Han-Nuri Kim
Spring Semester 2021

Barking Up the Wrong Tree: The Challenge of Classifying American Voters

Classification Tree Models in the Context of the 2020 U.S. Presidential Elections

Past literature on political polarization has time and again shown that the US Congress is becoming progressively homogenized in its policy positions whereas the differences between the two parties' stands on major policy issues are steadily increasing. This polarization however has not stopped at government officials, but extends to the parties' mass identifiers and activist bases (Layman, Carsey, and Horowitz 2006). Furthermore, studies emphasize that politics are growing more divided along orthodox-progressive lines (Hunter 1991). However, more recent literature points out that there has been an erroneous tendency to categorize Trump voters as homogenous bloc with similar tastes and preferences, which ignores the nuanced typologies of different Trump supporters[1] (Ekins, 2017). Motivated by the recent findings, this project attempts to build a statistical learning model that is able to accurately predict whether an individual voted for Trump in the 2020 U.S presidential elections or not. The research question can be formulated as follows: *"How good are statistical learning tools to classify Trump voters based on an individual's beliefs and socioeconomic status?"* This project will firstly describe the data, present the statistical method of choice and lastly compare and analyze the results of the models.

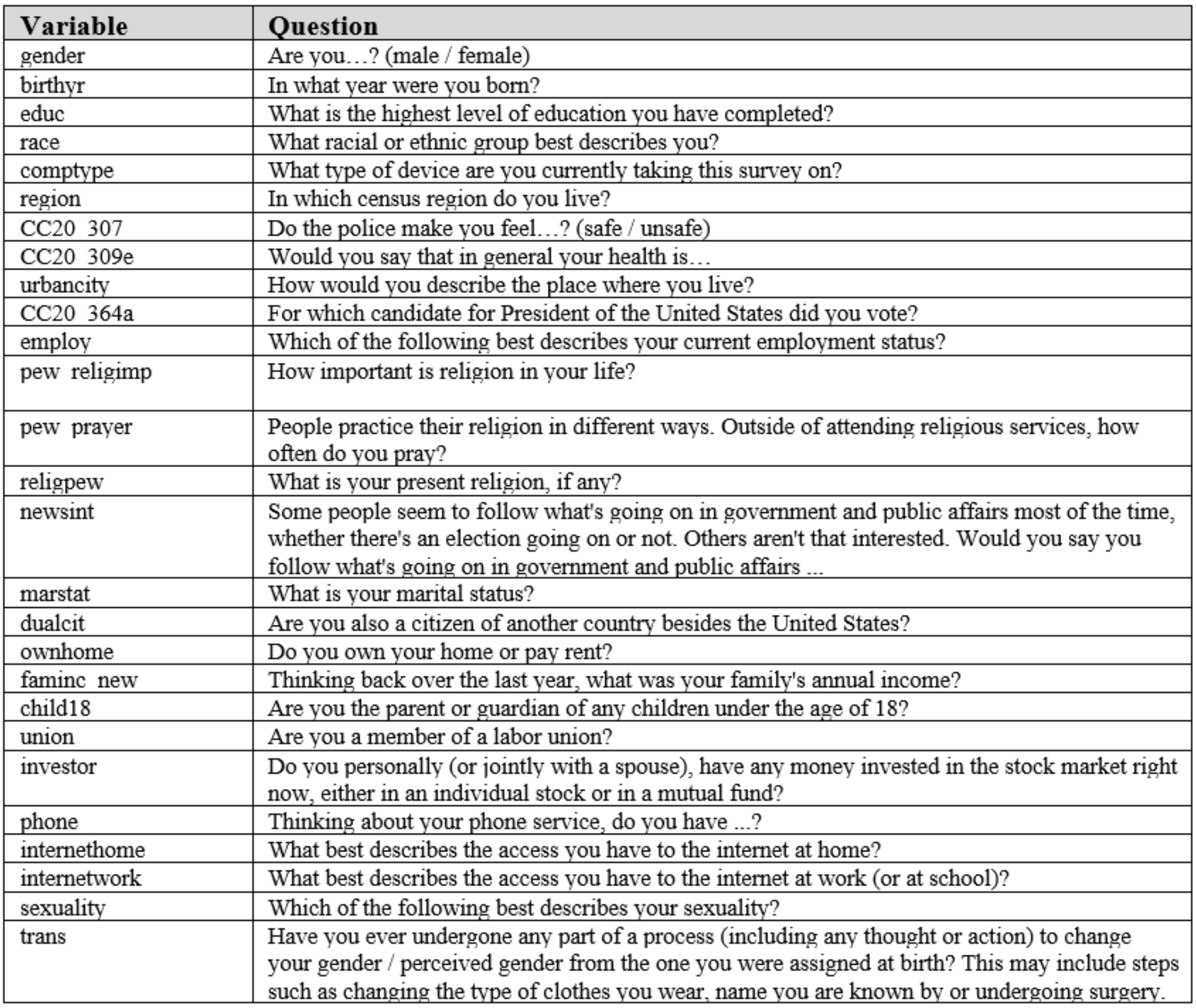
DESCRIPTION OF DATA

In the framework of this project, survey data from the Cooperative Election Study was used (Schaffner, Ansolabehere, and Luks 2021). This survey dataset of 61'000 respondents asked individuals questions concerning their socioeconomic status, their political beliefs and who they voted for in the aftermath of the 2020 presidential elections.[2] The covariates incorporate questions that give information on respondent's socio-economic status and beliefs, while the dependent variable is their 2020 presidential election vote. Concerning variable selection, not all questions were optimal to use for the analysis. Questions regarding party affiliation (democrat versus republican) and self-evaluation of an individual's ideology (liberal versus conservative) were dropped, on the basis that they are assumed to correlate very highly with the independent variable and which each other (multicollinearity). Furthermore, questions that allow for multiple answers were dropped to facilitate the analysis. For the next step, all missing observations of the independent variable were dropped. This step reduced the size of the data from approximately 61'000 observations down to 10'000 observations, which means that there was a significant amount of the population that declined to answer whether they voted for Trump or for Biden. This non-response bias represents an inherent flaw of the dataset that cannot be easily resolved, outside of classical imputation methods.

[1] Ekin's article distinguishes between 4 archetypical trump voters (Free marketeers, Anti-Elites, American Preservationists, Staunch Conservatives) based on a voter survey in 2016. For more information, see: https://www.voterstudygroup.org/publication/the-five-types-trump-voters

