FLATIRON SCHOOL

The Bias-Variance Tradeoff + Train_Test_Split()

8/27/2019

Today's Lesson

Learning Objectives

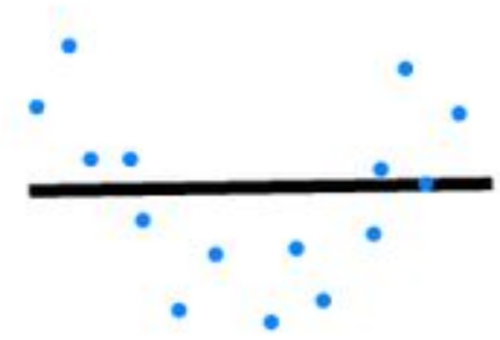
- Explain what bias, variance, and error are in the context of statistical modelling
- Explain how bias, variance and error are related via the bias-variance tradeoff
- Explain how a holdout set can be used to evaluate a model
- Use a test set to estimate model bias, variance and error

Activities

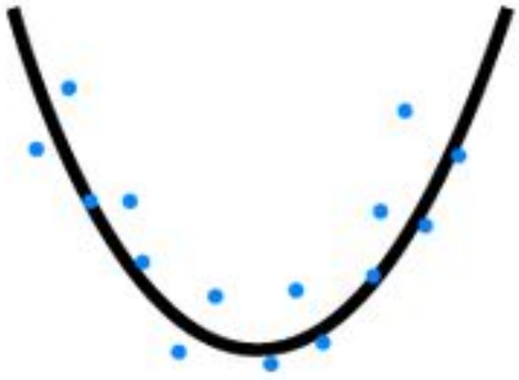
- Bias-Variance Partner Activity
- Bias-Variance Tradeoff Lecture
- Bias-Variance Partner Activity #2
- Train_Test_Split() Practice

Which model is best? Why?

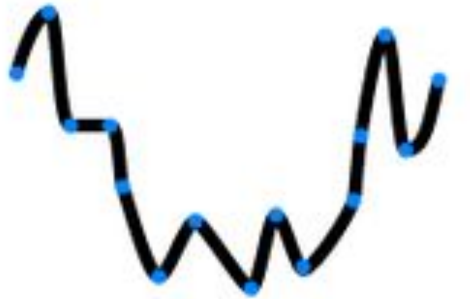
We have the commodity price of gold at the end of each market day for 3 weeks. Which model best describes the price of gold over time? What makes it "best"?



Linear Model



Quadratic Model



High-Order Polynomial Model

https://towardsdatascience.com/cultural-overfitting-and-underfitting-or-why-the-netflix-culture-wont-work-in-your-company-af2a62e41288

What makes a model good?

We don’t ultimately care about how well your model fits your data.

What we really care about is how well your model *describes the process that generated your data.*

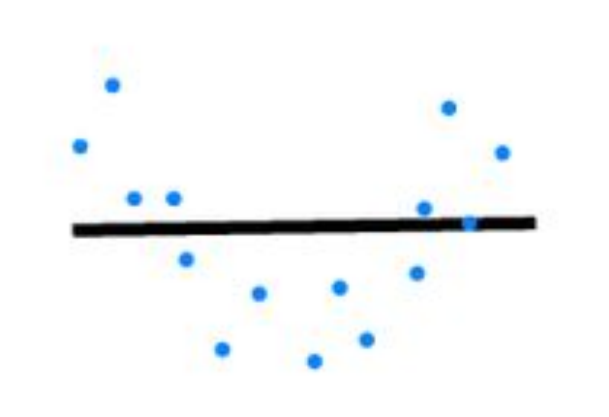
Why? Because the data set you have is but one sample from a universe of possible data sets, and you want a model that would work for *any* data set from that universe

What is a “Model”?

A “model” is a general specification of relationships among variables and parameters.

- E.G. Linear regression, or $PRICE = \beta_1 * TIME + \beta_0 + \varepsilon$

A “trained model” is a particular model with parameters estimated using some training data.

