Credit Risk Business Report - 2

* **Name**: Hemant Patidar
* **Batch**: PGPDSBA Online Sep\_A 2021
* **Date**: 17/07/2022

Table of Contents

[**Credit Risk – Milestone 1** 7](#_Toc108957039)

[**Executive Summary** 7](#_Toc108957040)

[**Introduction** 7](#_Toc108957041)

[**Data Description** 7](#_Toc108957042)

[**1.1 Outlier Treatment** 11](#_Toc108957043)

[**Analysis / Visualization:** 11](#_Toc108957044)

[**Identification:** 12](#_Toc108957045)

[**Count and Percentage:** 13](#_Toc108957046)

[**1.2 Missing Value Treatment** 15](#_Toc108957047)

[**Value Proportion –** 15](#_Toc108957048)

[**Treatment –** 16](#_Toc108957049)

[**1.3 Transform Target variable into 0 and 1** 16](#_Toc108957050)

[**Value counts –** 16](#_Toc108957051)

[**1.4 Univariate & Bivariate analysis with proper interpretation.** 16](#_Toc108957052)

[**Correlation Map** 17](#_Toc108957053)

[**Finding Significant variables (with VIF)** 17](#_Toc108957054)

[**Remaining significant variables –** 19](#_Toc108957055)

[**Univariate Analysis** – 20](#_Toc108957056)

[**Co\_Code** 20](#_Toc108957057)

[**Equity\_Paid\_Up** 20](#_Toc108957058)

[**Book\_Value\_Adj\_Unit\_Curr** 21](#_Toc108957059)

[**ROG\_Total\_Assets\_perc** 21](#_Toc108957060)

[**Creditors\_Vel\_Days** 21](#_Toc108957061)

[**Value\_of\_Output\_to\_Gross\_Block** 22](#_Toc108957062)

[**Bi-variate Analysis** 22](#_Toc108957063)

[**Co\_Code vs. Default** 22](#_Toc108957064)

[**Equity\_Paid\_Up vs. Default** 23](#_Toc108957065)

[**Book\_Value\_Adj\_Unit\_Curr vs. Default** 23](#_Toc108957066)

[**ROG\_Total\_Assets\_perc vs. Default** 24](#_Toc108957067)

[**Creditors\_Vel\_Days vs. Default** 24](#_Toc108957068)

[**Value\_of\_Output\_to\_Gross\_Block vs. Default** 25](#_Toc108957069)

[**1.5 Train Test Split** 27](#_Toc108957070)

[**1.6 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** 28](#_Toc108957071)

[**Logit Formula 1:** 28](#_Toc108957072)

[**Model 1 Summary:** 28](#_Toc108957073)

[**Logit Formula 2:** (after removing above 2 features from analysis) 29](#_Toc108957074)

[**Model 2 Summary:** 29](#_Toc108957075)

[**Classification Report (with threshold probability 0.5)** 30](#_Toc108957076)

[**Optimal Threshold** 31](#_Toc108957077)

[**1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model** 31](#_Toc108957078)

[**Credit Risk – Milestone 2** 32](#_Toc108957079)

[**1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach** 32](#_Toc108957080)

[**RF Classification Report (Train Data)** 32](#_Toc108957081)

[**1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model** 32](#_Toc108957082)

[**RF Classification Report (Test Data)** 33](#_Toc108957083)

[**Comparison Matrix (RF)** 33](#_Toc108957084)

[**1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach** 33](#_Toc108957085)

[**LDA Classification Report (Train Data)** 34](#_Toc108957086)

[**LDA Classification Report with Optimal Threshold (Train Data)** 34](#_Toc108957087)

[**1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model** 34](#_Toc108957088)

[**LDA Classification Report with Optimal Threshold (Test Data)** 34](#_Toc108957089)

[**Comparison Matrix (LDA)** 35](#_Toc108957090)

[**1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve)** 35](#_Toc108957091)

[**ROC Curve** 35](#_Toc108957092)

[**ROC AUC Score** 36](#_Toc108957093)

[**Metrics Comparison Report** 36](#_Toc108957094)

[**1.13 State Recommendations from the above models** 37](#_Toc108957095)

[**Market Risk Analysis** 38](#_Toc108957096)

[**Executive Summary** 38](#_Toc108957097)

[**Introduction** 38](#_Toc108957098)

[**Exploratory Data Analysis** 38](#_Toc108957099)

[**2.1 Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference** 38](#_Toc108957100)

[**Infosys** 38](#_Toc108957101)

[**Shree Cement** 39](#_Toc108957102)

[**2.2 Calculate Returns for all stocks with inference** 40](#_Toc108957103)

[**Infosys** 41](#_Toc108957104)

[**Shree Cement** 41](#_Toc108957105)

[**2.3 Calculate Stock Means and Standard Deviation for all stocks with inference** 42](#_Toc108957106)

[**2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference** 43](#_Toc108957107)

[**Stock Means (Average Weekly Returns)** 43](#_Toc108957108)

[**Stock Standard Deviation (Volatility)** 44](#_Toc108957109)

[**2.5 Conclusion and Recommendations** 45](#_Toc108957110)

[**Conclusion** 45](#_Toc108957111)

[**Recommendations** 46](#_Toc108957112)

List of Figures

[Figure I - Boxplot Analysis for Outliers 12](#_Toc108957113)

[Figure II - Post Outlier Treatment 15](#_Toc108957114)

[Figure III - Correlation Map 17](#_Toc108957115)

[Figure IV - Logit Regression Summary 19](#_Toc108957116)

[Figure V - Data Head 20](#_Toc108957117)

[Figure VI - Co\_Code Univariate Analysis 20](#_Toc108957118)

[Figure VII - Equity\_Paid\_Up Univariate Analysis 21](#_Toc108957119)

[Figure VIII - Book\_Value\_Adj\_Unit\_Curr Univariate Analysis 21](#_Toc108957120)

[Figure IX - ROG\_Total\_Assets\_perc Univariate Analysis 21](#_Toc108957121)

[Figure X - Creditors\_Vel\_Days Univariate Analysis 22](#_Toc108957122)

[Figure XI - Value\_of\_Output\_to\_Gross\_Block Univariate Analysis 22](#_Toc108957123)

[Figure XII - Co\_Code Bivariate Analysis 23](#_Toc108957124)

[Figure XIII - Equity\_Paid\_Up Bivariate Analysis 23](#_Toc108957125)

[Figure XIV - Book\_Value\_Adj\_Unit\_Curr Bivariate Analysis 24](#_Toc108957126)

[Figure XV - ROG\_Total\_Assets\_perc Bivariate Analysis 24](#_Toc108957127)

[Figure XVI - Creditors\_Vel\_Days Bivariate Analysis 25](#_Toc108957128)

[Figure XVII - Value\_of\_Output\_to\_Gross\_Block Bivariate Analysis 25](#_Toc108957129)

[Figure XVIII - Pair Plot Analysis 26](#_Toc108957130)

[Figure XIX - Heat Map of Significant Features Remaining 27](#_Toc108957131)

[Figure XX - Logit Regression Result (MODEL 1) 29](#_Toc108957132)

[Figure XXI - Logit Regression Result (MODEL 2) 30](#_Toc108957133)

[Figure XXII - Classification Report (Train Data) 30](#_Toc108957134)

[Figure XXIII - Classification Report (Test Data) 30](#_Toc108957135)

[Figure XXIV - Classification Report Test Data (On Optimal Value) 31](#_Toc108957136)

[Figure XXV - RF Classification Report (Train Data) 32](#_Toc108957137)

[Figure XXVI - RF Classification Report (Test Data) 33](#_Toc108957138)

[Figure XXVII - LDA Classification Report (Train Data) 34](#_Toc108957139)

[Figure XXVIII - LDA Classification Report with Optimal Threshold (Train Data) 34](#_Toc108957140)

[Figure XXIX - LDA Classification Report with Optimal Threshold (Test Data) 34](#_Toc108957141)

[Figure XXX - ROC Curve (Train Data) 35](#_Toc108957142)

[Figure XXXI - ROC Curve (Test Data) 36](#_Toc108957143)

[Figure XXXII - Infosys Prices over the Years 39](#_Toc108957144)

[Figure XXXIII - Shree Cement Prices over the years 40](#_Toc108957145)

[Figure XXXIV - Weekly Returns (Infosys) 41](#_Toc108957146)

[Figure XXXV - Weekly Returns (Shree Cement) 42](#_Toc108957147)

[Figure XXXVI - Average Weekly Return 44](#_Toc108957148)

[Figure XXXVII – Volatility 45](#_Toc108957149)

List of Tables

[Table 1 - Data Information 8](#_Toc108957150)

[Table 2 - NULL values 9](#_Toc108957151)

[Table 3 - Data Description 11](#_Toc108957152)

[Table 4 - Outlier Proportion 14](#_Toc108957153)

[Table 5 - Missing Value Proportion 16](#_Toc108957154)

[Table 6 - Median Values 16](#_Toc108957155)

[Table 7 - VIF of Dropped Features 18](#_Toc108957156)

[Table 8 - Remaining Significant Variable 20](#_Toc108957157)

[Table 9 - Data Description 20](#_Toc108957158)

[Table 10 - Performance Comparison Metrics 31](#_Toc108957159)

[Table 11 - RF Comparison Matrix (RF) 33](#_Toc108957160)

[Table 12 - Comparison Matrix (LDA) 35](#_Toc108957161)

[Table 13 - ROC AUC Score 36](#_Toc108957162)

[Table 14 - Metrics Comparison Report 36](#_Toc108957163)

[Table 15 - Weekly Returns (Top 5 Rows) 40](#_Toc108957164)

[Table 16 -Stocks' Average Return (Weekly) 42](#_Toc108957165)

[Table 17 - Stocks' Volatility 43](#_Toc108957166)

# **Credit Risk – Milestone 1**

### **Executive Summary**

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 Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

### **Introduction**

Milestone-1 of the FRA project will cover the basic model building for predicting credit risk.

