Modeling Political Identity

CS 229 Final Paper

Katharina Roesler and Ben Krausz

June 6, 2016

1 Introduction

Political scientists have done extensive research on the mindsets of Democrats, Republicans, and Independents, and have developed several theories as to why individuals identify as each. For instance, they have posited that "pure Independents" are distinct from partisans and "weak" Independents in their political attitudes (Magleby 2012).

In this project, we focus on self-identified Independents, who comprised 11 percent of Americans in 2012 and have a diversity of attitudes. Our goal is to identify features that predict who is Independent. We are interested in these features in their own right, in order to understand the mindset of Independents, as well as in order to build an effective model.

2 Data

We use data from the American National Election Studies (ANES), which surveyed Americans' political attitudes from 1948 to the present. This survey is the gold standard in political science and contains 55,674 observations and 951 attributes. We use only responses from 2012, in order to have comparable and consistent measures of each feature, resulting in a training set with 3,365 people and 88 attributes.[1] Please see our Github repository for all of our code.

3 Algorithms

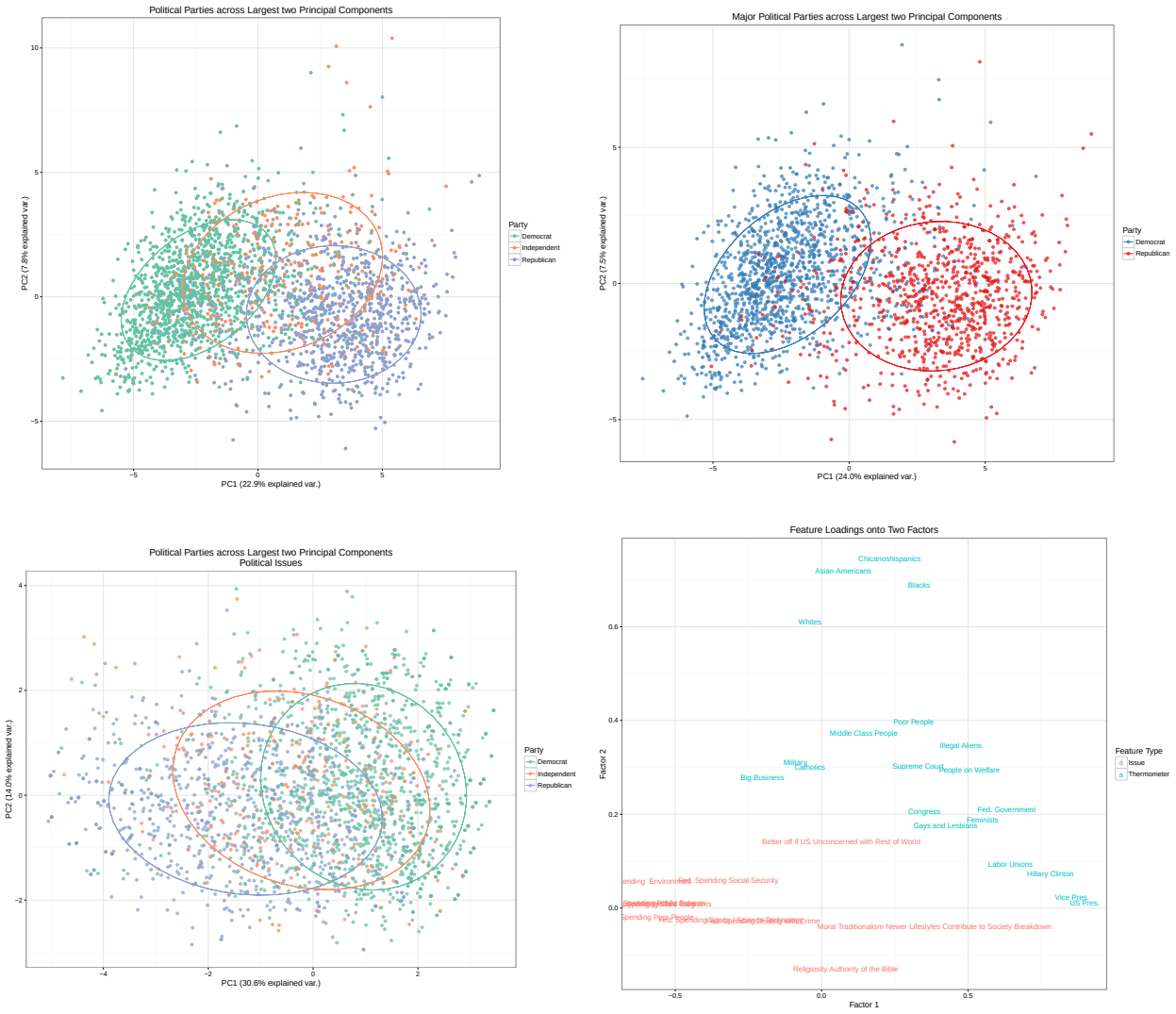
3.1 Principal Components Analysis

The goal of principal components analysis (PCA) is to represent the data with fewer dimensions, reducing noise in the data and capturing directions in which the data vary. For instance, one component identified might correspond to having a positive disposition, or distrusting the government.

