International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:14, No:8, 2020

Competitors' Influence Analysis of a Retailer by Using Customer Value and Huff's Gravity Model

Yepeng Cheng, Yasuhiko Morimoto

***Abstract*—**Customer relationship analysis is vital for retail stores, especially for supermarkets. The point of sale (POS) systems make it possible to record the daily purchasing behaviors of customers as an identification point of sale (ID-POS) database, which can be used to analyze customer behaviors of a supermarket. The customer value is an indicator based on ID-POS database for detecting the customer loyalty of a store. In general, there are many supermarkets in a city, and other nearby competitor supermarkets significantly affect the customer value of customers of a supermarket. However, it is impossible to get detailed ID-POS databases of competitor supermarkets. This study firstly focused on the customer value and distance between a customer's home and supermarkets in a city, and then constructed the models based on logistic regression analysis to analyze correlations between distance and purchasing behaviors only from a POS database of a supermarket chain. During the modeling process, there are three primary problems existed, including the incomparable problem of customer values, the multicollinearity problem among customer value and distance data, and the number of valid partial regression coefficients. The improved customer value, Huff's gravity model, and inverse attractiveness frequency are considered to solve these problems. This paper presents three types of models based on these three methods for loyal customer classification and competitors' influence analysis. In numerical experiments, all types of models are useful for loyal customer classification. The type of model, including all three methods, is the most superior one for evaluating the influence of the other nearby supermarkets on customers' purchasing of a supermarket chain from the viewpoint of valid partial regression coefficients and accuracy.

***Keywords*—**Customer value, Huff's Gravity Model, POS, retailer.

I. Introduction

IN today's supermarket business, the ID-POS database enables supermarkets to analyze customer behavior and adopt more targeted and personalized marketing strategies such as customer relationship management (CRM) [1], to improve the competitiveness of supermarkets. The ID-POS database digitally records customer ID, customer information, sales records, etc. Therefore, customer behavior is measurable by counting their daily shopping records as customer values. Generally speaking, customer value analysis, which is also known as recency, frequency and monetary (RFM) analysis analysis [1]-[3], mainly depends on three parametric indicators, customer shopping recency, frequency, and monetary. They can reflect the customer loyalty of a store. The models consist of RFM indicators with other statistical parameters that are trainable by clustering analysis [4] and other machine learning methods to investigate the customer shopping preference.

Tanaka et al. [5], proposed a model, including RFM indicators with the proportion of purchased products of each customer in a supermarket chain. They define the loyal customer by Decyl analysis [5], and then use logistic regression analysis [6]-[9] to find loyal customers and detect the loyal customers' preferences for each product simultaneously. Logistic regression analysis is widely used in parametric impact analysis. The coefficients of logistic regression mathematically considered as the parameters in the Odds ratio [10]. The Odds ratio can reflect the influence of variable parameters on a particular parameter. As a result, they built a loyal customer analysis model with high classification accuracy. There is a lot of the other customer's information in the ID-POS database, such as the customer's address. Therefore, Tanaka's method is also useful to detect different aspects of the customer's behavior. For example, the distance between a customer's home and all supermarkets in a city is computable. The influence of nearby competitors is discoverable by analyzing the relationship between distance and the customer's shopping amount of the target supermarket. The customers who live close to competitors are more likely to be influenced by them, resulting in decreased shopping amounts in the target supermarket. However, logistic regression cannot train the raw distance data without preprocessing directly since the multicollinearity problem [11] may occur between the distance data and RFM indicators. Therefore, it is essential to find a method that can transform the distance data into probability similar to Tanaka's work. The main contributions of this paper include as follows:

- This paper improves the original RFM values by RFM scores [4], [12] to build loyal customer analysis models effectively.
- This paper proposes a method based on Huff's gravity model [13] to solve the multicollinearity problem in logistic regression analysis.
- This paper presents another method named inverse attractiveness frequency to increase the accuracy of loyal customer classification and retailer competition analysis diversity.

The remainder of the paper is structured as follows: Section II analyzes the related work. The structures of the loyal customer analysis models are shown in Section III. Section IV explains analytical methods. Section V evaluates the three types of models by logistic regression analysis. Section VI concludes this paper.

Yepeng Cheng and Yasuhiko Morimoto are with the Department of Information Engineering, Graduate School of Engineering, Hiroshima University, Higashi-Hiroshima, Japan (e-mail: d185088@hiroshima-u.ac.jp, morimo@hiroshima-u.ac.jp).

II. Related Work

A. Retailer Competition Analysis

In economics, [14] applied the law of gravity in physics to analyze the retail industry, which indicated that consumers are willing to drive a further distance to larger retail stores for shopping. However, this law only considers the macro aspect and lacks the investigation of the micro aspect of consumer decision making in actual shopping activities. Reilly's law assumes that consumers will choose a fixed retail store for shopping. In fact, consumers expect to go shopping at two retail stores in close geographical locations simultaneously. Huff's gravity model [13] later reformed Reilly's law.

