CS 7646: MC3 –P1

[Arti Chauhan: Feb-18-2017]

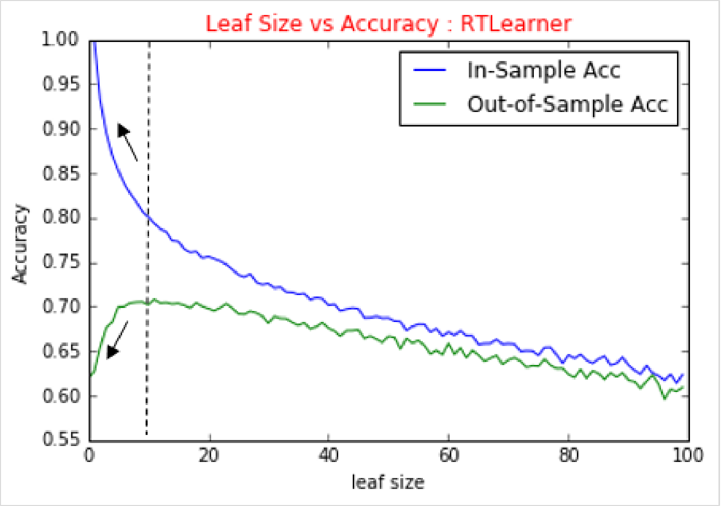
Q1.Does overfitting occur with respect to leaf_size? Consider the dataset istanbul.csv with RTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts.

- Dataset : istanbul.csv
- Learner : Random Tree
- Objective : show RMSE as function of leaf size for Train (In-sample) and Test set(out-of-sample)
- Methodology :
  a) For leaf size in range (1,100), compute in-sample and out-of-sample RMSE and accuracy.
  b) Shuffle data for each iteration of leaf size in step-a.
  c) Repeat step-a N times and compute average RMSE and Accuracy for each leaf size. N used for following chart is 100. This was done to smoothen the lines and increase the confidence in results.

Answer: Yes overfitting occurs with respect to leaf_size. Overfitting occurs when models gives good performance on training data but generalizes poorly to other data.

- For leaf size in range 100 to 10, RMSE for both in-sample (blue) and out-of-sample (green) error continue to improve but at around leaf_size=10 divergence between two trends occurs (Fig-1a).
- For leaf_size <= 10, in-sample error decreases significantly but out-of-sample RMSE starts to increase, exhibiting signatures of overfitting.
- Accuracy shows similar trend as RMSE.

