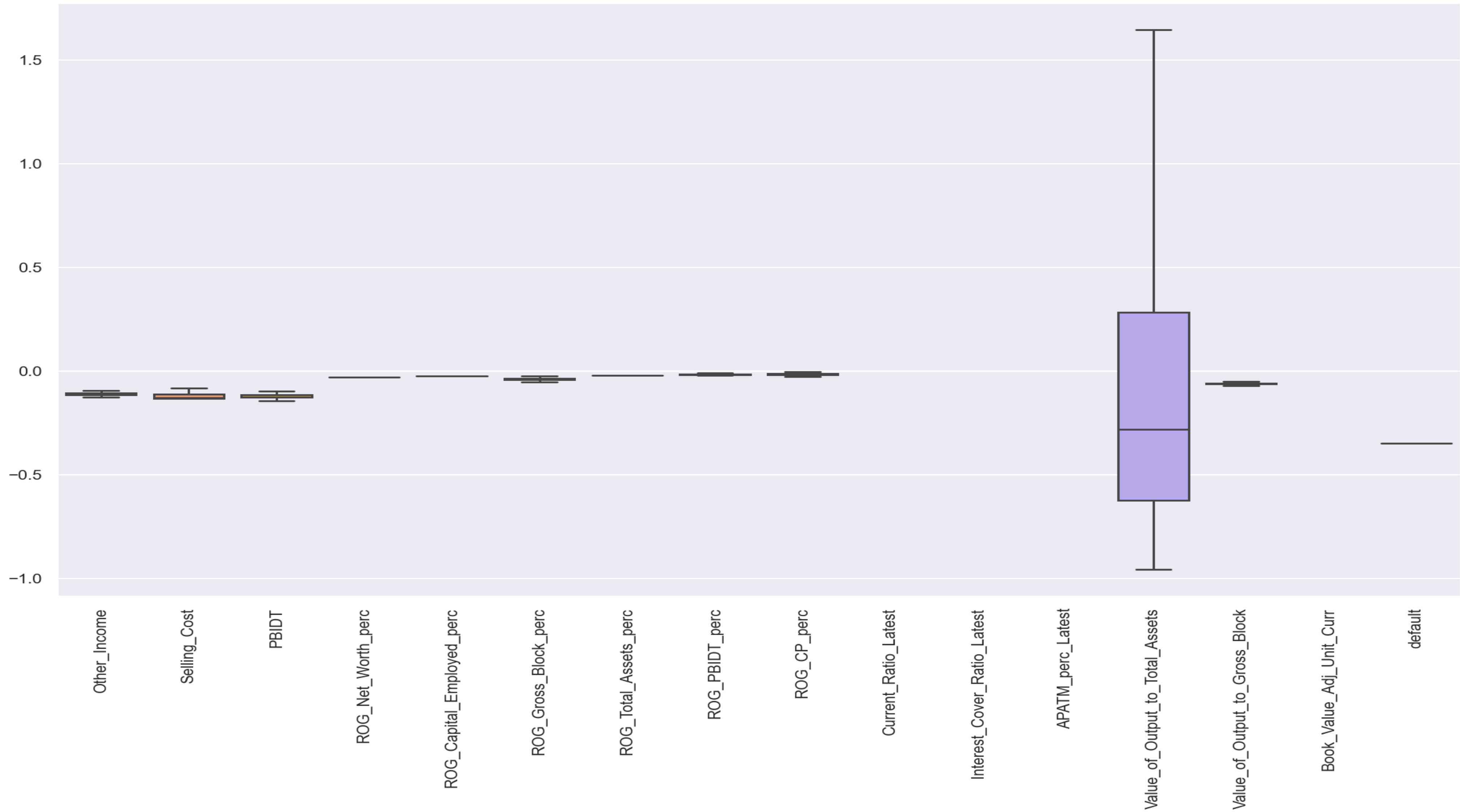


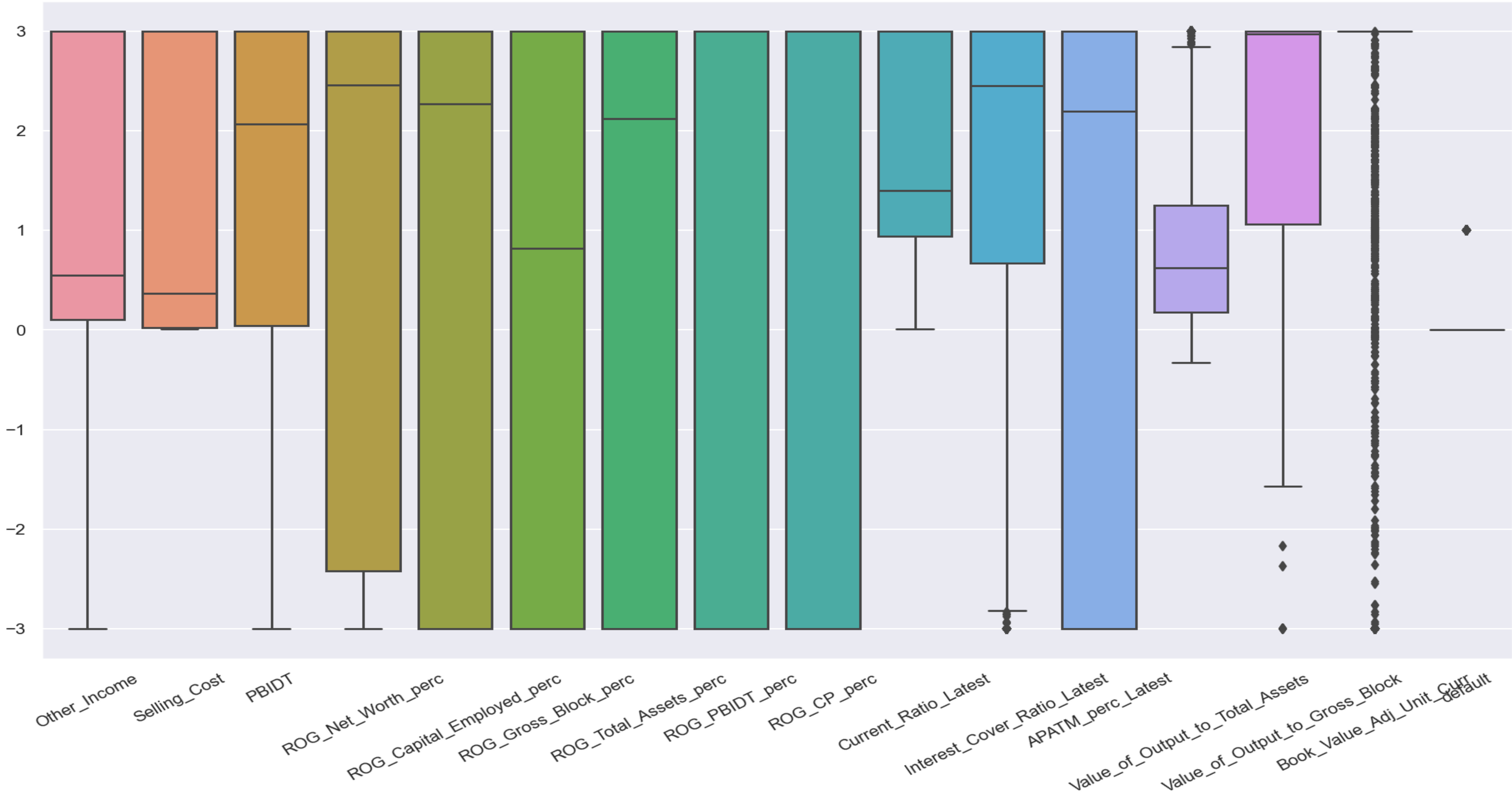
Figure 3. **Boxplot after IQR Treatment - Top 15 predictors - Z-scaled**



Back to Top

1.2.1 Z-score Treatment

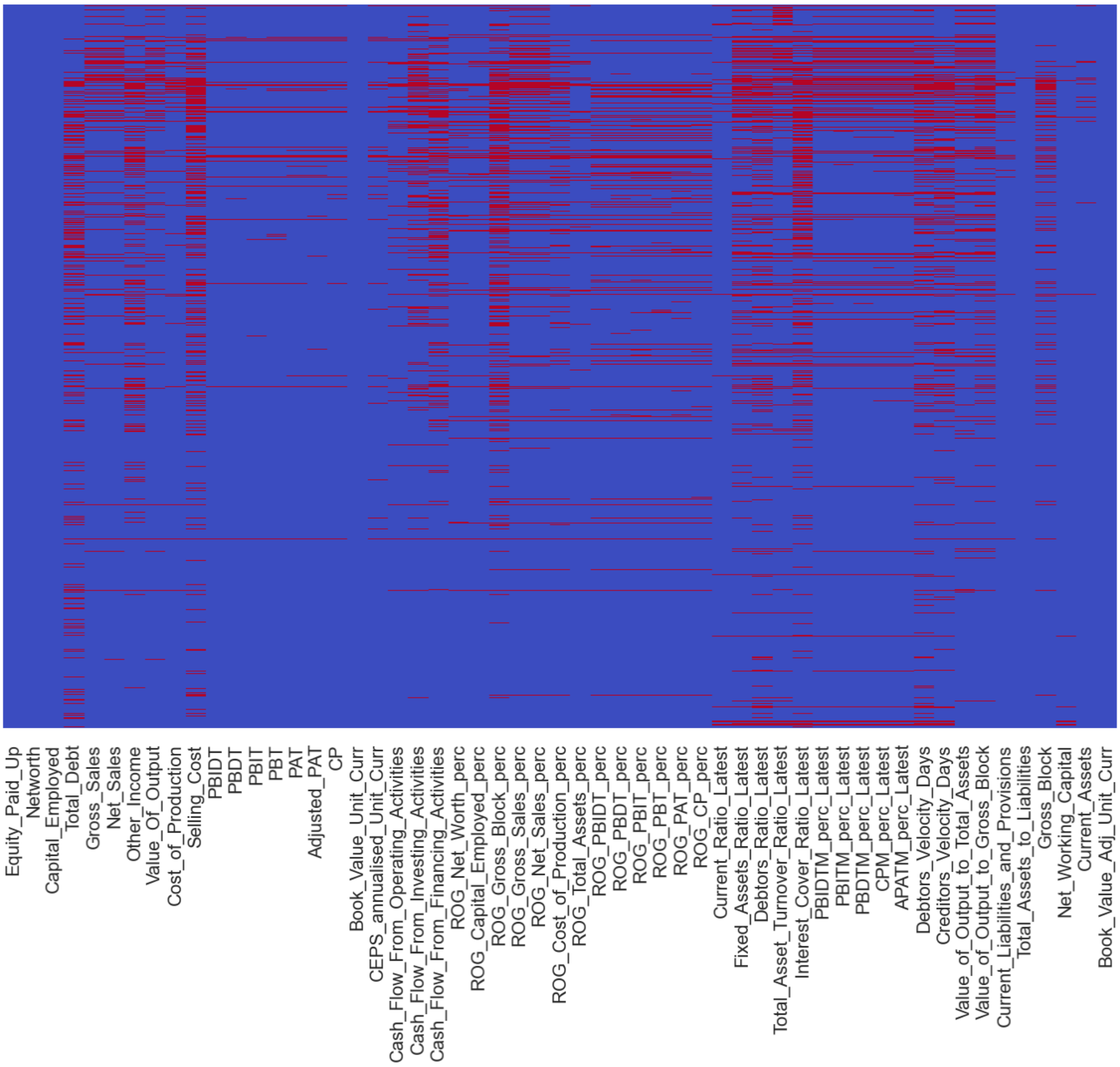
Figure 4. Correlation matrix of the seven variables with integer data type



