Race to the Ballot Box: Will Demographics Decide a Winner This Year?*

A Regressional Approach to Race and Political Preferences in the 2024 US Elections and Insights on the Next

Sameeck Bhatia Sean Chua Tanmay Shinde

November 2, 2024

This paper analyzes polling data for the 2024 U.S. Presidential election to predict the winner, while also capturing demographic sentiment toward the two main contenders - Donald Trump and Kamala Harris. We found that ….[NEED TO COMPLETE]. This study highlights that considering factors such as the reputation of the pollster, quality of the poll, bias, sample size, and demographic area are essential for analyzing polling data and thereby forecasting election outcomes. Highlighting these trends is crucial for understanding the evolving political landscape in the U.S., as they reveal how voter demographics and sentiments shape election outcomes.

Table of contents

1 Introduction **1**

2 Data **2**
2.1 Overview 2
2.2 Measurement 3
2.3 Outcome variables 4
2.4 Predictor variables 4

3 Model **4**
3.1 Overview 4
3.2 Model Structure 4
3.3 Predictors and Weights 5

*Code and data are available at: https://github.com/SameeckBhatia/2024-US-Elections.

3.4 Priors . . . 5
3.5 Assumptions and Limitations . . . 5
3.6 Software Implementation . . . 5
3.7 Model Validation . . . 6
3.8 Alternative Models . . . 6
3.9 Interpretation . . . 6

4 Results **6**

5 Discussion **11**
5.1 Harris Wins the Swing States and Overcomes the Bradley Effect . . . 11
5.2 More Urban States Tend to Lean Harris . . . 11
5.3 Limitations and Areas for Future Research . . . 12

Appendix **13**

A Additional Data Details **13**

B Additional Model Details **14**

C Additional Results Details **15**
C.1 Null Values . . . 15
C.2 Prediction Errors . . . 15

D Polling Methodology Overview and Evaluation for Siena/NYT **15**

E Idealized Methodology and Survey **16**
E.1 Sampling Approach and Strategy . . . 16
E.1.1 Stratification Variables [TO-DO] . . . 17
E.2 Budget Allocation . . . 17
E.2.1 **Survey Structure**: . . . 17
E.3 Data Validation . . . 18

References **19**

1 Introduction

As the 2024 U.S. Presidential election draws near, the race between former President Donald Trump and Vice President Kamala Harris has garnered widespread attention worldwide. With both candidates representing contrasting ideologies and policy priorities, appealing to distinct voter bases, understanding the public sentiment surrounding them becomes crucial in predicting the election's outcome. Polls and demographic analyses provide a lens through which we can examine the factors influencing voter preferences and gauge the political landscape

as it evolves. This paper aims to analyze the 2024 election polling data to not only forecast the winner but also to capture the demographic-wise public sentiment surrounding the two contenders.

The primary estimand for this analysis is the predicted support percentage for each of the two candidates, which represents the proportion of the electorate that is expected to favor one candidate over the other. This estimand will be calculated based on survey data or polling results, and it will allow us to estimate the level of support each candidate has within the population. This analysis will also provide insights into how various demographic factors or voter preferences may influence the final outcomes.

We use polling data from FiveThirtyEight, focusing on surveys from organizations such as YouGov, Siena/NYT, and Washington Post. We used a Bayesian approach to model the polling support percentages for both candidates, and understand how polling factors such as numeric grade, poll score, state, and sample size affect the predicted support percentage.

The results show that … [TO-DO]. These results are important because … [TO-DO].

The rest of the paper is structured as follows: Section 2 details the data and measurement process. Section 3 presents the model and justifies the choices made in the building of the chosen model. Section 4 presents the results, highlighting the relationship between different variables and the polling support percentage, and Section 5 discusses the implications of the findings for this as well as future election forecasting.

2 Data

2.1 Overview

The dataset used for this paper is the 2024 national-level presidential general election dataset from FiveThirtyEight (FiveThirtyEight 2024), which compiles polling data from various reputable pollsters, including YouGov, Siena/NYT, Ipsos, and more. This dataset offers a comprehensive snapshot of public opinion leading up to the 2024 U.S. Presidential election, capturing support for the two main contenders: Donald Trump and Kamala Harris. By focusing on national polling, this dataset provides insight into voter sentiment across a wide range of demographic groups and regions.

The FiveThirtyEight dataset is part of a broader landscape of polling data used in electoral predictions. It includes surveys conducted at the national level, focusing on general election matchups, and is continuously updated as new polls are released. The dataset covers polls taken from early in the election cycle to the final days before the election, capturing the shifts in public opinion over time. While there are other polling datasets available, such as those from RealClearPolitics (RealClearPolitics 2024) or individual pollsters, we selected the FiveThirtyEight dataset due to its high level of transparency, detailed methodology grades for each pollster, and advanced polling aggregation techniques, which adjust for biases and

historical pollster performance. These features provide a more refined and reliable dataset compared to others that might not incorporate such rigorous adjustments.

