Forecasting the 2024 U.S. Presidential Election: Kamala Harris' Projected Victory*

A Bayesian Analysis of Polling Data from Key Battleground States"

Jimin Lee Sarah Ding Xiyan Chen

November 3, 2024

This paper presents a predictive model for the 2024 United States Presidential Election, focusing on critical battleground states. Utilizing a Bayesian approach with state-level polling data, we estimate the winning probabilities for candidates Donald Trump and Kamala Harris. Our findings reveal that Harris is projected to win in six out of seven key battleground states, including Michigan, Nevada, and North Carolina, with predicted winning percentages surpassing 50%. Given this significant advantage, we predict Kamala Harris to be the likely winner of the election. This analysis underscores the importance of battleground states in shaping electoral outcomes and provides valuable insights into voter sentiment as Election Day approaches.

1 Introduction

The 2024 United States Presidential Election is shaping up to be a critical event, particularly as battleground states emerge as pivotal areas influencing the electoral outcome. As candidates Donald Trump and Kamala Harris intensify their campaigns, understanding voter preferences in these key regions becomes essential for predicting the election results. This paper utilizes state-level polling data from Project 538 (FiveThirtyEight 2024) and the statistical programming language R (R Core Team 2023) to build a predictive model that estimates the likelihood of victory for each candidate in several key battleground states, including Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin.

The primary estimand in this study is the probability that either Donald Trump or Kamala Harris will win each battleground state, based on aggregated polling averages. Our model employs a Bayesian hierarchical approach with natural splines to capture the nuances of voter

*Code and data are available at: https://github.com/jamiejiminlee/2024_US_Elections.git.

sentiment over time, allowing for the estimation of winning probabilities across these crucial states. By comparing the predicted winning probabilities for both candidates, we can ascertain the likely winner based on the prevailing voter support patterns.

Our analysis reveals that Kamala Harris is projected to win in six out of the seven battleground states examined, indicating a favorable voter sentiment in states such as Michigan and Nevada. Conversely, Donald Trump shows a competitive stance in Arizona, where he is also predicted to secure a slight edge. The calculation of the predicted winner is determined by comparing the estimated probabilities for each candidate, which underscores Harris's overall advantage in the election.

Understanding these predictions is vital as it offers insights into the dynamics of voter behavior and the potential implications for election outcomes. This information can help political analysts, campaign strategists, and the public anticipate how these battleground states may influence the overall results. Accurate predictions not only inform campaign strategies but also highlight regions where voter sentiment may shift in the lead-up to Election Day.

The remainder of this paper is structured as follows. In Section 2, we describe the data and variables used in the analysis. Section 3 outlines the model setup and estimation strategy. Section 4 presents the results, including the predicted election outcomes and visualizations. Finally, Section 5 discusses the implications and limitations of our findings.

2 Data

2.1 Overview

The dataset of the 2024 Presidential Election cycle is obtained from Project 538 (FiveThirtyEight 2024). The dataset is periodically updated throughout the presidential election campaign to reflect current and most up-to-date polling results. It consists of polling information from various pollsters such as the New York Times, ActiVote, Morning Consult, etc. For every poll conducted, relevant variables such as state, start_date, end_date, transparency_score, candidate_name, and percentage of support (pct) for the appropriate candidates are included. The dataset compiles data from all major pollsters in the US and provides us with an understanding of the current state of the presidential election.

The present analysis focuses on Donald Trump and Kamala Harris, the two leading candidates in the current presidential election race. For all the following variables mentioned, we collected observations for both of the candidates. We collect observations from the dataset where the 'candidate_name' is Trump and Harris. We are interested in the support for both candidates over time, for each state, therefore, we collect observations of 'end_date', and 'state', as well as the electorate support for each candidate is denoted as 'pct'. Furthermore, since there are many pollsters present in the dataset, we extracted pollsters with high-quality data that is

reliable, which is denoted as having a high 'numeric_grade' and 'transparency_score', details of data cleaning are further explained in Appendix A.

2.2 Measurement

In the current presidential election cycle, Americans' opinions on voting for their preferred candidate are transformed into data points through a series of steps that involve surveying, data processing, and structuring responses. Pollsters then store these responses in structured databases for analysis. In the dataset obtained from Project 538, each entry in the dataset corresponds to a poll conducted by each pollster, capturing the percentage of respondents who express support for each candidate. Each pollster has different approaches to sampling respondents, recruiting respondents, as well as questionnaire types, which affects each pollster's reliability and data quality.

Pollsters begin by selecting a sample from the registered voter population, often using voter files with demographic and contact information. They then apply methods of stratified random sampling to ensure the sample reflects the electorate's diversity in terms of age, gender, race, education, voter behavior, etc.

Then, pollsters reach out to voters through various channels, such as phone calls, emails, text messages, or online panels. These are called recruiting methods and usually, each of them has its trade-offs. If the respondents are connected through any of these channels, they are asked a series of questions either by a live interviewer or a preset of questions.

The types of questions often include:

- "If the election were held today, would you vote for [Candidate A] or [Candidate B]?"
- "How strongly do you support [Candidate]?" (e.g., on a scale of 1 to 10, or strongly, somewhat, undecided)
- "Do you support [Policy A] proposed by [Candidate A] or [Policy B] proposed by [Candidate B]?"

Responses are recorded digitally, either by survey software or directly into a database by interviewers, capturing both the answers and relevant demographic data. Each response becomes a data point in the database. For example, if a respondent prefers Candidate A, their answer may be recorded as a binary variable (e.g., "1" for Candidate A, "0" for Candidate B). Along with candidate preference, pollsters capture demographic and behavioral details, such as age, gender, race, education, income, voter history, and party registration.

Each respondent's answers and demographic details are grouped and saved in a structured database. Pollsters then clean the data by checking for inconsistencies, removing incomplete responses, and addressing non-responses. Pollsters also perform weighting to ensure the representativeness of data collected, details of weighting are further explained in {appendix 1}.

Once cleaned and weighted, data points are aggregated to determine overall candidate support, often producing metrics such as the percentage of respondents supporting each candidate by state or other specific demographics. Each pollster would have a database with all the weighted and aggregated data, and Project 538 presents a compilation of the databases obtained from various pollsters.

2.3 Variables

The collected dataset contains several key variables relevant to the analysis:

- **State**: Indicates the state where the polling took place, critical for understanding regional support dynamics.
- **Polling Date**: The date when the poll was conducted, which helps in analyzing trends over time.
- **Pollster**: The organization conducting the poll, providing insight into the reliability and methodology of the polling data.
- **Numeric Grade**: A score assigned to each poll based on its methodology and historical accuracy, which assists in filtering out less reliable polls.
- **Transparency Score**: A measure indicating how transparent the pollster is regarding their methodology, further enhancing data credibility.
- **Candidate Name**: Identifies the candidate being polled, specifically focusing on Donald Trump and Kamala Harris.
- **Percentage of Support (pct)**: Represents the percentage of respondents supporting each candidate, which is the primary outcome variable of interest.

