Demographic Influences on U.S. Election Outcomes Across Levels: A Machine Learning Approach

Benjamin Qiao

Abstract

Demographic factors such as race and education are strongly correlated with party affiliation in the United States. Using machine learning techniques, this paper uncovers how the relationship between those demographic factors and voting behavior differs *across different election jurisdictions* by analyzing those relationships across senatorial, gubernatorial, and presidential elections. Previous studies surveyed do not incorporate modern machine-learning techniques; incorporating this into the model allows for a numerical scoring of each demographic in its importance of determining election results. The result was a separate importance score per demographic factor for each election jurisdiction analyzed. Among other results, the analysis in this paper showed that racial demographics tend to have the strongest predictive power among all demographic features studied, but have greater importance in presidential elections than in gubernatorial or senatorial elections. This study found that despite large shifts in party platforms and candidate profiles, demographic patterns still hold strong predictive power in how voters will choose candidates in elections across different levels of government.

Introduction

Voters behave differently across various types of elections for different reasons. A governor cannot dictate U.S. foreign policy, and the president cannot manage state infrastructure projects. What people look for on the local level is different from what people look for on the national level; this paper seeks to quantify what that difference is.

Two things can be easily observed with data: the demographics of a county and a county's voting outcomes. In this research, a model was developed using socioeconomic demographics to predict election results from 2010 to 2020, and through analysis, the weight of each demographic parameter in its influence on election results across election levels was identified.

This study's findings confirm that different demographic factors do indeed significantly affect voting patterns depending on the level of office. This study also provides insights into the nuanced ways in which voter behavior and preferences vary by the type of election, contributing to a deeper understanding of electoral dynamics and informing more targeted political strategies.

Literature Review

The influence of demographics on political party affiliation and subsequently voting behavior is well established. In their study identifying how socio-demographic factors determine political party affiliation, Tanuzi and Franklin noted the interaction of specific demographic

patterns that align with predicting party affiliation. For example, this study found that for Black Americans there is a 15.5% higher likelihood that the respondent is affiliated with the Democratic Party on a 0.01 significance level, even after controlling for gender, age, education level, political views, children, income, family income, father's education, and marital status. This study found similar strong relationships between other demographic factors (beyond race) and party affiliation as well.

The trends of partisan affiliation among demographic groups have not changed substantially since 1994 according to a 2018 Pew Research Center study. Among the demographic groups studied,[1] there were only slight changes in partisan identification for the group, with 37% of women identifying with the Democratic Party in 1994 compared with 39% in 2017, for instance (Pew Research Center). These findings suggest that the *relationship between specific socio-demographic groups and party affiliation* has endured in the period studied.

There have been past studies confirming the existence of a difference in voting behavior depending on election level. Kurtbaş found that voters place different priorities on local elections compared to general ones; the study found that nearly half of the surveyed voters (48.4%) conducted little to no research before voting in the local elections. This research also found that older voters and those with lower education and income levels are more focused on local elections and that younger, more educated, and wealthier voters tend to prioritize general elections. These differences in voting behavior based on demographic differences highlight key areas of analysis later in this paper.

Researchers have previously identified some specific differences in how voters hold incumbent parties and politicians accountable for conditions while they are in office, although the differences identified differ. Atkeson and Partin concluded that governors and senators undergo different types of retrospective evaluations by the constituents, finding that governors are held more accountable for state economic conditions than senators, who are linked to the actions of the presidency. In contrast, an earlier 1990 study determined practically the opposite, concluding that voters hold their governor "neither responsible nor accountable" for their state's economic conditions, placing more responsibility on congressional and presidential incumbents (Stein).

Past research surveyed does not fully incorporate modern machine learning techniques in analyzing the differences in voting outcomes *across* election jurisdictions—this research aims to fill that gap. Machine learning allows the quantitative assessment of factors in contrast with previous methods, which have identified relationships but did not provide a numerical ranking of their relative importance. It also allows for large-scale analysis across multiple elections and jurisdictions, taking advantage of existing data sets instead of having to conduct labor-intensive data collection. Finally, machine learning models can handle complex, nonlinear relationships between variables more effectively than traditional statistical methods like linear regression.

