Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Segmentation of store customers to increase sales using ABC-XYZ-analysis and clustering methods

To cite this article: S A Evdokimova 2021 *J. Phys.: Conf. Ser.* **2032** 012117

View the article online for updates and enhancements.

You may also like

- System simulation application for determining the size of daily raw material purchases at PT XY
H L Napitupulu
- Outliers Detection on Fisheries Commodity Transaction from Local Market in Tual City based on The x-means Clustering
Syahibul Kahfi Hamid, Wellem Anselmus Teniwut, Roberto Mario Kabi Teniwut et al.
- Russian practice of using digital technologies in public procurement management in the construction industry
Vadim Koscheyev and Almaz Hakimov

IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 185.46.118.239 on 21/10/2021 at 18:04

Segmentation of store customers to increase sales using ABC-XYZ-analysis and clustering methods

S A Evdokimova[1]

[1] Voronezh State University of Forestry and Technologies named after G.F. Morozov, 8, Timiryazeva St., Voronezh, 394087, Russia

E-mail: evdsv@mail.ru

Abstract. To manage interaction with customers, it is necessary to divide them into groups depending on their buying activity. For this, the author uses segmentation (ABC and XYZ analysis) and clustering (K-means, X-means, Expectation-Maximization) methods in the intelligent software platform RapidMiner Studio. For calculations, data on sales of a store selling auto parts are used. The author notes such features of the trade in truck parts as the seasonality of maintenance and the irregularity of purchases in comparison to general-purpose goods, which affect the choice of the time interval in sales data. The ABC-XY-analysis showed that it is better to use only the results of the ABC analysis, taking into account the volume of purchases, and the XYZ-analysis based on the analysis of the frequency of purchases gives the wrong result due to the seasonality of purchases. The results of clustering using K-means and X-means methods are almost the same, but their disadvantage is that 95% of customers are classified as the cluster with the worst purchasing activity, and this separation cannot be used for marketing purposes. The Expectation-Maximization method gave the best division of the client base into clusters.

1. Introduction

Modern intelligent information technologies allow solving many data analysis tasks to improve business efficiency and support decision making [1]. Using intelligent technologies, you can analyze large amounts of data from various sources, identify patterns and predict economic performance. These tasks include analysis and forecasting of foreign economic activity, inventory management, customer basket analysis, and others [2-6].

Dividing customers into groups to study buying patterns and optimize customer experience is also an intellectual task. There are various methods and approaches to solving this problem. For example, the methods of customer segmentation based on joint ABC and XYZ analysis [7]. Clustering methods divide customers according to different criteria into an arbitrary number of clusters. The most popular method is K-means and its modified versions [8-11].

We will divide the customer base of the retail store, which sells parts for trucks, into groups. The peculiarities of the trade in goods for cars is the seasonality of technical maintenance of equipment, the amount of buyers' checks can vary by an order of magnitude, there is a fairly large interval between purchases compared to goods of general consumption. Therefore, the task is to use various methods of segmentation and clustering, analyze the results and choose the appropriate option.

To analyze the client base, we will use the RapidMiner Studio analytical system, which is an intelligent platform for analyzing a large amount of data.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Material and methods

The simplest and most popular method of customer segmentation is ABC-analysis, conducted on the Pareto principle - 20% of the effort gives 80% of the profit. The essence of ABC-analysis is to divide the customer base into categories depending on the percentage of the customer's purchases to the total sales:

- Category □ – 80%;
- Category □ – 15%;
- Category □ – 5%.

Working with category A customers is a priority for the company, category B customers are a middle level, and you should not waste time working with group C customers.

To conduct an ABC-analysis of the customer base for each customer, you should calculate the amount of purchases for the selected period, the share of sales from the total sales with a cumulative total, and based on the obtained values, each customer belongs to a certain category.

Often, ABC analysis is supplemented with XYZ analysis based on the segmentation of objects by the values of the coefficient of variation:

$$V = \frac{1}{\bar{x}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{1/2}; \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \tag{1}$$

where $\bar{x}$ is the average value of the studied indicator; x_i is the value of the studied indicator for the period i; n is the number of considered time periods.

If the values of the studied indicator are stable and take values close to the average value, then the coefficient of variation will be small. Buyers are divided into groups depending on the value of the coefficient of variation as follows:

- Group X – from 0 to 10%;
- Group Y – from 10 to 25%;
- Group Z – over 25%.

