JSDSS 日本ソーシャルデータサイエンス論文誌
第 4 巻　第 1 号（2020 年 3 月）

原著論文

Proposal of Loyal Customer Discriminant Model Based on RFM Concept —Empirical Analysis using Golf EC Site Data—

Hiroyuki OIKAWA[†1]　Kohei OTAKE[†2]　Takashi NAMATAME[†3]

Abstract: Customer loyalty measurements attract more attention from a viewpoint of customer relationship management; hence, there are some studies that have explored this field. However, none of these studies have considered the potential customer features. In this study, the customer behaviors out of and in the site were considered. This helped determine a wide variety of customer behaviors by adding the behavior data out of the site and has also helped identify a loyal customer more appropriately. We designed a model that can determine if a customer has high loyalty to the electronic commerce (EC) site. To achieve this design, we used the behavioral data out of the EC site in addition to recency, frequency, and monetary (RFM) data from the purchasing data. Furthermore, we adopted a logistic regression from previous studies to accomplish this model. Moreover, we analyzed the purchase history of the loyal customer identified in the discriminant model in order to understand the purchasing tendency. Three customer classifications were used in this study: Customer A, B, and C, and the precision of the RFM and our RFMO models was compared. The RFMO model proved to be more precise in classifying loyal customers and the other customers with high precision compared with the ordinal RFM model about Customer A and Customer B. In addition, it was shown that purchasing tendencies are different between the loyal customers.

Keyword: Marketing, EC Sites, Loyal Customer, Logistic Regression

1. Introduction

Recently, the size of the electronic commerce (EC) market has increased in Japan. According to the Ministry of Economy, Trade and Industry in Japan, the EC market size in Japan has increased by 3 times in eight years from 2010 to 2018 [1]. On the EC site, consumers can purchase conveniently at any time without any temporal or spatial constraints.

Owing to the recent prevalence of big data, the quantity and types of data that can be archived at a company unit have also increased remarkably. As a result, various analyses are conducted using big data on an EC site run by a company. Noteworthy is that customer loyalty measurements attract more attention from a viewpoint of customer relationship management (CRM).

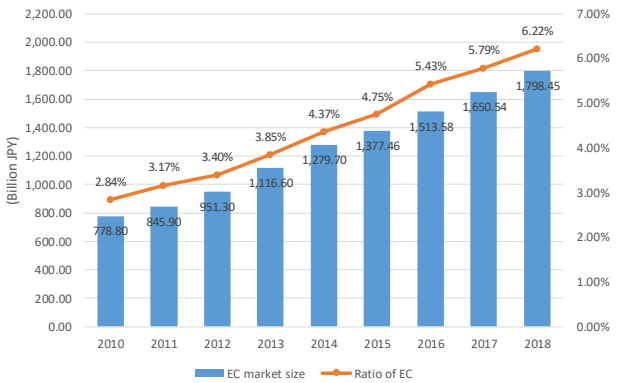

Figure 1 Transition in Japan EC market size

Owing to a reduction in semiconductor prices, companies accumulate a large amount of various data; this is called the Big Data era, which plays an important role in businesses.

The customer loyalty on the EC site is influenced by various factors and is measured using purchasing data (ID-POS data) or browsing data (access log). Above all, characterization by three indices of RFM is commonly used to classify a customer. Companies predominantly use recency (recent purchasing date), frequency (purchasing frequency), and monetary (purchasing amount of money), in short RFM, to measure customer loyalty. However, these indices are based on data obtained from customers at a certain point in time. Therefore, not all individual characteristics of customers are captured [2].

On the other hand, due to the big data era mentioned above, we were able to extract the behavioral data of the customer out of the site (the real world) from new services and social media. For example, on a golf EC site, there are a few new services such as golf play data (game score) and reservation data of the golf course. These data sets show customer's real behavior (behavior outside the EC site), which could not be known in the past. Thus, they can be useful when planning marketing promotions.

This paper comprises 9 sections. In Section 2, we explain the purpose of this study; Section 3 describes previous studies and the positioning of this study; Section 4 presents the data summary used in this study; Section 5 is an analysis of current conditions based on real onsite problems; in Section 6, we explain the

†1 NTT Communications Corp.
†2 Tokai University
†3 Chuo University (Correspondence Author: nama@indsys.chuo-u.ac.jp)

development of the model to use for this analysis; in Section 7, we explain purchasing trends of loyal customers' products; Section 8 presents the discussion of this study; and lastly, in Section 9, we draw our conclusion and mention future applications.

