MSDM 5054

Assignment 1

Problem 1

1.Show explicitly how fitting a cubic spline regression can be solved by linear regression with equality constraints.

For a given set of knots $(k_1, k_2, \ldots, k_m)$, The general form of a cubic polynomial for a segment can be written as: $S_i(x) = a_i + b_i(x - k_i) + c_i(x - k_i)^2 + d_i(x - k_i)^3$

We have 4(m-1) coefficients for m-1 segments; m constraints for the values of the spline at the knots; m-1 constraints for the first derivatives at the knots; m-2 constraints for the second derivatives at the internal knots

$S(k_j) = y_j$ for j = 1, …, m

$S_i'(k_i) = S_{i+1}'(k_i)$ for i = 1, …, m-1

$S_i''(k_i) = S_{i+1}''(k_i)$ for i = 1, …, m-1

We want to minimize $E(\beta) = \sum_{j=1}^{m}\left(S(k_j) - y_j\right)^2$ .Using Lagrange multipliers, we can solve the constrained optimization problem to find the coefficients β that best fit the data while satisfying the constraints.

$$\varsigma(\beta, \lambda, \mu, \nu) = E(\beta) + \sum_{j=1}^{m} \lambda_j\left(S(k_j) - y_j\right) + \sum_{i=1}^{m-1} \mu_i\left(S_i'(k_i) - S_{i+1}'(k_i)\right) + \sum_{i=2}^{m-1} \nu_i\left(S_i''(k_i) - S_{i+1}''(k_i)\right)$$

$$\frac{\partial \varsigma}{\partial \beta} = 0$$

$$\frac{\partial \varsigma}{\partial \lambda} = 0$$

$$\frac{\partial \varsigma}{\partial \mu} = 0$$

$$\frac{\partial \varsigma}{\partial \nu} = 0$$

2. Explain the difference between piecewise polynomial regression and local

polynomial regression.

Piecewise Polynomial Regression is a type of segmented regression method that uses different polynomial models across various intervals. Each interval model is independent, except for some requirements for continuity or smoothness at the boundary points.

Local Polynomial Regression is a non-parametric regression method that fits a polynomial model near each data point and then obtains the overall regression curve by weighing these local fits. The weights are usually determined by a kernel function, which makes the local fits near each data point have a greater influence on the overall regression.

3. **Does increasing the bandwidth h in local polynomial regression increase or decrease the bias? How about the variance? Explain with details**

Increasing the bandwidth h typically reduces the model's variance because a larger bandwidth means more data points are used for local fitting, thereby reducing the impact of random fluctuations. However, this can also lead to an increase in the model's bias, as a larger bandwidth might make the model too smooth to capture the true trends in the data.

