Easy Tree-sy
Everyday Applications of Decision Trees

CAS Ratemaking and Product Management

San Diego, CA

March 2017

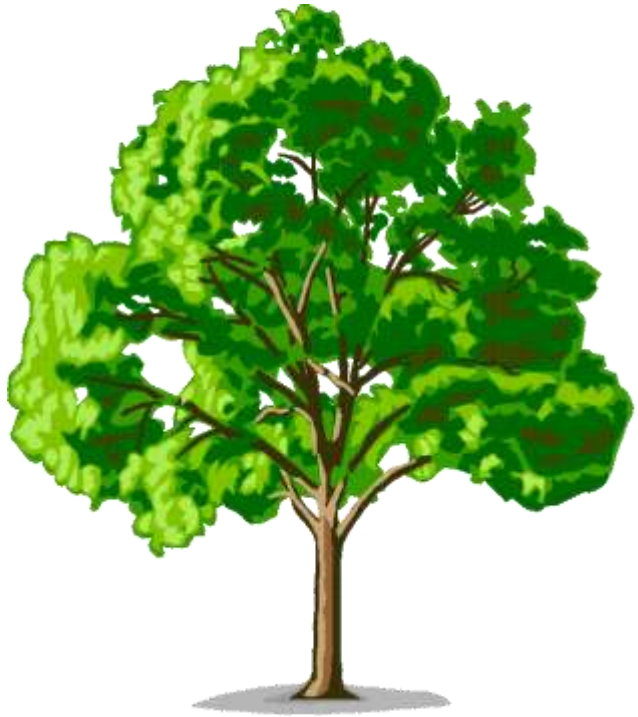
Linda Brobeck <lbrobeck@pinnacleactuaries.com>
Peggy Brinkmann <peggy.brinkmann@milliman.com>
Daniel Murphy <dmurphy@trinostics.com>

Antitrust Statement

- **The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.**
- **Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.**
- **It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.**

Introductions and Agenda

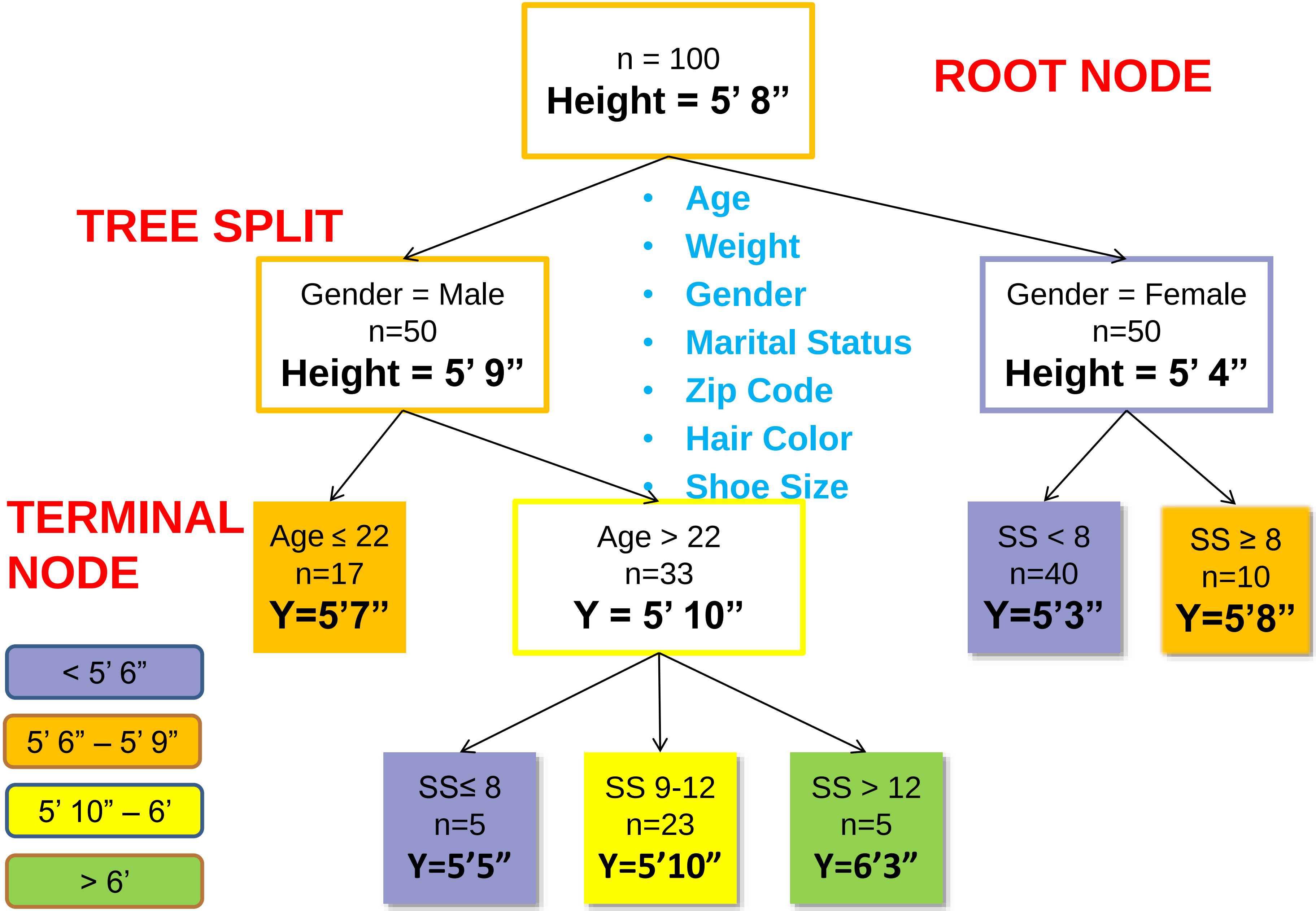
- Decision Tree Basics
  - An Example
  - Terminology
  - Objectives/Theory
- Applications of Decision Trees
- Case Study using Free Software
- Customization

