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: Limitation 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.

Categorizing and Refining Predictive Variables in Random Forest Models

The selection of parameters for the random forest algorithm fundamentally depends on the response variable and the predictor variables. To ensure the robustness and interpretability of our model, the response variable must be simplified and aligned with the conceptual framework established in prior research. In this case, utilizing a binary classification system based on a two-party structure is both logical and methodologically sound. This approach facilitates clear outcome interpretation and improves model performance. As part of our initial data preprocessing, observations where the majority winners at the county level were categorized as "Libertarian", "Other" or "Green" were excluded, as these categories represent a small proportion of total votes and deviates from the two-party paradigm that forms our analysis, ensuring we focus on dominant voting trends (see *Table 1*).

Initial Steps in Preprocessing the Dataset

The initial review of our county-level dataset revealed the presence of several unexpected values requiring further inquiry, including "Statewide Write In" for the state of Connecticut, "Maine UOCAVA" for the state of Maine, and "Federal Precinct" for the state of Rhode Island, (see *Table 2*). Upon investigation, it was found that the "Maine UOCAVA" record corresponds to votes submitted by

Uniformed Service and Overseas Citizens Absentee Voting Act (UOCAVA) voters, while the "Statewide Write In" for Connecticut represents votes for self-selected candidates not listed on the official ballot. The purpose of the "Federal Precinct" record in Rhode Island, however, remains unclear due to insufficient documentation or explanation provided by the data source.

Given that our analysis is focused on county-level voting trends, these records were excluded from further consideration. Nevertheless, their presence serves as a critical reminder of the complexities inherent in real-world datasets. Variations in data collection and reporting standards across states can introduce unexpected challenges, which must be addressed through thorough investigation and preprocessing.

Among the unexpected values identified in the dataset was the District of Columbia (see *Table 1*), a defined region encompassing a significant portion of the DC–VA–MD–WV Metropolitan Area population. Unlike the other excluded categories, this record represents a specific geographic area with a politically active resident population. Including it ensures that our analysis accounts for the unique voting behavior of this critical region. To address this, we conducted further research to confirm the Federal Information Processing Standard (FIPS) code corresponding to the District of Columbia. We determined that the appropriate FIPS code, 11001, should be applied to ensure its accurate representation within our county-level analysis framework.

Integrating Census Data and Addressing Missing Values

While the ideal scenario would involve relying on a single comprehensive data source to construct our random forest model, the complexity of our analysis necessitated integrating additional census data to improve the accuracy of approximations. To achieve this, we incorporated Citizen Voting Age Population (CVAP) data, which the Census Bureau generates using population estimates from the American Community Survey (ACS).

The datasets were merged using the Federal Information Processing Standard (FIPS) code for each county, covering the years 2008, 2012, 2016, and 2020. The distribution of missing values (NAs) in the merged dataset is shown in *Table 3*. The counts of NAs were highest in earlier years, with 3,154 missing entries in 2000 and 3,155 in 2004, compared to 39–40 NAs in more recent years (see *Table 4*). Although the number of missing values in earlier years was significant, we opted for full removal of these records. This decision was based on the compatibility of the datasets used in the merge, ensuring the analysis proceeded with the most reliable data available.

Resolving Data Conflicts and Aggregating Results

The next stage of preprocessing before beginning our exploratory data analysis (EDA) involved resolving conflicts in county-level data. Specifically, Jackson, Kansas City was recorded under both FIPS 20095 and 3600, while Bedford, Virginia used FIPS 51019 and 51515 (see *Table 5*). To streamline our analysis and following Allan Lichtman's emphasis on simplicity in *The 13 Keys to the White House*, we grouped the data by state to facilitate comparison with real-time results from the 2024 U.S. presidential election.

To further simplify the dataset and enhance the interpretability of our analysis, we aggregated county-level data into state-level majority winners. This adjustment reduces noise from third-party vote counts and aligns the dataset with the historical trend of major-party dominance in state-level outcomes. Historically, no third-party or independent candidate has achieved a majority in any state since 2008, with the most recent occurrence of third-party electoral votes dating back to George Wallace in 1968. This adjustment, therefore, has a negligible impact on the analysis while significantly enhancing clarity and interpretability.

Our resulting dataset accounts for total Democratic and Republican votes. These totals can be presented either as an aggregate across all states (see Table 6) or by individual state, depending on whether the data is grouped at the state level (see Table 7).

Addressing Alaska's Data Incompatibility

A distinct challenge with our county-level approach was encountered in Alaska. Alaska has 38 county-equivalent entries labeled as "District 1" through "District 40," excluding Districts 13 and 16 while also including District 99. Upon further research, we learned that Alaska's local governance is fundamentally distinct from other states. As documented by the Legislative Finance Division, Alaska comprises 19 non-unified boroughs and 19 home rule boroughs (State of Alaska, Legislative Finance Division, Dec. 2021, www.legfin.akleg.gov/InformationalPapers/21-028m-Local-Government-In-Alaska.pdf). Coupled with the absence of CVAP estimates, this structure rendered Alaska's data incompatible with our random forest model, leading to its exclusion.

Data Preparation and Feature Engineering for Random Forest Models

Feature Engineering for Enhanced Predictive Modeling

To enhance the performance of the random forest model, we engineered several features based on voter turnout and vote share data. Feature engineering involves creating new variables derived from existing ones to provide additional numeric information for predictive modeling. Voter turnout is a major predictor of election outcomes, making it a natural choice for transformation into additional metrics.

As noted by **Fatemeh Nargesian et al.** in their work *Learning Feature Engineering for Classification*, "Evaluation-based and exhaustive feature enumeration and selection approaches result in high time and memory cost and may lead to overfitting due to brute-force generation of features" (Nargesian et al. 2). Keeping this caution in mind, we designed a series of derived variables to enrich our dataset while mitigating the risks of overfitting.

The first category of variables includes **voter share metrics**, which express the share of total votes attributable to different groups. For instance, voter_share_major_party calculates the proportion of total votes received by the two major parties combined, while voter_share_dem and

voter_share_gop measure the shares for the Democratic and Republican parties individually. Additionally, voter_share_other represents the vote share for candidates outside the two major parties.

We also introduced **raw difference metrics** to quantify absolute differences in vote counts between competing parties. For example, rawdiff_dem_vs_gop measures the raw difference between Democratic and Republican votes, while variables like rawdiff_dem_vs_other and rawdiff_gop_vs_other calculate the difference between major party votes and those cast for third-party candidates.

Finally, we derived **percentage difference metrics** to express these raw differences as proportions of total votes. For instance, pctdiff_dem_vs_gop calculates the percentage difference between Democratic and Republican votes relative to the total votes cast.

The dataset also includes variables to capture **voter turnout metrics**, such as voter_turnout, which measures total voter turnout as a proportion of the Citizen Voting Age Population (CVAP). We further refined this with turnout metrics specific to each party, including voter_turnout_dem, voter_turnout_gop, and voter_turnout_other.

To summarize the outcomes, we created two key variables: **winning_party**, which identifies the party with the majority of votes in a given county, and **winning_party_binary**, a binary version of the winning_party variable for model use. Finally, we derived **pct_margin_of_victory**, which calculates the margin by which the winning party leads its closest competitor, expressed as a percentage of total votes.

Exploratory Insights on Engineered Features and Party Trends

The analysis of the dataset's variables highlights several notable patterns concerning voter turnout and electoral competitiveness at the state level, aggregated from county-level data. The CVAP estimates (cvap_est_*) for the years 2008–2020 exhibit a consistently right-skewed distribution, indicating that most states have smaller voting-age populations, while a select few—likely those containing major urban centers or metropolitan regions—account for disproportionately high values.

This same pattern is mirrored in the total votes (totalvotes_*), reinforcing the outsized influence of more populous states in overall election outcomes (See *Figure 1*).

The voter turnout metrics (voter_turnout_*, voter_turnout_dem_*, voter_turnout_gop_*) further reinforce these trends. Turnout rates consistently cluster around a mid-range of 40% to 60%, reflecting relatively stable participation across states over time. However, differences between Democratic turnout (voter_turnout_dem_*) and Republican turnout (voter_turnout_gop_*) are observable, with Democratic turnout exhibiting slightly broader variability. These deviations may indicate regional differences in political mobilization or varying levels of competitiveness.

It is important to note that these voter turnout metrics reflect only Democratic and Republican participation, as third-party data was excluded earlier in the preprocessing stage. This decision aligns with historical electoral trends, where third-party influence has been negligible since 2008.

The raw difference metrics (rawdiff_dem_vs_gop_*) and percentage difference metrics (pctdiff_dem_vs_gop_*) confirm the entrenched two-party system in U.S. elections. The majority of states are classified as either Republican- or Democratic-dominant, with minimal instances of third-party majorities. This outcome aligns with both historical patterns and the assumptions made during the preprocessing stage.

Overall, these observations highlight a polarized, yet stable electoral system dominated by major population centers. Future exploration should prioritize states where deviations from these patterns occur, as such anomalies may provide valuable insights into unique regional political dynamics, voter behavior, or emerging electoral trends that challenge the dominant two-party system.

State-Level Aggregation and Electoral College Dynamics

Although the Electoral College introduces additional complexity to the election process, *Figure 2* illustrates that, historically, the party with the majority of states voting in its favor often aligns with the

overall election winner. This pattern underscores the importance of understanding state-level behavior when modeling election outcomes.

As a reminder, our analysis initially began with county-level data, which was then aggregated to the state level as part of our methodology. This step allowed us to align our dataset more closely with the structure of the Electoral College, where state-level outcomes are the deciding factor. However, it is important to note that county-level voter turnout and voting habits may not always align with patterns observed at the state level, potentially creating deviations from the expected trends.

Moreover, while this analysis focuses on state-level voting patterns, it does not account for the differences between the popular vote and the Electoral College. These two structures operate independently, and the allocation of electoral votes—based on census results—can change from one election to another, further complicating predictions.

