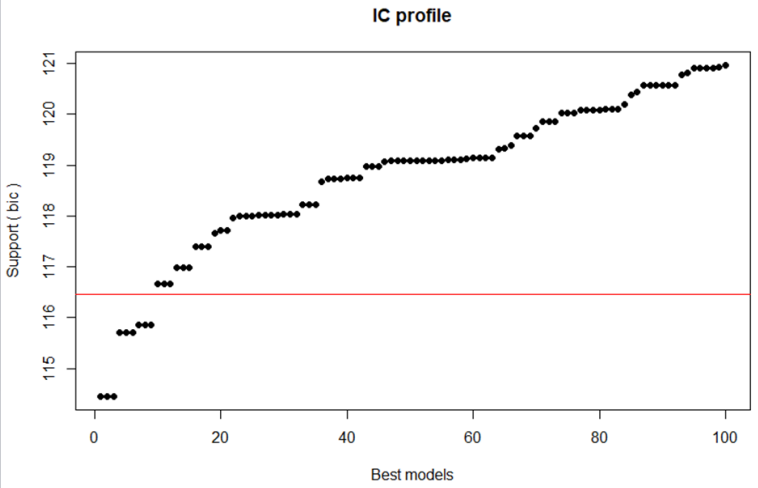
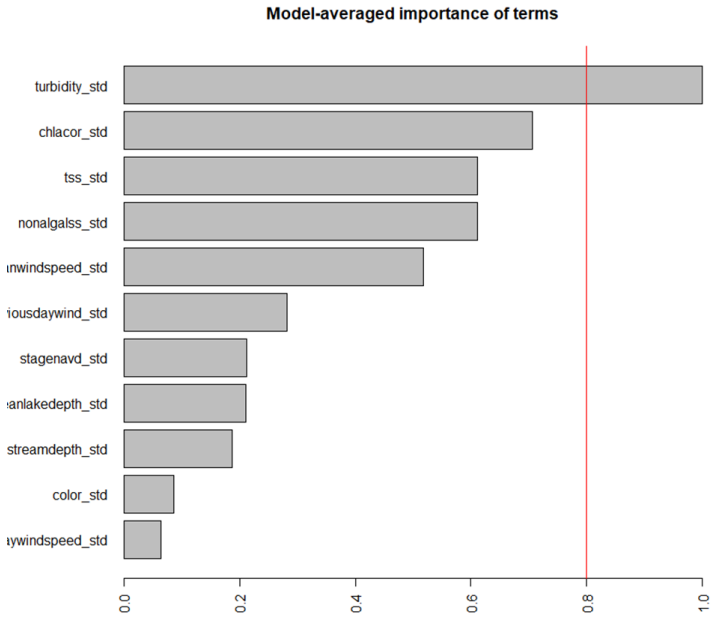
Advanced R: Statistical Machine Learning

Dan Schmutz, MS
Chief Environmental Scientist

Zoom Workshop for SJRWMD
September 24, 2020





Fundamental Challenges for Appropriate Application of Statistical Machine Learning

GPI

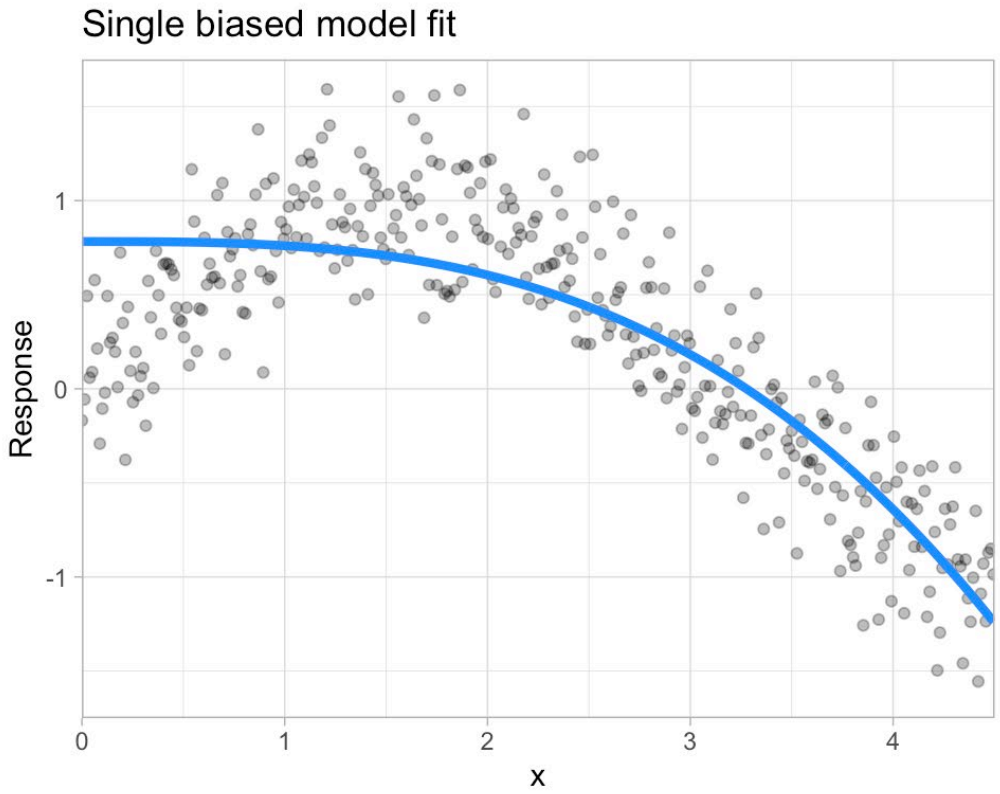
Bias-Variance Tradeoff: The Problem of Overfitting

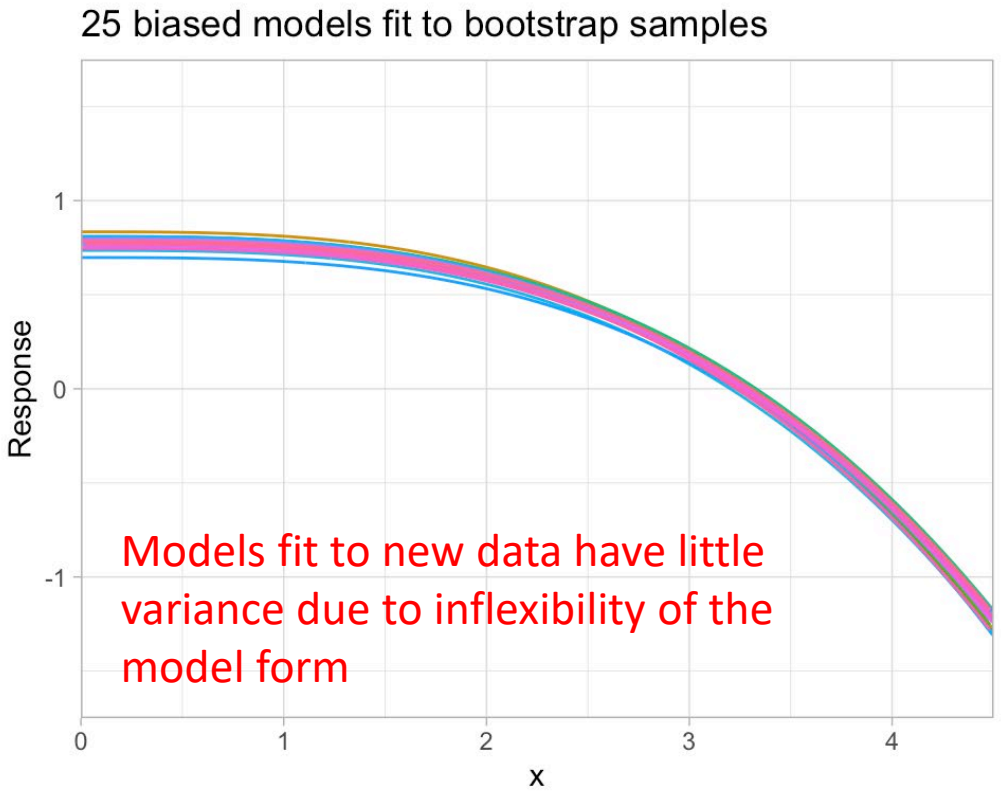
- Prediction errors can be decomposed into two important subcomponents: error due to “bias” and error due to “variance”.
- There is often a tradeoff between a model’s ability to minimize bias and variance.
- Understanding this tradeoff results in fitting more accurate models.

