PSTAT 131: 2016 Election Analysis Report

Gerardo Gonzalez, Chris Chiang, Eric Bletcher

Introduction

Before an election, determining vote predictions and results from polls is a difficult task. Generally, pollsters struggle to make polls and surveys that can accurately predict outcomes at county, state, and federal levels, as well as for demographic factors. Many bias and surveying problems can create over/under-representation for certain candidates. For example, online responses vastly underestimate republican voters, because people who tend to live in rural areas swing that direction and can't as easily access technological sources as the general population can. Trump's win in the 2016 election came as a big shock to many: most polls didn't suggest his victory but turnout problems, tossups, and geographical bias resulted in a vast underestimation of the percentage point difference, especially in many battleground states.

For this project analysis, we will try to complete two tasks for predicting outcomes of the 2016 election, using methods that we've learned in this class. For task 1, we will try to predict county level winners. We will use a logistic regression model and prepare the merged data for modeling by partitioning the data. This can open the opportunity for creating different things such as decision trees and ROC curves, and finally comparing our predictions to the actual result. For task 2, we will attempt to predict the winner of the popular vote. For this task, we will use a regression tree to estimate the proportion of voters in each county who voted for Hillary Clinton. With these estimates, we will find the number of votes for Clinton and Trump in each county and then find the sum of these to determine the estimates for each candidate's total votes received.

Materials and methods

Datasets

Our starting datasets consisted of an election dataset and a census dataset. The election dataset initially contained rows on three kinds of observational units: the county, the state, and the country. Each row displayed votes each candidate received in a county, state, or the nation. First, we divided the election dataset into three separate datasets, corresponding to county, state, and federal election information.

The census dataset described demographics and other statistics for each census tract. We had to aggregate the data to the county level by weighting each tract by its population.

Finally, we merged the county election data with the county-level census data to provide the main dataset used for our analysis. A few rows and columns of the resulting dataset are shown below:

county	candidate	state	votes	total	pct	Women	White
kent	Donald Trump	delaware	36991	74260	0.4981	51.79	63.68
kent	Hillary Clinton	delaware	33351	74260	0.4491	51.79	63.68
new castle	Hillary Clinton	delaware	162919	261507	0.623	51.65	59.72
new castle	Donald Trump	delaware	85525	261507	0.327	51.65	59.72

Methods

For task 1, we analyzed a merged county-level dataset in order to predict the winner of each county. We will merge the data for both the county winner and the census, then partition the result into 80% training and 20% testing partitions. Once we do this, we will train a logistic regression model on the training partition. What this will allow us to do is predict the most important demographic factors, which will aid us in creating a model to help us determine the winning candidate in each county. This logistic regression model is also able to be used to estimate any errors on the test partition as well. Once we have this model, this will show any significant variables that contributed to the odds of voting for a certain candidate in any particular case. It will predict the winner of a county based on these variables and compare it to the actual candidate. Easy to read charts such as misclassification tables and rates allow us to accurately sum up these results and compare them to each other.

For task 2, predicting the popular vote, we use a regression tree to estimate the proportion of voters in each county who voted for Clinton. Essentially, we partition the data into a number of regions by splitting on certain variables. Then, our estimate is the average Clinton proportion in that region. We only consider voters who voted for Clinton or Trump, so that our estimates of Clinton proportions also give us estimates of Trump proportions. Finally, we use these estimated proportions and the populations of each county to get estimated votes for Trump and Clinton for each county, then sum these up for national popular vote estimates.

Results

Task 1:

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Table 22: Logist Model Misclassification Error

	Donald Trump	Hillary Clinton
Donald Trump	0.9752	0.02484
Hillary Clinton	0.3049	0.6951

Table 23: Logistic Model Significant variables

Variables	Estimate	Exponential_coef	Pr(>|z|)
Service	0.3373	1.401	7.618e-12
Professional	0.2479	1.281	6.396e-11
Employed	0.1948	1.215	2.983e-09
Unemployment	0.1986	1.22	6.789e-07
PrivateWork	0.1025	1.108	1.349e-06
Drive	-0.2303	0.7943	1.894e-06
Production	0.1667	1.181	8.047e-05
Citizen	0.1055	1.111	0.0001528
IncomePerCap	0.0002387	1	0.000295
Carpool	-0.2118	0.8092	0.0007599
Office	0.1272	1.136	0.005071
Men	0.1405	1.151	0.005857
Intercept	-27.39	1.275e-12	0.006213
IncomePerCapErr	-0.0003048	0.9997	0.0234
WorkAtHome	-0.1664	0.8467	0.02533
FamilyWork	-0.9826	0.3743	0.02714
Income	-5.699e-05	0.9999	0.03578
MeanCommute	0.05013	1.051	0.04073

After training our logistic regression model on the training set and estimating the errors on the test partition, we can see that there are numerous significant variables whose p-value is less than 0.5. These consist of Service, Professional, Employed, Unemployment, PrivateWork, Drive, Production, Citizen, IncomePerCap, Carpool, Office, Men, Intercept, IncomePerCapErr, WorkAtHome, FamilyWork, Income, and MeanCommute. We used these factors to create a regression function to help predict Trump and Clinton winners in each county and summed the results. As an example of interpreting these variables, let's look at Service. We can see that a one percent increase in people in the service job field is associated with an increase in the odds of voting for Hillary Clinton by a factor of 1.401. This can be generalized to the rest of the variables.

