Texas 2018 Primary Models

Ben Gelman

Introduction

Knowing what level of turnout to expect in an election is important for campaign strategy and allocation of resources as well as for election-night coverage.

There are existing models of national or state-level election turnout for general elections, though those focused on primary races are rarer.

I propose a set of statistical models that can predict turnout for the Texas statewide primary elections (Gubernatorial and Senate races) in 2018, on the county level.

This is accomplished through linear regression and ensemble models (random forest, boosted trees, and extra trees) for both the Democratic and Republican primaries.

Background

Since 2010, Texas has run several primaries for Senator, Governor, and President, for both the Democratic and Republican parties.

With the exception of 2012, these elections were all held in early March.

In 2012, the primary was delayed to late May due to a redistricting dispute between the state legislature and several nonprofits (Fernandez).

Three of the competitors in the 2012 Presidential Primary, Rick Santorum, Mitt Romney, and Newt Gingrich.

Literature Review

The most relevant work is a 2020 presidential election turnout model created for Bloomberg News, by Dr. Andrew Therriault.

This model used data from the 2016, 2012, 2008, and 2004 presidential elections in order to predict turnout in 2020.

This model included past turnout, racial and economic data on the county level from the American Community Survey, and factors related to election competitiveness.

Therriault found that higher levels of past turnout, education, income, and homeownership predict primary turnout well.

Literature Review, cont.

A meta-analysis by Benny Geys, "Explaining voter turnout," found that past turnout and concurrent elections are statistically significant predictors of present turnout.

A 1985 study on Presidential primary turnout found that, on the state level, education was a strong predictor of turnout (Kenney, Rice 106-107).

Another paper on Senatorial primary turnout resulted in similar findings, with education and past turnout serving as significant variables in the model (Kenney 71).

A study in 2016 found that the presence of an incumbent also reduces turnout in a given election (Atkeson, Maestas 758).

Who Votes?

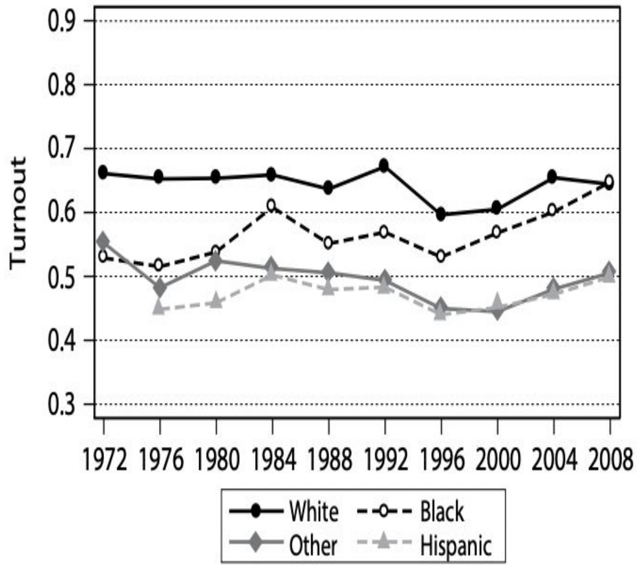


Figure 2.3. Turnout by Race, 1972–2008.

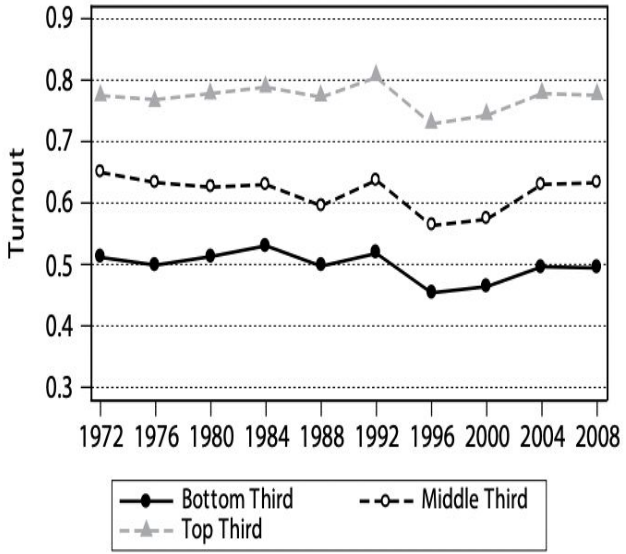


Figure 2.1. Turnout by Education, 1972–2008.

Data Sources

Census data includes ACS 5-year estimates from 2010, 2012, 2014, and 2016.

- The 5 year estimates collect data for the 5 years prior

Past election data is retrieved from the Texas Secretary of State website, with turnout defined as ratio of registered voters who cast a ballot.

Texas polling data comes from the Texas Politics Project at UT Austin.

National polling data is from Gallup.

Generic Ballot trends are from RealClearPolitics.

Census variables

‘male:’ and 'female': Percentage male and female

'White’, black,’ 'native', 'aapi: 'Hispanic', 'non-Hispanic White': Race by percentage

‘income': Mean household income

‘less than 10k','10-15k', '15-25', '25-35','35-50','50-75k','75-100', '100-150', '150-200','200+': income by bin

pct below PL': Percentage of population below poverty line

‘HH size’: Mean household size

‘hs',' college': Percentage with high school/college degree

'rentals', 'owner-occupied': Percentage of homes that are rented/owner-occupied

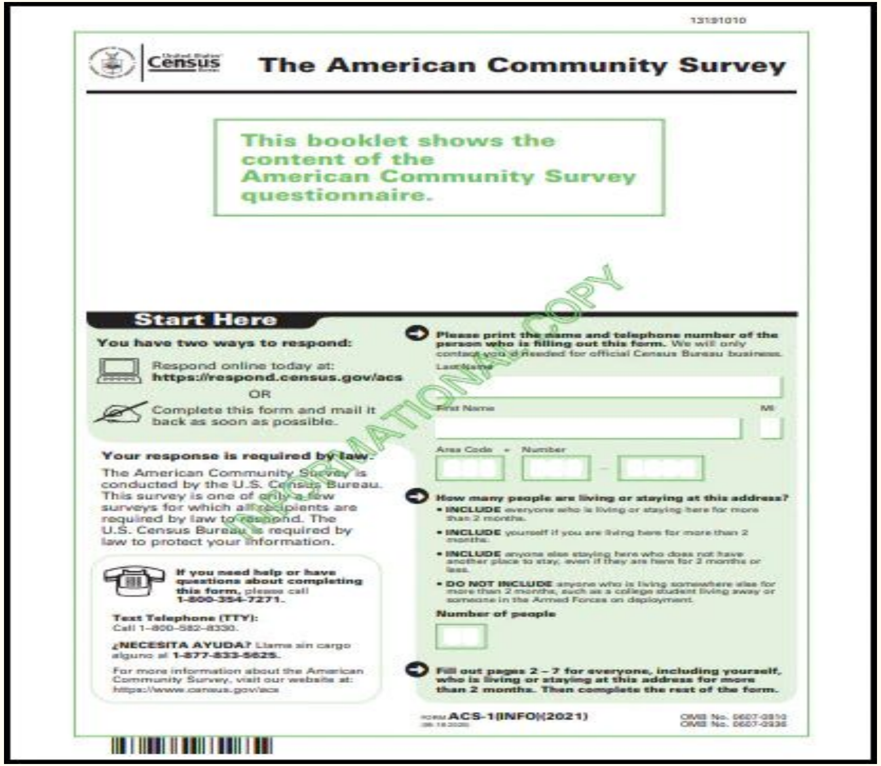
Census variables cont.