Huff's gravity model uses probability to describe the spatial relations between stores and consumers in a district. The attraction of a store to a given consumer is related to its size and geographical distance between them. The proportion of its attraction to all stores' total attractions in a region is the probability that a given consumer will purchase at this store. Retailers use this theory extensively for new site selection. However, the accuracy of the shopping store's preference for consumers is not precise enough. Although [15] considered other factors, except for retail store area and distance factors, to improve Huff's model, which is called the multiplicative competitive interaction (MCI) model, the accuracy improvement is still facing a bottleneck. Fig. 1 [16] shows the theoretical store trade area and customer location. The blue, green, yellow and red progression represents zones of increasing patronage probability. Different circles denote the different circular trade area of the retailer store.

B. Customer Analysis

In customer analysis, RFM analysis [1]–[3] aims to build a model that differentiates important customers from large transaction data. Chen et al. [17] proposed an extended model of RFM analysis for the challenge prediction problem of customers in the logistics industry. Later, the research that combines machine learning methods has also been reported. Tanaka et al. [5] considered the RFM and logistic regression analysis to detect loyal customers' preferences for various supermarket products. They set the month elapsed from a customer's last shopping record to the data statistic day, the frequency of a customer comes to store, and the purchase amount of a customer has spent in a time interval for R, F and M values, respectively. Although in Tanaka's work [5], they obtained relatively precise results, the partial regression coefficients of RFM indicators are incommensurable since they have different magnitude. The principal component analysis (PCA) [18] and normalization analysis [19] can also play a role in data magnitude reduction. However, they have the defects of poor interpretability. References [4] and [12] proposed a method that can uniform the magnitude of RFM values where they group the RFM values and give a score of each group, respectively.

Decyl analysis [5] is another analysis method that calculates the purchase ratio and the sales composition ratio of each rank by dividing the consumption of all customers into 10 equal parts based on purchase history data. By purchase ratio and composition ratio, it is possible to know a loyal customer group with a high contribution to sales. The purpose of Decyl analysis is to grasp a loyal customer group and concentrate on it to implement efficient marketing.

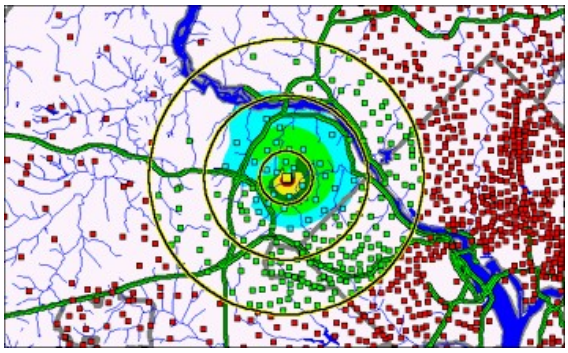

Fig. 1 The gravity based patronage probability model

C. Regression Analysis

Regression analysis is a statistical technique for estimating the relationship between dependent and a set of independent explanatory variables. Polynomial regression [20] is commonly used to analyze the curvilinear data when the power of an independent variable is more than one. It plays a crucial role in regression analysis because any function can be approximated piecewise by a polynomial. Zenker et al. [21] proposed a method including polynomial regression and response surface methodology for place marketing.

Logistic regression [6]-[9] is also a regression analysis method for the dichotomous problem. It quantifies the correlation between a series of independent variables and a dependent variable as a logit odds ratio. The logit odds ratio is the natural logarithm of an odds ratio that represents the influence of the fluctuation of a given variable on the dependent variable. Yeung and Yee [22] predicted consumer purchase propensity by logistic regression analysis. They demonstrate how logistic regression can be used to predict consumer behavior where the explanatory variables are dichotomous and interact with each other. Constantin [23] used a logistic regression model in supporting decisions of establishing marketing strategies for accommodation analysis. Tanaka et al. [5] built a loyal customer analysis model consist of original RFM values and the proportion of item purchasing of a customer. They define the loyal customer of a supermarket chain by Decyl analysis and tag them as target variables. After that, they use logistic regression analysis to find loyal customers and detect the item preference of them effectively.

III. Loyal Customer Analysis Model

A. RFM Analysis

RFM analysis [1]-[3] contains three indicators, how recently a customer has purchased (Recency), how often they purchase (Frequency), and how much they spend (Monetary). To solve the problem that different magnitude RFM indicators are incomparable in [5] and make the analysis results more

accurate, this study considers converting the original RFM values into the form of customer RFM scores based on [4], [12]. The analytic models set R value as days elapsed from last sales record to data statistics day, F value as the frequency of customer come to store and M value as the average of one time purchase amount in a time interval from first shopping day to data statistics day. R value is ordered by ascending and F, M value is ordered by descending. Each of them is divided into five groups according to top rank 20%, 20% to 40%, 40% to 60%, 60% to 80%, 80% to 100%, respectively. Each group of R, F and M value is scored from level 1 to 5 based on their group rank. If a customer owns three high RFM scores such as (5,5,5) or (5,4,4), this customer has a high loyalty in a store. Suppose a store has 100 customers, each with a different RFM value. The examples of the R, F and M score are shown in Tables I-III. The units of RFM values are days, times and amounts of money, respectively. Table IV shows the concrete structure of experimental data for the first proposal of this paper, the RFM type model.

