Overfitting and Cross validation

Cesar Acosta Ph.D.

Department of Industrial and Systems Engineering
University of Southern California

Model performance

How good is the regression model?

Model performance

How good is the regression model?

- *How well the model fits the data?*
- *How well the model predicts the data?*

Model performance

How good is the regression model?

- *How well the model fits the data?*
- *How well the model predicts new data?*

Model performance

How good is the regression model?

- *How well the model fits the data?* SSE R^2
- *How well the model predicts new data?*

Model performance

How good is the regression model?

- *How well the model fits the data?* SSE R^2
- *How well the model predicts new data?* MSPE

Overfitting

Regression assumption:

Expected values of Y follow a regression function

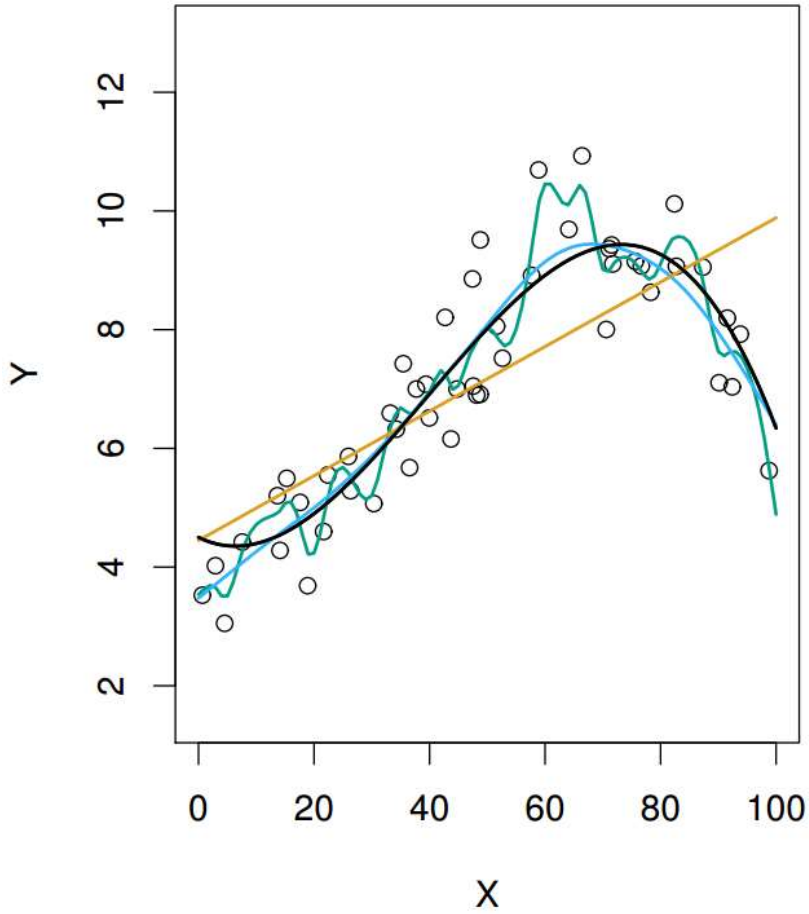
Best model:

Closest model to the regression function

Overfitting:

Model too close to data points

Overfitting - Example



Overfitting

What is overfitting?

Overfitting

What is overfitting?

Building a model
that follows the data too closely
resulting in poor predictions

Overfitting

How to avoid overfitting?

Overfitting

How to avoid overfitting?

Validation Set approach

Cross validation

Validation Set approach

Data set → *training set*

Data set → *test set*

Validation Set approach

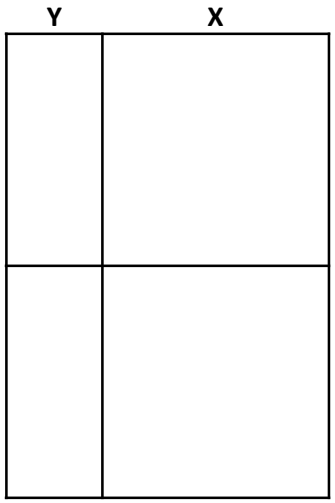
Data set → *training set* (*to find OLS line*)

