**Project: FRA Project (Milestone-2)**

Pavan Kalyan B | PGP DSBA Oct\_C’21 | SEP 2022

**Table of contents**

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

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

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

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

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

**1.13** State Recommendations from the above models

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

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

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

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

**2.5** Conclusion and Recommendations

**Problem Statement**

**This project contains details for Milestone-2**

**Credit Risk**

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.

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

**Hints**:

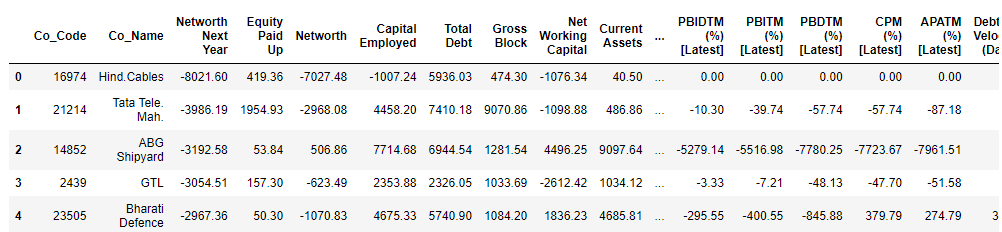
Dependent variable - We need to create a default variable which should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Test Train Split -   Split the data into Train and Test dataset in a ratio of 67:33 and use random state = 42. Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset.

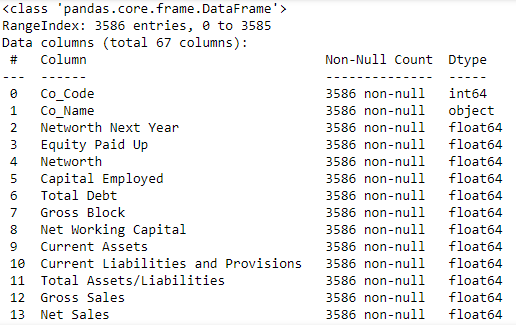
[Credit Risk Dataset](https://olympus.mygreatlearning.com/courses/63586/files/6377908/download?verifier=4yjeDOx5uSKLuSLlRGyX4rsW7L9WUj5ZLxPEyoV5&wrap=1)

[Data Dictionary](https://olympus.mygreatlearning.com/courses/63586/files/6377899/download?verifier=Hd9NBRcoMOWuPQS0ykPPeMhPWWWvOypE974llOGr&wrap=1)

**Reading the data and exploring.**

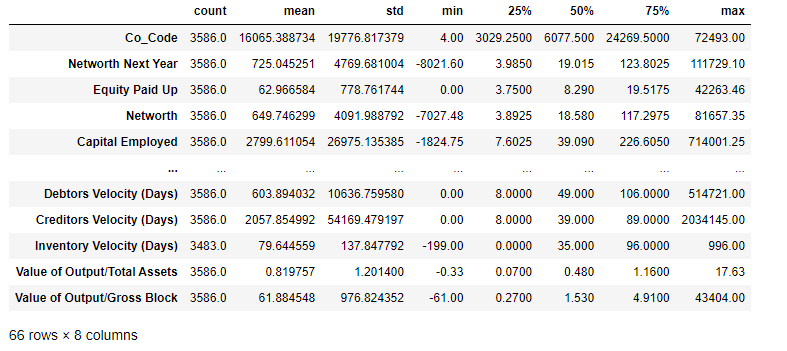
****

**Table 0.1.1 First five rows in the data**

****

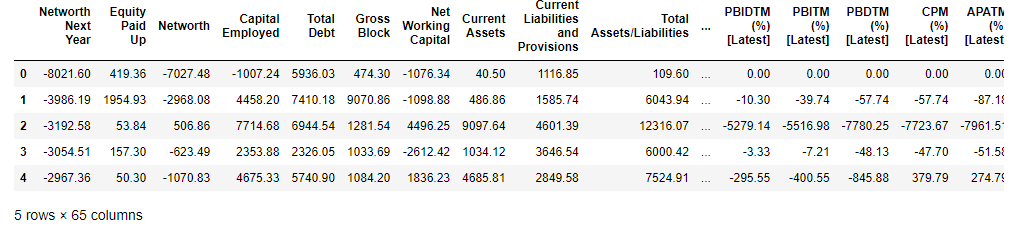
**Table 0.1.2 Data Info**

* We can observe that there are 67 columns and 3586 rows in the data
* It uses memory of 1.8+MB
* We have 63 - float64, 3 - int64, 1 - object data types

