Puzzling Election Results?

DNSC 6211: Programming for Analytics

Amit Nayak
Soomin Park
Junfei Zheng
Tianweibao Zheng
Ziqing Zhu

November 28th, 2016

Abstract

Our team collected a variety of data of each state in the United States. The variables we looked at include income levels, GDP, and unemployment rates, among various other variables. We then created one large data set including these variables and tried to run a regression analysis to see if we could develop some sort of formula to predict how the election results would turn out based on our regression equation. We chose to do this to see if such a regression equation could be attained given how surprising the result went in the eyes of many. Additionally, we examined the correlation matrix of variables to eliminate any redundant variables, and used visualization tools such as mapping and shiny to better envision the data that we received for each state. The regression equation we developed is statistically significant in that it does a significantly better job at predicting whether Trump or Hillary wins a state than a model that uses the averages for each variable.

Contents

1 Introduction

The 2016 presidential election will go down as one of the most interesting presidential elections in history. Behind all the chaos that went on with the debates and each candidate's issues and debacles, none of the polls had predicted a Trump victory. Even the remarkable website FiveThirtyEight, which had predicted 101 out of the 102 states (and DC) correctly in the last two presidential elections, failed to predict the outcome of this election cycle correctly. Our group thought it would be of vast interest not just to us, but countless Americans to see if we could determine variable(s) that could help create a regression equation to predict the correct outcomes in each state. Additionally, we want to identify variables that are highly correlated with one another and remove them from our regression model. Finally, we wanted to utilize visualization techniques such as mapping and shiny to help illustrate the data that we received for each individual state for each separate independent variable.

2 Background

As with any presidential election, it is always interesting to look at the breakdown of votes among the various demographic groups. This election cycle may have been even more interesting to investigate the breakdown among the various demographics given how divisive the two candidates can be perceived. Given Donald Trump's stance on certain groups of people, such as Mexicans, Muslims, and women to name a few, one would tend to believe that the breakdown among such demographics would be more pronounced than in past election cycles. As a team, we decided to investigate a multitude of variables, which include but are not limited to race, gender, income level, and education levels. If we are able to find a regression equation supporting our numerous variables, it can help us determine which variables strongly support a certain outcome, and which may play a more secondary role.

3 Method

We first used a web crawler in python to save web page contents into a **json** file. We then used the keyword extractor from the Monkeylearn library in Python, Twitter API and oAuth for data access and use by using the Tweepy library for visualizing main keywords using the Wordclouds library. Next, we utilized the **stepAIC()** function from the MASS package to perform stepwise selection. We then proceeded by trying to remove redundant variables via correlation analysis. We were able to do this using R's correlation table, as well as the matplotlib charts to add linear regression lines. We performed Python's matplotlib and seaborn to show pairwise relationships between variables and then drew the detailed correlations table using R to do detailed correlation analysis. We continued on by using the function **lm** to perform multiple linear regression in R. We then performed python's package sklearn to do cross validation of the models we got and drew comparisons of real and predictive values using matplotlib. We concluded this portion of the regression by using a decision tree to clarify the result. The rpart programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. Rattle provides an interface to R functionality for data mining. The last step was to use a Heatmap from the Plotly Module and Shiny with bubblechart from GoogleCharts to provide us with a visualization of the data.

4 Organization

Our group divided the work pretty evenly based on what each of our strengths were. Amit was responsible mostly for the documentation, presentations, videos, and analysis. Jeff and Tianweibao played an active role in the data preparation and modeling aspects, especially when it pertained to regression analysis, using R, ggplot, and matplotlib. Ziqing was responsible for the data preparation using python for web scraping and the map programming. Finally, Soomin was responsible for the twitter analysis using python twitter API libraries and oAuth, and R Shiny programming, as well as organization of the presentation.

