Project Outline

Select Data

Our group came together around a discussion of the situation of Virginia’s tenth congressional district. Which is considered an in-play district due to demographic shifts. Upon reviewing the data available from the Federal Election Commission, we rapidly expanded the scope of what was possible. After doing some research into data projects based upon FEC data, we discovered that most project were completely focused upon either predicting results or fundraising. Candidate spending was completely ignored.

With the decision to focus on spending, the question became, what aspect? Reviewing the data dictionaries provided by the FEC showed that the expenses are coded into 12 distinct categories. Upon further discussion, we felt that there was probably an ideal spending profile used by successful candidates.

Collect Data

The data collection process entailed several challenges. The FEC data came in two different formats, different layouts and varying time series. The actual expense and campaign information was available as .zip files dating back to 1996. We developed a python script to collect and decode the several .zip and create .csv files, while looping over several years. The next challenge was collecting the election results files which were available as Excel files, each of which was slightly different in layout; again a script was developed to link the information into a single table.

Clean Data

A detailed review of the information collected showed that there much data that was extraneous to our needs. By ruthlessly deleting features we have been able to reduce the size of the data set from in excess of 2.6 gigabytes with millions of data points to a much more manageable size of 77,632,639 lines of data

(insert here).

Develop Hypothesis

Is there an ideal spending profile of a house race. Compare the expenditures of winner versus losers, and what is the rate of increase of spending for each category.

Create Database

Statistical Analysis of data

Provide a statistical overview of the data, for example, what is the standard deviation for each spending category? In addition, what are the most expensive/cheapest races nationally? What is the average cost per vote, highest and lowest.

Data visualizations

What is the shape and profile of the raw data.

Machine learning

Using all of the available learning models, select the top three for further analysis

Cross validations

Run cross validations of the top three models

Data visualizations

Insert pictures here from yellowbrick

Final results

Summation, future considerations

Presentation

Create relevant slides