Vaibhav Bhatla

Professor Jacob Koehler

Data Bootcamp

12 May 2025

Predicting County-Level Partisan Preference in the United States

Reshaped by the 2016 United States (US) Presidential Election, the American electorate has drifted from traditional points of bipartisan consensus. Before the emergence of left-leaning and right-leaning populist tailwinds, both sides of the aisle in Washington shared an interest in promoting free trade, fiscal discipline, and an interventionist foreign policy. For instance, Republican President George H.W. Bush had negotiated the North American Free Trade Agreement (NAFTA), a historic piece of regional trade liberalization, and Democratic President Bill Clinton had implemented it, albeit with labor and environmental provisions (Rodriguez). However, this political landscape has been dramatically challenged by the rise of Donald Trump's 'Make America Great Again' movement within the Republican Party, fracturing the Democratic 'Blue Wall' coalition (Smith). Part of this seismic shift in the partisan affiliation of American voters has been conventionally attributed by political scientists and party operatives to the Democratic Party losing support from working-class demographics due to cultural and economic issues (Weisman).

To dive deeper into this departure from an empirical standpoint, this paper aims to examine the relationship between demographic subsets and partisan voting patterns using county-level data from the 2020 election through logistic regression, random forest, and decision tree classification models that predict the partisan lean of a county. For the most part, these models predominantly confirm that demographic inputs are structural indicators of a county's

partisan preference. As expected, I find that racial, occupational, and educational variables tend to have the strongest coefficients. However, surprisingly, most states do not meaningfully predict whether a county votes Democrat or Republican. Democratic and Republican Party strategists should employ the interactive features of this model to tailor their campaign strategies for each stakeholder. They will also benefit from mapping out the probability of each county in the United States by its likelihood of party affiliation, resulting in an efficient allocation of campaign funds and resources.

Although this model is trained on historical data, these findings will hold at least in the short term, should demographic inputs across counties remain stable. Therefore, this paper will conclude with some insights and recommendations for the policymakers and strategists of both parties, especially for the 2026 Midterm Elections and the 2028 Presidential Elections.

Relevant Inquiries

1. To what extent may demographic inputs predict a county's partisan preference?
2. Which demographic features display the most potent predicting power?
3. How might trends in the feature set affect presidential and congressional elections?

Exploratory Data Analysis

Before constructing predictive models, an overview of the dataset and a survey of high-level trends are essential to explore preliminary inquiries. This dataset is sourced from Kaggle, and it contains 32 columns of demographic and electoral data across 3,143 counties throughout the United States from the 2020 Presidential Election (Kaggle). At first glance, one might assume that a more national breakdown of demographics might be more helpful in analyzing national elections. Nonetheless, the granularity in county-level data lays the foundation for researchers to uncover underlying patterns in the preferences of a voter.

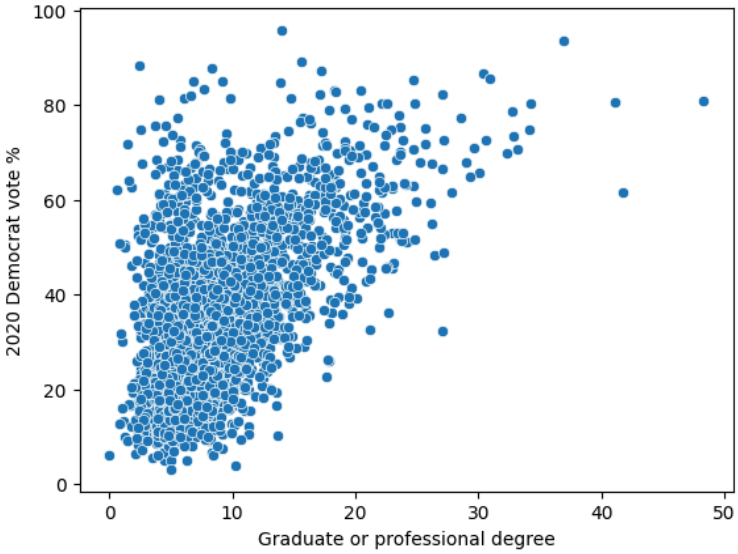
In ascertaining the relevant features within a dataset, one must look at the strength and direction of their correlation with partisan preferences. To streamline analysis, I have classified counties as Democratic if they have a 50% or greater vote share for the Democratic nominee and Republican if they have a greater than 50% vote share for the Republican nominee. Due to this binary framework with Democrats encoded as 0 and Republicans encoded as 1, features that are correlated with Democrats display a negative direction, and those with the Republicans highlight a positive direction in Figure 1.

