UNIVERSITY OF CALIFORNIA

SANTA BARBARA

DEPARTMENT OF PROBABILITY AND STATISTICS

DATA MINING FINAL PROJECT

2016 PRESIDENTIAL ELECTION ANALYSIS

Author:

Kaitlyn Boyle (PSTAT 131)
Perm Num. 8595332
Rahul Kasar (PSTAT 231)
Perm Num. 9599333

Supervisor:

Alexander Franks

June 14, 2018

Table of Contents

.... 1

2016 PRESIDENTIAL ELECTION ANALYSIS 1

Introduction 3

Background 3

Data Wrangling 6

Visualizaton 8

State Winner 8

County Winner 9

Minority vs Income 10

Transit vs Employed 11

Professional vs Services 11

Census.ct 12

Dimensionality Reduction 12

Clustering 14

Classification 15

Decision Tree 15

Logistic Regression 18

LASSO Regression 19

ROC Curve 20

Further Exploration 22

Random Forest 23

Support Vector Machines 25

Classification Tree 26

Logistic Regression 28

LASSO Regression 29

Conclusion 30

References 31

Introduction

The 2016 election was interesting and difficult to predict. Many statisticians could not predict the outcome, or predictions were wrong. Many predicted that Hillary Clinton would win in polls throughout the campaign, forecasts, and even exit polls. Donald Trump's victory was a surprise to many. Statisticians have been trying to understand what happened to make the polls miss. In this project, we are interested in looking at two datasets: an election dataset that includes county, state, and federal voting information, and information about the candidates that ran for President in the 2016 election. Our second dataset is census information, which includes county and state information on demographics, including race, income, mode of transportation, and form of employment amongst other demographics. We are interested in learning about the election information and discovering the important factors that help predict how people vote. We will do this through data visualization, dimensionality reduction, clustering, classification, and we will do further exploration of the datasets to discover exactly which factors are best at predicting voting behavior in swing states, red states, and blue states.

Background

One of the biggest issues in predicting voter behavior and election forecasting is sampling and polling error. The random sample chosen to predict the results may potentially have a higher concentration leaning one way or the other politically. This leads to a skewed sample which results in a skewed prediction. This sort of error can be estimated from past election results as this problem always persists. In addition, response rates to polls has been decreasing rapidly in recent years. If less people are responding, the people or companies conducting the polls are not getting the required information they need to construct accurate polls. We also know that there is response bias, meaning that the people who do respond to polls are typically passionate about the issue, and care more about the topic than the average person. They may be outspoken constituents of a particular area so they will respond with exaggerated answers. A problem that may have manifested in the 2016 election is untruthful responses from responders. Since the two candidates were so divided, voters for Donald Trump may have felt uneasy expressing their support for him and proceeded to lie in polls. We learned from Nate Silver's (2016) article "Why FiveThirtyEight Gave Trump A Better Chance than Almost Anyone Else", Trump won the election from undecided and late deciding voters. These would have been missed in the polls because they would not have responded, giving Clinton a lead in the polls before election day. This is referred to as non-response bias.

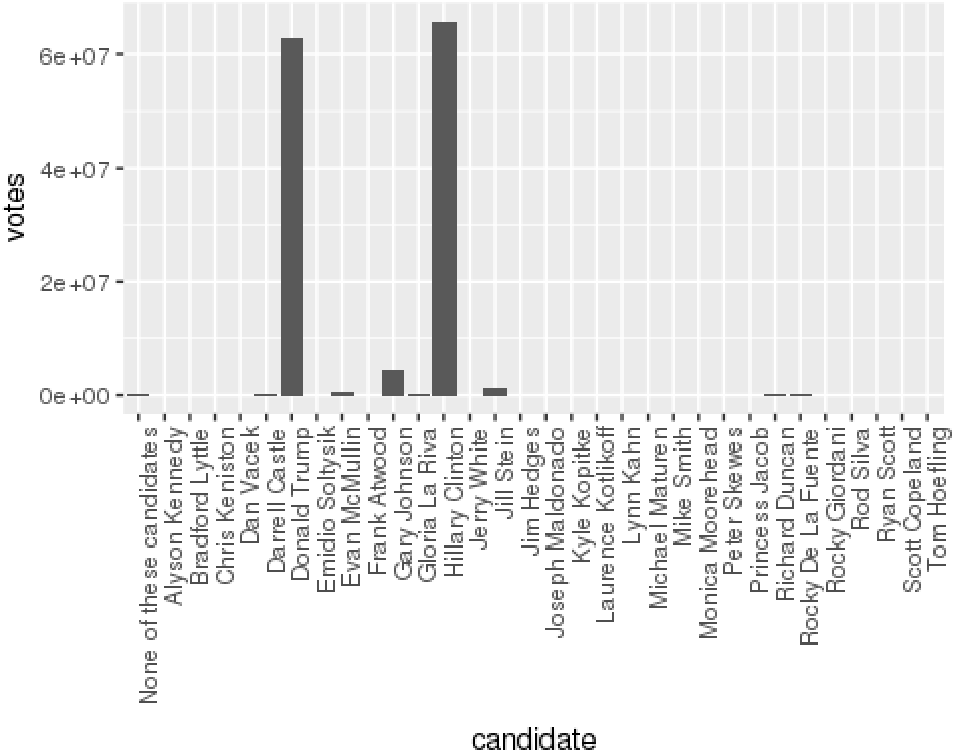
The problems discussed above are reducible error, error that can be estimated and factored into the prediction. A bigger issue with polls is the systematic error or the irreducible error. This error is either a flaw in the model or failure to understand the community sentiment on particular issues. Some polls hope that due to the variety of regions, systemic error will cancel each other out. But if they are correlated, it could add up to a very strong error.

Nate Silver gave Donald Trump higher odds that most other statisticians, although he still predicted that Clinton would win. According to the same 2016 article by Silver, his final model gave Trump a 29% chance of winning, compared with the New York Times model giving 15% chance of winning. Nate Silver's methodology differed in several ways. He based his model off the "accuracy of polling measures dating back to 1972" (Silver 2016). This means that his model reduces polling error unlike the other comparable models. His model also considered the possibility of Clinton underperforming in states in the Midwest and the Rust Belt, states that ended up being big deciders for Trump's win. The most important reason Silver's model gave Trump a higher chance of winning was because of their "assumption that polling errors are correlated” (Silver 2016). Polls in the same state will miss in the same direction, and states of similar demographics will also miss in the same direction. Silver’s model also differed in that it involved placing a heavy emphasis on undecided and late deciders for the election. This is non-response bias, and it is a difficulty that is hard to account for.

