The Application of Data Mining Technology in Customer Relationship Management of Commercial Banks

Qingsong Yu, Hong Jiang, Xiao Ma
School of Computer Science and Software Engineering
East China Normal University
Shanghai, China

Abstract—**Based on the marketing cases of commercial banks, with LOGISTIC regression, the customer response model is developed to forecast the response probability of target customers for different marketing campaigns. The model is verified by the comparison and assessment of K-S statistics tests between development group and validation group. Through the in-deep analysis of the characteristics of customers in the marketing campaigns, together with the marketing budget and model operation, the strategy of differentiating customer relationship management is proposed. The valuable data offered to the management level helps commercial banks to make more intelligent business decisions and provide quality financial services to the customers.**

Keywords- LOGISTIC; data mining; customer relationship management; commercial banks

I. INTRODUCTION

In the increasingly fierce competition and homogenization of financial activities, with the development of "big data", mobile Internet and cloud computing, the competitiveness of different banks has gradually shifted to the acquisition of customer resources and the improvement of customers' value to enterprises. Customer demand is the premise of enterprise profit, and customer satisfaction is not only an important index to improve customer loyalty, but also the source of enterprise benefit [1-3].

Commercial banks have accumulated a large amount of data, but the banking system was designed based on the traditional financial business, and customer preferences and lifestyle were not considered, so that the banks can only make more use of national policies and macroeconomic information in business decision-making. Data mining technology can make full use of the massive data and facilitate the effective use of big data analysis, which helps the bank to grasp the pulse of the business market, exploit big data for decision making, and develop into the forefront of the industry in the future.

Customer Relationship Management (CRM) aims to increase efficacy and efficiency in the acquisition and retention of profitable customers, starting from the construction and development of solid relationships. Exploiting CRM system in the bank has the following benefits:

- It can fully embody the management concept of "customer-centered", and regard the massive customer data as valuable assets.
- It can provide a good management tool for commercial banks, and exhibit advanced management ideas and excellent management modes.
- It can mine and retain high-quality customers, analyze customer's practical and potential demand, innovate service contents and modes, focus on service quality, provide distinctive differentiated services and maximize customer value.
- It can invest superior resources, build brand images, create special products and services, excavate high value customers, and constantly improve the benefit.

In this paper, the high performance data mining software is adopted and combined with the product demand. Based on the marketing cases, with historical marketing data as the samples of the model, LOGISTIC stepwise regression [4-5] is exploited in processing the variables to obtain the model equations. The score card is established to verify the model. The K-S statistics diagrams [6-7] are created for development group and validation group respectively to evaluate the performance of the model. The verified model is then put into use, and is monitored and evaluated as well as optimized to extend its service life span. The model can accurately orientate the target customer groups, deeply analyze the core business value, efficiently develop a personalized marketing strategy, flexibly make practical business operation, and effectively provide an edge tool for customer relationship management of commercial banks.

The remainder of this paper is structured as follows. Section II discusses related work in the literature. The development and evaluation of customer response model are proposed and analyzed in Section III. Section IV depicts the customer management strategy based on customer response model. Section V provides a conclusion and future work.

978-1-5386-8097-1/18/$31.00 ©2018 IEEE

II. LITERATURE REVIEW

Data mining is essentially the adoption of data analysis and processing to reveal the association relationship among data, forecast the future development trend, assist problem analysis and solution, and provide reference and support for decision-making. The traditional data analysis enterprises such as COGNOS, SPSS, SAS, SAP, Teradata, and Hyperion have better market share in data mining technology and markets. With the exponential increase of data, big data processing technology is the future development trend of data mining, which requires faster processing time, multi-type and multi-source data processing and analysis, high-speed marketing analysis, special big data computer, distributed computer cluster, and complex-structured data analysis [8-9].

