J Inf Process Syst, Vol.21, No.3, pp.328~341, June 2025
https://doi.org/10.3745/JIPS.04.0352 ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Scalable Representation of Customer Purchase Preferences through Co-Purchase History

Seonghyun Kim[1] and Doyeon Kwak[2,*]

Abstract

In the competitive e-commerce landscape, accurately measuring customer preferences and effectively representing customer segments are essential for driving personalized marketing and product offerings. Current data-driven methods often rely on resource-intensive algorithms, and there is a need for a systematic and scalable framework for extracting product sets that represent specific purchasing preferences. This study proposes an unsupervised, efficient framework that leverages purchase history data to derive product sets that best represent known customer segments and product categories. Utilizing an item-based top-N recommendation technique, the proposed method tracks co-purchase histories and generates relevant novel segment variants, capturing hidden purchase preference attributes and delivering a more accurate depiction of customer behavior. Evaluation with real-world customer data from a Korean retail and e-commerce platform network substantiates the practical applicability of the suggested framework in forecasting the probability of purchasing target products, outperforming other prediction techniques. By adopting this scalable and readily implementable approach, businesses can effectively make well-informed decisions regarding product offerings, promotional campaigns, and personalized recommendations, ultimately improving customer engagement and sales.

Keywords

Co-purchase Recommendation, Customer Lifestyle, E-commerce, Predictive Analysis, Purchase Preferences, Retail, RFM

1. Introduction

Knowing and catering to customer preferences is a crucial aspect of running a successful business. Digitized retail platforms, with their ability to collect copious amounts of customer data, are particularly well-positioned for predicting customer intentions. Customer data, which typically includes basic demographic and psychographic information as well as behavioral historical information, is utilized for identifying and categorizing customers into preference segments for targeted marketing strategies [1]. Furthermore, with the ongoing advancements in mobile Internet, determining how to recommend personalized products rooted in user behavior holds significant research implications and practical value in e-commerce platform [2,3]. It can be observed that a user's past purchases considerably influence their future buying decisions [4]. As research focusing on understanding users through their context information continues to gain prominence, extracting user characteristics from purchasing behaviors emerges as both a crucial and viable approach [5]. Nevertheless, data-driven models have inherent limitations as they rely

Manuscript received October 12, 2023; first revision December 26, 2023; accepted December 31, 2023.

* **Corresponding Author:** Doyeon Kwak (doyeon.kwak@cj.net)
[1] Department of Industrial Engineering, Ulsan National Institute of Science and Technology, Ulsan, Korea (withgrsh@unist.ac.kr)
[2] CJ AI Center, Seoul, Korea (doyeon.kwak@cj.net)

on past data and may lack the flexibility to adapt to constantly changing user preferences, which are often subtly represented in the data.

Existing techniques, such as unsupervised clustering, while revealing new customer segments, often impose rigid quantitative boundaries based on static data, leading to potential biases and incomprehensible results [6]. Moreover, advanced methods like collaborative filtering and graph neural networks, despite their effectiveness, demand substantial computational resources and struggle with scalability, posing challenges for businesses in keeping these models updated and cost-effective [7,8].

To overcome these limitations and fully harness the potential of data while characterizing customer tendencies based on tangible business insights, we propose an unsupervised method for extracting product sets that best represent known and explainable customer and product segments. We refer to the predefined customer segments and product categories as seed-customer segments and seed-product categories, which in our case correspond to the age groups of our customers and the lifestyles suggested by the products.

This is an efficient implementation of item-based top-N recommendation, suggested by Deshpande and Karypis [9], combined with extracting the products that best represent the seed-customer segments. By leveraging the purchase history data, we can deduce the most-purchased products by the seed-customers—predefined by existing features—as well as, extract the product set relevant to the selected products based on co-purchase frequency. Then, we can regroup customers based on the number of purchases of products in the relevant set. This mechanism generates a new segment variant that captures the hidden purchase preference attributes of the seed-customer segment by sampling selected products as hub nodes and the branching products as relevant sets, interconnected by co-purchasing customers, in a virtual bipartite network of product-customers as [10] shown in Fig. 1.

This method can also be applied to define customer segments based on the purchase of seed-products of specific categories, by monitoring customers' purchase frequencies from relevant product sets derived from these seed-products. As a result, we can obtain product-level features that represent the customer segments, enabling a more accurate depiction of their behavior and preferences within the data.

To evaluate our framework in a real-world setting, we have used customer data and purchase history to predict the likelihood of purchasing a specific set of target products, leveraging the augmented segment features generated by our method. We have obtained our data sample from a Korean retail and e-commerce platform network, which includes companies from diverse industries such as online and offline beauty and cosmetics sales, online fashion malls, movie theaters, restaurants, and bakery chains. To validate the effectiveness of the attained purchase preferences, we have compared the performances of various prediction models to our own, demonstrating its practical utility in real-life scenarios.

