**Predictive model evaluation for Donald Trump votes in the 2016 Election**

Andrew Graham

Ritchie School of Engineering and Computer Science, University of Denver

COMP 4448: Data Science Tools 2

Prof. Neba Nfonsang

November 14, 2022

**Predictive model evaluation for Donald Trump votes in the 2016 Election**

**Purpose**

The purpose of this project is to predict, based on demographics from surveys, whether someone voted for Donald Trump or not in 2016. The data will be used to see whether KNN, Naïve Bayes, or a Random Forrest will perform better in predicting whether a given individual voted for Trump.

**Significance**

The significance of this analysis is to gain an understanding of how certain demographics and views align with voting patterns. This will challenge and/or confirm assumptions about the 2016 election. Prediction models can be used in future elections to estimate the likelihood of a candidate getting votes in the future. This can also be used to predict what types of candidate voters would prefer in future elections.

**Research Question**

How does the performance of a Naïve Bayes Classifier compare with KNN Classifier, Random Forest Classifier, and Logistic Regression in predicting whether an individual voted for Trump in 2016 given survey answers on demographics, region, ideology, and views on racism?

**Data Description**

The dataset is the result of a 2016 survey conducted on individuals on whether they voted for Trump. The data includes demographics and survey question results. The original data set held 66,600 datapoints. However, this was condensed to 44,898 to only focus on data where the voter answered yes or no on voting for Trump. The datapoints where there was no answer were removed. The data has 1 continuous feature and 17 nominal. Of the 17 nominal features; 2 are purely categorical, 4 are binary choices, and 11 are ordinal.

***Output***

The output feature of this study is whether someone voted for Trump. This feature has two responses (1,0 representing ‘yes’ and ‘no’). In this set 40% are yes and 60% are no making this balanced set.

***Input***

The input features can be divided into 4 themes: demographics, ideology, religion, and racism. There are 6 demographic inputs including: age (continuous), sex, college degree, race, and household income. There are 2 ideology responses: one indicating how liberal or conservative, and the other how much they identify with either the republican or democrat parties. An additions 4 questions on religion asking about how much they attend church, prayer frequency, importance of religion, and born-again status. Finally, 4 questions regarding views on racism including: prevalence of racism, anger that racism exists, white advantages, and fear of other races. The data has numerical values for the answers to these questions based on scales as seen in *Appendix A.*

**Data Preprocessing**

**Data Preparation**

The data was first analyzed for any data inconsistencies such as misspellings, alternate spellings and other data issues. None were found. There were 7,337 rows having missing data, constituting approximately 16% of the data. A majority of these were from the household income question. The missing data was randomly distributed between both output classes and no internal correlations were found. Due to this and the fact that they were not predictive, these rows were removed so that 37,561 datapoints were left to be analyzed.

**Exploratory Data Analysis**

Correlation analysis showed high correlation between features in each feature theme of ideology, religion. and racism. Ideology had the strongest correlation to the output, followed by views on racism, then religion. Demographic data, with the exceptions of age and race had insignificant correlations.

***Demographics***

Age looked to be normal overall. The distribution for those voting for Trump had a mean around 55 while those who didn’t had a mean around 50. The lower age groups, those under 40, had a much higher percentage not voting for Trump, while over the age of 50 the difference was fairly even. Sex had a minor correlation of males more likely to vote for Trump and having a college degree had a minor correlation to not vote for Trump. Household income didn’t have a significant difference. There was a significant factor in race where if the respondent was not white, they overwhelmingly did not vote for Trump. The state of residence showed a few areas to vote against (New England and Hawaii for example), but most of the states were around 40-50% voting for, so not that significant in the positive.

***Ideology***

Donald Trump was the Republican candidate, so as expected the more republican and the more conservative the more likely the respondent voted for him. With the far ends overwhelmingly voting for and against.

***Religion***

Whether someone considered themselves a born-again Christian and how often they went to church had a minor correlation to voting for Trump. The importance of Religion and Prayer Frequency supplies the most correlation. Overall the more religious the respondent the more likely the voted for Trump

***Racism***

Opinions on one’s anger towards racism and how much they feared other races had minor correlations. This could have been tempered by respondent’s truthfulness and definitions. How rare one thought racism was had significant correlation were those who though it was rare were more likely to vote for Trump. The most significant factor, however, was the respondent’s view on whether white people have an advantage over others. Those who disagreed with this statement were overwhelmingly Trump voters.

**Feature Engineering**

The age category was binned into four groups: 18-35,36-50, 51-65, 65+. Since there was not a linear relationship between the output and this feature.

**Clustering**

Clustering was analyzed to see if it would improve predictive ability. K-means clustering with a k of 6 was used for this purpose. Two datasets were created, one with the clusters and one without.

