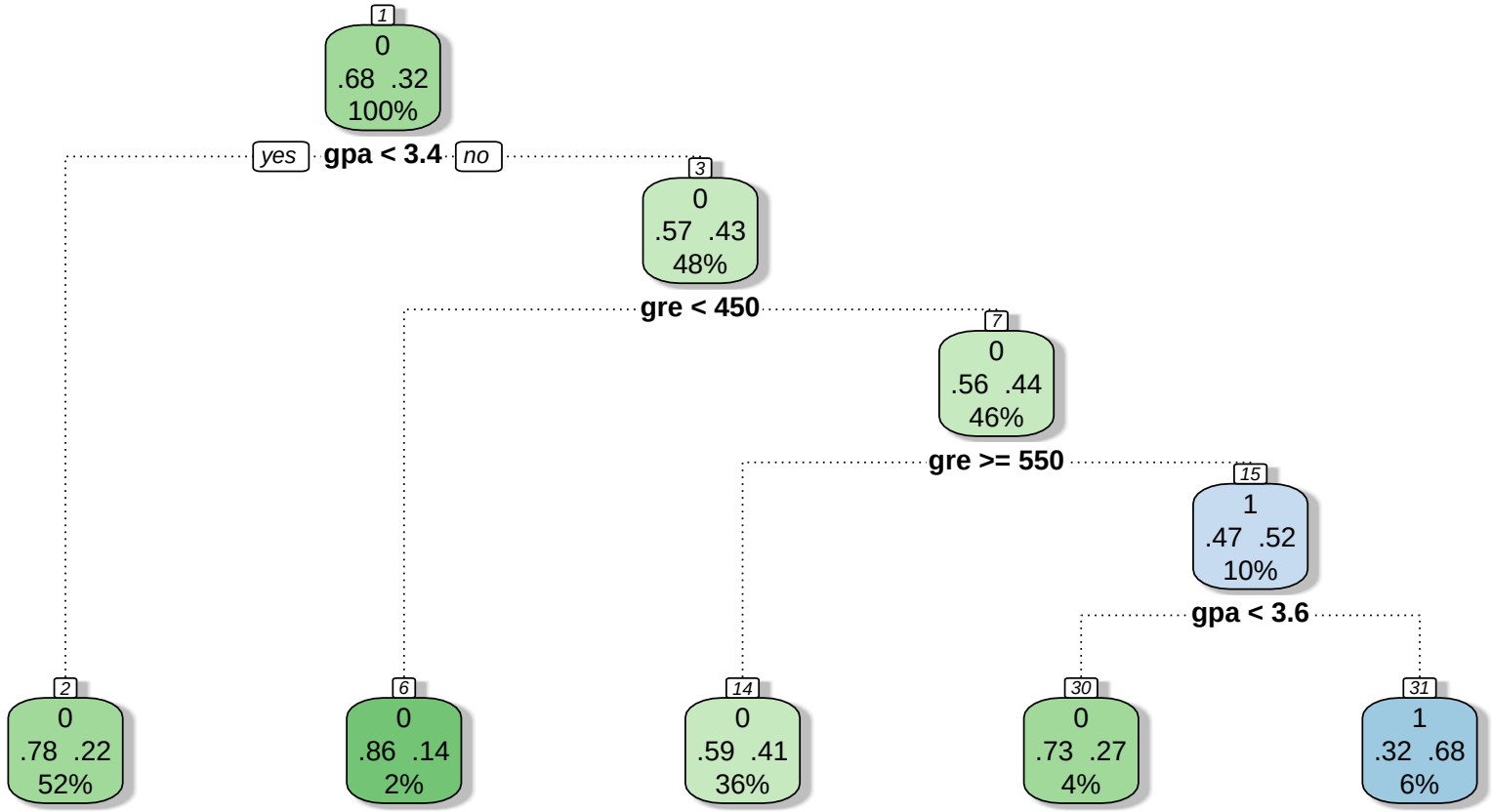
Decision Trees

Decision Trees

- Build a model to predict admit

admit	gre	gpa
0	380	3.61
1	660	3.67
1	800	4.00
1	640	3.19
0	520	2.93
1	760	3.00



Rattle 2023−Sep−23 18:40:06 sonou

Making Prediction

- Predict the outcome of an applicant with 700 GRE score and 3.5 GPA