We use the statistical programming language R (R Core Team 2023) to perform data cleaning and analysis on the dataset to ensure consistency and reliability. Polls were filtered to include only those related to Donald Trump and Kamala Harris, with a pollster grade of 2.8 or higher, and those with a negative pollscore, ensuring reliability and unbiasedness. The dataset was further narrowed by selecting polls conducted on or after July 21, 2024, to focus on the shift in sentiment over time after the Democratic Party announced Harris as their nominee for president in the 2024 elections.

2.2 Measurement

The process of converting real-world electoral phenomena into data involves careful measurement of public sentiment toward presidential candidates, in the context of the 2024 U.S. presidential election. The dataset is built from polling data collected by various established pollsters, such as YouGov and Siena/NYT, who use a variety of survey techniques to measure voter preferences. These polls aim to capture critical variables that reflect voter sentiment, such as the percentage of respondents favoring each candidate, the sample size of the poll, and the demographics of the voters surveyed.

The measurement begins with pollsters designing surveys aimed at accurately reflecting voter behavior and sentiment. A subset of the population is sampled using stratified random sampling to ensure that the sample accurately represents the diverse demographic makeup of the electorate. Responses are collected through various channels, including phone calls, online surveys, and face-to-face interviews. These polls capture voter sentiment around a candidate as the percentage of respondents favoring the candidate.

Once the polling is done, the dataset undergoes extensive cleaning and processing to handle missing values and any discrepancies in the data. Each row in the dataset represents the outcome of a poll, including details such as the date it was conducted, the pollster's methodology, and the percentage of support for each candidate. Platforms like FiveThirtyEight then assign a numeric grade and poll score to each poll to evaluate the reliability, quality, and historical accuracy of their polling methods. The numeric grade is a standardized rating assigned to pollsters based on their historical performance, transparency, and polling methodology. The poll score measures the accuracy, consistency, and bias in each poll, pollsters that show bias toward a particular party receive a positive score.

Thus, each entry in the dataset corresponds to the results of a specific poll, providing information on when it was conducted, the pollster's methodology, and key metrics such as biasdness, sample size, and geographic coverage. By converting complex voter opinions into structured data, we are able to systematically analyze trends in public sentiment, making it possible to predict the election's outcome with greater precision.

2.3 Outcome variables

We aim to predict our response variable `candidate_trump`, which is a binary variable with 0 and 1 representing a loss and win in a state respectively. The plot of our outcome variables is shown below:

TO-DO: It is important to understand what the variables look like by including graphs, and possibly tables, of all observations, along with discussion of those graphs and the other features of these data. Summary statistics should also be included, and well as any relationships between the variables.

2.4 Predictor variables

TO-DO: All variables should be thoroughly examined and explained. [PLOTS OF PREDICTOR VARIABLES NEEDED] - **Sample Size (`sample_size`)**: The number of respondents in the poll. - **State (`state`)**: A categorical variable for different U.S. states. - **Days to Election (`days_to_election`)**: Counts the number of days until Election Day starting from today's date (it also uses `end_date` from the original dataset). - **Voter Percentage of Candidate (`pct`)**: A continuous variable for the percentage of voters who plan to vote for Trump based on polling data.

3 Model

3.1 Overview

This section describes a predictive Bayesian model for estimating voter support for two candidates—Donald Trump and Kamala Harris—in the 2024 U.S. presidential election. The model separately estimates the proportion of support for each candidate in 37 states, including both swing states (e.g., Georgia, Arizona) and strongholds (e.g., California, Texas). The "Trump" model and "Harris" model each output the probability that a voter supports Trump or Harris, respectively, based on pollster reliability, state, and polling percentages.

3.2 Model Structure

The Trump and Harris models share a similar structure, represented by the following equations:

- **Trump Model**:

$$\text{predicted\_pct\_trump} = \beta_0 + \beta_1 \cdot (1|\text{pollster}) + \beta_2 \cdot (1|\text{state}) + \beta_3 \cdot \text{pct} \tag{1}$$

- **Harris Model**:

$$\text{predicted\_pct\_harris} = \beta_0 + \beta_1 \cdot (1|\text{pollster}) + \beta_2 \cdot (1|\text{state}) + \beta_3 \cdot \text{pct} \tag{2}$$

Here, β_0, β_1, β_2, and β_3 are the model coefficients. Here β_0 is the intercept or the baseline percentage, and β_1, β_2 are the mean percentage point change in the predicted percentage, while β_3 is the mean change in the response for a one percentage point change in the polling percentage. For each model, the indicator variable (1 if the candidate is Trump for the "Trump" model and Harris for the "Harris" model, 0 otherwise) represents the response, with pollster, state, and polling percentage as predictors. Random effects for "pollster" account for polling agency bias, and random effects for "state" capture state-level variations in voter support.

3.3 Predictors and Weights

- **Pollster:** Different polling agencies show varying levels of reliability. We assign weights to pollsters based on their sample sizes and a numeric grade (0 to 3, with 3 being the most reliable).
- **State:** By including states as a predictor, the model considers unique voter patterns by location, essential for identifying swing states and strongholds.
- **Polling Percentage (pct):** The polling percentage represents the actual polling data, which directly informs our estimate of voter support for each candidate.

