Lukas Froehlich
Waylin Wang

Data 401 - Project #4

Understanding the 2016 Election

Data Collection

The original dataset came from the a github repo that has scraped county level election data from townhall.com. This data had results for each county for years 2008, 2012, 2016. The demographic data was collected from the US Census API for the years 2011 through 2015.

Question

We had many questions we want to explore with this rich set of data, but we narrowed down to three we know we can answer confidently with classification and logistic regression. The questions are:

- Which demographics were most significant for voter turnout in 2016 compared to 2012?
- Which demographics contributed most to change in voter behavior from 2012 to 2016?
- How well can county majority be predicted given demographics?

Feature Transformation

A pattern we found was that most of the predictors were number of people that followed a particular demographic in a county. For these variables we decided to divide by the population of the county to turn these demographic variables into percentages of the county.

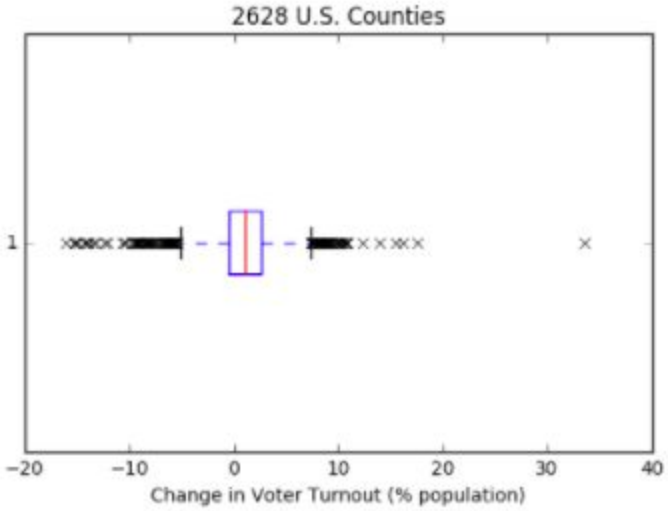
Exploring Voter Turnout

The question we were trying to answer was “Which demographics were most significant for voter turnout in 2016 compared to 2012?” The motivation for this question was that it would give us some exposure as to which groups should have voted but didn’t end up showing up. The election in 2012 is used generally as a baseline for expected voter behavior. By comparing 2016 election metrics to 2012 we are able to see unexpected behavior in this election.

This analysis begins by creating a response variable to represent the voter turnout change from 2012 to 2016.

```
change_voter_turnout = (total_votes_2016 - total_votes_2012) / population
```

After applying this function to each of the counties we are able to plot these value to get a sense of the shift in voter turnout.



mean	1.045871
std	3.064287
min	-16.037837
25%	-0.522142
50%	1.143138
75%	2.655468
max	33.625456

We then split the counties into two groups, those with very positive change in voter turnout (> Q3) and those with very negative change in voter turnout (< Q1). Once these two groups were established it became easier to work with the data and look for differences between the two groups. We wanted to see which demographics best distinguish these two groups, we did this by running the random forest classifier. Using random forest as our classifier gave us 77% accuracy when running cross validation. Random forest's also have the advantage of weighting the importances of each of the predictors. We then took the highest weighted demographics and compared their mean values between the groups to see how it affected the county.

Demographic	Mean value in low turnout counties	Mean value in high turnout counties
% Black	14.74	5.65
In State-Some College	13.79	10.8
% Black Female	7.51	2.76
Median Age Male	0.28	0.15
Total Reporting 0 Income	10.68	10.6