### **Data Description**

* Dataframe has 3586 records (rows) and 67 features (columns).
* Dataframe doesn’t have any duplicate.
* Dataframe has 66 numerical and 1 categorical variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Field Name** | **Non-Null Count** | **Data Type** |
| 0 | Co\_Code | 3586 | int64 |
| 1 | Co\_Name | 3586 | object |
| 2 | Networth\_Next\_Year | 3586 | float64 |
| 3 | Equity\_Paid\_Up | 3586 | float64 |
| 4 | Networth | 3586 | float64 |
| 5 | Capital\_Employed | 3586 | float64 |
| 6 | Total\_Debt | 3586 | float64 |
| 7 | Gross\_Block | 3586 | float64 |
| 8 | Net\_Working\_Capital | 3586 | float64 |
| 9 | Curr\_Assets | 3586 | float64 |
| 10 | Curr\_Liab\_and\_Prov | 3586 | float64 |
| 11 | Total\_Assets\_to\_Liab | 3586 | float64 |
| 12 | Gross\_Sales | 3586 | float64 |
| 13 | Net\_Sales | 3586 | float64 |
| 14 | Other\_Income | 3586 | float64 |
| 15 | Value\_Of\_Output | 3586 | float64 |
| 16 | Cost\_of\_Prod | 3586 | float64 |
| 17 | Selling\_Cost | 3586 | float64 |
| 18 | PBIDT | 3586 | float64 |
| 19 | PBDT | 3586 | float64 |
| 20 | PBIT | 3586 | float64 |
| 21 | PBT | 3586 | float64 |
| 22 | PAT | 3586 | float64 |
| 23 | Adjusted\_PAT | 3586 | float64 |
| 24 | CP | 3586 | float64 |
| 25 | Rev\_earn\_in\_forex | 3586 | float64 |
| 26 | Rev\_exp\_in\_forex | 3586 | float64 |
| 27 | Capital\_exp\_in\_forex | 3586 | float64 |
| 28 | Book\_Value\_Unit\_Curr | 3586 | float64 |
| 29 | Book\_Value\_Adj\_Unit\_Curr | 3582 | float64 |
| 30 | Market\_Capitalisation | 3586 | float64 |
| 31 | CEPS\_annualised\_Unit\_Curr | 3586 | float64 |
| 32 | Cash\_Flow\_From\_Opr | 3586 | float64 |
| 33 | Cash\_Flow\_From\_Inv | 3586 | float64 |
| 34 | Cash\_Flow\_From\_Fin | 3586 | float64 |
| 35 | ROG\_Net\_Worth\_perc | 3586 | float64 |
| 36 | ROG\_Capital\_Employed\_perc | 3586 | float64 |
| 37 | ROG\_Gross\_Block\_perc | 3586 | float64 |
| 38 | ROG\_Gross\_Sales\_perc | 3586 | float64 |
| 39 | ROG\_Net\_Sales\_perc | 3586 | float64 |
| 40 | ROG\_Cost\_of\_Prod\_perc | 3586 | float64 |
| 41 | ROG\_Total\_Assets\_perc | 3586 | float64 |
| 42 | ROG\_PBIDT\_perc | 3586 | float64 |
| 43 | ROG\_PBDT\_perc | 3586 | float64 |
| 44 | ROG\_PBIT\_perc | 3586 | float64 |
| 45 | ROG\_PBT\_perc | 3586 | float64 |
| 46 | ROG\_PAT\_perc | 3586 | float64 |
| 47 | ROG\_CP\_perc | 3586 | float64 |
| 48 | ROG\_Rev\_earn\_in\_forex\_perc | 3586 | float64 |
| 49 | ROG\_Rev\_exp\_in\_forex\_perc | 3586 | float64 |
| 50 | ROG\_Market\_Capitalisation\_perc | 3586 | float64 |
| 51 | Curr\_Ratio\_Latest | 3585 | float64 |
| 52 | Fixed\_Assets\_Ratio\_Latest | 3585 | float64 |
| 53 | Inventory\_Ratio\_Latest | 3585 | float64 |
| 54 | Debtors\_Ratio\_Latest | 3585 | float64 |
| 55 | Total\_Asset\_Turnover\_Ratio\_Latest | 3585 | float64 |
| 56 | Interest\_Cover\_Ratio\_Latest | 3585 | float64 |
| 57 | PBIDTM\_perc\_Latest | 3585 | float64 |
| 58 | PBITM\_perc\_Latest | 3585 | float64 |
| 59 | PBDTM\_perc\_Latest | 3585 | float64 |
| 60 | CPM\_perc\_Latest | 3585 | float64 |
| 61 | APATM\_perc\_Latest | 3585 | float64 |
| 62 | Debtors\_Vel\_Days | 3586 | int64 |
| 63 | Creditors\_Vel\_Days | 3586 | int64 |
| 64 | Inventory\_Vel\_Days | 3483 | float64 |
| 65 | Value\_of\_Output\_to\_Total\_Assets | 3586 | float64 |
| 66 | Value\_of\_Output\_to\_Gross\_Block | 3586 | float64 |

Table - Data Information

|  |  |
| --- | --- |
| **Field Name** | **NULL Values** |
| Book\_Value\_Adj\_Unit\_Curr | 4 |
| Curr\_Ratio\_Latest | 1 |
| Fixed\_Assets\_Ratio\_Latest | 1 |
| Inventory\_Ratio\_Latest | 1 |
| Debtors\_Ratio\_Latest | 1 |
| Total\_Asset\_Turnover\_Ratio\_Latest | 1 |
| Interest\_Cover\_Ratio\_Latest | 1 |
| PBIDTM\_perc\_Latest | 1 |
| PBITM\_perc\_Latest | 1 |
| PBDTM\_perc\_Latest | 1 |
| CPM\_perc\_Latest | 1 |
| APATM\_perc\_Latest | 1 |
| Inventory\_Vel\_Days | 103 |

Table - NULL values

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| Co\_Code | 3,586 | 16,065.39 | 19,776.82 | 4.00 | 3,029.25 | 6,077.50 | 24,269.50 | 72,493.00 |
| Networth\_Next\_Year | 3,586 | 725.05 | 4,769.68 | - 8,021.60 | 3.99 | 19.02 | 123.80 | 111,729.10 |
| Equity\_Paid\_Up | 3,586 | 62.97 | 778.76 | - | 3.75 | 8.29 | 19.52 | 42,263.46 |
| Networth | 3,586 | 649.75 | 4,091.99 | - 7,027.48 | 3.89 | 18.58 | 117.30 | 81,657.35 |
| Capital\_Employed | 3,586 | 2,799.61 | 26,975.14 | - 1,824.75 | 7.60 | 39.09 | 226.61 | 714,001.25 |
| Total\_Debt | 3,586 | 1,994.82 | 23,652.84 | - 0.72 | 0.03 | 7.49 | 72.35 | 652,823.81 |
| Gross\_Block | 3,586 | 594.18 | 4,871.55 | - 41.19 | 0.57 | 15.87 | 131.90 | 128,477.59 |
| Net\_Working\_Capital | 3,586 | 410.81 | 6,301.22 | - 13,162.42 | 0.94 | 10.15 | 61.18 | 223,257.56 |
| Curr\_Assets | 3,586 | 1,960.35 | 22,577.57 | - 0.91 | 4.00 | 24.54 | 135.28 | 721,166.00 |
| Curr\_Liab\_and\_Prov | 3,586 | 391.99 | 2,675.00 | - 0.23 | 0.73 | 9.23 | 65.65 | 83,232.98 |
| Total\_Assets\_to\_Liab | 3,586 | 1,778.45 | 11,437.57 | - 4.51 | 10.56 | 52.01 | 310.54 | 254,737.22 |
| Gross\_Sales | 3,586 | 1,123.74 | 10,603.70 | - 62.59 | 1.44 | 31.21 | 242.25 | 474,182.94 |
| Net\_Sales | 3,586 | 1,079.70 | 9,996.57 | - 62.59 | 1.44 | 30.44 | 234.44 | 443,775.16 |
| Other\_Income | 3,586 | 48.73 | 426.04 | - 448.72 | 0.02 | 0.45 | 3.64 | 14,143.40 |
| Value\_Of\_Output | 3,586 | 1,077.19 | 9,843.88 | - 119.10 | 1.41 | 30.90 | 235.84 | 435,559.09 |
| Cost\_of\_Prod | 3,586 | 798.54 | 9,076.70 | - 22.65 | 0.94 | 25.99 | 189.55 | 419,913.50 |
| Selling\_Cost | 3,586 | 25.55 | 194.24 | - | - | 0.16 | 3.88 | 5,283.91 |
| PBIDT | 3,586 | 248.18 | 1,949.59 | - 4,655.14 | 0.04 | 2.05 | 23.53 | 42,059.26 |
| PBDT | 3,586 | 116.27 | 956.20 | - 5,874.53 | - | 0.80 | 12.95 | 23,215.00 |
| PBIT | 3,586 | 217.66 | 1,850.97 | - 4,812.95 | - | 1.15 | 16.67 | 41,402.96 |
| PBT | 3,586 | 85.75 | 799.93 | - 6,032.34 | - 0.06 | 0.31 | 7.42 | 16,798.00 |
| PAT | 3,586 | 61.22 | 620.30 | - 6,032.34 | - 0.06 | 0.26 | 5.54 | 13,383.39 |
| Adjusted\_PAT | 3,586 | 60.06 | 580.43 | - 4,418.72 | - 0.09 | 0.21 | 5.34 | 13,384.11 |
| CP | 3,586 | 91.73 | 780.79 | - 5,874.53 | - | 0.74 | 10.91 | 20,760.20 |
| Rev\_earn\_in\_forex | 3,586 | 131.17 | 1,150.73 | - | - | - | 7.20 | 46,158.00 |
| Rev\_exp\_in\_forex | 3,586 | 256.33 | 4,132.34 | - | - | - | 6.99 | 193,979.73 |
| Capital\_exp\_in\_forex | 3,586 | 7.66 | 111.43 | - | - | - | - | 3,722.10 |
| Book\_Value\_Unit\_Curr | 3,586 | 157.24 | 1,622.66 | - 3,371.57 | 7.96 | 21.67 | 71.67 | 75,790.00 |
| Book\_Value\_Adj\_Unit\_Curr | 3,582 | 2,243.15 | 128,283.73 | - 33,715.70 | 7.06 | 18.93 | 60.01 | 7,677,600.29 |
| Market\_Capitalisation | 3,586 | 1,664.09 | 12,805.17 | - | - | 8.37 | 111.46 | 260,865.08 |
| CEPS\_annualised\_Unit\_Curr | 3,586 | 36.02 | 828.42 | - 1,808.00 | - | 1.15 | 8.77 | 45,438.44 |
| Cash\_Flow\_From\_Opr | 3,586 | 65.77 | 1,455.05 | - 25,469.23 | - 0.31 | 0.45 | 12.65 | 44,529.40 |
| Cash\_Flow\_From\_Inv | 3,586 | - 60.87 | 701.97 | - 23,843.45 | - 5.12 | - 0.12 | 0.12 | 3,732.98 |
| Cash\_Flow\_From\_Fin | 3,586 | 11.44 | 1,272.26 | - 38,374.04 | - 5.85 | - | 0.46 | 28,846.00 |
| ROG\_Net\_Worth\_perc | 3,586 | 1,237.62 | 41,041.93 | - 14,485.71 | - 1.49 | 1.84 | 11.36 | 2,144,020.00 |
| ROG\_Capital\_Employed\_perc | 3,586 | 2,988.88 | 126,472.87 | - 8,614.63 | - 3.84 | 1.38 | 12.59 | 7,412,700.00 |
| ROG\_Gross\_Block\_perc | 3,586 | 37.55 | 893.62 | - 116.12 | - | 0.25 | 6.72 | 47,400.00 |
| ROG\_Gross\_Sales\_perc | 3,586 | 242.67 | 6,103.53 | - 5,503.70 | - 8.08 | 3.31 | 21.53 | 320,200.00 |
| ROG\_Net\_Sales\_perc | 3,586 | 242.59 | 6,103.49 | - 5,503.70 | - 8.12 | 3.21 | 21.57 | 320,200.00 |
| ROG\_Cost\_of\_Prod\_perc | 3,586 | 310.49 | 5,573.22 | - 2,130.23 | - 7.24 | 4.42 | 23.12 | 267,150.00 |
| ROG\_Total\_Assets\_perc | 3,586 | 2,793.28 | 125,941.65 | - 136.13 | - 3.97 | 1.48 | 12.50 | 7,422,120.00 |
| ROG\_PBIDT\_perc | 3,586 | 375.85 | 23,278.40 | - 52,200.00 | - 23.36 | 4.57 | 47.88 | 1,386,200.00 |
| ROG\_PBDT\_perc | 3,586 | 336.38 | 20,353.40 | - 52,200.00 | - 30.60 | 3.37 | 52.92 | 1,208,700.00 |
| ROG\_PBIT\_perc | 3,586 | 374.70 | 22,462.79 | - 58,500.00 | - 31.35 | 2.13 | 50.14 | 1,338,000.00 |
| ROG\_PBT\_perc | 3,586 | 224.07 | 19,659.23 | - 78,900.00 | - 41.24 | 0.03 | 61.96 | 1,160,500.00 |
| ROG\_PAT\_perc | 3,586 | 112.23 | 13,480.52 | - 114,500.00 | - 43.73 | - | 65.35 | 774,200.00 |
| ROG\_CP\_perc | 3,586 | 221.09 | 13,980.20 | - 52,200.00 | - 29.51 | 4.62 | 52.91 | 822,400.00 |
| ROG\_Rev\_earn\_in\_forex\_perc | 3,586 | 37.23 | 658.67 | - 100.00 | - | - | - | 29,084.77 |
| ROG\_Rev\_exp\_in\_forex\_perc | 3,586 | 364.86 | 15,233.64 | - 100.00 | - | - | - | 894,591.69 |
| ROG\_Market\_Capitalisation\_perc | 3,586 | 63.68 | 1,047.93 | - 98.05 | - | - | 47.52 | 61,865.26 |
| Curr\_Ratio\_Latest | 3,585 | 12.06 | 108.41 | - | 0.88 | 1.36 | 2.77 | 4,813.00 |
| Fixed\_Assets\_Ratio\_Latest | 3,585 | 51.54 | 681.15 | - | 0.27 | 1.56 | 4.74 | 22,172.00 |
| Inventory\_Ratio\_Latest | 3,585 | 37.80 | 458.19 | - | - | 3.56 | 8.94 | 15,472.00 |
| Debtors\_Ratio\_Latest | 3,585 | 33.03 | 489.56 | - | 0.42 | 3.82 | 8.52 | 22,992.67 |
| Total\_Asset\_Turnover\_Ratio\_Latest | 3,585 | 1.24 | 2.67 | - | 0.07 | 0.60 | 1.55 | 57.75 |
| Interest\_Cover\_Ratio\_Latest | 3,585 | 16.39 | 351.74 | - 5,450.00 | - | 1.08 | 3.71 | 18,639.40 |
| PBIDTM\_perc\_Latest | 3,585 | - 51.16 | 1,795.13 | - 78,870.45 | - | 8.07 | 18.99 | 19,233.33 |
| PBITM\_perc\_Latest | 3,585 | - 109.21 | 3,057.64 | - 141,600.00 | - | 5.23 | 14.29 | 19,195.70 |
| PBDTM\_perc\_Latest | 3,585 | - 311.57 | 10,921.59 | - 590,500.00 | - | 4.69 | 14.11 | 15,640.00 |
| CPM\_perc\_Latest | 3,585 | - 307.01 | 10,676.15 | - 572,000.00 | - | 3.89 | 11.39 | 15,640.00 |
| APATM\_perc\_Latest | 3,585 | - 365.06 | 12,500.05 | - 688,600.00 | - | 1.59 | 7.41 | 15,266.67 |
| Debtors\_Vel\_Days | 3,586 | 603.89 | 10,636.76 | - | 8.00 | 49.00 | 106.00 | 514,721.00 |
| Creditors\_Vel\_Days | 3,586 | 2,057.85 | 54,169.48 | - | 8.00 | 39.00 | 89.00 | 2,034,145.00 |
| Inventory\_Vel\_Days | 3,483 | 79.64 | 137.85 | - 199.00 | - | 35.00 | 96.00 | 996.00 |
| Value\_of\_Output\_to\_Total\_Assets | 3,586 | 0.82 | 1.20 | - 0.33 | 0.07 | 0.48 | 1.16 | 17.63 |
| Value\_of\_Output\_to\_Gross\_Block | 3,586 | 61.88 | 976.82 | - 61.00 | 0.27 | 1.53 | 4.91 | 43,404.00 |

