Political Profiling using Feature Engineering and NLP *

Chiranjeevi Mallavarapu[1], Ramya Mandava[1], Sabitri KC[1], Ginger Holt[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX-75275, USA
[2] Facebook
{cmallavarapu, rmandava, skc}@smu.edu
gholt@facebook.com

Abstract. In present day scenario, forecast of election outcomes are based on public surveys. These are sometimes inaccurate.It poses a huge problem for Political parties to accurately spend their campaign budget.Hence, the need for a more accurate methodology to predict election outcome.In this paper, it has been identified that Dynamic public data from open data sources on politicians has influential features in determining the outcome of elections. Our model yielded 0.5 correlation between derived features and percentage of votes being predicted. Hence, the election outcomes are also dependent upon features such as a candidates affiliation with featured organizations, their presence in public events, their educational background etc,.

1 Introduction

Political election outcome is an important topic of discussion that occurs every year. News media and political analysts are always on the run to find a winning candidate, especially in constituencies where the competition is high. In general, forecast of political outcomes is based on voter polls through multiple mediums such as digital polls, telephone surveys etc. Sometimes, these predictions are inaccurate due to bias in user polls. This has a major impact on campaign planning and budgets spend for political parties. By having an unbiased prediction, it will help the political candidates and the parties to better manage their campaign spend.

In the recent past, political predictions have been inaccurate to the surprise of audience. Media and political analysts identified the problem stemming from the inadequacy of voter opinion and their overall participation at the time of polling. These inconsistencies make the political spending by the parties and candidates ineffective.

This paper presents a solution to this problem by collecting dynamic public data from open data sources on politicians. Plethora of open data is available in recent times that is relevant to this problem. It has been identified that Google's

* Supported by organization SMU

Knowledge Graph, Wikipedia, and NewsAPI provides a 360 degree view of data entities that can collectively address this issue. These data sources have been integrated into our analysis through restful APIs.

Feature engineering techniques have been applied and features such as a candidates affiliation with featured organizations, their presence in public events, their educational background etc,. showed high influence to predict percentage of votes. Random forest regression model with cross validation resulted a Pearson Correlation coefficient of 0.5.

Based on our analysis and modeling, it can be concluded that the voting percentage can be predicted based on a politician's association with noted organizations, their presence in public events, being part of news etc,. Our model performs above average in predicting the percentage of votes for politicians that have low to mediocre performance(up-to 30 percent votes within the state). However, our model doesn't perform well when predicting votes for lead runners.

NOTE: Update the this paragraph after rest of the sections are updated Section 2 provides a brief overview of some of the notable attempts to predict U.S. presidential elections. The section purposely does not cover any models that rely on polling and instead focuses specifically on models that use other methods. The emphasis is on models that contribute to explaining why an election outcome occurs

2 Prior Research

Other sections are here.

3 Data Acquisition and exploration

Based on initial assessment of multiple relevant data sources, it has been determined that open schematic data that is constantly updated is the best way to approach this problem. These factors resulted in identifying Google's Knowledge Graph, Wikipedia Data and NewsAPI as the leading data sources in their respective fields.

3.1 Data Collection

Googles Knowledge graph A knowledge graph, shows relationships between real world entities and describes them in an organized graph. It has potentially interrelating arbitrary entities with each other across various topical domains.[10] Google's knowledge graph has been leading the industry in this area.

Googles Knowledge graph[3] provides an API which can be used to download certain information about a politician. Google collects public information from

[3] https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

multiple sources and uses proprietary algorithms to rank and rate the information and structure it in a standard format that can be pulled using its knowledge graph API. The data is structured as per standard data types defined in standard schema organization[4].Key schema types in relation to politicians have been identified as : Person, Event and Organization.

Wikipedia Wikipedia[5] being an authentic open data source, it has been chosen to extract candidate biographic information such as educational background, date of birth, occupation and political history as available on their Wikipedia page. Wikipedia allows for edits from end users from all over the world, hence cross validating the data being entered and makes it authentic.

