Data, Democracy, and Discontent: Quantifying Populism

ST309 Group Project

Exam candidate number and contribution:

[redacted] – 1/3 of the project

[redacted] – 1/3 of the project

[redacted] – 1/3 of the project

Introduction

Populism has emerged as a significant discussion point in the context of democracy and economic inequality. In the United States, the rise of Donald Trump and Bernie Sanders shifted the political discourse onto populist, anti-elite and nativist lines. Similarly, in Europe, there has been political dishevelment with the Brexit referendum, the relative success of Marine Le Pen in the French 2017 elections, and the rise of the Alternative for Germany (AfD) in the 2017 German Elections. As such, understanding populism has gained traction in recent literature, and is timely in recognizing current political sentiments.

Cas Mudde (2017) describes populism as a "thin centred ideology", which makes it malleable to different political narratives. For example, populism can be utilised by multiple political ideologies with different viewpoints on social and economic distributions. Irrespective of political inclinations, populism pits "the people" against the "corrupt elite", distrusts state institutions, and rejects credentialled individuals.

Martin Wolf[i] and others argue that such views can primarily be attributed to economic reasons such as disillusionment with the current system of allocation, stagnating middle-class wages, and high levels of unemployment. On the other hand, Galston et al. (Brookings 2018) argue that populism can be traced back to the hyper-globalisation of the modern world resulting in anti-immigration views, xenophobia and insecurity. Such different perspectives on the origins of populism are at odds with each other, and testing which theory holds more credence can lead to different policy implications.

As such, using the 2016 US Elections voting patterns, populism can be cast as a data science problem. By understanding the strongest indicators of voting for a populist (Trump or Sanders), we can gauge whether populism can be rooted in ideological or economic reasons. Furthermore, by looking at right-wing populism (Trump) versus left-wing populism (Sanders), we can understand the difference in how populism can be framed by different ideologies. Such an analysis provides an essential insight into not only what causes populism but also provides a cross-section into the different types of populism.

Literature Review

While much has been said on populism, there is not much literature on quantifying populism using data science. Most notably, Inglehart and Norris (2016) highlight economic insecurity and the cultural backlash theses and argue that populism can predominately be accounted for by the latter. The Economist [ii] and YouGov use a statistical model in order to identify the strongest predictors of the 2018 mid-term elections, finding that religion is the best predictor of a person's vote. Our analysis, using data science and machine learning gives credence to these views – that cultural backlash is the prime driver

behind populism – but does this by quantifying the 2016 election dataset, when populism was at its highest.

Data Description:

I – Background:

The Views of the Electorate Research ("VOTER") Survey of 2016 was conducted by the survey firm YouGov, an international Internet-based market research and data analytics firm. The survey was the inaugural research product of the Democracy Fund Voter Study Group, a research collaboration funded by the bipartisan foundation, The Democracy Fund, whose aim is to improve the democratic process in the United States. The VOTER Survey of 2016 was conducted on-line between the 29th of November and 29th of December 2016, with 8000 adults (age 18+) participating. This 2016 survey was a built up upon a previous survey conducted by YouGov in 2011-12, in which a subset of the participants of the 2011-12 survey were approached again by YouGov to partake in the survey following the 2016 Presidential Elections. This is due to the essence of capturing the evolving views of the electorate and the implications of these changes that follow thus gives credence for a unique longitudinal analysis. All in all, the survey contains a total of 668 variables in the form of questions addressed to the 8000 participants, whose answers are given by a combination of categorical and hierarchical responses.

II – Explaining the Data Set:

With our underlying thesis in mind, the data set provides a holistic network of information – from the participant's cultural background to their economic wellbeing. All of these variables would indeed be instrumental in explaining the primary indicators for an individual to be voting for a candidate deemed a populist. The data set contains a vast amount of information with regards to the characteristics of the participants and their views and opinions that follow. It is worth noting that:

1. Variables whose names end with "_2016" capture responses of the 2016 survey – for example; the variable "healthcov2_2016" captures the type of health care coverage the individual possessed as of 2016.
2. Variables whose name end with "_baseline" or "_2012" capture responses of the 2011-12 wave of surveys the same participants of the 2016 undertook, and is used by the data set as a benchmark with regards to the initial views of the participants prior to engaging with the 2016 survey.

