

BobbyGerberick / Flatiron_Capstone

Q

+ ▾

🕒

🔗

📁

👤

<> Code

🕒 Issues

🔗 Pull requests

🎬 Actions

📁 Projects

📖 Wiki

🛡 Security

📈 Insights

⚙ Set

👁

🔗

☆

📄 MIT license

☆ 0 stars

🔗 0 forks

👁 1 watching

📈 Activity

🌐 Public repository

🔗 main ▾

⋮

🔗 Branches

🏷 Tags

👤 BobbyGerberick

Merge pull request #9 from BobbyGerberick/bobby ...


1 minute ago ⌚ 25

View code

☰ README.md

✎

Voter Data Analysis 🔗



Overview

As my capstone project for Flatiron School's Data Science program I built a model to predict how individuals would vote in a presidential election based on data from the 2012, 2016 and 2020 elections. I then used that model to analyze how broad categories of political issues and individual issues themselves influence an individual's vote. I also examined the accuracy of predictions based on basic demographic information like income, race, education etc.

Business Understanding

This type of modeling could be useful in a number of contexts. Most obviously for a campaign interested in focusing their efforts on individuals most likely to vote for them but it could also be useful for political parties and special interest groups who want to better understand their constituents and the public as a whole.

Data

My data comes from the American National Election Studies for 2020. The ANES is a national survey of voters in the United States, conducted before and after every presidential election. I used a subset of that data curated by the Inter-university Consortium for Political and Social Research. The full ANES survey data is publicly available for download from here: <https://electionstudies.org/>. You do have to make an account to access the data which you can do by clicking the login button in the top right corner of the home page. Once you have completed that process click on the Data Center tab at the top of the home page, then select the data set you would like (For example: 2020 Time Series Study) and then under the download data heading on the next page select the type of file you would like. The Inter-university Consortium for Political and Social Research's data is available here: <https://www.icpsr.umich.edu/web/pages/instructors/setups2020/> to individuals with an email address from with one of their member institutions. Once you have made an account with that email address, click on the "Find Data" tab at the top of the page and search for the data set i.e (Voting Behavior: The 2020 Election). The first result will take you to a page where you can download the data.

Data Preparation

To prepare my data for modeling I first dropped all rows where individuals did not vote or voted for a third party candidate. This left me with 6075 rows to work with. The columns are broken into 16 categories denoted by a letter in front of the question number. For example A01 and R15. Questions in categories A, D and E relate to past political behavior and opinions of current and former politicians. These are obviously strongly correlated with vote preference and are uninteresting in terms of analysis so were dropped. The data is categorical and so needed to be encoded. I used One Hot Encoding to avoid imposing a hierarchy where none should exist.

Methods

- Decision Tree
- Random Forest

- TruncatedSVD
- XGBClassifier

Results & Conclusions [↗](#)

The XGBClassifier was the best model with an accuracy score on the test data of 96.84%. Additionally I analyzed the accuracy scores for the XGBClassifier when I only showed it individual categories. I found that Political Engagement had the lowest score which is not surprising because with the country being so narrowly divided, both parties have similar levels of political engagement. The next lowest score is for the Demographics category. This is unfortunate because this information is often available at the state or county level. So being able to predict how a state or county will vote based on demographics alone would be useful. However the predictions using this data alone are still fairly accurate at 76.4%

Looking at the most accurate categories, the Trust in Government and Health Care & Policy categories have the highest score with both being above 94.5%. This tracks with what we know American Politics at the moment. Democrats and Republicans don't agree on much when it comes to health care policy. For example, a study from 2020 by a team from the Harvard T.H. Chan School of Public Health found that three quarters of democrats would like the federal government to ensure that all citizens have health insurance. In contrast, 79% of Republicans preferred a healthcare system that relies on private insurance. (<https://jamanetwork.com/journals/jama/fullarticle/2777394>)

The reliability of the Trust in Government category when it comes to predicting ones vote is a bit more troubling. It has long been the case that trust in government declines when an individual's preferred party is out of power in Washington as one can see in this analysis from Pew: <https://www.pewresearch.org/politics/2023/09/19/public-trust-in-government-1958-2023/>. However, as we saw following the 2020 election a lack of trust in institutions can quickly turn violent and deadly. The fact that this lack of trust is concentrated on one side of the political spectrum makes that an even more dangerous possibility.

Additionally I looked at individual columns and found that the model can predict with 87.1% accuracy who the respondent would vote for based on their opinion of the federal government's response to COVID-19. COVID-19 was obviously a major issue in the 2020 election however the fact that a single data point can get a prediction this accurate is noteworthy.

Next Steps [↗](#)

The most obvious next step would be to look at data from the 2012 and 2016 elections to see how the issues important to voters have changed. As I mentioned previously the healthcare and policy category was a good predictor for vote choice in 2020. This almost certainly impacted by the COVID-19 pandemic. Analyzing previous elections could help quantify that impact.

Demographic data is generally available and as this project demonstrated a somewhat accurate predictor of how an individual will vote. Improving that accuracy would be very useful to political parties and campaigns.

Finally, turnout among eligible voters in 2020 was 66% which is a high in recent U.S. history. Still, a third of eligible voters did not turnout. If we could analyze those potential voters and understand why they don't vote political parties could boost turnout among their voters or a nonpartisan group could work to boost turnout in general

Repository Structure [↗](#)

├── data ─┬── ANES_Raw_Data ─┬── ANESTimeSeriesStudy2016.DTA ─┬──
SETUPS2012SETUPS2012_Supplemental.do ─┬── 2012_Codebook.pdf ─┬── SETUPS2012.dta ─┬──
SETUPS2012_Supplemental.do ─┬── SETUPS2016 ─┬── 2016_Codebook-Pl.pdf ─┬── 2016_Codebook.pdf
├── Book2.xlsx ─┬── SETUPS2016.dta ─┬── SETUPS2016_Supplemental.do ─┬── SETUPS2020 ─┬──
2020Questions.xlsx ─┬── 2020_Codebook.pdf ─┬── Question_Categories.xlsx ─┬── SETUPS2020.dta ─┬──
SETUPS2020_Supplemental.do ─┬── .gitignore ─┬── LICENSE ─┬──
Predicting_Vote_Choice_Notebook.ipynb ─┬── README.md ─┬── presentation.pdf ─┬── requirements.txt

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 92.5% ● Stata 7.5%