* What is the use of training and testing sets when creating a random forest model?

Training Sets and Testing Sets are used to test the model’s validity. You can use the training set to set the initial assumptions of the model. The Testing data set tests how well that training data set trained the model. You can alter the ratio of training/testing data the data set is split into to adjust how well the Training dataset fits the model. Overfit models are not as robust as a less fit version of a model based on the same data.

* What are the accuracy, precision, and recall values for the model in Step 6? Hint: Use the confusion matrix to calculate these values.

I am going to assume that these values are requested of the Testing Dataset results. Using the Training Dataset won’t be as fruitful when testing a model’s robustness.

Table

Description automatically generated with medium confidence

True positive = 37, True Negative = 28, False positive = 13, False negative = 14

Accuracy = 70.65%

Precision = 74%

Recall = 72.55%

* Review the graph in Step 5. What happens to the training and testing classification errors as the number of trees grows?

As the number of trees grows the classification error stabilizes. The available data can only randomly generate so many classification trees before the averaging of all of those trees plateaus. This value can be used to determine the minimum amount of trees for the most accuracy. There are limited returns as you continue to increase the number of trees.

* How do we find the optimal number of trees to use in our random forest model using this graph?

Determine the point where the error rate plateaus and use that value. There is no need to use more trees than necessary.

* Review the graph in Step 10. What is the optimal number of trees to use in order to minimize the mean squared error?

18-20 trees would be optimal. 20 for good measure.

* What is the benefit of using a random forest model over a single decision tree?

A random forest model brings multiple decision trees to the table. Different decision trees are likely to be accurate in different scenarios. Real world datasets are so variable that it makes sense to take into account various models instead of applying only one.

* What difficulties, if any, are you having with the analyses for Project Two?

None. I finished Project Two today. The only issue I had was transposing some of the code for Classification Trees and the Loop for the Random Forest model. \*\*\*MAKE SURE TO CHECK ALL VARIABLE NAMES RELEVANT TO THE DATASET ARE PROPERLY TRANSPOSED\*\*\*. Specifically the column indicated by the “train.data” reference and the column names in the confusion matrix. 😊