Figure 1: Correlation of Features with Target (Partisan Preference).

Visually considering the magnitude of the features in Figure 1 implies that educational, racial, and occupational demographics are among the most pertinent for initial analysis. At the same time, variables like area in square kilometers do not tend to vary on a material level by party support and should only be briefly touched upon in the model and the final analysis.

Racial demographics may serve as a critical variable in determining the partisan preference of a county. As per Figure 1, there appears to be an incredibly potent relationship between a county's Non-Hispanic White population and the Republican Party, exhibiting a correlation of about 0.45, and an equally robust one between a county's Non-Hispanic Asian (-0.42) and Non-Hispanic Black (-0.37) populations and the Democratic Party. However, it is notable that the correlation between the percentage of other minority groups in a county and the Democratic Party is relatively smaller, which might suggest that not all racial demographics tend to indicate partisan affiliation. Despite gaining support across traditionally Democratic-leaning minority demographics in the 2024 election to some extent, the Republican Party is still considerably less diverse on an aggregate level (Frey et al.). In planning for future elections, Democratic strategists should strive to retain the support of minority demographics, and Republican strategists should strive to continue to expand their electoral base.

Shedding light on the relationship between attained education levels and partisan leans builds on the current scholarly discourse. According to the Pew Research Center, a greater share of higher-education individuals have tended to support the Democratic Party in recent years (Geiger). This educational shift may have fueled the Democratic platform's focus on social issues, ultimately failing to resonate with less formally educated individuals who were more concerned about the economy (Cohn).



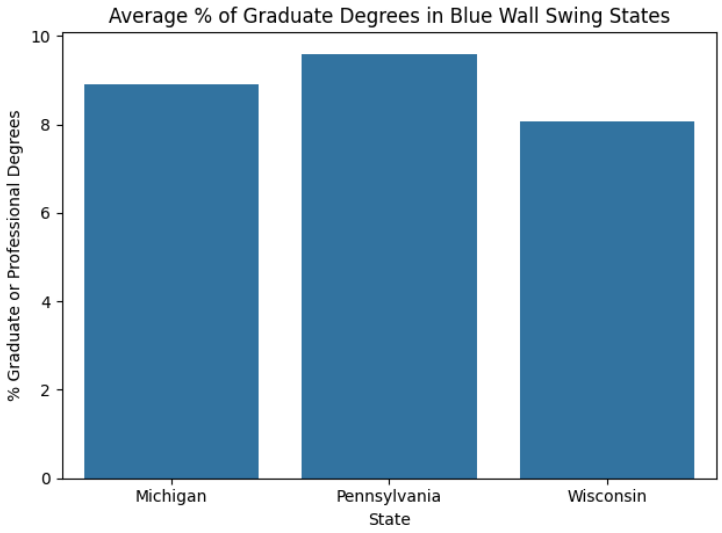
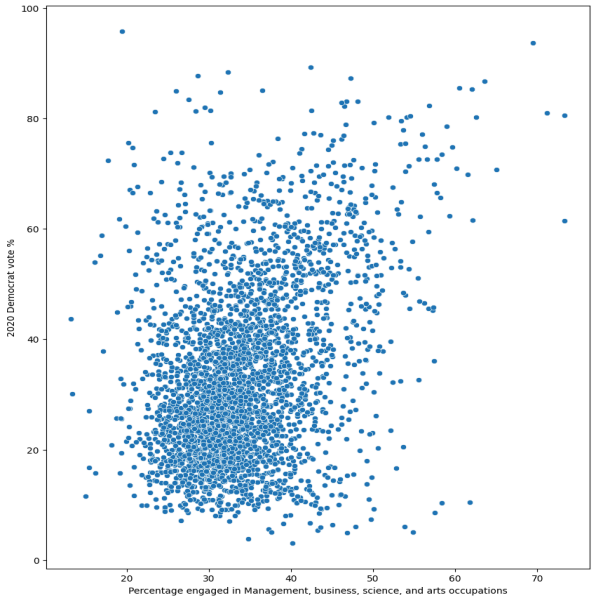


Figure 2: Graduate Degree % vs. 2020 Democrat Vote %; Figure 3: Graduate Degree % in Blue Wall Swing States

