**FINANCE AND RISK ANALYTICS**

**MILESTONE-1**

[**MILESTONE-2**](#Milestone) **(crtl+click)**

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
|  | Understanding data | 3 |
|  | Outlier Treatment | 9 |
|  | Null values Treatment | 9 |
|  | Univariate and Bivariate Analysis | 10 |
|  | Train test split | 12 |
|  | Choose significant variables | 12 |
|  | Build logistic model | 14 |
|  | Validate the Model on Test Dataset and state the performance matrices | 21 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
|  | Outlier before treatment | 9 |
|  | Outlier after treatment | 9 |
|  | Correlation plot | 10 |
|  | Histogram of dependent variable vs independent variables | 10 |
|  | Boxplot of default and predicted train data set | 18 |
|  | Confusion matrix for train data set with ut off 0.05 | 18 |
|  | Confusion matrix for test data set with ut off 0.05 | 19 |
|  | Confusion matrix for train data set with ut off 0.05 | 20 |
|  | Confusion matrix for test data set with ut off 0.05 | 21 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
|  | Head of the Credit Risk data | 3 |
|  | Basic information of Credit Risk data | 3 |
|  | Null values in Credit Risk data | 5 |
|  | Statistical information of Credit Risk data | 5 |
|  | Null value after treatment | 9 |
|  | VIF values | 12 |
|  | MODEL-1 summary | 14 |
|  | MODEL-2 summary | 16 |
|  | MODEL-3 summary | 17 |
|  | Classificationion report for training set with cut off 0.05 | 19 |
|  | Classificationion report for test data set with cut off 0.05 | 19 |
|  | Classificationion report for training set with cut off 0.05 | 20 |
|  | Classificationion report for test data set with cut off 0.05 | 21 |

**PROBLEM:**

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, reducing its chances of getting credit in the future and may have to pay higher interest on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a company’s financial statement 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.

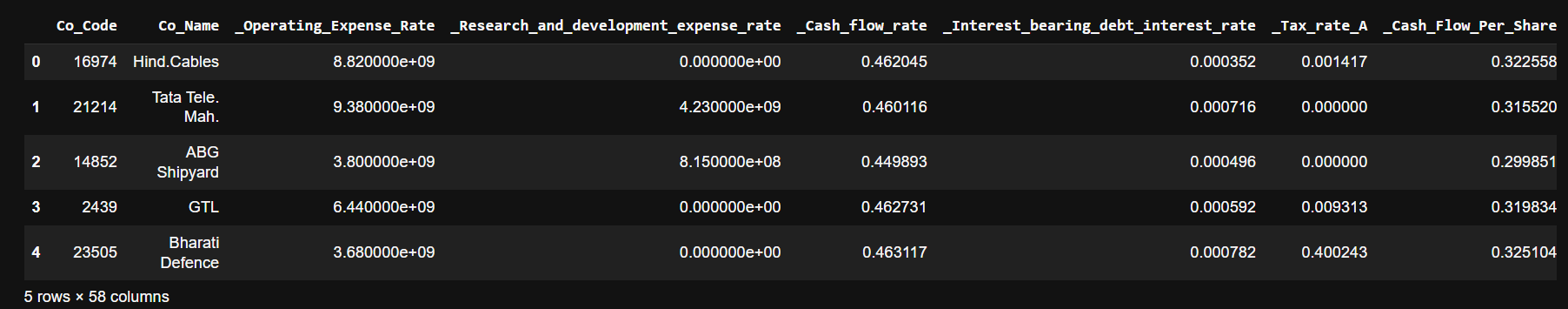
**Import the data:**

To check the data, we must import all the libraries and load the dataset.

We first use the read\_excel function from pandas to read the excel file CompData.xlsx.

First, see the first five rows of the data. This helps to know about the data and columns.

Table 1: Head of the Credit Risk data



Next, check the basic information of the dataset, this can be done by using info() function.

Table 2: Basic information of Credit Risk data

|  |  |  |  |
| --- | --- | --- | --- |
| \ | column | non null count | Dtype |
| 0 | Co\_Code | 2058 | int64 |
| 1 | Co\_Name | 2058 | object |
| 2 | \_Operating\_Expense\_Rate | 2058 | float64 |
| 3 | \_Research\_and\_development\_expense\_rate | 2058 | float64 |
| 4 | \_Cash\_flow\_rate | 2058 | float64 |
| 5 | \_Interest\_bearing\_debt\_interest\_rate | 2058 | float64 |
| 6 | \_Tax\_rate\_A | 2058 | float64 |
| 7 | \_Cash\_Flow\_Per\_Share | 1891 | float64 |
| 8 | \_Per\_Share\_Net\_profit\_before\_tax\_Yuan\_ | 2058 | float64 |
| 9 | \_Realized\_Sales\_Gross\_Profit\_Growth\_Rate | 2058 | float64 |
| 10 | \_Operating\_Profit\_Growth\_Rate | 2058 | float64 |
| 11 | \_Continuous\_Net\_Profit\_Growth\_Rate | 2058 | float64 |
| 12 | \_Total\_Asset\_Growth\_Rate | 2058 | float64 |
| 13 | \_Net\_Value\_Growth\_Rate | 2058 | float64 |
| 14 | \_Total\_Asset\_Return\_Growth\_Rate\_Ratio | 2058 | float64 |
| 15 | \_Cash\_Reinvestment\_perc | 2058 | float64 |
| 16 | \_Current\_Ratio | 2058 | float64 |
| 17 | \_Quick\_Ratio | 2058 | float64 |
| 18 | \_Interest\_Expense\_Ratio | 2058 | float64 |
| 19 | \_Total\_debt\_to\_Total\_net\_worth | 2037 | float64 |
| 20 | \_Long\_term\_fund\_suitability\_ratio\_A | 2058 | float64 |
| 21 | \_Net\_profit\_before\_tax\_to\_Paid\_in\_capital | 2058 | float64 |
| 22 | \_Total\_Asset\_Turnover | 2058 | float64 |
| 23 | \_Accounts\_Receivable\_Turnover | 2058 | float64 |
| 24 | \_Average\_Collection\_Days | 2058 | float64 |
| 25 | \_Inventory\_Turnover\_Rate\_times | 2058 | float64 |
| 26 | \_Fixed\_Assets\_Turnover\_Frequency | 2058 | float64 |
| 27 | \_Net\_Worth\_Turnover\_Rate\_times | 2058 | float64 |
| 28 | \_Operating\_profit\_per\_person | 2058 | float64 |
| 29 | \_Allocation\_rate\_per\_person | 2058 | float64 |
| 30 | \_Quick\_Assets\_to\_Total\_Assets | 2058 | float64 |
| 31 | \_Cash\_to\_Total\_Assets | 1962 | float64 |
| 32 | \_Quick\_Assets\_to\_Current\_Liability | 2058 | float64 |
| 33 | \_Cash\_to\_Current\_Liability | 2058 | float64 |
| 34 | \_Operating\_Funds\_to\_Liability | 2058 | float64 |
| 35 | \_Inventory\_to\_Working\_Capital | 2058 | float64 |
| 36 | \_Inventory\_to\_Current\_Liability | 2058 | float64 |
| 37 | \_Long\_term\_Liability\_to\_Current\_Assets | 2058 | float64 |
| 38 | \_Retained\_Earnings\_to\_Total\_Assets | 2058 | float64 |
| 39 | \_Total\_income\_to\_Total\_expense | 2058 | float64 |
| 40 | \_Total\_expense\_to\_Assets | 2058 | float64 |
| 41 | \_Current\_Asset\_Turnover\_Rate | 2058 | float64 |
| 42 | \_Quick\_Asset\_Turnover\_Rate | 2058 | float64 |
| 43 | \_Cash\_Turnover\_Rate | 2058 | float64 |
| 44 | \_Fixed\_Assets\_to\_Assets | 2058 | float64 |
| 45 | \_Cash\_Flow\_to\_Total\_Assets | 2058 | float64 |
| 46 | \_Cash\_Flow\_to\_Liability | 2058 | float64 |
| 47 | \_CFO\_to\_Assets | 2058 | float64 |
| 48 | \_Cash\_Flow\_to\_Equity | 2058 | float64 |
| 49 | \_Current\_Liability\_to\_Current\_Assets | 2044 | float64 |
| 50 | \_Liability\_Assets\_Flag | 2058 | int64 |
| 51 | \_Total\_assets\_to\_GNP\_price | 2058 | float64 |
| 52 | \_No\_credit\_Interval | 2058 | float64 |
| 53 | \_Degree\_of\_Financial\_Leverage\_DFL | 2058 | float64 |
| 54 | \_Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT | 2058 | float64 |
| 55 | \_Net\_Income\_Flag | 2058 | int64 |
| 56 | \_Equity\_to\_Liability | 2058 | float64 |
| 57 | Default | 2058 | float64 |

From the above table, we can understand that there are 2058 and 58 columns which are of float and int data types.

Table 3: Null values in Credit Risk data

The null values in the data are:

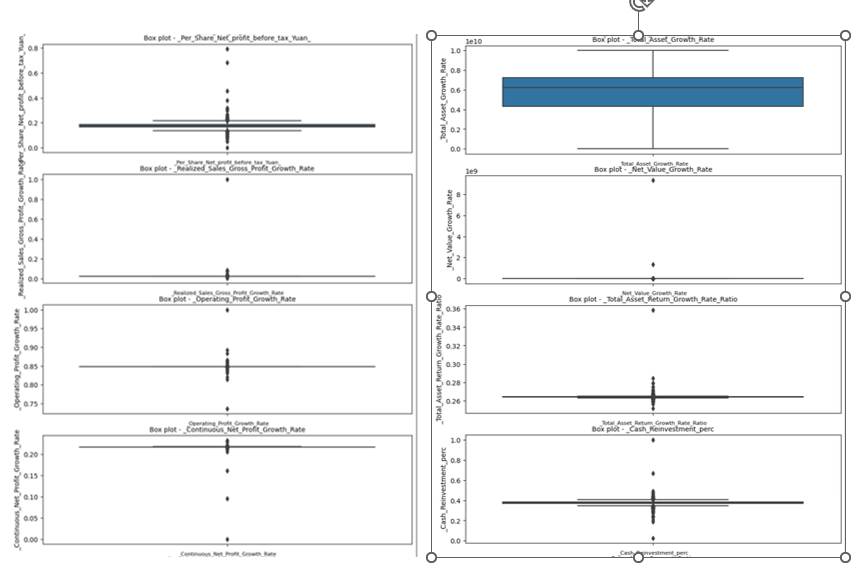
|  |  |
| --- | --- |
| \_Cash\_Flow\_Per\_Share | 167 |
| \_Total\_debt\_to\_Total\_net\_worth | 21 |
| \_Cash\_to\_Total\_Assets | 96 |
| \_Current\_Liability\_to\_Current\_Assets | 14 |

