ORIE 4741 Final Report

Haotian Liu, Siyu Yang

October 2016

1 Introduction

The result of the 2016 election took pundits and predictors by surprise, and there's been so much talk since Nov. 8 about what the prediction models got wrong. Many points to the inaccuracy of national and state-level polls as the main cause of the inaccuracies, since most prediction models are based on national polls, which according to estimate by Nate Silver, is off by about 2 percentage points. A natural question arise from here: if polls were subject to systematic biases, what can we know from other indicators about this election? This paper attempts to analyze two sets of data regarding 2016 election: economics indicators and twitter data. Using basic machine learning methods, we will test out what these data mean about the result of 2016 election, and see if they can tell us anything that the polls missed.

2 Economic Data

A prevailing argument on media is that people ”voted their wallet” – counties that are economically worse off voted Trump, as he is the person advocating for ”changes”. This section, I seek to analyze what economics indicators have to say about the by-county voting result of 2016 election.

2.1 Data Set and Methodology

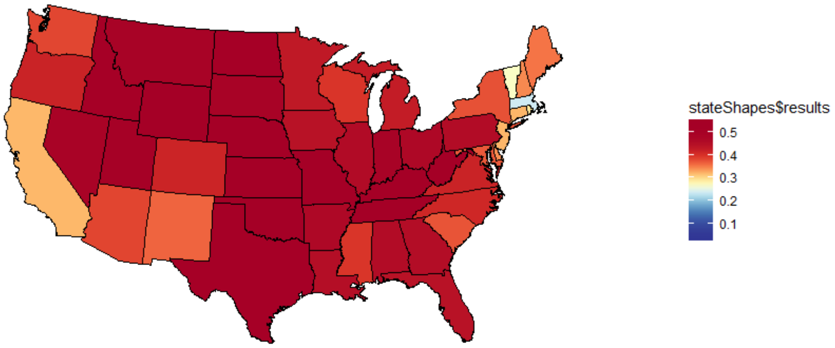
From US Census Bureau, we gathered the five following economics indicators in all counties from 2000 to 2016:

- GDP per capita
- GDP growth rate
- Unemployment
- Percentage below poverty line
- College education rate

For all three elections since 2000 (2004, 2008, 2012), we construct the data matrix X for every county with the five indicators above as well as voting result in the election before (for simplicity, we only took into account Democrat and Republican votes). In order to introduce sparsity, we choose a $l1$ loss function with quadratic regularizer and minimize:

$$\sum_{i=1}^{n} \frac{1}{n}|y - w^T X| + \sum_{i=1}^{n} w^2 \tag{1}$$

We test out the w on 2016 macroeconomic data, and visualize the predicted shift of vote from 2012 election as below:



The result is a almost uniform shift of votes from democrat to republican.

2.2 Model Review and Model Critique

It seems that on a county-by-county level, the macroeconomics data would predict a shift from democrat to republican candidate. However, what caused this predicted shift is questionable. Prior to 2008, the macroeconomic data in most counties exacerbated due to the financial crisis, and most voters chose Obama over McCain in that election. As a result, most macro-economic indicators are negatively correlated with democrat votes. In 2012 to 2016, as unemployment drop and GDP per capita picks up, the model predicted a movement away from democrat candidates.

3 Sentiment analysis

3.1 Data Set

Twitter provides an API endpoint to get relevant data based on keywords to be extracted and allows 1% of all the public tweets to be sampled. Taking advantage of the tool, we seek to predict how the people in each county of Florida will vote based on their tweeting behaviors.

Twitter API allows developers to download the tweets sent out in a given radius around a given location. United States Census Bureau provides the latitude and longitude of the geographic center of every single county in Florida, as well as the area of every county in square miles. We use a circle to approximately capture the tweets sent out in every county, with the center of the circle in the geographic center, and radius of the circle as the square root of its area.

Since Twitter Search API only allows access of a sample of recent Tweets published in the past 7 days, we downloaded all tweets sent from each county within the past week.

