Local Voters

An analysis of one district’s Registered Voter data.
Then building a model to predict likely voters.

Introduction - The Data

As part of his wife’s run for office, my friend has purchased the Registered Voter data for our city from the Registrar.

two data sets:

- one for every registered voter
- one for every household.

The registered voter data includes voting patterns for the last 3 full election cycles; November and June elections for 2012, 2014, and 2016.

Introduction - The Challenge

- Which households should I target for phone calling and mailing?
- Are there any gross voter trends that will help me with targeting?

What I did:

- Statistical analysis of the data:
  - Identifying various characteristics that impact a voters vote rate.
- Random Forest supervised machine learning model:
  - Training and optimizing
  - Predict likely voters for the June and November 2018 elections.

All data cleaning, analysis and modeling was completed in iPython Notebooks github: LocalVoters.

Data Cleaning and Wrangling

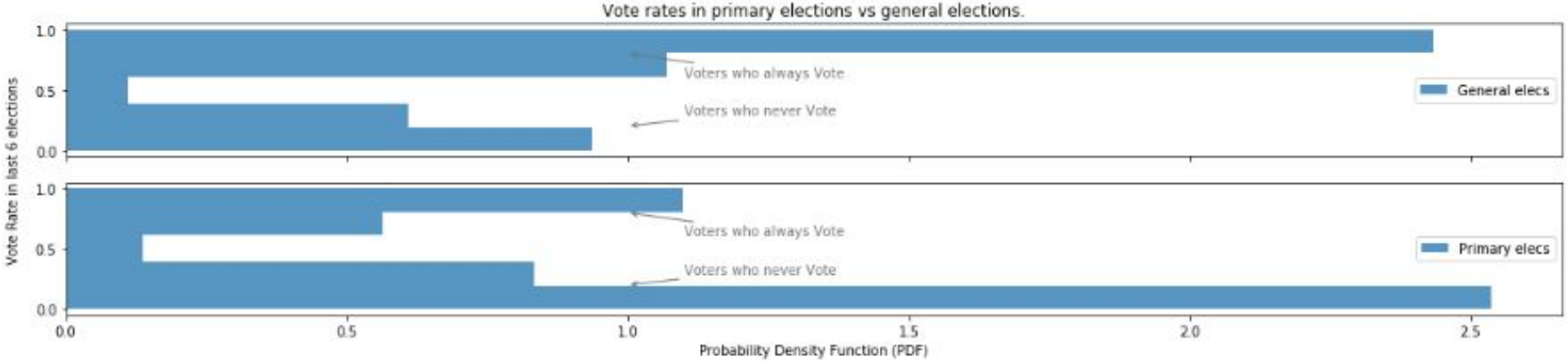
Voter Data: 13307 total rows and 56 columns of data.
Household Data: 6930 total rows and 23 columns of data.

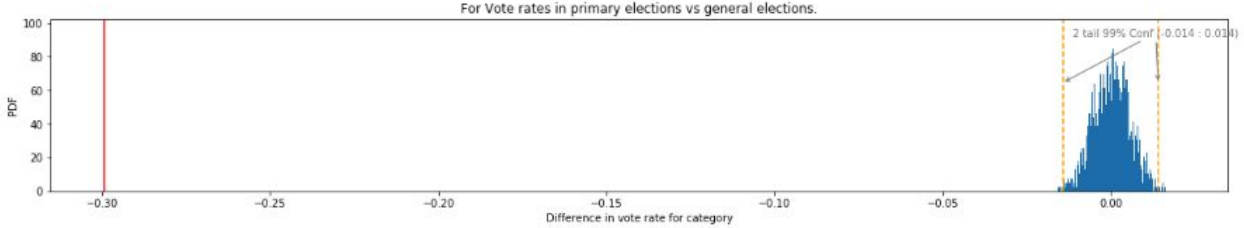
Steps:

