Research on the Influencing Factors of Precision Marketing in Commercial Banks from The Perspective of Data Mining

Tingyu Wu[a, &], Jiasheng Yang[*, &], Xinyi Xu[b], Qin Qin[c]

Yunnan University (Chenggong), Kunming, 650000, China

* Corresponding Author Email: jiashengy@mail.ynu.edu.cn, [a]TingyuWu@mail.ynu.edu.cn, [b]xuxinyi@mail.ynu.edu.cn, [c]qqin@mail.ynu.edu.cn

[&]These authors contributed equally to this work and should be considered co-first authors

Abstract. With the rapid development of financial technology, modern data mining techniques are becoming increasingly widespread in the banking industry. Telephone marketing, which offers interactivity, measurability, extensive spatial coverage, and database support, is an effective way for banks to promote their financial products. This study focuses on how to use existing customer data to explore the potential promotion value of bank financial products while meeting customer needs and to develop a telephone marketing model that is suitable for the era of big data. We begin by analyzing the data of a Portuguese bank to predict whether customers will sign a long-term deposit agreement and identify which features have a significant impact on the response variable. Univariate analysis is used to identify these features, and then the logistic regression model is employed to confirm them. Finally, the random forest model is used to identify the variables that have the greatest impact on customer response. After identifying these key variables, we analyze the reasons and provide improvement measures and optimization strategies for the telephone marketing promotion of bank financial products.

Keywords: Financial technology, Precision marketing, Univariate analysis, Logistic regression.

1. Introduction

A stable banking industry is crucial for a sound financial environment, and long-term deposits play a fundamental role in the risk management of banks and the development of their loan business. This study explores factors that influence customers' decisions to sign long-term deposit agreements through data analysis and proposes measures to attract long-term deposits.

2. Data Analysis

2.1. Descriptive Data Analysis

This study considered real data from a retail bank in Portugal from May 2008 to June 2013, consisting of 41,188 customer data points with 20 independent variables. The data analysis aims to identify key variables that have a significant impact on the customer response variable.

2.2. Univariate Analysis

In this section, we conduct univariate analysis by examining a feature and observing its differences between two classes. It is important to note that this analysis is not conclusive and should be confirmed through predictive modeling.

From the class distribution of the percentage of customers who signed a long-term deposit agreement, it can be observed that the response of "no" is about 8 times higher than the response of "yes", causing class imbalance. The bank's customers are mainly distributed in the age group of 30-45 years old, and age is not a decisive indicator of whether customers will sign a long-term deposit agreement. By observing the distribution and median of these indicators between the two response variables, we found that the daily indicator "euribor3m" is very useful in determining whether customers will sign a long-term deposit agreement, and the quarterly indicators "emp.var.rate" and "nr.employed" are also effective in predicting customer response. We removed the variable "duration" because it is difficult to predict and has a significant interference with the model fit. Although "managers" are the highest proportion of occupational types, further modeling is needed to determine their importance. The majority of customers are married, and this feature may be helpful in predicting whether customers will sign a long-term deposit agreement, but further analysis and modeling are needed to confirm its importance. The feature of weekly working hours may not have a significant impact on predicting whether customers will sign a long-term deposit agreement. The outcome of the previous marketing campaign may be very helpful in predicting whether customers will sign a long-term deposit agreement, as there are a large number of customers who subscribed to term deposits in the previous campaign.

2.3. Data preprocessing

In the data preprocessing, the "NA" values were removed, reducing the number of data points from 41188 to 30488. Given the 10 categorical variables, they were converted into numerical variables for ease of use with logistic regression. Finally, the data was split into a training set and a testing set with a ratio of 8:2, where the training set contains 24390 samples and the testing set contains 6098 samples.

3. Model Development and Solving

3.1. Model Selection

During the project initialization phase, we conducted data cleaning and transformation and used exploratory data analysis techniques and univariate analysis to summarize the key features of the dataset to identify how variables influence customer responses to long-term deposits. We then used two classification methods, logistic regression and random forest, to build and validate the predictive model.

3.2. Logistic Regression

3.2.1 Model Description and Development

Logistic regression is a widely used regression model for classification problems. Its basic idea is to establish a regression relationship between the independent variables and the dependent variable, and then map the linear regression results to the probability interval (0,1) using the logistic function.

This article used the dataset with all 20 features to build a logistic regression model, predicting the logarithmic linear relationship between the predictors and the response. 11 of the estimated coefficients of the predictors were statistically significant, indicating good model performance with an AIC value of 11587. On the test set, the model achieved a classification accuracy of 90.4%, although accuracy is not the best metric due to imbalanced classes. The ROC curve was plotted, and the AUC value was calculated to be 92.7%, with a true positive rate of 39.92% and a false positive rate of 2.49%. The stepwise regression algorithm was used to select the best model, which included 13 significant predictors and removed 6 less-contributing predictors, improving the model's predictive power.

Table 1 Models with full set of features and models with reduced features

full set of features			Reduced features		
variable name	Parameter Estimation	P value	variable name	Parameter Estimation	P value
Intercept	23.067	0.34	Intercept	37.284	0.02
job	0.013	0.06	job	0.012	0.07
marital	0.092	0.04	marital	0.068	0.09
euribor3m	0.805	0.00***	euribor3m	0.884	0.00***
nr.employed	-0.016	0.00***	nr.employed	-0.017	0.00***
AIC=11587			AIC=11579		

(Note: limited by the layout, only some variables are displayed)

3.2.2 Model Performance

Using logistic regression with simplified feature set: We found that the simplified model and the full model had the same classification accuracy of 90.4% and roughly the same AUC value rounded to the first decimal point, which is 92.7%. This suggests that the simplified model can improve classification accuracy while maintaining simplicity. The simplified model with 14 variables is even simpler and provides roughly the same classification accuracy.

