Too Much Sunshine, Not Enough Voting: Arizona in a post *Shelby* World

Final Group Project

Big Data Analytics

Elijah Castro, Madhuri Kashyap, Feynman Liu, Jason Teng

Abstract

In 2020, Arizona was one of four key swing states (alongside Georgia, Pennsylvania, and Wisconsin), that tilted a historic U.S. election in favor of Joe Biden over the incumbent president, Donald Trump. Perhaps not coincidentally, Arizona was also one of fifteen states freed from federal pre-clearance of voting law changes by a landmark Supreme Court decision in 2013, paving the way for a suite of new voting restrictions intended to limit minority voter participation. As the U.S. gears up for a likely repeat of the previous presidential election in 2024, the participation of minority voters in key swing states like Arizona will be a key determinants of election results. In this paper, we investigate the status of minority voter participation in Arizona in an attempt to understand the wider landscape of minority voter representation in modern American politics. Both for the sake of a healthy democracy in the long-term, and for the sake of understanding electoral politics in the short-term, understanding the disparity in voter participation between racial groups in the U.S. is key. By focusing on Arizona and setting our research against the backdrop of a landmark Supreme Court decision, our statistical analysis seeks to understand voter participation trends in a key swing state that could serve as a springboard towards understanding – and hopefully, alleviating – the barriers to minority voter participation that have been placed across the United States of America in the past decade.

Too Much Sunshine, Not Enough Voting: Arizona in a post *Shelby* World

Arizona, a key swing state in modern American electoral politics, presents a complex geopolitical landscape marked by its urban majority, a history of rapid suburban growth, and diverse natural features ranging from deserts to forested mountains. Economically, the state has transitioned from pastoral and agrarian bases to an industrial and technological focus; in recent years, it has been the site of much debate regarding the renewable energy transition, given that it is the US state with the highest average annual sunlight and thus theoretically a prime candidate for solar energy installations. Arizona's geography includes the Colorado Plateau, Basin and Range Province, and a unique Transition Zone, each contributing to its varied climate, demographics, and economic activities. The state's rich cultural and natural resources play a significant role in shaping its geopolitical dynamics, affecting voter turnout and engagement across different demographic variables (Byrkit et al., 2024).

Arizona is also one of fifteen states primarily affected by the landmark Supreme Court decision, *Shelby County v. Holder*, in which a 5-to-4 majority along party lines struck down the "preclearance regime" established by the Voting Rights Act (VRA) of 1965. The VRA had previously required certain states and jurisdictions to pre-clear any changes to their voting practices with the federal government, in an attempt to curb further implementation of discriminatory voting laws intended to limit the democratic participation of minority voters. In declaring a key formula used in Section 4(b) of the VRA unconstitutional due to obsolescence, the Supreme Court opened the doors to widespread changes in election laws in previously covered states, and all fifteen states, Arizona included, have taken advantage of that opportunity to pass laws restricting democratic participation, with disproportionate impact on minority communities (Bernini et al., 2023).

The post *Shelby County v. Holder* landscape has seen a myriad of restrictive laws passed to limit voter participations under the auspices of "preventing voter fraud". Such restrictions include removing or limiting access to online voter registration, Sunday voting, same day & automatic voter registration, and early voting. This restrictive legislation has been accompanied by logistical efforts to raise the bar for voting access, most notably the widespread shuttering of polling place, with over 1,000 polling places closed across the country in the first five years after the Supreme Court decision alone (Vasilogambros, 2018). Arizona's new restrictions on voting have been the subject of a more recent Supreme Court decision just a few years ago, *Brnovich v. Democratic National Committee*, in which the court upheld (again, strictly along party lines) Arizona's right to limit out-of-precinct voting and third-party ballot drop-offs. In her dissent in *Brnovich¸* Justice Elena called the ruling tragic, stating that Arizona is emblematic of how "too many states and localities are restricting access to voting in ways that will predictably deprive members of minority groups of equal access to the ballot box" (Garcia, 2024). Indeed, legal historian Mary Frances Berry noted that "Justice Kagan's dissent… makes clear that what the majority sees as inconveniences—such as transportation—are major burdens on the opportunity to choose a candidate of their choice for many voters, especially Native Americans and Latinos in Arizona" (Garcia, 2024).

Arizona is a microcosm for a wider movement of minority voter disenfranchisement in the wake of a landmark court decision that has reshaped the landscape of American politics over the last decade. Understanding the real influences of flagging minority voter participation in the U.S. and the phenomena's relationship to recent legal developments is crucial in crafting strategies to reverse this trend of minority voter disenfranchisement and build towards the high levels of voter participation rates enjoyed by other peer democracies necessary to sustain a

healthy democracy in the long-term. In this paper, we explore the landscape of voter participation in Arizona today. We investigate the relationships between race, gender, age, economic status, political affiliation, and voter behavior, and we demonstrate that the voter base of Arizona is disproportionately under-representing low-income and minority voters.

Data selection and handling

Data often starts messy, and this data set is no different. We wish to look at how we can use our data most effectively which will allow us to perform unbiased and more accurate analysis. Below are steps that we have taken to ensure that we maintain high data integrity before we dive into data exploration.