Data set → *test set* (*to test predictions*)

Validation Set approach

Training set

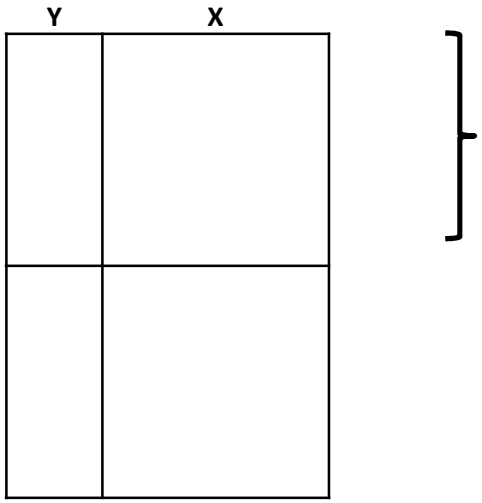
Test set



Validation Set approach

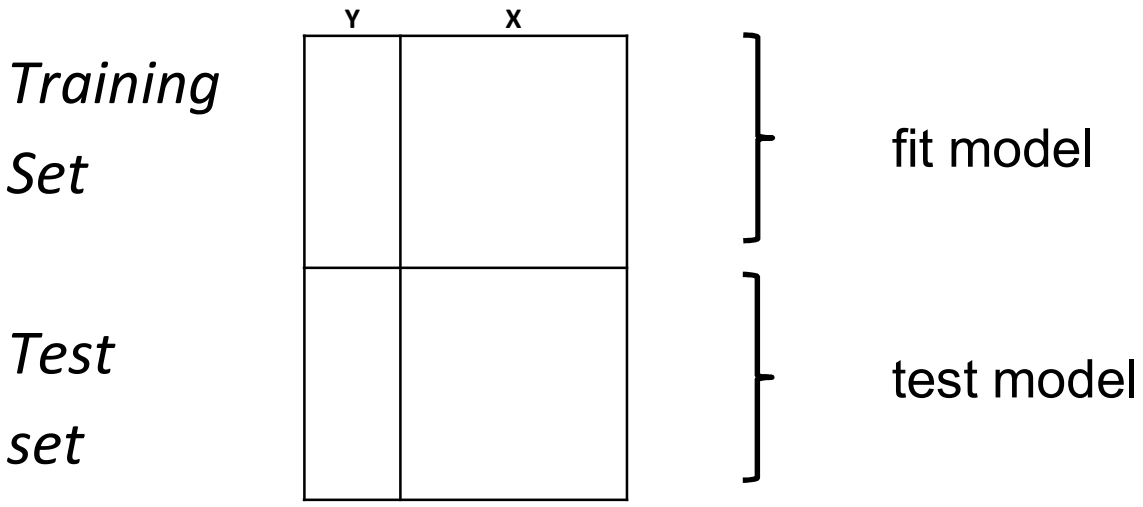
Training Set

Test set

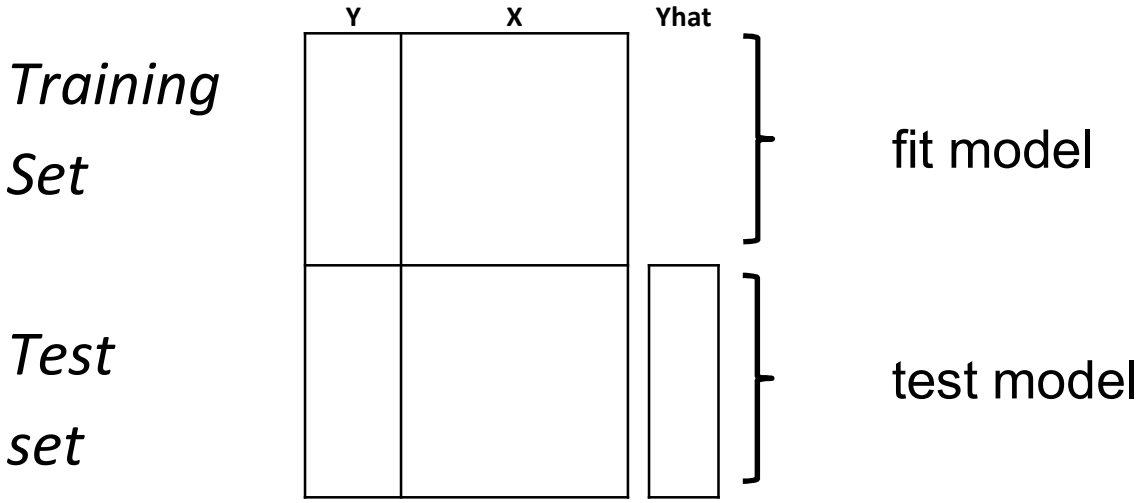


fit model

Validation Set approach



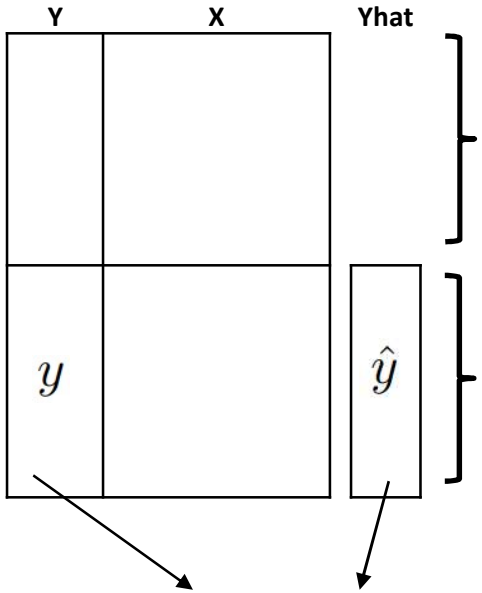
Validation Set approach



Validation Set approach

Training Set

Test set



fit model

