Fall 2021 – Project 3: Assess Learners

Michael Lukacsko
mlukacsko3@gatech.edu

Abstract—This assignment takes a deep dive into predictive modeling using a decision tree learner, a random tree learner, an ensemble learner/bag learner, and an insane learner. For each learner, the task was to train it, test it, and analyze the relationship between predicted and actual values of the dependent variable. This relationship is known as correlation. To accomplish this task, I used each learner's root mean square error (RMSE) with respect to an increasing leaf size.

INTRODUCTION

Predicative modeling in any form uses data to predict an outcome. In this assignment, as mentioned above, I employed 4 different predictive models to perform this task. First, I used a decision tree (DT) learner and random tree (RT) learner to create trees and query them for given values. I used 60% of the data to train each of the models and 40% to test. The DT learner and RT learner perform similarly, however, the RT learners split value uses the median value of a random feature. Next, I implemented a bag learner with a bag size of 20 bags – 10 DT learner bags and 10 RT learner bags. Each bag learner was trained and tested with a different subset of data. Finally, I implemented an insane learner that contained 20 bag learners where each bag learner employed 20 LinRegLearner instances. Again, I evaluated each learner's RMSE with respect to leaf size using in sample and out of sample errors to assess the correlation and review it for overfitting. For this report, I limited the learners to a leaf size of 1 to 75. I found that using a leaf size up to 75 was sufficient to assess the results. For the three experiments discussed below, it is my hypothesis that as RMSE increases correlation will decrease. Moreover, it is my belief that overfitting will occur where the in sample/out of sample output begins. This is because out of sample RMSE will be larger than in sample RSME at the start.

METHODS

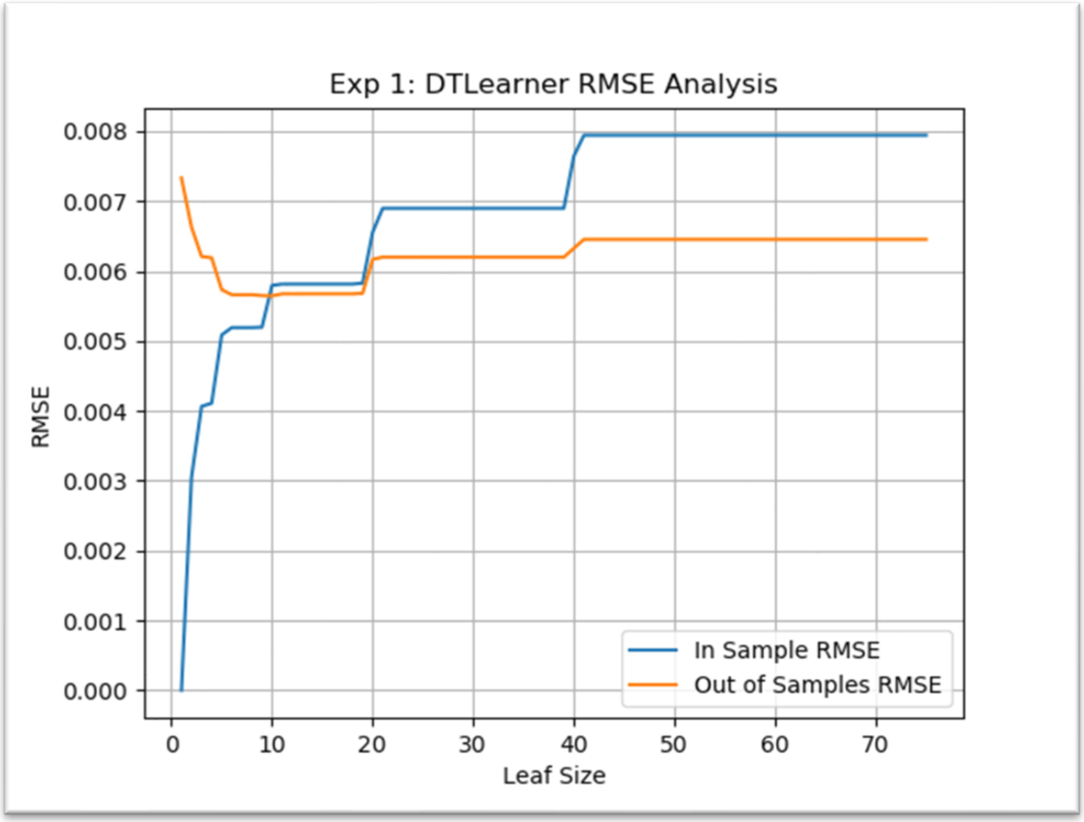
Each learner uses a subset of data to train and test before evaluating in sample and out of sample data for errors. All the testing and graphing take place in my testlearner.py file. For experiment 1, the DT learner is employed with a leaf size from 1 to 75. The first step in my implementation is to train the learner using the train_x data for in sample testing and train_y data for out of sample training and test it using a test_x and test_y subset of data. After training and testing the model, I calculate RSME and check correlation by returning the Pearson product-moment correlation coefficient. Based on correlation coefficient value returned, -1 to 1, I can measure the strength of the relationship between the testing and training data. Finally, I graph the output from in sample and out of sample testing using RMSE with respect to leaf size. Experiment 2 looks at effects of bagging versus overfitting. For this experiment I again used the DT learner, however, I implemented an ensemble learning model using bagging. The bag learner used here is made up of 10 DT learners. The process to train and test each learner is the same as described in experiment 1, however, the data is broken up such that a different subset is used to train and test each learner. Also similar is the in sample and out of sample evaluations, as well as the output being graphed using RMSE with respect to leaf size. The final experiment, experiment 3, looks at the bag learner RMSE when implementing a bag learner made up of 10 DT learners and 10 RT learners vs a single DT learner and RT learner. This metrics are calculated after creating each tree as described in experiment 1, however, I utilize the RT learner for this experiment and compare it accordingly with a standard RT learner