TABLE I
THE EXAMPLE OF R SCORE

Customer	R	R rank	Percentage	R group	R score
1	65	48	48%	40%-60%	3
2	354	87	87%	80%-100%	1
3	30	28	28%	20%-40%	4

TABLE II
THE EXAMPLE OF F SCORE

Customer	F	F rank	Percentage	F group	F score
1	28	80	80%	60%-80%	2
2	23	82	82%	80%-100%	1
3	96	20	20%	0-20%	5

TABLE III
THE EXAMPLE OF M SCORE

Customer	M	M rank	Percentage	M group	M score
1	1926	49	49%	40%-60%	3
2	150	95	95%	80%-100%	1
3	2111	44	44%	40%-60%	3

B. Huff's Gravity Model

The ID-POS database contains the customer information and two year customer shopping records of target supermarket chain A including supermarkets A1 and A2 in the experimental city. This paper converts the customer address to longitude and latitude and uses Euclidean distance [24] to compute the distance to each supermarket in the experimental city. There are seven supermarket chains in that city, including one target supermarket chain A and six competitive supermarket chains B, C, F, H, J and N. The chain name will combine with a number to denote each individual supermarket of supermarket chains. The horizontal axis of Figs. 7 and 8 shows all supermarkets in the experimental city. A method based on Huff's gravity model [13] is considered to convert the distance factors into uniform attractiveness probability.

$$hf_{ij} = \frac{\frac{s_j}{d_{ij}^{\alpha}}}{\sum_{j=1}^{n}\frac{s_j}{d_{ij}^{\alpha}}}, \sum_{j=1}^{n} hf_{ij} = 1 \quad (1)$$

In a district, the attraction of a retail store to a given customer is the ratio of its size denoted as s to distance between them denoted as d. Therefore, hf_{ij} indicates the attractiveness probability of customer i will go shopping at store j and α denotes the distance decline coefficient. If the size of every store is fixed, distance can convert into a uniform attractiveness probability. The attractiveness probability is a negative correlation to the distance. The customer obtains a significant influence when he lives close to a supermarket, and vice versa. The multicollinearity problem [2] will be avoided since the sum of the attractiveness probability of a customer is one, irrespective of the high or low of the RFM score. Table V shows the structure of experimental data for the second proposal, the RFM+ type model including RFM score and hf score.

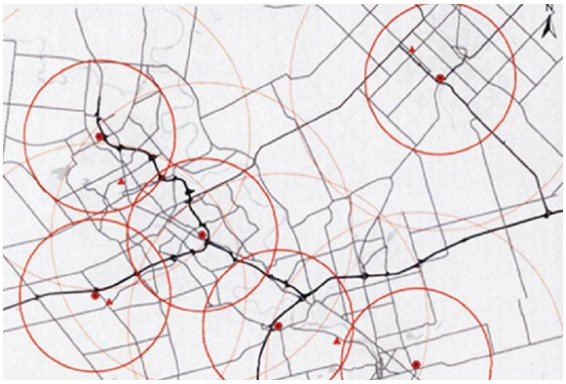

Fig. 2 The interaction between circular trade area of stores in a district

C. Inverse Attractiveness Frequency

The third proposal is based on [5]. They proposed inverse shop frequency to reduce the customer's item preference between different individual supermarkets of a supermarket chain and obtained precise results for loyal customer classification. This paper focuses on the influence of the different trade area radius of competitors on the customers of target supermarkets. Fig. 2 [25] shows an example of the interaction between the trade areas of stores. It reflects the competition between the stores in a district. For different stores, the customer's shopping preference and shopping frequency are affected by distance, store area, etc. Therefore, how to mathematically measure the influence of competitors is an important issue. The polynomial regression analysis [20] is used to detect the impact of competitors firstly. This paper conducts two experiments of polynomial regression analysis on the customer shopping data and distance data of A1 and A2. Fig. 3 shows the tendency of distance and each customer monthly purchased item number of A1. The closer the distance to A1, the higher the shopping quota is. Fig. 4 shows the comparison of A1 with two competitors. This paper chose the customers of A1 who live in a 3 km trade area radius of the

competitors, and the distance to the competitors is closer than to A1. The customers of A1 close to the competitors are affected since B and C are lower than A1. So do the tendencies in Figs. 5 and 6 for A2.

According to Figs. 3 and 5, this paper defines the customers that purchased item number per month is higher than the polynomial regression curve as the dominated customers by target supermarkets. The more significant the proportion of these customers surrounding the competitive supermarkets, the greater the influence the target supermarkets have in this area. Fig. 7 shows the proportion of dominated customers of A1 and A2 surrounding a 3 km radius of the trade area of all supermarkets in the experimental city. The difference between A1 and A2 is from 10% to 20%. The inverse attractiveness score of each supermarket is formulated to reduce this difference for feature quantity expression, as:

