**Research Article**

# Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers

Serpil Ustebay[1] , Ilkay YELMEN[2] , Metin ZONTUL[3]

[1] *Istanbul Medeniyet University, Computer Engineering Department, 34720 Kadikoy, Istanbul, Turkey*, serpil.ustebay@medeniyet.edu.tr, https://orcid.org/0000-0003-0541-0765

[2] *Istanbul Aydın University, Computer Engineering Department, 34295 Kucukcekmece, Istanbul, Turkey*, ilkayyelmen@stu.aydin.edu.tr, https://orcid.org/0000-0002-1684-9717

[3] *Istanbul Arel University, Computer Engineering Department, 34537 Buyukcekmece, Istanbul, Turkey*, metinzontul@arel.edu.tr, https://orcid.org/0000-0002-7557-2981

| Article Info | Abstract |
|---|---|
| | Customer segmentation is a customer grouping model based on common features and it directly relates with customer satisfaction of the companies. It provides access to the right customer with the right methods by knowing the customer better. Dealing with changes in a competitive market means airlines have to redefine customer segmentations, which translates from social-demography to a more complex behavioral approach that covers the entire travel experience and the way airlines deliver at every touch point. In this paper, a customer segmentation was performed using an airline ticket sales data and focused on two concepts such as customer loyalty and customer return. Customers with similar sales tendencies were clustered by using self-organizing map method and totally 15 clusters were obtained. In purchasing trends, the highest return was obtained in cluster 2 and the minimum return in cluster 6. Loyalty rate was calculated as 38% and it was seen that the most loyal customer profile was in the cluster 12. |

# Kendi Kendini Düzenleyen Haritalara Dayalı Müşteri Segmentasyonu: Havayolu Yolcuları Üzerine Bir Vaka Çalışması

| Makale Bilgisi | Öz |
|---|---|
| | Müşteri segmentasyonu, ortak özelliklere dayanan bir müşteri gruplama modelidir ve şirketlerin müşteri memnuniyeti ile doğrudan ilişkilidir. Müşteriyi daha iyi tanıyarak doğru müşteriye doğru yöntemlerle erişim sağlar. Rekabetçi bir pazardaki değişikliklerle başa çıkmak, havayolu şirketlerinin müşteri segmentasyonlarını yeniden tanımlamaları gerektiği anlamına gelir; bu da sosyal demografiden tüm seyahat deneyimini ve havayollarının her temas noktasında teslim şeklini kapsayan daha karmaşık bir davranışçı yaklaşıma dönüşür. Bu makalede, bir uçak bileti satış verileri kullanılarak bir müşteri segmentasyonu gerçekleştirilmiş ve müşteri sadakati ve müşteri iadesi gibi iki konsepte odaklanmıştır. Benzer satış eğilimlerine sahip müşteriler kendi kendini düzenleyen haritalar yöntemi kullanılarak kümelenmiş ve toplam 15 küme elde edilmiştir. Satın alma eğilimlerinde en yüksek getiri küme 2'de, minimum getiri ise küme 6'da elde edilmiştir. Sadakat oranı ise %38 olarak hesaplanmış olup, en sadık müşteri profilinin küme 12'de olduğu görülmüştür. |

## 1. INTRODUCTION

Customer segmentation concept was proposed by Wendell R. Smith, in 1956. Customer segmentation is an approach to classifying customers based on their preferences, demands, and other factors in the business model and targeted market conditions. While customers belonging to the same groups have certain similarities, customer groups divided into different segments have different characteristics [1]. The customer segmentation

model is a model created by clustering the customers according to the standards on the selected variables [5].

Airline industry is one of the most important sectors that need customer segmentation. Significant changes in the market environment strengthened competition in the airline industry [1, 17]. In this industry, low-cost carriers (LCC) have been established by penetrating significant parts of the markets. While the sector was previously dominated by government airlines, now it is shaped by the competitive attacks of companies. In this new environment, it is important to develop a customer and market-oriented sales strategy. Traditionally, airlines classify their passengers into the business and economy class and adapt their sales strategies to the flexibility for business passengers and the price for economy passengers. The competitive situation has created a change in the market in favor of customers who are aware of their needs. [1, 18]. Dynamically changing market and customer needs has led airline companies to get to know the customer better. For this, customers need to be segmented according to their purchase status.

