## Classification Techniques

### Logistic Regression

Logistic regression was performed in R. The details of the final best fit models for each of the 3 candidates can be found in Exhibit 4.1. Overall, the performance of the models at a cutoff value of 0.5 was pretty good except the model for Kasich with the following highlights for each candidate:

1. The model for Donald Trump had an accuracy of 76%, sensitivity of 59% and specificity 87%.
2. The model for Ted Cruz had an accuracy of 78%, sensitivity of 63% and specificity 85%.
3. The model for John Kasich had an accuracy of 96% but poor values for sensitivity at 100% and specificity at 0%. Hence, the model for Kasich was not useful at all.

The confusion matrices for the 3 models can be seen in Exhibit 4.2.

### Random Forest

Random Forest analysis was performed in R. The algorithm for a random forest involves sampling cases and variables to create a large number of decision trees. Each case is classified by each decision tree. The most common classification for that case is then used as the outcome. Random forests are grown using the randomForest () function in the random Forest package in R. The default number of trees is 500, the default number of variables sampled at each node is sqrt(M), and the minimum node size is 1.

Random forests tend to be very accurate compared with other classification methods. Additionally, they can handle large problems (many observations and variables), can handle large amounts of missing data in the training set, and can handle cases in which the number of variables is much greater than the number of observations.

Overall, the performance of the random forest analysis was pretty good for all candidates except for Kasich with the following highlights for each candidate:

1. The output for Donald Trump had an accuracy of 80%, sensitivity of 63% and specificity 90%.
2. The output for Ted Cruz had an accuracy of 83%, sensitivity of 57% and specificity 96%.
3. The output for John Kasich had an accuracy of 97% but poor values for sensitivity at 100% and specificity at 7%. Hence, the model for Kasich was less useful compared to other candidates.

The confusion matrices for the 3 candidates can be seen in Exhibits 5.1, 5.2 and 5.3.

### Support Vector Machines

Support vector machines(SVMs) are a group of supervised machine-learning models that can be used for classification and regression. They’re popular at present, in part because of their success in developing accurate prediction models, and in part because of the elegant mathematics that underlie the approach. Because predictor variables with larger variances typically have a greater influence on the development of SVMs, the svm () function scales each variable to a mean of 0 and standard deviation of 1 before fitting the mode by default. Unlike the random forest approach, the SVM is unable to accommodate missing predictor values when classifying new cases. We performed SVM in R.

Overall, the performance of the Support vector machines was pretty good for all candidates except for Kasich with the following highlights for each candidate:

1. The output for Donald Trump had an accuracy of 81%, sensitivity of 61% and specificity 92%.
2. The output for Ted Cruz had an accuracy of 81%, sensitivity of 51% and specificity 95%.
3. The output for John Kasich had an accuracy of 96% but poor values for sensitivity at 100% and specificity at 0%.

The confusion matrices for the 3 candidates can be seen in Exhibits 6.1, 6.2 and 6.3.

### Conditional Inference Trees

The Conditional Inference Tree is an important variant of the traditional decision tree. Conditional inference trees are similar to traditional trees, but variables and splits are selected based on significance tests rather than purity/homogeneity measures. The significance tests are permutation tests. We generated Conditional Inference Tree for each candidate in R.

Overall, the performance of the Conditional Inference Tree was pretty good for all candidates except for Kasich with the following highlights for each candidate:

1. The output for Donald Trump had an accuracy of 75%, sensitivity of 58% and specificity 84%.
2. The output for Ted Cruz had an accuracy of 79%, sensitivity of 66% and specificity 86%.
3. The output for John Kasich had an accuracy of 96% but poor values for sensitivity at 100% and specificity at 0%.

The conditional inference tree outputs for the 3 candidates can be seen in Exhibits 7.1, 7.2 and 7.3.

