Decision Tree

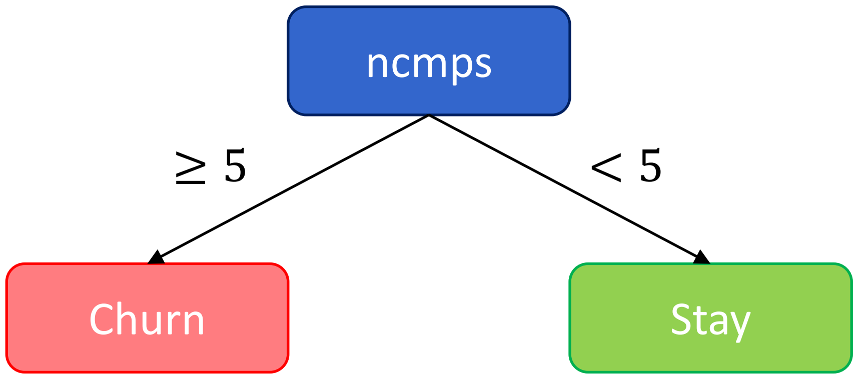
Ratchainant Thammasudjarit, Ph.D.

Learning Objectives

- Understand the theory, concepts and applications of tree classifier
- Understand the concepts of ensemble learning (bagging)
- Understand how to build a tree-based classifier using sklearn

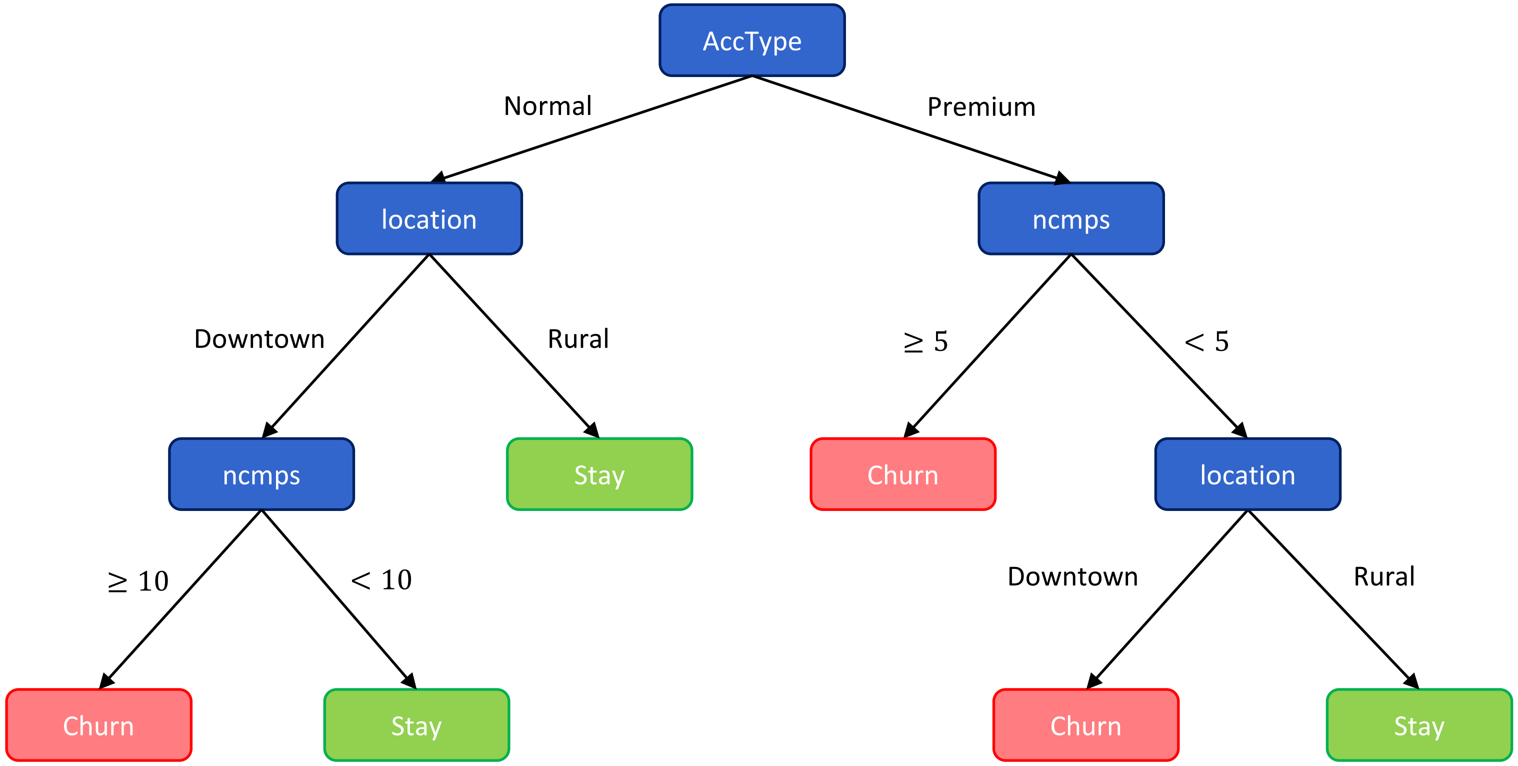
Basic Idea

- Decision Tree
  - Non-parametric model
  - A collection of rules organized in hierarchy