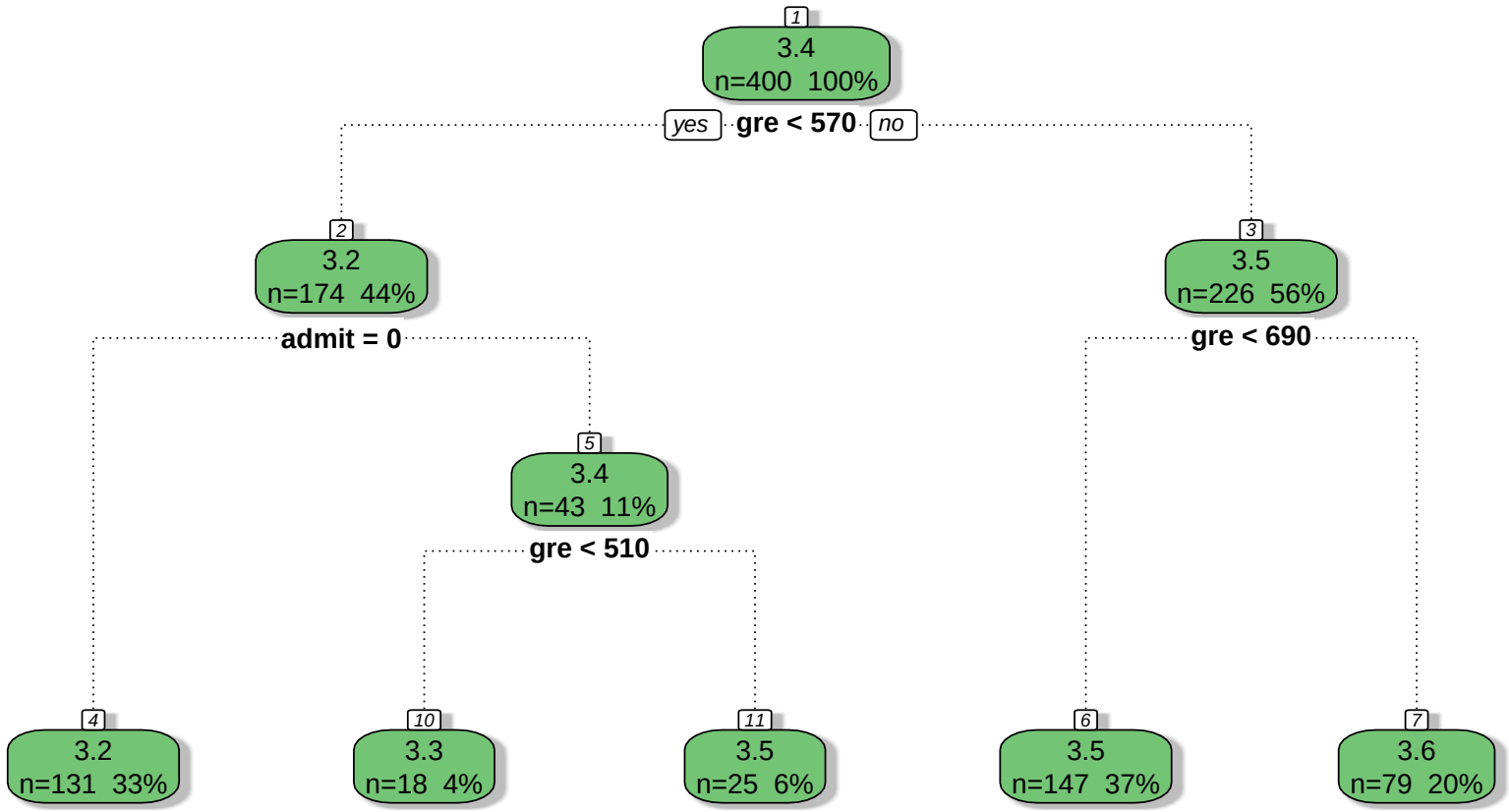
Decision Trees on the coordinate

- Plot the previous tree on the coordinate of GRE and GPA

Regression Trees

- Build a model to predict gpa

admit	gre	gpa
0	380	3.61
1	660	3.67
1	800	4.00
1	640	3.19
0	520	2.93
1	760	3.00



Rattle 2023−Sep−23 18:40:06 sonou

Example

