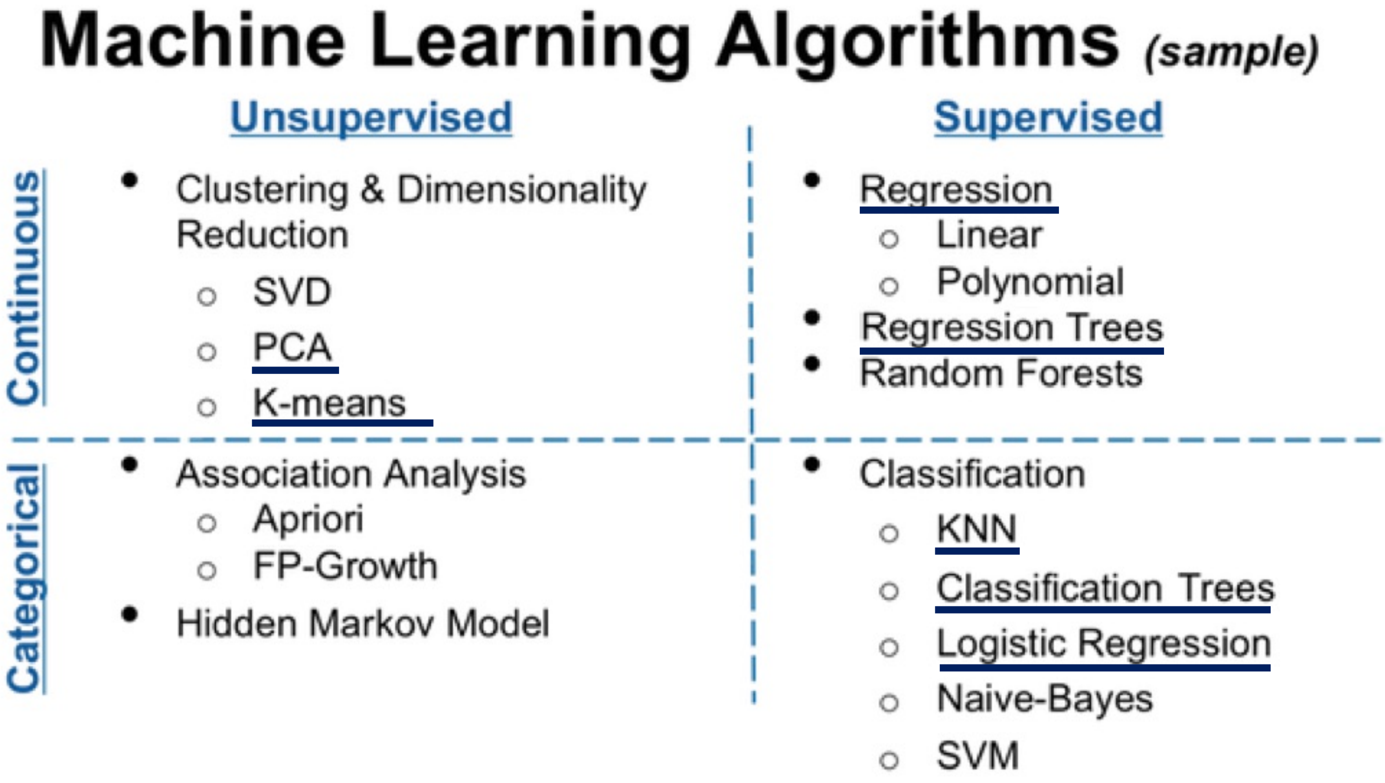
THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Decision Tress

