Bank Customer Churn Analysis and Prediction

Wenhui Zhang*
201911030129@mail.bnu.edu.cn

Business School, Beijing Normal Unversity, Beijing, 100875, China

Abstract. The term "Customer Churn" refers to the state in which the customer stops using products or services of a company. Although the bank will inevitably lose users, which is inevitable in the process of replacing the old and new banking users, the proportion and changing trend of lost users can indicate the bank's ability to retain users and the development trend of the bank. Therefore, it is necessary for banks to know the reasons leading a client to leave the company. In order to explore the factors affecting the loss of bank users, this paper selects a dataset obtained from Kaggle, using the methods of crosstabs analysis, independent samples T test, factor analysis and one-way ANOVA respectively. In addition, logistic regression is also used to predict customer churn. From the t-test, this paper finds that those who quit have lower credit scores, are older and have larger balances than those who don't. From the factor analysis, this paper finds that the feature of country and balance are the most explanatory factors. From the logistic regression, this paper finds that the percentages of correct predictions are 69.5% and 65.3% respectively before and after selecting main components.

Keywords: Banking, Customer Churn, crosstabs analysis, independent-samples t Test, factor analysis, one-way ANOVA, Logistic Regression.

1 Introduction

In the current fierce market competition prevailing in banking, it is particularly important for commercial banks to understand the loss of customers and the reasons for customers to stop trading with the bank. The loss of customers can be costly, as acquiring new customers is often much more expensive than retaining existing ones. Therefore, understanding and preventing customer churn is critical to the long-term viability of commercial banks.

In recent decades customer churn has become increasingly vital for industries like telecom, e-commerce and banking. The literature has explored the loss of users in different industries. Neslin et al. analyzed how the accuracy of the customer churn forecasting model is affected by methodological factors, finding that the observed differences in the accuracy of forecasts submitted by different groups can significantly change the profitability of churn management activities [1]. Ahn et al.focused on the mobile telecom service industry, analyzed the mediating effect of partial customer churn on the relationship between the determinants of customer churn and the total churn, and discussed how call quality related factors influenced customer churn [2]. Hadden et al.focused on the accuracy of churn management and reviewed some of the prevailing technologies [3]. Richter et al. pointed out that churn prediction had become an important business intelligence application for modern telecom operators, and found that the accuracy of churn prediction can be improved by analyzing customer interactions by assessing the social proximity of recent churn [4]. Jahromi et al. developed a

MSIEID 2022, December 09-11, Chongqing, People's Republic of China

DOI 10.4108/eai.9-12-2022.2327608

data mining technique for modeling customer churn in a B2B environment, identifying customers with a high likelihood of churn in the relatively near future [5]. Vafeiadis et al.studied several machine learning approaches to customer churn prediction in the telecom industry [6].

In existing studies, many methods have been explored and used to predict churn in different industries such as telecommunications and banking. Xia et al. proposed a structural risk-minimization support vector machine (SVM) to solve the problem of the predictive ability of machine learning methods, finding that the method enjoyed the best accuracy, hit rate, coverage rate and lift coefficient [7]. Burez et al. examined how we can better handle class imbalances in attrition forecasts, showing that oversampling can improve forecast accuracy, especially when using the AUC for assessment [8]. Xie et al. proposed a new learning method called IBRF and applied it to customer churn prediction, showing significantly higher prediction accuracy than other algorithms [9]. Tsai et al. (2009) combined two different neural network technologies, back-propagation artificial neural network (ANN) and self-organization mapping (SOM), to predict turnover, and found that the performance of ANN + ANN hybrid model was significantly better than that of SOM + ANN hybrid model [10]. Richter et al. proposed a new framework called "Group First Churn Prediction" to predict which user groups are most likely to lose [4]. Verbeke et al. applied two new data mining techniques, AntMiner+ and ALBA, to customer churn prediction modeling, and the results showed that ALBA improved the learning of classification techniques and improved performance [11]. Verbeke et al. studied the use of social network information to predict customer loss, finding that the social network effect had a significant impact on the performance of the customer loss prediction model, and the parallel model setting could improve the profits generated by retention activities [12].

