Voting: Demographics & Outcomes

A Data Science Bootcamp Project
Erdos Institute Summer 2024
Srijan Ghosh, Giovanni Passeri,
Yu (Alicia) Xiao, Li Meng

Overview

Question: Can we predict voting outcomes from demographic data? What factors should we consider? Which factors exert more influence?

Goal: Investigating and understanding possible factors influencing voting outcomes of the United States 2020 Presidential Election

Industry Use: Targeted Campaigning; Aid in policy-making

Data: American Community Survey 5-Year Database (2015-2020); Election result data from MIT Election Lab

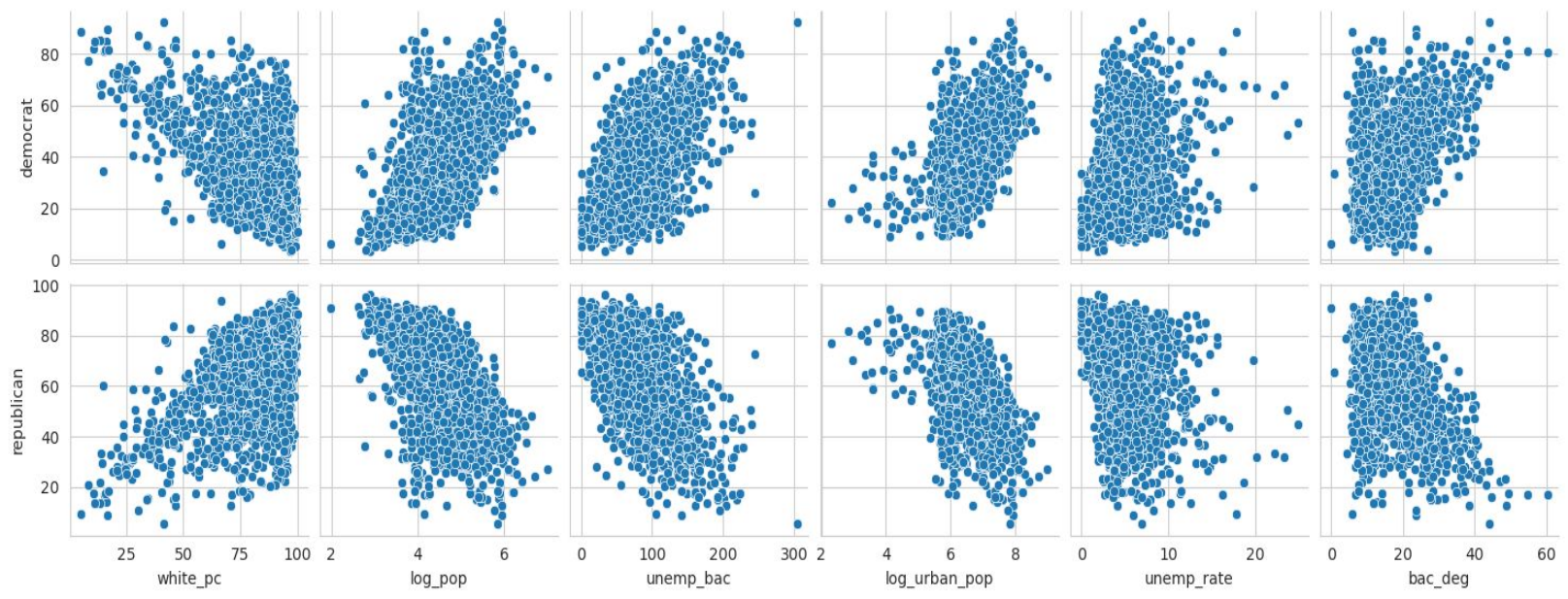
Data Cleaning and Processing

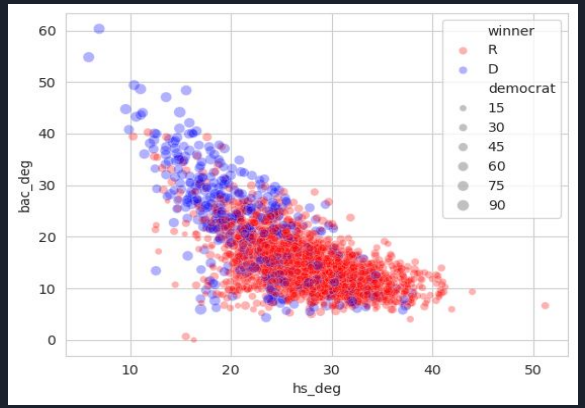
Did You Know:
Loving County, TX
677 sq mi; 64 people
No unemployment!

