Election Predictions from Local Demographic Data

Joel Gottsegen, Vignesh Venkataraman, Ben Weems - CS 221 - Stanford University

1 Introduction

Although the United States as a whole is a representative democracy, the state of California has a long-standing tradition of allowing citizens to vote directly on proposed legislation. These ballot propositions have significant statewide, and sometimes even national, implications: for instance, Proposition 8, a referendum on same-sex marriage, became a focal point for the battle over LGBTQ rights.

Given the impact of these propositions, it is unsurprising that activists and corporations spend massive amounts of time and money on "Yes" and "No" campaigns. Groups trying to prevent a recent medical insurance proposition from passing spent close to $60 million on advertising and get-out-the-vote (GOTV) initiatives. Clearly, there is significant interest in the outcomes of these referendums. The goal of our project is to create a system that accurately predicts whether or not a proposition will pass, in addition to providing voting-pattern insights that activists can use to plan their campaigns.

2 Task Definition

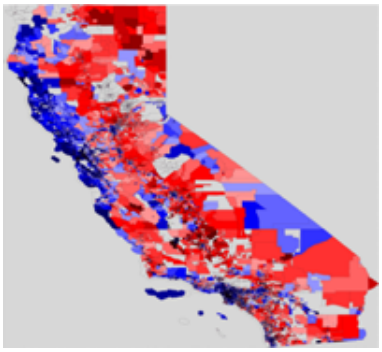

Figure 1: A typical voting distribution across the state of CA.

2.1 Issue Classification and Polarity

The key insight behind our project is that a voter's demographic background may be correlated with the way he or she votes on a particular class of issues. In order to harness this predictive power, we broke up the set of all propositions into a series of groups based on their subjects, henceforth referred to as "issue tags." These subsets were designed to be broad enough to allow for similar proposed legislation to be clustered together, while narrow enough for the clusterings to have intuitive meanings. Their definitions are as follows:

Infrastructure: Propositions concerning public infrastructure, including roads, buildings, and utilities.

Education: Propositions concerning schools and universities.

Crime: Propositions concerning prisons, convict rights, and crime prevention.

Gambling: Propositions concerning gambling regulations.

Politics: Propositions concerning the political process, including redistricting, legislative protocols, and campaign finance.

Environment: Propositions concerning environmental regulation, including pollutant control, park funding, and fuel efficiency standards.

Corporate: Propositions concerning corporate regulation, including insurance product requirements, corporate taxation, and employee benefit laws.

We hand-labeled the set of propositions based on their official descriptions. As an example, Proposition 24 from the 2010 election cycle had the following description:

> "This is a citizen-initiated state statute that would repeal three business tax breaks passed by the state legislature as part of negotiations of the 2008–10 California budget crisis."

Since the proposition deals with corporate tax policy, it fits best into the "corporate" issue tag. One nuance that we had to consider is that voting "Yes" on two propositions in the same category does not necessarily indicate agreement on the larger issue. To address this, we introduced the concept of "polarity". For each issue grouping, we define a general ideology corresponding to a positive polarity, and an ideology corresponding to a negative polarity. For the "corporate" group they are defined as follows:

Positive: In support of more corporate regulation and taxation.

Negative: In support of lowering corporate taxes and reducing regulation.

The concept of polarity, when paired with the issue tags, allows us to map a vote to a more general political sentiment.

2.2 Input-Output Behavior

The input to our system consists of demographic data for all California counties from 2006 to 2014, in addition to the proposition results by county over that same period. The output of the system is a set of trained classifiers, one for each issue tag. Each classifier takes in a vector of features extracted from demographic data for a particular county, and outputs a prediction of whether or not the proposition will pass in that county.

