FINDING THE RIGHT-SIZED MODEL WITH ENTERPRISE MINER: MODEL TUNING AND GENERALIZATION

Lorne Rothman, PhD, PStat
Principal Statistician, SAS Institute (Canada) Inc.
Lorne.Rothman@sas.com

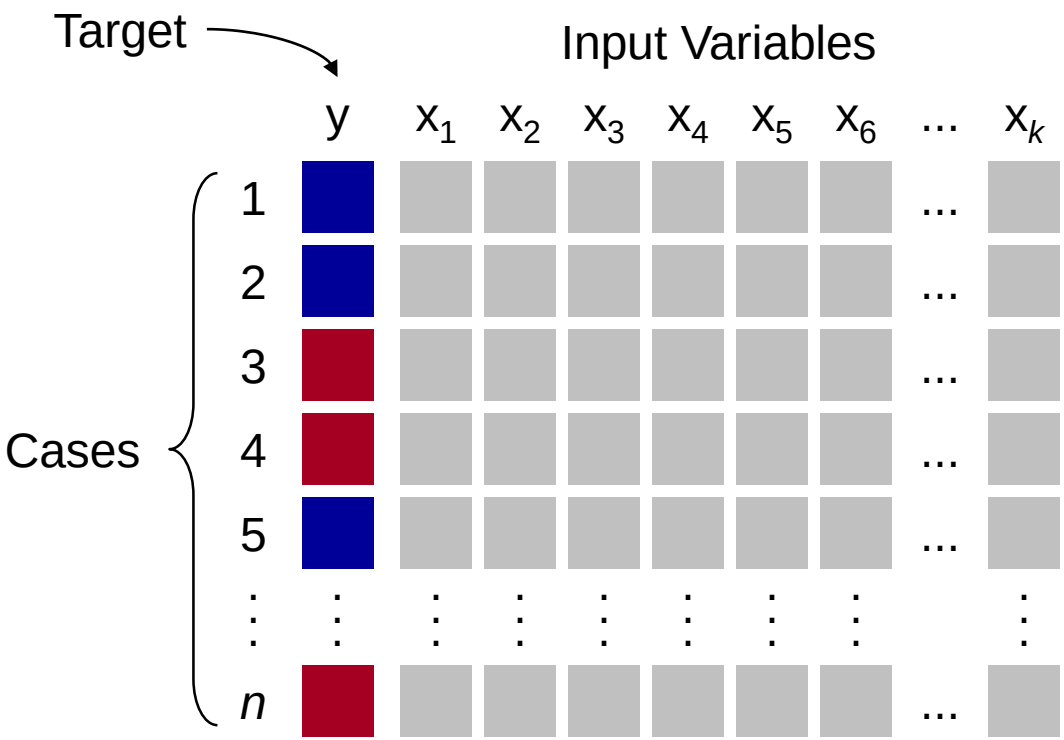
GENERALIZATION AND THE RIGHT SIZED MODEL: TOPICS

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble models

GENERALIZATION AND THE RIGHT SIZED MODEL: OVERVIEW

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble Models

MODEL BUILDING

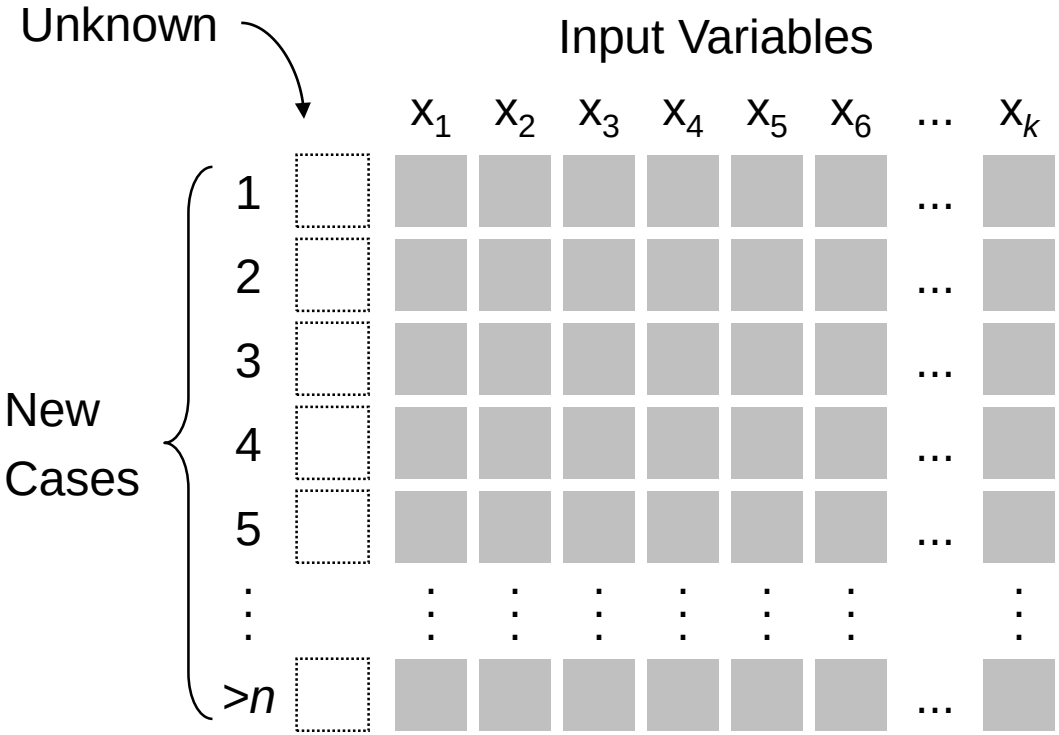


EXPLANATORY VS PREDICTIVE MODELING

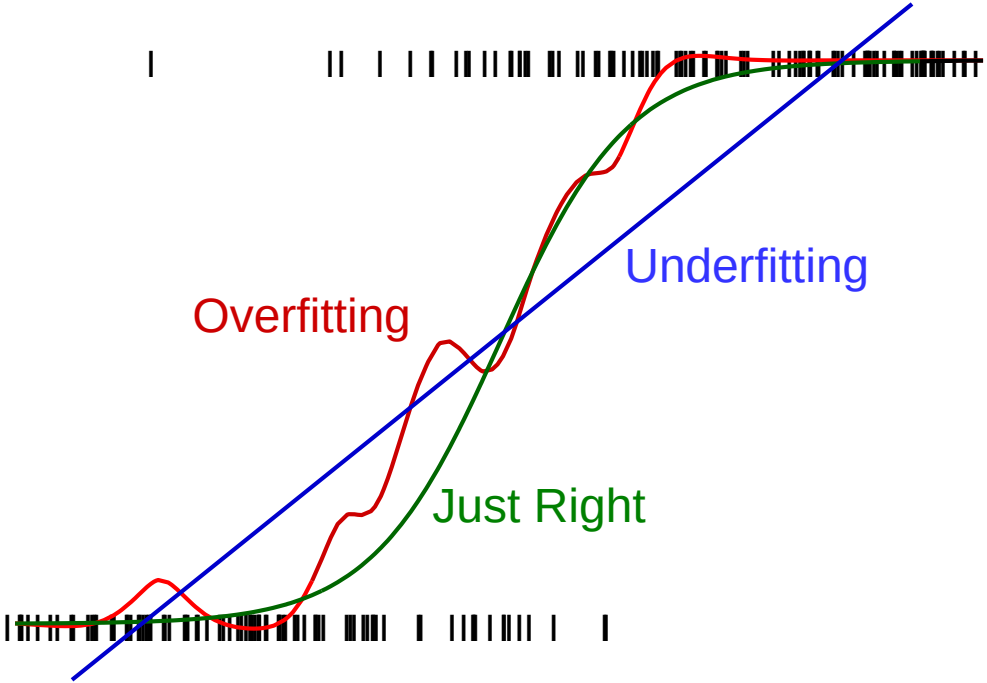
- **Traditional Explanatory Modeling**
  - Smaller samples
  - Small number of model inputs selected based on theory
  - Strong emphasis on probability theory (p-values) to make inferences from samples to the population
  - Purpose is explanation
  - Models are fit and assessed on the same data.

- **Predictive Modeling**
  - Larger samples
  - Large number of inputs selected based upon model performance
  - Less emphasis on probability theory; focus is on Empirical Validation to generalize results to the population
  - Purpose is prediction
  - Models are assessed on hold-out data.

