MSDS 422: Practical Machine Learning

Professor Anil D. Chaturvedi

Assignment 2: Evaluating Classification Models

Tiffany Duong

Summary

Machine learning methods can be extremely helpful when advising a company on which business plan to execute. In this assignment, a bank needed advice with targeted marketing segmentation for future telephone marketing campaigns. The purpose of the telephone marketing campaigns was to engage clients into subscribing to a term deposit, i.e. a certificate of deposit with the bank. Alongside to seeing the types of customers that subscribe, the bank also wanted to assess which group of banking clients appears to be the best target for direct marketing efforts.

Research Design

In order to help guide the bank with marketing planning, a combination of exploratory analysis and building and evaluating classification models will be implemented. Some assumptions were made: all clients were as close as possible to being independent of one another and the quality of the contact methods were consistent. There are 16 features (excluding the response variable) and 4521 observation within the dataset. Exploring the data, we see that only 11.52% of all clients subscribed to a term deposit – rather a small amount. Looking at the quality of those subscribed clients vs. not subscribed we see that they don’t differ much (age, bank balance, education, etc), however average contact time was significantly higher for those that have subscribed.

Technical Overview

Analysis on the survey data was performed using traditional statistical methods and Python, a high-level general program language. Several Python packages were used, including NumPy, pandas, matplotlib, and seaborn, and scikit-learn. Scikit-learn was heavily used when setting up the classification models as well as performing the cross validation to see model performance. The predictive variable we will be looking at is the response variable – whether the client subscribed to a term deposit. Three explanatory variables will be used: whether the client has a housing loan, if the client has a personal loan, and if the client has defaulted on credit. Because these three explanatory variables are all binary, we will make use of binary classifiers – the logistic regression and the Naïve Bayes classifier. We will be comparing to see which classification model performs better and to conduct the comparison, we will evaluate the models using k-fold (10 folds in this case) cross validation, using the area under the ROC curve as an index of classification performance.

Results

With two classification models built out for three explanatory variables to predict response, I would recommend implementing the Logistic Regression as it yielded a slightly higher score of 61.12%, compared to the 61.11% from the Naïve Bayes Classifier. With the Logistic Regression, there is a 61.12% chance that the classifier will accurate predict that based on the three features, the model will predict response as a true positive.

Given the additional research done on targeting client groups, I think that we should continue to contact clients through cellphone as opposed to the traditional telephone. We also want to be targeting our slightly older, married clients that have a positive bank balance as our subscribed bank client mix seem to trend in these characteristics.