The data set begins by identifying the diversity of the participants, from their age and race to their employment status and income. These data points are categorical in nature and allows us to properly navigate through the multitude of participants of different socioeconomic backgrounds. For example:

1. As illustrated in Figure 1 below, the variable "race_basline" shows us the proportion of participants by race:

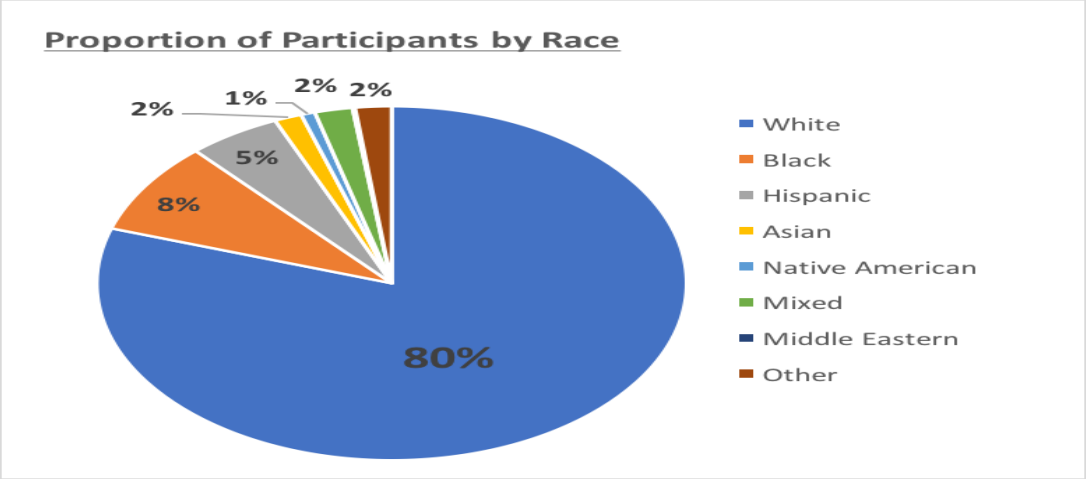

Figure 1 – Proportion of Participants by Race

2. The data set also includes the family income of the participants as of 2016. Signified by the variable "faminc_2016" in the data set - Figure 2 below shows the division of participants by income:

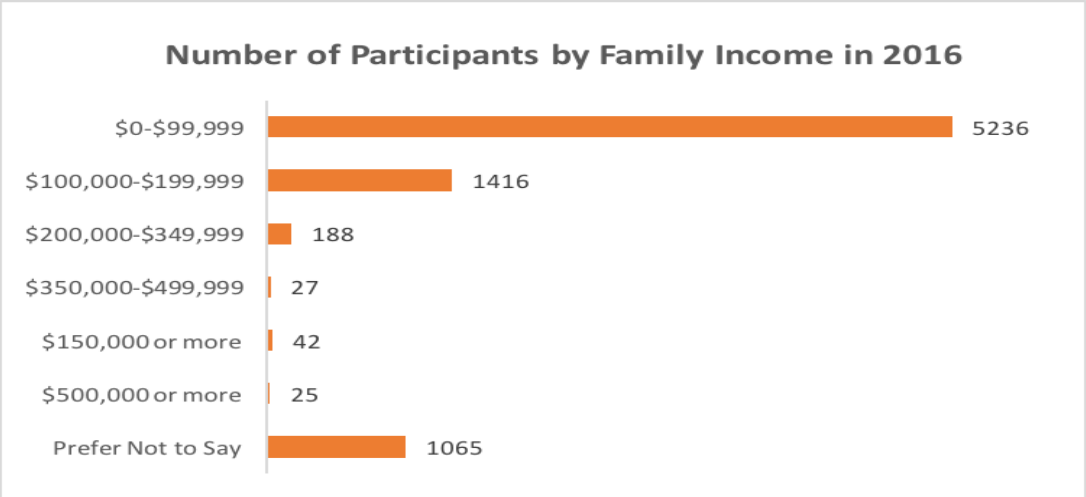

Figure 2 – Number of Participants by Family Income in 2016

With regards to the principal aim of the survey in capturing the views of the electorate, the overall structure of the data is centred upon the participant's responses to a series of qualitative and opinion-engaging questions ranging from the participant's views of healthcare to their individual sentiment towards the death penalty and trade policies.