Evaluation Feature Correlation for Model Integrity

Upon reviewing the correlation plot (see *Figure 3*), I aimed to explore the relationships between voter counts and the feature variables, while also monitoring for any potential signs of overfitting, particularly from variables with near-perfect correlations. While high correlations could indicate redundancy, the strong relationships between engineered features such as raw difference and percentage difference suggest they are mathematically aligned and beneficial for the Random Forest model by providing interpretable information about voting margins.

The expected negative relationship between Democratic turnout and Republican turnout reflects the competitive, zero-sum nature of two-party elections. Similarly, the strong correlation between CVAP estimates and total votes confirms that larger eligible populations tend to produce higher vote totals. While raw differences and percentage differences do not inherently identify races as "50/50," they provide valuable insight into vote margins, helping to measure how competitive an election is.

Quantifying Multicollinearity and Refining Model Features

Variance Inflation Factor (VIF) is a statistical measure used to quantify the extent of multicollinearity among predictor variables. VIF values of 1 indicate no correlation between a predictor and other variables, values between 1 and 5 suggest moderate correlation, and scores exceeding 5 (or sometimes 10) indicate high multicollinearity. As shown in Figure 3, the VIF values from our exploratory analysis reveal significant multicollinearity within the data. While high multicollinearity can pose challenges for models like linear regression, it is typically less problematic for tree-based methods such as Random Forest.

Given this, we identified multicollinearity (see Table 4) but confirmed that our chosen modeling approach remains appropriate. Prior to building the model, we will exclude non-predictive columns such as 'FIPS', 'county', and 'state'. These columns serve as identifiers or categorical labels rather than numerical predictors. Including such variables without proper encoding can unnecessarily increase dimensionality, especially when generating dummy variables, which can complicate the analysis without adding predictive value.

Implementation of Random Forest Model (Base Model)

Data Preparation and Splitting

We prepared the dataset by excluding any string or character-based variables from prior years, as these were unsuitable for predictive modeling. This included columns like 'winning_party_2008' and 'winning_party_2012', which were removed using the select(-c()) function. The 'winning' variables retained in the dataset were binary (0 and 1), derived during feature engineering, and converted into factors to ensure proper handling as categorical data. Additionally, the state variable was factored to accommodate its categorical nature.

The dataset was split into a training set (70%) and a testing set (30%) using random sampling. The train_indices variable, a list of row indices (e.g., [1] 31 15 14 3 42 43 ...), was generated to designate

rows for the training set, while the remaining rows were assigned to the testing set. A random seed (set.seed(123)) ensured consistent splits for reproducibility.

The Random Forest model was trained on the training set to predict winning_party_binary_2020. It was configured with ntree = 500, specifying the number of decision trees, and mtry = 5, controlling the number of variables considered at each split. These parameters balance the model's robustness and computational efficiency.

Evaluation of Model Performance

The Random Forest model achieved an Out-of-Bag (OOB) error rate of 2.86%, demonstrating strong predictive performance during training. The confusion matrix for the training dataset (Figure 4) reveals that the model successfully classified 16 samples as class 0 and 18 samples as class 1. A single misclassification occurred for class 0, resulting in a class error rate of 5.88%, while the error rate for class 1 was 0.00%. These results indicate that the model effectively captured the structure of the training data, learning the underlying patterns with minimal overfitting. The low OOB error rate further reinforces this conclusion, showcasing the model’s capacity to generalize within the training phase.

Testing on the hold-out dataset provided further insights into the model’s generalizability. The confusion matrix for the test data (Figure 5) demonstrates that 8 samples were correctly predicted as class 0 (True Negatives), while 6 samples were correctly classified as class 1 (True Positives). One sample was misclassified as class 1 instead of class 0 (False Positive), and no samples were misclassified as class 0 instead of class 1 (False Negatives). These results translate to an overall accuracy of 93.33%. Sensitivity, which measures the model's ability to correctly identify instances of class 0, was 88.89%, while specificity, the measure of correctly identifying instances of class 1, was 100.00%. The balanced accuracy, an average of sensitivity and specificity, was 94.44%. These metrics highlight the model’s effectiveness in accurately classifying both positive and negative cases within the test data.

The test performance metrics were further supported by additional statistical measures. The confidence interval for accuracy, at a 95% confidence level, ranged from 68.05% to 99.83%, reflecting the robustness of the model's predictions even with a relatively small test dataset. The Kappa statistic of 0.8649 indicates strong agreement between predicted and actual classifications, underscoring the model's reliability. These results confirm that the model generalizes well to unseen data and effectively maintains a balance between sensitivity and specificity.

Visual representations of the confusion matrices provide further clarity. Figure 4 illustrates the performance on the training dataset, emphasizing the model's accuracy in capturing the data's structure while maintaining low error rates. In contrast, Figure 5 focuses on the test dataset and highlights the model's predictive accuracy in real-world scenarios. The visual layout of the confusion matrices enables clear identification of true positives, true negatives, false positives, and false negatives, making it easier to interpret the model's strengths and weaknesses. These visuals serve as critical tools for communicating the results and validating the model's performance in an election forecasting context.

The Random Forest model demonstrates strong performance in both the training and testing phases, as evidenced by its low OOB error rate during training and the high accuracy, sensitivity, and specificity observed in testing. These metrics highlight the model's robustness and reliability in predicting election outcomes. The visualizations in Figures 4 and 5 provide further clarity by illustrating the distribution of true positives, true negatives, and misclassifications, offering valuable insights into the model's predictive strengths and areas for potential refinement. These results serve as a solid foundation for further analysis and exploration in subsequent sections of this report.

Overfit Check and Hyperparameter Tuning

To ensure the Random Forest model avoided overfitting, a 10-fold cross-validation methodology was applied to tune the mtry parameter. The tuning process evaluated mtry values of 2, 41, and 80, corresponding to increasing numbers of predictors considered at each split. The results, summarized in

Table 9, show that the model achieves its highest accuracy (97.50%) and Kappa statistic (0.95) when mtry = 41.

The selection of mtry = 41 balances model complexity and predictive performance, utilizing a significant proportion of predictors for splitting. Given the relatively small dataset, this choice ensures the model can leverage available features effectively while maintaining robustness. Additionally, the cross-validation methodology supports the conclusion that the model is well-generalized and unlikely to overfit.

Feature Importance

Mean Decrease Accuracy

Mean Decrease Accuracy (MDA) serves as a critical metric in Random Forest models for quantifying the importance of predictor variables. Specifically, it measures the reduction in predictive accuracy when the values of a given variable are permuted at random. Larger decreases in accuracy signify higher importance, whereas smaller decreases indicate limited predictive contribution.

These results are expected because these variables primarily represent raw counts (e.g., population estimates and voter turnout) or categorical identifiers, such as state names. They lack the deeper relational insights provided by engineered variables like pctdiff_dem_vs_gop, which captures percentage differences between party outcomes, or voter_turnout_dem, which focuses specifically on Democratic turnout. These derived metrics, designed to represent interactions between political parties, ranked higher in MDA because they provide more meaningful information for predicting election outcomes.

As shown in Figure 6, this distinction highlights the importance of feature engineering. Transforming basic data into relationship-focused metrics not only improves model performance but also enhances interpretability. MDA effectively illustrates how targeted features drive accuracy, underscoring their importance in this predictive framework.

Mean Decrease GNI

Mean Decrease Gini (MDG) serves as a critical measure of variable importance within Random Forest models. It evaluates the contribution of individual features to the model by assessing their role in improving the purity of decision tree splits. Variables with higher MDG values are more influential, as they significantly enhance the model's capacity to differentiate between classes. This metric is instrumental for feature ranking and selection, enabling the identification of variables that most substantially impact the model's predictions.

In this analysis, the variables with the highest MDG values are state, rawdiff_dem_vs_gop_2020, pctdiff_dem_vs_gop_2020, rawdiff_dem_vs_gop_2016, and pctdiff_dem_vs_gop_2016 (see *Figure 7*). These attributes reflect critical relationships, such as differences in vote counts or percentages between parties, which align closely with the model's goal of predicting election outcomes. As shown in Figure 6, these variables consistently rank higher than baseline variables, such as total voter turnout or population estimates, reinforcing their importance in identifying nuanced patterns in electoral data.

Adding Demographic Data to Enhance Model Granularity

Data Retrieval and Cleanup

To enrich the dataset with demographic variables, data was sourced from the American Community Survey (ACS) through the U.S. Census Bureau API. Data was retrieved for the years 2008, 2012, 2016, and 2020, corresponding to the election years under study. For the 2008 data, the 2006–2008 ACS 3-Year Estimates were used because 5-year estimates were unavailable for that period. Notably, the 3-year estimates were discontinued after 2009, making this dataset the most suitable option for this analysis.

The attributes retrieved included variables such as educational attainment (e.g., "Bachelor's degree," "Some college, no degree"), age groupings (e.g., "18 to 24 years," "65 years and over"), and gender ("Male," "Female"). The data was row-binded across years, sorted, and cleaned to remove any

missing values (NAs). Additionally, non-essential entries such as Puerto Rico were excluded to ensure alignment with the existing dataset of 49 states (Alaska was already excluded). The state variable was also standardized to facilitate seamless merging with the primary dataset.

Integrating Demographic and Existing Data

The cleaned demographic data was joined with the primary dataset using the state variable as the key, adding a new level of granularity to the Random Forest model. This integration included variables such as age distribution, gender proportions, and education levels for each state, enabling a deeper analysis of demographic factors and their influence on electoral outcomes.

After the merge, the dataset contained both raw demographic attributes and the previously engineered electoral variables, creating a unified structure for modeling. Table 10 provides a data dictionary summarizing the demographic variables and their integration with the existing electoral data.

Implementation of Random Forest Model (Final Model)

Data Preparation and Splitting

In this phase, we followed a similar data preparation and splitting procedure as implemented for the base model. The key steps included the exclusion of non-predictive string or character-based variables, such as historical 'winning_party' columns, and the conversion of categorical variables, including the state and binary 'winning' columns, into appropriate formats for modeling. The demographic variables, introduced in this iteration, were incorporated into the dataset to enhance granularity and predictive capacity.