With the advent of the era of online banking, financial services are omnipresent, and are gradually integrated with other commercial activities. Financial industry is expanding and extending, which leads to the mixed operation not only among financial institutions but also between the financial and nonfinancial institutions. Alipay, WeChat Pay, Faster Payments, etc. are challenging the traditional commercial banks. In order to maintain the competitiveness of the bank, every customer should be retained. To achieve the win-win situation among the bank and the customers, the bank should understand the real needs of the customers, excavate the customers' potential desire, reduce marketing costs, allocate reasonably all resources, and orientate precisely the target customer groups. Data mining technology can be exploited to gather, sift, analyze and dig out the hidden and valuable information from various types of data resources during the interaction between the bank and the client. Data mining technology plays an important role in customer relationship management. It has a significant and far-reaching impact on the implementation of CRM strategy and the improvement of decision-making level in commercial banks [10-12].

III. CUSTOMER RESPONSE MODEL OF COMMERCIAL BANKS

From the perspective of commercial banks, customer relationship management is a magic weapon for retaining old customers and developing new customers. Data mining is a strong technical support for scientific decision making, which can provide effective and valuable business information. Commercial banks have rich customer information and financial data resources. Mathematical modeling adopted in financial industry can provide a strong support for business decision-making.

The model will be developed after data acquisition, cleaning, conversion and sampling. In the commercial bank's marketing campaigns, LOGISTIC regression is one of the most widely used predictive modeling. Compared with LOGISTIC, decision tree and neural network will produce similar technology results, whereas decision tree may lead to the instability of prediction results, and there exists overfitting in neural network.

A. LOGISTIC Regression Model

The LOGISTIC regression model is expressed as follows:

$$\log it(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n = K \quad (1)$$

Or:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n . \quad (2)$$

Apply logarithm operation on both sides of the above two equations, we have:

$$p = \frac{1}{1+e^{-\eta}}, \quad (3)$$

where p is the probability of event occurrence, and 1- p is the probability that the event does not occur. In the LOGISTIC regression model, $odds = \frac{p}{1-p}$ is the odds ratio or cross-product ratio or relative odds, which is the ratio of response customer to nonresponse customer.

B. Variable Selection in Regression Model

When building a LOGISTIC regression model, the variables which have more important influence need to be sifted out from all the independent variables. There are four kinds of methods in the selection of variables: forward regression, backward regression, stepwise regression, and full model.

Based on the wide range and long time span of marketing campaigns in commercial banks, as well as the large amount of customer data, stepwise regression is used to filter the variables in this paper to balance the resource consuming and the running speed. The modeling information and the sifting results are shown in Table I, Table II, and Table III.

TABLE I. TESTING GLOBAL NULL HYPOTHESIS: BETA=0

Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ration	399899.594	12	<.0001
Score	527404.919	12	<.0001
Wald	290812.819	12	<.0001

TABLE II. ASSOCIATION OF PREDICTED PROBABILITIES AND OBSERVED RESPONSE

Parameter	Value	Parameter	Value
Percent Concordant	84.9	Somers' D	0.704
Percent Discordant	14.5	Gamma	0.708
Percent Tied	0.5	Tau-a	0.111
Pairs	664304976610	C	0.852

TABLE III. ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATE

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.3921	0.00790	91592.7955	<.0001
AGEQK(25-35)	1	0.0207	0.00564	13.4008	0.0003
AGEQK(35-45)	1	-0.2034	0.00561	1315.0639	<.0001
AGEQK(45-55)	1	0.1528	0.00583	734.6713	<.0001
AGEQK(>55)	1	0.0367	0.00858	18.2965	<.0001
GENDER(male)	1	-0.3440	0.00237	21140.4284	<.0001
CRLIMIT(10K-50K)	1	-0.6599	0.00692	9097.0124	<.0001
CRLIMIT(50K-100K)	1	0.7387	0.00695	11293.3926	<.0001
CRLIMIT(100K-300K)	1	1.1242	0.00809	19296.2144	<.0001
CRLIMIT(>300K)	1	1.0214	0.02190	2179.5100	<.0001
NET	1	0.0706	0.00249	800.3269	<.0001
HOLD	1	0.3855	0.00313	15135.8145	<.0001
XFSUM	1	5.612E-06	2.947E-08	36247.5318	<.0001

