Improving Polling Accuracy in US Presidential Elections

Using Machine Learning to Exclude Non-Voters from Polling Results

Derek Zhao

May 15, 2017

Contents

1 Introduction

Public opinion polling faces many challenges related to sampling methodologies, but these obstacles are compounded when the polling attempts to predict voter behavior in a national presidential election. In particular, even if a truly sufficiently large random sample of the population was interviewed, the pollster has limited means of determining which respondents will actually vote in the upcoming election. This uncertainty over who is likely to vote contributes a significant degree of error to polling results.

While prior electoral participation is a strong predictor of future participation, self-reported information on past voting behavior is frequently unreliable due to social desirability bias[1], and validating the interviewee's responses via a state voting registry is costly and time consuming. Furthermore, interviewees are often uncomfortable with providing the personal information necessary to make such validation possible.

Supervised machine learning, a form of statistical learning in which a predictive model fine-tunes itself based on data it is exposed to, provides a potential solution to the problem of identifying non-voters within a survey's sample.

1.1 Project Goals

This project aims to use machine learning techniques as a means for classifying voters and non-voters and assess the impact of such classifications on the accuracy of public opinion polls in recent US presidential elections. It will explore two main questions.

- Given a post-election survey that contains data on whether or not each respondent participated in that year's presidential election, which machine learning models are most successful in differentiating non-voters from voters?
- Given the best model for predicting whether survey respondents are voters or non-voters, does using the model's predictions to exclude likely non-voters lead to more accurate polling results?

1.2 Dataset

Supervised learning for classification requires labelled data, that is, data for which the correct class label is known. For example, in the classical email spam detection problem, a collection of emails would need to contain labels indicating whether each email constitutes spam in order for a model to learn how to classify said emails. In the context of a non-voter detection problem, traditional pre-election polls, like those conducted by Gallup or Quinnipiac University, do not qualify as labelled data because, by the nature of their timing, they cannot give any indication as to whether or not the respondent participated in the election.

However, American National Election Studies (ANES) conducts time series surveys of eligible voters[2]in which the same respondents are interviewed prior to an election as well as after the election, thus providing a posteriori data on their voting behavior. ANES is a publically funded election research organization that has conducted time series studies since 1948. Currently, surveys are conducted both in person and online with an approximate duration of 80 minutes.

Of particular interest is the ANES Cumulative Data File, a merged dataset containing common features from all time series studies conducted since 1948. Because each feature in the Cumulative Data File incorporates data for the same survey question from multiple time series studies, the dataset is particular suitable for:

[1] Social desirability bias is a type of response bias that describes the tendency of survey respondents to answer questions on sensitive topics in a manner that presents themselves in the best possible light. In the context of voting behavior, it takes the form of over-reporting electoral participation.

[2] ANES surveys are conducted from the sample universe of eligible voters, US citizens age 18 or older who are eligible to be registered to vote. This is not to be confused with a sample universe of registered voters or likely voters.

1. Training supervised learning models on historical data and testing the models' performance on a later election year.
2. Replicating results over several election years.

This project is concerned only with the national popular vote of recent presidential elections, so the relevant time series studies from the Cumulative Data File are from the years 2000, 2004, 2008, and 2012[3].

Finally, the surveys from the relevant years use a stratified sample rather than a simple random sample. Thus, accounting for the provided sampling weights is crucial to making inferences about the voting eligible population (VEP). While these sampling weights are unimportant to the process of detecting non-voters, they are incorporated into any calculations involving polling results.

1.3 Project Outline

The structure of this report is as follows.

- Section 2 details the steps taken in the data wrangling process. Decisions about how to represent certain features are made during data preparation, and the consequences are discussed.
- Section 3 presents some relevant insights discovered during exploratory data analysis. The section will examine the distributions of potentially important features as well as the relationship between certain features and respondent voting propensity.
- Section 4 discusses modifications to the feature set that were considered and describes an iterative method for finding the optimal subset of features for model training and respondent classification.
- Section 5 explores the performance of various supervised learning models for classifying voters and non-voters, given each models' optimal feature set. In addition to popular algorithms like logistic regression, random forests, and support vector machines, the section also explores the use of a soft voting classifier. The results here answer the first of the two aforementioned project questions.
- Section 6 presents a method for using the voter and non-voter classifications output by the best-performing models to improve the accuracy of the ANES surveys in estimating the support of presidential candidates. The results of this method are discussed and provide an answer to the second of the two aforementioned project questions.
- Section 7 summarizes the project's key findings and discusses avenues for continued research.

Figure 1 illustrates how the processes of sections two through six relate to one another when the models and survey estimates are evaluated for three election years, 2004, 2008, and 2012. Data wrangling and exploratory data analysis are conducted independently of the test election year, while feature selection and model tuning must be performed on different sets of training data to account for the different test years.

2 Data Wrangling

The ANES Cumulative Data File is a rectangular dataset consisting of 55,674 rows, each representing a respondent, and 953 columns, each representing a feature. The respondents span interviewees from all ANES time series studies from midterm and presidential election years between 1948 and 2012, inclusive. The features span all interview questions that appeared across multiple time series surveys.

The wrangling process for this project occurs in two steps. First, the dataset is filtered; a combination

[3] The ANES 2016 Time Series Study was recently made available, however its features and responses have yet to be recoded in a form compatible with the Cumulative Data File. While such recoding is certainly possible, it is outside the scope of this project.

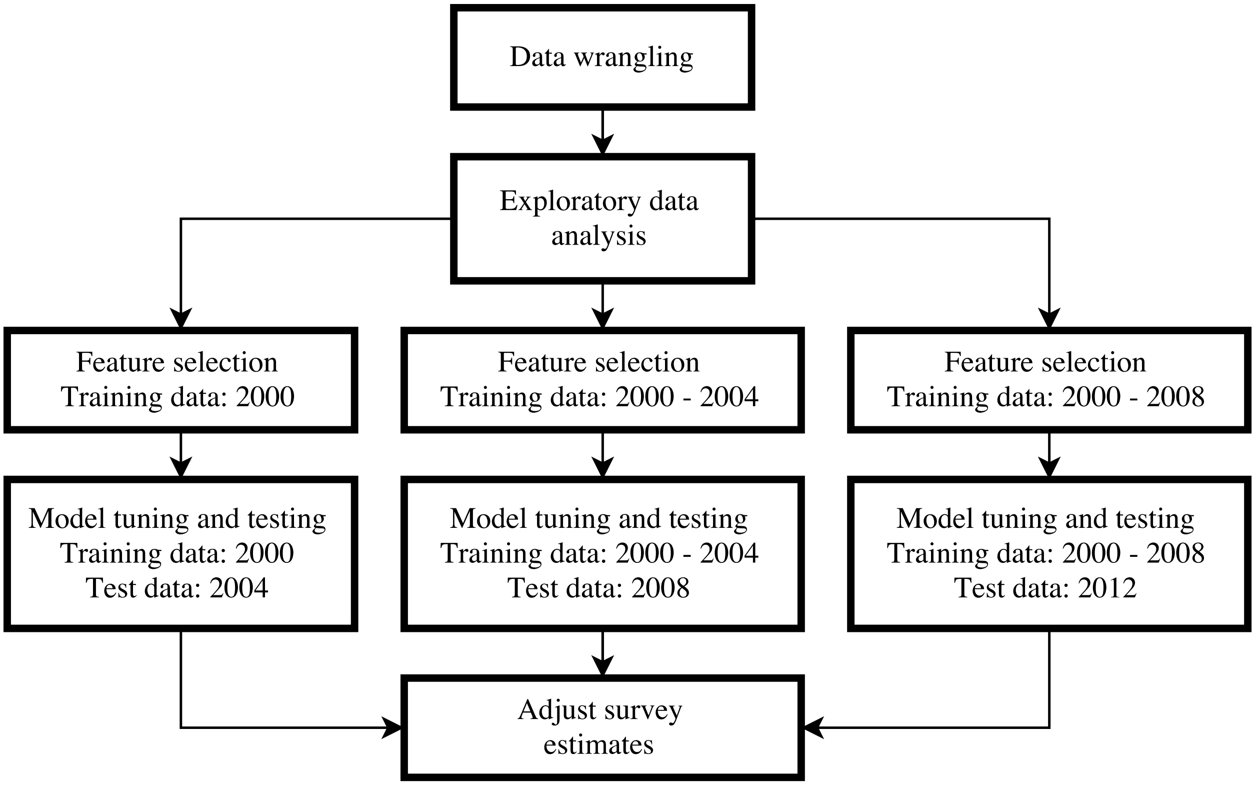
Data wrangling
Exploratory data analysis
Feature selection Training data: 2000
Feature selection Training data: 2000 - 2004
Feature selection Training data: 2000 - 2008
Model tuning and testing Training data: 2000 Test data: 2004
Model tuning and testing Training data: 2000 - 2004 Test data: 2008
Model tuning and testing Training data: 2000 - 2008 Test data: 2012
Adjust survey estimates

Figure 1: **Project flow**

of respondents and features are removed so that only relevant data remains. Second, the dataset is processeed; many features are re-encoded into a representation more suitable for analysis and machine learning later on.

2.1 Respondent Filtering

A large number of the 55,674 respondents are dropped from the dataset for a variety of reasons. Because this project is only concerned with those respondents interviewed for the four presidential elections from 2000 to 2012, all respondents interviewed prior to 2000 are removed, as are those interviewed during midterm elections. A significant number of respondents interviewed prior to the election did not participate in a follow-up post-election interview, or if they did, declined to answer questions regarding their voting behavior. Thus, they are dropped because whether they voted is unknown. Respondents that participated in abbreviated pre-election interviews are also removed from the dataset simply because they have too much missing data for meaningful comparison with other respondents. Table 1 summarizes the respondent filtering process.

2.2 Feature Filtering

A number of redundant features are manually dropped from the dataset. For example, each respondent is assigned three each of Type 0, Type 1, and Type 2 sampling weights, nine weights in total; however, all nine weights are the same for respondents interviewed in and after 2000, so only one of the nine features is necessary. Similarly, some features record the same attribute, but with different numbers of categories. For example, one *race summary* feature may contain three categories, {white, black, other}, while another *race summary* feature may contain five categories, {white, black, hispanic, asian, other}. This also constitutes a redundant set of features of which only one is necessary. Features that are administrative in nature, such as *interviewee ID*, *respondent ID*, *interviewer gender*, etc. are also removed from the dataset.

Reason	Number respondents removed	Number respondents remaining
Interviewed prior to 2000	42,908	12,766
Interviewed during midterm election year	1,511	11,255
No post-election data	1,022	10,233
Abbreviated pre-election interview	836	9397
No voting data	23	9,374

Table 1: **Respondent filtering**

While the features from the Cumulative Data File span all interview questions that appear in multiple ANES time series surveys, many features may correspond to interview questions that were no longer asked after 2000 or discontinued in the time from 2000 to 2012, resulting in significant amounts of missing data. Given that 9,374 respondents are of interest for this project, all features with fewer than 6,000 non-null values are dropped. Table 2 below summarizes how features are filtered.

Reason	Number features removed	Number features remaining
Redundant	12	941
Administrative	24	917
Missing data	654	263

Table 2: **Feature filtering**

2.3 Feature Encoding

Filtering the data as described results in a dataset of 9,374 respondents and 263 features. These features fall into one of five categories:

1. **Continous**: These features measure a number or amount of something. There are relatively few features of this type, and aside from replacing outliers with null values, no additional processing needs to be performed. Examples:
    - Age
    - Number of children
    - Number of political discussions in past week
2. **Quasi-continuous**: Theoretically, these features ought to be continuous, but in practice they are not. Quasi-continuous features correspond exclusively to one of two types of survey questions:
    - **Thermometer**: In these types of questions, the respondent is presented with a subject, either a person, group, or institution, and asked to rate this group on a scale from 0 to 100, with 0 indicating a negative sentiment and 100 indicating a positive sentiment. Examples:
        - Person: President, Vice President, Democratic Senate Candidate, ...
        - Group: Hispanics, Middle Class, LGBT, Christian Conservatives, Democrats, ...
        - Institution: Congress, Supreme Court, Republican Party, Military, Labor Unions, ...

- **Index**: Not technically survey questions, indices are constructed from various related *thermometer* questions. For example, a respondent that answers 90 for a *Democratic party thermometer* question and 40 for a *Republican party thermometer* question might have a *partisanship index* of $\frac{90+(100-40)}{2} = 75$. Examples:
  - Major party
  - Major party presidential candidates
  - Ideology

For *thermometer* questions, while respondents may select any whole number from 0 to 100, they typically prefer to answer in multiples of 10 or 25, resulting in irregular spikes and gaps uncharacteristic of a truly continuous feature. To address this, some smoothing is applied by dividing all *thermometer* and *index* values by 10 and then rounding to the nearest integer.

