Effective Machine Learning Models to Forecast The 2016 Presidential Election of The United States

Le Song
The Asian Institute of Digital Finance
and
Department of Statistics and Data Science
National University of Singapore
Singapore
le.song@u.nus.edu

Zhongyun Zhang
Department of Applied Mathematics and Statistics
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD, USA
zzhang222@jhu.edu

Abstract

The forecast of the result of the United States presidential elections is a popular topic every four years in the United States and worldwide. The forecast results were usually correct while their efficacy of them went away for the 2016 presidential election. Back at that time, most models argue that Hillary Clinton is much more possible to win the presidential election. However, Donald Trump became the president at last. In this paper, we investigate the poll-based data, find out how 2016 is different from before and develop and optimize the effective machine learning models for the 2016 presidential election. By comparing the result of our models and the test set split from the original data and calculating the error rate, and by further analysis through calculating their Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC), we justify the effectiveness of our machine learning models.

1. Introduction

Predicting voter behavior is complicated for many reasons despite the tremendous effort in collecting, analyzing, and understanding many available datasets.

According to the models from most well-known media before the presidential election, Hillary Clinton had a much greater chance of winning than Donald Trump the majority of the time. As an instance, Figure 1 is a line plot of time series that shows the chance of winning calculated by FiveThirtyEight.com, which is an American website that focuses on opinion poll analysis, politics, economics, and sports blogging. From this figure, we can see clearly how the models perform in the period from June 8th to November 8th (the date of the election).

Despite what is calculated by the models, on November 8th, Trump won the Electoral College with 304 votes compared to 227 votes for Clinton. Either the previous machine learning models could be modified, or the data of the 2016 presidential election has special parts that are different from the previous elections that worth be discussed.

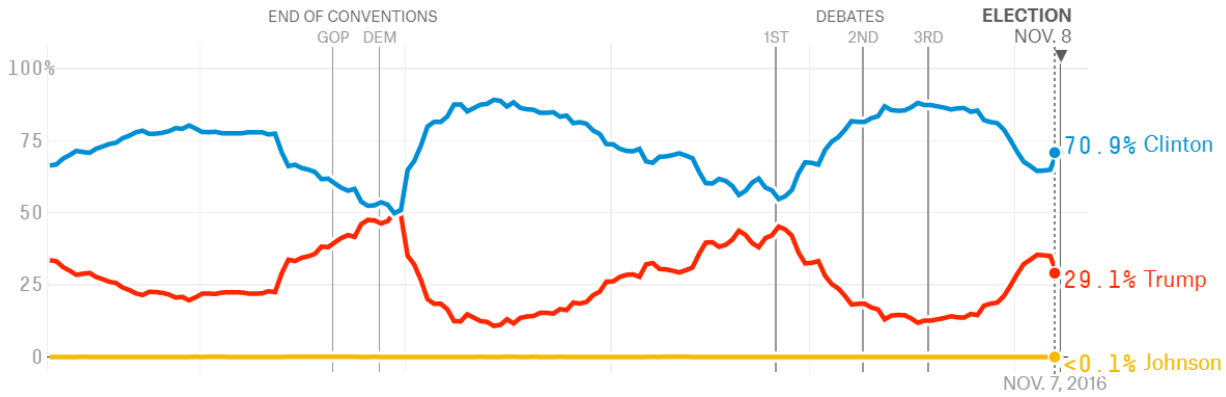


Figure 1: Chance of Winning Calculated by FiveThirtyEight.com in Time Series

2. Background

The model behind Figure 1 was developed by an American statistician, Nathaniel Read Silver (Nate Silver). The presidential election in 2012 did not come as a surprise. Some predicted the outcome of the election correctly including Nate Silver, and many speculated on his approach. Several statistical models and machine learning models were adopted by him to analyze the data, such as Bayesian methods, hierarchical clustering, and graph theory, adjusted predictions for states using statistical machine learning, and get the result based on how states with similar demographics are responding to polls. They took a further insight into the problems of the data, which we discussed in problem 1, by using such as actual percentage, the house effect, and sampling variation. They avoided many significant sources of errors and thus were able to successfully achieve good predictions for the 2012 election result.

However, some significant problems in the 2016 poll are that there is a high bias toward Hillary Clinton's winning the election. Not only tendentious poll questions but also ignored many voters for Donald Trump. Thus, we can see that certain demographics were under or overestimated in their support for Clinton or Trump. Moreover, the 2016 election is a much fiercer competition than previous elections, leading to a result that a candidate's behavior or comments on a certain event can show a strong reaction in polls, gaining new supporters as well as pushing old supporters away. Thus, what we should do is take these errors into account of the methodology and avoid biases, and therefore have a model that can predict more unforeseen outcomes.

3. Data Visualization and Discussion on the Rationality of the Data Source

In this paper, we use the raw election data provided on the FiveThirtyEight database as our data source. Almost all state and national polls are included in the raw dataset. We believe the completeness and the authenticity of the dataset are sufficient.

3.1 Election Data

In our dataset for the election, “fips” values denote the area (the nation, state, or county) that each row of data represents. For example, in Figure 2, the “fips” value of 6037 denotes Los Angeles County.

county	fips	candidate	state	votes
Los Angeles County	6037	Hillary Clinton	CA	2464364
Los Angeles County	6037	Donald Trump	CA	769743
Los Angeles County	6037	Gary Johnson	CA	88968
Los Angeles County	6037	Jill Stein	CA	76465
Los Angeles County	6037	Gloria La Riva	CA	21993

Figure 2: Data Rows for Los Angeles County

As shown in Figure 3, in the raw dataset, fip = 2000 represents Alaska, and we saw that it does not have any county inside the data. After we researched on it, it shows that Alaska does not have a county, which is noted with "NA". Therefore, we remove fip = 2000, since Alaska does not have county level votes.

county	fips	candidate	state	votes
NA	2000	Donald Trump	AK	163387
NA	2000	Hillary Clinton	AK	116454
NA	2000	Gary Johnson	AK	18725
NA	2000	Jill Stein	AK	5735
NA	2000	Darrell Castle	AK	3866
NA	2000	Rocky De La Fuente	AK	1240

Figure 3: Data Rows with fips = 2000

The dimension of our raw data after removing rows with fips=2000 is 18345 observations and 5 columns.

For the 2016 presidential election, there are 36 named presidential candidates in total. Figure 5 shows the plot for the top 10 candidates.