The model is solved by the stepwise regression and its parameters are estimated by the iterative method. The fitting of the model reaches the convergence status. The overall fitting test is conducted to compare whether the predictive variable is consistent with the nonpredictive variable. It is shown in Table I that the Likelihood Ratio is 399899.594 and its corresponding P value is less than 0.0001, which manifests that the modeling effect is significant. As for the statistical association analysis (see Table II), it can be seen the C statistic is 0.852, and 84.9% is from the Concordant, which can be interpreted as the association degree between predictive probability and observation value is very high. In general, if the association degree is more than 70%, the model can be put into actual use.

In addition, through parameter estimation (see Table III) by stepwise regression, parameter test variables AGEQK, GENDER, CRLIMIT, NET, HOLD and XFSUM, are eventually sifted out as the variables for customer response model: age, gender, credit line, online banking signed, aging of account, the amount of consumption (in 6 months). The P values corresponding to these six variables are all less than the significance threshold 0.05. Therefore, the variables in LOGISTIC stepwise regression are all significant, playing important roles in evaluating customer response model. Finally, the LOGISTIC regression model is obtained as in (4):

In (4), X_1 denotes the age of the customer ranging from 25 to 35, X_2 denotes the age of the customer ranging from 35 to 45, X_3 denotes the age of the customer ranging from 45 to 55, X_4 denotes the age of the customer above 55, X_5 denotes the gender of the customer, X_6 denotes the credit limit within 10K to 50K, X_7 denotes the credit limit within 50K to 100K, X_8 denotes the credit limit within 100K to 300K, X_9 denotes the credit limit more than 300K, X_{10} denotes online banking signed, X_{11} denotes the aging of account, and X_{12} denotes the amount of consumption (in six months).

In our LOGISTIC model, when the coefficient of a significance variable is positive, and the conditions for the other significance variables remain unchanged, the values of LOGIT will monotonically increase with the significance variables. By observing the coefficient of each significance variable, it can be concluded that the following features will improve the response rate of the customer: the customer is older than 45 years old, the credit limit is higher than 50 thousand, the signing of online banking, the aging of account over three years, and the consumption transaction is high in six months.

After frequency statistics, it is found that the probability of Type I error and Type II error is 7.32% and 1.02% respectively, as shown in table IV.

TABLE IV. SAMPLE PREDICTION IN RESPONSE MODEL (DEVELOPMENT GROUP)

Actual Customer Response	Model Prediction Results		Total
	Response Customer(1)	Nonresponse Customer(0)	
Response Customer(1)	232013	18317	250330
Nonresponse Customer(0)	27061	2626656	2653717

C. *Effect Evaluation of Model*

The effectiveness of the model is predicted after the customer response model is developed. In order to evaluate whether a model is excellent or not, the probability of the customer response should be predicted accurately and efficiently. As described above, if the capture rate of the response customer reaches 85%, only 40% of customers is needed for the model, as shown in Table V and Table VI.

$$\begin{aligned} \log it(p_i) = {} & -2.3921 + 0.0207X_1 - 0.2034X_2 + 0.1582X_3 + 0.0367X_4 - 0.344X_5 \\ & - 0.6599X_6 + 0.7387X_7 + 1.1242X_8 + 1.0214X_9 + 0.0706X_{10} + 0.3855X_{11} + 0.000005612X_{12} \end{aligned} . \quad (4)$$

TABLE V. CUSTOMER RESPONSE PROBABILITY (DEVELOPMENT GROUP)

Group	Customers	Cumu.Cust. Ratio	Response Customers	Response Rate	Cumu.Resp. Customers	Cumu.Res. Capture Rate	Cumu.Nonresp. Customers	Cumu.Nonresp. Capture Rate	Discriminative Rate
1	290404	10%	120801	42%	120801	48%	169603	6%	42%
2	290405	20%	48535	17%	169336	68%	411473	16%	52%

