Election 2024

Alexander Guarasci & Jacob Gilbert

October 24, 2024

This paper is about the election

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. 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 or Harris victory in each state. This approach allows us to analyze voter support patterns across the states and predict election outcomes more accurately.

The estimand in this study is the probability that Donald Trump or Kamala Harris wins 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

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…

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 towards candidates Donald Trump and Kamala Harris. Each entry in the dataset corresponds to a specific question around voting preference in a poll conducted by different polling organisations, which measure the percentage of respondents expressing their support for each candidate in their respective state.

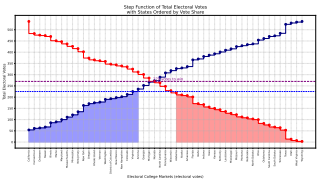
The polling data is collected through carefully structured survey questions, designed to elicit clear responses regarding voter preferences. By aggregating these individual responses, we are able to derive state-level sentiment for both candidates. This transformation of general voter sentiment into specific data points enables us to analyse the competitive landscape between Trump and Harris across swing states, providing a clearer understanding of electoral dynamics as they relate to the upcoming election.

2.1 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 his 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 (**538?**). In figure 3 we see the ammount that each party hold according to those sources, with the remaining swing states, seen in the middle of this graph set to determine if either Harris or Trump recieve over 270th electoral votes.

```
51
51
```

```
C:\Users\jgilb\AppData\Local\Temp\ipykernel_13496\2994737549.py:74: UserWarning: Tight layou
  plt.tight_layout()
```



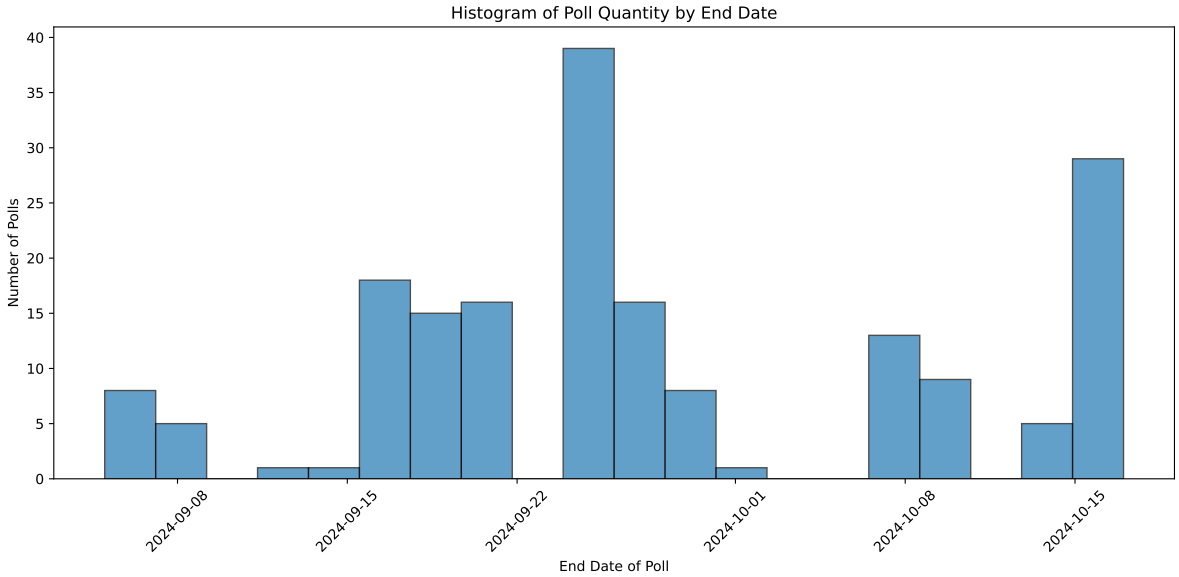
It's important to note that each pollster employs unique sampling techniques and methodologies that influence their Pollscore and Transparency Score. Pollscore is a measure of a pollster's historical accuracy, reflecting the quality of their sampling methods and data collection. a good score reflecting low propensity to be impacted by errors and biases that can occur in survey sampling. It also accounts for impacts that might bias a survey's score, such as adjustmenting 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 single star 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 Biden dropping out as a candidate on July 21st. 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

2.2 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 percentage of respondant favouring a candidate become the key variables 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 ar taking towards the build up of an election and thus ar taken by repondants at different times. Graph x 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 will adjust the weight that we place on polls that are closer to the election as these will be more indicative of how respondants will be voting on the day.



Another aspects 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 table x we can see the variety within the states polls. For use in the model the percentages in these polls is weighted for recenecy and size of the population.

3 Model {#sec-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 analy`
- `end_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:

$\text{recency\_weight} = \max(\text{current\_date} - \text{start\_date}) - (\text{current\_date} - \text{start\_date})$

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

Total Weight:

$\text{weight} = \text{sample\_size} \times \text{recency\_weight}$

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 va`
- `Independence: Observations are independent of one another.`

Limitations

1. `Polling Bias: The model is contingent upon the quality of the polling data. Polls with l`
2. `Dynamic Voter Sentiment: Voter preferences can change rapidly, particularly in the lead-`

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.

Conclusion

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 {#sec-results} The results of our analysis are as follows. We have Donald Trump predicted to win Arizona, Georgia and Michigan for a total of 42 of the 93 swing state votes. While Kamala Harris is projected to win Nevada, North Carolina, Pennsylvania and Wisconsin for the remaining 51 swing state votes. This results in a win for the democrats and the first female US president!

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?

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 favored 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?**). 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.

What Else Do We Learn About the World?

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 favor 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 favor 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 [https://www.oddsshark.com/entertainment/us-presidential-odds-2016-futures].

