US Election 2024*

Predicting election results

Robert Ford Michelle Ji Cher Ning

October 22, 2024

Compiled polling surveys regarding the 2024 United States presidential election were analyzed in this report based on data from FiveThirtyEight (FiveThirtyEight 2024b). From the analysis, bayesian and generalized linear models were built to forecast the outcome of the presidential election based on popular vote and the electoral college system. The models forecast Kamala Harris as the winner of the US presidency. Election forecasting is crucial for informing voters, guiding campaign strategies, and engaging the public in the electoral process. Additionally, it helps predict potential policy shifts and holds political institutions accountable, contributing to a more transparent and informed democracy.

1 Introduction

Every four years, the United States presidential election occurs and is one of the most significant political events internationally. The complex process of the election involves both the popular and electoral college system, in which voters cast ballots on election day in each state. The candidate that receives the popular vote by the voters wins the electoral votes for that state (Weber 2024). After the popular vote is accounted for, the candidate who receives the majority of the electoral votes from all states wins the presidential election. The number of electoral votes varies by state and is proportional to each state's respective population (Weber 2024). Specifically for the 2024 presidential election, the top candidates include Donald Trump and Kamala Harris, both with differing opinions on key topics such as immigration, technology, abortion, and transatlantic affairs. For example, regarding technology, both parties are concerned with the rapid progression of artificial intelligence's capabilities but differ in the response and regulation of it. Republicans believe that the current practices are unfair and disproportionately target their voices, whereas democrats generally are in favor of tighter regulation (NBC, n.d.). Another key issue important to voters and the candidates is abortion. As former president, he supported restrictions on abortion access and advocated for policies that align with the conservative movement's opposition to abortion rights (McKenzie Beard and Abbie Cheeseman and Justine McDaniel 2024). On the other hand, Harris has been a key supporter in the pro-choice movement and has been an advocate for protecting and expanding access to abortion services, as well as female autonomy (Chad de Guzman and Koh Ewe 2024). The outcome of this election will have profound consequences for various legislative policies and the future direction of the United States.

To forecast the 2024 United States presidential election, a generalized linear model and bayesian model were built. Our goal is to determine who will win the election based on the electoral college system. A key finding from our model is that we predict Kamala Harris to win the presidential election. This would be a historic win, as she would be the first woman to become US president. This would represent a significant

[*] Code and data are available at: https://github.com/Ford-Robert/us-presidential-election.

milestone in breaking barriers related to gender and racial representation in the highest levels of American politics, symbolizing progress toward a more inclusive and diverse political landscape. One drawback of our model stems from representativeness of the dataset, as it does not include all 50 states in the polls, which may impact the electoral college votes.

This paper is broken down into various sections, including Data, Modeling, Results, Discussion, Conclusion, and Appendices. This paper uses data made available by FiveThirtyEight (FiveThirtyEight 2024b) which compiles individual polling surveys based on state and methodology. Section 2 explores the data, highlighting key aspects that may be useful to future policymakers or campaign strategists, as well as details the variables present in the dataset. Section 3 presents the models that were built and used to forecast the election. Section 4 details the conclusions of our model and Section 5 explores possible implications and insights of our conclusions. Section A include an idealized methodology and survey that we could hypothetically run, with the task of forecasting the US presidential election. This section also includes an in-depth analysis of one specific pollster's methodology found within the larger dataset.

2 Data

The dataset used was obtained through FiveThirtyEight Interactives (FiveThirtyEight 2024b) and is focused on the United States presidential general polls for 2024. The polling data extracted are polling averages, where FiveThirtyEight (FiveThirtyEight 2024a) includes individual polling data from all publicly available polls that meet their ethical and methodological standards and at minimum, must test Harris versus Trump, the two primary presidential nominees for each political party. From the individual polls, FiveThirtyEight (FiveThirtyEight 2024a) will reweight and adjust the polls based on a few criteria, such as poll sample size, how recent it is, and house effects. The individual poll data is then added to the dataset, which is accessible for us to use. This dataset is crucial to help us build a model to forecast the 2024 US Presidential election and can help us analyze key features in voter habits for future political analysis.

