Application of Deep Neural Networks to Political Survey Data from the 2016 U.S. Presidential Election

Renzhi Cao, Brady Gartman, and Shukrit Guha
Department of Economics, New York University
Machine Learning in Economics, Spring 2018

May 8, 2018

Abstract

We use a multiple-layer perceptron (deep neural network) to predict which voters will vote against their affiliated political party. Our dataset is an exit poll of approximately 8,000 voters in the 2016 U.S. presidential election which includes a broad set of opinion and demographic responses. We find that it is relatively easy to predict who a person will vote for, but it is much harder to predict whether a voter will vote for or against their affiliated party even after accounting for their policy and ideological positions.

Keywords: machine learning, election outcomes, political behavior, deep neural networks

1 Introduction

The 2016 presidential election was one of the most contentious and unpredictable ones in recent decades. Donald Trump is the first president since 1876 to win an electoral college majority while losing the popular vote, and the details of this election process remain deeply controversial. The two major political parties in the United States, Democrats and Republicans, have different political ideologies and policies on various issues. By convention, we anticipate that people who identify themselves as members of one party typically vote along their partisan lines, i.e., once a candidate is chosen to represent their party they will prefer that candidate to candidates in other parties or who are unaffiliated. In reality, however, an individual's registered party may not perfectly indicate their perception of a candidate and therefore that individual may choose to vote for another candidate. For this reason, predicting which candidate will receive an individual's vote is a nontrivial problem, even if we know much of that individual's political affiliation.

The task we face is to use responses of voters from an electorate research survey which asked respondents to give their opinions on various topics to predict their voting actions in the 2016 presidential election. Our hypothesis is that this data will help predict whether voters will vote with or against their party and for whom they will actually vote based on their self-reported responses to questions concerning social, political, and economic issues, particular policies, America's place in the world, and their own demographics.

Previous research attempting to predict the outcome of presidential elections using machine learning (ML) tools is scarce, but a few existing works addressed similar problems. The application of statistical learning tools to politics and social science in general is amongst the most emerging topics in ML and successful application is still an open problem. Khashman and Khashman (2016) use support vector machines and a back propagation learning algorithm on a neural network model to anticipate the American congressional voting outcome. Khaze, Masdari, and Hojjatkhah (2013) use two-layer feed-forward network with a tan-sigmoid transmission function in input and output layers and anticipate participation rate in the coming presidential election in the Kohgiloye and Boyerahmad Providences of Islamic Republics of Iran with 91% accuracy. Gordon and Waxman presented a neural network model for forecasting the outcome of the 1992 U.S. presidential election which correctly predicted the Clintons victory at the 1993 World Congress on Neural Networks.

In the context of this research, a novel dataset, and the burgeoning field of ML, this paper seeks to predict complex voter behavior using deep neural networks. The paper is organized as follows: in section 2 we briefly describe the electorate research (VOTER) survey dataset. Section 3 we discuss our methods for selecting the input data and response variables and the design of the multiple layer perceptron model. In Section 4 we present the experimental results our study and compare the performance of two different models. Finally, we conclude by discussing the economic implications of our findings in Section 5.

2 Data

In this project, we use a sample of U.S. voters responses to a national panel survey conducted by the Democracy Fund on 8,000 respondents. The survey consists of a variety of questions about political beliefs, economic and social concerns, and status, including self-reported vote for president in the 2012 and 2016 elections. The size of the sample allows us to dig more deeply into smaller groups of American electorates than most other surveys [see Democracy Fund (2017)].

3 Methodology

Our general methodology was to compare model performance under different specifications in order to explore what types of survey questions could best predict respondent's voting choices. In order to make the problem of model specification with 167 variables manageable, we partitioned the response variables (response here refers to the answers respondents chose as part of the survey) in our our dataset into 7 categories based on subjective evaluations of the nature of the question being asked. Each model specification considered used a subset of these 7 response categories. Not all data was included in one of these categories - data which was collected after the election and data which was deemed redundant or was omitted from analysis. The created categories are shown in Table 1.