To ensure consistency and reproducibility, the dataset was split into training (70%) and testing (30%) sets using random sampling with a fixed seed (set.seed(123)). This approach aligns with the methodology established in the base model while accounting for the additional features introduced in the final dataset.

Evaluation of Final Model Performance

The performance of the final Random Forest model, which incorporated demographic variables, was nearly identical to the base model that used only engineered features. For the training data, the final model achieved an Out-of-Bag (OOB) error rate of 5.88%, compared to 2.86% for the base model. The slight increase in the error rate is likely attributable to random variations introduced during training, as only 70% of the data was used for this phase.

The confusion matrix for the final model's training data (see Figure 8) demonstrates that 15 instances of class 0 were correctly classified, with one misclassification, resulting in a class error rate of 6.25% for class 0. For class 1, 17 instances were correctly classified, with one misclassification, leading to a class error rate of 5.56%. This represents a marginal decrease in performance compared to the base model, which achieved class error rates of 5.88% and 0% for class 0 and class 1, respectively.

When evaluated on the test data, the prediction accuracy of the final model remained unchanged from the base model, achieving an overall accuracy of 93.33% (see Figure 9). The sensitivity (88.89%) and specificity (100%) were also identical, underscoring the consistency of the models in correctly predicting both positive and negative classes. The Kappa statistic of 0.8649 indicates strong agreement between predicted and actual classifications, further supporting the reliability of the predictions.

Although the addition of demographic data did not lead to significant changes in overall model performance, it provided an opportunity to assess feature importance within a broader context. This subsequent analysis revealed several noteworthy patterns, warranting further discussion in the following sections.

Feature Importance

The evaluation of feature importance for the final Random Forest model highlighted the "state" variable as the most influential predictor across both Mean Decrease Gini (see Figure 11) and Mean Decrease Accuracy (see Figure 10). This result emphasizes the critical role of state-level factors in

predicting electoral outcomes, consistent with the Electoral College's state-based structure. The "state" variable likely serves as a proxy for complex demographic, political, and historical patterns. Engineered features such as pctdiff_dem_vs_gop_2020 and rawdiff_dem_vs_gop_2020 also ranked highly, underscoring their ability to capture inter-party dynamics and voter alignment effectively.

Demographic variables, including bachelor's degree attainment and age groups, were ranked lower in importance, reflecting their more indirect influence on voter behavior. This observation supports the hypothesis that the polarized nature of the two-party system prioritizes features tied directly to partisan competitiveness. Although the addition of demographic data did not substantially alter model accuracy, their inclusion provides contextual insights that enhance interpretability, particularly when analyzing unique patterns in swing states or other regions.

Final Model Evaluation and Future Directions

Data Integration and Transformation

The live election results were sourced from Reuters' interactive website (Reuters, 2024), which categorizes states by their partisan alignment and level of competitiveness (see Figure 12). These results were merged with the predictions generated by the final Random Forest model to assess its accuracy. During this process, the model's output was transformed back into a human-readable format by removing engineered features and converting the predicted binary labels into their respective parties: "Democratic Party" and "Republican Party." This merged dataset is presented in Table 11, which summarizes the final model's predictions alongside the actual 2024 election results.

The transformation ensured that the dataset was directly comparable to the actual results, facilitating a comprehensive evaluation of the model's performance. The final dataset highlighted overall trends in prediction accuracy while allowing for a deeper dive into the most challenging states to predict.

Comparison of Predictions vs. Actual Results

The overall confusion matrix (see Figure 13) reveals that while the model successfully identified the winning party for the majority of states, there were key discrepancies in its predictions. Specifically, it misclassified 19 Democratic states as Republican, resulting in a skewed representation of the Democratic vote. Despite this, the model correctly predicted the outcomes for solidly partisan states, such as California and Alabama, aligning with the state-level polarization captured during feature engineering.

However, as shown in Table 12, the model's performance in swing states was notably poor. For critical states like Georgia, Pennsylvania, and Wisconsin, the model incorrectly predicted Republican victories, while the actual results favored the Democratic Party. Swing states are inherently volatile and difficult to predict due to their competitiveness and variability. Moreover, given their importance in U.S. elections, these states represent a critical benchmark for assessing the model's ability to capture nuanced dynamics. While this emphasis may reflect subjective prioritization, it highlights the need for targeted enhancements to improve predictive accuracy in these contexts.

Additionally, it is important to note that this analysis does not incorporate the Electoral College system, which ultimately determines U.S. presidential election outcomes. The predictions are based solely on state-level majority wins, reflecting the popular vote within each state. While this approach simplifies the analysis and allows for easier model evaluation, it limits the broader applicability of the results in accurately predicting real-world election dynamics. Incorporating Electoral College considerations in future models could better align predictions with actual election outcomes.

Future Work and Model Improvements

Future research should address several limitations identified in this study. First, a method for incorporating Alaska, which was excluded early in the analysis due to its unique data structure, should be developed to ensure the model accounts for all states. Second, the influence of polarized states,

where party alignment remains deeply entrenched, requires closer examination. These states may unfairly skew the model's predictions, emphasizing the need for a balanced approach.

Additionally, future iterations should assess the role of metropolitan areas, as their demographic and political influences likely play a significant role in determining state outcomes. Given the importance of these urban centers, incorporating granular data at the city or county level could enhance the model's predictive capacity. Finally, while Random Forest proved to be an appropriate modeling technique, further optimization of its parameters, such as the number of trees (ntree) and the number of variables considered at each split (mtry), may yield better results. These refinements, combined with a more nuanced feature set, could improve the model's accuracy, particularly in swing states and competitive regions.

References

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

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

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

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

Federal Election Commission. (n.d.). Official 2008–2020 Presidential Election Results. Retrieved from https://www.fec.gov

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

Legislative Finance Division. (2021). Local government in Alaska. *State of Alaska, Legislative Finance Division.* https://www.legfin.akleg.gov/InformationalPapers/21-028m-Local-Government-In-Alaska.pdf

Modeling Uncertainty: Exploring the Limits and Constraints of Predictive Models for Presidential Elections. [n.d.]

Nargesian, Fatemeh, et al. (2017). Learning feature engineering for classification. *University of Toronto, IBM Research, Georgia Institute of Technology.* https://www.ijcai.org/proceedings/2017/0352.pdf

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

Quantifying Predictive Ambiguity: Navigating Uncertainty, Constraints, and Theoretical Boundaries in Presidential Election Forecasting. [n.d.]

Reuters. (2024). 2024 U.S. election results. Retrieved from https://www.reuters.com/graphics/USA-ELECTION/RESULTS/zjpqnemxwvx/

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

U.S. Census Bureau. (2008). American Community Survey: 3-Year Estimates, Table B15001. Retrieved from https://api.census.gov/data/2008/acs/acs3/groups/B15001.html

U.S. Census Bureau. (2012). American Community Survey: 5-Year Estimates, Table B15001. Retrieved from https://api.census.gov/data/2012/acs/acs5/groups/B15001.html

U.S. Census Bureau. (2016). American Community Survey: 5-Year Estimates, Table B15001. Retrieved from https://api.census.gov/data/2016/acs/acs5/groups/B15001.html

U.S. Census Bureau. (2020). American Community Survey: 5-Year Estimates, Table B15001. Retrieved from https://api.census.gov/data/2020/acs/acs5/groups/B15001.html

U.S. Census Bureau. (n.d.). Guidance on estimates. Retrieved from https://www.census.gov/programs-surveys/acs/guidance/estimates.html

University of Missouri Science and Technology. (n.d.). Retrieved from https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=9187&context=masters_theses

Unicede AIR Worldwide. (n.d.). FIPS codes for District of Columbia. Retrieved from https://unicede.air-worldwide.com/unicede/unicede_district-columbia_fips.html#:~:text=FIPS%20codes%20for%20District%20of%20Columbia.

randomForest. (n.d.). *R documentation*. Retrieved from

https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.2/topics/randomForest

Tables

Table 1

Majority Winners by County

party	***count***
DEMOCRAT	*20906*
GREEN	*6035*
LIBERTARIAN	*4955*
OTHER	*19815*
REPUBLICAN	*20906*

Table 2

Unexpected Records

Unexpected Records		
state_po	county_name	county_fips
CT	STATEWIDE WRITEIN	NA
ME	MAINE UOCAVA	NA
RI	FEDERAL PRECINCT	NA
DC	DISTRICT OF COLUMBIA	NA

Table 3

Sample of NAs in Merged Dataset

Sample of NAs in Merged Dataset								
year	**FIPS**	**county_name**	**state**	**totalvotes**	**votes_dem**	**votes_gop**	**geoname**	**cvap_est**
2000	01001	AUTAUGA	ALABAMA	17208	4942	11993	NA	NA
2004	01001	AUTAUGA	ALABAMA	20081	4758	15196	NA	NA
2008	01001	AUTAUGA	ALABAMA	23641	6093	17403	Autauga County, Alabama	38010
2012	01001	AUTAUGA	ALABAMA	23932	6363	17379	Autauga County, Alabama	40545
2016	01001	AUTAUGA	ALABAMA	24973	5936	18172	Autauga County, Alabama	41305
2020	01001	AUTAUGA	ALABAMA	27770	7503	19838	Autauga County, Alabama	43905
2000	01003	BALDWIN	ALABAMA	56480	13997	40872	NA	NA
2004	01003	BALDWIN	ALABAMA	69320	15599	52971	NA	NA
2008	01003	BALDWIN	ALABAMA	81413	19386	61271	Baldwin County, Alabama	130865
2012	01003	BALDWIN	ALABAMA	85338	18424	66016	Baldwin County, Alabama	144120

Table 4

NA's in CVAP Estimates

NA's in CVAP Estimates	
year	**count**
2000	3154
2004	3155
2008	39
2012	40
2016	40
2020	39

