Forecasting the 2024 Election: Trump Dominates the South, Harris Leads in the Midwest*

Jimin Lee Sarah Ding Xiyan Chen

October 21, 2024

This paper presents a predictive model for the 2024 United States Presidential Election using state-level polling data. By aggregating high-quality polls, we predict the likely winner between Donald Trump and Kamala Harris in each state. A logistic regression model estimates the probability of Trump winning, with polling percentages as predictor variables. The model's results highlight key battleground states and offer insights into voter dynamics, while also reflecting on the limitations of polling data in forecasting political outcomes.

1 Introduction

This paper develops a predictive model for the 2024 US Presidential Election using state-level polling data (FiveThirtyEight 2024) to forecast the likely winner between Donald Trump and Kamala Harris. We aggregate high-quality poll data to create a logistic regression model that estimates the probability of a Trump or Harris victory in each state. By comparing polling percentages for both candidates, we aim to predict election outcomes based on voter support patterns across the states.

The estimand in this study is the probability that Donald Trump or Kamala Harris wins a given state, based on aggregated state-level polling averages. The binary outcome variable in our model indicates whether Trump (1) or Harris (0) is predicted to win in each state, and the predictor variables are the average polling percentages for each candidate.

Our model utilizes the Bayesian logistic regression model to predict election outcomes by comparing the average polling percentages for Trump and Harris in each state. The results

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

highlight the geographic distribution of support, with some states clearly favoring one candidate over the other. Swing states emerge as critical battlegrounds, where polling percentages are closely contested and could influence the final election result.

Accurate election predictions provide valuable insights into voter dynamics and help political analysts, campaigns, and the public anticipate electoral outcomes. By focusing on high-quality polling data, our model improves the reliability of predictions and highlights key regions where voter sentiment may shift, ultimately affecting the election.

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

We use the statistical programming language R (R Core Team 2023) to perform data cleaning and analysis using the datasets we obtained from the Project 538 online database (FiveThirtyEight 2024). The database provides a wealth of information on the current cycle of presidential general election polls. It consists of the polling results from various pollsters such as the New York Times, ActiVote, Morning Consult, etc. Information about the transparency of methodology as well as the quality of the data is present as well, denoted as transparency score and numeric grades, which are significant to our data cleaning process to ensure the quality of data. Polling results are specific to different states in the US, mostly focusing on the six battleground states that are the likeliest to determine the outcome of the elections. Thus, we chose to use states, candidate names, and percentage of support to perform statistical analysis and gather our results.

2.2 Measurement

The measurement of voter support in the dataset is derived from raw polling data obtained from the Project 538 online database (FiveThirtyEight 2024), which serves as a representation of potential election results for the 2024 U.S. Presidential Election. Each entry in the data corresponds to a poll conducted by each pollster, capturing the percentage of respondents who express support for Donald Trump or Kamala Harris. Each pollster has their approach of sampling respondents as well as their methodology of approaching respondents, which affects each pollster's numeric grades and transparency scores.

Therefore, to ensure the quality of the data, only polls from reputable pollsters with a numeric grade of 3 or higher and transparency scores above 6 were used for our paper. These polls measure voter preferences through structured survey questions, and the results are aggregated to create state-level averages for each candidate. The dataset thus transforms general voter sentiment into individual data points that reflect the competitive dynamics between Trump and Harris across different states.

2.3 Outcome variables

The outcome variable in our analysis is the binary variable winner, which represents the predicted winner of the 2024 US Presidential Election in each state. The value of winner is set to 1 if Donald Trump is predicted to win the state, and 0 if Kamala Harris is predicted to win. This binary outcome is determined by comparing the polling support for both candidates within each state. By setting up this binary variable, we aim to forecast which candidate will secure more votes in each state based on the aggregated polling data.

Figure 1 provides an overview of the average support for Trump and Harris across all states, along with the predicted winner for each state based on our model.

state	Trump_pct	Harris_pct	winner	predicted_winner
Arizona	46.48000	45.33333	1	Trump
Colorado	40.00000	NaN	NA	NA
Florida	53.00000	40.50000	1	Trump
Georgia	46.55000	44.36364	1	Trump
Michigan	44.52000	46.26667	0	Harris
Missouri	54.00000	41.00000	1	Trump
Montana	56.50000	39.00000	1	Trump
Nebraska	42.00000	50.75000	0	Harris
Nevada	46.64286	43.83333	1	Trump
North Carolina	46.81818	46.81818	0	Harris
Ohio	49.50000	44.16667	1	Trump
Pennsylvania	45.14474	47.00000	0	Harris
Texas	47.06250	43.57143	1	Trump
Virginia	41.00000	45.25000	0	Harris
Wisconsin	45.18000	48.81481	0	Harris
NA	44.45781	46.58824	0	Harris

Figure 1: Summary Table of Polling Averages by State

2.4 Predictor variables

The predictor variables used in the model are the aggregated polling percentages for Donald Trump (Trump_pct) and Kamala Harris (Harris_pct). These variables are calculated as the average support for each candidate, using polling data filtered to include only high-quality pollsters with a numeric grade of 3 or higher and transparency scores above 6. These averages reflect the level of support for each candidate across all polls in each state.