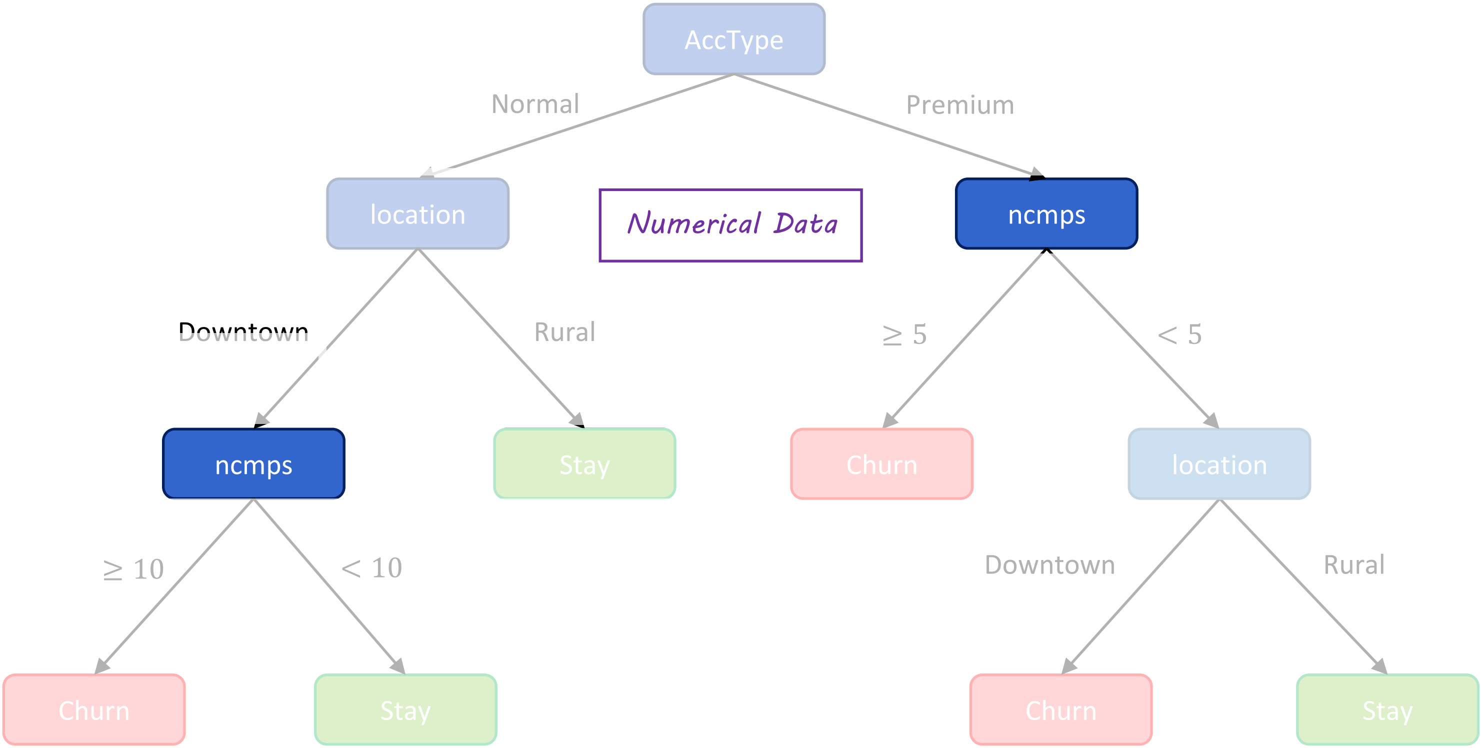
Basic Idea

- The more complicated decision tree



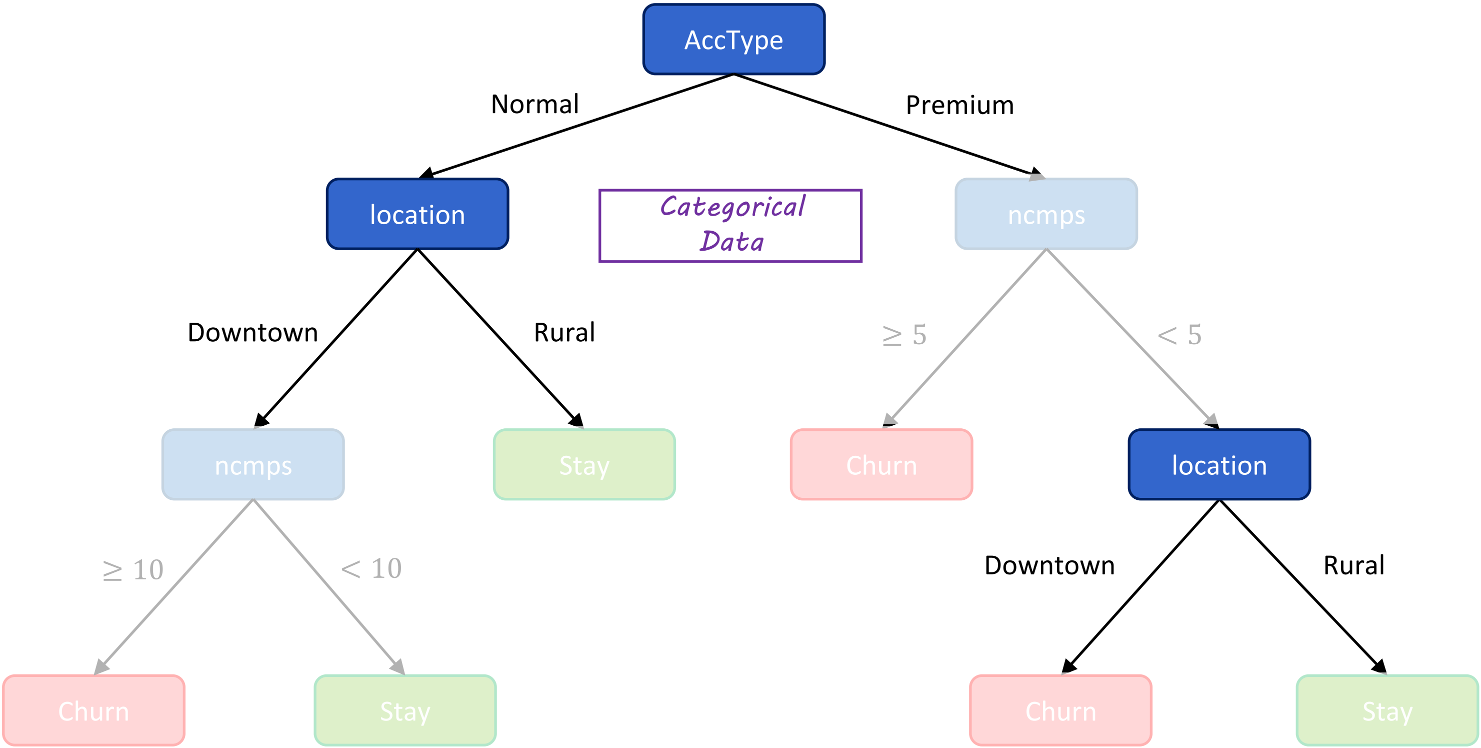
Basic Idea

- The more complicated decision tree



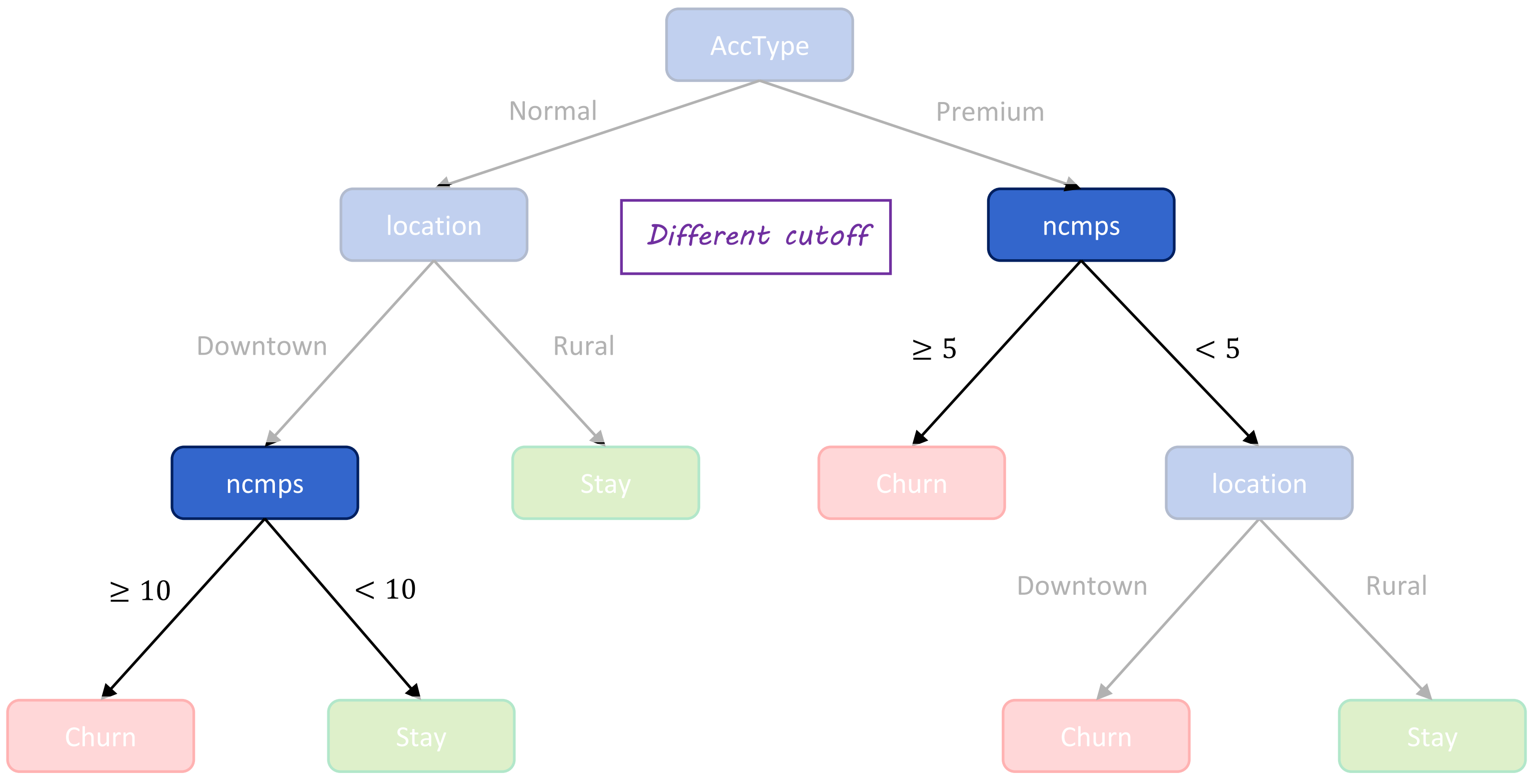
Basic Idea

- The more complicated decision tree



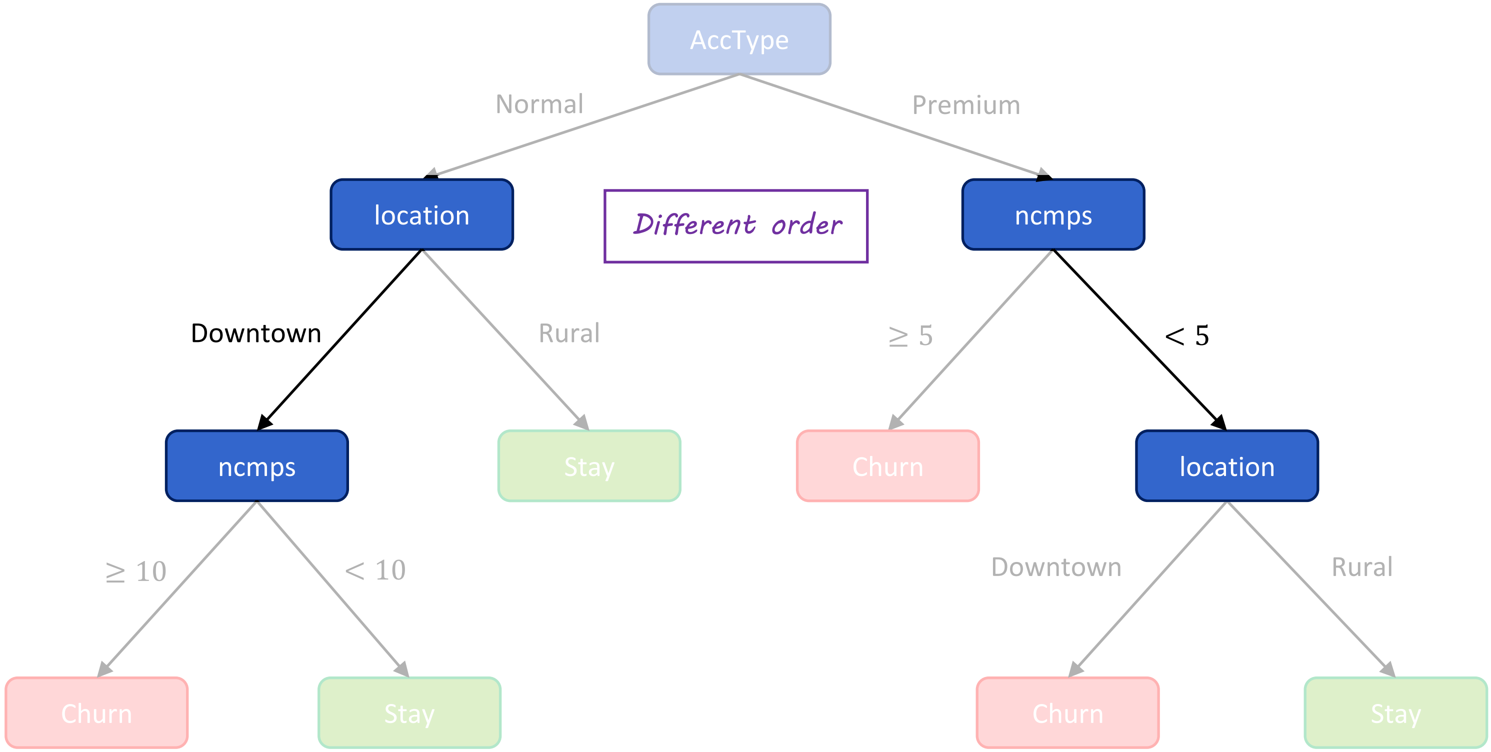
Basic Idea

- The more complicated decision tree



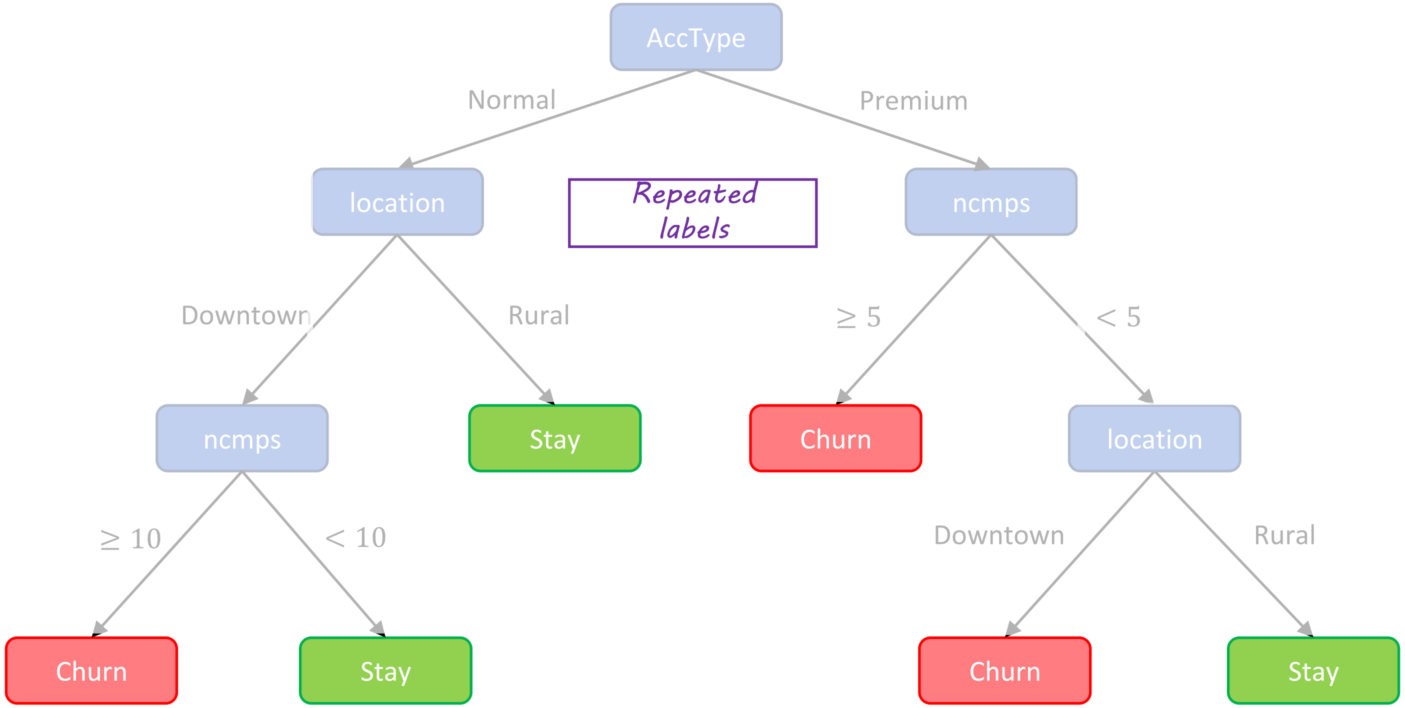
Basic Idea

- The more complicated decision tree

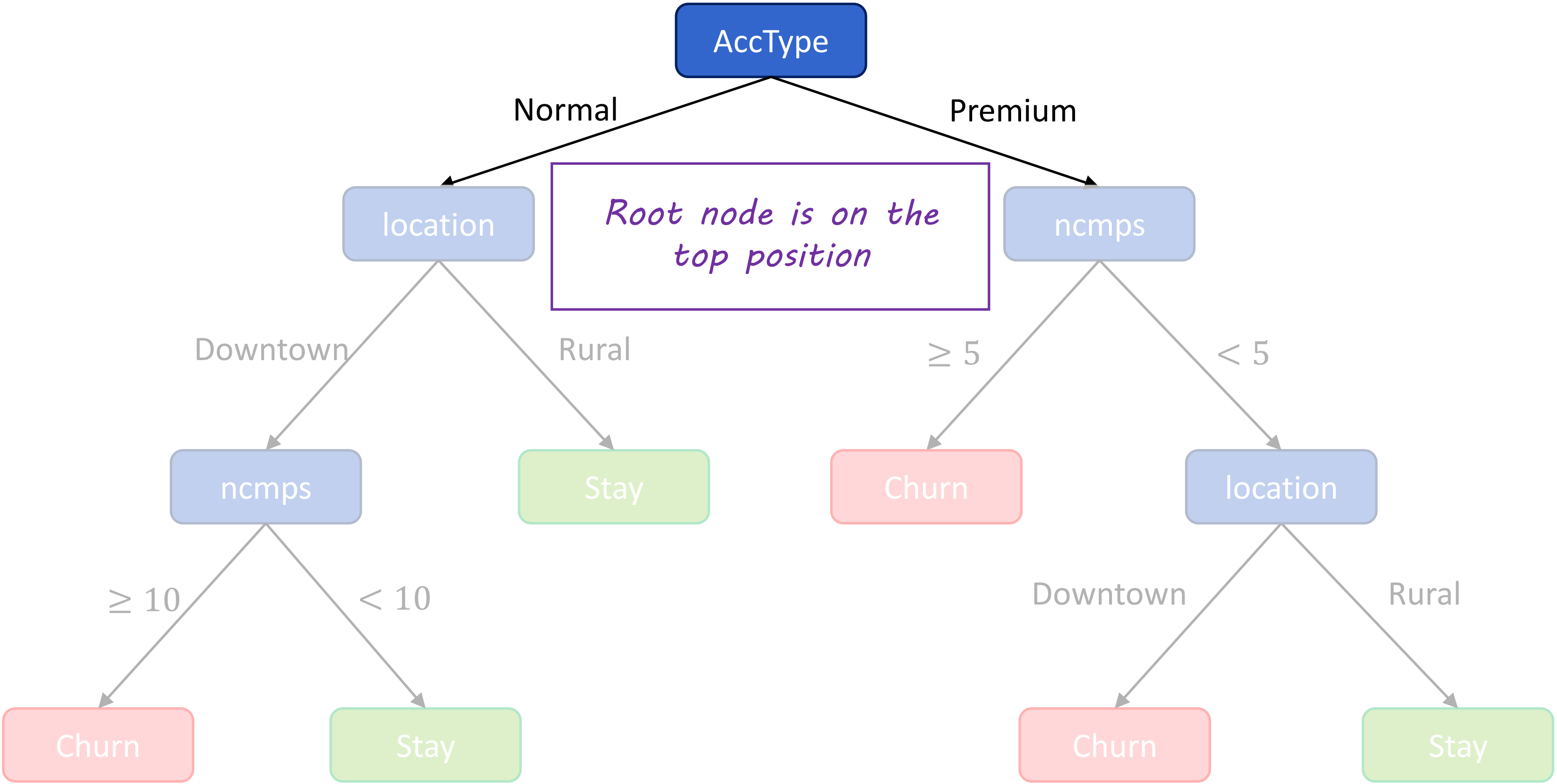


Basic Idea

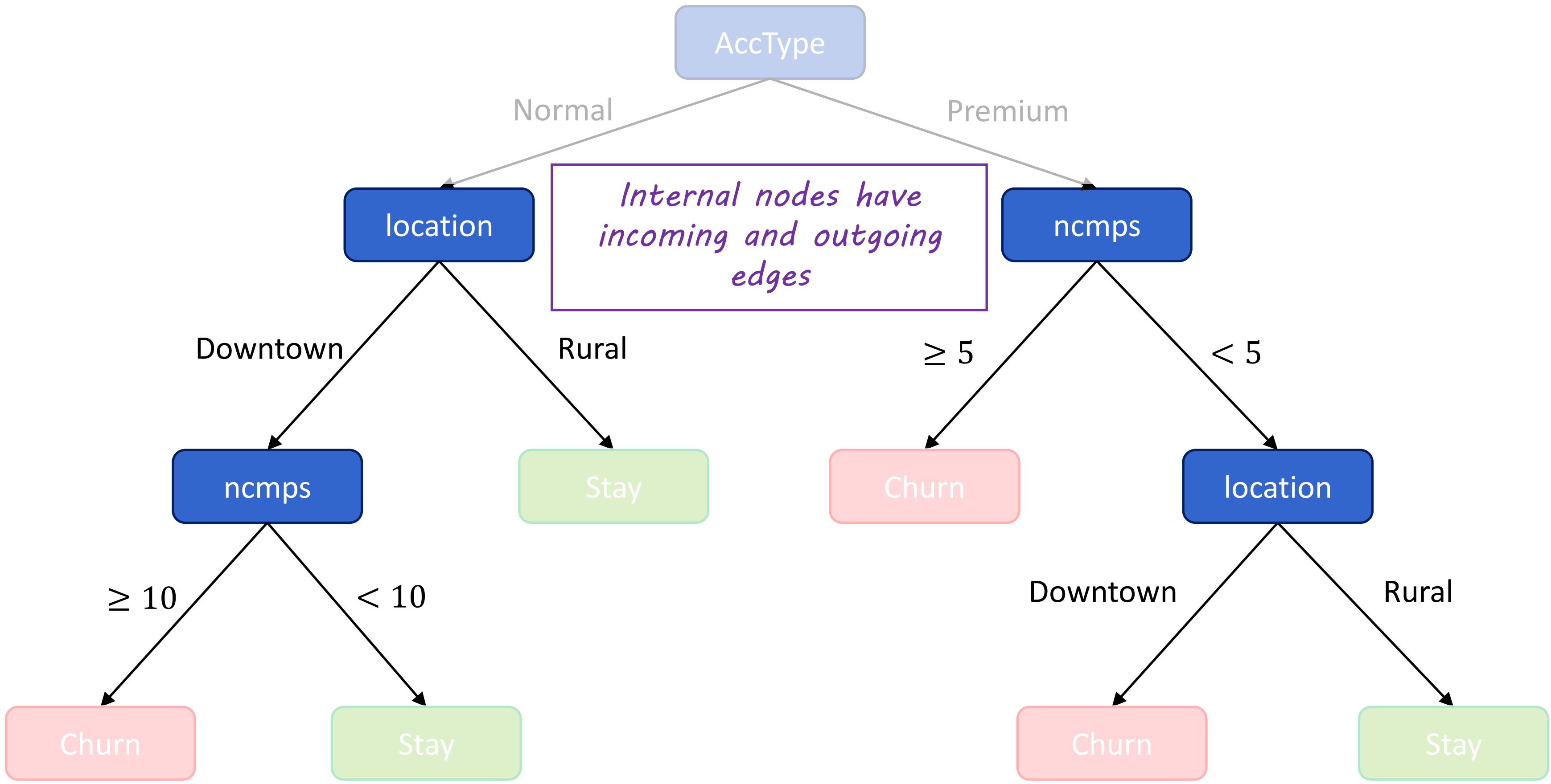
- The more complicated decision tree



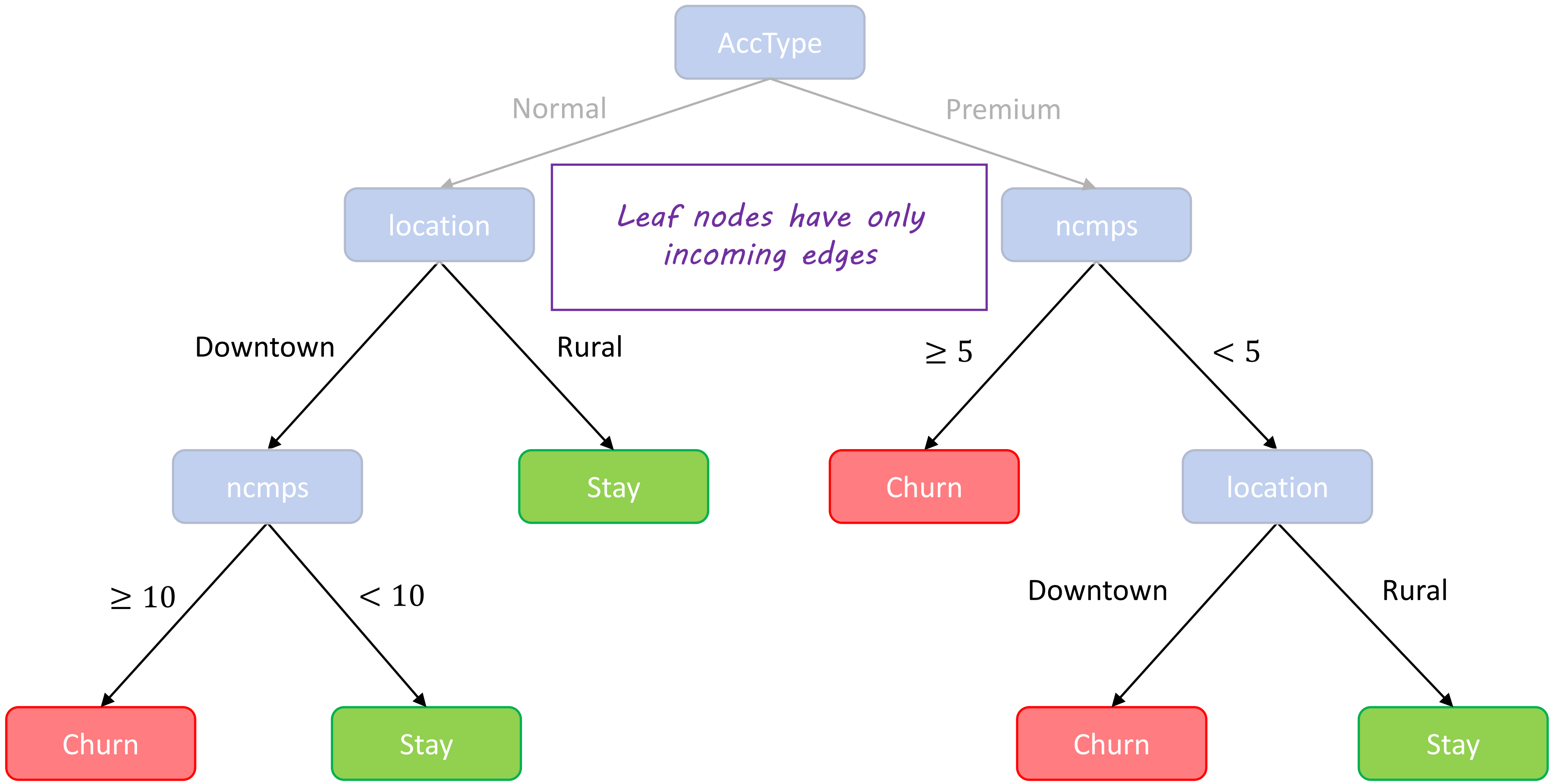
Terminologies



Terminologies



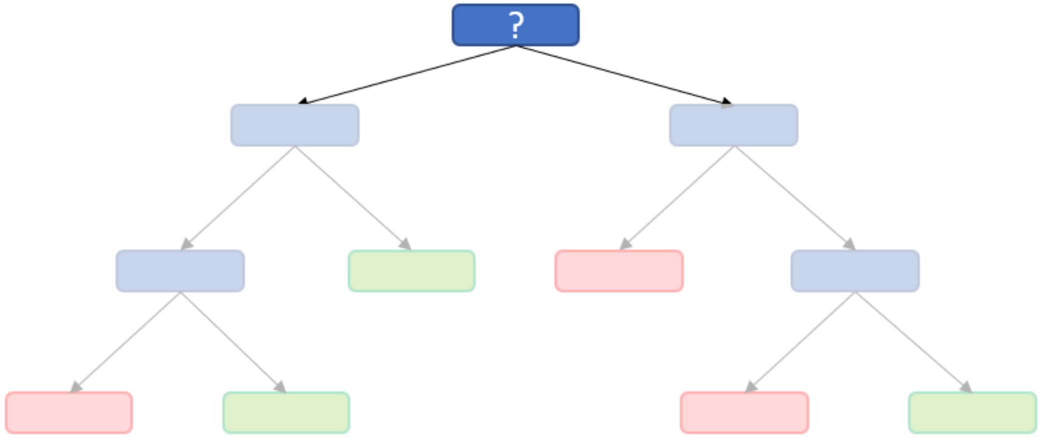
Terminologies



Impurity Measure