Table - Data Description

## **1.1 Outlier Treatment**

### **Analysis / Visualization:**

BOXPLOT helps in visualizing outliers present in numerical variables.

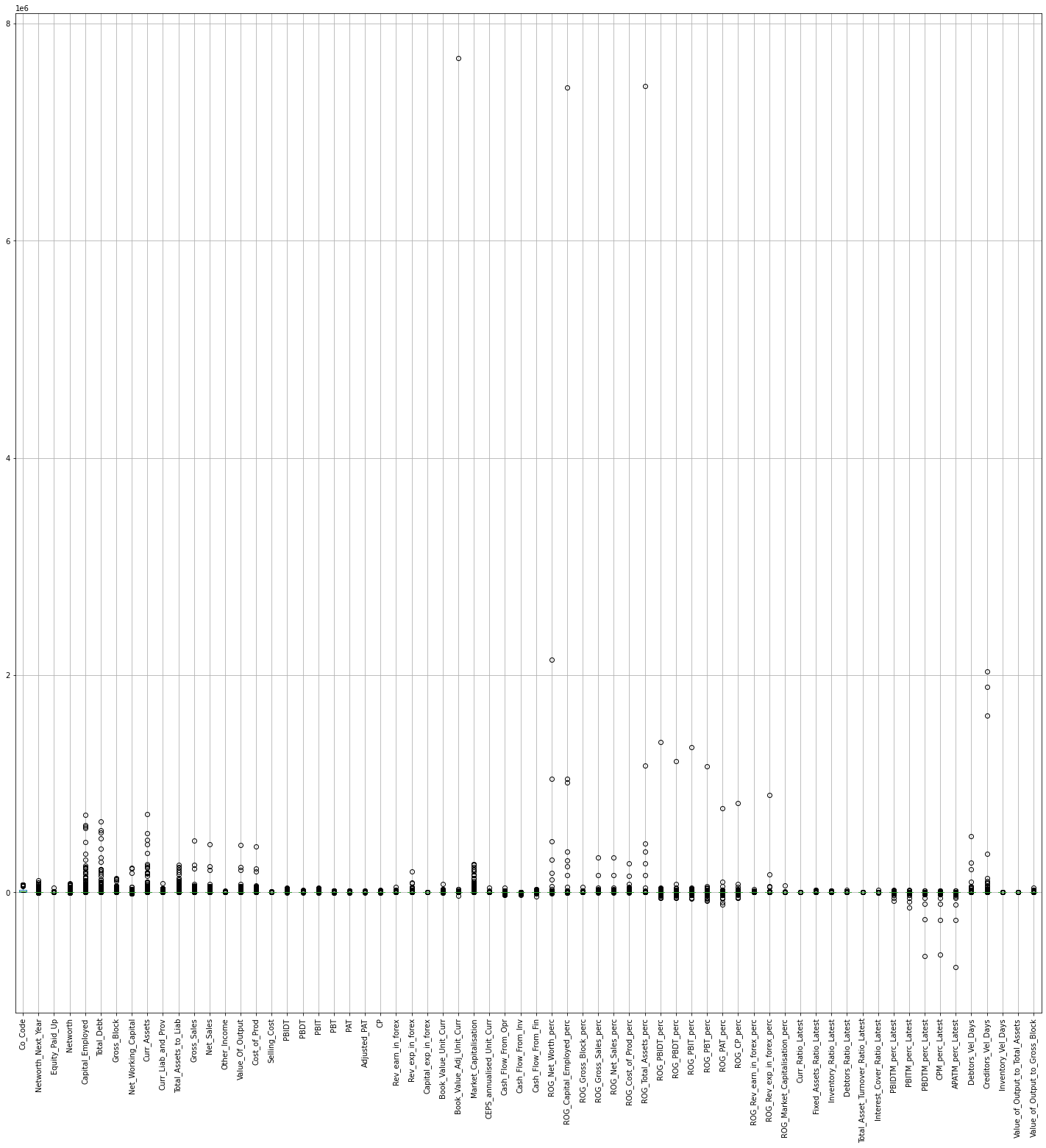


Figure - Boxplot Analysis for Outliers

By above figure, we can see that there are number of outliers present in the data.

### **Identification:**

If any value falls past threshold **Q1 - 1.5 IQR** (for lower values) or **Q3 + 1.5 IQR** (for upper values)

Where, Q1 – Median at 1st quartile (25%)

Q3 – Median at 3rd quartile (75%)

IQR – Inter quartile range (Q3 – Q1)

### **Count and Percentage:**

|  |  |  |
| --- | --- | --- |
| **Field Name** | **# Of Outliers** | **% Outliers in Data** |
| ROG\_Rev\_exp\_in\_forex\_perc | 1615 | 45.04% |
| ROG\_Rev\_earn\_in\_forex\_perc | 1317 | 36.73% |
| Cash\_Flow\_From\_Fin | 1005 | 28.03% |
| PAT | 959 | 26.74% |
| Adjusted\_PAT | 954 | 26.60% |
| PBT | 941 | 26.24% |
| APATM\_perc\_Latest | 933 | 26.03% |
| Cash\_Flow\_From\_Inv | 876 | 24.43% |
| ROG\_Gross\_Block\_perc | 830 | 23.15% |
| CP | 816 | 22.76% |
| PBDT | 815 | 22.73% |
| Cash\_Flow\_From\_Opr | 801 | 22.34% |
| ROG\_Net\_Worth\_perc | 747 | 20.83% |
| Rev\_earn\_in\_forex | 738 | 20.58% |
| Interest\_Cover\_Ratio\_Latest | 725 | 20.22% |
| PBIT | 720 | 20.08% |
| CPM\_perc\_Latest | 720 | 20.08% |
| PBITM\_perc\_Latest | 717 | 20% |
| PBDTM\_perc\_Latest | 695 | 19.39% |
| Capital\_exp\_in\_forex | 694 | 19.35% |
| Rev\_exp\_in\_forex | 693 | 19.33% |
| Networth\_Next\_Year | 676 | 18.85% |
| ROG\_Cost\_of\_Prod\_perc | 675 | 18.82% |
| PBIDT | 671 | 18.71% |
| ROG\_Gross\_Sales\_perc | 671 | 18.71% |
| ROG\_Net\_Sales\_perc | 667 | 18.60% |
| Networth | 650 | 18.13% |
| Market\_Capitalisation | 639 | 17.82% |
| ROG\_CP\_perc | 637 | 17.76% |
| ROG\_PBDT\_perc | 628 | 17.51% |
| Net\_Working\_Capital | 625 | 17.43% |
| ROG\_PBIT\_perc | 616 | 17.18% |
| ROG\_PBIDT\_perc | 611 | 17.04% |
| ROG\_PBT\_perc | 611 | 17.04% |
| Selling\_Cost | 605 | 16.87% |
| Other\_Income | 603 | 16.82% |
| CEPS\_annualised\_Unit\_Curr | 602 | 16.79% |
| ROG\_PAT\_perc | 598 | 16.68% |
| Capital\_Employed | 596 | 16.62% |
| PBIDTM\_perc\_Latest | 595 | 16.60% |
| Total\_Debt | 583 | 16.26% |
| Curr\_Liab\_and\_Prov | 581 | 16.20% |
| Curr\_Assets | 577 | 16.09% |
| Total\_Assets\_to\_Liab | 574 | 16.01% |
| ROG\_Capital\_Employed\_perc | 572 | 15.95% |
| Curr\_Ratio\_Latest | 565 | 15.76% |
| Cost\_of\_Prod | 560 | 15.62% |
| Value\_Of\_Output | 559 | 15.59% |
| Net\_Sales | 556 | 15.50% |
| Gross\_Sales | 554 | 15.45% |
| Gross\_Block | 540 | 15.06% |
| ROG\_Market\_Capitalisation\_perc | 497 | 13.86% |
| Fixed\_Assets\_Ratio\_Latest | 495 | 13.81% |
| Book\_Value\_Adj\_Unit\_Curr | 486 | 13.57% |
| Book\_Value\_Unit\_Curr | 485 | 13.52% |
| ROG\_Total\_Assets\_perc | 483 | 13.47% |
| Value\_of\_Output\_to\_Gross\_Block | 481 | 13.41% |
| Equity\_Paid\_Up | 448 | 12.49% |
| Debtors\_Vel\_Days | 398 | 11.10% |
| Creditors\_Vel\_Days | 391 | 10.90% |
| Inventory\_Ratio\_Latest | 375 | 10.46% |
| Debtors\_Ratio\_Latest | 371 | 10.35% |
| Co\_Code | 291 | 8.11% |
| Inventory\_Vel\_Days | 262 | 7.52% |
| Total\_Asset\_Turnover\_Ratio\_Latest | 201 | 5.61% |
| Value\_of\_Output\_to\_Total\_Assets | 150 | 4.18% |

Table - Outlier Proportion

There are,

* 5-10% outliers in 4 variables
* 10-15% outliers in 11 variables
* 15-20% outliers in 32 variables
* More than 20% outliers in 18 variables

The last 2 buckets have significant % of outliers, and all these features should not be a part model, as they might give biased result.