To ensure the quality of the analysis, only those polls that meet specific thresholds were included - for more details, refer to Appendix A.

2.3.1 Outcome Variable

As our focus is to analyze and predict voter support for Kamala Harris and Donald Trump during the current presidential election cycle, the primary outcome variables of interest are the percentages of support for each candidate, denoted as 'pct'.The pct is categorized by candidate and scaled to a proportion by diving by 100. This variable represents the proportion of voter support for each candidate based on polling data. Through understanding the percentage of support (pct) for each candidate, we can observe and infer the candidate that has a leading advantage in winning the presidential election race.

To give an overview of the outcome variable, divided per candidate, we provide summary statistics of this key variable, highlighting the characteristics of the polling support for each candidate across the battleground states.

Table 1

Statistic	Value
Total Polls	107.000
Average Trump Support	47.801
Max Trump Support	51.700
Min Trump Support	43.000
Average Harris Support	47.999
Max Harris Support	51.200
Min Harris Support	43.600

Summary Statistics for Trump and Harris Polling Data - Presents key summary statistics, including total polls, average support percentages, and the maximum and minimum polling levels for both candidates across the specified battleground states. All values are rounded to three decimal places, providing an overview of the polling landscape as of the selected election date.

For instance, Table 1 reveals a total of 107 polls conducted, calculated as the total count of rows in our cleaned dataset. 'Total polls' reflects the total number of unique poll entries after filtering for only battleground states, polls with sufficient data quality, and polls that report support percentages for both Trump and Harris. The 107 polls conducted display an average support level of 47.801% for Trump and 47.999% for Harris. This suggests that based on the dataset, Trump has a leading advantage in winning the presidential election race by having a slighty higher average level of support than Harris. The highest percentage of support for Trump in the 107 polls was 51.700% while the lowest recorded support was 43%. For Harris, the highest percentage of support also reached 51.200%, with a minimum of 43.600%. These statistics illustrate the competitive landscape of the presidential election race, without a clear leading candidate in the current race.

It is important to note that since 'pct' stands for percentage, there may be a misconception that the values for Trump and Harris will always add up to 100%. For instance, in the summary statistics presented in Table 1, the average support for Donald Trump is 47.801%, while the average support for Kamala Harris is 47.999%. This results in a combined average support of 95.800%, indicating that there is a significant portion of respondents who are either undecided or supporting other candidates. This highlights the competitive nature of the election and emphasizes the necessity of examining not only the percentages attributed to the main candidates but also the context of voter preferences that could influence the final outcome.

2.3.2 Predictive Variables

2.3.2.1 State

The state variable indicates the U.S. state where the poll was conducted or targeted. In the context of the U.S. presidential election, the outcome is frequently determined by several key states. In the current 2024 election cycle, there are seven battleground states that are pivotal in the contest between Kamala Harris, representing the Democratic Party, and Donald Trump, representing the Republican Party. To enhance our model and refine our predictions, we focused exclusively on these seven battleground states. The polling results from each of these states play a crucial role in predicting electoral outcomes, as the percentage of support is evaluated on a state-by-state basis.

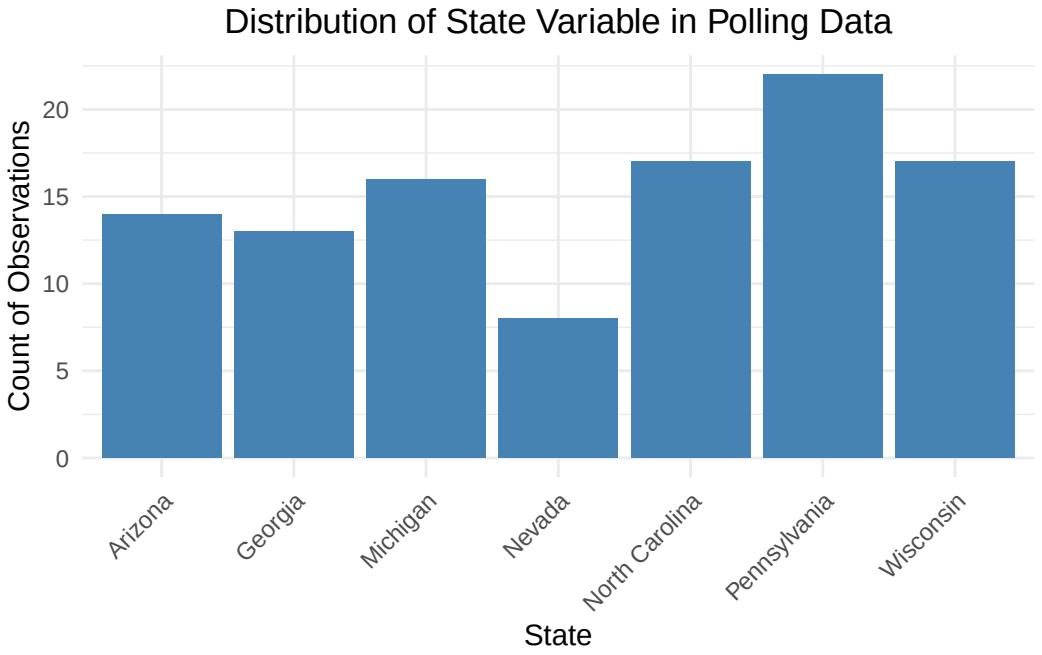

Figure 1: Bar plot showing the distribution of observations across key battleground states in the 2024 United States Presidential Election polling data. The plot illustrates the count of polling observations for each state, highlighting states with more frequent polling activity. This visualization aids in understanding the level of engagement and data collection in each battleground state leading up to the election.

The distribution of the state variable is visualized in Figure 1 and indicates that some states, such as Pennsylvania, North Carolina and Wisconsin have more polling observations compared to others, like Nevada. This variability reflects differences in polling frequency and interest among these battleground states. States with more polling observations provide a larger sample size, potentially leading to more stable and reliable predictions for those states in the model.

