Project – 2

**Company Loan Default Prediction**

Ananth Gopalakrishna Achar

PGP-DSBA 2020 - 2021

**Table of Contents**

**Credit Risk**

**Executive Summary.................................................................................................................................3**

**Data Description ......................................................................................................................................3**

**Exploratory Data Analysis........................................................................................................................5**

**Random Forest Classifier ........................................................................................................................8**

**Linear Discriminant Analysis..................................................................................................................10**

**Logistic Regression model......................................................................................................................12**

**Comparison of performance metrics from the three models......................................................................................................................................................15**

**Applying SMOTE.....................................................................................................................................16**

**Analyzing Market Risk and Returns**

**Problem summary..................................................................................................................................19**

**Exploratory data analysis.......................................................................................................................19**

**Stock Price Graph...................................................................................................................................20**

**Calculating Stock Returns......................................................................................................................22**

**Calculating Stock Means........................................................................................................................23**

**Calculating Stock Standard Deviation..................................................................................................23**

**Plot of Stock Means vs Standard Deviation……………………………………………………………………..…..24**

**Executive Summary:**

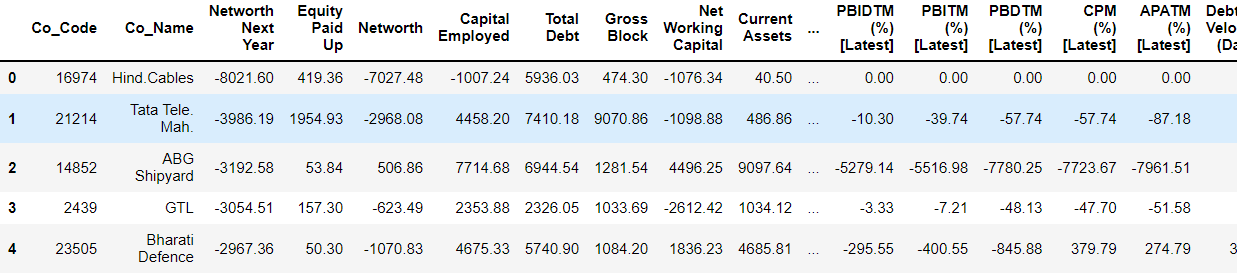
A bank’s loan data is given to us containing the list of variables taken from the Financial statements of the companies. We are expected to predict whether the company will default on their loan obligation or not. Attributes like credit worthiness, interest rates on the loan and financial stability of the company can be inferred from whether the company has defaulted or not.