- Creating a decision tree requires impurity measure for feature selection
    - Given a dataset, which feature will be selected for each node
        - **Root node**
        - Internal nodes

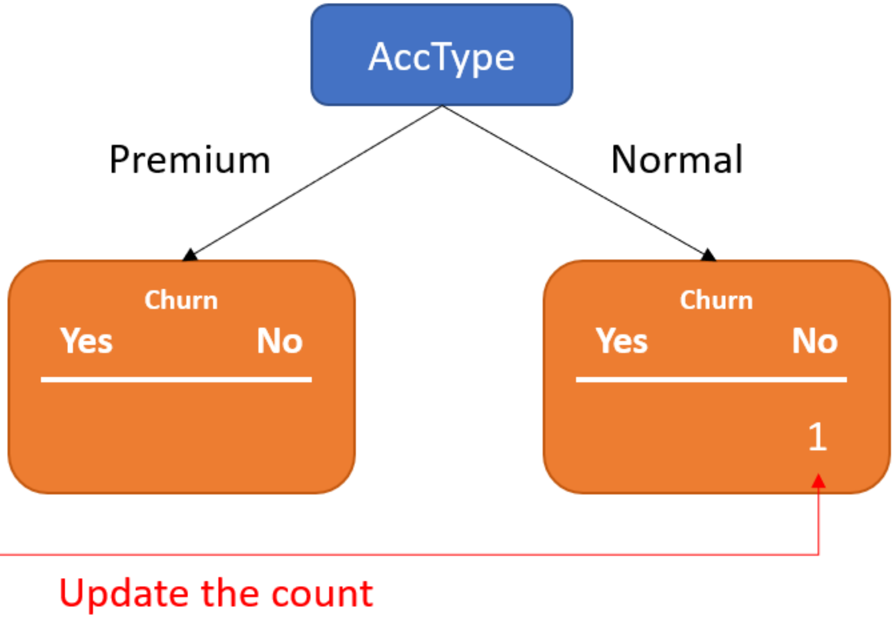
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - AccType

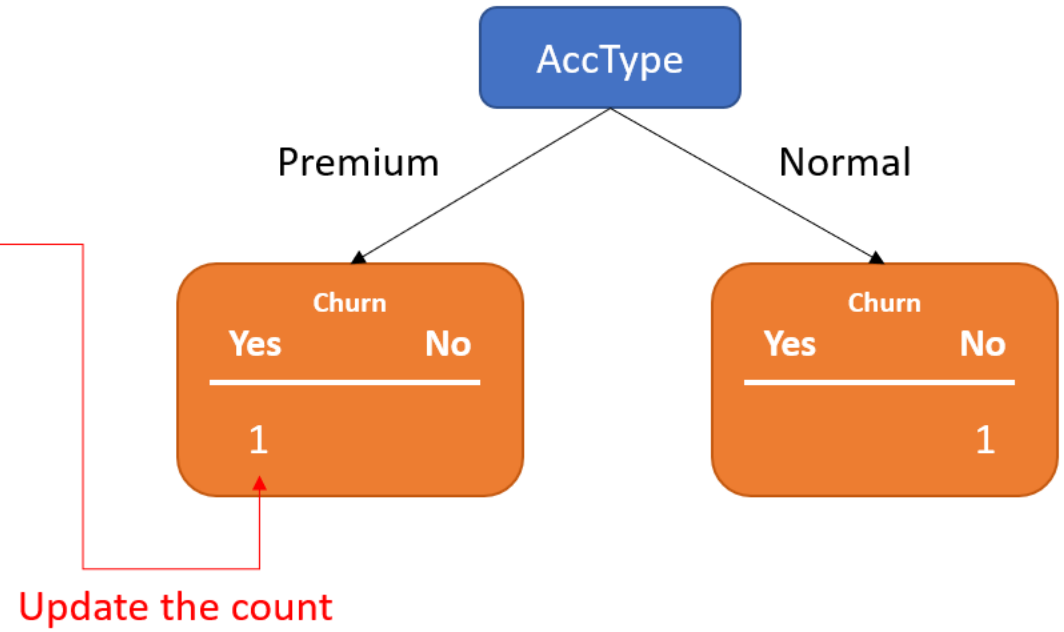
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - AccType

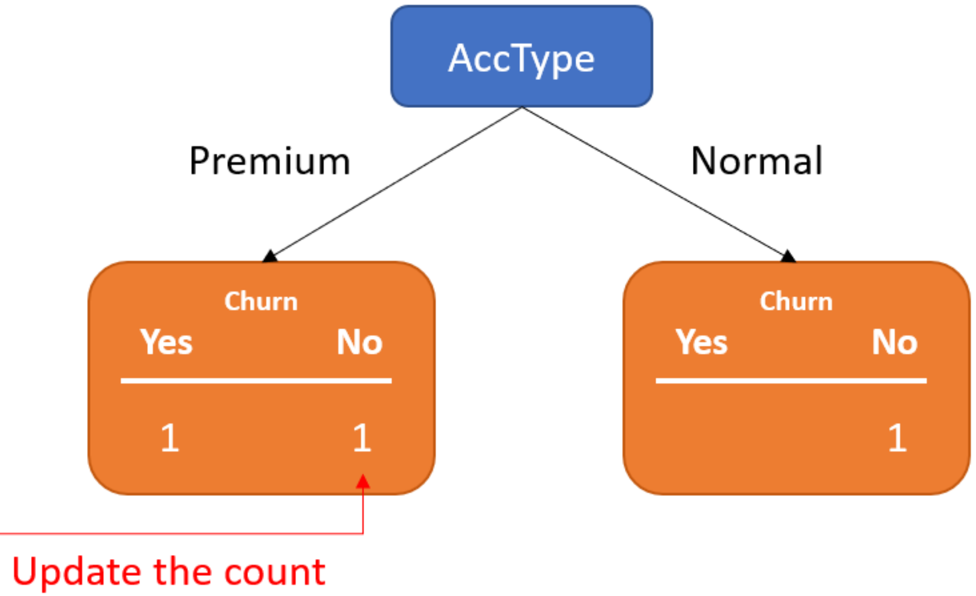
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - AccType

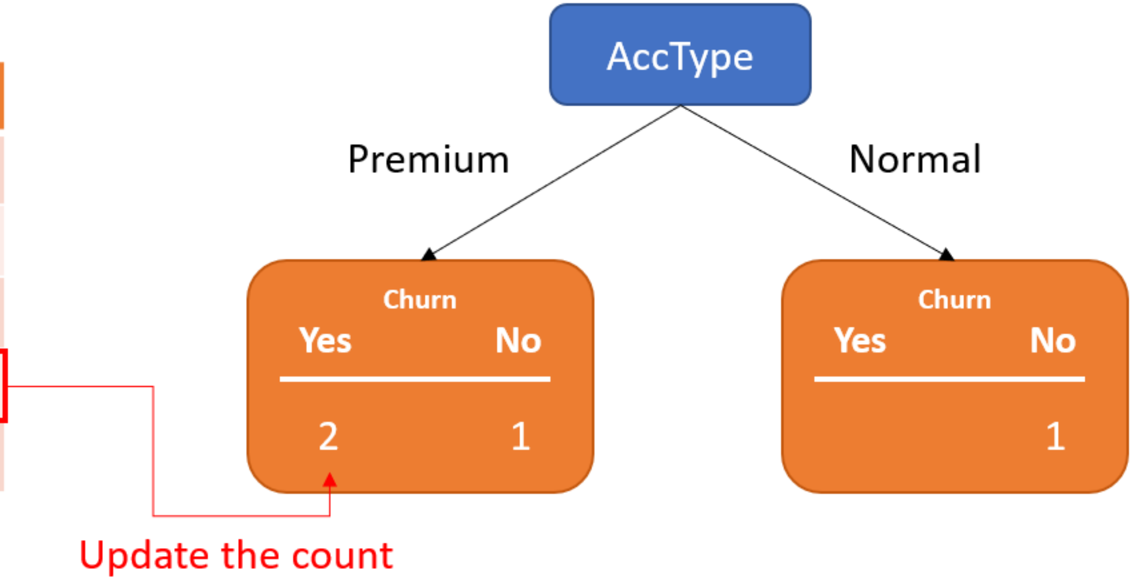
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - AccType

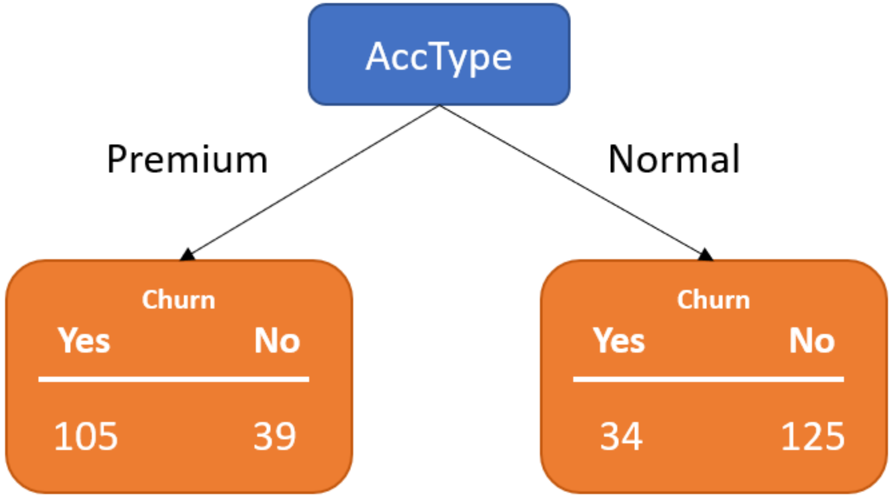
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - AccType

AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...