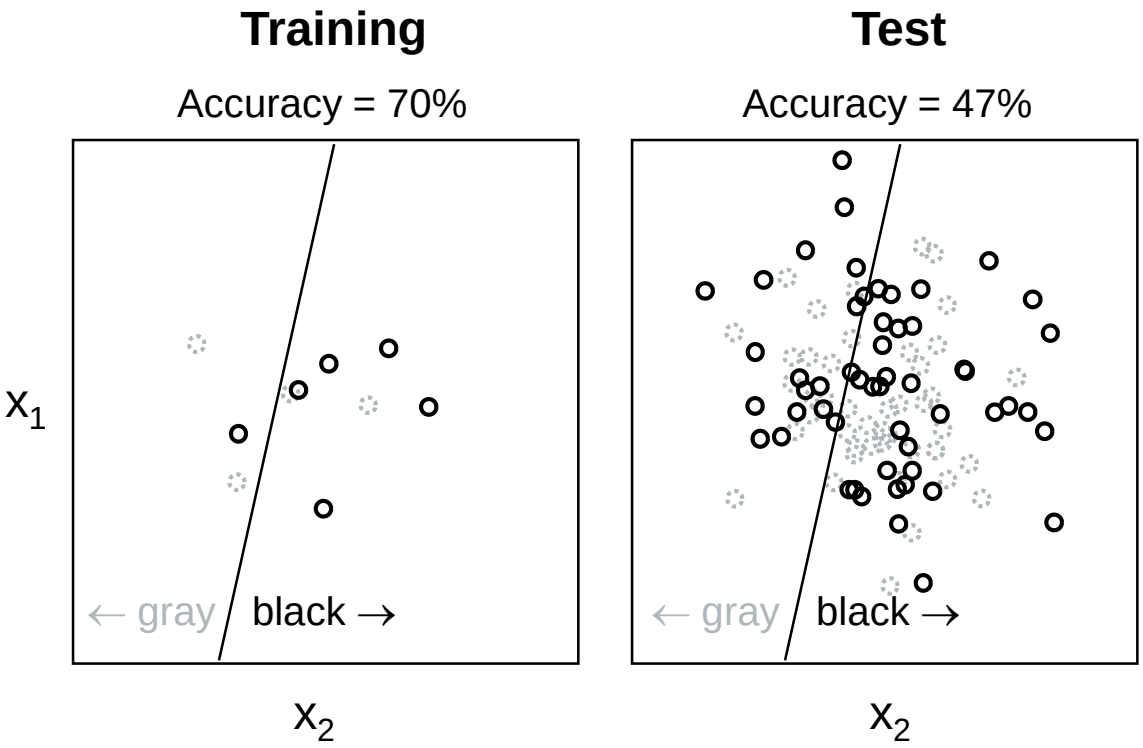
GENERALIZATION



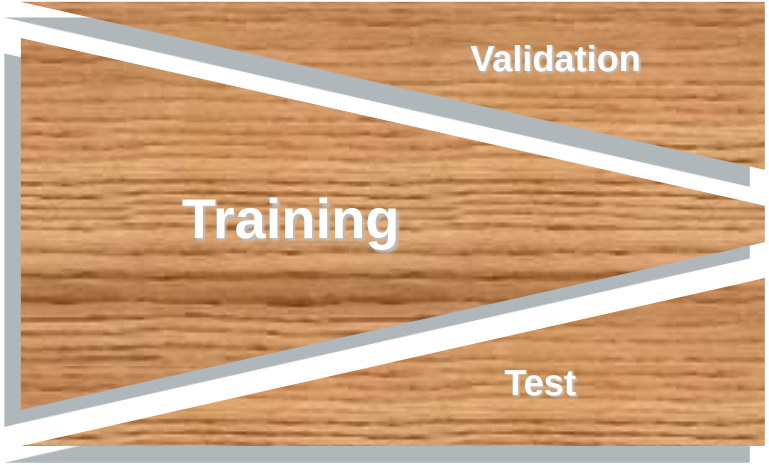
MODEL SELECTION



THE OPTIMISM PRINCIPLE



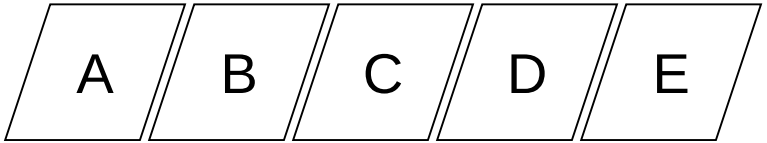
DATA SPLITTING



- Comparison
- Selection
- Tuning

- Final Assessment

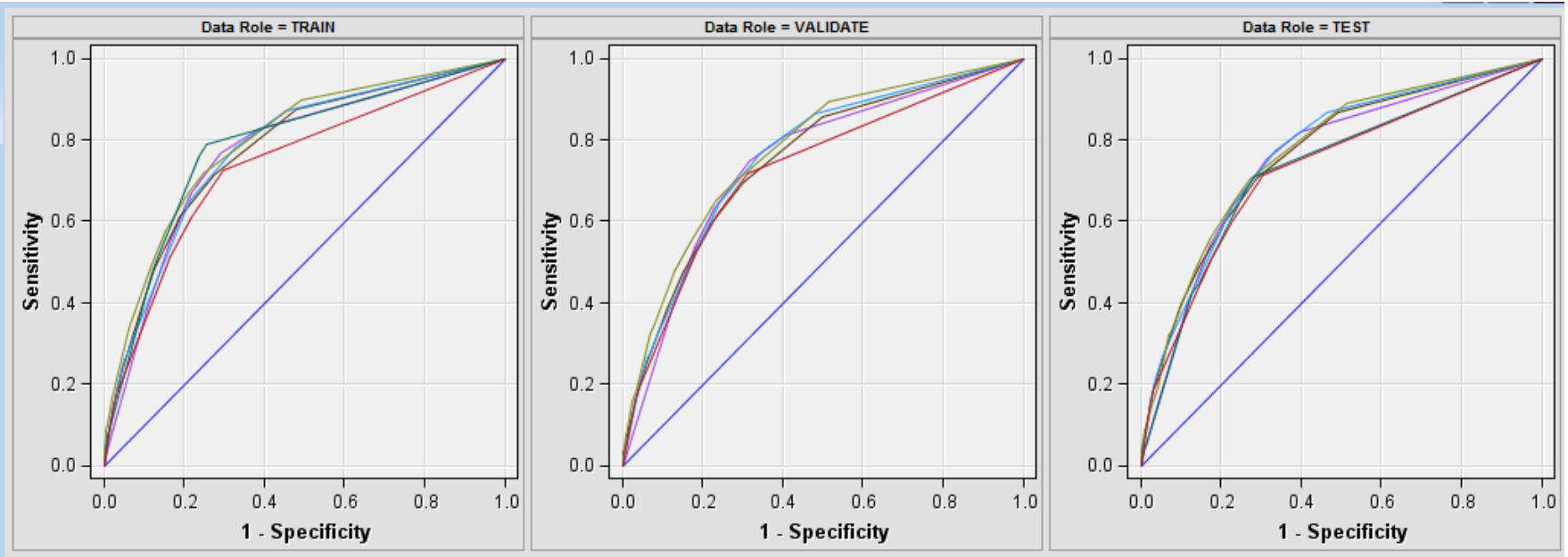
CROSS VALIDATION



	Train	Validate
1)	BCDE	A
2)	ACDE	B
3)	ABDE	C
4)	ABCE	D
5)	ABCD	E

**Enterprise Miner supports Cross validation for decision trees and leave-one-out cross validation for regression.

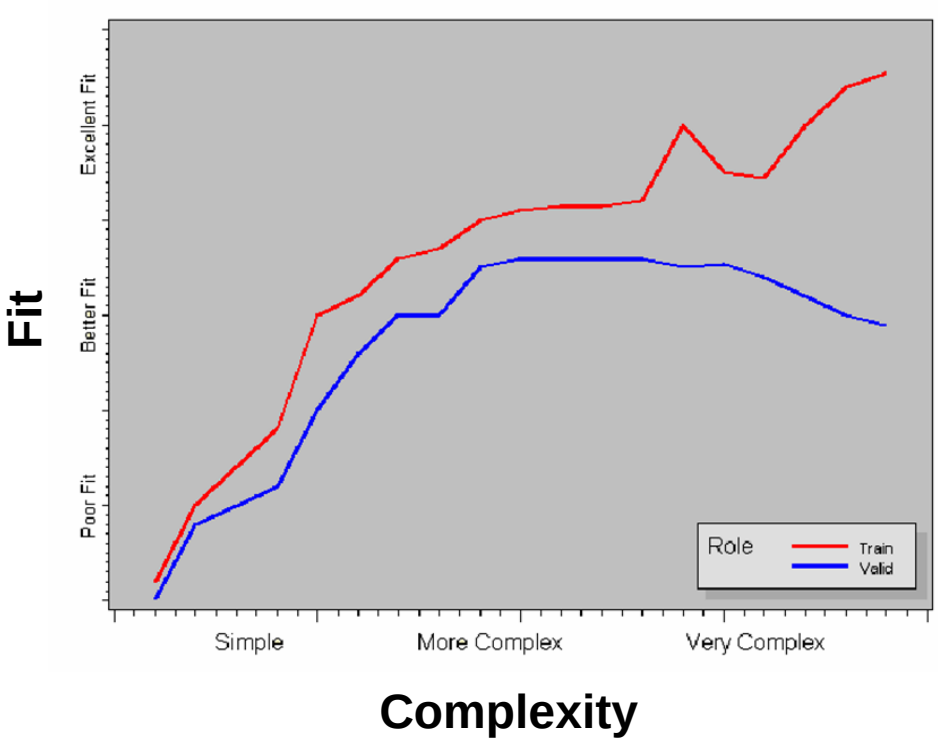
ASSESSMENT ON HOLDOUT DATA



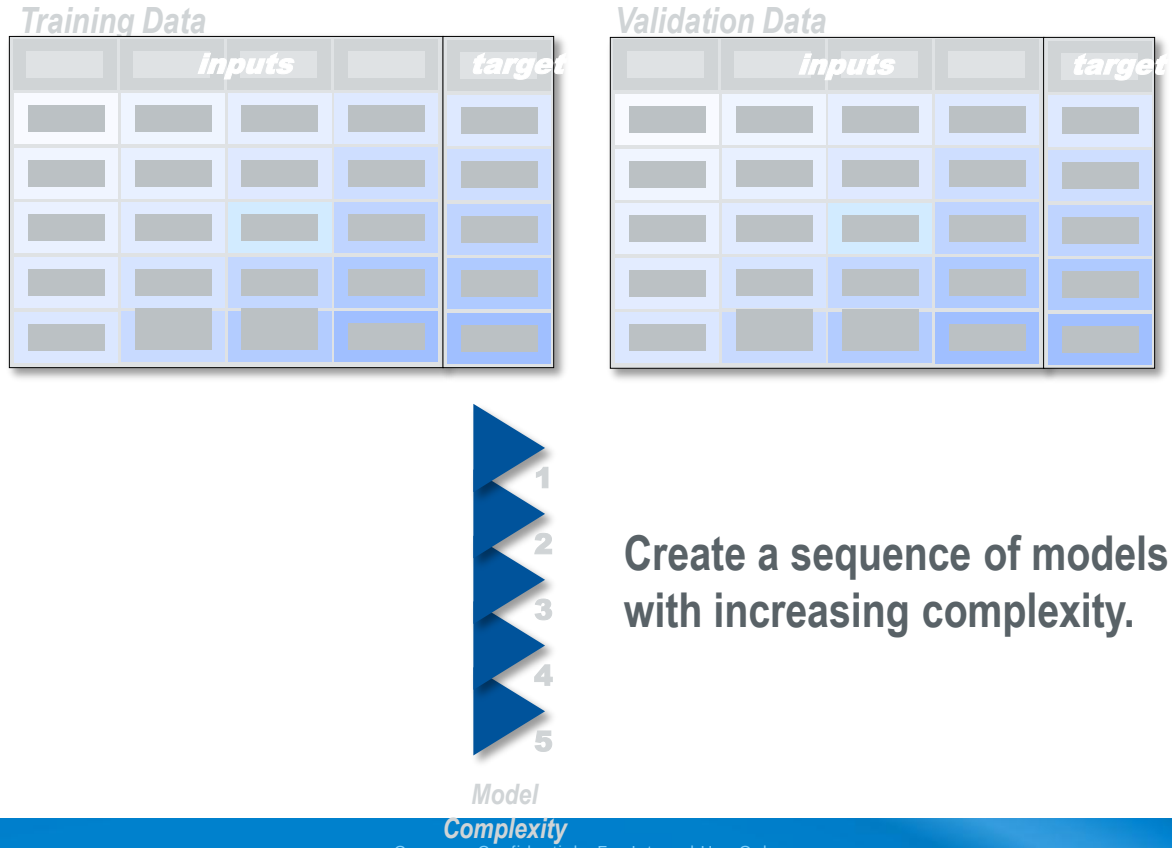
TOPICS

- Predictive modeling and empirical validation
- **Model tuning with Validation Data**
  - **Regression, Decision Trees, Neural Networks**
- Penalized measures of fit in Regression
- Ensemble Models