To further illustrate the distribution of support, a side-by-side bar graph is provided, comparing the average polling percentages for both candidates across all states. Figure 2 offers a clear comparison of the support levels, helping to visualize the competitive dynamics within each state. The side-by-side comparison represents the percentage of support each candidate has received, based on aggregated poll data from pollsters with high-quality scores. Trump's support is shown in red, while Harris's support is depicted in red.

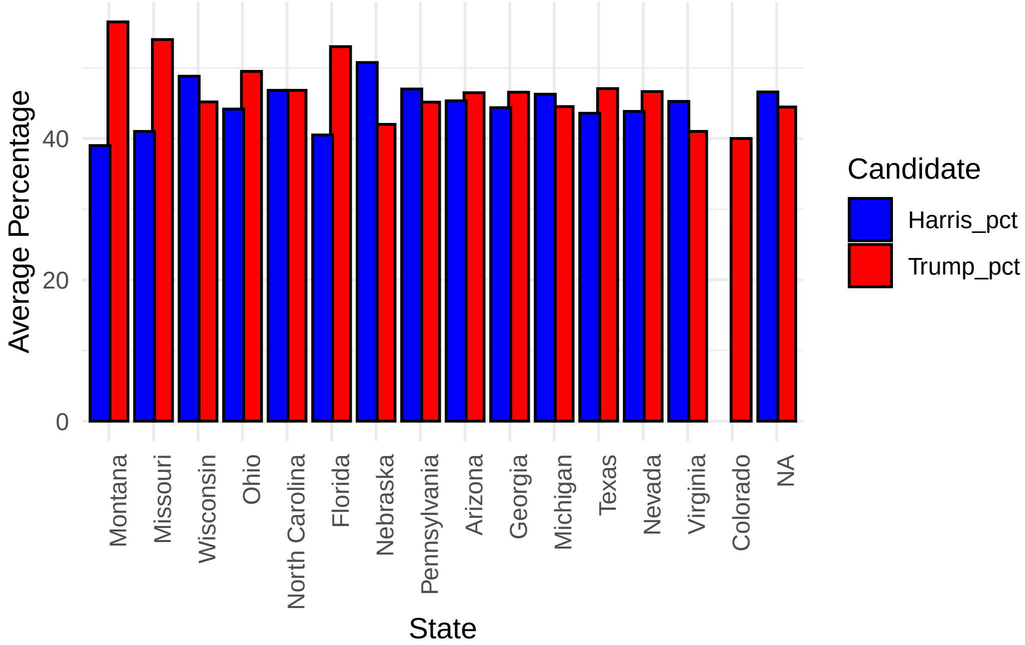

Figure 2: Average Poll Percentage by State

The comparison shows significant variability in support for each candidate. In some states like **Montana** and **Missouri**, Trump's support is notably higher, with a clear margin over Harris. In contrast, **Colorado** stands out as having nearly equal polling support for both candidates, suggesting a closely contested race in that state. Swing states such as **Florida**, **North Carolina**, and **Pennsylvania** exhibit relatively close polling percentages, indicating that the election outcomes in these states could be pivotal and difficult to predict. The plot also includes states like **Virginia** and **Texas**, where the support levels are more balanced, though Trump seems to hold a slight lead. The state labeled as **NA** at the end may indicate missing or incomplete data, emphasizing the need for comprehensive polling coverage in all states for more reliable predictions.