3.4 Priors

For this Bayesian model, we use a normal prior with a mean of 0.5 (50%) and a standard deviation of 0.25 (25%). This prior assumes a close race between Trump and Harris, allowing for a reasonable degree of uncertainty without overconfidence in either direction. These parameters reflect the consensus of a competitive 2024 election.

3.5 Assumptions and Limitations

The model assumes additional candidates besides Trump and Harris, addressed by separate models for each candidate. However, this dual-model approach may introduce error by not considering third-party effects directly. Furthermore, the model is limited to the 37 states included and cannot generalize predictions for other states (e.g., Idaho, Vermont).

3.6 Software Implementation

The model was developed in R using the rstanarm package for Bayesian modeling, including priors, posteriors, and predictions.

3.7 Model Validation

Model validation and performance metrics (F1 Score ~ 0.74, RMSE ~ 0.4) are available in `scripts/modeling.R`. A 70/30 train-test split was applied, with testing conducted exclusively on out-of-sample data. These scores indicate a reasonable balance of precision and recall, appropriate for the model's simplicity and data limitations.

3.8 Alternative Models

Alternative models considered included a single binary model (Trump vs. Harris) and a version with additional pollscore bias adjustment. While these models offer unique insights, they either oversimplify the election's dynamics or risk overfitting. The final model balances simplicity with predictive accuracy.

3.9 Interpretation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut elit leo, viverra ac enim non, pulvinar aliquam nisl. Nam accumsan ac nisi at consectetur. Curabitur luctus lacus eget risus rutrum, vel facilisis urna feugiat. Ut ac dictum velit, in blandit augue. Fusce sit amet vehicula dui. Suspendisse eleifend tempor rhoncus. Cras vehicula, nisl et molestie pulvinar, lorem ex dictum augue, nec mollis leo sem quis mauris.

4 Results

The table below summarizes the predicted support percentages for Donald Trump and Kamala Harris across the states where polling data is available. Due to the model being trained on a subset of the cleaned data and the specific approach used to split training and testing data, there are some missing values for each candidate. These missing values are expected and are discussed in further detail in the appendix. Additionally, in some cases, the sum of the predicted percentages for Trump and Harris slightly exceeds 100%, falling within a typical margin of error. This minor discrepancy is likely due to prediction error, which is also explored in the appendix.

Based on the predicted percentages, a winning candidate is identified for each state, assuming the winner takes all electoral votes for that state. As anticipated, both candidates are expected to secure wins in their respective stronghold states, with no predicted changes in non-swing states. This highlights the critical role of swing states, where a victory for either candidate could be decisive in the overall election outcome. The results underline the importance for both Trump and Harris to perform well in these battleground regions to secure their path to victory.