4.1 Workflow

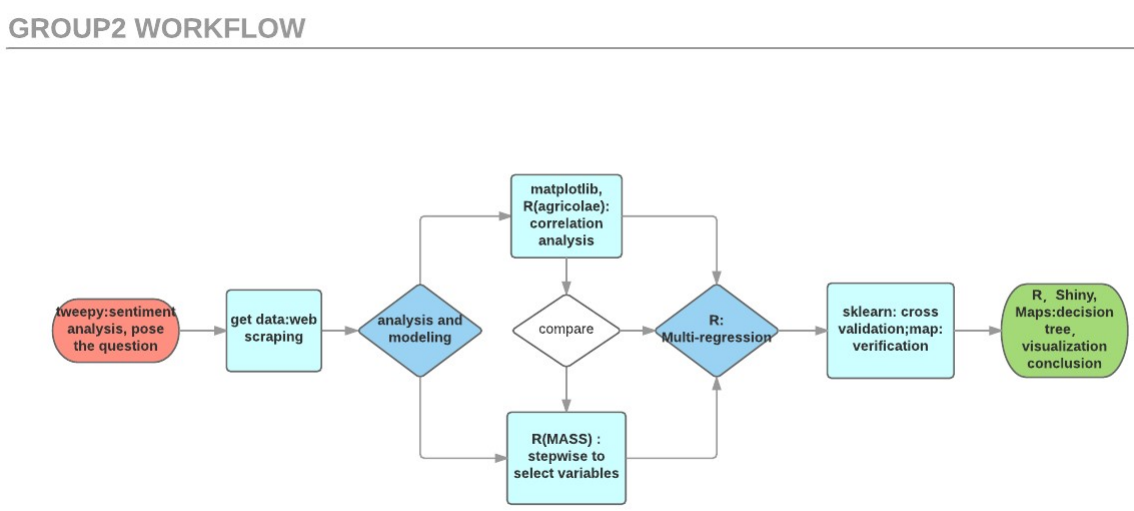

Figure 1: The project workflow

We first used tweepy to get a sentiment analysis from twitter. We then got our data through web scraping and from downloading the data (when web scrapping wasn't possible). We then did analysis and modeling by comparing the correlation analysis from matplotlib and R's agricolae package, with the stepwise regression we got from R's MASS function. This led to the development of the multivariate regression that we obtained through R. We then proceeded by verifying the model that we got by using sklearm, cross validation, and mapping to make sure that we weren't overfitting our model. We concluded our project by using shiny, a decision tree, and mapping to help us visualize our data and the results that we obtained in order to arrive at our final conclusion of all the work that we did.

4.2 Project structure

The main source we used was http://www.bea.gov/regional/downloadzip.cfm. This source enabled us to collect a vast array of data including the state-by-state breakdowns of GDP, income levels, and education levels. We also scraped data from CNN.com and statsamerica.org, while also doing a twitter analysis. These various sources enabled us to get the requisite

data for breaking down the various state populations among an assortment of demographics, including race, gender, income levels, and GDP to name a few. This helped enable us to develop a regression equation to help us solve our ultimate goal of attempting to predict the 2016 presidential winner in each of the 50 states and the District of Columbia.