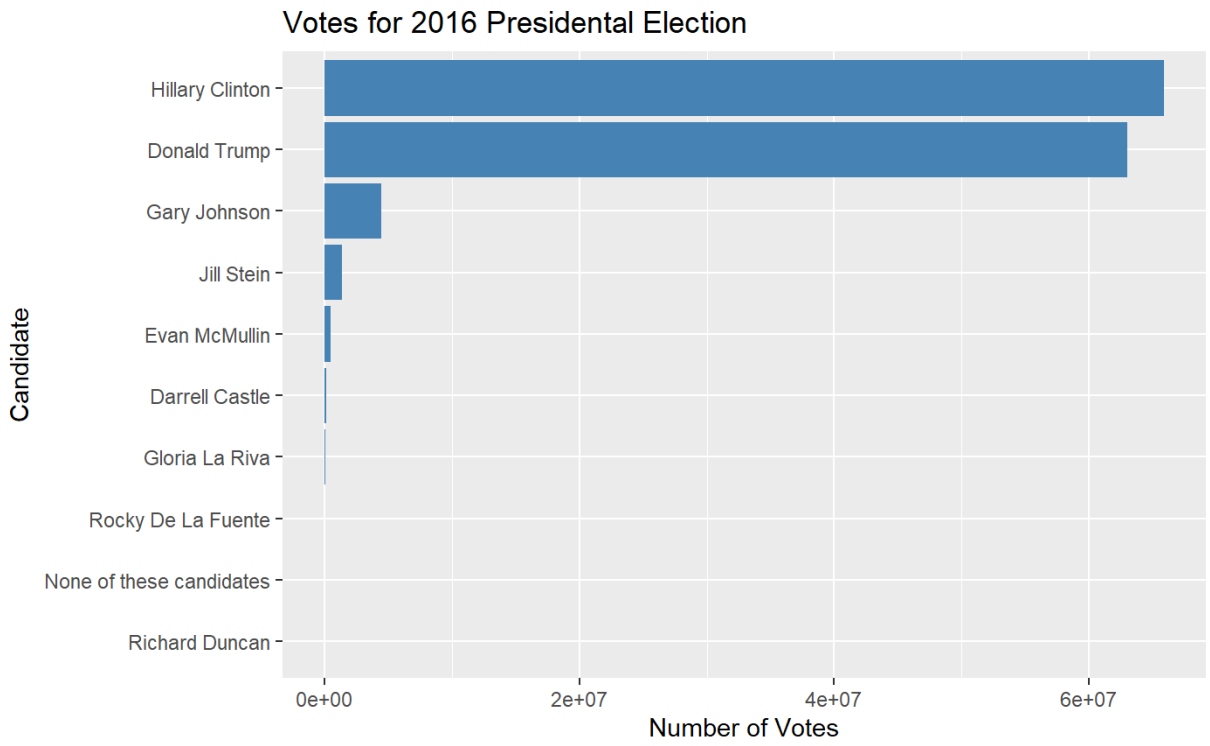


Figure 5: Votes for 2016 Presidential Election

3.2 Census data

Besides the election data, we include the census data for data processing.
There are 74001 rows and 36 features in total for this dataset.
Below shown in Figures 5 and 6 are the first few rows and columns of the census dataset.

State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	Asian	Pacific	Citizen	Income	IncomeErr
Alabama	Autauga	1948	940	1008	0.9	87.4	7.7	0.3	0.6	0.0	1503	61838	11900
Alabama	Autauga	2156	1059	1097	0.8	40.4	53.3	0.0	2.3	0.0	1662	32303	13538
Alabama	Autauga	2968	1364	1604	0.0	74.5	18.6	0.5	1.4	0.3	2335	44922	5629
Alabama	Autauga	4423	2172	2251	10.5	82.8	3.7	1.6	0.0	0.0	3306	54329	7003
Alabama	Autauga	10763	4922	5841	0.7	68.5	24.8	0.0	3.8	0.0	7666	51965	6935
Alabama	Autauga	3851	1787	2064	13.1	72.9	11.9	0.0	0.0	0.0	2642	63092	9585

Figure 5: First Few Rows of The Census Dataset

Citizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty	Professional	Service	Office	Construction
1503	61838	11900	25713	4548	8.1	8.4	34.7	17.0	21.3	11.9
1662	32303	13538	18021	2474	25.5	40.3	22.3	24.7	21.5	9.4
2335	44922	5629	20689	2817	12.7	19.7	31.4	24.9	22.1	9.2
3306	54329	7003	24125	2870	2.1	1.6	27.0	20.8	27.0	8.7
7666	51965	6935	27526	2813	11.4	17.5	49.6	14.2	18.2	2.1
2642	63092	9585	30480	7550	14.4	21.9	24.2	17.5	35.4	7.9

Figure 6: Example Columns for The Census Dataset

3.3 Distribution

We combine and modify the two datasets mentioned above based on the state name. We plot the distribution of the supporters of two major rivals in this campaign Donald Trump and Hillary Clinton. In Figure 7, red represents Donald Trump and blue represents Hillary Clinton.

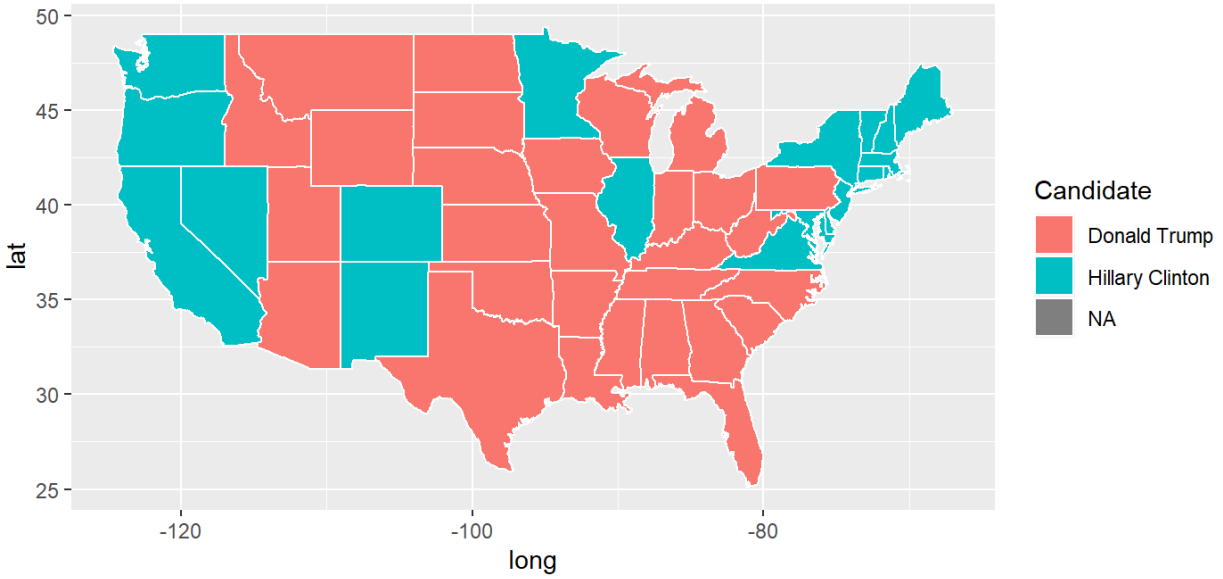


Figure 7: Distribution of the Votes

Furthermore, we did a detailed data processing for each subregion (county). In the United States, there are 3085 counties, and the results of the election are shown in Figure 8. And the same as in Figure7, red represents Donald Trump and blue represents Hillary Clinton.

We can find out clearly that the people living in rural areas tend to support Trump and the people living in metropolitan areas tend to support Clinton.