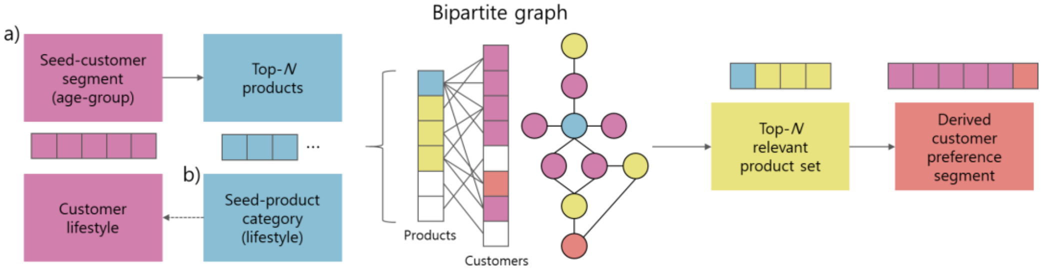

a)
Seed-customer segment (age-group)
Top-N products
Customer lifestyle
b)
Seed-product category (lifestyle)
...
Bipartite graph
Products
Customers
Top-N relevant product set
Derived customer preference segment

Fig. 1. Framework for extracting a relevant product set and the customer preference segment using purchase history data from (a) pre-defined seed-customer segment and (b) pre-defined seed-product category.

2. Related Work

2.1 Unsupervised Clustering in Customer Segmentation

Unsupervised clustering of customers into segments often generates quantitative boundaries among the customers based on rigid data. This technique may reveal previously unidentified customer segments, challenging assumptions about customer preferences and creating new market opportunities [11]. Techniques like density-based spatial clustering (DBSCAN) offer efficient methods for grouping spatial objects based on their density [12]. Another prominent method, k-means clustering minimizes the within-cluster variance relative to the mean centroid of the clusters [13]. Nevertheless, the inflexible segmentation boundaries, which are determined by the constraints of the source data as well as the selection of clustering algorithms and parameters, can lead to potential biases and incomprehensible results [11,13]. Particularly when many real-use-case datasets in e-commerce businesses are assumed to be constant and permanently-engraved, readily available demographic user data like age and gender, may not necessarily reflect one's evolving preferences.

2.2 Purchase-based Similarity Measures

Existing approaches in customer segmentation often employ advanced recommendation techniques that leverage purchase transaction data. One prominent method is collaborative filtering, which measures the similarity in purchasing patterns among customers to recommend relevant products [7,8]. This method typically involves analyzing large volumes of purchase data to identify patterns and preferences shared among different customer groups.

Another sophisticated technique is the use of graph neural network (GNN) based recommendation systems. These systems construct and train bipartite graphs of customer-product purchase relations, facilitating the calculation of the closeness of node pairs for an accurate similarity measurement. The strength of GNN lies in its ability to capture complex, nonlinear relationships within the data, offering a nuanced view of customer preferences.

However, both collaborative filtering and GNN-based systems come with their inherent challenges. A primary limitation of these methods is their resource intensity; they require significant computational power to maintain and update, leading to increased operational costs and scalability issues, especially for businesses with limited resources [14].

Despite the potential of data-driven customer segmentation in e-commerce, there is currently a lack of a systematic framework and generalized methods for effectively measuring customer preferences based on purchase history without resorting to resource-intensive algorithms. The ability to accurately select product sets that represent specific customer segments remains a challenge in real-life scenarios. This highlights the need for further research and development in creating more accessible and efficient tools for customer segmentation, particularly those that can adapt to the dynamic nature of consumer preferences in the digital marketplace.

2.3 Purchase Intentions Driven via Interpersonal Relationship

In the field of consumer behavior, the dynamic interplay between consumers and their relationships with brands, products, companies, and even other fans plays a pivotal role in shaping post-purchase intentions [15]. The enhancement of such intentions is intricately tied to the network of associations and interactions that consumers form in the marketplace. Existing literature explains how fostering relationships among individuals characterized by high levels of affinity and emotional resonance contributes significantly to reinforcing the identity and sense of belonging among members of a consumer segment [16]. This strengthening of inter-personal bonds and group coherence invariably acts as a catalyst in stimulating consumers' purchase intentions.

In our research, we aim to maintain a sense of shared identity and commonality among consumers by leveraging our novel features rooted in collaborative purchasing behaviors. These features are designed to uncover hidden purchase intentions, gauging them based on the degree of relevant purchases through the utilization of co-purchase networks derived from the comprehensive purchase history of customers.

3. Methods

3.1 Dataset Generation

In order to evaluate our framework in a real-world scenario, we have used customer data and purchase history from a Korean retail and e-commerce platform network. This network comprises companies from various industries, such as beauty and cosmetics, online fashion, movie theaters, restaurants, and bakeries.

Given the vast size and potential noise in the purchase-history data from the real-world retail and e-commerce platform network, we have employed several preprocessing techniques to obtain a representative sample. We focus on customers who have purchased our target products—online reservations for Christmas cakes—and their purchase history to extract co-purchased relevant products to derive customers with similar purchasing patterns. The observation duration of the purchase history is a 6-month period prior to a 2-month window for the target product reservations. We have also excluded irrelevant purchase data records, such as paper bags and shopping bags, which could skew the number of relevant product purchases. We included only customers who made between 5 and 80 purchases in the observation period to eliminate outliers. The final dataset includes 70,724 products with 102,920 purchase records from 95,946 customers, of which 6,974 had purchased the target product. The dataset predominantly consists of female customers (84.6%), attributed to retail services catering to this demographic, including beauty and cosmetics sales and fashion malls. The purchase frequency and total spending show a long-tail distribution, while the age distribution leans towards younger customers, with a median age of around 30.