Wikipedia rest API along with pywikibot is used to get full text of Wikipedia pages as well as standard templates of data from Wikidata for the 2016 Senates across all states. Even though mostly Wikipedias information about a politician is divided into biography, personal section and political career section, Pywikibot gives the necessary tools to extract individual elements.

2016 candidate Wikipedia pages Data about the politicians voting outcome has been collected from Wikipedia 2016 senator page, using JSON download to get the total number of votes they received to evaluate the accuracy of our features.

Newsapi.org News API[6] has been considered as an additional data source to crawl, index and monitor the top news related to every politician from over 30,000 news sources and blogs. It is a simple HTTP REST API for searching and retrieving any news article from the Internet based on different criteria like keyword, date published, language, domain etc. and can be sorted in different order for e.g.: date published, popularity of source, number of social shares etc. Hence it curates the data from thousands of different sources and serves as a great unbiased data source for our project. Information about the latest news on politician has been gathered from NewsAPI using their REST API.

3.2 Data Exploration

Our consolidated data set includes Wikipedia 2016 senator candidate **names** and **votes**, corresponding Google Knowledge Graph **Events** and **Organizations**, NewsAPI **Articles** and Wikipedia pages.

Data exploration has been started with California as our sample state which had 37 candidates that ran for Senate elections in 2016. Out of the 37 candidates, 7 candidates have a Wikipedia page, 13 candidates have presence on Google Knowledge Graph and 17 candidates have been mentioned in articles derived

[4] https://schema.org/
[5] https://en.wikipedia.org/
[6] https://newsapi.org/

from NewsAPI. Features for further analysis have been defined as **number of organizations** a candidate is associated with, **number of events** a candidate is associated with, **number of News articles** associated with a candidate and whether the candidate has **Wikipedia page** and **Google presence**. The data exploration then has been extended to the remaining states and candidates.

A pair-plot based on these derived feature data-set indicated a clear correlation between **number of Votes** and the features identified as shown in Fig 1.**NOTE: To be updated for draft 3 change the sns plot to show wiki and google features and remove no of opponents.**