There were two other potential datasets that could have been used, titled "Presidential Polling Averages" and "President Primary Polls", both found through FiveThirtyEight (FiveThirtyEight 2024a). The "Presidential Polling Averages" dataset only included data from the Biden versus Trump election in 2020 and was not updated to include data for the upcoming 2024 election and therefore inadequate. The "Presidential Primary Polls" dataset included very similar information as our chosen dataset, however, it included candidates that have since dropped out or were not major running candidates, such as Nikki Haley and Michelle Obama. It includes observations that are unnecessary and creates noise, therefore we did not choose this dataset either.

The variables used in the dataset include pollster name, pollster weight, method, state, sample size, pollscore, collection/end date, transparency score, candidate, support percentage, election date, days to election, initial weight, weight, and poll region. Pollster Name indicates which pollster the individual poll came from, which varies from individual companies, such as YouGov, to research institutions, such as Quinnipiac University and the weight is assigned by FiveThirtyEight depending on sample size and if they have multiple polls out in a short time. This is to adjust for any bias. Method is the way the poll was deployed, including but not limited to live phone, online panel, and text-to-web. State indicates which state the poll was conducted. State indicates what state the poll was conducted and sample size is the size of the poll. The pollscore is the rating FiveThirtyEight gives each indidivual poll, on factors such on ethical and methodological elements.The candidate variable indicates the candidate a certain percentage of people in the poll voted for and "support percentage" represents that specific number. The election date and days to election variables indicate how far out the poll was conducted with respect with the election date of November 5, 2024. We also reorganized the dataset to include a poll region, which divides the US into geographical regions such as rust belt, northeast, etc.

Table 1: First few observations of cleaned data set

pollster	numeric_grade	pollscore	end_date	transparency_score	question_id	method
St. Anselm	2.4	-0.5	2024-10-02	6	211550	Online Panel
St. Anselm	2.4	-0.5	2024-10-02	6	211550	Online Panel
St. Anselm	2.4	-0.5	2024-10-02	6	211550	Online Panel
St. Anselm	2.4	-0.5	2024-10-02	6	211550	Online Panel
St. Anselm	2.4	-0.5	2024-10-02	6	211550	Online Panel

state	sample_size	candidate	support	election_date	days_to_election	initial_weight	weight	pol_region
New Hampshire	2104	Harris	51	2024-11-05	34	0.8944272	1.041263	NEngland
New Hampshire	2104	Trump	44	2024-11-05	34	0.8944272	1.041263	NEngland
New Hampshire	2104	Stein	1	2024-11-05	34	0.8944272	1.041263	NEngland
New Hampshire	2104	Oliver	1	2024-11-05	34	0.8944272	1.041263	NEngland
New Hampshire	2104	West	0	2024-11-05	34	0.8944272	1.041263	NEngland

The table details the first few observations of the cleaned dataset, which we used to build our model.

For further analysis, the cleaned dataset was separated into two different datasets by beach name, for simplicity purposes. Data was analyzed through the R programming software (R Core Team 2023) and packages such as tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2014), and dplyr (Wickham et al. 2023) were used to help download, clean, simulate, analyze, and test the data.

Some pollsters are included in the dataset more than others based on the frequency of their conducted poll surveys and if they align with FiveThirtyEight's guidelines, which can be seen in Figure 1. Figure 1 highlights the distribution of the top 25 pollsters and how frequently they appear in the dataset, with Morning Consult being the most popular pollster with almost 500 appearances. Noting which pollsters are more trustworthy and provide reliable data is important for evaluating the quality of future analyses.

