Lab -3

PRML
AY 2020-21 Trimester - III

Random Forest and Bagging

Name : Akshaykumar Kanani (B19EE008)

Problem:

Dataset:- Consider the credit sample dataset, and predict whether a customer will repay their credit within 90 days. This is a binary classification problem; we will assign customers into good or bad categories based on our prediction.

Data Description:-

Features	**Variable Type**	**Value Type**	**Description**
Age	Input Feature	integer	Customer age
Debt Ratio	Input Feature	real	Total monthly loan payments (loan, alimony, etc.) / Total monthly income percentage.
Number_Of_Time_30-59_Days_Past_Due	Input Feature	integer	The number of cases when a client has overdue 30-59 days (not worse) on other loans during the last 2 years.
Number_Of_Time_60-89_Days_Past_Due	Input Feature	integer	A number of cases when the customer has 60-89dpd (not worse) during the last 2 years.
Number_Of_Times_90_Days_Late	Input Feature	integer	Number of cases when a customer had 90+dpd overdue on other credits
Dependents	Input Feature	integer	The number of customer dependents
Serious_Dlq_in_2yrs	Target Variable	Binary: 0 or 1	The customer hasn't paid the loan debt within 90 days

Perform the following tasks for this dataset:-

Question-1 (Random Forest): (Total 20 Marks)

1. Preprocessing the data. (5 Marks)
    a. Plot the distribution of the target variable.
    b. Handle the NaN values.
    c. Visualize the distribution of data for every feature.
2. Train the Random Forest Classifier with the different parameters, for e.g.:- (5 Marks)
    i. Max_features = [1,2,4]
    ii. Max_depth = [2,3,4,5]
3. Perform 5 fold cross-validation and look at the ROC AUC against different values of the parameters (you may use Stratified KFold function for this) and Perform the grid-search for the parameters to find the optimal value of the parameters. (you may use GridSearchCV for this) (5 Marks)
4. Get the best score from the grid search. (2 Marks)
5. Find the feature which has the weakest impact in the Random Forest Model. Briefly justify your answer. (3 Marks)

Question-2 (Bagging) : (Total 20 Marks)

6. Perform bagging-based classification using Decision Tree as the base classifier. (15 Marks)
    a. The number of trees to be considered is {2,3,4}.
    b. Perform 5 fold cross-validation using ROC AUC metric to evaluate the models and collect the cross-validation scores (use function cross_val_score for this).
    c. Summarize the performance by getting mean and standard deviation of scores
    d. Plot the model performance for comparison using boxplot.
7. Compare the best performance of bagging with random forest by plotting using boxplot. (5 marks)

Here is the Colab notebook attached for your reference.

Instructions:-

Please Submit the necessary code(s) (Notebook) and a PDF explaining and analyzing the steps in both the questions along with necessary plots/figures.

Note:- No submission will be accepted after the final deadline.

Answer :

Question 1 :
1-4 part of the question is done on code file.

5) from the graph of comparison between feachers we can clearly see that dependence has the lowest impact. Because it has only 0.3 partial impact which is less compared to other feachers. So we can say that dependence is the misleading column in our data set.

Other important detail on 1st question :
Best score from grid search = 0.8194986914865716
Best estimator grid value = 0.8606060606060606

Question 2:
6) A,B,D part : done on code file.
C part :

```
->for 2 trees mean : 0.749 standard deviation : 0.034
->for 3 trees mean : 0.751 standard deviation : 0.039
->for 4 trees mean : 0.765 standard deviation : 0.026
```

7)
Bagging : is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. Bagging is a special case of the model averaging approach.

by comparing the best classifier of bagging and random forest we get :

```
Accuracy of Random Forest Model : 0.861
Accuracy of Best model of Bagging : 0.765
```

So from this we can say that for this data set and randomized method Random Forest classifier is good compare to bagging method.
Normally bagging method is good compare to random forest but here we only use one decision tree as bagging classifier. So thats why we got higher accuracy on random forest model.

All other Details are present on Code file.
Thank you