# Classification Tree

# Carseats Data Set

- A data set containing sales of child car seats at 400 different stores.
- A data set with 400 observations on the following 11 variables.
- The variables are as follows:

1. Sales: Unit sales (in thousands) at each location.
2. CompPrice: Price charged by competitor at each location
3. Income: Community income level (in thousands of dollars)
4. Advertising: Local advertising budget for company at each location (in thousands of dollars)
5. Population: Population size in region (in thousands)

# Carseats Data Set

6. Price: Price company charges for car seats at each site

7. ShelveLoc: A factor with levels “Bad”, “Medium” and “Good” indicating the quality of the shelving location.

8. Age: Average Age of the local population

9. Education: Education level at each location

10. Urban: A factor with levels “No” and “Yes” to indicate whether the store is in an urban or rural location

11. US: A factor with levels “No” and “Yes” to indicate whether the store is in the US or not.

# Carseats Data Set

- We now recode “Sales” as binary variable.
- We create a dummy variable “High”, which takes on a value “Yes” if the sales exceed 8 (in thousands of units) and “No” otherwise.
- We will model “High” with the help of ten predictors.

# Classification Tree

- A classification tree is very similar to a regression tree except that we try to make a prediction for a categorical response rather than continuous one.
- In a regression tree, the predicted response for an observation is given by the average response of the training observations that belong to the same terminal node.
- In a classification tree, we predict that each observation belongs to the most commonly occurring class of the training observations in the region to which it belongs.

