Assignment 3: Assess Learners

XI LE
xi6@gatech.edu

Abstract—Using supervised machine learning algorithms, we create four learner models including DT Learner, RT Learner, Bag Learner and Insane Learner. Through a few experiments, train and query machine learning models and assess the performance and overfitting.

1 INTRODUCTION

This report briefly describe a few supervised machine learning algorithms and explain how to implement each of models through the predefined hyperparameter such as bag size and leaf size. All experiments try to compare the performance of trained models based on different quantitative metrics and analyze the in-sample and out-sample overfitting.

2 METHODS

Based on the Classification and Regression Trees (CARTs) theory, four kind of different machine learning models (DT Learner, RT Learner, Bag Learner and Insane Learner) get implemented and evaluated. DT Learner was designed as a single Decision Tree and predefined the best feature to split the value is the highest absolute correlation with the actual result value. RT Learner is very similar with DT Learner, the only difference is the best feature to split the value is random chosen. Bag Learner is kind of Bootstrap Aggregation Learner, which can apply any simple learner in this experiments using a various number of bags.Insane Learner is a more complicated aggregation Learner, since it contains a specific number of Bag Learner which each composed some simple learner. [1]

3 DISCUSSION

Istanbul Emerging Markets index is used as the basic database when these experiments are conducted. although this data is kind of time series data, we do not consider time series factor and data-time columns is removed. Therefore, the training data set is chosen randomly as well as the testing data. (Balch. and Romero., 2022)

[1] refer to https://gatech.instructure.com/courses/372882/assignments/1610466

3.1 Experiment1:DT Learner overfitting

Experiment1: overfitting happens in DT Learner's prediction when the leaf-size is 5. The direction of overfitting is more obvious when the leaf size become smaller. The reason is the model becomes more precise as the leaf size decreases to 1. As the Figure 1 shows when the leaf size is small less than 5, the model performs very well on the training data, but much more poorly on unseen testing data, regarding RMSE, the main metric. In the leaf size range from (5,1) overfitting happens. However, when the leaf size is beyond 9, no overfitting but the performance of the predicted model becomes stable.

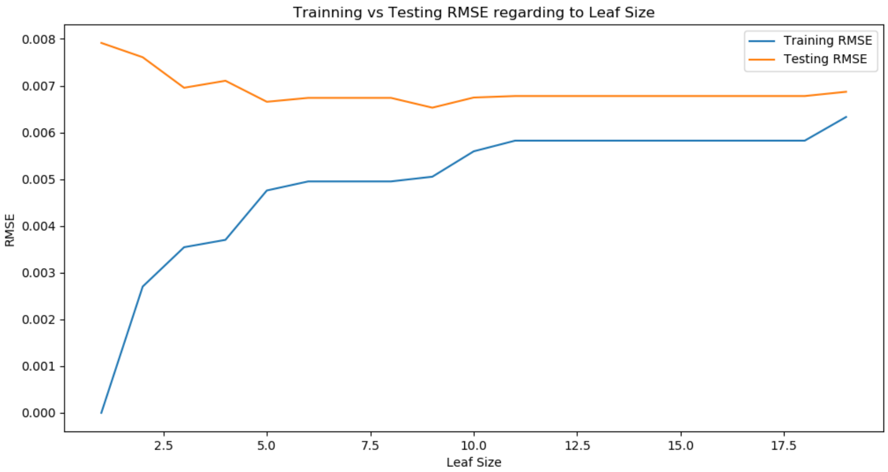

Figure 1—DT Learner Trainning vs Testing RMSE regarding to Leaf Size

Overfitting is important and the reasons are that it compromises the ability of models to generalize unseen data and provide a balance between the performance of the in and out sample data prediction.Although overfitting is quite important, mitigation of overfitting could help to train more precise models. There are a few methods to mitigate overfitting: Dataset should be reasonable split into separate training ,validation and testing sets. Especially when the dataset are auto correlatedsuch as time series data.

Choosing the simpler models with fewer parameters or reduce the complexity of the model by pruning in decision tree; Reduce the number of features in the dataset, such as manually chose reasonable factors; Increasing the training data, more training data could help the model generalize a better model and reduce the overfitting.

3.2 Experiment2:DT Learner VS Bag Learner

Bagging reduces overfitting by averaging out the predictions from multiple DT models, which reduce the variance of the predictions.Each simple model in a bagging ensemble is trained on a slightly different subset of the training data. This diversity among the models helps in reducing the overall overfitting. In this experiment, the bag number is fix to 20, As the Figure 1 and Figure2 states, the Bag Learner's performance is better than simple DT Learner,based on the lower RMSE value. In addition, overfitting is gone especially at previous point which leaf size is 5. But when the leaf size hits 6, the overfitting is still happened very slightly and then gone as the leaf size changes. Therefore, the Bag Learner performs great and overfitting reduces significantly but do not disappear fully. In sum, with the change of leaf-size, the overfitting does decrease but not eliminate. (Cao., 2023)

