PSTAT 131 Final Project

Leah Ding (3747821) and Ryan Gan (4227070)

December 12, 2019

Background

1. What makes voter behavior prediction (and thus election forecasting) a hard problem?

Solution: Voter behavior prediction (and thus election forecasting) can be a hard problem because human behavior, in general, is extremely volatile and unpredictable due to the innumerable amount of variables that could affect ones behavior. Oftentimes, these variables are extremely hard to measure, since they rely on the voter's background and mindset. No two voters are exactly the same, and it is extremely difficult to predict all the factors.

2. What was unique to Nate Silver's approach in 2012 that allowed him to achieve good predictions?

Solution: Nate Silver's approach in 2012 was unique because his prediction focused on the decisiveness in the public, something that other people didn't seem to look at. He used Bayes' Theorem and hierarchical modelling to calculate the probability of the support percentage being over 50% for each individual state on each day. He also incorporated the assumption that polling errors are correlated, and polls in other states can miss in the same direction. This approach allowed him to gather thorough data daily to increase the accuracy of his predictions.

3. What went wrong in 2016? What do you think should be done to make future predictions better?

Solution: There is a variety of reasons that the predictions were not accurate in 2016, the biggest being that voter behavior is incredibly unpredictable. In this specific election, media greatly influenced the results of the election.Silver tried to take this into account by labeling them as "shocks". However, it was still hard to cateogorize and quantify these variables. In the future, predictions can be made better by considering more voter behavior features such as voting late, or being undecided. Moreover, systematic polling errors can be evaluated more deeply and fixed, incorrectly collected/inaccurate reports on voter behaviors can be updated, and so forth.

Data

Election data

4. Report the dimension of `election.raw` after removing rows with `fips=2000`. Provide a reason for excluding them.

Solution: The dimension of 'election.raw'after removing rows with 'fips=2000' is 18345 x 5. A reason for excluding these rows is that the counties had NA values, which could negatively affect future calculations.

Census data

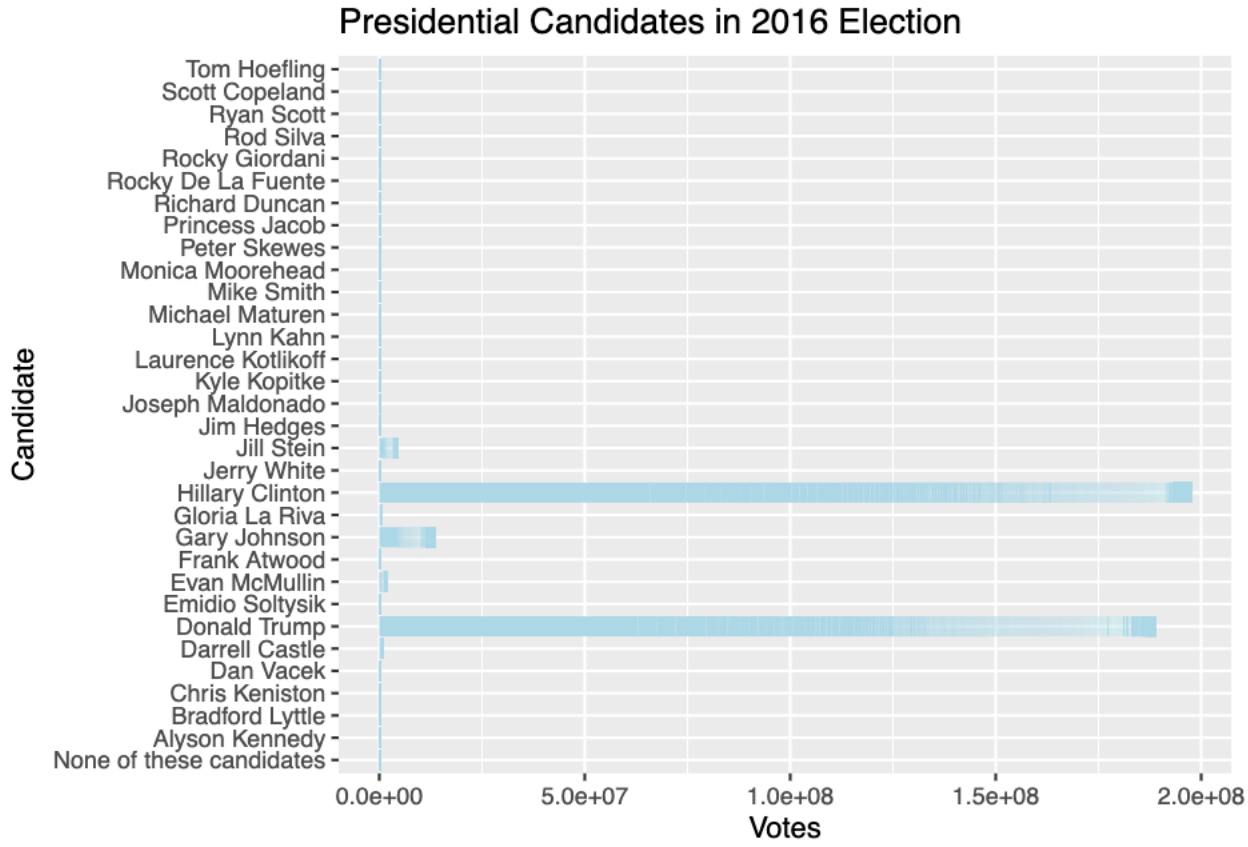
Data wrangling

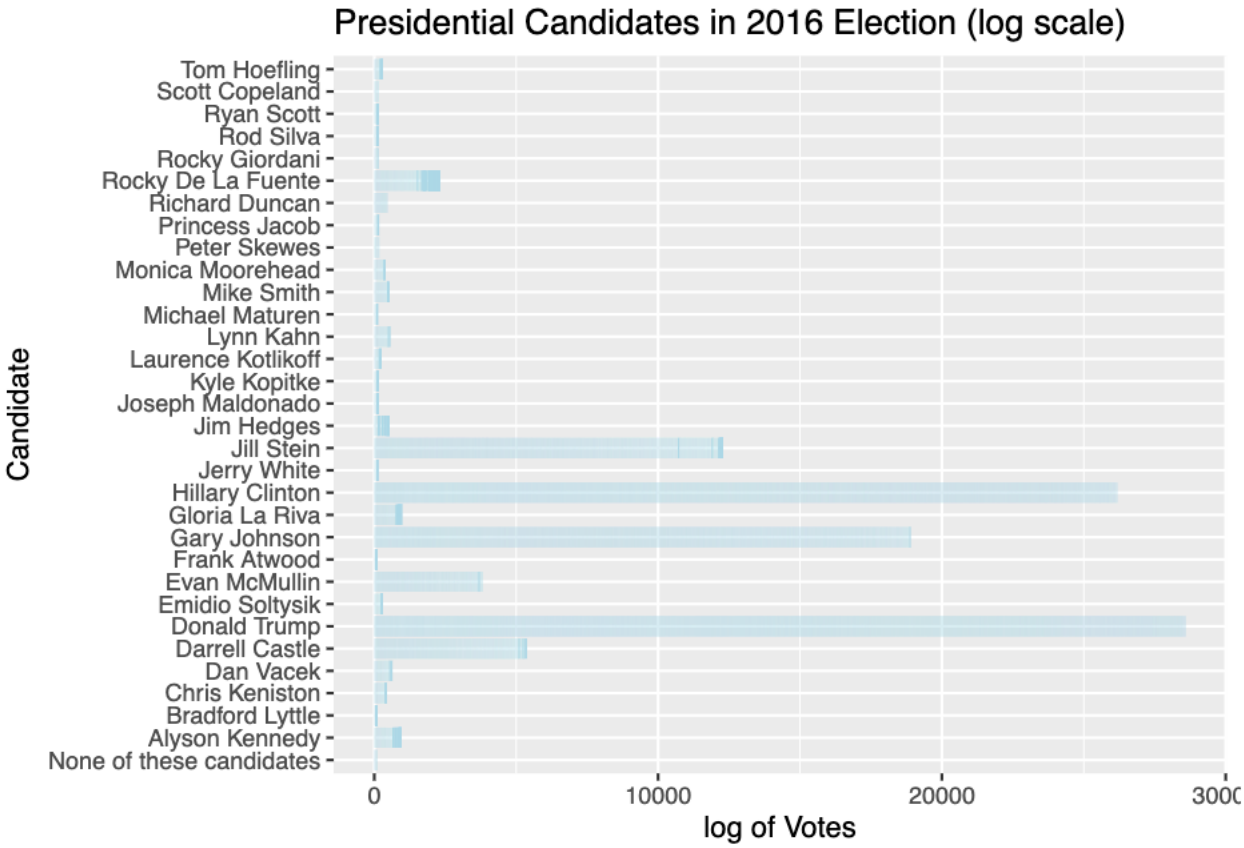
5. Remove summary rows from `election.raw` data: i.e.,

Solution: Election Federal data has 32 observations of 5 variables. Election State data has 298 observations of 5 variables. Election Data has 18,007 observations of 5 variables.

