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 model includes Wikipedia 2016 senator candidate **names** and **votes**, corresponding Google Knowledge Graph resulted **Events** and **Organizations** and NewsAPI **Articles**.

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. This initial exploration indicated these features being influential in the election outcome.After few iterations of data exploration, features have been further derived from the existing data points. Descriptive statistics for these features, and for all states and candidates are shown in Table 1

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

A pair-plot based on these derived feature data-set indicated a clear correlation between **number of Votes** and **number of Events, number of Organizations and number of News Articles** as shown in Fig 1.

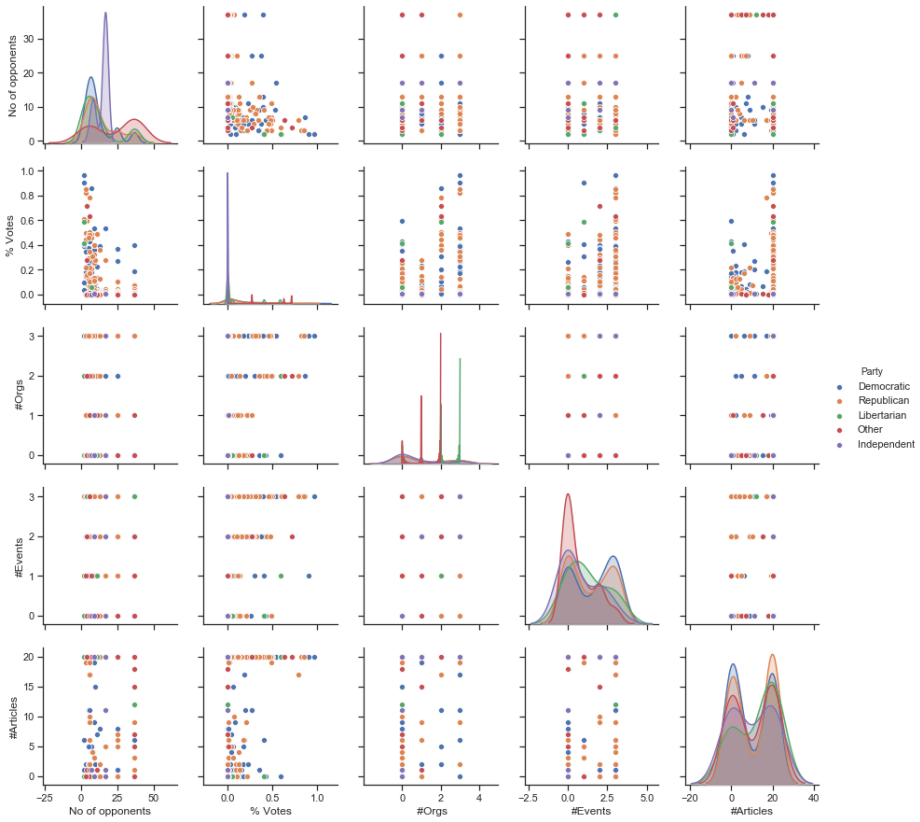

Fig. 1. Pair plot of engineered features

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. We see two different patterns here: One that would not change the votes based on increase in his/her number of news articles, another where we see a linear relationship between the number of articles and votes.

As observed in fig 1, candidates with highest votes (Kamala Harris and Loretta Sanchez) have a strong positive correlation to their associated number of organizations and events. However, this data is skewing rest of the information due to a great majority in the percentage of votes section. This pattern in data is understandable considering the demographic advantage of the top two candidates (Kamala Harris and Loretta Sanchez) being influential in determining their success, in relation to the state (CA) they competed in.

Percentage of votes vs Orgs As observed in fig 2, after excluding the top two candidates (based on votes), 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.