Group	Customers	Cumu.Cust. Ratio	Response Customers	Response Rate	Cumu.Resp. Customers	Cumu.Res. Capture Rate	Cumu.Nonresp. Customers	Cumu.Nonresp. Capture Rate	Discriminative Rate
3	290405	30%	22258	8%	191594	77%	679620	26%	51%
4	290405	40%	28877	10%	220471	88%	941148	35%	53%
5	290404	50%	9668	3%	230139	92%	1221884	46%	46%
6	290405	60%	6100	2%	236239	94%	1506189	57%	38%
7	290405	70%	6730	2%	242969	97%	1789864	67%	30%
8	290404	80%	3278	1%	246247	98%	2076990	78%	20%
9	290406	90%	2878	1%	249125	100%	2364518	89%	10%
10	290404	100%	1205	0%	250330	100%	2653717	100%	0%

TABLE VI. CUSTOMER RESPONSE PROBABILITY (VALIDATION GROUP)

Group	Customers	Cumu.Cust. Ratio	Response Customers	Response Rate	Cumu.Resp. Customers	Cumu.Res. Capture Rate	Cumu.Nonresp. Customers	Cumu.Nonresp. Capture Rate	Discriminative Rate
1	124459	10%	45046	36%	45046	43%	79413	7%	36%
2	124459	20%	23107	19%	68153	66%	180765	16%	50%
3	124460	30%	11539	9%	79692	77%	293686	26%	51%
4	124459	40%	11963	10%	91655	88%	406182	36%	53%
5	124458	50%	4005	3%	95660	92%	526635	46%	46%
6	124459	60%	2227	2%	97887	94%	648867	57%	37%
7	124459	70%	2788	2%	100675	97%	770538	68%	29%
8	124459	80%	1358	1%	102033	98%	893639	78%	20%
9	124460	90%	1392	1%	103425	100%	1016707	89%	10%
10	124459	100%	499	0%	103924	100%	1140667	100%	0%

It can be seen from Table V and Table VI that the higher the discriminative rate the better effect of the model. Through the analysis of Type I and Type II error, it can be concluded that the response predictive ability is accurate. The number of actual response customers in the first group is several times higher than that in the tenth group no matter for Development Group or Validation Group, which illustrates that the response model can better distinguish between the high response customers and the low response customers.

Finally, the K-S statistic diagrams are created for the development group and verification group of the customer response model respectively, where the horizontal coordinate denotes Cumulative Customer Percentage, and the vertical coordinate denotes the ratio of Cumulative Response Customers to the total number of Response Customers, as shown in Fig. 1 and Fig. 2 respectively.

Fig. 1 and Fig. 2 show that the maximum Discriminative Rate of the model reaches 53% in the 40th percentile, which means that the first 40% of data can capture 88% of the response customers. The results of K-S test for both development group and validation group are approximately the same, which proves that the customer response model is a better model no matter for the development group or validation group.

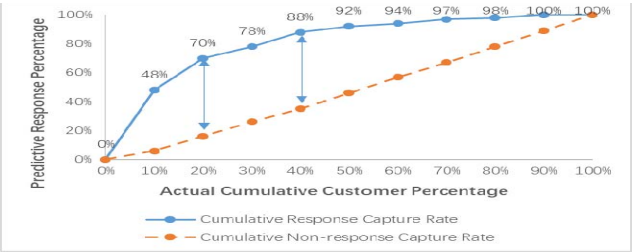

Figure 1. K-S for Customer Response Model (Development Group)

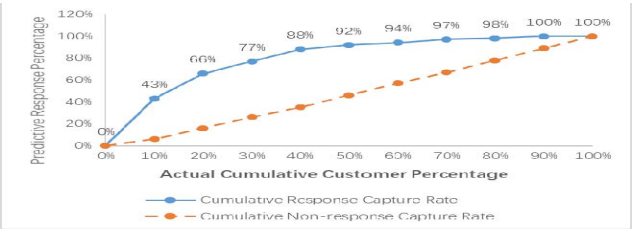

Figure 2. K-S for Customer Response Model (Validation Group)

IV. CUSTOMER MANAGEMENT STRATEGY BASED ON CUSTOMER RESPONSE MODEL

From the perspective of the development of commercial banks, the cost of developing a new customer is much higher than that of retaining an old customer, and promoting the loyalty of a new customer requires much more energy than retaining an old customer. The survival and development of commercial banks needs the support of stable target customers.