And rest can be fixed by setting threshold value (Q1 – 1.5 IQR) for lower values and (Q3 + 1.5 IQR) for upper values.

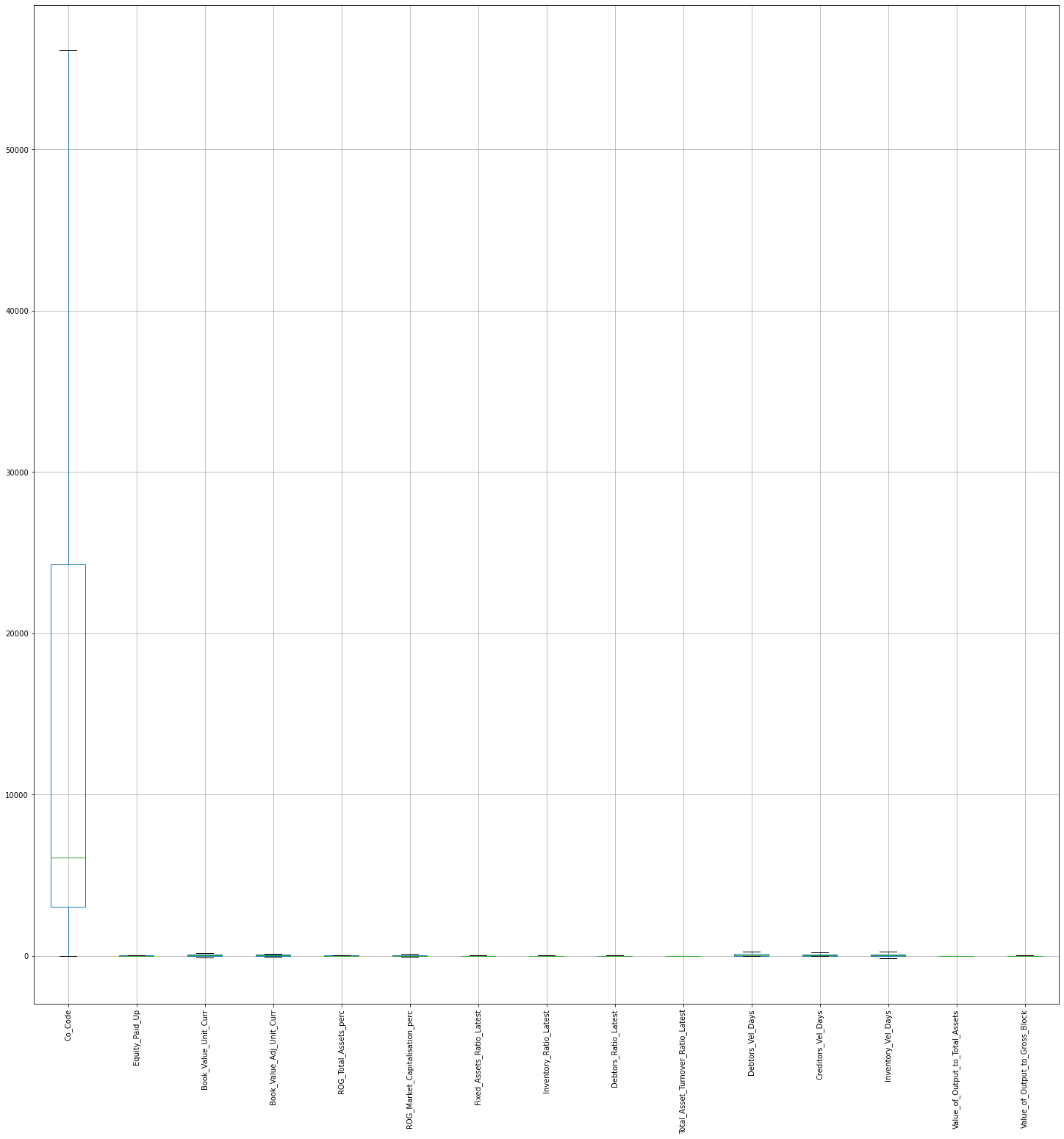


Figure - Post Outlier Treatment

## **1.2 Missing Value Treatment**

There are still 111 missing values within the database after removing 40 variables having more than 15%.

### **Value Proportion –**

Proportion of missing values help us to decide if the feature can still have significant data or not.

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Missing Values** | **% Missing in Data** |
| Book\_Value\_Adj\_Unit\_Curr | 4 | 0.11% |
| Fixed\_Assets\_Ratio\_Latest | 1 | 0.03% |
| Inventory\_Ratio\_Latest | 1 | 0.03% |
| Debtors\_Ratio\_Latest | 1 | 0.03% |
| Total\_Asset\_Turnover\_Ratio\_Latest | 1 | 0.03% |
| Inventory\_Vel\_Days | 103 | 2.96% |

Table - Missing Value Proportion

Investor\_Vel\_Days has 103 missing values which is ~3% of all data, and rest fields have as minimum as only 1 value missing…

### **Treatment –**

Since all the features are numerical and having outliers treated, we can either impute missing values for mean or median.

Median values on features that are missing values –

|  |  |
| --- | --- |
| **Field Name** | **Median Values** |
| Book\_Value\_Adj\_Unit\_Curr | 18.925 |
| Fixed\_Assets\_Ratio\_Latest | 1.56 |
| Inventory\_Ratio\_Latest | 3.56 |
| Debtors\_Ratio\_Latest | 3.82 |
| Total\_Asset\_Turnover\_Ratio\_Latest | 0.6 |
| Inventory\_Vel\_Days | 35 |

Table - Median Values

Missing values were imputed with above means on these features.

## **1.3 Transform Target variable into 0 and 1**

**Default** variable can be created as value of 1 when net worth next year is negative & 0 when net worth next year is positive.

### **Value counts –**

Default (0) 🡺 3198 (89.18%)

Default (1) 🡺 388 (10.81%)

## **1.4 Univariate & Bivariate analysis with proper interpretation.**

### **Correlation Map**

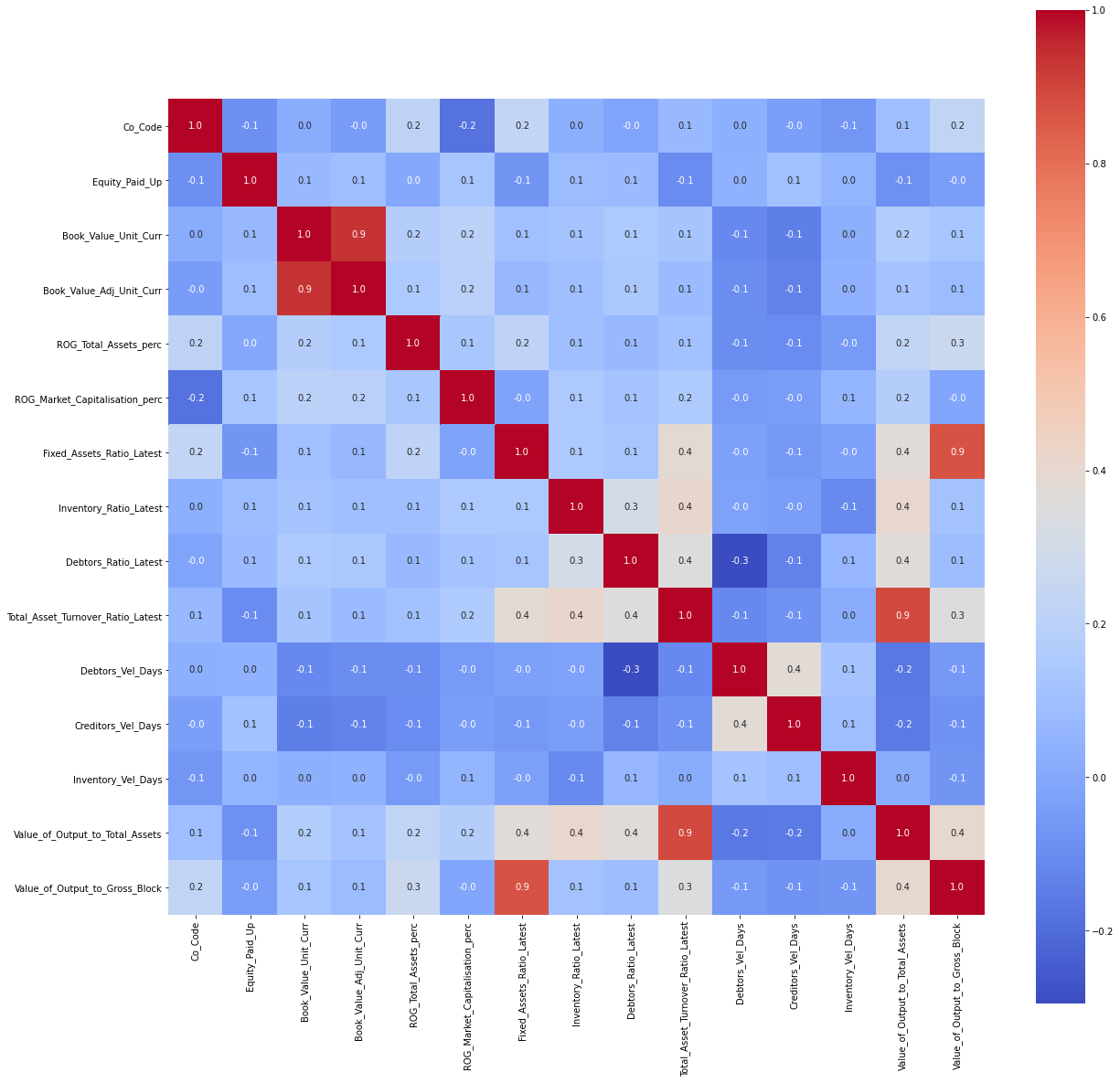


Figure - Correlation Map

By above correlation plot (heat map), we can infer –

* The blue blocks represent minimal correlation on the fields with each other
* The red blocks represent significant correlation on the fields with each other

### **Finding Significant variables (with VIF)**

**Variance inflation factor (VIF)** is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

We will be dropped any feature having VIF greater than 5, and the imputation would be recursive so that if one feature was dropped the effective VIF will be taken into consideration.

Dropped feature in order of recursive VIF being calculated –

|  |  |
| --- | --- |
| **Field Name** | **VIF** |
| Book\_Value\_Unit\_Curr | 13.59 |
| Value\_of\_Output\_to\_Total\_Assets | 11.65 |
| Fixed\_Assets\_Ratio\_Latest | 7.58 |

Table - VIF of Dropped Features

We are left with 13 features after eliminating features having more than 15% outliers (40) and VIF greater than 5.

