Lachlan Sinclair

3/8/2020

Assignment 1

Reflections

a. For the C5.0 tree there is a continual decrease in predictor importance that very loosely follows an exponential decay. The logistic regression predictor importance is far more piecemeal, looking at it there appears to be three separate groups that have relatively similar magnitudes of importance's. The first two predictors, house and overage in the LR model provide the same high value of importance. Then there are four that have a noticeably smaller amount of importance (between .04 and .09), this is then followed by another four with almost no importance (.01). The C5.0 trees predictors continually ramp down in importance before finally reaching 0. Aside from the first two values the order of the predictor's importance are entirely different. It is interesting that both methods placed such high values on house and overage, along with reasonably high values on leftover. Also, out of the lowest four predictors they had three in common, so regardless of the method used there are some attributes that provide a solid amount of information and those that don't.

b. The first thing I noticed in the C5.0 analysis was how the ratios of matrices values in the training are similarly represented in the testing sections. By this I mean that TN is greater than TP and FP is less than FN in both sections. As the problem statement mentions there is a 49.7 to 50.3 ratio of positive/negative instances in the data set. This adds some reasoning to the ratios I am seeing but isn't enough to fully account for the differences. Since mistakes are weighted the same, in theory we should see slightly higher values in TN and FN when compared to TP and FP respectively. However, this is immediately contradicted by the LR model. For the LR model FP was actually higher than FN which in turn explains how TP was higher than TN, even though the make up the data set had the 49.7 50.3 ratio. This difference is present in both the training and testing sections and could be explained by a variety of factors. It is also worth noting that the performance of the LR model was significantly higher than that of the C5.0 model, in both training and testing. After spending some time looking both results the fact than you can compare the sums of the FN+FP to understand the relative accuracy of the models and the meaning behind it really stood out to me(in case you glance over the sections that represent this).