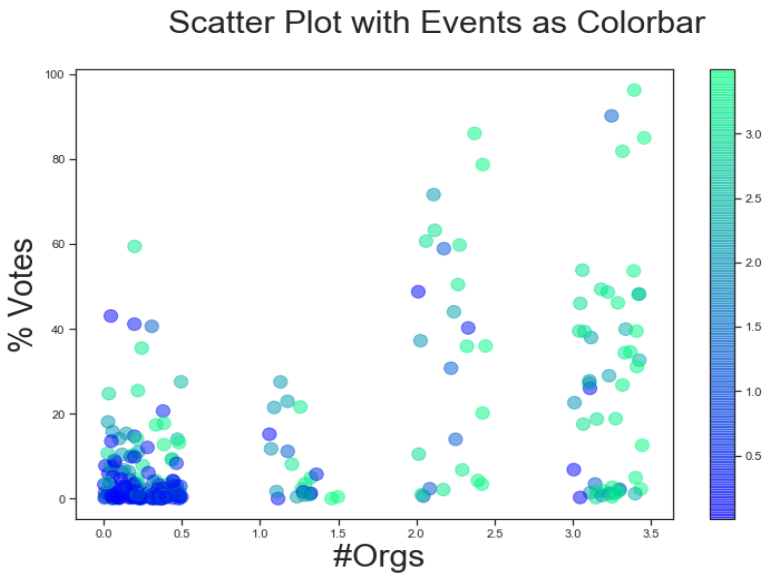

Fig. 2. *Pair plot of engineered features*

In order to understand the cluster of candidates that have non-zero votes with zero association with organizations, we included a color scheme based on number of Event and Articles that they are associated with as depicted in figures 3 and 4. It is evident from these plots that, all candidates that have 2 percent or more votes either have non-zero events or non-zero articles, suggesting that those features are influential in those cases. The only case where this was not true is where we see percentage of votes is 7.77 and the candidate is Duf Sundheim. It is also observed that this candidate has a political background.

