Problem Set 10

Matt Mullins

April 2018

1 Algorithm Results

Algorithm Type	F1 Result	Geometric Mean Result
Tree Model	0.8968421	0.6730932
Logit Model	0.8986422	0.6762722
Neural Network Model	0.9059357	0.7564711
Naive Bayes Model	0.8841864	0.7259584
kNN Model	0.8985478	0.7543678
SVM Model	0.9090018	0.7539106

2 Results Findings

For out-of-sample test results, we see that they are all to an accuracy of 90 percent, given to us by the f1 score each reported. With the best score going to Neural Network Algorithm and the worst being Naive Bayes (sadly). Since we are looking to determine whether or not someone has is a high-earner, I would say, depending on the sample size, all of the different methods would be nearly identical in effectiveness. Based on our tolerance level, 90 percent accuracy seems like a great place to start with these learning methods

For the geometric mean results, we have more variance between our 6 learning algorithms, ranging from the highest of 0.756 for Nueral Networks and the lowest being 0.673 for the Tree Model. I believe that this is indicating that the high-earners are equal to or above this percentage in the population. In other words, we would expect the number of high-earners in the sample to be above a percentage around 0.67-75. Or, that 25 percent of the sample is being classified as a high-earner.