Factors Shaping Young Voter Demographics in the U.S.

March 20, 2024

Final Report
Catherine Li
PSTAT 135

Introduction

As we delve into the realm of political engagement, one thing becomes clear: young adults, those aged 18-29, hold immense potential to shape the future of our democracy. Historically underrepresented at the polls, this demographic holds immense potential political influence yet remains largely untapped. Yet, understanding what influences these young voters is crucial for any campaign seeking to truly connect with them.

Targeting young voters is essential for several reasons. Firstly, this demographic stands at the forefront of societal change and technological advancement. They are digital natives, deeply connected to contemporary issues like climate change, social justice, and economic reform. Aligning with progressive ideals, they offer fertile ground for campaigns advocating for these issues. Moreover, investing in young voters cultivates a habit of civic engagement, ensuring a lasting impact on the democratic process.

Additionally, the sheer numerical strength of the 18-29 age group cannot be overlooked. As they come of age, their collective voting power grows significantly. However, their historically low turnout signals a gap between potential influence and actual participation. Tailoring campaign messages, policies, and strategies to appeal to young voters can bridge this gap, contributing to a more inclusive and representative democracy. Appealing to young voters is not just a strategic maneuver – it is a recognition of evolving societal values. By integrating their perspectives, campaigns demonstrate a commitment to inclusivity and forward-thinking.

In this project, we aim to uncover the key factors that influence voter turnout among young adults in the United States. By exploring elements like age, gender, ethnicity, education level, and household income, we hope to shed light on what drives these individuals to the polls – or what keeps them away.

Research Question and Objectives

Research Question:
Which key factors have the highest impact on young voter turnout across the United States?

Objectives:
(1) Explore socio-demographic factors including the relationship between age, gender, ethnicity, education level, household income, and voter turnout among young adults.
(2) Quantify the impact of socio-demographic factors and use statistical analysis techniques to quantify the effects of these factors.
(3) Identify which factors have the strongest correlation with voter turnout amongst young adults

(4) Recommend policy interventions and propose strategies aimed at increasing young voter engagement and participation.

Exploring our Data

Dataset Variable Selection
The dataset involves 700+ variables ranging from voter turnout of past elections to magazine purchase data. A variety of factors can thus be extracted to explain young voter turnout behavior across the United States. For the purpose of making the analysis manageable, we only investigate variables that are relevant to household and regional wealth, education, and racial information. The variable selection process thus involved manually selecting out variables based on this criterion.

Regarding the observation, after the variable selection is complete, to get a proper sampling of the United States youth behavior, we filtered between the ages of 19 and 29 and then sampled 10% of the dataset from each state. From there, we were able to use the concatenated dataset to conduct analysis and model fitting.

Tidying Our Data
To prepare for our analysis on the factors influencing voter turnout among young adults in the United States, we first had to clean and prepare our dataset. This involved several key steps aimed at tidying the data and enhancing its suitability for exploration and modeling.

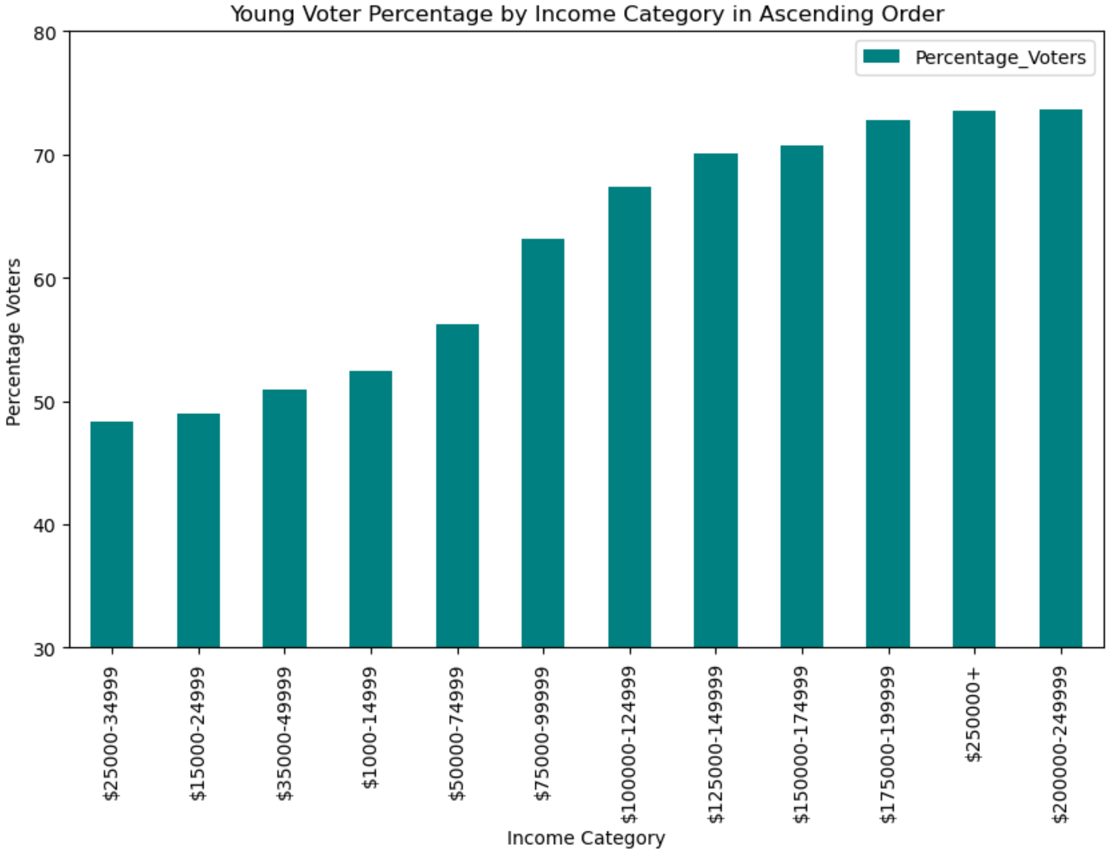
- *Data Cleaning*:
  - Our initial focus was on addressing any missing values and inconsistencies present in the dataset. Using Spark, we conducted a comprehensive assessment to identify columns with missing data. We also employed strategic approaches to handle these missing values, ensuring minimal impact on the integrity of our dataset. For categorical variables, we applied techniques such as imputation with mode values or dropping rows with missing entries, depending on the nature of the data. Similarly, for numerical variables, we utilized appropriate imputation strategies or removed records with missing values, thereby preserving the quality of our dataset.
- *Feature Engineering*:
  - We undertook feature engineering to derive new features or refine existing ones. This involved extracting pertinent information from certain columns, such as extracting state information from voter IDs and converting gender categories into binary representations. Furthermore, we transformed categorical variables into numerical formats using techniques like one-hot encoding or label encoding, facilitating their integration into our predictive models.
- *Preprocessing*:
  - Before model training, we took preprocessing steps to ensure the compatibility of our data with machine learning algorithms. Notably, numerical features were standardized to a common scale to prevent bias, while categorical variables were encoded into formats suitable for model ingestion. These preprocessing techniques not only enhance the interpretability of our models but also contribute to the robustness and efficiency of our analyses.
  - By meticulously addressing data quality issues and refining our features, we were able to lay down a solid groundwork for conducting insightful analyses into the factors influencing young voter turnout. This enables us to gain valuable insights and draw more meaningful conclusions.

