CSX415

Data Science Principals and Practice

Model Measurement

Christopher Brown
U.C. Berkeley / Decision Patterns LLC
Spring 2018

Berkeley
UNIVERSITY OF CALIFORNIA

Model Measurement

Berkeley
UNIVERSITY OF CALIFORNIA

Many of you have struggled with model measurements?

What is a good metric?

Which metric do I use?

How do I use that metric.

What data do I use to calculate performance?

What do I need to communicate model performance?

Metric Requirements

1. Quantify model performance (numeric)

2. Should be a scalar value

3. Allow for comparison of models ..
   having the same response

Be independent of the model methods

Be stationary wrt to the data.

Caveats

1. No single metric can tell you whether a model is good or bad.

2. Model performance is always **evaluated** wrt some other value.

- Existing model
- Naïve Model

3. The metric that you use for measuring model performance may not be the one used in the algorithm (search function)

Note

Loss ≠
Model Performance

For example:
Linear Regression uses SSE, often model evaluated with RMSE | MAE.

You need to understand this distinction…

--> algorithms may not be optimizing your performance measure.

Regression Models

Berkeley
UNIVERSITY OF CALIFORNIA

Regression metrics all have the same recipe.

They all have the form

```
(y - y_hat)            %>%   # error
   [abs|sq]            %>%   # fix sign
   [agg_fun(s)]        %>%   # aggreg.
   [scale|normalize]   %>%   #
```

→ one numeric value per model.

Binomial Metrics

Accuracy.

"How many did I get right"

$$\text{Accuracy} = \begin{Bmatrix} 1 |\ y = \hat{y} \\ 0 | y \neq \hat{y} \end{Bmatrix}$$

* Usually expressed as a ratio/percentage.

“How many did I get wrong”

Error (Rate).

$$\text{Error} = \begin{Bmatrix} 0 | \ y = \hat{y} \\ 1 | y \neq \hat{y} \end{Bmatrix}$$

- Usually expressed as a ratio/percentage.
- 1 - accuracy

Berkeley
UNIVERSITY OF CALIFORNIA

Many of you have decided to use accuracy …

Problematic because accuracy/error rate does:

- not describe the nuances story about your model.
- Does not account for the "state" of the system.

So What's Better?

Problematic because accuracy/error rate does:

- not describe the nuances story about your model.
- Does not account for the "state" of the system.

Cohen's Kappa

- Accuracy can be misleading because:
  1. Does not account for highly imbalanced classes

Cohen's Kappa compensates for these two.

→ Think of it as *adjusted* accuracy.

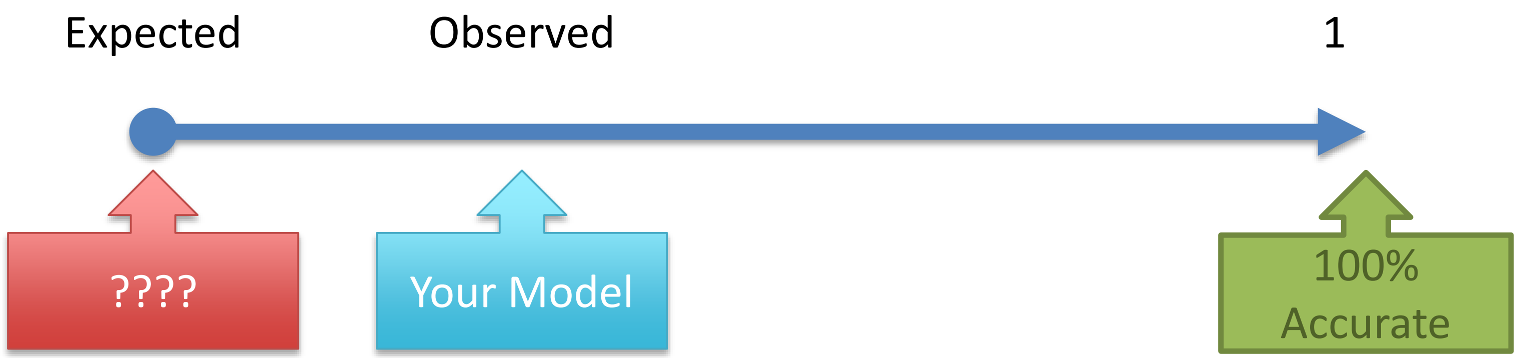
$$\kappa = \frac{O - E}{1 - E}$$

Ranges between 0-1

*Rule of thumb: Kappa values within (0.30-0.50)+ → good fit

What is this?

