Summary and Problem Statement

In this exercise, we are extending the set of classification models built on data from a bank marketing study.[1] Based on Assignment 2, we are adding SVM classifiers to the suite of models. To the 4 binary variables (3 explanatory, 1 response) used in assignment 2, we added “balance” which is a continuous variable. The bank wants to increase customers’ investments in term deposits by identifying the factors which, when used in marketing, yielded the best up-take in term deposits. Our goal: make a recommendation selected from the models built.

Methodology

The data were briefly explored using charts and summary outputs[2] and checked for completeness. The binary variables were coded to be 0 for a “no” value and 1 for a “yes”. New analysis was done using SVM linear and non-linear classification methods. Tuning experimentation was done by changing C. Results were evaluated comparing the average ROC value over 10-cross validation folds. Finally, models were run against synthetic test data to identify which “profile” (values for the explanatory variables) produced the highest probability of a “yes” response to the marketing campaign based on the training data.

Code Overview

After reading in the data, and checking for missing values, the binary yes/no columns were encoded then models were fit. Proceeding in order a logistic regression model was fit, followed by 2 SVM linear classifiers with different parameters, and SMV using RBF kernel and finally an SVM using the polynomial kernel were fit. For each model a ten-fold, cross-validation set of predictions are generated and used to compute an average ROC score for that model. The summary of the average ROC for each model is printed at the end of the fitting and evaluation loops. Finally, the models are tested using synthetic test data composed of 16 possible combinations of the 4 explanatory variables. Results are output to the console.

Results and Recommendations

Based strictly on the average ROC[3] values from the cross-validation, logistic regression yields the best results. Of the SVM attempts, the linear SVM classifier, using a C=1 (Linear_SVM_C=1) performed the best. Looking at the predicted probabilities for each of the models on the test data, we see that for each model, the "profile" of explanatory variables with the highest likelihood of responding "yes" to the campaign, has default = 1. The population of training data only has 76 instances of default = 1, so this is an interesting finding. I infer from this that the value of default is the most important consideration. The best-probability-per-model findings don't show any additional consistent explanatory variables.

My recommendation: generally, use logistic regression models for identifying customers to target for campaigns. In this specific case, a targeted out-reach to using the logistic regression model to customers who have the profile default = 1, loans=0, housing=0 and balance >0, is predicted to yield the best up-take. For the broadest possible up-take, consider out-reach to all customers who have had a default. We see in the Linear_SVM_C=1 model, a predicted up-take probability of 41% is achieved[4] with the profile: default, housing, & loan = 1 and balance >0. The logistic model had the best ROC numbers, but Linear_SVM_C=1 has the best predicted response. Market to both, or possibly to all the folks with a default, since default=1 is common to the best iteration for each model.

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] See console_output.html file in submission

[3] See console_output.html file in submission

[4] See console_output.html, near the end, for probabiities