Homework Chapter 8

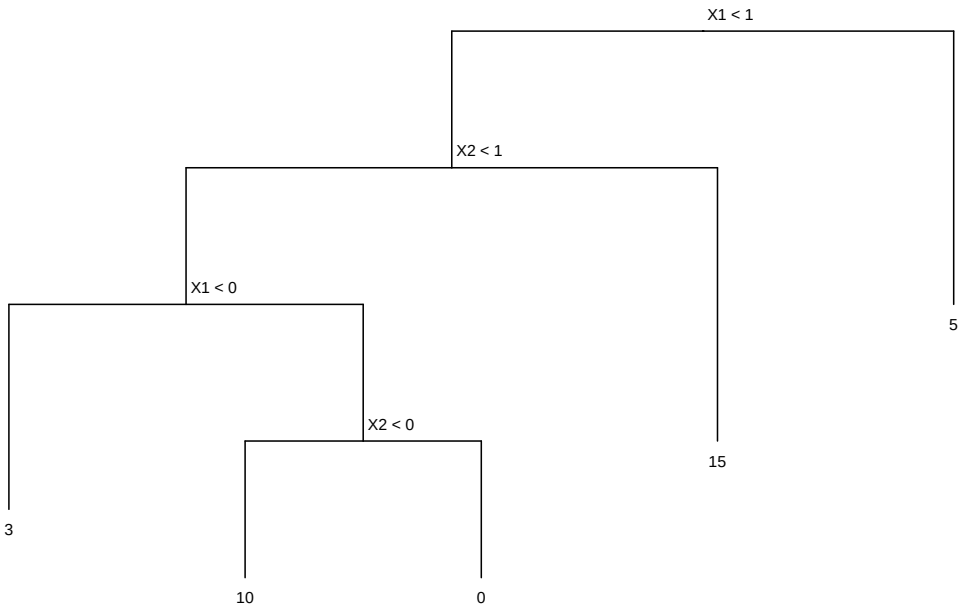
Jacob Thielemier

1 April 2024

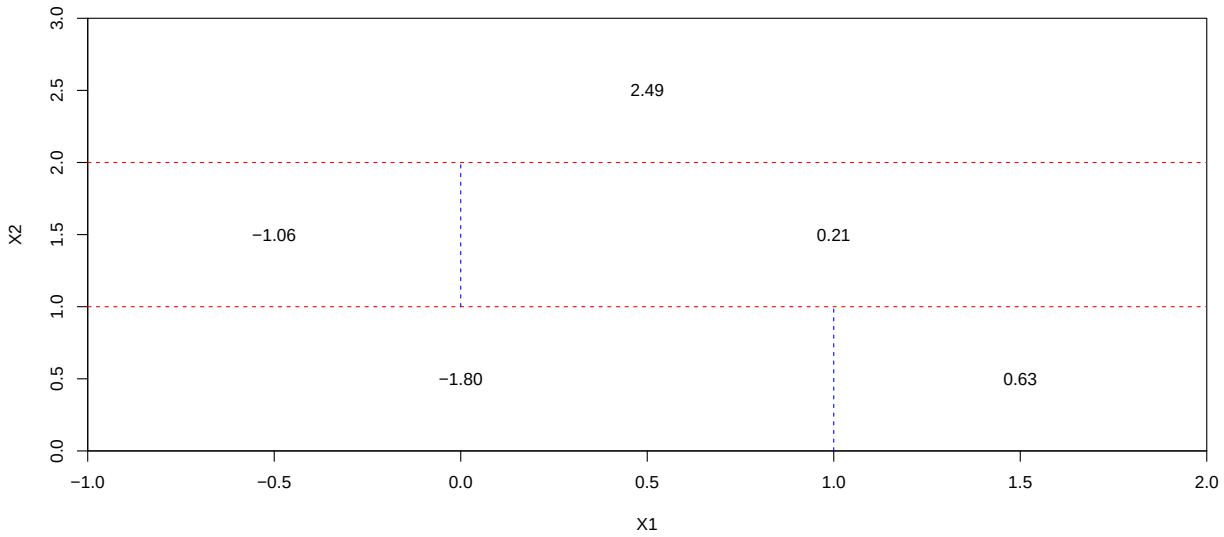
Question 1

- The randomForestSRC package in R provides several advancements and additional functionalities over the traditional randomForest package. Mainly, randomForestSRC implements Breiman's random forests for a wide variety of problems using fast OpenMP parallel processing. Key features include versatile data handling, advanced analytical methods, enhanced variable importance metrics, imputation methods, advanced clustering, parallel computing capabilities, and additional functionalities.