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). Additionally, the use of linear modeling may oversimplify the complex dynamics of voter behavior, as elections often involve non-linear 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.

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 behavior and election outcomes.

Apendix

(include some detail on how we are getting rid of biass. and focus on how we are spcfically uing stratifid and clustr sampling. how we will reach ou. how we will gt rsponce.)

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

In this revised approach, we will allocate the $100,000 budget not only for data collection and survey methods but also to incentivize participants to bet on the outcome of the election. This adds a monetary incentive for respondents to make accurate predictions, aligning their responses with their genuine expectations and potentially enhancing the predictive value of the data. Below is an outline of the methodology that integrates incentivized betting, sampling, recruitment, data validation, poll aggregation, and other features.

1. Sampling Approach

We will use stratified random sampling to capture a representative sample of the electorate across key demographics and regions. This ensures that our poll and betting pool reflects the diversity of the voting population, with emphasis on:

Demographic groups: Stratified by race, age, gender, education, and income. Geographic regions: Oversampling key battleground states (e.g., Pennsylvania, Michigan, Wisconsin,

Florida) to ensure robust state-level data while maintaining a representative national sample. Voter history: Stratified by past voter turnout, targeting likely voters, less frequent voters, and even non-voters to capture a full spectrum of voter engagement.

The goal is to survey 5,000 respondents, with a significant proportion from battleground states. By incentivizing respondents with the betting opportunity, the aim is to foster thoughtful responses, as they will have a monetary stake in their predictions.

2. Recruitment of Respondents

We will use a mixed-mode recruitment strategy, relying on both online and telephone methods to reach a broad audience:

Online recruitment: Participants will be recruited via digital platforms such as social media, news websites, and online political forums. Advertisements will emphasize the chance to bet on election outcomes and win money based on the accuracy of their predictions. Telephone recruitment: For older demographics and those less reachable online, telephone interviews will be conducted using voter registration rolls and commercial databases.

Incentives: The primary incentive is the betting pool, with $50,000 of the budget allocated to a reward fund for accurate predictions. Respondents will have the opportunity to place a bet (e.g., $10) on the candidate they believe will win, and those who predict correctly will share in the prize pool. This financial incentive should increase engagement and encourage respondents to think carefully about their answers.

3. Survey and Betting Design

The survey will gather standard polling data, but with the addition of a betting component:

Candidate preference: "If the 2024 U.S. Presidential Election were held today, would you vote for?" • Kamala Harris (D) • Donald Trump (R) • Other/Third-party candidate • Undecided Betting decision: After stating their candidate preference, respondents will be asked: "As a part of this survey you have been allotted $10 to bet on who you believe is most likely to win the election given the current odds. If the candidate of your choosing wins you will recieve a payout equal to the live odds. Which candidate would you like to place a bet on?" They will be given $10 to place on the candidate they believe is most likely to win.

Voter behavior and demographics: Questions about voter intention, past voting behavior, and demographic details will follow, ensuring the data can be properly stratified and analyzed.

Betting market engagement: "Have you participated in any betting markets (e.g., Polymarket, PredictIt) related to the election?" This will help analyze how participants in betting markets differ from typical poll respondents.

Election forecasting confidence: "On a scale of 1-10, how confident are you in your prediction of who will win the election?"

4. Data Validation

To ensure the quality of the data, several validation measures will be implemented:

• Consistency checks: Responses will be monitored for inconsistencies • Verification: Voter registration data will be cross-referenced to verify the eligibility of respondents. • Weighting: Data will be weighted to reflect the broader electorate using Census data on demographics and voting patterns.

5. Poll Aggregation and Modeling

After collecting the data, we will aggregate our polling results alongside other reliable polls and incorporate betting market data. This will allow us to adjust for discrepancies between public opinion polling and market-based predictions.

• Poll aggregation: Our results will be integrated with other national and state-level polls from reputable sources, such as Gallup and YouGov. Polls will be weighted based on sample size, recency, and historical accuracy. • Betting market integration: We will compare the betting behaviors within our sample to existing market data from platforms like PredictIt and Polymarket. This comparison will help adjust for potential biases in polling, as betting markets can reveal hidden preferences or voter reluctance to openly state their choices. • Modeling: We will use Bayesian hierarchical models to predict the election outcome at both the national and state levels. The model will integrate polling data, betting odds, and historical election trends. By comparing how betting and polling data align or diverge, we will refine the forecast, accounting for last-minute voter shifts or surprise outcomes.

6. Poll-Betting Market Discrepancy

A major component of this methodology is addressing the discrepancies between polling and betting markets. Our hypothesis is that participants are more likely to provide accurate predictions when they have a financial stake in the outcome.

• Betting behavior analysis: We will analyze how betting behavior correlates with stated voter preferences, identifying whether betting markets capture sentiment that traditional polling does not. • Incentive impact: By comparing the results of those who opted to bet versus those who did not, we can determine whether financial incentives lead to more accurate election predictions.

7. Budget Breakdown

The $100,000 budget will be allocated as follows:

• Respondent recruitment: $10,000 • Survey administration (online and telephone): $20,000 • Betting pool incentives: $50,000 (for those who accurately predict the election outcome) • Data validation and analysis: $10,000 • Poll aggregation and modeling: $10,000

Conclusion

This methodology uses a novel approach to incentivize accurate predictions by allowing respondents to bet on the election outcome. By combining traditional polling techniques with betting incentives, we aim to close the gap between polling data and betting market predictions, potentially uncovering hidden voter behaviors or biases that standard polling methods may miss.

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

appndix B

say what kind of ampling. reference rsearch on wha the sampling mthod is

References

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

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