A DATA MINING FRAMEWORK FOR PRODUCT BUNDLE DESIGN AND PRICING

By

Yiming Li

Submitted in partial fulfilment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
November 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Product bundling is a marketing strategy that has been widely studied in research literature and extensively used in practice. With the growing quantity of products and huge possible bundling combinations, it is necessary to develop algorithmic approaches to determine which items should be in a profitable bundle, which bundling strategy is most profitable, and what the proper price is for a bundle. Previous studies have put forward many approaches for bundle design, but they have many limitations. Consumer' behaviors may be not in accordance with their statements in a survey, thus the transaction data is a more reliable source to predict their purchase behaviors. As consumers' demand and market supply will fluctuate continuously, fail to consider price elasticity of demand (PED) will cause biases for prediction, where PED is used to measure consumers' abilities and willingness to pay for certain products. In this thesis, we propose a data mining framework which incorporates the time value of money in data mining tasks, and it is capable of determining the product combination and price of a bundle in order to maximize the revenue. We also apply association mining to generate meaningful candidate bundles and reduce computation cost. This framework analyzes consumer and product data, taking demand and inflation factors into consideration, to fill in the gaps as mentioned. We also demonstrate the efficiency of this data mining framework through experiments and simulations.

ACKNOWLEDGEMENTS

I gratefully acknowledge my supervisors Dr. Qigang Gao and Dr. Hai Wang. Their expert advice and insightful supervision helped me improve my research skills and make this thesis a potentially publishable paper. Also their continuous encouragement and support for my master study make me confident while studying and working. The skills of making detailed plan that I learned from Dr. Gao really benefit me a lot in both my study and life.

I would also give my deep thanks to Dr. Evangelos E. Milios and Dr. Vlado Keselj for their time and reviewing the thesis.

Finally, I would like to express my overwhelmed gratitude to my parents and my boyfriend for their essential support and dedication.

CHAPTER 1 INTRODUCTION

1.1 Background

Fierce competition always makes business adopt various promotion strategies to attract more consumers and outperform other competitors. To meet consumers' needs and expectations is the basic principle to survive in this competitive business environment. A transaction is a two-sided process from which both buyers and sellers want to get the most benefit. Thus, it is crucial to keep balance between them. With consumers' desire for buying related products at the same time, sellers have to provide combinations of products in order to facilitate the purchasing process. Bundling is a promotion strategy in which sellers provide multiple products or events as a single package with an attractive price [49].

Bundling has become a prevalent promotion strategy rapidly since it is capable of raising values to buyers and generating profit to sellers, which perfectly matches the objective of a transaction process. From the consumer's perspective, they will be able to save 8% on average through purchasing a bundle package with a discounted price [19], which is a key driver of bundling. Also, consumers can save search cost, which will increase their willingness to purchase since they can easily find all wanted products and services in a bundle package provided by the seller. Some people also prefer bundles because they can reduce compatibility risk among components [37]. From the seller's perspective, adopting bundling can help increase the number of buyers and thus increase sales [37]. Moreover, a newly released product will be noticed and accepted by consumers if it is bundled with an existing product [37]. The seller's cost, like packaging cost and distribution cost, can also be saved by offering several products as a bundle [12].

Besides, Dana and Spier use printer and ink as an example to explain that bundling can also promote quality improvement [17]. A bundle of a durable item with some nondurable ones is the most common format of bundling, like bundling a printer and ink or selling a computer with some accessories. Without bundling, under which the reputation mechanism is inoperative, sellers may have chances to reduce the quality of durable and infrequently purchased goods. But for frequently purchased items, like ink, which has known quality, the firm needs to keep their quality at a high level in order to hold the advantages over their competitors. However, with the bundling of printer and ink, the seller will have a strong motivation to produce high-quality durable goods to prevent customer churn [17].

1.2 Problem Formulation

Three bundling strategies have been widely studied in previous research. *Pure component*, or unbundling, is the traditional way in which consumers can only purchase products or services separately with their original prices [47]. It allows buyers to see the sales process clearly and pick up exactly the products they want. On the contrary, in the *pure bundling* strategy, sellers provide several products together as a bundle, and buyers can purchase only the whole bundle rather than individual products [47]. Combining these two strategies, the *mixed bundling* strategy is a more flexible one that the seller offers both individual products and the whole bundle, and a buyer can make a choice between purchasing the entire bundle or one part of the bundle package [47].

The consumer's reservation price, defined as the highest price that a consumer is willing to pay for a product, is a key factor in bundling. It has a great impact on not only deciding bundle combinations but also determining the optimal price for a bundle. The relationship between reservation price and the actual price of a product determines whether or not a consumer will make a purchase. Krishna *et al.* [33] identify that a bundle is more likely to be profitable if the sum of the standard deviation of the reservation price for each bundle component is greater than the standard deviation of the bundle. The two attributes - correlation and additivity - show how the reservation price affects the bundling process. Super additive reservation price occurs when the items in a bundle are complementary, like PC and printer, in which the reservation price for the bundle is greater than the sum of that for each item. While sub additive reservation price usually occurs for substitutes when the benefit for each element overlaps to some extent [47].

Two main tasks associated with bundling in the previous research literature are bundle design and bundle pricing. Suppose that there are N distinct products available for bundling, the $2^N-(N+1)$ possible bundling combinations (excluding the bundle with a single item) make this problem extremely complex, especially when N is large [21]. It is impossible for sellers to provide all possible combinations to consumers. Bundle design is a process of selecting bundle combinations to be promoted, which should be rational, practical, and in accordance with consumers' preferences. Unlike PC and printer, or car and insurance, which are the obvious combinations that people often purchase together, some surprising but really effective combinations, like beer and diaper, can only be obtained through the analysis of consumers' purchase behaviors [48]. The main objective of providing bundles is to increase sales, thereby producing more profit for sellers. Moreover, a more remarkable principle that needs to be considered in bundle design is to know exactly what consumers want. Although bundling the best seller with a low-demand product may improve the sales of the less popular one to some degree, it may cause waste if consumers buy goods they don't need. In order to avoid waste, it is more beneficial for sellers to provide flexible bundles that consumers can select along with their preferences and needs.

Bundle pricing is about deciding the optimal price for a bundle package. The objective of the term "optimal" can vary based on their different business goals, such as maximization of profit, revenue, attendance, or market share [18]. Different bundling strategies (i.e., pure and mixed bundling) may result in different optimal bundle prices.

Consider a simple example of two software in order to explore how sellers and buyers can benefit from bundling. Suppose two software A and B are priced at \$70 and \$120 respectively. A bundle of A and B is also offered at \$170, denoted by P_A, P_B, and P_{AB}. Now assume there are three consumers and their reservation prices are given in the first column of Table 1.1. R_A represents a consumer's reservation price for the product A. Similarly, R_B is the reservation for the product B and R_{AB} is for the whole bundle. The amounts shown in the table are the surplus of each consumer, which is the difference between his reservation price and the actual price of a product. A rational consumer will make a purchase only if the price of a product does not exceed the reservation price for that product. In the example shown in Table 1.1, both consumers C_1 and C_2 will purchase only one software when A and B are sold separately, but they are willing to purchase the bundle package consisting of both A and B in pure bundling or mixed bundling strategies. Hence, bundling increases sales for the seller. Moreover, consumer C_3 will purchase A and B anyway as the actual prices are lower than his reservation prices, but will be able to get a discount for the whole bundle package. Hence, bundling may benefit the buyer as well.

	Pure Component		**Pure Bundling**	**Mixed Bundling**		
Consumer (Reservation price)	$P_A = \$70$	$P_B = \$120$	$P_{AB} = \$170$	$P_A = \$70$	$P_B = \$120$	$P_{AB} = \$170$
C_1 $(R_A = \$75, R_B = \$100,$ $R_{AB} = \$175)$	$\underline{\$5}$	<0	$\underline{\$5}$	\$5	<0	$\underline{\$5}$
C_2 $(R_A = \$60, R_B = \$130,$ $R_{AB} = \$180)$	<0	$\underline{\$10}$	$\underline{\$10}$	<0	\$10	$\underline{\$10}$
C_3 $(R_A = \$80, R_B = \$125,$ $R_{AB} = \$190)$	$\underline{\$10}$	$\underline{\$5}$	$\underline{\$20}$	\$10	\$5	$\underline{\$20}$

[a] An underlined number means a nonnegative difference between the actual price and the reservation price, and a purchasing will occur.

Table 1.1 Purchasing behaviors with different bundling strategies

1.3 Research Issues

Data produced during business activities mirror consumers' purchase patterns. Data-driven methods refer to analyzing data in order to solve one or more following problems associated with bundle design and bundle pricing:

- **What are the consumers' preferences among available products?**
  Purchase behavior analysis is one of the main issues of association mining. By mining consumers' baskets, frequent itemsets and rules can be generated to help sellers understand typical purchase patterns. The *Apriori* algorithm is the best-known method for association rule mining. It is realized by identifying items that occur in a single transaction concurrently with high frequency. Other approaches include Eclat algorithm, FP-growth algorithm, etc.
- **How can consumers be segmented?**
  Customer segmentation is about finding similarities within a large heterogeneous market. Dividing customers into several groups helps the understanding of special needs of a particular group and enables the customized promotions. The criterion for distinguishing each segment can be tackled by clustering or classification techniques based on people's demographic characteristics, along with their purchase behaviors.
- **What are the sales patterns for some seasonal products?**

Seasonal product refers to goods that have remarkable seasonal characteristics while producing and selling. The sales price of this kind of product may fluctuate according to their supply and demand within a sales cycle. The consumer's reservation price will also vary from season to season. Therefore, it is necessary to generate season groups (monthly or quarterly) using clustering techniques, and the analysis of a certain period can take historical sales information within the same cluster as a reference.

- **What are consumers' valuations for certain products?**

  As the key input in bundle design, consumers' valuations can be obtained from conducting a questionnaire which asks participants about their reservation prices for some products directly, or mining historical sales data. The former is quite straightforward, but it may be unreliable if people never purchase the listed products or bundles. The latter is a more credible approach since the relationship between consumers' valuations and the actual price can be represented by their purchase behaviors.

- **What is the optimal bundle configuration?**

  Bundle configuration includes deciding which products should be combined as a bundle, how to price them in order to achieve business goals, and the choice among bundle strategies. Bundle selection is based on product attributes like its characteristics, sales amount, inventory, and consumers' purchase patterns.

- **How can sellers benefit from a bundling promotion, and how can buyers benefit from a bundling promotion?**

  Many previous studies focus on determining optimal bundle pricing in order to benefit the seller. As shown in Table 1.1, buyers may also benefit from a bundling promotion. It is worth to investigate how a bundling promotion benefits the seller, the buyer, or both.

1.4 Thesis Contribution

In this thesis, we propose a framework for solving the bundle design problem as well as the bundle pricing problem. The main contributions of this paper are summarized as follows:

- Many previous proposed methods on bundle pricing either make strong assumptions on the reservation prices (e.g., the reservation prices are known), or estimate the reservation prices based on consumers' survey data. Our proposed framework uses consumer/buyer's previous purchase behaviors rather than a marketing survey as the data source for estimating buyers' reservation prices. In contrast to the consumers' survey data, which are usually of small size and subjective, and may be inconsistent and incomplete, historical purchasing transaction data are of large size, accurate and objective.
- Our proposed framework also incorporates the time value of money in data mining tasks and analyzes the price elasticity of demand in order to obtain an accurate estimation of buyers' reservation prices. The estimated buyers' reservation prices serve as the basis for bundle design and bundle pricing. As the result, better bundle design and pricing strategies can be achieved.
- Our proposed framework is generic and is not limited to specific data mining algorithms. For examples, new association rule mining algorithms can be integrated into the proposed framework to improve the efficiency and effectiveness for determining the possible product combinations within a bundle.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 presents related literature about the existing data mining methods for bundle design applied in various research fields. Chapter 3 introduces our framework for bundle design and pricing. Chapter 4 shows the performance of the proposed approach through experiments and simulations. Chapter 5 remarks the conclusion and future work.

CHAPTER 2 RELATED WORK

2.1 Overview of Data Mining & Methods

2.1.1 Data Mining for Knowledge Discovery

Data mining is a core step in the knowledge discovery process. It refers to discovering unknown, valid, actionable patterns from a large amount of data. The essential steps in knowledge discovery process can by summarized as follows [24]:

- **Data cleaning and integration** – dealing with outliers, noise, and missing values in data. Data from multiple sources may be combined for discovering more comprehensive knowledge.
- **Data selection and transformation** – selecting useful data that are relative to the task, then transforming the data into the form that is appropriate for mining. The order of these two steps can be reversed in some cases.
- **Data mining** – defining an appropriate data mining task and algorithm to solve the problem. Major data mining tasks include classification, clustering, association mining, and regression. The objective is to generate hidden but valuable patterns from the dataset.
- **Pattern evaluation** – measuring the interestingness and completeness of a data mining algorithm by evaluating the generated pattern.
- **Knowledge presentation** – presenting the knowledge to users using visualization and representation techniques.