3. **Ordinal**: These features have a limited number of possible values, but these values have an intrinsic ordered relationship with one another. For example, features corresponding to survey questions where respondents must choose from {strongly disagree, disagree, neither agree nor disagree, agree, strongly agree}, also known as a Likert scale, are re-encoded from a string representation to an integer representation {1,2,3,4,5}. In cases where one of the possible responses, such as {I do not know} do not fit into an ordered relationship with the other responses, a new binary feature is created to record the presence of the unorderable response while all such responses in the original ordinal feature are converted to null values. Ordinal features fall into three main categories:

   - **Scale**: A large number of survey questions ask the respondent to place a person or institution on a seven-point ideological scale with respect to a policy debate. For example, in a *Democratic presidential candidate government services scale* question, respondents are asked where they believe the Democratic presidential candidate stands on the issue of providing government services by selecting a value from 1 to 7, with 1 corresponding to {Government should provide fewer services} and 7 corresponding to {Government should provide more services}. Examples:
     - Republican presidential candidate, health insurance ∈ [1 - Government insurance plan, 7 - Private insurance plan]
     - Democratic party, aid to blacks ∈ [1 - Government should help more, 7 - Blacks should help themselves]
     - Self, standard of living ∈ [1 - Government should guarantee everyone a good job and standard of living, 7 - Government should let each person get ahead on his own]
   - **Federal spending**: These types of questions all follow the same format: *Federal spending on [program] should* {be increased, stay the same, be decreased}. Program examples:
     - Aiding the poor
     - Dealing with crime
     - Public schools
   - **Other**: Examples:
     - Does respondent feel better or worse off in the past year? {Much better, Somewhat better, Same, Somewhat worse, Much worse }
     - When should abortion be allowed? {Never, In cases of rape or life of mother in danger, Need established, Always}
     - How much does government waste tax money? {A lot, Some, Not very much}

4. **Categorical**: Unlike ordinal features, categorical features correspond to survey questions where the possible responses do not have an ordered relationship with each other. Thus, each categorical feature is re-encoded using one-hot encoding, where the presence of each response is recorded as a binary feature. Examples:

   - Religion - {NA, Protestant, Catholic, Jewish, Other/None}

- Census region {Northeast, North, South, West}
- Work status {Working now, Temporarily laid off, Unemployed, Retired, Permanently disabled, Homemaker, Student}

Some categorical features include response categories that apply to very few respondents. In these situations, the sparse category is either combined with a larger category or converted into null values, decided on a case by case basis.

5. **Binary**: A special class of categorical features that correspond to survey questions for which there exists only two possible responses. Features values are converted from their string representation to either a 0 or 1. Examples:

   - Did respondent vote in the national election?
   - Does respondent care who wins the presidential election?
   - Does respondent own a home?

3 Exploratory Data Analysis

This section presents some of the most relevant characteristics of the dataset, including the distributions of various features, the relationships between certain features and the target variable (whether the respondent voted), and visualizations of the data following dimensionality reduction. Because the purpose of this analysis is to better understand the data itself rather than draw inferences about the voting eligible population, unless otherwise stated, sampling weights are ignored.

3.1 Unbalanced Data

The dataset is unbalanced in two important aspects. First, the number of respondents interviewed per election year is not constant, increasing dramatically from 2000 to 2012.

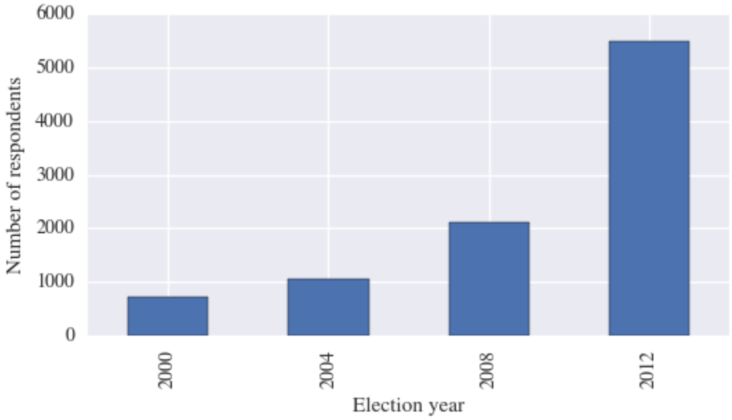

Figure 2: **Number of respondents per election year**

In all cases, the number of respondents interviewed in a particular election year is greater than the number of respondents interviewed prior, resulting in the non-ideal situation where all training sets contain less data than their corresponding test sets. Consequently, the ideal number of features for predicting vote participation in 2004 using 2000 data for training is much lower than that for 2012 using 2000 through 2008 data for training, as too many features would cause overfitting problems for a smaller training set.

Second, for each election year, the number of voters and non-voters is sufficiently different that accuracy

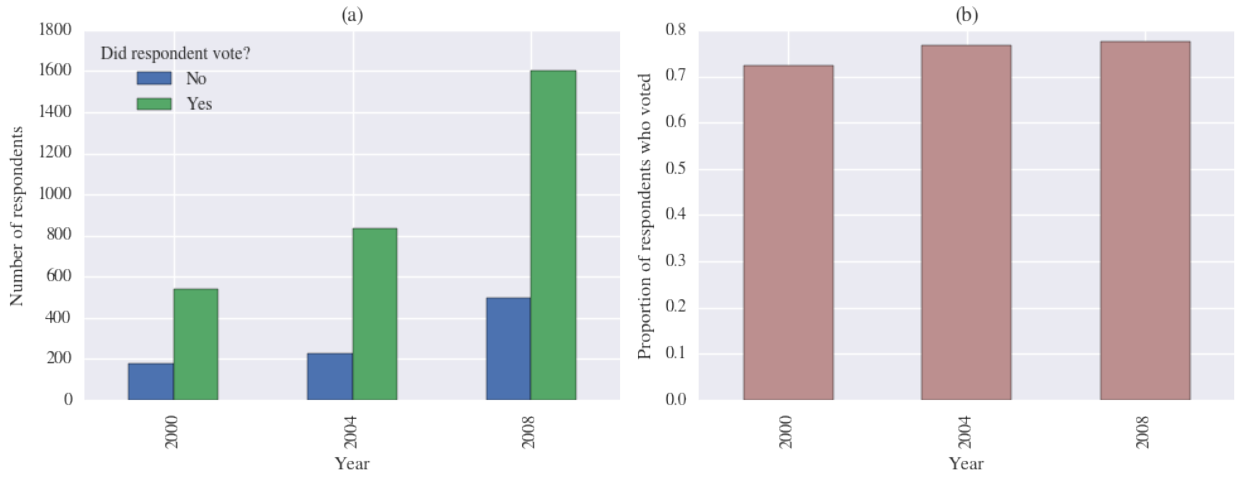

Figure 3: **Voter turnout by election year**. Because data from 2012 is used exclusively as test data, it is excluded from further analysis so that it cannot influence decisions that should be made solely by examining training data. (a) Far more respondents claim to have voted than those who do not. (b) Sampling weights are applied to infer the proportion of the VEP that voted in each election.

is no longer a meaningful measure of a predictive model's effectiveness, necessitating the use of a different performance metric, discussed later.

Of particular concern is the fact that the voter turnout statistics in Figure 3.b do not match widely accepted knowledge. Whereas the ANES data suggests that at least 0.7 and sometimes almost 0.8 of the VEP participated in the national election, well-respected data from the United States Elections Project (Figure 4) indicates that voter turnout is much lower, typically 0.6 of the VEP. This suggests that ANES voter turnout data is severely inflated, likely a result of social desirability bias, and that the ground truth the data presents is unreliable. A method for addressing this obstacle is discussed in Section 6.

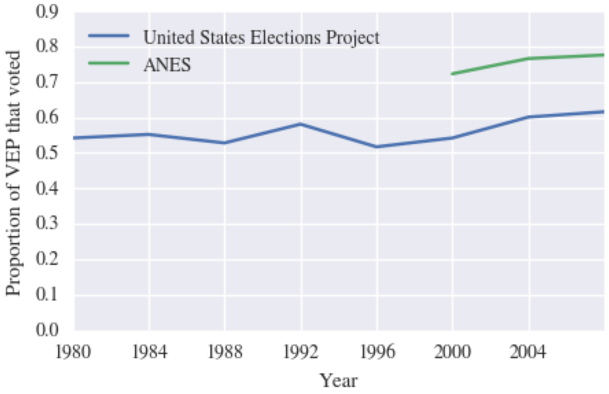

Figure 4: **Voter turnout data differs by source**. According to the United States Elections Project, between 0.54 and 0.62 of the VEP has participated in national elections from 2000 to 2008. However, according to ANES, this range is from 0.72 to 0.78.

3.2 Low-dimensional Visualization

Using principal component analysis (PCA) to project a dataset's feature space onto a three-dimensional subspace provides a convenient means of visualizing high-dimensional data. Applying this technique to ANES data from 2000 to 2008 (Figure 5) does not reveal any striking structure to the data, though it should be noted that only 13.9% of the total variance of the dataset is explained by the three principal components. It is possible that better separation between voters and non-voters exists in higher-dimensions where more variance is preserved, however this is not easily visualized. Nonetheless, there appear to be regions where voters are much more abundant as well as regions where voters and non-voters overlap but with greater concentrations of non-voters.

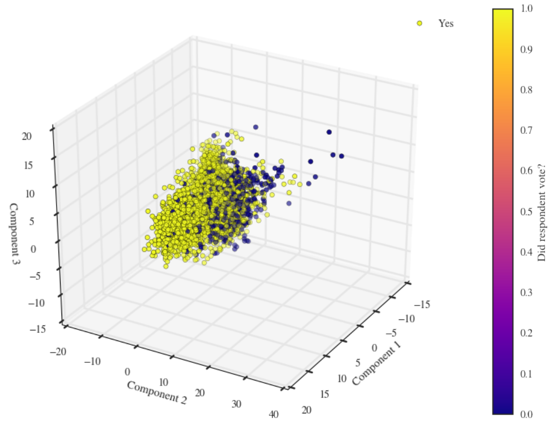

Figure 5: **Projecting respondent data onto a three-dimensional subspace using PCA**. The absence of obvious clustering or separation in the data shows how similar voters and non-voters can appear.

t-Distributed stochastic neighbor embedding (t-SNE), a non-linear dimensionality reduction algorithm, is optimized for compressing high-dimensional data to two or three dimensions and preserves more information than PCA, though it too shows similar results (Figure 6).

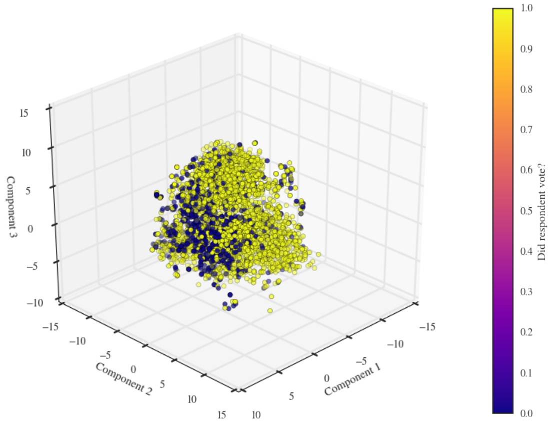

Figure 6: **Embedding respondent data onto a three-dimensional subspace using t-SNE**. Like the PCA visualization, the absence of obvious clustering or separation in the data shows how similar voters and non-voters can appear. Yet, the distribution of voters and non-voters is not completely random or uniform either.

3.3 Binary Features

The dataset contains 49 binary features, of which the 15 most correlated with vote participation are displayed in Table 3. Correlation is a more useful metric for determining the sole predictive power of a feature than the difference in the non-voter proportions between a feature's two categories because it accounts for situations where the number of respondents belonging to each category may be unbalanced. On the other hand, the difference in non-voter proportions is a better metric for determining how useful a feature may be in conjunction with other features.

For example, the proportion of respondents who did not donate money to a party or campaign and did not vote is 0.225 greater than the proportion of respondents who did donate money and did not vote. However, whether the respondent donated is ranked 15 by correlation because only a very small proportion of respondents actually donated to a party or campaign.

Rank	Feature	Correlation	Difference NV Proportions
1	Does respondent care who wins the presidential election?	0.362617	0.399264
2	Was respondent contacted by any major party?	0.255805	0.220867
3	Did respondent discuss politics with family or friends?	0.242067	0.290371
4	Does respondent dislike anything about the Republican party?	0.231964	0.199910
5	Did respondent try to influence the vote of others?	0.227229	0.193851
6	Does respondent believe there is a major difference between the two major parties?	0.217943	0.219666
7	Does respondent dislike anything about the Democratic Party?	0.216516	0.186011
8	Does respondent like anything about the Republican party?	0.204859	0.176908
9	Does respondent know which party controls the House majority?	0.200864	0.170431
10	Does respondent own a home?	0.197914	0.176171
11	Was respondent contacted by the Democratic party?	0.196247	0.181030
12	Does respondent like anything about the Democratic Party?	0.186178	0.162504
13	Was respondent contacted by the Republican party?	0.179923	0.178119
14	Did respondent display a political button or sticker during the campaign?	0.178994	0.194679
15	Did respondent donate money to a party or campaign?	0.170604	0.225182

Table 3: **Binary features most correlated with vote participation**. The features are displayed in descending order by absolute value of Pearson correlation coefficient. The third column shows the absolute value of the difference in non-voter proportions for that feature's two categories.

