Predicting the US Primary Election Results

Arjun Bora (110310134), Aditi Singh (110285096), Yunke Tian (109929662), Yinquan Hao ()

Motivation:

U.S. Presidential elections are not just important for the U.S. nationals, but also to the nationals of other countries. It becomes more important when the candidates have extreme ideas, strategies towards some particular issue, ethnicity and policy, like foreign policies.

This time it becomes even more exciting because of the fact that, if Hillary Clinton becomes the president, she would be the first women president; if Sanders become the president, he would be the first Jewish president and if Trump wins, he might be the first industrialist to become the president.

There can be lot of factors, like age, income, ethnicity, state, education level, sex that can affect the voting behavior of people. The more predictors we consider, the prediction becomes more accurate and challenging.

As the primaries elections are still going on, we can apply our model to the ongoing elections and verify our model. Thus, this is an interesting, easily verifiable on real results and this is why we are motivated to choose this project.

Approach:

Data Collection:

We are collecting data from many different sources:

1. **Twitter API** – to get the latest tweets (twitter has a limitation on the #tweets collected per user per query), using Python to download the twitter feeds using hash tags.
2. **Kaggle datasets** – GOP primary dataset and also the 2016 US election dataset (US primaries data).

Overview:

There are many factors that could influence the popularity of a candidate and the aim is to find out if there is some correlation between the demographic distribution of the country (age/education level/ race & ethnicity / income etc) and the candidate popularity, and if yes, then include these dimensions to come up with a multiple linear regression model to predict the final number of votes which would be acquired by the contesting candidates.

Factors: The below mentioned factors would be used to predict the votes in the primary for each candidate.

Data Source	Attribute	Reason
Twitter	Popularity of Candidates = (#tweets) x (weight of individual tweet)	The tweets would be weighted by the person's #follower. This would guarantee that genuine tweets only make through.
Twitter	Weighted Sentiment score based on the #retweets	This would add weights to the tweets and good tweets (the ones which are most retweeted) get more weightage in the final output than the others
Kaggle (2016 US Elections)	County wise demographics	There are several factors which characterize a particular county. We would be running PCA on the set of factors to get some 4-5 dimensions which characterize the counties best. These final dimensions would be used as factors in the multiple-linear regression model.
Kaggle (GOP debate analysis)	Sentiment scores based on first GOP debate data	A trend analysis in the decrease/ increase in the popularity of the candidates can be done. This would give us a view of the past data.

Further Analysis:

After getting the result for primaries, we want to predict the final election results.

After the result of primaries, voters can be divided into two parts:

1) Whose candidate choice won the primaries: We assume that those voters will vote for the same candidate in the final election.
2) Whose candidate choice lost the primaries: These voters now have two options
   i) Vote for the winning candidate of the same party his candidate for the primary belongs to, or
   ii) Change the party and vote for the winner of opposite party's candidate

Three Broad Categories: *Apart from the first phase of Probability/Discovery/Prediction for the primary elections we would also extend the model to incorporate the below mentioned points*

Probability: We, using the probability theory over the 2012 election data, will calculate the probability of a voter from a particular state will choose the extreme step (2) and change the party if his own candidate lost the primaries.

Discovery: We will find correlation between the demographic distribution and the final vote count of the winner of the primaries in a particular county.

Prediction: Applying this probability to our predicted primaries result, we can predict the number of votes for the final election, and thus predict Who Will be the president of the U.S.A. for 2016-2020.

Results Figure:

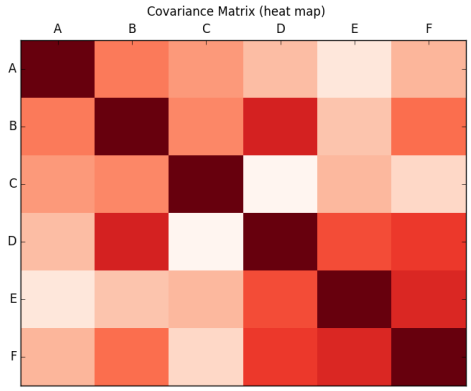


Fig1: Correlation heat map to find correlated Factors within demographics dimensions

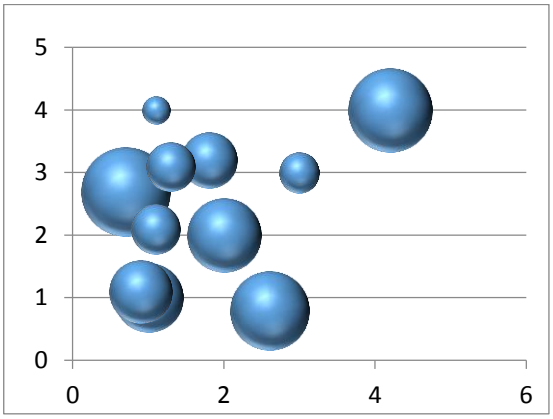
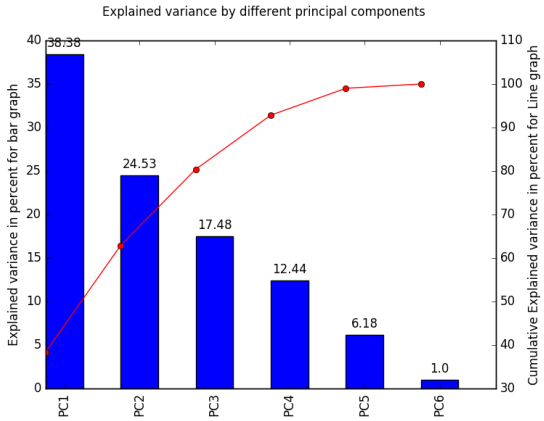


Fig2: scatterplots: relations between candidates &factors



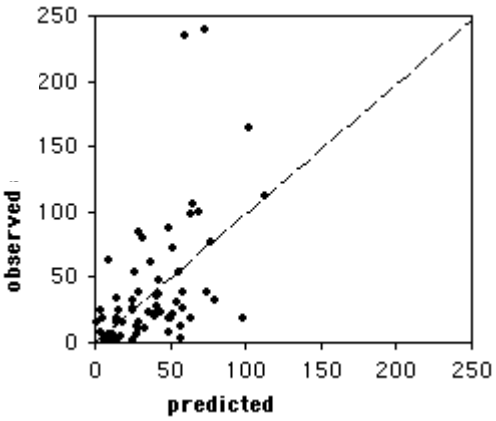


Fig4: shows the multiple linear regression model results wherein x axis would have votes predicted and y axis would have observed votes for each candidates. The accuracy is the output

Timeline:

Date	Complete
April 27	**DATA DISCOVERY PHASE:** Data collection & correlation between different factors. **PREDICTION PHASE1:** Run PCA and find out the relevant factors to be included into the multiple-linear model for predicting the primary results in the upcoming primaries.
May 3	**PREDICTION PHASE1:** Start the prediction and get the accuracy for few test counties/states.
May 7	**PREDICTION PHASE2:** Use the model of the primaries to work on the final election results.
May 12	**FINALIZE:** Finalize the model and report the accuracy