For example, the results of the ABC-analysis by the volume of customer purchases can be combined with the results of XYZ-analysis, which segments customers by the frequency of purchases made. As a result, there are 9 groups of customers, each of which should have its own management policy.

The disadvantage of using XYZ-analysis is its overestimation of the coefficient of variation due to seasonality or other reasons for unevenness in purchases. In this case, many regular customers with a large volume of purchases may fall into group Z, which will give incorrect results. The same problem can arise when segmenting the customers of the considered auto parts store for trucks.

In the methods above, the number of segments into which the customer base is split is known in advance. In cluster analysis, the number of groups into which a given set of objects should be divided is not known, and also the rules for assigning an object to a specific group are not known. Therefore, the clustering task belongs to the tasks of Data Mining, the solution of which is associated with the search for new data or patterns in a large array of source data. The advantage of cluster analysis is to divide objects into groups by sets of features of various types.

In general, the clustering problem is formulated as follows: the set of objects X is described by a set of parameters. It is necessary to divide it into disjoint sets (clusters) containing similar objects. The values of the object parameters by which the objects are divided into classes are not known. The number of clusters can be specified or they need to be found.

The solution to the clustering problem is ambiguous and depends on the chosen metric of the proximity distance of objects ρ. The Euclidean distance is often used as a metric:

$$\rho(p,q) = \left(\sum_{k=1}^{n}(p_k - q_k)^2\right)^{1/2}; \tag{2}$$

where n – the number of object parameters; p_k and q_k - values of the parameter k of objects p and q.

In clustering theory, there are a large number of algorithms that are divided into hierarchical and non-hierarchical methods. In turn, the hierarchical group includes agglomerative methods, when

clusters are created by sequentially combining individual objects or their clusters into larger ones, and divisive methods, when new clusters are formed by sequentially dividing large clusters into smaller ones. Clustering can be performed based on the maximum, minimum, or average distance.

The algorithms of non-hierarchical methods divide the source objects into k clusters, which are characterized by a dense location near some central points. The most popular are:

- k-means – clusters are formed near centroids. The operation of the k-means algorithm includes the following steps:

  1. Setting the number of clusters k.

  2. Randomly, in the original set of objects, k objects are selected which are considered to be the centers of the clusters.

  3. Distances from each object to the centers of the clusters are calculated by formula (2).

  4. Each object belongs to a cluster whose distance is the smallest.

  5. The cluster centroids are defined for each cluster. Each centroid coordinate is calculated by the formula:

$$\overline{x_i} = \frac{1}{m}\sum_{j=1}^{m} x_{ij}; \tag{3}$$

  where $\overline{x_i}$ is the coordinate i of the centroid, $i \in [1,n]$; n is the number of coordinates (parameters of objects); x_{ij} is the value of the attribute of the j object, $j \in [1,m]$; m is the number of objects.

  Centroids become new cluster centers and steps 3-5 are repeated. The algorithm stops if he coordinates of the centroids and cluster objects have not changed.

- k-mediods – clusters are formed near the medoids (the coordinates of the centroids are shifted to the nearest object). The center of the cluster is necessarily one of the objects. In this case, the algorithm will be as follows:

  1. Setting the number of clusters k.

  2. K objects are randomly selected, which are considered to be the centers of clusters.

  3. Each of the remaining objects belongs to the cluster, the distance to the center of which will be minimal:

$$\rho(p,q) = \sum_{k=1}^{n} |p_k - q_k|; \tag{4}$$

  where p_k and q_k are values of the parameter k of the objects p and q; n is the number of parameters of the objects.

  4. The object m, which is the center of the cluster, is replaced by the object o, which is included in the cluster. In this case, the penalty function is calculated as the sum of the distances, which should be minimized:

$$\cos t(x, md) = \sum_{j=1}^{m} |x_j - md|; \tag{5}$$

  where $cost(x, md)$ is the penalty function; x_j - the value of the attribute of the object j, $j \in [1,m]$, m is the number of objects in the cluster; md is the value of the medoid attribute.

  If the function (5) is incremented, the objects do not change. If the function (5) has decreased, then steps 3 and 4 are repeated. The process stops if no changes have occurred in the set of centroids.

- EM (Expectation-Maximization) – based on the maximum plausibility of parameters. The basic procedure of the EM algorithm includes two steps:

  1. E-algorithm (Expectation step) - the expected value of the likelihood function is determined by the formula:

$$z_{ij} = \frac{p(x_j \mid y_i)p(y_i)}{p(x_j)}; \tag{6}$$

where $p(x_j \mid y_i)$ is the probability of occurrence x_j in the cluster y_i; $p(y_i)$ and $p(x_j)$ - probability of occurrence y_i and x_j; z_{ij} - probability of finding object *j* in cluster *i*.

