Vahid Fazel-Rezai
Wenjia Wang
Lawrence Baker
Tanguy Marion

November 5th, 2017

IDS.131 Project Proposal

“In this imperfect world, the sovereign citizens of the first and greatest Electronic Democracy had, through Norman Muller (through him!) exercised once again its free, untrammeled franchise.”

~ Isaac Asimov, Franchise (1955)[1]

Introduction & Question

In 1955, Isaac Asimov imagined America as the first “Electronic Democracy” in his short story, Franchise. Set in 2008, as computers had become more powerful they were able to predict the correct outcome of elections with fewer and fewer votes. Eventually, with enough computing power, the outcome of the election could be accurately predicted with the opinions of a single individual.

We aren’t (quite) living in Asimov’s reality, but we do not require knowledge of every single vote to ascertain the outcome of an election, the typical swing states alone would be enough. Our research question is, for the election of the US president and based on past voter data, which group of voters is most representative of the whole population? Quantitatively, we would like to maximize the accuracy of predicting the election result while minimizing the number of counties and demographics we use to make the prediction.

Data:

We have found an online data set consisting of the results of US presidential elections (1). The data includes presidential elections every four years dating back to 1920. For each election, we have the number of votes towards each candidate at the county level. With this dataset, we can treat the county, year, and previous years’ vote counts as features.

We hope to enrich our features by expanding the county and year features with demographic data. For this, we have found another online data set based on the US Census (2)(3). This data includes fields collected at the household level, which we can aggregate to match the county granularity of the election data. We have access to a very wide variety of fields, including education level, work status, race, veteran status, health insurance etc. Selecting representative counties based these fields may improve our ability to predict the election result based on a small subsample of the population.

(1) http://library.cqpress.com/elections/download-data.php
(2) https://usa.ipums.org/usa/index.shtml
(3) https://www.nhgis.org/

[1] Asmiov, Issac, 1995. “Franchise”

Analysis and methods:

- **Pre-processing the data**

  We will first need to perform some pre-processing on the election data. Our data are separated by year and state, so initially we will need to clean them into a compatible format and join into one large dataset for further querying. We will also have to reformat and aggregate the demographic data to be able to link them with counties and years from the election data. Since we have county voting data and demographic data, we will perform state-level linear regressions in order to estimate how each demographic voted.

- **Divide the prediction in two steps**

  The prediction of the federal election can be broken into two steps: first we use pure vote data to select a subset of the counties for more detailed analysis - reducing the dimension of the problem and then we include demographic data for a more detailed analysis of the most important voter groups.

  The **first step** is the selection of swing counties. We can treat it as a dimension reduction problem, where dimensions are counties and the features are the vote counts for each county. We need to reduce the number of counties ultimately to the most important counties for determining the federal result. Since we cannot know which dimension reduction technique will work best for our dataset a priori, we will adopt a test and learn approach, performing sequentially the main techniques we learned during lectures and comparing the different results. We will reduce dimensions by different algorithms: principal component analysis, stochastic neighbour embedding and multi-dimensional scaling. We will then interpret these results to select a subset of important counties.

  The **second step** is a binary classification problem, where the features are the current year's results for the selected counties, and the label is the overall federal election winner (Democrat/Republican). We plan to classify this data using a set of different classifiers: logistic regression, LDA/QDA, linear SVM and kernelized SVM in order to model which the most important counties are. We will test the accuracy of each of these methods by splitting our data between train and test sets and select the method with the best test set accuracy. We will do that over a different set of regularization (L1, L2 and L1&L2) with hyperparameters optimized over cross validation error minimization.

  In addition, once we have identified important counties, we will use clustering algorithms, both on the subset of swing counties and the entire sample in order to see whether there are shared characteristics and which clusters can be disregarded by the current electoral system.

- **Evaluating the significance of our predictions**

  We will have a mix of options for our first and second steps, and we would like to evaluate which pairing as a whole makes the most accurate predictions. Thus, in addition to optimizing hyperparameters for each choice of model, we can apply our models to predict both state and federal winners for each year. We will determine the success of our methods using cross validation methods, such as leave-one-out.