'pop': Total population

'Over_18',' over_65': Percentage of population over 18 and over 65

'employed', 'in civilian labor force': Percentage employed or in labor force

'20-24', '25-34', '35-44','45-54', '55-59', '60-64’ '65-74', '75-84', '85+': Age bins



13191010

United States Census Bureau

The American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

Start Here

You have two ways to respond:

Respond online today at: **https://respond.census.gov/acs**

OR

Complete this form and mail it back as soon as possible.

Your response is required by law.

The American Community Survey is conducted by the U.S. Census Bureau. This survey is one of only a few surveys for which all recipients are required by law to respond. The U.S. Census Bureau is required by law to protect your information.

If you need help or have questions about completing this form, please call **1-800-354-7271.**

Text Telephone (TTY): Call 1-800-582-8330.

¿NECESITA AYUDA? Llame sin cargo alguno al **1-877-833-5625.**

For more information about the American Community Survey, visit our website at: https://www.census.gov/acs

➜ **Please print the name and telephone number of the person who is filling out this form.** We will only contact you if needed for official Census Bureau business.

Last Name

First Name MI

Area Code + Number

➜ **How many people are living or staying at this address?**

- **INCLUDE** everyone who is living or staying here for more than 2 months.
- **INCLUDE** yourself if you are living here for more than 2 months.
- **INCLUDE** anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.
- **DO NOT INCLUDE** anyone who is living somewhere else for more than 2 months, such as a college student living away or someone in the Armed Forces on deployment.

Number of people

➜ **Fill out pages 2 – 7 for everyone, including yourself, who is living or staying at this address for more than 2 months. Then complete the rest of the form.**

FORM **ACS-1(INFO)(2021)**

OMB No. 0607-0810
OMB No. 0607-0936

Election related variables

Incumbent: 1 if incumbent running, 0 if not.

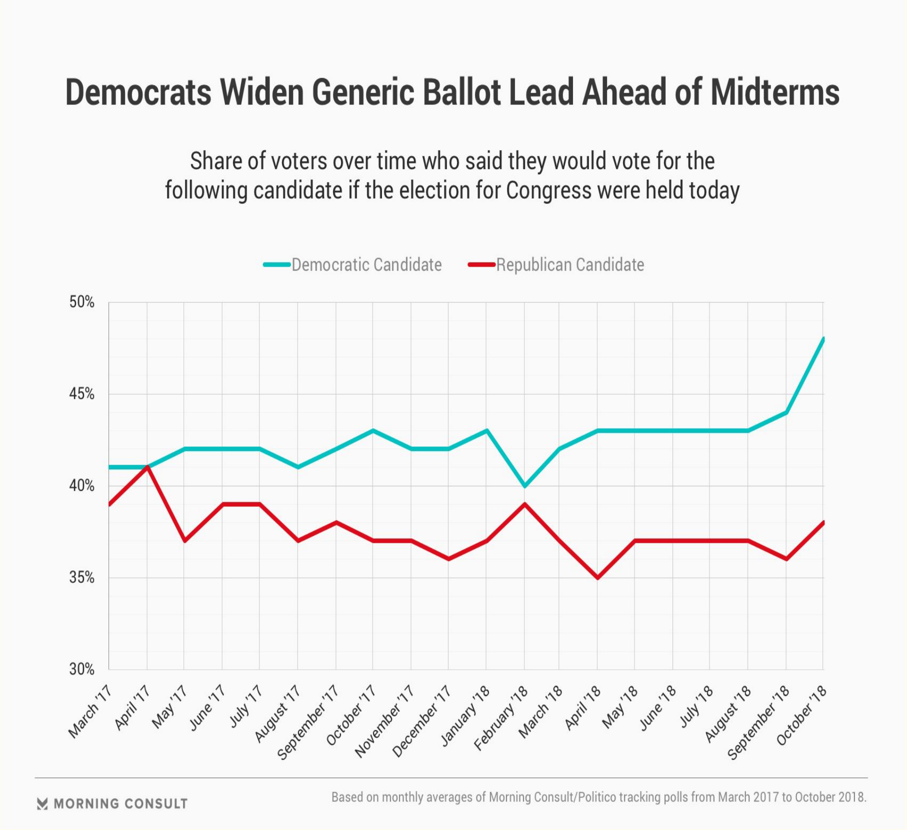
Presidential: 1 if in a presidential year, 0 if not.

Pres, Gov, Sen: binary variables for each type of election.

Gallup_approval_ED: election day presidential approval

Texas_approval_ED: Texas election day presidential approval.

Generic: generic ballot on election day

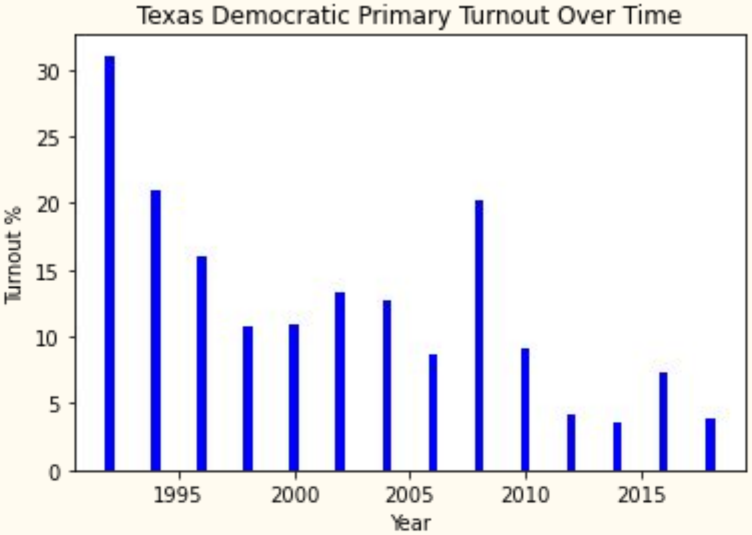


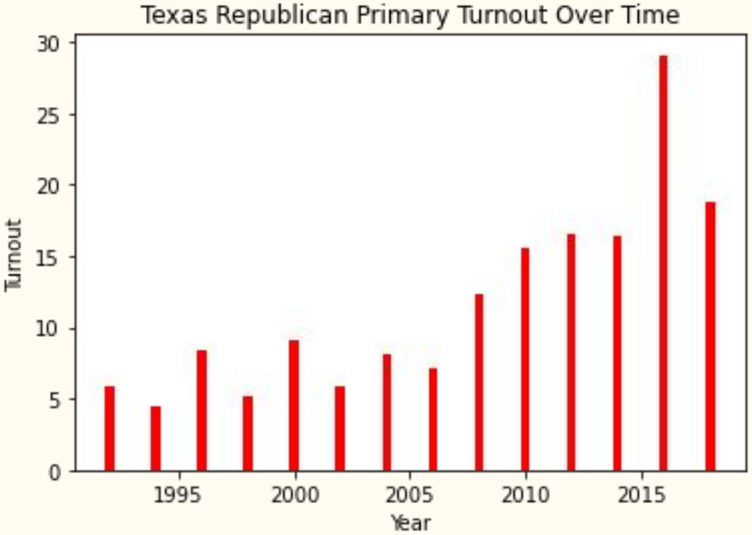
Democrats Widen Generic Ballot Lead Ahead of Midterms

