Easy Tree-sy
An Overview of Decision Trees

Underwriting Collaboration Seminar

New Orleans, LA

June 26, 2018

Linda Brobeck Lbrobeck@Pinnacleactuaries.com

Elaine George Elaine.George@Chubb.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.**

1

Introductions and Agenda

LEARNING OBJECTIVES

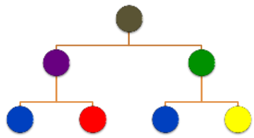
1. Explain the fundamentals of decision trees
2. Evaluate and decide when to apply decision trees to a underwriting business problem
3. Design a decision tree predictive modeling analysis

AGENDA

- Decision Tree Basics via an Example
- Underwriting Specific Applications of Decision Trees
- Predictive Modeling Concepts

2

Two Objectives:



Purity

➔ **Measure of variation**

Parsimony

➔ **Desire for simple**

3

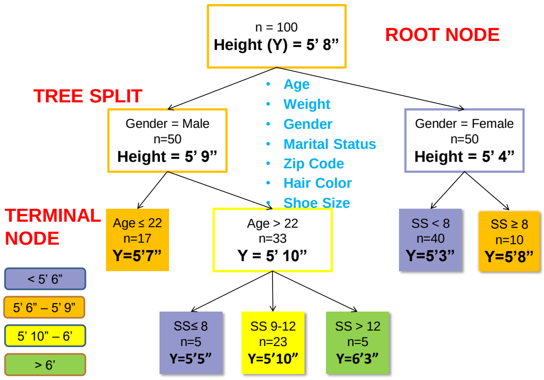
An Example

4

Estimate(predict) the height of an adult, given the following information:

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

5



6

Terminology

7

Target Response,
Predicted Outcome,
Dependent Variable

Y: Height

Explanatory/Independent Variables, Predictors, Features

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

8

If the Target Variable is:

Categorical

➔ **Classification Tree**

Continuous

➔ **Regression Tree**

9

UW Example

10

First Determine:

What is the business issue?

Renewal Guidelines
Non-Sufficient Funds Prevalence
ACV vs Replacement Cost
Inspection Ordering
Allocation of Resources

11

Business issue?

Renewal Guidelines

Predicted Outcome (Target)?

Commercial Auto Loss Ratio

Currently 80% overall LR

12

Explanatory/Independent Variables, Predictors

Vehicle Type
Principal Garaging
Business Type
Annual Mileage
Driving Record
Claim History

13

Individual Variable Analysis

Overall Loss Ratio = 80%

Vehicle Type (VT)

PPA = 71%; Heavy = 86%; Light = 82%

Principal Garaging (PG)

Urban = 80%; Rural = 81%; Suburb = 80%

Business Type (BT)

10 Classifications LR ranges 78% - 89%

14

Individual Variable Analysis

Overall Loss Ratio = 80%

Annual Mileage (AM)

Short = 81%; Average = 80%; Long = 80%

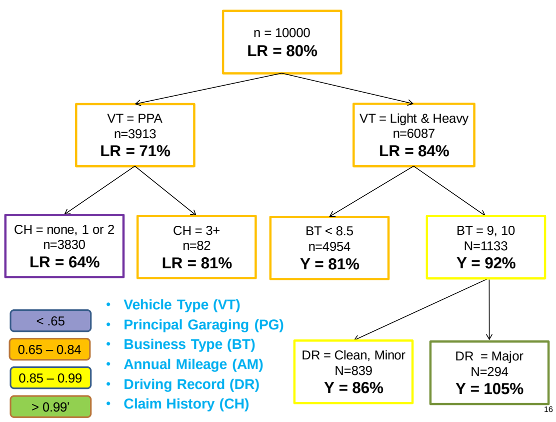
Driving Record (DR)

Clean = 78%; Minor = 78%; Major = 84%

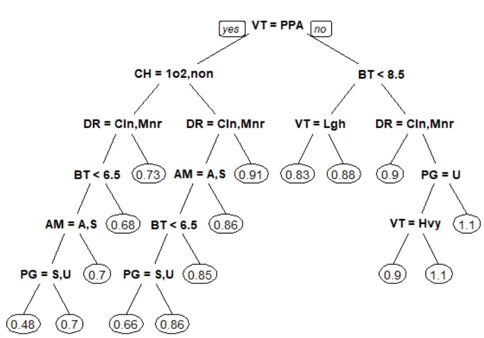
Claim History (CH)

None = 81%; 1-2 = 77%; 3+ = 83%

15



16



17

Theory

18

The Process

- **Splitting Procedure**
  The domain space of explanatory variables $X_1,\ldots X_n$ is split into two subsets where observed values in X_j belong to one of the subsets
  i.e. ***< s*** **or** ***>= s*** *OR* ***s₁=male s₂=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.