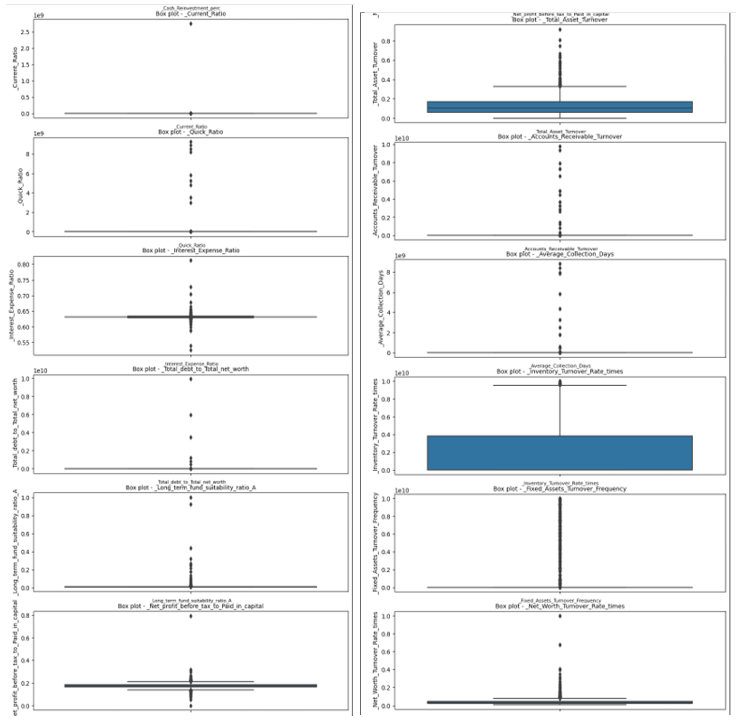
Table 4: Statistical information of Credit Risk data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Co\_Code** | 2058 | 1.76E+04 | 2.19E+04 | 4 | 3.67E+03 | 6.24E+03 | 2.43E+04 | 7.25E+04 |
| **\_Operating\_Expense\_Rate** | 2058 | 2.05E+09 | 3.25E+09 | 0.0001 | 1.58E-04 | 3.33E-04 | 4.11E+09 | 9.98E+09 |
| **\_Research\_and\_development\_expense\_rate** | 2058 | 9.32E+08 | 1.44E+09 | 0 | 0.00E+00 | 1.99E-04 | 1.55E+09 | 3.88E+09 |
| **\_Cash\_flow\_rate** | 2058 | 4.64E-01 | 7.46E-03 | 0.44814 | 4.60E-01 | 4.63E-01 | 4.68E-01 | 4.80E-01 |
| **\_Interest\_bearing\_debt\_interest\_rate** | 2058 | 4.84E-04 | 2.96E-04 | 0 | 2.76E-04 | 4.54E-04 | 6.63E-04 | 1.24E-03 |
| **\_Tax\_rate\_A** | 2058 | 1.10E-01 | 1.35E-01 | 0 | 0.00E+00 | 3.71E-02 | 2.16E-01 | 5.40E-01 |
| **\_Cash\_Flow\_Per\_Share** | 2058 | 3.20E-01 | 9.04E-03 | 0.30118 | 3.16E-01 | 3.21E-01 | 3.25E-01 | 3.40E-01 |
| **\_Per\_Share\_Net\_profit\_before\_tax\_Yuan\_** | 2058 | 1.76E-01 | 1.80E-02 | 0.13768 | 1.67E-01 | 1.76E-01 | 1.86E-01 | 2.15E-01 |
| **\_Realized\_Sales\_Gross\_Profit\_Growth\_Rate** | 2058 | 2.21E-02 | 9.34E-05 | 0.02192 | 2.21E-02 | 2.21E-02 | 2.22E-02 | 2.23E-02 |
| **\_Operating\_Profit\_Growth\_Rate** | 2058 | 8.48E-01 | 1.43E-04 | 0.84776 | 8.48E-01 | 8.48E-01 | 8.48E-01 | 8.48E-01 |
| **\_Continuous\_Net\_Profit\_Growth\_Rate** | 2058 | 2.18E-01 | 4.70E-05 | 0.21751 | 2.18E-01 | 2.18E-01 | 2.18E-01 | 2.18E-01 |
| **\_Total\_Asset\_Growth\_Rate** | 2058 | 5.29E+09 | 2.91E+09 | 0 | 4.32E+09 | 6.23E+09 | 7.22E+09 | 9.98E+09 |
| **\_Net\_Value\_Growth\_Rate** | 2058 | 4.63E-04 | 5.33E-05 | 0.00036 | 4.36E-04 | 4.55E-04 | 4.88E-04 | 5.67E-04 |
| **\_Total\_Asset\_Return\_Growth\_Rate\_Ratio** | 2058 | 2.64E-01 | 5.50E-04 | 0.26288 | 2.64E-01 | 2.64E-01 | 2.64E-01 | 2.65E-01 |
| **\_Cash\_Reinvestment\_perc** | 2058 | 3.78E-01 | 1.42E-02 | 0.34849 | 3.71E-01 | 3.79E-01 | 3.86E-01 | 4.08E-01 |
| **\_Current\_Ratio** | 2058 | 1.07E-02 | 6.08E-03 | 0 | 6.57E-03 | 8.95E-03 | 1.35E-02 | 2.39E-02 |
| **\_Quick\_Ratio** | 2058 | 6.70E-03 | 5.12E-03 | 0 | 2.95E-03 | 5.28E-03 | 8.90E-03 | 1.78E-02 |
| **\_Interest\_Expense\_Ratio** | 2058 | 6.31E-01 | 1.20E-03 | 0.62891 | 6.31E-01 | 6.31E-01 | 6.32E-01 | 6.33E-01 |
| **\_Total\_debt\_to\_Total\_net\_worth** | 2058 | 9.33E-03 | 7.02E-03 | 0 | 3.94E-03 | 7.27E-03 | 1.30E-02 | 2.66E-02 |
| **\_Long\_term\_fund\_suitability\_ratio\_A** | 2058 | 5.97E-03 | 1.10E-03 | 0.00413 | 5.16E-03 | 5.52E-03 | 6.42E-03 | 8.30E-03 |
| **\_Net\_profit\_before\_tax\_to\_Paid\_in\_capital** | 2058 | 1.75E-01 | 1.72E-02 | 0.13799 | 1.66E-01 | 1.75E-01 | 1.84E-01 | 2.12E-01 |
| **\_Total\_Asset\_Turnover** | 2058 | 1.24E-01 | 8.35E-02 | 0 | 6.15E-02 | 1.03E-01 | 1.68E-01 | 3.28E-01 |
| **\_Accounts\_Receivable\_Turnover** | 2058 | 1.48E-03 | 1.01E-03 | 0 | 7.45E-04 | 1.08E-03 | 1.85E-03 | 3.52E-03 |
| **\_Average\_Collection\_Days** | 2058 | 6.42E-03 | 3.94E-03 | 0 | 3.58E-03 | 6.00E-03 | 8.64E-03 | 1.62E-02 |
| **\_Inventory\_Turnover\_Rate\_times** | 2058 | 2.03E+09 | 3.07E+09 | 0 | 1.91E-04 | 1.91E+07 | 3.82E+09 | 9.54E+09 |
| **\_Fixed\_Assets\_Turnover\_Frequency** | 2058 | 5.71E-03 | 8.63E-03 | 0 | 2.28E-04 | 6.00E-04 | 8.42E-03 | 2.07E-02 |
| **\_Net\_Worth\_Turnover\_Rate\_times** | 2058 | 3.51E-02 | 1.97E-02 | 0.00887 | 2.05E-02 | 2.87E-02 | 4.44E-02 | 8.02E-02 |
| **\_Operating\_profit\_per\_person** | 2058 | 3.96E-01 | 1.00E-02 | 0.37713 | 3.91E-01 | 3.95E-01 | 4.01E-01 | 4.15E-01 |
| **\_Allocation\_rate\_per\_person** | 2058 | 1.74E-02 | 1.69E-02 | 0 | 4.67E-03 | 1.06E-02 | 2.46E-02 | 5.44E-02 |
| **\_Quick\_Assets\_to\_Total\_Assets** | 2058 | 3.42E-01 | 2.10E-01 | 0 | 1.73E-01 | 3.06E-01 | 4.85E-01 | 9.51E-01 |
| **\_Cash\_to\_Total\_Assets** | 2058 | 6.72E-02 | 6.05E-02 | 0.00018 | 2.17E-02 | 4.56E-02 | 9.40E-02 | 2.03E-01 |
| **\_Quick\_Assets\_to\_Current\_Liability** | 2058 | 7.30E-03 | 5.22E-03 | 0 | 3.62E-03 | 5.97E-03 | 9.61E-03 | 1.86E-02 |
| **\_Cash\_to\_Current\_Liability** | 2058 | 5.25E-03 | 5.67E-03 | 0.0001 | 1.09E-03 | 2.68E-03 | 7.54E-03 | 1.72E-02 |
| **\_Operating\_Funds\_to\_Liability** | 2058 | 3.46E-01 | 1.56E-02 | 0.31305 | 3.38E-01 | 3.45E-01 | 3.54E-01 | 3.79E-01 |
| **\_Inventory\_to\_Working\_Capital** | 2058 | 2.77E-01 | 6.59E-04 | 0.27596 | 2.77E-01 | 2.77E-01 | 2.78E-01 | 2.79E-01 |
| **\_Inventory\_to\_Current\_Liability** | 2058 | 9.02E-03 | 7.92E-03 | 0 | 2.89E-03 | 6.78E-03 | 1.28E-02 | 2.75E-02 |
| **\_Long\_term\_Liability\_to\_Current\_Assets** | 2058 | 6.81E-03 | 8.77E-03 | 0 | 0.00E+00 | 2.59E-03 | 1.05E-02 | 2.62E-02 |
| **\_Retained\_Earnings\_to\_Total\_Assets** | 2058 | 9.33E-01 | 1.24E-02 | 0.90831 | 9.28E-01 | 9.35E-01 | 9.41E-01 | 9.61E-01 |
| **\_Total\_income\_to\_Total\_expense** | 2058 | 2.32E-03 | 2.23E-04 | 0.00182 | 2.19E-03 | 2.30E-03 | 2.43E-03 | 2.80E-03 |
| **\_Total\_expense\_to\_Assets** | 2058 | 2.68E-02 | 1.90E-02 | 0.00085 | 1.27E-02 | 2.09E-02 | 3.53E-02 | 6.92E-02 |
| **\_Current\_Asset\_Turnover\_Rate** | 2058 | 9.06E-04 | 1.13E-03 | 0 | 1.50E-04 | 2.46E-04 | 1.26E-03 | 2.93E-03 |
| **\_Quick\_Asset\_Turnover\_Rate** | 2058 | 2.57E+09 | 3.45E+09 | 0 | 1.51E-04 | 3.79E-04 | 5.79E+09 | 1.00E+10 |
| **\_Cash\_Turnover\_Rate** | 2058 | 2.65E+09 | 2.82E+09 | 0.0001 | 1.74E-03 | 1.73E+09 | 4.55E+09 | 9.99E+09 |
| **\_Fixed\_Assets\_to\_Assets** | 2058 | 2.73E-01 | 2.17E-01 | 0 | 9.65E-02 | 2.14E-01 | 4.15E-01 | 8.93E-01 |
| **\_Cash\_Flow\_to\_Total\_Assets** | 2058 | 6.44E-01 | 2.17E-02 | 0.60218 | 6.33E-01 | 6.43E-01 | 6.54E-01 | 6.85E-01 |
| **\_Cash\_Flow\_to\_Liability** | 2058 | 4.60E-01 | 4.66E-03 | 0.45109 | 4.57E-01 | 4.59E-01 | 4.62E-01 | 4.68E-01 |
| **\_CFO\_to\_Assets** | 2058 | 5.81E-01 | 5.33E-02 | 0.45747 | 5.50E-01 | 5.83E-01 | 6.12E-01 | 7.05E-01 |
| **\_Cash\_Flow\_to\_Equity** | 2058 | 3.15E-01 | 3.89E-03 | 0.30714 | 3.13E-01 | 3.15E-01 | 3.17E-01 | 3.22E-01 |
| **\_Current\_Liability\_to\_Current\_Assets** | 2058 | 3.47E-02 | 1.80E-02 | 0 | 2.19E-02 | 3.27E-02 | 4.39E-02 | 7.68E-02 |
| **\_Liability\_Assets\_Flag** | 2058 | 3.40E-03 | 5.82E-02 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.00E+00 |
| **\_Total\_assets\_to\_GNP\_price** | 2058 | 4.88E-03 | 5.23E-03 | 0 | 9.12E-04 | 2.48E-03 | 7.00E-03 | 1.61E-02 |
| **\_No\_credit\_Interval** | 2058 | 6.24E-01 | 7.71E-04 | 0.62225 | 6.23E-01 | 6.24E-01 | 6.24E-01 | 6.25E-01 |
| **\_Degree\_of\_Financial\_Leverage\_DFL** | 2058 | 2.69E-02 | 2.85E-04 | 0.0264 | 2.68E-02 | 2.68E-02 | 2.70E-02 | 2.74E-02 |
| **\_Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT** | 2058 | 5.66E-01 | 1.18E-03 | 0.56355 | 5.65E-01 | 5.65E-01 | 5.66E-01 | 5.68E-01 |
| **\_Net\_Income\_Flag** | 2058 | 1.00E+00 | 0.00E+00 | 1 | 1.00E+00 | 1.00E+00 | 1.00E+00 | 1.00E+00 |
| **\_Equity\_to\_Liability** | 2058 | 3.50E-02 | 1.94E-02 | 0.00395 | 2.04E-02 | 2.85E-02 | 4.34E-02 | 7.80E-02 |
| **Default** | 2058 | 1.07E-01 | 3.09E-01 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.00E+00 |