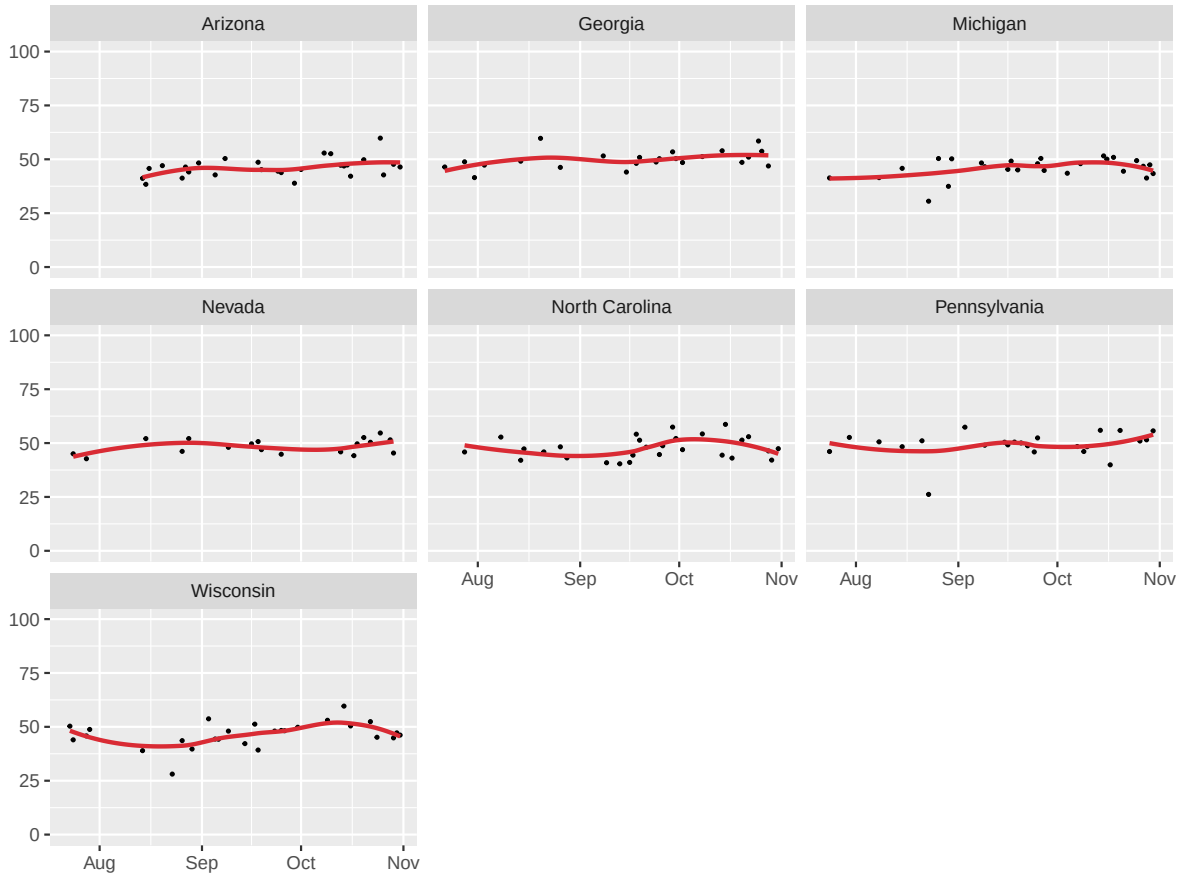
Table 1

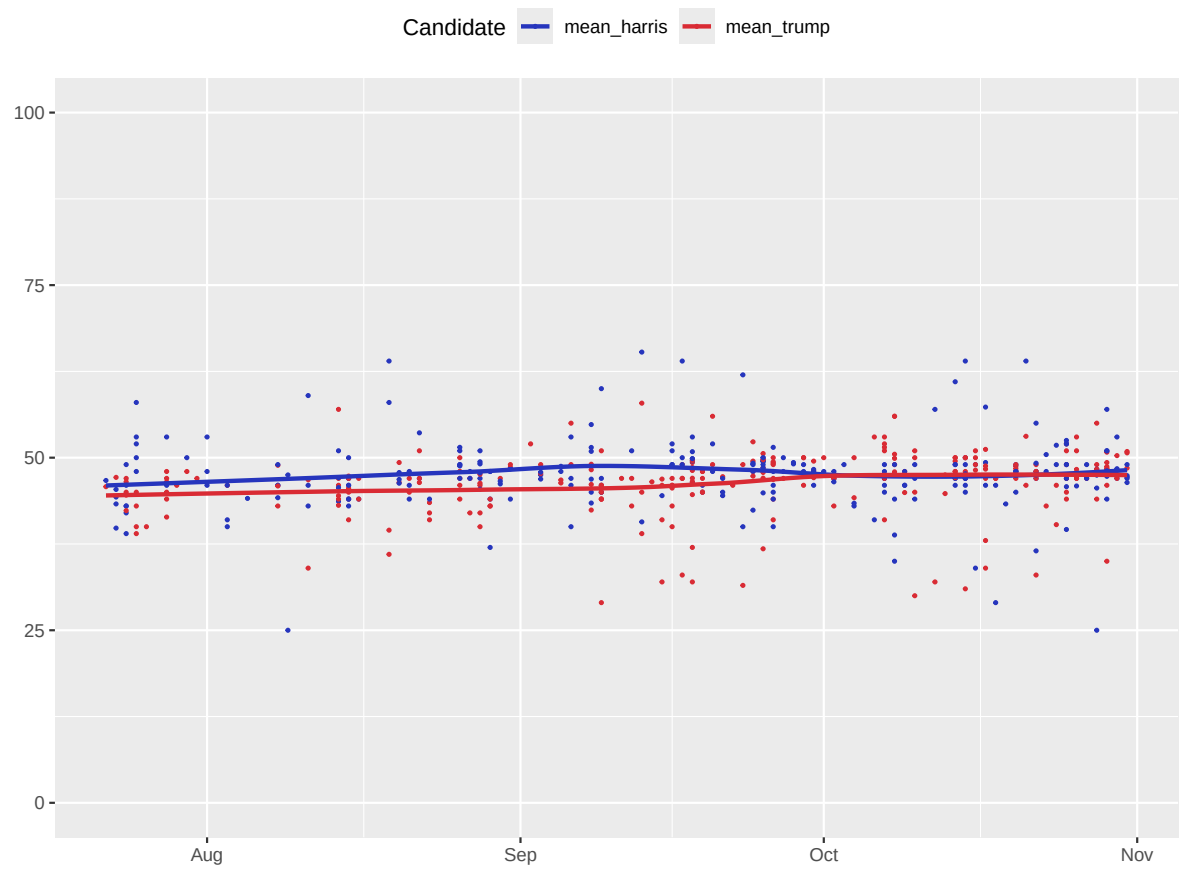
State	Harris %	Trump %	Electoral Votes	Winner
Alaska	46.52	48.80	3	Trump
Arizona	53.90	45.82	11	Harris
Arkansas	30.98	70.78	6	Trump
California	80.61	17.84	54	Harris
Colorado	67.80	NA	10	Harris
Connecticut	70.97	21.77	7	Harris
Florida	38.90	54.52	30	Trump
Georgia	46.87	49.44	16	Trump
Indiana	26.40	69.29	11	Trump
Iowa	NA	45.52	6	Harris
Kansas	NA	53.17	6	Trump
Maine	68.80	27.23	4	Harris
Maryland	87.04	18.60	10	Harris
Massachusetts	NA	9.98	11	Harris
Michigan	49.99	45.19	15	Harris
Minnesota	58.75	32.56	10	Harris
Montana	21.87	72.19	4	Trump
Nebraska	39.75	49.81	5	Trump
Nevada	50.79	48.62	6	Harris
New Hampshire	62.28	36.55	4	Harris
New Jersey	72.12	34.12	14	Harris
New Mexico	60.63	37.41	5	Harris
New York	74.88	30.80	28	Harris
North Carolina	50.18	47.67	16	Harris
North Dakota	36.15	57.17	3	Trump
Ohio	44.63	52.65	17	Trump
Oklahoma	8.65	NA	7	Trump
Oregon	NA	31.50	8	Harris
Pennsylvania	46.36	49.68	19	Trump
Rhode Island	64.45	NA	4	Harris
South Dakota	15.00	NA	3	Trump
Tennessee	21.82	72.25	11	Trump
Texas	40.37	56.66	40	Trump
Utah	7.25	66.18	6	Trump
Virginia	60.95	39.58	13	Harris
Washington	72.28	22.70	12	Harris
Wisconsin	50.24	46.47	10	Harris