FIT VERSUS COMPLEXITY

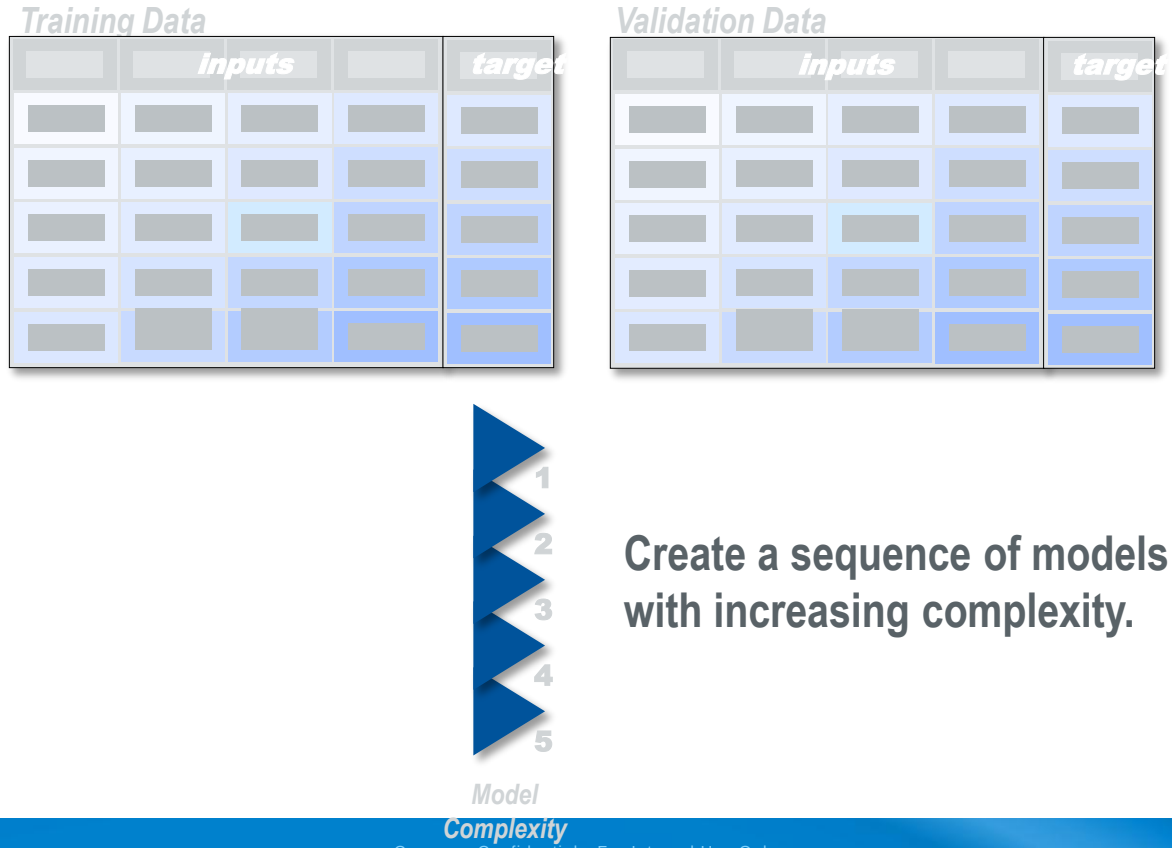


GENERAL MODEL TUNING



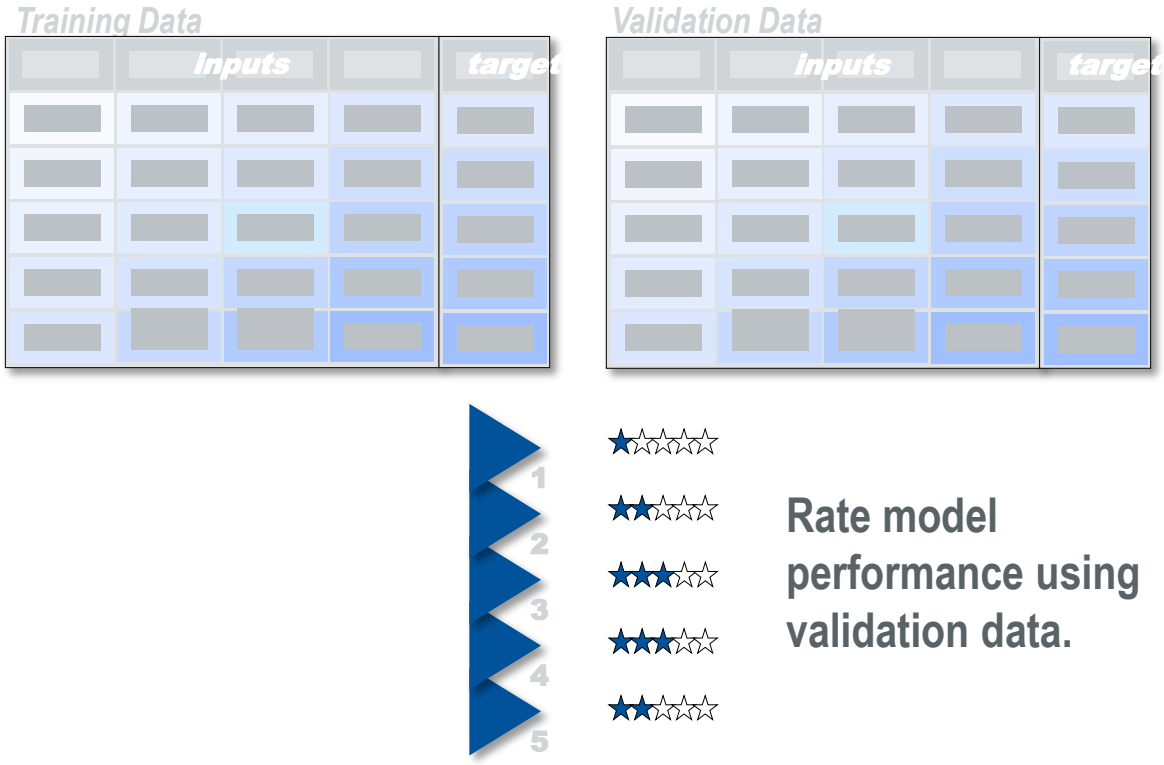
Create a sequence of models with increasing complexity.

GENERAL MODEL TUNING



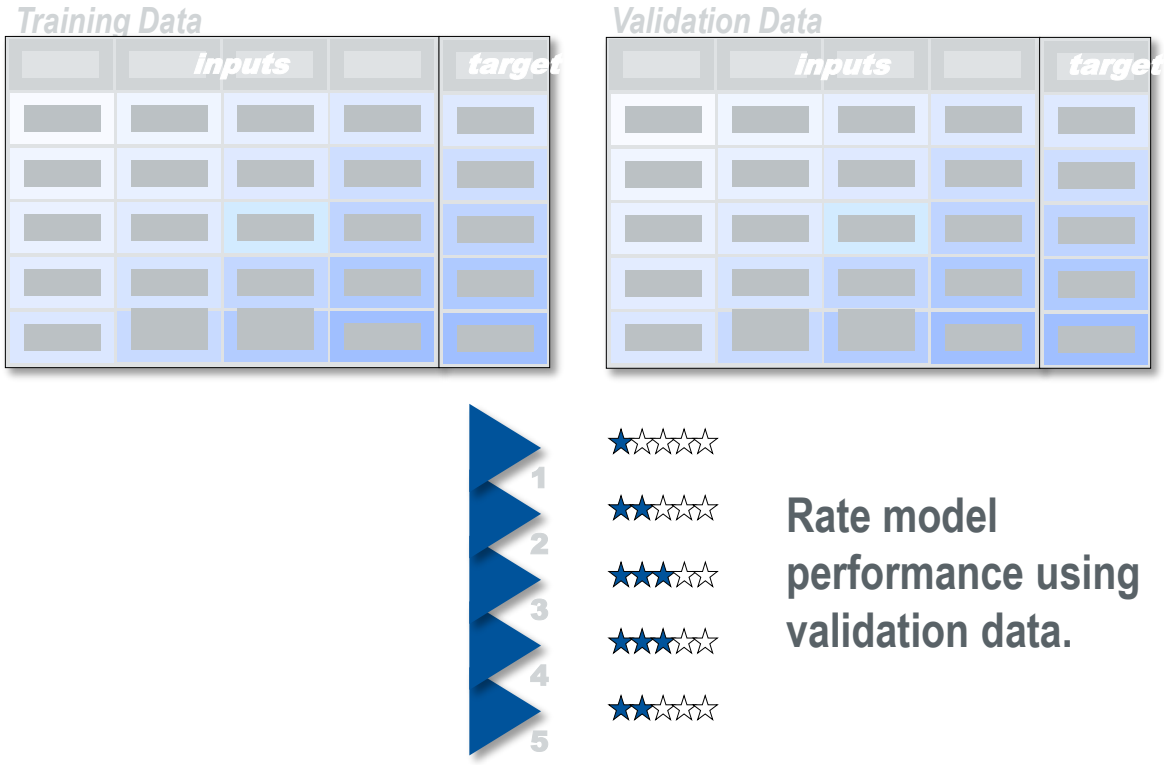
Create a sequence of models with increasing complexity.

GENERAL MODEL TUNING



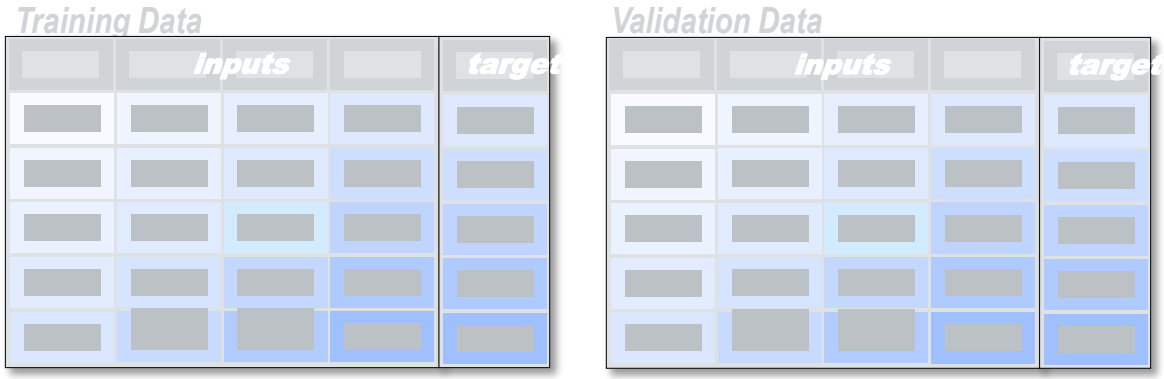
Rate model performance using validation data.

GENERAL MODEL TUNING



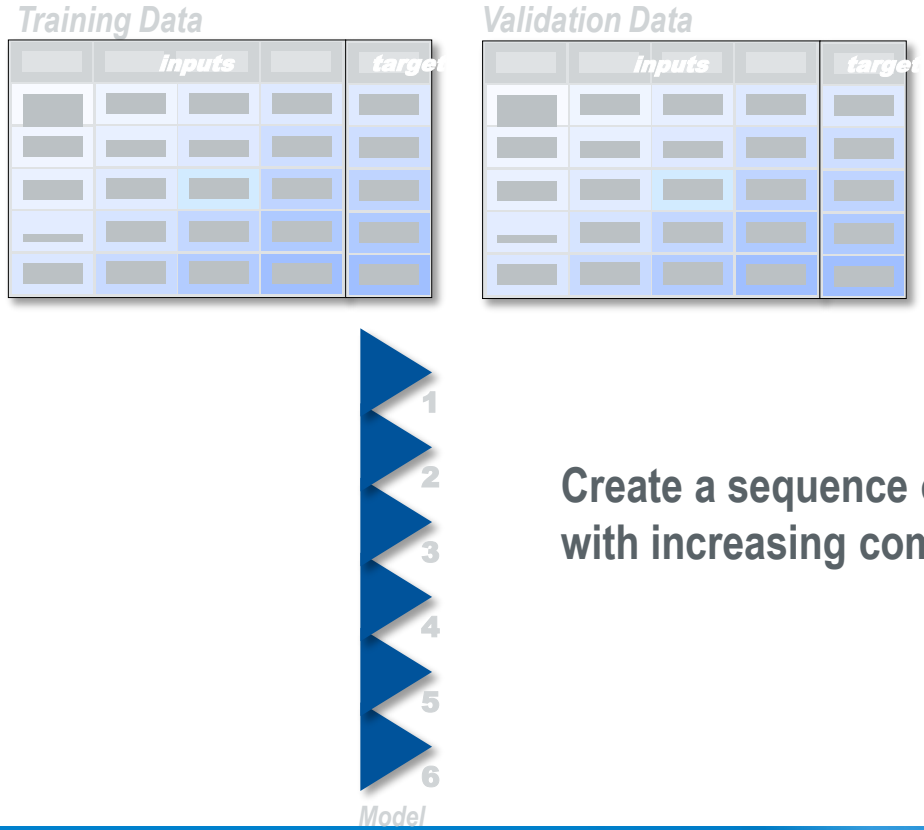
Rate model performance using validation data.