6. How many named presidential candidates were there in the 2016 election? Draw a bar chart of all votes received by each candidate.

Solution: There were 32 named presidential candidates in the 2016 election.

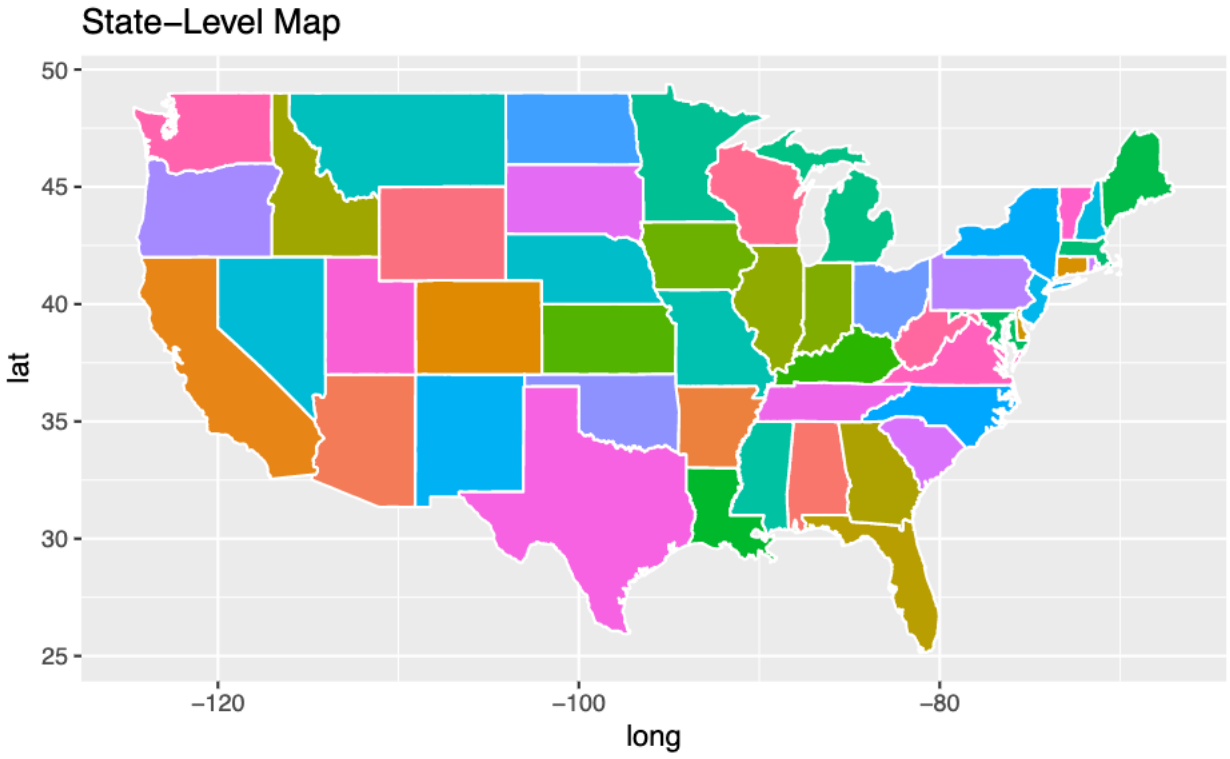




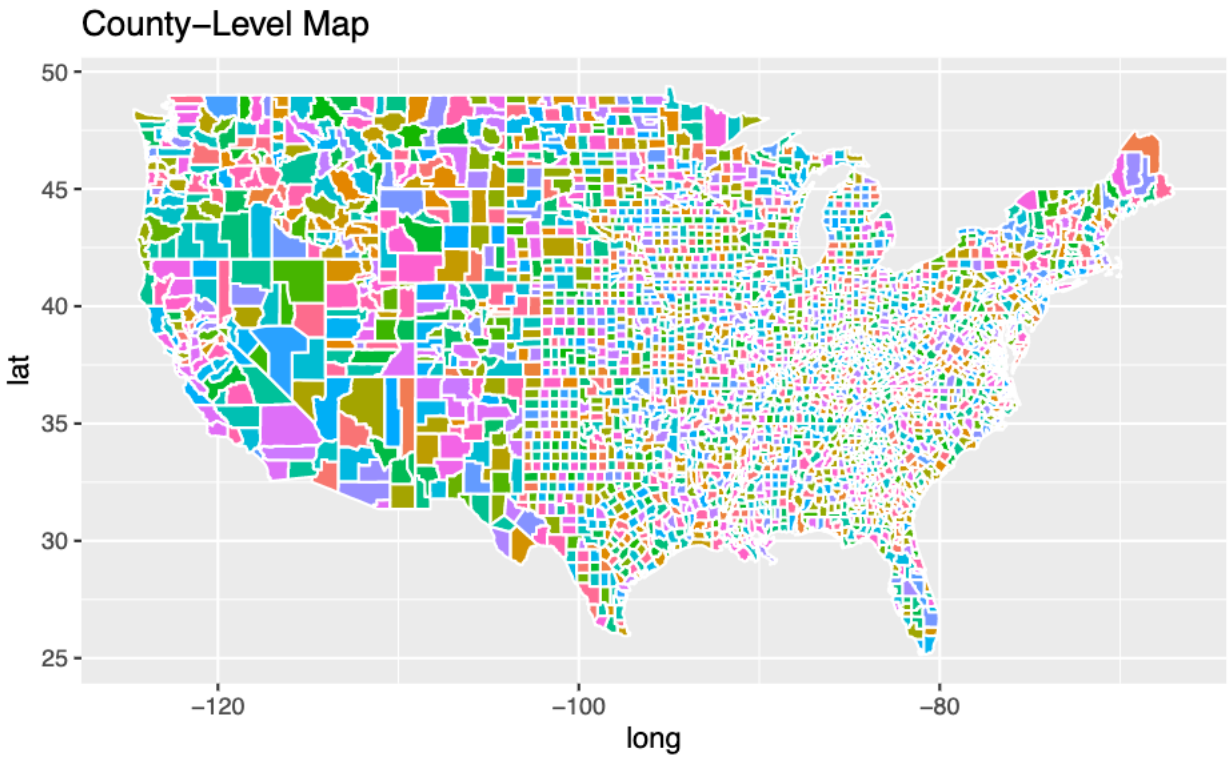
7. Create variables `county_winner` and `state_winner` by taking the candidate with the highest proportion of votes.

Solution: county_winner data has 3,111 observations of 7 variables and state_winner has 50 observations of 7 variables.

