UNIVERSITY OF CALIFORNIA, SANTA BARBARA

PSTAT231 FINAL PROJECT

12/17/2020

Analysis of 2016 United States Presidential Election

Director: Prof. Zhijian Li

Report by Jiwon Baik 9897331
Meiyu Li 6979074

Contents

1 Introduction

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

   Although polling information can tell us voters' behaviors, it is still hard to make an accurate prediction, since many factors are contributing to the complexity of an individual's choices. Comparing to factors like gender, age, or region, some factors are too complex to measure and may vary a lot in voters' decision making. Factors, like religious beliefs, scandals influence, mood, and atmosphere during the election, are hard to measure. These add more complexity to make voter behavior prediction. From Carl Bialik (2016), there is a lot we may ignore, such as the final margins in many states, whether the miss was due to systematic problems among pollsters. These are also the factors contributing to the difficulty of predicting voters' behaviors.

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

   From O'Hara (2012), we know that Nate used statistical and mathematical models such as hierarchical modeling which allows information to move around the model. Additionally, Nate tried to combine Bayes' theorem and graph theory, a combination of knowledge on statistics and mathematics, to calculate the probabilities of each level of support. Besides the technical and reliable models, Nate also concerned factors beyond polls, such as voters' employment, race, wealth, and even the bias of pollsters. Based on the experience of predicting previous election and reliable models, Nate achieved a good prediction in 2012.

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

   From Carl Bialik (2016), there were polling errors coming from statistical noise and from factors more difficult to quantify, such as nonresponse bias aka. Survivorship bias. Also, polls are vulnerable in all states to systematic errors, and these errors usually spread unevenly nationwide. Besides these, there is a phenomenon that women who voted for Trump might have been especially reluctant to tell pollsters, and this underestimates Trump's votes. Furthermore, any possible campaign elements could affect the predictions, so the only single number that looks similar to other polls is not favored.

2 Data

This section includes the answers for question 4-5, and 12. We are using datasets election.raw(18345 observations and 5 variables) and census(74001 observations and 36 variables). Some important attributes are attached below:

election.raw

- county (summary rows=NA)
- fips (federal-level summary rows=US, state-level summary rows=states values)
- candidate (32 in total)
- votes (number of votes received)
- state

Notes: the values in the variable fips denote the area (US, state, or county) that each row of data represent. We found out that the information for state=AK(Alaska) is duplicated. In other words, the counties in Alaska state has two rows for each. To avoid repetition, we removed the rows with fips=2000 which stands for the repeated information for state=AK(Alaska). The original dimension of election.raw was 18351, but after the duplicates removal, the dimension of election.raw is 18345 rows and 5 columns.

Based on the election.raw, we created three datasets as following: election_federal (federal-level summary), election_state (state-level summary), and election (only county_level data).

census

- Men (Number of men)
- Women (Number of women)
- White (percent of population that is white)
- Citizen (Number of citizens)
- Unemployment (percent of population that is unemployed

- Poverty (percent of population that is under poverty level)
- Drive (percent of population that is commuting alone in a car, van, or truck)
- Employed (Number of employees)
- ...

As census data contains high resolution information, more fine-grained than county-level, the data is aggregated into county-level data. Each sub county observations were aggregated by TotalPop (total population) weighted average.

Also, the census data had some variables that could be converted into percentage scale, which makes the comparison between variables could be easier. Those are Men, Employed, and Citizen. Also, the Hispanic, Black, Native, Asian, Pacific variables had the number of population from each racial group. The numbers of population were summed up to get the total population of Minority groups. And the Minority variable was converted into the percentage scale. The first five rows of census.ct are printed in the table (1). After converting the variables, it is checked whether there is another variable when summed up together the total percentage is 100. Because it means one of them is not necessary and better to be removed to improve efficiency. Even though Unemployment and Employed, or White and Minority didn't sum up to be 100, Men and Women were summed up to be 100. So the Women variable was deleted.

census.ct

- Men (percent of male population)
- White (percent of population that is white)
- Citizen (percent of citizens)
- Unemployment (percent of population that is unemployed
- Poverty (percent of population that is under poverty level)
- Drive (percent of population that is commuting alone in a car, van, or truck)
- Employed (percent of employees)
- ...

	State	County	Men	White	Citizen	Income	IncomeErr	IncomePerCap
1	Alabama	Autauga	48.43	75.79	73.75	51696.29	7771.01	24974.5
2	Alabama	Baldwin	48.85	83.1	75.69	51074.36	8745.05	27316.84
3	Alabama	Barbour	53.83	46.23	76.91	32959.3	6031.06	16824.22
4	Alabama	Bibb	53.41	74.5	77.4	38886.63	5662.36	18430.99
5	Alabama	Blount	49.41	87.85	73.38	46237.97	8695.79	20532.27

	IncomePerCapErr	Poverty	ChildPoverty	Professional	Service	Office	Production
1	3433.67	12.91	18.71	32.79	17.17	24.28	17.16
2	3803.72	13.42	19.48	32.73	17.95	27.1	11.32
3	2430.19	26.51	43.56	26.12	16.46	23.28	23.32
4	3073.6	16.6	27.2	21.59	17.96	17.47	23.74
5	2052.06	16.72	26.86	28.53	13.94	23.84	20.1

	Drive	Carpool	Transit	OtherTransp	WorkAtHome	MeanCommute	Employed
1	87.51	8.78	0.1	1.31	1.84	26.5	43.44
2	84.6	8.96	0.13	1.44	3.85	26.32	44.05
3	83.33	11.06	0.5	1.62	1.5	24.52	31.92
4	83.43	13.15	0.5	1.56	0.73	28.71	36.69
5	84.85	11.28	0.36	0.42	2.27	34.84	38.45

	PrivateWork	SelfEmployed	FamilyWork	Unemployment
1	73.74	5.43	0	7.73
2	81.28	5.91	0.36	7.59
3	71.59	7.15	0.09	17.53
4	76.74	6.64	0.39	8.16
5	81.83	4.23	0.36	7.7

Table 1: First 5 rows of census.ct

3 Exploratory Data Analysis

This section includes the answers for question 6 to 11.

3.1 Election result of each candidate

The Figure (1) includes multiple bar charts. From left to right, from top to bottom, please note that the range of the x-axis of the charts is getting increased. Each bar chart shows how many votes were received by each 32 candidate. So the candidates in the latter chart received more votes than candidates in the previous chart. The last chart has a color notification of the parties where the final contestants, Donald Trump and Hillary Clinton, are in. Interestingly, even though the winner of the election was Donald Trump, Hillary Clinton got more votes in total. So the election system matters a lot to define a winner.