[2] The whole list of questions can be found in the codebook: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/E9N6PH

By looking at the distribution of observations with Trump and Biden votes, another problem became apparent: From the 10'000 observations left, 78% of individuals had voted for Biden, whereas only 21% of observations belonged to Trump voters. This class imbalance is insofar problematic because it skews the learning algorithm in classification problems (Fernández et al. 2017). To account for this, I used the "ROSE" package, which uses a random over-sampling bootstrap technique to balance the classes (Lunardon, Menardi, and Torelli 2014). I over-sampled the minority class (Trump voters) to balance the dataset. However, it was important to carefully distinguish between training and testing dataset before applying random over-sampling. I only applied cross-validation through random over-sampling *after* splitting up into training and testing set and only on the training dataset, to avoid any data leakage. In the end, the data preparation left me with a balanced training sample of 12'000 observations and an imbalanced validation set of 2'000 observations with 27 covariates each. I did not discard the imbalanced train sample, but used it to compare the results to the balanced data.



Variable	Question
gender	Are you…? (male / female)
birthyr	In what year were you born?
educ	What is the highest level of education you have completed?
race	What racial or ethnic group best describes you?
comptype	What type of device are you currently taking this survey on?
region	In which census region do you live?
CC20 307	Do the police make you feel…? (safe / unsafe)
CC20 309e	Would you say that in general your health is…
urbancity	How would you describe the place where you live?
CC20 364a	For which candidate for President of the United States did you vote?
employ	Which of the following best describes your current employment status?
pew religimp	How important is religion in your life?
pew prayer	People practice their religion in different ways. Outside of attending religious services, how often do you pray?
religpew	What is your present religion, if any?
newsint	Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs ...
marstat	What is your marital status?
dualcit	Are you also a citizen of another country besides the United States?
ownhome	Do you own your home or pay rent?
faminc new	Thinking back over the last year, what was your family's annual income?
child18	Are you the parent or guardian of any children under the age of 18?
union	Are you a member of a labor union?
investor	Do you personally (or jointly with a spouse), have any money invested in the stock market right now, either in an individual stock or in a mutual fund?
phone	Thinking about your phone service, do you have ...?
internethome	What best describes the access you have to the internet at home?
internetwork	What best describes the access you have to the internet at work (or at school)?
sexuality	Which of the following best describes your sexuality?
trans	Have you ever undergone any part of a process (including any thought or action) to change your gender / perceived gender from the one you were assigned at birth? This may include steps such as changing the type of clothes you wear, name you are known by or undergoing surgery.

