Predicting Election Results in Swing States

Vaed Prasad(vsp22), Emily Weed(emw232), Alexander Salonga(as2739)

December 2020

1 Introduction

Unlike democratic elections in most nations where leaders are elected based on the popular vote of its constituents, presidential elections in the United States follow a unique electoral system outlined by the Electoral College. In this process each state is allocated a share of electoral votes that is roughly proportional to its population and, in predominantly every state, the winner of the statewide plurality collects all the electoral votes from that state. This structure presents a phenomena where the outcome of an entire presidential election can be decided based on the results of a handful of swing states, where either presidential candidate has a feasible opportunity of winning the statewide plurality. In this project we aim to leverage historical demographic, socioeconomic, and electoral data to predict which candidate will win these critical swing states in the 2020 United States Presidential Election. For additional details regarding our goal, please reference our proposal.

2 Exploratory Data Analysis for 2016 Data

2.1 Data Selection

In order to predict the results of the 2020 U.S. Presidential Election, we honed our focus to twelve key swing states: Michigan, Florida, Nevada, Texas, Minnesota, Wisconsin, Iowa, Ohio, North Carolina, Georgia, Arizona, and Pennsylvania. These states were selected by considering the closest states in the 2016 U.S. Presidential Election. For these twelve states we collected county level data containing racial composition, age and gender clusters, economic features, health conditions, occupation, educational level attained, etc. We plan to use this data to analyze their respective influence on the county's voting distribution for the Republican and Democratic presidential candidates in the 2016 election.

2.2 Data Cleaning

In Data Cleaning and Manipulating we compiled our various data sets into a single dataframe where each row represented a particular county. In order to facilitate the merging of our various datasets, we utilized the "Fips" feature as a primary key, as it is the unique identifier for all counties in the United States.

Some of the numerical data in the CSVs were encoded as type String. Thus, we also had to clean the data accordingly, converting them to type Int so we can later run regressions and measure errors for our model with these features.

One consideration we had to make when cleaning our data was that a number of features were represented as a raw count (i.e. "AGE1824-MALE", "AGE1824-TOT"), which would be heavily dependent on the county's population. As a result, we normalized these particular features by dividing their raw count by the "Total Population" column, so that we could consider a more relevant per capita representation of these features. Next, we pruned features with a large density of missing values by calculating the null ratios for each feature and dropping features that had a null ratio greater than 0.0001.

2.3 Feature Selection

After cleaning our data we had a large slate of merged, normalized, and filled data to consider for our final feature selection. In order to pick the most relevant features for our model we constructed a correlation plot to observe a feature's relationship with the Democrat/Republican vote share. We were aiming to select about 20 features (besides the 2012 vote share) to work with further so we chose the subset of 20 features with the largest | Correlation of x with "Democrat 2016" or "Republican 2016" |. We constructed the correlation plot for our final subset of features. See Figure 1 below.

2.4 Feature Transformations

We created scatter plots of our features against our target variables to assess the linearity of the relationships. We saw that every feature except "realGDP2016" showed a linear relationship, rather it be positive or negative, with the target variables. Figure 2 shows the distribution of points before performing any feature transformations. It is obvious that there are many counties with similar, smaller values for real GDP while a few larger counties skew the plot, making it difficult to interpret. To solve this we took the log_{10} of "realGDP2016." As shown in Figure 3, this creates a much more linear relationship between the variables, making it more useful

for our model. All of our features are continuous numerical variables so we did not have to deal with encoding any nominal, ordinal or text features.