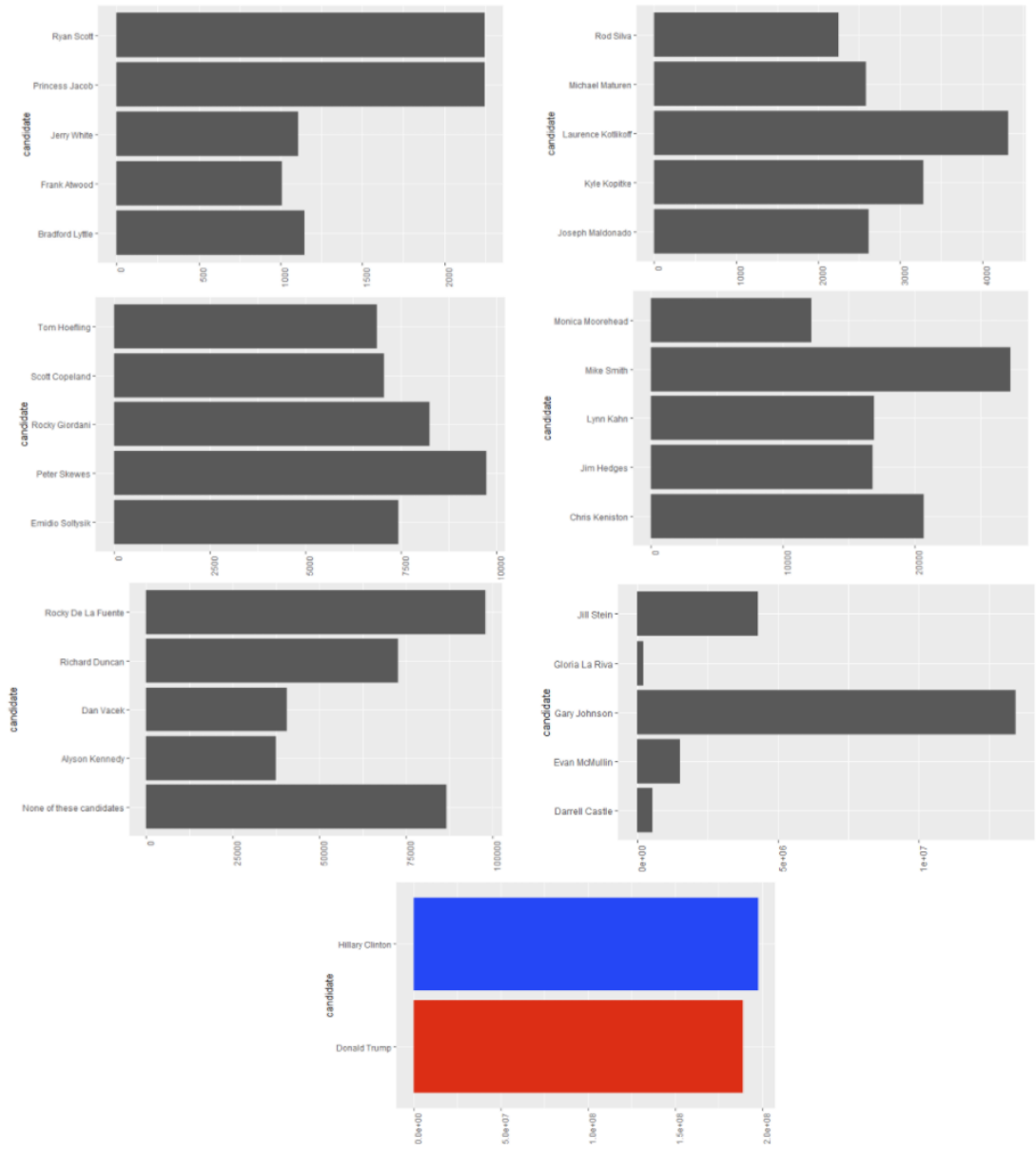

Figure 1: Bar charts of total votes received by each candidate

3.2 The election result of the final contestants

Figure (2) indicates the County-level base map used for the election result maps. Each polygon divided by the white line is county and colored based on the state. Figure (3) includes two election result maps in different spatial scale. The first map indicates the winner of each state. Donald Trump won in the red colored states and Hillary Clinton won in the blue colored states. Trump won in 29 states, and Clinton won in 19 states. The second map indicates the winner of each county. Donald Trump won in 2607 counties, and Hillary Clinton won in 462 counties. So

even though Hillary got more votes in total, Trump won in more electoral districts. The effect of electoral district could be rediscovered here again.

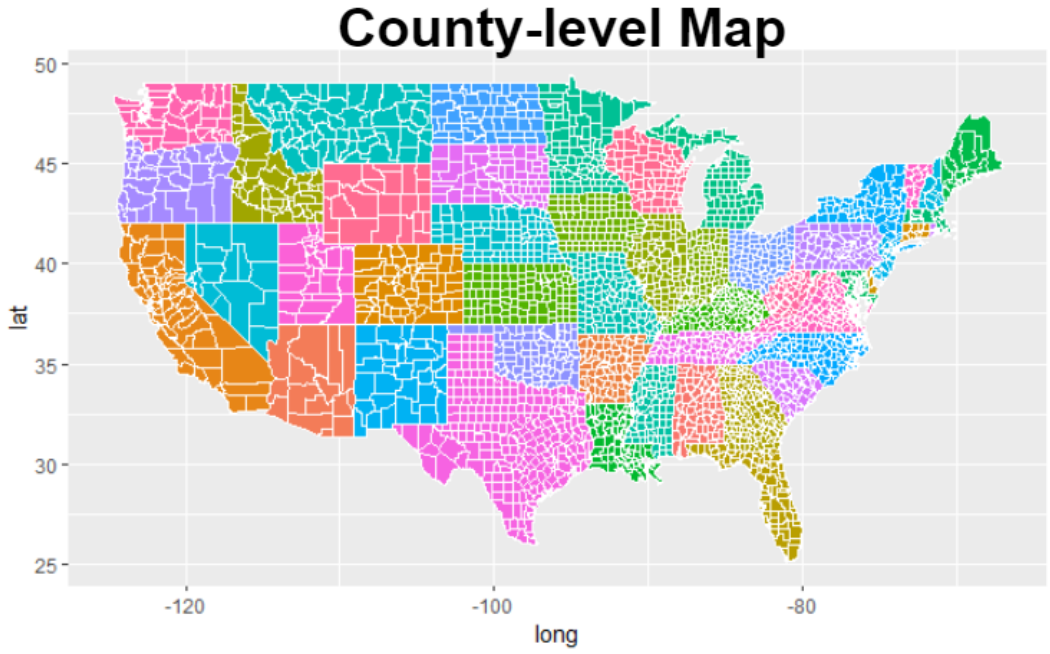

Figure 2: County-level map

3.3 Demographics explaining the election result

It is well known that demographics play a big role in elections. In this paper, we used six variables to plot the correlations: White, Employed, Professional, IncomePerCap, Citizen, and Transit. Figure (4) has the correlation plots of the varialbles. In the left part, scatter plots are drawn with each variable on the x-axis and y-axis. A county is represented as a point, in each scatter plot. Each point is colored to show which candidate won the county. In the right part, three values are presented. The first value in each cell is the overall correlation of the variables on the x-axis and y-axis.The remaining two values are correlations calculated by collecting only the cases where each candidate wins.

First thing to note is that the variables are highly correlated to each other. All the variables are correlated to each other very closely. All the correlations are very significant with at least 0.1 correlation coefficient. It's reasonable that more employed county has higher income per capita, and more professional people to get more employed. So they might cause inefficiency when put into the models.

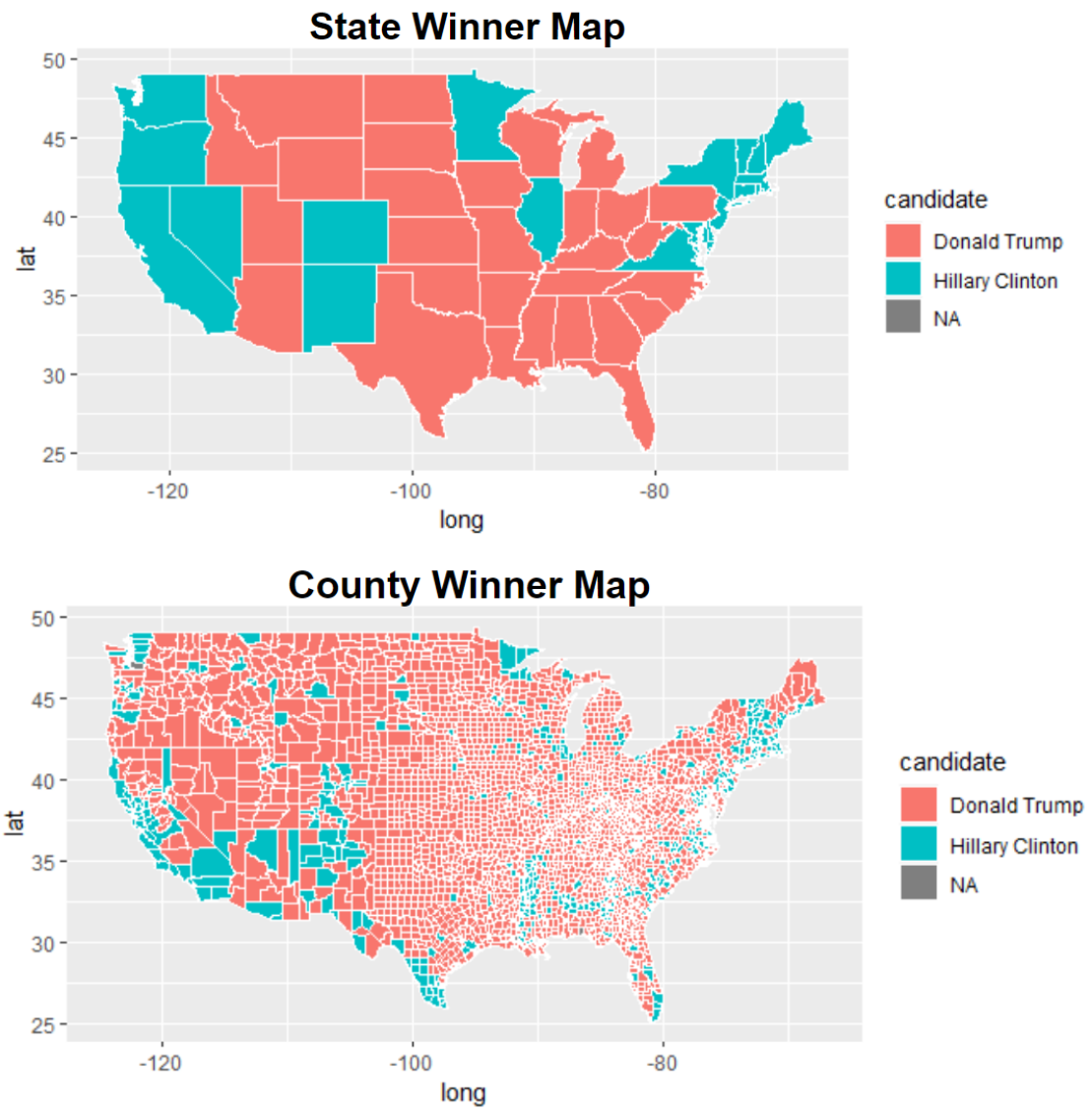

Figure 3: State and County Winners map

The transit variable has very small values, majority of them are very close to zero. So the last plot seems like not showing the density functions but actually there is.