METHODOLOGY

Due to the nature of my research question and the cross-sectional dataset, I decided to train various classification tree algorithms. Decision trees and their derivates in general allow for a great mix between interpretability and accuracy. Since my dataset almost exclusively makes use of categorial data, it made sense to choose a learning algorithm that mimics human decision-making to predict from a set of covariates of unseen observations whether a person is more likely to vote for Trump or for Biden. As for the procedure, instead of jumping directly to the most optimized learning technique, I started from a model with a single tree, and attempted progressively improved model prediction through different techniques and by increasing the model complexity each time. This gave me a clear outline how to proceed with the analysis. First, the implementation of a single decision tree model with manually tweaked parameters. Second, a random forest model with further hyper-tuning, and lastly, a gradient boosted decision tree model with variations in loss function and individual tree parameters. Five-fold cross-validation was applied to all models during the training process. Concerning the evaluation metric of a model, accuracy itself is not a very reliable method to assess classification algorithms, since it will be validated on the test-sample, which has class-imbalance. Therefore, I used the Kappa and the AUC metric, which control for class imbalance and distribution.

ANALYSIS

1. Single Tree Model with Pruning

For the single decision tree model, I started with the default specification of the package regarding the minimum number of observations at the nodes (20) and the leaves (10), the complexity parameter (set to zero), as well as the maximum depth of a subtree (30). Validating the model on the imbalanced test dataset resulted in a kappa value of 0.31. Next, I proceeded with tree pruning, where the cross-validation error is minimized, and then the subtrees are retrospectively cut back. This resulted in a slightly improved value of 0.36. However, the algorithm still wrongly classified Trump voters more than 50% of the time. Regarding the kappa values, I want to have at least a predictive accuracy of 0.4.[3] Visualizing this pruned tree was challenging, since the tree grew to a model with 255 splits. This made it impossible to recognize any individual characteristics of the tree. Therefore, I visualized a simpler tree with a maximum depth of 6, to give a brief overview of the relative variable importance.