**Data Description:**

|  |  |  |
| --- | --- | --- |
| **#** | **Field Name** | **Description** |
| 1 | Co\_Code | Company Code |
| 2 | Co\_Name | Company Name |
| 3 | Networth Next Year | Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities) |
| 4 | Equity Paid Up | Amount that has been received by the company through the issue of shares to the shareholders |
| 5 | Networth | Value of a company as on 2015 - Current Year |
| 6 | Capital Employed | Total amount of capital used for the acquisition of profits by a company |
| 7 | Total Debt | The sum of money borrowed by the company and is due to be paid |
| 8 | Gross Block | Total value of all of the assets that a company owns |
| 9 | Net Working Capital | The difference between a company's current assets (cash, accounts receivable, inventories of raw 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. |
| 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) |
| 12 | Total Assets/Liabilities | Ratio of total assets to liabilities of the company |
| 13 | Gross Sales | The grand total of sale transactions within the accounting period |
| 14 | Net Sales | Gross sales minus returns, allowances, and discounts |
| 15 | Other Income | Income realized from non-business activities (e.g. sale of long term asset) |
| 16 | Value Of Output | Product of physical output of goods and services produced by company and its market price |
| 17 | Cost of Production | Costs incurred by a business from manufacturing a product or providing a service |
| 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) |
| 19 | PBIDT | Profit Before Interest, Depreciation & Taxes |
| 20 | PBDT | Profit Before Depreciation and Tax |
| 21 | PBIT | Profit before interest and taxes |
| 22 | PBT | Profit before tax |
| 23 | PAT | Profit After Tax |
| 24 | Adjusted PAT | Adjusted profit is the best estimate of the true profit |
| 25 | Retained Earning | Earnings after adjusting to dividen |
| 26 | CP | Commercial paper , a short-term debt instrument to meet short-term liabilities. |
| 27 | Revenue earnings in forex | Revenue earned in foreign currency |
| 28 | Revenue expenses in forex | Expenses due to foreign currency transactions |
| 29 | Capital expenses in forex | Long term investment in forex |
| 30 | Book Value (Unit Curr) | Net asset value |
| 31 | Book Value (Adj.) (Unit Curr) | Book value adjusted to reflect asset's true fair market value |
| 32 | Market Capitalisation | Product of the total number of a company's outstanding shares and the current market price of one share |
| 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 |
| 34 | Cash Flow From Operating Activities | Use of cash from ongoing regular business activities |
| 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 |
| 36 | Cash Flow From Financing Activities | Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends) |
| 37 | ROG-Net Worth (%) | Rate of Growth - Networth |
| 38 | ROG-Capital Employed (%) | Rate of Growth - Capital Employed |
| 39 | ROG-Gross Block (%) | Rate of Growth - Gross Block |
| 40 | ROG-Gross Sales (%) | Rate of Growth - Gross Sales |
| 41 | ROG-Net Sales (%) | Rate of Growth - Net Sales |
| 42 | ROG-Cost of Production (%) | Rate of Growth - Cost of Production |
| 43 | ROG-Total Assets (%) | Rate of Growth - Total Assets |
| 44 | ROG-PBIDT (%) | Rate of Growth- PBIDT |
| 45 | ROG-PBDT (%) | Rate of Growth- PBDT |
| 46 | ROG-PBIT (%) | Rate of Growth- PBIT |
| 47 | ROG-PBT (%) | Rate of Growth- PBT |
| 48 | ROG-PAT (%) | Rate of Growth- PAT |
| 49 | ROG-CP (%) | Rate of Growth- CP |
| 50 | ROG-Revenue earnings in forex (%) | Rate of Growth - Revenue earnings in forex |
| 51 | ROG-Revenue expenses in forex (%) | Rate of Growth - Revenue expenses in forex |
| 52 | ROG-Market Capitalisation (%) | Rate of Growth - Market Capitalisation |
| 53 | Current Ratio[Latest] | Liquidity ratio, company's ability to pay short-term obligations or those due within one year |
| 54 | Fixed Assets Ratio[Latest] | Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating |
| 55 | Inventory Ratio[Latest] | Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company |
| 56 | Debtors Ratio[Latest] | Measures how quickly cash debtors are paying back to the company |
| 57 | Total Asset Turnover Ratio[Latest] | The value of a company's revenues relative to the value of its assets |
| 58 | Interest Cover Ratio[Latest] | Determines how easily a company can pay interest on its outstanding debt |
| 59 | PBIDTM (%)[Latest] | Profit before Interest Depreciation and Tax Margin |
| 60 | PBITM (%)[Latest] | Profit Before Interest Tax Margin |
| 61 | PBDTM (%)[Latest] | Profit Before Depreciation Tax Margin |
| 62 | CPM (%)[Latest] | Cost per thousand (advertising cost) |
| 63 | APATM (%)[Latest] | After tax profit margin |
| 64 | Debtors Velocity (Days) | Average days required for receiving the payments |
| 65 | Creditors Velocity (Days) | Average number of days company takes to pay suppliers |
| 66 | Inventory Velocity (Days) | Average number of days the company needs to turn its inventory into sales |
| 67 | Value of Output/Total Assets | Ratio of Value of Output (market value) to Total Assets |
| 68 | Value of Output/Gross Block | Ratio of Value of Output (market value) to Gross Block |

**Exploratory Data Analysis:**

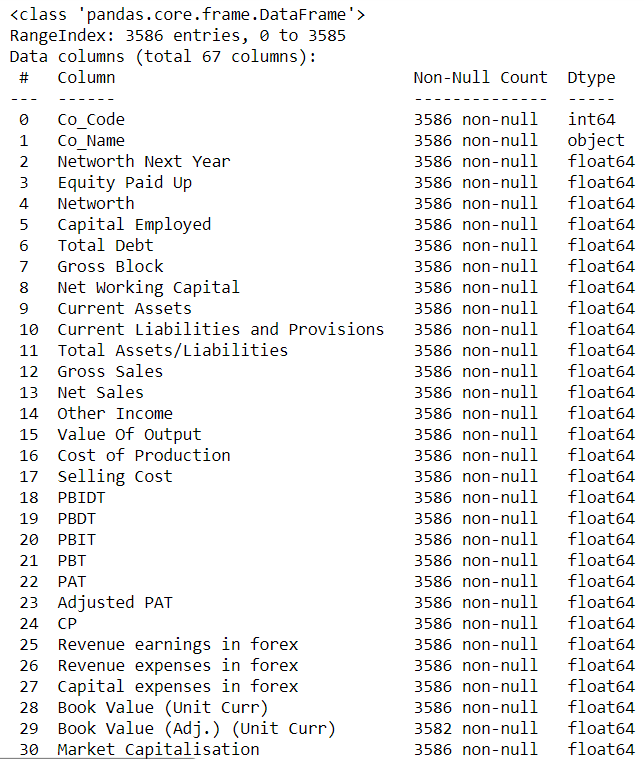
The head of the data looks like this, many columns are not shown here due to space constraint:

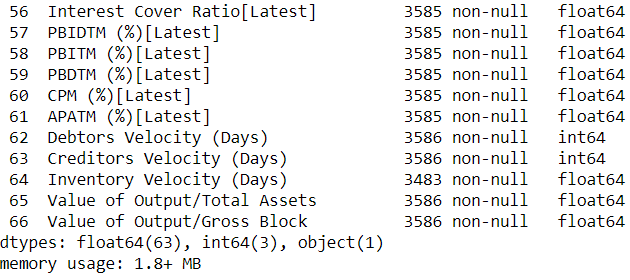


The shape of the data is as follows:



We then look at the info of the data which contains the name of each column, the number of non null entries and the data type:



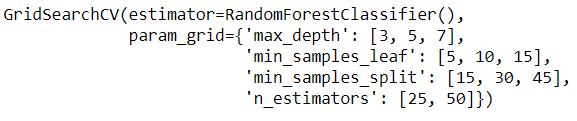
 

It can be inferred from the above output that there are some missing values and all the variables have a numerical data type of Float or Int except for the column: Co\_Name which is the name of the company.

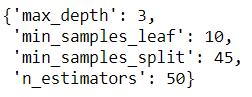
* A binary target variable as ‘Default’ is created using the Networth\_Next\_Year field such that if the Networth is positive, then the company has not defaulted (0) and if not, the company has defaulted (1).
* The missing values are treated using the median with the help of the SimpleImputer function.
* Outliers are present in almost every column and they are treated with the upper range and the lower ranges which correspond to third quartile + (1.5 \* IQR) and the first quartile – (1.5 \* IQR) respectively.
* High positive and negative correlations are observed.
* The data is then split into train and test with the test size being 33% and the train size of 67%, the random state is 42 and stratify as y since there is an imbalance in the target variable: Default.

**Random Forest Classifier:**

* From Scikit learn, we import the Random Forest Classifier and also the grid search cross validation.
* We use the following hyperparamters in the grid search CV: maximum depth as (3,5,7). This is defined as the longest path between a root node and a leaf node. Min samples leaf as (5,10,15) and this corresponds to the minimum number of samples that should be present in a leaf after the node is split. Min samples split as (15, 30, 45) this is the minimum number of observations required in a node in order to split the same. N estimators as (25,30) which is the number of trees used in building the forest.

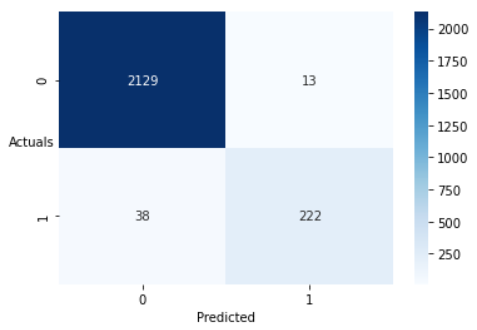


* The values for the best grid is found to be as:

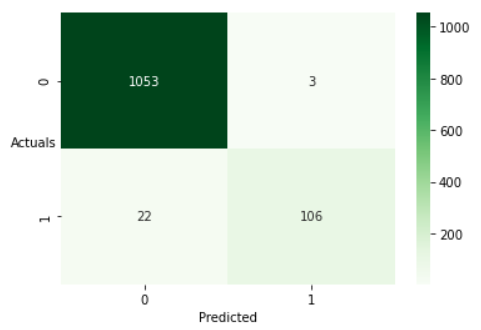


* Confusion matrix on the train and test data are as follows:

Train data:

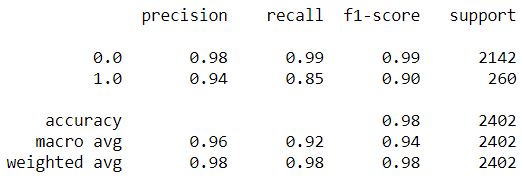


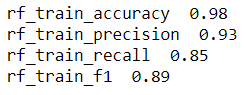
Test data:



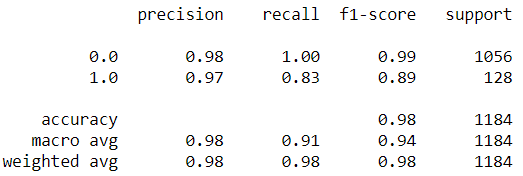
* Classification report:

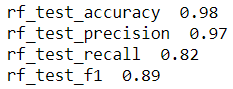
Train data:



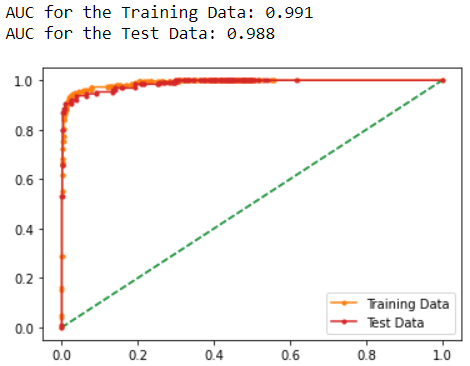


Test data:





* AUC and ROC for the train and test data:

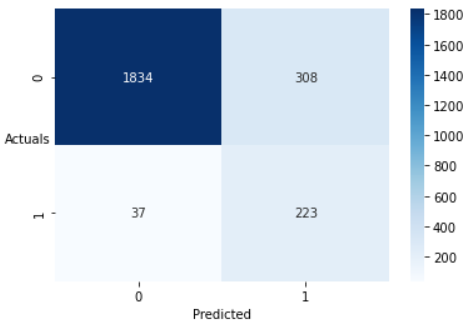


* The false positives in the confusion matrix are extremely low for both train and test. The false negatives are high on the train data and not low on the test data as well.
* Model accuracy on the train and test data are equal and is very good.
* For our business objective, the Recall metric is of great importance (how many of the actual defaulters is my model able to identify). It is 83% on the test data.
* Precision is also an important metric in our case as it will save us from falsely charging high interest rates on companies that are predicted to default but in actuality, they don’t. This means, that valuable companies might choose to leave the bank due to the high interest rates that might have been imposed on them, incorrectly though. In our model, the precision on the test data is 97% which more than that of the train data 94% which is very good.
* The AUC and ROC for the train and test data is extremely good.

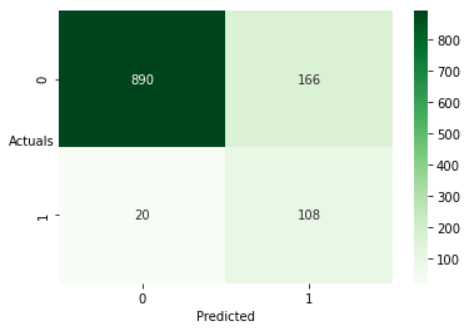
**Linear Discriminant Analysis model:**

* The Linear Discriminant Analysis model is then imported and the optimal threshold value is then calculated that will best separate the defaulters and the non-defaulters.
* It is determined as 0.11
* Confusion matrix

Train data:

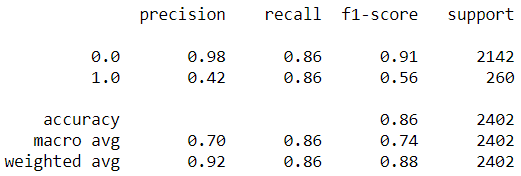


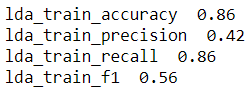
Test data:



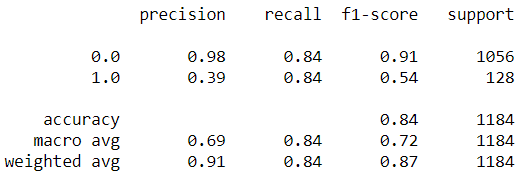
* Classification report:

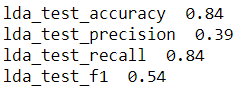
Train data:



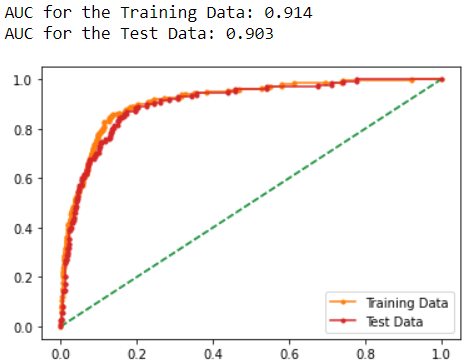


Test data:





* AUC and ROC for the train and test data:

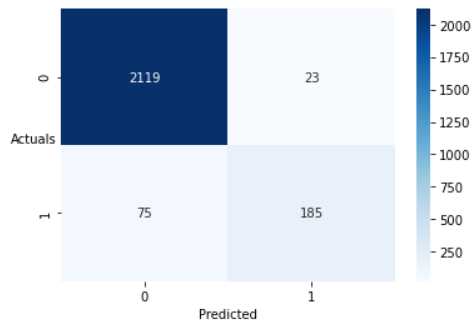


* The Recall values are the same as in the case of our Random Forest model but the model precision is very low. The model accuracy is also low. The model does not overfit but does not qualify to be deployed into production.
* The model has very high false positives in the confusion matrix for both the train and test data. The false negatives are also too much in number that can prove to be very costly for the bank.
* The AUC and ROC for the train and test data is actually not bad which is very surprising when considering the metrics from the classification report.