3.2 Feature Selection

- **Tweet Volume**: The number of tweets mentioning one candidate. An example of a Tweet mentioning Hillary would be:

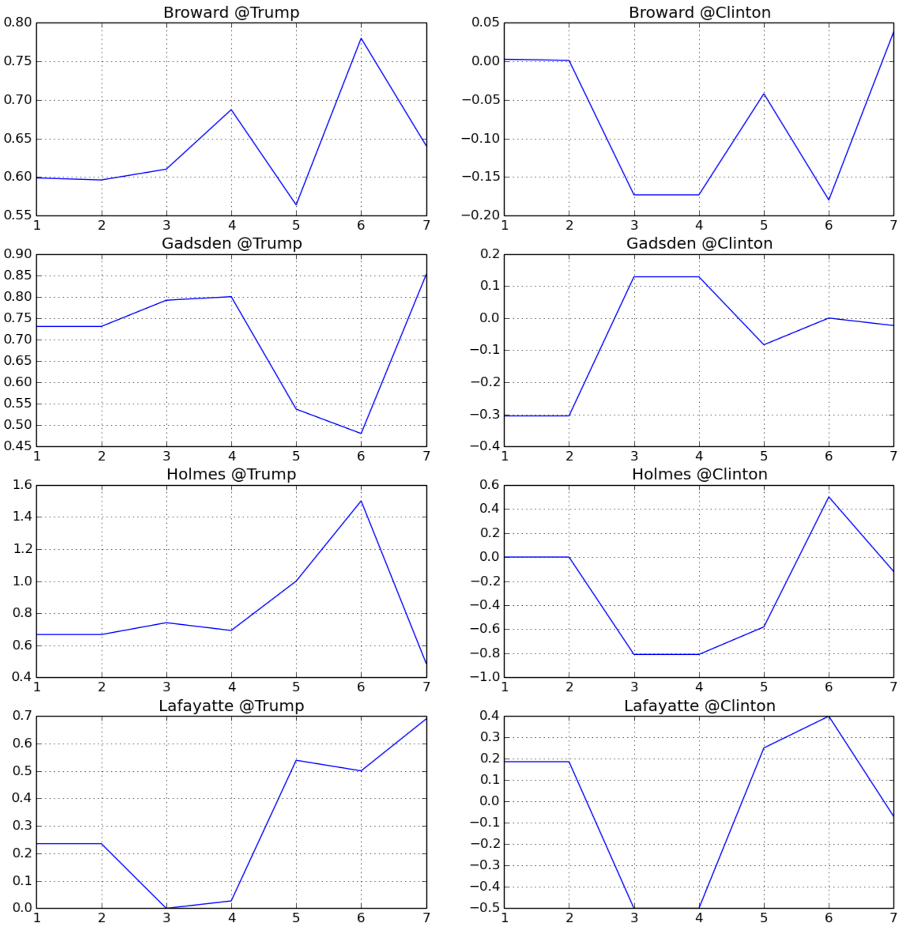
> *@YoungDems4Trump: Thanks to Anthony Wiener, Hillary can commit massive voter fraud...and still lose.*

- **Twitter Sentiment**: In first attempt, we conducted a lexicon based sentiment analysis. A list of English positive and negative opinion words is matched with the words used in the Tweets. Although lexicon based analysis does not work well in deciding the sentiment of individual tweets, past literature indicate that the prediction increase in accuracy as text volume increase.
- **Hashtag Volume**: We compiled a list of hashtags that are used prevalently to indicate support or opposition of candidates. For example, "#LockHerUp" would indicate an opposition of Hillary, and "#ImWithHer" indicates a support for the same candidate.
- **Retweet Volume**: The number of retweets of a tweet mentioning one candidate. This feature is important because it indicates the size of the impact of the tweet.
- **Favorites**: The number of favorites of a tweet mentioning one candidate. This feature is important because it indicates how much other followers agree or disagree with the given user.
- **Tweets with links**: Most tweets containing links are forwarding news about one candidate.

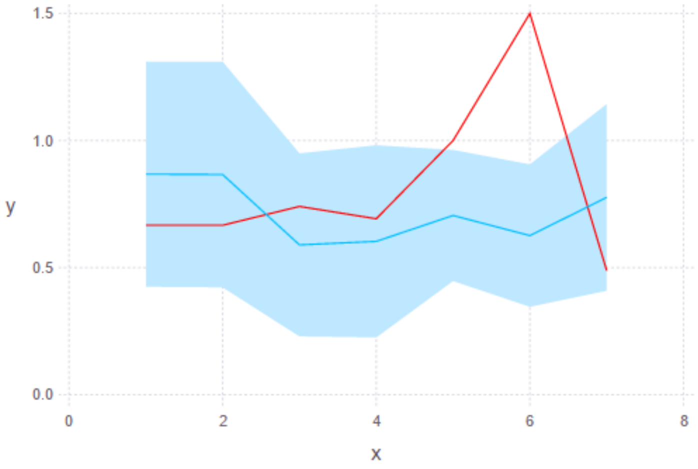
3.3 General Data Exploration: Is there a trend?

