Bank Customer Loss Forecast Analysis

Shengqian Zhou*

*Corresponding author. Email: shengqianzhou1225@gmail.com

University of Sydney, Sydney, NSW 2006, Australia

Abstract. With the development of banks, the competition is increasing among banks, and the loss of customers is a serious problem, which affects the profitability of banks. Therefore, it is necessary to judge the reasons for customer loss through a series of indicators. In order to explore which factors are related to customer churn, this paper uses Kaggle's open data set and the methods of contrastive analysis, factor analysis and binary logistic regression. Through comparative analysis, it is found that age, customer activity, the number of products owned by the customer in ABC Bank, the credit score of the customer and the account balance has an impact on the loss of customers. The accuracy of prediction by logistic regression is 67.3% and 68.5% respectively.

Keywords: Bank customer loss, Cross table, Factor analysis, Binary logistic regression

1 Introduction

In the increasingly competitive banking industry, bank customers can easily change banks. How to retain the existing customers has become a serious problem faced by every bank. Customer loss will have a negative effect in terms of bank profitability. The cost is huge because of losing a customer, and keeping them as long as possible can lead to increased profits. In the financial industry, 5% customer retention improvement rate can bring more than 25% profit increase [1]. In addition, when a certain number of customers are reached, it is more difficult and costly to create new customers [2]. Therefore, studying why customers are leaving and predicting whether they will leave is crucial for the long-term development of banks.

In the past, regression algorithms, decision tree algorithms, and neural networks have been used to analyze customer churn. Researchers believe that different technologies have different advantages in forecasting. Linear regression, decision tree and RIPPER models have better performance in predicting customer churn [3]. However, neural networks have higher accuracy than other machine learning models [4]. Using six algorithms, Respectively, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting Classifier and K-Nearest Neighbor predict customer loss. Accuracy, Precision, Recall, and F-Measure are used to judge the accuracy. Finally, it is concluded that Gradient Boosting Classifier and Random Forest have the best performance [5]. Using logistic regression, artificial neural networks and support vector machines predict customer loss, and compared these methods. Think logistic regression is better than other methods [6]. In terms of loss factors, taking the communication industry as an example, it is found that the change of customer status would have an impact on the loss of customers [7]. Factors such as the frequency of mobile banking would have an impact on customer loss [8]. Through decision tree and K-means, the paper analyzes the reasons for the

ICBBEM 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347318

loss of insurance companies' customers, and believes that the service, popularity and reputation of insurance companies are important factors affecting the loss of customers [9]. Taking the securities industry as an example, it is found that monetary value variables have an important impact on whether customers are losing [10]. Through k nearest neighbor, support vector machine, decision tree and random forest classifier, the conclusion is drawn that the customers of a few products need special attention of the bank [11].

Existing literature uses logistic regression, decision tree, neural network and other methods to predict the loss of bank customer, but few studies use the binary logistic regression model to analyze the factors affecting the loss of the bank customers in detail. Therefore, this paper uses ABC Bank data set, obtained from Kaggle to analyze the influence of relevant variables on customer loss through comparative analysis, factor analysis and logistic regression, and helps banks find out which factors need to be focused on to maintain existing customers. The rest of the article is shown below. The second chapter introduces the data set and research ideas, and the third chapter is the use of comparative analysis, factor classification and binary logistic regression. The fourth chapter is summary and suggestion.

2 Data and Methods

The data is a public data set obtained from the Kaggle website. Created in 2022, the data records the customer data of ABC Bank. The data involves various information about the customers, including personal information, the number of products purchased in the bank, credit score, whether they hold credit cards, the location of the customers, etc.

Some of the variables are explained below:

Churn: 1 indicates that the customer has lost, and 0 indicates that the customer has not lost.

Credit card: Customers with credit cards will be represented by 1, the customer without a credit card will be represented by a 0.

Active member: 1 represents active customers and 0 represents inactive customers.