In the increasingly complex and competitive business environment, systematic work and clustering customers by specific categories improve customer loyalty and expand the profitable customer portfolio, increasing business levels for long-term relationships [10].

There are various data mining methods to extract useful information from huge volumes of data. Especially in CRM applications there are large amount of customer data and so data mining is a useful area to understand customer behavior [19]. In data mining, the clustering method generates clusters for the given input data where data in one cluster is more similar when compared to data in other clusters [20]. The similarity is measured in terms of the distance between the data.

Customers also can be segmented using other Artificial Intelligence (AI) techniques such as Self Organizing Map (SOM), Bee Colony, Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). GA is part of the evolutionary computing. It is based on the natural evolutionary process that uses processes such as mutation and crossing. The main disadvantage of GA is that it is time consuming and makes it difficult for the solution to approach an optimal solution. On the other hand, SOM is a neural network based unsupervised algorithm. Neural network has interconnected processing elements and reaches input and output relationships. The algorithm is based on trial and error method and suffers from long training time [21].

The study aims to segment airline customers' using their purchasing behavior. Experimental studies have been carried out by SOM and made inferences about Customer loyalty and Customer Return.

The rest of this paper is structured as follows. In Section 2, literature review on customer segmentation methods is discussed; then the details of dataset are presented in Section 3. The data preprocessing, clustering, and experiments and results are presented in Section 4, 5 and 6 respectively. Finally, we conclude the paper with some future research plans in Section 7.

## 2. LITERATURE REVIEW

Classically, airline companies would segment their customers based only on demographic information. Segmentation of the customer is based on dividing the passengers by the purpose of the trip. Increasing the ancillary revenue for each seat sold has become an indispensable condition for airline companies. Therefore, it is necessary to segment the passengers according to their travels and develop the airline sales strategy according to this segmentation. Demographic segmentation does not include the needs and motivation of passengers to travel, but behavioral segmentation helps airlines understand their willingness to pay as and first-class and business passengers, as they tend to make reservations close to the date of departure, which underlies the price sensitivity for certain routes or periods [22] .

Big data impact on CRM is reviewed based on the critical success criterias. Results shows three contributions like previous contributions, past reviews and five propositions [14]. Neural networks and Naïve Bayes classification methods are used to create CRM structure to enhance the decision making for not lose the customers [15]. Streams of research in multiple areas like information management, operations, marketing and other areas are linked to create an integrated structure [16].

Data mining approach has been successfully used for CRM in the financial and telecommunications sectors. As an example, support vector machine (SVM) classification algoritm is used for customer credit scoring. On the other hand, Regression is used for customer loyalty [12]. In marketing, K-Nearest Neighbor, Decision Tree and Naive Bayes are used [13].

Gupta & Pathak have created a dynamic product pricing model by using shopping transactions obtained from an e-commerce company. The developed model aims to offer dynamic prices to different customers. K-Means clustering algorithm has been used to define customer purchase behaviors [3].

Tsai & Chiu has developed a new clustering algorithm, that is based on Genetic Algorithm to cluster customers, based on product related variables such as the purchased products and related monetary expenses from transactional customer histories [6].

Chen, Chiu, & Chang are identified patterns of customer behavior by using association rule method that applied to analyze the relationships of products purchased by customers at retail stores. Apriori algorithm is applied to discover the patterns of customer behavior. Apriori has generated 111 rules, for the first period and 78 rules for second-period [2].

Maryani et al. performed customer segmentation based on RFM model on Nine Reload Credit by applying K-means clustering method. They used 82,648 sales transactions [4].

Customer segmentation can also be used in e-pharmacy sector to increase the customer retention and it is one of the prerequisite for the essential CRM in the customer loyalty [8]. Due to rapid development of the market, customer segments can change over a time period. Carnein et al. proposed a new concept and a tool as 'Stream Clustering' to handle the customer segmentation changing issue. An important aspect in Stream Clustering is to indicate new clusters or revealing clusters and replacing the older ones [9].