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 employ a Bayesian logistic regression model, which allows for probabilistic predictions of binary outcomes, making it well-suited for the task of predicting election results. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

The binary outcome variable y_i is defined as 1 if Donald Trump is predicted to win state i, and 0 if Kamala Harris is predicted to win. The predictor variables are β_i, which represents the average percentage of polling support for Donald Trump in state i, and γ_i, the average percentage of support for Kamala Harris in state i. The relationship between these variables and the outcome is captured through a logistic link function.

$$\begin{aligned} y_i \mid \mu_i &\sim \text{Bernoulli}(\mu_i) \\ \mu_i &= \alpha + \beta_i + \gamma_i \\ \alpha &\sim \text{Normal}(0, 2.5) \\ \beta_i &\sim \text{Normal}(0, 2.5) \\ \gamma_i &\sim \text{Normal}(0, 2.5) \end{aligned}$$

We use weakly informative Normal priors for the intercept (α_i) and the coefficients (β_i) and γ_i), centered around zero with a standard deviation of 2.5. These priors were chosen to reflect the assumption that, before seeing the data, there is no strong reason to favor one candidate over the other across states, but to allow room for the data to adjust predictions as polling percentages differ. We run the model in R (R Core Team 2023) using the `rstanarm` package (Gabry et al. 2023). The model uses default priors from `rstanarm`, and estimates the likelihood of Donald Trump winning each state based on polling averages for both candidates.

3.2 Model Justification

We expect a positive relationship between polling support for each candidate and their probability of winning a state. As Donald Trump's polling percentage (β_i) increases, the likelihood of him winning state i rises. Similarly, as Kamala Harris's polling percentage (γ_i) increases, her chances of winning that state increase.

Bayesian logistic regression was chosen for its ability to model binary outcomes like win/loss predictions while incorporating prior knowledge and uncertainty, helping mitigate overfitting through regularization with priors. This model strikes a balance between flexibility and interpretability, given that polling percentages are continuous and the outcome is binary. The model assumes polling data accurately reflects voter intentions and that state outcomes are independent of each other, although it does not account for interactions between states or regional voting trends. Potential biases in polling data could also affect predictions. The model was implemented using the 'rstanarm' (Gabry et al. 2023) package in R (R Core Team 2023), which supports Bayesian inference with default priors for logistic regression.

4 Results

4.1 Predicted Election Winner By State

Figure 3 shows the predicted winner of the 2024 US Presidential Election by state on the US map (Albers et al. 2023), based on a logistic regression model using aggregated polling data for Donald Trump and Kamala Harris. States where Trump is predicted to win are shown in red, while states where Harris is predicted to win are displayed in blue. States in gray represent missing or incomplete polling data, where predictions were not possible.

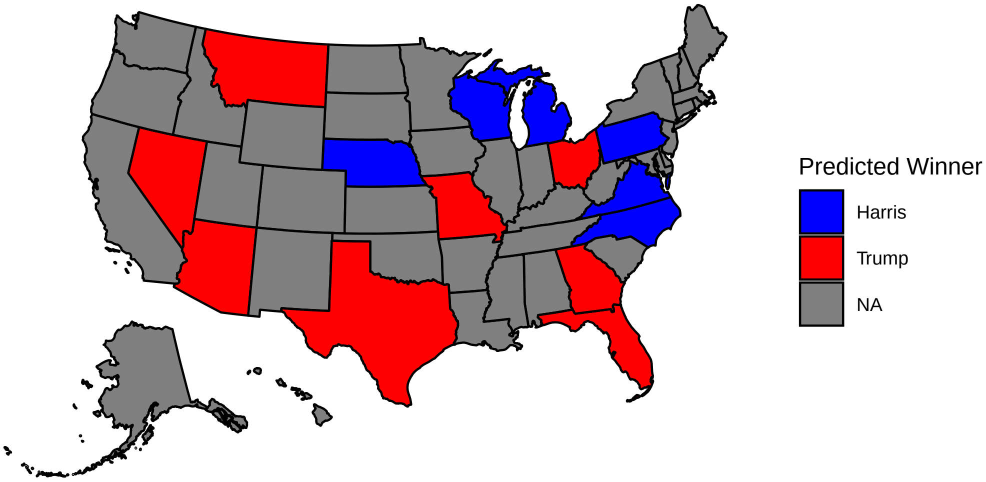

Figure 3: Predicted Winner of 2024 US Presidential Election by State

According to the model, Trump is expected to win key states, including Florida, Missouri, and Montana, while Harris is forecasted to lead in Michigan, Virginia, and Pennsylvania. The results highlight the geographic divide between the two candidates. Trump is predicted to perform well in southern and midwestern states, where his polling averages are significantly higher. Harris, on the other hand, is projected to win in the Midwest and along the East Coast, with stronger polling percentages in traditionally Democratic-leaning regions.

4.2 Summary of Predicted Election Results by State

Table 2 further breaks down the polling data by state, showing the average polling percentages for each candidate and the predicted win probability for Trump. For example, in Florida, Trump's polling average is 53%, and his predicted win probability is almost certain at 99.93%. Conversely, in Michigan, Harris leads with an average of 46.27% of the vote, resulting in a 96.77% predicted probability that she will win the state. In swing states like North Carolina and Nevada, where the polling percentages are closer, Trump is still predicted to win with probabilities of 94.36% and 51.23%, respectively. These battleground states show tighter polling averages and thus greater uncertainty in the predicted outcomes.