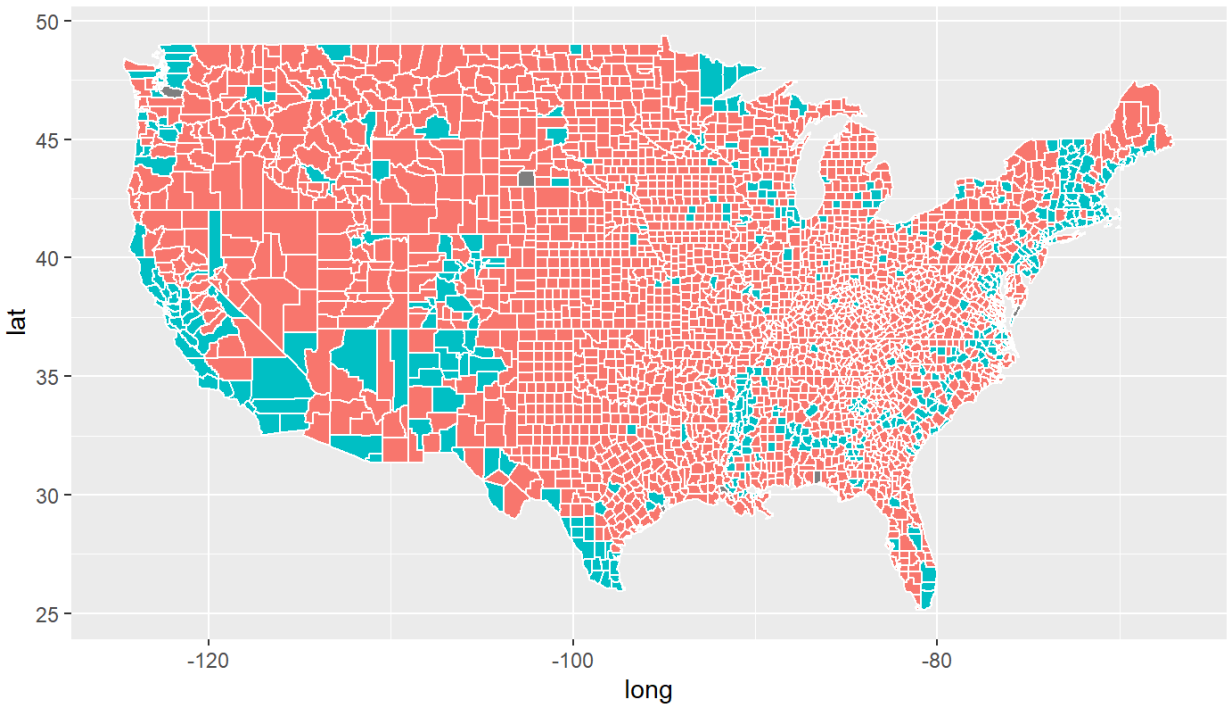


Figure 8: County Level Distribution of Votes

3.4 Bias in Data

Additionally, another crucial index for the election is the Income Per Capita (Figure 9) and the Poverty Level (Figure 10). The two indices are inversely proportional to each other. Considering these two charts together with the previous geographic distribution of the voters, we find that Trump is doing better in poorer areas than Clinton. As we know, people with higher income levels usually have a stronger voice in the media and polls to express their political views.

However, a weaker voice does not mean that the poorer people lose their right to vote. We propose a reasonable hypothesis that, for the 2016 presidential election, the elector groups that support each side (Trump versus Clinton) differ much more than before, which significantly raises the bias in the data source from the polls.

Figure 9: Income Per Capita Aggregated by States

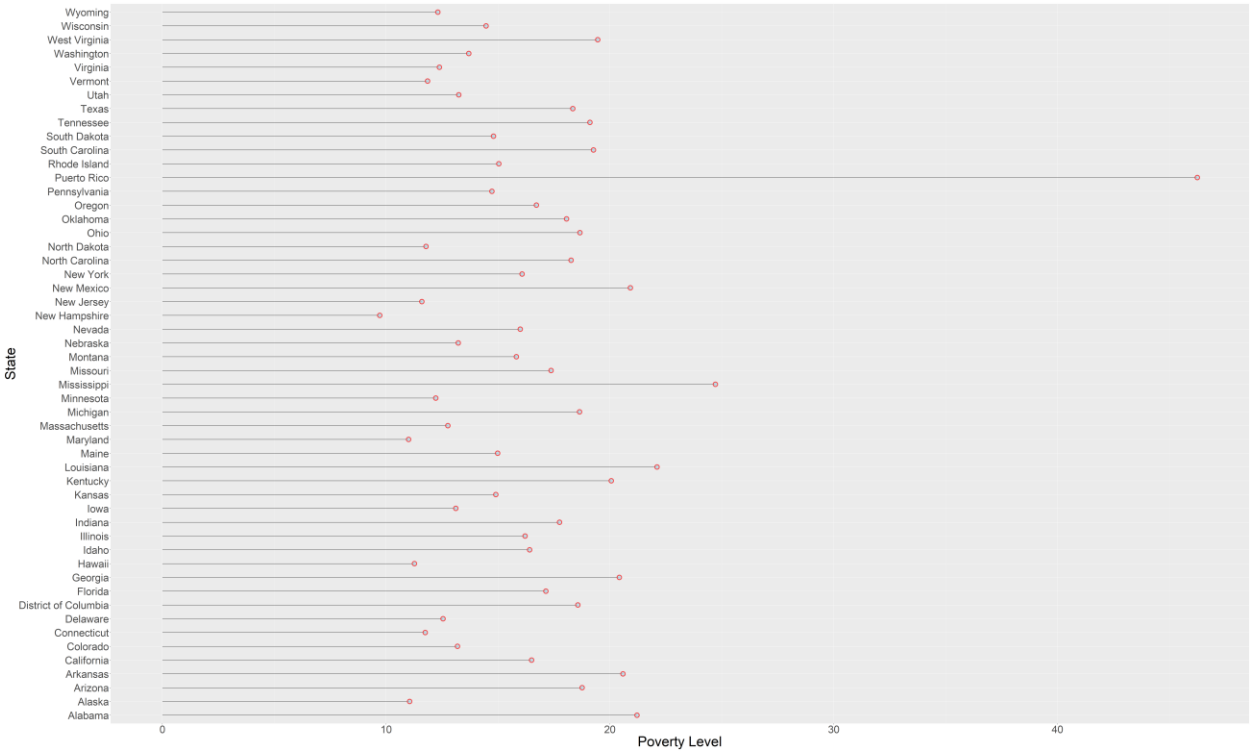


Figure 10: Poverty Level Aggregated by States

3.5 The reason that makes voter behavior prediction, and thus election forecasting, a hard problem.

To begin with, people are voting based on many inner factors contribute to the voter behavior, for example, race, gender, age, income, etc. Also, their decisions may vary due to media, political propaganda, family, and friends, etc. Thus, it is difficult to create an efficient, accurate sampling model.

Moreover, Voters are changing their opinions as time flies. Presidential candidates are always competing to get the votes from these states to gain votes from swinging counties. The propaganda, families, and friends' persuasion, and most importantly the political instability may vary one's mind. Candidates' committing to a certain issue important to voters can gain new supporters as well as push old supporters away, and these changes are hard to quantify.

Additionally, some votes are hard to predict due to unforeseen reasons. For example, occurrence events may affect whom voters will vote to. In addition, there is also bias while voting, people may not be the same opinion when they actually vote. When scientists use census data to predict the actual result, census data may not be inclusive and representative enough. Bias from pollsters will also influence the result, different questions can lead the same person to a different response. People's true opinions can only be represented at the millisecond when they vote at the ballot box. Therefore, election forecasting can be considered difficult.

4. Data Processing

The census data contains high resolution information, that is, more fine-grained than county-level. We aggregate the information into county-level data by computing the weighted average for TotalPop variable of each attribute for each county. Create the following variables:

To clean the census data, we start with the “census” dataset and filter out any rows with missing values. We convert {Men, Employed, Citizen} attributes to percentages, which the “meta data” seems to be inaccurate, and we compute Minority attribute by combining {Hispanic, Black, Native, Asian, Pacific}. After that we remove these variables after creating Minority and meanwhile, we remove {Walk, PublicWork, Construction}.
Many columns seem to be related, and, if a set that adds up to 100%, one column will be deleted.

For the Sub- County census data, we start with the original dataset from above. We take the existing columns and group them by two attributes, State and County. We perform add_tally() function to compute County Total value (the total population living in each county). Then, we compute the weight by dividing the value of the total population by the value of the County Total value.

After computing the weighted sum of the data, we get the data frame after processing shown in Figure 11.