We first ran PCA on the data with all 55 numeric features. The primary two components capture a considerable amount of variance related to political identity. That is, the first two components can be used to separate Democrats from Republicans, but Independents are similar to both Democrats and Republicans ("partisans") in terms of these components. Please see the figures below, which shows individuals' distribution along the primary two components for the full data set (left) and for Democrats and Republicans alone (right).

We next ran PCA on subsets of the data including only features relating to individuals' warmth to certain people and groups and towards certain issues, respectively. The first subset includes measures such as how warmly individuals feel towards Black people and Congress, while the latter includes measures such as support for abortion legalization and welfare spending. These components were even less effective at capturing variation in party identity, as is to be expected.

[1] Certain analyses use smaller subsets, due to algorithms' particular constraints. For instance, principal components analysis requires numerical features, for which reason we include only 51 features when conducting principal components analysis.



3.2 Factor Analysis

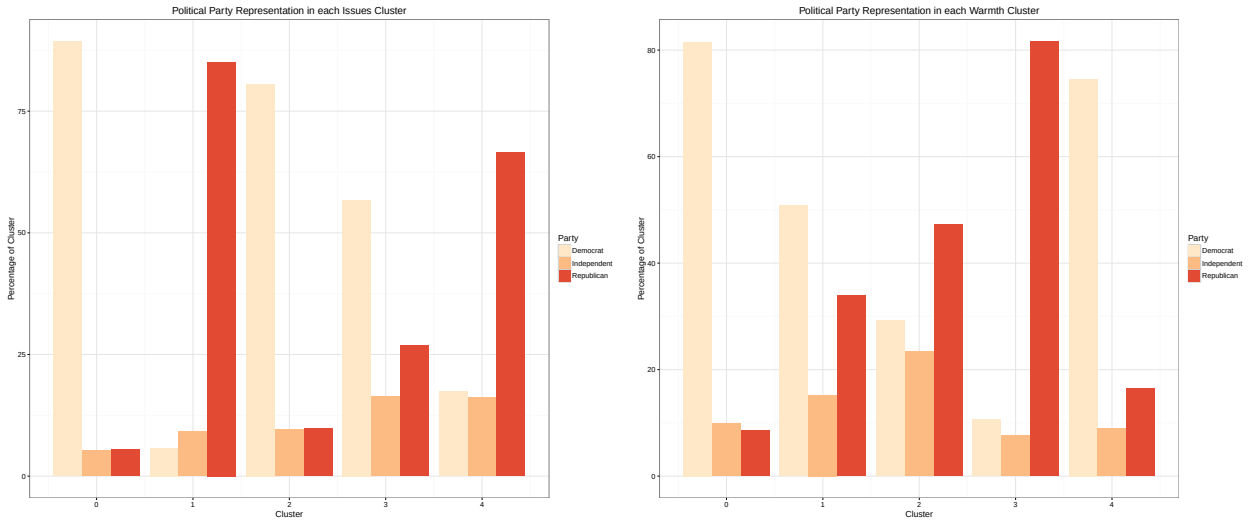
In order to verify that "warmth" features and attitudes to political issues are two core types of features, we ran factor analysis on a subset of the data with only these features. We confirmed that warmth features load similarly to one another on the two factors estimated, in relation to issues features. The plot above (right) shows how warmth and issues features loaded onto the two factors, with warmth features shown in blue and issues features shown in red. Warmth features are more positively related to factor 2 than are issues features, indicating that they are a different "type" of feature, and that analyzing these feature sets separately is appropriate.