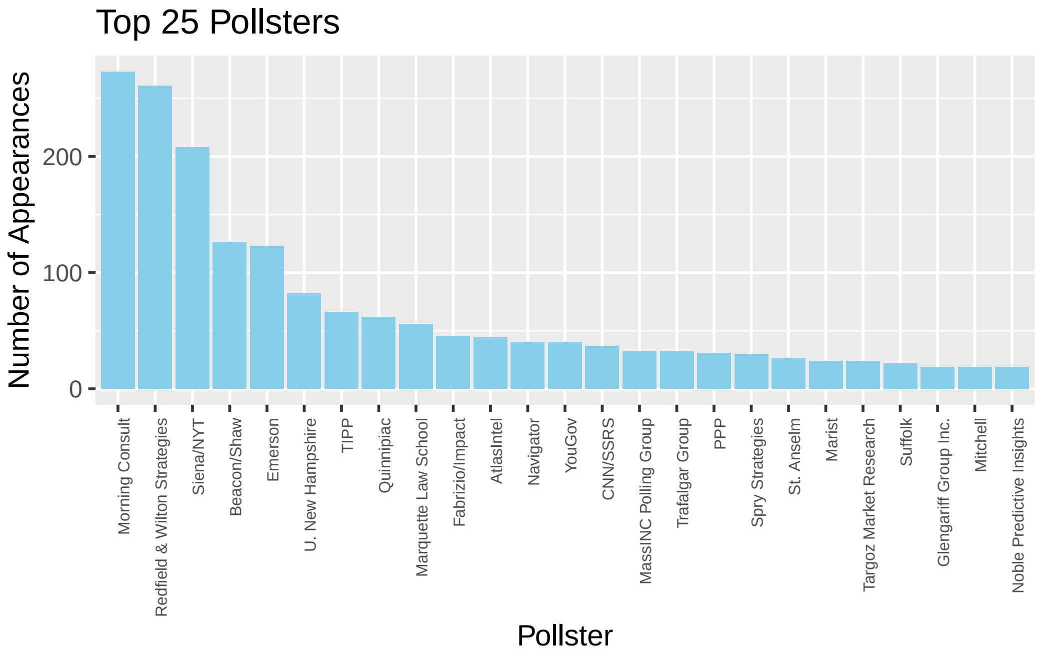

Figure 1: Top 25 pollsters in overall polling data

Figure 2 details the frequency of various polling methods used in the individual polling data, showing online panels are the most popular methodology used to deploy these polls, followed by live phone. The polling method plays a crucial role in reducing certain response biases. For instance, if the poll is conducted via an online panel with anonymous responses, participants may feel more comfortable expressing their true voting intentions, potentially reducing social desirability bias.