The diagonal cells show two density functions separated based on the winning candidate by each. With the diagonal part, we wanted to know which is the best variable where the distribution of Trump voters and Hillary voters was separated the most. It was the White variable. Even though it's not a clear separation, the two distributions could be distinguished the most compared to the other variables. So further spatial visualization was conducted on the variable.

Figure (5) has two maps indicating spatial distribution of White proportion in

Figure 4: Plots of four demographic variables: White, IncomePerCap, Professional, Employed

population in each county. First map indicates the proportion of white population in each county by the color scale of each county polygon. Red color indicates higher white people ratio, while blue color indicates lower ratio. And we could tell that the distribution is very similar to the county winner map presented in the Figure (3). In case the polygon representation distorts the truth due to the different size, the second map is created. The second map has points for each county. Each point is colored based on the candidate won in the county. And the point size is proportional to the white people ratio. So please note the different size of blue points and red points. Comparing the second map to the first map, we could easily tell that often blue counties are very hard to find due to the small size of the points. It indicates that in general, counties where Hillary won have less white people ratio, while counties where Trump won have higher white people ratio.

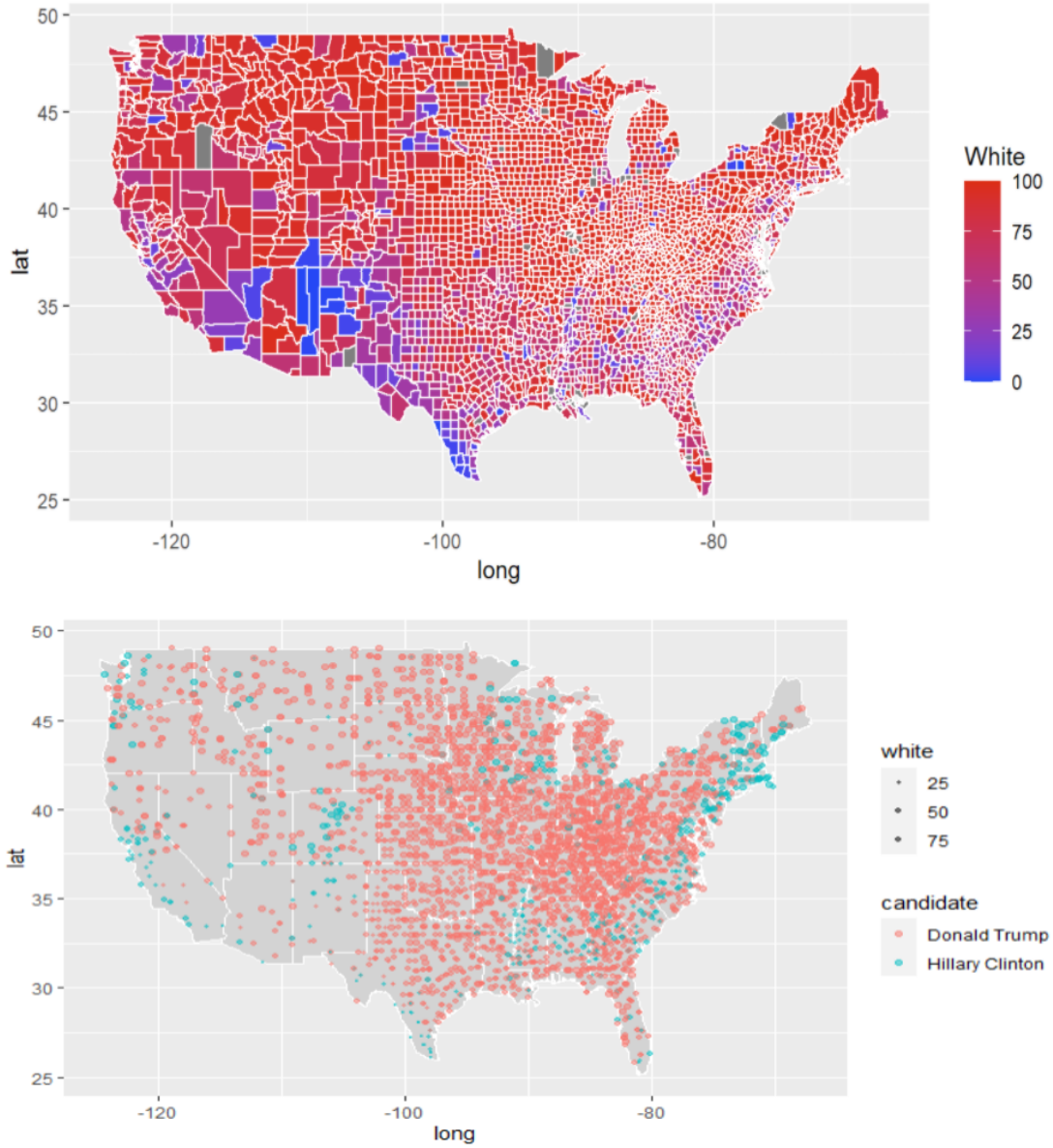

Figure 5: Spatial distribution of White proportion in population

4 Method

Basically, we first do principal component analysis and try to compute the principal components. Then, we use them to perform a change of basis on the data. After doing that, we do hierarchical clustering and explore the case study at San Mateo County. Furthermore, we randomly partition data into 80percent training set and 20percent testing set. Additionally, we select 5 different models to predict voter behaviors, and the models include decision tree, logistic regression, lasso logistic regression, random forest, boosting, and support vector machine. Based on the

output, we try to compare the test errors and ROC curves to reach our best model. In the end, we conduct an exploratory analysis of the purple counties-the counties which the models predict Clinton and Trump were roughly equally likely to win. After analysis, we share some thoughts about why these counties are hard to predict.

5 Result

5.1 Principal Component Analysis

This section is an answer to the question 13. The Principal Component Analysis(PC) was conducted on both county and sub-county level data. Regarding that the variables included in each data have different ranges of value, scaling and centering option was enabled in the PCA. For example, referring to the Table (1), the Income related columns have large values, while the population related percentage columns such as Men, White, Citizen have a value range 0 to 100. As PCA is very sensitive to the unit of data, the centering and scaling is needed to make the PCA works in a robust way.

5.1.1 The first and second Principal components for each data

The three features with the largest absolute values of the first principal component are "IncomePerCap", "Poverty", "ChildPoverty" in the county-level data, while in the subcounty-level data "IncomePerCap", "Poverty", "Professional". The PC1 and PC2 components values for each feature could be checked in table (2) and (3). Table (2) has the PCs from the county-level data, while Table (3) has the PCs from the subcounty-level data.

5.1.2 The features with opposite signs in the first Principal Component

The features with opposite signs are listed in table (4) and (5) for county-level and subcounty-level data respectively. The features that have opposite signs have negative correlation. For example, the Income, White, Employed, Professional features have a negative sign, while Poverty, Minority, Unemployment, Service features have a positive sign in the first principal component from the county-level data. They are the features with negative correlation between each other.

	PC1	PC2
Men	-0.0048	0.1355
White	-0.2177	0.2927
Citizen	-3e-04	0.2396
Income	-0.3226	-0.2074
IncomeErr	-0.1738	-0.3145
IncomePerCap	-0.3531	-0.1389
IncomePerCapErr	-0.1969	-0.2071
Poverty	0.3406	-0.056
ChildPoverty	0.3422	-0.0406
Professional	-0.252	-0.1094
Service	0.1802	-0.0589
Office	0.0115	-0.2453
Production	0.1212	0.1434
Drive	0.095	-0.0303
Carpool	0.0772	0.0369
Transit	-0.0765	-0.2778
OtherTransp	0.0086	-0.0591
WorkAtHome	-0.1725	0.2157
MeanCommute	0.0556	-0.1929
Employed	-0.3274	-0.003
PrivateWork	-0.0589	-0.182
SelfEmployed	-0.0939	0.3088
FamilyWork	-0.0463	0.2088
Unemployment	0.2876	-0.1589
Minority	0.2213	-0.289
CountyTotal	-0.0625	-0.2932

Table 2: First 2 principal components of census data at county-level

so we could infer that the correlation between the features with opposite signs is negative.