In summary, airline companies face changes in both the business environment and customer behavior. The flying class no longer seems to be a viable indicator for identifying heterogeneous customer segments. As the spectrums of choice in both classes are increasingly expanding, the traditional segmentation of passengers for business and entertainment becomes invalid [1]. New approaches are needed for customer segmentation to identify different segments that represent customer preferences and respond to targeted product or service offerings.

In this study, flight data belonging to passengers of an airline company in Turkey are used to categorize the airline passengers according to Customer loyalty and Customer Return by SOM algorithm. After data collection and data preprocessing, the SOM algorithm have been used to cluster the data. As a result, 15 clusters have been found.

## 3. DATASET

In this study, total 3.200.000 passenger sales data of an airline between 2017-2018 were used. The collected data attributes related with the ticket sales have converted to a new format for better clustering and shown in Table 2.

In our work, we used Microsoft Excel to convert data files to other formats and editing data. Python programming language and sompy library were used for SOM clustering.

**Table 1.** Features and descriptions of dataset

| Column Number | Description |
|---|---|
| 1 | Passenger ID |
| 2 | Passenger age |
| 3 | Total sales price between January-December 2018 |
| 4 | Total sales price in January 2018 |
| 5 | Total sales price between October-December 2017 |
| 6 | Total sales price between July and December 2017 |
| 7 | Total sales price between January-December 2017 |
| 8 | Total travel at January 2018 |
| 9 | Total number of trips between October-December 2017 |
| 10 | Total number of trips between July and December 2017 |
| 11 | Total number of trips between January and December 2017 |
| 12 | Average travel sales price (for 2017 and 2018) |
| 13 | Average value of sales made in October-December 2017 |
| 14 | Average value of sales in July-December 2017 |
| 15 | Average value of sales made in January-December 2017 |
| 16 | Days since first flight (Start January 2017 - End 2018 December) |
| 17 | Number of days since last flight (Start January 2017 - End 2018 December) |
| 18 | Number of last months since the last flight (Start 2017- End 2018 December) |

## 4. DATA PREPROCESSING

During the data preprocessing phase, from the 22 columns in the dataset, 18 suitable columns were selected for data analysis. All null values and all travel records of 0 between January 2017 and 2018 have been removed. Also some outliers, such as a very high ticket price with negative values were detected, and those with total sales prices between 0 and 10.000 TL were considered. Finally, a total of 148710 data and 18 attributes were obtained for use in clustering.

## 5. CLUSTERING

Self-organizing maps (SOM), is one of the unsupervised learning methods developed by Kohonen [26]. In a high-dimensional input space, it produces low-dimensional representations of training points.

Let S be a set of training points in an n-dimensional input space and the or latent space is (m − 1) dimensional. There are D = D1 x···x Dm−1 neurons in the latent space. Each neuron u = 1, ···, D has

- a predefined position in the latent space: $z^u = (z_1^U, \cdots, z_{m-1}^U)$, $z_i^U \in \{1, \cdots, D_i\}$ for i = 1, ···, m−1, and
- a representation or weight vector $w^u = (w_1^U, \cdots, w_n^U)$ in the input space.

Typical SOM topology has been illustrated in Fig. 1. The input space is n-dimensional and the latent space is 2- dimensional in the example. SOM purposes to find out the representation vectors which are embedded in the latent space through the training points. SOM learning begins with the assignment of a randomly selected training point from S to each neuron weight vector. The weight vectors are updated by repeating training points close to them [7].
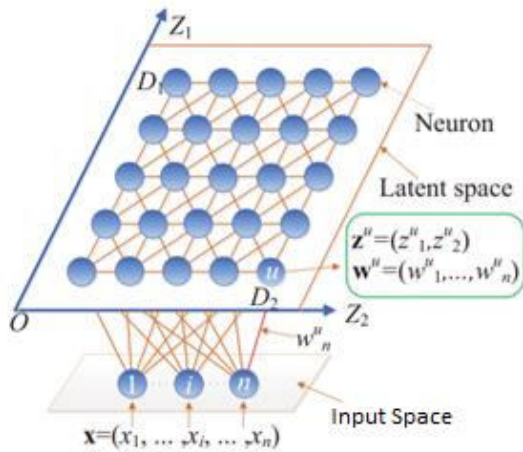


