Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

3.5 Ensembles of Trees

Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

3.5 Ensembles of Trees

Chapter 3: Decision Trees and Ensembles of Trees

- **3.1 Introduction**
- 3.2 Tree-Structure Models
- 3.3 Recursive Partitioning
- 3.4 Pruning
- 3.5 Ensembles of Trees

Essential Discovery Tasks

Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

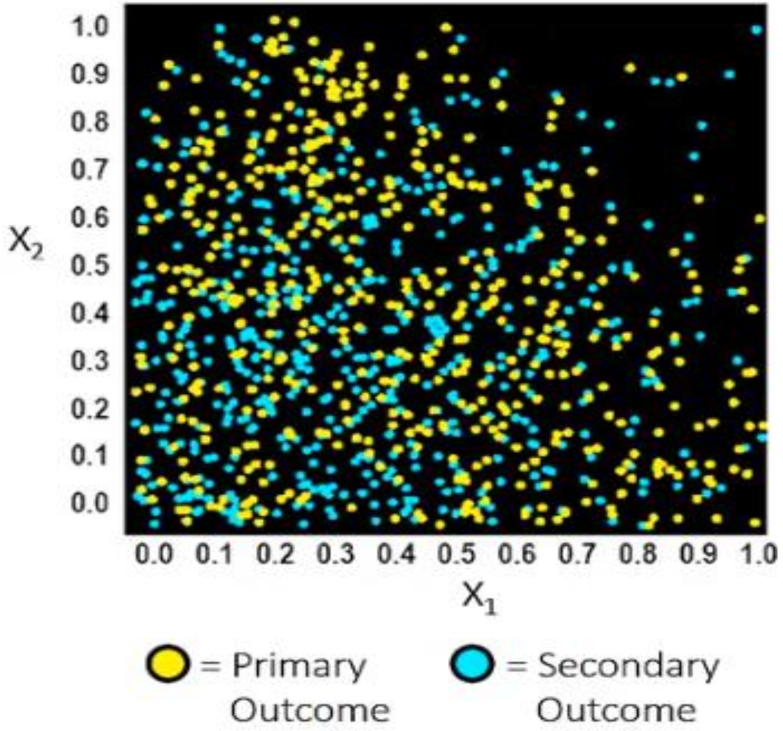
Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

Copyright © SAS Institute Inc. All rights reserved.

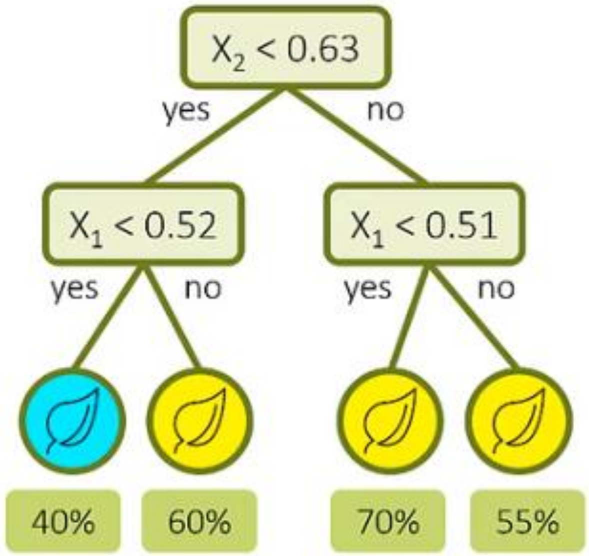
Basics of Decision Trees

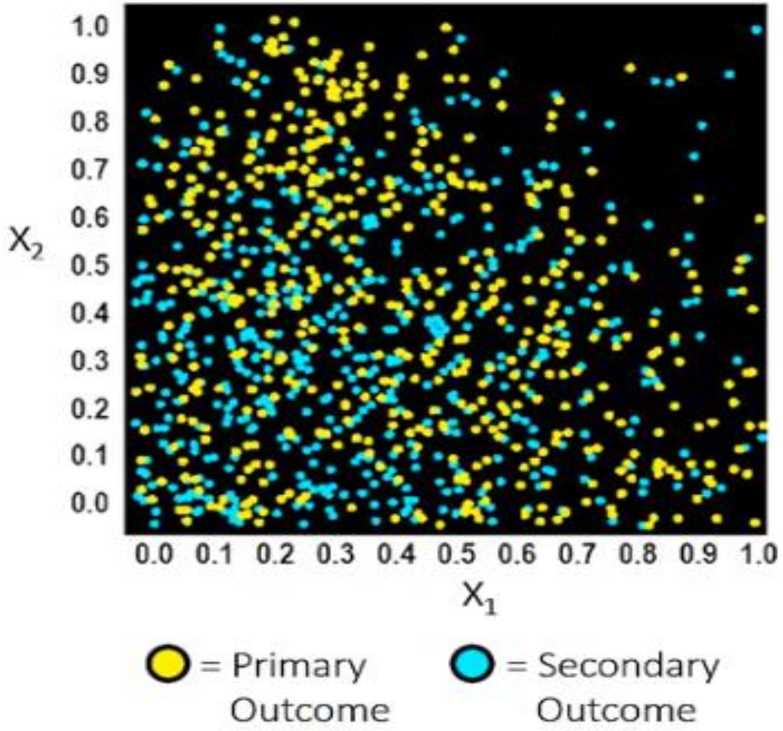


Copyright © SAS Institute Inc. All rights reserved.

Basics of Decision Trees

Analysis goal : Predict the outcome based on the based on the location on the plot





Copyright © SAS Institute Inc. All rights reserved.

Which component of a decision tree provides the predictions?

a. child nodes
b. interior nodes
c. leaf nodes
d. nodes in generation 0
e. root nodes

Copyright © SAS Institute Inc. All rights reserved.

Which component of a decision tree provides the predictions?

a. child nodes

b. interior nodes

c. leaf nodes

d. nodes in generation 0

e. root nodes

Copyright © SAS Institute Inc. All rights reserved.

Building a Decision Tree Model with Default Settings

In this demonstration, you use the CPML demo pipeline as a starting place in a new pipeline in the Demo project. You add a Decision Tree node and build a Decision Tree model using the default settings of the node.

Copyright © SAS Institute Inc. All rights reserved.

Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

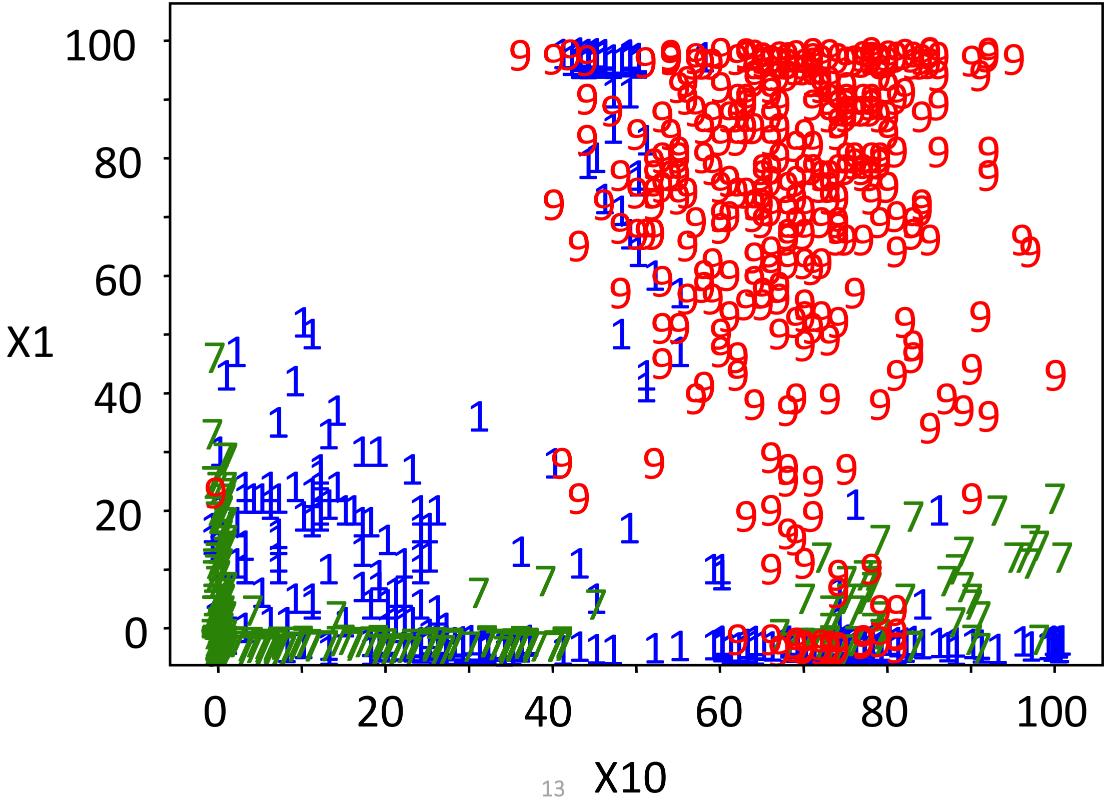
3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

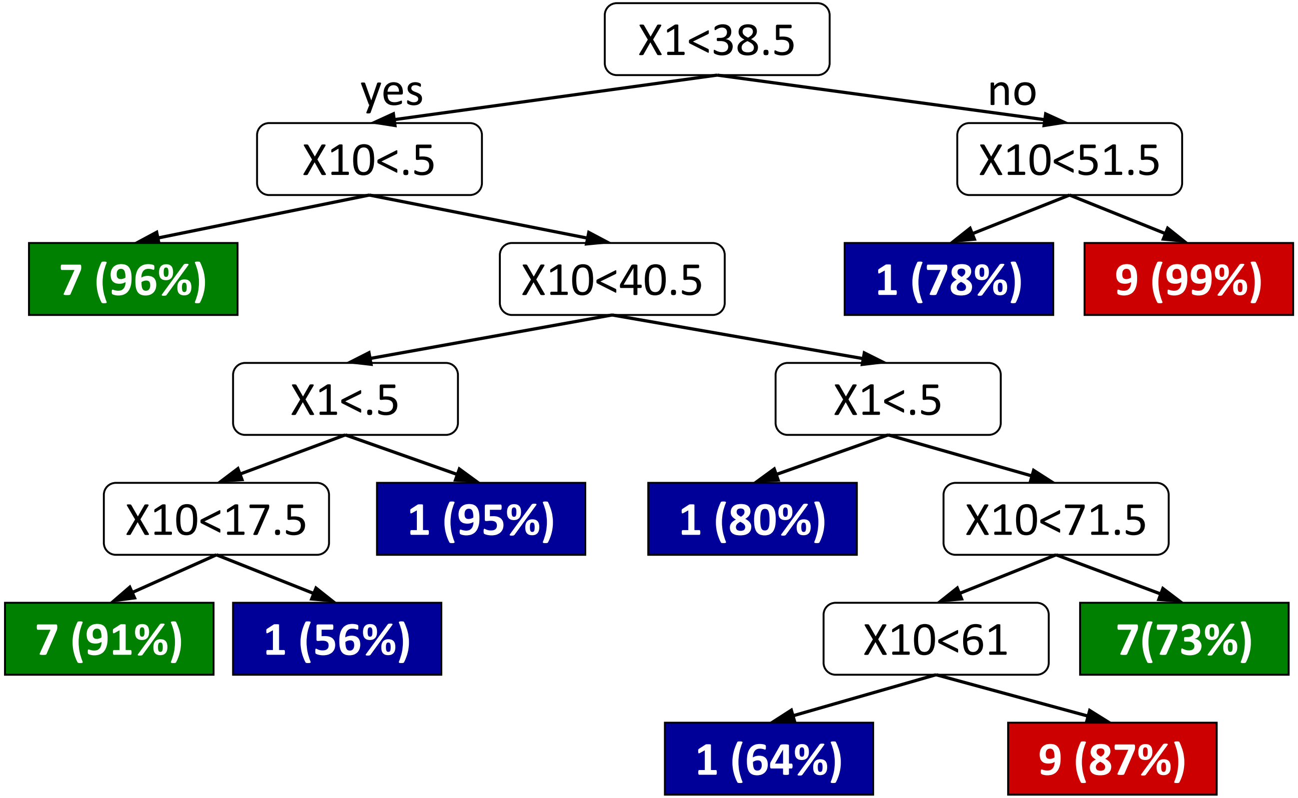
3.5 Ensembles of Trees

Supervised Prediction for a Nominal Target: Handwriting Recognition



Copyright © SAS Institute Inc. All rights reserved.

Classification Tree



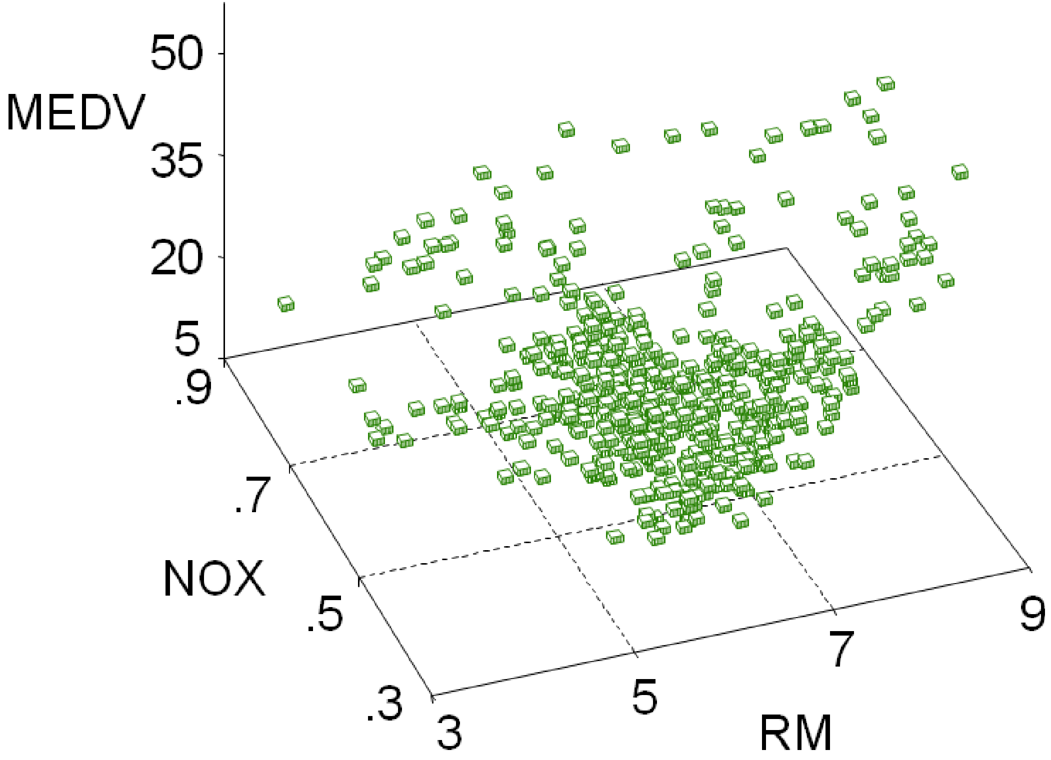
Copyright © SAS Institute Inc. All rights reserved.