- Data with no valuable information was dropped ~5 columns
- All PII was removed, sometime after use to create a feature (eg birthyear)
- Unique ID was created so existing id could be dropped (privacy)
- Election data/features was collected - eg size of majority in government
- Vote Rates calculated
- Calculated fields eg num of voters in Household, mixed affiliations in HH

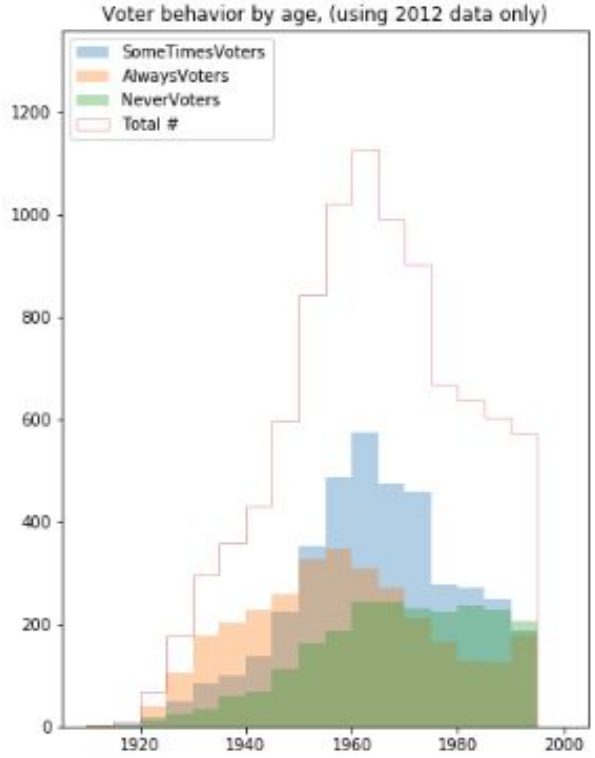
Analysis of the Data - Primaries

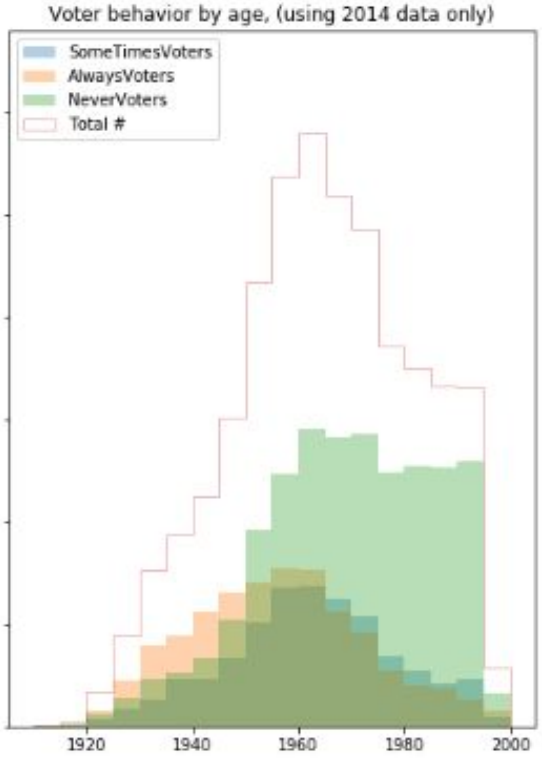
Voters are 30% less likely to vote in primary elections than general elections.