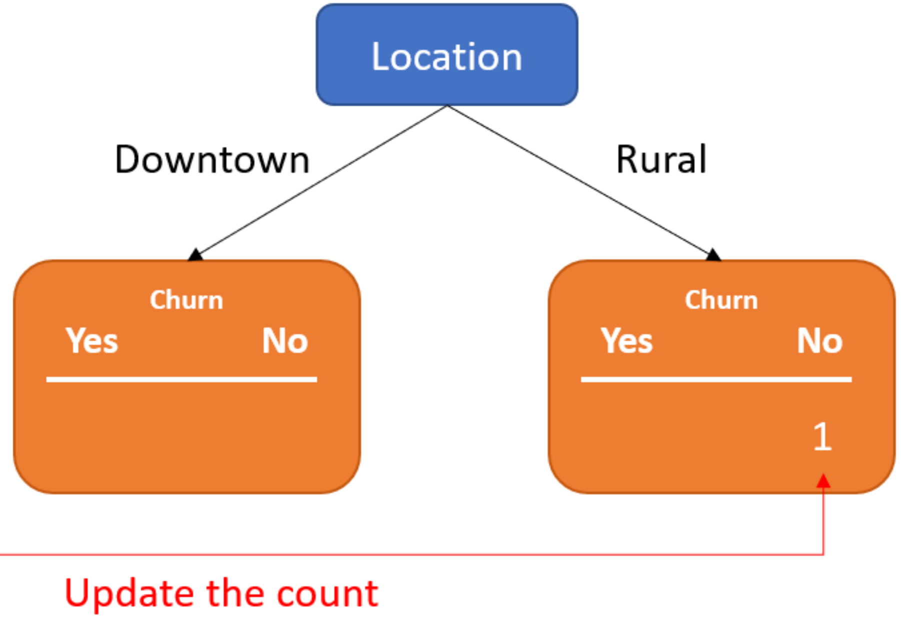


Suppose this is the final count from data

Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - Location

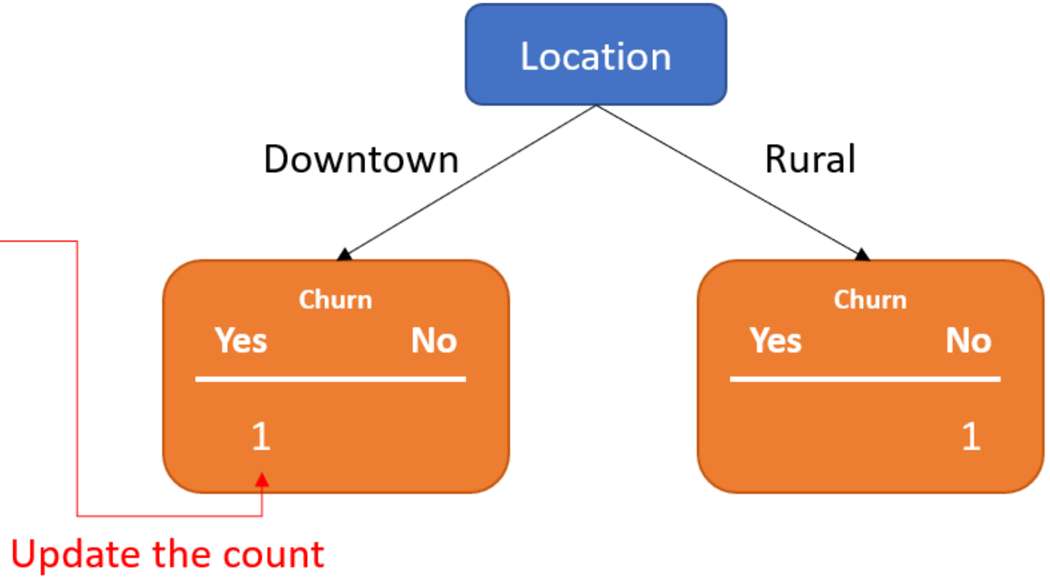
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - Location

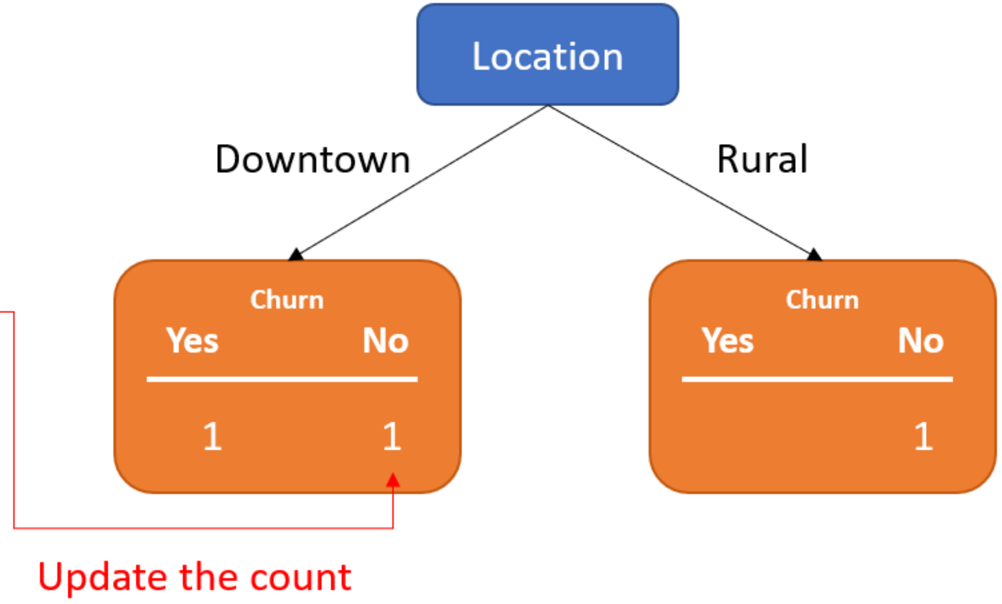
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - Location

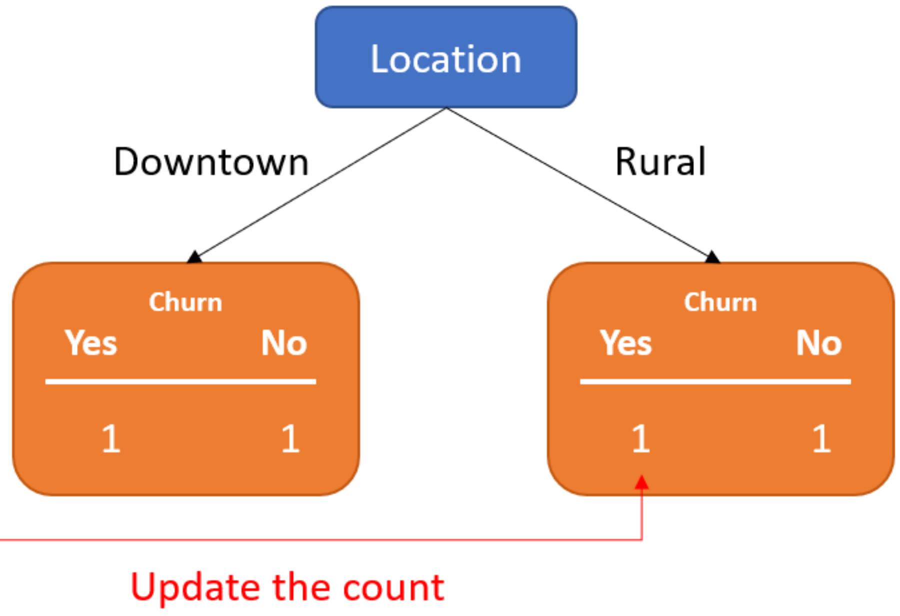
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - Location

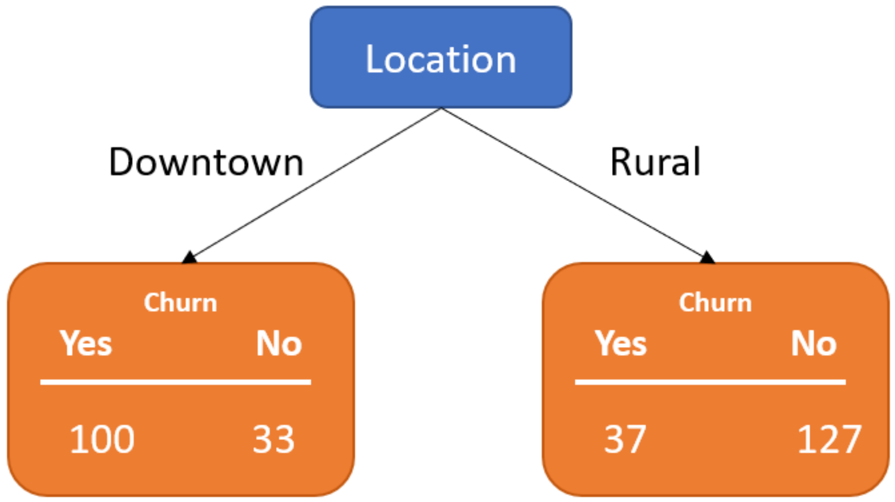
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - Location

AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...

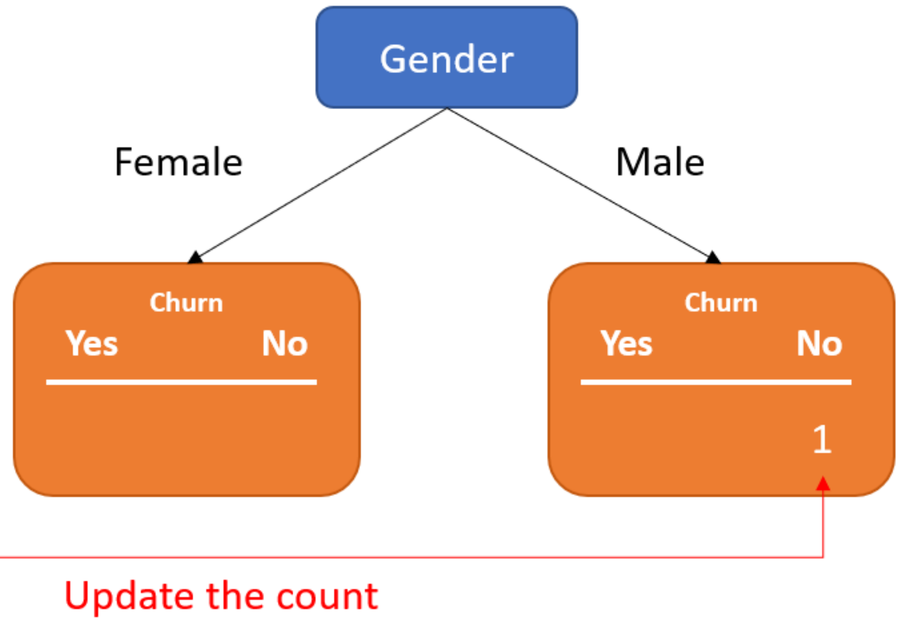


Suppose this is the final count from data

Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - Gender

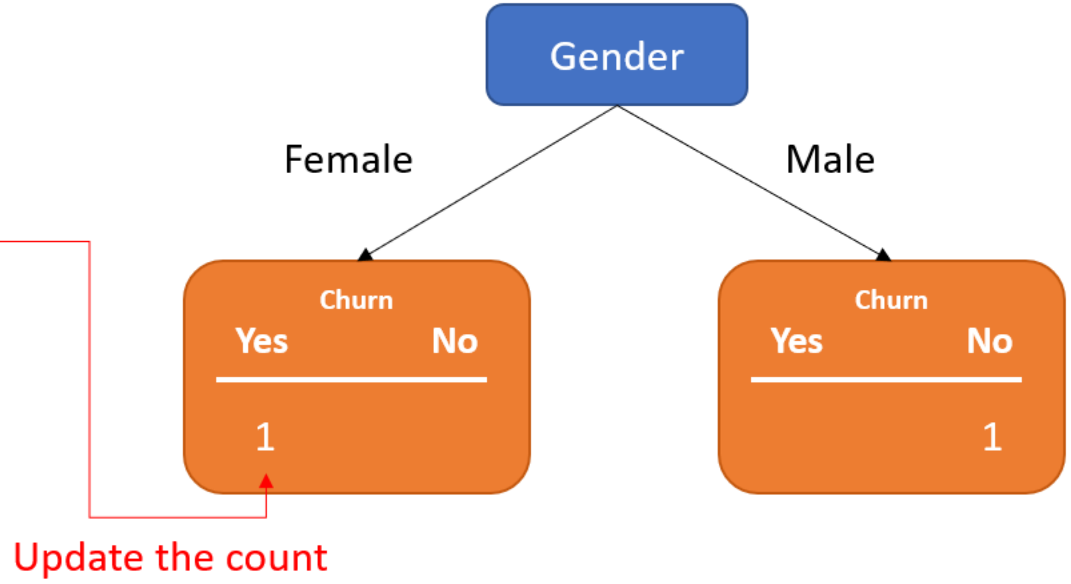
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
…	…	…	…



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - Gender

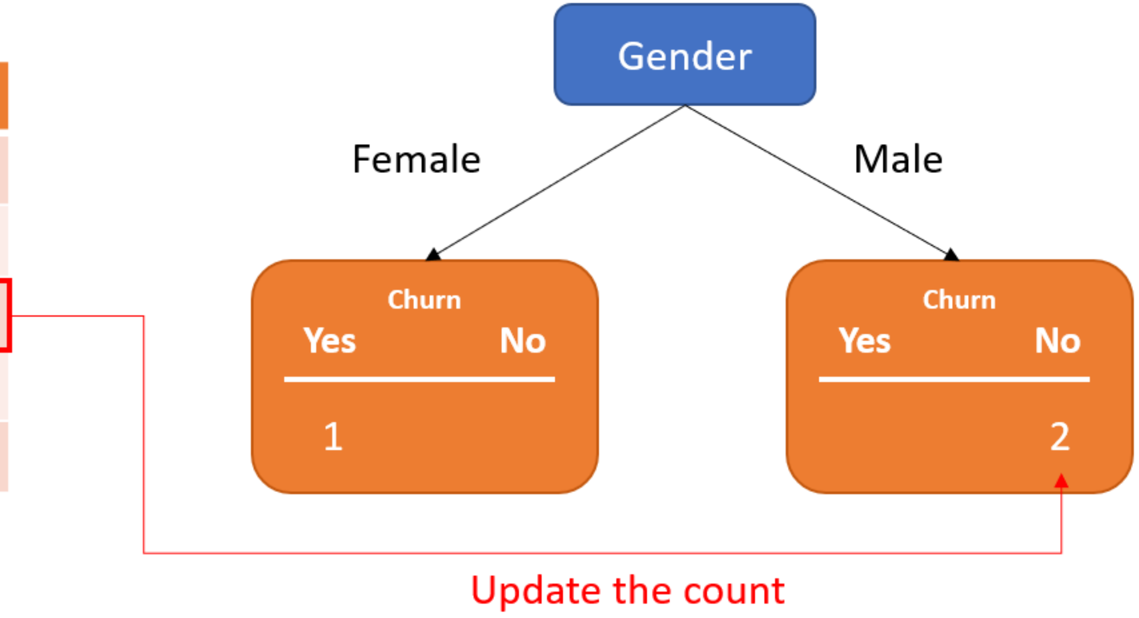
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - Gender

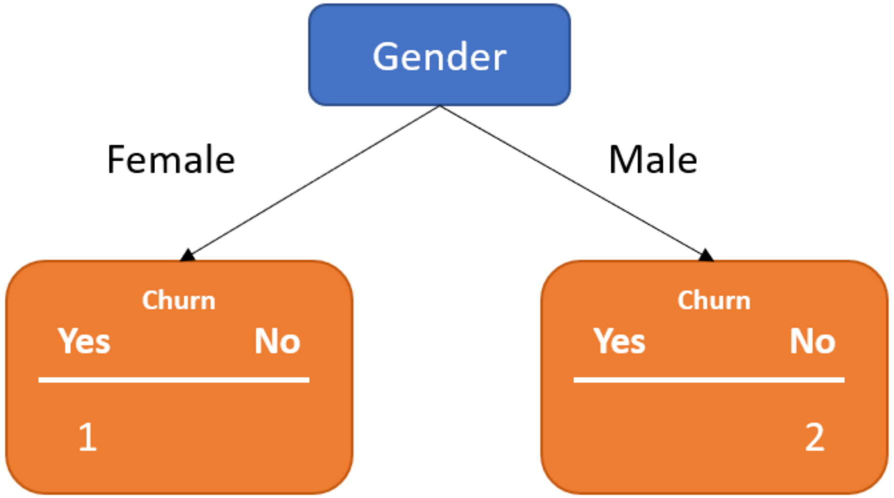
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
    - Gender

AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...