## Logistic Regression

### Summary results of the best fit model for Logistic Regression

*Best fit model for Donald Trump*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deviance Residuals:** | | | | | |
| Min | 1Q | Median | 3Q | Max |  |
| -3.0119 | -0.6553 | -0.3204 | 0.6087 | 3.2617 |  |
| **Coefficients:** | | | | | |
|  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |  |
| (Intercept) | 1.31E+01 | 3.10E+00 | 4.21 | 2.55E-05 | \*\*\* |
| Age775214 | -1.20E-01 | 2.20E-02 | -5.431 | 5.60E-08 | \*\*\* |
| Edu635213 | 1.08E-01 | 2.01E-02 | 5.374 | 7.68E-08 | \*\*\* |
| Lfe305213 | -9.91E-02 | 1.80E-02 | -5.497 | 3.86E-08 | \*\*\* |
| Lnd110210 | -2.50E-04 | 8.11E-05 | -3.087 | 0.002019 | \*\* |
| Man450207 | 1.26E-07 | 3.51E-08 | 3.588 | 0.000333 | \*\*\* |
| Nes010213 | 8.90E-05 | 2.19E-05 | 4.071 | 4.67E-05 | \*\*\* |
| Pop010210 | -8.71E-06 | 2.15E-06 | -4.059 | 4.92E-05 | \*\*\* |
| Pop645213 | -2.26E-01 | 3.83E-02 | -5.891 | 3.83E-09 | \*\*\* |
| Pvy020213 | -5.79E-02 | 1.98E-02 | -2.921 | 0.00349 | \*\* |
| Rhi225214 | -2.59E-01 | 2.42E-02 | -10.738 | < 2e-16 | \*\*\* |
| Rhi325214 | -2.13E-01 | 3.41E-02 | -6.236 | 4.48E-10 | \*\*\* |
| Rhi825214 | -2.47E-01 | 2.32E-02 | -10.646 | < 2e-16 | \*\*\* |
| Rtn131207 | -5.59E-05 | 1.99E-05 | -2.814 | 0.004887 | \*\* |
| Sbo015207 | -2.04E-02 | 6.66E-03 | -3.061 | 0.002207 | \*\* |
| Sbo315207 | -4.07E-02 | 2.02E-02 | -2.016 | 0.043814 | \* |
| Sbo415207 | -1.56E-01 | 2.26E-02 | -6.923 | 4.43E-12 | \*\*\* |
| Sex255214 | 1.45E-01 | 4.01E-02 | 3.623 | 0.000291 | \*\*\* |

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1945.8 on 1503 degrees of freedom

Residual deviance: 1245.6 on 1486 degrees of freedom

AIC: 1281.6

Number of Fisher Scoring iterations: 6

*Best fit model for Ted Cruz*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deviance Residuals:** | | | | | |
| Min | 1Q | Median | 3Q | Max |  |
| -3.3048 | -0.7422 | -0.4546 | 0.6802 | 2.9635 |  |
|  |  |  |  |  |  |
| **Coefficients:** | | | | | |
|  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |  |
| (Intercept) | -6.53E+00 | 2.13E+00 | -3.073 | 0.002122 | \*\* |
| Edu635213 | 7.46E-02 | 1.95E-02 | 3.838 | 0.000124 | \*\*\* |
| Hsd310213 | 1.97E+00 | 4.25E-01 | 4.627 | 3.71E-06 | \*\*\* |
| Lfe305213 | -1.15E-01 | 1.74E-02 | -6.637 | 3.19E-11 | \*\*\* |
| Inc110213 | -3.39E-05 | 1.37E-05 | -2.467 | 0.013621 | \* |
| Pop060210 | -1.13E-03 | 3.40E-04 | -3.31 | 0.000932 | \*\*\* |
| Pop645213 | -9.51E-02 | 3.13E-02 | -3.045 | 0.002329 | \*\* |
| Pop815213 | 1.49E-01 | 2.06E-02 | 7.252 | 4.12E-13 | \*\*\* |
| Pvy020213 | -1.15E-01 | 2.26E-02 | -5.068 | 4.01E-07 | \*\*\* |
| Sbo415207 | -8.19E-02 | 1.77E-02 | -4.637 | 3.54E-06 | \*\*\* |

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1847.2 on 1503 degrees of freedom

