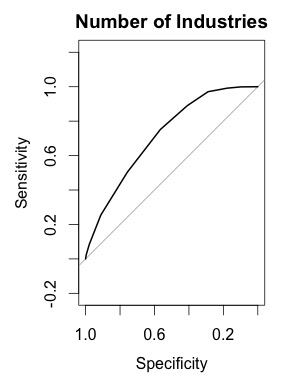
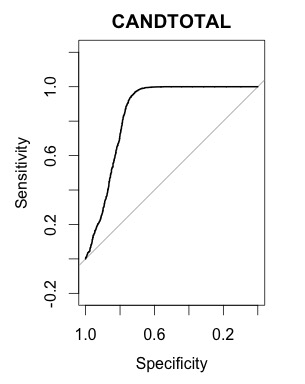
The first Null hypothesis was tested was: the is no difference between the total contributions that an incumbent gets and one that a challenger gets. We used our Merged Data set without outliers. We preformed a **student t-test** (not pairwise) to test this null hypothesis. We got a p-value of < 2.2e-16 and given that that this a social science analysis, the threshold should be .05. The p-value crosses this threshold and is well in the rejection region, so the p-value of extremely significant. So we reject the null hypothesis in favor of the alternative which is that there is a difference. Given the mean of the two categories of contributions, it is clear that the incumbent has a higher amount of the contributions compared to the challenger. This is also verifiable when compared to the association rules.

The Second Null hypothesis that tested involved a **logistic regression model**. The idea behind the model is to predict who the winner will be based on the Total amount of money raised and the Number of supporting industries. From the confusion matrix, we know the following. The Precision of the model is 0.8125881, the recall is 0.8662994, is F-measure is 0.8385846. Below are the ROC plots of both variables.



From the ROC curved, it seems like the logistic model predicts many false positives. This could signify that the model maybe be over fitting. This is probably because the model is predicting the values based upon itself. As excepted, the total amount of money raised seems to have a bigger impact on accuracy as opposed to Number of Industries.

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | lose | win |
| Lose | 1273 | 399 |
| win | 267 | 1730 |

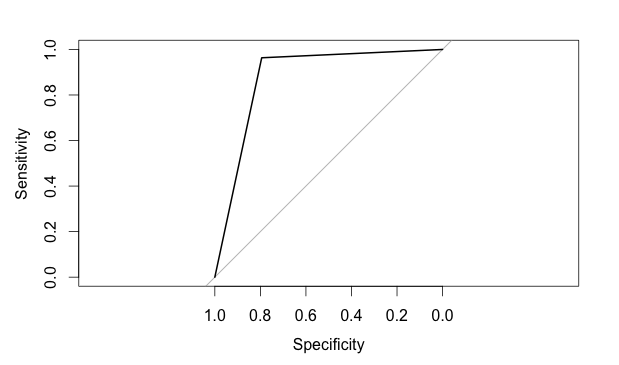
When running the decision three algorithm, the actual make-up of the candidate total was looks at in attempting to predict election winners. The Null hypothesis was that the nth number of total contribution had no effect of if a candidate won. By applying the decision tree, it became clear that the first ranked industry contributor total that more effect. However, it’s interesting to note from the results, that rank five also had an effect of winner ability.

A **third hypothesis** that we tested using the decision tree was that the incumbency status of a candidate should not affect whether or not the candidate would win. From the results of the decision tree, it became clear that incumbency status has any effect on who wins. The Decision Tree also expanded the first hypothesis by showing that the Total amount of contributions has an affect on. In addition, the Tree confirmed that the yrpercentchange variable has no effect on whether a candidate wins or loses.

From Cross Validating the Decision Tree

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| tree model | confusion matrix | accuracy | variance | TRUE | false positive | true positive |
| run1 | tree\_predict\_after\_prune  L W  L 222 15 W 45 258 | 0.889 | 0.099 | 300 | 0.150 | 0.850 |
| run2 | tree\_predict\_after\_prune  L W L 204 14 W 44 278 | 0.893 | 0.096 | 322 | 0.137 | 0.863 |

After running the decision tree 10 times, the mean(accuracy of predicting a winner) is 0.881 and variance is 0.105,



The Area under the ROC curve is 0.8929 with a 95% CI: 0.8672-0.9187. The area under the curve is pretty high, so the model seems to work pretty well.