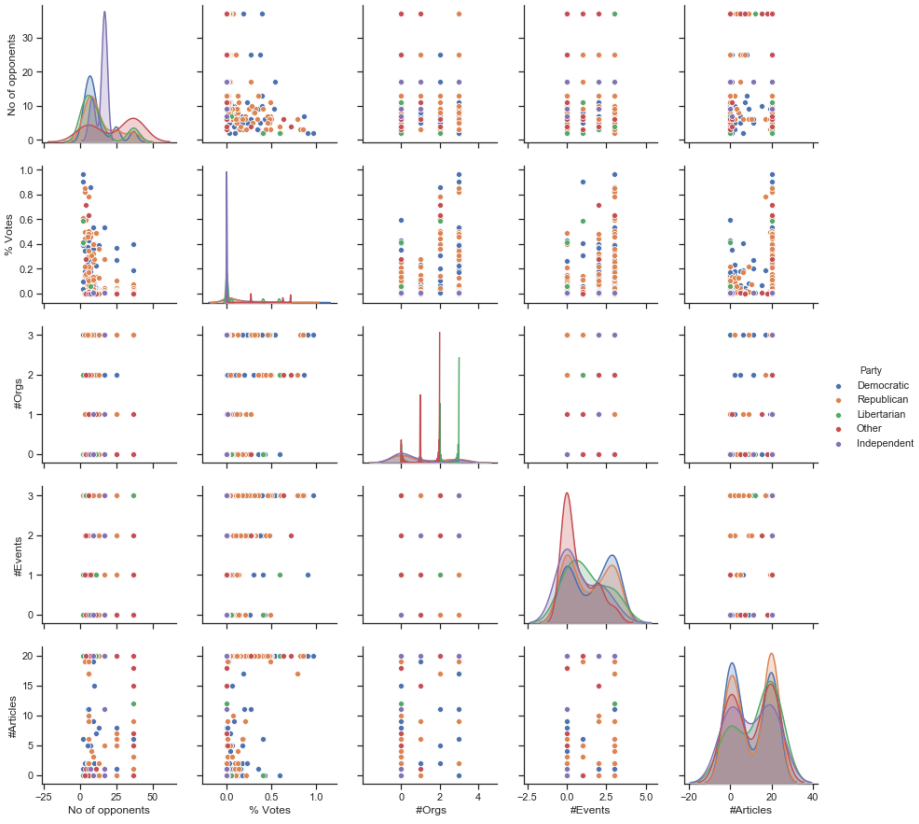

Fig. 1. Pair plot of engineered features

It is apparent that the percentage votes vary based on the number of events and organizations a candidate is associated with. However, there is no straight forward linear relationship. Two different patterns can be seen here on Votes vs Articles: One that would not change the votes based on increase in a candidate's number of news articles, another a positive correlation between the number of articles and votes can be observed.

Percentage of votes vs Orgs As observed in Fig.2, it is noticed that the people that have association with organizations seems to have a positive impact on the number of votes. It is also observed that people that don't have association with organizations also seems to have a certain effect on the votes. The plot has been color-coded with number of events as color-bar and it can be noted from Fig.2 that the higher the number of organizations, it is also likely for the candidate to have a higher number of events associated with him/her. Few instances can be observed where higher number of orgs are associated with lower number of events for a candidate, hence a need to separately explore events vs percentage of votes.

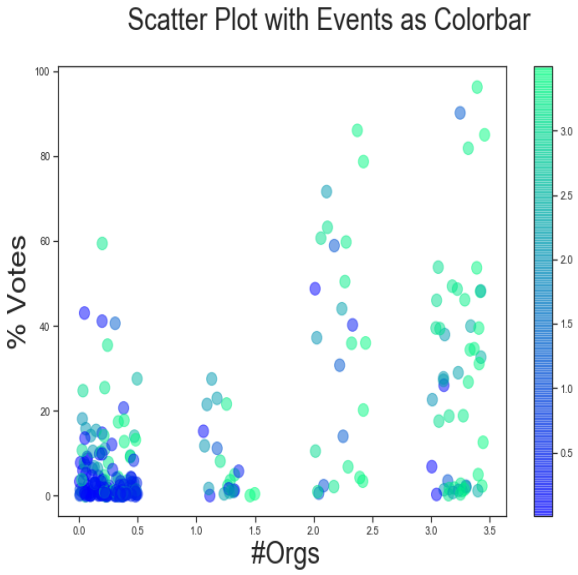

Fig. 2. *Pair plot of engineered features*

In order to understand the cluster of candidates that have non-zero votes with zero association with organizations, the included color scheme based on number of events that they are associated with as depicted in Fig.2. It is evident from these plots that, all candidates that have 2 percent or more votes either have mostly non-zero events, suggesting that events features is also influential in those cases.

Percentage of votes vs Events A positive correlation can be observed between the number of events and percentage of votes from Fig.3. It is also observed that when the number of events is more than 2, candidates associated with 2 or more organizations generally seems to have more percentage of votes.

