Prediction of Bank Fixed Deposits Based on Logistic Regression

Fei Xie
School of Accounting, The Australian National University, Canberra, 2601, Australia

Keywords: Prediction, Fixed Deposits, Logistic Regression.

Abstract: Fixed deposits, as the main source of bank funds, are also a prerequisite for the safety of bank operations and good market liquidity. Having more deposits is more conducive to the various investments of banks and the stable development of the national economy, and whether customers will participate in the fixed deposit business has become a focus of research by scholars from various countries. Logistic regression, as a typical binary discrimination method, is highly favored by scholars from various countries in terms of accuracy and interpretability. This article uses logistic regression to analyze the factors of customers themselves and the effectiveness of telemarketing and establishes a mathematical model, achieving good results with an accuracy of 89%, which has the significance of guiding customer selection. Meanwhile, this article believes that if more accurate data can be obtained and psychological data on personality, emotions, and other aspects can be added, the model's accuracy will be further improved.

1 INTRODUCTION

Deposits are a prerequisite for ensuring the safety of bank operations and good market liquidity. For banks, customer deposits are the main source of funds and the foundation for conducting other businesses. Having more deposits means that banks can have more funds to invest and lend to various types of income, which is beneficial for the banks themselves and the stable development of the national economy (Yang, 2014).

With the development of the Internet and computer hardware, banks have more and more customer information and data. To provide customers with better and more considerate services, banks can use this big data to identify customers with deposit intentions and preferences, abandon targeted marketing for some unwilling customers, and focus on targeted marketing for willing customers. This not only preserves the majority of deposit customer sources, allowing bank deposits to remain stable but also reduces the consumption of human and material resources on unintended customers, thereby improving the bank's efficiency (Deng 2023).

Telemarketing, as a product of technological development, is a two-way communication method that uses the phone to communicate with target customers. People can accurately convey information or requests to each other through the phone, which has now become an essential marketing method for major enterprises. Major commercial banks also established telephone service projects as early as the late 1990s. This year, major commercial banks have mainly divided their telephone marketing methods into two types: one is to actively contact customers for telephone marketing, and the other is to passively wait for customers to contact them before conducting telephone marketing. The first marketing method is mainly initiated by banks, promoting products based on pre-acquired customer information and further understanding customer needs; The second type is initiated by the customers themselves, such as product inquiries, confirmations, and complaints.

Telemarketing is believed to have been proposed by American scholar Juic Freestone in the 1970s. Since its development, it has become a research focus for domestic and foreign scholars to achieve precision marketing. At first, research on telemarketing focused on theory and techniques. Cain summarized regulatory legal norms related to telemarketing and regulated marketing contract issues (Cain 1996). Mann provided a detailed introduction to the specific process of telemarketing (Mann 2006). Hurst proposed nine tips for telemarketing in 2008. Subsequently, with the development of data mining technology, more and more researchers began to combine telemarketing with data mining. Accordingly, they proposed many research methods,

Xie, F.
Prediction of Bank Fixed Deposits Based on Logistic Regression.
DOI: 10.5220/0012842900004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 189-195
ISBN: 978-989-758-690-3

further improving the success rate of telemarketing. Hyeon used Bayesian network models to predict customer responses to bank telemarketing and designed a decision system to provide real-time decision support (Hurst 2008, Ahn & Ezawa 1997). Elsalamony applied logistic regression, a naive Bayesian algorithm, and a neural network model to analyze marketing effectiveness (Elsalamony 2014). Kim classified customers and used convolutional neural network models, decision tree models, and logistic regression models to predict the probability of successful customer marketing under different classifications, to find the optimal model for predicting telemarketing effectiveness (Kim et al. 2016). Jiang compared the predictive performance of various models on bank deposit data and found that the Forest model is one of the best-performing models in predicting this type of data (Jiang 2021). Liu proposed a fuzzy support vector machine (SVM) model and compared it with traditional SVM models regarding prediction performance. The results showed that the newly proposed fuzzy SVM model had better prediction performance (Liu et al. 2017). Jiang used various classification models such as Bayesian and logistic regression to predict the optimal consumer group for telemarketing and provided some suggestions for refined management and services of banks (Jiang 2018). Chun improved the unsupervised learning Kohonen network and proposed a Kohonen-supervised learning network model for telemarketing prediction (Yan et al. 2020).

