

## A novel model for product bundling and direct marketing in e-commerce based on market segmentation

Arash Beheshtian-Ardakani<sup>a</sup>, Mohammad Fathian<sup>b\*</sup> and Mohammadreza Gholamian<sup>c</sup>

<sup>a</sup>Master of Science in Information Technology Engineering, School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

<sup>b</sup>Professor at School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

<sup>c</sup>Assistant professor at School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran

### CHRONICLE

#### Article history:

Received September 16, 2016

Received in revised format:

October 22, 2016

Accepted April 25, 2017

Available online

April 26, 2017

#### Keywords:

Product bundling

Direct marketing

Market segmentation

Customer loyalty

Personalization

Electronic commerce

### ABSTRACT

Nowadays, companies offer product bundles with special discounts in order to sell more products. However, it is important to note that customers show different levels of loyalties to companies, and each segment of the market has unique features, which influences the customers' buying patterns. The primary purpose of this paper is to propose a novel model for product bundling in e-commerce websites by using market segmentation variables and customer loyalty analysis. RFM model is employed to calculate customer loyalty. Subsequently, the customers are grouped based on their loyalty levels. Each group is then divided into different segments based on market segmentation variables. The product bundles are determined for each market segment via clustering algorithms. Apriori algorithm is also used to determine the association rules for each product bundle. Classification models are applied in order to determine which product bundle should be recommended to each customer. The results demonstrate that the silhouette coefficient, support, confidence, and accuracy values are higher when both customer loyalty level and market segmentation variables are used in product bundling. Accordingly, the proposed model increases the chance of success in direct marketing and recommending product bundles to customers.

© 2018 Growing Science Ltd. All rights reserved.

## 1. Introduction

Electronic commerce (e-commerce) refers to using networks such as the Internet to purchase or sell products and services. Business-to-consumer (B2C) is a term in electronic commerce in which online transactions are made between businesses and individual consumers. Hence, companies can sell their products directly to the consumers (Turban et al., 2015). Marketing is the act by which companies create values for customers and build relationships with them for receiving values from the customers in return. In the market segmentation process, a market is separated into groups of customers with distinct characteristics that might require different marketing mixes. Companies can adopt appropriate

\* Corresponding author. +9821-73225048

E-mail address: [fathian@iust.ac.ir](mailto:fathian@iust.ac.ir) (M. Fathian)

marketing strategies based on the characteristics of different segments or select the market segments on which they want to focus. Therefore, they can perform target marketing at the right time and place. One of the most important marketing strategies is direct marketing. Direct marketing includes establishing direct connections with individual customers to obtain an instant response and long-lasting relationships with them. In this way, companies offer products directly to the customers (Kotler & Armstrong, 2014).

Personalization is the process in which a company decides what marketing mix is appropriate for individual customers (Arora et al., 2008). Product bundling refers to selling two or more products together, packaged at a lower price in comparison with the total cost of independent prices. Product bundling determines which products should be combined as a bundle (Beladev et al., 2016). A number of e-commerce websites offer product bundles with discounts to their customers. However, it is important to note that each segment of the market has its own characteristics and needs. Moreover, customers show different levels of loyalty to companies. Consequently, they should pay special attention to how they personalize their product bundles. They should also consider the level of customer loyalty and the characteristics of the segment in which the customer belongs.

The present study compares different scenarios of product bundling by considering customer loyalty and demographic and geographic variables to prove the superiority of the proposed model to other existing models. The rest of the paper is organized as follows. Section 2 includes the literature review. Section 3 describes the data mining techniques employed. Section 4 discusses the research methodology, and Section 5 presents the empirical analysis and results. Finally, Section 6 provides the conclusion of the paper.

## **2. Literature review**

The major variables for segmenting consumer markets are demographic, geographic, psychographic, and behavioral variables. For instance, loyalty status is a behavioral variable for market segmentation (Kotler & Armstrong, 2014). Several researchers have used these types of variables in data mining techniques for market segmentation (Kuo et al., 2002; Bloom, 2005; Lee & Park, 2005; Kuo et al., 2006; Huang et al., 2007). Lu and Wu (2009) proposed a customer segmentation method based on customers' transaction patterns. Hsu et al. (2012) used the idea of the hierarchy of the items consumed to segment customers. Dutta et al. (2015) categorized the data mining techniques employed in market segmentation into thirteen methods, such as neural network, RFM analysis, hierarchical clustering, and K-means.

A great number of researchers have employed recency, frequency, and monetary value (RFM) model to determine the level of customers' loyalty to the company. In this method, customers of a company are divided into different groups based on RFM variables. In RFM, R shows the recency of the last purchase, F represents the frequency of the purchases, and M signifies the monetary value of the purchases (Linoff & Berry, 2011). Using RFM analysis, companies can better understand their customers' profitability. As a result, they can adopt appropriate marketing strategies to deal with different customers (Chen et al., 2012).

Tsai and Chiu (2004) employed a RFM model to examine the profitability of market segments. Chen et al. (2012) used K-means and decision tree techniques to segment the customers of an online retailer based on RFM model. Shim et al. (2012) first determined important customers based on RFM analysis. Afterwards, they classified the customers into VIP and non-VIP categories by using data mining techniques. Subsequently, they identified the association rules from the VIP transactions. Sarvari et al. (2016) found out that the effectiveness of customer segmentation would be improved by using both of RFM and demographics variables.

Data mining techniques can be used to analyze customers' data and purchases in order to develop product packages (Karimi-Majd & Fathian, 2017). Stremersch and Tellis (2002) suggested that companies can increase their profits by using bundling strategies. They recognized two key dimensions in the categorization of bundling strategies: the focus of bundling on price or product and the form of bundling, whether none, pure, or mixed.

