### Association Rules

Association Rules Mining was performed using R on the republican primary results and county demographic data. The purpose of mining the data for association rules was to identify key demographic factors within a county for each of the top 3 republican candidates (Donald Trump, Ted Cruz and John Kasich) that results in a greater likelihood of a candidate either winning or losing the republican primary in that country. In order to do association rules mining, we chose 5 key demographic factors: percentage below the poverty line, percentage of white people (not including Hispanics or Latinos), percentage with Bachelor’s degree or higher, per capita income, and population density. These 5 demographic factors were identified on the basis of our analysis of the data in general and the results of multi-linear regression models wherein these factors were significant and influential in predicting the voting share for each candidate within the county. Before performing the data mining, we converted the data for the 5 variables into categorical factors by bucketing the data into 3 levels. Please see Exhibit 3.1 for more information on how the data for the 5 demographic factors was bucketed. Further, we created 3 separate data files for each of the 3 republican candidate that included the following:

1. The 5 demographic factors discussed above
2. An “Outcome” variable indicating if the candidate “Lost” or “Won” within a county
3. A single observation for each of the 1881 counties for which primary elections results were available.

Next, we generated association rules in R for each candidate and isolated the rules where the consequent was the “Outcome” variable. Please see Exhibits 3.2 and 3.3 to see the details of the rules generated for Donald Trump and John Kasich.

In conclusion, we found that the following demographic variables were influential in predicting the outcome for each of the 3 republican candidates:

1. For Donald Trump:
   1. The likelihood of winning increases moderately in counties with low education levels, low per capita income, low population density and predominantly white population.
   2. The likelihood of losing increases significantly in counties with medium education levels, lower poverty rate, high per capita income, low population density and medium white-only population.
2. For Ted Cruz:
   1. The likelihood of winning increases significantly in counties with medium education levels, low poverty rate, low population density and medium white population.
   2. The likelihood of losing increases moderately in counties with low education levels, low per capita income, high poverty rate, low population density and predominantly white population.
3. For John Kasich, no rules with any significant rules were generated even at a support and confidence of 0.5

We applied several other Classification Methods to the republican primary results to predict whether a certain candidate would either win or lose a county of interest (a categorical outcome) based on a set of demographic characteristics (prediction variables) of each county. The methods that we tried include logistic regression, support vector machines, random forest, and conditional inference trees. In order to perform this analysis, we created 3 separate data files for each of the 3 republican candidate (Donald Trump, Ted Cruz and John Kasich). Each file included the following:

1. 51 demographic factors that describe each of the 1881 counties for which primary election results were available.
2. An “Outcome” variable indicating if a particular candidate “Lost” or “Won” within that county
3. A single observation for each of the 1881 counties for which primary elections results were available.

# Bucket information:

Variables considered:

|  |  |  |
| --- | --- | --- |
| HigherEducation | | |
| Range | Value | Count |
| <=15 | 1 | 843 |
| > 15 and < 25 | 2 | 734 |
| >=25 | 3 | 304 |

|  |  |  |
| --- | --- | --- |
| PerCapitaIncome | | |
| Range | Value | Count |
| <=20000 | 1 | 681 |
| > 20000 and < 25000 | 2 | 762 |
| >= 25000 | 3 | 438 |

|  |  |  |
| --- | --- | --- |
| PopulationDensity | | |
| Range | Value | Count |
| <=60 | 1 | 1071 |
| > 60 and < 500 | 2 | 703 |
| >= 500 | 3 | 107 |

|  |  |  |
| --- | --- | --- |
| BelowPovertyLine | | |
| Range | Value | Count |
| <=15 | 1 | 605 |
| > 15 and < 23 | 2 | 878 |
| >= 23 | 3 | 398 |

|  |  |  |
| --- | --- | --- |
| WhiteOnly | | |
| Range | Value | Count |
| <=60 | 1 | 417 |
| > 60 and < 90 | 2 | 893 |
| >= 90 | 3 | 571 |

# John Kasich

Won % = 59/1881 = 3.1%

Loss % = 96.9%

Makes sense since he only won in the state of Ohio because he is the governor of Ohio and very popular. Our data did not have any evidence or support for this factor.

### Won

Support: .05, confidence: 0.1

No rules produced predicting an outcome of “Won”. Dataset has more losses than wins.

### Loss

For losses, the top 10 rules were as follows at supp=0.1, conf=0.3

|  |  |  |  |
| --- | --- | --- | --- |
| **rules** | **support** | **confidence** | **lift** |
| {BelowPovertyLine=3} => {Outcome=Lost} | 0.21159 | 1 | 1.032382 |
| {WhiteOnly=1} => {Outcome=Lost} | 0.221691 | 1 | 1.032382 |
| {BelowPovertyLine=3,WhiteOnly=1} => {Outcome=Lost} | 0.110048 | 1 | 1.032382 |
| {PerCapitaIncome=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.186071 | 1 | 1.032382 |
| {HigherEducation=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.154173 | 1 | 1.032382 |
| {PopulationDensity=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.152578 | 1 | 1.032382 |
| {PerCapitaIncome=1,WhiteOnly=1} => {Outcome=Lost} | 0.13025 | 1 | 1.032382 |
| {HigherEducation=1,WhiteOnly=1} => {Outcome=Lost} | 0.120149 | 1 | 1.032382 |
| {PopulationDensity=1,WhiteOnly=1} => {Outcome=Lost} | 0.141414 | 1 | 1.032382 |
| {PerCapitaIncome=1,WhiteOnly=2} => {Outcome=Lost} | 0.13344 | 1 | 1.032382 |