2 METHODOLOGY

2.1 Data Sources

The data in this article is taken from the direct marketing activity data of bank fixed deposits on the Kaggle website, including two datasets: the training set and the test set, with 45211 pieces of data in the training set and 4521 pieces of data in the test set. This data includes data related to direct telemarketing activities carried out by Portuguese banking institutions and a collection of various customer information data. This article will perform logistic regression on the training set to determine the model, and then use the test set to test the accuracy of the model in determining whether customers will choose to engage in fixed deposit business.

2.2 Variable Selection

This study aims to predict whether customers will engage in fixed deposit business. Therefore, the dependent variable y is whether customers choose a fixed deposit business, and it is a 0-1 variable. The proportion of y in the two datasets is shown in Figure 1.

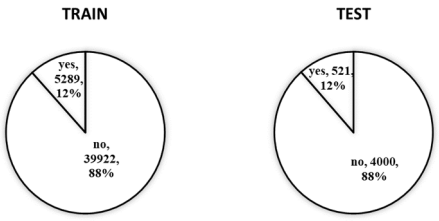

Figure 1: The proportion of fixed deposits (Picture credit: Original).

From the above figure 1, it can be seen that the majority of customers refuse to make fixed deposits, and the distribution of the proportion in the training and testing sets remains consistent. The independent variables are divided into two parts. The first part is customer-related data, including 8 items such as age, work, and marital status, as shown in Table 1.

From the below table, it can be seen that there are more middle-aged people aged 30 to 40 in terms of age distribution. In terms of year-end deposit balance in banks, there are both large deposits and large liabilities, and overall, customers have a positive year-end deposit balance. In terms of work, there are more Blue-collar and Management, both exceeding 20%, and almost all have stable sources of income. In terms of marital status, more than half of married individuals have a relatively happy overall family situation. In terms of education, there is almost no illiteracy, and more than half of the clients have reached the secondary level with a high level of education. In terms of credit, only a few customers have default records. At the same time, the number of customers with or without housing loans remains relatively stable, while the majority of customers do not have personal loans.

The second part is the telemarketing data for the customer, including the customer's contact information, the month and date of the last contact of the year, and a total of 8 items of data. The interval between the previous two marketing activities is -1, indicating that they have not been contacted before, as shown in Table 2.

Table 1: Customer Related Data.

Index	Descriptive Statistics			
	Mean	Mode	Maximum	Minimum
Age	40.94	32	95	18
Balance/€ (Average annual account balance)	1362	0	102127	-8019
Job	Admin (11.44%); Blue-collar (21.53%); Entrepreneur (3.29%); Housemaid (2.74%); Management (20.92%); Retired (5.01%); Self-employed (3.49%); Services (9.19%); Student (2.07%); Technician (16.80%); Unemployed (2.88%); Unknown (0.64%)			
Marital	Divorced (11.52%); Married (60.19%); Single (28.29%)			
Education	Primary (15.15%); Secondary (51.32%); Tertiary (29.42%); Unknown (4.11%)			
Default (If having a record of breach of contract)	No (98.18%); Yes (1.82%)			
Housing (If having a housing loan)	No (44.42%); Yes (55.58%)			
Loan (If having a personal loan)	No (83.98%); Yes (16.02%)			

Table 2: Telemarketing Data.

