Predicting the 2024 U.S. Presidential Election: Trump's Likely Victory Based on Polling Analysis*

A Bayesian Approach Reveals Strong Support for Donald Trump Across Key Battleground States, Highlighting Critical Trends in Voter Preferences Leading Up to Election Day

Jimin Lee Sarah Ding Xiyan Chen

November 3, 2024

This paper presents a predictive model for the 2024 United States Presidential Election, utilizing state-level polling data to focus specifically on key battleground states. By employing a Bayesian approach, we estimate the outcomes between Donald Trump and Kamala Harris, incorporating high-quality polls from reputable sources. Our Bayesian hierarchical model, featuring time-dependent splines and state-specific random effects, effectively captures shifts in candidate support over time. The analysis reveals that Trump is projected to win in critical battleground states such as Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, highlighting his advantage in these pivotal regions. These findings underscore the significance of battleground states in determining electoral outcomes, particularly in a highly polarized political landscape. We also discuss the model's performance, accuracy, and limitations, reflecting on the unique challenges inherent in forecasting electoral outcomes in these crucial areas.

1 Introduction

EDIT & ELABORATE - gpt generated The 2024 United States Presidential Election presents a unique opportunity to examine the dynamics of voter preferences in key battleground states. With a highly polarized electorate and the stakes at an all-time high, accurately predicting the outcome of this election is crucial for political analysts, campaigns, and the public. This paper develops a predictive model that leverages state-level polling data to forecast the likely

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

winner between Donald Trump and Kamala Harris, using a Bayesian hierarchical approach to accommodate variations across states.

The central estimand of this study is the probability of victory for each candidate in battleground states, defined by the binary outcome variable indicating a predicted win for Trump (1) or Harris (0). To achieve this, we utilize aggregated polling percentages as predictor variables, allowing us to capture shifts in voter sentiment over time. By employing logistic regression and accounting for state-specific random effects, our model not only identifies trends in polling data but also highlights critical regions that could determine the overall election outcome.

In our analysis, we focused on key battleground states—Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin—where the margins are often razor-thin and can swing the election in favor of one candidate or another. Our findings indicate that Trump is predicted to win in these crucial states, positioning him as the likely overall victor in the 2024 election. This is significant as it underscores the importance of battleground states in shaping the electoral landscape and highlights the intricate dynamics of voter preferences that can influence the final result.

The structure of this paper is as follows: we begin by detailing the data sources and methodology used for the analysis. Next, we present the results, focusing on the predicted probabilities of each candidate winning in the targeted battleground states. Finally, we discuss the implications of our findings and the limitations of our model, contributing to a better understanding of voter dynamics in the 2024 U.S. Presidential Election.

2 Data

2.1 Overview

This paper utilizes the dataset of the 2024 Presidential Election cycle, obtained from (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 pollster, relevant variables such as state, polling date, transparency score, candidate name, and percentage of support for the appropriate candidate 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'. 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 quality are further explained in {additional data details}. The electorate support for each candidate is denoted as 'pct'.

All of the variables mentioned above are used during the steps of data cleaning, simulation, testing, and analysis. We use the statistical programming language R [cite R]… to perform data simulation, testing, cleaning, and analysis.

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

To give an overview of the data, we provide summary statistics of the key variables, highlighting the characteristics of the polling support for each candidate across the battleground states.

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

Figure 1: Summary Statistics for Trump and Harris Polling Data. This table 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.

?@tbl-summarystatistics reveals a total of 34 polls conducted, with an average support level of 46.321% for Trump and 46.661% for Harris. Trump's support peaked at 51.000%, while the lowest recorded support was 42.500%. For Harris, the maximum support also reached 51.000%, with a minimum of 42.000%. These statistics illustrate the competitive landscape in these critical states, emphasizing the narrow margins that could ultimately influence the election outcome. The data reflects the volatility of voter sentiment leading up to the 2024 Presidential Election, underscoring the importance of these battleground states in determining the final result.

