Project 3: Assess Learners

Eddie Figone

Abstract—This project explores the performance of decision tree, random tree, and bagging models in the context of overfitting as measured through RMSE, MAE and R^2. Through experimentation, the decision tree is shown to have issues with overfitting, but less so than random trees. Overfitting is also shown to be mitigated through bagging.

1 INTRODUCTION:

This paper seeks to explore the performance of decision tree, random tree, and bagging models in the context of overfitting. Overfitting is the phenomenon that as the model better fits the training data, the model. Three experiments were conducted to explore this. In experiment 1, a decision tree was built with varying lead size to explore the effects of overfitting. One could expect that the smaller the leafs, the closer the model will fit the training data, and lead to more overfitting. Experiment 2 examines overfitting in a bagging model, and if it can be reduced or eliminated. It is likely that bagging would have such an effect as it averages values out across the bags, which will help avoid matching the training data too closely. Finally, in experiment 3, a decision tree is compared to a random tree in regard to R^2 and Mean Absolute Error to see if one is superior. It would stand to reason that a decision tree will fit the data closer, though it remains to be seen if that is always better.

2 Methods:

A rudimentary decision tree model was built in python using recursive logic. For the data being trained on, the model will find the best feature to split on (the one with the highest correlation to the response) and then take the median of the values, with values equal and less to the median feeding the left branch and the rest going to the right. The model will continue to split until the data being split reaches a predetermined leaf size. The random tree model is identical but picks a random feature to split on rather than one with the highest correlation. The bagging model takes any other model as input and builds a number of versions (or bags) using different allotments of training data (taken with resampling). These models were then trained on the *Istanbul.csv* file, and then evaluated both

the training and testing data with varying parameters. These experimental results were vectorized and graphed in the experiment sections below.

3 Discussion:

3.1 Experiment 1:

This experiment seeks to answer the question of if overfitting occurs in regards to leaf size, and where and how overfitting presents itself. Overfitting is when the model performs better on in sample (training) data at the expense of performance on out of sample (testing or external) data. For this experiment, a decision tree was trained on the *Istanbul.csv* and its performance on the training and testing data was recorded via RMSE on leaf sizes ranging from 1 to 20. The results are charted below:

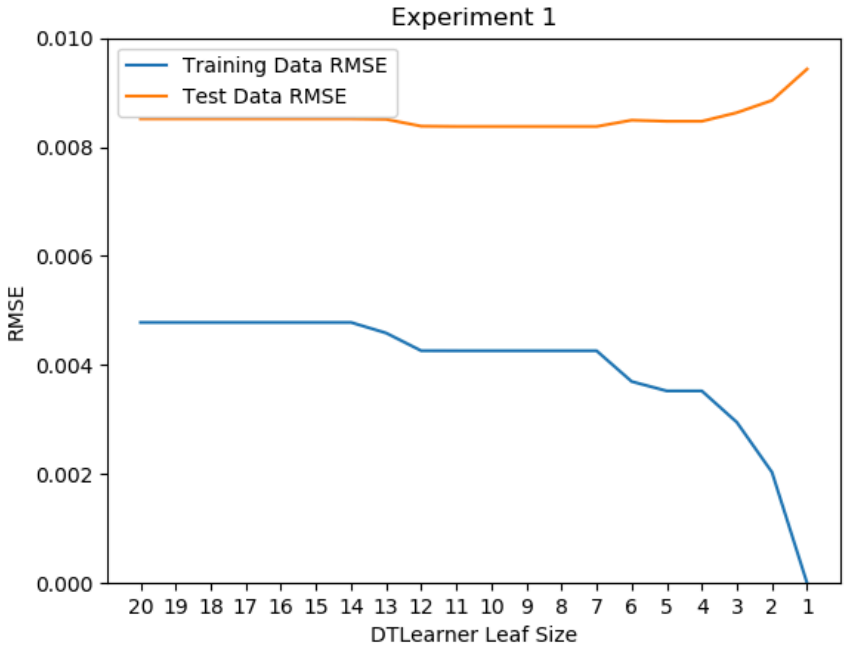

Figure 1–The chart above tracks the RMSE performance of the decision tree model trained on the Istnabul.csv on in sample and out of sample data, with respect to minimum leaf size.

As we see in the chart above and in the data, the model's performance on the in sample data continues to improve with smaller leaves, however out of sample performance begins to worsen when leaf size goes below 5. This is a textbook example for overfitting, in other words, this model begins to overfit with leaf size 4 or less. We see that the overfitting continues to worsen as the leaf sizes decrease below for. This trend in performance is what one would expect, as the model begins to more fully fit the data it is trained on, it will be too fit to the data and not handle new data well. Overfitting can be mitigated by observing in sample and out of sample performance, and ceasing to improve in sample fit when a decrease in out of sample performance is observed.