Residual deviance: 1426.9 on 1494 degrees of freedom

AIC: 1446.9

Number of Fisher Scoring iterations: 5

*Best Fit model for John Kasich*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Deviance Residuals:** | | | | |  |
| Min | 1Q | Median | 3Q | Max |  |
| -4.8915 | 0.0623 | 0.1285 | 0.2505 | 1.101 |  |
|  |  |  |  |  |  |
| **Coefficients** | | | | | |
|  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |  |
| (Intercept) | 23.68884 | 5.865074 | 4.039 | 5.37E-05 | \*\*\* |
| Edu685213 | 0.078356 | 0.030805 | 2.544 | 0.010971 | \* |
| Edu635213 | -0.27663 | 0.046475 | -5.952 | 2.65E-09 | \*\*\* |
| Hsd310213 | 2.959502 | 1.464228 | 2.021 | 0.043259 | \* |
| Sbo015207 | -0.03688 | 0.016099 | -2.291 | 0.021988 | \* |
| Age775214 | 0.357502 | 0.104545 | 3.42 | 0.000627 | \*\*\* |
| Pop715213 | -0.12277 | 0.052252 | -2.35 | 0.018794 | \* |
| Pop060210 | -0.00052 | 0.000246 | -2.102 | 0.035544 | \* |
| Pst120214 | 0.185744 | 0.071809 | 2.587 | 0.009692 | \*\* |

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 404.47 on 1503 degrees of freedom

Residual deviance: 319.69 on 1495 degrees of freedom

AIC: 337.69

Number of Fisher Scoring iterations: 8

### Confusion matrix output for each model

*Confusion matrix for 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

*Confusion matrix for Ted Cruz*

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Ted Cruz | Not Ted Cruz |
| Ted Cruz | 218 | 37 |
| Not Ted Cruz | 45 | 77 |

Sensitivity = 0.63

Specificity = 0.85

Positive Predictive Value = 0.63

Negative Predictive Value = 0.85

Accuracy = 0.78

*Confusion matrix for John Kasich*

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | John Kasich | Not John Kasich |
| John Kasich | 0 | 14 |
| Not John Kasich | 1 | 362 |

Sensitivity = 1

Specificity = 0

Positive Predictive Value = 1

Negative Predictive Value = 0

Accuracy = 0.96

## 

## Random Forests

### Confusion matrix output for Donald Trump

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Donald Trump | Not Donald Trump |
| Donald Trump | 213 | 23 |
| Not Donald Trump | 52 | 89 |

Sensitivity = 0.63

Specificity = 0.9

Positive Predictive Value = 0.63

Negative Predictive Value = 0.9

Accuracy = 0.8

### Exhibit 5.2: Confusion matrix output for Ted Cruz

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Ted Cruz | Not Ted Cruz |
| Ted Cruz | 244 | 11 |
| Not Ted Cruz | 52 | 70 |

Sensitivity = 0.57

Specificity = 0.96

Positive Predictive Value = 0.57

Negative Predictive Value = 0.96

Accuracy = 0.83

### Exhibit 5.3: Confusion matrix output for John Kasich

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | John Kasich | Not John Kasich |
| John Kasich | 1 | 13 |
| Not John Kasich | 0 | 363 |

Sensitivity = 1

Specificity = 0.07

Positive Predictive Value = 1

Negative Predictive Value = 0.07

Accuracy = 0.97

## Exhibit 6: Random Vector Machines

### Exhibit 6.1: Confusion matrix output for Donald Trump

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Donald Trump | Not Donald Trump |
| Donald Trump | 218 | 18 |
| Not Donald Trump | 55 | 86 |

Sensitivity = 0.61

Specificity = 0.92

Positive Predictive Value = 0.61

Negative Predictive Value = 0.92

Accuracy = 0.81

### Exhibit 6.2: Confusion matrix output for Ted Cruz

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Ted Cruz | Not Ted Cruz |
| Ted Cruz | 243 | 12 |
| Not Ted Cruz | 60 | 62 |

Sensitivity = 0.51

Specificity = 0.95

Positive Predictive Value = 0.51

Negative Predictive Value = 0.95

Accuracy = 0.81

### Exhibit 6.3: Confusion matrix output for John Kasich

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | John Kasich | Not John Kasich |
| John Kasich | 0 | 14 |
| Not John Kasich | 0 | 363 |

