Analysis of financial product purchases based on logistic regression

Yuzhen Wang¹,², Jingqiao Qin²

¹Silk Road Economic Research Institute, Lanzhou University of Finance and Economics, Lanzhou, Gansu, 730020, China
²School of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou, Gansu, 730020, China

*Corresponding author’s e-mail: wangyzh@lzufe.edu.cn

Abstract. In order to analyse the different characteristics of different users in the purchase of financial products and improve the accuracy of the identification of potential customers, this paper uses the logistic regression method to mine the data of customer historical transaction based on comparing the performance indexes of various classification methods. By analyzing the internal relationship between product consultants' communication with customers, personal loans and deposits, and the number of on-site exchanges, we can find potential customers and make accurate predictions on customer purchasing behaviors to help wealth management companies better tap new customers.

1. Introduction

As people's awareness of financial management becomes stronger, more and more people hope to obtain stable and generous income by purchasing financial products. On the other hand, in a highly competitive market, wealth management companies must effectively identify potential customers by understanding their customers’ behavioral characteristics in order to gain greater competitive advantage. At present, the research results in this field are also relatively rich, such as: Dong Jiyang[1] through the characteristic analysis of the lost fund customers, accurately identify and predict the customers who are about to lose, analyze the reasons for their loss, and provide feasible countermeasures for the company to retain and remedy. Wang Anning[2] and others analyzed the car review texts of SUV models, and proposed that adding product parameters to build an emotional feature preference model can more accurately predict customers’ perceived preferences and provide a theoretical basis for corporate marketing and product design. Zhao Yi[3] It is proposed that when purchasing Internet wealth management products, it is not only necessary to rationally analyze the purchase behavior of customers, but also to consider the psychological deviation of investors. In addition, the analysis methods and model selection of customer characteristics are currently relatively diversified, mainly including traditional analysis Methods and machine learning methods, such as: Luo Hongxuan[4] and others used the decision tree CHAID algorithm model to predict the power outage sensitivity of all customers of a municipal power supply bureau in Guangdong Province, effectively identify the special power users, and solve the problems of this population. It can be seen that the analysis of customer characteristics has become a very important issue in enterprise customer relationship management. However, there are relatively few research results using logistic regression to analyze and predict customer characteristics. Therefore, this article uses logistic regression model to analyze the sales data of financial products of a
wealth management company. Analyze the influence of customer characteristics on product purchases, and predict the purchase behavior of new customers.

2. Research method

This article uses a logistic regression model to analyze historical transaction data of a certain wealth management product, with the purpose of analyzing whether customer characteristics affect the occurrence of purchase behavior. Since whether to purchase financial products is a binary discrete type, through comparative analysis, a logistic regression model is selected. The specific research model is as follows:

$$\ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{16} X_{16} + \varepsilon$$  \hspace{1cm}(1)$$

Among them, the dependent variable is whether the financial product will be purchased, p means to buy financial product, 1-p means not to buy financial product; independent variable $X_i$ represents 16 characteristics such as age, education status, loan status, and communication duration, $\beta_i$ represents self. The coefficient of the variables, $\varepsilon$ is the random disturbance term.

In order to improve the accuracy of the model in identifying customers, the loss function of logistic regression is introduced. The specific model is as follows:

$$J(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log( \hat{p}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{p}^{(i)})$$  \hspace{1cm}(2)$$

Among them, $y^{(i)}$ indicates the classification label of the training data set, $\hat{y}^{(i)}$ indicates the classification label of the test data set, and $\hat{p}^{(i)}$ indicates whether to purchase financial products. The relationship is as follows:

$$\hat{y}^{(i)} = \begin{cases} 1, & \hat{p}^{(i)} \geq 0.5 \\ 0, & \hat{p}^{(i)} \leq 0.5 \end{cases}$$  \hspace{1cm}(3)$$

After the loss function is obtained, gradient descent is used to continuously approximate the optimal solution, and the minimum value of the loss function is obtained by deriving the formula (2).

