Using Voter File Data to Study Electoral Reform

Group 5 - Project 64

Carolyn Kolaczyk
NYU Center for Data Science
ck3419@nyu.edu

Doma Ghale
NYU Center for Data Science
dg2491@nyu.edu

Jin Ishizuka
NYU Center for Data Science
ji721@nyu.edu

Rodrigo Kreis de Paula
NYU Center for Data Science
rk4197@nyu.edu

Mentor: Julia Payson
NYU Department of Politics
julia.payson@nyu.edu

Co-instructor: Julia Kempe
NYU Center for Data Science
kempe@cims.nyu.edu

Abstract

The goal of this project is to use voter file data consisting of nearly 190 million records to learn about how electoral reform affects participation in U.S. cities. This will involve wrangling and mining large datasets to uncover patterns in the demographic features associated with voter turnout across cities with and without ranked-choice-voting (RCV). We also develop predictive models to forecast the likely effects of RCV adoption on the turnout for different subgroups of voters.

1 Introduction

Many US cities have begun adopting an alternative voting system called ranked-choice-voting (RCV), in which voters rank candidates in order of preference rather than simply voting for one. Proponents of RCV argue that this is a more fair system than the typical "winner takes all" approach in US politics. Critics claim that it is confusing and might discourage turnout, especially among minority groups (DeSilver, 2021). However, the effects of RCV on voter turnout have not been extensively researched. In this paper, we aim to explore how RCV affects voter turnout, with a focus on turnout across different demographics.

We perform exploratory analysis and predictive modeling to better understand the voting patterns and differences in demographics between RCV and non-RCV cities. We create city-level voter profiles to visualize voter turnout across both demographic groups and RCV versus non-RCV elections. Additionally, we examine the relative importance of various demographic features in predicting voter turnout using random forest models.

We are able to draw some insightful conclusions regarding the demographic profile of voters for each election type (general, consolidated general, and local or municipal) and for RCV vs. non-RCV elections. We find that RCV elections record a higher overall turnout in local/municipal elections, showing that it could potentially increase voters' participation in elections with more than two polarized candidates. However, outside of local and municipal elections, we find that ranked-choice-voting does not seem to have a significant impact on overall voter turnout. All in all, we believe that our initial exploratory analysis of the voter file dataset may prove useful for future research on similar topics.

This is the Final Report of the discipline DS-GA 1006 - Capstone Project and Presentation of the Masters of Science in Data Science at New York University. Fall 2022.

2 Related work

The issue of examining factors that affect voter turnout is not a novel one. Hajnal, et al. (2022) examined efforts to improve turnout in local elections. Specifically, they examined the effectiveness of a movement to shift the timing of local elections so they are held on the same day as national contests. By examining data on election timing along with voter demographic data, they found that on-cycle elections in California not only increased local election turnout, but also resulted in an electorate that was considerably more representative in terms of race, age, and partisanship.

For our project, we aim to continue exploring the factors that influence turnout by focusing on the rising popularity of ranked-choice-voting and its effects on voting patterns. Information on the effectiveness of ranked-choice-voting and other alternative voting systems has been largely speculative with limited statistical analysis. We aim to begin exploring the effects of ranked-choice-voting by examining voter files, a relatively novel dataset. These databases provide a nationwide overview of voter registration and election turnout but are lesser known by the general public and have not received much scholarly attention (DeSilver, 2018). By leveraging this extensive and relatively under-utilized dataset, we hope to uncover new and insightful findings and shed light on the effects of alternative approaches to voting.

3 Tasks

City Voter Profiles
Using voter file data containing both voting history and demographic information, we explore how RCV affects voter turnout, overall and across different demographics. We use this data to create city voter profiles, which summarize turnout and demographic information per city election, for a subset of recent elections in each city. We focus our exploration on 8 states, selecting all cities in these states that have adopted RCV, as well as using a similarity measure to select non-RCV cities to include in our analysis. We select a subset of recent elections for each city and create aggregated turnout statistics for each. A majority of our work focuses on data cleaning and processing, due to the historically infrequent use of the voter file data.

Feature Importance Analysis
For the second part of our analysis, we use a combination of voter file and census data to examine which demographic features most strongly predict voter turnout. We train several random forest models and examine the feature importance scores for each. Specifically, we look at (1) which features predict overall turnout and (2) whether there is a difference in the features that predict turnout in RCV versus non-RCV cities.

4 Experimental evaluation

4.1 Data

City Voter Profiles
For this project, we sampled from eight states which have implemented ranked-choice-voting: California, Colorado, Maine, Maryland, Minnesota, New Mexico, Utah, and Vermont. In order to obtain a sample of demographically similar RCV and non-RCV cities, we implemented a cosine similarity function. For each RCV city, we selected the five non-RCV cities in the same state with the most similar demographics, using our cosine similarity function on census data at the city level. We decided on sampling the five most similar non-RCV cities for each RCV city after discussions with our mentor. Future researchers can experiment with this number. We modified our selection threshold to ensure that we had at least 30 non-RCV cities per state, if possible. If a state had data on less than 30 cities in total, we included all cities in that state. If a state had less than 6 RCV cities, we selected a larger sample of the most similar non-RCV cities to get closer to our target of 30 non-RCV cities per state. The demographic features used for determining similarity were population, median age, median household income, average home value, median rent, percentage of the population with a college education or above, unemployment rate, percentage of the population identifying as white, and percentage of the population identifying as Hispanic (code link). This resulted in 30 RCV cities and 183 non-RCV cities for our analysis, whose geographic distribution is exhibited in Figure 1.

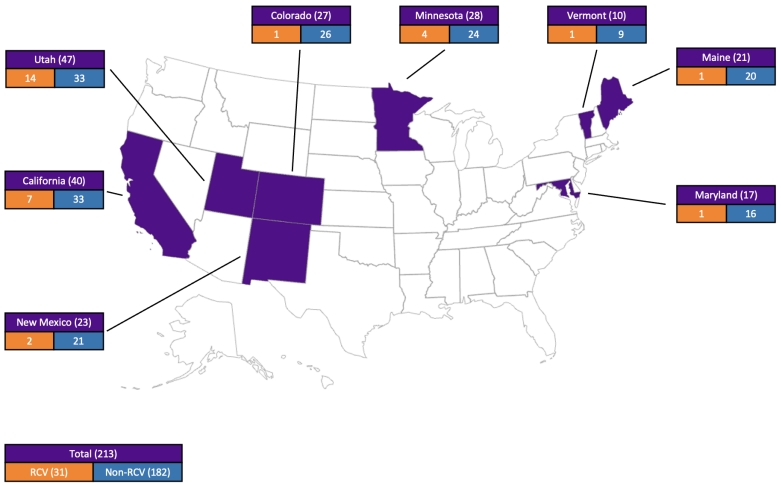

Figure 1: Pooled States and number of RCV and non-RCV Cities utilized

Our voter-level data comes from the L2 Political Academic Voter File (link here), which is a continuously updated database of registered voters in the US. The voter file contains two types of files: Demographic and Vote History. The demographic file consists of 691 detailed socio-demographic variables for each voter. The vote history file contains information on which elections each voter has voted in since 1994.

The size of both the demographic and vote history files presented some early challenges. For instance, California's demographic file was 59 GB, while its vote history file was 29 GB. To make the files more manageable, we pre-selected 14 features of interest from the demographic file: Voter ID, address information (city, county), demographic information (birth date, ethnicity/race, education, gender), official registration date, and financial information (income, number of donations and total donations amount). Similarly, we selected three election types from the vote history file: general, consolidated general, and local/municipal. We converted both of our data files to parquet format for more efficient columnar storage (code link). We then merged our sub-sampled demographic and vote history files on VoterID.

Since ranked-choice-voting is a relatively new form of voting, we limited our analysis to the four most recent elections per election type per city, occurring in 2008 or later. However, we encountered two problems when attempting to select these four most recent elections using the merged file: (1) a mismatch between the city names in the census file and L2 file, and (2) low voter counts for some cities in some elections.

In order to address the inconsistencies in the city naming conventions between the census and L2 datasets, we used the FuzzyWuzzy Python package. The algorithm uses Levenshtein distance, which calculates the number of edits required in one string to produce another string. For each city in the census file, we used the package to generate the top-10 closest city names in the L2 file (code link). We then verified which of the ten potential L2 city names was the correct match for the census city name by simply entering each name into Google Maps to see if they indeed referenced the same location.

The second issue we faced during our data processing was unreasonably low voter counts for some elections in some cities. After some investigation, we found that the issue came from the fact that the vote history file simply tracks an individual's history of participation in different elections and does not track whether an individual moves to a different city between elections. Therefore, there were cases where a voter currently residing in California may have been credited with voting in an election in New Mexico. This election in New Mexico would then incorrectly appear in our merged dataset as a California election with almost no other voters. To remedy this issue we simply discarded any elections with a voter turnout under 5% of the city population (code link). After this adjustment, our data was ready to create voter turnout profiles for each city, state, and election.

Feature Importance Analysis

To prepare data for our random forest models, we selected five target variables from the merged file discussed previously: overall turnout, turnout among white voters, the average age of voters, turnout

for voters with income < 50k, and turnout for voters with income > 100k. We also used additional demographic information from the census data as features. These features included state, election type, RCV (binary flag indicating whether a city uses RCV voting), city population, percentage of city population identifying as white, percentage of the population with a college degree or higher, median age, and median income. These census features were merged with our selected target columns to use as input to our models.

The final merged data file can be found on this link, and all of our code used for data cleaning, aggregation, visualization, and models can be found on this link.

