PSTAT 131 Final Project

Daniel White

12/10/2019

Background

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

Voter behavior prediction can be extremely difficult because there are many factors to take into consideration when dealing with how voters think, making it hard to come up with an accurate and efficient sampling model. Through news outlets, online articles, and other multimedia sources, political information is thrown into the faces of the public. Whether it is an article headline seen on social media taking a jab at a candidate, or a commercial on television taking a controversial stance in order to make a connection with the candidate's target audience, these can have a drastic effect on the opinions of a voter and can ultimately change who they vote for. It can also be difficult obtaining samples that accurately represent voter demographics. Some people may easily just decline to answer certain questions, or feel pressured to give one answer when they actually believe a different answer. These are a few obstacles pollsters are faced with when predicting voter behavior and attempting to correctly forecast elections.

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

For the 2012 presidential election, rather than calculating a probability for an overall percentage, Nate Silver used Bayes' Theorem to examine the full range of probabilities of a candidate obtaining each percentage of the vote. He also used a hierarchical modelling technique and incorporated a time series component to update the model each day leading up to the election.

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

Because every prediction based on polling revolved around probability, there is always some degree of error that will be present and unavoidable. Polling error was believed to be the culprit for the incorrect predictions for the 2016 presidential election. In an article by FiveThirtyEight, they state that polling error can account for 2-3% from statistical noise and nonresponse bias. Although it is possible for state polls and forecasts to "miss" in different directions, thus cancelling each other out, it is not what occured in 2016. It is likely that a large amount of polls and forecasts "missed" in the same direction for the 2016 election, which could be due to overestimated or underestimated demographics numbers for supporting one candidate over the other. It is evident that minimizing polling error would lead to more accurate predictions, and can be done through improvements to the statistical techniques being used.

Election Data

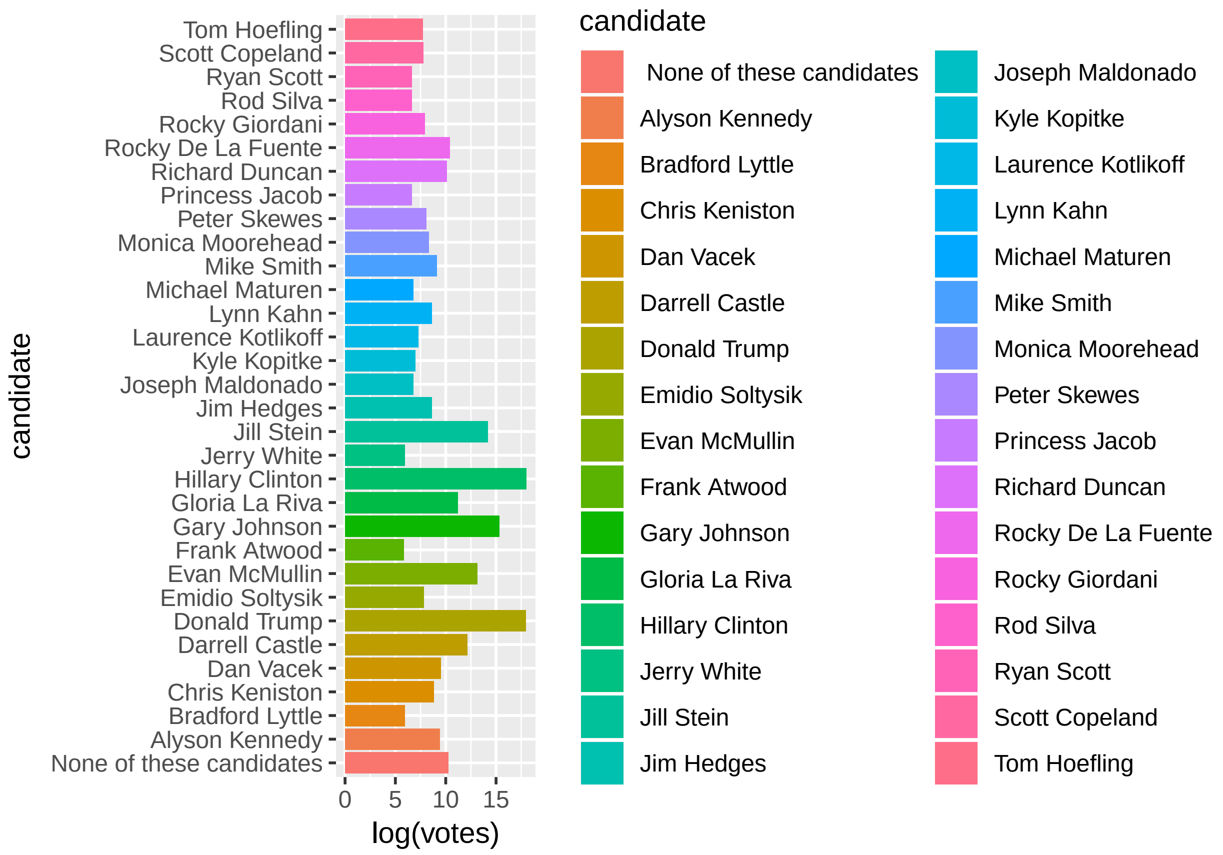
Dimension of election.raw After Removing Rows with 'fips' Equal to 2000.

After removing rows containing fips=2000, there are 18,345 rows and 5 columns in our data set. These rows are excluded because they contain NA values.