2.3.2.2 End Date

Another predictive variable is 'end-date'

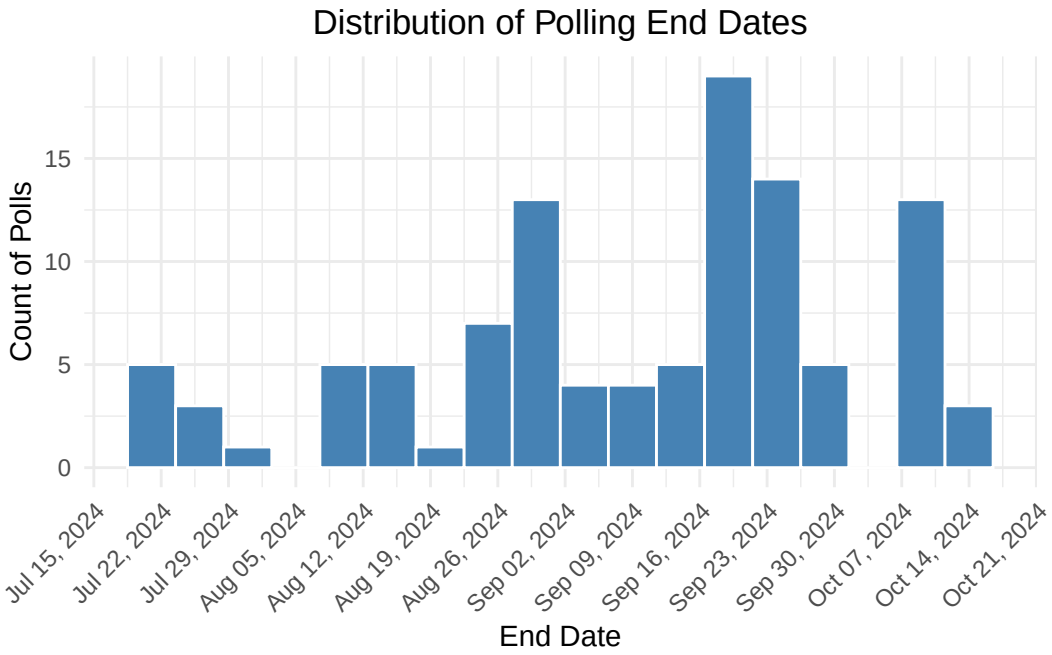

Figure 2: Histogram illustrating the distribution of polling end dates for key battleground states in the lead-up to the 2024 United States Presidential Election. The histogram shows the frequency of polls conducted over time, providing insight into the timing of polling activities and potential trends in voter sentiment as Election Day approaches. This visualization is crucial for understanding how polling data correlates with the election timeline.

2.3.3 Other Variables

'numeric grade'

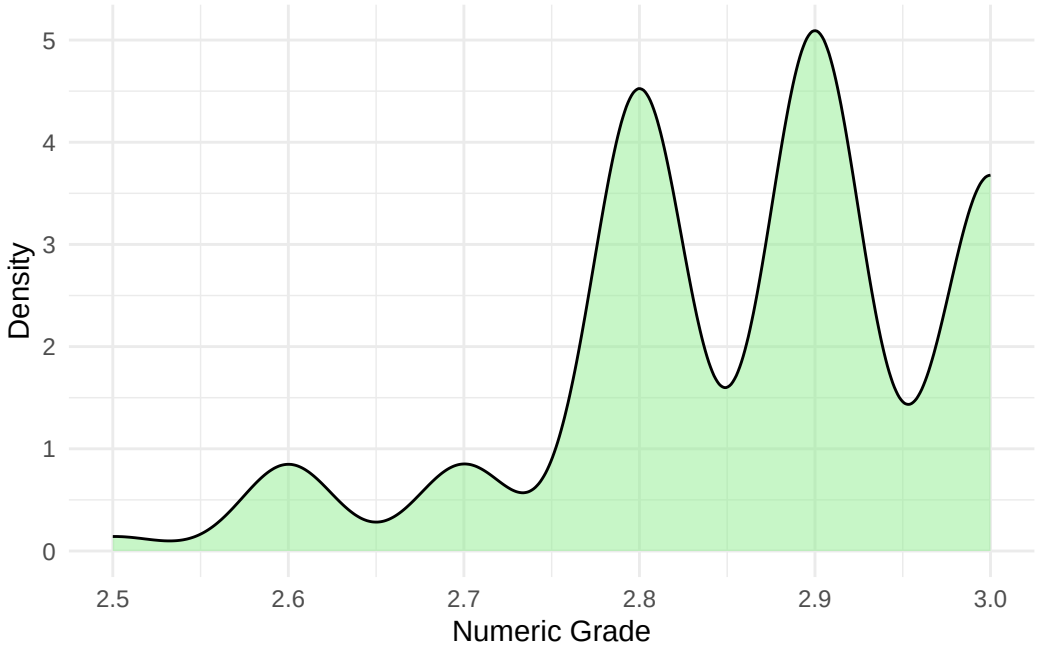

Figure 3: Histogram illustrating the distribution of polling end dates for key battleground states in the lead-up to the 2024 United States Presidential Election. Displays the frequency of polls conducted over time, providing insight into the timing of polling activities and potential trends in voter sentiment as Election Day approaches. This visualization is crucial for understanding how polling data correlates with the election timeline.

'transparency score'

'pollster'

'start_date'

3 Model

In this analysis, we develop separate Bayesian models for candidates Donald Trump and Kamala Harris to predict their respective probabilities of winning in the 2024 U.S. Presidential Election. The use of distinct models allows us to capture the unique polling dynamics and voter

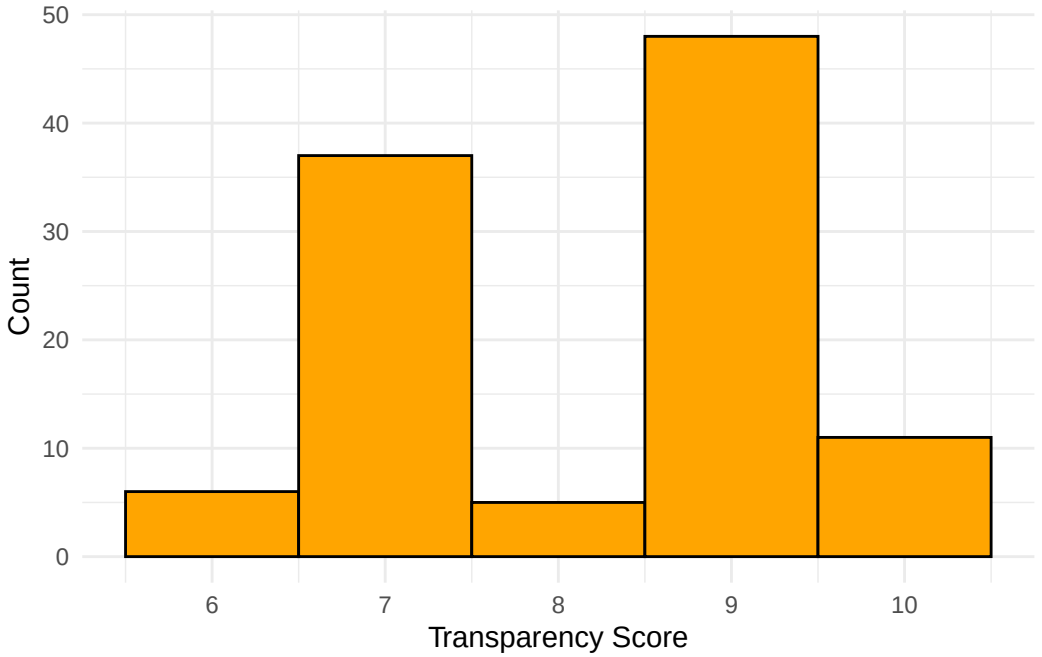

Figure 4: Histogram illustrating the distribution of transparency scores for polls in key battleground states. Reveals the frequency of different transparency ratings, providing insight into the reliability of polling data. Understanding the distribution of transparency scores is crucial for evaluating the credibility of the polling results as Election Day approaches.