2.3 Literature Review

Previous literature on the subject is largely concerned with votes that fall along party lines, or "red vs. blue" voting. Propositions are unique because they are often nuanced and do not fall strictly along party lines. This adds a degree of difficulty in deciding who to target in a campaign for or against a certain proposition, because you cannot simply target Democrats or Republicans. As a result, some literature has been published concerning the demographics of individual propositions that were particularly impactful, such as Proposition 8. However, we have not found a comprehensive review of the demographic patterns across all propositions. This demographic classifier over all propositions is what our analysis contributes to the existing literature.

`http://www.ropercenter.uconn.edu/elections/how_groups_voted/voted_12.html`
`http://www.thetaskforce.org/static_html/downloads/issues/egan_sherrill_prop8_1_6_09.pdf`

3 Data

As indicated by the task definition above, there are two pieces of data that are vital to our project: demographic data and election data.

3.1 Demographic Data

This project is highly dependent on demographic data specific to election years. The official United States Census is taken every 10 years, making it unsuitable as a source for yearly demographic data. Fortunately, the US Census Bureau also performs an American Community Survey (ACS) every year, providing estimates for a wide variety of demographic metrics. For this project, we scraped the US Census Bureau's ACS 1-Year projections between 2006 and 2012. The data was presented in .csv format on a per county basis; this was allowed us to map the demographic data to election results at the county level.

3.2 Election Data

Election data was slightly less easy to come by, given the highly fragmented nature of California's Secretary of State website. However, we were able to find PDFs of election results on a per-county basis for each year from 2006 to 2012. We hand-copied the data from these PDFs into .csv format, thus completing our dataset acquisition.

3.3 Dataset Sanitization

The challenge of sanitizing the dataset was significant. As is typical with US government organizations, the Census Bureau changed its data formatting on a yearly basis, making data extraction from the ACS projections quite difficult. Again combing through the .csv files, we manually relabelled each demographic metric's header and deleted any columns that were not present in all years. Even with these removed, there were still 595 demographic metrics to choose from for each year between 2006 and 2012.

A final wrench thrown into this process was the absence of demographic data projections for several counties; we realized that locations like Alpine County, which have total populations in the hundreds to low thousands, were not pollable by the Census Bureau with any sort of statistical accuracy and were thus left out of the demographic projections. Rather than try to acquire this data, we chose to exclude these counties from our datasets.

4 Approach

4.1 Algorithm Choice and Evaluation Metric

There were two primary approaches that we considered: support vector classification and support vector regression. The former approach outputs a binary response; for a given demographic input, would the voting result be yes or no? The latter approach would provide a continuous response; for the same demographic input, what percentage of people would vote yes, and what percentage of people would vote no? We chose to focus on support vector classification, as that most closely mirrors how voting in California works (i.e. it only takes a one vote margin of victory for a proposition to pass). Margin of victory makes no difference in the eyes of the state government. Classification can also be computed faster, according to the scikit

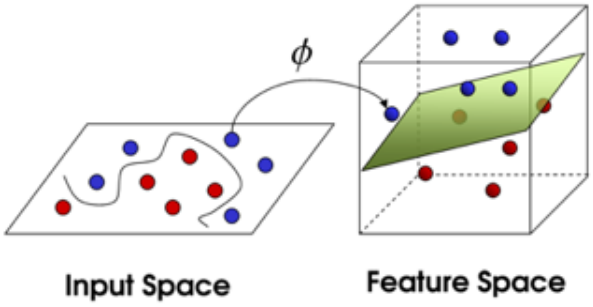

Figure 2: A visual representation of Support Vector Classification.

documentation that we consulted. Additionally, support vector classification has a very simple evaluation metric: did the prediction match the true value of an example?

