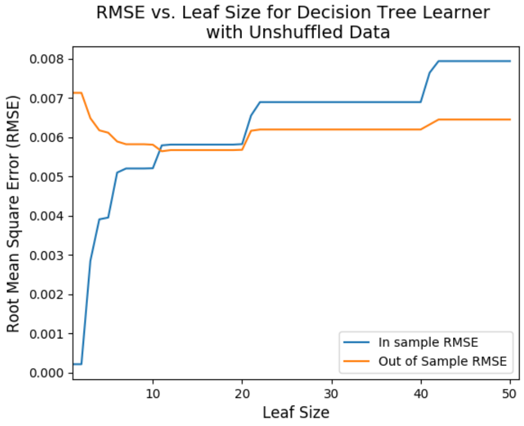
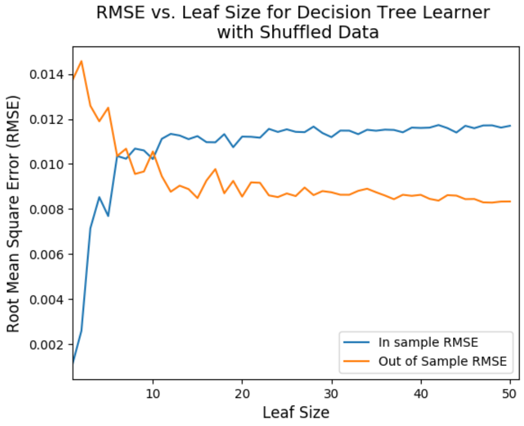
Matthew L. Miller
mmiller319

1. For the Decision Tree Learner, overfitting clearly occurs when the leaf size is low, especially less than 10. An indication of overfitting is when the RMSEs of the in-sample and out-of-sample data diverge. In the results of my experiments with the Decision Tree Learner, the in-sample RMSE increases from less than 0.001 to approximately 0.005 as the leaf size increases from 1 to 5. Meanwhile, the RMSE of the out-of-sample data decreases from 0.007 to about 0.0065 for the same leaf sizes.

I also ran experiments with the Decision Tree Learner where the testing and training data were shuffled prior to each iteration. The results of both shuffled and unshuffled data show a similar pattern in regard to overfitting occurring at smaller leaf sizes, especially leaf sizes less than five or so. However, the RMSEs for both in-sample and out-of-sample data are higher when the data is shuffled. The RMSE of the unshuffled in-sample data reached a value of 0.008 when the leaf size exceeded 40, while the unshuffled out-of-sample RMSE was approximately 0.006 for leaf sizes greater then 10. For the shuffled data, however, the in-sample RMSE plateaued around 0.010 - 0.012 when the leaf size exceeded 10 while the out-of-sample RMSE started at over 0.014 for leaf sizes close to one and leveled off at approximately 0.008 for leaf sizes greater than 20. The most likely cause of the higher RMSEs is the element of randomness that is introduced by shuffling the data.

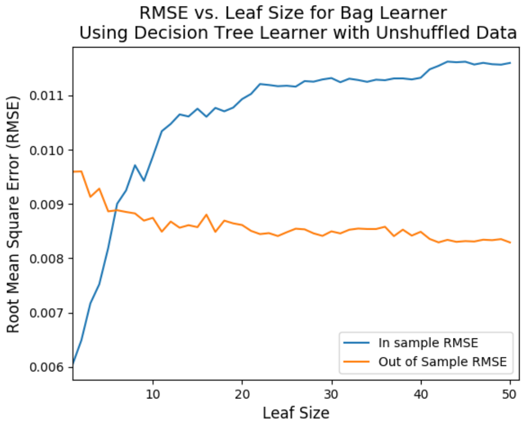

RMSE vs. Leaf Size for Decision Tree Learner with Unshuffled Data

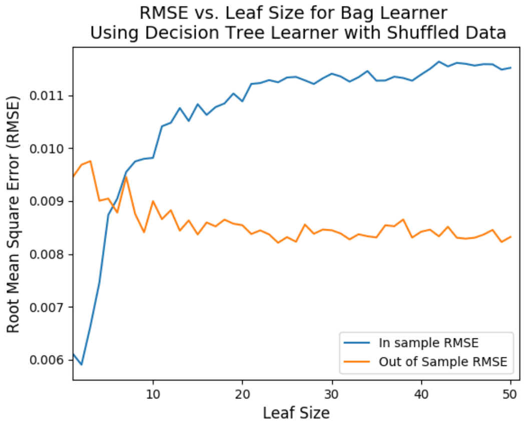

RMSE vs. Leaf Size for Decision Tree Learner with Shuffled Data

2. My experiments with bagging found no effect on the occurrence of overfitting. The divergence of the RMSEs for in-sample and out-of-sample data still occurred for small leaf sizes, especially when the leaf size was five or less. I used a bag size of 20 and yet the overfitting appears nearly the same as it did without bagging.

One difference, however, was that the RMSEs were clearly higher when using bagging, both for shuffled and unshuffled data. The in-sample RMSEs for both shuffled and unshuffled data had a low of about 0.006 for a leaf size of one and plateaued at 0.011 - 0.012 when the leaf size was greater than 20. The out-of-sample RMSEs for both shuffled and unshuffled data had a high of approximately 0.0095 when leaf size was close to one and leveled off between 0.008 - 0.009.

