Lachlan Sinclair

3/14/2020

Week 10 Assignment 2

Reflections

a. As the influence of false positives rises, the values for the logistic regression model don't see much change. From 10-13x the values retrieved from the test set and the AUC values are barely changed. The 10x predictor does differ from the other three, in terms of the values obtained via the lift chart, but again this difference is very small. The predictor importance is different in all 4 rows however many of the variables are reused. The C5.0 results were less static, their AUC's values on the training set were across the board higher than the LR's, however their AUC values for the test set were actually worse than the LR AUC test set values. The training lift max value also followed this pattern. I would say this is a sign that the C5.0 predictor is more overfit than the LR predictor, having your predictor perform far better on the training set in comparison to the test set is a strong indicator of overfitting. Unfortunately, I didn't have time to get extra samples, but the overall trend looks like this difference in performance on the training and test set increased as the influence of the positive instances was increased. The C5.0 also had far more variety its top ten predictors, some variables did make appearances in both types of predictors, such as 210, which happened to be in the top 2 in all the LR predictors in the 10x-13x rows.

b. If we were to consider a situation where we want to identify a sub population of positives (for the 11x increase of positive instance influence) I would say that the C5.0 tree would be the better classifier. As the sub population gets very small, say 1%, my runs show the lift of the tree classifier was 3.8952 and the lift for the LR model was 3.7087. The tree model also outperforms the LR model at 2% of instances. While the difference isn't massive the tree classifier clearly out preforms the logistic regression model when selecting this specific small number of instances. The LR predictors lift curve starts off lower and decreases at quicker rate but then transfers to a less steep slope. The tree model doesn't have as steep of a decreasing slope at the start of the lift graph and starts off higher for the first 2%, but it doesn't transition to a shallower slope as quickly as the LR predictor. This difference in layout of the lift graphs means the tree model can provide a higher concentration of true positive instances when a small portion of the instances are selected. Having more true positive instances located at the start of the lift graph makes the C5.0 classifier more suited for a situation where tightly identifying a sub population of true positives is desired.