Yang and Lai (2006) compared different product bundling models for a publisher's website. They showed that the publisher can provide better product bundles by integrating shopping-cart and browsing data instead of using only browsing data or order data. Miguéis et al. (2012) employed VARCLUS algorithm to cluster the products of a retailing company and to determine the customers' lifestyles. They used their proposed model to assign each product item to a unique cluster. Consequently, they determined the type of lifestyle in each cluster by finding the product categories that formed the largest part of each cluster. Each customer was then assigned to the lifestyle cluster whose shopping basket had the closest similarity to the customer's past purchases.

Liao et al. (2011) applied the algorithms of K-means and association rules to the databases of a company to propose solutions for direct marketing. Liao et al. (2011) used SPSS Two-Step clustering algorithm to group customers. They achieved three clusters and named them low-, medium-, and high-frequency consumption groups according to the characteristics of each group. Subsequently, they applied Apriori algorithm to determine which categories of products are preferred more in the hypermarket studied. They suggested that managers can use this piece of information to bundle their products.

Beladev et al. (2016) proposed a model for product bundling and price bundling by integrating demand functions, collaborative filtering techniques, and price modeling. They employed a recommender system platform to identify bundles and prices, maximizing both the likelihood of buying a product bundle by the customer and the revenue earned by selling that product bundle. Cataldo and Ferrer (2017) presented a programming model to find the optimal composition and pricing of multiple product bundles.

To achieve better marketing and customer relationship management (CRM) strategies, researchers have investigated market segmentation variables and have analyzed customer loyalty levels. However, they have not tried to provide product bundles that suit the characteristics of each market segment by considering both market segmentation variables and customer loyalty analysis. In the present study, a novel model is proposed for product bundling by considering both market segmentation variables and customer loyalty analysis.

### **3. Data mining techniques**

Data mining is the process of analyzing large amounts of data to discover meaningful information and patterns from a data set (Kantardzic, 2011). Cross-industry standard process for data mining (CRISP-DM) can be applied for fitting the data mining process into the overall business (Larose & Larose, 2015).

#### *3.1. Classification*

A classification technique finds a model that describes data classes and predicts the class labels of objects. There are several classification models, such as the decision tree, artificial neural networks, K-nearest neighbors, Bayesian network, support vector machine (Han et al., 2011), and logistic regression (Kantardzic, 2011).

The accuracy measure can be employed to assess the performance of the classification models. The accuracy is the proportion of true results to the total number of cases examined. The accuracy measure is shown in Eq. (1) (Han et al., 2011).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

### 3.2. Clustering

A clustering technique clusters a set of objects so that the objects in each cluster have a number of common properties (Han et al., 2011). There are several clustering algorithms, such as the K-means (Kantardzic, 2011), SPSS Two-Step (SPSS Inc., 2001), and Kohonen self-organizing feature map (SOFM or SOM) (Tan et al., 2005). The quality of the clustering algorithms can be obtained according to the silhouette coefficient given in Eq. (2). In this equation, A is the average distance of the objects within a cluster from each other, and B is the minimum average distance of the objects within the cluster from other clusters.

$$Silhouette = \frac{(B - A)}{\text{Max}(A, B)} \quad (2)$$

The value of the silhouette coefficient is between [-1, 1]; 1 represents the highest clustering quality, and -1 signifies the lowest clustering quality. To measure the quality of a clustering algorithm, the average silhouette coefficient value of all objects in the data set can be used (Han et al., 2011).