19

Measures for Splitting Criteria

Significance — Measures Independence
- Numeric purity
- p-values of Chi-square variance reduction

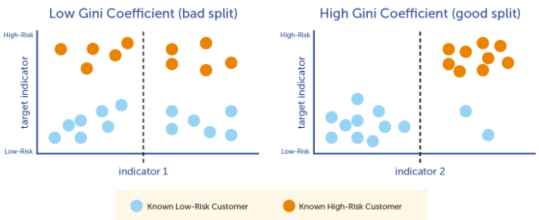
Entropy — Measures Disorder
- Categorical
- Measures pureness of the level

Gain Ratio — Measures Gain in Intrinsic Information
- Information Gain = Entropy (parent) – Weighted sum of Entropy (children)
- Penalizes large values/splits

Gini — Measures Misclassification
- Max = 1 – (1 / # of classes)
- Minimum = 0 (all records belong to one class)

20

Gini Coefficient



21

Stopping Criterion

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

22

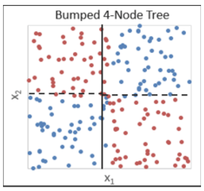
Advantages

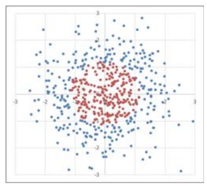
- Non-parametric
- Simple to understand / Easy to interpret
- Automatic variable and interaction selection
- Handles missing values and outliers

23

Limitations

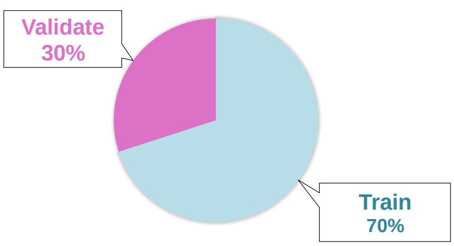
- Over-fit and Instability
- Some relationships difficult to find



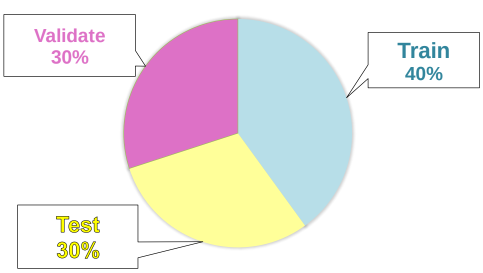


Validating Results - Avoiding Over Fit

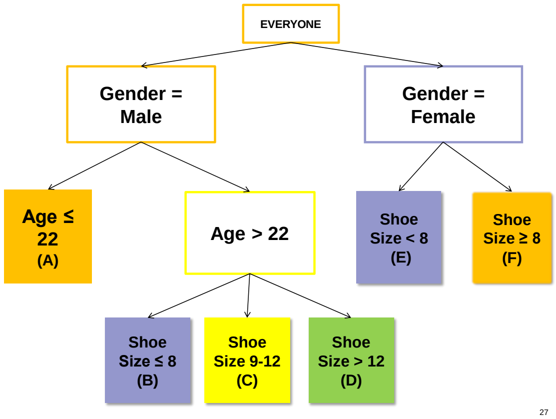
The validation dataset ensures a way to accurately measure your model's performance.

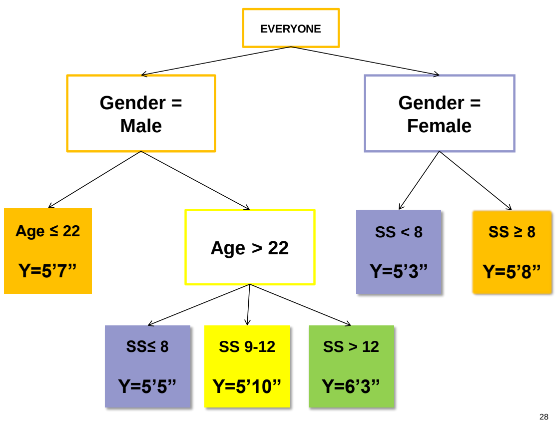


Validating Results - Avoiding Over Fit



Large datasets can be split into 3 unique subsets.





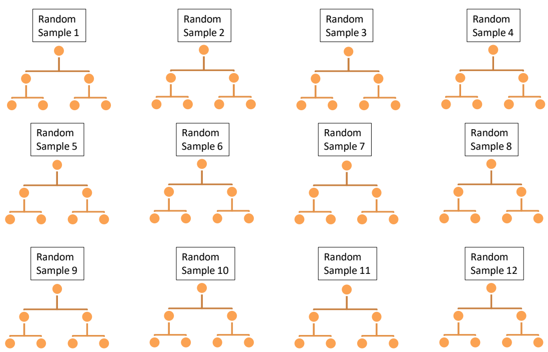
If there is time…

Ensembles

Combine many weak classifiers in order to strengthen the overall result

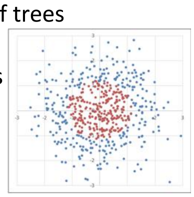
- **Bagging (Bootstrap Aggregating)**
- **Boosting**
- **Stacked Generalization (Blending)**

Random Forests



Boosting

- **Gradient Boosting**
  - Sequential based on residual of prior tree
- **Multiplicative Boosted Trees**
  - Multiplicative residuals
  - Multiplicative combining of trees
- **AdaBoost**
  - Iteratively changes weights of training observations based on errors of previous prediction



Appendix

R

```
Library(rpart)

Library(rpart.plot)

tree <- rpart(LR~VT+PG+BT+AM+DR+CH, data=dataset, weights =
WP)

PRP(tree)
```

34

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

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

- $E(L_k)$ 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

35

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

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

- 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

36

Bibliography

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

37