GPI

Model with High Bias

- Bias is the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.



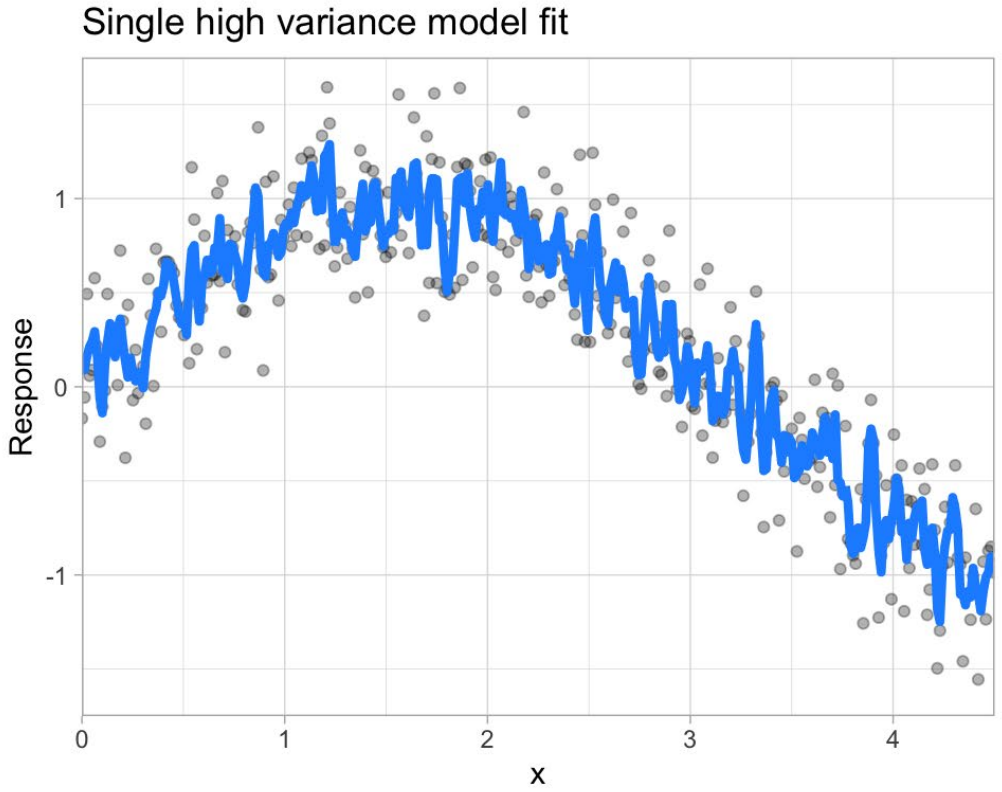


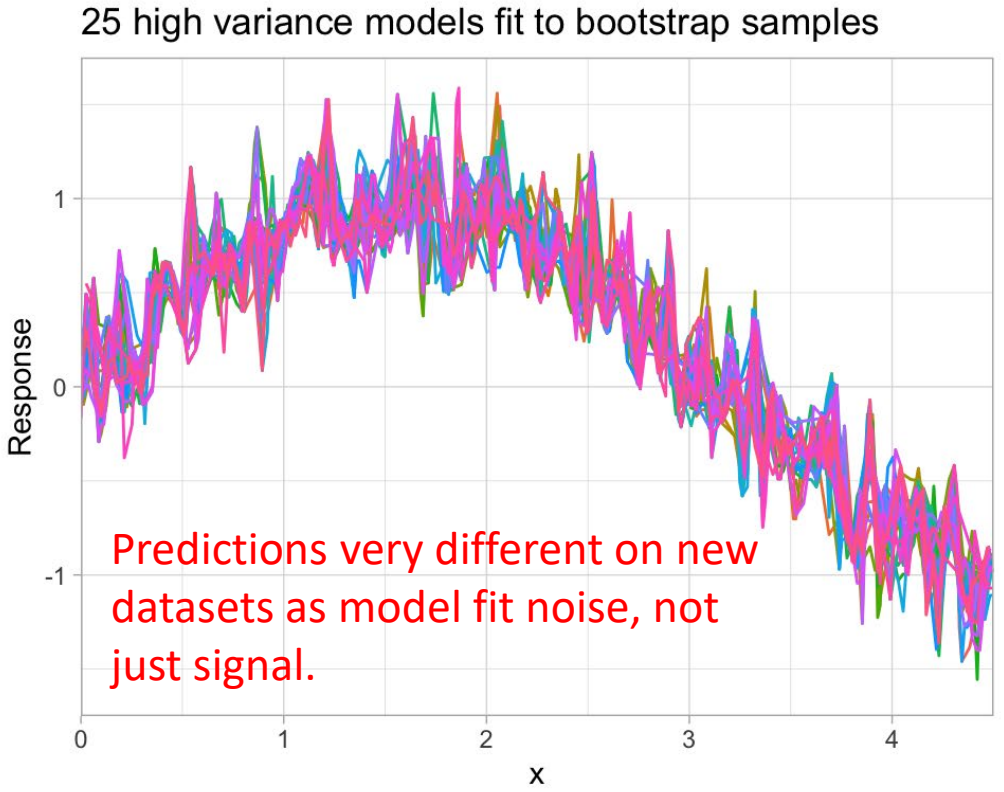
https://bradleyboehmke.github.io/HOML/process.html

GPI

Model with High Variance

- Error due to variance is defined as the variability of a model prediction for a given data point. High variance models are very flexible but prone to overfitting to a specific dataset.

