Assignment 2 - Classification Models for Bank Marketing Study

1. Executive Summary & Problem Definition

Telephone marketing campaigns provide a way of reaching out to customers and informing them of services being offered to grow the business. A banking company wants to perform a new targeted marketing campaign and is interested in identifying factors that affect client responses in subscribing to new term deposit being offered. We need to assist the bank in narrowing down the scope of the study from all customers to those most likely to be interested in the program from previous campaigns. In doing so it will help to maximize ROI or return on investment in this upcoming campaign.

2. Research & Design

We can draw upon results from previously executed studies that tracked detailed information from prior calls, and important clientele demographics (age, job, marital status, education, etc.). Our analysis focused on two key areas, which align with our recommendation back to the bank management on:

1. Determine the best group of banking clients to target for future marketing campaigns
2. Run classification methods and determine the most appropriate machine learning model to identify the profile of people likely to subscribe.

3. Technical Overview

In target marketing, we need to identify factors that are useful and determine how to use those factors in modeling techniques to identify and quantify relationships between explanatory variables (customer demographics, past marketing campaign data) and outcome variable (likelihood to enroll in a term deposit). We approached through EDA and modelling techniques:

a. Exploratory Data analysis

We ran exploratory analyses using simple cross tabulations across various customer demographics and likelihood of them having applied for a bank term deposit. We also explored the variables of call metadata with the client looking for patterns to improve yes response rate.

b. Modeling and Feature selection

For this binary classification problem, we prototyped three probabilistic machine learning models for comparison; a logistic regression classifier, a naive Bayes Gaussian and naive Bayes Bernoulli classifier. These were used to predict if the client subscribed to a term deposit. We fitted models with predefined features: housing loan, default credit, and personal loans. Each feature is independent and equally weighted at start. Each model was evaluated with accuracy and ROC area-under-curve indices.

c. Cross Validation

We used a shuffle split cross validation method to have some consistency into our modeling. No logical stratification existed in the feature selection. The method maintained a training set with 80% of the data and a test set using the rest. We set the number of splits to 5 and erred on the side of a smaller number of splits. While we could have used more splits, we chose a smaller number due to the skewed dataset. Finally, a shuffle split has the advantage over K-fold cross-validation because this lessens the chance our model is split with an irregularly distributed number of targets, which would produce erratic results.

4. Conclusion and Recommendations

EDA: We identified several customer attributes that are associated with clients who have applied for bank term deposits in the past. While the majority of the customers have not applied for a term deposit (only ~11.5% of respondents) we saw a strong increase in yes responses to term deposit for students and retired clients (23% or double the average). The age group we believe was closely correlated to job status as the 20-25 age group and those 60 years of age responded yes to term deposits much higher than 25-59 ages. However, sample sizes in these groups were relatively small to full dataset where a larger dataset is recommended for these particular areas. Not having a housing loan was the strongest of the 3 predefined variables for the modeling part of the assignment. Clients who are college educated, not married, and don't have a personal loan have a small probability of success but we were careful not to overfit. Outside of demographics we did see that clients contacted more than once improved chance of success along with duration of last call, but not to contact too often as that lowered probability. These trends suggest that marketing campaigns targeted towards banking customers based on the findings above may be beneficial to increase call success rate, yielding higher revenue per campaign and lower operating expenses to maximize ROI.

Model selection: Out of the three models run using cross validation the Bernouli Naive Bayes would be the best model selected for this dataset because logistic regression does not perform well on an imbalanced dataset with only a few, binary variables. However, we have to be careful with the poor ~60 % AUC score returned. Our suggestion is to evaluate and implement the suggestions from EDA to improve yes response rate first prior to implementing this model. The very low 11.5 % acceptance term deposit acceptance rate led to all three models returning a very poor performance measure score for the three predefined exploratory binary variables where the banking firm is going to produce limited success implementing this model in current state.