2024 US Presidential Election Forecast Model*

My subtitle if needed

Viet Nguyen Yihang Xu Doran Wang

November 2, 2024

This paper uses aggregate polling data and 'STH' modeling methods to predict the 2024 U.S. presidential election outcome. The Model use the factors related to the polls on Donald Trump's support rate. The study also includes a deep-dive analysis of methodology on selected pollster, the New York Times (NYT), and then an idealized methodology and survey for predicting elections on a limited budget.

Table of Contents

*Code and data are available at: https://github.com/vietng04/2024-US-Presidential-Election-Forecast.

1 Introduction

The 2024 U.S. Presidential Election Forecast Model aims to provide a comprehensive analysis of polling data to predict potential outcomes in the upcoming election. Utilizing a robust dataset that includes over 50 variables, such as pollster ratings, sample sizes, election dates, candidate affiliations, and polling methodologies, this model seeks to offer a nuanced understanding of voter preferences and trends across the country.

Key factors, such as pollster transparency scores, polling population characteristics, and election stages, are incorporated to ensure accuracy and relevance. By leveraging this data, the model can account for regional dynamics, partisan leanings, and shifts in voter sentiment, providing detailed insights into the evolving electoral landscape. This forecast model represents a data-driven approach to understanding the political climate as candidates navigate the road to the presidency.

2 Data

The data used in this paper came from the FiveThirtyEight (FiveThirtyEight 2024). Data were cleaned and analyzed using the open source statistical programming language R (R Core Team 2023). Libraries `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `knitr` (Xie 2022), and `dplyr` (Wickham et al. 2023) were used for simulating, cleaning and testing. Graphics were made using `ggplot2` (Wickham 2016).

2.1 Sample Data

Table 1: Sample Cleaned Data

pollster	display_name	pollster_rating_id	numeric_grade	pollscore
Siena/NYT	The New York Times/Siena College	448	3	-1.5
Siena/NYT	The New York Times/Siena College	448	3	-1.5

Table 1: Sample Cleaned Data

pollster	display_name	pollster_rating_id	numeric_grade	pollscore
Siena/NYT	The New York Times/Siena College	448	3	-1.5
Siena/NYT	The New York Times/Siena College	448	3	-1.5
Siena/NYT	The New York Times/Siena College	448	3	-1.5

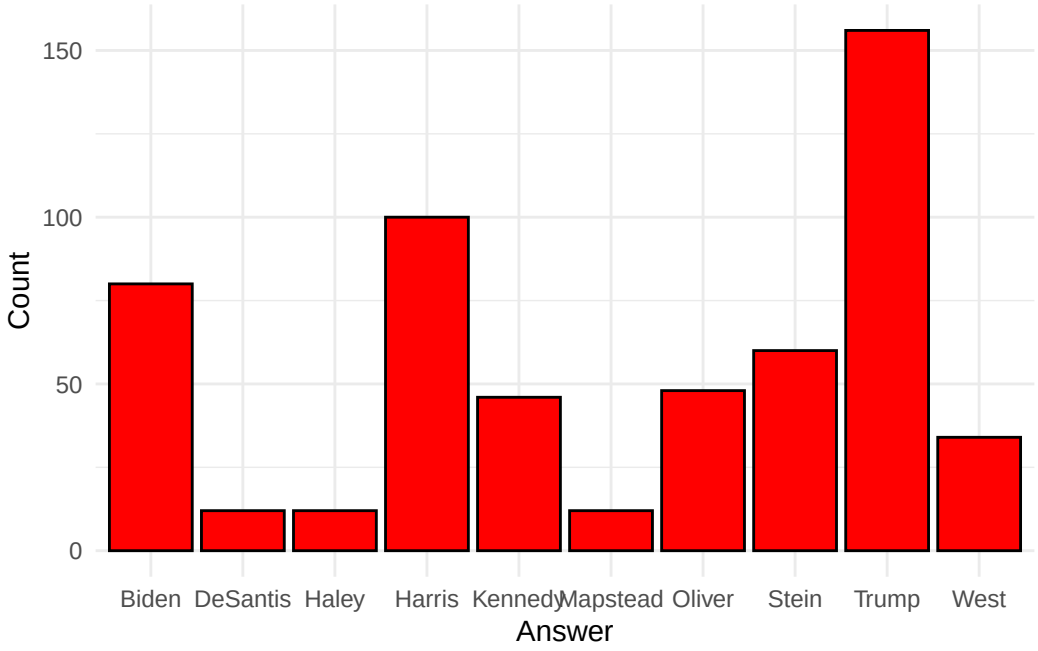

Figure 1: Distribution of Answers

3 Model

In the 2024 U.S. Presidential Election Forecast Model, the percentage of voter support for a candidate, represented by the variable pct, is modeled as the response variable. The predictors in this model include the candidate's political affiliation (party), the size of the sample surveyed (sample_size), the rating of the pollster based on their historical accuracy (numeric_grade), and the geographic location of the poll, represented by the variable state. By incorporating these predictors, the model aims to account for both the qualitative and quantitative factors