The polls in the 2016 election were largely wrong. They all gave Clinton a better chance of winning, and according to Silver, even Trump’s own campaign only gave him a 30 percent change of winning ahead of the election. There are many reasons why polls in 2016 were bad, and much of the reason was polling error. Some states had weak polls, meaning they either had too few, not recent enough, or not good enough, all factors which increase error. Also, if a state poll misses, its forecast will miss in the same direction. This is because polling error is correlated so the error will not cancel out, it will only increase. Another reason polls were bad is due to how polls were sampling. According to Carl Bialik and Harry Enten in their article “The Polls Missed Trump. We Asked Pollsters Why”, people voting for Trump were embarrassed to report this, and Trump performed better in the polls where people were responding on the phone to an automated voice, rather than a live person (Bialik and Enten 2016). Also, the polls were bad because there was a lower than expected turnout, especially amongst Democrats. Trump also performed better with people who were late deciders, or who had been supporting a third-party candidate until the final election. There is also a problem with media budgets not commissioning for polls. With a smaller budget, media companies cannot pay for as many polls, which increases error. With more polls, you are seeing better predictions and lower error. Journalists tend to communicate election forecasting models to a general audience in terms of percentages and odds of winning. They may report that Clinton is rising in the polls, for example, without going into detail about how that prediction was reached or how the models were made. The general audience does not want to know about the prediction model, and only wants to know what the forecasts are. We think that journalists tend to only report the percentages, and if the candidate is rising or falling.

Data Wrangling

To begin our project, we must make our raw data easier to work with. We start out by splitting up our election data into three different sets: federal summary data, state summary data, and county data. We named each of these new data frames election_federal, election_state, and election, respectively. We found that there were 31 candidates total in the 2016 presidential election. Below you can see a bar chart of all of the candidates and how many votes they received in the popular vote of the election.



We can see that Hillary Clinton had the most popular votes, followed by Donald Trump, Gary Johnson, Jill Stein, and Evan McMullin, respectively. The rest of the candidates had no votes, or too few votes to be shown on the bar chart. We then created new variables county_winner and state_winner to find which candidate had the most votes in each county or state. We put these winners into a new column in the election_state and election data frames. We selected the winner by finding the candidate who had the highest proportion of votes in each state or county. This will help us to create a visualization for which candidate won in each state.

State Winner

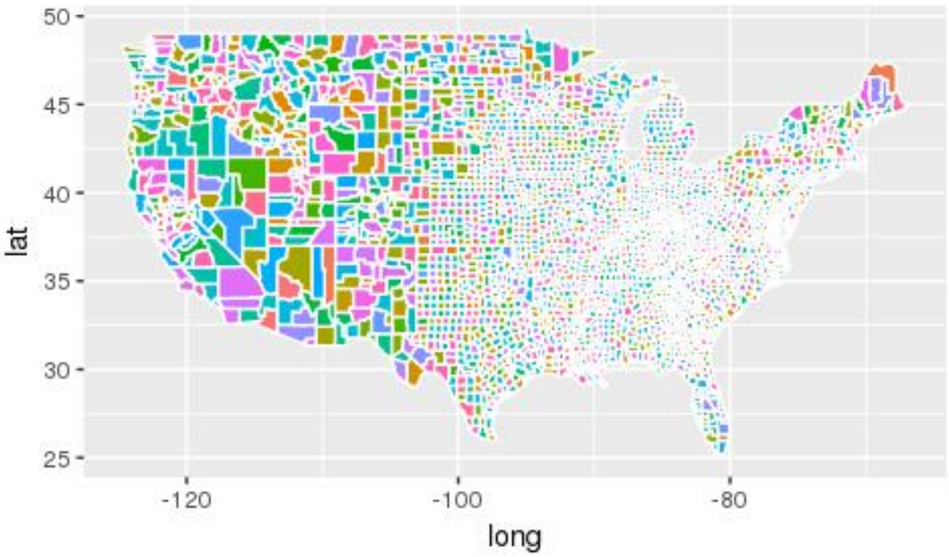
county	fips	candidate	state	votes	total	pct
NA	CA	Hillary Clinton	CA	8753788	135691978	0.0645122
NA	FL	Donald Trump	FL	4617886	135691978	0.0340321
NA	TX	Donald Trump	TX	4685047	135691978	0.0345271
NA	NY	Hillary Clinton	NY	4556124	135691978	0.0335770
NA	PA	Donald Trump	PA	2970733	135691978	0.0218932
NA	IL	Hillary Clinton	IL	3090729	135691978	0.0227775

County Winner

county	fips	candidate	state	votes	total	pct
Los Angeles County	6037	Hillary Clinton	CA	2464364	135691978	0.0181615
Cook County	17031	Hillary Clinton	IL	1611946	135691978	0.0118794
Maricopa County	4013	Donald Trump	AZ	747361	135691978	0.0055078
Harris County	48201	Hillary Clinton	TX	707914	135691978	0.0052171
San Diego County	6073	Hillary Clinton	CA	735476	135691978	0.0054202
Orange County	6059	Hillary Clinton	CA	609961	135691978	0.0044952

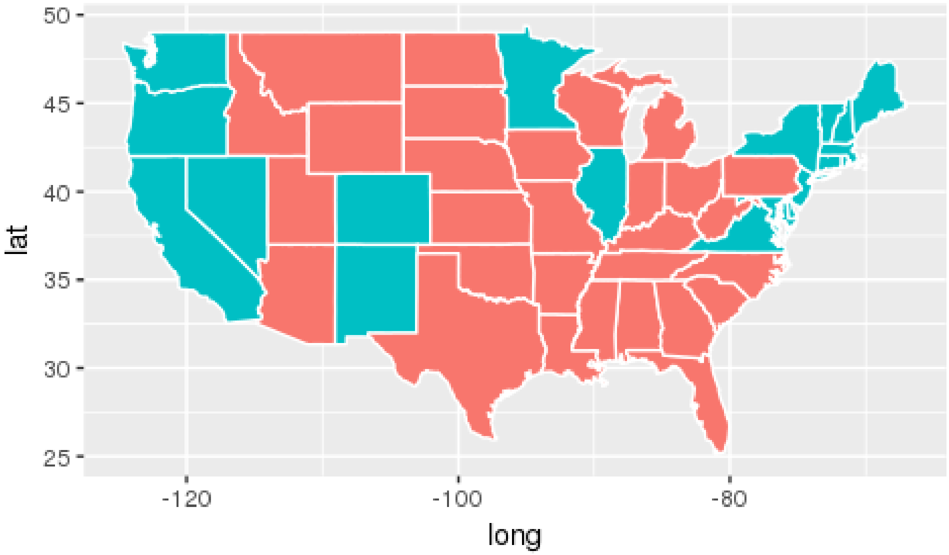
Visualizaton

We first want to visualize the map of the United States by each county. We used ggplot and the counties data frame to create a map of the United States colored by county, as seen below.