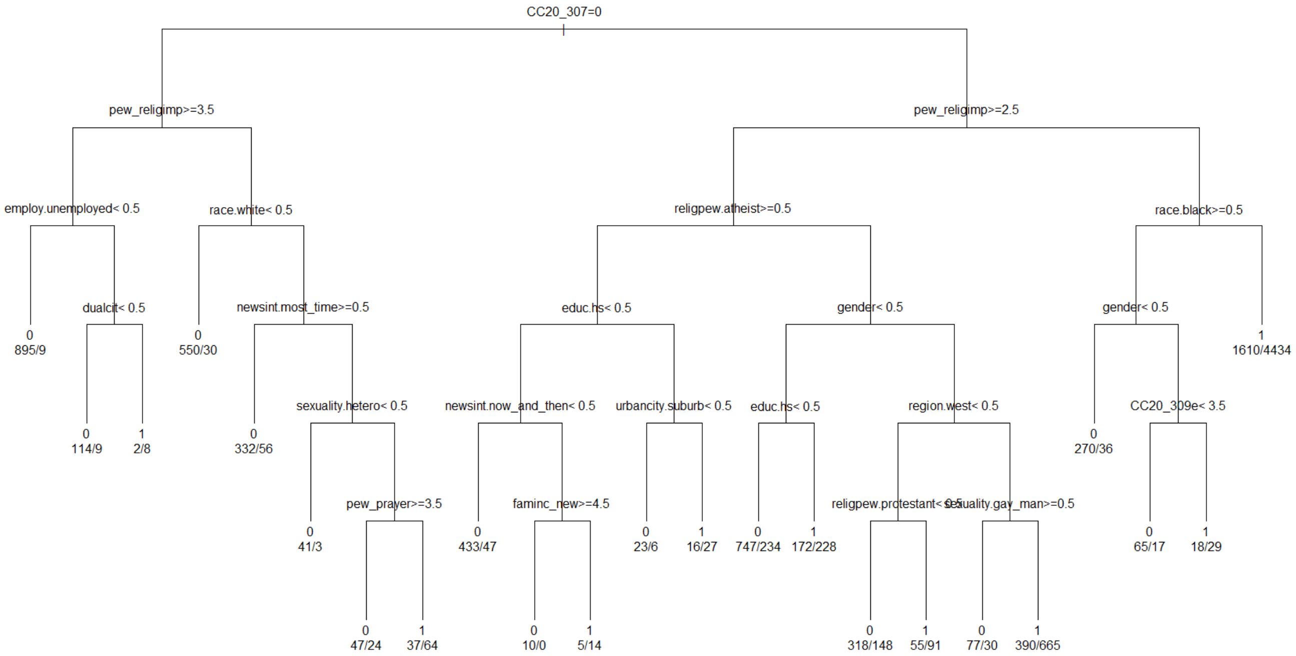


Figure 1 Decision Tree with maximum depth of 6

[3] According to the Queen's University School of Medicine website, kappa values starting from 0.4 are considered moderate. URL: https://elentra.healthsci.queensu.ca/assets/modules/reproducibility/kappa_values.html#:~:text=Kappa%20Values&text=Generally%2C%20a%20kappa%20of%20less,of%20%3E0.75%20represents%20excellent%20agreement

We can derive relative variable importance from the simplified tree. It shows that beliefs about the police and religion play an important factor when classifying voters. Further varying minimum observations, maximum depth and different complexity parameters did not improve the model beyond the default specifications.

2. Random Forest Approach

For the random forest approach I tweaked the mtry parameter that decides how many random variables the model takes at each split. For the balanced training sample, I found the optimal number of m to be 13, with an accuracy of 0.40. With the original, imbalanced training sample I was able to achieve a much higher accuracy of 0.91 with an m-value of 20. This result is to be expected, since the imbalanced original data tends to overestimate its own accuracy due to covariate shifting.[4] Therefore I trusted the first model more.