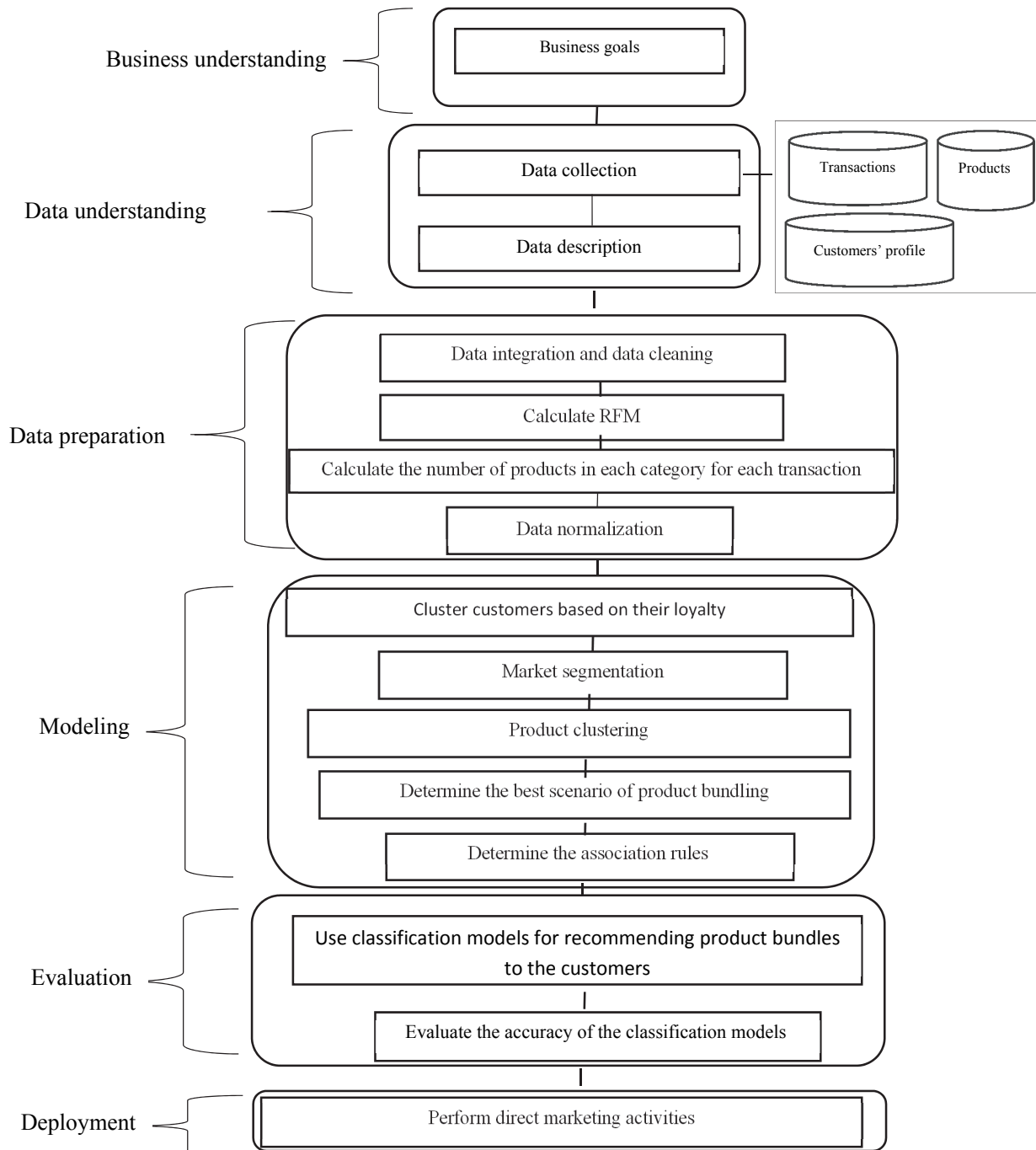
### 3.3. Association rules

Association rules show the relationships between the item sets and discover the buying patterns in customers' transactions. Consequently, they can be employed in market basket analysis. An association rule is shown in the form of  $X \Rightarrow Y$ , where X and Y are the item sets and are not equal. X is the antecedent and Y is the consequent of an association rule. Association rules must satisfy the minimum thresholds of the support and confidence measures in order to be interesting. The rule support is the proportion of the number of transactions including X and Y to the total number of transactions. The support shows the frequency of the patterns occurring in the rule. The rule confidence is the proportion of the number of transactions including X and Y to the number of transactions including X. The confidence signifies the strength of implication of the rule. The Apriori algorithm can be applied to extract the association rules from the transactions data (Kantardzic, 2011).

## 4. Research methodology

The steps of the research methodology are demonstrated in Fig. 1. In the business-understanding phase, the business goals of the e-commerce company are studied. In the data-understanding phase, the customer transaction data are collected from the databases of the company, and the descriptions of these data are defined. In the data-preparation phase, the data collected are preprocessed to be used in data mining algorithms. First, data integration and data cleaning are performed. Afterwards, the values of the RFM model for the customers are calculated. In addition, in each consumer transaction, the number of products in each category is calculated. The z-score formula is then used in order to normalize these values. In the modeling phase, the model proposed for product bundling is presented. First, the customers are clustered based on their loyalty level, and then the procedure of market segmentation is performed based on the geographic and demographic variables. In the product-clustering step, the product bundles are determined by applying the clustering algorithms to the customers' transactions in

each market segment. Following that, the Apriori algorithm is used to determine the association rules between the products in each product bundle. In the evaluation phase, a classification model is employed to predict the number of a suitable product bundle for the customers. Thus, the company can recommend personalized product bundles to its customers. In the deployment phase, the company performs direct marketing by sending the information of the personalized product bundles to its customers via emails.



**Fig. 1.** Research methodology

## 5. Empirical analysis and results

In this section, all the steps in each phase of the research methodology are investigated in detail. IBM SPSS Modeler 18 software is also employed to implement these steps.

### 5.1. Business understanding

In the present research, the company selected is an electronic retailer selling foods, drinks, proteins, junk foods, healthcare products, and kitchen tools. The goal of the company is to personalize its product bundles for the customers. The company also aims to recommend product bundles with different discounts to its customers based on their loyalty level and the segment of markets to which they belong.

### 5.2. Data understanding

The data for this research were collected from the electronic transactions of the company with its customers. These data include 541910 records received from 4340 customers from December 2014 to December 2015 in the cities of Golestan province in Iran. The information used in the present study were obtained from the three databases of customers' profile, transactions, and products. The attributes of these databases are presented in Table 1.

**Table 1**  
Attributes of the databases of the company

Database	Attributes
Customers' profile	Customer ID, Customer Name, Gender, Address, Phone number
Products	Product code, Product Name, Price, Category
Transactions	Invoice No., Product code, Customer ID, Quantity, Transaction Date

The product categories of the company and some examples of the products belonging to each product category are given in Table 2.

**Table 2**  
Product categories

Category	Products
Foods	Dairy, Oil, Beans, Grain, Rice, Flour, Sauce, Canned, Conserve, Compote, Soup
Drinks	Water, Beverage, Energy drink, Tea, Coffee, Juice
Proteins	Meat, Chicken, Fish, Sausage, Egg, Burger
Junk foods	Chips, Snack, Biscuit, Cake, Chocolate, Gum, Candy
Healthcare products	Detergents, Napkin, Tissue, Toothpaste, Toothbrush, Shampoo, Soap, Hand cream
Kitchen tools	Washing gloves, Kitchen sponge, Freezer bag, Garbage bag, Aluminum foil, Plastic dinnerware

### 5.3. Data preparation

The primary keys of the databases of transactions, products, and customers' profile are used to merge them and achieve an integrated data set. Three records of the integrated data set is shown in Table 3.

**Table 3**  
The integrated data set of the company

Record No.	Customer ID	Gender	Product code	Invoice No.	Quantity	Invoice Date	Unit Price	Total	Category	City Code
1	11850	M	14029	506365	6	2014-12-01 08:26	3.390	20.340	Drinks	1001
2	11850	M	12752	506365	2	2014-12-01 08:26	7.650	15.300	Foods	1001
3	11850	M	13123	506365	6	2014-12-01 08:26	2.550	15.300	Healthcare products	1001