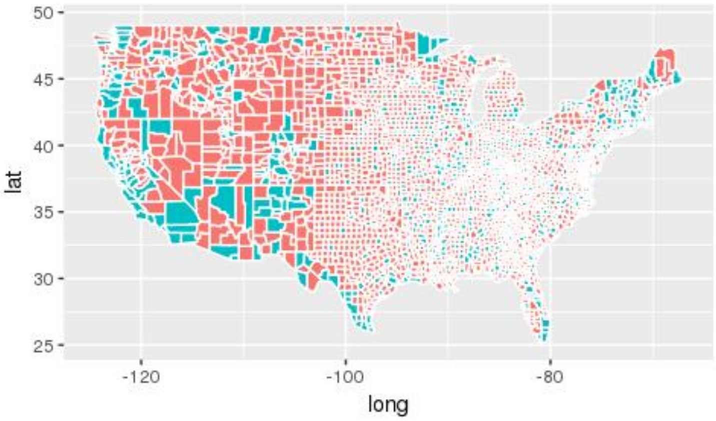
State Winner

We then wanted to visualize which candidate won in each state. To do this, we had to combine the states data frame with the state_winner variable we created earlier using the left_join() function. From here, we used ggplot again to create a map of the United States and colored by the winning candidate in each state, as seen below. The red states are where Donald Trump won, and the blue states are where Hillary Clinton won.



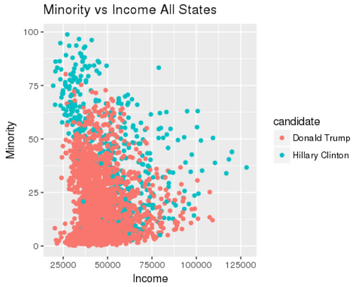
County Winner

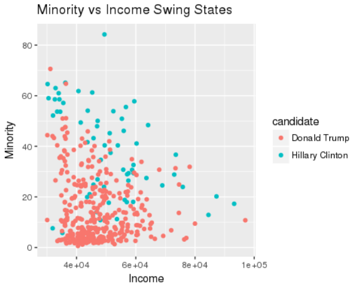
Our next step was to visualize by county winner, using the same method we used to visualize the state winner. We can see this graph below. Again, the red counties are where Donald Trump won, and the blue states are where Hillary Clinton won.

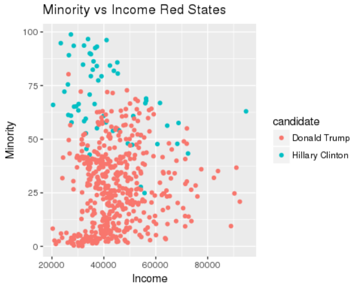


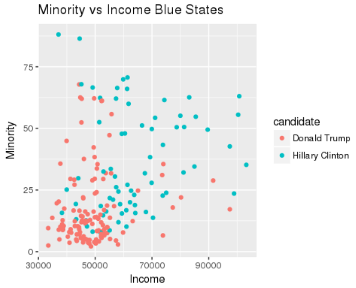
For our visualizations, we will be trying to see what the important factors are in swing states. We created separate datasets swing, red, and blue, in addition to the dataset containing all states. Note, the 3 new data sets contain a smaller subset of each state classified as swing, red, and blue, where a swing state is one that could go either Democrat or Republican, a red state generally votes Republican, and a blue state generally votes Democrat. We are interested in testing variables Minority vs Income, Transit vs Employed and Professional vs Service and seeing how swing states behave in these relationships.

Minority vs Income



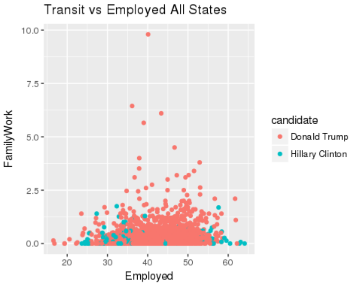


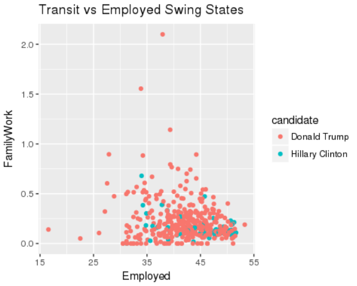


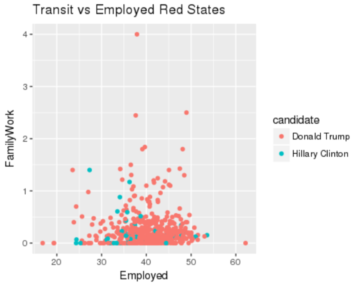


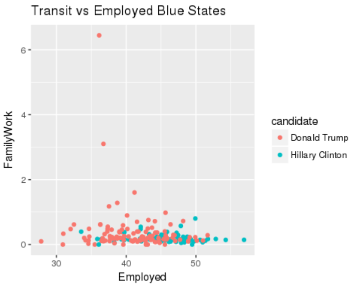
From this chart, we see that there is some separation between classes for Minority vs. Income. In the swing states, we see poor white people are likely to vote for Trump and low income high minority. In swing states we do see high income, high minority swing towards Clinton. In red states, you need to have over 65% Minority to vote blue. In the blue states, we see high income, and these people are more likely to vote democrats. Wealthy minorities are more likely to live in big states like California and New York.

Transit vs Employed


Transit vs Employed All States

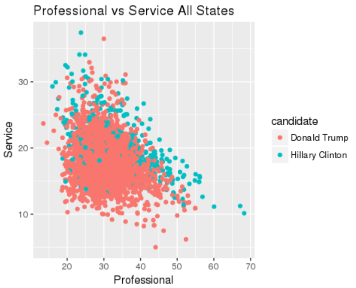

Transit vs Employed Swing States

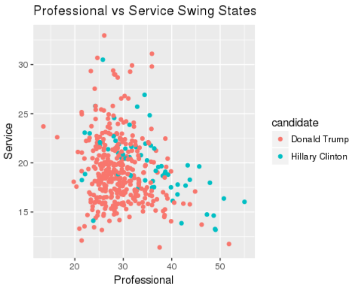

Transit vs Employed Red States

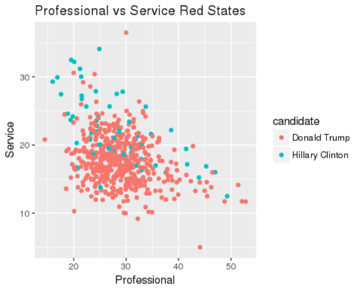

Transit vs Employed Blue States

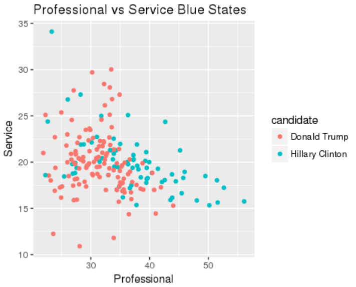
Swing state behavior on these two predictors is not that different compared to the rest of the states. We see that Employed and FamilyWork are evenly split.

Professional vs Services


Professional vs Service All States


Professional vs Service Swing States


Professional vs Service Red States


Professional vs Service Blue States

In swing states, there are a lot more service workers, and they voted strongly for Trump. We think this is a major indicator of where the election was won as Trump seemed to flip a lot of these service workers to him. In blue states, as expected, there are a lot more professionals who swayed the vote for Clinton.