4.3 Figures and Tables

State	Trump %	Hillary %	Income	GDP	Unemployment Rate	Education Level	White Ratio	M to F Index	LGBT Ratio
Alabama	62.9	34.6	38030	207248	5.4	24.2	65	94.33	2.8
Alaska	52.9	37.7	56147	50592	6.9	29.7	58	108.52	3.4
Arizona	49.5	45.4	39156	296555	5.5	27.7	51	98.74	3.9
Arkansas	60.4	33.8	38252	125759	4	21.8	74	96.45	3.5
California	33.1	61.6	53741	2513807	5.5	32.3	39	98.83	4
Colorado	44.4	47.3	50899	320176	3.6	39.2	69	100.48	3.2
Connecticut	41.2	54.5	68704	262240	5.4	38.3	70	94.83	3.4
Delaware	41.9	53.4	47633	69105	4.3	30.9	63	93.94	3.4
District of Columbia	4.1	92.8	73302	125838	6.1	56.7	37	89.52	10
Florida	49.1	47.8	44429	911025	4.7	28.4	54	95.6	3.5
Georgia	51.3	45.6	40306	508372	5.1	29.9	52	95.38	3.5
Hawaii	30	62.2	48288	81595	3.3	31.4	19	100.32	5.1
Idaho	59.2	27.6	38392	65915	3.8	26	84	100.39	2.7
Illinois	39.4	55.4	50295	784365	5.5	32.9	63	96.24	3.8
Indiana	57.2	37.9	41940	341092	4.5	24.9	80	96.83	3.7
Iowa	51.8	42.2	45902	173444	4.2	26.8	85	98.07	2.8
Kansas	57.2	36.2	47161	148473	4.4	31.7	75	98.45	3.7
Kentucky	62.5	32.7	38588	197649	5	23.3	85	96.85	3.9
Louisiana	58.1	38.4	42947	240028	6.4	23.2	58	95.9	3.2
Maine	45.2	47.9	42799	58125	4.1	30.1	91	95.84	4.8
Maryland	35.3	60.5	55972	371899	4.2	38.8	50	93.63	3.3
Massachusetts	33.5	60.8	62603	485474	3.6	41.5	73	93.66	4.4
Michigan	47.6	47.3	42812	478773	4.6	27.8	76	96.28	3.8
Minnesota	45.4	46.9	50871	336729	4	34.7	82	98.52	2.9
Mississippi	58.3	39.7	34771	108971	6	20.8	56	94.44	2.6
Missouri	57.1	38	42300	298158	5.2	27.8	80	96.01	3.3
Montana	56.5	36	41809	45786	4.3	30.6	90	100.8	2.6
Nebraska	60.3	34	48544	114327	3.2	30.2	79	98.51	2.7
Nevada	45.5	47.9	41889	143995	5.8	23.6	50	102	4.2
New Hampshire	47.3	47.6	55905	74936	2.9	35.7	91	97.35	3.7
New Jersey	41.8	55	59949	580377	5.3	37.6	56	94.84	3.7
New Mexico	40	48.3	37938	90978	6.7	26.5	38	97.66	2.9
New York	37.5	58.8	58670	1474493	5	35	58	93.76	3.8
North Carolina	50.5	46.7	40759	510248	4.7	29.4	62	95	3.3
North Dakota	64.1	27.8	55950	51503	3	29.1	85	102.15	1.7
Ohio	52.1	43.5	43566	618982	4.8	26.8	79	95.39	3.6
Oklahoma	65.3	28.9	45573	174776	5.2	24.6	75	98.03	3.4
Oregon	41.1	51.7	43783	221308	5.5	32.2	74	97.98	4.9
Pennsylvania	48.8	47.6	49745	698517	5.7	29.7	76	95.06	2.7
Rhode Island	39.8	55.4	50018	58295	5.6	32.7	72	93.43	4.5
South Carolina	54.9	40.8	38302	203631	4.9	26.8	64	94.73	2.9
South Dakota	61.5	31.7	47881	47216	2.9	27.5	83	100.14	4.4
Tennessee	61.1	34.9	42094	322944	4.6	25.7	73	95.11	2.6
Texas	52.6	43.4	46947	1568247	4.8	28.4	44	98.41	3.3
Utah	46.3	27.7	39308	150621	3.4	31.8	81	100.93	2.7
Vermont	32.6	61.1	48587	31171	3.3	36.9	94	97.06	4.9
Virginia	45	49.9	52052	492761	4	37	62	96.34	2.9
Washington	37.7	55.1	51898	455943	5.6	34.2	69	99.26	4
West Virginia	68.7	26.5	36758	71992	5.8	19.6	93	97.25	3.1
Wisconsin	47.9	46.9	45914	311191	4.1	28.4	78	98.53	2.8
Wyoming	70.1	22.5	56081	36219	5.3	26.2	85	104.07	2.9