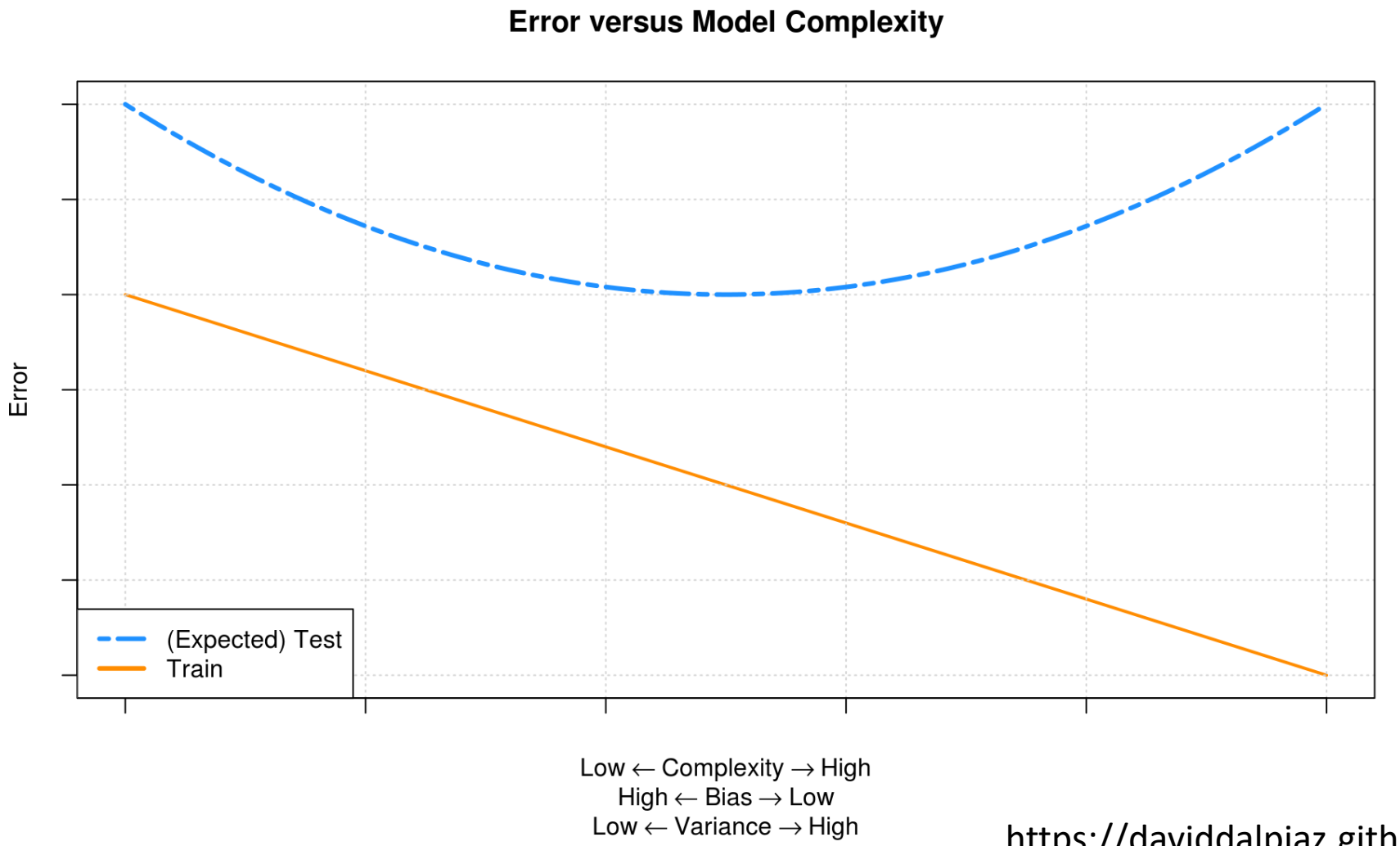




https://bradleyboehmke.github.io/HOML/process.html

GPI

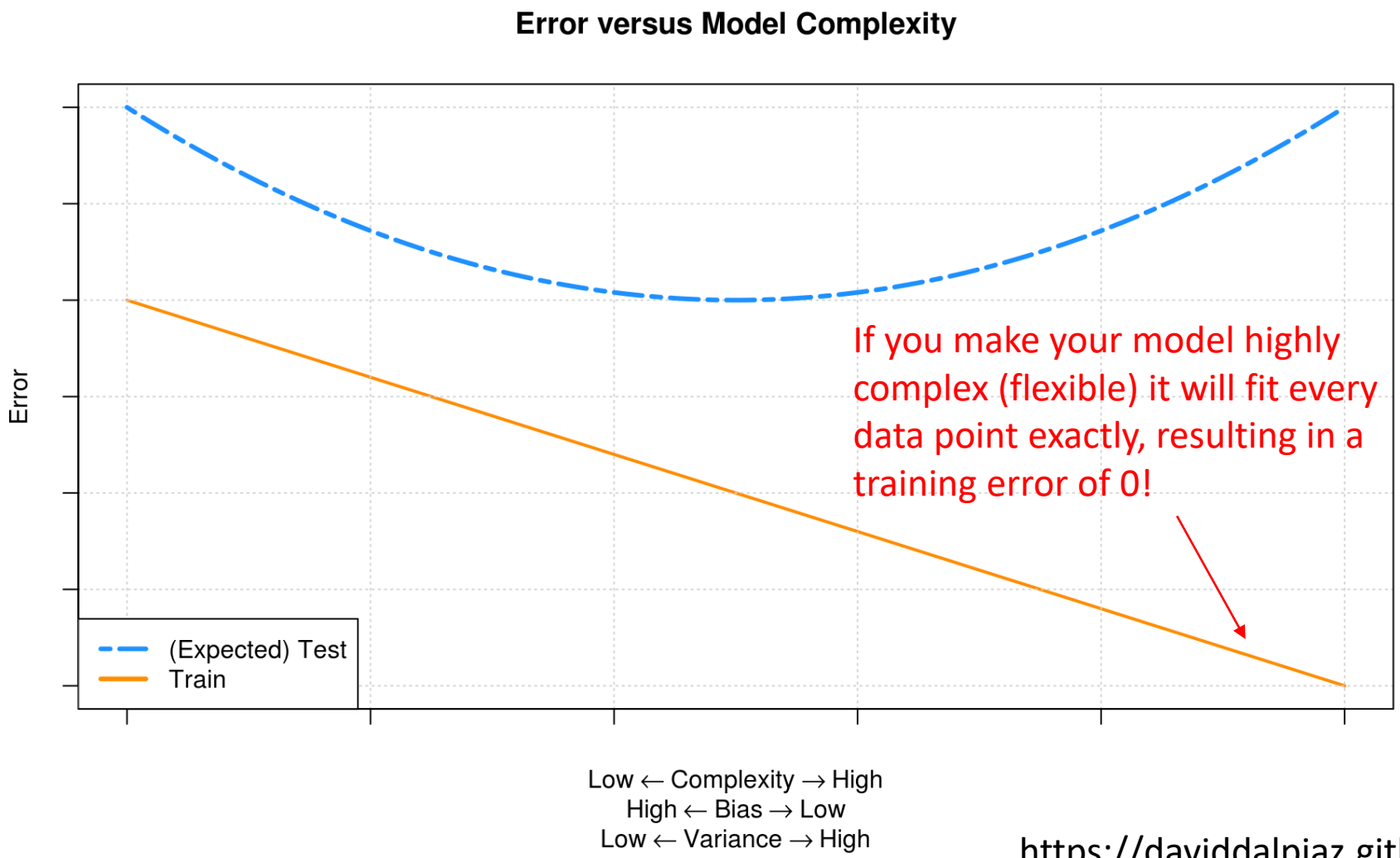
Finding the Middle Path: Bias Variance Tradeoff