Figure 1 lends credence to this relationship, as there appears to be a slight direct relationship between the percentage of a county's population with a graduate or professional degree and the share that voted for Democrats in the 2020 Presidential Election. Narrowing the scope to solely counties in formerly Democratic-leaning swing states like Pennsylvania, Michigan, and Wisconsin, Figure 2 suggests that these regions had a relatively small percentage (8-10%) of individuals with graduate or professional degrees, which may have played a role in their support for Trump and the Republicans in 2020. Despite these promising patterns, Figure 12 in the appendix highlights the overlapping nature of the confidence intervals of graduate degree support by partisan lean, implying that this relationship may not be as significant as the scatterplot suggests. As such, in the modeling stage, education level should be assessed along with other indicators, and it cannot be the sole factor in the model.



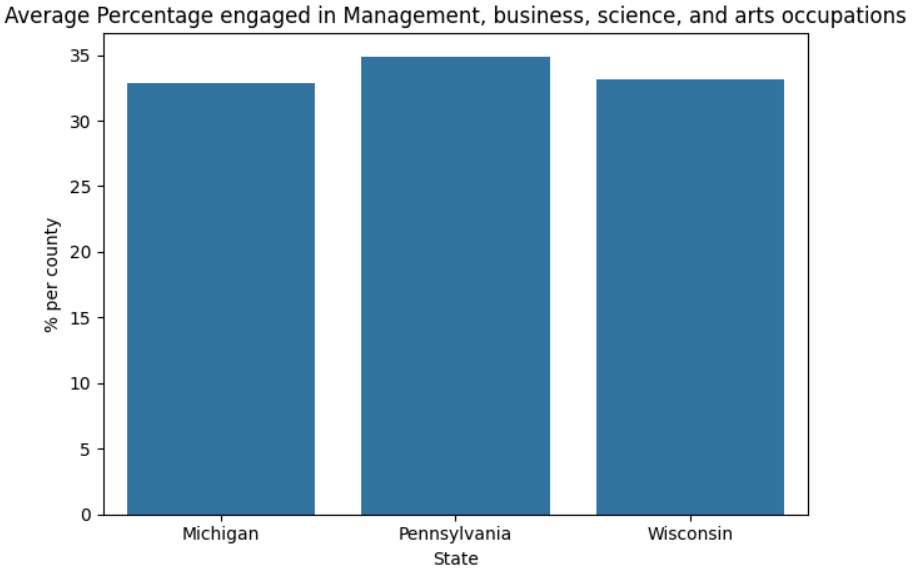


Figure 4: Management % vs. 2020 Democrat Vote %; Figure 5: Management % in Blue Wall Swing States

Implicitly diving deeper into the conversation of educational levels, it is prudent to examine the implications of occupation on partisan preferences. Figures 1 and 4 suggest that the percentage of management-centric occupations in a county may be directly linked to the vote percentage for the Democratic Party. This parallel reinforces the findings in the educational section of the exploratory data analysis, highlighting the shift of the Democratic Party's base. However, this transition has exacerbated electoral implications, as the Democrats will need to recalibrate their party platform to appeal to voters in the formerly safe Blue Wall States once again, as per Figure 5.

Conducting this exploratory data analysis on racial, educational, economic, and occupational variables highlights that Michigan, Pennsylvania, and Wisconsin may have voting trends due to their similar demographic characteristics. This analysis also provides context for the model generation stage, as it may provide insight into which features might be more or less significant in the prediction process.

Methodology

Although an exploratory data analysis is meaningful to gauge high-level trends and visually assess relationships between different variables, political party strategists need statistically significant data for their analysis. For this reason, I have crafted three models that can be employed to predict the partisan lean of a county with multiple demographic inputs: Logistic Regression, Random Forest, and Decision Tree Classification.

As mentioned above, I had cleaned and processed the data to turn the Democrat vote percentage column into a binary column, as the American electoral system has a winner-take-all system as opposed to a proportional representation one found in parliamentary systems. Furthermore, I split my demographic data into a feature set, containing all columns except county names, and a target column that denoted the political party preference by county. As my feature set had states as an input, a categorical variable, I used OneHotEncoder to prevent any difficulties in the analytical steps. I also converted certain columns from string to float to ensure that numerical variables were not unintentionally classified as categorical ones. At the end, I bundled the numerical and categorical features into a single preprocessor that I would employ for all three models.

