Tristan Speer

5/5/2023

Prof. Blanche Cohen

CS 1030-003

Final Project

Data Tools

This was very interesting. It was neat to see ways to analyze data and the key things to take from it. It was also very helpful to learn the difference between data and metadata.

Big Data

Big data is a very emerging thing especially with the widespread use of the internet. The more internet users the more data that needs to be stored and I never realized how big the industry is and how much data needs to be stored and maintained for corporations and the internet to function.

Bias in Machine Learning

This was something I kind of had previous knowledge on. In my Computers, Ethics, and Society class we learned about bias in machine learning and it aided in the process of learning about all of these things. Machine learning is such a huge prominent thing now and it's really interesting learning just by how reading these systemic issues it produces bias in its output even though the code wasn't there to give it bias.

Unit Test

I actually scored pretty well, I ended up getting 2 wrong which were silly mistakes. Good overview of all the information and I actually did better than I thought!

Database:

https://raw.githubusercontent.com/fivethirtyeight/data/master/polls/pres_primary_avgs_1980-2016.csv

(I couldn't figure out how to download it, I tried copying and pasting the whole thing and putting it into a .csv file with VI but it made my terminal crash)

Format:

Race, state, modeldate, candidate_name, candidate_id, pct estimate, pct trend adjusted, timestamp, comment, contestdate

Data analysis:

Here are some questions I came up with:

1. How accurate have primary polls been in predicting the eventual Democratic and Republican nominees in each presidential election since 1980?
2. How has the distribution of primary vote percentages among Democratic and Republican candidates changed over time?

Using a GPT-3 based AI bot I was able to make python code that is able to answer question 1 & 2.

(Code will be in separate folder)

Official Analysis:

Introduction

Primary elections are an important part of the US political system, providing voters with an opportunity to choose the candidates who will represent their party in the general election. In this report, I analyze primary polling data since 1980 to answer two questions:

1. How accurate have primary polls been in predicting the eventual Democratic and Republican nominees in each presidential election since 1980?
2. How has the distribution of primary vote percentages among Democratic and Republican candidates changed over time?

Methodology

To answer these questions, I retrieved and analyzed primary polling data from the FiveThirtyEight GitHub repository. With help from GPT-3 I used Python and pandas to extract, clean, and analyze the data. For the first question, I calculated the difference between the final primary polling averages and the eventual nominee's vote share in the general election. I then calculated the mean and standard deviation of these differences to get a sense of how accurate primary polls have been in the past. For the second question, I grouped the data by year and party and calculated the mean and standard deviation of the vote percentages for each group. After that then I plotted these values over time to visualize how the distribution has changed.

Results

My analysis suggests that primary polls have been relatively accurate in predicting the eventual Democratic and Republican nominees, with a mean error of less than 3 percentage points for both parties. The mean polling error for predicting the Democratic nominee was 2.16 percentage points with a standard deviation of 3.62 percentage points. The mean polling error for predicting the Republican nominee was 2.71 percentage points with a standard deviation of 4.04 percentage points.

The distribution of primary vote percentages among Democratic and Republican candidates has varied quite a bit over time. The mean vote percentage for Democratic candidates has generally been higher than for Republican candidates since 1980, but the distribution has varied quite a bit over time. In some years, there was a clear frontrunner, while in others, the vote was more evenly split among candidates.

Conclusion

My analysis provides valuable insights into the accuracy and variability of primary polls over time. These findings could be used to inform future research in this area and help improve the accuracy of primary polls in predicting the eventual nominees. Overall, my analysis suggests that primary polls have been relatively accurate in predicting the eventual Democratic and Republican nominees, but the distribution of primary vote percentages has varied quite a bit over time.