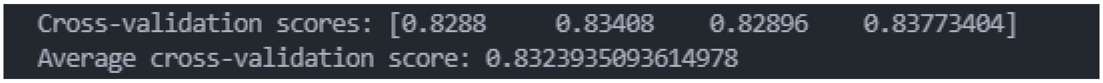
Stone Barrett

CSE 590: Intro to Machine Learning

Homework 3

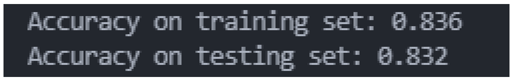
In this assignment, we will be exploring the effectiveness of three different classification algorithms when applied to a given IMDB rating dataset. The IMDB dataset has been heavily preprocessed, with each of the 25000 entries having 1000 binary features to sift through. The three algorithms in question are the Multinomial Naïve Bayes classifier, the Random Forest classifier, and the Gradient Boosted Regression Trees classifier. The goal of this experiment is to determine the optimal algorithm for this dataset based on computational time and prediction accuracy as well as the optimal parameters for each algorithm.

Figures



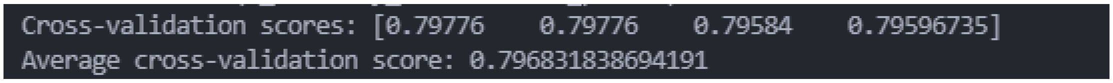
```
Cross-validation scores: [0.8288     0.83408    0.82896    0.83773404]
Average cross-validation score: 0.8323935093614978
```

Figure 3.1



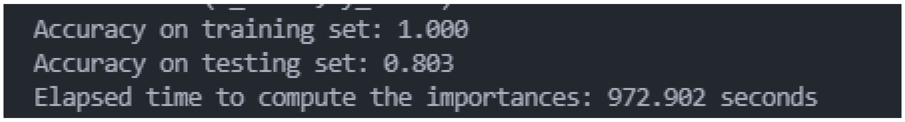
```
Accuracy on training set: 0.836
Accuracy on testing set: 0.832
```

Figure 3.2



```
Cross-validation scores: [0.79776    0.79776    0.79584    0.79596735]
Average cross-validation score: 0.796831838694191
```

Figure 3.3



```
Accuracy on training set: 1.000
Accuracy on testing set: 0.803
Elapsed time to compute the importances: 972.902 seconds
```

Figure 3.4

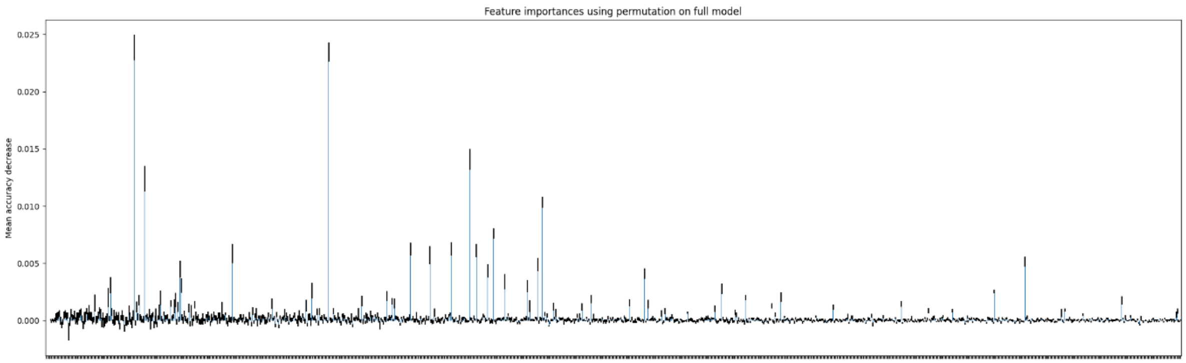
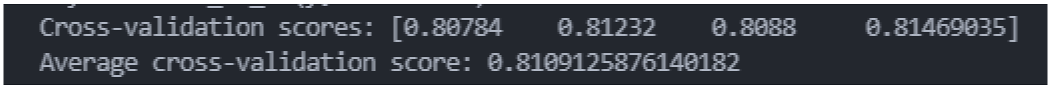
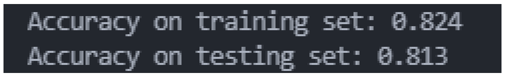


Figure 3.5



```
Cross-validation scores: [0.80784    0.81232    0.8088     0.81469035]
Average cross-validation score: 0.8109125876140182
```

Figure 3.6



```
Accuracy on training set: 0.824
Accuracy on testing set: 0.813
```

Figure 3.7

Problem 1a: Multinomial Naïve Bayes

The first algorithm we'll be testing on the provided dataset is the Multinomial Naïve Bayes (MNB) classifier. Before running the algorithm in isolation, we will run it using cross-validation. K-fold cross-validation is when the data is sectioned or "folded" into k pieces in order to see if there is a tendency for part of the data to be better fit than others. For this assignment, we will be using 4 folds. Figure 3.1 shows the accuracy scores for each fold of the data as well as the average of those scores. Running the algorithm without cross-validation shows a nearly identical accuracy in Figure 3.2. This algorithm shows good results for this dataset, as the training and testing accuracies are very close and no overfitting or underfitting were observed. This model was ran very quickly and without much computational power needed.

Problem 1b: Random Forest

Next, we will look at the Random Forest (RF) classifier. The RF algorithm uses a combination of many decision trees that were built with different features of focus. Random number generation (RNG) is used to make sure each tree is unique (or at least very different) – hence the name "Random Forest". Figure 3.3 shows the 4 cross-validation scores and their average when ran with this algorithm. It is of note that the accuracy is slightly lower than MNB's. Figure 3.4 shows the accuracy scores of the training data and testing data without having been cross-validated. Given that the training set score is 100%, we know that there is overfitting of the highest level happening with this model. Increasing the number of estimators from 30 to a higher value could potentially help with the overfitting issue generally speaking, though RF is prone to overfitting with datasets of this size and complexity. In this case, testing this algorithm with 100 estimators did not result in much of a difference. Next, we will inspect feature importance just for curiosity's sake. Figure 3.4 also shows how long it took to compute the graph output in 3.5. Since there are literally 1000 unnamed features, it is impossible to identify which specific ones are more important. However, Figure 3.5 does show that some features are significantly more important than others. The RF algorithm required fairly little time and computational power.

Problem 1c: Gradient Boosted Regression Trees

Lastly, we will be examining the effectiveness of the Gradient Boosted Regression Trees (GBRT) classifier. Figures 3.6 and 3.7 show the cross-validation scores and average and the training and testing scores without cross-validation. It is of note that these scores are only slightly lower than MNB and no overfitting or underfitting are observed. Reducing max depth and reducing learning rate actually made the results worse in this case. This algorithm took slightly longer to run than the other two.

Conclusions

After testing these three models on the provided dataset, it would seem there is a clear answer to which algorithm is best fit. The Multinomial Naïve Bayes classifier ran slightly faster than the other two while providing slightly higher accuracy and no observed instances of overfitting or underfitting.