4.2 Methodology

City Voter Profiles
We created demographic voter profiles at the city and election levels using the pre-processed data obtained through the processes described in 4.1. For each city and election, we calculated voter turnout information across demographics (code link). This process involved determining the proportion of eligible voters who turned out to vote among different voter groups. These groups included racial/ethnic affiliation, income level, education level, age, and donation activity. We retrieved an estimate of the total number of residents of voting age using the count of voters over 20 years old from the census data.

We then used these city voter profiles to create our visualizations displaying the difference in turnout for RCV and non-RCV elections across demographics. We created visualizations for all states combined and separated, based on RCV versus non-RCV category, and further broken down by election type and different demographics and/or election date. Below is the summary of the visualizations we performed.

1. Average turnout in local or municipal, consolidated general, and general elections.

   We used the proportion of voters' age 20 and above from the census file to compute an estimate for the voting age population and plotted a bar graph by election type.
2. Distribution of the average gap in turnout between general and local or municipal elections plotted as a bar graph by election type.
3. Average turnout for asian, black, hispanic, other, and white voters in local or municipal, consolidated general, and general elections plotted as a bar graph by election type.
4. Average gap in turnout between white and non-white voters in local or municipal, consolidated general, and general elections plotted as a bar graph by election type.

   We combined all races/ethnicities other than white to create the non-white race/ethnic category.
5. Average voter age in local or municipal, consolidated general, and general elections plotted as a bar graph by election type and election date.

   The age column in the demographic file is a static column, so we used the birth date of each voter to calculate their age at each election date, by the difference between both features.
6. Average voter income in local or municipal, consolidated general, and general elections, plotted as box and whiskers, and histogram.
7. Average percent of voters making contributions in local or municipal, consolidated general, and general elections as a bar graph by election type.

Feature Importance Analysis
Our primary objectives in the predictive modeling portion of our project were two-fold. First, we wanted to examine whether RCV influences voter turnout. Second, we wanted to determine whether or not there were significant differences in the features that predict turnout in RCV versus non-RCV cities. To do this, we trained several random forest models.

We chose to use random forest models for several reasons. First, random forests use an accessible framework that can be easily understood. This can be particularly useful for those in the political science field who may not have a background in mathematics or computer science but simply wish to gain a better understanding of this data. Second, the data we used for our modeling was relatively

minimal, consisting of only 355 rows and 17 columns. Given the relatively simple nature of our data, we felt that implementing a straightforward approach such as random forest would be suitable. Lastly, we wanted to use a model that could make the most out of our limited data. The random forest framework makes it easy to implement a bootstrap approach, where data are randomly sampled with replacement. By repeatedly subsampling the data in this manner, we can, in theory, simulate sampling from a larger dataset with the same statistical properties as our original sample.

As a first step in our analysis, we implemented a random forest model using the full RCV and non-RCV datasets. We split the data into training and test sets using an 80-20% split. We then conducted hyperparameter tuning using a 5-fold cross-validation over a set of 45 different combinations of hyperparameters. After fitting a model on the training data using the optimal combination of parameters, we then examined the model's feature importance scores to determine if a city's having RCV voting was an important contributor towards voter turnout.

As a follow-up analysis, we trained two more random forest models. This time we split our data into two separate datasets, one from RCV cities and one from non-RCV cities. Our RCV dataset consisted of 52 rows and 16 columns, while our non-RCV dataset consisted of 303 rows and 16 columns. Given the small sample sizes, we forwent splitting our data into training and test sets for this analysis. We then trained one random forest model on the RCV dataset and one random forest model on the non-RCV dataset using the previously described bootstrap approach. After training, we compared feature importance scores between the two models to determine if RCV could potentially influence the factors that predict voter turnout.

We also implemented models to predict a variety of other outcomes (e.g., turnout among white voters, the average age of voters, and turnout for low and high-income voters). These outcomes are not the focus of this paper, but the feature importance scores for these models are included in the Appendix for reference.

4.3 Results

City Voter Profiles

For this paper, we will focus our discussion of city voter profiles on our most salient observations. A complete collection of all of the visualizations we constructed from our city voter profiles can be found in the Appendix. Additionally, for all analyses, we will note that there were 37 cities and 151 records (10%) for local or municipal elections, 119 cities and 469 records (31%) for consolidated elections, and 213 cities and 872 records (58%) for general elections.

In Figure 2, we compare overall voter turnout in RCV versus non-RCV cities across election types. We find that non-RCV cities record stronger voter turnout than RCV cities during general and consolidated general elections. Interestingly, however, RCV cities report higher turnout during local or municipal elections. This observation may suggest that ranked-choice-voting is potentially more effective at encouraging turnout in elections with more than two polarized candidates.

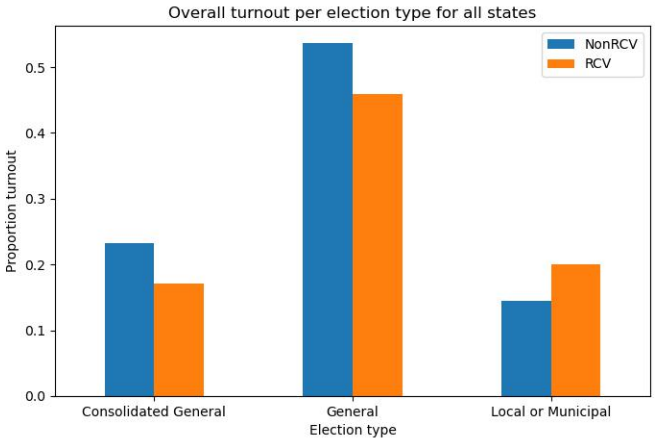

Figure 2: Voter turnout per election type for pooled states RCV vs non-RCV

In Figure 3, we examine the average voter age for RCV versus non-RCV cities across election types. Here we find minimal differences. The average voter age appears slightly lower for RCV cities in general and consolidated general elections but slightly higher for non-RCV cities in local or municipal elections. However, with our limited sample size, it is difficult to come to any definitive conclusions on this matter.

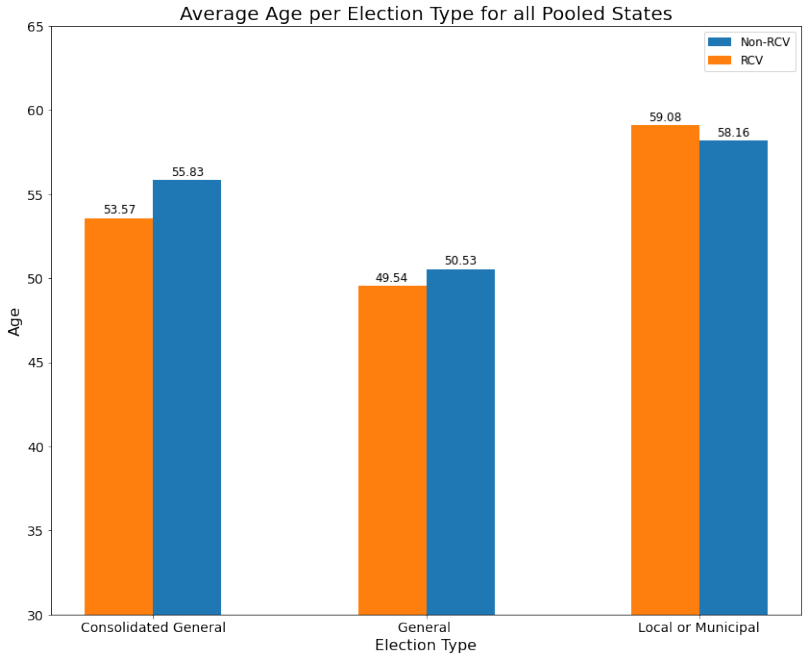

Figure 3: Average Age per Election Type for all Pooled States

In Figures 4 and 5 we examine the distribution of voter income across election types and between RCV and non-RCV cities. In Figure 4, we see that voter incomes for general and consolidated general elections are more widely dispersed and tend to skew toward a slightly higher average voter income. When examining RCV versus non-RCV cities, we find that RCV cities tend to record a slightly higher average voter income than non-RCV cities.

