Final Project - Milestone 3

Data Description

Outcome variable

We aim to predict the percentage of votes for the Democratic and Republican presidential candidates in each county across 51 states weighted by each county's population. The outcome variables for the training data are derived from the total population in the county times the percentage of votes for the Democratic and Republican presidential candidates, respectively, in the 2016 presidential election. The model's predictions of the presidential election result in each state is calculated by summing all predicted votes of the counties in the state and dividing by the state's population; the electoral votes are assigned to the democratic party if this final percentage is larger than 50% for the Democratic candidate, and vice versa for the Republicans. Thus, our prediction is a continuous numeric value for 3139 counties. In the base model section, we explore whether using Democratic or Republican votes as our outcome variable provides more consistent results. We believe that using the county-level votes can allow us to have a detailed understanding of the underlying demography, socioeconomic factors, and regulations that drive presidential election results. The vote turnout by county is retrieved from opendatasoft (https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county).

For predictions of the house of representatives, we summarize precinct-level votes to county-level. The precinct-level 2016 data is from the MIT election lab Data section (https://electionlab.mit.edu/data). We will use votes per county for Republican candidates (2767 counties) because there are more missing data for the number of votes for Democratic candidates (2282 counties) in the data set. There are four missing states: Alaska, District of Columbia, North Dakoda, and Vermont; yet each state only has one house of representatives and their party membership is constant over time. Thus, we will impute them for our data. We will aggregate votes per state for Republicans and divide it by the total number of votes per state then multiply it by each state's apportion of house representatives. This will become our final prediction of the proportion of house of representatives for each party.

Voting Consistency

Often, certain states have the tendency to vote for a particular political party. It is likely that this consistency carries into future elections. Therefore, it is worth taking this into consideration to predict our elections results in the year of 2020. To approximate into county-level data, we looked at both years of 2008 and 2012 and calculate the voting percentage ratios (continuous variable) for either democrats or republicans. If the ratio is smaller than 1, it means there is a decreased voting percentage for that political party from 2008 to 2012. If the ratio is greater than 1, it suggests there is an increased voting percentage for that particular party from 2008 to 2012. Voting percentage data are pulled from (open source opendatasoft).

Base demographics

We include 5 categories of demographic data and 2016 data for our training set and 2018-2019 data for our test set.

1. Race, ethnicity, gender (source: County Population by Characteristics: 2010-2019 https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html):
    a. We include 12 continuous numeric variables describing population per race/ethnicity group by gender by county: total male, total female, White American male/female, Black or African American male/female, Black or African American male/female, American Indian and Alaska Native male/female, and Native Hawaiian and Other Pacific Islander male/female
2. Education status (source: Economic Research Service- Educational attainment for the U.S., States, and counties, 1970-2018; https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/):
    a. We include 5 continuous numeric variables describing education status per county: % of population with high school diploma, at least bachelor's degree, graduate degree, and school enrollment.
3. Poverty (source: same as race, ethnicity, gender category):
    a. 2 continuous numeric variables for poverty characteristics per county: population of poverty estimated for all ages and median household income in dollars.

4. Insurance status (source: 2008 - 2018 Small Area Health Insurance Estimates using the American Community Survey):
    a. 2 continuous numeric variables per county: numbers of uninsured, numbers of insured per county
5. County business pattern (source: County Business Patterns: 2016 and 2018 https://www.census.gov/data/datasets/2016/econ/cbp/2016-cbp.html):
    a. 19 Continuous numeric variables per county for total numbers of establishments in the following business categories: undefined, Administrative, Agriculture, Construction, Educational Services, Finance, Food services, Health care, Information, Management, Manufacturing, Mining, Other, Professional Service, Real Estate, Recreation, Retail, Transportation, Utilities, and Wholesale.

Financial data

The amount of money a candidate spends on their campaign may predict how likely they are to win, since candidates that spend more money will likely be able to reach more people and sway more voters. Because there is no state or county level data of presidential spending for the election, we will be using the total spending within each state by house representatives as an approximation. Our data is from the FEC (https://www.fec.gov/data/browse-data/?tab=candidates), specifically the amount of money each candidate received and spent during their campaign, and had left afterwards.

Voting accessibility

Accessibility to voting, or level of voter suppression, could help predict the election. Populations that are unable to vote in certain states tend to have shared demographic or socioeconomic traits, and thus be more likely to vote for a certain candidate. This inability to vote may result from restrictions placed on the degree of identification necessary to cast a vote, or how complex the process is to send in a mail-in ballot. To represent these predictors, we used the following data:

1. Strictness of Voter ID Laws, Voter Registration Laws, Felony Disenfranchisement Laws, and Early Voting (source: The Guardian https://www.theguardian.com/us-news/ng-interactive/2019/nov/07/which-us-states-hardest-vote-supression-election)
    a. Each law is represented with an ordinal categorical variable coded to represent strictness
2. Rates of early votes (source: United States Elections Project http://www.electproject.org/early_2016)
    a. This data is only available on the state level, so it is estimated by calculating the percentage of the population that voted early at a state level.

Economic uncertainty

Economic uncertainty could impose a big influence on presidential and house representative elections. Here, in order to delineate the relationship between economic uncertainty and election results, we decided to look into the unemployment rate (continuous variable). Data of 2016 and 2019-2020 are pulled from the U.S. Bureau of Labor Statistics (https://www.bls.gov/data/#unemployment).