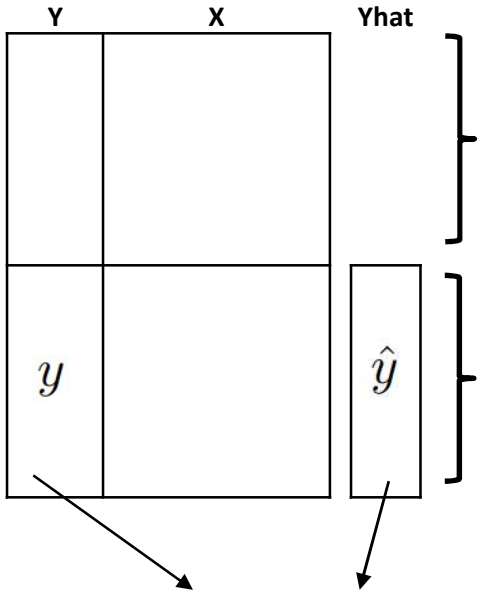
test model

$$MSPE = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}$$

Validation Set approach

Training Set

Test set



fit model

train MSE

test model

$$MSPE = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}$$

test MSE

Prediction performance

The model prediction performance can be estimated by

- Validation Set approach
- Cross Validation
  - LOOCV (Leave-One-Out Cross-validation)
  - k-Fold Cross Validation