$$iaf_{i,s} = \log \frac{c_i}{d_{i,s}} \tag{2}$$

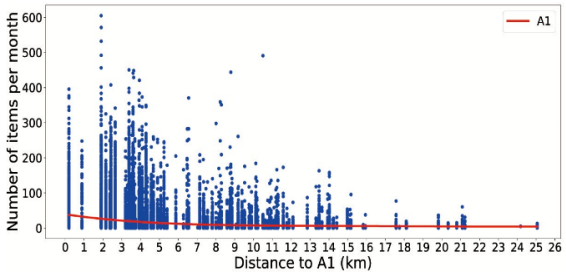

Fig. 3 The correlation between distance and sales for A1

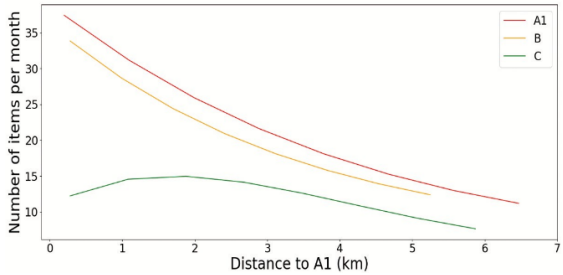

Fig. 4 The comparison of A1 with 2 competitors

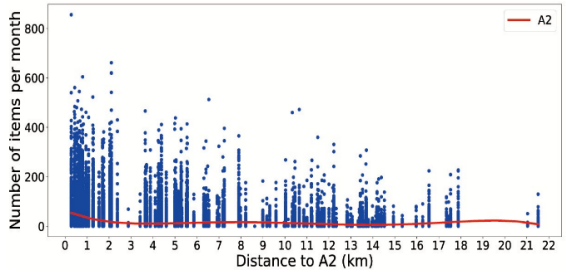

Fig. 5 The correlation between distance and sales for A2

where, c_i denotes the total number of customers in the specific radius of the trade area of the ith competitive supermarket. $d_{i,s}$ indicates the number of dominated customers by target supermarket s surrounding the specific radius of the trade area of the ith competitive supermarket. iaf vector is defined by each customer of the target supermarket chain A. The high proportion in Fig. 7 will have a low iaf value, which means that the target supermarket has a powerful impact on customers surrounding competitive supermarkets. Oppositely, a weak impact has a high value. In this paper, the customers purchased at affiliated target supermarkets may be far from some competitors since this study detects almost all supermarkets in a city. The value of c_i or $d_{i,s}$ perhaps zero. The Laplacian smoothing is considered to avoid this situation.

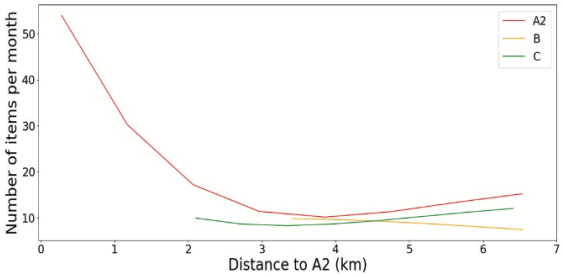

Fig. 6 The comparison of A2 with 2 competitors

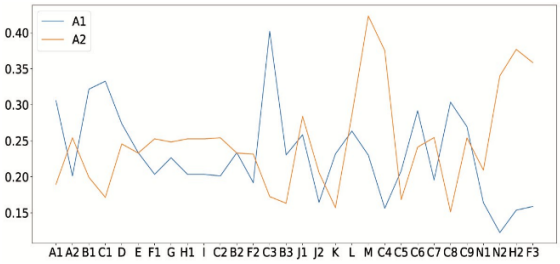

Fig. 7 The proportion of dominated customers of A1 and A2

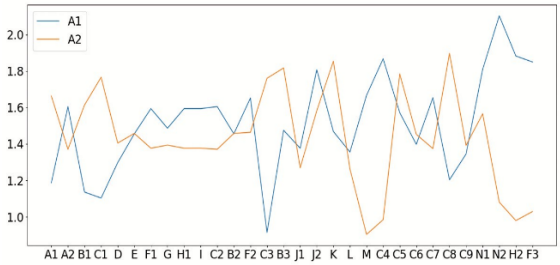

Fig. 8 The iaf score of A1 and A2

$$iaf_{i,s} = \log \frac{c_i + n\lambda}{d_{i,s} + \lambda} \tag{3}$$

where, denotes the total number of supermarkets in a city. The λis a smoothing coefficient. The hf score matrix of A1 and A2 is computed by (1). The hf-iaf score matrix of A1 and A2 is generated by the corresponding iaf score vector element-wise multiplied by the related hf score matrix of A1 and A2, respectively. In consideration of the heterogeneity of the influence tendency of A1 and A2 shown in Fig. 7, the hf-iaf score is used to acquire the feature quantity expression of the

customers for the third proposal. Fig. 8 demonstrates the iaf score of A1 and A2, where the radius of the trade area of all supermarkets is 3 km. The trend is reversed from Fig. 7. Table VI shows the structure of experimental data for the RFM++ type model including RFM score and hf-iaf score.

D. Decyl Analysis

In economics, there is a theory which is known as 80/20 rule that 20% of customers account for 80% of sales.

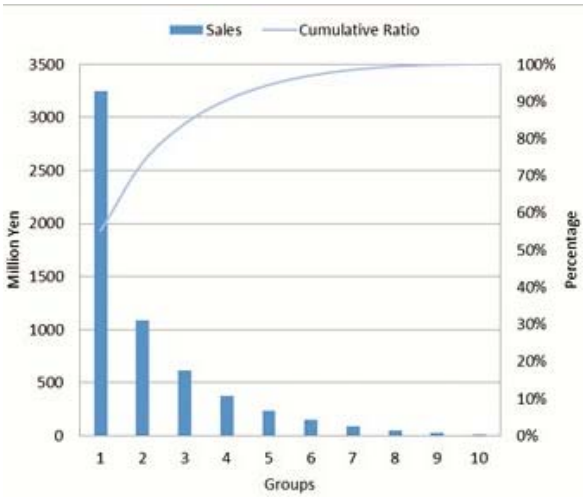

Fig. 9 Pareto chart of customer Decyl analysis

TABLE IV
THE STRUCTURE OF EXPERIMENTAL DATA FOR RFM TYPE MODEL

Customer	R score	F score	M score
1	5	5	1
2	5	3	5
⋮			
n	4	3	2

TABLE V
THE STRUCTURE OF EXPERIMENTAL DATA FOR RFM+ TYPE MODEL

Customer	R score	F score	M score	hf score vector
1	5	5	1	⋯
2	5	3	5	⋯
⋮				
n	4	3	2	⋯