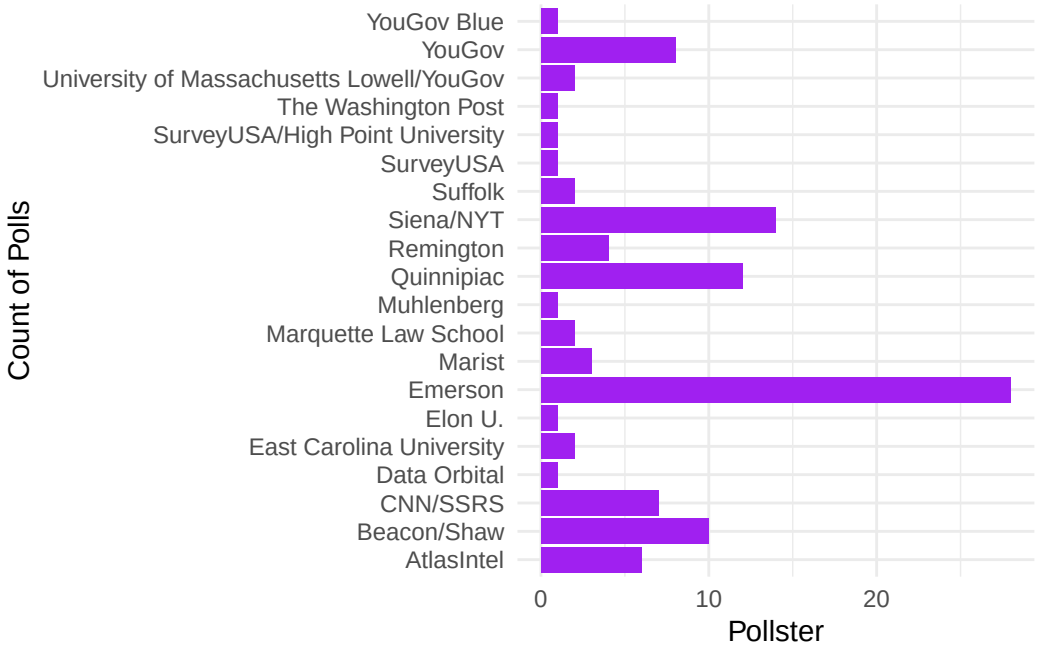

Figure 5: Horizontal bar graph illustrating the distribution of polls by various pollsters in key battleground states. Displays the count of polls conducted by each pollster, highlighting the most frequently utilized polling organizations. Understanding the diversity of pollsters is essential for assessing the reliability and variance in polling data as Election Day approaches.

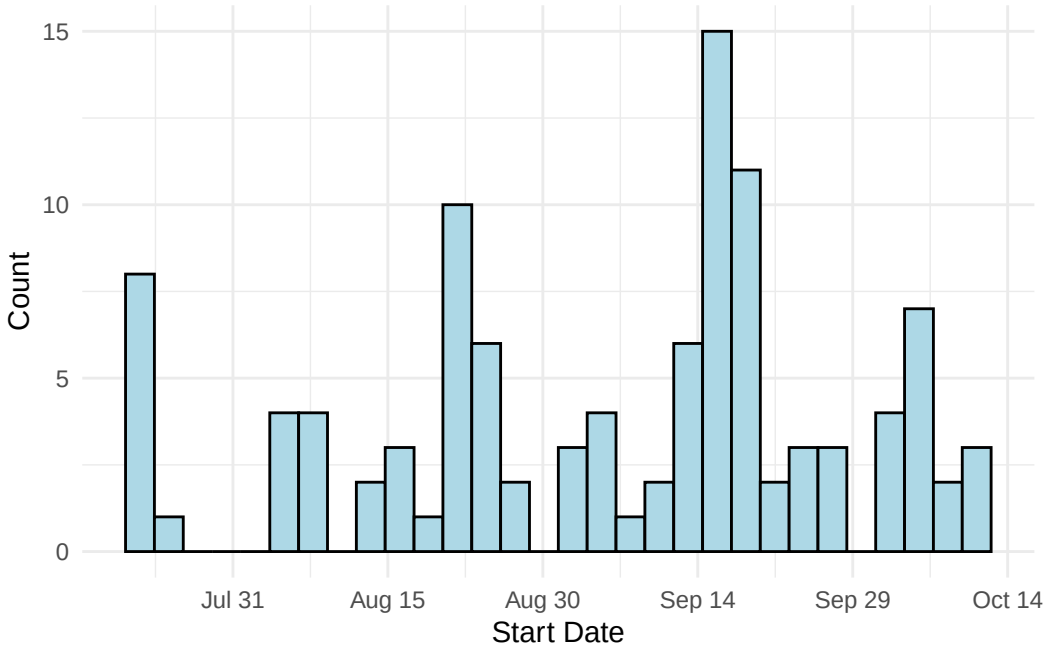

Figure 6: Histogram illustrating the distribution of polling start dates for key battleground states in the lead-up to the 2024 United States Presidential Election. This visualization displays the frequency of polls conducted over time, providing insights into when polling activities were initiated and potential trends in voter sentiment as Election Day approaches. Understanding the timing of polls is crucial for evaluating the relevance of the polling data.

preferences for each candidate across key battleground states, which include Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. By analyzing polling data collected from various sources, we aim to provide a comprehensive assessment of each candidate's electoral prospects.

The models are built using state-level polling data, which is aggregated to ensure robustness in the predictions. Each candidate's model incorporates temporal factors through natural splines applied to the numeric representation of the end date, allowing us to observe changes in voter support over time. This approach provides a nuanced understanding of how public sentiment evolves as the election date approaches. Further background details, model specifications, and diagnostics are included in Appendix B.

3.1 Model Setup

The primary estimand in our models is the probability of victory for each candidate in the selected battleground states, expressed through a logistic regression framework. The model is formulated as follows:

$$P(\text{Victory}_i) = \frac{e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num}_i + u_{\text{state}}}}{1 + e^{\beta_0 + \beta_1 \cdot \text{end\_date\_num}_i + u_{\text{state}}}}$$

Where:

- $P(\text{Victory}_i)$ is the predicted probability of winning for candidate i.
- β_0 represents the intercept, reflecting the baseline log-odds of winning.
- β_1 is the coefficient for the natural spline transformation of the numeric end date (end_date_num_i), capturing nonlinear trends in voter support over time.
- u_{state} is the random effect associated with each state, accounting for variations in voter preferences.

