Quantifying Predictive Ambiguity: Navigating Uncertainty, Constraints, and Theoretical Boundaries in Presidential Election Forecasting

Gabriella Martinez and Gabriel Campos

City University of New York, School of Professional Studies

DATA 698: Masters Research Project

Author Note

This research was conducted as part of a course project with no external funding.

Correspondence concerning this article should be addressed to Gabriella Martinez and Gabriel Campos, City University of New York, School of Professional Studies, 119 W 31st St, New York, NY 10001

Emails: gabriel.campos77@spsmail.cuny.edu and gabriella.martinez@spsmail.cuny.edu

Abstract

Data science has become integral to modern business practices, with organizations across various sectors implementing continuous monitoring and predictive analytics to enhance performance and meet key performance indicators (KPIs). Recent advancements have advanced statistical and data-driven methodologies, allowing for the integration of both quantitative and qualitative indicators in outcome forecasting. However, certain sectors, such as election outcome prediction, remain complex due to the multitude of influencing factors. Despite these challenges, improving the prediction of presidential elections is crucial for a comprehensive understanding the democratic processes. Employing data science techniques in election forecasting not only represents a valuable learning opportunity for students but also contributes to the broader field of political analysis.

Historically, election forecasting has depended on polling and surveys, however both methods are susceptible to bias and unrepresentative samples of the broader electorate. More recent methods, like social media analysis, attempt to address this but face similar issues. This thesis aims to address these challenges by developing a data-driven model using historical voter data, with a focus on Random Forest algorithms, as taught in the CUNY School of Professional Studies curriculum.

By integrating demographic data such as age, education, and party affiliation, using county-level census data and voter turnout records, the study aims to better understand the complexities of voter behavior. Additionally, the research will explore the social and political dynamics not incorporated into the model however, significant in influencing outcomes. The final analysis will also reflect on the broader social and political factors that might affect election outcomes. By comparing this approach to traditional models, the thesis will offer valuable insights into election forecasting challenges and improvements.

Keywords: key performance indicators, presidential election forecasting, voter behavior analysis, Random Forest algorithm, predictive modeling, categorical data integration, polling bias, social media, count-level census data, voter turnout, political analysis.

Quantifying Predictive Ambiguity: Navigating Uncertainty, Constraints, and Theoretical Boundaries in Presidential Election Forecasting

Election forecasting presents ongoing challenges for both political scientists and data analysts due to the complex nature of voter behavior. Traditional tools such as opinion polls and surveys have long been used, but their predictive accuracy is frequently undermined by biases and unrepresentative sampling. The increasing polarization of the political landscape as noted by Ethan Rosen, Associate General Counsel of PredictIt, during the *'Predicting Elections in an Unstable Political Environment*' panel at the Columbia University Political Analytics Conference (March 22,2024), has further complicated the forecasting process. Rosen remarked that 'this era is more unstable than previous eras' and emphasized the issue of 'hyperpolarization that we're dealing with in our society', which underscores the limitations of conventional forecasting methods.

As data science becomes more integrated into political analysis, algorithms like Random Forest have gained attention for their ability to analyze complex data. This thesis uses Random Forest to predict presidential election outcomes by incorporating detailed demographic information such as age, education, and party affiliation. However, the focus is not on optimizing or refining the algorithm itself. Instead, the study aims to compare the effectiveness of the algorithm against traditional forecasting tools, like polls and media predictions, as well as the actual election results. Through this comparison, the study will highlight the differences in accuracy between a data-driven algorithmic approach and more conventional prediction methods.

This research will also examine the role of social and political dynamics not typically included in data-driven models. Through a comparative analysis of historical forecasting techniques and this proposed approach, the thesis aims to identify both strengths and weaknesses. Ultimately, the goal is to contribute to the broader discourse on electoral predictability, addressing critical shortcomings and advancing the field of political data analysis.