Table 5

Duplicate FIPS Code Entries for Selected Counties

Duplicate FIPS Code Entries for Selected Counties								
year	**state**	**FIPS**	**county_name**	**totalvotes**	**votes_dem**	**votes_gop**	**cvap_est**	**geoname**
2008	MISSOURI	29095, 36000	JACKSON, KANSAS CITY	339266	210824	124687	481045	Jackson County, Missouri
2008	VIRGINIA	51019, 51515	BEDFORD	38564	12225	25917	56350	Bedford County, Virginia
2012	MISSOURI	29095, 36000	JACKSON, KANSAS CITY	311566	183953	122708	493440	Jackson County, Missouri
2012	VIRGINIA	51019, 51515	BEDFORD	40230	11434	28206	58850	Bedford County, Virginia
2016	MISSOURI	29095, 36000	JACKSON, KANSAS CITY	301876	168972	116211	506340	Jackson County, Missouri
2016	VIRGINIA	51019, 51515	BEDFORD	42525	9768	30659	61205	Bedford County, Virginia
2020	MISSOURI	29095, 36000	JACKSON, KANSAS CITY	333063	199842	126535	523040	Jackson County, Missouri
2020	VIRGINIA	51019	BEDFORD	48669	12176	35600	62435	Bedford County, Virginia

Table 6

Aggregate Totals of Democratic and Republican Votes

Aggregate Totals of Democratic and Republican Votes			
year	**total_dem**	**total_gop**	**result**
2008	69,324,684	59,734,854	Democratic Party
2012	65,628,040	60,500,800	Democratic Party
2016	65,724,133	62,814,943	Democratic Party
2020	81,109,594	74,028,963	Democratic Party

Table 7

Sample Aggregate Totals of Democratic and Republican Votes by State

Sampe Aggregate Totals of Democratic and Republican Votes by State					
state	*year*	*totalvotes*	*votes_dem*	*votes_gop*	*cvap_est*
ALABAMA	*2008*	*2099819*	*813479*	*1266546*	*3481380*
ALABAMA	*2012*	*2070353*	*795696*	*1255925*	*3600120*
ALABAMA	*2016*	*2123367*	*729547*	*1318250*	*3671115*
ALABAMA	*2020*	*2323282*	*849624*	*1441170*	*3782980*
ARIZONA	*2008*	*2293475*	*1034707*	*1230111*	*4110885*
ARIZONA	*2012*	*2299254*	*1025232*	*1233654*	*4444230*
ARIZONA	*2016*	*2604277*	*1161167*	*1252401*	*4812760*
ARIZONA	*2020*	*3385294*	*1672143*	*1661686*	*5000090*
ARKANSAS	*2008*	*1086617*	*422310*	*638017*	*2090155*
ARKANSAS	*2012*	*1069468*	*394409*	*647744*	*2152350*
ARKANSAS	*2016*	*1129896*	*380494*	*684872*	*2195865*
ARKANSAS	*2020*	*1219069*	*423932*	*760647*	*2211560*
CALIFORNIA	*2008*	*13561900*	*8274473*	*5011781*	*22329310*
CALIFORNIA	*2012*	*13038547*	*7854285*	*4839958*	*23881285*
CALIFORNIA	*2016*	*14181595*	*8753788*	*4483810*	*25232630*
CALIFORNIA	*2020*	*17500881*	*11110250*	*6006429*	*25916215*
COLORADO	*2008*	*2401361*	*1288576*	*1073589*	*3403825*
COLORADO	*2012*	*2569217*	*1322998*	*1185050*	*3679115*
COLORADO	*2016*	*2780220*	*1338870*	*1202484*	*3979310*

Sampe Aggregate Totals of Democratic and Republican Votes by State

state	*year*	*totalvotes*	*votes_dem*	*votes_gop*	*cvap_est*
COLORADO	*2020*	*3256980*	*1804352*	*1364607*	*4194465*
CONNECTICUT	*2008*	*1647085*	*1000291*	*628041*	*2493100*
CONNECTICUT	*2012*	*1557885*	*905083*	*634892*	*2564230*
CONNECTICUT	*2016*	*1644920*	*897572*	*673215*	*2600980*
CONNECTICUT	*2020*	*1823857*	*1080831*	*714717*	*2638020*
DELAWARE	*2008*	*412412*	*255459*	*152374*	*638160*
DELAWARE	*2012*	*413937*	*242584*	*165484*	*674335*
DELAWARE	*2016*	*442997*	*235603*	*185127*	*704105*
DELAWARE	*2020*	*504010*	*296268*	*200603*	*733785*
DISTRICT OF COLUMBIA	*2008*	*265853*	*245800*	*17367*	*435875*

Table 8

Variance Inflation Factor

Variance Inflation Factor (VIF) Results	
	x
totalvotes_2008	*12668.3908*
totalvotes_2012	*12694.3444*
totalvotes_2016	*7599.7554*
cvap_est_2008	*148251.5428*
cvap_est_2012	*359757.1275*
cvap_est_2016	*134479.5925*
cvap_est_2020	*29345.9999*
voter_turnout_2008	*731.9125*
voter_turnout_2012	*989.6403*
voter_turnout_2016	*174.6884*
voter_turnout_2020	*823.5184*
voter_turnout_dem_2008	*2021.3224*
voter_turnout_dem_2012	*2140.8185*
voter_turnout_dem_2016	*1248.5868*
voter_turnout_dem_2020	*4274.2918*
voter_turnout_gop_2008	*1046.6863*
voter_turnout_gop_2012	*1622.7741*

Variance Inflation Factor (VIF) Results	
	x
voter_turnout_gop_2016	*1075.2029*
voter_turnout_gop_2020	*926.9023*
pctdiff_dem_vs_gop_2008	*1768.3352*
pctdiff_dem_vs_gop_2012	*2541.5297*
pctdiff_dem_vs_gop_2016	*3328.2442*
pctdiff_dem_vs_gop_2020	*2357.2987*
rawdiff_dem_vs_gop_2008	*379.9912*
rawdiff_dem_vs_gop_2012	*427.1657*
rawdiff_dem_vs_gop_2016	*998.3352*
rawdiff_dem_vs_gop_2020	*655.8737*

Table 9

Hyperparameter Tuning Results for Random Forest Base Model

Hyperparameter Tuning Results for Random Forest Base Model		
mtry	Accuracy (%)	Kappa
2	94.17	0.89
41	97.50	0.95
80	97.50	0.95

Table 10

Variable Descriptions for Final Model Dataset

Variable Descriptions for Final Model Dataset		
Variable_Name	*Description*	*Data_Type*
state	*State name or abbreviation.*	*Character*
totalvotes_2008	*Total votes cast in 2008.*	*Numeric*
totalvotes_2012	*Total votes cast in 2012.*	*Numeric*
totalvotes_2016	*Total votes cast in 2016.*	*Numeric*
totalvotes_2020	*Total votes cast in 2020.*	*Numeric*
cvap_est_2008	*Citizen voting age population estimate for 2008.*	*Numeric*
cvap_est_2012	*Citizen voting age population estimate for 2012.*	*Numeric*
cvap_est_2016	*Citizen voting age population estimate for 2016.*	*Numeric*
cvap_est_2020	*Citizen voting age population estimate for 2020.*	*Numeric*
voter_turnout_2008	*Voter turnout as a proportion of CVAP in 2008.*	*Numeric*
voter_turnout_2012	*Voter turnout as a proportion of CVAP in 2012.*	*Numeric*
voter_turnout_2016	*Voter turnout as a proportion of CVAP in 2016.*	*Numeric*
voter_turnout_2020	*Voter turnout as a proportion of CVAP in 2020.*	*Numeric*
voter_turnout_dem_2008	*Democratic voter turnout as a proportion of CVAP in 2008.*	*Numeric*
voter_turnout_dem_2012	*Democratic voter turnout as a proportion of CVAP in 2012.*	*Numeric*
voter_turnout_dem_2016	*Democratic voter turnout as a proportion of CVAP in 2016.*	*Numeric*
voter_turnout_dem_2020	*Democratic voter turnout as a proportion of CVAP in 2020.*	*Numeric*
voter_turnout_gop_2008	*Republican voter turnout as a proportion of CVAP in 2008.*	*Numeric*
voter_turnout_gop_2012	*Republican voter turnout as a proportion of CVAP in 2012.*	*Numeric*

Variable Descriptions for Final Model Dataset		
Variable_Name	*Description*	*Data_Type*
voter_turnout_gop_2016	*Republican voter turnout as a proportion of CVAP in 2016.*	*Numeric*
voter_turnout_gop_2020	*Republican voter turnout as a proportion of CVAP in 2020.*	*Numeric*
pctdiff_dem_vs_gop_2008	*Percentage difference between Democratic and Republican votes in 2008.*	*Numeric*
pctdiff_dem_vs_gop_2012	*Percentage difference between Democratic and Republican votes in 2012.*	*Numeric*
pctdiff_dem_vs_gop_2016	*Percentage difference between Democratic and Republican votes in 2016.*	*Numeric*
pctdiff_dem_vs_gop_2020	*Percentage difference between Democratic and Republican votes in 2020.*	*Numeric*
rawdiff_dem_vs_gop_2008	*Raw vote difference between Democratic and Republican votes in 2008.*	*Numeric*
rawdiff_dem_vs_gop_2012	*Raw vote difference between Democratic and Republican votes in 2012.*	*Numeric*
rawdiff_dem_vs_gop_2016	*Raw vote difference between Democratic and Republican votes in 2016.*	*Numeric*
rawdiff_dem_vs_gop_2020	*Raw vote difference between Democratic and Republican votes in 2020.*	*Numeric*
winning_party_2008	*Party with the majority of votes in 2008.*	*Character*
winning_party_2012	*Party with the majority of votes in 2012.*	*Character*
winning_party_2016	*Party with the majority of votes in 2016.*	*Character*

Variable Descriptions for Final Model Dataset		
Variable_Name	*Description*	*Data_Type*
winning_party_2020	*Party with the majority of votes in 2020.*	*Character*