Share of voters over time who said they would vote for the following candidate if the election for Congress were held today

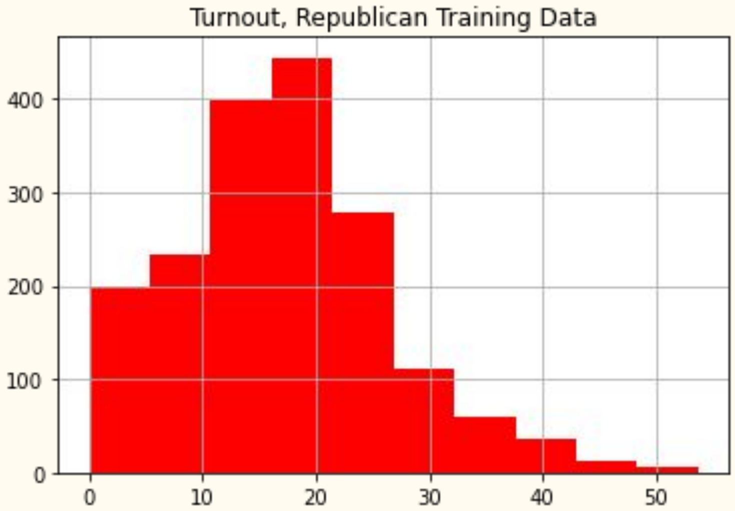
Based on monthly averages of Morning Consult/Politico tracking polls from March 2017 to October 2018.