Accuracy



RMSE

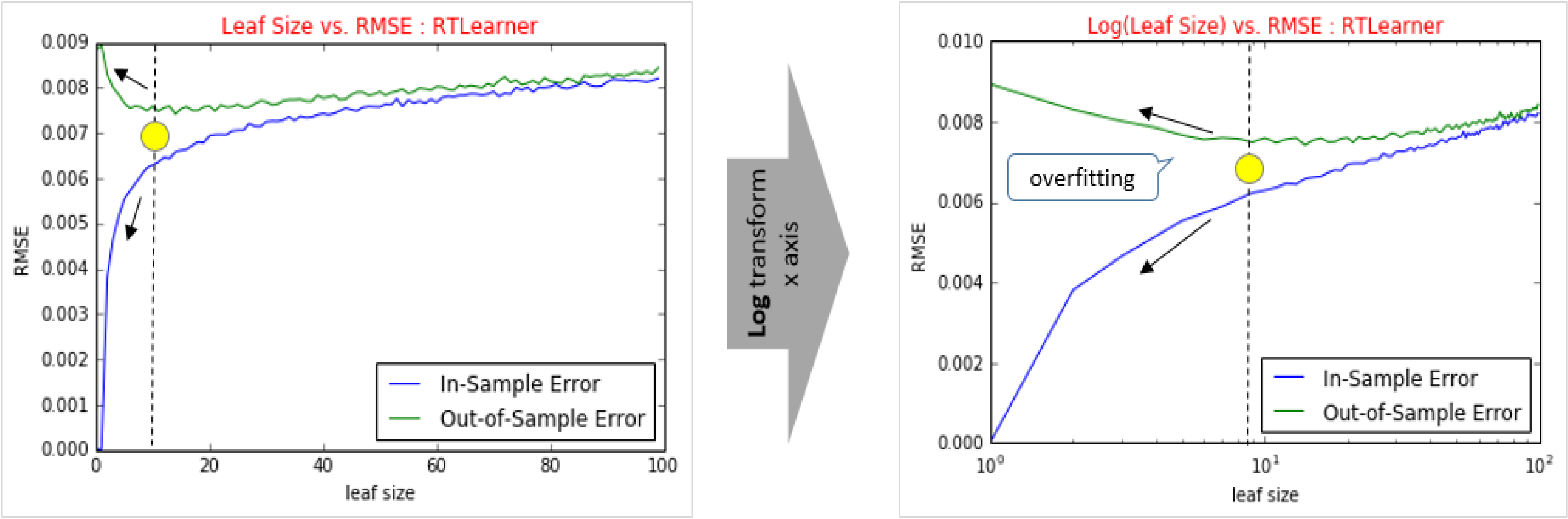


Fig 1-a

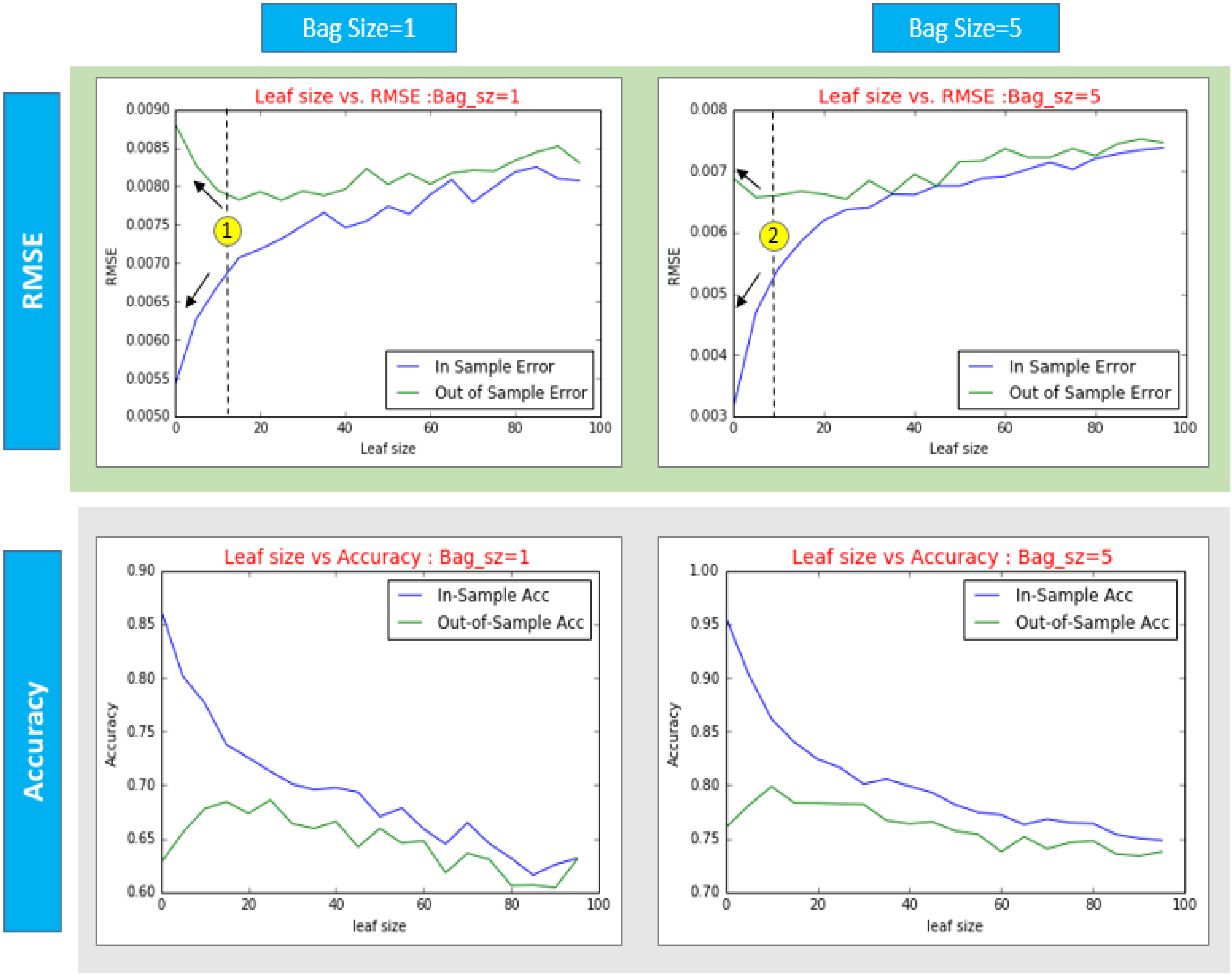
Q2.Can bagging reduce or eliminate overfitting with respect to leaf_size? Fix the number of bags and vary leaf_size to investigate. Provide charts and or tables to validate your conclusion.

