Good Data Science Is Validating

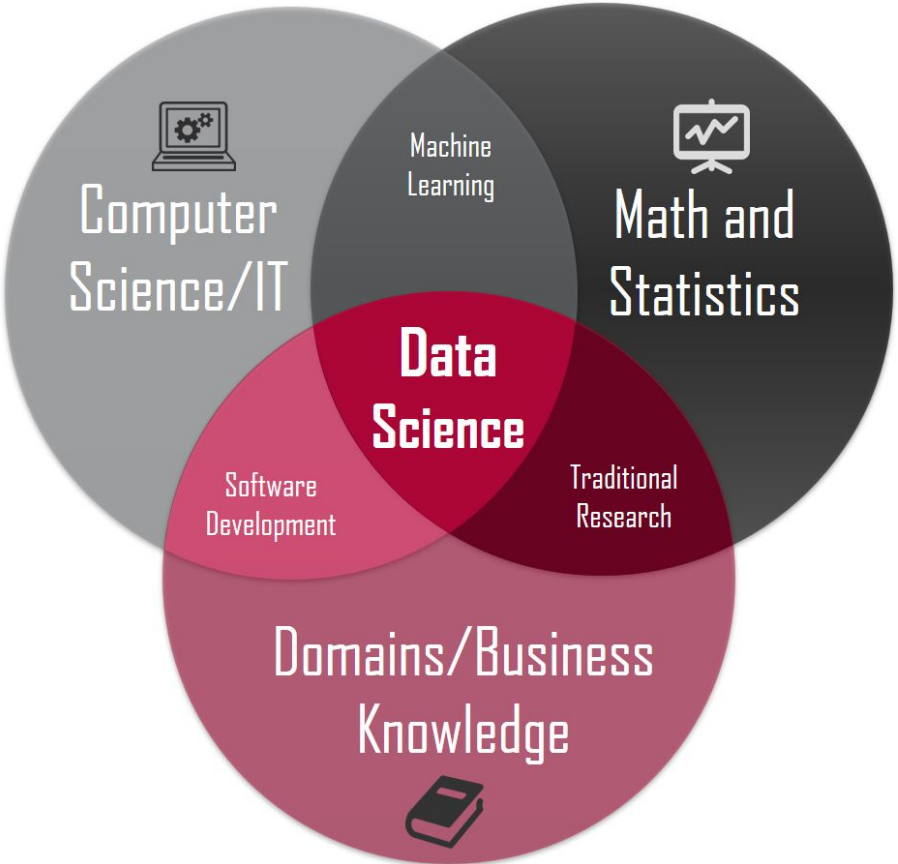
Peter Casey, Ph.D.

A little about me...

- PhD in Political Science from WashU
  - Failed Ryan Moore's first exam
  - Passed the class with an A-
- Work at intersection of data science and politics / public policy
  - Democratic National Committee, 2014 - 2016
  - The Lab @ DC, 2017 - 2019
  - Catalist, 2019 - 2021
  - Data Science for Social Good, 2021-2022

What does a Data Scientist do?

- Lots of smart people have answered this question
- A Data Scientist makes informed decisions about how to get actionable information from data
- One way that data scientists do this is predictive modeling



Planning a predictive model

In predictive modeling, there are several decisions a Data Scientist needs to make

- What data to use
- How to construct the outcome
- What features to use and how to construct them
- How to select which features to include
- Which models to test and how to tune model parameters
- Whether to use nested models or whether to calibrate
- **How to validate the model**

Validation

- How you know your model is working
- How you know your model generalizes to other observations (not overfit)
- Check that the model is performing well across subgroups within your population (checking for bias)
- Check that the model can be used for what you want to use if for

Validation Decisions

- Choose a performance metric (or metrics)
- Choose how you'll construct your validation set(s)
- Figure out what subsamples you may want to validate (where might your model be biased)

Choosing your Performance Metric

Performance metrics

- Focus on classifiers
  - More common
  - More performance metrics to choose between
- Choosing a performance metric
  - Are you trying to…
    - Efficiently target a costly intervention? (Precision)
    - Identify as many instances of an outcome as you can? (Recall)
    - Distinguish between two different outcomes? (ROC-AUC)
    - Fit closely to actual probabilities? (Brier Score, Log Loss)

