A Data Mining Framework for Product Bundle Design and Pricing

Yiming Li
Faculty of Computer Science
Dalhousie University
Halifax, Canada
email: ym510041@dal.ca

Hai Wang
Sobey School of Business
Saint Mary's University
Halifax, Canada
email: hwang@smu.ca

Qigang Gao
Faculty of Computer Science
Dalhousie University
Halifax, Canada
email: qggao@cs.dal.ca

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 products should be in a profitable bundle, and what the proper price is for a bundle. In this paper, we propose a new data mining framework for product bundle design and bundle pricing. This framework 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 profit. We also demonstrate the efficiency of this data mining framework through experiments and simulations.**

Keywords-data mining; bundling; bundle design; bundle pricing; marketing strategy

I. INTRODUCTION

To meet consumers' needs and expectations is the basic principle for sellers to survive in the current fierce competitive business environment. Sellers often adopt various promotion strategies to attract more consumers/buyersto increase their revenues. Bundling is a promotion strategy in which sellers provide multiple products or events as a single package with a discounted price [21]. Bundling has become prevalent as it can benefit bothbuyers and sellers.

From the consumer/buyer's perspective, bundling can provide benefits such as

- 8% monetary savings on average [8].
- Saving search cost, which will increase the willingness to purchase [17].

From the seller's perspective, selling bundles can benefit them from the following aspects:

- Increasing the number of buyers and thus increasing sales [8].
- Easier for newly released products to be noticed and accepted by consumers/buyers [17].
- Saving packaging and distribution cost [4].

In this paper, we propose a data mining framework for solving the bundle design problem. Our framework can help sellers obtain a good knowledge of their consumers by analyzing their purchase patterns and reservation price in different time periods, as well as design profitable bundles with proper price and strategies to increase sales. We demonstrate that the proposed framework is able to achieve significantly better performance on analyzing the *price elasticity of demand* (PED) and estimating the buyers' *reservation prices*. The PED measures the change of quantity demanded of a product with respect to the changes of the price, with other things being equal [20]. The *reservation price* of a buyer is the highest price that the buyer is willing to pay for a particular product [22]. 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 PED in order to obtain accurate estimation of buyers' reservation prices. Considering time as a factor can help with understanding consumers' purchase behaviors in different time periods and designing proper bundles to meet consumers' varying requirements, which is missing in previous bundling studies. The estimated buyers' reservation prices serve as the basis for bundle design and bundle pricing.
- Our proposed framework is generic and does not limit to specific data mining algorithms. For example, new association rule mining algorithms can be integrated into our proposed framework to improve the efficiency and effectiveness for determining the possible product combinations within a bundle.

The remainder of this paper is organized as follows. Section II formulates the bundle design problem and the bundle pricing problem. Section III describes our data mining framework for bundle design and pricing. Section IV shows the performance of the proposed approach through experiments and simulations. Section V remarks the conclusions and the future work.

II. RELATED WORK AND PROBLEM FORMULATION

Bundling has been extensively studied and applied in retailing [2][13][15][16], entertainment [6][22], e-commerce [1][3][16][19], travel planning [9][10], telecommunication [23] as well as the service sector [12][18].

Two main problems associated with bundling in the previous research literature are the bundle design problem and the bundle pricing problem. Suppose that there are N distinct products available for bundling, the $2^N-(N+1)$ possible product combinations for bundling (excluding the bundles with a single product) make this problem extremely complex for sellers,

 ISBN: 978-1-61208-558-6

especially when *N* is large [9]. Bundle design is a process of selecting product combinations to be promoted and sold as a bundle, which should be rational, practical, and in accordance with consumers' preferences. The main objective for providing bundles is to attract more buyers and hence increase the sales for sellers.

Moreover, a more remarkable principle that needs to be considered in bundle design is to know exactly what consumers/buyers want. It is more beneficial for sellers to provide flexible bundles that consumers can choose along with their preferences and needs.

Bundle pricing is about deciding the optimal price for a bundle. Objective of the term "optimal" here can vary based on their different business goals, such as maximization of profit, revenue, attendance, or market share [7].