EDA - put in graphs here analysis from each graph - how did this help us narrow down our research question and final objectives? - major avenues to investigate: Education, Income, Race

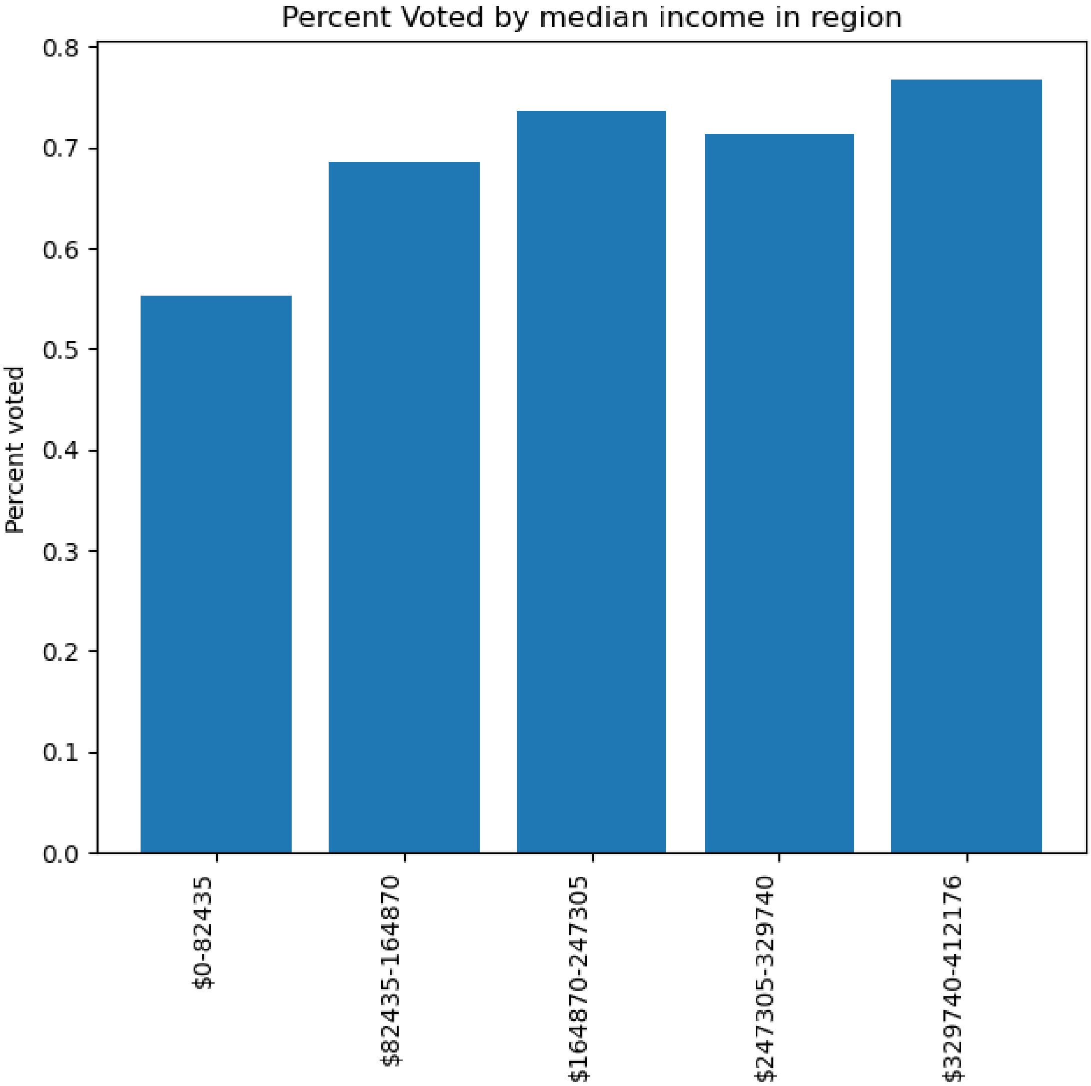
The goal of this EDA is to introduce the variables used in the analysis and gain a better understanding of their relationship to voter turnout.