Because of the large number of data variables, it is impossible to accurately judge which variables have an impact on the loss of customers. Therefore, comparative analysis is first adopted to determine which variables will have an impact on the problem, and then factor analysis is carried out on the screened variables, and binary logistic regression is carried out on the screened variables and the variables after factor analysis. Thus, the relationship between each variable and customer loss is obtained.

3 Results

3.1 Contrastive analysis

First, we test whether the discrete variables in the data have an impact on customer churn. The test results are shown in Table 1. Table 1 shows whether a credit card has an impact on the loss of customers; whether an active customer has an impact on the loss of customers and whether the number of products the customer has in ABC Bank has an impact on the loss of customers.

Table 1. Results of Discrete Variables * Bank Churn Crosstabulation

Value		df	Asymptotic significance (2-sided)
Results of Credit Card and Customer Churn by Pearson Chi-square Test	.509	1	.475
Results of Active Customer and Customer Churn by Pearson Chi-square Test	243.760	1	<.001
Results of Product Number and Customer Churn by Pearson Chi-square Test	1503.629	3	.000

From Table 1, we can see that only the progressive significance (on both sides) of credit card and customer churn is greater than 0.05 and the other two are less than 0.05, indicating that in addition to whether have a credit card, whether customers are active or not and the number of products owned by customers in ABC Bank have a significant impact on the loss of customers. Next, the continuity variables in the data are checked. For continuous variables, normality test is required first.

Table 2. Normality test results of continuous variables

	churn	Kolmogorov-Smirnov			Shapiro-Wilk		
		statistic	df	sig.	Statistic	df	Sig.
age	0	.119	7963	<.001			
	1	.027	2037	.001	.998	2037	.006
credit_card	0	.447	7963	.000			
	1	.443	2037	.000	.576	2037	<.001
balance	0	.268	7963	.000			
	1	.186	2037	<.001	.870	2037	<.001
estimated_salary	0	.055	7963	<.001			
	1	.062	2037	<.001	.953	2037	<.001

According to Table 2, we can know that the customer's credit score, account balance, estimated salary and age do not conform to the normal distribution. In order to know whether these three variables have an impact on customer loss, non-parametric test of these three variables is needed. Table 3 shows the results of the nonparametric tests.

Table 3. Normality Test Summary

Null Hypothesis		Test	sig	Decision
1	The customer's account balance has no effect on customer churn	lndependent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
2	Customer age distribution has no effect on customer churn.	lndependent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

3	The customer's estimated salary has no effect on customer churn	lndependent-Samples Mann-Whitney U Test	.227	Retain the null hy-pothesis. iscorrece
4	A customer's credit score with the bank has no effect on cus-tomer churn.	lndependent-Samples Mann-whitney U Test	0.02	Reject the null hy-pothesis.

According to Table 3, only the third column accepts the null hypothesis, the first, second, and fourth columns all reject the null hypothesis. We can conclude that estimated salary does not have an impact on customer churn, while customer reputation score, age and account balance will have a significant impact on customer turnover. Therefore, we have obtained a total of 5 variables that will affect the loss of customers, which are age, whether the customer is active or not, the customer's product count in ABC Bank, the credit score of the customer and the account balance.

3.2 Factor analysis

Through observation, we found that there may be correlations between the selected variables. For example, younger customers may be more active than older customers or active customers may have more products at the bank. Therefore, factor analysis can be used to convert these indicators into fewer and unrelated indicators, and study the impact of these indicators on whether customers are lost.