Question 2

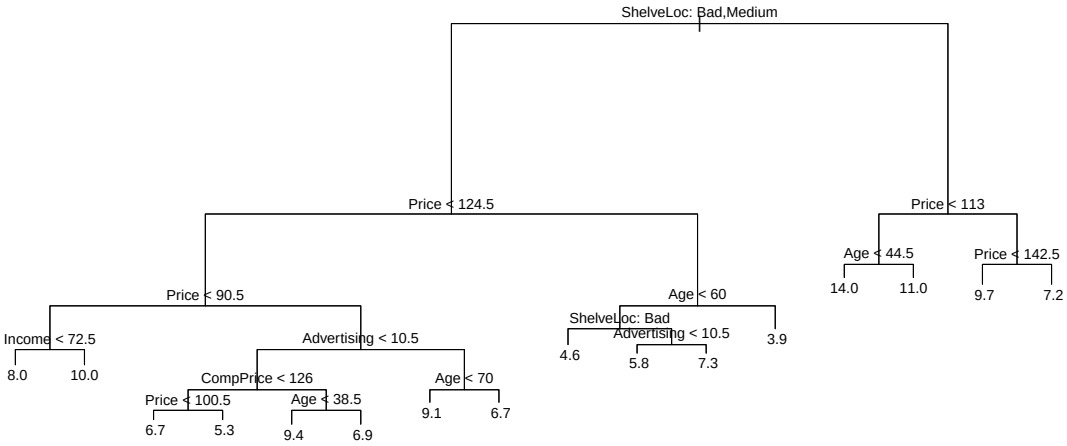


Part (a)



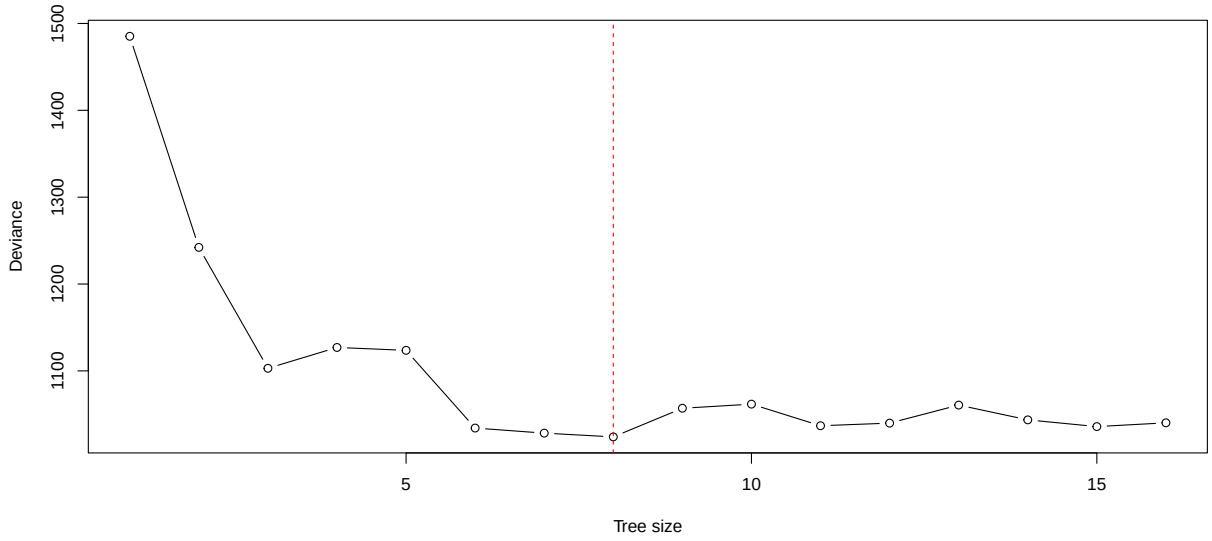
Part (b)

