Innovations in Systems and Software Engineering
https://doi.org/10.1007/s11334-020-00372-5

ORIGINAL ARTICLE

Product recommendation for e-commerce business by applying principal component analysis (PCA) and *K*-means clustering: benefit for the society

Soma Bandyopadhyay[1] · **S. S. Thakur**[1] · **J. K. Mandal**[2]

Received: 20 February 2020 / Accepted: 10 August 2020

Abstract
Recommender system is a computer-based intelligent technique which facilitates the customers to fulfill their purchase requirements. In addition to this, it also helps retailers to manage the supply chain of their business and to develop different business strategies keeping in pace with the current market. Supply chain management (SCM) involves the streamlining of a business's supply-side activities to remain competitive in the business landscape. Maximizing the customer value is another important activity of SCM to gain an advantage in the market. In this work, the *K*-Means clustering algorithm has been used for the effective segmentation of customers who have bought apparel items. PCA has been used for dimensionality reduction of different features of products and customers. The main focus of this work is to determine the different possible associations of customers in terms of brand, product, and price from their purchase habits. The result shows that the clusters made by the algorithm based on PCA and *K*-Means are similar and the results are acceptable on the basis of feedback received from existing customers and satisfies the customers' requirements based on the amount of money (price range) the customers want to spend while doing online shopping. The features of products purchased by customers were combined together to generate a unique product key for business, and a model was prepared to segment products based on the volume of products sold and revenue generated, and the price of products sold and revenue generated. This work, in the long run, will help business houses to build a sustainable, profitable, and scalable e-commerce business. Environmental, social, and economic aspects are important to make e-commerce more sustainable for the benefit of the society.

Keywords E-commerce · Principal component analysis · *K*-Means clustering · Supply chain management · Stock keeping unit · Recommendation systems

1 Introduction

As every customer is having individual preferences, collecting datasets is the most important and challenging task. While giving recommendations about any product, the loyalty of the customer needs to be taken into consideration. In the last decade, recommendation systems have become an integral part of e-commerce business to promote product sales and thus become a popular research field in the present era. E-commerce business is electronic transactions over the internet which is focusing on products as well as services. The sustainability of an e-commerce business depends on the financial progress of the enterprises. It has been observed that though the e-commerce businesses have enhanced their performance in short term through investment and financing, there are some market risks involved which may even lead to bankruptcy. The prime motto of online business is increasing sales and maximizing profit. To maximize sale, one method employed by an online business is called the conversion of browsers into customers. It is a process wherein a customer actually purchases or does some action apart from a simple visit. To increase the conversion rate of browsers into customers, some strategies are employed by the website. In

✉ S. S. Thakur
subroto_thakur@yahoo.com

Soma Bandyopadhyay
somabanmuk@yahoo.co.in

J. K. Mandal
jkm.cse@gmail.com

[1] MCKV Institute of Engineering, Howrah, West Bengal, India

[2] University of Kalyani, Nadia, West Bengal, India

Published online: 25 August 2020

particular, page visits, purchase history, duration of the visit, conversion rate, and number of orders made by the customers are analyzed for the purpose of recommendation. Customers' buying habits can be analyzed by the relationships of different products they put into their shopping basket while shopping. This can help retailers to know which products consumers most frequently purchase so that they can develop a better marketing strategy. It has been observed that a certain category of customers has inclination toward a particular brand. The recommendation system plays an important role by generating suggestions about products to the users. It can also suggest a particular brand of product according to the customers' requirements. It takes the feedback of the customers about the item, which a user has already purchased earlier. According to the input provided by the user, recommendation systems apply certain algorithms to generate users' ratings on a particular product/item. In general, the recommendation system suffers from sparsity and scalability problem which reduces the quality of predictions. Another major challenge in present e-commerce business is the categorization of the products precisely and efficiently. Clustering the products into different groups or effectively reducing the dimensionality of data by applying PCA can solve these problems. By applying PCA, the dimension reduction is done to a large extent and hence algorithmic complexity is reduced. In this work, two unsupervised machine learning techniques PCA and *K*-Means algorithm have been applied to cluster the customers who purchase apparel online. *K*-Means clustering has been applied to segment the customer using the monthly income of the customer and the price of the item which they have purchased. This model can recommend apparel products to customers using an online recommender system. Using PCA, all the information contents of the particular user have been summarized as the principal components and the clustering of customer have been done based on that. By comparing these two results, it has been observed that similar clusters were formed. In addition to that, a segmentation model was prepared to cluster the items on the basis of the stock keeping unit (SKU). The model can run through the sales volume of the item and revenue generated, and the sale price of the item and revenue generated. This model will help the retailer in identifying the maximum selling item, the most revenue generated item, or the item which can make loss of business. Furthermore, this model will help the retailer in identifying which item needs more marketing focus. SKU is a unique numerical identifying number that is used to identify the product, product size or type, and the manufacturer in the retail industry. It is a part of a backend inventory control system that can help the retailer to get low-inventory alerts and restock items that are forecasted by the model. Though we have collected the data from the customers only, the model was prepared to recommend the retailer also. We have prepared the stock keeping unit according to the product, price, and brand information. So, this model can be extended to all categories of customers as well as retailers too.

