Summary and Problem Statement

In this exercise, we are fitting two different classification models using data from a bank marketing study.[1] Four variables (3 explanatory, 1 response) of the 17 variables contained in the full data set were used. All 4 of the variables used are binary, meaning they take a yes/no or 1/0 value. The data were collected by a Portuguese bank via 17 phone marketing campaigns conducted between May 2008 and November 2010. Information on the sampling method used to select call recipients is not provided, so the composition of the sample relative to the population is not known. The bank wanted to increase customers' investments in term deposits by identifying the factors which, when used in marketing, yielded the best up-take in term deposits.

Methodology

The data were explored using charts and summary outputs.[2] Analysis was done using logistic regression and naïve Bayes methods. Results were evaluated comparing ROC and Accuracy numbers for the two models; ROC charts were created[3]. Confusion matrices[4] based off the cross-validation predicted probabilities were also made. F1 scores were also calculated off the cross-validated predictions.

Code Overview

Exploratory visualizations come first, then all code related to Logistic model, followed by all code for the naïve Bayes (NB) model. For the Logistic regression model, the "l1" penalty term was used as it gives better results than "l2"; possibly because of the imbalance in the response variable. The NB model uses the BernoulliNB method as the most appropriate for the binary data. All other model parameters were the defaults. ROC and accuracy values are reported for

each of the cross-validation folds[5]; the ROC values were then averaged to produce a single value for comparisons. To generate the confusion matrices and the F1 scores, I iterated over the results of the cross-validation prediction probabilities. For each probability pair (Pr(0) and Pr(1)) if the probability of 0 was at or above .88, output class 0, otherwise output class 1 as the predictions. A threshold of .88 was used to match the frequencies in the training data.

Results and Recommendations

The models fit to this data are not likely to be good predictors. The ROC values, as seen in the individual cross-validation fold results[2] and as charted[3], show the models are not much of an improvement over the 50% "random guess" line; a plot of Precision[6] versus Recall confirms this. Using just 3 explanatory variables may be part of the problem. Another factor that may contribute to the poor performance is the fact response variable is imbalanced[7]. The "no" or 0 values are 88% of the training set; hence, a prediction of 0 will be correctly roughly 88% of the time. There is little difference between the two models; average accuracies are the same, 88% which is the same as just always predicting 0. The average ROC score for the naïve Bayes model is only .001 better than that for Logistic regression; there is little to recommend one model over the other.

Rather than attempt to put either model into production, I recommend 1) collecting more data and re-building the models, or 2) expanding the models to use more than the current 3 explanatory variables or 3) trying other techniques, perhaps Random Forests. While the average naïve Bayes ROC number was slightly higher than that of Logistic regression, the results are so similar[4] that I would try both techniques with the additional data and/or expanded feature set before selecting between them.

[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 output.html file in submission
[3] See roc_curve_LR.pdf and roc_curve_NB.pdf
[4] See conf_LR.pdf and conf_NB.pdf
[5] See output.html file in submission
[6] See pr_curve_LR.pdf
[7] See count_of_vars.pdf