Capstone Project 1 Proposal

Polling data analysis and extrapolation

Problem Statement: Tens of millions of Americans vote for their House and Senate representatives every two years and their president every four years. The last presidential election saw 128 million people cast their votes. Polling organizations like Gallup, the Pew Research Center, and YouGov attempt to poll America and determine in advance which way these elections will go. Due to the heterogenous nature of the country and cost of conducting wide, far-reaching polls, poll quality can vary greatly. This project seeks to examine a broad slice of polling data against election results to determine which polls are trustworthy and, more importantly, what makes a poll trustworthy.

Who Might Care: Anyone who follows polling groups would be interested in knowing which ones are highly rated, but polling organizations could use this data to efficiently improve their polling methods. For example, if sample size is shown to be the most accurate indicator of a poll's accuracy, then they can allocate resources more intelligently. This polling model would also be of interest to political candidates: knowing which polls are more accurate allows them to make more informed decisions about which areas to campaign in.

Dataset: The website FiveThirtyEight collects and analyzes polling data from hundreds of different pollsters. This data is publicly available via a GitHub page and can be compared against final election results. To work with recent data in an acceptable scope, this project will focus on data from the 2018 midterms and attempt to assess the effect of factors such as poll size and time before election on poll accuracy. If scope needs to be increased, data from prior election years can be added.

Problem: How can we separate good polling from bad polling to determine which polls of the American public accurately reflect how they will vote in a coming election? Furthermore, how can we improve how we conduct polling to maximize accuracy and precision while minimizing costs? Is it more cost-efficient to broaden the scope of a poll and contact more people, or focus on specific types of polling over others, such as door-to-door polling versus telephone polls? Can we quantify the effect of time before election on polling accuracy?

Outcomes: The goal of this project is to deliver a model that predicts the accuracy of a poll based on factors such as: Time before election, number of people surveyed, type of poll conducted, type of election, and geographical location. This will allow for an assessment of poll quality without needing to compare the poll to the final results. It will also facilitate improvements in polling by clearly predicting how altering survey methodology affects poll accuracy. To accomplish this, the model will need to isolate the effect of each of these factors on overall poll accuracy. The number of different factors may make it difficult to fully isolate specific variables. It may be valuable to test multiple models and select the best.

Known issues: Not all relevant variables are captured by the FiveThirtyEight dataset. Each polling organization conducts polls differently – therefore, there may be significant differences in polling methodology that this model can only track by pollster identity, not quantify. Expanding the scope of this project to track individual pollsters may resolve this issue, but it is likely that each individual pollster will not provide enough data to meaningfully assess their accuracy.