For instance, the variable named "envwarm_2016" aims to gauge the participant's views with regards to whether or not global warming is actually happening. The participants are given choices ranging from "Definitely is happening" to "Definitely is not happening" with the option to respond with "Don't know". It is clear to us that the response of the participant's opinion to this topic (as with all the other topics) are recorded in a hierarchical manner. In addition, there are also variables that begin with "ft_"

– these variables illustrate the participant's "feeling thermometer" towards certain sections of society (other races or specific professions, i.e. "ft_muslim_2016"). This variable required the participants to give a numerical score, ranging from 1-100, with regards to their own personal sentiment towards the section of society in question.

All in all, while some of the data points in this data set are categorical, the bulk of the observations contained in this data set are hierarchical responses by the participants that reflected each individual's opinion regarding a specific issue in question. We acknowledged that some transformations of the data set are needed in order to conduct the analysis for our premise on R, and the process undertaken is highlighted in the section below. In conclusion, the VOTER 2016 Survey does achieve its goal in gauging the evolving views of the United States electorate, and thus, allows us to conduct our study in explaining the drivers towards an individual voting for a populist candidate. The meaning behind each variable in the data set can be found in the guide for the VOTER 2016 Survey data – this can be found together with the original data in the uploaded folder with this submission or on the original website.

Data Cleaning

I – Removing Variables on Excel

Upon our preliminary analysis, we encountered certain underlying issues with the variables that would hinder our progress when running the functions on R. The first variable on the data set, named "inputstate_2016", discloses the participant's state of origin. Seeing as the United States has 50 states in total, this variable alone would not allow a simple tree function to be ran on R, due to the function not being able to handle more than 32 factor levels. With this, a transformation of the responses on Excel of the participants to "regional codes" (ranging from 1-9) had to be done, in correspondence to the participant's origin. These regional codes were outlined and obtained from the US Census Bureau and thus were inputted and converted into our data set under a new variable name called "Region_2016". For that specific reasoning, the same thing had to be done to other variables – i.e. "birthyr_baseline" was changed to "agerange".

Furthermore, the data set also contains a set of variables that we felt were crucial to our analysis, though were recorded not only in a hierarchical manner with more than 32 factor levels (thus arising the aforementioned problem with the tree function) but also in an inconsistent manner. The set of variables in question are related to those that start with "ft_". As mentioned previously, this variable records the participant's feelings towards particular sections of society. Upon examining the responses to these variables, we found that, though the majority of the participants answered as required (by simply inputting their own numerical score), there was a significant amount of participants who added on specific comments next to their numerical score recorded (an example of such response being "25 – Unfavorable"), thus making the data points fairly inconsistent. With this, we've converted the responses

of these variables by grouping a range of numerical scores into a specific category. For example, responses ranging between "0 to 45" would be categorized under "Not Favorable" – thus transforming these responses from a potentially 100 different ways of answering to only 3 specific categories ("Neutral" and "Favourable" being the other two). Thus, following these conversions on Excel, it has successfully allowed us to conduct the analysis of our thesis on R, with the potentially significant and explanatory variables included from the data set.

II - Removing Variables on R

The first attempt at data cleaning in R simply involved dealing with missing values. At a first pass attempt, none of our go-to functions (such as *tree)* for analysis were viable due to the huge number of missing values in some variables. Furthermore, if the majority of observations in a variable is just missing values, they are unlikely to be very useful in generating accurate predictions or have any explanatory power. Therefore, we decided to remove all variables with more than 50% missing values. Next, we decided to eliminate data points from 2011 and 2012 because they are outdated in describing the same person: some people may have gotten a job or lost one between 2012 and 2016, graduated from or dropped out of college, or even have a child within that timeframe. This allows us to avoid any concerns, where outdated characteristics and opinions become predictors to preferences in 2016. We eliminated randomizers and open-ended variables for when people answer "Other" in the survey – they serve no purpose in our analysis and nuances have been captured by the levels within their original variables.

We also removed variables which were determinants of voting preferences but don't form on the two purported drivers of populism. The set of variables named "PARTY_AGENDA…" are possible predictors of voting preference – if a person strongly believes in a particular agenda on either major party, this is likely to be captured in their vote during the 2016 presidential election. However, they are slightly unhelpful for a two reasons:

1. It is as if we are trying to predict voting preferences by directly asking people about their voting preferences. What we want to know is *why* they believe in a democratic agenda. In the language of econometrics - they are essentially bad controls because said variables are also an outcome of the variables we would like to use to predict populist preferences such as identity and economic factors.
2. The variables which directly capture party and political preferences say nothing about our overarching inquiry – what truly drives populism: economic factors or identity politics? For example, the variable "fav_trump" (directly measuring whether people had a favourable opinion of Donald Trump) is surely a strong indicator of whether you voted (or will vote) for Trump but tells us nothing about which driver of populism *and* is akin to basically asking about your voting preferences in the presidential election.