3. Empirical analysis

3.1. Data pre-processing

The data analyzed in this article comes from Alibaba Cloud (https://tianchi.aliyun.com/). The data is preprocessed by reasonable use of numpy, pandas, and matplotlib packages. The processed data set is shown in Figure 1.

![Figure 1. Training data set](image.png)

3.2. Preliminary characterisation

In this paper, a heat map is used to visually display the data to initially judge the relationship between customer characteristics and product purchases. According to the heat map, we can see: the correlation coefficient between duration (the duration of communication between product consultants and
customers) and Y is 0.4, and the two are weakly correlated. The correlation coefficient between days (the number of days the product consultant has not contacted the customer) and the previous (the number of times the product consultant communicates with the customer) is 0.41, and the two have a weak correlation. In addition, the correlation between most features is not obvious, and it is impossible to accurately determine the relationship between customer features and product purchases. Therefore, considering the large degree of data skew, model training cannot be directly performed, and it is necessary to process the data skew first.

3.3. Handling skewed data
According to the sample data, using the count function to quantitatively analyze the number of samples purchased and the number of samples not purchased are 2064 and 15543, respectively. It can be seen that the sample data is highly skewed, and the direct use of the sample data will affect the accuracy of the analysis of the purchase of wealth management products. In order to balance the data, under-sampling was performed on samples of non-purchasing wealth management products and over-sampling was performed on samples of purchased wealth management products. First calculate the sampling number ratio to be 7:1 based on the ratio of the two types of data, and then sample the two types of data separately. 50% of the samples that are not purchased are under-sampled, and the purchased samples are over-sampled by 3.6 times, and the final number of samples after processing are 7771 respectively.

3.4. Model comparison and training

3.4.1. Model comparison
Because model accuracy is the focus of model training, three models are used to compare and analyze the Precision Rate, Recall Rate and F1_Score Index to determine the optimal classification model.

- Logistic regression model
The results of the logistic regression analysis are shown in Figure 2. According to Figure 2, its precision rate is 0.83, recall rate is 0.69, and F1_Score is 0.48. According to the confusion matrix shown in Figure 3, on the training set sample, there are 5,644 samples with correct predictions, 1298 samples with incorrect predictions, and 605 customers would buy the product.

- Decision tree model
The result of the decision tree model is shown in Figure 4. According to Figure 4, its precision rate is 0.80, a recall rate is 0.75, and F1_Score is 0.47. According to the confusion matrix shown in Figure 5, on the training set sample, there are 5,389 samples with correct predictions, 15,04 samples with incorrect predictions, and 654 customers would buy the product.
The results of the SVM model is shown in Figure 6. According to Figure 6, its precision rate is 0.30, a recall rate is 0.95, and F1_Score is 0.24. According to the confusion matrix shown in Figure 7, on the training set sample, there are 145,9 samples with correct predictions, 529 samples with incorrect predictions, and 835 customers would buy the product.

Based on the above comparative analysis, it can be concluded that the effect of logistic regression is the best, so this paper uses logistic regression model to analyze the main characteristics that affect customers' purchase of financial products, and predict whether customers will purchase financial products, so as to accurately identify potential customers.

### 3.4.2. Model training

The balanced data were used to train the logistic regression model by substituting logistic regression() function. To test the accuracy of the model, the ROC curve (AUC value) and ROC_AUC score were used to analyze the prediction performance of the model. It is known that the more the ROC curve connects to the upper left corner, and the AUC>0.5, the higher the ROC_AUC score, indicating that the model prediction accuracy is better than the random classifier. By calculating the ROC curve of the data, the curve is close to the upper left corner, and AUC = 0.85. By using the lr_s.predict_proba (test_X)[:,1] method to calculate the predicted sample probability value, the output ROC_AUC score is 85.15. Therefore, the model has high accuracy, which further proves that it has good application value and predictive performance in analyzing customer characteristics and predicting customer purchase behavior.
3.5. Analysis of results