TABLE VI
THE STRUCTURE OF EXPERIMENTAL DATA FOR RFM++ TYPE MODEL

Customer	R score	F score	M score	hf-iaf score vector
1	5	5	1	⋯
2	5	3	5	⋯
⋮				
n	4	3	2	⋯

Decyl analysis is derived from this theory. This paper refers to [5] and uses Decyl analysis to define loyal customers of supermarket chain A. Decyl analysis arranges customers in descending order of customer's consumption and then divides them into ten equal groups in terms of headcount, as shown in Fig. 9. The top three groups generated 80.01% sales in the first year. Therefore, loyal customers of supermarket chain A are defined as the top three groups. These three groups are considered as the target variables of logistic regression analysis to build the model for loyal customer classification in the first year. Decyl analysis also conducts on the customer shopping data in the second year for testing the built models.

IV. ANALYTICAL METHODS

A. Logistic Regression Analysis

The formula of logistic regression [6]-[9] is defined as follows: where p_c is a probability that customer c is a loyal customer, ω denotes partial regression coefficients, x indicates explanatory variables, and d represents bias.

$$p_c = \frac{1}{1+e^{(\omega_1 x_{1,c}+\omega_2 x_{2,c}+\cdots+\omega_k x_{k,c})}} \quad (4)$$

This paper uses logistic regression analysis for detecting the influence degree of the competitive supermarket on loyal customers of target supermarket chain A. The top three customer groups of Decyl analysis are tagged as target variables for the loyal customer classification. For the explanatory variables, this study adopts three RFM score indicators and the transformed distance factors by the proposed methods of this paper.

B. Evaluation Criteria

This paper uses the Accuracy, the Precision, the Recall and the F1-score as evaluation criteria. The formulas are defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (8)$$

Let TP be true positives that samples correctly classified as positive, FN be false negatives that samples incorrectly classified as negative, FP be false positives that samples incorrectly classified as positive, and TN be true negatives that samples correctly classified as negative. For supermarket competition analysis, this study detects the value of all partial regression coefficients and rejects all cases where the statistical significance level (p value) [19] is greater than 5%.

V. EXPERIMENT

A. Experimental Data

This study uses two-year ID-POS data of a supermarket chain A including A1 and A2 in a middle-size city in Japan. There are 40,977,672 sales records in the ID-POS data where the number of IDs and categorized products is 176076 and 2251, respectively. In addition, there are 30 supermarkets including two target supermarkets in that city.

TABLE VII
THE EXPERIMENTAL DATA OF TARGET SUPERMARKET CHAIN A

	First year (2016.04-2017.03)		Second year (2017.04-2018.03)
	Train data	Validation data	Test data
A	82029	27344	111772
A1	32690	10897	44020
A2	49339	16447	67752

Table VII shows the detail of experimental data. The ID-POS data is segmented into the first year and the second year. The first-year data are divided into training data and validation data to build the models. The second-year data (test data) are applied to test the models.

TABLE VIII
THE ACCURACY ANALYSIS FOR CHAIN A

	Accuracy	*Precision*	*Recall*	*F1-score*
RFM-A	91.89%	79.82%	97.47%	0.878
RFM+A	92.05%	80.28%	97.67%	0.881
RFM++3km-A	92.06%	80.31%	97.73%	0.882
RFM++4km-A	92.10%	80.46%	97.73%	0.883
RFM++5km-A	92.23%	80.78%	97.82%	0.885

TABLE IX
THE RFM MODEL FOR CHAIN A

Variables	***Coefficients***	***P values***
Intercept	-20.01	0.0%
R score	0.26	0.0%
F score	3.47	0.0%
M score	1.96	0.0%

TABLE X
THE RFM+ MODEL FOR CHAIN A

Variables	***Coefficients***	***P values***
Intercept	-22.27	0.0%
R score	0.25	0.0%
F score	3.77	0.0%
M score	2.06	0.0%
A2	0.15	0.0%
A1	0.07	0.0%
C2	0.06	0.1%
C7	0.03	1.5%
C6	0.03	3.2%
G	-0.03	0.0%
C9	-0.03	0.1%
C5	-0.04	0.0%
C3	-0.04	0.1%
B2	-0.04	0.0%
J2	-0.05	0.0%
E	-0.06	0.0%
C8	-0.06	0.0%
K	-0.06	0.0%
D	-0.44	0.0%

B. Experimental Procedure

The experiments are executed on Windows 8 with 2.50 GHz Intel Core i7 and 8 GB memory and conducted on the Python environment. For all cases, the store area is fixed as 1000 m^2 and the distance decline coefficient is fixed as 2. The smoothing coefficient is fixed as 1.

