Murali b

CREDIT RISK

**CONTENTS Page numbers**

**Table**

**1.1 Head of the data 3  
1.2 Tail of the data 3  
1.4 Describe the data 4  
1.29 Train set prediction 21  
1.30 Test set prediction 22**

**Figure**

**1.9 Removing before outliers 8  
1.10 After removing outliers 9  
1.14 Univariate analysis 12  
1.15 Correlation 13  
1.16 Heatmap 13  
1.27 AUC value&ROC curve of test set 20  
1.28 AUC value&ROC curve of train set 21  
1.33 Confusion matrix of test&train set 23  
1.34 AUC value&ROC curve of test set 24  
1.35 AUC value&ROC curve of train set 25  
1.44 AUC value&ROC curve of train set 26  
1.45 AUC value&ROC curve of test set 29**

**Output :**

**1.3 Shape of the data 4  
1.5 Column names 5  
1.6 Renaming column names 6  
1.7 Duplicate values 6  
1.8 Datatype details 6  
1.11 Missing values 10  
1.12 Missing values treatment 11  
1.13 Removing the columns 12  
1.17 Shape of the train data 14  
1.18 Shape of the test data 14  
1.19 Random Forest Classifier 14  
1.20 Best params 14  
1.21 Train set predictions 15  
1.22 Test set predictions 16  
1.23 Important columns 17  
1.24 Classification report of train set 18  
1.25 Classification report of test set 19  
1.26 Confusion matrix of train and test set 19  
1.31 Classification report of train test 22  
1.32 Classification report of test test 23  
1.36 Selector features 25  
1.37 Selector ranking 25  
1.38 Confusion matrix of train set 25  
1.39 Classification report of train test 26   
1.40 Confusion matrix of test set 26  
1.41 Classification report of test test 27  
1.42 Predictions on train set 27  
1.43 Predictions on test set 28**

**Credit Risk**

**Problem Statement**

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 can handle 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.

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

**Hints**:

Dependent variable - No need to create any new variable, as the 'Default' variable is already provided in the dataset, which can be considered as the dependent variable.

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.

Top 5 rows of the data:

A screenshot of a computer

Description automatically generated with low confidence  
Table 1.1 Head of the data

Lat 5 rows of the data:

A screenshot of a computer

Description automatically generated with low confidence

Table 1.2 Tail of the data

Shape of the data:   
  
Output 1.3 Shape of the data

There are 2058 observations with 58 columns.

Describe the data:   
A screenshot of a computer

Description automatically generated with low confidence  
Table 1.4 Describe the data

Columns names:

A screenshot of a computer screen

Description automatically generated with low confidence

Output 1.5 Column names

Renaming columns names:

As we can see column names has ‘\_’ so let’s fix it. After removing ‘\_’ then the column names looks like:



Output 1.6 Renaming column names.

Checking for duplicate values:   
  
Output 1.7 Duplicate values

We can see that no duplicate values present in the dataset.

Datatypes details:   
A screenshot of a computer screen

Description automatically generated with low confidence

Output 1.8 Datatype details

We see that 53 columns with float datatype , 4 columns with integer datatypes and 1 column with object type.

Outlier treatment:

A picture containing text, screenshot, parallel, diagram

Description automatically generated  
Fig 1.9 Before removing outliers.

We can see outliers present in almost every variable so let’s remove the outliers.

After removing the outliers:

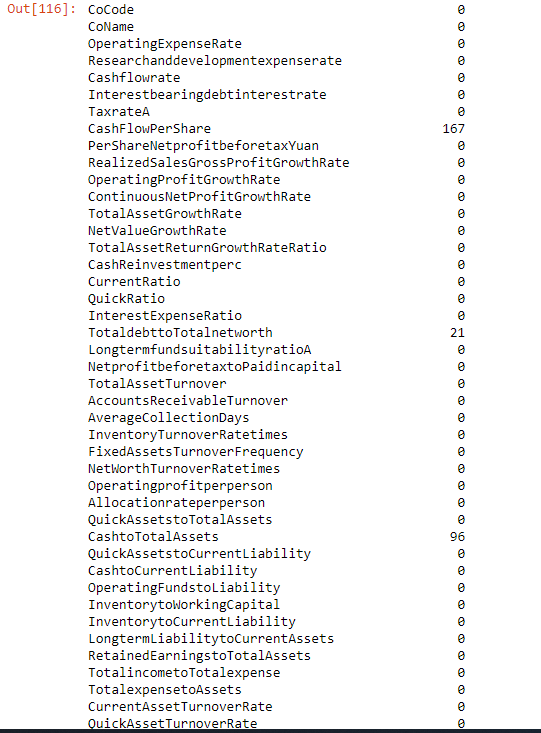
A picture containing text, screenshot, diagram, square

Description automatically generated

Fig 1.10 After removing outliers.