Upon this preprocessing stage, I transitioned into model selection, considering a diverse array of models that would be the best fit for this experimental design. As the target set reflects binary outcomes, Linear Regression would not be an ideal fit at this time due to its primary strength in predicting continuous variables. Adept in binary classification, Logistic Regression allows me to determine the probability that a county leans one way or another based on a set of input variables. Political strategists may manipulate the input settings in this model to determine relevant shifts in partisan preferences. As a mechanism for fine-tuning the parameters in this

Logistic Regression Model, I utilized GridSearch to determine the model that results in the best accuracy. To supplement the findings of the Logistic Regression model, I crafted Random Forest and Decision Tree models to further verify the significance of features within my dataset. However, these models do not indicate the direction of the relationship between the features and the target, highlighting the necessity of Logistic Regression in this study. Setting an even playing field to assess the results of all three models, I have constructed two confusion matrices on the training and testing data, coupled with an accuracy score that might suggest the predictive capabilities of this model.

Results and Interpretation

The three aforementioned predictive models shed light on the behavior of certain features in the context of the entire model. Before considering the directional characteristics of the feature set as exemplified by the Logistic Regression model, it may be beneficial to consider solely its magnitude.

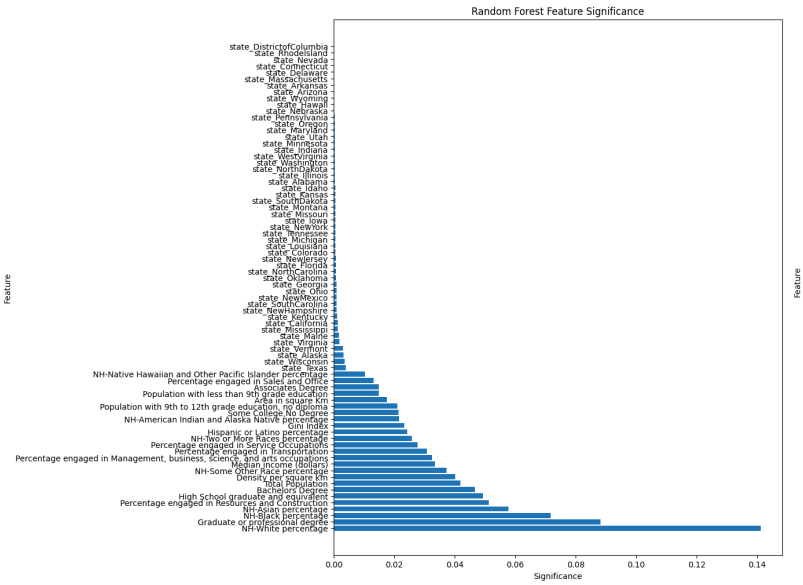


Figure 6: Random Forest Feature Significance; Figure 7: Decision Tree Feature Significance

In line with the findings in the Exploratory Data Analysis, Figures 7 and 8 suggest that racial, occupational, and educational features tend to carry the most predictive power in determining the partisan classification of a county. These visual depictions highlight that the two features with the most outsized weight in the model are the percentage of Non-Hispanic White individuals in a county and the percentage of graduate or professional degrees in a county. Although the Random Forest and Decision Tree models do not shed light on the direction that these voters sway, the Logistic Regression model and the Exploratory Data Analysis suggest that the former group leans towards the Republican Party and the latter leans towards the Democratic Party. In this light, neither party can afford to lose support in this core base.