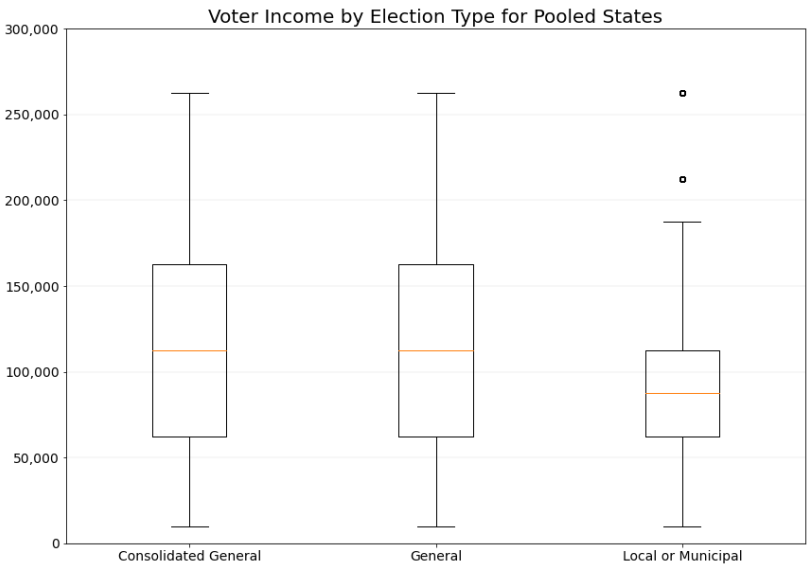

Figure 4: Voter income by election type

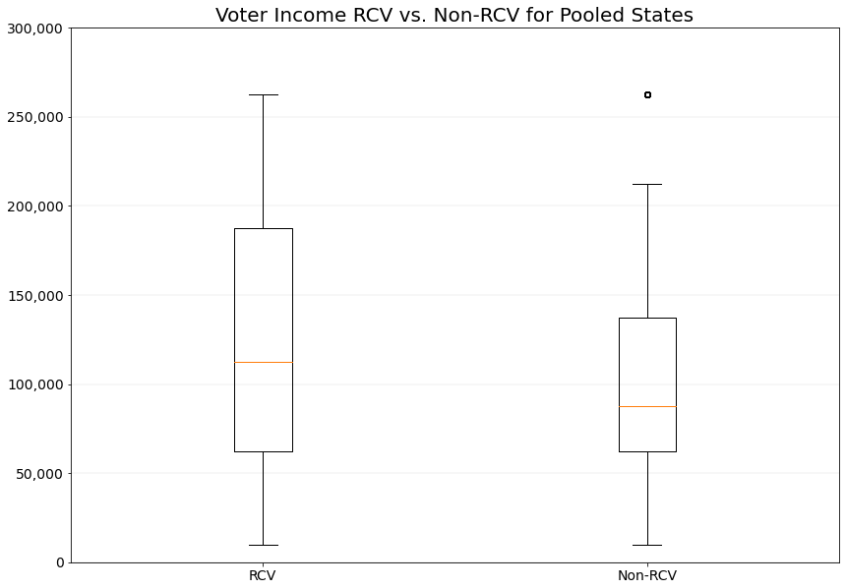

Figure 5: Voter income in RCV versus non-RCV cities

In Figure 6, we examine turnout by race across election types. We find that local and municipal elections tend to attract a more racially diverse body of voters. While general and consolidated elections record white voter turnout to be 11 percentage points and 8 percentage points higher respectively than the racial group with the second highest turnout, local and municipal elections record a disparity of less than 3 percentage points between white voter turnout and the next highest racial group's turnout.

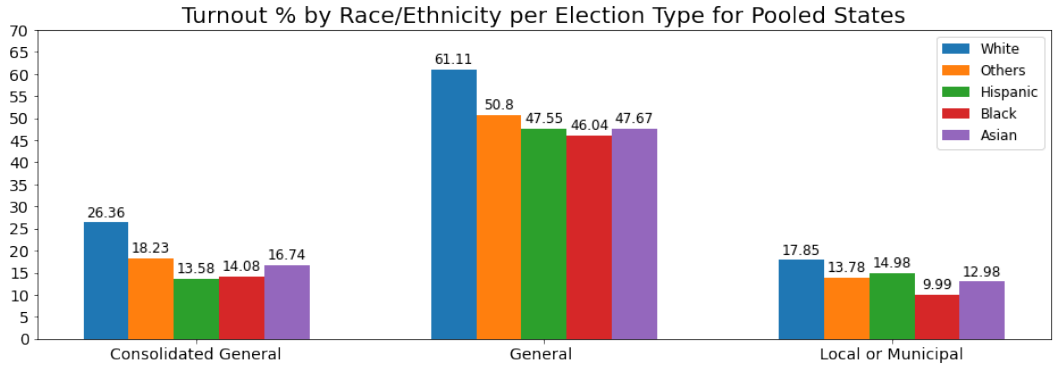

Figure 6: Average voter turnout per election type and race

Feature Importance Analysis

As a first step in our analysis of the features that predict voter turnout, we examined how ranked-choice-voting compared to other factors. After training and tuning a random forest model to predict turnout, we were able to achieve an R-squared score of 0.926 on our test set. We then examined the feature importance scores of the model. We found that the dominating features for predicting turnout were the election types (i.e., whether the election was a general, consolidated general, or local/municipal). Other demographic factors such as race, income, age, population, and state were comparatively minimal. Notably, we found that RCV was among the least influential factors when predicting turnout in our model. These results are displayed in Figure 7.

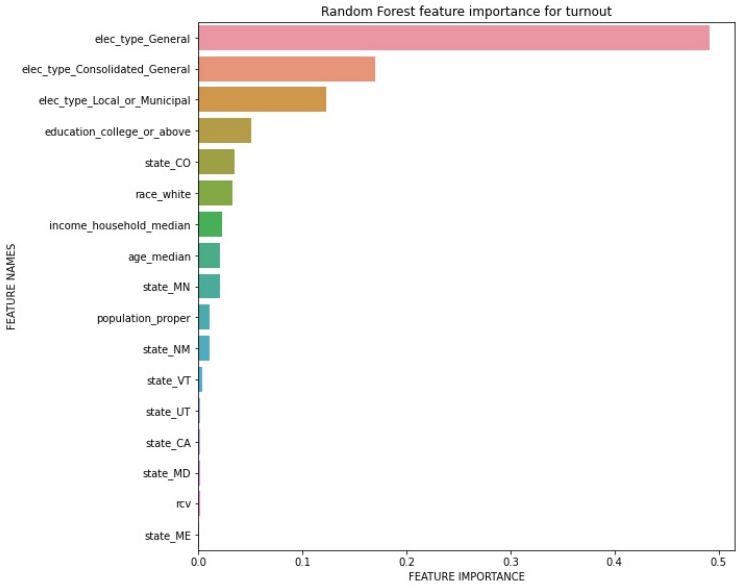

Figure 7: Feature importance scores for random forest model predicting voter turnout

For the second part of our analysis into feature importance for predicting turnout, we investigated whether feature importance scores differed between RCV and non-RCV cities. For this, we trained two models. One using data from RCV cities (R-squared=0.869) and one using data from non-RCV cities (R-squared=0.938). We then compared the feature importance scores from these two models and visualized their results in Figure 8.

We found that the results for RCV and non-RCV cities were similar, as shown in Figure 8. Across both categories of cities, we found that the dominating predictor of turnout was whether or not the election was a general, consolidated general, or local/municipal election. Similar to our first model, we saw that other demographic variables appeared to have minimal influence on turnout regardless of whether or not a city has implemented ranked-choice-voting.

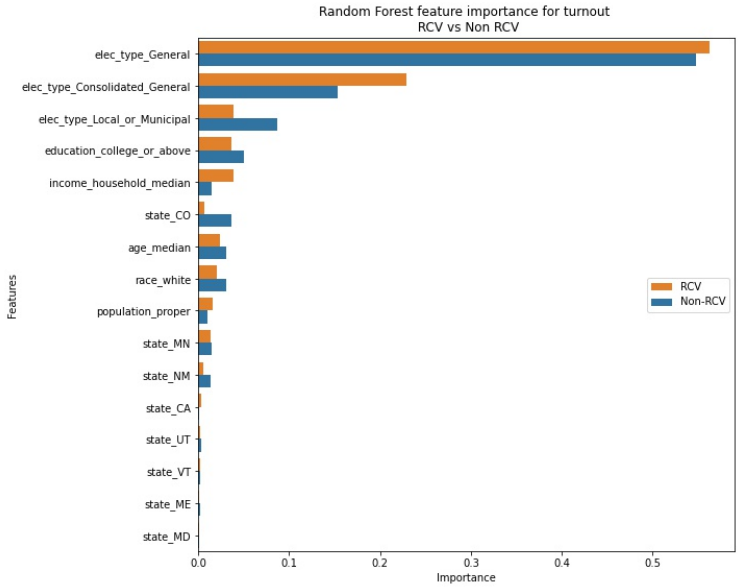

Figure 8: Comparison of feature importance scores for random forest models predicting voter turnout in RCV versus non-RCV cities

5 Conclusions and Future Work

Our project aimed to explore the effects of ranked-choice-voting on voter turnout by using the L2 voter file. In our approach, we first created city-level profiles of election activity for a sample of RCV and non-RCV cities and used these profiles to generate visualizations of the underlying patterns in this relatively novel dataset. We then further explored the factors that predict turnout by implementing several random forest models and examining the resulting feature importance scores.

As a whole, our results were mixed in their ability to provide support for or against the adoption of ranked-choice-voting. One of our most notable findings was a tendency for RCV cities to report higher voter turnout for local and municipal elections than non-RCV cities. These results could potentially be used to bolster support for ranked-choice-voting as they may indicate that this alternative voting system may provide a more appealing voting method in situations where there are more than two primary candidates. However, when examining the features that predict turnout in our random forest models, we found that ranked-choice-voting seemed to have minimal influence on overall turnout. Despite mixed results, we believe that our exploratory analysis of the relatively under-researched voter file dataset provides a promising starting point for future research in the area.

In terms of future work, we would like to expand our analysis to include more states and cities, as our current analysis focuses on 213 cities from 8 US states. Including more cities and states might provide us with a more reliable and significant understanding of the impact of RCV voting on turnout across different demographics. We would also be interested in creating models to predict voter turnout at the individual level, given that the L2 Voter File data already contains information at the individual voter level. In this sense, we would like to create models to predict whether a specific individual will turn out to vote at a particular election using information about both the individual's demographic background as well as the individual's voting history.

6 Lessons learned

One of the first challenges we faced was working with the demographic and vote history files for each state. Given the large size of the data, we were not able to combine all states into a single file. Instead, we worked with each state individually by sampling only the relevant features and converting all datasets to parquet files. Only after extensive pre-processing and the creation of summary voter profiles for each state were we able to combine all information into a single file for further analysis. Through this experience, we learned to appreciate the value of having an organized database that we could query in order to obtain only the data needed for analysis.

Since our project was largely exploratory in nature, many of the issues we encountered revolved around data pre-processing. One issue we encountered was that some city names in census data did not match city names in the L2 file. Additionally, there were often different variations of the same city name within the L2 file. We were able to fix these problems by first using the FuzzyWuzzy package to compile a list of probable city name matches and then manually reviewing these potential matches with Google Maps to verify that they indeed referenced the same city. Although such a process is not scalable if we wanted to expand our sample of cities, we believe that these issues provided us with useful experience for future projects as data cleaning and pre-processing are constant challenges in data science. It also showed us the importance of standardization across multiple sources of data.

