New Voter Insights and Analysis

Joshua Kurdys
github.com/jkurdys/new_voter_insights

April 2, 2021

source

About Me

- This is my normal office attire and I don’t know these people
- When I’m not doing data science, I enjoy cosplay with my family
- Prior to data sciencing, I repaired elevators and studied philosophy (in that order)

Motivation

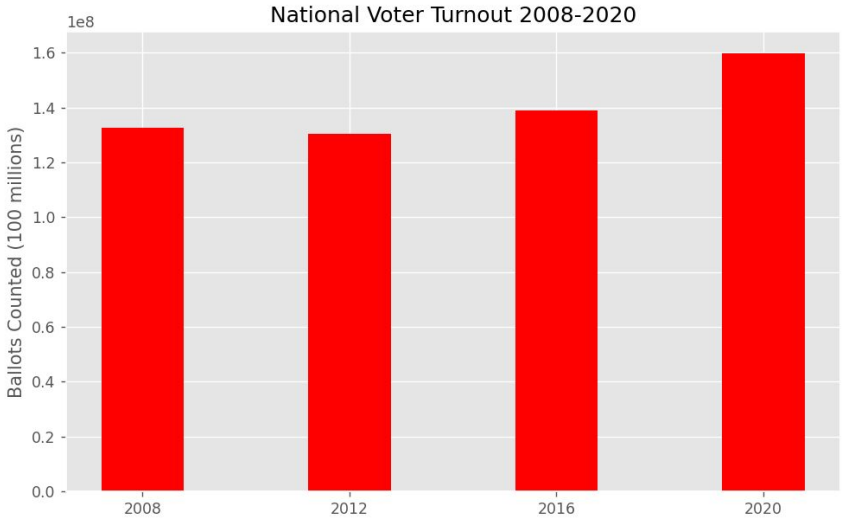
source

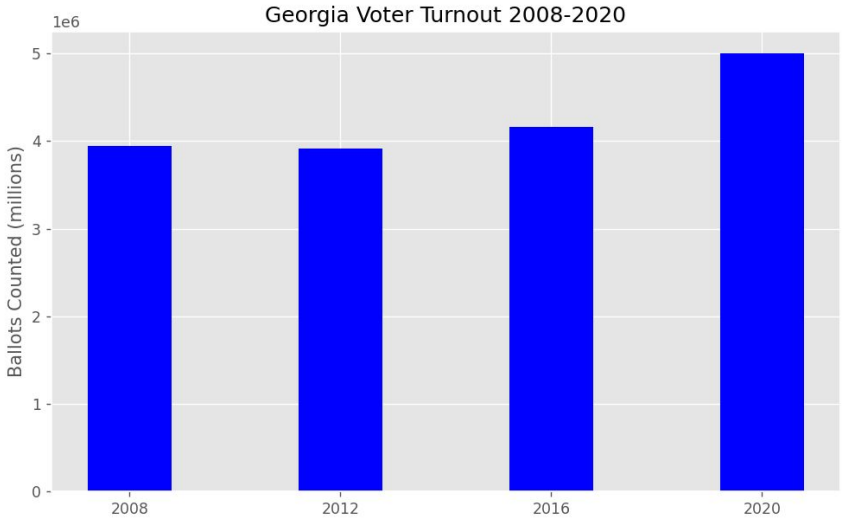
Given the historic turnout of the 2020 General Election, I wanted to isolate the features of new voters so voter registration efforts might better understand the types of individuals who fit that profile.

Using inferential logistic regression on voter registration records, **this project aimed to identify the features of voters that could confidently distinguish new voters from the rest of the electorate.**

