ANES Election Data Classification (CS289 Project)

Axel Amzallag and Kevin Sun

1. Background and Introduction

The United States has been experiencing high levels of polarization in politics over the past two decades. Most voters now vote reliably for the same party in presidential elections no matter the candidate, and the number of persuadable voters in the United States has been decreasing over time. As the parties have divided politically, there have been questions asked about whether voters for the Democratic and Republican party have also been diverging. If both parties continue to become more different from each other and the country continues to divide more along political lines, it could have important ramifications for the future of American life and politics.

If it is possible to identify which party a voter is likely to choose in an upcoming election for based on multiple different groups of attributes, then it may be the case that political identity is becoming tied to beliefs that are seemingly unrelated to political attitudes. If that were true, both the Democratic and Republican parties could more effectively target voters who are likely to vote for their party, even if that potential voter does not have past voting history or has never been registered to vote. Political scientists and demographers would also be interested in this knowledge; one potential use would be to see whether voters are similar along more personality and social dimensions than they used to be. If that were the case, they could study people are living in politically homogeneous neighborhoods and how often they interact with a politically diverse set of neighbors and social peers. If political views can be predicted along many personality and social dimensions, then it may lead to citizens interacting with a less representative group of peers.

In this analysis, we try to analyze a representative sample of voters in the 2016 election along multiple dimensions (Social Class, Political Activity, Racial Attitudes, etc.) using a machine learning approach with random forests. To this end, we use data from the 2016 American National Election Survey (ANES), which asks participants a variety of questions about their lives, from their political beliefs to their personality to their child-rearing habits. Participants were interviewed before and after the election, and how they voted was also recorded. In order to be certain that the voters surveyed didn't change their mind about which candidate to select, only the post-election interview was used in this analysis. The interview questions were grouped into categories and a machine learning model was fit to each category separately. In this way, we can analyze with more granularity the connection between voting behavior and more deeply held beliefs and attitudes.

2. Methodology

2.1. Data

The ANES Time Series post-election data has over 371 variables, each one of them asking a question to the participants of the the study. The questions are quite varied, and include ones that are overtly asking about party identification (e.g Did respondent donate money to Hillary Clinton's campaign?), some are clearly about government policy (e.g Did health care law affect respondent's health care costs?), and some are about topics that seem unrelated to politics (e.g Which one do you think is more important for a child to have: Independence or respect for elders?). We are targeting the latter two types of questions for this analysis, and we're particularly interested in questions of the third type. We manually went through the 371 questions asked to participants and put them into the following subjects:

- Community
- Gender
- Health Care and Science
- International Relations
- Nationalism
- Personality/Values
- Views on Political System
- Racial Attitudes
- Financial/Social Class Attitudes
- Miscellaneous

Out of these subjects, four contain almost exclusively questions unrelated to political attitudes: Nationalism, Community, Views on Political System, and Personality/Values.

The questions in these subjects tend to be more about overarching values and beliefs of the individual, for example what kind of person they are, how they raise their children, and how they feel when they see an American flag. The remaining subjects contain many questions directly related to policies, although none asking about party identification.

The majority of questions are categorical, where the respondent marks 1-Agree completely, 2-Agree partially, and so on. Only feeling thermometer questions are continuous, where each respondent rates their feeling toward a group/concept (0 indicates strong dislike, 50 neutral, 100 strong like).

2.2. Processing

Using the `pandas` library in Python, we began the processing by extracting our outcome variable, V162034a, from the data set. This question asks what Presidential candidate the respondent voted for in the 2016 election. We filtered the data to only include respondents that voted for 1 - Hillary Clinton and 2 - Donald Trump. We then subject the data into different sets corresponding to the categories mentioned above. While doing so, we took note of which subject features were continuous (e.g. Feeling thermometer questions).

For each subject, we first filter out any observations with more than 40% missing values across the selected features. We then use `sklearn`'s `preprocessing` tools to scale the data and imputed our missing values using a K-nearest neighbors imputer before scaling back.

Then, we use one-hot encoding on only the categorical features. We then combined our encoded features with the continuous features.

2.3. Fitting

Because our problem is classification using mostly categorical data, we focus on decision trees as our fitting mechanism. Decision trees are known to work well with combinations of categorical and quantitative features, and allow for arbitrarily complicated decision boundaries, which we expected since our features have multiple categories. While we began by fitting single decision trees, we opted to pursue ensemble methods for better fits - namely random forests and AdaBoost.

Random forests fit a large number of decision trees, utilizing bagging and randomized subset selection for each tree. It then predicts the class based on the features for each tree, and averages the overall posterior probability before rounding to the final predicted class.

