International Journal on Data Science and Technology
2018; 4(1): 35-41
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20180401.15
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Using Logistic Regression Model to Predict the Success of Bank Telemarketing

Yiyan Jiang

School of Data Science, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:
elaine97yy@yahoo.com

To cite this article:
Yiyan Jiang. Using Logistic Regression Model to Predict the Success of Bank Telemarketing. *International Journal on Data Science and Technology*. Vol. 4, No. 1, 2018, pp. 35-41. doi: 10.11648/j.ijdst.20180401.15

Received: April 19, 2018; **Accepted**: May 17, 2018; **Published**: June 21, 2018

Abstract: Term deposit is always an essential business of a bank and a good market campaign plays an essential role in financial selling. Nowadays, the telephone marketing, which can assist consulting institution to extract potential clients, has been one of the most general marketing campaigns. Previous research shows that data mining has gradually stood out on the era of Big Data and has been incorporated to deal with massive data precisely. The purpose of this study is to predict the success of bank telemarketing to select the best consumer set. A relationship is observed between success and other factors through constructing logistic regression model. To validate the effectiveness of prediction, some basic classification models have been compared in this study, including Bayes, Support Vector Machine, Neural Network and Decision Tree. As a result, the prediction accuracy and the area under ROC curve prove the logistic regression model performs best in classifying than other models. All of the experiments are implemented by R language software. And the experimental results can provide some suggestions and instructions towards the management of the bank.

Keywords: Term Deposit, Data Mining, Prediction, Logistic Regression Model, R Language

1. Introduction

Strengthening cost control and improving working efficiency is the basic ways for a company to realize profit maximization and obtain the ideal economic benefit in market operation. Undoubtedly, good strategy is the key to long-term economic operation. Nowadays, commercial banks implement telemarketing campaign to optimize the allocation of resources, satisfy the needs of customers thus enhancing the productivity of companies. Through a marketing campaign about contacting clients on telephone directly, the bank intends to select the best set of clients. It is beneficial for narrowing the range of potential customers, elevating the rate of success as well as reducing the cost of the marketing process efficiently.

In this study, a logistic regression model (LR) [1-3] is used to find the related factors of successful deposit subscription and predict the success rate of bank telemarketing. However, the raw data in the study is unbalanced, so this problem is handled in the means of random sample. Moreover, the information of clients is available, which is more likely to increase the likelihood of experiment prediction, finding target clients and realizing the goal of profit maximization. To validate accuracy of prediction, the comparison between logistic regression model and other four data mining models, i.e. Bayes [4-6], Support Vector Machine (SVM) [7-9], Neural Network (NN) [10-12] and Decision Tree (DT) [13-15] is carried out at the end of study. And all the processes of models are implemented in R language.

The remainder of this paper is organized as follows. Section 2 gives some examples of the related work about data mining. In section 3, the dataset is described and how to process the raw data and establish the model are presented as well. After the experiment, the results are analyzed in section 4. And in the last section, the conclusion is drawn and some suggestions are raised about achieving certain purpose.

2. Related Work

With the development of the internet and modern information technology, the popularity of big data has increased greatly, indicating that the ability to extract and analyze such great amounts of data is essential in the information age. That is why researchers have paid more attention to data mining and shown great interest in many different fields [16]. Nowadays, the big data are available and some potentially useful and related information can be drawn to attain objective for decision making and prediction.

With respect to marketing campaign, to improve the success of telemarketing, data mining helps to construct various models to solve the problems about bank direct marketing. Moro et al. [17] proposed a data mining approach to analyze the probability of success. For the sake of extracting the key information, feature selection was emphasized by employing NN. Moro et al. [18] also came up with a concept of lifetime value (LTV), including recency, frequency and monetary value to improve the return and investment about bank marketing. The study aimed to extract additional knowledge, according to the benefit from past contacts history. By using neural networks, they found that the presence of the historical data could improve the accuracy for targeting the deposit subscribers. Moreover, two highly relevant LTV features were concluded, which were useful for bank contact center to improve performance of telemarketing.

The logistic regression model is used in this study. It was mainly focused on the relationship between the success of telemarketing (output variable) and some positive attributes. Besides, in this paper, based on the established model, the prediction is implemented to find some potential clients who will really subscribe the deposits. In this case, the work efficiency and the performance of bank will be both enhanced greatly.