Leaves of a Classification Tree

Leaf	Pr(1|x)	Pr(7|x)	Pr(9|x)	Decision
1	.03	.96	.01	7
2	.09	.91	.00	7
3	.56	.44	.00	1
4	.95	.05	.00	1
5	.80	.10	.10	1
6	.64	.09	.27	1
7	.00	.13	.87	9
8	.10	.73	.17	7
9	.78	.01	.21	1
10	.01	.00	.99	9

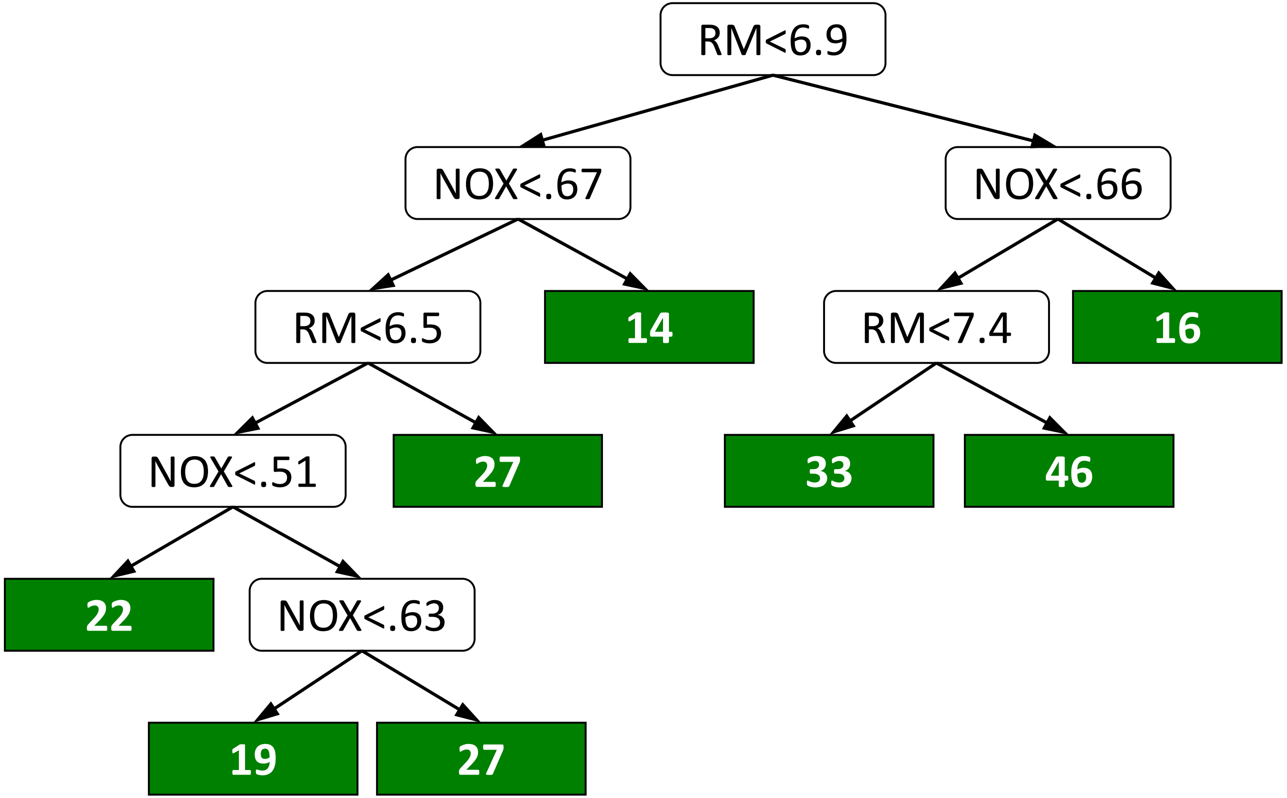
Copyright © SAS Institute Inc. All rights reserved.

Supervised Prediction for an Interval Target: Median Home Value



Copyright © SAS Institute Inc. All rights reserved.

Regression Tree



Copyright © SAS Institute Inc. All rights reserved.

Leaves = Boolean Rules

If RM ∈ {*values*} and NOX ∈ {*values*}, then MEDV=*value*.

Leaf	RM	NOX	Predicted MEDV
1	<6.5	<.51	22
2	<6.5	[.51, .63)	19
3	<6.5	[.63, .67)	27
4	[6.5, 6.9)	<.67	27
5	<6.9	≥.67	14
6	[6.9, 7.4)	<.66	33
7	≥7.4	<.66	46
8	≥6.9	≥.66	16

Copyright © SAS Institute Inc. All rights reserved.

3.01 Multiple Choice Poll

Which of the following statements is true regarding decision trees?

a. To predict cases, decision trees use rules that involve the values or categories of the input variables.
b. Decision trees can handle only categorical targets.
c. The predictor variables can appear only in a single split in the tree.
d. The splits in decision trees can be only binary.

Copyright © SAS Institute Inc. All rights reserved.

3.01 Multiple Choice Poll – Correct Answer

Which of the following statements is true regarding decision trees?

a. To predict cases, decision trees use rules that involve the values or categories of the input variables.
b. Decision trees can handle only categorical targets.
c. The predictor variables can appear only in a single split in the tree.
d. The splits in decision trees can be only binary.

Copyright © SAS Institute Inc. All rights reserved.

Improving a Decision Tree Model by Changing the Tree Structure Parameters

In this demonstration, you change the default settings of the Decision Tree node that was just added in the Starter Template pipeline. You modify the tree structure parameters and compare this model performance to the models built earlier in the course.

Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true regarding decision trees?

a. To predict cases, decision trees use rules that involve the values or categories of the input variables.

b. Decision trees can handle only categorical targets.

c. Each predictor variable can appear only in a single split in the tree.

d. A decision tree for a categorical target can provide decisions but not estimates.

Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true regarding decision trees?

a. To predict cases, decision trees use rules that involve the values or categories of the input variables.

b. Decision trees can handle only categorical targets.

c. Each predictor variable can appear only in a single split in the tree.

d. A decision tree for a categorical target can provide decisions but not estimates.

Copyright © SAS Institute Inc. All rights reserved.

Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

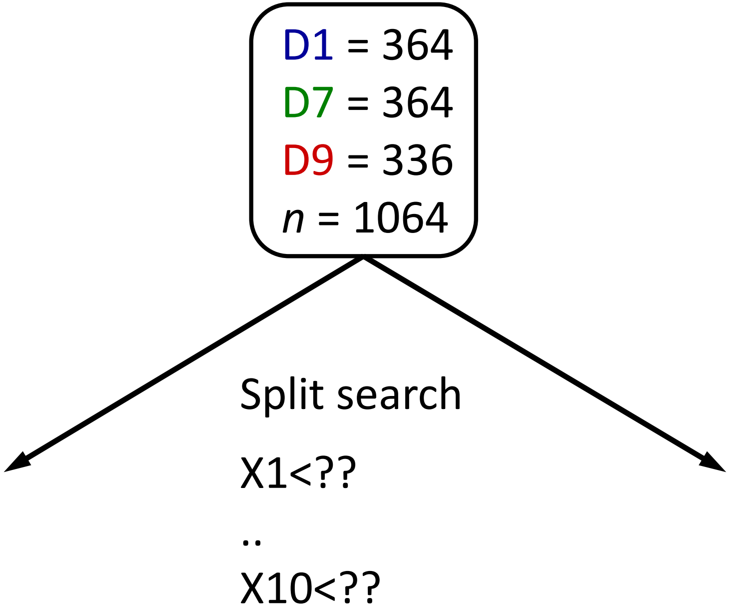
3.5 Ensembles of Trees

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

Copyright © SAS Institute Inc. All rights reserved.

Root-Node Split



1. **Identifies the candidates split.**

Copyright © SAS Institute Inc. All rights reserved.

Root-Node Split

D1 = 364
D7 = 364
D9 = 336
n = 1064

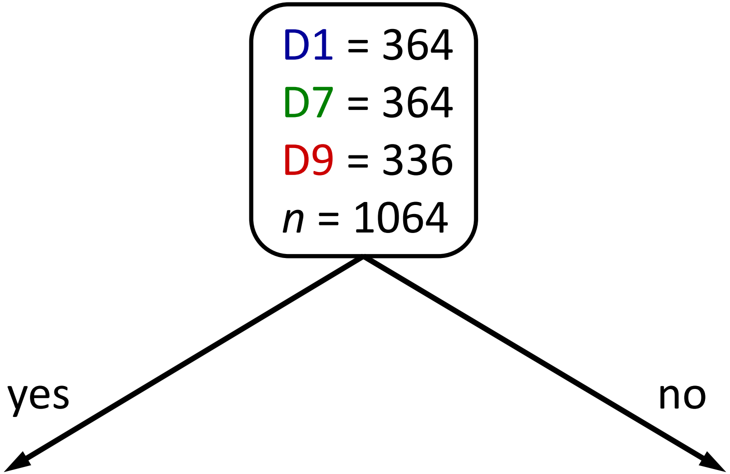
Split search

X1<??
..
X10<??

1. **Identifies the candidates split.**

Copyright © SAS Institute Inc. All rights reserved.

Root-Node Split

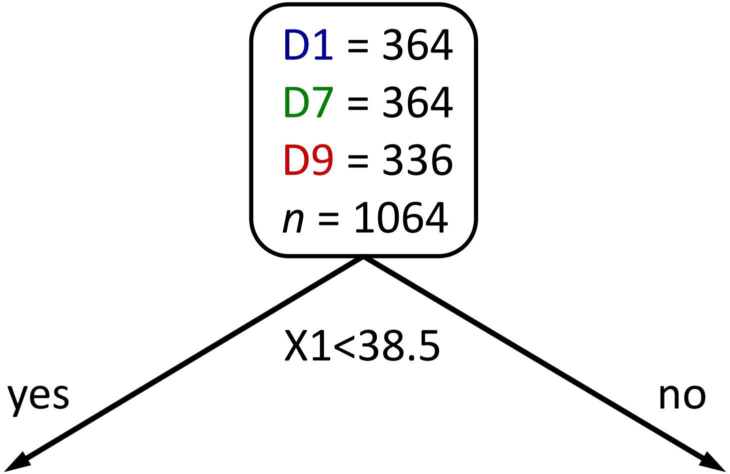


1. **Identifies the candidates split.**

Splitting criterion measure the variable and the target is to reduce the variability

Copyright © SAS Institute Inc. All rights reserved.

Root-Node Split

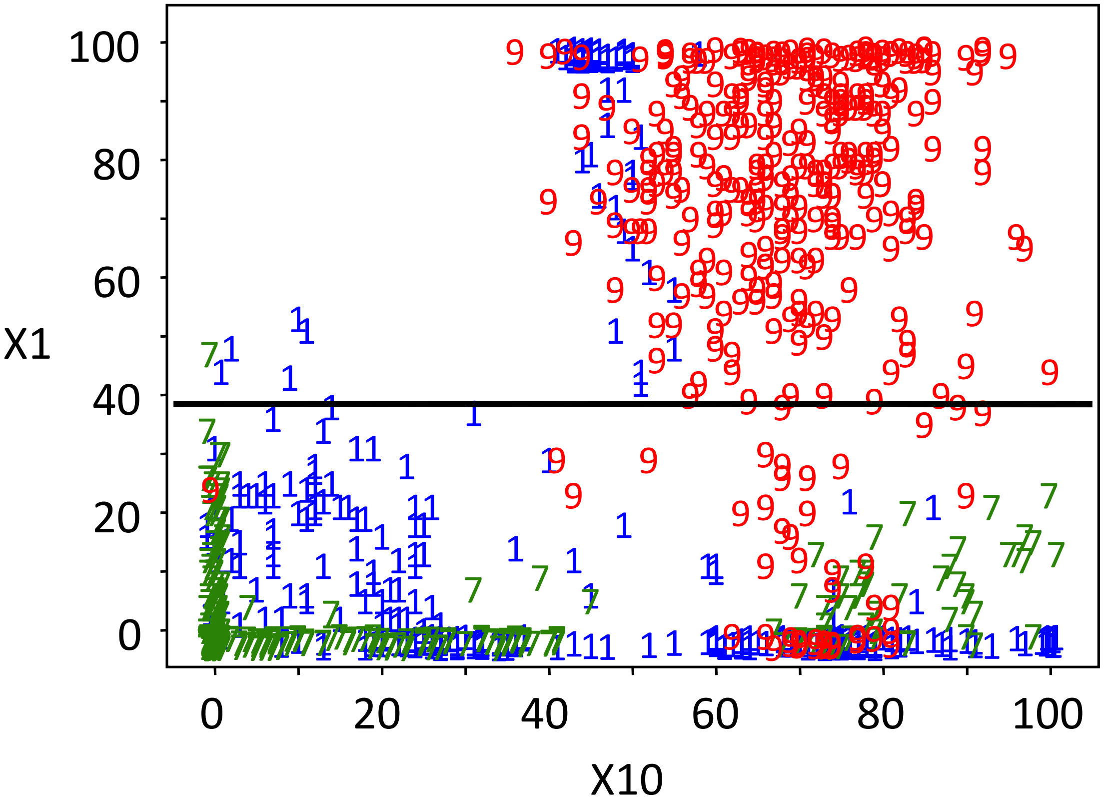


1. **Identifies the candidates split.**
2. **Select a split**

Splitting criterion measure the variable and the target is to reduce the variability

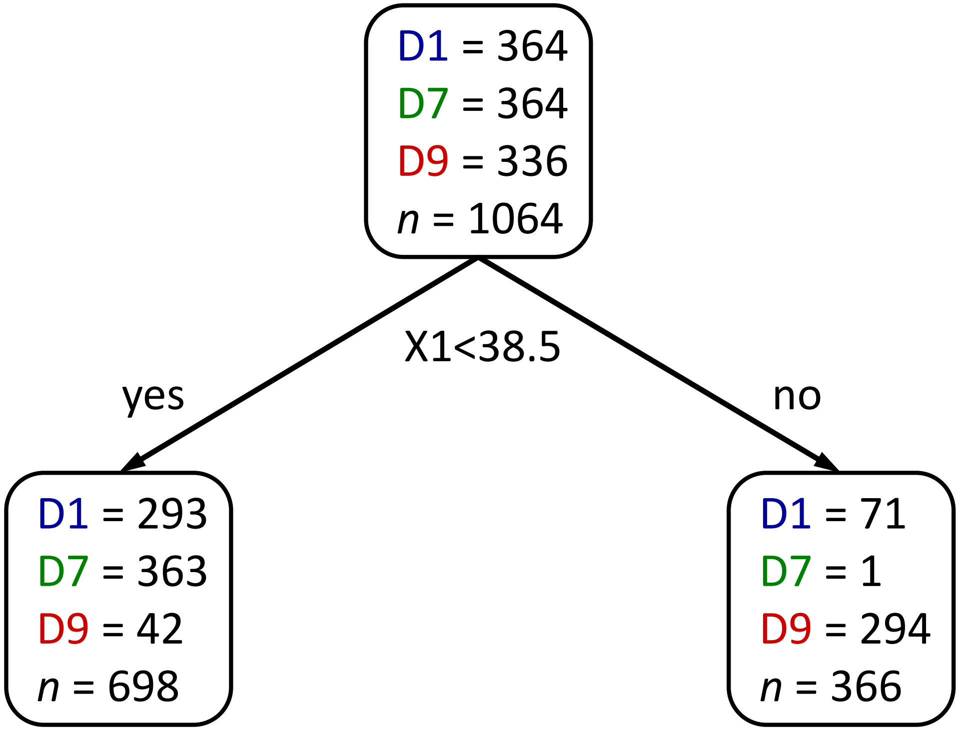
Copyright © SAS Institute Inc. All rights reserved.

1-Deep Space



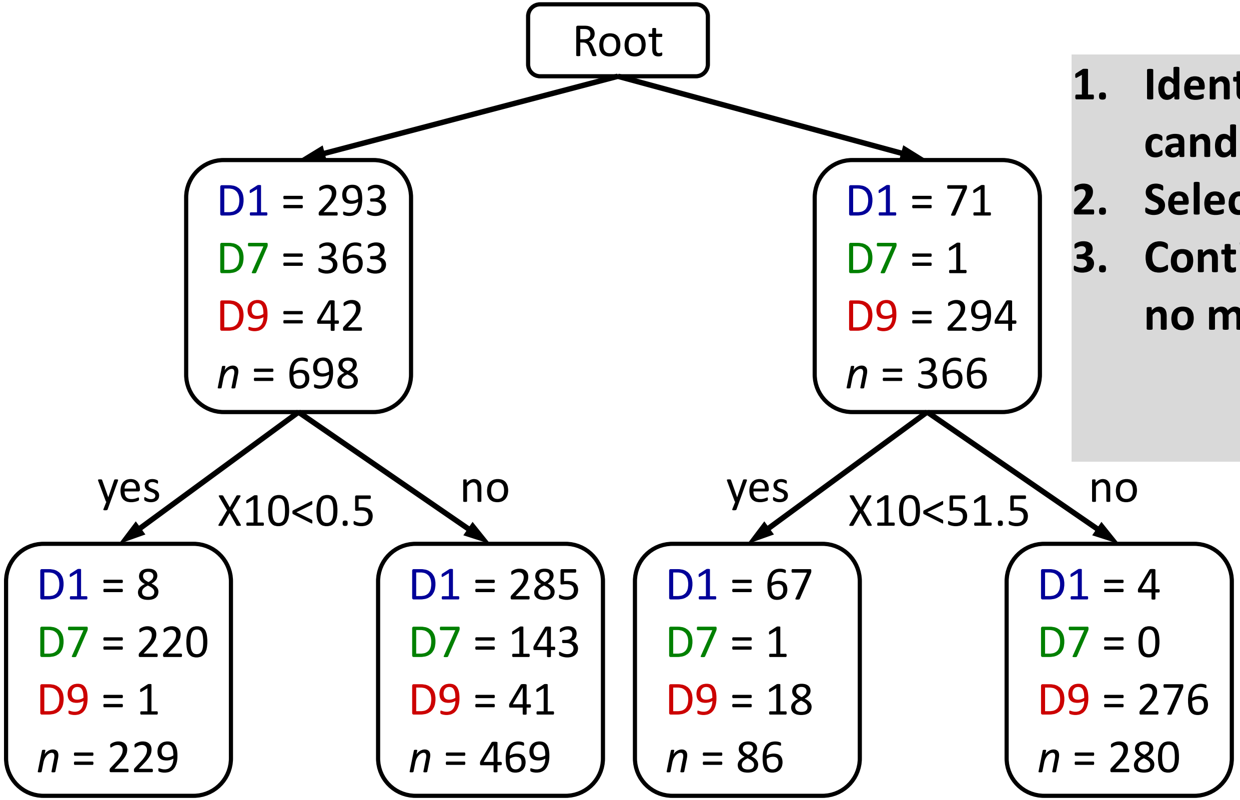
Copyright © SAS Institute Inc. All rights reserved.

