Predicting the 2024 U.S. Presidential Election: A Data-Driven Analysis of Swing State Outcomes

Forecast Model Suggests Harris as the Probable Winner of the majority of swing states

Alexander Guarasci & Jacob Gilbert

November 3, 2024

This paper develops a predictive model for the 2024 U.S. Presidential Election, analyzing state-level polling data to forecast outcomes between Donald Trump and Kamala Harris. Using logistic regression, the model estimates the likelihood of Trump's success in swing states by weighting polls based on recency and sample size. Findings indicate that Harris has a higher probability of winning in most swing states, though close margins suggest targeted campaign efforts could alter outcomes in pivotal areas. This analysis provides insight into electoral dynamics, highlighting how data-driven approaches can clarify potential election outcomes and inform strategic political decision-making.

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, focusing on swing states that will likely decide the election result. By aggregating high-quality polls that account for recency and sample size, we create a logistic regression model that estimates the probability of a Trump victory in each state. This approach allows us to analyze voter support patterns across the states and predict election outcomes from polling data.

The estimand in this study is the probability that Donald Trump will win a given state, derived from 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, with predictor variables comprising the weighted average polling percentages for both candidates. We assign greater importance to more recent polls and those with larger sample sizes, enhancing the reliability of our estimates.

Our model employs a logistic regression framework to predict election outcomes based on these weighted averages. The results reveal the geographic distribution of support for each candidate. We have focused our analysis on the swing states where polling percentages are closely contested and could significantly influence the final election result. Accurate election predictions provide valuable insights into voter dynamics, helping political analysts, campaigns, and the public anticipate electoral outcomes. By focusing on high-quality polling data and incorporating weights for recency and sample size, our model enhances prediction reliability and identifies key regions where voter sentiment may shift, ultimately impacting the election.

The structure of the paper is as follows: Section 2 describes the data collection process, detailing the sources, variables of interest, and criteria for selecting high-quality polls. Section 3 introduces the logistic regression model used for prediction, explaining the variables, weights, and methodology applied to forecast the election outcome. In Section 4, we present and analyze the model's predictions, highlighting the probabilities of candidate success across key states. Section 5 discusses the model's implications, limitations, and the observed contrasts between polling data.

2 Data

The data used in this paper was gathered from FiveThirtyEight (FiveThirtyEight 2024) a website that aims to use "data and evidence to advance public knowledge". The programming language for data analysis, visualization and statistical investigation was Python (Van Rossum and Drake 2009) along with the packages Matplot (Hunter 2007), Seaborn (Waskom 2021), Numpy (Harris et al. 2020), Pandas (team 2020), Jupyter (Jupyter 2023), Tabulate (Korolev 2023), and Sklearn (Buitinck et al. 2013).

2.1 Measurement

The voter support data used in this analysis comes from raw polling information sourced from the Project 538 online database (FiveThirtyEight 2024). This dataset provides predictions for candidate support for the 2024 U.S. Presidential Election by aggregating various polls that capture public sentiment toward candidates Donald Trump and Kamala Harris. Each entry in the dataset corresponds to a specific question about voting preference in a poll conducted by different polling organizations, which measures the percentage of respondents expressing their support for each candidate in their respective state. the surveys utilized one or multiple of the following methods for response gathering: Online Ad, Online Panel, Live Phone, text-to-web, mail-to-web, email, Probability panel, IVR. The diversity of the methods of collection supplies a colourful array of polls capturing all facets of voters.

The polling data is collected through carefully structured survey questions, designed to elicit clear responses regarding voter preferences. By aggregating these individual responses, we are

able to derive state-level sentiment for both candidates. This transformation of general voter sentiment into specific data points enables us to analyze the competitive landscape between Trump and Harris across swing states, providing a clearer understanding of electoral dynamics as they relate to the upcoming election.

2.2 The Dataset

A swing state in the context of a US election is defined as "a state where the number of Democratic and Republican voters is about the same, that has an important influence on the result of the election of the United States President". States that do not fall under this definition are not likely to "swing" to another party as we approach the election, and thus their outcome can be reasonably assumed. Thus the election result is normally decided by which party wins just a handful of states in the electoral college system. By concentrating on these swing states, we aim to capture the most critical and uncertain areas of the electoral map, where voter preferences are most likely to sway the final result. Thus our analysis restricts the datasets to state-level polls in the following states: Arizona, Pennsylvania, North Carolina, Georgia, Nevada, Michigan, and Wisconsin; determined to be swing states in 2024 FiveThirtyEight (2024). In Figure 1 we see the current distribution of electoral votes that each party hold according to those sources as of October 20th, with the remaining swing states, seen in the middle of this graph set to determine if either Harris or Trump receives over 270th electoral votes.