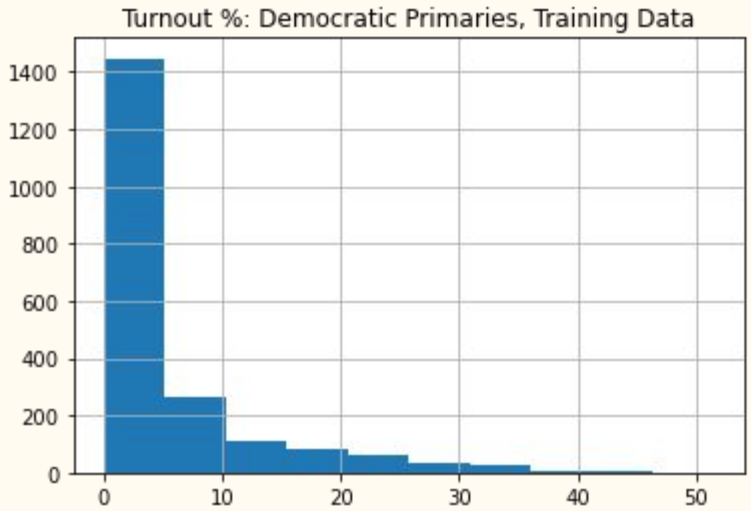
EDA: Turnout over time



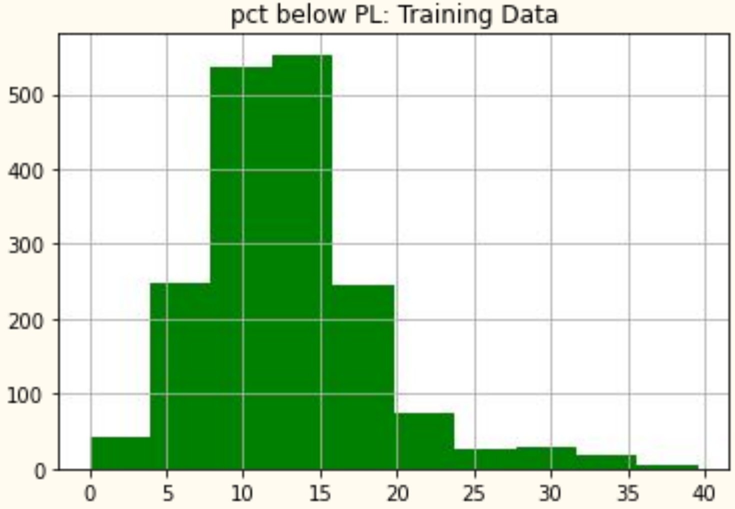


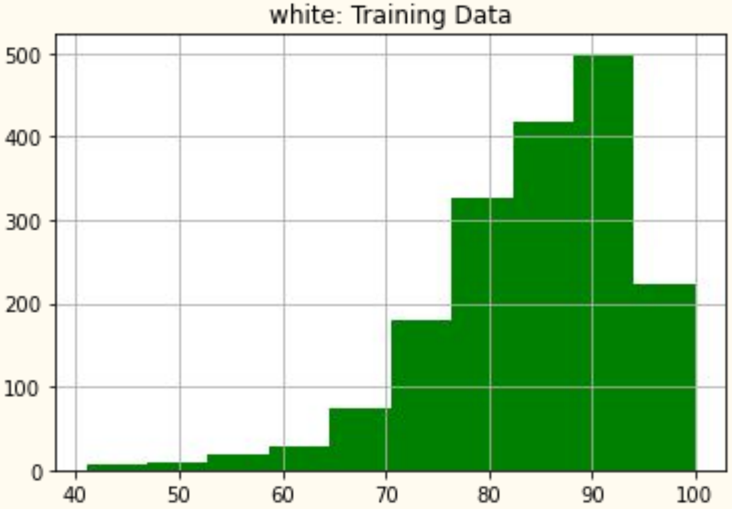
EDA: Feature Distributions





EDA: Other Features





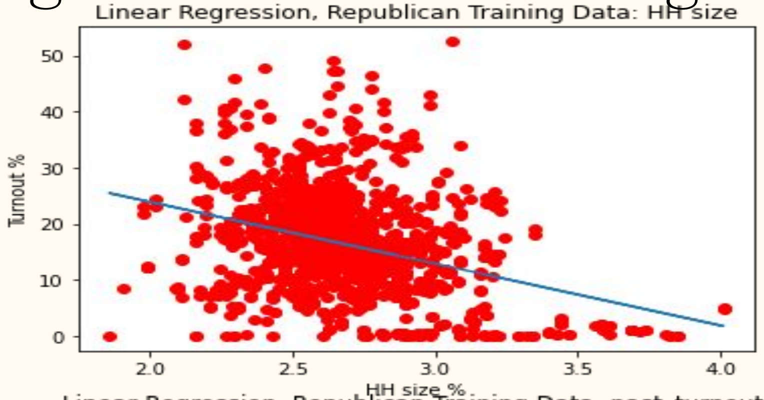
Approach to modeling

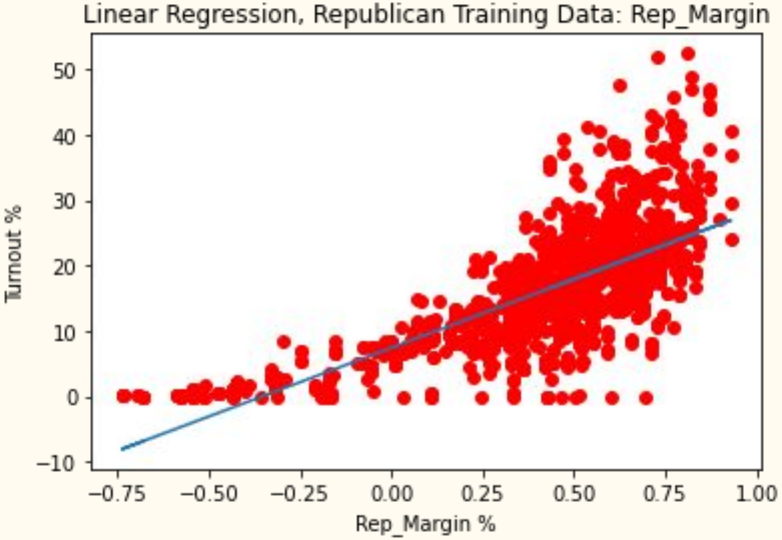
I used a validation-testing approach based on how Dr. Therriault tested his 2020 turnout model.

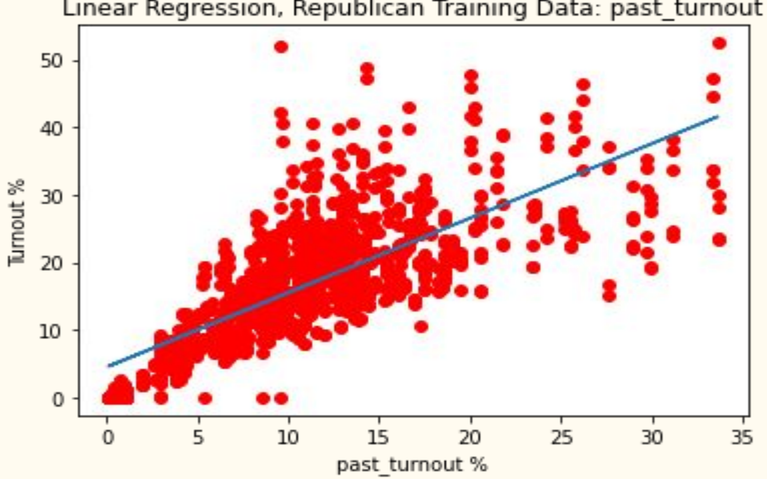
For each kind of model, I withheld one year's worth of data at a time for training, and then used the elections from the missing years' data for testing.

I created 6 different types of models: Full linear regression, a less complex version of linear regression with fewer variables, single variable linear regressions, random forest, boosted trees, and extra trees.

Single variable linear regression


Linear Regression, Republican Training Data: HH size


Linear Regression, Republican Training Data: Rep_Margin


Linear Regression, Republican Training Data: past_turnout

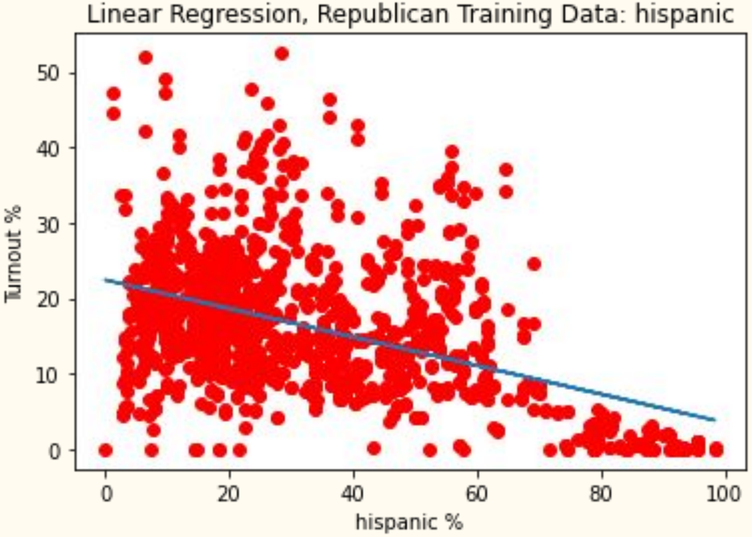
Single variable LR takeaways

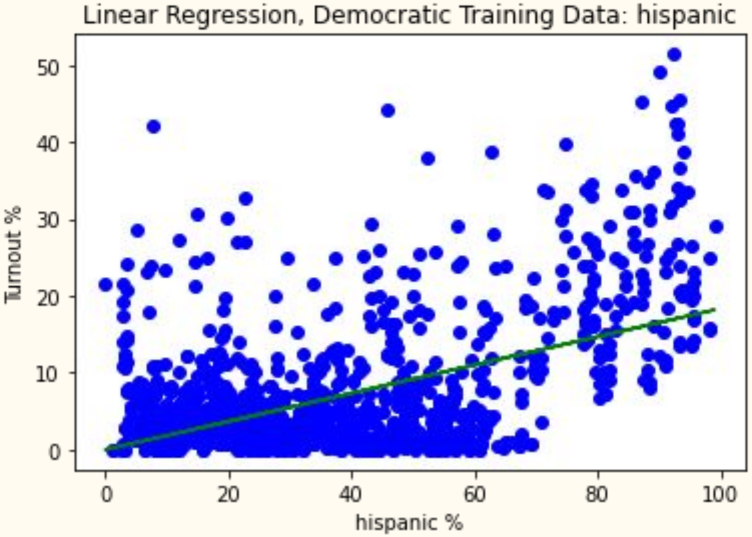
Most variables showed very little explanatory power on their own.

Variables that did have notable R^2 values included:

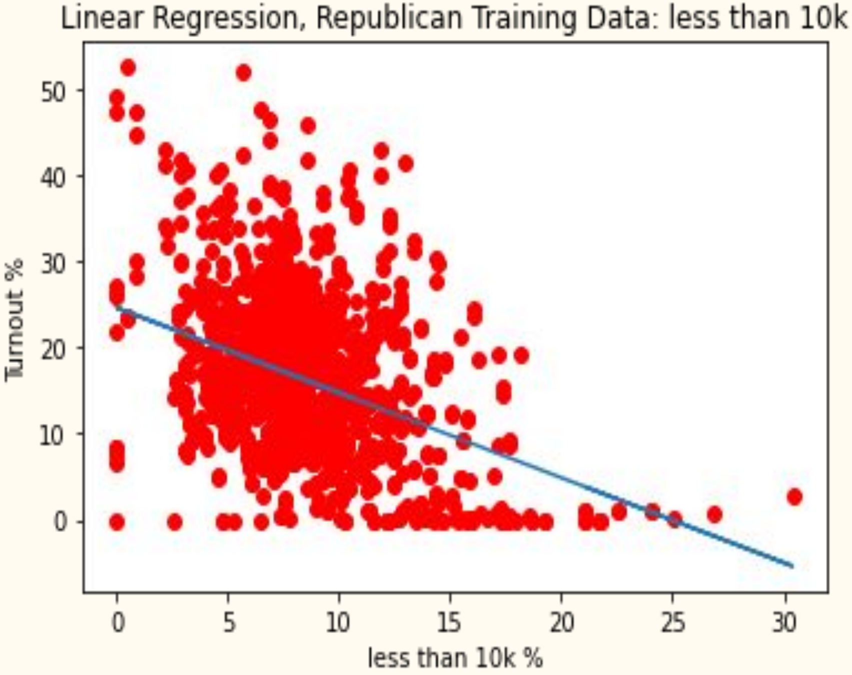
- Most recent presidential margin of the county
- Hispanic percentage, non-Hispanic White percentage
- High school education rate,
- Mean of past primary turnout,
- Percentage of the population below poverty line, percentage of the population making less than $10,000 per year,
- Household size