The data-cleaning process is done by eliminating the missing and noisy values. After implementing this process, the number of records decreased from 541910 to 406742. In the attribute-construction step, recency, frequency, and monetary value are calculated from each customer's transactions in order to be used in clustering the customers into loyalty levels. In addition, the number of products in each product category is determined for each customer transaction to be used in product clustering. For example, in a consumer transaction, the two items of dairy and oil from the foods category exist. After calculating RFM values and the number of products in each category, they should be normalized prior to being used in the clustering algorithms. Accordingly, the z-score formula is used for data normalization. The z-score formula is shown in Eq. (3). In z-score normalization, the value ( $v$ ) of an attribute ( $A$ ) is normalized to a new value ( $V'$ ) based on the mean ( $A'$ ) and the standard deviation ( $\sigma_A$ ) of the attribute ( $A$ ) as follows (Han et al., 2011):

$$V' = \frac{v - A'}{\sigma_A} \quad (3)$$

#### 5.4. Modeling

To propose our model, the K-means clustering algorithm is applied to the variables of recency, frequency, and monetary value in order to group customers based on their loyalty level. Afterwards, the market segmentation variables are used in each customer loyalty level. The customers are partitioned based on the cities where they live. Subsequently, for each city, the customers are separated based on their gender. The product bundles are determined by applying the K-means clustering algorithm to the customers' transactions in each market segment. Finally, the Apriori algorithm is employed to determine the association rules in each product bundle. Fig. 2 shows the proposed model and Fig. 3 illustrates an example for these steps.

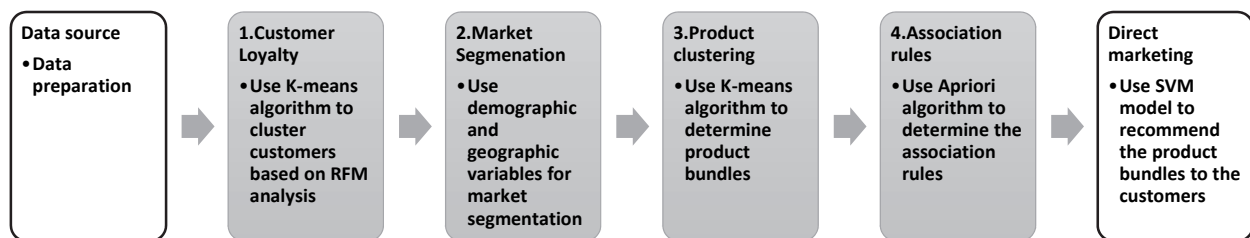


Fig. 2. Proposed model

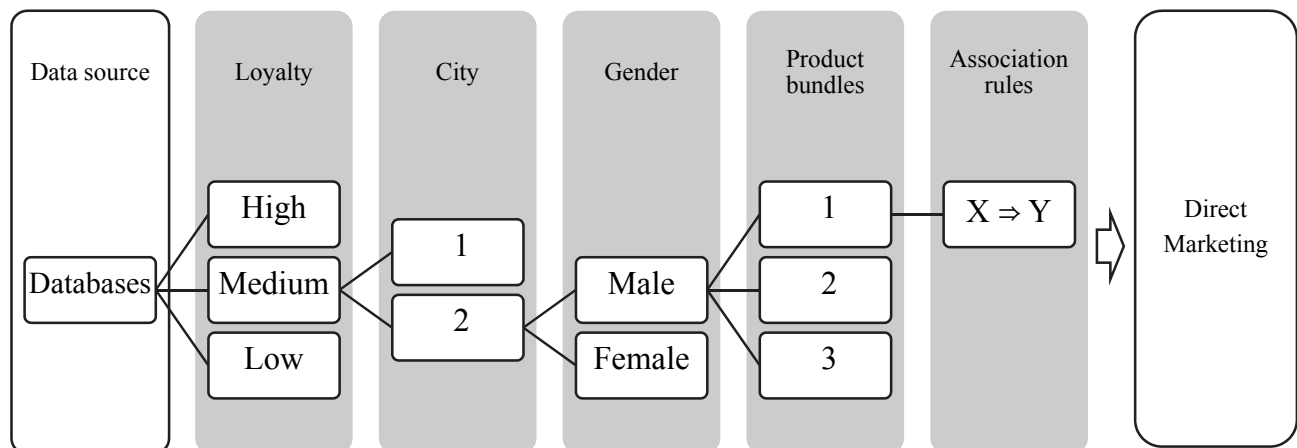


Fig. 3. An example of the steps of the proposed model

### 5.4.1. Customer loyalty

In this step, the K-means, SPSS Two-Step, and Kohonen SOM clustering algorithms are applied to the RFM values, and the amount of silhouette value for each algorithm is calculated. The number of clusters in K-means algorithm must be defined before executing the algorithm. Therefore, silhouette values are compared for different numbers of clusters in K-means algorithm. Table 4 indicates the results.

**Table 4**  
Comparing silhouette values of K-means algorithm for customer loyalty

Number of clusters	Silhouette
2	0.6614
3	0.7091
4	0.5542
5	0.5165

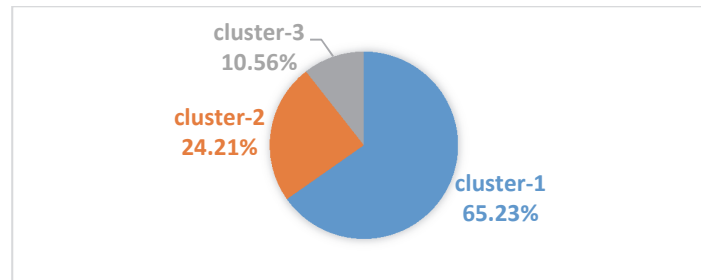
Table 5 compares the silhouette value and the number of clusters for the clustering algorithms of K-means, SPSS Two-Step, and Kohonen SOM to determine customers' loyalty levels.

**Table 5**  
Comparison of the clustering algorithms for customer loyalty

Algorithm	Number of clusters	Silhouette
K-means	3	0.7091
SPSS Two-Step	4	0.524
Kohonen SOM	12	0.404

By considering the table above, it can be observed that the K-means algorithm with three clusters has the highest silhouette value. Therefore, the K-means algorithm with three clusters is used for clustering the customers into their loyalty level groups.

Fig. 4 demonstrates the percentage of each cluster of customer loyalty clusters. These clusters are the levels of customer loyalty to the company. Cluster-1, cluster-2, and cluster-3 show the medium-, low-, and high-loyalty level of the customers, respectively. The majority of the customers belong to the medium loyalty level.



