Challenge 1 — Iowa Caucus Predictions

Jeffrey Barrera & Jacob Fenton

January 31, 2016

Predictions

We predict XXX to be the winner of the 2016 Iowa Republican caucus.

Candidate	Vote Share
Donald Trump	0
Ted Cruz	0
Marco Rubio	0
Jeb Bush	0
Ben Carson	0
Chris Christie	0
Rand Paul	0
Mike Huckabee	0
John Kasich	0

Methodology

Features

In-state polling

We believe the best indication of a candidates' standing in Iowa is in-state polling conducted close to the caucus. Coming up with a weighted poll average is not trivial, and we considered several approaches: linear regression, lowess regression and non-parametric regression with gaussian and epanechnekov kernels.

As training data we used polls aggregated for Iowa by pollster.com (which was later acqured by The Huffington Post) for 2008 and 2012. (All of the code used in this project is available at: https://github.com/jeffbarrera/

iowa-caucus/). We wrote python scripts (2008 ; 2012/16) to standardize the data across years and add one key variable: the number of days before the Iowa Caucus is held.

We also wrote scripts to test several techniques under a variety of circumstances: lowess (results), and first-order non-parametric regression with gaussian (results) and epanechnikov (results) kernels. We tested each regression with a variety of bandwidths (or f values in the case of lowess) and compared the final estimated point with the actual polling result. The most accurate estimation result in terms of overall MSE for 2008 and 2012 was obtained using an epanechnikov kernel with a bandwidth of 13 days, which gave us an MSE of 7.29.

One additional complication is that prior years' polling included results the day of the caucus; the most recent poll results we have access to are several days ahead of the caucus. Thus we had to interpolate from a point several days ahead of the caucus to a final result; we tested several approaches and found the best to be TK TK.

National Polling

In their "polls-plus" model, 538 uses national polling as a contrarian indicator, based on data suggesting that candidates who poll better in a particular state than they do nationally tend to do better than their statewide polls. We adopted a similar approach, applying the WHICH technique we used to estimate statewide polling trends to national polling for each candidate.

Prediction Markets

TKTK

Maybe regress endorsements after all?

TKTK

Poll weights by pollster rating?

Hmm, the house effects are in terms of dem vs rep. Not sure how that works out here.

Features Not Included

We considered a number of additional features, but chose not to include them for various reasons:

Campaign Finance Data

Reports filed with the Federal Elections Commission give some insight into a candidates' fundraising and spending, but we chose to disregard these as not predictive of vote share for several reasons:

- This year is different! One candidate (guess who) has been the beneficiary of millions in "earned media" — coverage that's not paid for. Because he's been so effective in winning earned media, he hasn't sought contributions in the same manner as other candidates. Thus many of the usual governing assumptions (probably) don't hold.
- Candidates' reports are filed at a significant lag. Quarterly reports covering the fourth quarter of 2015 are due Jan. 31, but do not reflect any spending or fundraising that took place in 2016. Polling data is generally much more current than that.
- The 2012 and 2008 Iowa caucuses were held Jan. 2 (two days after the end of a filing period) whereas the 2016 caucuses are held Feb. 1 (a month after the end of the most recently available candidate spending data). Thus a relationship between financial figures for 2008 and 2012 wouldn't necessarily hold true for 2016.
- The way that campaigns spend money is in flux and increasingly money spent is excluded from public accounting. Increasingly spending from a candidates' principal campaign committee is overshadowed by money spent by independent expenditure only committees (aka super PACs). In 2012, super PACs primarly spent money on media buys, but increasingly these "outside" groups are taking on tasks previously handled by candidate committees, including last-minute voter targetting and mobilization. Other money spent by non-profit groups is not publically reported at all, and anecdotal reports suggest this type of spending is rising.

Endorsements

Fivethirtyeight uses a weighted endorsements system to help predict primary results. What their data show[footnote], however, is that there are far fewer endorsements this year than in previous cycles.

Crosstabs available in polls

Most reputable polls provide results cross-tabulated by various demographic groups. Unfortunately, we were unable to find any easily available aggregation of poll crosstabs (and the inconsistent approach pollsters take would make this a considerable challenge). Nonetheless, we believe this might be a useful indicator. Were this data available in bulk we might be able to make different assumptions about the electorate. Data suggest many potential voters who say they plan to participate in caucuses do not actually do so; we believe voter subgroups' lie to pollsters at a differential rate, introducting a meaningful bias into polls.

Model

Since we had a limited number of prior observations on which to build our model (the 13 main candidates who ran in the 2008 and 2012 Iowa Republican Caucuses), we used a LASSO regression to determine how to weight our features. We used R to calculate β coefficients at different values of λ, and then used leave-one-out cross-validation to see how each model performed out-of-sample. We selected the model with the lowest out-of-sample MSE, in the hope that this would be predictive in 2016 but would also avoid overfitting to the 2008 and 2012 data. We then used this model to predict vote shares for each of the 2016 candidates. Finally, since the sum of these predictions likely would not exactly equal 100%, we scaled them proportionally to produce our estimates of vote share.

Limitations

The biggest limitation of our model is that it does not include any features intended to measure voter turnout or the "ground game" — the campaigns' efforts to identify supporters and get them to show up to caucus. Historically, this has been a critical aspect of winning in Iowa **CITE**: since caucuses typically have low turnout rates, effectively mobilizing supporters can have a big impact on a candidate's vote share. However, we were unable to find a good way to measure organizing operations or predict turnout. Campaign finance data could have shed some light on the resources campaigns have on the ground, but this data is unavailable for the critical three weeks immediately before the caucuses. Crosstabs may have told us more about turnout intentions among different subgroups of voters, but we were unable to obtain this data in machine-readable formats at the scale necessary to be a valuable predictor.

Ultimately, we had to rely on assumptions about the relationship between polling and turnout. Since most of the polls we're aggregating are of "likely voters," we hope that the polling firms are at least somewhat

accounting for projected turnout when they weight responses in their polls. We then hope that the relationship between these polls of likely voters and the final vote share is reasonably consistent across elections, and thus that the coefficients in our LASSO model will capture this relationship. These are both admittedly rather tenuous assumptions, but in the absence of better data will have to suffice.