2. Purpose of this study

In this study, we propose a model that can classify a loyal customer with a high loyalty to the EC site. We therefore used behavioral data out of the EC site in addition to RFM from the purchasing data. The behavioral data capture a wide variety of customer behavior hence we can identify a loyal customer more appropriately. To develop the model, we used a logistic regression adopted from previous studies.

In addition, we analyzed the purchase history of the loyal customer identified in the discriminant model. This helps us understand the purchasing tendency of the loyal customer.

Finally, we proposed a marketing tool by using the analysis of the purchasing tendency and the discriminant model. In this study, we examine golf EC sites to identify loyal customers.

3. Previous studies and relevance of this study

Previously, the evaluation index of RFM from the purchasing data was used to measure the customer's loyalty. The index has proved to be effective.

In addition to the RFM index, Mahboubeh et al. [3] used an index to evaluate the customer's loyalty based on the number of items purchased; however, it showed no significance. As observed in the result, the RFM index had a certain interpretability.

On the other hand, Abe [2] reported that the RFM index is widely used to determine excellent customers at a site in CRM. However, these indices are significantly influenced by the time of observation; thus, it is not enough to use only the index of RFM when measuring the customer's loyalty.

As a result, various behavioral data are used in addition to RFM. Doi et al. [4] proposed a model that determines excellent customers in a store using the history of the system that registers the check-in to stores by smartphone applications. The proposed model and the discriminant model are highly accurate. The results show that other types of data in addition to the purchasing data can be acquired.

As Abe [3] points out, the RFM index depends on how it is observed; thus, it is an inadequate tool to classify customers. Instead, we need an index that shows the customer's behavior out of site, uses behavioral data in the site similar to the purchasing data, and accesses logs when measuring the loyalty of the customers. Behavioral data from out of the site accumulate in the enterprise unit in the current big data era. However, there are very few applications wherein these data were utilized.

Recently, data acquisition from various services has become possible thanks to the widespread portal site. In addition to the purchasing data that have been used to measure customer's loyalty in the past, reservation data in a golf course and play score data of EC site members for a course were employed as the behavioral data from out of the site.

4. Data summary

In this study, we focus on one of the largest golf portal sites in Japan. We used the purchasing, reservation, and the play score data of the golf course from A.

4.1 Purchasing data

The purchasing data refer to the data that record the purchasing behavior history in the golf EC site. Information such as the purchase date, product information, price, and the amount is considered.

4.2 Golf course reservation data

The reservation data refer to the data that record the reservation history of the golf course. Information such as the location for the golf course, reservation date, and the play date is recorded.

4.3 Golf score data

The golf score data refer to the data used to record the score at the time of the play and the score after every 1 round of each user. The players registered the score themselves when they played.

Only the data that satisfy the following conditions and are common to a subscriber ID will be analyzed. The periods considered are as follows.

- Purchase Data
  - Period: September 1, 2013 to August 31, 2015
- Golf course reservation data
  - Period: September 1, 2001 to August 31, 2015
- Golf score data
  - Period: September 1, 1970 to August 31, 2015

5. Analysis of the current condition

We first analyzed the loyal customers in Company A after acquiring data following our discussion with the marketing manager. Company A runs the operation of a golf portal site. The following points were noted from our discussion:

Table 1 Data summary.

Purchase Data (No. of transaction)	2,179,348
Member (No. of ID)	49,068
Golf score data (No. of transaction)	17,340,121
Golf course reservation data (No. of transaction)	9,308,890

1. The repeaters had previously contributed immensely to the rate of sales but have decreased in recent years.
2. Other companies have intensified the competition of attracting customers.
3. Specific customers are afforded custom services (incentives).
4. It is difficult to detect loyal customers.

Points 1 and 2 above show how increasingly difficult it is to attract loyal customers in Company A. Regarding point 3, periodic magazines were mailed; nonetheless, they proved ineffective. With regard to point 4, although Company A has acquired the data pertaining to customer attributes, it has not realized any concrete marketing campaign. Generally, gathering new customers requires more advertising costs than retaining loyal customers. Thus, the marketing manager of Company A said that the specification of a loyal customer is very important for them to save advertisement expenses and make profits.
After the discussion, we realized that a customer might become loyal in two ways:

I. Through years of continuous purchases
II. After a sudden change in purchasing behavior

We define customer [I] as a “continuous loyal customer” and customer [II] as a “hidden loyal customer.” We have designed a model to identify each loyal customer ([I] and [II]) to understand their purchasing tendency.