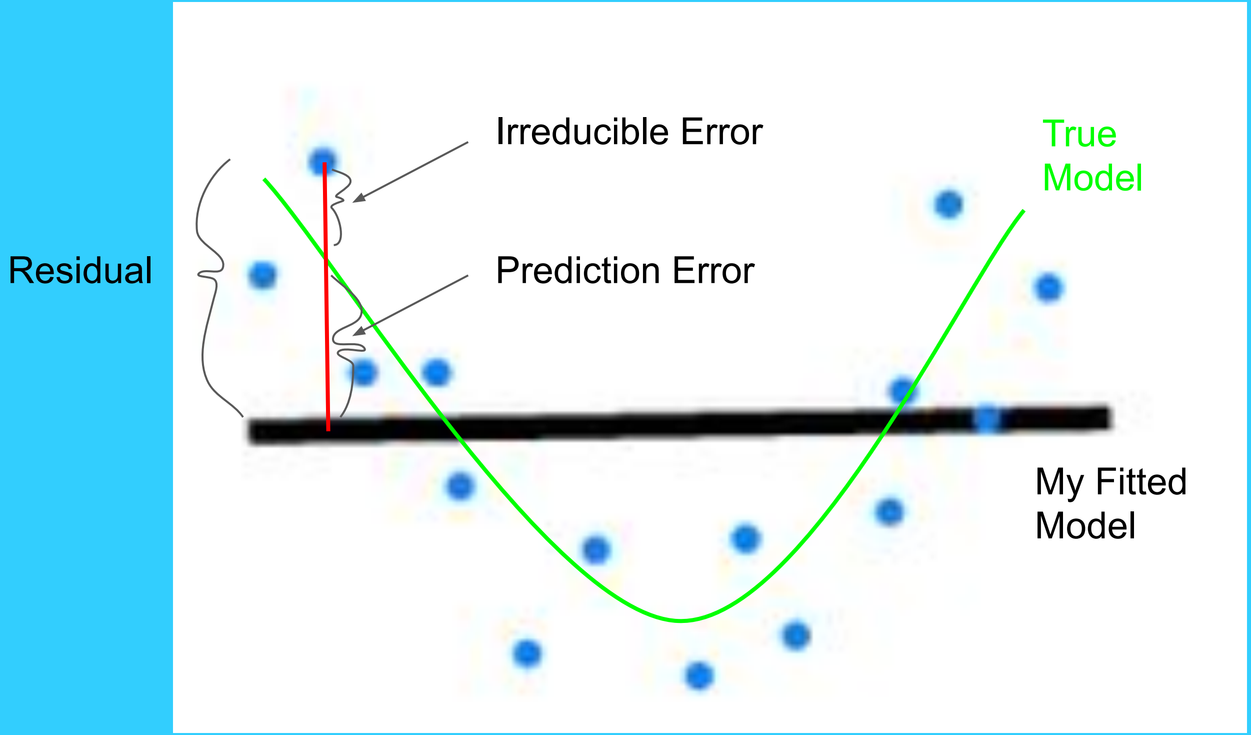


Expected Value

The expected value of a quantity is the weighted average of that quantity across all possible samples

What is the expected value of a roll on a six-sided die?

Defining "Error"



Irreducible Error

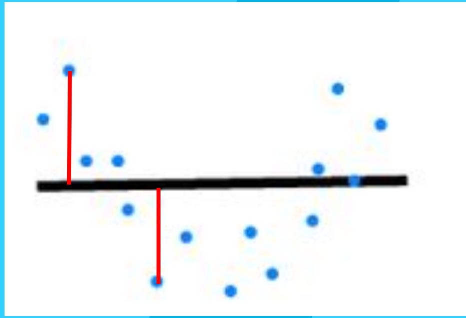
True Model

Residual

Prediction Error

My Fitted Model

Defining “Error”



For regression, “error” usually refers to prediction error or to residuals

- Errors are approximated by residuals

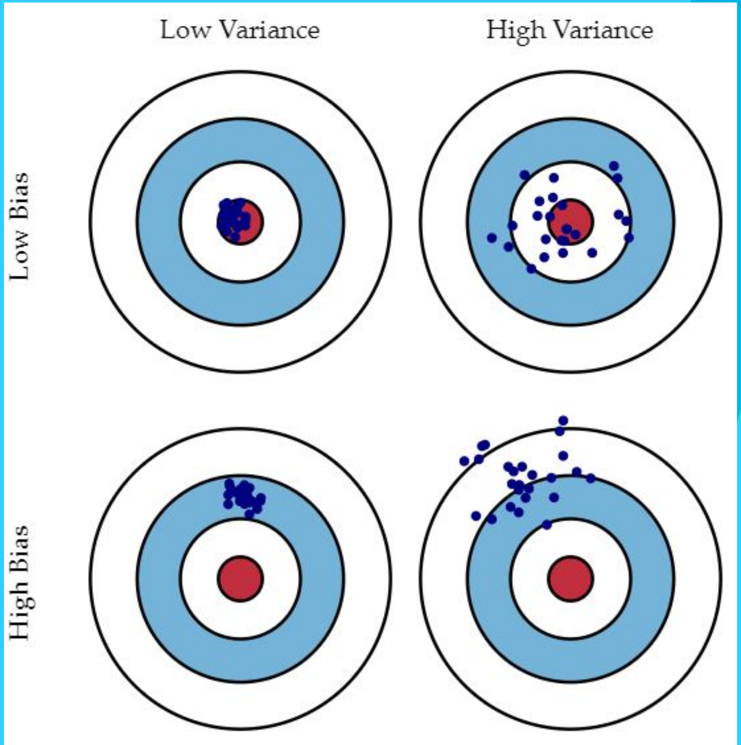
Regression fit statistics are often called “error”