By reviewing the existing literature, it can be found that the existing literature mainly studies how to predict customer churn, but few users conducted a detailed analysis of the characteristics and influencing factors of customer churn. Therefore, this article uses the data from the Kaggle website to analyze the characteristics and influencing factors of customer churn to help banks make better decisions on customer retention. The rest of the paper is organized as follows. The data used in the research is described in Section 2, analysis process using comparative analysis, independent samples T test, single factor ANOVA, factor analysis and the modeling process based on logistic regression are presented in Section 3 to Section 7. In Section 8, we conclude.

2 Data Description

The dataset is obtained from Kaggle whose name is “Bank Customer Dataset for Churn prediction”. The dataset consists of 10000 records and 11 features (see Table 1), and this paper splits the features into Predictors and Target variables. Predictor variables contain information about customers, and the target variable refers to customer abandonment. Predictor variables can be divided into 2 groups. One is customer behavior including Balance, NumofProducts, IsActiveMember, CreditScore, and HasCrCard. The other is customer demographics, including Age, Gender, Geography, EstimatedSalary.

The interpretation of some variables and value rules are as follows:

- Exited: If the customer has withdrawn his/her account from the bank, then exited = 1; otherwise exited = 0.
- Gender: 0 if it is male, and 1 if it is female.
- Geography: country of residence, The Geography variable has three values: 0 if it is France and 1 if it is Spain, and 2 if it is Germany.
- Tenure: how long has he/she had a bank account in the bank.
- IsActiveMember: whether a customer is an active member or not.1 if active member, 0 if inactive member.
- HasCrCard: whether a customer has a credit card or not. 1 if the customer has credit card, 0 if not.
- NumOfProducts: the number of products that a customer has bought through the bank.
- CreditScore
- Age
- Balance
- EstimatedSalary

In this dataset, we look at the "Exited" column to see if that customer is churned or not. Our goal is then to use all other features (e.g. CreditScore, Geography, etc.) to predict customer churn. The numerical features in this dataset are credit score, age, tenure, balance, number of products, and estimated salary. It is helpful to look at the statistical summary shown below.

Table 1. Data description.[Owner-draw].

	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000	10000	10000	10000	10000	10000	10000	10000
mean	38.9218	5.0128	76485.889	1.5302	0.7055	0.5151	100090.2399	0.2037
std	10.48781	2.89217	62397.405	0.58165	0.45584	0.49979	57510.49282	0.40277
min	18	0	0	1	0	0	11.58	0
max	92	10	250898.09	4	1	1	199992.48	1

3 Crosstabs Analysis

To describe the interaction between Exited variable and other categorical variables, we use cross-tabulation (or "crosstab" for short). For Exited and Geography, we could see from Table 3 that there was a statistically significant difference in nationality between those who exited and those who did not ($p<0.05$). From Table 2, of those who did not quit, 21.3% were German and 52.8% were French; of those who did, however, 40.0% were German and 39.8% were French. 16.2% of the French and 16.2% of the Spanish exited, compared with 32.4% of the

Germans exited. This can be explained by the possibility that the French are more conservative, and the Germans are more radical. For Exited and Gender, we could see from Table 4 that there was a statistically significant difference in gender between those who exited and those who did not ($p<0.05$). From Table 2, 57.3% of those who did not quit were men, while only 44.1% of those who did. While 16.5 percent of men quit, 25.1 percent of women did not quit, which meant men were more conservative than women. For Exited and NumOfProducts, we could see from Table 3 that there was a statistically significant difference in the number of products between those who exited and those who did not ($p<0.05$). From Table 2, 69.2% of those who quit bought only one product, while 46.2% of those who stayed in bought only one product. However, only 17.1% of people who quit were those who bought 2 products, which meant buying more products indicates greater loyalty to the bank. For Exited and IsActiveMember, we could see from Table 4 that there was a statistically significant difference in how active the users were between those who exited and those who did not ($p<0.05$). From Table 2, 44.5% of those who did not quit were inactive members, while 63.9% of churners were inactive members. While 14.3% of active members quit, 26.9% of inactive members did. This indicates that active members are less likely to quit than inactive members. We can conclude that the more active users are, the higher their recognition and trust in the bank.

Table 2. Crosstab of Exited and other features [Owner-draw].