Data Wrangling

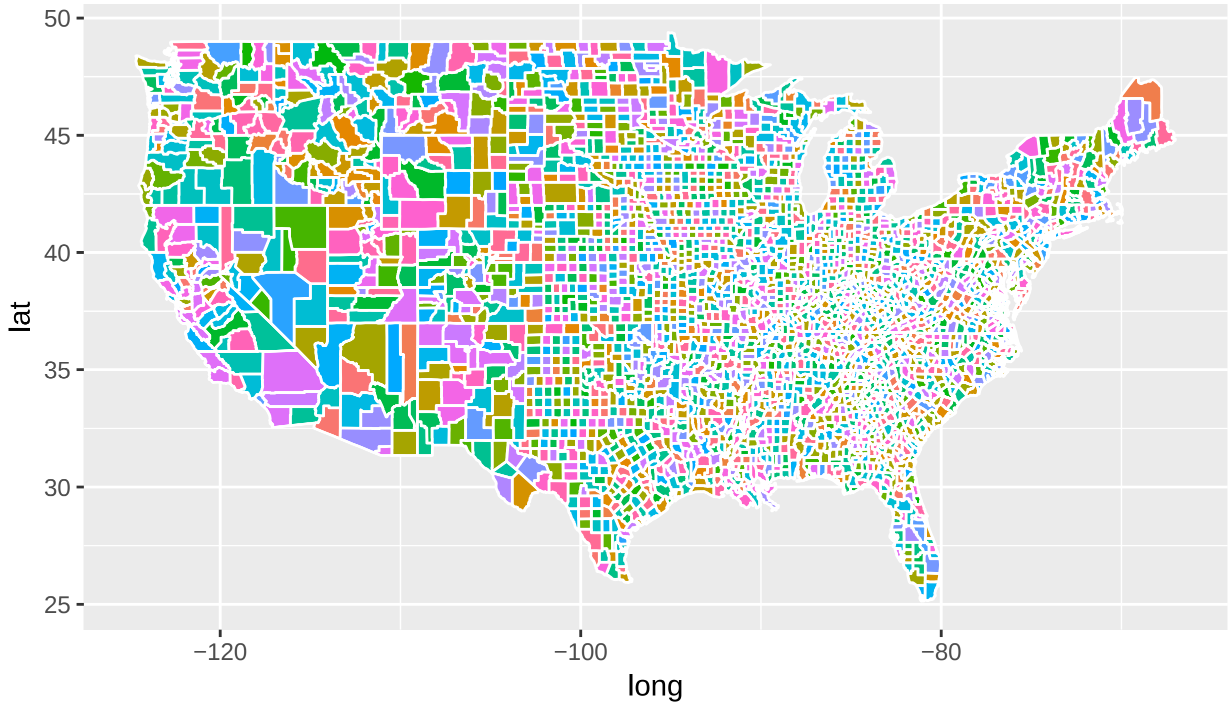
How many named presidential candidates were there in the 2016 election?

There were 31 named presidential candidates in the 2016 presidential election, as shown in the plots below.

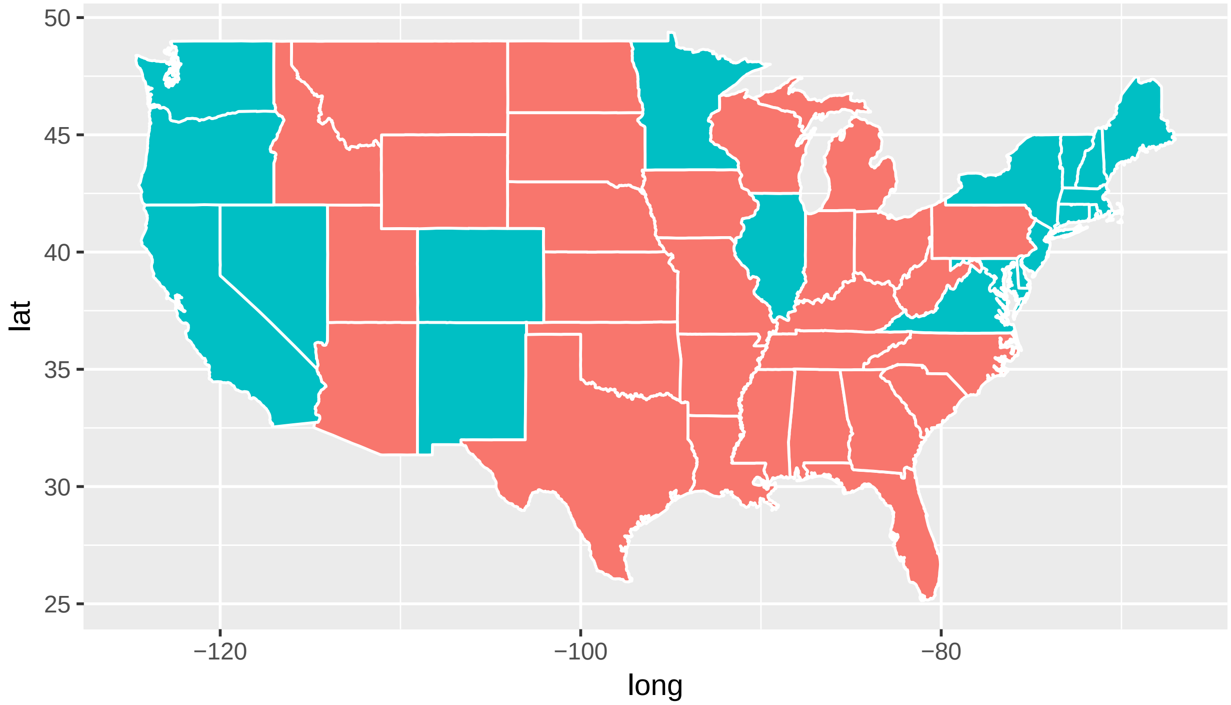

candidate
Tom Hoefling
Scott Copeland
Ryan Scott
Rod Silva
Rocky Giordani
Rocky De La Fuente
Richard Duncan
Princess Jacob
Peter Skewes
Monica Moorehead
Mike Smith
Michael Maturen
Lynn Kahn
Laurence Kotlikoff
Kyle Kopitke
Joseph Maldonado
Jim Hedges
Jill Stein
Jerry White
Hillary Clinton
Gloria La Riva
Gary Johnson
Frank Atwood
Evan McMullin
Emidio Soltysik
Donald Trump
Darrell Castle
Dan Vacek
Chris Keniston
Bradford Lyttle
Alyson Kennedy
None of these candidates
0
5
10
15
log(votes)
candidate
None of these candidates
Alyson Kennedy
Bradford Lyttle
Chris Keniston
Dan Vacek
Darrell Castle
Donald Trump
Emidio Soltysik
Evan McMullin
Frank Atwood
Gary Johnson
Gloria La Riva
Hillary Clinton
Jerry White
Jill Stein
Jim Hedges
Joseph Maldonado
Kyle Kopitke
Laurence Kotlikoff
Lynn Kahn
Michael Maturen
Mike Smith
Monica Moorehead
Peter Skewes
Princess Jacob
Richard Duncan
Rocky De La Fuente
Rocky Giordani
Rod Silva
Ryan Scott
Scott Copeland
Tom Hoefling