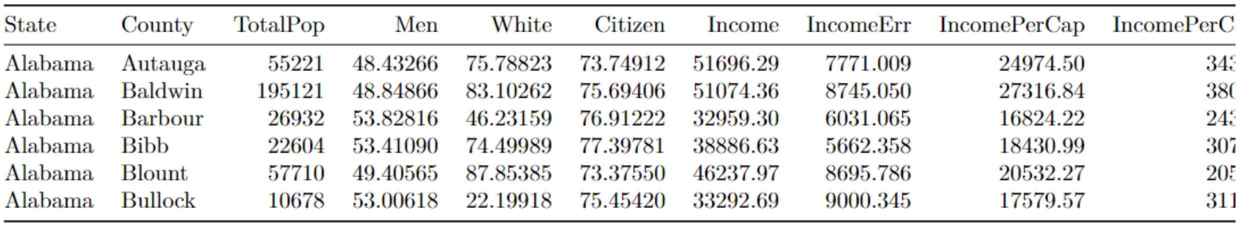
Our next step was to clean up the census data. We created a new data frame called census.del, which is the cleaned-up version of the census data frame. We did this by filtering out rows with missing values, converted the columns of Men, Employed, and

Citizen to percentages, created a Minority variable combining Hispanic, Black, Native, Asian, and Pacific, and we removed the variables Walk, PublicWork, and Construction.

We then summarize the subcounty data by creating a new data frame, named census.subct. We use data from census.del, and group the data by state and county. We used add_tally() to find the CountyTotal, which is the aggregated data for each county. We then computed the weight by dividing TotalPop by CountyTotal, which is the weighted average of each attribute for each county.

Next, we create one more new data frame called census.ct. We started with census.subct, and used the function summarize_at() to compute the weighted sum for the county census data. We printed the first few rows of census.ct below.

Census.ct



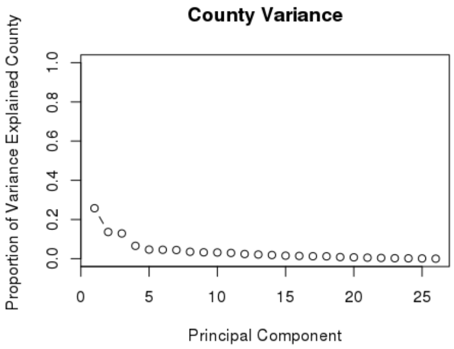
State	County	TotalPop	Men	White	Citizen	Income	IncomeErr	IncomePerCap	IncomePerC
Alabama	Autauga	55221	48.43266	75.78823	73.74912	51696.29	7771.009	24974.50	34:
Alabama	Baldwin	195121	48.84866	83.10262	75.69406	51074.36	8745.050	27316.84	38(
Alabama	Barbour	26932	53.82816	46.23159	76.91222	32959.30	6031.065	16824.22	24:
Alabama	Bibb	22604	53.41090	74.49989	77.39781	38886.63	5662.358	18430.99	307
Alabama	Blount	57710	49.40565	87.85385	73.37550	46237.97	8695.786	20532.27	20!
Alabama	Bullock	10678	53.00618	22.19918	75.45420	33292.69	9000.345	17579.57	311

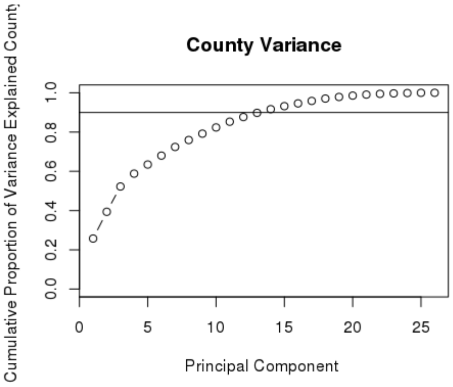
Dimensionality Reduction

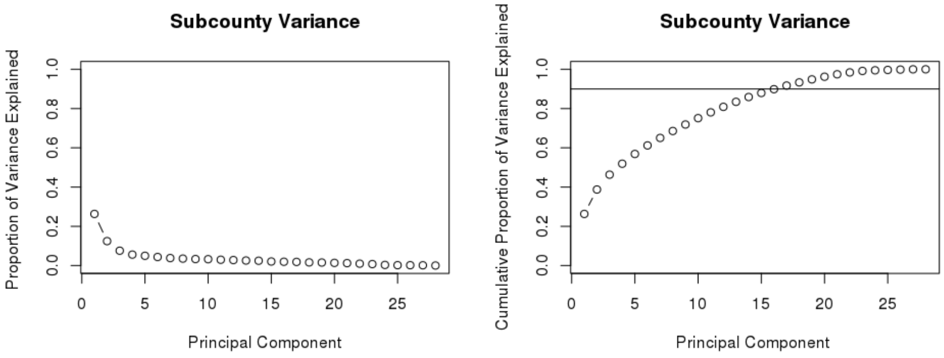
Principal component analysis is used for dimension reduction which is useful in identifying the directions of highest variance as well as identifying which predictors have the strongest impact in those high variance directions. PCA helps in throwing away noise and extracting important features. Dimension reduction can also help visualize high dimensional data using 2 or 3 principal components. It is important to scale and center data while conducting PCA, otherwise results will be incorrectly interpreted. We chose to scale because spreads of predictors are different in different states and counties. For example, New York has a very large spread of income compared Kansas where the income level is similar across the state. By scaling, predictors with higher spreads do not have as dominant of an impact they would have had if they were not scaled. If the data was not centered, the first principal component would correspond to the direction of the mean, instead of the direction with the most variance.

The first principal component shows which direction has the most variance. For sub-county, the highest absolute loadings for the first component were Poverty, Income and Professional. These predictors all pertain to information regarding jobs within a sub-county. The highest absolute loadings for the first PC in county data were Poverty, Income, and Employed which all have to do with money and money spread with a county. The second principal component is the direction with the second most variance. Transit, Minority and Drive were the highest loadings for sub-county, which means the direction has a lot to do with how people get to work in their county. Self Employed, TotalPop, and White are the highest loadings for county which is quite different than what was found for sub-county.

According to the figures below, which plots the Cumulative Proportion of variance explained, 13 principal components explain 90% of the county data and 16 principal components explain 90% of subcounty data.

