The Elusive ‘Likely Voter’: Improving Prediction of Who Will Vote

Anthony Rentsch, University of Massachusetts Amherst (arentsch@umass.edu)

Background

While many explanations have been offered for the “polling misses” of 2016, this research looks at likely voter models – tools used by pollsters to predict which survey respondents are most likely to make up the electorate and, thus, whose responses should be used to calculate election predictions. I evaluate a number of different likely voter models and provide recommendations on how to use them to effectively communicate election predictions under a range of different turnout estimates.

Data and Methods

I use CCES surveys from 2008, 2010, 2012, 2014, and 2016 to construct and evaluate a series of likely voter models using 2016 as a test case. This research is unique in its

1. Scope – simple and complex models at both national and state levels
2. Inclusion of structural election variables, such as presidential approval, state of the economy, incumbency, and polarization
3. Focus on communicating results in a probabilistic manner

Models

- Vote intent
- Vote intent + vote history
- Perry-Gallup index
    - Composite index of vote intent, vote history, political interest, voter registration status, and eligibility to vote in previous election
- Logistic regression
- Random forests

Results

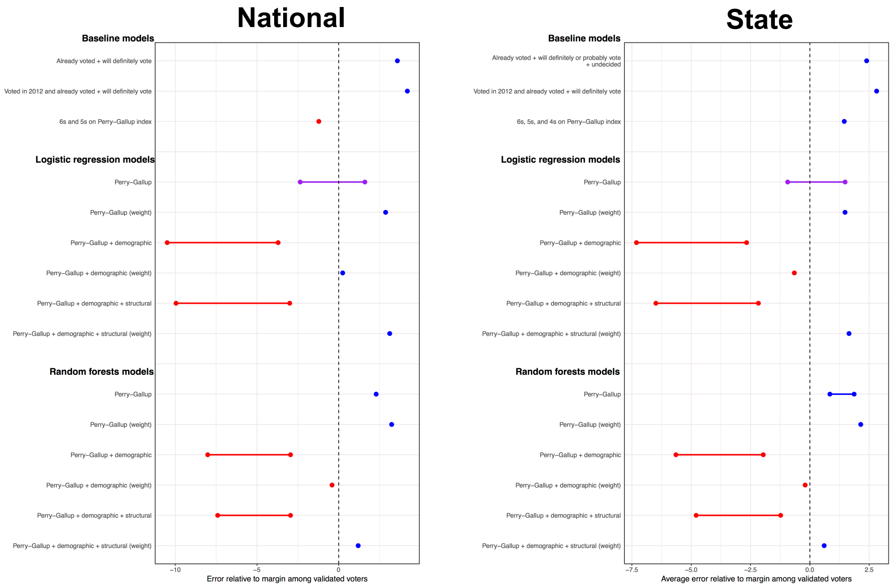

Figure 1. Comparison of errors for national-level (left) and state-level (right) likely voter models

Perry-Gallup index

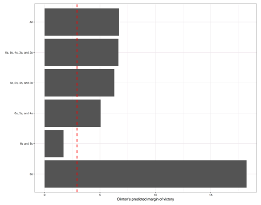

Figure 2. Comparison of national-level Perry-Gallup index models

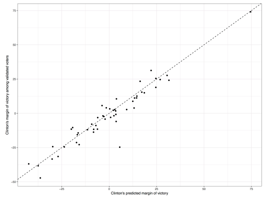

Figure 3. Predicted state margins for 6s and 5s on index against margin among validated voters

Logistic regression and random forests

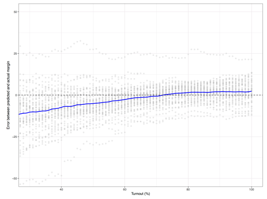

Figure 4. Average error between state-level predicted and actual margin among validated voters by turnout using random forests

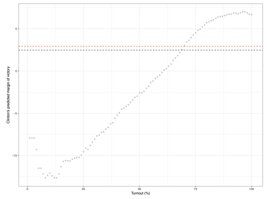

Figure 5. National margin by turnout using logistic regression (red dashed line is margin among validated voters and black line is vote-propensity-weighted margin among all respondents)

Possible implementation

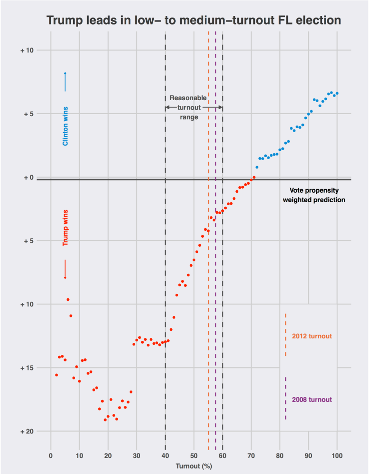

Figure 6. Mock visualization of a Florida poll using a random forests model trained on Perry-Gallup index and demographic variables

Conclusions

- ✓ Partitioning likely voters using Perry-Gallup index or weighting preferences using logistic regression- and random forests-based vote propensity scores lead to more accurate polling estimates
- ✓ Communicating results probabilistically using a range of turnout scenarios can add valuable information for poll consumers

Future Directions

- Regularization to identify key predictors for more parsimonious models
- Explicitly adjust turnout predictions for subgroups based on local demographics and trends in voter enthusiasm
- Update individual turnout estimates in-cycle using Bayesian framework