Visualization

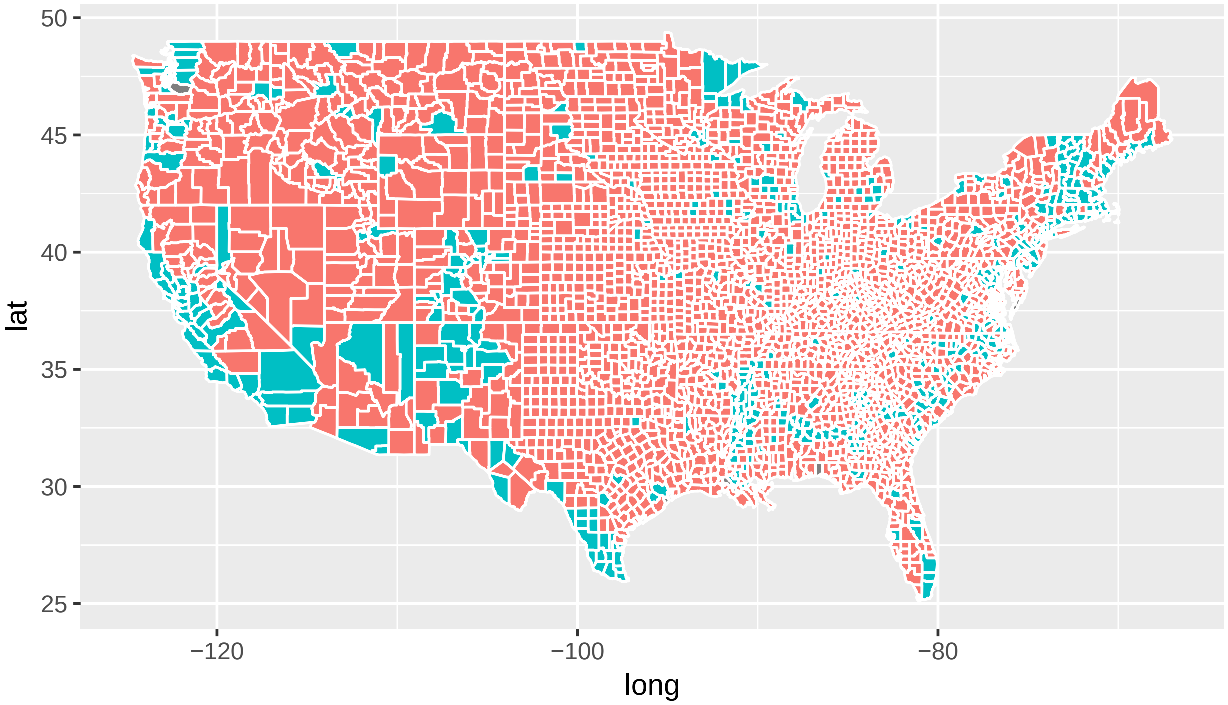
County-level Map



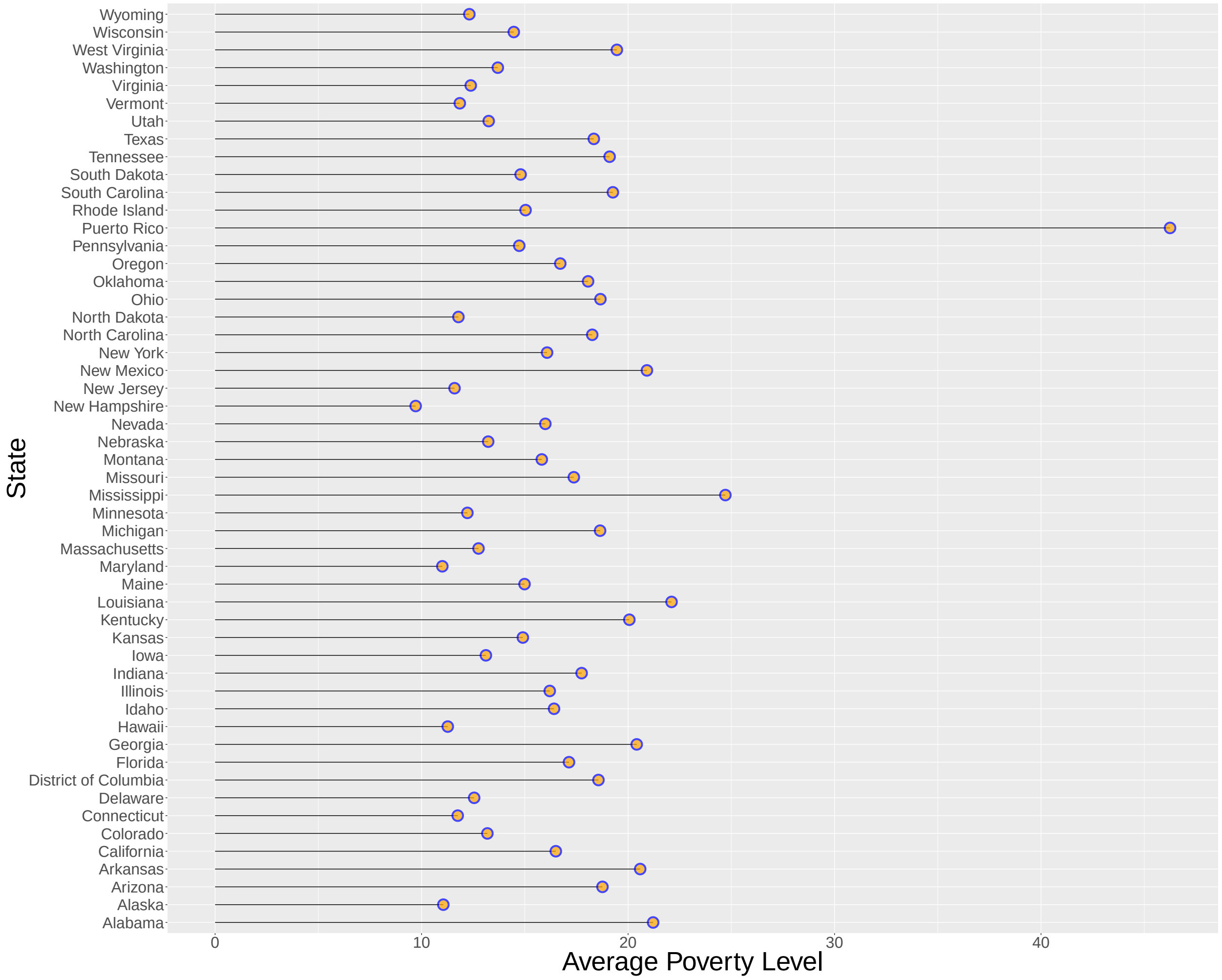
Map of Winning Candidate by State



Map of Winning Candidate by County



Average Poverty Level For Each State (including Puerto Rico)