6. Proposed model

Figure 2 below presents the outline of the analysis of this study. Moreover, it describes the specific creation summary of the model.

6.1 Statistical model

To control the fluctuations in purchasing behavior due to seasonality, the target data for the analysis were separated into two periods, as shown in Table 2.

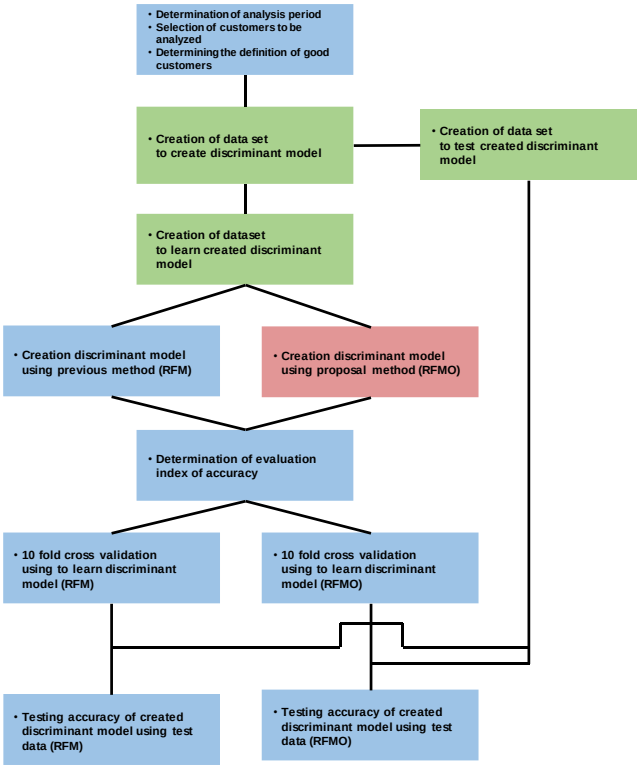

Figure 2 Flowchart of analysis

Table 2 Analysis period

Period No.	Analysis period
Period 1	September 1, 2013 ~ August 31, 2014
Period 2	September 1, 2014 ~ August 31, 2015

First, the number of purchase categories and the accumulated number of purchases were used as the linear measurement for customer loyalty. The kind of clubs (driver), clothes (shirt, pants, and skirt), a ball, and a glove fall in purchase categories. The accumulated number of purchases indicates the total number of purchases in 1 year. Customers were segmented in each period following the conditions outlined in Table 3.

Table 3 Ranking conditions of customer

Rank	Condition
upper	5 or more number of purchases and 5 or more in the number of product categories
middle	Except for upper and lower
lower	2or less number of purchases and 2 or less in the number of product categories

In this study, we analyzed customers who switch like A, B, and C, as shown in Figure 3.

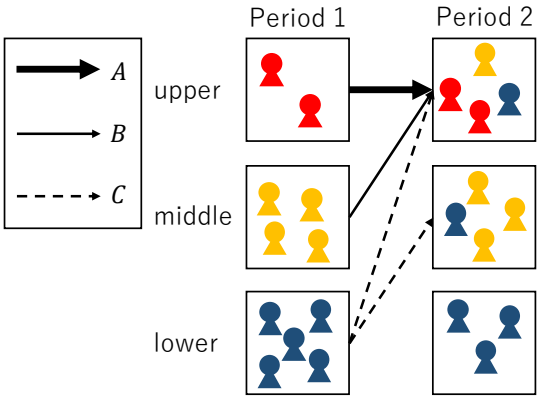

Figure 3 State of the customer segments switch

We regarded customers who switch like A as continuous loyal customers, and B and C as hidden loyal customers.

A: *Continuous loyal customer*
The customer who purchases as an upper customer together in sites during both periods. Thereafter, we call a Continuous loyal customer as Customer A.
B: *Hidden loyal customer-1*
The customer who was a middle customer in the first period but became an upper customer in the second period. Hereafter, we call Hidden loyal customer-1 as Customer B.
C: *Hidden loyal customer-2*
The customer who was a lower customer in the first period, but became an upper customer or a middle customer in the second period. Hereafter, we call Hidden loyal customer-2 as Customer C.