2.1.2 Major Data Mining Methods and Algorithms

Major data mining methods and algorithms are organized into the following categories, which are also depicted in Figure 2.1.

A) Classification

Classification is a data mining technique in which a classifier is built to predict the categorical label of target attributes for new each data tuple [24]. It is a supervised learning method, which means the classifier learns regular patterns from training data containing a set of attributes and the associated labels. Then the classifier is evaluated based on its performance on a set of unseen test data. If the accuracy is acceptable, the learnt model will be then applied to class prediction for newly generated data. Commonly used algorithms for classification include:

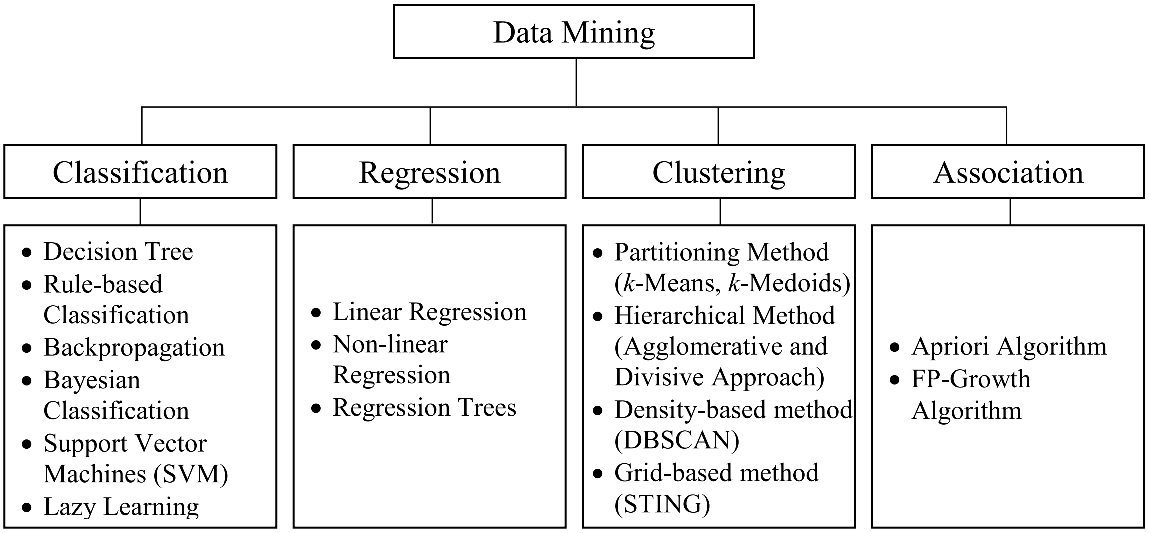

Data Mining
Classification
Regression
Clustering
Association
• Decision Tree
• Rule-based Classification
• Backpropagation
• Bayesian Classification
• Support Vector Machines (SVM)
• Lazy Learning
• Linear Regression
• Non-linear Regression
• Regression Trees
• Partitioning Method (*k*-Means, *k*-Medoids)
• Hierarchical Method (Agglomerative and Divisive Approach)
• Density-based method (DBSCAN)
• Grid-based method (STING)
• Apriori Algorithm
• FP-Growth Algorithm

Figure 2.1 Major data mining methods and algorithms

Decision Trees. A decision tree is a tree-like structure. Each internal node contains a test on an attribute, each branch of an internal node represents a result of the test, and each leaf node represents a class label [22]. ID3 and C4.5 are two basic algorithms for learning decision trees. Both of them are capable of processing a set of attributes and produce a class label of target for each of input data tuples. However, ID3 uses information gain as the measurement to select split attributes, while C4.5 uses gain ratio as the measurement to avoid splitting on an attribute with too many values. Besides, C4.5 also applies tree pruning to avoid overfitting.

Backpropagation. Deep learning is turned out to have good performance on analyzing high-dimensional data by using a deep graph consists of multiple processing layers [34]. A typical neural network contains an input layer, an output layer, and at least one hidden layer. A neural network becomes "deep" when it contains multiple hidden layers. Backpropagation is one of famous neural network learning algorithms. It proceeds the attributes iteratively and trains parameters in the network by comparing the predicted label with the actual one. It usually takes a long time for training, but it has a high tolerance for noise [24].

Bayesian Classification. Bayesian classifiers calculate the probability that a data tuple belongs to each class, and classify it to the class with the highest probability. Bayesian classification is based on Bayes' theorem [24], described as:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

where X is the set of attributes, H represents a hypothesis, and $P(H|X)$ is the probability that a tuple belongs to a certain class. Bayesian classifiers make the classification problem a simple statistical task. Naïve Bayesian algorithm, assuming all attributes are independent of each other, is a widely used algorithm that simplifies the computations involved in the classifying process. However, when the relationship among attributes cannot be ignored, the Naïve Bayesian algorithm becomes less efficient. A Bayesian Network consists of a directed acyclic graph and a conditional probability table [24]. It is capable of predicting the probability of each class that a given data tuple belongs to by considering the joint conditional probability distribution between variables.

Support Vector Machines (SVMs). SVMs search a linear hyperplane with the largest margin that can optimally separate data into two classes. Original training tuples are mapping into a higher dimension using an appropriate kernel function. Besides classification, SVMs can also be used for linear and non-linear regression [24].

B) Regression

Regression is another supervised learning method. The difference between regression and classification is that, instead of predicting categorical labels in classification, regression is about predicting a continuous value for a data tuple. Regression is preferred when the variables used in prediction have continuous values. It is good at discovering the relationship between independent variables and the dependent variable. The following are some typical algorithms used in the regression.

Linear Regression. Linear regression is the simplest form of regression. It is to fit a straight line for training data between independent variables $(x_1, x_2, \cdots, x_n)$ and the dependent variable y.

$$y = w_0 + \ w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

Non-linear Regression. Non-linear regression is realized by adding polynomial terms to the basic linear function to fit a non-linear function on training data [24].

Regression Tree. A regression tree is similar to a decision tree in classification, but each leaf in regression tree contains continuous-valued prediction rather than a class label.

C) Clustering

Unlike classification and regression, clustering is an unsupervised learning technique. It is a process that separates the given dataset into several clusters with higher similarity within a group and lower similarity between groups.

Partitioning Method. Partitioning methods simply partition all objects into clusters [22]. The most famous algorithms are k-Means and k-Medoids. Both of them select k objects initially as the centroid of each cluster, assign remaining objects to the one where they are most similar, and then calculate the new centroid of each cluster [24]. The process repeats until there is no change in cluster distribution, or the error is lower than the threshold. However, the centroid in k-Means is the mean of all objects in a cluster, while in k-Medoids is an actual object in data collection. That makes k-Medoids more tolerant to outliers.

Hierarchical Method. Hierarchical methods aim to cluster data objects using a tree-like structure [6]. An agglomerative approach is a bottom-up approach. It sets each object as a cluster initially, then merge similar clusters until the stop condition is satisfied. A divisive approach is a top-down method, which divides a large cluster into small ones.

Density-based Method. DBSCAN, as the most popular density-based clustering algorithm, searches data points connected by their neighborhoods within a given radius which contains a minimum number of objects [6]. The clustering process stops when there are no objects can be added to any clusters. The density-based method is capable of detecting clusters with arbitrary shapes, and it is not sensitive to noise and outliers. Other density-based methods include OPTICS and DENCLUE.

D) Association

Association analysis was first introduced in [2], which has been widely used in market basket analysis, medical diagnosis, bioinformatics, and web mining. It aims to find association rules from the dataset that satisfy the given minimum support and confidence. Support of item A and B is defined as the percentage of transactions that contain both A and B. Confidence of an association rule $A \Rightarrow B$ is the percentage of transactions containing both A and B over the total number of transactions containing A [24]. A typical association mining can be divided into two steps. The first one is to find a set of frequent itemsets where items often occur simultaneously, then is to generate strong association rules from these frequent itemsets [24]. [32] reviews basic concepts and major existing association rule mining techniques, among which the Apriori algorithm is the most popular one.

Apriori. Apriori was introduced by R. Agrawal and R. Srikant in 1994 [3]. It avoids generating too many unnecessary candidate itemsets by setting up constraints that subsets of a frequent itemset must be frequent [3]. Larger candidate itemsets are generated based on smaller frequent itemsets instead of considering transactions in the database. The efficiency of the Apriori algorithm can be improved by partitioning the dataset into nonoverlapping partitions or removing transactions that do not contain any frequent k-itemsets [24].

FP-Growth. The FP-growth algorithm can find frequent itemsets without generating candidates. It generates frequent 1-itemsets using the same method in Apriori, then sorts them in an ascending or descending order, and proceeds transactions in the same order to construct an FP-tree [24]. The FP-growth algorithm reduces the search cost significantly when mining both short and long frequent patterns.

2.1.3 Business Demand for Data Mining Solutions

Companies are gathering a huge amount of data in recent years. To make these data valuable, lots of companies has adopted data mining techniques to analyze data, discover hidden knowledge, and make correct decisions. Data mining solutions are well-fitted in the business environment, from data selection and preprocessing to applying data mining techniques and knowledge presentation [38].

Applying data mining in the retailing industry can help retailers know more information about their customers' preferences, purchase patterns, and trends, thereby provide high-quality customer services. Customer relationship management (CRM) aims to analyze customers' data to maintain existing customers and develop new customer strategies, where the classification and clustering techniques are widely used. Another example of data mining application used in retailing industry is market basket analysis. By applying association mining to customers' purchase history, retailers can have a good knowledge about customers' purchase patterns, thereby adopting appropriate promotions to attract more customers and increase sales.

Besides retailing industry, finance and insurance are also areas that involve lots of data mining solutions. Classification and clustering techniques are successfully applied in credit scoring when people apply for a loan from banks, and risk identification to analyze the risk level of an insured. Other applications include predicting customer profitability, fraud detection, and customer retention [38].

2.2 Overview of Bundling

Data mining methods have been applied to bundling. The related research work on various data mining methods is shown in Table 2.1. This section surveys the data mining related research for bundling in these research fields.

2.2.1 Retailing

A) Traditional Retailing Industry

Retailing is the largest industry for the research of bundling, which is represented in various patterns, like buy A and B together at a special price, or buy A get B 50% off. The main purpose for retailers to adopt bundling is to attract more consumers and maximize their own revenue or profit. For physical products, variable cost (e.g., material and labor) cannot be ignored in most cases so that revenue maximization and profit maximization are supposed to be treated differently.

Research Fields	Examples	Models
Traditional Retailing industry	• PC system (computer, printer and monitor)	• General bundle utility model^{1} [14] • Joint distribution using copula model^{1} [35] • Heuristic modeling^{1} [27] • Reserved bundling pricing^{1} [39]
E-commerce	• Amazon.com	• Association mining^{5} [29][36][40] • Dynamic bundling model^{5} [8][28] • Genetic model^{1} [9] • Graph-matching model^{1} [18]
Entertainment industry	• Seasonal tickets • Music and dance performance • TV shows	• Balance model^{3} [20]; • Hierarchical Bayesian Model^{3} [10] • Probabilistic model^{1} [5][46] • Bundle Selection Heuristic (BSH) modeling^{1} [50] • Industry model^{1} [16]
Information goods	• Books • Music albums	• Value-creating model^{4} [42] • Markov-based approach^{3} [11]
Travel products	• Hotel, flight ticket and car rental • City Pass	• Conjoint model^{3} [23] • Markov decision process and data envelop analysis^{3} [21]
Telecommunication	• Cellular phones and service plans	• Nonlinear mixed-integer programming^{1} [51] • Conjoint model^{3} [30]
Services	• Home service (cleaning and ironing)	• Mixed integer linear model^{1} [25] • Cluster-based model^{4} [41]

[a] Associated data mining methods: 1-Conventional statistical method 2-Classification 3-Regression 4-Clustering 5-Association

Table 2.1 Research fields of the existing data mining methods for bundle design and bundle pricing

Based on [20], Chung and Rao extend it into a general model, which can deal with products from different categories [14]. According to their comparable levels, attributes are divided into fully comparable, partial comparable and noncomparable attributes. They also take the heterogeneity among consumers into consideration by analyzing their preferences of bundles using a mixture distribution model. Estimated price is obtained from consumers' stated choice data about their preferred bundle combinations and willingness to pay through a survey. The Bundle Utility Model fills the gap in market segmentation and bundle for multi-category products, and it is proved outstanding among any other previous bundling models.