1.1 Experiment 1

Questions: Research and discuss overfitting as observed in the experiment. (Use the dataset Istanbul.csv with DTLearner). Support your assertion with graphs/charts. (Do not use bagging in Experiment 1). At a minimum, the following question(s) that must be answered in the discussion:

- Does overfitting occur with respect to leaf_size?
- For which values of leaf_size does overfitting occur? Indicate the starting point and the direction of overfitting. Support your answer in the discussion or analysis. Use RMSE as your metric for assessing overfitting.

Answer: To test whether overfitting occurs with respect to leaf size, I tested my DT learner using the Istanbul.csv data and a varying leaf size from 1 to 75 and output the data to the Exp 1: DTLearner RMSE Analysis figure below. What I notice from the graphed data is that, in the area where leaf size starts at 1 and increases to around 7, in sample RMSE is increasing while the out of sample RMSE is decreasing. Hence, I can confidently say that overfitting does occur with respect to leaf size where the leaf size is equal to 1 to about 7. This area of overfitting is expected because, as discussed by Prof. Balch, learners typically start with a lower in sample RMSE and a higher out of sample RMSE before generalizing.


Exp 1: DTLearner RMSE Analysis
RMSE
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
10
20
30
40
50
60
70
Leaf Size
In Sample RMSE
Out of Samples RMSE