Root-Node Split



Copyright © SAS Institute Inc. All rights reserved.

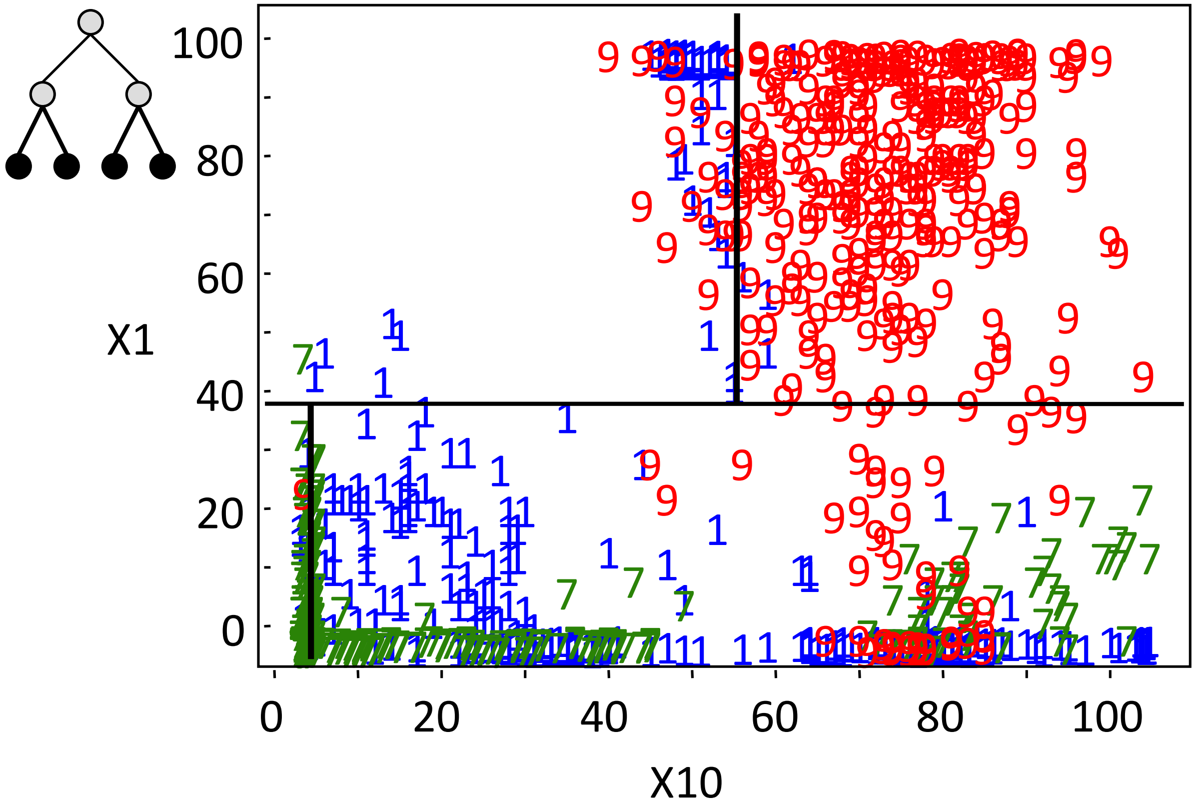
Depth 2



1. **Identifies the candidates split.**
2. **Select a split.**
3. **Continuous… until no more growth.**

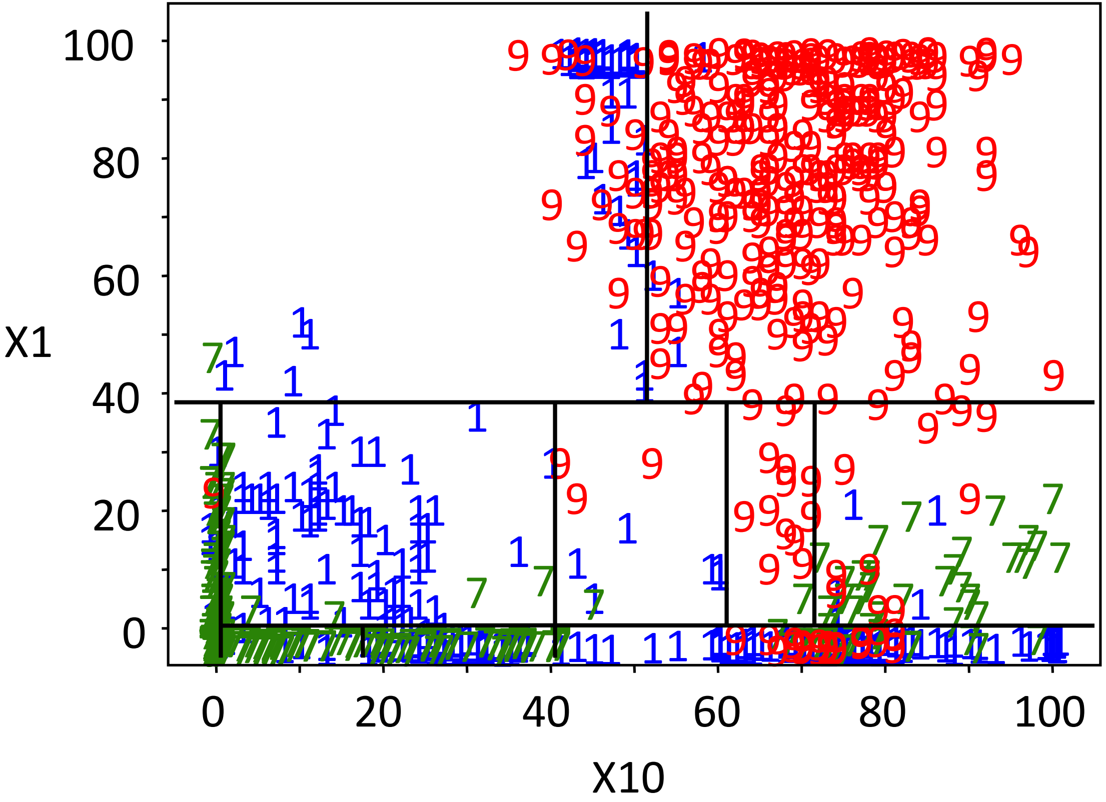
Copyright © SAS Institute Inc. All rights reserved.

2-Deep Space



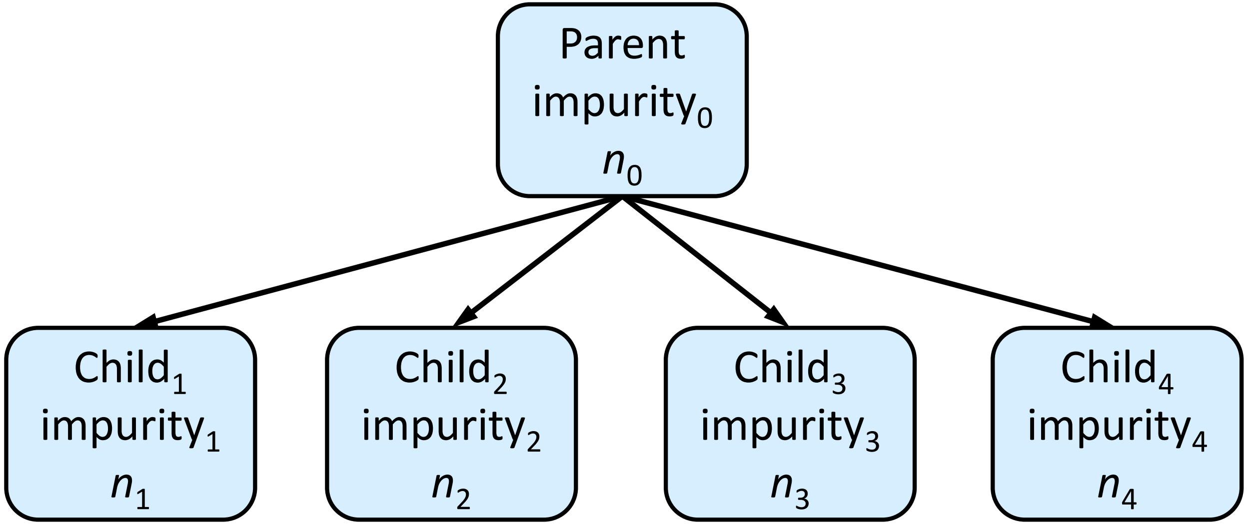
Copyright © SAS Institute Inc. All rights reserved.

Split Characteristics



Copyright © SAS Institute Inc. All rights reserved.

Impurity Reduction Measures



Delta impurity

$$\Delta i = i(0) - \left(\frac{n_1}{n_0} i(1) + \frac{n_2}{n_0} i(2) + \frac{n_3}{n_0} i(3) + \frac{n_4}{n_0} i(4) \right)$$

Copyright © SAS Institute Inc. All rights reserved.

The Gini Index and Impurity

$$1-\sum_{j=1}^{r} p_j^2 = 2\sum_{j<k} p_j p_k$$

0 pure node

1 impure node

high diversity, low purity

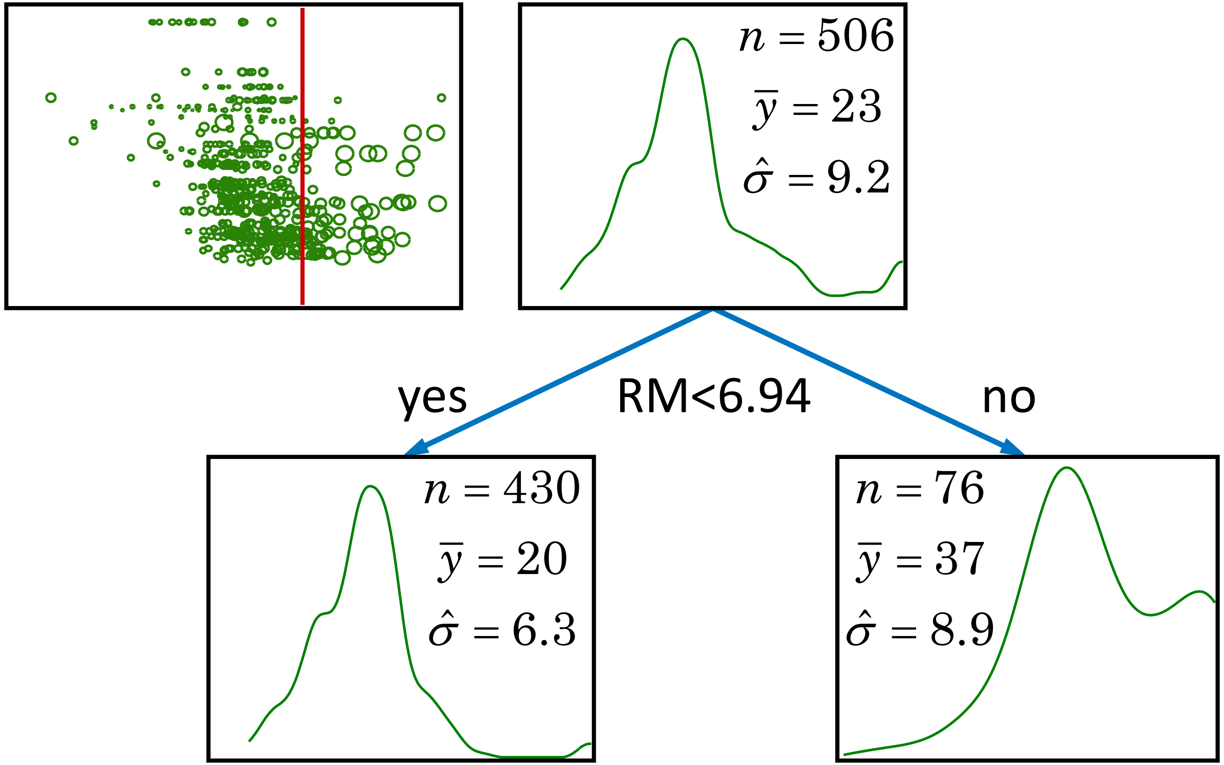
Pr(interspecific encounter) = $1-2(3/8)^2-2(1/8)^2 = .69$

low diversity, high purity

Pr(interspecific encounter) = $1-(6/7)^2-(1/7)^2 = .24$

Copyright © SAS Institute Inc. All rights reserved.

Variance Reduction



Copyright © SAS Institute Inc. All rights reserved.

Suppose you build a decision tree to predict whether a customer makes a purchase on the internet (yes or no). A leaf has perfect purity when it contains which of the following?

a. all events

b. all nonevents

c. either all events or all nonevents

d. an equal number of events and nonevents

Copyright © SAS Institute Inc. All rights reserved.

Suppose you build a decision tree to predict whether a customer makes a purchase on the internet (yes or no). A leaf has perfect purity when it contains which of the following?

a. all events

b. all nonevents

c. either all events or all nonevents

d. an equal number of events and nonevents

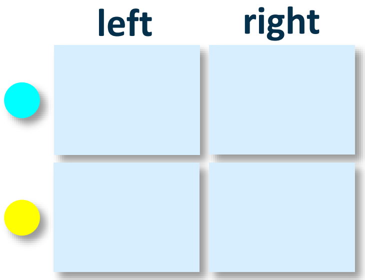
Copyright © SAS Institute Inc. All rights reserved.

Split Criteria in Model Studio

- Categorical target
  - CHAID – Chi-square Automatic Interaction Detection
  - Chi-Square
  - Entropy
  - Gini
  - Information gain ratio
- Interval target
  - CHAID
  - *F* test
  - Variance

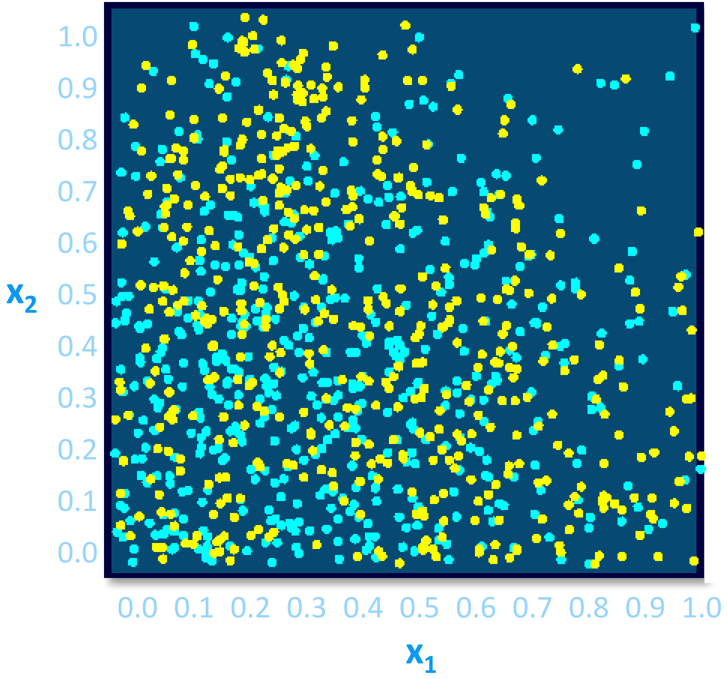
Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search



Classification Matrix

Calculate the logworth of every partition on input x_1.

