Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2017 Proceedings | Wuhan International Conference on e-Business

Summer 5-26-2017

Customer Churn Prediction Based on BG / NBD Model

Huan Li
School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China, 16125355@bjtu.edu.cn

Zhongliang Guan
School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China, zlguan@center.njtu.edu.cn

Ying Cui
School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China, 16125351@bjtu.edu.cn

Customer Churn Prediction Based on BG / NBD Model

Huan Li[1], *Zhongliang Guan*[2], *Ying Cui*[3*]

[1]School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China

[2]School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China

[3]School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China

Abstract: With the rapid development of information technology, most enterprises have built e-commerce platform, which promotes the revolution of operation mode. The focus of competition gradually becomes the customers rather than the products under the increasingly fierce market competition of the E-commerce model. Because of the non-contractual relationship between the customers and the e-commerce platform, maintaining the stable customer relationship becomes the necessary condition for the e-commerce enterprises to get profit. So predicting the customer churn accurately plays an important role in the development of e-commerce enterprises. In this paper, the BG / NBD model is used to analyze the historical transaction records of an e-commerce platform in order to analyze and predict the purchase behavior of the existing customers, and identify the pre-losing customers, which helps the enterprises to implement the more effective strategies of CRM and restore the pre-loss customers timely.

Keywords: E-commerce; BG/NBD model; customer churn, predicting; strategies of restoring

1. INTRODUCTION

In recent years, with the rapid development of Internet technology, Internet thinking gradually assimilates into the various aspects, such as economy, politics, culture, education and so on, which causes the rapid rise of the e-commerce. E-commerce breaks through the limitation of the space and time to make customers get information conveniently that increases greatly the selectivity[1]. Because of the great freedom and convenience, there exists serious phenomenon of customer churn. Customers who just click in the page may become the competitor's resources. The customer relationship is different between the E-commerce and the entity enterprise. It is a kind of contractual relationship between the enterprises and customers in which the link is maintained by many factors. However, the electronic commerce is a kind of virtual network platform[2] and it is the non-contractual relationship with customers. The bond is very fragile between the electric business platform and customer. So how to manage the customer relationships is a challenge and opportunity for e-commerce business.

In the fierce market competition, the enterprise products and services more and more tend to homogeneity. Many enterprises have realized the problem and developed the customer relationship instead of products. The customer relationship management (CRM) becomes one of the most concerning issues in the enterprise management strategies[3]. Based on the characteristics of e-commerce and non-contractual relationship with clients, stable and long-term customer relationship affects the development of the enterprise[4]. Therefore, how to accurately predict customer churn should get e-commerce enterprise's attention[5,6]. The quantitative researches of e-commerce are relatively few at present because that there are too many factors to effect the customers' behaviors in the e-commerce.

2. THEORETICAL REVIEW

In the research field of customer non-contract random transaction behavior, Ehrenberg[7] first proposed the negative binomial distribution model (NBD model) to better fit the customer's trading behavior and more

* Corresponding author. Email: 16125355@bjtu.edu.cn (Huan Li) , zlguan@center.njtu.edu.cn (Zhongliang Guan), 16125351@bjtu.edu.cn (Ying Cui)

accurately predict the times of customers' shopping at a certain time. But the model assumes that the customer is continuously active. This assumption is clearly inconsistent with the reality against the customer life cycle theory that interest about products can change over time. Later, Morrison and Schmittlein[8] extended the hypothesis on the basis of the original model, that the customer's trading status in a period of forecasting period will be some changes, in 1987 proposed forecasting customer behavior probability model Pareto / NBD, and they improved the formation of the SMC model in 1994[9]. The model predicts the customer's activity level based on the customer's historical transaction data, and deduces the customer's loss problem according to the possibility of the customer's purchase. However, the accuracy of the model is usually reflected in the user group level, rather than the level of individual users, and the model of super-complex computing, generalization capability is weak[10]. Which limits the application of the model in practice and promotion. Subsequently, Peter S Fader and others[11] on the Pareto / NBD model in the calculation method to further simplify the proposed BG / NBD model used to simply predict the number of future customer transactions and customer activity, and with the Pareto / NBD model Compared with the predicted results, the calculation process is easier to achieve, but the accuracy is similar.

