

Prediction of Term Deposit in Bank: Using Logistic Model

Enjing Jiang^{1, †}, Zihao Wang^{2, *, †}, Jiaying Zhao^{3, †}

¹College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou, China

²Tandon School of engineering, New York University, New York, U.S.

³College of Pittsburgh, Sichuan University, Chengdu, China

*Corresponding author: zw1580@nyu.edu

†These authors contributed equally

Abstract. Telemarketing remaining to be one of the most popular direct marketing techniques, which is urged to make accurate predictions. It is critical for banks to have a mature machine learning approach for marketing forecasting problems. Through the analysis of the time deposit telemarketing data of Portuguese commercial banks on Kaggle.com, the bank can more accurately locate the target customers, so as to improve the efficiency of increasing the amount of time deposit business. Firstly, the data are preprocessed, coded and shuffled to divide the test set and training set. Secondly, logistic regression is carried out to eliminate irrelevant independent variables through correlation analysis, which makes the prediction more accurate. The results of logistic regression were compared with the results of decision tree and retrograde comparison. Accuracy and Area Under Curve (AUC) were used as evaluation indexes. The results demonstrate the effectiveness of the logistic regression model and the elimination of irrelevant variables. This method can be applied to the actual marketing problems of banks to accurately find target customers and improve the accuracy of marketing.

Keywords: term deposit; logistic regression; machine learning.

1. Introduction

1.1 Background

Marketing is a process which display the products by appealing and attractive ways to clients and customers to serve their requirements. There are two kinds of ways of marketing, which are direct marketing and Mass marketing [1]. Mass marketing is to market massively without any target, whereas direct marketing is to concentrate on a small group of customers with a specific goal. Telemarketing is one of a direct marketing way using the telephone, internet, or fax to potential customers [2]. Telemarketing remains to be one of the most popular direct marketing techniques because the potential effectiveness of human-to-human individual contact is sometimes quite the opposite of many impersonal and robotic marketing messages relayed through social and digital media.

Mass campaigns involve a significant amount of waste because only a very small proportion of customers that are communicated with show an interest in buying the product. However, executing direct telemarketing also requires a huge investment by the business as large call centers need to be contracted to contact clients directly [3].

When competitions increase, it is more important for the companies to keep track of customers' preference so as to win in the market. In the current business climate, customer preferences are very complicated and often change dynamically, therefore difficult to collect. Direct marketing campaigns are essential methods for enhancing the economic gain of a firm in two respects: obtaining new customers and creating additional yield from present customers. Notably, many financial services providers made use of telemarketing strategies to attract new customers, and provide better services to existing customers by satisfying their special needs. The bank is one of the most important parts of the national economic structure. It can offer loan and deposit services to customers. Term deposits are cash investments held at a financial institution and are a major source of revenue for banks.

However, there are limited research about banks' telemarketing in subscribe deposit. In this case, our study aims to predict the accuracy of the telemarketing practice of banks for selling long-term deposits.

1.2 Related research

Many researchers have developed a number of methods for term deposit [4-6]. In Hou's study, machine learning was applied to the creation of prediction models for bank deposit subscriptions using five algorithms (naive bayes, decision trees, random forests, support vector machines, and neural networks) [4]. Khan Mohd Zeeshan's study uses machine learning to suggest a suitable model to the banks in order to offer an explainable AI-based solution for predicting potential term deposit customers [5]. Dutta Shawni devised a method applying convolutional-GRU to create a system for term deposit likelihood prediction [6].

There are studies on credit management and bank default in the publications that use logistic regression. Some proposed algorithms based on logistic regression helped to provide a strong statistical background to trust prediction [7].

1.3 Objection and motivation

The algorithm of logistic regression has been more mature, and the prediction is more accurate. Numbers of questions regarding term deposits remain to be addressed. Most studies have only relied on machine learning like neural trees, and they are not conclusive because they did not using logistic regression to select attributes and predict. Therefore, using logistic model to predict term deposits in a bank is a worth studying topic, which fills in the gaps in the research.