Figure 2: State-by-State Data

Figure 3: 1st Twitter WordCloud

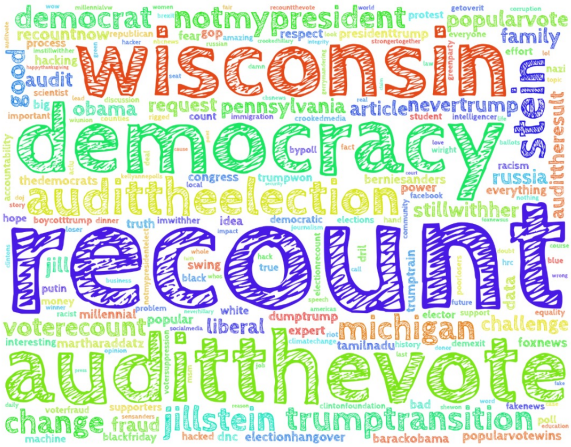

Figure 4: 2nd Twitter WordCloud

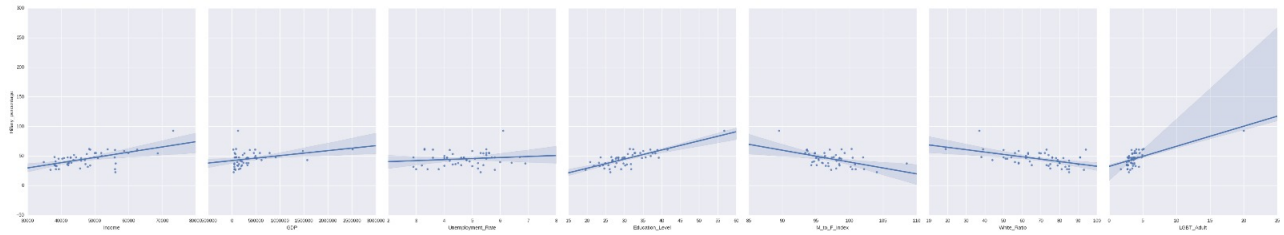

Figure 5: Single Linear Regression for Clinton

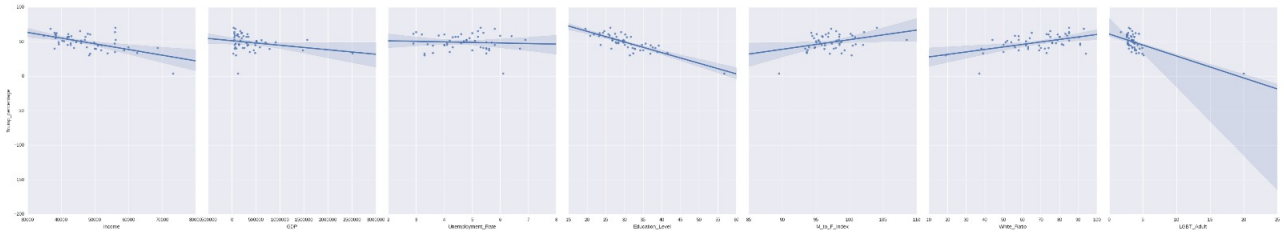

Figure 6: Single Linear Regression for Trump

```
> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Hillary.percentage ~ Income. + GDP. + Unemployment.Rate + Education.Level +
    White.Ratio + M.to.F.Index + LGBT.Ratio

Final Model:
Hillary.percentage ~ GDP. + Education.Level + White.Ratio + M.to.F.Index +
    LGBT.Ratio


                Step Df Deviance Resid. Df Resid. Dev      AIC
1                                       43   1144.508 174.6561
2          - Income.  1 2.127739        44   1146.636 172.7509
3 - Unemployment.Rate  1 5.410736        45   1152.047 170.9910
```

Figure 7: Stepwise Regression for Clinton

```
> step1$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Trump.percentage ~ Income. + GDP. + Unemployment.Rate + Education.Level +
    White.Ratio + M.to.F.Index + LGBT.Ratio

Final Model:
Trump.percentage ~ Income. + GDP. + Education.Level + White.Ratio +
    LGBT.Ratio


                Step Df Deviance Resid. Df Resid. Dev      AIC
1                                       43   1174.175 175.9613
2      - M.to.F.Index  1 1.688724        44   1175.864 174.0346
3 - Unemployment.Rate  1 3.703405        45   1179.567 172.1949
```

Conclusion:

Figure 8: Stepwise Regression for Trump

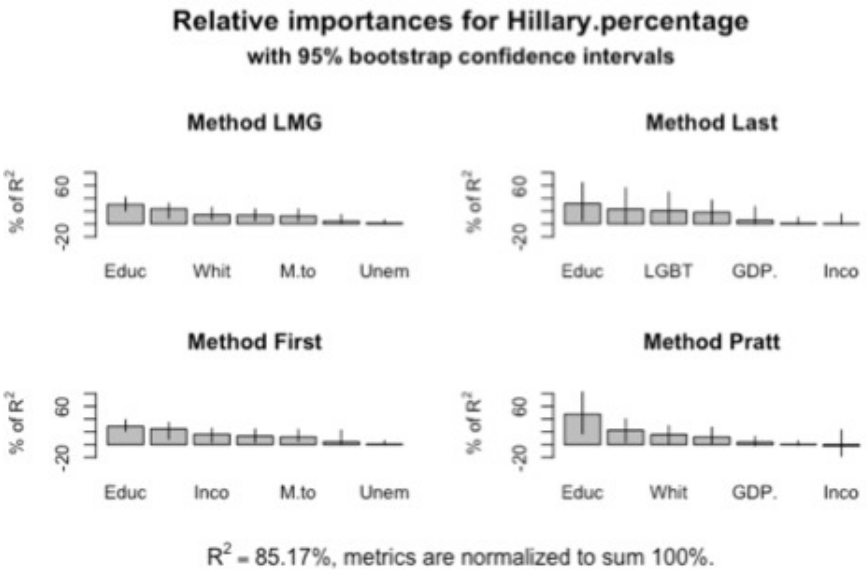

Figure 9: Variable Impact on Model for Clinton

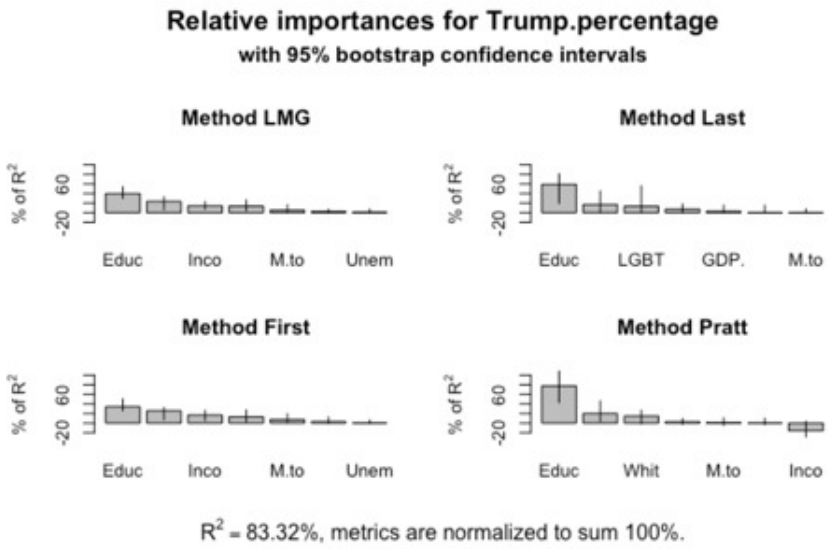

Figure 10: Variable Impact on Model for Trump

```
> cor(election[,c(4:10)],method="pearson")
                    Income.         GDP. Unemployment.Rate Education.Level White.Ratio M.to.F.Index   LGBT.Ratio
Income.          1.00000000  0.185697574       -0.04245328       0.8124473  -0.1620168   -0.1110420  0.434192871
GDP.             0.18569757  1.000000000        0.15708893       0.1192266  -0.3910250   -0.1532202 -0.001754075
Unemployment.Rate -0.04245328  0.157088931        1.00000000      -0.1477452  -0.4043772   -0.1061634  0.146856231
Education.Level    0.81244726  0.119226646       -0.14774518       1.0000000  -0.2193178   -0.3029699  0.610995020
White.Ratio       -0.16201677 -0.391025004       -0.40437719      -0.2193178   1.0000000    0.1567227 -0.316154984
M.to.F.Index      -0.11104199 -0.153220228       -0.10616344      -0.3029699   0.1567227    1.0000000 -0.344215426
LGBT.Ratio         0.43419287 -0.001754075        0.14685623       0.6109950  -0.3161550   -0.3442154  1.000000000
```

Figure 11: Correlation matrix of all our independent variables

```
Call:
lm(formula = Hillary.percentage ~ Education.Level + White.Ratio +
    M.to.F.Index + LGBT.Ratio, data = election)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3504  -2.8903   0.1377   3.1665  11.9085

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     107.83264   25.85945   4.170 0.000133 ***
Education.Level   0.98092    0.14857   6.603 3.60e-08 ***
White.Ratio      -0.23527    0.04672  -5.036 7.79e-06 ***
M.to.F.Index     -0.88197    0.24980  -3.531 0.000954 ***
LGBT.Ratio        2.62890    0.83944   3.132 0.003019 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.165 on 46 degrees of freedom
Multiple R-squared:  0.841,     Adjusted R-squared:  0.8272
F-statistic: 60.84 on 4 and 46 DF,  p-value: < 2.2e-16
```

Figure 12: Final Regression Model for Hillary Clinton

```
Call:
lm(formula = Trump.percentage ~ Education.Level + White.Ratio +
    LGBT.Ratio, data = election)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9049  -4.7800   0.3942   3.7334  11.2421

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     78.09813    5.60551  13.932  < 2e-16 ***
Education.Level -1.13518    0.15218  -7.460 1.65e-09 ***
White.Ratio      0.21058    0.04816   4.372 6.76e-05 ***
LGBT.Ratio      -2.55086    0.84989  -3.001  0.00429 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.33 on 47 degrees of freedom
Multiple R-squared:  0.8103,    Adjusted R-squared:  0.7982
F-statistic: 66.93 on 3 and 47 DF,  p-value: < 2.2e-16
```