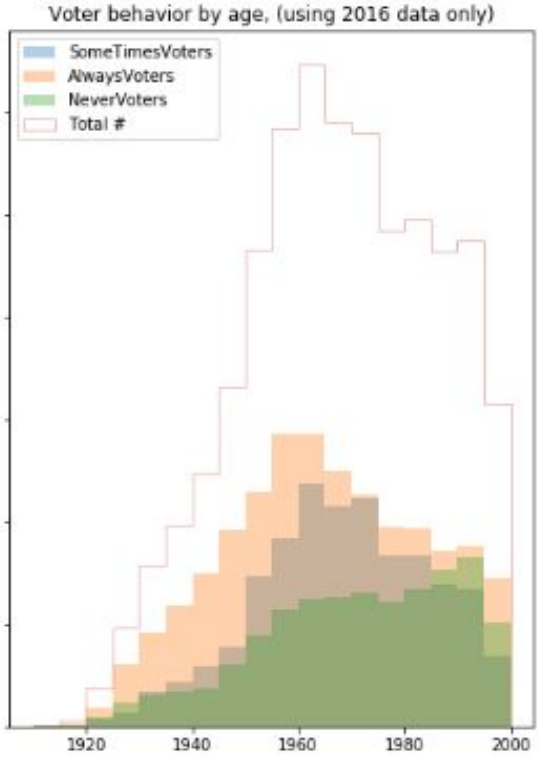




Analysis of the Data - Age







Analysis of the Data - Summary

- Voters are 30% less likely to vote in primary elections than general elections.
- Older Voters are more likely to vote
- Voters with a party affiliation are 12.4% more likely to vote than unaffiliated voters.
- Voters with Democrat and Republican affiliations were more likely to have a higher vote rates than someone belonging to one of the minor parties or holding No Party Preference.
- Holding a permanent Absentee Ballot increases the chance of you always voting by ~20%
- No significant difference in Male and female vote rate
- Voters who live in apartments are less likely to vote than voters whose address does not include an apartment number (obs: ≈4.9%).
- If everyone in your household is affiliated with a political party you are 12% more likely to vote than if you don't live in such a household.
- Living in a household with people who are affiliated with another party meant you were 4.9% (SS at 99%) less likely to vote than if you lived in a household where all the voters in the household were affiliated with the same party.

Building a Predictive Model

Steps:

- Reshaped data into one row per election vs one per voter
- Checked class balance:
  - 30701 in vote cast class
  - 30329 in didn’t vote class
- Removed features with zero variance
- Used get_dummies to one-hot encode multi class features
- Final Training Data :
  - 61030 voter/election rows
  - 530 features columns

Initial Random Forest Model

- With a standard hold out test/train split of the data we were able to achieve
  - 90.5% accuracy

	importance
VScorePct	0.202490
nVScorePctInHH	0.134275
us_repre_maj	0.042719
lastCycleVoteRatehh	0.040229
lastElecVoteRate	0.039531
lastElecVoteRatehh	0.037942
etype_Primary	0.034404
BirthYear	0.028919
lastCycleVoteRate	0.027851
OldestInHouseBirthYear	0.026430

The top features are:

- An engineered weighted combination of which elections someone had voted in considering which they were available for.
- The size of the majority in the US house of Representatives is a surprisingly important feature - pos related to Rep maj and a Dem district
- The next set of features are Vote Rates in the last election & cycle and the type of election
- BirthYear and whether you are the Oldest in your household are also important

Tuning the Random Forest

Used a randomized grid search, 5 fold cross validation and 100 iterations.

Random Forest Feature	Parameters explored during random grid search	Best Value
n_estimators	[4, 8, 16, 32, 64, 100, 250, 400, 550, 700]	400
max_features'	['sqrt', 'log2']	‘sqrt’
max_depth	[10, 20, 30, 40, 50, 60, None]	30
min_samples_split	[2, 5, 10]	5
min_samples_leaf	[1, 2, 4]	1
bootstrap	[True, False]	False

- With a standard hold out test/train split of the data we were able to achieve
  - 93.5% accuracy (+3%)

Predicting Voters using Model

Created prediction features set containing 26614 rows of E7 and E8 data

Used the model fully trained on all labeled data and the above prediction features

- 4032 people will vote in E7 the 2018 Primary
- 5861 will vote in E8 the 2018 General.

Expect the E8 results to be an underestimation as everyone's 'last election' vote will incorrectly be -1 for E8 at this time. Running the model again between the E7 and E8 elections (ie in July or August 2018) would enable more accurate predictions to be made.

I use the Vid and my stored lookup table to share who the likely voters are.