4. (b) Piecewise Constant Regression

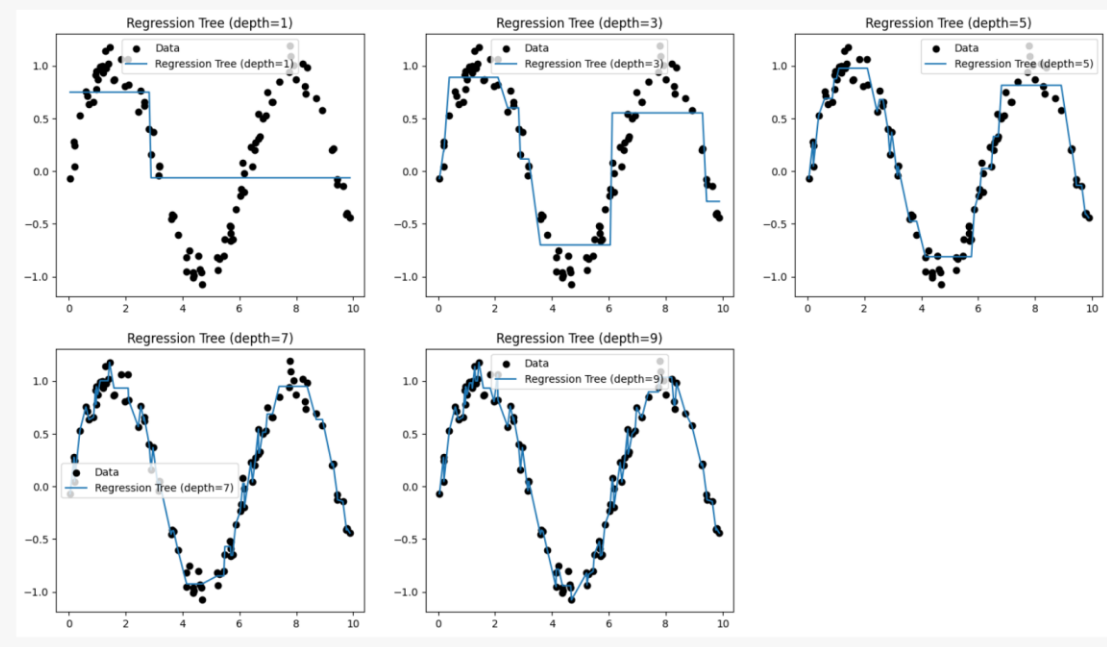
Regression trees fit data by recursively partitioning the data into smaller regions and using constant values in each region. This is like the idea of piecewise constant regression, which uses different constants across different intervals. The difference lies in the way the intervals are determined: regression trees use a tree structure to define these intervals, while piecewise constant regression may employ different methods to define the intervals.

5. When a tree gets finer, how does the bias and variance behave?

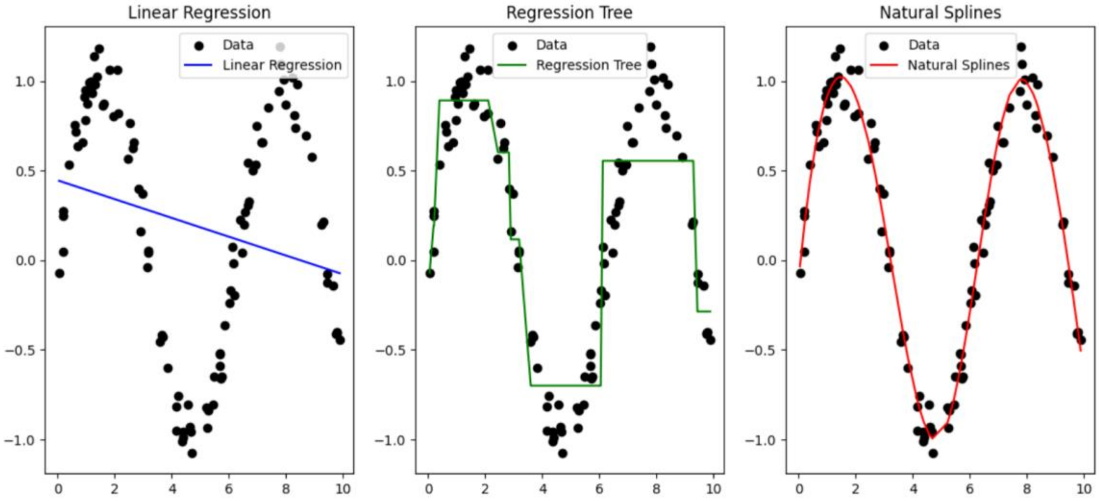
As the tree gets finer, the variance of the model typically increases because each leaf node contains fewer data points, making the model more sensitive to the fit of these data points. At the same time, the bias of the model may decrease because a more refined tree can better capture the complex structure of the data.

6. **From what I've been told through lecture, (b) Regression Tree should be selected.** However, I think it needs more detailed cases to be discussed more strictly. In a simple case as below, different depth of regression tree does make sense. As we can see, when the depth is enough deep, the bias could be quite small (0.001 when depth=9). It is easy to explain because this method can capture non-linear relationships and interactions

between variables.



However, in a normal choice, natural spline seems to be less biased more naturally.



