Classification Tree Assignment. (20 points)

* Ali Ovlia, Webster University, Math/Computer Science Department.
* Textbook

OBJECTIVE

* To understand classification tree
* To build and evaluate classification tree model

SOFTWARE PREREQUISITES

* MS Excel
* R

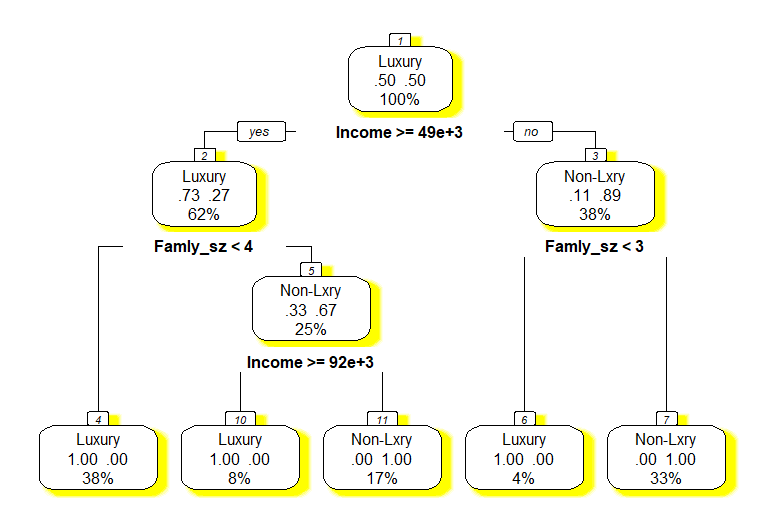
Dataset

1. eBayAuctions.csv

1. Understanding Classification tree

A dataset has 24 records. Calculate the following (10 Points)

* 1. How many observations belong to the Luxury
  2. Gini value for nodes 10 and 11
  3. For node 5
     + 1. Gini
       2. Gini of split
       3. Split Gain
  4. How many observations having salary less than $49,000, with 2 children driving luxury car. Explain.



**Solution:**

1. Given total observations = 24

For each node we need calculate the total observations

Consider the leaf luxury nodes only

At node 4, 38% of 24 observations = 9

At node 10, 8% of 24 observations = 2

At node 6, 4% of 24 observations = 1

Total 12 observation belong to luxury

1. Gini value for nodes 10 and 11

the values of p1=0/2 and p2 =2/2

Gini value for node 10 = 1- [(0/2) ^2+(2/2) ^2] = 0

Gini value for node 11 = 1- [(0/4) ^2+ (4/4) ^2] =0

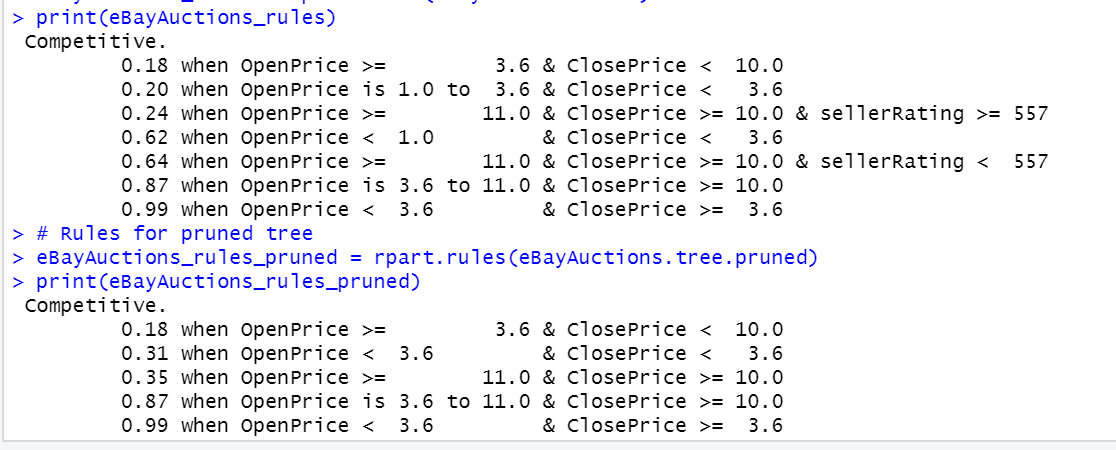
1. Gini of node 5

2. Building Classification tree Model

The file eBayAuctions.csv contains information on 1972 auctions that transacted on eBay.com during May-June 2004. The goal is to use these data to build a model that will classify auctions as competitive or noncompetitive. A competitive auction is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (his/her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price at which the auction closed. The task is to predict whether or not the auction will be competitive. Data Preprocessing. Convert variable Duration into a categorical variable. Split the data into training (60%) and validation (40%) datasets. partition the data into train (60%) and validation (40%) sets.