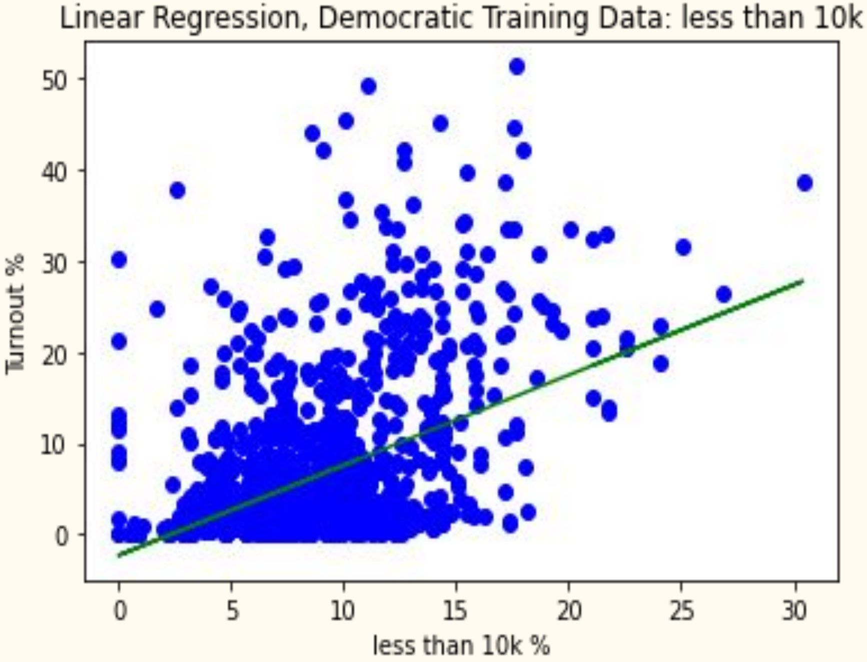
Differences between Republican and Democratic models





Differences between Republican and Democratic models

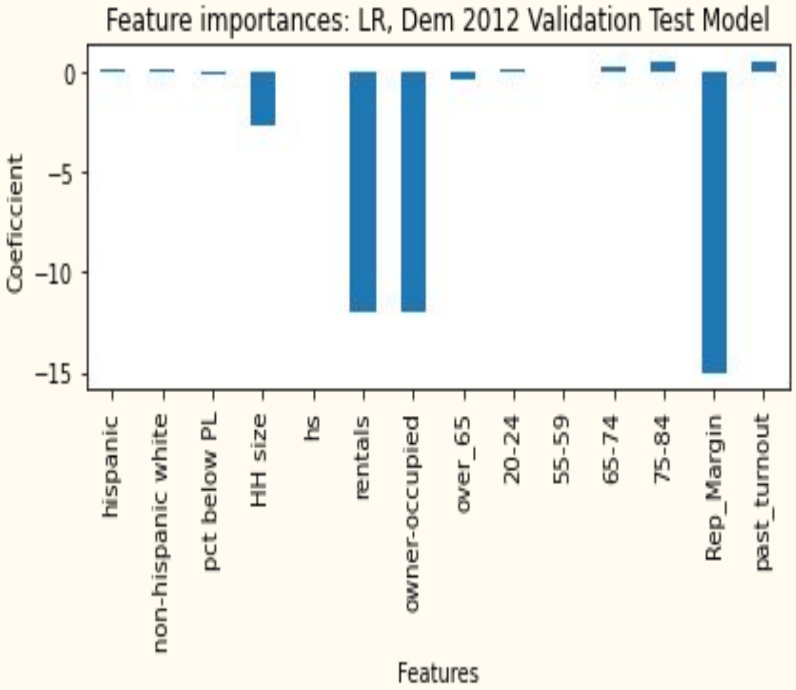




Validation Testing: Linear Regressions

The full linear model achieved scores between 52% and 61% for the Republican models, and 5% and 62% for the Democratic ones.

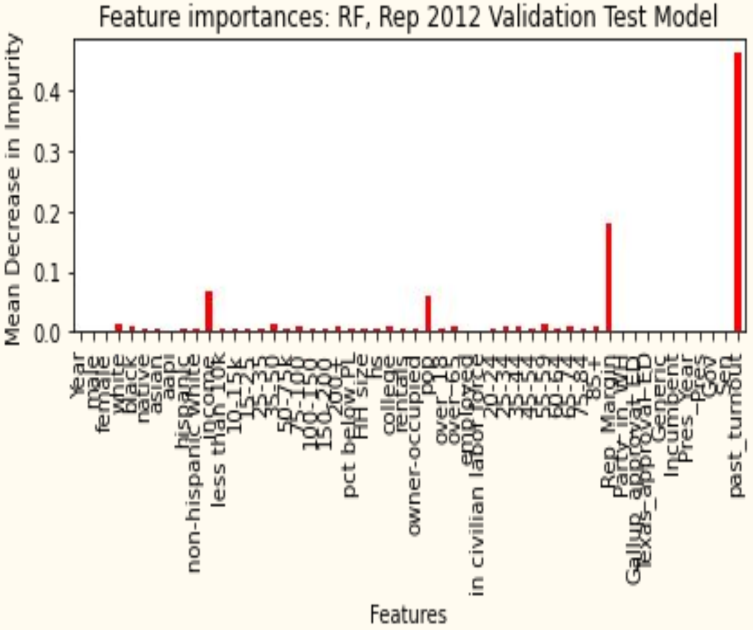
The smaller models performed worse for Republican races, with scores between 57% and 66%, though for Democratic races the scores were 47% and 72%.

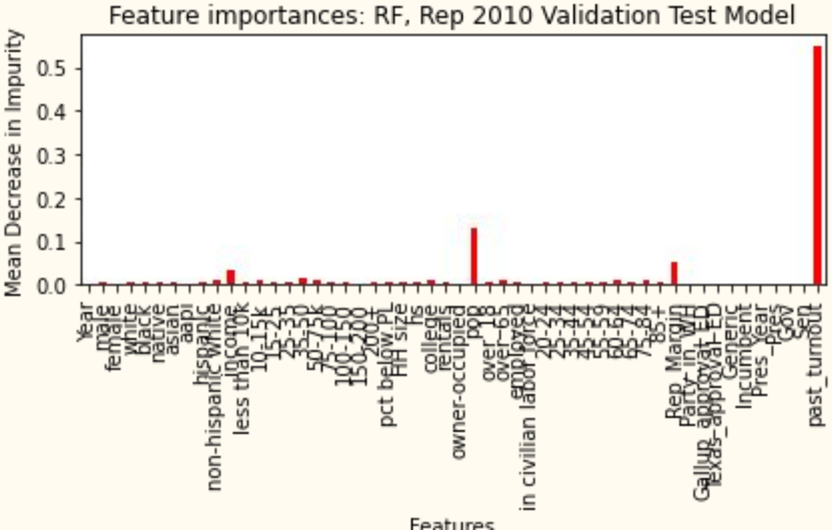


Validation Testing: Random Forest

Random Forest Models performed better, with Democratic R^2 ranging from 51% for the 2010 Democratic governor's race to 74% for the presidential race in 2012.