The two-year ID-POS data of the target supermarket chain A is divided into the first year as current customer information and the second year as future customer information. RFM score indicators will combine with 30 converted distance indicators of competitors to build feature quantities as shown in Tables IV-VI. Decyl analysis will also be conducted on 2-year ID-POS data to define the loyal customers, respectively. This research uses the customer ID, 33 feature quantities as explanatory variables and loyal customers as target variables to build experimental data in logistic regression analysis. The first-year experimental data are divided into two pieces, 75% for training data and 25% for validation data. The oversampling and undersampling problems [26] are judged that it is unnecessary in this experiment. The first-year experimental data are utilized to construct the models. The constructed model will be implemented on the second-year data (test data) to classify the loyal customers. There are two stages of the experiments in this research. The first stage is an experiment of loyal customer classification on the entire supermarket chain A. The second stage is an experiment of loyal customer classification on individual supermarkets of the target supermarket chain A. Both of the stages contain three types of model analysis, the RFM, RFM+ and RFM++ type model. The evaluation of the model is carried out from the viewpoints of accuracy, precision, recall rate, classification accuracy (F1-score), and feature understanding of loyal customers.

TABLE XI
THE RFM++3KM MODEL FOR CHAIN A

Variables	***Coefficients***	***P values***	**Variables**	***Coefficients***	***P values***
Intercept	-21.90	0.0%	E	-0.38	1.8%
R score	0.28	0.0%	C5	-0.39	0.1%
F score	3.74	0.0%	C9	-0.40	1.6%
M score	2.12	0.0%	B3	-0.41	0.0%
A2	0.85	0.1%	H1	-0.45	0.5%
C2	0.37	0.0%	C4	-0.46	0.0%
C7	0.22	0.0%	C3	-0.49	0.4%
A1	0.19	0.0%	G	-0.49	0.0%
B1	0.13	0.0%	B2	-0.50	0.0%
C6	0.10	0.0%	L	-0.51	0.0%
I	-0.03	0.0%	C8	-0.54	0.0%
J1	-0.03	0.0%	H2	-0.58	0.2%
N1	-0.05	0.0%	J2	-0.62	0.0%
M	-0.11	0.0%	N2	-0.88	3.6%
C1	-0.18	0.0%	K	-1.13	0.0%
F3	-0.19	0.0%	F2	-1.76	2.6%
F1	-0.23	0.0%	D	-5.39	0.0%

C. Experiment of Supermarket Chain

Table VIII presents the classification results of the proposed models for target supermarket chain A. In the experiments, this paper generates five models. 'RFM-A' model only includes three RFM score indicators. 'RFM+A' model contains three RFM score indicators and 30 hf score indicators. This research chooses 3 km, 4 km and 5 km radius of the trade area of each supermarket in the experimental city to generate the three 'RFM++A' models for supermarket chain A. The RFM++ type model is superior to the other two type models for loyal customer classification from the viewpoint of four evaluation criteria.

Table IX shows the partial regression coefficients of the

RFM type model for target supermarket chain A. The loyal customers have the attribution of high RFM scores since three partial regression coefficients are positive. This is consistent with intuitive understanding. By unifying the magnitude of the three indicators, the F value seems more critical for loyal customers because its value is maximal. However, it is difficult to thoroughly analyze the influence of supermarket competition by these three indicators.

TABLE XII
THE RFM++4KM MODEL FOR CHAIN A

Variables	*Coefficients*	*P values*	Variables	*Coefficients*	*P values*
Intercept	-23.51	0.0%	F1	0.70	0.0%
R score	0.28	0.0%	C9	0.59	0.1%
F score	3.81	0.0%	H1	0.56	0.1%
M score	2.14	0.0%	H2	0.45	2.6%
F2	1.63	0.0%	G	0.45	0.4%
B1	1.54	0.0%	B3	0.45	0.0%
I	1.47	0.0%	C5	0.41	0.4%
A2	1.34	0.0%	B2	0.40	0.3%
A1	1.31	0.0%	C4	0.37	0.0%
C2	1.21	0.0%	C3	0.25	0.0%
N2	1.10	1.1%	E	0.23	0.0%
C6	1.09	0.0%	J2	0.20	0.0%
F3	0.94	0.0%	C8	0.17	0.0%
C7	0.93	0.0%	N1	-0.13	0.0%
M	0.91	0.1%	K	-0.30	0.0%
J1	0.84	0.0%	C1	-0.63	0.0%
L	0.81	1.5%	D	-4.49	0.0%

Table X presents the competition analysis by the partial regression coefficients of the RFM+ type model. The coefficients are ordered by descending except for RFM scores and intercept. All competitors give loyal customers of the target supermarket chain A negative influence except for supermarket C2, C6, C7. The most of statistical significance level is less than 5%. This paper omits the cases that the statistical significance level higher than 5%. The values of target supermarkets A1 and A2 are positive, which is consistent with the intuitive idea since the closer to them, the more likely it becomes a loyal customer. The loyal customers are most active affected by supermarket D.

Tables XI-XIII demonstrate the partial regression coefficients of three RFM++ type models. The competitors have a powerful impact on customers of supermarket chain A who live in their 3 km radius of the trade area because most of the coefficients of the 'RFM++3km-A' model are negative. When the trade area radius increased to 4 km for all supermarkets, the influence of competitors decreases because most coefficients of the 'RFM++4km-A' model are positive. This tendency is also found in the case of the 'RFM++5km-A' model. The values of coefficients become positive and greater than the 'RFM++4km-A' model.

