Name: Ajay Joshi

Ajoshi319@gatech.edu

MC2-Project 1

Overview

In this project, I've implemented Random Tree Learner and Bag Learner Class in python. This report covers RMS error and correlation comparison, detection of overfitting on various models of ensemble learning model. I have used Istanbul.csv data source for this report.

1. Detection of overfitting as Leaf size increases

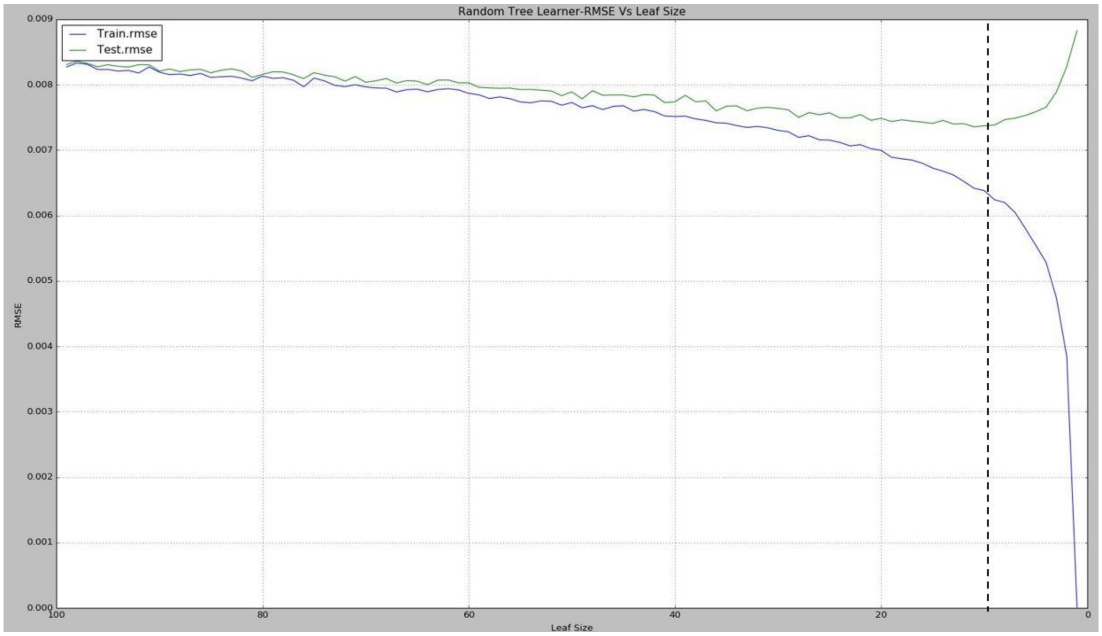

Figure 1: Random Tree Learner – RMSE Vs Leaf Size (1 to 100)

The above graph displays the movement of RMS errors of Train and Test data as a leaf size on Random Tree learner increases from 0 to 100. As we can see that, when a Leaf size is roughly around from 9 to 11, the graph starts diverging. It is because training (In sample) RMS error is decreasing and Testing (Out of sample) RMS error is increasing. Hence, occurred an overfitting.

The following data was taken from model that I generated. The table shows that as leaf size decrease, Train RMSE is decreasing and Test RMSE is increasing till the leaf size is 9.

Leaf	Train.rmse	Train.corr	Test.rmse	Test.corr
20	0.006998	0.743505	0.007487	0.703793
19	0.006892	0.752494	0.007439	0.709232
18	0.006868	0.754529	0.007464	0.707131
17	0.006849	0.755968	0.007445	0.709641

16	0.006800	0.760204	0.007428	0.710739
15	0.006727	0.766073	0.007408	0.712646
14	0.006678	0.769776	0.007456	0.708929
13	0.006621	0.774467	0.007400	0.713692
12	0.006524	0.781846	0.007407	0.713291
11	0.006419	0.789682	0.007357	0.718289
10	0.006381	0.792653	0.007373	0.717119
9	0.006242	0.802796	0.007385	0.717236
8	0.006198	0.805878	0.007467	0.710255
7	0.006044	0.816302	0.007490	0.709921
6	0.005798	0.832486	0.007532	0.708413

