Forecasting US Elections

Jerry Lin, Nicholas Lorio, Aadit Narayanaswamy, Junseo Park, Brandon Scolieri, Cong Yang

Background

We became interested in election forecasting after briefly getting acquainted with a legal election betting market called PredictIt. We believed that prices on the site for various races were influenced mostly by the opinions of speculators who were not approaching buying and selling bets from a quantitative perspective, and we wanted to create a rigorous model that gave us the ability to make bets that paid off at a higher average rate than the bets of the median "speculator."

While organizations like FiveThirtyEight already approached forecasting from a quantitative perspective, we did think there was room for improvement for a couple reasons. For one, because PredictIt is formally classified as an academic research project at Victoria University, they were only willing to share their data with other academic institutions and their affiliates. This meant that we could conceivably train our models on data that FiveThirtyEight simply was not allowed access to. Second, a quick glance at the research literature available gave us the impression that the use of machine learning models in election forecasting was scant. Third, we were initially quite ambitious in that we also wanted to incorporate information from Twitter and local news sources.

Ultimately, the only data sources we felt we could rely on were polling and FEC data that we acquired independently.

Project Goal

The goal of our project is to use polling and FEC data to predict winners of congressional and gubernatorial races in the United States. We plan to use this data to provide campaign donors with a mechanism that will manage their donation in a way that will have the most impact for the candidate of the donor's choice. Users can also make convenient direct campaign donations through the site.

Hypothesis

Using machine learning techniques on polling and FEC data, we can outperform betting market prices on PredictIt as well as the forecasts from professional news organizations like FiveThirtyEight.

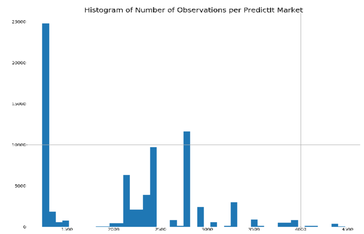
Data Acquisition

Initially, we wanted to use data from Twitter, news sites local to respective races, and PredictIt, but we quickly gained the impression that acquiring the necessary data from Twitter and local news going back the years we wanted was more effort than it was worth. As for PredictIt, the quality of the data we received from our partnership was very poor. There were less than 200 unique races, which was dangerous to generalize from. Additionally, there was unacceptable variation in the number of observations we had for each race. Some races had as few as ten observations while others had thousands.

When it came to creating a poll database, we were very surprised to learn there were no publicly available databases of polls for congressional and gubernatorial races. We reached out to FiveThirtyEight to see if they would be willing to share their polling data, and we were told in response:
We took their claim that the only way to create a polling database was to collect data by hand and their refusal to share their data as a challenge, and we decided to build a polling database from scratch ourselves. We did this by scraping archived versions of https://www.realclearpolitics.com/epolls/latest_polls/, which we found using wayback machine. When it came to matching winners, we were able to scrape those from wikipedia. Thankfully, there was a page corresponding to every year we were interested in, so there was no need to use an internet archive. That being said, sometimes formatting was inconsistent, and we were forced to edit the wikipedia pages that did not adhere to a standardized formatting amenable to scraping.

> "Unfortunately, no can do -- the polling database is an incredibly valuable piece of intellectual property. Hopefully you understand; good luck with your project!"
> - FiveThirtyEight Employee

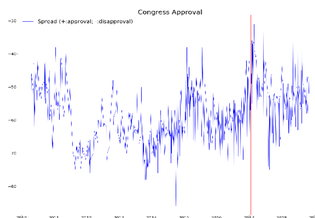
As for the FEC data, they had a very convenient API which was unfortunately limited to 1000 calls an hour, with each call containing 100 results per page. We used Amazon Web Services to run code for several days that slowly but surely built us a database for all the races we were interested in dating back to 2010.

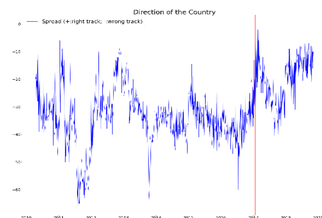


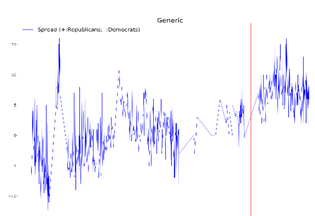
Assumptions

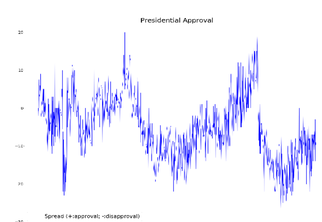
- The most recent poll in any given race is not necessarily a definitive measure of voter sentiment because of sampling error.
- A sizable portion of voters decide whom they are voting for months in advance, and much of the variation in polling data over time for an individual race might be more of a result of sampling error than a legitimate change in sentiment.
- Polls offered on realclearpolitics.com are the most accurate polls available.
- There are very few, if any, errors on wikipedia pages detailing results of elections.
- Political donations that are not reported to the FEC do not represent a sizable percentage of all political donations.

Exploratory Data Analysis





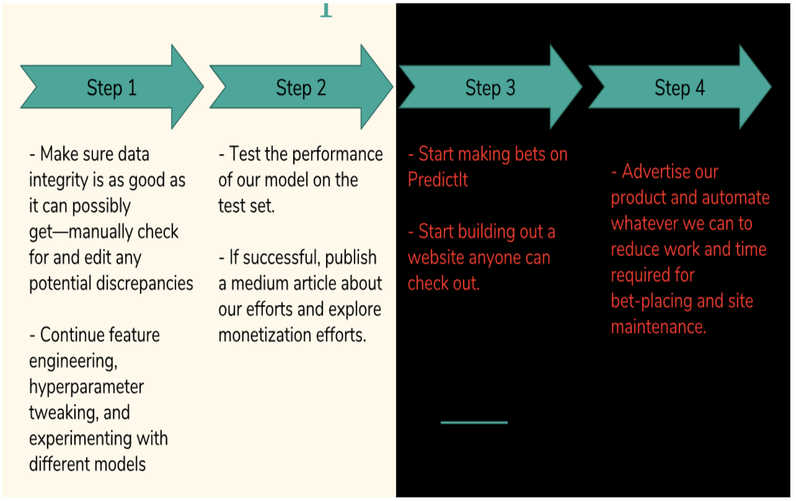




Result

We use four different machine learning methods, including logistics regression, perceptron, adaptive boosting, and random forest, to train the models to predict the U.S. election based on the polling information and FEC data. For the model trained by logistic regression, we end up getting an accuracy around 81.35% and 78.69% on training and testing set respectively. However, for the model trained by perceptron, we only get an accuracy around 38.52% for both training and testing set. The model built by adaptive boosting comes up with an accuracy of 96.11% on training set and 90.16% on testing set. Lastly, the random forest model has the best overall performance which is about 99.38% on the training set and 95.08% on the testing set. Based on the performance of our models, we are confident to conclude that our models could be used to accurately predict the winner of the U.S. election.

Next Steps



User Interface