On the other hand, customer churn prediction methods based on data mining, such as decision tree, artificial neural network, regression, Bayesian classifier, are also widely used in customer churn prediction problems[12]. The above data mining methods have advantages and disadvantages, ignoring the real life customer life cycle changes in the purchase of data for the non-linear processing capacity is poor, generalization capacity is weak, it is difficult to other similar issues to promote[13].

3. BG/NBD MODEL

There is a non-contractual relationship between the electricity supplier and the customer, the customer's purchase behavior is a random event, the time interval and the number of purchases are uncertain, the customer buying behavior may occur at any time, and the buying behavior between the customer and the customer is not mutual influences. BG / NBD model is used to predict the number of non-contractual customers in the next period of time and customer activity, suitable for forecasting customer activity in e-commerce. The following are five assumptions of the BG / NBD model:

（1）While active, the number of transactions made by a customer follows a Poisson process with transaction rate λ. This is equivalent to assuming that the time between transaction is distributed exponential with transaction rate λ,

$$f\left(t_j \middle| t_{j-1};\lambda\right) = \lambda^{e^{-\lambda(t_j - t_{j-1})}}, t_j > t_{j-1} \geq 0 \tag{1}$$

（2） Heterogeneity in λ follows a gamma distribution with pdf,

$$f\ (\lambda|r,\alpha)\ = \frac{\alpha^{r}\lambda^{r-1}e^{-r\alpha}}{\Gamma(r)}, \lambda > 0 \tag{2}$$

（3） After any transaction, a customer becomes inactive with probability p. Therefore the point at which the customer "drops out" is distributed across transactions according to a (shifted) geometric distribution with pmf,

$$P(inactive\ immediately\ after\ jth\ transaction) = p(1-p)^{j-1},\ j = 1,2,3,\cdots\cdots \tag{3}$$

（4） Heterogeneity in p follows a beta distribution with pdf,

$$f(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \quad 0 \le p \le 1 \tag{4}$$

Where $B(a,b)$ is the beta function, which can be expressed in terms of gamma functions:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

（5） The transaction rate λ and the dropout probability p vary independently across customers.

The BG / NBD model divides the customer's historical purchase data into two periods based on time: the observation period and the forecast period. In order to test the accuracy of the forecast data, the historical data is divided into three periods: T1, T2 and the forecast period, as shown in figure 1, different customers have the same length of the complete observation period, the observation period T1 and the forecast period, and the observation period T2 and the first time the customer purchase behavior Time-dependent. The model requires the following basic data:

X: the number of transactions in the observation period T2 (the number of customer purchases);

t_x: the time of the last transaction in the observation period (0 <= t_x <= T);

T: the length of the observation period T2 (from the customer's first purchase behavior to the observation period deadline).