We will still see when we build logit function, will it provide us more features that are insignificant.

**Logit Formula (f\_1)** == Default ~ Co\_Code + Equity\_Paid\_Up + Book\_Value\_Adj\_Unit\_Curr + ROG\_Total\_Assets\_perc + ROG\_Market\_Capitalisation\_perc + Inventory\_Ratio\_Latest + Debtors\_Ratio\_Latest + Total\_Asset\_Turnover\_Ratio\_Latest + Debtors\_Vel\_Days + Creditors\_Vel\_Days + Inventory\_Vel\_Days + Value\_of\_Output\_to\_Gross\_Block

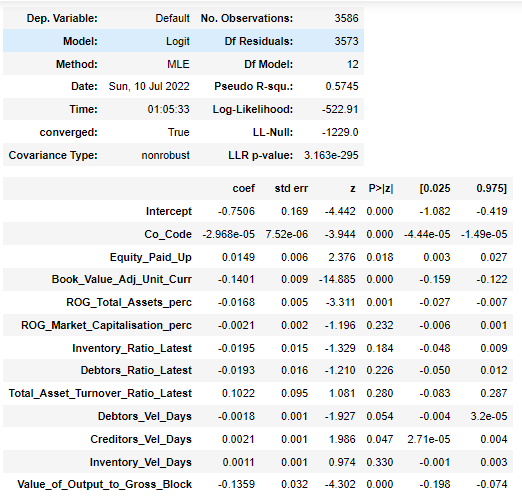


Figure - Logit Regression Summary

We can identify insignificant variables which are having P > 5 %

Hence, below features are insignificant –

* ROG\_Market\_Capitalisation\_perc
* Inventory\_Ratio\_Latest
* Debtors\_Ratio\_Latest
* Total\_Asset\_Turnover\_Ratio\_Latest
* Debtors\_Vel\_Days
* Inventory\_Vel\_Days

### **Remaining significant variables –**

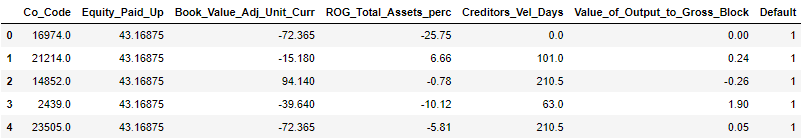


Figure - Data Head

|  |
| --- |
| Co\_Code |
| Equity\_Paid\_Up |
| Book\_Value\_Adj\_Unit\_Curr |
| ROG\_Total\_Assets\_perc |
| Creditors\_Vel\_Days |
| Value\_of\_Output\_to\_Gross\_Block |
| Default |

Table - Remaining Significant Variable

### **Univariate Analysis** –

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **count** | **mean** | **std** | **min** | **0.25** | **0.5** | **0.75** | **max** |
| Co\_Code | 3586 | 15036.73 | 17147.88 | 4 | 3029.25 | 6077.5 | 24269.5 | 56129.87 |
| Equity\_Paid\_Up | 3586 | 13.99 | 14.00 | 0 | 3.75 | 8.29 | 19.51 | 43.16 |
| Book\_Value\_Adj\_Unit\_Curr | 3586 | 38.13 | 50.12 | -72.36 | 7.065 | 18.92 | 59.96 | 139.43 |
| ROG\_Total\_Assets\_perc | 3586 | 4.27 | 16.37 | -28.68 | -3.97 | 1.47 | 12.5 | 37.20 |
| Creditors\_Vel\_Days | 3586 | 62.44 | 68.14 | 0 | 8 | 39 | 89 | 210.5 |
| Value\_of\_Output\_to\_Gross\_Block | 3586 | 3.35 | 4.10 | -6.69 | 0.27 | 1.53 | 4.91 | 11.87 |
| Default | 3586 | 0.10 | 0.31 | 0 | 0 | 0 | 0 | 1 |

Table - Data Description

### **Co\_Code**

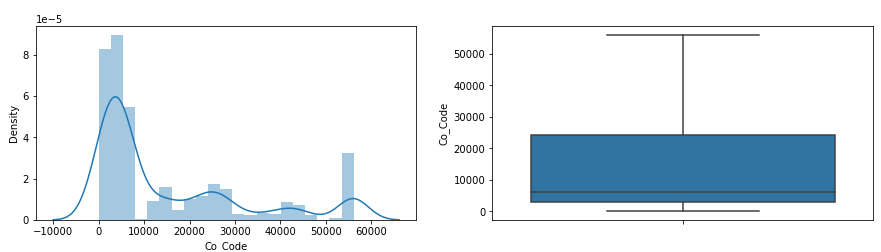


Figure - Co\_Code Univariate Analysis

The 50% median would lie below 10000, by descriptive analysis on data we can see it is 6077.5

### **Equity\_Paid\_Up**

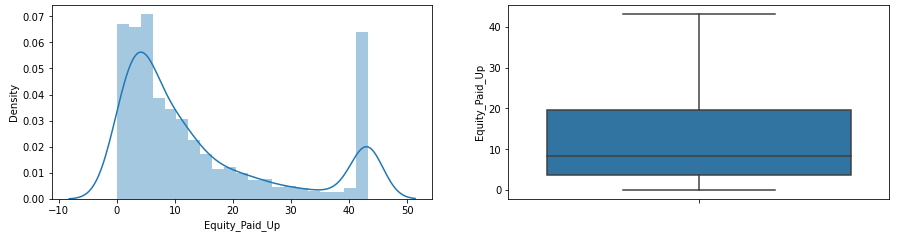


Figure - Equity\_Paid\_Up Univariate Analysis

The 50% median would lie below 10, by descriptive analysis on data we can see it is 8.29. Highest value >45 will increase mean value of data.

### **Book\_Value\_Adj\_Unit\_Curr**

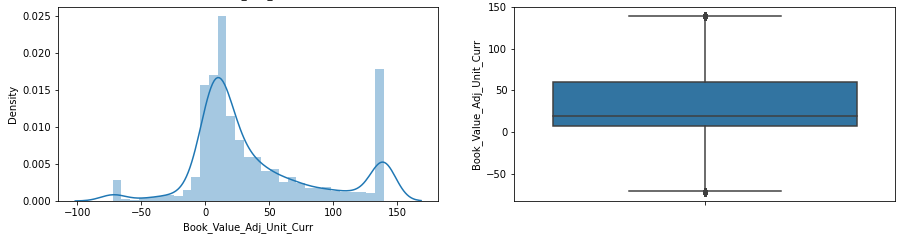


Figure - Book\_Value\_Adj\_Unit\_Curr Univariate Analysis

The data has mean 38.13 with standard deviation 50.12. 50% median is 18.92

### **ROG\_Total\_Assets\_perc**

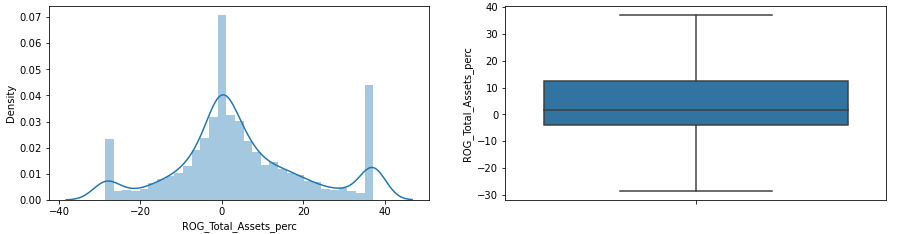


Figure - ROG\_Total\_Assets\_perc Univariate Analysis

The data has mean 4.27 with standard deviation 16.37. 50% median is 1.475

### **Creditors\_Vel\_Days**

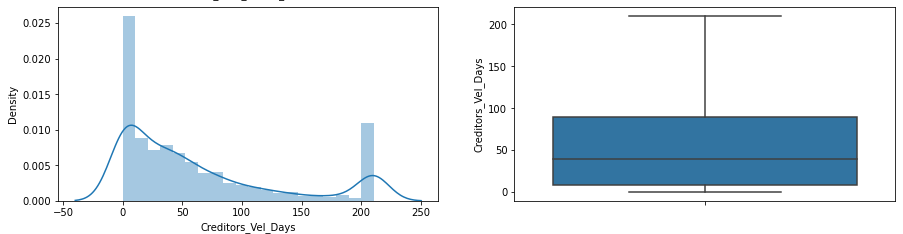


Figure - Creditors\_Vel\_Days Univariate Analysis

The data has mean 62.44 with standard deviation 68.14. 50% median is 39.0

### **Value\_of\_Output\_to\_Gross\_Block**

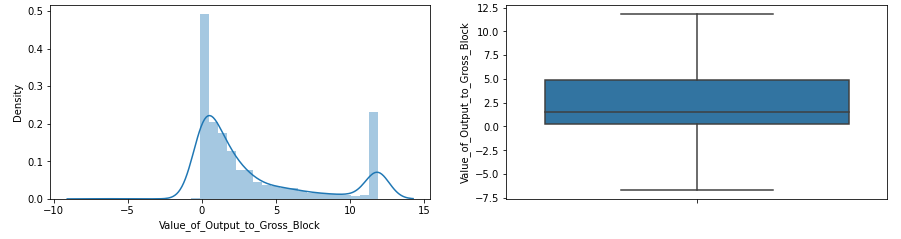


Figure - Value\_of\_Output\_to\_Gross\_Block Univariate Analysis

The data has mean 3.35 with standard deviation 4.10. 50% median is 1.53

## **Bi-variate Analysis**

### **Co\_Code vs. Default**

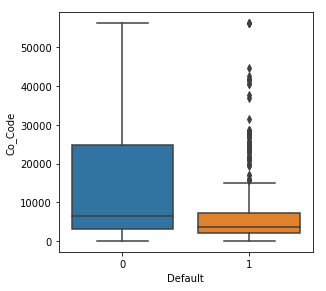


Figure - Co\_Code Bivariate Analysis

Defaults are having comparatively lower Co\_Code values than non-Defaults.

### **Equity\_Paid\_Up vs. Default**

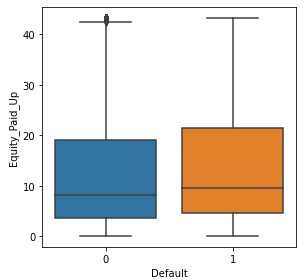


Figure - Equity\_Paid\_Up Bivariate Analysis

There is no significant difference identified in Equity\_Paid\_Up for Defaults and non-Defaults.

### **Book\_Value\_Adj\_Unit\_Curr vs. Default**

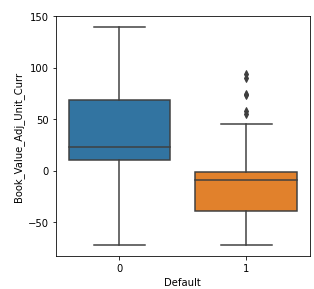


Figure - Book\_Value\_Adj\_Unit\_Curr Bivariate Analysis

Defaults are having comparatively lower Book\_Value\_Adj\_Unit\_Curr values than non-Defaults.

### **ROG\_Total\_Assets\_perc vs. Default**

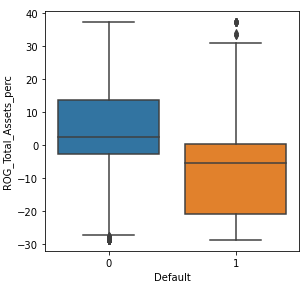


Figure - ROG\_Total\_Assets\_perc Bivariate Analysis

Defaults are having comparatively lower ROG\_Total\_Assets\_perc values than non-Defaults.