Household Income

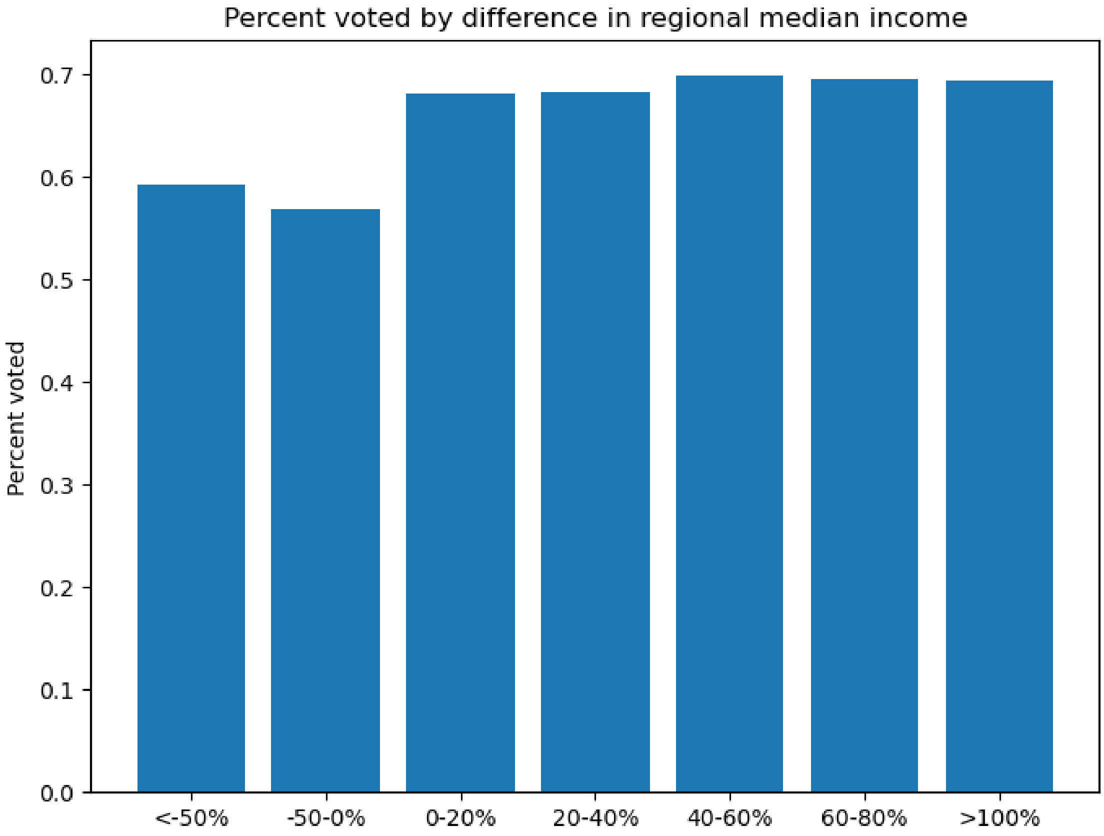
The bar graph presents the percentage of young voters by different income categories, arranged in ascending order of income. There is a noticeable increase in voter turnout as income levels rise, with the biggest difference between income levels being up to 20%. This may suggest that financial stability plays a role in one's ability to engage in the electoral process, or that higher income individuals are more invested in the political outcomes that might affect their economic status. The highest income categories show the greatest voter turnout. This could imply that higher income individuals have more resources and opportunities to vote. While the general trend shows an increase in turnout with income, the relationship does not appear strictly linear, as some adjacent income categories have similar voter turnout percentages. This suggests that while income is an important factor, it interacts with other variables to influence voter behavior. This data suggests that political campaigns and policymakers might need to address economic barriers when attempting to increase voter turnout among young adults.



We are also interested in how the area of living may affect voter turnout. As expected, voter turnout increases as regional median household income increases, but the relationship does not seem to exhibit a strong linear trend compared to direct household income. This trend is to be expected as high individual household income is correlated with median household income in the area. This indicates that median household income in the region contains some sort of caveat on top of household income.

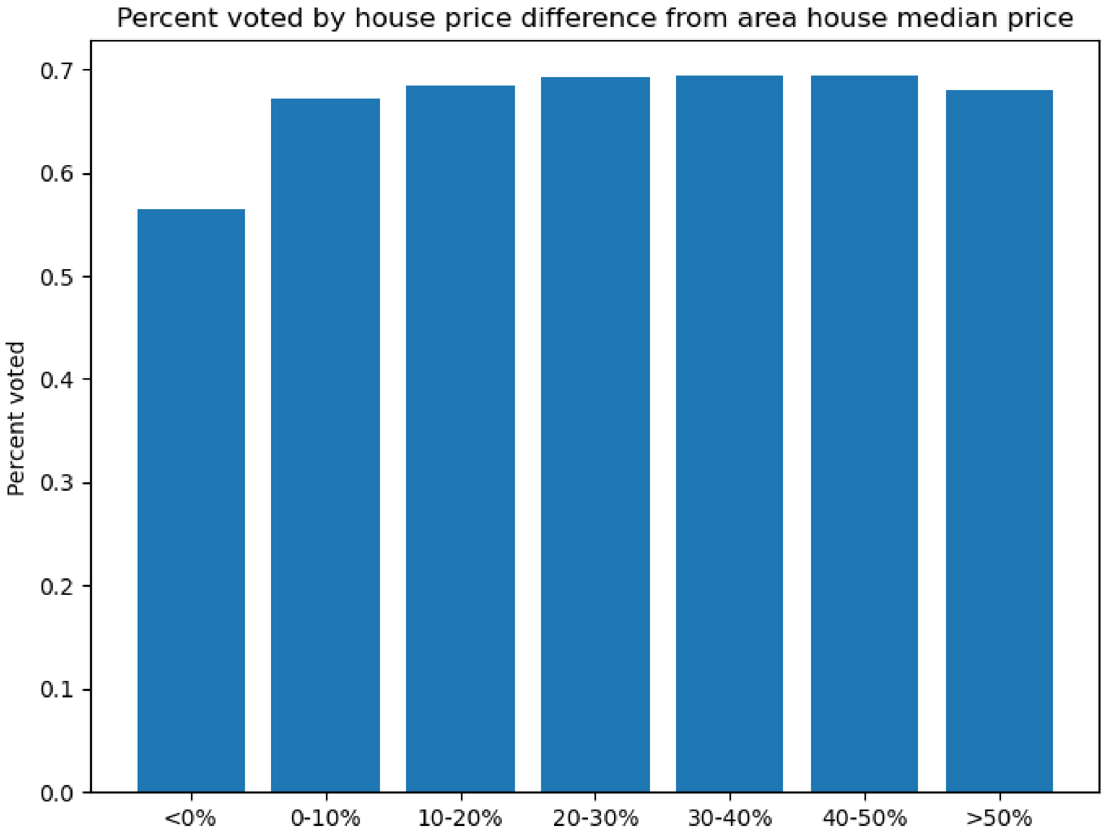


A natural question arises from this: How does the regional welfare affect voter turnout, once you account for household income? The dataset indicates that voter turnout does not change significantly if your household income is higher than the median household income in the region. However, a lower household income with respect to median household income in the region indicates a significantly lower voter turnout rate. Young people who earns less than their neighbors oftentimes have a lower voter turnout rate.