****

**Table 0.1.3 Summary of the data**

* We can observe that two columns need to be remove which doesn’t make any sense
* Co\_code and co\_name are the one need to be removed

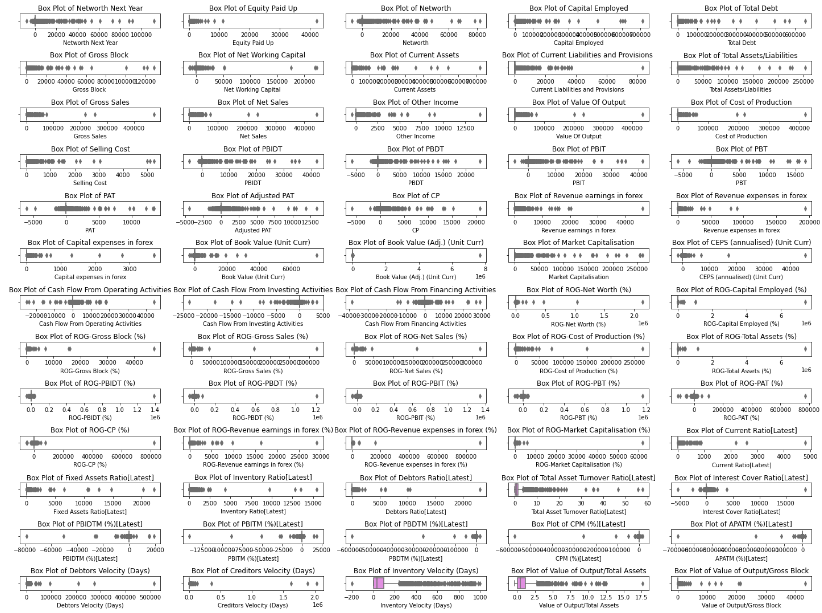
****

**Table 0.1.4 First rows after removing 2 columns**

* We can observe that 65 columns and rest rows are present in the data
* There are no duplicates in the data

**Outlier Treatment**

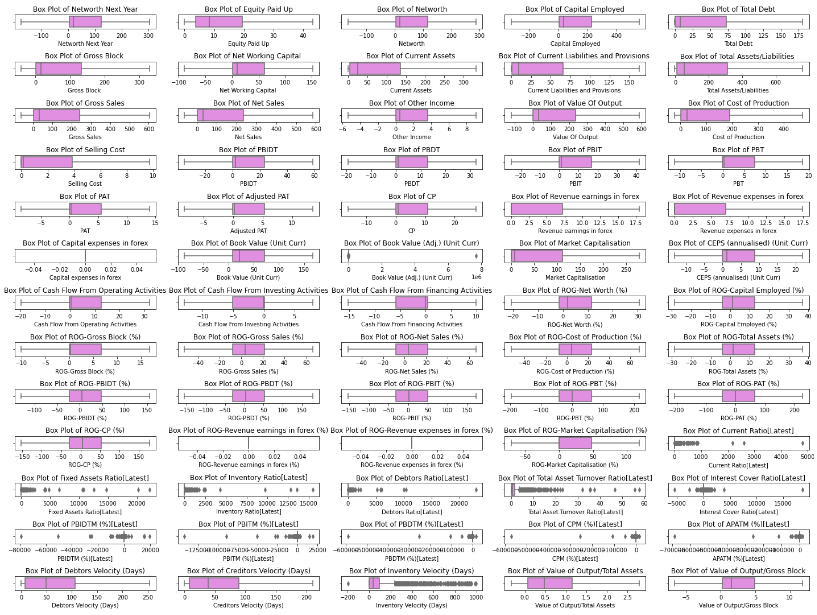
**Before Treating Outliers:**

****

**Fig 1.1.1 Boxplot showing outliers**

* We can observe that most of the columns has outliers in the data which needs to be treated