A tibble: 3,218 x 30 Groups: State [52]

State <chr>	County <chr>	TotalPop <dbl>	Men <dbl>	White <dbl>	Citizen <dbl>	Income <dbl>	IncomeErr <dbl>	IncomePerCap <dbl>	IncomePerCapErr <dbl>	Poverty <dbl>
Alabama	Autauga	4601.750	48.36759	73.15000	74.97871	49985.00	8035.583	24386.92	3703.667	14.075000
Alabama	Baldwin	6294.226	48.82690	83.45484	76.36530	48672.84	9268.065	26842.77	4007.871	14.364516
Alabama	Barbour	2992.444	52.22358	46.61111	76.31093	32367.67	5891.444	17104.56	2523.111	26.788889
Alabama	Bibb	5651.000	53.24573	77.50000	76.82747	40211.50	6142.750	18807.00	3340.000	15.600000
Alabama	Blount	6412.222	49.53062	87.80000	73.53580	45101.11	8920.111	20171.33	2052.556	17.300000
Alabama	Bullock	3559.333	51.78404	22.00000	76.04840	33445.33	8511.333	17735.00	3367.667	25.566667
Alabama	Butler	2261.556	46.65430	56.62222	77.14022	34192.33	7544.222	18892.89	2704.222	24.766667
Alabama	Calhoun	4165.857	47.77278	67.25357	76.22790	38152.89	6955.750	19977.61	2718.107	24.782143
Alabama	Chambers	3786.556	48.39813	55.04444	78.59200	33351.33	7061.667	21144.67	4270.000	22.466667
Alabama	Cherokee	4334.667	49.84339	91.63333	79.05828	37047.83	4712.667	21833.33	3388.833	19.450000
Alabama	Chilton	4868.778	49.40090	81.91111	73.87143	41878.11	9226.222	21859.00	3528.000	18.300000
Alabama	Choctaw	3348.750	47.71141	53.47500	78.87779	34029.00	7753.750	20312.25	2534.250	24.800000
Alabama	Clarke	2785.556	47.22199	55.17778	76.26482	31588.22	9327.333	19099.11	4849.889	26.644444
Alabama	Clay	3384.250	49.41073	81.72500	75.95981	33411.00	6562.750	18010.75	2269.750	17.750000
Alabama	Cleburne	3750.500	49.22456	92.55000	75.96225	37079.75	6322.250	19868.75	2523.500	17.475000
Alabama	Coffee	3634.571	49.18530	75.18571	75.17165	44105.86	7129.714	23157.79	2630.786	18.492857

1-16 of 3,218 rows | 1-11 of 30 columns Previous 1 2 3 4 5 6 ... 63 Next

Figure 11: The Data after Processing

5. Methods

5.1 Dimensionality Reduction

Recall Figure 11, we can observe that there are so many columns in the original dataset. The number of columns, or in the terminology of machine learning, the number of feature variables is too much to do modeling and analysis.

According to the book *Pattern recognition and machine learning*, in machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon. (Bishop, Christopher 2006)

Each feature is a dimension of our data frame. When analyzing a high-dimensional space, the data contains redundant information and noise information, which causes errors in practical applications such as image recognition and reduces the accuracy. Through the application of dimensionality reduction, we reduce the errors caused by redundant information and improve the accuracy of identification and other applications.

5.2 Principle Component Analysis

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables. (James, Witten, Hastie, Tibshirani, 2013).

The normalized linear expression of the combination of the features is:

$$Z_j = \phi_{1j} X_1 + \phi_{2j} X_2 + ... + \phi_{pj} X_p$$

where $X = (X_1, X_2, X_3, \ldots X_p)$ is the dimension of the data we have. Some parts of the polynomial are less significant. And the purpose of performing principal components analysis is to find out the most significant part and reduce the complexity of a large dataset.

The data before performing PCA has many features that have different scales. For instance, the "Income" feature is a value that has a size of ten thousand levels (larger than 30000), and in comparison, the "White" feature is shown in a percentage that has only two figures. The size of the "Income" feature is too large beside the "White" feature and makes it undiscoverable. The difference between them decreases the effectiveness of the result of the analysis of components.

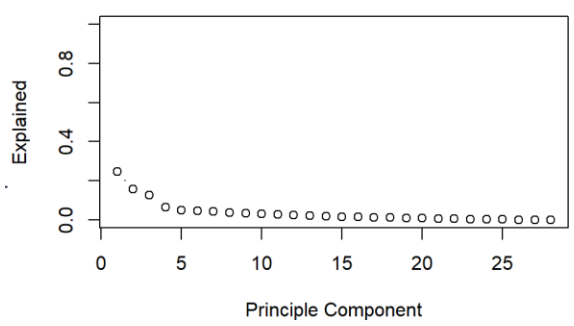
To figure out this problem, we shift the variables to zero centered scale and scale the value to have unit variance before the analysis takes place. And therefore, the variables in our dataset can have the standard deviation and the same mean.

After performing the PCA computation, we save the first two principal components PC1 and PC2 into a two-column data frame, and name it ct.pc and subct.pc dataset, respectively.

For county level data, the features with the largest absolute values of the first principal component are "IncomePerCap", "ChildPoverty", and "Employed", where the "ChildPoverty" variable has different signs compares to the other two. It means that the principal components have the direction same as the eigenvector. And for sub-county level data, the features with the largest absolute values of the first principal component are "IncomePerCap", "Poverty", and "Professional". From the result of the analysis of components, we can observe the most significant factor for the presidential election.

5.3 Proportion of Variance Explained

In order to determine the value of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. We plot the proportion of variance explained (PVE) (Figures 12 and 13) and cumulative PVE (Figures 14 and 15) for both county and sub-county analyses.



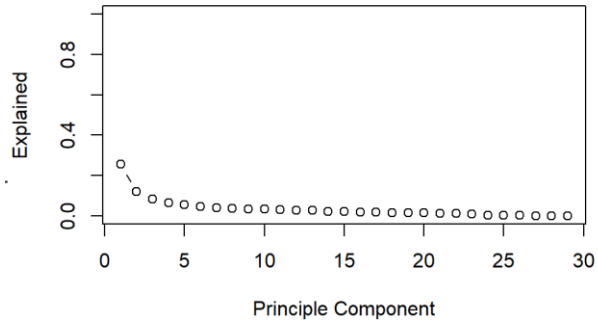


Figure 12: PVE plot for County Level

Figure 13: PVE plot for Sub-County Level

From the cumulative value PVE for both county and sub-county analyses. We find out that 14 columns of PC are needed for county level data to pass the 90% threshold and 16 columns of PC are needed to reach 90% for sub-county level data.

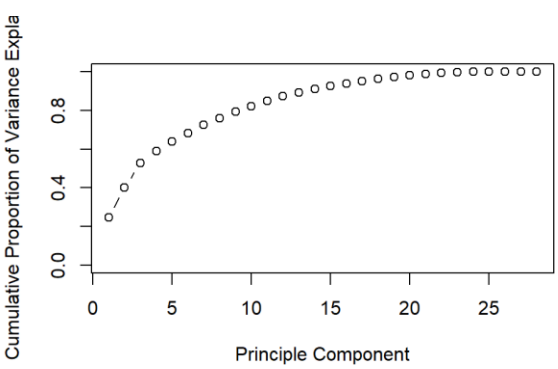


Figure 14: Cumulative PVE plot for County Level

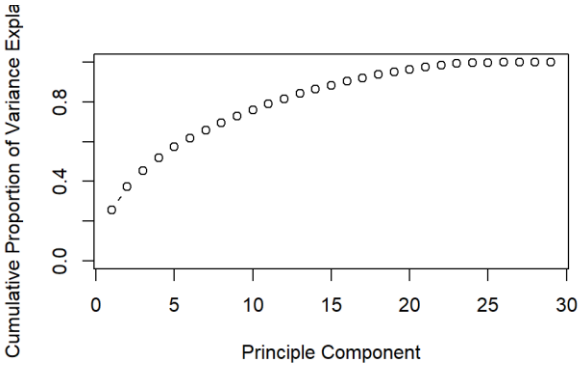


Figure 15: Cumulative PVE plot for Sub-County Level

5.4 Clustering

Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set. When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
With data processed by principal components analysis, we perform hierarchical clustering with complete linkage. We cut the tree to partition the observations into 10 clusters and then re-run the hierarchical clustering algorithm using the first several principal components of the county level after PCA processing as inputs instead of the original features.