**Data Splitting**

Each dataset was split into training and test sets using the same random number input to keep the splits the same for the clustered and non-clustered data. There was no scaling applied as all features were nominal.

**Model Building and Evaluation**

**Model Building (include code snippets**)

4 models were chosen to provide a baseline performance evaluation: Logistic Regression, Random Forest, K-Neighbors, and Naïve Bayes. All models used the default parameters. They were all evaluated using cross validation with 5 splits. Accuracy and F1 score were used to evaluate the initial run.

Text

Description automatically generated

The results showed no over fitting from the models. The results from the clustered and non-clustered data were essentially the same, so the non-clustered approach was chosen. Finally, logistic Regression and Random Forest both performed at .89 accuracy and .87 F1 scores, while KNN had .87 accuracy and Naïve Bayes at .81. From this the Random Forest and Logistic Regression were chosen to be optimized

**Model Optimization (Hyperparameter Tuning) And Model Selection**

Both Random Forest and Logistic Regression were tuned using a Gridcv search. Random Forest tuning parameters included the max depth, number of estimators, the max features and the minimum sample split. The Logistic Regression model tuned the vales of C.

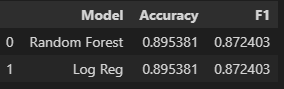
Text

Description automatically generated

Text

Description automatically generated

**Model Comparison (summary table)**

****

The optimizations did not improve the accuracy or the f1 score significantly. Also, the Random Forest and Logistic Regression performed equally, showing they both arrived at the same classifications. Since there seemed to be trivial improvement from optimization, it may be close to the accuracy limit for this dataset. The resulting performance is still high and seems to perform well in predicting votes for and against Trump.

**Conclusion**

From these results, the simplest predictive model would be that of Logistic Regression. The optimal C parameter being 0.1. Also, clustering is not needed as it does not improve performance. Random Forest also performs well but is more time consuming, so it is not the best choice. Finally KNN and Naïve Bayes underperformed.

**Lessons Learned**

Ideology, views on white advantages, and the race of the respondent were the most significant predictors of a vote for Trump. Income was not as significant as initially thought. There seems to be an accuracy ceiling of around 90% for this dataset. Simpler methods were found to be the best for this exercise

**Recommendations**

A Logistic Regression with a C parameter of 0.1 is the best predictive model for predicting whether a respondent voted for Trump with an accuracy of 89.5%. This model produces maximum accuracy and the fastest evaluation time. Added data from the source survey would be useful if a higher accuracy is desired.

**References**

(2012). Retrieved November 18, 2022, from Github.io website: https://vincentarelbundock.github.io/Rdatasets/csv/stevedata/TV16.csv

‌

(2012) The Individual Correlates of the Trump Vote in 2016 https://vincentarelbundock.github.io/Rdatasets/doc/stevedata/TV16.html (accessed 2022 -11 -18).

‌

**Appendix A**

**Data Description**

state

a character vector for the state in which the respondent lives

votetrump

a numeric that equals 1 if the respondent voted says s/he voted for Trump in 2016.

age

a numeric vector for age that is roughly calculated as 2016 - birthyr, as it's coded in the CCES data.

female

a numeric that equals 1 if the respondent is a woman

collegeed

a numeric vector that equals 1 if the respondent says s/he has a college degree

racef

a character vector for the race of the respondent

famincr

a numeric vector for the respondent's household income. Ranges from 1 (Less than $10,000) to 12 ($150,000 or more).

ideo

a numeric vector for the respondent's ideology on a liberal-conservative discrete scale. 1 = very liberal. 5 = very conservative.

pid7na

a numeric vector for the respondent's partisanship on the familiar 1-7 scale. 1 = Strong Democrat. 7 = Strong Republican. Other party supporters (e.g. libertarians) are coded as NA.

bornagain

a numeric vector for whether the respondent self-identifies as a born-again Christian.

religimp

a numeric vector for the importance of religion to the respondent. 1 = not at all important. 4 = very important.

churchatd

a numeric vector for the extent of church attendance for the respondent. 1 = never. 6 = more than once a week.

prayerfreq

a numeric vector for the frequency of prayer for the respondent. 1 = never. 7 = several times a day.

angryracism

a numeric vector for how angry the respondent is that racism exists. 1 = strongly agree (i.e. is angry racism exists). 5 = strongly disagree.

whiteadv

a numeric vector for agreement with statement that white people have advantages over others in the U.S. 1 = strongly agree. 5 = strongly disagree.

fearraces

a numeric vector for agreement with statement that the respondent fears other races. 1 = strongly disagree. 5 = strongly agree.

racerare

a numeric vector for agreement with statement that racism is rare in the U.S. 1 = strongly disagree. 5 = strongly agree.