A final issue we encountered was a limited sample of RCV and non-RCV cities. This proved to be a challenge during the modeling portion of our project. In order make the most out of our limited data, we chose to use random forest models whose framework allowed us to easily implement a bootstrapped approach. By repeatedly sampling the data with replacement, we were theoretically able to replicate sampling from a larger dataset with the same statistical properties as our original data.

All in all, this project presented a variety of unique challenges. Through trial and error, collaboration with each other, and mentorship and guidance by Prof. Julia Payson and Prof. Julia Kempe, we were able to overcome most of the issues presented to us, and we believe that we finish this project with a more well-rounded set of skills than when we started.

7 References

[1] DeSilver, D., Blazina, C., Chavda, J., & Leppert, R. (2021, June 29). More U.S. locations experimenting with alternative voting systems. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/06/29/more- u-s-locations-experimenting-with-alternative-voting-systems/

[2] DeSilver, D. (2018, February 15). Q&A: The growing use of 'voter files' in studying the U.S. electorate. Pew Research Center. https://www.pewresearch.org/fact-tank/2018/02/15/voter-files-study-qa/

[3] Hajnal, Z., Kogan, V., & Markarian, G. (2022). Who Votes: City Election Timing and Voter Composition. American Political Science Review, 116(1), 374-383. doi:10.1017/S0003055421000915.

[4] Nickerson, D. W. & Rogers, T (2014, Spring). Political Campaigns and Big Data. Journal of Economic Perspectives, Volume 28, Number 2, Pages 51–74.

[5] Rentsch, A., Schaffner, B. F., & Gross, J. H. (2019, Winter). The Elusive Likely Voter. Improving Electoral Predictions with more informed vote-propensity models. Public Opinion Quarterly, Vol. 83, No. 4, pp. 782–804.

[6] United States Cities Database. https://simplemaps.com/data/us-cities.

8 Student contributions

All four of us equally contributed to writing this report and developing the Midterm Presentation and the Poster. Below there is a detailed description of other tasks each of us performed during this project.

Carolyn: Helped write and organize code to create demographic profiles from individual voter level data. Wrote code to calculate voter turnout per income bracket at the city/election level, and voter turnout per education level (college vs. no college). Also wrote code to calculate an estimate of total voting population for each city using census data. Created visualizations for voter turnout across different election types and RCV vs non-RCV (Figures 2 and 9), and mean donation amount per person per election type (Figures 28 and 29), both overall and per state.

Doma: Updated Jin's code on the cosine similarity for California to take into account low number of cities and also expanded to seven other states. Converted csv files for all eight states to parquet format. Implemented fuzzy matching code to narrow down the list of possible mismatched city names and manually reviewed them using google maps. To create voter profile, calculated average income and average donations at city, election date and election type level. Implemented Prof. Payson's suggestion on how to compute the four most recent election dates for each city. Created visualization to find average turnout for each races and average gap between white and non-white voters at election type, state and RCV vs. non-RCV level Figures (6, 11 till 23). Took Jin's baseline random forest code and added hyperparameter tuning and cross-validation on his first model, which predicted overall turnout, and used the best hyperparameter for the next two models (RCV and non-RCV) to predict overall turnout. Also, combined two feature importance plots into one (Figure 8).

Jin: Implemented cosine similarity function to find the five non-RCV cities that are most demographically similar to each RCV city. Helped resolve issue with mismatched city names between the L2 and census data files. Created visualizations to compare differences in voter income by election type and between RCV and non-RCV cities. Implemented random forest models to evaluate differences in the features that predict voter turnout between RCV and non-RCV cities.

Rodrigo: Responsible for communicating with Prof. Julia Payson, the group, and Prof. Julia Kempe, organizing the schedules, and initially structuring the files in the shared folder and the working methodology. Implemented the coding part related to "age," including all the modifications required in the data, and the visualizations per state and consolidated for each election type (Figures 3, 24, 25, 26, and 27). Also drew the Map of pooled states divided by RCV and non-RCV cities (Figure 1).

Appendix

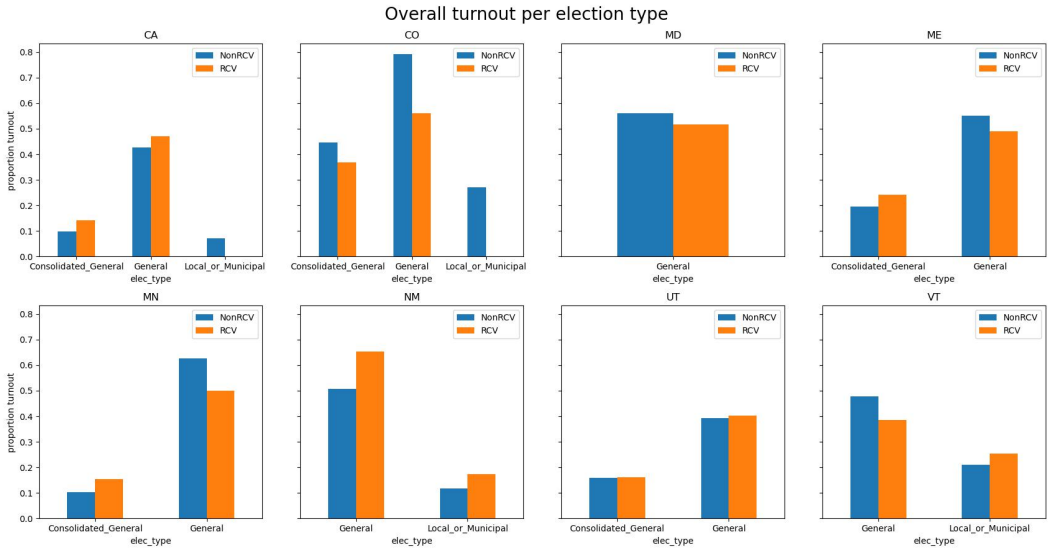

Figure 9: Voter turnout per election type by state RCV vs non-RCV

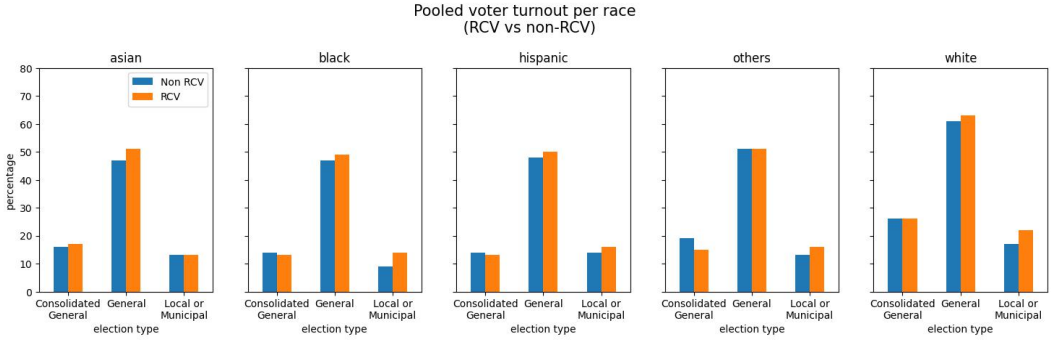

Figure 10: Average voter turnout per election type and race for RCV vs non-RCV cities