Furthermore, the study will consider past election trends, particularly the enduring influence of the two-party system, and explore its relevance in forecasting models. In addition, the accuracy of Random Forest predictions will be evaluated both at the national level and in key battleground states, where, as Joe Lenski, Exit Poll Director at Edison Media Research, noted during the Columbia University Political Analytics Conference (2024), 'narrow margins in very key states are determining the winners and losers.' This comparison will help assess whether the algorithm can provide greater precision in close races, where traditional polling often struggles. Finally, the study will reflect on broader socioeconomic factors that influence electoral outcomes.

Accounting for Intangibles: Limitations and Justification in Election Forecasting

The competence of voters in selecting presidential candidates is frequently scrutinized, with motivations often appearing disconnected from candidates' policy positions, historical actions, or stances on major issues. In *Predicting Elections: Child's Play!*, Antonakis and Dalgas invoke Plato's writings from The Republic to elucidate the flawed processes through which voters navigate their civic duty. Plato's allegory of a ship, captained by a figure who is physically imposing but lacks adequate vision and knowledge of navigation, serves as a metaphor for electoral behavior. The crew (i.e., voters), misled by appearances, are unable to select a capable captain (i.e., leader). This allegory exposes the limitations in rational voter behavior, a dynamic that is difficult to model using algorithms like Random Forest. Antonakis and Dalgas assert that voters may be swayed by superficial traits—such as a candidate's physical attractiveness or charisma—rather than substantive policy issues. However, this poses challenges: how do we measure charisma or attractiveness? Should facial features be the primary focus, or should factors like attire, speech clarity, and vocal tone also be considered? Quantifying these subjective qualities would necessitate advanced methodologies, such as developing a framework for

rating physical attractiveness or charisma. Moreover, voter behavior driven by such factors may not be easily captured through traditional surveys or polling.

In a broader academic discourse, Ahearn, Brand, and Zhou's (2023) research offers a substantial contribution to understanding the intersection between education and civic engagement. Their empirical findings demonstrate that while educational attainment is positively associated with voter turnout, particularly among marginalized groups, "civic returns to college do not hinge on its socioeconomic returns; instead, they appear to stem primarily from the college experience itself." This conclusion emphasizes the intrinsic value of higher education in fostering civic participation, beyond the economic advantages it may confer. Notably, their study further highlights that "individuals with a lower likelihood of attending college, who tend to have more disadvantaged backgrounds, experience greater increases in self-reported voting due to college attendance." This observation underscores the significant role that higher education plays in mitigating voter turnout disparities across socioeconomic lines. As such, while educational attainment is a critical predictor of voter turnout, it offers limited explanatory power when it comes to understanding individual voting preferences.

Historically, voter turnout has significantly impacted the accuracy of election predictions. As John R. Petrocik points out, "turnout was not the only source of error, but it displayed one of the largest correlations with accuracy." In his analysis of Crespi's (1984) study of 423 pre-election polls, Petrocik found that, on average, polls missed the actual vote distribution by nearly six percentage points. Interestingly, Petrocik observes that "polls which attempt to factor in turnout were not measurably better at predicting outcomes than polls which ignored it." Similarly, Traugott and Tucker's (1984) turnout predictor faced challenges in accurately forecasting who would vote. Their model, although sophisticated, suffered from social desirability bias, resulting in inflated turnout estimates. Petrocik acknowledges that "the difficulty of predicting turnout is a persistent problem," highlighting the gap between respondents' stated voting intentions and their actual behavior.

These insights are essential for data scientists seeking to refine predictive models by incorporating variables such as voter turnout and education. However, it is equally important to recognize the limitations of these variables and to avoid overly complex models that may obscure key findings with unnecessary qualitative factors.

Beyond employing historical election forecasting methods, the endurance of the two-party system remains a key element in understanding electoral outcomes. Tom Rice's review of Forecasting Presidential Elections by Steven J. Rosenstone highlights the influence of the electoral environment on voting behavior. According to Rosenstone, "If we can identify the important environmental factors and specify their impact on the vote, accurate predictions should be forthcoming." Central to this environment is party identification, a long-term force that plays a crucial role in shaping voter decisions. Gradual shifts in party loyalty can alter election outcomes, but these shifts are slow and often hard to predict.

