# Forecasting the Winner of the Upcoming US Presidential Election*

Yanzun Jiang, Chang Tong, Wenxin Xu

November 5, 2024

In this study, we develop a model to predict the outcome of the upcoming U.S. presidential election in 2024 using a simulated dataset representing U.S. voters. The paper uses polling data and statistical programming in R to generate and analyze voter opinions. The findings reveal key demographic trends that influence voting preferences, including age, gender, education, and employment status. This analysis is important because it helps us understand the underlying factors that drive voter behavior and thus contributes to a clear view of the dynamics of democratic elections.

## 1 Introduction

Predicting election results is an important task that helps people understand the direction of public choice and predict the future political landscape. The latest poll data from America's Political Pulse shows that this election is a competition between major candidates including Donald Trump, Kamala Harris and other major contenders, and the final results of the US election also affect the landscape around the world. In this paper, we aim to promote this understanding by developing a predictive model using polling data.

The main focus of this paper is to use a simulated dataset representing American voters, using data from polling methods. The model is built using statistical programming in R to predict the results of the upcoming US presidential election. The quantities estimated in this analysis depend on demographic predictors such as age, gender, education, race, employment status, and state of residence. To this end, the dataset was organized and adjusted, and rigorous testing was performed to ensure data quality. Our analysis reveals key demographic trends that affect voting behavior, such as (to be updated)

*Code and data are available at: https://github.com/Stary54264/US_Election_Prediction.

The structure of this paper is as follows: Section 2 We first outline the data collection and cleaning process and describe the results and predictors used in the analysis. **?@sec-model** introduces the predictive models and discusses the reasons for choosing them for election results prediction. **?@sec-result** then presents the main findings, displays the results, and discusses the implications of our findings. Finally, **?@sec-discussion** explains the results, highlights important trends and predictions, and concludes with a discussion of the reliability of the predictions and potential limitations of the model. These findings are important because they provide insights into the factors that influence voter decisions and enable us to better predict election outcomes.

# 2 Data

## 2.1 Overview

This article uses the statistical programming language R (**citeR?**) to simulate and analyze data. The dataset used in this analysis comes from the polling method, specifically America's Political Pulse to predict the results of the upcoming US presidential election. Following the practice of (**tellingstories?**), we consider the robustness of aggregate polling data to effectively measure voter opinion at the macro level.

This analysis is based on a simulated dataset with 500 observations on 7 variables. The data is specially collated to provide a representation of the US electorate. For quality assurance, performed multiple integrity tests to verify the properties of the dataset, such as the number of rows and columns, unique values, and missing data.

## 2.2 Measurement

# A Appendix

## A.1 America's Political Pulse Methodology Analysis

### A.1.1 Discussion of Methodology and Key Features

The target population for this survey includes American voters who have specific views on political issues, such as partisan hatred and violations of democratic norms. The sampling frame may be based on paid survey respondents on the YouGov survey platform. This survey involved 70,825 unique respondents and a total of 115,000 responses.

The survey uses an online panel provided by YouGov, and participants are paid for their responses, which is non-probability sampling. This method of recruitment is efficient and allows for a continuous flow of respondents. However, online recruitment allows for a large sample to be obtained quickly but may introduce self-selection bias. Those who are incentivized by payment and have access to the Internet are more likely to participate, which may limit representativeness.

There is no explicit mention of how non-response is managed in this survey. Without proper handling of non-response, the survey results may be skewed. This means that certain perspectives might be over represented, while others could be underrepresented, impacting the overall reliability and accuracy of the findings.

### A.1.2 Strengths and Weaknesses

Large, ongoing sample size with over 70,000 unique respondents enhances reliability. Weekly data collection allows for trend analysis over a significant period. Publicly available dataset increases transparency and opportunities for independent analysis.

The primary drawback is the lack of representativeness. The findings may not generalize well to the entire population, as participants self-select into the survey and may not reflect key population segments. Lack of clarity on how non-response is addressed means that there could be gaps in representativeness, especially if certain groups consistently opt out.

---

## A.2 Appendix 2 - Idealized Methodology and Survey

This appendix outlines an idealized methodology and survey design to forecast the upcoming U.S. presidential election within a budget of $100,000. The approach involves detailed planning for sampling, recruitment, data validation, poll aggregation, and the use of survey platforms. The survey aims for high representativeness and reliability while remaining within the specified budget.

### A.2.1 Sampling Methodology

Stratified random sampling will be used to ensure accurate representation of different population groups. The sample will be stratified according to key demographics such as age, gender, ethnicity, income, geographic region, and education level.

**Target Sample Size:** Approximately 5,000 respondents.

### A.2.2 Respondent Recruitment

Respondents will be recruited using a combination of online panels and non-web-based methods.

1. Reputable panel providers such as YouGov will be contracted to provide a diverse panel of respondents, accounting for 60% of the sample.
2. The remaining 40% will be recruited using SMS and offline invitations to ensure that respondents without internet access are included, helping to minimize bias that may arise from exclusively online recruitment.

Respondents will be offered a small monetary incentive ($5 per participant) to complete the survey to encourage participation and minimize dropout rates.

### A.2.3 Survey Platform

The survey will be implemented using Google Forms for its ease of use and low cost. However, data collected will be exported and processed using a secure database and analyzed in software like R or Python for advanced statistical modeling.

Google Form Link

### A.2.4 Survey Questions for U.S. Presidential Election Forecast

Thank you for taking the time to participate in this survey regarding the upcoming U.S. Presidential Election. This survey aims to understand voter preferences as we approach the election.

Your responses will contribute to a larger effort to forecast the election results and to gain insight into voter engagement across different demographic groups.

All responses are completely anonymous, and your privacy is of utmost importance. The survey should take approximately 3 minutes to complete.

Your participation is entirely voluntary, and you can choose to stop at any time. We greatly appreciate your time and honest input.

**Survey Questions: - Are you registered to vote in the upcoming presidential election?**
- Yes - No

- **If the election were held today, which candidate would you most likely vote for?**
  - Democrat
  - Republican
  - Third-Party Candidate
  - Not Planning to Vote
  - Undecided
- **How certain are you about your choice?**
  - Very Certain
  - Somewhat Certain
  - Not Very Certain
  - Not Certain at All
- **What is your gender?**
  - Male
  - Female
  - Non-binary/Other
  - Prefer not to say
- **What is your race or ethnicity?**
  - White/Caucasian
  - Black/African American
  - Hispanic/Latino
  - Asian
  - Native American/Indigenous
  - Other
  - Prefer not to say
- **What is your age group?**
  - 18-24
  - 25-34
  - 35-44
  - 45-54
  - 55-64
  - 65 and above
- **What is the highest level of education you have completed?**

  - High School Diploma or Equivalent
  - College
  - Bachelor's Degree
  - Graduate or Professional Degree
  - Other

- **Which political party do you most identify with?**
  - Democratic Party
  - Republican Party
  - Independent
  - Other/None

- **Do you think the current voting system is fair and secure?**
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree

### A.2.5 Data Validation

IP addresses and contact information will be checked to identify and remove duplicate responses. Once the data is collected, weighting adjustments will be applied to ensure that the sample matches U.S. population demographics according to Census data.

### A.2.6 Budget Allocation

The budget of $100,000 will be allocated as follows: - Physical and Online Recruitment: $30,000 - Respondent Incentives: $30,000 - Data Collection & Validation: $10,000 - Statistical Consulting/Data Analysis: $20,000 - Other (Unforeseen Expenses): $10,000

## B References