2. Detection of overfitting - fixed Leaf size =15 and changes in bags

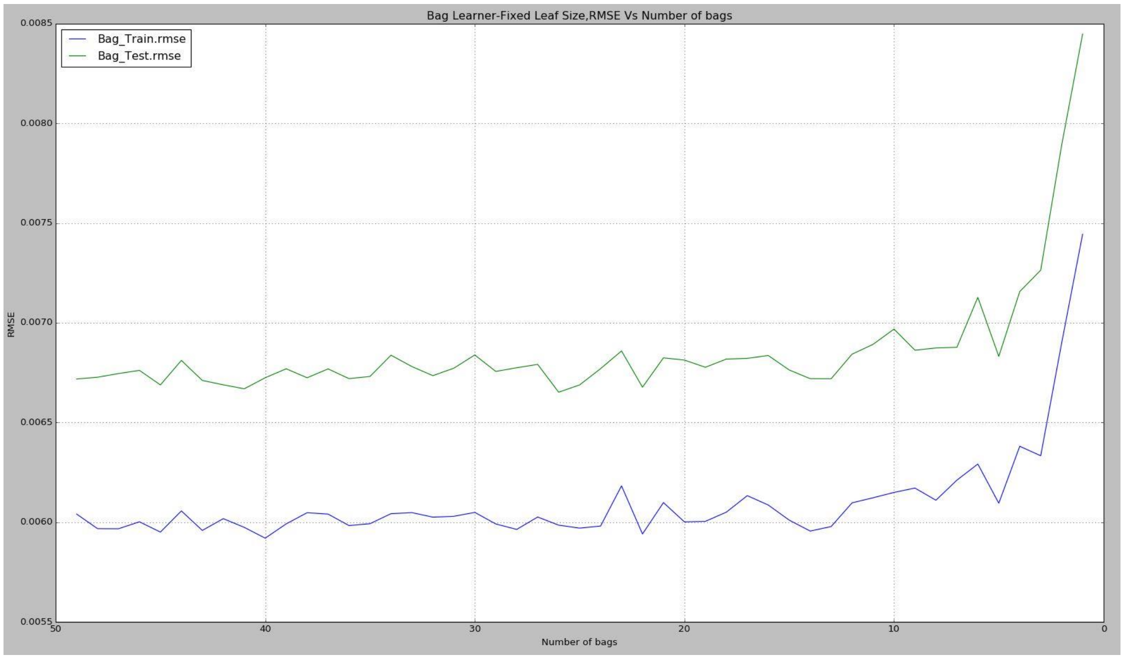

Figure 2a- Bag learner Leaf Size- RMSE Vs Bags

By looking at the graph, we can see that as number of bag increases, RMSE decreases on both training and test data. As the number of bags reaches to 0, there will high chance of overfitting.

By looking at the table below, we know that Linear regression learner performed best and followed by Bag learner, whereas single random tree performed worst with high RMS error. Bagging helped to lower variance result in less probability to overfitting.

Benchmark

Linear Regression Learner	Random Tree Learner (leaf Size =1)	Bag Learner
In sample results	In sample results	In sample results
RMSE: 0.00528577751921	RMSE: 0.0	RMSE: 0.00689895405376
corr: 0.893488459362	corr: 1.0	corr: 0.839222128139

```
Out of sample results          Out of sample results            Out of sample results
RMSE:  0.00403140544617  RMSE:  0.00845576262521       RMSE:  0.00476361177267
corr:  0.889492715519      corr:  0.492479357602          corr:  0.815517822536
```

Comparison of RMSE between bag learner (leaf = 15, bag = 50) and Random Tree Leaner (Leaf =15) in a 50 iteration.

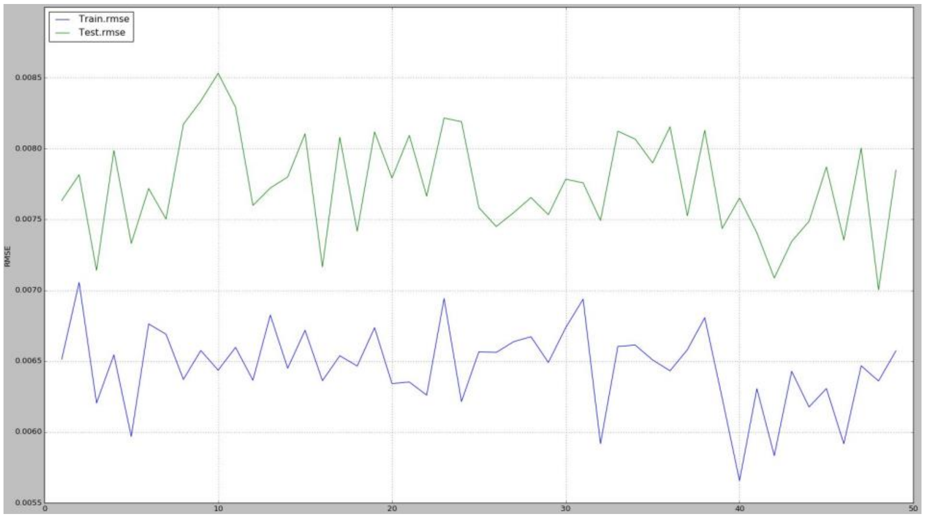

Figure 2b- Random Tree Leaner (Leaf =15) in a 50 iteration.

