Analysis of the Characteristics of Repeat Customer in a Golf EC Site

Yusuke Sato[1(✉)], Kohei Otake[2], and Takashi Namatame[2]

[1] Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
al3.bthf@g.chuo-u.ac.jp

[2] Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
{otake,nama}@indsys.chuo-u.ac.jp

Abstract. In recent years, acquisition of repeat customers is emphasized for EC sites. On the other hand, the defection rate from the first purchase to the second purchase is the highest. There are much attention to acquire the repeat customers in the EC sites in this situation. The purpose of this study is to clarify factors necessary for acquiring repeat customers. Especially, we construct models that predict whether or not to repurchase within a certain period using membership information variables, purchase behavior variables and web browsing behavior variables. Using these models, we extract characteristics relate to presence or absence of repurchase and propose marketing measures to promote to repurchase.

Keywords: Consumer behavior · Repeat customer · Logistic regression

1 Introduction

Due to recent advances in the current internet environment, the market size of EC (Electronic Commerce) market that trades products on the internet is in rapid expansion. In addition, competition for customer acquisition is occurring and acquisition cost of new customers is rising in this market. Therefore, acquisition of repeat customers who use the EC site continually is regarded as important. In the EC site market, there is a feature customer defection rate from the first purchase to the second purchase is the highest, and the subsequent rate customers who have purchased for the second time decreases [1]. Therefore, when considering acquisition of repeat customers, it is important to prevent separation from first purchase to second purchase. The transition of customer defection rate from the first purchase to the multiple purchases is shown below (Fig. 1) [2].

Hence, it is important to understand the behavior of repeat customers, and it allows decreasing of defection customers [3–6]. Especially, the target EC site of this study provides the system of make reservations for golf courses in addition to purchasing golf supplies. So, customer retention on the EC site without limiting total number and purchase price of items purchased at second purchase brings sales increase as a whole.

G. Meiselwitz (Ed.): SCSM 2017, Part I, LNCS 10282, pp. 223–233, 2017.
DOI: 10.1007/978-3-319-58559-8_19

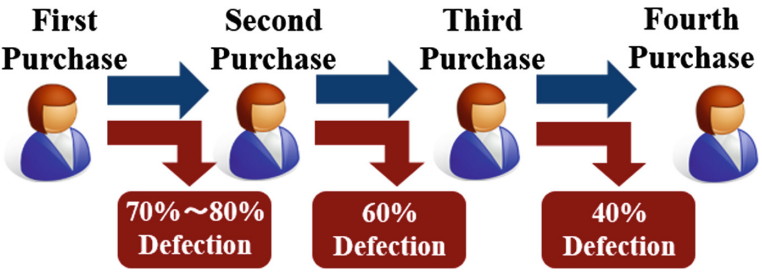

Fig. 1. Transition of customer defection rate

The purpose of this study is to clarify the factors specific to customers who repurchase through the analysis of behavior at the first purchase at the EC site.

Using the result of the analysis, we also propose marketing policy for the time of first purchase to encourage repurchase.

2 Data Sets

In this study, we target on the general 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. In this study, we used following data.

- Customer information data (age, sex, registration date, etc.)
- Purchase data (category of purchase items, purchase date, whether purchased item is used item or not, etc.)
- Access history data (login date and time, URL of access page, URL of referrer page, etc.)

  * Period for each data: 8 months from January to August, 2014.

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

The landing route and browsing page name included in the access log data is shown in Tables 2 and 3.

In this study, we analyzed the bellow customer. The reason for this is that we defined the first three months of the data period as the first purchase period and the six months from the first purchase month as the repurchase period.

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.

Table 2. Browsing page name

Browsing page name	Golf beginner page
Golf course reservation page	Golf style page
Golf news page	Golf trip page
Golf lesson page	Golf community page
Golf score management page	Golf event page
Golf movie page	Golf school page

Table 3. Landing route name

Landing route name
Landing from search engine
Landing from mail magazine
Landing from news site
Landing from Facebook
Landing from bookmark
Landing from golf information site
Landing from other EC site
Landing from golf brand site
Landing from Amazon
Landing from Yahoo! shopping
Landing from price comparison site
Landing from internet auction site of Yahoo!
Session disconnection or simultaneously starting a plurality of windows

[Customers]

- Customers who bought for the first time between January and March 2014
  * We exclude the customer who has passed for more than 2 years from registration.

So, the number of analyzed customers was 8,181, of which 3,228 customers repurchased within 6 months.

In this study, the purpose was to predict the presence or absence of repurchase within a certain period from the first purchase. When the objective variable to be predicted is binary, binomial logistic regression models are often used [7].

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 [8].