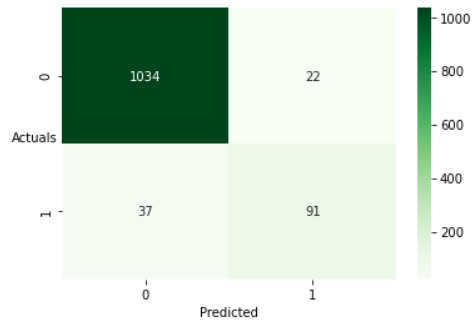
**Logistic Regression model:**

* We import the Logistic Regression model from the Scikit library and use the solver as ‘liblinear’ since it is good for small datasets like ours.
* It uses both L1 (Lasso) and L2 (Ridge) regularization technique. The Lasso regularization (L1 regularization) reduces the less important feature’s coefficients to zero, thus removing some features altogether. This is useful when we have a large number of features as in our case. The Ridge regularization (L2 regularization) helps to avoid the over-fitting issue.
* Confusion matrix:

Train data:

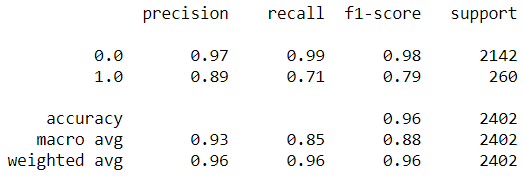


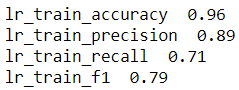
Test data:



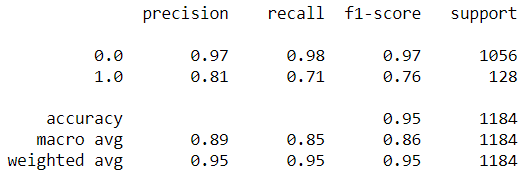
* Classification Report:

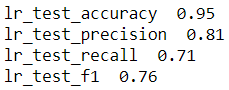
Train data:



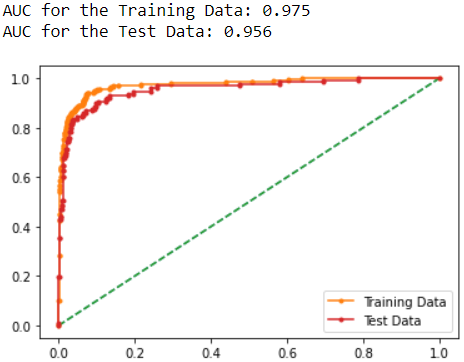


Test data:





* AUC and ROC for the train and test data:



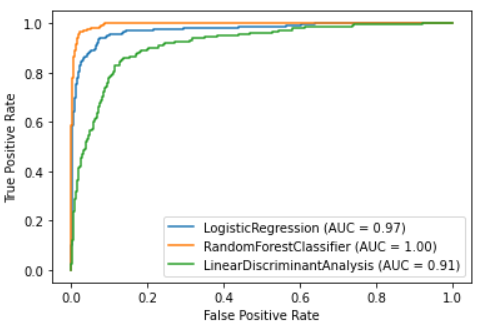
* This model has low false positives in the confusion matrix but has relatively high false negatives.
* The model accuracy is good but the recall values aren’t. Recall is low.
* There is a considerable difference in the precision values between the train and test data but although does not qualify to be branded as an overfit model.
* The AUC and ROC values are actually good for this model.

### Comparison of performance metrics from the 3 models:

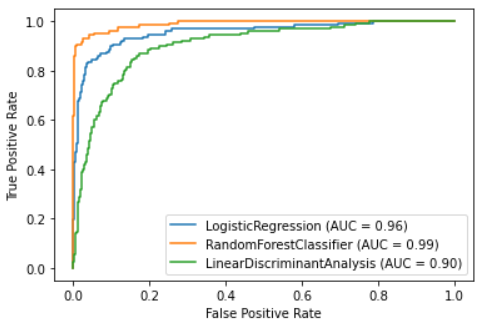
### 

#### Using AUC - ROC Curve:

On train data:



On Test data:

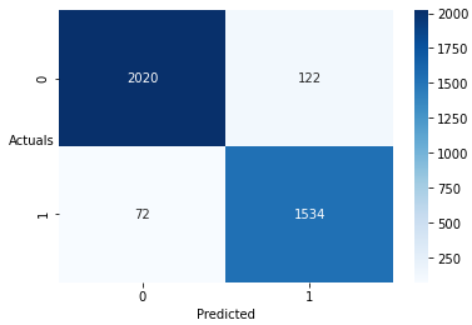


* The Random Forest model has far superior numbers in all the metrics except for in Recall. The Recall numbers for the LDA model are the highest in both the train and test data.
* When comparing different models, it is best advisable to consider the AUC-ROC scores as a deciding factor.
* The Random Forest models beats hands down all the other models and is considered as the best model for solving our problem statement.

