Good Data Science Should Be Validating

Peter Casey, Ph.D.
Director of Analytics
Catalist, LLC

A little about me...

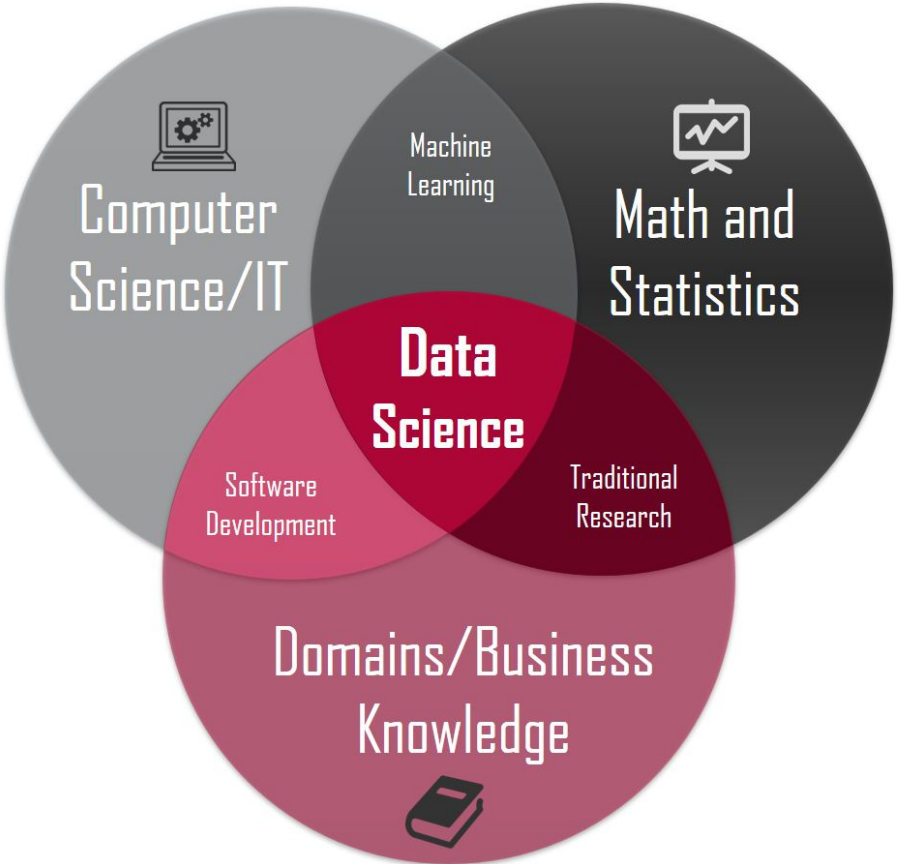
- PhD in Political Science from WashU
  - Failed Ryan Moore's first test
  - Passed the class with an A-
- Data Scientist for the DNC (2014-16)
  - Built 2016 turnout model
  - Built a bunch of demographic models
  - Survived Russian hack

A little about me...

- Senior Data Scientist at DC Govt
  - Office of the Chief Technology Officer (OCTO)
  - The Lab @ DC
  - Studied Rats
- Director of Analytics at Catalist
  - First org to collect and maintain national voter file (for the progressive movement)
  - Build predictive models, synthetics, analysis to support progressive campaigns

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



What does a Data Scientist do?

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 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
- Check that the model can be used for what you want to use if for

Validation Decisions

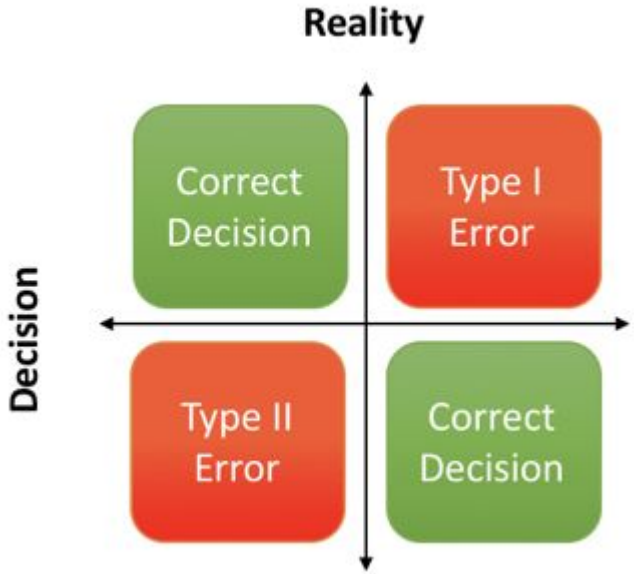
- Choose a performance metric
- Choose how you'll construct your validation set(s)
- Figure out what subpopulations you may want to validate

