Project for Data Mining (520-40) (Progress/Status)

RaviTeja Manda
CISC-520-40
Data Mining
Harrisburg University
Harrisburg, Pennsylvania
Email: VKMANDA@my.harrisburgu.edu

***Abstract*—The recent elections in the United States was one of the interesting elections in the recent history of the country and there are definitely a lot of intriguing questions to find answers. This is an approach to define and question the rationale behind forecasting models that try to predict the presidential race. Among umpteen data analyzing tools, data mining is one of the practices of looking data in a perspective to allow provide better insight in to the dataset. One of the interesting reasons for the need to look deeper in to the datasets is to help find answers to the relation between causality and correlation within the data. As data scientists love to point out, numbers do not lie, and understanding the reasoning behind the difference in national popular vote and electoral college vote outcome is always a fascinating area. The presidential race of 2016 has seen lot of polling agencies failing to predict the result accurately and this is generally unconventional to have as many models to be unsuccessful in forecasting the outcome.This is an approach to use statistical analysis and data mining to find the reasons for the major difference in polling models that have played a part in the presidential race of 2016 in United States. KeywordsBigData, Data mining, Election Polls, Popular vote, R, statistics, correlation, Deep learning, machine learning.**

Keywords*—*IEEEtran, journal, LaTeX, paper, template.

I. Introduction

It is a common practice to have different agencies to conduct opinion polls to find the opinion of a sample of population and then extrapolating it across the medium. And a prevalent practice for using them in public relations at the state and national level for predicting the pulse of any elections and a commonplace for presidential races. There are lot of different models for polling and each of them are done at various stages. Like any mathematical model that incorporates extrapolation in to its system, they are subject to introduction of errors and it is very important to reduce the inaccuracies.

The presidential elections of 2016 in United States have been particularly interesting for various factors and there are multiple forums on the global front that have closely followed them through the race. The candidates from both parties, Democratic and Republican parties, had significantly different views, policies and employed different campaigning styles. And for data enthusiasts, this election cycle has been unusually interesting for the different series of events that have resulted in the outcome of the elections.

A. Mining Polling Data

With the presence of high computing and storage capabilities and the ease of access to such machines, big data applications have helped save large amounts of time in solving complex problems. There are lot of patterns that can be followed and found within the data when looked in the intended way. Historically, for all political elections since 1936 there were multiple agencies that have conducted opinion polls to forecast and predict the president at various stages before the Election Day.

B. Intent of the Project

There was a huge disparity in multiple polling agencies results to the electoral vote outcome on the Election Day in the United States presidential elections of 2016. This effort would try to look in to the polling data and the electoral vote result and find out the reasons for the variance in the prediction models result and help find factors that could have resulted in returning erroneous results. Various individual and aggregate polling data with varying sample sizes in a broader range would be used to compare and study. Considering the political situation in United States being bipartisan, there are multiple models that do not include independent candidates in their studies. This effort will also take in to consideration of the independent candidates that were contesting in the presidential race while reviewing the data to analyze their impact on the overall result.

II. Understanding US Presidential Elections

United States is one of the leading nations in military power and is a major power source on global front. Also, it being considered as the leader of the free world makes their election cycle something that can have an influence on the entire world. The US presidential election involves picking a leader from each party, Democratic and The Republican Party who go through an indirect voting by ballots cast be eligible candidates for the members of the electoral college who in turn cast their direct electoral votes to pick the President and the Vice-President of the United States.

The votes cast by the eligible citizens of the different fifty states of United States and the District of Columbia are considered as the popular votes to pick the member of the Electoral College. And, the votes cast by the elected Electoral College member, referred to as the electoral votes pick the President and the Vice-President of the United States.

And, the other uncommon thing that this election cycle has observed is the result of the popular vote winner not winning the election race. This is happened only twice in the last hundred years of the United States Presidential election race. Most of the pollsters that have polled and surveyed through the election cycle have predicted the Democratic candidate Hillary to win the race, while the outcome has been completely different to have given the Republican candidate, Donald Trump the presidency.