The subsequent sections of this paper are organized as follows. In Sect. 2, the related research work has been discussed. Our proposed model and details of implementation have been described in Sect. 3. We have discussed about the dataset in Sect. 4. In Sect. 5, result analysis has been done. In Sect. 6, the conclusion and future work have been summarized.

2 Related research work

A genetic algorithm (GA)-based approach was proposed to optimize the *K*-Means clustering algorithm for the segmentation of the online shopping market [1]. A centering bunching clustering algorithm was proposed which shows better performance than *K*-Means or *K*-medoids algorithm on the Iris dataset [2]. It has been observed that collaborative filtering suffers with the problem of scalability, sparsity, and cold start problems. The problem of scalability and sparsity have been solved by developing a personalized recommendation system which uses user cluster and item cluster based collaborative filtering approach [3]. A personalized recommendation system based on Recency, Frequency, and Monetary (RFM) was proposed using *K*-Means clustering of item category under pervasive computing environment which is essential by real-time accessibility [4]. At present, it is observed that the e-commerce business is very competitive. Hence, it is equally important to retain existing customers as well as to add new customers according to the demand of the market [5]. Clustering has been broadly used as an unsupervised learning technique, which makes a great influence on researchers in the field of computer science [6, 7]. Still, there is a need to explore the application of the clustering technique in the recommendation system. It is important to partition choices into smaller sets for the future choice of a neighbor [7]. The categorization of buying patterns of different customers using the *K*-Means clustering algorithm helped a lot for future prediction [8]. *K*-Means clustering algorithm uses Euclidean distance, where the computation of distance is done by finding the square of the distance between each data point and then summing the squares to get the desired result [9]. E-commerce products from websites were classified using k-NN models. Before classification, hierarchical and *K*-Means algorithms were applied in the dataset. The second (classic in iterative optimization) version of *K*-Means iterative optimization reassigns points based on a more detailed analysis of effects on the objective function caused by moving a point from its current cluster to a potentially new one [8]. If a move has a positive effect, the point is relocated and the two centroids are recomputed. It

is not clear whether this version is computationally feasible because the outlined analysis requires an inner loop over all member points of involved clusters affected by centroids shifts. However, in case it is known that all computations can be algebraically reduced to simply computing a single distance and in this case, both versions have the same computational complexity. It was observed that the computation time is less when the *K*-Means clustering algorithm was applied before applying the k-NN classification model [10]. It has been found that high-dimensional data are transformed into lower-dimensional data via the principal component analysis (PCA) [11] which is an unsupervised dimension reduction used for information retrieval purposes. To convert higher dimensional data to lower dimension subspace, PCA is used; on the other hand, the *K*-Means algorithm is then applied on the subspace [12]. By the principal component analysis, dimension reduction is performed which automatically clusters the data [13]. In this work, to infer the relationship between Kansei words and sunglass specimens, PCA was applied [14]. The Expectation Maximization (EM) algorithm, an unsupervised machine learning technique is used to cluster the data, and the Adaptive Neuro-Fuzzy Inference System (ANFIS) is used for designing the prediction model [15]. An unsupervised method has been designed to detect shilling groups in a real e-commerce platform (e.g., Amazon China). In this work, frequent pattern mining is applied to generate candidate groups. An unsupervised ranking method based on principal component analysis (PCA) is then employed to capture noticeable outliers [16]. Marketing strategies were proposed based on ontology ideology. When the samples are sparse, there is a problem in the case of the *k*-nearest neighbor algorithm. These shortcomings may be overcome by adopting fuzzy classification, through importing membership function [17].