Data description. The data used in this project utilizes the VM2 Uniform Dataset from L2 Voter Files made available by University of Pennsylvania. It contains extensive demographic and voting history details of registered voters in the United States.

Variable selection. Below is our first initial selection on the range of variables we should consider. There were 726 possible selections and our team narrowed it down to 24 predictors with 1 key. The variables were selected to capture a wide range of demographic information

Demographic Information:
- LALVOTERID
- Voters_Gender
- Voters_Age
- Ethnic_Description
- Voters_PlaceOfBirth

Geographic Information:
- Voters_FIPS
- CountyEthnic_LALEthnicCode
- CountyEthnic_Description

Economic Information:
- Mailing_Families_HHCount
- CommercialData_Education
- CommercialData_EstimatedHHIncome
- CommercialData_EstimatedAreaMediaHHIncome
- CommercialData_AreaPcntHHMarriedCoupleWithChild
- CommercialData_AreaPcntHHSpanishSpeaking
- CommercialData_AreaMedianEducationYears
- CommercialData_PoliticalContributerInHome
- CommercialDataLL_Hispanic_Country_Origin

Voting Behavior:
- Vote_By_Mail_Area

Election Results:
- ElectionReturns_G12_Cnty_Vote_Obama_D
- ElectionReturns_G12_Cnty_Vote_Romney_R
- ElectionReturns_G16_Cnty_Vote_Clinton_D
- ElectionReturns_G16_Cnty_Vote_Trump_R

Political Information:
- Parties_Description
- Voters_VotingPerformanceEvenYearGeneral

Figure 1. Variables selected for analysis

while avoiding highly correlated information to reduce possibility of overfitting in our machine learning models later in this report.

Data extraction. The data itself is stored as parquet files in a designated Google Cloud bucket. Utilizing Google Cloud Shell, we pulled the data from this repository, targeting specifically Arizona state data for further analysis. Our team utilizes Google Cloud's Dataproc clusters to take advantage of its high scalability functions for large data and its seamless integration with Apache Spark. This allowed us to leverage advanced analytics and machine learning capabilities for big data.

Missing data - removal. We found that some of the selected columns have over 70% of the data missing and drop these. These columns include:

- Voters_PlaceOfBirth
- CountyEthnic_LALEthnicCode
- CountyEthnic_Description
- CommericalDataLL_Hispanic_Country_Origin
- Vote_By_Mail_Area
- CommercialData_PoliticalContributerInHome
- CommercialData_Education

The last variable of education is an important demographic variable but is unfortunately missing 46% of its observations, making the usage of imputation methods rather difficult.

Additionally, we implemented a data cleansing step where we excluded rows containing missing observations in four specific categories:

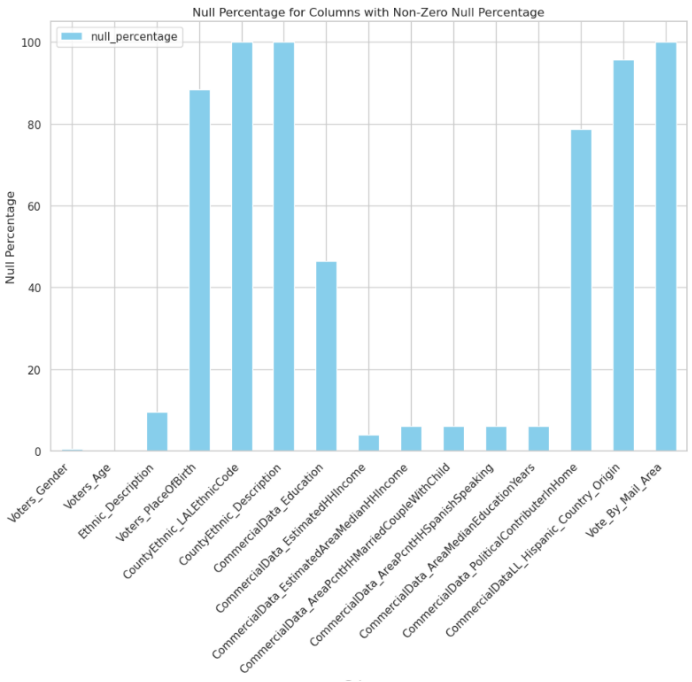

Figure 2. Graph representing percentage of null items in each selected predictor

- CommercialData_EstimatedAreaMedianHHIncome
- CommercialData_AreaPcntHHMarriedCoupleWithChild
- CommercialData_AreaPcntHHSpanishSpeaking
- CommercialData_AreaMedianEducationYears

This decision was motivated by the observation that missing data in any one of these categories corresponded to missing data across all four areas for a given observation. Recognizing the limited utility of such incomplete records in accurately characterizing the voter demographics, we opted to remove these rows entirely. By eliminating instances with significant data gaps, we ensure that our machine learning models are trained on a more robust and informative dataset. This targeted approach enhances the quality of our analyses by focusing exclusively on data points with substantial information, thereby improving the predictive capabilities of our models.

Missing data – imputation. We imputed the variables of age, gender, and estimated household income. Age was the only continuous variable and was imputed using the median (Fig. 3), while gender and household income were imputed by maintaining the proportions of each category.

