Report of assess learners

1. Does overfitting occur with respect to leaf_size?

Yes. Overfitting occurs when the in-sample error continues decreasing but the out-of-ample error begins to increase with the degrees of freedom. Shown in the following plot, the x-axis is the leaf size, y-axis is tHe RMSE, the curves represent the RMSE of in-sample data (blue color) and out-of-sample data (orange color) calculated by the DTLearner. As the leaf_size decreases (i.e. the degrees of freedom increases), the out-of-sample error and in-sample error become lower together at first but the out-of-sample error begins to increase while the in-sample error continue decreasing once the leaf size is reduced to around the value of 6, where the overfitting occurs.

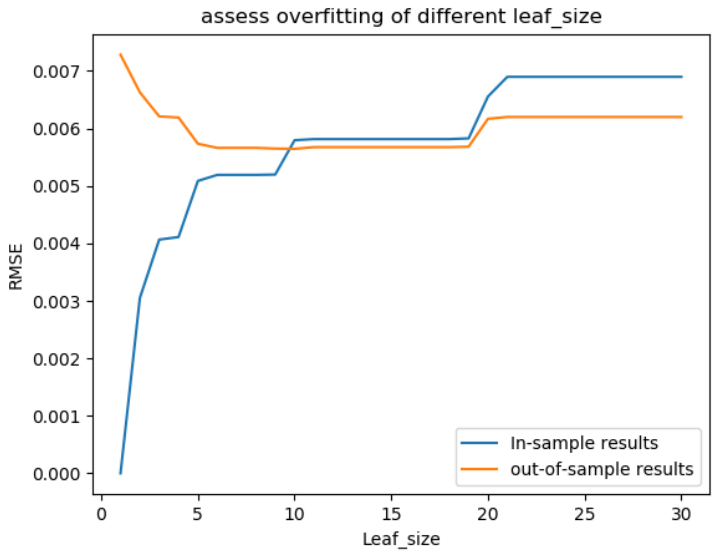

Fig.1. The assessment of overfitting of DTLearner with different leaf size

2. Can bagging reduce or eliminate overfitting with respect to leaf_size?

Yes. The bagging can reduce but cannot eliminate the overfitting. As shown in the Fig.2, we employ the BagLearner and DTLearner to calculate the RMSE with respect to leaf_size under the different values of bags. Compared with the Fig.1, the leaf size where the overfitting begins to occur become lower. In Fig.1, overfitting occurs when the leaf size is around 6. In Fig. 2, overfitting occurs when the leaf size is less than 4. It is also obvious that the curves become more smooth or stable as the bags increase. But if we continue increasing the bag number (e.g. 100), the results show that the overfitting becomes less obvious but still exists.

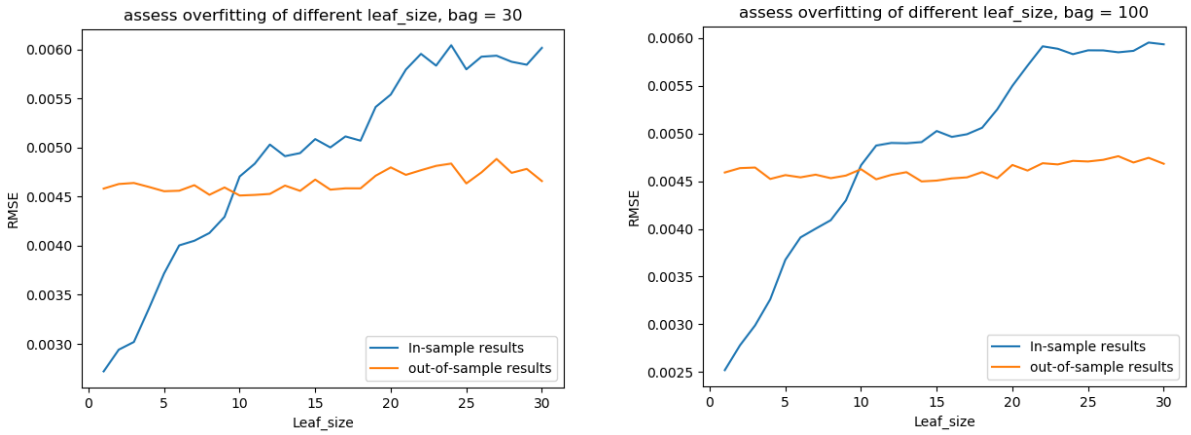

Fig.2. The assessment of overfitting of DTLearner with different bags

3. Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner).

We use the new data soure of 3_groups.csv to compare the performance. The RMSE and the correlation coefficients between the Y and predicted Y of in-sample and out-of-sample data are employed to compare the decision tree and random tree methods. The curves of the RMSE of out-of-sample and in-sample data between DT and RT are shown in the Fig. 3.a. The RMSE of the DT is less than that of the RT, no matter the in-sample or out-of-sample data. The curves of the DT are also smoother than those of the RT, which means the DT could give more stable and accurate results.

