**Problem 1: Conceptual Review**

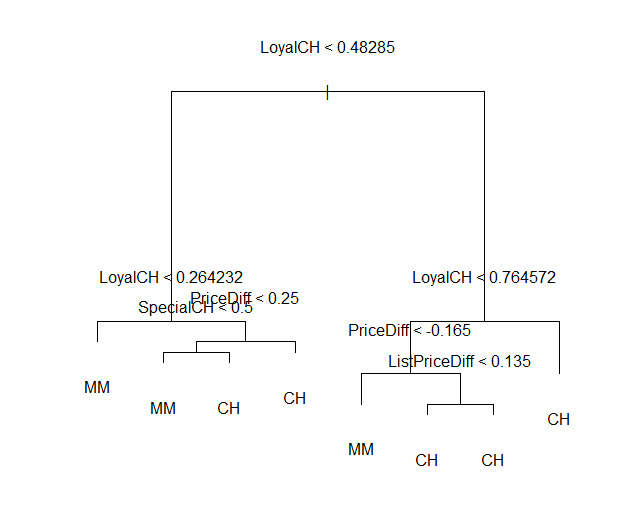
1. For majority vote approach, classify X as red since it is the most commonly occurring class from the 10 predictions (6 for red vs 4 for green). With the average probability approach, classify X as green as the average of the 10 probabilities is 0.45.
2. A diagram of a tree

   Description automatically generated with medium confidence
3. A paper with numbers and lines

   Description automatically generated
4. Bagging is a technique used to reduce the variance of a model by combining multiple independent models. The idea behind bagging is to create several bootstrap samples from the training data and train a separate model on each sample. The final prediction is then obtained by averaging or voting across the predictions of these individual models. In theory, bagging tends to reduce variance because it averages out the nuances of individual models. But there could be scenarios where bagging might result in a higher variance than an individual tree. For example, when the base models are highly overfitted, if each individual tree in the ensemble is overfitting the training data and making errors in different ways, combining them through bagging may not necessarily improve the overall performance. The diversity introduced by overfitting in each model may lead to an ensemble that is more sensitive to variations in the training set, resulting in higher variance.

**Problem 2: Basics of Decision Trees**

1. see Rscript
2. The number of terminal nodes is 8. Training error rate is 0.1525
3. One of the terminal nodes is node label 4. The split criteria is LoyalCH < 0.264232. This sub tree has 161 observations with a deviance of 112.800. The overall prediction of this sub tree is MM. About 11.180% of observations as CH and 88.820% of observations as MM.
4. Based on the tree plot, it looks like the most important splitting variable is LoyalCH. The top 3 nodes have splitting variable LoyalCH