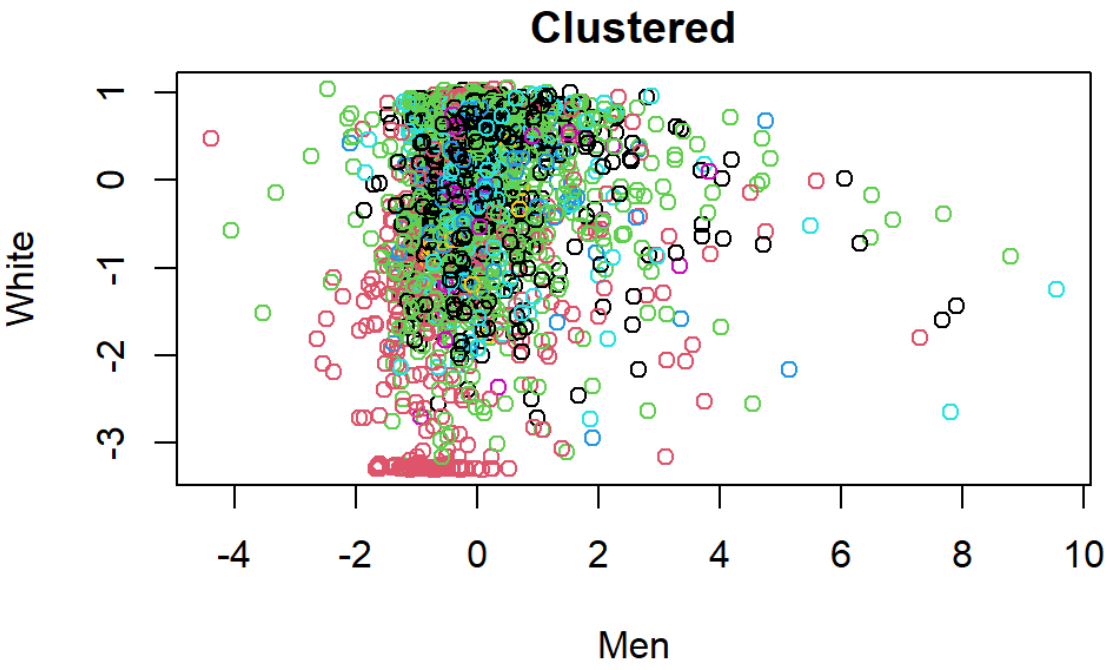


Figure 16: Clustered Data 1

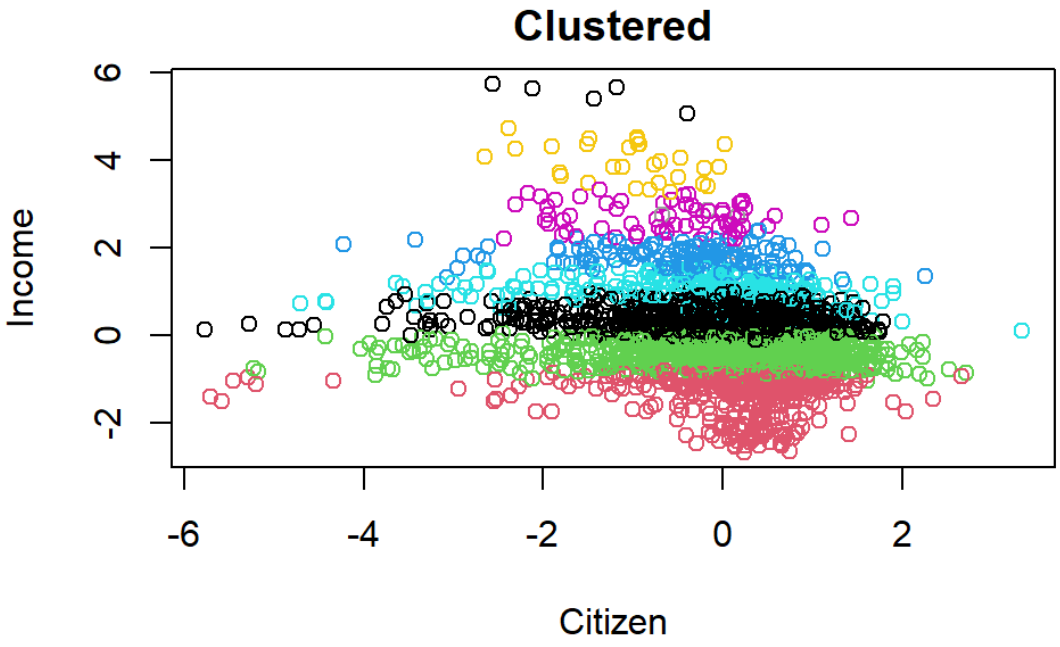


Figure 17: Clustered Data 2

For different choices of principal components, we get different results. The most significant group for the clustering is shown in Figure 16, where we see that from this unsupervised model,

the machine believes that “white” and “men” are the two most significant factors. The voters with a high eigenvalue from “white” and “men” tend to vote for the same candidate and the voters that are “not white” and “not men” tend to behave similarly.

The other two examples are in Figures 17 and 18, where we de-selected “white” and “men”, the clustering method then proposes “Income” and “Citizen” as a group. However, we can see that this time the cluster of data went into segments according to the “Income” level. In the meanwhile, the value of how long the voter has become a “Citizen” has a less significant influence on the result. This doesn’t necessarily mean the “Citizen” feature is absolutely not a strong indicator but shows its relative weakness when compared with the “Income” feature. A similar situation happens again in Figure 18, where we deselected “Citizen” and for the pair of “Poverty” and “Income”, “Income” is still relatively too strong and makes the clustered cloud of data distributed in obvious segments.