Therefore, the objective of this paper is to predict whether the clients will subscribe to the term deposit or not and introduce the product to potential customers. It can help a bank save time and marketing costs. There are three main parts to the study. First is data preprocessing, which is to code the dummy variables and divide the data into 80% train set and 20% test set. The second part is attributes selection using a logistic model, which includes correlation analysis and irrelevant variable removal. The third part is to establish and optimize the prediction model and compare it to the decision tree model by predicting accuracy and AUC indicators.

2. Method

2.1 Data Sources

The data is downloaded from Kaggle.com and it is related to marketing campaigns of a Portuguese bank on phone. The data showed about 14 attributes about the clients' information and the results of the clients' whether would subscribe to the term deposit or not. The date of data was from May 2008 to November 2010 and the number of instances was 45211.

The attributes of clients from the data can be categorized into three parts. The first is bank client data, which includes age, job, marital, education, credit default, yearly balance, housing loan, and personal loan. The second part is the information from the last contact with the clients: contact communication type, and last contact duration. The last part is some other information from clients: the number of times the bank contacted the customers during the campaign, the number of days that passed by after the bank called the customers last time, the number of times the bank contacted the customers before the campaign, result of the last marketing campaign.

2.2 Logistic regression

Logistic regression is a generalized linear regression analysis model, which belongs to supervised learning in machine learning and is used to solve dichotomous classification problems. The model is trained with several sets of given data, and one or more sets of given data are classified after training [8]. Each set of data is made up of p indicators. The input to logistic regression is the result of linear

regression as follows function (1), among it, β_i is the regression coefficient. X_i is the independent variable.

$$\begin{cases} y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad (1)$$

Using sigmoid function (2), the result of linear regression will be mapped to [0,1]. Assuming that 0.5 is the threshold, the value of p less than 0.5 will be 0 by default, and the value of p greater than 0.5 will be 1.

$$p = \frac{e^y}{1 + e^y} \quad (2)$$

2.3 Decision trees

One of the highly interpretable models that can carry out classification and regression tasks is the decision tree. [9]. A decision tree is a tree-like structure model similar to an inverted tree. Consider a scenario where there are classical machine learning models (such as linear regression and logistic regression) to perform regression and classification tasks, it is often necessary to make sure that the data used to train the model is free of all irregularities, such as missing values, outliers to deal with, and multicollinearity. A lot of data preprocessing needs to be completed. In decision trees, however, no data preprocessing is required. The decision tree is powerful enough to handle all of these problems to make a decision. Moreover, decision trees can deal with nonlinear data that classical linear models cannot. Therefore, decision trees are diverse enough to perform regression and classification tasks. Decision trees are constructed by asking a series of questions to the data to make a decision. It can be said that the decision tree mimics the human decision-making process. During tree construction, it divides the entire data into subsets of data until a decision is reached.

2.4 ROC and AUC

Receiver Operating Characteristic (ROC) and Area under the curve (AUC) are valuable indicators used for evaluating the performance in classification models. They clearly help determine and get to know the capability of a model in differentiating different classes.

A ROC curve measures the performance of a classification model by plotting the rate of true positives against false positives. AUC is the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative one by a classifier. Graphically speaking, AUC measures the area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1) [10]. AUC of 0 indicates that the predictions of the model is 100% wrong. AUC of 1 shows that the prediction is 100% correct. 0.5 means that the model is not able to separate classes at all. [0.5, 0.7] means that the model can separate classes but not very well. [0.7, 0.85] means that the model is good, and [0.85, 0.95] indicates that the model is excellent.

3. Results and Discussion

3.1 Data Preprocessing

3.1.1 Code for Dummy Variable

Apply the 14 attributes as X variables and the result of subscription as Y variable. The Y variables are binary: "yes", and "no". Coding "yes" as 1, "No" as 0. There are 8 attribute variables that need coding dummy variables. Use SPSS to code and the outcome is shown in Table 1.

