Analysis of the Characteristic Behavior of Loyal Customers on a Golf EC Site

Yue Su[1(✉)], Kohei Otake[2], and Takashi Namatame[3]

[1] Graduate School of Science and Engineering, Chuo University,
1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
al5.xwsr@g.chuo-u.ac.jp
[2] School of Information and Telecommunication Engineering, Tokai University,
2-3-23, Takanawa, Minato-ku, Tokyo 108-8619, Japan
otake@tsc.u-tokai.ac.jp
[3] Faculty of Science and Engineering, Chuo University,
1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
nama@indsys.chuo-u.ac.jp

Abstract. In recent years, with expansion and growth of electronic commerce (EC) market, it is expected that the competition of getting customers will be fierce. The EC company is required to find new customers who have the potential of becoming loyal customers as soon as possible. In this study, we analyze customers' behavior using customer membership information data, purchase records data and web access logs data on a golf EC site. Firstly, we evaluate the loyalty of customers using RFM analysis to divide customers into the loyal and general ones. Next, we perform logistic regression to discriminate loyalty by using the first-time purchase and browsing behaviors. Through our analysis, we built a model to predict loyal customers and clarify the characteristic behaviors of high loyal customers.

Keywords: Customer behavior · RFM analysis · Logistic regression

1 Introduction

In recent years, electronic commerce (hereinafter called "EC") continues to evolve at a rapid pace [1]. With expansion and growth of the EC market, it is expected that the competition of getting customers will be fierce. Choosing appropriate target customers is very important for expanding sales and improving profitability.

Therefore, the EC company is required to find new customers who have the potential of becoming loyal customers as soon as possible. Here, the first purchase date can be considered a point. We look forward to the common behaviors of these customers in their initial purchases. Customers raise customer satisfaction, so that companies improve sales and profits. It is desirable to have such a relationship between both sides that can benefit from each other.

Figure 1 shows the framework of customers hierarchy. First, customers visit the website. Upper-level customer purchase frequently and high amount. Then, finding these loyal customers and developing new loyal customers are very important strategies for the retail company.

G. Meiselwitz (Ed.): HCII 2019, LNCS 11579, pp. 475–485, 2019.
https://doi.org/10.1007/978-3-030-21905-5_37

Fig. 1. Framework of customer hierarchy

In this study, we focused on new customers and the purpose is to clarify the characteristic behaviors of high loyal customers using customer's membership information data, purchase data and access historical data.

2 Datasets

We target on a general electronic commerce website (hereinafter called "EC site") relating to golf. The EC site provides some services such as EC of golf equipment, reservations for golf courses, manage golf score, etc. From among these services, we used the following data.

- Customer information data (age, sex, registration date, etc.)
- Purchase history data (category of purchase items, purchase date, whether purchased item is brand-new or secondhand, etc.)
- Access history data (log in date and time, URL of access page, URL of referrer page, etc.)

The category name of the product included in the purchase data is shown in Table 1.

Table 1. Category name of item

Category	Item
Men's wear	Tops for men, pants for men, etc.
Lady's wear	Tops for women, pants for women, etc.
Golf club	Putter, iron, etc.
Accessory	Golf ball, golf glove, etc.
Other	Calendar etc.

Target Customer

In this study, we analyzed 5,553 customers who purchased for the first time from May 1, 2015, to July 30, 2015, and purchased more than twice a year from the initial purchase date. We exclude the customer who has passed for more than one year from registration.

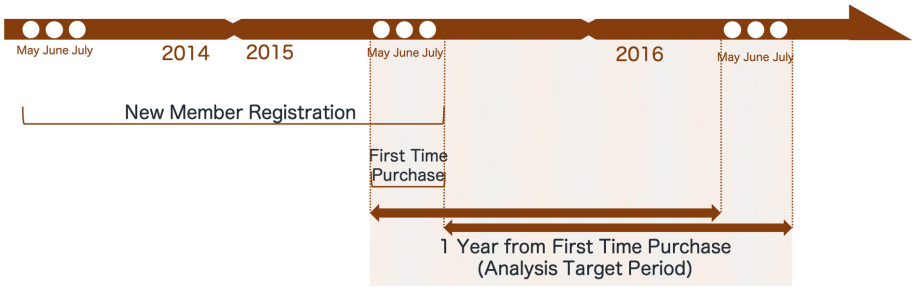

Fig. 2. Target period

In Fig. 2, we show the target period used in this research.