We use the tweet data from Nov.1 to Nov.7 and calculate the sentimental score for tweets related to both of the candidates. Specifically, we find that the average score for tweets mentioning Trump is much larger than those tweets referring to Clinton. This might be a indicator that, although different counties has their individual Twitter features, the majority of them are pro-Trump and this can be observed from the 2016 election result.

However, to harness the power of prediction using Tweet data, we need to show if there is a strong relationship between the sentiment score and the advantage that each candidate gets in their "turf". Therefore, we manage to find 4 counties that represent the largest differences in election result with two of them (Broward, Gadsden) being hyper-Democratic and two of them being hyper-Repulican(Holmes,Lafayatte). We track their sentimental score in a week for each candidate.

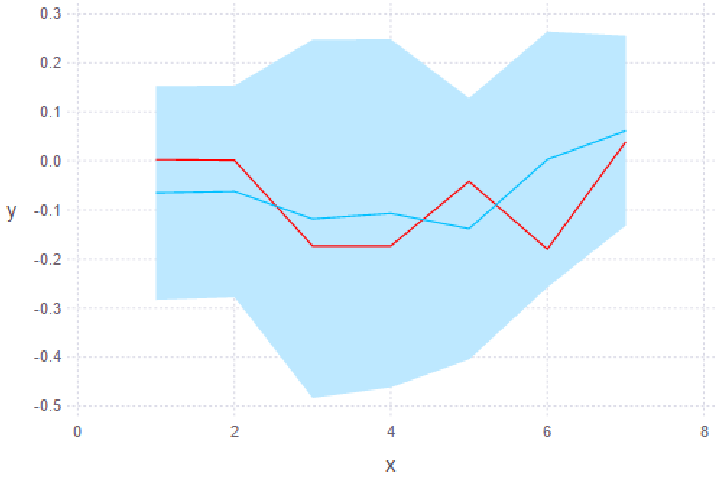


It can be seen that the sentiment scores are higher for Trump even in areas where Clinton finally wins. However, it seems extremely hard to generalize a pattern from the curve since no detectable regularity can be implied from them.
Nevertheless, if we push this generalization thought further, we can compare these hyper-partisan areas with the average score for all counties, to see if they stand out as extremes as well.
For instance, here is the comparison of Holmes to all counties in Florida for Trump:



The blue line is the average score for all counties across the week. The shaded area covers the +-1 standard deviation for the score. As we can observe, Trump's score surges two days before the election, which might imply that the score is a reflective feature of the result.

However, most other graphs do not agree with this argument, for instance, here is the plot for Broward, where Clinton finally wins by almost 40 percent:



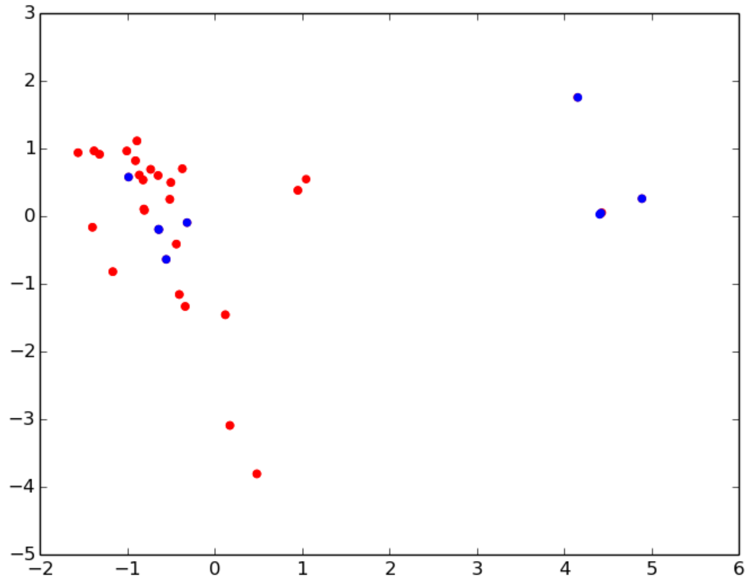
3.4 Regression Analysis:PCA and Classification

Since a clear pattern is lacking in sentimental scores along, we want to develop a effect model using all the features of our available tweet data. However, as mentioned in our mid-term report, we are convinced that a simple linear regression will yield statistically significant result because the variance in the output space is too random.