3.2 Baseline Features

Our baseline model employs 24 features, comprising demographic information (gender and age), and RFM analysis (Recency, Frequency, Monetary), along with purchase counts by brand code. Recency is measured from using the last day of the 6-month data as a reference point and subtracting each customer's most recent purchase date, while frequency and monetary values are derived from purchase count and total spend in this period. Purchase counts are aggregated from 19 brands using unique brand codes for each customer within the observed 6 months.

3.3 Aggregated Feature Function and Complexity Analysis

To derive quantitative results from our framework for the selected products, we have developed a method for calculating the cumulative frequency of purchasing products from relevant sets for an individual customer.

$$f(S, c) = \sum_{i=1}^{n} freq_c \left(rank_N \left(R_{s_i} \right) \right), \tag{1}$$

where S is selected product set from a seed-customer segment (seed-product category); c is customer c; n is the total number of selected products in S; $freq_c$ is purchase frequency by customer c; $rank_N$ is top-N most purchased products by all customers; R_{s_i} is set of relevant products purchased by customers who purchased s_i; and s_i is the iteration of each selected seed-product from 1 to n.

For each of the n selected products, we extract a top-N co-purchased relevant product set and the total purchase frequency for all product in the relevant set by a single customer, c. By aggregating the purchase frequencies of relevant sets R_{s_i} for all selected seed-products in S, we can obtain a single relative value that measures the relationship between customer c and the selected seed-product group S. Using this approach, we create aggregated features by consolidating individual item-level results into age-group and lifestyle categories, facilitating a more comprehensive analysis.

According to Big O complexity analysis, the comparison between a generic GNN model and the algorithmic function $f(S, c)$ underlines key differences in scalability and computational efficiency. Despite the fact GNN based recommenders are utilized to reduce the complexity of collaborative filtering methods, their complexity is still large [17]. The GNN model's complexity, denoted as $O(E \cdot B \cdot (L \cdot (C + P) \cdot (d + F)))$, embodies the inherent complexities of graph-based methodologies. This complexity is influenced by the number of epochs E, batch processing B, total number of nodes (total count of customers and products, $(C + P)$, network layers L, d the average node degree, and feature vector size F. This can be further summarized by removing constants as $O(L \cdot (C + P) \cdot (d + F))$ [18]. This reflects substantial computational demands, particularly in large, complex graph scenarios. Such a complexity profile suggests scalability challenges, primarily due to the intricate interactions of these elements in extensive graph networks.

Conversely, the function $f(S, c)$, with a Big O complexity of $O(P \log P + n \cdot (P \log P))$, demonstrates a more linear and manageable scalability. It includes the initial sorting of all products $(P \log P)$ to extract the seed-product set S, followed by a logarithmic factor for ranking and extract the top-N relevant product set R iterating n-times through the selected seed-products in S. This equation can be derived as $O(P \log P)$ when removing the constant parameter $(1 + n)$. Our approach is particularly effective for large datasets, where the scalability is primarily influenced by customer interactions and transaction processing, offering predictable computational growth.

While GNNs are effective in detailing the complex data relationships, they may encounter scalability issues due to the graph's size and depth, and training complexities. In contrast, our method provides a scalable and efficient solution for extensive datasets with large customer bases and product assortments.

3.4 Preference Feature Generation via Aggregated Feature Function

Utilizing the age data, we divide our customers into five different age groups (20–30, 30–35, 35–40, 40–50, and 50–70) to be used as seed-customer segments. These segments were chosen based on observed lifestyle shifts, such as marriage and having children, typically occurring in the 30–40 age range. Additionally, the 50–70 range was grouped together due to a lack of samples.

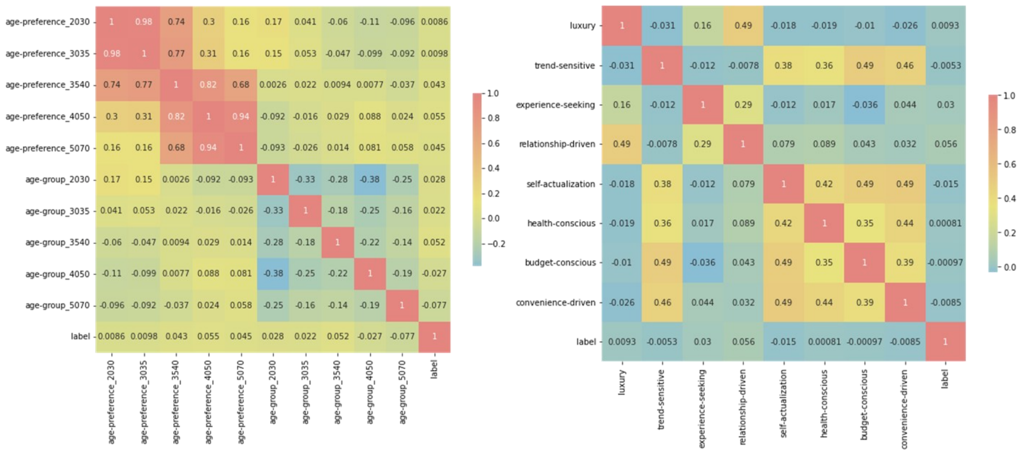
For each segment, we have identified representative products by extracting the top 200 products and excluding those popular in other age groups. This resulted in 75, 44, 37, 77, and 70 products for the five segments, respectively. According to Table 1, the total of selected 303 products to represent the five age-groups constitute 12.91% of the entire transaction frequencies and 6.90% of total monetary volume. We then define top-30 products most-frequently co-purchased within the 6-month by customers who bought the segment's representative product, limited to customers in the corresponding age group, as relevant product sets. The purchase count for each set in each age group is subsequently tallied. Since the numbers of representative products vary across age groups, we normalize them by dividing each purchase count by its respective group's representative product count. The resulting relevant product sets make up 38.86% of transaction frequencies and 23.98% of total monetary volume (Table 1).