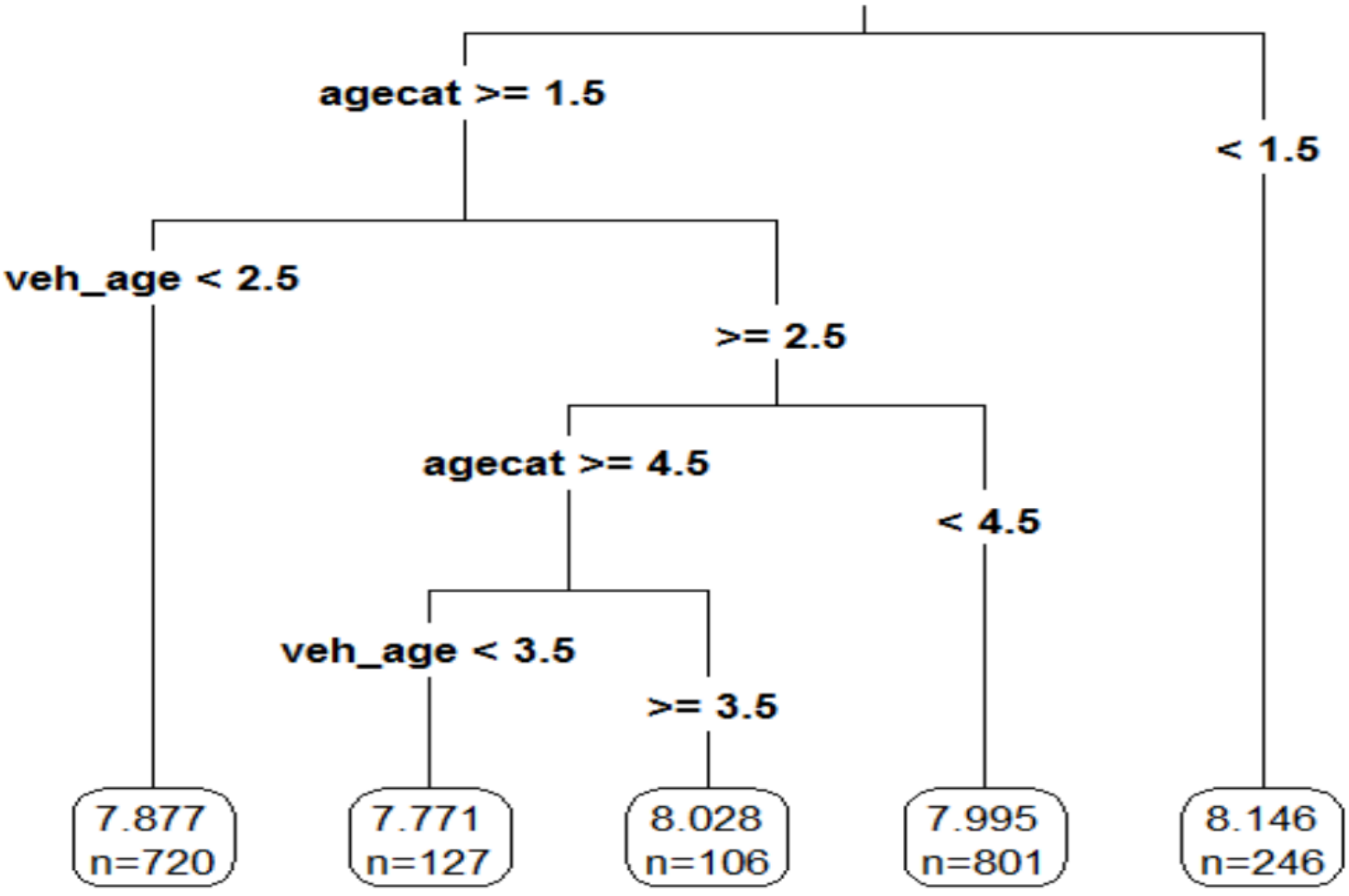
The regression tree shown below was produced from a dataset of auto claim payments. Age Category (agecat: 1, 2, 3, 4, 5, 6) and Vehicle Age (veh_age: 1, 2, 3, 4) are both predictor variables, and log of claim amount (LCA) is the dependent variable. Rank the estimated LCA of Autos I, II, and III.