2.3.1 Outcome Variable

The outcome variable in our model is 'predicted_winner', which is a binary indicator, specifically constructed to represent the predicted winner of the 2024 U.S. Presidential Election based on polling data from key battleground states. This variable categorizes the election outcome into two distinct categories: a value of "Trump" indicates that Donald Trump is predicted to win, while a value of "Harris" denotes a predicted victory for Kamala Harris. The classification is derived from a conditioning statement that assigns the predicted winner based on the probability of Trump winning; if this probability exceeds 0.5, the predicted winner is labeled as Trump, otherwise, it is labeled as Harris.

The 'predicted_winner' variable in our model is a constructed binary indicator derived from the pct variable based on our model. The "pct" variable represents the percentage of support for each candidate, Harris and Trump, within each state. To create the "winner" variable, we compare the "pct" values: if Trump's percentage of support in a given state is higher than Harris's, the "winner" variable is assigned a value indicating Trump as the predicted winner for the given state; otherwise, it is assigned to Harris. This construction allows us to translate individual support levels from the raw data into a binary prediction of each state's likely outcome.

2.3.2 Predictive Variable

The predictive variables for this paper are each candidate's support rate ('pct'), which directly influences the "winner" variable. The 'pct' variable represents the percentage of the vote or support that each candidate received in the poll. Understanding these polling percentages is crucial, as they provide insight into the electoral landscape and help evaluate which candidate may hold an advantage in the upcoming election. The battleground states presented in the table are particularly significant due to their historical tendency to swing between parties, making them vital focal points for campaign efforts. By examining the outcome variable through these polling percentages, we enhance our predictive modeling and gain a deeper understanding of the factors influencing the election. This comparison not only contextualizes the current political climate but also highlights the critical role these battleground states play in determining the overall election results. As the campaign unfolds, these insights will be invaluable for strategizing and gauging voter sentiment in these pivotal areas.

```
#| echo: false
#| warning: false
#| message: false

# Aggregating polling percentages for Trump and Harris by state
state_average <- clean_poll_data %>%
  group_by(state) %>%
```

```
  summarise(
    Trump_pct = mean(Trump_pct, na.rm = TRUE),
    Harris_pct = mean(Harris_pct, na.rm = TRUE)
  )


# Prepare table for display with rounding to 1 decimal place
average_table <- state_average |>
  mutate(
    Trump_pct = round(Trump_pct, 1),
    Harris_pct = round(Harris_pct, 1)
  ) |>
  select(
    State = state,
    "Trump %" = Trump_pct,
    "Harris %" = Harris_pct,
  )

# Render the table
knitr::kable(average_table)
```

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

Summary of fixed effects estimates from the Bayesian hierarchical models for both Donald Trump and Kamala Harris. The table includes the estimated coefficients, standard errors, and 95% confidence intervals for each predictor, illustrating how various factors influence polling percentages. The significant intercept and time-related variables are crucial for understanding candidate support dynamics over the election period.

?@tbl-predictivevar highlights the average support levels for both Donald Trump and Kamala Harris across key battleground states, including Wisconsin, North Carolina, Georgia, Arizona, Pennsylvania, Michigan, and Nevada. The values reflect the polling percentages for each candidate, based on 'pct', providing a straightforward comparison of the predictive variables in each battleground state. 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%. In reality, as seen in Arizona, the combined 'pct' for Trump and Harris is 92.9%, with the remaining 7.1% likely reflecting support for other candidates in the race.

2.4 Other Variables

WRITE SECTION

3 Model

The goal of our modeling strategy is to predict the winner of the 2024 US Presidential Election in each state based on polling data, while also assessing the likelihood of Donald Trump or Kamala Harris winning. We use a Bayesian hierarchical modeling approach with natural splines to capture time-dependent trends in support and random effects to account for state-level variability. This strategy enables us to model changes in candidate support over time across states, allowing for more nuanced, probabilistic predictions of each candidate's support levels and overall election outcomes.

Separate models are fit for each candidate's polling support, with state-specific random intercepts and a natural spline for time, allowing the model to capture state-level deviations from the national average as well as time-based changes in support. This hierarchical Bayesian approach enables partial pooling across states, stabilizing predictions for states with limited polling data. Further background details, model specifications, and diagnostics are included in Appendix B.

