Voting: Demographics & Outcomes

(Project Repository: https://github.com/srijanrodo/erdos_ds_project)

Project Overview: Many factors contribute to determining the political alignment of a population. In this project we focus on the US 2020 presidential elections, investigating the impact of the demographic factors in the outcome at the county level.

Stake Holders: This information is useful for geopolitical institutions or any investor who wants to understand which policies are more likely to be applied in the next future in a certain community of people.

Data Gathering: We collected our data from the American Community Survey 5-year database. This data is collected over a period of 5-years (in our case ending in 2019) from all over the US covering all counties. Initial exploration suggested the use of the following key indicators: total population, percentage of people in poverty, percentage of people with a bachelor's degree, ethnic distribution, the split into urban and rural areas, unemployment rate and a few related variables. We also collected the county level presidential election data from MIT Election Lab’s database.

Data Preprocessing: We first selected a collection of features from our data sets that seemed to affect the political alignment of the population. These features regarded population, average age, ethnicity, poverty, unemployment and education level.
After some more work on the data we realized that some feature were not actually affecting the elections outcome, as long as some interesting correlations. This lead us to the final set of features we used. More on data cleaning here:
https://github.com/srijanrodo/erdos_ds_project/blob/master/README.md

Modeling Approaches: We approached our problem with two different kinds of models:

- A regression approach, predicting the percentage of votes gained by the democratic candidate.
- A classification approach, predicting the winner of the election at a county level.

For regression we considered the following models:

1. Linear regression
2. Random Forest Regression
3. Gradient boosting via XGBoost *
4. Feed-forward neural network

For classification, we considered:

1. Random forest classification

2. Logistic regression
3. Support Vector Machines (SVC) *

Results: Comparing the performance of the models we examined, upon all combinations of features, we saw the best model is XGBoost for regression and SVC for classification.

XGBoost yielded a final mean square error score of 8.35 and an R^2 score of 0.63.
SVC yielded an accuracy score of 0.9197 (meaning a correct prediction in 91.97% of times)

The following features stood out as impactful for regression:

- Ethnic distribution (percentage of population classified as white)
- Education (percentage of people with a bachelor's degree)
- Older population (age 60 and above)
- Total population (logarithm of total population)

Overall our accuracy seemed to be better for the classification problem than the regression ones. This seems to indicate that the demographic indicators do have a high correlation with political alignment while the exact votes gained by a candidate is more variable and depends on additional predictors. Indeed, this is supported by the fact that most counties in the US have the same political alignment over different elections even though the exact number of votes vary.

Future Improvements: The downside of using demographics features is that any prediction made starting from them is not very meaningful, since they don’t typically change significantly in a four year period of time. For this reason it would be interesting to find some new features that could detect the sentiment of the population at a certain moment.
Good candidates for this could be:

- Sentiment analysis of social media posts made in a given county in a certain period of time before the election; sentiment analysis of local news.
- Analysis of voter turnouts and general political involvement (e.g. council meeting and political rally attendance)