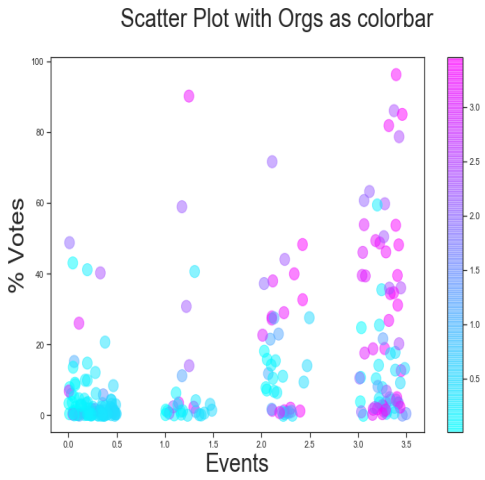

Fig. 3. Scatter plot of No: of Orgs vs. Percentage of Votes - Events

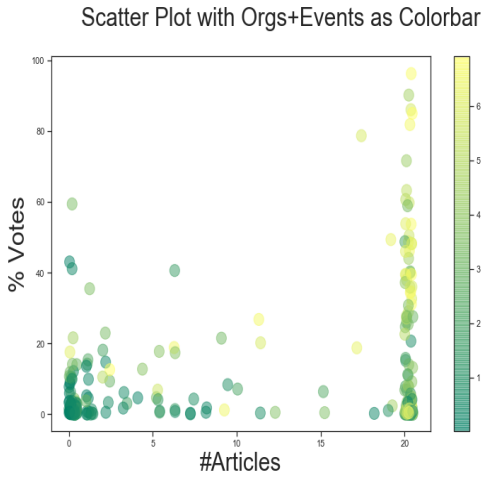

Fig. 4. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

Percentage of votes vs Articles As can be seen in Fig.4, there is a general positive correlation between number of articles and percentage of votes. However, for all the candidates where we noticed the number of articles are more than 7.5 but with small percentage of votes, it is verified that the articles do not belong to the political candidate but a different person with the same name. The candidates with low percentage of votes and with higher number of articles seem to have low number of events+articles associated with them. This confirms the findings from Fig.2 and Fig.3.**NOTE: To be updated for draft 3 This prompted us to further validate the accuracy of these articles from**

news API using NLP techniques which we will explore for the next revision of this document.

NOTE: To be updated for draft 3 Add exploration plots for Wiki and Google presence. Descriptive statistics for the features identified above, and for all states and candidates are shown in Table 3

Table 1. Table captions should be placed above the tables.

Stats	Votes	Wiki	Google	No:Orgs	No:Events	No:Articles
count	244	244	244	244	244	244
mean	0.131	0.422	0.627	0.897	1.368	10.520
std	0.196	10	10	10	10	10
min	0.000001	10	10	10	10	10
25 percent	0.008	10	10	10	10	10
50 percent	0.033	10	10	10	10	10
75 percent	0.174	10	10	10	10	10
max	0.961	10	10	10	10	10

4 Data modeling and Evaluation

Random Forest is a supervised learning algorithm. As the name suggests, it creates a forest and makes it somehow random. The "forest it builds, is an ensemble of Decision Trees, most of the time trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result as shown in Fig.5.[11]

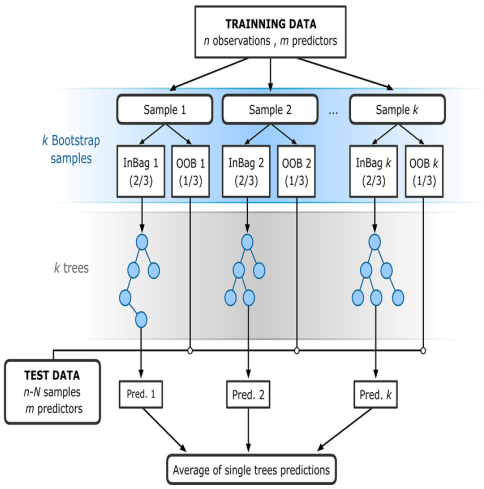

Fig. 5. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Based on the above analysis and data exploration, it has been determined that a model can be formulated using the five features as shown in Table 3 with percentage of Votes as the output variable.

As a first iteration, Random Forest Regressor with 150 estimators, yielded a Pearson correlation of 0.5493 and MSE of 0.0355. The data has been divided to have an 50/50 split between training and testing data.

Further, new features such as no: of opponents and party affiliation of the candidate have been added to the Random Forest model with cross validation(cv=5), that increased the Pearson correlation to 0.79 with MSE of 0.02.

New features have been derived based on words with highest(top 10) Term-Frequencies from the Wikipedia pages of candidates with greater than 50 percent of votes.The features values are **tf-idf** scores of the above identified top frequency words, as shown in Table 3. These features are added to the Random Forest model in addition to the ones from the previous models. This new model with 5-fold cross-validation on the subset of candidates with Wikipedia page(103 candidates) yielded a Pearson Correlation of 0.709 and MSE of 0.019.