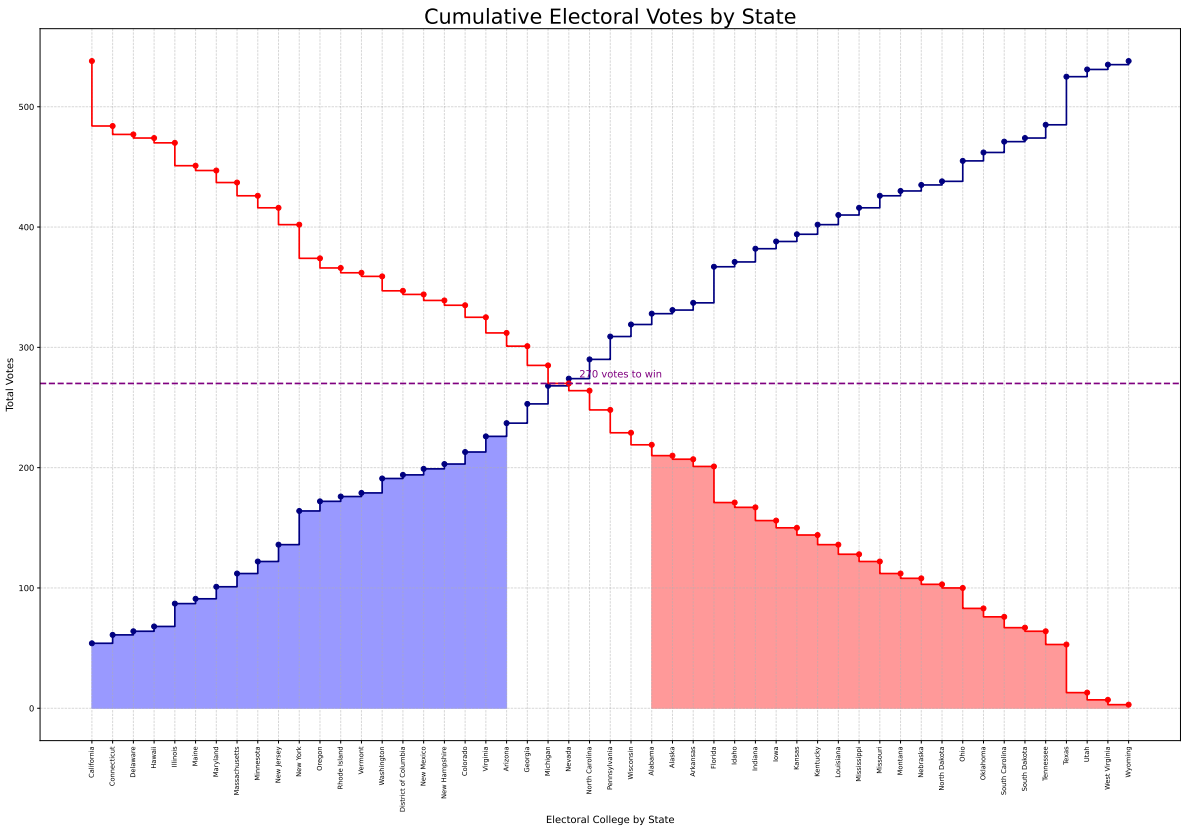

Figure 1: This figure shows the cumulative total of electoral votes by state, grouped by current party affiliation, accordng to 538. The chart distinguishes between Democrat, Republican, and swing states, with a color-coded fill under each party's cumulative curve. Key threshold of 270 electoral vote mark needed to secure a victory is diplayed. This visualization shows the democrats lead with 227 votes and the Republicans with 219 vot secured

It is important to note that each pollster employs unique sampling techniques and methodologies that influence their Pollscore and Transparency Score. Pollscore is a measure of a pollster's historical accuracy, reflecting the quality of their sampling methods and data collection. A good score reflects low propensity to be impacted by errors and biases in survey sampling. It also accounts for impacts that might bias a survey's score, such as adjustments for the differing difficulty of polling certain political races or elements of luck in samples by resampling polls. Transparency Score, on the other hand, measures how openly a pollster shares methodological details like question-wording, sampling methods, and weighting procedures. By combining these two metrics into a rating, we achieve an excellent overall assessment of a pollster's quality, capturing both their empirical accuracy and their commitment to methodological transparency.

We use this single measurement out of 3 to ensure the integrity of our analysis. Selecting only those polls that originate from reputable pollsters, specifically targeting those with a numeric grade of 2.5 or higher. This approach is used so that the data reflects a robust and credible representation of voter preferences.

In keeping with this aim, we focused our analysis on polls conducted after September. Early in the election cycle, polls can be significantly influenced by initial campaign events, such as the shocks from the first debates or unexpected changes in the candidate lineup. For example, Joe Biden dropping out as a candidate on July 21st severely impacted the odds of a Kamala Harris victory. These factors can introduce biases and fluctuations that do not necessarily reflect the enduring preferences of the electorate. By selecting polls from after September, we capture data from a period when voter opinions are more stable and better represent the current state of the race.