5.1.3 The Proportion of Variance Explained Plots

This section is an answer to the question 14. The Proportion of Variance Explained (PVE) plots show how much variance of the response variable is explained by each principal component. Figure (6) has the PVE and cumulative PVE plots of county-level data. Figure (7) includes the PVE and cumulative PVE plots of Sub county-level data. To explain 90% of variance, the county-level analysis requires 16 principal components, and sub county-level analysis requires 14 principal components. So at least 16 Principal components are needed to capture 90% of the variance for both the county and sub-county analysis.

	PC1	PC2
TotalPop	-0.0324	0.0079
Men	-0.0175	-0.0392
White	-0.2411	-0.3097
Citizen	-0.1613	-0.2296
Income	-0.3022	0.1546
IncomeErr	-0.1985	0.2275
IncomePerCap	-0.3178	0.1726
IncomePerCapErr	-0.212	0.2056
Poverty	0.3047	0.0585
ChildPoverty	0.2979	0.0324
Professional	-0.3062	0.147
Service	0.269	0.0588
Office	0.0141	-0.0656
Production	0.2064	-0.1958
Drive	-0.0794	-0.4022
Carpool	0.1625	-0.0387
Transit	0.0579	0.4089
OtherTransp	0.0453	0.1444
WorkAtHome	-0.173	0.1106
MeanCommute	-0.0095	0.2851
Employed	-0.221	0.0502
PrivateWork	0.0426	0.0181
SelfEmployed	-0.0702	0.0443
FamilyWork	-0.0155	-0.0297
Unemployment	0.253	0.0771
Minority	0.2426	0.3069
CountyTotal	0.0221	0.2815

Table 3: First 2 principal components of census data at subcounty-level

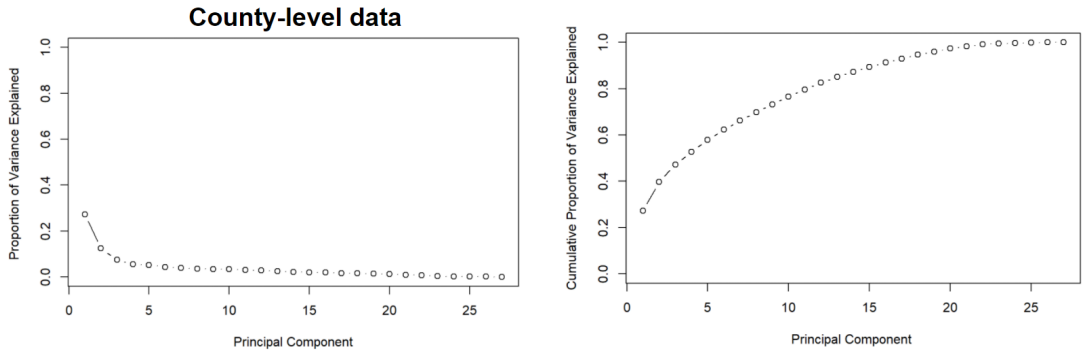

Figure 6: County-level PVE plots

	negative	positive
1	Men	Poverty
2	White	ChildPoverty
3	Citizen	Service
4	Income	Office
5	IncomeErr	Production
6	IncomePerCap	Drive
7	IncomePerCapErr	Carpool
8	Professional	OtherTransp
9	Transit	MeanCommute
10	WorkAtHome	Unemployment
11	Employed	Minority
12	PrivateWork	
13	SelfEmployed	
14	FamilyWork	
15	CountyTotal	

Table 4: The features from county-level data separated by Positive and Negative values in PC1

	negative	positive
1	Men	Poverty
2	White	ChildPoverty
3	Citizen	Service
4	Income	Office
5	IncomeErr	Production
6	IncomePerCap	Drive
7	IncomePerCapErr	Carpool
8	Professional	OtherTransp
9	Transit	MeanCommute
10	WorkAtHome	Unemployment
11	Employed	Minority
12	PrivateWork	
13	SelfEmployed	
14	FamilyWork	
15	CountyTotal	

Table 5: The features from subcounty-level data separated by Positive and Negative values in PC1

5.2 Hierarchical Clustering

This section is an answer to the question 15. In this paper, Hierarchical Clustering has been applied with two data. The first data is the original data in scaled version. As in the PCA part the data was scaled because of the different ranges

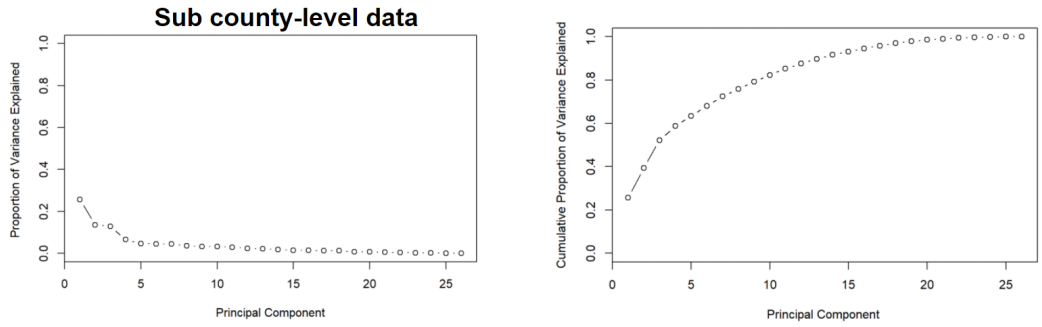

Figure 7: Sub county-level PVE plots

between columns. So to be fair, hierarchical clustering for the original data was also conducted with scaling. The other data is the 5 Principle Components. Using the 5 Principle components, each observation from the scaled original data turns into a sequence of 5 values.

5.2.1 Case study about San Mateo County

To examine which approach works better with Hierarchical Clustering, a case of San Mateo County is selected. To see whether the clustering is appropriate, the objective of the clustering analysis must be settled. In this paper, the voting result is in specific interest. So the better approach would be an approach resulting more similar clusters to the election result.

The real winner in the San Mateo county is Hillary Clinton. Based on the fact, we could check the distribution of the other clusters in the same cluster with San Mateo county. Figure (8) shows the 10 hierarchical clusters created with the scaled original data and 5 principal components data. It would be concluded that 5 Principal Components approach is more appropriate to detect a pattern in election result. The hierarchical clustering with scaled data classified San Mateo as class 2, which is indicated with gold color in the first plot. The distribution of the class 2 counties is less synchronized with the distribution of counties where Hillary Clinton won. On the other hand, the 5 PC approach classified San Mateo as class 7. The blue color in the second plot. The distribution of blue counties in the second plot synchronizes well with the distribution of counties where Hillary won. So the PCA approach worked better.