In light of this, we decided to remove a further 33 variables which, upon multiple iterations of decision trees and inspection, were more directly capturing voting preferences without telling us much about populism. Finally, we removed any remaining variables containing more than 32 levels – to adjust for the tree function that requires variables to have fewer than 33 unique values. It turns out that only two variables were removed: "race_other_2016" and "izip_2016" (description of race in 'other' and the respondents' respective ZIP code).

III – Manipulation for Analysis on R

Having arrived at 157 relevant and usable variables, we also converted our responding variable "presvote16post_2016" (who you voted for in the 2016 election) to "Vote" (Did you vote for Trump or Not?). The resulting data set had 8000 observations and 1606 missing values. A random sample of 4000 observations was extracted from the data set to be training data, while the other 4000 observations were left as testing data.

Upon preliminary analysis on trees (conducted as described in the methodology), we discovered that by omitting the rows with missing values in any variable, there was a significant increase in the accuracy of our predictive tree model. The useable data set shrinks to 3581 observations and 157 variables. The data set is smaller, but far more powerful than before – the percentage of wrong predictions decreased from 23.2% to 9.4%.

Methodology and Analysis

In this section, we will discuss the analytical methods deployed to form an understanding of voting preferences and later, evaluate on a more technical perspective, our approach and results. From the US VOTER Survey data, our proxy for populist preferences is a vote for Donald Trump in the Presidential elections. From the following analysis, we were able to:

1. Form a predictive model for voting for Trump.
2. Highlight the strongest drivers of populist voting.

I - Classification Decision Tree

Firstly, we built a classification model to group voters based on who they voted for in the US Presidential Elections in 2016 – voting for Donald Trump would be indicative of populist sentiment. We started with fitting the training data into a decision tree. We used a *classification* decision tree instead of a generalized linear model (regression) to form a prediction as most variables in our data were qualitative variables.

This is the motivation behind a classification decision tree: think of all the variables in our data set as a set of questions. For example, the variable "Region_2016" would be the question, "Which region do you live in?". The observations in the training data are partitioned at every *node* where a question is

asked – almost like a series of questions to determine your class. In this case, their class is the "Vote" variable: whether they voted for Donald Trump. The data is repeatedly partitioned according to the variables in the data until each partition that we have contain observations of the same class. By fitting a classification tree, we have a model for voting preferences based on the choices of the observations in our training data. In a nutshell, the model uses the training data to learn any relationships between the variables we have and the one we want to predict – forming a model upon completion.