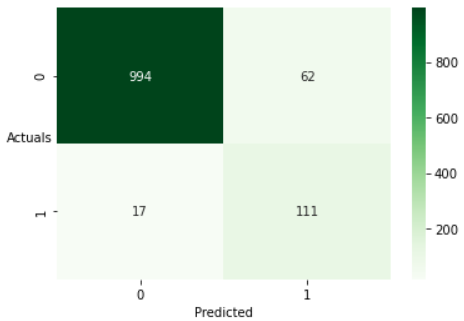
**Applying SMOTE:**

* We apply Synthetic Monority Oversampling Technique to treat the class imbalance in the target variable: default. We use the random state as 33 and the sampling strategy as 0.75 which means that for every 4 observations of the majority class (default = 0), it is going to add 3 observations of the minority class (default = 1).
* We use the Logistic Regression model with the parameters stated earlier (with solver as ‘liblinear) and look at the confusion matrix:

On Train data (resampled data):

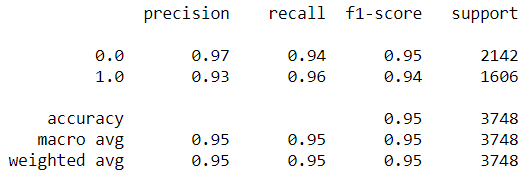


On Test data:

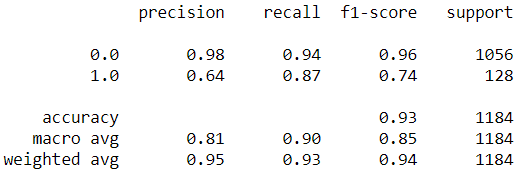


* Classification Report:

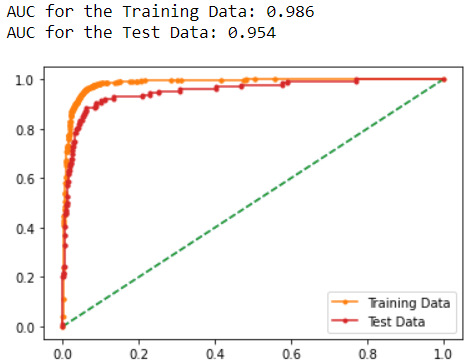
On Train data:



On Test data:



* AUC and ROC for the train and test data:



* The False negatives on the test data have reduced to the lowest ever in any of our models till now which is a very good sign. The False positives are actually very high on the test data which is a cause for concern in this model.
* On the train data, the numbers in the classification report look amazing, better than the Random Forest model on the recall metric front. But on the Test data, the precision numbers have gone for a toss. The recall numbers are still higher the Random Forest model. The accuracy is also pretty good on the train and test data.
* Looking at the AUC-ROC curve for the train and test data, it is very good. The model also does not overfit.
* The Random Forest model is still the best model because of its better precision scores in the test data. The business has to take a call on comparing the cost incurred due to misclassification of the positives and to have a good recall (i.e. correctly identifying the actual defaulters).

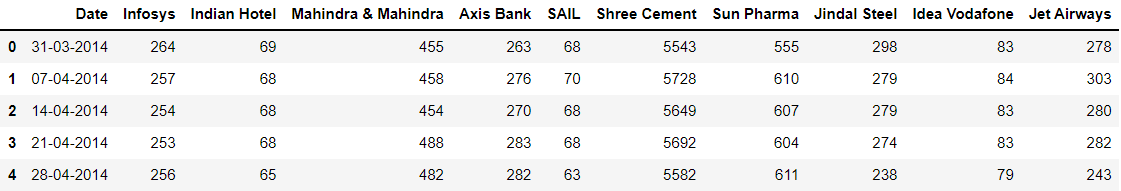
**Analyzing Market Risk and Returns:**

**Problem summary:**

We are required to analyze the data of stock prices of 10 different Indian stocks for period of 6 years with a weekly frequency.

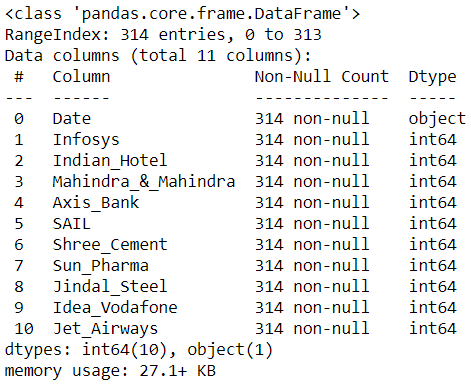
**Exploratory data analysis:**

The head of the data looks like this:



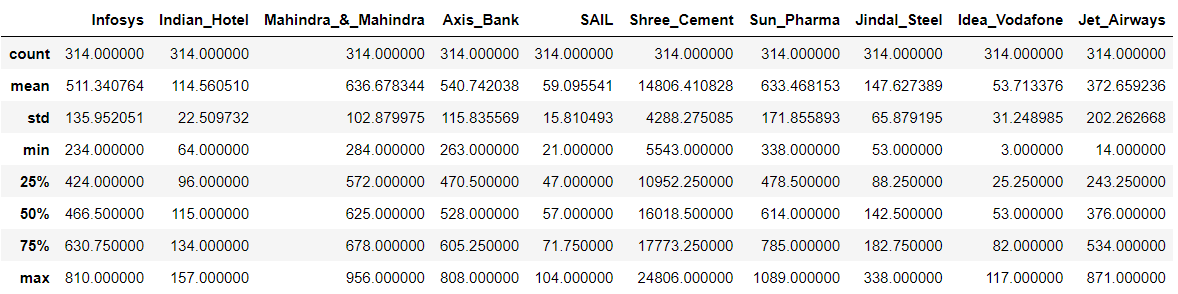
We then replace the spaces in the column names with an underscore to facilitate better readability.