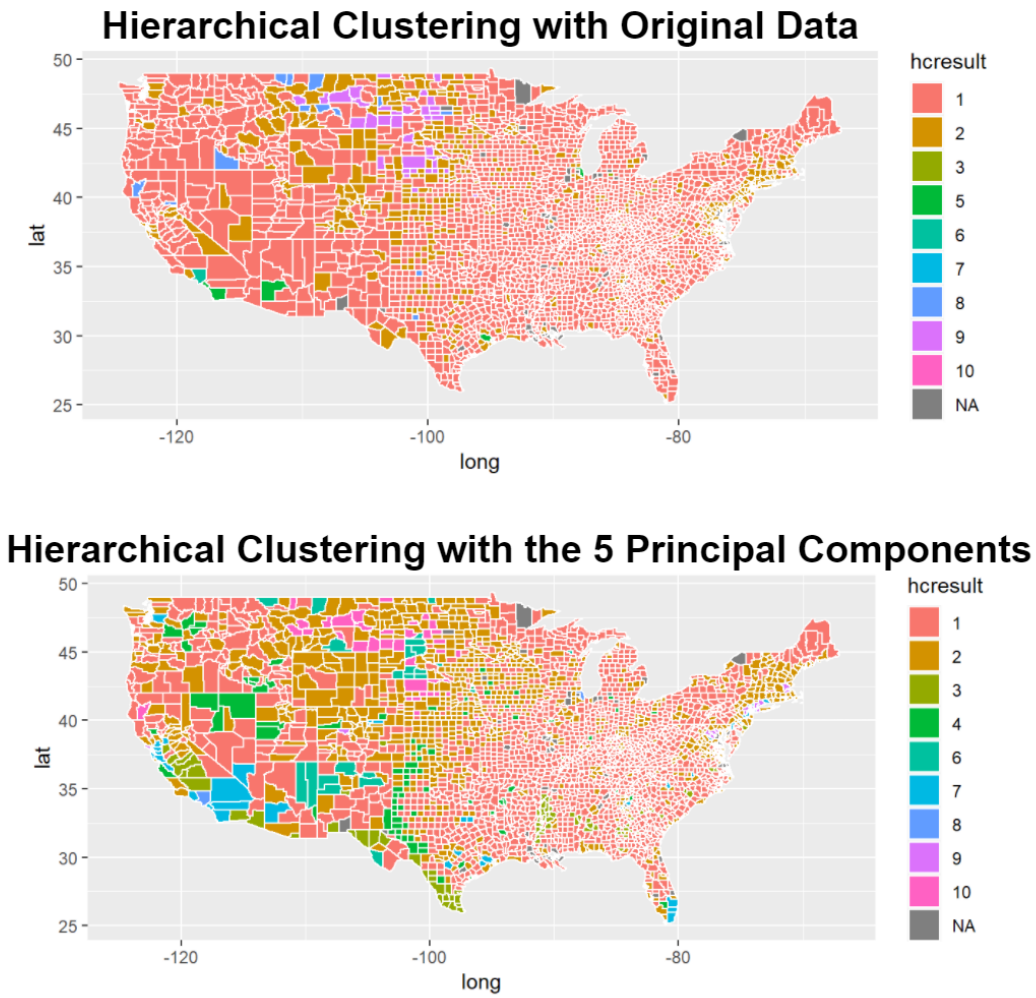

Figure 8: County-level PVE plots

5.3 Classification

5.3.1 Decision Tree

Specifically, this section is for question 16. We pruned tree to minimize the misclassification error and use folds defined in the handout for cross-validation. We got 17 terminal nodes in total. The misclassification error rate is 0.06474. In this case, our best tree size is 9. The variables actually used in the tree construction are transit, white, unemployement, citizen, production, drive, employed, selfemployed and countyTotal. In this question, we got two decision trees which are before and after pruning. The pruned trees look legible and clean, so we will use the pruned tree to analyze the behaviors of voters.

Also, we calculated test error=0.06504065 in decision tree.

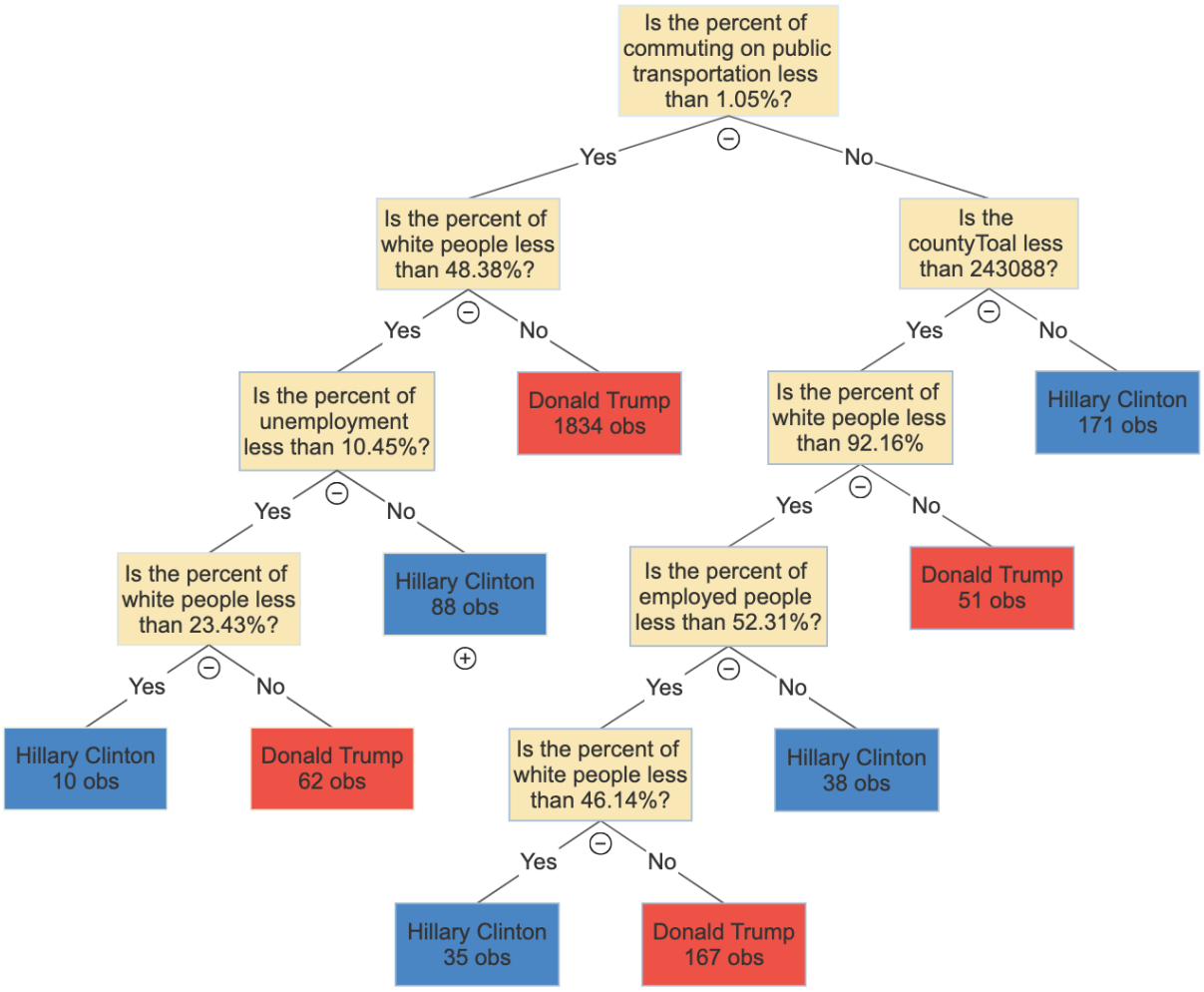

Figure 9: Decision Trees Story

5.3.2 Random Forest

Specifically, this section is a part of question 20. Random Forest is a kind of bagging method that decorrelates each tree. Instead of using all of the predictors at each split of the tree, the random forest algorithm only uses part of the predictors. We explored additional classification method, random forest and made further analysis. Here is the confusion matrix we got:

	Donald Trump(truth)	Hillary Clinton(truth)
Donald Trump(prediction)	524	22
Hillary Clinton(prediction)	7	62

We can calculate $TPR = \frac{TP}{TP+FN} = \frac{524}{524+7} = 0.99$ and type-I error $FPR = \frac{FP}{FP+TN} = \frac{22}{22+62} = 0.26$. We can conclude that the probability of detection is

pretty much higher and the probability of false alarm is low. Also, we calculated test error=0.04715447 in random forest.

Besides these, we also plot out the Accuracy and Gini for each variable. We found that the variable transit has the most decrease in Gini and Accuracy when we process the random forest. Thus the variable transit may be the most important predictor. Also family work, whether the voter works at home, private work may not be very effective for prediction.

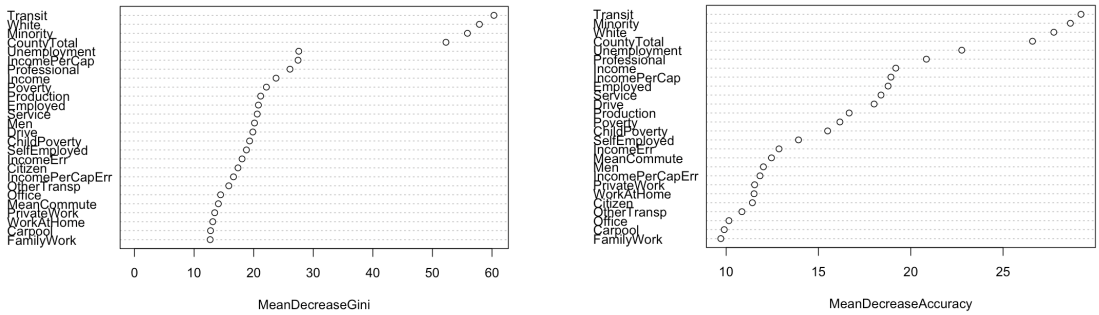

Figure 10: Gini and Accuracy

5.3.3 Boosting

Specifically, this section is part of question 20. We explore additional classification method, boosting and made some analysis. Boosting is an ensemble method for improving the model prediction. The idea of boosting is to train weak learners sequentially, each trying to correct its predecessor. Since boosting requires the response to be coded as $\{0, 1\}$, we change candidate=Hillary Clinton to 1 and candidate=Donald Trump to 0. Then we fit the boosting model with the training set, setting n.trees=500 and interaction.depth=4 to limit the depth of each tree. Combining the figure and data, we can see that Transit and White are by far the most important variable. he results we got from boosting is similar to what we got for random Forest. Transit is important for both random forest and Boosting method. Also, we predict the test set and obtain the confusion matrix. The test error for boosting is 0.39024390.