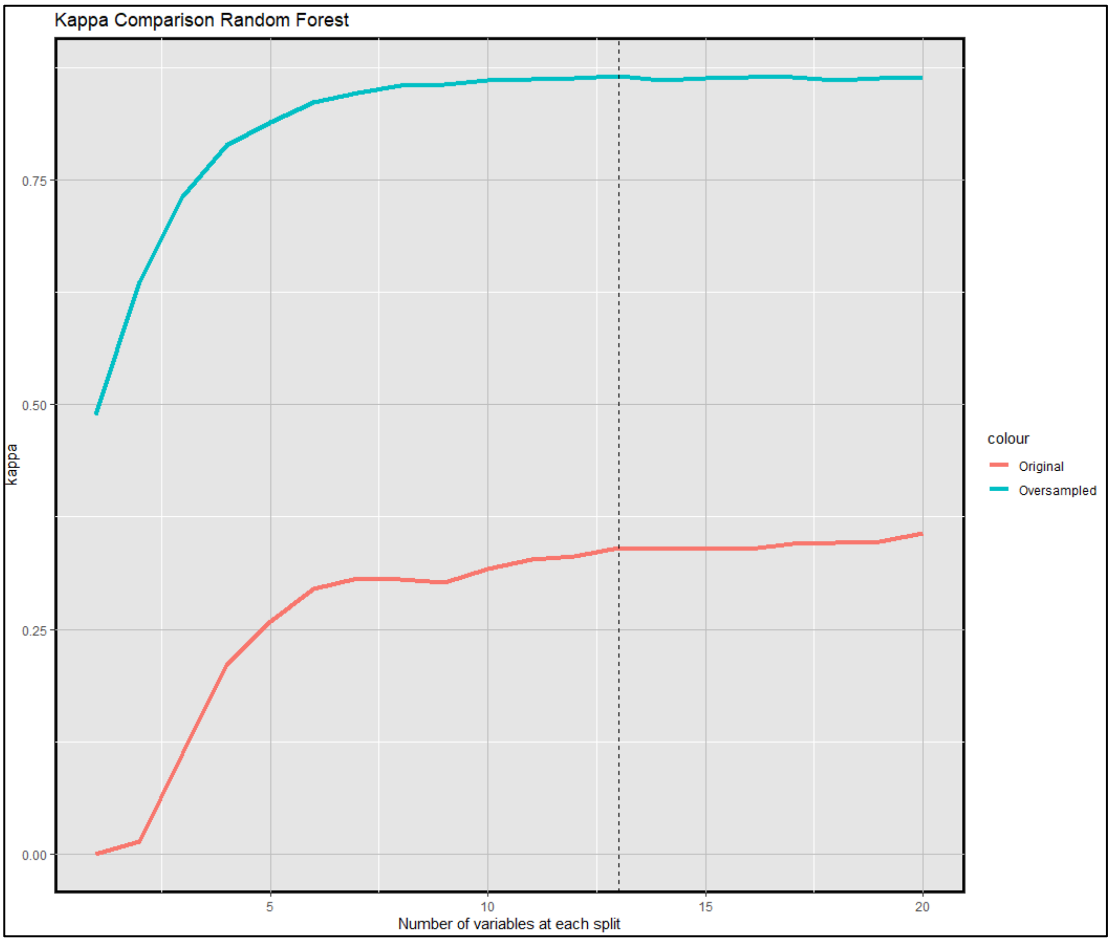


Figure 2 Kappa Comparison between original train data and oversampled train data

[4] More on dataset shifting and prior probabilities is discussed in this book: Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., & Schwaighofer, A. (Eds.). (2009). Dataset shift in machine learning. Mit Press.

3. Gradient Boosted Trees

For the last approach, I used gradient boosted trees to try to further improve my model accuracy from the random forest. I tuned the parameters in this order: Max. tree depth ➔ Step size shrinkage ➔Subsample parameter ➔ Subsample of each tree ➔ Gamma. Another feature used in gradient boosting was the early-stopping parameter. This function stopped the training process once the error term did not improve further to avoid overfitting and to save time. This parameter was fixed to 10.

