22 May 2022

FRA Milestone-2 Project

PGP-DSBA June-Batch

SOOREJ KUNNOOL BALAN

2022

Table of Contents

[1. Problem 3](#_Toc104132265)

[1.8 Build a Random Forest Model on Train Dataset. Also showcase your model building approach. 3](#_Toc104132266)

[1.9 Validate the Random Forest Model on test Dataset and state the performance matrices. Also state interpretation from the model. 5](#_Toc104132267)

[1.10 Build a LDA Model on Train Dataset. Also showcase your model building approach. 7](#_Toc104132268)

[1.11 Validate the LDA Model on test Dataset and state the performance matrices. Also state interpretation from the model. 9](#_Toc104132269)

[1.12 Compare the performances of Logistics, Radom Forest and LDA models (include ROC Curve). 11](#_Toc104132270)

[1.13 State Recommendations from the above models. 12](#_Toc104132271)

[2. Market Risk 13](#_Toc104132272)

[2.1 Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference. 14](#_Toc104132273)

[2.2 Calculate Returns for all stocks with inference. 15](#_Toc104132274)

[2.3 Calculate Stock Means and Standard Deviation for all stocks with inference 16](#_Toc104132275)

[2.4 Draw a plot of Stock Means vs Standard Deviation and state your inference 17](#_Toc104132276)

[2.5 Conclusion and Recommendations 17](#_Toc104132277)

[END 18](#_Toc104132278)

List of Tables List of Figures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table-1.1 RF- Model 1 – Train confusion matrix | 4 |  | Figure-1.1 RF- Model 1 – Test ROC Curve | | 6 |
| Table-1.2 RF- Model 1 – Train classification report | 4 |  | Figure-1.2 RF- Model 2 – SMOTE Test ROC Curve | | 7 |
| Table-1.3 RF- Model 2 – SMOTE Train confusion matrix | 4 |  | Figure-1.3 LDA- Model 1 – Test ROC Curve | | 10 |
| Table-1.4 RF- Model 2 – SMOTE Train classification report | 5 |  | Figure-1.4 LDA- Model 2 – SMOTE Test ROC Curve | | 11 |
| Table-1.5 RF- Model 1 –Test confusion matrix | 5 |  | Figure-1.5 ROC Curve Comparison | | 12 |
| Table-1.6 RF- Model 1 – Test classification report | 5 |  | Figure-2.1 Shree\_Cement stock price graph | | 14 |
| Table-1.7 RF- Model 2 –SMOTE Test confusion matrix | 6 |  | Figure-2.2 Idea\_Vodafone stock price graph | | 15 |
| Table-1.8 RF- Model 2 – SMOTE Test classification report | 6 |  | Figure-2.3 Stock returns: Mean vs std. | | 17 |
| Table-1.9 LDA- Model 1 – Train confusion matrix | 8 |  |  | |  |
| Table-1.10 LDA- Model 1 – Train classification report | 8 |  |  | |  |
| Table-1.11 LDA- Model 2 – SMOTE Train confusion matrix | 8 |  |  | |  |
| Table-1.12 LDA- Model 2 – SMOTE Train classification report | 9 |  |  | |  |
| Table-1.13 LDA- Model 1 –Test confusion matrix | 9 |  |  | |  |
| Table-1.14 LDA- Model 1 – Test classification report | 9 |  |  | |  |
| Table-1.15 LDA- Model 2 –SMOTE Test confusion matrix | 10 |  |  | |  |
| Table-1.16 LDA- Model 2 – SMOTE Test classification report | 10 |  |  | |  |
| Table-1.17 Model comparison | 11 |  |  | |  |
| Table-1.18 Logit feature importance | 12 |  |  | |  |
| Table-2.1 Input Data | 13 |  |  | |  |
| Table-2.2 Modified column names | 13 |  |  | |  |
| Table-2.3 Concise Data Summary | 13 |  |  | |  |
| Table-2.4 Central Measures of Tendency | 14 |  |  | |  |
| Table-2.4 Modified concise info | 14 |  |  | |  |
| Table-2.5 Modified concise info | 14 |  |  | |  |
| Table-2.6 Head of stock returns | 15 |  |  | |  |
| Table-2.7 tail of stock returns | 15 |  |  |  | |
| Table-2.8 Stock return data description | 16 |  |  |  | |
| Table-2.9 Stock return mean/std | 16 |  |  |  | |
| Table-2.10 Stock price movement | 18 |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |
|  |  |  |  |  | |

# 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 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 labeled field.

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

Hints :

Dependent variable - We need to create a default variable that 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.

 

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

First we will build random forest model on the train dataset and then on the oversampled training dataset created using SMOTE.