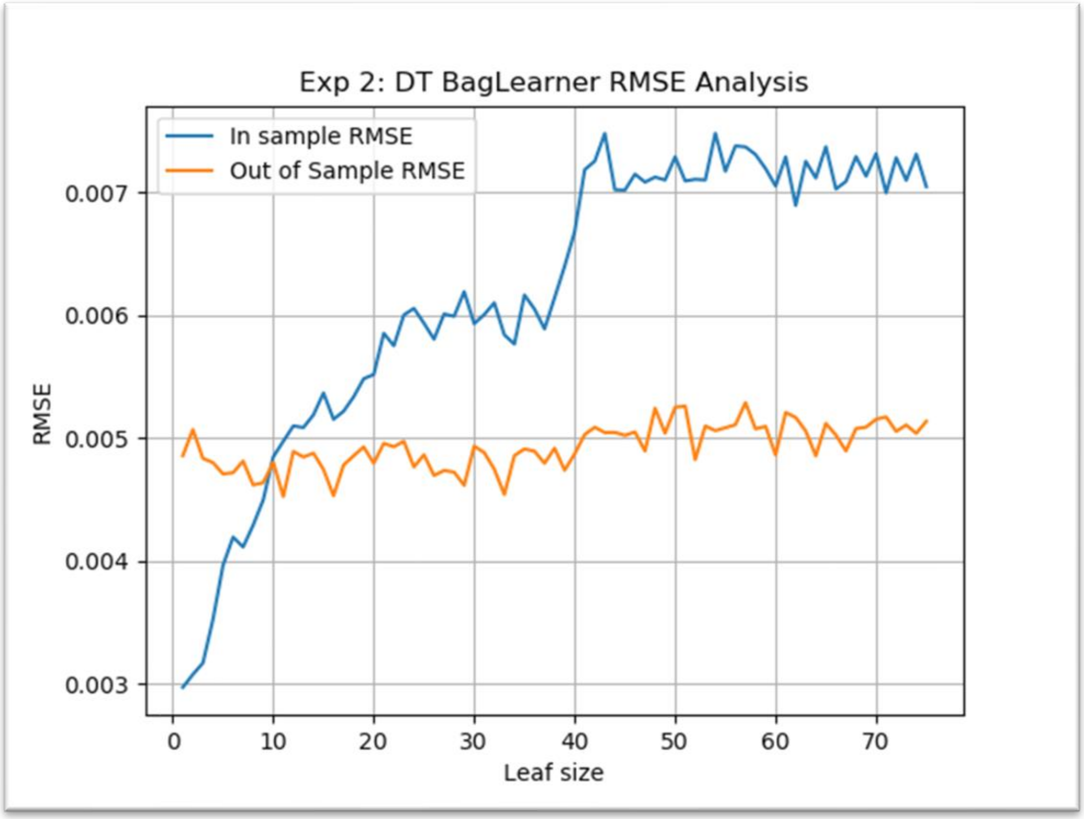
1.2 Experiment 2

Question: Research and discuss the use of bagging and its effect on overfitting. (Again, use the dataset Istanbul.csv with DTLearner.) Provide charts to validate your conclusions. Use RMSE as your metric. At a minimum, the following questions(s) must be answered in the discussion.

- Can bagging reduce overfitting with respect to leaf_size?
- Can bagging eliminate overfitting with respect to leaf_size?

To investigate these questions, choose a fixed number of bags to use and vary leaf_size to evaluate. If there is overfitting, indicate the starting point and the direction of overfitting. Support your answer in the discussion or analysis.

Answer: When overfitting occurs, one possible method to minimize or avoid its presence is to use an ensemble learner or bag learner. A bag learner is made up of multiple bags, each bad containing a predictive model. For this experiment, I implemented a bag learner that used 10 bags, each consisting of a DT learner. Like experiment 1, RMSE was graphed with respect to leaf sizes ranging from 1 to 75. Almost immediately I noticed that the out of sample RMSE was lower, less than 0.005, compared to testing a DT learner on its own in experiment 1 where it was greater than 0.007. Additionally, in the area where overfitting was observed in experiment 1, between leaf size 1 and 7, it no longer looks to me like the out of sample RMSE is decreasing. Rather, the out of sample RMSE for this bag learner looks linear in this range. Hence, I would say that by utilizing bagging, this experiment has effectively eliminated overfitting with respect to leaf size.


Exp 2: DT BagLearner RMSE Analysis
In sample RMSE
Out of Sample RMSE
RMSE
0.003
0.004
0.005
0.006
0.007
Leaf size
0
10
20
30
40
50
60
70