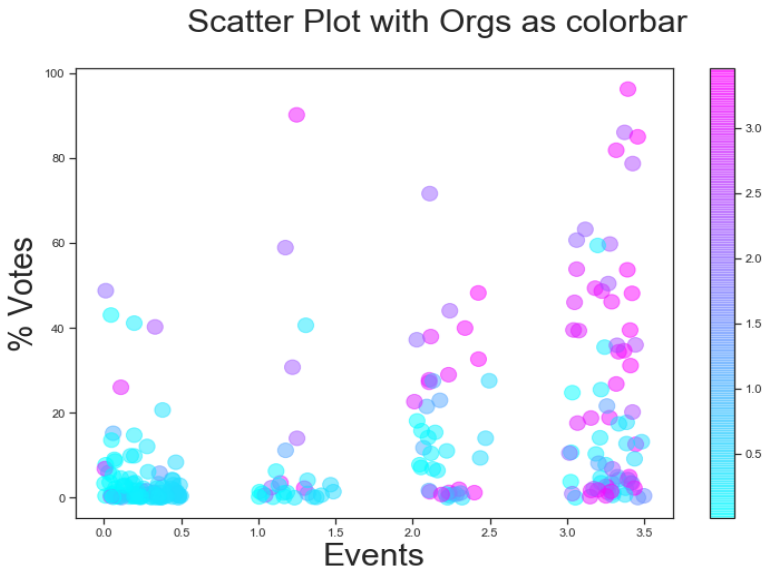

Fig. 3. Scatter plot of No: of Orgs vs. Percentage of Votes - Events

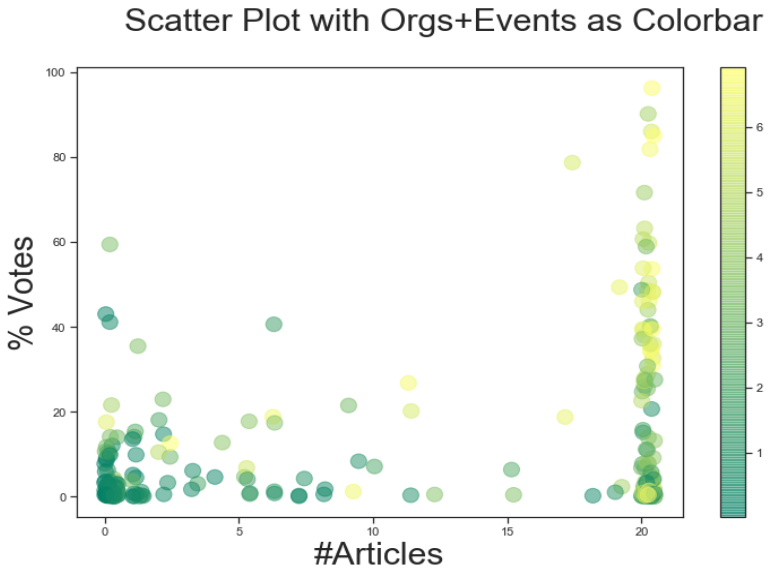

Fig. 4. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

Percentage of votes vs Events A clear positive correlation can be observed between the number of events and percentage of votes from fig 5. The one candidate with 8 percent of votes and zero events seems to be an outlier (Duf Sundheim).

Percentage of votes vs Articles As can be seen in fig 6, the one outlier with 7 percent votes is Duf Sundheim, whom we discussed in section 5.1. Otherwise, there is a general positive correlation between number of articles and percentage 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. **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.**

Based on the above analysis and data exploration, it has been determined that a model can be formulated using No: of Events, No: of Organizations and No: of Articles are input features and the percentage of Votes as the output variable.

4 Data modeling and Evaluation

The data extraction from generic data sources such as Google, Wikidata and NewsAPI has plenty of unstructured data. Quantitative features have been obtained from the unstructured data through feature engineering.

Based on the feature engineering that we did, we determined the most influential features are number of notable organizations a candidate is associated with, number of prominent events a candidate is associated with, number of news articles a candidate is mentioned about, if the candidate has a Wikipedia page, and if the candidate has a presence on Google Knowledge Graph.