11/27/2018

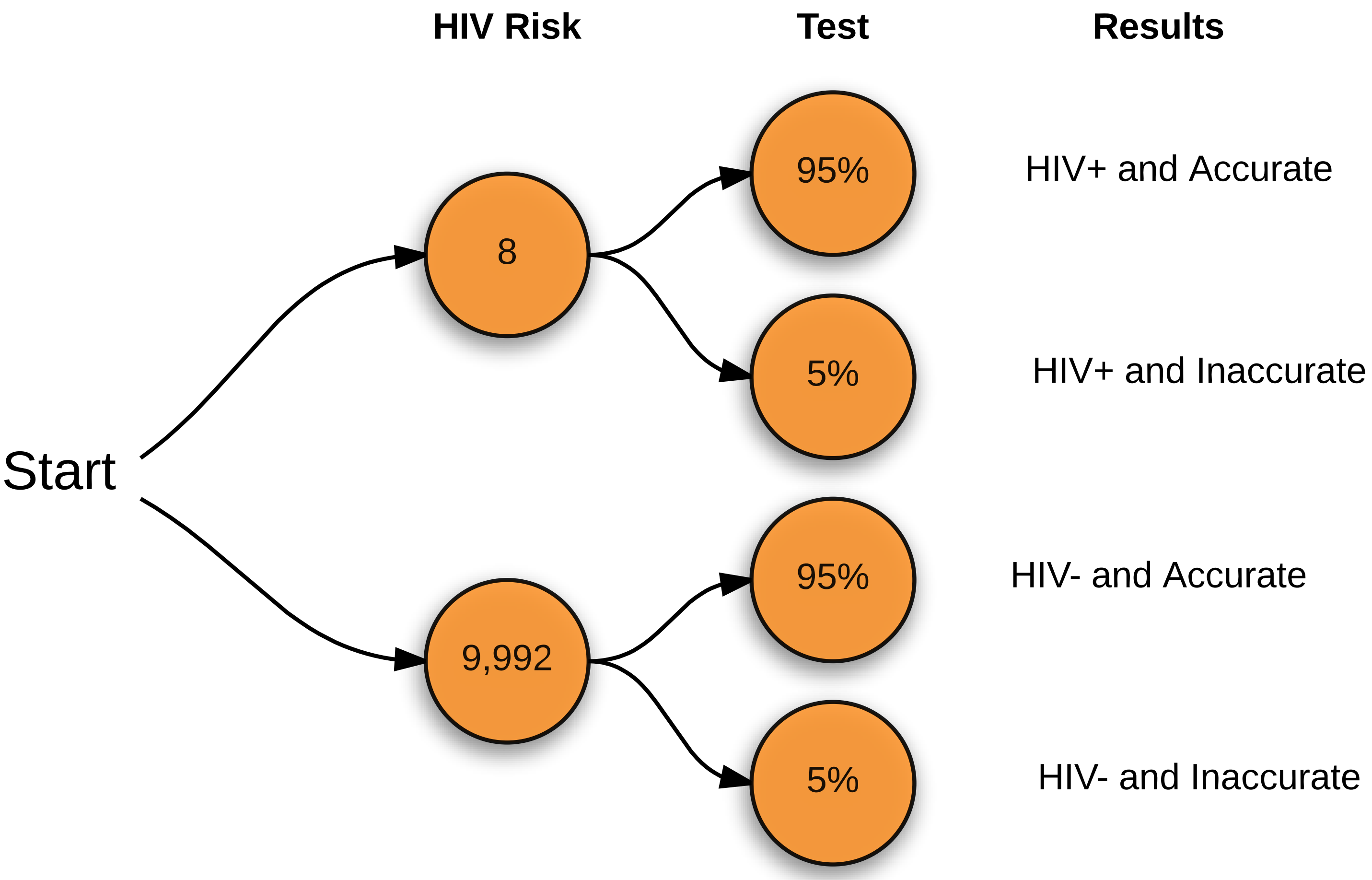
Trees



Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Regression Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Classification Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Decision Trees in Statistics