III. DATA-SET AND DATA-MINING TOOLS

Although there are multiple factors that can be found to be reasons for the predictive models from various pollsters to have gone wrong this particular study would concentrate majorly on the statistical problems that could have resulted in returning inaccurate results. To provide an example, comparing the census data to the polling size would help understand the errors that can be introduced while under-sampling and extrapolating datasets. At the same time, this model will also look in to major social networking data streams to understand any changes in how a particular candidate has increased or decreased their chances through their social networking profile propaganda. Although there is high speculation that the president-elect Donald Trump has used more social networking advertising than any of the presidential candidates, there are no analytics to prove that and this model will try to gather enough information to prove or disprove such a statement.

Causal and correlation study on the datasets of polling data and electoral vote results would be studied along with looking at the census data to understand the effects of undersampling and oversampling. R-programming would be used to simulate and build the model. The model will also try to validate and verify the credibility of the sample that is participating in the opinion polls. The model will also try to compare its results to any of the election data of previous presidential elections in the United States.

Electoral data would be downloaded from United States Electoral College available from the National Archives and Records Administration. Polling data would be downloaded from models ranging from being very accurate to completely inaccurate to provide for a better distribution of the data. Datasets that would be looked in to but not limited to would be from the Nate Silver Project 538, UPI/CVoter poll and the University of Southern California/ Los Angeles Times poll, 270 to win and the New York Times poll.

A. Data Collection

Data Collection is one of the important steps of Data Mining and in this entire project there was no data manipulation that was done. The data that was collected to study and understand the presidential elections of 2016 is from genuine sources and was only collected from the state board of elections websites. The census data was also collected from the Investigative Reporters and Editors and there was no tampering done with the data. The polling information that was collected was from National Polls data from the Princeton University. The last time the census data for the united states was collected in the year of 2010.

B. Data Cleaning

Data that was collected from all of the above mentioned sources was either collected in the form of a csv file or regular text files. Regular shell scripting was used to automate the collection of the census data from the websites and the csv files were processed to remove the headers before they were loaded on for further data analysis in R.

C. Data Preparation

Data Preparation is one of the significant steps in data analytics and is often neglected in the first stages of planning a Big Data application for mining. When thought through and planned for it ahead, it not only allows for significant cost savings on resources it also reduces the complexity of the models. Considering, there were multiple sources of data that was collected through the process, both structured (csv and excel formatted files) and unstructured (raw text files) there was a huge amount of effort in making the data available in single format before loading on to the data analytics tool, R.

D. Data Understanding

Data understanding and data analytics is often confused to be a single stage of data mining process but when the data is constantly added on to the data analytics engine, it is highty recommended to allow for proper understanding of what data is being loaded in to the system. This reduces the complexity of the model and allows for being able to differentiate and debug issues easily.

E. Data Analytics

Different things that were studied as part of the work done till now, included the following:

Collect census data from a genuine source.

Collect election results data from genuine sources.

Collect national poll data from multiple sources

Clean the collected data to be able to feed in to data analytics engine in a common format.

Data analytics engine that was used is R, and any shell scripting was used to write automated scripts to collect data from the mentioned online resources.

Started a twitter application to automatically collect twitter content from the user accounts of the presidential candidates and other significant hashtags used during the polling data.

Comparison of electoral votes and popular votes for various menioned states was done and plotted in various formats to allow for better understanding.