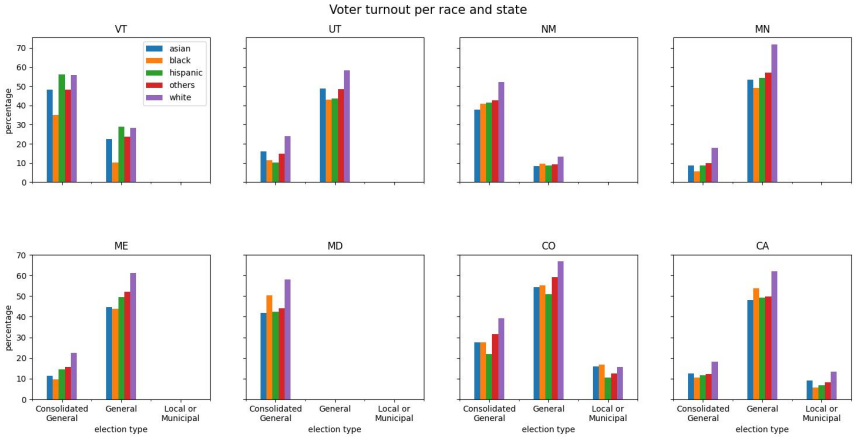

Figure 11: Average voter turnout per election type and race per state

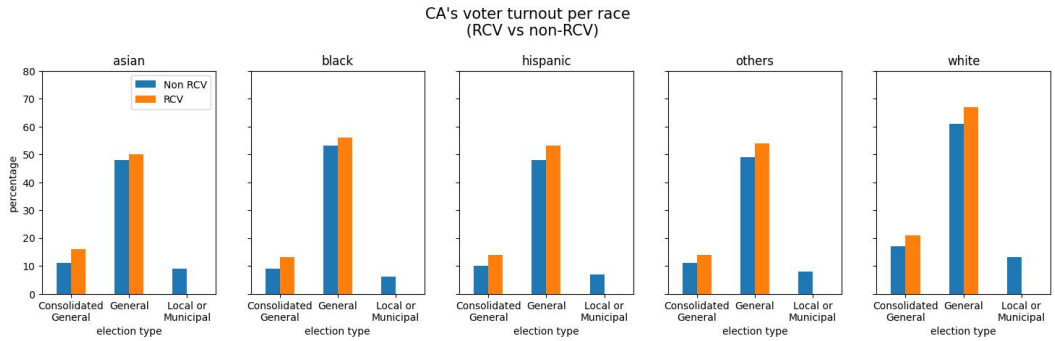

Figure 12: Voter turnout per election type for each race in California

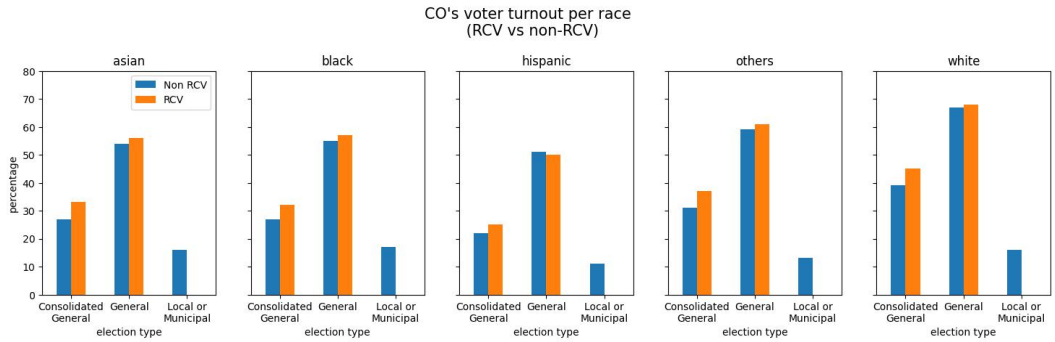

Figure 13: Voter turnout per election type for each race in Colorado

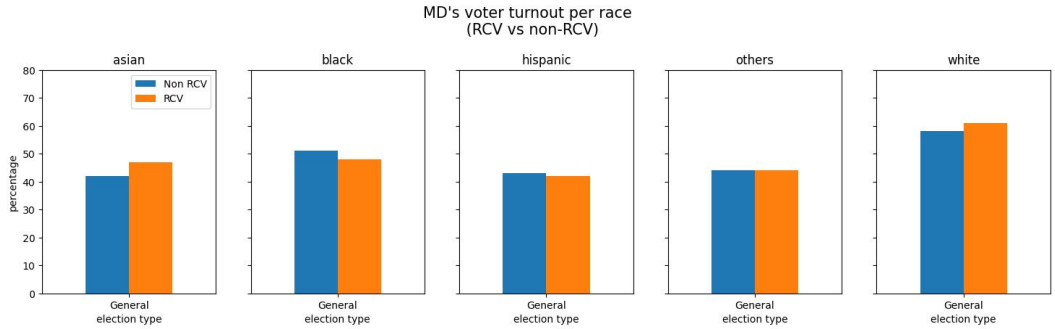

Figure 14: Voter turnout per election type for each race in Maryland

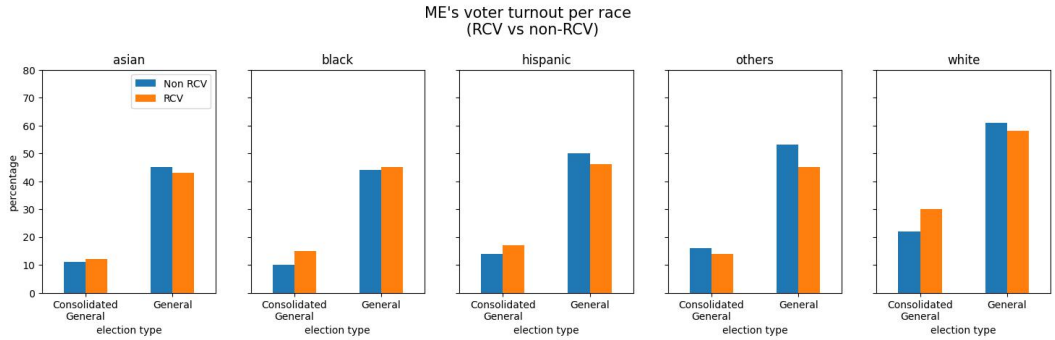

Figure 15: Voter turnout per election type for each race in Maine

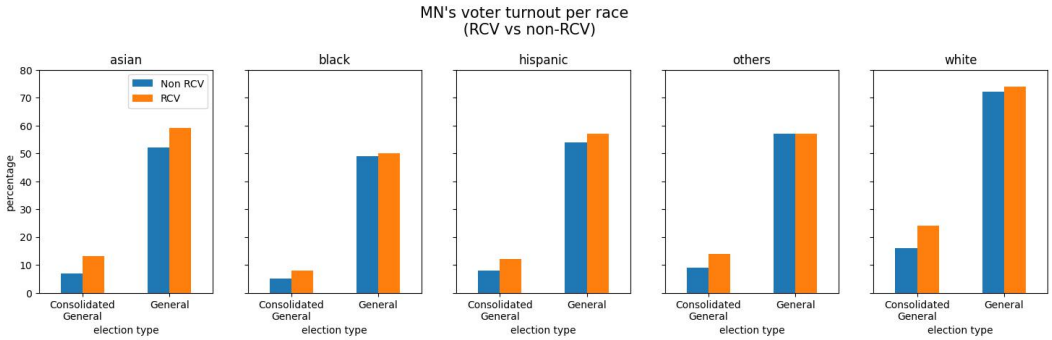

Figure 16: Voter turnout per election type for each race in Minnosota

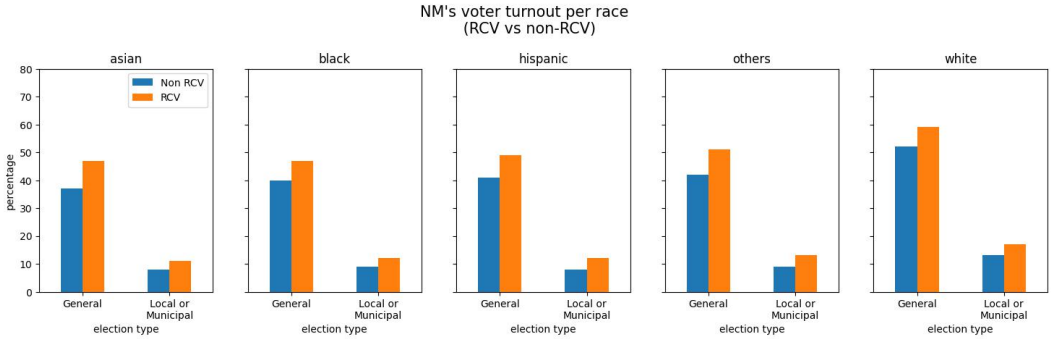

Figure 17: Voter turnout per election type for each race in New Mexico

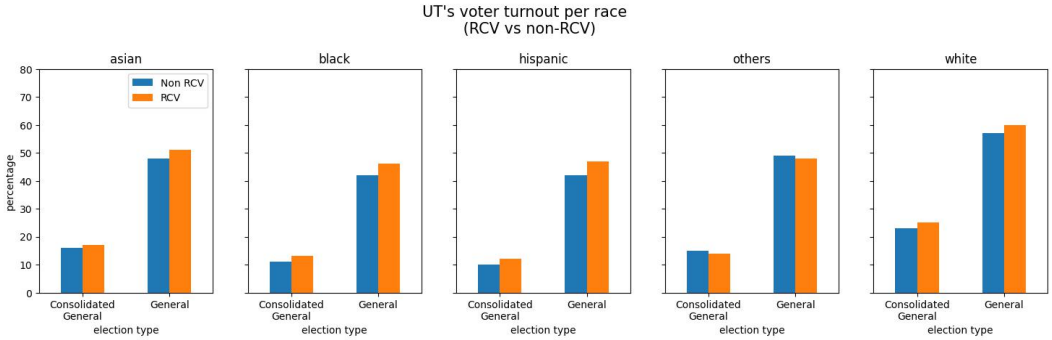

Figure 18: Voter turnout per election type for each race in Utah

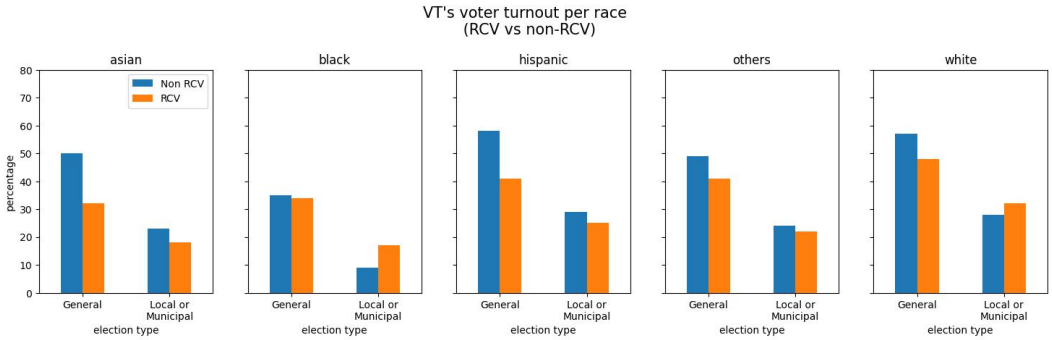

Figure 19: Voter turnout per election type for each race in Vermont

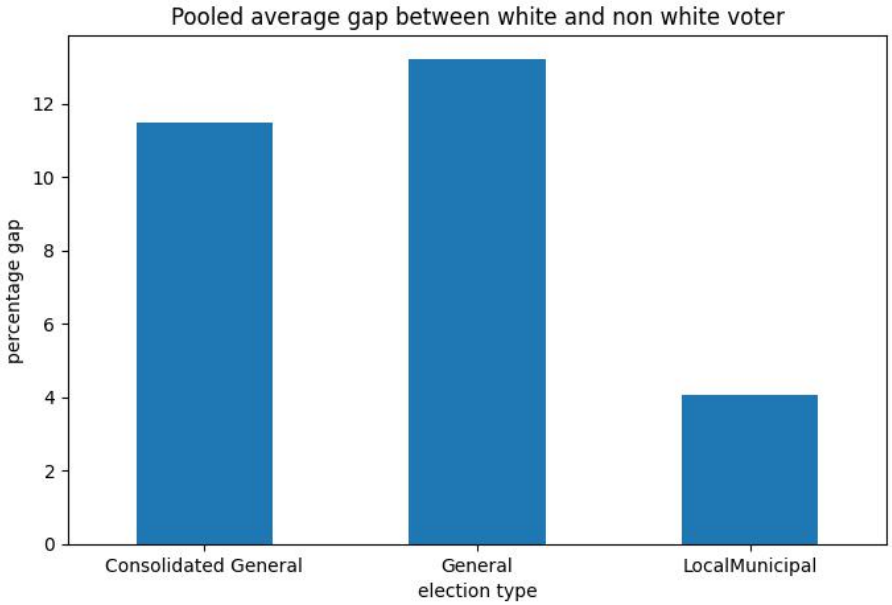

Figure 20: Average gap between voter turnout between white and non-white voters per election type