While factors like incumbency and regional voting patterns also influence outcomes, party affiliation remains the most reliable predictor. This idea was reinforced during the "*How to Spend $20 Billion: Media Strategy and Data Analytics*" panel at the Columbia University Political Analytics Conference 2024. Dr. Doug Usher and Lee Dunn discussed the inefficiencies of political advertising, citing John Wanamaker's famous remark: "Half the money spent on advertising is wasted, the trouble is I don't know which half." Dunn extended this observation to today's context, saying, "You might change that today to say 98% of what we spend on political advertising is wasted—we just don't know what 98%." Despite billions being spent on campaigns, only 5,000 to 10,000 voters are typically swayed. However, this small number can be critical in deciding close elections, highlighting the resilience of party loyalty and how difficult it is to sway voters en masse.

When developing election forecasts, it's important to account for this voter consistency. Party loyalty is deeply ingrained, and while it's possible to influence a small segment of the electorate, the challenge lies in identifying where this shift will occur, especially in tight races. Understanding this dynamic is crucial for generating accurate forecasts, as the small percentage of voters swayed can be the difference in a highly contested election.

Leveraging proven factors for enhanced Predictive Accuracy in election Models

The selected algorithm for our prediction is Random Forest, which operates by constructing numerous decision trees, each trained on distinct data points. The model aggregates the predictions of all the trees through a voting mechanism to arrive at a final outcome. To enhance precision beyond a binary comparison of electoral winners and losers, the analysis will incorporate voter turnout data, recognized as a significant factor in elections, albeit historically underutilized for predictive purposes. This approach allows for a detailed performance assessment in key states, examining both voter turnout and electoral results. The qualitative indicators employed will be limited to education and age. This decision is informed by prior research indicating that an excessive number of indicators can adversely affect results and specifically excludes variables that are challenging to quantify (such as attractiveness) or subjective, like Allan Lichtman's charisma variable in *The 13 Keys to the White House* (Madison Books, 1991). Striving for model simplicity is a primary objective, as we aim to evaluate what we believe are contributing or non-contributing variables. This methodology may serve as a foundational baseline for future studies, where additional data derived from newly established frameworks can be incorporated.

References

Ahearn, C. E., Brand, J. E., & Zhou, X. (2023). How, and for whom, does higher education increase voting? Research in Higher Education, 64(4), 574-597. https://doi.org/10.1007/s11162-022-09717-4

Abramowitz, A. I. (2010). The disappearing center: Engaged citizens, polarization, and American democracy. Yale University Press. https://www.jstor.org/stable/j.ctt1njms8.9

Columbia University. (2024, March 22). Predicting elections in an unstable political environment [Panel discussion]. Columbia University 2024 Political Analytics Conference.

Columbia University. (2024, March 22). Is polling dead? Facing the challenges in measuring the electorate [Panel discussion]. Columbia University 2024 Political Analytics Conference.

Columbia University. (2024, March 22). How to spend $20 billion: Media strategy and data analytics [Panel discussion]. Columbia University 2024 Political Analytics Conference.

Columbia University. (2024, March 22). Strategic voter mobilization and persuasion [Panel discussion]. Columbia University 2024 Political Analytics Conference.