We adopted an approach where variables were implicitly drawn to randomly sample from the empirical distribution generated from columns with missing variables. This method enabled us to maintain the original distribution of each respective category, thereby preserving the integrity of the data. Additionally, given sufficient resources, we could enhance our analysis by running multiple models to determine the bootstrap average of each unique model. This iterative process would further refine our understanding and improve the reliability of our findings.

For handling the variable representing ethnic descriptions, we encountered 82 distinct categorical indicators. To mitigate the impact of missing data, which constituted only a small fraction of our dataset (less than 10%), we opted to categorize these missing observations as a unique category labeled "None." This approach allowed us to retain the entirety of our dataset while ensuring that our model remains robust and effective. By treating missing values as a separate category, we enable our model to glean insights from the available 90% of the data. This strategy prevents the removal of potentially valuable information and encourages the model to discern patterns within the ethnic descriptions variable, even in the absence of certain observations. Furthermore, this

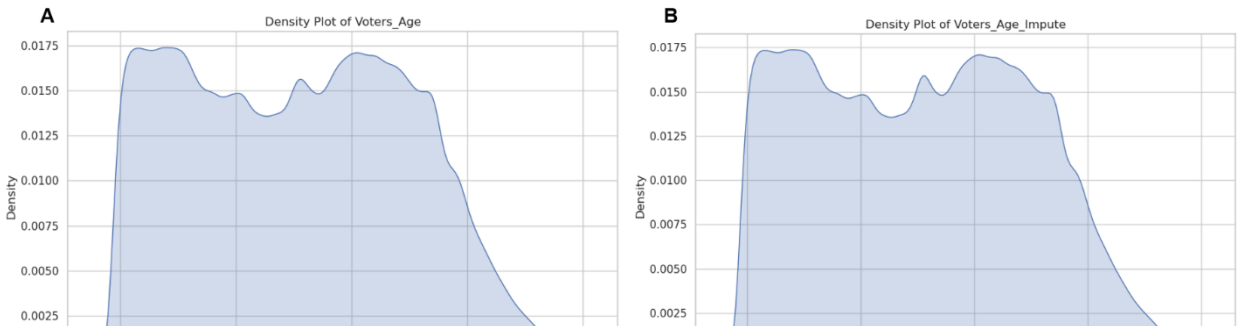

Figure 3. Density plots of voters' age (panel A) and imputed voters' age (panel B)

technique served to regularize the model, promoting better generalization and resilience against overfitting, particularly in large-data settings where such concerns may arise.

Exploratory data analysis

We begin our analysis of voting patterns in Arizona by overviewing the landscape of registered voters, starting with demographic information. In general, Arizona voters are slightly more female than male, overwhelming White or Hispanic, and predominantly middle-class.

We have self-report data on the ethnicity of registered voters in Arizona (Fig. 4), though nearly 10% of voters declined to report this information (represented here as 'None'). Arizona is predominantly European-American at 64%, and slightly more so than the national average of "White alone, not Hispanic or Latino" at 59% (US Census Bureau, 2023). Arizona voters are slightly more likely to be Hispanic than the national average (21% vs. 19% nationally), and all other ethnicities are under-represented relative to the national average. Notably, African American voters comprise only 1.4% of registered voters in Arizona, as opposed to 14% nationally, and Asian voters comprise only 2.3% of the voter base as opposed

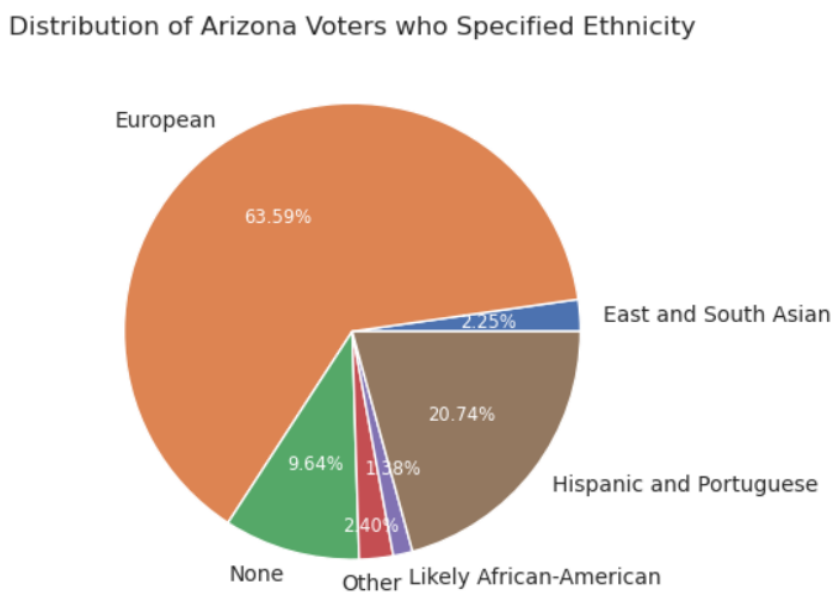

Figure 4. The ethnic background of Arizona voters. Almost two-thirds of registered voters self-report European-American.

to 6.3% nationally. Thus, Arizona voters are overwhelmingly likely to be either White or Hispanic, and that fact is reflecting in the political messaging of the two most prominent political parties, which have tailored their messages to either the White majority, the Hispanic majority-minority, or both, to the relative occlusion of other voters.

