Introduction to Machine Learning

Class

Tree Based Models

Topic

Introduction to Regression Tree

Decision Tree: Regression

Decision Tree can be used to do regression tasks

When the target variable is continuous decision tree regressor can be used

Prediction → Mean value of the target variable

Decision Tree: Regression

Example

Country	Rim	Tires	Type	Price
Japan	R14	195/60	Small	11.95
Japan	R15	205/60	Medium	24.76
Germany	R15	205/60	Medium	26.9
Germany	R14	175/60	Compact	18.9
Germany	R14	195/60	Compact	24.65
Germany	R15	225/60	Medium	33.2
USA	R14	185/75	Medium	13.15
USA	R14	205/75	Large	20.225
USA	R14	205/75	Large	16.145
USA	R15	205/70	Medium	23.04

Build a decision tree model to predict price

Price is a continuous variable

Regression tree

Recursively subset the data

Decision Tree: Regression

Example

Country	Rim	Tires	Type	Price
Japan	R14	195/60	Small	11.95
Japan	R15	205/60	Medium	24.76
Germany	R15	205/60	Medium	26.9
Germany	R14	175/60	Compact	18.9
Germany	R14	195/60	Compact	24.65
Germany	R15	225/60	Medium	33.2
USA	R14	185/75	Medium	13.15
USA	R14	205/75	Large	20.225
USA	R14	205/75	Large	16.145
USA	R15	205/70	Medium	23.04