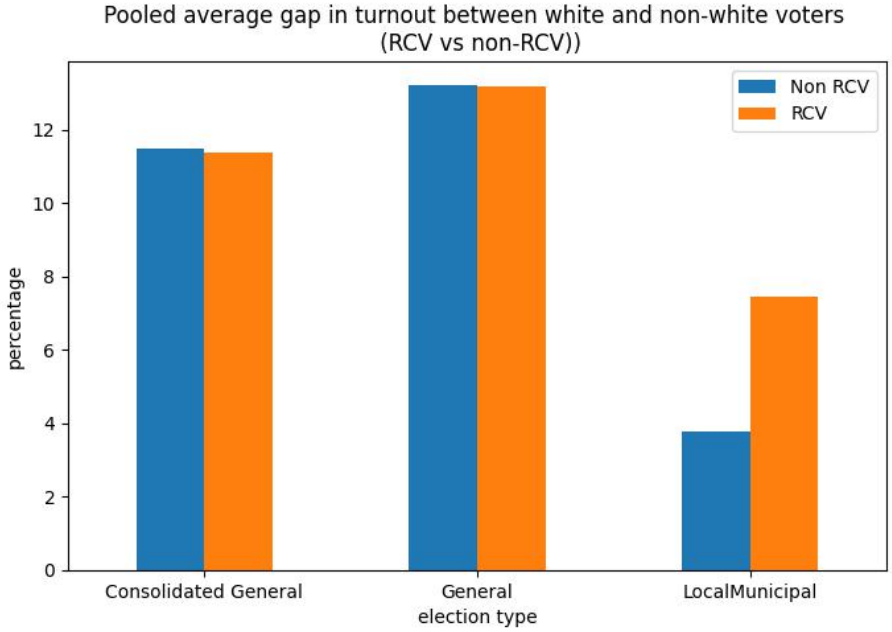

Figure 21: Average gap between voter turnout between white and non-white voters per election type for RCV vs non-RCV cities

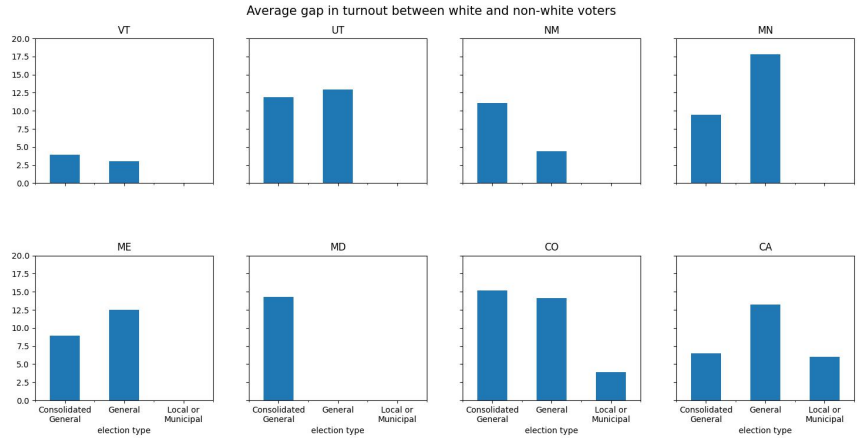

Figure 22: Average gap between voter turnout between white and non-white voters per election type and state

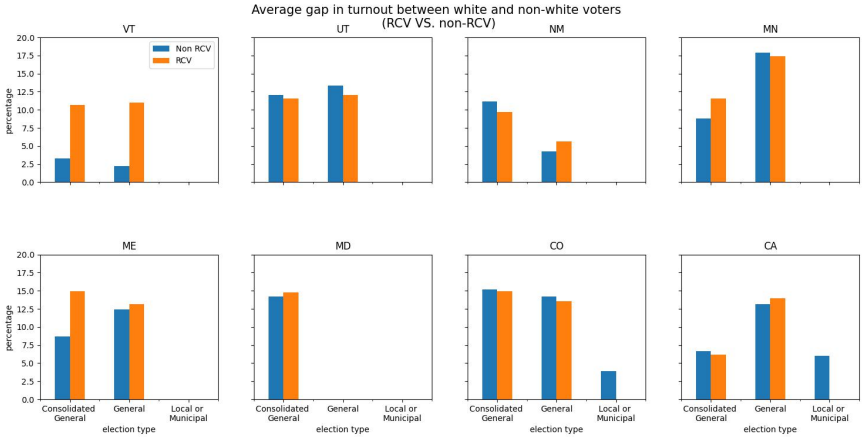

Figure 23: Average gap between voter turnout between white and non-white voters per election type and state for RCV vs non-RCV cities