**Figure 1.** An illustration of a 2-dimensional SOM [11].

## 6. EXPERIMENTS AND RESULTS

Before we start to experiment, data preprocessing has been applied on raw data set to obtain the high score in clustering phase. Then SOM algorithm has been applied to segment the airline passengers. Finally, customers are divided into a total of 15 clusters.

Table 2 shows the statistical information of the clusters obtained with SOM. The column of cluster count provides the number of customers assigned to that cluster, mean age column provides the average age of the customers, and the column of mean flight cost offers the average flight return of the customers. According to this table, the highest flight return is in cluster 2 with 3.417 TL, and the lowest flight return is in cluster 6 with 80 TL.

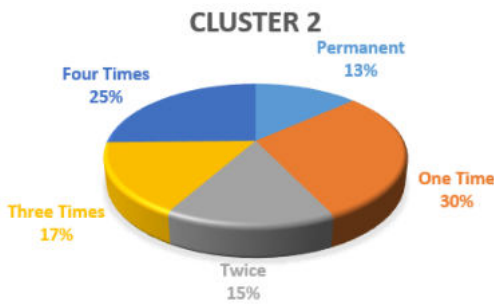**Table 2.** Statistical information of customers in clusters

| Cluster Number | Cluster Count | Mean Age | Mean Flight Cost (TL) |
|---|---|---|---|
| 0 | 12196 | 58 | 91 |
| 1 | 26636 | 29 | 86 |
| 2 | 3604 | 39 | 3417 |
| 3 | 9702 | 14 | 104 |
| 4 | 14675 | 51 | 122 |
| 5 | 14832 | 28 | 95 |
| 6 | 23222 | 31 | 80 |
| 7 | 7300 | 44 | 96 |
| 8 | 15285 | 28 | 88 |
| 9 | 4152 | 41 | 99 |
| 10 | 7701 | 46 | 100 |
| 11 | 3964 | 59 | 247 |
| 12 | 1449 | 16 | 515 |
| 13 | 893 | 42 | 1120 |
| 14 | 2418 | 39 | 646 |
| 15 | 680 | 49 | 408 |

Table 3 contains information about the purchase history of customers in clusters. The Permanent column shows the number of customers purchasing flights in each of the time segments in the dataset, and the one time column shows the number of customers who bought tickets only once in the time segments specified in the dataset. Similarly, the number of customers who bought tickets three times and four times is also indicated in the relevant columns.
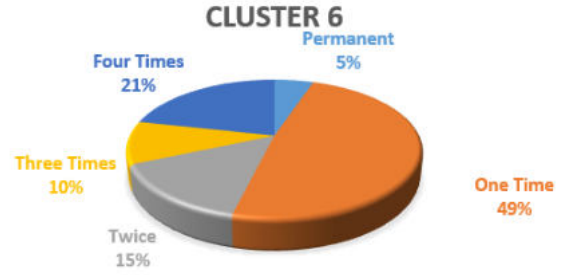
When the purchasing tendencies of the customers in the determined time periods are analyzed, the highest return is shown in the Figure 2 and the minimum return is shown in Figure 3.
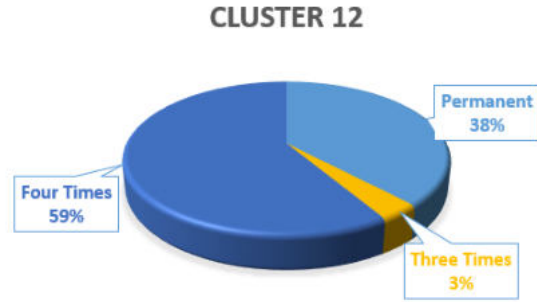
**Table 3.** Number of purchases in the clusters