3 Proposed model and implementation

In the present scenario, it has been observed that e-commerce product classification and recommendation of products to the customers are tough and challenging tasks, and its dominance in the research area is increasing manifold. In this work, the database of different products/items purchased by customers is taken into consideration. The block diagram of the proposed model for recommendation aimed at customers is shown in Fig. 1.

The block diagram of the proposed model for the recommendation of the retailer is shown in Fig. 2. In this proposed model, a unique product–brand–price was modeled as stock keeping unit. Product frequency count of a unique product–brand–price has been modeled as sales volume. Product price of a unique product–brand has been modeled as the sale price. Then, sales volume was multiplied by sale price to compute revenue generation.

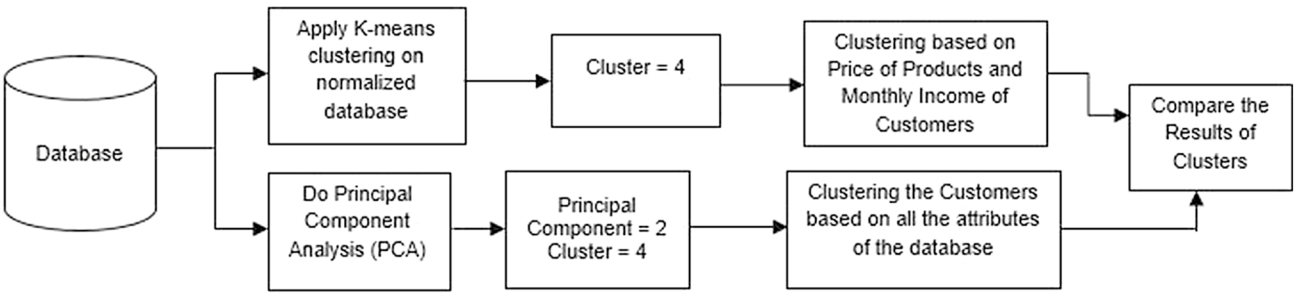

Fig. 1 Block diagram of proposed model for recommendation of customer

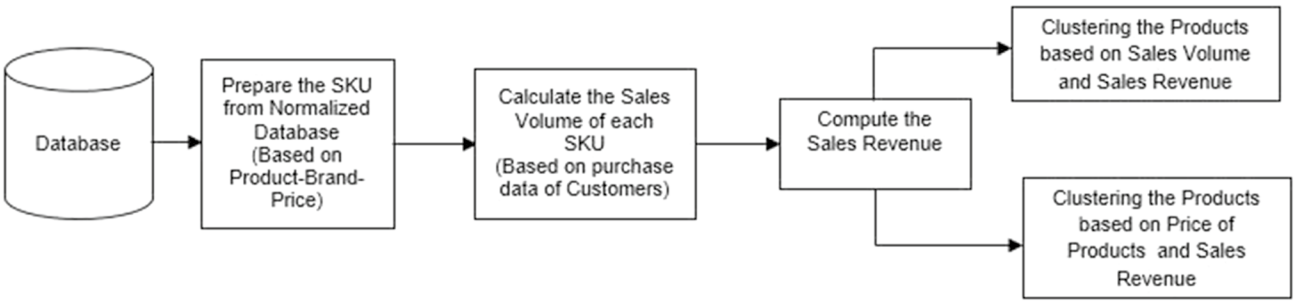

Fig. 2 Block diagram of proposed model for recommendation of retailer

During the festive season, customers normally buy different types of apparel products from various brands and the buying pattern of a person is different from that of other periods of the year and the choice of products and brand varies from person to person. Customers choose different brands as status symbols, some customers trust the quality, and others may choose products based on the prices offered. If the retailer can categorize the products, it will help them to fix their marketing strategy.

3.1 *K*-Means clustering algorithm

K-Means algorithm is an unsupervised learning algorithm, which is used to classify/group objects based on their attributes/features where K is a positive integer. This grouping is formed by minimizing the sum of squares of distances, between data and the corresponding cluster centroid. This method uses K clusters which are fixed at the beginning, and it is used to classify a given dataset [18].