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 bundle design [22]. The relationship between reservation price and the actual price of a product determines whether or not a consumer will make a purchase. Another factor in bundle design is bundling strategy. Three bundling strategies have been widely studied in previous research. *Pure component*, or unbundling, is the traditional way in which consumers/buyers can only purchase products or services separately with their original prices [20]. It allows consumers to see the sales process clearly and pick up exactly the product 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 [20]. Combining these two strategies, the *mixed bundling* strategy is a more flexible one that the seller offers both individual products and the whole bundle [20].

In this paper, we will focus on data mining techniques for solving (1) the bundle design problem, which is to determine what product combinations should be in a bundle, (2) the bundle pricing problem, which is to determine the "optimal" price for a bundle, given a specific bundling strategy (i.e., the pure component, pure bundling or mixed bundle strategy).

III. The Proposed Data Mining Framework

We propose a data mining framework for bundle design and pricing, which is illustrated in Figure 1. One of the important features of the proposed framework is to incorporate the time value of money for estimating consumers' reservation prices. The notations used in our proposed framework are listed in Table I.

TABLE I. NOTATIONS

N	The number of items for sale I = $\{i_1, i_2, \dots, i_N\}$
M	The number of consumers C = $\{c_1, c_2, \dots, c_M\}$
T	The set of transaction data generated by consumers. Observations that belong to a specific consumer *c* with an item *i* can be represented as $\{T_{c,i}\}_{c\in C, i\in I}$
S	The number of years that covered by transaction datasetY = $\{y_1, y_2, \dots, y_S\}$
p	Unit price of an item
v	The sales volume for an item
RI	An M×N matrix containing consumers' reservation price intervals for all products

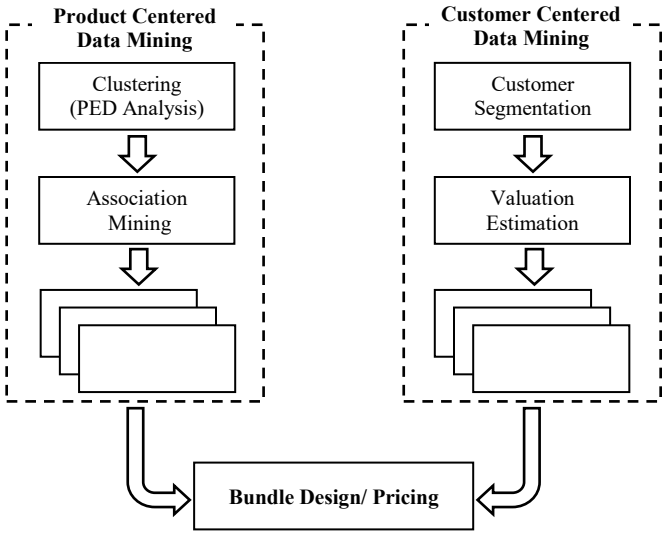


Figure 1. The data mining framework for products bundle design and pricing

Features of the Framework

1) The PED analysis

PED is used to measure the change of quantity demanded of a product or service in its price, with other things being equal. For elastic products, an increase in unit price will lead to fewer units sold, resulting in a downward-slopping curve in its graphic representation with quantity on the horizontal axis and price on the vertical axis.

The demand curve also expresses consumers' willingness and abilities to pay for a product in a given 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 Principle Component Analysis (PCA) and K-means algorithm to analyze the fluctuations of the consumer's reservation price.

Given a set of transaction data, sales volume and price for an item in a month can be extracted easily. The average price is calculated if 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 (1)), assuming p_m and $v_m$are unit price and volume of an item in the month *m*. The objective to use the 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.

$$\overline{p_m} = \frac{1}{S} * \sum_{m=1}^{S} p_m \quad (1)$$

$$\overline{v_m} = \frac{1}{S} * \sum_{m=1}^{S} v_m$$

The $(\overline{v_m}, \overline{p_m})$ 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. 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 the data, and transforming the original data into the data expressed in terms of new axes. We adopt PCA to find the principle component in downward-slopping 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 *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 value of K depends on a heuristic learning method using Within Cluster Sum of Squared Error (WCSSE), defined as

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

where p is an object in data collection, m_i is the mean value of all objects in cluster G_i[11]. 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 2. Black points are original points representing the relationship between quantity and unit price in each month. Colored points are the mapping result using PCA. Points in an oval are the ones being grouped in a cluster using K-means algorithm.