Table 4. Data Correlation

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.502
Bartlett's Test of Sphericity	Approx. Chi-Square	1069.128
	df	10
	sig.	<.001

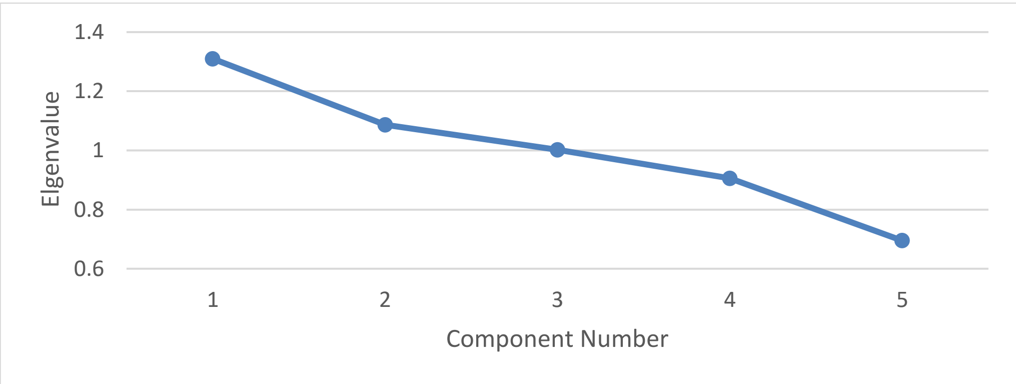

Fig. 1. Scree Plot

Table 5. The Degree to Which the Factors Explain the Data

Initial Eigenvalues				Extraction Sums of Squared Loadings
Component	Total	% of inter-pretive	Total Inter-pretive %	Total
1	1.310	26.198	26.198	1.310
2	1.087	21.737	47.935	1.087
3	1.002	20.037	67.972	1.002
4	.906	18.123	86.095	
5	.695	13.905	100.00o	

According to Table 4, the value of KMO test is 0.502>0.5; Spherical Bartley test were 0.01<0.05, indicating that factor analysis can be used for dimension reduction. According to Table 5, only three indicators are needed to interpret 67.972% of the data. According to Figure 1, only the initial eigenvalues of the first three indicators are greater than 1. Therefore, it is most appropriate to select three indicators.

Table 6. Factor Component List

	1	2	3
balance	.806		
products_number	-.804		
age		.742	
active_member		.731	
credit_score			.979

According to Table 6, account balance and product quantity have the most significant influence on whether customers are lost, while credit score has the least significant influence on whether customers are lost.

3.3 Binary logistic regression

The selected variables and variables obtained by factor analysis were respectively analyzed by binary logistic regression. The regression analysis obtained from the selected variables and the regression analysis of variables obtained by factor analysis is shown in Table 7.

Table 7. The Result of Binary Logistic Regression Prediction

Observed		Predicted (before factor analysis)			Predicted (after factor analysis)		
		0	1	Percentage	0	1	Percent-age
churn	0	5363	2600	67.3	4934	3029	62.0
	1	642	1395	68.5	845	1192	58.5
Overall Percentage				67.6			61.3

According to Table 7, the accuracy of predicting whether customers will lose by using this model reaches 67.3% and 68.5%, indicating that the error of logistic regression is very small and it is suitable for binary logistic regression. In addition, the accuracy rates of binary logistic regression after factor analysis were 62% and 58.5% respectively.

Table 8. Chi-Square Tests

	B	S.E.	Wald	df	sig.
balance	.000	.000	117.953	1	<.001
age	.073	.003	825.747	1	<.001
credit_score	-.001	.000	5.120	1	.024
active_member(1)	1.083	.057	362.324	1	<.001
products_number	-.023	.046	.246	1	.620
Constant	-4.906	.236	433.643	1	<.001

According to Table 8, we can see that active members have the largest B value, indicating that whether customers are active or not has the greatest impact on the loss of customers. Inactive customers are more likely to lose than active customers.

Table 9. Active_member*products_number Crosstabulation

	Value	df	Asymptotic significance (2-sided)
Results of Active Member and Products Number by Pearson Chi-square Test	17.194	3	<.001

Table 10. Active_member*products_number Crosstabulation

		1	2	3	4	Total
active_member	0	2521	2144	153	31	4849
	1	2563	2446	113	29	5151
Total		5084	4590	266	60	10000

According to Table 9, the significance is less than 0.05, indicating that customer activity has an impact on the number of bank products owned. According to Table 10, the number of banking products owned by active customers is higher. Active customers have a total of 5,151 products, inactive customers can have a total of 4,849 products. Most bank customers have one or two products, and very few have four.