State	Variable	Mean	Std	Min	Max	Count
Arizona	Numeric Grade	2.9	0.1	2.7	3	26
	Sample Size	905.3	283.6	500	1440	26
Georgia	Numeric Grade	2.8	0.1	2.6	3	26
	Sample Size	1002.7	273.7	682	1441	26
Michigan	Numeric Grade	2.8	0.1	2.6	3	22
	Sample Size	914.4	271.7	500	1529	22
Nevada	Numeric Grade	2.7	0.1	2.6	2.9	9
	Sample Size	884.1	187.5	652	1171	9
North Carolina	Numeric Grade	2.8	0.1	2.5	3	33
	Sample Size	978.1	263.3	589	1674	33
Pennsylvania	Numeric Grade	2.9	0.1	2.6	3	38
	Sample Size	1077.6	387.3	450	2048	38
Wisconsin	Numeric Grade	2.9	0.1	2.6	3	30
	Sample Size	891.9	171.8	680	1312	30

Figure 2

Figure 2: Summary statistics of key sample variables by state. Shows high sample size and numeric grade among all polls of swing states.

2.3 Variables of Interest

In our logistic regression model predicting the 2024 U.S. Presidential Election outcome, the primary variables of interest are derived from aggregated state-level polling data. The percent-

age of respondents favouring a candidate is the key variable used for prediction. Within the collection of polls in our sample, the independent variable `trump_pct` represents the percentage of respondents who indicate support for Donald Trump, and `harris_pct` respectively.

Polls are taken towards the build-up of an election and thus are taken by respondents at different times. Figure 2 shows how the sample of polls are distributed over time in which it is clear that we have a collection of polls from a wide range of periods. Because sentiment changes over time we adjust the weight that we place on polls that are closer to the election as these are more indicative of how respondents will be voting on the day.

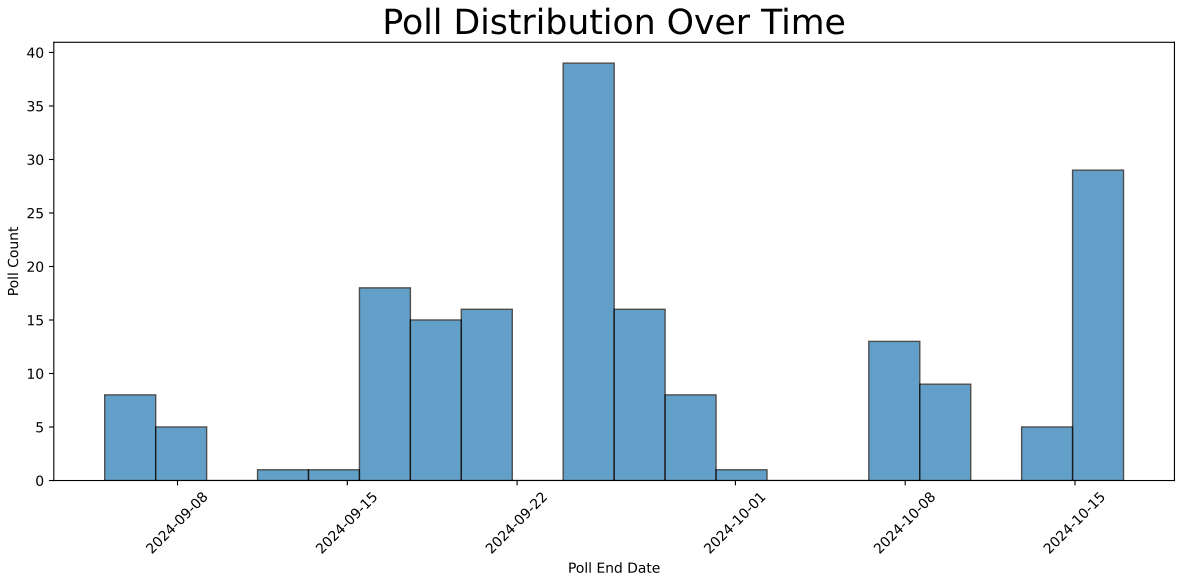

Figure 3: A histogram of the poll quantities, aggregated by the end date of each poll, shows the distribution of polling data over time. This chart reveals the frequency of polls leading up to the election and provides insight into polling activity trends, with potential spikes reflecting periods of increased public interest or campaign activity.

Another aspect of the polls we wish to consider is the population size of each. Larger polls result in a stronger reduction in randomness in the response. The larger the sample size the more the response reflects the true population of the state, this is known as the law of large numbers. In Figure 3 we can see the variety within the state polls. For use in the model, the percentages in these polls are weighted for recency and size of the sampled population.

3 Model

This paper develops a logistic regression model to predict the likelihood of Donald Trump winning various states in the upcoming 2024 U.S. Presidential Election based on aggregated

polling data. The model leverages polling percentages, recency of the polls, and sample sizes to provide a robust probability estimate of Trump's chances against Kamala Harris.