1.3 Missing Value Treatment

- Missing values in the raw data are very less, about 0.05%
- But there are large number of zero values, which are mostly placeholders for missing values, about 15%
- Also, these zero values add no more value to predictions
- But also mainly, large number of zero values in any feature cause 'Linear Algebra Error' while using StatsModel
- Hence, it is of paramount importance to treat these zero values
- Firstly, we drop all those features with zero values greater than 30%
- Then, we convert all other zero values to Missing Values (NaN values)
- These transformed and original missing values together are imputed using KNN Imputer (n_neighbors=10)
- A visual of all these missing values is give below - after dropping variables

Figure 5. Box plot with outliers



1.4 Transform Target variable into 0 and 1

- We check the financial health of companies
- We'll base our prediction on Company's health on whether they will have a positive Net-worth next year or negative
- Hence, We consider 'Networth Next Year' as our Default Variable
- So, we call negative values as Default = 1
- And, zero or positive values as Default = 0
- We convert accordingly - Below is the sample

Table 2 Target Variable - First 5

Default	Networth_Next_Year
1	-8021.6
1	-3986.19
1	-3192.58
1	-3054.51
1	-2967.36

Table 3 Target Variable - Last 5

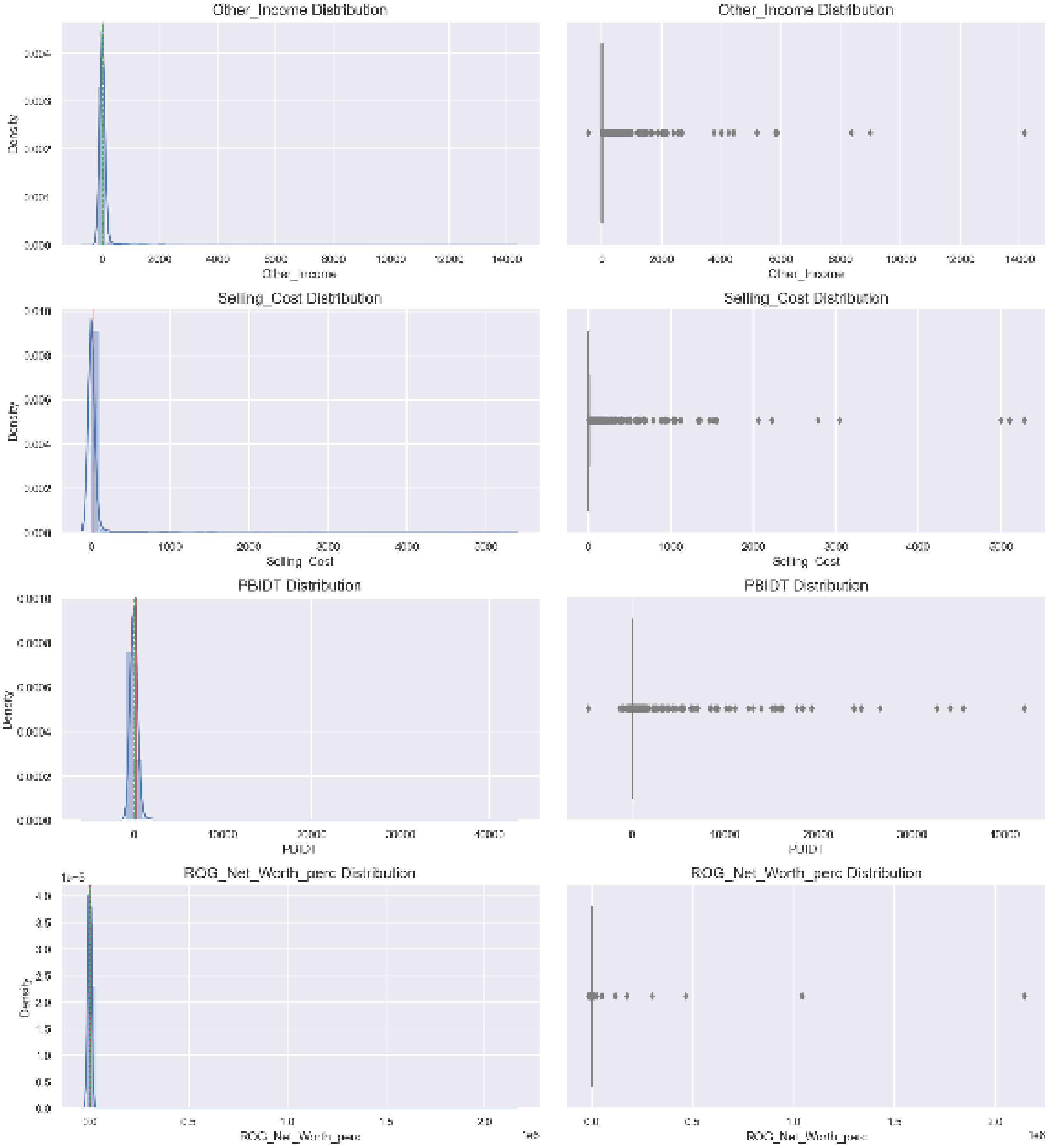
Default	Networth_Next_Year
0	72677.77
0	79162.19
0	88134.31
0	91293.7
0	111729.1

Back to Top

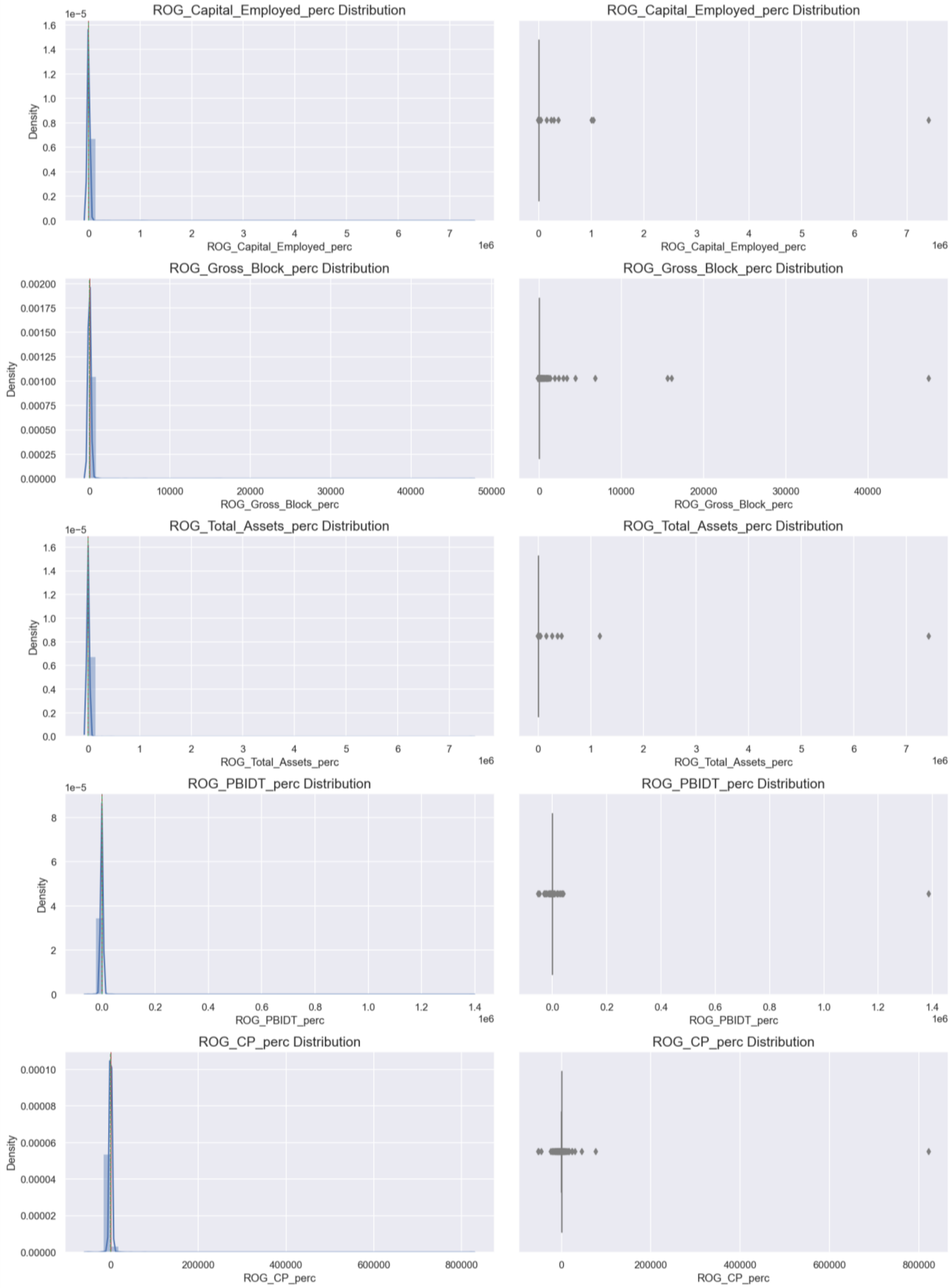
1.5 Univariate (4 marks) & Bivariate (6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