Prediction performance

Compare models based on MSPE

Model with the smallest MSPE

is best in terms of prediction performance

k-Fold Cross Validation

dataset

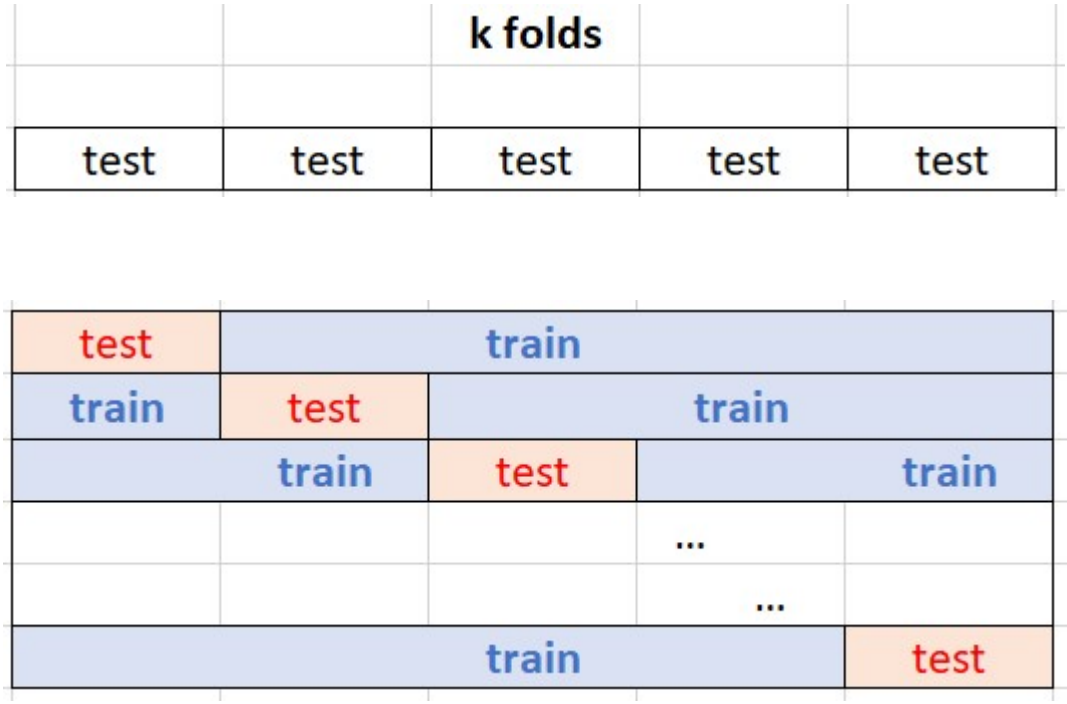
k-Fold Cross Validation

dataset

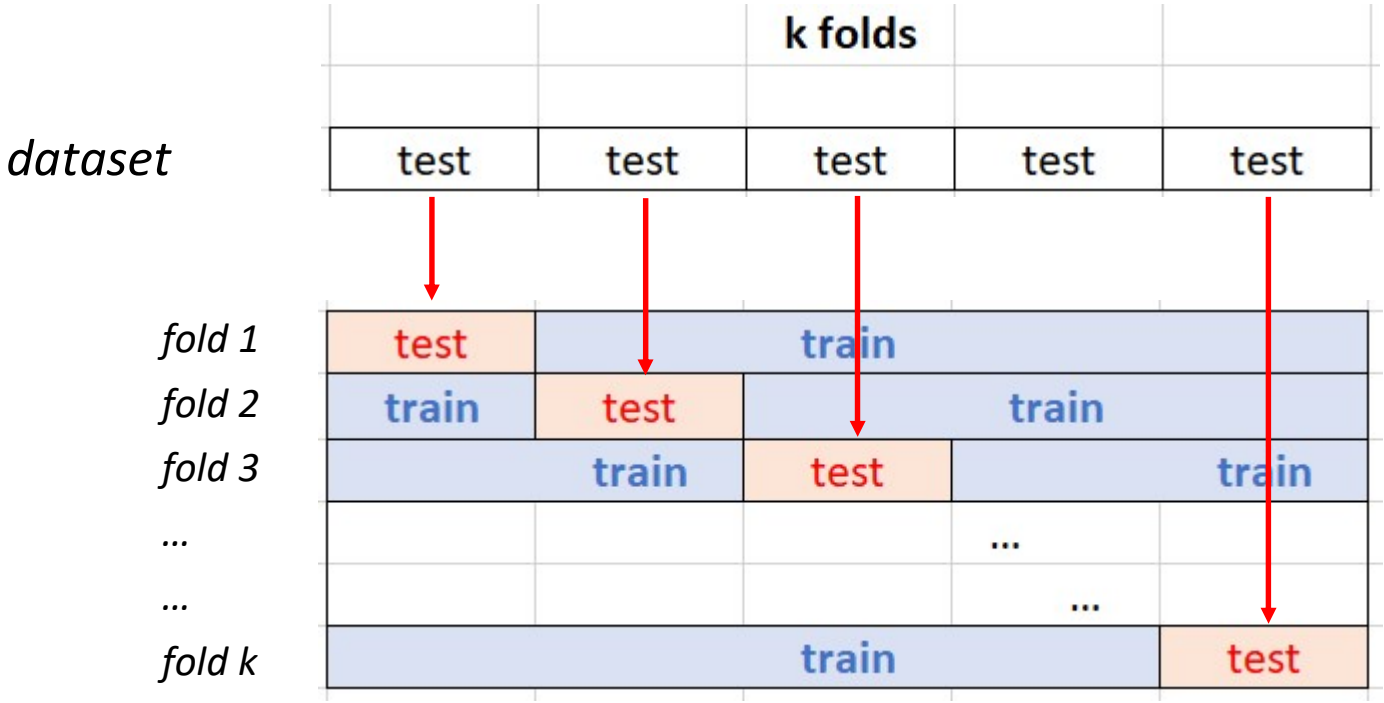