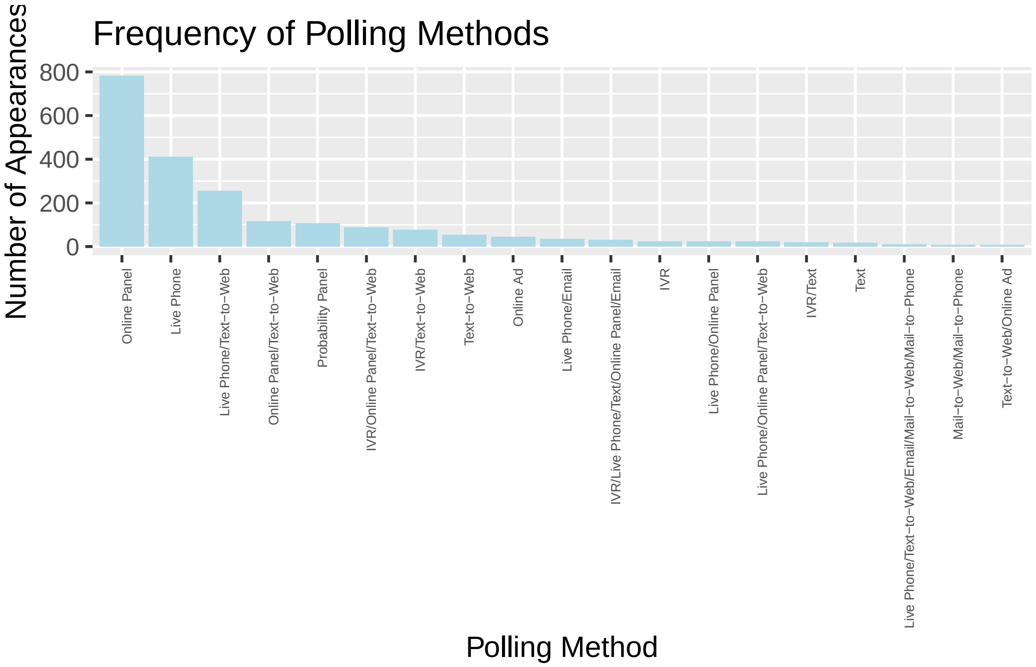

Figure 2: Frequency of Various Polling Methods

Finally, Figure 3 shows the number of votes per candidate. As expected, Harris and Trump have the highest number of votes, with Harris having slightly more votes based on the surveyed polls than Trump. Other candidates with a noticeable number of votes in the graph include Robert F. Kennedy and Jill Stein. However, their totals are significantly lower than those of Harris and Trump, making them likely insignificant in the election. It is also important to mention that Biden still appears in these polls, and therefore in Figure 3, due to the lag time in poll data processing by FiveThirtyEight. However, since he has dropped out of the race, he will not be included in our model.

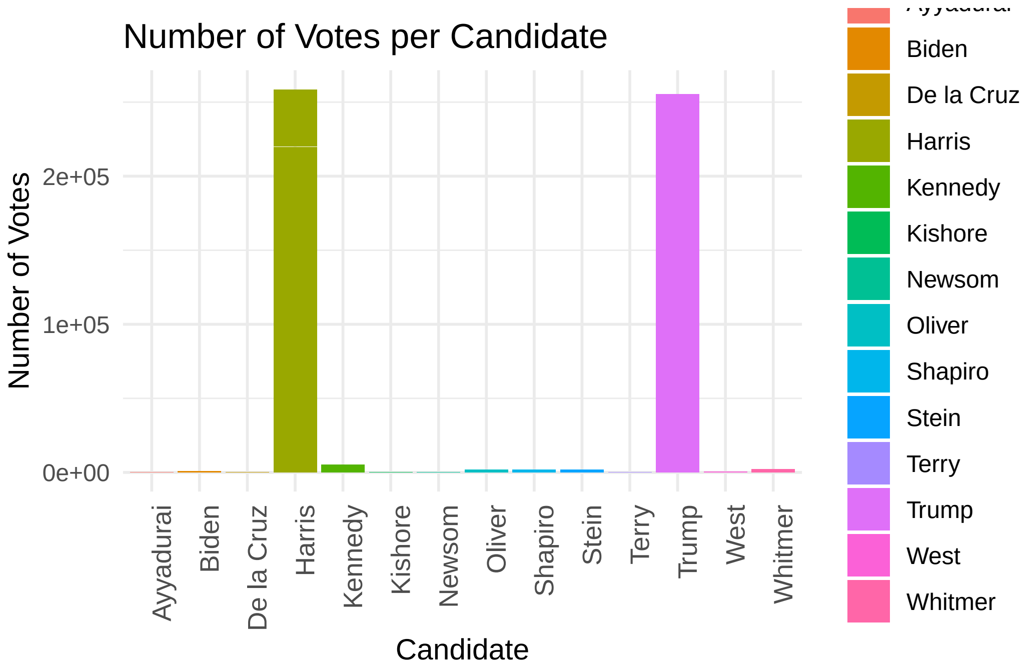

Figure 3: Number of Votes per Candidate

3 Model

3.1 Model set-up

We built three unique models, this allows us to choose the one that most closely fits the data historically, and weigh the pros and cons of each model to decide which one works best for us. This section will go through the process of how each of these models was developed and how we selected our final model. The first model was a conventional GLM, and the second was a Bayesian model with various differences and caveats.

For the GLM model, we began by defining the 538 political regions [REF], and we fit a Harris and Trump GLM model for each of those regions. The intuition behind this is that voters in these regions are likely to vote in similar ways, therefore by pooling their preferences together we can get a more accurate model than one built at the State level. In other words, there would be too much noise in the State level model. Furthermore, some states have little to no polling data making the building of a model in these cases impossible (NOTE: how do we deal with these states later if there's no polling?). We then apply the polling data to these fitted regional models at the state level, to gauge Harris and Trump support. This allows us to predict the winner of each state, which informs us how many electoral college votes a candidate will get, and thus the winner of the election.