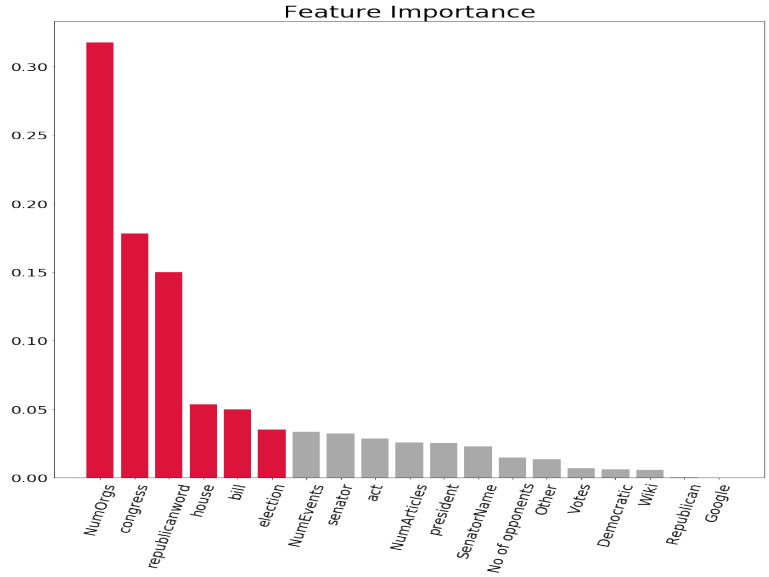

Fig. 5. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

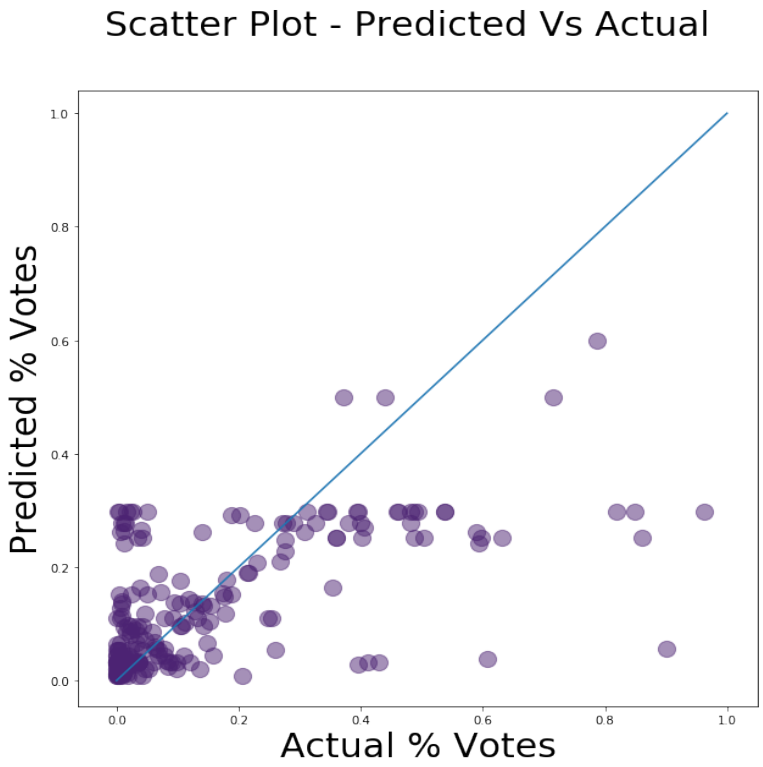

Fig. 6. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

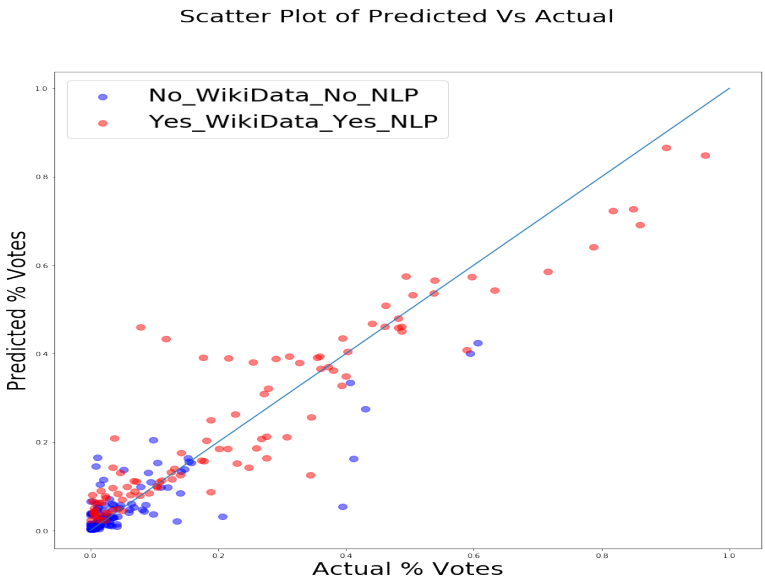

Fig. 7. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

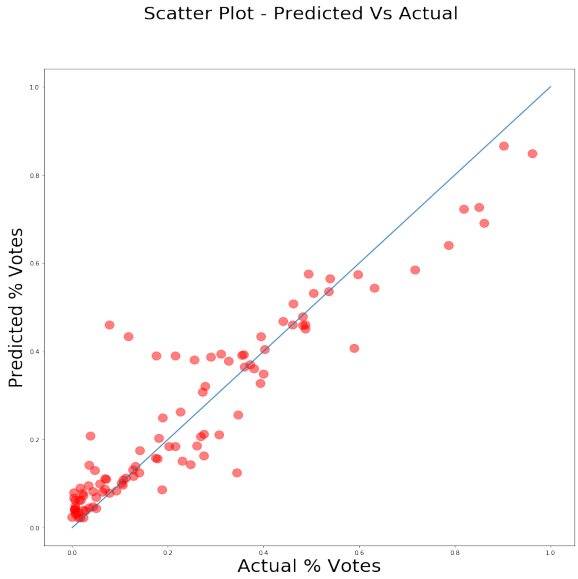

Fig. 8. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

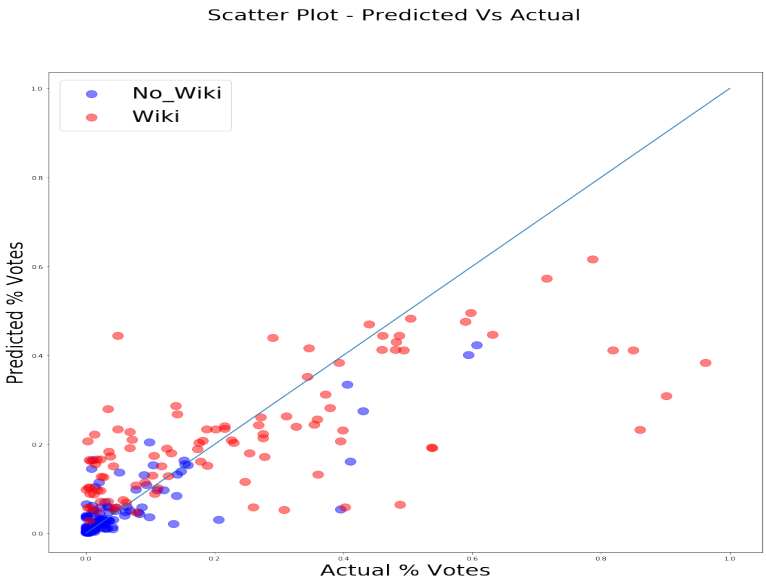

Fig. 9. Scatter plot of No: of Orgs vs Percentage of Votes - Articles

Using this feature set, we used the Random Forest Regressor with cross validation to come up with a model to predict the candidate votes. The model

yielded an accuracy of 93 percent with MSE = .00412 and Pearson coefficient = 0.3(out of bag R square = 0.53). The data has been divided to have an 80/20 split between training and testing data.

	Actual %Votes	Predicted %Votes
0	0.846729	0.742934
1	1.228981	19.172042
2	4.307776	6.743827
3	0.257151	0.739715
4	0.000426	0.742934
5	0.418646	0.742934