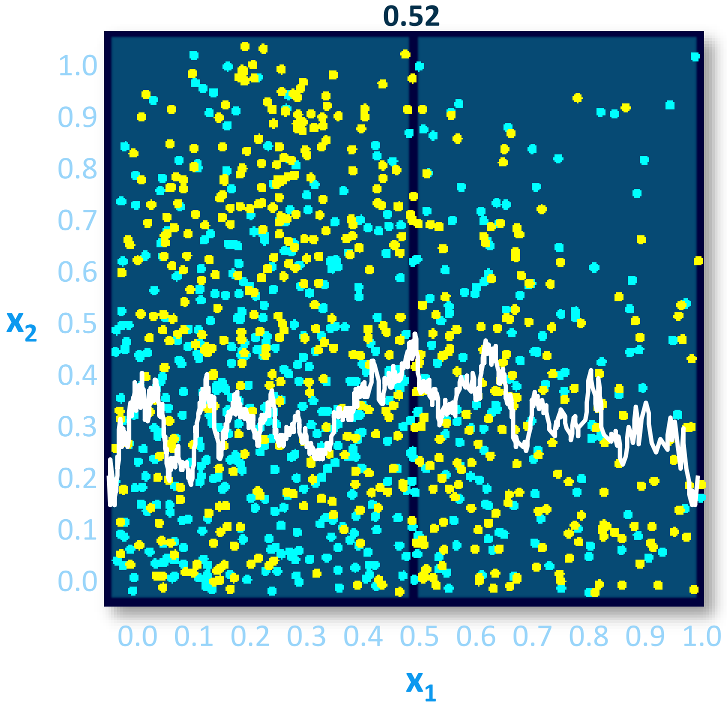


Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search

	left	right
	53%	42%
	47%	58%

max logworth(x_1)
0.95



Select the partition with the maximum logworth.

Copyright © SAS Institute Inc. All rights reserved.

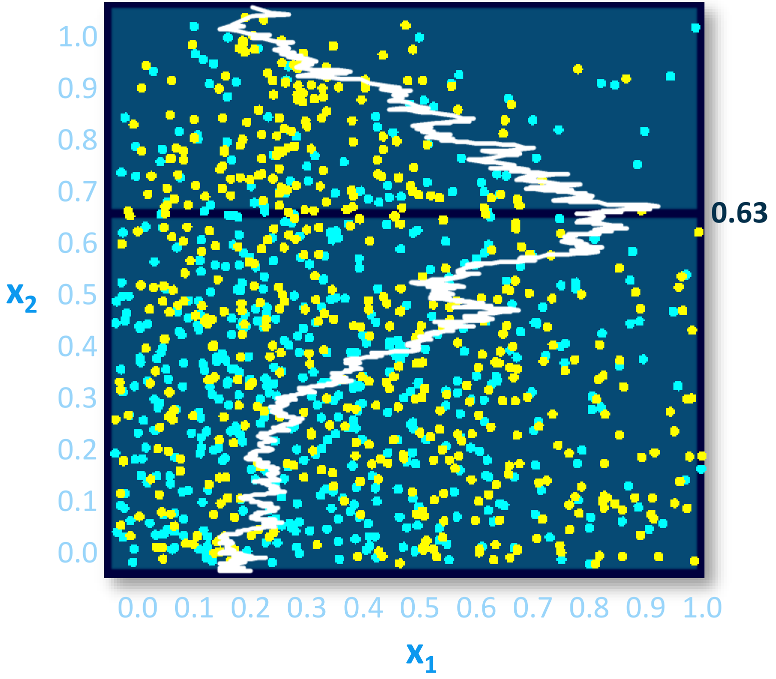
Decision Tree Split Search

	left	right
	53%	42%
	47%	58%

max
logworth(x_1)
0.95

	bottom	top
cyan	54%	35%
yellow	46%	65%

max
logworth(x_2)
4.92



Copyright © SAS Institute Inc. All rights reserved.

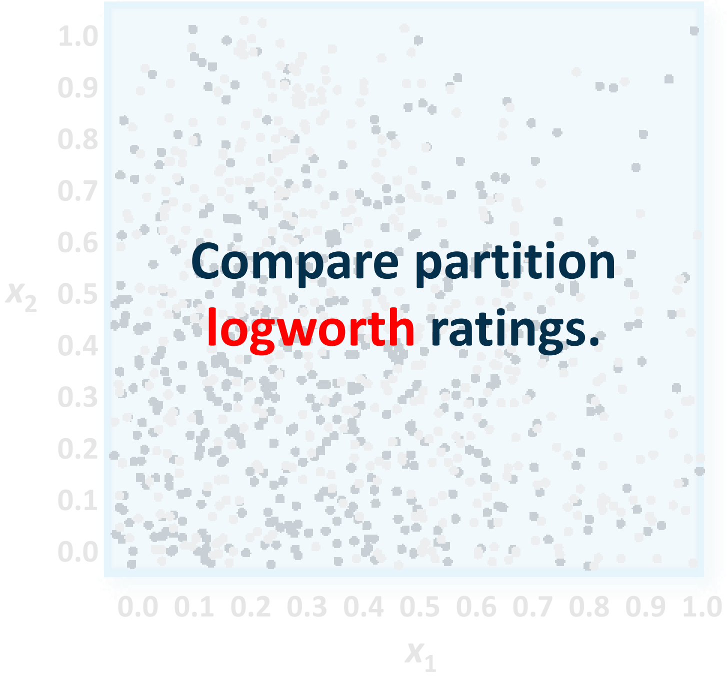
Decision Tree Split Search

	left	right
	53%	42%
	47%	58%

max logworth(x_1) **0.95**

	bottom	top
	54%	35%
	46%	65%

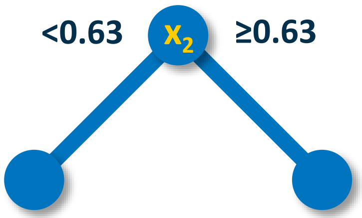
max logworth(x_2) **4.92**

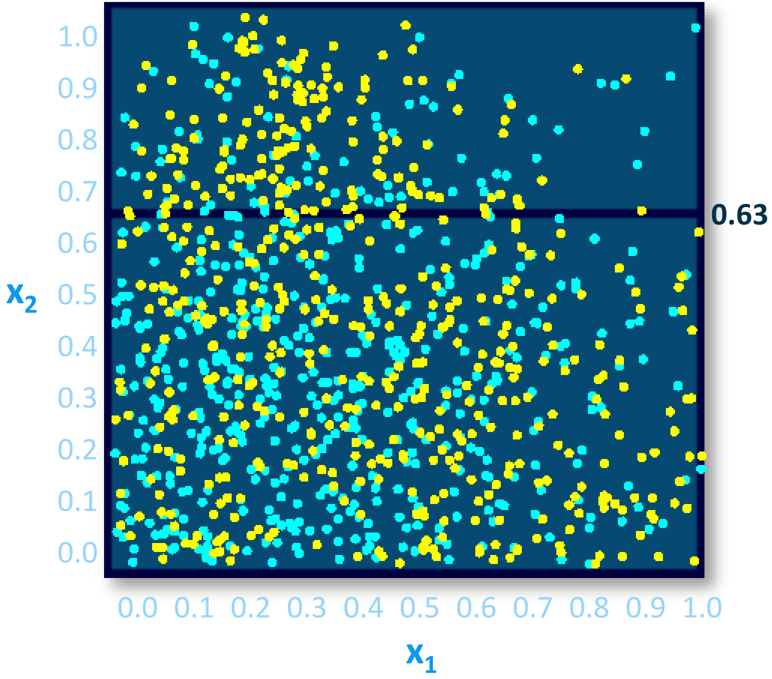


Compare partition logworth ratings.

Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search

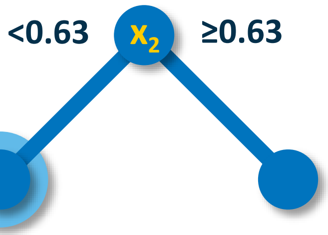




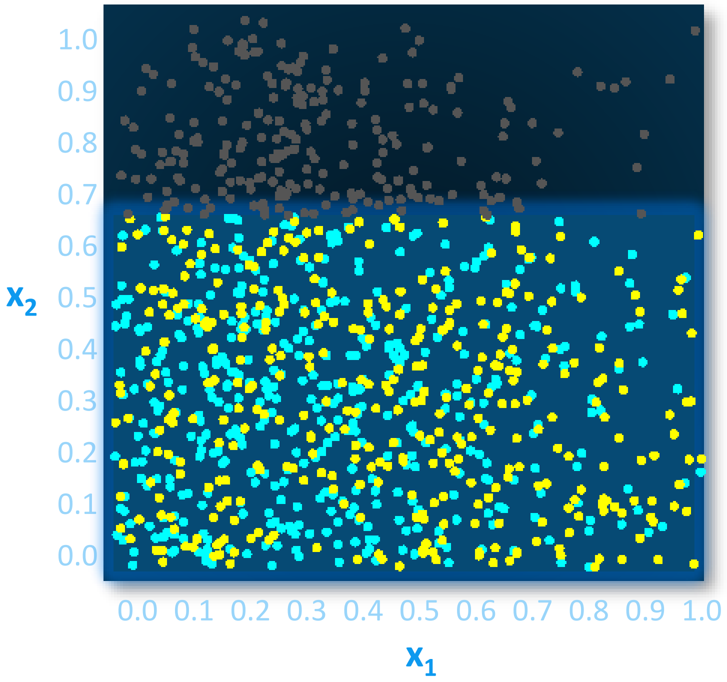
Create a partition rule from the best partition across all inputs.

Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search



Repeat the process in each subset.

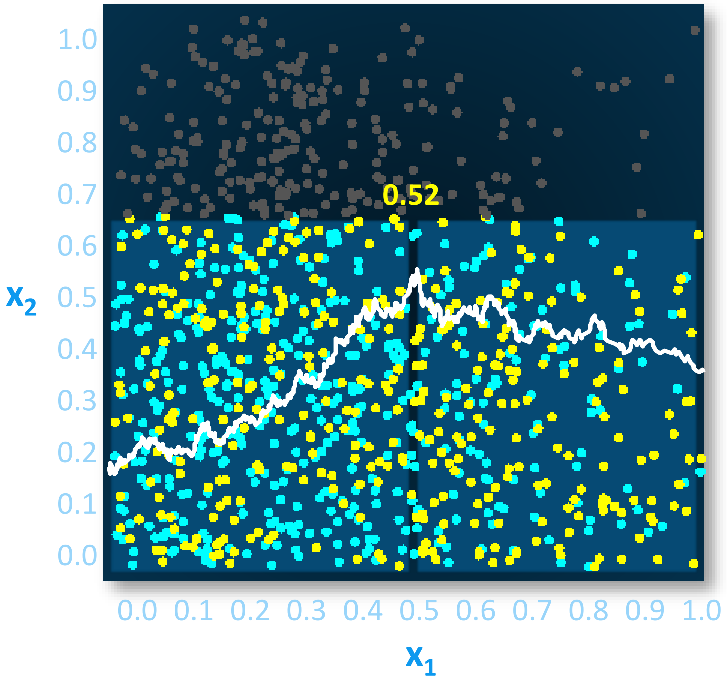


Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search

	left	right
cyan	61%	55%
yellow	39%	45%

max logworth(x_1)
5.72



Copyright © SAS Institute Inc. All rights reserved.

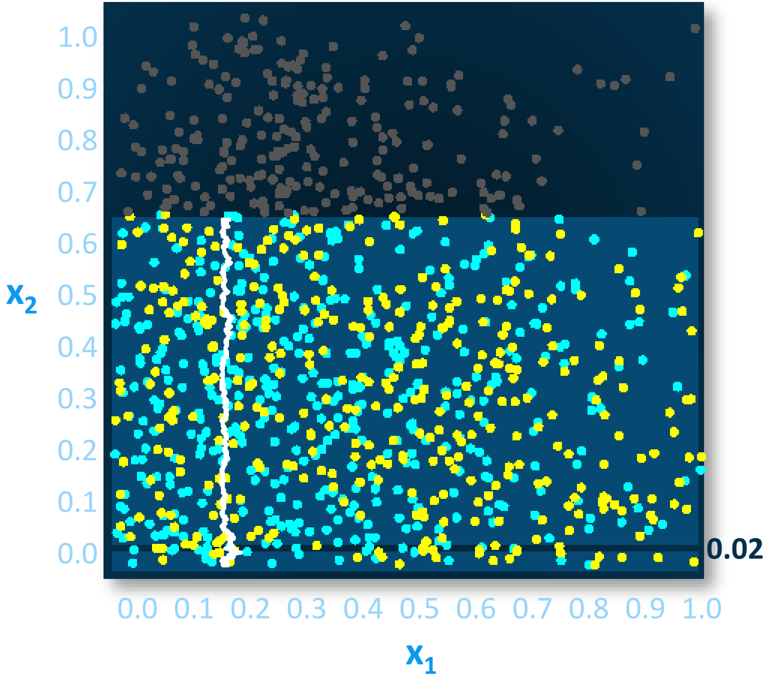
Decision Tree Split Search

	left	right
	61%	55%
	39%	45%

max logworth(x_1) 5.72

	bottom	top
	38%	55%
	62%	45%

max logworth(x_2) -2.01



Copyright © SAS Institute Inc. All rights reserved.

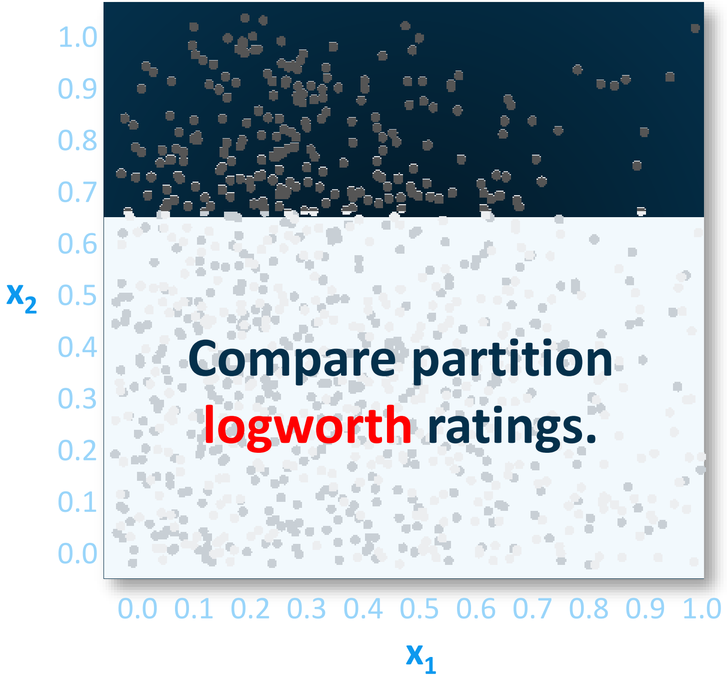
Decision Tree Split Search

	left	right
	61%	55%
	39%	45%

max logworth(x_1)
5.72

	bottom	top
	38%	55%
	62%	45%

max logworth(x_2)
-2.01



Copyright © SAS Institute Inc. All rights reserved.

