Lachlan Sinclair

3/15/2020

Week 10 Assignment 1

Reflections

a. The predictor importance from the two predictor importance charts show that the positive instances (1) do not have that many similarities, such that those similarities are also not present in the negative instances. The C5.0 tree produced just a negative one in the column beside the predictor importance chart. And the predictor importance chart said, “the predictor importance is not available”, I also checked the summary page which showed the tree depth was 0. I interpreted this as this predictor determined it was better to not use any nodes that partition the instance (other than one possibly to set them all to -1), but rather blindly sets every value to negative. The logistic regression predictor did end up using predictors, the predictors used had incredibly low importance values which adds to the claim the positive instances do not have a high consistency. Given that we didn’t compensate for roughly 93% of the data being negative instances, it makes sense that unless there was an extremely strong correlation between some attributes and the positive instances the tree model would decide its best to not apply any logic. The regression model will use an algorithm that penalizes it for errors corresponding to the distance from the “line” to the incorrect instance, so it would be unreasonable to expect it to use a “line” such that it has all of the data on one side, so it is forced to make an attempt.

b. Both models in terms of the percentage correct are incredibly close, the tree model performed slightly better on the testing set and slightly worse on the training set. Looking at the coincidence matrix for the tree model it is clear that it did not make any positive predictions on either set. The logistic regression model does make attempts at assigning positive instances and doesn’t do so all that well, on the testing set the true positives are about half the size of the false positives which isn’t great. However due to the size of this set this has little impact on the comparative overall percent of correct assignments. The AUC shows that the logistic regression model performed slightly better than random with a value of .675 on the testing set. The tree model had a value of .5 for the AUC which is equivalent to the random guessing, and from the reading this means it doesn’t provide discrimination for the classifier. So given this dataset, in these configurations arguably the LR predictor performs better if you would like to find at least a few of the TP values and are willing to make a few more mistakes to do so. The tree model would be best if the overall accuracy is the greatest concern.