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

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

As an explanatory variable used in the model construction, we created three variables from membership information data, nine variables from purchasing behavior at first purchase, and 27 Web-browsing behaviors at the first purchase date. Details of the explanatory variables are shown in Tables 4 and 5.

Table 4. Dmographic variables and purchasing behavior variables used in the model construction

Type of variable		Variable name	Data type
Objective variable		Whether customer repurchase within 6 months from first purchase or not	0 or 1
Explanatory variable	Demographic variables	Age	Integer
		Number of days from membership registration to first purchase	Integer
		Whether customer is an mail magazine subscriber or not	Integer
	Purchasing behavior variables	Total amount at first purchase	Integer
		Total number of items purchased at first purchase	Integer
		Whether customer purchased men's wear item at the first purchase or not	0 or 1
		Whether customer purchased lady's wear item at the first purchase or not	0 or 1
		Whether customer purchased golf club item at the first purchase or not	0 or 1
		Whether customer purchased accessory item at the first purchase or not	0 or 1
		Whether customer purchased other item at the first purchase or not	0 or 1
		Whether customer purchased used item at the first purchase or not	0 or 1
		Whether customer purchased sale item at the first purchase or not	0 or 1

* Demographic Variables was created by membership information data
Purchasing Behavior Variables was created by purchase data
Access History Variables was created by Web browsing data.

Although the number of target customers in this research was 8,181, at the time of model construction, we randomly sampled the number of non-repurchased customers by setting the number equal to the number of repurchased customers.

Furthermore, in order to verify the prediction accuracy of the model, we set 70% of the data as training data and 30% as the test data, for each non-repurchased customer

Table 5. Access history variables used in the model construction

Type of variable		Variable name	Data type
Explanatory variable	Access history variables	Browsing frequency of golf course reservation page	Integer
		Browsing frequency of golf news page	Integer
		Browsing frequency of golf lesson page	Integer
		Browsing frequency of management golf score page	Integer
		Browsing frequency of golf movie page	Integer
		Browsing frequency of golf beginner page	Integer
		Browsing frequency of golf style page	Integer
		Browsing frequency of golf trip page	Integer
		Browsing frequency of golf community page	Integer
		Browsing frequency of golf event page	Integer
		Browsing frequency of golf school page	Integer
		Whether landing from search engine or not	0 or 1
		Whether landing from mail magazine or not	0 or 1
		Whether landing from news site or not	0 or 1
		Whether landing from Facebook or not	0 or 1
		Whether landing from bookmark or not	0 or 1
		Whether landing from golf information site or not	0 or 1
		Whether landing from other EC site or not	0 or 1
		Whether landing from golf brand site or not	0 or 1
		Whether landing from Amazon or not	0 or 1
		Whether landing from Yahoo! Shopping or not	0 or 1
		Whether landing from price comparison site or not	0 or 1
		Whether landing from internet auction site of Yahoo!	0 or 1
		Session disconnection or simultaneously starting a plurality of windows	0 or 1
		Average number of page view at first purchase date	Integer
		Average login time of all session at first purchase date	Integer
		Number of login at first purchase date	Integer

and each repurchased customer. As a result, the datasets used in the model construction was split as follows (Table 6).

In addition, in order to grasp the characteristics of repurchased customers more precisely, we constructed repurchase prediction model for each purchase item category

Table 6. Datasets used in the model construction

	Training data	Test data	Total
Non repurchased customers	2260	968	3228
Repurchased customers	2260	968	3228
Total	4520	1936	6456

such as wear item, golf club item and accessory item at first purchase. This is because the behavior at the first purchase is considered different depending on the purchase category. Purchasing behavior variables and number of datasets (training data and test data) used in these model construction are shown in Tables 7 and 8.

Table 7. Purchasing behavior variables used in model construction for each purchase category

Variables	Wear model	Club model	Accessory model
Total purchase amount of each item at first purchase	○	○	○
Total number of items at first purchase	○	○	○
Whether customer purchased wear item at the first purchase or not	×	○	○
Whether customer purchased golf club item at the first purchase or not	○	×	○
Whether customer purchased accessory item at the first purchase or not	○	○	×
Whether customer purchased other item at the first purchase or not	○	○	○
Whether customer purchased used item of each item at the first purchase or not	○	○	○
Whether customer purchased sale item of each item at the first purchase or not	○	○	○

Table 8. Datasets used in repurchase prediction model for each purchase category

	Training data	Test data	Total
Wear item model	1608	690	2298
Club item model	1356	580	1936
Accessory item model	1948	834	2782

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 a following table and we calculated prediction accuracy of the constructed model by using following equations (Table 9).