3.1.2 Divide the data sets and test sets

The division of data after shuffling data. In order to prevent overfitting, 80% of the data is divided into the training set, and the remaining data is divided into the training set.

Table1. Code the data and the outcome

X	Count	Code	X	Count	Code
admin.	5171	0	failure	4901	0
blue	9732	1	other	1840	1
entrepreneur	1487	2	success	1511	2
housemaid	1240	3	unknown	36959	3
management	9458	4	cellular	29285	0
retired	2264	5	contact telephone	2906	1
self-employed	1579	6	unknown	13020	2
services	4154	7	divorced	5207	0
student	938	8	marital married	27214	1
technician	7597	9	single	12790	2
unemployed	1303	10	housing no	20081	0
unknown	288	11	yes	25130	1
primary	6851	0	loan no	37967	0
secondary	23202	1	yes	7244	1
tertiary	13301	2	default no	44396	0
unknown	1857	3	yes	815	1

3.2 Logistic Regression and Variable Selection

In this problem, the Logistic regression model is selected due to the binary of Y variables: “Yes” and “No”. The specification is established (3)(4), and the meaning of the symbol is shown in Table 2.

$$\begin{aligned}
 y_n = \beta_0 + \sum_{i=1}^{11} \beta_i B_n(i) + \beta_{12} A_n + \beta_{13} C_n(1) + \beta_{14} C_n(2) + \beta_{15} D_n(1) + \beta_{16} D_n(2) + \beta_{17} D_n(3) \\
 + \beta_{18} E_n + \beta_{19} F_n + \beta_{20} G_n + \beta_{21} H_n + \beta_{22} I_n(1) + \beta_{23} I_n(2) + \beta_{24} J_n + \beta_{25} K_n \\
 + \beta_{26} L_n + \beta_{27} M_n + \beta_{28} N_n(1) + \beta_{29} N_n(2) + \beta_{30} N_n(3)
 \end{aligned}
 \quad (3)$$

$$p_n = \frac{e^{y_n}}{1 + e^{y_n}}
 \quad (4)$$

Table 2. The meaning of the symbol

Symbol	A _n	B _n	C _n	D _n	E _n	F _n	G _n	H _n
Attribute	age	job	marital	education	default	balance	housing	loan
Symbol	I _n	J _n	K _n	L _n	M _n	N _n	β	y _n
Attribute	contact	duration	campaign	pdays	previous	poutcome	coefficient	subscribe

Running the Logistic model on SPSS and analyze the coefficient between the attributes and the Y variable, and the results are shown in Table 3,

Table 3. Results

	B	sig.		B	sig.
age	0.002	0.47*	education(1)	0.166	0.009*
job		0***	education(2)	0.404	0***
job(1)	-0.393	0***	education(3)	0.271	0.007**
job(2)	-0.485	0***	default(1)	-0.166	0.304
job(3)	-0.527	0***	balance	0	0***
job(4)	-0.229	0.001**	housing(1)	-0.769	0***
job(5)	0.341	0***	loan(1)	-0.563	0***
job(6)	-0.382	0***	contact		0***
job(7)	-0.312	0***	contact(1)	-0.083	0.251
job(8)	0.504	0***	contact(2)	-1.156	0***
job(9)	-0.273	0***	duration	0.004	0***
job(10)	-0.219	0.044*	campaign	-0.11	0***
job(11)	-0.338	0.141	pdays	0	0.517
marital		0***	previous	0.011	0.104
marital(1)	-0.159	0.006**	poutcome		0***
marital(2)	0.161	0.014**	poutcome(1)	0.249	0.004**
education		0***	poutcome(2)	2.298	0***
education(1)	0.166	0.009*	poutcome(3)	0.242	0.007**
education(2)	0.404	0***	Constant	2.572	0***

