Turnout Modeling for Alamance County

Cara Nix

Goal

Predict whether an individual will vote in the 2020 General Election in Alamance County, North Carolina.

Data Exploration

On NC website:

- Voter Files (with active registrants, 2/year, one on election day in Nov and one on Jan 1st).
- Voter History Files (vote history going back through 2015).

Downloaded all voter files since 2008 -- quickly realized that was too massive of an undertaking and **vote history data** only went back to 2015.

Decision point: I chose to train my model on 2016 active registrants even though a 2018 file does exist. This is because Presidential electorates look different than Midterm electorates.

Data Questions

How many active people on file in 2016? **~87k**

How many active people on file in 2020? **~120k**

How many people join the file between 2016 and 2020? **~35k**

How many people leave the file? **~20k**

Quick Data Analysis: Race

2016 Voter File (for training)

Race	% of File
White	72%
Black	20%
Undesignated	4%
Other	2%
Two or More Races	<1%
Asian	<1%
American Indian	<1%

2020 Voter File

Race	% of File
White	65%
Black	19%
Undesignated	11%
Other	2%
Two or More Races	<1%
Asian	<1%
American Indian	<1%

Quick Data Analysis: Sex

2016 Voter File (for training)

Sex	% of File
Female	54%
Male	43%
Undesignated	2%

2020 Voter File

Sex	% of File
Female	50%
Male	41%
Undesignated	9%

Quick Data Analysis: Age

2016 Voter File (for training)

Age	% of File
41-65	45%
65+	23%
26-40	20%
18-25	12%

2020 Voter File

Age	% of File
41-65	41%
65+	23%
26-40	21%
18-25	14%

Quick Data Analysis: Party

2016 Voter File (for training)

Party	% of File
Democrat	39%
Republican	34%
Unaffiliated	27%
Libertarian	<1%

2020 Voter File

Party	% of File
Democrat	35%
Republican	33%
Unaffiliated	30%
Libertarian +Other	<1%

Modeling Plan (in a nutshell!)

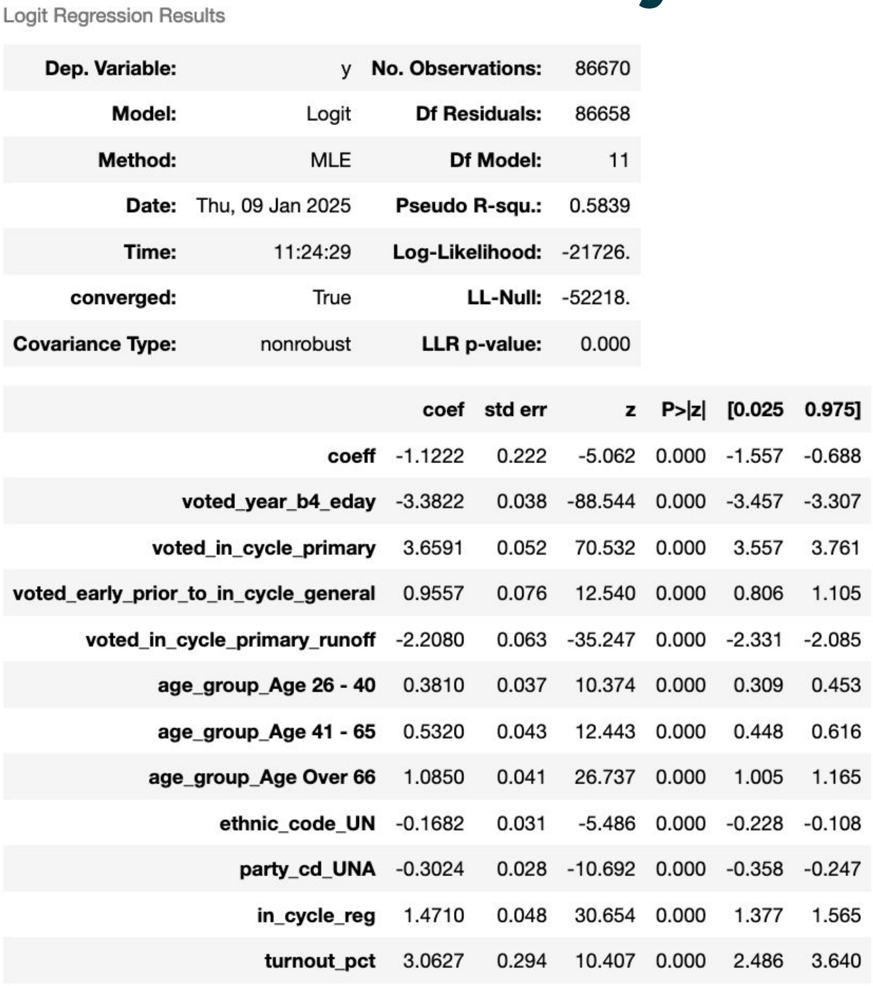
Idea is to use logistic regression to train model on the 2016 voter file using core demographics provided, i.e. race, sex, age group, political party + including data from the vote history file; picking features to include in our model that are likely to impact voter turnout.