Logistic regression was widely and effectively used when creating a discriminant model for customers in previous studies [5, 6, 7, 8]. Therefore, in this study, we propose a logistic regression model to distinguish between a continuous loyal customer and a hidden loyal customer. The continuous loyal customer probability p_i is expressed by the following equation (1):

$$p_i = \frac{\exp\{b_0 + \sum_j b_j x_{ij}\}}{1 + \exp\{b_0 + \sum_j b_j x_{ij}\}} \quad (1)$$

where x_{ij} refers to the factors that affect loyalties and β_j refers to parameters for each explanatory variable (β_0 is intercept).

Next, we will report on the explanatory variables used in identification. We created two kinds of models: one is a model using each explanatory variable of RFM and the other is a model that includes an index of O, which shows customers outside the site behavior. Here, the O index refers to a general term for the behavior variables of each customer; it was made using some reservation data and play score data. In particular, the play score data contain detailed information on the golfer's play per round. Specifically, there are data on the hole score, which indicate the number of occurrences of birdies, pars, bogies, and triple bogies, and data on the play content such as the number of OBs and the number of fairways. We created variables to use in the O index; these variables are listed in Table 4.

Table 4 Explanatory Variables for index O

Type of data	Explanatory Variable
Purchase data	RFM score
Golf score data	No. of Played
	Statistical values of total score
	Statistical values of hole score ratio (e.g. birdie, par, bogey)
	Statistical values of play content ratio (e.g. OB, fairway, bunker)
Golf course reservation data	No. of reserved
	No. of types which have been reserved
	No. of cancel after reserved

Here, we calculated the average, maximum, minimum, and median values for the three statistical values of the "total score," "hole score ratio", and "play content ratio" using the play history for each user. These statistical variables are treated as different explanatory variables in this study.

From the variables shown in the table, we choose explanatory variables, which are to be used in the model as the index of O. We have assigned loyal customers as the objective variable and have conducted a logistic regression using all the explanatory variables. This was achieved by choosing the combination of explanatory variables with the lowest Akaike's Information Criterion (AIC). Table 5 lists the explanatory variables of the index of O that were applied to each model.

6.2 Model evaluation

In this section, we conduct a comparative verification using the model created to identify all three kinds of loyal customers. RFM and RFMO models are used as explanatory variables.

Table 5 Explanatory Variables

Customer-A	Customer-B	Customer-C
total-score-Median	total-pat-Max	total-score-Min
rate-birdie-Max	rate-par-Mean	total-pat-Mean
rate-par-Min	rate-double-bogey-Mean	total-pat-Min
rate-bogey-Mean	rate-double-bogey-Min	total-pat-Median
rate-bogey-Max	rate-double-bogey-Max	rate-par-Median
rate-bogey-Median	rate-triple-bogey-Min	rate-double-bogey-Max
rate-double-bogey-Mean	rate-fairway-Mean	rate-triple-bogey-Median
rate-triple-bogey-Min	rate-fairway-Max	rate-penalty-Max
rate-fairway-Min	rate-penalty-Min	rate-bunker-Mean
rate-ob-Mean	rate-penalty-Max	rate-bunker-Min
rate-ob-Min	rate-bunker-Min	rate-sandsave-Min
rate-ob-Median	rate-bunker-Median	rate-par-on-Max
rate-penalty-Mean	rate-bogey-on-Max	rate-bogey-on-Max
rate-penalty-Median	reservation-count	reservation-count
rate-sandsave-Min		
rate-bogey-on-Mean		
rate-bogey-on-Max		
rate-bogey-on-Median		
reservation-count		
number-of-types-of-golf		

6.2.1 Evaluation method

The proposed model is evaluated in this section. In addition, we compare the precision, evaluated by 10-fold cross validation using period 1 to assess the validity of the analysis for each model. The evaluation results are shown using the confusion matrix in Table 6. In addition, we used accuracy, precision, recall, and F-value as indicators to evaluate precision. These indicators are shown in equations (2) to (5).

Table 6 Confusion matrix of the classification results

		Actual Result	
		True	False
Prediction Result	Positive	TP	FP
	Negative	FN	TN

Accuracy: The ratio of the predicted real result in the whole

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

Precision: The ratio of the predicted true results in the whole

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall: The ratio of predicted true result in the whole

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

F-value: Harmonic mean of Precision and Recall

$$\text{F} - \text{Value} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

Moreover, we inspected these models using period 2. In addition, we adopted the model with the highest F-value from the result of the 10-fold cross validation for the inspection. We show the mean of all the evaluation indexes in Table 7 as the precision value of each model.