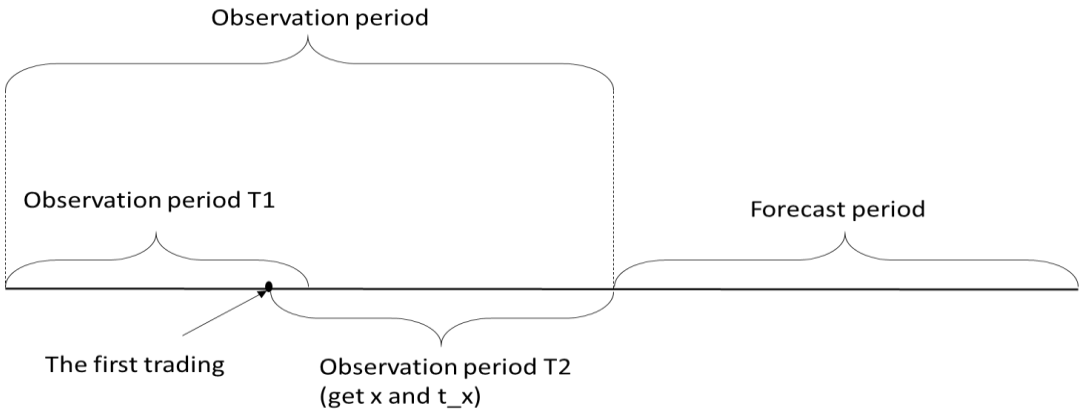

Figure 1. Time division diagram

4. MODEL CONSTRUCTION

4.1 Data preprocessing

By the impact of human influence and the environment, the reality of the data are mostly flawed and impure. The data cleaned-up is the high-quality. The data with low quality and low value density can greatly reduce the computational efficiency and impact the calculation result of the algorithm[14]. Therefore, the data must be preprocessed to improve the data quality and adjust the data structure before data mining.

This data is a website of the transaction data, including the 12,146,637 shopping records of 100000 customers from April 1, 2010 to March 31, 2011. Then the useful fields in the data: the customer ID, the purchase time, and the consumption amount are extracted, and the results are shown in Table 1. The consumption records from the 2000 customers who are randomly selected are regarded as a research sample for the operability of research. The records from April 1 to June 16, 2010 are used as observation data to estimate the parameters needed by the model and prepare for the follow-up model implementation. Based on the forecast

period from June 17 to December 31, 2010, the forecasted purchase data are fitted to the actual purchase data to analyze the accuracy of the BG / NBD model.

Table 1. Data table after extraction of valid field (only part)

customer_id	visit_date	visit_spend	customer_id	visit_date	visit_spend
2	2010/4/1	5.97	64823	2010/10/11	82.31
2	2010/4/6	12.71	64823	2010/10/30	131.82
2	2010/4/7	34.52	64823	2010/11/4	13.14
12955	2010/7/14	109.94	64823	2010/11/5	91.18
12955	2010/7/17	9.69	64823	2010/11/13	12.42
12955	2010/7/20	12.6	64823	2010/11/15	48.35

In order to build the model, we need to collate the collected sample data and remove the useless, invalid and fuzzy data. The records are aggregated and counted to get model data: x, t_x, T. And the results are shown in Table 2.

Table 2. Preprocessed data(only part)

ID	x	t_x	T	ID	x	t_x	T
197	10	11.00	11.00	697	8	10.43	9.57
279	15	10.57	11.86	793	18	10.29	11.86
349	29	10.43	11.29	867	13	11.00	11.86
416	11	9.71	11.29	915	10	10.86	10.14
430	7	9.71	11.71	939	13	10.71	11.57
568	12	10.14	11.14	941	13	11.00	11.00
694	14	10.29	11.00	944	20	11.00	11.57

4.2 Parameter estimation

Since the later use of the program to solve the LL solution for the maximum likelihood function and the best value of the four parameters, you can first four parameters r, alpha, a and b initialized as: r = 1.0, alpha = 1.0, a = 1.0, B = 1.0, calculate the LL value through the formula (5), and then use the "Solver" tool in excel to get the maximum likelihood estimation of the four model parameters by maximizing the log-likelihood function. The results are shown in the following table 3:

$$\Gamma(r,\alpha,a,b|X=x,t_x,T)=\frac{\Gamma(r+x)\alpha^r}{\Gamma(r)}\cdot\frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}\cdot\left(\left(\frac{1}{\alpha+T}\right)^{r+x}+\delta_{x>0}\left(\frac{a}{b+x-1}\right)\left(\frac{1}{\alpha+t\_x}\right)^{r+x}\right) \quad (5)$$

Table 3. Parameter estimation

LL	r	alpha	a	b
-13710.01177	3.908368417	2.206899333	0.000256398	1.407709477

4.3 Repetitive purchase expectations

By using the four parameter values r, alpha, a, and b, we can predict the expected future repeat purchases of 2000 customers in time t by using formula (6).

