MA515 PROJECT REPORT

Foundations of Data Science

A Project Report Submitted by

Anubothula Hema Priya
2020MCB1228

Mathematics and Computing

Under the supervision of
Dr.Arun Kumar

Department of Mathematics

Indian Institute of Technology, Ropar

December 1, 2022

Contents

Problem Statement :

Do Exploratory data analysis on the data. Predict the bankruptcy using logistic, LDA and QDA. Construct the AUC and ROC and compare results from different methods.

Basic Exploratory Analysis :

- Find the summary of the data using data.describe() command. It gives count, mean, standard deviation, minimum and maximum values of each column.
- The data.shape command gives dimension of data
- The basic information like non null values count and data type of each column can be obtained by data.info() function.
- Plot the heat map to visualize the correlation between features.

Data Preprocessing :

1. Removing Missing Data :

We will check for any NaN values and remove them if they are any. In given data, there are no NaN values.

2. Dividing columns into predictor and target variables. (X and y)

3. Remove highly correlated features using a threshold of 0.9.

4. Splitting the dataset into training and testing :

Take test_size = 0.25, divide the dataset into training and testing.

5. Feature Scaling :

If some column values are much larger than others we will scale those column values using Standard Scaler.

Steps for implementing classifiers :

1. Fit the X_train, y_train on the classifier chosen.

2. Predict the values based on training and testing data.

3. Calculate and plot confusion matrix for training and testing data.

4. Calculate accuracy for training and testing data.

Observations :

1. Logistic Regression as Classifier
   - Logistic Regression Training Performance:
     [[4924 22]
     [125 43]]
   - Logistic Regression Testing Performance:
     [[1639 14]
     [43 9]]

   - Logistic Regression Training Accuracy: 97.12553773953853%
   - Logistic Regression Testing Accuracy: 96.65689149560117%

2. LDA as Classifier
   - LDA Training Performance:
     [[4900 46]
     [119 49]]
   - LDA Testing Performance:
     [[1632 21]
     [39 13]]
   - LDA Training Accuracy: 96.77356276886977%
   - LDA Testing Accuracy: 96.48093841642229%

3. QDA as Classifier

- QDA Training Performance:
  [[1810 3136]
  [0 168]]
- QDA Testing Performance:
  [[669 984]
  [15 37]]
- QDA Training Accuracy: 38.678138443488464%
- QDA Testing Accuracy: 41.407624633431084%

Comparison :

- Based on Training Accuracy:
  1) QDA performs worst among the three.
  2) Logistic Regression has a slightly greater accuracy than LDA (~1%)
- Based on Testing Accuracy:
  1) QDA performs worst among the three.
  2) Logistic Regression and LDA have nearly the same accuracy.

Here, Logistic Regression and LDA both are good models but Logistic Regression is better (based on testing accuracy) compared to LDA but slightly overfitted (based on training accuracy).

AUC and ROC curve :

The area under the ROC curve (AUC) results are considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6.

AUC Scores :

- AUC Score of Logistic Regression = 0.8916189678440132
- AUC Score of LDA = 0.9209246591279259
- AUC Score of QDA = 0.5675287356321839

Conclusions :

- The LDA is considered to be an excellent classifier among three based on the given dataset.
- Logistic Regression is a good classifier whereas QDA is a poor or failed classifier.