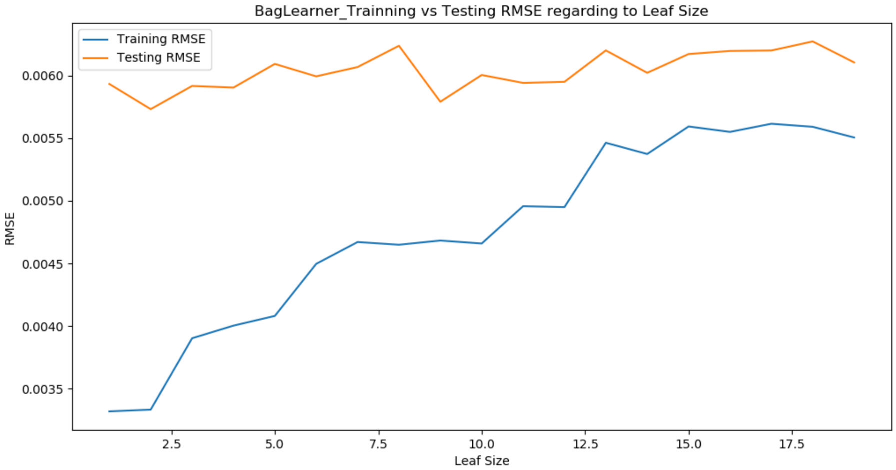

***Figure* 2**—Bag Learner Trainning vs Testing RMSE regarding to Leaf Size

3.3 Experiment 3:which model is better

In experiment3, without using RMSE in last two experiments, R-Square and Mean Absolute Error are used as two metrics to measure the DT and RT learner's performance. As Figure3 and Figure4 show: The DT Learner seems to have lower MAE on the training data, which indicates it may be fitting the training data better than

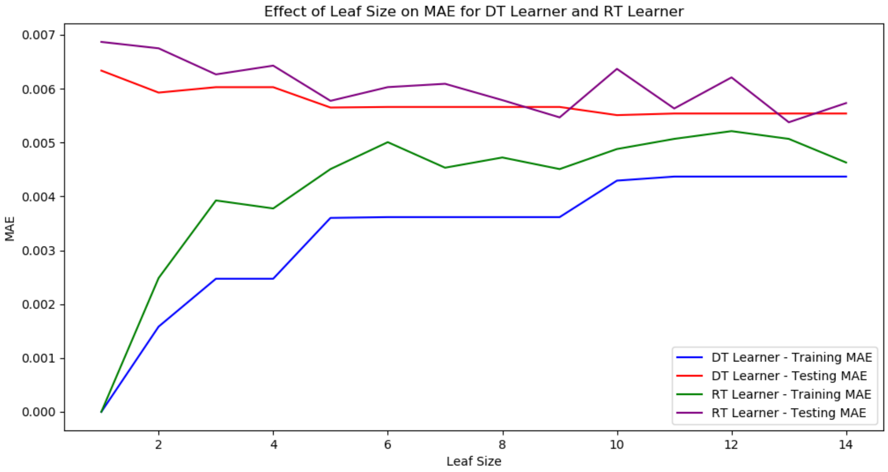

Figure 3—Effect of Leaf Size on MAE for DT Learner and RT Learner

the RT Learner. However, this does not fully translate to better prediction as the MAE for the DT Learner on the test data is very close to that of the RT Learner. In terms of R^2, the DT Learner shows a higher score both on the training data and the test data, but still show the overfitting in both RT and DT learners.

Therefore, The DT Learner seems to have better performance based on the selected measures of MAE and R^2, particularly in this context of training data. The likely reason for this could be the specific selection of features in the DT Learner.

In fact, it is hard to say which learner is always be superior to others across all possible scenarios. The DT Learner might perform better in some scenarios where the training data is representative of the entire data distribution or the dataset has high auto correlation. On the other hand, the RT Learner may be superior in scenarios where the data contains a lot of noise, or it is prone to overfitting.The choice between a DT Learner and an RT Learner should be based on the specific characteristics of the data, the complexity of the problem, and the goals of the modeling effort. (Cao., 2023)

4 SUMMARY

Through implement of these four machine learning algorithms and a few experiments,we can see each learner's behavior and performance under different predefined assumption and parameters. Overfitting usually happens,but can be

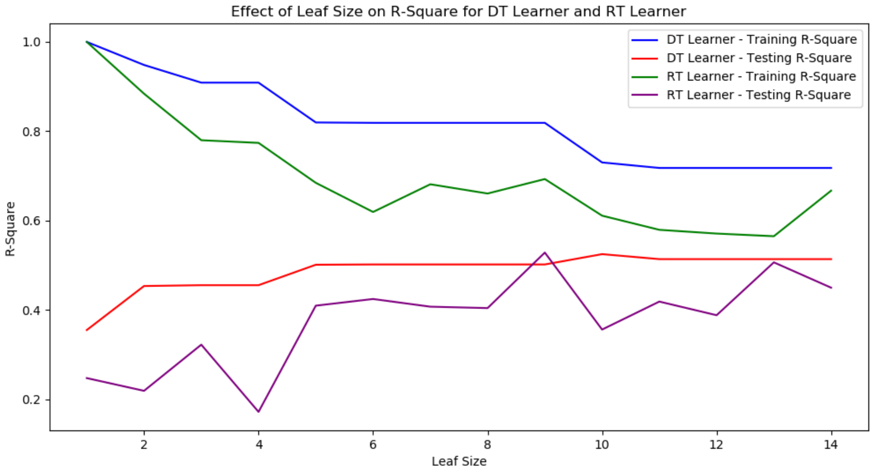

Figure 4—Effect of Leaf Size on R-Square for DT Learner and RT Learner

significantly reduced by bagging or boosting in which are time-consuming and expensive way, however not fully removed. In addition,it is not easy to tell which one learner performs better than others since the different dataset and different measurement scope.

5 REFERENCES

1. Balch., Tucker and Romero., Philip J. (2022). “What Hedge Funds Really Do”. In: *Market-Making Mechanics*. Business Expert Press.
2. Cao., Larry (2023). “Handbook of Artificial Intelligence and Big Data Applications in Investments Paperback”. In: *Trading with Machine Learning*. CFA Institute Research Foundation.