Republican model performance ranged from 62% for the 2014 Gubernatorial primary to 70% for the 2010 Senate primary.



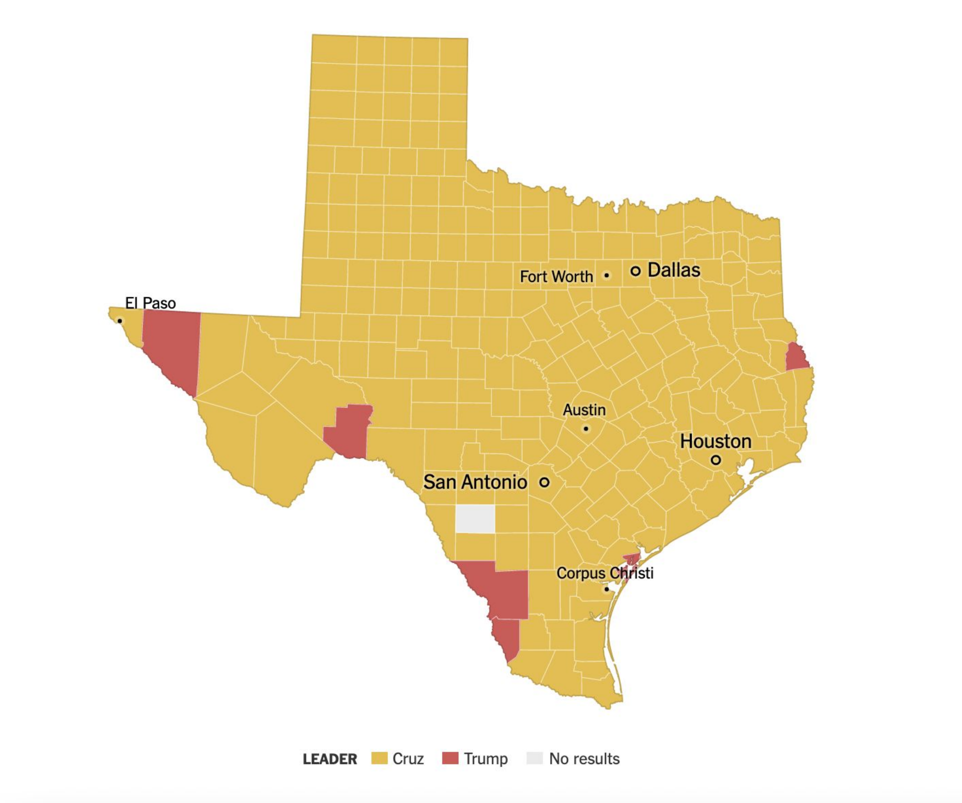


Republican 2016 Data

Initially, Republican validation test models were not functioning well, and I found that removed 2016 from the training set vastly improved performance.

This is likely due to the fact that turnout in the 2016 Republican primary was almost double every other year (~29%), as it was a very high-profile and competitive primary between Trump and Cruz.

This led me to remove 2016 from the training set for the final 2018 models.



Model Testing: Linear Regression

Full Linear models performed well for Republicans, both with R^2 of 64%, but very poorly for the Democratic model.

Both full and mini linear models tend to overestimate Republican turnout and underestimate Democratic turnout.

Mean Absolute Errors were about 4 for both Republican models and 2.6 for Democratic ones.

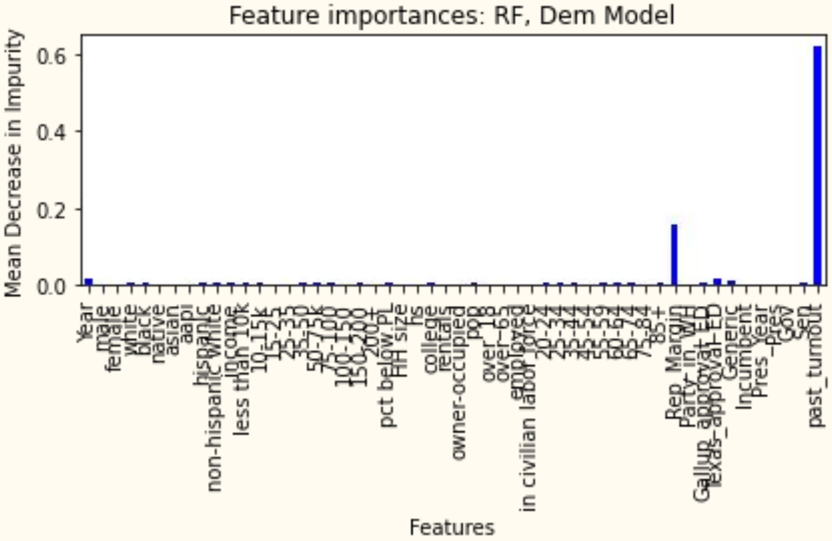
Mini Linear Regressions performed about the same for Republican races, and achieved scores of 69% and 67% for the Democratic races.

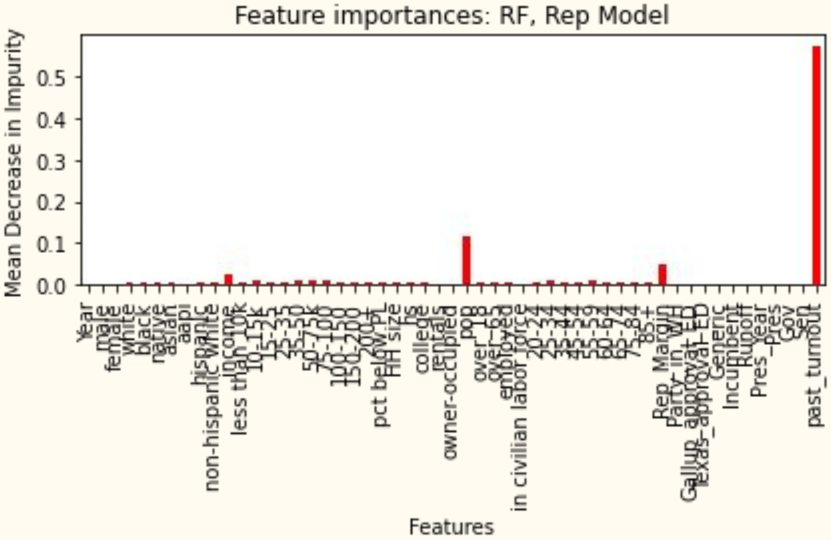
The largest coefficient of the Democratic model was by far presidential margin, though the Republican ones had larger coefficients for % of housing rented/owned.

Model Testing: Random Forest

Random Forest modeling led to dramatic improvements for the Democratic models, with R^2 scores of 86% and 82% for the Senate and governor's races.