			X = Geography			X = Gender		X = NumOfProducts				X = IsActiveMember		Total
			0	1	2	0	1	1	2	3	4	0	1	
Exited	0	Count	4204	2064	1695	4559	3404	3675	4242	46	0	3547	4416	7963
		% within Exited	52.80%	25.90%	21.30%	57.30%	42.70%	46.20%	53.30%	0.60%	0.00%	44.50%	55.50%	100.00%
		% within X	83.80%	83.30%	67.60%	83.50%	74.90%	72.30%	92.40%	17.30%	0.00%	73.10%	85.70%	79.60%
		% of Total	42.00%	20.60%	17.00%	45.60%	34.00%	36.80%	42.40%	0.50%	0.00%	35.50%	44.20%	79.60%
	1	Count	810	413	814	898	1139	1409	348	220	60	1302	735	2037
		% within Exited	39.80%	20.30%	40.00%	44.10%	55.90%	69.20%	17.10%	10.80%	2.90%	63.90%	36.10%	100.00%
		% within X	16.20%	16.70%	32.40%	16.50%	25.10%	27.70%	7.60%	82.70%	100.00%	26.90%	14.30%	20.40%
		% of Total	8.10%	4.10%	8.10%	9.00%	11.40%	14.10%	3.50%	2.20%	0.60%	13.00%	7.40%	20.40%
Total		Count	5014	2477	2509	5457	4543	5084	4590	266	60	4849	5151	10000
		% within Exited	50.10%	24.80%	25.10%	54.60%	45.40%	50.80%	45.90%	2.70%	0.60%	48.50%	51.50%	100.00%
		% within X	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
		% of Total	50.10%	24.80%	25.10%	54.60%	45.40%	50.80%	45.90%	2.70%	0.60%	48.50%	51.50%	100.00%

Table 3. Result of Chi-Squared Test of Exited * Geography and Exited * NumOfProducts [Owner-draw]**.**

	Exited * Geography			Exited * NumOfProducts		
	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)
Pearson	301.255	2	0.000	1503.629	3	0.000

Chi-Square							
Likelihood Ratio	280.341	2	0.000		1399.07	3	0.000
Linear-by-Linear Association	236.43	1	0.000		22.865	1	0.000
N of Valid Cases	10000				10000		

Table 4. Result of Chi-Squared Test of Exited * Gender and Exited * IsActiveMember [Owner-draw].

	Exited * Gender					Exited * IsActiveMember				
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	113.449	1	0.000			243.760	1	0.000		
Continuity Correction	112.919	1	0.000			242.985	1	0.000		
Likelihood Ratio	113.044	1	0.000			245.83	1	0.000		
Fisher's Exact Test				0.000	0.000				0.000	0.000
Linear-by-Linear Association	113.438	1	0.000			243.736	1	0.000		
N of Valid Cases	10000					10000				

4 Independent Samples t Test

We use the Independent Samples t Test to compare two sample means to determine whether the population means of exited and non-exited are significantly different. Then we examine the effects of credit score, age, balance, tenure, and estimated salary on those who quit and those who don't. From Table 6, there was a significant difference in mean credit score, age and balance between people who quit and people who did not ($p < .001$). Those who quit have lower credit scores, are older, and have larger balances than those who don't. The average age for churned customers was 7.43 years old younger than the average age for non-churned customers, and the average balance for churned customers was 18,364 less than the average balance for non-churned customers (see Table 5). It is possible that when people have small balances, they don't care much about which bank they keep them in. Managing balances becomes more important only when balances are large, so more users will quit. People who did not quit had significantly higher credit scores, as a customer with a higher credit score was less likely to leave the bank due to more benefits in this bank.

Table 5. Basic information about the group comparisons [Owner-draw].

	Exited = 0				Exited = 1			
	Count	Mean	Std. Deviation	Std. Error Mean	Count	Mean	Std. Deviation	Std. Error Mean
CreditScore	7963	651.85	95.654	1.072	2037	645.35	100.322	2.223
Age	7963	37.41	10.125	0.113	2037	44.84	9.762	0.216
Balance	7963	72745	62848	704	2037	91109	58361	1293

Tenure	7963	5.03	2.881	0.032	2037	4.93	2.936	0.065
EstimatedSalary	7963	99738	57406	643	2037	101466	57912	1283

Table 6. Results of the Independent Samples t Test [Owner-draw].