Then, we will apply the model to the 2020 voter file to predict whether or not a person in Alamance county, NC will vote in the general election!

Model Features

- **Race**
- **Age**
- **Gender**
- **Political Party**
- **In-cycle registration:** Did registrant register in the year of the Presidential election? Shows recent excitement to vote.
- **Previously Vote Early:** Did registrant vote early before the General election? Shows inclination to vote.
- **Voted in election in previous year:** In previous election cycle (off-cycle) did registrant vote in election? Shows past history of voting in off-cycle elections.
- **Past-cycle precinct-level Turnout:** What was turnout in registrants’ precinct from past Presidential election? If precinct has high turnout individual more likely to turnout.

Model v3 Summary



Logit Regression Results

Dep. Variable:	y	**No. Observations:**	86670
Model:	Logit	**Df Residuals:**	86658
Method:	MLE	**Df Model:**	11
Date:	Thu, 09 Jan 2025	**Pseudo R-squ.:**	0.5839
Time:	11:24:29	**Log-Likelihood:**	-21726.
converged:	True	**LL-Null:**	-52218.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
coeff	-1.1222	0.222	-5.062	0.000	-1.557	-0.688
voted_year_b4_eday	-3.3822	0.038	-88.544	0.000	-3.457	-3.307
voted_in_cycle_primary	3.6591	0.052	70.532	0.000	3.557	3.761
voted_early_prior_to_in_cycle_general	0.9557	0.076	12.540	0.000	0.806	1.105
voted_in_cycle_primary_runoff	-2.2080	0.063	-35.247	0.000	-2.331	-2.085
age_group_Age 26 - 40	0.3810	0.037	10.374	0.000	0.309	0.453
age_group_Age 41 - 65	0.5320	0.043	12.443	0.000	0.448	0.616
age_group_Age Over 66	1.0850	0.041	26.737	0.000	1.005	1.165
ethnic_code_UN	-0.1682	0.031	-5.486	0.000	-0.228	-0.108
party_cd_UNA	-0.3024	0.028	-10.692	0.000	-0.358	-0.247
in_cycle_reg	1.4710	0.048	30.654	0.000	1.377	1.565
turnout_pct	3.0627	0.294	10.407	0.000	2.486	3.640

Model Interpretation

Note: I do not go over the interpretation for each significant feature! The coefficients have been exponentiated to be interpreted via odds.

`turnout_pct:` A one-unit increase in historical precinct-level turnout percentage leads to a 21.4 increase in odds of turning out to vote.

`age_group_Age 26 - 40:` Being in age group 26-40 leads to a 1.5 increase in odds of turnout relative to the reference age category, 18-25 year olds.

`in_cycle_reg:` Registering in-cycle leads to a 4.4 increase in odds of turnout.