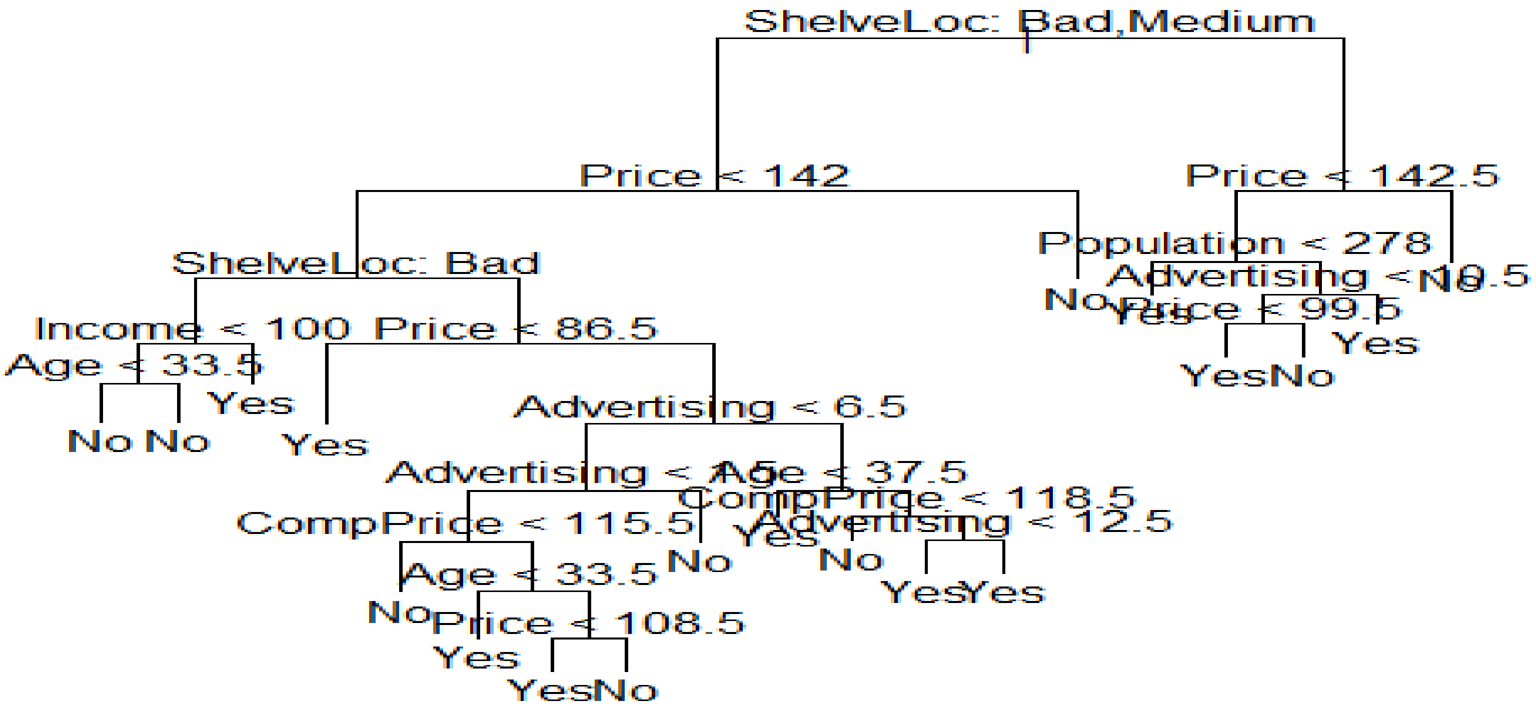
# Classification Tree

- The tree is grown in exactly the same manner as with a regression tree
- However in classification tree, minimizing MSE no longer makes sense.
- A natural alternative is classification error rate.
- The classification error rate is simply the fraction of the training observations in that region that do not belong to the most common class.
- There are several other different criteria available as well, such as the “gini index” and “cross-entropy”.

# Carseats Data Set

- We split the observations into a training data set and a test data set.
- Both the training set and the test set contain 200 observations.
- We next build a tree using the training set, and then evaluate its performance based on the test data.

# Carseats Data Set: Unpruned Tree



# Confusion Matrix based on Test Data

| | True High status | | | |
|---|---|---|---|---|
| Predicted High Status | | No | Yes | Total |
| | No | 88 | 28 | 116 |
| | Yes | 28 | 56 | 84 |
| | Total | 116 | 84 | 200 |

$$Sensitivity = \frac{56}{84} = 66.67\%$$

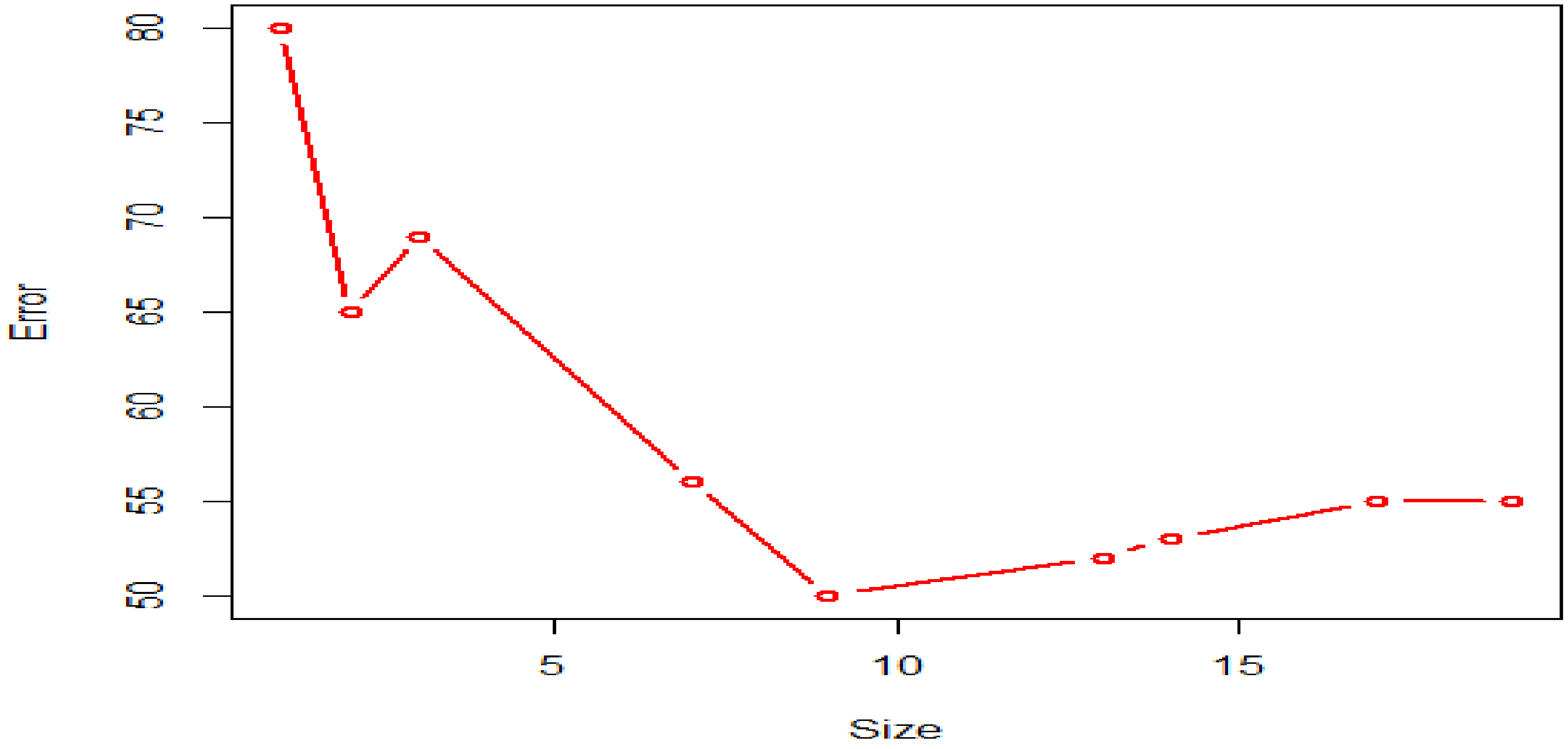
$$Specificity = \frac{88}{116} = 75.86\%$$