Trees

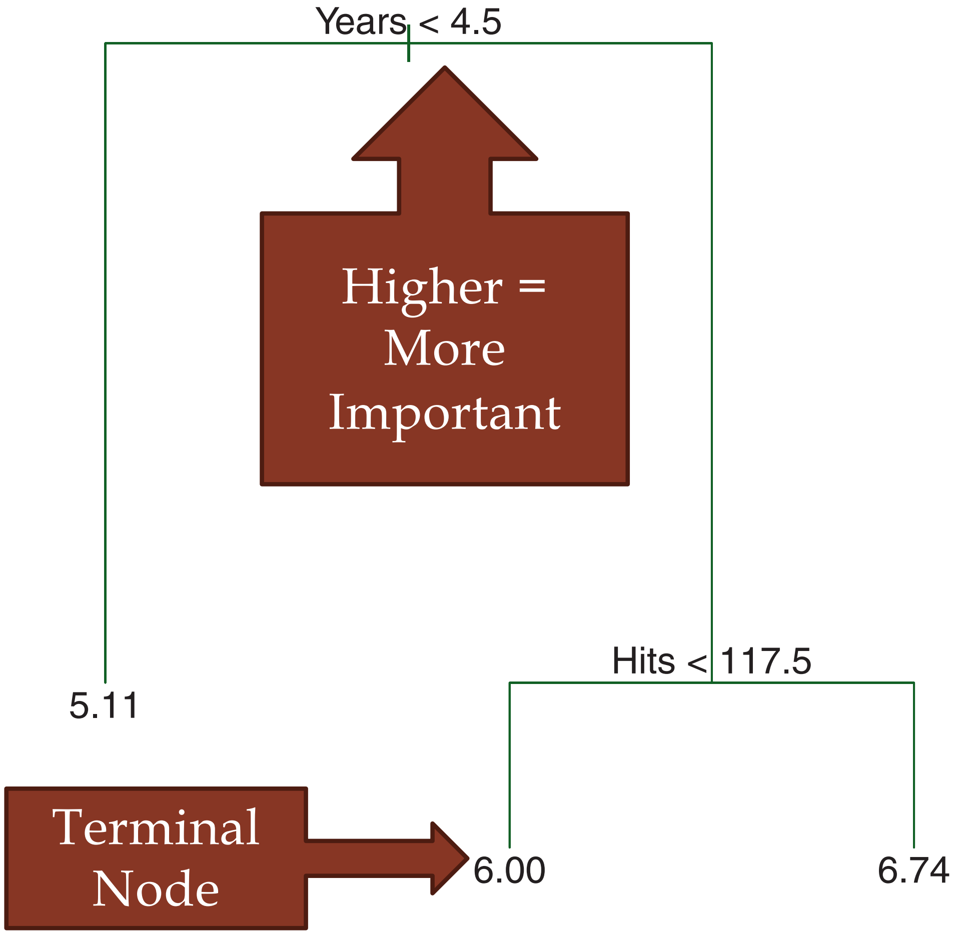
- In many situations, decisions are made by hierarchical rules. (Is your team going to playoff?)
- Which variable is most important?
  - ✓ Decision trees can give good hints
- Categorical dependent variable: Classification tree
- Quantitative dependent variable: Regression tree
- Predictors: can mix categorical & quantitative
- Trees can be tuned and pruned.
- ***Random Forests*** – with many many trees, averaged, tuned, randomized (cross-validation etc)

Regression Tree Example

- Predict baseball player salary
  - Example by Rebecca C. Steorts, Duke University
- Use **binary recursive splitting**
- The predictor solution space is divided into regions
- Branches are created by **cutpoints** that divide the space into regions
- Each region is defined by its average values of the variables (indept and/or dept), c.f. K-means
- Cutpoints are determined by minimizing MSE of both new regions.
- Cuts continue recursively