		k folds		
test	test	test	test	test

k-Fold Cross Validation

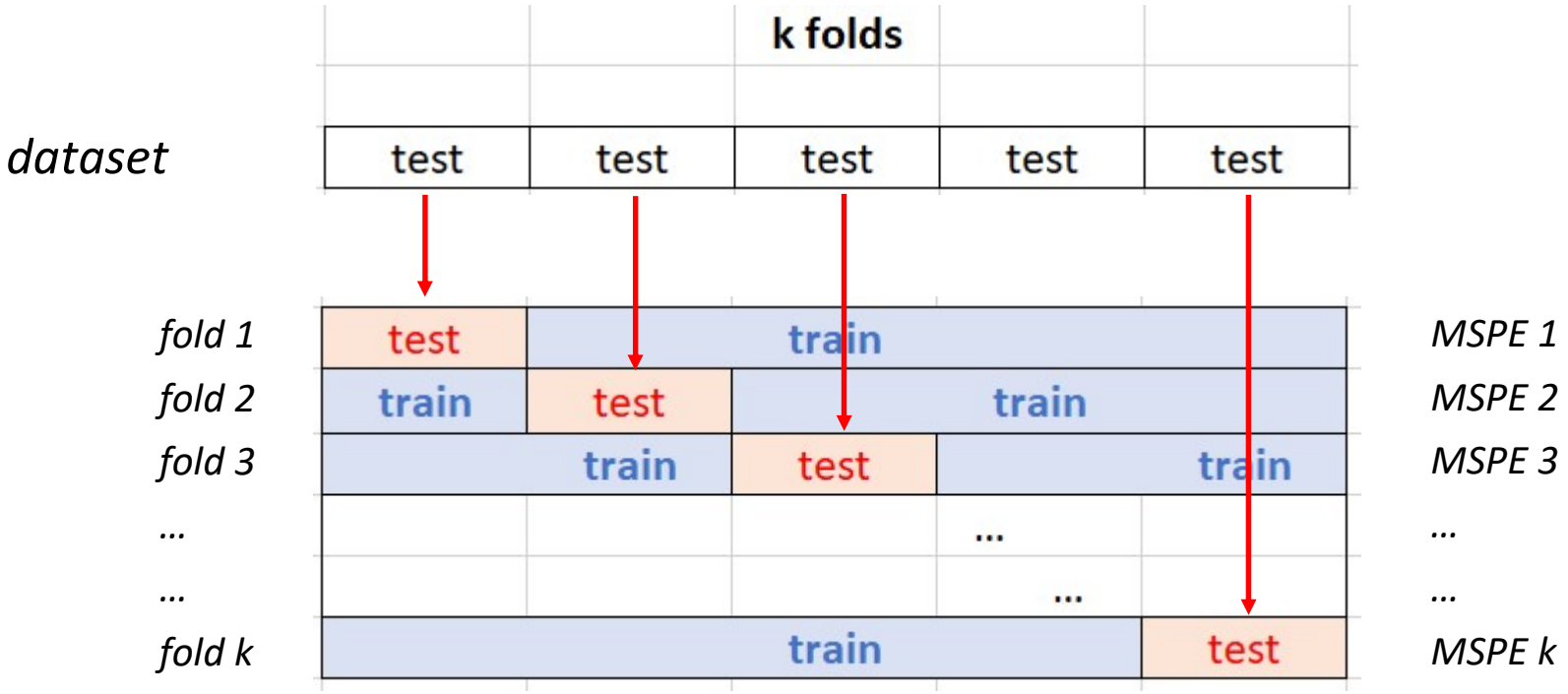
dataset



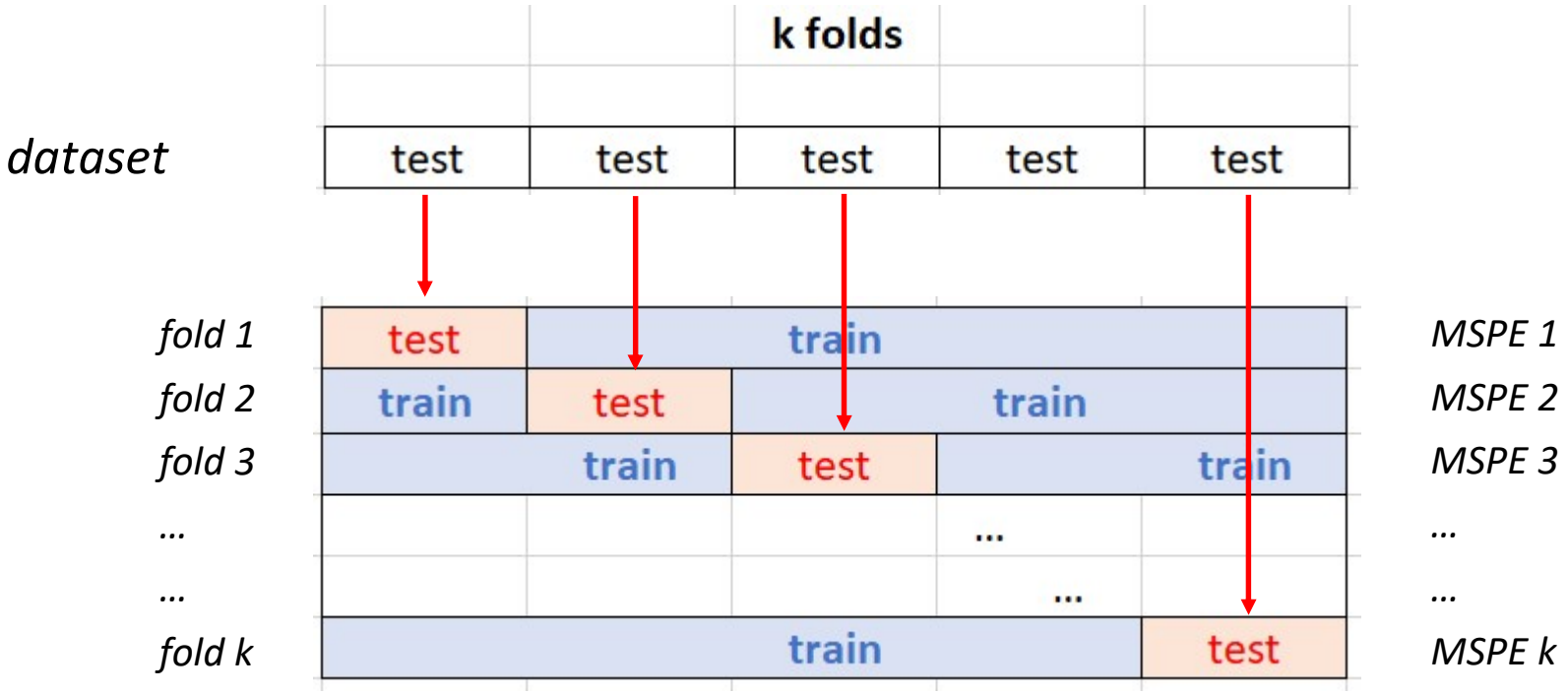
k-Fold Cross Validation



k-Fold Cross Validation



k-Fold Cross Validation



MSPE (average)

k-Fold Cross Validation

Data set

- *training set* $n\left(1-\frac{1}{k}\right)$ *obs*
- *test set* $n\left(\frac{1}{k}\right)$ *obs*

k-Fold Cross Validation

k=5 folds

Data set

- *training set* $n\left(1-\frac{1}{k}\right)$ *obs*
- *test set* $n\left(\frac{1}{k}\right)$ *obs*

