Brandon O'Briant
PREDICT 422-DL Practical Machine Learning
Assignment 2

Predicting Term Deposit Participation: Classification Model Comparison

Introduction

Telephone marketing campaigns can provide the means to reach customers and inform them of banking services. These campaigns can be useful if they provide a positive return on investment. To evaluate if the telephone marketing campaign to recruit participation in term deposit services at your bank will be successful, using classification machine learning models to help determine which customers will participate in term deposits can help reduce the cost of employee time and organizational resources.

Exploratory Data Analysis Methods

The data used in this evaluation was obtained from a previous telephone marketing campaign. The dataset consists of 4521 respondent's answers (rows) to 17 marketing questions (columns). Non-values were dropped if present in data. Removal of outliers was not performed since the raw observations related to a meaningful scale. Model data consists of three explanatory variables, *default*, *housing*, and *loan*, were used and one response variable, *response*, resulting in 4521 rows and 4 columns. Descriptive statistics were conducted, such as the mean, 0.2128, standard deviation, 0.4092, median, 0.000, and variance 0.1675 for model data. The area under the receiver operating characteristic curve (ROC AUC) was used as an index for classification performance classification machine learning models. The mean ROC AUC for the logistic regression model is 0.6079, and the mean for ROC AUC for naives Bayes model is 0.6081.

Brandon O'Briant
PREDICT 422-DL Practical Machine Learning
Assignment 2

Overview of Programming Work

Python was used exclusively for the analysis of the telephone direct marketing data, including the use of the following packages: Pandas and Numpy for data handling, and Scikit Learn for machine learning and model evaluation metrics. Telephone direct marketing campaign data was presented to us as a CSV file and loaded into the program using Pandas. Explanatory and response variables were transformed from categorical responses to binary: 0 for no, 1 for yes. Subsets of data were taken via standard DataFrame slice operations via indexes. Descriptive statistics were calculated via built in functions in the Numpy environment. K fold cross validation design, with ten folds using the AUROC as index for classification performance, was used for both logistic regression and naives Bayes classification machine learning models and was implemented using the Python Scikit Learn environment. All the results were saved to separate .txt. files.

Results and Recommendations

Examining the average AUROC for the logistic regression classification method, 0.6079, and the naives Bayes classification method, 0.6081, we see that the naives Bayes model performs slightly better for predicting customers that will participate in term deposits when using three explanatory variables, *default*, *loan*, and *housing*. Those most likely to participate also have no defaults. Thus, I recommend the naives Bayes method to direct telephone direct marketing campaigns towards those with no defaults. Also, I recommend further models to be explored with possible more explanatory variables being introduced to help better explain an individual's response.