Table 1. Comparison of monetary volume and transaction frequency ratios within each age-group segment and across the entire dataset: for selected products and for relevant product sets

Age group segment[a] (yr)	Selected products				Relevant product sets			
	Within segment		Total		Within segment		Total	
	Monetary volume	Trans freq.	Monetary volume	Trans freq.	Monetary volume	Trans freq.	Monetary volume	Trans freq.
20–30	0.0364	0.0532	0.0094	0.0175	0.1963	0.2596	0.0507	0.0854
30–35	0.0190	0.0262	0.0033	0.0055	0.1739	0.2282	0.0304	0.0474
35–40	0.0125	0.0234	0.0015	0.0031	0.1492	0.2576	0.0176	0.0340
40–50	0.0219	0.0566	0.0054	0.0121	0.1172	0.2689	0.0290	0.0574
50–70	0.0218	0.0488	0.0041	0.0053	0.1202	0.2597	0.0227	0.0280
All	-	-	0.0690	0.1291	-	-	0.2398	0.3886

[a] Seed-customer segments.

The resulting user-level features represent purchase preference scores reflecting each age group, indicating the individual's preferential position in terms of their purchasing behavior within their respective demographic. This approach allows for the categorization of individuals into age-preference segments, providing a more nuanced understanding of their purchasing tendencies.



(a)

(b)

Fig. 2. The correlation matrix (a) between existing age-group and aggregated age-preference segments with the target value and (b) among aggregated lifestyle-preference segments with the target value.

As demonstrated in Fig. 2(a), we have compared the correlation among age-preference segments and age-groups, as well as the target label. Our analysis reveals that age-preference segments exhibit strong correlations when derived from proximate age-groups, and weaker correlations with more distant age-groups. Importantly, there was no substantial correlation with the target label, as the target should not be inherently highly correlated to a specific preference. When examining the correlation between age-preference segments and age-groups, a relatively modest, but visible correlation is observed for corresponding age-group and age-preference pairs. This outcome can be attributed to our derivation framework, which relies on co-purchase similarity to generate significantly diverse variant features. These unique feature distinctions contribute to improved model prediction capabilities.

We have conducted an evaluation of the 300 most commonly purchased products in order to pinpoint items that represent particular lifestyle segments. From this pool, we meticulously selected 83 seed-products that effectively encapsulate the characteristics of eight distinct lifestyles. Below are the categories and representative seed-products, with the quantity selected in parentheses:

- Luxury-goods (2): This lifestyle is characterized by indulging in high-end experiences, such as IMAX movies and upscale dining at family restaurants.
- Trend-sensitive (16): This category focuses on individuals who embrace fashionable products, such as color cosmetics and men's cosmetics. Notably, men's basic cosmetics are classified as trend-sensitive, due to social norms.
- Experience-seeking (5): This lifestyle places emphasis on engaging in outdoor activities, including cafe visits, to enrich one's experiences.
- Relationship-driven (13): This category encompasses product groups that suggest purchases made for a broader social group or imply an individual's inclusion within a family unit, such as baby probiotics supplements.
- Self-actualization (13): This lifestyle entails prioritizing essential items like basic cosmetics, which contribute to an individual's personal growth and fulfillment.
- Health-conscious (19): This lifestyle is characterized by emphasis on maintaining a healthy lifestyle, opting for products like diet supplements and organic snacks.
- Budget-conscious (15): This lifestyle involves seeking cost-effective options, such as purchasing items in bulk, taking advantage of promotional offers and discounts.
- Convenience-driven (3): This category emphasizes achieving convenience through products like cleansing tissues and remover pads, which simplify daily routines.

We then have established a collection of 30 relevant products, frequently co-purchased with each seed-product, within a 6-month observation window. The initial top-300 products are deemed too generic and lacks the specificity needed to represent lifestyle traits, hence, are excluded from the relevant sets. We then calculate the cumulative purchase frequencies of the products within each set to create lifestyle-specific features. The total of 83 seed-products consist of 3.53% of total monetary spending and 5.35% of entire transaction frequencies, while the relevant product sets show 16.96% of total monetary spending and 24.33% of entire transaction frequencies, as shown in Table 2.

Table 2. The ratio of total monetary value and transaction frequency of the seed-products for each lifestyle category, and the ratio of total monetary value and transaction frequency of the co-purchased relevant product sets derived from seed-products for each lifestyle category

	Selected products			
	Monetary volume	**Trans freq.**	**Monetary volume**	**Trans freq.**
Lifestyle categories[a]	0.0353	0.0535	0.1696	0.2433

[a] Seed-product categories.

Next, we grouped these set-level features into categories that shared common lifestyle characteristics. We then aggregated the set features for each lifestyle category and normalized these features by dividing them by the total number of features within each category, akin to age group normalization. The final outcome yielded lifestyle preference scores for each individual based on their purchase patterns.