7. Write down the formal and explicit definition of variable importance measure for random forest using math notations. How can we do variable section/model selection in random forest?

Let T be the set of trees in the random forest, and let N be the total number of trees.

For each tree t∈T, let Impurity(t) be the impurity measure. For a feature j, the mean decrease impurity (MDI) would be:

$$MDI_j = \sum_{t\in T} \sum_{n\in N_t} \frac{N_n}{N} \Delta Impurity_{j,n}$$

Where:

N_t is the number of nodes in tree t

N_n is the number of observations reaching node n

$\Delta Impurity_{j,n}$ is the decrease in impurity for feature j at node n.

The final importance score is then normalized:

$$Importance_j = \frac{MDI_j}{\sum_j MDI_k}$$

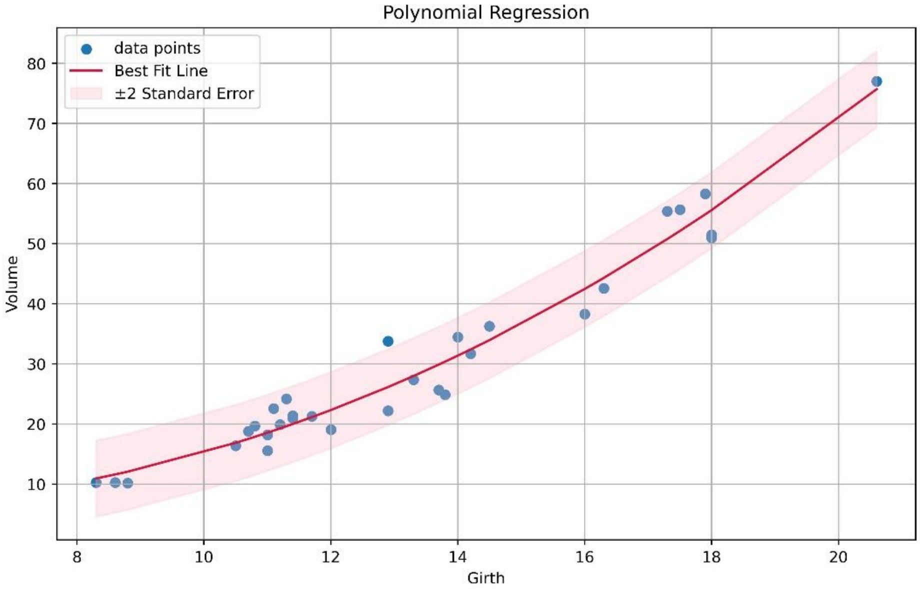
For the Mean Decrease Accuracy (MDA), the procedure involves measuring the accuracy of the model and assessing the drop in accuracy when a feature is permuted. Let $\text{Accuracy}_{\text{original}}$ be the accuracy of the model on the validation set using all features, and $\text{Accuracy}_{\text{j,perm}}$ be the accuracy after permuting feature j:

$$MDA_j = Accuracy_{original} - Accuracy_{j,perm}$$

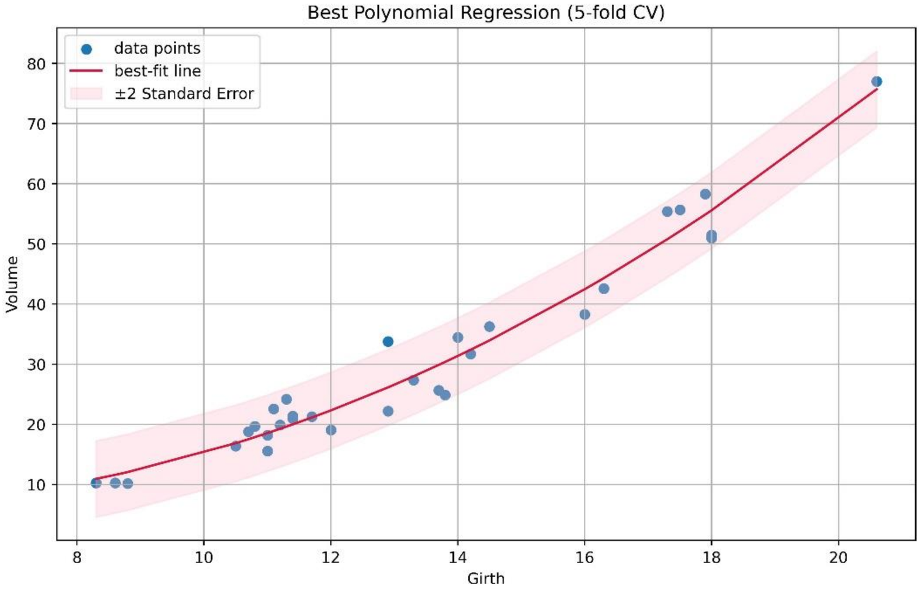
Variable Selection/Model Selection in Random Forests can be performed using the variable importance measures described above. Firstly, compute importance scores (such as MDI or MDA) for all features. Then select features: set a threshold for the importance scores. Features with scores above this threshold are selected for the model. After that, cross-validation to assess model performance with the selected features to validate the effectiveness of the selected subset. Repeat the process by refining the feature set based on model performance until an optimal subset is identified.

Problem 2

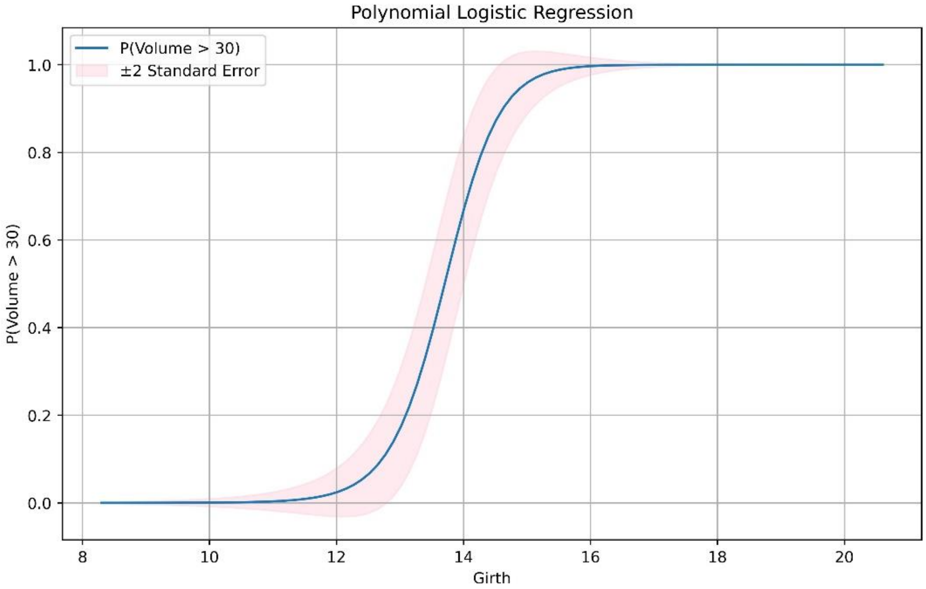
1. The model with the largest adjust R-squared.