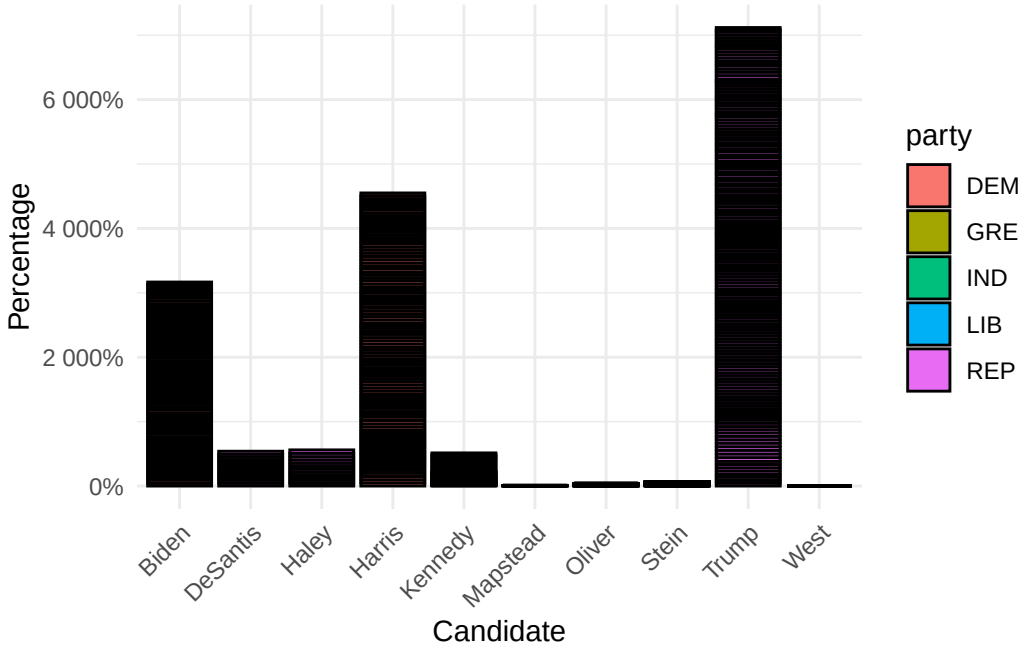

Figure 2: Percentage of Support for Each Candidate

that influence polling outcomes. The party variable captures the effect of political affiliation on voter support, while sample_size adjusts for the variability in poll precision. Numeric_grade reflects the reliability of the pollster, and state introduces regional variations in voting preferences. This model structure enables a detailed analysis of the factors driving election polling percentages across different states and political contexts.

3.1 Model set-up

$$\text{pct} = \beta_0 + \beta_1 \cdot \text{pollscore} + \beta_2 \cdot \text{sample size} + \beta_3 \cdot \text{numeric grade} + \epsilon$$

where β_0 is the intercept, β_1, β_2, β_3, β_4 are the coefficients associated with the predictors, and ϵ represents the error term. This model accounts for both qualitative variables, like the party and state, as well as quantitative variables, such as sample size and pollster rating.

3.1.1 Model justification

```
Call:
lm(formula = pct ~ pollscore + sample_size + numeric_grade, data = clean_data)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-30.90 -25.78  11.21  16.38  26.94

Coefficients: (2 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   40.758456   5.410781   7.533 2.01e-13 ***
pollscore            NA         NA      NA       NA
sample_size   -0.016314   0.007813  -2.088   0.0372 *
numeric_grade        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.33 on 558 degrees of freedom
Multiple R-squared:  0.007753,  Adjusted R-squared:  0.005975
F-statistic:  4.36 on 1 and 558 DF,  p-value: 0.03724
```

4 Results

5 Discussion

Appendix

A Discussion of the survey's methodology of Siena/NYT

A.1 Population, Frame, and Sample

-Population:

-Frame:

-Sample: a topline sample size of at least 100

A.2 Sample Recruitment Method

A.3 Sampling Approach's trade-offs

-Strengths:

-Weaknesses:

A.4 How is non-response handled

A.5 What is good and bad about the questionnaire

A.5.1 Advantages

A.5.2 Disadvantages

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. *2024 National Presidential Poll Results*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.