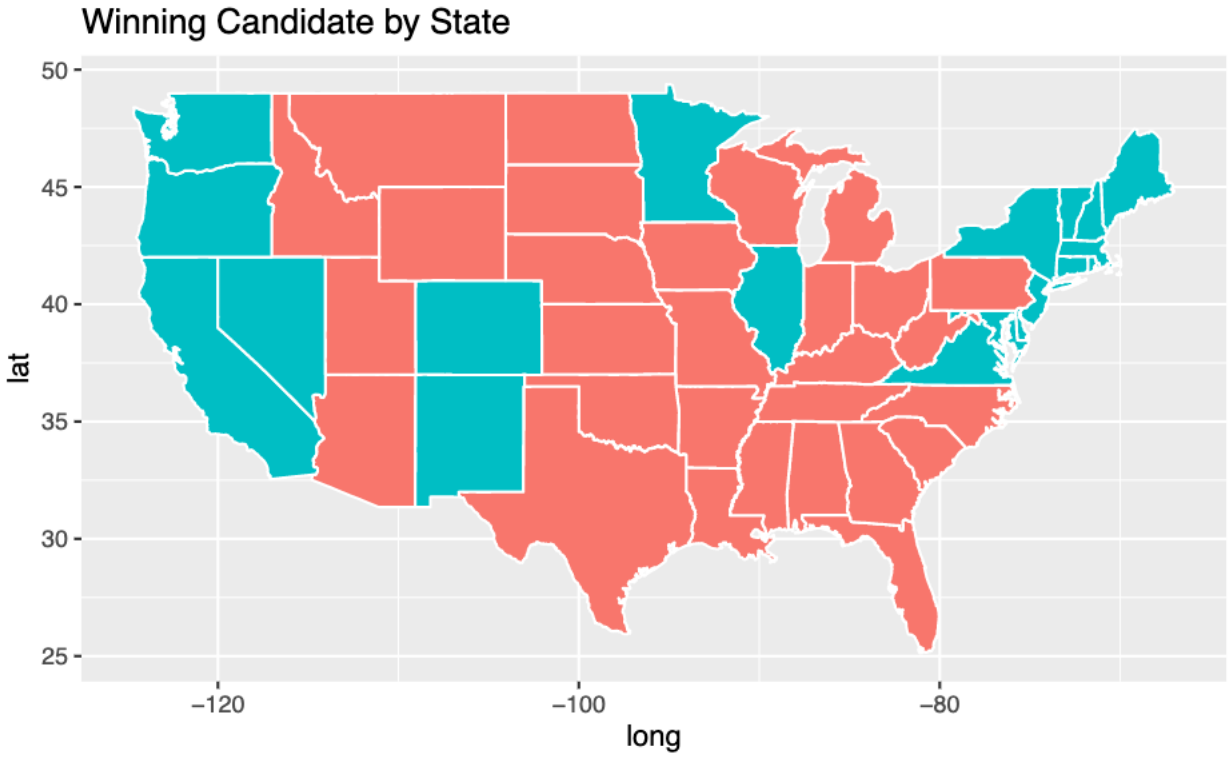
Visualization



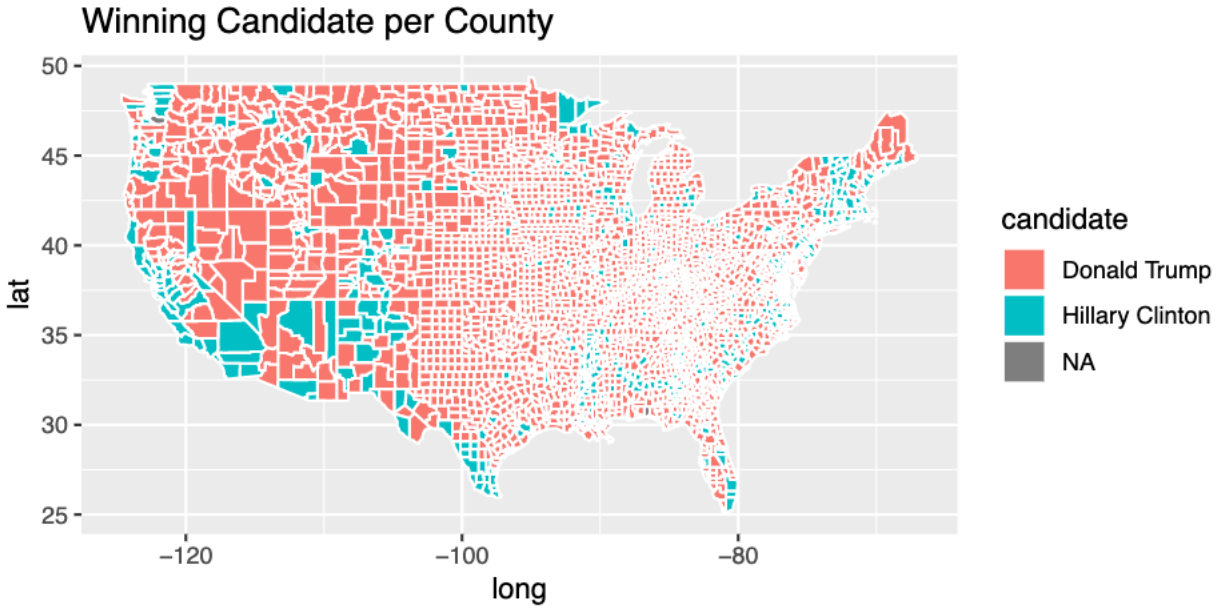
8. Draw county-level map and color by county