Logistic regression model is often utilized to model the probability of an experimental unit that falls into a particular class based on information measured on the experimental unit. In general, logistic regression model is a usual statistical model for discriminant analysis and classification. It is widely applied in various fields, including marketing management [19], medical fields [20], engineering [21] and so on. And in the end of study, the logistic regression model was compared with other four classification models for data mining, such as Bayes, SVM, NN and DT, to validate the efficiency and effectiveness of the model.

3. Data Pre-Processing and Modeling

In this section, overall experiment processes are introduced in detail. The model framework is elucidated in Figure 1 below, which displays two main parts of experiment: data pre-processing and modeling. First of all, the raw data are described as follow.

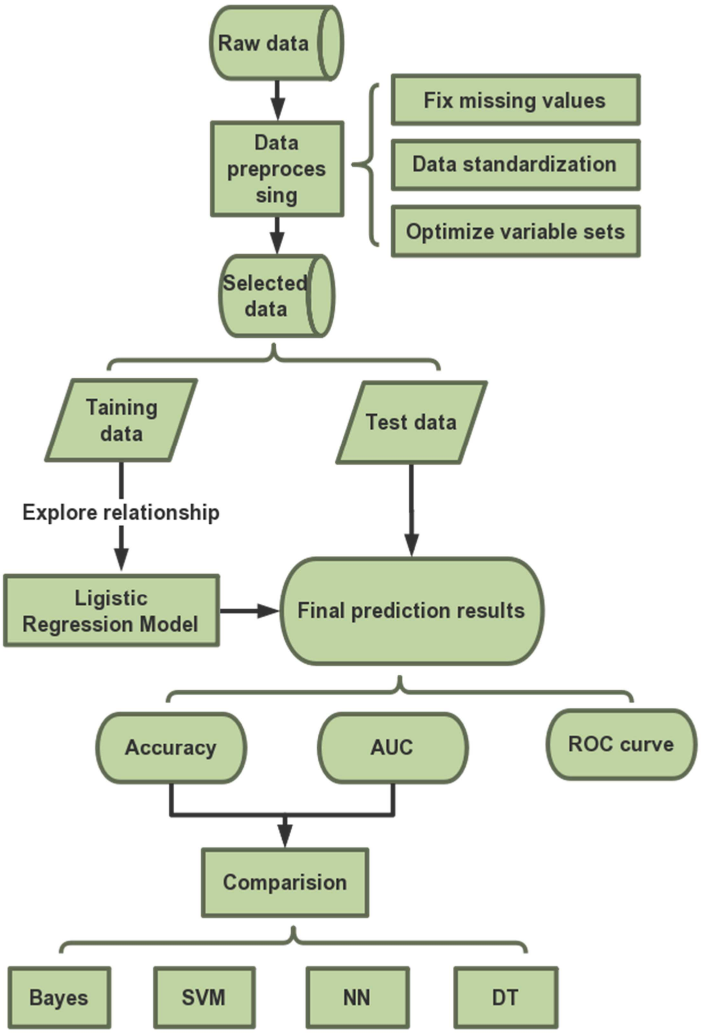

Figure 1. Model framework.

3.1. Data Description

The purpose of this paper is to detect the relationship between the success of bank telemarketing and the rest of attributes, including bank client data, social and economic context attributes, etc., as well as to predict the success radio of bank telemarketing. The databases can be downloaded from the study of Moro et al [17]. The data, which are collected from a Portuguese retail bank, are related to the success of bank telemarketing. In this data set, the same client was more than one contacted to the bank, in order to access if the term deposit would be subscribed. There are 4119 instances and 21 attributes in this data set. Except the output target y (subscription), there are 20 attributes, containing age, job, marital, education, default, etc.

The detailed descriptions about attributes are presented in Table 1 [17].