What we're optimizing for

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Accuracy

- The proportion of correct predictions
- (TP + TN) / (TP + TN + FP + FN)
- Commonly used, but *usually not the right metric*
  - Treats false positives and false negatives *equally*. Usually, we care more about one than another.
  - Usually, sets threshold for positive and negative prediction at 0.5, but this may not be the best threshold value (more on this later)

Precision

- The proportion of the model's positive classifications that are actually positive
- **Positive Predictive Value:** TP / (TP + FP)
- How often the model's positive guesses are correct
- Useful when you have limited resources and want to identify the best targets for your intervention (e.g., inspecting for rodents, reaching out to voters)

Precision @ N

- Precision typically looks at the proportion of correct classifications with a predicted probability over 0.5
- When resources are severely constrained, and you know how many people you want to reach out to you may want to use **Precision @ N**
- First identify the number N you want to target
- Then look at precision for the N targets with the highest predicted probability

Recall (True Positive Rate)

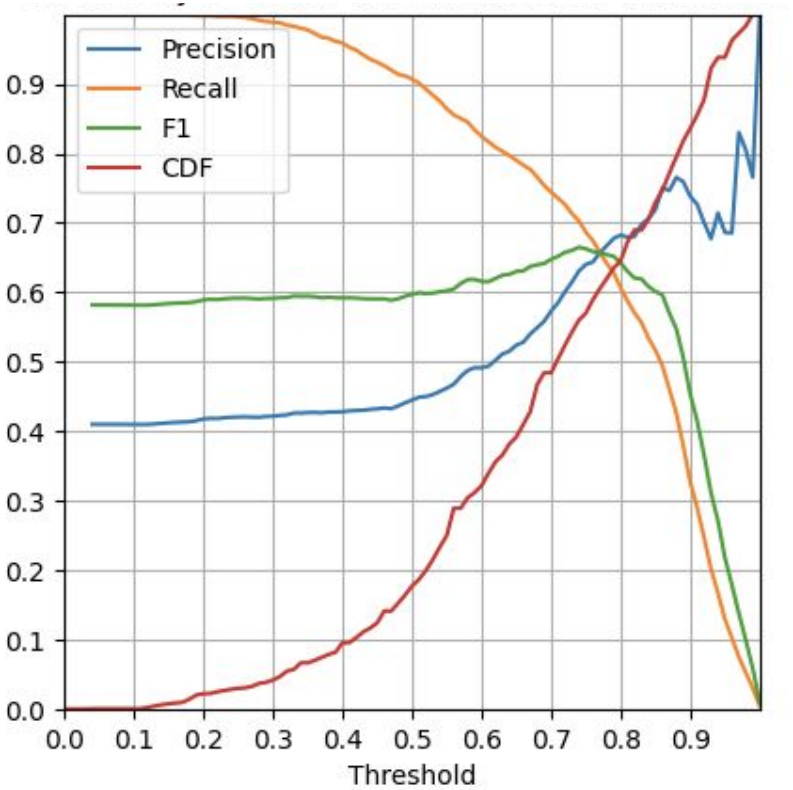
- **Recall** (also known as **sensitivity**) is the the proportion of actual positives that the model correctly identifies
- **True Positive Rate**: TP / (TP + FN)
- Useful when the cost of your intervention is low but the cost of missing a positive case is high (e.g., fraud)

Fbeta Score

- Usually there is some cost or trade-off to false positives, so we want to balance our recall against our precision
- Fbeta allows you to choose the balance between precision and recall
  - The F1-Score (beta=1) provides an equal balance between precision and recall
  - Beta = 0.5 values precision twice as much as recall
  - Beta = 2 value recall twice as much as precision
- Only returns score at a single threshold
  - Most implementations set threshold at 0.5
  - This may not be the best threshold value