The only notable feature of the gender breakdown of Arizona voters (Fig. 5) is the over-representation of female voters relative to the national average: female voters comprise ~53% of the Arizona voter base, whereas only 50.4% of U.S. citizens are female (US Census Bureau, 2024).

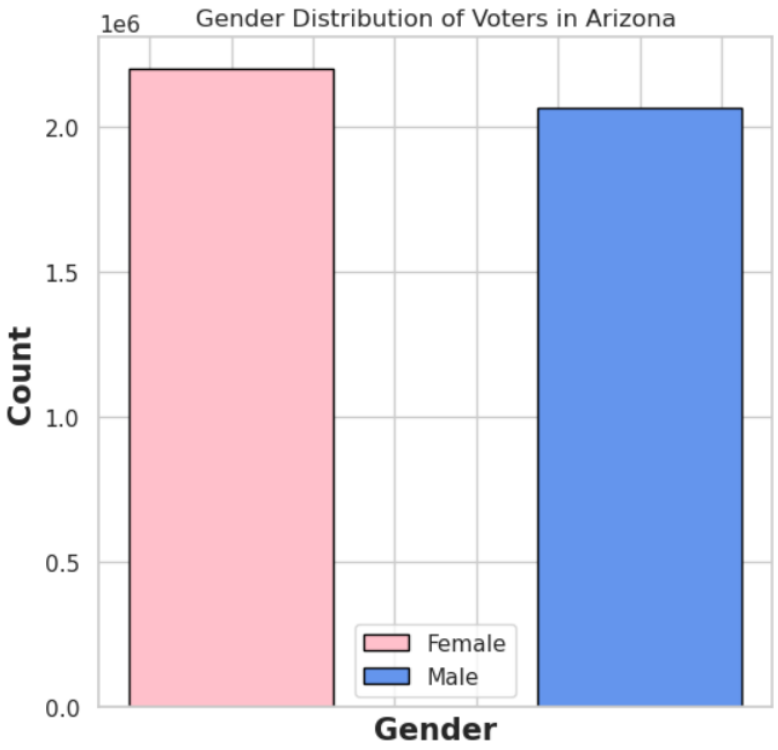

Figure 5. The gender background of Arizona voters: female voters are slightly over-represented relative to the national average (53% of Arizona voters vs. 50.4% of US citizens).

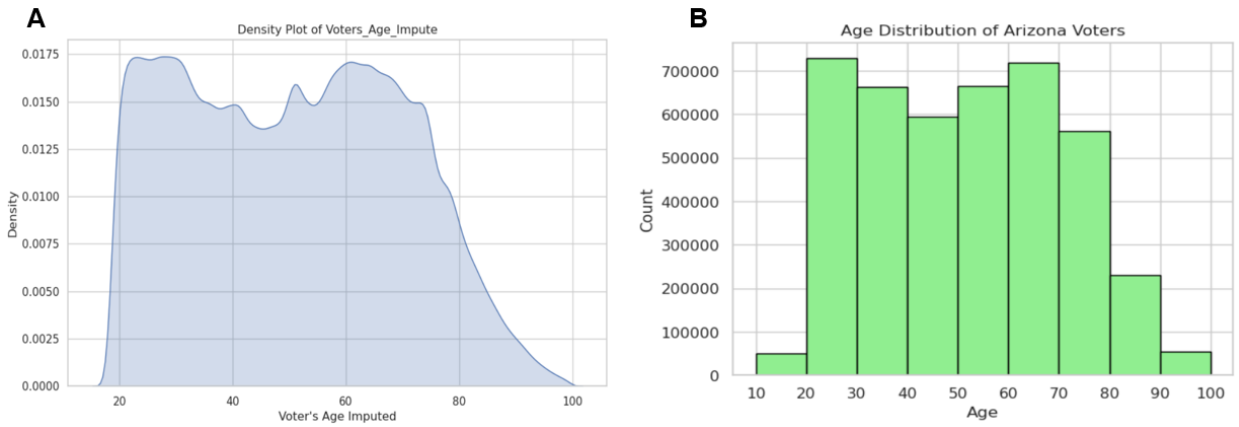

Figure 6. Two views into the age breakdown of registered voters in Arizona. The shape is slight bi-modal and reflects the under-participation of young voters in Arizona generally as the left-hand peak would be ~20% higher than the right peak if all citizens voted at equal rates.

Now we turn to a breakdown of the age of registered Arizona voters. As you can see in Figure 6, the density curve of age has a slightly bi-modal distribution, with population peaking at 20-25 years and again at 55-65 years. This is, in fact, in line the bimodal distribution of age ("population pyramid") for the U.S. as a whole and is likely attributable to many complicated macro-economic factors and does not reflect Arizonan idiosyncrasies. As a digression, top theories for the current U.S. bimodal age distribution – which stands in contrast to the wide-tailed bell curve centered at 40 that is the European Union's age distribution – include depressed fertility rates during the economically-troubled 1970's, as well as new trends in age-of-first-birth fertility curves showing bimodal peaks for women at age 19 and again at 29, potentially contributing to the non-normal shape seen in both U.S. and Arizona voter age curves. Overall, relative to the national average, voters aged 20-30 are the most under-represented relative to the national average. Amongst Arizona voters, there is a roughly 1:1 ratio between voters in the third decade of their life as in their seventh, whereas that ratio is nearly 1.25:1 across all U.S. citizens. This likely represents the under-participation of young voters in American electoral politics that has been reported nationally for some time (U.S. Census Bureau, 2024).