We then look at the info of the data to check the data types:



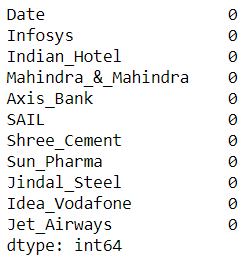
The ‘Date’ column is read as an object datatype which then needs to be converted to the date time datatype. The other variables are read as an integer datatype.

We then look at the descriptive statistics of the data:



It can be inferred that for all the companies the data is approximately symmetric.

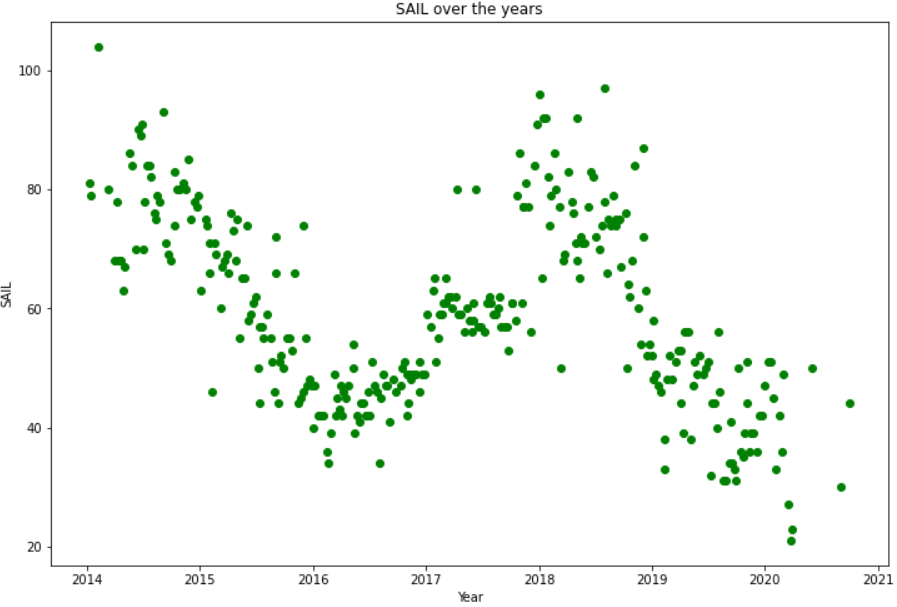
We then look for the presence of missing values and see that there are none:



**Stock price graph:**

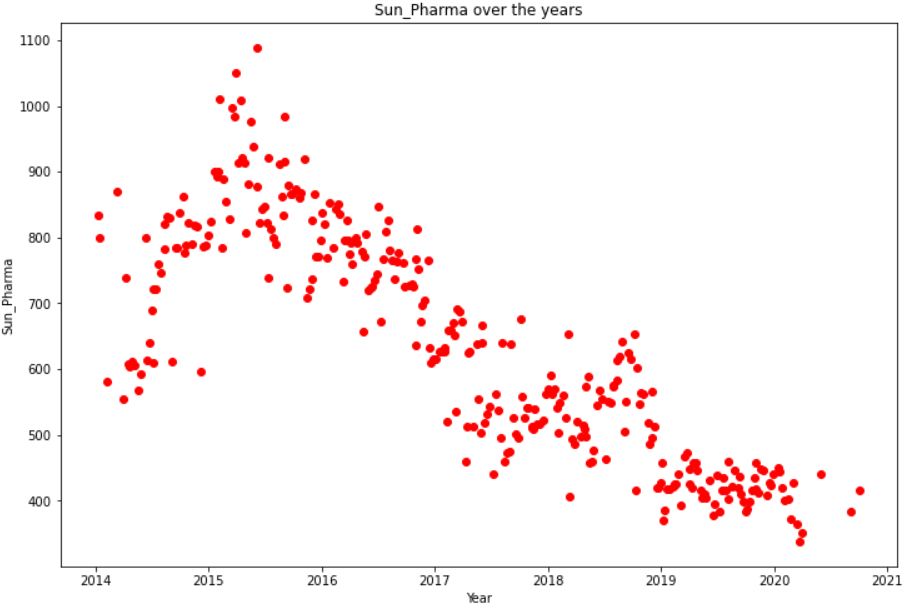
We then look at the behaviour of the stock price over time for

SAIL:



* The stock prices have been varying with time. A definite trend can be seen which is varying. The stock prices were highest in the start of 2014 and then a steep and steady decline was observed until the mid of 2016. Again, after this the stock prices increased to more than 96 in the mid of 2018 post which the prices again see a decline until the present times.