Figure 13: Final Regression Model for Donald Trump

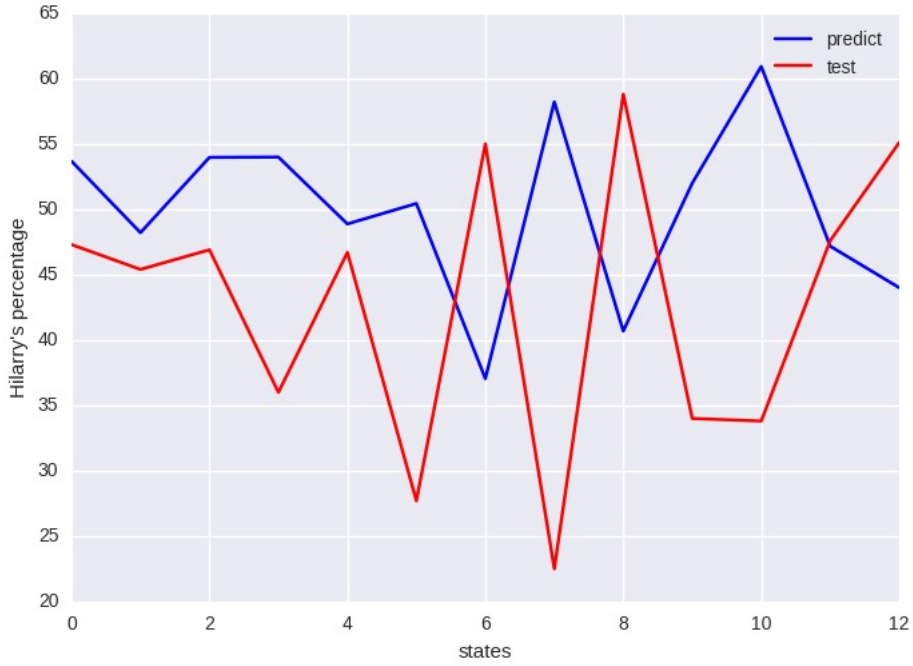

Figure 14: Comparison of Clinton Model with the Actual Outcomes

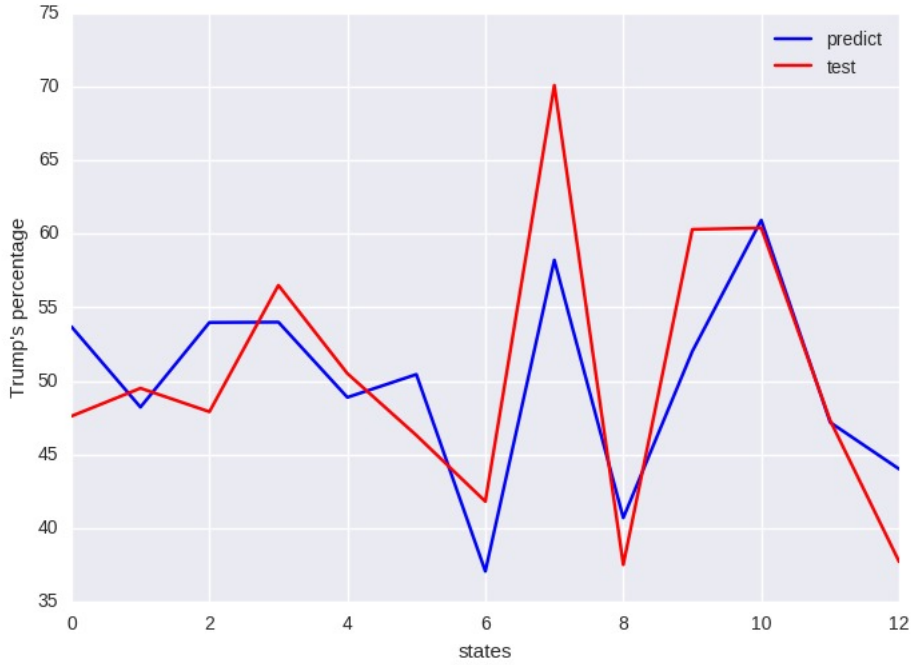

Figure 15: Comparison of Trump Model with the Actual Outcomes

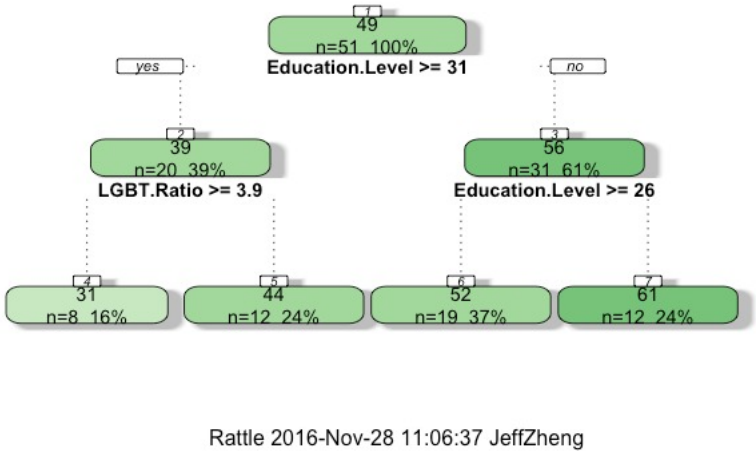

Figure 16: Stepwise Model for Hillary Clinton

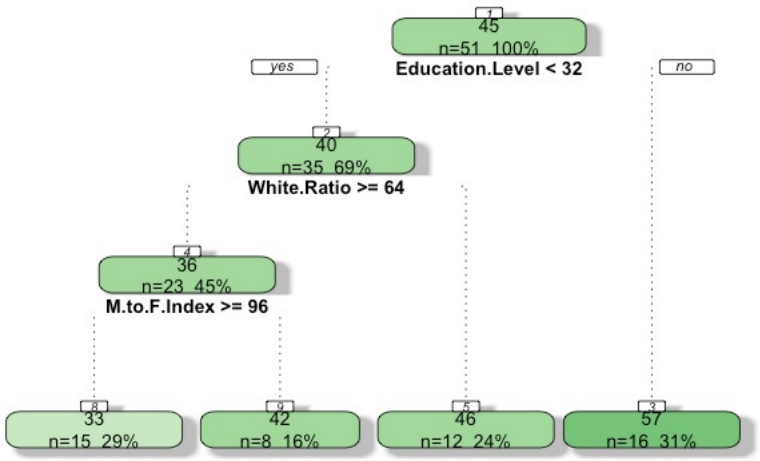

Figure 17: Stepwise Model for Donald Trump

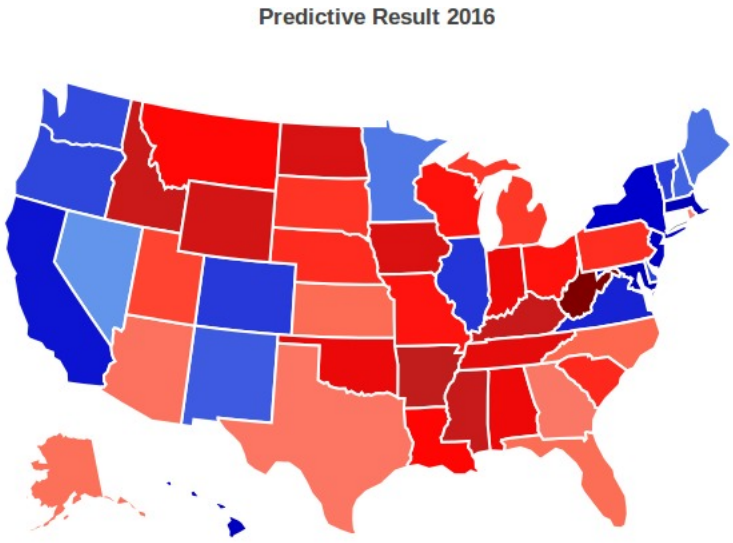

Figure 18: Map our Model Predicts

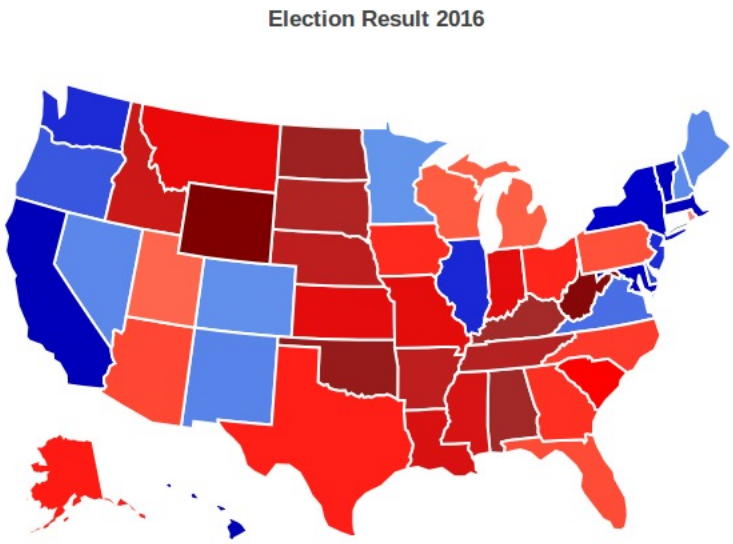

Figure 19: Actual Election Results Map

We started by doing a keyword twitter analysis. More common words are larger in our wordcloud here. There are numerous noteworthy words here, such as protest, audittheelection, and recount. We then scraped all our data into one easy to read data table. Next, we performed simple linear regressions of each variable to find the correlation between each independent variable and the support rate of each candidate. We then developed a stepwise regression model to eliminate variables. For both models we were able to go from 7 variables initially to 5 in our final models here. We then performed four different tests on each model to calculate the contribution that each variable had on our variable. We then got a correlation matrix to help remove even more variables. Variables with a correlation value closer to 1 are highly correlated, so we can remove one of those variables, since they become redundant in our model. We then got a decision tree to break down our data at the cleanest breaks in support rates among all our variables. We then got a map to compare our predictive model with the actual model and you can see that we performed pretty well, getting all 50 states and DC predicted correctly. We finally got our multivariate linear regression. In both models we got our final model down from 7 variables to 3 or 4. Additionally, both models have a high R-squared value of over 80 percent and an extremely statistically significant p-value. We then verified our model to account for the possibility of overfitting. We can see here that our Trump model performed slightly better, but that both models performed pretty well.