A copula model describes the dependence and probability distribution of multiple variables. Letham *et al.* [35] model a joint distribution using a copula model to derive consumer's valuation for products from mining sales transaction data. An inference procedure is introduced to predict the expected profit for unseen bundles with up to two items. Their data-driven approach offers both theoretical and experimental basis for bundle pricing problem in retailing industry [35].

Vertically differentiated products are defined as the products that vary on the quality attribute and are provided at different prices, whereas horizontally differentiated products are the ones with the same price and vary on other features like color or fat content [27]. Honhon and Pan show bundles of vertically differentiated products can significantly increase retailers' profit using their heuristic strategy with the setting of positive variable costs and non-uniform distribution of consumers' reservation price. They also demonstrate the factor that affects the choice of bundling strategies with abundant supply is the cost-quality ratios of components, while consumer's valuations affect the choice of bundle strategies when supply is limited [27].

Reserved product pricing (RPP) is a form of co-promotion that provide the second product with a discount to single product buyers, which has an advantage in gathering consumers' information of their decision making. It is proved optimal among other three bundling strategies when at least half of consumers do not anticipate the second stage discount [39].

B) E-commerce

As a rising industry, shopping online has an advantage that, not only the purchase history, consumers' all online behaviors are traceable. Therefore, dynamic bundling is much easier to realize in an online environment. Online recommendation system (ORS) is now prevalently used by electronic retailers since it can facilitate consumer's decision making process by recommending products they are likely to buy.

Association mining is a widely used and productive method for discovering interest patterns from business sales data. Karageorgos and Rapti use association mining to generate substitution and complementarity associations, realized with cosine similarity measure on product attributes and the *Apriori* algorithm for consumers' historical purchases respectively [29]. Bundles are retrieved from generated associations with a certain threshold. Rapti *et al.* extend it by introducing constraints and rules in bundling process, meaning the bundles can be produced from association dataset along with consumers' requests [40].

Liu and Zhang introduce association mining into ORS to make it more efficient and personalized [36]. They use Adaptive Resonance Theory (ART) model to generate consumer clusters based on their profile and habits, then analyze transaction data to find the most likely purchase combination of hot, general, and dull sales products for a given consumer. The recommend decision is made upon his/her purchase pattern and association mining result. However, the main focus of ORS is consumers' preferences, few ORS concentrates on the seller's profit. This problem is put forward by a real-time bundle pricing model in [8], [28]. Their models can calculate the optimal bundle price at each stage of consumers' decision making process when adding to or deleting items from the shopping cart.

The user-generated rating data crawled from Amazon.com was mined in [18] for solving the bundling configuration problem. The item's sales price multiplied by a coefficient calculated based upon their rating is used to estimate consumer's reservation price under the linear relationship assumption. Items and bundles are regarded as nodes and edges in a graph so that bundling configuration issue can be solved as graph matching. Using matching-based method and greedy algorithm, they develop the optimal solution for

2-sized bundling and heuristic solution for k-sized bundling under both pure bundling and mixed bundling strategy.

Besides profit maximization, minimization of dead stock is another objective of bundling. Birtolo *et al.* propose a genetic approach which takes both buyer requirements and product availability into account [9]. High fitness bundles of furniture (e.g. bed) and accessories (e.g. lamp) are generated using their attributes like color and material, and availability as constraints.

2.2.2 Entertainment Industry

Entertainment is another main field where bundling has been studied. The major difference between bundling in retailing and entertainment industry is that most bundles in former are complements while substitutions in the latter one since they usually have some overlaps, like seasonal sport or performance tickets.

Farquhar and Rao introduce an evaluation rule for the multiattributed items in a subset of TV shows [20]. They classify these attributes into five categories based on the level they contribute to the balance of subsets, and elicit two variables to represent the dispersion and centroids for each attribute. By applying these concepts, an evaluation model is developed which can measure the balance of a subset. This model is fully operational since more than a half of predictions are matched with subjects' judgments collected from a questionnaire. However, due to the assumption that all products in the bundle share the same set of attributes, this model can only be applied to homogeneous products. Bradlow and Rao combine Hierarchical Bayesian Model and this Balance Model, using subjects' preference selection data from sets of bundles among eight popular magazines to solve consumer priority and magazine assortments problems [10].

Probabilistic Approach is a crucial method of bundle pricing, which does a good job at determining optimal price of a bundle as well as each component. Venkatesh and Mahajan build a probabilistic model, considering consumers' time and reservation price as two central dimensions in consumers' decision making process, to compute the optimal price in each sale strategy and their corresponding profits for sellers [46]. Their model is applied to the entertainment industry including ten performances. Through their experiments, mixed bundling is proved to be the most profitable strategy, which is consistent with Schmalensee’s work in 1984.

Ansari *et al.* use the same dataset as [46], but they extend the situation to a non-profit case [5]. The difference between profit and non-profit organization is that the goal of the latter one is to maximize attendance rather than profit. They also use their probabilistic model to determine the optimal number of events being scheduled instead of fixing it to ten. For the case they studied, the result shows that non-profit organization is prone to adopt lower price and a greater number of events than a profit-maximizing firm based on the usage maximizing and non-deficit constraint.

Selling season tickets for several events exclusively first and allowing people to purchase tickets for a single event in a later date is a sale strategy that can help firms ensure maximum seats. For this scenario, Yakıcı, Özener and Duran raise a Bundle Selection Heuristic (BSH) method for selecting the best bundle based on two measures – demand and time factor [50]. That is, the potential revenue of a bundle will increase if the bundle consists of a high demand event and a low demand event with a longer time gap.

Crawford and Yurukoglu mine television rating data, price, and market shares to predict the household preference on television channels and bundle purchases [16]. They estimate consumers’ willingness to pay for a channel using their watching frequency. Aggregated channel and cable system data are used in their industry model to predict household preference among possible combinations of demographic groups. Under a à la carte (according to the menu) situation, content and distribution of television channels are bargained among distributors, causing their input costs change. The industry model in [16] catches this variation to estimate social welfare in a more accurate way.

2.2.3 Digital Information Goods

The unique cost structure that being reproduced and distributed easily with lower cost makes digital goods different with consumable products [1]. The profit maximization and revenue maximization can be treated equally on the occasion of ignorable variable cost increment with sales amount growing, which also makes digital goods more ideal for bundling. Large bundles of information goods can significantly benefit sellers as well as increase their competitiveness [7]. Hiller analyzes the bundles of DVD services and streaming offered by Netflix, which is a company in America that provides on-demand Internet streaming media, and finds that commercial success, distribution network, and the age of release of films are the characteristics of information goods that make the mixed bundling strategy profitable [26].

Similar to [28] [36] that combine bundling algorithm with recommendation system, Somefun and Poutré apply this integration to information goods [42]. A recommendation system is used to trace consumers' profiles and generate their preferences without referring to privacy. The learning method combines genetic algorithm (GA) for bundle definition adjustment and Amoeba algorithm for pricing. A very small price deduction positively correlated with bundle size is applied to the bundle price. But unlike the profit-generating approach, a value-creating method is introduced in which the three key value drivers – transaction efficiency, complementarities, and lock-in – are targeted for generating value.

An *iPrice* collaborative pricing system consisted of collaborative prototyping module, optimal price estimation module, and version revisionary module is developed in [11]. A Markov-based approach is used in the first module to predict consumers' needs and transition among three needs categories. User profile and price history are taken into consideration for estimating optimal price for a bundle. Their system performs well in mining consumers' actual needs using ERG theory [4] as a basis.

2.2.4 Travel Products

Bundling of travel products let travelers plan their vacation in a convenient way since such a bundle may help them schedule all tickets, hotels, and the attractions they want to visit. An understandable example of travel products bundling is Toronto City Pass. Comparing to buying tickets individually, consumers can save about 45 percent with purchasing the city pass which contains tickets of the five best attractions in Toronto [15]. The city also can attract more tourists at the same time.

Conjoint analysis is a method that measures the effect of two or more independent variables on the dependent variable. Understanding consumers' preferences and decision making process can help sellers make proper strategies. Goldberg, Green, and Wind modify the conjoint model to a hybrid categorical conjoint analysis model which can deal with correlated attributes and used it to analyze consumers' preferences for hotel amenities from six facets [23].

Ferreira and Wu point out that although the conjoint model provides a method for bundling, it utilizes a non-dynamic approach [21]. To fill in this gap, they adopt a time dependent function to develop a dynamic bundle-pricing model for travel packages containing flight tickets, hotel reservation and car rental promoted by an online agency. In their research, bundle selection is modeled by Data Envelope Analysis and pricing problem is done by Markov decision process.

2.2.5 Telecommunication

Besides cost, the main factor that drives consumers to buy telecommunication bundles is usage convenience, which may be caused by the complexity of provided services with technical features [37]. A bundle of a cellular phone and service plans has been widely provided by telecommunication service providers to attract more consumers as well as promote wireless telecommunication services. The cellular phone is sold with a discount but users must subscribe to a service plan with a minimum price. The optimal reduction in the price of the cellular phone is determined by [51] using nonlinear mixed-integer programming to maximize the total profit for providers.

Service plans can vary with different amounts of calling minutes, data, and text messaging. Collecting 116 respondents from an online survey about mobile plans and monthly fees, Klein and Jakopin use a conjoint model to analyze consumers' willingness to pay and the criterion when they state their preferences. The result shows the integration of calling minutes and data gains the utility of a bundle most while free messaging gains least [30].

2.2.6 Services

Service industry develops rapidly in recent years. Kohlborn *et al.* state that services are non-standard and perishability, and the process of production and consumption cannot be separated [31]. It is better than physical goods on creating values in consequence of these special characteristics.

Razo-Zapata *et al.* propose a value-oriented framework to help educational parties provide suitable educational service bundles to candidates so that they can fill in some skill gaps while looking for jobs. Services are clustered first in the light of functional consequences, and solution clusters without overlapped functions are grouped to generate potential service bundles [41].

Mixed integer linear program is an approach which performs well in solving the optimal bundle pricing problem with the exponential growth in possible bundle combinations due to the increasing number of products being sold. Hanson and Martin formulate an optimization model by using mixed integer linear program which requires consumers' reservations, customer segments size, and sellers' unit variable cost as input, to determine the optimal solution based on a questionnaire for home services such as laundry and ironing [25]. Their approach can also check the discrimination among different customer segments' reservation prices for the same bundle.

2.2.7 Summary

Many studies have target estimating the consumer's reservation price and bundling problem, some of them also involve customer segmentation and product clustering according to attributes, popularity, and value.

Although many techniques have been previously proposed as described before, there is no study consider time as a dimension in their analysis currently. None of them analyze consumers' demands and the fluctuated value of money for consumers' reservation estimation. The consumer's reservation price is a key factor in bundling problem. It is difficult to be predicted because consumers' willingness and abilities to pay for a single product are different. Moreover, their reservations will also fluctuate in different time periods.

CHAPTER 3 METHODOLOGY AND FRAMEWORK DESIGN

We propose a data mining framework for bundle design and pricing which is illustrated in Figure 3.1. One of the important features of the proposed framework is to incorporate the time value of money for estimating consumers' reservation prices. We consider the actual value of money in different years rather than using the historical money directly. Our framework can fill in the following gaps:

- Method for estimating consumer's reservation price. Sometimes buyers' behaviors may not be in accordance with their statements in a questionnaire. Using their actual purchase records can avoid these differences. Therefore, we use transaction data as the source to estimate the highest prices that consumers' want to pay for the certain products, which is more reliable than consumer feedback based approaches.
- Variation in market demand and the consumer's reservation price in different time periods. We adopt data mining techniques for price elasticity of demand (PED) analysis to discover fluctuations in consumers' demands, which aims to detect time periods in which consumers have different willingness and abilities to pay for a certain product.
- Loss of real value of currency. Real value of historical currency fluctuates due to inflation, leading to bias when estimating the consumer's valuation. We use inflation rate to map historical currency to present value, which can reflect a consumer's actual purchasing power.

Basic market profile and notations are listed below.

N: The number of items for sale I $= \{i_1, i_2, \ldots, i_N\}$

M: The number of consumers $\mathrm{C} = \{c_1, c_2, \ldots, c_M\}$

T: The set of transaction data generated by consumers. Records that belong to a consumer *c* with a product *i* can be represented as $\{T_{c,i}\}_{c\in C, i\in I}$

S: The number of years covered by the transaction dataset Y $= \{y_1, y_2, \ldots, y_S\}$

p: The unit price of a product

v: The sales volume for a product

RI: An M×N matrix containing price intervals with each one represents the range of a consumer's reservation price for a product.