4.2 Approach Details

Our goal required the creation of a classifier for each issue tag. In order to accomplish this, we constructed a series of design matrices, one for each issue tag. For each issue tag, we chose a number of features (the details of feature selection are explored in the Experiments section). For each proposition in the issue tag, our training data is the set of examples constructed from the demographic and election results data. Each example consists of a single county's demographic datums in the features picked for the specific issue tag, and the overall result of the vote ("yes" or "no"), adjusted for polarity. Thus, if there are two propositions contained in an issue tag, the design matrix would be each county's important demographic features for the years the propositions took place, and the target vector would be each county's net voting decision on each of the propositions. Running SVC on this dataset will produce a per-issue-tag classifier which takes as input demographic data and produces as output the predicted results of a proposition that falls into the specific issue tag that the classifier pertains to. In order to achieve good results with our small dataset, we chose to cross-validate, splitting each issue tag's propositions up such that 75% were used as a training set and 25% as a test set. After training the classifier, we randomly re-rolled the training and test set splits and averaged the error over ten to twenty iterations to better estimate the accuracy of our algorithm.

4.3 Challenges

There are a number of challenges associated with this choice of model. The biggest problem is the general lack of data: given the election and demographic data that we were able to pull (which only dated back to 2006), we had a total of 100 propositions over 40 counties, meaning that in total, there were only 4,000 training examples to disperse amongst each of our issue tags. Some tags ended up with just 4 or 5 issues in them, which leading to a small enough training set for overfitting to be a real concern.

Another issue is that which issue tag a proposition should have is sometimes ambiguous. For instance, Proposition 39 from the gubernatorial election in 2012 was a measure intended to help the environment by raising corporate income tax rates; however, this falls into both the corporate and environment issue tags! As a result of this ambiguity, we had to do additional research to find out what lines of reasoning were followed by the promoters and dissenters of this proposition, eventually concluding that the corporate impact would have been far more significant than any environmental gains (the issue was ultimately tagged as corporate). This was by no means the only issue that suffered from these ambiguities; we had to do significant research on several issues before settling on their "proper" classification. A final challenge is feature choice; given

```
def test_features(issue_tag, features):
    training_issues_hash = get_all_training_issues()
    train_issues = training_issues_hash[issue_tag]
    test_size = len(train_issues) / 4
    random.shuffle(train_issues)

    test_issues = []

    for i in range(test_size):
        test_issues.append(train_issues.pop())

    tag = { "name": "Random", "type": "Percent", "demographics": features }
    model,design_matrix, target_matrix = build_classifier_model(train_issues, tag)
    test_design_matrix,test_target_matrix = combine_design_matrices(test_issues, tag)
    test_target_matrix = convert_to_binary_target(test_target_matrix)
    train_error, test_error = test_classifier_model(model, design_matrix, target_matrix,
                                                    test_design_matrix, test_target_matrix)

    return train_error, test_error
```

Figure 3: A code outline of our proposition-level algorithmic procedure.

roughly 600 options to choose from, how do we know which features are relevant, and how do we know which combination of features will produce accurate results? The age-old machine learning adage is that your classifier is only as good as the data you feed it and the features you pick; these are fairly common issues that we tried to address as we conducted our experiments.

4.4 Baseline

Recollecting our key insight, that demographic data can have a significant effect on voting patterns and election results, the baseline against which our results can be evaluated is relatively naive. Instead of plugging in varied demographic data, we use as input an identical vector for each and every county (i.e. the feature vector [1]). From a higher level perspective, this is akin to training a model while assuming that all counties have the same demographic makeup. The target variable for each county is how the county voted on the issue in question. In essence, the classifier trained from this input data set will predict the majority label (i.e. "yes" or "no", as determined by the majority of counties statewide), which seems like a reasonable place to start. This baseline will be fairly accurate across the state but will not be able to offer demographic-specific or location-specific results.

4.5 Oracle

Similarly, we can define an oracle (i.e. a "cheating" method) to try to give an upper bound on performance. The oracle we chose is, for a given proposition and county, a classifier trained on all the county demographic and election data for the proposition in question EXCEPT for the county we are interested in. The logic here is that each proposition is unique, so the results of the proposition itself is the best training set. Additionally, demographics will have similar effects on voting patterns across counties. Thus, excluding the current county from consideration will make this a great, but not perfect, measure of accuracy for the proposition.