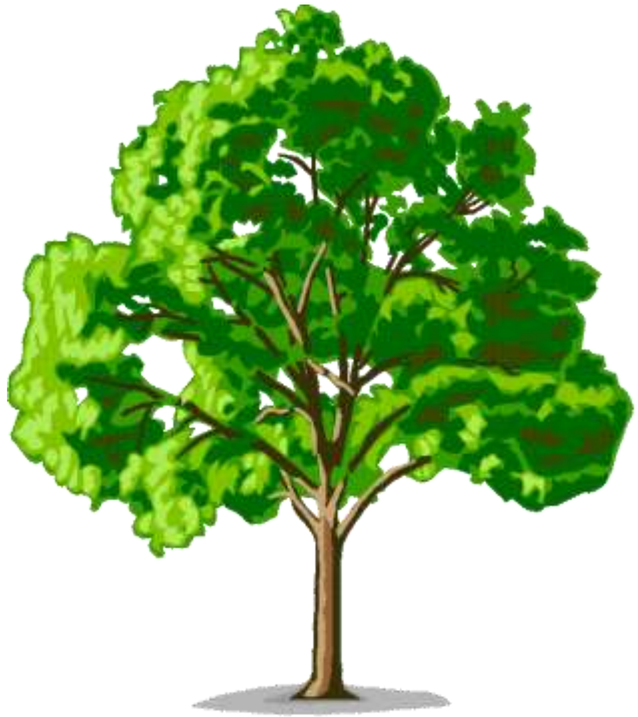


An Example

Estimate the height of an adult, given the following information:

- Age
- Weight
- Gender
- Marital Status
- Zip Code
- Hair Color
- Shoe Size


n = 100
Height = 5’ 8”
ROOT NODE
TREE SPLIT
Age
Weight
Gender
Marital Status
Zip Code
Hair Color
Shoe Size
Gender = Male
n=50
Height = 5’ 9”
Gender = Female
n=50
Height = 5’ 4”
TERMINAL NODE
Age ≤ 22
n=17
Y=5’7”
Age > 22
n=33
Y = 5’ 10”
SS < 8
n=40
Y=5’3”
SS ≥ 8
n=10
Y=5’8”
< 5’ 6”
5’ 6” – 5’ 9”
5’ 10” – 6’
> 6’
SS≤ 8
n=5
Y=5’5”
SS 9-12
n=23
Y=5’10”
SS > 12
n=5
Y=6’3”



