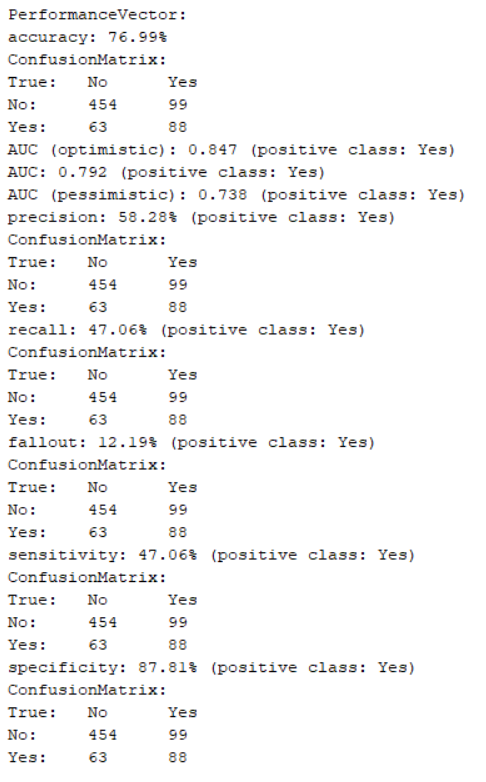
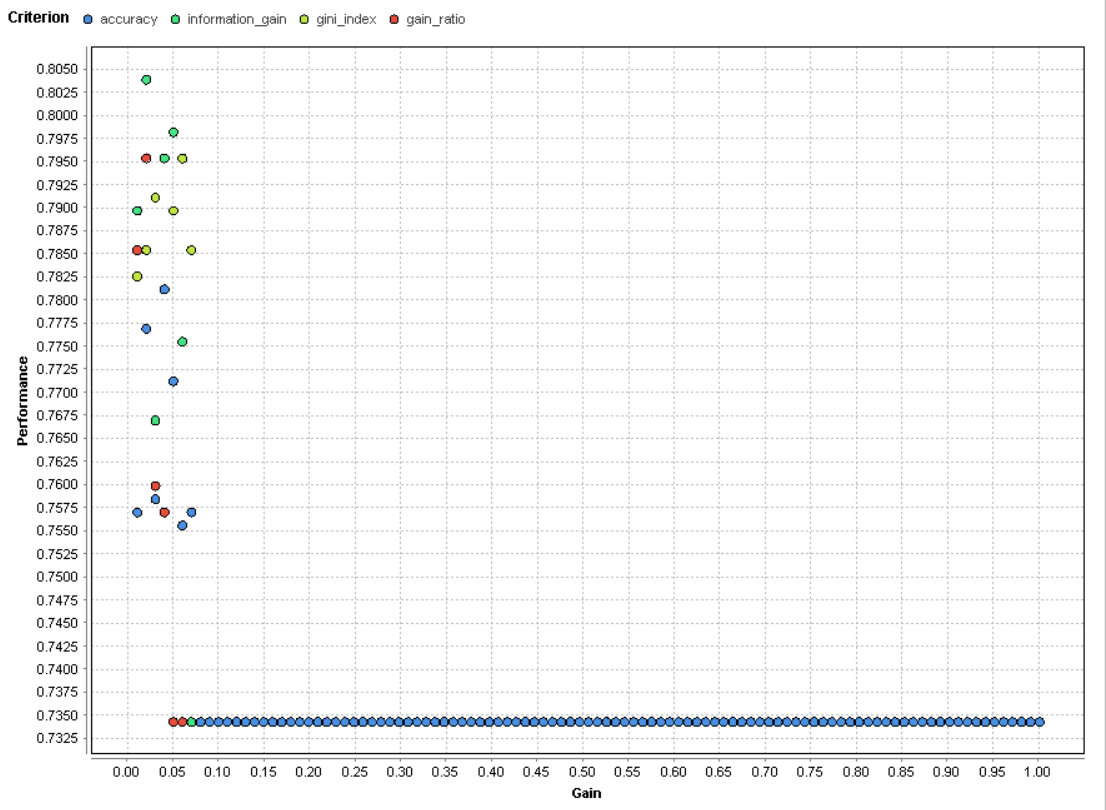
Nick Iudiciani