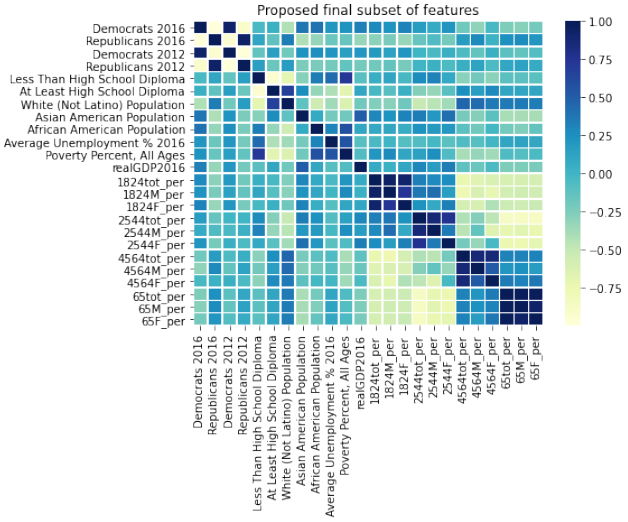

Figure 1: Correlation Plot of Final Feature Selection

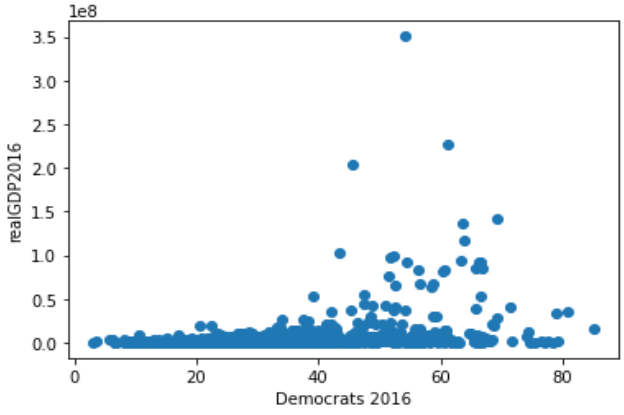

Figure 2: County Level Real GDP vs. Democrat Vote Share

3 Modeling on 2016 Data

3.1 Model Construction

We plan to use 5-fold cross validation to get a better understanding of the test error before performing any predictions with our test set. This will also give us a better sense of whether our model is overfitting or underfitting because we can get a better estimate of the test error and adjust our model accordingly.

We are using a Linear Regression Model. Initially we tried to see how well our model performed without using previous year's data on Democrats/Republicans vote share. We suspected that this would not have a very high accuracy but it serves as a baseline for our future models. In Model Construction we got training and validation R^2 around 0.55, train mean squared error (MSE) and validation MSE both around 75. The R^2 demonstrates the amount of variation in the response variable the features explain (closer to 1 is better). These errors are very high for such a specific objective as ours, but we can see that our model is neither overfitting or underfitting here because the training and the validation values are very similar for R^2 and MSE.

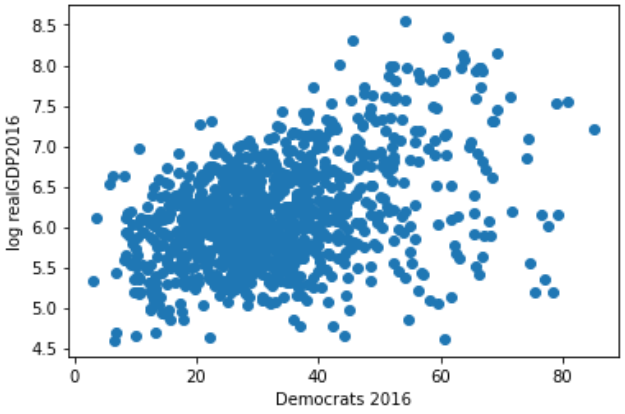

Figure 3: Log County Level Real GDP vs. Democrat Vote Share

Next we added the vote share from the previous presidential election which provided strong insight on the likely outcomes of the 2016 election. Our R^2 values and MSE values greatly improved as was expected (roughly .978 and 4.5 respectively for both training and validation). We also checked the previous year's vote share by itself to get a sense of the impact of our other features we were including. It gave better results than the original model (without vote share from the previous presidential election) but was not as successful as the model with all the features. This indicates that we should investigate using a combination of all the features, and that the demographic and economic features do in fact contribute to our model's accuracy.