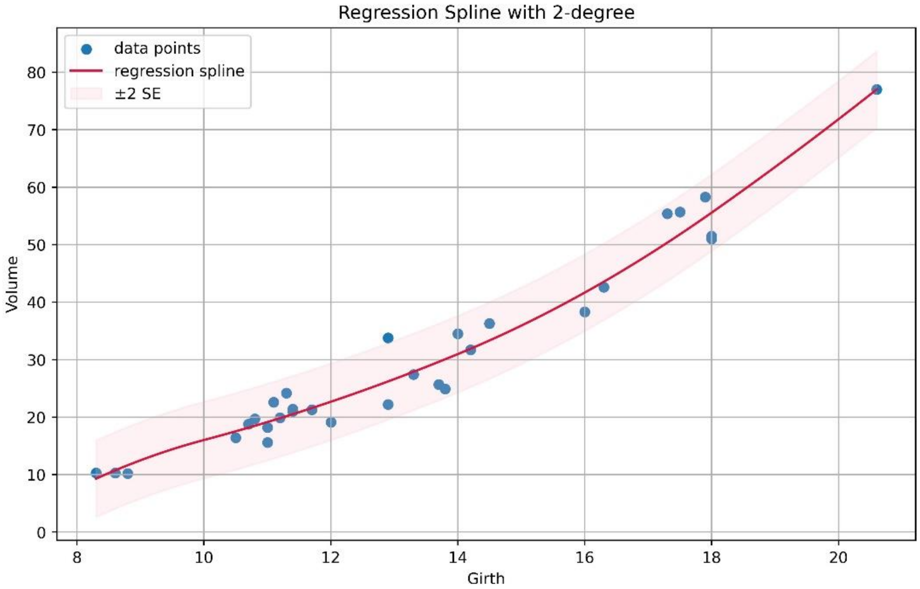
the model using 5-CV error



2.

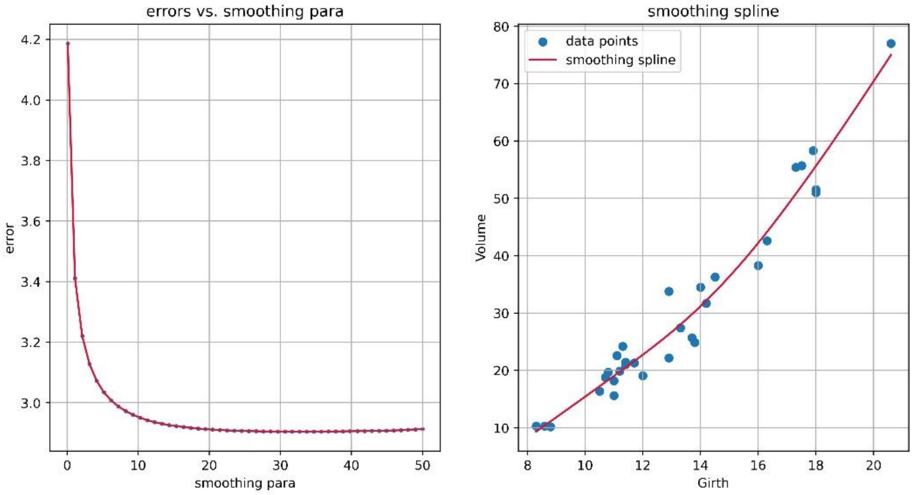


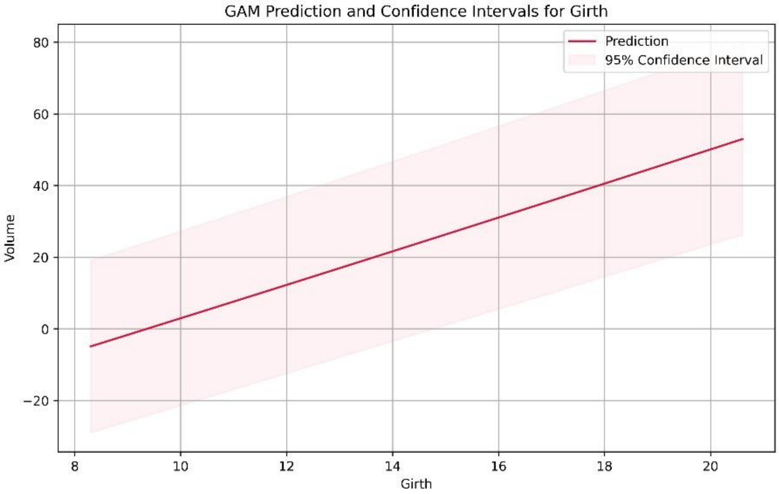
3.