9. Now color the map by the winning candidate for each state.

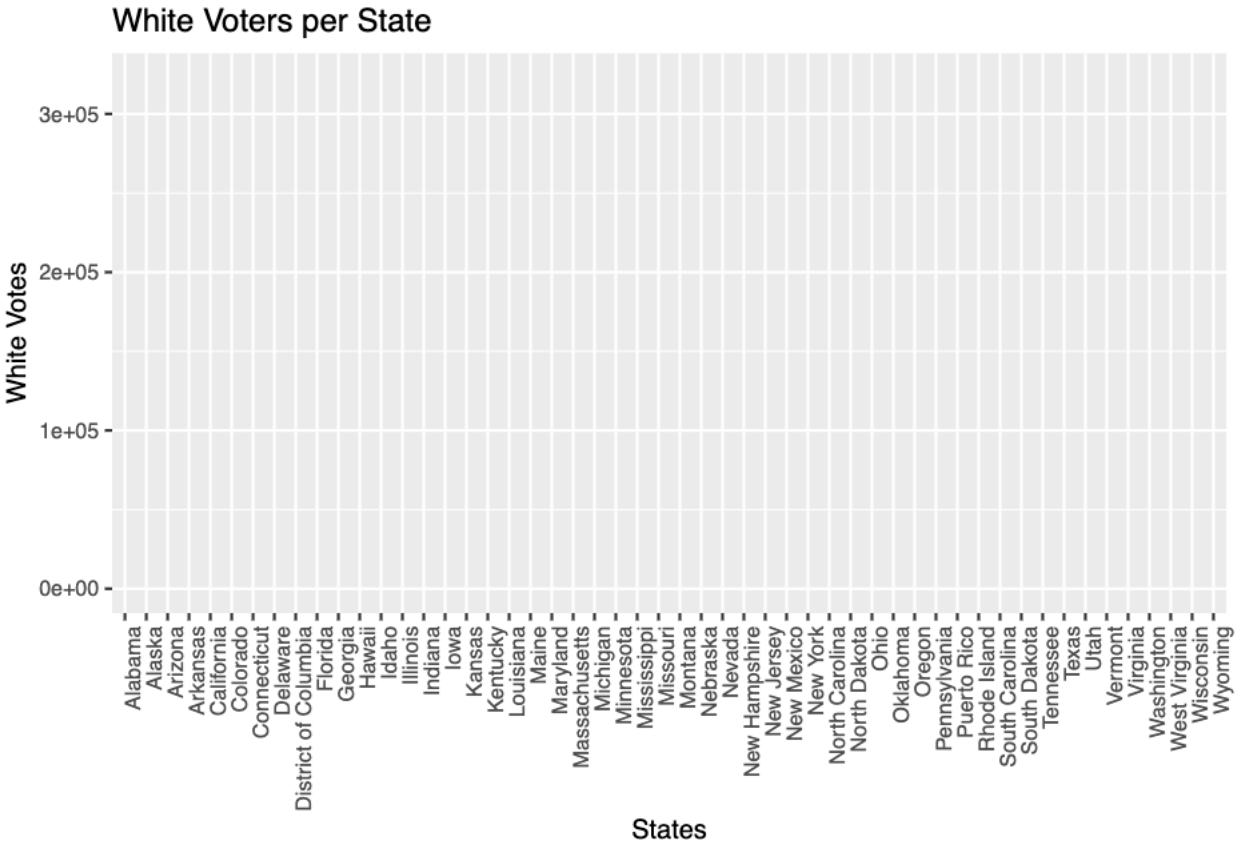


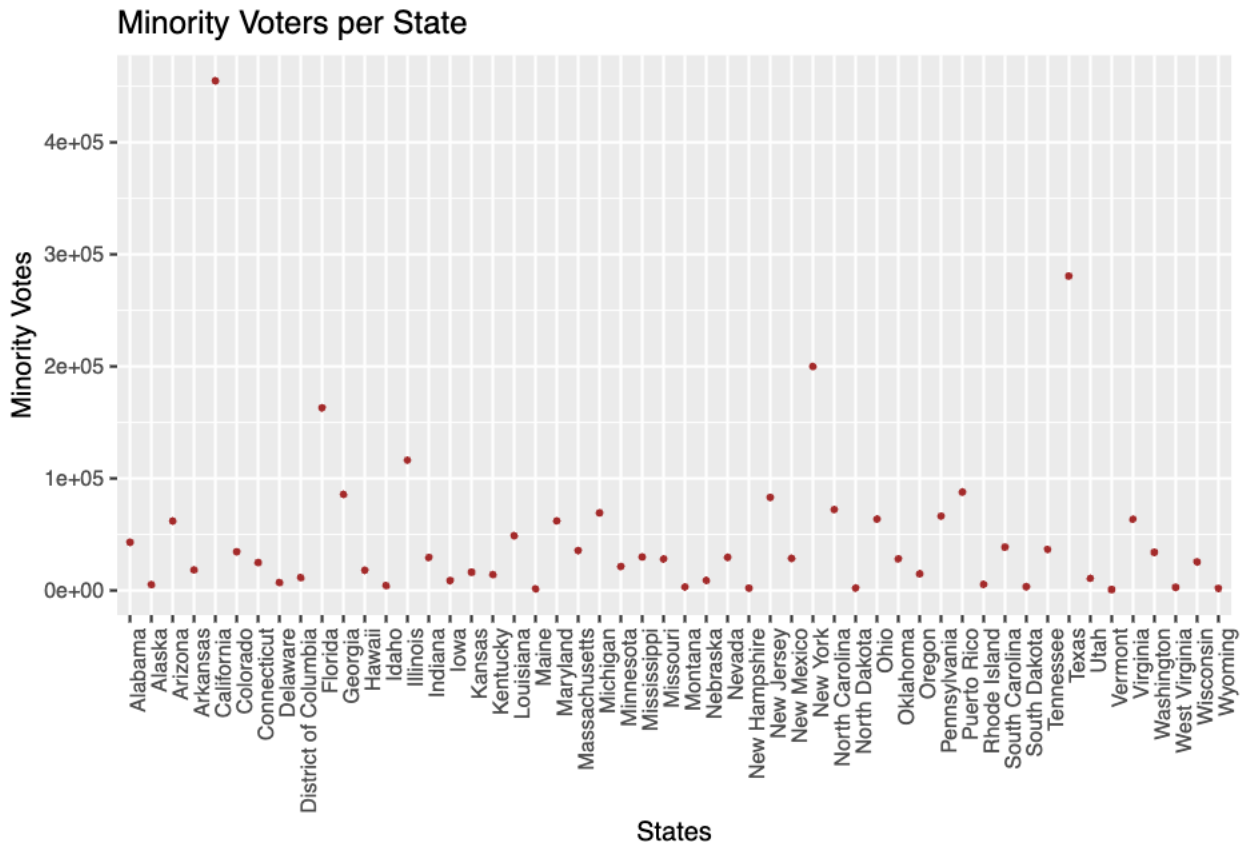
10. The variable `county` does not have `fips` column. So we will create one by pooling information from `maps::county.fips`.



11. Create a visualization of your choice using census data.

Solution: We decided to plot the number of White People who voted in each state and the number of Minorities who voted in each state. It clearly shows that minorities are less likely to vote, which could be due to a variety of reasons.


White Voters per State
White Votes
3e+05
2e+05
1e+05
0e+00
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
States



12. In this problem, we aggregate the information into county-level data by computing `TotalPop`-weighted average of each attributes for each county. Create the following variables: Clean census data, Sub-county census data, and County census data

Solution: Census.del data has 72,720 observations of 28 variables. Census.subct data has 72,720 observations of 30 variables. Census.ct data has 3,218 observations of 28 variables.

#Dimensionality reduction

13. Run PCA for both county & sub-county level data.

Solution: We chose to center and scale the features before running PCA because otherwise, most of the principal components that we observed would be driven by a weighted variable that has the largest mean and variance. Thus, rendering it impossible to scale the other variables evenly.

What are the three features with the largest absolute values of the first principal component?

Solution:

For county level PCA data, the three largest absolute values of the first principal component are IncomePerCap, ChildPoverty, and Poverty.

For sub-county level PCA data, the three largest absolute values of the first principal component are IncomePerCap, Professional, and Poverty.

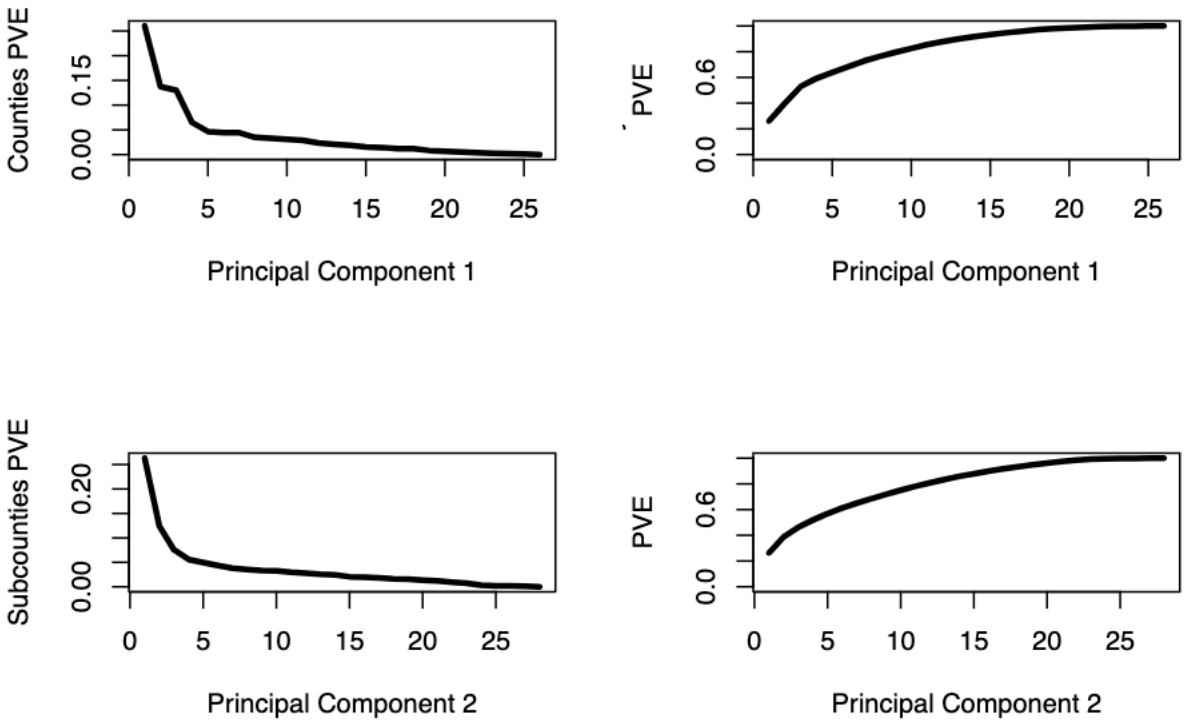
Which features have opposite signs and what does that mean about the correlation between these features?

Solution: For both PCA datas, many features in each one contain a negative sign. For example, in county level PCA data, Poverty and ChildPoverty both are negative values. This means that the correlation of these features are negative with the features that are positive values (ie. Income and Poverty). The same

applies for sub-county level PCA data. If features are the same sign as one another, than they are positively correlated with one another.

14. Determine the number of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses.

Solution: 14 is the minimum number of PCs needed to capture 90% of the variance for the county analysis, and 17 is the minimum number of PCs needed to capture 90% of the variance for the subcounty analysis.




```
## [1] 14
```

```
## [1] 17
```

Clustering

15. With `census.ct`, perform hierarchical clustering with complete linkage.

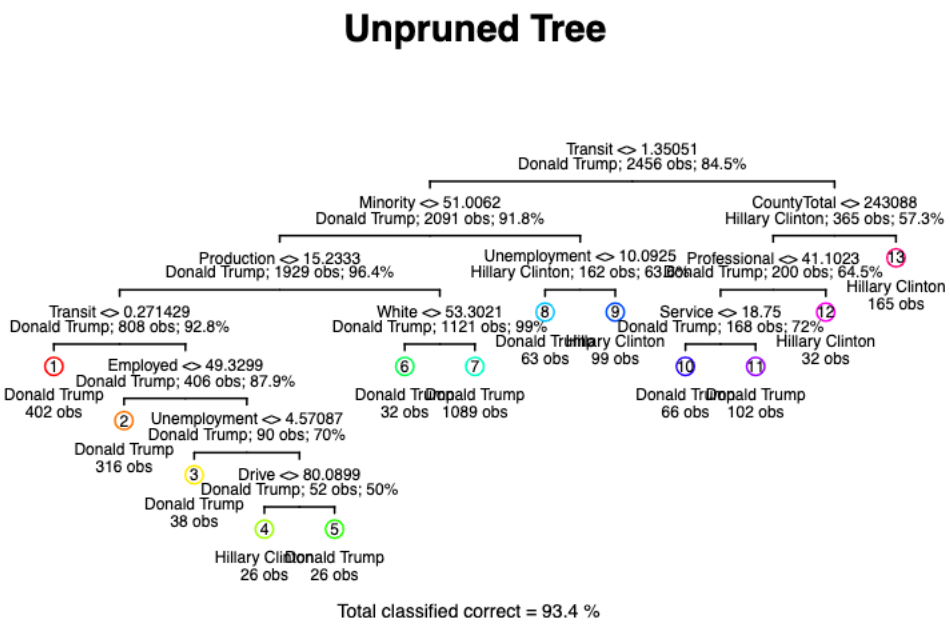
Compare and contrast the results. For both approaches investigate the cluster that contains San Mateo County. Which approach seemed to put San Mateo County in a more appropriate clusters? Comment on what you observe and discuss possible explanations for these observations.

