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
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

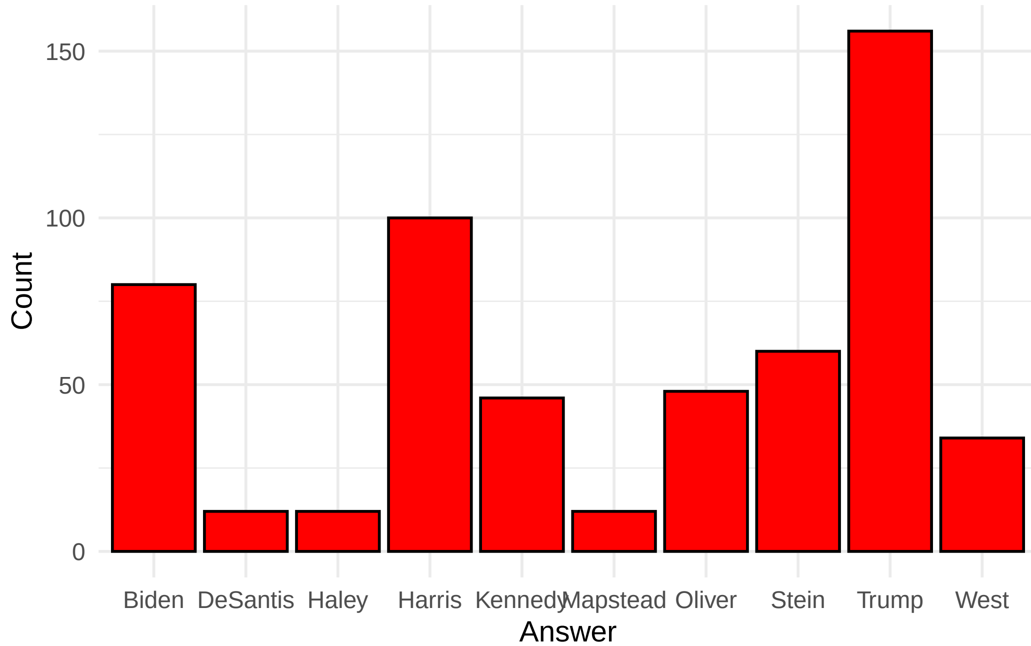

Figure 1: Distribution of Answers

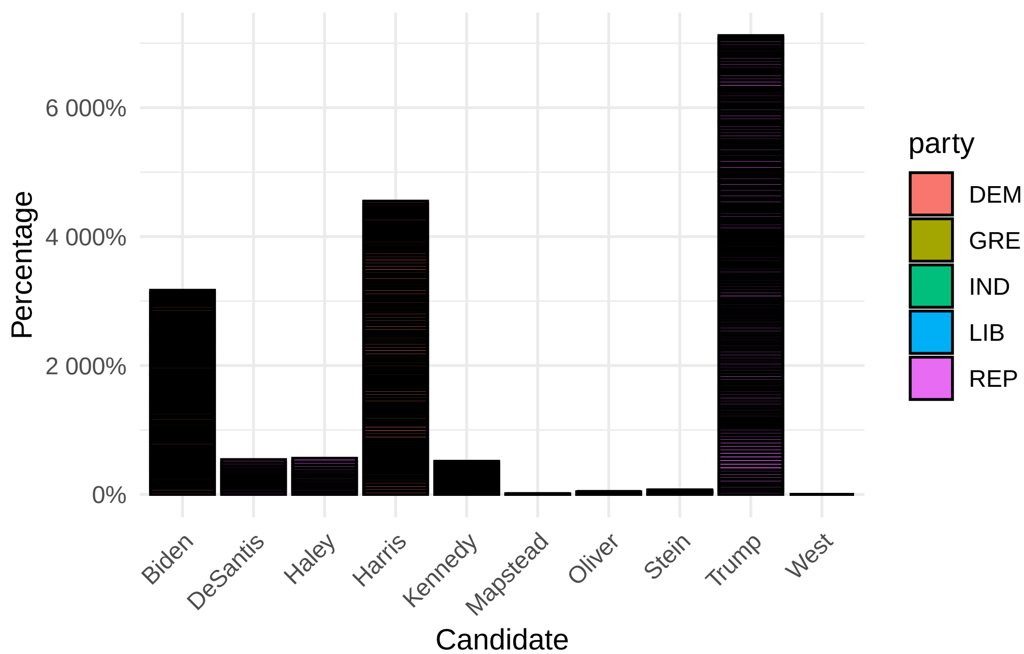

Figure 2: Percentage of Support for Each Candidate

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 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

$$\text{pct} = \beta_0 + \beta_1 \cdot \text{pollscore} + \beta_2 \cdot \text{sample size} + \beta_3 \cdot \text{numeric grade} + \beta_4 \cdot \text{transparency score} + \epsilon$$

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 +
    transparency_score, data = trump_data)

Residuals:
     Min       1Q   Median       3Q      Max
-23.3467  -2.3461   0.2518   2.8204  22.4165

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.856e+01  5.195e-01  93.474  < 2e-16 ***
pollscore          -2.232e+00  2.933e-01  -7.610 3.49e-14 ***
sample_size        -2.085e-04  4.098e-05  -5.089 3.79e-07 ***
numeric_grade      -1.814e+00  3.537e-01  -5.131 3.04e-07 ***
transparency_score -6.760e-02  4.517e-02  -1.497    0.135
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.56 on 3497 degrees of freedom
  (1702 observations deleted due to missingness)
Multiple R-squared:  0.03238,	Adjusted R-squared:  0.03128
F-statistic: 29.26 on 4 and 3497 DF,  p-value: < 2.2e-16