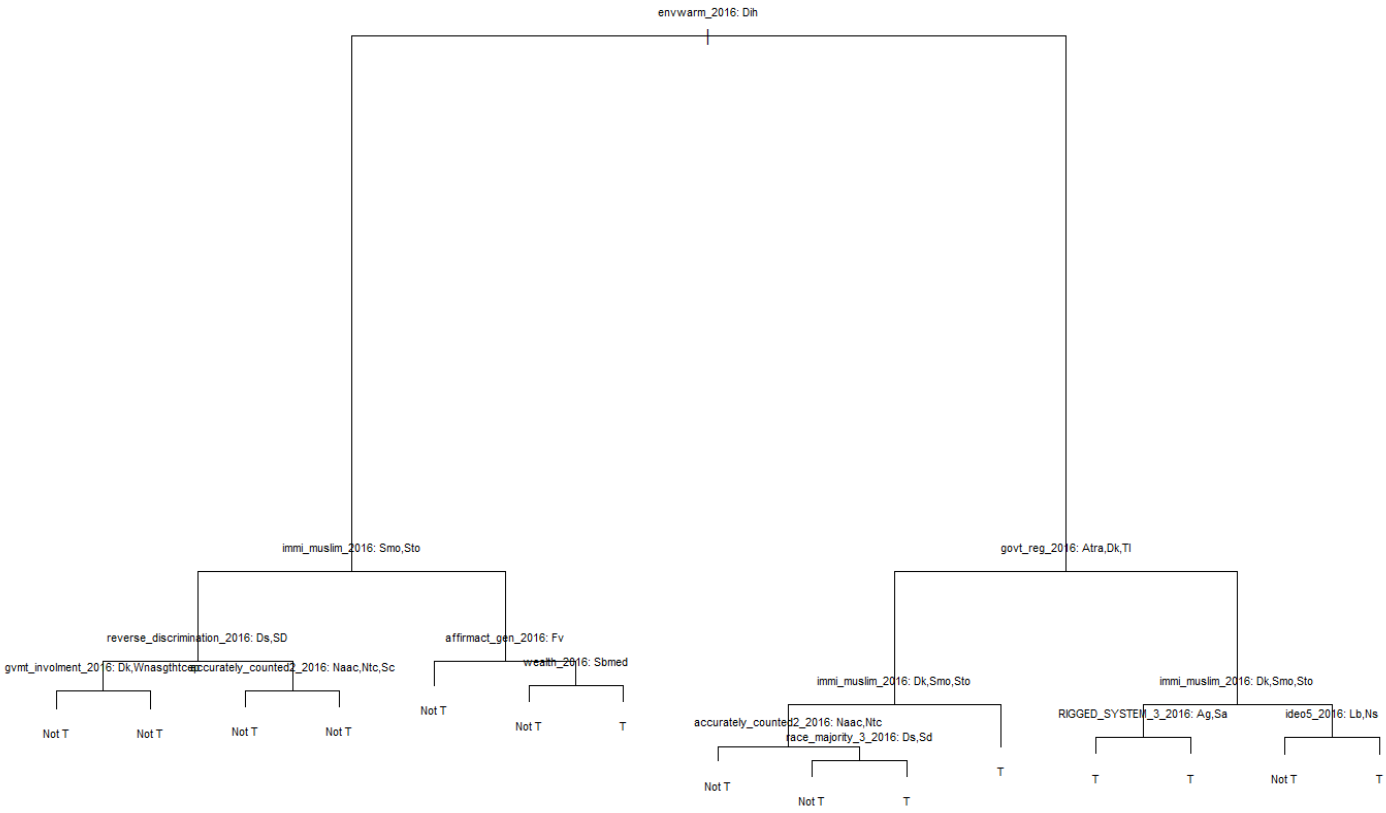

Figure 3.1 – Shows a plot of our decision tree.

For each variable, if respondents chose the response(s) next to the name of the variable, then they flow to the left of the variable. For example, "Dih" next to envwarm_2016 means that if respondents thought that global warming "**D**efinitely **i**s **h**appening", they flow to the immi_muslim_2016 node. The data guide attached explains the shorthand for every node.

Using our model, we used the *predict* function in R to determine which observations in the testing data would vote for Trump, and which wouldn't. The result was as follows:

Firstly, we confirmed that the testing data is roughly balanced, avoiding concerns of imbalanced data which could've biased our results.

Testing Data on Variable 'Vote'	
Not Trump	Trump
1083	708

Figure 3.2 – Shows the number of observations for each level in "Vote.test2".

By comparing the prediction data and our actual data below, we could measure how effective this model is at doing the job. Refer to the tables below for guidance:

Prediction/Actual	Not Trump	Trump
Not Trump	Negative	False Negative
Trump	False Positive	Positive

Prediction/Actual	Not Trump	Trump
Not Trump	987	72
Trump	96	636

Figure 3.2 – Cells meanings and corresponding values according to predictor and test data.

By calculation: the error rate – percentage of predictions which turn out to be false negatives or positives – was only 9.38%.

With regards to our goal in obtaining a good predictor, the classification tree is a much better model than a straightforward 50/50 guess on voting preferences. We have also used R to prune the classification tree to choose the number of terminal nodes that give us the smallest standard deviation in our results – here, the optimal number is 15. Therefore, it can be concluded with some, but not much certainty, that the 15 variables used in the classification tree in **Figure 3.1** are the strongest indicators of populist preferences. According to the table below, it is very clear that identity-related predictors such as opinions on Trump's Muslim ban ("immi_muslim_2016") dominate in numbers. This lends some credibility to our thesis that populism is a function of identity politics as opposed to economic disillusionment.