2) Customer segmentation

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. Customers' 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) Valuation estimation

A consumer's reservation price for the same 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.

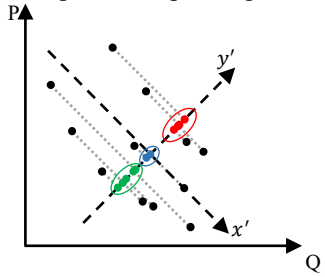


Figure 2. Process of PCA and K-means

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 historical price can be calculated using (3).

$$\mathrm{PV} = \mathrm{p} \times (1 + r)^n \qquad (3)$$

If we are going to estimate consumers' reservation price and generate profitable bundles in the month m, only the months that 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 the transaction set, pick up the records that happened in the month in G_m along with their timestamp and price mapped to present value. We assumed his valuation of a given product equals to its price when he made the first purchase. The relationship between the consumer's reservation price R and the number of purchases np forms the following function $R = (1 + \theta)^{np} \times PV$. Each successful transaction makes his 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 assumed their valuation were less than the posted 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 play the same role in estimation when $\beta = 1$. Table II shows the purchase records for a consumer $c \in C$ with an item$i \in I$. A consumer's approximate reservation is estimated using (4).

$$R_{c,i} = \frac{\sum_{j=1}^{S} \sum_{m_k \in G_m} w_{y_j,m_k} \times v_{y_j,m_k}}{\sum_{j=1}^{S} \sum_{m_k \in G_m} w_{y_j,m_k}} \qquad (4)$$

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.

 ISBN: 978-1-61208-558-6

TABLE II. PURCAHSE RECORDS FOR CONSUMER *C* WITH PRODUCT *I*

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

Assuming the unit price for the item *i* is p_i, we create several intervals with each 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 (4) 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 value $RI_{c,i}$ represents the reservation price interval of the consumer*c* for the item *i*. We set the value to 0 for a consumer with the item he has never purchased.

However, since the valuation matrix only contains reservation price for individual items, we still need to predict their willingness to pay for a bundle *b*, which consists of more than one products. A recognized function deriving a consumer'svaluation for a bundle $R_{c,b}$ from its components $R_{c,i}$ proposed by Venkatesh and Kamakura is shown in (5) [20].

$$R_{c,b} = (1+\lambda) \times \sum_{i \in b} R_{c,i} \quad (5)$$

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, such as 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, $\lambda < 0$ indicating buyers do not want to pay as much as the total price when purchasing separately. λ equals to 0 when there is no relationship among the components in a bundle.

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 business but meaningless to consumers. Through basket analysis, we can find 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 often being purchased together obtained through *Apriori* algorithm with the minimum support *min_sup*.

b) Bundle design and pricing

Bundling configuration including determination of the bundling strategy and price is done based on the potential bundle set *B* (frequent itemsets in the *Apriori* algorithm) and consumers' valuation matrix *RI*. Unlike previous researches, which set the bundling strategy and its constraints as prerequisites, we calculate the revenue under 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 made several assumptions, which were also used in the previous studies [7].

- **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 products with price lower than 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 used in [7], $P(h_{c,i}|p_i, R_{c,i})$ represents the probability of the occurrence of *c* purchases *i* ($h_{c,i} = 1$) with the price p_i and his reservation price$R_{c,i}$. However, we develop it to P_{pc}, P_{pb} and P_{mb} under different promotion strategies.