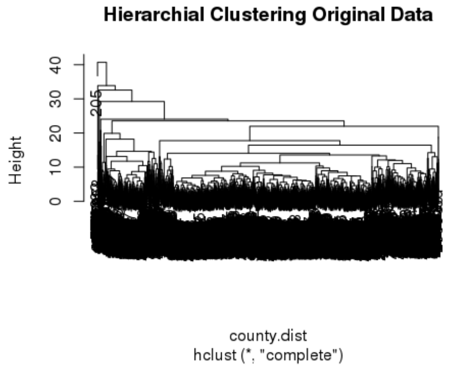


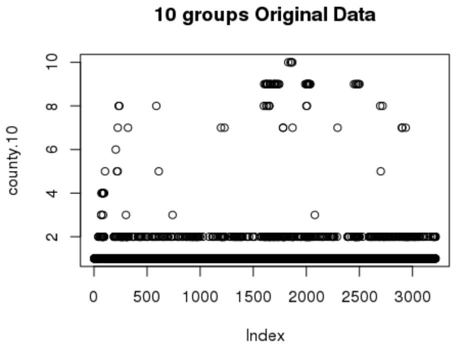


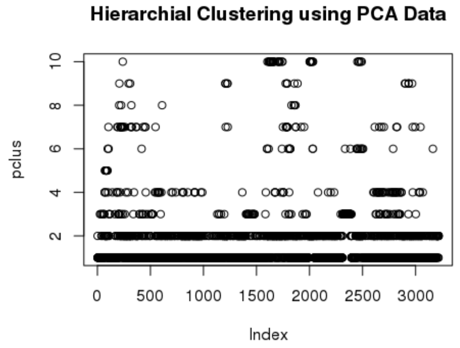


Clustering

Hierarchical clustering is an unsupervised learning clustering method that considers distances between individual observations to group them together. We ran hierarchical cluster using complete linkage on our original data and cut the resulting output into 10 groups. We then repeated the process with the first five principal components. We were interested in which method classified San Mateo County better, and to determine this, we considered the distance from the cluster mean. We found the cluster mean by taking the average of all observations for both groups and then computing Sum of Square Errors as our distance metric. Making sure to scale the variables and dividing by N, we found that the distance to the cluster center for the cluster created from the original data put San Mateo in a better cluster because it had a smaller sum of square error. A look at the cluster shows that a lot of nearby counties, such as Alameda, Marin, and Santa Clare were also placed in this group. These counties have very similar demographics across the board and vote similarly. In the cluster from PCA, these other counties were not placed in the group. The sum of squares error for the normal group was 1134293559, and the sum of squares error for the PCA group was 12112405811.



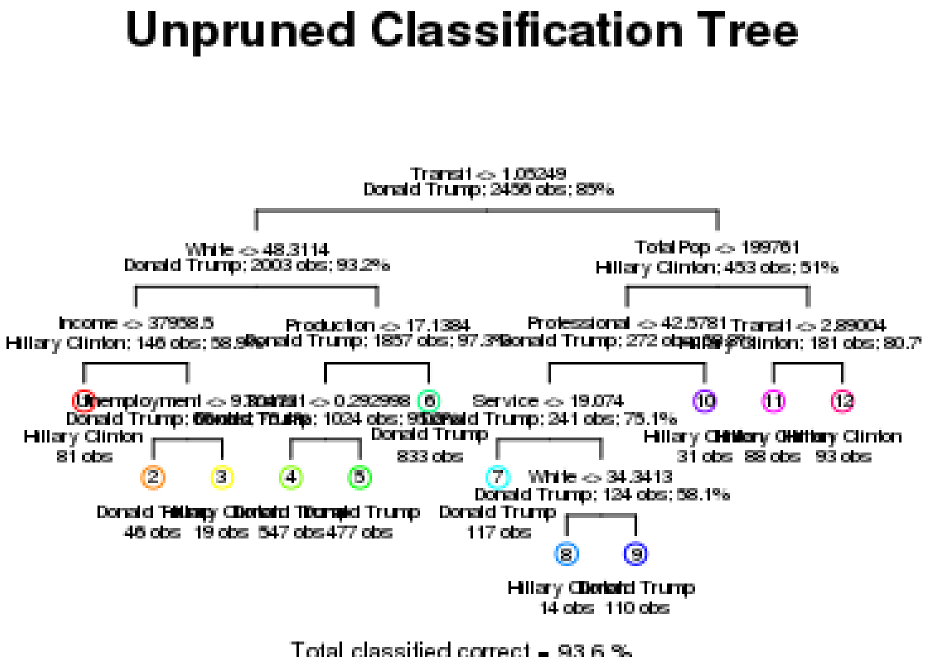




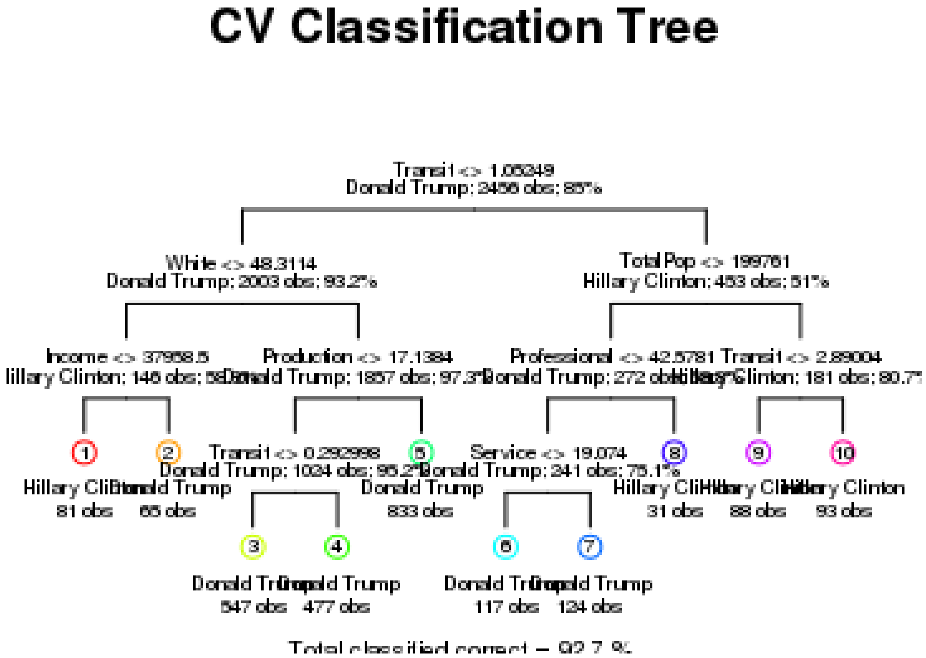
Classification

Decision Tree

The decision tree classification algorithm is a very interpretable classification method that splits the segment space of outcomes into many rectangular regions. The algorithm conducts a greedy step by step splitting of the segment space in which it makes the split that reduces the classification error in the short run. The figure below is the unpruned tree, which does a great job fitting the training data.



Unpruned decision trees have very high variance as slight changes in the data lead to big changes in prediction accuracy. We ran cross validation to see that a tree with 10 leaf nodes is the optimal size. The optimal tree is pictured below.



The most important result from the tree classification is the importance of the variable 'Transit'. At first, this does not seem like the most important variable in determining the way counties vote, but it makes sense after further thought. Conservative voters, even the ones who are poor, generally own their own vehicle. This is necessary as a lot of them live in rural areas that vote Republican the majority of the times. In contrast, many urban places have high concentration of liberal voters as well as public transportation. Also, poor liberal voters are more likely to take public transport than poor conservative workers. According to the algorithm, transit does the best job of initially dividing the outcome space. Percentage of White and TotalPop are the next important predictors. Going down the tree, educated professionals, and poor minorities are more inclined to vote for Hillary. People who are white, not educated but work in the service and manufacturing industry are more likely to vote for Trump. It is interesting that seemingly key variables like poverty and employment rate are not in the tree, but it is essentially included in the Transit split and Professional Splits. We include the records matrix so far with the training and testing error of the classification tree.

