Yevhenii Repetskyi
Master student,
Economic Cybernetics Department,
Taras Shevchenko National University of Kyiv,
Kyiv, Ukraine
evgeniy_repetsky@knu.ua

E-COMMERCE CUSTOMER SEGMENTATION APPROACH BASED ON K-MEANS CLUSTERING ALGORITHM

Abstract

This investigation is devoted to the application of k-means clustering algorithm in the case of e-commerce customer segmentation. For the purpose of modeling and analysis a data set, which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail was used. The elbow method for defining amount of clusters was applied. The silhouette score for the measuring of model quality was used.

Key words: e-commerce, customer segmentation, clustering, elbow method, k-means.

Introduction

The rapid development of communication technologies is currently transforming many processes in modern society. The business sector is not an exception. The Internet, as the most accessible and convenient system for the global exchange of information between users, has not only proven its viability, but is also beginning to replace other methods and channels of communication, which is due to the lower cost of services, high data transmission speed, and a wider range of information presented and transmitted.

At the present stage of development of market relations, a special role is assigned to the tasks of improving the innovative component of the entrepreneurial activity of enterprises in order to increase the efficiency of business. The emergence of the global computer network was marked by the emergence of a new communication environment and a market with a large number of potential consumers with a fairly high level of income. E-commerce in general and internet marketing in particular today

perform communication functions and provide opportunities for concluding transactions, making purchases and making payments.

Today the stability of business development largely depends on the loyalty of the audience, which affects sales volumes, distribution and profit levels especially in the case of e-commerce. Therefore, it is important for companies to know their customers, build long-term relationships with them, take into account their needs, benefits and priorities. It is also necessary to increase the competitiveness of products and services, which is possible through the formation and management of customer loyalty, since it is customer loyalty that is one of the main factors in the competitiveness of an enterprise.

In order to build long-term relationships, a company can influence consumer behavior, but for this it needs to know how to effectively work with each of them. For this, it is necessary to properly segment the company's customers. There are a lot of customer segmentation approaches. One of the most popular are once that based on data science techniques.

Literature overview

There are a lot of modern researches devoted to the usage of data science models in customer segmentation.

Shaik, Nittela, Hiwarkar and Nalla in their investigation use the k-means clustering algorithm to detect the patterns of customer segment based on e-commerce big-data. Euclidian distance for the modeling was used. The features for the clustering were frequency of transactions, total quantity of the item (product) bought and total sum purchased by the customer. The considered optimal number of clusters was 3. Authors came to the conclusion, that k-means algorithm has a good efficiency in the case of big data, but it has some restrictions, such as need for the prior defining the fixed number of clusters [3].

The work of Kamthania, Pahwa and Madhavan is devoted to the construction of business intelligence tool for market segmentation based on k-mode clustering algorithm. They used customer behavior analysis and geographical information as input features. The model was evaluated silhouette score. The number of cluster

considered is 31. The dataset used by authors refers to 200 users of dummy e-commerce website developed and hosted online as a part of the experiment conducted for their research. The proposed technique could simplify strategy decisions and be especially useful for the small e-commerce business owners or growing startups [2].

Data Preparation and Preprocessing

For the purpose of modeling and analysis a data set, which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail was used.

The data set contained 8 variables: invoice number, product (item) code, product (item) name, quantities of each product (item) per transaction, invoice date, product (item) price per unit, customer identifier, country name. Data was preprocessed in order to eliminate missing values and outliers.

Based data set variables number of invoices, average unit price and average amount of units indicators for each customer were created. These indicators were used as features in the clustering model. Since, the range of values of data may vary widely, it becomes a necessary step in data processing to scale it before using in the model. The range of features data was scaled between 0 and 1.

Results

In this investigation the k-means clustering algorithm was used. All the calculations were made using the Python programming language.

For the purpose of cluster number choice, the elbow method was used. The optimal clusters number in our case is 4 (see Figure 1).

Based on the simulation results, clusters of buyers were obtained. The silhouette score of the model, which determines how clusters are differ from each other, is 0.511. Thus, we can conclude, that the model is of rather good quality. The average feature values for each of the clusters are shown in the Table 1.