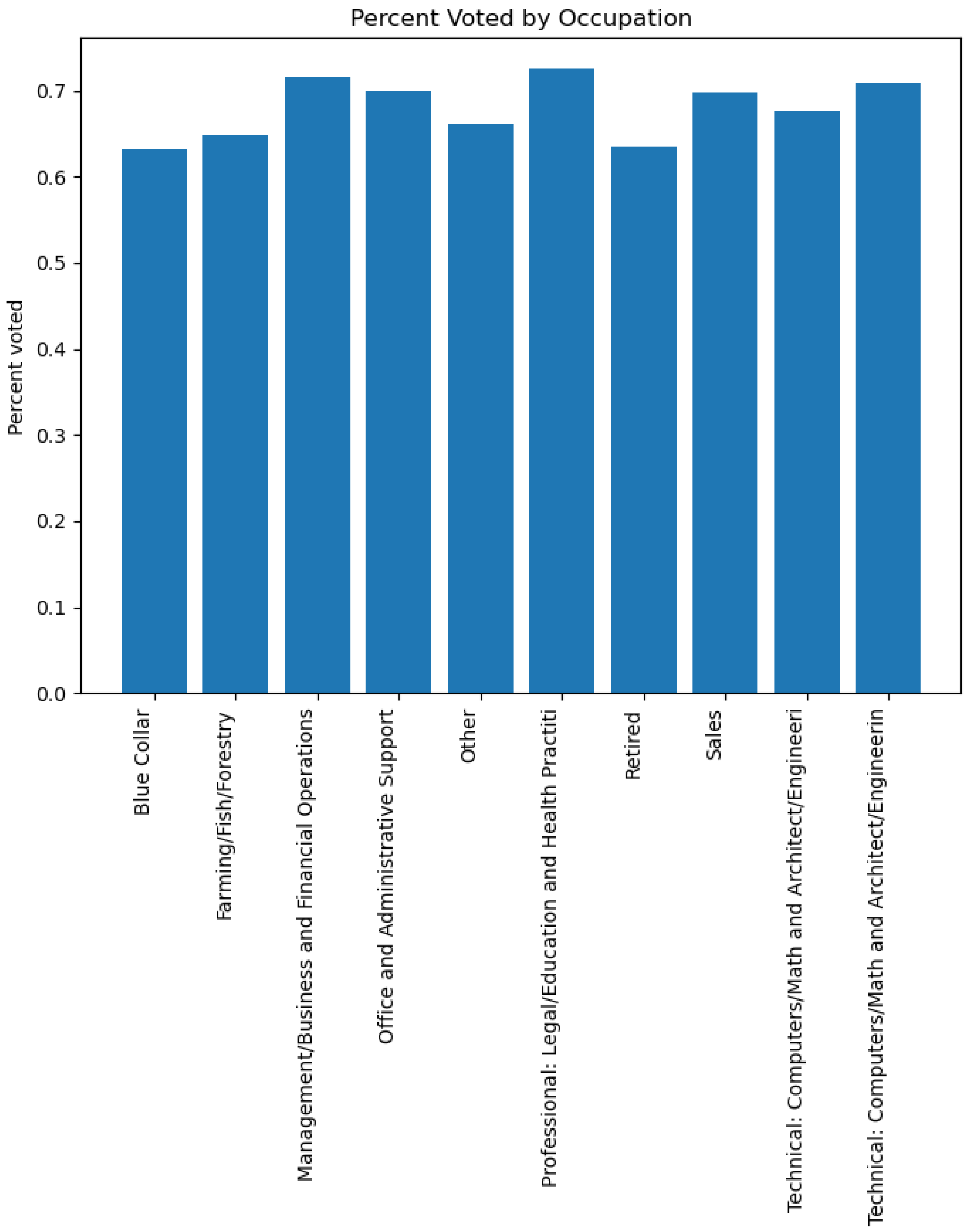
Housing Price

To further explore this, we will take a look at another related variable: household price. Doing the same as we did before, young people who live in housing prices lower than the regional median house price tend to have less voter turnout, but if they live in housing prices above the median house value, their voting behavior stays constant and does not change as the housing price increases.