Total Population =10
Average price = 21.9

R14

Yes

Total Population = 6
Average Price= 17.50

Price
11.95
18.9
24.65
13.15
20.22
16.14

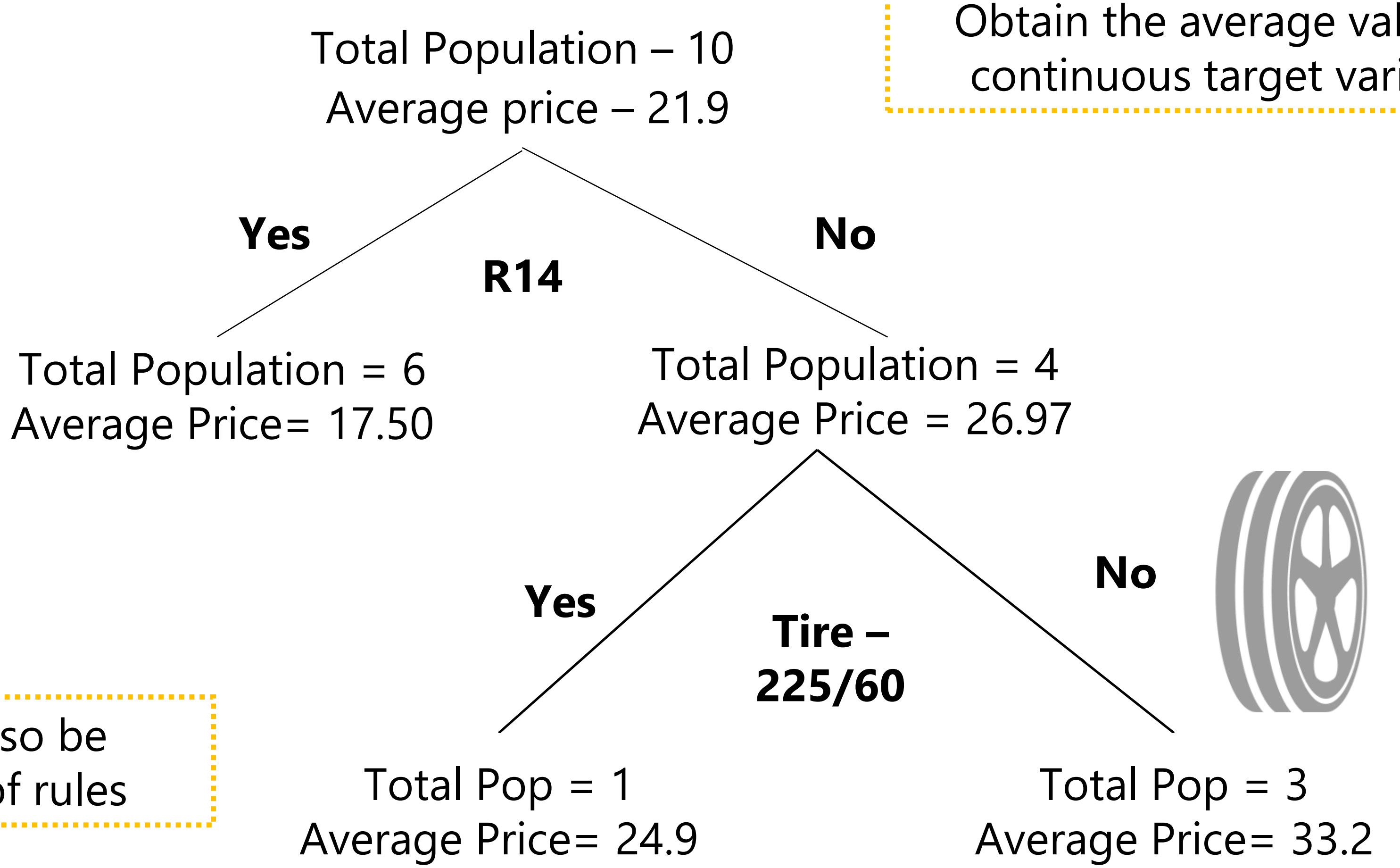
No

Total Population = 4
Average Price = 26.97

Price
24.76
26.90
33.20
23.04

Decision Tree: Regression

Obtain the average value of continuous target variable



A regression tree can also be summarised as a series of rules

Purity Metrics

How does a regression tree algorithm pick up which variable to split on?

Predictions need to be accurate

The prediction is the average value of target variable in decision node

Higher the accuracy of prediction, the better the split is

Mean Squared Error (MSE) or Residual Sum of Square (RSS) as a proxy of accuracy in each node

Purity Metrics

Country	Rim	Tires	Type	Price
Japan	R14	195/60	Small	11.95
Japan	R15	205/60	Medium	24.76
Germany	R15	205/60	Medium	26.9
Germany	R14	175/60	Compact	18.9
Germany	R14	195/60	Compact	24.65
Germany	R15	225/60	Medium	33.2
USA	R14	185/75	Medium	13.15
USA	R14	205/75	Large	20.225
USA	R14	205/75	Large	16.145
USA	R15	205/70	Medium	23.04

MSE or RSS helps in deciding which variable to choose for a split

Purity Metrics

Country	Rim	Tires	Type	Price
Japan	R14	195/60	Small	11.95
Japan	R15	205/60	Medium	24.76
Germany	R15	205/60	Medium	26.9
Germany	R14	175/60	Compact	18.9
Germany	R14	195/60	Compact	24.65
Germany	R15	225/60	Medium	33.2
USA	R14	185/75	Medium	13.15
USA	R14	205/75	Large	20.225
USA	R14	205/75	Large	16.145
USA	R15	205/70	Medium	23.04

Total Population =10
Average price = 21.9

R14

Yes

Total Population = 6
Average Price= 17.50

Price
11.95
18.9
24.65
13.15
20.22
16.14

No

Total Population = 4
Average Price = 26.97

Price
24.76
26.90
33.20
23.04

Purity Metrics

Country	Rim	Tires	Type	Price
Japan	R14	195/60	Small	11.95
Japan	R15	205/60	Medium	24.76
Germany	R15	205/60	Medium	26.9
Germany	R14	175/60	Compact	18.9
Germany	R14	195/60	Compact	24.65
Germany	R15	225/60	Medium	33.2
USA	R14	185/75	Medium	13.15
USA	R14	205/75	Large	20.225
USA	R14	205/75	Large	16.145
USA	R15	205/70	Medium	23.04

Total Population =10
Average price = 21.9

Country - Germany

Yes

Total Population = 4
Average Price= 25.91

Price
26.90
18.90
26.45
33.20

No

Total Population = 6
Average Price = 18.21

Price
11.95
24.76
13.15
20.22
16.14
23.04

Purity Metric

Total Population =10
Average price = 21.9

R14

Yes

Total Population = 6
Average Price= 17.50

No

Total Population = 4
Average Price = 26.97

Total Population =10
Average price = 21.9

Country - Germany

Yes

Total Population = 4
Average Price= 25.91

No

Total Population = 6
Average Price = 18.21

Rim or Country?