Table 11

Final Model Predictions vs. Actual 2024 Election Results

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Alabama	0.34	0.65	Republican	Republican Party	Republican Party	TRUE
Alaska	0.41	0.55	Republican	Republican Party	NA	NA
Arizona	0.47	0.52	Competitive	Republican Party	Republican Party	TRUE
Arkansas	0.34	0.64	Republican	Republican Party	Republican Party	TRUE
California	0.58	0.38	Solid Democrat	Democratic Party	Democratic Party	TRUE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Colorado	0.54	0.43	Solid Democrat	Democratic Party	Democratic Party	TRUE
Connecticut	0.56	0.42	Solid Democrat	Democratic Party	Democratic Party	TRUE
Delaware	0.57	0.42	Solid Democrat	Democratic Party	Democratic Party	TRUE
District Of Columbia	0.90	0.06	Solid Democrat	Democratic Party	NA	NA
Florida	0.43	0.56	Lean Republican	Republican Party	Republican Party	TRUE
Georgia	0.49	0.51	Competitive	Republican Party	Democratic Party	FALSE
Hawaii	0.61	0.37	Solid Democrat	Democratic Party	Democratic Party	TRUE
Idaho	0.30	0.67	Republican	Republican Party	Republican Party	TRUE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Illinois	0.55	0.44	Solid Democrat	Democratic Party	Democratic Party	TRUE
Indiana	0.40	0.59	Republican	Republican Party	Republican Party	TRUE
Iowa	0.43	0.56	Republican	Republican Party	Republican Party	TRUE
Kansas	0.41	0.57	Republican	Republican Party	Republican Party	TRUE
Kentucky	0.34	0.65	Republican	Republican Party	Republican Party	TRUE
Louisiana	0.38	0.60	Republican	Republican Party	Republican Party	TRUE
Maine	0.52	0.45	Lean Democrat	Democratic Party	Democratic Party	TRUE
Maryland	0.63	0.34	Solid Democrat	Democratic Party	Democratic Party	TRUE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Massachusetts	0.61	0.36	Solid Democrat	Democratic Party	Democratic Party	TRUE
Michigan	0.48	0.50	Competitive	Republican Party	Democratic Party	FALSE
Minnesota	0.51	0.47	Competitive	Democratic Party	Democratic Party	TRUE
Mississippi	0.38	0.61	Republican	Republican Party	Republican Party	TRUE
Missouri	0.40	0.58	Republican	Republican Party	Republican Party	TRUE
Montana	0.38	0.58	Republican	Republican Party	Republican Party	TRUE
Nebraska	0.39	0.59	Republican	Republican Party	Republican Party	TRUE
Nevada	0.47	0.51	Competitive	Republican Party	Democratic Party	FALSE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
New Hampshire	0.51	0.48	Lean Democrat	Democratic Party	Democratic Party	TRUE
New Jersey	0.52	0.46	Solid Democrat	Democratic Party	Democratic Party	TRUE
New Mexico	0.52	0.46	Lean Democrat	Democratic Party	Democratic Party	TRUE
New York	0.56	0.44	Solid Democrat	Democratic Party	Democratic Party	TRUE
North Carolina	0.48	0.51	Competitive	Republican Party	Republican Party	TRUE
North Dakota	0.31	0.67	Republican	Republican Party	Republican Party	TRUE
Ohio	0.44	0.55	Republican	Republican Party	Republican Party	TRUE
Oklahoma	0.32	0.66	Republican	Republican Party	Republican Party	TRUE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Oregon	0.55	0.41	Solid Democrat	Democratic Party	Democratic Party	TRUE
Pennsylvania	0.49	0.50	Competitive	Republican Party	Democratic Party	FALSE
Rhode Island	0.56	0.42	Solid Democrat	Democratic Party	Democratic Party	TRUE
South Carolina	0.40	0.58	Republican	Republican Party	Republican Party	TRUE
South Dakota	0.34	0.63	Republican	Republican Party	Republican Party	TRUE
Tennessee	0.34	0.64	Republican	Republican Party	Republican Party	TRUE
Texas	0.42	0.56	Lean Republican	Republican Party	Republican Party	TRUE
Utah	0.38	0.59	Republican	Republican Party	Republican Party	TRUE

Final Model Predictions vs. Actual 2024 Election Results						
State	**Democrat**	**Republican**	**type**	**actual_2024**	**prediction_2024**	**correctly_predicted**
Vermont	0.64	0.32	Solid Democrat	Democratic Party	Democratic Party	TRUE
Virginia	0.52	0.46	Lean Democrat	Democratic Party	Democratic Party	TRUE
Washington	0.57	0.39	Solid Democrat	Democratic Party	Democratic Party	TRUE
West Virginia	0.28	0.70	Republican	Republican Party	Republican Party	TRUE
Wisconsin	0.49	0.50	Competitive	Republican Party	Democratic Party	FALSE
Wyoming	0.26	0.72	Republican	Republican Party	Republican Party	TRUE

Table 12

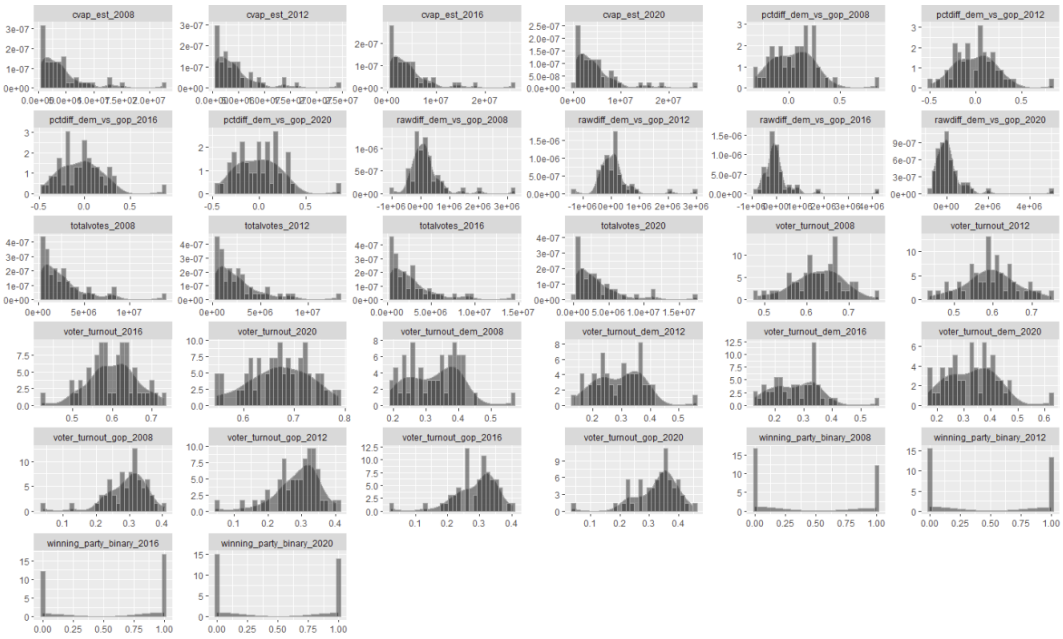
Predicted vs. Actual Outcomes for Swing States (2024)

Predicted vs. Actual Outcomes for Swing States (2024)						
State	***Democrat***	***Republican***	***type***	***actual_2024***	***prediction_2024***	***correctly_predicted***
Georgia	*0.49*	*0.51*	*Competitive*	*Republican Party*	*Democratic Party*	*FALSE*
Michigan	*0.48*	*0.50*	*Competitive*	*Republican Party*	*Democratic Party*	*FALSE*
Nevada	*0.47*	*0.51*	*Competitive*	*Republican Party*	*Democratic Party*	*FALSE*
Pennsylvania	*0.49*	*0.50*	*Competitive*	*Republican Party*	*Democratic Party*	*FALSE*
Wisconsin	*0.49*	*0.50*	*Competitive*	*Republican Party*	*Democratic Party*	*FALSE*

Figures

Figure 1.

Distribution of Voter Turnout and Related Metrics Across States (2008-2020)



Note. Data compiled from Citizen Voting Age Population (CVAP) estimates and county-level election results from 2008–2020.

Figure 2.

Histogram of Party Vote Count by State

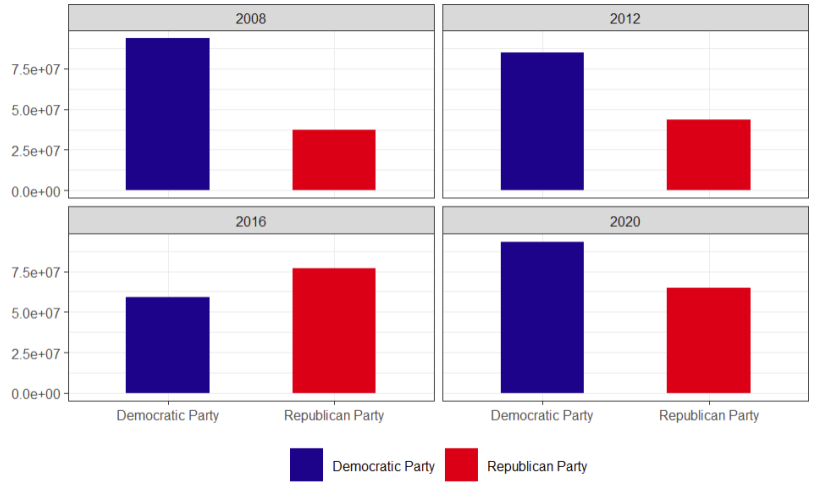


Figure 3.

