FinalReport

TBD

2025-11-04

R Markdown

Title: A Predictive Model of USA Presidential Election by States.

Abstract:

In this project, we aim to create a predictive model based on State data that aims to accurately predict the party that a state will vote for in the presidential electionbased on demographic data and past voting history. Our question is "Are we able to somewhat accurately predict a states voting tendency based on demographic data using logistic regression? If so, which variables are most important to the model?". **Conclusion to be done**

Introduction:

The purpose of this analysis is to find certain how certain characteristcs in indivudals, and how their demographic background impacts their voting tendencies. By creating a succesful predictive model, one is able to act on these predictions, wether it is to change them or to ensure that they happen. Knowing which states have a chance of fillping to either political party is also extremely important, as it is in these states that presidential candidates have the hardest battles. Not only can this benefit politicians, but also the voters, as the model highlights what factors are more imporant to the way in which they vote, and will therefore force these politicians to adress these factors. If we were to find that unemployment rate seems to be heavily associated with a voter voting for the Democratic party, the democratic party should aim to adresss their constituents and lower unemployment.

Data:

The data available is mostly demographic data relating to the 50 states, meaning that there are 50 observations (plus territories and overall USA data that will not be used). The datasets are as follows:

Demographic Data: (https://www.kff.org/state-health-policy-data/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)

Income Data: (https://hdpulse.nimhd.nih.gov/data-portal/social/table?age=001&age_options=ageall_1&demo=00011&demo_options=income_3&race=00&race_options=race_7&sex=0&sex_options=sexboth_1&socialtopic=030&socialtopic_options=social_6&statefips=00&statefips_options=area_states)

Unemployment Rates: (https://www.bls.gov/web/laus/laumstrk.htm)

Commuter Mode: (https://www.bts.gov/browse-statistical-products-and-data/state-transportation-statistics/commute-mode)

Gun Ownership: (https://worldpopulationreview.com/state-rankings/gun-ownership-by-state)

Covid Deaths: (https://catalog.data.gov/dataset/provisional-covid-19-death-counts-rates-and-percent-of-total-deaths-by-jurisdiction-of-res)

In order to clean and aggregate the data, we performed a few transformations. Most importantly, for the commuter mode data set, the rows were formatted as (State_Name_Commuter_Mode), where it was

necessary to modify it so that the state became the row on its own and the commuter mode into a column for every state. All conversions were made so that we could have the data for each state.

The following are the variables for each of our observations:

[1] “Location” “White”
[3] “Black” “Hispanic”
[5] “Asian” “American Indian or Alaska Native”
[7] “Native Hawaiian or Pacific Islander” “Multiple Races”
[9] “Total” “Bicycle”
[11] “Walked” “Taxi, motorcycle, or other”
[13] “Public transportation” “Worked at home”
[15] “Carpool” “Drove alone”
[17] “stateFlagCode” “GunOwnership_PercentageOfHouseholdsThatOwnGuns_pct_2022” [19] “GunOwnership_NumOfGunLicenses_num_2022” “data_as_of”
[21] “Group” “data_period_start”
[23] “data_period_end” “COVID_deaths”
[25] “COVID_pct_of_total” “pct_change_wk”
[27] “pct_diff_wk” “crude_COVID_rate”
[29] “aa_COVID_rate” “crude_COVID_rate_ann”
[31] “aa_COVID_rate_ann” “footnote”
[33] “end_date” “Unemployment.Rate.August.2025”
[35] “Median.Household.Income” “Voting.Results.2016”
[37] “Voting.Results.2020” “Voting.Results.2024”
[39] “Electoral.College.Votes”

Further data cleaning will be necessary in order to ensure that only the most influential variables are chosen for our logisitic regression model.

Visualization: Provide preliminary visualization in the form of histograms, density plots, scatter plots, boxplots, and numerical summaries of the data, depending on the type of analysis. What do these visualizations suggest about the hypothesis, model or simulation you are aiming to perform? If you are interested in including further visualization, describe them here.

Analysis:

We plan to implement a logistic regression model, while also possibly implemeting a K-Nearest Neighbors model for comparison of performance. We plan to use backwards selection in order to implement and discern only the most influential variables to our model. This way, not only can we build the most accurate model, but we can find the issues and demographics that are most important to how a state might vote. We expect to find that demographic variables such as race and median household income will play a significant role in the way a state votes, it is often that class and ethnic backgrounds demonstrate an affinity for either political party.