`voted_in_cycle_primary:` Voting in the in-cycle Presidential Primary leads to a 38.8 increase in odds of turnout.

`voted_early_prior_to_in_cycle_general:` Voting early in a different election prior to general election leads to a 2.6 increase in odds of turnout.

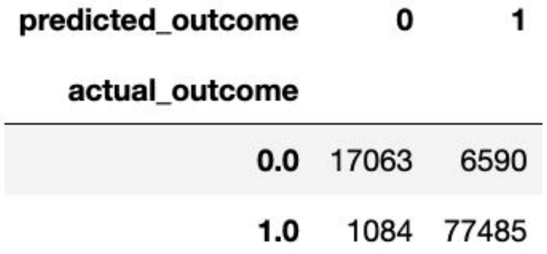
`voted_year_b4_eday:` Voting in the year before the general election leads to a 1-.03 **decrease** in odds of turnout.

`ethnic_code_UN:` having ethnic code unassigned leads to 1-.73 **decrease** in odds of turnout relative the reference category, Hispanic and Latino.

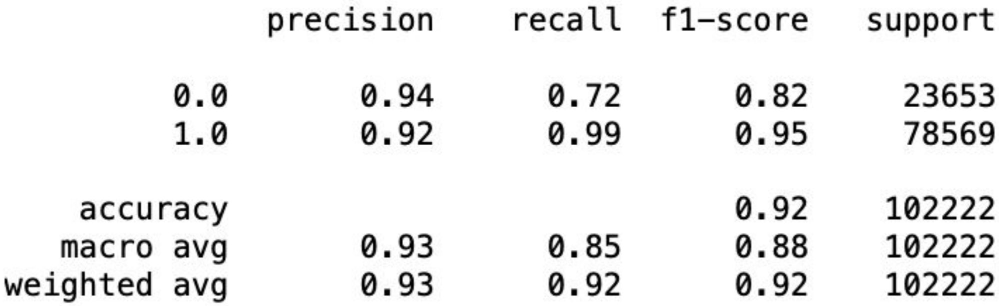
Most of these features match our previous expectations.

Evaluation of Model (on 2020 voter file!)

Note: This is the “best” version of the model which I chose to include. This model includes only the features which were significant from 2016 model output.



predicted_outcome	0	1
actual_outcome		
0.0	17063	6590
1.0	1084	77485

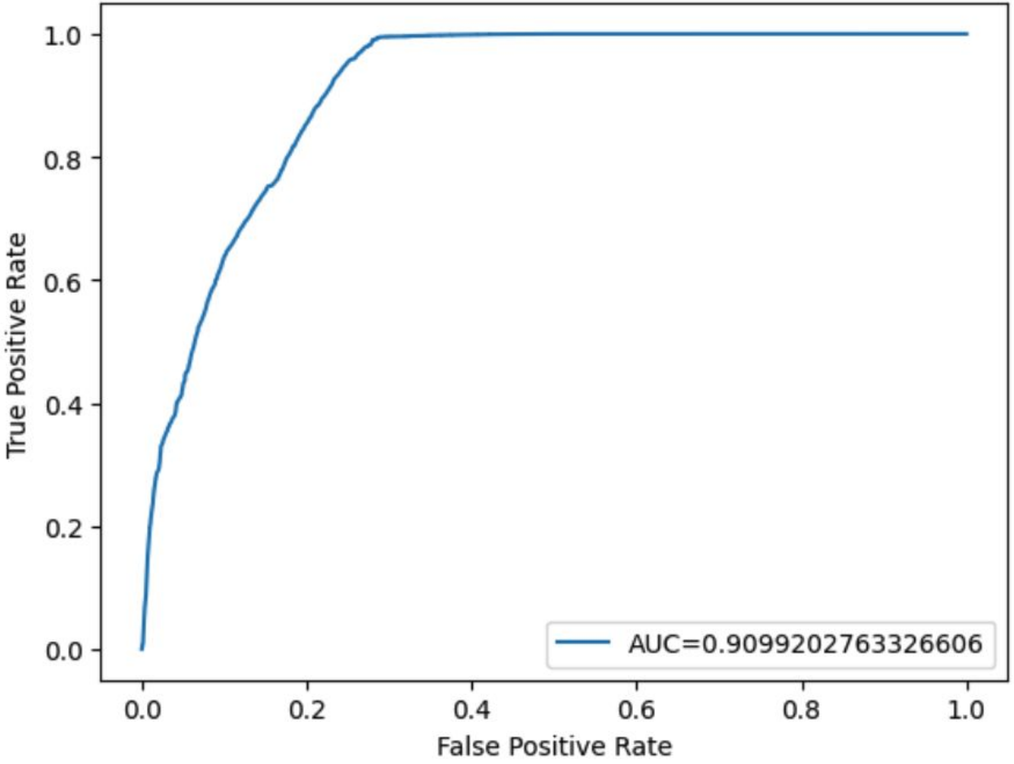


```
              precision    recall  f1-score   support

         0.0       0.94      0.72      0.82     23653
         1.0       0.92      0.99      0.95     78569

    accuracy                           0.92    102222
   macro avg       0.93      0.85      0.88    102222
weighted avg       0.93      0.92      0.92    102222
```