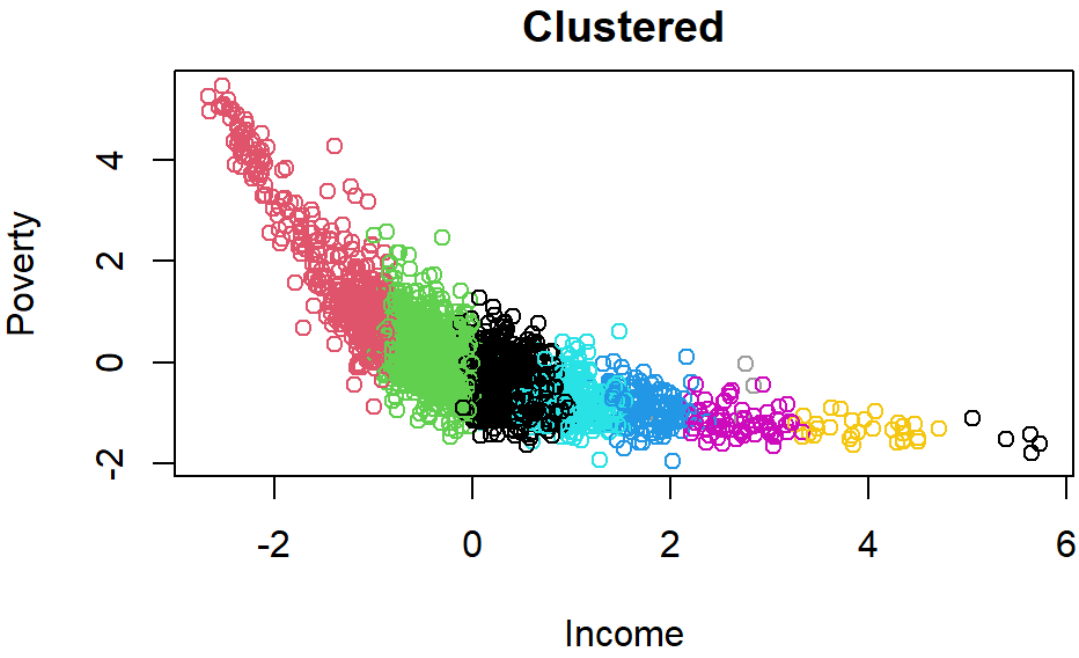


Figure 18: Clustered Data 3

By using the original census data, we find out that the first 2 principal components gain the results for more clusters. The majority of results fall in the 2nd cluster. When using the first 2 components, the majority of results are spread between 1 to 6 clusters. As an instance of further investigation, we compare the Euclidean distance of San Mateo County from the centroid of its cluster. San Mateo county ends up in cluster one when using the census data. Euclidean distance of San Mateo County is far less when we perform clustering with original data compared to using the data after PCA. This shows that PCA might not be working well for the clustering method for San Mateo County. This is most likely because we are using not enough information to build this unsupervised model. When we use the first 2 principal components, San Mateo ends up in cluster 7. The counties that end up in cluster 7 must have

similar levels of the first 2 principal components when it comes to sub-county level census data.

5.5 Decision Tree

Tree-based methods are one of the most simple and useful method for interpretation. A tree is built by splitting the source set, constituting the root node of the tree, into subsets—which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. (S., Shai; B., Shai, 2014).

We first plot a decision tree shown in Figure 19 without pruning, where we have a quite long and complicated tree. There are 14 nodes in total for this "tree before pruned". And this tree led to an overfitting result with a high-error rate. To avoid overfitting, we tested and modified several models and pruned our tree to be 6 nods and get a simplified tree shown in Figure 20.

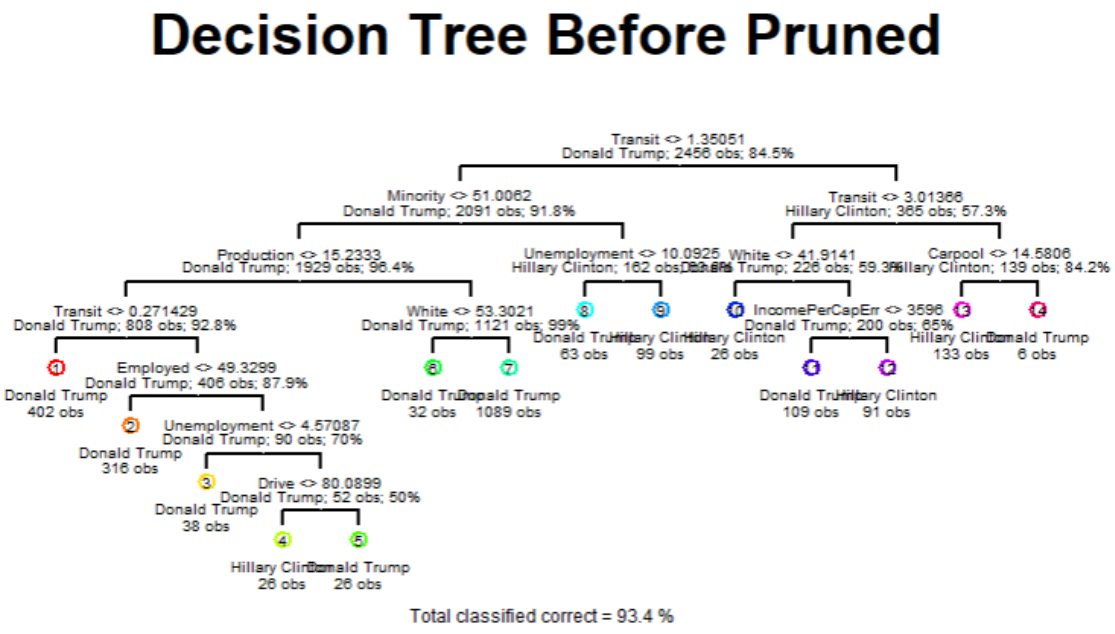


Figure 19: Decision Tree Before Pruned

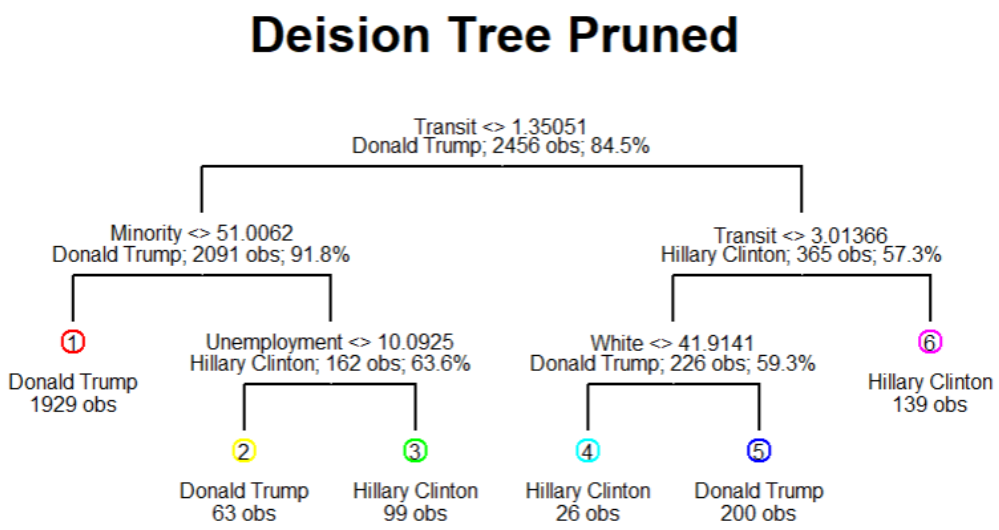


Figure 20: Decision Tree Pruned

For the tree after being pruned, we get a better result that the residual mean deviance becomes 0.4687 (1148 divided by 2450) and the misclassification error rate become 0.0794 (195 / 2456). "Transit" is the first split for our pruned tree, this is because urban areas with heavier transit are more likely to vote for Hillary Clinton and the areas with a smaller volume of transit tend to Donald Trump. And this becomes the deciding factor again in the second row, this shows its significance. After this, "minority" is to be the second important sign where the minority are more likely to be Hillary Clinton's supporters and vice versa. we can see that this as Trump supporters tend to be white while minorities are more likely to vote for Clinton. And besides that, the unemployment rate also affects the voters' decision but is not as significant as the former factors.