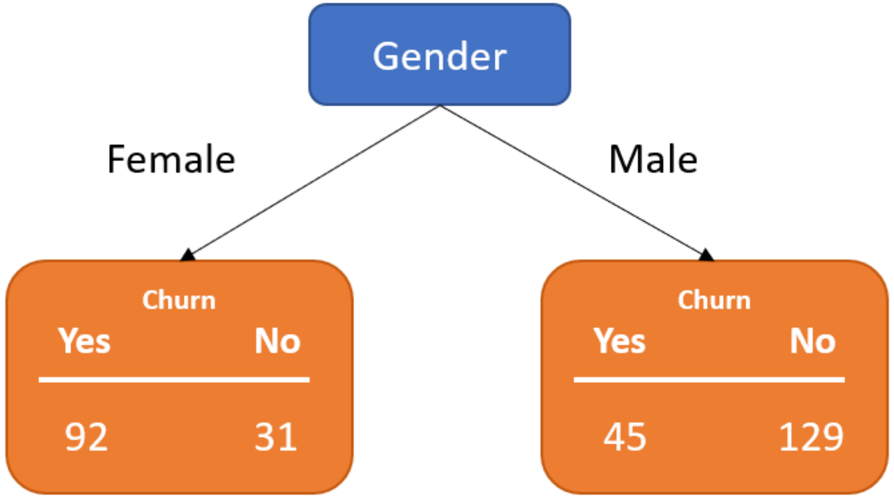


Since we don't know whether this customer is male or female, we skip counting

Impurity Measure

- Each feature will be evaluated how well it predicts the class label (Churn)
  - Gender

AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...

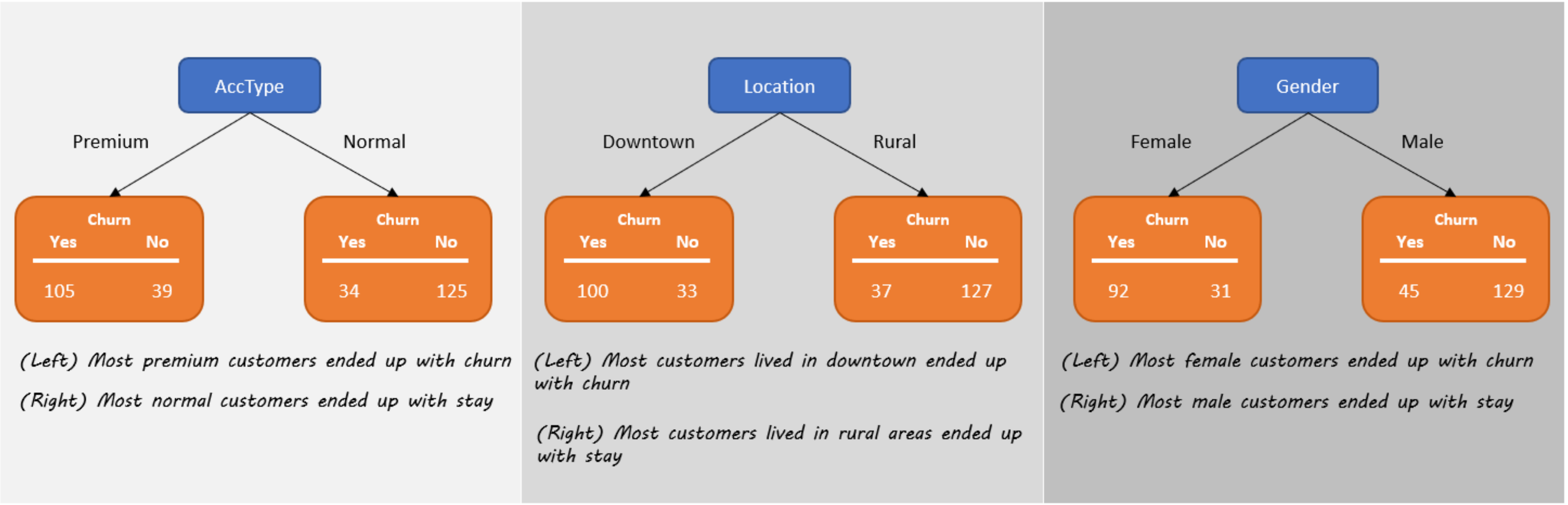


Suppose this is the final count from data

Impurity Measure

- Which feature is the best to be the root node
  - None of them can 100% separate yes from no
  - This is called **impure**

All features are imperfect to separate churn (yes, no)



Gini Impurity

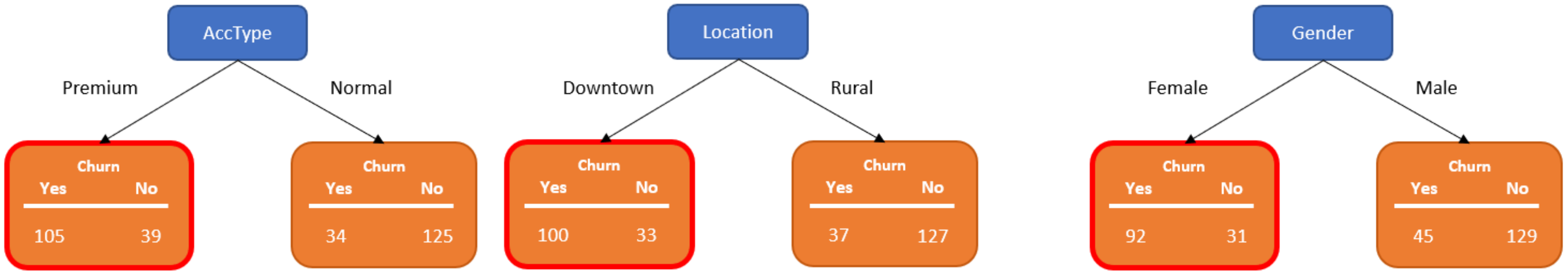
- Let x be any value in a feature $\mathbf{x}$
- y be any possible value in the target $\mathbf{y}$
- Gini Impurity of x is $G(x)$

$$G(x) = 1 - \sum_{y} P(y|x)^2$$

- The total Gini Impurity of a feature $\mathbf{x}$ is the weighted average of $G(x)$ for all x
- The lower total Gini Impurity the better feature to separate distinct values in $\mathbf{y}$

Gini Impurity

- Calculation Example



$$G(premium) = 1 - P(yes|premium)^2 - P(no|premium)^2$$

$$= 1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2$$

$$= 0.395$$

$$G(downtown) = 1 - P(yes|downtown)^2 - P(no|downtown)^2$$

$$= 1 - \left(\frac{100}{100+33}\right)^2 - \left(\frac{33}{100+33}\right)^2$$

$$= 0.373$$

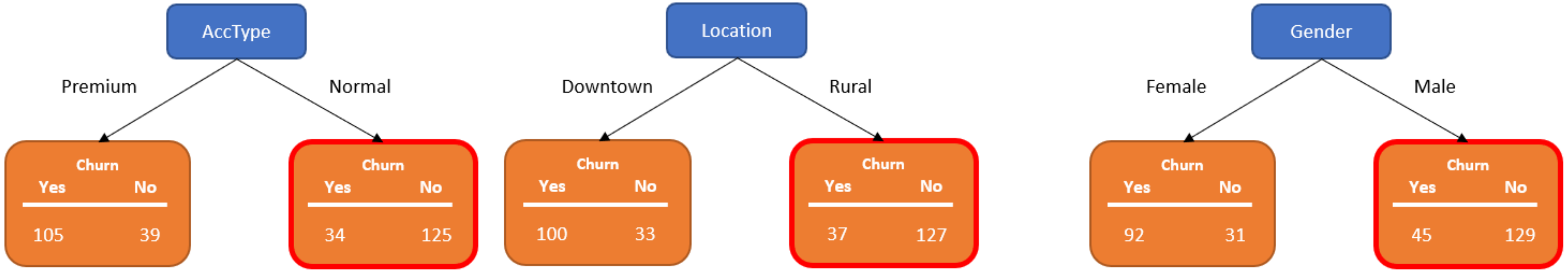
$$G(female) = 1 - P(yes|female)^2 - P(no|female)^2$$

$$= 1 - \left(\frac{92}{92+31}\right)^2 - \left(\frac{31}{92+31}\right)^2$$

$$= 0.377$$

Gini Impurity

- Calculation Example



$$G(premium) = 1 - P(yes|premium)^2 - P(no|premium)^2$$

$$= 1 - \left(\frac{34}{34+125}\right)^2 - \left(\frac{25}{34+125}\right)^2$$

$$= 0.336$$

$$G(downtown) = 1 - P(yes|downtown)^2 - P(no|downtown)^2$$

$$= 1 - \left(\frac{37}{37+127}\right)^2 - \left(\frac{127}{37+127}\right)^2$$

$$= 0.349$$

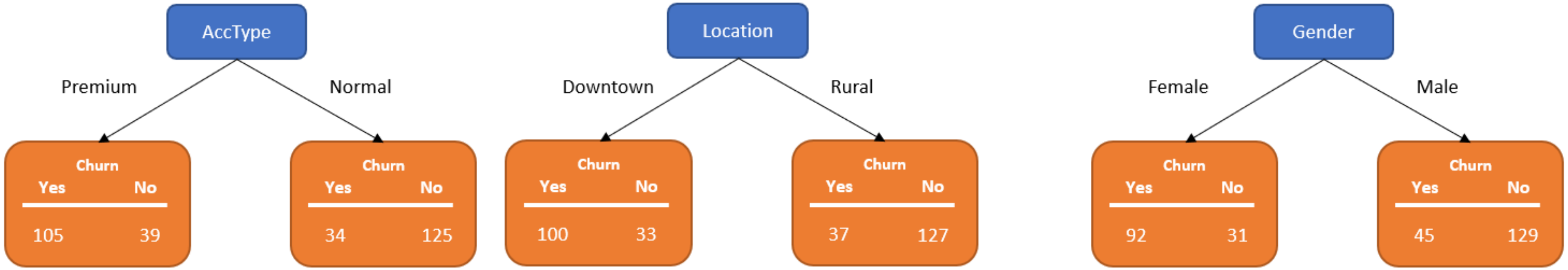
$$G(female) = 1 - P(yes|female)^2 - P(no|female)^2$$

$$= 1 - \left(\frac{45}{45+129}\right)^2 - \left(\frac{129}{45+129}\right)^2$$

$$= 0.383$$

Gini Impurity

- Calculation Example



$$G(acctype) = \left(\frac{144}{144+159}\right)(0.395) + \left(\frac{159}{144+159}\right)(0.336)$$
$$= 0.364$$

$$G(location) = \left(\frac{133}{133+164}\right)(0.373) + \left(\frac{164}{133+164}\right)(0.349)$$
$$= 0.360$$

$$G(gender) = \left(\frac{123}{123+174}\right)(0.377) + \left(\frac{174}{123+174}\right)(0.383)$$
$$= 0.381$$

Location has the lowest Gini Impurity. Therefore, select the location to split data

Tree Induction

- Step 1: Calculate Gini Impurity for all available features
- Step 2: Select the feature with lowest Gini Impurity to be the root node
- Step 3: Exclude the feature selected as the root node, for each branch, do
    - Step 3.1: Calculate Gini Impurity for all remain features
    - Step 3.2: Select the feature with lowest Gini Impurity to be the next internal node
    - Step 3.3: Exclude the selected feature from that remaining features of that branch
    - Step 3.4 Repeat 3.1 to 3.3 until satisfies stopping criteria

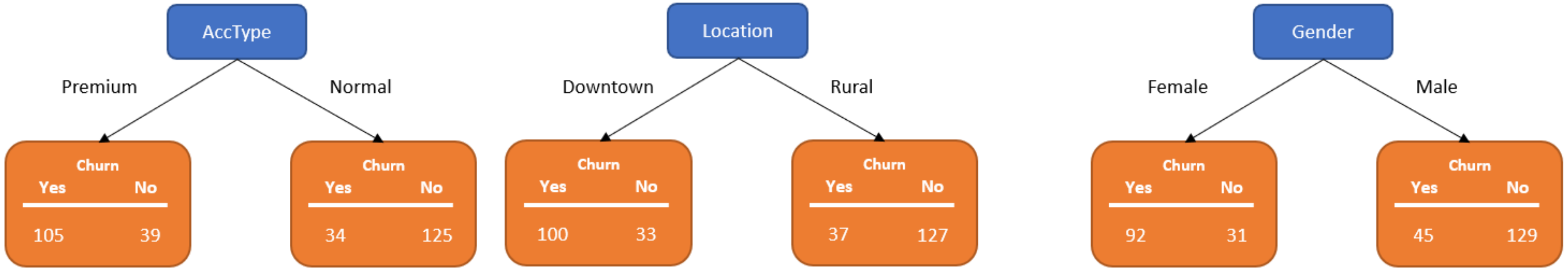
Tree Induction

- Example: Given the following data, create a decision tree

AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
…	…	…	…

Tree Induction

- Step 1: Calculate Gini Impurity for all available features
  - Assume that each node return the following split



$$G(acctype) = \left(\frac{144}{144+159}\right)(0.395) + \left(\frac{159}{144+159}\right)(0.336)$$
$$= 0.364$$