https://daviddalpiaz.github.io/r4sl/

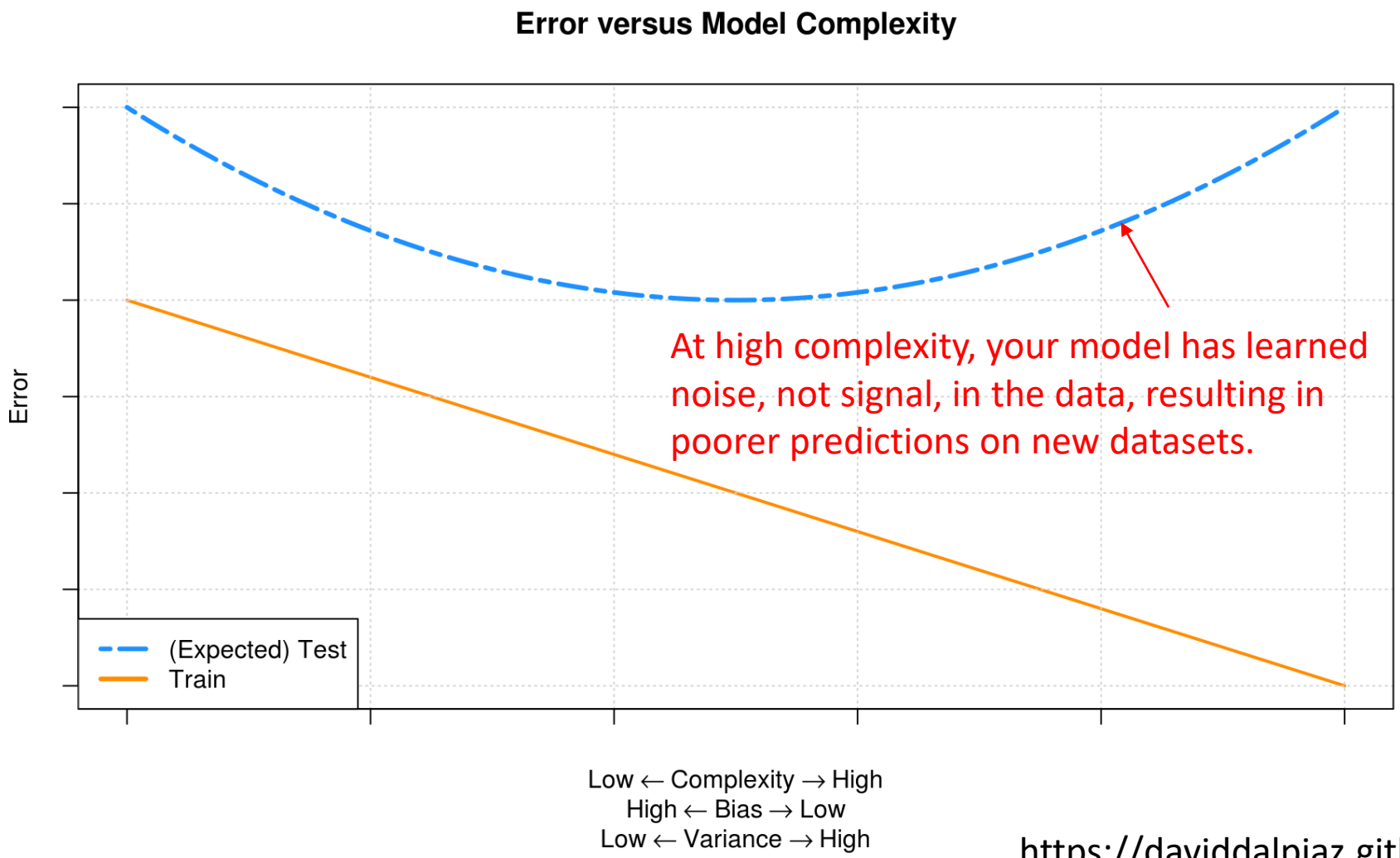
GPI

Finding the Middle Path: Bias Variance Tradeoff



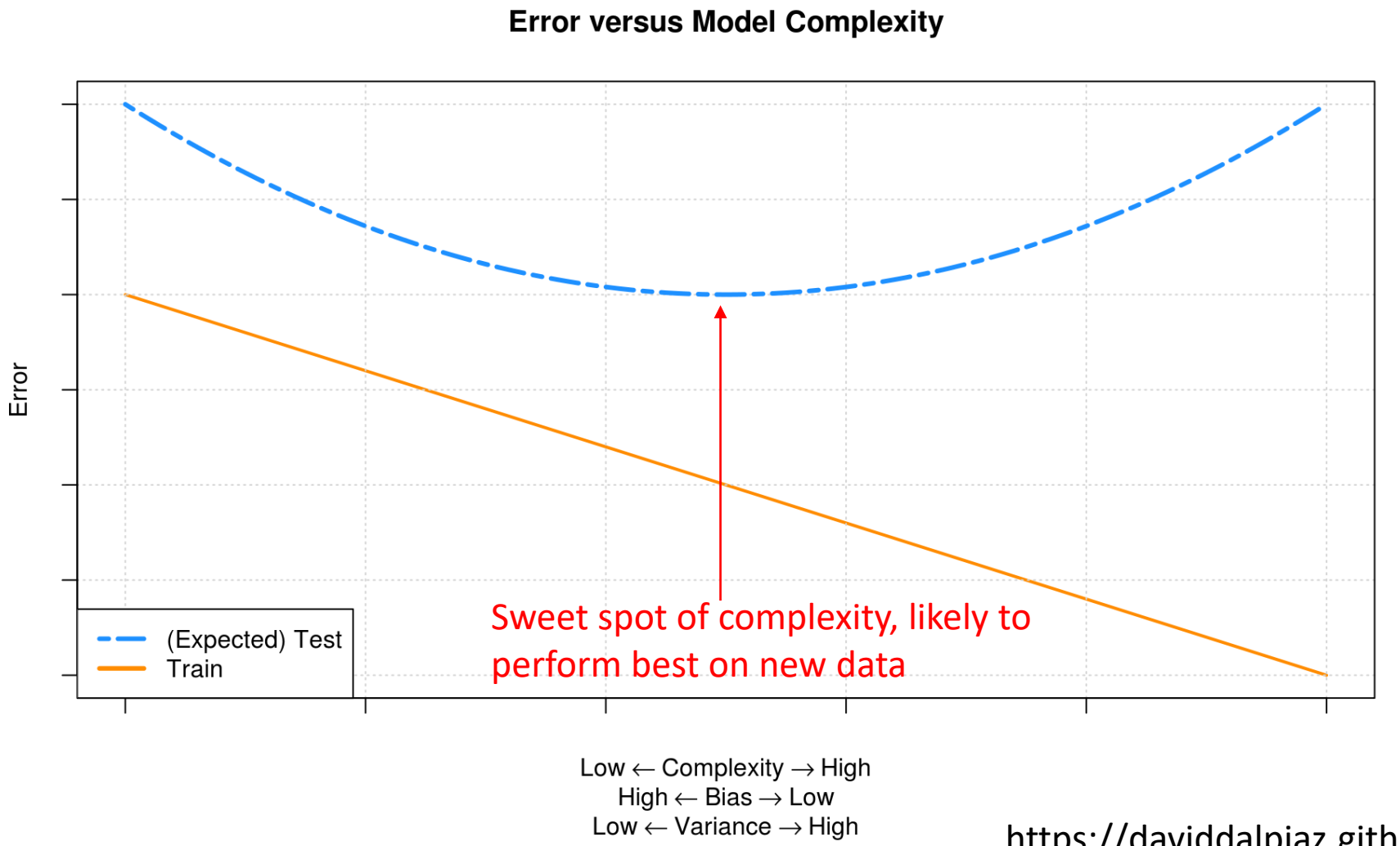
https://daviddalpiaz.github.io/r4sl/

Finding the Middle Path: Bias Variance Tradeoff



https://daviddalpiaz.github.io/r4sl/

Finding the Middle Path: Bias Variance Tradeoff

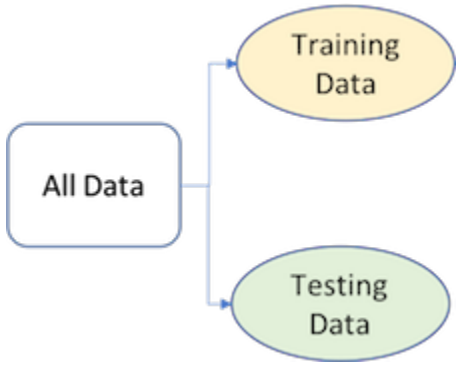


https://daviddalpiaz.github.io/r4sl/

GPI

Avoiding Overfitting in Practice

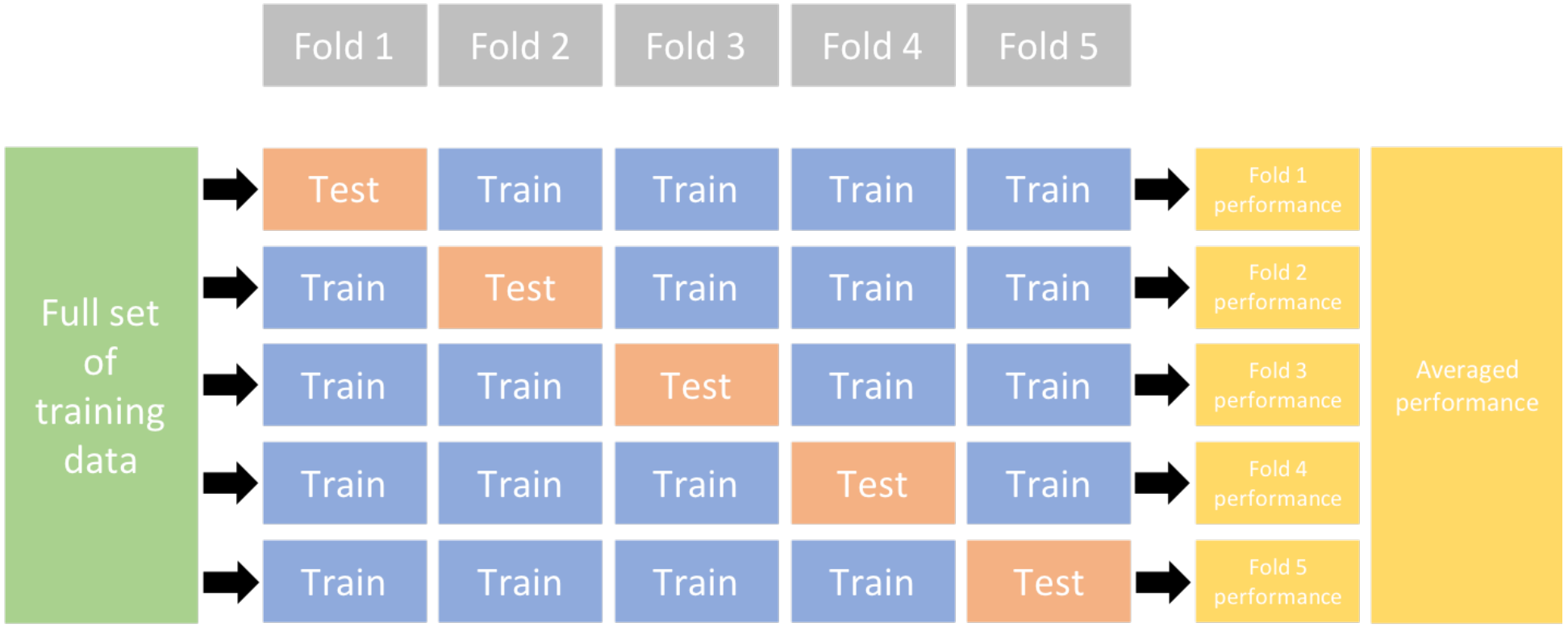
- Use a train/test split (if you have plenty of data, check performance on a test dataset you keep locked up until you're done)
  - Train data used for model selection and fitting (many iterations)
  - Test data used for evaluation (also called hold-out sample. Used only once!)
  - Typical splitting fractions
    - 80:20
    - 70:30



- Use a train/validation/test split
  - Validation set is used for evaluation of model parameter settings on a new dataset, prior to a final evaluation on test data
  - Typical splitting fractions: 70:20:10
- Crossvalidation