GENERAL MODEL TUNING



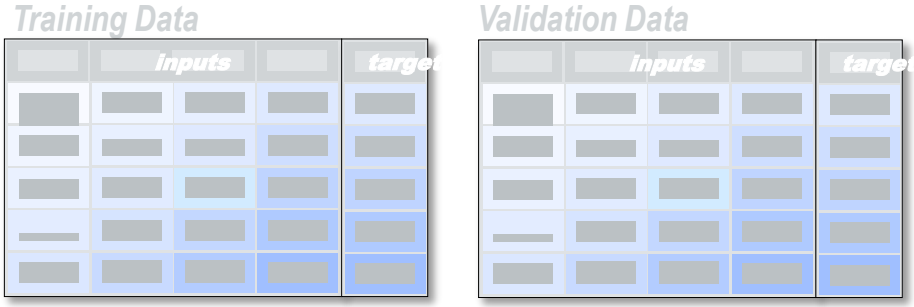
Select the simplest model with the highest validation assessment.

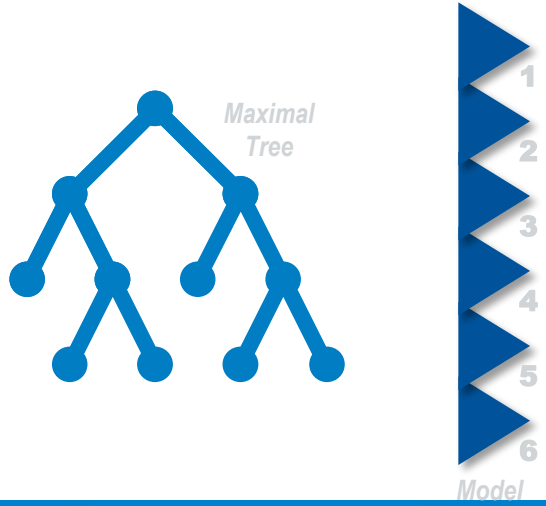
MODEL TUNING: DECISION TREES



Create a sequence of models with increasing complexity.

MODEL TUNING: DECISION TREES

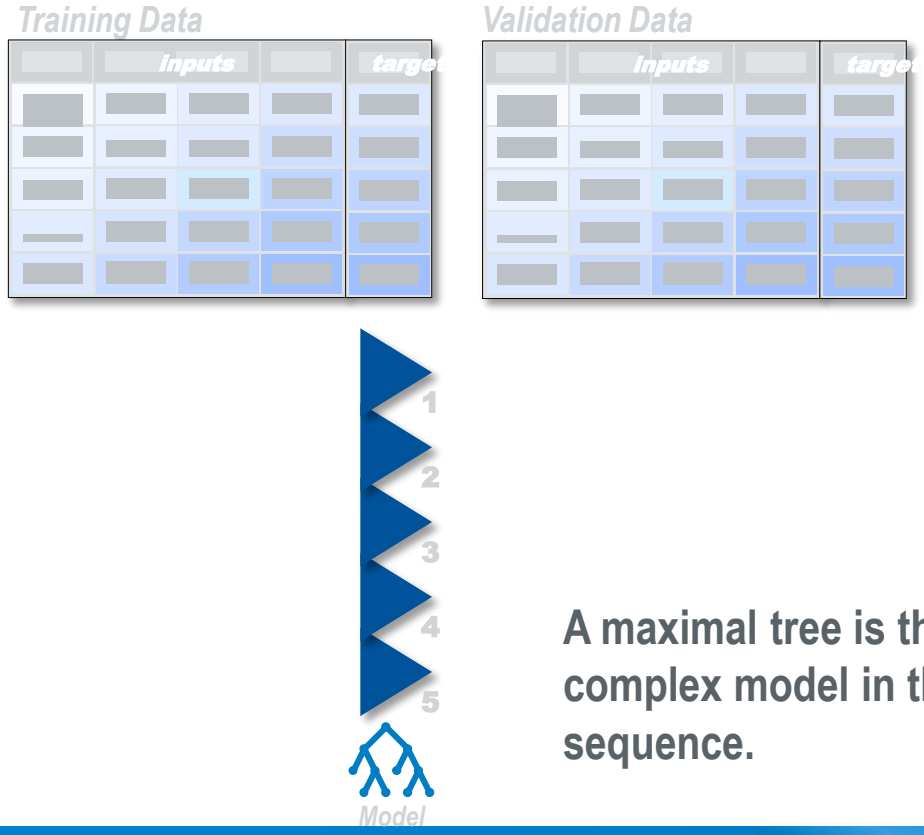




Create a sequence of models with increasing complexity.

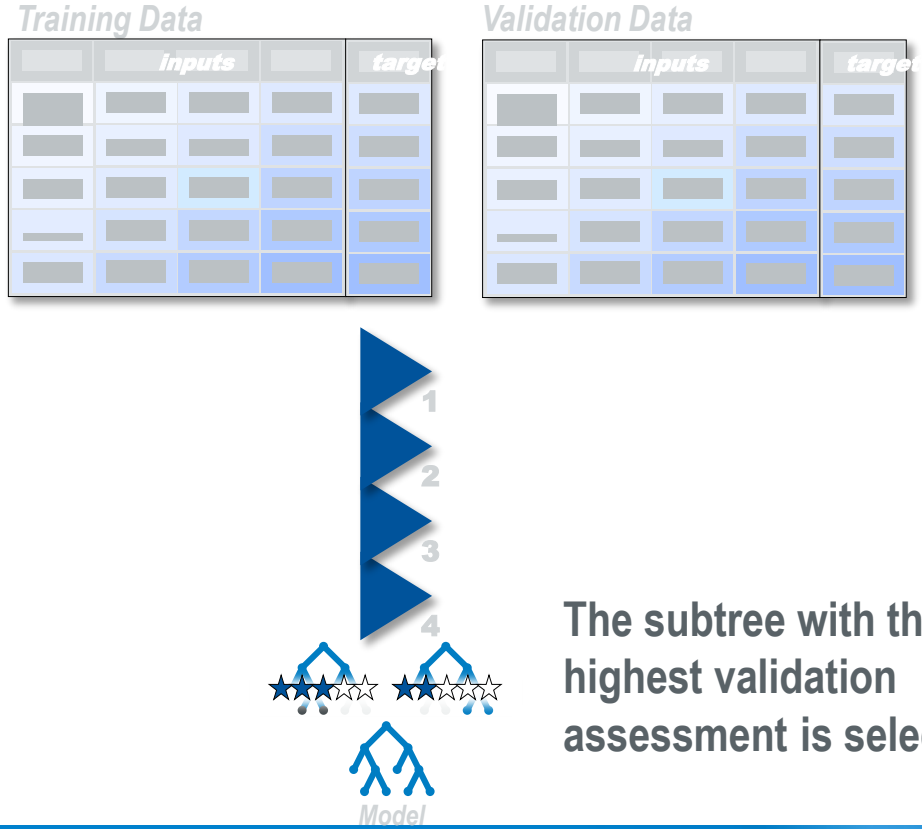
A maximal tree is the most complex model in the sequence.

MODEL TUNING: DECISION TREES



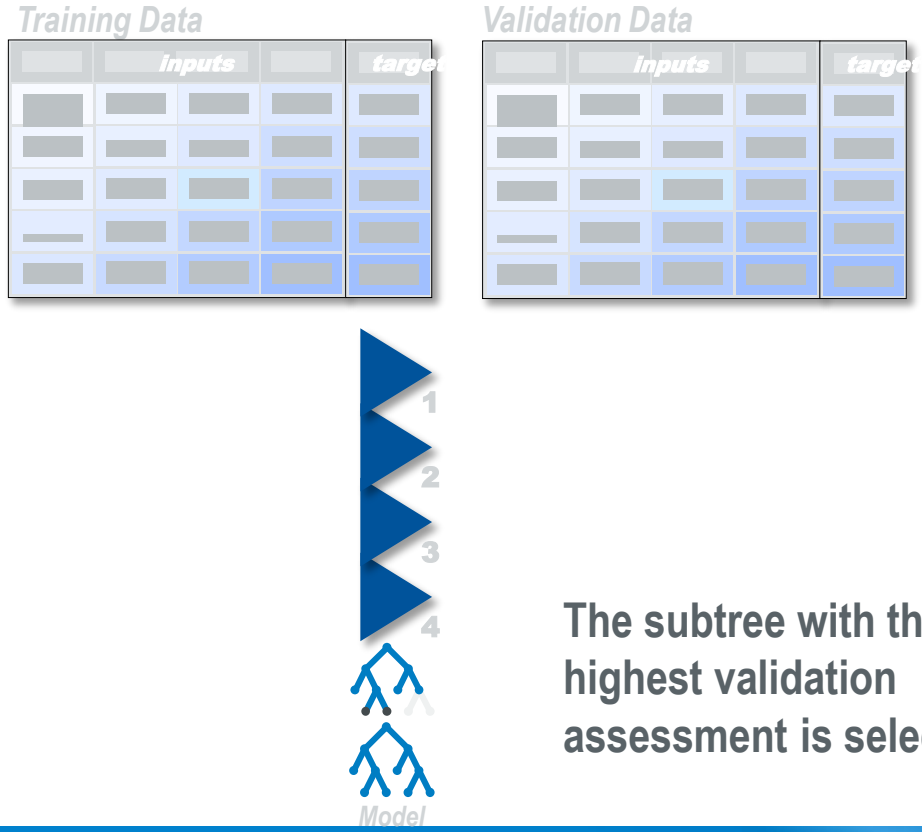
A maximal tree is the most complex model in the sequence.