For each subject, we fit both a random forest and AdaBoost ensemble classifier and assessed validation performance. For both ensemble learners, we tuned the splitting criteria, between gini impurity and cross-entropy, and also the maximum depths of the trees. Tuning was done using 5-fold cross-validation, and models were evaluated on the validation accuracy using 5-fold cross-validation. For random forests, the max depth varied from 1 to 20, while the AdaBoost depth varied from 1 to 5. Note that both the random forest and AdaBoost learners were tuned separately, and independently for each subject.

Finally then took the three highest-accuracy subjects and fit a random forest and AdaBoost classifier on all features from the combined subject to see what the increase in accuracy would be. We again used 5-fold validation for hyperparameter selection and validation accuracy measurement.

Table 1. Results for variable groupings using RandomForest, along with the best hyperparameters found using cross-validation. Criterion and Max Depth refer to the decision tree parameters.

SUBJECT	SCORE	CRITERION	MAX DEPTH
COMMUNITY	0.561009	ENTROPY	3
GENDER	0.785625	GINI	10
HEALTHCARE/ SCIENCE	0.845159	GINI	10
INTL RELATIONS	0.802111	ENTROPY	10
NATIONALISM	0.801784	GINI	20
MISCELLANEOUS	0.759336	GINI	10
PERSONALITY/ VALUES	0.781652	GINI	10
POLITICAL SYSTEM	0.704434	GINI	10
RACE	0.830738	ENTROPY	20
SOCIAL/ FINANCIAL	0.832184	ENTROPY	20

3. Results

3.1. Random Forest

Using the random forest classifier (Table 1) on all of each subject separately, we found that the majority of our categories had validation accuracies between 0.70 and 0.84. The only subject with a validation accuracy close to chance was the Community category (accuracy = 0.56). This was expected since the Community category did not contain any questions directly related to government policy. Seven of the subjects that we created had validation accuracies between 0.78 and 0.84, despite some containing many questions directly related to policy (e.g Healthcare/Science) and some containing almost no questions directly related to policy (e.g Personality/Values).

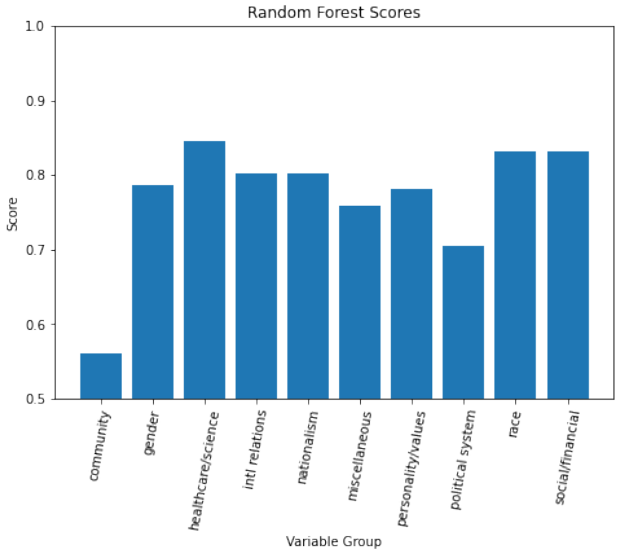

Figure 1. Validation accuracy for our tuned Random Forest ensemble learner across each question subject.

The similarity of all of these validation accuracies is an indication that voting behavior cuts across many different dimensions. If voters chose their candidates based exclusively on their economic impact, we would expect that the subjects related to economic policy, such as Social/Financial and Healthcare/Science, to be much more predictive than other subjects. However, the difference between these and other subjects is quite small, signaling that voters can behave in predictable ways based on other subjects such as racial and gender beliefs.

One other result to highlight in particular is the validation accuracy of the Personality/Values and Nationalism subjects. While they were thought not to contain any questions related to policy, they nevertheless had relatively high validation accuracies – 0.78 and 0.80, respectively. This could be evidence supporting the hypothesis that personality and values are becoming more intertwined with political beliefs, and that most citizens within a political party are becoming more similar along personality and social dimensions. The homogeneity of the parties along this axes could create a political environment unlike those of the past in the United States.

3.2. AdaBoost

Using the AdaBoost classifier (Table 2) on all of each subject separately, we found that the majority of our categories had validation accuracies between 0.70 and 0.84, the same at the random forest classifier. There was no appreciable improvement in validation accuracy from the AdaBoost classifier. This can be seen when we compare the validation accuracies side-by-side in Figure 3.

Table 2. Results for variable groupings using AdaBoost, along with the best hyperparameters found using cross-validation. Criterion and Max Depth refer to the decision tree parameters.