The numbers for the RFMO model are larger than in the RFM model, particularly in Accuracy and Precision out of the 4 evaluation indexes. However, the numbers for the RFM model became larger than in the RFMO model for the Recall and F-value. We distinguished each loyal customer by the respective models and conducted a comparative inspection using the inspection data. The obtained result is shown in Table 8 to Table 10.

Table 7 Accuracy results of 10-fold cross validation

Model		Accuracy	Precision	Recall	F-value
Customer-A	RFM model	0.654	0.594	0.920	0.736
	RFMO model	0.861	0.841	0.900	0.879
Customer-B	RFM model	0.553	0.530	0.951	0.680
	RFMO model	0.590	0.589	0.659	0.608
Customer-C	RFM model	0.590	0.668	0.359	0.486
	RFMO model	0.697	0.683	0.737	0.709

Table 8 Comparison results of the four values of the discriminant model in Customer-A

Model	Accuracy	Precision	Recall	F-value
RFM model	0.450	0.235	0.974	0.378
RFMO model	0.838	0.517	0.901	0.657

Table 9 Comparison results of the four values of the discriminant model in Customer-B

Model	Accuracy	Precision	Recall	F-value
RFM model	0.283	0.077	0.996	0.142
RFMO model	0.612	0.127	0.938	0.224

Table 10 Comparison results of the four values of the discriminant model in Customer-C

Model	Accuracy	Precision	Recall	F-value
RFM model	0.674	0.069	0.095	0.080
RFMO model	0.470	0.026	0.071	0.038

From the results obtained, we performed the comparison, which is the result of the obtained confusion matrix at the RFM and RFMO models for each loyal customer.

Customer A: As Table 8 shows, the Accuracy of the RFMO model is higher than that of the RFM model. The value of Recall is similar for both models, and the precision of the RFMO model is higher than that of the RFM model. Therefore, the F-value of the RFMO model became high.

Customer B: As Table 9 shows, the Accuracy of the RFMO model is higher than that of the RFM model. The value of Recall is similar for both models, and the precision of the RFMO model is higher than that of the RFM model. Therefore, the F-value of the RFMO model became high.

Customer C: As Table 10 shows, the Accuracy of the RFM model is higher than that of the RFMO model. The Precision of the RFM model is higher than that of the RFMO model. Similarly, the Recall value of the RFM model is higher than that of the RFMO model. Therefore, the F-value of the RFM model became higher than that of the RFMO model. However, the F-value of the two models is relatively low because both the Precision and Recall values were also low.

From the comparative inspection discussed above, the precisions of classification of the RFMO model were higher than in the RFM model for Customer A and Customer B. On the other hand, the F-value was lower in Customer C. Therefore, we find out that both models did not grasp Customer C. As a result, we checked the data set that was divided into loyal customers and other customers. Figure 4 shows the boxplot of explanatory variables for two data sets.

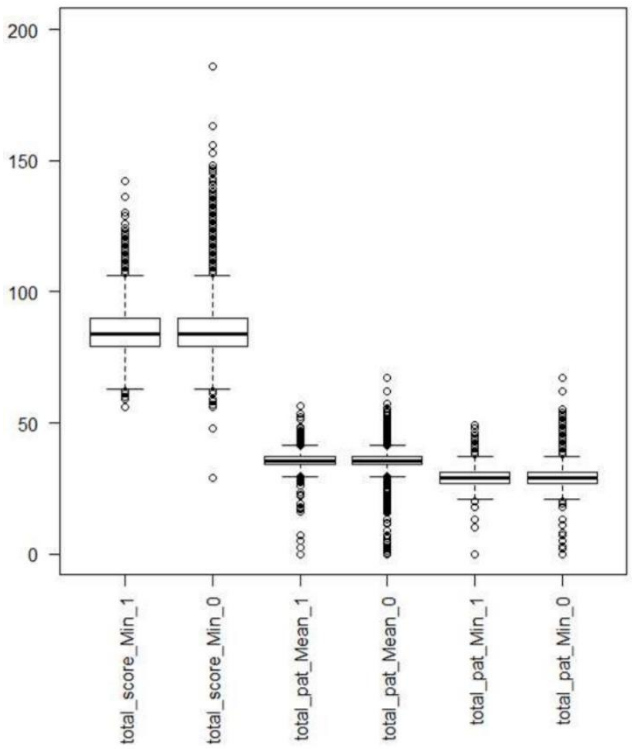

Figure 4 Boxplot of each loyalty variable in Customer-C of teacher data

As shown in Figure 4, the data set in each explanatory variable yielded no change. Therefore, regarding Customer C, the variable used for this model is not that significant to differentiate between it.