$$E(X(t)|r,\alpha,a,b)=\frac{a+b-1}{a-1}\left[1-\left(\frac{\alpha}{\alpha+t}\right)^r 2F_1\left(r,b;a+b-1;\frac{t}{\alpha+t}\right)\right] \quad (6)$$

Table 4. Expected repeat purchase(only part)

In t days	Repeat purchase times	In t days	Repeat purchase times
1	0.069130079	6	1.403570641
2	0.250206239	7	1.675101152
3	0.553226528	8	1.942959965
4	0.844519322	9	2.207980781
5	1.12723302	10	2.470783962

4.4 Individual customer forecasts

Using the customer's past behavioral information and parameter estimates, use formula (7) to predict the future purchases of a particular customer, and the results are shown in Table 5 and figure 2.

$$E(Y(t)|X = x, t_x, T, r, \alpha, a, b) = \frac{a+b+x-1}{a-1} \cdot \frac{\left[1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} {}_2F_1\left(r+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t}\right)\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}} \tag{7}$$

Table 5. Comparison of actual and predicted values(only part)

Customer ID	Forecast values	Actual values	Customer ID	Forecast values	Actual values
197	18	17	915	12	18
279	31	22	939	22	20
349	38	40	941	16	21
416	17	18	944	28	28
430	17	13	985	27	41
568	25	19	1288	17	20
694	21	22	1446	46	35
697	18	16	1477	12	28
793	37	25	1484	18	18
867	19	20	2429	42	41

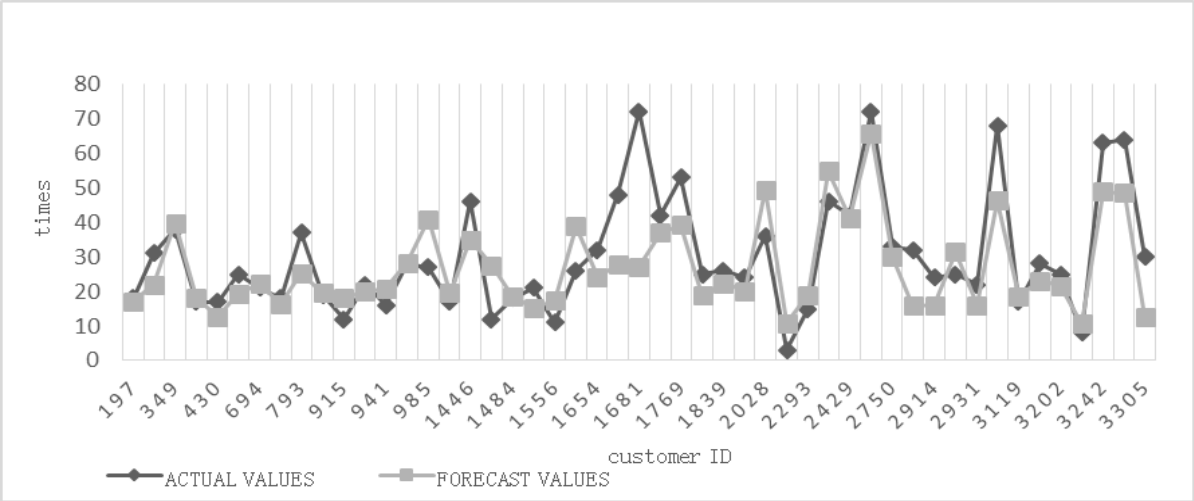

Figure 2. The line chart of the actual values and predicted values for each customer (only part)

The purchase times in the future period are got by calculating. When the times of real purchases and predicted purchases are closer for each customer, the calculation result of the formula (8) is closer to 0%, and the model is more accurate. The percentage of each customer is placed in a different result range. The results are shown in the table 6. 74.90% of the customers are in the range of 0% to 40%. The percentage of customers is very small in the range of 45% and above. So it proves that the predictability of BG/NBD model is better.

$$\text{range} = \frac{\text{predicted purchase times} - \text{actual purchase times}}{\text{actual purchase times}} * 100\% \tag{8}$$