5 Experiments

5.1 Generating Issue Tags

Our group analyzed all of the propositions from the past 8 years and brainstormed logical clusterings. We worked together to decide on a group of issue tags which would allow each bucket to hold enough propositions for training to be meaningful, while also being small enough such that all propositions concerned similar legislation. This required tweaking throughout the process as we realized that some tags did not encompass enough propositions, and others were too broad. For example, we originally had a bucket that encompassed LGBT rights, but we soon found out that Proposition 8 was the only proposition that would be categorized within this tag. We also referenced Ballotpedia (a well-reputed internet resource) for ideas on possible issue tags.

5.2 Assigning Propositions to Tags

The process of assigning issues to tags was also done manually. We sat down and assigned each issue to one of the buckets we had generated. This was sometimes difficult because a proposition was not fully encompassed by one subject, but due to our SVM model we were required to choose only one. We reached mutual agreement on all tags.

5.3 Feature Selection

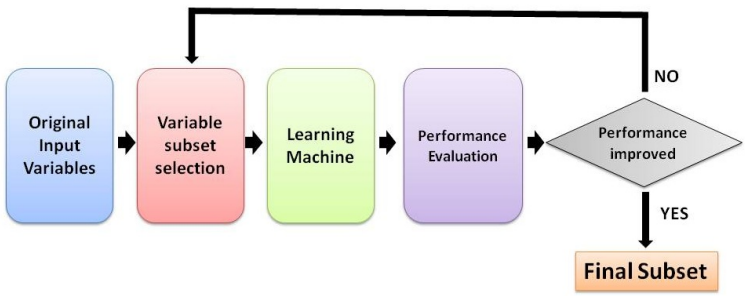

Figure 4: The workflow followed to select the optimal set of features for each issue tag.

For each issue tag, we had to specify a unique set of demographic features which would impact how a region would vote of propositions concerning that issue. We had approximately 600 features available to us, but using all of these would have resulted in overfitting, so we needed a subset of these demographics. To do this, we developed a multithreaded python script which would choose the best features available to us by an exhaustive method. Because each run of the model was expensive, and there was high standard deviation in testing error across runs, it would be impossible to try all combinations of features. Instead, for each demographic feature, we found the test error of that isolated feature: if it was better than the baseline we kept it, if it was worse we got rid of it. We ran this recursively, eliminating features each time and averaging the test error of each feature over all runs until there were 10 or fewer features remaining. The remaining features are used as our feature vectors. This method makes the assumption that features impact voting patterns in isolation, meaning that there are no synergistic or antagonistic effects between features and that

selected features are not redundant. Additionally, limiting each feature vector to 10 or fewer dimensions helps ward off overfitting, a significant problem given our relatively small dataset.

6 Results

The results achieved relative to the baseline are as follows (training plus 20 iterations of cross validation for each issue tag's classifier):

Issue Tag	**Baseline Test Error**	**Algorithm Test Error**
Infrastructure	25.38%	16.38%
Education	30.63%	22.67%
Crime	53.90%	34.65%
Gambling	0.0%	0.0%
Politics	43.67%	34.50%
Environment	46.88%	35.18%
Corporate	35.38%	28.50%

The features used by each classifier are as follows. Note that all features are the percentage of the population that is in the described demographic:

Issue Tag	**Features**
Infrastructure	Lived in same house 1 year ago, Lived in different house 1 year ago, Elementary school children, Nursing school, Children, Family Households, Spouses, Households with children, Single-parent households
Education	Grandparents, Ukranian-descent, Hungarian-descent, European-descent, Greek-descent, Born out of state
Crime	Slovak-descent, English-speaking-only, Danish-descent, Divorced, Householders who own home, Single males, Native population, Oceanian
Politics	State resident for over 1 year, Women who have birthed in the past year, Disabled citizens, Married males, Latin Americans, Non-US citizens, European-born, Year of entry of foreign-born population
Environment	College-educated, Danish-descent, Asian-language speaking, North-American born, Foreign-born, US-born
Corporate	Nonfamily households, In state more than 1 year, Women who have birthed in last year, Disabled, Graduate educated, Family households, Foreign-born, Foreign-born in country more than 15 years