6	2.247042	1.292832
7	0.391597	0.739715

Fig. 10. *Random Forest Regression: Actual vs. Predicted*

Fig 7. shows the predicted percentage of votes against the actual given by our Random forest Regressor. It can be noted that both values are close except for row 1(Candidate: Ron Unz), where the predicted votes exceeded the actual by 17 percent due to the high number of feature values we engineered. Upon further analysis we noted that he decided to run for 2016 senate elections in the last minute which might not have had a good reach to the CA population of voters [15].

5 Results

Conclusions are here.

6 Analysis

Conclusions are here.

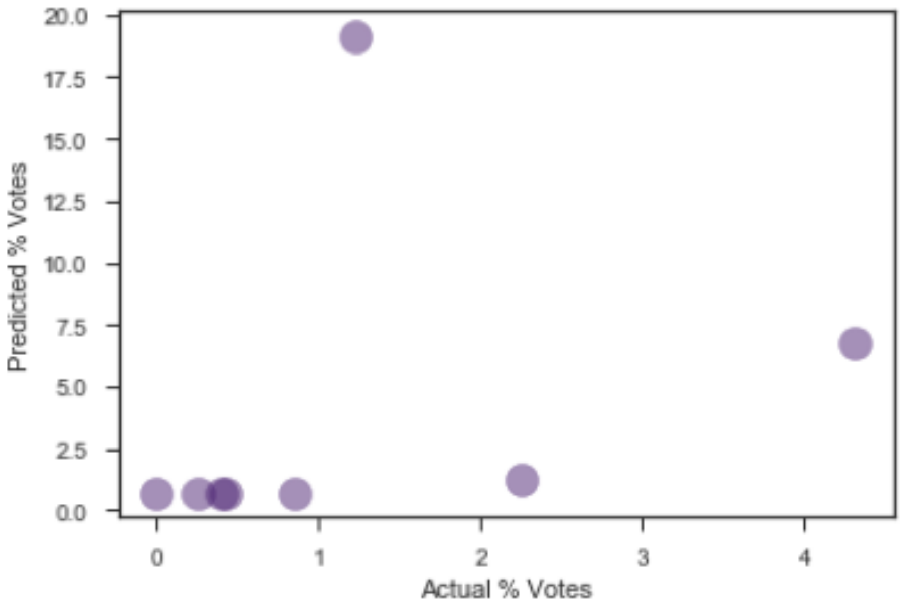

Fig. 11. *Random Forest Regression: Actual vs. Predicted scatter plot*

7 Ethics

Conclusions are here.

8 Conclusions

Conclusions are here.

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.
[10]H.Paulheim. Knowledge Graph Refinement: A Surcery of Approaches and Evaluation Methods, Semantic Web Journal,(Preprint):1-20, 2016.