That whether or not the respondent cares who wins the presidential election is the most correlated with whether or not he will vote is hardly surprising. More interesting, however, is that contact from a major party is the second most correlated, suggesting that outreach efforts may be important to turning out favorable voters. Additionally, of these fifteen features, only one does not correspond to a question about politics. Rather, it corresponds to a demographic question: "Does the respondent own a home?", which if he does not, he is 17.6% more likely to be a non-voter. Also notable is how disliking anything about one of the major parties (rank 4 and 7) correlates to vote participation more than liking anything about one of the major parties (rank 8 and 12).

Examining the binary features least correlated with vote participation yields the most surprising results. Whether the respondent approves of the President's job performance is entirely uncorrelated with whether he will vote, perhaps because the President is so ubiquitous in contemporary culture that nearly everyone has an opinion about him, non-voters included. Whether the respondent believes the presidential race in his state will be close is also surprisingly uncorrelated with vote participation, defying the conventional wisdom that perceived closeness boosts voter turnout.

Rank	Feature	Correlation	Difference NV Proportions
49	Does respondent approve of the President's job performance?	0.000696	0.000612
48	Does respondent believe LGBT should be allowed to adopt children?	0.001901	0.001645
47	Are there more than one candidates in respondent's house race?	0.002436	0.003116
46	Does respondent believe presidential race in his state will be close?	0.012238	0.011754
45	Are respondent's parents native-born?	0.038676	0.047312
44	Is religion important to respondent?	0.052240	0.052466
43	Is anyone in respondent's household a member of a union?	0.054508	0.067455
42	Has the President made respondent feel afraid?	0.059212	0.050596
41	Is the respondent from the south?	0.066156	0.057289
40	Has the Democratic presidential candidate made respondent feel hopeful?	0.069300	0.059564

Table 4: **Binary features least correlated with vote participation**. The features are displayed in ascending order by absolute value of Pearson correlation coefficient.

Of all binary features, the highest ranked are only weakly correlated with vote participation, with the remaining showing either a very weak or no correlation at all (Figure 7.a). Thus, no binary feature by itself is effective at predicting vote participation. Examining the difference in non-voter proportions shows that all but two features have differences below 0.25 (Figure 7.b), so the overall predictive power of binary features is rather limited.

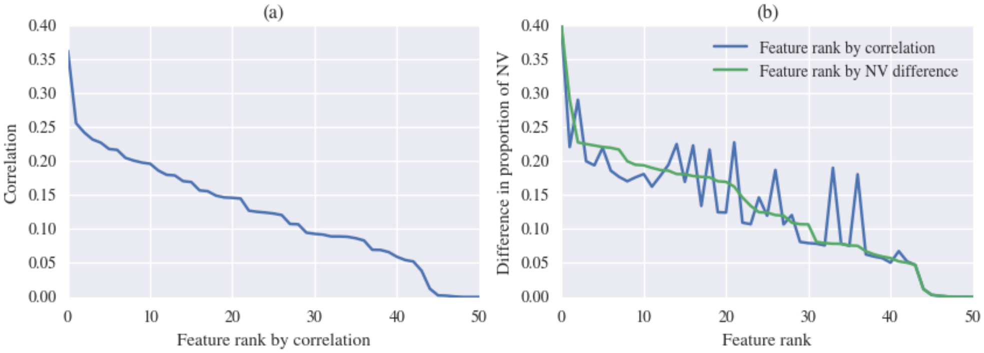

Figure 7: **Correlation and feature rank**. (a) All correlation coefficients of binary features with vote participation are quite low. (b) Most differences in non-voter proportions are also quite low. The peaks and valleys produced from ranking the difference of non-voter proportions by their corresponding correlations shows how features with high differences can be significantly less correlated with vote participation than features with lower differences.

3.4 Continuous and Quasi-Continuous Features

The dataset contains only a handful of continuous features, of which only age happens to be significant. Conventional wisdom says that young people are low-propensity voters, and the data confirms this, with the median non-voter age about 10 years lower than median voter age (Figure 8.a), a trend that holds true for each election year (Figure 8.b).

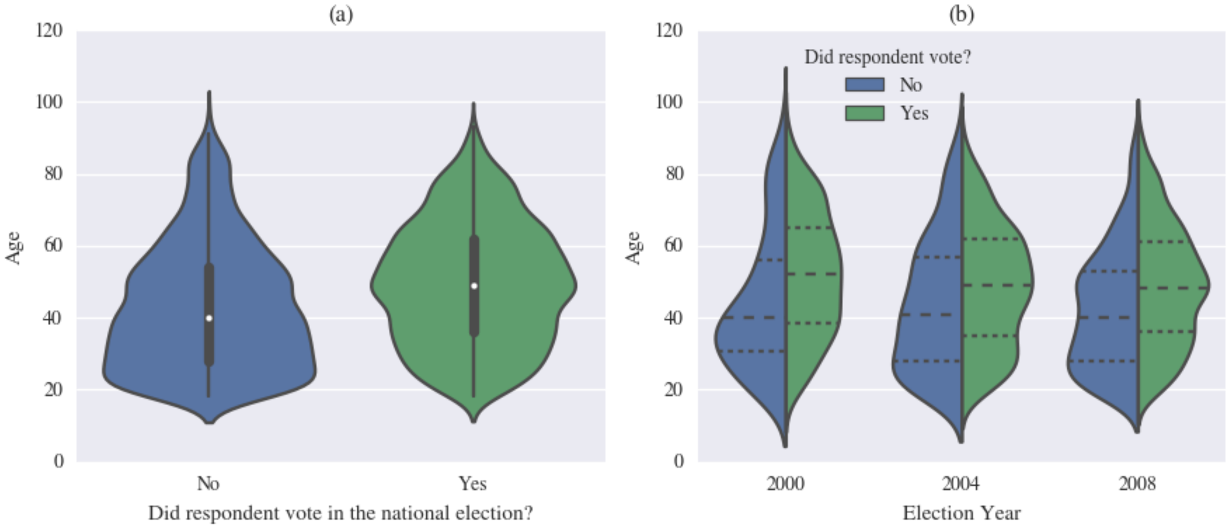

Figure 8: **Age distribution of voters and non-voters**

Quasi-continuous features are far more numerous and consist exclusively of *index* features and *thermometer* features from which they are built. Table 5 displays all pairs of features with correlations greater than or equal to 0.6. Because *index* features are derived from *thermometer* features, it is natural that they are highly correlated with their base features. Such a strong correlation, however, is also indicative of redundant information which must be addressed during feature selection. Of the correlated pairs that do not involve *index* features, the vast majority involve high-profile politicians or their political parties. However, Table 5 is more notable for the features not present, for example, interest groups such as big business and labor unions, and groups of people, such as illegal aliens, LGBT, and Christian fundamentalists.

In fact, Table 6 displays traditionally anti-correlated feature pairs and their correlation coefficients. One expects in a country as seemingly politically polarized as the United States to find significant negative correlations in such feature pairs, but while it is unsurprising that the Democratic and Republican presidential candidates form the most anti-correlated pair, the negative correlation is only moderate. More significantly, the correlation is signifcantly weaker for the major political parties themselves, and weaker even still, to the point of statistical insignificance, for the major political ideologies (liberals vs. conservatives). This supports the theory that politics is far less about issues or even ideology and far more about what some have called identity and tribalism. This notion is further supported by the striking degree to which senate and house candidates are uncorrelated, suggesting that many Americans may not be ideological but nonetheless become temporarily polarized during presidential election cycles.

Figure 9 shows the distributions of all 39 quasi-continuous features. Most distributions are roughly normal or skew normal, but a few, namely racial groups, the military, and middle class, are decidedly one-sided, with the vast majority of respondents indicating some degree of positive sentiment. Also interesting is the number of subjects for which a strikingly large number of respondents have no opinion, resulting in spikes in the middle of many distributions. These spikes are most prominent in distributions for house candidates and groups associated with social issues (Christian fundamentalists, LGBT, feminists).

Rank	Feature 1	Feature 2	Correlation Coefficient
1	Republican Presidential Candidate	Index: Presidential Candidates	-0.85
2	Republican VP Candidate	Index: VP Candidates	-0.85
3	Democratic Presidential Candidate	Index: Presidential Candidates	0.84
4	Index: VP Candidates	Index: Presidential Candidates	0.77
5	Challenger House Candidate	Index: Democratic House Candidate	0.77
6	Democratic VP Candidate	Index: VP Candidates	0.76
7	Republican Presidential Candidate	Republican Party	0.73
8	US Vice President	US President	0.77
9	Democratic Presidential Candidate	Democratic Party	0.71
10	Democratic Presidential Candidate	Index: VP Candidates	0.70
11	Republican VP Candidate	Republican Presidential Candidate	0.70
12	Republican Presidential Candidate	Index: VP Candidates	-0.68
13	Democratic VP Candidate	Democratic Presidential Candidate	0.67
14	Index: Presidential Candidates	Republican Party	-0.67
15	Index: Republican VP Candidate	Index: Presidential Candidates	-0.66
16	Republican VP Candidate	Republican Party	0.66
17	The Federal Government	Congress	0.66
18	Republican VP Candidate	Index: VP Candidates	0.65
19	Hillary Clinton	Democratic Party	0.66
20	Asian Americans	Hispanics	0.65
21	US Vice President	Republican VP Candidate	0.65
22	Index: VP Candidates	Republican Party	-0.64
23	Index: Presidential Candidates	Democratic Party	0.64
24	US President	Republican Presidential Candidate	0.62
25	Incumbent House Candidate	Republican House Candidate	0.62
26	Incumbent House Candidate	Democratic House Candidate	0.62
27	Hillary Clinton	Democratic Presidential Candidate	0.60
28	Democratic VP Candidate	Democratic Party	0.60

Table 5: **Most correlated quasi-continuous features**. The feature pairs are displayed in descending order by absolute value of Pearson correlation coefficient.

Feature 1	Feature 2	Correlation Coefficient
Democratic Presidential Candidate	Republican Presidential Candidate	-0.617542
Democratic VP Candidate	Republican VP Candidate	-0.530565
Democratic Party	Republican Party	-0.469287
Liberals	Conservatives	-0.306816
Democratic Senate Candidate	Republican Senate Candidate	-0.254249
LGBT	Christian Fundamentalists	-0.230465
Democratic House Candidate	Republican House Candidate	-0.159287
Big Business	Labor Unions	0.015196

Table 6: **Expected anti-correlated features**. The Pearson correlation coefficients here are calculated using sampling weights.

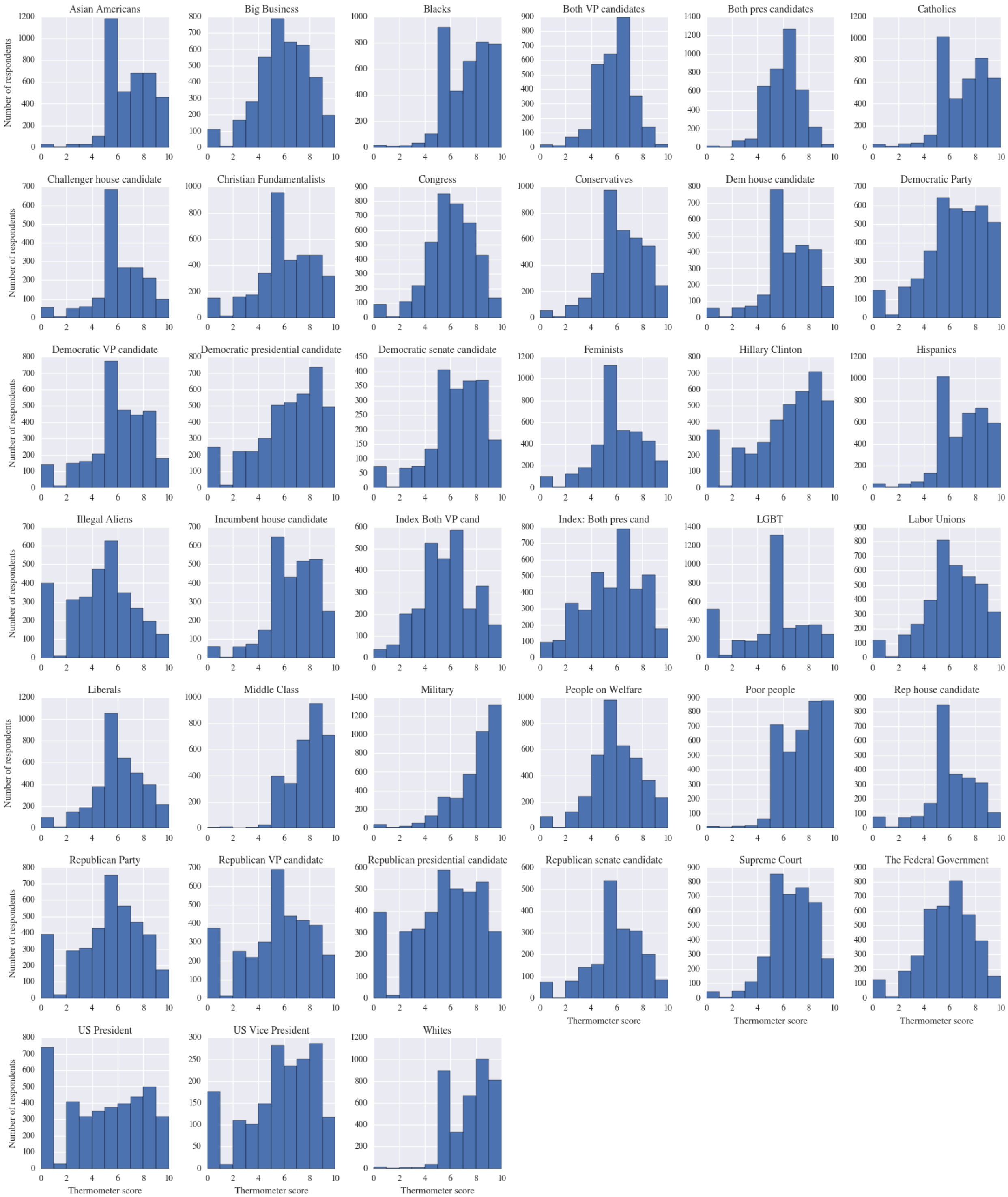

Figure 9: **Histograms of quasi-continuous features**.