Looks like model is more prone to false positives than false negatives, i.e. we predict more people will vote that do not vote than we predict people will not vote who do. This is good for targeting voters, as we aren’t missing a lot of people expected to vote. However, we would still be allocating resources on (some) people who did not wind up voting.

Overall, model performs well!

Evaluation of Model v3: ROC and AUC



Model Performance by Subgroup

Note: Full confusion matrices and classification reports are available in code.

- Model performed better for Non-White registered voters (higher accuracy).
- Model performed slightly better for Non-Male registered voters (higher accuracy & recall).
- Model performed similarly for Democrats and Republicans.

Importantly, there were no drastic changes in precision or accuracy when split by subgroup, indicating the model is performing sufficiently across important demographic cuts.

Next Iteration Feature Ideas

Midterm Voter in Previous Cycle

Vote History from 2014 (unfortunately) isn’t available on the NC website, so I couldn’t include a variable for previously voted in most recent midterm.

- The expectation is that Midterm voting → Increased Odds of Turnout in Presidential

FEC Data

FEC Data from 2015-2016 and 2019-2020 is available, so could match based on zip code and first, middle and last name to voter file for a feature which is 1 if a registered voter made a political contribution and 0 otherwise.

- The expectation is that political contribution → Increased Odds of Turnout in Presidential