Table 1. *The pre-processed variables.*

Attribute	Description (Domain)
Bank client data (Input variables)	
age	client's age (numeric: from 18 to 88)
job	client's job (nominal: admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed or unknown)
marital	client's marital status (nominal: divorced, married, single, unknown)
education	client's education (nominal: basic. 4y, basic. 6y, basic. 9y, high.school, illiterate, professional.course, university.degree, unknown)
default	credit default (nominal: no, yes, unknown)
housing	housing loan (nominal: no, yes, unknown)
loan	personal loan (nominal: no, yes, unknown)
Contacting attributes (Input variables)	
contact	contact communication type (nominal: cellular, telephone)
month	last contact month of year (nominal: jan, feb, mar, apr, may, jun, jul, aug, sep, nov, dec)
day	last contact day of the week (nominal: mon, tue, wed, thu, fri)
duration	last contact duration in seconds (numeric: from 0 to 3643)
Other attributes (Input variables)	
campaignn	number of contacts performed during this campaign and for this client (numeric: from 1 to 35)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric: from 0 to 999)
previous	number of contacts performed before this campaign and for this client (numeric: from 0 to 6)
poutcome	outcome of the previous marketing campaign (nominal: failure, nonexistent, success)
Social and economic attributes (Input variables)	
emp.var.rate	employment variation rate (numeric: from -3.4 to 1.4)
cons.price.idx	consumer price index (numeric: from 92.2 to 94.77)
cons.conf.idx	consumer confidence index (numeric: from -50.8 to -26.9)
euribor3m	euribor 3-month rate (numeric: from 0.635 to 5.045)
nr.employed	number of employees (numeric: from 4964 to 5228)
Desired target (output variables)	
y	succeed to subscribe the term deposit (binary: yes or no)

To extract specific information of clients, we depict some charts to make descriptive statistics, which demonstrates the distribution and other statistical issues visually. And the descriptions are as follow:

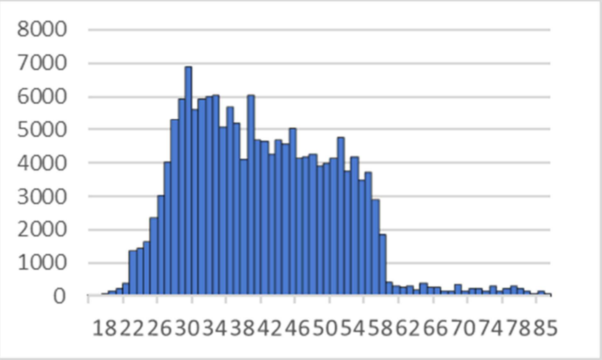

Figure 2. *The distribution of clients' age.*

The distribution of clients' age is displayed in Figure 2. It is obvious that the distribution of clients' age is roughly acknowledged as skewness distribution, which is skewed to the right. Besides, the consulted clients aged between 30 and 40 are in vast majority. It is concluded that the main target of counseling is middle-aged people.

The distribution of the clients' jobs is shown in Figure 3. As can be seen from this pie chart, the bank consults order of time deposit mainly with administrators, blue-collar workers and technicians. Among them, administrators account for 24% of total clients. The second is blue-collar workers, who accounted for 21% of the total, only four percent more than the number of technicians.

As for education, the proportion of which is illustrated in the Figure 4. It's clear from the chart that this figure is divided into five categories. The tertiary accounts for the highest proportion (44%) of the total figure, while the secondary constitutes 36% of all clients. The number of clients who have received primary education ranks in the second place, reaching 657 and making up 16%. By contrast, only 4% of the respondents' education background is unknown. However, there is still one respondent who is totally unlettered.