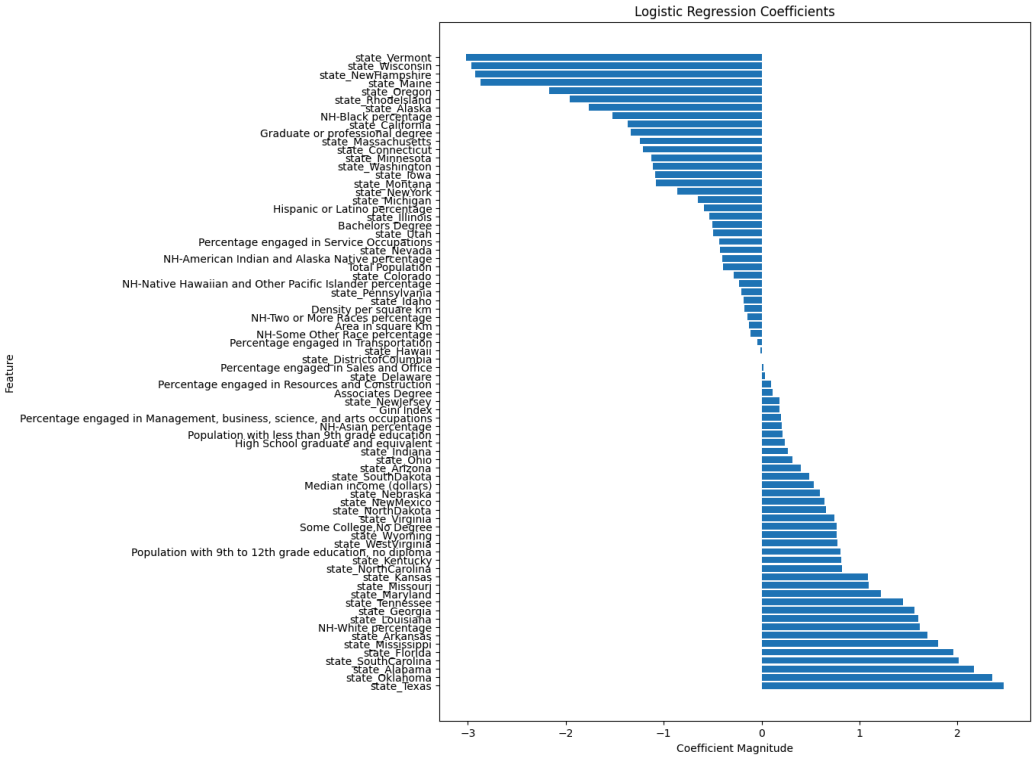


Figure 8: Logistic Regression Coefficients

Nevertheless, the relevance of geographical indicators like states, population size, and population density is under question, thanks to contrasting findings between the Random

Forest/Decision Tree models and the Logistic Regression models. Although the Random Forest/Decision Tree models suggest that the state a county lies within carries a negligible weight in the predictive model, the Logistic Regression model attaches some of the highest coefficients to it. This discrepancy may be explained by the fact that Logistic Regression models fail to consider secondary relationships between features, unlike Random Forest/Decision Tree models. In some ways, this finding may suggest that the partisan preferences of individual states are not set in stone, generating political opportunities on a bipartisan basis. Along with focusing on swing states, partisan strategists should tailor their campaign messaging to swing demographics as well.

Model	Logistic Regression	Random Forest	Decision Tree
Accuracy	88.87%	92.68%	88.23%

Figure 9: Model Accuracies

Based on these features, weighted by significance, the Logistic Regression, Random Forest, and Decision Tree models were effective in predicting partisan affiliation to some extent. From an accuracy standpoint, all three models yielded similar outcomes around 90%. Although the target set is significantly imbalanced, where 83.45% of counties nationwide are Republican and only 16.55% are Democratic, all three models have been adjusted to ensure that the class weights are balanced. Therefore, we can complement this accuracy analysis with a precision and recall one.

Model	Logistic Regression	Random Forest	Decision Tree
Democratic Precision	61.5%	84.5%	62.5%
Democratic Recall	87.5%	68.3%	72.11%

Figure 10: Democratic Party Precision and Recall

. Comparing these models, the Random Forest has the highest precision metric, which implies a high accuracy in its predictions, but it cannot adequately find all of the Democratic counties. However, Logistic Regression has the highest recall metric, which suggests that it effectively searches for most of the Democratic counties with the inputs given. At the same time, the Decision Tree model is mediocre, albeit not poor, in both precision and recall, suggesting that party strategists might be better suited to balance the findings of the former two models. Further information can be found in the confusion matrices attached below in the appendix. Therefore, I would recommend that a strategist primarily use Logistic Regression and Random Tree, leaving Decision Tree Classification as a last resort.

Interactive Prediction: Case Studies

Thanks to the relatively robust accuracy, precision, and recall metrics of these binary classification models, party strategists might be able to employ them in their decision-making calculus for future electoral races. Let us consider the predictions of the three models in an interactive setting with New York County, New York, a Democratic bastion with 86% of the votes for the Democratic candidate, Bucks County, Pennsylvania, a significant swing county with 51.66% votes for the Democratic candidate, and Baker County, Florida, a Republican stronghold with merely 14.49% of the votes for the Democratic candidate as case studies. The demographic details of these counties are attached in the CSV dataset from Kaggle.