3.3. Random Forest

3.3.1 Model Establishment

Each decision tree is built based on randomly selected samples and features, and pruning is not performed during the construction process. Random forest can reduce overfitting and improve model generalization. In prediction, multiple decision trees predict the same sample and the average or voting result is taken as the final prediction.

Based on experience, we chose m=4 and ntree=50 to implement the random forest method in this article. The confusion matrix was calculated and the model achieved a classification accuracy of 91.5%.

3.3.2 Model Performance

Our random forest model performed well on the test data, with an AUC value of 94.8%. "Duration" was observed to have high importance through the predictive variable importance plot. Although the importance plot can be used for feature selection, high-cardinality variables may have the issue of multiple testing and potential collinear variables that could cause bias. Therefore, we constructed new logistic regression and random forest models and removed the variable.

3.4. Model Improvement

We built new LR and RF models by removing the "duration" variable. For the LR model, we trained it on all predictors except for "duration" using backward stepwise regression to select the best linear combination of predictors that gave the lowest AIC value while not compromising classification accuracy. The new full model trained on 19 predictors, while the stepwise regression selected the same 13 predictors as the simplified model that included "duration".

The full model, simplified model, and new RF model that included "duration" all had low classification accuracy of about 89.0%. This indicates that "duration" had a detrimental effect on model performance and would lead to inaccurate and unrealistic predictions if used in real-world applications.

In comparison, we found that RF performed better than LR in handling nonlinearity and overfitting in high-dimensional data. However, after removing the "duration" predictor, the LR model performed slightly better than the RF model, suggesting that "duration" has nonlinear features that reduce classification accuracy when included. Additionally, "duration" may be a dependent variable, and removing it can further improve classification accuracy.

4. Conclusion

The above data analysis draws our attention to two main characteristic variables, euribor3m and nr. employee. Euribor3m has a positive impact on customers signing long-term deposit agreements, while nr.employed has a negative impact. We analyzed the reasons for this.

Firstly, interbank lending carries higher risks, so interbank lending rates are usually higher than depositors' deposit rates, so the interbank borrowing rate is usually higher than the deposit rate for savers. If the three-month euro interbank offered rate increases, the cost of borrowing from other banks will increase, and banks will turn to increase the deposit rate to attract savers for long-term deposits, thereby increasing the number of long-term deposit customers.

Secondly, according to the law of diminishing marginal returns, as the number of employees increases, the marginal output efficiency shows a decreasing trend. Too many employees may lead to low efficiency of bank work, leaving a bad impression on savers, which can cause loss of core customer resources and have a negative impact on the bank's reputation. Too many employees are difficult to manage, and their quality is uneven, which makes it difficult to guarantee the effectiveness of their telephone marketing. These factors will reduce the number of long-term deposit customers of the bank.

5. Suggestions

Based on the above analysis, we have identified the key factors that affect customers' long-term deposits. Banks should take reasonable measures to promote their financial products accurately in response to these important factors.

Banks should fully utilize financial technology to establish their own databases and regularly update data. Valuable decision-making information can be mined from the database to improve the bank's decision-making efficiency and make timely adjustments to marketing strategies. Banks can provide personalized services to customers based on the information provided by the database, scientifically classify customers, and accurately grasp their needs.

Facing the fierce competition of Internet finance, banks should implement precision marketing strategies and rely on data mining technology to promote marketing diversification and personalization. They should identify the behavior patterns of customers signing long-term deposit agreements and win the battle for customer resources. When the three-month Euribor interbank lending rate rises, banks can moderately increase the interest rates for depositors within a reasonable range. With other factors remaining unchanged, an increase in interest rates can to some extent attract customers' long-term deposits.

Banks should streamline their organizational personnel and provide regular professional telephone sales training to employees to continuously improve their work capabilities. They should fully utilize their talent advantages, improve employee labor efficiency, and enable employees to provide refined services.

Acknowledgements

The authors gratefully acknowledge the financial support from Provincial University Students' Innovation and Entrepreneurship Training Program Project (S202210673062) "Legal Home - A Legal Forum System Based on Django".

References

[1] Elsalamony H A. Bank direct marketing analysis of data mining techniques [J]. International Journal of Computer Applications, 2014, 85(7): 12-22.

[2] Moro S, Laureano R, Cortez P. Using data mining for bank direct marketing: An application of the crisp-dm methodology [J]. 2011.

[3] Canhoto A. Portuguese banking: A structural model of competition in the deposits market [J]. Review of financial economics, 2004, 13(1-2): 41-63.

[4] Mendes V, Rebelo J. Structure and performance in the Portuguese banking industry in the nineties [J]. Portuguese Economic Journal, 2003, 2: 53-68.

[5] Cabrita M R, Bontis N. Intellectual capital and business performance in the Portuguese banking industry [J]. International Journal of Technology Management, 2008, 43(1-3): 212-237.

[6] Balan V J R V S. A review on data mining in banking sector [J]. American Journal of Applied Sciences, 2013, 10(10): 1160-1165.

[7] Jin W, He Y. Three data mining models to predict bank telemarketing[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2019, 490(6): 062075.