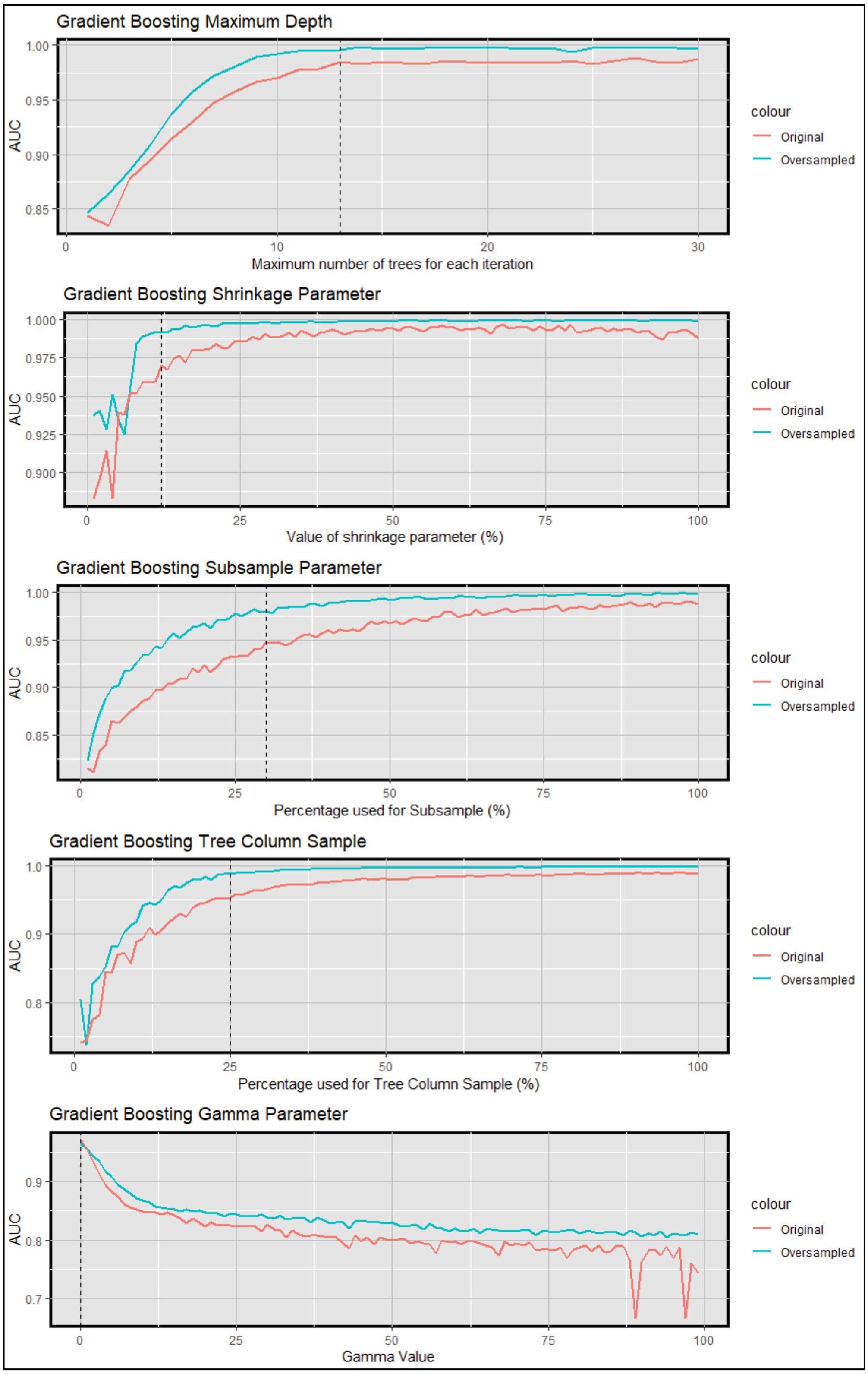


Figure 3 Parameter Tuning for Gradient Boosted Trees

After I tuned all parameters, I evaluated both the balanced and the imbalanced train sample with my hold out sample. The final kappa value for the balanced training set was 0.45, which is a further improvement from the random forest model. The value for the original imbalanced dataset was 0.68, which is significantly better than for the oversampled data. Thus, we can clearly see that oversampling the test data made the model put more weight on the smaller class, while at the same time reducing overall accuracy when validating on the test dataset.