2. M-algorithm (maximization step) - the maximum likelihood is estimated using the formula:

$$u_i = \frac{\sum_{j=1}^{n} z_{ij} x_i}{\sum_{j=1}^{n} z_{ij}}. \quad (7)$$

The EM algorithm works until the increment of the likelihood function becomes less than the specified value.

In practice, various heuristic or statistical functions are used to compare the results of clustering. For example, the average intra-cluster distance is often used, which should be minimal:

$$Q = \frac{1}{k} \sum_{i=1}^{k} \sum_{p,q \in C_k} \rho(p,q) \to \min; \quad (8)$$

where Q is the intra-cluster distance; k – number of clusters; $\rho(p,q)$ - the distance between two objects *p* and *q* of the C_k cluster.

3. Results and Discussion

We will analyze the customer database of a store selling auto parts for trucks, containing data on sales made by 560 customers. For this, data processing scripts were created in the analytical system RapidMiner Studio.

First, we will conduct an ABC analysis by sales volume, as a result of which 97 buyers will be classified as category A, 146 buyers in category B, and 317 buyers in category C. Thus, 17.3% of customers from the total number of customers of the store provide 80% of the profit, 26.1% of customers – 15% of the profit and 56.6% of customers-only 5% of the profit, i.e. about half of the customers make purchases for small amounts. Group A mainly includes commercial organizations and individual entrepreneurs who have several trucks, in groups B and C - more individuals.

Sales of auto parts depend on the season, commercial organizations usually carry out maintenance of the machines immediately for all equipment. Therefore, the XYZ analysis taking into account the frequency of purchases gives the wrong result (Table 1). A large group of buyers who shop for both large and small amounts is in category Z. Using the results of the XYZ analysis will lead to erroneous management of customer interactions.

Table 1. ABC-XYZ-analysis results.

	X	Y	Z	Total
A	1	-	96	97
B	10	-	136	146
C	132	11	174	317
Total	143	11	406	560

Due to the specifics of car parts sales, we will highlight three parameters in the sales dataset that characterize customer purchases to which we will apply clustering methods:

- Parameter 1 – The amount of purchases for the year, rub;
- Parameter 2 – The number of unique products purchased in a year;
- Parameter 3 – Number of purchases per year.

In many clustering statements in RapidMiner Studio, you need to specify the number of resulting clusters, so first we will perform clustering using the X-Means operator, which heuristically determines the number of clusters. It turned out three clusters, including 529, 26 and 5 buyers (Table

2), the cluster tree is shown in Figure 1.

Table 2. Parameters of cluster centroids using the X-Means method.

Cluster	Parameter 1	Parameter 2	Parameter 3	Number of customers	Average distance to cluster center
Cluster 0	21371.252	6.824	2.055	529	1.83
Cluster 1	297122.692	72.231	9.192	26	1.42
Cluster 2	817254.862	253.800	19.800	5	3.72

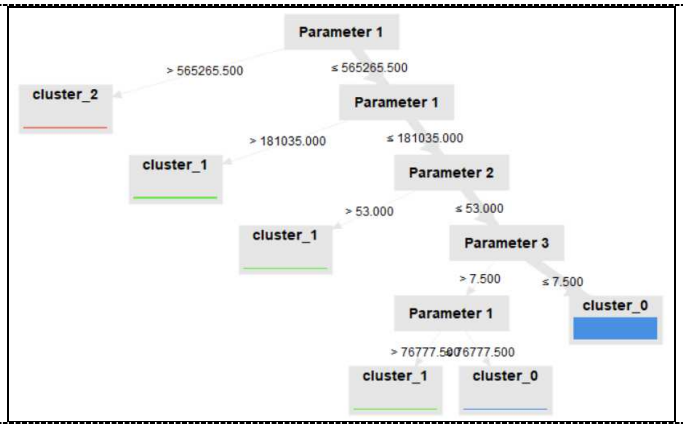

Figure 1. X-Means Cluster Tree.

The results of clustering using the K-means method (Table 3) almost coincided with the results of clustering using the X-Means method, and the application of the EM method gave a completely different division into clusters (Table 4).

Table 3. Parameters of cluster centroids using the K-Means method.