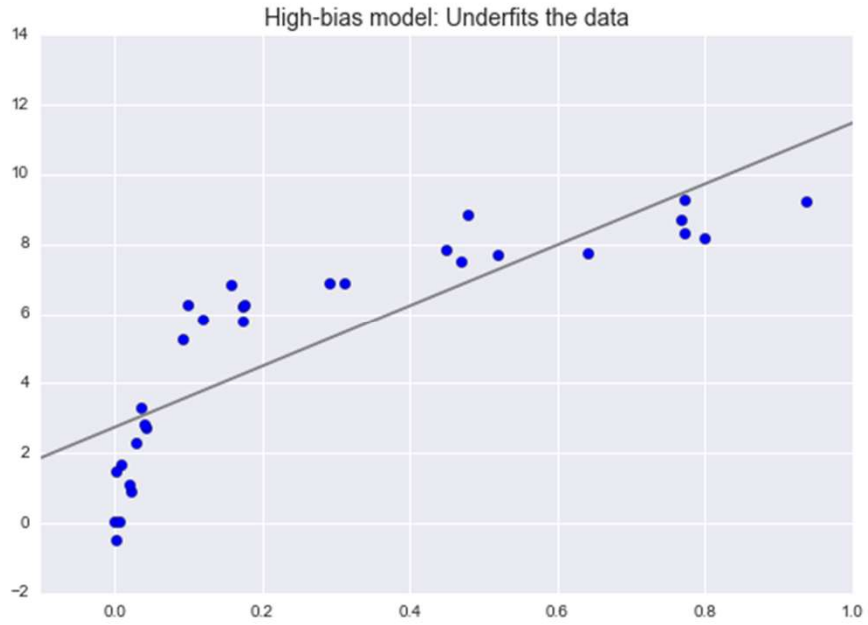
k-Fold Cross Validation

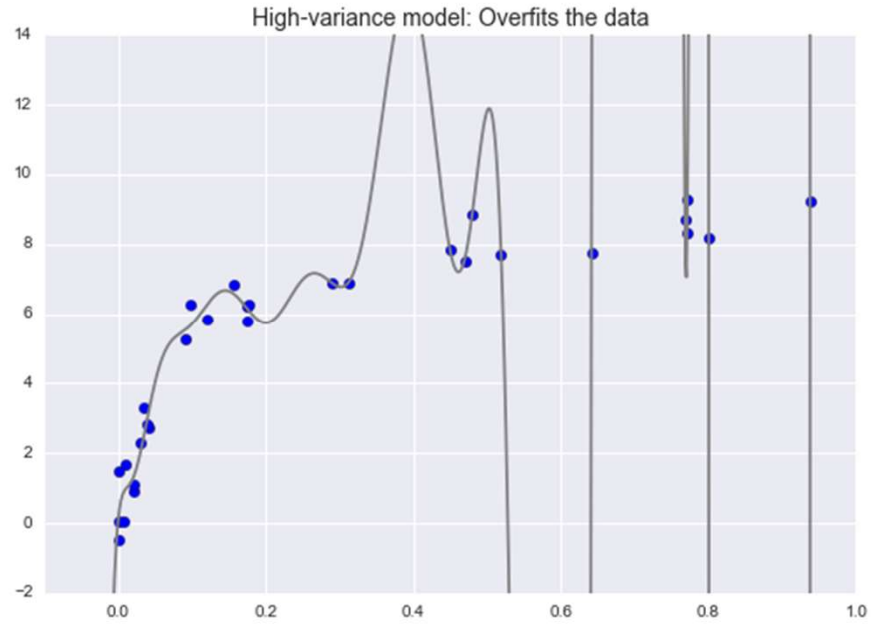
k=5 folds

Data set	*training set*	$n\left(1-\frac{1}{k}\right)$ *obs*	*80%*
	test set	$n\left(\frac{1}{k}\right)$ *obs*	*20%*