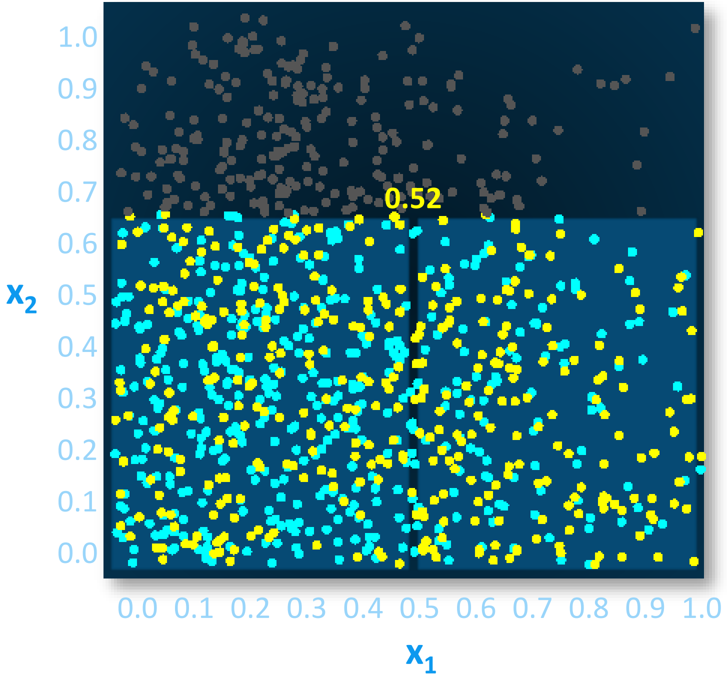
Decision Tree Split Search

	left	right
	61%	55%
	39%	45%

max logworth(x_1)
5.72

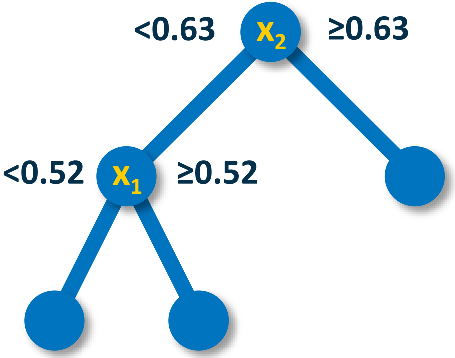
	bottom	top
	38%	55%
	62%	45%

max logworth(x_2)
-2.01

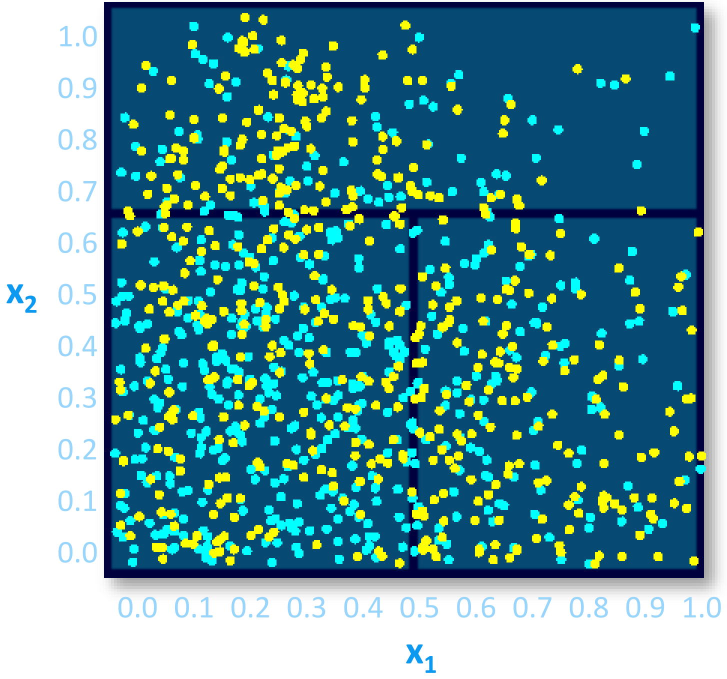


Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search

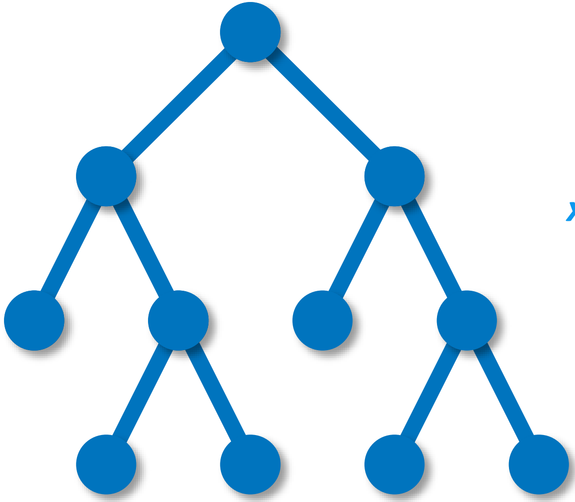


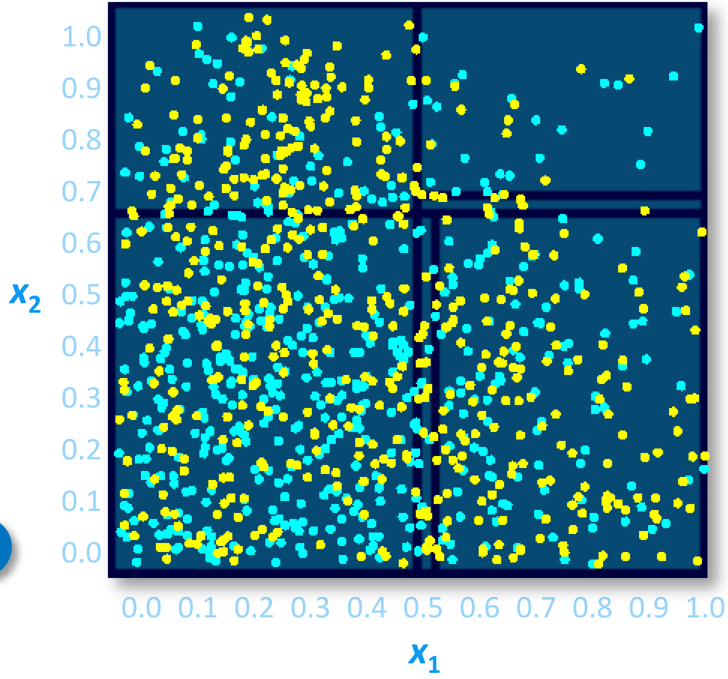
Create a second partition rule.



Copyright © SAS Institute Inc. All rights reserved.

Decision Tree Split Search





Repeat to form a maximal tree.

Copyright © SAS Institute Inc. All rights reserved.

If you build a decision tree model to predict a binary target, the rules in that decision tree are limited to two-way splits.

a. True

b. False

Copyright © SAS Institute Inc. All rights reserved.

If you build a decision tree model to predict a binary target, the rules in that decision tree are limited to two-way splits.

a. True

b. False

Copyright © SAS Institute Inc. All rights reserved.

How do decision trees address the curse of dimensionality?

?????

Copyright © SAS Institute Inc. All rights reserved.

How do decision trees address the curse of dimensionality?

The split search process <u>reduces the number of inputs</u> in the model by eliminating irrelevant inputs.

Irrelevant inputs do not appear in any splitting rules in the decision tree.

Copyright © SAS Institute Inc. All rights reserved.

Handling of Missing Values in Decision Trees

- **Use in search:** missing values are used as a value.
  - **Nominal inputs**: treat missing values as a separate level.
  - **Ordinal inputs**: require modification of the split search strategy for missing values by adding a separate branch adjacent to the ordinal levels.
  - **Interval inputs**: treat missing values as having the same unknown nonmissing value.
- **Additional options**:
  - **Largest branch**: assign observations to the largest branch.
  - **Most correlated branch**: assign observations to the branch with the smallest residual sum of squares among observations that contain missing values.
  - **Separate branch**: assign observations to a separate branch.

Copyright © SAS Institute Inc. All rights reserved.

3.02 Multiple Choice Poll

Which of the following statements is true regarding decision trees?

a. The recursive partitioning used to construct decision trees leads them to being uninterpretable.
b. The optimal split for the next input considered is the one that minimizes the logworth function for that input.
c. The maximal decision tree is usually the one used to score new data.
d. The logworth of a split can sometimes be negative.

Copyright © SAS Institute Inc. All rights reserved.

3.02 Multiple Choice Poll – Correct Answer

Which of the following statements is true regarding decision trees?

a. The recursive partitioning used to construct decision trees leads them to being uninterpretable.

b. The optimal split for the next input considered is the one that minimizes the logworth function for that input.

c. The maximal decision tree is usually the one used to score new data.

d. The logworth of a split can sometimes be negative.

Copyright © SAS Institute Inc. All rights reserved.

Improving a Decision Tree Model by Changing the Recursive Partitioning Parameters

In this demonstration, you change more settings of the Decision Tree node in the Starter Template pipeline. You modify the recursive partitioning parameters and compare this model performance to the models built earlier in the course.

Copyright © SAS Institute Inc. All rights reserved.

Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

3.5 Ensembles of Trees

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

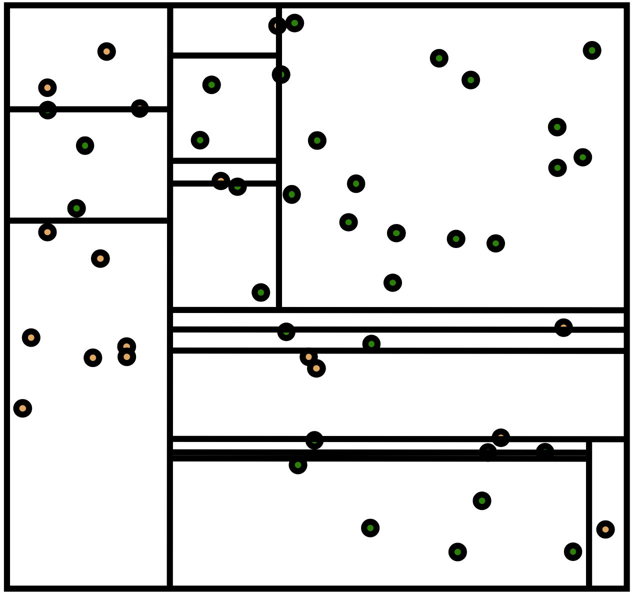
Copyright © SAS Institute Inc. All rights reserved.

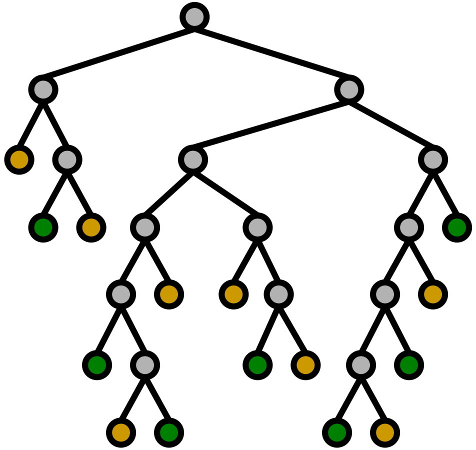
Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

Copyright © SAS Institute Inc. All rights reserved.

Maximal Tree





Copyright © SAS Institute Inc. All rights reserved.

Pruning Options

- **Subtree method**: specifies how to construct the subtree in terms of subtree methods.
  - **C4.5**: The pruning is done with a C4.5 algorithm (class target only).
    - **Confidence** – specifies the binomial distribution confidence level to use to determine the error rates of merged and split nodes.
  - **Cost complexity**: The subtree with a minimum leaf-penalized ASE is chosen.
  - **Reduced error**: The smallest subtree with the best assessment value is chosen.

Copyright © SAS Institute Inc. All rights reserved.

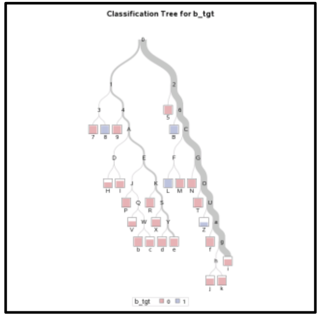
Pruning Options

- **Selection method**: specifies how to construct the subtree in terms of selection methods.
  - **Automatic**: specifies the subtree for the selected subtree pruning method.
  - **Largest**: specifies the full tree.
  - **N**: specifies the largest subtree with at most N leaves.
    - **Number of leaves**: specifies the number of leaves that are used in creating the subtree when the subtree selection method is set to N.
- **Cross validation folds**: specifies the number of cross validation folds to use for cost-complexity pruning when there is no validation data.
  - **1–SE rule**: specifies whether to perform the one standard error rule when performing cross validated cost complexity pruning.

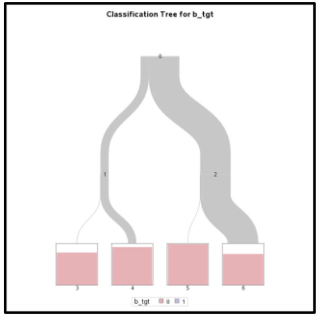
Copyright © SAS Institute Inc. All rights reserved.

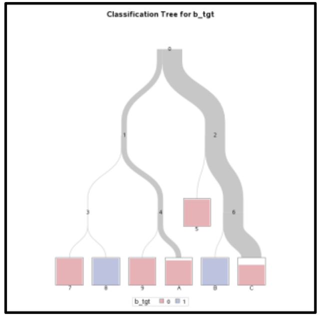
Bottom-Up Pruning

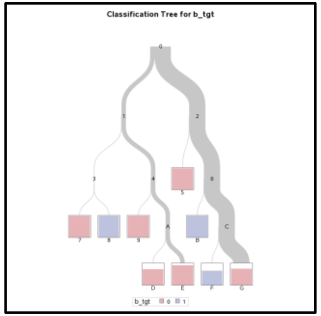
1. Grow a maximal tree:

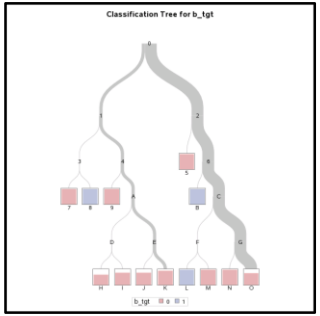


2. Prune to create optimal sequence of subtrees:





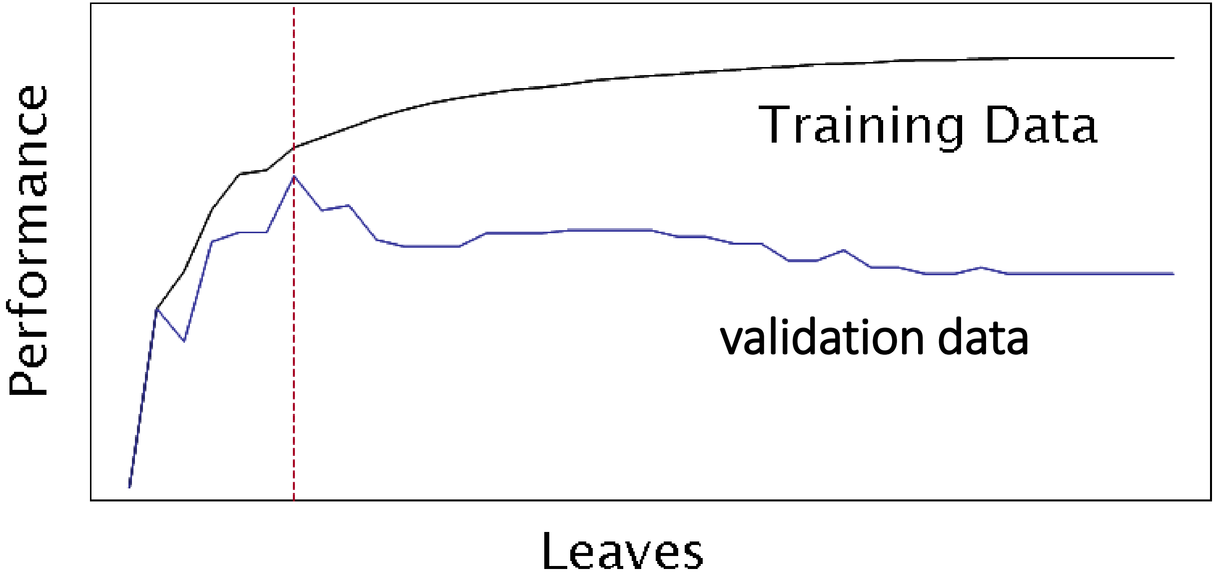




Copyright © SAS Institute Inc. All rights reserved.

Bottom-Up Pruning

3. Choose the best tree on validation data:



Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true about pruning decision trees?

a. The goal of pruning is a tree with low bias and high variance.

b. Pruning starts by identifying a sequence of candidate subtrees, one for each possible number of leaves, and then selects the best of the candidates.

c. In bottom-up pruning, the subtree with the best performance on training data is selected.

d. Top-down pruning is usually more time-consuming but is considered more effective than bottom-up pruning.

Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true about pruning decision trees?

a. The goal of pruning is a tree with low bias and high variance.

b. Pruning starts by identifying a sequence of candidate subtrees, one for each possible number of leaves, and then selects the best of the candidates.

c. In bottom-up pruning, the subtree with the best performance on training data is selected.

d. Top-down pruning is usually more time-consuming but is considered more effective than bottom-up pruning.

Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

Copyright © SAS Institute Inc. All rights reserved.

Autotuning

- Search for the best combination of values in different properties:
  - Maximum depth
  - Interval input bins
  - Split criteria (class and interval targets)
  - Search method
    - Bayesian, Genetic algorithm, Latin hypercube sample, Random
  - Validation method
    - Partition, cross validation
  - Objective function (class and interval targets)

Copyright © SAS Institute Inc. All rights reserved.

3.03 Multiple Choice Poll

Which of the following statements is true regarding decision trees?

a. A well-fit tree has low bias and high variance.

b. Accuracy is obtained by multiplying the proportion of observations falling into each leaf by the proportion of those correctly classified in the leaf and then summing across all leaves.

c. In bottom-up pruning, the subtree with the best performance on training data is selected.

d. Top-down pruning is usually slower but is considered more effective than bottom-up pruning.

Copyright © SAS Institute Inc. All rights reserved.

3.03 Multiple Choice Poll – Correct Answer

Which of the following statements is true regarding decision trees?

a. A well-fit tree has low bias and high variance.
b. Accuracy is obtained by multiplying the proportion of observations falling into each leaf by the proportion of those correctly classified in the leaf and then summing across all leaves.
c. In bottom-up pruning, the subtree with the best performance on training data is selected.
d. Top-down pruning is usually slower but is considered more effective than bottom-up pruning.

Copyright © SAS Institute Inc. All rights reserved.

Improving a Decision Tree Model by Changing the Pruning Parameters

In this demonstration, you change the default settings of the Decision Tree node in the Starter Template pipeline. You modify the pruning parameters and compare this model performance to the model built earlier in the course.