Question 3



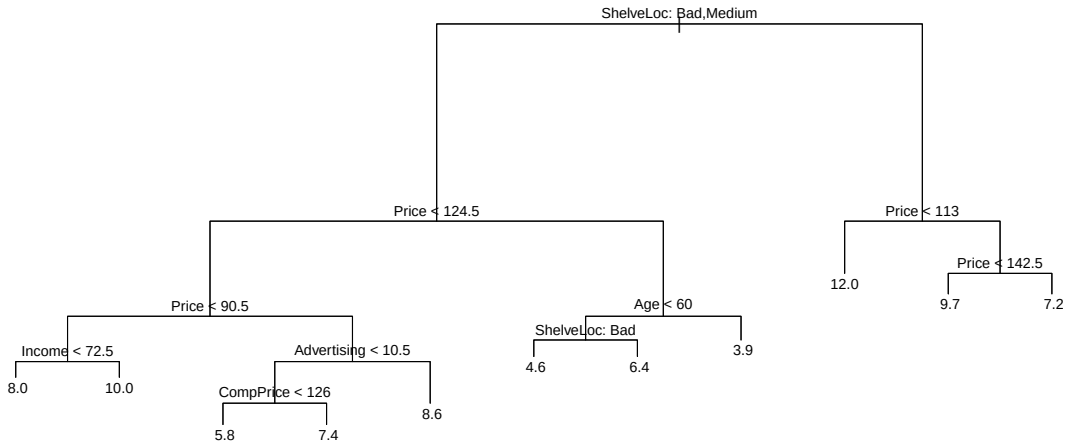
Part (a and b)

- Shelve location and Price are the most important predictors, same as with the classification tree. The Test MSE is: 3.04



Part (c)

- Pruning improves performance very slightly (though this is not repeatable in different rounds of cross-validation). Arguably, a good balance is achieved when the tree size is 11.



Part (d)

- The test error rate is ~2.1 which is an improvement over the pruned regression tree.

Part (e)

- The test error rate is ~2.2 which is an improvement over the pruned regression tree, although not quite as good as the bagging approach.

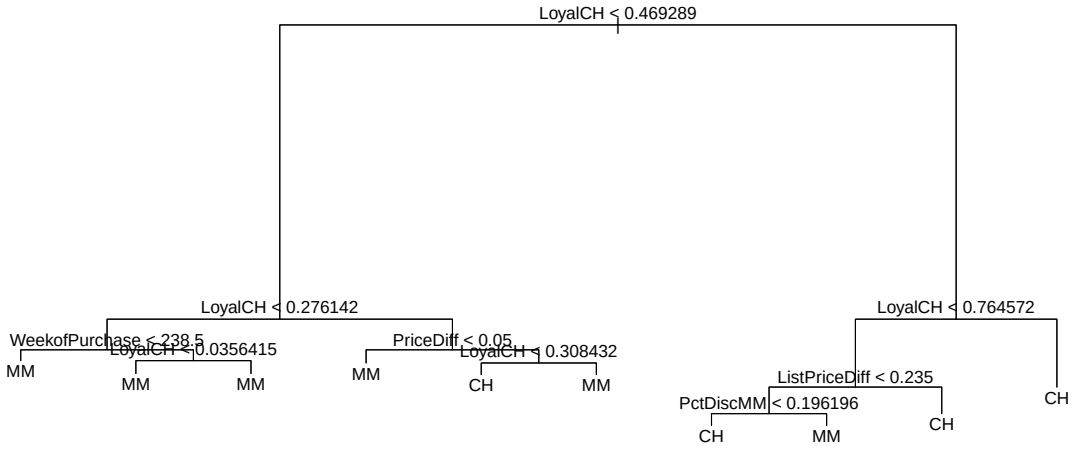
Part (f)

- Using BART, the test error rate is ~1.6 which is an improvement over random forest and bagging.

Question 4

Part (a, b, c, and d)