  - N-fold (5 or 10-fold common, I prefer 10-fold)
  - Leave-one-out-cross-validation (LOOCV) – sometimes time-intensive

GPI

5-fold crossvalidation



https://bradleyboehmke.github.io/HOML/process.html

GPI

Generalization Error

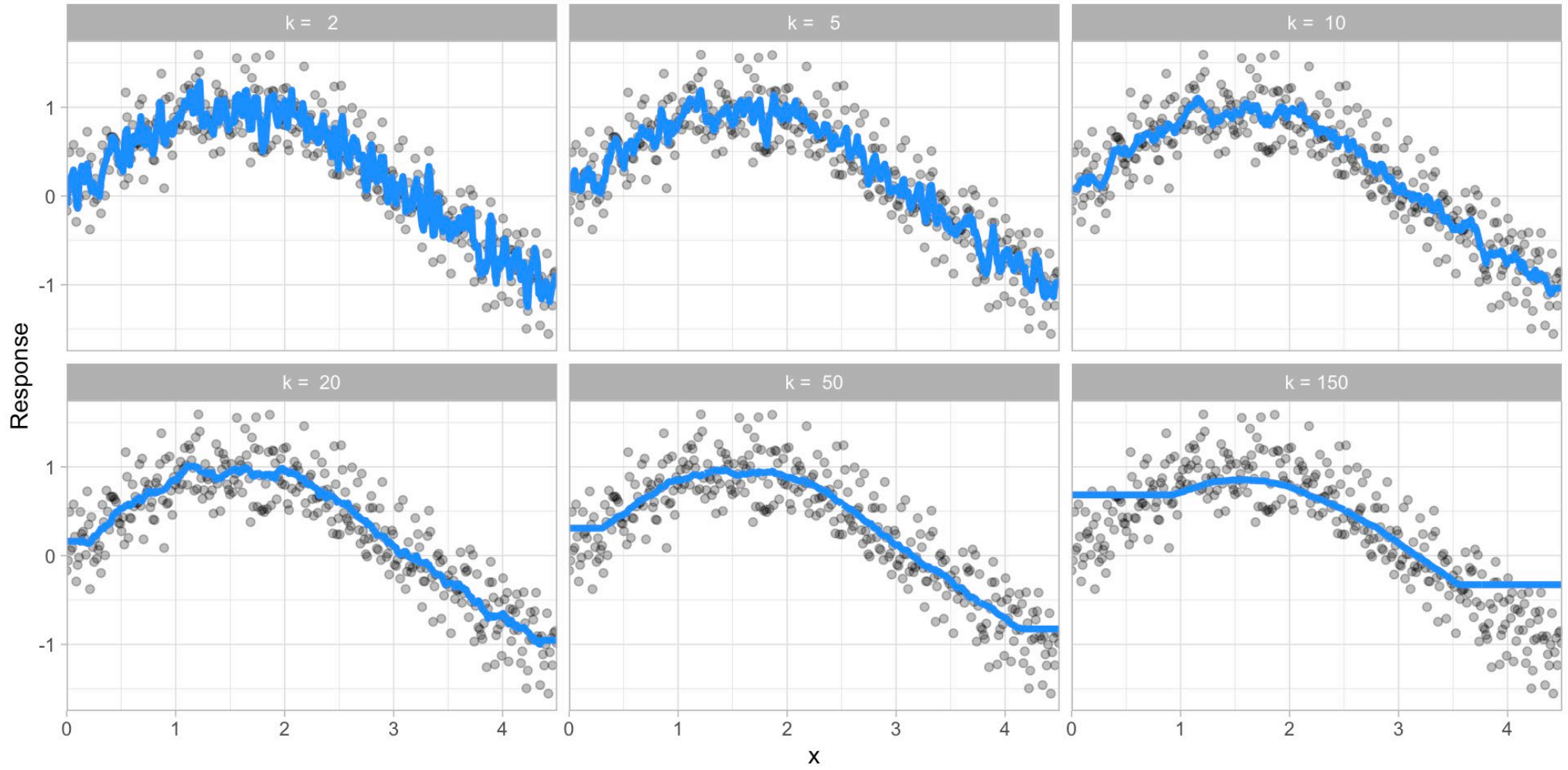
- Some type of generalization error is critical to model evaluation
  - Out-of-sample (hold-out sample) test dataset
  - Crossvalidated error
  - Out-of-bag - some algorithms generate bootstrapped samples and provide estimate of performance on those not in the sample

What is a Model Hyperparameter?

- A setting that is external to the model and whose value is not automatically estimated from the data
- In contrast to our usual understanding of model parameters like the slope coefficients in regression which are estimated as part of the model-fitting process
- They are often specified by the practitioner.
- They can often be set using heuristics (i.e., rule of thumb). What has worked in the past? What do the package developers recommend?
- They are often tuned for a given predictive modeling problem using crossvalidation.