[1] Gender, race, education, generation, religion, urban/rural county voter

Data/Methods

This study used a predictive model leveraging county demographics to predict county election results.

The county demographic data used in this study is readily available online from 2010 onwards in the US Census Bureau data set from its American Community Survey program. The demographics considered in the predictive model of this study are median household income, unemployment rate, education (proportion of the county population with a bachelor's degree or higher), age (proportion of the county population aged 60 years and older), place of birth (proportion of the county population that was born inside the United States), minority population (proportion of the non-white county population), and the population density. These specific demographics were selected in line with previous studies.

This study used the 5-year studies and its estimates from 2010 to 2020 (the years chosen correspond with each general election: 2010, 2012, 2014, 2016, 2018, and 2020) for every demographic except population density. The 1-year estimates, while more accurate, are only published for geographies with a population of 65,000 or more. Because of data availability reasons, this study used data for population density *only* from a 2020 Census Bureau dataset, which is used as a parameter in models for every year analyzed.[2] This still captures the essence of whether or not a county is rural or urban, as most county populations have not changed drastically since 2010.

The county election results included in this study come from a dataset created by Algara and Amlani. They include Democratic raw votes, Republican raw votes, and the total two-party votes for each county. A new parameter *result* was created which is the proportion of Democratic votes to the total two-party votes. Using a proportion of votes instead of a binary result (win or loss) allows the model to consider the scale of the victory or loss to adjust the parameter weights accordingly.

Variable	Observations	Mean	Std Dev	Min	Max
Result (Presidential)	5915	0.35	0.15	0.03	0.93
Result (Gubernatorial)	5259	0.38	0.14	0.01	0.93
Result (Senatorial)	7844	0.37	0.15	0.01	0.84

Independent Variables

Variable	Observation	Mean	Std Dev	Min	Max

[2] Used for data availability reasons: www2.census.gov/geo/docs/reference/ua/

	s				
Median Household Income	12134	44881.45	11679.56	10550.00	125672.00
Age 60 yrs and Older	12134	14.98	435	0.00	51.40
Population Density	12134	105.22	166.25	0.04	999.95
Bachelors or higher	12134	18.56	8.40	0.00	68.40
Native Pop. Ratio	12134	0.96	0.05	0.28	1.0
Unemployment Rate	12134	8.11	4.00	0.00	36.00
Percent Minority	12134	15.54	16.20	0.00	96.10

Figure 1. Data Summary Statistics

Both datasets include common metadata including *year* and the county *FIPS* code, which uniquely identifies every county. This allows for the linking of each election result in a county from a specific year to the corresponding demographics of the county in that same year. The Pandas library within Python was used to clean and combine the data. Any observations with missing or highly abnormal data were excluded from this analysis.[3]

Year	**FIPS**	**Median household income**	**Age 60 years and older**	**…**	**Result**
2010	01003	$50,147	15.5	…	0.191781
2010	01005	$33,129	11.0	…	0.487186
2010	01007	$41,770	9.7	…	0.244469
…	…	…	…	…	…

Figure 2. Sample Data Structure

[3] Missing or unusual data includes uncontested elections and demographic data that does not exist for the indicated year and county.

A random forest algorithm was implemented with the Python library *scikit-learn*, a standard, and interpretable machine learning model, using the variable *result* as the outcome variable of the model and the demographics as predictors.

Analysis and Discussion

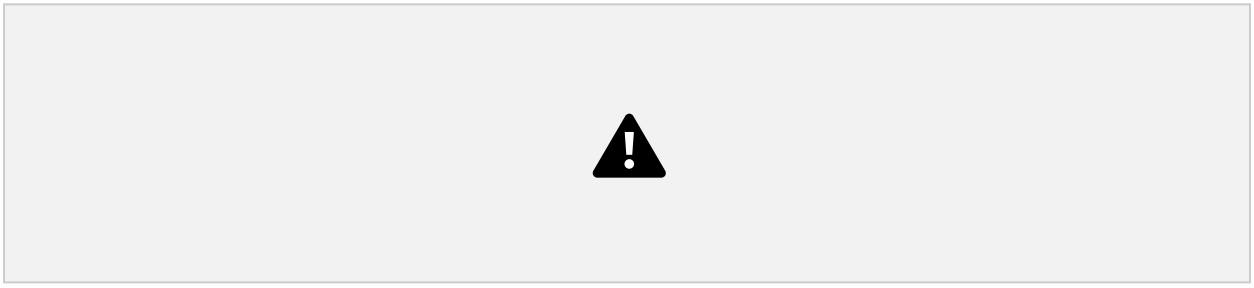

Figure 3. Presidential Election Results (r^2 = 0.57)