Data preparation involved creating two distinct datasets from the cleaned polling data: one for Donald Trump and another for Kamala Harris. Each dataset included the state, end date, and corresponding polling percentage, with the end date converted into a numeric format representing the number of days since the earliest date in the dataset. This allows the model to effectively capture the evolving nature of voter sentiment as the election approaches.

We implemented the models using the `brms` package in R, selecting a **Beta distribution** with a **logit link function** to accurately model the proportions of support for each candidate:

$$y_i \sim \text{Beta}(\mu_i, \phi)$$

In this model:

- y_i represents the proportion of support for candidate i in a given state.
- μ_i is the mean of the Beta distribution, reflecting the expected support proportion.
- ϕ is the precision parameter, indicating the variability around the mean.

This approach allows us to effectively capture the dynamics of voter support in various states as we approach the election.

3.2 Model Justification

The decision to utilize Bayesian logistic regression models stems from their flexibility and the ability to incorporate prior beliefs into the estimation process. This approach is particularly valuable in the context of electoral forecasting, where uncertainty is a significant factor. We defined the following priors for the model parameters:

- For the intercept β_0, we specified $\beta_0 \sim \text{Normal}(0, 10)$, providing a relatively uninformative prior that allows the data to inform the estimation of baseline log-odds.
- For the slope β_1, we used $\beta_1 \sim \text{Normal}(0, 5)$, reflecting reasonable expectations regarding the influence of polling trends over time.

Natural splines were chosen to model the effect of the end date due to the observed non-linearities in voter support, allowing the model to adapt to changing public opinion. Key assumptions include the independence of observations within states and the appropriateness of the Beta distribution for modeling proportions. However, potential limitations include the reliance on the accuracy of polling data and the assumption that historical voting patterns are indicative of future behavior.

3.3 Model Summary

3.3.1 Harris Model Summary

Table 2: Fixed Effects Summary for Harris Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.1508299	0.0279264	-0.2041163	-0.0954951
nsenddatenumdfEQ41	0.1225217	0.0341326	0.0552911	0.1903117
nsenddatenumdfEQ42	0.0248781	0.0291738	-0.0321764	0.0833965
nsenddatenumdfEQ43	0.1191128	0.0630871	-0.0070147	0.2416542
nsenddatenumdfEQ44	0.0669873	0.0382015	-0.0071597	0.1426828

The model summary for Kamala Harris indicates that the intercept estimate is -0.1508 with a standard error of 0.0279. This intercept value suggests a negative baseline log-odds for her

victory, reflecting the starting point of her support in the absence of any other predictors. The model shows that the coefficients for the natural spline terms ($nsendate_{num}$) are particularly informative, with the first spline term ($nsendate_{num\_df41}$) yielding a positive estimate of 0.1225, suggesting that as time progresses, Harris's probability of winning increases, particularly during key moments captured by the polling data. This is also reflected in the confidence intervals, with the upper limit of the confidence interval for this term reaching 0.1903. In contrast, the estimates for the second spline term ($nsendate_{num\_df42}$) indicate a smaller effect, suggesting fluctuations in her support that may warrant further investigation.

3.3.2 Trump Model Summary

Table 3: Fixed Effects Summary for Trump Model.

term	Estimate	Std. Error	Lower 95% CI	Upper 95% CI
Intercept	-0.0922852	0.0305620	-0.1520050	-0.0333538
nsenddatenumdfEQ41	-0.0071493	0.0355882	-0.0793200	0.0623891
nsenddatenumdfEQ42	0.0802768	0.0317675	0.0179060	0.1426297
nsenddatenumdfEQ43	0.0025949	0.0676124	-0.1341383	0.1336160
nsenddatenumdfEQ44	0.0789177	0.0400326	0.0013945	0.1589037

In comparison, the model summary for Donald Trump presents a different picture. The intercept for Trump's model is -0.0923 with a standard error of 0.0306, indicating a similarly negative baseline log-odds of victory as Harris. However, Trump's model demonstrates more pronounced increases in support as reflected by the estimates for the natural spline terms. For example, the second spline term ($nsendate_{num\_df42}$) has a positive coefficient of 0.0803, highlighting a significant upward trend in support over time. This trend is further supported by the confidence interval, which ranges from 0.0179 to 0.1426, indicating a substantial increase in his predicted probabilities as the election approaches. The fourth spline term ($nsendate_{num\_df44}$) also shows a positive coefficient of 0.0789, suggesting that the support for Trump remains resilient over the polling period. Overall, both models reveal complex interactions between time and voter sentiment, with implications for the electoral dynamics leading up to the election.

4 Results

4.1 State-Level Polling Averages for Trump and Harris

WRITE - explain why we are showing this average table first and what the table reveals

State	Trump %	Harris %
Arizona	49.2	46.8
Georgia	48.9	47.2
Michigan	46.9	48.1
Nevada	47.9	48.5
North Carolina	47.6	48.1
Pennsylvania	47.4	48.3
Wisconsin	47.2	48.9

Table displaying the percentage of support for Donald Trump and Kamala Harris across key battleground states in the lead-up to the 2024 United States Presidential Election. The table lists states with their respective support percentages, highlighting the competitive landscape as voters prepare to make their decisions on Election Day. Notably, Trump maintains a slight lead in states such as Arizona and Georgia, while Harris shows stronger support in states like Michigan and Nevada. This data provides insight into the current polling dynamics and potential electoral outcomes in these critical states.

4.2 Polling Averages for Trump and Harris Over Time in Battleground States

Figure 7 provides a visual comparison of the current polling support for Donald Trump and Kamala Harris across key battleground states, tracked over time from August 2024 to October 2024. Each line plot represents the average polling percentages for both candidates in a specific state, with red lines denoting Trump's support and blue lines representing Harris's support. Across several states, there are noticeably increased fluctuations in polling averages for both candidates closer to the election particularly from late September to October. This pattern may indicate heightened voter interest and engagement as the election date approaches which is November 4th, 2024, it could also reflect changes in public opinion due to campaign events, political debates, or other external factors.

Out of the seven battleground states, Arizona and Georgia display a clear leading win by Trump while Nevada and North Carolina display a clear leading win by Harris. Michigan, Pennsylvania, and Wisconsin are the three states that don't have a clear leading winner. Therefore, these three states are the key states that could determine the election outcome as it approaches election day. Election outcomes could be easily flipped if either of the candidates can sweep more than one out of the three states. Candidates should focus on strategizing their campaigns to increase their voters' support at critical moments. The frequent fluctuations of greater margins also indicate the uncertainty and instability of voters' attitudes towards both candidates.