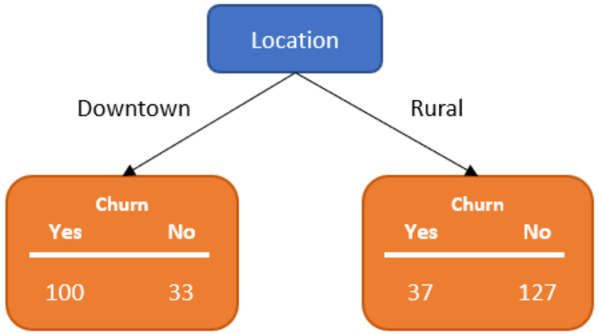
$$G(location) = \left(\frac{133}{133+164}\right)(0.373) + \left(\frac{164}{133+164}\right)(0.349)$$
$$= 0.360$$

$$G(gender) = \left(\frac{123}{123+174}\right)(0.377) + \left(\frac{174}{123+174}\right)(0.383)$$
$$= 0.381$$

Location has the lowest Gini Impurity. Therefore, select the location as the root node

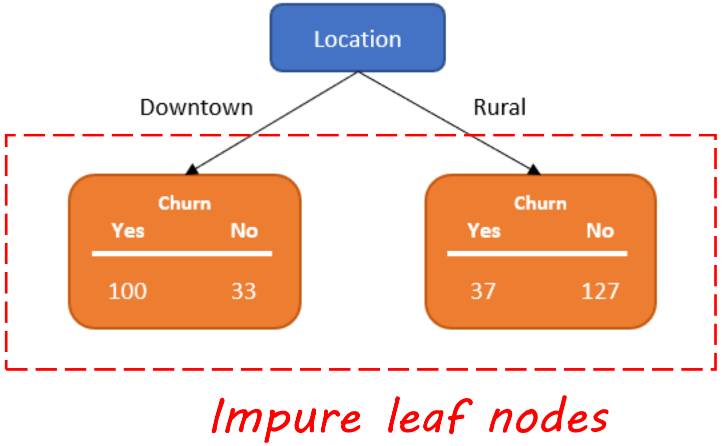
Tree Induction

- Step 2: Select the feature with lowest Gini Impurity to be the root node
  - Our current tree



Tree Induction

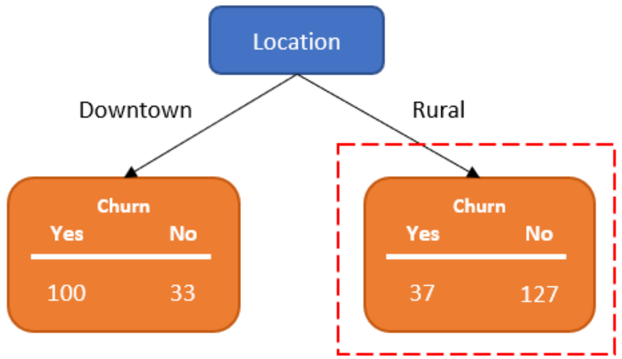
- Our current tree



Tree Induction

- Step 3: Exclude the feature selected as the root node, for each branch, do

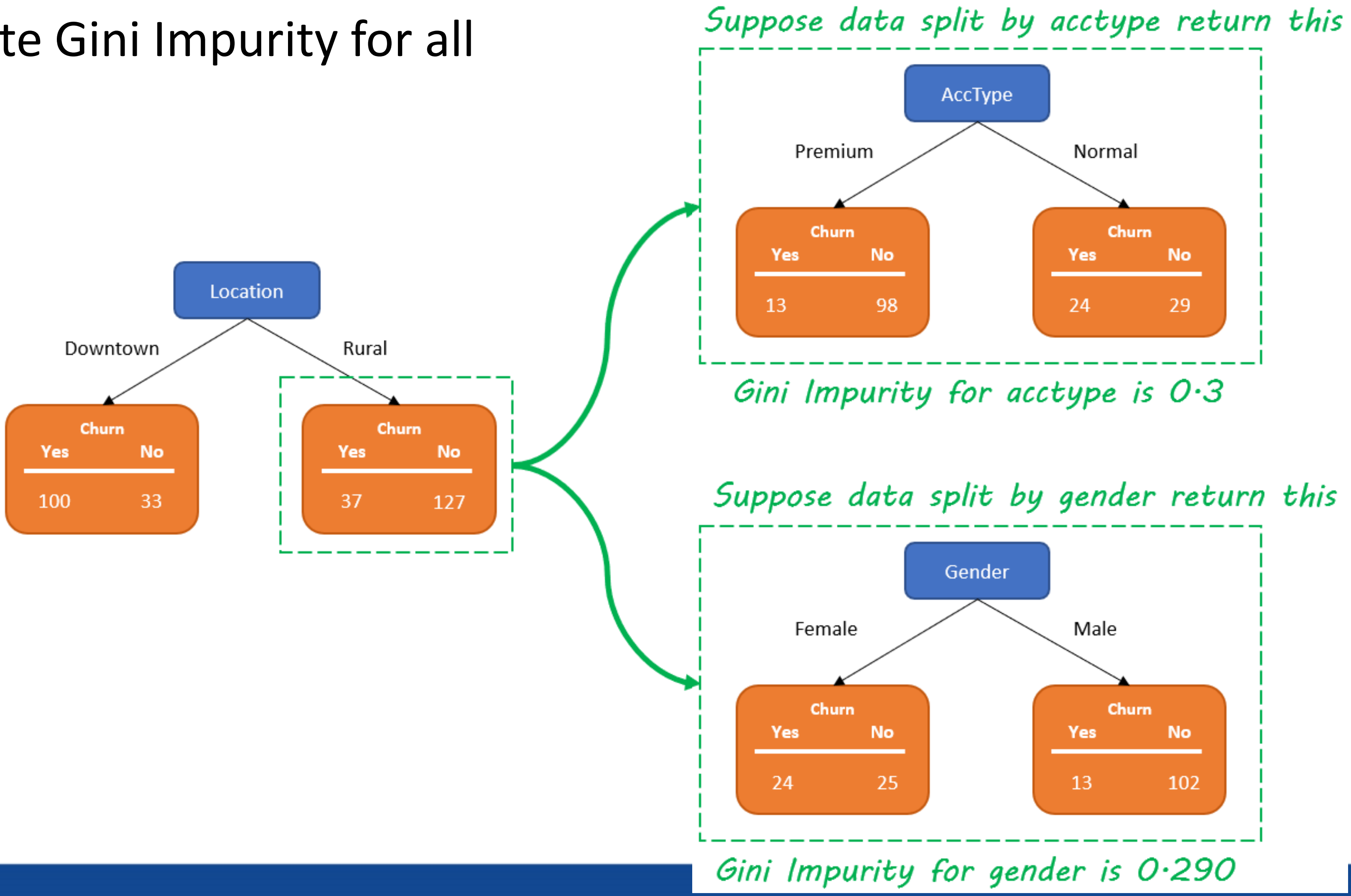
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Next step:
How well acctype and gender separate these 164 samples

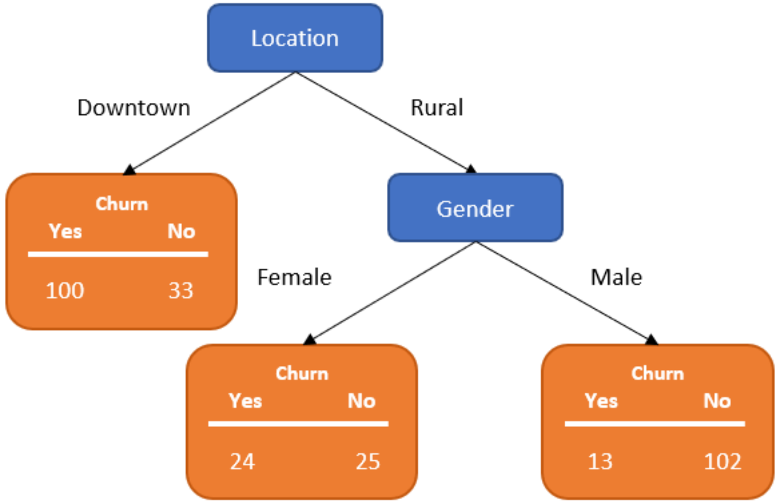
Tree Induction

- Step 3.1: Calculate Gini Impurity for all remain features



Tree Induction

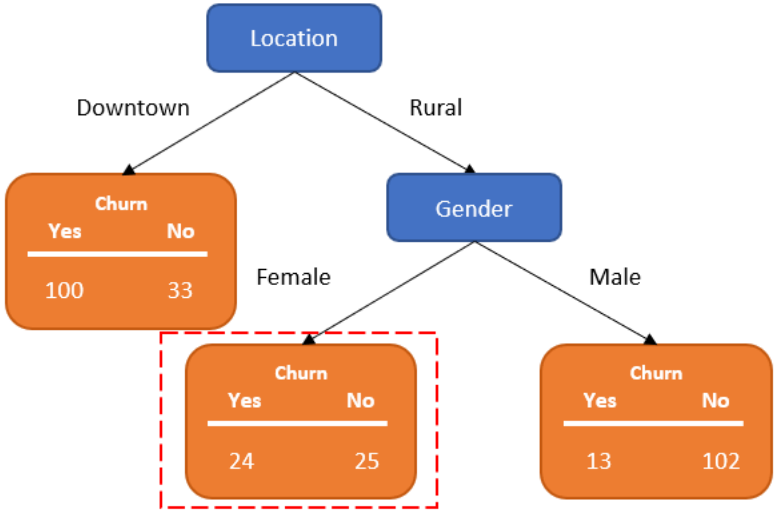
- Step 3.2: Select the feature with lowest Gini Impurity to be the next internal node
    - Our current tree



Tree Induction

- Step 3.3: Exclude the selected feature from that remaining features of that branch

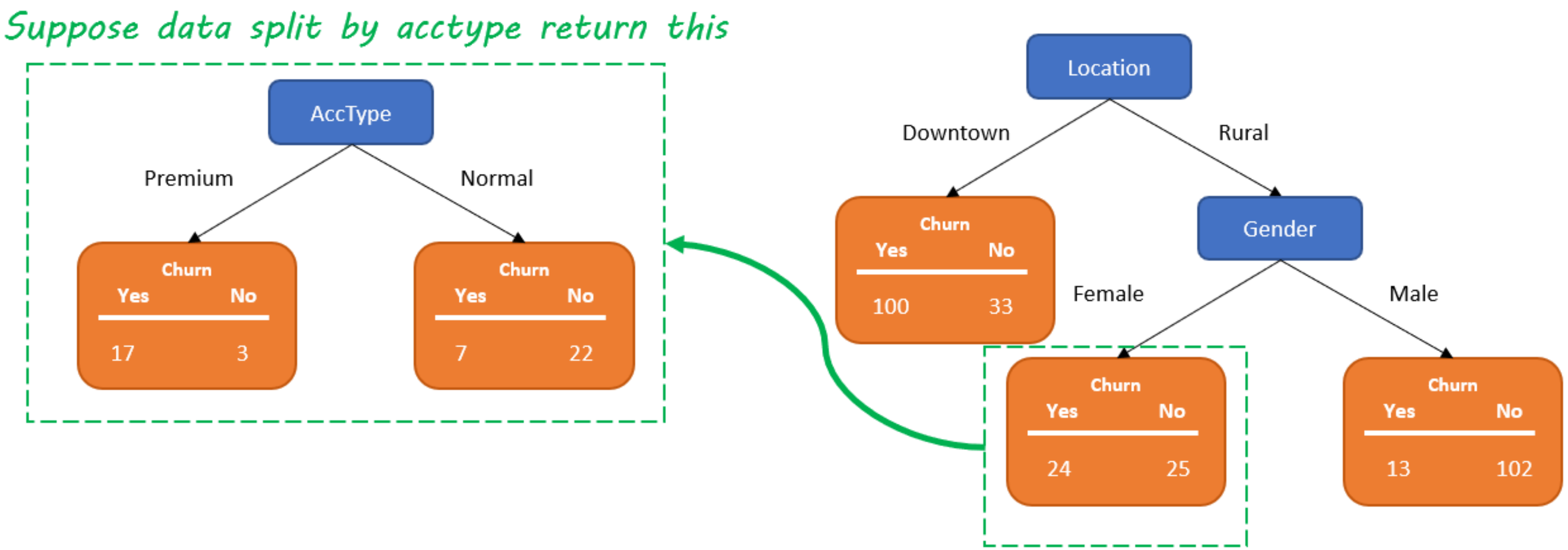
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



We have only acctype as an unused feature
How well the acctype these 49 samples?

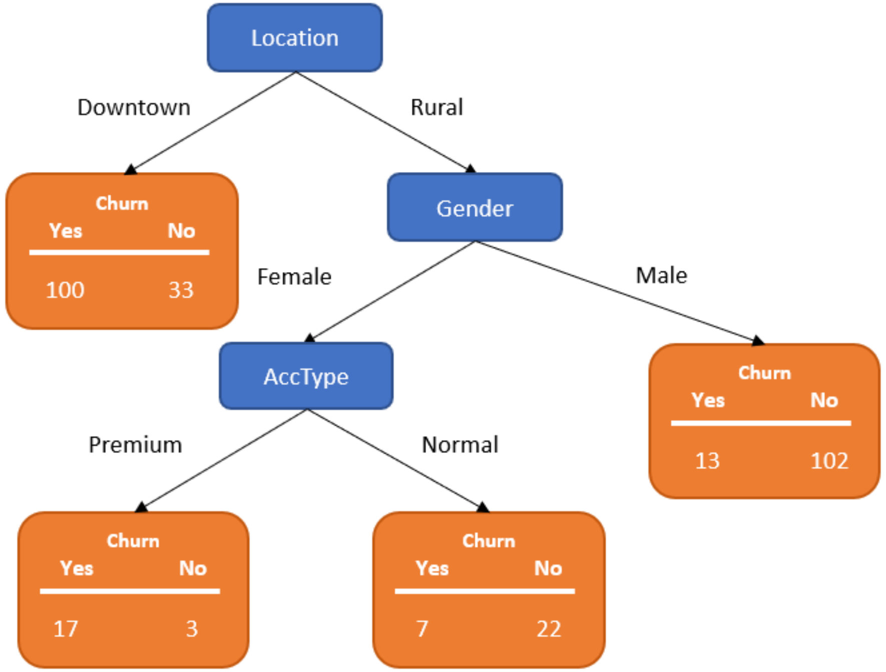
Tree Induction

- Step 3.1 (Repeat): Calculate Gini Impurity for all remain features



Tree Induction

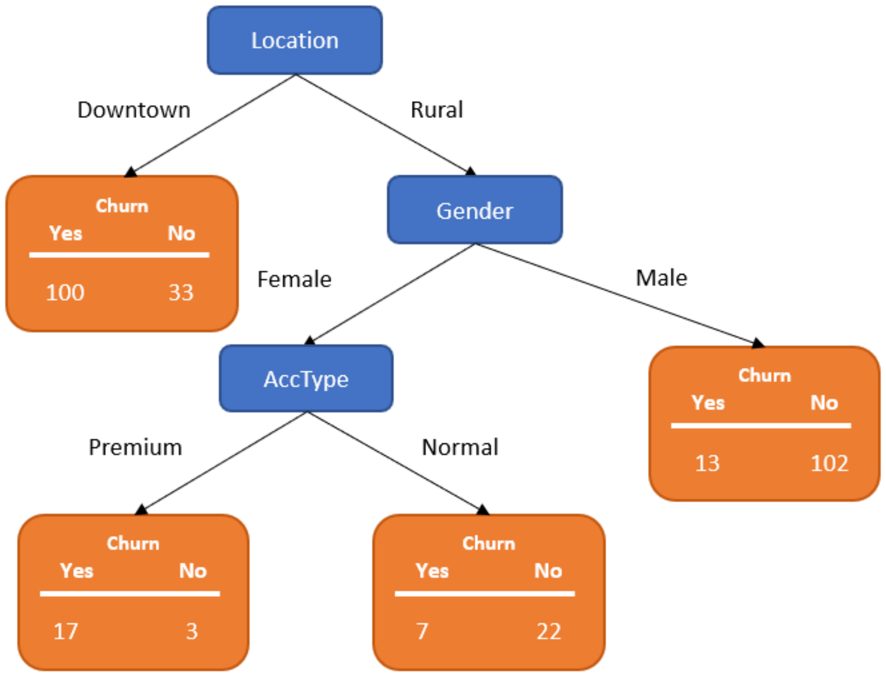
- Step 3.2 (Repeat): Select the feature with lowest Gini Impurity to be the next internal node
  - Our current tree