For each possible combination in *B*, we calculate the maximum revenue it can create under 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(6).

$$r_{pc} = \sum_{i \in b} \sum_{c \in C} p_i \times P_{pc}\left(h_{c,i}|p_i, R_{c,i}\right) \quad (6)$$

where

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

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 need to be determined. Given all consumers' reservation prices for a bundle, we set cut-points p_b to calculate the number of consumers who will make a purchase and the corresponding revenue using (7). The one that makes r_{pb} maximized is chosen as the unit price for the bundle b.

$$r_{pb} = \sum_{c \in C} p_b \times P_{pb}(h_{c,b}|p_b, R_{c,b}) \quad (7)$$

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 (5) 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 posted prices imply $p_{XY} - p_X = \$6$, which is beyond his valuation of Y. Therefore, we set selection conditions shown below.

$$r_{mb} = \sum_{c \in C} [p_b \times P_{mb}(h_{c,b}|p_b, R_{c,b}) + \sum_{i \in b} p_i \times P'_{mb}(h_{c,i}|p_i, R_{c,i}, P_{mb})] \quad (8)$$

where

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

and

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

With all calculations finished, next step is the simple comparison of the results of (6) – (8) 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 confliction. 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, confliction will exist since components of A are also provided individually.
- Revenue maximization. With the prerequisite $\cup B = \mathrm{I}$, 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 that provides the highest absolute gain will be chosen for promotion, and then removed from the pickup pool along with the bundles that have items overlapped with it. Having the new set of candidate bundles, we still choose the one with the most absolute gain and repeat the process above until there is no bundle left. This method has no effect on bundling strategies 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.

Figure 3 summarizes the aforementioned features for bundle design and bundle pricing.

IV. Experiment And Evaluation

This section describes the simulation and experiments we did to test the effectiveness of our framework.

A. Simulating Transaction Data

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

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 transaction dataset for the following years is obtained based on the one generated in the last step by taking the inflation rate into consideration.

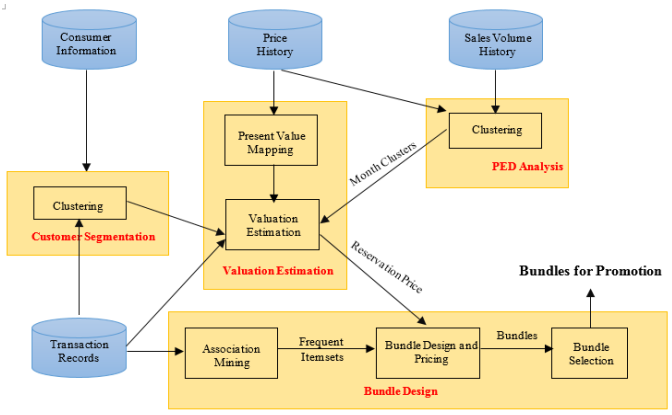


Figure 3. The system architecture of the features of the proposed framework

1) 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, consumer's reservation price is given by a normal distribution with mean of p_{base} and 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 uniform distribution is that we want to generate consumers' reservation price 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. 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. 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 a reservation price 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 our simulation are listed in Table III.

2) 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, form our simulated transaction set. Table IV shows the number of transactions in each year filtered by normally and uniformly distributed reservation price matrix respectively.

B. Training and Evaluation

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

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.

TABLE III. PARAMETER SETTINGS

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 IV. NUMBER OF TRANSACTIONS FILTERED BY RESERVATION PRICE MATRIX

	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

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 that have similar situation with January, and uses them in prediction.

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 in all years and extract its price and sales volume in each month. Figure 4 shows the statistic under different reservation price distributions in the first three years. Fluctuations in each year form a relatively stable pattern, which keeps repeating during the period. 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 did not 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.

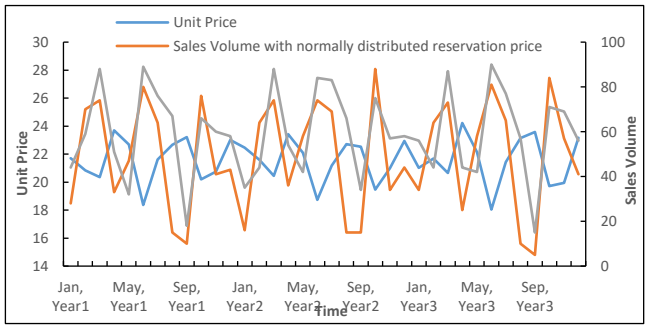


Figure 4. Unit price and sales volume for product PRO028 unider normally and uniformly distributed reservation price

