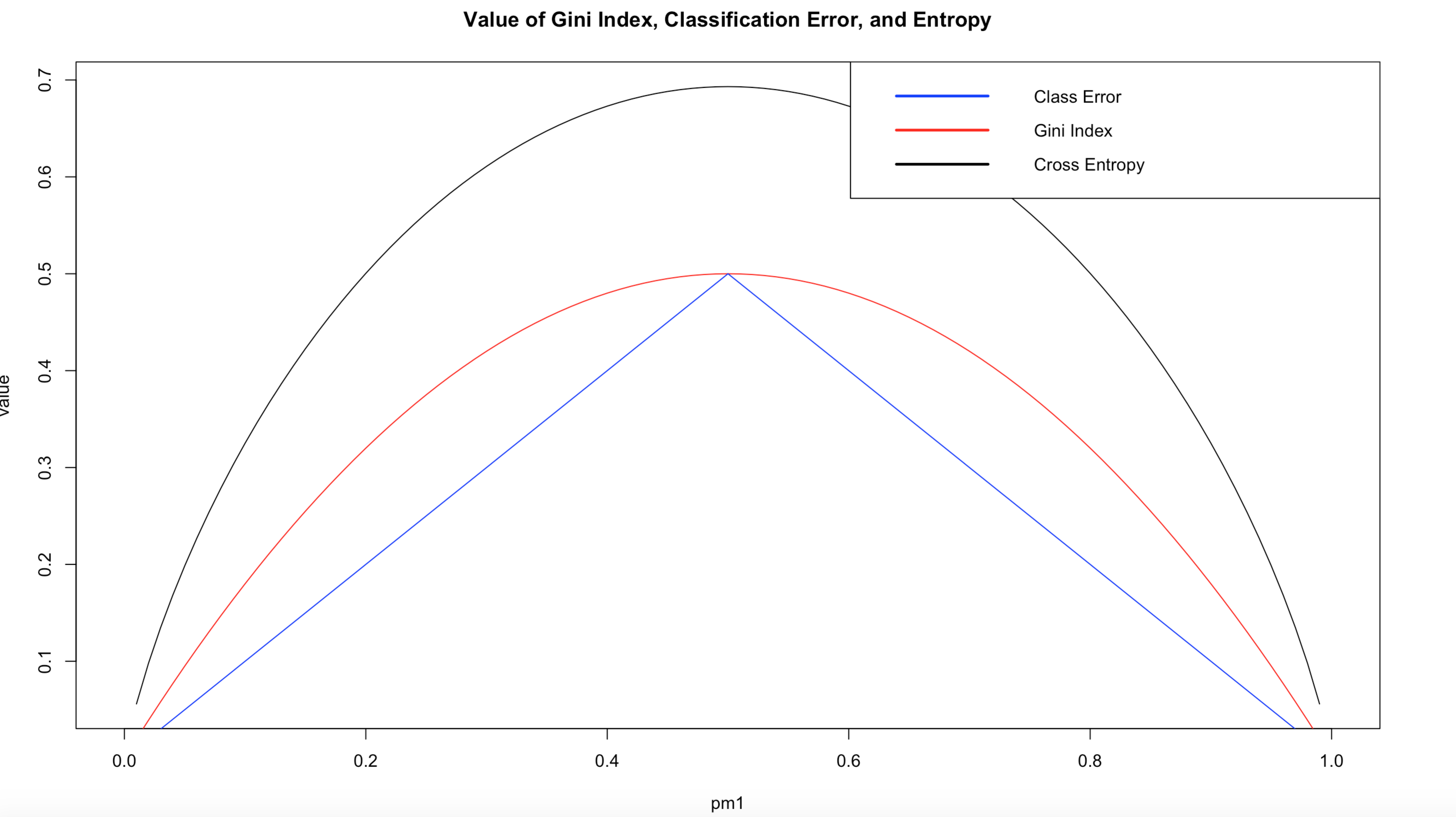
Andrew Hillard

Data 3

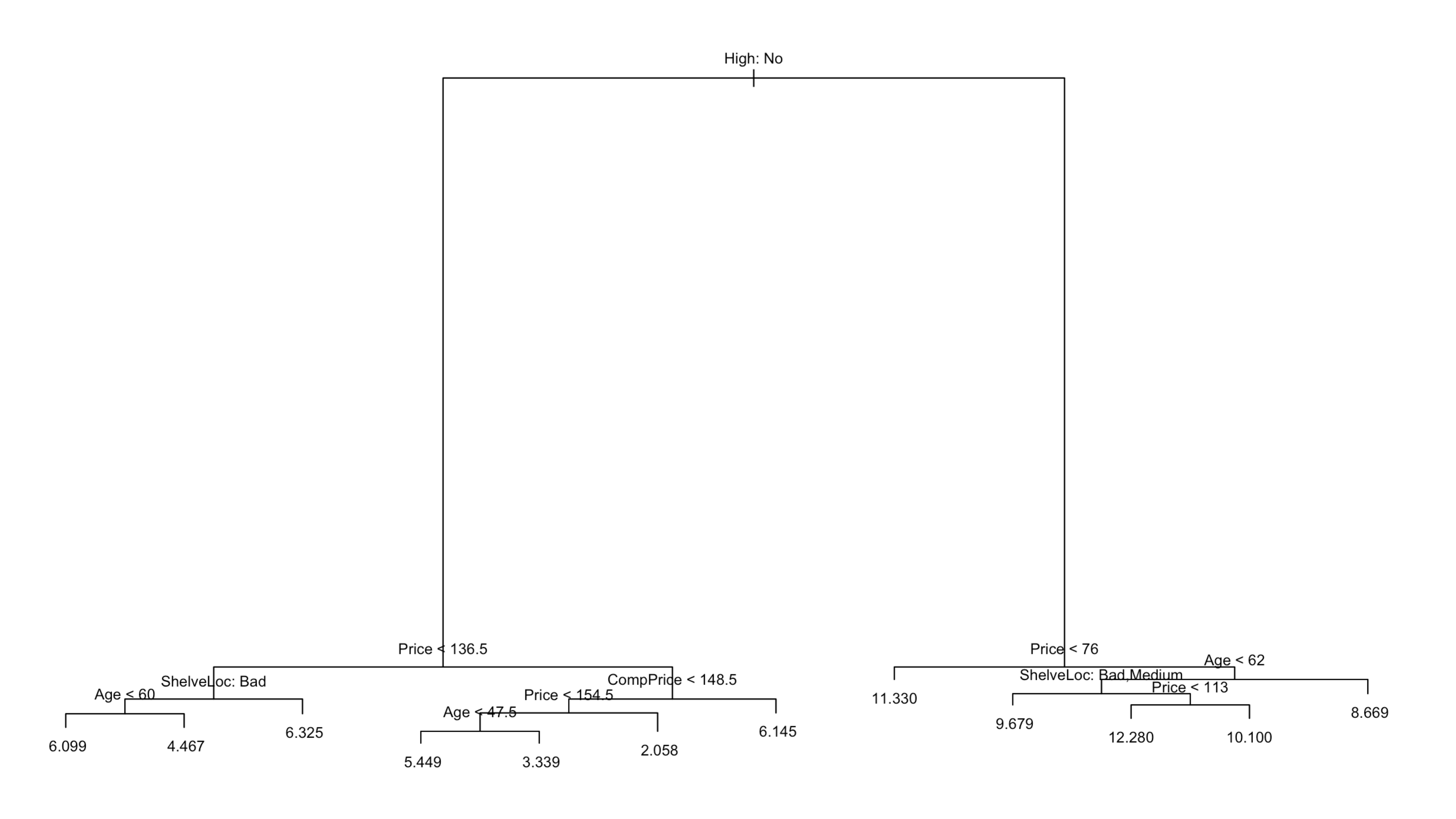
Homework 5

Problem 1



Problem 2

a. The tree diagram plot is given below. The high and price variables determine the top three splits in the tree diagram and are therefore the most important variables. When high is no, the predicted sales are lower. When high is yes, the predicted sales are higher. After this first split in the tree diagram, the predictions become too complicated to make generalizations from the tree diagram. The test MSE is 2.793222.