	Levene's Test for Equality of Variances		T-test for Equality of Means					
	Equal variances assumed		Equal variances assumed			Equal variances not assumed		
	F	Sig.	t	df	Sig.(2-tailed)	t	df	Sig.(2-tailed)
CreditScore	5.5	0.019	2.71	9998	0.007	2.635	3051	0.008
Age	9.129	0.003	-29.767	9998	0.000	-30.419	3248	0.000
Balance	193.16	0.000	-11.936	9998	0.000	-12.471	3348	0.000
Tenure	1.639	0.201	1.4	9998	0.162	1.384	3114	0.166
EstimatedSalary	0.928	0.335	-1.21	9998	0.226	-1.203	3137	0.229

5 Factor Analysis

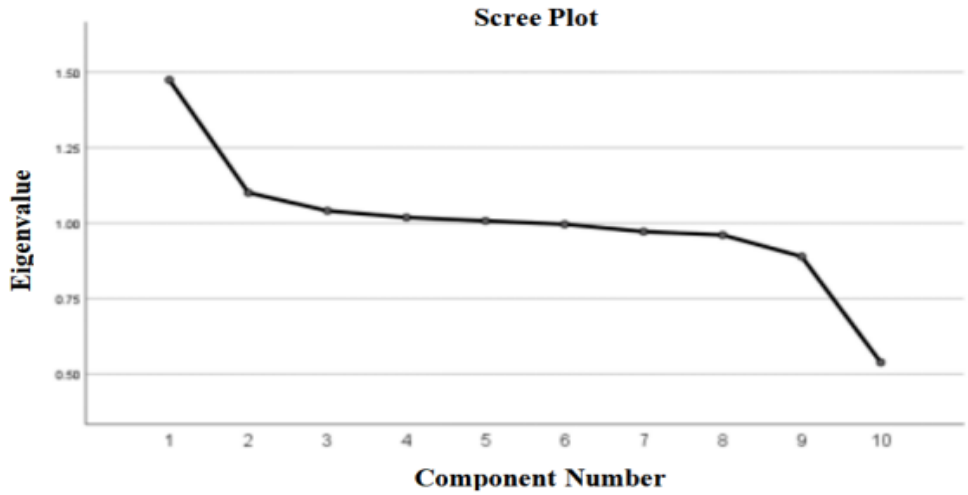

Figure 1. Scree plot [Owner-draw].

Through the above analysis, we found that some of the variables we examined were correlated with each other. Therefore, the method of factor analysis can be used to convert these variables into fewer comprehensive indicators that are unrelated to each other, and then investigate their impact on whether to quit.

The result of Barlett's tests of sphericity is 0.00<0.05, which means these 10 variables are related to each other. From Figure 1, we can see after the eighth component, the eigenvalues drop very low, so eight main components are extracted. As shown in Table 7, the proportion of variance that could be explained by these eight components is 85.719%. According to the rotated component matrix in Table 8, CreditScore and Geography are the main factors affecting whether the user quits or not, while estimated salary is the least important factor.

Table 7. Total Variance Explained [Owner-draw].

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings
	Total	% of Variance	Cumulative %	
1	1.474	14.744	14.744	1.294
2	1.101	11.008	25.753	1.171
3	1.041	10.409	36.162	1.086
4	1.019	10.19	46.351	1.013
5	1.007	10.075	56.426	1.004
6	0.996	9.964	66.39	1.003
7	0.972	9.721	76.111	1.000
8	0.961	9.608	85.719	1.000
9	0.889	8.895	94.614	
10	0.539	5.386	100	

Table 8. Communalities and Rotated Component Matrix [Owner-draw].

	Communalities		Rotated Component Matrix	
	Initial	Extraction	Coefficient	Component
CreditScore	1.000	0.969	0.911	1
Geography	1.000	0.847	0.676	1
Gender	1.000	0.902	0.939	2
Age	1.000	0.652	0.748	3
Tenure	1.000	0.973	0.724	3
Balance	1.000	0.724	0.948	4
NumOfProducts	1.000	0.883	0.984	5
HasCrCard	1.000	1.000	0.985	6
IsActiveMember	1.000	0.624	1.000	7
EstimatedSalary	1.000	0.999	1.000	8

Note: Extraction method: Principal Component Analysis.