I: An Auto in Age Category 1 and Vehicle Age 4

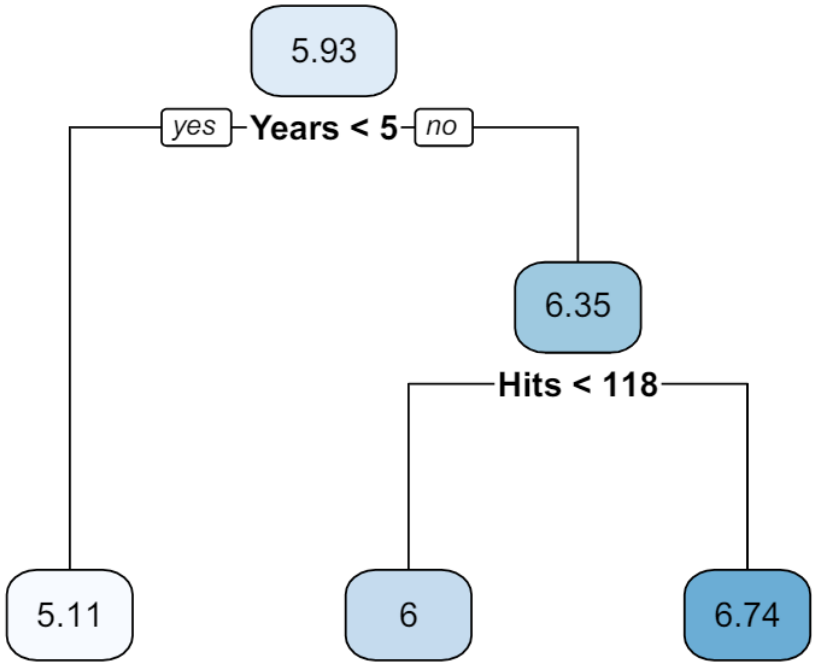
II: An Auto in Age Category 5 and Vehicle Age 5

III: An Auto in Age Category 5 and Vehicle Age 3

Calculate the estimated LCA of Autos I, II, and III.