Which variable helps in creating a more accurate prediction?

Purity Metric

Total Population =10
Average price = 21.9

Yes **R14** **No**

Total Population = 6
Average Price= 17.50

Total Population = 4
Average Price = 26.97

Total Population =10
Average price = 21.9

Yes **Country - Germany** **No**

Total Population = 4
Average Price= 25.91

Total Population = 6
Average Price = 18.21

Use Mean Squared Error (MSE) or
Residual Sum of Square (RSS)

$$MSE = \frac{1}{n}\sum(y_i - \mu)^2$$

MSE is just the average of RSS

Nothing but variance in the values of
target in variable in a node

Purity Metric

Total Population =10
Average price = 21.9

R14

Yes

Total Population = 6
Average Price= 17.50

Price	Pred.
11.95	17.50
18.9	17.50
24.65	17.50
13.15	17.50
20.22	17.50
16.14	17.50

No

Total Population = 4
Average Price = 26.97

Price	Pred.
24.76	26.97
26.90	26..97
33.20	26.97
23.04	26.97

Total Population =10
Average price = 21.9

Country - Germany

Yes

Total Population = 4
Average Price= 25.91

Price	Pred.
26.90	25.91
18.90	25.91
26.45	25.91
33.20	25.91

No

Total Population = 6
Average Price = 18.21

Price	Pred.
11.95	18.21
24.76	18.21
13.15	18.21
20.22	18.21
16.14	18.21
23.04	18.21

MSE tries to find out how accurate a prediction is in each node

Purity Metric

$$MSE = \frac{1}{n}\sum(y_i - \mu)^2$$

Total Population =10
Average price = 21.9

R14

Yes

Total Population = 6
Average Price= 17.50

Price	Pred.
11.95	17.50
18.9	17.50
24.65	17.50
13.15	17.50
20.22	17.50
16.14	17.50

$$\frac{1}{6}(11.95 - 17.50)^2 + (18.90 - 17.50)^2 + .. + (16.14 - 17.50)^2$$

No

Total Population = 4
Average Price = 26.97

Price	Pred.
24.76	26.97
26.90	26..97
33.20	26.97
23.04	26.97

$$\frac{1}{4}(24.76 - 26.97)^2 + (26.90 - 26.97)^2 + .. + (23.04 - 26.97)^2$$

Total Population =10
Average price = 21.9

Country - Germany

Yes

Total Population = 4
Average Price= 25.91

Price	Pred.
26.90	25.91
18.90	25.91
26.45	25.91
33.20	25.91

No

Total Population = 6
Average Price = 18.21

Price	Pred.
11.95	18.21
24.76	18.21
13.15	18.21
20.22	18.21
16.14	18.21
23.04	18.21