Tree Induction

- Step 3.3 (Repeat): Exclude the selected feature from that remaining features of that branch

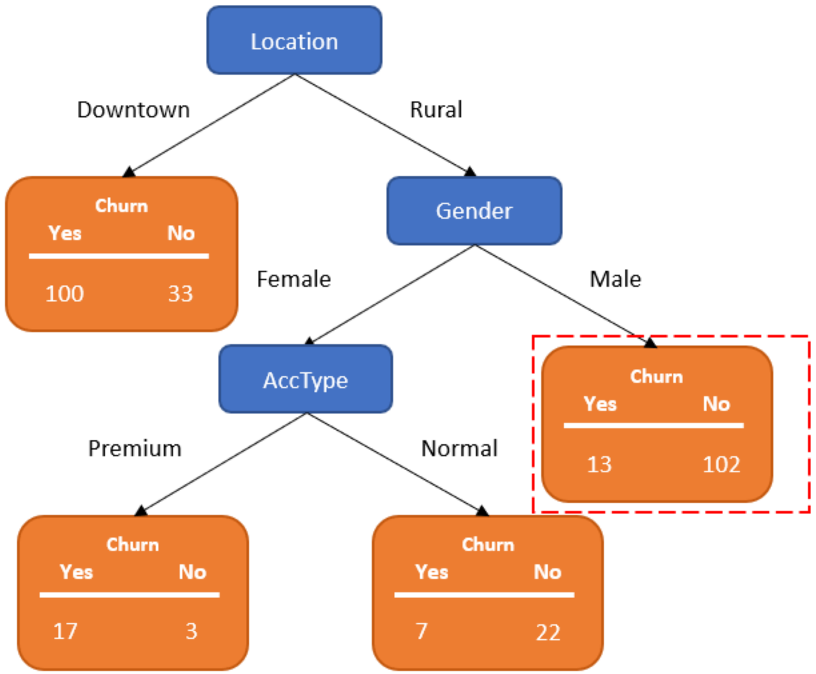
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



Tree Induction

- Step 3.1 (Repeat): Calculate Gini Impurity for all remain features

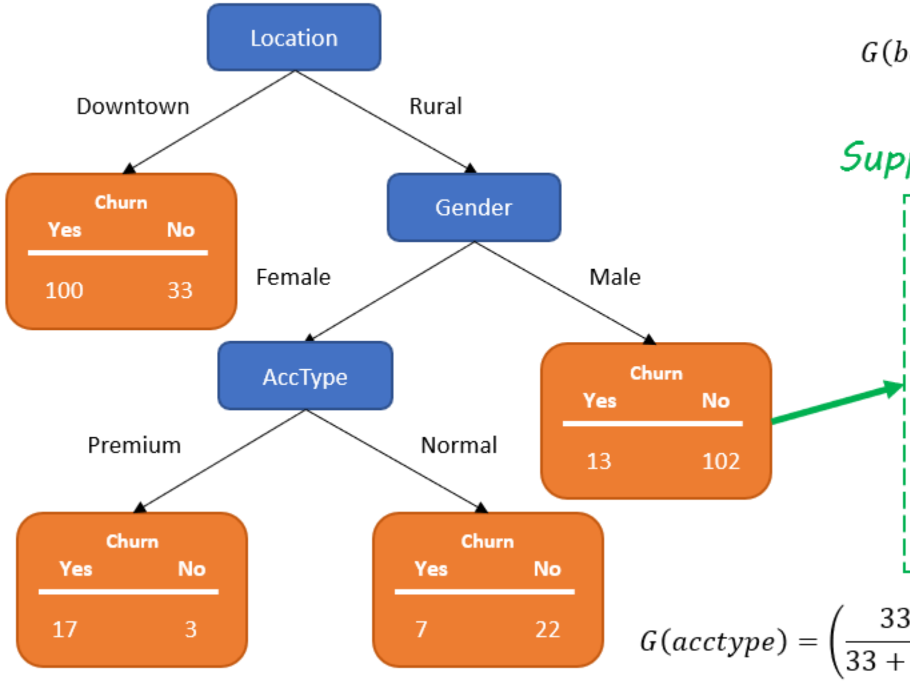
AccType	Location	Gender	Churn
Normal	Rural	Male	No
Premium	Downtown	Female	Yes
Premium	Downtown	Male	No
Premium	Rural	?	Yes
...	...	...	...



What if we use the acctype to separate these 115 samples?

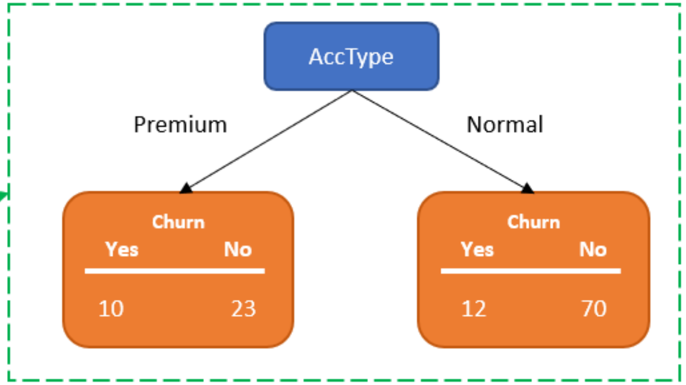
Tree Induction

- Step 3.2 (Repeat): Select the feature with lowest Gini Impurity to be the next internal node



$$G(before) = 1 - \left(\frac{13}{13+102}\right)^2 - \left(\frac{102}{13+102}\right)^2 = 0.200$$

Suppose data split by acctype return this



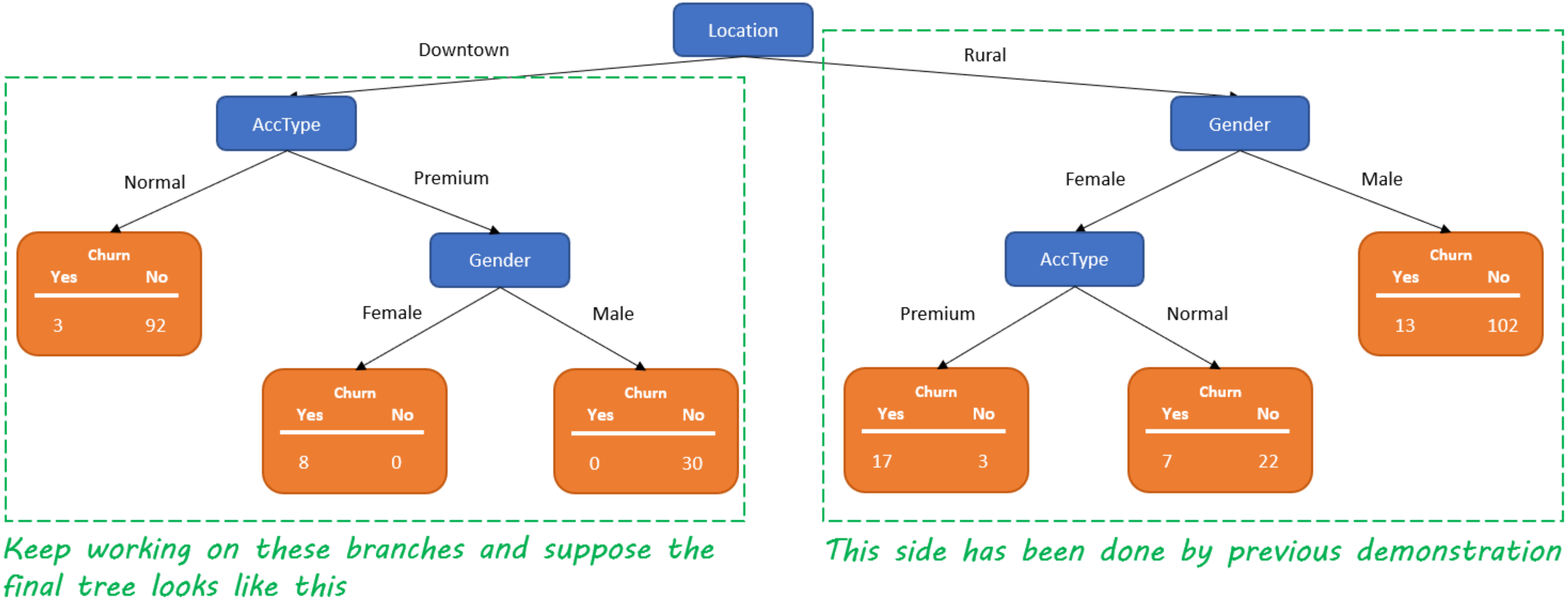
$$G(acctype) = \left(\frac{33}{33+82}\right)\left[1 - \left(\frac{10}{10+23}\right)^2 - \left(\frac{23}{10+23}\right)^2\right] + \left(\frac{82}{33+82}\right)\left[1 - \left(\frac{12}{12+70}\right)^2 - \left(\frac{70}{12+70}\right)^2\right]$$

$$= 0.299$$

$G(before) < G(acctype)$ *then stop splitting*

Three Induction

- Repeating step 3.1 to 3.3 for the left branch



Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
  - Step 1: Sort $\mathbf{x}$

ncmps	Churn
8	Yes
5	Yes
10	Yes
7	No
3	No

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
    - Step 2: Calculate the middle points between rows

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

4.0
6.0
7.5
9.0

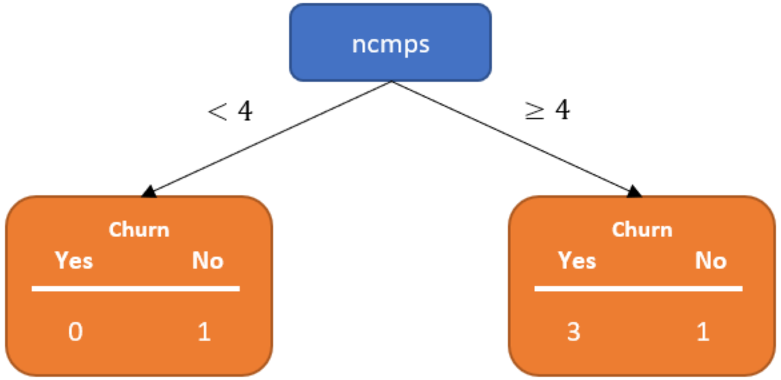
Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
  - Step 3: Calculate Gini Impurity for each middle points

	ncmps	Churn
	3	No
Gini = ?		
	5	Yes
Gini = ?		
	7	No
Gini = ?		
	8	Yes
Gini = ?		
	10	Yes

Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
    - Step 3: Calculate Gini Impurity for each middle points



$$G(ncmps) = \left(\frac{1}{1+4}\right)\left[1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2\right] + \left(\frac{4}{1+4}\right)\left[1 - \left(\frac{3}{3+1}\right)^2 - \left(\frac{1}{3+1}\right)^2\right]$$

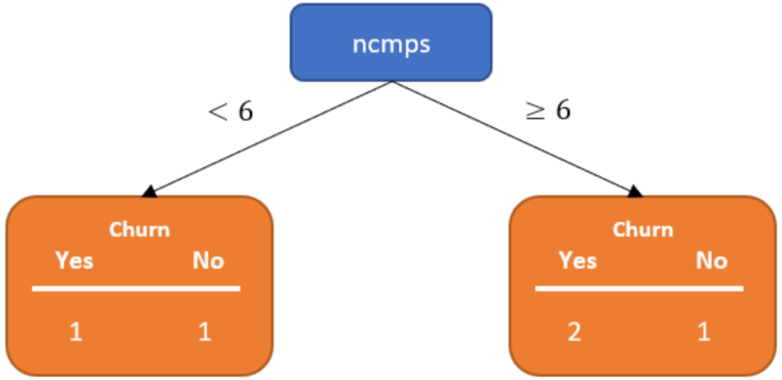
$$= 0.300$$

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

4.0
6.0
7.5
9.0

Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
  - Step 3: Calculate Gini Impurity for each middle points



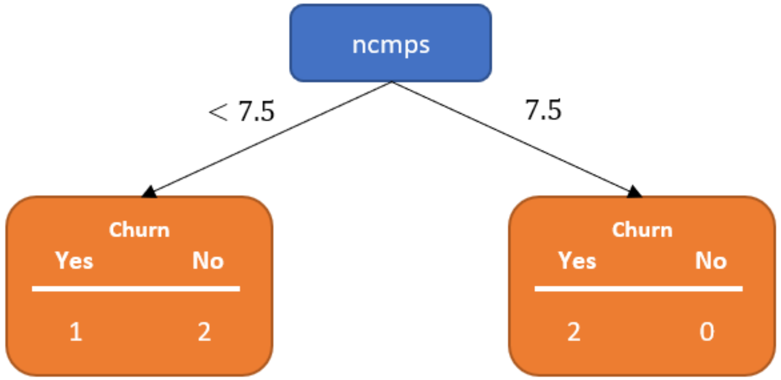
$$G(ncmps) = \left(\frac{2}{2+3}\right)\left[1-\left(\frac{1}{1+1}\right)^2-\left(\frac{1}{1+1}\right)^2\right]+\left(\frac{3}{2+3}\right)\left[1-\left(\frac{2}{2+1}\right)^2-\left(\frac{1}{2+1}\right)^2\right]$$
$$= 0.467$$

4.0
6.0
7.5
9.0

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
    - Step 3: Calculate Gini Impurity for each middle points



$$G(ncmps) = \left(\frac{3}{3+2}\right)\left[1 - \left(\frac{1}{1+2}\right)^2 - \left(\frac{2}{1+2}\right)^2\right] + \left(\frac{2}{3+2}\right)\left[1 - \left(\frac{2}{2+0}\right)^2 - \left(\frac{0}{2+0}\right)^2\right]$$