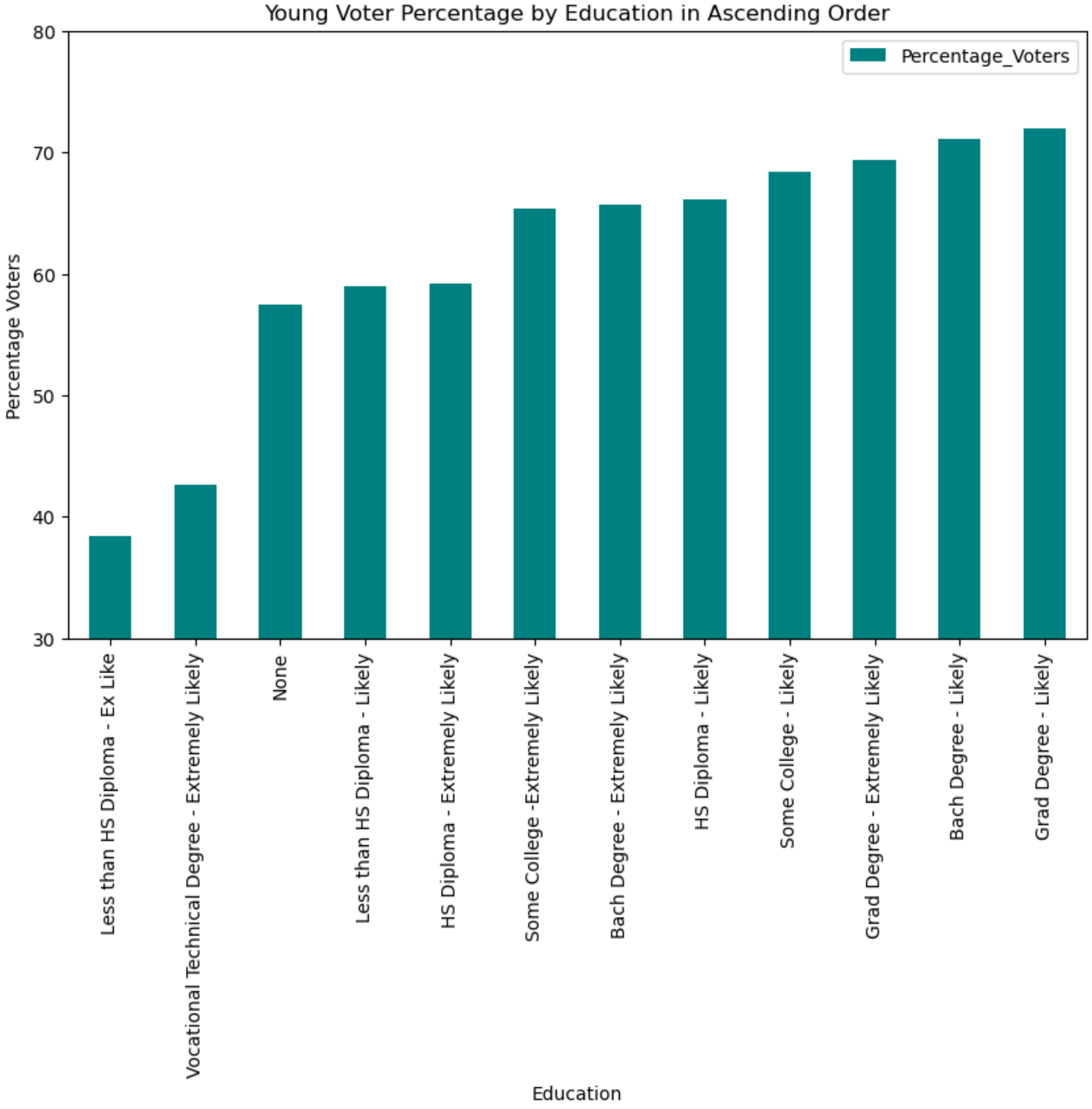
Occupation

Do people's occupation indicate voter turnout? The chart shows that voter turnout is relatively uniform across different occupations, with percentages mostly hovering above the 60% mark. This implies a broadly consistent level of civic engagement across these occupational categories. For political campaigns and organizations aiming to increase voter turnout, this chart suggests that focusing efforts on occupational categories alone might not be as effective as targeting based on other factors like education, age, or ethnicity. However, due to the majority of the data being missing and the relatively low difference, this variable will be excluded from the model fitting.



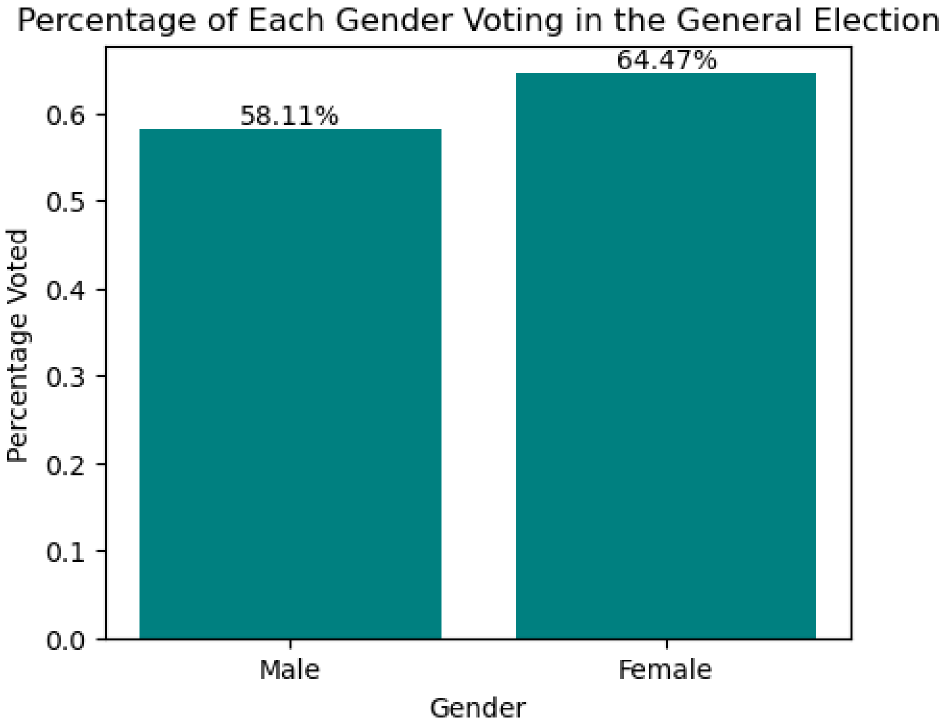
Education

This bar graph illustrates the percentage of young voters classified by their levels of education, with education levels sorted in ascending order of voter turnout. The bar graph shows a clear trend: as the level of education rises, so does the likelihood of voting. This suggests that educational attainment may be a strong predictor of political engagement among young adults. The trend could reflect that individuals with higher education levels are more informed or feel more empowered to participate in the democratic process. They might also have greater access to resources that facilitate voting, such as flexible work schedules or better access to voting information. The pronounced jump in voter turnout among individuals with a ‘High School Diploma’ compared to those with ‘Less than a High School Diploma’ or ‘Vocational Technical Degree’ may point to a critical intervention point.