Generally speaking, $p < 0.05$ proves to be significant. The age, default, pdays, previous have the p value > 0.05 , which shows the 4 attributes have insignificance with Y variables. Remove the 4 insignificant attributes from the logistic model, and run the logistic model again on SPSS, the optimized logistic model has all X variables significant. The logistic model specification after optimization is function (5).

$$\begin{aligned}
 y_n = & -2.424 - 0.396B_n(1) - 0.489B_n(2) - 0.524B_n(3) - 0.228B_n(4) + 0.368B_n(5) \\
 & - 0.385B_n(6) - 0.314B_n(7) + 0.49B_n(8) - 0.274B_n(9) - 0.22B_n(10) \\
 & - 0.331B_n(11) - 0.161C_n(1) + 0.145C_n(2) + 0.16D_n(1) + 0.397D_n(2) \\
 & + 0.27D_n(3) - 0.768G_n - 0.568H_n - 0.072I_n(1) - 1.156I_n(2) + 0.004J_n \\
 & - 0.109K_n + 0.256N_n(1) + 2.289N_n(2) - 0.321N_n(3)
 \end{aligned}
 \tag{5}$$

3.3 Comparison between Models

3.3.1 Test the Model and Robustness Analysis

This paper tests and compares the prediction accuracy of both the logistic regression model and the decision tree model after coding by Python. In both models, the prediction is done based on the original data, and the data with irrelevant variables (age, default, previous, p days) removed. So, there are basically four scenarios that need to be tested. Four scenarios are used to evaluate the performance of both models.

- 1) Logistic regression model under complete data.
- 2) Logistic regression models when irrelevant attributes are removed.
- 3) Decision tree model under complete data.
- 4) Decision tree model when irrelevant attributes are removed. Prediction accuracy and AUC.

3.3.2 Using prediction accuracy as criterion

When using prediction accuracy as the criterion to choose prediction models and data types, the four scenarios respectively are tested, as shown in Table 4. Firstly, the logistic regression model under

complete data is tested. After running the code on python, the prediction accuracy is shown to be 0.88698. The result is higher than 0.85, indicating that the prediction accuracy is quite good. Then, the data type is changed to the reduced data (four irrelevant variables are removed), and the result shows that the value remains the same. So, prediction accuracy won't be affected by data types when using a logistic regression model to do the prediction.

The Decision tree model is tested under complete data. Although its prediction accuracy is still higher than 0.85, it is much smaller than the previous value (0.88698). So, if prediction accuracy is used to evaluate both models, the logistic regression model turns out to be the better one. When the data type changes to reduced data, the prediction accuracy decreases once again.

Table 4. Prediction Accuracy under Different Data and Two Prediction Methods

	Logistic regression	Decision tree
Complete data	0.88698	0.85868
Reduced data	0.88698	0.8537

3.3.3 Using AUC as criterion

Using AUC as a criterion and testing the four scenarios again, as shown in Table 5 and Figure 1-4. Firstly, when the logistic regression model under complete data is tested, the area under the ROC curve is 0.8524. F2.D in this paper shows that when AUC ranges from [0.85, 0.95], the model is excellent. So, using a logistic regression model to do the forecast is excellent. Then, four irrelevant attributes are removed. Running the model again, AUC increases a little. This reveals that using reduced data to do the prediction is better than predicting using complete data under the logistic model.

Next, the decision tree model is tested under complete data. The statistical result of python shows that AUC decreased to 0.67257. So, using a decision tree to predict is less accurate than the logistic regression model, which is confirmed again in the four graphs below. At last, running the model under reduced data and found that AUC was further reduced.

Based on the above results, the conclusion is reached, that if prediction accuracy is used to evaluate the models and data types, the logistic regression model is better than the decision tree model and irrelevant attributes won't affect the accuracy here. If using AUC to evaluate, the logistic regression model under reduced data is the best forecasting method.

