Active Polling: Improving Presidential Election Predictions by Targeting Polls with Active Learning

Raymond Gilmartin, Sean Roelofs
CS 229, Spring 2019, Stanford University

Background: Polling

Current Polling is Insufficient

- Most publicly-available polls come from large survey research companies, (e.g., Pew and Gallup) which release election polls for free to promote their brand [1].
- Data given away for free can't be targeted, so publicly-available polls are taken at state- or national-level. Geographically-targeted polls are informative, but expensive. They are only available to political parties and well-funded campaigns.

Active Polling May Offer Improvements

- Once a survey research company polls a county, active learning could help choose which counties should be targeted next for polling. Not all counties are identical, so pollsters should choose counties which are expected to be informative.
- Once pollsters have exhausted their funds and cannot conduct any more polls, machine learning can be used to predict results for counties that weren't polled.

Data: Predictors from Publicly-Available Information

Sources

- Ideally, data for all predictors should be free and publicly available.
- **Demographic data:** Aggregated 2015 U.S. Census American Community Survey [2].
- **Historical election data:** Dataset from the Harvard Dataverse, aggregates county-level election returns for all states and presidential elections from 2000-2016 [3].

Features

- **Demographic data:** General demographics include total population and breakdown of population by sex and race. Economic demographics include median income, poverty rate, unemployment rate, and employment by industry.
- **Historical election data:** Consists of Democratic and Republican presidential candidate vote counts for each county and historical election year.
- Where applicable, some features were removed to avoid heteroskedasticity.

Data: Response From Simulated Polls

- Studies on polling have shown that there is justification for simulating polls results as a normal distribution around the true election results [4].
- So, for each true election result, $y^{(i)}$, we have the first element, $y_1^{(i)}$, equal to the percentage of the county's residents who voted in that year's presidential election and the second element, $y_2^{(i)}$, equal to the percentage of voters in the county who voted for the Democratic Party's candidate in that year's election.
- Simulate polling data as: $p_1^{(i)} \sim y_1^{(i)} + N(0,\ 0.04)$ and $p_2^{(i)} \sim y_2^{(i)} + N(0,\ 0.02)$.
- Using 2 percentage points as the variance for the Democratic candidate's polling share is motivated by the aforementioned literature on election polling. Using 4 percentage points as the variance for the turnout is the educated guess we use because we wish to be conservative in our polling estimates—guessing a higher amount of variance than truly exists rather than a lower amount.

Models

- The active learning algorithm will proceed by alternately fitting a model based on the counties for which it has polling data, then using a querying method to choose which counties to poll next.
- 5 querying methods and 2 model types were tested, in various combinations.

Querying Methods for Active Learning

- **Random selection:** Choosing random counties to poll next, used as the baseline.
- **Selection by committee:** Train "student" models on random subsets of counties polled so far. Poll counties for which the students' predictions have highest variance.
- **K-means:** Fit a k-means clustering model to the predictors and poll one county from each cluster to ensure that the algorithm polls a diverse sample of counties.
- **Entropy selection:** For a random forest, use each tree to predict the response for each county. Poll the counties with the highest variance among those predictions.
- **Committee with K-means initialization:** Choose the initial sample using the k-means approach outlined above, then proceed by using committee selection.

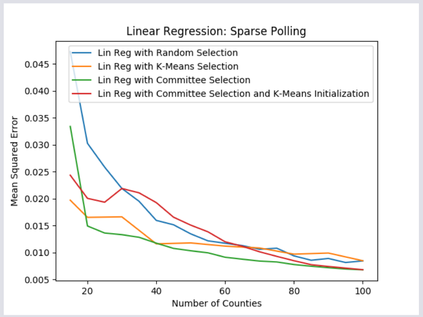
Model 1: Linear Regression

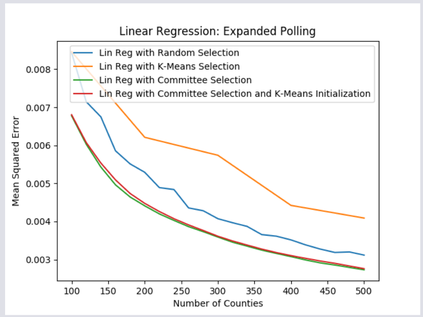
- Advantages: Easy to fit. Easy to interpret the way predictors affect election results.
- Disadvantages: Not flexible. May predict values outside of [0, 1] for percentages. Can be fixed by changing predictions greater than 1 or less than 0 to be 1 or 0, respectively.