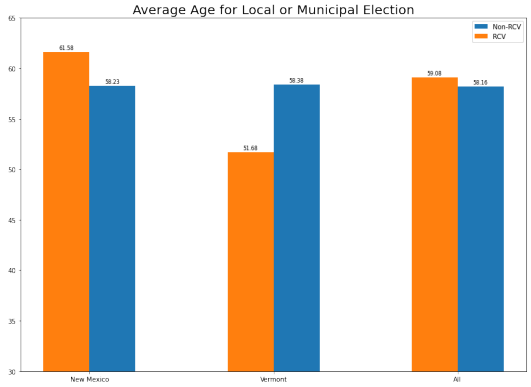

Figure 24: Average age for local or municipal election

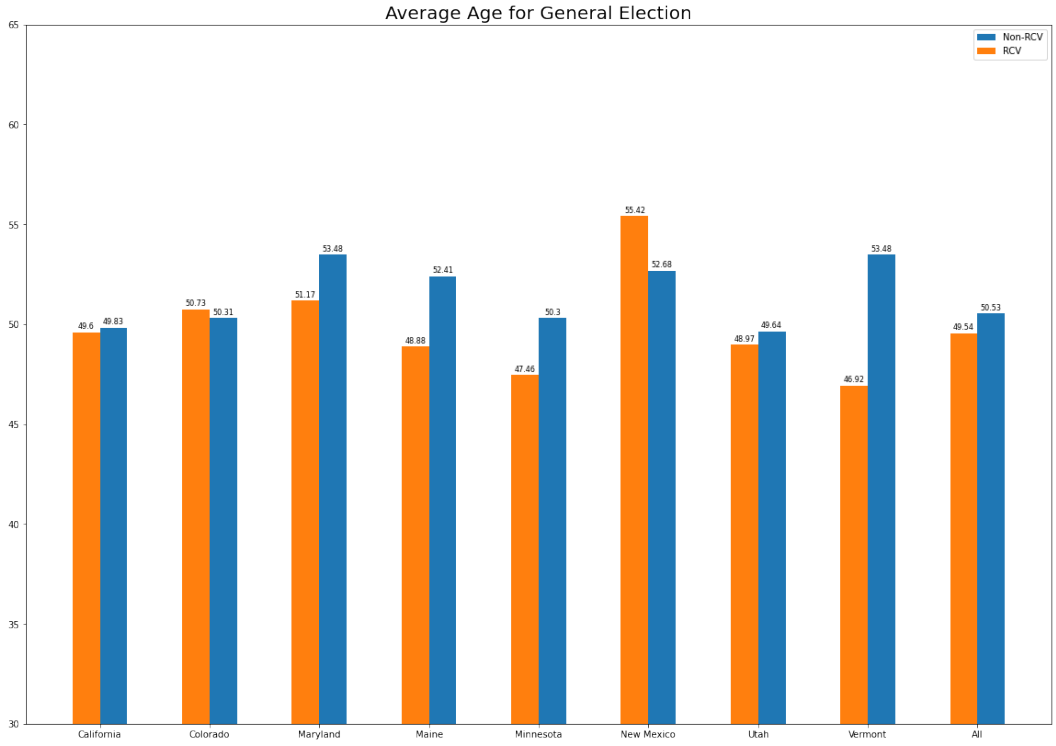

Figure 25: Average Age for General Election

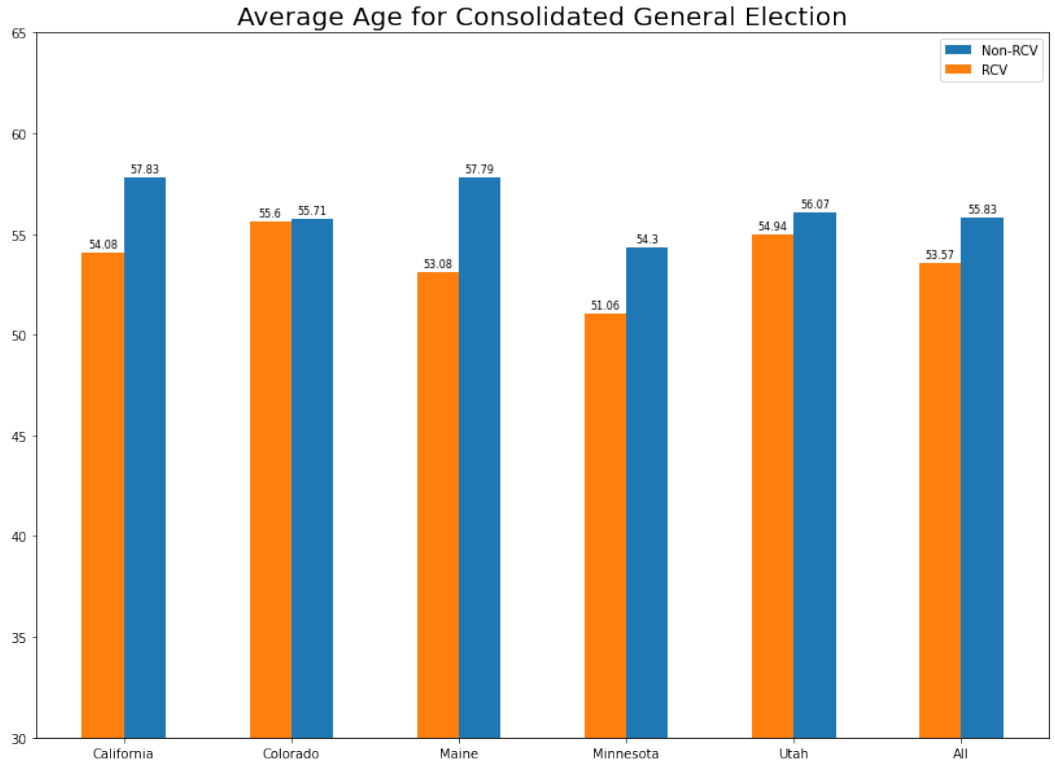

Figure 26: Average age for consolidated general Election

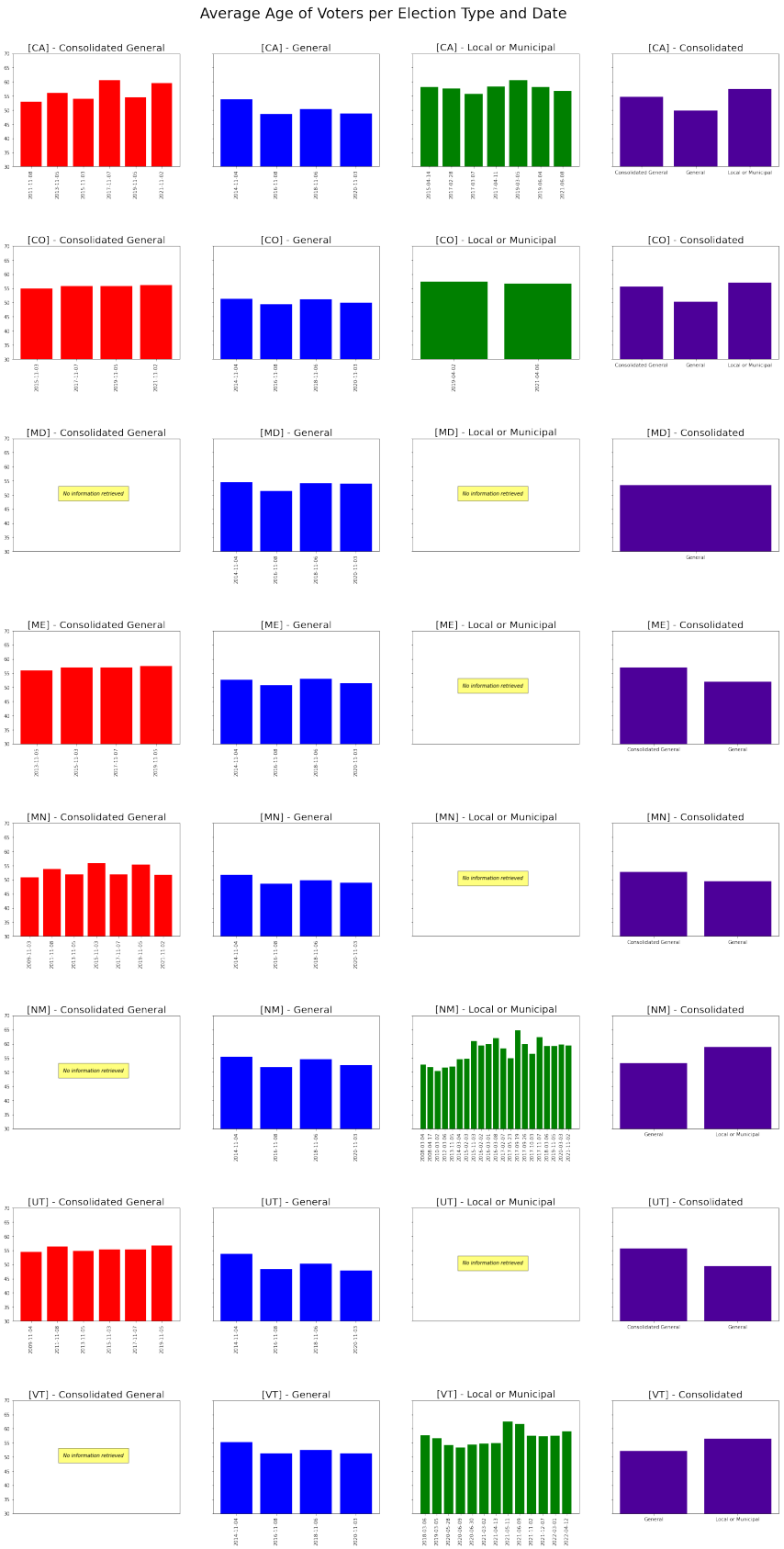

Figure 27: Average age of voters per election type and date

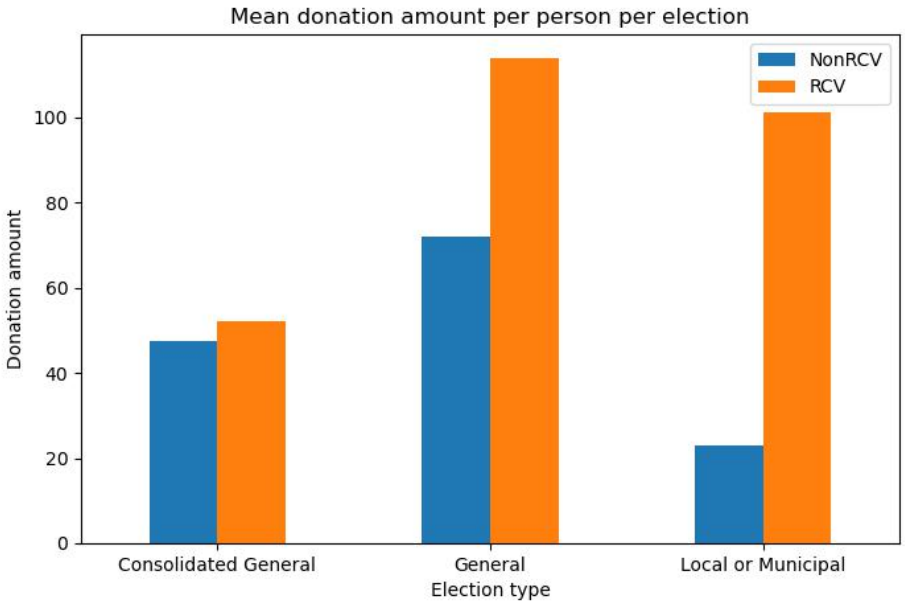

Figure 28: Average donation per election type for RCV vs non-RCV cities

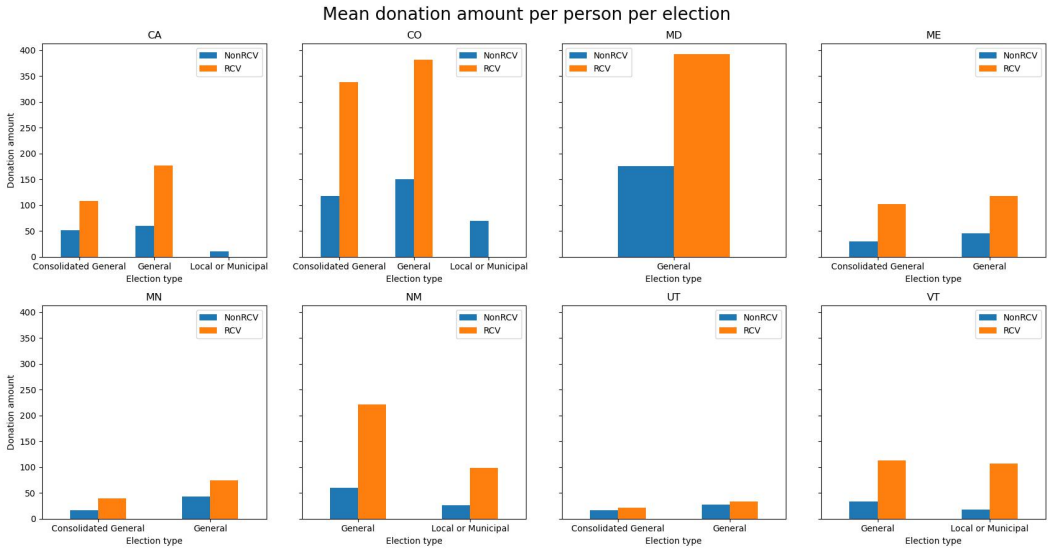

Figure 29: Average donation per election type for RCV vs non-RCV cities per state