Example Hyperparameter Tuning for k-nearest neighbor regression

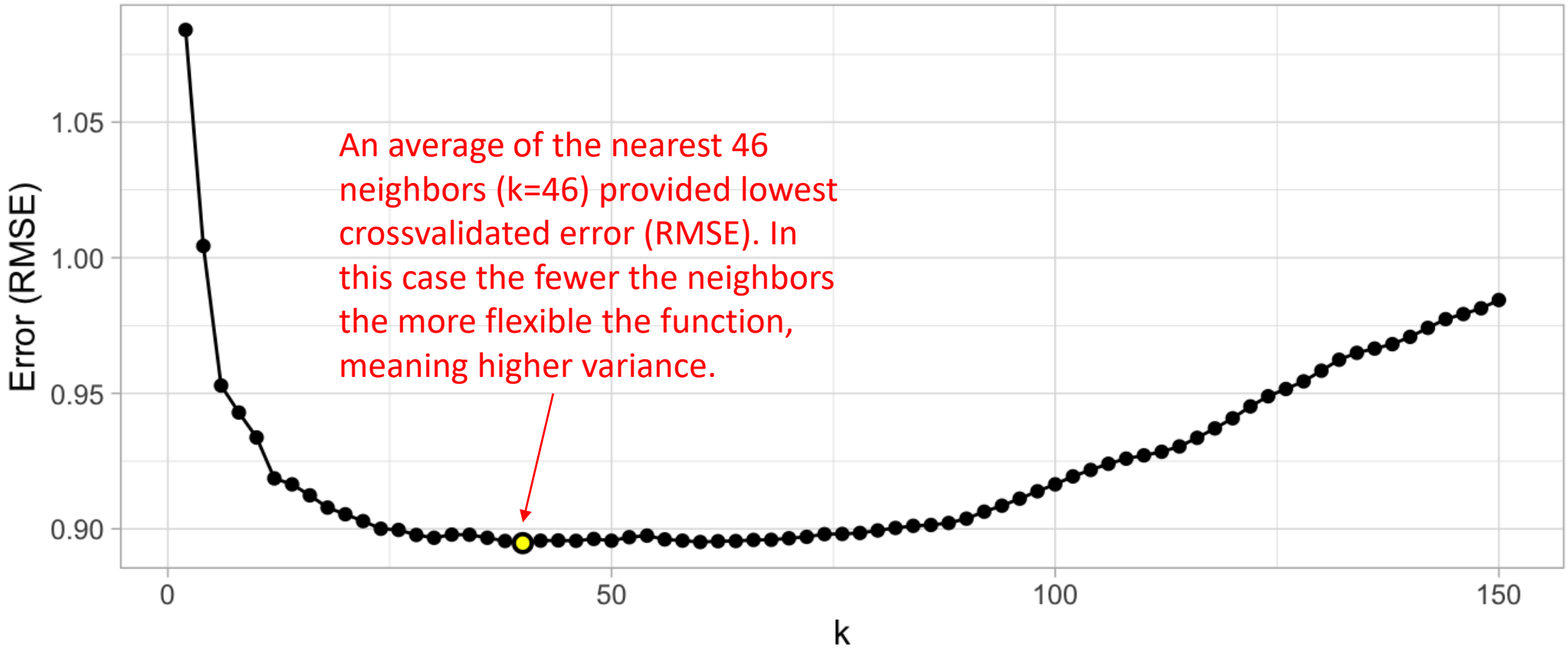
k-nn algorithm calculates value as the mean of the k-nearest neighbors



https://bradleyboehmke.github.io/HOML/process.html

GPI

Example Hyperparameter Tuning



https://bradleyboehmke.github.io/HOML/process.html

GPI

The Curse of Dimensionality

Sample of points get sparse in high dimensions

1d = units
2d = units squared
3d = units cubed
etc.

- As the number of dimensions (i.e., variables) grow, observations move farther apart in data space, reducing the number of samples in a region to make good inferences. This is particularly a problem when one or more variables are worthless.
  - Imagine that books with similar subjects are organized physically near each other on a floor of the library and similar in terms of content, but if we allow the dimension of floor of a multistory library building, then books immediately above and below that floor are relatively near physically, but no longer near in terms of content.

Reducing the Effects of the Curse of Dimensionality

- Drop features which aren't important to your prediction results or use a method which is less sensitive to extraneous factors (e.g., random forest)
- Reduce the number of features using a dimensionality reduction technique such as Principal Components Analysis

Algorithm Choice: Interpretability versus Flexibility

- Simpler algorithms (like linear regression) tend to be highly interpretable, having only a few parameters (particularly if only a few, noncorrelated variables are used)…although correlation is not sufficient to assume causation without experimental manipulation.
- Complex algorithms (like random forest and neural networks) result in many parameters and appear to some to be “black boxes”, hiding the process by which inputs become outputs.

GPI

Algorithm Choice: There is no Free Lunch!

- "…if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems."

David H. Wolpert and William G. Macready. 1997. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 1, NO. 1, APRIL 1997

GPI

Feature Engineering

- Part of the art of machine learning is providing the chosen algorithm with an appropriate representation of the problem at hand.
- As an example, during Exploratory Data Analyses (EDA) an investigator could recognize in an analysis of airplane flight data that there is a weekly pattern, but that certain holidays depart from the usual pattern and therefore deserve their own special coding.

Performance Evaluation: Measuring the Quality of the Fit: Regression*

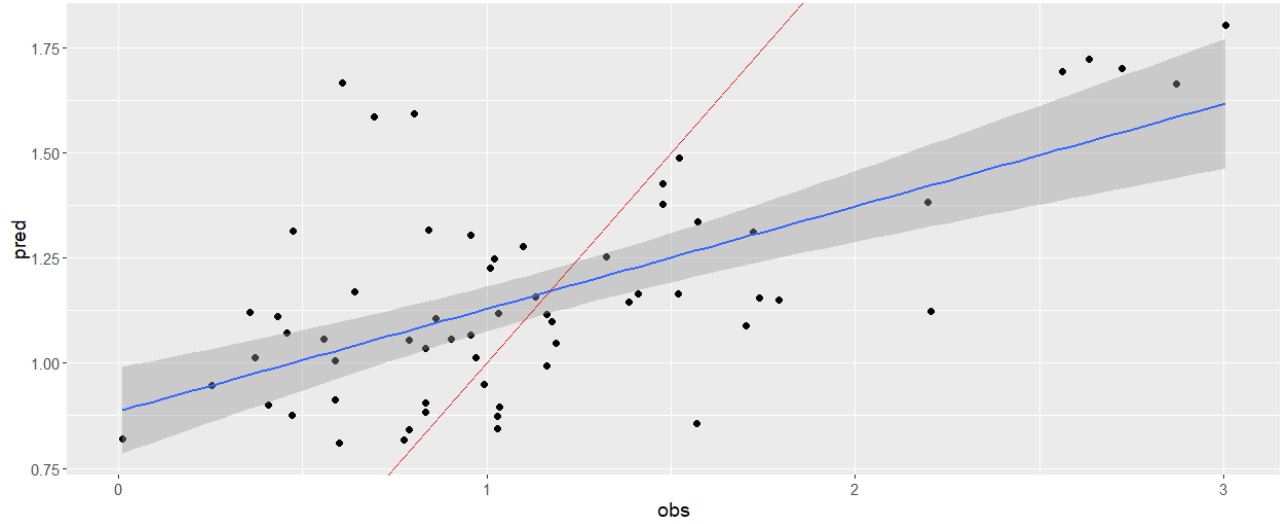
Can be calculated on both training and testing data, although testing or cv results more realistic.