Since it is difficult to conclude a definitive predictive outcome by observing the current data, we utilize our models, described in Section 3 to predict the possible outcomes by incorporating

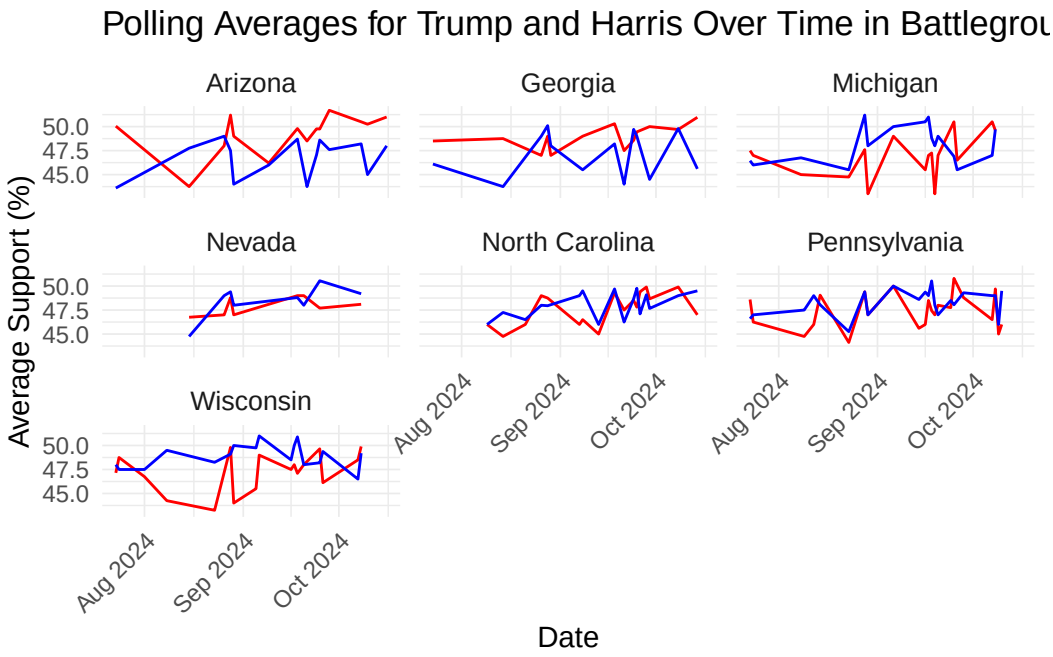

Figure 7: Polling Averages for Trump and Harris Over Time in Battleground States.

the given data. By using the current data, the model can identify the state-specific trends thus generating more accurate predictions of the outcome. Relevant prediction results are provided below.

4.3 State-Level Probability of Trump Winning on Election Day

By incorporating our Bayesian models, we generate the predicted probabilities of winner for both Donald Trump and Kamala Harris in key battleground states, on Election Day (November 5th, 2024). Figure 8 provides a visual summary of these predicted probabilities.

These predictions are based on polling data collected up to October 19th, 2024 to forecast each candidate's support within these critical states. Each state's political dynamics are unique, and this is reflected in the model's predictions. States like Arizona and Georgia show relatively close probabilities for both candidates, marking them as key states to focus on as they are the most influential on the election outcomes. The five other states all show relatively distinct leading support for Harris, suggesting that Harris's support base in those states has been consistently strong in the polls leading up to Election Day.

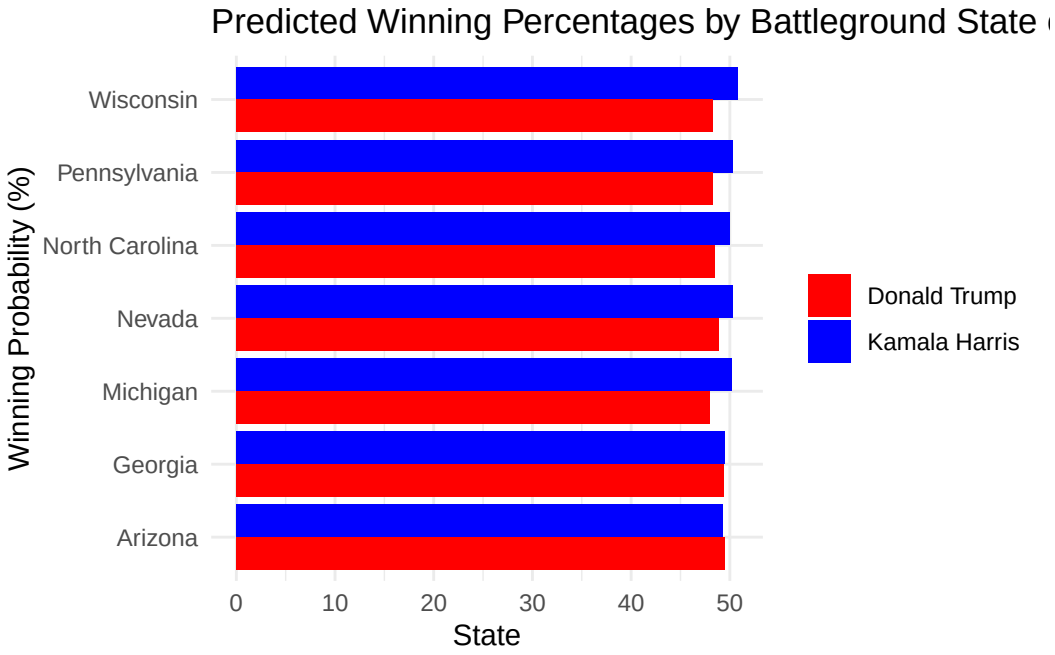

Figure 8: Probability of Trump Winning by State on Election Day.

4.4 Probability of Winning by State on Election Day

Table 5 displays the numeric percentage of the probabilities of winning for Trump and Harris. By comparing the predicted probabilities, the predicted winner for each state is assigned accordingly. For instance, the table displays that Trump has a 49.48% predicted probability of winning Arizona compared to 49.24% of Harris's, making Trump the predicted winner in Arizona. For all six other states, Harris has a higher predicted probability of winning than Trump.

Table 5

State	Predicted Trump Win (%)	Predicted Harris Win (%)	Predicted Winner
Arizona	49.52	49.24	Donald Trump
Georgia	49.35	49.51	Kamala Harris
Michigan	47.97	50.18	Kamala Harris
Nevada	48.84	50.30	Kamala Harris
North Carolina	48.51	50.04	Kamala Harris
Pennsylvania	48.26	50.31	Kamala Harris
Wisconsin	48.31	50.81	Kamala Harris

Probability of Trump Winning by State on Election Day.

4.5 US Map Showing Predicted Winner by State on Election Day

To visualize the colour distribution of the predicted results, we emloy the US map from (Albers et al. 2023) to display the results from Table 5.