Now, we can see that there are no outliers present in the dataset.

Missing Value Treatment:

  
Output 1.11 Missing values

We see 167 missing values in Cashflowrate , 21 missing values in TotaldebttoTotalnetworth , 96 missing values in CashtoTotalAssests and 14 missing values in CurrentLiabilitytoCurrentAssets. So, we have to fix it further.

After removing Missing values:

A screenshot of a computer screen

Description automatically generated with low confidence

Output 1.12 After treating missing values.

So, all the missing values have been treated by imputing the values with median so now no missing values present in the dataset.

Removing the columns:

We have removed CoCode and CoName columns from the dataset.

A screenshot of a computer

Description automatically generated with low confidence  
Output 1.13 Removing the columns.

Univariate Analysis:

A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.14 Univariate Analysis

Correlation:   
A picture containing text, screenshot, number, font

Description automatically generated

Fig 1.15 Correlation

Heatmap:

A screenshot of a computer screen

Description automatically generated with low confidence  
Fig 1.16 Heatmap

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

Random Forest Classifier

Shape of the train data:



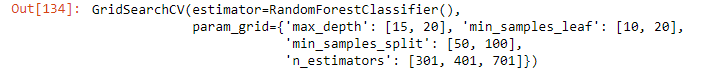
  
Output 1.17 Shape of train data

Shape of the test data:



  
Output 1.18 Shape of test data

Random Forest Classifier:

  
Output 1.19 Random Forest Classifier

Grid search best params:

A picture containing text, font, screenshot, white

Description automatically generated  
Output 1.20 Best params

Train set prediction:

A screenshot of a computer

Description automatically generated with low confidence  
Output 1.21 Train set probability

A screenshot of a computer

Description automatically generated with low confidence  
Output 1.22 Test set probability A screenshot of a computer

Description automatically generated with medium confidence

Output 1.23 Important columns

We see most important columns are NetValueGrowthRate , NetprofitbeforetaxtoPaidin capital , PerShareNetprofitbeforetaxYuan , RetainedEarningstoTotalAssets and TotalincometoTotalexpense.

We have fitted the model to the GridSearchCV for random forest model by including the model with both independent variables and dependent variables. So, we shall use Scikit\_Learn, RandomisedSearchCV function and we hereby thus will create a grid of ranges involving hyperparameters and sampled randomly from the grid which thus performs K-fold CV with all possible combinations of values.

Detailed information about the variables used are:

Max\_features = which means maximum number of features which is considered for splitting the node.  
n\_estimators = which shows the number of trees in the forest.   
max\_depth = which shows the number of levels in each decision tree.  
min\_samples\_leaf = it will show the minimum number of data points allowed in a leaf node.   
min\_samples\_split = which shows the minimum number of data points which are placed in a node before the nodes are split.

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

Classification report of the train set:

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.24 Classification report of the train set

Train data inferences:

Precision: 83%  
Accuracy: 93%  
Recall: 46%  
F1 value: 59%

Classification report of the test set:

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.25 Classification report of the test set

Test data inferences:

Precision: 65%  
Accuracy: 92%  
Recall: 36%  
F1 value: 46%

Confusion matrix of train and test set:

A picture containing text, screenshot, diagram, number

Description automatically generated  
Output 1.26 Confusion matrix of the train and test set

Training data Inferences:

True Negative: 1210

False Positive: 15

False Negative: 82

True Positive: 71

Test data Inferences:

True Negative: 1034

False Positive: 7

False Negative: 23

True Positive: 120

AUC value and ROC curve of test data:

A picture containing line, text, plot, diagram

Description automatically generated  
Fig 1.27 AUC value and ROC curve of test data:

AUC value and ROC curve of train data:

A picture containing text, line, plot, diagram

Description automatically generated  
Fig 1.28 AUC value and ROC curve of train data:

Recall values has been increased from 36% to 46% and f1 score value also increased from 46% to 59% hence the model is good to proceed further for analysis.