The Bayesian model is necessarily more complicated. A major advantage of a Bayesian model is the ability to include prior information, encoded in the aptly named Prior. In the US elections states often vote very similarly [REF] across decades. We began to reflect this by classifying all states as either, Deep Red, Light Red, Swing, Light Blue, Dark Blue, to reflect their historical voting bias based on this list [REF]. With the Priors encoded we fit a Harris and Trump model using Bayes R package [REF]. We then synthesized a testing dataset for all the states, and used this data to get predictions of support for both candidates in each State. As was done in the previous model this gave us predictions on who would win each state, which allowed us to calculate the elector votes each candidate was likely to get and therefore the winner of the election.

3.1.1 Model justification

4 Results

MODEL NOT PERFECT WRITE AFTER DONE

The GLM predicts Harris winning with [EV VOTES], with Trump only winning [EV VOTES]. The Bayesinal model also predicts a Harris victory with [EV VOTES] and Trump with [EV VOTES].(Maybe better as a table).

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

A Appendix

One particular poll within our data sample is one conducted between October 11-14 by Beacon Research and Shaw & Company Research (cite Fox article). Targeting the sample frame of registered American voters, this poll collected data on 1,110 participants' responses on 51 questions through either live phonecall interviews or an online survey. The sample was recruited by applying the probability proportionate to size method on the nationwide voter file of registered voters' phone numbers, meaning that the contacted individuals would be proportionally representative of the number of voters per state (cite Fox article). This ensures that the voices represented by poll results would match the ones involved with the real election as much as possible.

The key finding from this poll is that 48% of participants favoured Harris while 50% favoured Trump, with the lead being consistent in the larger sample of registered voters and smaller subsample of likely voters. Despite this, responses also indicate that Harris is narrowly leading in seven swing states, meaning that Democrats could potentially win by electoral college while losing the popular vote (cite Fox article). This report also notes that current results are Trump's highest approval ratings since Biden dropped out of the race, while support for Harris is at its lowest.

One strength of this questionnaire is that information regarding the respondent's confidence level and commitment level was also collected. With questions such as "How often do you make a point to read or listen to the news?" and "Are you certain to support that candidate, or do you think you may change your mind and support someone else?", this poll is able to gain a more indepth view of how easily these participants may be swayed in the time between the polling and the real election (cite Fox article). As well, probabilistic statistical models based on past voting history, interest in current election, age, education, race, ethnicity, church attendance, and marital status were used to predict the respondents' likeliness to vote; interestingly, results found that the results for the full sample vs subsample of 870 likely voters varied by a ±3%, which can be quite significant given how tight the race is currently (cite Fox article).

On the other hand, one potential weakness of this poll is that it was sponsored by Fox News, a news site that has historically been Republican-leaning. Though from the reported methodology alone, no evident biasing of the pollster can be observed, this potential source of biased reporting must be noted as it can affect the type of analyses that occurred and which key findings are the focus of news reports. As well, another potential issue is the lack of indication regarding how non-responses of "Don't know" were handled during analysis (cite Fox article). This is a critical area to pay attention to as it is an option offered on every question, selected by up to 4% of participants on key questions such as "If the presidential election were today, how would you vote?" (cite Fox article).

References

Chad de Guzman and Koh Ewe. 2024. "A Guide to Kamala Harris' Views on Abortion, the Economy, and More." https://time.com/7001208/kamala-harris-views-abortion-economy-immigration-israel-gaza/.

FiveThirtyEight. 2024a. "FiveThirtyEight Latest Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

———. 2024b. *FiveThirtyEight Presidential General Polls Data*. https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

McKenzie Beard and Abbie Cheeseman and Justine McDaniel. 2024. "Harris Vs. Trump on Abortion: Where They Stand on the Issue." https://www.washingtonpost.com/politics/interactive/2024/trump-harris-abortion/.

NBC. n.d. "Harris and Trump: Compare Where They Stand on Key Issues."

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Weber. 2024. "Electoral College Information." https://www.sos.ca.gov/elections/electoral-college#:~:text=Under%20the%20%22Electoral%20College%22%20system,more%20%22votes%22%20it%20gets.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.