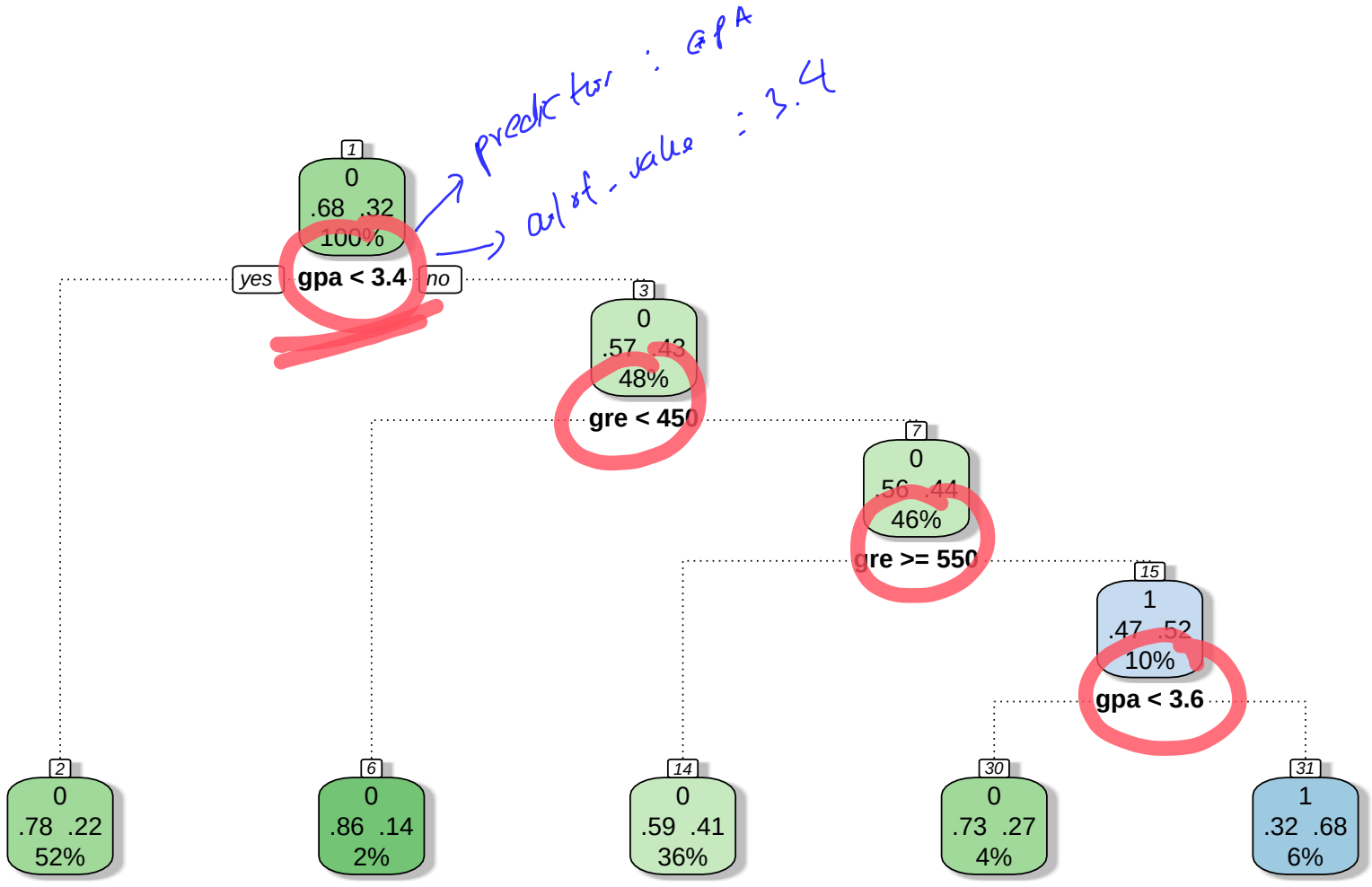

agecat >= 1.5
< 1.5
veh_age < 2.5
>= 2.5
agecat >= 4.5
< 4.5
veh_age < 3.5
>= 3.5
7.877
n=720
7.771
n=127
8.028
n=106
7.995
n=801
8.146
n=246

Consider the following regression tree to predict the log salary of a baseball player.



- Calculate the predicted salary for a player who has played 10 years and has 100 hits.

Growing a Tree



Rattle 2023−Sep−23 18:40:06 sonou

- A tree is a combination of a sequence of splits
- Every split of a tree is defined by
  - The predictor it split at
  - The cutoff value
- For example, a split at gpa (predictor) of 3.5 (cutoff value)
- How to determine the first split, the second split, and so on?

- For the first split: Consider all possible split (a combination of all possible predictors and cutoffs values) then choose the best one
- For the second split: Consider all possible split after the first split then choose the best one
- And so on