Variable	Relative Influence
Transit	21.2751433
White	20.2777808
CountyTotal	9.5264701
Citizen	4.6022123
Minority	4.4294446
Unemployment	4.0667729
Employed	3.8394432
Professional	3.7361700
Service	3.2003940
Poverty	2.4854296
...	...

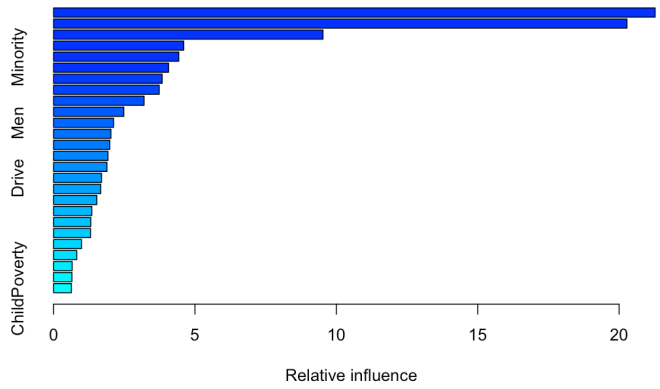

Figure 11: Boosting Method

	Donald Trump(truth)	Hillary Clinton(truth)
Donald Trump(prediction)	294	3
Hillary Clinton(prediction)	237	81

5.4 Logistic Regression

This section is specifically for questions 17-19. To solve the binary classification problem, we build a logistic regression model on training data. We calculate the training error=0.07043974 and test error=0.06341463. Based on the p-values for each variable, we found the significant variables are proverty, childPoverty, transit, selfEmployed, and so on. These are consistent with what we are doing for decision tree part. To interpret some coefficients in terms of a unit change in the variables,

we choose the numeric variable transit. To interpret the coefficient transit, in table 7, for fixed other values, the percentage of commuting on public transportation increases one unit, the logit function increase $\beta_1 = 0.07578$ units.

5.4.1 Lasso Logistic Regression

We notice that we get a warning: glm.fit: fitted probabilities numerically 0 or 1 occurred. This is an indication that we have perfect separation, a sign indicating we are overfitting. Now we try to use cv.glmnet function to run 1-fold cross valiadation and select the best regularization parameter for the logistic regression with LASSO penalty. We found the optimal $\lambda = 5e - 04$. The nonzero coefficients are {Men, White, Citizen, Income, IncomeErr, IncomePerCap, IncomePerCapErr, childPoverty, Poverty, Professional, Service, Office, Production, Drive, Carpool, Transit, WorkAtHome, MeanCommute, Employed, PrivateWork, FamilyWork, Unemployment, CountyTotal}. Comparing to unpenalized logistic regression, lasso logistic regression tend to give zero coefficients, since lasso performs shrinkage and subset selection. Also, in this case, logistic regression has lower test errors than lasso regression. For lasso logistic regression, the training error=0.06962541 and test error=0.06504065.

Besides, comparing the test errors, we also compute the ROC curves for decision tree, logistic regression, and Lasso logistic regression using predictions on the test data.

So far, we have done the decision trees, logistic regression, and lasso regression. To answer question 19, we compare these three methods and explore the advantages and disadvantages of these three.

For decision tree, we may not require normalization of data or scaling of data as well. Also, missing values in the data also do NOT affect the process of building a decision tree to any considerable extent. Additionally, decision tree is very intuitive and easy to explain. However, decision tree is not stable, since a small change in the data can cause a large change in the structure of the decision tree. Comparing to logistic regression, decision tree's calculation can go far more complex. Since anything may happen or data is not stable in the election case, so decision tree may not be the best method to get our model.

For logistic regression, it makes no assumptions about distributions of classes and it can easily extend to multiple classes. In this election case, we have multiple variables, logistic regression can consider all variables together. However, it may

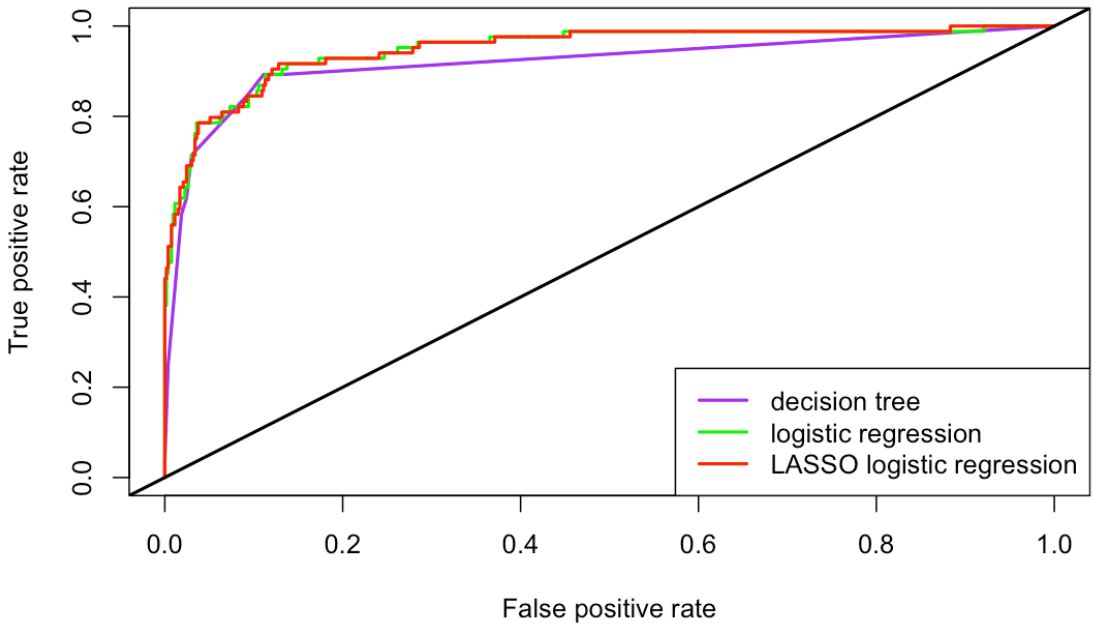

Figure 12: ROC Cruves

not perform very well if the data is not linearly separable. Also, it usually predicts the discrete variable, so we need to transform our data into character for this case.

For lasso regression, as any regularization method, it can avoid overfitting and it can do feature selection. However, The model selected by lasso is not stable and it is hard to intrepret. When there are highly correlated features, lasso may randomly select one of them. In this case, if we have some highly correlated variables, we may miss information while doing lasso regression.

5.5 Support Vector Machine

This section is also part of Question 20. Support vector machine is a machine learning algorithm, which aims to find the optimal hyperplane to appropriately classify the data points of different classes. In this case, we try svm with linear kernel, radial kernel and poly kernel by using svm() function and setting the same gamma=0.5 and same cost=1. Here are the test errors for each kernel:

kernel	test errors
linear	0.06666667
radial	0.13333333
poly	0.12195122

As we can see, when the kernel is linear, the test error is least. Then, we summarize the tune() function with a range of list(cost=c(0.001, 0.01, 0.1, 1, 10)) and get the best model. The number of support vectors is 473. The cost=0.1 has the smallest error. The best support vector machine model is when kernel=linear and cost=0.1. After we get the best.performance for our model, the more accurate error is 0.07084619 which is very closer to what we previously got for kernel=linear, 0.06666667.

So far, we have all the possible methods and here is the test errors for each method. As we can see, the random forest has the smallest test error. However, it is too absolute to see random forest is the best model, We still need to consider other possibilities. As we discussed before, lasso logistic regression has bigger area under the ROC curve (AUC), but the test error of lasso regression is a little bit bigger than the logistic regression.

METHODS	test errors
decision tree	0.06504065
logistic regression	0.06341463
lasso logistic regression	0.06504065
boost	0.39024390
random forest	0.04715447
support vector machine	0.07084619

Besides this, we also have our records of training errors and test errors for the decision trees, logistic regression, and lasso logistic regression. Comparing these three test errors, we found logistic regression has the smallest test error. Logistic regression may be better than lasso regression and decision tree.