The correlation analysis of aggregated lifestyle-preference segments, as illustrated in Fig. 2(b), reveals a distinct separation between two broad lifestyle groups: luxury-goods, experience-seeking, and relationship-driven, versus rest of the lifestyles. This discrepancy can be attributed to customers indulging in various lifestyles rather than being confined to a single specific lifestyle. Our selection of products in the relationship-driven lifestyle cater to strong familial relationships, which is often associated with customers who possess a better financial background to accommodate luxury goods and new experiences. In contrast, the second group primarily focuses on individual-level purchases, often targeting younger customers without direct ties to family-level spending.

3.5 Experiments

To evaluate the performance improvements incorporating the preference features based on seed-customer segments and seed-product categories, we conduct nine experimental cases as follows. We have compared the effectiveness of our feature addition with customer demographic and RFM behavioral features, as well as with features derived from unsupervised clustering techniques. Furthermore, we have carried out prediction experiments employing aggregated preference segment-level features. These features were derived by aggregating the purchase frequencies of relevant product sets, allowing us to access and validate customers' general purchasing patterns.

The experimental cases are:

- Baseline features (gender, age, RFM, brand code)
- Baseline + cluster features
- Baseline + age-preference features (based on age group seed-customer segments)
- Baseline + lifestyle-preference features (based on lifestyle seed-product categories)
- Baseline + age-preference + lifestyle-preference
- Baseline + lifestyle-preference + lifestyle-preference + clustering
- Baseline + aggregated age-preference segments
- Baseline + aggregated lifestyle-preference segments
- Baseline + aggregated age-preference segments + aggregated lifestyle-preference segments.

Table 3. Various baseline models compared using the SMOTE oversampled dataset

Model	Accuracy	AUC	Recall	Precision	F1-score	Kappa	MCC
XGBoost	0.9617	0.9793	0.9257	0.9981	0.9605	0.9234	0.9258
LightGBM	0.9573	0.9782	0.9169	0.9982	0.9558	0.9146	0.9177
Extra Tree	0.9475	0.9835	0.9346	0.9601	0.9472	0.895	0.8953
Random Forest	0.9469	0.9803	0.9145	0.9788	0.9456	0.8939	0.8959
Decision Tree	0.9122	0.9122	0.9176	0.9091	0.9133	0.8244	0.8245

The train dataset is sampled to have approximately 1:1 ratio for the target label, which then is split into train and validation set, for training models and comparing the model performances.

We have selected XGBoost classifier, which demonstrated strongest performances in the baseline experiments in terms of accuracy at 0.9617 and Matthews correlation coefficient (MCC) at 0.9258, as shown in Table 3. When training models, the dataset's imbalance can be detrimental. Given that the target represented only approximately 7% of the entire dataset and was highly imbalanced, we performed stratified splitting of the dataset and applied an oversampling technique (SMOTE) on the training set to acquire approximately 1:1 ratio of the binary target label. Afterwards, we have split the sampled data into a train and a validation set for model selection test. Note that the model performance results in Table 3 are that of the sampled train data for the final prediction model. The final prediction model utilizes the full, but unbalanced dataset, which should result in lower performance metrics.

In the experimental cases involving baseline and cluster features, we employ the k-means clustering method as an unsupervised segmentation technique. We have used the purchase frequencies of the top 200 most-purchased products to cluster customers into various segments. The optimal number of clusters (k) was determined to be 3, with a silhouette score of 0.4492 as shown in Table 4. However, using only a single customer-product relation resulted in a large, bulky cluster that was difficult to subdivide into smaller segments, as the clustering algorithm struggled to identify distinct characteristics.

Table 4. Silhouette scores for the number of clusters (k) ranging from 2 to 6

	k = 2	**k = 3**	**k = 4**	**k = 5**	**k = 6**
Silhouette score	0.4359	**0.4492**	0.3187	0.3127	0.3200

Boldface values denote the highest silhouette score achieved across the evaluated cluster configurations.

4. Results

Binary classification is conducted to predict which customers have made purchases of the specified target products. Contrary to the model validation, which is also used for model selection, the full unbalanced test dataset is utilized in the final analysis to accurately assess the prediction performance of each experimental case. Consequently, the performance metrics may appear to be lower, but this does not necessarily indicate inferior performances on the validation metrics.

Customer segmentation based on clustering has demonstrated an improvement in purchase prediction performance from the baseline model, but the degree of improvement was not high. In terms of the AUC metric, the XGBoost baseline model has scored 0.7006, while the additional cluster features only have improved the model to 0.7090. However, features based on seed-customer segments (age-preference features) and seed-product categories (lifestyle-preference features) show a higher performance boost compared to the baseline model. Notably, a significant increase in performance can be observed, with AUC values of 0.7006 for the baseline-only case, increasing to 0.9509 for age-preference features, 0.9000 for lifestyle-preference features, and an even higher performance of 0.9604 when both preference feature sets are used together. This demonstrates that our features exhibit high performance compared to the k-means method.