Terminology

Target Response, Predicted Outcome, Dependent Variable

Y: Height

Explanatory, Predictor, Independent Variables

X_i: Age, Gender, Marital Status
Zip Code, Hair Color, Shoe Size

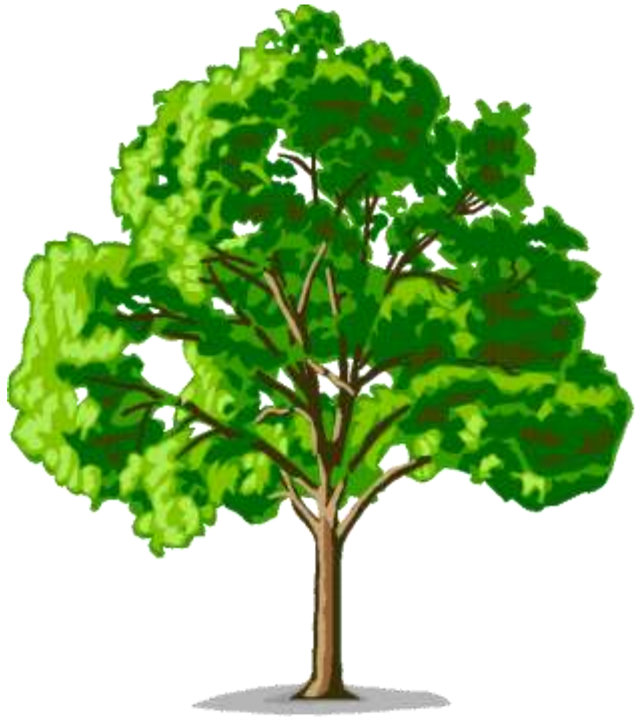
If the Target Variable is:

Categorical

➔ **Classification Tree**

Continuous

➔ **Regression Tree**



Objectives/Theory

Two Objectives:

Purity

➔ **Measure of variation**

Parsimony

➔ **Desire for simple**

The Process

- **Splitting Procedure**

  The domain space of explanatory variables $X_1,\ldots X_{\mathbf{n}}$ is split into two subsets where observed values in X_j belong to one of the subsets