Sensitivity = 1

Specificity = 0

Positive Predictive Value = 1

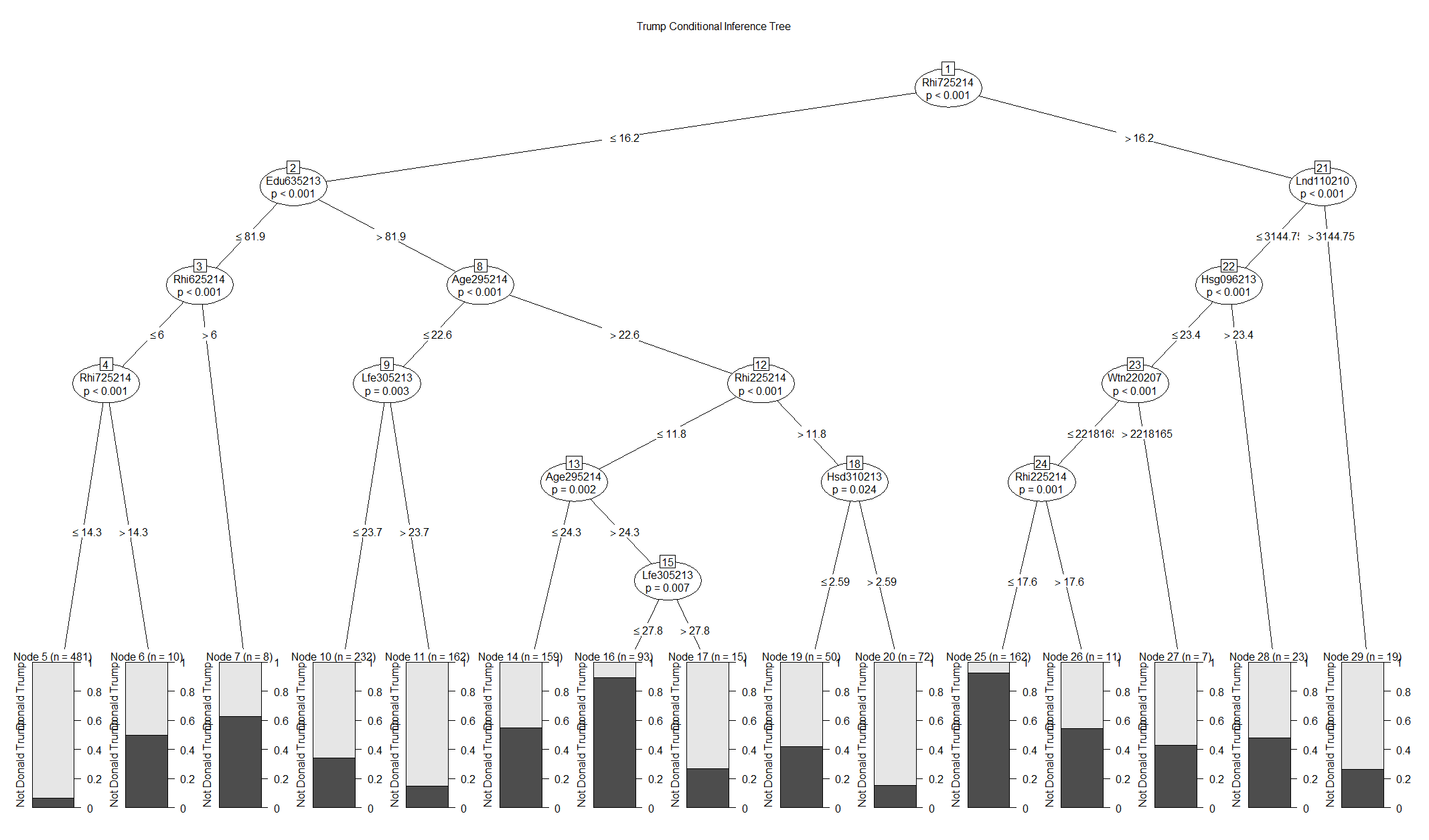
Negative Predictive Value = 0

Accuracy = 0.96

## Exhibit 7: Conditional Inference Tree

### Exhibit 7.1: Conditional Inference Tree for Donald Trump

Conditional Inference Tree for Donald trump



Confusion Matrix for Donald Trump

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Donald Trump | Not Donald Trump |
| Donald Trump | 199 | 37 |
| Not Donald Trump | 59 | 82 |