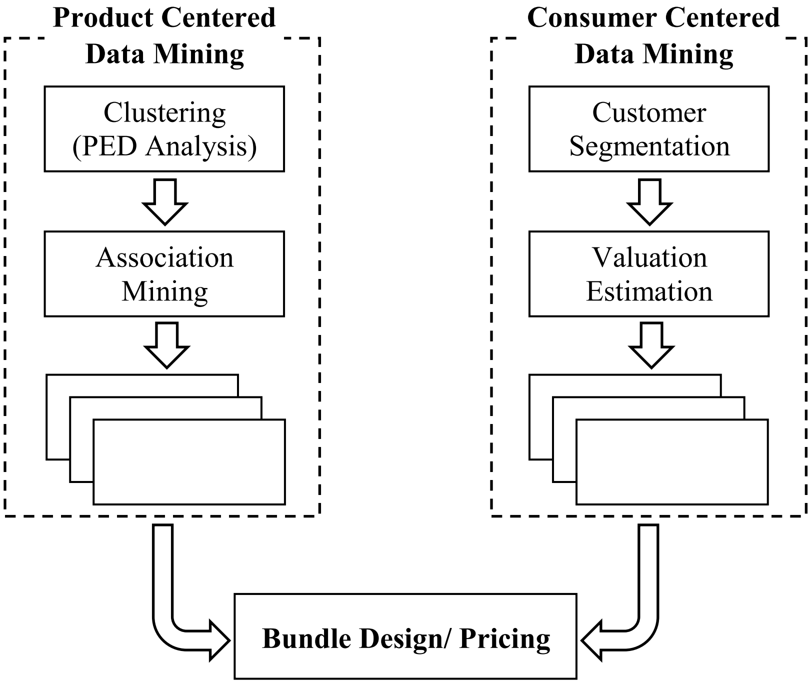

Product Centered Data Mining
Clustering (PED Analysis)
Association Mining
Consumer Centered Data Mining
Customer Segmentation
Valuation Estimation
Bundle Design/ Pricing

Figure 3.1 The data mining framework for product bundle design and pricing

3.1 Features of the Framework

As depicted in Figure 3.1, this framework consists of three main components. Product centered data mining analyzes the price elasticity of demand and consumers' baskets to generate frequent itemsets. Consumer centered data mining obtains customer segments and accurate estimation of buyers' reservation prices based on PED analysis. The results of association rule mining and valuation estimation serve as the basis for bundle design/pricing to determine the product combination and the price of a bundle.

3.1.1. Price Elasticity of Demand (PED) Analysis

Price elasticity of demand (PED) is used to measure the change of quantity demanded of a good or service in its price, with other things being equal [43]. For elastic products, an increase in unit price will lead to fewer units sold, resulting in a downward-sloping curve in its graphic representation with quantity on the horizontal axis and price on the vertical axis.

A demand curve expresses the relationship between the price of a given product and the consumers' willingness and abilities to pay for this product with that price in a period of time. That is, with consumers' reservation prices and other determinants remaining the same, changes of unit price lead to movements along the same demand curve. However, a change in consumers' reservations will cause a positive or negative shift in demand curves. Based on these economic concepts, we adopt Principal Component Analysis (PCA) and k-Means algorithm to analyze the fluctuations of consumers' reservation prices by discovering the direction of demand curves and generating month clusters.

Given a set of transaction data, sales volume and price for a product in a month can be extracted easily. The average price is treated as the sales price if the unit price changes within a month. As a result, we can get a list for each product which contains the year, month, sales volume, and unit price. Next step is to calculate the average sales volume and price in the same month within S years (see Equation (1)), assuming v_{y_j,m_k} and p_{y_j,m_k} are sales volume and unit price of a product in the month m_k in year y_j. The objective to use mean instead of individual ones is to avoid bias due to some random factors including weather, holidays, or unexpected events. For example, if the weather in a year gets warm much earlier than other years, the sales of short sleeve shirts will start increasing and reach the peak in advance.

$$\overline{v_{m_k}} = \frac{1}{S} * \sum_{j=1}^{S} v_{y_j,m_k} \quad (1)$$

$$\overline{p_{m_k}} = \frac{1}{S} * \sum_{j=1}^{S} p_{y_j,m_k}$$

The $(\overline{v_{m_k}}, \overline{p_{m_k}})$ pairs for all 12 months may be distributed in more than one parallel demand curves in its graphic representation if all other determinants stay equal. Each data point represents the relationship between the average unit price $\overline{p_{m_k}}$ and sales $\overline{v_{m_k}}$ in a month. The following step is to find the months on the same or very close curves. PCA is a common-used method for dimensionality reduction, which is achieved by detecting the directions of the first several largest variances in data and transforming original data into the data expressed in terms of new axes. We adopt PCA to find the principle component in downward-sloping direction, which represents the trend of demand curves for elastic goods, then build a new axis x' in this direction and another axis y' as orthogonal to the first one. By mapping data points to the y' axis, points on the same curve are closer while points on different curves are far away from others.

Then k-Means is applied to discover month clusters using the transformed data points. Each one of them represents a month. The procedure of k-Means is shown in Figure 3.2. k-Means aims to find data clusters with large intracluster similarities and small intercluster similarities. The similarity is measured by the distance between a data point and the centroid of the cluster it belongs to. The value of K is required as an input, which varies for different products and depends on a heuristic learning method using Within Cluster Sum of Squared Error (WCSSE) as the measurement, defined as

$$\mathrm{E} = \sum_{i=1}^{K} \sum_{p \in G_i} |p - m_i|^2 ,$$

where p is an object in data collection, m_i is the mean value of all objects in a cluster G_i [24]. With K increasing, the first one that makes WCSSE smaller than a threshold will be set as the number of month clusters $\mathrm{G} = \{G_1, G_2, \dots, G_K\}$. Each cluster contains an uncertain number of months and the cluster which includes the month m is denoted as G_m.

The process and result of PCA and k-Means can be illustrated using Figure 3.3. Black points are original data representing the relationship between sales volume and unit price in each month. Colored points are the mapping result by PCA. Points in an oval are the ones being grouped in a cluster using k-Means algorithm.

```
Input:
    ▪ D: a data collection containing n observations,
    ▪ K: the number of clusters.

Output: A set of K clusters $\mathrm{G} = \{G_1, G_2, \dots, G_K\}$.

Algorithm:
BEGIN
    randomly choose K objects as initial cluster centres $m = \{m_1, m_2, \dots, m_K\}$.
    repeat
        foreach object $p \in D$
            set $p$ to the cluster $G_i \leftarrow arg \min_i |p - m_i|^2$
        foreach cluster $G_i \in G$
            $m_i$ ← the mean value of all objects belong to $G_i$
    until no changes;
END
```

Figure 3.2 The k-Means clustering algorithm

3.1.2. Customer Segmentation

Customer relationship management (CRM) has been widely used among sellers to develop new customers, enhance the relationship between existing customers and retain profitable customers [44]. Customer segmentation is one of the essential tasks in CRM which divides customers into several segments based on their demographic and geographic information, purchase behaviors, or survey statements.

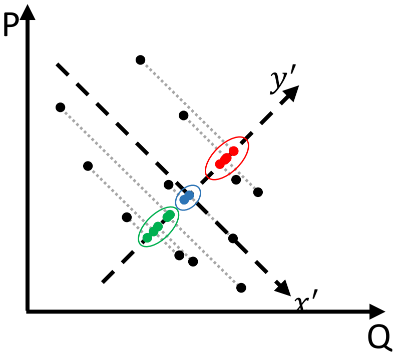


Figure 3.3 Process of PCA and *k*-Means

Clustering techniques have been applied to solve customer segmentation problem due to its efficiency and ability to process large datasets. In our research, we adopt *k*-Means algorithm to discover customer segments since it is efficient in modeling and capable of producing understandable results. Consumers' information including gender, age and income provided while registration, along with transaction records, are transformed into features in the clustering process. Similar to PED analysis, a WCSSE threshold is set to determine the optimal number of customer segments.

3.1.3. Valuation Estimation

A consumer's reservation price for a product may be various in different periods depending on trackable factors like season and demand, and some unpredictable factors as well. Sales price is determined by market supply and demand, which will be affected by the cost of material, technology, and inflation. These two variables are uncertain, but the relationship between them can be represented by consumers' purchase records. It is assumed consumers are rational. In other words, a consumer's reservation price for an item is equal to or greater than the unit price if he made a purchase. Therefore, we use historical transaction data to estimate their valuations.

Due to inflation, the price levels of goods and services reveal a sustained increase over a period of time. It may lead to a loss of real value if we use unit price five years ago directly. Therefore, we map historical currency to present value to eliminate the effect of inflation. Assuming the average inflation rate is *r*, *n* is the number of year gap between the original year and the target, the present value PV of a historical price can be calculated using Equation (2).

$$\mathrm{PV} = \mathrm{p} \times (1 + r)^{n} \tag{2}$$

If we are going to estimate consumers' reservation prices and generate profitable bundles in the month m, only the months which belong to the same cluster G_m will be considered in following steps. For a consumer $c \in C$ and an item $i \in I$, we extract his purchase records $T_{c,i}$ from transaction set, pick up the records which happened in the month in G_m along with their timestamp and price mapped to present value. We assume their valuations of a given product equals to its price when they made the first purchase. The relationship between a consumer's reservation price and the number of purchases *np* forms the following function $R = (1 + \theta)^{np} \times PV$. Each successful transaction makes their valuation increased by θ ($\theta > 0$). For example, if the unit price mapped to present value for an item is $PV = \$2$ and $\theta = 0.1$, a consumer's reservation price when he made the first purchase was *\$2*, which increased to *\$2.2* at the second purchase and *\$2.42* at the third time. But for the month with no purchase, we assume their valuations were less than the actual price and dropped exponentially by θ. We order all records according to the year and month sequence and assign each year a weight. For the year y_j, the weight is $w_{y_j} = \beta^{j-1}$. If $\beta > 1$, earlier months are assigned smaller weights and later months have larger ones, representing the latest purchases have more impact on their future behaviors. Whereas the former purchases influent their future decisions more if $\beta < 1$. All months have the same weight in the estimation process when $\beta = 1$. Table 3.1 shows the purchase records for a consumer $c \in C$ with an item $i \in I$. A consumer's approximate reservation is estimated using Equation (3).

$$R_{c,i} = \frac{\sum_{j=1}^{S} \sum_{m_k \in G_m} w_{y_j,m_k} \times R_{y_j,m_k}}{\sum_{j=1}^{S} \sum_{m_k \in G_m} w_{y_j,m_k}} \tag{3}$$

Year	**Month**	**Purchase or not**	**Price (Present Value)**	**Reservation Price**	**Weight**
y_1	m_1	Y	PV_{i,y_1,m_1}	$R_{y_1,m_1} = PV_{i,y_1,m_1}$	$w_{y_1} = \beta^0$
y_1	m_2	Y	PV_{i,y_1,m_2}	$R_{y_1,m_2} = (1+\theta) \times PV_{i,y_1,m_2}$	$w_{y_1} = \beta^0$
y_1	m_3	N	PV_{i,y_1,m_3}	$R_{y_1,m_3} = (1-\theta) \times PV_{i,y_1,m_3}$	$w_{y_1} = \beta^0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_j	m_k	Y	PV_{i,y_j,m_k}	$R_{y_j,m_k} = (1+\theta)^{np} \times PV_{i,y_j,m_k}$	$w_{y_j} = \beta^{j-1}$
y_j	m_{k+1}	N	$PV_{i,y_j,m_{k+1}}$	$R_{y_j,m_{k+1}} = (1-\theta)^{nnp} \times PV_{i,y_j,m_{k+1}}$	$w_{y_j} = \beta^{j-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3.1 Purchase record for consumer *c* with product *i*

Considering that the reservation price is an extremely subjective factor, and some unpredictable factors may cause bias during estimation, we use an interval to represent a consumer's reservation price instead of a single value. Assuming the sales price for the item i is p_i, we create several intervals with each one covers $0.05 \times p_i$. Examples of intervals are $[0.9 \times p_i,\ 0.95 \times p_i)$, $[0.95 \times p_i,\ p_i)$, and $[p_i,\ 1.05 \times p_i)$. The interval of estimated value of Equation (3) is treated as the consumer's reservation price interval. The results for all consumers and items form an M×N valuation matrix *RI*, in which the interval $RI_{c,i}$ represents the reservation price range of consumer *c* for item *i*. We set the range to $[0,\ 0.05 \times p_i)$ for a consumer with the products he has never purchased.

However, since the valuation matrix only contains the reservation price for individual items, we still need to predict their willingness to pay for a bundle *b* which consists of multiple products. A recognized function deriving a consumer's valuation for a bundle $R_{c,b}$ from its components $R_{c,i}$ proposed by Venkatesh and Kamakura is shown in Equation (4) [45].

$$R_{c,b} = (1 + \lambda) \times \sum_{i \in b} R_{c,i} \tag{4}$$

The $R_{c,i}$ here is the median of the interval that a consumer's reservation price belongs to. The coefficient λ indicating the bundle's type among complementary, substitutes, and independent. If the bundle is complementary, i.e., PC and printer, a consumer's willingness to pay for this bundle is higher than the sum of each composition, then $\lambda > 0$. However, for substitutes like seasonal sports tickets, $\lambda < 0$ indicates buyers do not want to pay as much as the total price when purchasing separately. And λ is supposed to equal to 0 when there is no relationship among the components in a bundle.