Records Matrix

	train.error	test.error
tree	0.0728827361563518	0.0912052117263844
logreg	NA	NA
LASSO	NA	NA

Logistic Regression

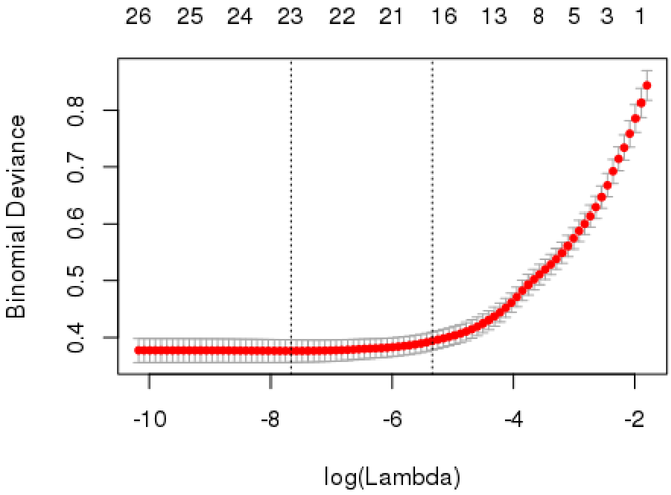
Logistic regression differs from Decision Trees in that it is a soft classifier. Logistic regression returns the odds of a county voting for Hillary Clinton. A look at the summary of our fit shows that the highest significant coefficient is FamilyWork which measures unpaid family work. The way to interpret this is if percentage of FamiyWork increases, the odds that county votes for Clinton decreases by $e^{-1.745e-02}$. This is essentially a measure of stay at home parents, a luxury that a lot of conservative families can afford and participate in. What is interesting is that FamilyWork was not a key variable in our decision tree at all. In fact, Transit which was the most important variable in the tree, is not significant at all in our logistic regression. Another important predictor that decreased the odds of voting for Clinton was the income level of the county and percentage of white folk in the county. Predictors that increased the odds for Clinton were Carpool and Professional. So even though the variables in the logistic regression chose different important variables, the story remains the same. Professionals and people who lived in urban areas (Carpool) are more inclined to vote for Clinton. White people and people who work in the service and production industry are more likely to vote for Trump. We include our updated records matrix below, with the addition of training and test error for the logistic regression.

Records Matrix

	train.error	test.error
tree	0.0728827361563518	0.0912052117263844
logreg	0.0639250814332247	0.0765472312703583
LASSO	NA	NA

LASSO Regression

LASSO regression penalizes overfitting of normal regression models by adding a penalty term to coefficients. This helps with overfitting because when a model has large coefficients, this is a sign that it will not perform well with new data. LASSO regression has the advantage over Ridge regression because it can make certain coefficients zero. LASSO estimates are ‘sparse’ because it eliminates predictors that are not useful for predictions. We used cross validation to determine that 0.0002945269 is our best lambda, as shown by the graph below.



We used the best lambda to fit our model and evaluated the coefficients and see that predictors like Minority, Child Poverty and Self Employed were exactly 0. Other predictors like Income and TotalPop were very close to 0. Predictors that kept a high coefficient were Family Work, Unemployment, Professional and Drive. The values of these coefficients did not change significantly. The major significance is that LASSO regression tells us that Minority, Child Poverty and Self Employed are not important in determining which way a state will vote. This is interesting because poverty and minority statistics seems like major talking points during elections but according to our data, they are not significant predictors. Both conservatives and liberal factions face poverty so it is not a good determinant of the way the county voted. We include our updated records matrix with the LASSO regression training and test errors below.

Records Matrix

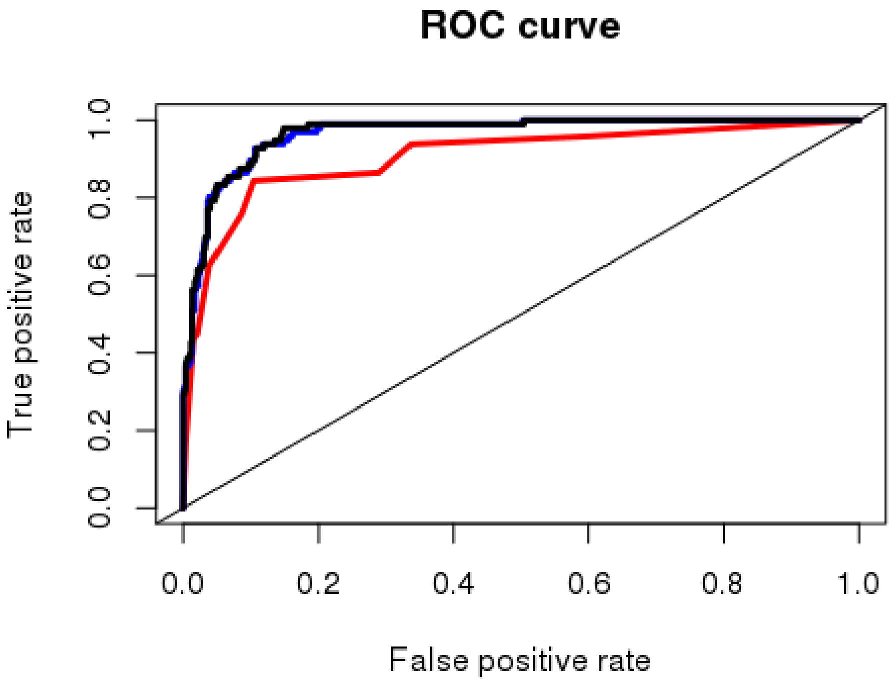
	train.error	test.error
tree	0.0728827	0.0912052
logreg	0.0639251	0.0765472
LASSO	0.0651466	0.0781759

ROC Curve

The figure below shows the ROC curves for the 3 different models that we have fit. ROC curves measure the True positive rate vs False positive rate where Clinton is the positive. The main problem with decision trees is that it has very high variance. It is very likely that the same decision tree will not perform well with new test data. The main point of classification is to help predict new elections, so a single decision tree will not perform well on the 2020 presidential election. However, decision trees give very interpretable results that can be helpful in determining why a county votes a certain way. We think this variability can be solved with bootstrapping and the fitting of a random forest, a problem discussed in the Further Exploration section.

The strength of logistic regression is that it is a soft classifier which is valuable because it gives predictions some margin of error. When a publication comes out and says a certain county or state *will* vote for Clinton, when that prediction is incorrect, the publication is ridiculed for being wrong. However, if the publication says that a county will vote for Clinton at a certain probability or odds, they add more margin of error for incorrect answers. Soft classifiers also give a more interpretable classifier than a simple classifier. If the probability of a county voting is split 51 Democrat, 49 Republican, the interpretation is that the county could go either way in the actual election. Hard classifier would simply state the the county would vote democrat.

