U431888754 tigeryi@bu.edu
Sicheng Yi (Tiger Yi)

Project Proposal - Covid, Polls, Election

Background/Motivation

The goal of this project is to try to see how some of the factors, such as Covid 19, influence the past 2020 U.S. general election. The polls have largely overestimated Democratic performance in the general election, according to multiple national news such as Washington Post:
https://www.washingtonpost.com/politics/2020/12/03/was-2020-really-disaster-polling/

But why is the result of the most polls biased toward Democrats? Did Covid-19 play a major role in biasing the poll and influencing the election? Or is it more of the demographic difference between various parts of the country?

And is it possible to predict and check some of the 2020 election results (%margin difference between REP and DEM) based on other 2020 results in the training set, the previous election result from 2016, the county level demographic data, and the Covid-19 positive rate on the 3,000+ counties level?

Questions

The overall question will be how did the Covid-19 impact the U.S. 2020 elections in general?

Did Covid-19 contribute to the polling error/bias of the election polls?
Many news sources suggest that Covid-19 skews the polls in democrat's favor because they are more likely to answer polls during the pandemic. And can we confirm this claim by some statistical analysis?
https://www.vox.com/policy-and-politics/2020/11/10/21551766/election-polls-results-wrong-david-shor

As the positive rate rises and the number of positive cases rises, how does this trend impact the margin of the polls between DEM and REP?

What is the correlation between the polling error vs. covid positive rate? And is the correlation statistically significant enough? (Captured as R^2)

U431888754 tigeryi@bu.edu
Sicheng Yi (Tiger Yi)

Different counties in the U.S. have been impacted by Covid differently, did the counties that got hit by Covid the most see a big shift in the election margin?

Did other traditional factors, such as the difference between gender, race & ethnicity, employment & economy contribute to the result of the election more than Covid-19?

How exactly did the Covid impact the election? How the highly Covid affected areas voted differently compared to the relatively safer areas? Are those severely affected counties overall more Democratic or Republican? Are those counties having more minority presence?

Can we use the 2016 result, covid data, and demographic data on the county level to predict and recreate the result of the 2020 election? And is Covid as a factor statistically significant enough to be included in the model to predict the 2020 election results?

Try to split up the counties into a training set and test set to predict the final margin for the 2020 general election, based on existing 2020 results in the training set, the county demographic data and the covid data, as well as the previous results from the 2016 general election. To see if we can get a reasonable model to forecast the final percent of margin for the rest of the counties in the test set.

If I can do prediction for rating score on the amazon movie reviews in midterm, I believe I might be able to do something similar in predicting the election results based on modeling.

How to answer the questions:

1. Datasets:
All my datasets will be downloaded into my google drive that i share with you. Link:
https://drive.google.com/drive/folders/19lNytOmjVmR7TMJJ2mieXGBSDe3QS3hI?usp=sharing

2. Source:

Election polls provided by FiveThirtyEight (2 datasets from here)
https://github.com/fivethirtyeight/data/tree/master/election-forecasts-2020
https://projects.fivethirtyeight.com/2020-general-data/presidential_polls_2020.csv
https://projects.fivethirtyeight.com/2020-general-data/presidential_poll_averages_2020.csv

U431888754 tigeryi@bu.edu
Sicheng Yi (Tiger Yi)

General election results by Cook Politics as CSV on state level:
https://cookpolitical.com/2020-national-popular-vote-tracker

General election results from Kragle on the county level
https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv

Covid datasets from Kaggle:
https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_states_covid19_daily.csv

Covid data from Kaggle, JHU tracker
https://www.kaggle.com/headsortails/covid19-us-county-jhu-data-demographics?select=covid_us_county.csv

U.S. Demographic data by county:
https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv
https://www.kaggle.com/etsc9287/2020-general-election-polls?select=county_statistics.csv

Election, Covid, Demographic by County (this dataset has a lot of information columns)
https://www.kaggle.com/etsc9287/2020-general-election-polls?select=county_statistics.csv

3. What to do with those datasets?

The first stage is to combine several of those datasets in order to answer some of the questions. Make sure to get rid of the columns containing useless information.
The data cleaning step is mostly done by Pandas. Need to pivot longer/wider and join multiple datasets for both the state and county level.