Model Overview

The model can be expressed mathematically using the logistic function:

$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$

Where:

- $P(Y = 1|X)$ represents the probability of Trump winning (i.e., Y=1).
- β_0 is the intercept term.

β_1 and β_2 are the coefficients for the predictor variables, which include:
- X_1 : The average polling percentage for Trump (trump_pct).
- X_2 : The average polling percentage for Harris (harris_pct).

Data and Features

The dataset used in this model is sourced from Project 538, focusing on high-quality polling data. Key features include:

- trump_pct: The percentage of respondents supporting Trump.
- harris_pct: The percentage of respondents supporting Harris.
- sample_size: The size of each poll, which influences the weight of that poll in our analysis.
- start_date: The date when the poll began, which is used to calculate recency.

Weight Calculation

To ensure that more recent and larger polls have a higher influence on the model, we calculate a weight for each poll as follows:

Recency Weight:

recency weight $= \max(\text{current date} - \text{start date}) - (\text{current date} - \text{start date})$

This transformation ensures that more recent polls receive a larger weight.

Total Weight:

weight = sample size × recency weight

The weighted polling percentages for Trump and Harris are then calculated for each state, allowing for a more nuanced aggregation of voter sentiment.

Logistic Regression Implementation

The logistic regression model is implemented using the LogisticRegression class from the sklearn library. The predictor variables are defined as:

- X = [trump_pct, harris_pct]

The binary outcome variable is defined as follows:

$y = 1$ if trump_pct > harris_pct (Trump is predicted to win) y = 0 otherwise.

Model Validation and Assumptions

The assumptions underlying this model include:

- Linearity: The log-odds of the outcome variable are linearly related to the predictor variables.
- Independence: Observations are independent of one another.

Limitations

1. Polling Bias: The model is contingent upon the quality of the polling data. Polls with lower numeric grades or transparency scores are excluded, potentially introducing bias if high-quality polls are not representative of the entire electorate.

2. Dynamic Voter Sentiment: Voter preferences can change rapidly, particularly in the lead-up to the election, and the model may not fully capture these shifts if not updated frequently.

Alternative Models Considered

Alternative models, such as Bayesian logistic regression, were considered. While Bayesian models allow for the incorporation of prior distributions and might provide additional insights through credible intervals, they require more complex implementation and careful selection of priors. Given the nature of the data and the objective of clear interpretability, the logistic regression model was chosen for its straightforward application and interpretability.

Ultimately the logistic regression model provides a structured approach to predict the likelihood of Trump winning in various states, based on recent and high-quality polling data. By focusing on key features such as polling percentages and sample sizes, the model captures the competitive dynamics between candidates and offers valuable insights into potential election outcomes. Future iterations of this model can incorporate real-time data updates to enhance predictive accuracy as the election approaches.

4 Results

Regression results

State	Trump %	Trump SD	Harris %	Harris SD	Trump Winning %
Arizona	49.85	1.24	47.04	2.17	86.85
Georgia	49.39	1.78	47.66	2.52	71.18
Michigan	48.57	2.14	47.72	2.12	53.09
Nevada	47.90	1.50	48.63	1.19	21.34
North Carolina	48.48	1.41	48.58	1.55	32.31
Pennsylvania	47.68	1.81	48.53	1.29	19.67
Wisconsin	47.76	2.00	48.92	1.72	15.52

Figure 4

Figure 4: Is a results table for the model described above.It shows support for Trump and Harris at the state level according to the available polls. Small standard deviations show consistent results, yet the race is very close in many states. Percentages above 50% show the favoured candidate in each state.

In Arizona and Georgia, the model predicts the strongest likelihood of a Trump victory, with a mean polling percentage showing Trump over 2 percentage points ahead in Arizona. In Georgia, Trump's mean polling percentage is 49.39% against Harris's 47.66%. The data thus resulted in a strong 86.85% and 71.18% likelihood of Trump winning in Arizona and Georgia respectively.

Trump is also predicted to win in Michigan, although by smaller margins as shown in this state. As a result, the model predicts Trump winning with only a 53% chance, showing his odds are near to a coin flip. With Arizona, Georgia and Michigan Trump is predicted a total of 42 of the 93 swing state votes.