1.5.1 Univariate Analysis

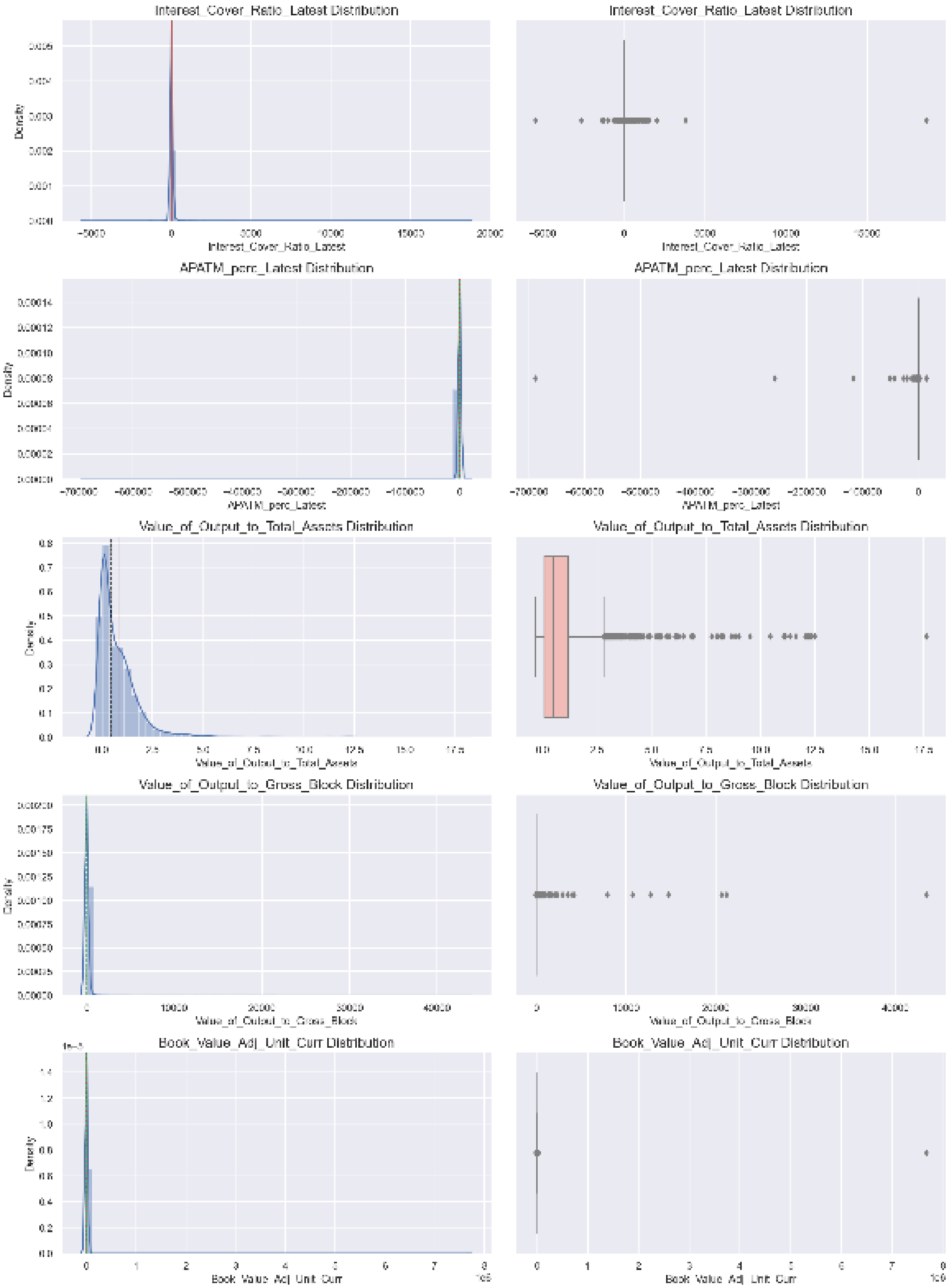
Figure 6. Distribution of top 15 variables - Z-Scaled



Back to Top


ROG_Capital_Employed_perc Distribution
Density
ROG_Capital_Employed_perc
ROG_Capital_Employed_perc Distribution
ROG_Capital_Employed_perc
ROG_Gross_Block_perc Distribution
Density
ROG_Gross_Block_perc
ROG_Gross_Block_perc Distribution
ROG_Gross_Block_perc
ROG_Total_Assets_perc Distribution
Density
ROG_Total_Assets_perc
ROG_Total_Assets_perc Distribution
ROG_Total_Assets_perc
ROG_PBIDT_perc Distribution
Density
ROG_PBIDT_perc
ROG_PBIDT_perc Distribution
ROG_PBIDT_perc
ROG_CP_perc Distribution
Density
ROG_CP_perc
ROG_CP_perc Distribution
ROG_CP_perc

Back to Top

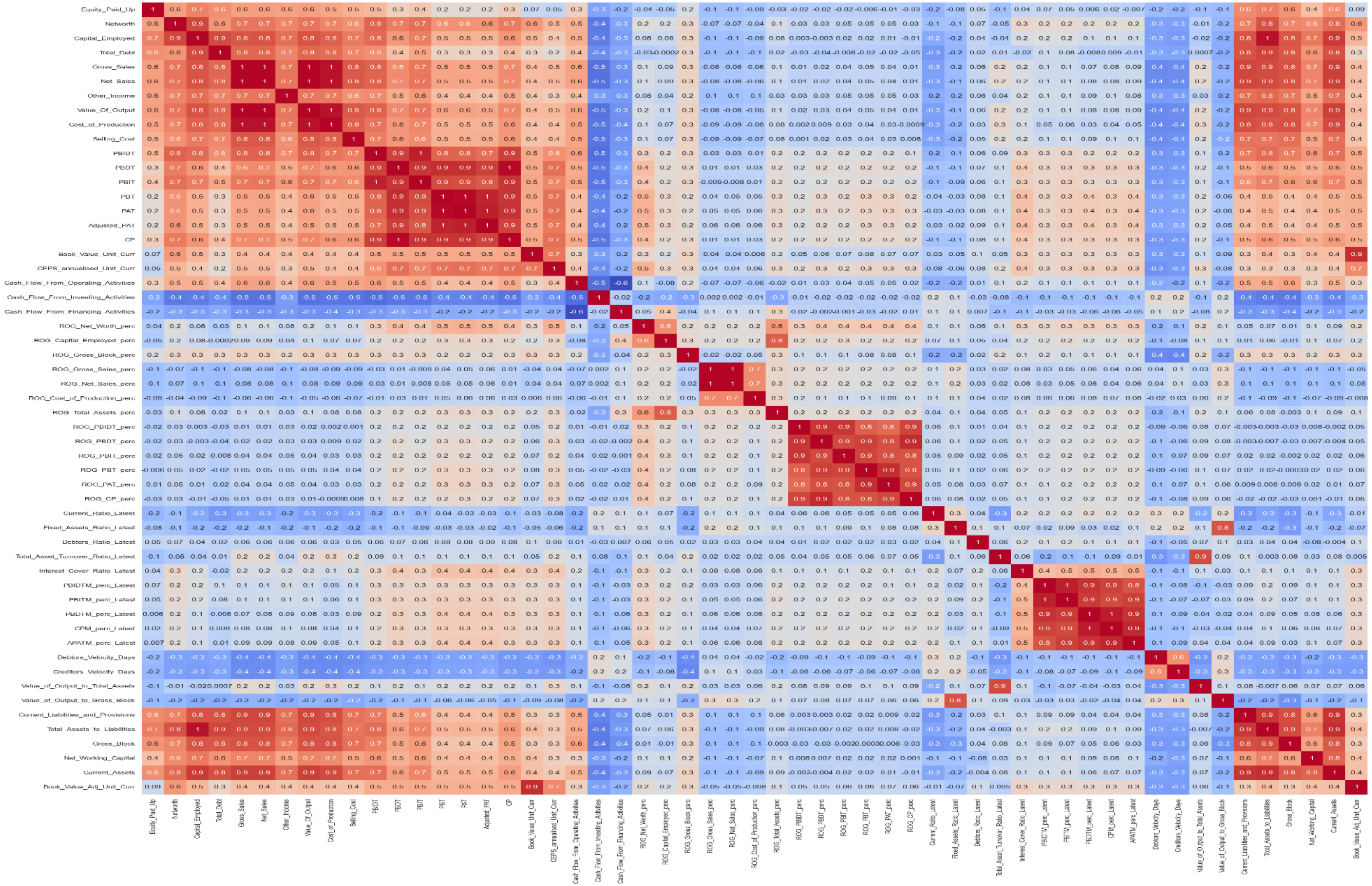

Interest_Cover_Ratio_Latest Distribution
Interest_Cover_Ratio_Latest Distribution
Density
Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest
APATM_perc_Latest Distribution
APATM_perc_Latest Distribution
Density
APATM_perc_Latest
APATM_perc_Latest
Value_of_Output_to_Total_Assets Distribution
Value_of_Output_to_Total_Assets Distribution
Density
Value_of_Output_to_Total_Assets
Value_of_Output_to_Total_Assets
Value_of_Output_to_Gross_Block Distribution
Value_of_Output_to_Gross_Block Distribution
Density
Value_of_Output_to_Gross_Block
Value_of_Output_to_Gross_Block
Book_Value_Adj_Unit_Curr Distribution
Book_Value_Adj_Unit_Curr Distribution
Density
Book_Value_Adj_Unit_Curr
Book_Value_Adj_Unit_Curr