$$Total\ Error\ Rate = \frac{56}{200} = 28\%$$

# Cross Validation

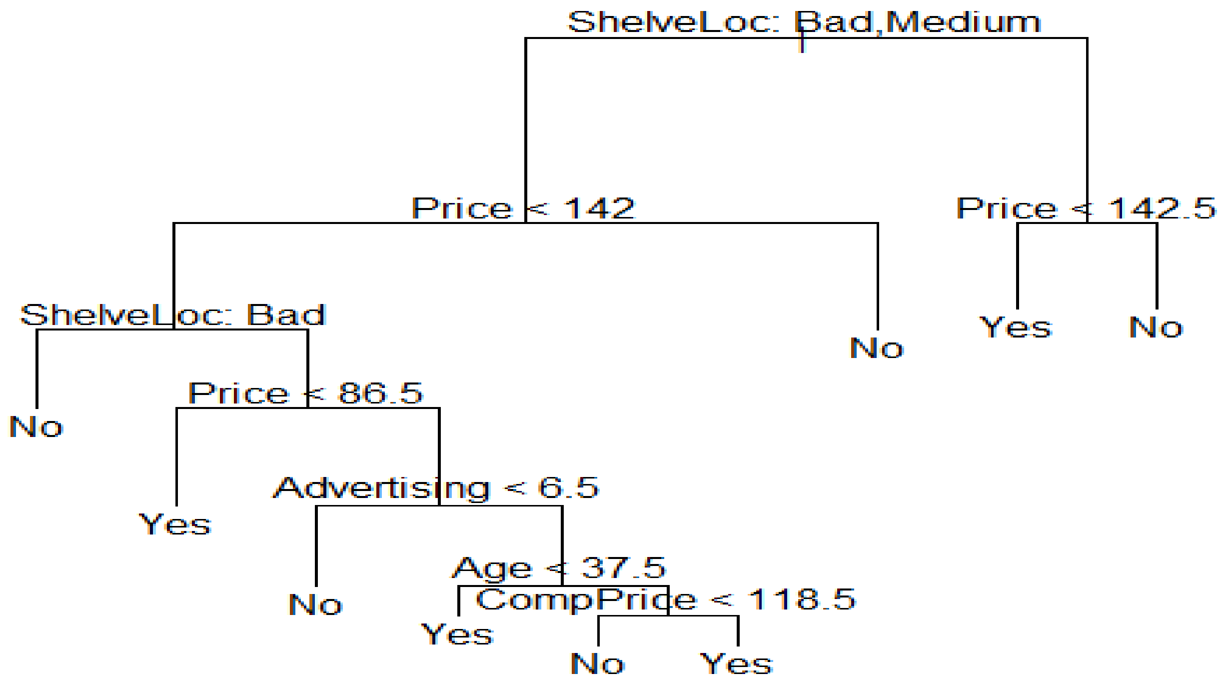
- We now consider whether pruning the tree leads to a better performance.
- We decide the optimal level of tree complexity using cross-validation.

