HW 1 Report

1. Implementation
   1. The main issue was to decide the data structure to store the decision tree. That structure had to be fairly clean and also had to lend itself to being traversed to produce a forecasted/scored value. Other choices included the following.
      1. What to do when there were no more splits but more than one row of data remained. In this case, I return the average response value.
      2. I found that in some cases, a split would return the same amount of information gain. In these cases, I choose a feature to split on at random.
      3. Need to decide how to measure fit – I’ll be working with the correlation between response and forecast because my algorithm returns fractional values.
   2. Four additional features as follows
      1. The first name – produces some pretty terrible results on test data.
      2. The last name – likewise produces bad results.
      3. If the middle initial is a vowel.
      4. If the name ends in a consonant.
   3. Correlation between response and forecast is 91.1% on training data.
   4. Correlation between response and forecast is 92.2% on test data.
   5. Maximum depth of the tree is 6 (unless we count the root in which case it is 7).
2. Limiting Depth
   1. See the Figure 1 below for accuracy summary. Based on Figure 2, I would allow the tree to go all the way to a depth of six. The tree can only go to a depth of six because there are only six features. Also, the test accuracy doesn’t suffer at all as we go up to six features.
   2. Results are no different than presented in question (1) as the training and test data are the same and results were reported on a max depth of six.
   3. Results are the same as far as I can tell. In general, it would be a good idea to limit the depth of the tree to avoid fitting too much of the noise in the data set. As an example, I did try adding a feature that was just the first given name. Results showed a very good fit on the training data but correlation between response and prediction were very bad on the test data (only about 35% correlation). The given data did not present any options other than returning a tree with max depth equal to the number of features.

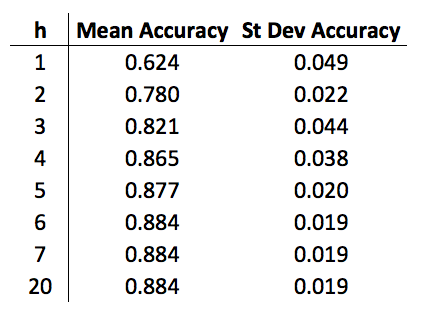


Figure Accuracy Summary

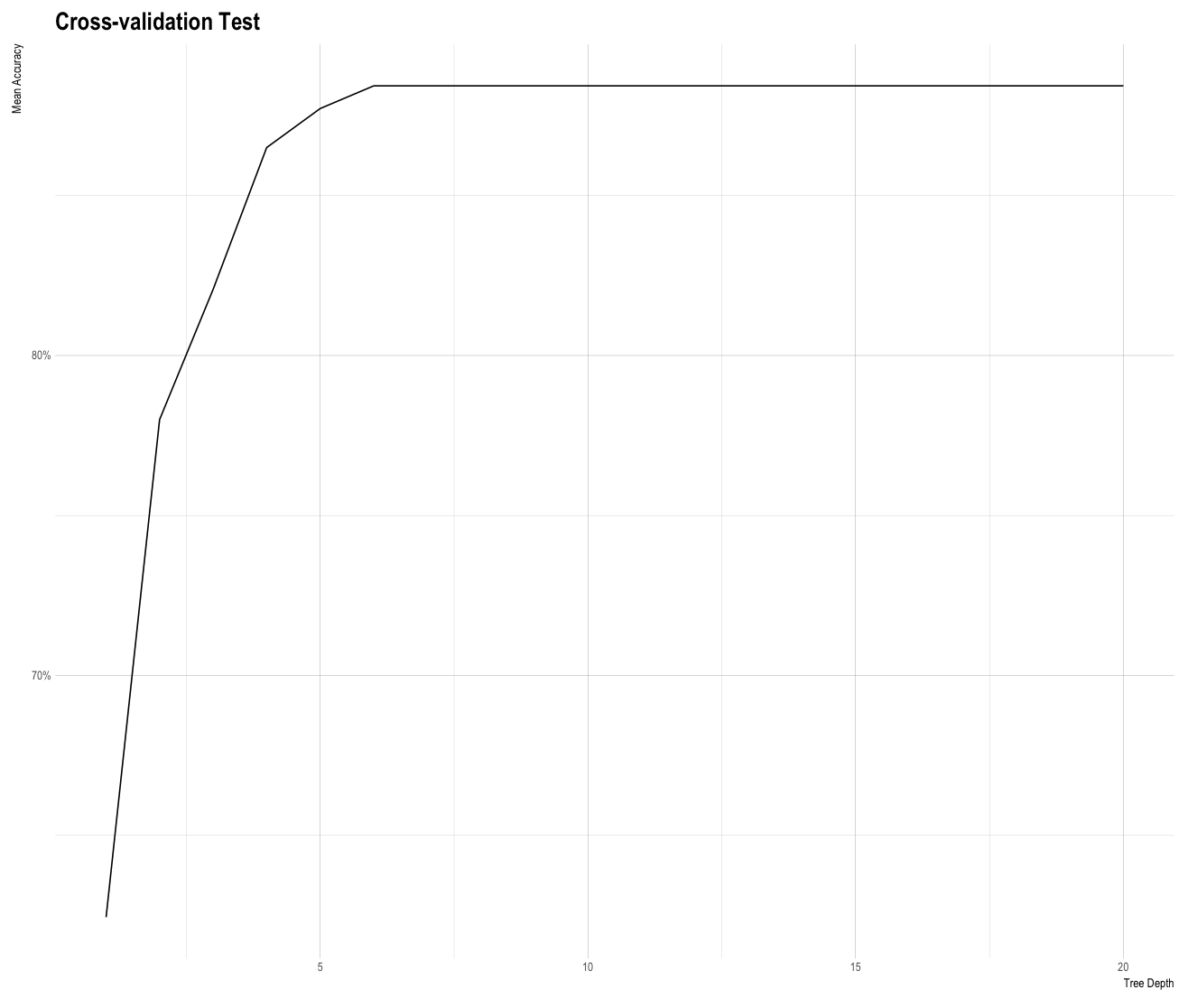


Figure Cross-validation Chart