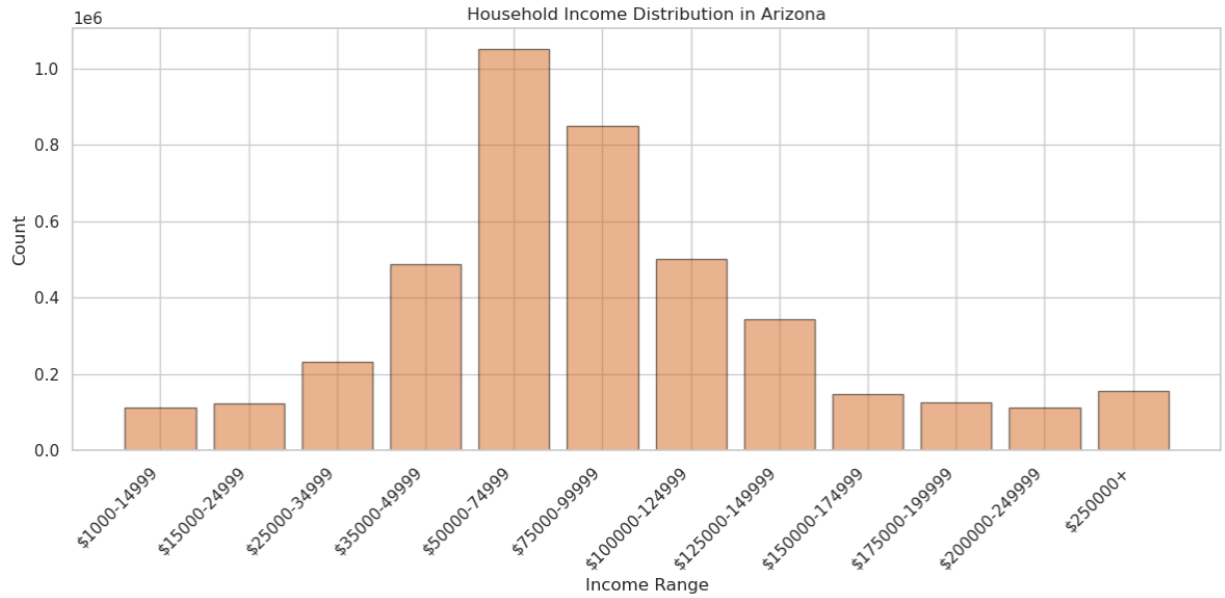

Figure 7. Household income distribution for registered voters in Arizona.

Household income data for registered Arizona voters reflects the right-tailed distribution that one expects for income data as a whole (Fig. 7). The median household income for registered Arizona voters is roughly equivalent to the national average of $75,000, and reflects a voter base that is predominantly middle-class, as is the case the American electorate as a whole (U.S. Census Bureau, 2024).

Results

Having cleaned and prepared our data and having canvassed the landscape of registered voters in Arizona, we then turn to machine learning to uncover patterns in our database between voter background and voter behavior. Specifically, we attempted to answer two questions: given a voter's background information, can we predict whether or not they will vote in an upcoming

election? And can we predict who they will vote for? To this end, we trained and optimized two machine learning models – a Random Forest model and a Logistic Regression model – on each of our two questions, so that we had a comparison case and so that we could be reasonably sure we were modeling our parameter space properly. Random Forest models are known for their flexibility, resistance to overfitting, and predictive power in a wide variety of contexts, though that comes at the expense of expensive computation and opaque interpretability inherent in an ensemble model. A Logistic Regression model, on the other hand, is inexpensive to train and its output is easily interpreted in this context ("how likely is this voter to vote for X candidate?"), though it suffers from potentially shaky assumptions of non-collinearity between predictors and an inability to represent more complex, non-linear relationships in the data.

Having already settled on a set of predictors as described above, for both models, we undertook a hyperparameter grid search to optimize predictive power. For our Random Forest Classifier model (hereto referred to as "RF"), we tested the most impactful hyperparameters: number of trees, max depth per decision tree, and node impurity measure, Gini vs. entropy (due to limited computational resources, we had to keep our grid search relatively constrained). Using standard 5-fold cross validation on a random 80%-20% train/test data split, we arrived at an optimal model up 50 trees, max depth of 20, and a Gini impurity measure. Notably, our test-set accuracy for the RF was 96%, up from 66% using default parameters. For our Logistic Regression model, our hyperparameter grid search using a similar cross-validation approach arrived at optimal parameters of 100 max iterations, a regularization parameter of 0.001, and an elastic net regularization parameter of 1.0 (thus an L1 penalty, also known as Lasso regularization). These optimal parameters improved test-set accuracy from 62% to 89%. Indeed,