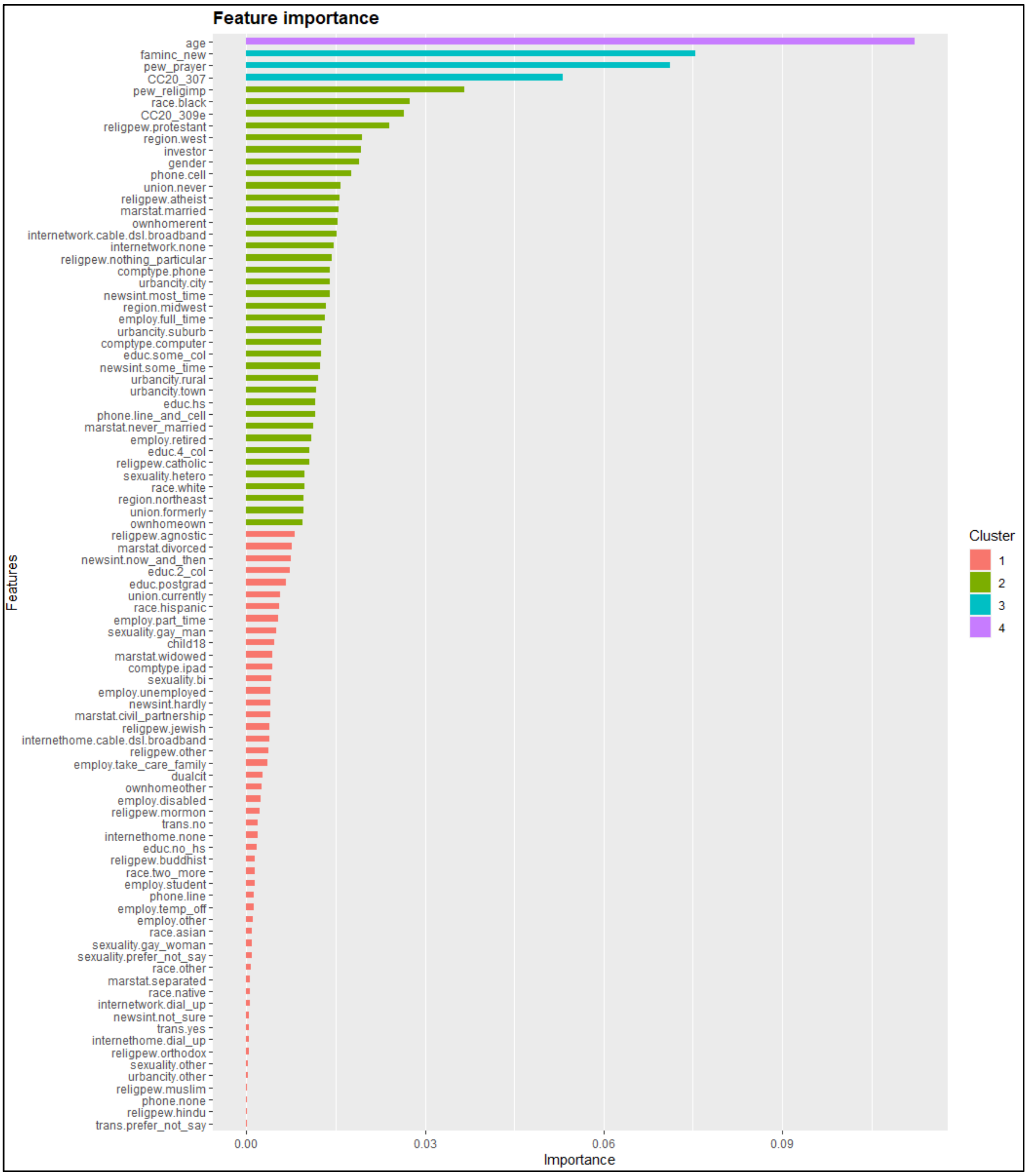


Figure 4 Feature Importance of all covariates

CONCLUSION

The results give us an insight on what factors might play an important role whether someone is a Trump supporter or not. The most important factors were age, family income, religious behavior and whether individuals felt safe in the presence of police or not. The kappa values received from test validations across all models still showed a great margin of uncertainty with regards to predictive power.
In this project I was able to steadily improve the predictive power of the model, starting from single trees, and later all the way to gradient boosted trees. Special emphasis was given to the question of over-sampling versus keeping the imbalanced train data as it is. My models have shown consistently that training data of the same class proportions as the hold out sample will yield better predictive power than balancing the training data beforehand. The question which model is better boils down to how we expect data from the real world to look like. We would expect the distribution of Trump and Biden voters in the US to be much more balanced than 80-20. Therefore, I can say that the model trained on the balanced dataset is less biased and is more appropriate to use on further unseen data. There are other limitations imposed on the model that cannot be resolved on its own, with the most important one being non-response bias in the survey data. Furthermore, the decision to over-sample the training data is also subject to bias, since new observations were created synthetically. As for further avenues for analysis, this project suggests re-training the models by imputing the missing responses of the independent variable with highly colinear covariates like party affiliation and ideology, which would certainly get rid of the non-response bias.

To conclude, it became abundantly clear that classifying voters is not in any way an easy task. With even the most computationally powerful methods like gradient boosting and countless ways of tweaking the parameters, I was not able to reach high levels of predictive precision. This result is a tentative affirmation of past literature that emphasizes the futility of bundling together groups of people based on their political preferences.

Word count: 1794

REFERENCES

Ekins, Emily. 2017. "The Five Types of Trump Voters." *Democracy Fund Voter Study Group, June. Available (accessed 21 June 2017) at: https://www. voterstudygroup. org/reports/2016-elections/the-five-types-trump-voters*.

Fernández, Alberto, Sara del Río, Nitesh V. Chawla, and Francisco Herrera. 2017. "An Insight into Imbalanced Big Data Classification: Outcomes and Challenges." *Complex & Intelligent Systems* 3(2): 105–20.

Hunter, James Davison. 1991. "Culture Wars: The Struggle to Define America: Making Sense of the Battles over the Family." *Art, Education, Law, and Politics*.

Layman, Geoffrey C., Thomas M. Carsey, and Juliana Menasce Horowitz. 2006. "PARTY POLARIZATION IN AMERICAN POLITICS: Characteristics, Causes, and Consequences." *Annual Review of Political Science* 9(1): 83–110.

Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. 2014. "ROSE: A Package for Binary Imbalanced Learning."

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. *Cooperative Election Study Common Content, 2020*. Harvard Dataverse.