Index	Descriptive Statistics			
	Mean	Mode	Maximum	Minimum
Day (The last contact date of the year)	15.81	20	31	1
Duration/second (Last communication time)	258.16	124	4918	0
Campaign (The number of contacts)	2.76	1	63	1
Pdays/day (Time interval between last contact)	40.18	-1	871	-1
Previous (Accumulated contact times)	0.58	0	275	0
Contact	Cellular (64.77%); Telephone (6.43%); Unknown (28.80%)			
Month (The month of the last contact of the year)	Jan (3.10%); Feb (5.86%); Mar (1.06%); Apr (6.49%); May (30.45%); Jun (11.81%); Jul (15.25%); Aug (13.82%); Sept (1.28%); Oct (1.63%); Nov (8.78%); Dec (0.47%)			
Poutcome (The results of the last marketing campaign)	Failure (10.84%); Other (4.07%); Success (3.34%); Unknown (81.75%)			

From the table above, it can be seen that the bank's telemarketing time can reach 2 to 4 minutes, allowing for simple introductions and communication with customers. For this telemarketing client, there are many new customers. For old customers, banks usually maintain contact once a month or so to maintain customer stickiness. Overall, banks will make 2-3 contacts with each customer to ensure that marketing activities are communicated to them. In terms of contact information, most customers are accustomed to using cellular, which is more convenient and also makes it easier for the bank to contact the customers themselves. In terms of contact months, banks prefer to contact customers in May, with much more contact times than in other months. From the previous marketing results, it can be seen that the number of customer rejections to marketing activities is much higher than the number of acceptances. At the same time, the vast majority of the previous marketing results are unknown, indicating that it is highly likely that these customers have refused to contact the bank, so they will no longer handle any business with the bank in the future.

2.3 Model Selection

Because the data results in this article are only binary data for "accepting fixed deposit projects" and "not accepting fixed deposit projects", the direct linear regression method does not apply to the data in this article, and the logistic regression method should be used instead.

Logic functions were initially introduced by Belgian mathematician Pierre François Verhulst in the mid-19th century as a tool for modeling the growth of biological numbers. Logistic regression, as a method of processing binary data, usually performs well and is the most commonly used and simplest analysis method. The formula for the logistic regression model is as follows:

$$Pr(Y=1|X)=\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}} \quad (1)$$

By transforming it, it can be concluded that:

$$\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}=e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p} \quad (2)$$

At this point, $\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}$ is called the probability ratio. By taking the logarithm of both ends of the equation, it will become to:

$$log\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \quad (3)$$

At this point, the right end of the equation becomes a general linear model, and its coefficients can be explained. Meanwhile, the random forest method has relatively better results in data classification, so this article selects random forest as a method for comparison.

Random forest generates a large number of decision trees by sampling sample units and variables. For each sample unit, all decision trees classify it sequentially. The mode of all decision tree prediction categories is the prediction result of the random forest for that unit.

3 RESULTS AND DISCUSSION

3.1 Data Processing

Because the data selected in this article as independent variables such as Job and Education are categorical data, which is not conducive to the process of logistic regression, this article will number Job, Marital, Education, Contact, Month, and Poutcome, with the unknown item numbered as 0. The remaining projects are gradually numbered according to their importance under this classification, and the detailed numbering is shown in Table 3.

From Table 3, it can be seen that the Poutcome indicator is different from other indicators in that it includes two situations: Failure and Other, so it is not possible to simply sort them in natural number order. This article believes that Failure should maintain the same weight as Success, and the Other situation should be better than Unknown. Therefore, the value assigned to Failure is -2, while others are still assigned normally.

Table 3: Number of Each Index.

Index	Number Situation
Job	Unknown (0); Unemployed (1); Student (2); Retired (3); Housemaid (4); Blue-collar (5); Services (6); Self-employed (7); Technician (8); Management (9); Admin (10); Entrepreneur (11)
Marital	Divorced (1); Single (2); Married (3)
Education	Unknown (0); Primary (1); Secondary (2); Tertiary (3)
Contact	Unknown (0); Telephone (1); Cellular (2)
Month	Jan (1); Feb (2); Mar (3); Apr (4); May (5); Jun (6); Jul (7); Aug (8); Sept (9); Oct (10); Nov (11); Dec (12)
Poutcome	Failure (-2); Unknown (0); Other (1); Success (2)

3.2 Model Evaluation

The results obtained from logistic regression of the above numerical variables are shown in Table 4:

Table 4: Logistic Regression Results for All Variables.