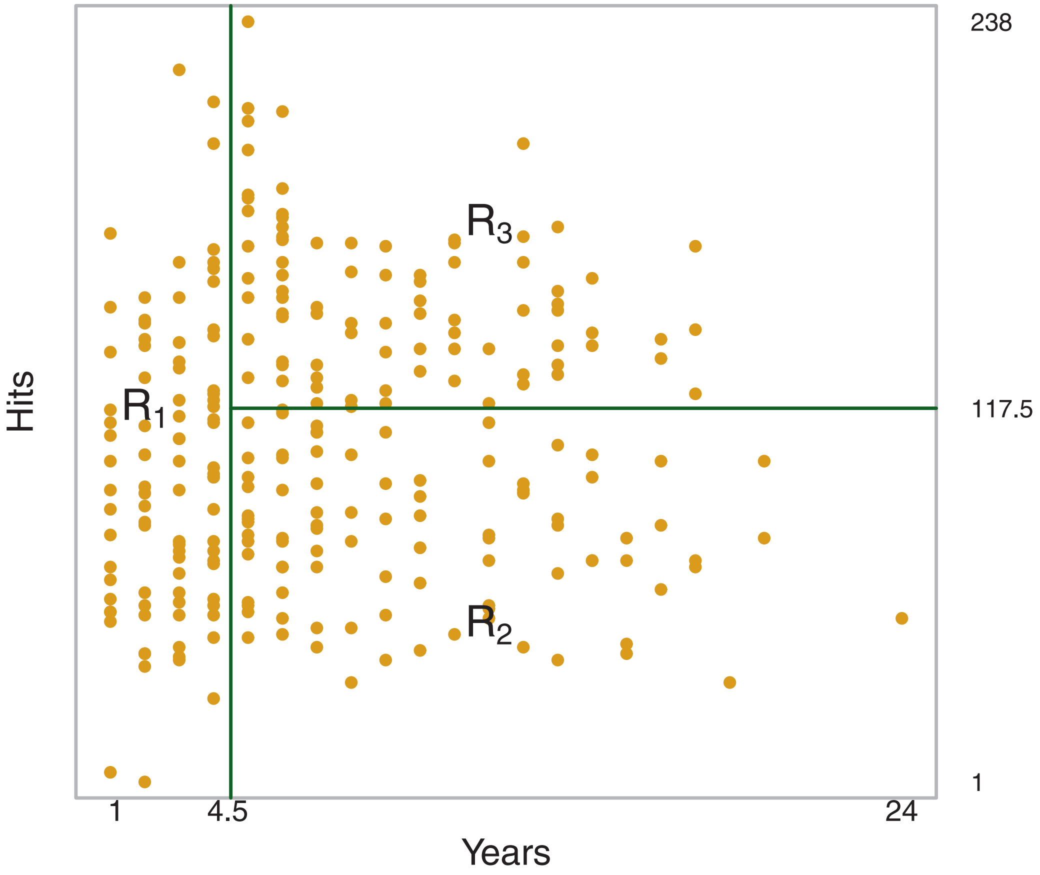
Regression Tree Example

Recursive Binary Splitting



Higher = More Important

Terminal Node



Regression Tree Example

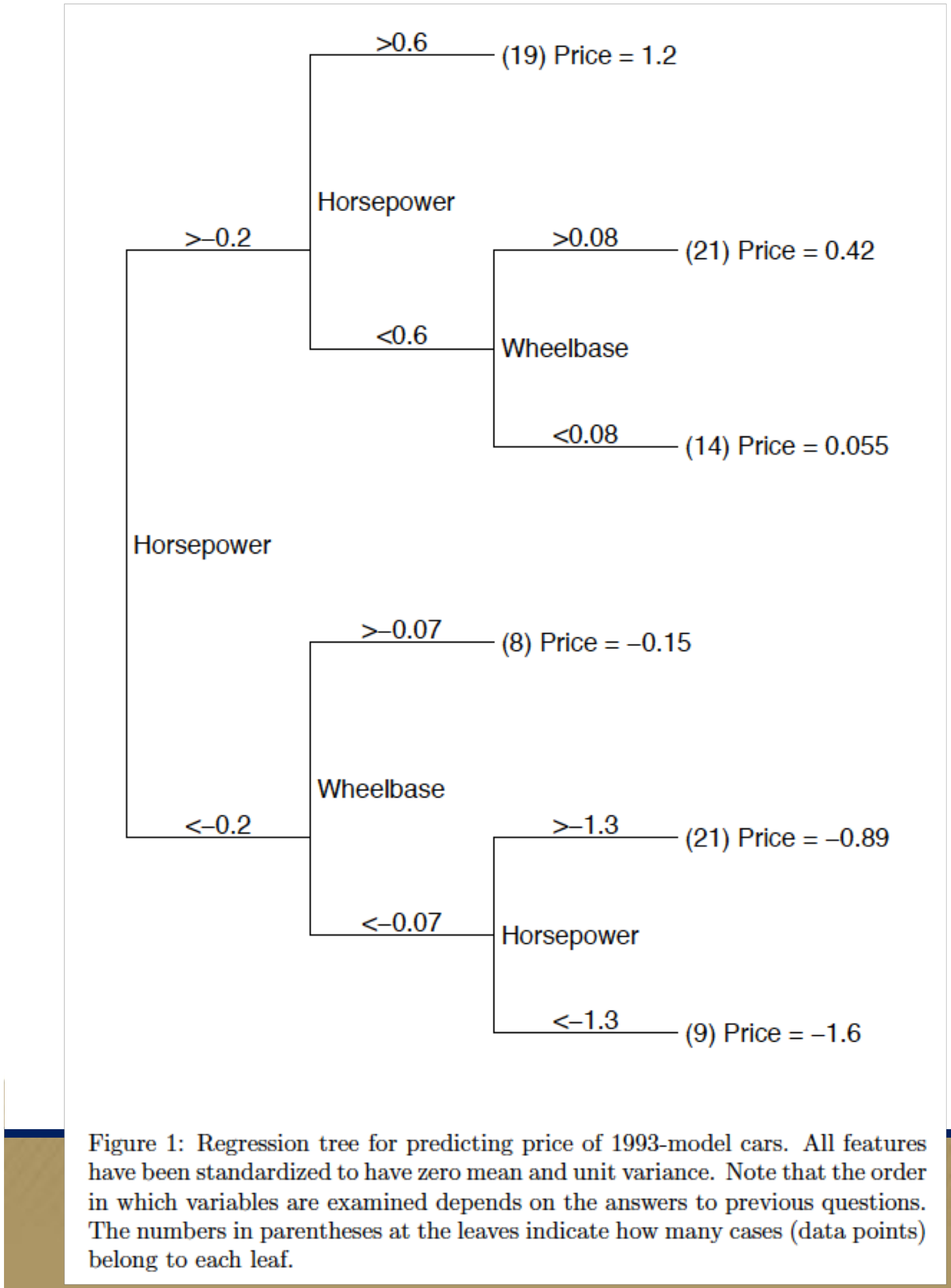


Figure 1: Regression tree for predicting price of 1993-model cars. All features have been standardized to have zero mean and unit variance. Note that the order in which variables are examined depends on the answers to previous questions. The numbers in parentheses at the leaves indicate how many cases (data points) belong to each leaf.

Predicting price of 1993-model cars.

All variables standardized.

Note the order of variables examined depends previous questions.

The numbers in parentheses at the leaves indicate how many cases (data points) belong to each leaf.

Source: CMU Data Mining lectures http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