7 Analysis

7.1 Interpretation and Discussion

As presented in the Results section above, across all but one issue tag ("society") our approach outperforms the baseline by significant margins. At face value, this appears to verify our supposition that demographic data can serve as a good predictor of voting outcomes. Our algorithm does especially well when used as a testbed for a variety of features; while a human may be able to juggle one or two demographic variables and look for correlations between them and decreased test error, but with this system we can test, retest, recombine, and shuffle various different feature vectors and work towards finding the "best" one. In fact, the value of this system isn't just in predicting results, but also in finding out what features have significant impact on prediction accuracy. Learning what these features are, from a qualitative rather a quantitative sense, can help campaign managers and interested parties better understand what makes the voting populace act the way it does. This data could prove to be useful for other applications in the future.

Some of the features are fairly intuitive, given their issue tag. For example, many casual political observers would have guessed that college-educated people are more likely to support environmental regulations. The true value of the project is in the unexpected features. Even the most savvy political pundit would have had difficulty predicting that people of Slovak and Danish descent would be especially likely to support anti-crime legislation. One could explore the sociological reasons for these particular groups holding these views, but this is beyond the scope of our project and we make no attempt to do so here. What we do invite the reader to do is explore our table of features and make note of the granular and unexpected nature of many inputs to our models. This is the value of the machine-learning approach–it would be extremely difficult for a human to manually pick through hundreds of potential features for each of the models. Our approach allows the best predictors to rise to the surface.

It is also worth mentioning that some of the correlations shown in this data could be purely coincidental; that is to say, with such a small training and test set, there is high potential for pseudo-random features to yield remarkable results. An expansion and extension of this project, to include many more propositions and many more years of demographic data, would certainly allow for more concrete and rigorous conclusions to be reached.

7.2 Successes

Our biggest success was undoubtedly triumphing over the baseline in the vast majority of issue tags as a whole. Many of our success rates are on par with professional polling and projection companies, which is certainly a good trend. The issue tag model worked remarkably well, especially when considering the small numbers of propositions that we were able to assign to each issue tag. That success at least suggests that we were right to assume that demographics will vote similary on issues from the same bucket. We were also able to nail down some consistently useful features, like "% families with children" or "% English as a second language," which had high correlations to successful predictions across many different issue tags. We were also able to verify our initial hypotheses (that election results could be predicted, and demographics could be the key to unlocking those predictions). A final great success was the data that we collected and sanitized; although we chose to use it to predict elections, the utility of having demographic data about every county in California from 2006 onwards cannot be overstated.

7.3 Areas for Improvement

The most glaring need for improvement is shown by the performance of the "society" issue tag, which already starts with a high baseline error (45%) and manages to do worse when our algorithm is applied

(finally weighing in with a ludicrous 52.225% test error). The struggles of our classifier to perform in this issue tag make a ton of sense when retrospectively looking at all the issues that have gone into that issue tag; with "society" serving as our catch-all category, it contains propositions that range from LGBTQ rights to progressive tax reform to updated state building codes. With such a diverse and disparate set of topics contained within itself, it is very hard for the "society" classifier to really learn anything relevant, and the test error balloons. One of the first things we would look at doing in the future is splitting up this issue tag into more segregated tags; however, given our small dataset, this was not achievable with just propositions from 2006 onwards. To actually perform the split would require many more datapoints and thus much more data.

Another area that coud use some improvement is our feature selection. With 595 options to choose from, it was very hard to be exhaustive in our attempts to pick the "best" features and combinations of features. Improved feature selection is a bit of a machine learning cliche, but our project could truly benefit from it given extra time and more powerful computing resources. Deep learning might also apply to this situation, as neural networks and other self-teaching strategies could learn automatically what features are most useful.