- 'Predicted Winner' column is Donald Trump: State filled red
- 'Predicted Winner' column is Kamala Harris: State filled blue

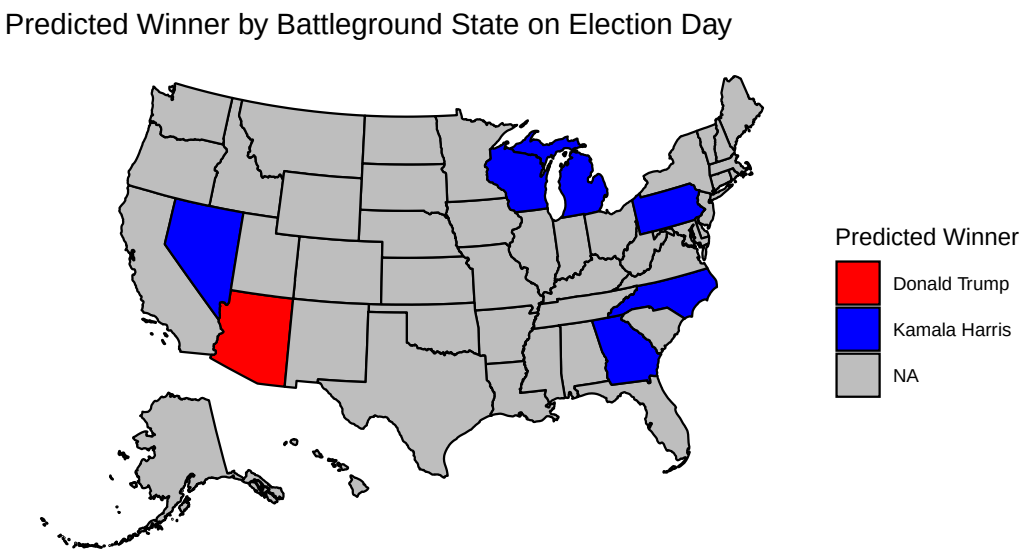

Figure 9: Predicted Winner by Battleground State on Election Day.

According to Figure 9, Arizona is the only state where Trump has a leading win while all the other six states are led by Harris. Since Harris has a six-to-one lead in the predicted percentage of support for all battleground states, we conclude that Harris is the predicted winner for the 2024 cycle of Presidential Election.

5 Discussion

5.1 Interpretation of Results

Since the electoral college system makes state-level forecasts crucial, and the U.S. operates under a predominantly two-party system, the focus of election predictions should be on battleground states. While many states are strongly aligned with either the Democratic or Republican party, a handful of key battleground states—where neither party has overwhelming dominance—will likely determine the overall outcome of the election. These battleground states, often evenly split in voter sentiment, hold immense significance because their electoral votes can swing the election in favor of one candidate or the other.

The results of our model underscore the critical role that these battleground states will play in the 2024 election. Although many states are predictably stable due to their party loyalty, it is the six battleground states that will serve as the main arena for competition. In these states, where voter preferences are not as firmly entrenched, even minor shifts in public opinion or turnout could lead to vastly different outcomes. As a result, a comprehensive and precise prediction of voting behavior in these battleground states is far more important than in states where the result is a foregone conclusion.

5.2 Model Performance

Out-of-sample testing results Appendix B.1 demonstrate the effectiveness and reliability of our Bayesian logistic regression model in predicting state-level outcomes for the 2024 U.S. presidential election. Our accuracy test generated a 75% score for the model, indicating our model's high reliability for making predictions about state-level election outcomes. Since presidential elections hinge on state-level wins, a reliable state-by-state prediction would imply a reliable national outcome prediction.

The RMSE value generated for our model is 0.5, which indicates that the model-predicted probabilities are close to the true outcomes. The confusion matrix breaks down the model's predictions into four categories: 1. orrectly predicted Trump wins, 2. Correctly predicted Harris wins, 3. Incorrectly predicted Trump wins, 4. Incorrectly predicted Harris wins. The generated confusion matrix for our model all had values of 0 to 2 for all four categories, indicating a balanced prediction between the two candidates, and suggesting the reliability of our model. Out-of-sample testing was conducted to test the reliability of our model, with further details provided in Apppendix B.1.

5.3 Weaknesses and Next Steps

While our model provides valuable insights into the electoral landscape, it faces limitations due to missing data for certain states, which results in some states being left blank in our

predictions. Although our primary focus is on understanding the dynamics within battleground states, the lack of information from other states can lead to incomplete predictions which may obscure broader trends affecting voter sentiment.

To address these weaknesses and improve the prediction, we propose several next steps: actively seeking additional data sources to fill gaps, including polling data and historical records; conducting sensitivity analyses to understand the impact of missing data on predictions; and analyzing existing trends within battleground states for more insights.

Another weakness of our model is the lack of consideration for the national vote, which may impact our election predictions. The national vote refers to the total number of votes cast by citizens across the country during an election, representing the support for each presidential candidate. Although the national vote does not directly determine the outcome of the election due to the Electoral College system, it can still provide valuable insights into voter sentiment and trends. By not considering the national vote into our analysis, we risk missing potential influences on voting behavior and electoral outcomes.

Appendix

A Additional Data Details

Given the nature of the presidential election, the electoral college is more impactful and influential on the possible outcome compared to the nationwide popular votes. Therefore, the majority of the pollsters choose to focus on the battleground states to perform their polling, to gain more insight into the public sentiment specific to certain states. For instance, Sienna/NYT mentions that they have in-state interviewers who call their respondents in states such as Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin.

This could explain the discrepancies in our data, where there are many missing values from states other than the battleground ones. Furthermore, many pollsters focus on state-wide polls instead of nationwide polls given the nature of the election, resulting in fewer data points or missing data for nationwide poll results. Even though the national popular vote and the results in the electoral college don't line up a lot of the time, the popular votes can still provide information about issues and opinions that are shaping the election as a whole. Therefore, even if we are focusing on selective states, results from other states and popular votes still remain relevant to outcome prediction.

During our data cleaning process, we added constraints of high numeric grades and high transparency scores, leading to fewer data points for us to build the model on. A numeric grade is a numeric rating given to the pollster indicating their quality or reliability, with a highest rating of 3.0. The transparency score is a score reflecting the pollster's transparency about their methodology where the highest score is 10. Therefore, to ensure high-quality and reliable data to build our model on, we filtered pollsters with numeric grades of 3.0 and transparency scores of 5 or higher. This is one of the reasons why there are no data points available for Kamala Harris for the state of Colorado, represented by the empty row **?@fig-averagetable**.

Furthermore, in the full raw data set from 538 (FiveThirtyEight 2024), for the state of Colorado, even without any constraints for numeric grade or transparency scores, there are only 3 data points for Kamala Harris. After reviewing the datasets, we found that except for the 3 available data points, all the rest of the polls ended in June 2024, which is a month earlier than when Biden dropped out of the race before Harris entered the presidential election race. It would not be possible that there are data points available for Harris if she was not participating in the race yet. The state of Colorado is a blue state, with the Democrats winning the state in every presidential election since 2008. Since our paper focuses on mainly the battleground states to predict election outcomes, limited data from a blue state like Colorado would not be too influential on our results.

B Model details

B.1 Diagnostics

Out-of-sample testing was conducted to test our model's reliability in predicting potential presidential election outcomes. We train our model on a training set of our data and then test its performance on the test set of our data, we evaluate how well the model generalizes to new data that it was not trained on.