Table 2. Additional Features - based on Term Frequencies(NLP)

congress	president
republican	bill
election	act
house	committee
senator	senate

Table 3 shows the various models used and the corresponding parameters optimized for each model.

Table 3. Model Evaluation

Iteration	Data split	no: of estimators	MSE	Pearson Correlation
1	Train-Test: 50-50	150	0.014	0.79
2	Train-Test: 50-50	150	0.014	0.79
3	Train-Test: 50-50	150	0.014	0.79

5 Results and Analysis

Based on the above modeling and evaluation, for iteration 1, as depicted in Fig 5., it shows the predicted percentage of votes against the actual given by Random Forest Regressor. It can also be noted that the prediction percentage of votes are mostly being flattened at 30 percent. Considerable amount of the predictions are not in alignment with the Actual percentage of votes(away from the 45 degree line).

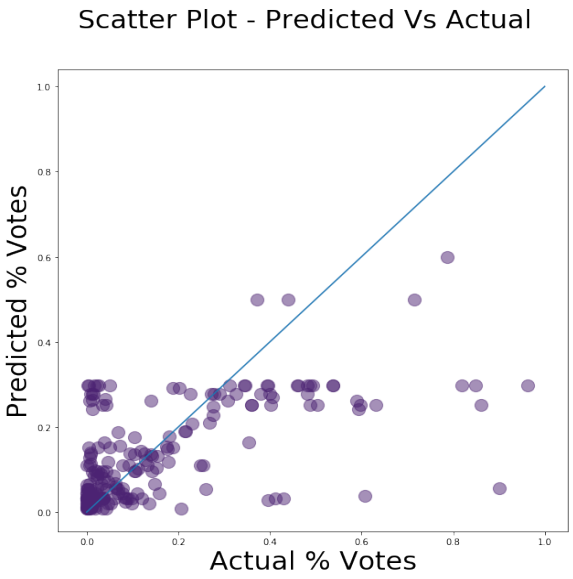

Fig. 6. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

As mentioned in the previous section, for iteration 2, after including new features such as 'No: of opponents' and 'Party affiliation', as shown in Fig.6, the prediction accuracy has improved in comparison to previous model. However, candidates with very high percentage of Actual votes(greater than 80 percent) are not in alignment with the predictions(away from the 45 degree line).

As indicated in Fig.6, it is noticed that majority of these candidates with high percentage of actual votes have a Wikipedia page associated with them. The red points represent candidates with Wikipedia page associated with them and the blue points represent candidates without a Wikipedia page associated with them.

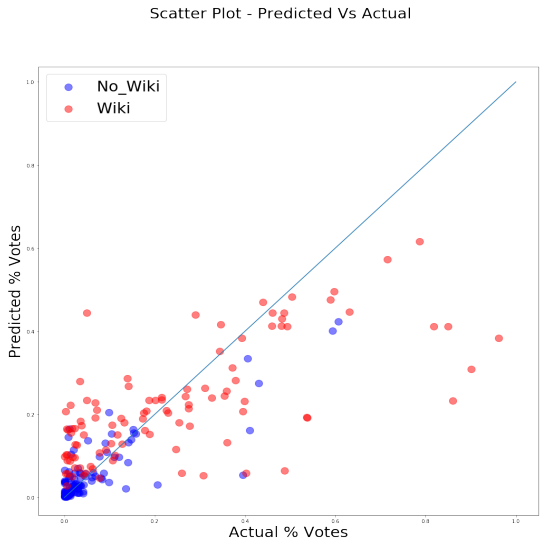

Fig. 7. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

For iteration 3, after including NLP features for the subset of candidates with Wikipedia page(103 total) and as shown in Table 2, the results of Predicted vs Actual votes are shown in Fig.7. Predictions for candidates with very high percentage of Actual votes(greater than 80 percent) improved in comparison to the previous iteration. In addition, majority of the predictions are very close to the actual percentage of votes.