# of Housing Units	6.5	4.5
Total Born In State education	44.11	36.86

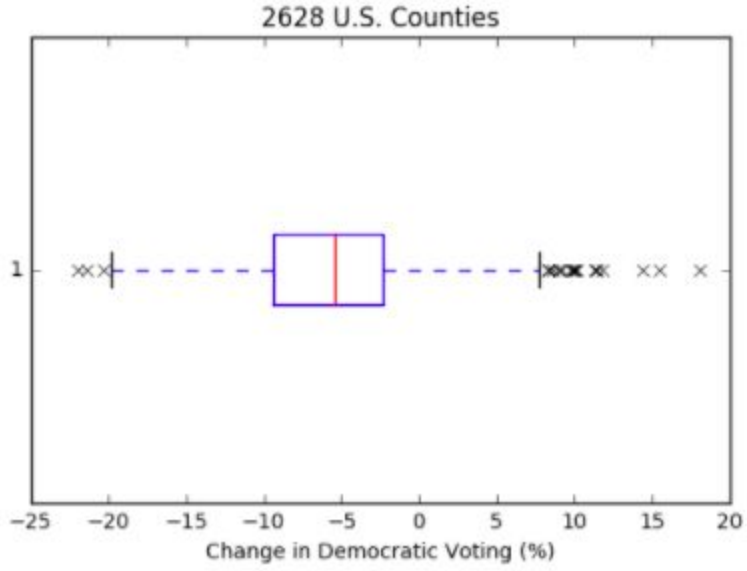
These results are powerful because they can be used as evidence to draw greater conclusions in the election. For example, there is a clear indication that the % black vote was a considerable separator between counties with low turnout and high turnout. Furthermore, counties with a high percentage of black voters had considerably lower turnout than counties with lower percentages. This could be attributed to Obama being in the last election. Another interesting observation is that counties with high volumes of educated people did not come to vote as much as they did last year.

Exploring Voter Behavior

The goal of this section is to gain an understanding of which demographics were most significant in the shift of voter behavior in counties. This response variable used was calculated as follows:

```
        voting_ratio_2012 = dem_2012 / total_2012
        voting_ratio_2016 = dem_2016 / total_2016
voting_ratio_change = ratio_voting_2016-ratio_voting_2012
```

Once we had this predictor we followed the same procedure as in the voter turnout section.



mean	-5.676965
std	5.351930
min	-21.940109
25%	-9.387943
50%	-5.361564
75%	-2.358579
max	18.064662

With the random forest we were able to get 94.2 prediction accuracy. The greatest contributing variables were:

Demographic	Mean value for county that voted more GOP	Mean value for county that voted more DEM
% White Not Hispanic Male	45.33	32.55
Foreign Born below 100% poverty	0.4	1.9
Foreign Born at/above 1.5X poverty line	1.2	6.23
% White Not Hispanic Female	45.62	33.38
Citizen by Naturalization	0.82	3.88
HS or equal	27.76	17.68
Total Born In State education	49.82	32.99

From this we can use it as supporting evidence to start drawing conclusions. One potential explanation for the election was that it was a white backlash against a changing government, and this is supported by counties with higher % white population changing to vote more GOP. Another interesting observation is that foreigners voted more so towards democratic, possibly out of fear for Trump's strong stance against immigrants.

We also tried using Elastic Net and logistic regression to try to see the change in feature selection for predicting election outcome at the county level. The idea is to have Elastic Net select the best features to use for logistic regression and see how well the regression model did. We can then compare the features selected and their corresponding coefficients to see how each feature affected the outcome.

We at first tried Lasso to select the best features, but Lasso was dropping most of our 130 features, so we moved on to try basis expansion with Lasso. Turns out that didn't work very well either. We were left with less than 5 features with Lasso, so we decided to go with Elastic Net instead. Since Elastic Net factors in L1 and L2 errors, it's a more reliable technique anyways. Running through the 2012 demographic data, we got about .879 cross validated f1 score with 27 features. The coefficients are displayed below.

Currently Married	-0.317283
Moved Within Same State	-0.280607
Other State-HS or equal	-0.252629
Less Than HS	-0.234204
Graduate	-0.164781
Total Reporing 0 Income	-0.137280
# White Not Hispanic	-0.097915
2 vehicles	-0.087223
Same house 1 year ago	-0.070133
Median Income	-0.004497
Median Income Born Outside US: Native	-0.001337
Median Income Born Other State	0.000057
Median Income Foreign Born	0.000827
Median Income Born In State	0.001208
Same house 1 year ago White	0.037653
100% below povery line	0.069505
Work outside County	0.078911
Foreign born naturalized	0.104808
Citizen by Naturalization	0.104808
Never Married	0.112771
Work outside State	0.134887
# of Housing Units	0.154294
Income 35-50K	0.213100
Education Count	0.296345
Income 50-65K	0.335249
Other State-Graduate	0.363238
Walked	0.590713

From the coefficients, we can infer that (1 being Democrat and 0 being Republican) People living in metropolitan areas with an above average income and college education voted more for the Democratic Party in the 2012 election. On the other side, we see that married individuals, especially couples with one person stay home, with 2 cars and people without a college education that have been living in the same house for the past year is more likely to vote Republican. This aligns well with the stigma that the Democratic Party have their strongholds in the more liberal and metropolitan areas, while the Republican Party have the strongholds in more conservative regions.