  i.e. **< *s* or >= *s*** *OR* ***s_1=male s_2=female***

- **Improvement Value**

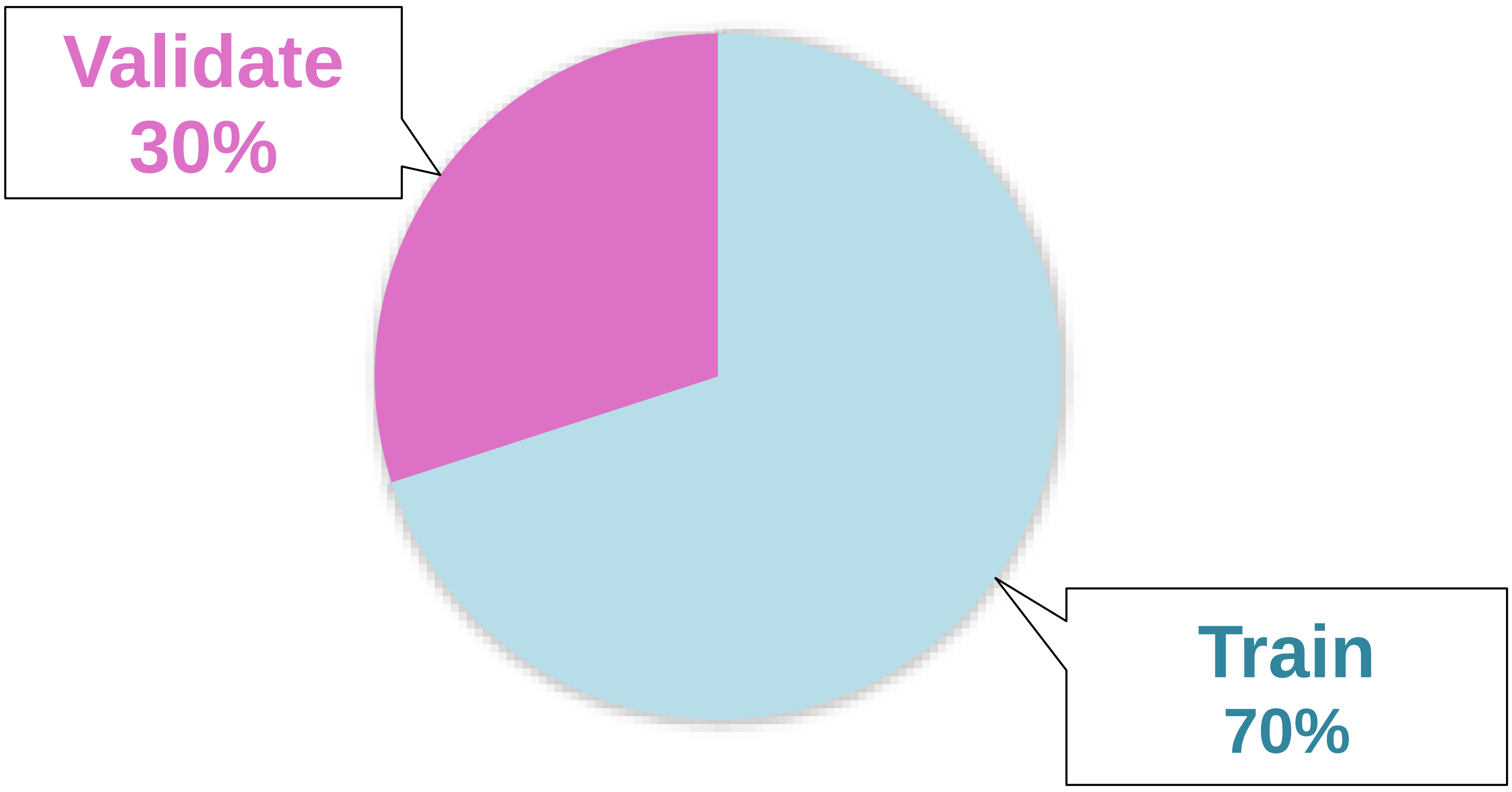
  The dimensions ***j*** and ***s*** above are chosen to minimize the error in the prediction among all such binary (two-leveled) trees. Process is iterated.