3.1.4. Bundle Design

A) Association Mining

Since the number of products available in a market is large, which creates numerous possible combinations, considering all potential bundles will cost too much computation. Some combinations may be profitable to sellers but meaningless to buyers. Through basket analysis, we can find that the relationship between some merchandises really exists since they always appeared in a single transaction simultaneously, but they are independent seemingly. However, for the items that consumers never or seldom purchased together, this kind of bundles is pointless.

Therefore, we only consider the itemsets that are often being purchased together obtained through association rule mining. By setting the minimum support *min_sup* and confidence, association rule mining detects all frequent itemsets which reach the support threshold and generates strong association rules from the frequent itemsets. The *Apriori* algorithm is the most well-known approach for association mining, which is applied in our framework to find frequent itemsets. It first finds all individual items that satisfy *min_sup* to constitute the frequent 1-itemsets L_1, then generates frequent 2-itemsets L_2 by calculating $L_1 \bowtie L_1$ and removing itemsets with support lower than *min_sup.* In the following process, frequent k-itemsets are produced by calculating $L_{k-1} \bowtie L_{k-1}$ and removing itemsets with infrequent subsets and support lower than *min_sup*. This procedure is repeated until L_k is empty. Figure 3.4 shows the pseudo-code for the *Apriori* algorithm [24].

Input:

- D: a collection of transaction data,
- *min_sup*: the minimum number of support count.

Output: A set of frequent itemsets *L*

Algorithm:

```
BEGIN
    $L_1$← individual items with count > min_sup
    $k \leftarrow 2$
    while ($L_k \neq \emptyset$) do
        foreach $l_1, l_2 \in L_{k-1}$
            if ($l_1[1] = l_2[1] \wedge l_1[2] = l_2[2] \wedge \ldots \wedge l_1[k-2] = l_2[k-2] \wedge l_1[k-1] > l_2[k-1]$) then
                $c \leftarrow l_1 \bowtie l_2$
                if $s$ is a subset of $c$, $\forall s \in L_{k-1}$, then
                    add $c$ to $C_k$
        $L_k$ ← itemsets in $C_k$ with count > min_sup
        $k \leftarrow k + 1$
        add $L_k$ to $L$
END
```

Figure 3.4 Pseudo-code for the Apriori algorithm

B) Bundle Design and Pricing

Bundling configuration including determination of bundle combinations, price, and strategies is done based on the potential bundle set *B* (the frequent itemsets in the *Apriori* algorithm) and consumers' valuation matrix *RI*. Unlike previous studies, which set the bundling strategy and its constraints as prerequisites, we calculate the revenue in each of pure component, pure bundling and mixed bundling, and choose the one with the highest revenue gain instead of restricting a bundle to a specific strategy ahead. Price for a bundle under each promotion is set as the one that can maximize the seller's revenue.

We make several assumptions which were used in previous studies [18].

- **Single Unit.** Each consumer purchases up to one unit for each item or bundle.
- **Single price.** Each item or bundle has exact one sales price.
- **No budget constraint.** Consumers do not have budget constraint while shopping.
- **No supply constraint.** The market can provide as much as consumers need. The occasion of "Out of Stock" will not be considered in this paper.

In practice, the consumer's rationality will make them purchase the product with a price not exceeds their valuations. We use the variable $h_{c,i}$ to denote the purchase behavior of the consumer c with the item i. $h_{c,i} = 1$ when c takes i, and $h_{c,i} = 0$ if the purchase does not happen. $h_{c,b}$ achieves the similar purpose but shows the relationship between the consumer c and the bundle b instead of an individual item. Following the probabilistic variable using in [18], $P(h_{c,i}\,|p_i, R_{c,i})$ represents the probability of the occurrence of c purchases i ($h_{c,i} = 1$) with the price p_i and his reservation price $R_{c,i}$. But we extend it to P_{pc}, P_{pb}, and P_{mb} in different promotion strategies.

For each possible combination in B, we calculate the maximum revenue it can create in each bundling strategy.

Pure Component. This is an unbundling strategy which is adopted in conventional market. Price for each commodity p_i is provided by sellers. The corresponding revenue r_{pc} is obtained by Equation (5).

$$r_{pc} = \sum_{i\in b} \sum_{c\in C} p_i \times P_{pc}(h_{c,i}|p_i, R_{c,i}) \tag{5}$$

where

$$P_{pc}(h_{c,i}|p_i, R_{c,i}) = \begin{cases} 1, & if\ p_i \le R_{c,i} \\ 0, & otherwise \end{cases}$$

Pure Bundling. Comparing with the pure component, this is a similar situation with bundles replacing individual items. The most significant difference is that the price for a bundle p_b is a variable which needs to be determined. Given all consumers' reservation prices for a bundle (see section 3.1.3), we set cut-points p_b to calculate the number of consumers who will make purchases and the corresponding revenue using Equation (6). The one which makes r_{pb} maximized is chosen as the sales price for the bundle b.

$$r_{pb} = \sum_{c\in C} p_b \times P_{pb}(h_{c,b}|p_b, R_{c,b}) \tag{6}$$

where

$$P_{pb}(h_{c,b}|p_b, R_{c,b}) = \begin{cases} 1, & if\ p_b \le R_{c,b} \\ 0, & otherwise \end{cases}$$

Mixed Bundling. This is a more complicated situation since both individual items and bundles are offered. Prediction of a consumer's choice among a bundle and its components is essential to estimating revenue. Taking the scenario containing two products X and Y as an example. A consumer's valuation $R_X = \$10$ and $R_Y = \$5$. We set λ in Equation (4) to -0.1 so that his reservation price for the bundle of X and Y is $R_{XY} = \$13.5$. If both of them are sold as $p_X = p_Y = \$7$ and $p_{XY} = \$13$, we predict that he tends to choose X rather than the bundle since the actual prices imply $p_{XY} - p_X = \$6$, which is beyond his valuation of Y. Therefore, we set selection conditions shown below.

$$r_{mb} = \sum_{c \in C}\left[p_b \times P_{mb}\left(h_{c,b}|p_b, R_{c,b}\right) + \sum_{i \in b} p_i \times P'_{mb}\left(h_{c,i}|p_i, R_{c,i}, P_{mb}\right)\right] \tag{7}$$

where

$$P_{mb}\left(h_{c,b}|p_b, R_{c,b}\right) = \begin{cases} 1, & if\ p_b \leq R_{c,b}\ and\ for\ \forall s: p_b - p_s \leq R_{c,(b-s)}, \\ & s\ is\ a\ subset\ of\ b \\ 0, & otherwise \end{cases}$$

and

$$P'_{mb}\left(h_{c,i}|p_i, R_{c,i}, P_{mb}\right) = \begin{cases} 1, & if\ p_i \leq R_{c,i}\ and\ P_{mb}\left(h_{c,b}|p_b, R_{c,b}\right) = 0 \\ 0, & otherwise \end{cases}$$

With all calculations finished, next step is the simple comparison of the results of (5) – (7) and choose the strategy with the highest one for promotion.

C) Bundle Selection

Bundle selection is necessary for eliminating redundant bundles and ensuring maximum revenue to sellers. We adopt this step for the following objectives:

- Avoid conflict. Promotion strategy for each bundle is selected according to their potential gain in revenue. If a combination *A* is assigned to pure bundling but one of its subsets is assigned to the mixed bundling, conflict will exist since components of *A* are also provided individually.
- Revenue maximization. With the prerequisite $\cup \mathrm{B} = \mathrm{I}$, various configurations can be issued, but we aim to find the one with the highest revenue gain.

We use a greedy approach for bundle selection to find the eligible bundle configuration. We select bundles from all frequent itemsets based on their absolute revenue gain. The itemset which provides the highest absolute gain will be chosen for promotion, then removed from the pickup pool along with the bundles which have items overlapped with it. Having the new set of candidate bundles, we still choose the one with the highest absolute gain and repeat the process above until there is no bundle left. This method has no effect on the bundling strategy so that all selected bundles are enrolled in the one where they are optimized. It can prevent confliction among bundling strategies since all bundles are non-overlapped.

3.2 Framework Architecture

Our data-driven framework applies clustering and association mining techniques for analyzing consumer data and product data. The four major components of proposed data mining framework are shown in Figure 3.5. *Customer Segmentation* aims to discover similarities among customers and improve accuracy when estimating the reservation prices for new customers. *PED analysis* takes item data including historical price and sales volume as input, then generates month clusters which will be considered in *Valuation Estimation*. For a target month which we are going to analyze, the consumer's reservation price will be predicted based on only the months in the same cluster with it. *Association Mining* is used to generate frequent itemsets as candidate bundles and reduce computation cost. The frequent itemsets, along with reservation price matrix are processed in *Bundle Design and Pricing* to determine the optimal bundling strategy and proper price for each combination. Final promoted bundles are picked up from candidates by *Bundle Selection* using revenue maximizing and non-overlapping criterions. These four processes form the central part of this data mining framework.

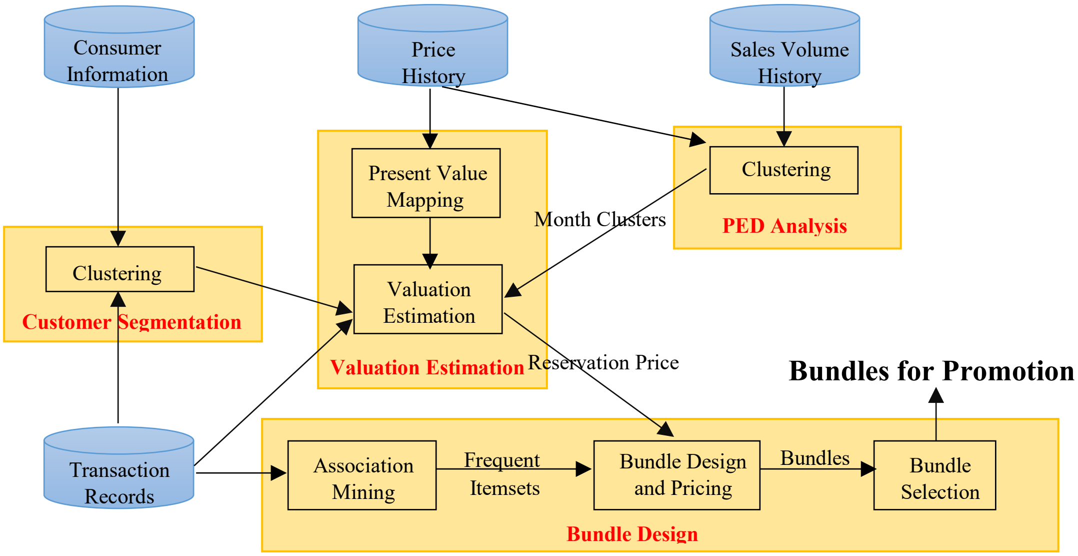


Figure 3.5 Framework architecture

CHAPTER 4 EXPERIMENT AND EVALUATION

4.1 Simulation Transaction Data

Based on our proposed framework, a consumer's reservation price is estimated based on the consumer's historical purchase behaviors. However, there are no publicly available transaction data sets covering multiple years. We used simulation data set to demonstrate the efficiency of our framework.

4.1.1 Candidate Transactions

Given the number of consumers M and products N, we first generate the consumer set C and product set I, and randomly pick up a base price p_{base} for each product. Then we generate 12 monthly candidate transaction datasets in a year with each one consists of the Cartesian product of C and I, along with a price for each combination. Considering some dynamic factors like seasonality and holidays, the price for a product in a certain month is produced by multiplying its base price and a seasonal coefficient, which is randomly generated in the range of -α to α. That is, the sale price for a product $p_i \in [(1-\alpha) \times p_{base}, (1+\alpha) \times p_{base}]$. Since the seasonal coefficient is randomly picked up for each product in each month, different seasonal patterns can be found in the candidate transaction dataset for different products. Candidate transactions for the following years are obtained based on the one generated in the last step by taking the inflation rate into consideration.

4.1.2 Reservation Prices

We also generate a consumer's reservation price matrix with size $M \times N$. Each row represents a consumer and each column represents a product. For a product i, consumers' reservation prices are given by a normal distribution with a mean of p_{base} and a standard deviation of $\sigma \times p_{base}$, or a uniform distribution between $(1-3\sigma) \times p_{base}$ and $(1+3\sigma) \times p_{base}$. The reason for choosing $1 \pm 3\sigma$ as boundaries of the uniform distribution is that we want to generate consumers' reservation prices with same range using different distributions. The reservation price matrices are used to filter candidate transactions and evaluate our algorithm as a benchmark.