Solution: Before using PCA, clustering decreases form 2584 to 13 in the first 5 clusters and then proceeds to decrease to 1, increase to 14, and decrease to 11. When we recluster with PCA, however, there is a trend of decreasing in the first 3 clusters to increasing from cluster 4 to 5, and then decreasing from cluster 6 to 8, and increasing until the 10th.

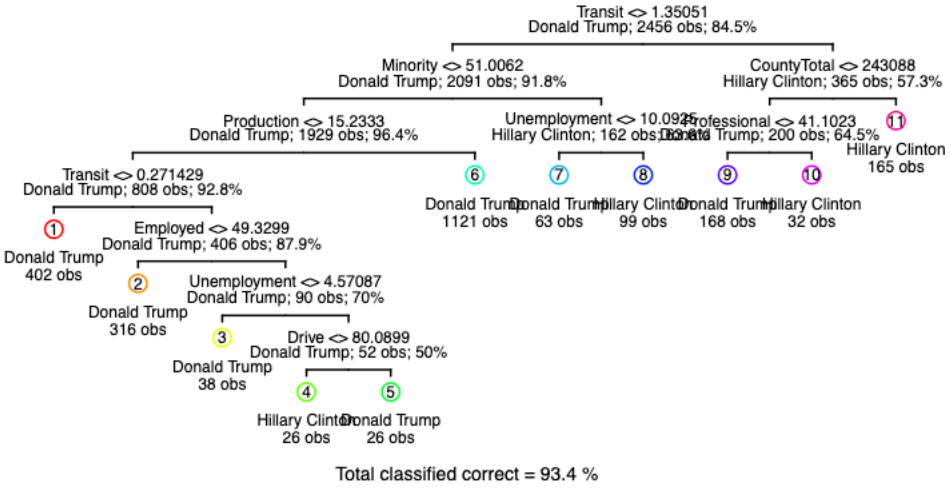
From these trends, we believe that San Mateo is placed into Clust 2, group 9. Clust 2, the complete linkage cluster, is more appropriate since a complete link is less susceptible to noise and outliers in contrast to the other method (PCA), which means San Mateo belongs more in Group 9 than in Group 8. In addition, a smaller distance from the mean displays a more appropriate cluster and contains less variance than variables further away from each other.

Classification

16. Decision tree: train a decision tree by `cv.tree()`.



Pruned Tree



	train.error	test.error
tree	0.0663681	0.0731707
logistic	NA	NA
lasso	NA	NA

Intepret and discuss the results of the decision tree analysis. Use this plot to tell a story about voting behavior in the US.

Solution: From observing, our unpruned decision tree had a 94.1% classification success, which is a good indication for a data set that is as large as ours. Looking closely, we notice that for people who tend to use the transit less, they are more likely to vote for Donald Trump if they have a medium to high income and they're white. This contrasts to Hillary Clinton being popular in counties, minorities, and people who have a low to medium income or are unemployed.

On our pruned decision tree, we notice that it received a total of 93.9% classification success, 0.2% less than our unpruned tree. This could be due to pruning tending to decrease the number of variables the decision tree has, which can affect the accuracy of bigger data sets like ours. But this 0.2% difference is insignificant since it is a very small difference. The pruned decision tree overall provides a more clear visualization since almost all of our previous observations from the unpruned tree still applies. In addition, it also helps us easily observe the different factors that can affect a voter's decision.

17. Run a logistic regression to predict the winning candidate in each county.

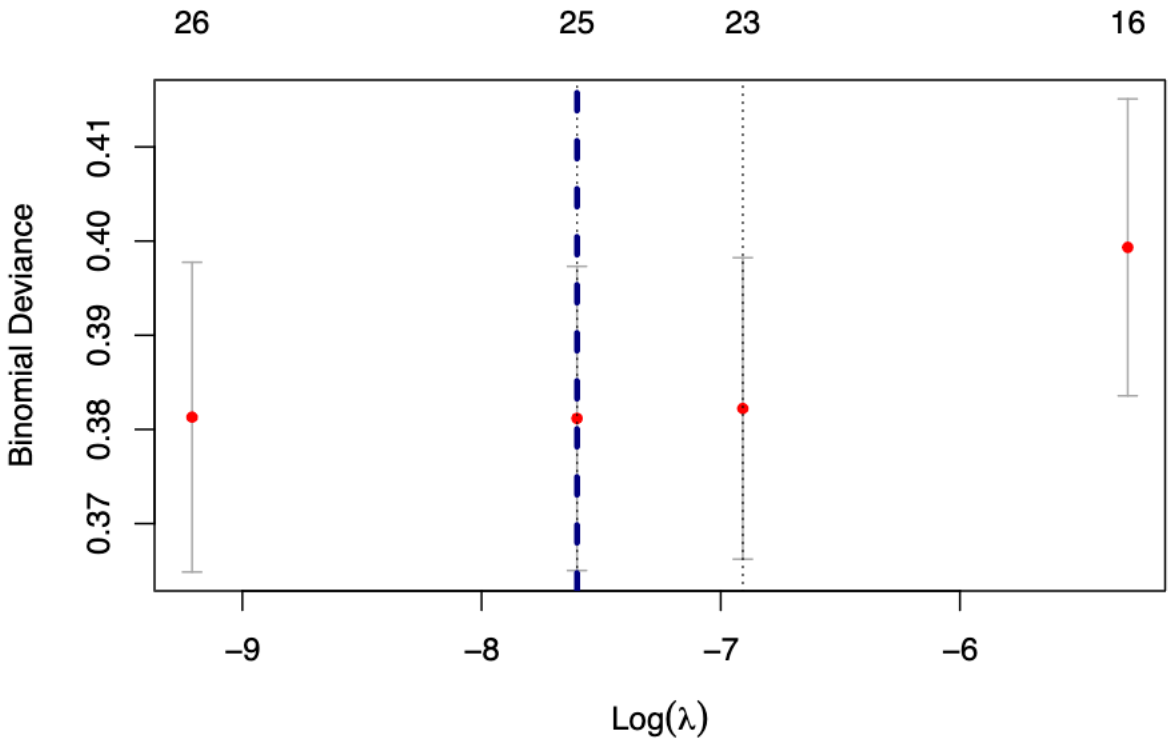
	train.error	test.error
tree	0.0663681	0.0731707
logistic	0.0720684	0.0715447
lasso	NA	NA