**After Treating Outliers:**

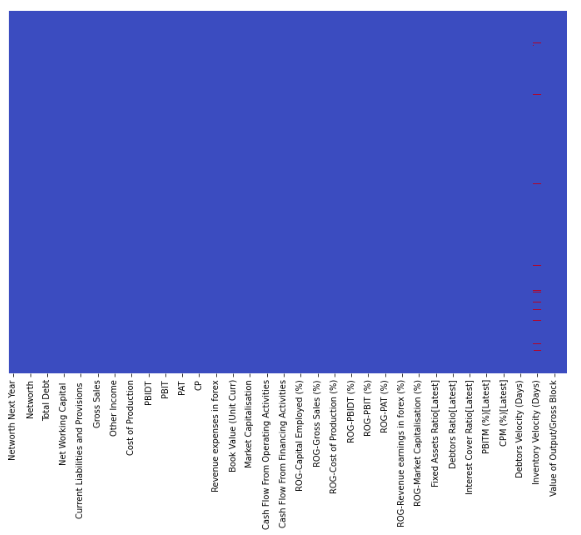
****

**Fig 1.1.2 Boxplot showing no outliers**

* We can observe that there no outliers after the treatment except in few cases

**Missing Value Treatment**

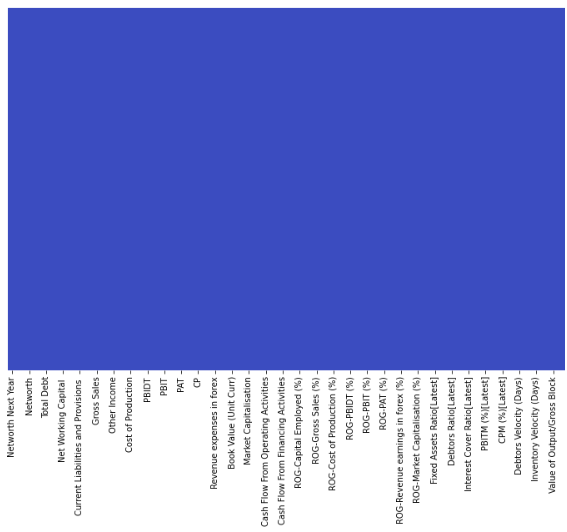
**Before Treating Missing values:**



**Fig 1.2.1 Missing values in the data**

* We can observe that there are missing values in the data in few of the columns
* More number of missing values are identified in the inventory velocity (Days) column and less missing values in rest columns

**After Treating Missing Values:**

****

**Fig 1.2.2 Data without missing values**

* We have replaced the null values with the knn imputation

**Transform Target variable into 0 and 1**

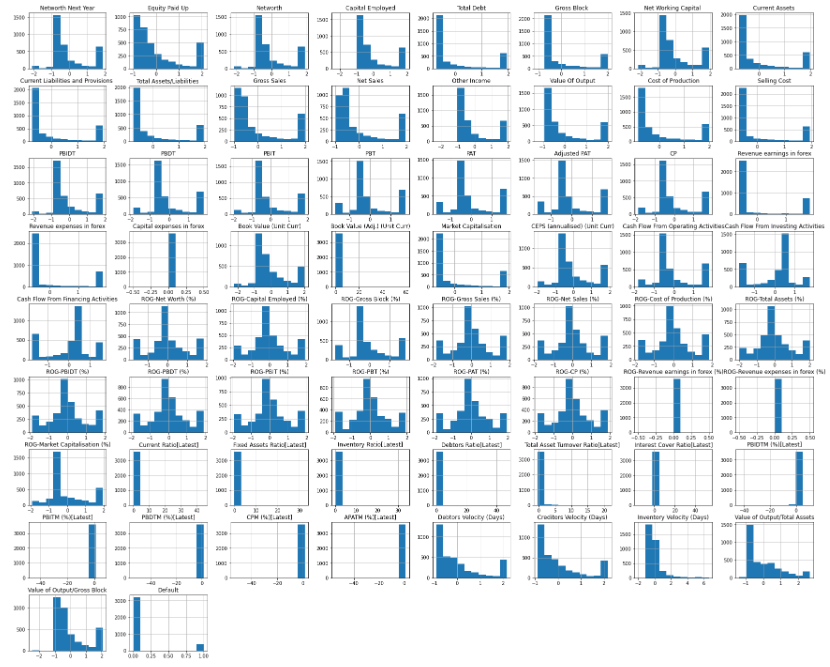
****

* There are 3198 positive values and 388 negative values in the data
* We make positive values to be fixed as 0 and negative values to be fixed as 1