Explanatory Variables

We considered the impact factors to the first purchase using the above data. Based on the result, we created the explanatory variables such as customer's member information (5 variables), purchasing behavior at the time of initial purchase (11 variables) and web browsing behavior at the initial purchase date (13 variables) [4].

Details of the explanatory variables are shown in Tables 2, 3 and 4.

Table 2 presents demographic variables created by membership information data.

Table 2. Demographic variables used in the model construction.

Variable name	Data type
Gender (male = 1, female = 0)	0 or 1
Age	Integer
Whether customer lives in capital or not	0 or 1
Whether the member registration date matched the initial order date or not	0 or 1
Whether customer updated membership or not	0 or 1

Table 3 demonstrates purchasing behavior variables created by purchase data.

Table 3. Purchasing behavior used in the model construction.

Variable name	Data type
Total number of items purchased at the first-time purchase	Integer
Total amount at the first-time purchase	Integer
Average amount at the initial order date	Integer
Whether customer purchased lady's wear item at the initial order date or not	0 or 1
Whether customer purchased men's wear item at the initial order date or not	0 or 1
Whether customer purchased golf club item at the initial order date or not	0 or 1
Whether customer purchased accessory item at the initial order date or not	0 or 1
Whether customer purchased the other item at the initial order date or not	0 or 1
Whether customer purchased brand-new item at the initial order date or not	0 or 1
Whether customer purchased secondhand item at the initial order date or not	0 or 1
Whether customer purchased sale item at the initial order date or not	0 or 1

Table 4 shows Access History Variables created by web browsing data.

Table 4. Access history variables used in the model construction.

Variable name	Data type
Average login time of all session at the initial order date	Integer
Number of log in at the initial order date	Integer
Average number of page view at first purchase date	Integer
Whether browsing golf lesson page or not	0 or 1
Whether browsing golf course reservation page or not	0 or 1
Whether browsing golf movie page or not	0 or 1
Whether browsing golf news page or not	0 or 1
Whether browsing golf style page or not	0 or 1
Whether browsing golf second-hand goods shop page or not	0 or 1
Whether browsing golf gear page or not	0 or 1
Whether browsing golf new goods shop page or not	0 or 1
Whether browsing management golf score page or not	0 or 1
Whether browsing golf event page or not	0 or 1

3 Analysis of Loyal Customer

In this study, we analyze the behavior of the initial order date for customers who purchase more than once a year using customer membership information data, purchase records data and web access logs data on a golf EC site.

As an analysis, firstly we evaluated customer loyalty for new customers by RFM analysis. We determined customers' loyalties with three purchasing behavior indicators (Recency, Frequency, Monetary) and categorized them as loyal customers and general customers based on this.

Next, we created variables related to the initial purchase and exploratory behavior and constructed a discrimination model of customer loyalty by logistic regression analysis. Through these analyses, we worked to grasp the characteristics of customers with high loyalties at the initial order date.

3.1 RFM Analysis

RFM analysis is one of the most common approaches in database marketing. RFM analysis is a proven marketing model for behavior-based customer segmentation. It groups customers on recency, frequency, and monetary value can indicate customer.

RFM analysis segments customers on recency, frequency, and monetary value can indicate customer We evaluated the loyalty of customers using RFM analysis to divide customers into loyal and general ones [2]. Commonly, the F in RFM analysis is determined by the number of purchases. Here, we defined F by the total number of logins instead of the number of purchase, because frequent browsing behavior is also relates to customer's loyalty for the website.

RFM stands for the three dimensions:

- Recency: Period since last purchase
- Frequency: Total number of logins within the period
- Monetary: Amount of purchase within the period

The approach to RFM is to assign a score for each dimension on a scale from 1 to 5. The maximum score represents the preferred behavior.

Customers are divided into five scales equally for each of recency, frequency, monetary. The maximum score of RFM stands for the three dimensions:

- Recency: The maximum score (5) represents the shortest number of days that have passed since the customer last purchased within a year.
- Frequency: The maximum score (5) represents the longest number of logins within a year.
- Monetary: The maximum score (5) represents the highest value of all purchases within a year.

3.2 Binomial Logistic Regression

The purpose of this study is to predict the high loyal customers by using the initial purchase and browsing behaviors. When the objective variable to be predicted is binary, binomial logistic regression models are often used.