**Fig. 4.** Percentage of customer loyalty clusters

### 5.4.2. Market segmentation

The company studied in this paper keeps in its databases the addresses of the customers and their gender attributes. Therefore, the city attribute is used for geographic segmentation and the gender attribute for demographic segmentation. There are eight scenarios of product bundling in our research, as shown in Table 6. The ✓ symbols in Table 6 indicate which variables are used in these scenarios.



**Table 6**

Different scenarios based on customer loyalty level and market segmentation variables

Scenario	Loyalty	City	Gender
1			
2		✓	
3			✓
4	✓		
5		✓	✓
6	✓		✓
7	✓	✓	
8	✓	✓	✓

### 5.4.3. Product clustering

In this step, the product bundles are determined for the eight scenarios specified in the previous step. The K-means, SPSS Two-Step, and Kohonen SOM clustering algorithms are applied to the customers' transactions, and the amount of silhouette value for each algorithm is calculated. Applying the clustering algorithms to the customers' transactions, the products, which are usually purchased with a similar transaction pattern, are placed in the same cluster. Hence, the product bundles can be determined in each of the eight scenarios. Two different situations, which were randomly selected, are considered to propose our model and show the results of performing the proposed model in these situations. Table 7 presents these situations.

**Table 7**

Two example situations used in the proposed model

Situation	Customer loyalty level	City code	Gender
1	Medium	1002	Male
2	High	1001	Female

In what follows, the steps of scenario 1 and scenario 8 are presented to compare the product bundles determined in these scenarios. situation 1 is used to investigate the proposed model. The final results of situation 2 are then presented to compare them with those of situation 1.

Table 8 shows the silhouette values for different numbers of clusters in K-means algorithm for scenario 1 in situation 1.

**Table 8**

Comparison of the silhouette values for different numbers of clusters in K-means algorithm for scenario 1 in situation 1

Number of clusters	Silhouette
2	0.3236
3	0.1896
4	0.3105
5	0.2885
6	0.2761
7	0.3308
8	0.2206
9	0.3094
10	0.2874

Table 9 indicates the silhouette value and the number of clusters for the K-means, SPSS Two-Step, and Kohonen SOM clustering algorithms for scenario 1 in situation 1.

**Table 9**

Comparison of the clustering algorithms for scenario 1 in situation 1

Algorithm	Number of clusters	Silhouette
K-means	7	0.3308
SPSS Two-Step	4	0.285
Kohonen SOM	13	0.102

The K-means algorithm with seven clusters is used for clustering the products into product bundles in this scenario. Fig. 5 demonstrates the percentage of each cluster of product clusters for scenario 1 in situation 1. These clusters are the product bundles for scenario 1 in situation 1.

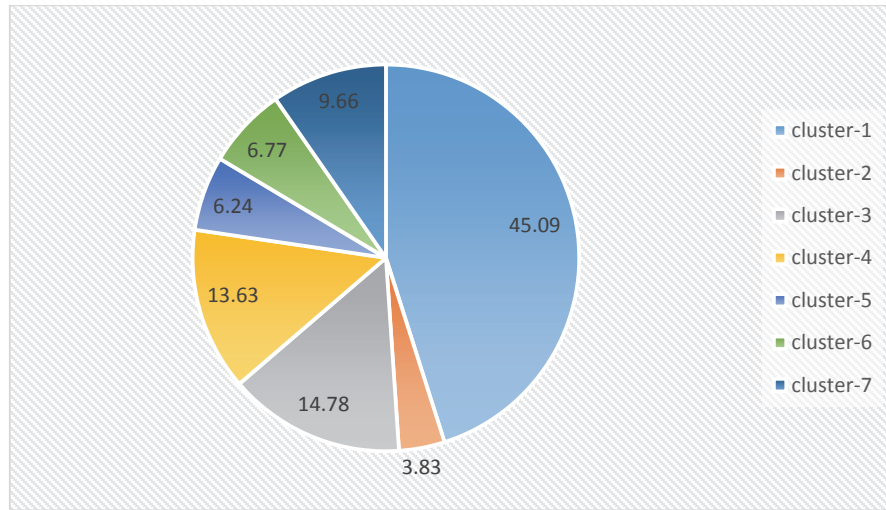
**Fig. 5.** Percentage of product clusters for scenario 1 in situation 1

Table 10 presents the percentages of the product categories in each product bundle for scenario 1 in situation 1. In this scenario, junk foods and healthcare products categories form the largest parts of the product categories.

**Table 10**

Percentages of product categories in each product bundle for scenario 1 in situation 1

Product bundles	Proteins (%)	Kitchen tools (%)	Junk foods (%)	Healthcare products (%)	Foods (%)	Drinks (%)
cluster-1	3.922	4.724	30.606	32.349	9.205	19.194
cluster-2	5.674	6.531	26.615	32.850	10.396	17.933
cluster-3	3.160	3.722	49.536	20.658	6.010	16.914
cluster-4	7.480	3.570	20.194	35.284	14.406	19.065
cluster-5	4.361	6.005	41.985	22.468	7.193	17.988
cluster-6	6.905	6.288	20.008	35.720	12.498	18.580
cluster-7	4.376	13.711	26.201	30.382	8.072	17.258

**Table 11**

Comparison of the silhouette values for different numbers of clusters in K-means algorithm for scenario 8 in situation 1

Number of Clusters	2	3	4	5	6	7	8	9	10
Silhouette	0.4429	0.501	0.456	0.4432	0.3407	0.3547	0.3241	0.3183	0.3268

Table 11 shows the silhouette values for different numbers of clusters in K-means algorithm for scenario 8 in situation 1.

Table 12 indicates the silhouette value and the number of clusters for the K-means, SPSS Two-Step, and Kohonen SOM clustering algorithms for scenario 8 in situation 1.