7.4 Error Analysis

7.4.1 General pattern

We will analyze crime features to identify some of the weaknesses and strong suits of our classifier. The crime features we used are detailed in the feature selection table in section 6. A key feauture in this set was the percentage of householders who owned their own home.

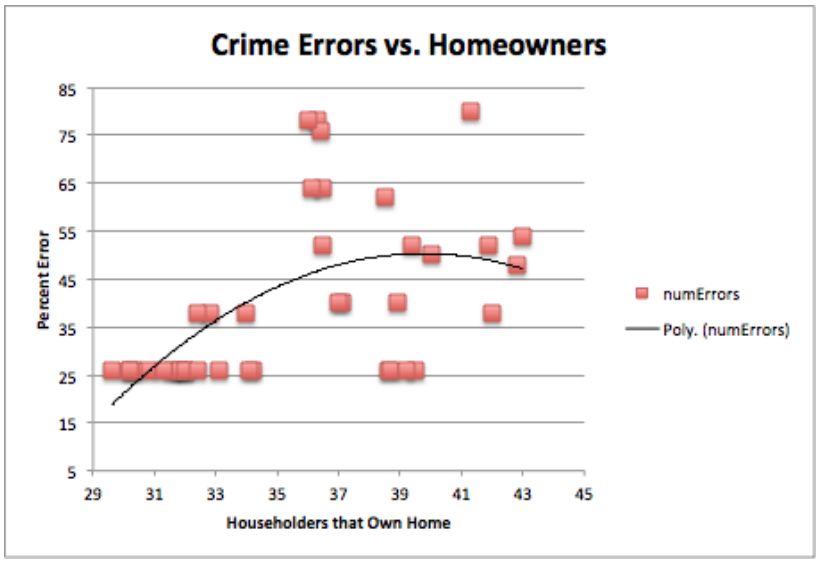

Figure 5: A graph showing percentage error in the crime feature tag versus the percentage of householders that own their own home.

The graph shown above represents the percent error of a specific county on crime issues when testing and training was done exclusively on the feature of percent of householders who own their home. Each dot represents an individual county.

The trendline suggests that at low percentages of homeowners, the error rate is small, while the classifier does worse as the number of homeowners rises. However, we also see that the trendline begins to dip at the higher rate of homeowners. This suggests that there is a middle area at which we see the high error. The results of these counties are unpredictable because they are in a toss-up region. However, at the extremes, the feature is a strong indicator of a region's voting patterns. By creating feature vectors with more than one feature, we are hopefully able to classify regions that fall into the toss-up region by other features. As we add features, the toss-up region shrinks as counties get pushed one way or the other.

7.4.2 Example Error

The model predicted San Mateo County's voting pattern incorrectly for all propositions in the crime bucket. Let's look into why this might have occurred:

If we use the percent of population that speaks only English at home as the only feature, San Mateo gets about half of propositions wrong. From an examination of trained weights, we see that counties with low populations with English as the only language spoken in the home, such as San Mateo, have a tendency to vote against stricter crime measures. If we look at some cases it gets wrong, we see that San Mateo county is more likely to vote No on crime measures (more likely to vote against stricter punishments). So, as expected, our predictor is wrong when we expect San Mateo to vote no on a crime measure but it actually votes Yes.

8 Conclusion

The electoral system is the basis for American politics, and the proposition system provides a unique opportunity for citizens to vote directly on the legislation that they care about. In this project we have built a system that predicts whether a proposition will pass and identifies demographics that are sympathetic to particular types of legislation. We hope that these models will be useful for political strategists and activists as they develop their campaign strategies. We also hope to further expand on this approach with a more ambitious undertaking, perhaps one that has nationwide context as opposed to being exclusively focused on California.