# Ted Cruz

Won% = 579/1881 = 30.78%

Loss% = 1302/1881 = 69.22%

supp=0.1, conf=0.3

### Won

|  |  |  |  |
| --- | --- | --- | --- |
| Rules | Support | Confidence | Lift |
| {HigherEducation=2,PopulationDensity=1} => {Outcome=Won} | 0.10579479 | 0.537837838 | 1.747276292 |
| {PopulationDensity=1,WhiteOnly=2} => {Outcome=Won} | 0.106858054 | 0.45890411 | 1.490843921 |
| {BelowPovertyLine=1} => {Outcome=Won} | 0.127060074 | 0.395041322 | 1.283372586 |
| {HigherEducation=2} => {Outcome=Won} | 0.153110048 | 0.392370572 | 1.274696108 |
| {PopulationDensity=1} => {Outcome=Won} | 0.208931419 | 0.366946779 | 1.192101711 |

### Loss

|  |  |  |  |
| --- | --- | --- | --- |
| Rules | Support | Confidence | Lift |
| {HigherEducation=1,WhiteOnly=3} => {Outcome=Lost} | 0.145136 | 0.916107 | 1.323501 |
| {HigherEducation=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.136098 | 0.882759 | 1.275322 |
| {HigherEducation=1,PerCapitaIncome=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.129718 | 0.880866 | 1.272588 |
| {BelowPovertyLine=2,WhiteOnly=3} => {Outcome=Lost} | 0.117491 | 0.873518 | 1.261972 |
| {HigherEducation=1,PopulationDensity=1,BelowPovertyLine=3} => {Outcome=Lost} | 0.110048 | 0.866109 | 1.251268 |

# Donald Trump

Won in 1215 counties out of 1881 = 64.6%

Loss = 35.4%

### Won

supp=0.2, conf=0.3

|  |  |  |  |
| --- | --- | --- | --- |
| rules | support | confidence | lift |
| {HigherEducation=1,PerCapitaIncome=1} => {Outcome=Won} | 0.240297714 | 0.821818182 | 1.272296296 |
| {PerCapitaIncome=1} => {Outcome=Won} | 0.281233386 | 0.776798825 | 1.202599663 |
| {PerCapitaIncome=1,PopulationDensity=1} => {Outcome=Won} | 0.210526316 | 0.76300578 | 1.181245986 |
| {HigherEducation=1} => {Outcome=Won} | 0.341307815 | 0.761565836 | 1.179016739 |
| {WhiteOnly=3} => {Outcome=Won} | 0.226475279 | 0.746059545 | 1.155010702 |

### Lost

supp=0.1, conf=0.3

|  |  |  |  |
| --- | --- | --- | --- |
| rules | support | confidence | lift |
| {HigherEducation=2,PopulationDensity=1} => {Outcome=Lost} | 0.106326 | 0.540541 | 1.526662 |
| {BelowPovertyLine=1} => {Outcome=Lost} | 0.155237 | 0.482645 | 1.363145 |
| {PerCapitaIncome=3} => {Outcome=Lost} | 0.112174 | 0.481735 | 1.360576 |
| {PerCapitaIncome=2,PopulationDensity=1} => {Outcome=Lost} | 0.100478 | 0.468983 | 1.324559 |
| {PopulationDensity=1,WhiteOnly=2} => {Outcome=Lost} | 0.10739 | 0.461187 | 1.302542 |

# Random Forest output:

### Donald Trump

Predicted

Actual Donald Trump Not Donald Trump

Donald Trump 213 23

Not Donald Trump 52 89

Accuracy = 0.8

### Ted Cruz

Predicted

Actual Not Ted Cruz Ted Cruz

Not Ted Cruz 244 11

Ted Cruz 52 70

Accuracy = 0.83

# Logistic Regression

### Donald Trump

Predicted

Actual Donald' Trump Not Donald Trump

Donald Trump 205 31

Not Donald Trump 58 83

Sensitivity = 0.59

Specificity = 0.87

Positive Predictive Value = 0.59

Negative Predictive Value = 0.87

Accuracy = 0.76

### Ted Cruz

Predicted

Actual Ted Cruz Not Ted Cruz

Not Ted Cruz 218 37

Ted Cruz 45 77

Accuracy = 0.78

# Support vector Machines

### Donald Trump

Predicted

Actual Donald Trump Not Donald Trump

Donald Trump 218 18

Not Donald Trump 55 86

Accuracy = 0.81

### Ted Cruz

Predicted

Actual Not Ted Cruz Ted Cruz

Not Ted Cruz 243 12

Ted Cruz 60 62

Accuracy = 0.81