METHODS	train.error	test.error
decision tree	0.06555375	0.06504065
logistic regression	0.07043974	0.06341463
lasso logistic regression	0.06962541	0.06504065

5.6 EDA on "Purple" counties

This section is an answer to the question 20, going further. "Purple" Counties are the counties where the probability of Hillary's wining and Trump's winning is almost the same. So it was hard to predict who's going to win the counties. As the symbolic color of Trump is red and Hillary is blue, the middle color Purple is

included in it's name. In this paper, the purple counties are redefined in a practical meaning. Counties which the models predict Clinton and Trump were roughly equally likely to win. The two models used to detect Purple counties in this paper are logistic regression and random forest. Because the two models showed the best performances in the previous method comparison part. The test errors were the smallest. And decision tree is a classification model, while logistic regression is a regression model. So they would capture different aspects of purple counties. A county which was predicted the probability of Trump's winning between 0.45 to 0.55 by logistic regression or random forest was detected as purple county. So when any one model thinks the county as purple county, it become a purple county.

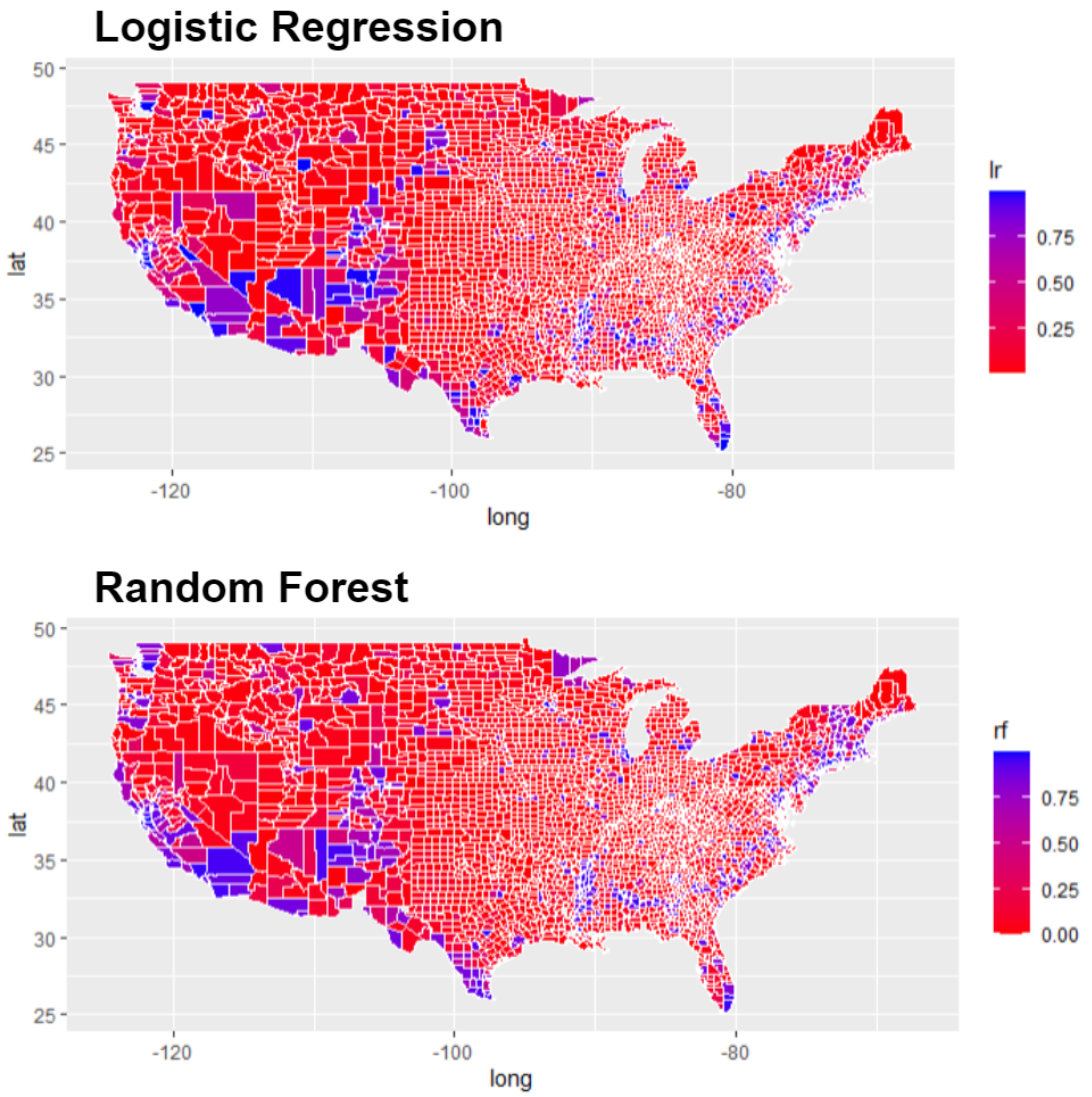

Figure 13: Logistic Regression and Random Forest models probability prediction of Hillary's winning

Figure (13) shows the spatial distribution based on the predictions of each model, Logistic Regression and Random Forest. The red color indicates higher probability of Trump's winning in the county. The blue color indicates the county where probability of Hillary's winning is higher. Logistic Regression predicted 53 counties to be purple, while Random Forest predicted 20 counties to be purple. They are mainly similar to each other. Figure (14) shows the spatial distribution of detected purple counties from each model. We could tell that they are predicting different counties as purple counties because the regression analysis and classification analysis are strong at catching different traits. So any one model detected a county as a purple county, the county will be treated as a purple county.

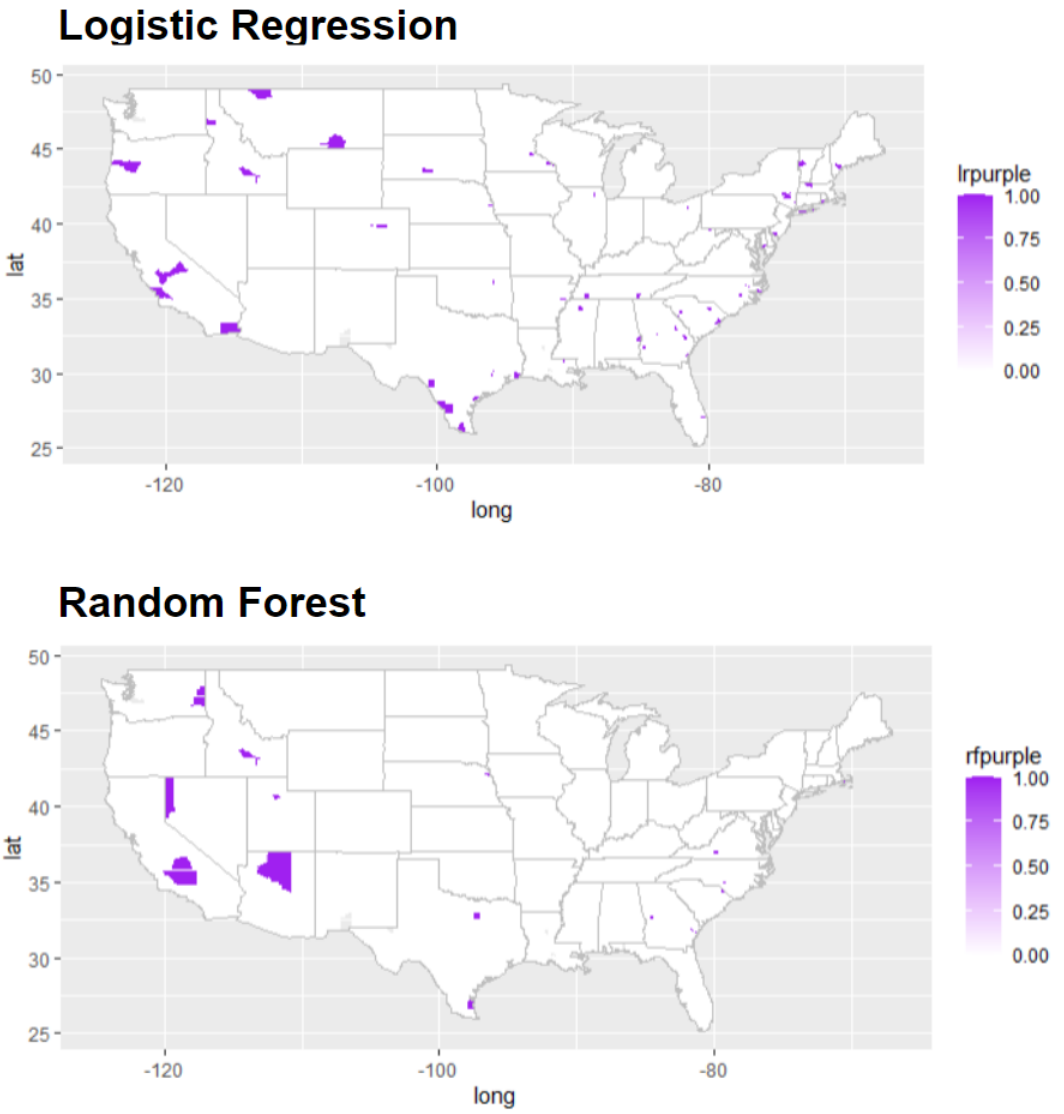

Figure 14: Purple counties from Logistic Regression and Random Forest models

(a) purple county: the probability of winning is between 0.45 to 0.55

	ST	CT	candidate	White	UNEM	IncomePC	Profess
1	alabama	russell	Hillary Clinton	49.81	11.25	19881.73	28.4
2	arizona	coconino	Hillary Clinton	54.67	9.65	24308.23	33.65
3	arkansas	st francis	Hillary Clinton	41.22	11.39	15970.47	23.88
4	california	fresno	Hillary Clinton	31.2	13.56	20495.21	26.84
5	california	imperial	Hillary Clinton	12.69	17.84	16508.42	22.67
6	california	kern	Donald Trump	36.82	13.05	21038.28	23.96