Logistic regression suffers from the overfitting problem as coefficients for the predictors could get very large which leads to high variance in the predictions. To deal with this overfit, LASSO regression adds a penalty to the coefficients. This helps eliminate coefficients that are determined to be useless. However, in a model as complex as election data, the complete elimination of coefficients could be risky as the predictors could actually have some value. LASSO stated that poverty is not an important predictor, but in an election a politician's attitude on poverty is a major point of interest. Users of the model have to be careful and not blindly follow its results. Ultimately, LASSO performs the best out of these three models on the test data as evidenced by the highest Area Under the Curve (AUC table), followed by logistic regression and decision trees. The red ROC curve is the classification tree, the blue ROC curve is the logistic regression, and the black ROC curve is LASSO regression.



AUC Table

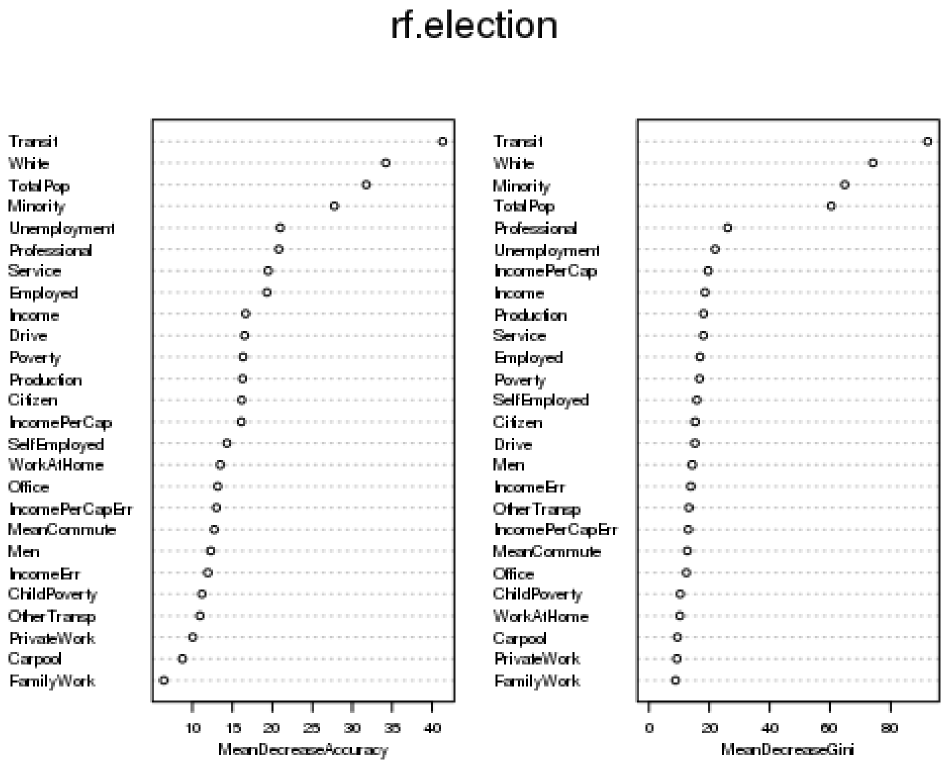
	AUC
tree	0.9050233
logreg	0.9639237
LASSO	0.9653113

Further Exploration

We are interested in learning about how the swing states, red states and blue states differ in what factors are most important for predicting the winner. We define a swing state as a state in which the winner could be either a Democrat or a Republican, and the expected winner is not determined until the election. A red state would be a state in which a Republican almost always wins, and a blue state would be a state in which a Democrat almost always wins. Elections are determined in the swing states. We picked the largest swing states, blue states, and red states with the most electoral college votes because it would be easier to notice trends. Swing states were especially important in the 2016 election, and according to Harry Enten in “Almost Every Swing States is a Must Win for Trump Now”, he states that if Trump were to lose certain swing states, he would most likely lose the election (Enten 2016). We saw in the results of the election that Trump won the swing states. We decided to group the states into new data frames. We included California, Washington, New York, and Oregon as blue states; Texas, Georgia, Kentucky, and Virginia as red states; and Michigan, Ohio, Florida, Pennsylvania, Nevada, and North Carolina as swing states. We knew through prior knowledge which states to include in these categories. We used data from the election.c data frame because it included both election data and census data. We wanted to know which variables were most important in determining how the state overall voted. We are interested in seeing the change in coefficients between our original models and our sub-setted models. We explored this through support vector machines, decision boundaries, random forests, classification trees, logistic regression, and LASSO regression.

Random Forest

We started out by creating a random forest to find out which variables were important in predicting an election winner. The random forest reduces variance through bootstrapping, and decorrelating the trees after bootstrapping. We know that if there is a very strong predictor, it will be the first split on every tree. We subsample predictors, which is how we decorrelate the trees. This is how we get larger variance reduction when bagging. Our random forest included all 50 states plus Puerto Rico and Washington D.C. We used our training data set and regressed candidate on all other variables and plotted a random forest with misclassification error and the Gini index, as seen below.



We can see that the Gini index does a better job of telling us which variables are most important. In the misclassification error plot, we have about six variables of importance: transit, white, total population, minority, unemployment, and professional. In the Gini index plot, we have reduced the variables of importance to four: transit, white, minority, and total population. This tells us which variables are most important in determining an election winner over the whole US. We also see that the error rate for the random forest is very low, which tells us that this is a good classification model for learning which variables are most important in predicting an election winner. We add a fourth line to the records matrix and include the training and test error for the random forest. We add a fifth line as well, to add the SVM training and testing error when we get them, which will be our next step.

Records Matrix

	train.error	test.error
tree	0.0728827361563518	0.0912052117263844
logreg	0.0639250814332247	0.0765472312703583
LASSO	0.0651465798045603	0.0781758957654723
Random Forest	0.0012214983713355	0.0456026058631922
SVM	NA	NA

Support Vector Machines

We then used support vector machines (SVM) to create decision boundaries on all variables. The SVM is good to use here because it is a classifier that introduces some slack, or in other words, allows some observations to be misclassified. This makes the SVM a soft margin classifier and can lead to better overall classification. Our SVM classification was on all 50 states plus Puerto Rico and Washington D.C. We created a prediction table with the SVM classifications using training data. We then regressed candidate on all variables to try to find a decision boundary. However, our plot was uninterpretable, and we did not learn anything new from the plots. We decided to move on from SVM classification, because the results we were getting were not helpful in determining the most important variables. We calculated the testing and training error rates of the SVM and found that the error rates were very low. This shows us that SVM is a good classification model to use for this type of data, but the graphs were unreadable and did not tell us much about the important variables in this case, and we decided to move on to other types of classification. We include the records matrix below with the addition of the SVM training and testing error. We can compare all of these types of errors because we used the full dataset with each of these types of classification.