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

Train test prediction:

A screenshot of a computer

Description automatically generated with low confidence  
Table 1.29 Train test probability

Test set prediction:

A screenshot of a computer

Description automatically generated with low confidence  
Table 1.30 Test test probability

We use sklearn library in Linear Discriminant Analysis and LDA is mainly used for modelling differences in groups that is separating two or more classes and it is a dimensionality reduction technique which is used for the supervised classification problems. Also, it is used to showcase the features in higher dimension space into a lower dimension space. We use GridSearchCV and parameters used are: max\_iter , penalty , solver , tol are used to find the best grid for prediction of the model.

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

Classification report of the train set:

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.31 Classification report of the train set

Train data inferences:

Precision: 60%  
Accuracy: 90%  
Recall: 33%  
F1 value: 42%

Classification report of the test set:   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.32 Classification report of the test set

Test data inferences:

Precision: 41%  
Accuracy: 89%  
Recall: 39%  
F1 value: 40%

Recall value has been already increased from 33% to 39% and f1 score also increased from 40% to 42% hence the model is good for analysis.

Confusion matrix of test and train set:

A picture containing text, screenshot, diagram, font

Description automatically generated  
Fig 1.33 Confusion matrix of test and train set

Train data Inferences:

True Negative: 1192

False Positive: 33

False Negative: 103

True Positive: 50

Test data Inferences:

True Negative: 576

False Positive: 37

False Negative: 41

True Positive: 26

AUC value and ROC curve of train data:

A picture containing text, line, plot, diagram

Description automatically generated  
Fig 1.34 AUC value and ROC curve of train data

AUC value and ROC curve of test data:

A picture containing text, line, plot, diagram

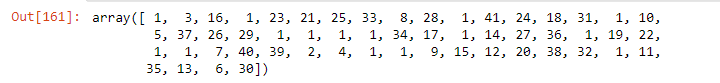
Description automatically generated  
Fig 1.35 AUC value and ROC curve of train data

Logistic regression:

Selector features

  
Output 1.36 selector features

Selector ranking:

  
Output 1.37 Selector ranking

Confusion matrix of train set:

A picture containing text, font, design

Description automatically generated  
Output 1.38 Confusion matrix of train set

Train data Inferences:

True Negative: 1190

False Positive: 35

False Negative: 147

True Positive: 6

Classification report of train set:

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.39 Classification report of train set

Train data inferences:

Precision: 15%  
Accuracy: 87%  
Recall: 0.04%  
F1 value: 0.06%

Confusion matrix of test set:

  
Output 1.40 Confusion matrix of test set

Train data Inferences:

True Negative: 587

False Positive: 26

False Negative: 64

True Positive: 3

Classification report of test set:

A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.41 Classification report of test set

Test data inferences:

Precision: 10%  
Accuracy: 87%  
Recall: 0.04%  
F1 value: 0.06%

Probability of train set:

A screenshot of a computer

Description automatically generated with low confidence

Output 1.42 Train set predictions

Probability of test set:

A screenshot of a computer

Description automatically generated with low confidence  
Output 1.43 Test set predictions

AUC value and ROC curve of train data: A picture containing text, line, plot, diagram

Description automatically generated

Figure 1.44 AUC value and ROC curve of train data

AUC value and ROC curve of test data:

A picture containing text, line, plot, diagram

Description automatically generated  
Figure 1.45 AUC value and ROC curve of test data

Mainly for categorical response data, we use logistic regression to estimate the probability of the binary response which occurs mainly on predictor variables. RFE is used to configure and to use as it is effective way of selecting features columns in a train set that are used to predict the target variables. It uses filter-based feature selection internally and works by searching for a subset of features by starting with all features in the train set and thus removes features until unless we get the desired number. We shall fit the model and thus analysis the ranking features based on importance and removing the least important features by re-fitting the model and shall keep on repeating until we get the number of features.

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

We can see from the above analysis that ROC curve is not overfitted nor unfitted and Random Forest with GridSearchCV performed better with high recall value and F1 score value. So Random Forest is the best model which can be used for credit risk analysis. The model with selected important features can predict a high probability of default. The credit report analysis will depict the details of the credit worthiness of a potential customer. So, we can classify the model based on defaulters like very high risk, high risk, medium risk and low risk.

1.13 State Recommendations from the above models

We were able to get the recall value without overfitting the model and so the model is a good model as we have fixed the outliers, missing values and can be improved if the data received is of qualified features with less missing values. To predict the best model, we sued recall and precision values by minimizing the Type 2 errors so we can predict many actuals defaults in the model.