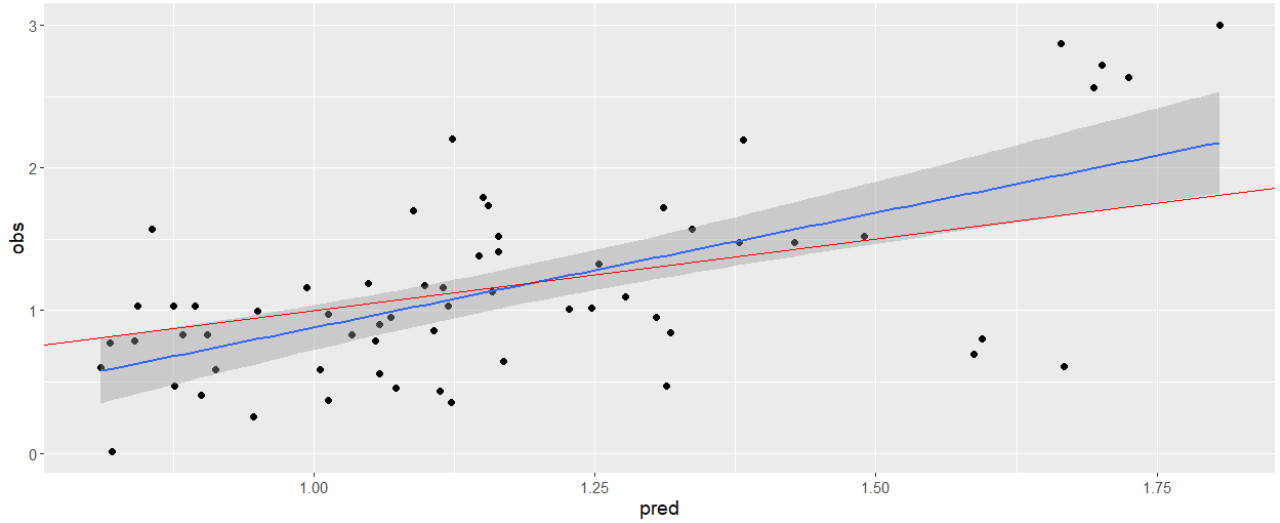
- RMSE – Root Mean Square Error – represents the average residual error in units of original measurements
  - ((Sum(Obs – Exp)^2)/n)^0.5
- r^2 – Pearson Product Moment Correlation between observed and expected model predictions (warning: commonly used but not as reliable as the RMSE because it is dependent on the range of data)
- Comparison plot of the modeled and the actual values (looking for bias)

* Regression means any continuous output algorithm, not just linear regression.

GPI

Which one is better to plot: Pred vs. Obs or Obs vs. Pred?

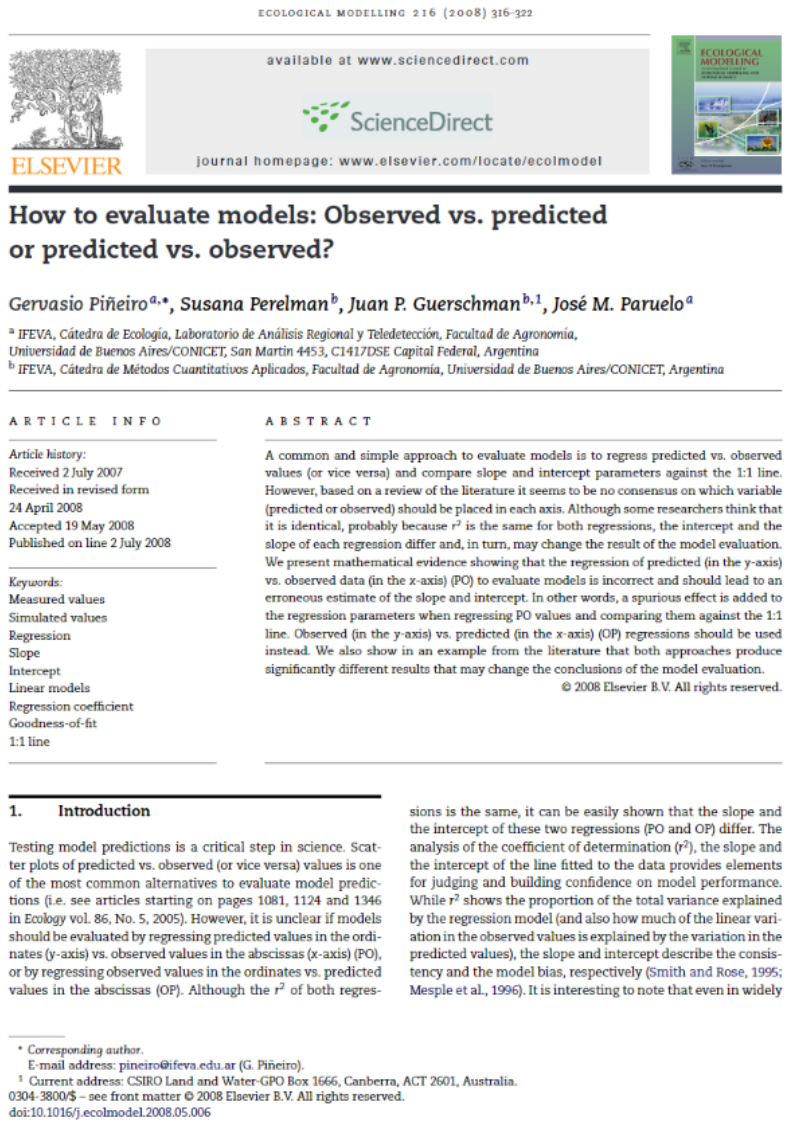




GPI

Yes, it matters!

- About half of publications get it wrong
- Obs (y) vs Pred (x) allows proper comparison with the 1:1 line
- r^2 is unaffected but the slope and intercept are different



ECOLOGICAL MODELLING 216 (2008) 316–322

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

How to evaluate models: Observed vs. predicted or predicted vs. observed?

Gervasio Piñeiro[a,*], *Susana Perelman*[b], *Juan P. Guerschman*[b,1], *José M. Paruelo*[a]

[a] IFEVA, Cátedra de Ecología, Laboratorio de Análisis Regional y Teledetección, Facultad de Agronomía, Universidad de Buenos Aires/CONICET, San Martín 4453, C1417DSE Capital Federal, Argentina

[b] IFEVA, Cátedra de Métodos Cuantitativos Aplicados, Facultad de Agronomía, Universidad de Buenos Aires/CONICET, Argentina

ARTICLE INFO

Article history:
Received 2 July 2007
Received in revised form
24 April 2008
Accepted 19 May 2008
Published on line 2 July 2008

Keywords:
Measured values
Simulated values
Regression
Slope
Intercept
Linear models
Regression coefficient
Goodness-of-fit
1:1 line

ABSTRACT

A common and simple approach to evaluate models is to regress predicted vs. observed values (or vice versa) and compare slope and intercept parameters against the 1:1 line. However, based on a review of the literature it seems to be no consensus on which variable (predicted or observed) should be placed in each axis. Although some researchers think that it is identical, probably because r^2 is the same for both regressions, the intercept and the slope of each regression differ and, in turn, may change the result of the model evaluation. We present mathematical evidence showing that the regression of predicted (in the y-axis) vs. observed data (in the x-axis) (PO) to evaluate models is incorrect and should lead to an erroneous estimate of the slope and intercept. In other words, a spurious effect is added to the regression parameters when regressing PO values and comparing them against the 1:1 line. Observed (in the y-axis) vs. predicted (in the x-axis) (OP) regressions should be used instead. We also show in an example from the literature that both approaches produce significantly different results that may change the conclusions of the model evaluation.