Classification Tree

Impurity of a Node

node A is more uncertain than node B

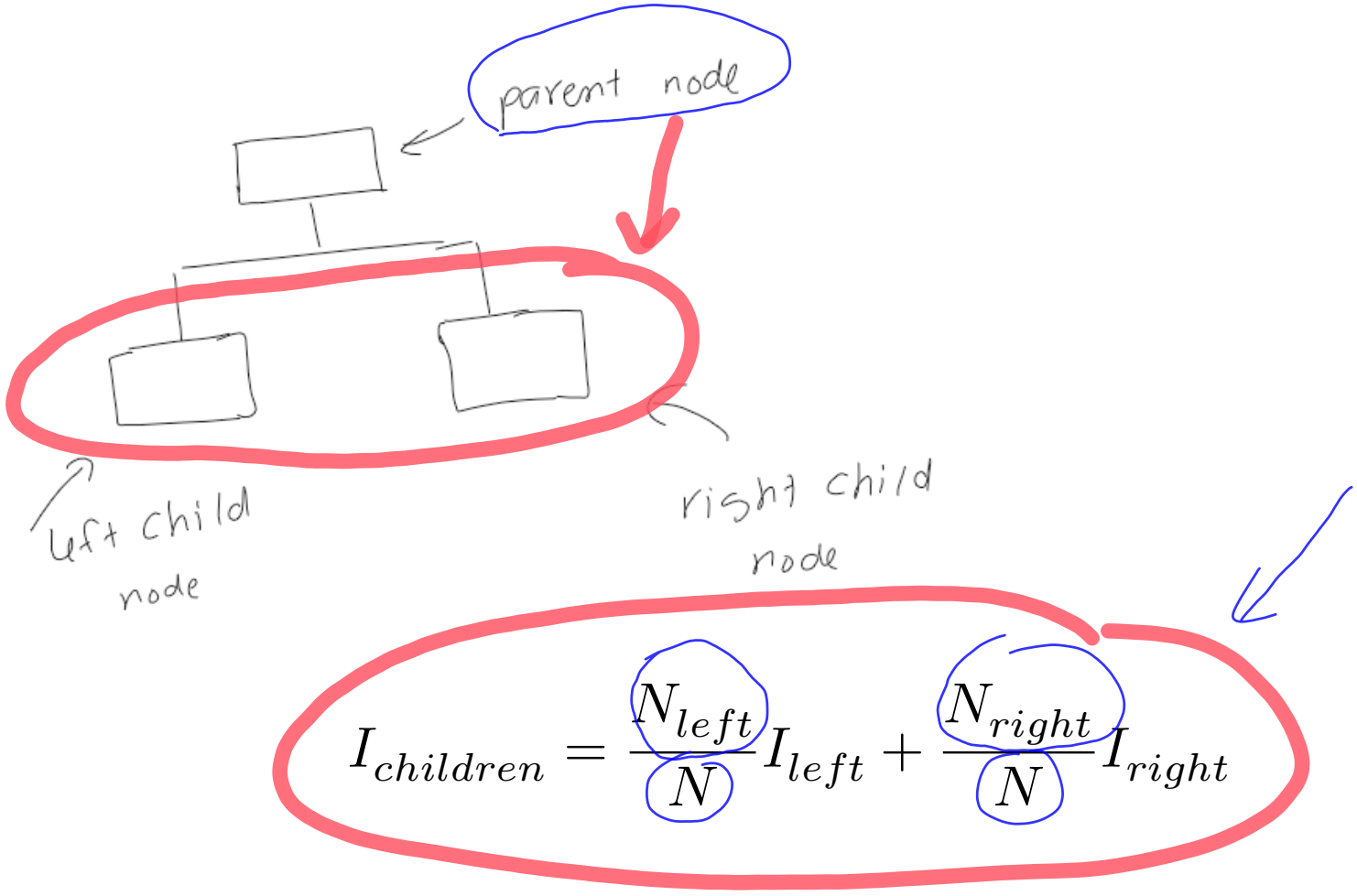
- Impurity can be measured by: classification error, Gini Index, and Entropy.
- Let p_0 and p_1 be the proportion of class 0 and class 1 in a node.