- Learner : Bagging with Random Tree Learner
- Objective : show RMSE vs. leaf_size for **a fixed bag_size**
- Methodology :
  a) For leaf_size in range (1,100), compute in-sample and out-of-sample RMSE and Accuracy for bag_size=1.
  b) Shuffle data (Istanbul.csv) for each iteration in step-a.
  c) Repeat step–a 40 times and compute average RMSE and Accuracy for each leaf_size.
  d) Now, repeat step a to c for bag_size=5,15,20 and 50

Answer: yes, Bagging does help in reducing/eliminating over fitting.

- With bag_size=1, there are clear signs of over fitting for leaf_size in range 1-10. Both in and out of sample RMSE continue to improve as we move from leaf size 100->10. But in leaf size 1-10 region, two trends diverge. (Fig-2a, left charts.)

- As bag_size is increased from 1 to 5, overfitting reduces significantly. (Fig-2a, right charts.)
- For larger bag_size (such 20 or 50) no signs of overfitting are seen. (Fig-2b)



1. Overfitting is evident in leaf size range 1-10 in top left chart where in-sample RMSE (blue) decreases but out-of-sample RMSE starts to increase (green).

2. Overfitting reduces significantly in the same region for bag_size =5 (top right chart)

Fig-2a

Fig-2b below shows RMSE and Accuracy results for Bag size **15, 20 and 50**. Little to no overfitting is seen for these values of bag size. It can be seen in top charts, green line (out-of-sample error) doesn't increase as blue line (in-sample error) starts to decrease dramatically in 1-10 leaf_size region