Now we performed the same analysis with the 2014 data since that was the most recent year we could find with complete datasets from US Census. 34 features were selected by Elastic Net for this set of demographics. Our model was

Less Than HS	-0.265917
# White Not Hispanis Female	-0.262218
Moved Within Same State	-0.246613
Moved From Different State	-0.219069
Born Other State at/above 1.5X poverty line	-0.208518
Total Reporing 0 Income	-0.191759
Income > 75K	-0.181936
Currently Married	-0.168618
No Vehicle Owned	-0.156645
Same house 1 year ago	-0.090056
Graduate	-0.063102
Work in County	-0.043940
Bachelor	-0.039522
In State-HS or equal	-0.029615
# Hispanic	-0.010952
# White Not Hispanic	-0.008495
Median Income	-0.005299
1-1.5X of poverty line	-0.005140
Median Income Born Other State	-0.001758
Median Income Born Outside US: Native	0.000871
Median Income Foreign Born	0.003692
100% below povery line	0.023298
Same house 1 year ago White	0.026474
# Black Female	0.029098
Work outside County	0.036297
Foreign born naturalized	0.174013
Citizen by Naturalization	0.174013
Work outside State	0.196733
Never Married	0.202886
2 vehicles	0.220181
Education Count	0.289691
Other State-Bachelor	0.427380
Walked	0.574626
Other State-Graduate	0.836953

From the 2014 data, we noticed the same general trend with Democrats winning the metropolitan areas and Republicans winning the others.

However, it is way more interesting to see when the coefficients are put side by side below for contrasting.

	Field	2016_coef	2012_coef	diff
1	# White Not Hispanis Female	-0.262218	0.000000	-0.262218
3	Moved From Different State	-0.219069	0.000000	-0.219069
4	Born Other State at/above 1.5X poverty line	-0.208518	0.000000	-0.208518
6	Income > 75K	-0.181936	0.000000	-0.181936
8	No Vehicle Owned	-0.156645	0.000000	-0.156645
11	Work in County	-0.043940	0.000000	-0.043940
12	Bachelor	-0.039522	0.000000	-0.039522
13	In State-HS or equal	-0.029615	0.000000	-0.029615
14	# Hispanic	-0.010952	0.000000	-0.010952
17	1-1.5X of poverty line	-0.005140	0.000000	-0.005140
18	Median Income Born Other State	-0.001758	0.000057	-0.001814
19	Median Income Born Outside US: Native	0.000871	-0.001337	0.002207
23	# Black Female	0.029098	0.000000	0.029098
29	2 vehicles	0.220181	-0.087223	0.307405
31	Other State-Bachelor	0.427380	0.000000	0.427380
34	Other State-HS or equal	0.000000	-0.252629	0.252629
35	Median Income Born In State	0.000000	0.001208	-0.001208
36	# of Housing Units	0.000000	0.154294	-0.154294
37	Income 35-50K	0.000000	0.213100	-0.213100
38	Income 50-65K	0.000000	0.335249	-0.335249