Think of it as a journey to 100% Accuracy (or 0% Error).



$$\kappa = \frac{How\ Far\ You've\ Come}{How\ Far\ You\ Can\ Go}$$

There is a more nuanced way to think about it.

Total Population

	Observed	
Total Population	Positive	Negative

Total Population

Predicted	Positive
	Negative

Berkeley
UNIVERSITY OF CALIFORNIA

(BINOMIAL) CONFUSION MATRIX

Total Population		Observed	
		Positive	Negative
Predicted	Positive		
	Negative		

Total Population		Observed	
		Positive	Negative
Predicted	Positive	“True” Positive	“False” Positive (Type I Error)
	Negative	“False” Negative (Type II Error)	“True” Negative

Total Population		Observed	
		Positive	**Negative**
Predicted	**Positive**		
	Negative		

Use this as a tool to reason about your problem and evaluate the choices.

Total Population		Observed	
		Positive	Negative
Predicted	Positive	+ benefit	- cost
	Negative	- cost	+ benefit

Total Population

Observed

Positive

Negative

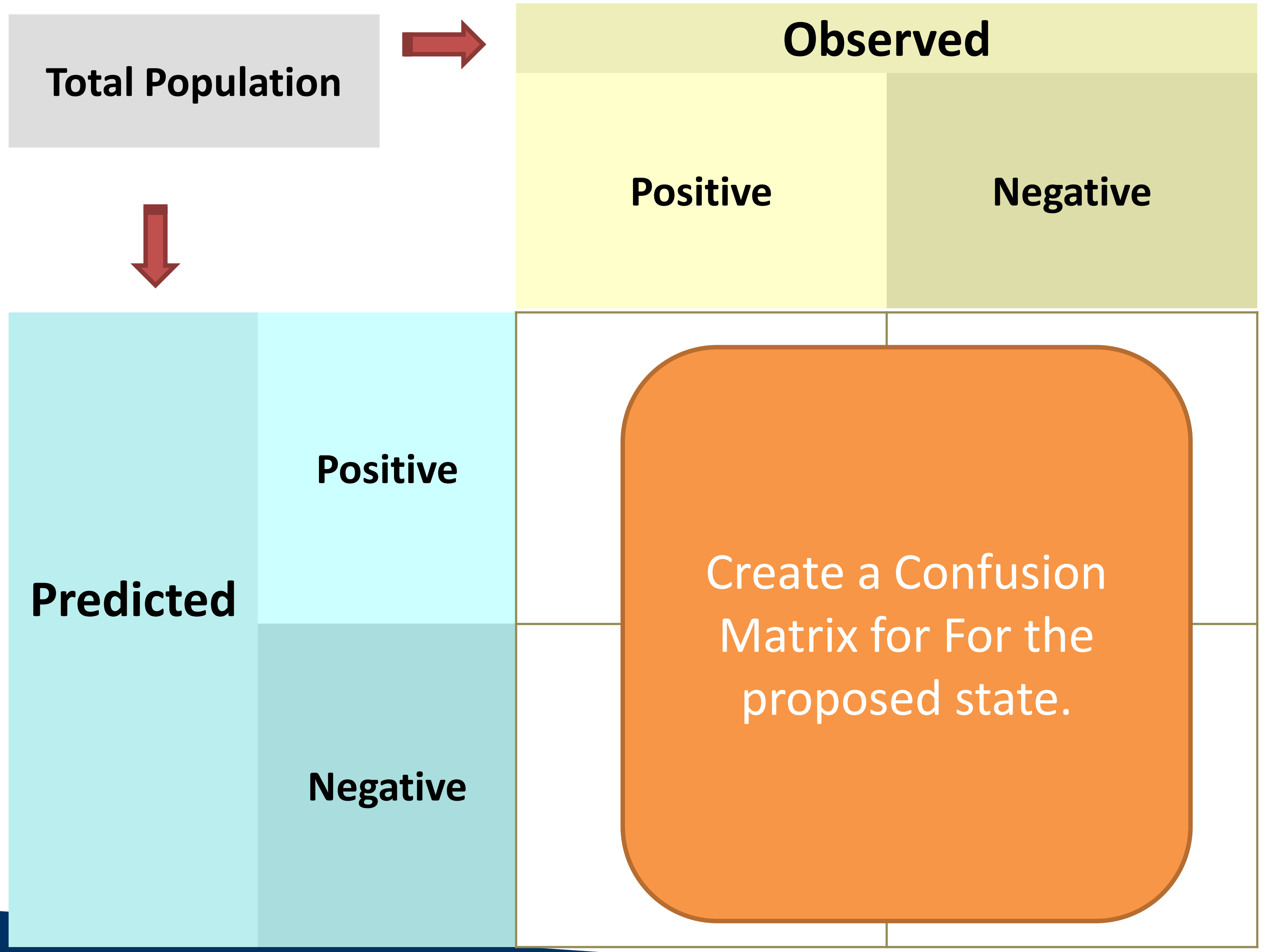
Predicted

Positive

Negative

Create a confusion Matrix for Current State.

Example


Total Population
Observed
Positive
Negative
Predicted
Positive
Negative
Create a Confusion Matrix for For the proposed state.

Note

Loss ≠
Model Performance

Most algorithms optimize using 0-1 Loss.

$$\text{Loss} = \begin{cases} 0 | \; y = \hat{y} \\ 1 | y \neq \hat{y} \end{cases}$$

Enumerating the TP, FP, FN, TN isn't enlightening …

Need to normalize by the number of observations …

but generally not the total true positives

Alternatives: “**Rates**” Normed by Observed

Total Population		Observed: Positive	Observed: Negative