Correlation Plot of Voter Count and Feature Variables

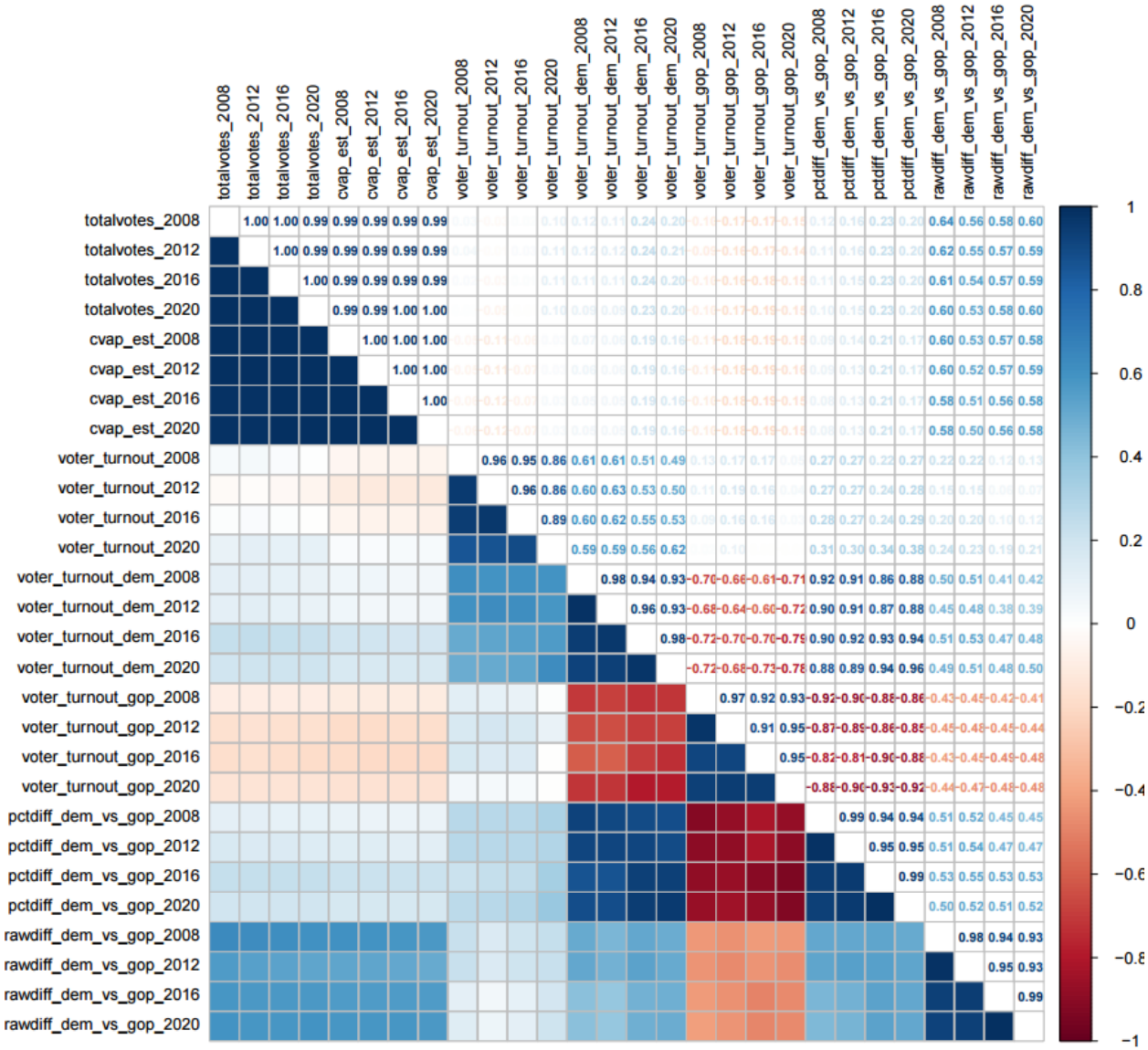


Figure 4.

Confusion matrix for training data (Base Model)

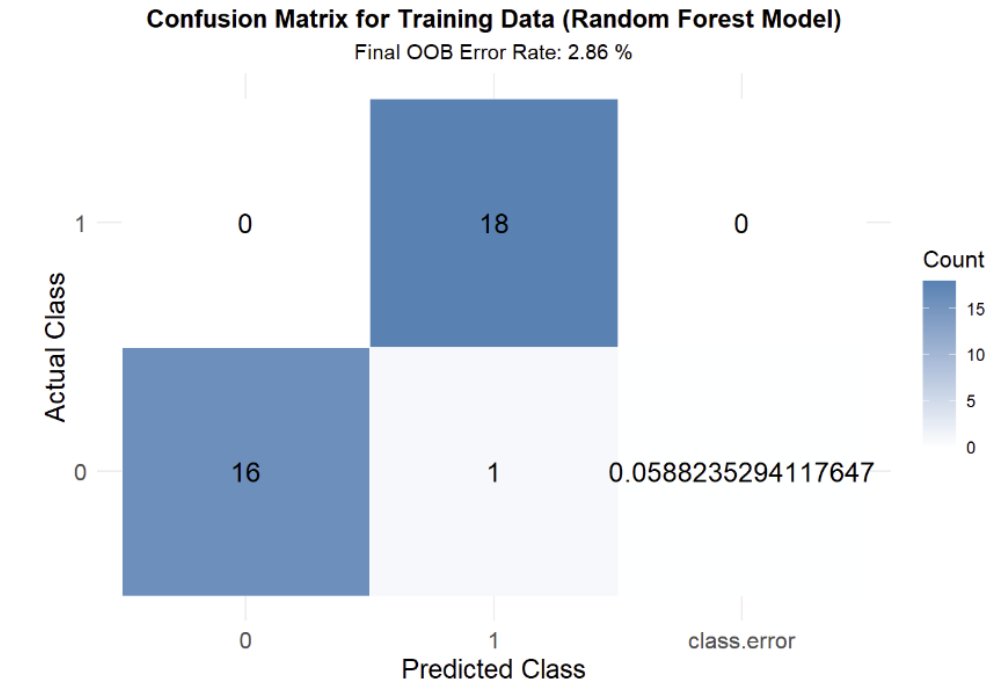


Figure 5.

Confusion matrix for test data (Base Model)

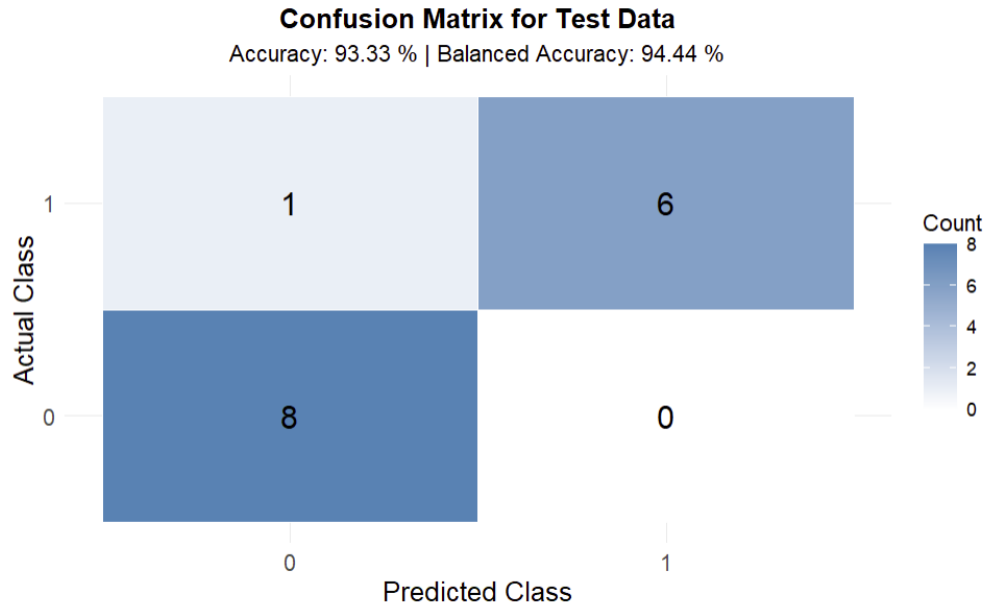


Figure 6.

Feature Importance: Mean Decrease Accuracy for Random Forest (Base Model)