Sensitivity = 0.58

Specificity = 0.84

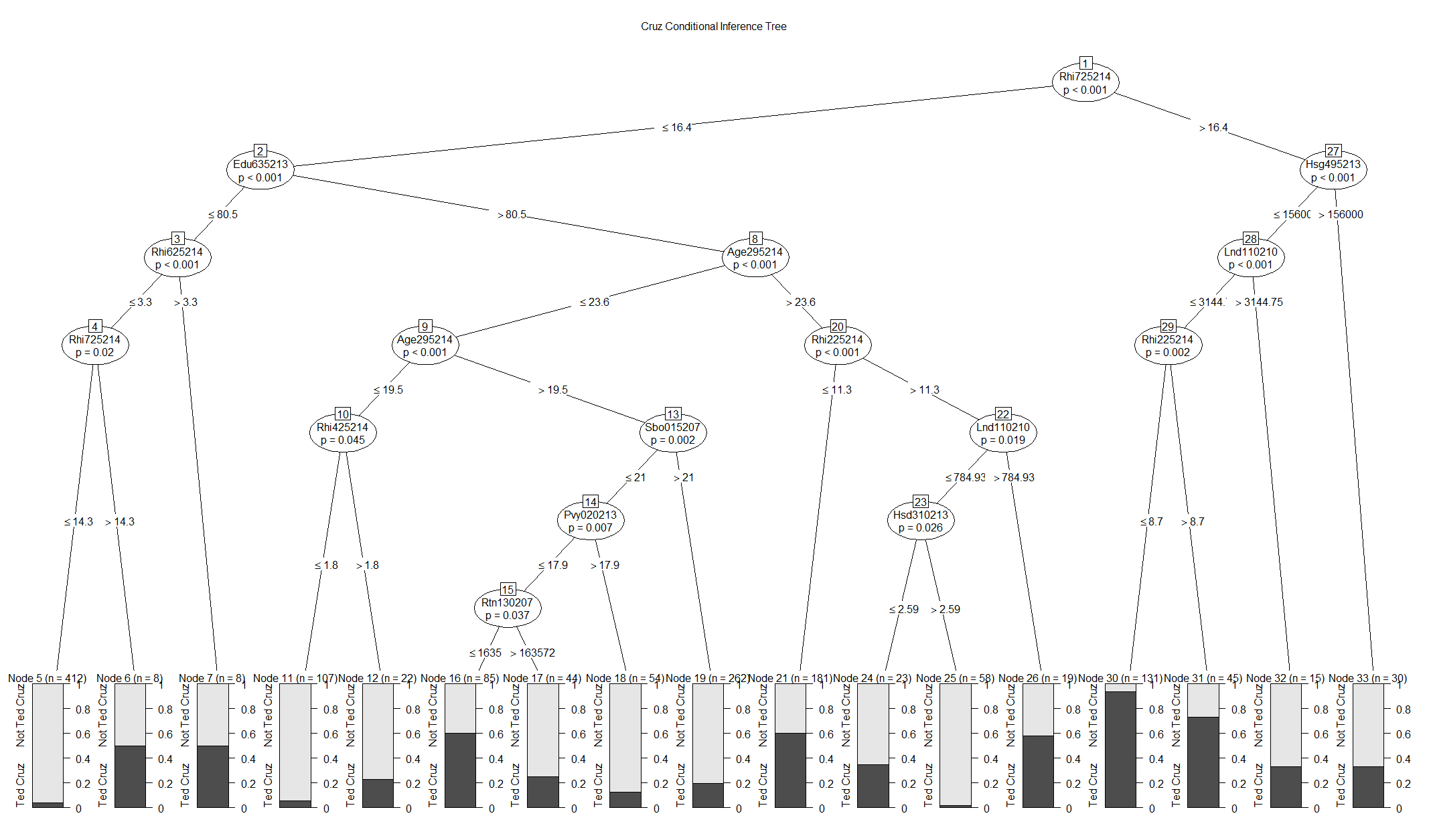
Positive Predictive Value = 0.58

Negative Predictive Value = 0.84

Accuracy = 0.75

### Exhibit 7.2: Conditional Inference Tree for Ted Cruz

Conditional Inference Tree for Ted Cruz



Confusion Matrix for Ted Cruz

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | Ted Cruz | Not Ted Cruz |
| Ted Cruz | 219 | 36 |
| Not Ted Cruz | 42 | 80 |

