Vote Preference by Charles English. November 6, 2018

Introduction to Vote Preference

Welcome to Vote Preference, a web app that predicts the outcome of state-wide ballot measures.

Vote Preference uses machine learning models trained on the results of past ballot initiatives to predict the outcome of future votes. Specifically, county-level results are almost always available from Secretaries of State websites from across the country. The goal is to predict the percentage of yes votes given a ballot question. Care is taken to standardize questions across state lines and questions are selected to ensure consistency for similar questions across state lines. For example, increasing the minimum wage from $7 to $8 in one state is quite similar to an increase of $7.25 to $8.25 in another state. In contrast, adding a few slot machines to a race track is not the same as opening 3 casinos in 3 locations in the state. Models are trained using basic information/features describing the individual counties. These features are divided into political, economic, racial, and social indicators as listed below:

Political:

percentage of votes Clinton received in excess of Trump in 2016[1]

Economic:

percent of uninsured people[2]
per capita income[2]
unemployment rate[2]
income inequality (calculated as: 80th percentile income/20th percentile income) [2]
percent with some college education[2]
poverty rate[2]

Racial:

percentage of Whites[3]
percentage of African-Americans[3]
percentage of Hispanics[3]
percentage of Asians[3]

Social:

percentage of Seniors[2]

percent US Citizens[2]
percent rural population[2]

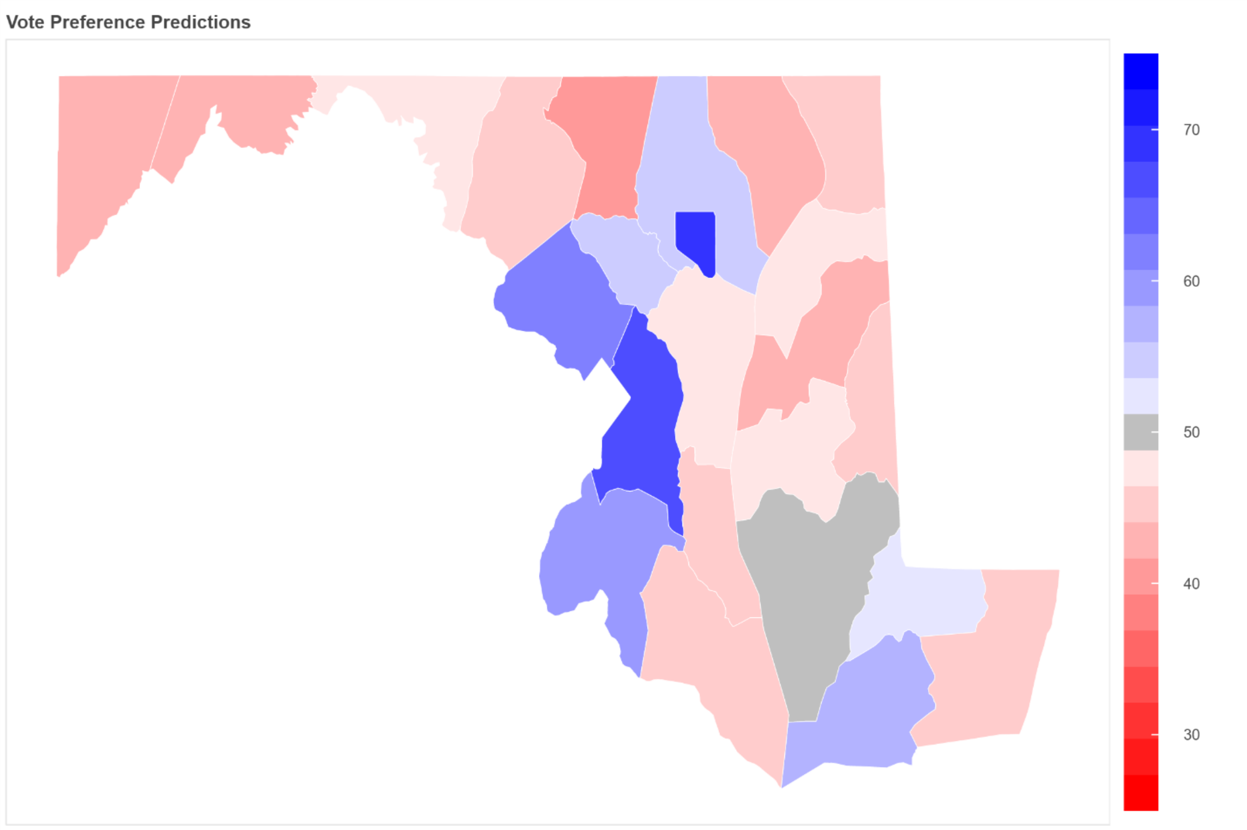
Data sources:

1. From secretaries of states websites, 1 website per state
2. From https://datausa.io/map/, for the latest year available (2017 estimate)
3. US Census data, estimate for 2017

The goal is to build a model that can not only make reliable predictions as to whether a given ballot question has a good chance at the polls, but to also provide an explanation as to why certain counties are predicted to vote the way they do. This way, users, such as advocacy groups, could choose to target not only specific counties, but also specific communities they might otherwise have ignored, yielding an advantage over opposing groups in an actual vote. See the Maryland minimum wage example below for an explanation of how this works.

