Stat Learning Group 3 Project Proposal

Shisham Adhikari, Grayson White, Maggie Slein

2 November 2020

Background

In the last two decades, the polarization of American politics has yielded unexpected victories by minority political groups, both in the 2000 and 2016 election. The Republican party represents less Americans by volume than both the Democratic party and the Independent parties, yet has continued to secure political power as a result of the electoral college. The two unprecedented victories in 2000 and 2016 by candidates who lost the popular vote but won the presidency as a result of the electoral college have been attributed to highly variable party affiliation in what have been dubbed "swing states". In these states, the split of Democratic to Republican votes is fairly even, which makes winning their popular votes crucial for ultimately winning their electoral votes, in the electoral college's "winner takes all" framework. As these highly influential and dynamic "swing states" shift with each passing election, a better of understanding and prediction of their political party tendencies becomes increasing important. To better understand how external factors drive political party affiliation and ultimately predict political party affiliation, we are interested using techniques that combine many model types to provide more accurate conclusions.

Research question

How can we use a combination of models to predict political party affiliation using small sample sizes?

Our group wishes to understand the external factors related to political party affiilation, and to understand how well we can predict political party affiliation by using a combination of the modeling techniques learned in class.

Description of data type

To answer our question, we aim to use datasets with a large sample size and several predictors (though n > p preferably). The data set would include political party affiliation (Democrat, Republican, Independent, etc.) and several predictors (age, gender, socio-economic status, location, etc.)

Candidate data sets

Our first choice dataset is the GSS (General Social Survey) dataset that features political party affilation with several predictors (6110 to be exact) that include marital states, age when married, level of education, age, as well as a subset with relevant predictors (college degree, age, political party affiliation, income, etc.) for to answer our research question.

Our second dataset of choice comes from the "politicaldata" package in R: "us_pres_polls_history", which features election polling data from 1980 through the 2016 election about party popularity over the course of the last 36 years.

Potential obstacles

We have already run into issues with finding the right data to answer our question. The first dataset features data that truly address our question, while the second dataset really only tracks the increasing divide and polarization of the Democratic and Republican parties. It may be hard to find dense datasets that tie political party affiliation to categorical attributes in a tidy way.

We could run into issues with the computational complexity of our approach and of our datasets, which could lead to an inconclusive result. Stacking these different models with a wide variety of predictors may not reveal a conclusive result.