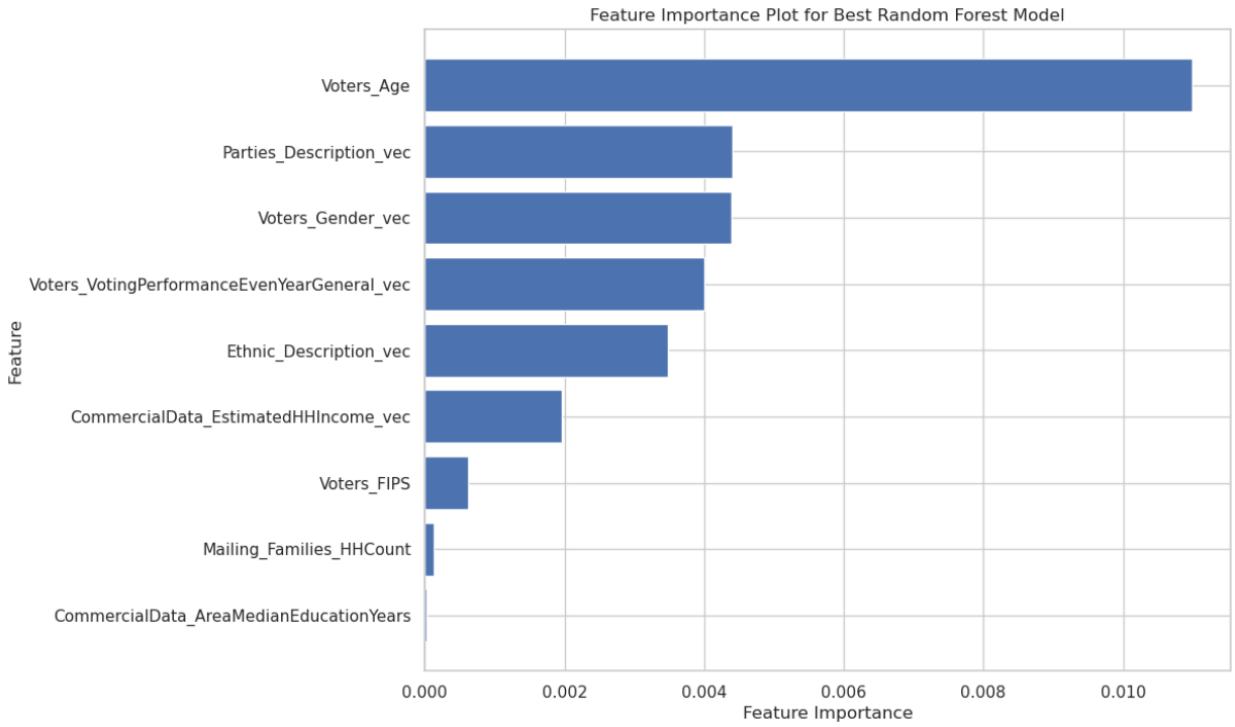

Figure 8. Feature importance scores for a Random Forest Classifier predicting voter turnout (i.e., whether or not an individual voter will vote).

across both our questions, a similar hyperparameter optimization arrived at the same hyperparameter's for each model both times.

We began by trying to predict, for each voter, whether they voted at all in the 2016 election, the most recent in our dataset. After tuning and optimizing both modes, we generated feature importance scores for both, using a Gini impurity measure for the RF model and F-scores for the Logistic Regression. While the Logistic Regression model achieved a commendable accuracy of approximately 89%/94% in predicting 2016 U.S. election voting behavior and voter turnout respectively, the RF model ultimately outperformed it with accuracies of roughly 94%/96.2% respectively, indicating a more robust model in terms of prediction for this particular election result. The higher accuracy of the Random Forest model suggests that its approach to handling the complexity and interactions between features is more effective for this dataset, and thus became our primary focus.

As you can see above, voter age was by far the most significant predictor (meaning, had the highest average impact of the model's final prediction) on whether an individual voter was likely to vote in 2016. This reflects the truism of the "disaffected youth" in American electoral politics and is likely downstream from the same trend in young voter under-participation that we saw in the under-representation of young voters in the demographic information we overviewed above. The voter's political party and gender were the next most significant predictors: these factors might be capturing the impact of party loyalty and gender-specific political mobilization on voter engagement. The next most significant predictor is a measure of past voting behavior, indicating past voting is indeed predictive of future turnout and perhaps reflecting a sense of civic duty present in some Arizona voters. Our fifth-most significant predictor is ethnic description, indicating that ethnic identity can have a bearing on electoral participation, potentially due to community-level mobilization efforts or shared concerns that drive voter turnout within certain ethnic groups. Our final predictor with notable model impact was household income, indicating that the basic logistical impact of increased affluence does indeed increase a voter's likelihood to participate in the democratic process. Our remaining predictors, including county-code, family size, and county-level educational data, did not have much impact on the model's final predictions. Overall, these feature significance measures support the conclusion that ethnicity and economic reality do have a tangible impact on a registered voter's likelihood to vote in Arizona, but also indicate that other relevant demographic info such as age and gender may be more important determining factors overall.