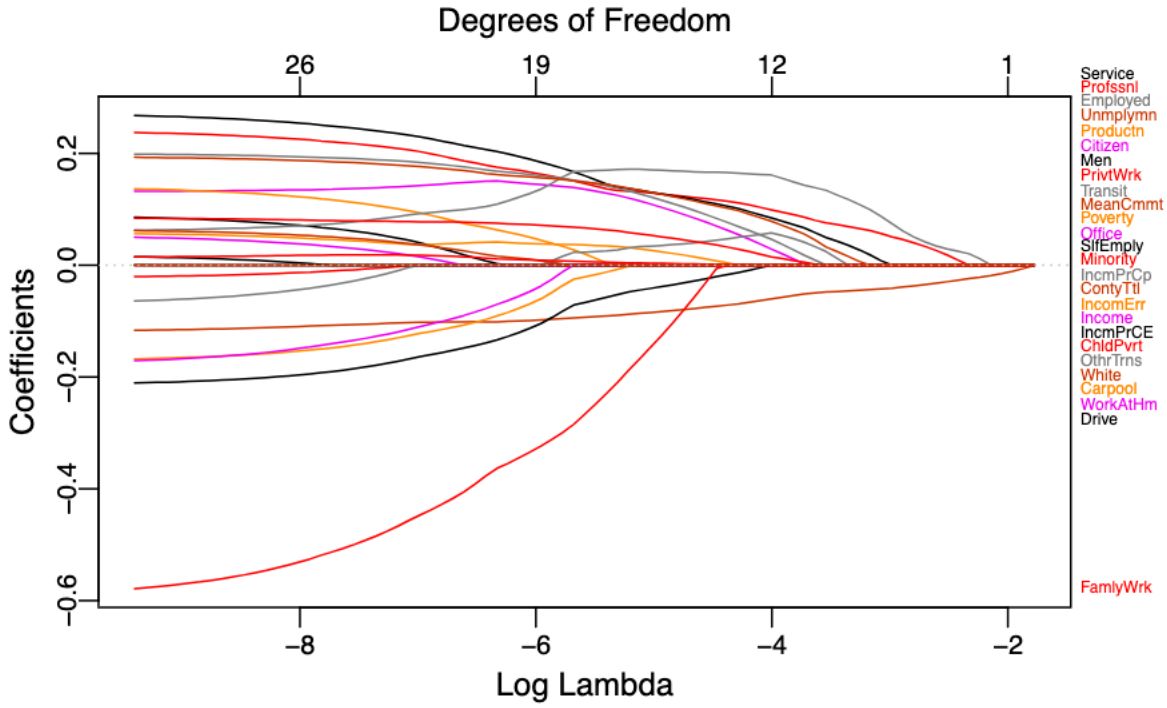
What are the significant variables? Are the consistent with what you saw in decision tree

analysis? Interpret the meaning of a couple of the significant coefficients in terms of a unit change in the variables.

Solution: The significant variables are the following: Citizen, IncomePerCap, Professional, Service, Production, Drive, Carpool, Employed, PrivateWork, Unemployment. These variables are for the most part consistent with what was observed in the decision tree. For example, Employment and Unemployment are significant since they determine what social class the person belongs to and that made a difference in which candidate the voters chose. Voters who Carpooled, like those who used Transit, are probably more liberal and money conscious since they are more environmentally aware and want to save money as well. The Drive category showed a more financially stable group of people who can afford to drive for themselves.