EDA

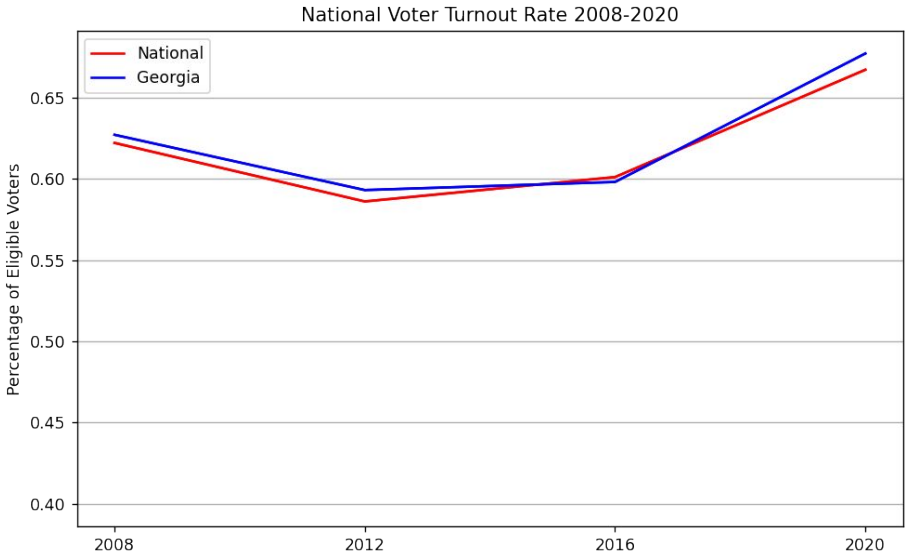
- Turnout up in recent elections at both national and state level
- Georgia is a suitable case study that mirrors national trends in turnout





EDA

- Trends for turnout rates approximate trends for ballots counted
- Georgia mirrors national trends with slight differences
- In the last four election cycles, higher than national average Georgia turnout corresponds to Democratic Presidential victories



Results

- Logistic regression analysis of the data shows no significance for the majority of features
- Predictably, voter status and age were significantly associated with voter registration
- Unfortunately, these trivial predictors contribute little to the model's predictive power

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.3755	9.31e+04	-4.7e-05	1.000	-1.82e+05	1.82e+05
C(voter_status)[T.1]	5.5953	0.147	37.940	0.000	5.306	5.884
C(AI)[T.1]	-0.4067	6.13e+04	-6.63e-06	1.000	-1.2e+05	1.2e+05
C(AP)[T.1]	-0.1788	6.11e+04	-2.92e-06	1.000	-1.2e+05	1.2e+05
C(BH)[T.1]	-0.8289	6.24e+04	-1.33e-05	1.000	-1.22e+05	1.22e+05
C(HP)[T.1]	-0.4579	6.2e+04	-7.39e-06	1.000	-1.21e+05	1.21e+05
C(OT)[T.1]	-0.3711	6.22e+04	-5.97e-06	1.000	-1.22e+05	1.22e+05
C(U)[T.1]	-0.9155	6.16e+04	-1.49e-05	1.000	-1.21e+05	1.21e+05
C(WH)[T.1]	-0.8654	6.18e+04	-1.4e-05	1.000	-1.21e+05	1.21e+05
C(F)[T.1]	-0.8820	9.1e+04	-9.69e-06	1.000	-1.78e+05	1.78e+05
C(M)[T.1]	-0.7163	9.17e+04	-7.81e-06	1.000	-1.8e+05	1.8e+05
C(O)[T.1]	0.7206	9.15e+04	7.88e-06	1.000	-1.79e+05	1.79e+05
C(rural)[T.1]	-1.3781	1.48e+05	-9.34e-06	1.000	-2.89e+05	2.89e+05
C(urban)[T.1]	-1.4026	1.48e+05	-9.5e-06	1.000	-2.89e+05	2.89e+05
C(military)[T.1]	-1.1956	1.47e+05	-8.11e-06	1.000	-2.89e+05	2.89e+05

Dep. Variable:	y	No. Observations:	6181531
Model:	Logit	Df Residuals:	6181500
Method:	MLE	Df Model:	30
Date:	Thu, 01 Apr 2021	Pseudo R-squ.:	0.04466
Time:	13:49:44	Log-Likelihood:	-8.9655e+05
converged:	False	LL-Null:	-9.3846e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

Results

- Although selecting only the trivial features improved p-values, it had no impact on prediction
- Other reconfigurations of the feature set had little impact

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	6181531
Model:	Logit	**Df Residuals:**	6181527
Method:	MLE	**Df Model:**	3
Date:	Thu, 01 Apr 2021	**Pseudo R-squ.:**	0.03627
Time:	15:16:05	**Log-Likelihood:**	-9.0442e+05
converged:	True	**LL-Null:**	-9.3846e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.6516	0.148	-51.858	0.000	-7.941	-7.362
C(voter_status)[T.1]	5.5735	0.147	37.794	0.000	5.285	5.863
C(rural)[T.1]	0.0157	0.005	3.154	0.002	0.006	0.025
age	-0.0278	0.000	-204.667	0.000	-0.028	-0.028

