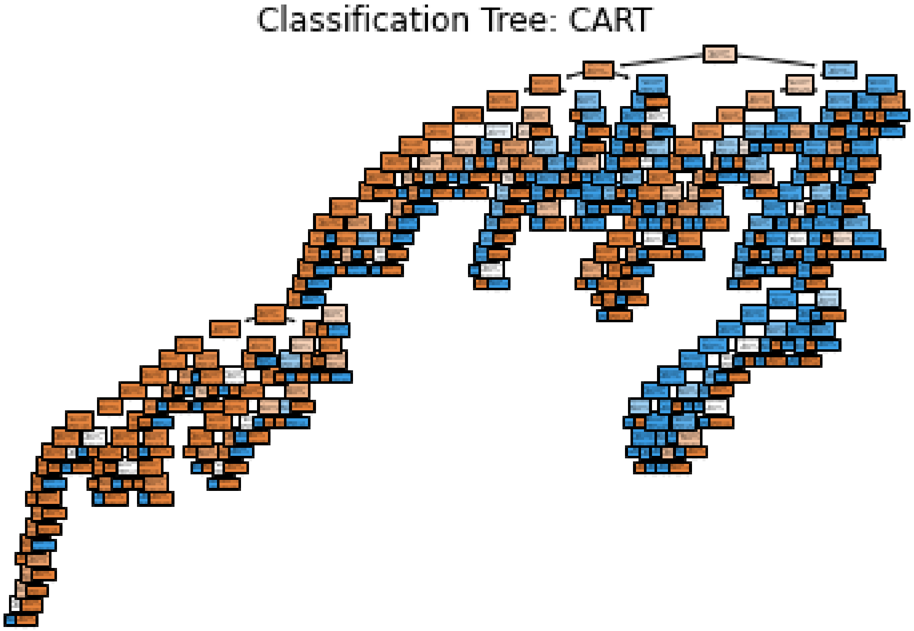
Decision Tree and Random forest for email spam classifier

After imputing missing values with zero and splitting the data set into training and test sets while shuffling, a CART model was built on the training data. The resulting fitted classification tree is quite large as no maximum depth or number of trees was specified for this model. While this may result in a model that can better represent the possible intricacies of the data, it is also possible that this will result in overfitting. The fitted classification tree can be seen below:

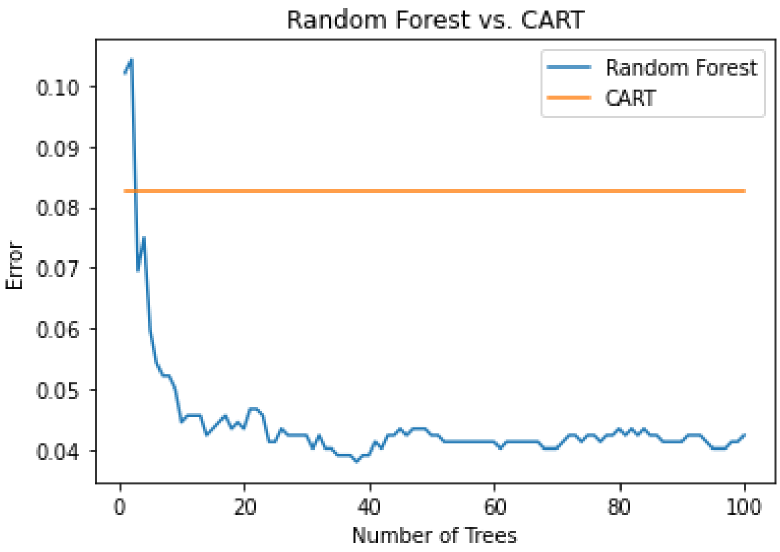


A random forest model with no maximum depth or number of trees similar to the CART model was fit on the training data. Next, predictions were made using the test set of the CART and random forest models, and the test accuracy of each model was calculated to determine the misclassification error rate. These error rates can be seen below:

CART Test Error: 0.08251900108577637

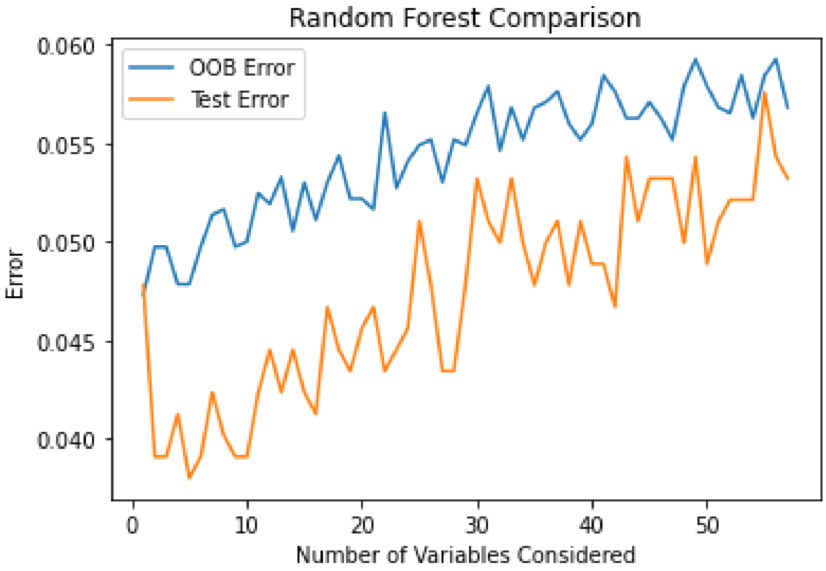
Random Forest Test Error: 0.04234527687296419

From this, it can be seen that the random forest model achieved a lower test error rate than the CART model, which is to be expected as the random forest model is able to avoid overfitting to the noise in the data better than a large CART model. To understand the effect of the inclusion of more trees in the random forest model, the plot of the curve of the test error versus the number of trees for the random forest can be seen below, which the test error for the CART model as a reference:



From this, it can be seen that the inclusion of more trees in the random forest model quickly decreases the test error rate below the threshold of the CART model, which is likely overfit to the noise in the training data. It should be noted, however, that the inclusion of more trees past approximately 40 trees did not see a large improvement in the performance of the random forest model showing that this is likely an appropriate choice for the number of trees for this model.

The number of variables selected at random to split was tested between the minimum value, one, and the maximum value, all the variables. A random forest model was fit for each of choice of the parameter, and the OOB error and test error were calculated for each. These error estimates can be seen below against the range of values for the parameter v that were tested:



From this, it can be seen that the OOB error and test error was quite similar and behaved in a similar fashion as the number of variables considered increase, with the OOB error being slightly higher than that of the test error for each model. This slight discrepancy is to be expected, however, since the OOB error is an estimate of the error in which all of the trees in the random forest are not considered when calculating the error of each observation whereas all the trees would be considered when computing the test error. From this plot, it appears that as more variables are considered to be split randomly, both the OOB error and test error tend to increase. This is likely due to there being more correlation between pairs of bagged trees as the number of variables considered increase, which in return limits the benefit of averaging the trees in the random forest.