Decision-tree-based model's performance often improve with each added feature, as they enable more refined branching points. To test the effectiveness of our framework without the excessive feature amplification, we also have evaluated cases with aggregated preference features consisting of age-preference segments with the same number of segments as age-groups and aggregated features of lifestyle-preference segments with the same number of lifestyle product categories. Although the final XGBoost model performances of aggregated features decreased to AUC of 0.8813 for age-preference and 0.8730 for lifestyle-preference, these results were still much higher than that of the baseline model and the case using additional cluster features. When the aggregated feature sets are used together, the

Table 5. The model performances of the experimental cases for XGBoost classifier using the complete test dataset

Model	Accuracy	AUC	Recall	Precision	F1-score	Kappa	MCC
Baseline (BL)	0.9319	0.7006	0.0093	0.3939	0.0182	0.0151	0.0520
BL + Cluster (C)	0.9318	0.7090	0.0086	0.3636	0.0168	0.0137	0.0472
BL + Age (A)	0.9520	0.9509	0.4573	0.7342	0.5636	0.5397	0.5566
BL + Lifestyle (L)	0.9413	0.9000	0.2595	0.6729	0.3745	0.3500	0.3944
BL + A + L	0.9549	0.9604	0.4667	0.7787	0.5836	0.5613	0.5820
BL + A + L + C	0.9555	0.9619	0.4832	0.7765	0.5957	0.5735	0.5916
BL + Aggr. Age (AA)	0.9389	0.8813	0.2129	0.6499	0.3207	0.2972	0.3490
BL + Aggr. L (AL)	0.9366	0.8730	0.2108	0.5892	0.3105	0.2849	0.3270
BL + AA + AL	0.9427	0.9049	0.3011	0.6731	0.416	0.3905	0.4258

AUC metric improves to 0.9049. The entirety of the model performances is shown in Table 5.

The SHAP feature importance analysis for the XGBoost model with aggregated age-preference and lifestyle-preference features show our generated features exhibit high decision power when predicting the likelihood of purchasing target products, compared to many of the baseline features (Fig. 3).

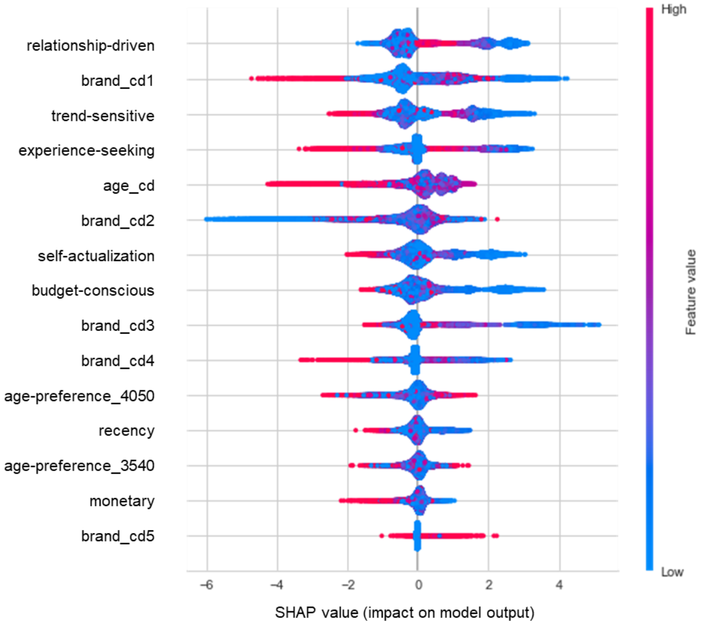


Fig. 3. Top-15 SHAP feature importance plot for the experimental case of the baseline model with aggregated age-preference and aggregated lifestyle-preference features.

To further test the generality of the proposed method, we have evaluated the performance improvements due to the aggregate features across various targets. For this, five products with over 1,000 purchases in the 2-month prediction window, which were not included in the aggregate feature generation, are randomly selected as targets. The performance of XGBoost and Extra Tree models using k-means cluster features was compared against the models employing our aggregate features. An average of the metric performances is presented in Table 6.

Each new target was specific to a single product, in contrast to previous targets encompassing a range of items. This specificity required the models to precisely predict purchases for each product, leading to highly imbalanced data. To address this, the SMOTE sampling method was employed during training. The finalized models were then applied to actual test data. The results indicated lower performance metrics for recall and precision due to the imbalanced nature of the data, predicting a single product purchases. However, models using aggregate features demonstrated significantly better performance in terms of AUC, F1-score, Kappa, and MCC, key indicators of overall model efficacy. In comparison, baseline models with k-means cluster features showed negligible predictive power. Despite challenging conditions of extreme imbalance, models with the aggregate features exhibited notable improvements.

Table 6. Comparison of models using conventional k-means clustering features (C) to our aggregate features (AA, AL) for predicting five randomly selected targets

Model	Accuracy	AUC	Recall	Precision	F1-score	Kappa	MCC
XGBoost BL + C	0.9789	0.5001	0.0005	0.0294	0.0009	0.0003	0.0011
Extra Tree BL + C	0.9605	0.5067	0.0333	0.0342	0.0337	0.0135	0.0135
XGBoost BL + AA + AL	0.9791	0.5380	0.0778	0.4840	0.1340	0.1290	0.1879
Extra Tree BL + AA + AL	0.9607	0.6033	0.2304	0.1701	0.1957	0.1761	0.1782

The performance metrics for each of the five targets are averaged to provide a comprehensive evaluation for each model.

5. Discussion

In the context of real-life scenarios and data, determining preference segments based on existing demographic segments and product categories through purchase history data reveals much more information regarding the purchase intent. By analyzing customer purchasing behavior, we have generated alternative age-preference segments that represent the core preferences of different age demographics and extended lifestyle-preference segments, encompassing co-purchased product sets that elucidate these lifestyles. Each preference designation significantly enhances conventional models, both individually and in combination, yielding more refined performances. Given the large number of generated features, the preference features can be aggregated based on their preference segments to reduce the resource cost of operating the model, while providing generalized measures of customer's segment involvement. This approach still demonstrates superior performance compared to conventional clustering techniques.