Predicted	Positive	True Positive Rate (TPR), **Sensitivity**, Recall True Positives / **Observed Positives**	False Positive Rate (FPR), Fall-Out False Positives / **Observed Negatives**
Predicted	Negative	False Neg. Rate (FNR), Miss rate False Negatives / **Observed Positives**	True Neg. Rate (TNR), **Specificity** (SPC) True Negatives / **Observed Negatives**

Berkeley UNIVERSITY OF CALIFORNIA

Alternatives: Norm by Predicted

Total Population		Observed	
		Positive	Negative
Predicted	**Positive**	Pos. Predictive Value (PPV), **Precision** $\frac{\text{True Positives}}{\textbf{Predicted Positives}}$	False Discovery Rate (FDR) $\frac{\text{False Positives}}{\textbf{Predicted Positives}}$
	Negative	False Omission Rate(FOR) $\frac{\text{False Negatives}}{\textbf{Predicted Negatives}}$	Negative Predictive Value (NPV) $\frac{\text{True Negatives}}{\textbf{Predicted Negatives}}$

More Fun …

https://en.wikipedia.org/wiki/Sensitivity_and_specificity
https://en.wikipedia.org/wiki/precision_and_recall

Classification Performance

- Accuracy … problems?
- Confusion Matrix
  - `table`
  - `caret::confusionMatrix`
  - `ModelMetrics::confusionMatrix( actual, predict, cutoff )`

1.You need 2.

Why?

18 Separate Measure, 20+ different names …

Which do you use?

2. Have to be same type

3. Is determined by the specific application

What is important?

What does the industry use?

Berkeley
UNIVERSITY OF CALIFORNIA

Use the one that suits your needs.

1.You need 2.

Why?

2. Have to be same type

3. Is determined by the specific application

What is important?

What does the industry use?