18.Use the cv.glmnet function from the glmnet library to run K-fold cross validation and select the best regularization parameter for the logistic regression with LASSO penalty.






```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

	train.error	test.error
tree	0.0663681	0.0731707
logistic	0.0720684	0.0715447
lasso	0.0757329	0.0699187

What is the optimal value of in cross validation? What are the non-zero coefficients in the LASSO regression for the optimal value of ? How do they compare to the unpenalized logistic regression? Save training and test errors to the records variable.

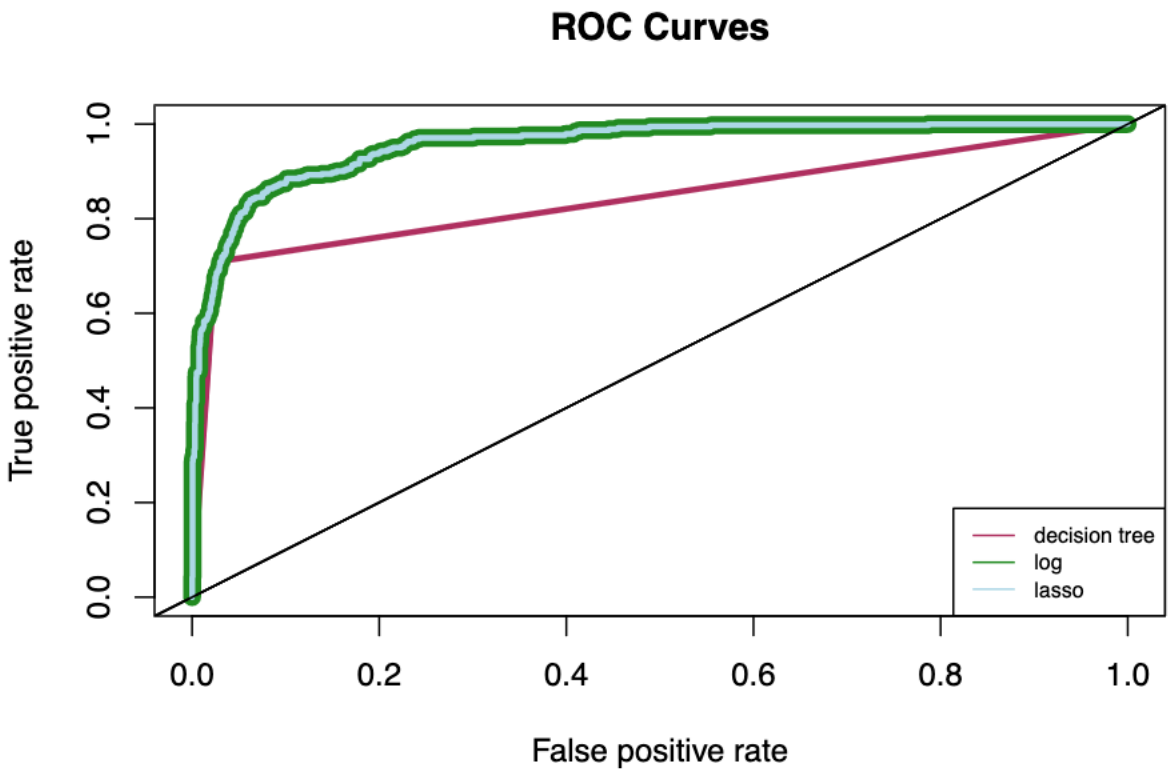
Solution: The optimal value in cross validation is 0.0001. The non-zero coefficients in the LASSO regression for the optimal value of were all of the variables excluding Transit, Self-Employed, and Minority because many of the variables in our dataset affect the outcome. The LASSO regression is used for data sets with not enough data, which has high variance estimates. This is in contrast to logistic regression, which is better for big data. We use the LASSO regression to use the shrinkage method and reduce the variance. However, because our data set is large and many of our variables influence our outcome, they don't have a coefficient of zero. The largest non-zero coefficients that we had were Men, Office, MeanCommute, and PrivateWork, which are the same variables with the highest estimates from the logistic regression.

In contrast to the unpenalized logistic regression, the LASSO regression has less variables to work with because some of the variables equal 0.

In conclusion, the LASSO and logistic regression fits look very similar. The errors are so close to each other since there is already enough data to estimate the coefficients to high accuracy. This means the LASSO regression does not provide any extra insight in contrast to a different scenario with a smaller data set.

19. Compute ROC curves for the decision tree, logistic regression and LASSO logistic regres-

sion using predictions on the test data.



Based on your classification results, discuss the pros and cons of the various methods. Are the different classifiers more appropriate for answering different kinds of questions about the election?

Solution: From the ROC curves, the logistic regression and LASSO methods give the highest true positive rate since both curves hug the upper left corner the most. We also calculated the AUC (Area Under Curve) for each method to accurately determine which method gives the best predictions. The AUC for the decision tree method is the lowest at 0.8297726 while the AUC values for logistic regression and the LASSO method are the highest and the exact same value at 0.9528419. This confirms that the logistic regression and LASSO methods are the best predictive models.

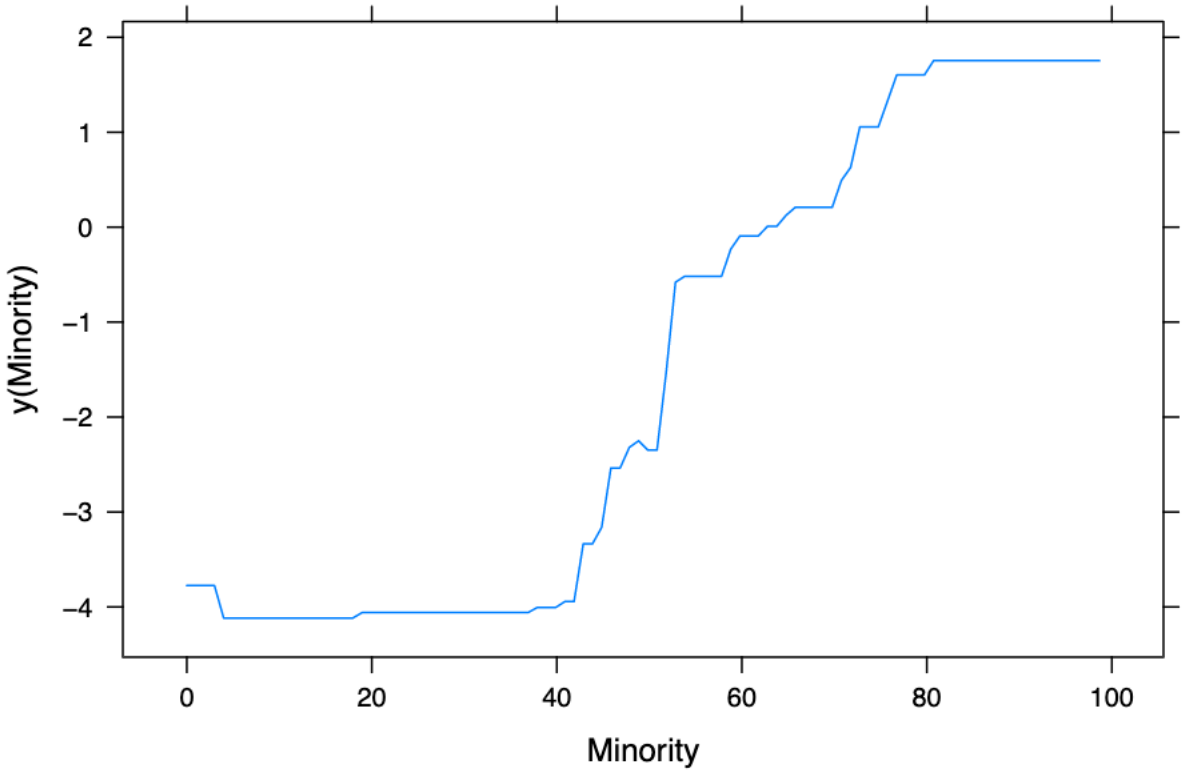
We noticed the decision tree fails to address understanding voter behavior (which entails narrowing down the most variables) as well as the LASSO and logistic regression model, making these two classifiers more appropriate. The LASSO was able to do so with our large data set and the logistic regression model helped us identify the best candidate in the election for each variable group through separation. This gives us more insight on voter behavior.

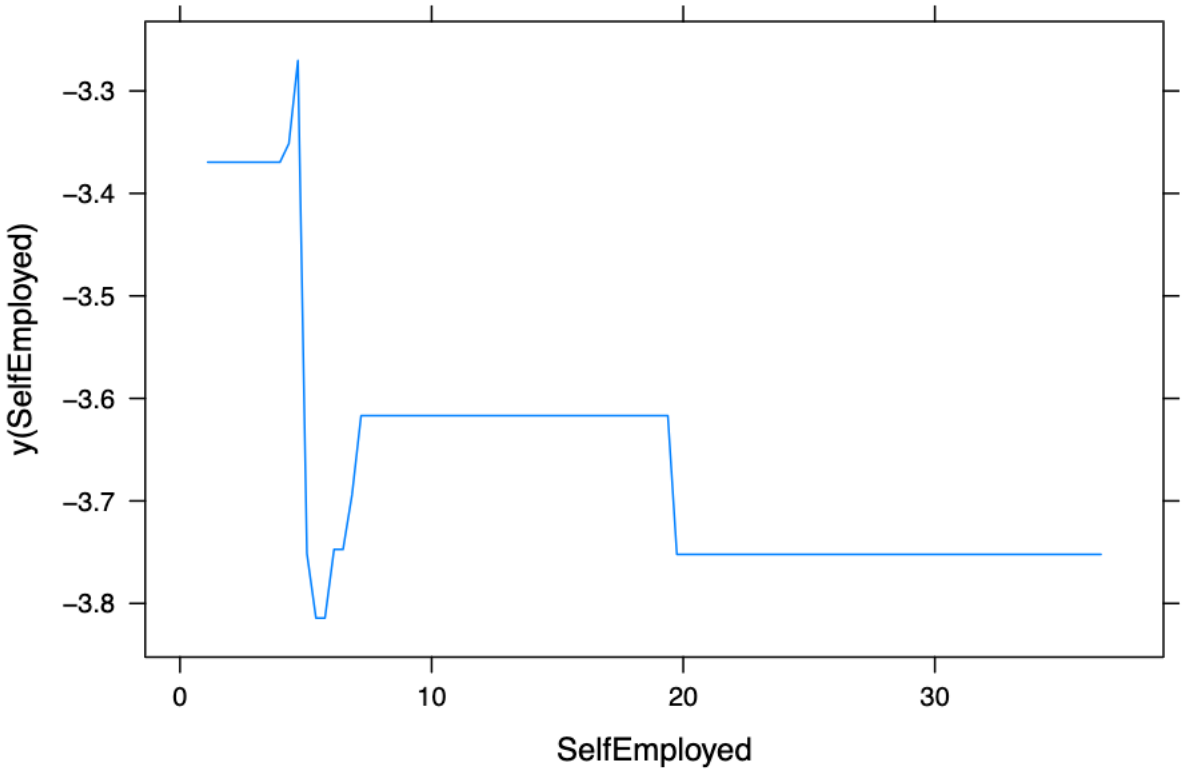
Taking it further

20. Interpret and discuss any overall insights gained in this analysis and possible explanations. Use any tools at your disposal to make your case: visualize errors on the map, discuss what does/doesn't seems reasonable based on your understanding of these methods, propose possible directions (collecting additional data, domain knowledge, etc).