4 Conclusion

The results of binary logic analysis are basically in line with our cognition, and customers with high activity are less likely to be lost. According to the results in Table 9 and Table 10, the active customers have more bank products, and they have more opportunities to obtain satisfactory service and product experience, which makes the active customers less likely to lose. By reducing the dimension, it can be seen that the account balance and the number of products have the greatest impact on the loss of bank customers, which can be interpreted as that customers with high account balance and more products may be hindered by the bank when they leave the bank, while customers with low account balance and fewer products can easily leave the bank. In a word, Customers with deposits have a lower attrition rate than those without. In addition, the logistic regression results after factor analysis, the accuracy of the model are not as good as the regression results before factor analysis, which may be because factor analysis reduces the available variables, thus reducing the accuracy of the model. However, the accuracy is only slightly decreased, so the factor analysis results are still worthy of reference.

Based on the above analysis, for banks, to retain customers, they need to improve customer activity and deposit amount through various marketing methods, and actively guide customers to increase their dealings with the bank.

There are several ways to improve the enthusiasm of bank customers. To begin with, Banks can offer ultra-personalized services. To be specific, banks recommend some financial products more suitable for customers according to their preferences, so as to increase customers' dependence on banks. 72% of customers believe that personalized service is very important in the financial services industry [12]. Hence, it will greatly reduce the situation of customers leaving. In addition, it is highly necessary to banks can set up special departments. Specific, banks can set up special customer experience departments to study customers' experiences on different products through customer data. This is mainly because customers want more personalized service. In these ways, the problem of bank customer loss can be addressed to some extent. As for the limitations of this paper, because only one bank's customer group is selected for analysis, the conclusion may not be suitable for all banks.

References

[1] Reichheld, F. (2001). Prescription for cutting costs. https://static1.squarespace.com/static/618efa3c26282b35b68344af/t/61b2c9e8a0c4f0311b116adc/1639107048571/A+Prescription+for+Cutting+Costs.+Bain+/26+Company/2C+2011.pdf

[2] Kim, M.-K., Park, M.-C., & Jeong, D.-H. (2004) The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy., 28: 145–159.

[3] Li, F., Lei, J., Tian, Y., Punyapatthanakul, S., & Wang, Y. (2011) Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking. Proceedings of the Ninth Australasian Data Mining Conference.,121:119-124

[4] Mundada, M. R. (2019) Enhanced Deep Feed Forward Neural Network Model for the Customer Attrition Analysis in Banking Sector. International Journal of Intelligent Systems and Applications., 11(7): 10–19.

[5] Dalbah, L. M., Ali, S., & Al-Naymat, G. (2022). An Interactive Dashboard for Predicting Bank Customer Attrition. In: 2022 International Conference on Emerging Trends in Computing and Engineering Applications. pp. 1-6.

[6] DUR, R., KOÇER, S., & DÜNDAR, Ö. (2022) Evaluation of Customer Loss Analysis for Marketing Campaigns in the Banking Sector. Politeknik Dergisi, 1–1.

[7] Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006) Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy., 30: 552–568.

[8] Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016) Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation., 2: 1-13.

[9] Allahyari, R., 1+, S., & Rodpysh, K. (2016) Applying Data Mining to Insurance Customer Churn Management. International Proceedings of Computer Science and Information Technology., 30: 82-92.

[10] Ahn, Y., Kim, D., & Lee, D.-J. (2019) Customer attrition analysis in the securities industry: a large-scale field study in Korea. International Journal of Bank Marketing., 38, 561–577.
[11] Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. In: 2020 4th international conference on electronics, communication and aerospace technology. pp. 1196-1201.
[12] Ritz M, (2021). Capco Study: 72% of Customers Rate Personalization as “Highly Important” in Today’s Financial Services Landscape. https://www.businesswire.com/news/home/20210526005143/en/