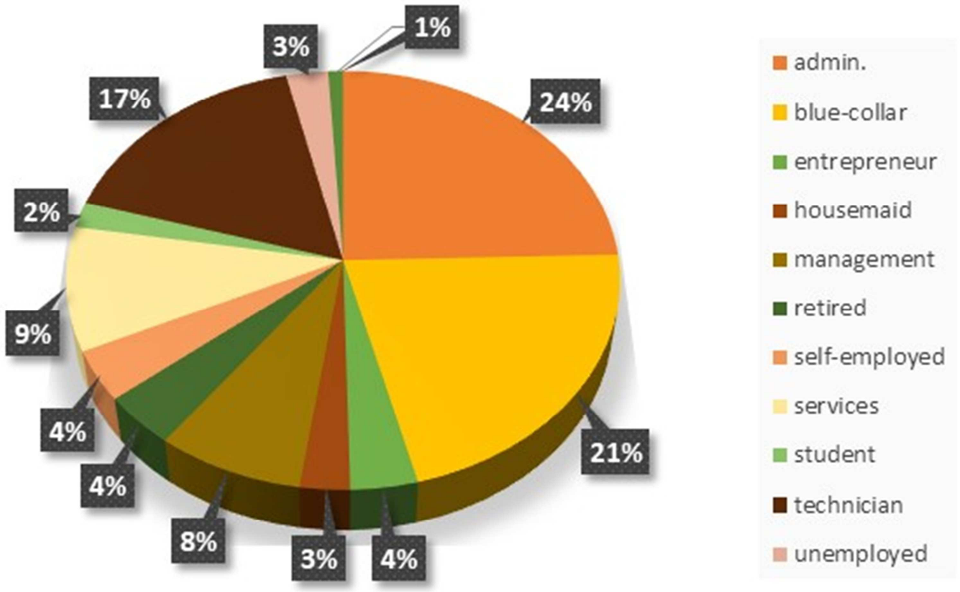

Figure 3. *The proportion of clients' jobs.*

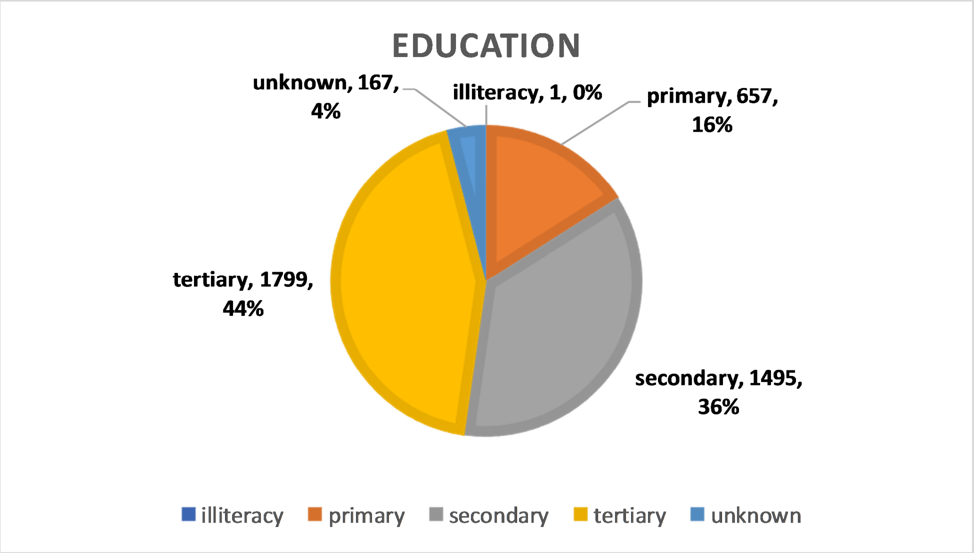

Figure 4. *The proportion of clients' education.*

After some adjustments of certain data levels, the Table 2 shows the basic statistical features of all the numeric data. What are presented in the Table 2 are gained by R language.

Table 2. *The basic statistical features of the data.*

Attributes	Mean	Std	Min	25%	50%	75%	max
age	40.11	10.31	18	32	38	47	88
duration	257	255	0	103	181	317	3643
previous	0	0.5	0	0	0	0	6
emp.var.rate	0.1	1.563	-3.4	-1.8	1.1	1.4	1.4
cons.price.idx	93.58	0.579	92.2	93.1	93.8	94	94.8
cons.conf.idx	-40.5	4.595	-50.8	-42.7	-41.8	-36.4	-26.9
euribor3m	3.621	1.734	0.635	1.334	4.857	4.961	5.045
nr.employed	5166	73.7	4964	5099	5191	5228	5228

From Table 2, it can be seen that “duration” and “nr.employed” have the highest standard deviation, which indicates the last contact duration and quarterly number of employees are the key points that need be focused on in this study. In order to avoid the error caused by different scales of numeric data, the numeric data need be standardized. This can be done by converting raw data to a dimensionless one.

3.2. Data Preprocessing

To improve the predictive effect of our proposed model, the raw data, which are often redundant, uncertain or inconsistent in general data mining field, are processed and optimized in this section. Therefore, it is crucial to preprocess the data before develop the predictive model. The following steps have been done to achieve optimization.

3.2.1. Fixing Missing Values

There are several missing values existing in the dataset about clients’ information, like the categorical data including default of credit, jobs, and marital status. About 1012 of the 4119 records contain the missing values, accounting for a quarter of the total. In this case, these records are still kept and are named “unknown”, which may decrease the risk of small sample and little information.