- Distribution of Z-Scaled data of top 15 variables given
- Colored vertical lines in the distribution indicate central tendencies
- ‘Selling Cost’ - has max companies around its mean. They have Right Skew with outliers on higher side.
- ‘PBIDT’ - ‘Profit before Int Depreciation and Tax’ - max companies are around the mean with a prominent right skew. This indicates that there are still many companies with high PBIDT
- ‘Cash Flow from Operating Activities’ - normal distribution with max companies lying around the mean
- ‘ROG Networth’, ‘ROG Capital Employed’, ‘ROG Total Assets’, ‘ROG PBIDT’, ‘ROG PBT (Profit Before Tax)’, ‘ROG CP’, ‘Current ratio Latest’, ‘Interest Cover Ratio Latest’, ‘Value of Output to Total Assets’, ‘Net Working Capital’, ‘Book Value Adjusted’ - these variables have max density of companies around its mean with right skew. This indicates outliers on the higher side.
- ‘APATM (After Tax Profit Margin)’ - has max density around its mean and a prominent left skew. This indicates that there are many companies have their Net Profit on the lower side of the distribution - Possible indication of default
- Largely, it is observed that there are many companies with good margin and financials before tax and all other costs. But, after costs are considered, they slide to the lower half - Shows they need to work on their costs and bottom line

Back to Top

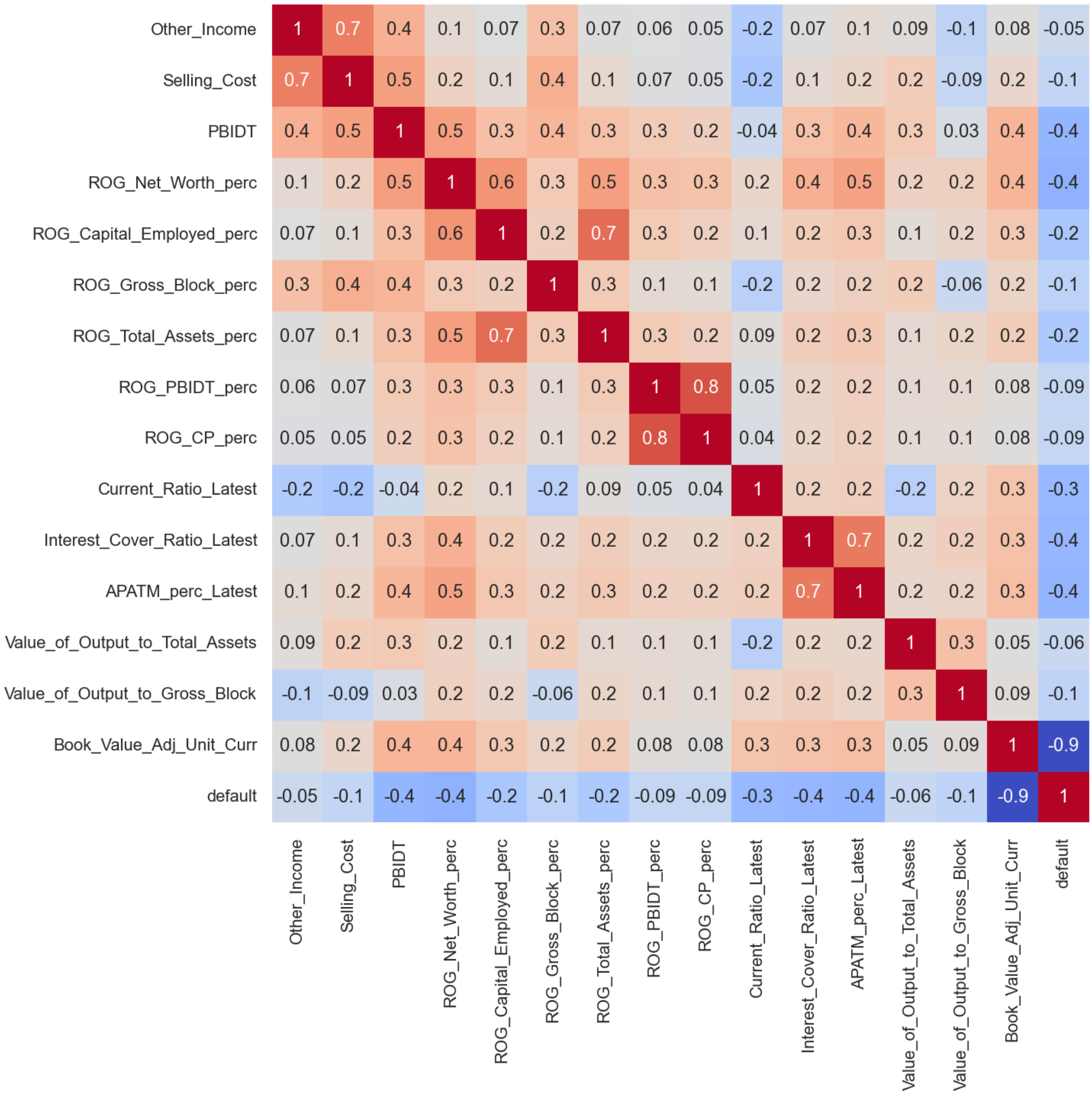
1.5.2 Bivariate Analysis

Figure 7. Correlation Heatmap for 55 variables



- There are a lot of red patches seen. This indicates high correlation between many variables
- Highly correlated features cause multi-collinearity which affect the interpretability of Logistic Regression model. They are best removed.
- We use Variance Inflation factor method and remove all variables with VIF > 5. This is done recursively, one-by-one
- Correlation heat-map of top 15 predictors and 1 Target, used to get the best model is given below:

Figure 8. Correlation Heatmap for top 15 variables



- 'ROG-Capital Employed and ROG-Total Assets' - 'ROG-PBT and ROG-PBIDT' - 'ROG-CP and ROG-PBIDT' - 'ROG-CP and ROG-PBT'
  - The above pairs of features show high correlation
  - It looks obvious as they seem derived or direct functions of each other

- Target variable 'default' has high negative correlation with 'Book Value Adj'
  - This indicates as Book Value rises, Probability of Default falls

1.6 Train Test Split

- We use train_test_split function from scikit-learn library to split the data into train and validation sets
- We split in the ratio of 67-33 - 67% in Training Set and 33% in Testing (Validation) Set
- We seed this split at random_state=42
- So, after split, Out of Total 3586: **Train Set has 2402 observations**
  Test Set has 1184 observations

1.7 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach

PREPROCESSING DATA FOR ALL MODELS

- We had large number of 66 predictor variables in the raw data
- 2 unique identifiers - Company Code and Name - were dropped
- There are large number of zero values in the data. We drop all variables with zeros > 30% (9 vars)
- Highly correlated features exist in the data, which cause multi-collinearity. This indicates existence of redundant features
- We perform Outlier treatments using IQR and Z-score methods
- We drop all features recursively one-by-one with VIF > 5 (VIF - Variance Inflation Factor)
- We are left with - 27 variables after IQR Outlier treatment and 23 variables after Z-score treatment

LOGISTIC REGRESSION MODELS

- We build multiple Logistic Regression models with different approaches and strategies. We test each model on Test set and fine-tune to improve Recall and Precision of default = 1
- We use StatsModel and SciKitLearn libraries to build and test these models
- **Model 1:**
  - 27 vars, IQR Outlier treated
- **Model 2:**
  - From 27 vars - insignificant vars dropped with p-values > 0.05 (final 10 vars)
  - IQR Outlier treatment
- **Model 3:**
  - 23 vars, Z-score Outlier treatment
- **Model 4:**
  - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
  - Z-score Outlier treatment

- **Model 5:**
    - Above 9 vars, Z Score treatment
    - Regularising the model by Hyper-parameter tuning with GridSearch over 10 folds over following: {'penalty':['l2','none', 'l1'], 'solver':['lbfgs', 'liblinear', 'sag', 'saga', 'newton-cg'], 'tol':[0.0001,0.00001]}
    - Best Parameters were found as follows: {'penalty': 'none', 'solver': 'lbfgs', 'tol': 0.0001}
- **Model 6:**
    - From 23 vars - insignificant vars dropped with p-values > 0.05 (final 9 vars)
    - Z-score Outlier treatment
    - Check for optimum threshold to get max Recall for default=1
    - This is obtained by maximising the difference between True Positivity rate and False Positivity rate (tpr - fpr) and Optimum Threshold = 0.084
- **Model 7:**
    - 23 vars, Z-score Outlier treatment
    - Extracting top 15 features using Recursive Feature Elimination (RFE)
- **Model 8:**
    - 23 vars, Z-score Outlier treatment
    - This model gave the best metrics on Test Set
    - StatsModel report of Model 9 given below -
    - We note that 'Book_Value_Adj_Unit_Curr' has the highest negative coefficient suggesting that this variable has the highest negative impact on Probability of Default
    - Also, 'Selling_Cost' has the highest positive coefficient suggesting that this variable has the highest positive impact on Probability of Default