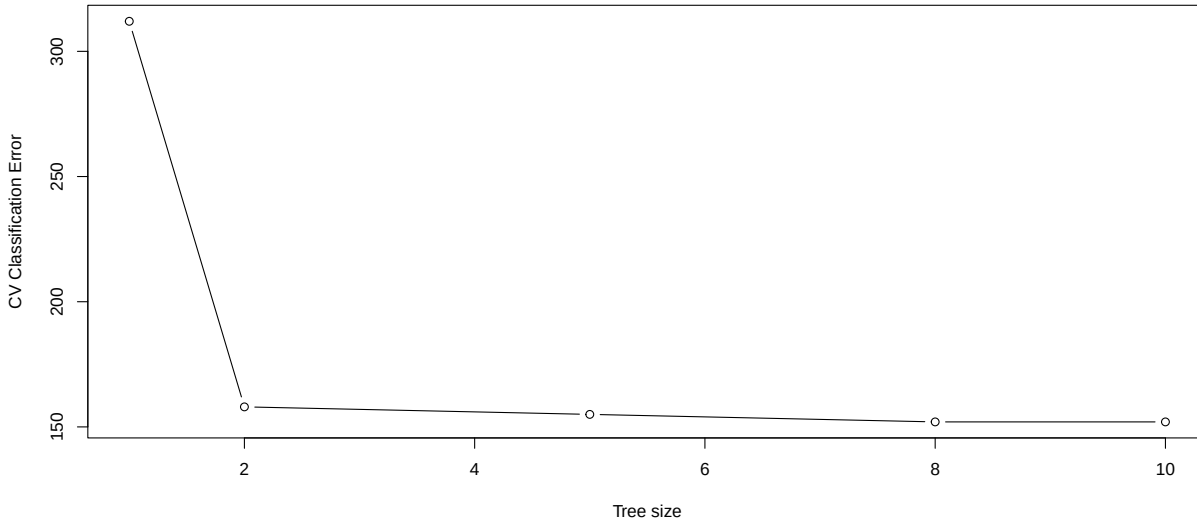
- The training error rate is 0.15, and there are 10 terminal nodes. The residual mean deviance is high, and so this model doesn't provide a good fit to the training data.
- Branch 8 results in a terminal node. The split criterion is `WeekofPurchase < 238.5` and there are 49 observations in this branch, with each observation belonging to MM. The final prediction for this branch is `MM`.



- `LoyalCH`(Customer brand loyalty for Citrus Hill) is the most important variable. Only five variables out of 18 are used.

Part (e)

- Test error rate of 0.21. This is higher than for the training set and is as expected.



Part (f, g, and h)

- Trees with 10 or 8 terminal nodes have the lowest CV Classification Errors.

Part (i and j)

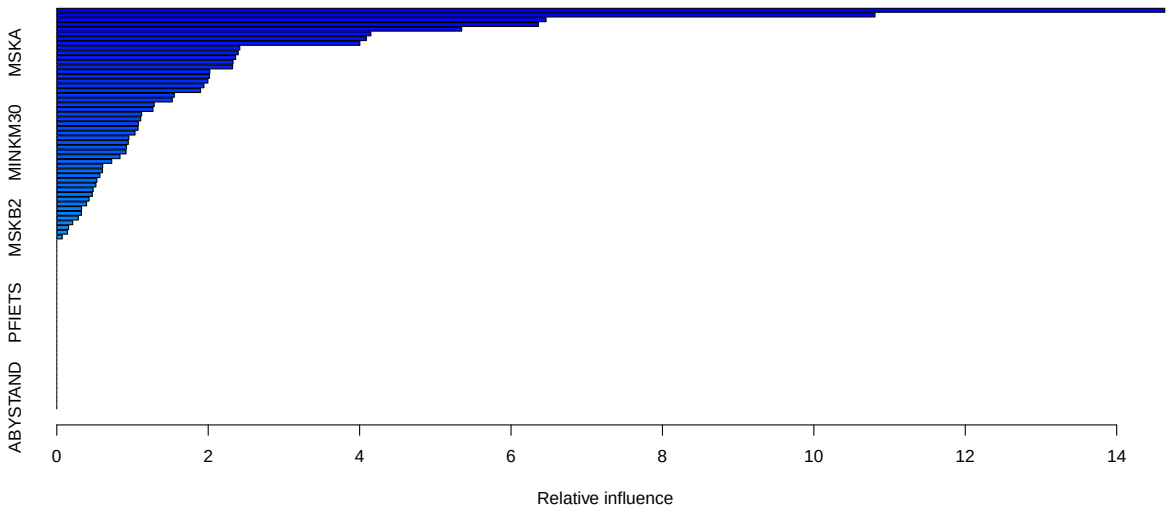
- Training error rate of 0.16. Slightly higher than using the full tree.

Part (k)

- Test error rate of 0.207. Pretty much the same as using the full tree, however, we now have a more interpretable tree.

Question 5

Part (a)



Part (b)

- PPERSAUT and MKOOPKLA appear to be the most important variables.

Part (c)

- Overall, the boosted model makes correct predictions for 92.2% of the observations. The actual number of "No" is 94% and "Yes" is 6%, and so this is an imbalanced dataset. A model simply predicting "No" on each occasion would have made 94% of the predictions correctly. However, in this case we are more interested in predicting those who go on to purchase the insurance. The model predicts "Yes" 158 times, and it is correct on 35 of these predictions - so 22.2% of those predicted to purchase actually do so. This is much better than random guessing (6%).
- Logistic regression predicts "Yes" 408 times, and it is correct on 58 occasions - so 14.2% of those predicted to purchase actually do so. This model is better than random guessing but is worse than the boosted model.