Erdos2024 Voter Prediction Project

Voting Time vs. Voter Turnout

By: Avi Steiner, Chase Kimball, Davis Stagliano

We also acknowledge early efforts of Brian Hepler on voter turnout before he left the program.

Problem Statement:

The problem being explored is to use demographic and geographic data to predict voter turnout, in order to see what variables are indicators of voting access and activity.

Summary

We restricted our analysis to just the city of Chicago, and pulled demographic data from the US Census Bureau, voting data from the Illinois Board of Elections, geographical precinct data from the City of Chicago, polling station data from the Center for Public Integrity, and transit times from the Google Maps API. Our baseline model was a simple average of voter turnout across precincts in our training set. The vast amount of our time and effort was spent on sourcing, cleaning, and engineering data. For our analysis, we compared linear and logistic regressions, as well as an XGBoostRegressor ensemble method. Although the logistic and linear regression models perform similarly (and both outperform XGBoost), the logistic model is "philosophically" better. In particular, it reflects the nature of the problem (i.e. how probable is it that a person in a precinct will vote), and it prevents us from predicting voter turnout percentages >100% or <0%. Our results also indicate that features like level of educational attainment and income are the most important in predicting voter turnout, as opposed to polling accessibility.

While this document provides an overview, we point the reader to our notebooks for more detail. In particular: - Our main analysis on the processed and cleaned data can be found in analysis.ipynb - Our significant data processing can be found in project_preproc/ - preproc_1_census_data_chicago.py pulls down and processes census data for the City of Chicago from the 2010 census - preproc_2_voting_turnout.py processes precinct returns for the City of Chicago and computes a voter turnout percentage - preproc_3_combining_with_precinct_data.py combines all of the above with the geographical precinct data, converts census tract-wise statistics into precinct-wise form, and pulls transit times to polling centers from the googlemaps API - reading_data.py contains utility functions used in the last script - EDA_0.ipynb contains old and eventually unused analyses from back when the project had a different direction and was focused on this paper on using persistent cohomology to identify regions with low access to polling locations - EDA.ipynb contains vital data exploration we did. There you can find demonstrations of our population averaging scheme, initial plots of census tract data, and a confirmation from using a secondary source that anomylous precinct returns (above 100%) were actually in the data. Note that it is expected that this is due to same-day voter registration. Given the timing of the project (2024 elections) we did not hear back from the Chicago Board of Elections about our inquiry into the matter.

1. Potential Stakeholders

- **Election Authorities and Government Agencies**: Local election boards, state/federal commissions, and voter outreach offices interested in optimizing voter turnout.
- **Policymakers and Legislators**: City officials and state legislators focused on shaping policies to improve voter access and turnout.
- **Civil Rights and Advocacy Organizations**: Groups like the ACLU and NGOs advocating for voting rights, aiming to address disparities in voter access.
- **Community Leaders and Activists**: Grassroots activists and civic organizations focused on voter mobilization in underserved areas.
- **Academic and Research Institutions**: Researchers and think tanks studying voting behavior and policy solutions related to voter access.

2. Key Performance Indicators (KPIs):

- Root mean-squared error for predicted average voter turnout per precinct.
- F-scores to identify variables useful in predicting voter turnout
- Geographic Distribution of Voter Turnout: Visualization of turnout rates relative to polling site coverage.
- Polling Site Access: Measured as average travel time or distance to the nearest polling site.

3. Dataset Identification:

Our data is sourced from a variety of public sources: - **Census Data**: This includes statistics on the level of education, age, race, income level, home-to-work transit method, and employment from the Census Bureau - **Voter Turnout Data**: Voter turnout rates in the 2016 general election by precinct, sourced from the Illinois Board of Elections - **Precinct Maps**: Boundaries of the 2069 precincts that compose the City of Chicago under the 2012-2022 districting. In 2022 the City was redistricted, reducing the number of precincts greatly. This data was sourced from the City directly - **Polling locations**

4. Dataset Description and Problem Statement:

Dataset Description:

The dataset is composed of three primary components: 1. **Polling Site Accessibility**: This dataset includes polling station locations and travel times (using Google Maps API or other transportation data). 2. **Voter Turnout Data**: Voter turnout data includes information about how many eligible voters actually cast their votes in elections. The data is aggregated at the precinct level, providing turnout rates as a percentage of registered or eligible voters. 3. **Demographic Data**: Demographic information includes socioeconomic and population data from the US Census Bureau. These details provide context for voter eligibility and participation and include factors like income, race, education, and population density.

Discussion of Dataset Issues:

1. The data from the Census had multiple missing or invalid fields.
    - For average household income, if the variance of the data for the tract was larger than the average of the tract, roughly -$6,000,000 was entered instead. We resolved this by cleaning these rows from our data set
    - In total, only a handful of the 2000+ precincts exclusively overlapped with census tracts with missing data. Those that overlapped with a mix of tracts with and without data had their demographics inferred from only those tracts with data.
2. The formatting of the Census was far too granular.
    - While a few, important, data fields were collated into a total number, most were left in extremely granular forms. For example, there is no topline result for the number of people in a tract with a bachelor's degree. That data is only available broken down first by sex and then by age. Finding the total number of bachelor's degrees in a tract required summing over ten non-consecutive fields. This had to be repeated for every education level reported in the census.
3. Census data is reported by Census tract, which is a different partition of Chicago than the precincts which report voting results
    - We opted to perform a population-based weighting scheme. This is done by finding the intersection of each precinct with each census tract, assuming each census tract is of constant population density, and then getting a population-weighted average of each statistic across the tract intersections that make up each precinct.
4. Finding exact data on the eligible voting population, especially broken down enough to be of use, was not feasible in our timeframe.
    - As such, we did our models based on total population. While this will certainly have an impact on our data, our hope is that the impact is roughly equal across all precints, and that in comparing them, the effect will be nullified. The only precint where this assumption likely does not hold is home to the Cook County Correctional Facilities, which has an exceptionally large inmate population. We left it in our analysis, but better voter eligible data would likely have an impact on that precinct.
5. Some precincts reported above 100% turnout. It is expected that this is due to same-day voter registration. Given the timing of the project (2024 elections) we did not hear back from the Chicago Board of Elections about our inquiry into the matter. See EDA.ipynb for a confirmation that the anomaly is really in the data by computing it from a second source.

5. Analysis

We ran the following models, did hyperparameter tuning, compared against baselines, and explored feature importance. Please seee our analysis notebook for details. Please also see EDA_0.ipynb and EDA.ipynb for additional preliminary data exploration.

- Average voter turnout for Chicago in our training set was roughly 71%. This was used as our baseline model. The baseline model had an RMSE of about 9.5%
- Linear Regression
- Logistic Regression
- XGBoost

6. Conclusion and Next Steps

Comparing to our baseline, we found the linear and logistic regressions improved the RMSE by about 45%. The XGBoost tree only improved the RMSE by 42%. All models agreed on which factors were the most important, namely level of education, followed by income, with populations with higher degrees of education and more income having higher voter turnout. While racial identity had some correlation, it was varied, and travel time to vote had small coefficients all around.

The next step would be to expand the model into other major cities for the 2016 election, and then to compare with the 2024 election, once that data is available. Other cities, such as LA and NYC, with know and serious traffic problems, might yield different results. It is entirely possible the relationship between voting time and turnout depends on the city. Additionally, since the pandemic changed how people can and do vote, it is possible that travel time is even less important now in 2024, correlating with the rise in mail in voting options.