3.3 K-Means

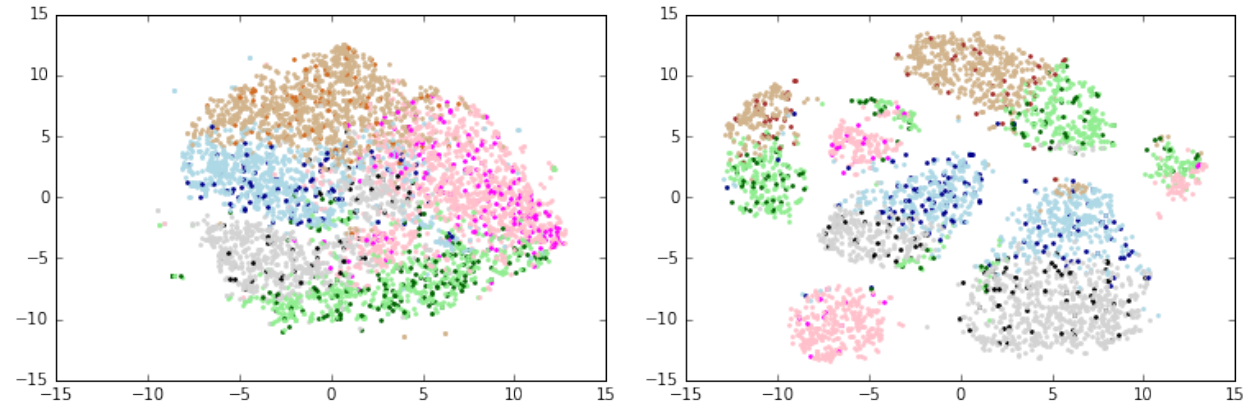
We theorized that it might be possible to split political Independents into a number of different groups based on their political ideologies. The ANES dataset has two types of political ideology features: issues and political thermometers. Issues features involve questions such as "when should abortion be allowed: always, sometimes, or never?" Political thermometers ask respondents to rate their warmth towards groups such as blacks, Big Business, and labor unions on a 0-97 point scale. We ran k-means for each feature group,

trying cluster sizes between 1 and 20 and plotting their inertia. We created 5 clusters for each features set, resulting in clusters displayed below using T-SNE.

We next identified the clusters with the greatest number of Independents. For the political thermometer k-means, the cluster with the largest percentage of Independents was a a cluster of people who had negative opinions of every single group asked about (23 percent Independent), and the second largest was a cluster of people who mostly felt neutral towards all groups asked about (15 percent Independent). For the issues k-means, the clusters with the largest percentages of independents were social conservatives/fiscal liberals ("left Christians"), and social liberals / fiscal conservatives ("Libertarians").



Issues (left) and thermometer (right) clusters scatterplot (bolded points represent independents):



Note that though the aforementioned clusters contained the most independents, they did not contain overwhelming majorities.

3.4 Hypotheses

Based on these results, we theorized that Independents would be relatively more likely to fall into one of four groups: "pessimists," "moderates," "left-Christians," and "Libertarians." We conducted a t-test for each of these criteria and compared the percentage of Independents and partisans in each category. We found that Independents were more likely to be Moderates, Left-Christians, and Libertarians, but not "left-Christians." [2] In fact, we found that about 47% of independents fit in to at least one of the first three groups, compared

[2] Using a 95% confidence level.

to 26% of Democrats, and 35% of Republicans. This difference is statistically significant result but does not enable high-confidence classification of individuals' party identity.

3.5 Boosting

Principal components analysis and k-means clustering suggest that individuals' political attitudes and warmth towards particular groups only partially explain their political identities. For this reason we use boosting with decision stumps to "sift" through our data and determine which features best predict Independence. We ran 2,000 iterations of boosting with a logistic loss function on a training set with 2,524 observations and 88 attributes. We specified an "interaction depth" of 3, meaning that two- and three-way interactions are included. We find that error on the validation set is smallest at 332 iterations, after which the model begins to overfit. We next ran boosting with an exponential loss function, which had the lowest validation error at 356 iterations.

Boosting with logistic and exponential loss functions resulted in very similar rankings of features. Both identified "state" as the feature with the most "relative influence," whose inclusion in the model most reduced the loss across all iterations. Other influential features are how interested the individual is in the presidential election, whether he or she voted, and how warmly he or she feels towards Congress.