Index	Coefficients	z value	Pr(>|z|)
(Intercept)	-3.464	-26.628	0.000***
Age	0.003	2.241	0.025*
Job	-0.022	-3.186	0.001**
Marital	-0.155	-6.295	0.000***
Education	0.166	6.774	0.000***
Defaultyes	-0.328	-2.029	0.042*
Balance	0.00002	4.679	0.000***
Housingyes	-0.960	-25.268	0.000***
Loanyes	-0.668	-11.640	0.000***
Contact	0.587	22.171	0.000***
Day	-0.005	-2.193	0.028*
Month	-0.011	-1.699	0.089
Duration	0.004	64.387	0.000***
Campaign	-0.131	-13.015	0.000***
Pdays	0.003	16.927	0.000***
Previous	0.050	6.204	0.000***
Poutcome	0.524	27.433	0.000***

Note: *** ‘p<0.001’; ** ‘p<0.01’; * ‘p<0.05’

From the above table, it can be seen that at a 95% confidence interval, only the Month indicator does not meet the test criteria. This article believes that the possible reason is that during the last time the bank contacted customers, the number of contacts in mid-May was much higher than in other months. As the

month is close to the middle of the year, May is also close to the middle position in the numbering system. Therefore, its impact on the entire model is difficult to explain with a simple linear trend and is not significant in the test. The solution to this article is to delete the indicator and rebuild the model to ensure that all used indicators pass the test.

Table 5 shows the model validation results after deleting Month: From table 5, it can be seen that in the model after removing Month, the p-values of all indicators are less than 0.05, indicating that the model has passed the test and fits the data well. Chi-square tests will be conducted on these two models to ensure that the model still has the same degree of fit as the full indicator model. The test results are shown in Table 6.

From the chi-square test results in table 6, it can be seen that the p-value is 0.08934, which is greater than 0.05, and the result is not significant. Therefore, it can be considered that the fitting degree of the model without the Month indicator is as good as that of the model with all indicators.

Table 5: Logistic Regression Results without Month.

Index	Coefficients	z value	Pr(>|z|)
(Intercept)	-3.506	-27.452	0.000***
Age	0.003	2.113	0.035*
Job	-0.023	-3.279	0.001**
Marital	-0.156	-6.361	0.000***
Education	0.165	6.738	0.000***
Defaultyes	-0.332	-2.051	0.040*
Balance	0.00002	4.573	0.000***
Housingyes	-0.953	-25.229	0.000***
Loanyes	-0.670	-11.669	0.000***
Contact	0.582	22.108	0.000***
Day	-0.005	-2.283	0.022*
Duration	0.004	64.382	0.000***
Campaign	-0.132	-13.098	0.000***
Pdays	0.003	17.112	0.000***
Previous	0.051	6.204	0.000***
Poutcome	0.523	27.409	0.000***

Note: *** ‘p<0.001’; ** ‘p<0.01’; * ‘p<0.05’

Table 6: The Chi-square Test

Df	Deviance	Pr(>Chi)
1	2.886	0.089

To ensure that the above logistic regression does not have excessive deviation, which leads to singular standard error tests and imprecise significance tests, this article conducts an excessive deviation test on the data after removing the Month indicator, with a value of 0.52, which is far less than 1. Therefore, the above logistic regression does not have excessive deviation, and its results are reliable.

3.3 Explain Model Parameters

In logistic regression, the logarithmic odds ratio of the response variable is y=1. Therefore, the meaning of its regression coefficient is the change in the logarithmic odds ratio of the response variable that can be caused by a change in one unit of the predictor variable when other predictor variables remain unchanged. Due to the poor interpretability of the logarithmic odds ratio, this article exponentiates it, and the results are shown in Table 7:

Table 7: Exponential Coefficient.