Purity Metric

$$MSE = \frac{1}{n}\sum(y_i - \mu)^2$$

R14

Total Population = 10
Average price = 21.9

Yes

Total Population = 6
Average Price = 17.50

MSE – 18.67

No

Total Population = 4
Average Price = 26.97

MSE – 14.78

$$\frac{6}{10} * 18.67 + \frac{4}{10} * 14.78 \quad \boxed{= 17.114}$$

Country - Germany

Total Population = 10
Average price = 21.9

Yes

Total Population = 4
Average Price = 25.91

MSE – 26.21

No

Total Population = 6
Average Price = 18.21

MSE – 23.22

$$\frac{4}{10} * 26.21 + \frac{6}{10} * 23.22 \quad = 24.416$$

Rim is better than country at producing more accurate predictions

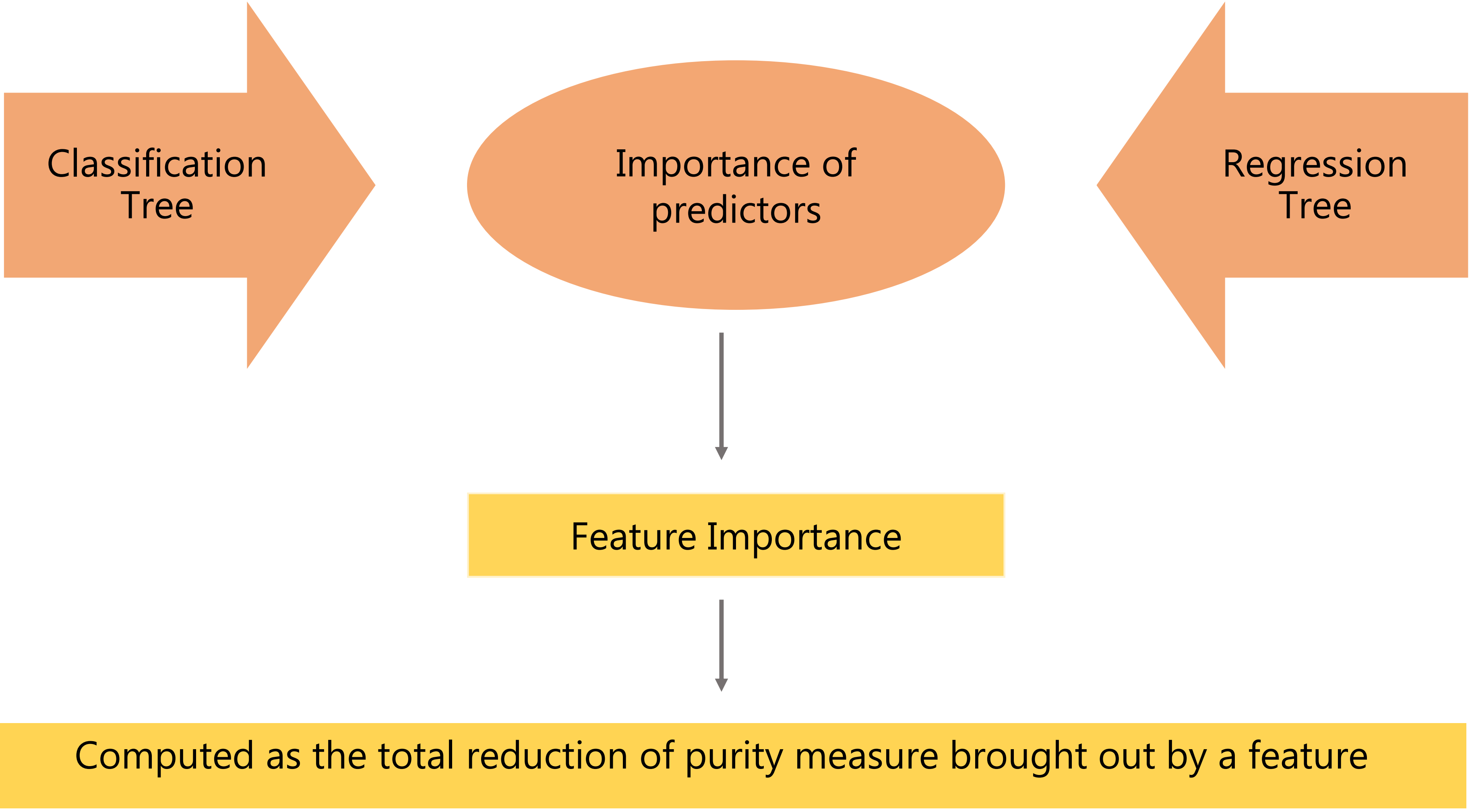
Hyperparameters

Regression Tree

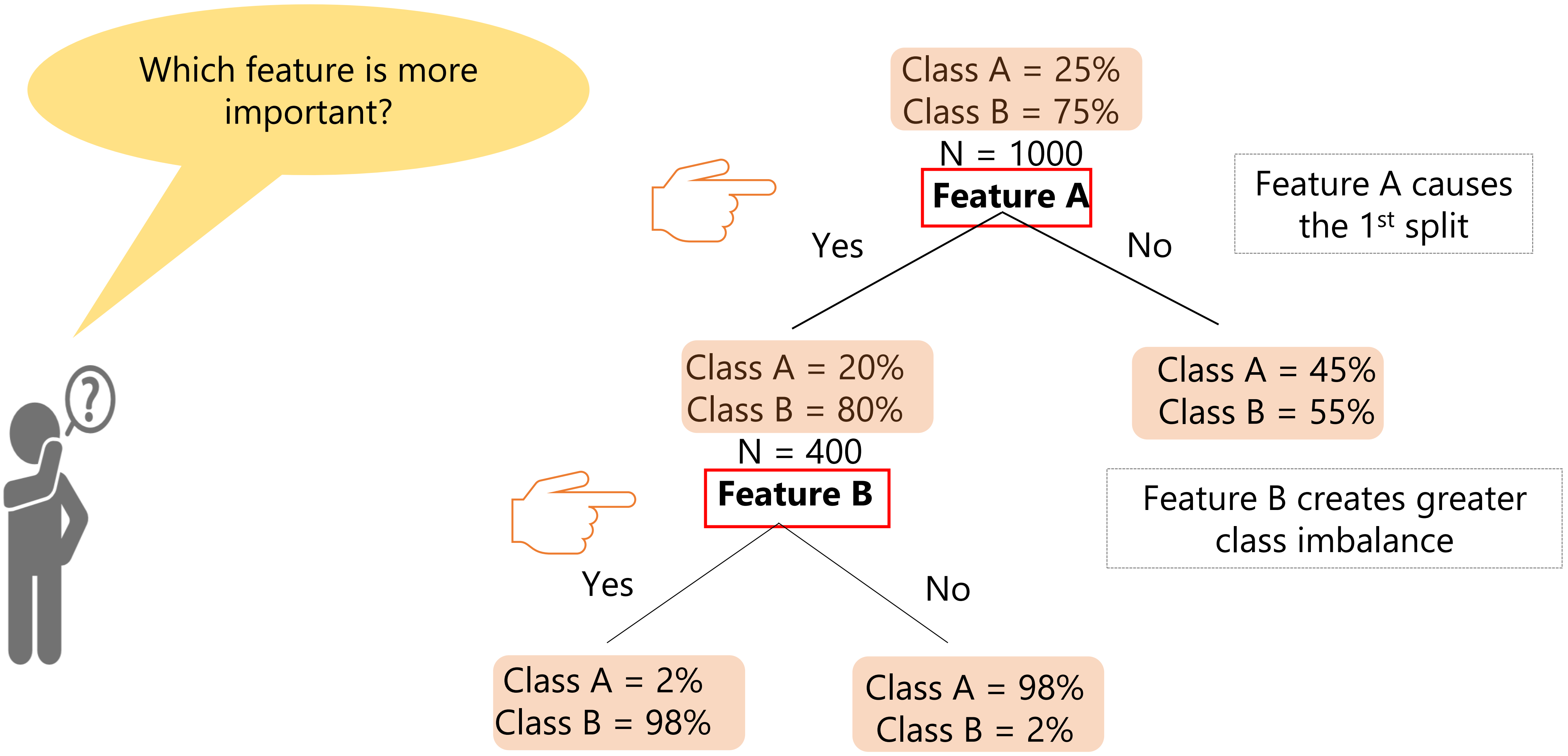
- Depth of tree
- Number of observations in terminal node