$$\text{By Classification Error: } I = min\{p_0, p_1\}$$

$$\text{By Gini Index: } I = 1 - p_0^2 - p_1^2$$

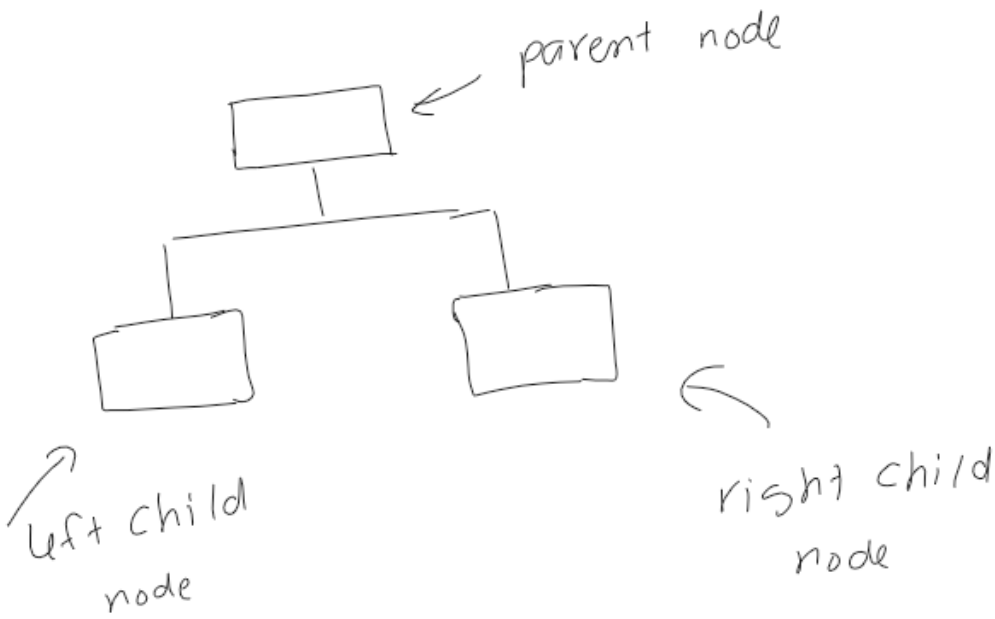
$$\text{By Entropy: } I = -p_0 \log_2(p_0) - p_1 \log_2(p_1)$$

Children Impurity of a Slit



- N_{left} and N_{right} are the number of points in the left child node and right child node, respectively.
- $N_{left} + N_{right} = N$

Impurity Gain of a Split



$$IG = I_{parent} - I_{children}$$

- IG is Impurity Gain of the split

- The split with the highest impurity gain is the best
- Coming from the same parent node, the split with the lowest total children impurity is the best

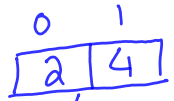
Example 1

admit	gre	gpa
0	380	3.61
1	660	3.67
1	800	4.00
1	640	3.19
0	520	2.93
1	760	3.00

- Split 1: Split at $gre \geq 700$
- Split 2: Split at $gre \geq 3.5$

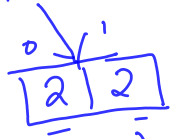
Which split is the best? use Gini Index

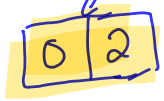
admit	gre	gpa
0	380	3.61
1	660	3.67
1	800	4.00
1	640	3.19
0	520	2.93
1	760	3.00



gre < 700

gre ≥ 700





$p_0 = \frac{0}{2} = 0$

$p_1 = \frac{2}{2} = 1$

$I_{left} = 1 - 0^2 - 1^2 = 0$

$p_0 = \frac{2}{4}, p_1 = \frac{2}{4}$

$I_{right} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \frac{1}{2}$

$I_{children\ 1} = \frac{2}{6} \cdot 0 + \frac{4}{6} \cdot \frac{1}{2} = \frac{1}{3} = \boxed{\frac{3}{9}}$ ← better

Split 2 :

gpa < 2.5

gpa ≥ 2.5

$p_0 = \frac{1}{3}$

$p_1 = \frac{2}{3}$

$p_0 = 1/3$

$p_1 = \frac{2}{3}$

$I_{left} = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$

$I_{right} = \frac{4}{9}$

$\frac{3}{6} \cdot \frac{4}{9} + \frac{3}{6} \cdot \frac{4}{9} = \boxed{\frac{4}{9}}$

$I_{children\ 2} =$

Example 2

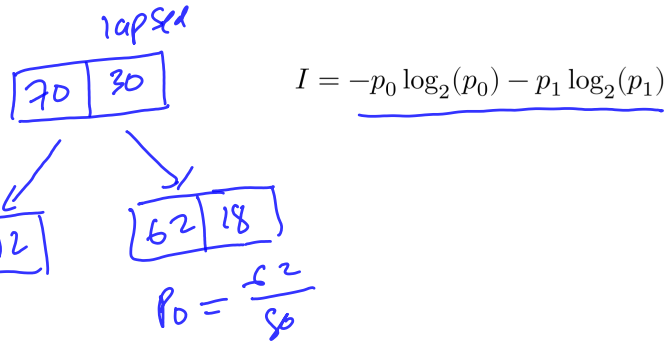
A classification tree is being constructed to predict if an insurance policy will lapse. A random sample of 100 policies contains 30 that lapsed. You are considering two splits:

- Split 1: One node has 20 observations with 12 lapses and one node has 80 observations with 18 lapses.
- Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 22 lapses.