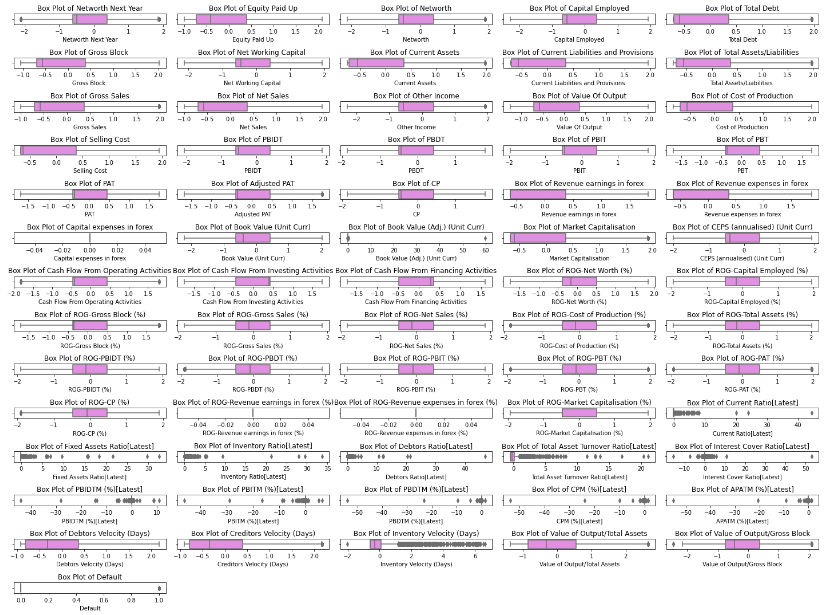


**Univariate & Bivariate analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)**

**Univariate Analysis:**

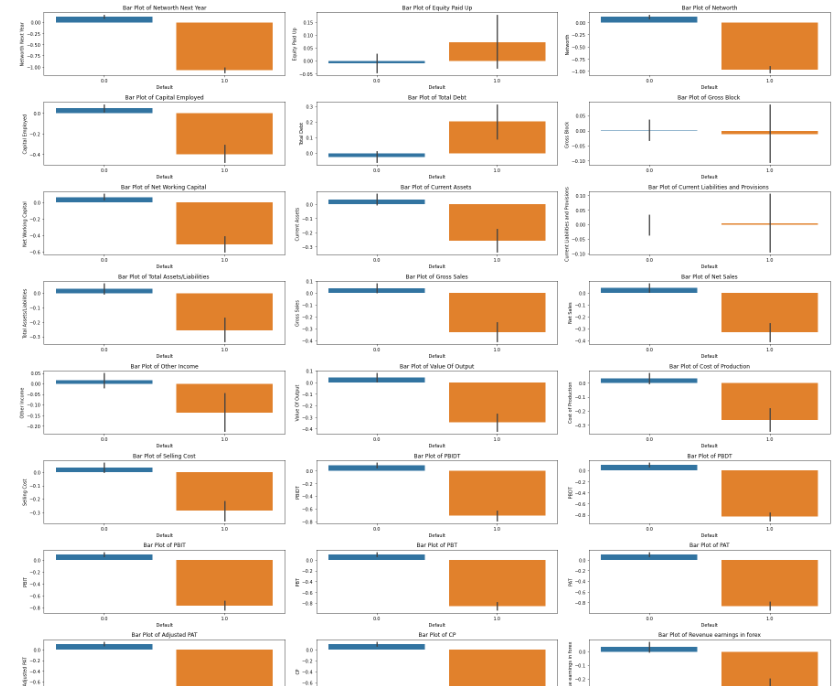
****

**Fig 1.4.1 Histogram for all columns**

****

**Fig 1.4.2 Box plot of all columns**

**Bivariate Analysis:**

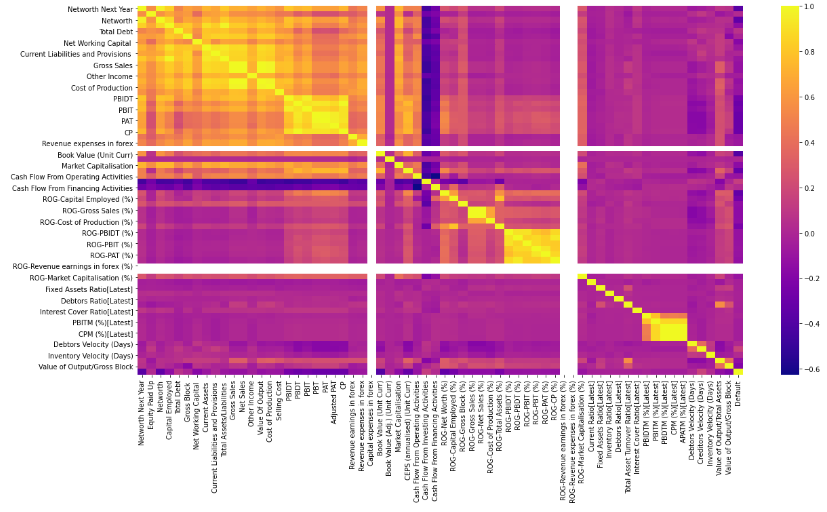
****

**Fig 1.4.3 Box plot shown for few columns**

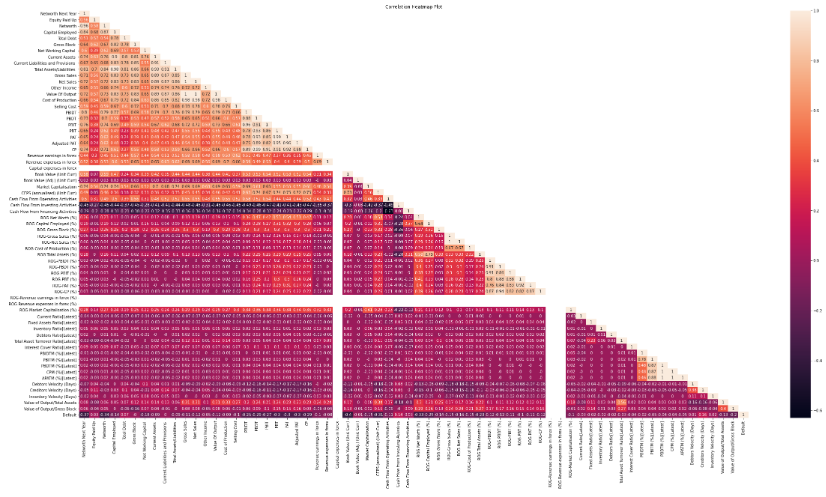
* Created bar plots for few columns based on the new column “Default”

**Multivariate Analysis:**

**Correlation plot:**



**Fig 1.4.4 Heat map**

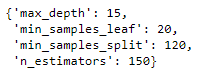
****

**Fig 1.4.5 Triangular Heat map**

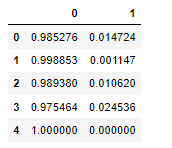
**Train Test Split**

****

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