Having shed light on our primary question of the role of demographics on voter turnout overall, we decided to see if our tuned models could help predict voter *behavior* as well. Specifically, we tried to see how well we could predict whether a voted would vote for Hillary

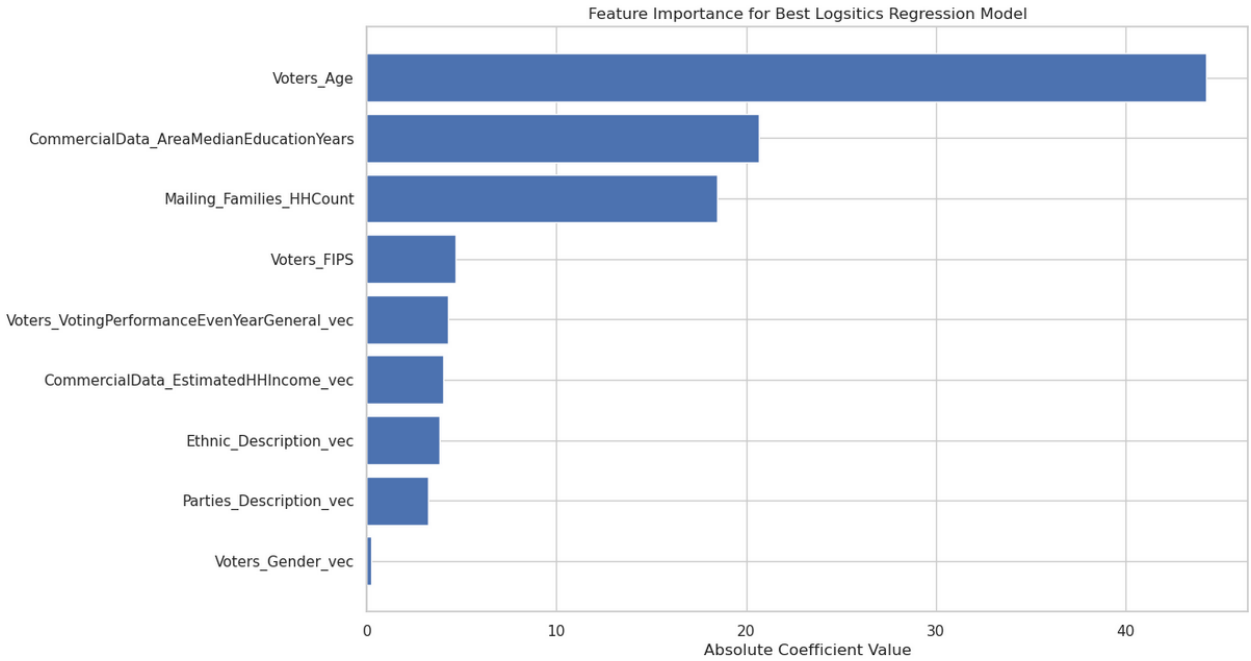

Figure 9. Feature importance scores for a Logistic Regression model predicting **voter turnout** (i.e., whether or not an individual voter will vote).

Clinton in the 2016 presidential election (stylized in our dataset as "ElectionReturns_G16_Cnty_Vote_Clinton_D"). Notably, both the overall accuracy and feature importance measures were almost identical to that of our voter-turnout RF Classifier model. While this could reflect similar underlying trends in both voter engagement and party preference, it could also indicate an inherent rigidity in our RF model, which could potentially struggle with true out-of-sample data (e.g. 2020 election predictions) as a result).

We then turned to our Logistic Regression model, to understand if the slightly less accurate model picked up similar patterns in the data as the RF model. As you can see below, the Logistic Regression model did differ in its identifiably important features both between dependent variables (DV's, voter turnout vs. behavior), as well as from our RF model. Notably, for both our DV's, voter age remains by far the most significant predictor, as was the case with our RF model. Unlike the Random Forest, however, the Logistic Regression picks up family head count as a significant predictor for both DV's. For voter turnout, county-level education

