Election 2024

Alexander Guarasci & Jacob Gilbert

October 30, 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 Matplot (Hunter 2007), Seaborn (Waskom 2021), Numpy (Harris et al. 2020), Pandas (team 2020), Jupyter (Jupyter 2023), Tabulate (Korolev 2023), 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 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.

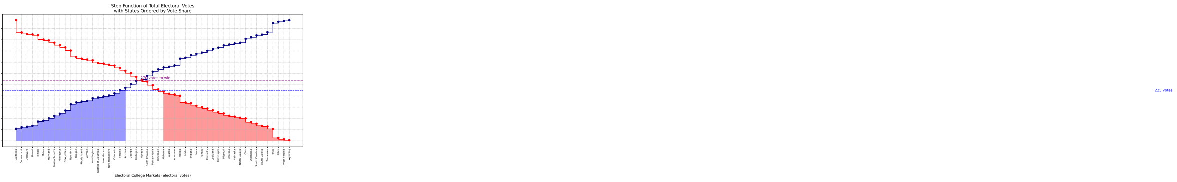
The polling data is collected through carefully structured survey questions, designed to elicit clear responses regarding voter preferences. By aggregating these individual responses, we are able to derive state-level sentiment for both candidates. This transformation of general voter sentiment into 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.

2.2 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 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 FiveThirtyEight (2024). 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
```

```
/var/folders/0d/ldk0yf0s73jfz51ll7sgchgr0000gn/T/ipykernel_19108/2994737549.py:74: UserWarnin
  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.

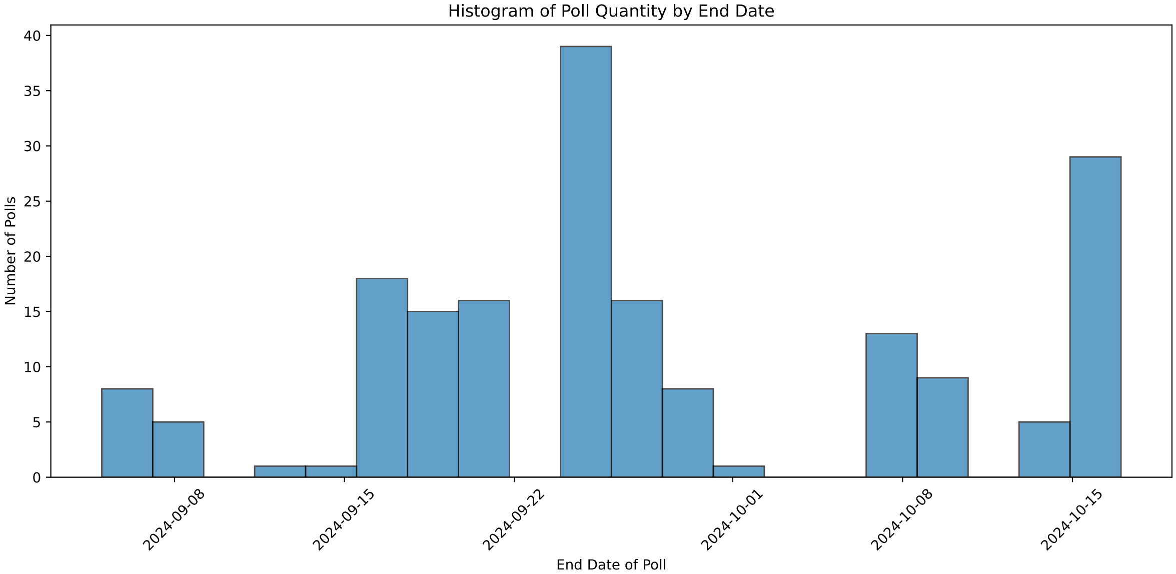
In keeping with this aim, we focused our analysis on polls conducted after September. Early in the election cycle, polls can be significantly influenced by initial campaign events, such as the shocks from the first debates or unexpected changes in the candidate lineup, for example Biden dropping out 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

2.3 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 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

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:

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

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 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

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.

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

1. 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.

2. 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 minmize the attrition of the respondants, 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 and family, as well as their community are expecting to win.

3. 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.

4. 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.

5. 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.

6. Question Design and Goal Allignment

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 be 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 surveys anonymity to minimize social desireablity 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.

appndix B

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

References

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.

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.

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.

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

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.

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.