Given a dataset of n data points $x_1, x_2, \ldots, x_n$ such that each data point is in $\mathbf{R^d}$, the problem of finding the minimum variance of the dataset into K clusters is that of finding K points $\{m_j\}$ $(j = 1, 2, \ldots, k)$ in $\mathbf{R^d}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}[\min_{f} d^2(x_i, m_j)] \tag{1}$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j. The points $\{m_j\}$ $(j = 1, 2, \ldots, K)$ are known as cluster centroids. The problem in Eq. (1) is to find K cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized.

The basic step of *K*-Means clustering is discussed below:

Step 1 Determine the number of clusters K at the beginning.

Step 2 Determine the coordinates of the centroids/centers of these clusters. Any random data point can be taken as the initial centroid or first K data points can also be used as the initial centroids.

Step 3 Determine the distance of each data point to the centroids.

Step 4 Group the object based on minimum distance (find the closest centroid). Each record is assigned to the nearest cluster using a measure of distance (Euclidean distance).

Step 5 Keeping the same number of clusters, the new centroid of each cluster is calculated, by taking the arithmetic mean of all the data points which belong to that cluster.

Step 6 Repeat step (2) to step (5) until the clusters stop changing or until the data points stay in the same cluster.

The major work of this algorithm is customer segmentation and having knowledge about customers' needs. This will play an important role in building a successful venture. Customer segmentation is the process where the segregation of data is performed based on different interests of potential customers. The whole process involves various actions that help to predict future actions. We can get a huge amount of unstructured data; so once these data are provided to the *K*-Means algorithm, it effectively analyses and integrates the data to give a structured appearance, and if the user requests for a particular set of information, then properly analyzed data are provided.

3.2 Principal component analysis (PCA)

PCA is a statistical technique, which is used to reduce the dimensionality of a dataset consisting of many correlated variables. It retains the variation present in the dataset, up to the maximum extent, and uses the ideas of variances and co-variances. The following steps are performed to reduce the dimensionality of the dataset.

Step 1 Standardization or scaling the data into a comparable range.

Step 2 Computing the covariance matrix to identify the correlation and dependencies among the features in a dataset.

Step 3 Calculating eigenvectors and eigenvalues from the covariance matrix to determine the largest to smallest eigenvalues of the eigenvectors.

Step 4 Computing the principal component. (The eigenvector with the highest eigenvalue is most significant and taken into consideration and thus the first principal component is formed. The principal components which are lesser significant are removed to reduce the dimension of the data.)

Step 5 Rearrange the original data with the final principal components and thus reducing the dimension of the dataset.

PCA helps to identify the correlation and dependencies among the features in a dataset. A covariance matrix expresses the correlation between the different variables in the dataset. It is essential to identify heavily dependent variables because they contain biased and redundant information, which reduces the overall performance of the model. Mathematically, a covariance matrix is a $p \times p$ matrix, where p represents the dimensions of the dataset. Each entry in the matrix represents the covariance of the corresponding variables. Here are the key takeaways from the covariance matrix:

1. The covariance value denotes how co-dependent two variables are with respect to each other.
2. A negative covariance denotes the respective variables are indirectly proportional to each other.
3. A positive covariance denotes that the respective variables are directly proportional to each other.

Eigenvectors and eigenvalues are the mathematical constructs that are computed from the covariance matrix to determine the principal components of the dataset. Principal components are the new set of variables that are obtained from the initial set of variables. These are computed in such a manner that newly obtained variables are highly significant and independent of each other. The principal components compress and possess most of the useful information that was scattered among the initial variables. If any dataset is of five dimensions, then five principal components are computed such that the first principal component stores the maximum possible information and the second one stores the remaining maximum information and so on. Eigenvectors and eigenvalues are two algebraic formulations computed as a pair. For every eigenvector, there is an eigenvalue. The dimensions in the data determine the number of eigenvectors that are needed to calculate.

The idea behind eigenvectors is to use the covariance matrix to understand where in the data there is the most amount of variance. Since more variance in the data denotes more information about the data, eigenvectors are used to identify and compute principal components. Eigenvalues, on the other hand, simply denote the scalars of the respective eigenvectors. Therefore, eigenvectors and eigenvalues compute the principal components of the dataset. After computing the eigenvectors and eigenvalues, we have arranged them in descending order, where the eigenvector with the highest eigenvalue is the most significant and thus forms the first principal component. The principal components of lesser significances can thus be removed in order to reduce the dimensions of the data. The final step in computing the principal components is to form a matrix known as the feature matrix that contains all the significant data variables that possess maximum information about the data. The last step in performing PCA is to rearrange the original data with the final principal components, which represents the maximum and the most significant information of the dataset. In order to replace the original data axis with the newly formed principal components, the transpose of the original dataset is multiplied by the transpose of the obtained feature vector.