3.2.2. Normalizing the Data

Due to the different qualities of the indicators, standardized processing generally plays a crucial role in transforming raw data into dimensionless index, that is, each index value is at the same scale level.

3.2.3. Optimal Variable Sets

First, the level of some categorical variables is reclassified. For example, the level of education variable is reclassified from 9 levels to 5 levels, named illiteracy, primary, secondary, tertiary and unknown respectively.

Correlation coefficient is a statistical index reflecting the degree of correlation between variables. Therefore, for validating the feasibility of the 20 attributes to predict the success of bank telemarketing, the correlation coefficients between output variable y (the success of subscription) and 20 input variables are acquired through the R language. six irrelative variables with correlation coefficients less than 0.05 are excluded and a model interpreted in an easy way can be obtained. The correlation coefficient values of the remaining 14 independent variables are exhibited in Table 3.

Table 3. *The correlation coefficients between y and 14 relative attributes.*

Attribute	Correlation coefficient
duration	0.4186
nr.employed	-0.3492
pdays	-0.3282
euribor3m	-0.2986
emp.var.rate	-0.2832
previous	0.2557
contact	-0.1374
poutcome	0.1234
cons.price.idx	-0.0983
default	-0.0766
campaign	-0.0761
education	0.0642
age	0.0604
cons.conf.idx	0.0544

3.3. Modeling

The logistic regression model proposed in this study is implemented through the R language. This classification model focuses on predicting the target through telemarketing to sell the term deposits.

Bank marketing data set is divided into two parts, 80 percent of which is for training and 20 percent of which is for testing. The training data is used to model a fitted and logical model, the function of which is to discover potential predictors. As for testing data, it is utilized to calculate the accuracy of the model prediction, which is able to show the efficiency and effectiveness of the model. In this paper, five widely-used classification models are employed, such as Bayes, LR, SVM, NN and DT.

With these models, the author can explore the relationship between selected data sets and predict the target. Comparing accuracies of model predictions and calculating the value of area under ROC curve (AUC), the discrimination effects of the models mentioned above can be measured. And the depiction of ROC curve is displayed simultaneously. All the results indicate that LR has better performance than other classification models eventually.

The specific experimental results are shown in the following section, most of which are obtained by executing R language.

4. Experimental Results

4.1. Correlation Analysis

By running code in the R language software, we exclude 4 predictors with weak correlation and show the correlation coefficients between desired target and remaining predictors in Table 3, the results of which are sorted in descending order.

Given that all the above correlation coefficients are greater than 0.05, these high related variables can be utilized in the subsequent experiment about modeling and prediction. It is obvious that three attributes of the maximum correlation coefficients are “duration”, “nr.employed” and “pdays”. In other words, last contact duration, quarterly number of employees and the amount of days that passed by after the last contacting have the most significant effects on subscription. It seems that last contact duration has a positive impact on the success of subscription, yet the numbers respectively about employees and passed days are negatively correlated to the target variable.

4.2. Prediction Analysis

The experimental results of prediction are displayed in the following table. A confusion matrix is plotted to demonstrate the match between the predicted value and the real value. In this study, successfully subscribing the time posits is regarded as interesting category called "Positive", while the others (fail to subscribe) is considered as "Negative". As results shown in Table 4, the sensitivity of LR is 44.05%, which means the clients of such proportion, whom are convinced successfully to subscribe the term deposits, are predicted as "1" (Success). And the result of specificity indicates 97.45% of clients who refuse to subscribe the deposit are predicted as "0" (Failure). The overall accuracy of this classification model proposed in this study is 92.03%, which indicates that the logistic regression model performs well in predicting the success of bank telemarketing.

Table 4. *The prediction result of the logistic regression model.*

		Predicted Value		Class Precision
		0 (Failure)	1 (Success)	
Real Value	no	725	19	97.45%
	yes	47	37	44.05%
	class recall	93.91%	66.07%	

The classification accuracy of the logistic regression model: 92.03%

At last, in order to validate the efficiency and effectiveness of LR, we contrast four other classification models, including Bayes, SVM, NN and DT. The results of accuracy and AUC are displayed in the Table 5.