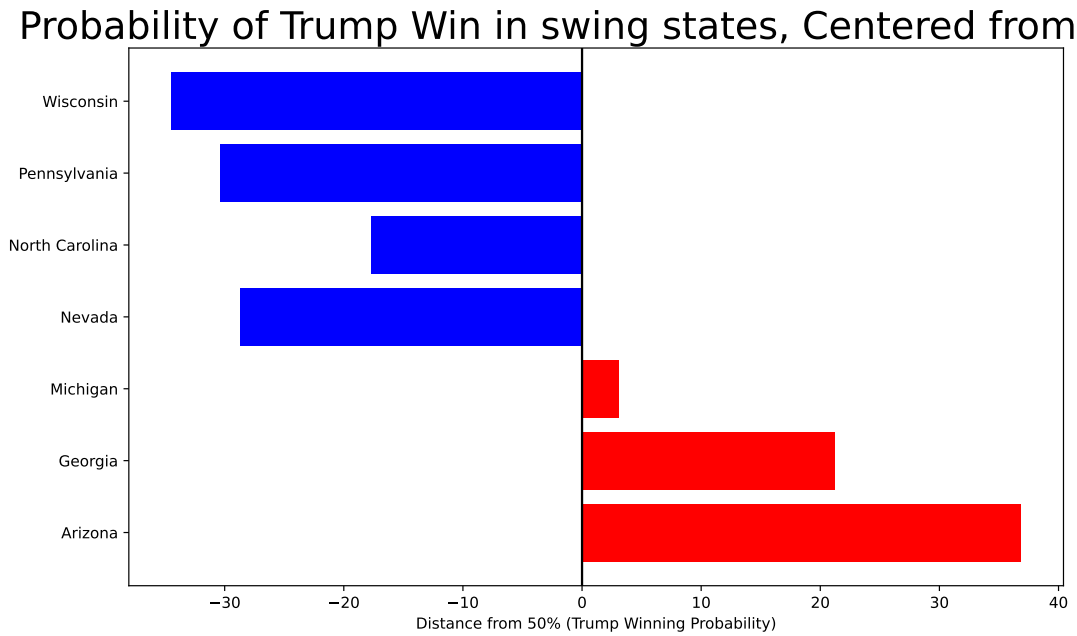

Figure 5: A centered bar chart where each state has a bar reflecting Trump's probability of winning. The bars extend to the right or left from the center line at 50%, with the distance representing the magnitude of deviation from an even chance. This chart visualizes Trump's relative strength in each swing state and highlights that he s only favoured in Georgia and Arizona.

Kamala is predicted to win the remaining states and is given fairly high chances of locking these up. despite only winning Nevada, North Carolina and Pennsylvania by under a percentage point in the model Kamala's probability of winning those electoral votes is very high. This is due to the small standard deviations shown in Figure 5 making the predicted outcome of a victory more secure.

Wisconsin is shown to be her strongest and most likely victory of the swing states, with Trump only having a 15% chance here.

Overall this gives Harris 51 of the 93 swing state votes, our model expecting a win for the Democrats and the first female US president!

Trump overall probability

Taking The probabilities from the model above we can look at the outcomes of the 207 possible combinations of wins and losses for Trump in these swing states and deduce his overall probability of reaching 270 electoral votes. Based on the modelled probabilities his current expected vote is 41.05, below the threshold of 51 he needs based on our assumptions.

By separating the outcomes into those that give Trump the required electoral college votes, we can obtain the probability that Trump seizes a second term, using probability theory, and assuming independence among the state outcomes in our approach. When accounting for all possible state outcome combinations, the model calculates that Trump has approximately a 22.5% chance of securing at least 51 electoral votes from these key swing states. This probability aggregates the likelihoods across all combinations where the sum of electoral votes meets or exceeds the 51-vote threshold. The analysis reveals that despite high individual probabilities in certain states, the overall chance of Trump amassing sufficient electoral votes from the swing states is less than one in four.

5 Discussion

What Is Done in This Paper?

In this paper, we create a logistic regression model to predict the outcome of the 2024 U.S. Presidential Election, focusing on state-level polling data to estimate the probability of Donald Trump winning against Kamala Harris. By aggregating recent and reliable polls for swing states and weighting them by sample size and recency, we aim to provide a data-driven analysis of voter preferences. Our model is constructed to identify the likelihood of Trump securing a majority in each state, ultimately offering a prediction for the overall election result.

What Do We Learn About the World?

The result of the model tells us that overall the polls in October are favouring Harris to win the election. From the perspective of the republican party if they wish to "turn the tide" this model highlights which of the swing states are more likely to flip than others. North Carolina is the premier state predicted to go blue with the highest likelihood of Trump winning. It would be recommended to allocate resources here to try and win a state from Harris. As far as retaining currently favoured states, Michigan is in a weak position for Trump according to the model, nearing a coin flip. As this state holds 15 electoral College votes, the 4th most of the swing states, it is a key state to win. Assuming the model's correct prediction that Michigan is currently held by Trump, expect the Democrats to attempt to tip the scales in their favour there.

One of the most significant insights from this analysis is the discrepancy between polling data and the betting markets. While the majority of reputable polls suggest that Kamala Harris is favoured on both a national level and in most swing states, betting markets like Polymarket imply a 67% chance of Trump winning the election as of October 23, 2024 (Polymarket 2024). This stark disconnect raises questions about the accuracy of traditional polling methods and whether betting markets, which are financial tools driven by market forces, may offer a more precise reflection of public sentiment.