Regression Tree Example

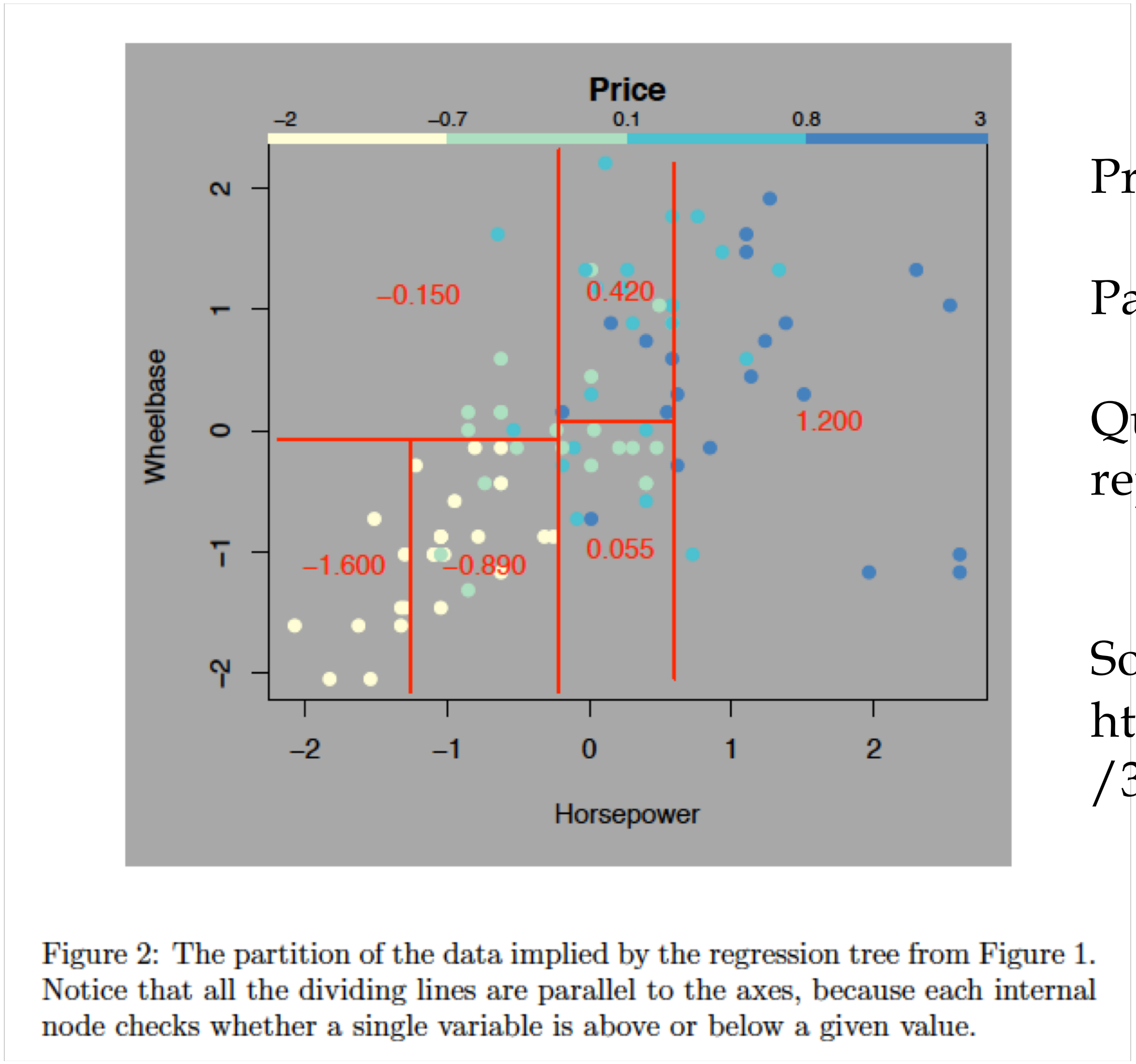


Figure 2: The partition of the data implied by the regression tree from Figure 1. Notice that all the dividing lines are parallel to the axes, because each internal node checks whether a single variable is above or below a given value.