Stopping Criterion

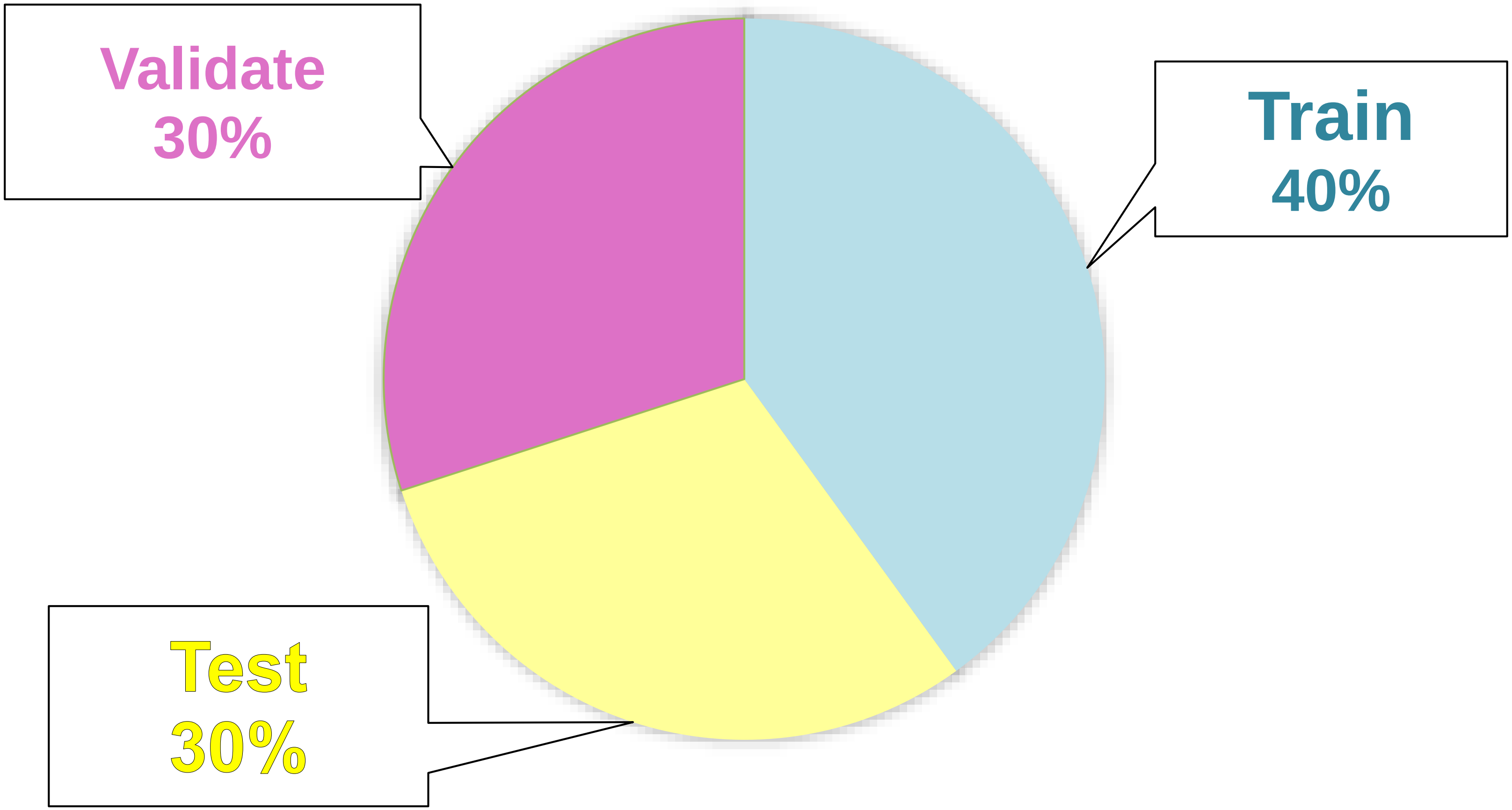
- No stopping criterion
- Minimum leaf (node) size
- Maximum number of levels or splits
- Let data determine the stopping criterion (see Appendix)

Validating Results - Avoiding Over Fit

The validation dataset ensures a way to accurately measure your model's performance.