Table 1: Constructed Classes of Variables

Variable Category	Number of Variables
Social Issues	47
Political Issues	10
Economic Issues	12
Issues of General Importance	23
Policy Debates	14
Demographics	24
Issues specific to American Citizens	18

Categories were constructed as logically as possible. Social, Political, and Economic Issues were questions based on those respective issues, whereas Policy Debates were questions based on particular policy issues. Issues of general importance were not directly related to an issue or policy but rather about an idea or general sentiment about the world. Issues Specific to American Citizens were questions about American exceptionalism and patriotism, and Demographic questions are self-explanatory. These variables were included in the analysis because they were observable before the election, either through a similar survey or through social network data mining or another mechanism, and because they could have a credible connection to one's voting choice.

We then constructed two response candidates to predict. The first of these was a 3-level unordered categorical response indicating whether a respondent voted for or against the party they respectively identified with or whether they voted for neither party. This response was the more interesting one, as predicting what causes which voters to break from their party is a complex and relevant problem. The second response variable we created indicated which candidate the respondent actually voted for in the 2016 election. This response had 7 unordered levels, 5 of which were candidate names and the other two were options for those who were either Not sure who they voted for or chose to vote for another, unlisted candidate.

The raw data required extensive preparation beyond that discussed above. Since most of the data in the survey was in the form of qualitative responses, there were naturally several categorical variables with multiple categories each. We anticipated that using this data directly by converting them into dummy variables would become cumbersome as the dimensions of our input matrix would increase drastically. In order to mitigate this problem as much as possible we first converted qualitative responses that had reasonable ordering into ordered factors. For example, for a question with responses “Agree”, “Don't Know”, and “Disagree”, we replaced these character responses with the integers -1,0 and 1 to reflect an increasing order of agreement. It is important to note that we did not apply this method to variables for which there was no clear distinction between categories, or ones for which some responses were ordered but at least one response had no reasonable place in that ordering. By converting these variables into numeric vectors with ordered inputs to represent the ordered categories of qualitative responses in the original dataset, we were able to reduce the number of dimensions of the final input dataset significantly.

In order to increase the number of observations to be used in the dataset, we also had to deal with a considerable number of NA values that were present in the original dataset. Our initial attempt at trying to reduce the number of NAs was to simply replace the NAs with “Dont know” as the categorical response value realized when respondents chose not to answer specific questions. This substitution was not arbitrary in that we only substituted them in variables where “Don't know” was already one of the categories. Theoretically, there is little difference between a person that admits to not knowing something or holding a particular opinion and a person who declines to respond. This method of imputation is subject to scrutiny in that we are essentially imposing our own structure to the dataset that was originally not there. However, given that the respondents chose not to answer certain questions even though they answered more than 100 other specific questions, we deduced that “Don't know” would be a reasonable approximation to their true answers to those unanswered questions. We also imputed values for 29 other variables using imputation by singular value decomposition [see Schmitt, Mandel and Guedj (2015)]. 23 of these variables had numerical inputs ranging from 0-100 in the original dataset while the other 6 had ordered responses. The optimal k values chosen to impute these missing data were calculated by cross-validating the two subsets of data after removing NAs.

After the above preprocessing procedure was complete less the NA removal, the final input dataset contained 7611 observations with 167 variables. After omitting observations which had NAs that could not be imputed due to lack of data for that observation, the number of observations was reduced to 7310, a loss of 301 observational units.

The machine learning model chosen for this analysis was a a 10-by-10 multiple layer perceptron deep neural network (DNN). The network was implemented using the R package RSNNS [(see Bergmeir and Benitez (2012)] with 1000 training iterations and a decay parameter equal to 0.01. The model was specified on both the 3-level and the 7-level response variables using each individual subset of the data in Table 1. We trained the model on a random sample of 75% of the total number of observations (roughly 5,483 observations) and tested the model performance on the rest (i.e.1,827 observations). Following the initial results of this technique, we respecified the model to include all of the subsets (i.e. all of the categories in Table 1 in a single model). The reasoning for this is that a model with a maximal number of predictors should perform no worse than a model with only a subset of those predictors, therefore it is worth determining the upper limit on the prediction performance one can attain with a given dataset and modeling technique. The results of this expanded specification are discussed in the next section.