Grid search procedure to compute the appropriate values of these hyperparameters

Feature Importance

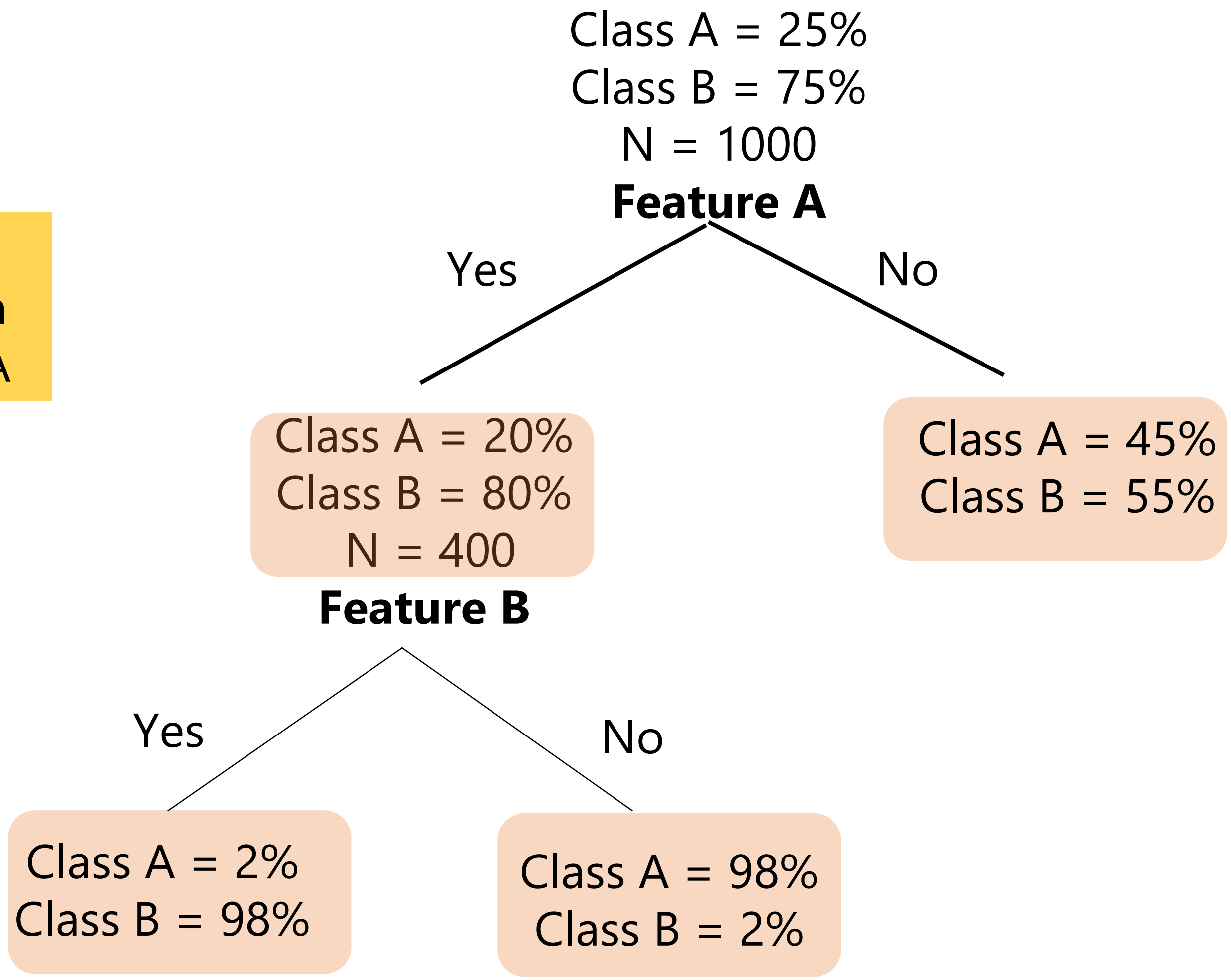


Feature Importance



Feature Importance

Proportion of classes are **more disproportionate** in Feature B than in Feature A