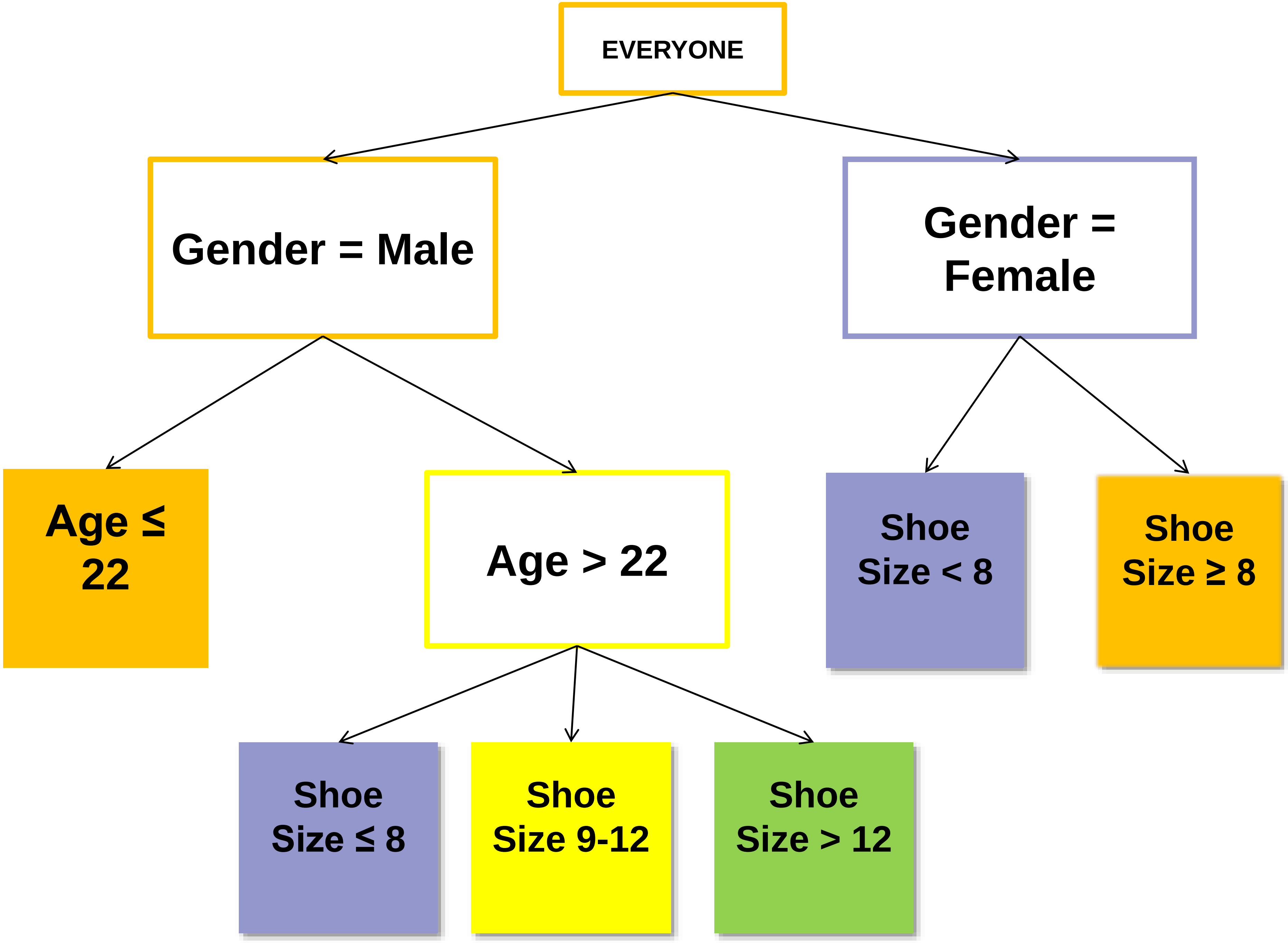


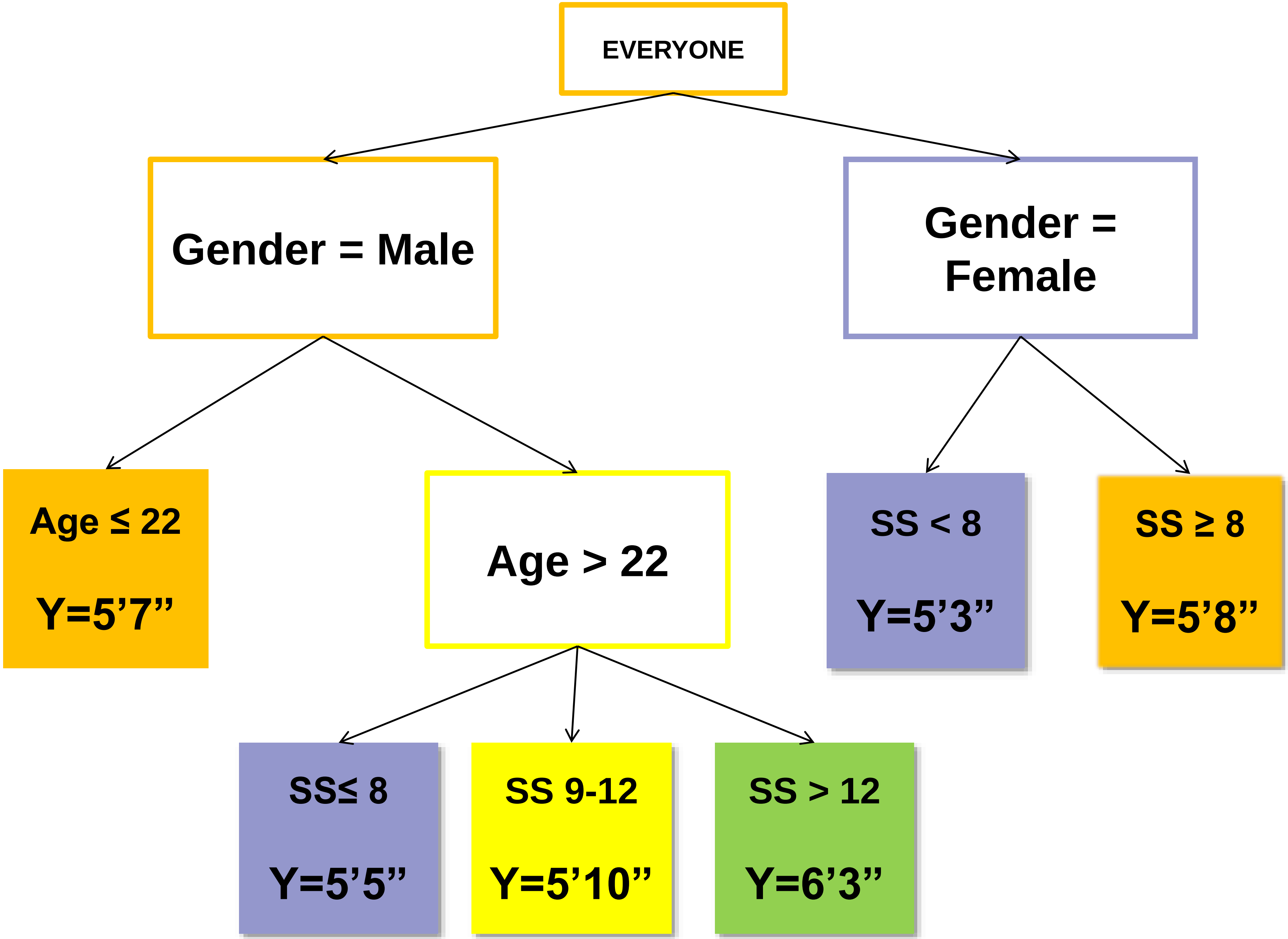
Validating Results - Avoiding Over Fit



Large datasets can be split into 3 unique subsets.

If there is time…


EVERYONE
Gender = Male
Gender = Female
Age ≤ 22
Age > 22
Shoe Size < 8
Shoe Size ≥ 8
Shoe Size ≤ 8
Shoe Size 9-12
Shoe Size > 12


EVERYONE
Gender = Male
Gender = Female
Age ≤ 22
Y=5’7”
Age > 22
SS < 8
Y=5’3”
SS ≥ 8
Y=5’8”
SS≤ 8
Y=5’5”
SS 9-12
Y=5’10”
SS > 12
Y=6’3”

Appendix

Stopping Criterion – Regression Trees

- To begin, we need to define an error function $E()$ on any leaf of a tree. Think of $E()$ as a measure of how far the predicted are from observed
- Then, for a fixed $\alpha > 0$, find that tree T that minimizes

$$C_\alpha(|T|) \quad = \quad \sum_{k=1}^{|T|} E(L_k) + \alpha|T|$$

- $E(L_k)$ is the error contributed by the kth leaf and α is a parameter that rewards parsimony
- One can see that minimizing the cost complexity criterion $C_\alpha()$ requires a balance between predictive power and parsimony to be struck

Stopping Criterion – Regression Trees (cont.)

- Define
  1. $|L_k| = \sum_{\substack{i=1 \\ \mathbf{x}_i \in L_k}}^{K} w_i$
  2. $\bar{y}_k = \frac{1}{|L_k|} \sum_{\substack{i=1 \\ \mathbf{x}_i \in L_k}}^{K} w_i y_i$
- A standard choice for $E()$ is

$$E(L_k) = \sum_{\mathbf{x}_i \in L_k} w_i (y_i - \bar{y}_k)^2$$

- There are other standard functions for $E()$, for example
  1. $E(L_k) = \sum_{\mathbf{x}_i \in L_k} w_i |y_i - \bar{y}_k|$
  2. $E(L_k) = \sum_{\mathbf{x}_i \in L_k} w_i |y_i - \bar{y}_k|^p$ for $1 < p < 2$
- User may have choice on what functional form $E()$ may take depending on the software

Bibliography

- Hastie, T. et al. (2011) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*, Springer, New York.