4 Results

We found the predictions for the 3-level response (whether respondents voted for or against their respective parties) were quite poor for any of the individual variable subsets. Given this poor performance, we considered the full model as discussed above.

Figure 1: ROC Curve for 3-level Response Model

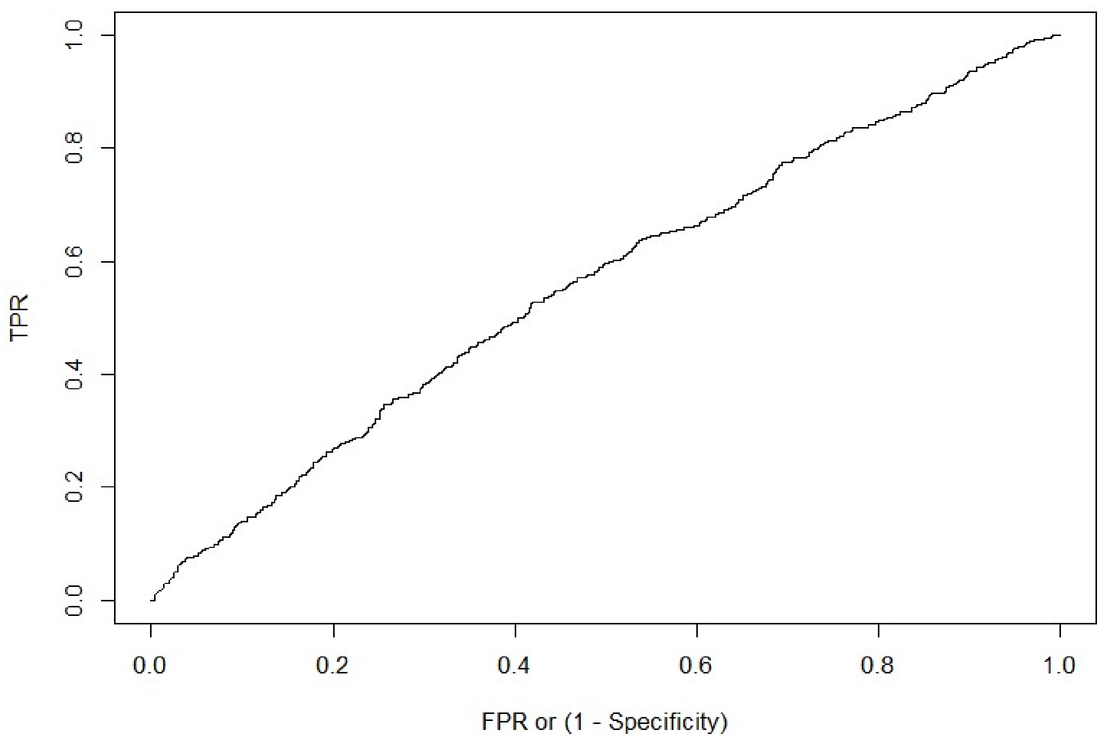

Figure 2: ROC Curve for 7-level Response Model

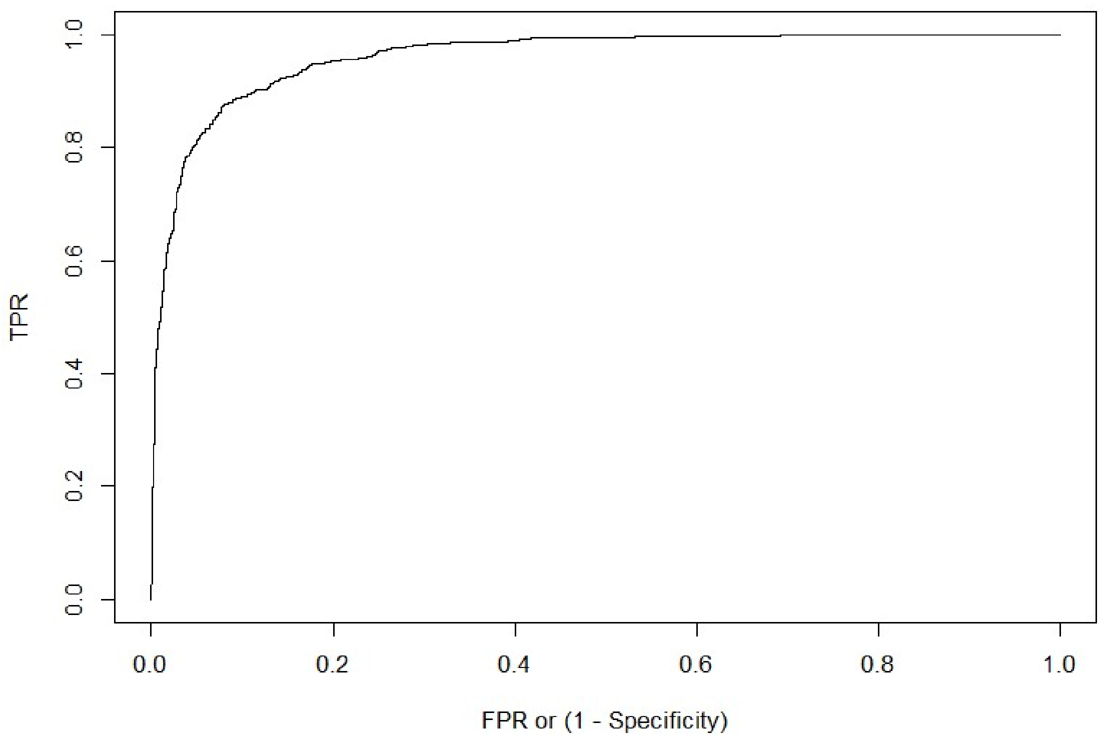

Table 2: Confusion Matrix for 3-level Response

		Predicted		
		Neither	Vote Against	Vote For
	Neither	23	49	61
Actual	Vote Against	80	153	272
	Vote For	139	270	781

Table 3: Confusion Matrix for 7-level Response

		Predicted				
		Trump	Johnson	Clinton	Stein	Other
	Didn't Vote	2	1	5	0	1
	Trump	831	14	43	6	11
	McMullin	1	0	0	0	1
Actual	Johnson	22	13	14	0	2
	Clinton	54	10	712	1	8
	Stein	4	2	20	0	0
	Other	25	1	13	0	0

Figure 1 and Table 2 illustrate the poor prediction ability of the full DNN model for the 3-level response. A model that predicts well on the test set should have a receiver operating characteristic (ROC) curve with minimal area between the curve and the left and top boundaries. Poorer models are closer to the diagonal line, which is equivalent to unconditional random class assignment. Since the ROC curve is so close to this diagonal line, this model's prediction accuracy is not significantly different from random guessing. The confusion matrix presented in Table 2 conveys this same point in a different way - values in the off-diagonal are misclassifications, and their sum divided by the sum of all entries in the matrix is the misclassification rate, which was 47.6%. In contrast, using the same full set of inputs to predict who a respondent voted for resulted in far more accurate predictions, as shown in this model's ROC curve in Figure 2 and confusion matrix in Table 3. Note that Table 3 omits columns for which no predictions were made - the DNN did not predict anyone in the test set would vote for Evan McMullin or abstain from voting. The misclassification rate of this model was 14.9%.

Despite running a relatively complex neural network (i.e. with 10 hidden layers, each with 10 neurons), our prediction for whom a respondent voted was still less satisfactory than one might expect, given the well-documented ability of DNNs to capture complex nonlinear relationships reliably. However, in comparison to the performance of a similar neural network on the 3-level response, the 7-level response model performs quite well. We were unable to cross-validate the neural network to obtain the optimal learning rate and number of neurons and layers due to the size of the dataset and unbalanced classes, thus we run the risk of having overfitted the model. This is a clear limitation of our study as we were unable to conclusively determine whether the model performs poorly due to poor model fit or as a result of unknown underlying factors influencing the decisions of voters.

