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 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, which serves as a representation of potential election results for the 2024 U.S. Presidential Election. Each entry in the dataset 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.

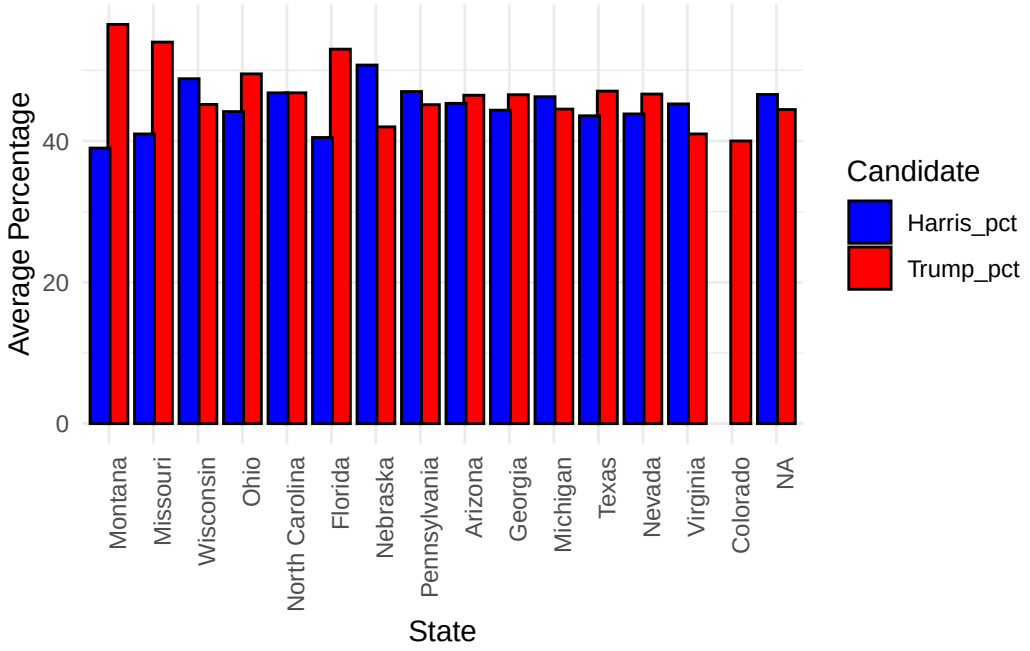

Figure 2: Average Poll Percentage by State

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.

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

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 package in R, 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 a 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.

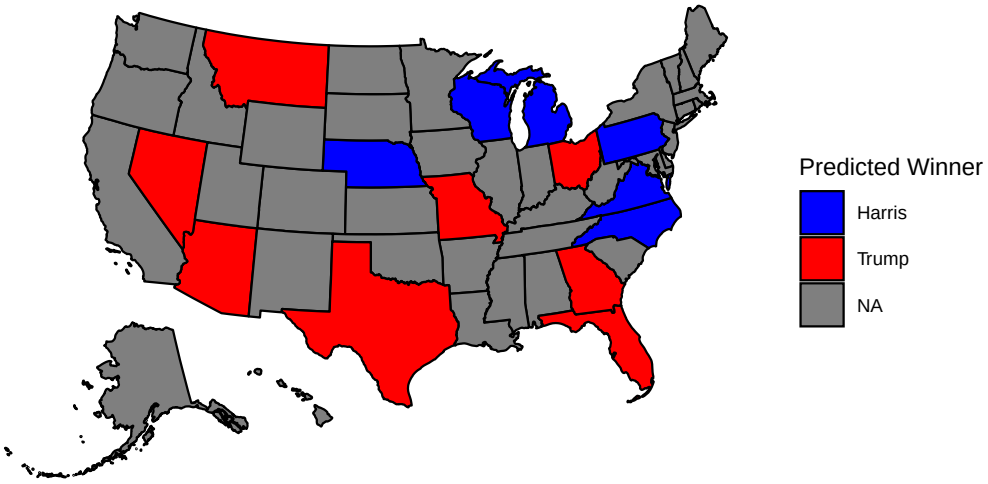

Figure 3: Predicted Winner of 2024 US Presidential Election by State

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

Placeholder - The results of our model reveal clear geographic divides in voter sentiment, with Donald Trump predicted to win in many southern and midwestern states, while Kamala Harris is expected to perform well in the Midwest and on the East Coast. The predictions align with historical voting patterns, with Trump's strongest support coming from traditionally conservative regions and Harris garnering more support in Democratic-leaning areas. Notably, several battleground states, such as Nevada and North Carolina, show a close contest, reflecting the competitive nature of the 2024 election. These results suggest that the election may hinge on the outcomes in these key states, where even small shifts in polling percentages could influence the final result. The predicted probabilities in states like Michigan, Virginia, and Pennsylvania indicate a strong chance for Harris, while Trump's significant lead in states like Florida and Missouri solidifies his dominance in certain regions. Overall, the model provides valuable insights into the competitive landscape, highlighting the states that could be decisive in determining the election outcome.