### **Creditors\_Vel\_Days vs. Default**

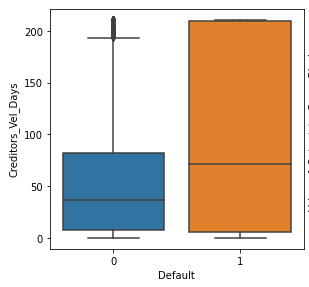


Figure - Creditors\_Vel\_Days Bivariate Analysis

There is no significant difference identified in Creditors\_Vel\_Days for Defaults and non-Defaults.

### **Value\_of\_Output\_to\_Gross\_Block vs. Default**

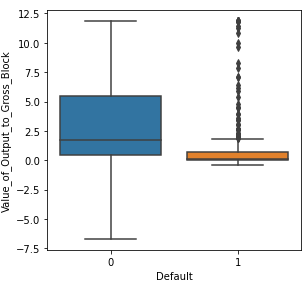


Figure - Value\_of\_Output\_to\_Gross\_Block Bivariate Analysis

Defaults are having comparatively lower Value\_of\_Output\_to\_Gross\_Block values than non-Defaults.

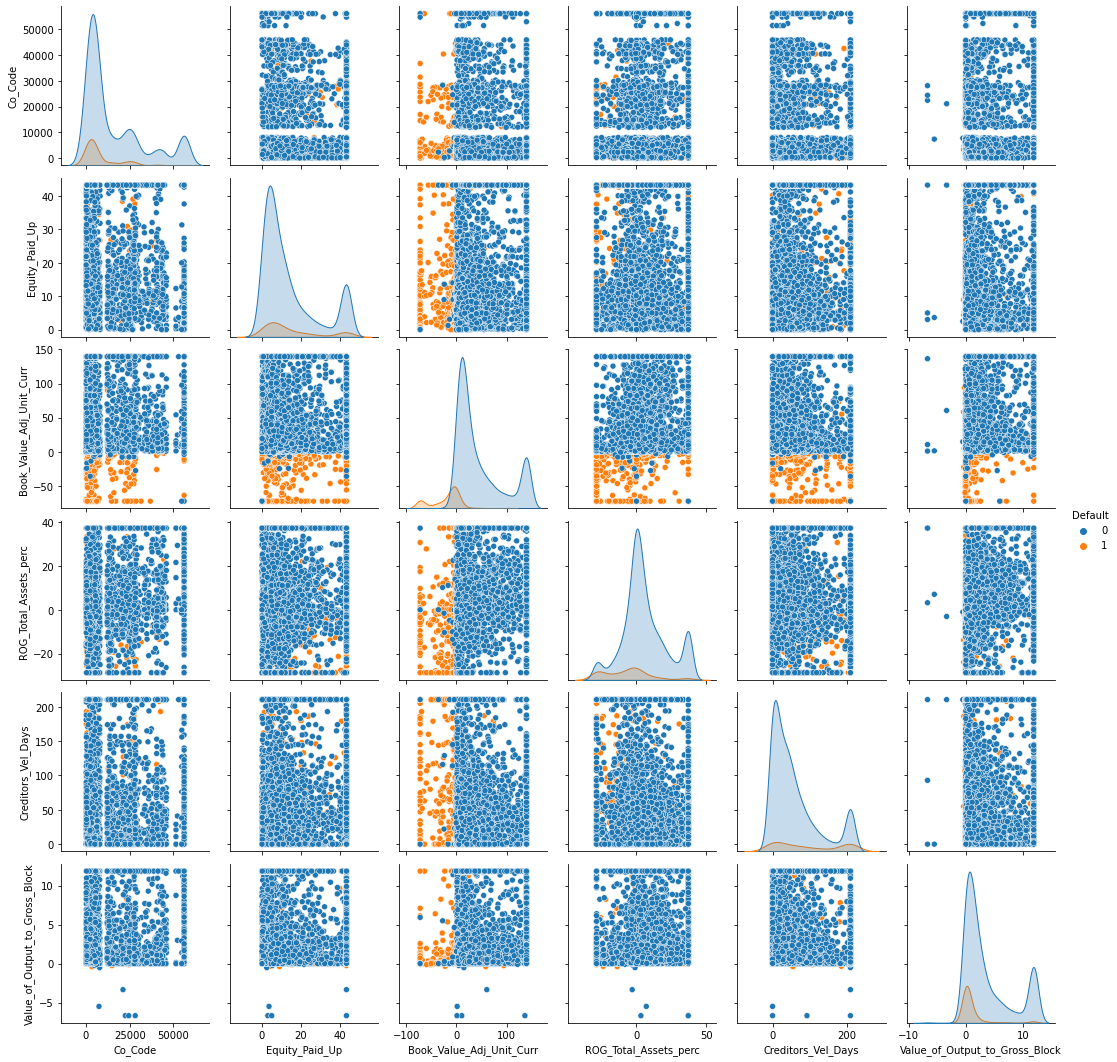


Figure - Pair Plot Analysis

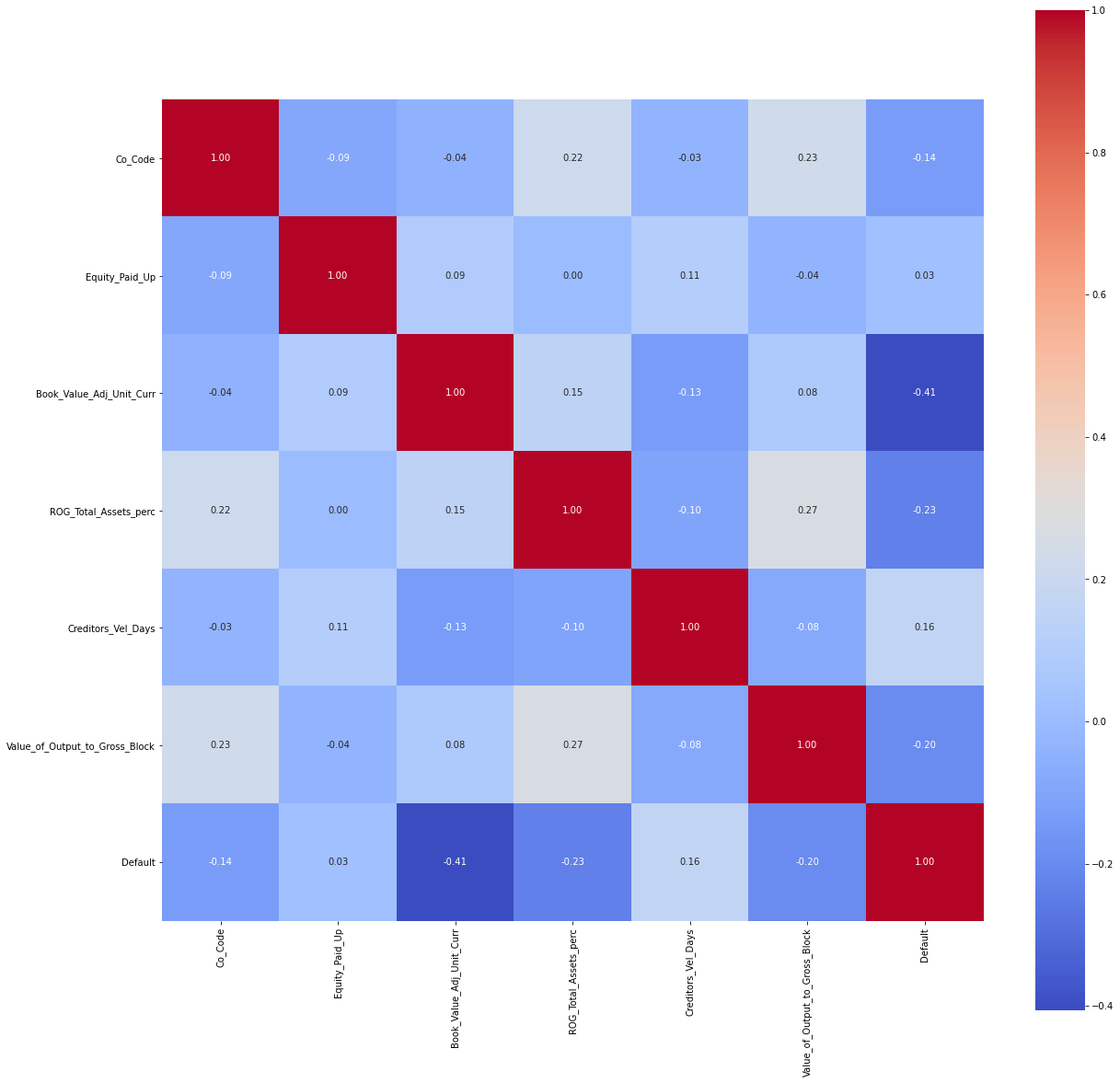


Figure - Heat Map of Significant Features Remaining

## **1.5 Train Test Split**

We are dividing data in 67 – 33 ratios, where 67% would be train test and remaining 33% would be test data.

For keeping the same seed, we have used random state as 42

Train data –

Shape: 2402 with 6 features

Default Proportion: 89.80% non-defaults and 10.20% defaults

Test data –

Shape: 1184 with 6 features

Default Proportion: 87.93% non-defaults and 12.07% defaults

## **1.6 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**

In order to build a logit regression from statsmodel, we need to derive a logit formula which will LOGIT model can be built.

We have started off with our most important features identified, which are below –

|  |
| --- |
| Co\_Code |
| Equity\_Paid\_Up |
| Book\_Value\_Adj\_Unit\_Curr |
| ROG\_Total\_Assets\_perc |
| Creditors\_Vel\_Days |
| Value\_of\_Output\_to\_Gross\_Block |
| Default |

### **Logit Formula 1:**

Default ~ Co\_Code + Equity\_Paid\_Up + Book\_Value\_Adj\_Unit\_Curr + ROG\_Total\_Assets\_perc + Creditors\_Vel\_Days + Value\_of\_Output\_to\_Gross\_Block

### **Model 1 Summary:**

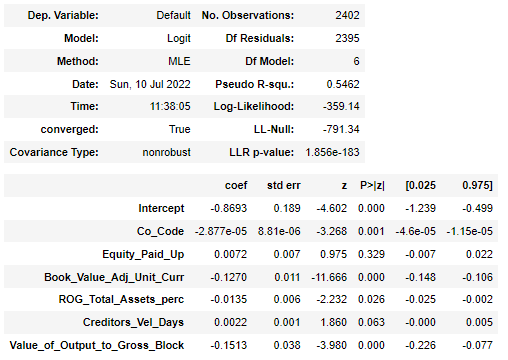


Figure - Logit Regression Result (MODEL 1)

* There are 2 insignificant features (having P value > 5%) should be excluded –
  + Equity\_Paid\_Up
  + Creditors\_Vel\_Days

### **Logit Formula 2:** (after removing above 2 features from analysis)

Default ~ Co\_Code + Book\_Value\_Adj\_Unit\_Curr +ROG\_Total\_Assets\_perc + Value\_of\_Output\_to\_Gross\_Block

### **Model 2 Summary:**

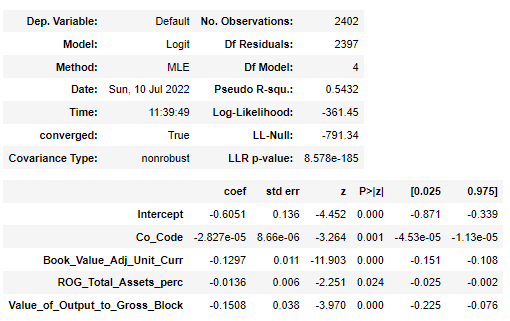


Figure - Logit Regression Result (MODEL 2)

There is no insignificant variable now. We can proceed with this model…

### **Classification Report (with threshold probability 0.5)**

**On Train Data –**

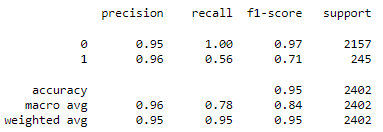


Figure - Classification Report (Train Data)

**On Test Data –**

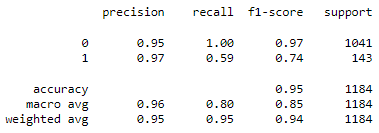


Figure - Classification Report (Test Data)

* The model has good precision and f1-score
* But recall is significantly lower on defaults
* The model is not overfit, as the metrics have been improved on test data

### **Optimal Threshold**

We can find optimal threshold with roc\_curve values. And we find optimal probability as **0.1979 on test data**

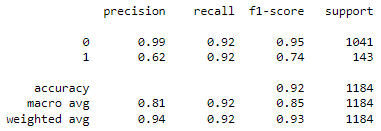
****

Figure - Classification Report Test Data (On Optimal Value)

* Recall has been significantly improved but precision and f1-score has been decreased.