Vahid Fazel-Rezai
Wenjia Wang
Lawrence Baker
Tanguy Marion

November 5th, 2017

Literature review:

There is significant literature on a related problem, that of predicting an election results from a poll. We have a different dataset, the results from many years as opposed to the poll from a single year, but the problems are comparable. A strength of our approach is that we have many years of data, and so we can iron out anomalous results and we are not subject to polling bias (we can look at real vote-counts). The weakness is that not all Republicans/Democrats run on the same platform and societal values shift over time, so an election many years in the past is likely to be quite different from a more recent one. We can weight our data based on recency to mitigate this issue.

Unscientific polling can date back to 1824 presidential election, where talliers are taken in taverns, militia offices and public meetings to gauge people's perceptions towards candidate [2]. The first scientific polling methodology namely "quota sampling" was invented by George Gallup and adopted in the 1936 pre-election poll. The idea was to force the polling sample to fit in a certain national profile by placing quotas on the sample size for each characteristics categories, e.g. sex, race, age, so that the polling sample proportionally reflect the characteristics of the electorate at large. Despite three consecutive successful predictions, the quota-sampled polling failed disastrously during the 1948 Dewey - Truman presidential election[3].

Immediately after the 1948 polling disaster, a group of social scientists and statisticians led by Frederick Mosteller then performed a thorough review on the polling procedures and outlined eight metrics to assess polling accuracy, among which "Mosteller Measure 3" and "Mosteller Measure 5" are the most commonly used metrics. Mosteller Measure 3 compares the average absolute difference between the poll estimate for each leading candidates in the final estimate against the difference in their margins in actual vote, while Mosteller Measure 5 compares the absolute value of the difference between the margin separating the two leading candidates in the poll against the differences in the election result[4].

There have been multiple published attempts to calculate the relative power of voters to decide the presidency (and presumably many unpublished by party strategists). The simplest analyses argue that the states with the greatest ratio of electoral votes per person have the most power - noting that Wyoming has 3.2 times as many electoral votes per person as the average state. However, these reductive analyses miss the importance of swing states - Wyoming is reliably Republican, and therefore individuals have relatively little say over who their state supports.

[2] Smith, Tom W. 1990. ''The First Straw? A Study of the Origins of Election Polls.'' Public Opinion Quarterly 54(1):21–36.

[3] Elizabeth A. Martin, Michael W. Traugott, Courtney Kennedy; A Review and Proposal for a New Measure of Poll Accuracy, *Public Opinion Quarterly*, Volume 69, Issue 3, 1 January 2005, Pages 342–369, https://doi.org/10.1093/poq/nfi044

[4] Michael W. Traugott; The Accuracy of the National Preelection Polls in the 2004 Presidential Election, *Public Opinion Quarterly*, Volume 69, Issue 5, 1 January 2005, Pages 642–654, https://doi.org/10.1093/poq/nfi061#

538 uses a “tipping point” methodology: the power of the voter is proportional to the probability that their state ‘tips the balance’ in favour of one candidate, divided by the number of voters in that state[5], other groups have refined this attempt by incorporating demographic data and how likely each demographic is to vote[6]

Timeline:

To-do	**Deadline**
Hand in project proposal	Nov 5
Pre-processing the data	Nov 14
Exploration & Computation on county selection	Nov 19
Exploration & Computation on classification	Nov 24
Hand in intermediate status report/ Update questions, project definition and objectives	Nov 26
Be done with the county selection step	Nov 29
Be done with the classification step	Dec 2
Assess the significance of results/ Draw conclusions	Dec 7
Write the report and the presentation	Dec 10
Hand in project report	Dec 13
Project presentation	Dec 13 - Dec 20

[5] Silver, 2016 “A User’s Guide To FiveThirtyEight’s 2016 General Election Forecast” taken from https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/

[6] McLaughlin & Stohr, 2016. “How we calculated how much your vote counts” taken from https://splinternews.com/how-we-calculated-how-much-your-vote-counts-1793862047