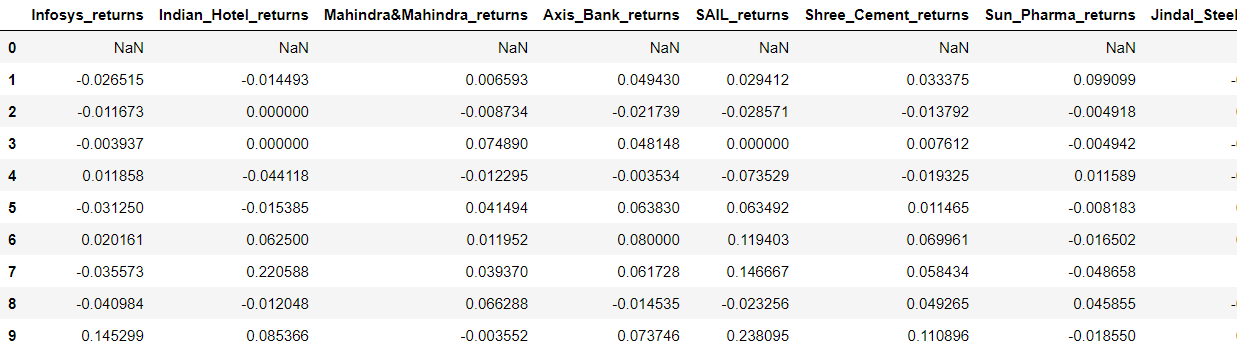
Sun\_Pharma:



* The stock prices of Sun\_Pharma rose from around an average of 700 in the year 2014 to 1100 in the mid of 2015. Post this there has been a constant decline until the present times. There also seems to be a short increase in the stock prices towards the end of 2018.

**Calculating stock returns:**

We have used the percentage change between the current week and the previous week to calculate the stock returns for each company. The head of the first 10 rows of data looks like this:

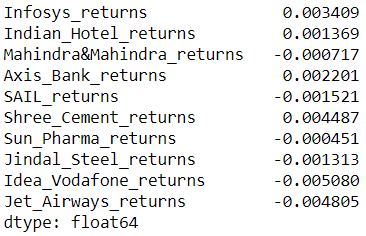


The first row is NaN because it does not have a previous value with which a percentage difference can be calculated. A lot of negative returns are observed for a lot of companies in the head of the data.

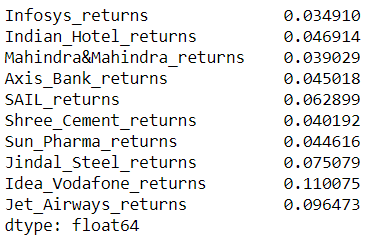
**Calculating Stock Means:**

We then calculate the means of stock returns for each company.

Negative mean returns are observed for Mahindra & Mahindra, SAIL, Sun Pharma, Jindal Steel, Idea Vodafone and Jet Airways which is a bad sign.

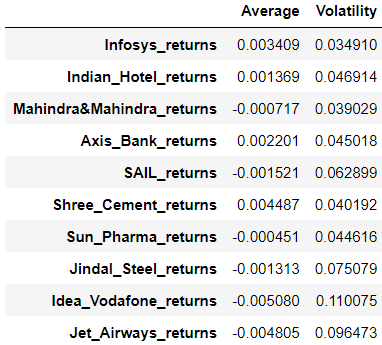


**Calculating Stock Standard Deviation:**



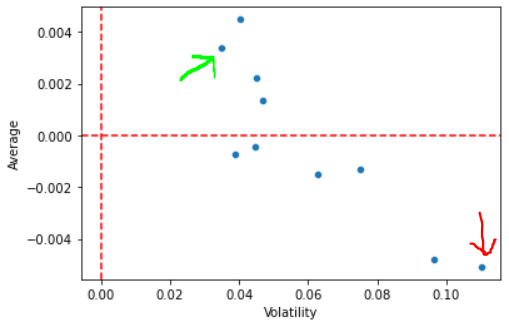
Idea Vodafone has the highest standard deviation on stock returns followed by Jet Airways.

We then look at the stock means and standard deviations for different companies together:

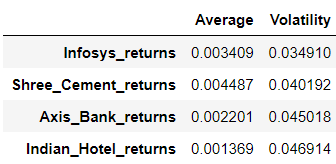


The volatility of Idea Vodafone is the highest and with negative average returns.

**Plot of Stock means vs Standard deviation:**



* The company that is indicated by the green coloured arrow in the above plot is the best as it has the lowest volatility with a decently good average return.
* The company that is indicated by the red coloured arrow in the above plot is the worst as it has the highest volatility with the highest negative returns.
* We shall also sort the stocks that have average returns as positive and the volatility in the ascending order:



These are the best companies to invest in from our data.

The best one being Infosys with the lowest volatility and the second highest returns.