CS 109A: Final Project
Protopapas, Rader, Tanner

Predicting the 2020 Election

Problem statement

The General Election, to be held on November 3, 2020, will shape the future of the country. The 2016 Presidential election was poorly predicted by the experts, which was due to many contributing factors (see The Atlantic). In hindsight, incorporating more uncertainty in the prediction modeling (and more dependency between voting districts) would have prevented such poor performance. But can these known issues from the 2016 election be incorporated in order to improve forecasting in 2020, or will history just repeat itself? And will voter rights, COVID, and social issues impact voter turnout and thus the election?

Project goal: The primary goal of this project is to predict the outcomes of the 2020 Election: predicting the results of the Presidential election all federal House of Representative election races in November (Senate seats are a bonus) using data from before the election as features, to mimic a true forecasting model. Any piece of information from before election day can be used to perform this prediction, but the suggested feature set should include past elections, polling data, and demographic information of congressional districts (see below).

For an example prediction model for 2020, see fivethirtyeight.

Data Recources

1. **Election Results**
   There are a few sources of recent election results: 2016 Presidential Election on Wiki 2018 House Election on Wiki, FEC's 2018 Report, Archived Election Results since 1982.

2. **Polling Data**
   Recent (and a little historical) polling data can be found at Real Clear Politics and Fivethirtyeight.

3. **Congressional District Data**
   Demographic data on Congressional districts can be found on Census.gov, and geospatial data are found on Data.gov.

4. **Other Useful Data**
   MIT's Election Data Repository. Fivethirtyeight's Data Repository Voting Laws.

High-level project goals

1. Obtain publicly available data from various public sources (some scraping will definitely be required).

2. Build a predictive model for the Presidential and House election in November (not using the actual 2020 election results).

3. Use the 2020 results to determine why and when/where the predictive model was right or wrong.
4. Determine which of the polls got it right or wrong, and why.

References

1. COVID-19 and the 2020 Election
2. Voter turnout in 2020
3. The changing composition of the electorate
4. What went wrong with the 2016 election