Visualizations and findings from EDA (Reference Jupyter notebook for graphs)

Response Variables

Figure 1 - Histograms of % votes for Republicans and Democrats per county

% votes for Republicans are left skewed and % votes for Democrats are close to normal distribution

Figure 2 - Histograms of Number of Votes Weighted by County Population for Republicans and Democrats per County

Democrats and Republicans have a similar distribution of votes weighted by county population in each county

Figure 3 - Histograms of log-transformed numbers of votes weighted by county population for Republicans and Democrats per county

Both parties' log-transformed counts of votes weighted by county population are normally distributed

Voting Consistency

Figure 4 - Histogram of Percentage Votes Ratio Between 2012 and 2008 for Both Parties

This suggests that a majority of counties are shifting their voting habits with the trend moving to vote more for republicans in 2016 presidential elections.

Base Demographics and Business

Figure 5 - Histogram of all 41 demographic and business variables adding economic uncertainty
Raw data is right-skewed thus warranting log transformation to better see the trend
Figure 6 - Heatmap of all 41 demographic and business variables adding economic uncertainty
All demographic based data are moderately associated with each other (pearson correlation lower or equal to 0.3) thus showing that regularized method should be employed to deal with collinearity
Figure 7 - Scatter plot of all 41 demographic and business variables adding economic uncertainty against % votes for Democrats and Republicans
Each scatter plot shows an opposite trend of demographic variables with % votes for Democrats vs Republicans in 2016 besides weaker trends in median household income and school enrollment.
Figure 8 - Scatter plot of all 41 demographic and business variables adding economic uncertainty against numbers of votes for Democrats and Republicans weighted by county population
Using the weighted by county population increases correlation between outcome variable for both parties

Financial data

Figure 9 - Distributions of Campaign Financial Features
There is no clear trend for any financial predictors on each of the different voting response variables.
Figure 10 - Heatmap of all 6 financial variables
Moderate correlations are seen among variables with a few opposite correlations between Repulican and Democrat spendings. Correlations again point us to regularized-model approach.
Figure 11 - Scatter Plot of all Financial Features and Different Response Variables
The trends between financial investment of campaigns with % votes do not seem to be strong. Trend is not obvious.

Economic uncertainty

Figure 12 - Histogram of Unemployment Rate (%) in 2016 and 2019-2020
It seems the median and the mode of unemployment rate in 2016 are falling between 4-6 percent. When compared to the unemployment rate 2019-2020, first of all, there are way fewer people reporting during 2019-2020. Second, the unemployment rate is much higher in 2019-2020 with respect to mean and median.
Figure 13 - Scatter Plot of Unemployment Rate (%) vs %Votes
It seems that when the unemployment rate is below 10 percent, it is more likely to vote for republicans than democrats in 2016 presidential elections.

Voting accessibility

Figure 14 - Violin Plots of Voter Distribution of Each Level of Law Strictness
None of the violin plots seem to show a significantly different rates of voting for a particular party between counties that have different degrees of voter suppression. While more strict voter ID and voter registration laws seem to lead to slightly increased rates of Republican voting, these differences are minor.
Figure 15 - Histogram of County Advanced Voting Frequencies in 2016
While there doesn’t seem to be a clear distribution, the majority of states in 2016 had 20% or less of their population vote in advance, either through mail or through early in-person voting.

Correlation of All Features

Figure 16 - Heatmap of All Features
The range of correlations between our features is low to moderate, with pearson correlation ranging from -0.4 to 0.3, suggesting that our model can benefit from regularization.

Revised project question

After exploring the data, our project and question have remained largely the same, with only the instantiations of variables changing and the addition of the voting consistency variable into our model. Therefore, our project question is as follows: does accounting for various demographic, legal, and behavioral variables allow a model to predict the 2020 presidential election and federal House of Representative election with higher accuracy than models did the 2016 election?

Baseline model

Presidential election models with Republican presidential votes as response (See Figures 17 - 22 and Table 1)

Outcome:	*# votes for Republicans weighted by county population*	*Log transformed # votes for Republicans weighted by county population*	*% votes for Republicans*
Decision Tree	Train R^2 = 1.000 Val R^2 = 0.8159	Train R^2 = 0.9871 Val R^2 = 0.9639	Train R^2 = 0.6148 Val R^2 = 0.4939
Ridge Regression	Train R^2 = 0.9799 Val R^2 = 0.9491	Train R^2 = 0.6248 Val R^2 = 0.1755	Train R^2 = 0.6792 Val R^2 = 0.5194

Presidential election models with Democratic presidential votes as response (See Figures 23 - 26 and Table 2)

Outcome:	*# votes for Democrats weighted by county population*	*Log transformed # votes for Democrats weighted by county population*	*% votes for Democrats*
Decision Tree	Train R^2 = 0.9999 Val R^2 = 0.9081	Train R^2 = 0.9787 Val R^2 = 0.9518	Train R^2 = 0.7429 Val R^2 = 0.5305
Ridge Regression	Train R^2 = 0.9944 Val R^2 = 0.9834	Train R^2 = 0.7211 Val R^2 = 0.4123	Train R^2 = 0.6494 Val R^2 = 0.5683

House of Representatives model with Republican house representative votes as response

	# votes for Republicans weighted by county population
Decision tree based (3 fold Cross validation)	Train R squared= 0.8720, Validation R squared = 0.7291
Ridge Regression (3 fold Cross validation)	Train R squared= 0.9399, Validation R squared = 0.8564

Conclusion

Decision tree and ridge regression performs equally well for predicting both presidential and house of representatives votes, so we will proceed with a regularized tree-based model for predicting both outcomes.