Variable	Type
envwarm_2016	Identity
wealth_2016	Economics
immi_muslim_2016	Identity
reverse_discrimination_2016	Identity
gvmt_involment_2016	Identity
accurately_counted2_2016	Identity
affirmact_gen_2016	Identity
govt_reg_2016	Identity
race_majority_3_2016	Identity
RIGGED_SYSTEM_3_2016	Identity
ideo5_2016	Identity

Figure 3.4 – Variables used in classification tree.

Having said that, it we wanted to find a way to quantify each variable's importance to fulfil the second goal of this project. Furthermore, a decision tree is prone to overfitting: the tree is made to account for noise just to explain all the variation in the data, without actually increasing (and almost always decreasing) predictive power. The non-robustness of decision trees demands that we use a more sophisticated analytical method.

II – Random Forest Model

Given that the data set contained a large number of variables (668 at the beginning and 157 after cleaning), we also used a Random Forest model to further consolidate our analysis. A random forest model is a more sophisticated version of the decision tree. The fundamental idea is that a random forest combines multiple decision trees (hence, the name) into a single model and takes the average result from the pooled predictions. In the case of classifiers, like our model, the random forest will consider the *majority vote* for our predicted variable "*Vote"*. This accounts for anomalous predictions by individual decision trees, **because the average given large numbers, converge to the expectation.** The word 'random' comes from the fact that each decision tree only considers a subset of variables to be listed as 'questions' and would only use a random subset of the training data. Naturally, this increases diversity of the forest, allowing us to form more robust predictions.

Our prediction model was formed on R the same way we formed a decision tree, but using the *randomForest* function instead. The results obtained were as follows:

Prediction/Actual	Not Trump	Trump
Not Trump	1048	62
Trump	35	646

Figure 4.1 – Corresponding values for predictor and actual data.

By calculation, the error rate is much lower at 5.4%. Very obviously, the model is almost twice as good at predicting voting preferences on the testing data compared to the single decision tree. From the error rate alone, this is a more effective, and hence preferable model at predicting voting preferences.
In addition, the random forest model has a unique function that allows us to quantify the importance of some variables in determining voting preferences. This gives us an added layer of certainty the tree model could not provide.

Variable	Mean Gini Decrease	Type
immi_muslim_2016	57.71233375	Identity
wealth_2016	52.59628677	Economics
govt_reg_2016	46.65258772	Identity
envwarm_2016	46.03576828	Identity
ideo5_2016	41.12829142	Identity
gvmt_involment_2016	38.01105609	Identity
police_threat_2016	29.59764866	Identity
ft_blm_2016	27.98982982	Identity
envpoll2_2016	26.58025828	Identity
imiss_l_2016	24.96886131	Identity

Figure 4.2 – Table of the most important predictors.

For a single partition in a decision tree, the Gini importance measures how much heterogeneity was reduced by the variable through that split. Remember that a decision tree aims to classify the data into a binary – "T" or "Not T". The greater the decrease in heterogeneity (fewer mix of observations that voted and didn't for Trump compared to before) after the partition, the greater a variable's Gini importance. The mean decrease Gini measures how important each variable is over *all* splits that occurred in the tree or forest. This is a good measure for how significant each variable was in determining voting preference.

Now, there is greater nuance with this result: whilst wealth is the only economic factor, its Mean Gini Decrease is very high – second highest at 52.6. Therefore, it plays a crucial role in determining voting preferences for populism. However, if we take the total summation of the top 10 variables alone, the importance of economic disillusionment (by taking "wealth_2016") still pales in comparison to identity in determining populist preferences. Numerically, it is 338.7 compared to 52.6 – supporting our initial thesis on the main driver of populism

III – Methodological Evaluation

The qualitative evaluation on our results will be done in the further section. At this juncture, we first investigated the weaknesses and/or limitations of our approach.

We would've preferred to use another model as well – a logistic regression using the *glm* function in R – for robustness. However, there were two limitations here:

1. The first is that a logistic regression fundamentally requires the data set to be numeric. Most of our variables are factors: ranking of political sentiments, employment, and so on. Whilst most if not all those variables could've been converted into ordered or unordered numeric ranks, we faced a time constraint and couldn't possibly complete the conversion for 668 unique variables within the time frame that we had.
2. Furthermore, a conversion to unordered or ordered numbers might involve some loss in nuance in our model. For example, using the *GLM* on "Region_2016" gives us no meaningful information. The average of 1-9 regions conditional on what we observed in the training data set is merely an arbitrary number – we could've assigned entire different numbers or change the rankings and the value of our predictor may also be very different (by definition).

