DSI- 06 Homework 7:

Introduction with Statistical Learning with Applications to Python (ISLP) pg. 365

9. This problem involves the OJ data set which is part of the ISLP package.

1. Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
2. Fit a tree to the training data, with Purchase as the response and the other variables as predictors. What is the training error rate?
3. Create a plot of the tree, and interpret the results. How many terminal nodes does the tree have?
4. Use the export\_tree() function to produce a text summary of the fitted tree. Pick one of the terminal nodes, and interpret the information displayed.
5. Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
6. Use cross-validation on the training set in order to determine the optimal tree size.
7. Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.
8. Which tree size corresponds to the lowest cross-validated classification error rate?
9. Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.
10. Compare the training error rates between the pruned and unpruned trees. Which is higher?
11. Compare the test error rates between the pruned and unpruned trees. Which is higher?