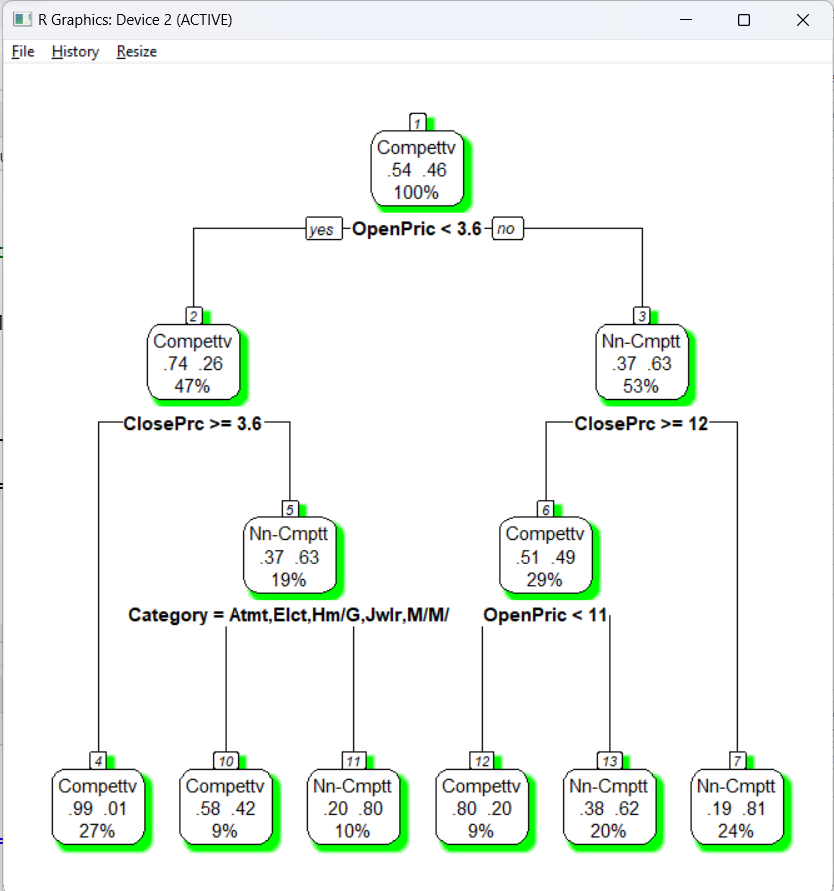
1. Fit a classification tree using all predictors, using the best-pruned tree. To avoid overfitting, set the minimum number of records in a terminal node to 50 (in R: minbucket = 50). Also, set the maximum number of levels to be displayed at seven (in R: maxdepth = 7). Write down the results in terms of rules. (Note: If you had to slightly reduce the number of predictors due to software limitations, or for clarity of presentation, which would be a good variable to choose?)