| Cluster Number | Permanent | One Time | Twice | Three Times | Four Times |
|---|---|---|---|---|---|
| 0 | 1538 | 2965 | 2627 | 1851 | 3215 |
| 1 | 3446 | 6810 | 6584 | 3937 | 5859 |
| 2 | 482 | 1063 | 552 | 594 | 913 |
| 3 | 458 | 3565 | 1880 | 1807 | 1992 |
| 4 | 1375 | 5116 | 2214 | 1765 | 4205 |
| 5 | 662 | 6679 | 2847 | 1977 | 2667 |
| 6 | 1199 | 11306 | 3392 | 2320 | 5005 |
| 7 | 755 | 2280 | 1946 | 940 | 1379 |
| 8 | 671 | 7135 | 2609 | 2182 | 2688 |
| 9 | 173 | 1914 | 827 | 540 | 698 |
| 10 | 337 | 3443 | 1323 | 1193 | 1405 |
| 11 | 400 | 1044 | 572 | 744 | 1204 |
| 12 | 550 | 0 | 0 | 50 | 849 |
| 13 | 120 | 248 | 164 | 163 | 198 |
| 14 | 350 | 821 | 286 | 338 | 623 |
| 15 | 126 | 5 | 34 | 200 | 315 |



**Figure 3.** Number of purchases in cluster 6



**Figure 4.** Number of purchases in cluster 12

Loyalty rate of customers in cluster 12 is calculated as 38% and the most loyal customer profile is in this cluster. Cluster 12 is shown in the Figure 4. A total of 95% of the customers in this set could not purchase tickets at most once in all of the specified time periods. When the total flight cost is analyzed, the average ticket amount paid by those in this cluster per flight is 515 TL. This amount is the 4th highest amount. When the purchasing tendency in cluster 2 is analyzed, it is seen that the cluster has a customer profile that usually purchases a high amount of tickets at once. When the customer profile in the set 6 is analyzed, it is seen that the rate determined as the loyalty rate (permanent and four-time ticket purchase) is not very high.

## 7. CONCLUSION AND FUTURE WORK

Customer segmentation is a model that aims to divide customers into groups with similar behavior, needs, and expectations. It enables to get to know the customers better and to offer better marketing strategies to the right customers.

In this study, customers who purchased tickets between 2017-2018 in the airline company have been analyzed. Customers were divided into segments with the SOM algorithm according to their purchasing tendencies. It is aimed to create a risk map of the relevant group by examining customer profiles per cell in the SOM map designed in 4x4 size and to increase customer satisfaction with special campaigns to be created. The study also focused on two concepts companies should address in customer analysis. These concepts are



**Figure 2.** Number of purchases in cluster 2

customer loyalty and customer return. Loyal customers are the profile of customers who always purchase flight tickets from this company, but do not have a very high profit return, and are satisfied with the business process of the company. The group with high returns is generally defined as a one-time purchasing group. It is determined that this customer group is tourism companies. The ability of firms to achieve high returns depends on analyzing the target audience in detail and determining the new target audience. The proposed approach for customer behavior clustering can help managers to develop correct market strategy.

In the next study, a model will be developed that predict the possible sales of customers in further years by using Deep Learning methods**.**

## 8. REFERENCES

[1] F. Alamdari and K. Mason, "The future of airline distribution," *Journal of Air Transport Management*, vol. 12, pp. 122-134, 2006.

[2] M. C. Chen, A. L. Chiu, and H. H. Chang, "Mining changes in customer behavior in retail marketing," *Expert Systems with Applications*, vol. 28, pp. 773-781, 2005.

[3] R. Gupta, and C. Pathak, "A machine learning framework for predicting purchase by online customers based on dynamic pricing," *Procedia Computer Science*, vol. 36, pp. 599-605, 2014.

[4] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq and E. A. Pratama, "Customer segmentation based on RFM model and clustering techniques with K-means algorithm," *Proceedings of the 3rd International Conference on Informatics and Computing*, ICIC 2018, 2018. pp. 1–6.