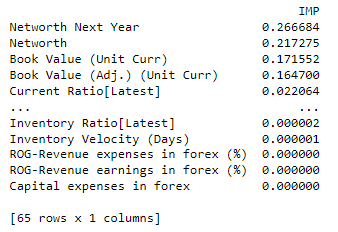


* The above are the best fit parameters for the random forest model
* Hence, we will use these parameters to get the good accuracy for the model



**Table 1.8.1 Training Predictions**

* The above table is the sample prediction for the training data

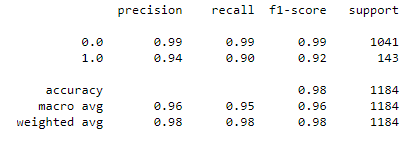


**Table 1.8.2 Important Features**

* The above table represents the important features in the data with the importance values

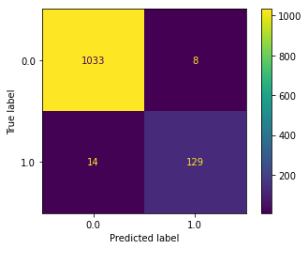
**1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model**

**Classification metrics on testing data set:**



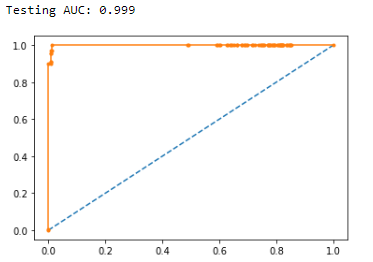
**Table 1.9.1 Testing data accuracy**

**Confusion matrix on testing dataset:**

****

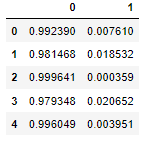
**Fig 1.9.1 confusion matrix on testing**