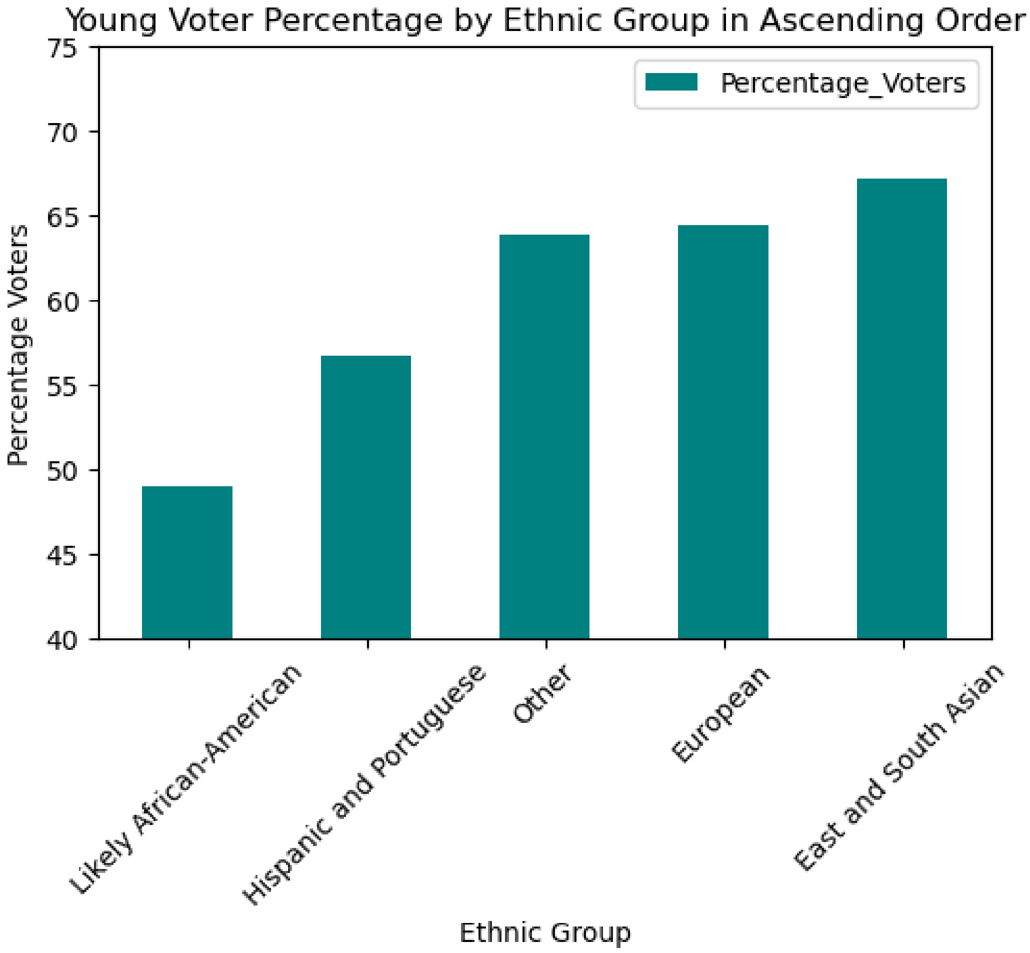


Gender

The bar chart illustrates a significant gender disparity in voter turnout, with 64.47% of females voting compared to 58.11% of males. This 6.36 percentage point difference underscores the higher political engagement among female voters within the dataset’s demographic scope.



Ethnicity: The graph shows the percentage of young voter turnout across various ethnic groups, sorted in ascending order. The visual data suggests a disparity in voter turnout rates among different ethnic demographics. A clear variance in voter turnout among different ethnic groups is present. The 'Likely African-American' group has the lowest turnout, with disparity upward of 20%, while the 'East and South Asian' group has the highest. The data points to a possibility for inclusive strategies that consider the unique circumstances and challenges of each ethnic group. Engagement efforts might need to be culturally tailored to effectively connect with and mobilize these distinct populations. We might also need to explore how socioeconomic factors intersect with ethnicity to impact voter turnout. It's possible that within each ethnic group, variations in income or education levels could further influence these turnout rates.



Models

Logistic Model

- Introduction to the Model:

  The logistic regression model we developed can be used as a predictive analytics tool to understand and predict voter turnout. Logistic regression is a basic classification technique that estimates the probability of a binary outcome based on one or more predictors. It is particularly suitable for scenarios like voting, where the outcome is dichotomous: a person either votes or does not vote.

  Our model uses a range of input variables, from socioeconomic indicators such as estimated household income and home values to demographic information such as age and race, as well as the gender and place of residence of voters. Analyzing historical voting data from this perspective, the model calculates coefficients for each feature, quantifying their impact on the likelihood of voter turnout. A positive coefficient indicates a feature associated with a higher probability of voting, while a negative coefficient indicates a deterrent effect.

  The strength of logistic regression lies in its interpretability and the meaningful probabilities it provides, resulting in a sophisticated balance between simplicity and predictive power. It provides a model that is both easy to understand and informative, providing a crucial probabilistic understanding for strategizing voter participation campaigns or developing public policy.

Random Forest Model

- Introduction to the Model:

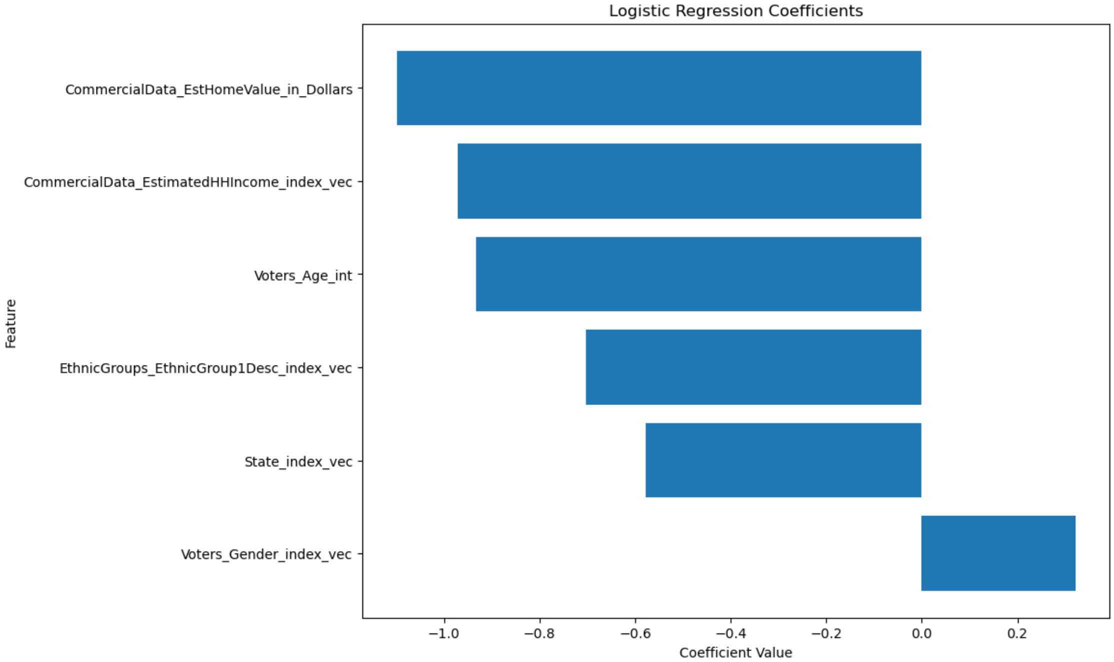
  In understanding and predicting voter turnout, our Random Forest model utilizes the collective decision-making power of multiple decision trees as an advanced analytical tool. This ensemble approach is capable of handling datasets with complex interactions between features with its high accuracy.At the heart of the model is a series of decision trees, each trained on a random subset of the data, that work in concert to improve overall prediction accuracy while reducing the risk of overfitting associated with a single decision tree. The Random Forest algorithm accomplishes this by aggregating the results of individual decision trees to provide a more complete picture of the factors that influence the likelihood of a voter running for office.