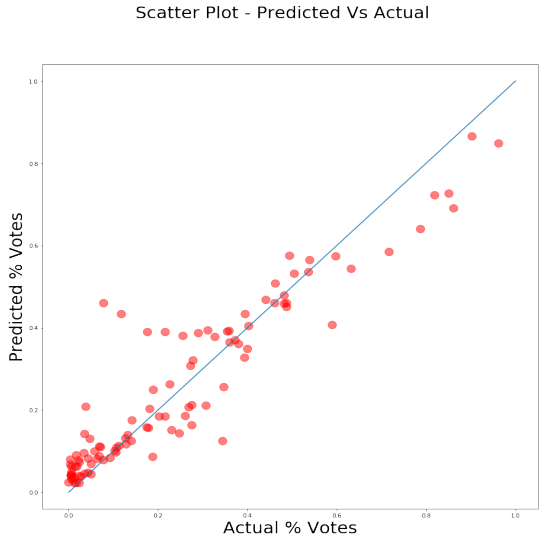

Fig. 8. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

After it has been determined that this model holds good,the Random Forest Regressor prediction was plotted against the actual votes as indicated in Fig.8. The proximity of prediction in black line to the actual votes in red dots further validates the model visually.

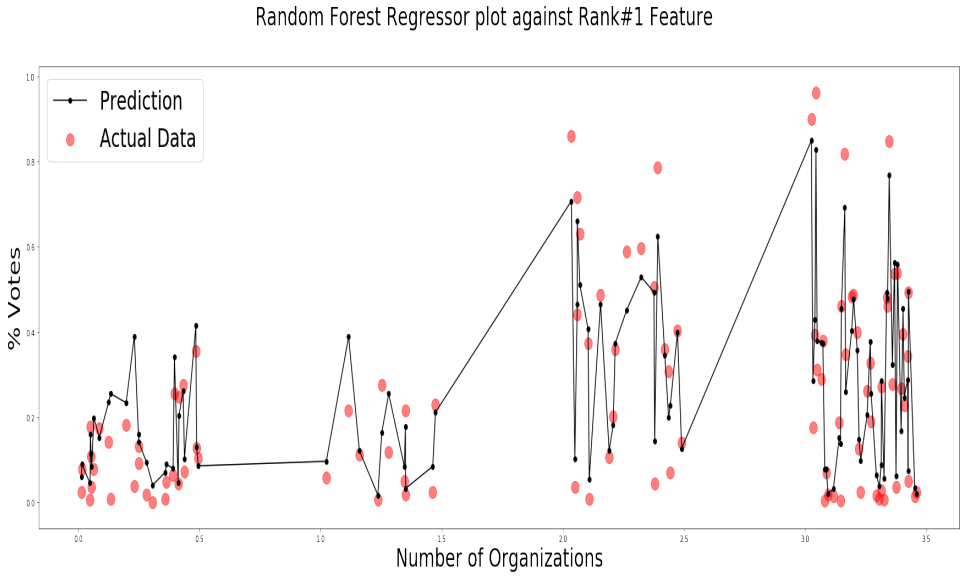

Fig. 9. Random Forest Regressor - Prediction vs Rank 1 feature

The prediction data between the values of 2 and 4 for number of organizations is further zoomed-in and it indicates the model's flexibility in handling the complex scenarios between those data points as shown in Fig.9.