Region	Logistic Regression	Random Forest	Decision Tree
New York County, NY	Democratic (100%)	Republican (97%)	Democratic (99%)
Baker County, FL	Republican (100%)	Republican (100%)	Republican (100%)
Bucks County, PA	Democratic (70%)	Democratic (84%)	Democratic (94%)

Figure 11: Predictions (Solid Democrat, Solid Republican, Swing Cases)

As per Figure 11, all three models had reasonably predicted the partisan swings of the counties in each case study, irrespective of the strength of their actual partisan lean. Although these models will not be perfectly accurate for all case studies, their recommendations may still be valuable for policymakers and strategists when coupled with those of well-established instruments. Further research is also critical to ascertain whether retrospective political data trends hold in forward-looking predictive models. If so, these models may carry the potential to play a meaningful role in shaping county-by-county political assessments.

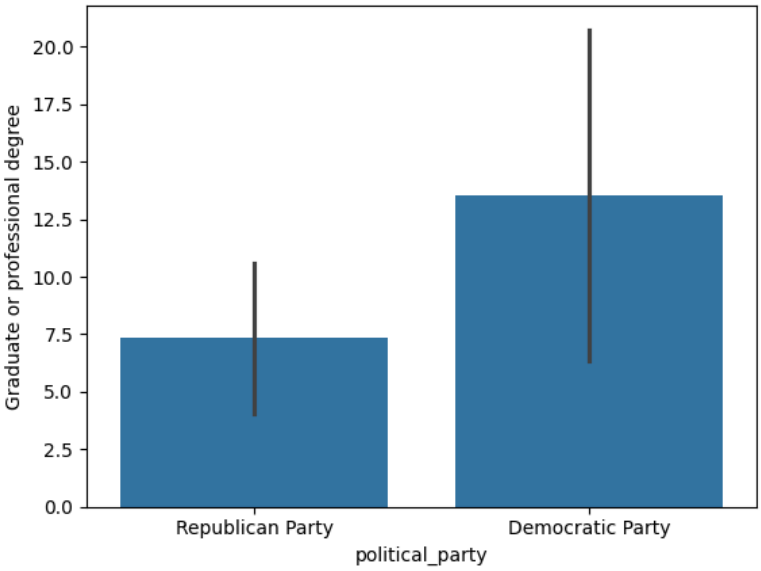
Summary of Findings and Conclusion

From the perspective of an exploratory data analysis and three predictive models, this paper provides insight into the electoral patterns of demographic groups.

- Racial, occupational, and educational features are meaningfully predictive
- Geographic features (population density and area) are largely insignificant
- Demographic composition of a county should be considered on a bipartisan basis

Democratic strategists should employ the findings in these models to determine the most effective way to appeal to the working-class voters that were the cornerstone of their former Blue Wall base. Although the Republican Party has returned to the White House in 2024, its strategists would do well to continue to focus on expanding the party's appeal across racial demographics. Owing to the Random Forest and Decision Tree Classification models, this paper also finds that strategists should narrow their scope to swing counties in swing states, as the latter may not carry that much weight when considered in tandem with other variables. As the American political climate undergoes a rapid transformation from neoliberalism to populism, both parties should ensure they adapt accordingly.