**AUC & ROC on testing dataset:**

****

**Fig 1.9.2 AUC & ROC on testing**

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

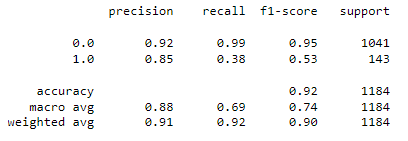
****

**Table 1.10.1 Training predictions on LDA model**

* We can observe that as the model built, these are the training predictions
* Based on these training predictions, we are testing it

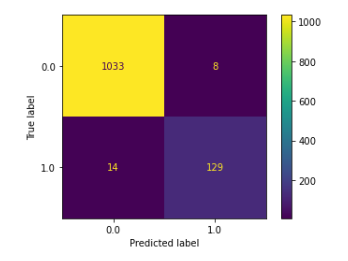
**1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model**

**Classification metrics on testing data set:**

****

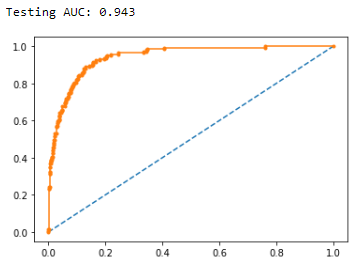
**Table 1.10.1 Testing data accuracy**

**Confusion matrix on testing dataset:**

****

**Fig 1.10.1 confusion matrix on testing**

**AUC & ROC on testing dataset:**

****

**Fig 1.10.2 AUC & ROC on testing**

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Data** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **AUC** |
| Logistic | Train | 100 | 100 | 100 | 100 | 100 |
| Logistic | Test | 98 | 99 | 99 | 99 | 99 |
| RFC | Train | 98 | 99 | 99 | 99 | 99 |
| RFC | Test | 98 | 99 | 99 | 99 | 99 |
| LDA | Train | 93 | 94 | 99 | 96 | 95 |
| LDA | Test | 92 | 92 | 99 | 95 | 94 |

* By looking into the table, we can observe that logistic and random forest performs good and gets better predictions
* Comparing all the three models, Logistic regression model is quite has greatest accuracy and we choose it as the best model

**1.13 State Recommendations from the above models**

* It is observed that credit risk in based on the credit worthiness of the customer
* The model which we have selected highly selects the default cases from the data
* Here are the steps to further classify the defaulters
  + Low Risk
  + Medium Risk
  + High Risk
  + Very High Risk

We will focus more on the very high risk and high risk factors to overcome the credit risk analysis.

**Problem Statement:**

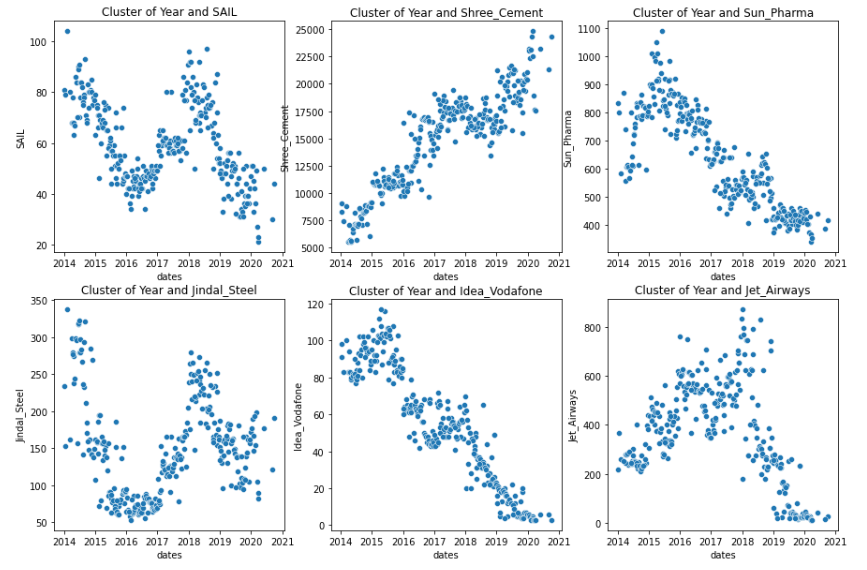
**Market Risk**

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.