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut elit leo, viverra ac enim non, pulvinar aliquam nisl. Nam accumsan ac nisi at consectetur. Curabitur luctus lacus eget risus rutrum, vel facilisis urna feugiat. Ut ac dictum velit, in blandit augue. Fusce sit amet vehicula dui. Suspendisse eleifend tempor rhoncus. Cras vehicula, nisl et molestie pulvinar, lorem ex dictum augue, nec mollis leo sem quis mauris. In lacus massa, maximus vel vehicula in, tincidunt sed lorem. Proin commodo pulvinar dictum.

Aliquam erat volutpat. Praesent in erat sit amet purus auctor posuere ut vitae leo. Suspendisse potenti. Interdum et malesuada fames ac ante ipsum primis in faucibus. Ut blandit vel orci vel aliquam. Curabitur a turpis vestibulum enim interdum consectetur at et dui. Praesent ut mauris maximus, auctor nisl ac, eleifend sapien. Morbi non varius urna. Ut placerat vestibulum nisi in lacinia.

Quisque pulvinar, enim a volutpat scelerisque, odio arcu condimentum velit, convallis vehicula velit augue id odio. Aenean ligula augue, mollis sit amet mi eu, ullamcorper aliquam massa. Aenean rutrum diam purus, eu porta mi blandit vitae. Phasellus luctus eleifend orci, nec aliquet mauris eleifend id. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