Index	Exponential Coefficient
(Intercept)	0.0300
Age	1.0033
Job	0.9772
Marital	0.8554
Education	1.1790
Defaultyes	0.7178
Balance	1.0000
Housingyes	0.3854
Loanyes	0.5118
Contact	1.7900
Day	0.9952
Duration	1.0040
Campaign	0.8767
Pdays	1.0030
Previous	1.0518
Poutcome	1.6875

From the above table, it can be seen that the impact of each indicator on whether the customer will choose a fixed deposit project can be roughly divided into three categories: almost no impact, slight impact, and significant impact. Among them, the indicators with almost no impact are Age, Job, Balance, Day, Duration, Pdays, and Previous, and only Day has the opposite direction of influence. These indicators are data that are difficult to artificially change, such as Age, and unpredictable data, such as Duration, which have little impact on the final results. The indicators with a slight impact are Marital, Education, Default, and Campaign, which will subjectively affect the judgment of customers. For example, if expenses increase after marriage, customers are more willing to choose current deposits for future needs. If the number of contacts is too frequent within a certain period, customers may feel bored, etc. The indicators that have a significant impact are Housing, Loan, Contact, and Poutcome, which are data that are highly correlated with fixed deposits. If a customer has both

a housing loan and a personal loan, it is almost difficult for them to make a fixed deposit.

3.4 Model Prediction

Apply the above model parameters to the test set, and mark the final calculation results less than or equal to 0.5 as 0 and the rest as 1. The confusion matrix can be obtained as shown in Table 8:

Table 8: Confusion Matrix.

		Prediction Category	
		0	1
Real Category	No	3916	84
	Yes	390	131

According to the above table, the accuracy of the model is 89.52%, indicating strong predictive ability. Meanwhile, based on this test, draw the receiver operating characteristic (ROC) curve, as shown in Figure 2:

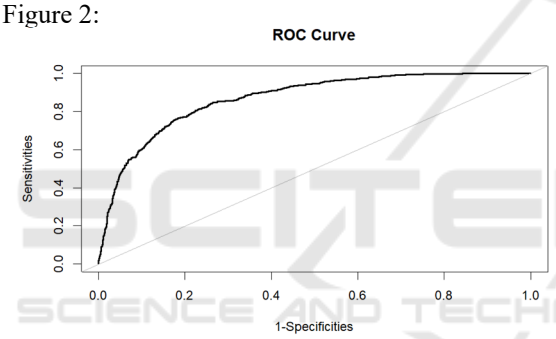

Figure 2: The ROC Curve (Picture credit: Original).

The above figure 2 shows that the model has a good prediction performance, with an area under the curve (AUC) of 0.87, which is close to 1. Therefore, it can be considered that the model has a good fitting classification effect.

3.5 Random Forest

Meanwhile, this article also compared the random forest method with logistic regression. Among them, the random forest method in this article is to randomly select 5 variables from each node of each tree and generate 500 decision trees. The following is the result of the operation using the random forest algorithm. The confusion matrix is shown in Table 9:

Table 9: Confusion Matrix from Random Forest.

		Prediction Category	
		0	1
Real Category	No	4000	0
	Yes	15	506

From the above table, it can be seen that the random forest method has significantly improved the prediction of results, with an accuracy of 99.67%. At the same time, the ROC curve can also be obtained as shown in Figure 3:

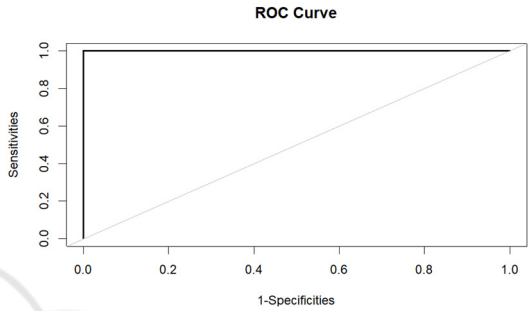

Figure 3: The ROC Curve from Random Forest (Picture credit: Original).