In terms of machine learning models, Generalized Linear Regression, Lasso Regression and Random Forest models proved to provide the most meaningful results. Using these models, whichever of these 3 models has the best average cross-validation score is selected for display. These models are chosen due to their relative ability to explain what drives votes for or against the ballot question issue. Because so many of the features are correlated with each other, Lasso Regression proves to be extremely useful in explaining predictions. Generally, Lasso Regression picks out a single political, single economic, single racial, and a single social factor for its predictions and weights them accordingly.

Let’s work through an example, let’s propose a minimum wage increase in the state of Maryland. Selecting the “Increase Minimum Wage” issue and typing in “MD” for the state, we are presented with the map below:

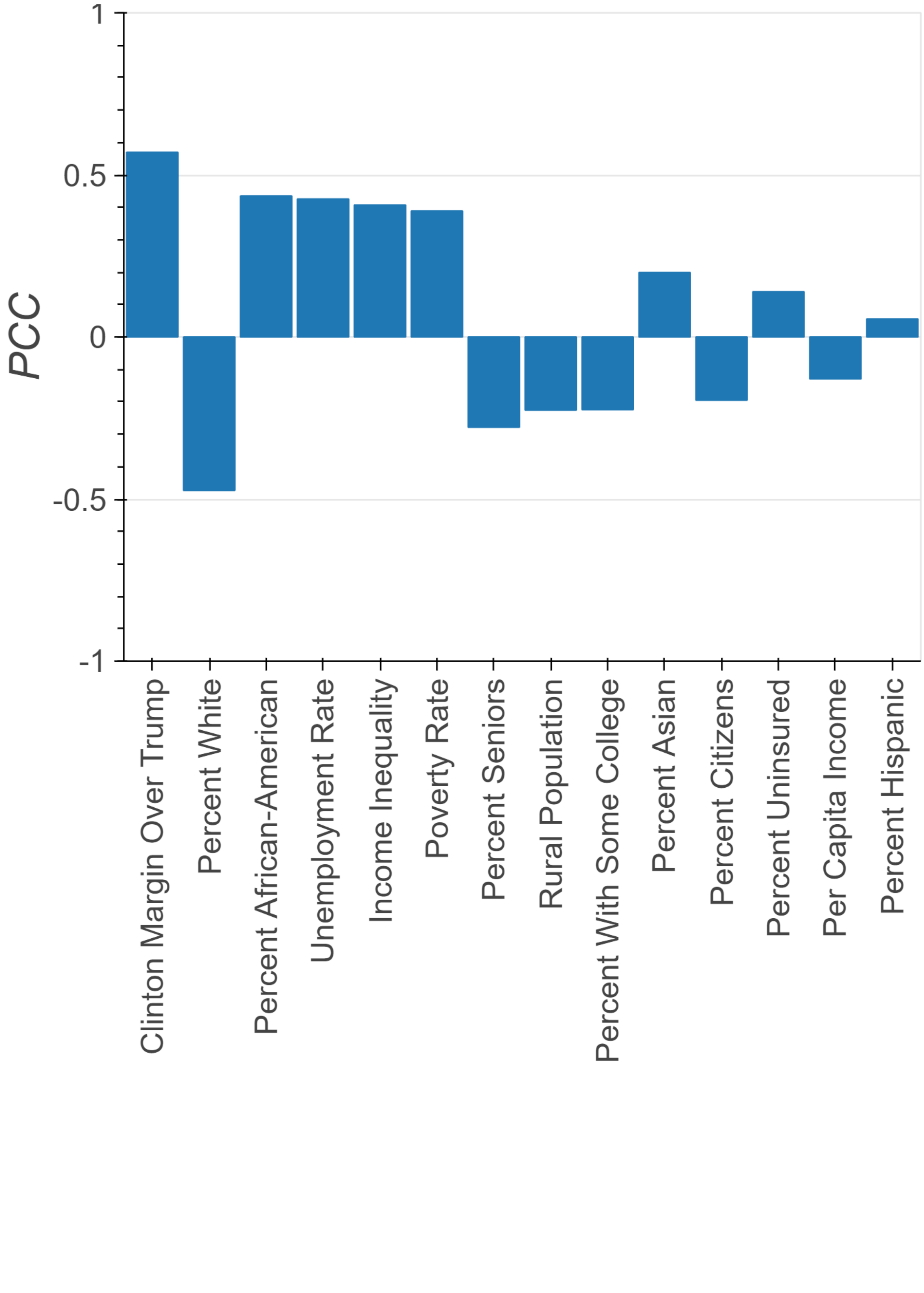


In the map, deeper shades of red indicate less support, while deeper shades of blue indicate counties with greater levels of support. Gray indicates a vote split of 50-50. Intuitively, we see that more democratic counties give darker shades of blue and more republican counties have various shades of red, perhaps "yes" votes for minimum wage increases are simply effects of political affiliation?

Vote preference calculates correlations between the percentage of "yes" votes and each of the 14 features listed above. For a minimum wage increase, the result is shown on the next page. Features are listed in order of highest absolute correlation/anti-correlation values. Unsurprisingly, the percentage of excess votes Clinton received over Trump was the most correlated with the percentage of "yes" votes on the minimum wage increase votes. However, the correlation is, while significant, only moderate (0.57). A second glance at the map above tells something else is at play here. The eastern shore of Maryland is heavily republican; however, we see that the lower eastern shore counties (Wicomico, Dorchester, Somerset) are predicted to have less opposition to the minimum wage increase than similarly pro-Trump counties. Why? Looking at the plot, support for Trump or Clinton is not the only feature moderately correlated with the percentage of "Yes" votes. Instead we see that percent Whites, percent African-Americans, Unemployment Rate, Income Inequality, and Poverty Rate all have correlations/anti-correlation absolute values above 0.35. While percentage of Whites and African-Americans are highly correlated with Trump/Clinton support, the others are economic indicators, suggesting counties with higher unemployment rate, greater income inequality, and greater poverty are more receptive to minimum wage increases than wealthier ones. Therefore, if we were running a campaign to get “yes” votes, we may want to target more impoverished communities on the Eastern Shore that voted for Trump. This could prove a key advantage in a tight vote. In contrast, if we wanted a “no” vote, we could instead target middle-class and high-income Republican communities in our “get out the vote” efforts.

Conclusion

Thank you for your interest in vote preference. While the guide above presents a brief look at how to use and analyze Vote Preference, there was a lot more in my analysis. Thus, in the below appendix, I present an analysis for those interested.


PCC
1
0.5
0
-0.5
-1