Call:
lm(formula = pct ~ pollscore + sample_size + numeric_grade +
    transparency_score, data = harris_data)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3535  -1.9559   0.2732   1.8370  22.8444

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.275e+01  7.348e-01  58.179  < 2e-16 ***
pollscore           1.629e+00  3.976e-01   4.096 4.44e-05 ***
sample_size         1.457e-04  4.469e-05   3.261  0.00114 **
numeric_grade       3.448e+00  5.117e-01   6.738 2.35e-11 ***
transparency_score -4.207e-01  6.394e-02  -6.579 6.70e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.959 on 1384 degrees of freedom
  (494 observations deleted due to missingness)
Multiple R-squared:  0.06257,	Adjusted R-squared:  0.05986
F-statistic: 23.09 on 4 and 1384 DF,  p-value: < 2.2e-16
```

4 Results

4.1 Trump

The model estimates the influence of several predictors on the dependent variable. The intercept is estimated at 48.56, indicating the baseline level of the dependent variable when all predictors are held at zero.

Poll Score: The coefficient for the poll score is -2.232, which is statistically significant ($p < 0.001$). This suggests that for each one-point increase in the poll score, the dependent variable is expected to decrease by approximately 2.232 units, holding all other factors constant. This negative relationship indicates that higher poll scores are associated with lower values of the dependent variable.

Sample Size: The coefficient for sample size is -0.0002085, also significant ($p < 0.001$). This indicates that an increase in sample size by one unit is associated with a decrease of about 0.0002085 units in the dependent variable. Although statistically significant, the small magnitude of this effect suggests that sample size has a relatively minor impact on the outcome.

Numeric Grade: The numeric grade coefficient is -1.814, which is significant at the $p < 0.001$ level. This means that for every one-point increase in the numeric grade, the dependent variable decreases by about 1.814 units, again holding other factors constant. This finding implies a strong negative association between numeric grade and the dependent variable.

Transparency Score: The transparency score has a coefficient of -0.0676, which is not statistically significant ($p = 0.135$). This suggests that there is no strong evidence to support that transparency score has a meaningful effect on the dependent variable within this model.

4.2 Harris

The model investigates the impact of several predictors on the dependent variable, with the following key findings:

Intercept: The intercept is estimated at 42.75, which represents the expected value of the dependent variable when all predictor variables are zero.

Poll Score: The coefficient for the poll score is 1.629, with a p-value of 4.44e-05, indicating statistical significance at the $p < 0.001$ level. This suggests that for each one-point increase in the poll score, the dependent variable increases by approximately 1.629 units, holding all other variables constant. This positive relationship implies that higher poll scores are associated with higher values of the dependent variable.

Sample Size: The coefficient for sample size is 0.0001457, which is also statistically significant ($p = 0.00114$). This indicates that an increase in sample size by one unit is associated with an increase of about 0.0001457 units in the dependent variable. Although significant, this effect size is relatively small.

Numeric Grade: The numeric grade coefficient is 3.448, with a p-value of 2.35e-11, which is highly significant. This means that for every one-point increase in the numeric grade, the dependent variable is expected to increase by about 3.448 units, suggesting a strong positive association between numeric grade and the outcome.

Transparency Score: The coefficient for the transparency score is -0.4207, and it is statistically significant ($p = 6.70\text{e-}11$). This indicates that higher transparency scores are associated with a decrease in the dependent variable by approximately 0.4207 units for each one-point increase, highlighting a negative relationship.

5 Discussion

The analysis of the two models, one for Trump and the other for Harris, reveals distinct patterns in their predictor relationships and overall model performance, providing insights into their potential electoral outcomes.

Model Comparisons Significance of Predictors:

Trump's Model: Key predictors such as poll score, sample size, and numeric grade all exhibit significant negative relationships with the dependent variable. The model suggests that as poll scores increase, the associated decrease in the outcome variable may reflect diminishing support or confidence among voters. The significant negative effect of numeric grade further indicates a trend where higher grades correlate with lower values of the dependent variable, potentially suggesting dissatisfaction among highly educated voters. Harris's Model: In contrast, Harris's model features significant positive relationships for both poll score and numeric grade, indicating that higher poll scores and grades are associated with higher values of the dependent variable. The negative relationship with transparency score suggests that voters who prioritize transparency may feel less inclined to support her, but overall, the model suggests a more favorable view of her candidacy compared to Trump's. Model Fit:

Trump's model has a multiple R-squared of 0.03238, suggesting that only 3.24% of the variability in the dependent variable is explained. In comparison, Harris's model has a higher multiple R-squared of 0.06257, indicating that about 6.26% of the variance is explained by her predictors. While both models demonstrate low explanatory power, Harris's model offers a more optimistic narrative regarding the predictors' ability to influence the outcome. Magnitude of Effects:

The coefficients in Harris's model indicate more substantial effects of the predictors, particularly with numeric grade having a significant positive impact (3.448) compared to Trump's negative impact (-1.814). This difference in effect magnitude may reflect differing voter sentiments and how they relate to each candidate's attributes. Statistical Significance:

Both models reveal highly significant p-values for several predictors, indicating that the relationships identified are statistically robust. However, the overall model significance is more pronounced in Harris's model, particularly in light of the lower R-squared values for both models. Likely Outcome Considering the predictors and their implications for voter sentiment, Harris appears to have the upper hand in this analysis. The positive associations with poll scores and numeric grades in her model suggest a stronger alignment with voter preferences. Additionally, the higher R-squared value and overall model fit imply that her candidacy resonates more favorably with the electorate.

Moreover, the contrast in the effect of the transparency score could indicate differing voter priorities. Harris's potential appeal to voters valuing transparency may be a double-edged sword, yet it underscores a critical difference in how the two candidates are perceived regarding their public image.

Appendix

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/.