The Republican model also improved, though not by as much, with scores of 70% for both models.





Model Testing: Boosted and Extra Trees

Boosted Trees performed about the same for Democratic models, and slightly worse for Republican ones, with a score of 67% for both models.

Extra Trees was slightly worse for the Democratic models, scoring 79% and 78% for the Senate and governor’s races.

It was the best for Republicans, with scores of 72% and 73%.

All 3 tree-based models continued to underestimate Democratic primary turnout and overestimate Republican turnout.

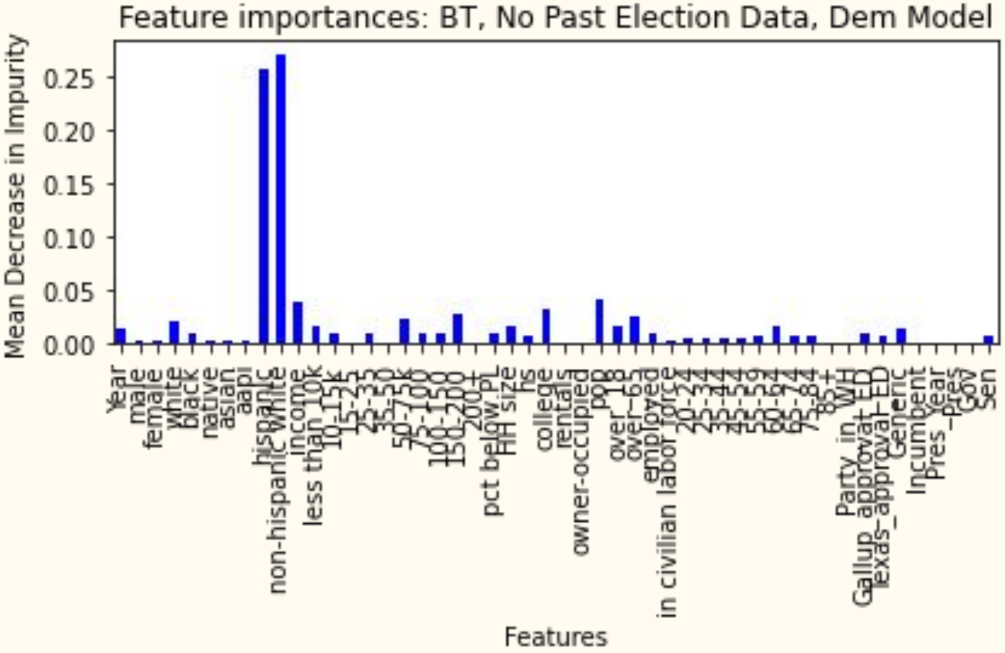
Feature Importance: Democratic Models

As shown in the feature importance graphs, there is a high degree of asymmetry in which features are playing large roles in the models, with mean of past turnout and most recent presidential margin consistently pulling the most weight.

I experimented with removing this data and seeing how the model performed, using the models that performed best originally, in this case the boosted trees model.

R^2 values dropped to 73% and 66% for the 2018 senate and governor's models.

The variables for Hispanic % and Non-Hispanic White % begin to demonstrate much more importance in this scenario.



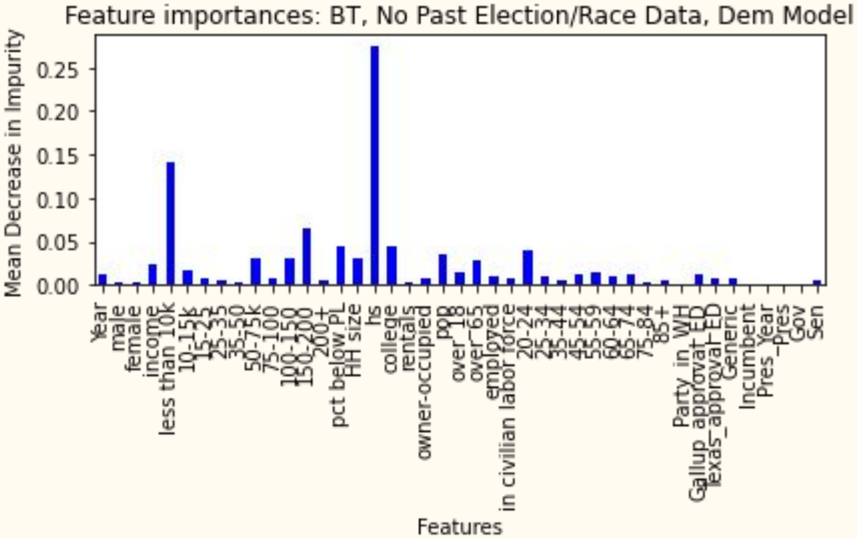
Feature Importance: Democratic Models cont.

After seeing the large role that race begins to play in the models if election data is removed, I conducted one more experiment in which I removed all race-related data from the model.

The result was a further drop in R^2, with scores of 53% and 52% for the two models.

High school completion rates, and % of the population making less than $10,000 begin to act as the most influential variables.

This is evidence toward the idea that the racial variables contain information not found in the demographic/economic ones.

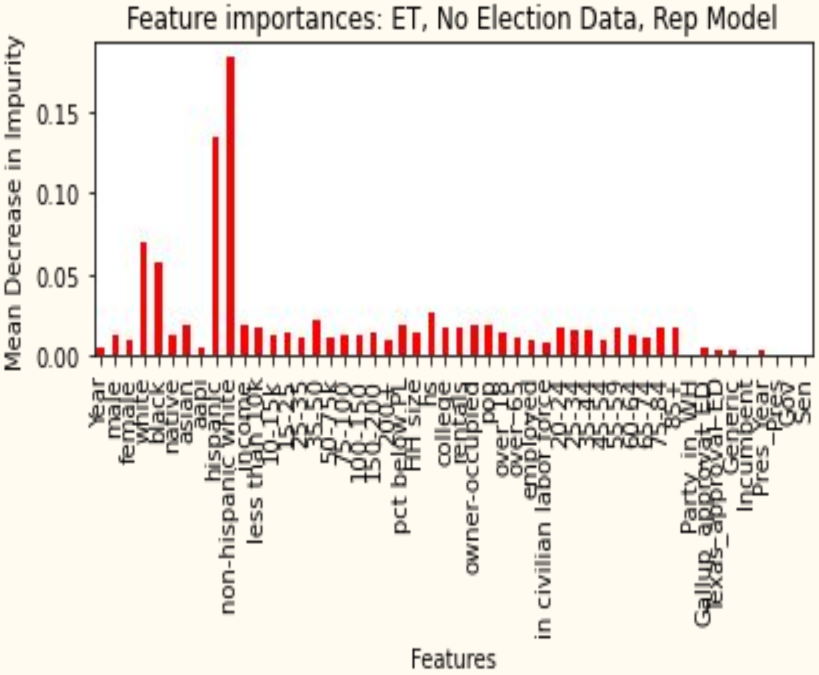


Feature Importance: Republican Models

I then did the same with the Republican Extra Tree models.

This led to a similar result: Hispanic % and Non-Hispanic White % become the most important features.

The R^2 values dropped to 52% and 54% for the Senate and governor’s primary models.