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

- True positives (TP): Correct positive predictions
- False positives (FP): Incorrect positive predictions
- True negatives (TN): Correct negative predictions
- False negatives (FN): Incorrect negative predictions



Precision

- The proportion of the model's positive classifications that are actually positive
- 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
- This is especially useful when cutting lists of locations to inspect or people to reach out to

When to use Precision @ N

1. Prioritizing locations to inspect
    a. Rodent infestations
    b. Housing code violations
2. Cutting lists for voter outreach
    a. Turnout / vote propensity
    b. Contact models, like phone quality

Recall

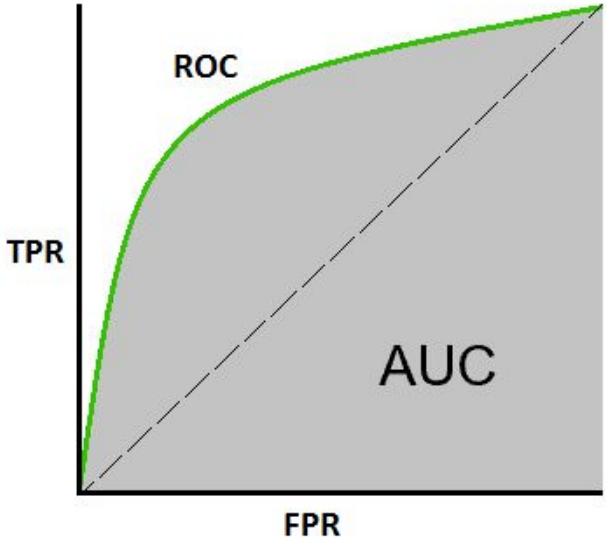
- **Recall** (also known as **sensitivity**) is the the proportion of actual positives that the model correctly identifies
- **True Positive Rate**: TP / (TP + FN)
- Useful when the cost of your intervention is low but the cost of missing a positive case is high (e.g., fraud)

F1-Score

- If the cost of your intervention is VERY low and the cost of missing a positive case is VERY high, then why use a predictive model at all?
- Usually there is some cost or trade-off to false positives, so we want to balance our recall against our precision
- The F1-Score is the harmonic mean of precision and recall
- Drawbacks:
  - Gives equal importance to precision and recall (different misclassifications may have different costs)
  - Measures precision and recall at a specific threshold (usually 0.5)

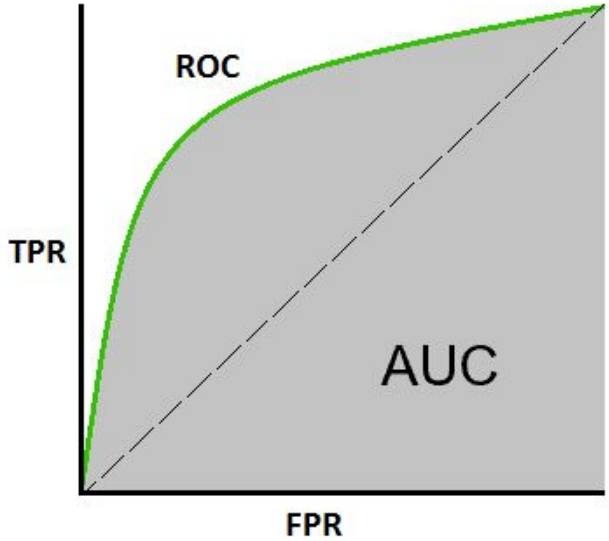
ROC-AUC

- 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** (false positive rate: FP / 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 when you want the model to distinguish between two possible outcomes (e.g., when you want to distinguish between Democrats and Republicans)



Calibration

- How well predicted probabilities fit to the actual probability of an outcome
- Brier score is essentially the average difference between our predicted probability and the observed outcome (similar to MSE)
- 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

- 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
- However, if you develop your model iteratively, you can also risk overfitting to your testing set
- Therefore, you should always have an independent validation set to compare your predictions to

Validation Set

- Important to think about your validation set
- May be a subset of the data you have on hand for training the model
- Better may be 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
- Data collected in the field (field validation)
  - E.g., Rat Project, Collecting field IDs during voter outreach
- Using survey data collected independently that should yield similar outcomes to your own data-generating process
- Validating against aggregate data

Example: Independently-Collected Validation Set

- 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

Example: Comparing to aggregate outcomes

- Collect data on precinct-level election outcomes
- Compare aggregate turnouts scores to turnout in that precinct during election
- Compare aggregate support scores of people who voted in that election to precinct-level outcomes

Cross-validation

- Carving your data into three different datasets can be difficult if you have limited data
- One option may be 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 accuracy (the model is more accurate 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
- 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 re-offend when they don't and that a white person will not re-offend when they do
  - A similar thing may happen in other models, like those predicting election turnout

Questions?