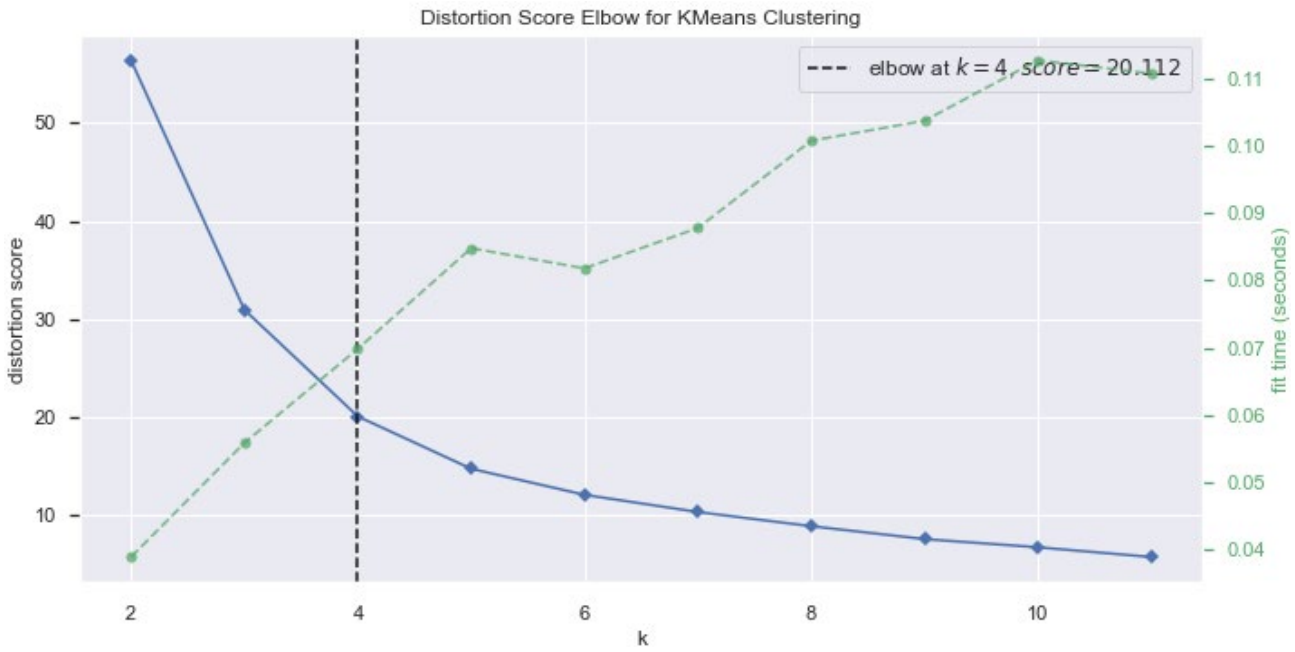

Figure 1. Elbow method

Source: Compiled by author based on his own calculations

Table 1

Average feature values of clusters

Cluster	Average feature value		
	Number of Invoices	Average Unit Price	Average Amount of Units
0	4.86	2.69	264.49
1	3.15	4.52	66.28
2	4.50	2.69	400.94
3	4.46	3.06	155.14

Source: Compiled by author based on his own calculations

The 3-dimensional plot of clusters is shown in Figure 2.

Conclusions

Customer segmentation is one of the most important aspects of a company's success. In this investigation an application of k-means algorithm for the purposes of customer segmentation was proposed. The data on e-commerce transaction were used. The obtained results demonstrated the good quality of the model built. Therefore, they can be the basis for the marketing activities as well as strategy development decisions.