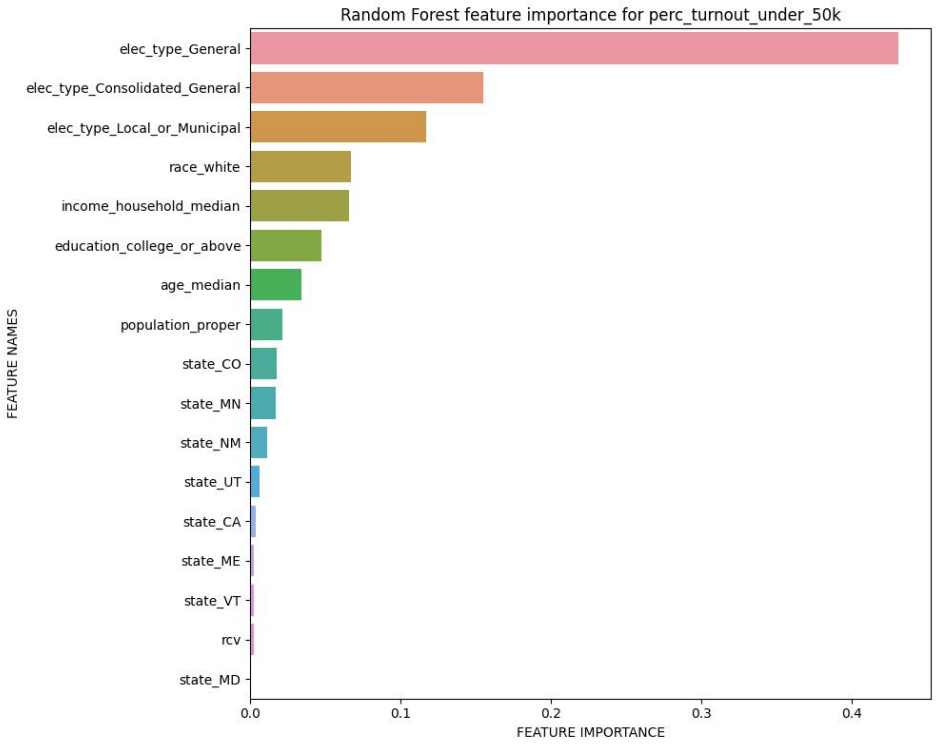

Figure 30: Feature importance scores for random forest models predicting turnout for voter with income under 50k

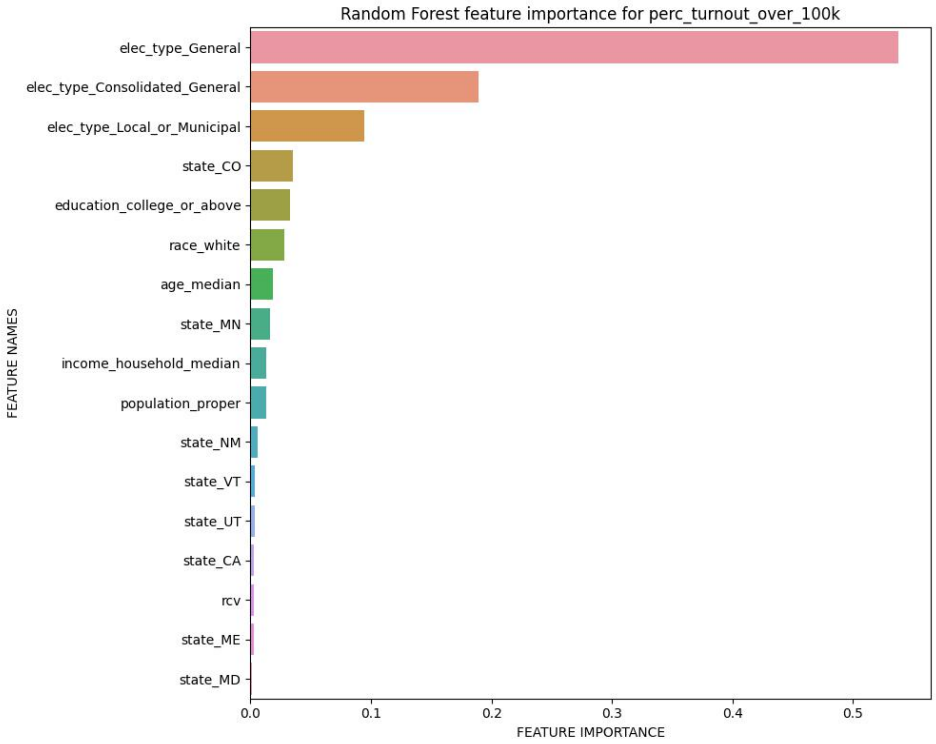

Figure 31: Feature importance scores for random forest models predicting turnout for voter with income over 100k

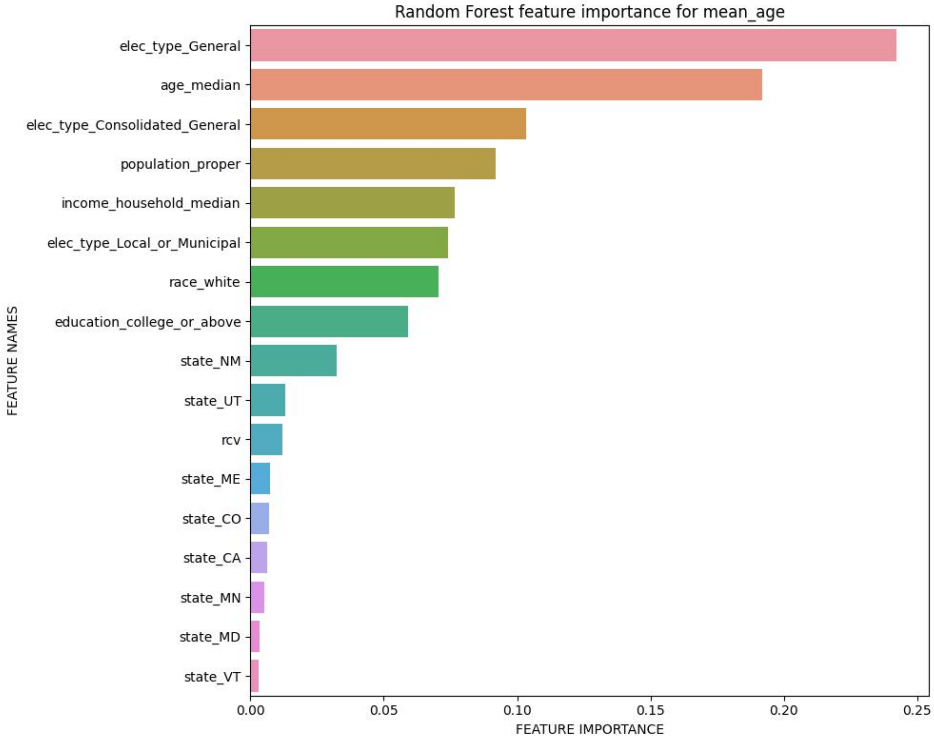

Figure 32: Feature importance scores for random forest models predicting average age of voters

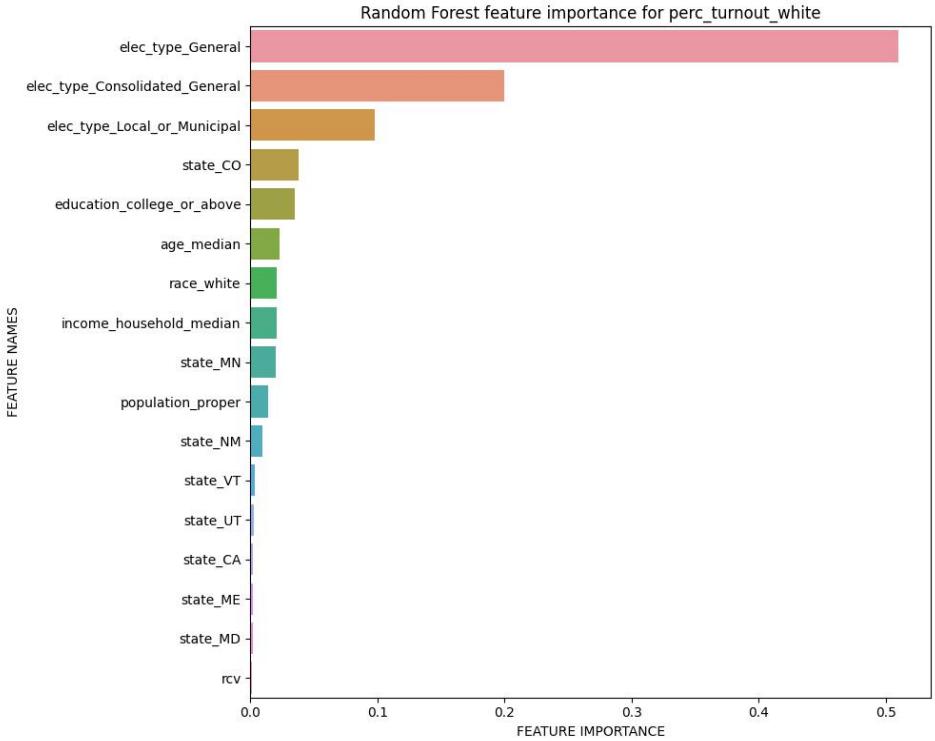

Figure 33: Feature importance scores for random forest models predicting white voter turnout