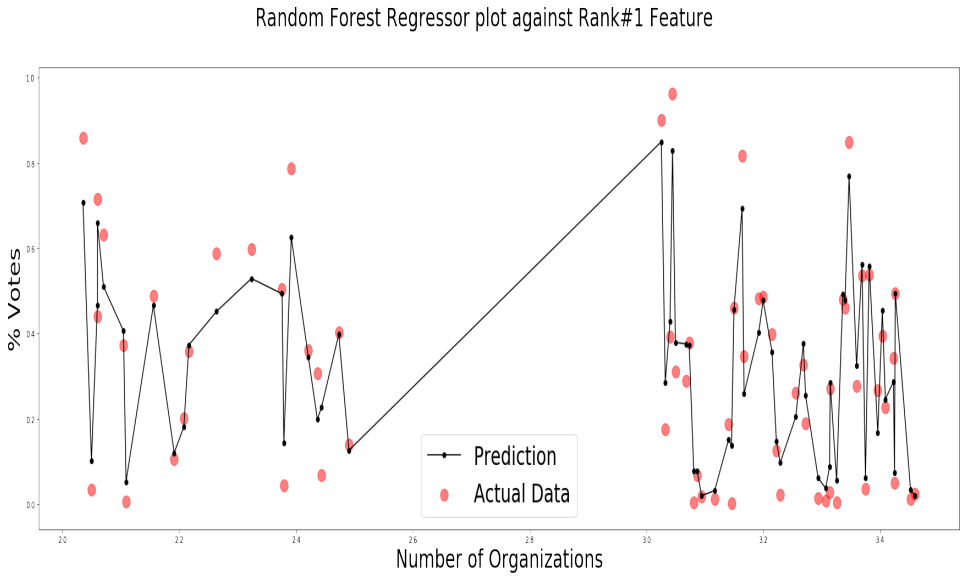

Fig. 10. Random Forest Regressor - Prediction vs Rank 1 feature

Fig.10 shows the results of the ensemble model, that has both sets of data(candidates with and without Wikipedia pages) and the associated features. It is observed that the candidates with high percentage of Actual votes are being predicted more accurately than the previous models due to the addition of NLP features. The candidates with no Wikipedia data associated with them tend to have lower percentage of votes as can be seen in Fig.8.

It is also noted that the 5 candidates(red tail on the left) with lower percentage of Actual votes are not being predicted well. However, these results were not consistent and changed over various iterations of 5-fold cross validation of the Random Forest model, while others being consistent. This is an indication of limited amount of data. But this variation is seen only on less than 5 percent of the data points.

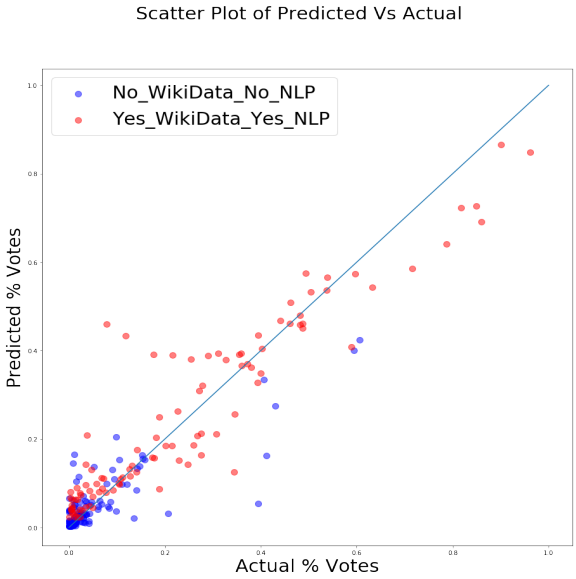

Fig. 11. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

Feature ranking as shown in Fig.11, indicates that '**Number of organizations**' and other NLP features such as '**congress**', '**republicanword**', '**house**', '**bill**' and '**election**' explains 80 percent variation in the percentage of votes(response).