Because state seems to be highly related to political identity, we mapped states' percentage of Independents in 2012. Please see our interactive map online, in which you can choose which year and political party you would like to view. Overall, it seems that a few states have particularly high proportions of Independents, including Colorado and New Hampshire. However, please keep in mind that our sample size is not very large[3], so this relationship may be largely due to noisy data.

3.6 Logistic Regression

Having identified key features that are most useful for predicting whether an individual identifies as an Independent, we build a logistic regression classifier using several feature sets. First, we estimate models with the principal components we found, with the factors we found, and with only warmth or issues features. Next, we estimate a model with all features, as well as models with only the ten most influential features identified by our boosting models. Coefficients from the logistic regression model including the top ten features from boosting with a logistic loss function are shown below. It seems that people who are less interested in national elections are less likely to be Independent, as are people who voted in the past national election.

Table 1: Logistic Regression predicting being Independent

	Model 1
Interest in Pres. Election	-1.251^{***}
Warmth to Congress	-0.010^{***}
Warmth to US President	-0.004
Voted in National Elections	-0.718^{***}
Social Class	-0.145^{***}
Warmth to Supreme Court	-0.001
Gov. Pays Attention to People	-0.278^{**}
Warmth to Middle Class	-0.008^{*}
US Better Off Unconcerned with World	-0.310^{**}
(Intercept)	-9.090
Num. obs.	3,365

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

[3]For instance, we have only 69 people in Colorado

We compared the error rates for logistic regression models with each feature set, as shown below. All models performed similarly poorly, severely underestimated the percentage of Independents when evaluated on the training set. This is likely because our features only explain a very small portion of the variance in political identity, and incorrectly estimating that someone is Independent is more "expensive" than estimating that they are partisan, given how rarely individuals are Independent. That is, the "safe" bet is to classify every individual as Independent and only predict Independence given an overwhelming amount of evidence.

Table 2: Logistic Regression Error Rates for each Feature Set

	Error Rate	Estimated Positive	True Positive
Principal Components	11.61	1.35	12.36
Warmth Principal Components	12.36	0.08	12.36
Issues Principal Components	12.36	0.00	12.36
Two Factors	12.36	0.00	12.36
Four Hypotheses	11.58	0.00	11.58
All Features	11.26	2.35	12.30
All Features (Test Set)	12.31	13.00	11.32
Top 10 Logistic Boosting Features	11.71	1.07	12.30
Top 10 Logistic Boosting Features (Test Set)	11.57	12.10	11.32
Top 10 Logistic Exponential Features	11.50	1.10	12.30
Top 10 Logistic Exponential Features (Test Set)	10.95	12.26	11.32

However, our models estimated approximately the correct proportion of Independents on the test set, as "random" noise in the new data caused our model to estimate false positives just as often as false negatives. This is to be expected and is reassuring in a sense, as our models will not drastically underestimate the proportion of Independents in unseen data.

4 Conclusions and Limitations

We find that classifying Independents based on their political philosophies is very difficult and prone to high error rates. Nonetheless, we find that Independents are more likely to be moderate, pessimistic, and/or libertarian than partisans, and are more likely to be from certain states.

Although we feel that this analysis is somewhat informative, we also feel that we must be very cautious when interpreting our results. In addition, we feel it is important to keep in mind that analyses involving 88 attributes include only individuals who answered 88 questions completely. In our case, that means including only 3,365 out of 4,748 survey respondents and hoping that these 3,365 are representative of the entire sample (and ideally the entire United States population). Also, we only looked at a single year in our study. According to Professor Morris Fiorina, independents often change whether they classify themselves as "leaners" or "true" independents between election years. Our analysis of independents might be tied to the specific circumstances of 2012.

5 References

J.H. Friedman (2001). "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* 29(5):1189-1232.

D.B. Magleby and C. Nelson (2012). "Independent Leaners as Policy Partisans: An Examination of Party Identification and Policy Views," *The Forum* 10(3).

G. Ridgeway (2007). "Generalized Boosted Models: A guide to the gbm package." saedsayad.com

G. Ridgeway (2015). "Package 'gbm'." cran.r-project.org

M. Fiorina (2014). "Are Independents Truly 'Closet Partisans'? Some Facts About Political Independents"