MODEL TUNING: DECISION TREES



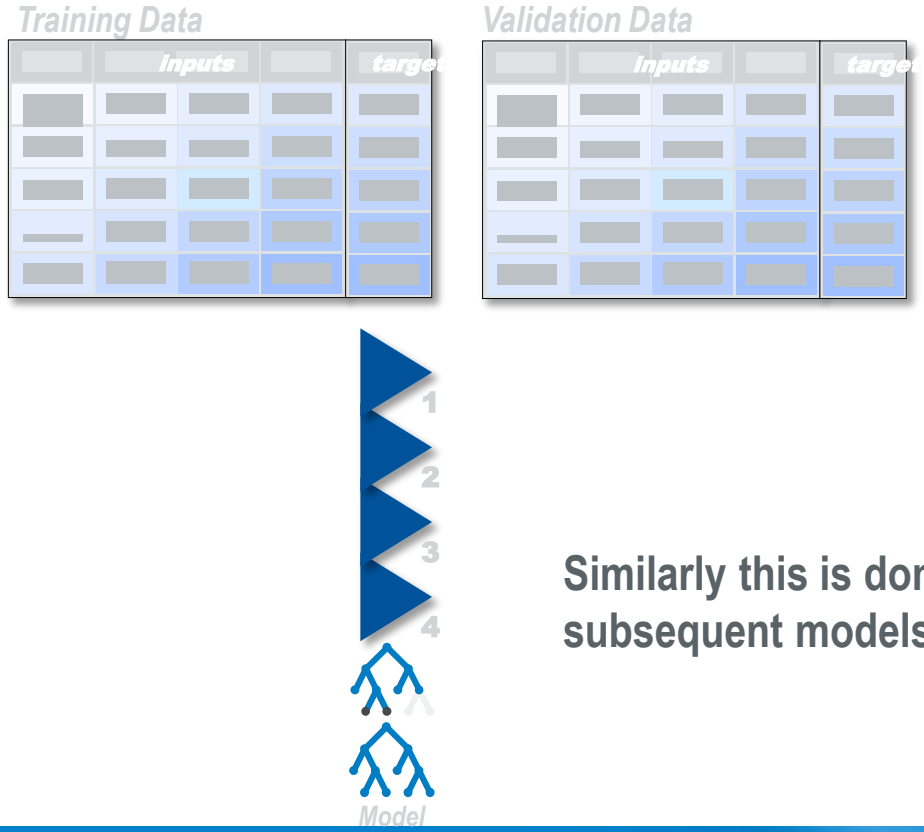
The subtree with the highest validation assessment is selected.

MODEL TUNING: DECISION TREES



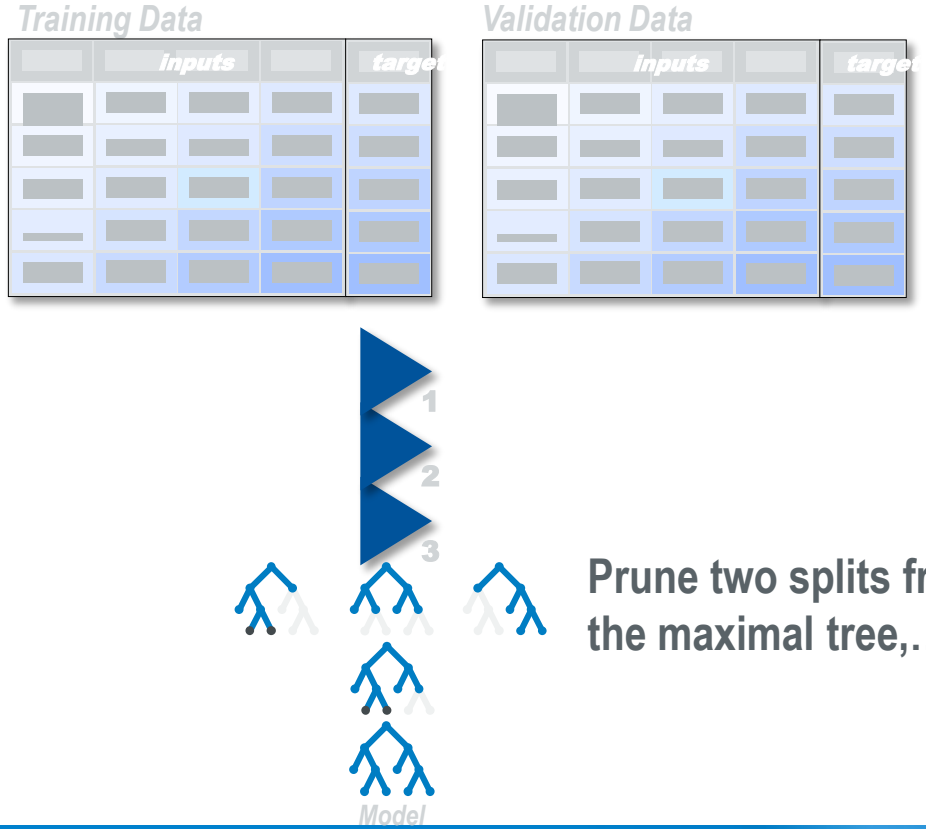
The subtree with the highest validation assessment is selected.

MODEL TUNING: DECISION TREES



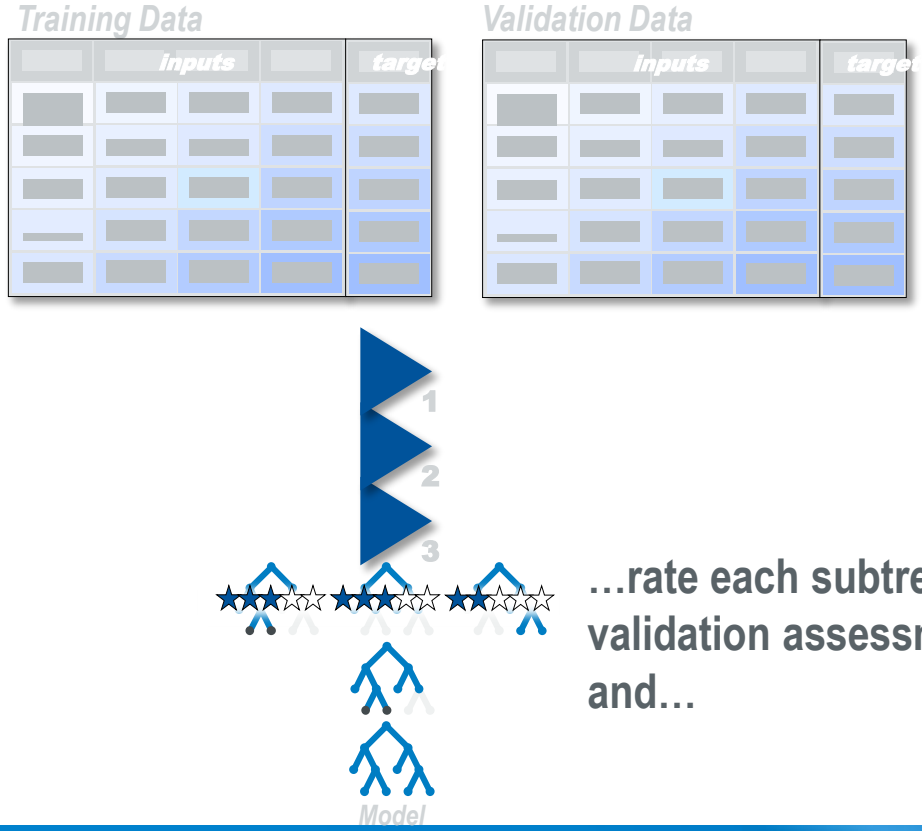
Similarly this is done for subsequent models.

MODEL TUNING: DECISION TREES



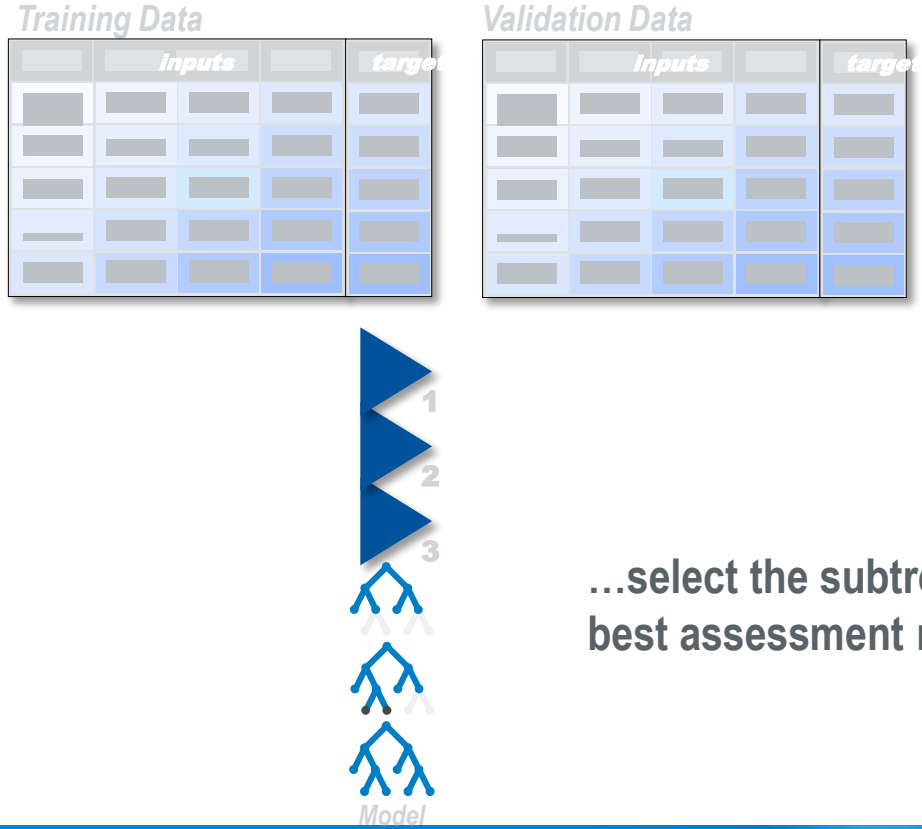
Prune two splits from the maximal tree,…

MODEL TUNING: DECISION TREES



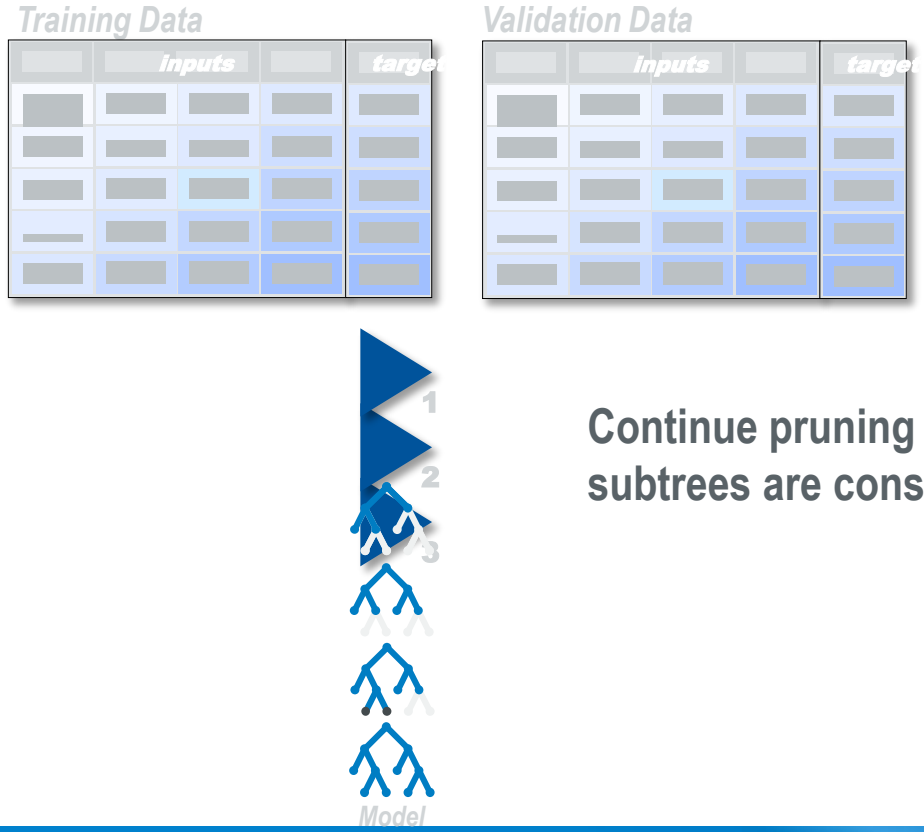
…rate each subtree using validation assessment, and…