**Outlier Treatment**

Figure 1: Outlier before treatment



****

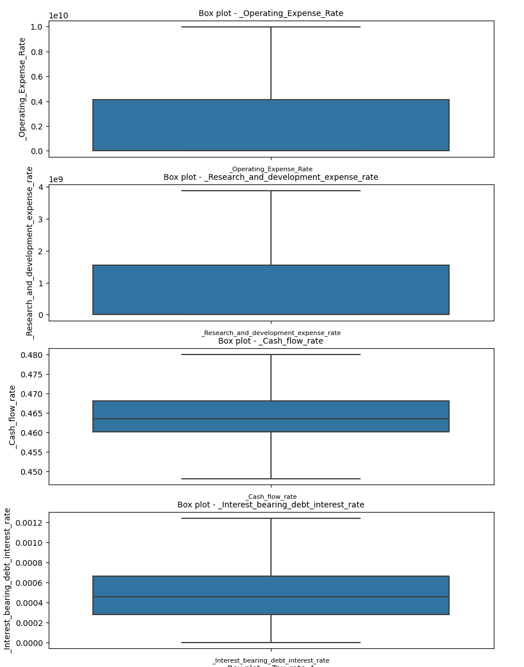
From the above boxplots we could observe that there are many outliers in the data.

Here we can observe outliers in all the variables.These are to be treated.

The data points which are greater than the maximum value are compressed to the maximum value.

The data points which are lesser than the minimum value are compressed to the minimum value.

Figure 2: Outliers after treatment



From the above the boxplots, we could see that all the outliers are treated.

**Null Values treatment:**

The null values are imputed with the median. This is done by Simple Imputer.

All the null values are treated.

Table 5: Null value after treatment

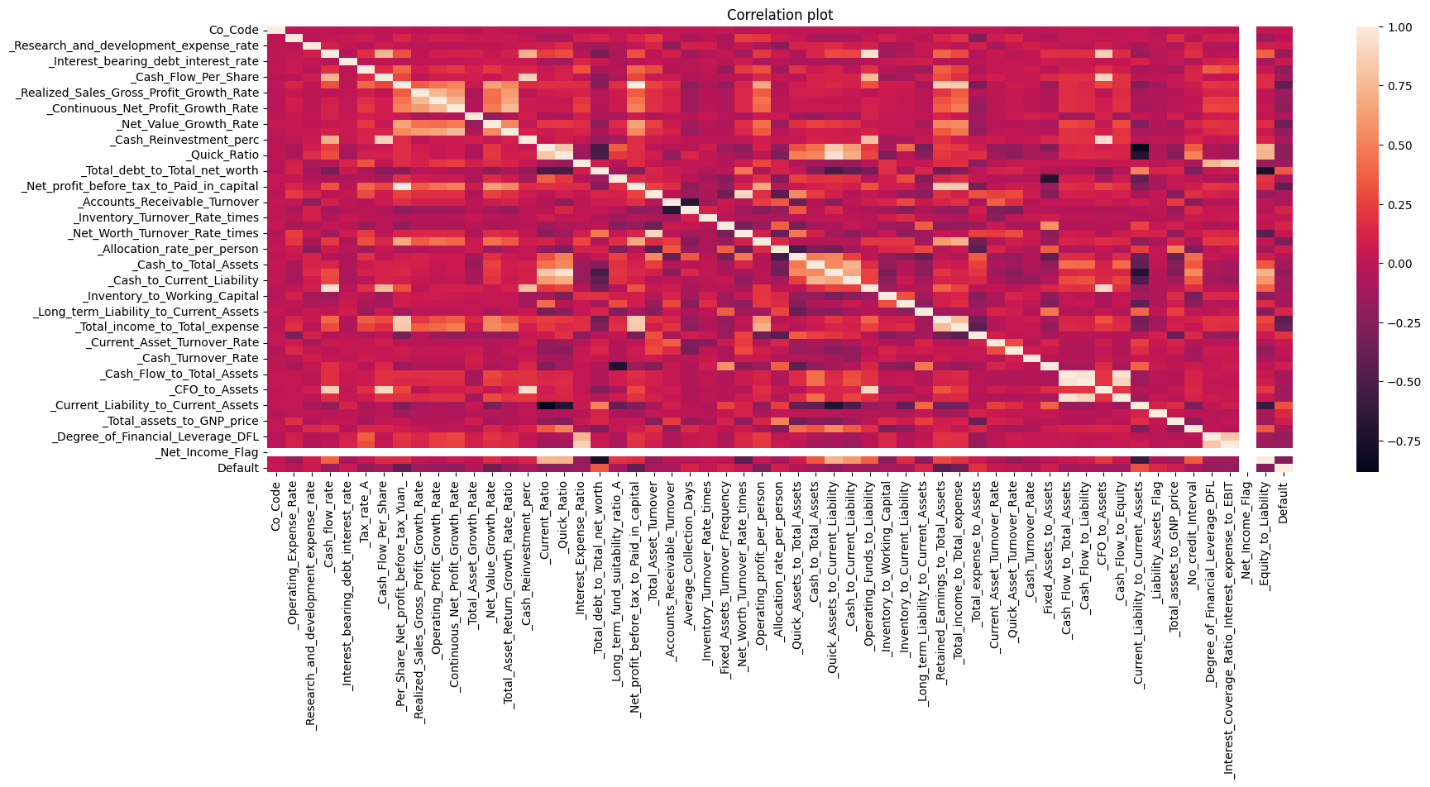
|  |  |
| --- | --- |
| \_Cash\_Flow\_Per\_Share | 0 |
| \_Total\_debt\_to\_Total\_net\_worth | 0 |
| \_Cash\_to\_Total\_Assets | 0 |
| \_Current\_Liability\_to\_Current\_Assets | 0 |

**Univariate and Bivariate Analysis:**

There are 1838 non defaulters (0) and 220 defaulters.

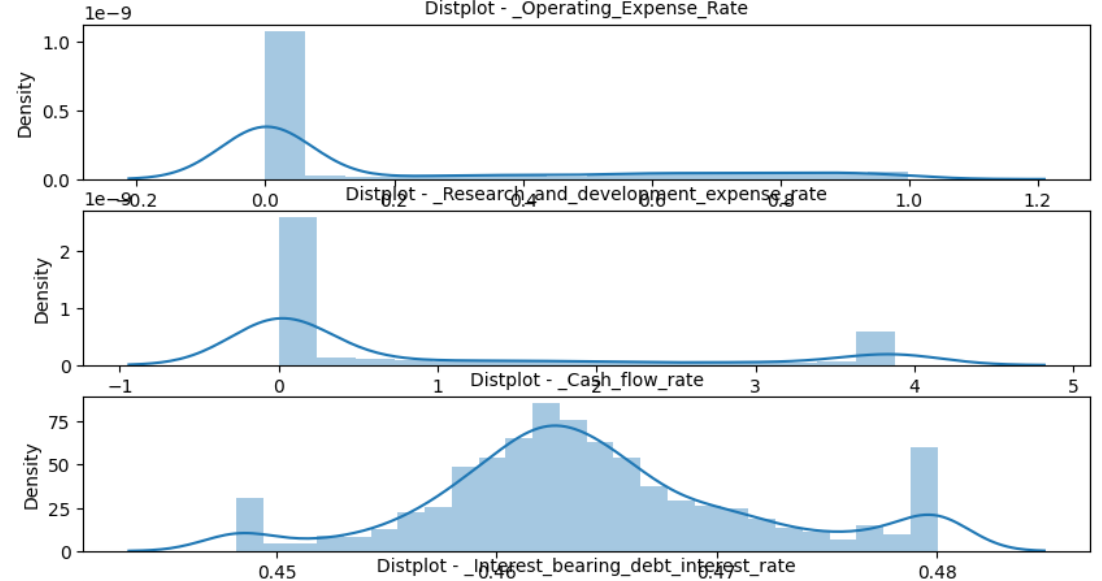
That is 89% non defaulters and 11 defaulters.