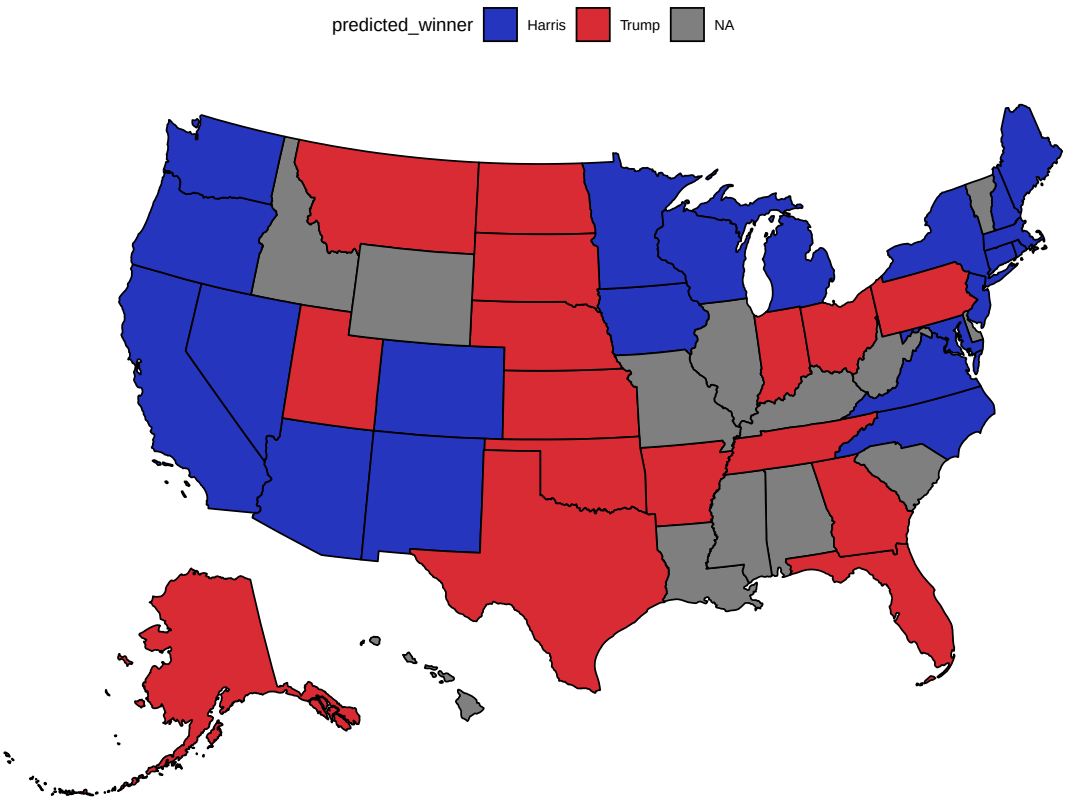




Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut elit leo, viverra ac enim non, pulvinar aliquam nisl. Nam accumsan ac nisi at consectetur. Curabitur luctus lacus eget risus rutrum, vel facilisis urna feugiat. Ut ac dictum velit, in blandit augue. Fusce sit amet vehicula dui. Suspendisse eleifend tempor rhoncus. Cras vehicula, nisl et molestie pulvinar, lorem ex dictum augue, nec mollis leo sem quis mauris.

Aliquam erat volutpat. Praesent in erat sit amet purus auctor posuere ut vitae leo. Suspendisse potenti. Interdum et malesuada fames ac ante ipsum primis in faucibus. Ut blandit vel orci vel aliquam. Curabitur a turpis vestibulum enim interdum consectetur at et dui. Praesent ut mauris maximus, auctor nisl ac, eleifend sapien. Morbi non varius urna. Ut placerat vestibulum nisi in lacinia.

Quisque pulvinar, enim a volutpat scelerisque, odio arcu condimentum velit, convallis vehicula velit augue id odio. Aenean ligula augue, mollis sit amet mi eu, ullamcorper aliquam massa. Aenean rutrum diam purus, eu porta mi blandit vitae. Phasellus luctus eleifend orci, nec aliquet mauris eleifend id. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut elit leo, viverra ac enim non, pulvinar aliquam nisl. Nam accumsan ac nisi at consectetur. Curabitur luctus lacus eget risus rutrum, vel facilisis urna feugiat. Ut ac dictum velit, in blandit augue. Fusce sit amet vehicula dui. Suspendisse eleifend tempor rhoncus. Cras vehicula, nisl et molestie pulvinar, lorem ex dictum augue, nec mollis leo sem quis mauris.

Aliquam erat volutpat. Praesent in erat sit amet purus auctor posuere ut vitae leo. Suspendisse potenti. Interdum et malesuada fames ac ante ipsum primis in faucibus. Ut blandit vel orci vel aliquam. Curabitur a turpis vestibulum enim interdum consectetur at et dui. Praesent ut mauris maximus, auctor nisl ac, eleifend sapien. Morbi non varius urna. Ut placerat vestibulum nisi in lacinia.

5 Discussion

5.1 Harris Wins the Swing States and Overcomes the Bradley Effect

Named after Thomas Bradley, a Los Angeles Mayor who underperformed his polls in the 80s, the Bradley Effect is when a candidate does poorly on polls due to his or her gender and/or race (Silver 2024). From the Table 1, we can see that Harris wins the majority of the seven swing states,states where "both major parties enjoy similar levels of support among the voting population (The Telegraph 2024)." With these states being the most pivotal toward Harris's path to victory, our prediction of Harris overcoming the Bradley effect, in spite of being a Black woman, bodes well for her in the upcoming election. It is important to realize that due to the nature of the Electoral College, it is possible to not win the popular vote but win the election, and these swing states play a major factor in this regard. Real-world polls, however, show Trump narrowly winning four out of seven swing states (however, the predicted margin of victory still lies within standard error margins of these polls).

Table 2: Swing States Polling Data as of October 31, 2024

State	Poll_1	Poll_2	Poll_3	Poll_4	Margin
Pennsylvania	47.7	47.7	48.3	48.3	R+ 0.6
North Carolina	47.5	47.5	48.6	48.6	R+ 1.1
Michigan	48.4	48.4	47.2	47.2	D+ 1.2
Wisconsin	48.5	48.5	47.8	47.8	D+ 0.7
Georgia	47.4	47.4	49.2	49.2	R+ 1.8
Arizona	46.8	46.8	48.9	48.9	R+ 2.1
Nevada	48.0	48.0	47.9	47.9	D+ 0.1

It is important to note that the margin of victories in each swing state are within the 2 to 5% margin of error of most polls; it cannot be said for certain who wins these states until Election Day.

5.2 More Urban States Tend to Lean Harris

From our results in Table 1, we see that Harris has significant leads in states like California and Maryland; from Table 3, notice that these states have a relatively high level of urban population, with 94.2% and 85.6% respectively. One reason for this could be attributed to the higher proportion of educated people in urban areas as opposed to rural ones. This allows people to gain more knowledge in diverse topics such as politics and climate change (This Nation 2021) and adopt a progressive mindset, which is in line with the ideals of the Democratic Party. In this regard, heavily urban-populated areas tend to be more culturally diverse and open to liberal ideas (Savat 2020). Moreover, the article states that "people living

in rural areas tend to have traditional values and be resistant to new ideas" which goes against the progressive mindset imbibed by Democrats.