Figure 10 contains the same features but displays the proportion of non-voters within each bin. For example, non-voters comprise 0.3 of all respondents who answered 5 for *Catholics*. Examining the features from this angle, two interesting patterns emerge:

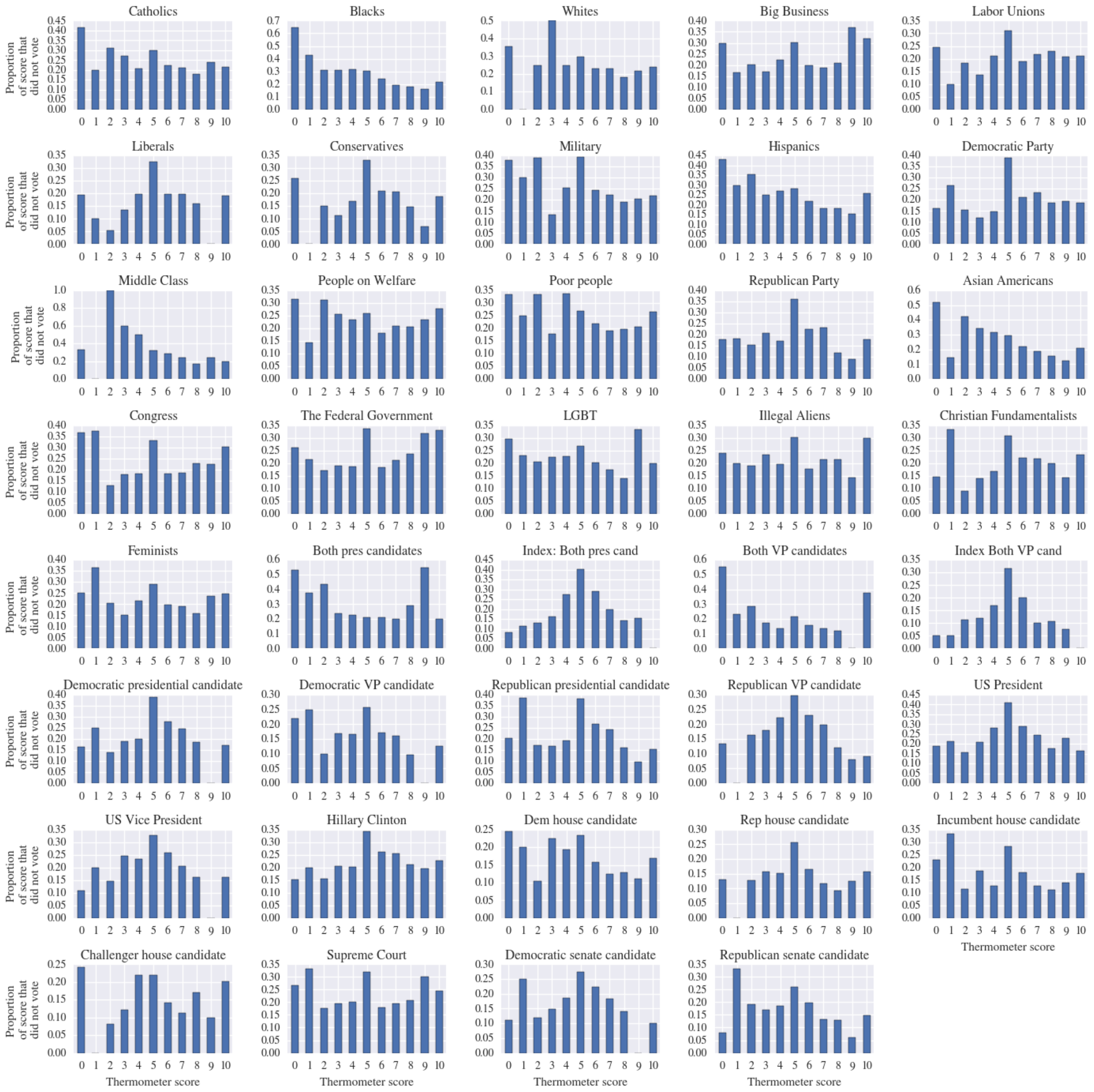

Figure 10: **Proportion of non-voters by bin**

- The more negative the respondent's sentiment toward racial minorities, the more likely the respondent is a non-voter, as can be seen by the negative linear relationship between non-voter proportions and *thermometer* score for *Blacks*, *Asians*, and *Hispanics*. However, this does not mean these features are useful for prediction because only a small segment of respondents display negative sentiment towards minorities.

- The more neutral the respondent's sentiment towards prominent politicians, the more likely the respondent is a non-voter. Features like *Republican VP Candidate* show how respondents who answer with *thermometer* scores closer to 0 or 10 have far lower non-voter proportions than those with scores closer to 5.

3.5 Ordinal and Derived Binary Features

The 22 *scale* features correspond to how a respondent places a subject along an ideological spectrum.

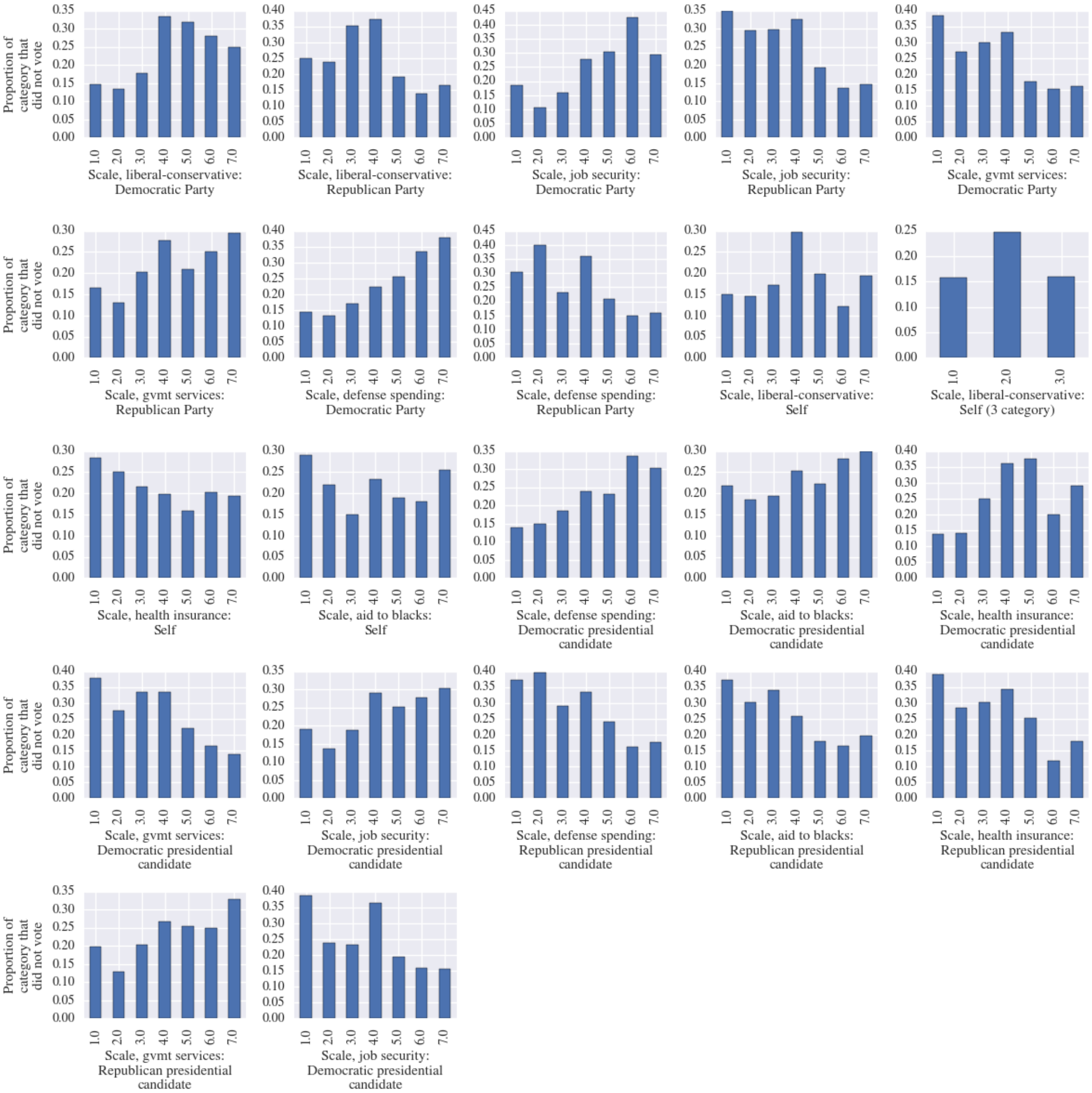

Figure 11: **Proportion of non-voters per ordinal response in *scale* features**. The corresponding scales are:

- Liberal-conservative: [1 - Extremely liberal, 7 - Extremely conservative]
- Job security: [1 - Gvmt should guarantee everyone a good job and standard of living, 7 - Gvmt should let each person get ahead on his own]
- Government services: [1 - Gvmt should provide fewer services, 7 - Gvmt should provide more services]
- Defense: [1 - Gvmt should increase military spending, 7 - Gvmt should decrease military spending]
- Health insurance: [1 - Gvmt should provide insurance plan, 7 - Insurance plans should be private]
- Aid to blacks: [1 - Gvmt should help more, 7 - Blacks should help themselves]

Scale questions in which a political party or politican serve as the subject illuminate the respondent's perceptions of the subject's ideological position on an issue, but they also double as a measure of political knowledge: does the respondent know what a politician's beliefs or a party's platforms are? In many cases, the more incorrect the respondent's answer, the more likely he is to be a non-voter. For example, Democrats generally believe that military spending should be reduced, and in *Defense spending, Democratic party*, the non-voter proportion increases for responses further away from this position on the scale. Likewise the Republican party is in favor of private insurance plans over government-provided insurance, so in *Health insurance, Republican presidential candidate*, respondents who placed the candidate on the wrong side of the spectrum have greater non-voter proportions than those who did not.

Another class of ordinal features correspond to questions about spending preferences on various federal programs (Figure 12). Those who tend to stay home on election day also happen to support Democratic policy priorities like welfare, child care, education, and environmental protection. But while these turnout margins can make all the difference in an election, the differences in non-voter proportions are not large enough to meaningfully predict vote particiaption.