Records Matrix

	train.error	test.error
tree	0.0728827	0.0912052
logreg	0.0639251	0.0765472
LASSO	0.0651466	0.0781759
Random Forest	0.0012215	0.0456026
SVM	0.0175081	0.0537459

Classification Tree

Our next step was to subset the data into swing states, red states, and blue states. We created test sets and training sets to perform logistic regression and LASSO regression as well as fitting cross validation classification trees. We fit three different classification trees, one each for swing states, red states, and blue states. We found that the best number of splits for each tree was nine. The first split for each category was transit. The classification trees help us visualize the demographics of how people vote in each type of state. For example, in the swing states, we can see that if the rate of people taking public transportation is low, and has a low population of white people, they will vote for Hillary Clinton. If there is a low rate of people taking public transportation, a high rate of white people, and people working in an office, they will vote for Trump. We can follow the classification trees to see the demographics and predict what type of people vote for which candidate in each type of state.

Figure 1: Swing States

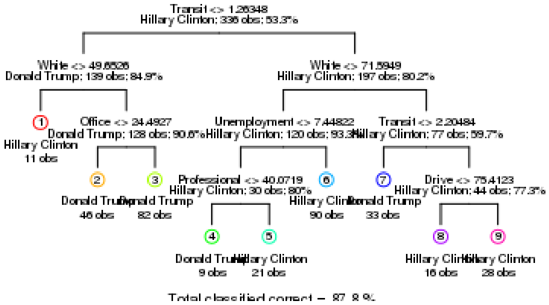


Figure 2: Blue States

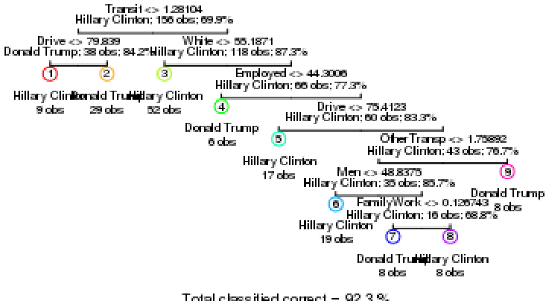
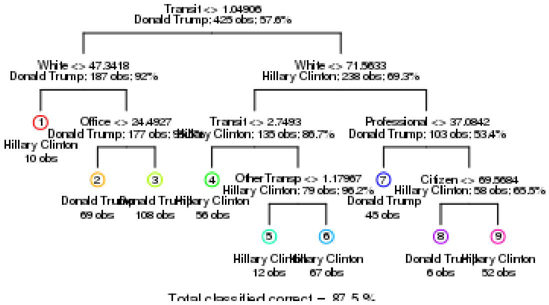


Figure 3: Red States



Logistic Regression

We then moved on to the fitting logistic regressions for each type of state. We compared the coefficients from the logistic regression of the subset data frames with the logistic regression of all the states together. We learned that in swing states, variables such as minority, transit, and carpool have higher coefficients and higher significance levels than the coefficients for all the states. We see that family work and professional are no longer significant as they are with all the states. Predictors such as unemployment and service remain significant across both regressions. We observed in the blue states that predictors such as professional, service, and production had large coefficients and has significant p-values. Many of the predictors related to mode of transportation were significant. Income had a very small coefficient. The fact that many of the predictors seemed to be important here showed us that the predictors in a blue state do not matter as much, because people in a blue state will vote Democrat no matter what. This could be because of high population rates, societal pressures to vote Democrat in these states, the amount of Democrat propaganda versus Republican propaganda in blue states, and differing beliefs and values in general in blue states. When we performed logistic regression on the red states, we learned that carpool, family work and white population are very significant predictors. The family work predictor is the rate of unpaid family work, which we interpreted as the percentage of people who have at least one stay-at-home parent. The fact that white population was also another important predictor showed us that the type of people who can afford family work are wealthier white people. This makes sense as a demographic that would typically vote Republican. Overall, the logistic regression gave us a lot of information based on the coefficients about which predictors are most important in each type of state.

LASSO Regression

Our last step was to perform LASSO regression on the swing, red, and blue states. When we performed the LASSO on the blue states, many of the coefficients went to 0. The positive coefficients included poverty, service, employed, and unemployment, meaning these were the most important predictors of who would vote for Hillary Clinton in a blue state. These predictors make sense for demographics of people who would vote Democratic. The negative coefficients were men, white, office, and drive, meaning the higher the rate, the less likely to vote for Hillary Clinton. These also make sense of demographics of people who would vote Republican in a blue state. For the red states, the positive, highest coefficients were citizen, professional, service, production, work at home, employed, self-employed, and unemployment. These make sense for people who would democratic in a red state. The negative coefficients white, office, drive, carpool, other transportation, and family work. These make sense for people who would vote Republican in a red state. For the swing states, the highest positive coefficients were professional, service, employed, and unemployment. The highest negative coefficients were men, white, office, drive, carpool, mean commute time, self-employed, and family work. Many of these predictors imply wealthy, white males, which make sense as a demographic who would vote Republican in a swing state.

We learned from the coefficients in logistic regression and lasso regression that the predictors almost do not matter in blue states, because they are going to vote Democratic no matter what demographics they fall into. However, we saw that in red states demographics mattered more on who would vote for which candidate. We saw in the swing states that there was a specific type of person who would be voting Republican, which was wealthier, white, professional males. Performing different types of classifications on each type of state helped us to compare and learn more about the demographics of people in the states, and what type of people were likely to vote for a Republican candidate or a Democratic candidate.

Conclusion

The 2016 election took a lot of people by surprise as most of the polls and public predicted a Clinton victory. This shows the importance of considering variables beyond what is in the data. It is really difficult to predict how people respond to information in real time as it becomes available. Models and polls give us a prediction of how the effect will be, but it is tough to figure out the significance of the effect. For example, FBI's report that it would be looking into Hillary Clinton's dealings in Benghazi two weeks before the election played a very major role in deciding the election. It is very difficult for polls to accurately decide how this stimulus will sway the vote.

As we also saw in the 2016 election, the presidency is not won by popular vote, but rather the electoral college. And to win the electoral college, performance in swing states must be really strong. A closer look at swing states showed that income, white, and male were major indicators in the election. Trump was able to resonate with these voters because he was able to identify his base and tell them what they wanted to hear. A major knock on Hillary Clinton throughout the election was that she did not focus on the right people in Swing States. A result from this analysis may help the opposing democratic party win the election in 2020 if it able to identify the people who can sway back to them. The biggest use of studying the 2016 election is to see what each side could do to win in 2020. Classification models provide a lot of key indicators that campaigns can use. However, it is very important to realize that these classifications were based on old data, and demographics as well as sentiment are constantly changing.

References

Bialilk, C. and Enten, H. (2016, November 9). The Polls Missed Trump. We Asked Pollsters Why. *FiveThirtyEight.* Retrieved from https://fivethirtyeight.com/features/the-polls-missed-trump-we-asked-pollsters-why/

Enten, H. (2016, October 28). Almost Every Swing State Is A 'Must Win' For Trump Now. *FiveThirtyEight.* Retrieved from https://fivethirtyeight.com/features/almost-every-swing-state-is-a-must-win-for-trump-now/

Silver, N. (2016, November 11). Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else. *FiveThirtyEight.* Retrieved from http://fivethirtyeight.com/features/why-fivethirtyeight-gave-trump-a-better-chance-than-almost-anyone-else/