5.3 Limitations and Areas for Future Research

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum. Nullam feugiat odio metus. Cras id convallis ante, ut ornare velit. Mauris turpis purus, porttitor eu leo quis, suscipit euismod ligula.

Appendix

A Additional Data Details

Table 3: 2020 Percentage of Urban Population for each US State

state	urban_percent_2020
Alabama	57.7
Alaska	64.9
Arizona	89.3
Arkansas	55.5
California	94.2
Colorado	86.0
Connecticut	86.3
Delaware	82.6
Florida	91.5
Georgia	74.1
Hawaii	86.1
Idaho	69.2
Illinois	86.9
Indiana	71.2
Iowa	63.2
Kansas	72.3
Kentucky	58.7
Louisiana	71.5
Maine	38.6
Maryland	85.6
Massachusetts	91.3
Michigan	73.5
Minnesota	71.9
Mississippi	46.3
Missouri	69.5
Montana	53.4
Nebraska	73.0
Nevada	94.1
New Hampshire	58.3
New Jersey	93.8
New Mexico	74.5
New York	87.4
North Carolina	66.7
North Dakota	61.0

Table 3: 2020 Percentage of Urban Population for each US State

state	urban_percent_2020
Ohio	76.3
Oklahoma	64.6
Oregon	80.5
Pennsylvania	76.5
Rhode Island	91.1
South Carolina	67.9
South Dakota	57.2
Tennessee	66.2
Texas	83.7
Utah	89.8
Vermont	35.1
Virginia	75.6
Washington	83.4
West Virginia	44.6
Wisconsin	67.1
Wyoming	62.0

B Additional Model Details

The model evaluation relies on the F1 Score and RMSE (Root Mean Squared Error) as these metrics capture key aspects of prediction accuracy and reliability, aligning well with the paper's objectives. The F1 Score, as the harmonic mean of precision and recall, is particularly useful here because it balances the model's ability to correctly identify strong support for either candidate, ensuring both Trump and Harris's predicted support is captured with minimal false positives and false negatives. This is essential in a political forecasting model, where the correct classification of state support is critical for predicting election outcomes. Meanwhile, RMSE provides insight into the overall prediction accuracy by measuring the average deviation of predictions from actual values. This helps gauge how well the model can approximate polling percentages, especially for swing states where slight errors can significantly impact projected outcomes. Together, these metrics ensure that the model is both precise in classification and accurate in magnitude, supporting the paper's goal of making robust, actionable election predictions.

C Additional Results Details

C.1 Null Values

Some states in the summary table have null values in the predicted percentages for each candidate, which is due to the subsampling and splitting approach used in the model training process. Specifically, the model was trained on a subset of the cleaned dataset found in `data/analysis_data/analysis_data.parquet`, and when training and testing data were split, certain states did not have sufficient representation in both subsets. This led to gaps in the predictions for some states, as the model could not generalize well for locations with limited or no data during training. While these null values do not affect the core results, they highlight a limitation in the sampling approach, especially in states with fewer polling data points, and are further discussed with suggestions for improvement in the appendix.

C.2 Prediction Errors

In some cases, the predicted percentages for Trump and Harris sum to slightly more than 100%, which reflects prediction error within an acceptable margin. This overestimation is consistent with the margin of error typically seen in polling averages, indicating that while the model is effective overall, minor discrepancies arise from rounding and the inherent variability in polling data. These errors can also result from the weights applied to pollsters based on their reliability scores, which may amplify certain pollster biases. To maintain accuracy, the model incorporates error margins similar to those in the polling data, ensuring predictions are within a reasonable range, with further analysis and adjustments explained in the appendix.

D Polling Methodology Overview and Evaluation for Siena/NYT

The poll was conducted from Sept. 29 to Oct. 6 and surveyed 3,385 likely voters nationwide and found that Harris led Trump by 49 percent to 46 percent, a slight lead that is within the poll's margin of error (2.4 points). According to (The New York Times 2024c), the national poll includes separate polls of 622 voters in Florida, and 617 voters in Texas. The weight given to each of these groups in the national poll has been adjusted so that the overall results are reflective of the entire country. Polls were also conducted by telephone, using live interviewers, in both English and Spanish, with about 98% of respondents were contacted on a cellphone (The New York Times 2024b). For battleground polls, voters were called from Arizona, Georgia, Michigan, Nevada, Pennsylvania and Wisconsin (The New York Times 2024a) — swing states that have a significant share in electoral college votes). In light of this, responses were also weighted to consider over- and under-represented voters to ensure that each demographic is adequately represented.

This poll employs a random sampling approach of registered voters using telephone interviews, with a focus on both landlines and cellphones. As such, this generally allowed for a representative sample. In addition, NYT/Siena also used voter registration files help ensure proper balance between political parties, and telephone polls have historically proven to be effective in gauging public opinion from recent elections. However, one particular drawback is that telephone response rates are extremely low; in fact, for this poll, "fewer than 2% of contacted individuals participate in the survey."