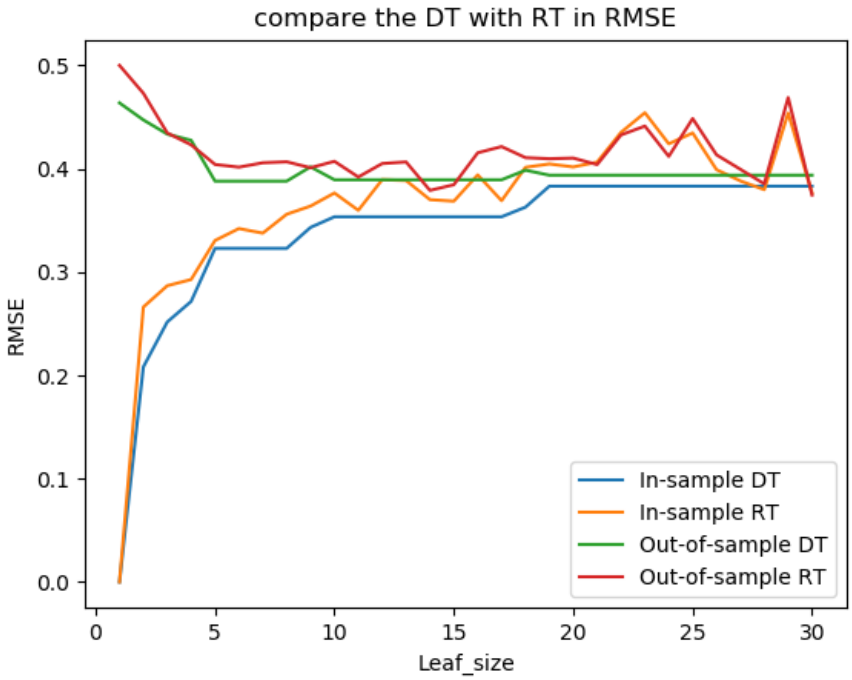

Fig.3.a. The comparison of DTLearner and RTLearner in RMSE

The curves of the correlation coefficients are shown in the Fig. 3.b. The correlation coefficient values of the DT is higher than those of the RT, no matter the in-sample or out-of-sample data, which means the

results of DT are more relevant with the actual data. Similarly, the curves of the DT are also much smoother than those of the RT, which means the DT performs more stable.

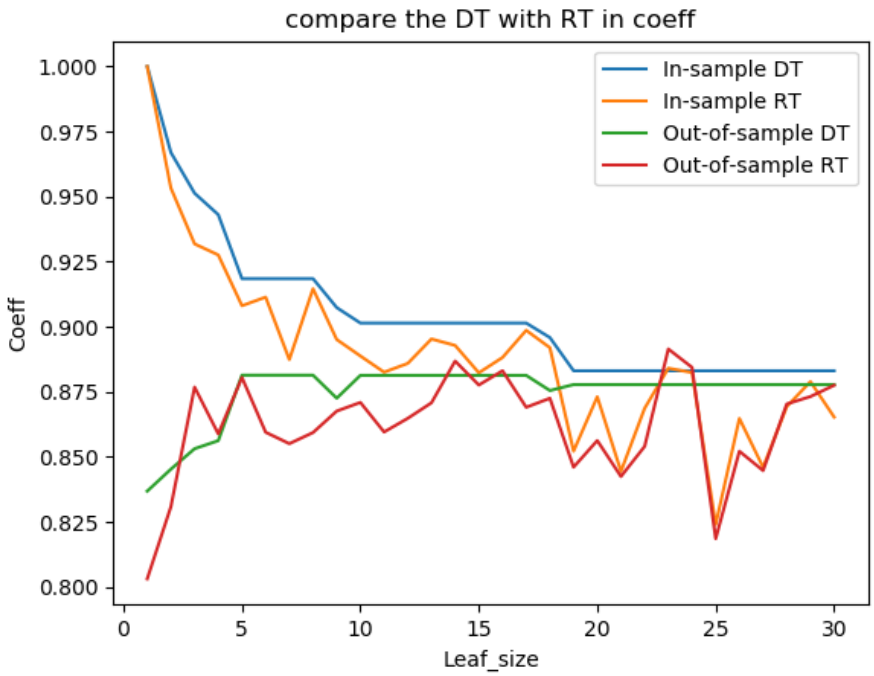

Fig.3.b. The comparison of DTLearner and RTLearner in coeff

The running time of RT is shorter than that of the DT but the difference is so slight that can be ignored. In summary, the DT performs much better than the RT, predicting more accurate and stable results.