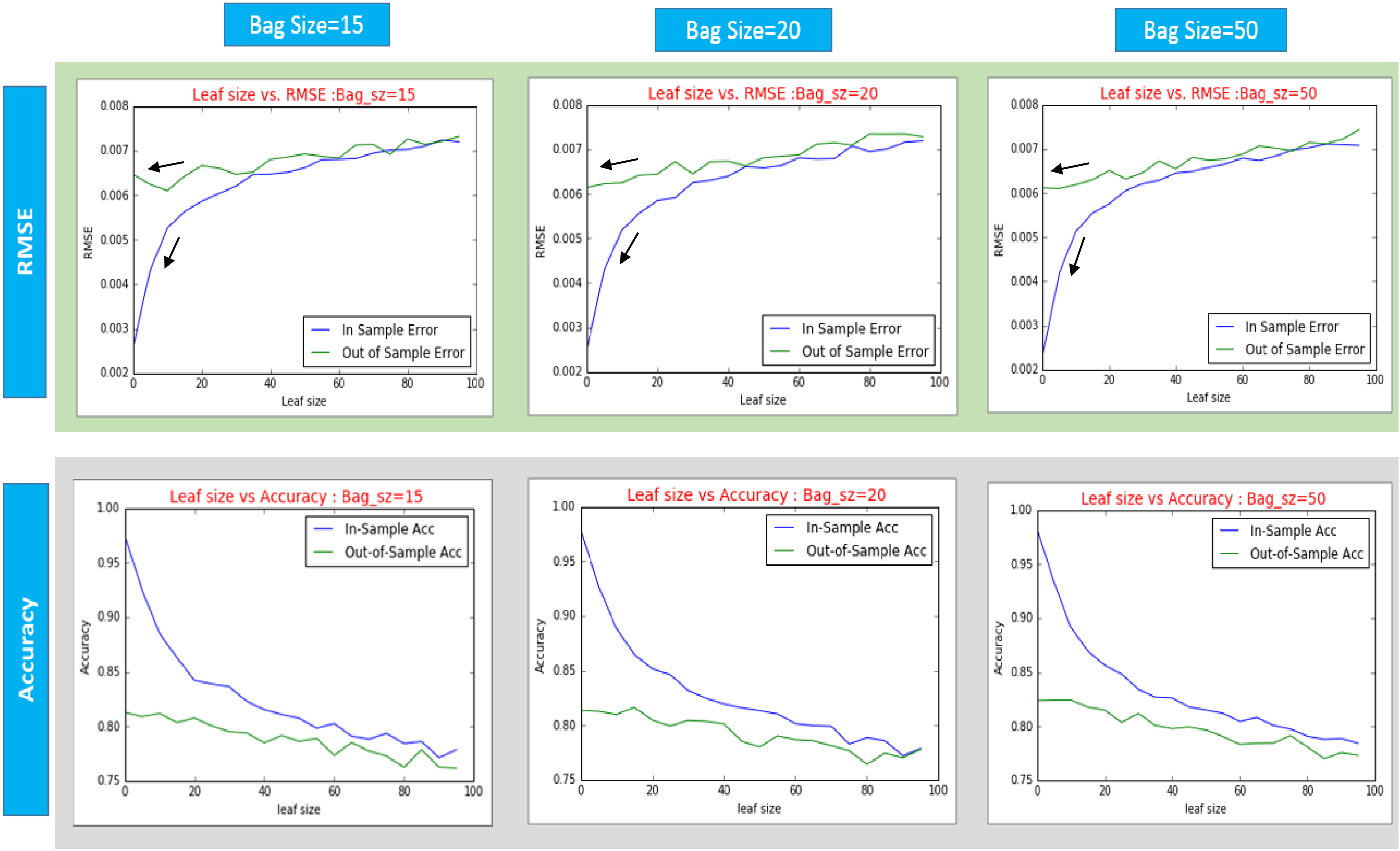


Fig-2b

Q.3Does overfitting occur with respect to number of bags? Choose some leaf_size and keep it fixed. How does RMSE vary as you increase the number of bags? Support your assertion with graphs/charts.

- Learner : Bagging with Random Tree Learner
- Objective : show RMSE vs. bag_size for **a fixed leaf_size**
- Methodology :
  a) For bag_ size in range (1,100), compute in-sample and out-of-sample RMSE and Accuracy, keeping leaf_size fixed at 1.
  b) Shuffle Istanbul.csv data for each iteration in step-a
  c) Repeat step-a 40 times and compute average RMSE and Accuracy for each bag_size.
  d) Now repeat step a to c for leaf_size=10, 20,50

Answer:

Experiments performed don’t show overfitting with respect to number of bags. (Fig 3a)

1. RMSE and Accuracy improves for both in and out of sample set for bag size in range 1 to 15 but after that both trends plateau out.
2. There is very little change in RMSE and Accuracy as bag size is increased from 20 to 100 for all four leaf sizes.
3. Overfitting is characterized by improvement in in-sample RMSE but degradation in out-of-sample RMSE. Data below shows no such trend and thus supports the assertion that there is no overfitting with respect to number of bags.