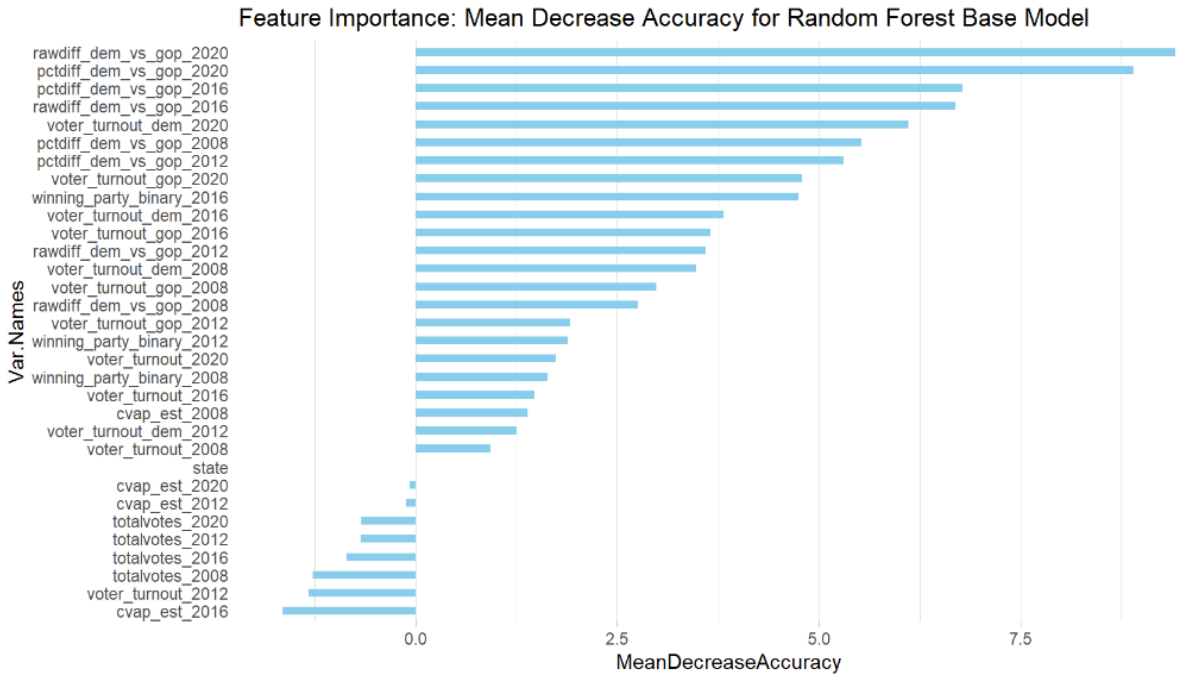
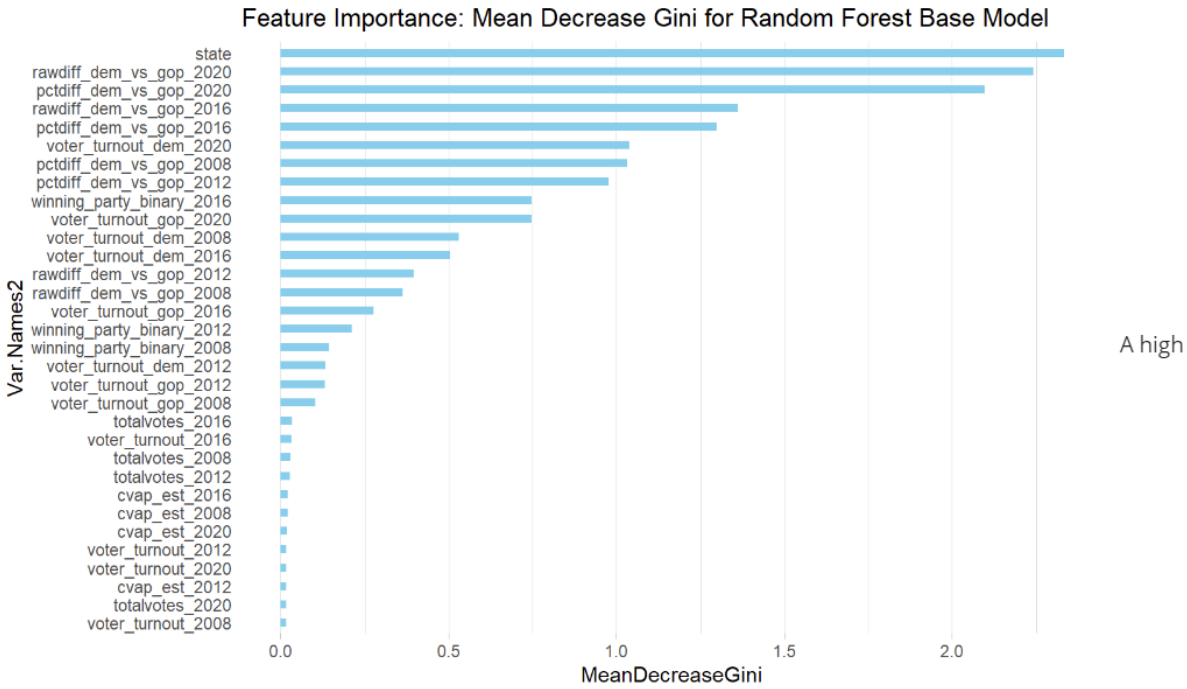


Figure 7.

Feature Importance: Mean Decrease Gini for Random Forest (Base Model)



A high

Figure 8.

Confusion Matrix for Training Data (Random Forest Final Model)

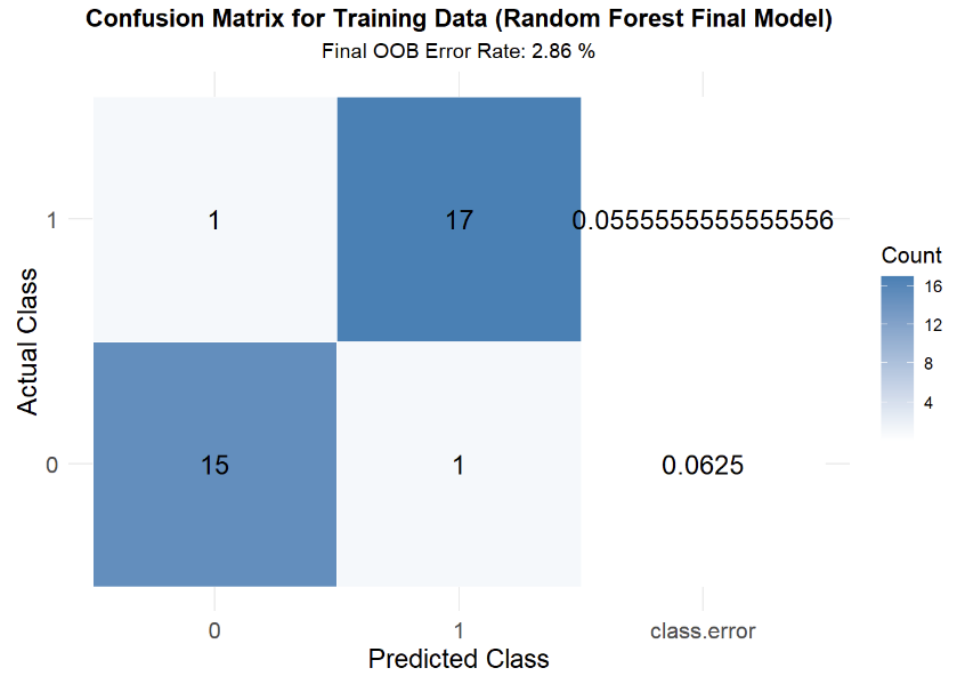


Figure 9.

Confusion Matrix for Test Data (Random Forest Final Model)

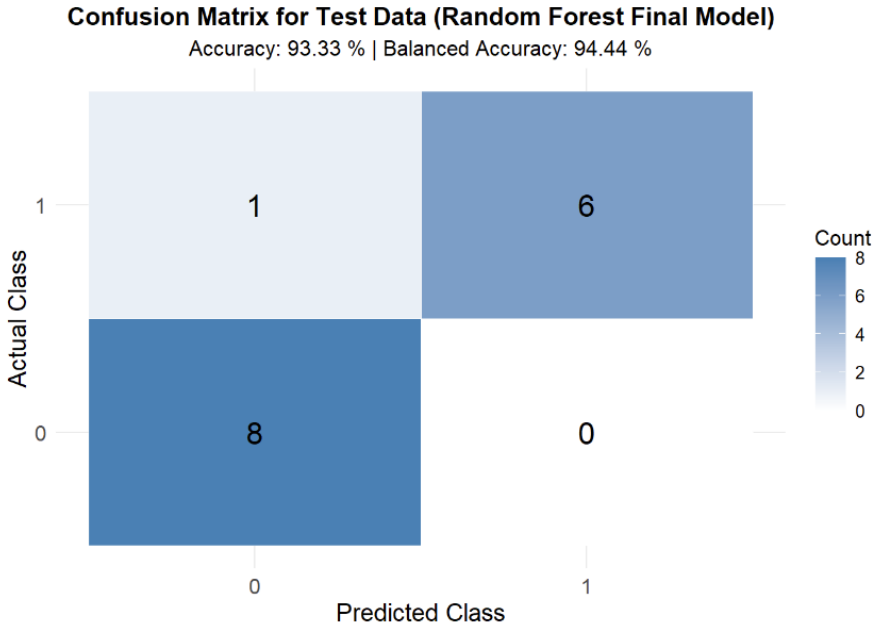


Figure 10.

Feature Importance: Mean Decrease Accuracy for Random Forest (Final Model)

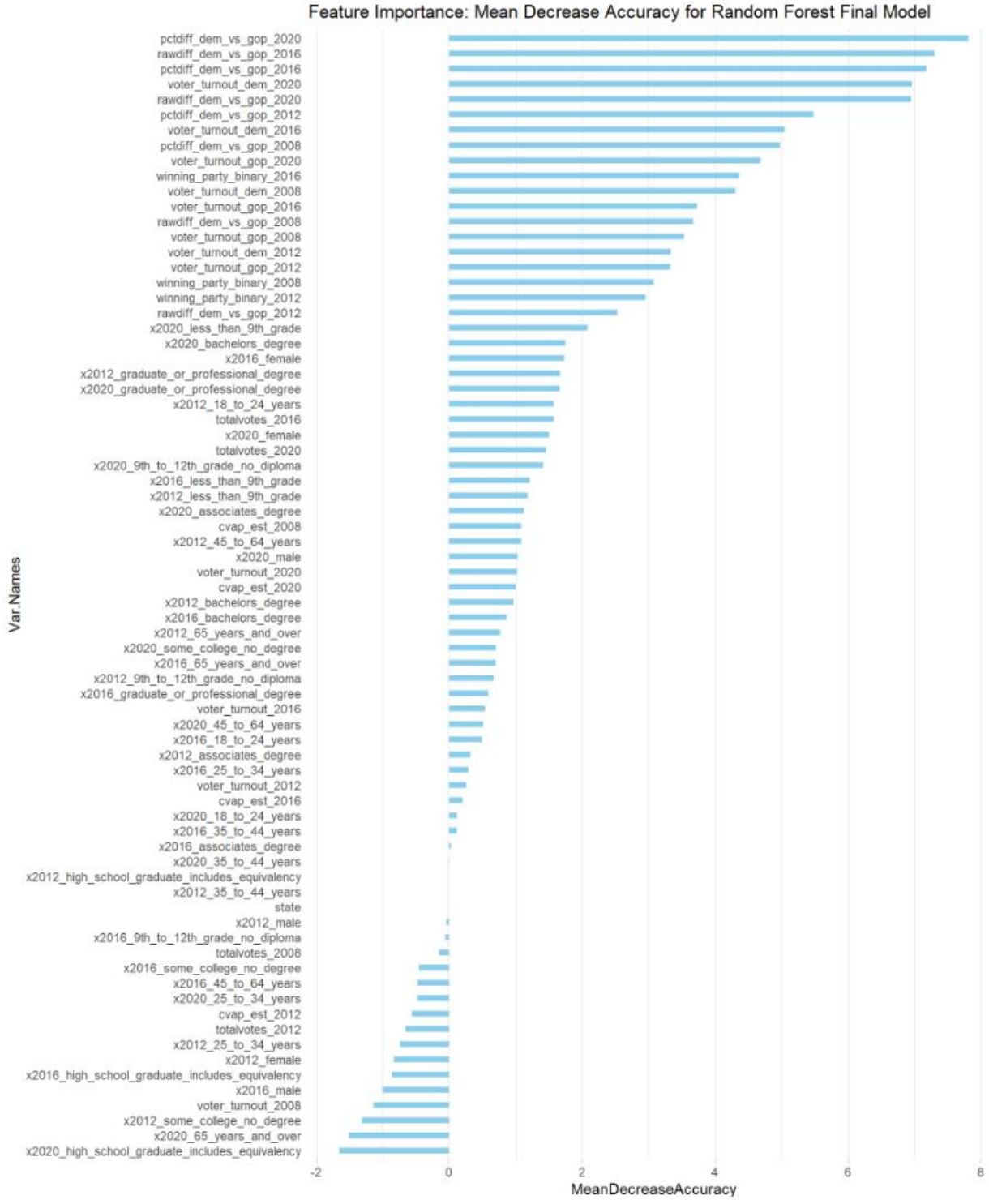


Figure 11.