	Coefficient	Std. Error	Z	P>|z|	[0.025]	0.975]
Other_Income	0.3180	0.101	3.145	0.002	0.120	0.516
Selling_Cost	0.6835	0.119	5.754	0.000	0.451	0.916
PBIDT	-0.3132	0.056	-5.553	0.000	-0.424	-0.203
ROG_Net_Worth_perc	-0.4148	0.049	-8.388	0.000	-0.512	-0.318
ROG_Capital_Employed_pe rc	0.4119	0.054	7.656	0.000	0.306	0.517
ROG_Gross_Block_perc	0.0368	0.047	0.790	0.429	-0.055	0.128
ROG_Total_Assets_perc	-0.2154	0.055	-3.939	0.000	-0.323	-0.108
ROG_PBIDT_perc	0.2171	0.057	3.835	0.000	0.106	0.328
ROG_CP_perc	-0.1513	0.057	-2.677	0.007	-0.262	-0.041
Current_Ratio_Latest	-0.3419	0.107	-3.199	0.001	-0.551	-0.132
Interest_Cover_Ratio_Lates t	-0.3591	0.059	-6.077	0.000	-0.475	-0.243
APATM_perc_Latest	-0.2973	0.055	-5.371	0.000	-0.406	-0.189
Value_of_Output_to_Total_A ssets	-0.0817	0.177	-0.461	0.645	-0.429	0.266
Value_of_Output_to_Gross_ Block	0.2140	0.092	2.329	0.020	0.034	0.394
Book_Value_Adj_Unit_Current	-1.6181	0.071	-22.933	0.000	-1.756	-1.480

- **Model 9:**
  - 23 vars, Z-score Outlier treatment
  - Extracting top 15 features using Recursive Feature Elimination (RFE)
  - Balancing default labels (0s and 1s) 50-50 using Over Sampling technique - SMOTE
  - Check for optimum threshold to get max Recall for default = 1
  - This is obtained by maximising the difference between True Positivity rate and False Positivity rate (tpr - fpr)
  - Optimum Threshold = 0.4246
- **Model 9:**
  - Z-Score Outlier treatment, Top 15 features through RFE
  - Class balancing 50 - 50 using SMOTE
  - Optimum Threshold = 0.5
  - Dropping insignificant vars with p-values > 0.05

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

Performance Metrics of all models on Test Dataset:

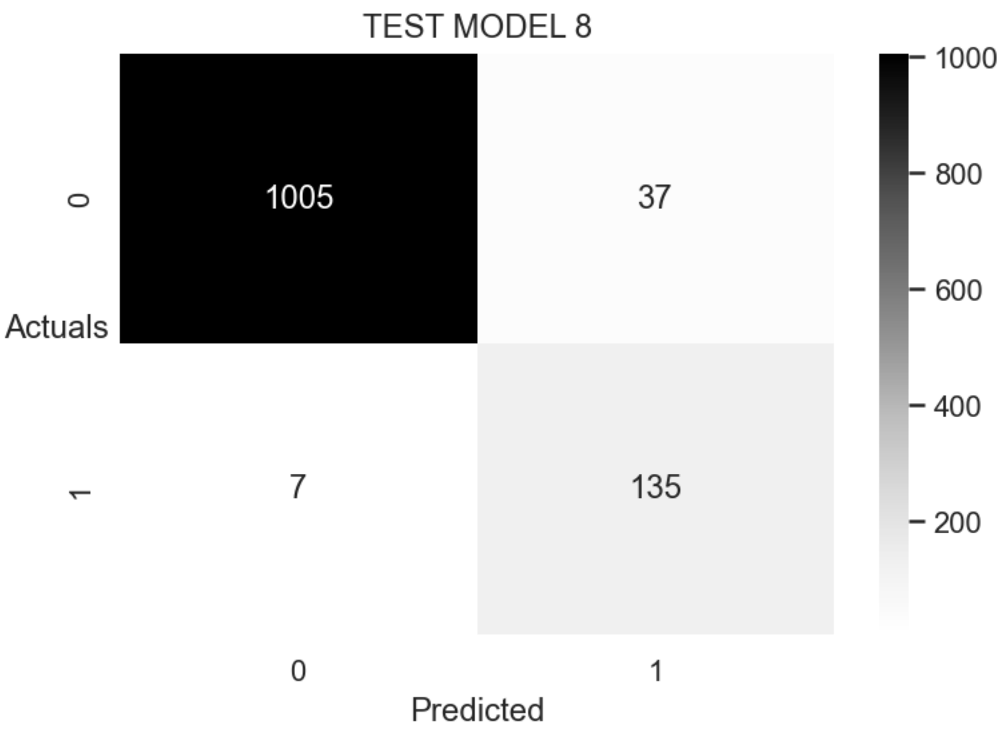
Table 4 All Model Performance Comparison

	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
Model 1	99	23	61	38
Model 2	99	23	60	37
Model 3	88	90	97	89
Model 4	88	89	97	88
Model 5	88	89	97	88
Model 6	95	71	95	81
Model 7	87	89	97	88
Model 8	95	78	96	86
Model 9	95	75	96	84
Model 10	92	78	96	85

Table 5 All Model Performance Comparison

	Precision	Recall	F1-score	Support
0	0.99	0.96	0.98	1042.00
1	0.78	0.95	0.86	142.00
accuracy	0.96	0.96	0.96	0.96
macro avg	0.89	0.96	0.92	1184.00
weighted avg	0.97	0.96	0.96	1184.00

Figure 9. Confusion Matrix of Model 8



Back to Top

Chapter 2. FRA Project (Milestone-2)

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

- We build a Random Forest model with GridSearch CV. We test each model on Test set and fine-tune to improve Recall and Precision of default = 1
- Parameters considered include 'max_depth': [1, 3, 5, 7, 9],
  'min_samples_leaf': [5, 10, 15, 20],
  'min_samples_split': [5, 15, 30, 45],
  'n_estimators': [25, 50]

- Best parameters include 'max_depth': 7,
  'min_samples_leaf': 15,
  'min_samples_split': 45,
  'n_estimators': 50

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

- Performance Metrics of the model on Train Dataset:

Table 6 Performance Metrics of the model on Train Dataset:

	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
RFCL	86	95	98	90

- Performance Metrics of the model on Test Dataset:

Table 7 Performance Metrics of the model on Test Dataset:

	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
RFCL	88	95	98	91

Back to Top

Figure 10. Confusion Matrix of RF Model

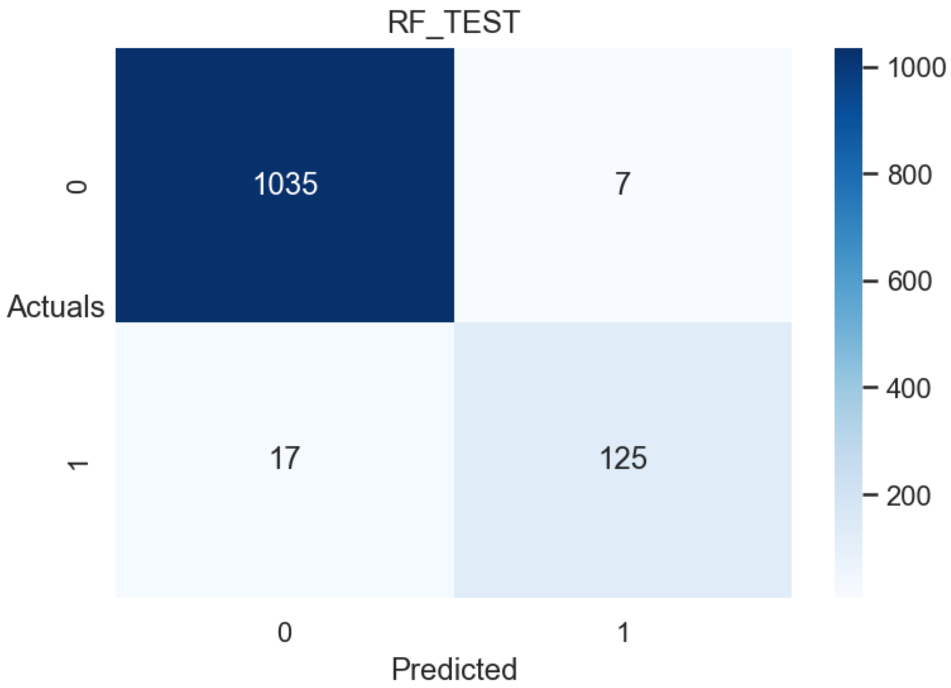


Figure 11. Train ROC Curve

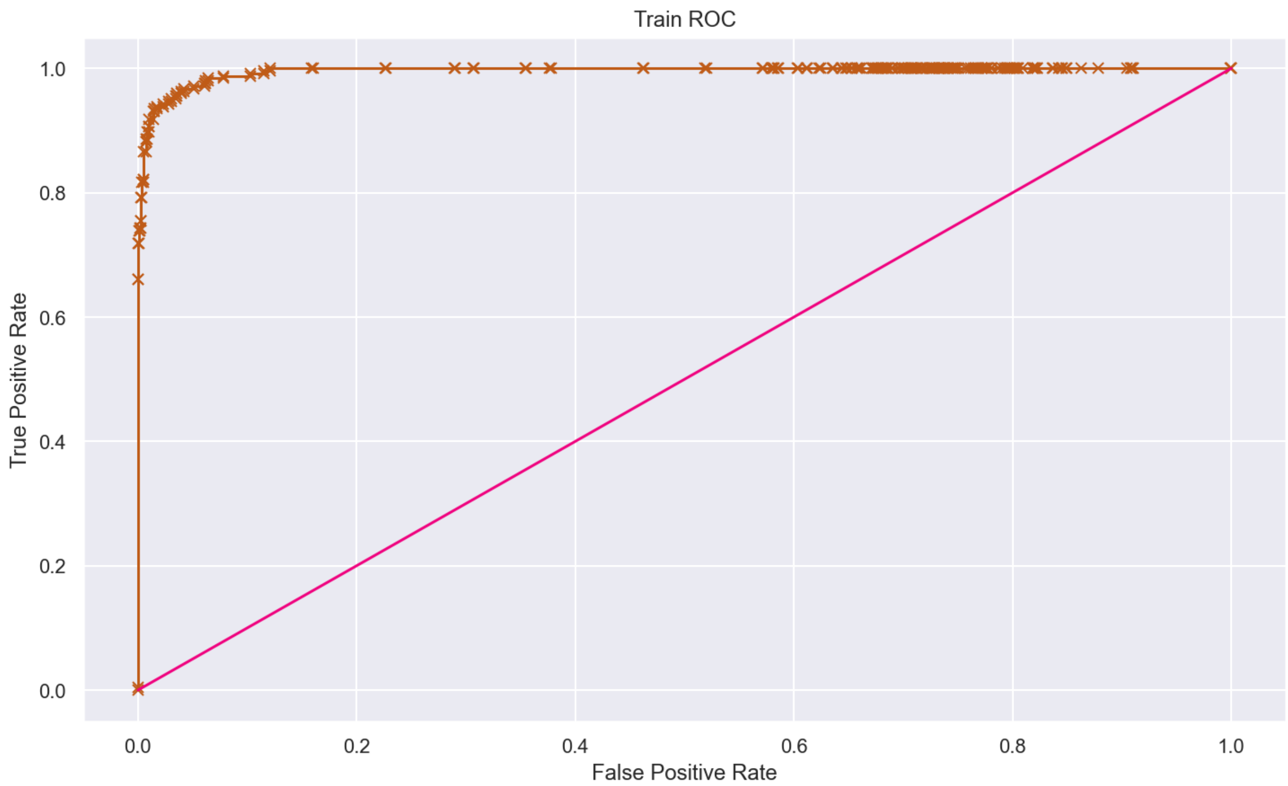
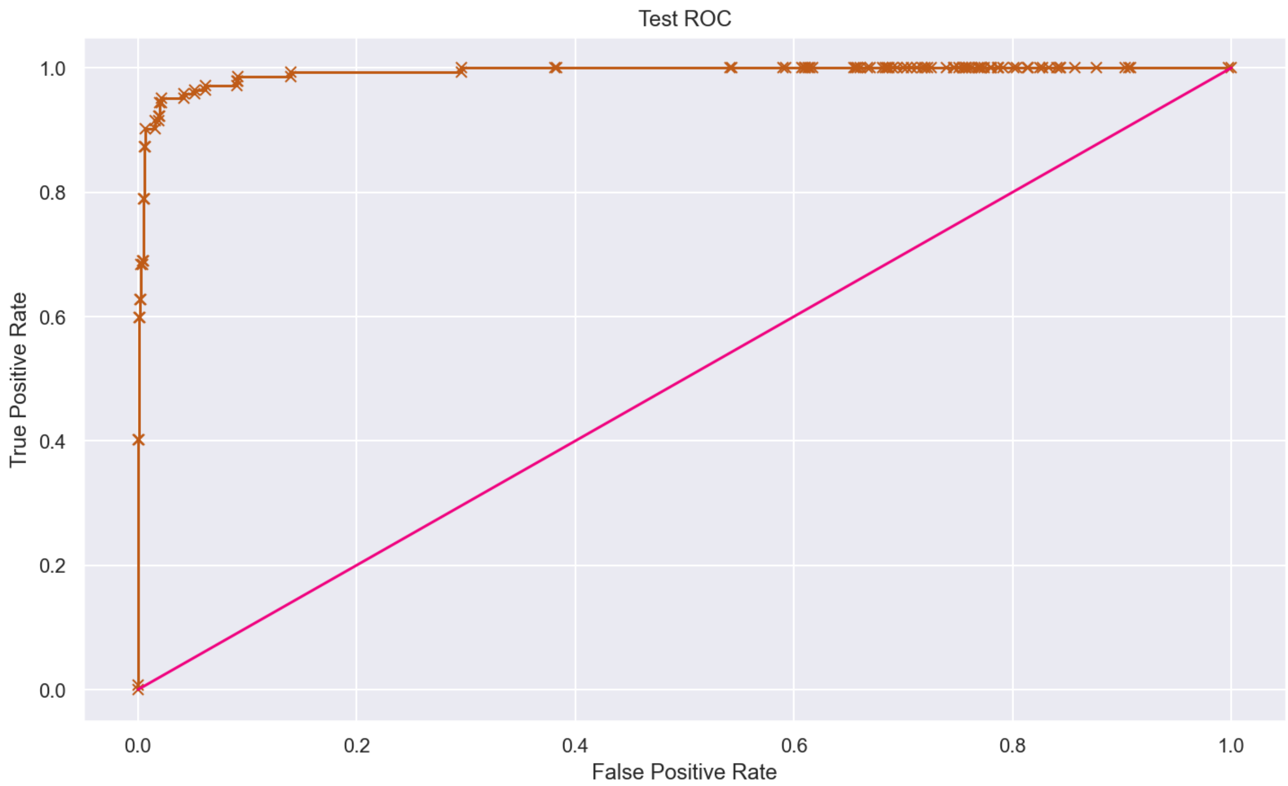


Figure 12. Test ROC Curve



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

- We build the LDA model post splitting the data and also build an ROC curve
- We also calculate the optimum threshold, which came to be 0.13140895885131224

2.4 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model

- Performance Metrics of the first model on Train & Test Dataset:

Table 8 Performance Metrics of the first model on Train & Test Dataset:

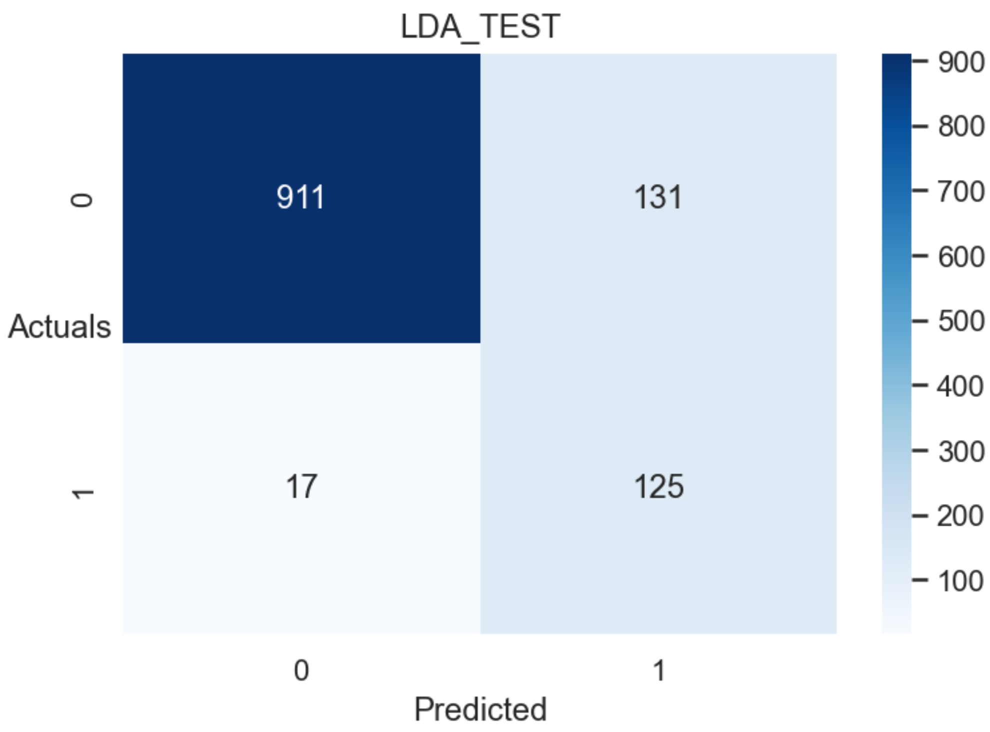
	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
Train	57	81	94	67
Test	61	78	93	68

- Performance Metrics of the optimized model on Train & Test Dataset:

Table 9 Performance Metrics of the optimized model on Train & Test Dataset:

	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
Train	87	48	89	62
Test	88	49	87	63