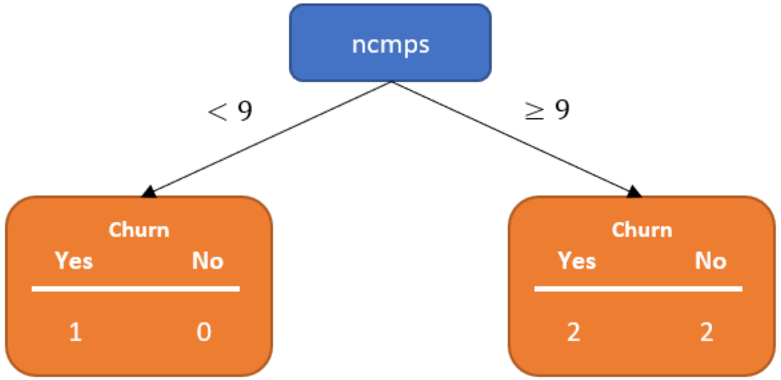
$$= 0.267$$

4.0
6.0
7.5
9.0

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

Dealing with Numerical Feature

- Let **x** be any numerical feature
    - Step 3: Calculate Gini Impurity for each middle points



$$G(ncmps) = \left(\frac{1}{1+4}\right)\left[1 - \left(\frac{1}{1+0}\right)^2 - \left(\frac{0}{1+0}\right)^2\right] + \left(\frac{4}{1+4}\right)\left[1 - \left(\frac{2}{2+2}\right)^2 - \left(\frac{2}{2+2}\right)^2\right]$$

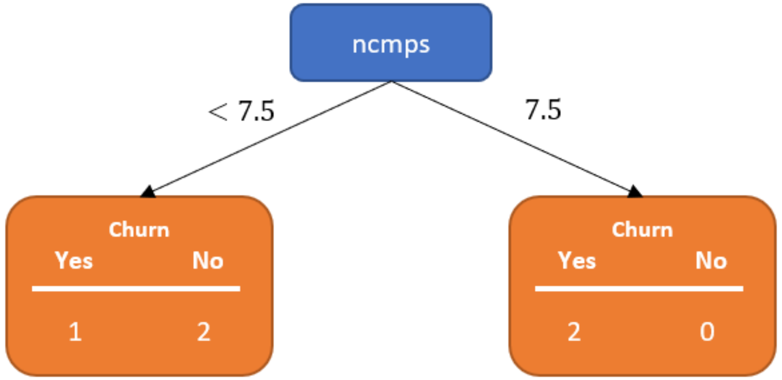
$$= 0.400$$

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

Middle points: 4.0, 6.0, 7.5, **9.0**

Dealing with Numerical Feature

- Let $\mathbf{x}$ be any numerical feature
    - Step 4: Choose the cut point that returns the lowest Gini Impurity



$4.0, G(ncmps) = 0.300$

$6.0, G(ncmps) = 0.467$

$7.5, G(ncmps) = 0.267$

$9.0, G(ncmps) = 0.400$

ncmps	Churn
3	No
5	Yes
7	No
8	Yes
10	Yes

Choose 7·5 as the cutoff value because it returns the lowest Gini Impurity

Dealing with Ordinal Feature

- If $\mathbf{x}$ is an ordinal feature
  - Process the same as a numerical feature excepts
    - Finding middle values is not required
    - Gini impurity measure for the largest value is not required if $\geq$ is used
    - Gini impurity measure for the smallest value is not required if $\leq$ is used

Dealing with Ordinal Feature

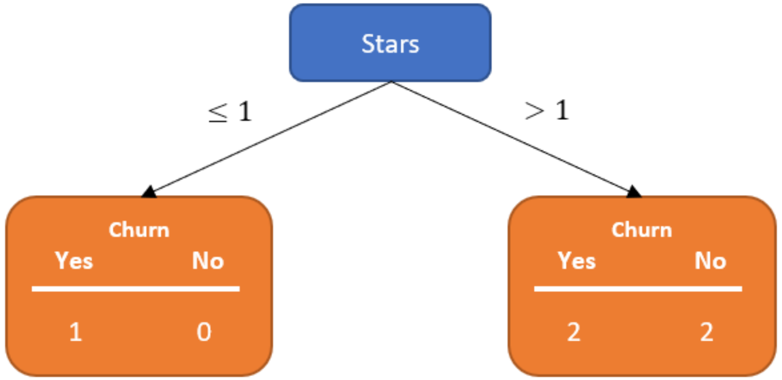
- Let $\mathbf{x}$ be any ordinal feature
  - Step 1: Sort $\mathbf{x}$

Stars	Churn
1	Yes
2	Yes
3	Yes
4	No
5	No

Dealing with Ordinal Feature

- Let $\mathbf{x}$ be any ordinal feature
  - Step 2: Calculate Gini Impurity for each x

Stars	Churn
1	Yes
2	Yes
3	Yes
4	No
5	No



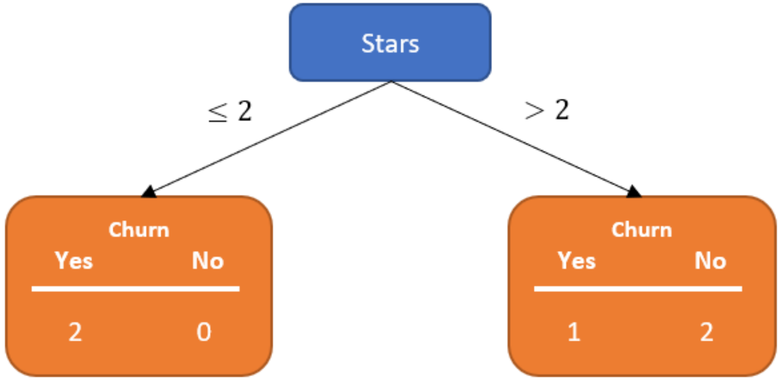
$$\mathrm{G}(stars) = \left(\frac{1}{1+4}\right)\left[1 - \left(\frac{1}{1+0}\right)^2 - \left(\frac{0}{1+0}\right)^2\right] + \left(\frac{4}{1+4}\right)\left[1 - \left(\frac{2}{2+2}\right)^2 - \left(\frac{2}{2+2}\right)^2\right]$$

$$= 0.400$$

Dealing with Ordinal Feature

- Let $\mathbf{x}$ be any ordinal feature
    - Step 2: Calculate Gini Impurity for each x

Stars	Churn
1	Yes
2	Yes
3	Yes
4	No
5	No



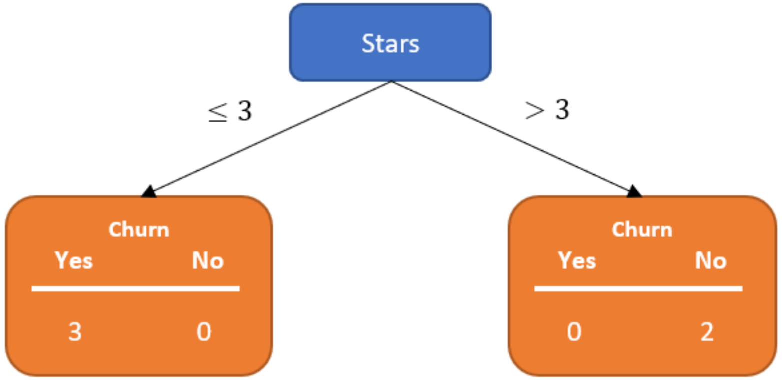
$$G(stars) = \left(\frac{2}{2+3}\right)\left[1 - \left(\frac{2}{2+0}\right)^2 - \left(\frac{0}{2+0}\right)^2\right] + \left(\frac{3}{2+3}\right)\left[1 - \left(\frac{1}{1+2}\right)^2 - \left(\frac{1}{1+2}\right)^2\right]$$

$$= 0.266$$

Dealing with Ordinal Feature

- Let $\mathbf{x}$ be any ordinal feature
  - Step 2: Calculate Gini Impurity for each x

Stars	Churn
1	Yes
2	Yes
3	Yes
4	No
5	No



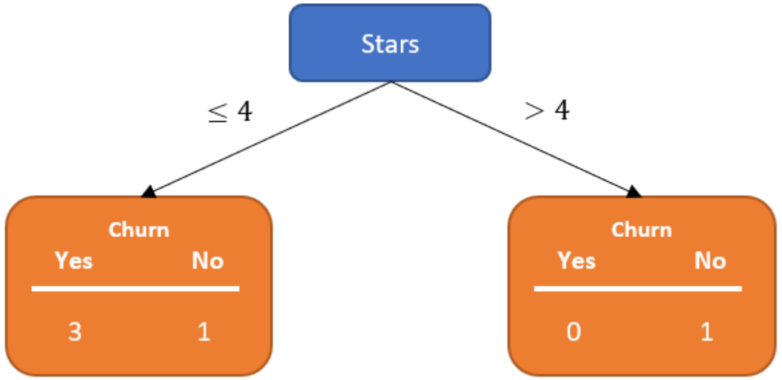
$$G(stars) = \left(\frac{3}{3+2}\right)\left[1 - \left(\frac{3}{3+0}\right)^2 - \left(\frac{0}{3+0}\right)^2\right] + \left(\frac{2}{3+2}\right)\left[1 - \left(\frac{0}{0+2}\right)^2 - \left(\frac{0}{0+2}\right)^2\right]$$

$$= 0.000$$

Dealing with Ordinal Feature

- Let $\mathbf{x}$ be any ordinal feature
  - Step 2: Calculate Gini Impurity for each x

Stars	Churn
1	Yes
2	Yes
3	Yes
4	No
5	No



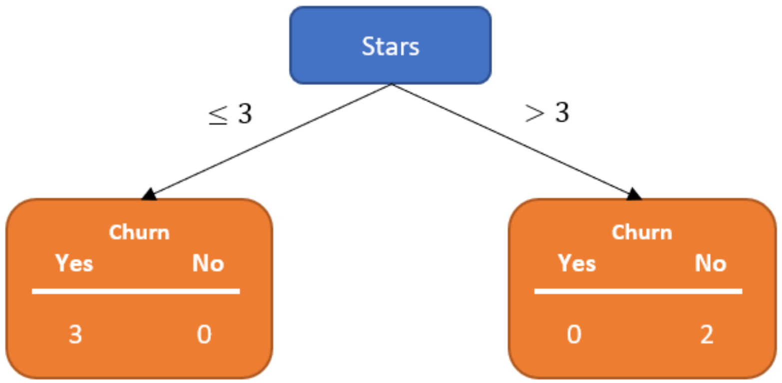
$$G(stars) = \left(\frac{4}{4+1}\right)\left[1 - \left(\frac{3}{3+1}\right)^2 - \left(\frac{1}{3+1}\right)^2\right] + \left(\frac{1}{4+1}\right)\left[1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2\right]$$

$$= 0.300$$

Dealing with Ordinal Feature

- Let $\mathbf{x}$ be any ordinal feature
  - Step 2: Calculate Gini Impurity for each x

	Stars	Churn
$\mathrm{G}(stars) = 0.400$	1	Yes
$\mathrm{G}(stars) = 0.266$	2	Yes
$\mathrm{G}(stars) = 0.000$	3	Yes
$\mathrm{G}(stars) = 0.300$	4	No
	5	No

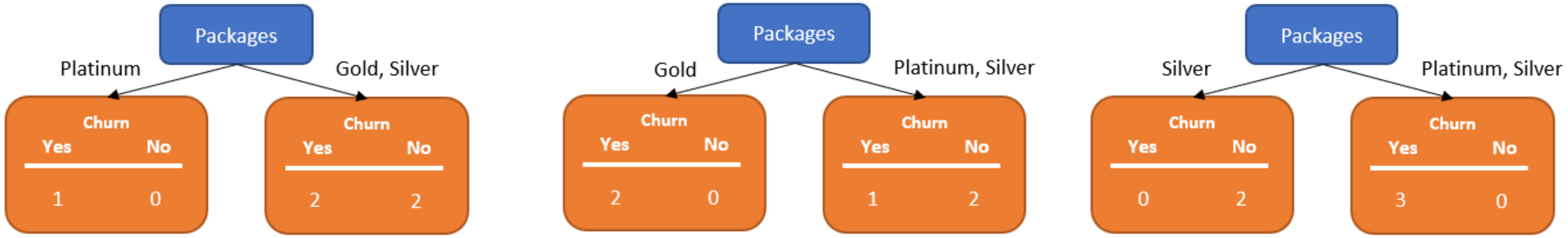


Choose 3 as the cutoff value because it returns the lowest Gini Impurity

Dealing with Categorical Feature

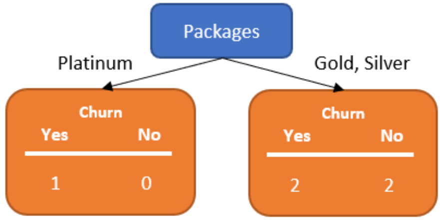
- Let $\mathbf{x}$ be any categorical feature
  - Step 1: Create all possible combinations

Packages	Churn
Gold	Yes
Platinum	Yes
Gold	Yes
Silver	No
Silver	No



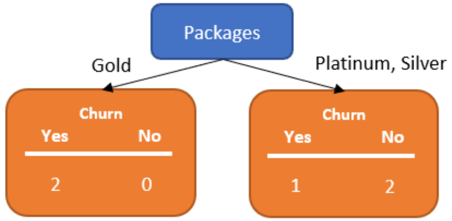
Dealing with Categorical Feature

- Let $\mathbf{x}$ be any categorical feature
  - Step 2: Choose the combination that returns the lowest Gini Impurity



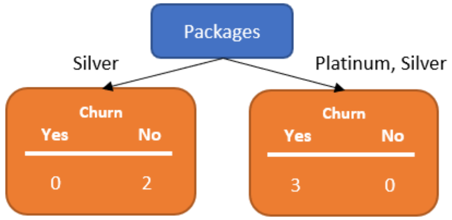
$$G(packages) = \left(\frac{1}{1+4}\right)\left[1-\left(\frac{1}{1+0}\right)^2-\left(\frac{0}{1+0}\right)^2\right]+\left(\frac{4}{1+4}\right)\left[1-\left(\frac{2}{2+2}\right)^2-\left(\frac{2}{2+2}\right)^2\right]$$

$$= 0.400$$



$$G(packages) = \left(\frac{2}{2+3}\right)\left[1-\left(\frac{2}{2+0}\right)^2-\left(\frac{0}{2+0}\right)^2\right]+\left(\frac{3}{2+3}\right)\left[1-\left(\frac{1}{1+2}\right)^2-\left(\frac{2}{1+2}\right)^2\right]$$

$$= 0.266$$



$$G(packages) = \left(\frac{2}{2+3}\right)\left[1-\left(\frac{0}{0+2}\right)^2-\left(\frac{2}{0+2}\right)^2\right]+\left(\frac{3}{2+3}\right)\left[1-\left(\frac{3}{3+0}\right)^2-\left(\frac{0}{3+0}\right)^2\right]$$

$$= 0.000$$

Choose this combination because it returns the lowest Gini Impurity

Summary

- The key concepts to take away
  - Decision tree can take any data type
  - Decision tree skips counting missing value is detected
  - Splitting relies on impurity measure
  - The best split returns the most purity