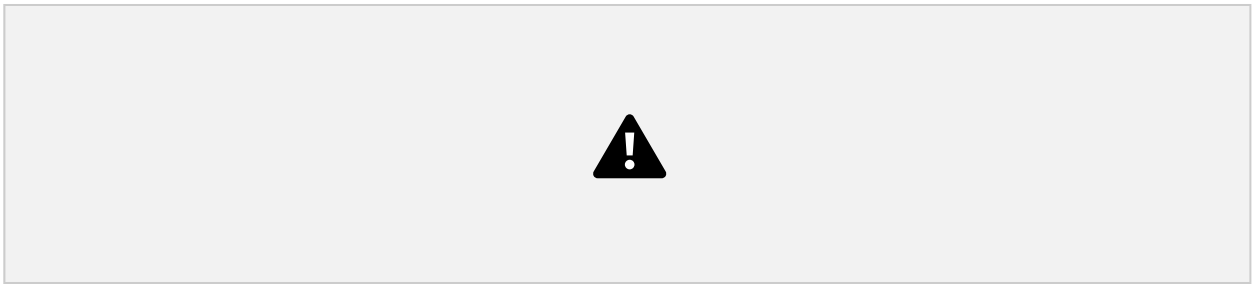

Figure 4. Gubernatorial Election Results (r^2 = 0.38)

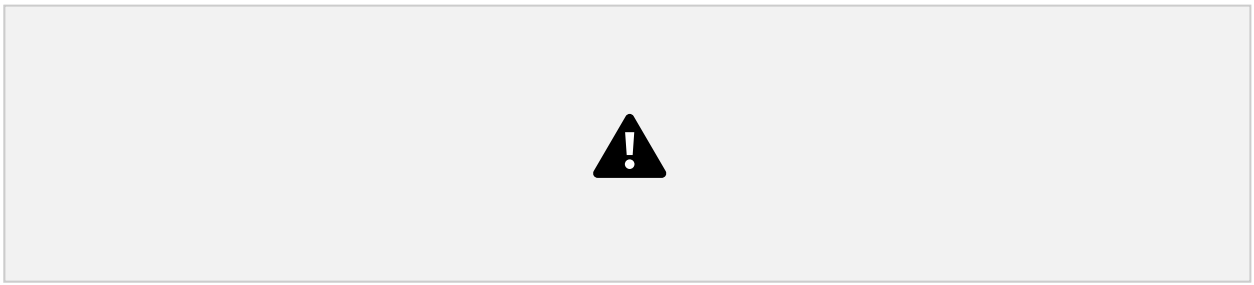

Figure 5. Senatorial Election Results (r^2 = 0.46)

Variables	Presidential	Gubernatorial	Senatorial
Median Income	0.10	0.15	0.11
Age	0.07	0.08	0.08
Pop. Density	0.15	0.16	0.16
Education	0.16	0.15	0.14
Native Pop.	0.10	0.11	0.11

Unemployment	0.14	0.14	0.18
Percent Minority	0.28	0.21	0.23

Figure 6. Relative Feature Importance

The data collected in this study reveals interesting trends that could be expanded upon in future research. The differences in the feature importance of most variables are marginal, but three stand out: income, unemployment, and minority population. There could be a variety of causes for these differences: governors may be held more accountable for regional economic conditions, like state income taxes, while the voter turnout among the minority population may be greater for presidential elections.

Other variables have extremely similar feature importance. Age, population density, native population, and education all predict the outcome of each election level by a similar amount. These variables likely are not affected as much as the aforementioned variables by the different characteristics of each office.

The exact *causes* of these differences or similarities are not clear, additional analysis would be helpful to understand the underlying patterns before arriving at any firm conclusion.