Aggregate Census Data Information into County-Level Data and Print First Few Rows.

```
## # A tibble: 6 x 28
## # Groups:   State [1]
##   State County TotalPop   Men White Citizen Income IncomeErr IncomePerCap
##   <chr> <chr>     <dbl> <dbl> <dbl>   <dbl>  <dbl>     <dbl>        <dbl>
## 1 Alab~ Autau~    4602.  48.4  73.1    75.0 49985      8036.       24387.
## 2 Alab~ Baldw~    6294.  48.8  83.5    76.4 48673.     9268.       26843.
## 3 Alab~ Barbo~    2992.  52.2  46.6    76.3 32368.     5891.       17105.
## 4 Alab~ Bibb      5651   53.2  77.5    76.8 40212.     6143.       18807
## 5 Alab~ Blount    6412.  49.5  87.8    73.5 45101.     8920.       20171.
## 6 Alab~ Bullo~    3559.  51.8  22      76.0 33445.     8511.       17735
## # ... with 19 more variables: IncomePerCapErr <dbl>, Poverty <dbl>,
## #   ChildPoverty <dbl>, Professional <dbl>, Service <dbl>, Office <dbl>,
## #   Production <dbl>, Drive <dbl>, Carpool <dbl>, Transit <dbl>,
## #   OtherTransp <dbl>, WorkAtHome <dbl>, MeanCommute <dbl>,
## #   Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Unemployment <dbl>, Minority <dbl>
```

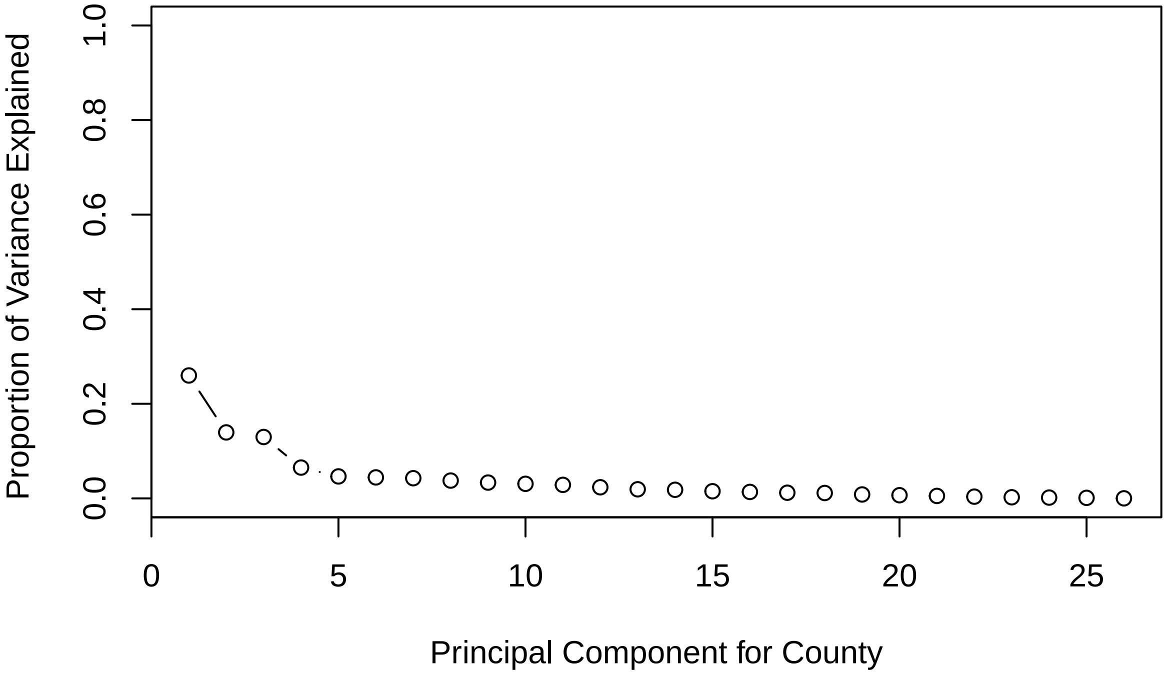
Dimensionality Reduction

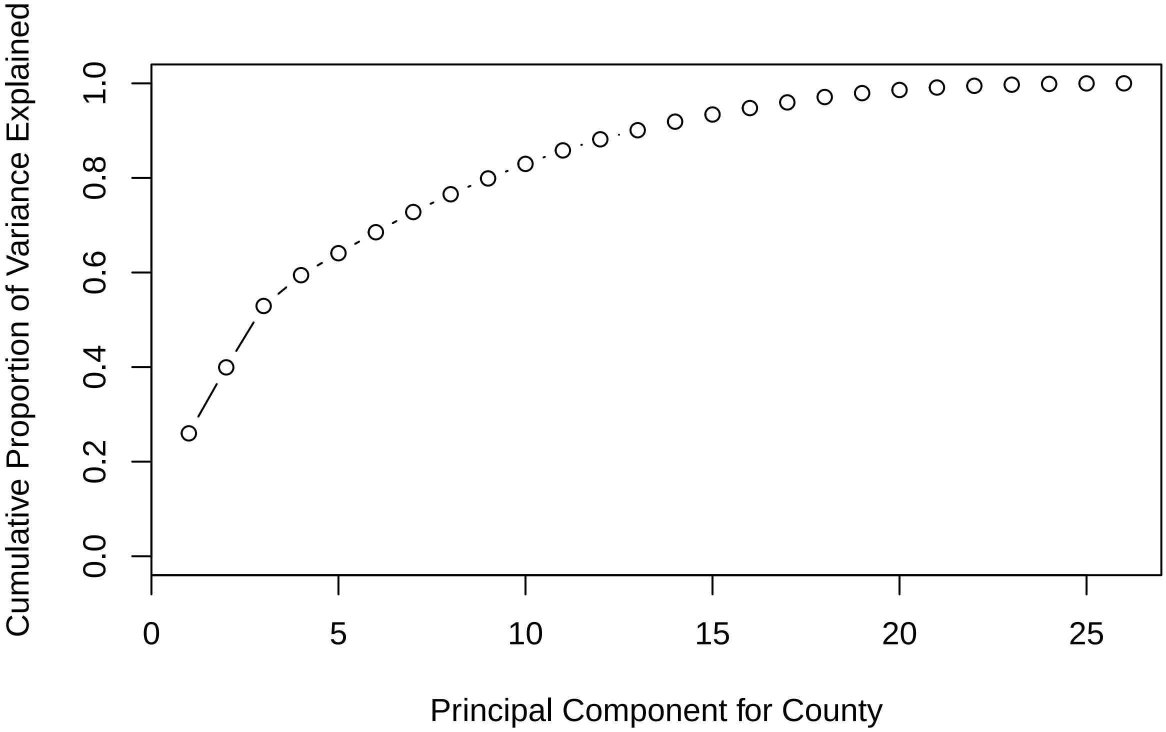
PCA for County and Sub-County Level Data