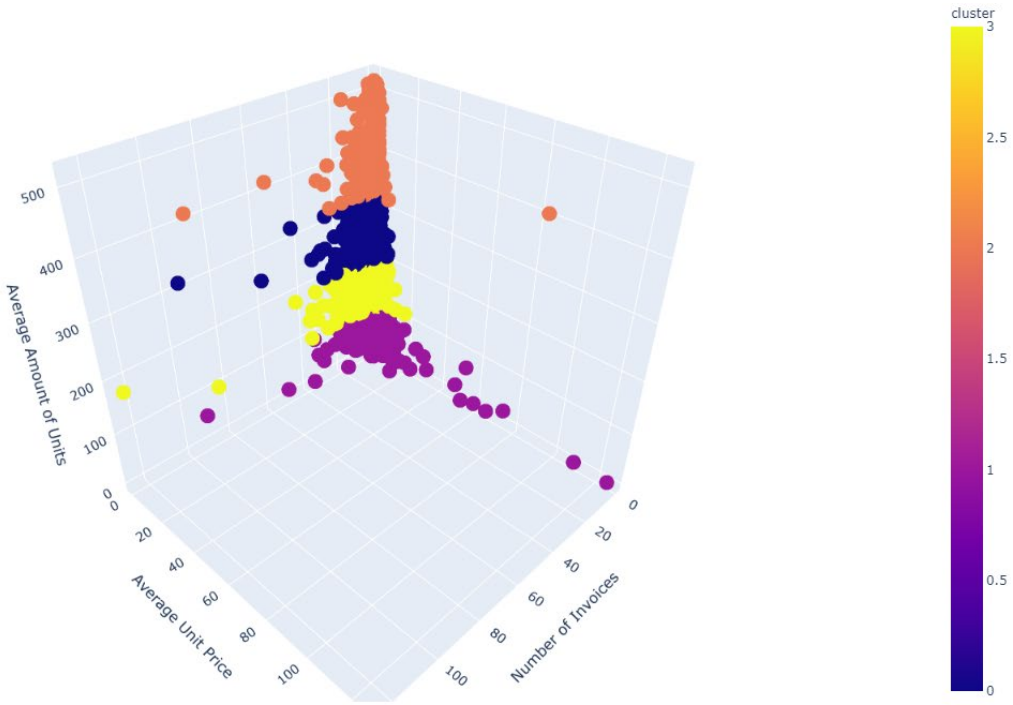

Figure 2. 3D plot of clusters

Source: Compiled by author based on his own calculations

References

1. Chunhui Yuan, Haitao Yang (2019) 'Research on K-Value Selection Method of K-Means Clustering Algorithm', *Multidisciplinary Scientific Journal,* 2(16), 226-235 [Online]. Available at: https://www.mdpi.com/2571-8800/2/2/16/pdf (Accessed: 25 November 2020).
2. Deepali Kamthania, Ashish Pahwa and Srijit S. Madhavan (2018) 'Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business', *Journal of Computing and Information Technology,* 26(1), pp. 57–68 [Online]. Available at: https://www.researchgate.net/publication/326403359_Market_Segmentation_Analysis_and_VisualizationUsing_K-Mode_Clustering_Algorithm_for_E-Commerce_Busines s/link/5b56814445851507a7c408b1/download (Accessed: 25 November 2020).
3. Indivar Shaik, Swapna Suhasini Nittela, Trayabak Hiwarkar, Srinivas Nalla (2019) 'K-means Clustering Algorithm Based on E-Commerce Big Data', *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* 8(11), [Online]. Available at: https://www.ijitee.org/wp-content/uploads/papers/v8i11/K21210981119.pdf (Accessed: 25 November 2020).

4. Juni Nurma Sari, Lukito Edi Nugroho, Ridi Ferdiana, P. Insap Santosa (2016) 'Journal of Computational and Theoretical Nanoscience', *Multidisciplinary Scientific Journal,* 22(10), [Online]. Available at: https://www.researchgate.net/publication/313737530_Review_on_Customer_Segmentation_Technique_on_Ecommerce (Accessed: 25 November 2020).
5. Sajid Naeem, Aishan Wumaier (2018) 'Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K', *International Journal of Computer Applications,* 182(31), [Online]. Available at: https://www.researchgate.net/publication/331045887_Study_and_Implementing_K-mean_Clustering_Algorithm_on_English_Text_and_Techniques_to_Find_the_Optimal_Value_of_K (Accessed: 25 November 2020).
6. Trupti M. Kodinariya, Prashant R. Makwana (2013) 'Review on determining number of Cluster in K-Means Clustering', *International Journal of Advance Research in Computer Science and Management Studies,* 1(6), [Online]. Available at: https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering (Accessed: 25 November 2020).