## **1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **Threshold 0.5** | | **Threshold 0.1979** | |
| **Train Data** | **Test Data** | **Train Data** | **Test Data** |
| Precision | 0.96 | 0.97 | 0.59 | 0.62 |
| Recall | 0.56 | 0.59 | 0.89 | 0.92 |
| F1-Score | 0.71 | 0.74 | 0.71 | 0.74 |

Table - Performance Comparison Metrics

We are evaluating the model on its’ metrics over **Defaults**, as predicting false Default can lead to company losing some business. Hence the key metrics we would be looking for is **Recall**.

**Interpretation and Decision –**

* High precision on default threshold 0.5, but lower on optimal threshold 0.1979
* Good recall 92% on optimal threshold but lower on default threshold 59%
* F1 Score stays the same on both default and optimal threshold

We will predict our target variable against optimal threshold.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END – Milestone 1 of 2 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# **Credit Risk – Milestone 2**

## **1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach**

We had split data into 67:33 ratio of training and testing.

Since random forest would have overfitted (having excellent accuracy at train and reduced on test) if we build without any parameters tuned,

We are building RANDOM FOREST model with hyper tuned parameter as below –

* + max\_depth: [3, 5, 7]
  + min\_samples\_leaf: [5, 10, 15]
  + min\_samples\_split: [15, 30, 45]
  + n\_estimators: [25, 50]

Through grid search we derived the best parameters out of them as below –

* + max\_depth – 7
  + min\_samples\_leaf – 5
  + min\_samples\_split – 30
  + n\_estimators – 25 (# of trees)

### **RF Classification Report (Train Data)**

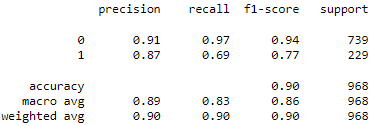


Figure - RF Classification Report (Train Data)

* Model has good precision and recall on non-default predictions
* Recall has been impacted for defaulted predictions
* Model accuracy is decent 90%

## **1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model**

### **RF Classification Report (Test Data)**

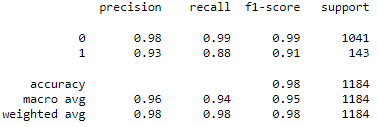


Figure - RF Classification Report (Test Data)

### **Comparison Matrix (RF)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **Non - Defaults** | | **Defaults** | |
| **Train Data** | **Test Data** | **Train Data** | **Test Data** |
| Precision | 0.91 | 0.98 | 0.87 | 0.93 |
| Recall | 0.97 | 0.99 | 0.69 | 0.88 |
| F1-Score | 0.94 | 0.99 | 0.77 | 0.91 |

Table - RF Comparison Matrix (RF)

INTERPRETATIONS –

* Metrics have been significantly improved on testing data
* Precision & recall are higher on both non-defaults and defaults prediction
* Model is slightly underfitted as the test data seem to have higher Precision and Recall as well as F1-Score.

## **1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach**

Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems.

It is used for modelling differences in groups i.e. separating two or more classes. It is used to project the features in higher dimension space into a lower dimension space.

We have built LDA model with default parameters –

Solver – Single value decomposition

Tolerance – 0.0001

### **LDA Classification Report (Train Data)**

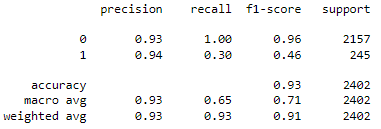


Figure - LDA Classification Report (Train Data)

Recall is very poor on with default threshold… We will find optimal threshold to improve Recall –

Optimal Threshold -- 0.130

### **LDA Classification Report with Optimal Threshold (Train Data)**

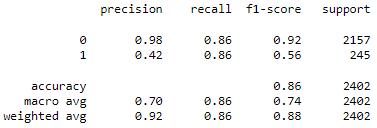


Figure - LDA Classification Report with Optimal Threshold (Train Data)

* Recall has been improved from 0.30 to 0.86 at cost of precision which has came down from 0.94 to 0.42
* Average Accuracy has come down from 93% to 86%

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

### **LDA Classification Report with Optimal Threshold (Test Data)**

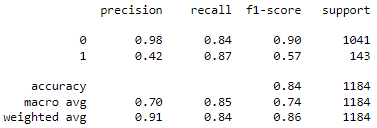


Figure - LDA Classification Report with Optimal Threshold (Test Data)

### **Comparison Matrix (LDA)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **Non - Defaults** | | **Defaults** | |
| **Train Data** | **Test Data** | **Train Data** | **Test Data** |
| Precision | 0.98 | 0.98 | 0.42 | 0.42 |
| Recall | 0.86 | 0.84 | 0.86 | 0.87 |
| F1-Score | 0.92 | 0.9 | 0.56 | 0.57 |

Table - Comparison Matrix (LDA)

INTERPRETATION –

* Model has better precision and recall on non-Defaults
* LDA has < 50% precision on both Train and test data for Defaults prediction
* Model is not over or under fitted

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

### **ROC Curve**

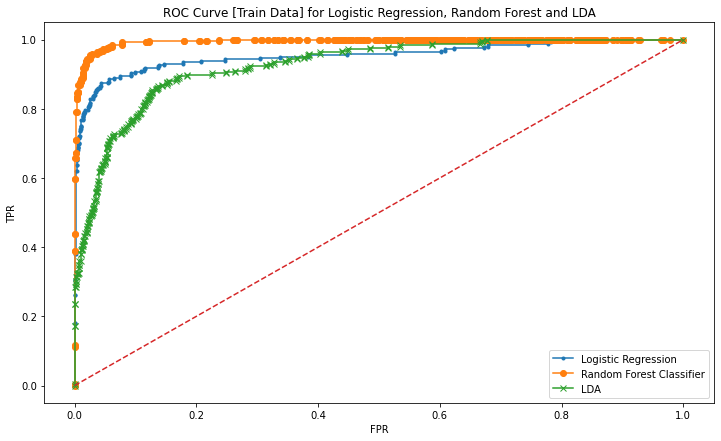


Figure - ROC Curve (Train Data)

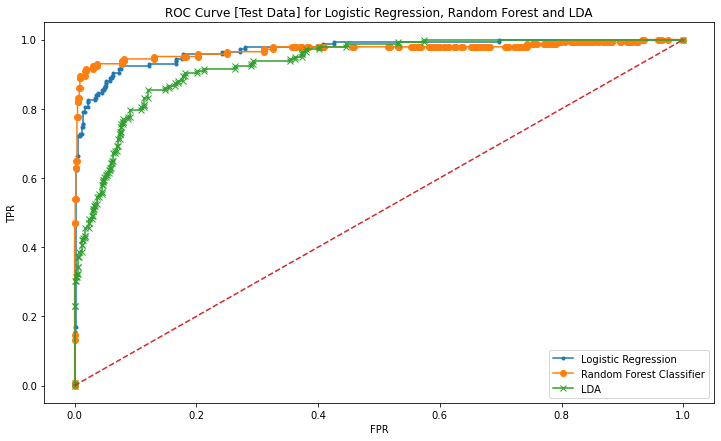


Figure - ROC Curve (Test Data)

From above 2 plots, we can infer –

* Random Forest classifies better than Logistic regression and LDA
* Logistic regression model performs better than LDA throughout TPR/FPR

### **ROC AUC Score**

|  |  |  |
| --- | --- | --- |
| **ROC AUC SCORE** | **Train Data** | **Test Data** |
| **Logistic Regression** | 0.953 | 0.969 |
| **Random Forest** | 0.994 | 0.97 |
| **Linear Discriminant Analysis** | 0.925 | 0.929 |

Table - ROC AUC Score

* **Random Forest** has better ROC AUC score on train and test data compare to other models
* **LDA** model has lowest ROC AUC score among all 3 models

### **Metrics Comparison Report**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metrics (On Defaults)** | **Logistic Regression Model** | | **Random Forest Model** | | **Linear Discriminant Analysis** | |
| **Train Data** | **Test Data** | **Train Data** | **Test Data** | **Train Data** | **Test Data** |
| Precision | 0.59 | 0.62 | 0.94 | 0.93 | 0.42 | 0.42 |
| Recall | 0.89 | 0.92 | 0.87 | 0.88 | 0.86 | 0.87 |
| F1-Score | 0.71 | 0.74 | 0.90 | 0.91 | 0.56 | 0.57 |

Table - Metrics Comparison Report

INTERFERANCE –

* Random Forest is best model among these 3, and have > 0.9 model score
* Logistic Regression Model has better recall compared to others.
* LDA model has lowest precision score and f1-score

**Random Forest Model** would be our best model to predict defaults!

## **1.13 State Recommendations from the above models**

From Table 14 - Metrics Comparison Report and important features derived, we can make few recommendations to business –

* From company’s perspective, we should identify true defaults and avoid false positives, hence recall of the model should be better. And **Logistic Regression Model** has highest recall, so it would be our model for this case.
* From Investor Perspective, model should be able to precisely predict Defaults, and **Random Forest Model** among all has better precision rate for this case.
* The companies being evaluated for default/loan, they should keep their total assets increasing. Rate of growth plays key role in default prediction.
* They should also pay-out suppliers on time. Creditor\_Vel\_day is an important field plays part in default prediction.

# **Market Risk Analysis**

## **Executive Summary**

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks.

## **Introduction**

We have stock’s weekly price for below companies –

* + - Infosys
    - Indian Hotel
    - Mahindra & Mahindra
    - Axis Bank
    - SAIL
    - Shree Cement
    - Sun Pharma
    - Jindal Steel
    - Idea Vodafone
    - Jet Airways

We will calculate the mean and standard deviation on the stock returns and share insights.