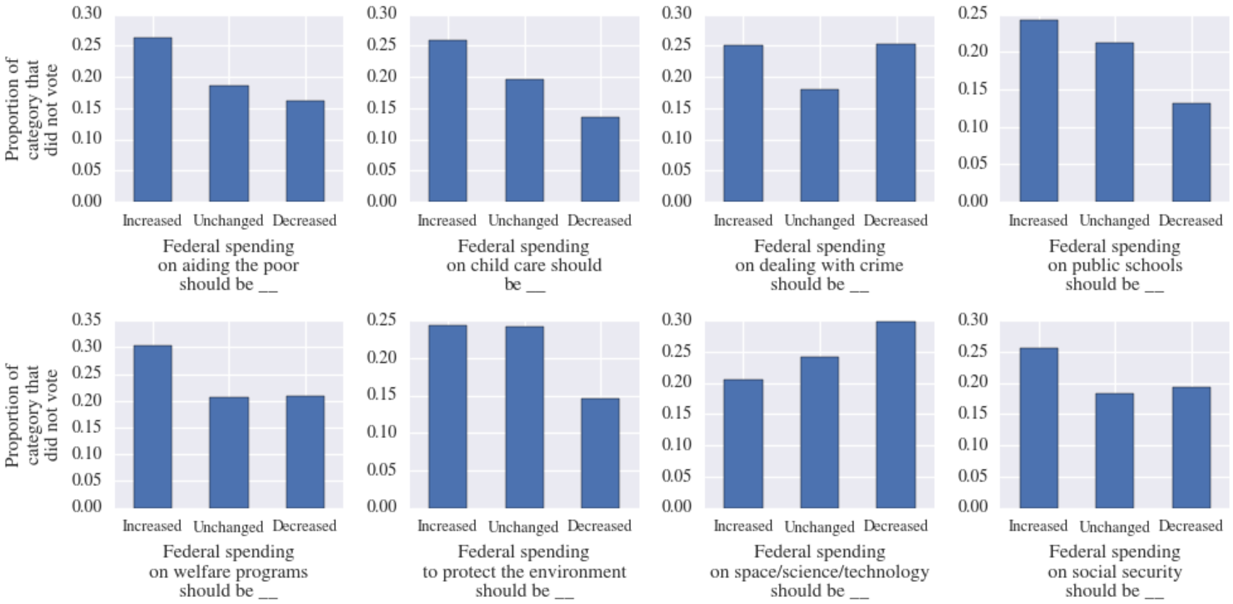

Figure 12: **Proportion of non-voters per spending preference**

Of the remaining ordinal features, those with the most interesting relationships with vote participation are displayed in Figure 13. Demographic characteristics like higher levels of education, frequency of church attendance, and social class are correlated with increased vote participation, as are ideological characteristics like higher levels of partisanship and interest in the election. Notably, the extent to which the respondent believes elections result in responsive government is also correlated with likelihood to vote, though the difference in proportions of non-voters is less significant than in more traditional features like strength of partisanship.

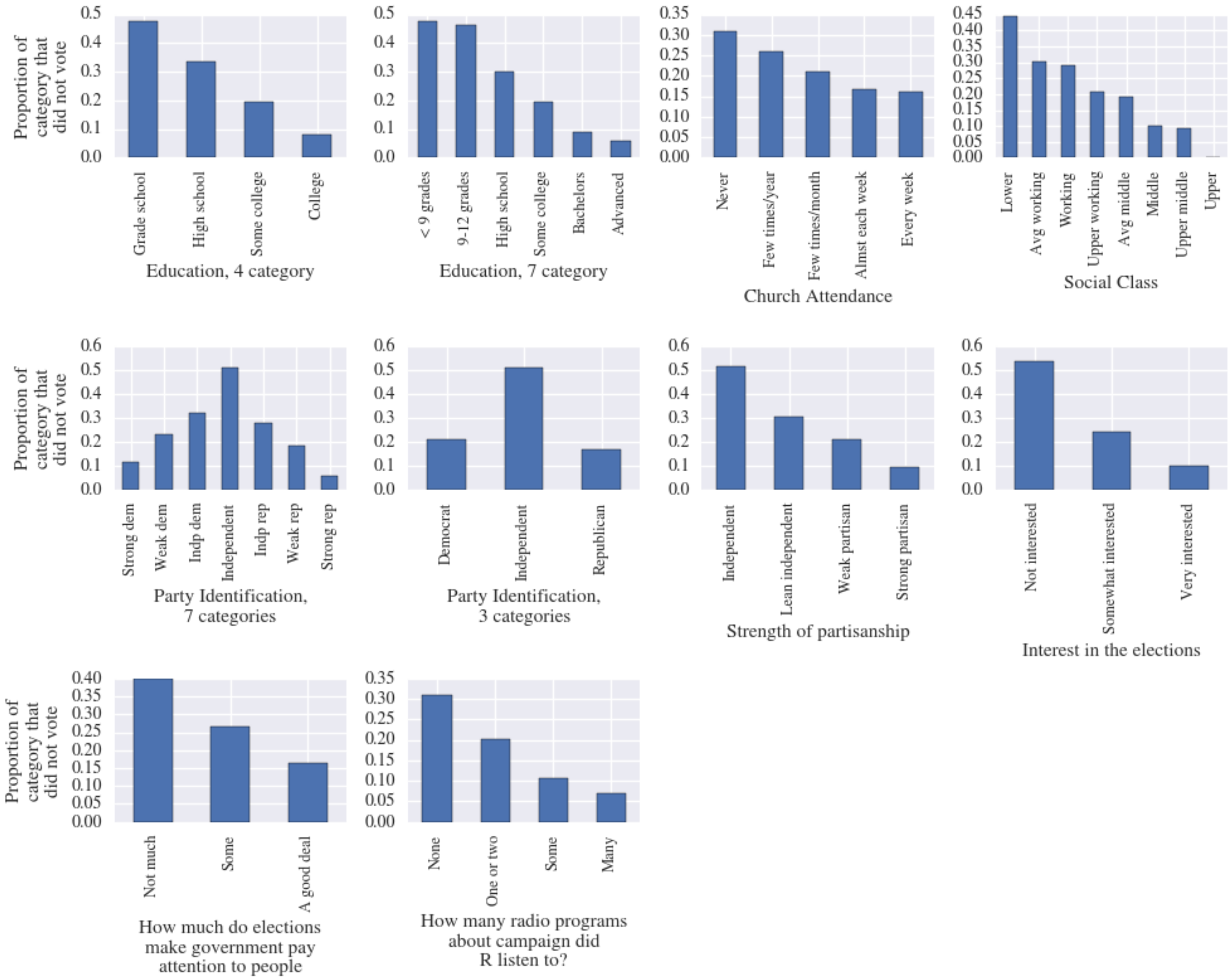

Figure 13: **Proportion of non-voters per response for miscellaneous ordinal features**

Most ordinal features include a non-orderable response choice for respondents who wish to answer ”don't know” for a given question. Given that in many features, the proportion of non-voters is highest for neutral responses, it is unsurprising that ”don't know” responses contain similarly high proportions of non-voters, as shown in Figure 14. ”Don't know” responses for each relevant ordinal feature are encoded into a new binary feature, and while the difference in non-voter proportions for these derived features is significantly higher than that of other binary features in the dataset, it is important to note that the number of ”don't know” responses for each feature is usually quite small (less than 200 responses).

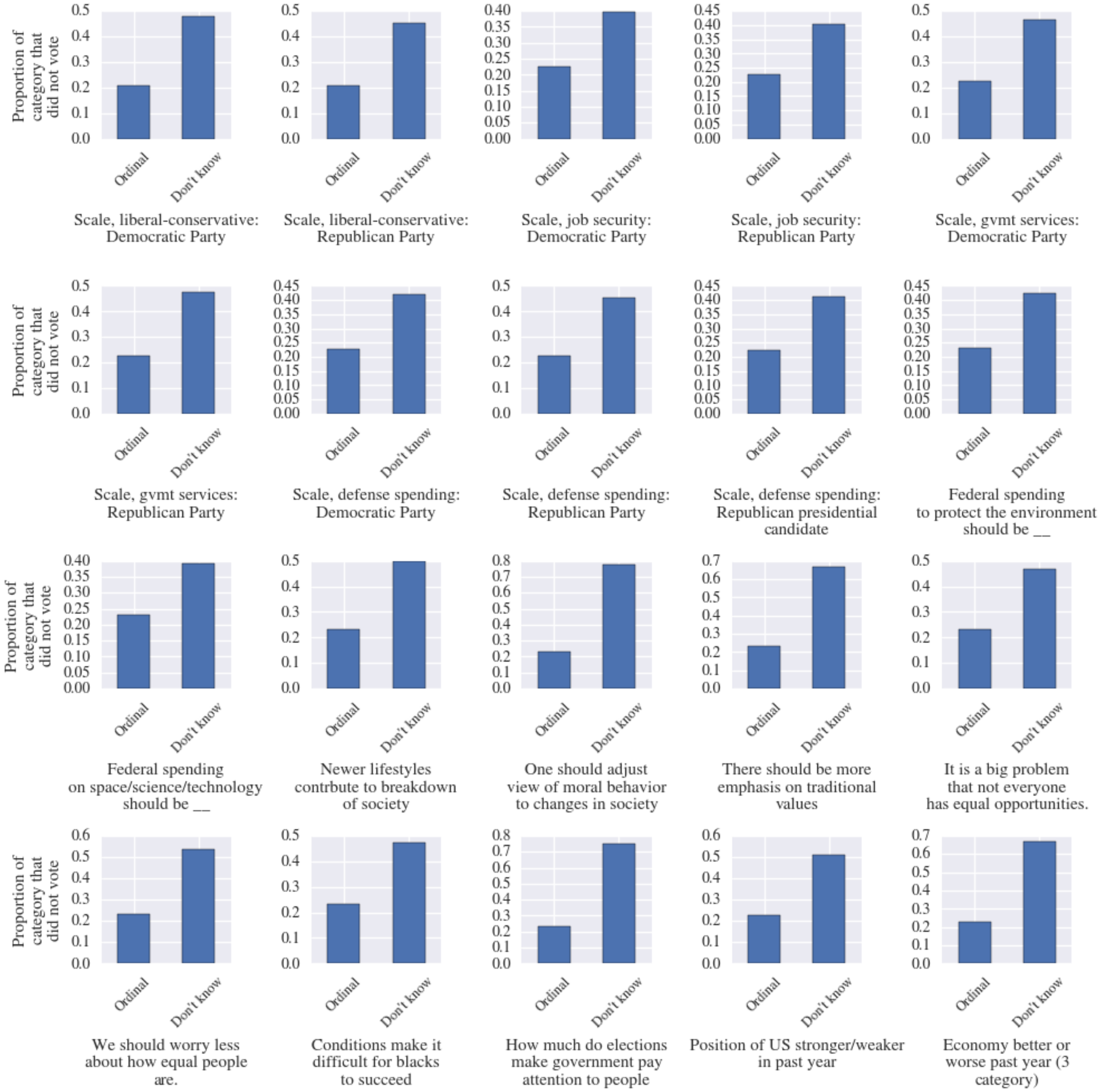

Figure 14: **Respondents who answer ”don't know” have higher non-voter proportions**

Figure 15 shows how a respondent's likelihood to not vote increases with each additional ”don't know” response, up to twelve such responses. Beyond twelve, however, the number of respondents per response total is too small for meaningful interpretations.

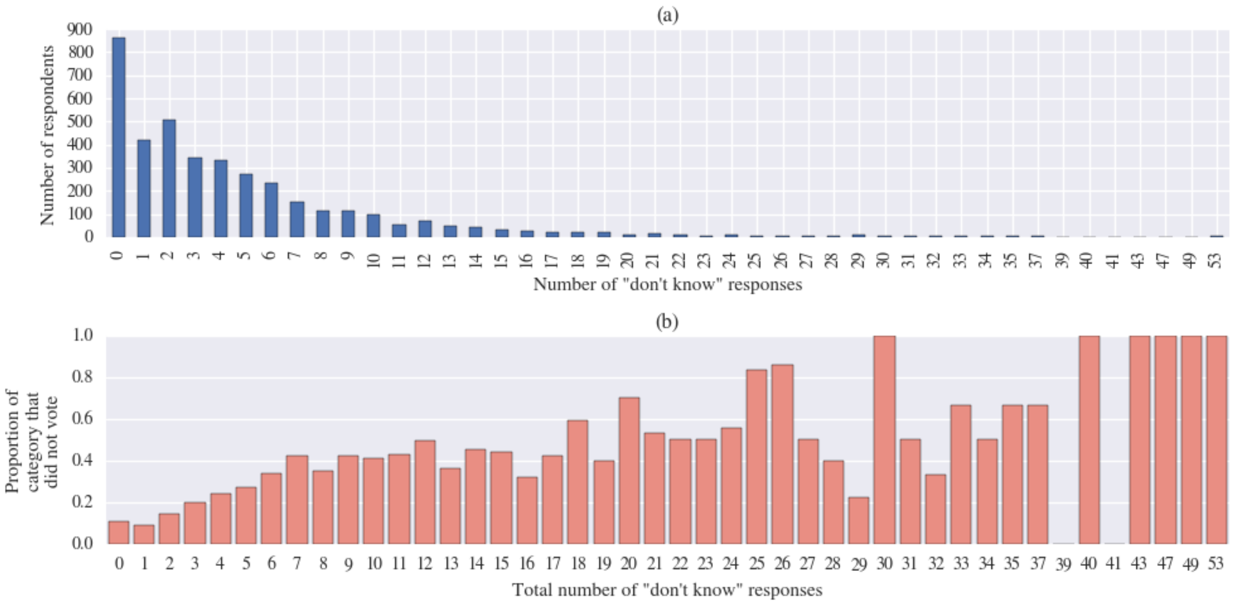

Figure 15: **Non-voter proportions by total number of ”don’t know” responses**

3.6 Categorical Features

Of the categorical features, three reveal interesting insights about vote participation. Figure 16 shows that whites and blacks, the largest and second largest racial groups, respectively, have the lowest non-voting rates, while Hispanics, the third largest racial group, have the highest non-voting rate. Asians, Pacific Islanders, and Native Americans do not fare much better than Hispanics, however they are a much smaller proportion of the voting eligible population.