5.2 Model Performance

Placeholder - The logistic regression model used in this analysis successfully captures the relationship between polling percentages and the likelihood of a candidate winning a state. The coefficients for the predictor variables (polling percentages for Trump and Harris) indicate a strong relationship between candidate support and the probability of winning. The model outputs probabilities for each state, which are useful for gauging the certainty of predictions, especially in closely contested states. The model's performance is reinforced by the robustness of the results in states where one candidate has a clear polling lead, providing accurate predictions. However, states with narrower margins reflect greater uncertainty, which is critical in understanding election dynamics. Convergence diagnostics and model checking were conducted to ensure reliable estimates, with further details and alternative models provided in Appendix B.

5.3 Weaknesses and Next Steps

Placeholder - Despite the model's predictive power, there are several weaknesses that should be noted. First, polling data itself is prone to biases, such as underrepresentation of certain demographic groups or inaccuracies in sampling. These issues can distort the predictions, particularly in states with fewer or lower-quality polls. Additionally, voter turnout and last-minute shifts in public opinion are not accounted for, which can significantly alter the outcome, especially in close races. The model also assumes that polling percentages directly translate to votes, overlooking other factors like voter mobilization efforts and campaign dynamics that may influence election outcomes.

Placeholder - To improve the accuracy and reliability of election forecasts, future iterations of the model could incorporate additional variables beyond polling data. Factors such as voter turnout models, economic conditions, and social media sentiment analysis may provide a more holistic view of voter behavior. Additionally, real-time polling data integration could allow for dynamic updates to predictions as the election nears, improving the timeliness and accuracy of forecasts. Moreover, regional adjustments to account for differences in polling methodology or demographic shifts would help to address some of the biases inherent in polling data. Expanding the model to include these aspects would offer a more comprehensive understanding of the election landscape and enhance the precision of predictions.

Appendix

A Additional data details

It is important to note that several states remain uncolored (in gray), indicating missing or incomplete polling data for those states. These uncolored states reflect the absence of reliable or sufficient polling information required for an accurate prediction. This underscores the limitations of the data used in the model, where predictions are only made for states with adequate polling coverage. As a result, certain regions lack predictions, which could be addressed by including more comprehensive polling data in future analyses.

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 overfitting 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.

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

# Out-of-Sample Testing of Our Model
# Load necessary libraries
library(caret)
```

```
Loading required package: lattice
```

```
Warning: package 'lattice' was built under R version 4.2.3
```

```

Attaching package: 'caret'
```

```
The following objects are masked from 'package:rstanarm':

    compare_models, R2

The following object is masked from 'package:purrr':

    lift
```

```r
library(rstanarm)
library(ggplot2)
library(dplyr)


# Split the data into training (70%) and test (30%) sets
set.seed(123)  # For reproducibility
trainIndex <- createDataPartition(state_average$winner, p = 0.7, list = FALSE)
train_data <- state_average[trainIndex, ]
test_data <- state_average[-trainIndex, ]

# Train the Bayesian logistic regression model on the training data
logistic_model_train <- stan_glm(winner ~ Trump_pct + Harris_pct, data = train_data,
                                 family = binomial, prior = student_t(df = 7),
                                 chains = 4, iter = 2000, refresh = 0, verbose = FALSE)

# Make predictions on the test set
test_predictions <- posterior_predict(logistic_model_train, newdata = test_data, draws = 1)

# Convert predictions to binary outcome (1 if Trump wins, 0 if Harris wins)
test_pred_binary <- ifelse(colMeans(test_predictions) > 0.5, 1, 0)

# Calculate accuracy
accuracy <- mean(test_pred_binary == test_data$winner)
accuracy_message <- paste("Test Accuracy: ", round(accuracy * 100, 2), "%")
print(accuracy_message)
```

```
[1] "Test Accuracy:  75 %"
```

```r
# Optionally, calculate RMSE
rmse_value <- sqrt(mean((test_pred_binary - test_data$winner)^2))
rmse_message <- paste("Test RMSE: ", round(rmse_value, 4))
print(rmse_message)
```

```
[1] "Test RMSE:  0.5"
```

```r
# Confusion Matrix for additional diagnostics
conf_matrix <- table(Predicted = test_pred_binary, Actual = test_data$winner)
print(conf_matrix)
```

```
         Actual
Predicted 0 1
        0 2 0
        1 1 1
```

Our accuracy test generated a 75% score for the model, which means that our model is quite reliable 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. correctly 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.

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/.

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