From Table 22, the misclassification errors, we see the percentage of people we had predicted to elect each candidate based on demographics, based on training partitions we have introduced. For example, 97.52% of predicted county-level Trump winners were actually true. Many of the counties that we had predicted to be in favor of Clinton, were wrong and were actually in favor of Trump, which was 30.49%. The true positive rate was 69.51%, while the true negative rate was 97.52%

Table 24: Predicted vs Actual Candidate for San Diego 2016 (continued below)

county	candidate	predicted_candidate	Men	Women	White
san diego	Hillary Clinton	Hillary Clinton	50.12	3221	46.96

Table 25: Table continues below

Citizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr	Poverty
66.32	69943	10850	31282	3983	14.48

Table 26: Table continues below

ChildPoverty	Professional	Service	Office	Production	Drive	Carpool
17.13	39.26	20.13	23.69	8.689	76.55	9.594

Table 27: Table continues below

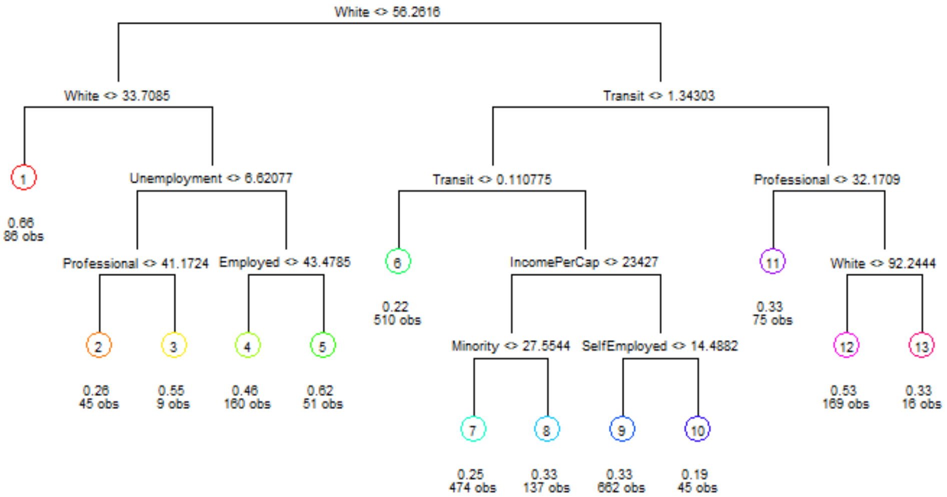
Transit	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork
3.108	1.923	6.315	25.29	45.52	76.86

SelfEmployed	FamilyWork	Unemployment	Minority
7.666	0.1605	8.948	49.75

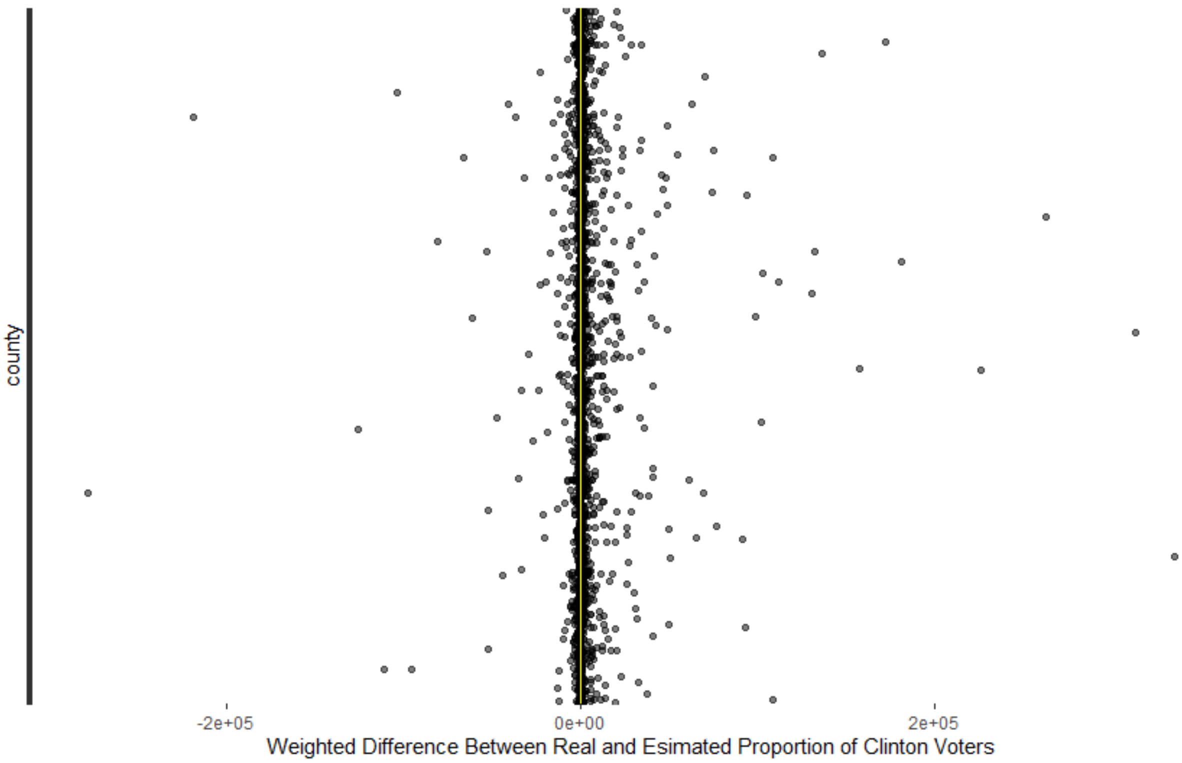
As for the prediction for each country, we will use San Diego as an example to see if our predictions worked. As you can see from table 24, Hillary Clinton had been the actual candidate from San Diego county. You can also see that we had correctly predicted in our model that Hillary Clinton would be the winning candidate for San Diego county. This is expected, as San Diego County disproportionately votes democratic, so although Trump voters are not properly and representatively polled, it would not make a difference in this county. Most of the errors we saw from the 30.49% of false Clinton counties would be in regions where the percentage margins are a lot closer, as compared to the result-accurate SD county.