MIST4060-201

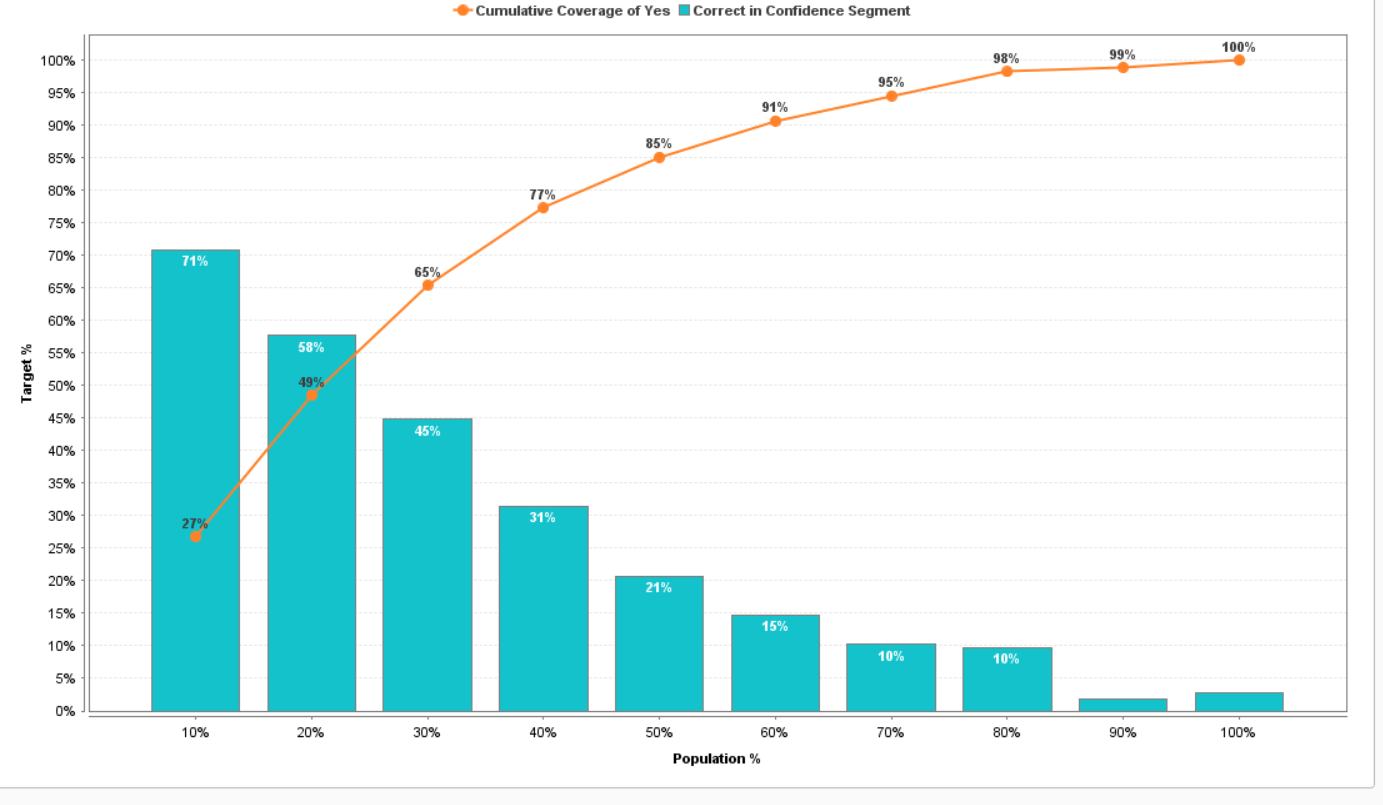
Prof. Amit Deokar

Homework Assignment 2

**Performance Vector Analysis**

At right is the performance vector. After running the optimization (Grid) operator in a separate process for the decision tree, I was given that “information gain” was the optimized criterion and 0.0199 was the optimal minimal gain. I discerned these as true because the minimal gain of 0.0199 is close to the minimal gain we had initially used of 0.01. The chart (below) shows that accuracy was the most used criterion (blue dots), but it yielded the lowest performance, unlike information gain (green dot at the top of the chart).

**Lift Chart Analysis**

 The lift chart shown below allows us to compare what a machine learning model would do against randomized guesses of information and confidence. The bars are in decreasing order to show that the more population of the churn data set that we reach, the less confident the model would be at predicting the churn rate. The last two columns break this pattern of decreasing order likely because both population sizes are so similar and yet the confidences are both 0 or near 0. The orange line however shows something different. It shows the cumulative coverage of customers who churned if we only used its corresponding bar (blue) as its confidence. For example, the third blue bar shows 45% of the target population. The third orange point shows 65% coverage. This means that the model would correctly identify a churn rate of 65% if we only used 45% of the population (or 45% of the highest confidences for the target class).

**ROC Curve Analysis**

A ROC curve is designed as a nested operator to compare the sensitivity (true positive rate on y-axis vs false positive rate on x-axis) of different types of operators, in this case a decision tree with the optimal parameters, a k-nearest neighbors and Naïve Bayes. The chart shown below is a ROC chart with optimistic confidence, though being neutral or pessimistic would produce different charts. Whichever curve follows the y-axis the most, that test is the most accurate (in this case the decision tree). However, envision a 45 degree angle cutting the graph in half diagonally. As the curves come closer to that imaginary line, the less accurate that test becomes (which would be the k-NN green curve cutting into the dotted line I drew on the graph).