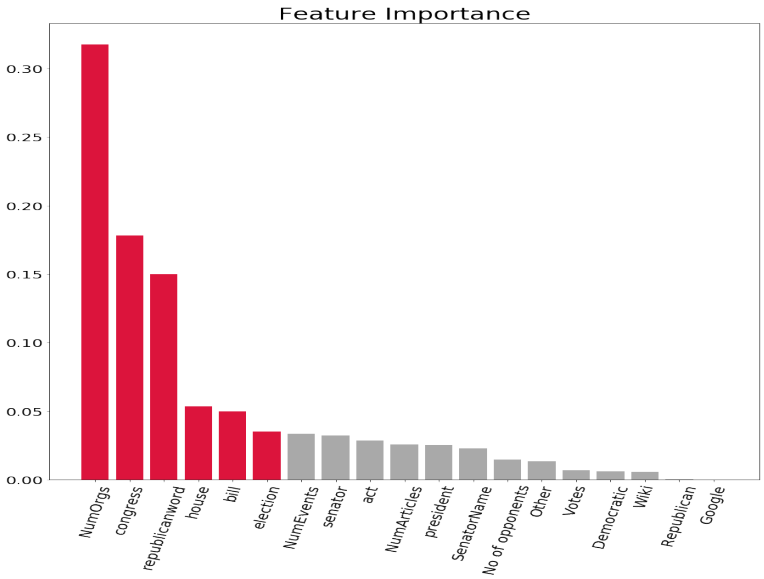

Fig. 12. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

The final set of features are as indicated in Table 4(Features for base model without NLP) and Table 5(Features for base model with NLP).

Table 4. Base Features - without NLP

Google	No: of Articles
Wiki	No: of Opponents
No: of Events	Party
No: of Organizations	

Table 5. Additional Features - with NLP

congress	president
republicanword	bill
election	act
house	committee
senator	senate

6 Ethics

The three data sources used for our project(Google Knowledge Graph, Wikipedia, News API) are public and open data sources. None of these require authentication or approval by any entity.

Public data from social media platforms that might have legal implications has not been considered for our analysis. Data anonymity has been maintained throughout our analysis as well as documentation.

7 Conclusions and Future Work

Our final model is an ensemble of two Random Forest Regressor models with different feature sets as shown in Table 4 and Table 5. The feature set being used depends on whether a candidate has a Wikipedia page or not.

It has been noticed that candidates with Wikipedia presence are more likely to win. 80 percent of variation in response is determined by the identified 6 features as shown in Fig 11.

This analysis can be extended to take user input and be converted to a recommendation system for indecisive voters. This can also be extended to predict other election scenarios. New NLP features based on News articles can be engineered for the extended model that might have much deeper understanding using sentiment analysis of the content, thus deriving insights into candidates stand on open issues.

Acknowledgments

Authors would like to thank YYYYY.

References

1. A. Einstein, On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat, Annalen der Physik 17, pp. 549-560, 1905.
1] http://journals.sagepub.com/doi/pdf/10.1177/2053951716645828 [2] https://link.springer.com/article/10.1007/s00146-014-0549-4 [3] https://en.wikipedia.org/wiki/UnitedStatesSenateelections,2 016 [4]https : //www.quora.com/What are the free news feed APIs [5]https : //newsapi.org/docs/get started [6]http : //www.euclid project.eu/iBook [7]http : //knowledgegraph.info/ [8]https : //everypolitician.org/unitedstatesofamerica/senate/download.html [9]https : //vote ca.org/Intro.aspx?State = CAId = CAHarrisKamalaD [10]https : //www.mysociety.org/files/2014/11/FOIImpactP art2Practitioner Study 06.pdf [11]https : //newsapi.org/docs [12]http : //docs.everypolitician.org/technical.html [13]http : //docs.everypolitician.org/repostructure.html [14]Stoffa,J.,Lisbona,R.,Farrar,C.,Martos,M.(2018). Retrievedfromhttps : //scholar.smu.edu/cgi/viewcontent.cgi?article = 1005context = datasciencereview; [15]http : //www.unz.com/runz/a last minute decision to enter theussenateraceincalifornia/
[10]H.Paulheim. Knowledge Graph Refinement: A Surcery of Approaches and Evaluation Methods, Semantic Web Journal,(Preprint):1-20, 2016.
[11]Rodriguez-Galiano, Victor Snchez Castillo, Manuel Dash, Jadunandan Atkinson, Peter Ojeda-Zujar, Jose. (2016). Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Biogeosciences. 13. 3305-3317. 10.5194/bg-13-3305-2016.