Capstone Project B

This is the initial submission for my Capstone Project of the Data Science Career Track

Problem statement:

I approached this challenge as a consultant would, I reached out to a friend active in local politics knowing that their wife is about to run for city councilor. I asked if they had data I could help them analyze.

Customer:

Politically active people who want to help run effective candidate marketing campaigns:

1. Which households should I target for phone calling and mailing?
    a. Who is likely to vote
    b. What is their party registration so I can tailor my message
2. Are they any gross voter trends that will help me with targeting:
    a. Do people in apartments vote as much as others?
    b. What is the nature of very frequent voters?

Data:

As part of preparing for his wife's run for office my friend has purchased the Registered Voter data for our city from the Registrar. He has kindly shared it with me (initially just District 3). It consists of two data sets, one for every registered voter and one for every household. The registered voter data includes voting patterns for the Nov and Jun elections for 2012, 2014, and 2016 the past 3 full cycles.

Key Questions for the Data:

- Who is most likely to vote in the next election.
- What voter characteristics best correlate with voting behavior
- What prediction accuracies can be reached with the Voter features we have available

Initial Exploration:

I created the following ipython notebook exploring the data and what each column contains.
http://localhost:8888/notebooks/LocalVotersRaw/The%20Data%20By%20Column.ipynb

No PII (Personally Identifiable Information) is displayed in the notebook

Analysis:

Remove PII and engineer features to keep valuable Voter attributes without exposing PII:
Clean data and deal with missing values
Calculate a Voter Score and or other modelling methods to indicate propensity to vote
Run full analysis on each variable and how it related to voter behavior
Create predictive models using first 5 elections - to predicted voting behavior in the 6th

Data cleaning and Feature creation initial thoughts:

There are a number of fields that can be cleanly removed from the data however I want to review missing data first and consider clean up as well as feature engineering before dropping the PII.

Specifically:

Gender

there are 5223 missing entries in the Gender column and it may be worth looking at first names to see if any of these can be used to populate Gender before we drop the First Name.

Address:

full street address needs to be dropped as PII - however creating a building type from AptNumber, HouseNumberSuffix etc could be insightful

Birth Date:

Clearly PII, however birth year or generation membership could both be interesting indicators of voting behavior

Identifiers:

I need to be able to back out who someone is in order to provide actionable data to my friend. So although I want to remove the current Household_Id and Voter ID (as these maybe identifiable if you have the right context) I will replace them with my own randomly generated identifier - keeping a mapping for use in delivering results to the customer.

Election

There is data for 6 elections, the votes for each election for each elector need captured and manipulated to allow easy capture of election type (General, Primary), whether the election was a ‘Presidential’ or ‘mid-term’ also the vote type (V, A, N) needs to be separated from the Ballot type (‘DEM’, ‘REP’ etc)

To be removed (PII):

Voter ID - replaced by my own random ID
Affidavit
Name - used for Gender Identification and then removed
Address - used for Building categorization (Single Family Home, Condo, Townhouses, Multi Family Home) and then removed
ImageID - not entirely clear what this number is, but images tend to be private
Phone
Birth Date - to be cleaned into birth year and then dropped
Mailing address - to be cleaned into a boolean - mailing address different to other address’ and dropped
Household_ID - replace by my own random ID

To be removed (no value):

Abbr - an integer in range 1 and 116 (unclear relevance)
LastVoted - Only 6 entries
Military - only 9 - positives in the data set - too sparse to be valuable ??
LTDate - internal to Registrar's office
PermCategory - unknown meaning (even to guy in Registrar's office!)