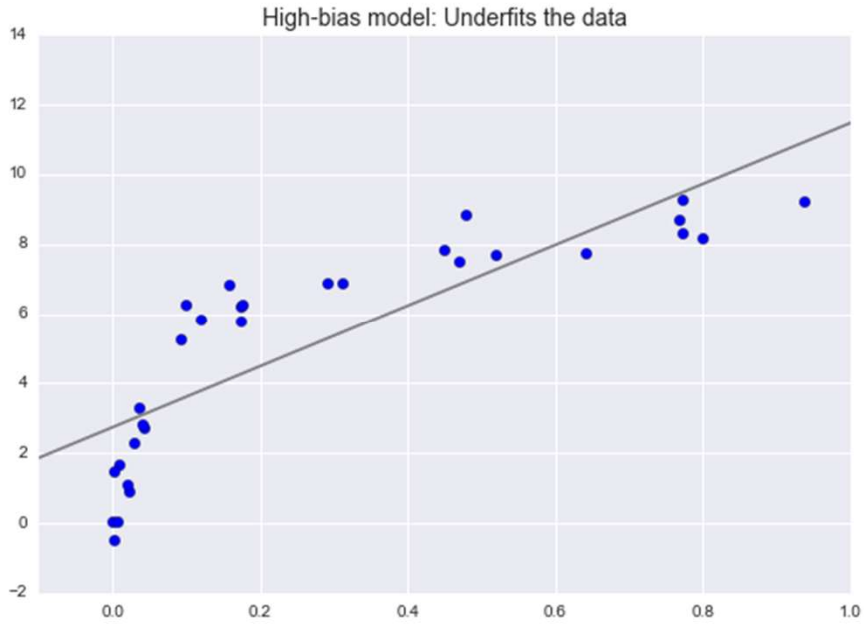
Bias – Variance trade-off

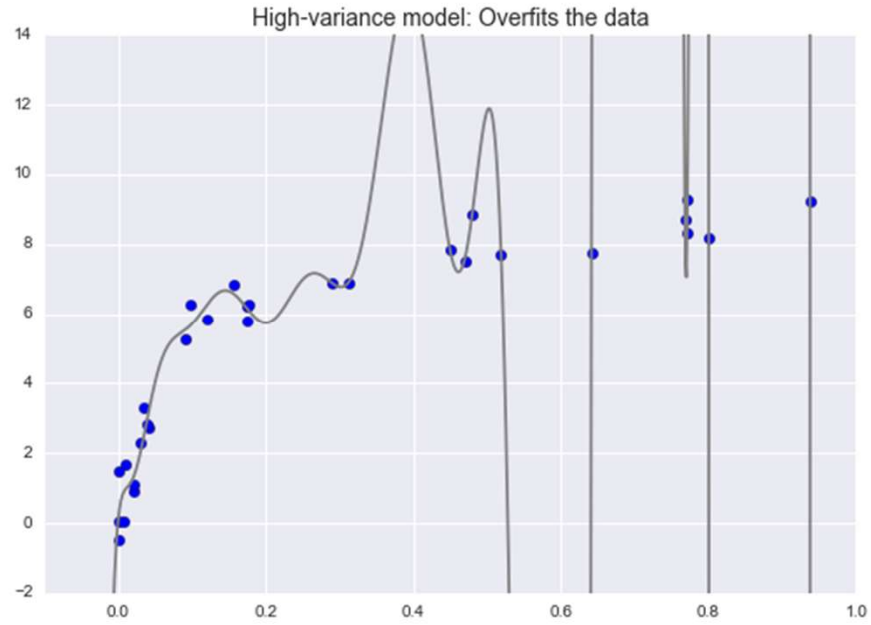
Bias – Variance trade-off





Bias – Variance trade-off





We want a balance between model bias and variability

Bias – Variance trade-off

We want a balance between model bias and variability