MODEL TUNING: DECISION TREES



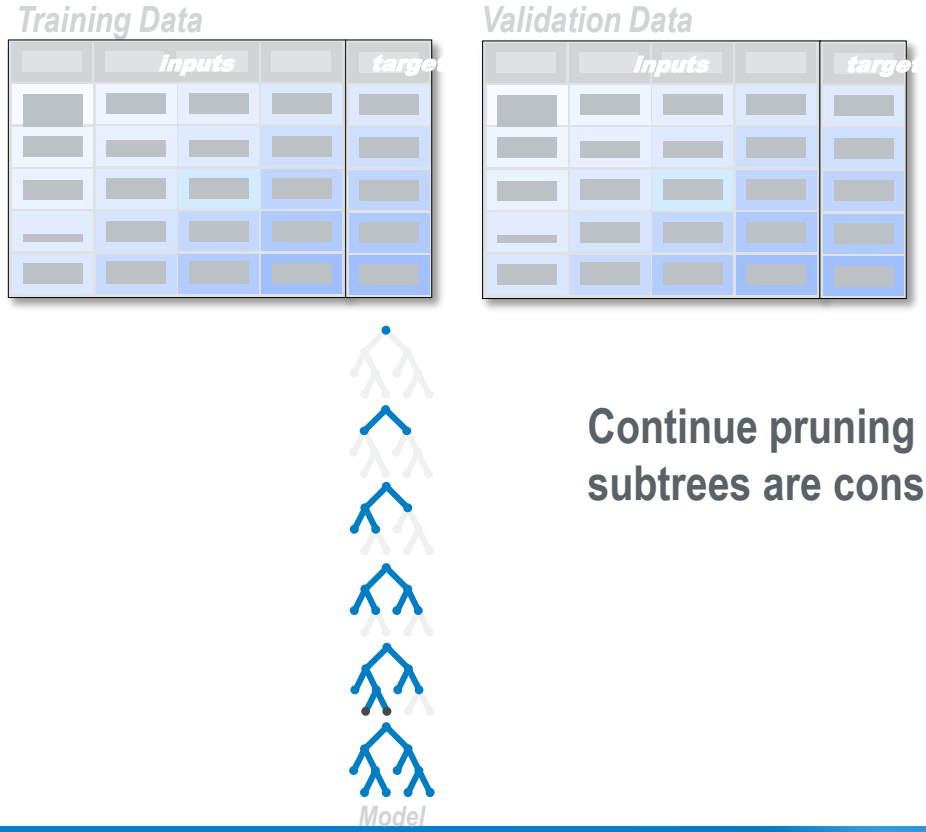
…select the subtree with the best assessment rating.

MODEL TUNING: DECISION TREES



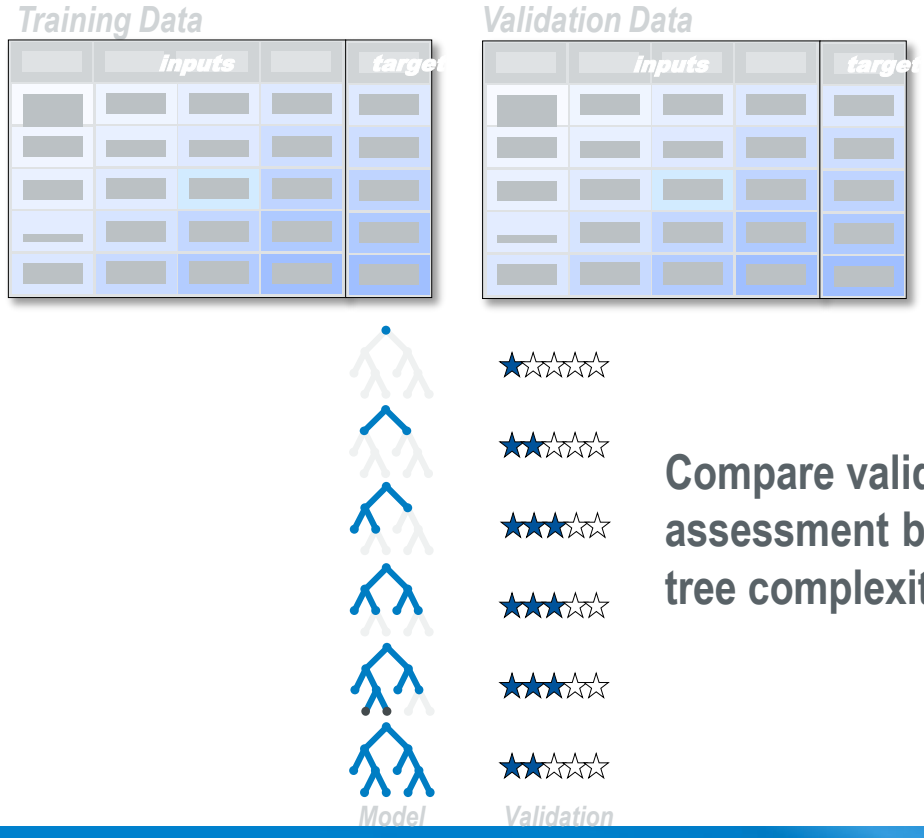
Continue pruning until all subtrees are considered.

MODEL TUNING: DECISION TREES



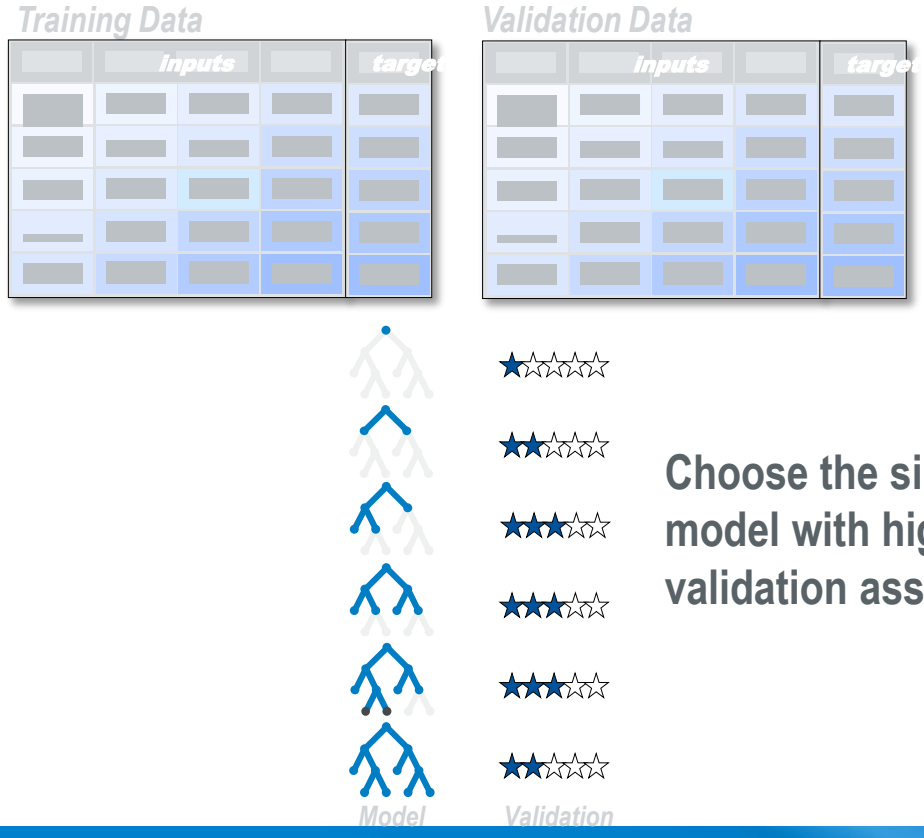
Continue pruning until all subtrees are considered.

MODEL TUNING: DECISION TREES



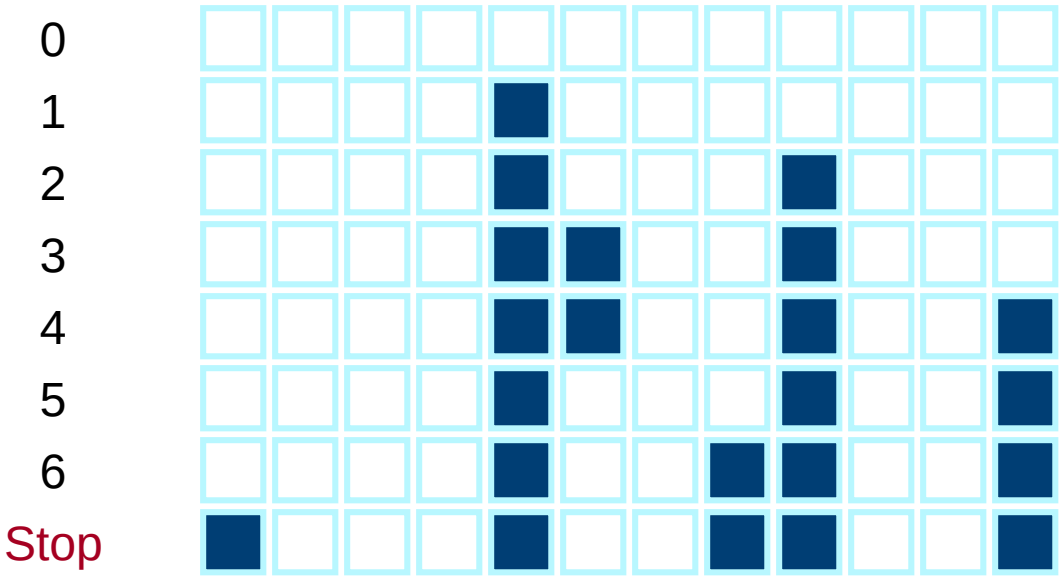
Compare validation assessment between tree complexities.

MODEL TUNING: DECISION TREES

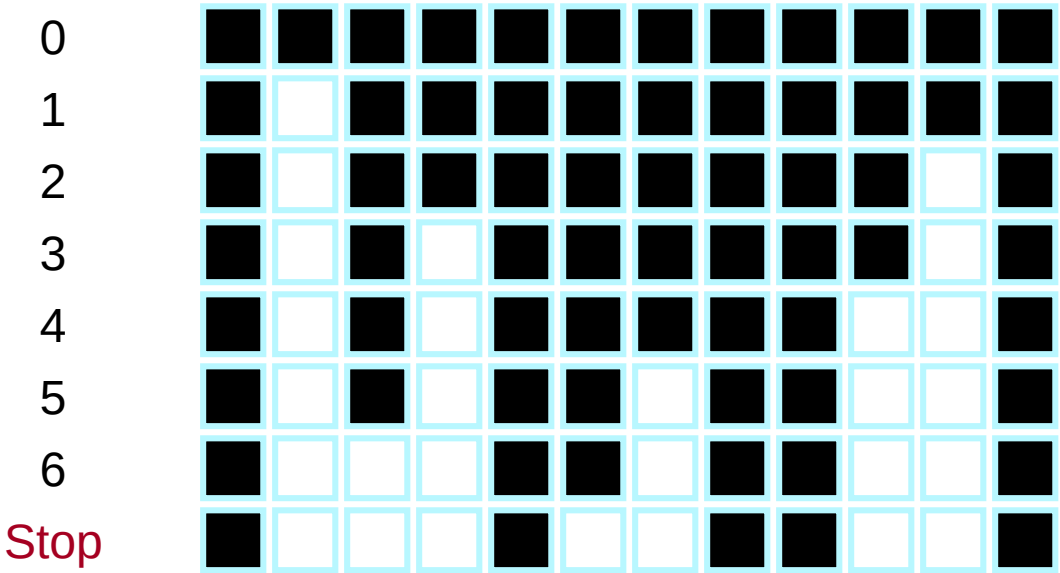


Choose the simplest model with highest validation assessment.