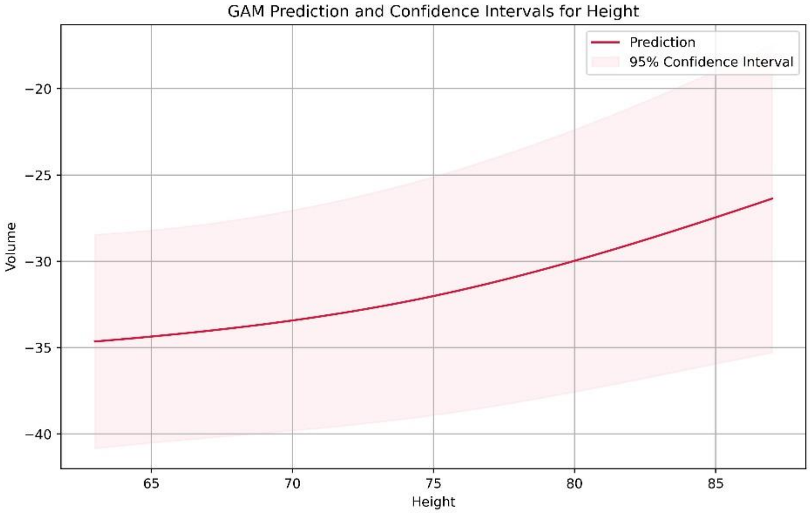
4. The smoothing level (32.69) is chosen by Cross-Validation.

The used degrees of freedom: 3.87


errors vs. smoothing para
error
4.2
4.0
3.8
3.6
3.4
3.2
3.0
0
10
20
30
40
50
smoothing para
smoothing spline
data points
smoothing spline
Volume
80
70
60
50
40
30
20
10
8
10
12
14
16
18
20
Girth

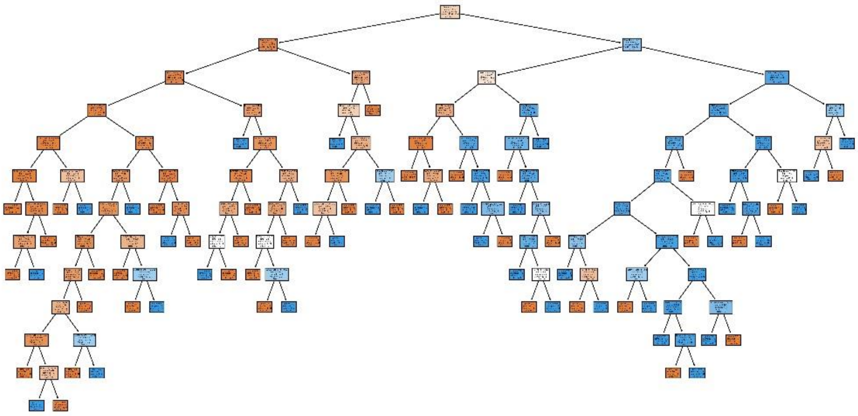

GAM Prediction and Confidence Intervals for Girth
Prediction
95% Confidence Interval
Volume
80
60
40
20
0
−20
8
10
12
14
16
18
20
Girth

5.


GAM Prediction and Confidence Intervals for Height
Prediction
95% Confidence Interval
Volume
−20
−25
−30
−35
−40
65
70
75
80
85
Height

Problem 3

1.