Table 2: Predicted 2024 US Presidential Election outcomes by state, based on polling percentages for Trump and Harris

State	Trump Polling (%)	Harris Polling (%)	Predicted Trump Win Probability	Predicted Winner
Arizona	46.48000	45.33333	0.5167401	Trump
Florida	53.00000	40.50000	0.9999069	Trump
Georgia	46.55000	44.36364	0.8805457	Trump
Michigan	44.52000	46.26667	0.0345724	Harris
Missouri	54.00000	41.00000	0.9998728	Trump
Montana	56.50000	39.00000	0.9999780	Trump
Nebraska	42.00000	50.75000	0.0001831	Harris
Nevada	46.64286	43.83333	0.9465966	Trump
North Carolina	46.81818	46.81818	0.1361986	Harris
Ohio	49.50000	44.16667	0.9872218	Trump
Pennsylvania	45.14474	47.00000	0.0190336	Harris
Texas	47.06250	43.57143	0.9767476	Trump
Virginia	41.00000	45.25000	0.0759898	Harris
Wisconsin	45.18000	48.81481	0.0046061	Harris
NA	44.45781	46.58824	0.0200494	Harris

The results of our predictive model suggest that Donald Trump is likely to win key states such as Florida, Missouri, and Montana, while Kamala Harris is expected to secure victories in states like Michigan, Virginia, and Pennsylvania. Close contests are predicted in battleground states like Nevada and North Carolina, where Trump holds a slight edge.

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.

Additionally, during our data cleaning process, we added constraints of high numeric grades as well as high transparency scores which led to fewer data points for us to build the model on. For example, for the state of Colorado, when we set the transparency score to higher than 5, and the numeric grade to larger and equal to 3, there are no data points for Kamala Harris. This is represented by the empty row in Figure 1.

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

Figure 4 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.

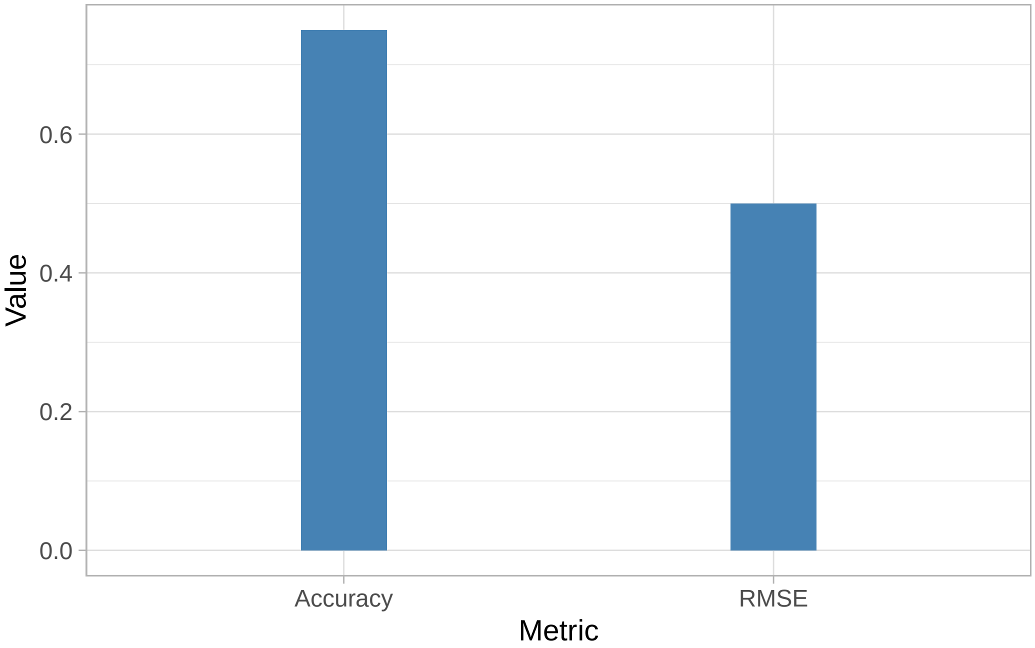

Figure 4: Model: Diagnostics: Accuracy and RMSE

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

Table 3 displays the model's predictive performance in distinguishing between state-level wins for Trump (coded as 1) and Harris (coded as 0).

Table 3: Confusion Matrix of Test Predictions

	Actual 0 (Harris)	Actual 1 (Trump)
Predicted 0 (Harris)	2	0
Predicted 1 (Trump)	1	1

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 **?@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/.

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.