3.1 Model set-up

The binary outcome variable y_i represents whether Donald Trump is predicted to lead in state i based on polling averages for both candidates over time. Here, $y_i = 1$ indicates a Trump lead in state i, and $y_i = 0$ indicates a lead for Kamala Harris. The model estimates the support levels for each candidate based on polling data using a Gaussian likelihood function, with time trends captured by a natural spline and state-level variation captured through random intercepts.

The predictor variables are $Trump\_pct_i$ and $Harris\_pct_i$, which represent the predicted percentages of support for Donald Trump and Kamala Harris in state i. A separate model is fit for each candidate's support over time, capturing the time-varying nature of support with a natural spline function and allowing for state-specific random effects.

The model is specified as follows:

[

$$
\begin{aligned}
\text{Trump\_pct}_i &\sim \text{Normal}(\mu_{\text{Trump},i}, \sigma_{\text{Trump}}) \\
\mu_{\text{Trump},i} &= \alpha_{\text{Trump}} + f(\text{end\_date\_num}_i) + u_{\text{state}}^{\text{Trump}} \\
u_{\text{state}}^{\text{Trump}} &\sim \text{Normal}(0, \tau_{\text{Trump}}) \\
\text{Harris\_pct}_i &\sim \text{Normal}(\mu_{\text{Harris},i}, \sigma_{\text{Harris}}) \\
\mu_{\text{Harris},i} &= \alpha_{\text{Harris}} + f(\text{end\_date\_num}_i) + u_{\text{state}}^{\text{Harris}} \\
u_{\text{state}}^{\text{Harris}} &\sim \text{Normal}(0, \tau_{\text{Harris}})
\end{aligned}
$$

]

Where:

- Trump_pct_i and Harris_pct_i are the polling support levels for Donald Trump and Kamala Harris in sta (1)
- $f(\text{end\_date\_num}_i)$ is a natural spline function of time (measured as end_date_num, with 4 degrees of (2)
- α_{Trump} and α_{Harris} are intercepts for the models, representing the baseline support levels for each candida (3)
- $u_{\text{state}}^{\text{Trump}}$ and $u_{\text{state}}^{\text{Harris}}$ are state-level random effects, accounting for differences in baseline support across sta (4)

For both models, we use weakly informative Normal priors. The intercepts $ *{Trump} $ and $* {Harris} $ are given Normal priors with a mean of 0 and a standard deviation of 10, reflecting prior uncertainty about the baseline support levels. The time trend coefficients are also given Normal priors with mean 0 and standard deviation 5, allowing for flexibility in capturing the temporal changes in support. State-level random effects for each candidate are modeled with a Normal distribution centered at 0, with the variance estimated from the data.

The model was implemented in R using the `brms` package, which allows for Bayesian inference using `Stan` as the backend. This approach captures the dynamic nature of polling support over time, while accounting for state-specific variability and allowing us to estimate the probability of Trump or Harris leading in each state.

3.2 Model Justification

We expect a positive relationship between each candidate's polling support over time and their likelihood of winning a given state. As Donald Trump's polling percentage in a state increases, his probability of leading in that state rises, while higher support for Kamala Harris increases her probability of winning. By modeling each candidate's support trends over time, we capture shifts in voter sentiment that reflect campaign dynamics, national events, and other time-dependent factors influencing support levels.

A Bayesian hierarchical model was chosen for its ability to handle the complexity of election forecasting, where both temporal trends and state-specific variations play significant roles. Bayesian inference allows us to incorporate prior beliefs and uncertainties about candidate support, while the hierarchical structure enables partial pooling across states. This is particularly useful for states with sparse data, as the model can "borrow strength" from the national trend, stabilizing predictions.

The inclusion of a natural spline for time allows us to model non-linear changes in support, capturing important shifts in polling trends as the election approaches. Random effects for each state account for baseline differences in support levels, reflecting the unique political climate of each state and reducing the risk of overfitting by allowing state-level deviations from the national average.