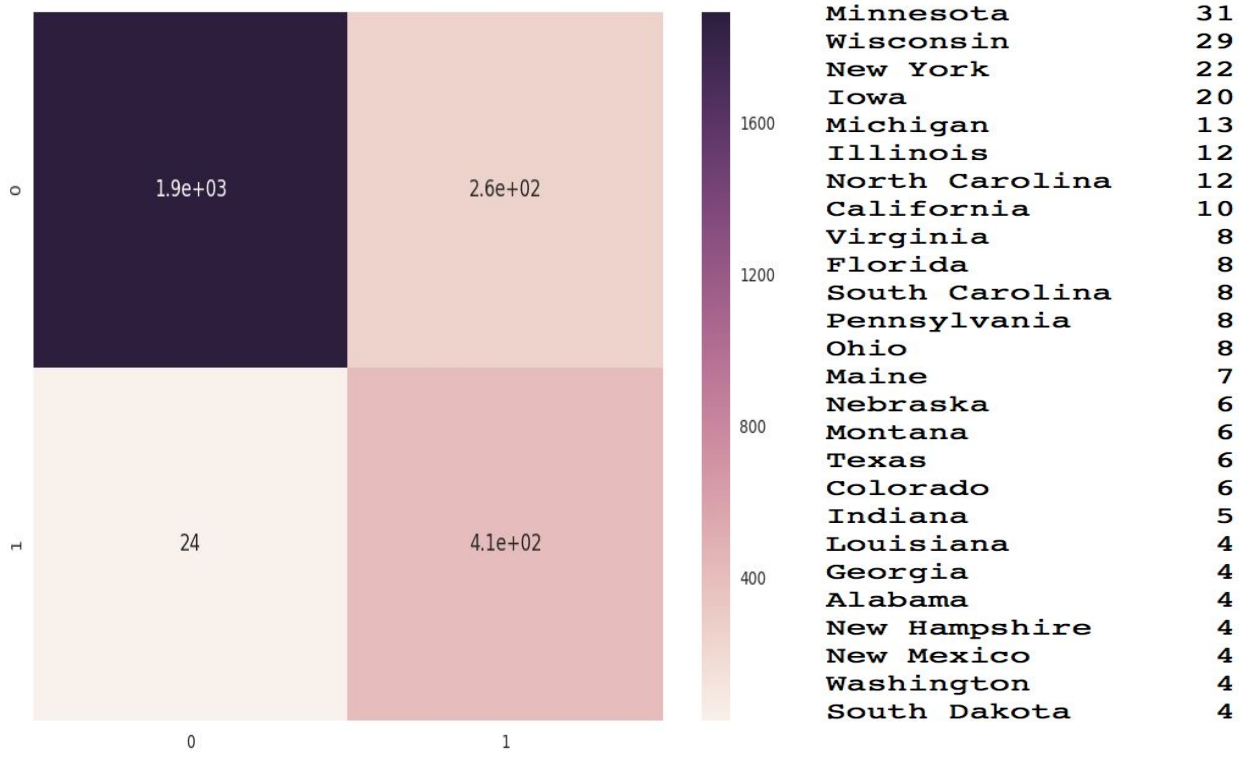
We noticed that there are more features being introduced into the model for 2014 data set, suggesting a change in support for both parties. In the chart above, we noticed that females started to show more participation in the political process. Caucasian women sided more strongly with the GOP while African American women sided more with the Democratic Party. Also we can see come cleavages forming between where people received their 4-year college education, with people that received their degree from out of state heavily favoring the Democratic Party and people that received their degree from in-state universities voting towards the Republican Party. We noticed that income between 35-65 K has been dropped and income > 75 K added. This trend is probably due to the economic strength currently in the United States, so people have higher income than from 2012. Since Republicans tend promote tax cuts, people moving from the lower bracket into higher income might be more supportive of paying less tax. Another interesting shift is the movement of having two vehicles from supporting Republicans to Democrats.

Building Models to Predict Election Outcome

For this section, we tried some clustering and classification techniques to replicate what went wrong.