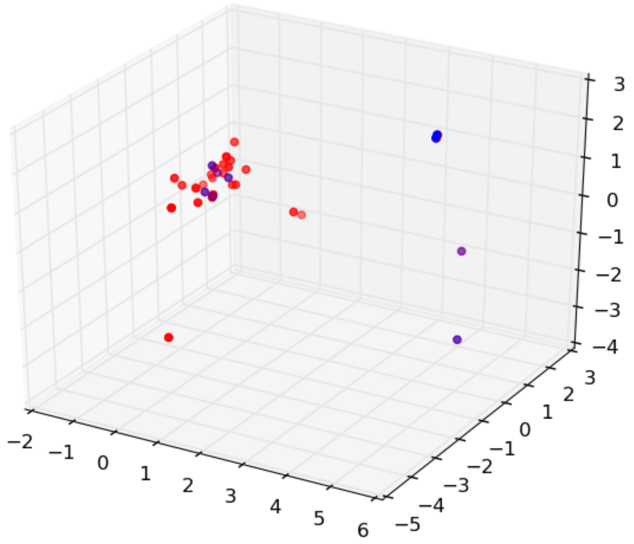
Therefore, we seek to transform the problem to a classification problem instead of regressing on real-number election result (percentage of received votes). We denote 1 for counties where Trump wins and -1 for counties where Clinton wins.

Also, we consider using a PCA method to reduce the dimension of the feature space. Even though the data is NOT low-ranked, PCA is still an effective method to transform the feature space while preserving as much covariance as possible. In this case, we are using the first two and three principal components derived using the SVD.

Here is an example of the PCA result using only the first two components for Clinton. The feature space is the tweet data mentioning Clinton on Nov.2. The location is decided from the component score of the first two components and the color defines the final winner of the county.



As we can see, though not ideal, the data is well separated to two ”clusters”.
Here is the example of using three components on the same dataset:



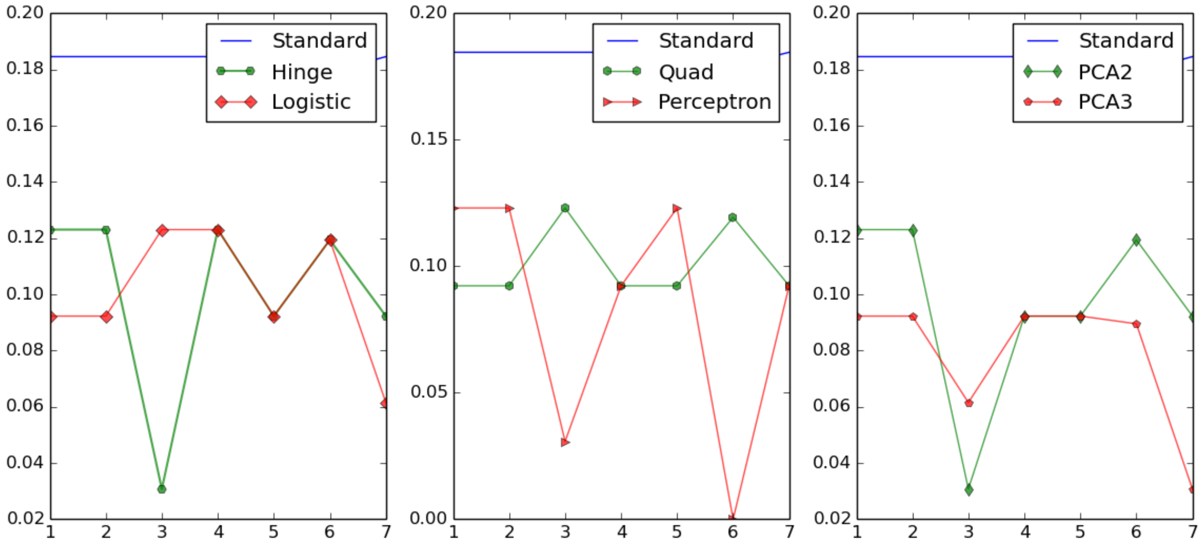
3.5 Classification: Method Comparison

Finding confidence in PCA method, we decide to compare this method with other popular methods to see if the result has a lower prediction error. Also, we want to examine if the prediction error changes over the week.
We choose six different methods:

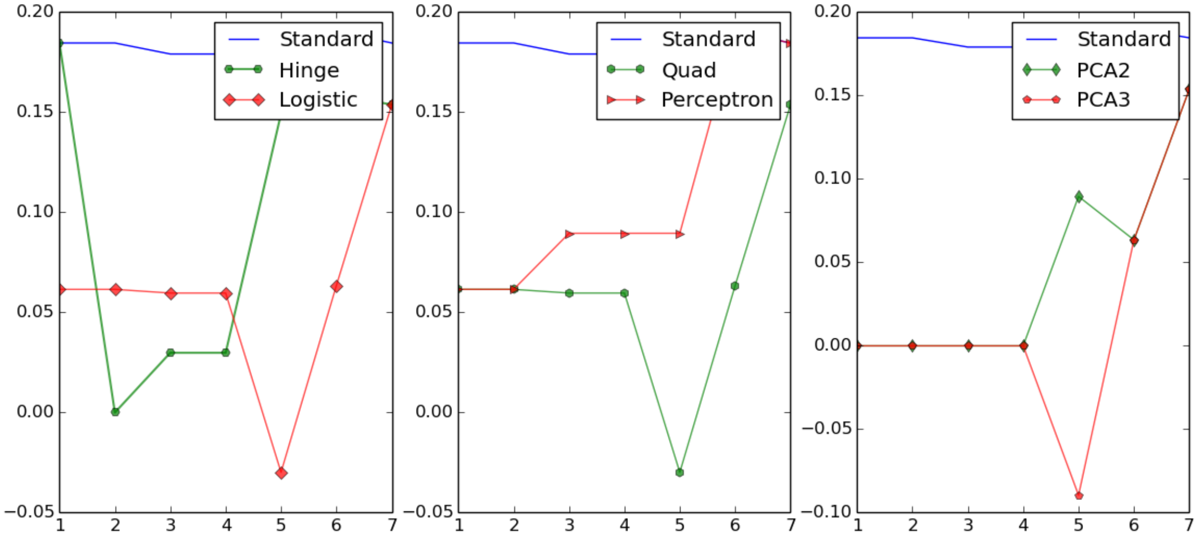
- PCA2: using the first two principal component scores, with a logistic loss
- PCA3: using the first three principal component scores, with a logistic loss
- Hinge: using the normal feature space with a hinge loss

- Logistic: using the normal feature space with a logistic loss
- Quad:using the normal feature space with a quadratic loss
- Perceptron: using the normal feature space and feed into a perceptron algorithm
- Standard: we predict that trump win all counties. This is used as a measurement for the predictive power of other methods.

Here is what we get by comparing the six methods across the week(X-axis) for tweets related to Trump.



Tweets related to Clinton:



3.6 The problem of silent majority, or is there one?

Ever since the election ends, there has been heated discussion as to the failure of polling and modern data science methods. Nate Silver, the God of political prediction and the father of fivethirtyeight.com, even posted an article on his website addressing the limits of predictive modeling even though his website gives Trump a better chance than anyone else.

Among all the problems that political modeling encounters, one of them is mentioned frequently as the root of all evil: the problem of silent majority. Addressed both in media and in research studies, the problem mainly argues that the poll and all other data science methods depending on poll does not reflect the public opinion either because somehow people are afraid to express their support to Trump or because voice these supporters cannot be heard by the general public.

But before we dive into solving this problem, we must ask ourselves if this problem really exists.

Looking at our prediction error using different methods, it is not hard to find that prediction error using tweets related to Trump is generally lower than using the tweets related to Clinton, **especially when it comes closer to the election day**. Even though both sets of methods are merely used for training (with no available test data), it is clear that Tweets mentioning Trump, instead of Clinton, actually contains much more valuable information of the public whereas the problem of the silent majority believes otherwise. In other words, both supporting and opposing forces express themselves via Tweets much more fully than polling statistics reveal. The problem of silent majority, to the point of this analysis, is nonexistent.