Table 6. The table of percentage

Range	Number Of Customers	Percentage	Cumulative Percentage	Range	Number Of Customers	Percentage	Cumulative Percentage
0-5%	207	10.35%	10.35%	50%-55%	61	3.05%	87.45%
5%-10%	191	9.55%	19.90%	55%-60%	32	1.60%	89.05%
10%-15%	185	9.25%	29.15%	60%-65%	35	1.75%	90.80%
15%-20%	212	10.60%	39.75%	65%-70%	19	0.95%	91.75%
20%-25%	195	9.75%	49.50%	70%-75%	16	0.80%	92.55%
25%-30%	197	9.85%	59.35%	75%-80%	12	0.60%	93.15%
30%-35%	174	8.70%	68.05%	80%-85%	11	0.55%	93.70%
35%-40%	137	6.85%	74.90%	85%-90%	14	0.70%	94.40%
40%-45%	111	5.55%	80.45%	90%-95%	9	0.45%	94.85%
45%-50%	79	3.95%	84.40%	95%以上	5	0.25%	95.10%

5. PREDICTION ANALYSIS OF CUSTOMER CHURN

5.1 The importance of churn prediction

Customer churn will make a significant negative impact to enterprises, not only lose the potential profit, but also lose the customer's trading opportunities affected[15]. And some loss of customers is inevitable, enterprises to achieve "zero customer loss" is unrealistic[16]. The cost of developing a new customer is far greater than the cost of maintaining an old customer, the loss of customers affect the development of new customers and it is not conducive to the development and growth of enterprises, so managers should take active measures to improve customer satisfaction, Thereby improving customer loyalty and control customer churn rate in a lower range.

It is possible to save the loss of customers. Studies have shown that sales to the loss of customers, every four customers will be successful one, and sales to potential customers and target customers every 16 customers have a success. Therefore, the return for the loss of customers is easier successful than for new customers[17]. Therefore, the analysis of the reasons for the loss of customers and make reasonable recommendations can extend the customer's life cycle and create more profits.

5.2 Advice

There are two aspects for the customers churn, including the electricity business and the customers. It is difficult for the electric business to change the customer's own reasons, but it is possible to influence customer behavior by improving their own building[18]. E-commerce platform is the best way for the enterprises to communicate with customers, which not only can provide services to customers, but also enable enterprises to get more information. Customer relationships becomes easier to maintain through e-commerce platform. There are some advice to maintain better customer relationships.

5.2.1 Website building

Web site building should consider the following four aspects. First, the aesthetics of the site: a beautiful page

attracts customers' eyes and the desire to browse; Second, the convenience: help customers quickly find the services or goods needed, and obtain the most important information with the least time. Third, the security: e-commerce are mostly used online payment, high security to enable customers to buy at ease, otherwise, low security will give enterprises and customers a huge economic losses. Finally, the performance: to ensure the speed of the site and system's stability, in order to make the site more efficient operation.

5.2.2 Communication with customer service

That customers communicate with customer service is the only way to interact with the website, so the communication ability of customer service determines the customer's evaluation of a site. Customer service should provide multi-angle and full range services to customers with the "customer first" philosophy and leaves a good impression and comfortable shopping experience to customers, in order to improve customer satisfaction and loyalty. Secondly, online customer service not only can reduce the operating costs of enterprises, but also can collect comprehensively and analysis of customers' information, which transforms customers' information into enterprise knowledge.

5.2.3 Data mining of customers' information

It is enormous for the daily registration volume, searching volume and transactions volume in e-commerce platforms, which brings the massive data. The data is the valuable assets of the enterprise. We can analyze customers' purchase behavior and personal preference to classify customers by using data mining technology, in order to make website easily push information to meet customers' individual needs[19]. It not only gives the site a good reputation, but also helps enterprise earn profits while attracting more customers.

6. CONCLUSIONS

In an increasingly competitive environment, mining and maintaining high-value customers which have a long-term stable relationship with enterprises is one of the effective ways to help business get the sustainable competitive advantages and help managers have a more clear Cognitive of their own level of customer value, at the same time, it can position customers of high value of and make limited resources to get reasonable optimization configuration.