We set the number of consumers and products as 100 in the simulation. Therefore, candidate transaction dataset has 10,000 records for each month and 120,000 records for each year. To achieve PED analysis, we generate transactions covering ten years so that the sales can reveal a relatively stable pattern. The seasonal coefficient is set to 0.2, representing the unit price for a single item can fluctuate within the range of 20 percent in different months. The standard deviation of normally distributed reservation price is set to $0.1 \times p_{base}$. This setting can ensure most consumers have chances to make a purchase because 97.5% of consumers have reservation prices greater than the possible lowest unit price. Accordingly, uniformly distributed reservation price follows $U(0.7 \times p_{base}, 1.3 \times p_{base})$. The parameter settings in the simulation are summarized in Table 4.1.

Parameters	Meaning	Value
M	The number of consumers	100
N	The number of products	100
S	Transaction length (years)	10
α	Seasonal coefficient	0.2
σ	Standard deviation of normal distribution	0.1

Table 4.1 Parameter settings

4.1.3 Transaction Filtering

According to the consumer rationality assumption, consumers will only purchase the products with price not exceeding their reservation prices. That makes some transactions in our candidate datasets unreasonable. Therefore, we remove the transactions in which the sales price is greater than the corresponding consumer's reservation price. The remaining transactions, along with a transaction ID for each record, constitute our simulated transaction set. Table 4.2 shows the number of transactions in each year filtered by normally and uniformly distributed reservation price matrix respectively.

4.2 Training and Evaluation

Several experiments are implemented to test each part of our framework. We first use our model to estimate the consumer's reservation price using simulated transaction set. The results were used for exploring the best bundling configuration.

	Normal Distribution	Uniform Distribution
year 1	59,529	59,055
year 2	59,356	59,050
year 3	59,556	59,056
year 4	59,057	58,681
year 5	59,194	58,887
year 6	59,580	59,094
year 7	59,227	59,113
year 8	59,270	58,820
year 9	59,320	59,024
year 10	59,720	59,398
Total	593,809	590,178

Table 4.2 Number of transactions filtered by reservation price matrix

4.2.1 Reservation Price Estimation

In order to evaluate the accuracy of the proposed model, we compare the estimated reservation price with the matrix we generated. Our model is also compared with other two methods. The all-month estimation model does not consider the time dimension so that it uses historical transactions in all months for prediction. On contrary, the same-month estimation model uses only the transactions in the same month with the one being predicted. For example, if we are going to estimate consumer's reservation price in January, the all-month estimation model uses the whole year transactions in each year, while the same-month estimation model uses only historical transactions generated in January for estimation. However, our model analyzes previous sales records, discovers the months which have similar situations with January, and uses them in estimation.

We use the estimation result for a single item instead of the whole dataset to reveal the comparison of different models more clearly. We pick up transactions of the product PRO028 and extract its price and sales volume in each month. Figure 4.2 shows the statistic in the first five years under different reservation price distributions.

Input:

- N: the number of products;
- M: the number of consumers;
- S: the number of years covered by transaction dataset;
- r: average inflation rate in S years.

Output: A set of transaction records.

```
Algorithm:
BEGIN
    generate product_id for each product $i$ in $I = \{i_1, i_2, \cdots, i_N\}$ ;
    generate consumer_id for each consumer $c$ in $C = \{c_1, c_2, \cdots, c_M\}$;
    randomly generate a price for each product $P = \{p_1, p_2, \cdots, p_N\}$;
    generate candidate transaction set $CT$ = yearlyTransaction_gen (I, C, P);
    for each $c \in C$ and $i \in I$
        randomly generate a reservation price $R_{c,i}$;
        add $R_{c,i}$ to R;
    for each $t(c, i, p_{i,k}, y, k)$ in CT
        if ($p_{i,k} \leq R_{c,i}$) then
            generate a random tranction_id to t;
            add t to T;
    return T;
END

procedure yearlyTransaction_gen (I: product list; C: consumer list; P: price list)
BEGIN
    generate candidate transaction set for the first year $y_1$
        $CT_1$ = monthlyTransaction_gen (I, C, P, $y_1$);
    add $CT_1$ to CT;
    for each year $y_s$, $s \in \{2, 3, \cdots, S\}$
        for each $t(c, i, p_{i,k}, y, k)$ in $CT_1$
            generate a new transaction $t'$ with
                $c' = c, i' = i, p'_{i,k} = p_{i,;} \times (1 + r)^{s-1}, y' = y_s, k' = k$;
            add $t'$ to $CT_s$;
        add $CT_s$ to CT;
    return CT;
END

procedure monthlyTransaction_gen (I: product list; C: consumer list; P: price list; y: year)
BEGIN
    for each month $k \in \{1,2, \cdots, 12\}$
        for each $i \in \{i_1, i_2, \cdots, i_N\}$
            randomly generate a seasonal rate $pr \in [-0.2, 0.2]$;
            calculate price of $i$ in month $k$ $p_{i,k} = p_i \times (1 + pr)$;
            generate a transaction $t(c, i, p_{i,k}, y, k)$ for each consumer;
            add t to $CT_1$;
    return $CT_1$;
END
```

Figure 4.1 Pseudo-code for transaction dataset simulation

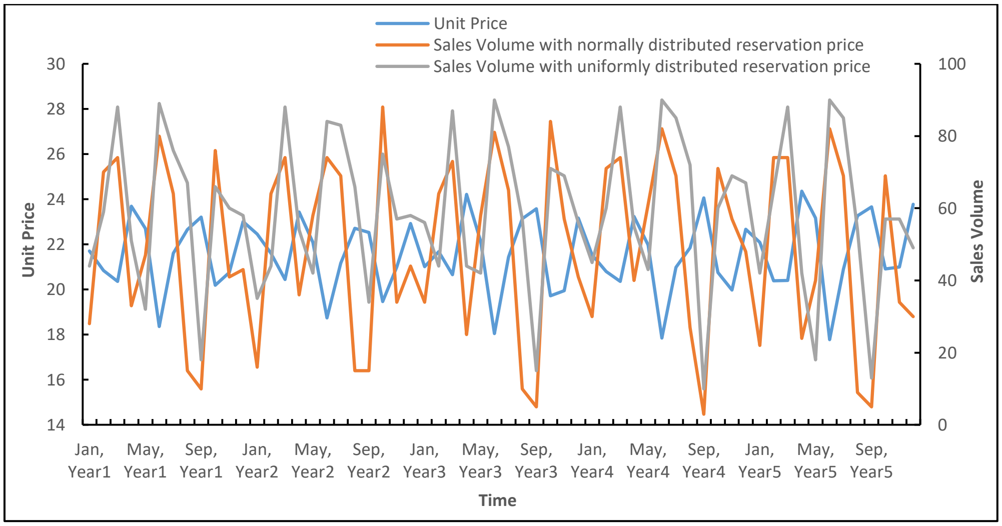
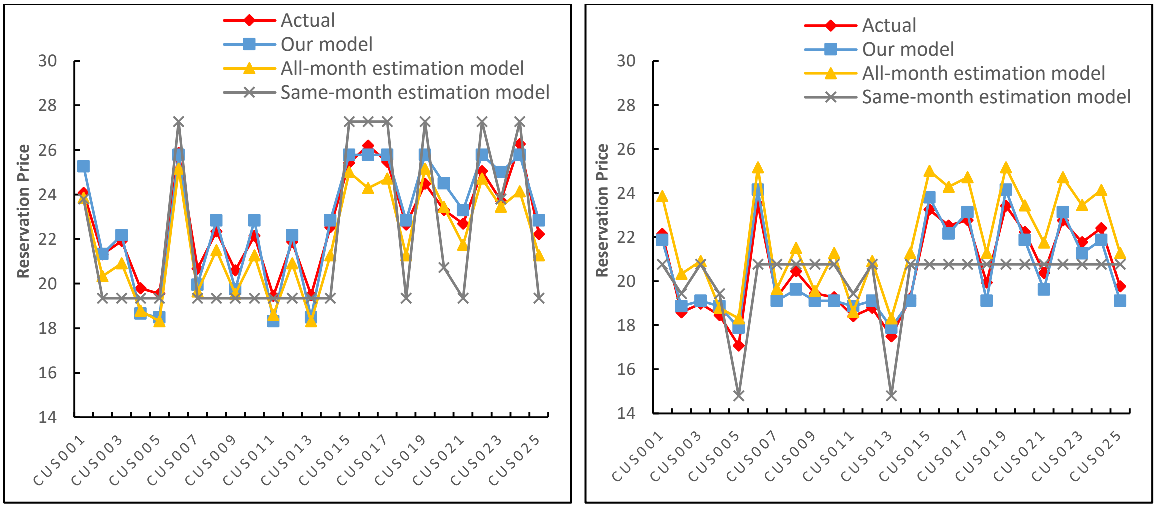


Figure 4.2 Unit price and sales volume for product PRO028 in the first five years under normally and uniformly distributed reservation price

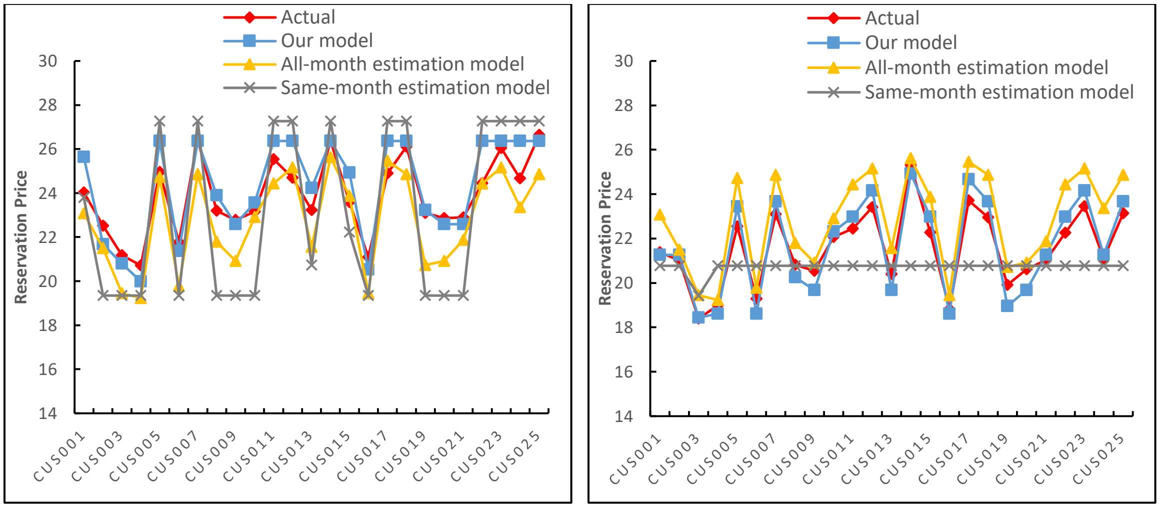
From Figure 4.2 we can easily find that PRO028 has significant fluctuations in both unit price and sales volume during a year. Fluctuations in each year form a relatively stable pattern, which keeps repeating over and over again during the period we analyze. Usually, the sales will rise up with a lower price and drop down with a higher price when the consumer's reservation price stay stable. However, by comparing the trend of unit price and sales, we find the relatively low price in January didn't bring a high volume. Instead, its volume is lower than that in December which has a higher unit price. A similar situation also occurs in April and September. These contradictions are caused by various reservation prices while making purchases in different months.

We picked up the estimation result for 25 consumers in two months which are separated into different clusters by *k*-Means in PED analysis. Unit price in April always stays high, resulting in lower sales every year. While June has an opposite situation with lower unit price and higher sales. From Figure 4.3, we find the consumer's reservation price in April is higher than that in June. However, since the all-month estimation model considers transactions in the whole year, the estimation reveals an approximate range, but it is always same and stays in middle regardless of the target month. Therefore, estimated reservation price is lower than the actual value in April while higher than that in June. The same-month estimation model only takes the transaction in the same month as the target one, resulting in insufficient purchase behaviors for estimation. This model is successful in separating consumers with significant differences but fails to differ the ones with similar reservation price. Compared to these two models, our model using clustering for PED analysis can not only estimate the consumer's reservation price in a single month more accurately, but also distinguish the months in which consumers may have different valuations and purchase behaviors. Similar biases are also detected in the result with uniformly distributed reservation price, which is shown in Figure 4.4.



a. Estimation result in April b. Estimation result in June

Figure 4.3 Estimated result under normally distributed reservation price for 25 consumers



a. Estimation result in April b. Estimation result in June

Figure 4.4 Estimated result under uniformly distributed reservation price for 25 consumers

To show the improvement of our model over all products, we plot the average error of 100 products obtained by six models (three for each reservation distribution) in each month (see Figure 4.5). For the product with a high price, we allow a relatively wide range of bias, while the tolerance for cheap products is much smaller. Therefore, we use Mean Absolute Percentage Error (MAPE) as the measurement (see Equation (8)).

$$\text{MAPE} = \frac{1}{M*N}\sum_{n=1}^{N}\sum_{m=1}^{M}\frac{|R_a - R_p|}{R_a} \tag{8}$$

For both normally and uniformly distributed reservation price, our model achieves the best performance with MAPE around 3.5%. The possible bias means if a consumer's actual reservation for a single item is \$50, our estimation falls within the range of \$48 and \$52. The performance of the all-month estimation model is much better than the same-month model, ranking in the middle in comparison. The major reason for a higher bias is the failure in distinguishing potential variances of reservation price in different months. Insufficient purchase records make the same-month estimation model the worst one. MAPEs are always greater than 7%, representing the bias can be up to \$3.5 when a consumer's actual reservation equals to \$50.

