BIOS 735 Project Proposal
Predicting the Success of Bank Telemarketing Using Client Information

Shuang Du, Michael Jetsupphasuk, Camille Liu and Kai Xia

BIOS 735 Team 6 "Robotcall"

Keywords: Telemarketing · predictive modeling · classification

1 Introduction

Telemarketing is a method of marketing directly to customers via phone or the Internet in order to sell goods or services. Telemarketing has increasingly been used in industrial sales organizations such as banks and insurance companies. Implementing this method successfully is difficult [1]. To improve the performance in telemarketing, predictive analytics could play an essential role in the success of telemarketing operations [2]. A data driven approach could be used in a decision support system to maximize the success rate of sales using key metrics from potential customers. A well-trained predictive model could potentially classify telemarketing targets using demographic information such as education, income, or marriage status. By training models on historical data including demographic information and success labels, a predictive classification model can be constructed using either parametric methods such as logistic regression or non-parametric methods such as random forests (RF) [3] or support vector machines (SVM) [4]. In this proposal, we will evaluate the performance of both parametric and non-parametric statistical learning methods in predicting the success of telemarketing from bank data. Our primary outcome of interest is whether a client would agree to subscribe a term deposit after the targeted telemarketing campaign.

2 Data Description

This study will use real telemarketing data from a Portuguese retail bank, from May 2008 to June 2013. The dataset was provided by the UCI Machine Learning Repository as a study example for classification problems. The original research study using this dataset was described in [2]. The dataset includes a total of 45,211 instances of phone calls. There are 16 features, or demographic variables, that include both numeric and categorical data types. Missing data is treated as its own category and called "unknown". Detailed information for each variable is shown in Table 1.

Table 1. Description of variables in telemarketing dataset

Feature Name	Description
age	numeric
job	type of job (categorical)
marital	marital status (categorical: "married","divorced","single";)
education	(categorical: "unknown","secondary","primary","tertiary")
default	has credit in default? (binary: "yes","no")
balance	average yearly balance, in euros (numeric)
housing	has housing loan? (binary: "yes","no")
loan	has personal loan? (binary: "yes","no")
contact	contact communication type (categorical: "unknown","telephone","cellular")
day	last contact day of the month (numeric)
month	last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
duration	last contact duration, in seconds (numeric)
campaign	number of contacts performed during this campaign and for this client (numeric)
pdays	number of days that passed by after last contacted from a previous campaign (numeric)
previous	number of contacts performed before this campaign and for this client (numeric)
poutcome	outcome of previous campaign (categorical: "unknown","other","failure","success")
outcome	has the client subscribed a term deposit? (binary: "yes","no")

3 Study Aims

The primary aim of this project is to build predictive statistical and machine learning models to predict whether a client would agree to subscribe a term deposit of future telemarketing campaigns using historical data. By using the statistical models and techniques we learned from module 2 of this course, we will construct the parametric logistic regression model by optimizing through maximum likelihood estimation (MLE). We will then build machine learning models and compare the results of performance in terms of prediction accuracy to the logistic regression. The secondary aim is the build a GitHub repository which will help improve our skills in algorithm implementation and software development.

4 Methods

4.1 Logistic regression

We will use the logistic regression modelling framework. We have $i = 1, 2, ..., n$ subjects. We observe the outcome Y_i for subject i, and we observe the subject's covariates x_i.

Assume $y_i|x_i \sim Binomial(1, p_i)$ for $i = 1, ..., n$; $logit(p_i) = log(\frac{p_i}{1-p_i}) = x_i^T\beta = \eta_i$ for $i = 1, ..., n$. Thus, the log-likelihood function of β for the logistic regression model is given by

$$l_n(\beta) = \sum_{i=1}^{n} \left\{ y_i x_i^T \beta - log(1 + exp(x_i^T \beta)) \right\} \tag{1}$$

We will use Newton-Raphson algorithm to obtain the ML estimate $\hat{\beta}$ in this GLM model by

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} - l''(\beta^{(t+1)})^{-1} l'(\beta^{(t)}) \\ &= \beta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (Y - p^{(t)}) \\ &= \beta^{(t)} + h^{(t)}\end{aligned} \tag{2}$$

where $p^{(t)}$ depends on $\beta^{(t)}$ and W is a diagonal matrix with the ith diagonal element having value equal to $p_i(1-p_i)$. $W^{(t)}$ is a function of $p^{(t)}$.