Model 2: Random Forests

- Advantages: Random forests are flexible. Won't predict values outside of [0, 1].
- Disadvantages: More computationally intense to fit than linear regression model. Has no coefficients, so interpreting effect of predictors on election results may be difficult.

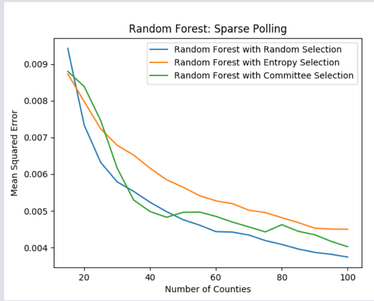
Results and Discussion: Linear Regression

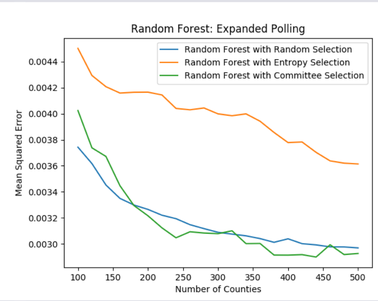




- "Sparse Polling": K-means and committee selection perform better than random selection.
- "Expanded Polling": Committee selection and committee selection with k-means initialization perform better than random selection.
- Improvements from committee selection are consistent as more counties are polled, since polling high-variance counties continues to improve prediction as more counties are polled.
- Improvements from k-means are limited to the "Sparse Polling" range. As more counties are polled, it is less important to explore a diverse array of counties and more important to improve predictions in the counties that are most important to the model's overall performance.

Results and Discussion: Random Forest





- "Sparse Polling": Entropy and committee selection perform better than random selection only when the number of counties chosen is 15, corresponding to 10 initial random points and 5 points chosen using the querying method.
- "Expanded Polling": Committee selection performs slightly better than random selection, when 200 or more counties are polled, which would likely be prohibitively expensive.
- Selection methods may be failing with random forests because the forest model itself relies on random sampling to fit the data. The forest may perform poorly because the polled counties don't represent a random sample of the data.

Conclusions and Future Work

- Active learning seems to provide benefits when using linear regression models, which may be used in practice because of their ability to clearly express relationships between predictors and responses. If explaining relationships in elections is the goal, this approach to polling might be useful.
- However, if the goal is prediction, random forests dramatically outperform linear regression. This can be observed by comparing the mean squared errors across the graphs for random forests and linear regression. And when using random forests, the querying methods tested here did not improve predictions over random polling.
- The most promising possibility for future work is to apply a wider array of querying methods to this problem. It is possible that some methods could improve predictions from the random forest.
- Another possibility for future work is to consider elections as a classification problem. If the goal is to predict a winner or loser in each county classification models could be used instead of regression models. Classification models offer clearer ways to estimate variance and uncertainty among unlabeled points, making querying methods easier to apply.

References

[1] M. Fahey and E. Chemi, "Most Election Pollsters Aren't Really in it for the Money," *CNBC*, para. 3, October 18, 2016. [Online], Available: https://www.cnbc.com/2016/10/18/most-election-pollsters-are-not-really-in-it-for-the-money.html [Accessed May 24, 2019].

[2] Kaggle user MuonNeutrino, "US Census Demographic Data," *kaggle.com*, March, 2019. [Online]. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data. [Accessed May 16, 2019].

[3] MIT Election Science Data Lab, "County Presidential Election Returns," *Harvard Dataverse*, October 11, 2018. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ. [Accessed May 16, 2019].

[4] H. Shirani-Mehr, D. Rothschild, S. Goel, and A. Gelman, "Disentangling Bias and Variance in Election Polls," *Journal of the American Statistical Association*, vol. 113, no. 522, Feb., pp. 607-614, 2018.