With regards to our data, we dealt with missing values by removing observations containing missing values in any variable. In an ideal scenario, we would've liked to use other methods to deal with missing values such as the K-nearest neighbour method – this wasn't viable given our timeframe as it was necessary for us to convert the variables to numeric as well. In an even more ideal scenario, we would've liked to gather our own data to control for any possible measurement errors and bias. We cannot be entirely sure whether there was anything systematic about the observations which had missing values.

The dearth of good quality data is also the reason why our model for left-wing populism, by using voting for Bernie Sanders in the democratic primary as a proxy, was ineffective at prediction. Referring to the R-script, there were only 1679 observations to work with upon fully cleaning the data. The resulting tree model had an error rate of 31% while the random forest fared only slightly better with an error rate of 28%. The importance function showed that no variable had double-digit importance bar one at 11. Therefore, we made a judgement call to exclude this from our report.

Finally, on evaluating the performance of our model, we relied on the error rate and confusion matrix to conduct our analysis. This is good compared to the 50/50 model but it would've been better had we been able to use and ROC and the AUC in order to solidify our choice between the two models. Unfortunately, the predictor is neither numeric nor ordered, given that our data is a survey. Therefore, this wasn't possible.

Conclusion

According to the data, the strongest indicator of populism is whether an individual supported Trump's Muslim ban. This can be contextualised with the nativist and anti-immigrant rhetoric employed by Trump and his campaign team, and the conspiratorial nature of the claims made in the run-up to the 2016 elections.

Other relevant predictors also seem largely ideological. For example, views on reverse discrimination, global warming, or the accuracy of the electoral count etc. all appear important in determining whether a person voted for Trump or not. Broadly, all this is consistent with the view espoused by Galston et al. who argue that "populism draws strength from public opposition to mass immigration, cultural liberalization, and the perceived surrender of national sovereignty…" (Brookings, 2018)

However, equally, it is worth noting that populism is not completely devoid of ground economic realities. This is because while most of the strong indicators are ideological, the wealth of an individual still plays some role in determining whether someone voted for Trump or not. This gives some credence to the views of Martin Wolf and others who point at stark income inequalities and stagnating middle-class wages as the prime drivers of populism. Furthermore, while our analysis largely points towards ideology as the main force behind populism, it is hard to discount that some ideologies are formed *because of* economic discontent.

This research proposes a novel new way to look at populism, utilising data science and machine learning, and attempts to quantify the main drivers of populism. However, while this analysis can be used to understand the 2016 US elections, it might be hard to extrapolate these results onto different countries and different time-periods. For example, it could entirely be possible that in other elections

ground economic realities play a much larger role in electoral dynamics. As such, for further avenues of research, it could be interesting to look at different elections or referenda – examining the main predictors for the Brexit referendum or the French and German elections can greatly enrich the literature on the topic. Lastly, further research can also be done on right-wing versus left-wing populism by, for example, looking at the determinants of a Sanders vote.

Bibliography

1. Berlet, C. and Lyons, M.N., 2018. *Right-wing populism in America: Too close for comfort*. Guilford Publications.

2. Galston, W.A., 2018. *The Populist Challenge to Liberal Democracy*

3. Hawkins, K.A., Carlin, R.E., Littvay, L. and Kaltwasser, C.R. eds., 2018. *The Ideational Approach to Populism: Concept, Theory, and Analysis*. Routledge.

4. Mudde, C. and Kaltwasser, C.R., 2017. *Populism: A very short introduction*. Oxford University Press.

5. Norris, P. and Inglehart, R., 2019. *Cultural Backlash and the Rise of Populism: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press.

6. Rodrik, D., 2018, May. Is Populism Necessarily Bad Economics. In *AEA Papers and Proceedings* (Vol. 108, No. May, pp. 196-199).

7. The Democracy Fund Voter Study Group., 2017. *The 2016 Views of the Electorate ("VOTER") Survey.*

8. The United States of America Census Bureau., 2017. *Census Bureau Region and Division Codes and Federal Information Processing System (FIPS) Codes for States*

[i] https://www.ft.com/content/5557f806-5a75-11e7-9bc8-8055f264aa8b

[ii] https://www.economist.com/graphic-detail/2018/11/03/how-to-forecast-an-americans-vote