Appendix



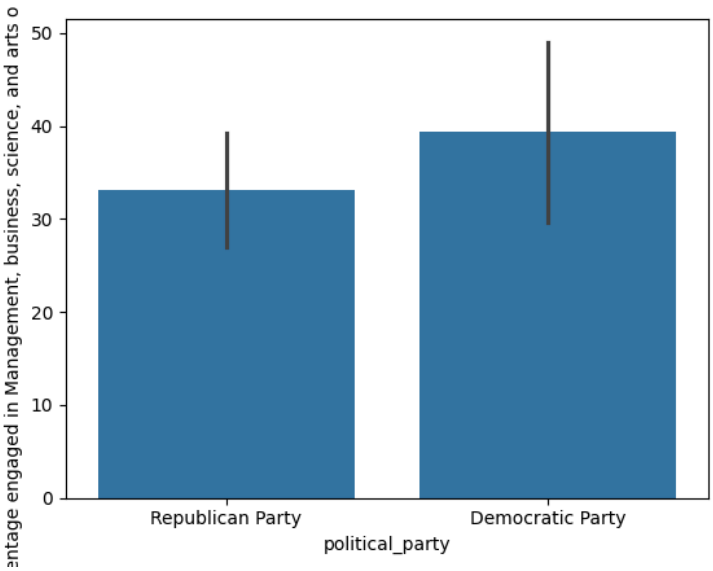
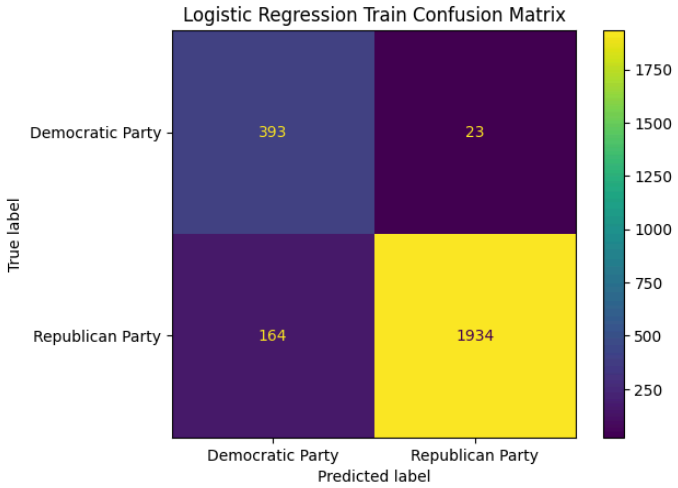


Figure 12: Mean % Graduate Degree by Political Party, Figure 13: Mean % Management et. al by Political party



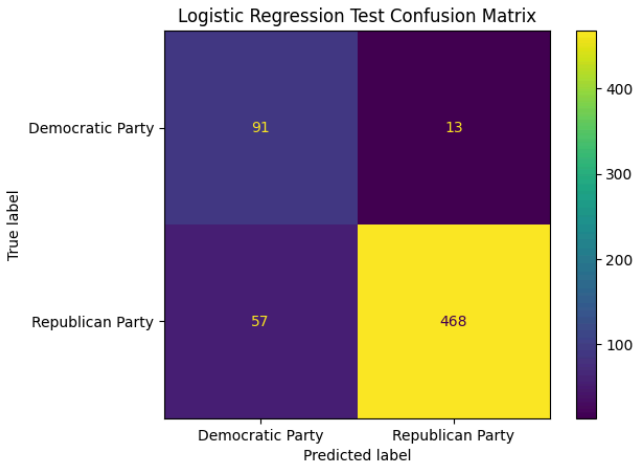


Figure 14: Logistic Regression Train Confusion Matrix, Figure 15: Logistic Regression Test Confusion Matrix