The Binomial logistic regression model is a type of classifier that performs class discrimination. By interpreting significant explanatory variables in the constructed model, it is possible to clarify the characteristics that affect the presence or absence of

repurchase. In the binomial logistic regression analysis, the customer's repurchase probability p_i is expressed by the following equation [3].

$$p_i = \frac{\exp\left\{\sum_{j=0}^{m} \beta_j X_{ij}\right\}}{1 + \exp\left\{\sum_{j=0}^{m} \beta_j X_{ij}\right\}} \tag{1}$$

X_{ij} : Factors affecting repurchase ($X_{i0} = 1$)
β_j : Parameters for each explanatory variable (β_0 is intercept)

We prepared variables related to demographic variables, initial purchase behavior and exploratory behavior (Tables 2, 3 and 4) and constructed a discrimination model of customer loyalty by binomial logistic regression analysis. Here, we label the loyal customer as 1, and the general customer as 0.

In logistic regression analysis, when the explanatory variable is excessive, it may be difficult to interpret the regression equation, or the versatility of prediction of the objective variable may decrease. It may occur multicollinearity problem due to some variables have a high correlation. Therefore, in this study, to select true effective variables, we used stepwise method based on Akaike's Information Criterion (AIC).

In order to confirm the discrimination accuracy of the model, we divided the data used in the logistic regression analysis into two groups (Group A, Group B), and performed a 2-fold cross-validation method.

The cross-validation method is mainly used in settings where the purpose is a prediction, and one wants to estimate how accurately a predictive model will perform in practice.

In order to confirm the prediction accuracy of the constructed model, we performed hold-out validation by using the training data and test data. Specifically, we created a confusion matrix like Table 5 and we calculated prediction accuracy of the constructed model by using the following equations.

Table 5. Confusion matrix

		Predicted class	
		Positive	Negative
Actual class	True	True Positive (TP)	True Negative (TN)
	False	False Positive (FP)	False Negative (FN)

Accuracy (ACC): Percentage of the total number correctly predicted among the total number predicted.

$$\mathrm{ACC} = \frac{TP + TN}{FP + FN + TP + TN} \tag{2}$$

Precision (PRE): Percentage of the total number that is a positive class actually among the total number predicted positive class.

$$\mathrm{PRE} = \frac{TP}{TP + FP} \tag{3}$$

Recall (REC): Percentage of the total number predicted positive class among the total number that is a positive class actually

$$\mathrm{REC} = \frac{TP}{FN + TP} \tag{4}$$

F-measure: harmonic mean of PRE and REC

$$\text{F-measure} = 2 \times \frac{PRE \times REC}{PRE + REC} \tag{5}$$

4 Results and Discussions

In this section, we show our analyzing results and discuss them.

4.1 RFM Analysis

Customers were divided into five equal scales equally for each of recency, frequency, monetary. Categories for each attribute of RFM are shown in Table 6.

Table 6. Categories for each attribute of RFM

Score	Recency (/days)			Frequency (/times)			Monetary (/yen)		
5		~	34	326	~		60950	~	
4	35	~	97	160	~	325	27667	~	60949
3	98	~	198	77	~	159	14001	~	27666
2	199	~	307	30	~	76	6801	~	14000
1	308	~			~	29		~	6800

Although the number of target customers in this research was 5,553, at the time of model construction, we randomly sampled the number of general customers by setting the number equal to the number of loyal customers.

The number of datasets (Group A, Group B) used in these model constructions are shown Table 7.

Table 7. Datasets used in prediction model

	Target customers	Analysis data		
		Group A	Group B	Total
Loyal customers	961	480	481	961
General customers	4592	480	481	961
Total	5553	960	962	1922

4.2 Binomial Logistic Regression

In each iteration, the model will be fit to one group of the data, and used to predict the other group.

We built two models that predicts loyal customer for the customers using binomial logistic regression analysis with AIC based the stepwise selection method.

The evaluation indicator for confirming the prediction accuracy are shown Table 8.

Table 8. Evaluation indicator of model for customers (%)

	Training data: Group. A	Training data: Group. B	Average
ACC	82.22%	82.40%	82.31%
PRE	84.60%	83.16%	83.88%
REC	78.79%	81.25%	80.02%
F-measure	81.59%	82.19%	81.89%

Table 9. Partial regression coefficients.

Explanatory variables	Partial regression coefficient	
(Intercept)	−5.460	***
Whether customer updated membership or not	0.798	**
Whether the member registration date matched the initial order date or not	0.754	***
Total number of items purchased at the initial order date	2.590	***
Average amount at the initial order date	0.736	***
Whether customer purchased lady's wear item at the initial order date or not	0.405	
Whether customer purchased men's wear item at the initial order date or not	0.744	**
Whether customer purchased golf club item at the initial order date or not	0.887	***
Whether customer purchased accessory item at the initial order date or not	0.709	*
Whether customer purchased sale item at the initial order date or not	0.724	*
Average login time of all session at the initial order date	0.179	.
Whether browsing golf gear page or not	0.449	

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Both models are over accuracies. Since the conventional researches on the EC site had the accuracies about 60%, it can be said that this research gained sufficient prediction accuracy.