Copyright © SAS Institute Inc. All rights reserved.

Exercise

This exercise reinforces the concepts discussed previously. You use Model Studio to build a Decision Tree model using the Autotune feature.

Copyright © SAS Institute Inc. All rights reserved.

Chapter 3: Decision Trees and Ensembles of Trees

3.1 Introduction

3.2 Tree-Structure Models

3.3 Recursive Partitioning

3.4 Pruning

3.5 Ensembles of Trees

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

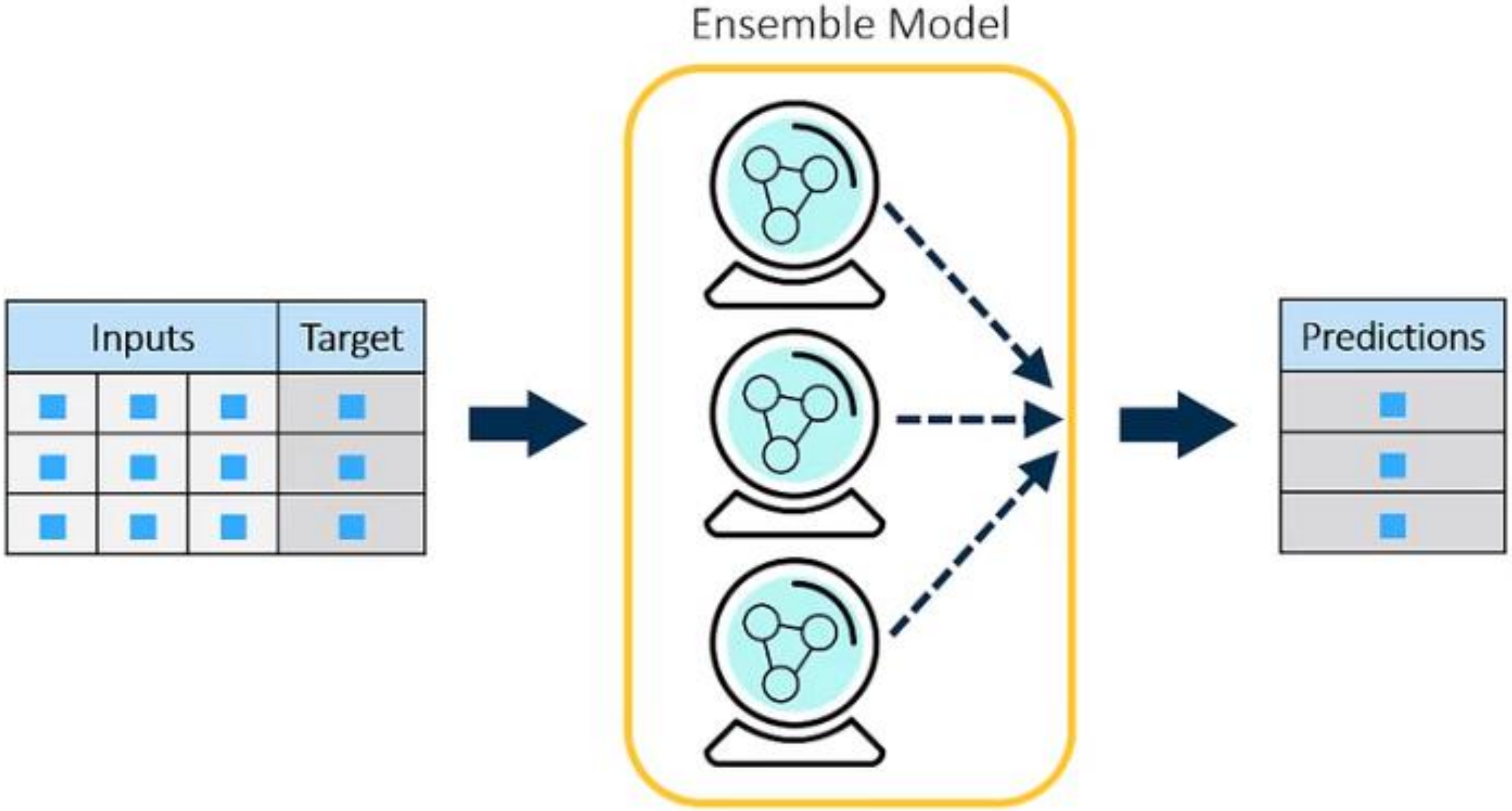
Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm.
- Improve the model.
- Optimize complexity of the model.
- Regularize and tune hyperparameters of the model.
- Build ensemble models.

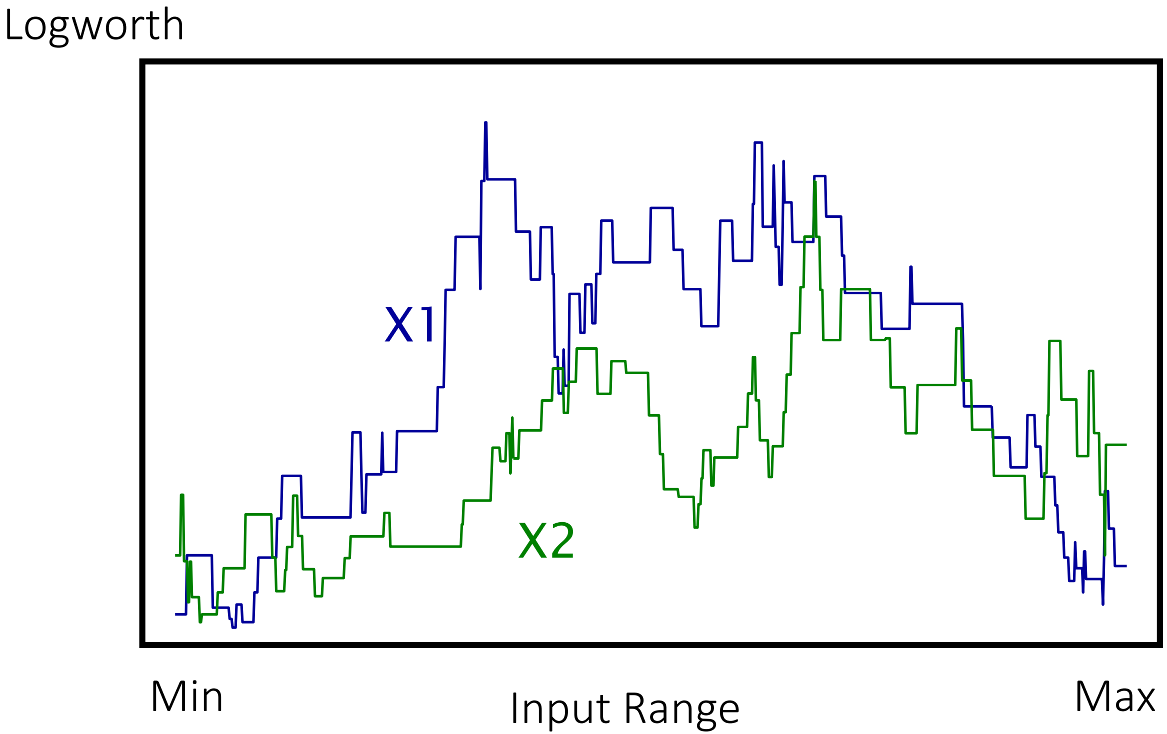
Copyright © SAS Institute Inc. All rights reserved.

Building Ensemble Models



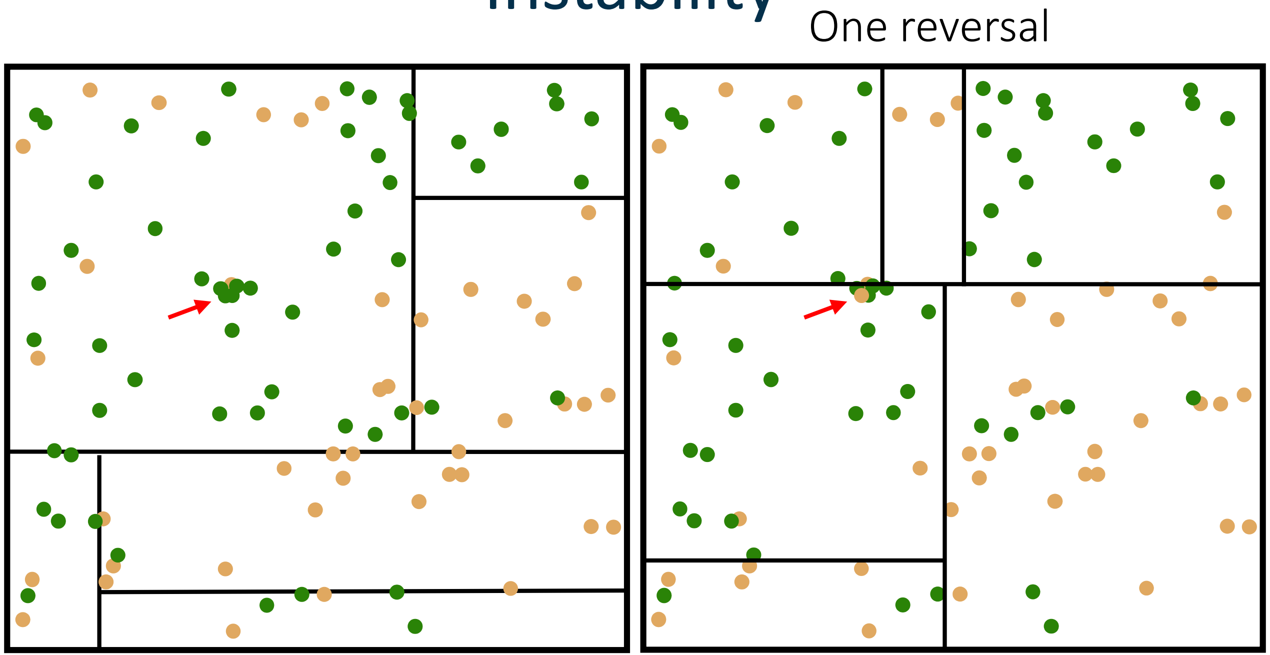
Copyright © SAS Institute Inc. All rights reserved.

Competitor Splits



Copyright © SAS Institute Inc. All rights reserved.

Instability



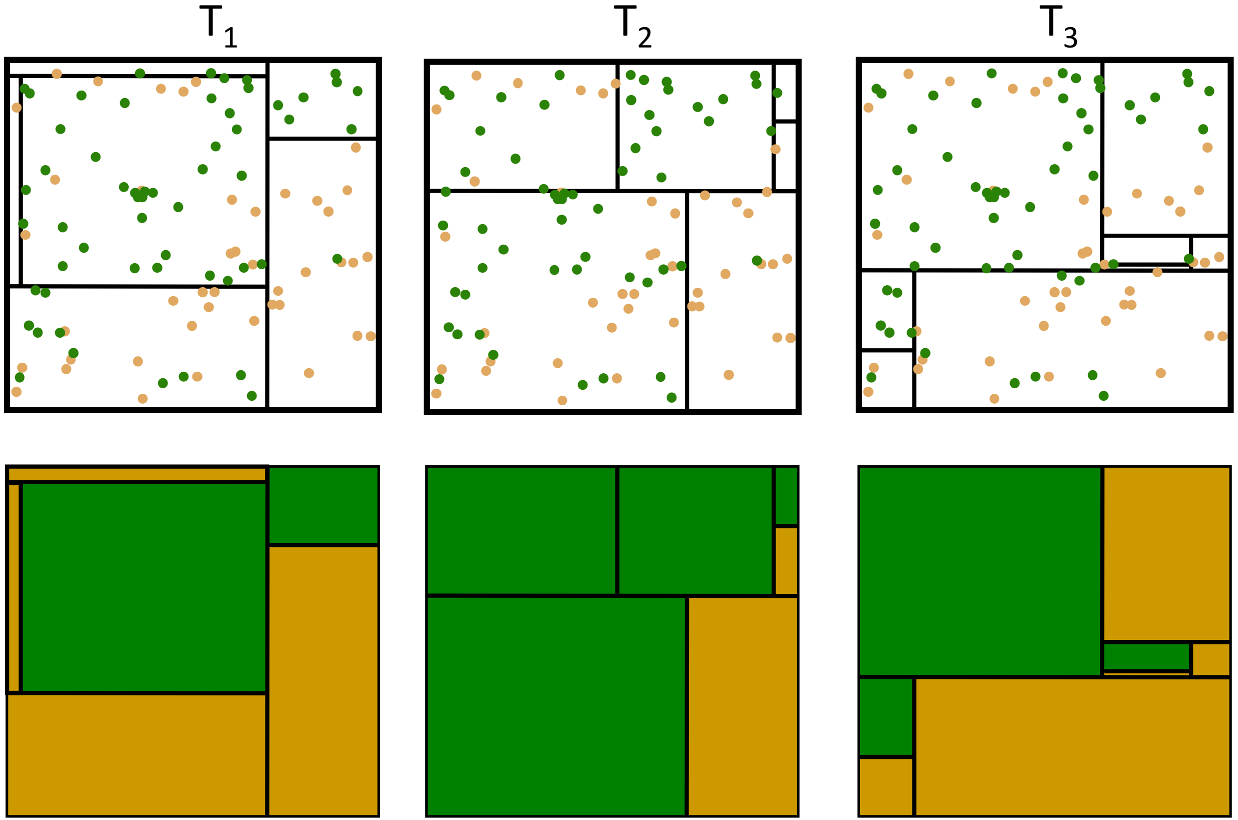
Accuracy = 81%

Accuracy = 80%

Copyright © SAS Institute Inc. All rights reserved.

Perturb

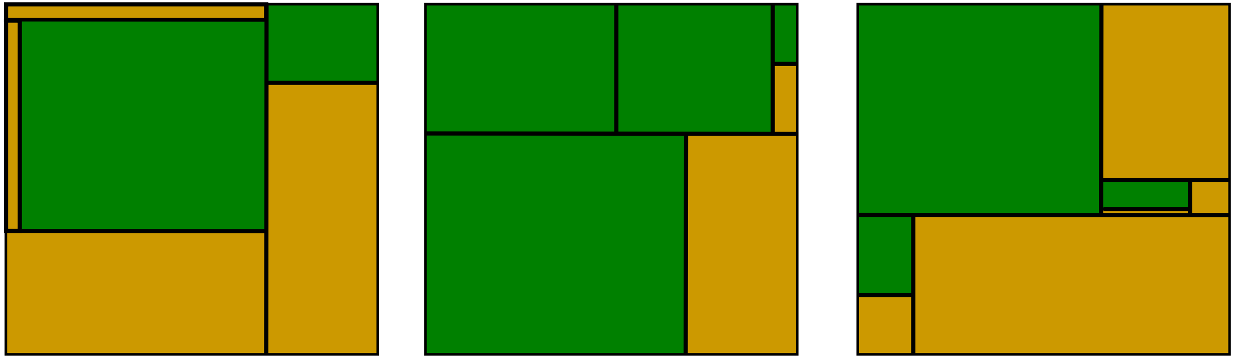
creates different models



Copyright © SAS Institute Inc. All rights reserved.

Perturb

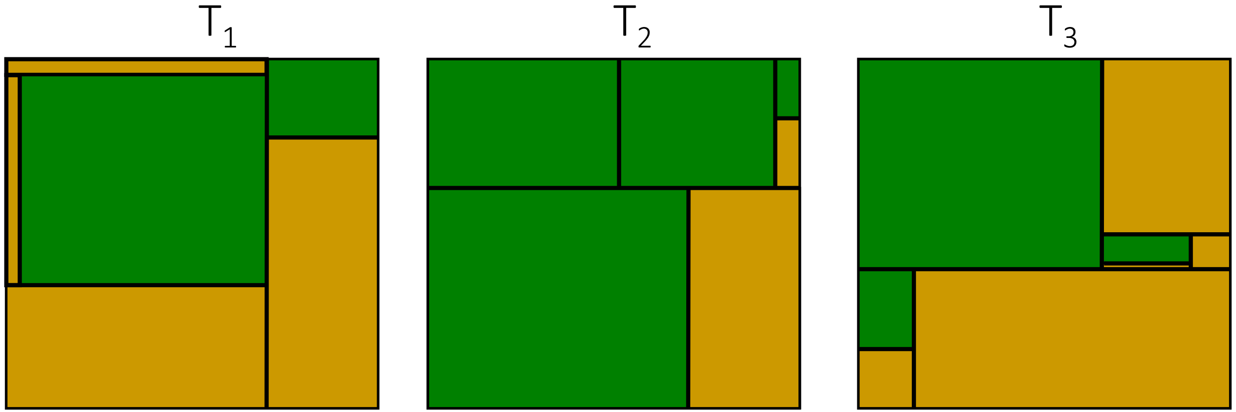
1. **Perturb**: Create different models.
   - resampling
   - subsampling
   - adding noise
   - adaptively reweighting
   - randomly choosing from the competitor splits



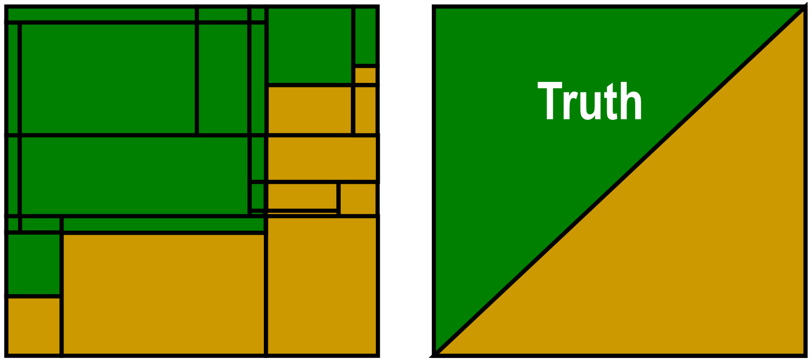
Copyright © SAS Institute Inc. All rights reserved.