4 Dataset collection

The majority of e-commerce consumers fall under the age group of 25–36 years and the highest spenders are the customers above the age of 37 years [19]. It has been observed that in India distribution of online customers is based on the age group. The database for this work is created by collecting data using online Google Forms, designed from our side. The form consists of a set of attributes required for the proposed work. The form was sent to employees of our institution, alumni as well as to the parents of the present students to get a robust variation of feedback from all categories of customers. Before framing the questionnaire, a pilot survey was conducted. The data collection period was from August to October which is the festive period in West Bengal. During this period, people normally purchase different apparel, so sales during this time are quite important for retailers also. Total 1856 records were collected for analysis out of which 30 datasets/records were incomplete. After normalization, the total numbers of records available are 1826 only. Table 1 depicts how the dataset has been collected from individual customers.

This dataset contains information about the brand and product which the customers have purchased online during Indian festive periods like Durga Puja, Diwali, and Eid. The customers were given a set of questions where they also mentioned the satisfaction level about the particular brand and products which they have purchased. In addition to these, the customers also mentioned their income and whether they have used a credit card or not while purchasing. The dataset contains the data of eight products of fourteen different brands, out of which nine brands were products of men, five brands were products of women, and among them two of the products are of both men and women. The products are also of six different sizes.

Table 1 Dataset collection

Survey ID	Product	Brand	Size (in inch)	Price (in Rs.)	Gender (Male/Female)	Credit Card (Yes/No)	Satisfaction	Monthly Income (in Rs.)
1	Shirt	Brand8	40	1499	Male	No	4	45,000
2	Trouser	Brand4	38	4499	Male	Yes	5	95,000
3	Kurti	Brand10	34	799	Female	Yes	5	55,000
……	……	………	……..	……	……	……..	…..	……..
……	……	……..	……..	…….	……	……..	…..	……..
……	……	……..	……..	…….	…….	……..	…..	
1826	T-Shirt	Brand7	40	3399	Male	No	3	65,000

5 Result analysis

To give equal importance to all features, all the continuous features of the dataset have been scaled and all the categorical attributes of the dataset have been represented numerically using one hot encoding scheme.

In the dataset, the price of the products and monthly salary are in Indian Rupees. These data have been standardized such that its distribution has a mean value 0 and a standard deviation of 1. All the features of a product such as type of product, brand, and price were combined together to generate a business key, and their purchase count was aggregated to get the sales volume of our model. These business keys are considered as the SKU of the retailers for the retailer recommendation model. Sales revenue has been calculated by multiplying the frequency count of the unique product with the purchased price of that product. The sales volume of each unique product was also standardized. The purchase frequency count of each unique products is considered as the sales volume of the particular product.

The elbow method has been used to determine the optimum number of clusters K. Distortion has been calculated as the average of the squared distances from the cluster centers of the respective clusters. We have iterated the value of K from 1 to 11 and calculated the distortion for each value of K in the given range. Values of K have been plotted on the horizontal (X) axis and the distortion on the vertical (Y) axis. When the value of K has been increased, the centroids started to become closer to the centroids of the clusters. The improvements declined at some points rapidly, creating the elbow shape. That point is the optimum value for K.

5.1 Clustering of customers by *K*-Means clustering algorithm

Figure 3 shows the segmentation of customers by the K-Means clustering algorithm. Here, we can segment the customers into four categories according to the price of products they have purchased and their monthly income.

From the elbow method, we have found that the optimum value of K is 4. Cluster 0, cluster 1, cluster 2, and cluster 3 contain 468, 231, 525, and 602 customers, respectively. Cluster 0 shows the categories of customers whose income is moderately low (in the range of Rs. 30,000–56,000) and the price of their purchased product is moderately low (in the range of Rs. 599–2799). Cluster 1 shows the categories of customers whose income is moderately high (in the range of Rs. 69,000–120,000) and the price of their purchased product is moderately high (in the range of Rs. 2399–5000). Cluster 2 shows the categories of customers whose income is moderately high (in the range of Rs. 56,000–87,000) and the price of their purchased product is moderately low (in the range of Rs. 599–2799). Cluster 3 shows the categories of customers whose income is moderately low (in the range of Rs. 37,000–67,000) and the price of their purchased product is moderately high (in the range of Rs. 2499–5000).