# Cross Validation




We select a tree with 9 terminal nodes.

# Pruned Tree


ShelveLoc: Bad,Medium
Price < 142
Price < 142.5
ShelveLoc: Bad
No
Yes
No
No
Price < 86.5
Yes
Advertising < 6.5
No
Age < 37.5
CompPrice < 118.5
Yes
No
Yes

# Confusion Matrix based on Test Data for Pruned Tree

| | True High status | | | |
|---|---|---|---|---|
| Predicted High Status | | No | Yes | Total |
| | No | 94 | 24 | 118 |
| | Yes | 22 | 60 | 82 |
| | Total | 116 | 84 | 200 |

$$Sensitivity = \frac{60}{84} = 71.43\%$$

$$Specificity = \frac{94}{116} = 81.03\%$$

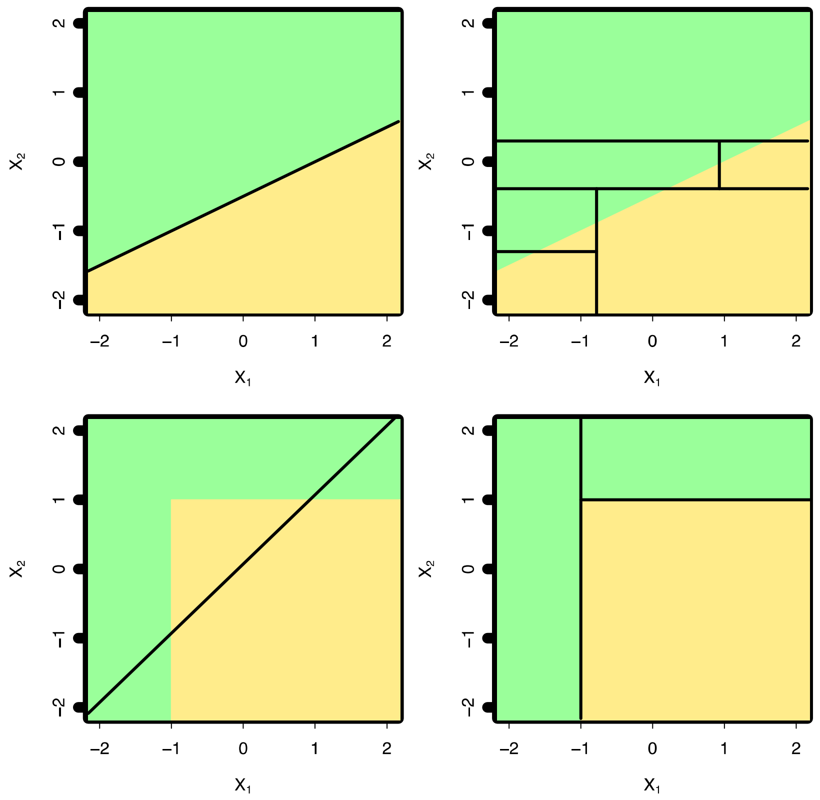
$$Total\ Error\ Rate = \frac{46}{200} = 23\%$$

# Trees vs. Linear Models

- Which model is better?
  - If the relationship between the predictors and response is linear, then classical linear models such as linear regression would outperform regression trees.
  - On the other hand, if the relationship between the predictors is non-linear, then decision trees would outperform classical approaches.

# Trees vs. Linear Model: Classification Example

- Top row: The true decision boundary is linear
  - Left: linear model (Better)
  - Right: decision tree
- Bottom row: The true decision boundary is non-linear
  - Left: linear model
  - Right: decision tree (Better)



# Advantages and Disadvantages of Decision Trees

- Advantages:
  - Trees are very easy to explain to people (even easier than linear regression).
  - Trees can be plotted graphically, and hence can be easily communicated even to a non-expert.
  - They work fine for both classification and regression problems.

- Disadvantages:
  - Trees don’t have the same prediction accuracy as some of the more flexible approaches available in practice.