**Table 12**

Comparison of the clustering algorithms for scenario 8 in situation 1

Algorithm	Number of clusters	Silhouette
K-means	3	0.501
SPSS Two-Step	2	0.449
Kohonen SOM	9	0.238

The K-means algorithm with three clusters has the highest silhouette value for scenario 8 in situation 1. Therefore, the K-means algorithm with three clusters is used for clustering the products into product bundles in this scenario. Fig. 6 demonstrates the percentage of each cluster of product clusters for scenario 8 in situation 1. These clusters are the product bundles for scenario 8 in situation 1. Most of the customers' buying patterns in this scenario are assigned to cluster-1, as shown in Fig. 6. Therefore, cluster-1 is the most important product bundle for scenario 8 in situation 1.

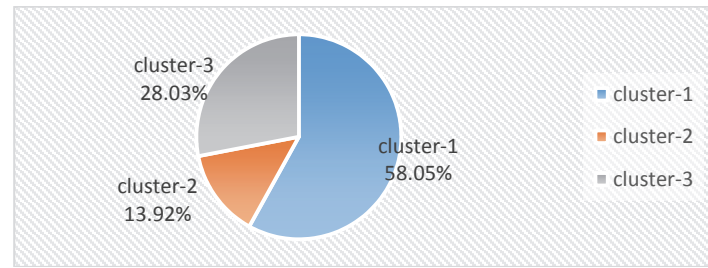
**Fig. 6.** Percentage of product clusters for scenario 8 in situation 1

Table 13 presents the percentages of the product categories in each product bundle for scenario 8 in situation 1. In this scenario, foods and kitchen tools categories form the largest parts of the product categories. Foods category also has the highest percentage of products in cluster-1 product bundle. Therefore, the company should consider the importance of these categories in product bundling for scenario 8 in situation 1.

**Table 13**

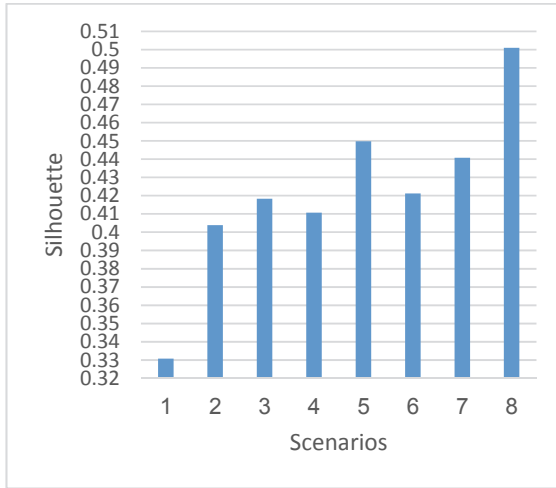
Percentages of product categories in each product bundle for scenario 8 in situation 1

Product bundles	Proteins (%)	Kitchen tools (%)	Junk foods (%)	Healthcare products (%)	Foods (%)	Drinks (%)
cluster-1	0.312	0.000	26.219	2.614	41.202	29.653
cluster-2	0.555	5.096	8.022	2.926	58.274	25.126
cluster-3	0.065	49.773	15.010	1.105	20.078	13.970

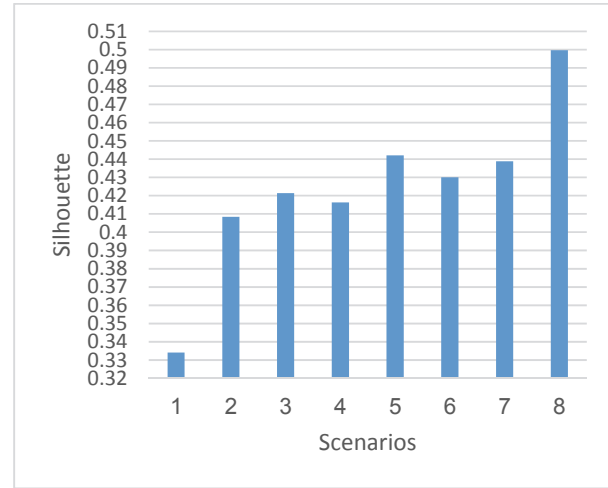
Considering the above-mentioned results, if customer loyalty level and market segmentation variables are used, the number of product bundles in each market segment, the products in each product bundle, and the percentages of the product categories in each product bundle will change.

Table 14 shows the maximum silhouette values of the K-means, SPSS Two-Step, and Kohonen SOM clustering algorithms for the scenarios examined in situation 1. In Table 14, scenario 8 is our proposed model in this study, and other scenarios are similar to the models in previous works of literature. The maximum silhouette value in each scenario is obtained, using the K-means algorithm. Hence, K-means algorithm is employed for product clustering in our proposed model.

Fig. 7 compares the silhouette values of K-means algorithm for the scenarios examined in situation 1.



**Fig. 7.** Comparing silhouette values of K-means algorithm for all eight scenarios in situation 1



**Fig. 8.** Comparing silhouette values of K-means algorithm for all eight scenarios in situation 2

**Table 14**

Comparison of the clustering algorithms for all eight scenarios in situation 1

Scenario	Algorithm	Number of clusters	Silhouette
<b>1</b>	K-means	7	0.3308
	SPSS Two-Step	4	0.285
	Kohonen SOM	13	0.102
<b>2</b>	K-means	7	0.4039
	SPSS Two-Step	4	0.385
	Kohonen SOM	12	0.175
<b>3</b>	K-means	4	0.4184
	SPSS Two-Step	3	0.408
	Kohonen SOM	12	0.109
<b>4</b>	K-means	5	0.4107
	SPSS Two-Step	4	0.361
	Kohonen SOM	12	0.160
<b>5</b>	K-means	4	0.4498
	SPSS Two-Step	3	0.439
	Kohonen SOM	12	0.198
<b>6</b>	K-means	4	0.4213
	SPSS Two-Step	3	0.409
	Kohonen SOM	12	0.161
<b>7</b>	K-means	4	0.4408
	SPSS Two-Step	3	0.396
	Kohonen SOM	9	0.251
<b>8</b>	K-means	3	0.501
	SPSS Two-Step	2	0.449
	Kohonen SOM	9	0.238

The highest silhouette value belongs to scenario 8. Therefore, the best distinct product bundles can be obtained, provided that both customer loyalty and market segmentation variables are employed. To investigate the validity of the proposed model, situation 2 is also considered to determine the product bundles. Fig. 8 compares the silhouette values of K-means algorithm for the scenarios examined in situation 2. In this situation, like situation 1, the highest silhouette value belongs to scenario 8.