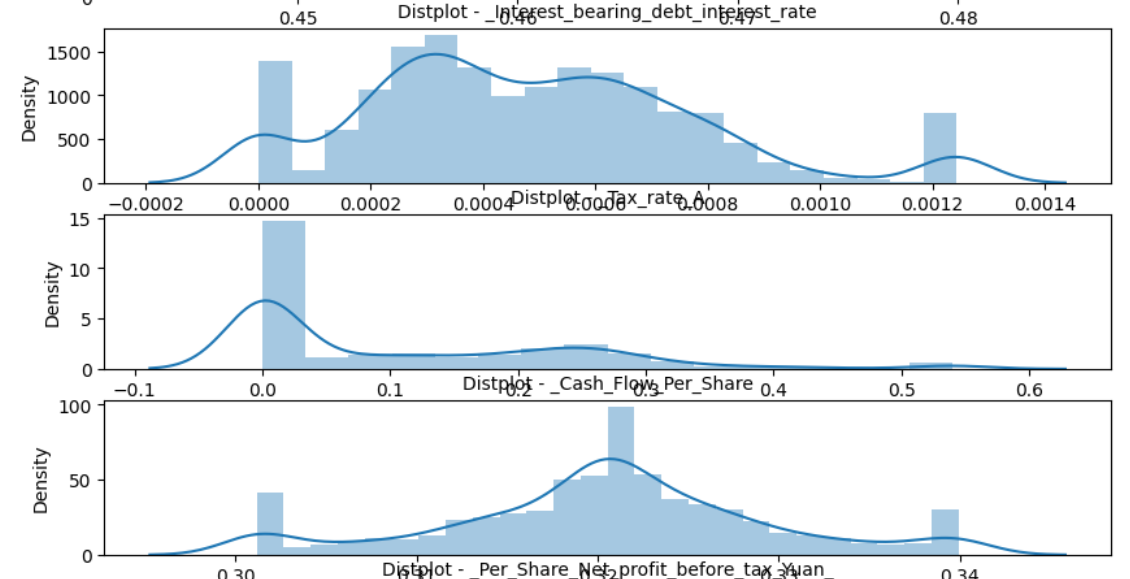
Figure 3: Correlation plot

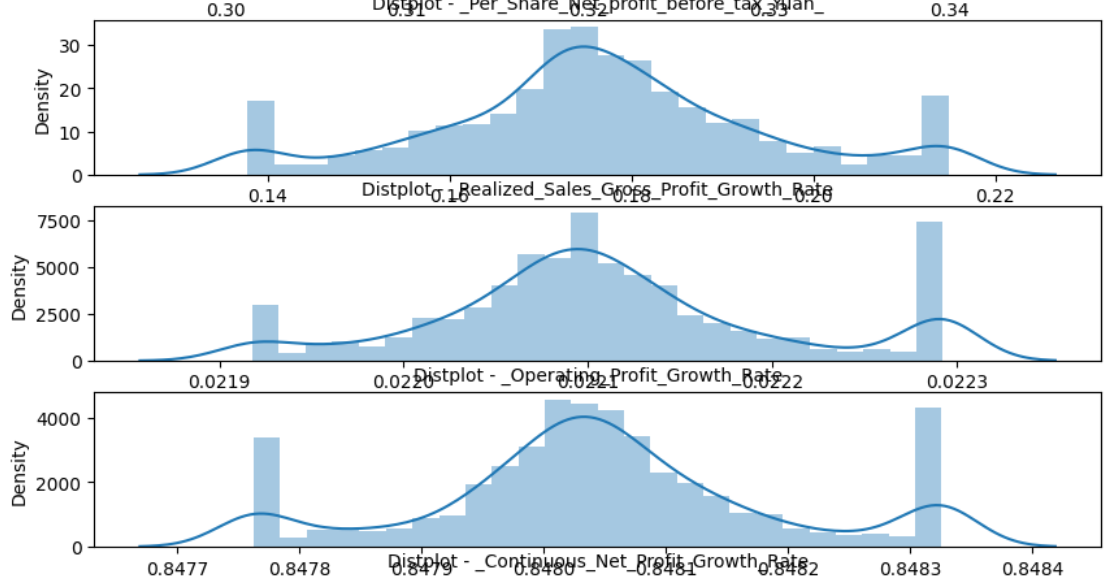
****

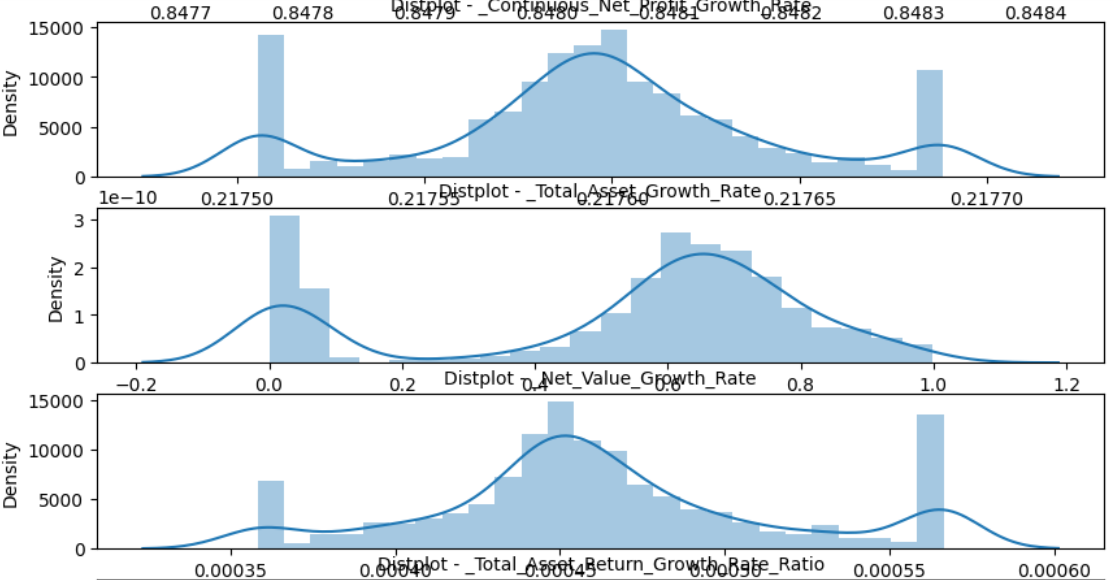
**The Bivariate analysis is done for Default column and all the columns**

Figure 4: Histograms of all variables vs Default



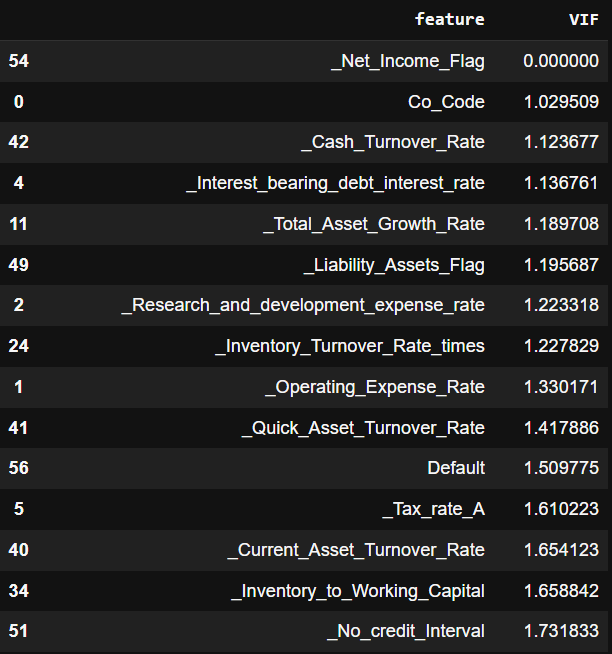


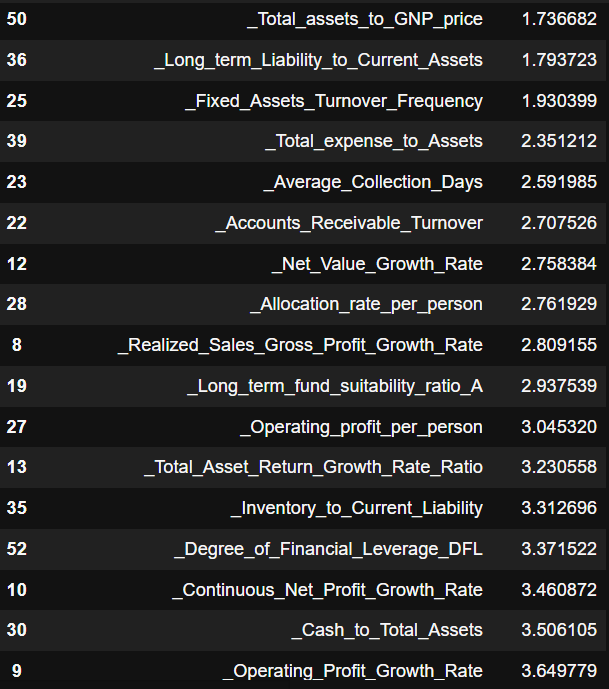


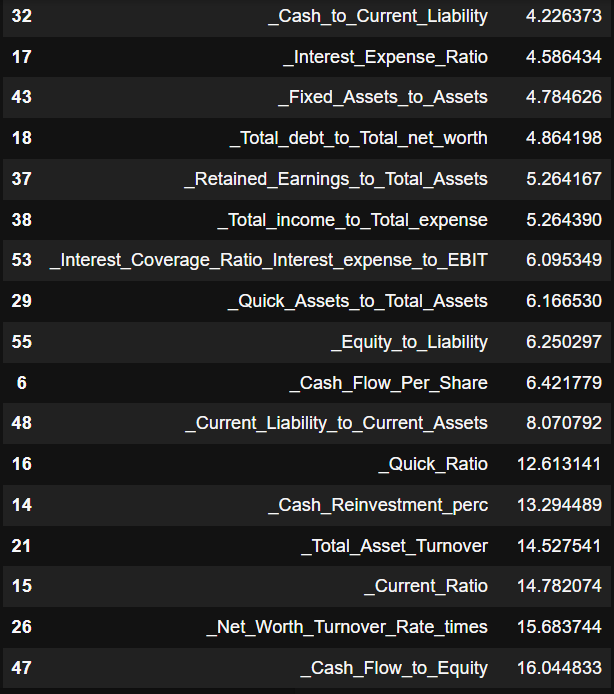


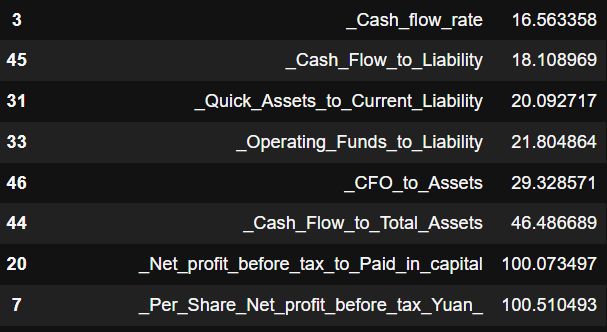
**Important Features:**

**Table 6 : VIF values**

****

****

****

****

The important variables are selected by the Variance inflation factor.

Variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression.

* A value of 1 indicates there is no correlation between a given explanatory variable and any other explanatory variables in the model.
* A value between 1 and 5 indicates moderate correlation between a given explanatory variable and other explanatory variables in the model, but this is often not severe enough to require attention.
* A value greater than 5 indicates potentially severe correlation between a given explanatory variable and other explanatory variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

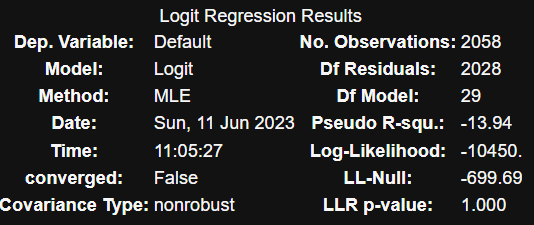
Let’s consider the variables with less 5 VIF. The variables with VIF more than 5 have high correlations which effects the model which is not good.