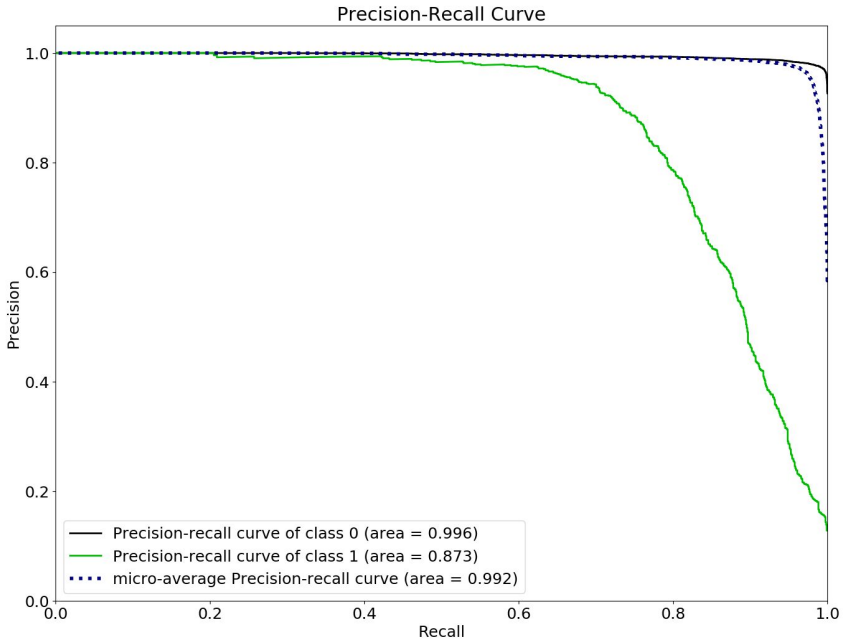
Combining Performance Metrics

- May want to consider precision and recall curves on their own when evaluating:
  - Model performance
  - Choosing model decision threshold
- Example
  - At 0.8
    - Precision: ~0.7
    - Recall: ~0.6
    - CDF: ~0.65, or 35% of sample



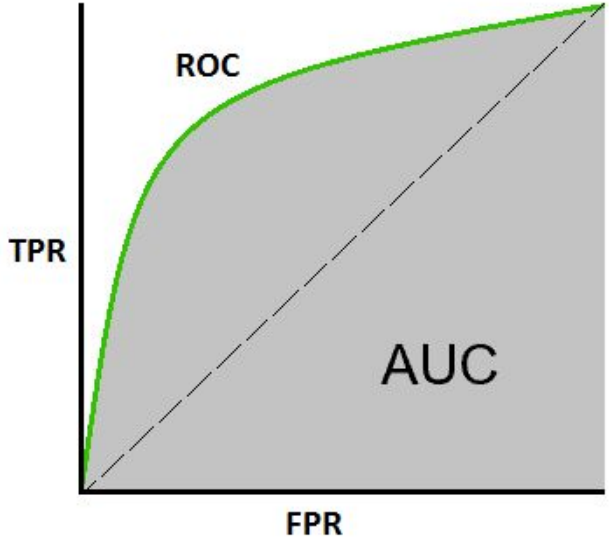
Precision-Recall Curve

- A good way to evaluate decision threshold
- Best to use if:
  - Your classes are very imbalanced (i.e., very few observations of positive outcome)
  - You care more about positive classifications than negative
- Useful evaluation metrics
  - Area under the curve (AUC)
  - Average precision: weighted mean of precision at each threshold