SUBJECT	SCORE	CRITERION	MAX DEPTH
COMMUNITY	0.554924	GINI	1
GENDER	0.788881	ENTROPY	1
HEALTHCARE/ SCIENCE	0.832999	GINI	1
INTL RELATIONS	0.810218	GINI	1
NATIONALISM	0.800988	ENTROPY	1
MISCELLANEOUS	0.762172	GINI	1
PERSONALITY/ VALUES	0.777187	GINI	1
POLITICAL SYSTEM	0.704444	ENTROPY	1
RACE	0.824590	ENTROPY	1
SOCIAL/ FINANCIAL	0.835017	GINI	1

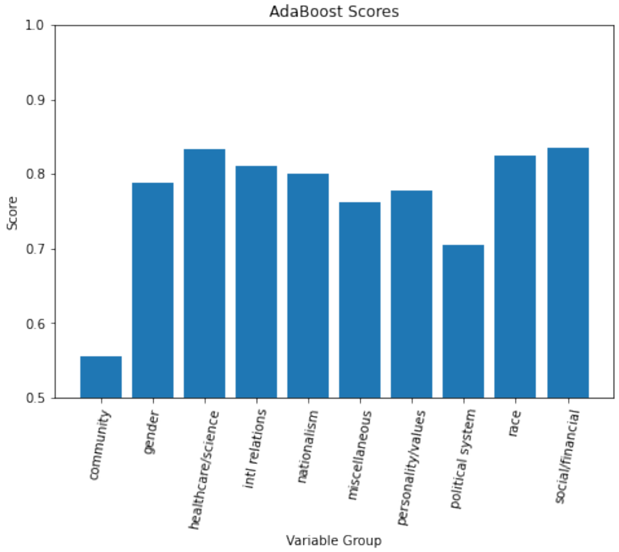

Figure 2. Validation accuracy for our tuned AdaBoost ensemble learner across each question subject.

3.3. Joint Analysis

For our final analysis, we looked at the top three subjects together, Healthcare/Science, Race, and Social/Financial, to see if together they improved the accuracy of our classifiers. In both the random forest and AdaBoost classifiers, we found that overall validation accuracy increased to 0.89 when the classifiers were fit utilizing three of these subjects.

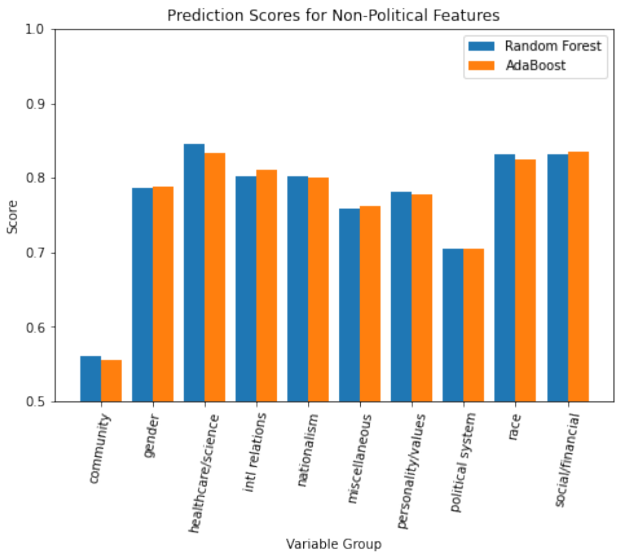

Figure 3. Validation accuracy for our tuned AdaBoost and tuned Random Forest ensemble learners across each question subject.

Table 3. Results for combined variable groupings (Healthcare/Science, Race, and Social/Financial) using tuned AdaBoost and Random Forest ensemble learners. The score was from 5-fold cross validation.

Model	Score
Random Forest	0.891321
AdaBoost	0.895776

4. Conclusion

In this project, we used random forest and AdaBoost ensemble learners to determine political outcomes (specifically, whether a person votes for the Democratic of Republican presidential candidate), using their responses to nonpolitical questions. These responses were categorized into various nonpolitical topics, and contained both categorical and ”continuous” responses. We use one-hot encoding for categorical features and trained random forests and AdaBoost learners on each topic. We found that, across all topics, our best random forest and AdaBoost performed very similarly. We combined the three questions subjects with the highest predictive power and were able to improve our overall validation accuracy, but still did not see a significant difference between the AdaBoost and random forest classifier.

This project creates many opportunities for further analysis with this dataset. For further work, we would analyze the feature importances of the various questions used for each topic. This would allow further insight into which non-political factors of a respondent’s life most strongly influence their voting tendency. After identifying such features, we would collect the strongest predictors from each category and fit models with those features and determine if that would improve our predictive accuracy.