Table 6: Example of five counties which are purple

Table (6) has the logistic regression coefficients and p-value of each coefficient. So p-value is from Z test, testing the null hypothesis that the coefficient is zero. There are 17 variables which are statistically significant to have non-zero coefficient. As the response variable is 0 or 1, the estimated coefficients are very small. But it doesn't mean that the variable is not related to the response variable. The valid variables are Citizen, Professional, Service, Production, Drive, Employed, PrivateWork, Unemployment, IncomePerCap, Income, Carpool, Intercept, White, FamilyWork, MeanCommute, WorkAtHome, IncomePerCapErr, Men. These variables could be compared to the variables in Figure (10). It has the variable importance plot of random forest. The top 9 variables contributed to the mean decrease of Gini are Transit, White, Minority, CountyTotal, Unemployment, IncomePerCap, Professional, Income, Poverty. Threre are four variables which are valid in logistic regression, and also contributing to the random forest than the others. They are White, Unemployment, IncomePerCap, and Professional.

So further exploratory analysis was applied on the four variables which are valid and influential in both models, logistic regression and random forest. Figure (16) and Figure (17) are the correlation plots. The difference is whether each observation is separated due to the purple county or not. What we want to check is whether there are any variables that makes the prediction hard so that making the county purple. Figure (16) shows the regressed lines between two variables. The left side of the scatter plots in the first column indicates non-purple counties and right side of the scatter plots indicates purple counties. Clear decreasing or increasing pattern could be found between purple and candidate, and purple and White only. So in purple counties, it came out for people to vote for Hillary. Also, purple counties tend to have less White proportion in its population. In Figure (17) we could take a closer look at each variable where observations are separated by the purple county or not. The distinction between the density plot is more prominent in White and Unemployment variables than the other variables. And

	Estimate	Std..Error	Pr...z..	valid
Citizen	0.1302	0.028	0	valid
Professional	0.2802	0.0387	0	valid
Service	0.3242	0.0476	0	valid
Production	0.1668	0.0409	0	valid
Drive	-0.2097	0.0462	0	valid
Employed	0.2056	0.0335	0	valid
PrivateWork	0.101	0.0219	0	valid
Unemployment	0.2073	0.0397	0	valid
IncomePerCap	3e-04	1e-04	1e-04	valid
Income	-1e-04	0	0.0014	valid
Carpool	-0.1736	0.0593	0.0034	valid
(Intercept)	-24.8708	9.4517	0.0085	valid
White	-0.1689	0.0676	0.0125	valid
FamilyWork	-0.8873	0.3765	0.0184	valid
MeanCommute	0.0562	0.0242	0.0204	valid
WorkAtHome	-0.1657	0.072	0.0213	valid
IncomePerCapErr	-4e-04	2e-04	0.0356	valid
Men	0.0959	0.0482	0.0468	valid
Office	0.0759	0.0447	0.0892	not
Poverty	0.0474	0.0405	0.2419	not
CountyTotal	0	0	0.4141	not
Transit	0.0758	0.0942	0.4213	not
OtherTransp	-0.0626	0.095	0.5102	not
ChildPoverty	-0.0158	0.0246	0.5197	not
Minority	-0.0305	0.0651	0.639	not
SelfEmployed	0.0197	0.0468	0.6739	not
IncomeErr	0	1e-04	0.9583	not

Table 7: Logistic Regression Model coefficients and p-value

it corresponds to the top 2 highest correlation coefficients with purple variable in the previous plot.

Based on the plots, and model summaries, it could be infer that less white proportion in each county, and more Unemployment rate tends to make the county purple. It is also related to the less IncomePerCapita, and less proportion of people who are working in Professional vocations. It makes sense because Less white proportion means more possibility of racial diversity which leads to the confusion in predicting the winner. More unemployment, less professional vocations and income per capita could be related to various styles of life and different interests among people. And it makes the voting behavior hard to converge.

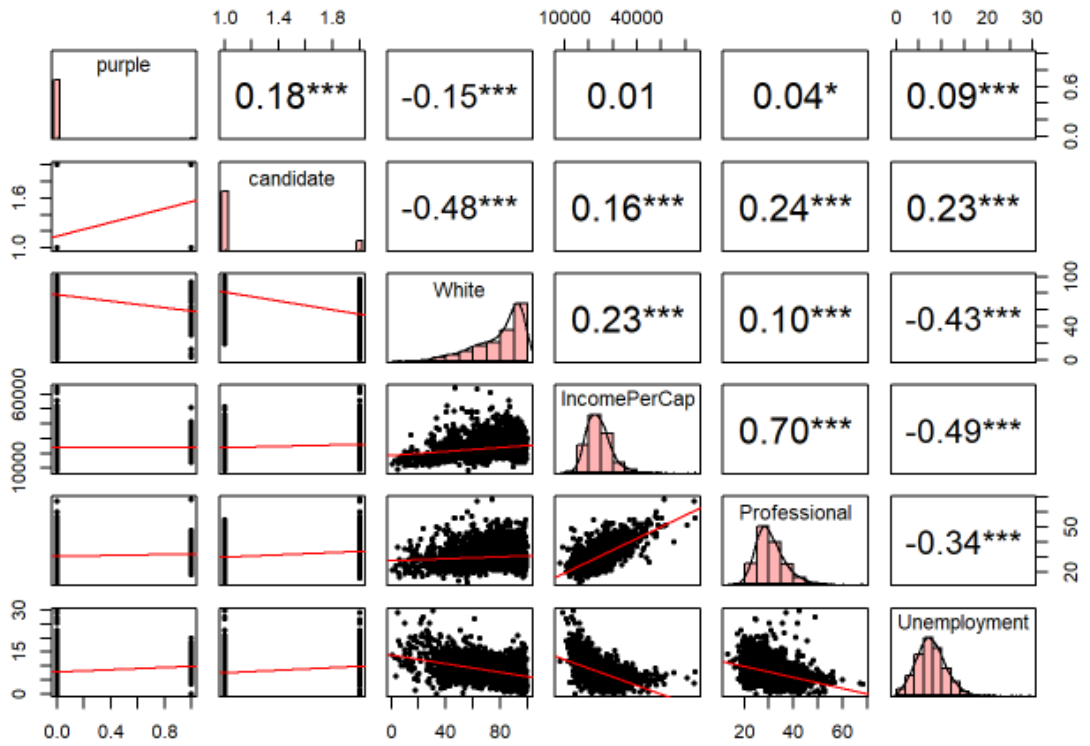

Figure 16: Correlation plots between purple binary variable and the others

Figure 17: Correlation plots with separation between purple and non purple counties

6 Conclusion

After processing 6 methods including decision tree, random forest, boosting, logistic regression, lasso regression, support vector machines, we compare the test errors and conclude that random forest has the smallest test error. Random forest is the best method to help us find our best model. It adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, random forest searches for the best feature among a random subset of features. Also, random forest avoids the overfitting problem that may occur in logistic regression. From the model, we can see transit, white, minority are significant variables, and this is consistent with what we got for decision tree and boosting. The percentage of commuting on public transportation is a main factor of 2016 election. Also, the percentage of voters of white race is also a significant factor that can influence the prediction. This may be due to the percentage of white voters is pretty large. Besides these two, important variables including citizen and unemployment also have huge influence on prediction of voters' behaviors.

References

Carl Bialik, H. E. (2016). The Polls Missed Trump. We Asked Pollsters Why. *FiveThirtyEight*.

O'Hara, B. (2012). How did Nate Silver predict the US election? *the Guardian*.