To show the improvement of our model over all products, we plot the average squared loss of 100 products obtained by six models (three for each reservation distribution) in each month in Figure 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) defined in Equation (9) as the measurement. 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. Performance of the all-month estimation model are 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 variance 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.

$$\text{MAPE} = \frac{1}{M*N}\sum_{n=1}^{N}\sum_{m=1}^{M}\frac{|R_a - R_p|}{R_a} \quad (9)$$

2) Moving validation

To validate the accuracy of model in prediction, we adopt the "moving" validation approach introduced in Chu and Zhang's work [5]. 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 behavior. 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 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 Figure 6(a), MAPE decreases 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 in-sample length longer than four rises again. The increase is more obvious in 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 higher bias causing by product replacement and upgrading. Considering these factors and average MAPE shown in Figure 6(a), we fixed the length of in-sample to 4 years and the out-of-sample covers year5 to year10. MAPEs of prediction for sales in these six years are plotted in Figure 6(b). Prediction error fluctuates in a small range, representing our model can produced a stable result with the moving in-sample. This "moving" validation schema can evaluate the stability and reliability of the proposed model.

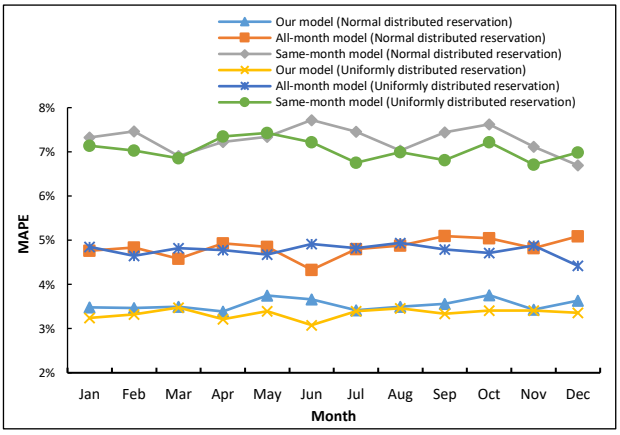


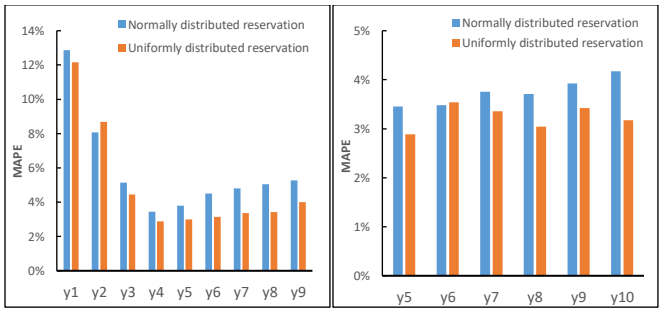
Figure 5. Average Mean Absolute Percentage Error (MAPE) of 100 products in each month using different models

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 that creates the most revenue gain is selected as the bundling strategy for promotion. Table V 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 that create more revenue in pure bundling occupy a larger proportion.

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



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

Figure 6. Average Mean Absolute Percentage Error (MAPE) of annual sales prediction uding different in-samples

TABLE V. THE NUMBER OF BUNDLES WITH DIFFERENT *MIN_SUP* VALUES

min_sup	Before bundle selection				After bundle selection			
	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

Revenue Gain. One is to measure how much the sellers can benefit from bundling. We compare the revenue created by bundling against 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 posted price [7]. 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 7 shows the revenue and surplus gain with different *min_sup* values. We also calculate the bundling efficiency, which 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. Regarding the revenue gain and bundling efficiency, bundling itemsets that are frequently purchased together but separately may be more profitable. Therefore, we choose $min\_sup = 0.04$ as the default setting in the rest of this paper.

Experiment result shows suitable itemsets can be sold as bundles. Revenue gain created by bundling is around 46.8% and surplus gain is around 71.7% comparing with selling products individually.