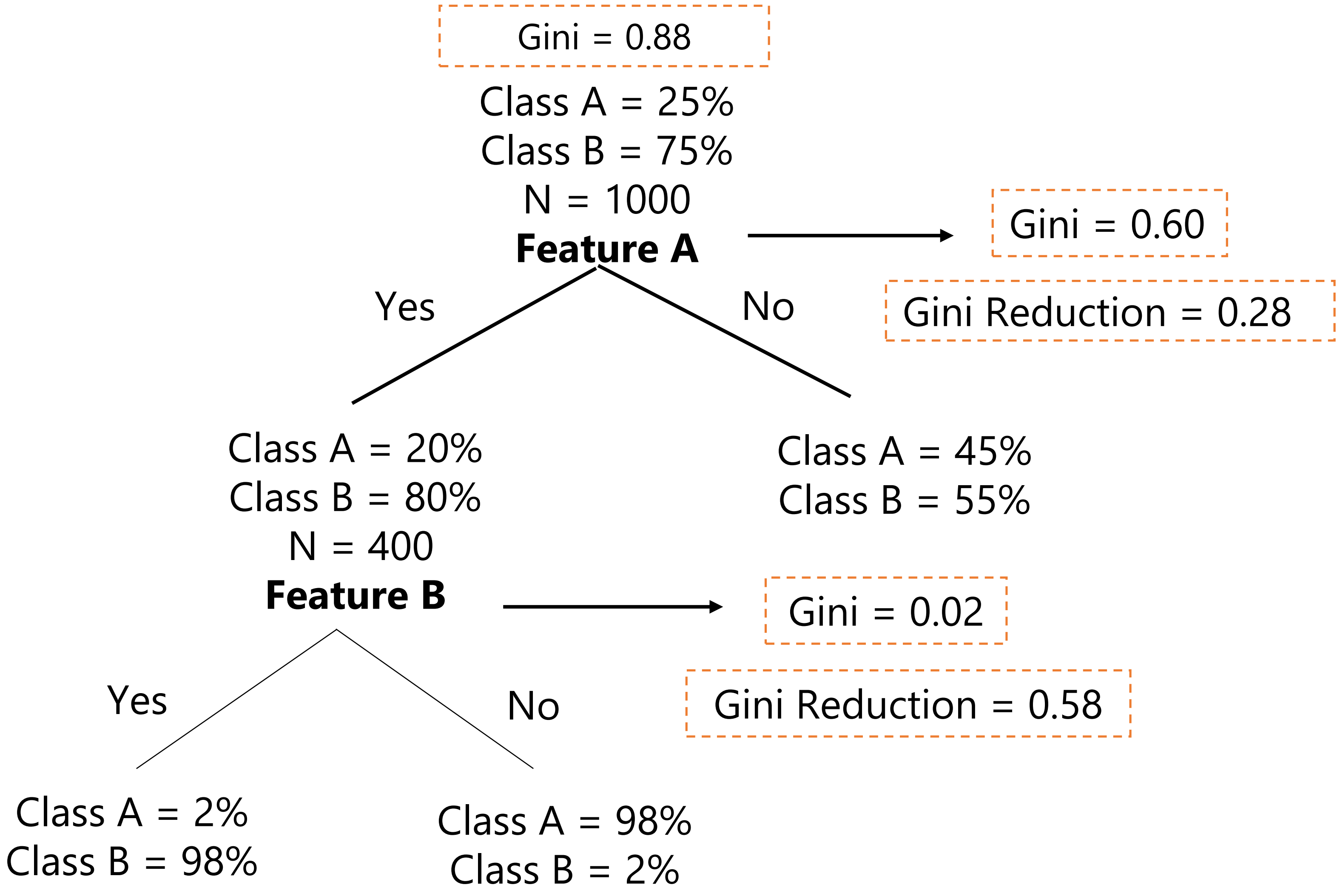


Feature Importance

In Variable Importance both the sequence of the split and the purity of a node should be considered