Table 15 indicates the works which have the closest similarity with our proposed model. In our proposed model, customer loyalty level and market segmentation variables are used before the product-clustering step. Therefore, the product bundles can be determined with higher accuracy.

**Table 15**  
Comparison of the proposed model with similar works

Model	Loyalty analysis	Demographic analysis	Geographic analysis	Product clustering	Product bundling
Liao et al. (2011)		✓			✓
Miguéis et al. (2012)				✓	
Stremersch and Tellis (2002)					✓
Beladev et al. (2016)					✓
Sarvari et al. (2016)	✓	✓			
The proposed model	✓	✓	✓	✓	✓

#### 5.4.4. Determining the association rules

The Apriori algorithm is used to determine the association rules between the products and to discover which products are frequently bought together in each product bundle. These association rules assist the company in selecting the best product items for bundling in a package. Table 16 shows the first five association rules for scenario 1 in situation 1 sorted according to the confidence value.

**Table 16**  
Association rules for scenario 1 in situation 1

Antecedent (Product code)	Consequent (Product code)	Confidence (%)	Rule Support (%)
22698	22697	89.45	2.104
22699			
22698	22699	84.783	2.104
22697			
22697	22699	83.422	1.683
22423			
22698	22697	82.734	2.482
22698	22699	78.417	2.352

Table 17 shows that the first five association rules for scenario 8 in situation 1, in which customer loyalty level and market segmentation variables were used, are sorted according to the confidence value.

**Table 17**  
Association rules for scenario 8 in situation 1

Antecedent (Product code)	Consequent (Product code)	Confidence (%)	Rule Support (%)
22698	22697	91.139	28.8
22699			
22629	22727	90.909	4.0
22726			
22698	22699	90.0	28.8
22697			
22197	22699	90.0	3.6
22698			
22197	22697	90.0	3.6
22698			

The tables above show that the values of support and confidence measures in scenario 8 are more than scenario 1 in the same situation. Therefore, the products are bundled more appropriately in scenario 8.

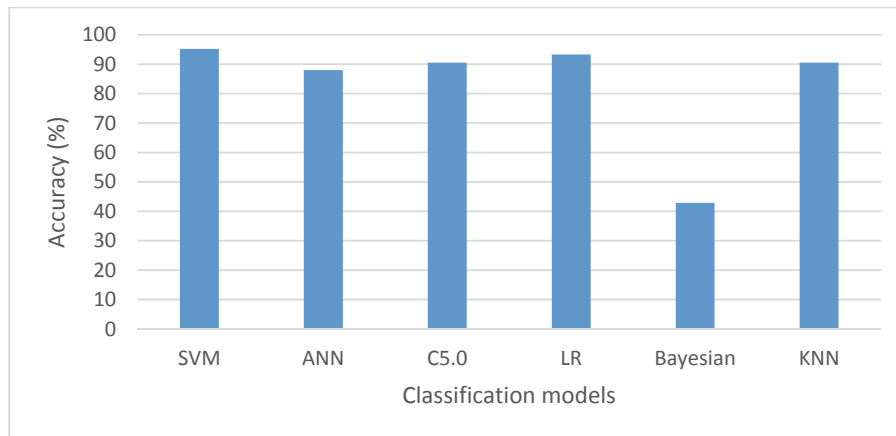
#### 5.5. Evaluation

Previous steps showed that if customer loyalty level and market segmentation variables are used, better product bundles will be obtained. Therefore, scenario 8 is the best scenario of product bundling. The

product bundles determined in this scenario are recommended to the customers of the company. In what follows, the efficacy of this scenario to recommend product bundles to customers is evaluated.

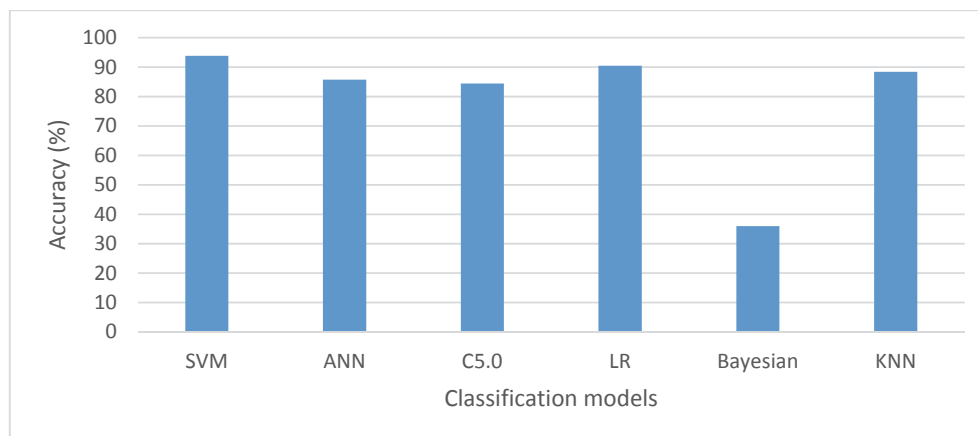
Artificial neural network (ANN), Bayesian network, K-nearest neighbor (KNN), logistic regression (LR), C5.0 decision tree, and support vector machine (SVM) are applied to recognize which product bundles should be recommended to the customers. The product bundle numbers obtained from the product-clustering step are also considered as the class labels in these classification models.

Each customer's transactions are given to the classification models, and the output of these classification models will be the class label of the product bundle, which has the closest similarity to the customer's transaction pattern. Sixty-seven percent of the transaction data are randomly selected for training and the rest for testing the classification models. Fig. 9 demonstrates the accuracy results of the classification models in situation 1. The SVM model has the highest accuracy among these classification models.