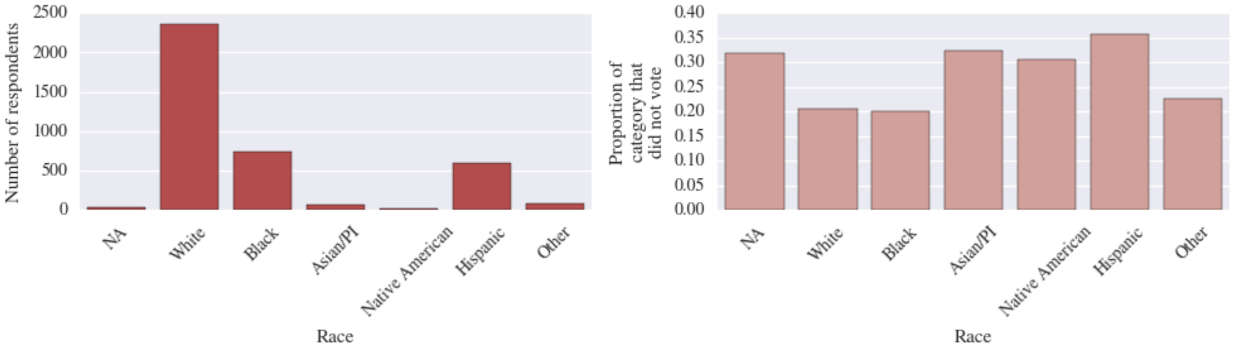

Figure 16: **Non-voter proportions by ethnicity**

Figure 17 shows that Jewish people have a drastically lower non-voting rate than any other religious group. Because Jews are also the most affluent religious group in the US[4], this supports previous data

[4] Masci, David. ”How Income Varies among U.S. Religious Groups.” *Pew Research Center*. N.p., 11 Oct. 2016. Web. 30 Apr. 2017.

showing that higher education levels and social class correlate with lower non-voting rates. Also notable is that atheists and other religions, as a group, have the highest non-voting rates, supporting previous data correlating increased church attendance with increased vote participation.

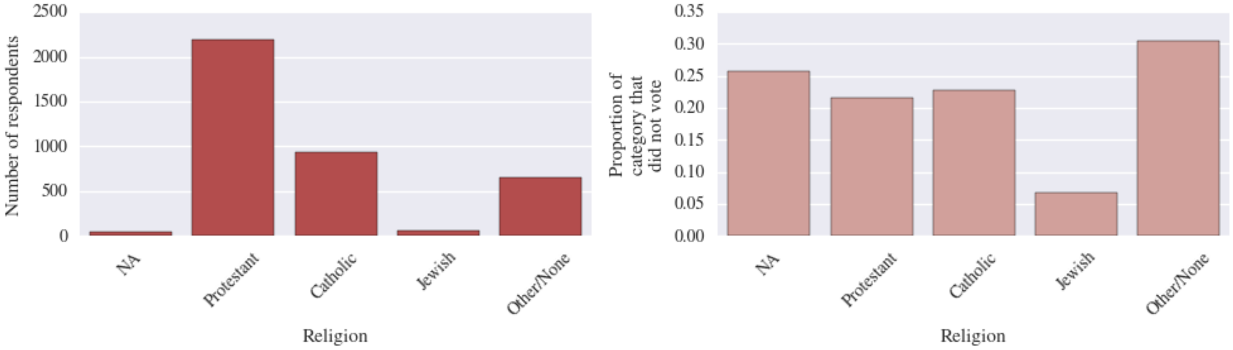

Figure 17: **Non-voter proportions by religion**

Finally, vote intention is the most important indicator of vote participation, with almost 0.9 of all respondents who answered ”will not vote” staying home on election day. However, vote intention is far from a perfect predictor, as only roughly 500 respondents indicated they do not plan to vote out of a total of 903 non-voters.

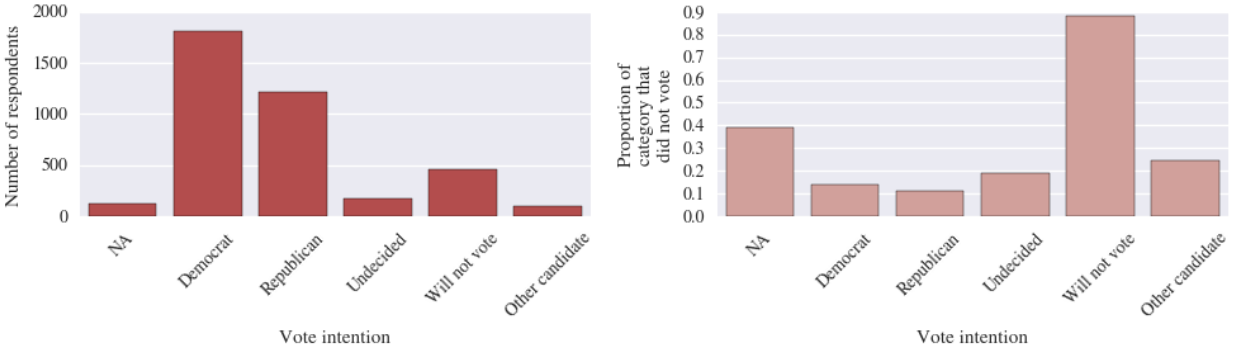

Figure 18: **Non-voter proportions by vote intention**

4 Feature Selection

Because the number of respondents increases drastically with each following election year, the amount of training data available for each target year is dramatically different. Thus, a feature set suitable for a target year such as 2012 for which a large amount of training data from preceeding years is available may contain an excess of features that causes overfitting for a target year such as 2004 for which only a few hundred respondents from 2000 comprise the training data. Testing across target years, then, requires an automated process for feature selection.

Such a process is also helpful for selecting features from various combinations of feature set modifications to find an optimal combination. Additionally, ideal feature sets vary not only with respect to the amount of training data available, but also with respect to the model used to make predictions. This section mainly explores the performance of four different models, logistic regression, adaptive boosting, Bernoulli Naive

Bayes, and support vector machines, each of which is tested across various feature sets produced from different combinations of feature set modifications and across each target year.

4.1 Baseline Performance

Prior to applying a model, two additional steps are performed on the data: missing value imputation and normalization. Missing values for a given feature are replaced by the median value of that feature, then the distribution for all continuous and ordinal features are normalized to the standard normal distribution. These two preprocessing steps are always applied prior to any application of a predictive model.

Given the unbalanced nature of the dataset, where far more respondents voted than those who did not, the F1-score is a more useful measure of predictive performance than accuracy[5]. Table 7 shows baseline F1-scores for six initial models obtained from performing 5-fold cross-validation[6] on each training set, with non-voters treated as positive cases.

	Years included in training set								
	2000			2000-2004			2000-2008		
	Mean	Std.	Range	Mean	Std.	Range	Mean	Std.	Range
Logistic Regression	0.549	0.029	0.081	0.613	0.032	0.081	0.625	0.031	0.080
Adaptive Boosting	0.625	0.085	0.253	0.619	0.031	0.075	0.629	0.025	0.067
Bernoulli Naive Bayes	0.633	0.0368	0.103	0.634	0.008	0.023	0.613	0.074	0.027
Support Vector Machine	0.533	0.076	0.212	0.568	0.050	0.142	0.574	0.033	0.084
Guassian Naive Bayes	0.479	0.043	0.117	0.574	0.021	0.052	0.550	0.072	0.189
Random Forest	0.426	0.191	0.523	0.518	0.031	0.092	0.524	0.086	0.032

Table 7: **Baseline F1-scores for each model across training sets**

With the exception of one case, Gaussian Naive Bayes and random forests consistently perform worse than the other four models and are consequently excluded from further testing. Of the remaining models, Bernoulli Naive Bayes and adaptive boosting are most resilient to large changes in training set size, though adaptive boosting becomes less stable with smaller training sets (as observed by the increases in standard deviation). On the other hand, logistic regression and support vector machines perform noticeably worse as training set size decreases.

4.2 Recursive Feature Elimination

Recursive feature elimination (RFE) with logistic regression is used to address the challenge of selecting an optimal feature set given a particular training set. In RFE, a logistic regression model is fit on the training data and a 3-fold cross-validated F1-score is produced for that feature set. Each feature is assigned a coefficient, and the feature with the smallest absolute coefficient is removed from the training set. The procedure is repeated until all features are exhausted, and the feature set resulting in the highest cross-validated F1-score is then chosen.

Table 8 shows the improved F1-scores for each model obtained from performing 5-fold cross-validation on training sets with features pruned using RFE. The process selects not only the optimal features, but the optimal number of features commensurate with the training set size. Thus, while all training sets started with 366 features regardless of target year, only 17 features are selected for the smallest training set (2000) while 115 are selected for the largest (2000-2008).

[5] The F1-score is the harmonic mean of precision p and recall r, defined $F_1 = \frac{2pr}{p+r}$, where precision is the proportion of true positives out of all predicted positives and recall is the proportion of true positives out of all positives.

[6] K-fold cross-validation allows for models to be tested with only training data, using a randomly selected $\frac{1}{K}$ of the data for testing performance. This process is repeated K times for an average score.

	Years included in training set								
	2000			2000-2004			2000-2008		
Number of features selected	17			33			115		
	Mean	Std.	Range	Mean	Std.	Range	Mean	Std.	Range
Logistic Regression	0.642	0.047	0.130	0.611	0.042	0.123	0.670	0.024	0.060
Adaptive Boosting	0.666	0.023	0.061	0.626	0.036	0.093	0.639	0.025	0.075
Bernoulli Naive Bayes	0.688	0.041	0.104	0.625	0.017	0.048	0.637	0.030	0.092
Support Vector Machine	0.557	0.067	0.155	0.561	0.053	0.152	0.595	0.039	0.121

Table 8: **F1-scores for each model across training sets with recursively eliminated features**

4.3 Feature Modifications

A modest number of pre-RFE modifications to the feature set are considered to further improve the performance of the predictive models.

A. **Adding thermometer intensity**: For many thermometer features, the proportion of responses corresponding to non-voters decreases as the response category becomes less neutral. This represents a non-linear relationship between voting propensity and the feature in question that a logistic regression is not suited to exploit. Thus, a set of new features are derived by taking the absolute value of the difference between the respondent’s thermometer score and the neutral value corresponding to that feature.

B. **Adding ordinal intensity**: As with the case for thermometer features, for certain ordinal features, the proportion of responses corresponding to non-voters also decreases for less neutral response categories. Thus, a process similar to that used for thermometer intensity is applied to generate new intensity features from ordinal features.

C. **Adding total ”don’t know” responses**: For many features, the proportion of respondents that answered ”don’t know” who are non-voters is quite high. Furthermore, up to a certain limit, the more times a respondent answers ”don’t know,” the more likely the respondent is to be a non-voter. Thus, a feature that squares the sum of all ”don’t know” responses is generated. The squaring is necessary to avoid collinearity issues that may affect the stability of logistic regression.

D. **Dropping the first one-hot feature**: In the same way that a 2-categorical (binary) variable can be represented using 1 feature, a d-categorical variable can be represented using $d-1$ features following one-hot encoding. Therefore, the first feature of every one-hot encoding is removed to avoid collinearity issues when using logistic regression.

E. **Dropping correlated features**: Highly correlated features that contain redundant data may also present collinearity issues when using logistic regression, so it is necessary to consider removing features to break correlation clusters. The process involves identifying a cluster of correlated features (where each feature has a correlation of at least 0.85 with at least one other member of the cluster), and iteratively using the mutual information score[7] to remove the least predictive feature from the cluster until no feature is highly correlated with any other remaining feature in the cluster.

However, given the inherent instability of RFE, different combinations of the above modifications result in the selection of different feature sets and therefore different F1-scores. Furthermore, the best combination of modifications for one model may not necessarily yield equally desirable results for another. Therefore, every combination of modifications is run through RFE, with each resulting feature set tested with 5-fold cross-validation on four different models. This allows for the selection of an optimal feature set for each model.

[7] The mutual information between two discrete features X and Y is defined as $I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$ where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively. Conceptually, it measures how much knowing X reduces uncertainty about Y and vice versa.

	Years included in training set								
	2000			2000-2004			2000-2008		
	Mean F1	Pre-RFE	# feat	Mean F1	Pre-RFE	# feat	Mean F1	Pre-RFE	# feat
Logistic Regression	0.720	ACDE	16	0.717	AD	75	0.676	AE	122
Adaptive Boosting	0.698	ADE	20	0.694	ABC	61	0.663	ABCE	89
Bernoulli Naive Bayes	0.693	ADE	20	0.669	B	50	0.651	CD	83
Support Vector Machine	0.605	BCE	32	0.608	ABCE	111	0.600	ACDE	140

Table 9: **Optimal feature sets and resulting F1-scores**. This table shows the mean F1-score calculated through 5-fold cross validation, the feature modifications performed prior to selection using RFE, and the number of features selected for each model. The pre-RFE modifications follow the labels introduced at the beginning of subsection 4.3.