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	6181531
Model:	Logit	**Df Residuals:**	6181527
Method:	MLE	**Df Model:**	3
Date:	Thu, 01 Apr 2021	**Pseudo R-squ.:**	0.0002149
Time:	15:48:58	**Log-Likelihood:**	-9.3826e+05
converged:	False	**LL-Null:**	-9.3846e+05
Covariance Type:	nonrobust	**LLR p-value:**	3.980e-87

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.3818	3.31e+04	-7.2e-05	1.000	-6.48e+04	6.48e+04
C(rural)[T.1]	-0.9519	3.31e+04	-2.88e-05	1.000	-6.48e+04	6.48e+04
C(urban)[T.1]	-0.9339	3.31e+04	-2.82e-05	1.000	-6.48e+04	6.48e+04
C(military)[T.1]	-0.5653	3.31e+04	-1.71e-05	1.000	-6.48e+04	6.48e+04

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	6181531
Model:	Logit	**Df Residuals:**	6181523
Method:	MLE	**Df Model:**	7
Date:	Thu, 01 Apr 2021	**Pseudo R-squ.:**	0.005159
Time:	15:32:25	**Log-Likelihood:**	-9.3362e+05
converged:	False	**LL-Null:**	-9.3846e+05
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.6436	4.47e+04	-5.92e-05	1.000	-8.75e+04	8.75e+04
C(AI)[T.1]	-0.1697	4.47e+04	-3.8e-06	1.000	-8.75e+04	8.75e+04
C(AP)[T.1]	-0.0464	4.47e+04	-1.04e-06	1.000	-8.75e+04	8.75e+04
C(BH)[T.1]	-0.6874	4.47e+04	-1.54e-05	1.000	-8.75e+04	8.75e+04
C(HP)[T.1]	-0.1258	4.47e+04	-2.82e-06	1.000	-8.75e+04	8.75e+04
C(OT)[T.1]	-0.0954	4.47e+04	-2.14e-06	1.000	-8.75e+04	8.75e+04
C(U)[T.1]	-0.7726	4.47e+04	-1.73e-05	1.000	-8.75e+04	8.75e+04
C(WH)[T.1]	-0.7736	4.47e+04	-1.73e-05	1.000	-8.75e+04	8.75e+04

Conclusions and Next Steps

- Regression analysis failed to identify any connection between the geographic and demographic data found in Georgia Voter Registration Records
- No geographic or demographic features of this dataset seem to be overrepresented among new voters
- An expanded dataset of the entire voting age population may isolate the features of new voters from the general population
- Utilizing the .geojson files in the dataset could provide an avenue for utilizing more granular geographic data

Questions?

Joshua Kurdys

jkurdys@gmail.com
LinkedIn
Github

Tech Stack:

- Python (Numpy, Pandas, Matplotlib, Scikit-learn)
- SQL (PostgreSQL)
- Linux (Terminal)
- Git (Github)

Appendix

Data Preparation:

- Null values and missing data
- Uninterpretable “object” data types
- Redundant categories
- One-hot encoding

Before:

#	Column	Non-Null Count	Dtype
1	land_district	62601	object
2	land_lot	62597	object
3	status_reason	59	object
4	county_districta_name	4906	object
5	county_districta_value	4906	object
6	county_districtb_name	26491	object
7	county_districtb_value	26491	object
8	city_school_district_name	24949	object
9	city_school_district_value	24949	object
10	city_dista_name	0	float64
11	city_dista_value	0	float64
12	city_distb_name	0	float64
13	city_distb_value	0	float64
14	city_distc_name	0	float64
15	city_distc_value	0	float64
16	city_distd_name	0	float64
17	city_distd_value	0	float64
18	party_last_voted	1666	object

After:

#	Column	Non-Null Count	Dtype
1	voter_status	6183143	int64
2	AI	6183143	int64
3	AP	6183143	int64
4	BH	6183143	int64
5	HP	6183143	int64
6	OT	6183143	int64
7	U	6183143	int64
8	WH	6183143	int64
9	F	6183143	int64
10	M	6183143	int64
11	O	6183143	int64
12	rural	6183143	int64
13	urban	6183143	int64
14	military	6183143	int64
15	cd_1	6183143	int64
...	...	...	...
29	cd_99999	6183143	int64
30	age	6183143	int64

Appendix

Assumptions of logistic regression

- Binary dependent variable
- Independent observation
- No collinearity
- Linearity of continuous independent variables
- Large sample size