Cluster	Parameter 1	Parameter 2	Parameter 3	Number of customers	Average distance to cluster center
Cluster 0	21844.100	6.951	2.072	531	1.86
Cluster 1	351820.512	82.000	9.538	26	1.58
Cluster 2	790102.000	311.333	25.667	3	2.67

Table 4. Parameters of cluster centroids using the EM method.

Cluster	Parameter 1	Parameter 2	Parameter 3	Number of customers	Probability of belonging
Cluster 0	282282.388	73.834	8.132	49	0.99
Cluster 1	30850.195	9.879	2.707	261	0.99
Cluster 2	3578.269	1.886	1.242	250	1

Let's compare the results:

- Application of clustering methods X-Means, K-Means, and Expectation-Maximization gave the number of buyers with the best activity 5, 3, and 49, respectively;

- the largest number of buyers under the X-Means and K-Means methods is assigned to the group with the worst activity – 529 and 531, which is 95% of the total number of buyers. On average, these customers visit the store twice a year and buy 7 different products;
- according to the EM clustering method, customers with the worst purchasing activity visit the store once a year and buy 1-2 products. There are 250 such clients (45% of the total number).

4. Conclusion

Thus, the buyers of the truck auto parts store were divided into clusters by means of segmentation and clustering methods in the analytical system RapidMiner Studio. According to ABC analysis, customers are divided into groups containing 97, 146 and 317 customers, according to the volume of purchases. Due to the unevenness and seasonality of the purchases made, the XYZ analysis gave an incorrect result.

The clustering results obtained using the K-means and X-means methods cannot be used to manage customer interaction, because 95% of customers are assigned to one cluster. The best division of the customer base was provided by the Expectation-Maximization method, according to which the group with the best buying activity included 49 buyers, the group with the average - 261 and the worst - 250 buyers.

References

[1] Zinoveva I S, Yakovlev A V and Pecherskaya O A 2019 Methods of application of intellectual technologies of decision support for maximizing economic effectiveness of regional economy in the conditions of its sustainable development *Advances in Intelligent Systems and Computing* **726** 337-343 DOI: 10.1007/978-3-319-90835-9_39

[2] Evdokimova S A and Kopylova V S 2017 Analysis of the automation of foreign trade activities organization *Modeling of systems and processes* vol 10 **1** 20–23 DOI: 10.12737/article_5926f7b1761f69.28618968

[3] Gao L and Dou H 2020 Inventory management of railway logistics park based on artificial neural network *Journal Europeen Des Systemes Automatises* **53(5)** 715–723 DOI:10.18280/jesa.530514

[4] González R G, Gonzalez-Cava J M and Méndez Pérez J. A. 2020 An intelligent decision support system for production planning based on machine learning *Journal of Intelligent Manufacturing* **31(5)** 1257-73 DOI: 10.1007/s10845-019-01510-y

[5] Vahidi F M, Etebarian A, Azmi R and Ebrahimzadeh D R 2021 An analytics model for TelecoVAS customers' basket clustering using ensemble learning approach *Journal of Big Data* **8** 36 DOI: 10.1186/s40537-021-00421-1

[6] Tripathi Sh, Bhardwaj A, Poovammal E 2018 Approaches to clustering in customer segmentation *International Journal of Engineering &Technology* 7(3.12) 802–807 DOI: 10.14419/ijet.v7i3.12.16505

[7] Stojanović M and Regodić D 2017 The significance of the integrated multicriteria ABC-XYZ method for the inventory management process *Acta Polytechnica Hungarica* **14** 29–48 DOI: 10.12700/APH.14.5.2017.5.3.

[8] Al Malki A, Rizk M M, El-Shorbagy M A and Mousa A A 2016 Hybrid genetic algorithm with K-Means for clustering problems *Open Journal of Optimization* **5** 71–83 DOI: 10.4236/ojop.2016.52009

[9] Rayala V and Kalli S R 2021 Big data clustering using improvised fuzzy C-means clustering *Revue d'Intelligence Artificielle* **34(6)** 701-708 DOI:10.18280/RIA.340604

[10] Syakur M A, Khotimah B K, Rochman E M S and Satoto B D 2018 Integration K-means clustering method and elbow method for identification of the best customer profile cluster *IOP Conference Series: Materials Science and Engineering*, **336(1)** 012017 DOI:10.1088/1757-899X/336/1/012017

[11] Li L, Zhou X, Li Y, Gu J and Shen S 2020. An improved genetic algorithm with Lagrange and density method for clustering *Concurrency Computation* **32(24)** 101 DOI:10.1002/cpe.5969