**Bank Direct Marketing Campaign Prediction in Machine Learning & Causal Inferences**

**Qianhui Li**

**Introduction**

Nowadays, marketing has become an integral part of banks to promote their goods and services. However, banks always have limited visibility into their customers, which confines their ability to get the best out of the marketing business. The bank is an exemplary sector in which the promotional activity is highly competitive. The 2007-2009 global economic crisis increased competition among banks for deposit retention and marketing campaigns, due to credit restrictions pressure on banks (Moro, Cortez & Rita, 2012). Therefore, there is a need for efficient marketing campaigns with lesser contacts but keep the number of clients subscribing the deposit to some extent. Additionally, it is essential to identify the deterministic factor that has causal effect on customers’ responses.

There are two common approaches to promotions, which are mass marketing and direct marketing. In mass marketing, banks do not need to build direct relationships with customers, instead, they broadcast their promotional message through television, radio, and newspapers. Due to the high ineffectiveness in mass marketing, banks are shifting to direct marketing. Direct marketing selects targeted customers and focuses more on customers’ specific needs for specific product and service offers (Elsalamony, 2014).

**Literature Review**

The machine learning method can be applied to direct marketing, which utilizes customers’ historical purchasing data and predictive models to measure whether a customer will respond to an offer or not (Sing’oei & Wang, 2013). It compensates the increasing costs in marketing promotion and decreasing customer response rates. Besides, although direct marketing such as telemarketing is an interactive and powerful tool, it annoys customers sometimes (Vajiramedhin & Suebsing, 2014). The machine learning prediction can eliminate this problem by extracting knowledge from raw data and predicting customers’ responses more accurately.

The machine learning methods have rising popularity for prediction in bank direct marketing. Morgan & Sonquist (1963) first introduced classification trees in their work, and it gained popularities in marketing analytics. Grzonka, Suchacka, & Borowik (2016) reviewed tree-based classification methods, and the best predictive result was obtained from random forests. Also, Miguéis et al. (2017) used random forests to predict customers’ response to direct marketing campaigns. Besides, Artificial Neural Networks (ANNs) is wildly accepted in prediction and classifications. The previous study showed that neural networks has its advantage of allowing identifying links among factors, and not be based on “a priori” assumption (Bishop, 2005). In the research conducted by Ali & Özgür (2013), they compared the analytical results between ANN and logistic regression (LR), and it showed those two algorithms achieved the identical accuracy, but ANN ran faster than LR. Furthermore, Moro, Laureano & Cortez (2012) suggested in their study that the Support Vector Machine can achieve high predictive performances compared with Naïve Bayes and Decision Trees.

**Objective**

In my analysis, I decided to make use of Logistic Regression (LR), Lasso, Decision Tree (DT), and Neural Network (NN) to predict customers’ responses. I choose Lasso and NN, because they were seldom used in previous studies, and I would like to try something new. For LR and DT, they are popular in previous researches, and I would like to compare them with Lasso and NN. I proposed to compare the accuracy of those techniques and find the best model for predicting customers’ responses. Moreover, I would like to explore the causal inference and identify the deterministic covariant that has a causal effect on customers’ decision.

**Data**

My proposed methods performance were assessed by using the real data from the UCI Machine Learning Repository (Moro et al., 2014). The dataset was obtained from a Portuguese banking institution from May 2008 to November 2010, and the marketing campaigns were based on phone calls. It was often the case that more than one contact to the same client was required to obtain whether the bank term deposit would be (‘yes’) or not (‘no’) subscribed. The detailed descriptions about variable meanings were clearly identified, thus there were no missing values. The dataset involves 41,188 phone contacts in total with 20 input variables and 1 output variable, which will be listed in Table 1. There are two types of input variables, which are numerical and categorical, and details are listed below. The classification goal is to predict if the customer will or not (yes=1/no=0) subscribe to the term deposit (response variable).

**Table 1. Variable Descriptions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number** | **Variable Name** | **Description** | **Type** |
| 1 | Age | Age of the customer | numeric |
| 2 | Job | Type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') | categorical |
| 3 | Marital | Marital status ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) | categorical |
| 4 | Education | Education status('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown') | categorical |
| 5 | Default | Has credit in default? ('no','yes','unknown') | categorical |
| 6 | Housing | Has housing loan? ('no','yes','unknown') | categorical |
| 7 | Loan | Has personal loan? ('no','yes','unknown') | categorical |
| 8 | Contact | Contact communication type ('cellular','telephone') | categorical |
| 9 | Month | Last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec') | categorical |
| 10 | Day\_of\_week | Last contact day of the week (categorical: 'mon','tue','wed','thu','fri') | categorical |
| 11 | Duration | Last contact duration, in seconds. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. | numeric |
| 12 | Campaign | Number of contacts performed during this campaign and for this client (includes last contact) | numeric |
| 13 | pdays | Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted) | numeric |
| 14 | Previous | Number of contacts performed before this campaign and for this client | numeric |
| 15 | poutcome | Outcome of the previous marketing campaign ('failure','nonexistent','success') | categorical |
| 16 | emp.var.rate | Employment variation rate - quarterly indicator | numeric |
| 17 | cons.price.idx | Consumer price index - monthly indicator | numeric |
| 18 | cons.conf.idx | Consumer confidence index - monthly indicator | numeric |
| 19 | euribor3m | euribor 3 month rate - daily indicator | numeric |
| 20 | nr.employed | Number of employees - quarterly indicator | numeric |
| 21 | y | Has the client subscribed a term deposit? | (binary: 'yes','no') |

**References(Haven’t change style, incomplete…)**

Moro, Sérgio & Cortez, Paulo & Rita, Paulo. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. 62. 10.1016/j.dss.2014.03.001.

Morgan J. N., Sonquist J. A. (1963) Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association, Vol. 58, Issue 302, 415-434.

Elsalamony, H. (2013). Bank Direct Marketing Analysis of Data Mining Techniques. International Journal of Computer Applications (0975 – 8887), Volume 85, No 7

C. M. Bishop, Neural networks for pattern recognition, Oxford University press, New York, 2005.

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, *Elsevier, 62*, 22-31, June 2014

Sing’oei L. & Wang J. (2013). Data Mining Framework for Direct Marketing: A Case Study of Bank Marketing. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013

Vajiramedhin C. & Suebsing A. (2014). Feature Selection with Data Balancing for Prediction of Bank Telemarketing. Applied Mathematical Sciences, Vol. 8, 2014, no. 114, 5667 – 5672. Retrived from <http://dx.doi.org/10.12988/ams.2014.47222>

Grzonka, D., Suchacka, G. & Borowik, B. (2016). Application of Selected Supervised Classification Methods to Bank Marketing Campaign. Information Systems in Management. 5. 36-48.

Miguéis, V.L. & Camanho, A.s & Borges, José. (2017). Predicting direct marketing response in banking: comparison of class imbalance methods. Service Business, 11:831–849, DOI 10.1007/s11628-016-0332-3

Koç, A.A. & Yeniay, Özgür. (2013). A Comparative Study of Artificial Neural Networks and Logistic Regression for Classification of Marketing Campaign Results. Mathematical and Computational Applications, Vol. 18, No. 3, pp. 392-398

Moro, S., Laureano, R., & Cortez, P. (2012). Enhancing Bank Direct Marketing through Data Mining. 41th EMAC Conference Marketing to Citizens Going beyond Customers and Consumers. ISBN: 978-989-732-004-0