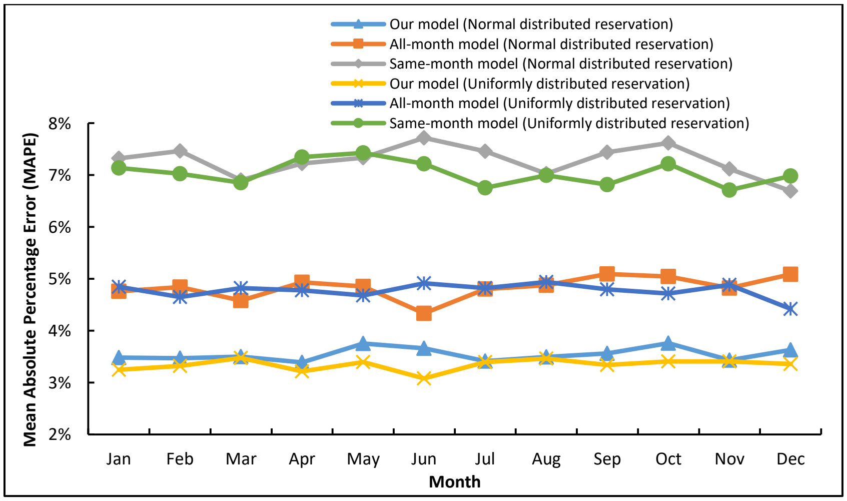
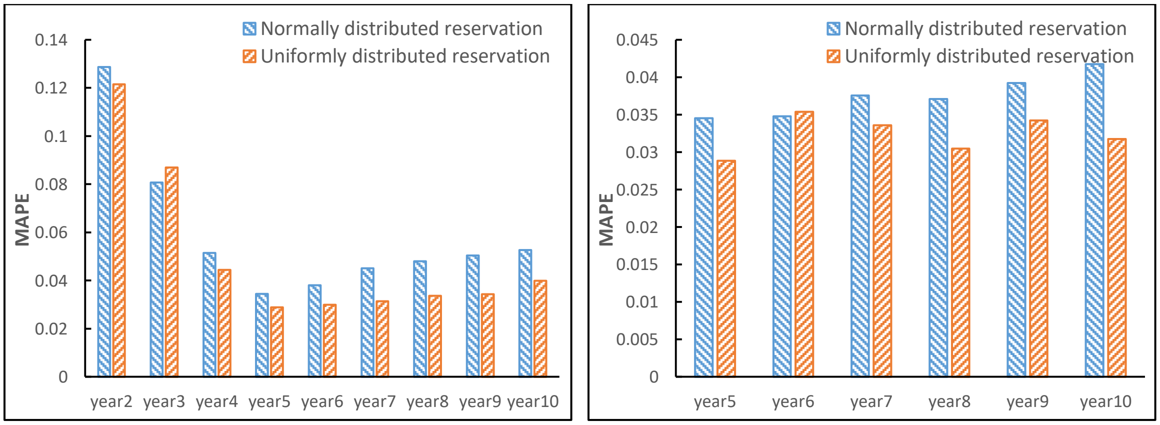


Figure 4.5 Average Mean Absolute Percentage Error (MAPE) of 100 products in each month using different models

4.2.2 Moving Validation

To validate the accuracy of the model in prediction, we adopt the "moving" validation approach introduced in Chu and Zhang's work [13]. That is, using the monthly sales and unit price in several continuous years (in-sample) to estimate the consumer's reservation price and predict the yearly sales in the following year (out-of-sample). We adopt in-sample with both variable and fixed length to explore the effect of in-sample length on the accuracy of predicting future purchase behaviors. For each in-sample, months are re-clustered using the corresponding sales and unit price so that the estimation can eliminate the effect of dynamic factors but catch the trend if it tends to stable. Figure 4.6(a) shows the average MAPE for the annual sales of all products using different in-sample lengths. The annual sales in year2 is predicted using only transactions in the year1, and the sales in year3 is predicted using transactions in both year1 and year2, and so on. As shown in the figure, MAPE decreases a lot with the length of in-sample growing until it reaches the lowest in year5, which means it is optimal to use transactions in previous four years to predict consumers' behaviors in the next year. MAPE with an in-sample length longer than four rises again. The increase is more obvious in the normally distributed reservation. A longer in-sample period can eliminate the effect of dynamic factors like climate change and special events. However, regarding the product lifecycle, an overlong in-sample may result in a higher bias causing by product replacement and upgrading. Considering these factors and average MAPE shown in Figure 4.6(a), we fixed the length of in-sample to four years and the out-of-sample covers year5 to year10. MAPEs of prediction for sales in these six years are plotted in Figure 4.6(b). Prediction error fluctuates in a small range, representing our model can produce a stable result with the moving in-sample. This "moving" validation schema can evaluate the stability and reliability of the proposed model.



a. In-sample with variable length b. In-sample with fixed length

Figure 4.6 Average MAPE of annual sales prediction with different in-samples

4.2.3 Bundle Design

Our bundle design algorithm is based on frequent itemsets obtained by association mining. The choice of three bundling strategies is made by comparing the absolute revenue gain created by each strategy. The one which creates the most revenue gain is selected as the bundling strategy for promotion. Table 4.3 shows the number of bundles before and after bundle selection with different *min_sup* values when the bundling coefficient is set to 0 by default. We only consider the itemsets with more than one item, because a bundle with only one item is equivalent to selling it individually. With *min_sup* increasing by 0.005 each round, the number of frequent itemsets decreases exponentially, as well as the number of bundles in each strategy.

In order to avoid overlapping and confliction among bundles, we adopt bundle selection based on the absolute revenue gain they provide. Only a small part of frequent itemsets are selected as eligible bundles. When the *min_sup* is relatively small, most frequent itemsets are more profitable in mixed bundling than in pure bundling. With the *min_sup* growing, the itemsets which create more revenue in pure bundling occupy a larger proportion.

	Before bundle selection				**After bundle selection**			
min_sup	*Total*	*Pure components*	*Pure bundling*	*Mixed bundling*	*Total*	*Pure components*	*Pure bundling*	*Mixed bundling*
0.025	3429	28	862	2539	48	0	9	39
0.03	2352	20	672	1660	47	0	7	40
0.035	1400	9	457	934	39	0	9	30
0.04	696	3	243	450	28	0	6	22
0.045	284	0	117	167	18	0	5	13
0.05	93	0	48	45	8	0	5	3
0.055	23	0	13	10	4	0	2	2

Table 4.3 The number of bundles with different *min_sup* values

To evaluation the effect of this algorithm regarding the revenue maximization objective, we use the following measurements.

Revenue Gain. One is to measure how much the sellers can benefit from bundling. We compare the revenue created by bundling over the baseline which is the revenue created by selling products individually. Revenue gain is the percentage of growth over the revenue of pure components.

Surplus Gain. Another is to evaluate how much consumers can benefit from bundling. A consumer's surplus is the difference between his reservation price and the product's actual price [18]. A higher surplus gain shows the improvement in consumer's willingness to pay and satisfaction. Similar to revenue gain, surplus gain is represented by the percentage of growth in surplus of bundling over pure components.

Figure 4.7 shows the revenue gain with different *min_sup* values. We also calculate the bundling efficiency with is the average gain generated by each bundle. Revenue can be increased by more than 10% by only four bundles with two products in each one when *min_sup* is set to 0.055. As *min_sup* decreases by 0.005 each round, revenue gain rises up with a decreasing rate. Although the revenue gain with a smaller *min_sup* is higher than that with a larger *min_sup*, bundling efficiency drops down a lot, indicating the higher revenue gain is the result of the growing amount of eligible bundles rather than efficiency. Bundling efficiency reached the peak when *min_sup* is set to 0.05, where each bundle can generate around 4% revenue gain on average. This also happens to surplus gain.

Experiment result shows itemsets which are frequently purchased together but sometimes separated are more profitable to be sold as bundles. Regarding the revenue gain and bundling efficiency, we choose $min\_sup = 0.04$ as the default setting in the rest of this paper. Revenue gain is around 46.8% and surplus gain is 71.8% compared with selling products individually.

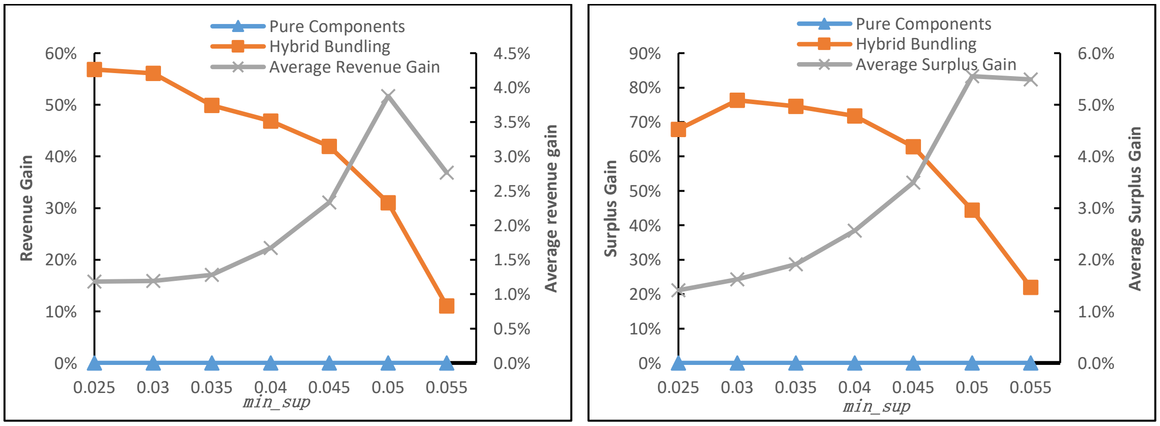


Figure 4.7 Experiments with different *min_sup* values

Bundling Coefficient. The bundling coefficient λ in our research can reveal the type of bundling. Figure 4.8 shows the effect of λ on revenue and surplus gain respectively. The line of hybrid bundling is the experiment result using our model. The other two lines show the revenue/surplus gain created by pure bundling and mixed bundling among qualified bundles.

A negative λ means the consumer's reservation price for a bundle is lower than the sum of reservation for each component (subadditivity), which happens to substitutes. When λ is smaller than -0.15, mixed bundling is the only source of revenue gain. The advantage of mixed bundling becomes outstanding because it can offer bundles to the consumers with higher reservation price while offering components to others. However, the revenue gain may be at the expense of consumer surplus since there is no surplus gain revealed. Such bundles are not desired regarding consumer satisfaction for a long term. Revenue and surplus gain come from pure bundling increase gradually, but they are still much lower than that provided by mixed bundling. Therefore, mixed bundling is more profitable for substitutes.

A positive λ applies when items in a bundle are complementary, where consumers have super additive reservations. Overall revenue and surplus gain augment with a higher λ. From Figure 4.8 we can also find, pure bundling is very sensitive to the increase of λ. Revenue and surplus gain created by pure bundling climb dramatically until pure bundling becomes the most profitable strategy for all qualified bundles. Mixed bundling becomes less desirable since consumers tend to purchase bundles instead of components. Our result agrees with Do, Lauw, and Wang's research [18].