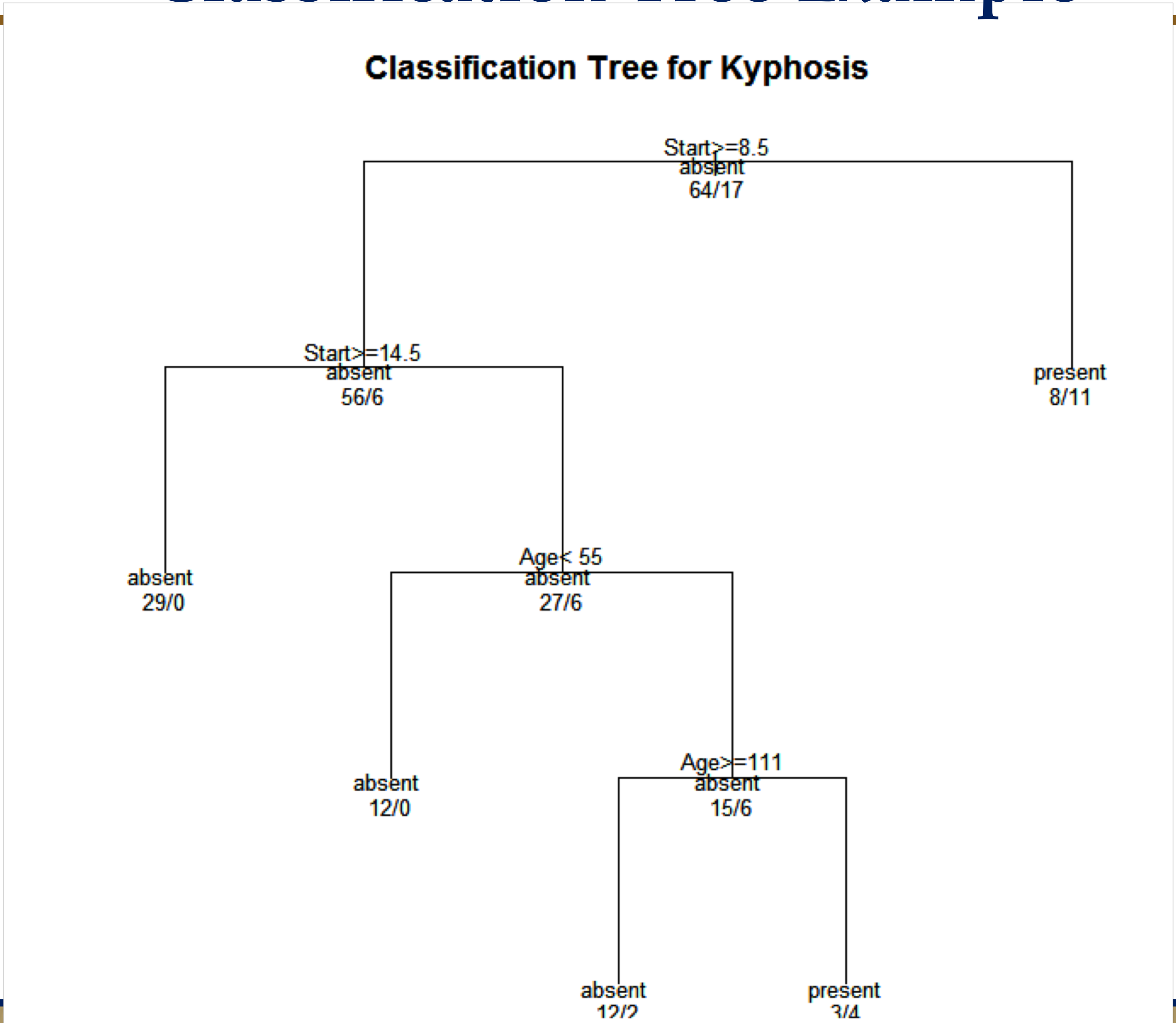
Predicting price of 1993-model cars.

Partitions of solution space

Quantitative response variable represented by color.

Source: CMU Data Mining lectures http://www.stat.cmu.edu/~cshalizi/350-2006/lecture-10.pdf