5 Discussion and Conclusion

Given the well-established complexity of collective human behavior and the eccentricities of the 2016 U.S. presidential election, it is not surprising that predicting voting outcomes is difficult. That said, it is counterintuitive that, given the comprehensive nature of the survey data we utilize, we were not able to observe a systematic meaningful difference between someone that votes with their party, someone who votes against their party, and someone who chooses not to vote or to vote for a third party. It is possible that this shortcoming is a result of our modeling methodology, and future work should consider implementing more robust models. However, the ability of DNNs to capture complex relationships in general is well-supported empirically, suggesting that data limitations or problem intractability may be better explanations for poor prediction performance.

The dataset is known to be partially artificially constructed, with corrections made to make the sample demographically representative. The nature of this construction is beyond the the scope of this paper, but if there are problems with this procedure they would only be exacerbated by our attempts to glean meaning from the data. However, we would expect the most likely types of construction errors to reinforce biases in the dataset, which would give us false confidence in the ability of our model to accurately predict reality.

For example, if the actual sample failed to get representative data for a particular class of voter, what data that was collected for that class would be extrapolated (increased in relative weight) for that class. This would reinforce whatever signal or noise was actually present in those observations, leading to increased model variance. Of course, as is always the case in modern ML and data science in general, more data would help improve our confidence in the the feasibility of this research.

Ultimately, it is possible that this problem is simply intractable. Perhaps voters are truly irrational and any model which assumes a logical consistency between beliefs/opinions and voting behavior are obsolete, at least in the present political climate. Perhaps voters are in fact rational but their underlying motivations for choosing one candidate over others is something not captured (even indirectly) by the types of questions asked in this survey. In this case, the problem is referred to the fields of social and political psychology and behavioral economics, and ML and traditional economics has little to say.

Bibliography

Bergmeir, C. and Benitez, J.M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software 46(7) 1-26. Retrieved from http://www.jstatsoft.org/v46/i07/

Borisyuk R, Borisyuk G, Rallings C, Thrasher M. (2005). Forecasting the 2005 General Election: A Neural Network Approach. British Journal of Politics and International Relations 7(2) 199-209.

Democracy Fund Voter Study Group. View of the Electorate Research Survey, December 2016. Release 1: August 28, 2017. Washington DC: Democracy Fund Voter Study Group. Retrieved from https://www.voterstudygroup.org/

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.(2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. New York, NY: Springer.

Jill GS.(2008). Election Result Forecasting Using Two Layer Perceptron Network. Journal of Theoretical and Applied Information Technology 4(11) 1019-1024.

Khashman, Zeliha, Adnan Khashman.(2016). Anticipation of Political Party Voting Using Artificial Intelligence. Procedia Computer Science 102 611-616. Retrieved from https://doi.org/10.1016/j.procs.2016.09.450

Khaze, Seyyed Reza, Mohammad Masdari and Sohrab Hojjatkhah.(2013). Application of Artificial Neural Networks in Estimating Participation in Elections. International Journal of Information Technology, Modeling and computing(IJITMC) 1(3) 23-31.

Luna, Guillermo De Ita, Aurelio Lpez Lpez and Josu Prez Lucero.(2014). Predicting Preferences of Voters from Opinion Polls by Machine Learning and Game Theory. Research in Computing Science 77 121-131.

Schmitt, P., Mandel, J., and Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics and Biostatistics 6.1. Retrieved from https://www.omicsonline.org/open-access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.pdf

Supplementary Material

Original survey data and included documentation As downloaded from `https://www.voterstudygroup.org/publications/2016-elections/data` (.csv, .pdf)

Altered dataset All preprocessing done outside of R - run the R files on this rather than the raw data. (.csv)

R routine For processing the data and fitting the models (.R)