The differences in the relationship between social/demographic characteristics and party voting behavior could help interest groups and parties tailor campaign messaging to better align with select voters. While the actual causal effect of demographic differences on election outcomes across office levels is not vast, they may be significant enough to warrant further investigation and potential campaign strategy changes.

Limitations and Future Work

Several obstacles arose during this study. Data availability was a challenge. For example, religion is a notable demographic variable influencing voter party affiliation, but the US Census Bureau does not collect information about religion due to federal laws. Future studies including religion would need to be smaller, limited by samples collected from private researchers. High-quality election data spanning Presidential, Gubernatorial, and Senatorial elections were also sparse. With higher quality and more prevalent election data, this study could analyze additional offices and jurisdictions beyond Presidential, Gubernatorial, and Senatorial elections. Ideally, additional studies could be conducted on the municipal level, where the differences in voter behavior are likely much more pronounced than in the larger-scale elections used in this study.

In the future, this study could be expanded to include data from the 20th century. There is much less demographic data available from the US Census Bureau before 2005 given that is when the American Community Survey (ACS) began (this study used data from 2010 onward because that is the data easily accessible online). A similar analysis without data from the ACS would have to rely on decennial census records, which are only available for every ten years.

A more retrospective study would also have to take shifting political party platforms and the resulting demographic affiliation shifts into account. For example, the Republican Party under Ronald Reagen and Donald Trump are distinctly different; their supporters are likely also distinctly different. Because of party differences through time, it would likely be necessary to develop separate models dividing the "generations" of political parties, as the importance of each demographic may have shifted significantly over the past few decades. Future studies could compare the importance of each demographic in determining voting outcomes across *party generations* to uncover specific changes in party platforms.

Of note is also how the coefficient of determination (r^2) for each model varies significantly. In the models with a lower r^2, other demographic variables significant to voter behavior are likely missing (omitted variables). Future analysis could incorporate more demographic variables overall into the models.

Conclusion

The base findings of this study do indeed confirm the results of previous studies: demographic factors predict election results distinctly across office levels. This study identified income, unemployment, and minority population as the demographics with the biggest differences in influencing election results, with age, population density, native population, and education having little to no difference.

This study mainly discusses the *association* of certain demographic variables with election results. This research is not intended to discover *causality*, but rather to identify interesting relationships for further analysis in additional studies. Future research could expand upon how the demographic variables discussed in this study directly relate to voter priorities.

Acknowledgements

I would like to express my sincere gratitude to my mentor Brian Lewis for his support throughout the research process. His insights into the data collection and analysis process have been invaluable for the writing of this paper.

References

Algara, Carlos, and Sharif Amlani. *Replication Data for: Partisanship & Nationalization in American Elections: Evidence from Presidential, Senatorial, & Gubernatorial Elections in the U.S. Counties, 1872-2020*. 1.1, Harvard Dataverse, 3 Sept. 2022. *Harvard Dataverse*, https://doi.org/10.7910/DVN/DGUMFI.

Atkeson, Lonna Rae, and Randall W. Partin. "Economic and Referendum Voting: A Comparison of Gubernatorial and Senatorial Elections." *The American Political Science Review*, vol. 89, no. 1, 1995, pp. 99–107. *JSTOR*, https://doi.org/10.2307/2083077.

Census Bureau Data. https://data.census.gov/. Accessed 1 Oct. 2024.

Kurtbaş, İhsan. "The Factors Influencing Voting Preferences in Local Elections " An Empirical Study "." *International Journal of Humanities and Social Science*, vol. 5, no. 9, Sept. 2015, pp. 197–210.

Pew Research Center. "Trends in Party Affiliation among Demographic Groups." *Pew Research Center*, 20 Mar. 2018, https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/.

Stein, Robert M. "Economic Voting for Governor and U. S. Senator: The Electoral Consequences of Federalism." *The Journal of Politics*, vol. 52, no. 1, 1990, pp. 29–53. *JSTOR*, https://doi.org/10.2307/2131418.

Tanuzi, Besjon, and Justin Franklin. *How Do Socio-Demographic Factors Determine the Political Party Affiliation of Americans? Do Race and Region Affect Differently for Similar Political Views Levels?* 4693183, 12 Dec. 2022. *Social Science Research Network*, https://doi.org/10.2139/ssrn.4693183.