Figure 13. Confusion Matrix of Optimized LDA Model



Back to Top

Figure 14. Train ROC Curve

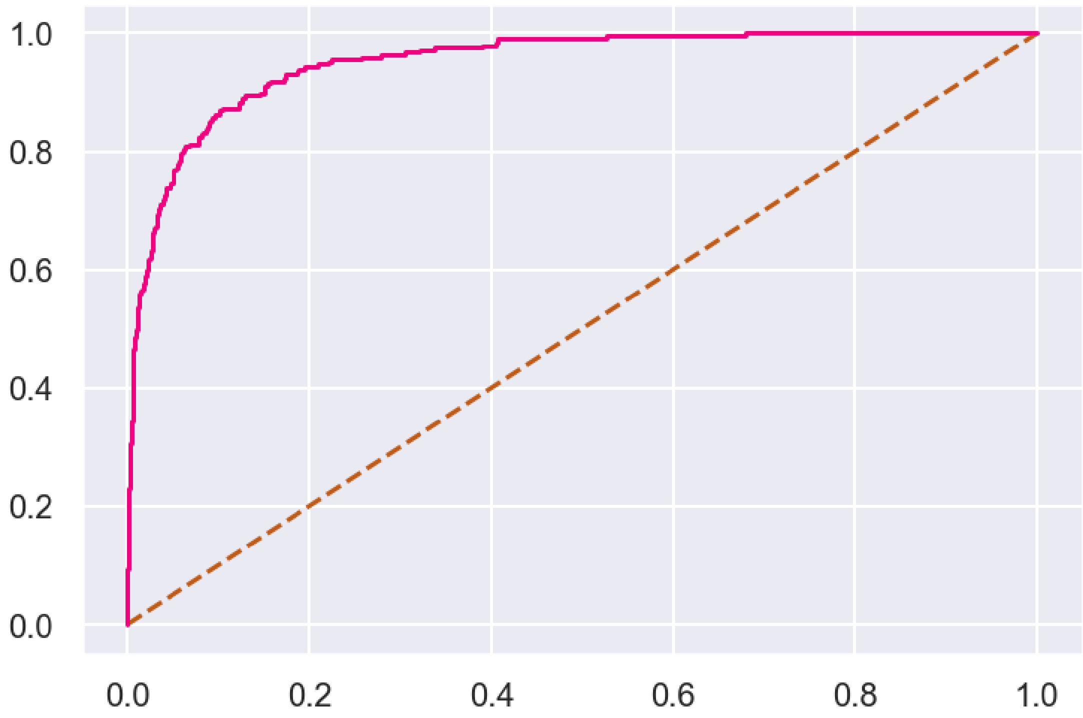
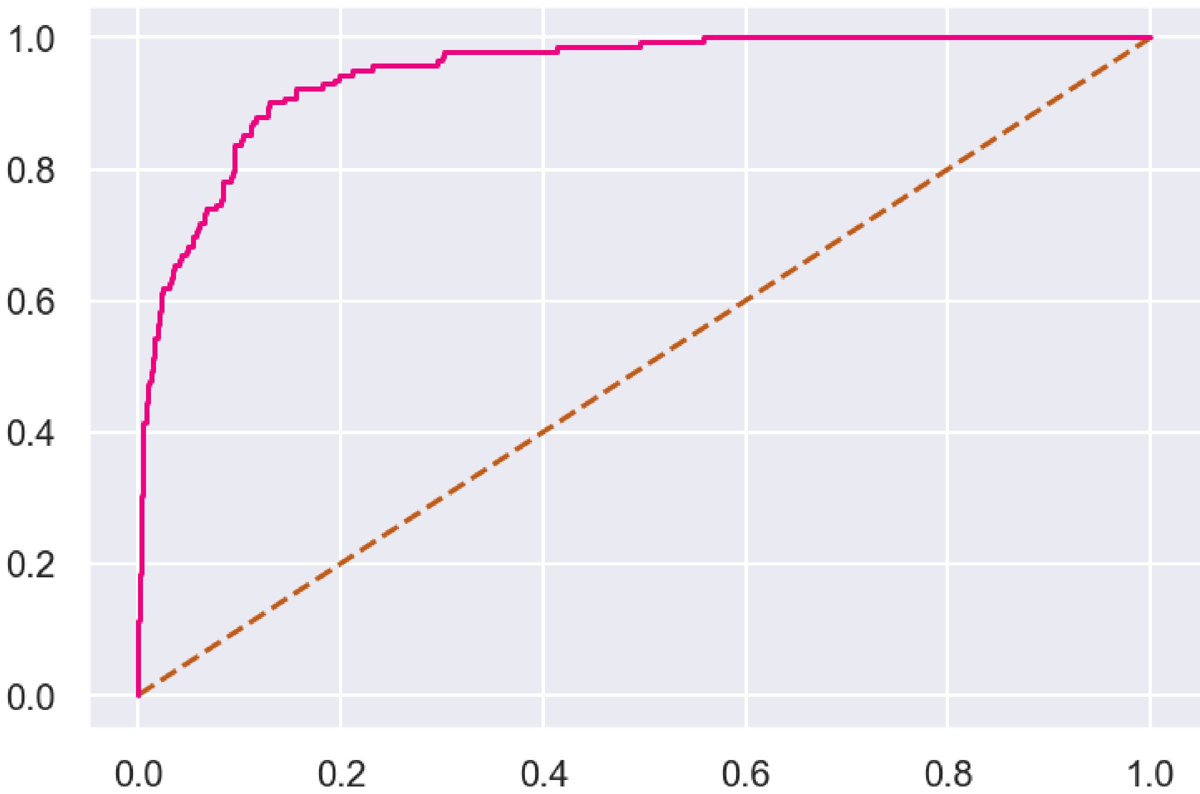


Figure 15. Test ROC Curve



2.5 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve)

Table 10 All Model Performance Comparison

	RECALL FOR 1 (in %)	PRECISION FOR 1 (in %)	ACCURACY (in %)	F-1 FOR 1 (in %)
Random Forest Model	88	95	98	91
LDA Model	88	49	87.5	63
Best LR: Model 8	95	78	96	86

2.6 State Recommendations from the above models

- Recall of 95% means - 95% of Actual Defaults were Predicted Correctly
- Precision of 78% means - 78% of Predicted Defaults were Actual
- For this modelling, we needed to predict as many of Actual Defaults as possible and minimise Type 2 errors foremost
- Hence Recall and then Precision was considered in choosing the best model
- In Table 10 above, coefficients of all variables indicate the weightage of that variable in predicting Default
- Positive coefficient means, if all else is equal, then higher value of this variable will lead to higher likelihood of default
- Negative coefficient means, if all else is equal, then higher value of this variable will lead to lower likelihood of Default
- For the above table, we can clearly see that th recall for logistic regression model is highest and the precision of the random forest model is higher as compared to other models.
- This means that majority of the random forest model predicted defaults, actually happened and the logistic regression model predicted correctly the actual defaults

Chapter 3. FRA Project (Milestone-2)

3.1 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.

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

Figure 16. Stock Price vs. Time: Shree Cement

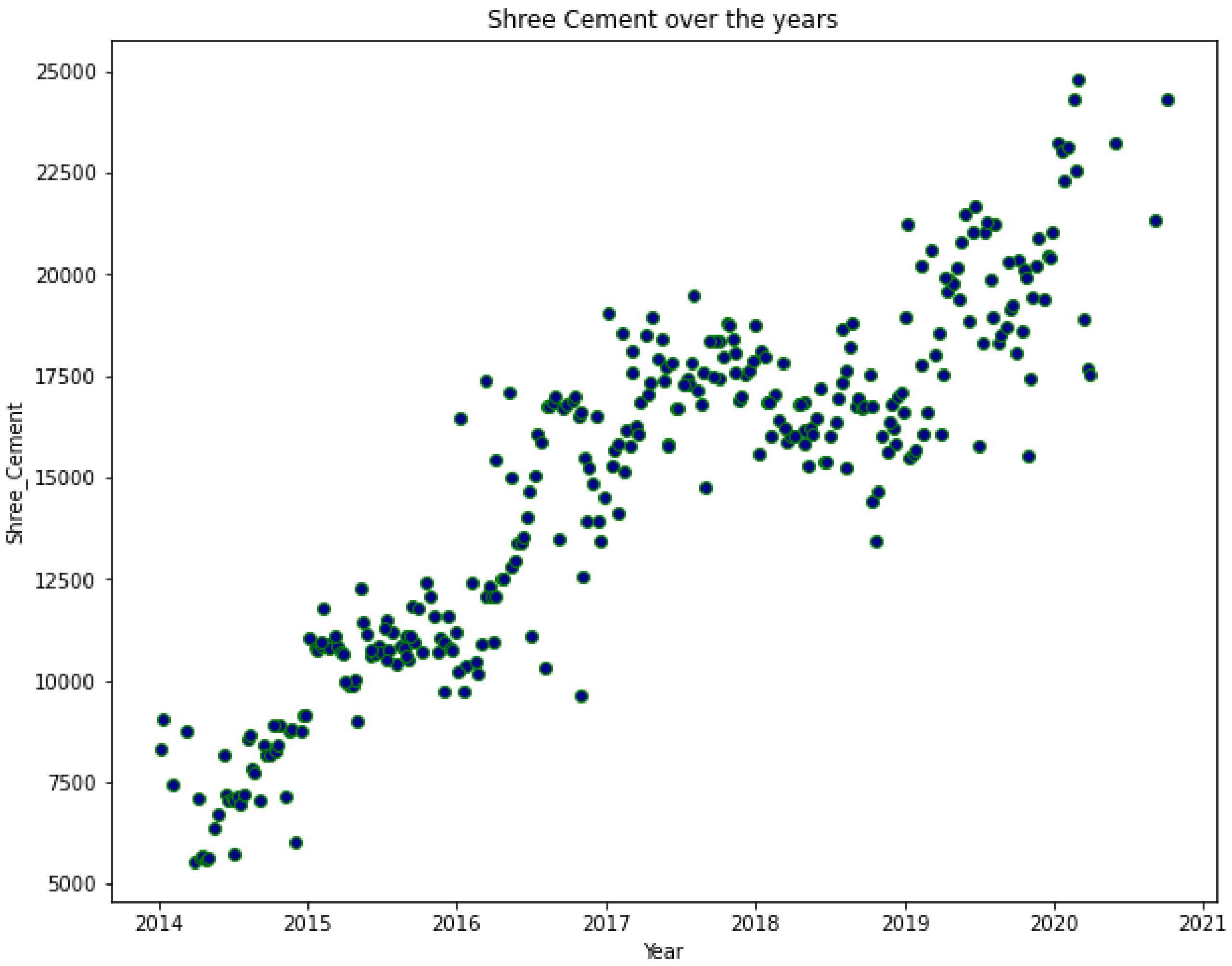
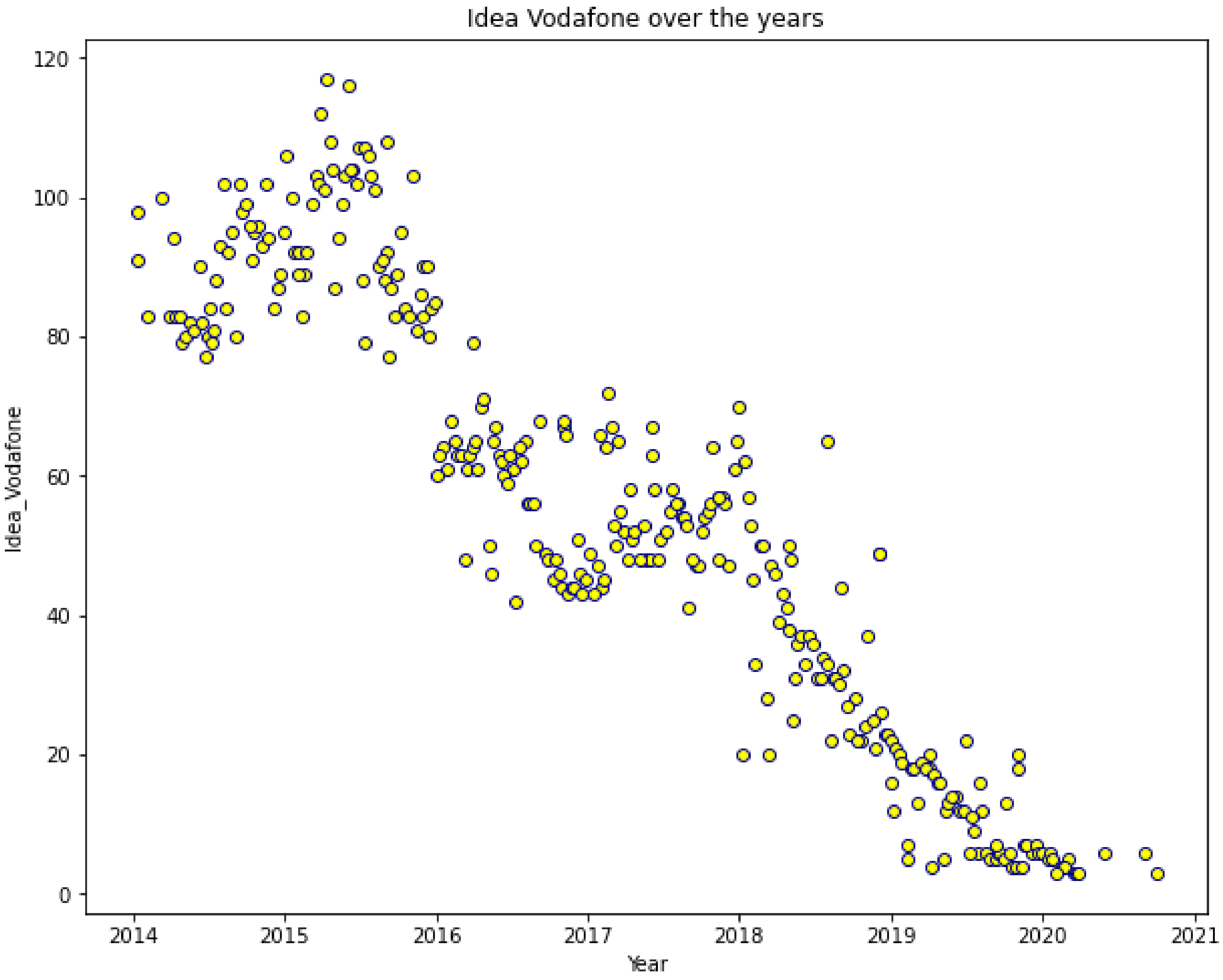