Feature Importance: Mean Decrease Gini for Random Forest (Final Model)

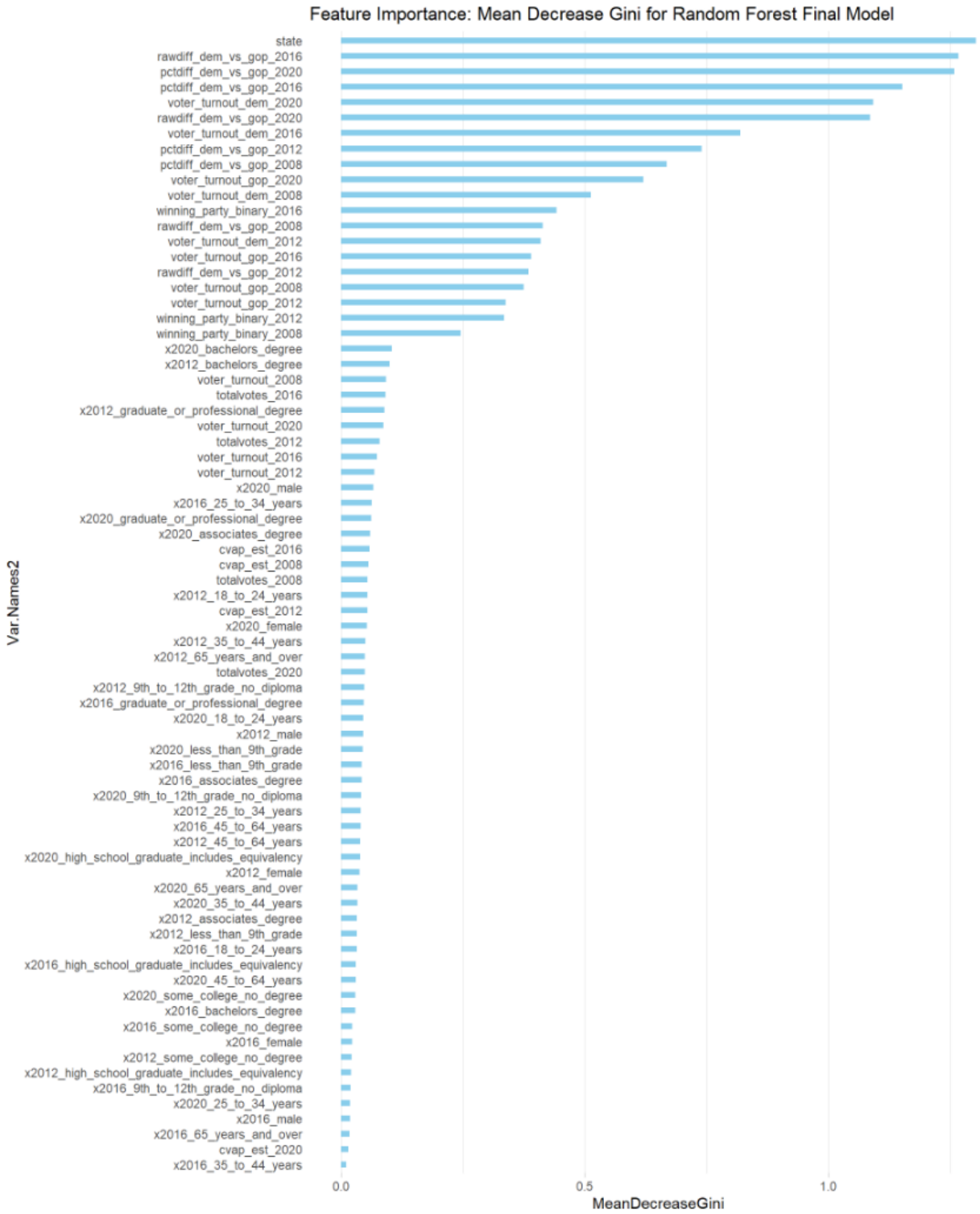
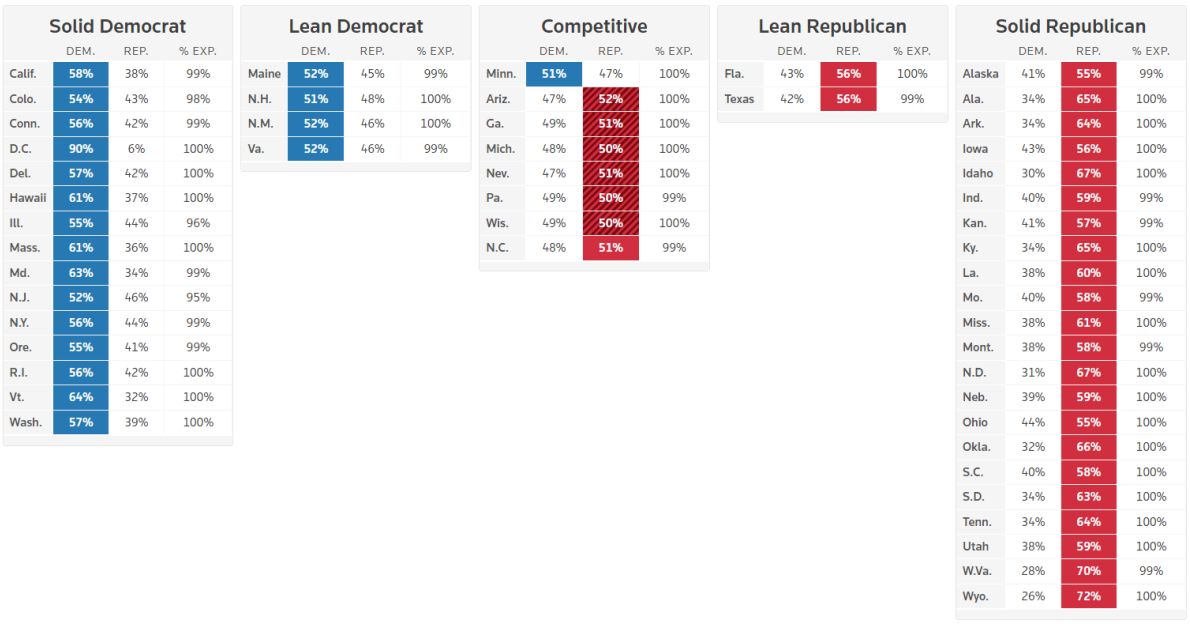


Figure 12.

Reuters Live results with State categorization



Solid Democrat

	DEM.	REP.	% EXP.
Calif.	58%	38%	99%
Colo.	54%	43%	98%
Conn.	56%	42%	99%
D.C.	90%	6%	100%
Del.	57%	42%	100%
Hawaii	61%	37%	100%
Ill.	55%	44%	96%
Mass.	61%	36%	100%
Md.	63%	34%	99%
N.J.	52%	46%	95%
N.Y.	56%	44%	99%
Ore.	55%	41%	99%
R.I.	56%	42%	100%
Vt.	64%	32%	100%
Wash.	57%	39%	100%

Lean Democrat

	DEM.	REP.	% EXP.
Maine	52%	45%	99%
N.H.	51%	48%	100%
N.M.	52%	46%	100%
Va.	52%	46%	99%

Competitive

	DEM.	REP.	% EXP.
Minn.	51%	47%	100%
Ariz.	47%	52%	100%
Ga.	49%	51%	100%
Mich.	48%	50%	100%
Nev.	47%	51%	100%
Pa.	49%	50%	99%
Wis.	49%	50%	100%
N.C.	48%	51%	99%

Lean Republican

	DEM.	REP.	% EXP.
Fla.	43%	56%	100%
Texas	42%	56%	99%

Solid Republican

	DEM.	REP.	% EXP.
Alaska	41%	55%	99%
Ala.	34%	65%	100%
Ark.	34%	64%	100%
Iowa	43%	56%	100%
Idaho	30%	67%	100%
Ind.	40%	59%	99%
Kan.	41%	57%	99%
Ky.	34%	65%	100%
La.	38%	60%	100%
Mo.	40%	58%	99%
Miss.	38%	61%	100%
Mont.	38%	58%	99%
N.D.	31%	67%	100%
Neb.	39%	59%	100%
Ohio	44%	55%	100%
Okla.	32%	66%	100%
S.C.	40%	58%	100%
S.D.	34%	63%	100%
Tenn.	34%	64%	100%
Utah	38%	59%	100%
W.Va.	28%	70%	99%
Wyo.	26%	72%	100%

Figure 13.

2024 Election results Confusion Matrix