Classification of Success for Bank Telemarketing Campaigns

Kaavya Gowthaman, Niyati Chopra and Kaushik Holla Vaderhobli Madhava Krishna

The data is related to direct marketing campaigns of a Portuguese banking institution, based on phone calls (Moro et al.). The goal of the campaigns was to get the clients to subscribe to a term deposit. There are 20 input variables and one binary output variable (y) that indicates whether the client subscribed to a term deposit with values 'yes' and 'no'.

The input variables can be divided into four categories: bank client data, data related to the last contact of the current campaign, social and economic context attributes, and other attributes. Bank client data contains variables containing information about the client. It includes variables indicating age, job, marital status, education, whether they have credit in default, whether they have a housing loan, and whether they have a personal loan. Data related to the last contact of the current campaign contains variables indicating the mode of communication, the month of last communication, the day of the week when the last contact was made, and the last call duration. Social and economic context attributes contain variables with the quarterly employment variation rate, monthly consumer price index, monthly consumer confidence index, number of employees, and the Euribor 3 month rate. (Moro et al.)

Other attributes include the number of previous contacts with the client during the current campaign, number of days since the last contact for the previous campaign, number of contacts performed before the current campaign for the client, and the outcome of the previous marketing campaign. The goal of the project is to classify with high accuracy whether the campaign will be successful or not given a set of input variables.

The above data parameters will be used to predict the outcome of the marketing campaign for a given customer. The ggplot2 package will be used for basic visualization, exploratory data analysis, and the tidyverse package will be used to wrangle and clean the data. Some data wrangling techniques that will be used are the imputation of missing/NA data values and conversion of categorical variables to numeric variables using one-hot encoding.

For classification, models like Logistic Regression, Random Forests, K-Nearest Neighbours and Neural Networks will be used. The initial challenges might include data wrangling and feature selection. Since there are 20 variables to fit the models and predict the outcome of the survey, it would be a challenge to select only those features which have a significant impact on the response variable. To do this, feature engineering will be carried out to create new features from existing ones.
After performing exploratory data analysis, it was found that social and economic context attributes affect the outcome the most as seen in Figure 1 and Figure 2. The bank client data and data related to the last campaign had little to no effect on the outcome. This will be used as a basis to start building machine learning models.

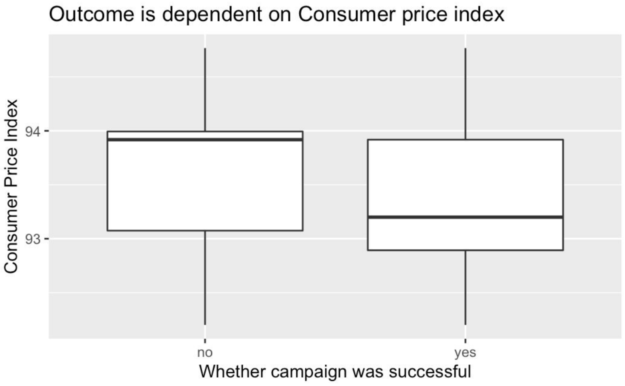

Figure 1. Distribution of Consumer Price Index

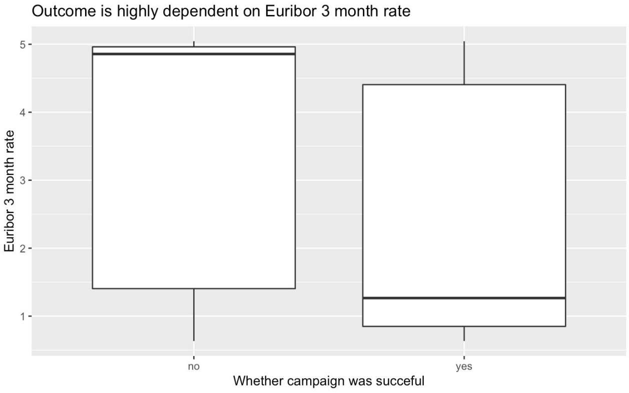

Figure 2. Distribution of Euribor 3 Month Rate

References

Moro, Sérgio, et al. “A Data-Driven Approach to Predict the Success of Bank Telemarketing.” *Decision Support Systems*, vol. 62, 2014, pp. 22–31, doi:10.1016/j.dss.2014.03.001.