This model assumes that polling data reasonably reflects voter intentions and that each state's outcome can be modeled independently of others, though regional trends are partially addressed by state-level random effects. Potential biases inherent in polling data, such as underrepresented demographics or sampling error, could affect our predictions. The model was implemented in R using the brms package, which provides Bayesian inference via Stan, allowing for flexible, probabilistic predictions of each candidate's support across states.

3.3 Model Summary - Harris Model

Table 3: 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

EDIT - explain model summary table

3.4 Model summary - Trump Model

Table 4: 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

EDIT - explain model summary table

4 Results

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

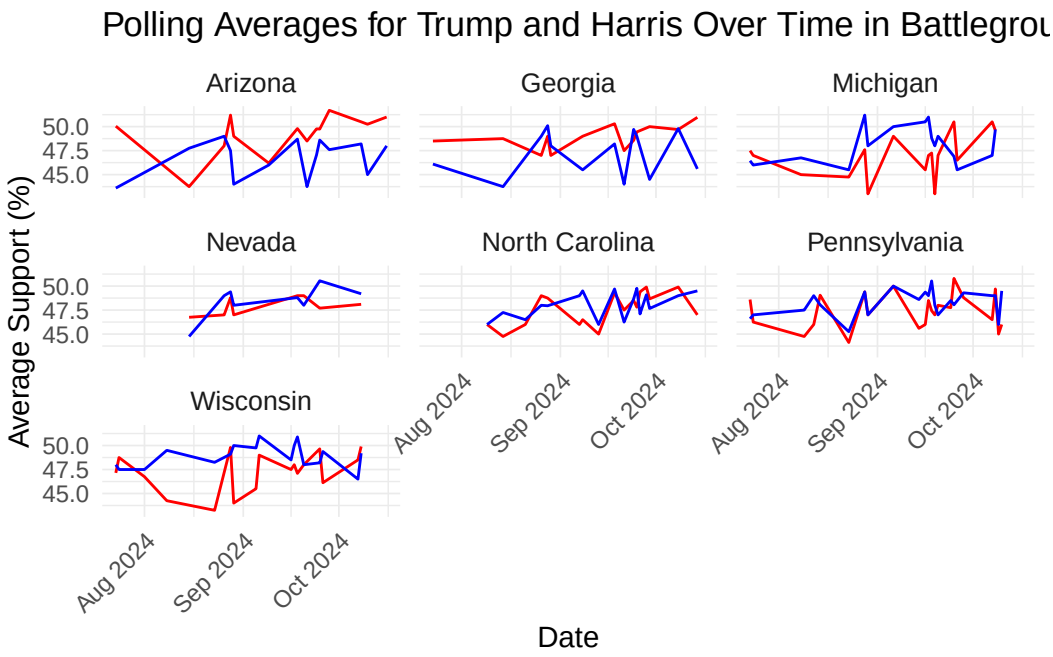

Figure 2: Polling Averages for Trump and Harris Over Time in Battleground States.

EDIT - explain results on graph

4.2 Probability of Winning by State on Election Day

Table 5

State	Predicted Harris Win (%)	Predicted Trump Win (%)	Predicted Winner
Arizona	49.48	49.23	Donald Trump
Georgia	49.31	49.51	Kamala Harris
Michigan	48.01	50.28	Kamala Harris
Nevada	48.85	50.28	Kamala Harris
North Carolina	48.49	50.05	Kamala Harris
Pennsylvania	48.21	50.30	Kamala Harris
Wisconsin	48.30	50.76	Kamala Harris

Probability of Trump Winning by State on Election Day.

