Matthew Burgess
Capstone Project 1
January 20, 2020

Data Science Capstone 1: Reliability of US Election Polls

Problem Statement:

In Presidential, Senatorial, and Congressional elections across the United States, all campaigns and media outlets rely on polling. Exit polls help determine decisions on election day. Polls of Iowa in presidential primaries help determine the front-runner for a given race. But what makes a poll reliable? I hope to use polling data from the past ~20 years in order to build a model that can help predict whether a poll is a reliable judge of the outcome of an election.

Data Set:

I plan on using data from the data news site FiveThirtyEight on polls. It is roughly 10000 rows by 25 columns from polls since 1998. Please see the link here to get to the CSV file. It not only includes data of polls of race by party, but also the year, partisan lean of both the pollster and the district, and real percentage of votes obtained by a candidate.

Application:

While there is no real client for this, being able to estimate the reliability of polls is something that is done commonly. Whether or not a pollster can be trusted carries massive consequences for the consumer of that poll, the campaigns participating in the poll, and the pollster themselves. The famous Selzer & Co. Poll with the Des Moines Register is a poll many people look to due to its notoriety. However, as data science continues to enter into the political arena, the reliability of a poll can help us narrow down the reliability of polls even in local congressional races.

Outcomes:

The main outcomes from this project are to have a model that can tell if a poll is a reliable source of information. This information can then be used by campaigns or media to legitimize or delegitimize a pollster's influence, and determine the reliability of polls.