## **Exploratory Data Analysis**

* We have 314 records of prices and 10 companies
* Dataframe does not have any null value

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

### **Infosys**

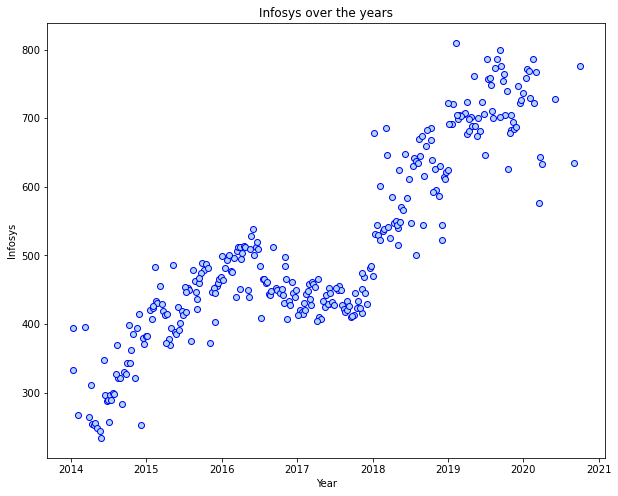
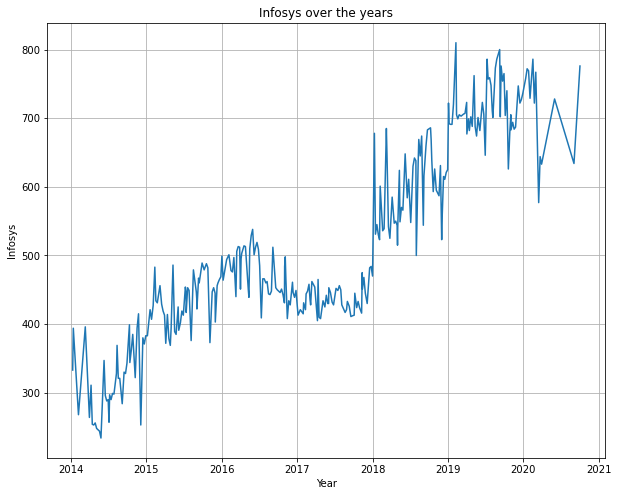


Figure - Infosys Prices over the Years



* Price trend is upwards for Infosys from 2014 – 2021
* The market was sideways for Infosys from 2016 – 2018, and investors might haven’t got any return.

### **Shree Cement**

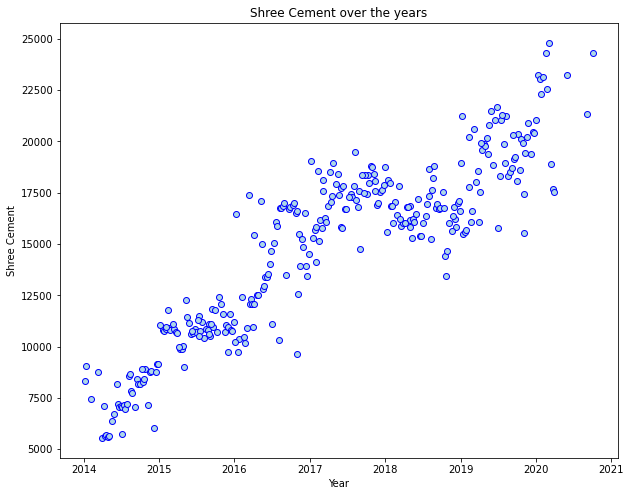
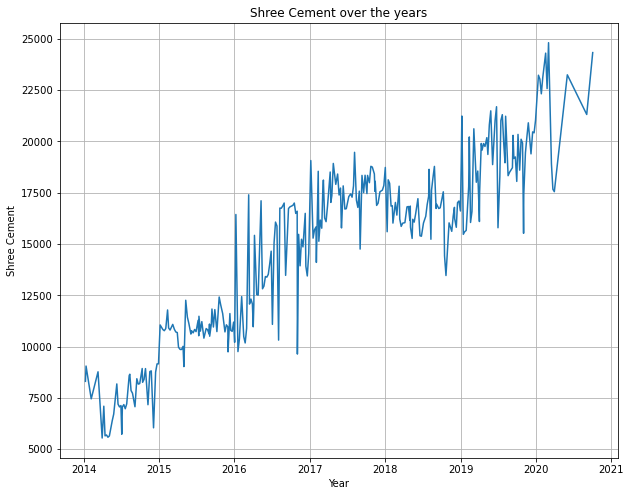


Figure - Shree Cement Prices over the years



* Shree Cement’s Price moved from 5000 to 25000 in 2014 – 2021
* The price moved upwards for Shree Cement over the years

## **2.2 Calculate Returns for all stocks with inference**

To calculate returns we will take a log and one step difference from previous price.

Since we have weekly price data, so returns would also be **weekly returns**

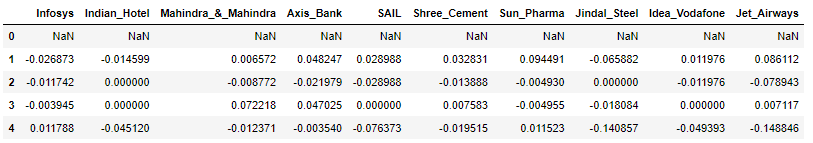


Table - Weekly Returns (Top 5 Rows)

### **Infosys**

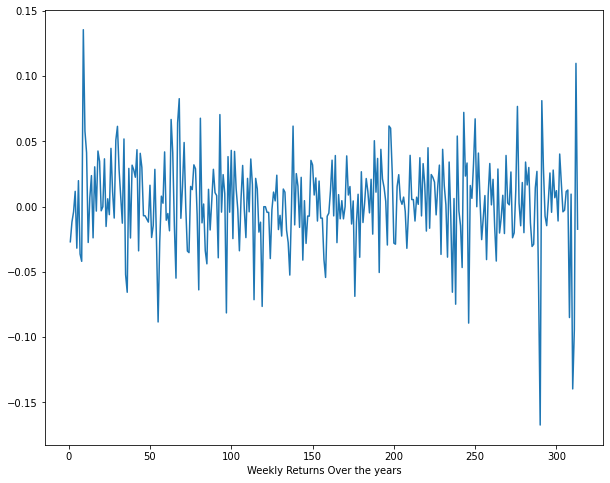


Figure - Weekly Returns (Infosys)

### **Shree Cement**

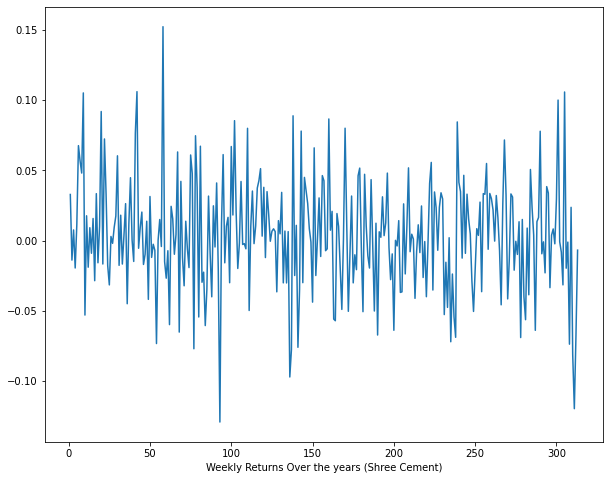


Figure - Weekly Returns (Shree Cement)

## **2.3 Calculate Stock Means and Standard Deviation for all stocks with inference**

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

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

|  |  |
| --- | --- |
| **Stock Name** | **Avg Returns (Mean)** |
| Infosys | 0.002794 |
| Indian\_Hotel | 0.000266 |
| Mahindra\_&\_Mahindra | -0.001506 |
| Axis\_Bank | 0.001167 |
| SAIL | -0.003463 |
| Shree\_Cement | 0.003681 |
| Sun\_Pharma | -0.001455 |
| Jindal\_Steel | -0.004123 |
| Idea\_Vodafone | -0.010608 |
| Jet\_Airways | -0.009548 |

Table -Stocks' Average Return (Weekly)

* Shree Cement and Infosys has good average weekly returns.
* Idea Vodafone and Jet Airways are not performing well and have poor weekly returns

|  |  |
| --- | --- |
| **Stock Name** | **Volatility (Std Dev)** |
| Infosys | 0.03507 |
| Indian\_Hotel | 0.047131 |
| Mahindra\_&\_Mahindra | 0.040169 |
| Axis\_Bank | 0.045828 |
| SAIL | 0.062188 |
| Shree\_Cement | 0.039917 |
| Sun\_Pharma | 0.045033 |
| Jindal\_Steel | 0.075108 |
| Idea\_Vodafone | 0.104315 |
| Jet\_Airways | 0.097972 |

Table - Stocks' Volatility

* Idea Vodafone and Jet Airways are very much volatile
* Infosys, Shree Cement and M&M are less volatile compared to other stocks in the list.

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

### **Stock Means (Average Weekly Returns)**

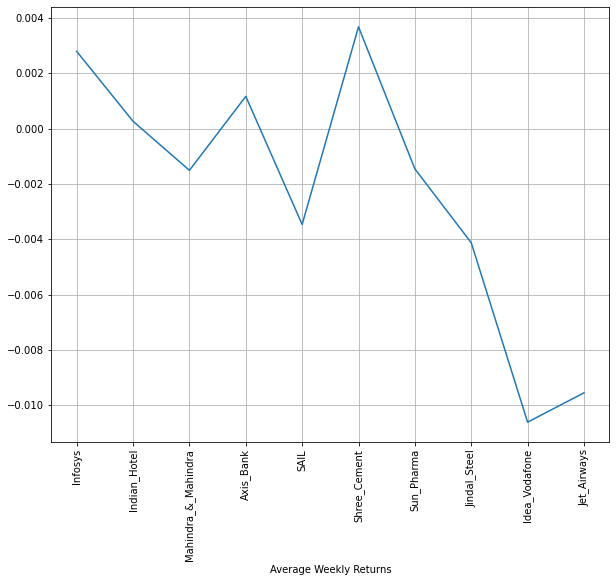


Figure - Average Weekly Return

From above plot, we can infer –

* Shree Cement has higher average weekly return, compared to others
* Idea vodafone has lowest average returns among all stocks
* Jet Airways also doesn’t have good returns
* Infosys has high average returns

### **Stock Standard Deviation (Volatility)**

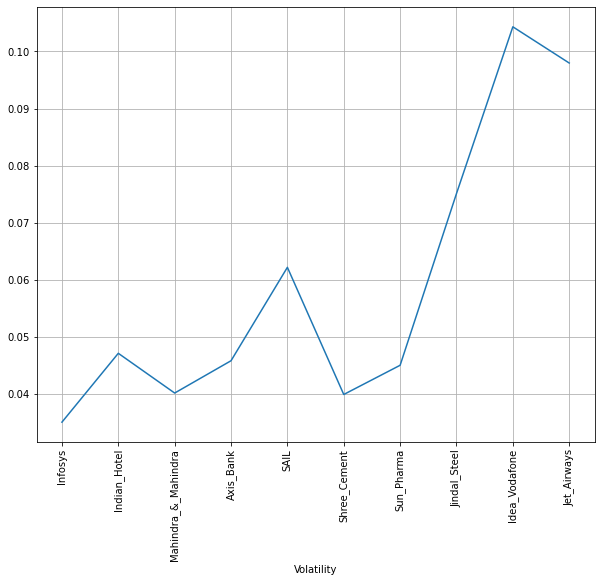


Figure – Volatility

* Idea vodafone and Jet Airways are highly volatile
* Infosys has lowest volatility among all stocks
* Shree Cement and M&M also has lower volatility

## **2.5 Conclusion and Recommendations**

### **Conclusion**

* Idea vodafone and Jet Airways are in downfall, and not good stocks for investing
* Infosys and Shree Cement has performed well during the years.
* Average returns of Infosys and Shree Cement has been good, they would be good for investing

### **Recommendations**

* Traders usually picks volatile stocks to generate profits in short term, for them Idea vodafone and Jet Airways would be ideal picks.
* Investors should avoid bad performing stocks and set goal for long term wealth creation; hence Infosys and Shree Cement would be best pick for them.
* In finance section, Axis bank gave good return compared to M&M, and their volatility is also comparable, hence investors should align towards Axis bank.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*