  The model is meticulously trained on a variety of features, including demographic data such as age and gender, as well as socioeconomic indicators such as income and home values. Each tree in the forest votes on the turnout prediction, with the majority vote being the final prediction, thus eliminating anomalies and leveraging the power of collective decision-making.

  In the dynamic environment of electoral politics, where a myriad of factors come into play, our random forest model is a powerful framework that allows us to dissect and understand the myriad determinants of voter participation. By doing so, it serves not only as a predictive tool, but also as a lens through which we can observe and interpret the intricate tapestry of democracy.

Analysis and Findings

Logistic Model:



- Model Performance:
  The model had an accuracy of about 63.58 percent, indicating that it correctly predicted whether voters would turn out to vote two-thirds of the time. The accuracy was 56.98% for non-voters and 65.05% for voters. This means that the model is more reliable at predicting who will vote than those who don't.

  The recall rate, or the ability of the model to find all data points of interest, was 26.64% for the negative class and 87.16% for the positive class. This model is significantly better than non-voters in identifying voters. This may be especially useful in situations where maximizing voter identification is more important than avoiding false positives.

  The F1 score is the harmonic average of accuracy and recall, with a negative class of 36.31% and a positive class of 74.50%, reflecting the difference in model performance between the two classes.

  Finally, the ROC AUC score of 63.98% represents the model's ability to distinguish between voter and non-voter categories. An AUC of 50% indicates no discrimination (equivalent to a random guess), while an AUC of 100% indicates complete discrimination. The model has an AUC score of about 64%, showing a moderate ability to distinguish between the two classes.

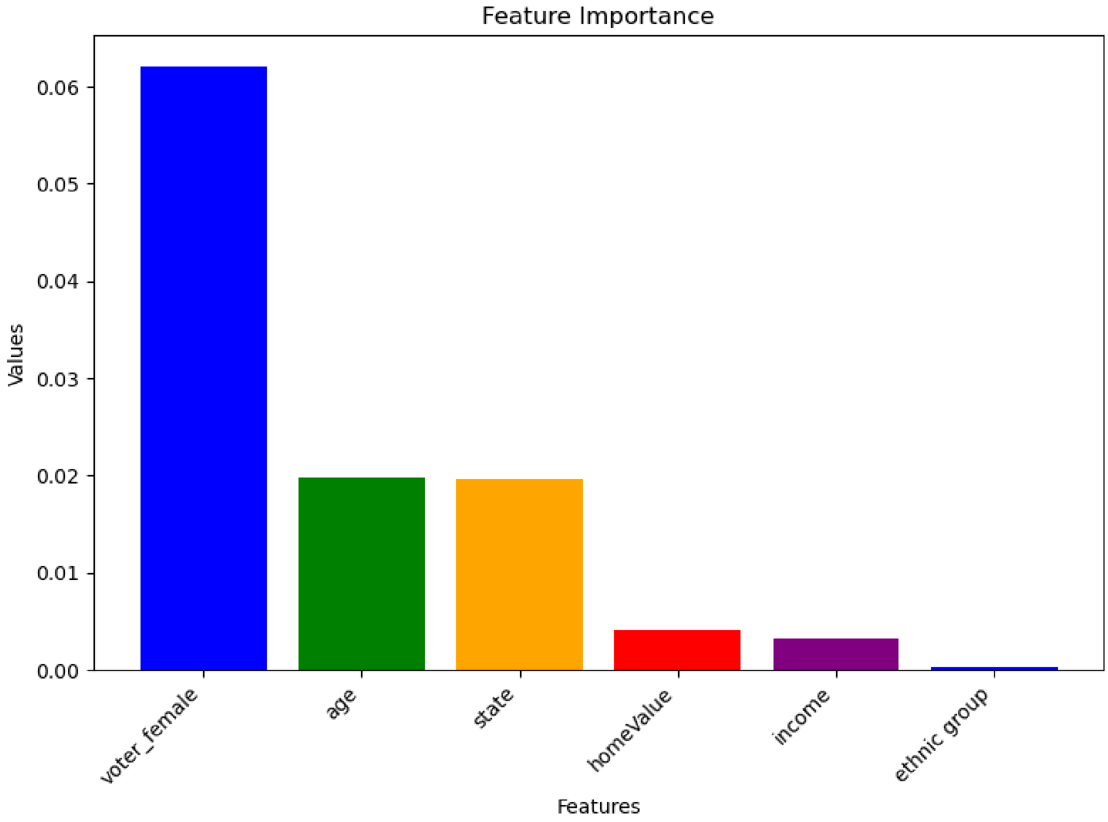
- Feature Importance in Regression:
  CommercialData_EstHomeValue_in_Dollars: This feature has the highest positive coefficient, suggesting that as the estimated home value increases, so does the likelihood of voting. This could be attributed to a greater sense of community investment among homeowners or a correlation between property value and other socio-economic factors that encourage political participation.

  CommercialData_EstimatedHHIncome_index_vec: Next in importance, but with a negative coefficient, this feature indicates that higher household income categories may be less likely to vote. It's a surprising insight that could be explored further, possibly reflecting economic differences in political engagement or the impact of income on voter's priorities or available time.

- Conclusion and Practical Application:
  The logistic regression model has provided meaningful insights into voter turnout, emphasizing certain socio-demographic and economic factors as pivotal in influencing an individual's likelihood to vote. With its moderate accuracy and ROC AUC, and particularly high recall for the positive class, the model proves useful in identifying potential voters, although it does suggest that improvements could be made to enhance its precision and overall balance in classification.