The logistic regression model is trained on the specified features (imp\_features) and the target variable (df['Default']). The model is fitted to the data, and the returned result object contains information about the estimated coefficients, statistical tests, and other model-related properties.

Let’s split the data in the ratio 67:33. That is 67% train data and 33% test data.

Let’s consider only the variables with less than 5 value of VIF.

Table 7: MODEL 1 summery



****



The above table is the summary provides information about the estimated coefficients, standard errors, z-values, p-values, and confidence intervals for each predictor variable in the model. The coefficients represent the estimated effects of each predictor on the log-odds of the "Default" variable. The p-values indicate the statistical significance of each predictor, where p < 0.05 suggests a significant effect.

Dep. Variable: Default: The dependent variable or target variable in the model is "Default".

• Method: MLE (Maximum Likelihood Estimation): The method used for estimating the model parameters is maximum likelihood estimation.

• Df Model: 29: The degrees of freedom for the model are 29.

• Pseudo R-squ.: -13.94: The pseudo R-squared value, which measures the goodness-of-fit of the model, is -13.94. However, a negative value suggests that the model is not a good fit for the data.

• Log-Likelihood: -10450.0: The log-likelihood value of the model is -10450.0.

• Converged: False: The model did not converge, indicating that the maximum number of iterations was exceeded or there may be issues with the data.

• LL-Null: -699.69: The log-likelihood value of a null model (without any predictors) is -699.69.

• Covariance Type: nonrobust: The covariance type used in the model is nonrobust.

• LLR p-value: 1.000: The p-value associated with the likelihood ratio test is 1.000, indicating that the model does not significantly improve the fit compared to the null model.

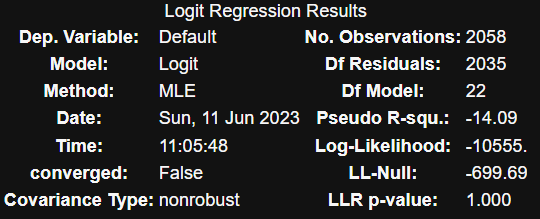
The table below the summary provides information about the estimated coefficients, standard errors, z-values, p-values, and confidence intervals for each predictor variable in the model. The coefficients represent the estimated effects of each predictor on the log-odds of the "Default" variable. The p-values indicate the statistical significance of each predictor, where p < 0.05 suggests a significant effect.

The p-values of '\_Liability\_Assets\_Flag','\_Quick\_Asset\_Turnover\_Rate','\_Tax\_rate\_A','\_Total\_expense\_to\_Assets','\_Interest\_Expense\_Ratio','\_Inventory\_to\_Current\_Liability','\_Fixed\_Assets\_to\_Assets' are greater than 0.05.

These variables are to be dropped.

After dropping thses variables we build other model.

Table 8: MODEL-2 summery:



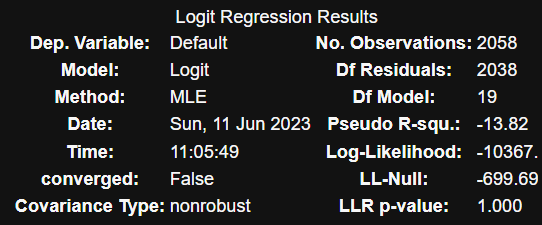


Still there are three variables with greater p-values.

These variables \_Inventory\_Turnover\_Rate\_times','\_Total\_assets\_to\_GNP\_price','\_Long\_term\_fund\_suitability\_ratio\_A' have higher p-values.

Let’s drop these variables and build model 3

Table 9 :MODEL-3 summery:

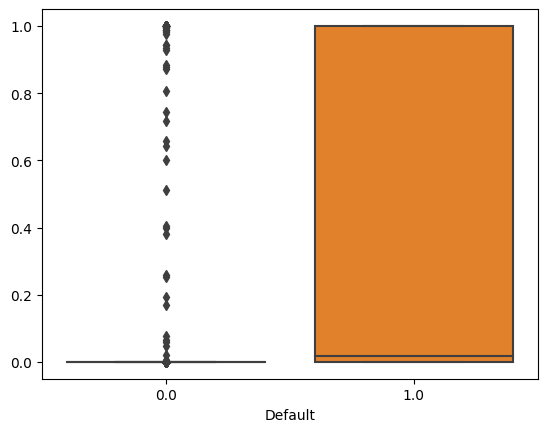




In the model-3, the p-values are less than 0.05.

Therefore the model-3 is the best model.

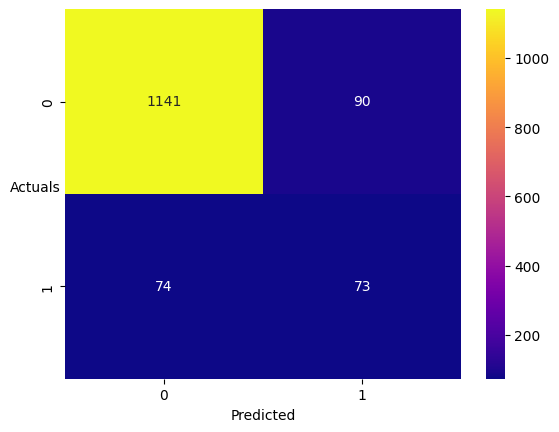
Figure 5: Boxplot of Default and predicted training data set



From the above boxplot, we need to decide on one such value of a cut-off which will give us the most reasonable descriptive power of the model. Let us take a cut-off of 0.07 and check.

Let us now see the predicted classes

Figure 6: confusion matrix for training set with cut-off 0.07



True Negative: 1141

False Positives: 90

False Negatives: 74

True Positives: 73

Table 10: Classification report for train set with cut off 0.08

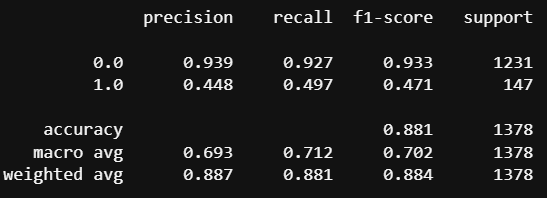
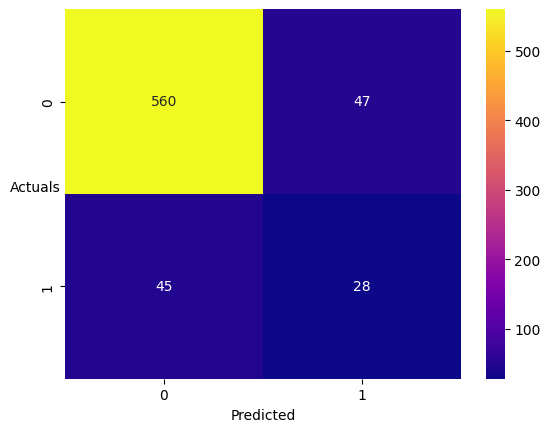


Figure 7: confusion matrix for test set with cut-off 0.07



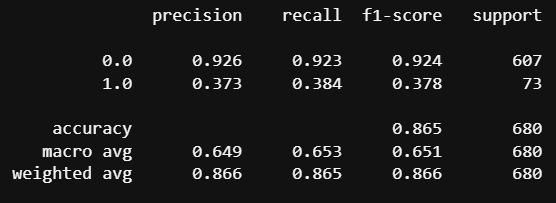
True Negative: 560

False Positives: 47

False Negatives: 45

True Positives: 28

Table 11: Classification report for test set with cut off 0.08



After checking the cut-off from 0.01 to 0.99

We got to know 0.05 is the cut-off with the best sensitivity.

Let’s set the cut off with 0.05 and check the confusion matric and classification report

Figure 8: confusion matrix for training set with cut-off 0.05

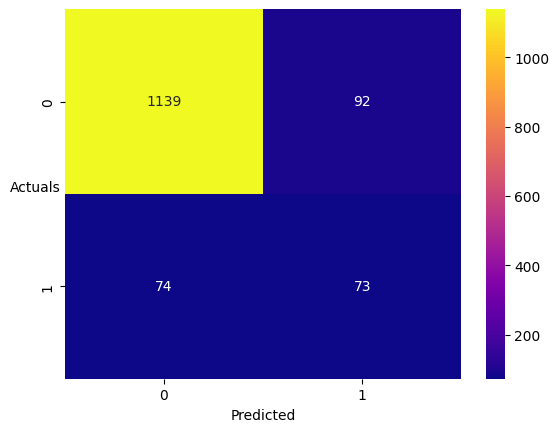


Table 12: Classification report for training set with cut off 0.05

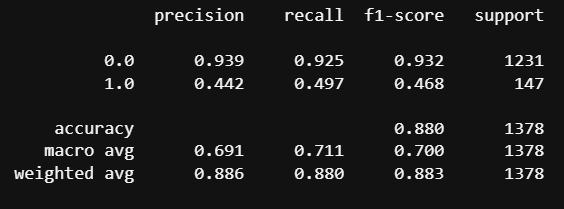


Figure 9: confusion matrix for test data set with cut-off 0.05

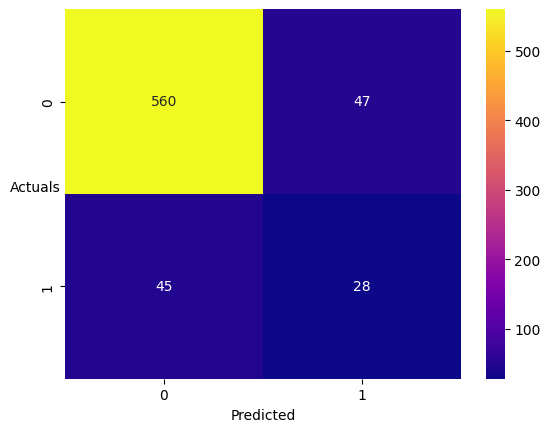
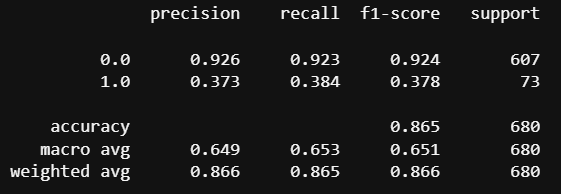


Table 13: Classification report for Test data set with cut off 0.05



**Conclusion:**

The logistic regression for the credit risk data of model 3 is the best model.

The sensitivity for train data is 47% which is very low and the sensitivity is only 38% on test data which is very low and is not the best model for the prediction of the credit risk.