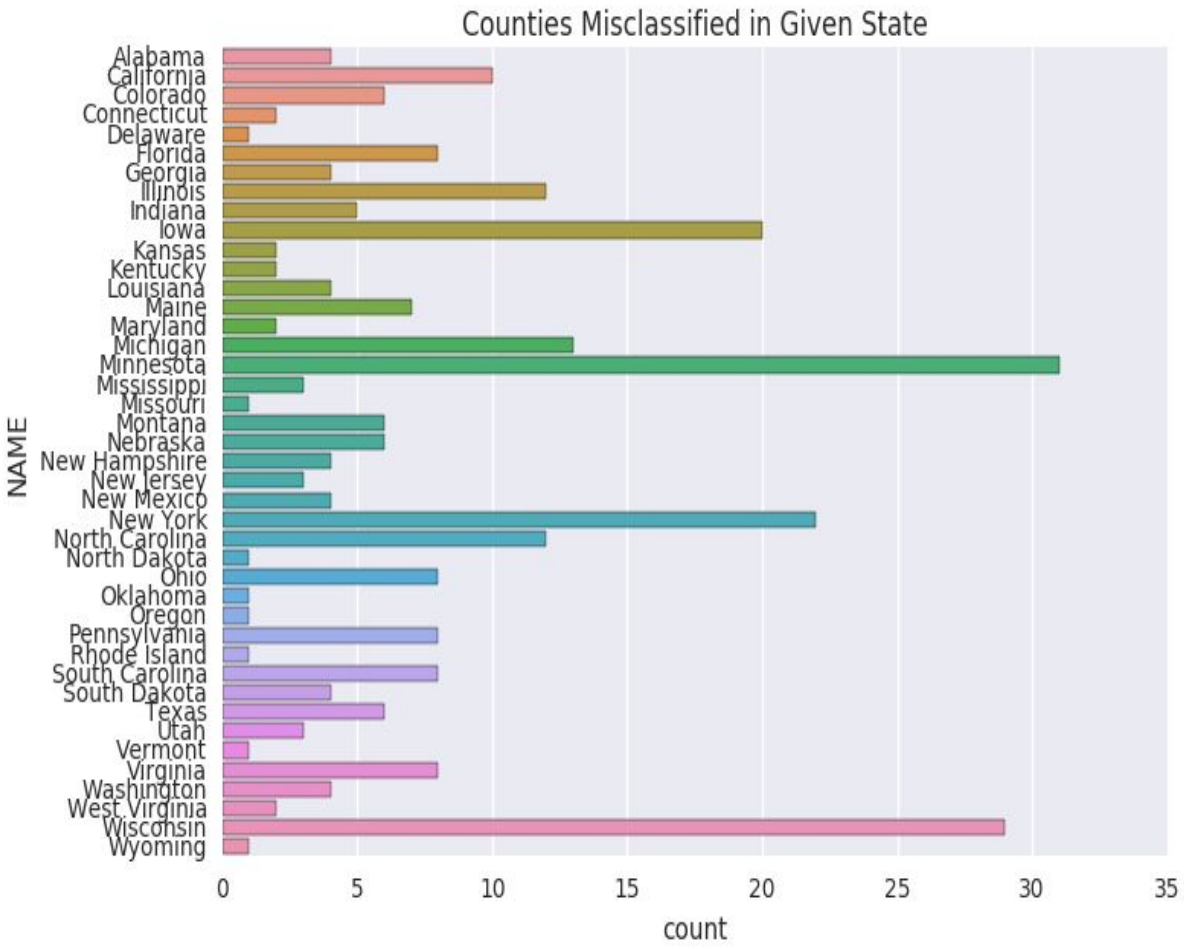
We tried the KMeans clustering algorithm on our data set and went with 3 clusters because we assumed that people with similar demographics tend to vote for the same party. We chose three clusters, representing Democrats, Republicans, and Other. Our clustering techniques used the same features selected by Elastic Net, and we were able to get 77 % accuracy, until we realized that it was putting everything into the Republican cluster and there are about 23% Democratic Counties. We plotted the 130 features and saw that many of the features have Democrats and Republicans mixed together and it was unlikely for us to be able to separate them into meaningful clusters.

We moved on to classification techniques instead of clustering. We are training with the 2012 demographics data. After a quick run to select which classification technique to use, we noticed that Ada Boost performed very well at 85% accuracy beating out the untuned neural network. However, when we started tuning the SciKit Learn implementation of neural network, we were able to achieve cross validated accuracy of 90%. We then used this model to predict election result for 2016 with the 2014 demographic dataset.



Minnesota	31
Wisconsin	29
New York	22
Iowa	20
Michigan	13
Illinois	12
North Carolina	12
California	10
Virginia	8
Florida	8
South Carolina	8
Pennsylvania	8
Ohio	8
Maine	7
Nebraska	6
Montana	6
Texas	6
Colorado	6
Indiana	5
Louisiana	4
Georgia	4
Alabama	4
New Hampshire	4
New Mexico	4
Washington	4
South Dakota	4

Interestingly enough, our confusion matrix showed that we classified 260 of the counties as Democratic, when in fact they voted more towards Republican in 2016.



In the diagram and table shown above, we can see most of the states that flipped from blue to red during the 2016 election. So needless to say, we have successfully created a classifier that erroneously predicted Clinton to win.

From the model we build, we also answered our third question that with demographics, even though we have decent accuracy in predictions, our models would fail because of how the electoral process work with each counties weighing differently. On top of that, since demographics change over time, it is also very difficult to maintain an up-to-date model to predict for future elections.