The RMSEs for the Bag Learner more closely resemble the RMSEs for the Decision Tree Learner with shuffled data. The most likely explanation is that the Bag Learner introduces are larger variance in the training and testing because it is generating a sample from the dataset for each bag, which produced an effect similar to shuffling the data for the Decision Tree Learner.

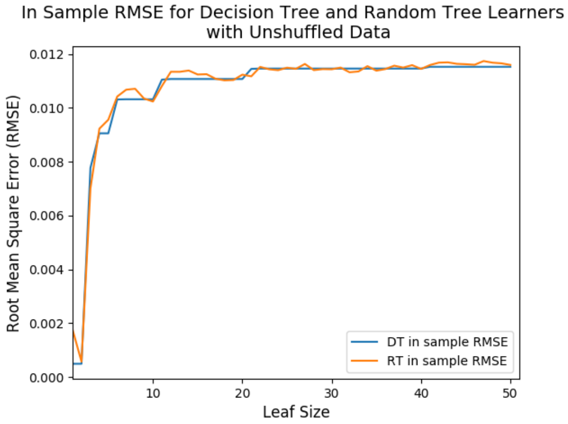


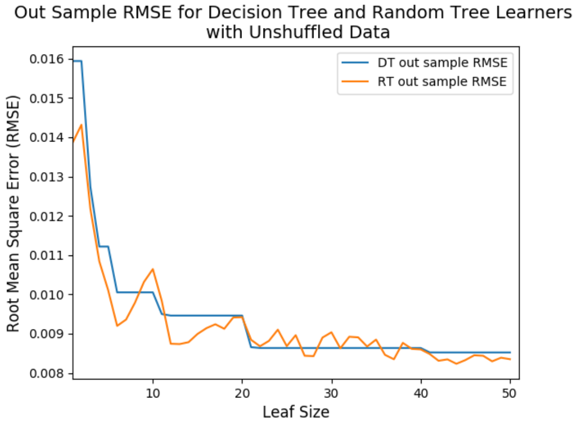


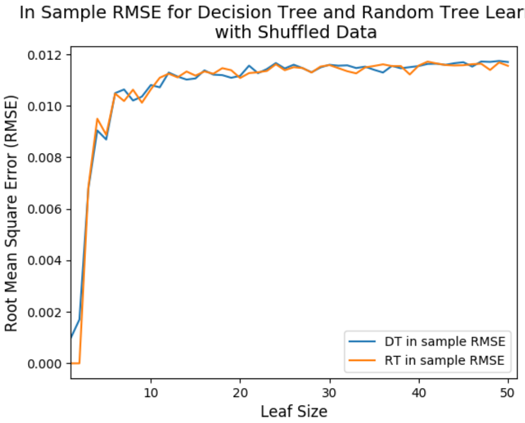
3. When comparing Decision Tree Learners with Random Tree Learners, it appears that when dealing with unshuffled input data, Decision Tree Learners perform better. Overall, the Decision Tree Learner has a lower and more consistent RMSE than the Random Tree Learner. This is especially true when comparing out-of-sample data. The Random Tree Learner may at times produce a lower RMSE, but it often produces a higher RMSE as well due to the greater variance introduced by the randomness. The classic Decision Tree, however, produced a smoother, more predictable RMSE and, when the leaf size is greater than 20, its RMSE appears to be generally lower than the Random Tree. A similar pattern holds for the in-sample RMSEs, where the RMSE of the Decision Tree is smoother and more consistent.

The performance of the Random Tree versus the Decision Tree differs when shuffled data is used. In this case, both methods have very similar RMSEs and both show a similar amount of variance in their RMSEs. This is true for both in-sample and out-of-sample data. Therefore, it is difficult to say that one is clearly better when the testing and training data is shuffled.

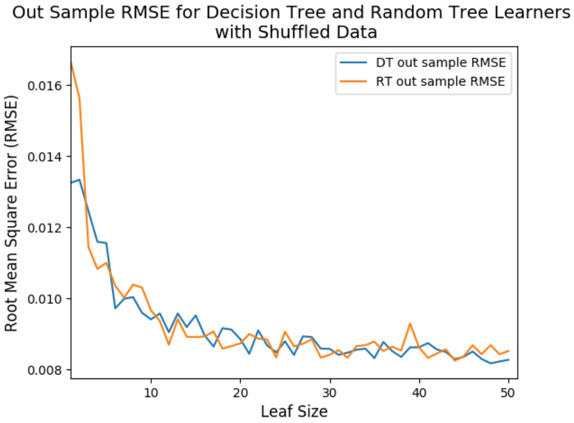
Overall, however, I would conclude that the classic Decision Tree is a better method to use, especially with unshuffled input data. The RMSE of the Decision Tree appears to be lower overall than the RMSE of the Random Tree as the leaf size exceeds 20. Also, the lower variation of the Decision Tree's RMSE makes it more consistent and likely a more reliable predictor.







In Sample RMSE for Decision Tree and Random Tree Learners
with Shuffled Data
Root Mean Square Error (RMSE)
0.000
0.002
0.004
0.006
0.008
0.010
0.012
10
20
30
40
50
Leaf Size
DT in sample RMSE
RT in sample RMSE



Out Sample RMSE for Decision Tree and Random Tree Learners
with Shuffled Data
Root Mean Square Error (RMSE)
0.008
0.010
0.012
0.014
0.016
10
20
30
40
50
Leaf Size
DT out sample RMSE
RT out sample RMSE