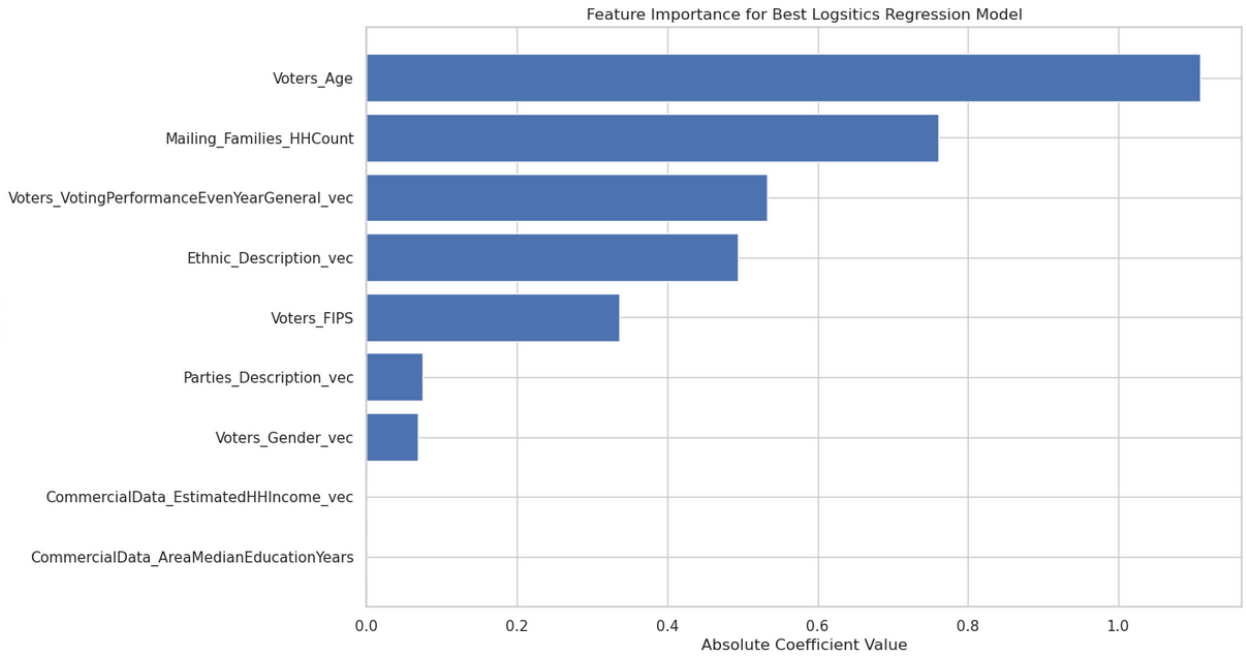

Figure 10. Feature importance scores for a Logistic Regression model predicting **voter behavior** (i.e., whether an individual voter will vote for Hillary Clinton).

data is the only other tangible predictor. For voting behavior, ethnicity, FIPS code, and prior voting behavior all became significant predictors. Further investigation and applications to wider datasets would be necessary to ascertain the robustness of these results: while some takeaways, such as gender being a motivating factor for *who* one votes for but not *whether* one votes, may jive with political scientists' intuition, but other takeaways, such as the predominance placed on household headcount, may not.

Discussion

In this report, we embarked on an exploratory journey to unravel the intricacies of minority voter participation within Arizona, against the backdrop of pivotal legal shifts that have reshaped the electoral landscape in the United States. Through meticulous data analysis and the application of advanced statistical methodologies, we dissected the dynamics at play in the state's

political arena, focusing on the ramifications of the Shelby County v. Holder decision on voting rights and access.

Our findings underscore a significant disenfranchisement of minority voters, compounded by legislative changes that have erected barriers to their participation in the democratic process. The introduction of restrictive voting laws under the guise of combating voter fraud has disproportionately impacted these communities, casting a shadow over the inclusivity and fairness of Arizona's electoral system. This analysis, while specific to Arizona, echoes a broader national concern about the erosion of voting rights and the imperative need for reforms to safeguard the bedrock principle of democratic participation.

Furthermore, our results reveal the profound influence of demographic factors on voting patterns, with age emerging as a pivotal determinant of electoral engagement. The underrepresentation of young voters, coupled with the disparities in participation among different ethnic and income groups, paints a picture of a fractured electoral landscape that fails to fully capture the diverse tapestry of the state's populace.

As we look ahead to future elections, it is imperative that policymakers, activists, and the broader community engage in a concerted effort to address these challenges. The restoration of robust federal oversight, the enactment of policies that enhance access to the ballot box, and the mobilization of grassroots efforts to encourage participation among disenfranchised groups are crucial steps towards reinvigorating the democratic spirit of Arizona and, by extension, the nation.

In conclusion, these results not only shed light on the current state of electoral participation in Arizona but also serve as a call to action. It underscores the urgent need for a

recommitment to the principles of equality, inclusivity, and justice in our electoral processes. As we move forward, let this paper be a catalyst for change, inspiring efforts to dismantle the barriers to voting and ensuring that every voice is heard and valued in the democratic dialogue that shapes our society.

References

Ang, D. (2019). Do 40-year-old facts still matter? Long-run effects of federal oversight under the Voting Rights Act. *American Economic Journal: Applied Economics*, *11*(3), 1-53.

Bernini, A., Facchini, G., & Testa, C. (2018). Race, representation and local governments in the US South: the effect of the Voting Rights Act. *Journal of Political Economy, 131* (4).

Burkimsher, M. (2017). Evolution of the shape of the fertility curve: Why might some countries develop a bimodal curve? *Demographic Research*, *37*, 295-324.

Bykrit, J. W., Hecht, M., McNamee, L. (2024, March 15). *Arizona.* Britannica. https://www.britannica.com/place/Arizona-state

Garcia, K. (2021, July 2). *Supreme Court decision rules Arizona's laws constitutional.* Penn Today. https://penntoday.upenn.edu/news/supreme-court-arizona-voting

United States Census Bureau. (2023, July 1). *QuickFacts.* https://www.census.gov/quickfacts/fact/table/US/PST045223

United States Commission on Civil Rights. (2018, July). *Voting Rights in Arizona.* https://www.usccr.gov/files/pubs/2018/07-25-AZ-Voting-Rights.pdf

Vasilogambros, M. (2018, September 4). *Polling Places Remain a Target Ahead of November Elections.* Pew. https://archive.ph/20200503150736/https://www.pewtrusts.org/de/research-and-analysis/blogs/stateline/2018/09/04/polling-places-remain-a-target-ahead-of-november-elections