In handling non-responses within their questionnaire, for some questions another "sub-question" related to the original is asked only once in order to get a definitive answer; this usually occurs in questions related to leaning (not outright) candidate support. By doing this, it allows NYT to more accurately gauge the public's opinions. However, at times, they leave non-responses as "Refused" or "Don't know" often on questions targeted to the individual person (e.g. How has ____ affected you personally?). At this point, there is little no use prodding further in case of a non-response. The overall questionnaire is of high quality (NYT/Siena is one of the most reputable polls after all), and it has its strong and weak points. First, most questions, often those on respondent identification or demographic, are unbiased and fair; in particular, questions on candidate preferences did not contain words suggestive of one candidate over another. However, some questions contain gender-coded language that may skew respondents' answers in favor of one candidate (e.g. In the question "cares more about people like you" the word "cares" is more often associated with females and so respondents are more likely to choose Harris as the answer, which was seen in the poll results with Harris receiving 49% of responses versus Trump's 41%).

E Idealized Methodology and Survey

E.1 Sampling Approach and Strategy

In predicting U.S. presidential election outcomes through polling, an ideal methodology would utilize stratified sampling to ensure a representative sample of the electorate. This approach would survey registered voters from all states, then the sample will be adjusted proportionally, reflecting key demographic factors such as education, race, gender, and the urban-to-rural population ratio. To achieve this, voter registration records and census data would serve as the basis for selecting respondents. Telephone polling would be employed, with calls made at different times of the day to mitigate potential biases from non-response patterns, as certain groups of individuals may be more reachable at specific times. Given that modern telephone polling often encounters low response rates, this methodology assumes a 10% response rate, consistent with findings, highlighting the challenges in obtaining representative samples from telephone-based surveys (Center 2017). Therefore, around 15,000 individuals will be contacted to obtain a desired sample size of 1500. Despite the low response rate, carefully deploying methods should help mitigate biases and improve the accuracy of election predictions.

E.1.1 Stratification Variables [TO-DO]

- **Age Groups**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female, Non-binary/Other
- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other
- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

E.2 Budget Allocation

- **Marketing and Outreach (Ads and Incentives)**: $50,000
- **Survey Design and Data Validation**:
  - Questionnaire Development: $5,000
  - Pilot Testing (survey run on small group of people): $5,000
  - Google Forms (with built-in validation rules): $5,000
- **Data Analysis and Software**:
  - Statistical Software: $5,000
  - Staff Fees (Analysts and Statisticians): $15,000
- **Financial Incentives**: $2,000
- **Miscellaneous Expenses and Contingency Fund**: $8,000

E.2.1 Survey Structure:

The survey will begin by introducing each potential participant to the topic and providing necessary context about the purpose of the research. This introduction aims to make the participants feel informed and comfortable with the process. The first question will ask whether the individual is willing to participate in the survey, emphasizing that participation is voluntary. It will also be clearly communicated that respondents may drop out of the survey at any point if they choose. Once participants agree, they will proceed to answer a series of demographic questions that will help ensure a diverse and stratified sample. These questions include:

- "What is your gender?"
- "What is your age?"
- "Which city do you live in?"
- "What race do you identify as?"

- "What is your highest degree of education?"
- "Will you be voting in the upcoming election?"
- "Which candidate will you be voting for?"

E.3 Data Validation

References

Center, Pew Research. 2017. "What Low Response Rates Mean for Telephone Surveys." 2017. https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/.

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

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

RealClearPolitics. 2024. "RealClearPolitics Polling Data." https://www.realclearpolitics.com/epolls/latest_polls/.

Savat, Sara. 2020. "The Divide Between Us: Urban-Rural Political Differences Rooted in Geography." https://source.washu.edu/2020/02/the-divide-between-us-urban-rural-political-differences-rooted-in-geography/.

Silver, Nate. 2024. "Election Polls and Results: Analyzing Trump Vs. Harris." *The New York Times*. https://www.nytimes.com/2024/10/23/opinion/election-polls-results-trump-harris.html.

The New York Times. 2024a. "Times/Siena Poll Methodology." https://www.nytimes.com/article/times-siena-poll-methodology.html.

———. 2024b. "Times/Siena Poll: Florida Toplines." https://www.nytimes.com/interactive/2024/10/13/us/elections/times-siena-poll-florida-toplines.html.

———. 2024c. "Times/Siena Poll: Likely Electorate Crosstabs." https://www.nytimes.com/interactive/2024/10/13/us/elections/times-siena-poll-likely-electorate-crosstabs.html.

The Telegraph. 2024. "What Is a Swing State? Key Battlegrounds in the 2024 US Election." https://www.telegraph.co.uk/us/politics/2024/10/30/what-is-swing-state-key-battlegrounds-us-election/.

This Nation. 2021. "Why Do Cities Vote Democrat?" https://www.thisnation.com/politics/elections/why-do-cities-vote-democrat/.