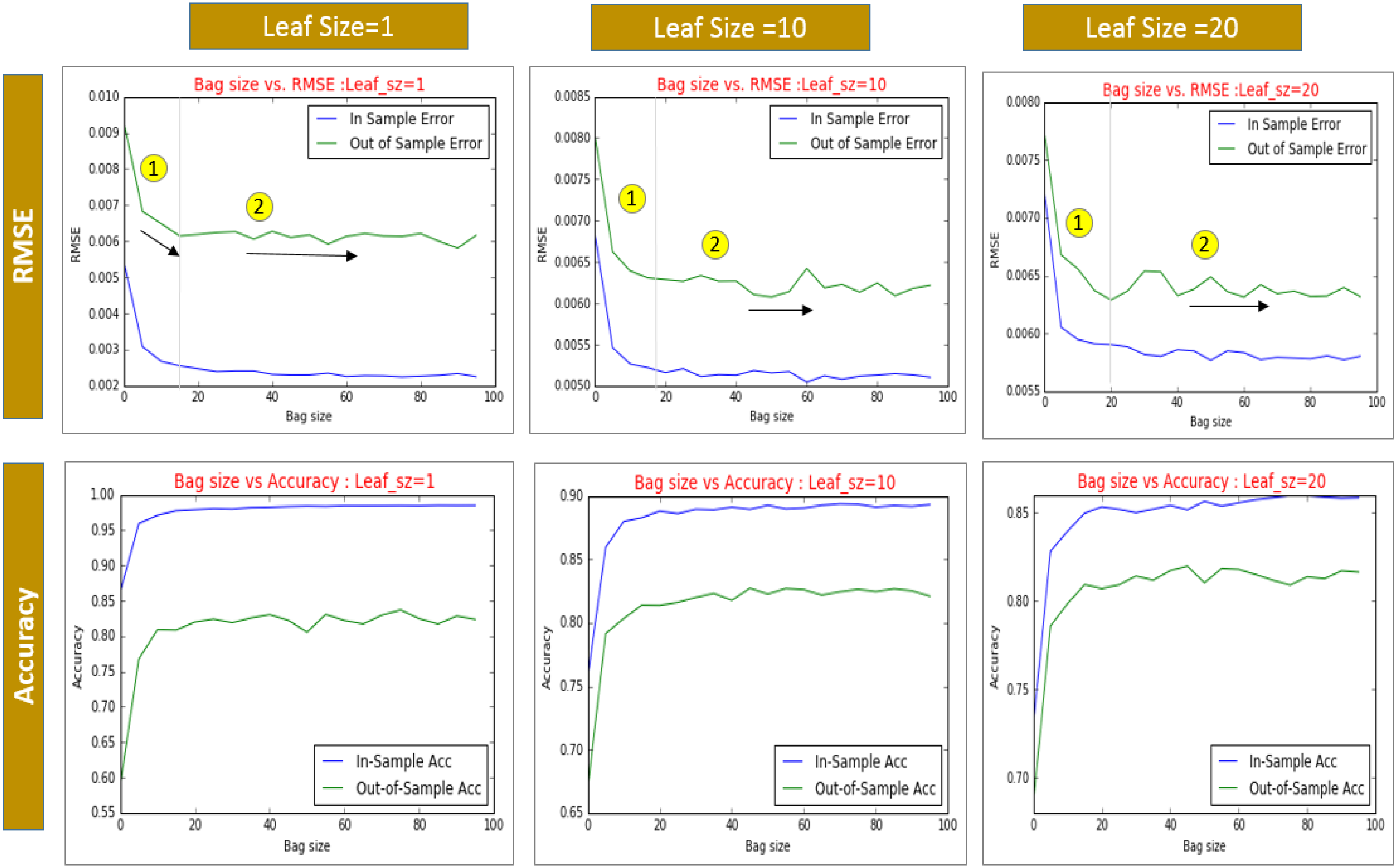

Leaf Size=1
Leaf Size =10
Leaf Size =20
RMSE
Accuracy
Bag size vs. RMSE :Leaf_sz=1
In Sample Error
Out of Sample Error
RMSE
Bag size
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0
20
40
60
80
100
1
2
Bag size vs. RMSE :Leaf_sz=10
In Sample Error
Out of Sample Error
0.0085
0.0080
0.0075
0.0070
0.0065
0.0060
0.0055
0.0050
1
2
Bag size vs. RMSE :Leaf_sz=20
In Sample Error
Out of Sample Error
0.0080
0.0075
0.0070
0.0065
0.0060
0.0055
1
2
Bag size vs Accuracy : Leaf_sz=1
In-Sample Acc
Out-of-Sample Acc
Accuracy
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Bag size vs Accuracy : Leaf_sz=10
In-Sample Acc
Out-of-Sample Acc
0.90
0.85
0.80
0.75
0.70
0.65
Bag size vs Accuracy : Leaf_sz=20
In-Sample Acc
Out-of-Sample Acc
0.85
0.80
0.75
0.70

Fig 3a