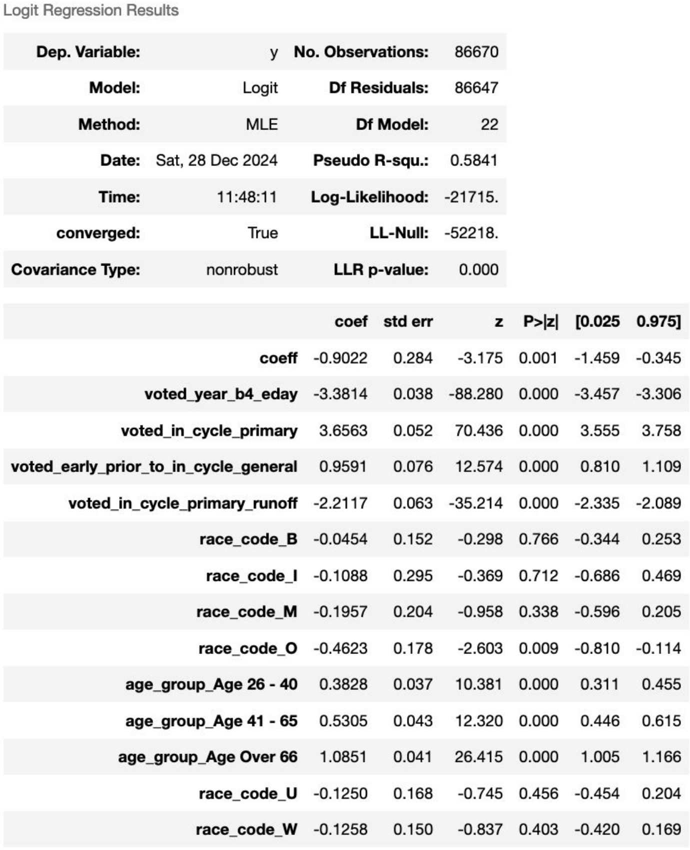
Include more vote history data

Further back, if possible! Also, I only used the vote history data from Alamance county, but could have matched to unique id’s statewide for vote history so if someone moved into Alamance county I’d have had their vote history.

Challenges I Faced

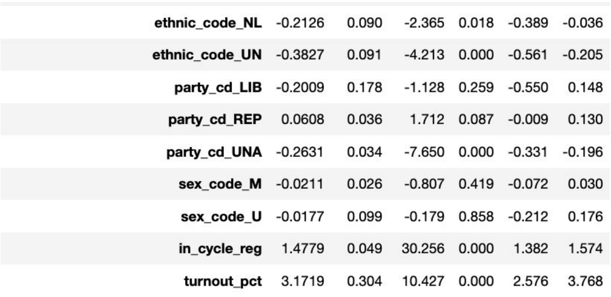
- For past-cycle precinct turnout feature the naming conventions changed between the precinct results and the voter file, so had to map the voter file names to the precinct result names. I used shapefile from NC to do so.
- Considered using Age as a continuous variable, but it performed better when bucketted.
- Reading in data needing “/t” as separator and “utf-16” as encoding was a joy.
- Mellville 4 Precinct! It didn’t exist one year and then did exist another, had to use average turnout for NA values.
- Surprisingly, the unique id’s never posed an issue which I was anticipating!

Full Model Output



Logit Regression Results

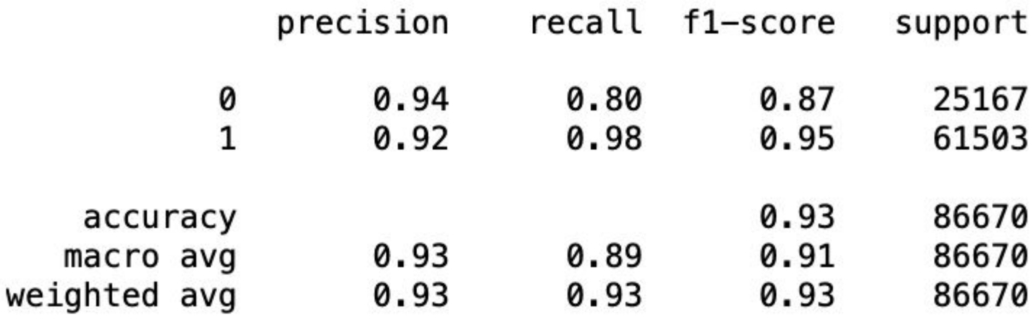
Dep. Variable:	y	**No. Observations:**	86670
Model:	Logit	**Df Residuals:**	86647
Method:	MLE	**Df Model:**	22
Date:	Sat, 28 Dec 2024	**Pseudo R-squ.:**	0.5841
Time:	11:48:11	**Log-Likelihood:**	-21715.
converged:	True	**LL-Null:**	-52218.
Covariance Type:	nonrobust	**LLR p-value:**	0.000