At first glance, one might expect that if polling data were more accurate than betting markets, arbitrage opportunities would emerge, allowing savvy participants to profit from discrepancies. However, this does not seem to be happening, which suggests that the markets may be pricing in information that the polls do not capture—perhaps reflecting shifts in voter sentiment, hidden preferences, or systematic biases in polling.

Another important consideration is the possible biases in both polling and betting markets. Poll respondents may not be a representative sample of the electorate; for example, Democrats could be more likely to respond to polls, skewing the results in favour of Harris. Meanwhile, individuals who participate in betting markets might form a subset of the population that is disproportionately supportive of Trump, which could explain why the implied odds heavily favour him. Furthermore, the rapid expansion and increased liquidity of betting markets in the last few years may have improved their efficiency, making them more reflective of real-world probabilities. However, historical data still suggest that polls have been accurate 78% of the time in predicting elections (FiveThirtyEight 2024), indicating that the reliability of betting markets remains questionable, especially in light of their poor performance during the 2016 election, when Trump's odds were listed at +475 (17%) just before his victory (Odds Shark 2016).

Weaknesses of the Model

While our model provides a structured framework for predicting the election, it has several limitations. First, the model is based on polling data available far in advance of the election, meaning there is a high degree of uncertainty, and the model may not fully capture late shifts in public opinion or external shocks (e.g., economic downturns or scandals). As polling results are based on public opinion, large shock events are expected, furthermore, scheduled events such as debates and rallies can be expected to have large impacts on the election that cannot be modelled here. Additionally, the use of linear modelling may oversimplify the complex dynamics of voter behaviour, as elections often involve nonlinear influences that are difficult to predict.

Moreover, our focus on swing states limits the model's applicability to the national picture. While swing states are crucial to the election outcome, non-swing states could offer additional insights into broader voter trends, and including them in future analyses could enhance the model's robustness. Training the model on a more comprehensive dataset that includes these

states and other predictive variables, such as demographic factors or economic indicators, could lead to a more accurate forecast.

Trump's low likelihood of winning shown by our results may under-represent his support, this claim is supported by the disparity with the betting markets. This could be a result of a common problem with polling data, response rate bias. In the previous two elections, where Donald Trump was also the republican candidate, his polling was far below the eventual results on election day. Post-mortems of both elections indicated evidence that those supporting Trump were not participating in polls and thus under-presenting Republican support (Clinton et al. (2021) , Kennedy et al. (2018)). Should this trend continue Trump could surprise the model and win some of these swing states. Another reason Trump could outperform the model is another bias often seen in surveys, one of social desirability. Trump is a unique candidate as, if elected, he would become the first president to be elected with a federal crime conviction. This along with other negative publicity the former host of TV's Apprentice has garnered since the start of his political campaign could bias voters from publicly or privately showing their support. Even with the quality polls used in this analysis, where steps are taken to ensure representative results (see Appendix B), we could be observing a skew in the results.

How Should We Proceed in the Future?

Looking ahead, future iterations of this model should incorporate more diverse data sources. Expanding the dataset to include polling data from all states, along with betting market information, could improve prediction accuracy. Moreover, experimenting with different types of predictive models, such as Bayesian logistic regression or machine learning models like random forests, could offer more sophisticated insights and account for non-linear interactions that our current model might miss.

It would also be valuable to study the interaction between polling data and betting markets more closely, potentially integrating them into a unified prediction model. This hybrid approach might help reconcile the discrepancies observed between the two sources and provide a more nuanced understanding of election dynamics. Additionally, out-of-sample validation and sensitivity analyses should be conducted to test the robustness of our model and adjust for overfitting.

Ultimately, while our model offers a strong foundation for predicting the 2024 U.S. Presidential Election, there is still much to learn. By refining the model and incorporating additional data, we can move closer to producing predictions that more accurately reflect voter behaviour and election outcomes.

Appendix A: Idealized Methodology and Survey for Forecasting the US Presidential Election with Incentivized Betting

Sampling Approach

To ensure that our sample represents the diversity of the US electorate and minimizes potential biases, we employ stratified and cluster sampling techniques. With a budget of $100,000, our goal is to gather data from 5,000 respondents, focusing on key swing states while preserving national representativeness. The sample will be stratified by:

- **Demographics**: Race, age, gender, education, and income.
- **Geography**: Emphasis on swing states with representation from urban, suburban, and rural areas.
- **Voter History**: Including respondents with varying voting patterns, such as frequent voters, occasional voters, and those who are less likely to vote.

To address geographic and political biases, cluster sampling will target diverse regions within each state (urban, suburban, and rural areas), capturing the variation in political leanings. Additionally, we will soft-launch the survey in order to catch potential issues.

Recruitment of Respondents

Respondents will be recruited through a multi-channel outreach strategy that combines online and telephone efforts to ensure a representative sample:

- **Online Recruitment**: Targeted ads on social media, political forums, and news websites will attract a wide audience, highlighting the unique opportunity to place a small bet on the election outcome.
- **Telephone Recruitment**: To capture older demographics and those less reachable online, we will conduct phone outreach using commercial databases and voter registration information.