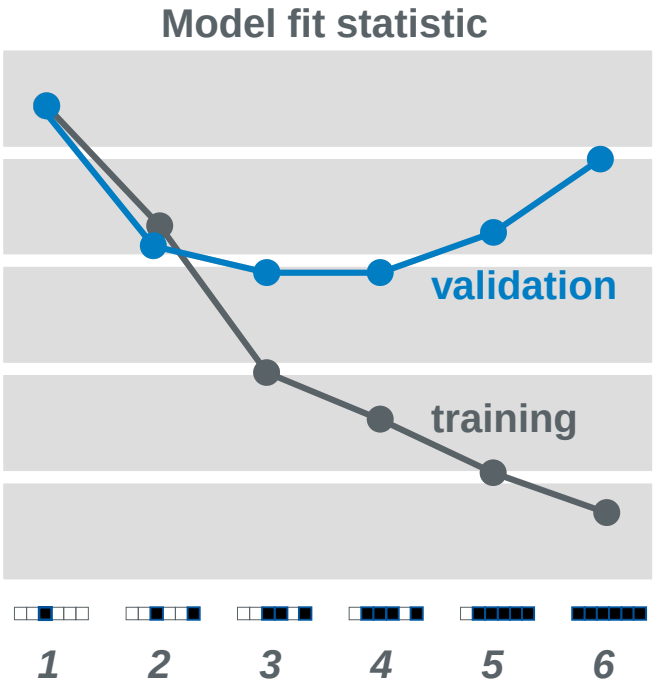
MODEL TUNING: REGRESSION STEPWISE SELECTION



MODEL TUNING: REGRESSION BACKWARD ELIMINATION

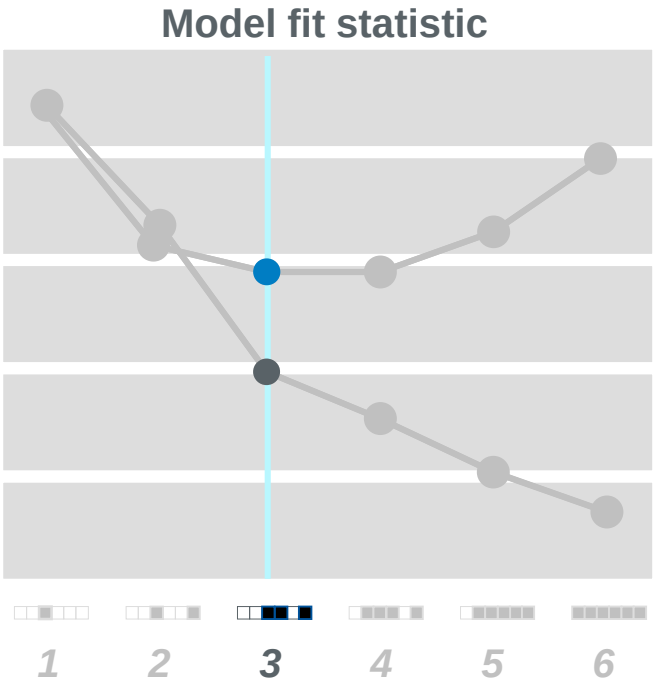


MODEL TUNING: REGRESSION



Evaluate each sequence step.

MODEL TUNING: REGRESSION



Evaluate each sequence step.

Choose simplest optimal model.

MODEL TUNING: NEURAL NETWORKS

$$\text{logit}(\hat{p}) = 0 + 0 \cdot H_1 + 0 \cdot H_2 + 0 \cdot H_3$$

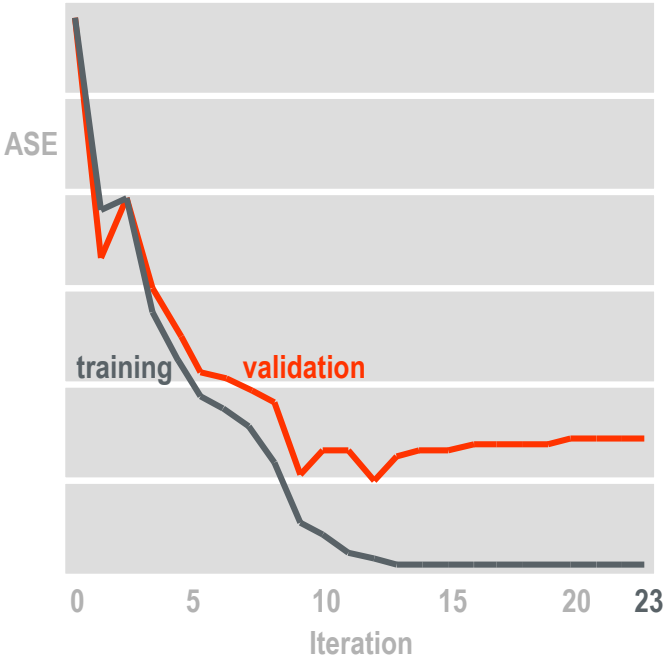
$$H_1 = \tanh(-1.5 - .03x_1 - .07x_2)$$

$$H_2 = \tanh(.79 - .17x_1 - .16x_2)$$

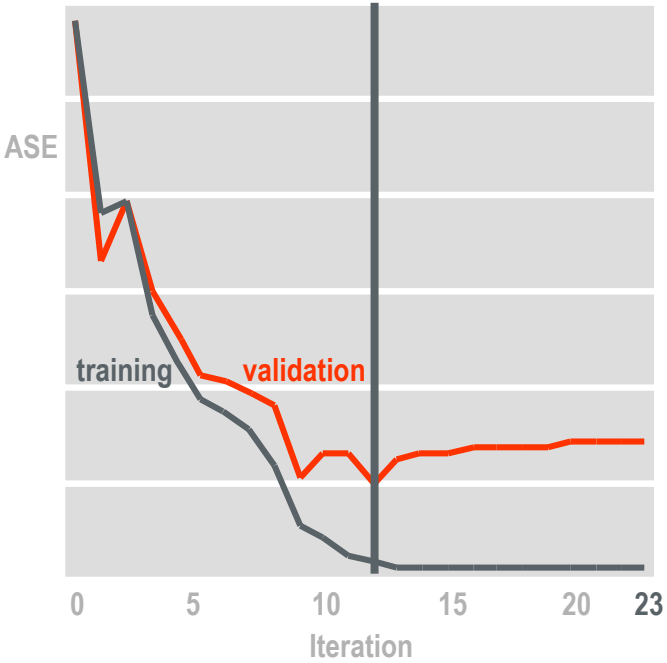
$$H_3 = \tanh(.57 + .05x_1 + .35x_2)$$

random initial
input weights and biases

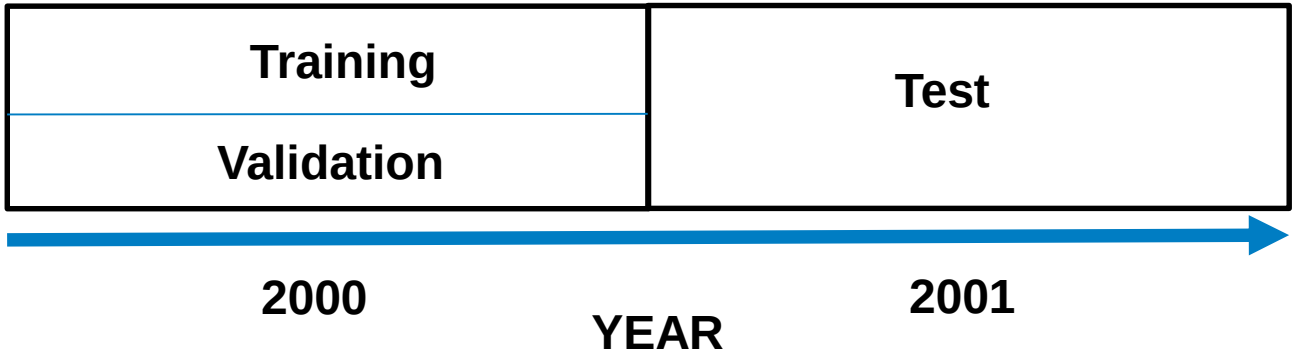
MODEL TUNING: NEURAL NETWORKS



MODEL TUNING: NEURAL NETWORKS



DEMONSTRATION DATA



This data focuses on North Carolina births for 2000 and 2001. The original data sets include more than 120,000 births in each year and contain data on the ethnicity, age, education level, and marital status of the parents; prenatal medical care received; and information about the mother's reproductive history including number of previous pregnancies and live births (State Center for Health Statistics 2001, 2002).

ETERPRISE MINER DEMONSTRATION: VALIDATION AND CROSS VALIDATION TUNING

PROPENSITY TO GIVE BIRTH TO A LOW BIRTH WEIGHT BABY

GENERALIZATION AND THE RIGHT SIZED MODEL: OVERVIEW

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- Ensemble models

PENALIZED MEASURES OF FIT: REGRESSION

- **For least square training: $SBC = N\ln(SSE/N) + P\ln(N)$**

  $AIC = N\ln(SSE/N) + 2P$

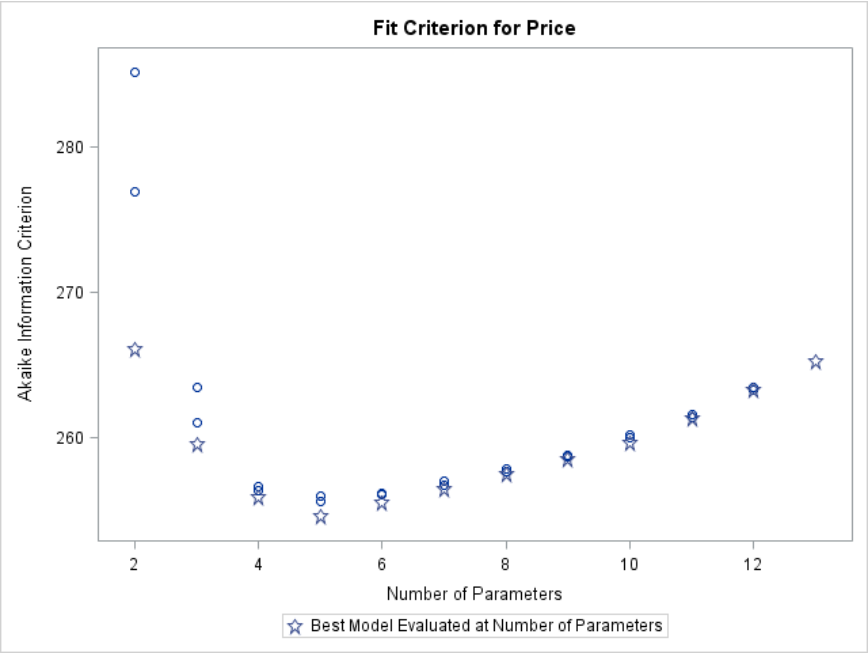
- **For maximum likelihood training: $SBC = -2LL + P\ln(N)$**