3.5.1. Analysis of the characteristics that influence customers to buy financial products

In order to speed up the solution of the gradient, firstly, the data are normalized to eliminate the dimensional influence of the data. Secondly, a trained logistic regression model is used to calculate the effect coefficient of customer characteristics, which is the influence of customer characteristics on the purchase of products. And the results are shown in Table 1. Finally, as can be seen from Table 1, the positive characteristics that affect customers’ purchase are: Product Advisor Duration, Product Advisor Uncontacted Days, Annual Account Balance, Previous Interactions with Customer, etc. The negative characteristics that affect customer purchase are: Number of times to purchase on-site communication, customer's personal loan and housing loan, whether there is any default record, etc. It can be seen from the data in Table 1 that long-term continuous communication can promote the occurrence of customer purchase behavior, while on-site communication can not promote the occurrence of customer purchase behavior, but weaken customer purchase behavior. The following conclusions can be drawn from the effect coefficient of customer characteristics: in the process of negotiation with customers, we should pay attention to step by step, regularly increase the number of exchanges with customers, let customers fully understand the financial products, give customers enough time to consider. Instead of cramming information into the customer during a live meeting, ignore the customer experience.

Table 1. Impact of features

| Positive impact | Negative effects |
|-----------------|-----------------|
| Job 0.324080    | Age -0.131434   |
| Duration 13.384815 | Education -0.115537 |
| Balance 2.194556 | Default -0.533155 |
| Housing 1.209377 | Loan -0.660784 |
| Pdays 3.089489  | Day -0.186029   |
| Previous 1.926053 | Campaign -4.333770 |

3.5.2. Analysis of the likelihood of customers purchasing financial products

- Individual customers

Logistic regression model can be used to predict the new customer data, judge the probability of buying behavior, and then identify potential customers effectively. First, the purchase rate of the customer was calculated using \( \text{lr\_s\_proct\_proba()} \) and the output was \([[[0.86572751, 0.13427249]]]\). Further analysis revealed that the probability of the customer not buying was approximately 86.6% and the probability of the customer buying was approximately 13.4%, so that the customer was an important potential customer.

- Group customers

The logistic regression model can be applied not only to the prediction of a single sample of data, but also to a batch of data. First, input the processed test data set, define tag 1 for customers with probability greater than 0.5 and tag 0 for customers with probability less than 0.5, and get the purchase situation of each customer. Then the customers with the tag of 1 are selected for key mining, and their customer characteristics are further analyzed, so as to improve the purchase rate of financial products.
4. Conclusion
The study of customer characteristics has always been one of the most important issues in customer relationship management. Accurately grasping customer features can effectively help enterprises identify potential customers and improve the accuracy of marketing. Therefore, this paper takes the historical transaction data of a financial company as the research object, uses a logistic regression model to analyze the behavior characteristics of customers when purchasing financial products, and predicts the purchase rate of financial products, so as to help financial companies better identify potential customers.

Acknowledgments
Fund Projects:
[1] Soft science project of Gansu science and Technology Department(Study on the evaluation and optimization of business environment in Gansu Province)
[2] Lanzhou University of Finance and Economics Major Tender Project (Lzufe2020A-001)
[3] Project of Silk Road Economic Research Institute of Lanzhou University of Finance and Economics (JYYZ201903)

References
[1] Dong, Jy. (2020) Prediction of Fund Customer Churn Based on Decision Tree Automation Feature Selection: Thinking in the Postepidemic Age.09, Shandong Social Science:74-80
[2] Wang, An. (2020) Customer Perception Preference Model Combining Characteristic Emotion and Product Parameters.09,28:199-208
[3] Zhao, Y. (2019) Research on factors influencing purchase intention and purchase behavior of Internet financial products.01, Yangzhou University:64
[4] Luo, Hx., Xiao, Y., Jin, X. (2020) Research on Outage Sensitivity Prediction Algorithm Based on Power Customer Clustering.10,45:106-112
[5] Xie, L., Chen, Wt., Cao, Hw., Xiao, J. (2020) Research on semi supervised customer key feature selection model based on Hall of comprehensive integration.07,32:236-245
[6] Jiang, Cq., Xu,Tg., Wang, Zh. (2020) An evaluation method of P2P network loan default risk based on rejection inference.06,34:165-172