**Table of Content**

|  |  |  |
| --- | --- | --- |
| S.No | Content | Pg.No |
|  | Logistic Regression Model-4 | 24 |
|  | Logistic Regression Model-5 | 25 |
|  | Build a Random Forest Model on Train Dataset. Also showcase your model building approach | 28 |
|  | Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model | 29 |
|  | Build a LDA Model on Train Dataset. Also showcase your model building approach | 30 |
|  | Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model | 31 |
|  | Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve) | 33 |
|  | State Recommendations from the above models | 33 |
|  | Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference | 35 |
|  | Calculate Returns for all stocks with inference | 35 |
|  | Calculate Stock Means and Standard Deviation for all stocks with inference | 36 |
|  | Draw a plot of Stock Means vs Standard Deviation and state your inference | 37 |
|  | Conclusion and Recommendations | 37 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| S.no | Figures | Pg.no |
|  | Confusion matrix for training set logistic regression | 26 |
|  | Confusion matrix for testing set logistic regression | 27 |
|  | ROC for logistic Regression | 27 |
|  | Important features | 28 |
|  | Confusion matrix for training set Random Forest | 28 |
|  | Confusion matrix for testing set Random Forest | 29 |
|  | ROC for Random Forest | 29 |
|  | Confusion matrix for training set LDA | 30 |
|  | Confusion matrix for testing set LDA | 31 |
|  | ROC for LDA | 31 |
|  | Confusion matrix for training set LDA with SMOTE | 32 |
|  | Confusion matrix for testing set LDA with SMOTE | 32 |
|  | ROC for LDA with SMOTE | 33 |
|  | Stock Price Graph | 35 |
|  | Stock vs Standard Deviation | 37 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| S.no | Tables | Pg.no |
|  | Important features using Decision Tree | 24 |
|  | Model-4 Summary | 24 |
|  | Model-5 Summary | 25 |
|  | Classification report for training set logistic regression | 26 |
|  | Classification report for testing set logistic regression | 27 |
|  | Classification report for training set Random Forest | 29 |
|  | Classification report for testing set Random Forest | 29 |
|  | Classification report for training set LDA | 30 |
|  | Classification report for testing set LDA | 31 |
|  | Performance Comparison Table | 32 |
|  | Head of MRA data | 33 |
|  | Basic information of the MRA data | 33 |
|  | Volatility | 34 |
|  | Value at Risk | 35 |
|  | Mean and Standard Deviation | 35 |

**Milestone-2**

**Logistic Regression Model 4:**

In Model 4, the **important features** are considered using **Decision Tree**. Using this we got the important features which are considered in building Logistic regression model.

Table 14:Important features using Decision Tree



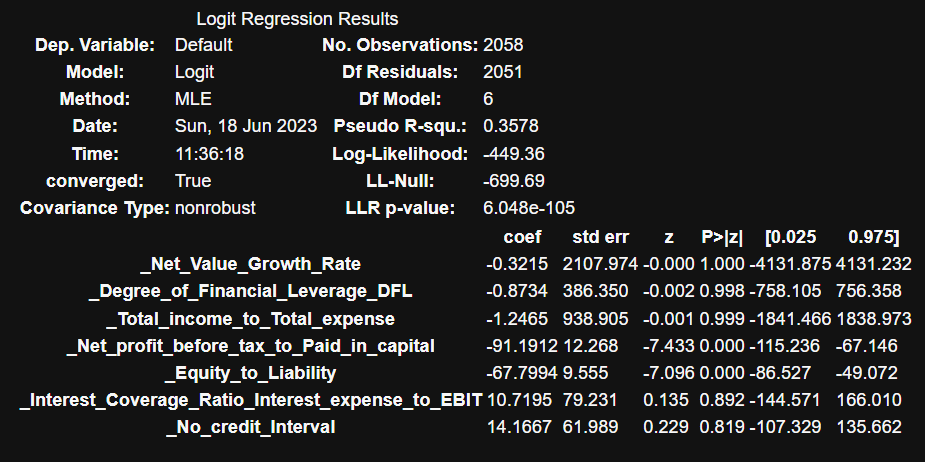
Using the important features that are greater than 0 are considered here as there are only few columns greater than 0.

This threshold can be changed according to the problem and features available.

'\_Net\_Value\_Growth\_Rate','\_Degree\_of\_Financial\_Leverage\_DFL','\_Total\_income\_to\_Total\_expense','\_Net\_profit\_before\_tax\_to\_Paid\_in\_capital','\_Equity\_to\_Liability','\_Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT','\_No\_credit\_Interval'

The above mentioned features are used for model.

Table 15: MODEL-4 Summary



The significance level or alpha values is 0.05. Considering this value check the p-value above.

We could say that 5 features have greater than 0.05.

When the p-values associated with the predictor variables in a logistic regression model are greater than 0.05, it suggests that these variables are not statistically significant predictors of the dependent variable.

In other words, there is insufficient evidence to conclude that these variables have a significant impact on the likelihood of the "Default" outcome.

Therefore, you may consider removing these variables from the model or exploring alternative explanatory variables that have stronger statistical significance.

But after removing these columns are will be left with only features to build a model with. This model will be meaningless.

Hence let’s now consider building a model after balancing the data using **SMOTE (Synthetic Minority Over-sampling Technique)** is a technique used to address class imbalance in datasets.

After balancing the data, the data has 3656 for training data and 3656 for testing data.

To this data we find the important features and fit the model.

**Important Features:**

\_Cash\_flow\_rate', '\_Operating\_Profit\_Growth\_Rate', '\_Continuous\_Net\_Profit\_Growth\_Rate', '\_Net\_Value\_Growth\_Rate', '\_Total\_debt\_to\_Total\_net\_worth', '\_Net\_profit\_before\_tax\_to\_Paid\_in\_capital', '\_Accounts\_Receivable\_Turnover', '\_Allocation\_rate\_per\_person', '\_Cash\_to\_Current\_Liability', '\_Inventory\_to\_Working\_Capital', '\_Retained\_Earnings\_to\_Total\_Assets', '\_Cash\_Flow\_to\_Equity', '\_Total\_assets\_to\_GNP\_price', '\_Degree\_of\_Financial\_Leverage\_DFL

Table 16: MODEL-5 Summary



In this logistic model, the p-value for all the features is less than 0.05. Therefore we might consider this model.

Lets check the confusion matrixs and classification report to know about the performance of the model.

Figure 10: Confusion matrix for training set

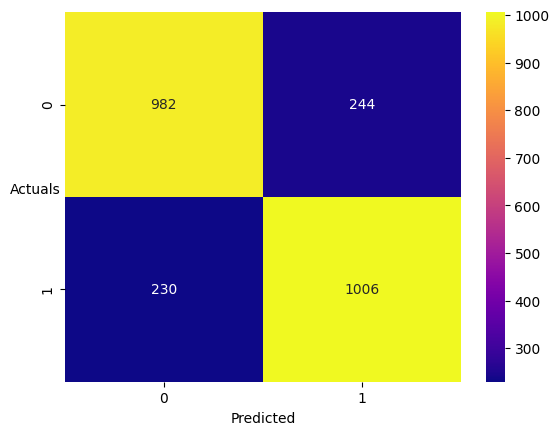
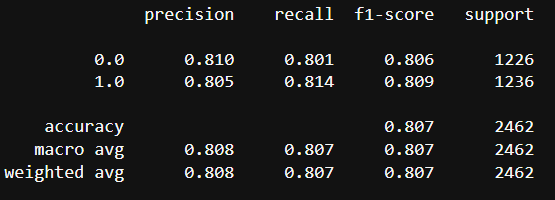


Table 17: Classification report for training set



**The Sensitivity (recall) = 81.4% => 81% (rounding off)**

**The Accuracy = 80.7% => 81% (rounding off)**

In general, a high sensitivity (recall) value is desirable when the cost of false negatives (misclassifying a positive instance as negative) is high.

On the other hand, accuracy is a more balanced metric that considers both true positives and true negatives.

It is suitable when the costs of false positives and false negatives are relatively similar or when there is no significant class imbalance.

A high accuracy value, approaching 1.0, indicates a high overall correctness of the model's predictions.

**In a credit risk problem, where the goal is to assess the creditworthiness of borrowers, the focus is typically on minimizing the risk of default or loan delinquency. Therefore, the metric that is often prioritized is the model's ability to correctly identify the positive class (i.e., borrowers who will default) while minimizing false negatives. In this case, sensitivity (recall) is usually considered more important than accuracy.**

Figure 11: Confusion matrix for test data

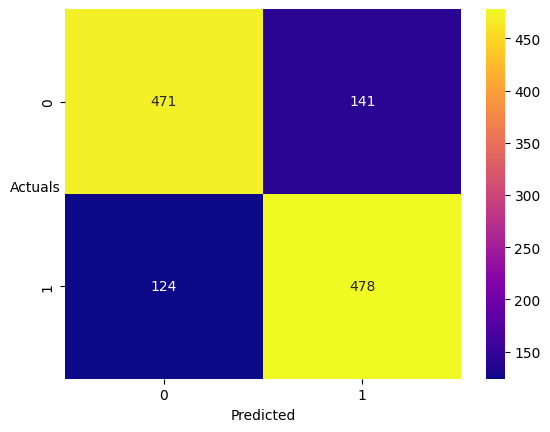
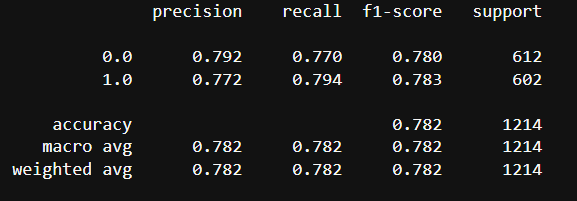


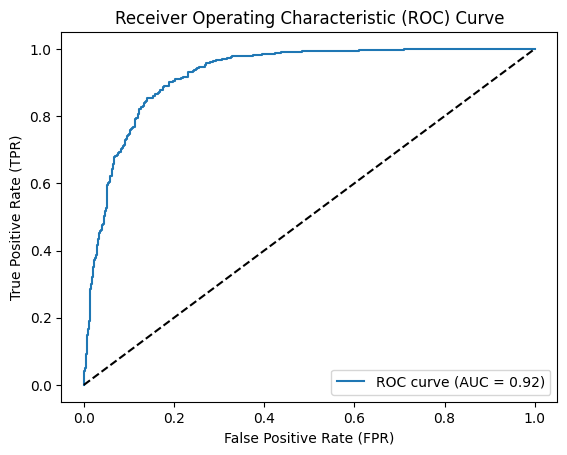
Table 18: Classification report for testing data



**The Sensitivity (recall) = 79.4% => 79% (rounding off)**

**The Accuracy = 78.2% => 78% (rounding off)**

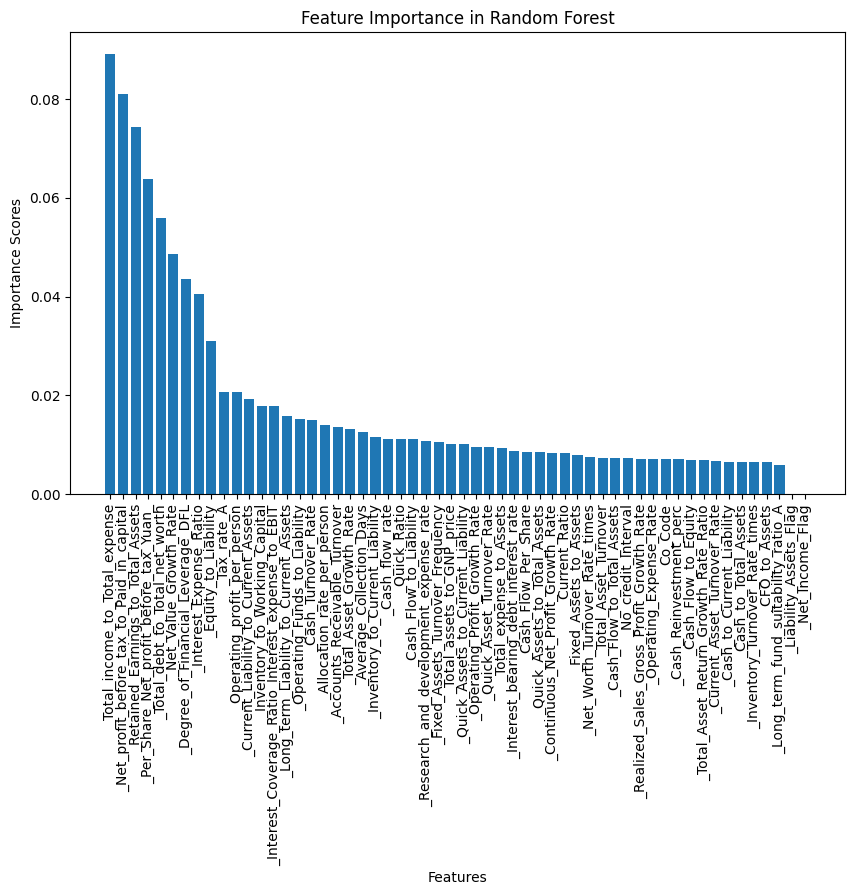
Figure 12: ROC for logistic regression



**RANDOM FOREST:**

The following are the important features.