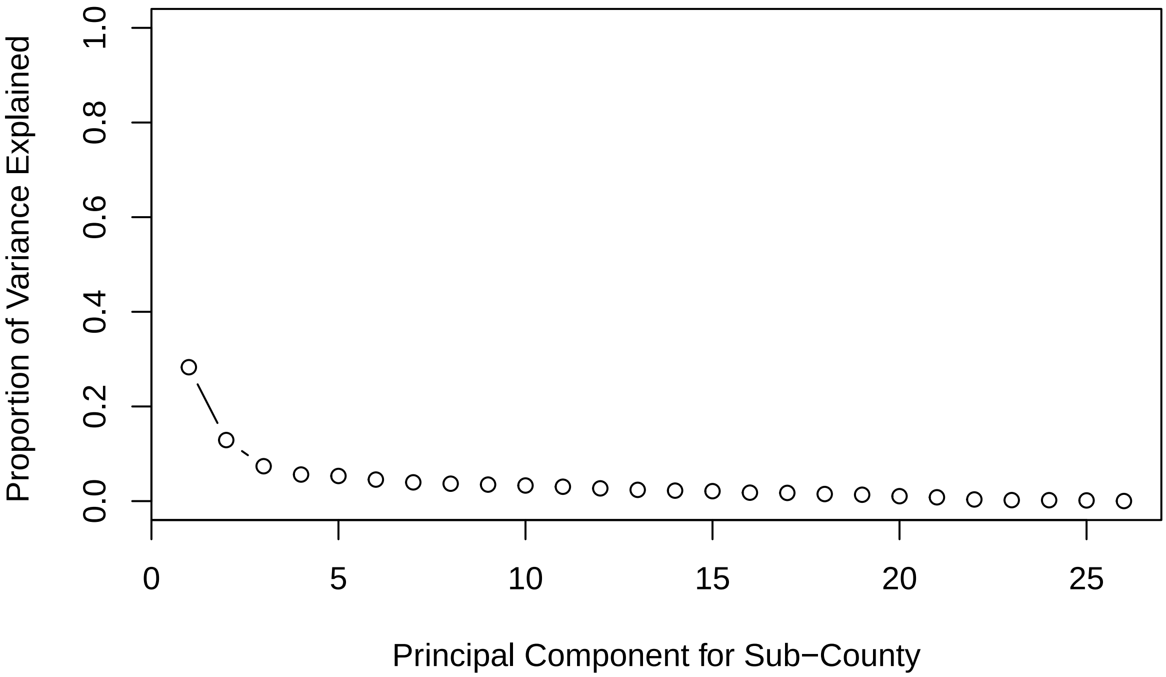
I chose to center and scale the features before running PCA because the data varies in size for some variables, and may have a different scale such as percentages, etc. The three features with the largest absolute values of the first principal component for county are IncomePerCap, ChildPoverty, and Poverty. The three features with the largest absolute values of the first principal component for sub county are IncomePerCap, Professional, and Poverty. The features with opposite signs when comparing county and sub county are TotalPop, Drive, Transit, MeanCommute and PrivateWork. The features with opposite signs have a negative correlation between the first two principal components, and vice versa.

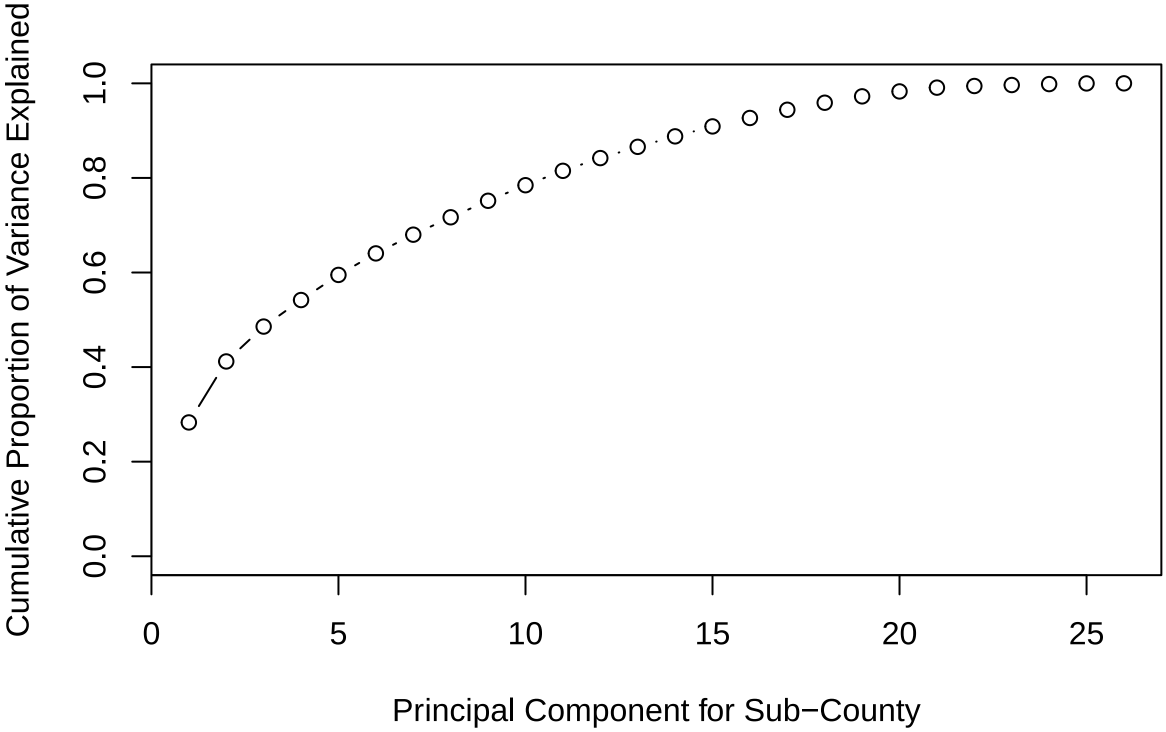
Minimum Number of PCs Needed to Capture 90% of Variance for County and Sub-County

The minimum number of principal components needed to capture 90% of the variance for the county and sub-county analyses are 13 and 15, respectively.