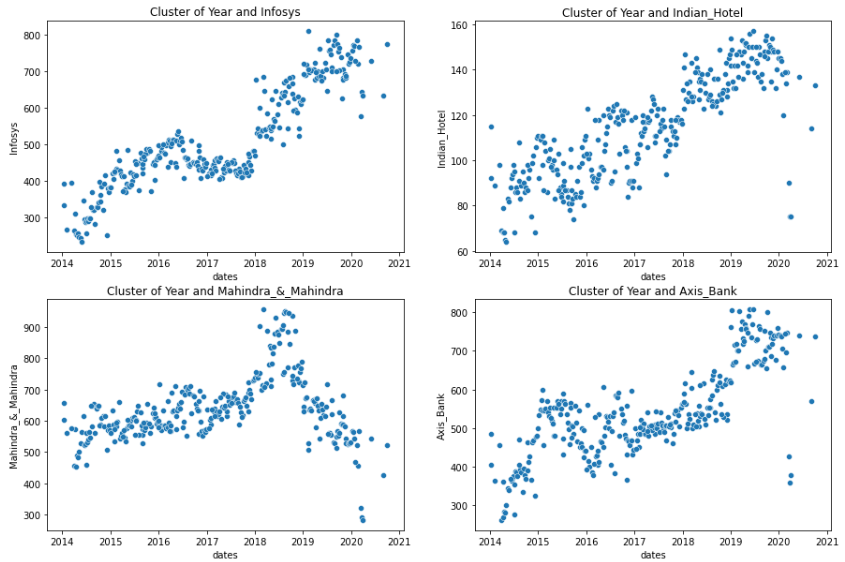
Please find attached the files to be referred.

[Market Risk Dataset](https://olympus.mygreatlearning.com/courses/63586/files/6377909/download?verifier=8w4fpSFt4PUXXPt3a9gXSkS8tus7jbFGuUkuSDBI&wrap=1)

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



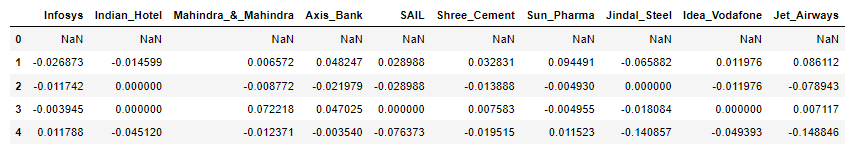
**Fig 2.1.1 Cluster of Stock prices vs Time**

****

**Fig 2.1.2 Cluster of Stock prices vs Time**

* As we look into fig 2.1.1 and 2.1.2, we can observe the clusters of stock prices and time
* There is a positive correlation for the stock prices of ‘Axis\_Bank’, ‘Infosys’ and ‘Shree\_Cement’ with respect to time
* There are negative correlation for the stock prices of ‘Sun\_Pharma’ and ‘Idea\_Vodafone’ with respect to time
* Rest all stock prices are moderately positive and negative correlated

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

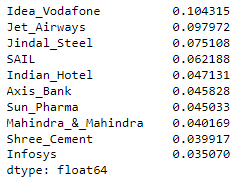
****

**Table 2.2.1 Returns by taking logarithmic**

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

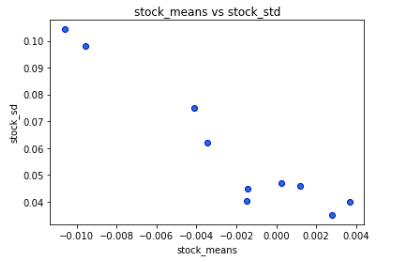
****

**Table 2.3.1 Stock Means**

****

**Table 2.3.2 Stock Standard Deviations**

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

****

**Fig 2.4.1 Stock Means vs Stock Standard Deviations**

* The plot is negatively correlated
* Stocks on the high has high volatile and low returns
* Stocks on the low has low volatile but high returns
* The graph will be useful for those who are investing newly

**2.5 Conclusion and Recommendations**

Traders mostly use several metrics to know about the risk and volatility in investments, they mostly use the standard deviation. Based on the standard deviation they make investments

* Standard deviations are more about the volatility
* If standard deviation is high, prices move wildly and investment will be risky
* If standard deviation is low, prices are stable or normal and we can invest with low risk

**Recommendations:**

* Stock Means vs Stock standard deviations plot asses the risk in investments and gives asses to reward ratio too
* Higher standard deviation, high risk in investments
* Lower standard deviation, low risk in investments

**================================================================**