Combine

create a single prediction

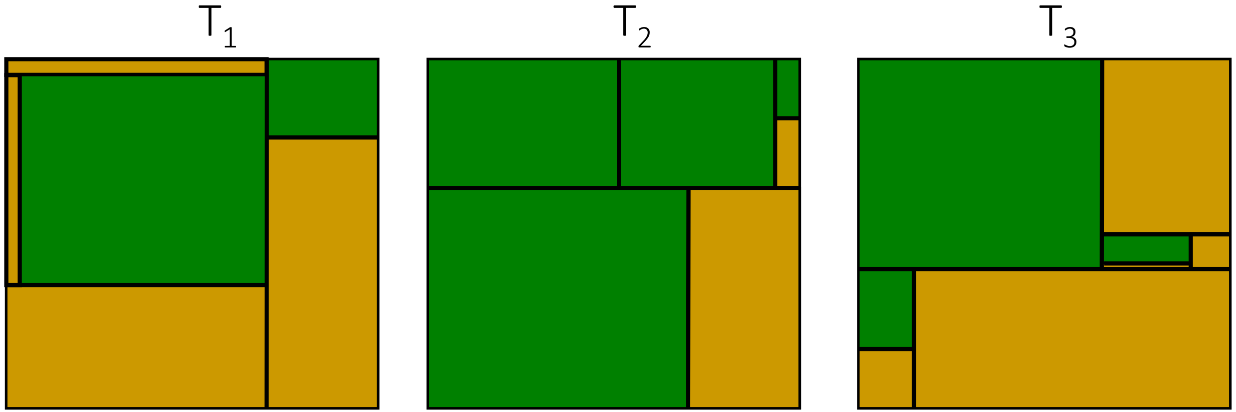


ave(T_1, T_2, T_3) =



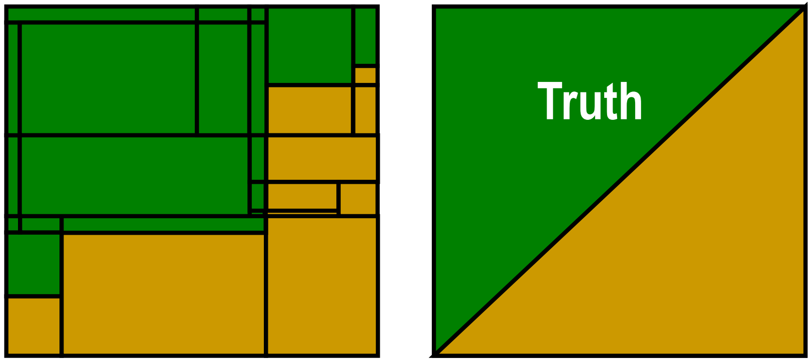
Copyright © SAS Institute Inc. All rights reserved.

Combine



Combine: Create a single prediction.

- voting
- weighted voting
- averaging

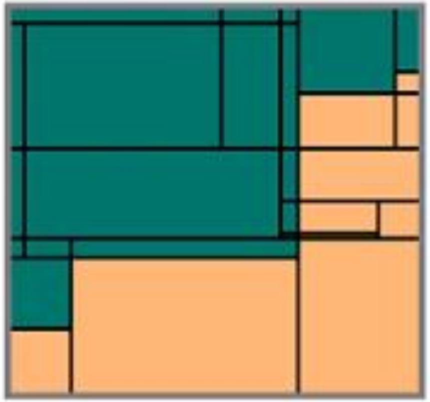


Copyright © SAS Institute Inc. All rights reserved.

P & C Methods

- bagging
- boosting
- ...

Ensemble Model



Bagging

1. Draw *k* bootstrap samples.
2. Build a tree on each bootstrap sample.
3. Combine the predictions by voting or averaging.

Forest

Copyright © SAS Institute Inc. All rights reserved.

Bagging

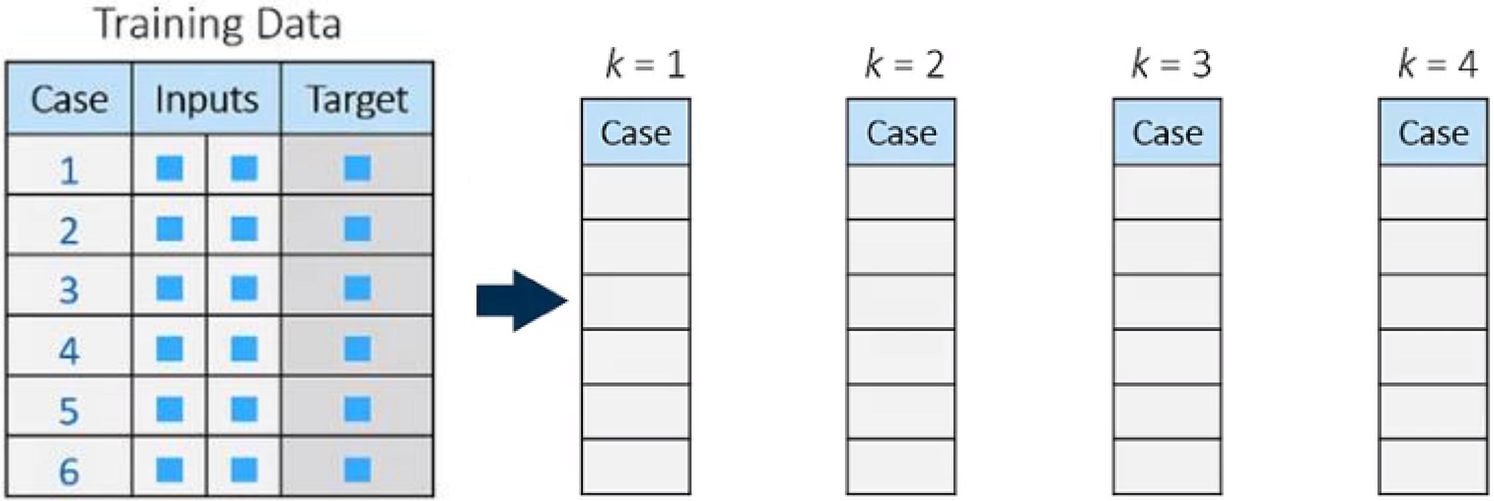
Training Data

Case	Inputs		Target
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

Copyright © SAS Institute Inc. All rights reserved.

Bagging

Step 1. Draw k bootstrap samples.



Copyright © SAS Institute Inc. All rights reserved.

Bagging

Step 1. Draw k bootstrap samples.

Training Data

Case	Inputs		Target
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

$k = 1$

Case	Freq
1	1
2	0
3	2
4	0
5	2
6	1

$k = 2$

Case	Freq
1	0
2	1
3	0
4	2
5	2
6	1

$k = 3$

Case	Freq
1	3
2	1
3	0
4	2
5	0
6	0

$k = 4$

Case	Freq
1	1
2	1
3	2
4	0
5	1
6	1

Copyright © SAS Institute Inc. All rights reserved.

Bagging

Step 2: Build a tree on each bootstrap sample.

Training Data

Case	Inputs		Target
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

$k = 1$

Case	Freq
1	1
2	0
3	2
4	0
5	2
6	1

$k = 2$

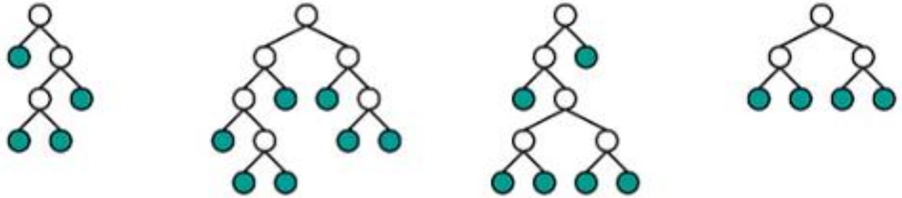
Case	Freq
1	0
2	1
3	0
4	2
5	2
6	1

$k = 3$

Case	Freq
1	3
2	1
3	0
4	2
5	0
6	0

$k = 4$

Case	Freq
1	1
2	1
3	2
4	0
5	1
6	1



Copyright © SAS Institute Inc. All rights reserved.

Bagging

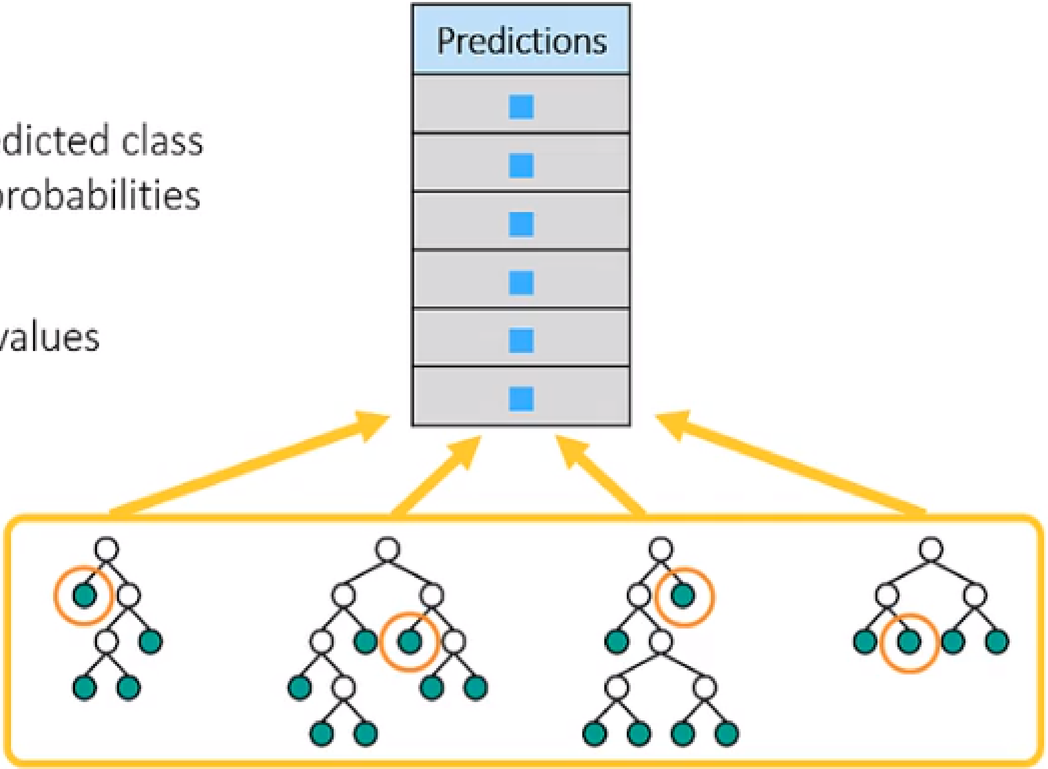
Step 3: Combine the predictions by voting or averaging.

Classification:

- plurality vote of the predicted class
- mean of the posterior probabilities

Interval:

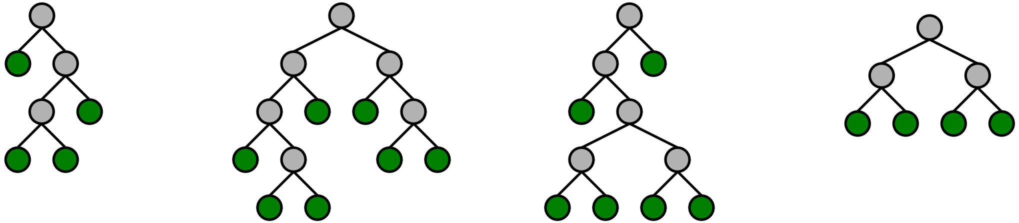
- mean of the predicted values



Copyright © SAS Institute Inc. All rights reserved.

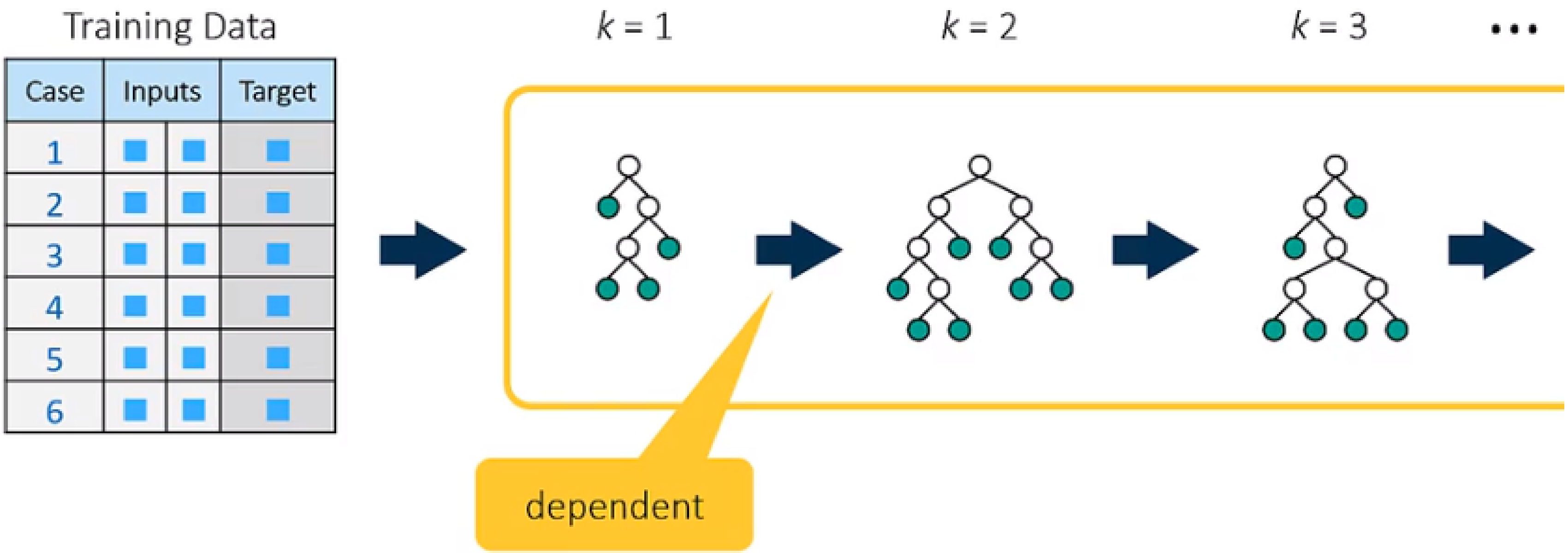
Bagging (Bootstrap Aggregation)

	k=1	k=2	k=3	k=4 …
case	freq	freq	freq	freq
1	1	0	3	1
2	0	1	1	1
3	2	0	0	2
4	0	2	2	0
5	2	2	0	1
6	1	1	0	1



Copyright © SAS Institute Inc. All rights reserved.

Boosting



Copyright © SAS Institute Inc. All rights reserved.

Boosting

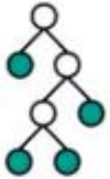
misclassification count

Training Data

Case	Inputs		Target
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

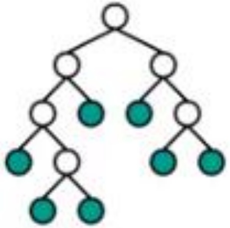
k = 1

Case	Weight	M
1	1	1
2	1	0
3	1	1
4	1	0
5	1	0
6	1	0



k = 2

Case	Weight
1	1.5
2	0.75
3	1.5
4	0.75
5	0.75
6	0.75



Copyright © SAS Institute Inc. All rights reserved.