**Rules Before pruning the tree and after pruning the tree.**

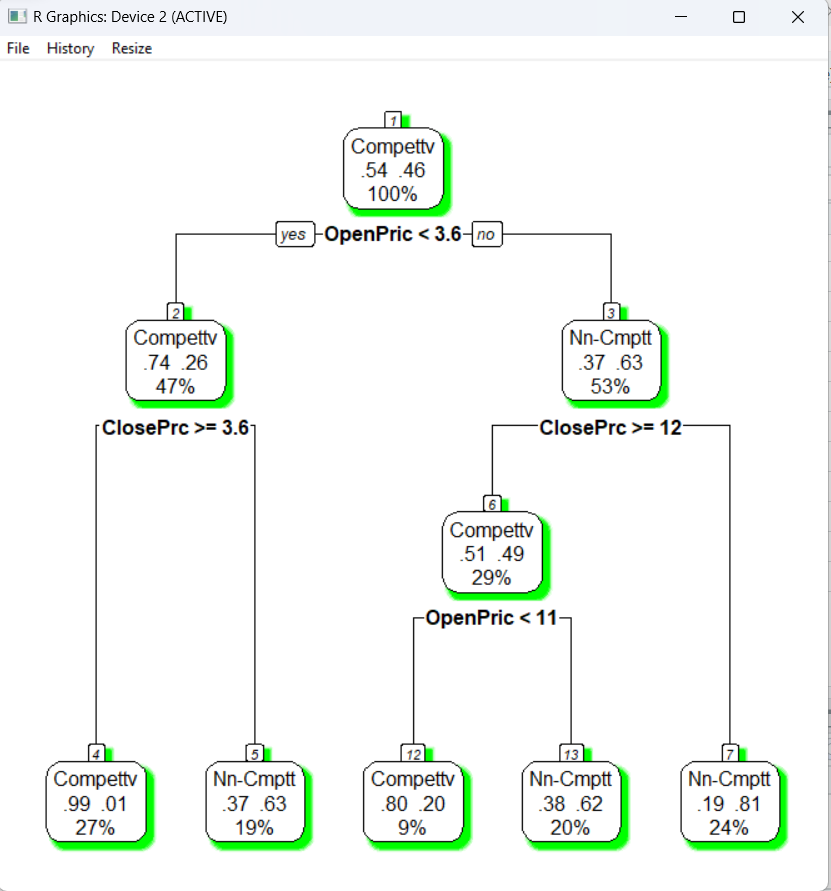


We can remove the endDay because it is least used in classifying the tree.

Before pruning the tree



After pruning the tree

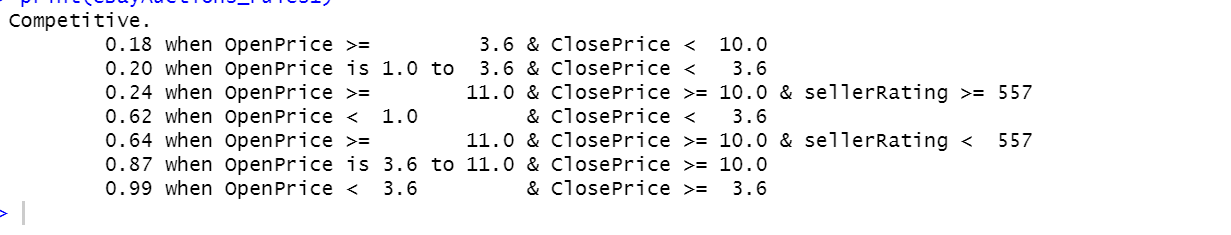


1. **Is this model practical for predicting the outcome of a new auction?**

Yes, we can use this model for predicting the outcome of new auction

1. Describe the interesting and uninteresting information that these rules provide.
2. Fit another classification tree (using the best-pruned tree, with a minimum number of records per terminal node = 50 and maximum allowed number of displayed levels = 7), this time only with predictors that can be used for predicting the outcome of a new auction. Describe the resulting tree in terms of rules. Make sure to report the smallest set of rules required for classification.

**The below rules after building the tree model with only three predictors sellerRating,ClosePrice,OpenPrice**



1. Plot the resulting tree on a scatter plot: Use the two axes for the two best (quantitative) predictors. Each auction will appear as a point, with coordinates corresponding to its values on those two predictors. Use different colors or symbols to separate competitive and noncompetitive auctions. Draw lines (you can sketch these by hand or use R) at the values that create splits. Does this splitting seem reasonable with respect to the meaning of the two predictors? Does it seem to do a good job of separating the two classes?
2. Examine the lift chart and the confusion matrix for the tree. What can you say about the predictive performance of this model?
3. Based on this last tree, what can you conclude from these data about the chances of an auction obtaining at least two bids and its relationship to the auction settings set by the seller (duration, opening price, ending day, currency)? What would you recommend for a seller as the strategy that will most likely lead to a competitive auction?