All content would be uploaded to github account under the project : Data Mining

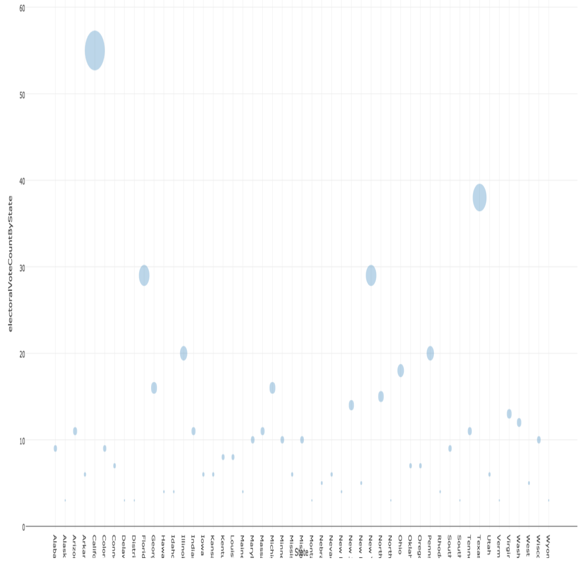
Different things that needs to be addressed in the next couple of weeks :

Study the timeline of different tweets and see the significant trend/change in the popular vote among the presidential candidates.

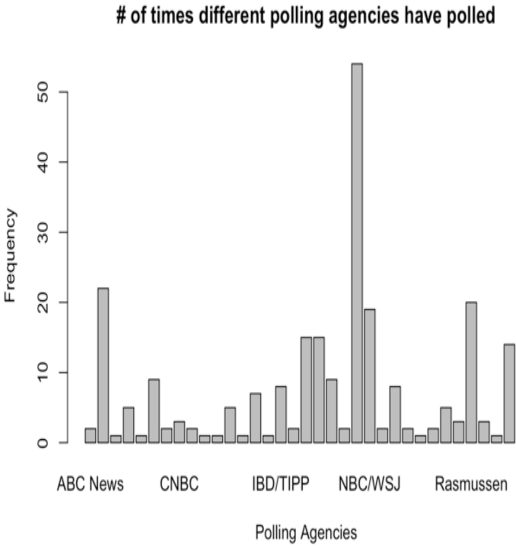
Correlation analysis on the mentioned significant changes in the trend of the change in popular votes to electoral vote results.

Understand a prediction model and study the reasons for their failures and map them to statistical distributions like under-sampling and over-sampling.