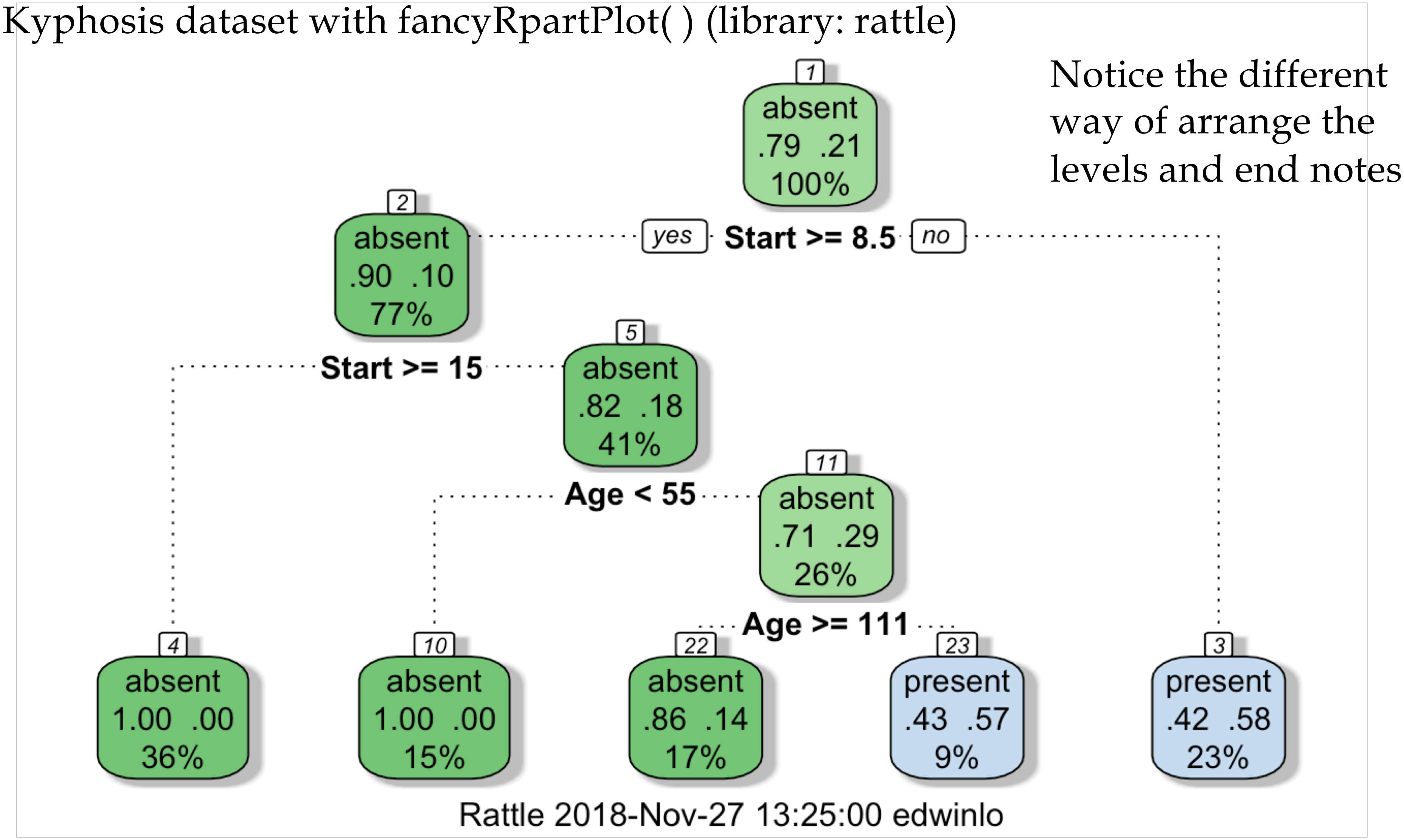
Classification Tree Example



Kyphosis dataset
Source: https://www.statmethods.net/advstats/cart.html

Classification Tree Example

Kyphosis dataset with fancyRpartPlot( ) (library: rattle)



Rattle 2018-Nov-27 13:25:00 edwinlo

Notice the different way of arrange the levels and end notes.

Random Forest

- Forest contains many many trees
- Each tree from different sample
- Each split chosen from random sample of possible predictors
- Usually $m = \sqrt{p}$ possible predictors
- Useful when many independent variables or one strong and many weak predictors
- If predicting within the space of the sample, decision tree might show better accuracy
- When predicting unexpected (untrained) data, random forest usually much better