Figure 13: Important features



Here in Random forest model, we doesn’t only consider important features. If we do so, we get highest accuracy but there will be much data lost.

Let’s consider all the features and build model.

Figure 14: Confusion matrix for training data

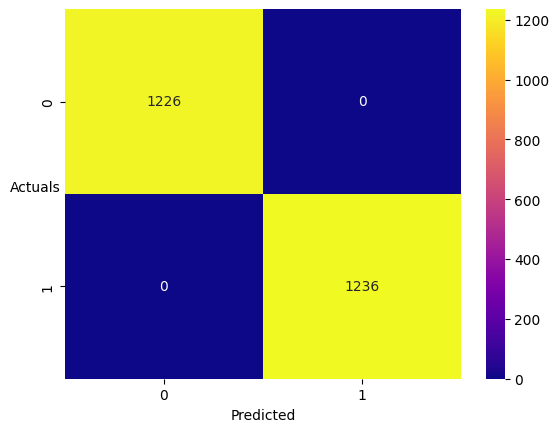


Table 19: Classification report for training data

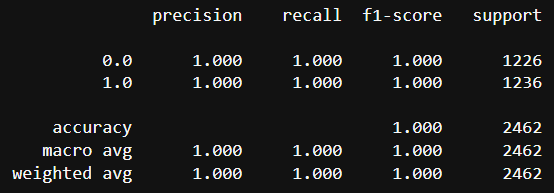


Figure 15: Confusion matrix for test data

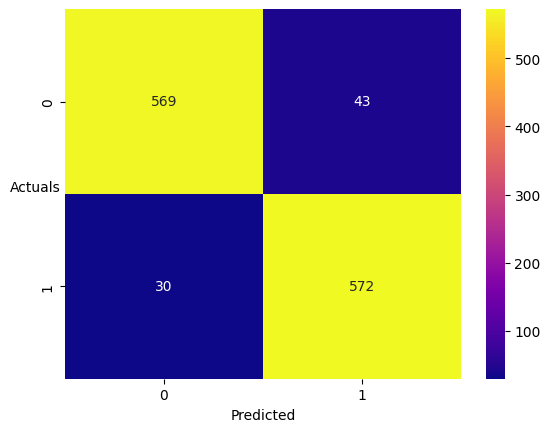


Table 20: Classification report for test data

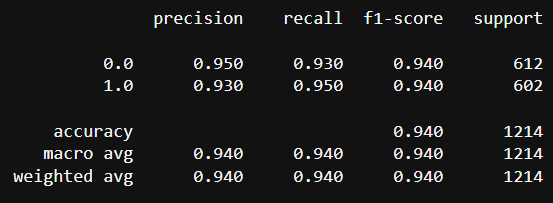
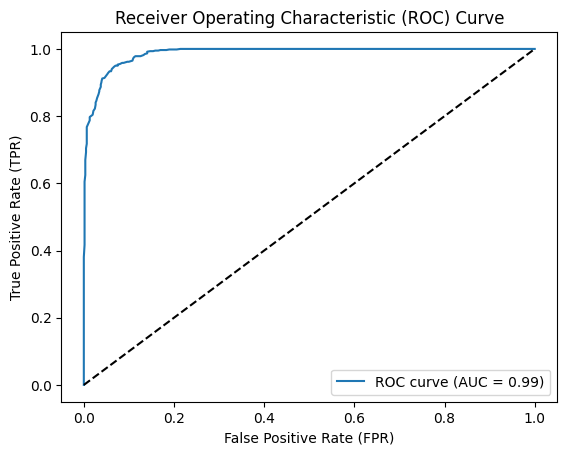


Figure 16: ROC for Random Forest



**LINEAR DISCRIMINANT ANALYSIS**

The LDA is build on the data set given which is not balanced and which is balanced using SMOTE.

For the data set which is not balanced, the performance is as following:

Figure 17: Confusion matrix training set

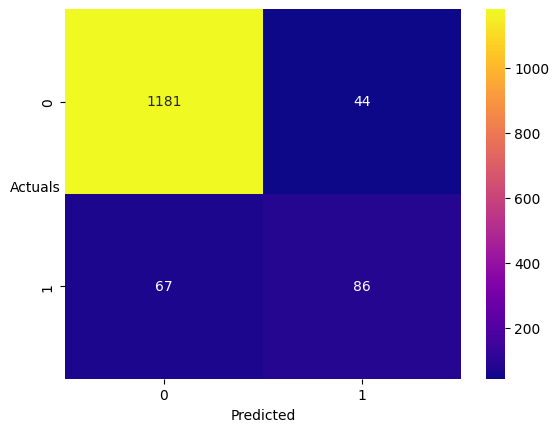


Table 21: Classification report training set

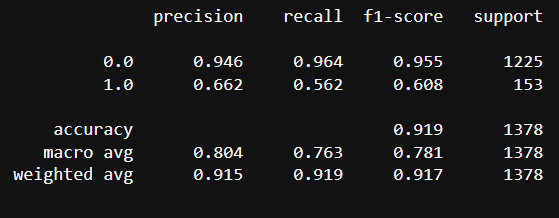


Figure 18: Confusion matrix for test data

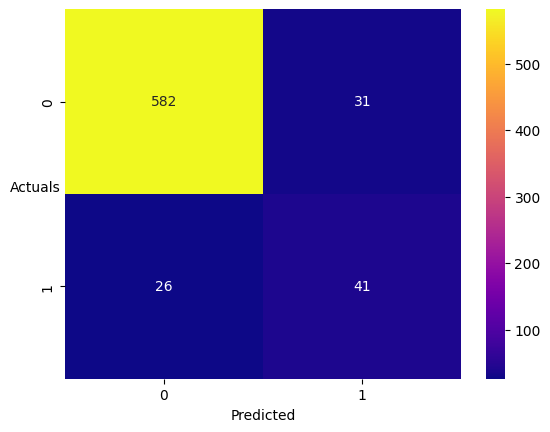


Table 22: Classification report testing set

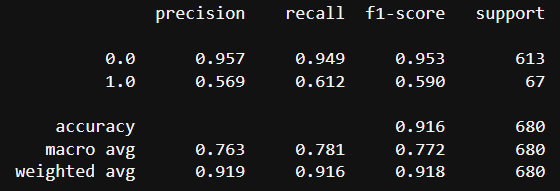
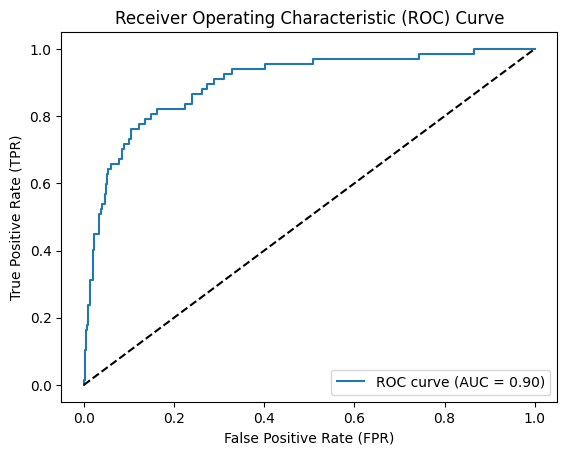


Figure 19: ROC for LDA



The performance of the balanced data model:

Figure 19: Confusion matrix training set

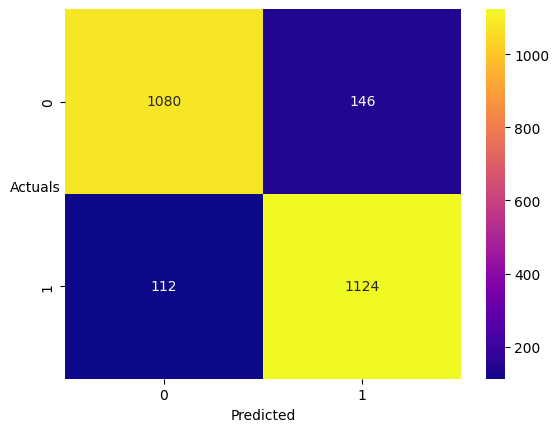


Table 23: Classification report training set

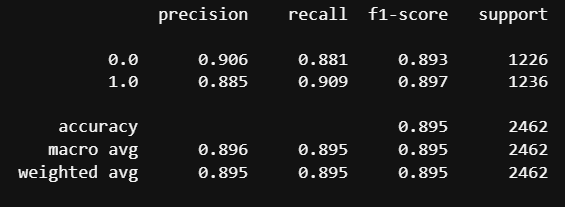


Figure 20: Confusion matrix for test data

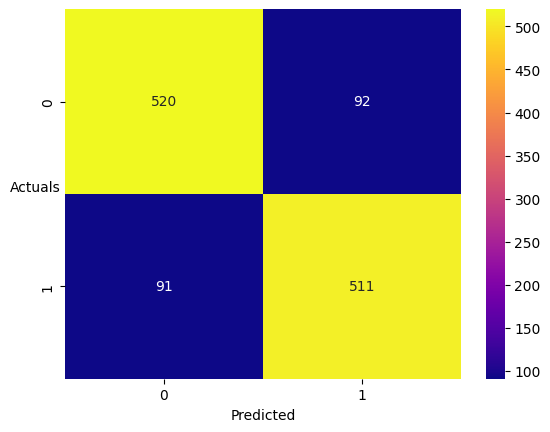


Table 24: Classification report testing set

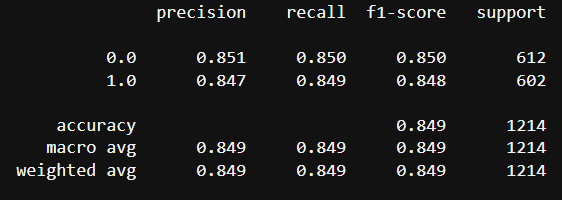
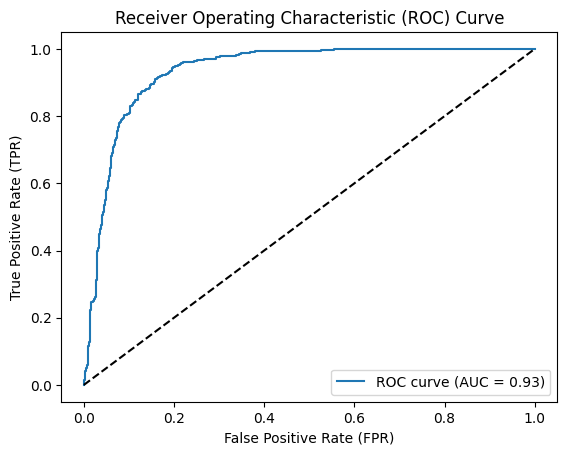


Figure 21: ROC for LDA



**Insights:**

Table 25: Performance comparison table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity (Recall) | AUC |
| Logistic Regression | 79 | 78 | 92 |
| Random Forest | 95 | 94 | 99 |
| LDA | 91 | 61 | 90 |
| LDA - SMOTE | 85 | 85 | 93 |
| Decision Tree (addition) | 95 | 96 | 79 |

By observing all the models, the above table can be called a performance comparison table.