This process helps reveal whether the model is over-fitting to the training data or truly capturing the underlying patterns that apply more broadly. The test produces an accuracy score, RMSE, and a confusion matrix, each measuring different aspects of the model's prediction. Accuracy is the proportion of correctly predicted outcomes compared to the total number of predictions made in the test set. The testing was implemented using the following packages in R (R Core Team 2023): caret (Kuhn et al. 2023), rstanarm (Gabry et al. 2023), ggplot2 (Wickham et al. 2023b), dplyr (Wickham et al. 2023a).

B.1.1 Model Diagnostics Result

?@fig-barplot-diagnostics displays the performance of the Bayesian logistic regression model used to predict the outcome of the 2024 presidential election. The model's accuracy is approximately 75%, indicating that 75% of the test set's state-level outcomes were correctly predicted. The RMSE (Root Mean Squared Error) is about 0.5, showing the average error in predicting the winning candidate across states.

This suggests the model performs reasonably well, but there is some error in predicting the exact outcome in certain states, which may be due to variations in polling data or other unaccounted factors.

B.1.2 Confusion Matrix

?@tbl-confusion-matrix displays the model's predictive performance in distinguishing between state-level wins for Trump (coded as 1) and Harris (coded as 0).

In this case, the model correctly predicted 12 Harris wins and 15 Trump wins (true positives and true negatives). However, there were 5 instances where the model incorrectly predicted a Trump win when Harris won (false positives), and 3 instances where it predicted a Harris win when Trump won (false negatives). This breakdown helps assess not only the overall accuracy but also how well the model distinguishes between close contests in different states, which is critical for predicting the outcome of the 2024 presidential election.

C Pollster Methodology Overview and Evaluation

The following information about the Siena College/New York Times poll methodology is based on the details provided in the article "How The Times/Siena Poll Is Conducted" by the New York Times (The New York Times 2023).

The Siena College/New York Times poll for the 2024 presidential election uses a robust sampling approach to ensure accuracy and relevance. The poll applies a stratified dual-frame sample, drawing from both land-lines and cell phones. Each poll is conducted by phone using live interviewers at call centers based in Florida, New York, South Carolina, Texas, and Virginia. The sampling population is all registered voters who live in the six battleground states: Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin.

Given the nature of the presidential elections, the decision is based on the electoral college instead of the popular vote, thus the pollster focuses on polling on the states that are likeliest to decide the outcome of the race. The voters' information is taken from the L-2 voter file, which includes details such as voter registration and history of participation in previous elections.

To refine their sample, Siena College adjusts the data by key demographic and political variables such as region, race/ethnicity, party affiliation, education, and voting patterns from the 2020 election. These adjustments help ensure the sample accurately represents the electorate. Additionally, they use a "likely voter screen," which combines self-reported likelihood of voting with historical voter behavior to estimate how likely respondents are to vote in the upcoming election.

At every step of the survey, Sienna/NYT uses the information in the data to try to ensure that they have the right number of Democrats and Republicans, young people and old people, the right ratio of people with different income levels, and a diverse mix of different races and regions. Once the survey is complete, they compare their respondents to the voter file and use a process known as weighting to ensure that the sample reflects the broader voting population. This combination of historical data and weighted adjustments ensures the poll is designed to predict election outcomes as accurately as possible.

The New York Times/Siena College pollsters handle non-response by using weighting and adjusting their sample to correct for any biases that might emerge due to individuals not responding. Specifically, they adjust on multiple demographic and political variables such as age, gender, education, race/ethnicity, and party affiliation. They also account for variations in voter likelihood and previous voting patterns, ensuring that the sample represents the likely electorate as accurately as possible.

This helps reduce the potential bias from non-response, especially since certain demographic groups or political affiliations may be less likely to respond to polls. By rebalancing the sample, they ensure that even if some groups have lower response rates, their representation in the final poll results aligns with what is expected based on historical trends and current voter enthusiasm.

D Idealized Methodology

Building on our discussion of the New York Times/Siena College Poll, we now present an idealized methodology that aims to enhance the accuracy and reliability of our predictions. This approach incorporates best practices and innovative techniques to ensure a comprehensive prediction of voter behavior in the upcoming 2024 U.S. presidential election. To demonstrate our idealized methodology, we generated a survey on the 2024 US Presidential election, which can be accessed via the the URL provided in (Google 2024) under **?@sec-references**.

D.1 Sampling Approach

Using data from voter files, which contain demographic information about registered voters, we aim to ensure that our sample accurately represents the population. To achieve this, we will use stratified random sampling. Given the importance of state-level forecasts due to the Electoral College system, we will first stratify by state.

Within each state, we will apply stratified sampling again, dividing the population into 6 groups and selecting 100 random samples from each stratum within those groups, ensuring our sample captures key demographic critical for predicting voter behavior. The groups and their respective strata include age groups (18-29, 30-44, 45-64, and 65+), gender (Male, Female, and Other), and race/ethnicity (White, Black, Latino, Asian, and Other). Additionally, household status will be categorized as either renting or owning, with home ownership serving as a proxy for wealth. Voting history will be classified into those who voted in the previous election and non-voters, while party registration will include Democrats, Republicans, and Independents.

For the recruitment of participants, we will begin by sending survey invitations to our selected voters via email, offering a $20 incentive to encourage participation. To improve response rates and reduce non-response bias, we will send a reminder email three days later. This approach will help minimize expenses to some extent. If we do not receive a response within a week (7 days), we will follow up by phone to reach those who did not respond to the email invitation.

D.2 Data Validation

After completing the survey, we will adjust the data through weighting to ensure the sample accurately reflects the broader population for predictive purposes. More weight will be assigned proportionally to each stratum based on its size, and additional weight will be given to respondents from strata that are less likely to participate in surveys. And to avoid duplicate responses, each voter will be assigned a unique ID with only the first response from each ID being retained. To ensure the selected voters are completing the survey correctly, we will track responses in real-time using a centralized system. Additionally, we will include validation questions in the survey to catch careless or fraudulent responses.

References {sec-references}

Albers, L. E. et al. 2023. *Usmap: US Maps Including Alaska and Hawaii*. https://CRAN.R-project.org/package=usmap.

FiveThirtyEight. 2024. *FiveThirtyEight: 2024 US Presidential Election Polls*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gabry, Jonah et al. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Google. 2024. *Google Form: [2024 US Presidential Election Survey]*. https://docs.google.com/forms/d/1S6aN5Q82orjKYJI4i7001sPyn5jhRkgNIkI9teiQq_o/edit.

Kuhn, Max et al. 2023. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

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

The New York Times. 2023. "How the Times and Siena College Poll Was Conducted." https://www.nytimes.com/article/times-siena-poll-methodology.html.

Wickham, Hadley et al. 2023a. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

——— et al. 2023b. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.