5 Discussion

The two most significant takeaways from our project were the multiple regression analysis and the visualization that we were able to obtain. The multiple regression analysis enabled us to figure out which variable(s) played the most significant role in our regression model. We were able to determine that two most significant factors in our model were the education level and the ratio of white people. Higher ratios of white people and lower education levels were associated with a higher support rate for Trump and vice versa for Clinton. The visualization methods we utilized, mapping and shiny, also helped created a very easy-to-read interpretation of the election results along each of the numerous variables that we utilized. For the shiny, we used the Google charts library and were able to provide a scatter plot chart for each of the independent variables with their relationship with the dependent variable (support rate for each candidate). Additionally, we were able to designate each state into one of the six basic regions in the US and adjust the size of each dot based on how many electoral votes that state was worth.

5.1 Learnings

There were a few things that we learned in addition to our main takeaways from this project. One precious lesson we learned from the project is how to build the model. Selecting a certain number of variables from thousands of possibilities is not easy. First, we had to take a glance over all the variables and then guess the possible driving factors that will be added into our initial formula. Then, identifying and creating the model required strict logic to execute. Finally, we needed to evaluate the final model to make sure that the result we obtained was a reliable one. We also learned to use plotly when created our map. We needed this to combine the two kinds of maps into our final map. We then divided the maps into two submaps and represented each with a different color.