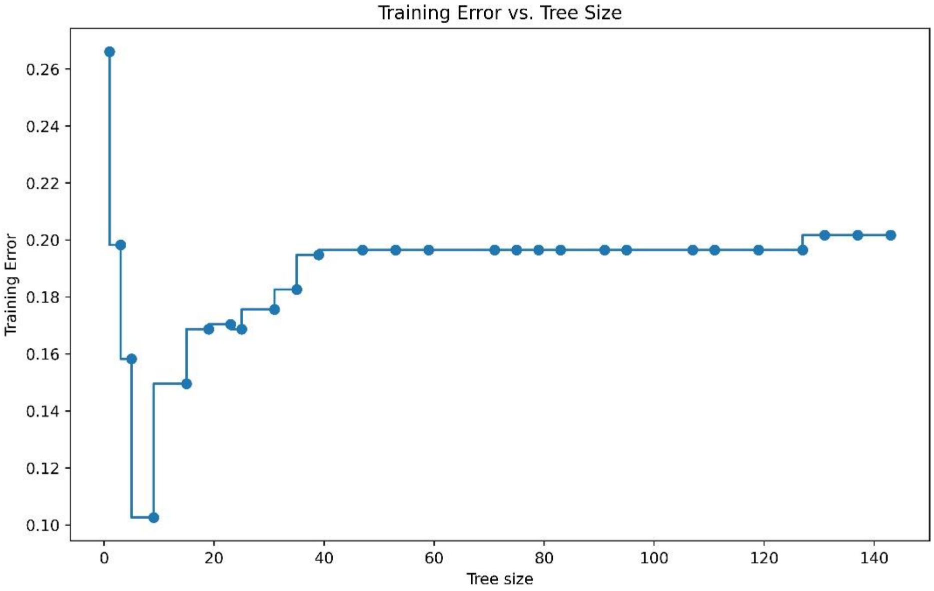
Tree Training error: 0.000

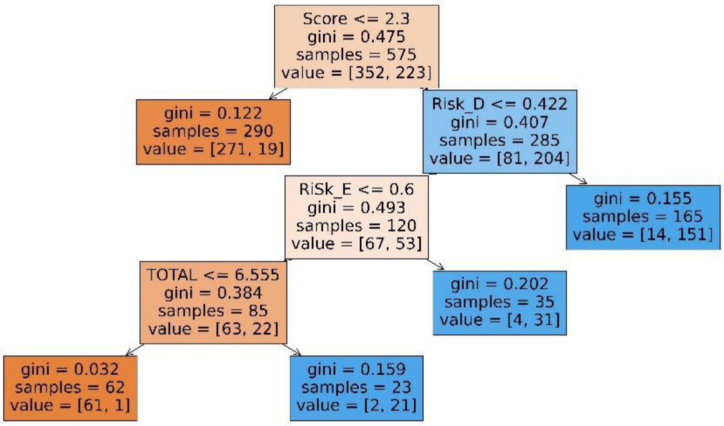
Tree Test error:0.107

Tree Confusion Matrix:

```
[[98 11]
 [10 78]]
```

2.





Test Error after Pruning: 0.05076

3. Random Forest training error (m=13): 0.002
4. Best m value: 14, with smallest training error rate 0.0017,
   Selected Random Forest Confusion Matrix:
   ```
   [[105    4]
    [  4   84]]
   ```
   Selected Random Forest Test Error:0.0406

5. Comparison Summary:

Initial Tree Test Error: 0.1066

Pruned Tree Test Error: 0.0508

Random Forest (m=13) Training Error: 0.0017

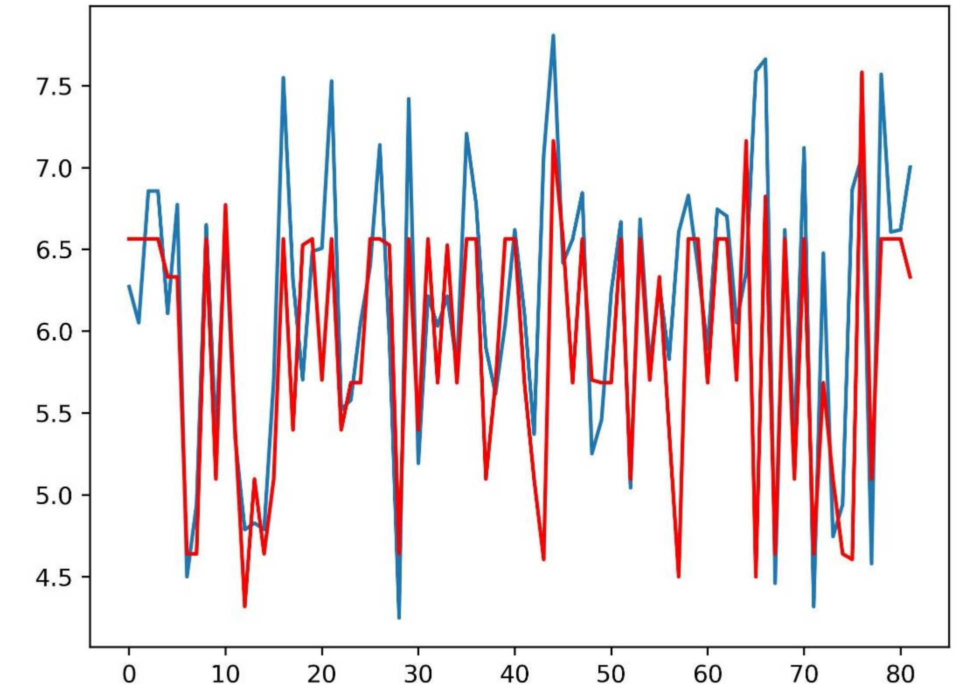
Best Random Forest (m=14) Test Error: 0.0406

