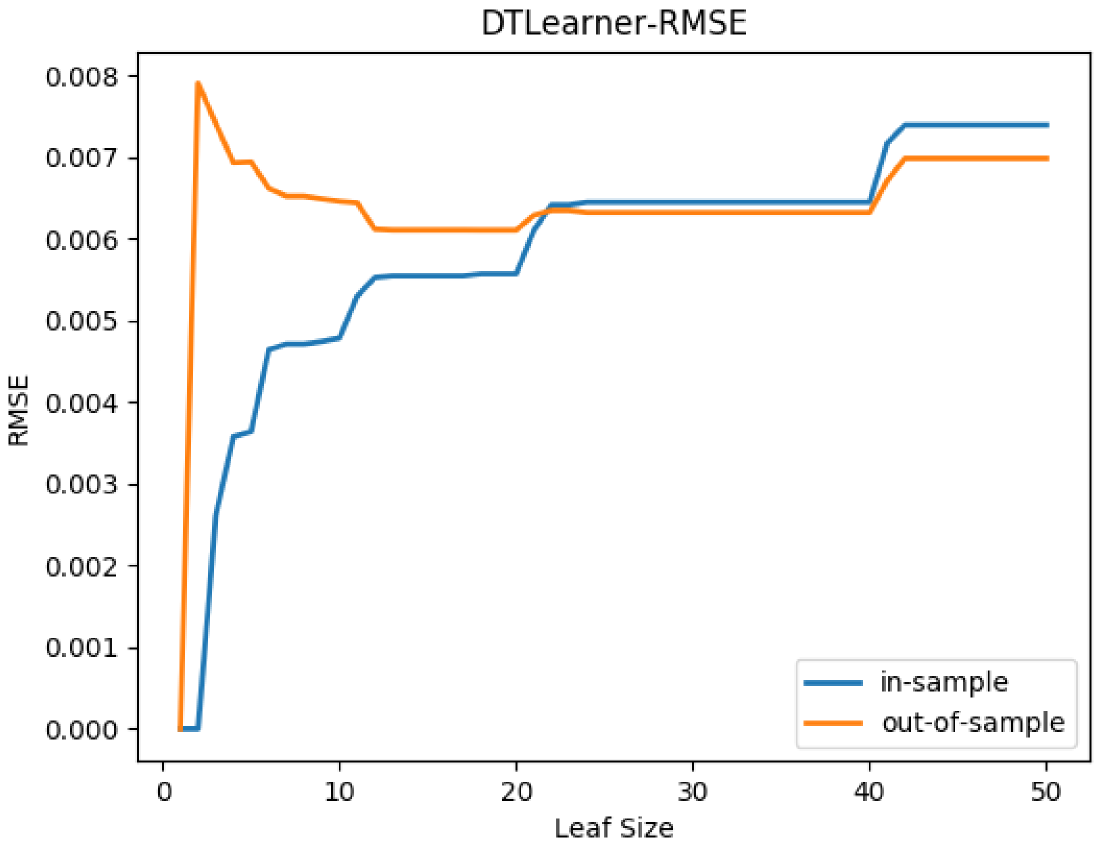
Hussain Mumtaz
CS 4646

Report: Assess Learners

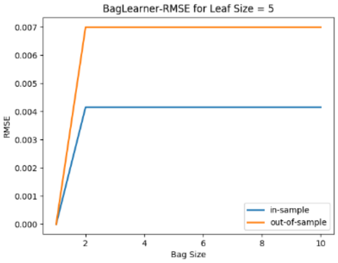
1. Does overfitting occur with respect to leaf_size? Consider the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).

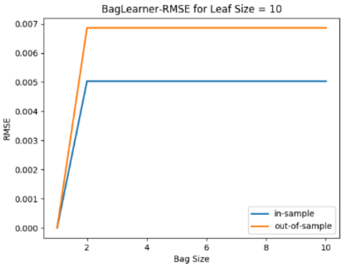
   Overfitting occurs with respect to leaf_size. The model overfits the data for approximately leaf_size < 15. Around that value the out-of-sample RMSE continues to go down before stabilizing and then increasing at which point the model begins to underfit. The lower the leaf_size the more the model overfits the training data. This makes sense because the lower the leaf_size is the more specifically the model is training to categorize the training data.

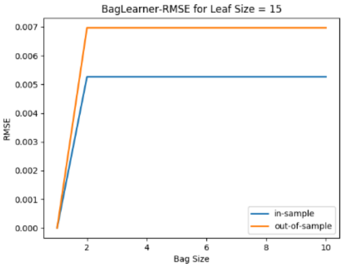


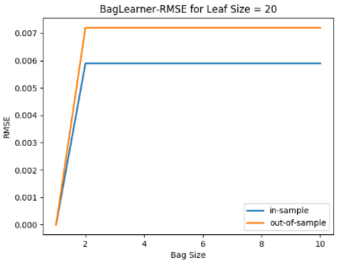
2. Can bagging reduce or eliminate overfitting with respect to leaf_size? Again consider the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts and/or tables to validate your conclusions.