[5] J. Wu, and Z. Lin, "Research on customer segmentation model by clustering," *In Proceedings of the 7th international conference on Electronic commerce,* 2005. pp. 316-318.

[6] C. Y. Tsai, and C. C. Chiu, "A purchase-based market segmentation methodology," *Expert Systems with Applications,* vol. 27, pp. 265-276, 2004.

[7] S. S. Haykin, Neural networks and learning machines, New Jersey: Pearson Education, Inc., 2009.

[8] M. Patak, H. Lostakova, M. Curdova, and V. Vlckova, "The e-pharmacy customer segmentation based on the perceived importance of the retention support tools," *Procedia-Social and Behavioral Sciences*, vol. 150, pp. 552-562, 2014.

[9] M. Carnein, and H. Trautmann, "Customer segmentation based on transactional data using stream clustering," *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*,Springer, Cham, 2019. pp. 280-292.

[10] K. Khalili-Damghani, K., F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," *Applied Soft Computing*, vol. 73, pp. 816-828, 2018.

[11] H. Zhang, A. Zhou, S. Song, Q. Zhang, X. Z. Gao, and J. Zhang, "A self-organizing multiobjective evolutionary algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 20, pp. 792-806, 2016.

[12] B. Lariviere, and D. Vandenpoel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert systems with applications,* vol. 29, pp. 472-484, 2005.

[13] T. Jiang, and A. Tuzhilin, "Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever?," *IEEE Transactions on knowledge and data engineering*, vol. 18, pp. 1297-1311, 2006.

[14] P. Zerbino, D. Aloini, R. Dulmin, and V. Mininno, "Big Data-enabled customer relationship management: A holistic approach," *Information Processing & Management*, vol. 54, pp. 818-846, 2018.

[15] T. F. Bahari, and M. S. Elayidom, "An efficient CRM-data mining framework for the prediction of customer behaviour," *Procedia computer science*, vol. 46, pp. 725-731, 2015.

[16] J. Sheng, J. Amankwah-Amoah, and X. Wang, "A multidisciplinary perspective of big data in management research," *International Journal of Production Economics,* vol. 191, pp. 97-112, 2017.

[17] K. Mason, K., "Observations of fundamental changes in the demand for aviation services," *Journal of Air Transport Management*, vol. 11, pp. 19-25, 2005.

[18] H. Lindstadt, B. Fauser, "Separation or integration? Can network carriers create distinct business streams on one integrated production platform?," *Journal of Air Transport Management*, vol. 10, pp. 23-31, 2004.

**[19]** E. W. Ngai, L. Xiu, and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert systems with applications*, vol. 36, pp. 2592-2602, 2009.

**[20]** J. Han, M. Kamber, *Data mining: Concepts and techniques*, Morgan Kaufmann Publisher, 2001.

**[21]** C. H. Cheng, and Y. S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert systems with applications,* vol. 36, pp. 4176-4184, 2009.

**[22]** B. Avram, "Airlines Customer Segmentation in the Hyper-Competition Era," *Expert Journal of Marketing*, vol. 7, 2019.

**VITAE**

**Serpil USTEBAY** received Ph.D. degree in Computer Engineering from Istanbul University in 2017. She is currently working as a researcher at Istanbul Medeniyet University. Her research interests are Data Mining, Artificial Neural Network, and Algorithms.

**Ilkay YELMEN** received BS degree in Software Engineering from the Istanbul Aydin University in 2013 and the MS degree in Computer Engineering from the Istanbul Aydin University in 2016. He is currently pursuing the PhD degree in Computer Engineering at the Istanbul Aydin University. His research interests are Machine Learning, Semantic Web and Natural Language Processing.

**Metin ZONTUL** received the BS degree in Computer Engineering from Middle East Technical University in 1993, the MS degree in Computer Aided Design, Production and Programming from Erciyes University in 1996 and Ph.D. degree in Numerical Methods from Cumhuriyet University in 2004. He is currently an Associate Professor at Istanbul Arel University. His research interests are Data Mining, Machine Learning, Information Retrieval and Natural Language Processing.