Task 2:

To predict the popular vote, we devised a regression tree model that would predict the proportion of Clinton voters in each county. We took the approach of creating a large initial tree, and then pruning the tree via cost-complexity pruning, with a cross-validation approach used to determine the optimal tuning parameters. The pruned tree is shown below:



As can be seen, the variables most often used to split the data were White, Transit, and Professional, with a few other variables being used only once. With the predictions from this tree, we computed the national popular vote for each candidate by applying this proportion to the voter population of each county. Our results were surprising- we predicted about 59,463,689 votes for Clinton, and 69,333,742 votes for Trump. This is, of course, inconsistent with reality, where Clinton convincingly won the popular vote. To investigate, we created a plot of the difference between the real Clinton proportion and our estimate, weighted by population:



We see significantly more counties with a large positive difference than a large negative one. So, our estimates tended to underestimate the proportion of Clinton voters in large counties, which explains our large undercounting of Clinton's votes and overcounting of Trump's.

Discussion

From what we have understood from our general insight from Project Stage 1 and our introduction, we start to recognize some of those ideas playing out in our report. The 2016 election was one where pollsters made very inaccurate predictions from their models and surveys, which resulted in massively underrepresented Trump voters. Looking at the results from Task 1, in the example of San Diego County, the model we used to predict county-level winners was correct. Given all of the variables that we found as the most important from regression methods, we predicted that Clinton would be the most popular candidate in SD county, which she actually was. However, this was not reflected by a lot of other counties who we would've predicted for Clinton. Looking at misclassification tables and various charts provided, almost a third of all predicted Clinton-majority counties ended up seeing results that favor Trump instead. On the other hand, most predicted counties in favor of Trump ended up being true (97.52%). Although Clinton was more popular in the 2.48% counties she was not expected to win, that rate could not compare to the counties she was predicted to win but actually lost. This makes sense, as overall, Trump won by large margins, especially in tossup states (on average, 7%) which is what we learned from extensive poll analysis mentioned in Project Stage 1, Part 0. Taking another analysis, we sought to predict the popular vote in Task 2. Some of the most important variables used in our pruned decision tree are white/minority, unemployment, transit and income per capita. Seeing as these factors are more indicative and create a larger margin of voter disparity in rural areas rather than urban areas, this led to an overcount of Trump's voters. In order to explain this effect, it was important to understand why Clinton's voters were so undercounted. The plot explaining the weighted difference determined an underrepresentation of larger counties, which tend to be an urban center (or main metro areas) where Clinton voters are decidedly the majority. This is much different from what we saw in our first task, but regardless, the popular vote isn't the statistic that our country uses to predict the overall winner of the presidency. The popular vote could not take into account who lives where, population densities or political boundaries, all of which create fair adjustments and representation.

At the end of the day, our two tasks sought to predict county winners and popular vote, but neither were exactly what we would have expected. Task 1 created an overrepresented Clinton voters by county, many of which were lost when it was not expected to be. On the other hand, Task 2 would suggest a win for Trump since we predicted his overall popularity: he did end up winning in that regard (by electoral college stratification) but contrary to our model, he lost the popular vote. These two varying (even contradicting) attempts at the 2016 election analysis exemplifies the difficulty in creating polls, and then formulating a certain model or statistic that helps us decide who wins, by what margin, and by what disparity. Understanding all of these interpretations, we now see how and why everyone was shocked when Trump ended up taking the presidency.