Solution: For this question, we decided to explore the classification methods: boosting, bagging, and random forests. Our goal is to fit these methods to our data and compare their respective final errors. The method with the smallest error will be the best model.

Boosting





	test.error
boosting	0.998374

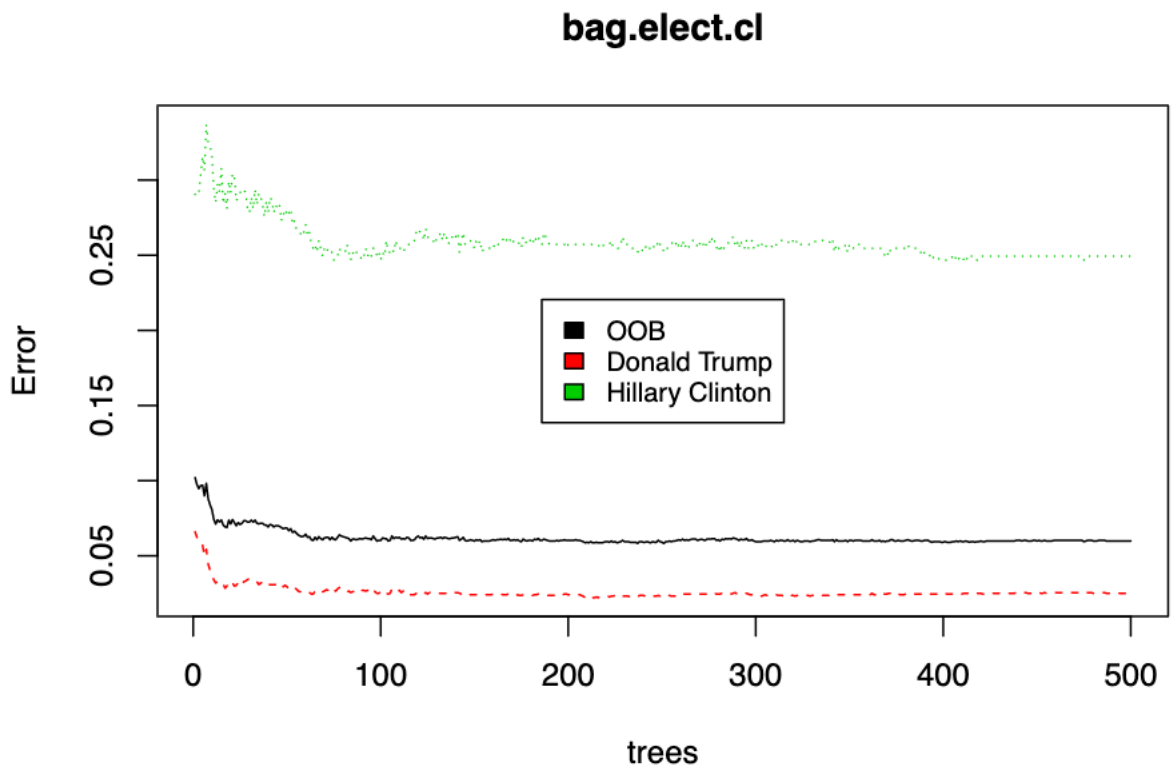
We first used the boosting method. In doing so, we received an error of 0.9983713, a significantly high error. This may be because boosting is more fit for smaller data sets and decision trees. From our relative influence graph, we are told that the variables Minority, Self-Employed, and Child Poverty are the more influential variables. However, this does not display all of our variables and will not provide enough information. It is also inconsistent with our random forest and the Self-Employment graph that we plotted below, in which the Self-Employed actually decreases.

In conclusion, the boosting method is not great at fitting our large data set compared to the logistic regression and other previously used methods. Its plots and graphs don't provide us with enough information on how to interpret the importance of these variables to our data and voter behavior.

Bagging

Next, we tested the bagging method. We observed an error of 0.04885, which is lower than the boosting method and the better method so far. We believe the small error occurred because bagging involves using large unpruned decision trees like our data set. However, this method only uses 2/3 of the total date, which may not provide as much insight on the importance of variance. But it reduces variance and is useful for larger data sets.

In conclusion , the bagging classification method is not as great as the logistic regression and decision tree. But it is better than the boosting method in this scenario.



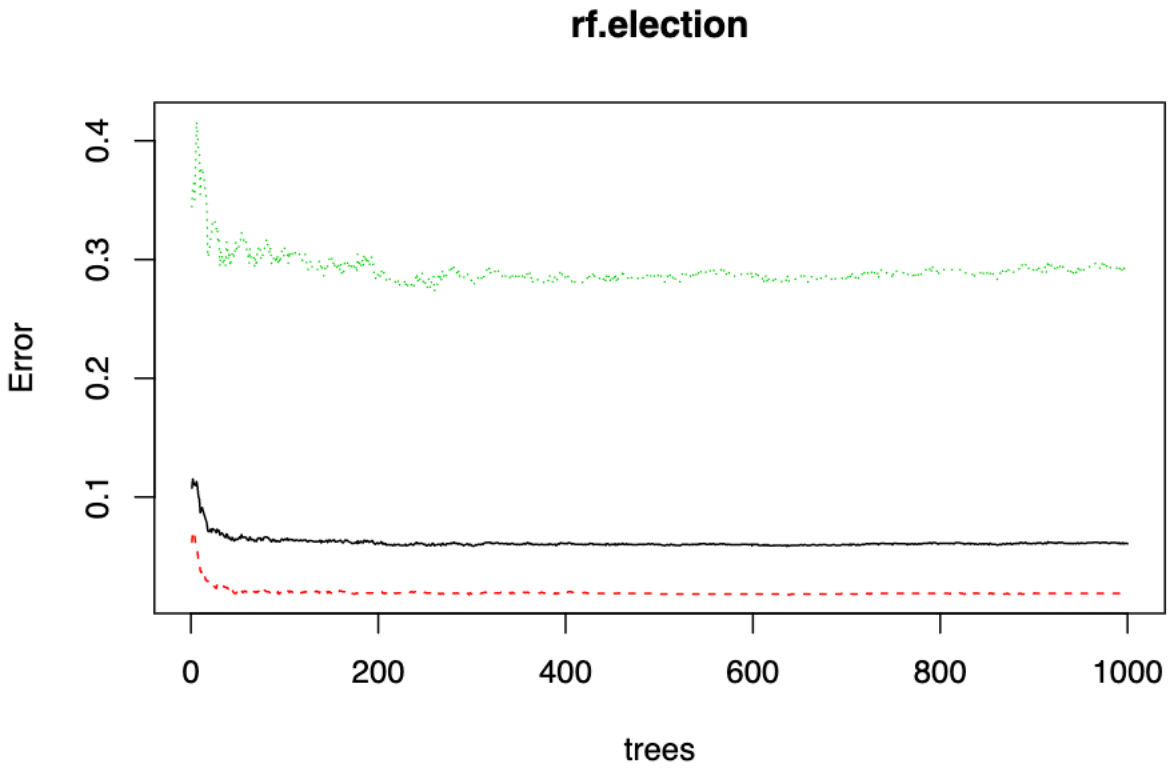
	test.error
boost error	0.9983740
bag error	0.0552846

Random Forest

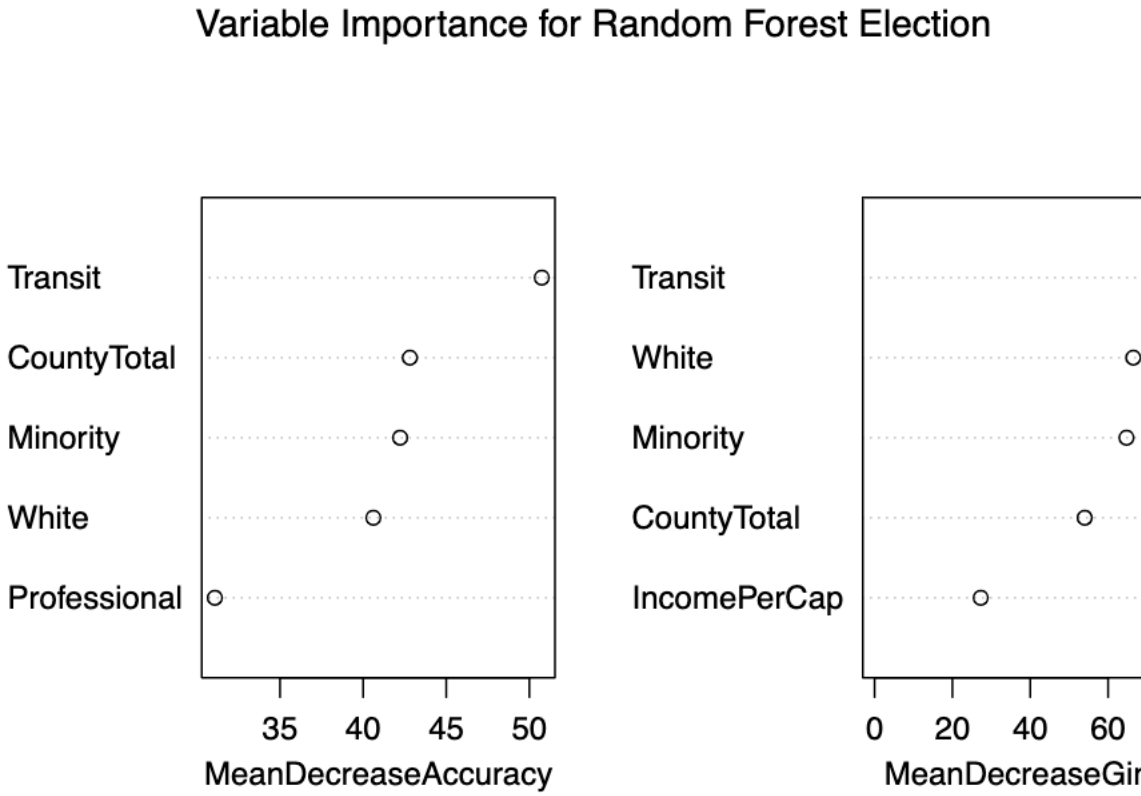
Finally, we computed random forest by creating more trees. We get an error of 0.04885, which is the same as the bagging method. This is a good result for a large data set despite it being small. Diving into the tree, we can see from the Variance Importance charts that the variables Transit, White and Minority play the biggest roles in decreasing the Gini impurity, which is one of the main goals of this method. Following closely are the variables County Total and Professional. This result is held up by previous methods in this project such as the decision trees. This shows the same variables with with the most important towards the top of the trees and the rest branching downwards.

In relation to the election, this makes sense since a voter's demographic as well as social-economic status played a role in their choice of presidential candidate. For example, most white people voted for Donald Trump while most minorities voted for Hillary Clinton.

In conclusion, the random forest tree is an informative method that helps identify strong predictors in the data set despite it being prone to over-fitting similar to the logistical regression model.



forest-1.pdf forest-1.bb



forest-2.pdf forest-2.bb

	train.error	test.error
tree	0.0663681	0.0731707
logistic	0.0720684	0.0715447
lasso	0.0757329	0.0699187

	test.error
boosting	0.9983740
bagging	0.0552846
random forest	0.0536585

From this question, we observed that of the three additional methods that we tested, boosting is the worst with an error close to 1. In contrast, bagging and random forest both have errors that were similar to the test errors of our previously used methods: decision tree, logistic regression, and LASSO.