**Model 1**: We will run random forest classifier for various parameters and perform hyper -parameter tuning using gridsearch on train dataset. Criterion we will consider both gini and entropy. In the train dataset we have 2402 observations. So we can start with min\_samples\_leaf as 1% of observations, so around 24. For the hypertuning we will consider 20,30 and 40 as the minimum sample leaf. The min\_samples\_split can be considered as 3 times split, hence we can use values 60,90 and 120. We will consider a minimum depth of 10,15 and 20.

On performing the gridsearch, with cross-validation 5 times on the train dataset, the optimal random forest classifier model has been identified as below:

RandomForestClassifier(criterion='entropy', max\_depth=10, min\_samples\_leaf=30, min\_samples\_split=60, random\_state=1)

On predicting the response with above model on training data gives the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 2123 | 19 |
| Actual Positive | 51 | 209 |

Table-1.1 RF- Model 1 – Train confusion matrix

Lets look at the classification report.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.2 RF- Model 1 – Train classification report

We can see that the overall accuracy is good at 0.97 and the recall for defaulters are at 0.8 which is decent.

We can comment about the effectiveness of the model once its results are verified against the test dataset.

**Model 2**: We will run random forest classifier for various parameters and perform hyper -parameter tuning using gridsearch on the oversampled train set using SMOTE. Criterion we will consider both gini and entropy. In the train dataset we have 4284 observations. So we can start with min\_samples\_leaf as 1% of observations, so around 40. For the hypertuning we will consider 30,40 and 50 as the minimum sample leaf. The min\_samples\_split can be considered as 3 times split, hence we can use values 90,120 and 150. We will consider a minimum depth of 10,15 and 20.

On performing the gridsearch, with cross-validation 5 times on the oversampled train dataset, the optimal random forest classifier model has been identified as below:

RandomForestClassifier(criterion='entropy', max\_depth=10, min\_samples\_leaf=40, min\_samples\_split=90, random\_state=1)

On predicting the response with above model on training data gives the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 2104 | 38 |
| Actual Positive | 80 | 2062 |

Table-1.3 RF- Model 2 – SMOTE Train confusion matrix

Lets look at the classification report.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.4 RF- Model 2 – SMOTE Train classification report

We can see that the overall accuracy is good at 0.97 and the recall for defaulters are at 0.96 which is better than model1.

We can comment about the effectiveness of the model once its results are verified against the test dataset.

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

We will use the 2 random forest models (Model 1 on train dataset and Model2 on the oversample train dataset using SMOTE) on the test dataset and verify if the models give good results.

**Model1**: We will predict the response of test dataset using model1. We get the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 1055 | 1 |
| Actual Positive | 37 | 91 |

Table-1.5 RF- Model 1 –Test confusion matrix

Let us look at the classification report of model 1 on test dataset. A screenshot of a computer

Description automatically generated with medium confidence

Table-1.6 RF- Model 1 – Test classification report

Let us look at the ROC curve and the ROC\_AUC score:

Chart, line chart

Description automatically generated

Figure-1.1 RF- Model 1 – Test ROC Curve

**Model2**: We will predict the response of test dataset using model2. We get the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 1034 | 22 |
| Actual Positive | 16 | 112 |

Table-1.7 RF- Model 2 –SMOTE Test confusion matrix

Let us look at the classification report of model 2 on test dataset.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.8 RF- Model 2 – SMOTE Test classification report

Let us look at the ROC curve and the ROC\_AUC score:

Chart, line chart

Description automatically generated

Figure-1.2 RF- Model 2 – SMOTE Test ROC Curve

Comparing the 2 models we can see accuracy is same at 0.97. But the ROC score for model 2 (0.927) is better than model 1 (0.855). Model 2 has a much better recall score of 0.88 than model1 at 0.71.

Both the models are not either over or underfit. We will consider Model2 as the better one in this scenario.

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

First we will build linear discriminant analysis(LDA) model on the train dataset and then on the oversampled training dataset created using SMOTE.

**Model 1**: We will run LDA for various parameters and perform hyper -parameter tuning using gridsearch on train dataset. Solver we will consider svd, lsqr and eigen. We will check for tolerance 0.01, 0.001… to 0.000001

On performing the gridsearch, with cross-validation 5 times on the train dataset, the optimal LDA model has been identified as below:

LinearDiscriminantAnalysis(solver=’svd’,tol=0.01)

On predicting the response with above model on training data gives the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 2117 | 25 |
| Actual Positive | 126 | 134 |

Table-1.9 LDA- Model 1 – Train confusion matrix

Lets look at the classification report.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.10 LDA- Model 1 – Train classification report

We can see that the overall accuracy is good at 0.94 but the recall is abysmal at 0.52.