Table 5. AUC under Different Data and Two Prediction Methods

	Logistic regression	Decision tree
Complete data	0.8524	0.67257
Reduced data	0.85588	0.66335

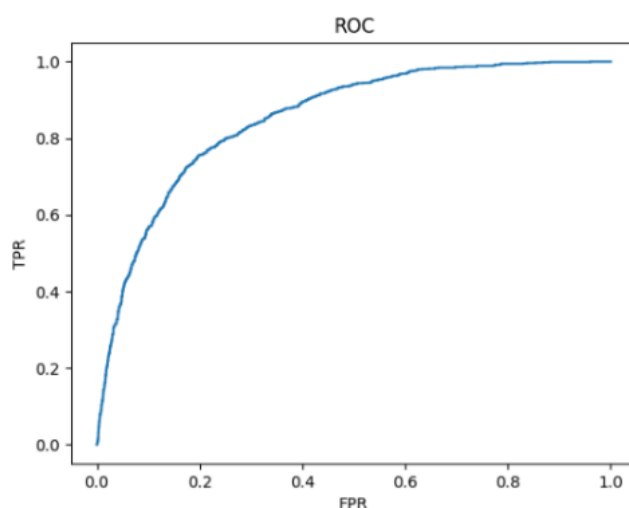


Fig.1 ROC Curve of logistic model under original data

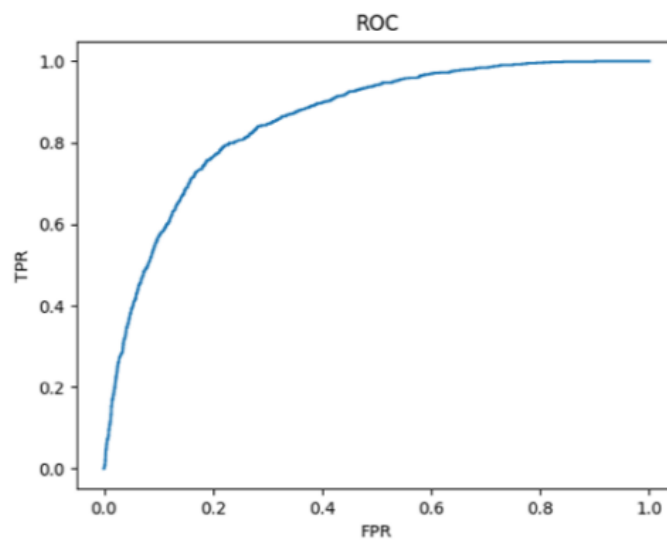


Fig.2 ROC Curve of logistic model under reduced data

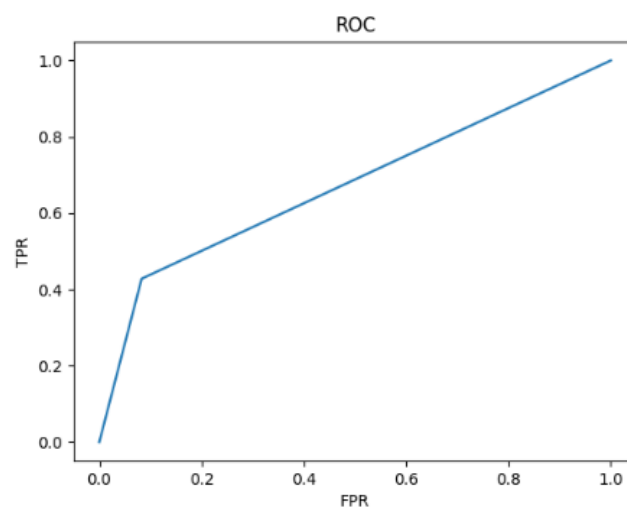


Fig.3 ROC Curve of decision tree model under original data

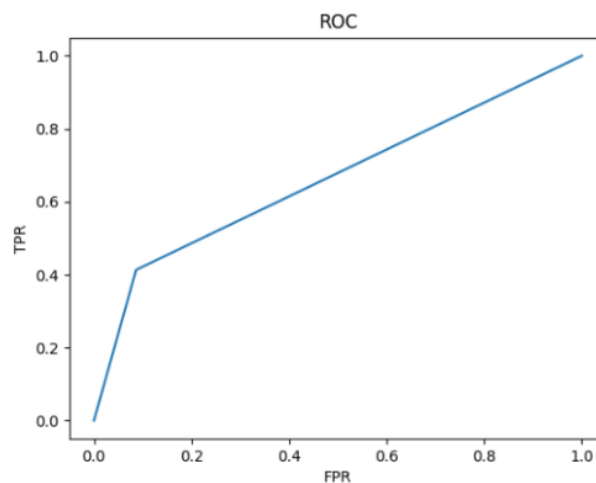


Fig.4 ROC Curve of decision tree model under reduced data

4. Conclusions

Under the trend of globalization, the traditional operation mode is not enough to bring businesses like education institutions, insurance companies, and banks competitive advantages, therefore various marketing strategies need to be used. Telemarketing is one of a direct marketing strategies in which a salesperson can learn whether the prospective customers would like to purchase products or services or not through several phone calls.

In this paper, the patterns and characteristics of customers who are more interested in the product are identified through the telemarketing method so as to explore the best strategies to improve the Portuguese bank's next direct marketing campaign. And the best prediction model by comparing prediction accuracy and AUC of alternative models is selected. Based on the results of python, the logistic regression model has higher prediction accuracy than the decision tree found. Besides, after removing the insignificant attributes, the optimized logistic model has a better performance. Therefore, the optimized logistic model is a good model for the bank to predict target customers who want to subscribe to a term deposit. Portuguese banks can use the model to identify the clients who will subscribe to the term deposit or not and can save time and money to call mass clients randomly. It is a good way for the bank to make telemarketing efficient.