- Sum of Squared Errors (SSE)
- Mean Squared Error (MSE)
- Calculated using residuals

Defining Model Bias and Variance

“Model Bias” is the expected prediction error from your expected trained model

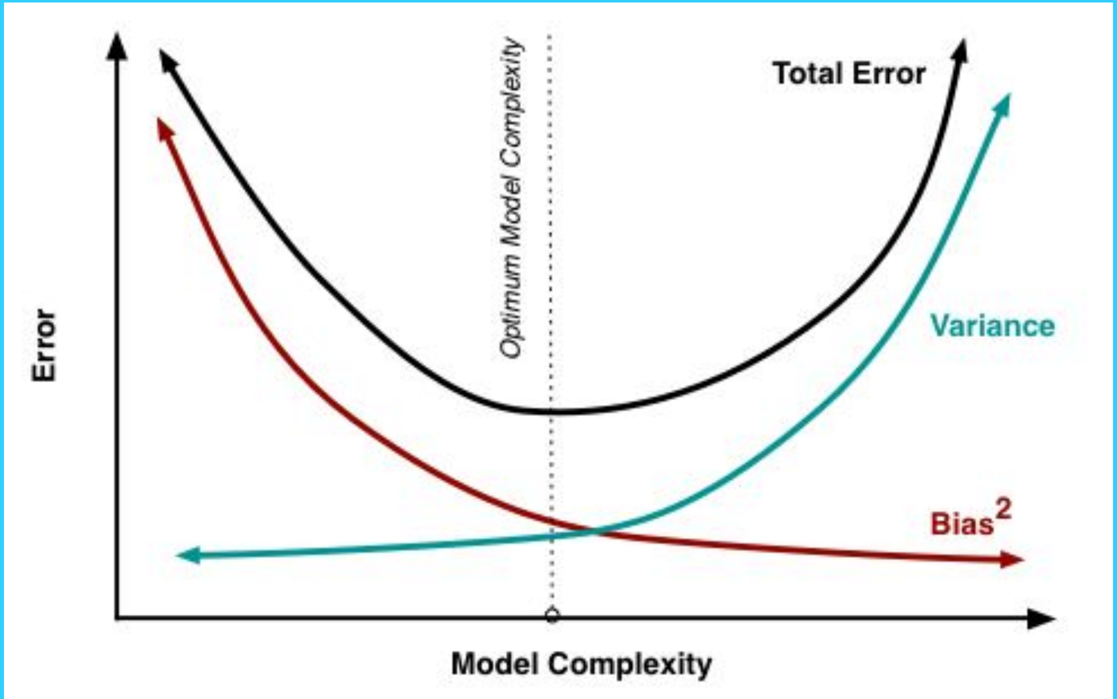
“Model Variance” is the expected variation in predictions, relative to your expected trained model



http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf
http://scott.fortmann-roe.com/docs/BiasVariance.html

The Bias-Variance Tradeoff

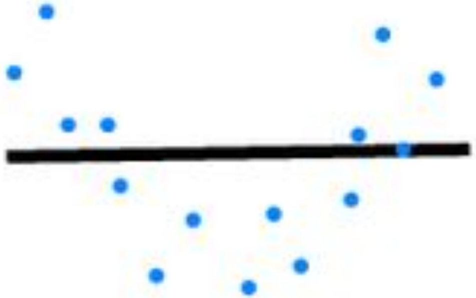
Total Error = Model Bias^2 + Model Variance + Irreducible Error



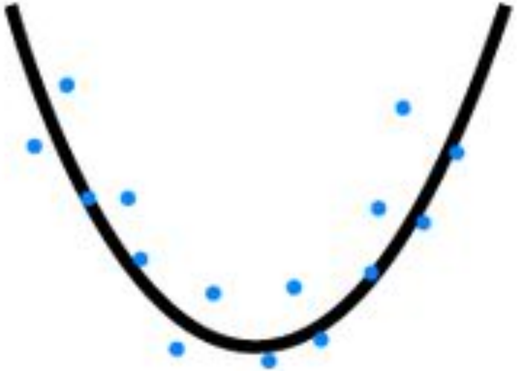
http://scott.fortmann-roe.com/docs/BiasVariance.html

Which model is best? Why?