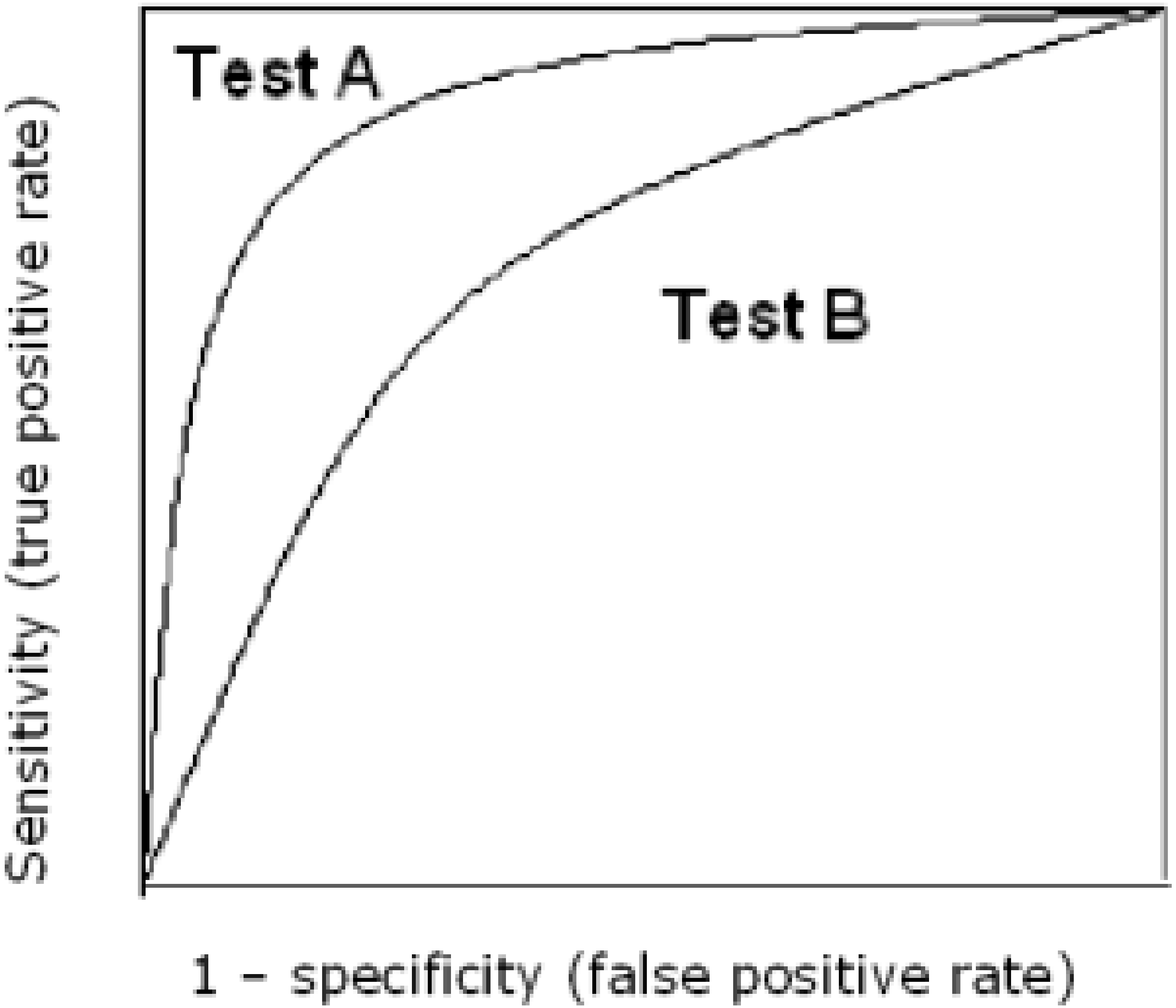
If Models How Are Classes Discriminated?

Models Put Out Class Probabilities …

The Distinction Between the Class is determined by the Data Scientist

Decision Point Analysis

ROC Curves



Test A
Test B
Sensitivity (true positive rate)
1 - specificity (false positive rate)

ROC Recipe

Order Observations According to model predictions probability…

Choose two metrics (for the axis)

Calculate cumulative metrics

Resource:
ROCR, pROC, plotROC (packages)

Mutli-nomial Classification

Berkeley
UNIVERSITY OF CALIFORNIA

Classification Performance

- `predict` methods can provide
  - Classes
  - Class probabilities

Better!

- Class probs → Classes?
  - Apply ***softmax*** function

$$\hat{p}^*_\ell = \frac{e^{\hat{y}_\ell}}{\sum_{l=1}^{C} e^{\hat{y}_l}}$$

- Probabilities often need post predict → calibrations (talk about this with deployment)

Even More Complication …

- Not all errors need count “*equivocal zone*” or “*intermediate zone*”
- *Prevalent when the model has three choices, e.g. A or B or Nothing.*

Assets

Model Performance

- **Determine** performance metric:
  - –Regression: RMSE, MAE, MAPE, …
  - –Classification: Accuracy, TPR, Kappa

- **Fit** Model

- **Calculate** statistic (“metric”) on data

Problems

“*training*” or “*apparent*” performance will:

– predict very well, unbelievably well

– Not generalize to *new data.*

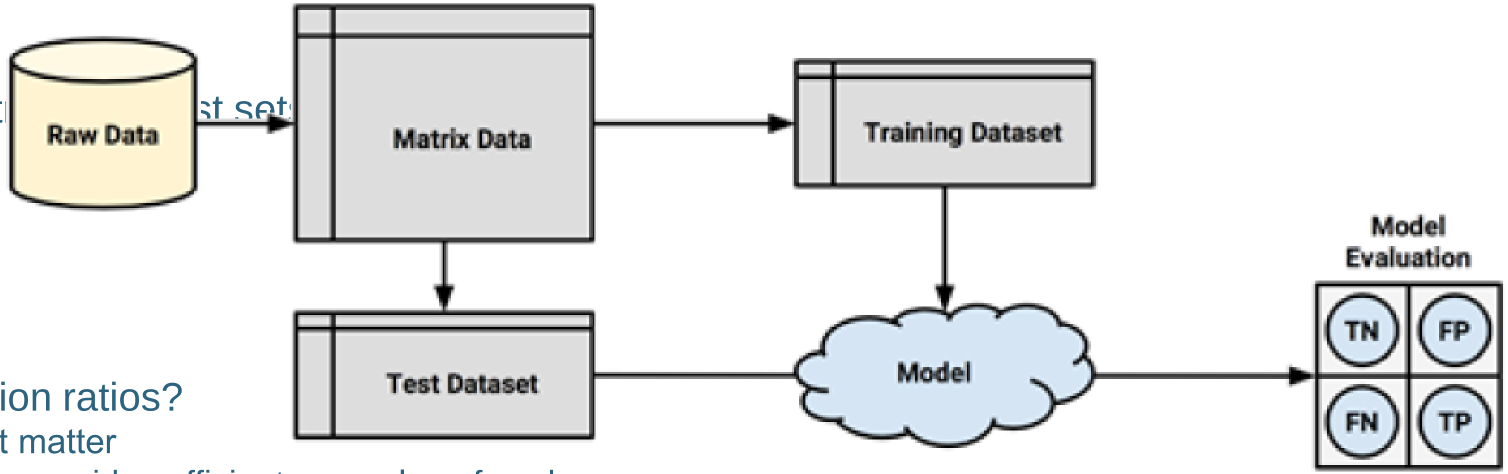
Cardinal Rule

DO <u>NOT</u> ESTIMATE PERFORMANCE ON YOUR TRAINING DATA

→ Need technique for unbiased estimate for calculating performance

1: Hold Out Method

- Partition data into t st sets



- What are the partition ratios?
  - Large N: doesn't matter
  - Small N: Need to provide sufficient examples of each case

How confident are you about your performance metric?

Berkeley
UNIVERSITY OF CALIFORNIA

Measurements and Statistics

Measurement : *quantification* of a phenomena

Statistic

measurement of a *stochastic* phenomena

Deterministic
≠
Stochastic

Examples:

```
• mean(x)   <- x measurements of stochastic process
               mean is a statistic
```

Exercise

calculate the mean of some measurement of (x)?

Exercise (cont.)

Question:

How good is your estimate of the statistic `mean(x)`?

Statistics

- “True” value unknown → uncertainty
- Uncertainty can be measured
  - Variance
  - Standard deviation
  - Confidence Interval
  - …
- Repeated measurements decrease the uncertainty

Works the same for models!

x : measure of model performance statistic?

Resampling

Kuhn – Two benefits of resampling

- Selection of optimal tuning parameter(s) – to come
- Unbiased estimate of model performance