Despite the fact that RFE uses logistic regression to assign coefficients for feature elimination, logistic regression does not necessarily perform better than the other models when only RFE is used. However, Table 9 shows that following a more exhausive feature selection process, logistic regression consistently performs better than the other three models. Furthermore, all models show an improvement in performance.

Also notable is how the performance of logistic regression, adaptive boosting, and Bernoulli Naive Bayes degrades as the training data includes respondents from more years. This suggests that data from each election year has particular characteristics that do not generalize well to other election years and that a substantial reduction in performance can be expected when predicting on test data.

5 Parameter Tuning and Testing Models

Before evaluating the four models on the test data using their respectively optimized feature sets, optimal hyper-parameters are found applying a cross-validated grid search on the training data. For each model, the hyper-parameters considered are:

- **Logisitc regression**: regularization strength, regularization penalty norm, positive classification threshold
- **Adaptive boosting**: learning rate, number of estimators, positive classification threshold
- **Bernoulli Naive Bayes**: additive smoothing, positive classification threshold
- **Support vector machine with RBF kernel**: regularization strength, kernel coefficient, positive classification threshold

All models are capable of outputting a positive classification probability rather than the classification itself and default to classifying a respondent as a non-voter if his non-vote probability is greater than 0.5. The positive classification threshold parameter sets the probability threshold above which a respondent is classified as a non-voter. Table 10 shows that using tuned models and optimized thresholds offers significant improvements in predictive performance on the training data.

5.1 Testing Model Performance on Unseen Data

While the performance of tuned models on training data with optimized feature sets appears to be good, the generalizability of the models beyond the training data can only be assessed by testing with completely new data from the target election year. Table 11 shows that doing so yields much lower performance, especially

	Years included in training set					
	2000		2000-2004		2000-2008	
	Optimized threshold	F1	Optimized threshold	F1	Optimized threshold	F1
Logistic Regression	0.31	0.739	0.38	0.772	0.34	0.728
Adaptive Boosting	0.497	0.719	0.497	0.736	0.499	0.688
Bernoulli Naive Bayes	0.35	0.722	0.276	0.722	0.364	0.665
Support Vector Machine	0.297	0.827	0.262	0.894	0.211	0.909

Table 10: **F1-scores for training data following hyper-parameter tuning**

	Target year of test set								
	2004			2008			2012		
	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Logistic Regression	0.612	0.604	0.620	0.614	0.531	0.727	0.595	0.606	0.585
Adaptive Boosting	0.611	0.694	0.546	0.617	0.522	0.699	0.584	0.532	0.565
Bernoulli Naive Bayes	0.636	0.647	0.624	0.618	0.525	0.749	0.538	0.655	0.456
Support Vector Machine	0.659	0.703	0.620	0.618	0.518	0.764	0.542	0.448	0.687

Table 11: **Model performance on test data**. For each model and target year, the model's performance is scored using the F1-score, precision, and recall.

for 2012.

Such a decline in performance is not entirely unexpected, as the characteristics of an electorate change with each election year, and low-dimensional visualizations in exploratory data analysis previously revealed the challenges of separating voters from non-voters. In light of such conditions, F1-scores of around 0.6, while not excellent, remain decent.

Also notable is the fact that no individual model performs the best across all three election years; Berneoulli Naive Bayers and and support vector machines may perform better for 2004 and 2008, but they are worse compared to logistic regression and adaptive boosting for 2012. This suggests that there does not exist an optimal individual model for non-voter classification. Rather, the optimal model for any election year must be decided on a case by case basis.

5.2 Soft Voting Classifier

To achieve additional gains in testing performance, a soft voting classifier is used to leverage the strengths of each individual model. Not to be confused with the problem at hand (whether or not a respondent voted), a hard voting classifier is an ensemble method that combines the predictions of individual models by having them "vote" on how to classify a given data point. In soft voting, each model outputs probabilities that a data point belongs to a certain classification category, and the weighted average of these probabilities then determines how the voting classifier classifies that data point. Tuning the soft voting classifier involves finding the specific weights that produce the best results, which is achieved once again through a simple grid search process.

	Target year		
	2004	2008	2012
Training data F1-scores	0.726	0.765	0.784
Test data F1-scores	0.619	0.647	0.611

Table 12: **Soft voting classifier performance on training and test data**. For each target the year, the training data consists of all respondents from previous years no earlier than 2000.

Table 12 shows the performance of tuned and threshold-optimized soft voting classifiers on training and test data. For target years 2008 and 2012, using a such a classifier provides better results on the test data than using any individual model, while for 2004, support vector machine remains the best option. While these test data F1-scores appear not as good as hoped, they must be considered in light of the fact that a much higher proportion of respondents indicated that they voted than historical data would suggest, likely due to social desirability bias. The next section discusses how to account for the effects of social desirability bias when using predicted classification results to reduce polling errors.

6 Testing Polling Errors

In Section 3, exploratory data analysis showed that voter turnout for presidential election years on or after 2000 never dipped below 70% and sometimes came close to 80%, a much higher turnout rate than that reported by the widely respected United States Elections Project, which suggests voter turnout to be much closer to 60% for those years.

If turnout data from USEP is to be believed, this means that anywhere from 16% to 20% of respondents from each year of ANES survey may have claimed to have voted despite not actually voting. This also suggests that the ground truth of the ANES survey is unreliable and may explain why the classification models tested in Section 5 do not perform better when evaluated with this ground truth.

Given that the ground truth of the voting behavior of ANES survey respondents is problematic, the F1-score becomes a poor means of evaluating a model. If in fact, the main purpose of predicting whether or not a respondent will vote is to aid in obtaining a more accurate prediction about candidate preferences in the voting population (i.e. "What percent of voters will vote for candidate X?"), a better measure of a model's effectiveness is whether or not it can reduce the error of a survey-based inference relative to the official popular vote.

6.1 Survey-based Prediction

Using just the ANES survey data for a particular election year, one could infer the presidential candidate preferences of the voting population by discounting respondents who indicate a pre-election intent of not-voting and applying sampling weights on the remainder of respondents. All the necessary data is contained within the two features, *Sampling Weight* and *Pre-election Intent*. Let w and v be the sampling weight and vote intention of a respondent, respectively. Then w_i and v_i denote the sampling weight and vote intention of the i^{th} respondent, respectively. Let $n \in v$ represent the intention of the respondent not to vote. Then P_j, the proportion of voting respondents who voted for candidate j, as predicted using only survey data, can be defined:

$$P_j = \frac{\sum_{i=1}^{m} w_i \delta_{v_i j}}{\sum_{i=1}^{m} w_i (1 - \delta_{v_i n})} \text{ ,where } \delta_{v_i j} = \begin{cases} 0, & \text{if } v_i \neq j. \\ 1, & \text{if } v_i = j. \end{cases} \tag{1}$$

.

6.2 Model-based Prediction

Where a survey-based prediction discounts respondents who indicate their intention to not vote, a model-based prediction discounts respondents predicted to be non-voters. In addition to *Sampling Weight* and *Pre-election Intent*, a *Predicted Vote/Non-vote* feature is required. Let w be the sampling weight, v be the vote intention, and $z \in \{0, 1\}$ be the model-generated non-vote/vote prediction of a respondent, where 1 indicates a predicted ***non-vote***. Let $n \in v$ represent the intention of the respondent not to vote. Then w_i, v_i, and z_i denote those properties of the i^{th} respondent, and P_j', the proportion of voting respondents who voted for candidate j, as predicted using a predictive model, can be defined:

$$P_j' = \frac{\sum_{i=1}^{m} w_i \delta_{v_i j} (1 - z_i)}{\sum_{i=1}^{m} w_i (1 - \delta_{v_i n})(1 - z_i)} \text{ ,where } \delta_{v_i j} = \begin{cases} 0, & \text{if } v_i \neq j. \\ 1, & \text{if } v_i = j. \end{cases} \tag{2}$$

.
Although the ANES survey's ground truth of whether respondents voted may be unreliable, it is only the case for one class of categories. That is, one expects a respondent claiming to have voted despite non-voting to be far more common than a respondent claiming to have not voted despite actually voting. Furthermore, a respondent who voted despite claiming no intention to vote would have no recorded candidate preference and cannot be included in the prediction. Thus, one can assume the ground truth for those respondents labelled as non-voters to be reliable even if the same cannot be said for those labelled as voters.

Because of this assumption, a model that produces many false positives is preferred over a model that produces many false negatives. In other words, a model can be altered to classify as many true positives (accurately predicted non-voters) as it can even if it means the model might produce many false positives (voters misclassified as non-voters) because the false positives may likely be those respondents who claimed to have voted but did not actually vote.

Concretely, this means developing a model with a high recall, which can be obtained by lowering the threshold for a positive classification. That is, if the probability that a respondent is a non-voter (as determined by the model's algorithm) is above a certain threshold, that respondent will be classified as a non-voter. Thus, a lower threshold means more respondents will be classified as non-voters, resulting in an increase in both true and false positives.

After a model is trained from the target year's corresponding training data, its classification threshold is optimized to predict with a recall of 0.8, and the model is applied to the target year's data to generate vote/non-vote predictions. The recall value of 0.8 is admittedly arbitrary but based on the following intuition: requiring a model to correctly classify all positive cases (resulting in a recall of 1.0) may require a threshold so low as to be meaningless, yet a model with too low a recall may not have an inclusive enough picture of what constitutes a potential non-voter. The effects of requiring different levels of recall on polling error is examined at the end of this section.

Conceptually, lowering the threshold to increase recall is akin to instructing the model to broaden its search for non-voters. The higher the resulting recall, the more complete a model's picture of what a non-voter looks like. With this broader picture, the model also classifies respondents who claim to have voted but display traits far too similar to those who have not. In this way, the model accounts for the suspiciously high proportion of respondents claiming to have voted.

6.3 Polling Error Analysis

Because the soft voting classifers yielded the best F1-scores for each target year's respective training data, they are used to generate vote/non-vote classifications in the model-based prediction. Figure 19 shows that survey-based predictions using ANES data consistently overestimates the proportion of votes for the Democratic and third party candidates while underestimating that of the Republican candidate. By correcting down the share of votes for the Democratic and third party candidate and correctung up that for the Republican candidate, the model-based method consistently produces more accurate predictions. The one case where the model-based prediction results in a greater error is for Barack Obama, the Democratic candidate in 2012, for whom the survey-based prediction is already quite accurate. However, this deficiency is offset by the near-perfect prediction for the Republican share of the vote as well as the substantial improvement for the third party share.

That the model-based prediction consistently lowers the Democratic candidate's vote share while raising that of the Republican candidate strongly suggests that respondents who prefer the Democratic candidate have a lower voting propensity and reinforces findings from Section 3 showing that non-voters tend to favor Democratic spending priorities. In fact, Figure 13 from Section 3 shows that self-identified Democrats have a non-voting rate of 4% more than Republicans, unsurprising given that major components of the Democratic coalition, such as young adults, non-religous adults, and non-black minorities, are characterized by low vote participation. One might expect non-voters to skew less Democratic in elections where the Democratic wins the presidency, but based on the model's corrective tendencies, this behavior appears to have remained the same in 2008 and 2012 when Barack Obama, the Democratic candidate, won both presidential elections.