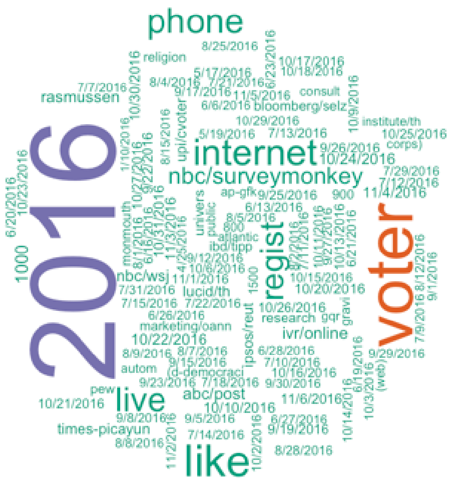
F. *Screenshots of the Work Till Now*



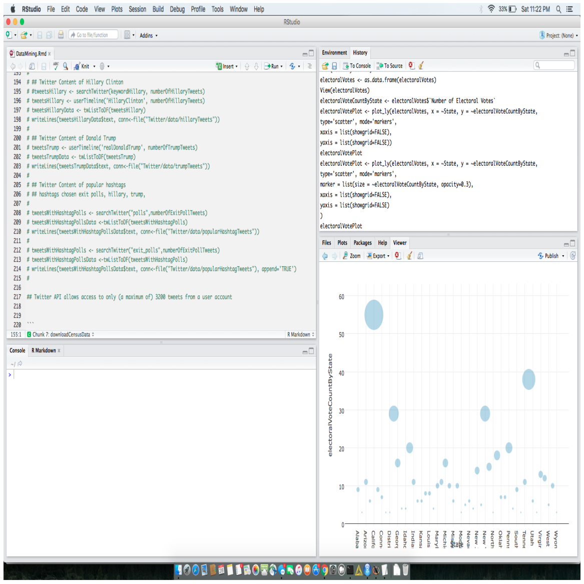
Electoral Vote Count Distribution



Polling Agency Frequency



Agency Frequency
Polling Agency Frequency



```
# ## Twitter Content of Hillary Clinton
# #tweetsHillary <- searchTwitter(keywordHillary, numberOfHillaryTweets)
# tweetsHillary <- userTimeline('HillaryClinton', numberOfHillaryTweets)
# tweetsHillaryData <- twListToDF(tweetsHillary)
# writeLines(tweetsHillaryData$text, conn<-file("Twitter/data/hillaryTweets"))
#
# ## Twitter Content of Donald Trump
# tweetsTrump <- userTimeline('realDonaldTrump', numberOfTrumpTweets)
# tweetsTrumpData <- twListToDF(tweetsTrump)
# writeLines(tweetsTrumpData$text, conn<-file("Twitter/data/trumpTweets"))
#
# ## Twitter Content of popular hashtags
# ## hashtags chosen exit polls, hillary, trump,
#
# tweetsWithHashtagPolls <- searchTwitter("polls",numberOfExitPollTweets)
# tweetsWithHashtagPollsData <-twListToDF(tweetsWithHashtagPolls)
# writeLines(tweetsWithHashtagPollsData$text, conn<-file("Twitter/data/popularHashtagTweets"))
#
# tweetsWithHashtagPolls <- searchTwitter("exit_polls",numberOfExitPollTweets)
# tweetsWithHashtagPollsData <-twListToDF(tweetsWithHashtagPolls)
# writeLines(tweetsWithHashtagPollsData$text, conn<-file("Twitter/data/popularHashtagTweets"), append='TRUE')
#


## Twitter API allows access to only (a maximum of) 3200 tweets from a user account



```
```

```
electoralVotes <- as.data.frame(electoralVotes)
View(electoralVotes)
electoralVoteCountByState <- electoralVotes$`Number of Electoral Votes`
electoralVotePlot <- plot_ly(electoralVotes, x = ~State, y = ~electoralVoteCountByState,
type='scatter', mode='markers',
xaxis = list(showgrid=FALSE),
yaxis = list(showgrid=FALSE))
electoralVotePlot
electoralVotePlot <- plot_ly(electoralVotes, x = ~State, y = ~electoralVoteCountByState,
type='scatter', mode='markers',
marker = list(size = ~electoralVoteCountByState, opacity=0.3),
xaxis = list(showgrid=FALSE),
yaxis = list(showgrid=FALSE)
)
electoralVotePlot
```

Programming Environment

IV. Conclusion

Learning different data analytics stages through working on presidential elections data of 2016 of the United States and simultaneously looking in to finding observations for the reasons that could have made the difference in the outcome of the presidential race is achieved. The source code and other analysis would be made available for public over github at handle, mvkraviteja and the repository Data Mining.

Acknowledgment

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] https://github.com/mvkraviteja/DataMining

[3] http://election.princeton.edu/code/data/2016$_{N}ationalPolls.csvhttps : //en.wikip$

[4]

[5] http://er.ncsbe.gov/contest$_{d}etails.html?election_{d}t = 11/08/2016county_{i}d = 0contest_{i}d = 1001$
```

[6] https://results.elections.myflorida.com/downloadresults.asp?ElectionDate=11/8/2016DATAMODE=

[7] http://historical.elections.virginia.gov/elections/view/80871/

[8] http://www.electionreturns.pa.gov/ENR$_N EW$

[9] http://www.sos.state.oh.us/SOS/elections/Research/electResultsMain/2016Results.aspx

[10] http://www.michigan.gov/sos/0,4670,7-127-1633$_8 722 - 397762 -$
$-, 00.html$

[11] http://elections.wi.gov/elections-voting/results

[12]