Resampling Strategies

- Repeated Holdouts
- K-Fold Cross Validation
- Bootstrap

Repeated Holdout

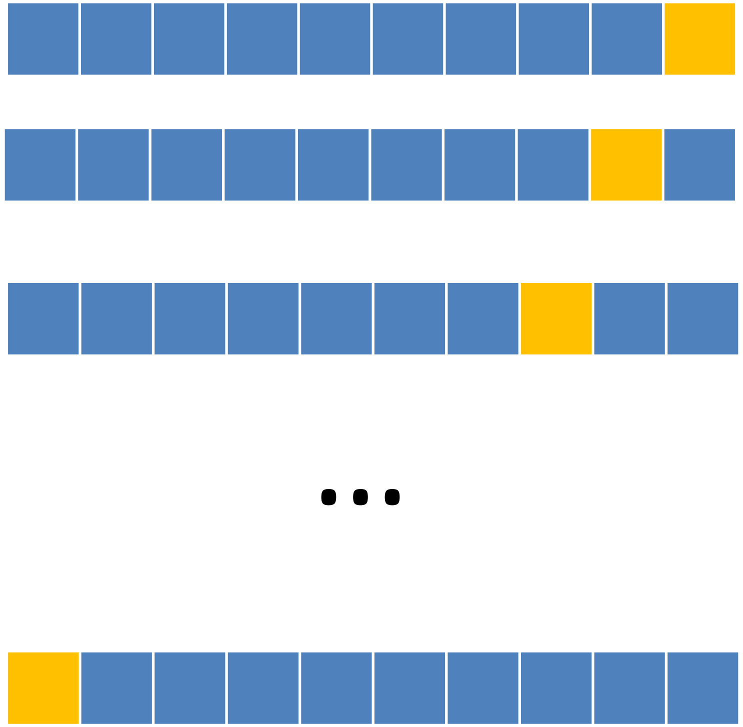
AKA Monte Carlo Splitting

- Split data 75%-25%
  - Fit Model
  - Calculate Performance Metric
  - Repeat with Different Split(K-times)

- Calculate Metric

$$Metric = AVG_i(metric)$$

10-Fold Cross Validation



LOOCV : K→n

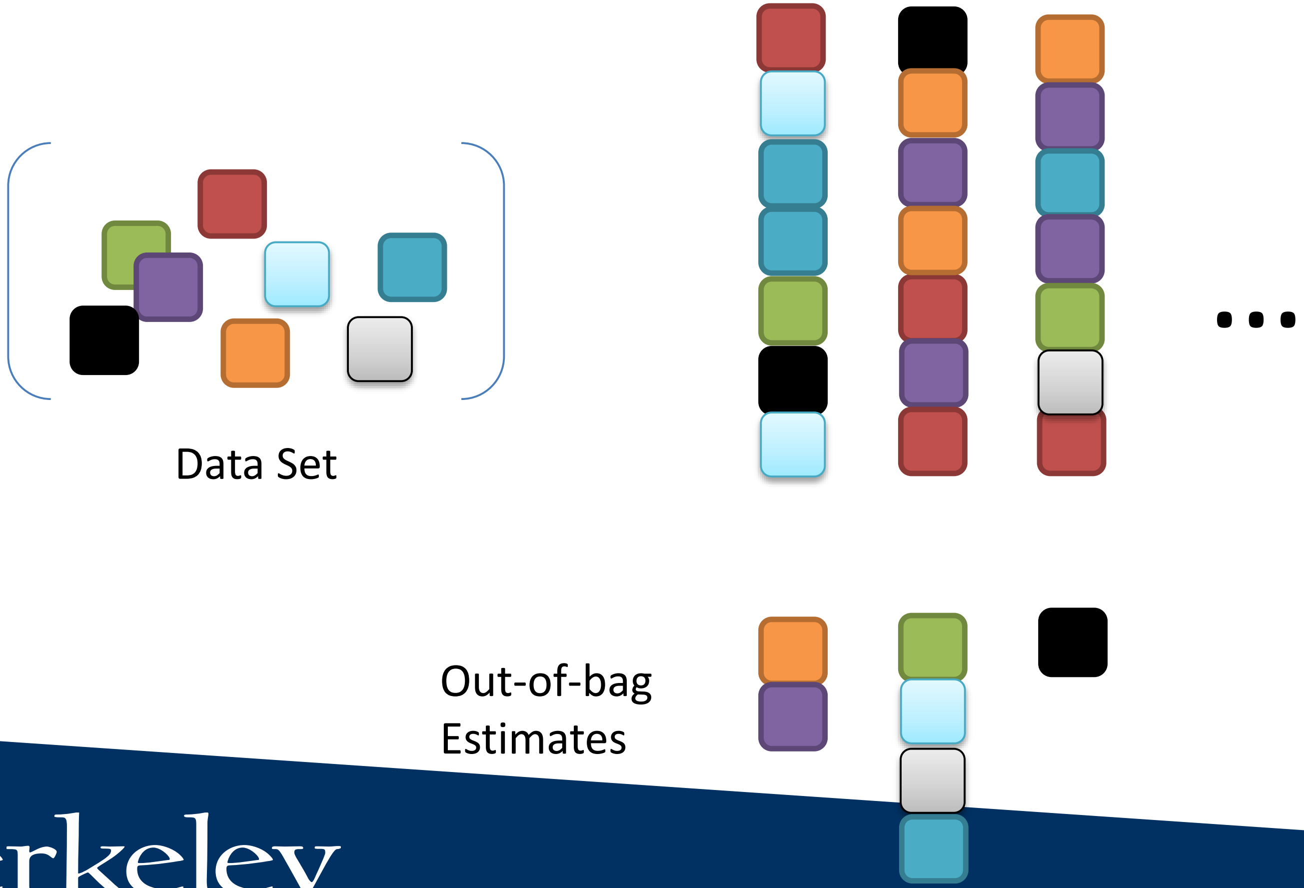
- Split the data set into 10 equal sized samples.
- Leave one sample out (fold)
  - Fit the model
  - calculate the metric on the fold
  - Repeat choosing another sample until done
- Calculate Metric