The total Gini index after a split is the weighted average of the Gini index at each node, with the weights proportional to the number of observations in each node.

The total entropy after a split is the weighted average of the entropy at each node, with the weights proportional to the number of observations in each node.

Determine the best split based on the total entropy.

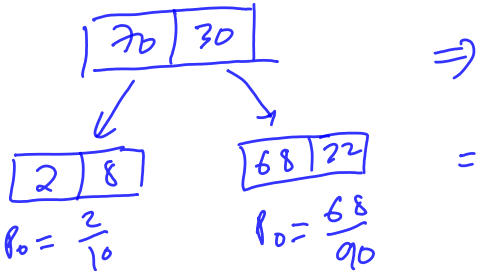


$$I = -p_0 \log_2(p_0) - p_1 \log_2(p_1)$$

$p_0 = \frac{8}{20}$ $p_1 = \frac{12}{20}$ $p_0 = \frac{62}{80}$ $p_1 = \frac{18}{80}$

$$I_{children} = \frac{20}{100}\left[-\frac{8}{20}\cdot\log_2\left(\frac{8}{20}\right) - \frac{12}{20}\cdot\log\left[\frac{12}{20}\right)\right] + \frac{80}{100}\left[-\frac{62}{80}\cdot\log_2\frac{62}{80} - \frac{18}{80}\log_2\frac{18}{80}\right]$$

$$= .5611$$



$p_0 = \frac{2}{10}$ $p_1 = \frac{8}{10}$ $p_0 = \frac{68}{90}$ $p_1 = \frac{22}{90}$

$\Rightarrow I_{children}$

$$= \frac{10}{100}\left[-\frac{2}{10}\cdot\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10}\right] + \frac{90}{100}\left[-\frac{68}{90}\cdot\log_2\frac{68}{90} - \frac{22}{90}\cdot\log_2\frac{22}{90}\right]$$

$$= .5506$$ better

Regression Trees

- The tree will search for all combination of predictors and cutoff value to decide the best split
- In Regression tree, the best split is the split that minimizes

$$\underbrace{\sum_{i:\mathbf{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2}_{\text{RSS of obs. in left branch}} + \underbrace{\sum_{i:\mathbf{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2}_{\text{RSS of obs. in right branch}}$$

- $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the means of the responses falling in to the left branch and right branch, respectively.

Example

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$

SPLIT 1

$X_1 < 4$: $Y = \{1.2, 2.1, 1.5 \ldots, 2.0, 1.6\}$

$X_1 \geq 4$: $Y = \{3.0\}$ $\quad \bar{Y}_2 = 3$

$$\bar{Y} = \frac{1.2 + 2.1 + 1.5 + 3.0 + 2.0 + 1.6}{6} = 1.9$$

$$\bar{Y}_1 = \frac{1.2 + 2.1 + 1.5 + 2 + 1.6}{5} = 1.68$$

$$RSS_1 = (1.2 - 1.68)^2 + (2.1 - 1.68)^2 + (1.5 - 1.68)^2 + (2 - 1.68)^2 + (1.6 - 1.68)^2 = .548$$

$$RSS_2 = (3 - 3)^2 = 0$$

total RSS $= .548$

split 2

$$R_1 = \{1.2, 2.1, 3, 1.6\} \Rightarrow \bar{R}_1 = 1.975$$

$$R_2 = \{1.5, 2\} \Rightarrow \bar{R}_2 = 1.75$$

$$RSS_1 = (1.2 - 1.975)^2 + (2.1 - 1.975)^2 + (3 - 1.975)^2 + (1.6 - 1.975)^2$$

$$RSS_2 = (1.5 - 1.75)^2 + (2 - 1.75)^2$$

$$\text{total } RSS = RSS_1 + RSS_2 = \boxed{1.9325} \leftarrow$$