Bundling coefficient. The bundling coefficient λ in our research can reveal the type of 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 reservationprice 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 comes 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 higher λ. From Figure 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 [7].

V. CONCLUSIONS

In this paper, we have proposed a data mining framework for bundle design and pricing. In this framework, we incorporate the time value of money for data mining tasks, and estimate the consumers' reservation prices based on historical purchasing data. 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. The main contribution of this research is to integrate various existing techniques into a single framework. Through simulations and experiments, we have demonstrated this framework is capable of solving the bundle design problem, as well as the bundle pricing problem. As this framework does not limit to specific data mining algorithms for its various sub-tasks, we plan to compare different algorithms within this framework in future. Furthermore, we will incorporate various objective and subjective measures to evaluate the effectiveness and performance of different algorithms.

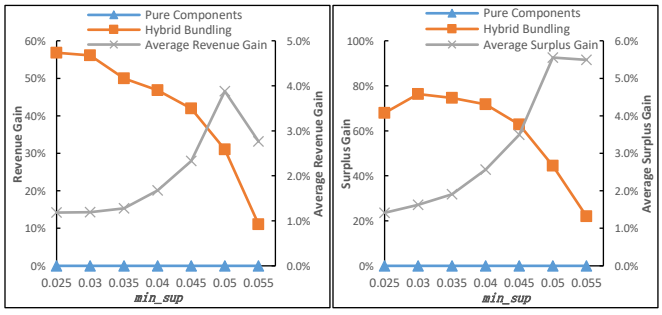


Figure 7. Experiments with different min_sup values

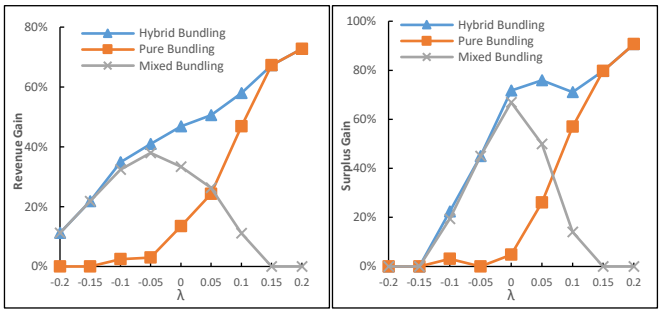


Figure 8. Experiments with different λ

REFERENCES

[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] M. Benisch and T. Sandholm, "A framework for automated bundling and pricing using purchase data," Auctions, Market Mechanisms, and their ApplicationsAnonymous, 2012.

[3] 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.

[4] 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.

[5] C. W. Chu and 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.

[6] G. S. Crawford and A. Yurukoglu, "The welfare effects of bundling in multichannel television markets." The American Economic Review, 102(2), 2012, pp. 643-685.

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

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

[9] 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.

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

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2006, pp. 402.

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

[13] D. Honhon and X. Pan, "Improving retail profitability by bundling vertically differentiated products," Working paper, University of Florida, Gainesville, FL, 2015.

[14] 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.

[15] B. Letham, W. Sun, and A. Sheopuri, "Latent variable copula inference for bundle pricing from retail transaction data," Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 217-225.

[16] G. R. Liu and X. Z. Zhang, "Collaborative filtering based recommendation system for product bundling," Proceedings of the 2006 International Conference on Management Science and Engineering, 2006, pp. 251-254.

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

[18] I. S. Razo-Zapata, J. Gordijn, andP. D. Leenheer, and H. Akkermans, "Dynamic cluster-based service bundling: A value-oriented framework," Proceedings of the 2011 IEEE 13th Conference on Commerce and Enterprise Computing, 2011, pp. 96-103.

[19] D. Somefun and J. La Poutré, "Bundling and pricing for information brokerage: Customer satisfaction as a means to profit optimization," Proceedings of IEEE/WIC International Conference, 2003, pp. 182-189.

[20] 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, pp. 211-232.

[21] 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.

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

[23] 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.

 ISBN: 978-1-61208-558-6