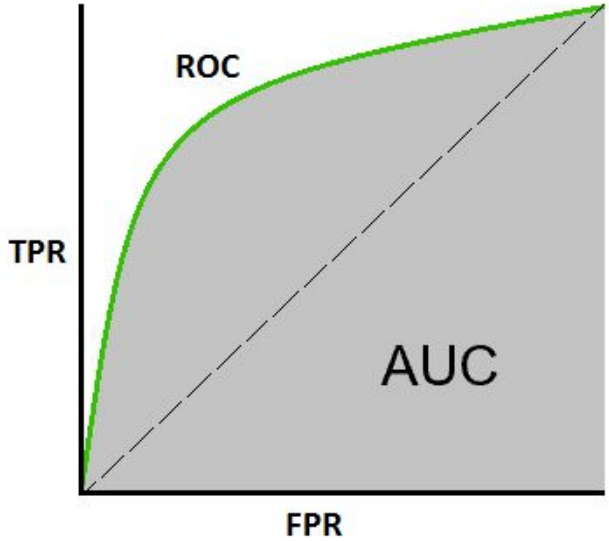
ROC-AUC

- A good metric for discrimination
- The Receiver Operating Characteristic is a graphical plot that can be used to assess how well a classifier distinguishes between two outcomes at different thresholds
- Illustrates the trade-off between **sensitivity** (recall, true positive rate) and **specificity** (true negative rate: TN / FP + TN)
- The false positive rate is 1 - true negative rate
- Trade-off between Type I and Type II error
- As TPR increases, so does FPR



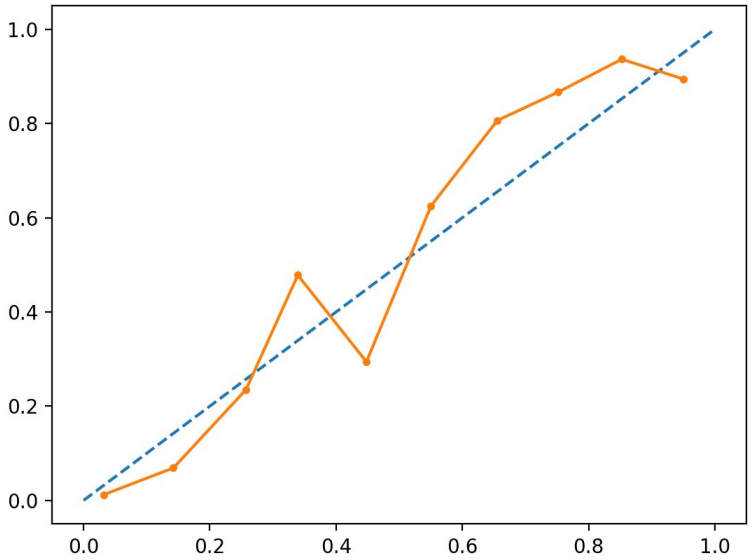
ROC-AUC

- Area under the curve gives us the probability (from 0 to 1) that a model will correctly distinguish between two classifier labels
- This metric is best if you:
  - Care about ranking
  - Care about distinguishing between positive and negative classes (e.g., distinguishing between Democrats and Republicans)
  - Sample is not heavily imbalanced



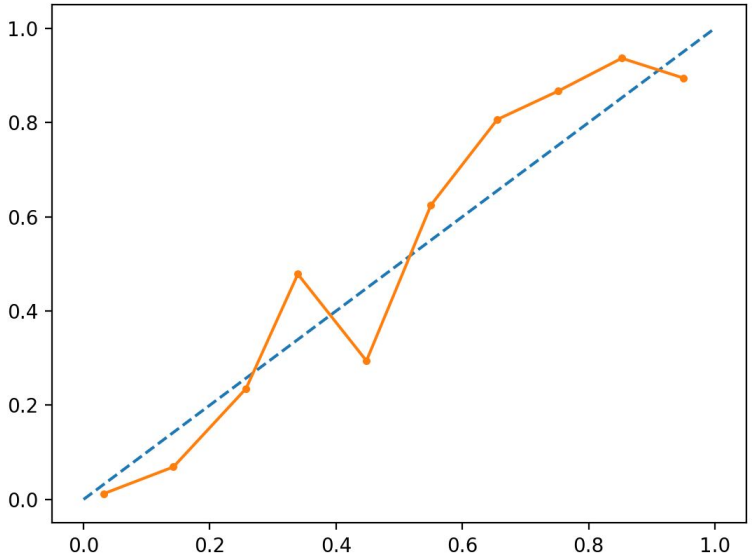
Calibration: Brier Score

- How well predicted probabilities fit to the actual probability of an outcome
- Calibration curve (seen right) is a graphical depiction of the fit of predicted to actual probabilities
- Brier score is essentially the average difference between our predicted probability and the observed outcome (similar to MSE)



Calibration: Log loss

- Similar to Brier score, but uses the log of probabilities to penalize large errors
- Useful if you're trying to fit closely to actual probabilities, like if you're trying to approximate the actual probability that someone will vote in order to make projections



Constructing your Validation Set

Constructing your validation set

- Choosing validation set(s)
  - What problem do you want to solve?
  - What structures in your data set could help you replicate the problem?
  - Where in your population are the greatest risks of error?