However, there are two shortcomings in this study: on the one hand, the customer buying behavior of this article only forecasts the number of consumers to buy, while the amount of consumption and the types of items purchased without research, the contents of the study is not extensive. On the other hand, the data used in this study is the recorded data which have occurred, while the potential customers who have visited the web pages and not make a purchase are neglected, and the research on the potential customers has great research value.

ACKNOWLEDGEMENT

This research was supported by the China Railway Corporation Science and Technology Research and Development Project: 2016D001-A and the Beijing to Support the Central University in Beijing to Build Together Project: B13H100050.

REFERENCES

[1] Fang ling(2016). Analysis on the Application of Big Data in Enterprise Marketing under New Situation[J]. Shangye jingji,11:92-93 (in Chinese).

[2] Zhang Yin,Chen Min,Liao Xiao-Fei(2013).Big Data Applications:A Survey[J]. Journal of Computer Research and Development,50:216-233 (in Chinese).

[3] Ma Te, Guo Yan-Hong, Dong Da-Hai(2011). Review and Prospect on Customer Lifetime-Value. Science-Technology and Management,13(6):89-93.

[4] Li Chunqing, Wang Suqiang. The application of randm models in forecasting commercial bank's saving deposits customer purchase behavior[C]. The Eight Wuhan International Conference on E-Business,2009,Vol 2:1316-1325.

[5] Shmueli G, Koppius OR (2011), Predictive analytics in information systems research[J]. MIS Quart,35(3):553–572.

[6] Indranil Bardhan, Jeong-ha (Cath) Oh, Zhiqiang (Eric) Zheng(2014).Predictive Analytics for Readmission of Patients with Congestive Heart Failure[J],Information Systems Research，26(1):1-21.

[7] Ehrenberg A S C .The Pattern of Consumer Purchases [J]. Applied statistics, 1959, 8:26-41.

[8] Sehmittlein D C, Morrison D G, Colombo R. Counting Your Customers：Who Are They and What Will They Do Next?[J]. Management Science,1987,33(1).

[9] Sehmittlein D C, Peterson R A, Customer Base Analysis：An Industrial Purchase Process Application[J]. Marketing Science, 1994,13(1).

[10] Shaohui Ma, Jinlan Liu(2006). Empirical and Applied Research of Pareto / NBD Model [J]. Management Science,6(10)：45-49 (in Chinese).

[11] Fader PS, Hardie BGS, Lee KL (2005), "Counting your customers"the easy way: An alternative to the Pareto/NBD model[J]. Marketing Sci,24(2):275–284.

[12] Chen Zhang-Liang(2009). Application and Research of Forecasting Decision Model Based on Data Mining. China Management Informationization,12(1):57-59 (in Chinese).

[13] Ying Wei-Yun,Lan Nan,Li Liu(2008). Customer Churn Prediction Algorithm for Unbalanced Data. Systems Engineering,26(11):99-104 (in Chinese).

[14] Jiawei Han,Micheline Kamber,Jian Pei(2012). Data Mining Concepts and Techniques [M]. Machinery Industry Press,2012 (in Chinese).

[15] D Jain, SS Singh. Customer lifetime value research in marking: A review and future directions, Journal of Interactive Marketing, 2002.

[16] Yao Pan(2008). Research on Customer Lifetime Value of Enterprise E - Commerce Websites. Trade Unions' Tribune,14(6):84-85.

[17] Xia Guo-en(2010). Research on Current Situation and Development of Customer Churn Prediction. Application Research of Computers,27(2):413-416.

[18] Chen Wei-Hua(2010). Research on Application of Multi - Channel Integration Strategy in Customer Relationship Management. Commercial Research,399:68-71 (in Chinese).

[19] Feng Zhi-Yan,Guoxun-Hua,Zeng Da-Jun(2013). On the Research Frontiers of Business Management in the Context of Big Data [J]. Journal Of Management Sciences In China,16(1): 1-9 (in Chinese).