Cumulative Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Principal Component for County


Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Principal Component for Sub−County



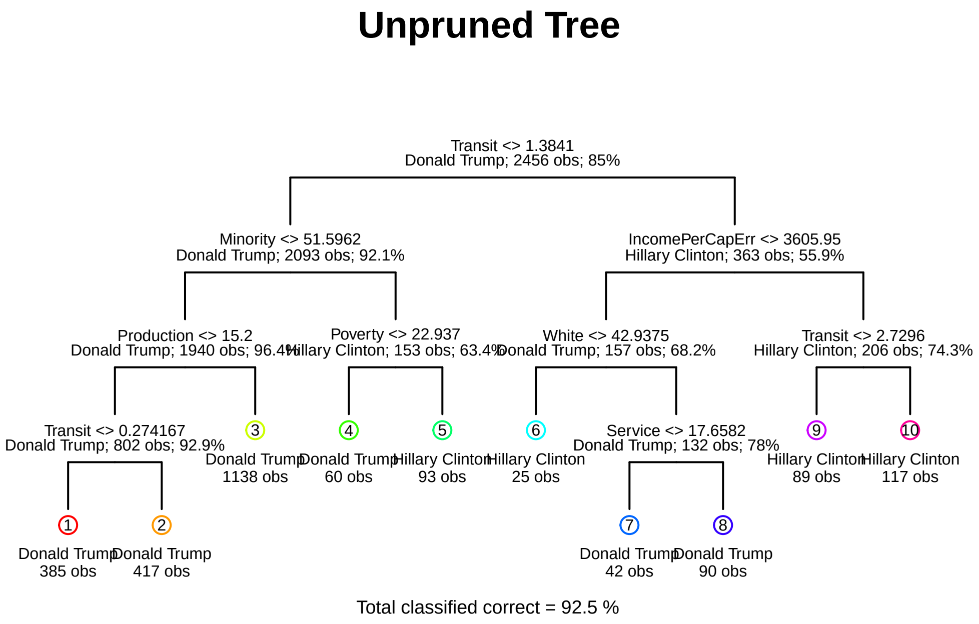
Clustering

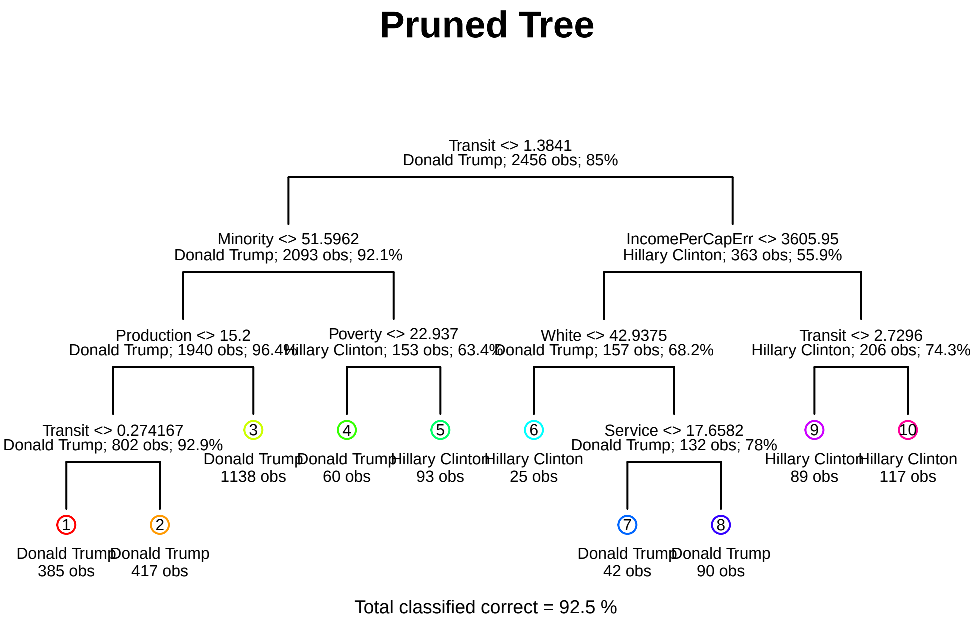
Hierarchical Clustering with Complete Linkage

Classification

Decision Tree

The pruned tree below shows the first split occurs at “Transit”. This could be because more urban, densely populated areas tend to lean more towards the Democratic party, in this case Hillary Clinton, while more rural areas tend to vote more towards the Republican party, in this case Donald Trump. Urban areas are more likely to contain more options for public transit, which seems to be a deciding factor for many people and proves to be a highly influential factor, as it shows up multiple times in the tree. White and Income also show great significance as well.


Unpruned Tree
Transit <> 1.3841
Donald Trump; 2456 obs; 85%
Minority <> 51.5962
Donald Trump; 2093 obs; 92.1%
IncomePerCapErr <> 3605.95
Hillary Clinton; 363 obs; 55.9%
Production <> 15.2
Donald Trump; 1940 obs; 96.4%
Poverty <> 22.937
Hillary Clinton; 153 obs; 63.4%
White <> 42.9375
Donald Trump; 157 obs; 68.2%
Transit <> 2.7296
Hillary Clinton; 206 obs; 74.3%
Transit <> 0.274167
Donald Trump; 802 obs; 92.9%
3
Donald Trump
1138 obs
4
Donald Trump
60 obs
5
Hillary Clinton
93 obs
6
Hillary Clinton
25 obs
Service <> 17.6582
Donald Trump; 132 obs; 78%
9
Hillary Clinton
89 obs
10
Hillary Clinton
117 obs
1
Donald Trump
385 obs
2
Donald Trump
417 obs
7
Donald Trump
42 obs
8
Donald Trump
90 obs
Total classified correct = 92.5 %



```
##           train.error test.error
## tree       0.07532573 0.08617886
## logistic           NA         NA
## lasso              NA         NA
```

Logistic Regression to Predict Winning Candidate in Each County

In the logistic regression model, the very significant variables are Citizen, Professional, Service, Production, Drive, Carpool, Employed, PrivateWork, and Unemployment. Other significant variables are White and WorkAtHome while less slightly significant variables include Income, IncomPerCapErr, MeanCommute, and FamilyWork. This is not consistent with what we saw in our decision tree analysis as the most significant variables there were Transit, White, and Income. Looking at some of the more significant variables and their coefficients, we see that a one unit increase in the Citizen variable (being a US citizen) increases the chances of voting for Donald Trump by about 11% while a one unit increase in the Carpool variable decrease the chance of voting for Donald Trump by about 26%.

```
##           train.error test.error
## tree       0.07532573 0.08617886
## logistic   0.07125407 0.07642276
## lasso              NA         NA
```