From these results, the analysis diversity of the RFM++ type model is superior to the other two type models. In addition, the RFM++ type model can grasp the impact of all nearby competitors since the statistical significance level is all less than 5%.

TABLE XIII
THE RFM++5KM MODEL FOR CHAIN A

Variables	*Coefficients*	*P values*	Variables	*Coefficients*	*P values*
Intercept	-23.79	0.0%	H1	0.98	0.0%
R score	0.27	0.0%	H2	0.97	0.0%
F score	3.77	0.0%	C9	0.96	0.0%
M score	2.11	0.0%	F3	0.91	0.0%
B1	1.84	0.0%	F2	0.86	0.0%
A2	1.83	0.0%	G	0.83	0.0%
A1	1.76	0.0%	N1	0.80	0.0%
M	1.72	0.0%	B2	0.77	0.0%
C2	1.67	0.0%	C5	0.76	0.0%
C6	1.48	0.0%	J2	0.68	0.0%
I	1.45	0.0%	E	0.66	0.1%
C7	1.43	0.0%	B3	0.59	2.7%
J1	1.26	0.0%	C8	0.54	0.1%
C4	1.25	0.0%	C3	0.53	0.9%
L	1.19	0.0%	K	0.26	0.0%
N2	1.06	1.0%	C1	-0.31	0.0%
F1	1.03	0.0%	D	-2.52	0.6%

D.Experiment of Individual Supermarkets

Tables XIV and XVII show the comparison of three models for A1 and A2. Similar to the experiments of supermarket chain A, this paper employs the first-year customer shopping data of A1 and A2 to build the RFM and RFM+ type models, and then classify the loyal customer in the second year. For the RFM++ type model, the constructed three models of target supermarket chain A, as shown in Tables XI-XIII, are used to classify the loyal customers of A1 and A2 in the second year. The results demonstrate the RFM++ type model is the most superior one in the loyal customer classification of individual supermarkets.

Tables XV and XVIII present the RFM type model for A1 and A2, respectively. Similar to the cases in supermarket chain A, the RFM scores in both cases are all positive. The RFM+ type model for A1 and A2 are shown in Tables XVI and XIX. It is interesting that even among A1 and A2 there is competition with each other. In the case of the RFM+ type model of A1, A2 has a negative value. So does the case in the RFM+ type model of A2. The statistical significance level is confirmed that most of the cases are less than 5%.

TABLE XIV
THE ACCURACY ANALYSIS FOR A1

	Accuracy	*Precision*	*Recall*	*F1-score*
RFM-A1	92.33%	79.71%	95.34%	0.868
RFM+A1	92.63%	80.51%	95.38%	0.873
RFM++3km-A	92.71%	80.81%	95.84%	0.877
RFM++4km-A	92.73%	80.81%	95.87%	0.877
RFM++5km-A	92.76%	80.94%	95.92%	0.878

TABLE XV
THE RFM MODEL FOR A1

Variables	*Coefficients*	*P values*
Intercept	-26.76	0.0%
R score	0.38	0.0%
F score	4.64	0.0%
M score	1.99	0.0%

TABLE XVI
THE RFM+ MODEL FOR A1

Variables	*Coefficients*	*P values*
Intercept	-25.89	0.0%
R score	0.39	0.0%
F score	4.51	0.0%
M score	2.09	0.0%
B1	0.26	0.0%
A1	0.05	0.0%
G	-0.03	1.1%
B2	-0.04	0.0%
C9	-0.04	0.1%
C8	-0.05	0.0%
E	-0.05	0.8%
C5	-0.05	0.0%
A2	-0.06	0.0%
K	-0.06	1.0%
J2	-0.06	0.0%
C1	-0.07	0.0%
D	-0.55	0.3%

TABLE XVII
THE ACCURACY ANALYSIS FOR A2

	Accuracy	*Precision*	*Recall*	*F1-score*
RFM-A2	91.48%	80.33%	94.80%	0.870
RFM+A2	91.58%	80.72%	96.53%	0.879
RFM++3km-A	92.85%	83.43%	97.12%	0.898
RFM++4km-A	92.98%	83.90%	96.80%	0.899
RFM++5km-A	93.05%	83.99%	96.74%	0.899

TABLE XVIII
THE RFM MODEL FOR A2

Variables	*Coefficients*	*P values*
Intercept	-23.15	0.0%
R score	0.23	0.0%
F score	3.85	0.0%
M score	2.40	0.0%

TABLE XIX
THE RFM+ MODEL FOR A2

Variables	*Coefficients*	*P values*
Intercept	-22.76	0.0%
R score	0.23	0.0%
F score	3.85	0.0%
M score	2.48	0.0%
A2	0.17	0.0%
B1	0.16	0.0%
C2	0.09	0.0%
B3	0.05	1.2%
F1	0.04	0.4%
B2	-0.02	0.0%
G	-0.03	0.1%
K	-0.04	3.8%
J2	-0.04	0.0%
M	-0.04	0.2%
C3	-0.05	0.8%
C8	-0.05	0.0%
A1	-0.09	0.0%
C1	-0.15	0.4%
D	-0.63	0.0%

VI. CONCLUSION

This paper constructs supermarket competition analysis models and makes loyal customer classification for a supermarket chain. In the experiments, this study estimates the RFM, RFM+ and RFM++ type model by accuracy and logistic regression coefficient analysis. All three model types can classify the loyal customers adequately. The RFM++ type model is superior to the other two type models from the viewpoint of accuracy and analysis diversity. The supermarket managers can grasp the influence degree of competitive supermarkets and understand the behavior of loyal customers. In the future, the presented models will be implemented on the sensitivity analysis of neural networks for supermarket competition analysis and loyal customer classification.

