Project Proposal for Data Mining Class (520-40)

RaviTeja Manda
Harrisburg University
Harrisburg, PA
Email: vkmanda@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.**

***Keywords*—BigData, Data mining, Election Polls, Popular vote, R, statistics, correlation, Deep learning, machine learning.**

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 of 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.

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

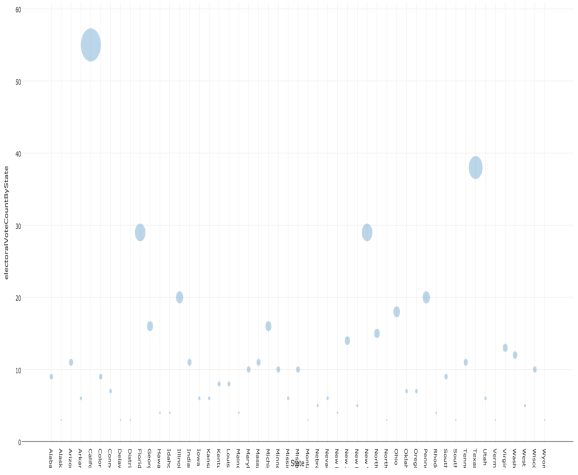
There was a huge disparity in multiple polling agencies results to the electoral vote outcome on the Election Day in the United States presidential elections of 2016. This effort would try to look in to the polling data and the electoral vote result and find out the reasons for the variance in the prediction 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. 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.



III. Conclusion

The model that is intended to be built would be using programming language R and its associated packages to visualize its findings about the reasons that it has found to have resulted in inaccurate prediction of 2016 presidential race in United States.

Appendix A

Causality and Correlation in R

Polling Data

Electoral College

Census Data

Acknowledgment

Will be taking in help from the professor of the class (Daqing Yun) for helping me steer and scope the work and any work used to help the study would be referenced accordingly.

References

[1]