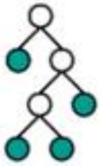
Boosting

Training Data

Case	Inputs		Target
1	■	■	■
2	■	■	■
3	■	■	■
4	■	■	■
5	■	■	■
6	■	■	■

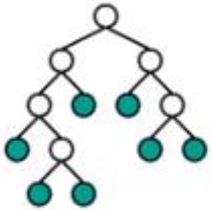
$k = 1$

Case	Weight	M
1	1	1
2	1	0
3	1	1
4	1	0
5	1	0
6	1	0



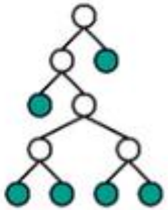
$k = 2$

Case	Weight	M
1	1.5	1
2	0.75	0
3	1.5	2
4	0.75	1
5	0.75	0
6	0.75	0



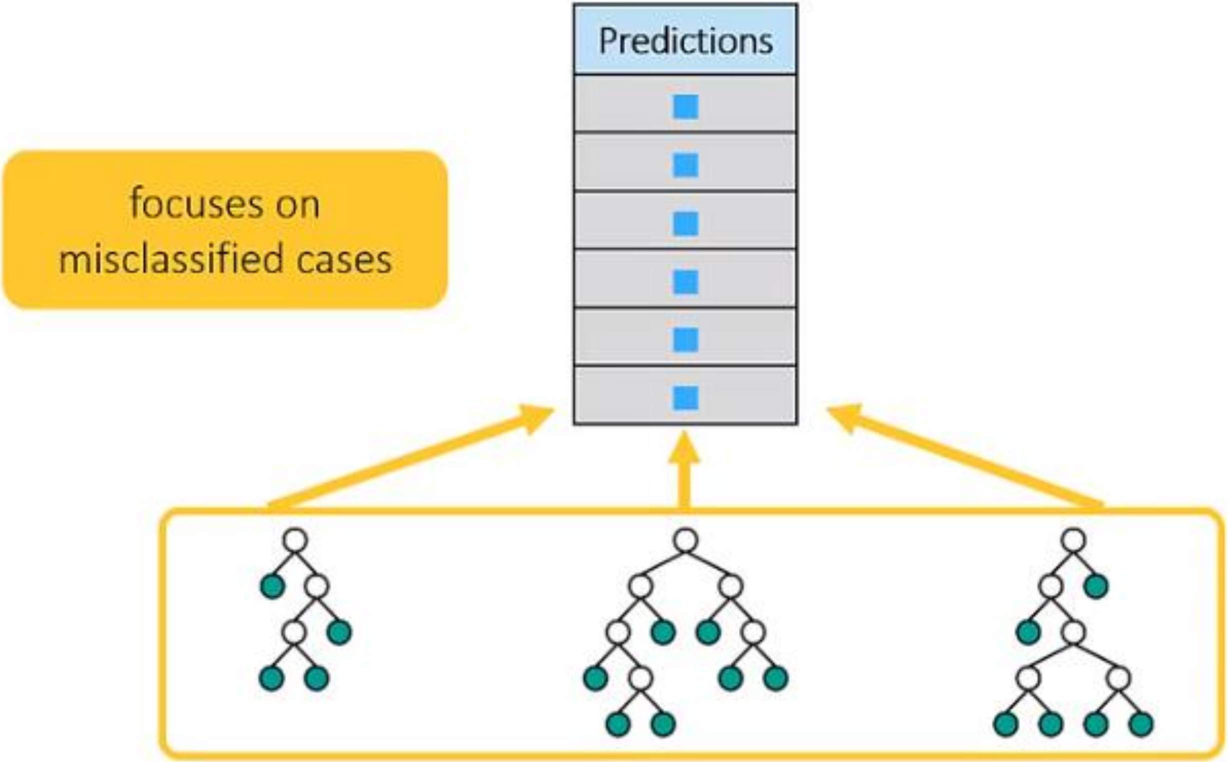
$k = 3$ …

Case	Weight	M
1	0.5	2
2	0.25	0
3	4.25	3
4	1.5	1
5	0.25	0
6	0.25	1



Copyright © SAS Institute Inc. All rights reserved.

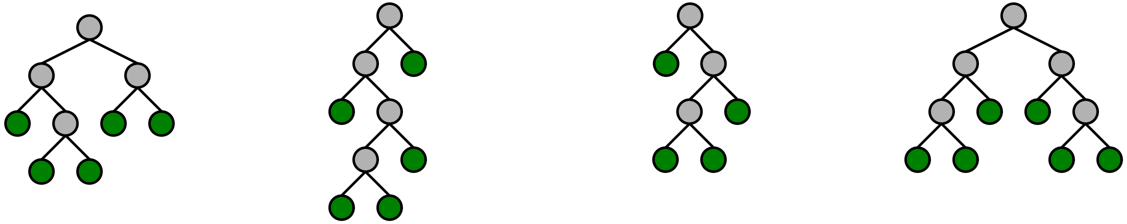
Boosting



Copyright © SAS Institute Inc. All rights reserved.

Boosting

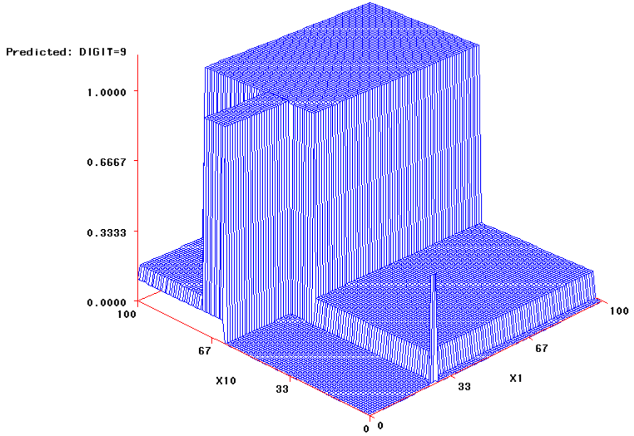
	k=1		k=2		k=3		k=4 ...
case	freq	m	freq	m	freq	m	freq
1	1	1	1.5	1	.5	2	.97
2	1	0	.75	0	.25	0	.06
3	1	1	1.5	2	4.25	3	4.69
4	1	0	.75	1	1.5	1	.53
5	1	0	.75	0	.25	0	.06
6	1	0	.75	0	.25	1	.51

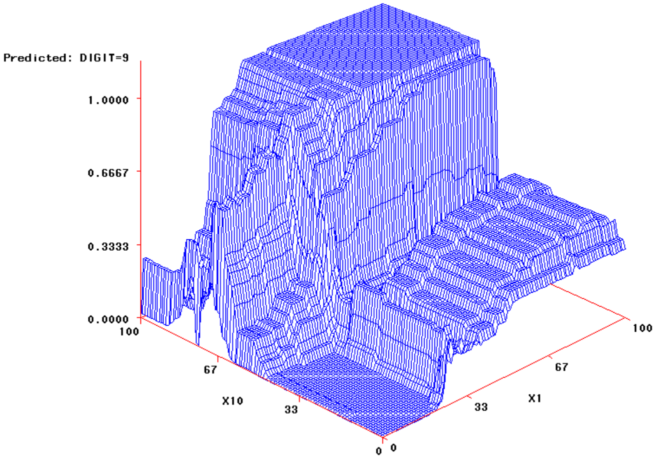


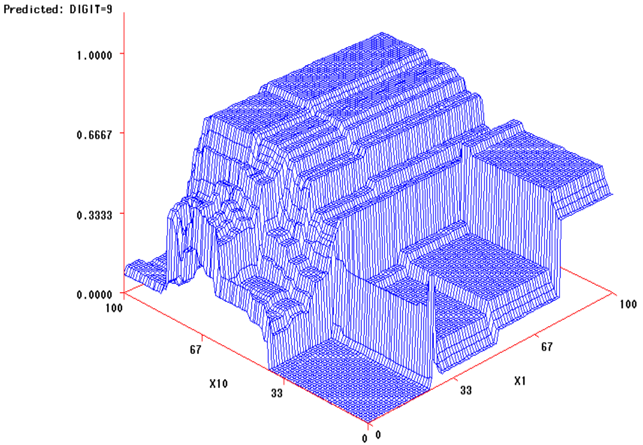
**Shown is Arc-x4, one method of boosting*

Copyright © SAS Institute Inc. All rights reserved.

Single, Bagged, and Boosted Tree







Gradient Boosting with Decision Trees

- The gradient boosting algorithm is similar to standard boosting, except at each iteration, the target is the residual from the previous decision tree model.
- At each step, the accuracy of the tree is computed.
- Successive samples are adjusted to accommodate previous inaccuracies.
- The model is a weighted ($\beta_1 \ldots \beta_M$) linear combination of (usually) simple models.

Copyright © SAS Institute Inc. All rights reserved.

Autotuning

- Search for the best combination of values in different properties:
  - Regularization (L1 and L2)
  - Learning rate
  - Number of inputs per split
  - Number of iterations (trees)
  - Subsample rate
  - Search method
    - Bayesian, Genetic algorithm, Latin hypercube sample, Random
  - Validation method
    - Partition, cross validation
  - Objective function (class and interval targets)

***Note**: Quantile binning usually does better than bucket binning, which is default.*

Copyright © SAS Institute Inc. All rights reserved.

Building a Gradient Boosting Model

In this demonstration, you add a Gradient Boosting node to the Starter Template pipeline. You build a default Gradient Boosting model, change some of the settings, and compare the model to the other models in the pipeline.

Copyright © SAS Institute Inc. All rights reserved.

Exercise

This exercise reinforces the concepts discussed previously. You use Model Studio to build a Gradient Boosting model using the Autotune feature.

Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true regarding tree-based models?

a. Small changes in the training data can cause large changes in the structure of a tree.

b. Ensemble models are used only with decision trees.

c. In the boosting algorithm, cases that are correctly classified are given more weight in subsequent models.

d. In the bagging algorithm, the training data are resampled without replacement.

Copyright © SAS Institute Inc. All rights reserved.

Which of the following statements is true regarding tree-based models?

a. Small changes in the training data can cause large changes in the structure of a tree.

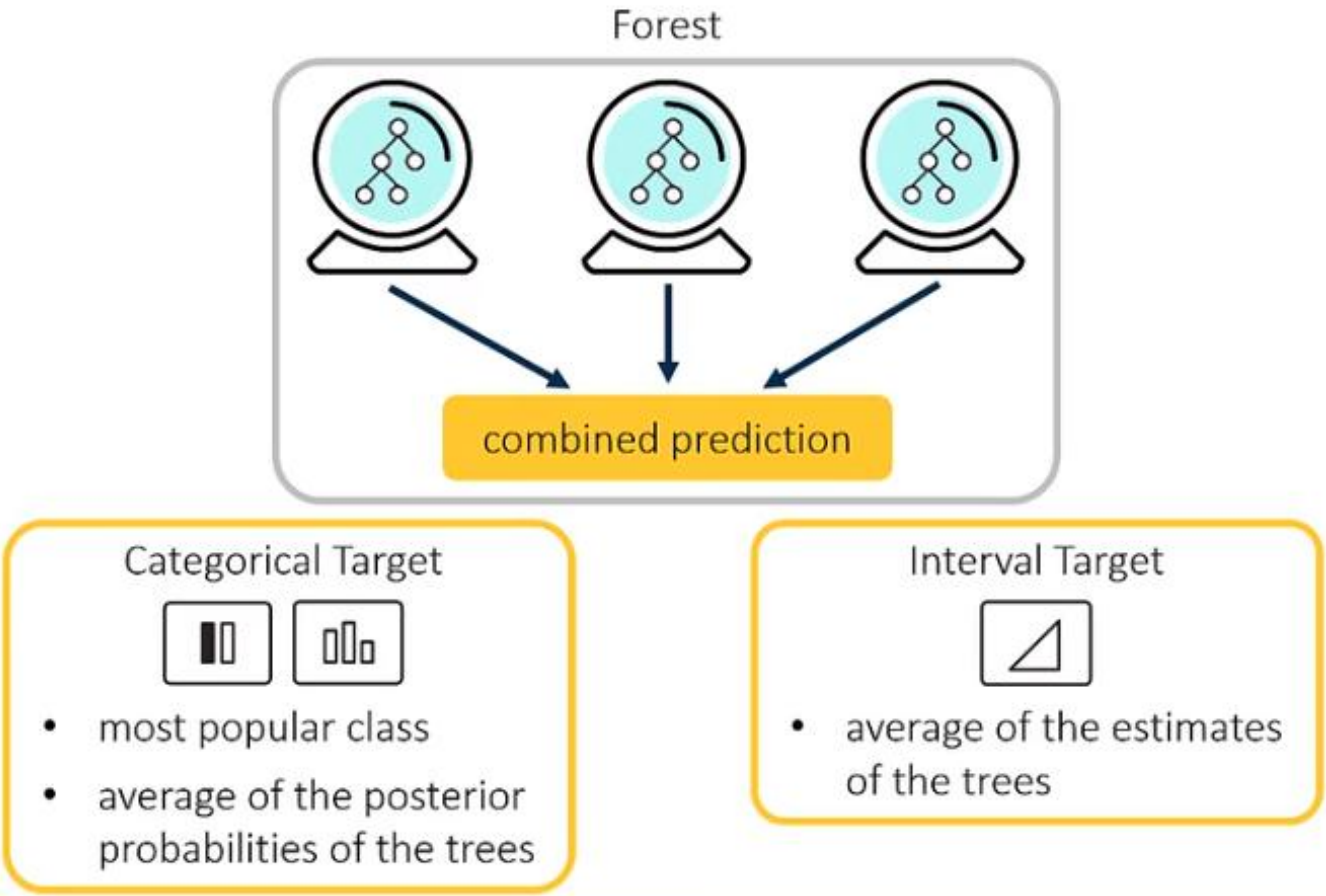
b. Ensemble models are used only with decision trees.

c. In the boosting algorithm, cases that are correctly classified are given more weight in subsequent models.

d. In the bagging algorithm, the training data are resampled without replacement.

Copyright © SAS Institute Inc. All rights reserved.

Forest Models



Copyright © SAS Institute Inc. All rights reserved.

Forest Models

- A *forest model* is an ensemble of classification, or regression, trees.
- Trees in the forest differ from each other in two ways:
  - Training data for a tree is a sample with replacement from all observations.
  - Input variables considered for splitting a node are randomly selected from available inputs. Only the variable most associated with the target is split for that node.

Copyright © SAS Institute Inc. All rights reserved.

Forest Algorithm

- Recall that bagging takes bootstrap samples of the rows of training data. All columns are considered for splitting at every step.
- The forest algorithm samples the rows ***and*** the columns at each step.
- The forest algorithm perturbs the training data more than the bagging algorithm.
- This increased variation among the trees in the ensemble often leads to improved predictive accuracy.

Copyright © SAS Institute Inc. All rights reserved.

Out-of-Bag Sample

- The out-of-bag sample refers to the training data that is excluded during the construction of an individual tree.
- Observations in the training data that are used to construct an individual tree are the bagged sample.
- Some model assessments such as the iteration plots are computed using the out-of-bag sample as well as all the training data.

Copyright © SAS Institute Inc. All rights reserved.

Autotuning

- Search for the best combination of values in different properties:
  - Maximum depth
  - Number of trees
  - In-bag sample proportion
  - Number of inputs per split
  - Search method
    - Bayesian, Genetic algorithm, Latin hypercube sample, Random
  - Validation method
    - Partition, cross validation
  - Objective function (class and interval targets)

Copyright © SAS Institute Inc. All rights reserved.

3.04 Multiple Choice Poll

Which of the following statements is true regarding tree-based models?

a. Small changes in the training data can cause large changes in the topology of a tree.

b. Ensemble models are used only with decision trees.

c. In the boosting algorithm, cases that are correctly classified are given more weight in subsequent models.

d. In the bagging algorithm, the training data is resampled without replacement.

Copyright © SAS Institute Inc. All rights reserved.

3.04 Multiple Choice Poll – Correct Answer

Which of the following statements is true regarding tree-based models?

a. Small changes in the training data can cause large changes in the topology of a tree.
b. Ensemble models are used only with decision trees.
c. In the boosting algorithm, cases that are correctly classified are given more weight in subsequent models.
d. In the bagging algorithm, the training data is resampled without replacement.

Copyright © SAS Institute Inc. All rights reserved.

Modeling a Binary Target with a Forest

In this demonstration, you add a Forest node to the Chapter 3 pipeline. You build a default Forest model, change some of the settings, and compare the model to the other models in the pipeline.

Copyright © SAS Institute Inc. All rights reserved.

Exercise

This exercise reinforces the concepts discussed previously. You use Model Studio to build a Forest model using the Autotune feature.

Copyright © SAS Institute Inc. All rights reserved.