The ROC curve shown in the above figure indicates that the accuracy of the random forest method is very good, and its AUC is almost 1. Therefore, it can be said that this method perfectly solves the problem of determining whether customers will engage in a fixed deposit business.

Overall, the random forest method is superior to ordinary logistic regression methods. On the confusion matrix, the random forest method did not make any errors in determining whether customers would make fixed deposits, indicating that the model believes that customers who make fixed deposits will make fixed deposits, and there are very few unclassified customers. Its model has high accuracy and can be used to accurately determine whether customers will engage in fixed deposit projects.

3.6 Discussion

Although the above model has passed the test and achieved good accuracy in testing, there are still some issues that have not been resolved in this article. Firstly, the effectiveness of individual indicators is insufficient. Due to the customer's desire to protect their privacy and the bank's respect for their privacy rights, there is always an unknown situation in the collection results of many indicators, and even under the Poutcome indicator, the proportion of Unknown

exceeds 80%. Although this article believes that the situation of Unknown is inevitable and cannot be simply removed from the results to ensure the validity of all information, if the information contained in the collected indicators can be ensured as much as possible, the model will also be more accurate and credible. Secondly, the data collection is not comprehensive enough. Whether to engage in fixed deposit business is a subjective choice, which is easily influenced by subjective factors such as the customer's personality and emotions. However, the indicators selected in this article are mostly objective data. Although the data in this article has good detection results, if this aspect of data can be added, the model should have better accuracy. Third, the model is not universal. Since the data selected in this article is from a Portuguese bank, it may only apply to that bank or country. For other banks or countries, it only guides at the modeling level and cannot truly be used for customer identification.

4 CONCLUSION

Based on the research results of this article, the following conclusion can be drawn: using logistic regression to analyze customer data of Portuguese banks can provide a guiding model to determine whether customers will handle fixed deposit business. At the same time, among the selected indicators, whether there is a housing loan, whether there is a personal loan, contact information, and the results of the last marketing activity have the greatest impact on whether to handle fixed deposit business. Customers who have housing loans and personal loans are less likely to apply for fixed deposit services, while customers who use convenient communication devices and participated in the last marketing campaign are more willing to apply for fixed deposit services.

Through the research in this article, some suggestions can be provided. For banks, the focus of engaging customers in the fixed deposit business is not on obvious indicators such as fund-related data, but on how to make good use of factors that are not easy to detect but can affect the results, such as the number of contacts. For old customers, more contact times can provide stronger customer stickiness, while for new customers, more phone calls may only lead to boredom. Only in this way can banks retain old customers as much as possible while seizing potential customers, increasing their fixed deposits, and thus gaining more benefits. In addition, whether customers will handle fixed deposit business is a subjective choice, and factors such as emotions and personality can be considered to obtain a more accurate model.

REFERENCES

B. Deng, Science Technology and Industry 23(10), 151-157 (2023).
C. G. Hurst, Direct Marketing An International Journal 2(2), 111-124 (2008).
C. Yan, M. X. Li and W. Liu, Applied Soft Computing Journal 92(2), 106-109 (2020).
H. A. Elsalamony, International Journal of Computer Applications 85(7), 12-22 (2014).
J. H. Ahn and K. J. Ezawa, Decision Support Systems 21(1), 21 (1997).
J. N. Mann, Process for telemarketing: US, (2006).
J. R. Yang, Financial Regulation Research 12, 12 (2014).
K. H. Kim, et al., IEEE 25(7), 314-317 (2016).
M. Liu, Y. M. Yan and Y. D. He, Springer International Publisher 6(1), 190-197 (2017).
R. M. Cain, Journal of Public Policy and Marketing 15(1), 135-141 (1996).
X. X. Jiang, Jiangsu Communication 6, 72-74 (2021).
Y. Y. Jiang, International Journal on Data Science and Technology 4(1), 35-41 (2018).