Table 5. *The validations of prediction effect among 4 classification models.*

Classification model	Accuracy	AUC
LG	92.03%	92.31%
Bayes	86.11%	75.90%
SVM	90.70%	66.31%
NN	89.98%	90.40%
DT	90.46%	92.12%

According to the output results calculated in R language, we can find the classification accuracies of LDM, SVM, NN and DT are 89.71%, 90.29%, 91.27% respectively. To make the comparison more explicit, the results are depicted in Figure 5.

To sum up, LR outperforms other classification models referred above. To offer further insight into the accuracy of the positive value by this classifier, the ROC curve is drawn and it is analyzed qualitatively while using AUC, which is shown in Figure 6.

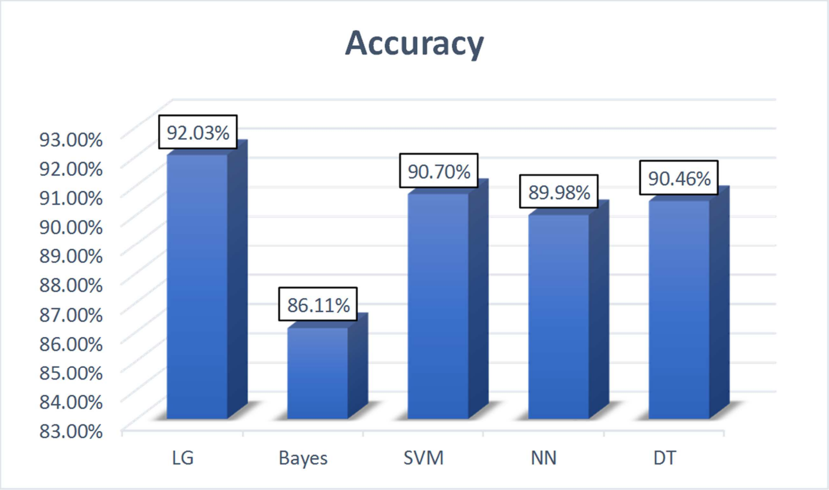

Figure 5. *The comparison on prediction accuracy among 5 models.*

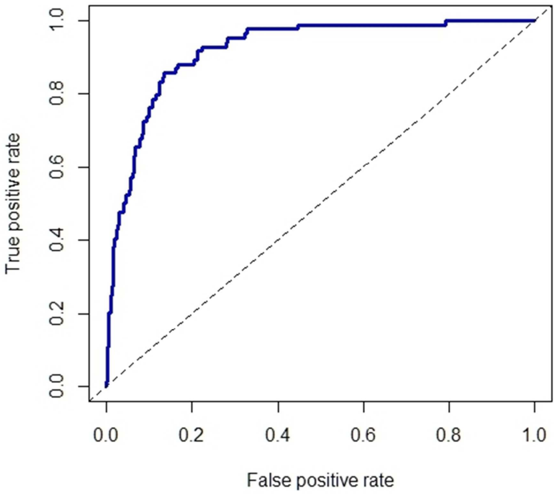

Figure 6. *ROC curve of LR.*

The ROC curve image shows the classification performance of three classifiers. The dotted line represents the classifier with no predictive value, the real curve represents the test classifier, and the blank area in the upper left corner reflects the performance of the perfect classifier.

As can be seen from Figure 6, the ROC curve occupies the upper left corner, which means the classifier (LR) used in this paper is closer to the perfect classifier. And the prediction of positive value is specific in some degree, with AUC of 92.31%.

5. Conclusions

In this study, a classification model of LR wa constructed to explore the relationship between the success of telemarketing and other attributes about clients' information, social-economic conditions and so on. Based on these highly related attributes, the success of telemarketing is predicted, whether the client accept to subscribe the term deposit or not. At the end of study, by comparing the classification accuracy and the results of AUC, it can be found that LR performs best in this study. Therefore, LR can be universally applied in our daily life, due to its strong practical significance. According to the correlation coefficients, it can be known that the last contact duration, the quarterly number of employees and the amount of days that passed by after the last contacting have great impact on the subscription of term deposit. As a whole, the bank should focus more on the duration and the frequency of contacting to enhance the communication with clients, instead of collecting information blindly. In addition, the bank with a good operation needs sufficient labor force. In this case, banks not only can improve the success of telemarketing by detecting particular clients specifically, but also can yield high returns on relatively small amounts of cost.