3.2 Experiment 2:

This experiment seeks to determine if bagging can reduce overfitting in regards to leaf size, and to explore if bagging can eliminate it entirely. The experimental set up is almost identical to experiment 1, with the sole difference being the use of a bagging model. The bagging model uses a decision tree like the previous experiment but will create a number of bags and average the result. Bag sizes from 1 to 20 were tried and their performance in RMSE were examined to find an optimal bag size.

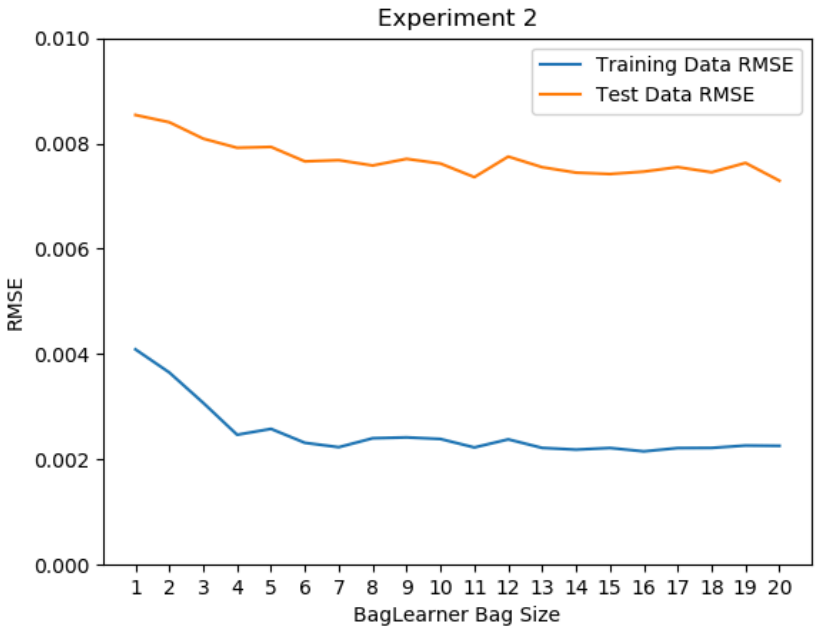

Figure 3–A bag learner was trained on the Instanbul.csv and its RMSE for bag sizes 1-20 was recorded on in and out of sample data to find the optimal value for this experiment.

A bag size of 11 was selected as that was where out of sample and in sample performance both seemed to flatline. Larger bags beyond 11 did not seem to impact performance.

Using this trained model, the experimental procedure of experiment 1 was repeated for experiment 2 examining overfitting (measured by in/out of sample RMSE) in regards to leaf size. The results are charted below and compared to figure 1:

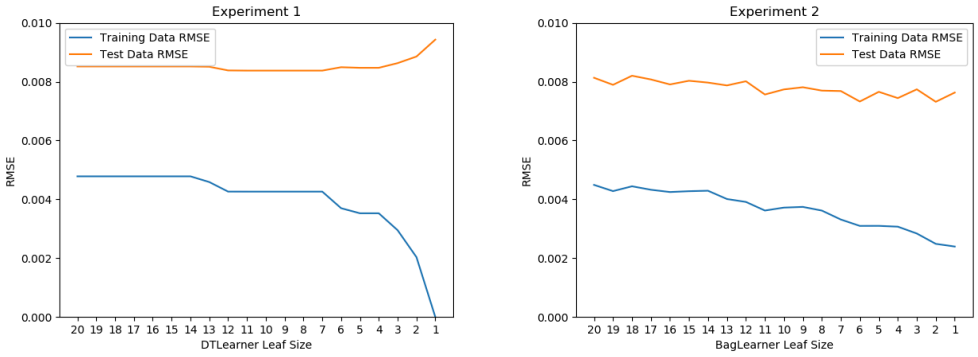

Figure 3–the results of experiment 1 (left) are compared to experiment 2 (right) using the same scale to demonstrate the reduced overfitting in experiment 2.

The chart on the right represents the bagging model. The bagging model displays a lack of overfitting which can be seen by the out of sample performance remaining effectively static as leaf size decreases and in sample performance increases. Even as the model fits the training data closer and closer, it does not trade performance on out of sample data unlike the previous decision tree model. It would seem that overfitting is eliminated from this model as even having one data point per leaf, the absolute minimum possible, does not cause a decrease in out of sample performance.