* In a credit risk problem, the goal is to assess the creditworthiness of borrowers.
* The focus is typically on minimizing the risk of default or loan delinquency.
* Therefore, the metric that is often prioritized is the model's ability to correctly identify the positive class (i.e., borrowers who will default) while minimizing false negatives.
* **In this case, sensitivity (recall) is usually considered more important than accuracy**.

So, by comparing the sensitivity of the model and also considering other factors, the Random forest is the best model.

This Random Forest model is built on balanced data (SMOTE).

The model built on imbalanced data has good accuracy but very low sensitivity which is to be considered in model building, for example, the LDA model is built on imbalanced data, which has 90% accuracy but 61% of recall which makes the poor. model

When we consider the Decision Tree, this model has good accuracy and good recall values but the AUC value is poor.

The AUC represents the overall performance of the model. It quantifies the model's ability to distinguish between the positive and negative classes. A higher AUC score indicates better discrimination between the classes, with 1 being a perfect classifier.

**Market Risk Analysis**

**Problem Statement:**

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights.

The data set is to be loaded and understand it.

Let’s first check few rows of data to understand the data set.

Table 26: Head of MRA data

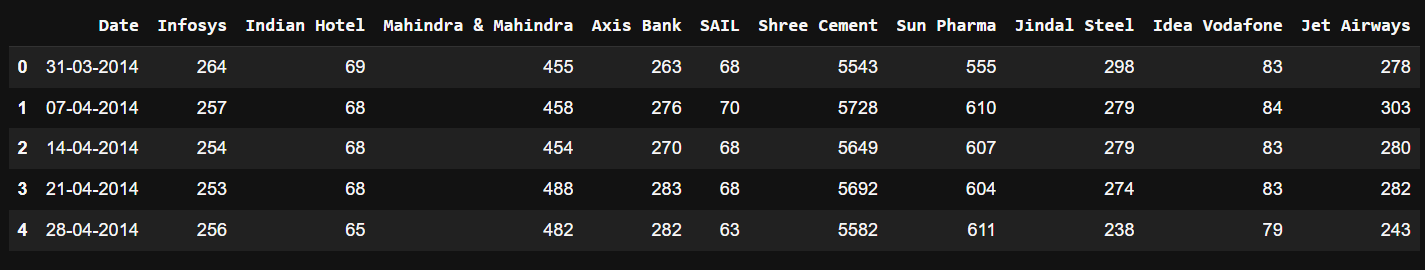
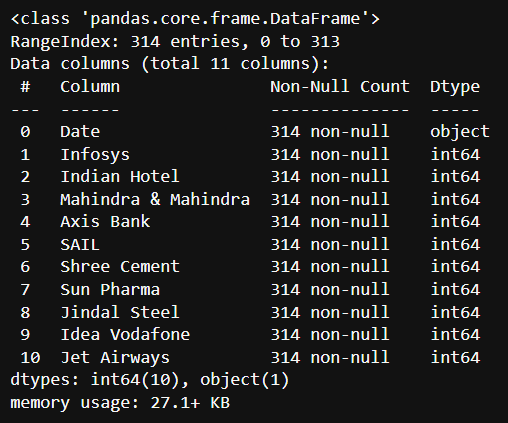


Table 27: Basic information of the MRA data



From the above tables, we can understand that the data set has 314 rows and 11 columns.

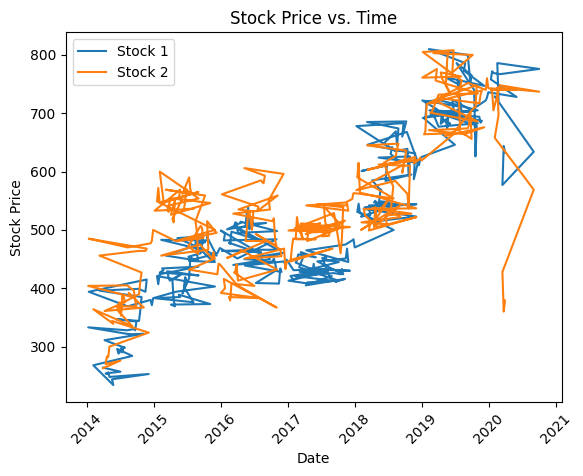
There are no null values in the data which is a good data set.

Also, the important thing to change is the data type of the Date column.

This column should be converted to Date time data type for further analysis.

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

Figure 22: Stock price Graph

****

Here, the Stock1 is Infosys and Stock2 is Axis Bank. The Stock Price Graph is plotted for two stocks.

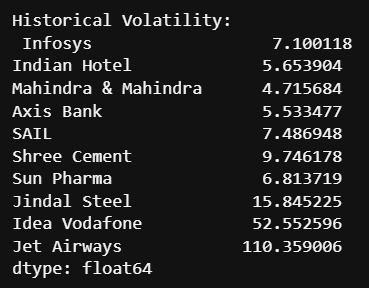
Infosy’s stock price is lower than Axis Bank’s stock till 2019, but later the stock price of Infosys got increased almost to the stock price of Axis Bank.

The stock price for both stocks is low in 2017-2018 and again drop in 2020- 2021.

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

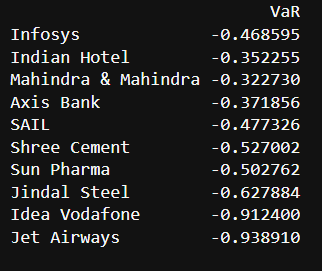
By calculating the returns, you can gain insights into the performance of each stock over time, analyze their volatility, and make informed investment decisions based on historical price movements.

Table 28: Volatility



We can observe the volatility of each stock, which provides an indication of the stock's price fluctuations or risk. Higher volatility suggests larger price swings, while lower volatility indicates more stable price movements.

Table 29: Value-at-Risk (VaR)



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

Table 30: Mean and Standard Deviation



Mean Returns: The mean returns represent the average performance of each stock over the given period.

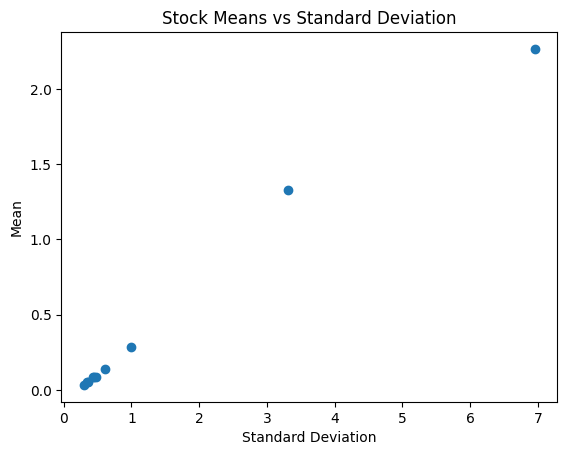
Positive mean returns indicate that, on average, the stock has generated a profit, while negative mean returns suggest an average loss.

Higher mean returns generally indicate better performance, but it's important to consider the risk associated with the stock as well.

Standard Deviation: The standard deviation measures the dispersion or volatility of returns for each stock.

A higher standard deviation indicates higher variability in the stock's returns, suggesting greater risk. Mean Returns:

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



**Conclusion and Recommendations**

**Insights:**

**Mean Returns:**

* Idea Vodafone and Jet Airways have significantly higher mean returns compared to other stocks. This suggests that, on average, these stocks have generated higher profits.
* Jindal Steel also has a relatively high mean return, indicating strong average performance.
* Other stocks such as Infosys, Indian Hotel, Mahindra & Mahindra, Axis Bank, SAIL, Shree Cement, and Sun Pharma have comparatively lower mean returns

**Standard Deviations:**

* Idea Vodafone and Jet Airways exhibit the highest standard deviations, indicating higher volatility and risk in their returns. These stocks are more prone to significant price fluctuations.
* Jindal Steel also has a relatively high standard deviation, suggesting greater variability in its returns and higher risk.
* Other stocks such as Infosys, Indian Hotel, Mahindra & Mahindra, Axis Bank, SAIL, Shree Cement, and Sun Pharma have lower standard deviations, indicating relatively lower volatility and risk.

In summary, based on the mean returns and standard deviations, stocks like Idea Vodafone and Jet Airways show higher potential returns but also come with higher risks due to their higher volatility.

On the other hand, stocks with lower standard deviations, such as Infosys, Indian Hotel, Mahindra & Mahindra, Axis Bank, SAIL, Shree Cement, and Sun Pharma, have a relatively lower risk but may offer lower average returns.

A lower standard deviation implies more stable and predictable returns. Investors often consider standard deviation as a measure of risk when comparing different stocks.

**Conclusion:**

The stock data represents the historical performance of various stocks, including Infosys, Indian Hotel, Mahindra & Mahindra, Axis Bank, SAIL, Shree Cement, Sun Pharma, Jindal Steel, Idea Vodafone, and Jet Airways.

The analysis included calculations of returns, volatility (standard deviation), mean returns, and Value-at-Risk (VaR) for the stocks.

The stocks exhibited varying levels of mean returns, standard deviations, and risk profiles.

**Recommendations:**

* Since different stocks have different risk and return characteristics, it is recommended to diversify the investment portfolio. By investing in a mix of stocks with varying risk levels, you can potentially reduce the overall risk and increase the chances of achieving more stable returns.
* Stocks with higher volatility and risk, such as Idea Vodafone and Jet Airways, may offer higher potential returns but are also more prone to significant price fluctuations.
* Assess your risk tolerance and align it with the risk associated with each stock.
* Stay updated with market conditions, industry trends, and company-specific news that can impact the performance of the stocks. Regularly review and adjust your portfolio based on changing market dynamics to optimize your investment returns.
* If you are unsure about making investment decisions or need expert guidance, consider consulting a financial advisor or investment professional. They can provide personalized advice based on your financial situation, risk tolerance, and investment goals.
* Stock investments are typically suited for the long term. Evaluate your investment horizon and consider holding stocks for an extended period to potentially benefit from long-term market trends and compounding returns.
* Remember, investing in the stock market carries inherent risks, and past performance is not indicative of future results. Conduct thorough research, evaluate your risk tolerance, and make informed investment decisions based on your financial goals and circumstances.