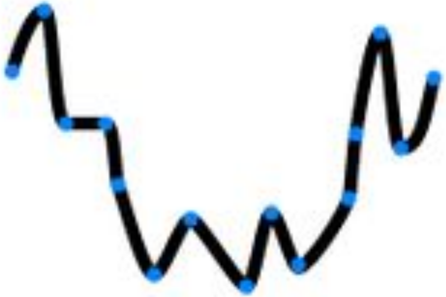
Write an explanation for why the quadratic model is “best”, using the terms bias, variance, and error. Share it with a neighbor, then synthesize your explanations.



Linear Model



Quadratic Model



High-Order Polynomial Model

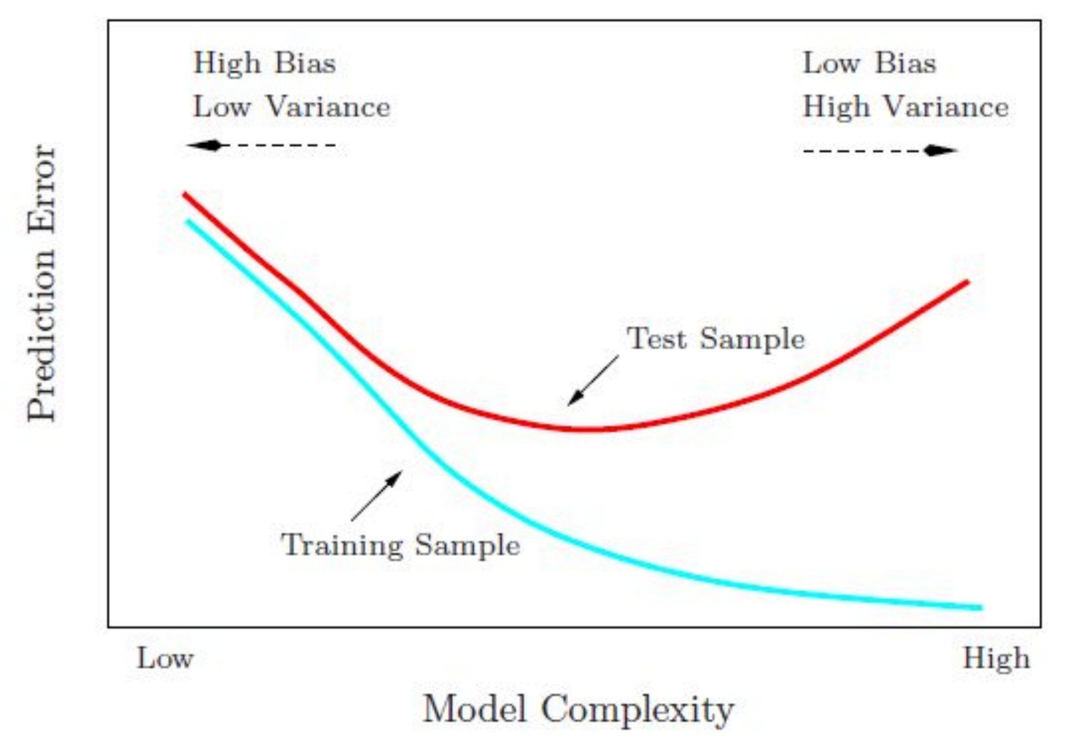
https://towardsdatascience.com/cultural-overfitting-and-underfitting-or-why-the-netflix-culture-wont-work-in-your-company-af2a62e41288

Using Test Samples

It is hard to know if your model is too simple or complex by just using it on training data.

We can save part of our training sample as a test sample and use it to monitor our prediction error.

This allows us to evaluate whether our model has the right balance of bias/variance.



https://stats.stackexchange.com/questions/152882/question-about-bias-variance-tradeoff

Train_Test_Split() Practice

Presented by David Braslow

FLATIRON SCHOOL