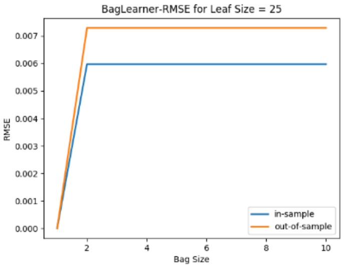
   Bagging does not reduce overfitting with respect to leaf_size with DTLearner. Consider the progress of DTLearner with leaf_size going from 1 to 50:

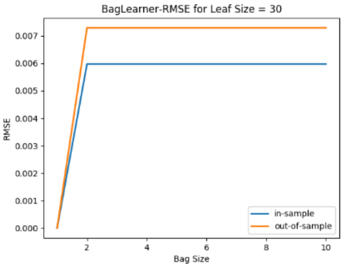


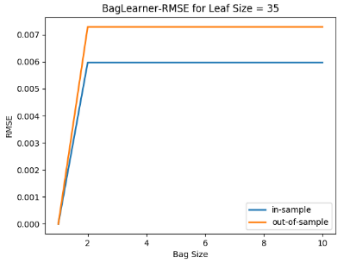


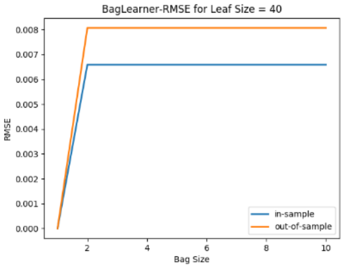


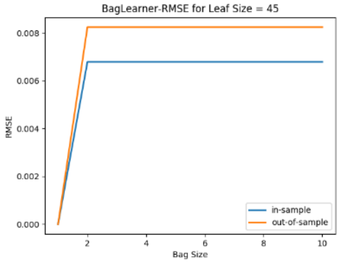


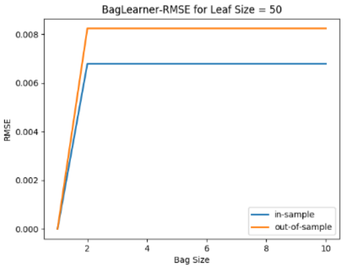








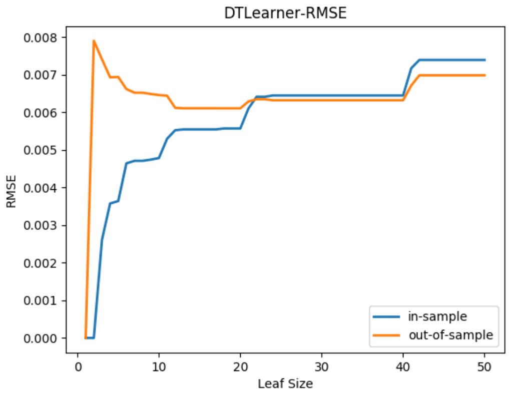


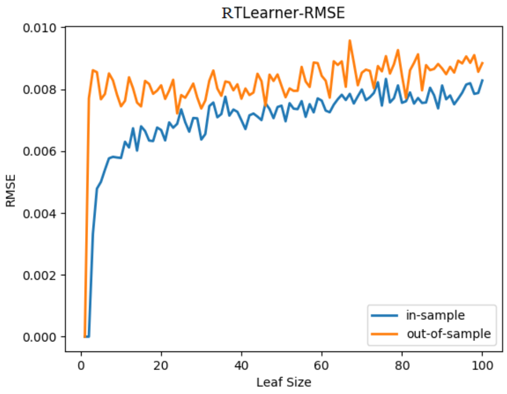


   The in-sample RMSE increases marginally as the leaf_size increases, but the out-of-sample RMSE changes roughly the same as a single DTLearner, so despite the in-sample RMSE increasing, the model generalizes to the testing data about the same. The aggregation the bagging does on DTLearner does not cause a big change in out-of-sample RMSE, because decision trees are created through a formulaic approach in which the best correlated variables are favored towards the top. Repeating this approach over and over produces the same or similar trees. Considering a single graph from above we see this as the number of bags increase the RMSE is the same. As the leaf_size increases, with a single DTLearner we see that the out-of-sample RMSE forms a basin-like shape. If we were to have the graphs above in a single graph showing leaf_size on the x-axis, we would likely see a similar shape.

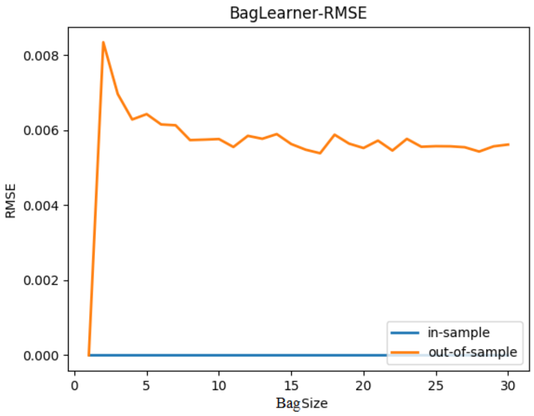
3. Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other?

If we look at the RMSE of a DTLearner versus the RMSE of a DTLearner you see that the DTLearner by itself is much more accurate:





The RTLearner however runs much faster in comparison. We can capitalize on this speed by using bagging over the RTLearner. We see that bagging random trees for a bag size greater than 5 we get results better than a single DTLearner.


BagLearner-RMSE
RMSE
Bag Size
in-sample
out-of-sample