5.6 Logistic Regression

In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling (H. David, L. Stanley, 2000). The equation of the logistic regression model is:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Where $$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

From the result shown in Figure 21, the larger coefficients in the logistic regression model mean a more significant influence on the result. Subtitle changes for the variable with a large coefficient can lead to a considerable change in predicted probability. There are a number of significant variables, for example, Professional, Citizen, IncomePerCap, Employed, production, etc. When we compare significant variables in the decision tree model, they are consistent with each other.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.507e+02  4.997e+02  -0.702 0.482790
TotalPop         3.115e-05  1.006e-04   0.310 0.756921
Men              9.180e-02  5.239e-02   1.752 0.079729 .
White           -1.347e-01  6.831e-02  -1.972 0.048602 *
Citizen          1.296e-01  2.842e-02   4.559 5.13e-06 ***
Income          -7.688e-05  2.758e-05  -2.788 0.005301 **
IncomeErr       -2.120e-05  6.155e-05  -0.344 0.730479
IncomePerCap     2.553e-04  6.610e-05   3.862 0.000112 ***
IncomePerCapErr -3.414e-04  1.618e-04  -2.110 0.034899 *
Poverty          5.760e-02  4.244e-02   1.357 0.174639
ChildPoverty    -2.133e-02  2.544e-02  -0.838 0.401769
Professional     2.450e-01  3.716e-02   6.591 4.35e-11 ***
Service          2.773e-01  4.366e-02   6.351 2.14e-10 ***
Office           6.069e-02  4.240e-02   1.432 0.152264
Production       1.449e-01  3.967e-02   3.652 0.000260 ***
Drive            2.551e+00  3.254e+00   0.784 0.433042
Carpool          2.592e+00  3.254e+00   0.797 0.425622
Transit          2.836e+00  3.252e+00   0.872 0.383153
Walk             2.770e+00  3.253e+00   0.851 0.394558
OtherTransp      2.693e+00  3.254e+00   0.828 0.407772
WorkAtHome       2.590e+00  3.254e+00   0.796 0.426005
MeanCommute      6.512e-02  2.430e-02   2.679 0.007373 **
Employed         2.004e-01  3.243e-02   6.181 6.37e-10 ***
PrivateWork      5.890e-01  3.796e+00   0.155 0.876681
PublicWork       5.006e-01  3.796e+00   0.132 0.895087
SelfEmployed     5.213e-01  3.792e+00   0.137 0.890670
FamilyWork      -1.030e-01  3.818e+00  -0.027 0.978478
Unemployment     1.959e-01  3.862e-02   5.071 3.96e-07 ***
Minority        -8.483e-04  6.603e-02  -0.013 0.989750
```

Figure 21: Coefficients from the Logistic Regression Model

5.7 Lasso method

Lasso is the abbreviation of the “least absolute shrinkage and selection operator”. It is a penalized logistic model that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

The mathematical expression of Lasso method is as below. Note that $\lambda \geq 0$ is a tuning parameter.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Where $$\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

Lasso is sometimes less fitted and leads to a relatively lower accuracy rate and higher error rate. However, this property is also useful when we need to avoid overfitting. In our model, we perform cross-validation to select the best regularization parameter The result of the Lasso method after log transformation is visualized below in Figure 22 and the training error and test error are 0.07166124 and 0.07154472, respectively.

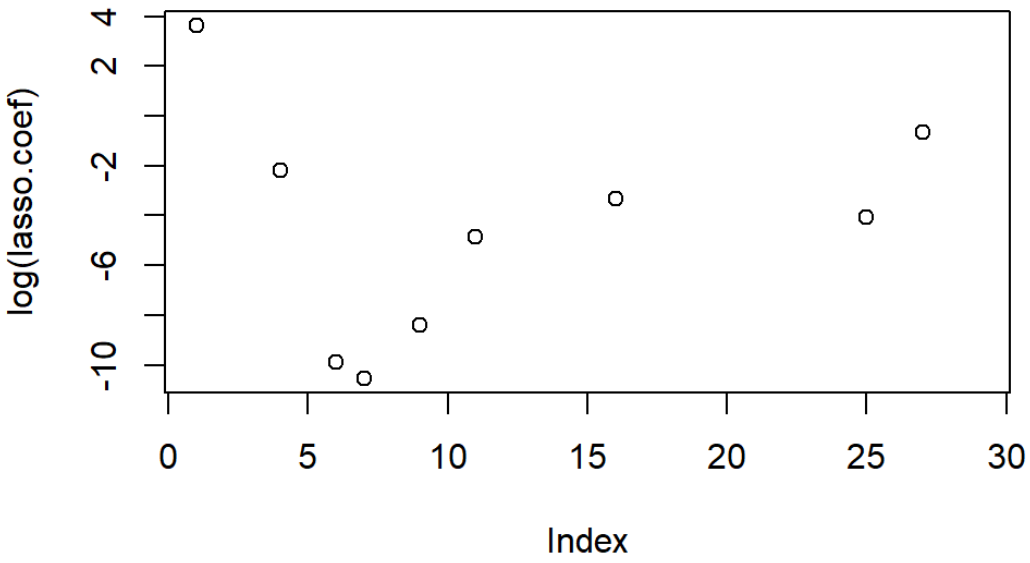


Figure 22: The Result of Lasso Method with Log Transformation

5.8 ROC Curve and Area under Curve

ROC is the abbreviation of Receiver Operating Characteristic. It is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Comparing the ROC curve, we can observe that the logistic model has the largest Area Under Curve (AUC), and the lasso penalized logistic model has almost the same AUC. Also, the logistic method has the lowest test error and the lowest train error which are the pros of it.

	AUC
tree	0.891769796430813
logistic	0.950251098556191
lasso	0.96174533444

Figure 23: The Area Under Curve

	train.error	test.error
tree	0.06596091	0.07317073
logistic	0.06683275	0.06829268
lasso	0.07166124	0.07154472

Figure 24: Comparison of Training Error and Test Error

By Comparing the value of training error and test error (Figure 24), we find that the logistic method has a test error and train error as low as the Lasso method. The lasso method can avoid overfitting as it is a regularization method, but the bad side of it is to be not stable. Another con for the Lasso model is when there are highly correlated features, the Lasso model may randomly select one of them or part of them. However, we have already cleaned our dataset, so this will not be a problem. The Area Under Curve of the decision tree method is slightly smaller but not too much, but it is easier to interpret and fast for inference. The cons can be its tendency to overfitting. Based on all those above, we believe the logistic model from above would be the most appropriate one.

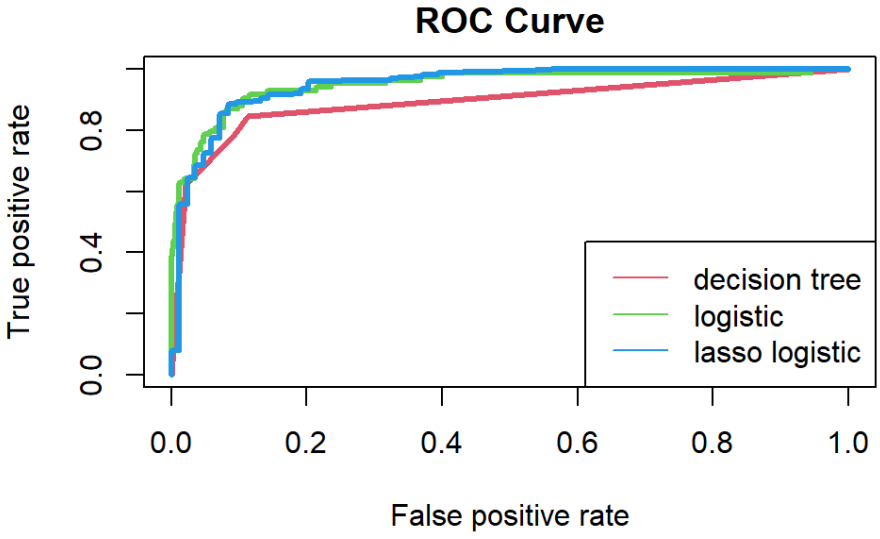


Figure 25: The ROC Curve

6. Further Analysis and Conclusions

Additionally, we explored additional classification and clustering methods: k-nearest neighbors model (KNN), Latent Dirichlet Allocation (LDA), support vector machines (SVM), random forest, and boosting. By investigating the result, output, and performance of different models. We can confirm that the logistic model in the previous sectors would be the most appropriate and effective model for this specific problem in the 2016 presidential election.

The error curve of the K-Nearest Neighbors Model is shown below in Figure 26. We list the result of the KNN and LDA model in Figure 27 and the performance of the SVM model is shown in Figures 28 and 29. It is obvious that the classification models in the previous section have lower training and test errors.

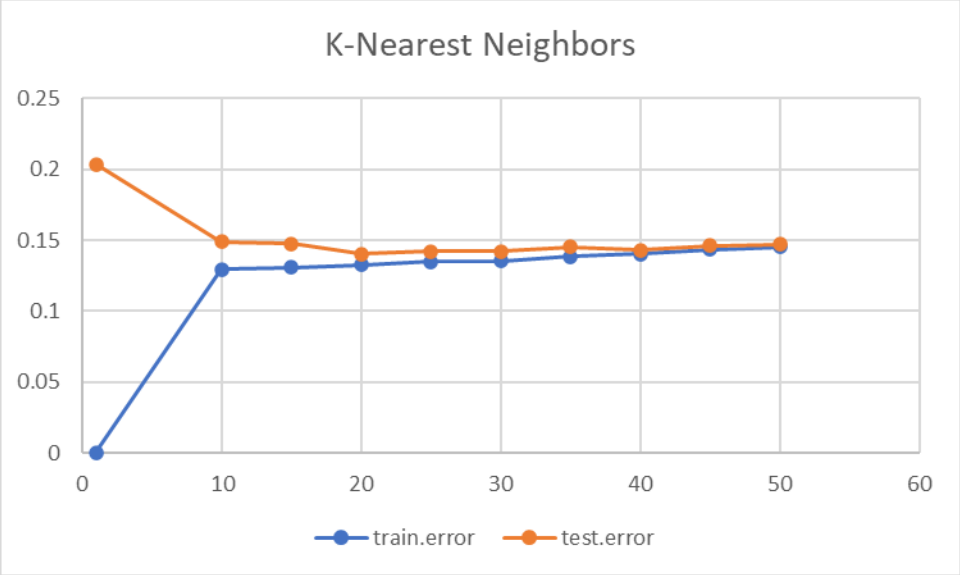


Figure 26: The Error curve of K-Nearest Neighbors Model

	train.error	test.error
KNN	0.18037459	0.19788274
LDA	0.07532573	0.07154472

Figure 27: The Training Error and Test Error for KNN and LDA