However, there are still some limitations and shortages of the paper. For example, other forecasting methods other than the logistic regression model and decision tree model into account have not been taken. In the future, other classification methods can be analyzed. For example, random forest, support vector machine, and Gaussian Naive Bayes to see if they perform better than the logistic model in terms of forecasting the Portuguese bank's next direct marketing campaign. Besides, the deposit policy hasn't been considered in this paper. Since the deposit policy keeps changing and it plays an essential role in predicting the future marketing campaign, it should always update the model based on the newest policy established.

References

- [1] Daniel Hopper, "12 Ways to Use Direct Marketing", 8 March 2021, 5 September 2022,
- [2] <https://www.business2community.com/marketing>.
- [3] Will Kenton, "Telemarketing", 30 July 2022, 5 September 2022, <https://www.investopedia.com/terms/t/telemarketing.asp#>
- [4] Moro S, Cortez P, & Rita P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 2014, 62(62), 22-31.
- [5] Hou Sipu et al. Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. *International Journal of Business Analytics (IJBAN)*, 2021, 9(1): 1-12.
- [6] Khan Mohd Zeeshan and Munquad Sana and Rao Thota Sree Mallikharjuna. A Study on Improving Banking Process for Predicting Prospective Customers of Term Deposits using Explainable Machine Learning Models. *Springer Nature Singapore*, 2022: 93-103.
- [7] Shawni Dutta and Samir Kumar Bandyopadhyay. Recommender System for Term Deposit Likelihood Prediction using Cross-validated Neural Network. *South Asian Journal of Social Studies and Economics*, 2021: 21-28.
- [8] Feslin Anish Mon Solomon and Godfrey Winster Sathianesan and R. Ramesh. Logistic Regression Trust–A Trust Model for Internet-of-Things Using Regression Analysis. *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, 2023, 44(2): 1125-1142.
- [9] M. Valavan and S. Rita. Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers. *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, 2023, 45(1): 231-245.
- [10] Hamdi Monia et al. Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model. *Alexandria Engineering Journal*, 2023, 62: 327-333.
- [11] Beschorner Kurt E. et al. Validating the ability of a portable shoe-floor friction testing device, NextSTEPS, to predict human slips. *Applied Ergonomics*, 2023, 106: 103854-103854.