7. Item trend of loyal customer

In this study, we know the purchase tendency of each loyal customer besides that of Customer C that has a small F-value. The tendency of the total purchase price for every category in each period of Customer A classified with the RFM and RFMO model is indicated using Figure 5.

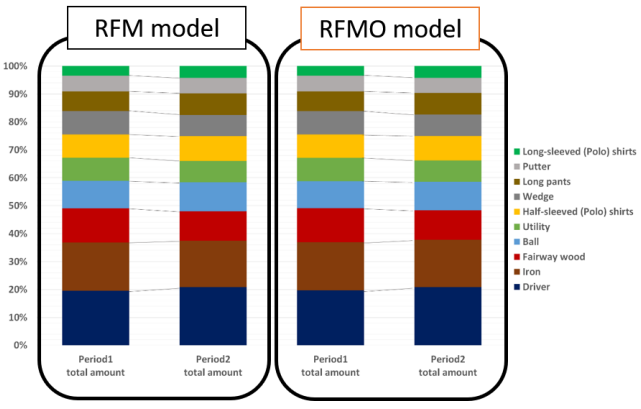

Figure 5 The purchase total amount ratio of Customer-A

As shown in Figure 5, the total purchase price of each category for the RFM model and RFMO model has a similar component ratio. For the above reasons, the loyal customers identified by each model find out that they could be customers with a similar purchase tendency. The percentage occupied by the dominating 5 categories in the total purchasing total amount of the top 10 items is approximately 60%, and the purchase tendency of the goods does not change in each period. In addition, club kinds are a dominant category. Therefore, this model shows that the identified Customer A is the customer who purchases high price range goods.

Figure 6 summarizes the tendency of the accumulated purchase price for every category in each period of Customer B.

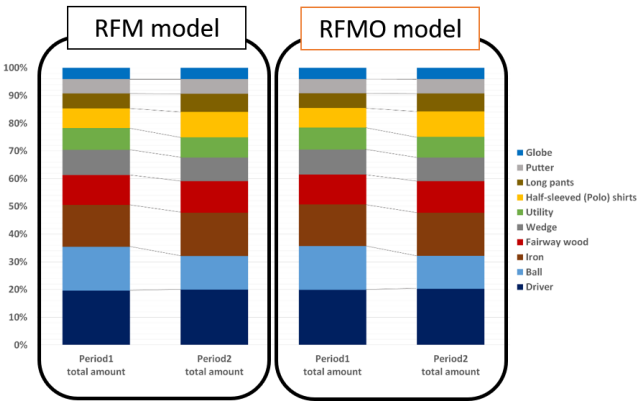

Figure 6 The purchase total amount ratio of Customer-B

Figure 6 compares the RFM and RFMO models; the purchase tendency calls period 2 from period 1, and there is no difference. For this reason, each model identified the same Customer-B. On the other hand, when comparing the purchase tendency of the goods in each period by the respective models, the names of the 5 dominating categories do not change. However, the ratio of the accumulating purchase price of the ball decreases and the percentage that the 5 dominating categories occupy decreases. As a result, Customer B could be shifting purchases to other consumables such as a ball.

8. Discussion

As seen in the result, the RFMO model can classify a loyal customer and the other customer with a higher precision than the RFM model regarding Customer A and Customer B. We assumed that the loyalty variable could be useful when discriminating a customer with high loyalty.

The knowledge imparted by this study can be applied in business. We used novice data techniques never used before. Owing to new indices, we were able to discriminate excellent customers. Customers who do not use the EC site in the portal site can also be classified as excellent customers in the future.

However, no model could specify Customer C in our analysis. We thought that it is not possible to make a model without customers who play golf have high level of proficiency. Golf course reservation and play score data were used when making the index that shows customer loyalty on the site. The customer segments from these data were extracted have a high level of proficiency with golf.