$$Metric = AVG_i(metric)$$

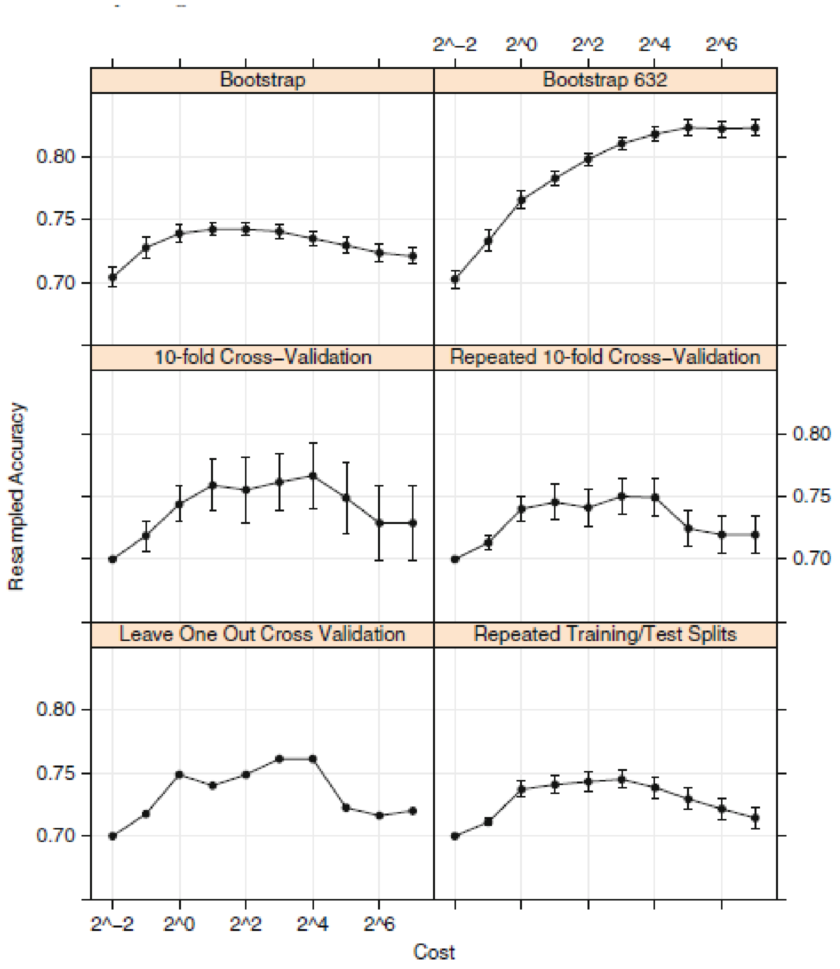
- 5 or 10-fold common

Bootstrap

- "Sampling with Replacement"



Which Is Best?



- There isn't one.