EDIT - explain table results

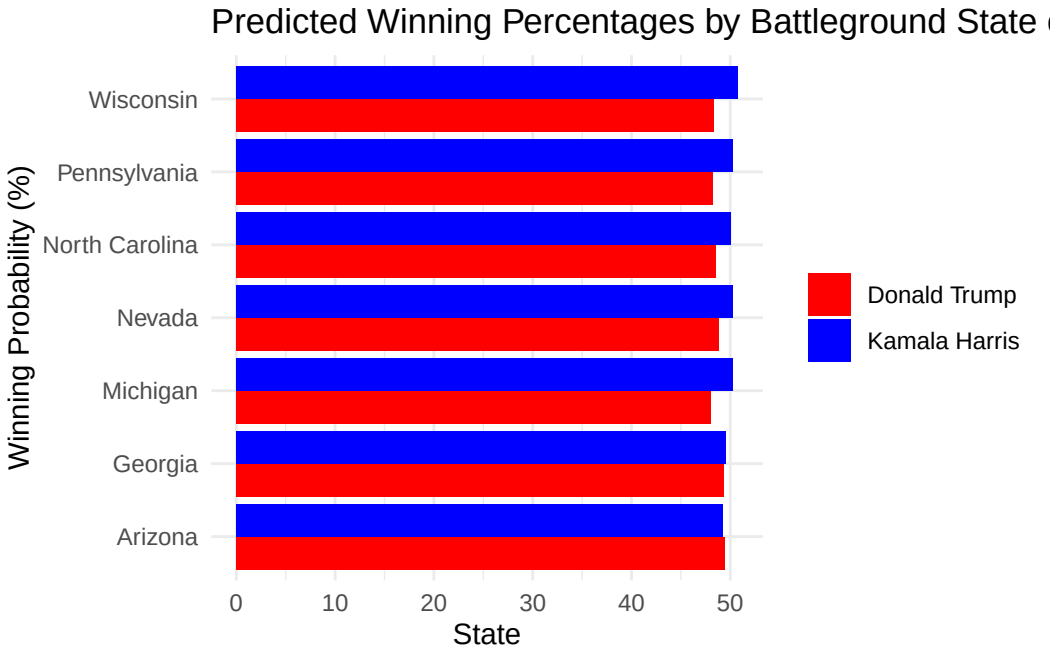

Figure 3: Probability of Trump Winning by State on Election Day.

EDIT - explain graph results

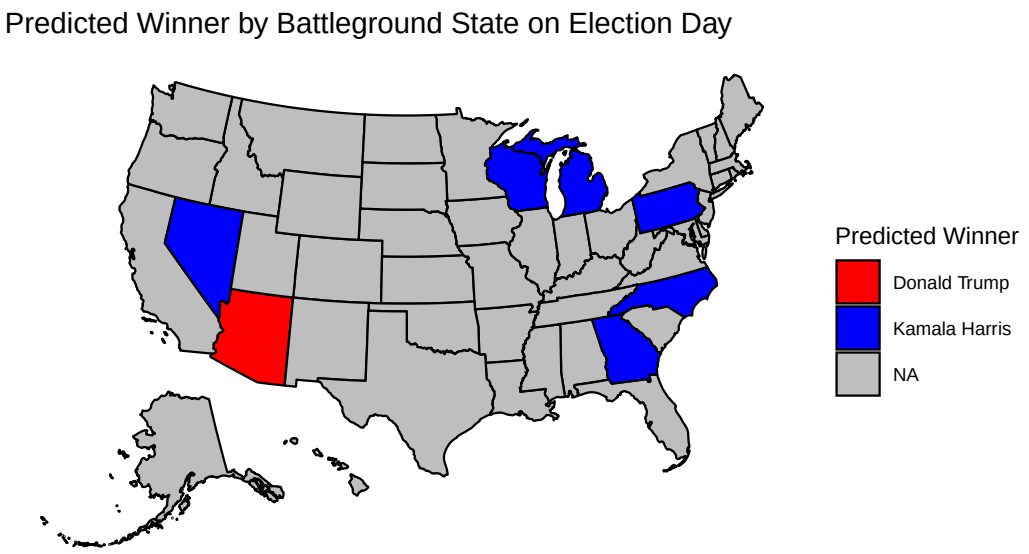

Figure 4: Predicted Winner by Battleground State on Election Day.

4.3 US Map Showing Predicted Winner by State on Election Day

EDIT - explain map results

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}

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.