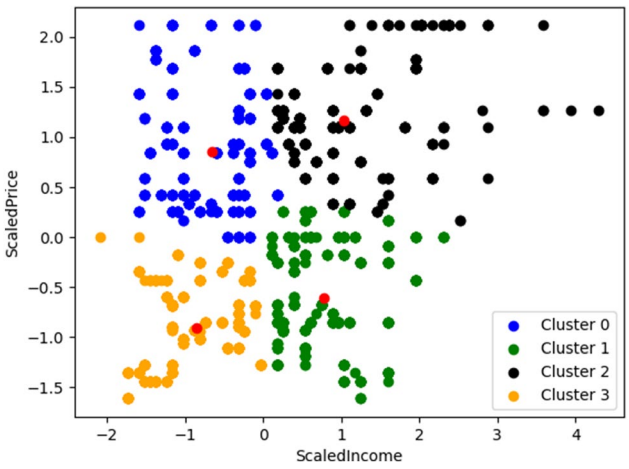

Fig. 3 Segmentation of customers by K-Means clustering algorithm

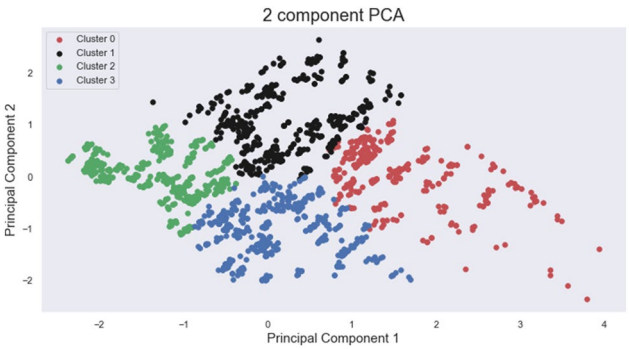

Fig. 4 Segmentation of customers using PCA

5.2 Clustering of customers using PCA

All the features like size of products, brand of products, price of products, customer satisfaction levels, and monthly salaries of the customers were considered to cluster the customers. All the categorical features like brand of products and size of products have been encoded with one hot encoding. Figure 4 shows the segmentation of customers using PCA. In this case, four clusters contain 454, 329, 518, and 525 data points, respectively.

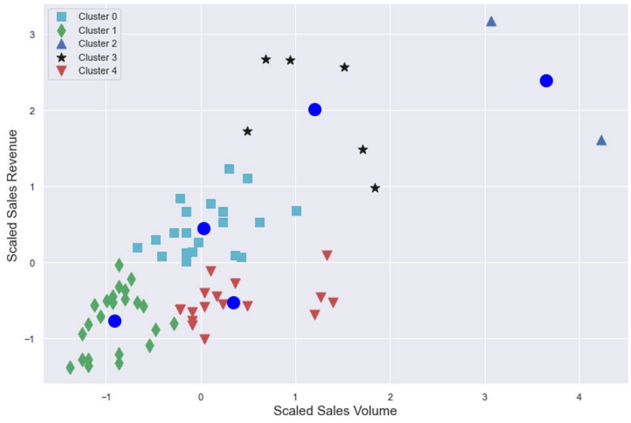

Fig. 5 Segmentation of products by *K*-Means clustering algorithm on the basis of sales volume and sales revenue

5.3 Clustering of products by *K*-Means clustering algorithm

Figure 5 shows the segmentation of products by the *K*-Means clustering algorithm on the basis of scaled sales volume and scaled sales revenue. From the elbow method, we have found that the optimum value of *K* is 5 in this case. Each data point of Graph represents the SKU of a particular type of product (product–brand–price).