Clinton Margin Over Trump
Percent White
Percent African-American
Unemployment Rate
Income Inequality
Poverty Rate
Percent Seniors
Rural Population
Percent With Some College
Percent Asian
Percent Citizens
Percent Uninsured
Per Capita Income
Percent Hispanic

Appendix

Below is additional analysis, issue by issue, that paint a picture of how Americans vote. This is based on the correlations, specifically the Pearson Correlation Coefficient (PCC) between the 14 features and the percentage of “yes” votes in the training data.

Motivation

This project originated in using data to learn how Americans vote on issues, rather than politicians or political parties. For example, are there issues where party affiliation does not matter? Are there issues where, for example, a person's income or race or religion is more predictive of how they will vote as opposed to political affiliation? Or, are there certain issues that are voted on in highly partisan ways?

Ballot initiative questions are a good way to answer these questions. Since ballot questions often focus on specific issues, such as increasing the minimum wage, taxes, labor rights etc., they are a good way to gauge how individuals think about these issues. Moreover, given that state-wide ballot questions are voted on across diverse sets of counties for an entire state, there is an opportunity to examine how popular different issues are based on the political, economic, racial, and social makeup of each county.

Let’s look at which issues are most politically partisan and which have nuances beyond a simple partly line vote.

Highly partisan issues

Of all the issues studied, the most classic party-line vote occurs on measures to increase tobacco taxes. The PCC for “yes” votes vs. Clinton-Trump margin is 0.91, indicating an extremely strong correlation. The Random Forest Model indicates a feature importance of 84% for this feature, while all 13 other features have at most a feature importance of about 2%.

Other issues, while not quite as partisan, still have votes that are split heavily on political lines. Many of these tend to be economic issues. For example, support for implementing so-called “right-to-work” laws (laws prohibiting mandatory membership in labor unions as a condition of employment) are correlated with pro-Trump votes. However, there is also a significant economic component to this vote, with more affluent voters opposing “right-to-work” laws and more impoverished voters supporting them. Based on this, it seems workers who may be desperate for a better job think closed (all-union) workplaces are an obstacle to their potential employment while perhaps more pro-Union, middle-class voters see “right-to-work” laws as a threat to them. Measures increasing the minimum wage see voters divide along similar lines, though, as in the example, more impoverished voters favor higher minimum wages rather than oppose them. There does appear to be a significant level of opposition among pro-Trump voters, but this does not appear to be significantly linked with any other feature. Ballot questions to prohibit mandatory health insurance (popular before 2016, as a means to oppose Obamacare) could be used to indirectly measure support for this health care law. From this, Obamacare is mostly opposed by pro-Trump voters; however, once again, there is a strong economic component, with areas with more unemployment and poverty seemingly more in favor of Obamacare.