	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
coeff	-0.9022	0.284	-3.175	0.001	-1.459	-0.345
voted_year_b4_eday	-3.3814	0.038	-88.280	0.000	-3.457	-3.306
voted_in_cycle_primary	3.6563	0.052	70.436	0.000	3.555	3.758
voted_early_prior_to_in_cycle_general	0.9591	0.076	12.574	0.000	0.810	1.109
voted_in_cycle_primary_runoff	-2.2117	0.063	-35.214	0.000	-2.335	-2.089
race_code_B	-0.0454	0.152	-0.298	0.766	-0.344	0.253
race_code_I	-0.1088	0.295	-0.369	0.712	-0.686	0.469
race_code_M	-0.1957	0.204	-0.958	0.338	-0.596	0.205
race_code_O	-0.4623	0.178	-2.603	0.009	-0.810	-0.114
age_group_Age 26 - 40	0.3828	0.037	10.381	0.000	0.311	0.455
age_group_Age 41 - 65	0.5305	0.043	12.320	0.000	0.446	0.615
age_group_Age Over 66	1.0851	0.041	26.415	0.000	1.005	1.166
race_code_U	-0.1250	0.168	-0.745	0.456	-0.454	0.204
race_code_W	-0.1258	0.150	-0.837	0.403	-0.420	0.169
ethnic_code_NL	-0.2126	0.090	-2.365	0.018	-0.389	-0.036
ethnic_code_UN	-0.3827	0.091	-4.213	0.000	-0.561	-0.205
party_cd_LIB	-0.2009	0.178	-1.128	0.259	-0.550	0.148
party_cd_REP	0.0608	0.036	1.712	0.087	-0.009	0.130
party_cd_UNA	-0.2631	0.034	-7.650	0.000	-0.331	-0.196
sex_code_M	-0.0211	0.026	-0.807	0.419	-0.072	0.030
sex_code_U	-0.0177	0.099	-0.179	0.858	-0.212	0.176
in_cycle_reg	1.4779	0.049	30.256	0.000	1.382	1.574
turnout_pct	3.1719	0.304	10.427	0.000	2.576	3.768

Model Performance (on 2016 data)

Because logistic regression was used we can “test” on our training set. Here’s some quick metrics. Overall model does well (which is as anticipated).



```
              precision    recall  f1-score   support

           0       0.94      0.80      0.87     25167
           1       0.92      0.98      0.95     61503

    accuracy                           0.93     86670
   macro avg       0.93      0.89      0.91     86670
weighted avg       0.93      0.93      0.93     86670
```

Model Summary (trained on 2016 data)

note: This is not the full model output, these are just the features which had significant coefficients at .005 level, indicating that they had an impact on voter turnout. The coefficients have been exponentiated to be interpreted via log odds, if a coefficient is bolded this indicates it had a negative impact on turnout. Full model summary available at end of slides.

voted_year_b4_eday : 0.03

voted_in_cycle_primary : 38.7

voted_early_prior_to_in_cycle_general : 2.6

voted_in_cycle_primary_runoff : 0.10

age_group_Age 26 - 40 : 1.4

age_group_Age 41 - 65 : 1.6

age_group_Age Over 66 : 2.9

ethnic_code_UN : 0.68

party_cd_UNA : 0.76

in_cycle_reg : 4.3

turnout_pct : 23.8

Model Interpretation (2016)

Note: I do not go over the interpretation for each significant feature!

`turnout_pct:` A one-unit increase in historical precinct-level turnout percentage leads to a 23.8 increase in odds of turning out to vote.

`age_group_Age 26 - 40:` Being in age group 26-40 leads to a 1.4 increase in odds of turnout relative to the reference age category, 18-25 year olds.

`in_cycle_reg:` Registering in-cycle leads to a 4.3 increase in odds of turnout.

`voted_in_cycle_primary:` Voting in the in-cycle Presidential Primary leads to a 38.7 increase in odds of turnout.

`voted_early_prior_to_in_cycle_general:` Voting early in a different election prior to general election leads to a 2.6 increase in odds of turnout.

`voted_year_b4_eday:` Voting in the year before the general election leads to a .03 **decrease** in odds of turnout.

`ethnic_code_UN:` having ethnic code unassigned leads to .68 **decrease** in odds of turnout relative the reference category, Hispanic and Latino.

Most of these features match our previous expectations.