Moreover, the amount of the purchasing data and play score data for Customer C in this study was insufficient, but the data were discriminated properly. Therefore, we thought that the precision of the model improves using data, which much acquire access log data and the data about social login.

When we grasped each loyal customer and looked for the purchase tendency of the goods, Customer A and Customer B showed different purchase tendencies.

From our comparison results, the total purchase amount of the ball in Period 2 of Customer B decreased from Period 1. The purchase tendency in the above categories changed from balls to other categories such as clubs and clothes.

We can deduce that Customer B focuses on playing golf or is interested in other categories. In addition, we assumed that Customer A is strongly committed to a club because it was previously devoted to golf as part of its lifestyle. We also assumed that not all loyal customers have the same loyalty to golf; hence, we developed a marketing policy on a small scale for each customer.

Therefore, when the customers who are in the upper segment of Customer B in Period 2, measures which treat the club as to acquire customers is effective. The following measures to attract customers were considered:

- Discount for a club
- Increasing point giving rate

The customer will therefore react to a marketing policy of a club and purchase goods. As described above, in order to increase a customer's loyalty to a store, a strong promotion and marketing policy are needed.

9. Conclusion and future applications

In this study, we designed a model that can classify and discover a customer with high loyalty by using customer behavioral data and purchase data. It was possible to use customer's loyalty variable for a customer with accumulated golf course reservation data and playing score data and the utility was also admitted. By

discriminating a loyal customer and analyzing the purchase data, the purchase tendency of a customer with high loyalty to an EC site can be known. Furthermore, it was possible to discern the purchase tendency for goods disregarded by loyal customers.

The data set used for a customer of a lower segment in period 1 was small and so could not be used as a loyalty variable. Therefore, a good model for application could not be extended to each model.

Recently, the portal site dealing with golf items focused on customer development using various contents such as lessons, news, and event information. In addition, it was possible to record the customer's reading history in a portal site.

Given the literature established in this study, a combination of data obtained through different contents and before purchasing is needed for effective analysis, in addition to the data used in this study. Another model should also be proposed for other commodities. This model will be more effective in analyzing the data obtained from the customers outside behavior.

Acknowledgement

We thank a golf portal site company for permission to use valuable datasets and for useful comments. This work was supported by JSPS KAKENHI Grant Number 19K01945 and 17K13809.

References

[1] Examination of Ministry of Economy, Trade and Industry, *Improvement of Information Infrastructure and Improvement of Service Infrastructure in Japan's Economic Society*, (Market Research on E-commerce) (2016). (in Japanese)

[2] Abe M.: “RFM Measures and Customer Lifetime Value: Investigating the Behavioral Relationship in a Non-Contractual Setting using a Hierarchical Bayes Model,” *Journal of the Japan Statistical Society*, Vol. 41, pp. 52-54, (2011), (in Japanese)

[3] Manboubeh, K., Kiyana, Z., Sarah, A. and Somayeh, A.: “Estimating Customer Lifetime Value Based on RFM Analysis of Customer Purchase Behavior: Case study,” *Procedia Computer Science*, Vol. 3, pp. 57-63 (2010).

[4] Doi, C., Katagiri, M., Ishii A., Araki, T., Inamura, H. and Ohta, K.: “Estimating Value of Customer Through Store Check-in Histories and its Application for Visitor Promotion,” *Journal of Multimedia, Distributed, Cooperative and Mobile Symposium 2016*, pp.735-741 (2016). (in Japanese)

[5] Suzuki, H., Mizuno, M., Sumita, U. and Saji, A.: “Characteristic Evaluation and the Financial Effect of the Excellent Customer Discrimination Technique for CRM,” *Department of Social Systems and Management Discussion Paper Series, University of Tsukuba*, No. 1123 (2005) (in Japanese)

[6] Nakahara, T., Uno, T. and Yada, K.: “Extracting Promising Sequential Patterns from RFID Data Using the LCM Sequence,” *14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, *Part III*, pp. 244-253 (2010).

[7] Hara, A., Takano, M. Shtykh, R. and Kawabata, T.: “Extracting Segments of Users with Conversion Probabilities for Internet Advertising,” *Proceedings of Annual Conference of Artificial Intelligence Society,* Vol. 29*,* pp.1-4 (2015).

[8] Rosales, R., Chang, H. and Manavoglu, E.: “Post-click Conversion Modeling and Analysis for Non-Guaranteed Delivery Display Advertising,” *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp.293-302 (2012).