The same situations were then applied via a lazy learner model: K- nearest neighbor. The general takeaway was that remove any one of the variables (amount from rank 1 industry to rank 5 industry) did not adversely affect the overall prediction of winners.

Confusion Matrix which regards to industry contribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *without attribute: rank1 amount of money* | *without attribute: rank2 amount of money* | *without attribute: rank3 amount of money* | *without attribute: rank4 amount of money* | *without attribute: rank5 amount of money* | *without attribute:*  Candtotal |
| L W  L 176 85  W 6 273 | L W  L 194 70  W 14 262 | L W  L 179 72  W 18 271 | L W  L 193 63  W 16 268 | L W  L 200 57  W 17 266 | L W  L 184 75  W 12 269 |
| Accuracy: 0.831482 | Accuracy: 0.844444 | Accuracy: 0.833333 | Accuracy: 0.853703 | Accuracy: 0.862963 | Accuracy: 0.8388889 |
| Variance: 0.14038 | Variance: 0.131602 | Variance: 0.139147 | Variance: 0.125125 | Variance: 0.118477 | Variance: 0.1354051 |

While it seems like certain attributes do not affect the overall score, the biggest takeaway is that a combination of these attributes have a high ability to predict which candidates can win.

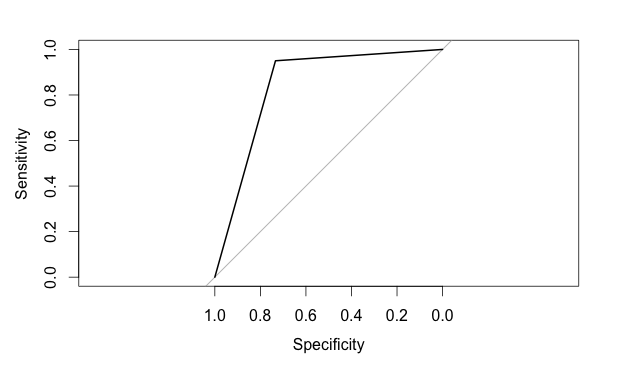


Figure ROC Curve

The area under the curve is 0.8419 with 95% CI: 0.8121-0.8717 (DeLong).

The area under the curve is pretty high, which means the model works pretty well.

The following confusion matrix was done for the Naïve Bayes example

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *without attribute: rank1 amount of money* | *without attribute: rank2 amount of money* | *without attribute: rank3 amount of money* | *without attribute: rank4 amount of money* | *without attribute: rank5 amount of money* | *without attribute:*  Incumbent | *without attribute:*  YrPercentChange |
| L W  L 252 38  W 27 223 | L W  L 228 32  W 36 244 | L W  L 249 37  W 20 234 | L W  L 247 36  W 27 230 | L W  L 235 40  W 31 234 | L W  L 134 17  W 135 254 | L W  L 240 36  W 37 227 |
| Accuracy: 0.8796296 | Accuracy: 0.8740741 | Accuracy: 0.8944444 | Accuracy: 0.88333 | Accuracy: 0.8685185 | Accuracy: 0.7185185 | Accuracy: 0.8648148 |
| Variance: 0.1060778 | Variance: 0.1102728 | Variance: 0.09458874 | Variance: 0.1032468 | Variance: 0.114406 | Variance: 0.2026249 | Variance: 0.1171271 |

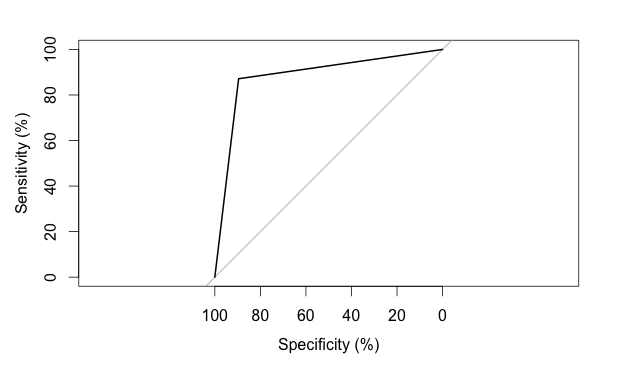


Figure ROC Curve

The Area under the ROC curve: 87.05% with 95% CI: 84.22%-89.87% (DeLong). The area under the curve is pretty high, which means the model works pretty well.

The results from Naïve Bayes confirmed that incumbency is an important factor in determine the winning candidate. This can be seen because the accuracy is significantly lower than when removing any other attributes.