K-fold cross validation

Higher Variance
Lower Bias

Bootstrap

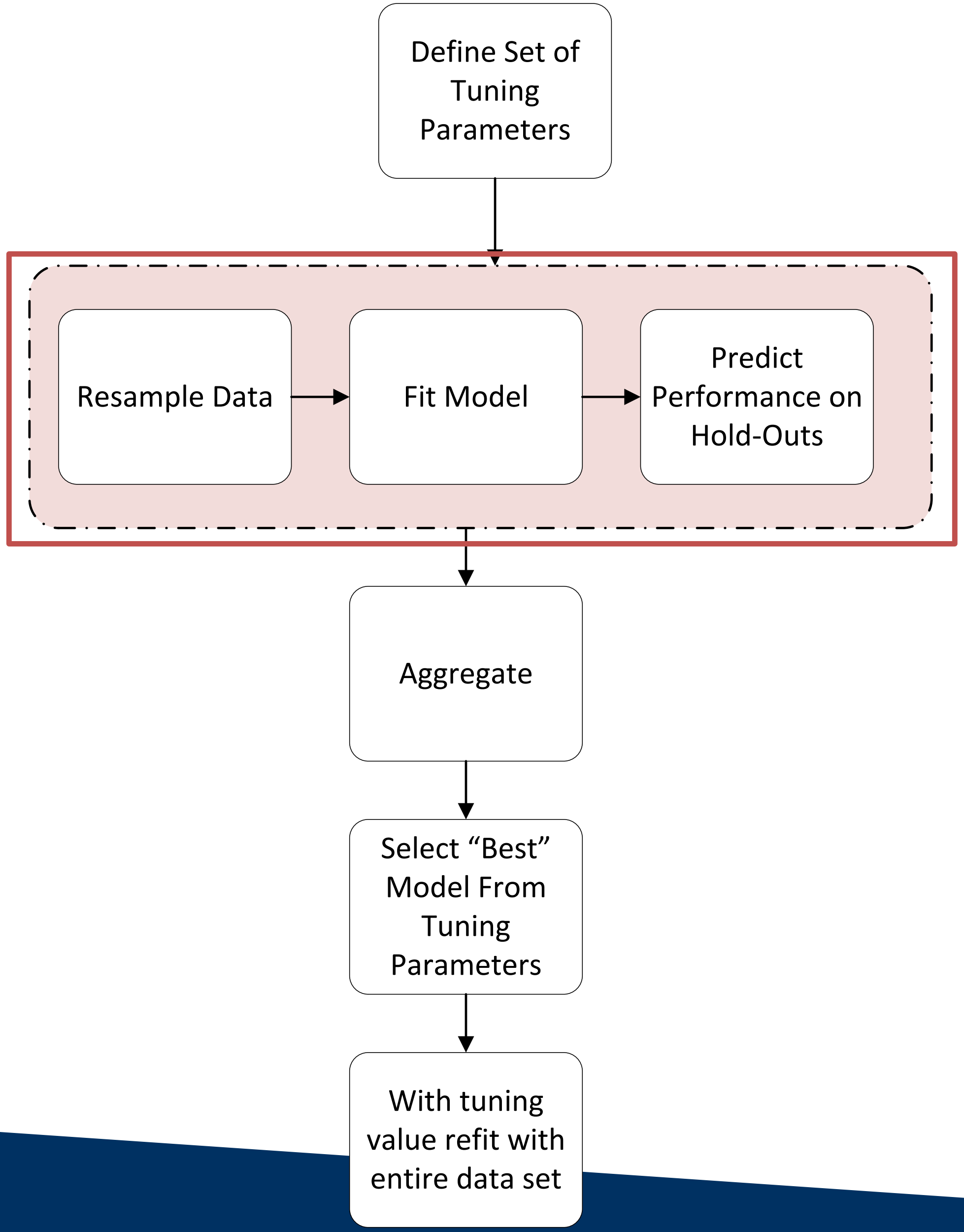
Lower Variance
Higher Bias

Better to employ resampling than worry about not resampling

Berkeley
UNIVERSITY OF CALIFORNIA

Resampling Process

Today's Focus



Berkeley
UNIVERSITY OF CALIFORNIA

Resampling

- Best Solution (n-permitting)
  – split data into training and test data
  – and do what Kuhn says.

Why(?)

- Easy to interpret defend
- Requires data not be consumed by model
- Computationally easy
- Is generally not (by itself) the most accurate → no confidence

MODEL PERFORMANCE IS <u>*NOT*</u> *Training* PERFORMANCE