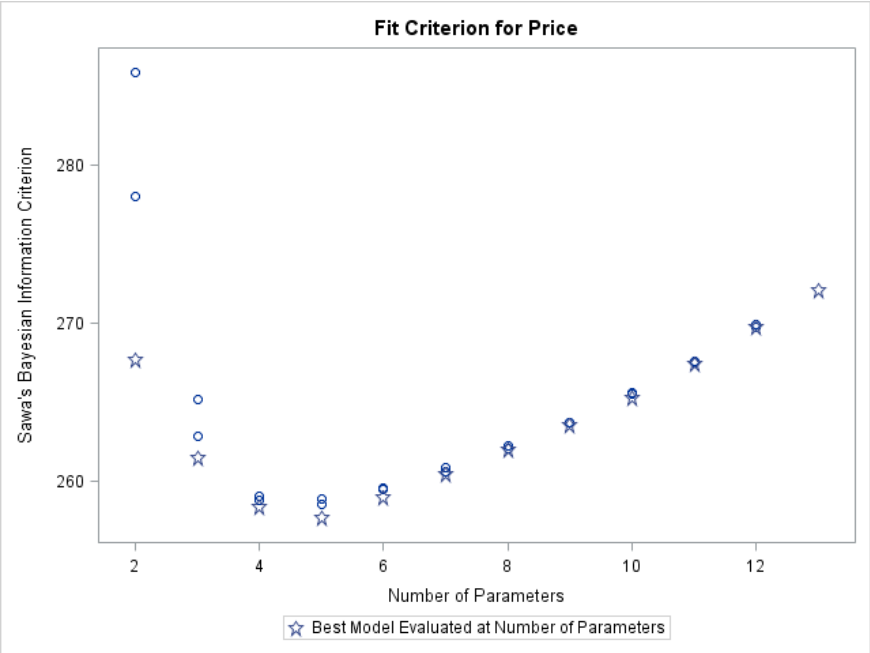
  $AIC = -2LL + 2P$

(N: sample size SSE: error sums of squares P: number of model parameters LL: log likelihood)

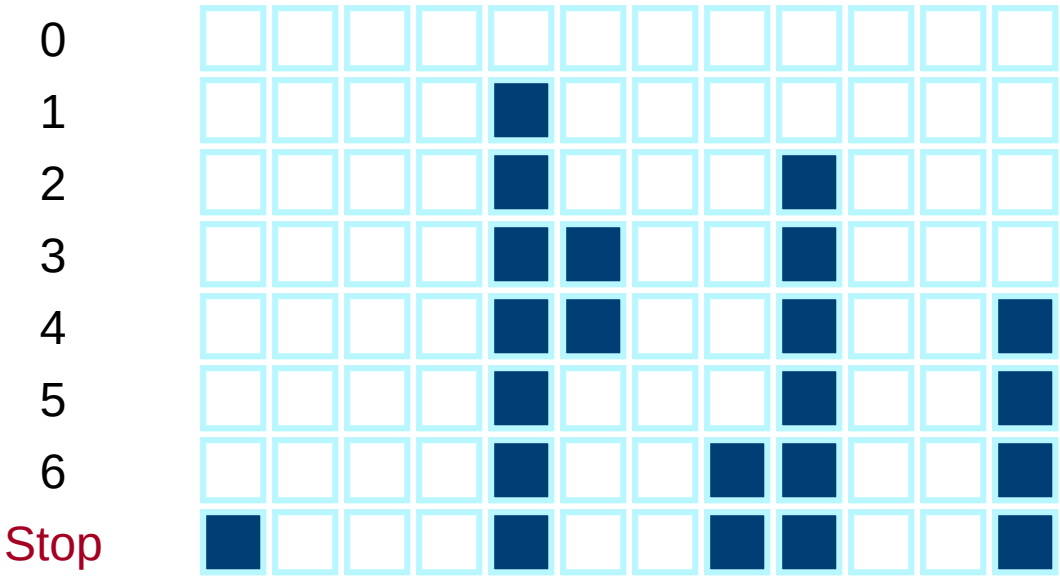
SMALLER IS BETTER

PENALIZED MEASURES OF FIT: REGRESSION





MODEL TUNING: REGRESSION STEPWISE SELECTION



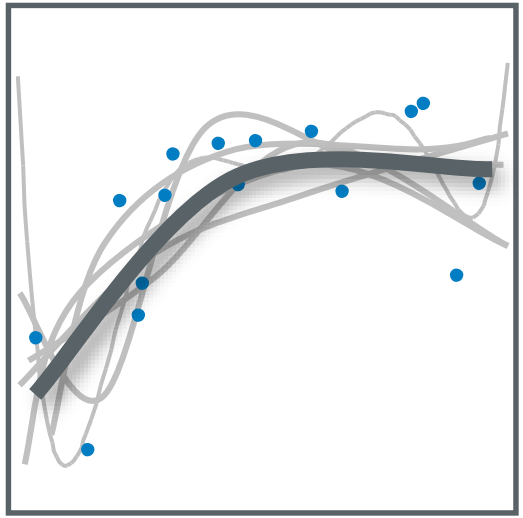
ENTERPRISE MINER DEMONSTRATION: PENALIZED FIT MEASURES

PROPENSITY TO DONATE TO CHARITY AMONGST LAPSED DONORS

GENERALIZATION AND THE RIGHT SIZED MODEL: OVERVIEW

- Predictive modeling and empirical validation
- Model tuning with Validation Data
  - Regression, Decision Trees, Neural Networks
- Penalized measures of fit in Regression
- **Ensemble models**

ENSEMBLE MODELS



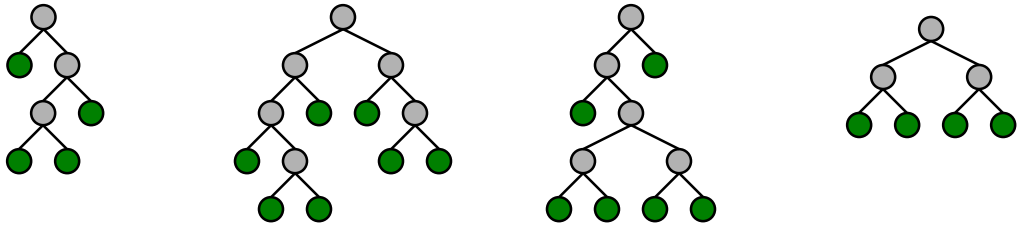
Combine predictions from multiple models to create a single consensus prediction.

ENSEMBLE MODELS

- Combined Models
- Bagging
- Boosting
- Gradient Boosting
- Random Forests

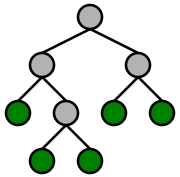
BAGGING

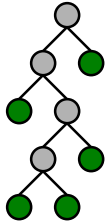
	k=1	k=2	k=3	k=4 …
case	freq	freq	freq	freq
1	1	0	3	1
2	0	1	1	1
3	2	0	0	2
4	0	2	2	0
5	2	2	0	1
6	1	1	0	1

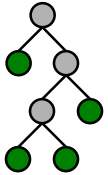


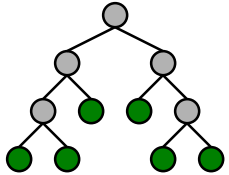
ARC-X4

	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	.5	1	.11
5	1	0	.75	0	.25	0	.06
6	1	0	.75	0	.25	1	.11

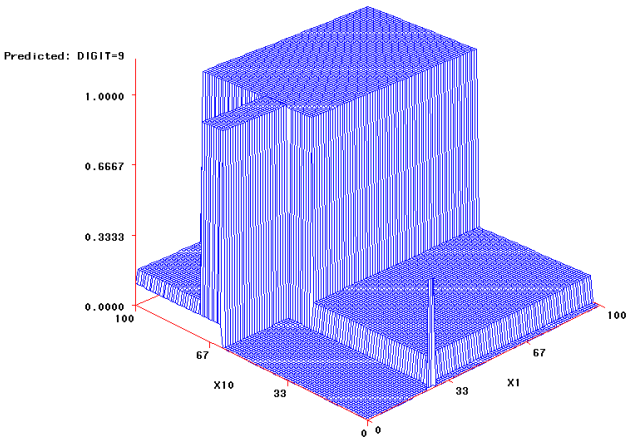


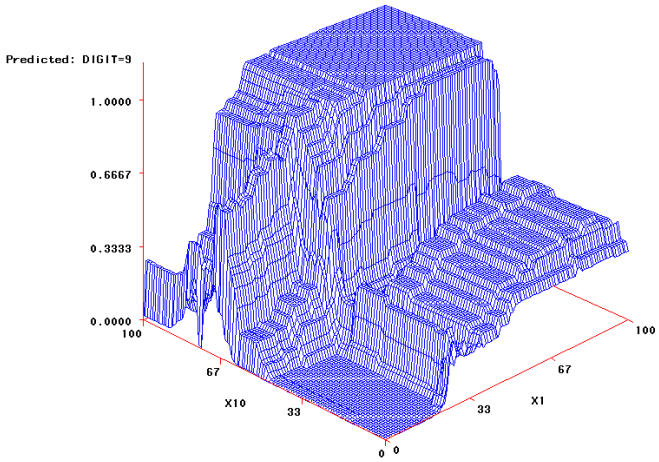


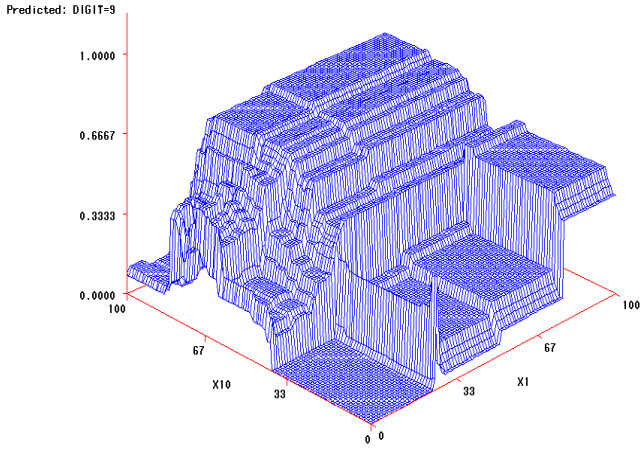




SINGLE, BAGGED, AND BOOSTED TREE







GRADIENT BOOSTING WITH DECISION TREES

Model

$$F_M(x) = F_0 + \nu\beta_1 T_1(x) + \nu\beta_2 T_2(x) + \ldots + \nu\beta_M T_M(x)$$

where *M* is the number of iterations.

Update Formula

For *m* = 1 to *M*, do…

$$F_m(x) = F_{m-1}(x) + \nu\beta_m T_m(x)$$

Examples

***m*=1**

$$F_1(x) = F_0 + \nu\beta_1 T_1(x)$$

***m*=2**

$$F_2(x) = F_0 + \nu\beta_1 T_1(x) + \nu\beta_2 T_2(x)$$

and so on.

ETERPRISE MINER DEMONSTRATION

PROPENSITY TO BUY INSURANCE PRODUCT AMONGST BANK CUSTOMERS

THANK YOU!

Lorne Rothman, PhD, PStat
Principal Statistician, SAS Institute (Canada) Inc.
Lorne.Rothman@sas.com