If You’re Single…You’re Probably a Democrat

(and other insights into US demographics and voting inclination)

Overview

Voting behaviors depend, to a significant degree, on news and events leading up to the election; these are often unpredictable and introduce variance that undermines the accuracy of election forecasts. Yet, it is common knowledge that certain demographic characteristics are strong predictors of voting tendencies (e.g., rural areas tend to vote Republican). In this project, we employ machine-learning methods to measure the predictive power of demographic characteristics (race, gender, education, socio-economic status, marital status) in determining voting outcomes, focusing in particular on the US popular presidential vote.

Problem formulation:

For a county and presidential election, we define the ‘*voting inclination*’ as the difference between the (percentage of) republican and democrat votes. We seek evidence for the following:

Thesis: Demographic data is a strong predictor of voting inclination across large timescales.

Data Collection:

County presidential election returns for the years 2008, 2012, 2016, and 2020 were downloaded from the MIT Election Data and Science Lab. County demographic data for the years 2008, 2012, 2016, and 2020 were downloaded using the data tool offered by IPUMS NHGIS; the data is sourced from the US Census Bureau's ACS 5-year surveys.

Model Selection and Training:

For subsets S of { 2008, 2012, 2016, 2020 }, we trained three regression models to predict voting inclination on the data from S. Our models were Ridge Regression (baseline), K Nearest Neighbor regression and Random Forest Regression. To mitigate overfit, we performed five independently shuffled 10-fold cross validation splits, resulting in fifty training, holdout splits. For each model, we carried out hyperparameter tuning in concert with cross-validation and (for Ridge and KNN) recursive feature elimination. Finally, to balance out the individual weaknesses of the models, we computed a weighted average of their predictions (in essence, a custom voting regressor in which the base models are trained on different sets of features).

Key Performance Indicators and Interpretation

Each model was used to predict voting inclinations for years not in S, and we took as our KPIs the *mean-squared-error* and (classification) *accuracy,* i.e., the fraction of counties for which we correctly “called” the election. We pause here to emphasize that, in the lens of our problem formulation, we view a model that achieves a performance rating close to its original mean-CV rating as evidence for our thesis (i.e., demographic data is a strong predictor of voting inclination across large timescales.)

Results:

The classification accuracy score of the models remained relatively stable between the cross-validation and test sets, in most cases not exceeding 3-4%. We found two remarkable patterns, which went against our intuition, while still providing positive evidence for our thesis:

1. In a majority of cases, the test accuracy turned out better than the CV accuracy. For example, the Random Forest Regressor trained in 2008 had CV accuracy of 0.855; when

we tested it on 2012, 2016, and 2020, we found accuracies of 0.888, 0.872, and 0.83, respectively.

2. The accuracy scores tend to be better when predicting forward in time. For example, among all the models trained in 2016, the test accuracies for the year 2020 ranged between 0.866 and 0.939, whereas the test accuracies for 2008 ranged between 0.759 and 0.808.