The accuracy is high when group A is used as training data. Table 9 shows the partial regression coefficients.

There are 11 variables selected from 29 candidate variables.

From Table 9, we can see that variables created from purchase data are selected much. In addition, the confusion matrix for the test data of this model is shown in Table 10.

Table 10. Confusion matrix of model for customers

		Predicted class	
		Positive	Negative
Actual class	Positive	390	90
	Negative	79	401

4.3 Discussions

We selected the explanatory variables which the coefficient of the significant probability of less than 0.05. There are 8 explanatory variables selected (Table 11).

Table 11. Estimated value of selected partial regression coefficient

Explanatory variables	Partial regression coefficient
Total number of items purchased at first purchase	2.590
Whether customer purchased golf club item at the first purchase or not	0.887
Whether customer updated membership or not	0.798
Whether the member registration date matched the initial order date or not	0.754
Whether customer purchased men's wear item at the initial order date or not	0.744
Average amount at first-time purchase	0.736
Whether customer purchased sale item at the initial order date or not	0.724
Whether customer purchased accessory item at the initial order date not	0.709

Overall, since all the partial regression coefficients are positive numbers, it was found that the higher the value of all the selected variables, the more likely to become loyal customers.

In all the variables, total number of items purchased at the initial order date is the highest partial regression coefficient. It seems that the loyalties will be improved by raising customer satisfaction such as giving coupons or gifts to customers with high purchase quantities at the initial order date.

Since partial regression coefficient of "Whether the member registration date matched the initial order date or not" is positive as well, we considered that customers who were interested for a long time and took a long time to purchase. From this result, it seems that recommendations of similar items promote purchase.

It seems that recommending the items of men's wear, golf club, accessory on sale items to the customers registered as a member and did not purchase leads to promotion of purchasing.

It is considered that it is necessary to improve the loyalty of customers by recommending goods to be compared without limiting prices at the initial purchase.

4.4 Verification

We verified with the data of the same period two years later using the prediction model built this time. The results are shown in Tables 12 and 13.

Table 12. Confusion matrix of model for customers

		Predicted class	
		Positive	Negative
Actual class	Positive	894	135
	Negative	946	3483

Table 13. Evaluation indicator of model for customers (%)

ACC	PRE	REC	F-measure
80.19%	48.59%	86.88%	62.32%

Here, although high prediction accuracy was obtained, the precision was low. It is considered that this model distinguishes loyal customers and general customers well, but it could not confirm loyal customers correctly.

5 Conclusion

In this study, we determined customers' loyalties by RFM analysis and constructed a discrimination model of customer loyalty by logistic regression analysis to find characteristic behavior of loyal customers on a golf EC site.

Through our analyses, we built a useful model to predict loyal customers using the web access logs and purchase records data at initial purchase on a golf EC site. As a result, we could clarify the initial purchase and browsing behavior of high loyal customers and tried to propose marketing measures. Even for the data after two years, the model we made this time got a high accuracy.

However, we are conducting a prediction from the data at one point in this study. It is important to check the prediction accuracy of loyal customers by analyzing the data at the transition time.

Acknowledgment. We thank Golf Digest Online Inc. for permission to use valuable datasets and for useful comments. This work was supported by JSPS KAKENHI Grant Number 16K03944 and 17K13809.

References

1. Ministry of Economy, Trade and Industry: Foundation for Data-Driven Society in Japan (Market Survey on Electronic Commerce) (2018). (in Japanese)
2. Nakamura, H. (ed.): Market Segmentation - Discovery of Sales Opportunities Using Purchase History Data, Hakuto Shobo (2008). (in Japanese)
3. Yamashita, H., Suzuki, H.: Analysis of purchasing behavior of customers focusing on sale items: logistic regression analysis with consideration of clustering of binary data. Commun. Oper. Res. Soc. Jpn. **60**(2), 81–88 (2015). (in Japanese)
4. Sato, Y., Namatame, T., Otake, K.: Analysis of the characteristics of repeat customer in a golf EC site. In: International Conference on Social Computing and Social Media, SCSM 2017: Social Computing and Social Media. Human Behavior, pp. 223–233 (2017)