We can comment about the effectiveness of the model once its results are verified against the test dataset.

**Model 2**: We will run LDA for various parameters and perform hyper -parameter tuning using gridsearch on the oversampled train set using SMOTE. Solver we will consider svd, lsqr and eigen. We will check for tolerance 0.01, 0.001… to 0.000001.

On performing the gridsearch, with cross-validation 5 times on the oversampled train dataset, the optimal LDA model has been identified as below:

LinearDiscriminantAnalysis(solver=’svd’,tol=0.01)

On predicting the response with above model on training data gives the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 1844 | 298 |
| Actual Positive | 95 | 2047 |

Table-1.11 LDA- Model 2 – SMOTE Train confusion matrix

Lets look at the classification report.

A picture containing table

Description automatically generated

Table-1.12 LDA- Model 2 – SMOTE Train classification report

We can see that the overall accuracy is good at 0.91 and the recall for defaulters are at 0.96 which is better than model1.

We can comment about the effectiveness of the model once its results are verified against the test dataset.

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

We will use the 2 LDAmodels (Model 1 on train dataset and Model2 on the oversample train dataset using SMOTE) on the test dataset and verify if the models give good results.

**Model1**: We will predict the response of test dataset using model1. We get the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 1041 | 15 |
| Actual Positive | 68 | 60 |

Table-1.13 LDA- Model 1 –Test confusion matrix

Let us look at the classification report of model 1 on test dataset.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.14 LDA- Model 1 – Test classification report

Let us look at the ROC curve and the ROC\_AUC score:

Chart, line chart

Description automatically generated

Figure-1.3 LDA- Model 1 – Test ROC Curve

**Model2**: We will predict the response of test dataset using model2. We get the below confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 898 | 158 |
| Actual Positive | 14 | 114 |

Table-1.15 LDA- Model 2 –SMOTE Test confusion matrix

Let us look at the classification report of model 2 on test dataset.

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.16 LDA- Model 2 – SMOTE Test classification report

Let us look at the ROC curve and the ROC\_AUC score:

Chart, line chart

Description automatically generated

Figure-1.4 LDA- Model 2 – SMOTE Test ROC Curve

Comparing the 2 models we can see accuracy of model 1 is higher than model 2. But the ROC score for model 2 (0.87) is better than model 1 (0.72). Model 2 has a much better recall score of 0.89 than model1 at 0.47.

Both the models are not either over or underfit. We will consider Model2 as the better one in this scenario.

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

Lets compare the performance of logistic regression, Random Forest and LDA models.

Table

Description automatically generated

Table-1.17 Model comparison

The above table shows the performance of the best model amongst logistic regression(LGR), random forest(RF) and linear discriminant analysis(LDA) against the train and test data.

We can see that in each of the models, SMOTE has corrected the imbalance of data and delivered better results.

Between each models, we can see that accuracy between train/test is lesser than 0.1, which tells us that the models are neither overfit or underfit.

If we compare the performance of the model metrics against the test dataset, we can see that Random Forest has higher values than other models in accuracy, precision, recall(which is almost same for all models), f1-score and roc\_auc\_score. The ROC curve below shows that the second curve which is random forest is able to predict most of the data correctly.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Figure-1.5 ROC Curve Comparison

## State Recommendations from the above models.

From the above models we have seen that Random Forest with SMOTE oversampling on training dataset has yielded the best predictive results on test dataset.

Graphical user interface, table

Description automatically generatedFrom the statsmodel summary done as part of logistic regression, we can make the below inferences:

* Book\_value\_Adj\_Unit\_Curr has the highest negative coeffiecient which indicates that as the value of this predictor increases the chance for default decreases by 10% (i.e. 1-e^(-3.6150)).
* Total\_debt and Other\_Income has positive coefficient which tells us that higher these values the chances of payment default also increases.
* As the Current\_RatioLatest increases, which translates to company’s agility clearing of short term dues, it reduces the chance of defaulting.
* Groos\_Block has the highest positive coefficient and increase in this factor will lead to less chances of payment default.

Table-1.18 Logit feature importance

# 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.



First lets read the dataset and check few records.

Graphical user interface, application

Description automatically generated

Table-2.1 Input Data

Let us correct the headings to have no spaces.

A screenshot of a computer

Description automatically generated

Table-2.2 Modified column names

The number of rows (observations) is 314.

The number of columns (variables) is 11.

Lets look at the concise summary of data:

Text

Description automatically generated

Table-2.3 Concise Data Summary

We have one date variable(as object, will need to typecast as date variable), and stock price of 10 stocks.