Next, we would like to perform variable selection using penalized lasso logistic regression [5]. The objective function is the penalized negative binomial log-likelihood, and is

$$-\left[\frac{1}{N}\sum_{i=1}^{n}\left\{y_i x_i^T \beta - log(1 + exp(x_i^T \beta))\right\}\right] + \lambda(||\beta||_1) \tag{3}$$

We will use Iteratively Reweighted Least Squares (IRLS) [6] to perform optimization, and select the tuning parameter using AIC/BIC.

4.2 Support Vector Machines

Support vector machines (SVM) are a popular machine learning method for binary classification. SVM classifies observations by separating the feature space into two classes with the goal of maximizing the number of correct classifications while allowing for some mis-classifications for identifiability and to prevent overfitting. The form of the decision boundary is controlled by choice of the kernel function. We will fit SVM models using linear and radial kernels.

Formally, SVM solves the following Lagrangian dual

$$\begin{aligned}&\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ &\text{s.t. } 0 \le \alpha_i \le C, i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0\end{aligned} \tag{4}$$

where $y \in \{-1, 1\}$ are the classification labels, α are the Lagrangian multipliers, C is a tuning parameter chosen by cross-validation, and $K(x, x')$ represents the kernel function. The linear kernel corresponds to the dot product between x and x', and the radial kernel corresponds to $K(x, x') = \exp\left(\gamma||x - x'||^2\right)$ where γ is a tuning parameter that will be chosen by cross-validation.

We will fit SVM using the R package `e1071`.

4.3 Model evaluation

To avoid overfitting and select the best model between logistic regression and SVM, we will hold out a test set by randomly sampling about 10 percent of the data. To select the best model within each class, we will split the remaining 90 percent of the data into training and validation sets in a 9 to 1 ratio (i.e. perform 10-fold cross-validation).

For model performance evaluation metrics, we will compute the Confusion Matrix and the F1 score, and calculate the area under the curve (AUC) for the ROC curve to evaluate the classification performance of the models.

4.4 Missing Data

The only missingness in the dataset exists in the categorical features so we will treat missingness as its own category, called "unknown." There are no missing observations in the numeric features or the outcome variable.

5 Analysis Plan

The analysis part will be divided as follows: 1. Data preparation 2. Feature selection 3. Model Fitting and Prediction and 4. Conclusions.

For data preparation: Based on the original data set collected from UCI Machine Learning repository, we will first conduct a quality check on the data set and remove noisy data if necessary. An exploratory data analysis will be done to help get a better overview of the whole data set and to facilitate the following featuring engineering process.

For feature selection: Feature Selection is one of the main components of feature engineering. It is the process of selecting the most important features to input in machine learning algorithms. We use this technique to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model. We plan to use lasso regularization for variable selection in the logistic model and SVM. We will also discuss on the business impact of feature selection in this model.

For model fitting and prediction: We will apply the parametric logistic regression model and other machine learning models such as SVM on the data set. The model performance on the test set would be an important criteria for selecting the optimal model for the further predictions.

For the conclusion: Based on the previous model performance, we will evaluate the accuracy of both parametric statistical learning and non-parametric machine learning methods in predicting the success of telemarketing from bank data. We will also discuss and explain the performance differences among the models we have applied in this project.

References

[1] Judith J Marshall and Harrie Vredenburg. Successfully using telemarketing in industrial sales. *Industrial Marketing Management*, 17(1):15–22, 1988.

[2] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[3] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[6] Paul W Holland and Roy E Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827, 1977.