Machine Learning Algorithm Selection

Finn Tomasula Martin

COSC-4557

---

1 Introduction

There are many important aspects when creating machine learning models. One of the most important is the algorithm used to build the model. The algorithm can heavily influence the result of model as it is what dictates how the learning is carried out. So, an important question to ask when planning to build a machine learning model is, what algorithm should I use? You may be able to answer this question by examining the nature of your problem as some times one algorithm may be more naturally suited to a specific problem than others. But often times it can be hard to tell what algorithm will work the best. So, another method is to create models using a set of algorithms and then comparing their performance. In this report, we will go over the process of evaluating a few algorithms on the same dataset and deciding which one we should use for our model.

2 Data

For this report we will be taking a look at the winequality-red.csv dataset. This dataset contains 1,599 observations on various red wines. One of the variables present is “quality” which gives a quality score to each observation on a scale of 3 – 8 (3 being the lowest quality and 8 being the highest). For our purposes we will add a new variable “quality.bin” which will attribute a quality of “good” or “bad” depending on if the observation’s quality was above the average or not. We will be using this variable as our target and the remaining 11 variables will be the features.

3 Experimental Setup

We will be running our analysis in R using the following libraries: caret, randomForest and e1071. The only data processing that needs to be done is the addition of the quality.bin target variable discussed in the data section. We will be comparing the performance of three algorithms: logistic regression, random forest and SVM. To evaluate the algorithms we will use 10-fold cross validation and classification accuracy. To do this have to split our data into 10 random but consistent folds. Then for each fold we will do the following process: Set the current fold as our test set and all the others as our training set. Build the model using the training set. Predict the target variable of all observations in the test set using the model. Compute the classification accuracy by comparing the predicted target with

the actual target from the test set, summing the correct predictions and dividing that by the total observations. By the end we will have a list of the classification accuracy from each of the 10 folds. We will run this process on each model and then compare them using the average classification accuracy and visualize the distribution in a box plot.

4 Results

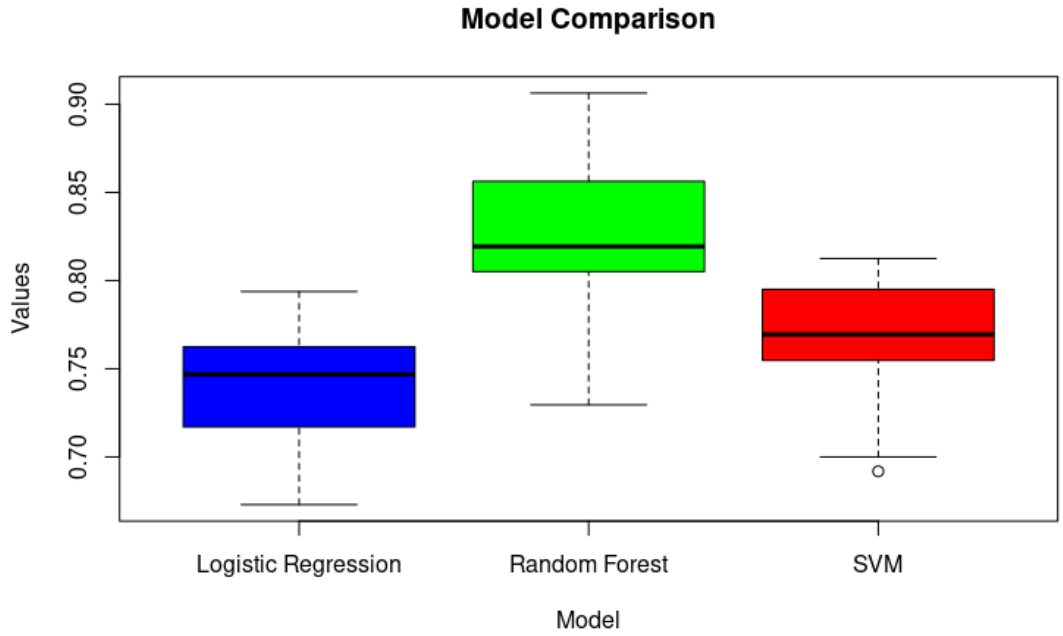
After running our analysis, we are left with the following results:

Logistic Regression: 0.717, 0.752, 0.739, 0.781, 0.673, 0.7421, 0.713, 0.756, 0.794, 0.763
Average: 0.743

Random Forest: 0.805, 0.820 0.857, 0.856, 0.730, 0.805, 0.788, 0.825, 0.819, 0.906
Average: 0.821

SVM: 0.755, 0.776, 0.795, 0.781, 0.692, 0.755, 0.700, 0.763, 0.806, 0.813
Average: 0.764

As you can see random forest yielded the best results with a classification accuracy of 0.821. We can further qualify this observation by taking a look at a box plot depicting the model.



Once again, we see that random forest performs the best for this dataset. Based on these results we can conclude that the best algorithm between these three is random forest and that is the one we should use for our model.

5 Code

All code is in file, AlgorithmSelection.R

https://github.com/COSC5557/ml-algorithm-selection-ftomasul/blob/main/AlgorithmSelection.R

Sources:

https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/

https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/