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

Table 9. Confusion matrix

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

$$\text{ACC} = \frac{TP + TN}{FP + FN + TP + TN}$$

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

$$\text{PRE} = \frac{TP}{TP + FP}$$

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

$$\text{REC} = \frac{TP}{FN + TP}$$

F-measure: harmonic mean of PRE and REC

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

3 Analysis of Repeat Customer

We built a model that predicts repurchase for the entire customer using binomial logistic regression analysis with stepwise selection method. We selected explanatory variables of coefficient of significant probability less than 0.05.

From Table 10, we can see that variables created from Web browsing data are selected much. In addition, the confusion matrix for the test data of this model and the evaluation indicator for confirming the prediction accuracy are shown in Tables 11 and 12.

Subsequently, we built discriminate model focusing only customer who purchased each product category such as wear item, golf club item and accessory item at the time of first purchasing. Table 13 shows the explanatory variables that selected by the model construction for each purchase category.

From Table 13, in all three models, variables of whether landing from bookmark or not, average number of page view at first purchase date and number of login at first purchase date are selected commonly. In addition, the confusion matrix for the test data of these three models and the evaluation indicator for confirming the prediction accuracy are shown in Tables 14, 15 and 16.

Table 10. Estimated value of selected partial regression coefficient

Explanatory variables	Partial regression coefficient
(Intercept)	−0.023
Age	0.082
Number of days from membership registration to first purchase	−0.076
Mail magazine registration	0.087
Total amount at first purchase	0.137
Whether customer purchased used item at the first purchase or not	−0.120
Browsing frequency of golf course reservation page	0.085
Browsing frequency of golf community page	0.089
Whether landing from news site or not	0.060
Whether landing from bookmark or not	0.284
Whether landing from other EC site or not	−0.098
Average number of page view at first purchase date	0.129
Number of login at first purchase date	0.243

Table 11. Confusion matrix of model for entire customer

		Predicted class	
		Positive	Negative
Actual class	Positive	626	342
	Negative	372	596

Table 12. Evaluation indicator of model for entire customer (%)

ACC	PRE	REC	F-measure
63.1	51.2	64.7	57.2

In comparison with accuracy of model for entire customer, it can be seen that there is no difference in accuracy of model between any models (Table 17).

4 Discussions

First, we consider the model predicting repurchase for entire customers. We could see that customers who purchased for the first time immediately after membership registration are leading to repurchase. It is considered important for acquiring repeat customers to promote golf equipment early after membership registration and to shorten the number of days until initial purchase. Moreover, since the partial regression coefficient of purchase of used items is negative, it seems that it is possible to encourage repurchase by recommending new item at the first purchase. Furthermore, since the partial regression coefficients of the e-mail magazine registration, browsing frequency of page other than shopping page and the landing from the news site are

Table 13. Estimated value of selected partial regression coefficient for each purchase category

Explanatory variables	Partial regression coefficient		
	Wear model	Club model	Accessory model
Intercept	−0.010	−0.043	−0.017
Age	0.227	–	–
Number of days from membership registration to first purchase	–	–	−0.098
Total purchase amount of each item at first purchase	–	0.128	–
Total number of items at first purchase	–	0.186	–
Whether customer purchased used item of each item at the first purchase or not	–	−0.256	–
Whether customer purchased sale item of wear item at the first purchase or not	0.132	–	–
Browsing frequency of golf course reservation page	–	–	0.110
Browsing frequency of management golf score page	0.151	–	0.133
Browsing frequency of golf style page	–	–	0.187
Browsing frequency of golf community page	–	–	0.207
Whether landing from bookmark or not	0.255	0.185	0.246
Whether landing from Amazon.com or not	−0.128	–	–
Whether landing from Yahoo! Shopping or not	–	–	−0.164
Average number of page view at first purchase date	0.133	0.128	0.233
Number of login at first purchase date	0.343	0.258	0.262

Table 14. Confusion matrix of model for customers who purchased wear item

		Predicted class	
		Positive	Negative
Actual class	Positive	226	119
	Negative	119	226

Table 15. Confusion matrix of model for customers who purchased golf club item

		Predicted class	
		Positive	Negative
Actual class	Positive	187	103
	Negative	113	187

Table 16. Confusion matrix of model for customers who purchased accessory item

		Predicted class	
		Positive	Negative
Actual class	Positive	226	191
	Negative	131	286

Table 17. Evaluation indicator of model for customers who purchased each category (%)

	Wear item model	Club item model	Accessory item model
ACC	65.5	62.8	61.4
PRE	50.0	51.4	44.1
REC	65.5	64.5	54.2
F-measure	56.7	57.2	48.7

positive, it can be said the customer who is highly interested in golf on a daily basis repurchased. From this, it seems that continuing attraction of customers' interests by periodically distributing e-magazines and news related to golf after membership registration will lead to a reduction of defection rate. Regarding that the estimated value of the partial regression coefficient of the whether landing from other EC site or not is negative and that the partial regression coefficient of whether landing from bookmark is positive, it is inferred that the customer is not using other EC site and uses only this EC site. It seems that these customers already settle in the EC site for purposes other than purchasing.