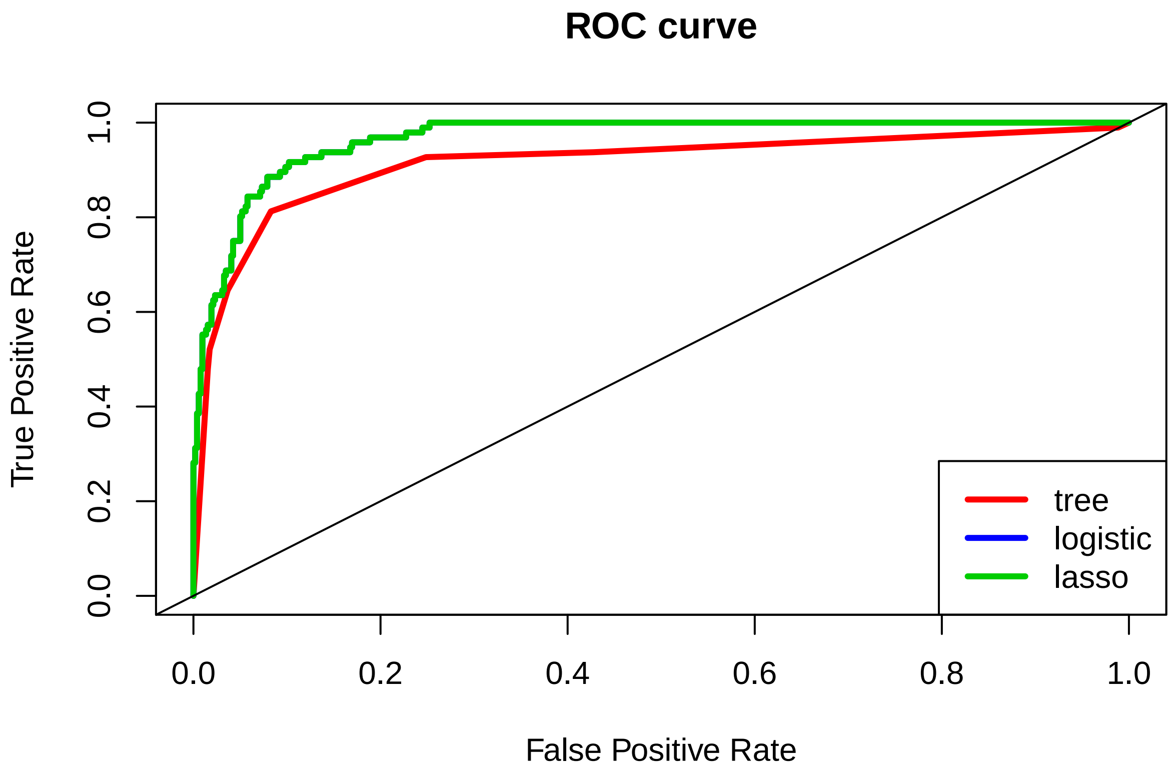
Lasso Penalty

The optimal value of lambda in cross validation is $\lambda = 0.001228762$. The non-zero coefficients in the LASSO regression for the optimal value of λ include all except Income, ChildPoverty, SelfEmployed and Minority. In the unpenalized logistic regression model, Income has a coefficient of -3.585e-05, ChildPoverty has a coefficient of -2.652e-02, SelfEmployed has a coefficient of 3.477e-02, and Minority has a coefficient of -5.859e-02.

```
##          train.error test.error
## tree      0.07532573 0.08617886
## logistic  0.07125407 0.07642276
## lasso     0.07043974 0.07967480
```

ROC Curves for Decision Tree, Logistic Regression and LASSO Logistic Regression

The model with the highest AUC value is the LASSO penalized logistic model. We would conclude that this model has the best predictive power. The AUC value for the decision tree is very close to the LASSO, and let us visualize how decisions are made within each model. The decision tree proves to be better for answering questions regarding how different aspects of a person's life can affect their vote in the election.



Taking it Further

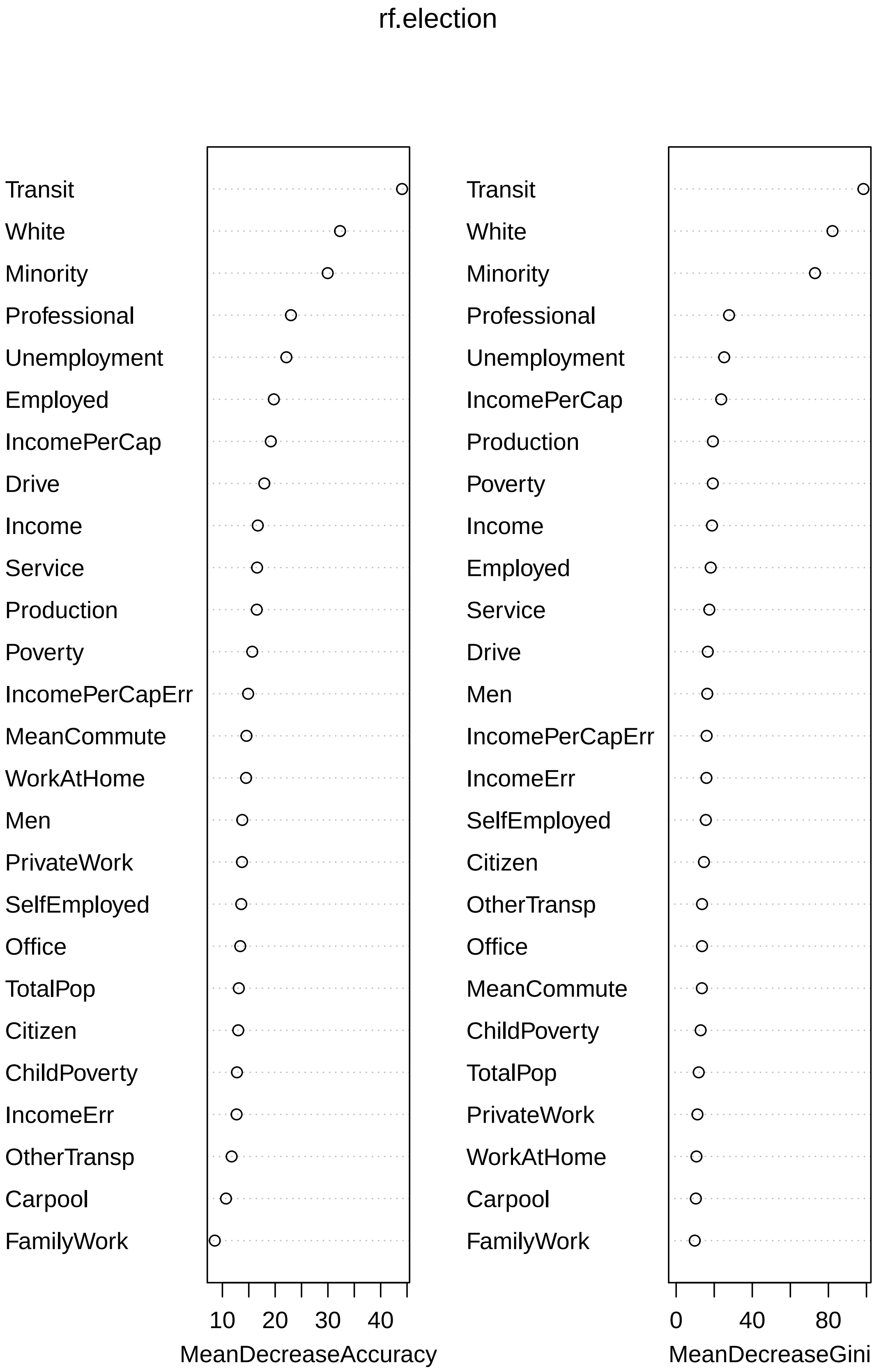
I found it very interesting how different models can have large differences in each variable, as shown in above. With that said, there are several other models that could be considered in this analysis, including SVM, Random Forest, and Boosting to names a few. One question that comes to mind is if any of these other models will outperform the aforementioned models.