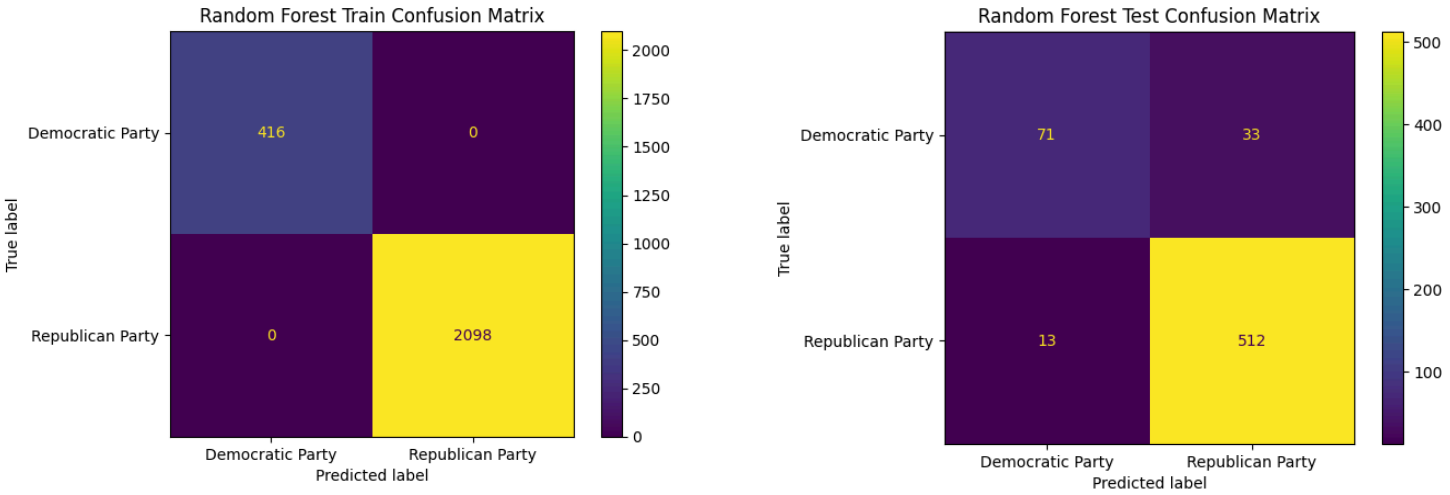


Figure 16: Random Forest Train Confusion Matrix, Figure 17: Random Forest Test Confusion Matrix

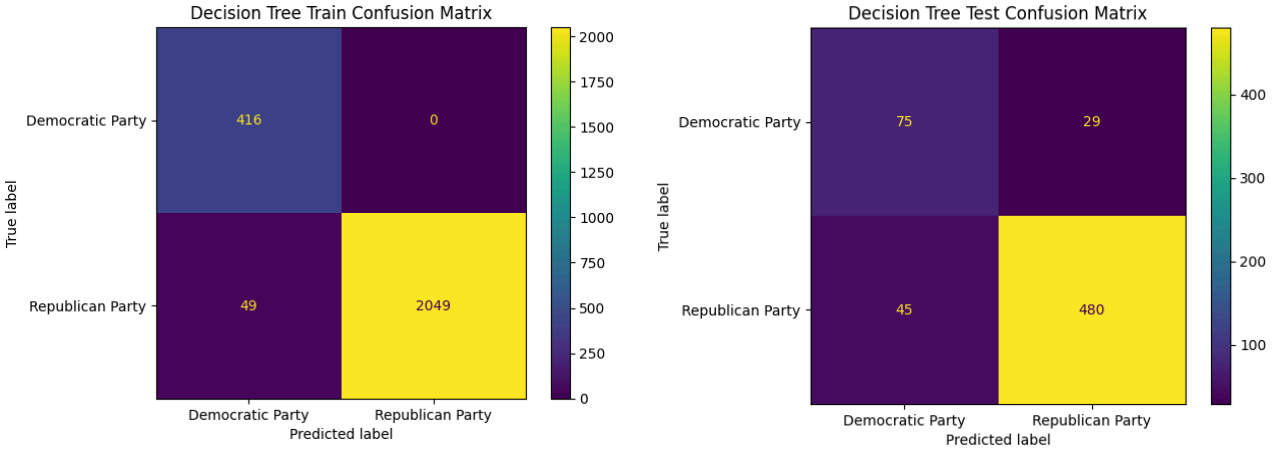


Figure 18: Decision Tree Train Confusion Matrix, Figure 19: Decision Tree Test Confusion Matrix

Works Cited

Cohn, Nate. *How Educational Differences Are Widening America's Political Rift*, www.nytimes.com/2021/09/08/us/politics/how-college-graduates-vote.html.

Frey, William H., et al. "Trump Gained Some Minority Voters, but the GOP Is Hardly a Multiracial Coalition." *Brookings*, 6 Feb. 2025, www.brookings.edu/articles/trump-gained-some-minority-voters-but-the-gop-is-hardly-a-multiracial-coalition/.

Geiger, Abigail. "A Wider Ideological Gap between More and Less Educated Adults." *Pew Research Center*, Pew Research Center, 26 Apr. 2016, www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/.

Rodriguez, Peter. "Playing the Long Game." *The Business School at Rice University (Rice Business)*, business.rice.edu/wisdom/expert-opinion/why-president-george-hw-bush-went-bat-nafta. Accessed 12 May 2025.

Smith, Mitch. *Democrats Again Banked on the 'Blue Wall.' It Crumbled.*, www.nytimes.com/2024/11/06/us/politics/trump-harris-blue-wall-election.html.

"US Election Dataset." *Kaggle*, 6 Nov. 2024, www.kaggle.com/datasets/essarabi/ultimate-us-election-dataset.

Weisman, Jonathan. *How the Democrats Lost the Working Class - The New York Times*, www.nytimes.com/2025/01/04/us/politics/democrats-working-class.html. Accessed 12 May 2025.