ORIE 4741 Midterm Report

Haotian Liu, Siyu Yang

October 2016

1 Introduction

Due to its unique electorial college system, American presidential election cannot be accurately predicted without proper insight into the state level voting patterns. For years, people have attempted to capture the capricious mood of the general public and create a faithful representation of people's opinions about presidential elections. Now, fortunately, people are recording their opinions and saving them une masse in social media. However, due to the formidable quanitity and messy forms of the data, traditional statistical tools no longer serves well, and this is where machine learning comes in.
In this paper, we seek to use the Tweets captured within three weeks before the election to predict the voting outcome of the state of Florida – one of the largest and least predictable swing states.

2 Data Selection

2.1 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. Also, due to large amount of computation needed for extracting information from tweet data, we only her present relevant analysis on tweets related to Trump. In the future, we shall present tweets related to Clinton in a similar fashion.

2.2 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.

2.3 Baseline Selection

A major challenge in the methodology is the selection of baselines to fit the dataset upon. Since the election has not been held yet, we need to use another dataset to approximate the results. During discussion, we thought of two datasets that could be used as baselines:

- The voting result of each county in the primary election.
- The early voting result of each county, as released by the State of Florida.

Both of these baselines have inherent biases. Since Twitter Search API does not allow access of data as far back as the primaries, the first baseline does not capture how the public sentiment has changed in the past 8 months.
The second baseline is the most up-to-date information on Florida voters' voting behavior. However, the type of people who votes early are not necessarily the same type of people who votes on the general election. Additionally, the state of Florida only released the registered party of the voter (Democrat or Republican), instead of which candidate they actually voted for. According to a widely cited survey by YouGov, 91 percent of Democrats said they planned to vote for Hillary Clinton, while 82 percent of Republicans planned to vote for Donald Trump. Eight percent of those Republicans said they would vote for Clinton, while five percent of Democrats said they planned to vote for Trump. We adjusted the early voting result according to this survey.

3 Data Collected

We use the **TwitterR** package to search for tweets for the past week, via the *searchTwitter()* function. For the 67 counties in Florida, inputting each of their location, specified by their respective longitude, latitude and radius, gives us a list of tweets accessible from last week. However, due to the heavy cost of calling, compiling and computing data, we capped 3000 tweets for each county, thereby leaving out 7 counties with real large number of unmined tweets for future analysis.

4 Methodology

4.1 Feature Engineering

We find that it is hard to use sheer volume of tweets to predict a voting result, which is usually a percentage. Therefore, a percentage of all the factors is used to reflect a certain trend within the whole county, instead of a given community of tweet users. However, it must be mentioned that two assumptions are made here.

- Each county has the similar proportions of tweet users
- Different users are independent of each other.

All the features in **section 2.2** is divided by the population of that county. Since features are real numbers, we apply no further transformations.

4.2 Correlation

Before we begin our model selection,it is easy to notice that many of the features are highly correlated and it is reasonable to argue that since many of the features definitively depends on the volume of tweet. The high correlation is to be expected. Nevertheless, this correlation might interferes with our future analysis, which is to be improved before the final presentation of this project.

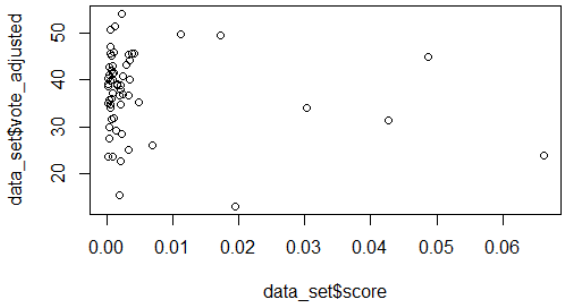
4.3 Model Fitting

Primarily, we fit different kinds of linear regressions on the data. To avoid overfitting the data, we set the degree of the regression to be one.
However, linear regression alone seems to yield relatively poor results, with none of the features having statistical significance. One possible explanation for such bad fit might be the existence of some high-leverage data points (see, for instance, the *sentiment score* vs *vote adjust* result)

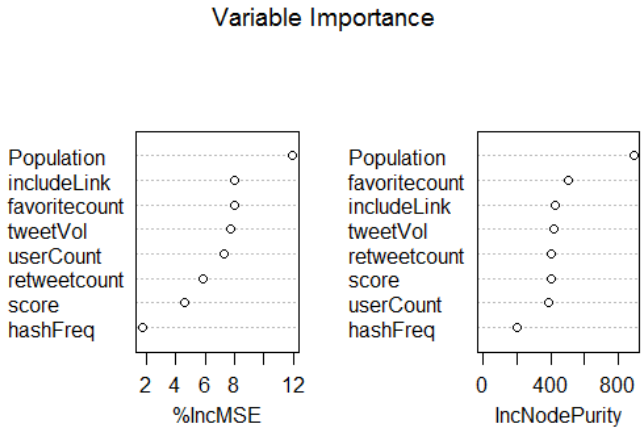
4.4 Model Selection

We use two methods to do model/feature selection:

- LASSO
- Random Forest Variance Importance Selection



Using LASSO, we are able to choose a model with lower variance overall. However, this method is not helpful because our best model will exclude 5 of our 7 features.
Random Forest Variance Importance Selection is another method to differentiate the significance level of different features. It is similar to bootstrapping data while calculating the reduction of mean square error when each feature is thrown out of the model. Here the plot on the left indicates the relative importance of our features.



5 Next Steps

5.1 Improving Sentiment Accuracy

We used a basic lexicon based sentiment score to capture the sentiments, which does not discriminate between some most frequently appeared words (e.g. "Crooked" is a term used heavily to express opposition to Hillary, and "Rigged" is a term used to express support for Trump) and other generic terms (e.g."Bad."). In reality, some terms should be assigned a higher weight in helping categorize the sentiment of the tweet.

5.2 Cleaning and Re-forming Data

We need to take data of possible wide ranges into account. One possible, thought not optimal way to do this is to normalize all the data. Furthermore, we shall compare the results computed from:

- *Raw Data*
- *Percentage/Ratio Data*
- *Adjusted data given new information about the population*
- *Adjusted data via a Bayesian method*

5.3 Improving Model

We will look into different types of model for significant fitting, prediction and interpretation. An alternative being considered is to use a logistic regression model.