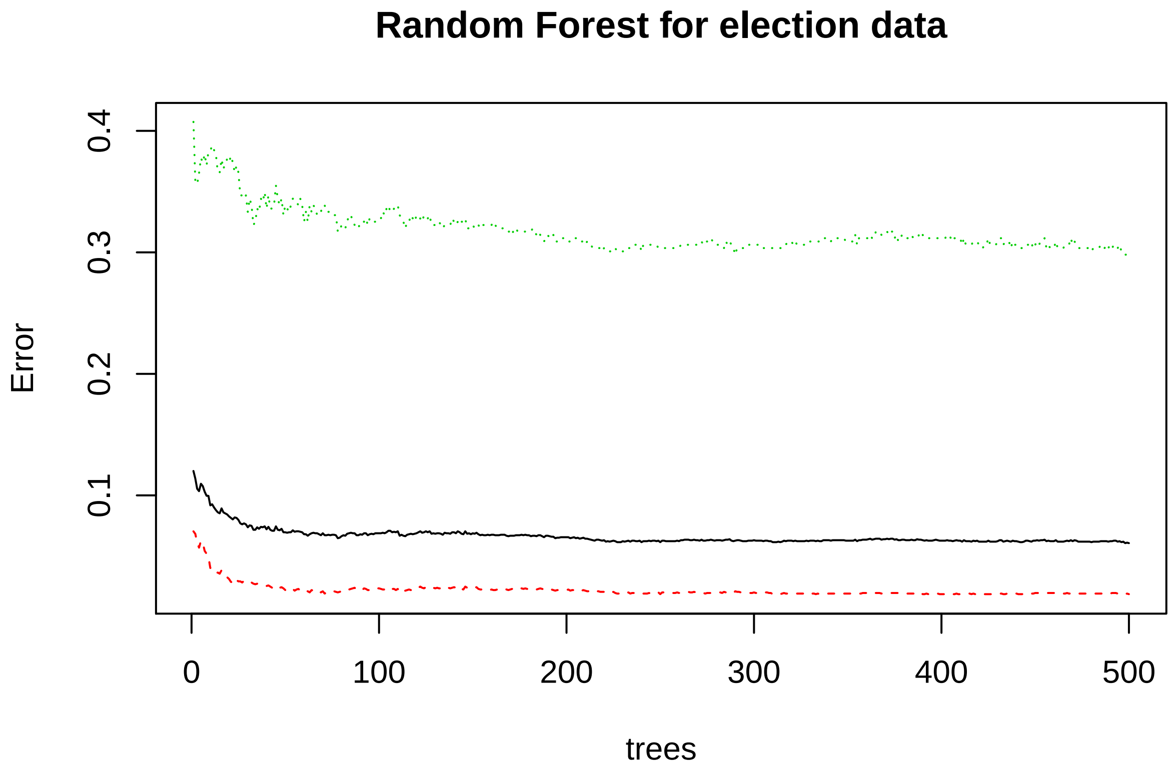
SVM

In fitting an SVM model, the optimal cost for the best model proves to be 0.01. Using this value in our model, I then viewed the absolute size of the coefficients, and saw that the coefficients for White, Minority, Transit, Professional, Drive, and Employed have the largest influence on voters decisions.

Random Forest

Using the random forest model, the top 5 influential predictors are Transit, Minority, White, Professional and Unemployment. This makes sense, considering the initial split from the decision tree is on the Transit variable. Again, because the use of public transportation proves to be of great concern in big cities with dense populations, it falls in line with our previous analysis of why this factor is so important. White was also a main splitting variable included in the decision tree from earlier as well. The out of bag estimate for the error rate is 6.19%.


rf.election
Transit
White
Minority
Professional
Unemployment
Employed
IncomePerCap
Drive
Income
Service
Production
Poverty
IncomePerCapErr
MeanCommute
WorkAtHome
Men
PrivateWork
SelfEmployed
Office
TotalPop
Citizen
ChildPoverty
IncomeErr
OtherTransp
Carpool
FamilyWork
10
20
30
40
MeanDecreaseAccuracy
Transit
White
Minority
Professional
Unemployment
IncomePerCap
Production
Poverty
Income
Employed
Service
Drive
Men
IncomePerCapErr
IncomeErr
SelfEmployed
Citizen
OtherTransp
Office
MeanCommute
ChildPoverty
TotalPop
PrivateWork
WorkAtHome
Carpool
FamilyWork
0
40
80
MeanDecreaseGini



Boosting

The boosting model shows Transit as the variable with the most influence. The five most influential predictors are Transit, White, Minority, Unemployment and Professional, varying just slightly from the top five predictors for the random forest model. This model also minimizes test error the best according to the confusion matrix shown below. This model along with the random forest model show to be very promising in building an efficient algorithm for predicting elections, and should be incorporated for future presidential elections.

```
##              train.error test.error
## svm          0.074511401 0.07479675
## randomforest 0.001221498 0.05853659
## boosting     0.027280130 0.04552846
```