Very popular non-partisan issues

These issues tend to almost always succeed at the polls and thus have very poor predictive models, since they almost always have high base levels of support. They tend to be measures funding community institutions such as schools and parks. For example, issuing Bonds to fund the government tend to be more opposed by pro-Trump counties than more pro-Clinton counties (approximately 60% vs 70% "yes" votes), but nearly always succeed. Funding for Parks is also nearly universally popular. Requiring secret ballots for unionization tend to almost always succeed, with only the most pro-Clinton, African-American, and impoverished areas opposing it by slim margins.

Controversial non-partisan issues

These are issues for which the level of support for Trump or Clinton does not appear to be the most significant driving factor for "yes" votes. They tend to be social or religious issues that pit more affluent, secular, educated voters against more religious rural voters in conjunction with some urban African-American (for some issues). African-Americans also tend to be a highly religious demographic.

Legalizing medical marijuana tends to be divided between more socially conservative, rural votes and more socially liberal, affluent voters, regardless of political preferences. It also enjoys a relatively high level of base support. Looking at extending legal marijuana legalization to recreational use, the base level of support drops. In contrast to medical marijuana, the models indicate that while more affluent, educated votes support legalization, there is a far more significant political divide where pro-Trump voters are more skeptical while more pro-Clinton voters tend to vote "yes". Support for legalizing physician assisted euthanasia divides among similar lines. Interestingly, heavily African-American areas tend to vote more conservatively on this issue; for example, Vote Preference predicts heavily African-American Baltimore City to yield 44% "yes" votes on this issue despite its heavy support for Clinton in 2016. In contrast, liberal, highly affluent Montgomery County is predicted to be the place of highest support in MD. Initiatives that ban public funding for abortion or allow public funding to go to religious organizations tend to only be favored by rural voters and perhaps some African-American voters.

One of the weirdest issues in terms of measuring who supports it are votes to expand casino gaming. It is one of the only issues whose outcome is not most correlated with the 2016 presidential election results and is relatively controversial (i.e. votes can be close to 50-50). Instead, adding casinos or slot machine measures is anti-correlated with the poverty rate and correlated with the percentage of rural population. It is not entirely clear from the data what drives "yes" votes for casino expansions and likely require more analysis to understand the mechanics of how people vote. The R^2 values are relatively decent (0.61), and the cross-validation scores are comparable to other reasonably predicting issues, such as increasing the minimum wage. This indicates the model have some predictive power, but more exploration needs to be done to determine the root motivation of why people vote the way they do on this issue. Casino expansions may be a NIMBY (Not In My Backyard) issue, something not accounted for entirely in the model. For example, voters who are highly rural may not see the casino affecting them, but someone else, and vote "yes" since the casino increases state revenues without increasing taxes. In contrast, counties with higher levels of unemployed voters may see their communities targeted as a convenient place to site the casino and must personally deal with the negative consequences of having the casino sited in their communities, such as increased traffic, crime, and addicted gamblers. On the other hand, some poorer communities may welcome casinos, as lower per capita income suggests more "yes" votes.

Issues where models failed with no explanation

Several issues resulted in very poor models. Usually, models were poor when either: the issue is very popular among most Americans (see the section above on very popular non-partisan issues) or none of the features are predictive. The biggest example of the later case is increasing the sales tax. No feature had an absolute PCC value above 0.2 and thus no good model was trained to predict how sales tax increases are voted on. However, correlations do not tell the full story again. The base level of support for a sales tax increase is relatively low.

Clusters of Voters

Then, what are the different types of voters. The analysis suggests 4 main axes: political (i.e. pro-Trump or pro-Clinton), economic (rich vs. poor), racial (often white vs. non-white, often subsumed into other axes), and social/religious (secular vs. religious). I tried using unsupervised learning techniques, such as dimensionality reduction and clustering methods, for example, k-means clustering to validate these measures. Unfortunately, I was unable to conclusively obtain these clusters. Several problems include most counties being white and rural (imbalanced data set), clustering is by counties rather than individual voters. In the end, I had hopes to use dimensionality reduction to engineer new features; however, I never obtained good models from this process. This aspect of the project is ongoing and requires far more work, as of now.