Upon analyzing the feature importance as illustrated in Fig. 3, we observe distinct positive impacts for the 40–50 and 35–40 age-preference segments. A closer examination of the relationship between the age-preference features and the lifestyle-preference features in Fig. 4 reveals that the 40–50 age-preference is closely aligned with the relationship-driven lifestyle-preference segment. This lifestyle segment demonstrates a propensity for purchasing the target products, Christmas cake reservations intended for family celebrations. The 30–40 age-preference segment exhibits positive correlations with all lifestyle-preferences, suggesting that it represents the primary customer segment with generally high purchasing power, also likely to purchase target products.

In terms of lifestyle-preference segments, the aggregated relationship-driven, trend-sensitive and experience-seeking segments have a significant impact on the prediction model. The relationship-driven segment is more inclined to be part of a family with young children, and it is prone to engage in family-oriented activities such as Christmas celebrations. This results in an increased likelihood of reserving target products.

The relationship-driven segment also corresponds well with the experience-seeking segment, which takes account for various restaurant dining, cafe and bakery purchases, and is likely to lead to higher likelihood of cake reservation. These lifestyle-preferences tend to correspond to generally older age-preference groups with higher purchasing power (Fig. 4). On the other hand, trend-sensitive, self-actualization and budget-conscious are closely related and show negative impact on the purchasing of the target products, which generally correspond to younger audiences. The relationship-driven segment also aligns well with the experience-seeking segment, which accounts for individuals with various restaurant dining experiences, cafe-related purchases, and bakery purchases, ultimately leading to a higher probability of the cake reservation. These lifestyle preferences are generally associated with older age demographic groups, as illustrated in Fig. 4, and appear to represent a higher purchasing power.

Conversely, the trend-sensitive group exhibits a strong correlation with health-conscious, budget-driven, self-actualizing, and convenience-seeking individuals, who demonstrate a lower likelihood of purchasing the target products, specifically Christmas cake reservations. This incongruity with their purchasing habits can be attributed to the misalignment with the lifestyle preferences of these consumer segments. As illustrated in Fig. 3, a discernible decrease in the propensity to purchase the target products is observed among individuals with higher self-actualization and budget-conscious inclinations.

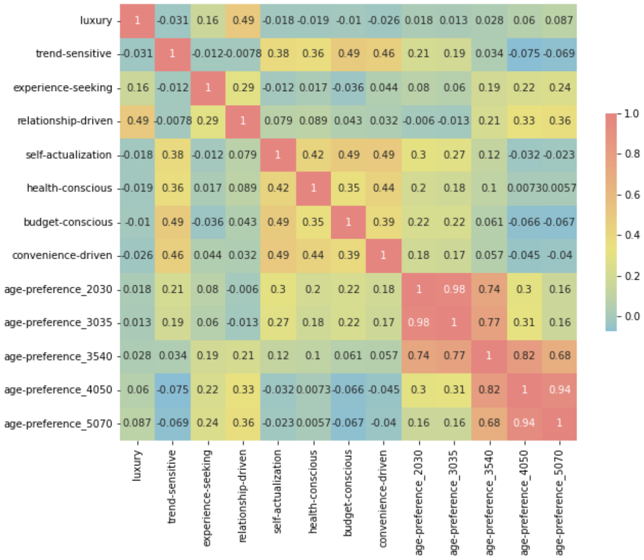


Fig. 4. Correlation matrix between the aggregated age-preference features and the aggregated lifestyle-preference features.

6. Conclusion

In our proposed framework, we implement the extraction of relevant product sets through item-based top-N recommendations, which are based on the co-purchases by predefined customer segments and of predefined product categories. This method has proven effective in extracting preference information and predicting purchase likelihood within the context of a real business scenario. We further generated alternative preference segmentation revealing the core preferences of age demographics and expanded product category data into lifestyle-preference segments, which are proven to be effective both independently and in combination.

However, as our dataset only allows for validation within a single case of target product purchasing, we are prevented from effectively validating through multi-year horizons. Additionally, our chosen comparison method, k-means clustering, was confined to the top-300 most purchased products, and generated a mere three distinct clusters. Consequently, this method fails to capture the same level of information as the age-preference and lifestyle-preference features, which more effectively retain purchase preference information.

Other limitations include the arbitrary determination of the number of seed-products, which is set at 303 for age-preferences and 83 for lifestyle-preferences in our study. The number of products within relevant product sets, 30 in our case, may also vary according to specific business contexts. Furthermore, the seed-product selection for the lifestyle category necessitated a certain level of data understanding.

Despite these limitations, our framework and experimental results demonstrate a resource-efficient alternative for extracting purchase-preference information in retail businesses. This study contributes to the ongoing development of preference-based recommendation techniques, minimizing the calculations necessary for extracting the relevant product sets using top-N method. Future research could concentrate on broadening the application of this framework across various retail contexts over multiple cycles, effective techniques for optimizing parameters to enable more granular segmentation. Moreover, incorporating additional demographic and sociographic variables could augment the predictive power of the proposed approach. Our findings lay the groundwork for further exploration and development of innovative strategies to distill purchasing behaviors into known customer attributes.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

None.