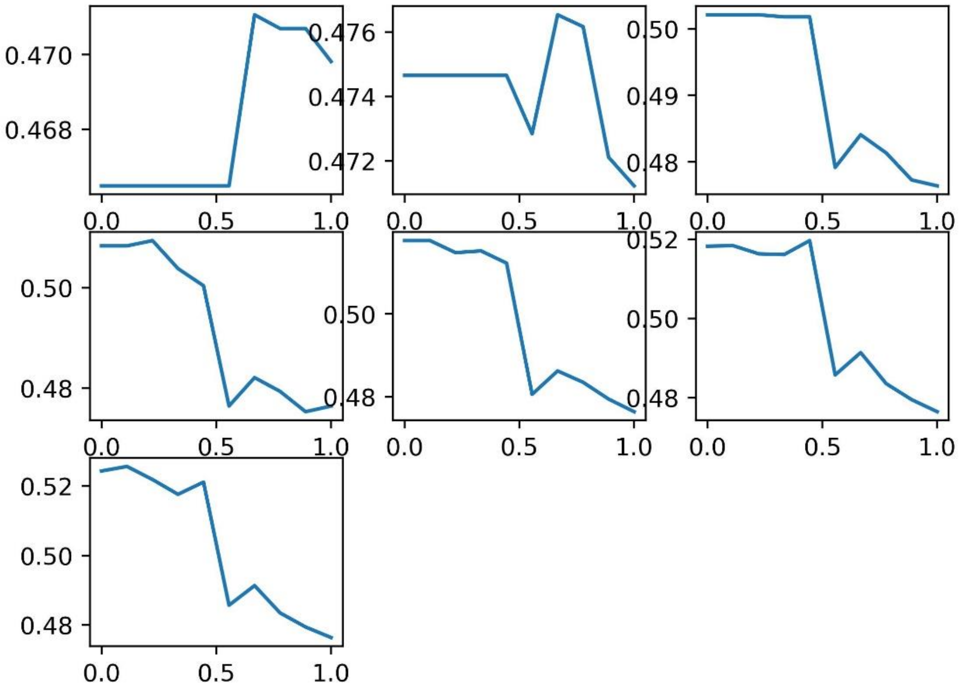
Pruning helps with decreasing the bias when predicting (Test Error Rate significantly fell from 0.106 to around 0.05). Random Forest performs even well with a little pros in test error.

Problem 4

After Tuning with max_depth from 4 to 11 and min_loss_threshold between 0 and 1, my model ends with Max Depth: 4.0 and Min Loss Threshold: 0.0.

Best Test Error: 0.47641696233738035


7.5
7.0
6.5
6.0
5.5
5.0
4.5
0
10
20
30
40
50
60
70
80


0.470
0.468
0.476
0.474
0.472
0.50
0.49
0.48
0.50
0.48
0.50
0.48
0.52
0.50
0.48
0.52
0.50
0.48
0.0
0.5
1.0