There are no missing values in the dataset.

Lets look at the central measures of tendency:

Table

Description automatically generated

Table-2.4 Central Measures of Tendency

Let us convert ‘Date’ column from object to datetime and check the concise summary:

Text

Description automatically generated

Table-2.5 Modified concise info

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

Chart, scatter chart

Description automatically generated

Figure-2.1 Shree\_Cement stock price graph

Figure 2.1 shows the price graph of Shree\_Cement from the year 2014 to 2021. Overall we can see a increasing trend in price from 5543 to 24806. In the periods, 2015-2016 and 2018-2019 the stock has maintained its rates or dipped slightly, but in all other time periods it has a very positive trend.

Chart, scatter chart

Description automatically generated

Figure-2.2 Idea\_Vodafone stock price graph

Figure 2.2 shows the price graph of Idea\_Vodafone from the year 2014 to 2021. Overall we can see a decreasing trend in price from 80 to 3. In the periods,2014 to mid 2015 the price has increased and reached max of 120. From then on till 2021 the price has decreased sharply, except for 2016 to 2017 when the price has not moved much.

## 2.2 Calculate Returns for all stocks with inference.

We can calculate return by taking the log of values and calculating the difference of price value with price value of previous week.

Lets look at the head and tail of this data.

A screenshot of a computer

Description automatically generated with medium confidence

Table-2.6 Head of stock returns

Table

Description automatically generated

Table-2.7 tail of stock returns

The returns are all calculated on the base of natural log. If we look at table 2.6, and consider rows 312 and 313, we can see Infosys price has reduced less than say Mahindra&Mahindra in the same week, on their respective rates.

Let us look at the central measures of tendency of the stock returns:

Graphical user interface, application, table

Description automatically generated

Table-2.8 Stock return data description

If we look at the mean, it shows over the time period, on a weekly basis the price of Infosys is a positive value and hence the price is on an increasing trend, whereas the weekly mean is negative for Mahindra&Mahindra, which says that the price should be on a negative trend.

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

Lets calculate the stock return means and standard deviation for all stocks.

Graphical user interface, text, application

Description automatically generated

Table-2.9 Stock return mean/std

The average mean shows the trend of overall price movement whereas standard deviation indicates the volatility of the stock. If mean is positive the stock shows price appreciation and if negative then price depreciation. Similarly standard deviation when low, shows that the price does not fluctuate that drastically, whereas higher standard deviation indicates a stock which is very volatile and with high price fluctuations.

Lets compare Infosys and Indian\_Hotel stock We can see mean of Indian\_Hotel is lesser than Infosys by an order of 10. This means that Infosys price has appreciated much more from its starting price than Indian\_Hotel has. If we look at the standard deviation Infosys has lesser than Indian\_Hotel which tells us that Infosys stock movement is much more stable than Indian\_Hotel. If one were to choose investment between both the stocks for a longer duration, Infosys is a good choice as it has higher price appreciation and is more stable.

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

Lets look at the plot of mean vs std.

Chart, scatter chart

Description automatically generated

Figure-2.3 Stock returns: Mean vs std.

All stocks that are below y-axis value of 0, have their price depreciated, whereas all stocks above have price appreciated.

All stocks with least value of x axis are better in terms of stability and price volatility is lesser.

If we look at price appreciation alone the top 3 stocks are Shree\_Cement, Infosys and Axis Bank whereas if we see the top 3 stocks in terms of price volatility, it is Infosys,Shree\_Cement and Mahindra&Mahindra.

Idea\_Vodafone has the lowest mean and highest standard deviation, which tells us the price for this stock has fallen the maximum and the stock is the most volatile, hence this stock should be least preferred of the lot.

## 2.5 Conclusion and Recommendations

Let us also calculate the overall price change in stocks.

Graphical user interface, application

Description automatically generated

Table-2.10 Stock price movement

As we have surmised in section 2.4, Idea\_Vodafone stock has crashed, the price reducing by 96% in the period 2014 to 2021, whereas Shree\_Cement price has increased the most among the stock by 216% in the period 2014 to 2021.

For long term investment in stocks, it os safer to invest in stock with lower volatility and higher average mean stock return. Shree\_Cement, Infosys and Axis Bank seems to be the pick of the lot from this perspective.

Volatile stocks can also be invested with short term duration in mind, as the price varies a lot and if the trade is performed diligently, profits can be earned. Intraday trading and options are mostly traded with this fact in mind.

The stock returns plot can be used by portfolio managers to get a better sense of the stocks in their portfolio and devise strategies accordingly.

# END