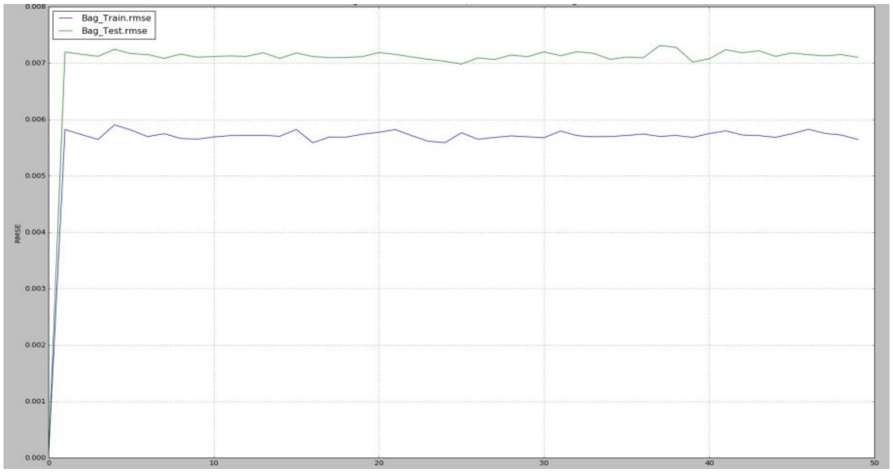

Figure 2c- bag learner (leaf = 15, bag = 50) in a 50 iteration

By looking at the figure 2b and 2c, we can see that single random tree tends to have higher test's RMS error than that of Bag learner.

3. Detection of overfitting - fixed Bag size and changes in Leaf Size

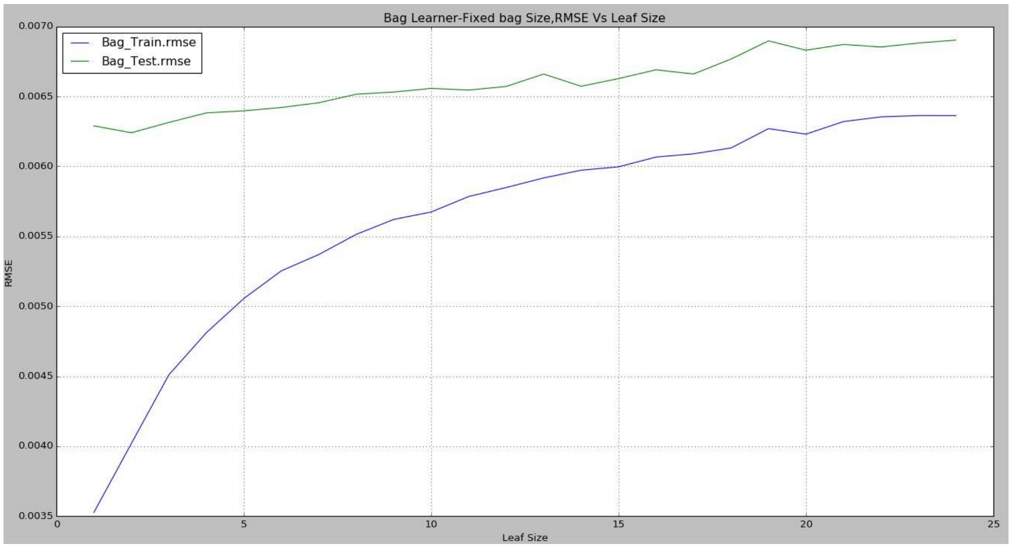

Figure 3: Bag Learner (Bag= 15), RMSE vs Leaf Size

In the graph, above, I've set the bagging set to 15 and iteration from a leaf size 1 to 50. Per my observation, as the leaf size increases, in sample and out of sample RMSE starts increasing. However, when a leaf size is around 1, there's slight decline of out of sample RMSE and increase of in sample RMSE which can be overfitting behavior. By looking at the RMSE trend, we can tell that there's high chance of overfitting as a leaf size decreases.