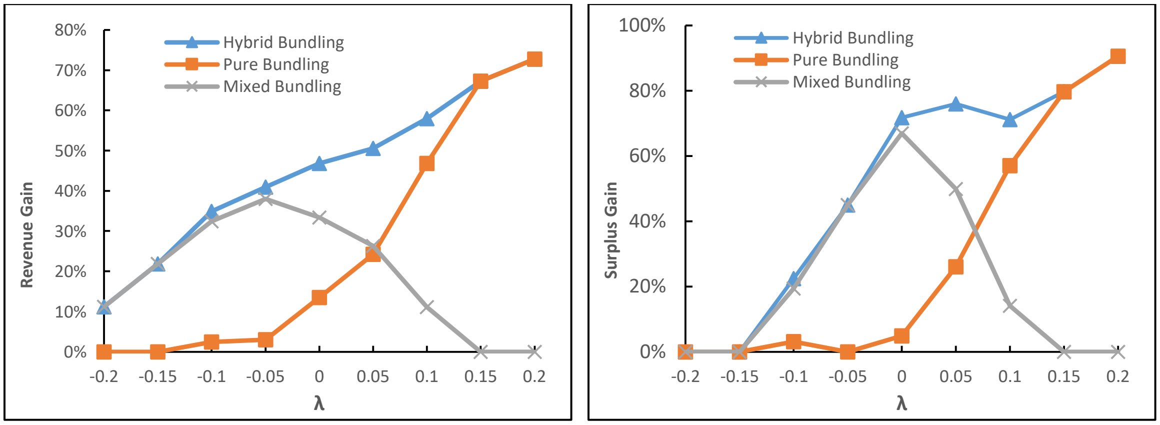


Figure 4.8 Experiments with different λ

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this paper, we have proposed a data mining framework for bundle design and pricing. The objective of this framework is to estimate the consumer's reservation price using data-driven methods and solve the bundle configuration problem.

All previous studies either make strong assumptions on the consumers' reservation prices or estimate the consumers' reservation prices based on a small amount of marketing surveys. In this framework, we incorporate the consumers' reservation prices based on historical purchasing data in order to reduce the bias caused by possible changes in the consumers' willingness and abilities to pay in different time periods. Through simulations and experiments, we have demonstrated this framework is capable of estimating consumers' reservation prices accurately, as well as solving the bundle design and the bundle pricing problem. The Mean Absolute Percentage Error (MAPE) of our model is 1.5% lower than that using transactions in the whole year and 3.5% lower than that using transactions in the same months when estimating consumers' reservation prices. By applying ANOVA test with $\alpha = 0.05$ significance, the p-value for all six models with different distributed reservations is less than 2×10^{-16}, which shows these models are significantly different. By predicting future sales using the estimation results, we find the MAPE drops with the increasing of in-sample length, but rises again with an overlong in-sample. Therefore, it is essential to choose an appropriate in-sample length when predicting consumers' purchase behaviors in future.

Instead of using all possible combinations as candidate bundles, we apply association mining with different minimum supports to generate frequent itemsets and select non-overlapping itemsets with the highest absolute revenue gain as qualified bundles. Experiment result shows itemsets which are frequently purchased together but sometimes separated are more profitable to be sold as bundles. Revenue gain created by bundling is around 46.8% and surplus gain is around 71.8% comparing with selling products individually.

5.2 Future Work

We recommend future work in the following aspects:

- As this framework is not limited to specific data mining algorithms for its various sub-tasks, we plan to explore and compare different data mining algorithms within this framework in future.
- We apply PCA and *k*-Means to detect the direction of demand curves and find month clusters with similar reservation price within each one, which can achieve the best performance when the demand curves are straight lines. However, a straight demand curve rarely occurs in reality. When the curvature of a demand curve is relatively high, months on different curves may be grouped into the same cluster, resulting in bias when estimating the consumer's reservation price. In future, different clustering technologies may be adopted to detect month clusters more accurately in the situation of high-curvature demand curves.
- Seasonality is one of the major tasks in time series analysis, which is about detecting the regularity of repetition over a fixed period of time in a dataset. Generally speaking, seasonality can be tackled by ARIMA model and seasonal coefficient. The former one adjusts a mathematical model to fit the random sequence generated by the time pass. The latter one calculates the coefficient for each month which can eliminate the effect of random factors and overall trend. Seasonality analysis can be added to the framework as an additive feature to understand consumers' behaviors in different seasons better.
- When estimating the reservation price of a new consumer, or the one with few purchase records, consumers in the same segments can be referred. Customer segmentation can use features including both profile and purchase history.

Bibliography

[1] G. Adomavicius, J. Bockstedt and S. P. Curley. “Bundling effects on variety seeking for digital information goods,” Journal of Management Information Systems, 31(4), 2015, pp. 182-212.

[2] R. Agrawal, T. Imieliński, & A. Swami, “Mining association rules between sets of items in large databases,” Acm Sigmod Record, Vol. 22, No. 2, 1993, pp. 207-216.

[3] R. Agrawal, & R. Srikant, “Fast algorithms for mining association rules,” 20th international conference on very large databases, VLDB, Vol. 1215, 1994, pp. 487-499.

[4] C. P. Alderfer. “Existence, relatedness, and growth: Human needs in organizational settings,” 1972.

[5] A. Ansari, S. Siddarth and C. B. Weinberg. “Pricing a bundle of products or services: The case of nonprofits,” Journal of Marketing Research, 1996, pp. 86-93.

[6] P. Andritsos, Data clustering techniques. Rapport technique, University of Toronto. Department of Computer Science, 2002.

[7] Y. Bakos and E. Brynjolfsson. “Bundling and competition on the internet,” Marketing Science, 19(1), 2000, pp. 63-82.

[8] M. Benisch and T. Sandholm. “A framework for automated bundling and pricing using purchase data,” in Auctions, Market Mechanisms, and their Applications Anonymous, 2012.

[9] C. Birtolo, D. De Chiara, M. Ascione and R. Armenise. “A generative approach to product bundling in the e-commerce domain,” Nature and Biologically Inspired Computing (NaBIC), IEEE, 2011, pp. 169-175.

[10] E. T. Bradlow and V. R. Rao. “A hierarchical bayes model for assortment choice,” Journal of Marketing Research, 37(2), 2000, pp. 259-268.

[11] W. Chang and S. Yuan. “A markov-based collaborative pricing system for information goods bundling,” Expert Systems with Applications, 36(2), 2009, pp. 1660-1674.

[12] P. Chiambaretto and H. Dumez. “The role of bundling in firms' marketing strategies: A synthesis,” Recherche et Applications en Marketing (English Edition), 27(2), 2012, pp. 91-105.

[13] C. W. Chu, & G. P. Zhang. “A comparative study of linear and nonlinear models for aggregate retail sales forecasting,” International Journal of production economics, 86(3), 2003, 217-231.

[14] J. Chung and V. R. Rao. “A general choice model for bundles with multiple-category products: Application to market segmentation and optimal pricing for bundles,” Journal of Marketing Research, 40(2), 2003, pp. 115-130.

[15] City Pass. Internet: http://www.citypass.com/toronto, 2016

[16] G. S. Crawford and A. Yurukoglu. “The welfare effects of bundling in multichannel television markets,” The American Economic Review, 102(2), 2011, 643-685.

[17] J. D. Dana and K. E. Spier. “Do tying, bundling, and other purchase restraints increase product quality?” International Journal of Industrial Organization, 2015.

[18] L. Do, H. W. Lauw and K. Wang. “Mining revenue-maximizing bundling configuration,” Proceedings of the VLDB Endowment, 8(5), 2015, pp. 593-604.

[19] H. Estelami. “Consumer savings in complementary product bundles,” Journal of Marketing Theory and Practice, 7(3), 1999, 107-114.

[20] P. H. Farquhar and V. R. Rao. “A balance model for evaluating subsets of multiattributed items,” Management Science, 22(5), 1976, pp. 528-539.

[21] K. D. Ferreira and D. D. Wu. “An integrated product planning model for pricing and bundle selection using markov decision processes and data envelope analysis,” International Journal of Production Economics, 134(1), 2011, pp. 95-107.

[22] M. Gera, & S. Goel, “Data Mining-Techniques, Methods and Algorithms: A Review on Tools and their Validity,” International Journal of Computer Applications, 113(18), 2015.

[23] S. M. Goldberg, P. E. Green and Y. Wind. “Conjoint analysis of price premiums for hotel amenities,” Journal of Business, 1984, pp. S111-S132.

[24] J. Han, & M. Kamber. Data Mining: Concepts and Techniques, 2006.

[25] W. Hanson and R. K. Martin. “Optimal bundle pricing,” Management Science, 36(2), 1990, pp. 155-174.

[26] R. S. Hiller. “Profitably bundling information goods: Evidence from the evolving video library of Netflix,” 2015.

[27] D. Honhon and X. Pan. “Improving retail profitability by bundling vertically differentiated products,” 2015.

[28] Y. Jiang, J. Shang, C. F. Kemerer and Y. Liu. “Optimizing e-tailer profits and customer savings: Pricing multistage customized online bundles,” Marketing Science, 30(4), 2011, pp. 737-752.

[29] A. Karageorgos and E. Rapti. “Dynamic generation of personalized product bundles in enterprise networks,” On the Move to Meaningful Internet Systems: OTM 2013 Workshops. 2013.

[30] A. Klein and N. Jakopin. “Consumers’ willingness-to-pay for mobile telecommunication service bundles,” Telematics and Informatics, 31(3), 2014, pp. 410-421.

[31] T. Kohlborn et al., “Conceptualizing a bottom-up approach to service bundling,” Presented at Advanced Information Systems Engineering. 2010.

[32] S. Kotsiantis, & D. Kanellopoulos, “Association rules mining: A recent overview,” GESTS International Transactions on Computer Science and Engineering, 32(1), 2006, pp. 71-82.

[33] A. Krishna, D. R. Lehmann and C. Mela. “Impact of bundle type, price framing and familiarity on purchase intention,” Journal of Business Research, 33, 1995, pp. 57-66.

[34] Y. LeCun, Y. Bengio, & G. Hinton, “Deep learning,” Nature, 521(7553), 2015, pp. 436-444.

[35] B. Letham, W. Sun and A. Sheopuri. “Latent variable copula inference for bundle pricing from retail transaction data,” Presented at Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014.

[36] G. R. Liu and X. Z. Zhang. “Collaborative filtering based recommendation system for product bundling,” 2006 International Conference in Management Science and Engineering, IEEE, 2006.

[37] K. Mikkonen, H. Niskanen, M. Pynnönen and J. Hallikas. “The presence of emotional factors: An empirical exploration of bundle purchasing process,” Telecommunications Policy, 39(8), 2015, pp. 642-657.

[38] R. Peter, “Data Mining Solutions for the Business Environment,” Database Systems Journal, 4(4), 2013, pp. 21-29.

[39] A. Prasad, R. Venkatesh and V. Mahajan. “Product bundling or reserved product pricing? Price discrimination with myopic and strategic consumers,” International Journal of Research in Marketing, 32(1), 2015, pp. 1-8.

[40] E. Rapti, A. Karageorgos and G. Ntalos. “Adaptive constraint and rule-based product bundling in enterprise networks,” 2014 IEEE 23rd International WETICE Conference, 2014, PP. 15-20.

[41] I. S. Razo-Zapata et al., “Dynamic cluster-based service bundling: A value-oriented framework,” 2011 IEEE 13th Conference on Commerce and Enterprise Computing, 2011, pp. 96-103.

[42] D. Somefun and J. La Poutré. “Bundling and pricing for information brokerage: Customer satisfaction as a means to profit optimization,” Presented at Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on. 2003.

[43] P. R. Thimmapuram, J. Kim, A. Botterud, and Y. Nam, “Modeling and simulation of price elasticity of demand using an agent-based model,” Innovative Smart Grid Technologies (ISGT), IEEE, 2010, pp. 1-8.

[44] C. Tsai, Y. Hu and Y. Lu. “Customer segmentation issues and strategies for an automobile dealership with two clustering techniques,” Expert Systems, 32(1), 2015, pp. 65-76.

[45] R. Venkatesh and W. Kamakura. “Optimal Bundling and Pricing under a Monopoly: Contrasting Complements and Substitutes from Independently Valued Products,” Journal of Business, 76(2), 2003.

[46] R. Venkatesh and V. Mahajan. “A probabilistic approach to pricing a bundle of products or services,” Journal of Marketing Research, 1993, pp. 494-508.

[47] R. Venkatesh, and V. Mahajan, “The design and pricing of bundles: a review of normative guidelines and practical approaches,” Handbook of pricing research in marketing, 2009, pp. 232.

[48] H. Wang and S. Wang. “A knowledge management approach to data mining process for business intelligence,” Industrial Management & Data Systems, 108(5), 2008, pp. 622-634.

[49] M. S. Yadav and K. B. Monroe. “How buyers perceive savings in a bundle price: An examination of a bundle's transaction value,” Journal of Marketing Research, 1993, pp. 350-358.

[50] E. Yakıcı, O. Ö. Özener and S. Duran. “Selection of event tickets for bundling in sports and entertainment industry,” Computers & Industrial Engineering, 74, 2014, pp. 257-269.

[51] B. Yang and C. Ng. “Pricing problem in wireless telecommunication product and service bundling,” European Journal of Operational Research, 207(1), 2010, pp. 473-480.