Train, Test, Validate

- When training a predictive model, it’s often desirable to have three different data sets
  - Training
  - Testing
  - Validation
- The purpose of the testing set is to ensure that your model does not overfit to the training set and therefore performs well on data out of sample
- If you model iteratively, you can also risk overfitting to your testing set
- **Always** have an independent validation set

Validation Set

- Important to think about your validation set
- May be a subset of the data you have on hand for training the model
- May be better to compare your model's performance to data collected through a separate data-generating process
- Examples:
  - Data collected independently that yields similar outcomes to your own data-generating process
  - Data collected in the field
  - Aggregate data

Independently-Collected Validation Sets

- Survey data collected independently that should yield similar outcomes to your own data-generating process
- Data collected in the field (field validation)
  - E.g., Rat Project, Collecting field IDs during voter outreach

Example: Independently-Collected Survey Data

- When developing a model of support for a policy issue like reproductive choice, one may compare model predictions to the responses to a survey not used in model training
- Responses may be to similar survey questions, or to different questions that we would be expect to be correlated with the response used for training
- For example:
  - Do you think abortion should be legal?
  - To what extent do you agree: Safe, effective, and affordable methods of abortion care should be available to women in their community?
- We would expect people who agree with the former are more likely to agree with the latter, so this could be a good validation.

Example: Field Validation

- Rodent inspection field validation
  - Randomly-selected 100 city blocks for inspection with a predicted probability over 0.5
  - Rodent Control inspected each block and recorded if they found rat burrows
  - Compared proportion of locations with rat burrows to predicted probabilities
- Phone quality score validation
  - Select ~100k phone numbers (10k / decile)
  - Collect phone dispositions (connected / disconnected, right / wrong person)
  - Compare phone dispositions to model predicted probabilities
- Automating pipelines to continuously validate models against incoming data

Cross-validation

- Simple train-test approaches are actually a special case of cross-validation
- Cross-validation usually involves splitting your training set into multiple cross-sections (often 3 to 5, sometimes more), holding out one cross-section and training the model on the rest
- This can also be difficult with a smaller data set

Cross-validation

- Cross-validation is especially powerful when you have natural cross-sections in your data that:
  - Replicate the kind of prediction you're trying to make, or
  - Represent subgroups in which your model may perform differently
- One good example is time-series cross-validation: dividing a data set into months or years and then predicting future outcomes based on models trained on past data
- Others may include geographic or cohort cross-validation

Validating against subgroups

- Models sometimes perform differently on subgroups in your data, and may perform better on some subgroups than others
- It is important to validate your model for subgroups where the model may perform differently
- Common disparities:
  - Differences in performance (the model performs better for some subgroups than others)
  - Differences in errors (the model tends to overpredict the outcome for some subgroups and underpredict it for others)

Validating against subgroups: Examples and risks

- Support for abortion and race (Differences in accuracy)
  - Black women tend to be more liberal and to support Democrats
  - However, Black women also show less support for abortion than other women
  - Many models of support for choice are poor at predicting support among Black women
  - If a person were to use such a model to mobilize women around reproductive rights, they may do a poor job identifying and mobilizing Black women who care about reproductive rights
- Recidivism (Differences in errors)
  - Recidivism is more common among Black people in part because of an unjust history of over-policing
  - Models tend to overpredict the likelihood that Black people will commit another crime and underpredict that white people will commit another crime
  - In other words, models tend to predict that a Black person will commit a crime when they don't and that a White person will not commit a crime when they do
  - If a person used such a model to make bail or sentencing decisions, they may overly penalize Black defendants compared to White defendants

Questions?