However, the classification model of LR proposed in this study still has some defects. For example, the amount of data is limited. The data are collected from a Portuguese retail bank, so the conclusion that is drawn just applies to this banking institution, and the model is exactly filled with restriction on fitting data and occasions. Moreover, LR is suitable for a small amount of feature space, which implies this model is generally ill-fitting with multiple variables. Last but not least, other types of classification model which have been referred or not referred in this study should be further developed. Therefore, the experiment still needs improvement, which means the author can apply a model adapting to even larger database, and detect more data mining models for prediction, improving the fitness and the accuracy in the meantime.

References

[1] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," Journal of the American Statistical Association, 1978, vol. 73, pp. 699-705.

[2] G. A. Morgan, J. J. Vaske, J. A. Gliner and R. J. Harmon, "Logistic regression and discriminant analysis: use and Interpretation," Journal of the American Academy of Child and Adolescent Psychiatry, 2003, vol. 42, pp. 994-997.

[3] S. Lisawadi, M. K. A. Shah and S. E. Ahmed, "Model selection and post estimation based on a pretest for logistic regression models," Journal of Statistical Computation and Simulation, 2016, vol. 86, pp. 3495-3511.

[4] A. E. Ades, M. Sculpher, A. Sutton and K. Abrams, et al., "Bayesian methods for evidence synthesis in cost-effectiveness analysis," Pharmacoeonomics, vol. 24, pp. 1-19.

[5] W. Dumouchel, "Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting System," American Statistician, 1999, vol. 53, pp. 177-190.

[6] K. J. Friston, V. Litvak, A. Oswal, A. Razi, K. E. Stephan, B. C. M. van Wijk, G. Ziegler and P. Zeidman, "Bayesian model reduction and empirical Bayes for group (DCM) studies," Neuroimage, 2016, vol. 128, pp. 413-431.

[7] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," IEEE Transactions on Geoscience and Remote Sensing, 2004, vol. 42, pp. 1335-1343.

[8] A. M. Andrew, "An introduction to support vector machines and other kernel‐based learning methods," Kybernetes, 2002, vol. 32, pp. 1-28.

[9] L. V. Utkin, A. I. Chekh, and Y. A. Zhuk, "Binary classification SVM-based algorithms with interval-valued training data using triangular and Epanechnikov kernels," Neural Networks, 2016, vol. 80, pp. 53-66.

[10] H. S. Hippert, C. E. Pedreira and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," IEEE Transactions on Power Systems, 2001, vol. 16, pp. 44-55.

[11] S. R. Presnell and F. E. Cohen, "Artificial neural networks for pattern recognition in biochemical sequences", Annual Review of Biophysics and Biomolecular Structure, 1993, vol. 22, pp. 283-298.

[12] L. Wang, B. Yang, Y. Chen, X. Zhang and J. Orchard, "Improving neural-network classifiers using nearest neighbor partitioning," IEEE Transactions on Neural Networks and Learning Systems, 2017, vol. 28, pp. 2255-2267.

[13] M. Núñez, "The use of background knowledge in decision tree induction," Machine Learning, 1991, vol. 6, pp. 231-250.

[14] S. K. Murthy, "Automatic construction of decision trees from data: a multi-disciplinary survey," Data Mining and Knowledge Discovery, 1998, vol. 2, pp. 345-389.

[15] R. Wang, S. Kwong, X. Z. Wang and Q. Jiang, "Segment based decision tree induction with continuous valued attributes," IEEE Transactions on Cybernetics, 2017, vol. 45, pp. 1262-1275.

[16] M. S. Chen, J. Han and P. S. Yu, "Data mining: an overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering, 2002, vol. 8, pp. 866-883.

[17] S. Moro, P. Cortez and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, 2014, vol. 62, pp. 22-31.

[18] S. Moro, P. Cortez and P. Rita, "Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns," Neural Computing and Applications, 2015, vol. 26, pp. 131-139.

[19] T. M. P. D. John, P. S. P. D. Theodore and L. E. P. D. Terry, "Supply chain management and its relationship to logistics, marketing, production, and operations management," Journal of Business Logistics, 2008, vol. 29, pp. 31–46.

[20] M. Strano and B. M. Colosimo, "Logistic regression analysis for experimental determination of forming limit diagrams," International Journal of Machine Tools and Manufacture, 2006, vol. 46, pp. 673–682.

[21] Y. Yorozu, M. Hirano, K. Oka and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Translation Journal on Magnetics in Japan, 1987, vol. 2, pp. 740–741.