3.3 Experiment 3:

Experiment 3 seeks to compare the performance of a decision tree to a random tree in regards to mean absolute error and R^2. The format of experiment 1 is

repeated again to create a decision tree and random tree model and evaluate the performance with regard to mean absolute error and R^2, as seen below:

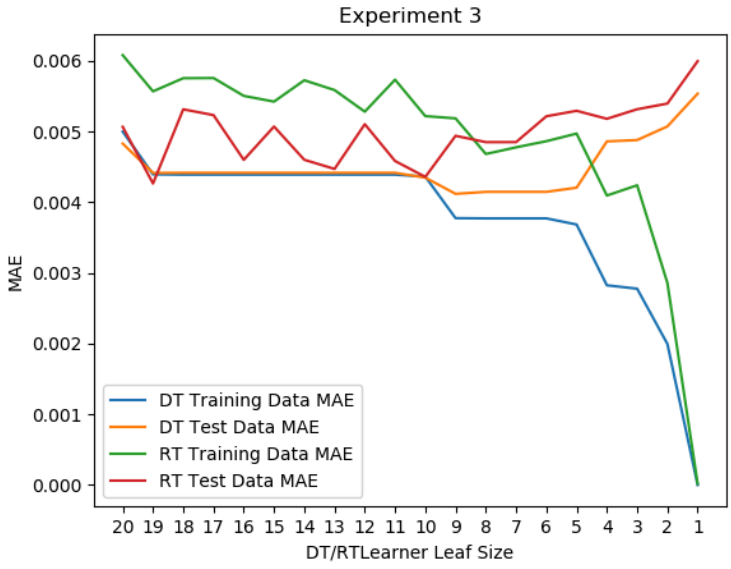

Figure 4–A decision tree and random tree model are trained on Istanbul.csv and have their in and out of sample performance in regards to mean absolute error compared.

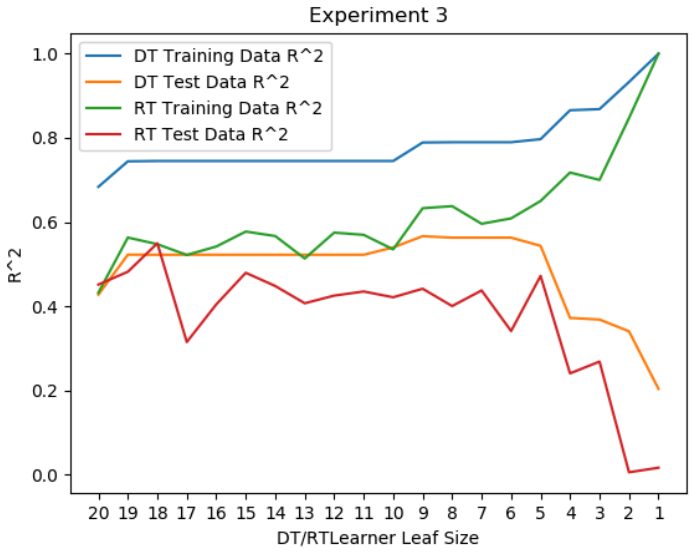

Figure 5–A decision tree and random tree model are trained on Istanbul.csv and have their in and out of sample performance in regards to R^2 compared.

Both models show overfitting, as the leaf size decreases, the in sample performance improves and the out of sample worsens for both metrics. However, in both metrics overfitting happens earlier for the random tree. It seems the decision tree is more resistant to overfitting. The decision tree also has a lower overall mean absolute error than the random tree, however, the random tree model outperforms the decision tree in regards to overall R^2. It is difficult to all around define one model as better, but the decision tree handles overfitting better which is one of the metrics we are most interested in. There may however be scenarios where we are not concerned with overfitting and the random tree may perform better (such as in regards to R^2).

4 Summary:

The initial hypothesis for experiment 1 was that smaller leaf size leads to more overfitting, and the most extreme overfitting would occur at leaf size 1, which was verified through the models trained. For experiment 2, it was theorized that it is likely that bagging would reduce overfitting, though potentially not remove it entirely. The results of experiment 2 show that bagging does seem to remove, or at least seriously decrease overfitting, which makes sense as numerous models will reduce the fit to the training data. It should be noted that this experiment was conducted with 11 bags, it would be interesting to further explore results with significantly less and more bags and compare the results. Finally, in experiment 3 it was hypothesized that a decision tree will fit the data closer than a random tree, but not likely be better in all ways. This hypothesis was supported by the findings. The decision tree handled overfitting better, but model performance was somewhat split with the random tree performing better in regards to R^2. It is unlikely for there to be a scenario where one model is always better, for example there may be some applications where overfitting is an important measure to keep track and others where it has little bearing.