**Fig. 9.** Accuracy results of the classification models in situation 1

To investigate the validity of the proposed model, situation 2 is also evaluated to recommend the product bundles to the customers. Fig. 10 shows the accuracy results of the classification models in situation 2.



**Fig. 10.** Accuracy results of the classification models in situation 2

In this situation, like situation 1, the SVM model has the highest accuracy among the classification models. Therefore, SVM classification model is used in our proposed model to determine which product bundle should be recommended to each customer.

## 5.6. Deployment

By deploying the proposed model, companies can personalize the product bundles based on the customers' characteristics and needs. In addition, the steps of the proposed model should be repeated regularly, for example monthly, to utilize recent data stored in the databases of companies.

Direct marketing of the product bundles can be performed both online and offline. For online marketing, email marketing can be performed by sending the information about the suitable product bundles to customers' emails. For offline marketing, direct mail marketing can be performed by sending catalogues of product bundles to customers' home addresses.

## 6. Conclusion

In this study, a novel model was proposed for product bundling and direct marketing in e-commerce based on market segmentation. Initially, the data in the databases of an e-commerce company were collected and preprocessed. Subsequently, based on RFM model, the customers were clustered into loyalty levels, using K-means algorithm. Afterwards, market segmentation was executed, using demographic and geographic variables. The product bundles were then determined, using K-means algorithm. The association rules in each product bundle were also determined by Apriori algorithm. Finally, SVM classification model was applied to recommend the product bundles to the customers.

The proposed model is a new path for personalization of product bundles, and marketing managers can employ it in their decisions. By applying the proposed model, the product bundles are more precisely determined. Companies can perform direct marketing and recommend suitable product bundles to their customers by considering different customer loyalty levels and characteristics of each market segment. They can also offer different types of promotions such as discounts and coupons to the customers. Therefore, companies can attract new customers, motivate the existing customers to purchase, and retain the valuable customers.

As future works, a response model can be constructed for the direct marketing model proposed in this paper. In addition, price bundling models can be proposed by considering the customers' characteristics and needs. Finally, to propose a real-time system for recommending product bundles to customers, recommender systems can be implemented by using market segmentation and customer loyalty analysis.

## References

- Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., ... Zhang, Z. J. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, 19(3), 305–321.
- Beladev, M., Rokach, L., & Shapira, B. (2016). Recommender systems for product bundling. *Knowledge-Based Systems*, 111, 193–206.
- Bloom, J. Z. (2005). Market segmentation. A neural network application. *Annals of Tourism Research*, 32(1), 93–111.
- Cataldo, A., & Ferrer, J. (2017). Optimal pricing and composition of multiple bundles: A two-step approach. *European Journal of Operational Research*, 259(2), 766–777.
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), 197–208.
- Dutta, S., Bhattacharya, S., & Guin, K. K. (2015). Data Mining in Market Segmentation: A Literature Review and Suggestions. In K. N. Das, K. Deep, M. Pant, J. C. Bansal, & A. Nagar (Eds.), *Proceedings of Fourth International Conference on Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing, Volume 335* (pp. 87–98). New Delhi: Springer India.

- Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham, MA: Elsevier Science.
- Hsu, F. M., Lu, L. P., & Lin, C. M. (2012). Segmenting customers by transaction data with concept hierarchy. *Expert Systems with Applications*, 39(6), 6221–6228.
- Huang, J. J., Tzeng, G. H., & Ong, C. S. (2007). Marketing segmentation using support vector clustering. *Expert Systems with Applications*, 32(2), 313–317.
- Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Karimi-Majd, A.-M., & Fathian, M. (2017). Extracting new ideas from the behavior of social network users. *Decision Science Letters*, 6(3), 207–220.
- Kotler, P., & Armstrong, G. (2014). *Principles of Marketing* (15th ed.). Harlow, England: Pearson Education.
- Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. *Expert Systems with Applications*, 30(2), 313–324.
- Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of Self-Organizing Feature Map and K-Means Algorithm for Market Segmentation. *Computers & Operations Research*, 29(11), 1475–1493.
- Larose, D. T., & Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Lee, J. H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications*, 29(1), 145–152.
- Liao, S. H., Chen, Y. J., & Hsieh, H. H. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications*, 38(5), 6059–6069.
- Liao, S. H., Chen, Y. J., & Lin, Y. T. (2011). Mining customer knowledge to implement online shopping and home delivery for hypermarkets. *Expert Systems with Applications*, 38(4), 3982–3991.
- Linoff, G. S., & Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.). Indianapolis, IN: John Wiley & Sons.
- Lu, T.-C., & Wu, K.-Y. (2009). A transaction pattern analysis system based on neural network. *Expert Systems with Applications*, 36(3), 6091–6099.
- Miguéis, V. L., Camanho, A. S., & Falcão e Cunha, J. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, 39(10), 9359–9366.
- Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, 45(7), 1129–1157.
- Shim, B., Choi, K., & Suh, Y. (2012). CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert Systems with Applications*, 39(9), 7736–7742.
- SPSS Inc. (2001). *The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation* [White paper-technical report].
- Stremersch, S., & Tellis, G. J. (2002). Strategic Bundling of Products and Prices: A New Synthesis for Marketing. *Journal of Marketing*, 66(1), 55–72.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Boston, MA: Pearson Education.
- Tsai, C. Y., & Chiu, C. C. (2004). A purchase-based market segmentation methodology. *Expert Systems with Applications*, 27(2), 265–276.
- Turban, E., King, D., Lee, J. K., Liang, T.-P., & Turban, D. C. (2015). *Electronic Commerce: A Managerial and Social Perspective* (8th ed.). Cham, Switzerland: Springer International Publishing.
- Yang, T. C., & Lai, H. (2006). Comparison of product bundling strategies on different online shopping behaviors. *Electronic Commerce Research and Applications*, 5(4), 295–304.