  The model's tendency to better predict who will vote rather than who will not could be particularly advantageous in applications where the cost of false negatives (failing to identify a voter) is higher than that of false positives (incorrectly identifying a non-voter as a voter). Political campaigns and organizations could leverage this aspect to focus resources and outreach efforts on individuals who are on the fence but more likely to vote, according to the model's predictions.

Random Forest Model Performance:



- Model Performance:
For our Random Forest regression model predicting voter turnout, we focus on how closely the model's predictions match actual voter participation rates. Our Random Forest regression model has demonstrated a robust ability to predict voter turnout, achieving an RMSE of 0.47358743671452386, which provides a measure of the average deviation of the predictions from the observed values,an MAE of 0.454078756649114, indicating the average absolute difference be-

tween predicted values and actual turnout figures,and an MSE of 0.22428506021383313 gives insight into the variance of the predictions.For example, an RMSE of 0.474 suggests that, on average, the model's predictions deviate from the actual turnout figures by 47.4 percentage points.

- Feature Importance in Regression:
  One of the valuable outcomes of using Random Forest regression is gaining insights into feature importance — understanding which variables most significantly impact voter turnout. By analyzing the feature importance scores generated by our model, we can identify the key factors that influence the likelihood of voter participation. Voters_Gender_index_vec: Gender seems to have the highest importance score, indicating a strong relationship between a voter's gender and the likelihood of them turning out to vote. This could reflect differences in political engagement across genders or varying effects of gender-based issues on voter turnout. State_index_vec: The voter's state is the second most important feature, which may highlight the impact of local political climates, state-specific issues, or varying election laws that influence voter turnout.

- Conclusion and Practical Application:
  The Random Forest regression model has provided a quantifiable insight into the factors that influence voter turnout. The most significant features — gender and geographical location — suggest that demographic and regional characteristics play pivotal roles in predicting voter behavior.

  Gender, as the most prominent feature, may indicate varied levels of engagement or the differing impact of political issues across genders. Meanwhile, the significance of the voter's state implies that local political environments and state-level policies are crucial determinants of turnout. The importance of estimated home values and household incomes highlights socioeconomic status as a key factor, suggesting that financial stability may affect an individual's capacity or motivation to vote. Age, as expected, correlates with different life stages and possibly varying degrees of political activism or apathy.

Recommendations

- The Insights Gained from the Logistic Model can Inform:

  The insights from the logistic regression model can inform various stakeholders in the political ecosystem. For policymakers, the model's findings could guide the development of initiatives aimed at increasing voter engagement among demographics that are less likely to vote. For electoral campaigns, understanding which factors contribute to higher voter turnout can help in crafting targeted messages and selecting the most effective channels for communication.

  Moreover, the model can serve as a base for civic organizations to design voter education programs that address the specific needs and barriers identified through the model. For instance, a high estimated home value's positive correlation with voting could inspire programs that make voting more accessible in areas with lower property values.

- The Insights Gained from the Random Forest Model can Inform:

  Targeted Outreach: Political campaigns and civic organizations might customize their outreach efforts based on gender and state, recognizing that these factors significantly affect voter turnout. Tailored messaging that addresses specific concerns or motivates particular demographics could prove more effective.

  Policy Development: Policies aimed at increasing voter accessibility and engagement could be crafted with an understanding of the socioeconomic barriers highlighted by the model. This could involve creating more voting locations in areas with lower home values or considering income-related constraints when scheduling elections.

Civic Education: Age being a factor, civic education initiatives could be designed to engage younger or older voters, depending on which age group is less likely to vote, fostering a sense of civic duty and awareness across generations.

Community Programs: The influence of ethnicity suggests that community-based programs that cater to specific cultural contexts might be instrumental in encouraging voter participation within diverse ethnic groups.

- Overall Recommendations:
  - Focus on property ownership and civic engagement:
    Since higher home values are positively correlated with voter turnout, campaigns should target homeowners and emphasize issues related to property and community development. Additionally, in areas with lower property values, campaigns could partner with local organizations to facilitate voter participation, perhaps through community events or forums that address local issues and policies.
  - Tailor messages to income levels:
    With higher household income categories unexpectedly showing a lesser likelihood to vote, campaigns should investigate further and potentially tailor messages that resonate with this demographic's unique concerns. Efforts could include addressing policies that impact higher-income earners or creating more convenient voting options for busy professionals.
  - Enhance accessibility for all income brackets:
    To address socioeconomic barriers, campaigns should advocate for and support policies that make voting more accessible to everyone, regardless of income. This might involve advocating for more flexible voting hours, offering free transportation to polling places, and promoting early voting and mail-in ballots.
  - Educational Programs for Civic Engagement:
    With education appearing as a determinant of voter turnout, the campaign could benefit from educational initiatives that focus on the importance of voting and civic participation. This could be accomplished through partnerships with schools, universities, and non-profit organizations.

Conclusion

Our analysis reveals intriguing patterns about the sociodemographic dynamics of political engagement. Notably, female voters exhibit a higher turnout rate than their male counterparts, suggesting a greater level of political engagement among women—-a trend underscored by the variable importance in our predictive models. Additionally, a strong correlation exists between high home values and increased voter turnout. This suggests an actionable strategy for political campaigns to target their messaging towards homeowners with higher property valuations. Data aggregation supports this strategy, indicating a 70% turnout rate for homeowners compared to 55% for renters.

Contrary to initial assumptions and in contrast to the positive correlation between property value and voting, we discovered that higher household income does not necessarily predict higher voter turnout. This counterintuitive finding points towards a complex relationship between income levels and electoral participation, warranting further investigation into underlying causes.

It is critical to contextualize these insights within the diverse population of the United States, considering the vast disparities in housing costs, living standards, political engagement, and state-specific regulations. The significance of state location in our random forest model underscores the need for geographically-focused strategies in political campaigning.

In conclusion, our study provides a foundational understanding of the factors influencing voter turnout and presents data-driven opportunities for enhancing political engagement. While these findings offer valuable insights for crafting targeted campaign strategies, they also highlight the complex nature of voter behavior, emphasizing the necessity for campaigns to adopt a tailored, culturally competent approach that resonates with the unique needs and concerns of various voting groups. Moving forward, campaigns that harness this knowledge and act upon it with sensitivity and specificity have a good chance at increasing democratic participation.