

Week 2 Assignment

By: Valon Tika

Summary and Problem Definition

Companies can leverage the use of telemarketing campaigns as a route to reach out to the market and to offer numerous services. This can lead to higher sales for a firm due to penetration of the market by the way of calling to customers. Companies in the financial services industry see this as a great opportunity to sell financial products to retain higher gross margins and to have more assets to allocate to possibly retain more earnings. One way to achieve this is to target customers that are probably more likely to participate in a given program through the use of a telemarketing campaign, rather than random dialing to customers.

Research Design, Measurement and Statistical Method

For this analysis, we can use the data from previous executed marketing campaigns and observe the data that was retained within those campaigns such as job occupation, housing, education, whether they have a personal loan, and other key pieces of information about the customer. It also contains financial related information such as outstanding loan amounts and average account balances.

There was around 11.5% of total respondents that said yes in this marketing campaign. Out of those 11.5%, the average balance was around \$1,570 dollars. Most of the individuals that response yes was contacted a little over 2 times on average during the campaign. Very few individuals that had a balance between \$5,000 to \$10,000 had a 15% on average of being interested in the campaign compared to the 75% of respondents who's account balances were under \$5,000.

Overview of Programming Work

The analysis of the marketing campaign data was performed in Python, a general-purpose programming language that provides many tools and packages to perform data analysis, data modeling, and predictive modeling in an efficient manner. One specific variable that seemed to show some impact to interest was the average balance, which was \$1,403 that were not interested vs \$1,572 for those that were. Balance, housing and loans were economic metrics deemed as the larger factors of interest. Those three metrics were used to compare 2 classification models. Naïve Bayes and Logistic Regression binary classifiers were chosen due to the robustness of these models for this scenario.

The classifiers will be different based on the model used. Performance of the models will be measured by using a standard performance metric, the ROC curve, or Receiver Operating Characteristic curve. This metric perspective of accuracy of predictions that will be relative to the TPR (true positive rate), as if we dial a customer on a false-positive, the negative outcome would be that they will reject, although they might not.

Review of Results – Recommendations

Since the ROC models can't be compared due to the training data being heavily biased for one class, training accuracy can be a valid criterion to evaluate the model. Training accuracy on logistic regression is 88% and 87% on Gaussian Naïve Bayes, hence the logistic regression is just 1% better than Naïve Bayes. Also, logistic regression performs much better on test data, but we can't use test data for comparison. The training accuracy is actually the mean of training + validation accuracy