```
No. of variables tried at each split: 5

        OOB estimate of  error rate: 6.43%
Confusion matrix:
               Donald Trump Hillary Clinton class.error
Donald Trump           2029              46  0.02216867
Hillary Clinton         112             269  0.29396325

Call:
svm(formula = candidate ~ ., data = trn.cl, kernel = "linear",
    cost = 0.01, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.01

Number of Support Vectors:  589
```

Figure 28: Performance of the SVM model

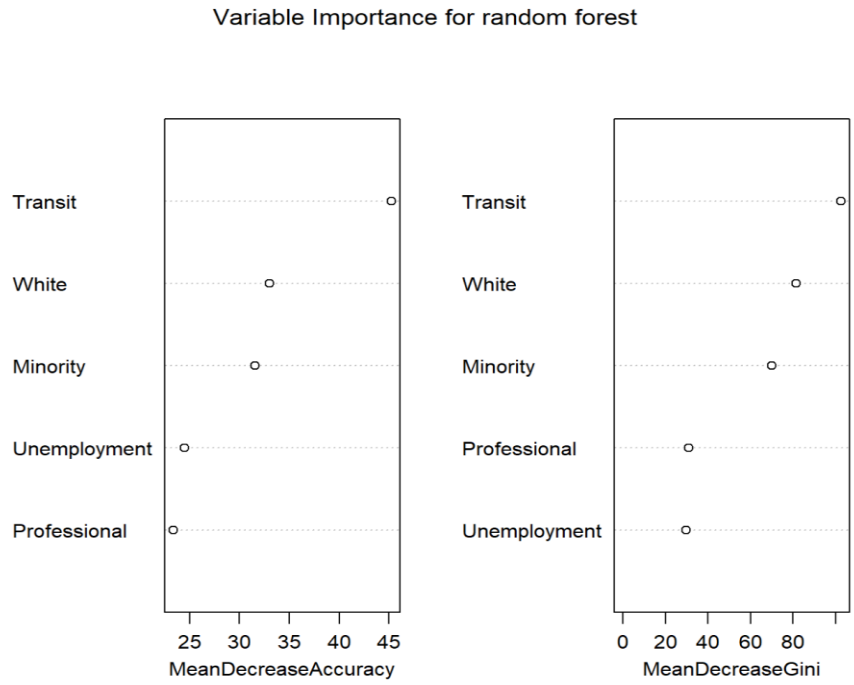


Figure 29: Variance Importance for Random Forest

From the boosting model, we reproved the result from the decision tree model that transit was the most influential predictor. We boost the first five most influential predictors from the most influential were Transit, White, Minority, Unemployment and Professional; compared to random forests where the top five predictors were also Transit, Minority, White, Unemployment and Professional.

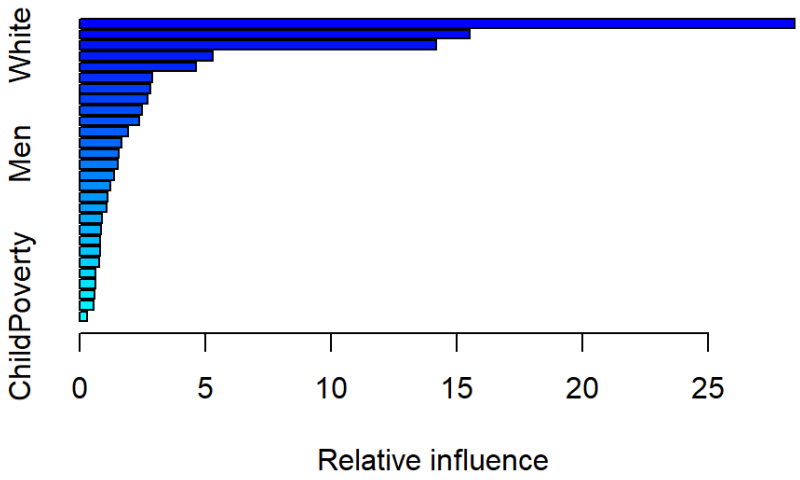


Figure 30: Boosting Model

By investigating the result, output, and performance of different models. We can confirm that the logistic model in the previous sectors would be the most appropriate and effective model for this specific problem in the 2016 presidential election.

References

1. N. Silver, **2016-election-forecast**, https://projects.fivethirtyeight.com/2016-election-forecast, 2016.
2. N. Silver, **Election Forecast: Obama Begins With Tenuous Advantage,** 2012**.**
3. Bishop, Christopher. **Pattern Recognition and Machine Learning**. ISBN 0-387-31073-8. Springer, Berlin, 2006.
4. G. James, D. Witten, T. Hastie, R. Tibshirani, **An Introduction to Statistical Learning**, ISBN 978-1-4614-7137-0, New York, 2013.
5. S. Shai; B. Shai. **Understanding Machine Learning**. Cambridge. 2014
6. H. David, L. Stanley. **Applied Logistic Regression** (2nd ed.). Wiley. ISBN 978-0-471-35632-5, 2000.
7. MATLAB & Simulink, **Detector Performance Analysis Using ROC Curves**. www.mathworks.com. 2016.