Figure 17. Stock Price vs. Time: Idea Vodafone



- Looking at the stock prices with the given timeline, Shree Cement prices have obviously increased over the period. There is a clear that there is a positive trend and we can also infere that the company might be doing well with growing stock prices.
- There was drop in the prices in mid-2018 to 2019; however, the Shree Cement stock prices started going up after first quarter of 2019
- The stock prices for Idea Vodafone were increasing from 2014 to mid-2015. However, the prices started to decline dramatically, tll the end of 2016. However, the prices started going up again till 2018, since then the prices have drastically declined to nearly zero or just above zero.

Back to Top

3.3 Calculate Returns for all stocks with inference

- We have calculated the returns for the all the stocks for the given period of time

Table 11 Stock Returns (Top 5)

	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways	Avg. Prices
0	NaN	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	0.032110
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943	-0.014989
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117	0.010306
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846	-0.024257

Table 12 Stock Returns (Last 5)

	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways	Avg. Prices
309	0.009649	-0.110348	0.030305	-0.057580	-0.087011	0.023688	0.072383	-0.053346	-0.287682	-0.127833	0.020528
310	-0.139625	-0.051293	-0.093819	-0.145324	-0.095310	-0.081183	-0.043319	-0.187816	0.693147	-0.200671	-0.084428
311	-0.094207	-0.236389	-0.285343	-0.284757	-0.105361	-0.119709	-0.050745	-0.141830	-0.693147	-0.117783	-0.125030
312	0.109856	-0.182322	-0.091269	-0.173019	-0.251314	-0.067732	-0.076851	-0.165324	0.000000	-0.133531	-0.066114
313	-0.017228	0.000000	-0.031198	0.051432	0.090972	-0.006816	0.040585	-0.081917	0.000000	0.000000	-0.005759

- The first row of the dataset after calculating retuns is NaN as we do not have stock prices for the companies before 31st March 2014, which is the first row in the dataset
- We have analyzed the means and volatility in the next question, wherein we will analyze the stock returns and compare them with the average prices of these stocks

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

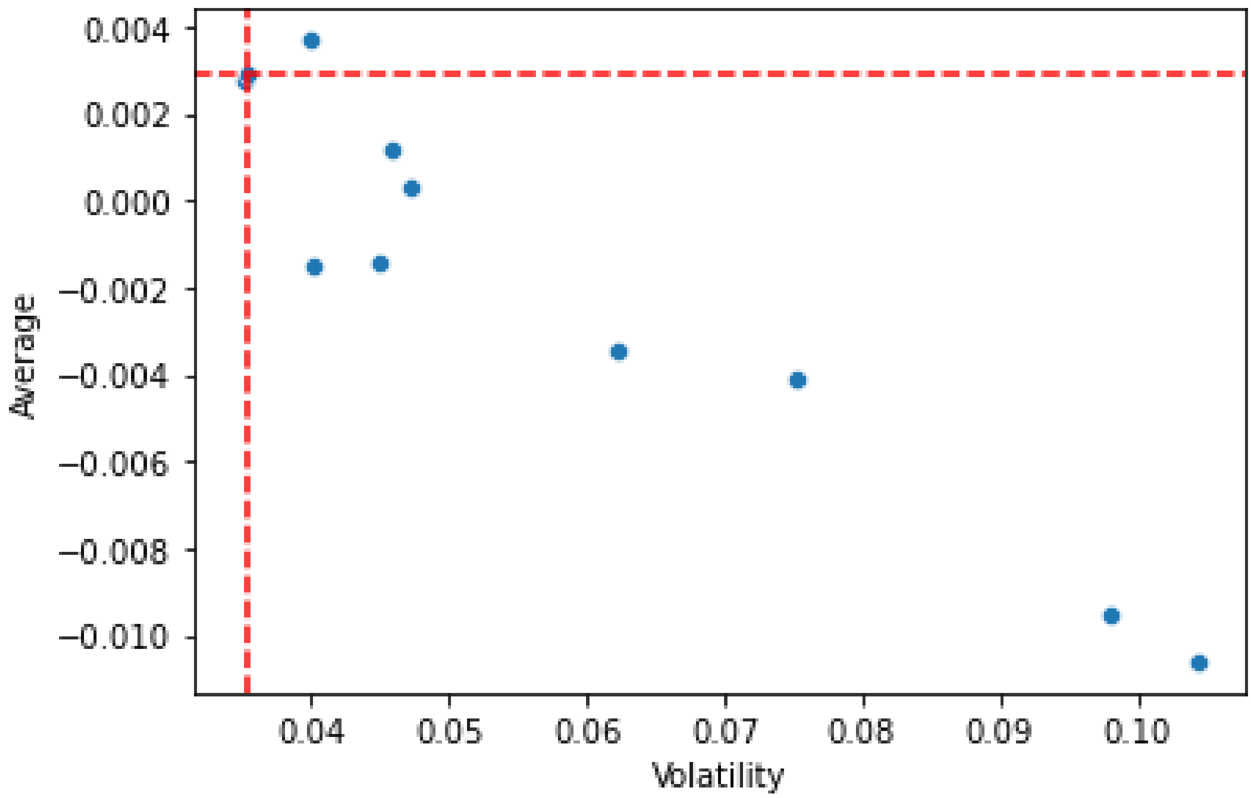
Table 13 Stock Returns: Means (Average) & Std. Deviation (Volatility)

	Average	Volatility
Infosys	0.002794	0.035070
Indian_Hotel	0.000266	0.047131
Mahindra_&_Mahindra	-0.001506	0.040169
Axis_Bank	0.001167	0.045828
SAIL	-0.003463	0.062188
Shree_Cement	0.003681	0.039917
Sun_Pharma	-0.001455	0.045033
Jindal_Steel	-0.004123	0.075108
Idea_Vodafone	-0.010608	0.104315
Jet_Airways	-0.009548	0.097972
Avg. Prices	0.002879	0.035459

- We can see that the average returns for majority of the stocks are negative. The volatility of Infosys and Shree Cement is lowest followed by Mahindra & Mahindra and Sun Pharma
- As all the stocks belong from different industries, we can club them together and analyze if particular industry stocks were doing well or not over the given period
- The mean of the average prices is 0.002879 and standard deviation is 0.35459

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

Figure 18. Stock Means (Average) vs. Standard Deviation (Volatility)



- Considering the means and volatility of the average prices as reference lines:
    - The volatility of Infosys and Shree Cement is extremely low and the average prices of the two stocks is close to the mean reference line of the average prices
    - Jet Airways and Idea Vodafone have high volatility and with lowest returns
    - The average returns for all the stocks are nearly negative, it could be because of the COVID-19 pandemic occurred during the end of 2019 to 2021

3.6 Conclusion and Recommendations

- Jet Airways and Idea Vodafone are two lowest performing stocks with low returns and high volatility
- The aforementioned two stocks with worst performance need to be removed from the investment list as it will most certainly incur losses against the investment
- Shree Cement is by far the best stock among the portfolio, which has been providing comparateively good returns with low volatility
- Infosys, India Hotel, and Axis Bank have positive returns on investment apart from Shree Cement. These stocks also show low volatility as compared to others and they hold lower risk than the ones with higher standard deviation and low returns
- Based on the analysis, Shree Cement is the best stock to buy as it doesn't hold a lot of risk and provide better returns as compared to others

The End