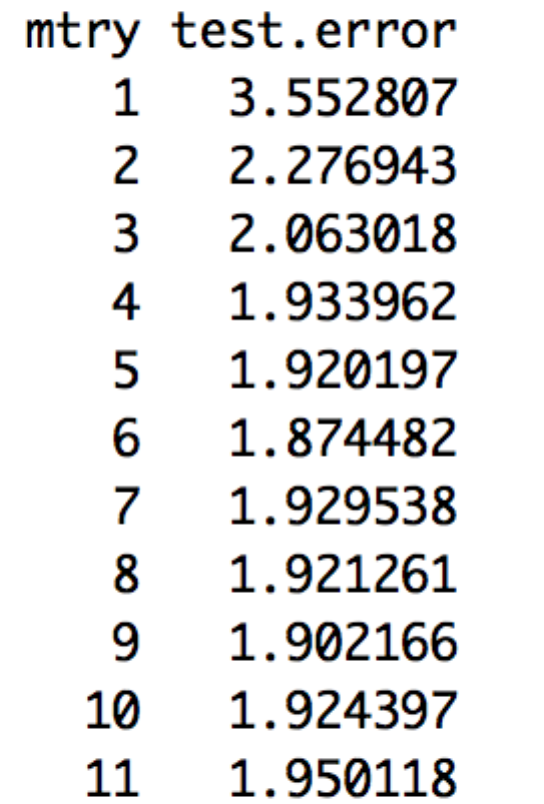


b. The pruned tree has a test MSE of 2.8444887. Therefore, pruning the tree does not improve the test MSE.

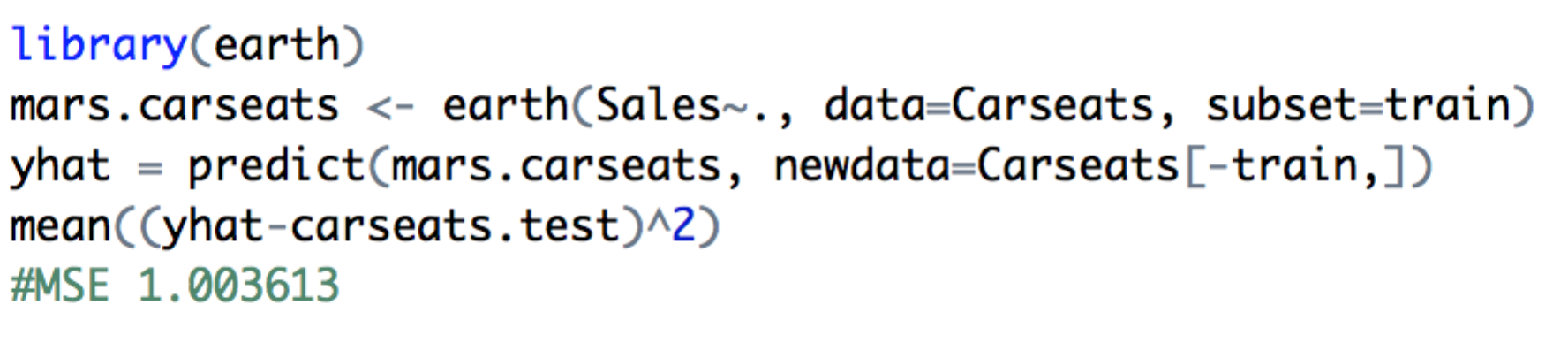
c. The test MSE is 1.944736. The high and price variables are the most important variables according to the importance function. Both high and price have %IncMSE and IncNodePurity values that are much greater than the other variables.

d. The test MSE is 1.269095. I used the Gaussian distribution, 31607 trees, and an interaction depth of 1. I chose the values for “n.trees” and “interaction.depth” by using a for loop. The for loop iterated over different values of interaction depth. For each interaction depth, I returned the cross validation error (5-fold) and used the gbm.perf to return the optimal number of trees. Out of all the tested values of interaction depth, the lowest cross validation error was 1.471904, which occurred when the interaction depth was 1 and the optimal number of trees was 31607.

e. The test MSE was 1.874482 when “mtry” equaled 6. The most important variables were high and price, similar to part c. The cross validation error decreased as the value of “mtry” increased to 6. As “mtry” increased beyond 6, the cross validation error increased and then stabilized around 1.92.



f. The test MSE is 1.003613. I used the earth function in the earth package, which automatically optimizes the number of splines in the MARS model. Therefore, the code is simple and given below.



Problem 3

The model that performed the best on the Pima dataset was CART with Boosting, which had a test misclassification equal to 0.2078313. Below gives the results and parameters used to construct the four models. Whenever possible, the seed was set to 1 before the model was constructed.

* CART with Bagging: Test Misclassification is 0.246988.
* CART with Boosting: Test Misclassification is 0.2078313. The optimal boosting model was created using the gbm function with n.trees = 4452 and interaction.depth = 1. The values for n.trees and interaction.depth were found using a for loop using the method described in Problem 2 Part d.
* Random Forests: Test Misclassification is 0.2319277. The optimal Random Forest had mtry = 2. The value for mtry was determined using a for loop that iterated over possible values of mtry and returned the misclassification value for each iteration.
* MARS: Test Misclassification is 0.253012. Used earth optimized function.