It has been observed from cluster 2 that item3 of Brand1 and item2 of Brand4 have been purchased by more in numbers, hence generating maximum revenue though their prices are moderate. Similarly, the products of cluster 4 have been sold in medium level and generated revenue in medium level. In the case of cluster 1, either low price products generated low revenue or high price products because of less sale generated low revenue. Cluster 3 indicates that products were sold moderately high, and hence revenue generation was good. From the datasets of cluster 0, it can be concluded

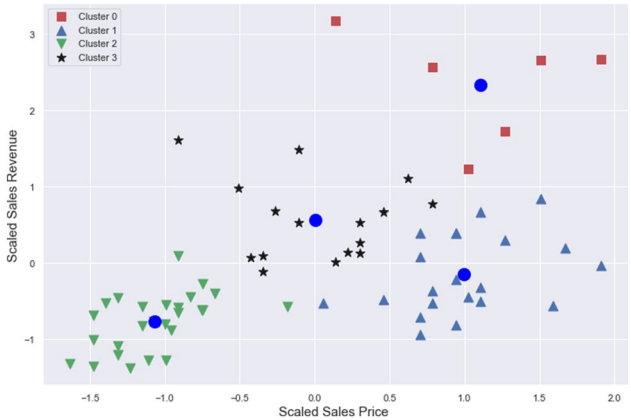

Fig. 6 Segmentation of products by *K*-Means clustering algorithm on the basis of sales price and sales revenue

that the product price is moderately high and sales volume is also moderately high so they also can generate good revenue. Figure 6 depicts the segmentation of products by the *K*-Means clustering algorithm on the basis of the scaled sales price of the product and scaled sales revenue. From the elbow method, we have found that the optimum value of *K* is 4 in this case. Each data point of Graph represents the SKU of a particular type of product (product–brand–rice).

From the figure, it is evident that cluster 0 contains the products that have high price and generate high revenue. Similarly, the products of cluster 2 have a moderately high price but due to low sales volume their revenue generation is low. Cluster 3 shows the products which have moderate sales value and generate moderate revenue. Cluster 1 indicates that though the sale price is high because of low sale, the products could not contribute much to generate sales revenue.

Figure 7 shows the percentage of each brand in our database. Though we have used the popular apparel brands of Men and Women products, we have not mentioned the name of the brand in our proposed model, as the model can work with any type of apparel product.

We have suggested to purchase products according to our proposed recommendation model. Again, the dataset was collected from online apparel customers who have purchased from August to October 2019. The recommendation result was validated with 165 customers. In this case also, it has been found that 19.5% of products were purchased from Brand1 whereas 4.5% products were purchased from Brand14.

6 Conclusion and future work

In the proposed work, an e-commerce-based online recommendation system has been implemented for customers by clustering different apparel products based on their brands, size, price, and income of the customers. In our dataset, the monthly income of the customers varies from Rs. 30,000 to Rs. 1,200,000. The ultimate goal of this work is to segment

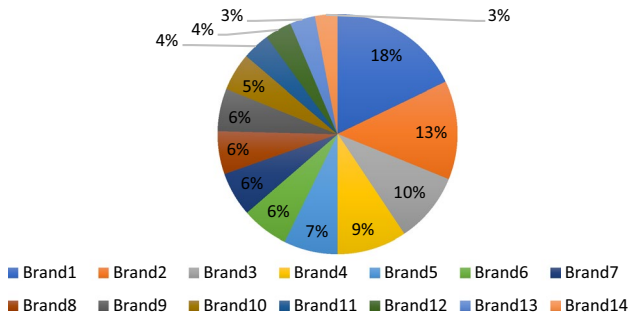

Fig. 7 Percentage of each brand in our database

users into smaller groups, which can be viewed as groups of users, who like to purchase the same type of garments. PCA helps in dimensionality reduction. *the K*-Means clustering is one of the most popular and common methods of clustering which is often applied to get an idea of the structure of the dataset. The objective of the *K*-Means algorithm is to group data points into distinct non-overlapping subgroups. It has been observed that using principal component analysis gives similar results for product recommendation as that of the *K*-Means clustering algorithm.

The database was created using Google Forms with input from the alumni, staff, and other stakeholders of our institution, and recommendation was done on the local intranet. The same will be made available on the internet for normal customers where the secrecy of data is to be handled with care. Though in the present work the recommendation was done based on a small dataset which is available to us, in the long run, the same concept will help the businesses and customers. It is believed that customers' satisfaction needs to be maximized so that the consumers continue buying the same brand of products rather than the competitive brand. Thus, by maintaining brand equity maximum sales can be achieved. At a global level, customer loyalty is positively related to the profitability and long-term growth of a company. The companies need to consistently analyze the sale of different products to increase business profitability. Brand–item clusters may be informed to the companies after considering the results found by the analysts. In this work, the datasets were collected from the customers' level only. In future, the database of the apparel SKU will be collected from retailers and our model for recommendation to the retailers will be applied on that. Proper business analysis and sales forecasting can help the retailer to understand the customers' behavior, what items sell, and what do not. The product segmentation of our proposed model will help the retailer to maintain the optimum stock level. They can get the cluster of items which are not going to sell anyway and will not overspend on it. This will reduce wastes and benefit our society as green computing is concerned [20].