I can compare the daily state polling data against daily state Covid data to graph for each state how the Covid positive rate affects the margin of the polls (DEM-REP).

Try to compare the Covid daily positive rate against the daily poll bias/error to see if there is any correlation when performing a linear regression.

The polling error is obtained with a few steps. There is existing daily polling data from each state to show what is the percentage of voters for REP or DEM, so we can easily obtain the polling margin by subtraction. Then we can compare this daily polling margin vs. the actual 2020 election result margin to know exactly how big the polling error is overtime.

In order to compare, we need to make scatter plots for the Covid positive rate and vote margin of the polls (and 2nd graph for the polling error), then we need to run linear regression and check R^2 if the model is statistically significant. Visualization with EDA will help, but also need the model and number. The comparison can be horizontal (across time) or vertical (across states). The data I have is cross sectional data, so I need to take advantage of that.

Try to split up the counties into a training set and test set to predict the final margin for the 2020 general election, based on existing 2020 results in the training set, the county demographic data and the covid data, as well as the previous results from the 2016 general election. To see if we can get a reasonable model to forecast the final percent of margin for the rest of the counties in the test set.

Basically to see if we have a training set and test set. With a few other factors, can we reasonably predict the test set? and how well is our prediction vs. the actual result based on MSE?

Also in order to know what variable to be included, it will be wise to run multi linear regression to see which variables are statistically significant enough.

Sklearn package must be used extensively for statistical analysis.

4. Limitation Bias

Some of the data is on county level others on the state level make it difficult to combine dataset. It would be nice to have polling data on the county level, but only state level data is available.

The demographic data is back in 2018 and some of the counties have changed names so there will be lots of missing value. Unfortunately, this is the latest data I can obtain on the county level.

I have already noticed for the covid data, when you aggregate county into state level, a few data points would disagree. So the data itself isn't 100% perfect. But this should be acceptable as the human error across different datasets is rare still.

The polling data for the time I will use the weighted average of all polls provided by FiveThirtyEight. The poll data itself already has too much bias, the details of how each poll is conducted can't be verified.

The poll itself, in nature, collects only a small sample of likely voters, which already incur a significant sampling error. Then the way most polls are collected on phone calls, or online, which is another sampling bias because those people can't be a perfect representation of the general population. There are certain voters who have little public trust and never answer polls, and they will still vote in the election, just not answer the poll. Another bias is out of control which is people can lie to the polls, but those instances are rare.

5. How to determine the success of this project?

(1) Must have successfully gather and process the dataset in those area:
Election Result (county and state level, 2020 and 2016)
Election Poll (state level, daily)
Covid Data (daily on state level, the election day data on the county level)
Demographic Data (on the county level)
Need to have really clean data to work with more smoothly. Need to take care of missing values and absurd values.

(2) Must know how to use those datasets for analysis. Need DS tools.

Must combine, for daily state level, election, poll, and covid data.

Too examine how Covid impacts the poll margin, and how Covid impacts poll error.
Need to have both visualization and statistical analysis for statistical correlation.

On the county level, try to combine the Covid, demographic, election, poll data

If high positive or death cases / population counties have a big shift in voting pattern from 2016 to 2020. To see whether Covid plays a role in the shift of margin between DEM and REP

Some areas are disproportionately hit by Covid, positive or death cases / population, are those tend to have some association with demographic data such as population number itself, race ethnicity, etc.

To split the county level dataset into a training set and test set. To see based on the known 2020 results in the training set, plus 2016 results, some demographic data, and covid, will the model be able to predict the 2020 result for the rest of the test set? How big is the Root mean square error compared to the actual result?

U431888754 tigeryi@bu.edu
Sicheng Yi (Tiger Yi)

(3) Better documentation of the report of what is the process of me getting to the final result? The process is very messy as I try to figure things out. But in the end, I need to file a readable report of what I have done so far in the project.