Sensitivity = 0.66

Specificity = 0.86

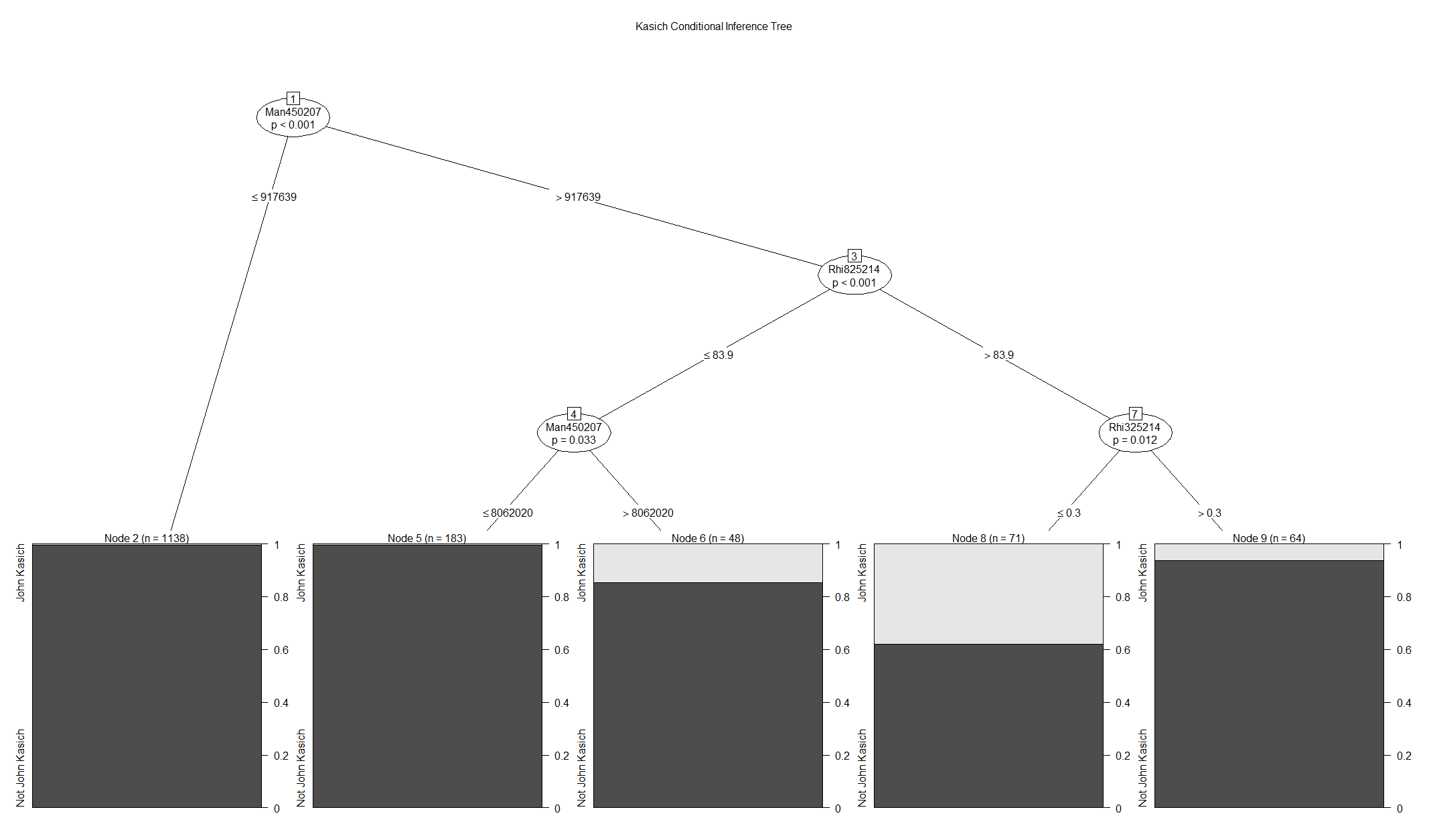
Positive Predictive Value = 0.66

Negative Predictive Value = 0.86

Accuracy = 0.79

### Exhibit 7.3: Conditional Inference Tree for John Kasich

Conditional Inference Tree for John Kasich



Confusion Matrix for John Kasich

|  |  |  |
| --- | --- | --- |
|  | **Predicted** | |
| **Actual** | John Kasich | Not John Kasich |
| John Kasich | 0 | 14 |
| Not John Kasich | 0 | 363 |

Sensitivity = 1

Specificity = 0

Positive Predictive Value = 1

Negative Predictive Value = 0

Accuracy = 0.96