1.3 Experiment 3

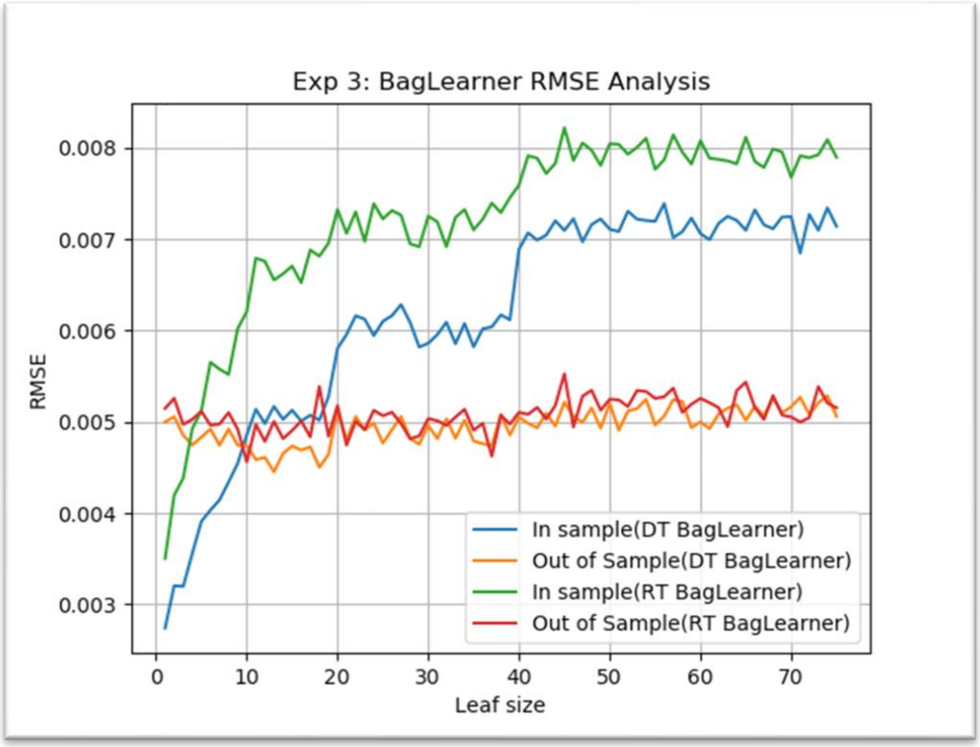
Experiment 3

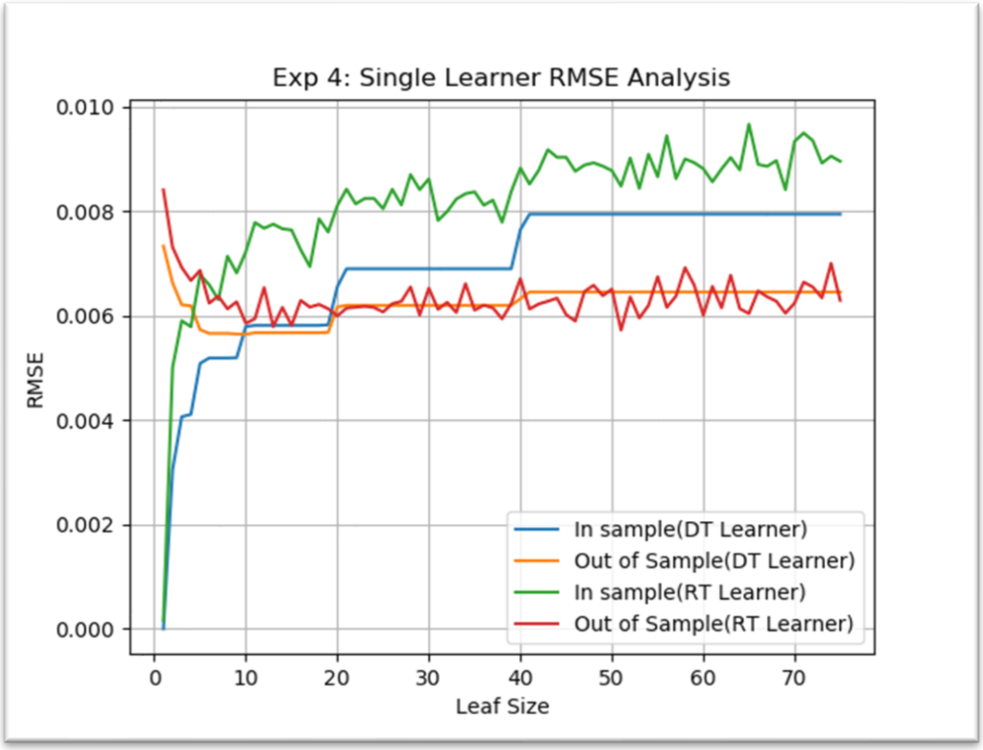
Question: Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). For this part of the report, you must conduct new experiments; do not use the results of Experiment 1. Importantly, RMSE, MSE, correlation, and time to query are not allowed as metrics for this experiment.

- Provide at least two new quantitative measures in the comparison.
- Using two similar measures that illustrate the same broader metric does not count as two separate measures. (Note: Do not use two measures for the accuracy or use the same measurement for two different attributes – e.g., time to train and time to query are both considered a use of the "time" metric.)
- Provide charts to support your conclusions.
- At a minimum, the following question(s) must be answered in the discussion.

- In which ways is one method better than the other?
- Which learner had better performance (based on your selected measures) and why do you think that was the case?
- Is one learner likely to always be superior to another (why or why not)?

Answer: As previously mentioned, the RT learner and DT learner is built using a split value that is determined either a randomly in the case of a random tree or by finding the feature with the highest correlation if using a decision tree. Because of this, it is often the case that a random tree might not use the best feature to split on. As a result, it can be true that the RMSE of a random tree will have a higher RMSE as opposed to the decision trees. To analyze the in sample and out of sample RMSE metrics, I employed a bag learner containing 20 bag – 10 DT learner and 10 RT learners – for my first test and a single DT learner, and RT learner for the second test. As can be seen in the Exp 3: BagLearner RMSE analysis plot that follow, out of sample RMSE for each learner runs almost parallel. However, when analyzing the in sample RSME, the in sample RMSE for the RT learner is higher and remains that way with a leaf size tested from 1 to 75. Again, this is expected behavior because of the way a RT learner decides the feature to split on. In the Exp 4: Single Learner RMSE Analysis plot, the observation is even more pronounced. The RT learner out of sample RMSE has significant variations for all leaf sizes tested between 1 and 75 while the DT learners RMSE is much flatter.


Exp 3: BagLearner RMSE Analysis
RMSE
0.008
0.007
0.006
0.005
0.004
0.003
0
10
20
30
40
50
60
70
Leaf size
In sample(DT BagLearner)
Out of Sample(DT BagLearner)
In sample(RT BagLearner)
Out of Sample(RT BagLearner)


Exp 4: Single Learner RMSE Analysis
RMSE
0.010
0.008
0.006
0.004
0.002
0.000
0
10
20
30
40
50
60
70
Leaf Size
In sample(DT Learner)
Out of Sample(DT Learner)
In sample(RT Learner)
Out of Sample(RT Learner)