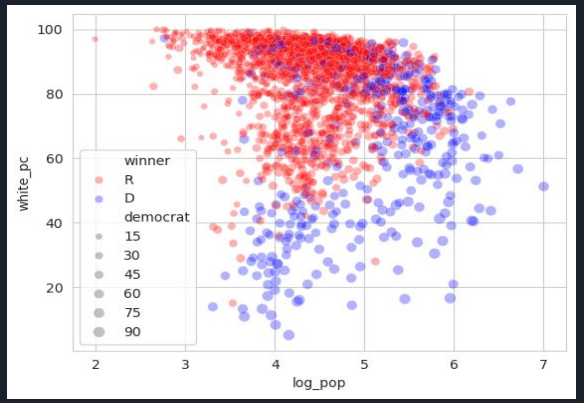
Population
Ethnicity
Poverty
Education
Urban - Rural
Age Percentage

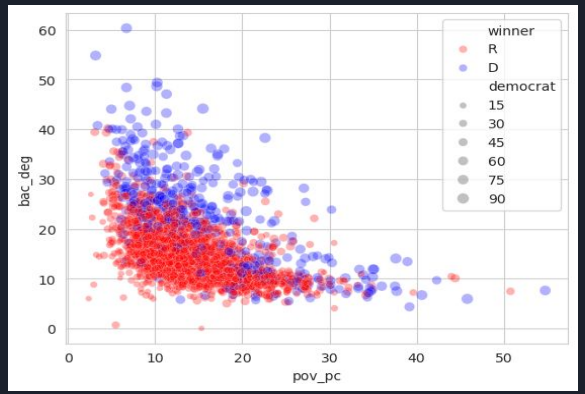
- log_pop: Logarithm (base 10) of the total population.
- white_pc: People living in the county belonging to ethnicity white (this is highly correlated to people of other ethnicities, hence we only consider white_pc as a feature).
- urban: People living in urban areas (equal to 100 - people living in rural areas)
- log_urban_pop: Logarithm of the total urban population
- log_unemp_rate: Logarithm of the unemployment rate in the county
- bac_deg: Percentage of adults holding a bachelor degree or higher
- hs_deg: Percentage of adults holding with a high school degree as highest educational attainment
- pop_18_30_pc: Percentage of young people (18 to 30 years of age)
- pop_60_up_pc: Percentage of older people (60 years or older)
- pov_pc: Percentage of people living under the poverty limit
- unemp_bac: Product of unemp_rate and bac_deg

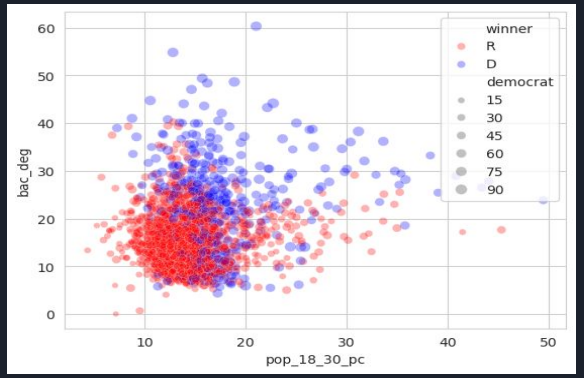
Exploratory Data Analysis




bac_deg
hs_deg
winner
R
D
democrat
15
30
45
60
75
90


white_pc
log_pop
winner
R
D
democrat
15
30
45
60
75
90

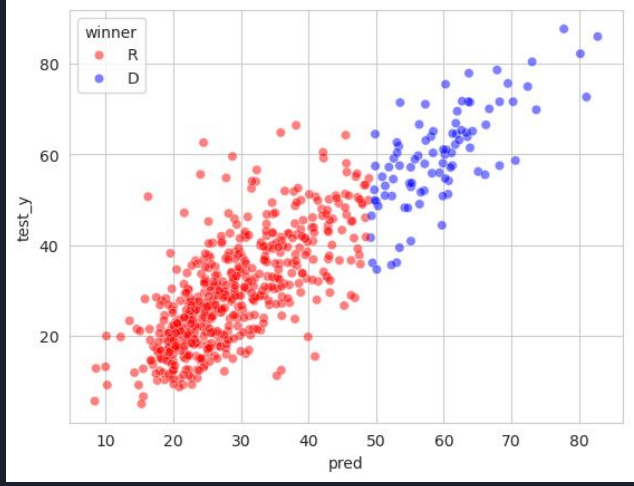

bac_deg
pov_pc
winner
R
D
democrat
15
30
45
60
75
90


bac_deg
pop_18_30_pc
winner
R
D
democrat
15
30
45
60
75
90

Modeling-Regression

TO PREDICT VOTING RESULT PERCENTAGE

Model	RMSE	R^2
Baseline	15.96	0
Linear Reg	9.19	0.5
Random Forest	9.15	0.33
XGBoost	8.21	0.63
Neural Network	8.27	0.63



After final training and testing on the initially split dataset, XGBoost yields a RMSE of 8.35 and a R^2 of 0.63.