6 Logistic Regression

Logistic regression is primarily a binary classification approach which is a simple and rapid way to predict results. We use logistic regression to predict customer churn rate using original factors and factors after dimension reduction respectively. As can be seen from Table 9, the accuracy of logistic regression prediction for those who quit and those who did not was 69.4% and 69.9% respectively. After the dimension reduction the figures were 65.0% and 66.4% respectively, which meant that after dimension reduction, the accuracy of prediction was slightly lower than that before dimension reduction, probably due to more compressed information. Overall, the accuracy of the prediction remained high even after the dimension reduction.

Table 9. Accuracy of Logistic Regression In Predicting Customer Churn Rate [Owner-draw].

Observed		Predicted (Original)			Predicted (after dimension reduction)		
		Exited		Percentage	Exited		Percentage
		0	1	Correct	0	1	Correct
Exited	0	5524	2439	69.4	5173	2790	65
	1	613	1424	69.9	684	1353	66.4
Overall Percentage				69.5			65.3

7 One-way ANOVA

As there may also be correlations between variables that influence whether to quit, the paper checked the relationship between the number of products and other factors. From Table 10 we conclude that Age and Balance are statistically significant. From Table 11 we can see the results of the post-hoc tests. In order to make the results more intuitive, we use the means plot to visualize the results of multiple comparisons output. The points on the chart are the averages of each group. It is easy to see from Figure 2 that the number of products is generally positively correlated with age. We conclude that the average age is significantly different for people who bought fewer products(1 or 2) and people who bought more products(3 or 4).

From Figure 3, it is shown that those who bought two products had the least balance, and those who bought one or four products had the most. There was no significant difference in the balance between those who bought one product and those who bought four. The reason may be that the ones who buy four products have enough money, and the ones who buy one product are saving money, so they buy less products. Those who buy two products have the lowest balance, and the corresponding age is also the lowest. This may be because young people don't save much money and have a smaller balance.

Table 10. ANOVA output [Owner-draw].

	Between Groups			Within Groups			F	Sig.
	Sum of Squares	df	Mean Square	Sum of Squares	df	Mean Square		
Credit Score	24844.854	3	8281.618	93384414.85	9996	9342.178	0.886	0.447
Age	16735.035	3	5578.345	1083095.813	9996	108.353	51.483	0.000
Balance	5272774279925	3	1757591426643	33657694043806	9996	3367116251	521.987	0.000
Estimated Salary	8514474466	3	2838158155	33062745910094	9996	3307597630	0.858	0.462
Tenure	19.379	3	6.46	83618.982	9996	8.365	0.772	0.509

Table 11. Multiple comparisons output [Owner-draw].

Dependent Variable:		Age					Balance				
(I) NumOfProducts	(J) NumOfProducts	Mean Difference (I-J)	Std. Error.	Sig.	95% Confidence Interval		Mean Difference (I-J)	Std. Error.	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound				Lower Bound	Upper Bound
1	2	1.919*	0.212	0.000	1.5	2.33	46673*	1181	0.000	44357	48989
	3	-3.522*	0.655	0.000	-4.81	-2.24	23094*	3650	0.000	15939	30248
	4	-6.010*	1.352	0.000	-8.66	-3.36	4819	7535	0.523	-9952	19589
2	3	-5.442*	0.656	0.000	-6.73	-4.16	-23579*	3659	0.000	-30753	-16406
	4	-7.930*	1.353	0.000	-10.58	-5.28	-41854*	7540	0.000	-56634	-27074
3	4	-2.488	1.488	0.094	-5.4	0.43	-18275*	8293	0.028	-34531	-2018

*. Sig.value of Mean Difference is 0.05.

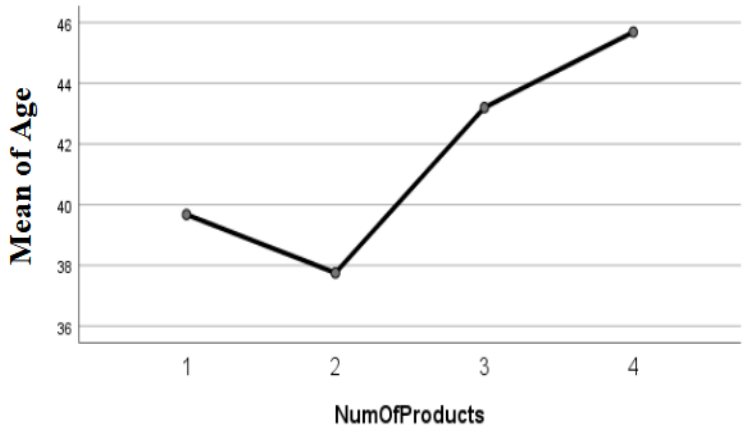

Figure 2. Age vs. NumOfProducts Means plot [Owner-draw].