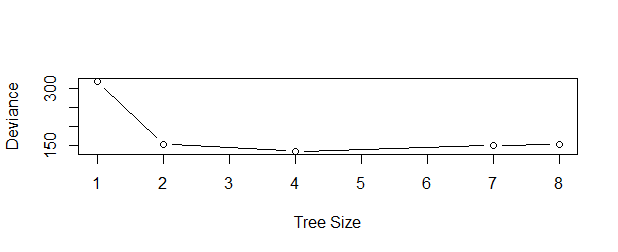


1. The confusion matrix is:  
   A black background with white text

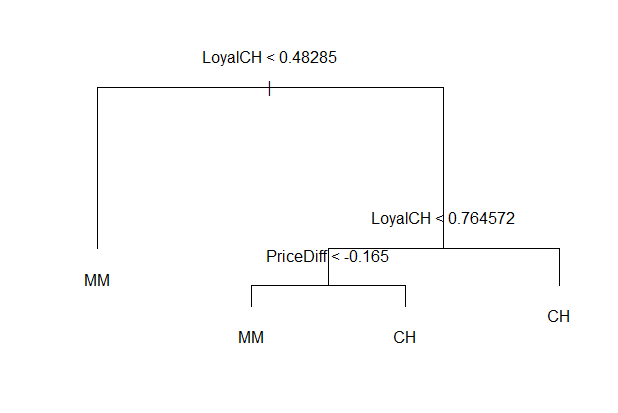
   Description automatically generated

The test error rate is 0.1740741

1. Plot with tree size on the x-axis and cross-validated classification error rate on the y-axis:



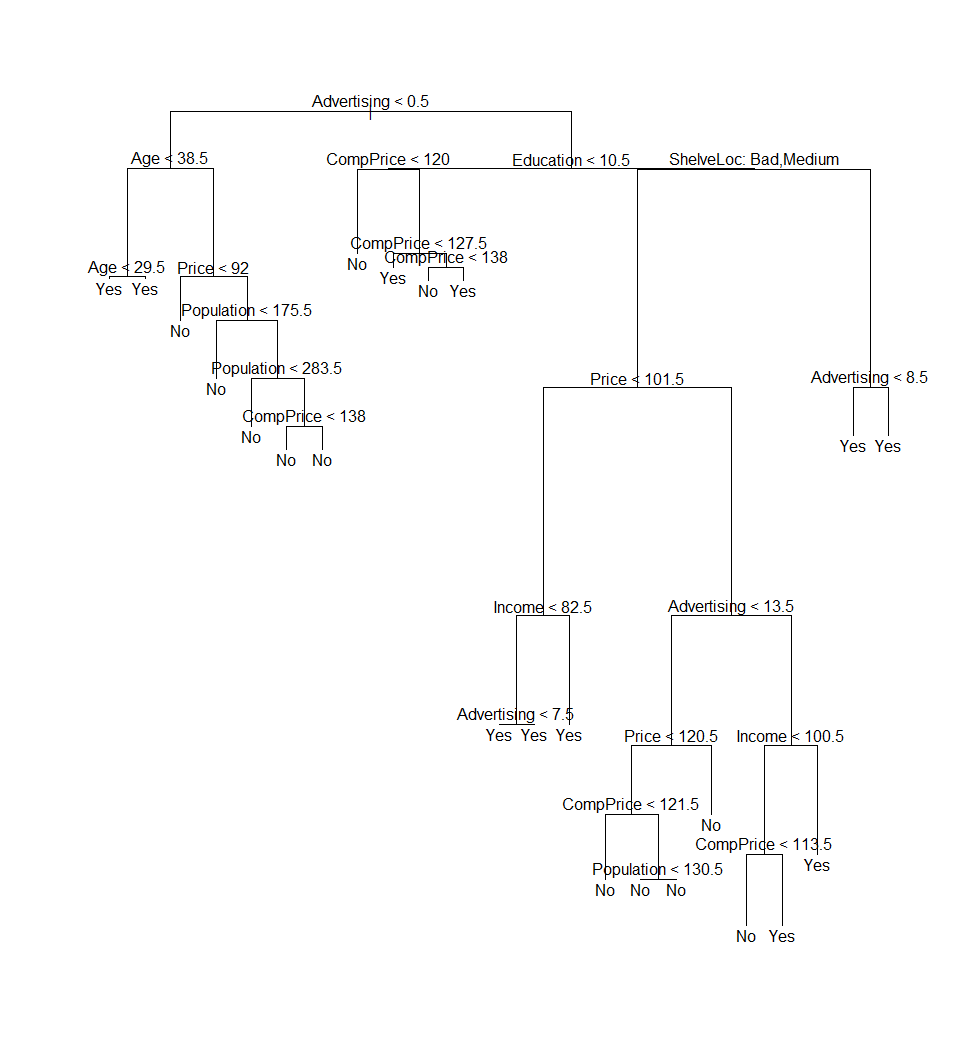
1. From the plot, tree size 4 has the lowest error.
2. My optimal tree size obtained using cross-validation was 4.