5.2 Challenges

We encountered a wide variety of challenges throughout the duration of this assignment. One basic issue was choosing only a certain amount of variables to examine. As is commonly known, there are countless different variables that can be examined when it comes to presidential elections. After choosing the variables we wanted to investigate, we then had to web scrape the data, where we came across a few more obstacles. Some sources didn't allow us to scrape their data, while a few others gave us unusable HTML code, instead of data. Utilizing multiple variables also gave us additional challenges when running our shiny. We were only able to investigate each independent variable's relationship independently with their correlation on the dependent variable. We were unable to find a way of running a multivariate shiny program. Finally, when running our twitter analysis, we came across a variety of time-sensitive issues when trying to gather the twitter analysis.

6 Conclusion

We were able to obtain a regression model of all the relevant independent variables that we examined. Additionally, we were able to see which variables had the most significant impacts on the election and what they meant in terms of the support rate for each of the candidates. The two most significant variables that we found were education level and the ratio of white people in each state. We were able to determine that higher education levels and a lower ratio of white people was correlated with a higher support rate for Clinton and a lower support rate for Trump and the exact opposite also held true. We were further able to visualize these trends using the visualization tools of mapping and shiny. Here we were able to better examine these trends of each independent variable and see their correlation with the support rate of each candidate.

7 References

Bea.gov
CNN.com
statsamerica.org
Twitter.com