Incentives for Engagement

A betting pool is a central part of this methodology, with $50,000 allocated as rewards for respondents who accurately predict the election outcome. Each participant will receive a $10 allowance to bet on their predicted winner (Trump or Harris), with payouts based on live odds, creating a direct incentive for respondents to make thoughtful, informed predictions. The purpose of this is twofold, it aims to minimize the attrition of the respondents, so that more people complete the survey, and it also will provide us insight into who people think will

win rather than just who people want to win. Ideally, the purpose of this is so that people take into consideration who their friends, family, and their community, are expecting to win.

Data Validation and Bias Reduction

To further reduce biases, several measures will be implemented for data validation and weighting:

- **Cross-Referencing Survey Responses**: Survey responses will be cross-referenced with voter registration data to validate eligibility.
- **Weighting**: Data will be weighted to reflect broader population demographics, adjusting for any imbalances in representation across age, race, gender, and geographic factors. This ensures that the results better mirror the entire electorate.

Poll Aggregation

Data collected through this survey will be aggregated with other national and state-level surveys to improve accuracy. The aggregation process will account for:

- **Sample Size and Recency**: More recent polls and those with larger sample sizes will be weighted more heavily.
- **Pollster Reliability**: Historical poll accuracy and transparency scores will further inform the weight of each poll.

Budget Allocation

The proposed $100,000 budget will be distributed as follows:

- **Recruitment**: $10,000 for targeting and engaging respondents across multiple platforms.
- **Survey Administration (Online and Phone)**: $20,000 for both online and phone-based survey collection.
- **Betting Pool**: $50,000 allocated to reward accurate predictions and enhance response quality.
- **Data Validation and Analysis**: $10,000 for verification and weighting.
- **Modeling**: $10,000 for analyzing and forecasting based on aggregated data.

Question Design and Goal Alignment

The goal of our survey is to predict who will win the election, this is the research question that guides our survey. In order to achieve this, our questions will isolate the relevant variables by asking only one thing at a time (ceteris paribus). We will also use item-specific scales over agree-disagree formats to avoid acquiescence bias. Furthermore, we will randomize response options in order to mitigate response order bias and emphasize the survey's anonymity to minimize social desirability bias.

This approach, inspired by Stantcheva's guide on survey creation (Stantcheva 2023), leverages financial incentives to align respondent predictions with their genuine expectations. By combining innovative sampling, incentivized engagement, and rigorous data validation, this methodology aims to bridge the gap between traditional polling and betting market predictions, ultimately enhancing the accuracy of our election forecast.

This survey can be found at: https://docs.google.com/forms/d/e/1FAIpQLSc5LYCAd0OmiIc-LHRFUVgVGuSMKzayVWZB7VOrsrDOe5742w/viewform?vc=0&c=0&w=1&flr=0

Appendix B: Siena / New York Tims Pennsylvania poll

This section discusses the Siena College and the New York Times poll from the database of polls used in the above analysis. The New York Times (NYT) polls often emphasize the swing states making these very useful in this paper's analysis. This section will focus specifically on one poll done in Philadelphia, but those in other states follow the same methodology and patterns. This consistency is important for the New York Times when they are reporting their results as it makes comparisons of states reliable. In going over the methodology this appendix aims to understand what about this pollster makes them high quality and where they might improve.

Sampling method

The poll employs a response rate-adjusted stratified sampling method using the L2 voter file, which contains detailed demographic information on registered voters. Initially, the interviewees are selected randomly from a national list of registered voters, responses are then weighed. The stratification, often referred to as weighing, is applied to the random sample collected in order to properly reflect the entire demographic. The stratification accounts for:

- Statehouse district
- Political party affiliation
- Race
- Gender
- Marital status

- Household size
- Turnout history
- Age
- Home ownership The sampling is conducted in order to address telephone coverage around the state as well as nonresponse rates. Records are selected by state and sampling is separated for Pennsylvania as a whole and its major city Philadelphia. The voter file is then ratified and weights are based on historically modelled response rate and coverage.

Collection

Times/Siena polls are conducted by telephone, using live interviews. About 96% of interviewees were contacted by cellphone, and interviews were conducted in both English and Spanish. In the Philadelphia polls occurring between September 11th and 16th, 240,000 calls were made to nearly 118,000 likely voters (The New York Times (2024)). This high call volume is key to maximizing response from all groups. In a further effort to get a good sample additional effort is made to contact under-represented groups.

Questionnaire design

The questions themselves included more than just the information on who a participant would vote on. Followup questions were given either because or regardless of what their previous answers were. One example would be follow-up questions for those who answer "don't know/ refused" in order to understand some reasoning behind this answer. Those who did select an option also were asked questions to get a clear image of their sentiment or reasoning. The benefit of these additional questions is to bring context to the initial binary question.