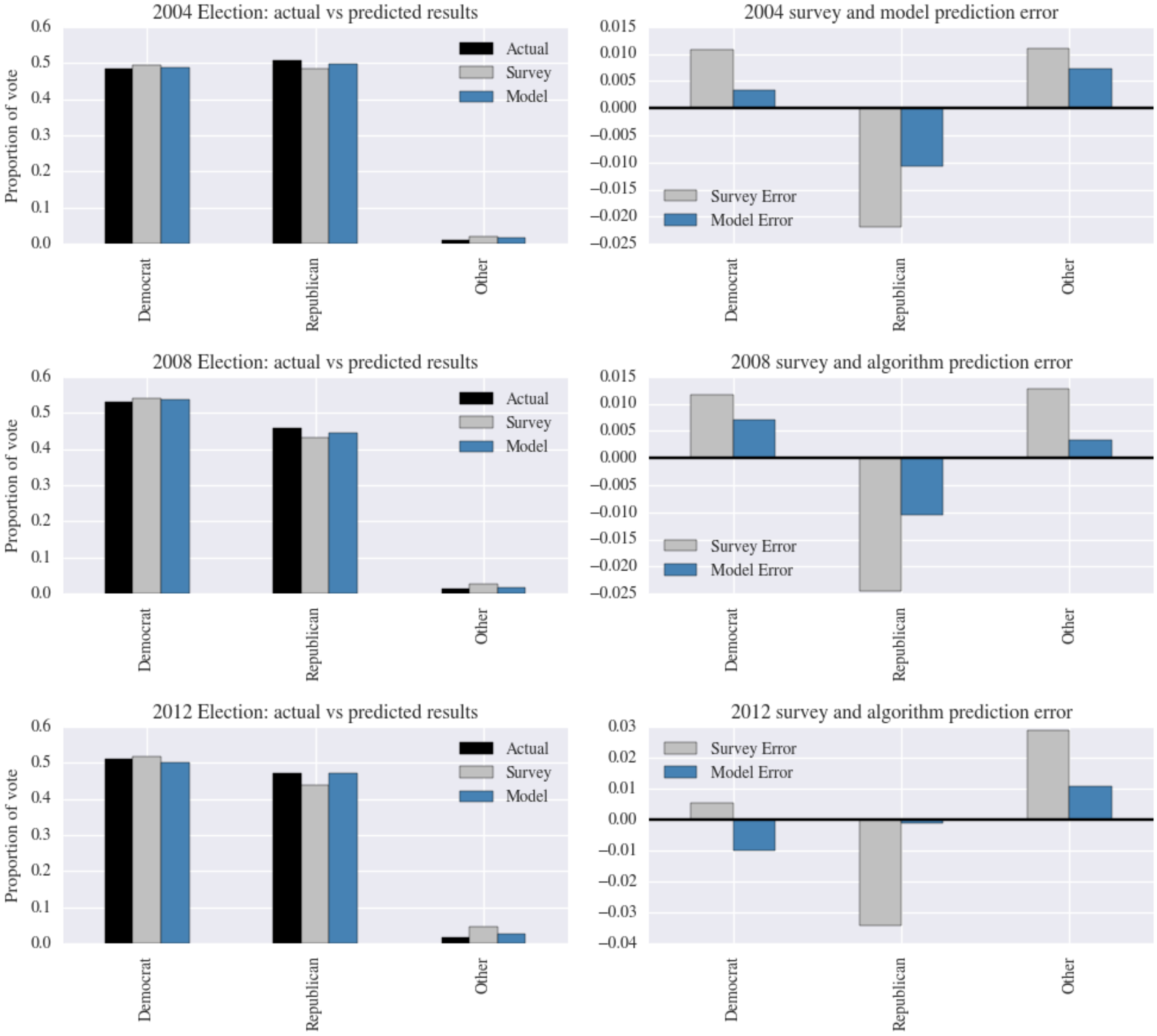

Figure 19: **Actual and predicted popular vote results and corresponding errors**. The left plots show the proportion of voters that voted for each presidential candidate according to actual popular vote data, a survey-based prediction, and a model-based prediction. The right plots show the errors of the two predictive methods measured against the actual popular vote.

Oftentimes correctly predicting the margin by which one candidate leads another is more important than predicting absolute proportions, and here the model-based predictions are also consistently more accurate than survey-based predictions (Figure 20). In particular, a survey-based prediction using the ANES data would have projected a narrow popular vote win for Democratic presidential candidate John Kerry in the 2004 election, and in fact most polls leading up to election day were projecting a similar result. However, filtering out predicted non-voters results in a model-based prediction that correctly projects a popular vote win for President George W. Bush, the Republican incumbent at the time.

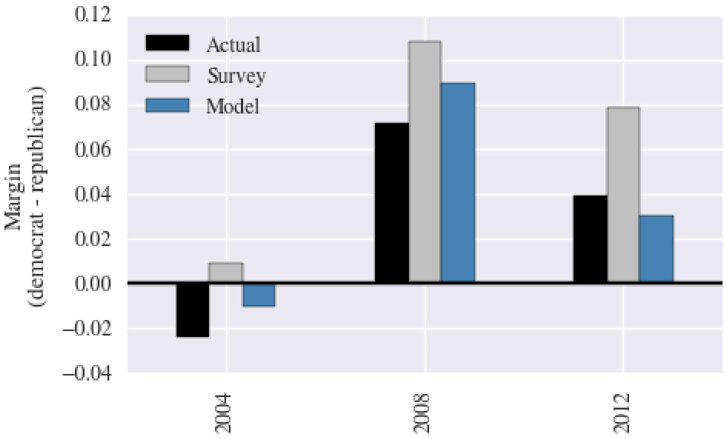

Figure 20: **Actual and predicted popular vote margins**. The margin is calculated as the proportion of votes cast for the Republican candidate substracted from the proportion of votes cast for the Democratic candidate.

6.4 Recall Analysis

While the results show that using a soft voting classifier to filter out likely non-voters produces superior results to relying on survey data alone, they come with the caveat that the model's positive classification threshold was altered to yield an arbitrary recall value of 0.8 on the training data. This begs the question of how sensitive model-based predictions are to different recall values. Figure 21 shows how the vote proportion predictions for each target year change with respect to recall. A recall of 0.0 corresponds to a positive classification threshold of 1.0. That is, only those respondents for whom the model assigns a non-voting probability of 1.0 are classified as non-voters and thus have their voting intent discounted from the vote prediction. But because the model never assigns a non-vote probability of 1.0 to any respondent, a recall of 0.0 essentially reproduces the survey-based prediction. For 2004, a training recall of 0.8 happens to be quite a good value as it leads to a prediction quite close to the Democratic candidate's actual vote share. Furthermore, the model's predictions are rather stable and reasonably accurate for a relatively large range of recall values. From recalls of 0.6 to 0.95, the model would have accurately projected the George W. Bush the winner of the popular vote despite survey data suggesting the contrary, and within this range, the model's predictions for the Democratic and Republican shares of the popular vote never deviates too far from the actual result.

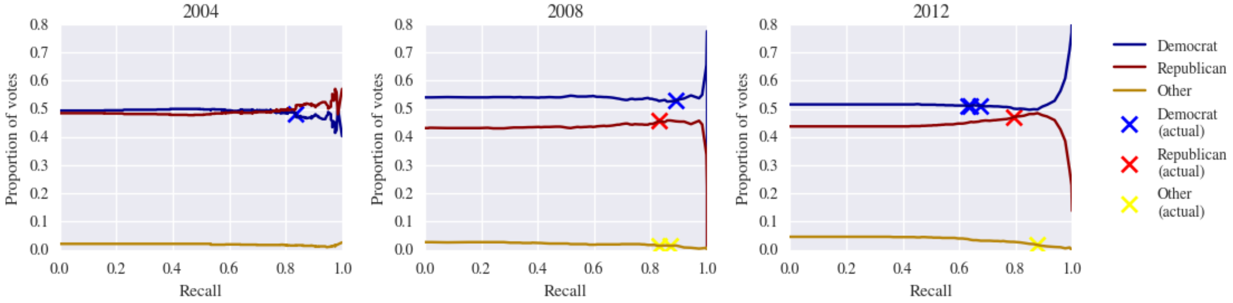

Figure 21: **Effect of training recall on predicted vote results**. The crosses indicate the recall and vote proportion for which a perfect prediction is made.

For predicting the 2008 popular vote, a training recall of greater than 0.6 is required for the effects of the model to be visible. Any training recall from within the range of 0.65 and 0.95 generates an improvement over the survey-based prediction, and once again, a training recall of 0.8 appears to work well. Notably, the recall value necessary for a correct prediction of the democratic share of the vote is quite close, within 0.07, to that for the republican share.

Of the three target years tested, the model's performance with respect to training recall is least stable for 2012. Whereas the models for previous presidential elections yield training recall ranges for which the predicted vote shares for democrats and republicans do not deviate too much, the model for 2012 has no such range. This is partly due to the fact that the most predictive feature in the training data for whether or not a respondent will vote, their vote intention, is much less predictive in 2012 (Figure 22).

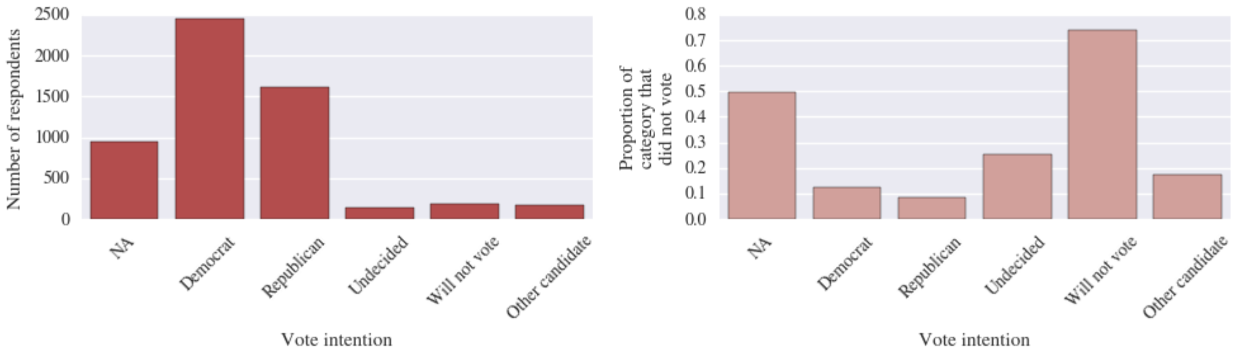

Figure 22: **Non-voter proportions by vote intention for 2012**. The proportion of respondents indicating they will note vote is much lower than in previous election years, and the non-vote proportion of respondents indicating they will not vote is lower than in previous years (Figure 18).

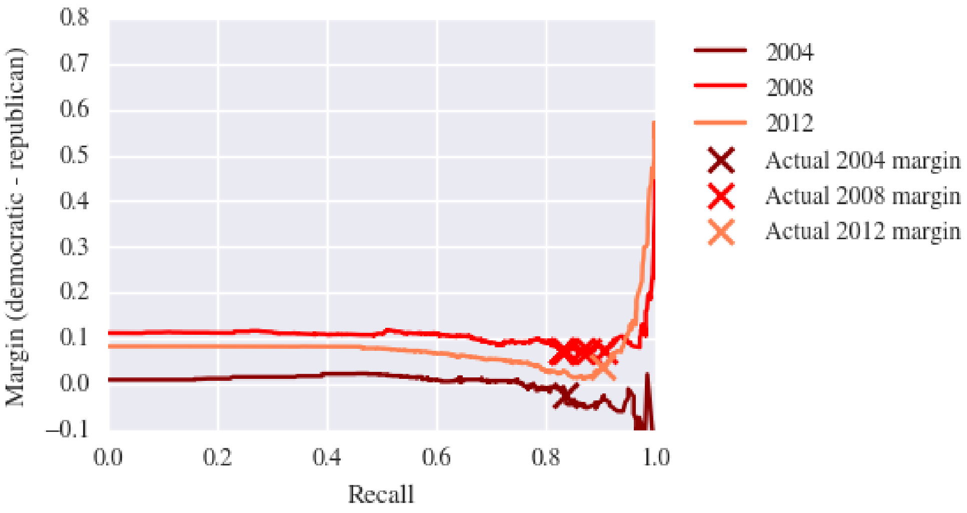

Figure 23: **Effect of training recall on predicted vote margin** The margin is calculated as the proportion of votes cast for the Republican candidate substracted from the proportion of votes cast for the Democratic candidate. The crosses indicate the recall and margin for which a perfect prediction is made.

Finally, examining how different values for training recall affect the popular vote margins shows that recall values between 0.6 and 0.9 reliably result in more accurate predictions of the vote margin for all three election years (Figure 23). Thus, while the favorable results presented earlier in this section relied on soft voting classifiers with positive classification thresholds modified to predict training data with a recall of 0.8, choosing a different recall value from within a relatively wide range of 0.6 to 0.9 would have produced improved predictions as well. This shows that using a modified soft voting classifier to filter out likely non-voters is an effective method for refining a poll's accuracy and is robust against varying assumptions about appropriate training recall conditions.

7 Conclusion

In investigating the effectiveness of using machine learning techniques to classify voters and non-voters, this project shows that while no single model stands out definitively as the most effective for all three presidential election years in question, models taken in combination via a soft voting classifier yields the highest F1-scores on test data from 2008 and 2012. The performance of soft voting classifiers initially appears unimpressive, but given the unreliability of ANES respondents' claims about their past voting behavior due to social desirability bias, an alternate evaluation method is devised, in which predicted non-voters, as classified by the soft voting classifier, are excluded from the survey when making statistical inferences about the popular vote distribution. The resulting predictions are significantly more accurate than when using survey data alone and remain so even if different assumptions are made about how to aggressively the model should classify non-voters.

Broadly speaking, this exploratory project can be regarded as an initial proof of concept: that machine learning can be applied to national election surveys to enhance their accuracy in predicting the popular vote. Practically, however, accurately predicting the popular vote is of limited use as American presidential elections are decided not by who wins the most votes but rather who wins the electoral college. Thus, whether this same method yields similar results for state-wide and local elections remains a crucial but unanswered question.

Many pollsters currently develop likely voter models using the Perry-Gallup likely voter index, in which the respondent's answers to seven questions determine whether they are classified as likely voters. Because the ANES survey was not meant to be an election poll but rather a broader study, it does not include a complete set of Perry-Gallup questions. Yet another avenue for further study, then, is how a machine learning method of identifying non-voters compares with traditional methods of determining the likelihood to vote.

Given the importance of polling to news organizations in gauging public sentiment and to candidates when devising campaign strategies, the continuing difficulty of developing likely voter models presents a prime opportunity for the fruitful application of state-of-the-art machine learning techniques. This project demonstates just one instance in which machine learning can provide better polling results, and others are strongly encouraged to test this study's reproducibility with different voter files as well as extend and improve upon it with even more powerful methods.