Hillygus, D. S. (2011). The evolution of election polling in the United States. The Public Opinion Quarterly, 75(5), 962-981. https://www.jstor.org/stable/41345918

Lichtman, A. (2020). The keys to the White House: Forecast for 2020. Harvard Data Science Review, 2(4). https://doi.org/10.1162/99608f92.baaa8f68

Pasek, J. (2015). The polls–review: Predicting elections: Considering tools to pool the polls. The Public Opinion Quarterly, 79(2), 594-619. https://www.jstor.org/stable/24546379

Rice, T. W. (1985). Predicting presidential elections: A review essay. The Western Political Quarterly, 38(4), 675-686. https://www.jstor.org/stable/448620

Footnotes

[1]For APA reports, add footnotes manually on their own page following references. Do not use the **Insert Footnotes** method on the **References** tab as they will not be formatted correctly. For APA formatting requirements, it's easier to type your own footnote references and notes. To format a footnote reference, select the number and then, on the **Home** tab, in the **Styles** gallery, click **Footnote Reference**. The body of a footnote, such as this example, uses the **Normal** text style. If you delete this sample footnote, don't forget to delete its in-text reference at the end of the sample Heading 2 paragraph on the first page of body content in this template.

Tables

Table 1

Table Title

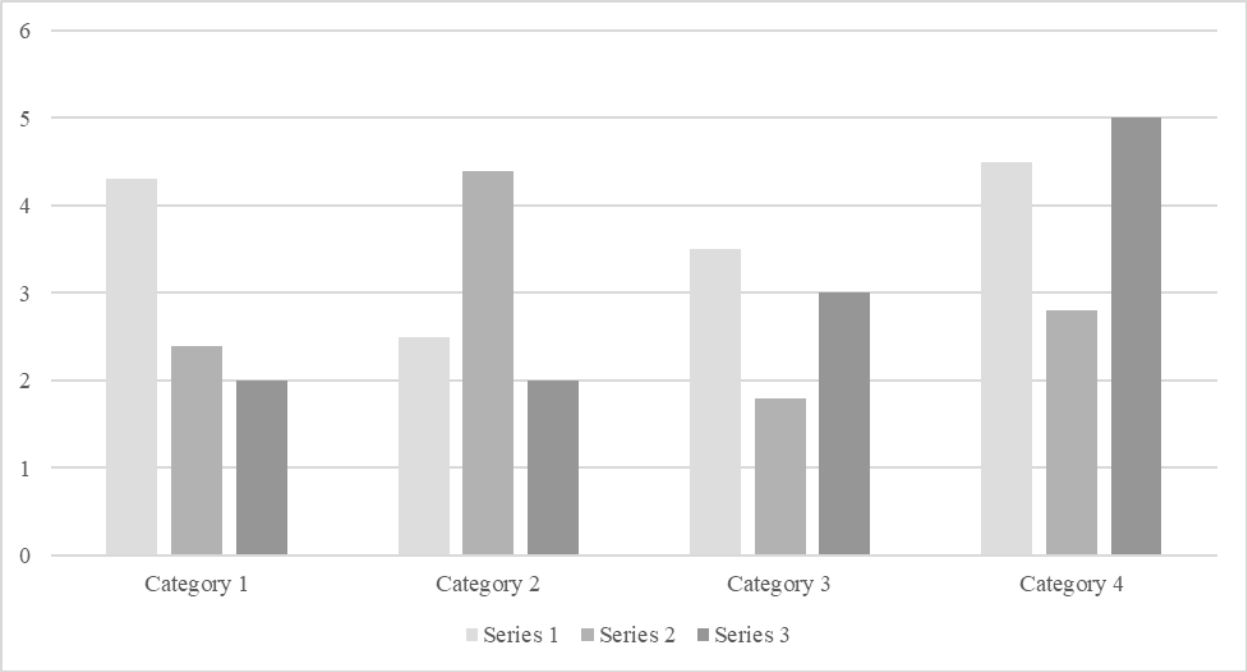
Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: Place all tables for your paper in a tables section, following references and footnotes. Start a new page for each table, include a table number and table title for each, as shown. All explanatory text appears in a table note that follows the table, like this one. Use the **Table/Figure** style, available on the **Home** tab, in the **Styles** gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been set up for this template that fits APA guidelines. To insert a table, on the **Insert** tab, click **Table**.

Figures Title

Figure 1.

Include all figures in their own section, following references, footnotes, and tables. Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.



For additional information on APA Style formatting, please consult the APA Style Manual, 7th Edition.