In order to provide better, faster and suitable financial services, the prediction model is established for the classification of the customers in the marketing campaigns by using data warehouse, data mining and other modern information technology, which can accurately determine the target customer groups for different marketing campaigns or services, and strengthen the scientific, efficient, and comprehensive relationship management of the target customer groups.

As described and analyzed above, female customers account for only 39% of the total sample data, but up to 53% (see Fig. 3) are among the response customers, which demonstrates that the marketing campaign is more attractive for female customers. As for the customers whose credit limits are greater than 100 thousand, the response rate of male customers is higher than that of the female, especially for the male customers whose credit li limits are greater than 300 thousand, and the difference between them is as high as 24% (see Fig. 4). In-depth analysis discovers that 89% of the male customers are married, and consume mostly by using credit cards in offshore large shopping malls, restaurants and entertainment, whose single consumption is much higher than that of the single customer. The response rate of the male customers with high credit limit is so high that it may be a kind of family activity such as traveling abroad, which infers that the influencing factors of the female cannot be ignored.

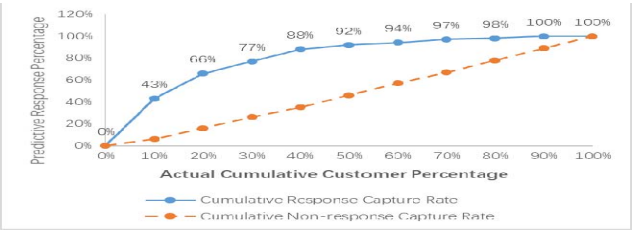

Figure 3. Customer gender analysis

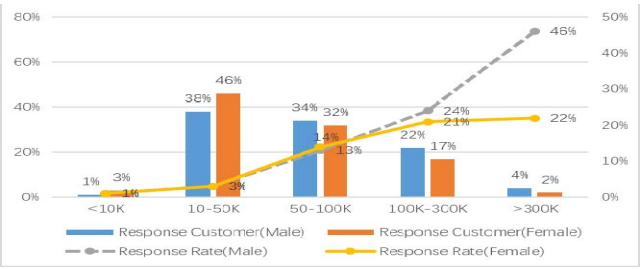

Figure 4. Customer credit line and gender analysis

Among female response customers, the customers aged from 35 to 55 years old account for 63%. As for the response rate, the female customers over 45 years old are 18% and 15% respectively (see Fig. 5), far exceeding the average response rate of 12%. Therefore, in order to promote the brand and broaden the audience, the business departments of the bank should carry out promotional marketing strategy for the female customers aged from 25 to 45 years old, whose characteristics are young and white-collar, and whose career is just started or slightly blooming but with strong consuming willingness. The banks need to strengthen the cultivation of such customers and design corresponding business campaigns for them. Although the response rate may be low temporarily, there exist a huge number of such potential customers, which is helpful to improve the reputation and communication efficiency of the bank. If the bank wants to increase the business volume, it should focus on the female customers over 45 years old, who have better economic foundation and good consumption concept.

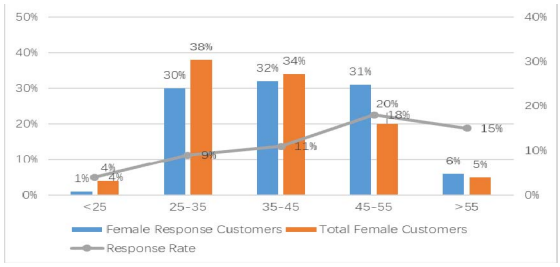

Figure 5. Age analysis of female customers

V. CONCLUSION AND FUTURE WORK

Valuable information and rules from mass customer data are obtained, and the implementation of bank customer relationship management is fully achieved. The main innovations of this paper are summarized as follows: compared with correlation analysis and decision tree, logistic regression model can orientate the target customers more effectively and more accurately in the bank customer relationship management. The response model can evaluate accurately each customer and retain tightly valuable customers.