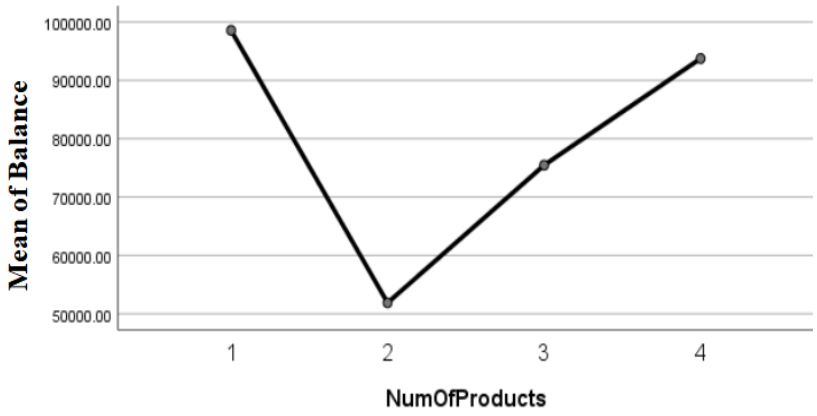

Figure 3. Balance vs. NumOfProducts Means plot [Owner-draw].

8 Conclusion

From the t-test, we find that those who quit have lower credit scores, are older, and have larger balances than those who don't. Through factor analysis, we find that country and balance are the main factors affecting customer churn, and salary and whether people have a credit card do not directly affect customer churn. From the logistic regression, this paper finds that the percentages of correct predictions are 69.5% and 65.3% respectively before and after selecting the main components. From one-way ANOVA we can also find that the number of products people buy can be affected by age and balance.

From the conclusion above, this paper provides a method to check the customer's exit intention in time. When the bank realizes that some users tend to stop the service, the bank can come up with some marketing strategies to engage customers and keep them in touch with the bank. In this way, the bank can successfully retain some customers, thereby reducing the customer churn rate. Of course, the research in this paper only considers the data set of a certain bank in a certain period, so the results obtained may be biased from those obtained by multiple banks in the real situation.

REFERENCES

[1] Neslin S A, Gupta S, Kamakura W, et al. Defection detection: Measuring and understanding the predictive accuracy of customer churn models[J]. Journal of marketing research, 2006, 43(2): 204-211.

[2] Ahn J H, Han S P, Lee Y S. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry[J]. Telecommunications policy, 2006, 30(10-11): 552-568.

[3] Hadden J, Tiwari A, Roy R, et al. Computer assisted customer churn management: State-of-the-art and future trends[J]. Computers & Operations Research, 2007, 34(10): 2902-2917.

[4] Richter Y, Yom-Tov E, Slonim N. Predicting customer churn in mobile networks through analysis of social groups[C]//Proceedings of the 2010 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2010: 732-741.
[5] Jahromi A T, Stakhovych S, Ewing M. Managing B2B customer churn, retention and profitability[J]. Industrial Marketing Management, 2014, 43(7): 1258-1268.
[6] Vafeiadis T, Diamantaras K I, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction[J]. Simulation Modelling Practice and Theory, 2015, 55: 1-9.
[7] Xia G, Jin W. Model of customer churn prediction on support vector machine[J]. Systems Engineering-Theory & Practice, 2008, 28(1): 71-77.
[8] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction[J]. Expert Systems with Applications, 2009, 36(3): 4626-4636.
[9] Xie Y, Li X, Ngai E W T, et al. Customer churn prediction using improved balanced random forests[J]. Expert Systems with Applications, 2009, 36(3): 5445-5449.
[10] Tsai C F, Lu Y H. Customer churn prediction by hybrid neural networks[J]. Expert Systems with Applications, 2009, 36(10): 12547-12553.
[11] Verbeke W, Martens D, Mues C, et al. Building comprehensible customer churn prediction models with advanced rule induction techniques[J]. Expert systems with applications, 2011, 38(3): 2354-2364.
[12] Verbeke W, Martens D, Baesens B. Social network analysis for customer churn prediction[J]. Applied Soft Computing, 2014, 14: 431-446.