Catalist

Data Science for Campaigns

Peter Casey, Ph.D.
Director of Analytics

About Catalist

- Started in 2006 as the Data Warehouse
- First organization to collect and maintain a national voter file for political campaignins
- Serves progressive organizations and campaigns, including:
  - Labor unions
  - Large umbrella campaigns
  - Issue organizations
  - Voter registration organizations

Our Data

- The national voter file
  - Compiled from state and county voter files from across the country
  - Records of every registered voter in America going back to 2006
  - Contains information like name, address, voting district, and vote history
  - Also frequently includes information like sex, age, race, and party affiliation

- Combined with data from other sources, including:
  - Commercial data
  - Census data
  - License and occupation data
  - Survey data
  - Predictive scores

Catalist Analytics Team

- Composed of
  - Data Scientists
  - Data Engineers
  - Analysts
  - Analytics Fellows

- Use the Catalist data file to create
  - Predictive models
  - Reporting and analysis
  - Research

Predictive Models

- Civic behavior, like voting, donating, activism
- Political support for parties, issues, etc.
- Political identity, like partisanship, ideology
- Demographics, like age, race, ethnicity, religion
- Life events, like marital status, education, children at home

Vote Propensity

- Voter file includes whether a person voted or not (but not *how* they voted) in each election, including general, primary, special, etc.
- Use a person's vote history (linked across states and over time) to predict their likelihood of voting in a future election
- However, we're always using *past elections* to predict future elections
- Challenging: Who says 2020 will be like 2016? Was 2016 like 2012 or 2008? Was 2018 like 2014 or 2010?

Vote Choice History

- A special support model that predicts how a person voted in past elections by updating individual-level party support scores with precinct-level election outcomes
- Predictions for even even-year presidential, Congressional, Senate, and gubernatorial election from 2008 to 2018
- Use for two products:
  - **Vote Choice Index** – Combines individual voters' Vote Choice History scores to create index of their propensity to vote for Democratic candidates
  - **Vote Choice History Reports** – Geographic aggregates of individual Vote Choice History scores to provide better-than-exit-poll estimates of vote margins among different groups of voters

Probabilistic Race Model

- Don’t have race data for most registered voters
- Predict race by modeling survey responses
- Most race information people have from the voter file is modeled, but is presented as an assignment
- Probabilistic Race model:
  - Makes clear that race assignment is a prediction with error
  - Allows campaigns to cut lists that give them a sense of how many voters from each racial group they will talk to
  - Allows campaigns to aggregate geographically to get a better estimate of the racial make-up of voting districts

Catalist Major Challenges in Political Data Sci

- Effectiveness on models
- Bias in Machine Learning
- Digital Space

Effectiveness of Models

- Vote Propensity models are not the same as Mobilization models
- Support models are not the same as Persuasion models
- A person's likelihood of voting or supporting a candidate is not the same as their likelihood of doing that because a campaign contacted them
- To build Mobilization or Persuasion models, we need data from randomized controlled trials, which is costly and difficult to collect
- Need further research on the value of Vote Propensity and Support models for identifying voters campaigns can mobilize and persuade

Bias in Machine Learning

- Extensive research showing that model pick up trend in data that can lead age, racial, and gender bias, among others
- When decisions are informed by models that are biased, it can bias those decisions
- Especially concerning for progressive campaigns:
  - Models may be biased against underrepresented communities
  - Models may be biased against the people campaigns want to mobilize
- Unfortunately there has not been much research on bias in models frequently used by campaigns, like vote propensity
- Need for research and ideas about addressing bias, such as:
  - Leaving out data that correlates with characteristics like race (very difficult)
  - Train-then-mask
  - Training campaign workers to cut lists that are diverse and inclusive

Digital Space

- Campaigns are moving further and further into the digital space
- However, digital data and analysis has not been incorporated heavily into campaigning
- Digital data is difficult to match back to the voter file
- Need to think of way to leverage data and analysis in the digital space apart from the voter file
- Digital outreach alone could be valuable for fundraising, online actions (like petition signing), and even persuasion or mobilization

Catalist

Thanks!

There’s a lot left to do

Questions?

Ideas?