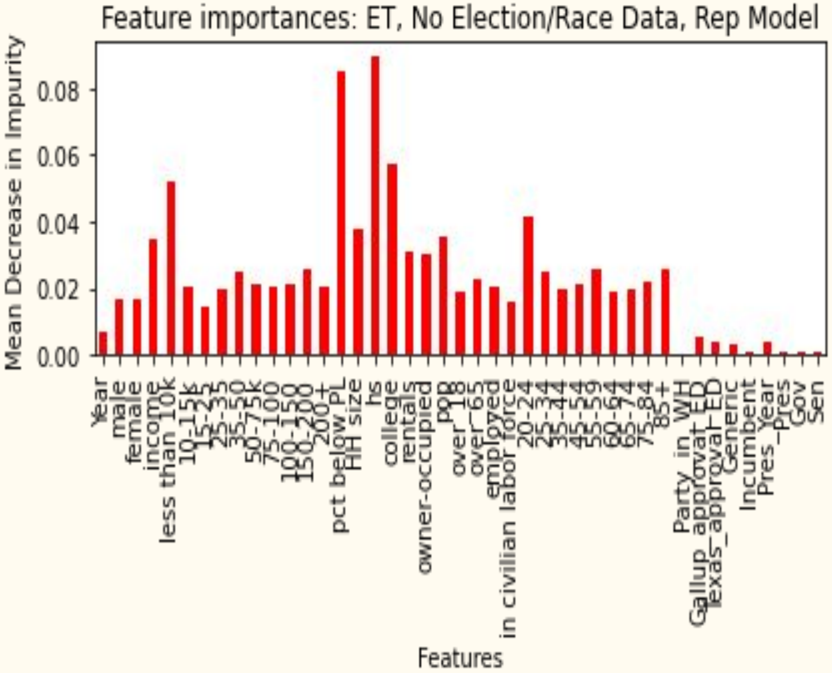


Feature Importance: Republican Models cont.

After removing race data, the feature importance becomes much more equally distributed.

HS education %, poverty, household size, and making less than 10k per year see large increases in importances.

Interestingly, R^2 values stay about the same, at 53% each. This implies that, at least for these Republican models, the predictive information of the racial data is also contained in these other variables.



Limitations

- Small sample size: I only included elections from 2010-2018.
- The removal of 2016 data for the Republican model could prove to be a good choice if it remains an anomaly. 2020 was a more normal year, though it was not competitive at all.
- Should Trump not run again in 2024, however unlikely that may be, then there is potential for another highly competitive race with high turnout.
- I did not study any years with any particularly competitive Democratic presidential primaries, though the 2014 governor's primary was competitive.
- The 2012 Republican Senate primary was also quite close.
- There is evidence that competitive presidential primaries lead to high Democratic turnout, as happened in 2008 and 2020, so that hypothesis could be tested again in 2024/2028.

Conclusions

Turnout for both Democratic and Republican primaries in Texas can be modeled with a relatively high degree of accuracy.

In general, Democratic models performed better, and had larger improvements between the linear regression and more complex tree-based models.

For both parties, information about past turnout and presidential election margins in the county played a large role in the predictions, followed by race.

In Republican models, the contributions of the racial variables could be supplemented by more education/class related variables without a drop in R^2, though this was not the case for the Democratic ones.

Works Cited

Atkeson, Lonna Rae, and Cherie D. Maestas. "Presidential Primary Turnout 1972–2016." *PS: Political Science & Politics*, vol. 49, no. 04, Oct. 2016, pp. 755–60. *DOI.org (Crossref)*, https://doi.org/10.1017/S1049096516001608.

"Censusdata." *PyPI*, https://pypi.org/project/CensusData/.

Election Results, https://www.sos.state.tx.us/elections/historical/index.shtml.

"Explaining the Bloomberg News 2020 Election Turnout Model." *Bloomberg.com*, Bloomberg, www.bloomberg.com/graphics/2020-us-election-results/methodology.

Fernandez, Manny. "Texas: Primary Election Set for May 29." *The New York Times*, The New York Times, https://www.nytimes.com/2012/03/02/us/texas-primary-election-set-for-may-29.html.

Geys, Benny. "Explaining Voter Turnout: A Review of Aggregate-Level Research." *Electoral Studies*, vol. 25, no. 4, Dec. 2006, pp. 637–63. *DOI.org (Crossref)*, https://doi.org/10.1016/j.electstud.2005.09.002.

Kenney, Patrick J. "Explaining Primary Turnout: The Senatorial Case." *Legislative Studies Quarterly*, vol. 11, no. 1, Feb. 1986, p. 65. *DOI.org (Crossref)*, https://doi.org/10.2307/439909.

Kenney, Patrick J., and Tom W. Rice. "Voter Turnout in Presidential Primaries: A Cross-Sectional Examination." *Political Behavior*, vol. 7, no. 1, 1985, pp. 101–12. *DOI.org (Crossref)*, https://doi.org/10.1007/BF00987264.

Krogstad, Jens Manuel, et al. "Key Takeaways about Latino Voters in the 2018 Midterm Elections." *Pew Research Center*, Pew Research Center, 27 Aug. 2020, https://www.pewresearch.org/fact-tank/2018/11/09/how-latinos-voted-in-2018-midterms/.

Works Cited cont.

Leighley, Jan E., and Jonathan Nagler. *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States*. Princeton University Press, 2014. *DOI.org (Crossref)*, https://doi.org/10.1515/9781400848621.

McElwee, Sean, et al. "Is America More Divided by Race or Class?" *The Washington Post*, WP Company, 7 Dec. 2021, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/12/how-do-race-ethnicity-and-class-shape-american-political-attitudes-heres-our-data/.

Nownes, Anthony J. "Primaries, General Elections, and Voter Turnout: A Multinomial Logit Model of the Decision to Vote." *American Politics Quarterly*, vol. 20, no. 2, Apr. 1992, pp. 205–26. *DOI.org (Crossref)*, https://doi.org/10.1177/1532673X9202000204.

Pedregosa, Fabian et al. "Scikit-learn: Machine learning in Python". *Journal of machine learning research* 12.Oct (2011): 2825–2830. Print.

"Polling." *The Texas Politics Project*, https://texaspolitics.utexas.edu/polling.

"The Politics of Financial Insecurity." *Pew Research Center - U.S. Politics & Policy*, Pew Research Center, 20 Aug. 2020, https://www.pewresearch.org/politics/2015/01/08/the-politics-of-financial-insecurity-a-democratic-tilt-undercut-by-low-participation/.

"Presidential Job Approval Center." *Gallup.com*, Gallup, 11 Nov. 2021, https://news.gallup.com/interactives/185273/presidential-job-approval-center.aspx.

RealClearPolitics - Generic Congressional Ballot. https://www.realclearpolitics.com/epolls/other/generic_congressional_vote-2170_1.html.

Texas Election Results 2016 – The New York Times. https://www.nytimes.com/elections/2016/results/texas.

U.S. Census Bureau. *2006-2010 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2010

U.S. Census Bureau. *2008-2012 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2012

U.S. Census Bureau. *2010-2014 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2014.

U.S. Census Bureau. *2012-2016 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2016

U.S. Census Bureau. *2013-2018 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file]. 2018