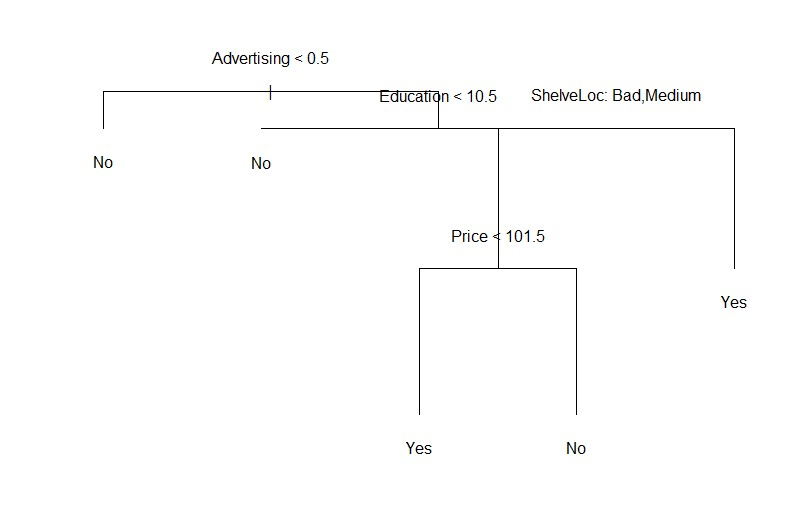
1. The training error rate for the unpruned tree is 0.1525. The training error rate for the pruned tree is 0.165. The training error rate of the pruned tree is higher because by pruning, we have reduced the flexibility in the model. Bias has increased, but variance of the model has reduced.
2. The test error rate for the unpruned tree is 0.1740741. The test error rate for the pruned tree is 0.1851852. By pruning, we have increased the interpretability of the tree model but accuracy didn’t improve. We have reduced the flexibility of the model by pruning it, but the reduction in variance of the model couldn’t overcome the increase in bias and resulted in a small increase in test error rate.

**Problem 3: Bagging and Random Forests**

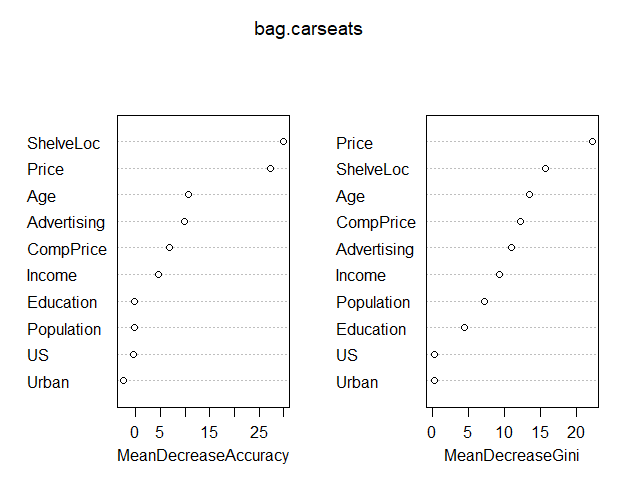
1. See Rscript
2. The classification tree is listed below. The training error is 0.185. The test error is 0.335.



1. The optimal size tree is 5. The test error for the pruned tree is 0.325



1. The test error is 0.175. The most important variables are Price and ShelveLoc.



1. Table of m and their corresponding test errors

A screenshot of a computer screen

Description automatically generated

1. In the context of random forests, choosing the value of m that gives the smallest test error may not always be appropriate. This is because the goal of random forests is to introduce randomness and decorrelation among trees, and the choice of m is part of this randomness. Instead of focusing solely on the smallest test error for a specific m, include out of bag error to perform a more comprehensive tuning to optimize the overall model performance.
2. I used k-fold cross validation with k=5. This gave me the following test errors:

A screenshot of a computer

Description automatically generated

The cross validation indicated that m=4 is the optimal m.

1. (i) The 4th observation appears in 300 out of the 500 bootstrapped trees in the random forest model.

(ii) No is the majority class with a proportion of 0.9375. Yes is the minority class with a proportion of 0.0625

(iii) The OOB error estimation is 0.195