REFERENCES

[1] M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study," Procedia Computer Science 3, pp. 57-63, 2011.
[2] A. M. Hughes, Strategic database marketing, Chicago: Probus Publishing Company, 1994.
[3] H. C. Chang, H. P. Tsai, "Group RFM analysis as a novel framework to discover better customer consumption behavior," Expert Systems with Applications 38, pp. 14499-14513, 2011.
[4] J. Wu, Z. Lin, "Research on customer segmentation model by clustering," In Proceedings of the 7th ACM ICEC international conference on electronic commerce, 2005.
[5] T. Tanaka, T. Hamaguchi, T. Saigo, K. Tsuda, "Classifying and Understanding Prospective Customers via Heterogeneity of Supermarket Stores," International Conference on Knowledge Based and Intelligent Information and Engineering Systems, pp. 956-964, 2017.
[6] D. W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed. John Wiley \& Sons, Inc, 2000.
[7] T. Tjur, "Coefficients of determination in logistic regression models," American Statistician: pp. 366-372, 2009.
[8] D. A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, pp. 128, 2009.
[9] F. E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer, 2010.
[10] J. A. Morris, M. J. Gardner, "Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates," British Medical Journal, 1988.
[11] D. E. Farrar, R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," Review of Economics and Statistics, 49, issue 1, pp. 92-107, 1967.
[12] J. Correia, RFM-analysis, GitHub repository, 2016. (Online). Available: https://github.com/joaolcorreia/RFM-analysis
[13] D. L. Huff, "Defining and Estimating a Trade Area," Journal of Marketing, vol. 28, pp. 34-38, 1964.
[14] W. J. Reilly, The law of retail gravitation, New York: Knickerbocker Press, 1931.
[15] M. Nakanishi, L. G. Cooper, "Parameter estimation for a multiplicative competitive interaction model-least squares approach," Journal of Marketing Research, 11, pp. 303-311, 1974.
[16] D. B. Segal, "Retail Trade Area Analysis: Concepts and New Approaches," The Journal of Database Marketing, vol. 6, no. 3, pp. 267-277, 1999.
[17] K. Chen, Y. H. Hu, Y. C. Hsieh, "Predicting customer churn from valuable B2B customers in the logistics industry: a case study," Information Systems and e-Business Management, vol. 13, no. 3, pp. 475-494, 2015.
[18] H. Abdi, L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433-459, 2010.
[19] D. Freedman, R. Pisani, R. Purves, Statistics: Fourth International Student Edition. W.W. Norton & Company, 2007.
[20] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability, Chapman &

Hall/CRC, 1996.
[21] S. Zenker, T. Gollan, N. V. Quaquebeke, "Using Polynomial Regression Analysis and Response Surface Methodology to Make a Stronger Case for Value Congruence in Place Marketing," Psychology and Marketing, vol. 31, issue 3, pp. 184-202, 2014.
[22] Ruth M. W. Yeung, Wallace M. S. Yee, "Logistic Regression: An advancement of predicting consumer purchase propensity," The Marketing Review, vol. 11, no. 1, 2011.
[23] C. Constantin, "Using the Logistic Regression model in supporting decisions of establishing marketing strategies," Bulletin of the Transilvania University of Braşov Series V: Economic Sciences, vol. 8, issue 57, no. 2, 2015.
[24] H. Anton, Elementary Linear Algebra, 7th ed. John Wiley & Sons, pp. 170-171, 1994.
[25] C. Cui, J. Wang, Y. Pu, J. Ma, G. Chen, "GIS-based method of delimitating trade area for retail chains," International Journal of Geographical Information Science, vol. 26, no. 10, pp. 1863-1879, 2012.
[26] A. I. Marqués, V. García, J. S. Sánchez, "On the suitability of resampling techniques for the class imbalance problem in credit scoring," Journal of the Operational Research Society, vol. 64, no. 7, pp. 1060-1070, 2013.

Yepeng Cheng received the B.S.E. degree from the Dalian Minzu University, Dalian, China in 2013 and M.S. degree from the Hiroshima University, Higashihiroshima, Japan in 2018. He is currently a PhD student in the Department of Information Engineering, Graduate School of Engineering at the Hiroshima University where he is working with Professor Yasuhiko Morimoto. His research interests include data mining, machine learning and deep learning, especially logistic regression analysis, time series prediction.

Yasuhiko Morimoto is a Professor at Hiroshima University. He received his B.E., M.E. and Ph.D. degrees from Hiroshima University in 1989, 1991 and 2002 respectively. From 1991 to 2002, he had been with IBM Tokyo Research Laboratory where he worked for data mining project and multimedia database project. Since 2002, he has been with Hiroshima University. His current research interest include data mining, machine learning, geographic information system and privacy-preserving information retrieval.