Acknowledgements

This paper is the extended version of "A Scalable Framework Leveraging Co-Purchase History for Optimal Representation of Core Purchase Preferences in Customer Segments," originally presented at the International Conference on Big Data, IoT, and Cloud Computing (BIC 2023).

References

[1] C. F. Lin, "Segmenting customer brand preference: demographic or psychographic," *Journal of Product & Brand Management*, vol. 11, no. 4, pp. 249-268, 2002. https://doi.org/10.1108/10610420210435443

[2] J. Xu, Z. Hu, and J. Zou, "Personalized product recommendation method for analyzing user behavior using DeepFM," *Journal of Information Processing Systems*, vol. 17, no. 2, pp. 369-384, 2021. https://doi.org/10.3745/JIPS.01.0069

[3] F. Bao, W. Xu, Y. Feng, and C. Xu, "A topic-rank recommendation model based on Microblog topic relevance & user preference analysis," *Human-centric Computing and Information Sciences*, vol. 12, article no. 10, 2022. https://doi.org/10.22967/HCIS.2022.12.010

[4] D. Y. Hong, G. Y. Kim, and H. H. Kim, "Deep learning-based personalized recommendation using customer behavior and purchase history in e-commerce," *KIPS Transactions on Software and Data Engineering*, vol. 11, no. 6, pp. 237-244, 2022. https://doi.org/10.3745/KTSDE.2022.11.6.237

[5] F. P. Guo and Q. B. Lu, “Contextual collaborative filtering recommendation model integrated with drift characteristics of user interest,” *Human-centric Computing and Information Sciences*, vol. 11, article no. 8, 2021. https://doi.org/10.22967/HCIS.2021.11.007

[6] C. Jin and E. C. Malthouse, “On the bias and inconsistency of k-means clustering,” 2016 [Online]. Available: https://www.researchgate.net/publication/287829457_On_the_bias_and_inconsistency_of_K-means_clustering

[7] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, “Collaborative filtering recommender systems,” *Foundations and Trends in Human–Computer Interaction*, vol. 4, no. 2, pp. 81-173, 2011. http://dx.doi.org/10.1561/1100000009

[8] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, “Graph neural networks for social recommendation,” in *Proceedings of the World Wide Web Conference*, San Francisco, CA, USA, 2019, pp. 417-426. https://doi.org/10.1145/3308558.3313488

[9] M. Deshpande and G. Karypis, “Item-based top-n recommendation algorithms,” *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 143-177, 2004. https://doi.org/10.1145/963770.963776

[10] K. Kafkas, Z. N. Perdahci, and M. N. Aydın, “Discovering customer purchase patterns in product communities: an empirical study on co-purchase behavior in an online marketplace,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 7, pp. 2965-2980, 2021. https://doi.org/10.3390/jtaer16070162

[11] P. Anitha and M. M. Patil, “RFM model for customer purchase behavior using K-Means algorithm,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 5, pp. 1785-1792, 2022. https://doi.org/10.1016/j.jksuci.2019.12.011

[12] H. J. Jang and B. Kim, “KM-DBSCAN: density-based clustering of massive spatial data with keywords,” *Human-centric Computing and Information Sciences*, vol. 11, article no. 43, 2021. https://doi.org/10.22967/HCIS.2021.11.043

[13] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: a comprehensive survey and performance evaluation,” *Electronics*, vol. 9, no. 8, article no. 1295, 2020. https://doi.org/10.3390/electronics9081295

[14] S. Wang, Y. Zheng, and X. Jia, “SecGNN: privacy-preserving graph neural network training and inference as a cloud service,” *IEEE Transactions on Services Computing*, vol. 16, no. 4, pp. 2923-2938, 2023. https://doi.org/10.1109/TSC.2023.3241615

[15] C. W. Ho and Y. B. Wang, “Re-purchase intentions and virtual customer relationships on social media brand community,” *Human-centric Computing and Information Sciences*, vol. 5, article no. 18, 2015. https://doi.org/10.1186/s13673-015-0038-x

[16] Y. Liu, X. Luo, and Y. Cao, “Investigating the influence of online interpersonal interaction on purchase intention based on stimulus-organism-reaction model,” *Human-centric Computing and Information Sciences*, vol. 8, article no. 37, 2018. https://doi.org/10.1186/s13673-018-0159-0

[17] J. Wang, H. Mei, K. Li, X. Zhang, and X. Chen, “Collaborative filtering model of graph neural network based on random walk,” *Applied Sciences*, vol. 13, no. 3, article no. 1786, 2023. https://doi.org/10.3390/app13031786

[18] G. Li, M. Muller, B. Ghanem, and V. Koltun, “Training graph neural networks with 1000 layers,” *Proceedings of Machine Learning Research*, vol. 139, pp. 6437-6449, 2021.

Seonghyun Kim **https://orcid.org/0009-0007-5698-3586**

She is currently an undergraduate in Industrial Engineering at Ulsan National Institute of Science and Technology (UNIST). She is expected to receive her B.S. degree in early 2024. She is interested in data analytics and data science.

Doyeon Kwak **https://orcid.org/0000-0001-6693-2714**

He has received his B.S. degree in Biochemistry at University of California, Los Angeles (UCLA) in 2008. He has completed his M.S. degree in Bio-engineering in 2011, M.S. and Ph.D. degrees in Culture Technology in 2013 and 2018, respectively, at Korea Advanced Institute of Science and Technology (KAIST), Korea. Currently he is working at CJ AI Center, Korea as a researcher.