Second, we consider the model constructed using only customers who purchased wear items. Since partial regression coefficient of whether customer purchase discount items is positive, it is considered effective as measure to encourage repurchase recommending discount items of wear items at the first purchase. Regarding that the partial regression coefficient of the browsing frequency of management golf score page at the first purchase date is positive, it is inferred that customer using the score management function of the EC site during the period until the first purchase or is interested in the score management function repurchased. Considering that the EC site provides score management app, it seems that concentrating on product recommendation in the app will lead to a reduction of defection rate.

Third, we consider the model constructed using only customers who purchased club items. It seems that the customer purchased high price club or didn't purchase used clubs repurchased. In other words, with respect to purchasing of clubs, a reduction of defection rate is expected by recommending without limiting price.

Finally, we consider the model constructed using only customers who purchased accessory items. We can observe that customers who purchased for the first time immediately after membership registration are leading to repurchase. From this result, in the purchasing accessory items whose average price is inexpensive compared to other item categories, it is considered as effective measure for reduction of defection rate that to urge early purchase after membership registration. In addition, since the partial regression coefficient of the browsing frequency of golf course reservation page is positive, the customer purchases inexpensive accessories items on the way to reserve a golf course repurchased. In other words, by recommending expendable items such as golf balls to the customer who is likely to reserve a golf course, customer retention can be expected. Moreover, since customers who browse pages other than shopping page much repurchased, it can be inferred that customers who purchased on impulse when visiting the EC site for purposes other than purchasing repurchased. Therefore, considering the low price of the accessory item, it seems that prompting unplanned purchasing promotes the acquisition of repeat customers.

In addition, since partial regression coefficient of average number of page view at first purchase date is positive in all of the four models constructed, it seems that repurchase is promoted by implementing measures to make customer stay at the EC site as long as possible. Since partial regression coefficient of the number of login on the first purchase date is positive as well, we considered that customers who took a long time to purchase repurchased. From this, it seems that recommendations of similar items promote repurchase.

5 Conclusion

In this study, we extracted the characteristic of customers who repurchase and tried to propose marketing measures. Especially, we built model that predict repurchase within a certain period by binomial logistic regression analysis. As a result of model, we could clarified the characteristics related to repurchase. Moreover, we built models predicting repurchase focusing only customer who purchased each product category such as wear item, golf club item and accessory item at the time of first purchasing. As a result of these model, it was found that characteristic of customers who repurchase are different for each category and we could propose marketing measures to promote to repurchase in detail. However, the prediction accuracy of the constructed model in this research is not sufficient and there is room for improvement. We think that we can build a more precise prediction repurchase model by incorporating variables of behavior before and after the first purchase into the model.

Acknowledgment. We thank Golf Digest Online Inc. for permission to use valuable datasets and Mr. Kazuhiro Fukunaga and Mr. Katsuyuki Mitsuyama for their useful comments.

References

1. DIAMOND. http://diamond.jp/articles/-/36261
2. LTV-Lab. https://wakuten.net/
3. van den Poel, D., Wouter, B.: Predicting online-purchasing behavior. Eur. J. Oper. Res. **166** (2), 557–575 (2005)
4. Blanca, H., Julio, J., Martín, M.J.: Customer behavior in electronic commerce: the moderating effect of e-purchasing experience. J. Bus. Res. **63**(9–10), 964–971 (2010)
5. Yamashita, H., Suzuki, H.: Analysis of purchasing behavior of customers focusing on sale items: logistic regression analysis with consideration of clustering of binary data. J. Oper. Res. Soc. Jpn. **60**(2), 81–88 (2013). (in Japanese)
6. Hamuro, Y., Nakanishi, M., Yamamoto, S.: Analysis of web access log data by the classification by aggregating integrated emerging pattern. J. Oper. Res. Soc. Jpn. **53**(2), 75–84 (2008). (in Japanese)
7. Hisamatsu, T., Togawa, T., Asahi, Y., Namatame, T.: Proposal of finding purchase sign model in an EC site. J. Oper. Res. Soc. Jpn. **58**(2), 94–100 (2013). (in Japanese)
8. Toshiro, T., Kazue, Y., Haruyoshi, T.: Logistic Regression Analysis: Practice of Statistical Analysis Using the SAS (Statistical Analysis System). Asakura Publishing, Shinjuku-ku (1998). (in Japanese)