LOGISTIC regression is adopted to identify the response probability of the customers in marketing campaigns. It is verified that the model has the features of high accuracy, good explanation and excellent efficiency. The sample data are divided into the development group and the validation group. It is found that the difference between the two groups is small, and the development data is put into the model to obtain the final results and parameters.

Type I and Type II error tests are conducted respectively for the development group and validation group, to inspect the fault tolerance of response prediction for the customers. Higher Type I and Type II errors indicate that the parameters of the model need to be adjusted. K-S statistics diagrams are created to verify the discriminative ability of the model for the customers.

The customer model can effectively dig out the hidden, unknown information and relationship, which can help the commercial banks to classify the valued customers, analyze the consumption features of the customers, search accurately target customer groups, develop timely new products and services to meet the needs of the customers of different levels, retain more customers,, and ensure that the customers receive good financial benefits and commercial efficiency which achieves a win-win situation.

Data mining technology brings people more and more convenient business, life and work, however, the problems of information privacy and data security involved are gradually highlighted. Therefore, commercial banks should pay more attention to the protection of privacy data and ensure that the sensitive data in the bottom layer will not be leaked out.

REFERENCES

[1] E. Ascarza, S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, etc., "In pursuit of enhanced customer retention management: review, key issues, and future directions," Customer Needs and Solutions, vol. 5, pp. 65-81, November 2017.

[2] H. K. Yau, and H. Y. H. Tang, "Analyzing customer satisfaction in self-service technology adopted in airports," J Market Anal, vol. 6, pp. 6-18, March 2018.

[3] F. L. Cabanillas, F. M. Leiva, and J. S. Fernandez, "A global approach to the analysis of user behavior in mobile payment systems in the new electronic environment," Service Business, vol. 12, pp. 25-64, March 2018.

[4] Vinicius Veloso de Melo, and Wolfgang Banzhaf, "Automatic feature engineering for regression models with machine learning: An evolutionary computation and statistics hybrid," Information Sciences, Volumes 430-431, pp. 287-313, March 2018.

[5] R. Maranzato, M. Neubert, M. Neubert, and A. Lago, "Fraud detection in reputation systems in e-markets using logistic regression and stepwise optimization," ACM SIGAPP Appl Comput Rev, vol. 11, pp. 14-26, November 2010.

[6] T. B. Arnold, and J. W. Emerson, "nonparametric goodness-of-fit tests for discrete null distributions," The R Journal, vol. 3, pp. 34-39, December 2011.

[7] P. Khrueasom, and A. Pongpullponsak, "The integrated model of the Kolmogorov–Smirnov distribution-free statistic approach to process control and maintenance," J King Saud Univ Sci, vol. 29, pp. 182-190, April 2017.

[8] M. Mitik, O. Korkmaz, P. Karagoz, H. Toroslu, and F. Yucel, "Data mining approach for direct marketing of banking products with profit/cost analysis," Review of Socionetwork Strategies, vol. 11, pp. 17-31, June 2017.

[9] Zhang Shengwei, "Research on customer management strategy of commercial banks based on big data," Economic Research Guide, vol. 4, pp. 67-69, February 2017.

[10] A. O. Jaraba, J. J. C. Fierro, and E. Centeno., "Analyzing relationship quality and its contribution to consumer relationship proneness," Service Business, vol. 1, pp. 1-21, January 2018.

[11] M. F. Gholami, F. Daneshgar, G. Beydoun, and F. Rabhi., "Challenges in migrating legacy software systems to the cloud: an empirical study," Information Systems, vol. 67, pp. 100-113, July 2017.

[12] J. J. Cambrafierro, E. Centeno, A. Olavarria, and R. Vazquezcarrasco, "Success factors in a CRM strategy: technology is not all," J Strat Market, vol. 25, pp. 1-18, March 2017.