1. Introduction

Testing model predictions is a critical step in science. Scatter plots of predicted vs. observed (or vice versa) values is one of the most common alternatives to evaluate model predictions (i.e. see articles starting on pages 1081, 1124 and 1346 in Ecology vol. 86, No. 5, 2005). However, it is unclear if models should be evaluated by regressing predicted values in the ordinates (y-axis) vs. observed values in the abscissas (x-axis) (PO), or by regressing observed values in the ordinates vs. predicted values in the abscissas (OP). Although the r^2 of both regressions is the same, it can be easily shown that the slope and the intercept of these two regressions (PO and OP) differ. The analysis of the coefficient of determination (r^2), the slope and the intercept of the line fitted to the data provides elements for judging and building confidence on model performance. While r^2 shows the proportion of the total variance explained by the regression model (and also how much of the linear variation in the observed values is explained by the variation in the predicted values), the slope and intercept describe the consistency and the model bias, respectively (Smith and Rose, 1995; Mesple et al., 1996). It is interesting to note that even in widely

* Corresponding author.
E-mail address: pineiro@ifeva.edu.ar (G. Piñeiro).
[1] Current address: CSIRO Land and Water-GPO Box 1666, Canberra, ACT 2601, Australia.
0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.05.006

GPI

Performance Evaluation: Measuring the Quality of the Fit: Classification

- Classification Matrix (a.k.a. Confusion Matrix)
- Accuracy
- Kappa
- Receiver Operator Characteristic Curve (ROC) and Area Under the Curve (AUC)

Misclassification or Confusion Matrix

True Positives

	Death (0)	Survived (1)	Totals
Predicted Death (0)	a	b	a+b
Predicted Survived (1)	c	d	c+d
Totals	a+c	b+d	

True Negatives

a, b, c, d are counts of individuals

- Sensitivity (a/a+c): Probability that a person who died will be modeled as dying.
- Specificity (d/b+d): Probability that a person who survived will be modeled as surviving.
- Accuracy = (a+d /a+b+c+d): Overall probability of a correct outcome.
- Positive Predictive Value (a/a+b): Probability that a model prediction of death is accurate.
- Negative Predictive Value (d/c+d): Probability that a model prediction of survival is accurate.

Tech tip: PPV and NPV change with predicted prevalence of stress types in the population

GPI

Kappa

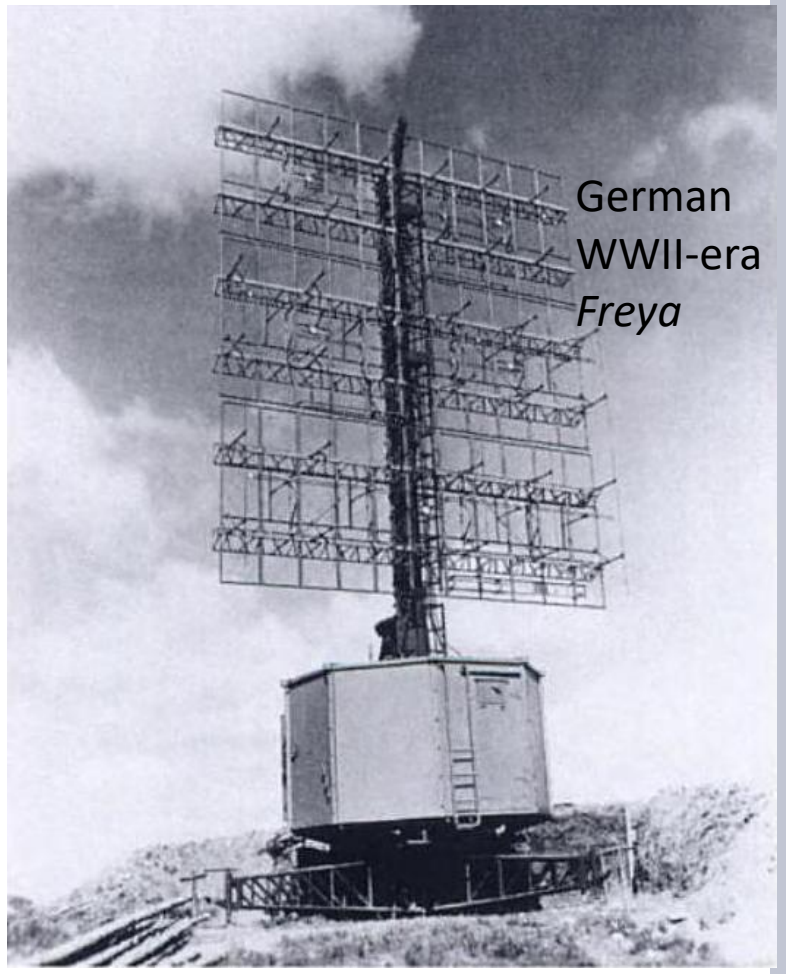
- Accuracy is easy to understand, but it can be a poor reflection of model performance if classes are highly unbalanced.
- As an example, if the rate of occurrence is only 2% in the population, then my model could simply guess all absences and have an accuracy of 98%!
- Kappa ranges from -1 to +1 (perfect agreement) and supposedly weights the accuracy by the randomly expected accuracy, but current thinking is that it is not useful for model comparison*.

* Delgado R, Tibau X-A (2019) Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE 14(9): e0222916. https://doi.org/10.1371/journal.pone.0222916

GPI

ROC Curve

- Signal Detection Theory and its Receiver Operating Characteristic (ROC) Curve developed rapidly in WWII in response to the need to maximize correct detection responses and minimize false detection responses (difficult)
- Represents a way of assessing model performance across all cutoff thresholds
- For example, LDL cholesterol of 160 might be considered high, suggesting treatment, but if we lower that number, we'll capture more true positives but also more false positives



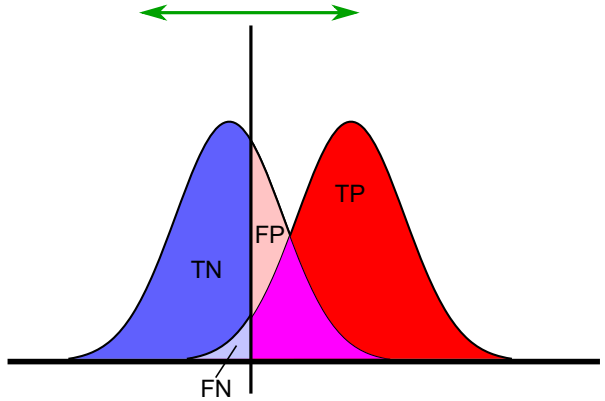
German WWII-era *Freya*

The US National Archives and Records Administration - Foto 111-SC 269043, Public Domain, https://commons.wikimedia.org/w/index.php?curid=6115058

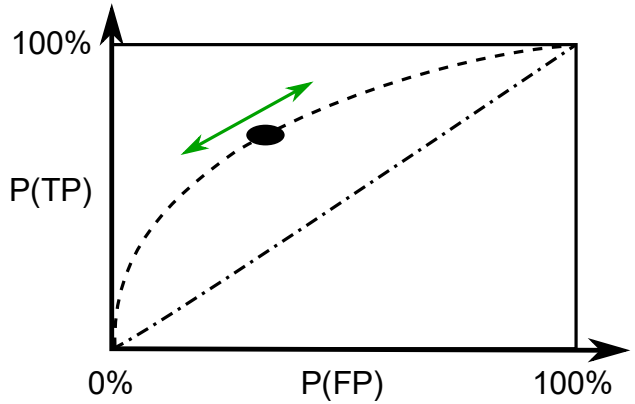
GPI

ROC Curve

- Closer the curve runs to the upper left corner (all true positives and no false positives), the better.
- The Area Under the Curve is an overall measure of how much curve deviates from the chance 1:1 line.



TP	FP
FN	TN



GPI

Area Under the (ROC) Curve: General Guidance

- .90-1.0 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail
- <.50 is worse than guessing so you just flip the assignment of your model's classification of what constitutes success!

Another perspective is the AUC represents the probability that a randomly selected success will show a higher model predicted probability than a randomly selected failure.

GPI

Prior to Statistical Machine Learning

- You have spent significant time cleaning and getting to know your data via Exploratory Data Analyses.
- Practiced some feature selection (i.e., retaining only those variables you wish to include in your modeling process).
- Recoding variable names and values so that they are meaningful and more interpretable.
- Recoding, removing, or some other approach to handling missing values.

Iterative Modeling Process

- Splitting data (withholding a test set or validation and test set) or establishing the type of crossvalidation to use (or both).
- Model 1
  - Training model using training set and crossvalidation
  - Tuning model using crossvalidation
  - Evaluating model performance
  - Characterizing what features of data that model has learned
  - Possibly performing additional feature engineering to address model deficiencies
  - Retraining and evaluating model performance
- Model 2, Model 3, etc. (repeat Model 1 steps)
- Evaluate all models on held-out test data

GPI