Modeling-Classification

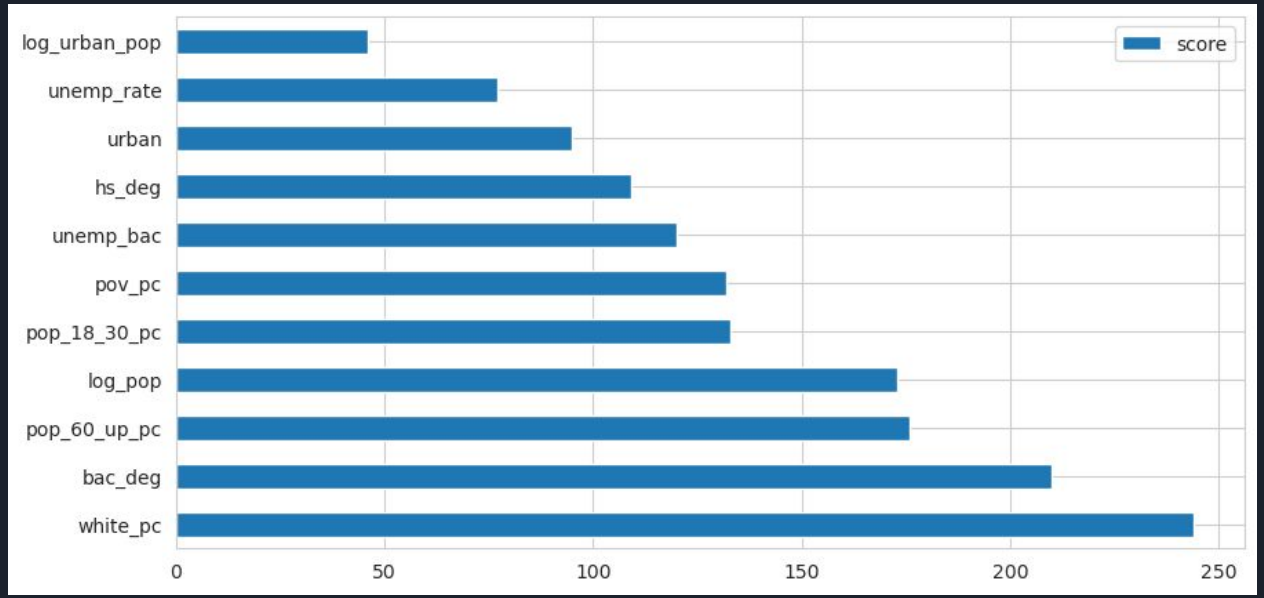
TO PREDICT VOTING RESULT DEMOCRATIC VS REPUBLICAN

Model	Accuracy Score
Baseline	0.7127
Logistic Regression	0.9061
Random Forest	0.9226
SVC	0.9266

After final training and testing, our classifier of choice SVC has an accuracy score of 0.9197

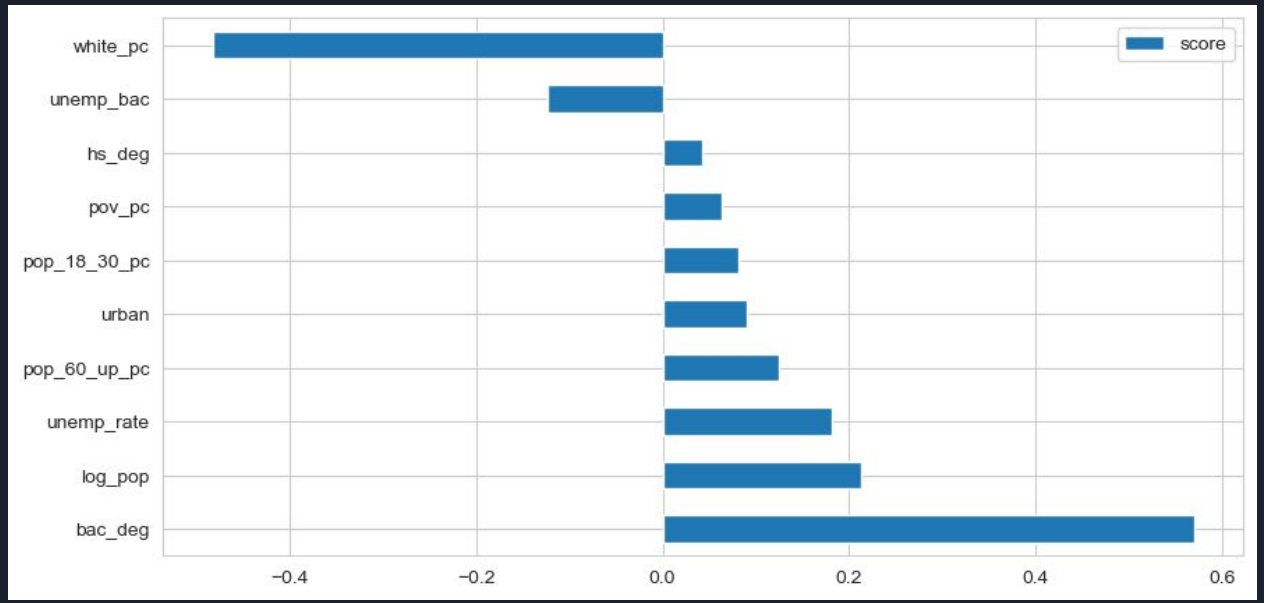
Feature Importance

XGBoost (fitted for Democratic votes percentage)



Feature Importance

LINEAR REGRESSION (fitted for Democratic votes percentage)



Future Investigations

- Suburban
- Religion
- Voter Turnout

  Council meeting involvement, political rally turnout, etc

- Sentiment Analysis

  Local News, Twitter/social media, new policy reception