Methodology pros and cons

The Random contact method in which people are contacted does not provide random sampling, as non-response has made this method on its own more redundant. (Bailey (2023)). This is clearly a factor here as the response rate was below 2%. The inclusion of response rate adjustments accounts for the likelihood that certain groups are more or less likely to respond to surveys, and is an efficient way to counter these issues ((**brick2013Nonresponse?**)). The stratification also allows for a more representative expectation from the poll ensuring all significant groups are accurately represented in the sample as they are in the population (Kish (1965)), Making the means generated more accurate.

With such low response rates the possibility for risk, in spite of the effort made by the pollster, is that non-response and choice are correlated, this is known as non-response bias. It is not uncommon to assume that non-response should occur at random, especially as this poll documents repeatedly calling each likely voter, however in the case of elections there is previous

evidence of a bias. Research into the previous two elections and the polls proceeding then showed a high probability that non-responders were more likely to vote for Donald Trump, who was the Republican candidate in both 2016 and 2020 (Clinton et al. (2021), Kennedy et al. (2018)). With Trump running again in 2024, his popularity being underestimated is a distinct possibility, as is mentioned in the discussion section.

The benefit of how the survey was taken is the detail added from the additional questions asked, some key questions allow for further stratification based on the answers. For example, survey information is gathered on self-reported likelihood to vote, and obviously this gives respondents answers far more weight if they are. One documented phenomenon worth mentioning is social desirability bias, that people are more inclined to answer with more socially acceptable answers. This could occur throughout the questionnaire, however it is very likely that respondents overestimate their likelihood to vote (Holbrook and Krosnick (2009)). The pollster's use of both modelled turnout probabilities based on historical data and self-reporting improves the accuracy of predicting who is likely to vote dramatically. The poll's use of likely voters still brings some uncertainty to its results, as the results may be impacted by non-voters, however, Siena/NYT took a methodology that aims to negate this as much as possible. Another benefit of live interviews is that clarification of answers is possible and the additional depth of the data that this provides. Qualitative insights into voter's choices or indecision create a more robust and significant dataset. By comparison, this is far more valuable than a singular binary choice quiz you may see online.

In conclusion, while the pollster's methodology is robust and incorporates several best practices in survey research, inherent challenges such as low response rates and the possibility of nonresponse and social desirability biases must be acknowledged. This poll's methodology makes great efforts to mitigate these biases through its weighting methods, yet these factors could affect the accuracy of the poll results. Nonetheless, The poll provides valuable insights and remains a useful resource for analyzing voter preferences, especially due to its aforementioned weighting strategy. The poll's transparency in methodology makes it an even more reliable resource and enhances the credibility and comparability of the findings.

what is included in the poll?

Methodology and why?

References

Bailey, Michael A. 2023. "A New Paradigm for Polling." *Harvard Data Science Review* 5 (3). https://doi.org/10.1162/99608f92.9898eede.

Boyce, Lily, Lazaro Gamio, Eli Murray, and Alicia Parlapiano. 2024. "Tracking the Swing States for Harris and Trump." *The New York Times*. https://www.nytimes.com/interactive/2024/us/elections/presidential-election-swing-states.html.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, et al. 2013. "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project." In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–22.

Clinton, Joshua, Jennifer Agiesta, Megan Brenan, Chase Burge, Michael Connelly, Ariel Edwards-Levy, Bernard Fraga, et al. 2021. "Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls." American Association for Public Opinion Research.

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

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Holbrook, Allyson L., and Jon A. Krosnick. 2009. "Social desirability bias in voter turnout reports: Tests using the item count technique." *Public Opinion Quarterly* 74 (1): 37–67. https://doi.org/10.1093/poq/nfp065.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Jupyter, Project. 2023. "Jupyter: Open Source Tools for Interactive Computing." https://jupyter.org.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua Clinton, Claire Durand, Charles Franklin, Kyle McGeeney, et al. 2018. "An Evaluation of the 2016 Election Polls in the United States: AAPOR Task Force Report." *Public Opinion Quarterly* 82: 1–33.

Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons.

Korolev, Sergey A. 2023. "Tabulate: Pretty-Print Tabular Data in Python." https://pypi.org/project/tabulate/.

Odds Shark. 2016. "US Presidential Odds 2016 Futures." https://www.oddsshark.com/entertainment/us-presidential-odds-2016-futures.

Polymarket. 2024. "Presidential Election Winner 2024." https://polymarket.com/event/presidential-election-winner-2024?tid=1730675985444.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 443–80. https://doi.org/10.1146/annurev-economics-090122-040538.

team, The pandas development. 2020. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

The New York Times. 2024. "Cross-Tabs: September 2024 Inquirer/Times/Siena Poll of the Pennsylvania Likely Electorate." https://www.nytimes.com/interactive/2024/09/19/us/politics/times-siena-inquirer-poll-pennsylvania-likely-electorate.html.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.