References

1. Kyoung-jae K, Hyunchul A (2008) A recommender system using GA *K*-Means clustering in an online shopping market. Expert Syst Appl 34(2):1200–1209
2. Shinde SK, Kulkarni U (2012) Hybrid personalized recommender system using centering-bunching based clustering algorithm. Expert Syst Appl 39(1):1381–1387
3. Gong S (2010) A collaborative filtering recommendation algorithm based on user clustering and item clustering. J Softw 5(7):745–752
4. Cho YS, Moon SC, Jeong S, Oh IB, Ryu KH (2014) Clustering method using weighted preference based on RFM score for personalized recommendation system in u-Commerce. In: Jeong YS, Park YH, Hsu CH, Park J (eds) Ubiquitous information technologies and applications, vol 280. Lecture Notes in Electrical Engineering. Springer, Berlin
5. Mohamed F, Mohamed C (2014) Application of data mining in e-Commerce. J Inf Technol Res 7(4):79–91
6. Ester M, Kriegel H P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining, Portland, Oregon, pp 226–231
7. Pitsilis G, Zhang X, Wang W (2011) Clustering recommenders in collaborative filtering using explicit trust information. In: Wakeman I, Gudes E, Jensen CD, Crampton J (eds) Trust Management V. IFIPTM 2011. IFIP Advances in Information and Communication Technology. Springer, Berlin, pp 82–97
8. Choi K, Yoo D, Kim G, Suh Y (2012) A hybrid online-product recommendation system: combining implicit rating-based collaborative filtering and sequential pattern analysis. Electron Commer Res Appl 11(4):309–317
9. Oyelade OJ, Oladipupo OO, Obagbuwa IC (2010) Application of *k*-Means clustering algorithm for prediction of students academic performance. Int J Comput Sci Inf Secur 7(1):292–295
10. Mathivanan NMN, Ghani NA, Janor RM (2018) Improving classification accuracy using clustering technique. Bull Electr Eng Inform 7(3):465–470
11. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
12. Zha H, Ding C, Gu M, He X, Simon H (2001) Spectral relaxation for *K*-Means clustering. In: Advances in neural information processing systems 14 (NIPS'01), pp 1057–1064
13. Ding C, He X (2004) *K*-Means clustering via principal component analysis. In: Proceedings of the international conference on machine learning, pp 225–232
14. Chuan NK, Sivaji A, Shahimin MM, Saad N (2013) Kansei engineering for e-commerce sunglasses selection in Malaysia. In: Procedia—Social and Behavioral Sciences, the 9th international conference on cognitive science 2013, vol 97, pp 707–714
15. Nilashi M, Ibrahim OB, Ithnin N, Sarmin NH (2015) A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. Electron Commer Res Appl 14(6):542–562
16. Wang Y, Wu Z, Bu Z, Cao J, Yang D (2016) Discovering shilling groups in a real e-commerce platform. Online Inf Rev 40:62–78
17. Shang W, Zhu H, Huang H, Qu Y, Lin Y (2006) The improved ontology kNN algorithm and its application. 2006. ICNSC '06. In: Proceedings of the 2006 IEEE international conference on networking, sensing and control, ICNSC2006, pp 198–203
18. Celebi M, Kingravi H, Vela P (2013) A comparative study of efficient initialization methods for the *k*-Means clustering algorithm. Expert Syst Appl 40(1):200–210
19. Consumers above 37 years are highest e-commerce spenders: WATConsult's e-commerce report. https://bestmediainfo.com/2018/08/consumers-above-37-years-are-highest-e-commerce-spenders-watconsult-s-e-commerce-report/ Accessed 30 Aug 2018
20. Ahmed AI (2018) Understanding the factors affecting the adoption of green computing in the Gulf Universities. Int J Adv Comput Sci Appl 9(3):304–3011