The subset of features we decided that would be best to proceed with was all those included in Table 1, excluding the gendered breakdown of age groups. Training a model on these gives very similar training and validation R^2 and MSE values as our full model with all the features. We ran this model on our test set and analyzing these error values. Figure 4 shows a scatter plot of the predicted values vs the actual values with the points colored by how far off they are off from the linear regression model (dotted line). Figure 5 shows a histogram of Actual - Predicted for every value in the test set. Since this perform the same as our other one with all the features, choosing the smaller, less complex set of features is preferable as sparse models are better to work with.

Variable	Type
Democrats 2016	Continuous
Republicans 2016	Continuous
Democrats 2012	Continuous
Republicans 2012	Continuous
No HS Diploma	Continuous
Least HS Diploma	Continuous
White	Continuous
Asian American	Continuous
African American	Continuous
Avg. Unemployment	Continuous
Poverty Percent	Continuous
Real GDP	Continuous
18-24 Total	Continuous
18-24 Male	Continuous
18-24 Female	Continuous
25-44 Total	Continuous
25-44 Male	Continuous
25-44 Female	Continuous
45-64 Total	Continuous
45-64 Male	Continuous
45-64 Female	Continuous
65+ Total	Continuous
65+ Male	Continuous
65+ Female	Continuous

Table 1: Final Feature Selection

4 Exploratory Data Analysis for 2020 Data

4.1 Data Selection

Our next step was to collect data for 2020. We needed to find all the same features we used to train our final model for 2016 data in order to be able to run this model on 2020 data and assess the accuracy.

Explain here what was available and how this changes our objective Go from here to explain data cleaning (pretty much what I already did in the 2020 data collection/cleaning notebook) and then split to explain each of our new approaches

4.2 Data Cleaning

5 Future work TODO

Initially, we only ran our models using the Democratic vote share as our target variable because we believed that it was complementary to the Republican vote share; however, we plan to also run this model where the Republican vote share is the target variable to identify if there are any discrepancies.

It was noted that different methods of voting were heavily emphasized in determining the outcome of the 2020 election, thus for each county/state we will try to determine the vote share in terms of who voted by mail, had an absentee ballot, voted in person on election day, and participated in early voting.

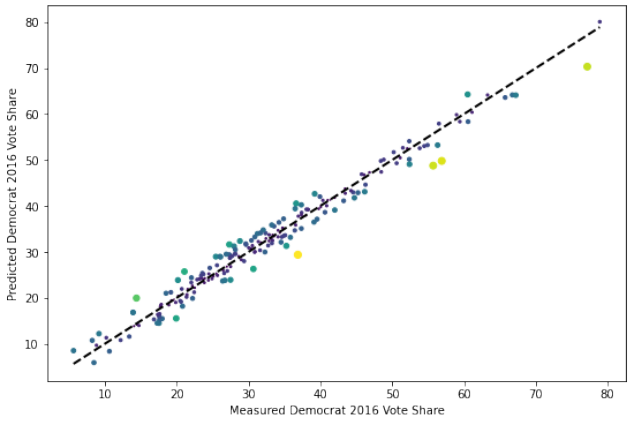

Figure 4: Correlation Plot of Final Feature Selection

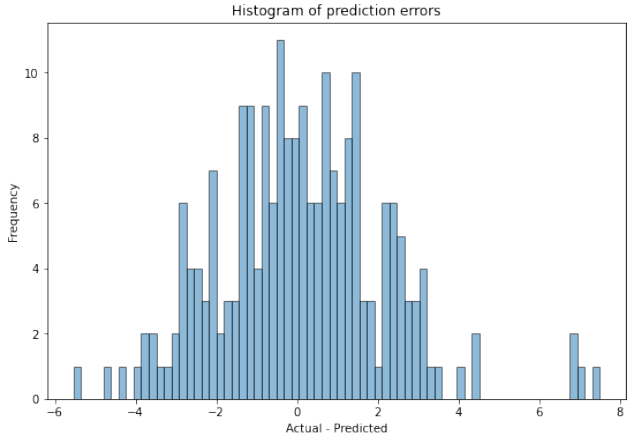

Figure 5: Correlation Plot of Final Feature Selection

6 References

- County Level 2016 Electoral Results
- County Level Age and Gender Distributions
- County Level Economic Data