Feature A precedes Feature B

Feature B creates greater node purity



Feature Importance

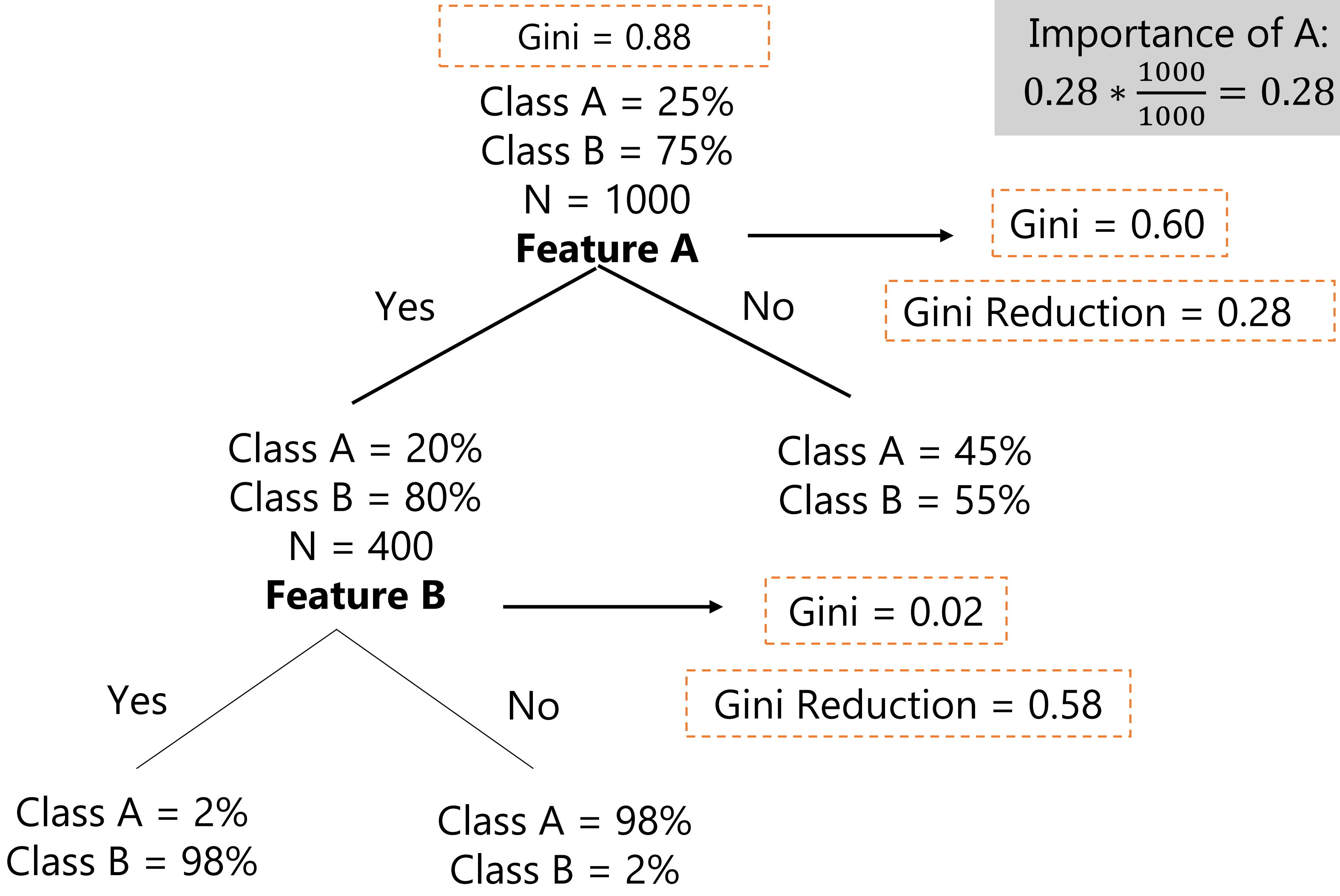
Importance of A: Decrease in Gini * Proportion of data

Decrease in Gini
Ability of a variable to create class imbalance compared to preceding split

Proportion of data
Sequence in which variable causes the split

More observations will pass through the node caused by an early split

Importance of A:
$0.28 * \frac{1000}{1000} = 0.28$



Feature Importance

Importance of B: Decrease in Gini*Proportion of data

Importance of B:

$$0.58 * \frac{400}{1000} = 0.23$$

Gini = 0.88

Class A = 25%
Class B = 75%
N = 1000
Feature A

Gini = 0.60

Gini Reduction = 0.28

Yes

Class A = 20%
Class B = 80%
N = 400
Feature B

Low

Gini = 0.02

Gini Reduction = 0.58

Yes

Class A = 2%
Class B = 98%

No

Class A = 98%
Class B = 2%

No

Class A = 45%
Class B = 55%

Feature Importance

Weigh the decrease in Mean Squared Error and Residual Sum Square appropriately

↓

Feature Importance

Recap

1. Decision tree – Regression
2. Purity Metric
3. Hyperparameters
4. Feature Importance