Classification of e-commerce customers based on Data Science techniques

Olena Piskunova and Rostyslav Klochko [0000-0003-2690-2785]

Kyiv National Economic University named after Vadym Hetman, Kyiv, Ukraine

EPiskunova@kneu.edu.ua,rostislav.klochko@gmail.com

Abstract. Currently, most organizations are trying to build their data-driven strategies investing heavily in developing their own intelligent decision-making systems. But there are many small online retailers in the economy who are looking to implement business intelligence systems but still lack the necessary knowledge and expertise to do it. The article provides an example of using data science techniques for classification online store customers by their purchasing activity. The analysis of different approaches allowed us to propose the solution of this problem in two stages. At first, we segmented our e-commerce customers by RFM metrics using the k-means method. The algorithms for automated selection of the number of clusters and the initial selection of group centers are applied. There were 6 groups of clients highlighted: first cluster - lost clients; cluster 2 is a new wholesale buyer; cluster 3 - customers that the company may soon lose; cluster 4 active retail buyer; cluster 5 new retail customers; cluster 6 is an active wholesale buyer. In the second stage, with help of machine learning algorithms the customers' classification system was built. The presence of the second stage is conditioned by the need to take into account the constant updating of the client base and accumulation of new information. Tenfold cross-validation was performed to avoid retraining models. The analysis of calcu-?ations by 5 classification methods allowed us to give the advantage of the "random forest" method. To perform the analysis and all calculations this study uses R programming language and RStudio system.

Keywords: clusterization, classification, rfm – model, e-commerce, machine learning, data science.

1 Introduction

Retail is one of the fastest-growing sectors of the Ukrainian economy. Today it is an almost unique market in Ukraine that has a lot in common with perfect competition. There are thousands of players in the segment that realize millions of different products. Most of them are small and medium-sized businesses. Year by year it becomes more and more difficult to win the loyalty of new customers and retain the loyalty of regular customers. Therefore, the ability to offer an individual approach to each of the clients in the coming years will be the only condition for successful business activity.

In view of the rapid economy digitalization, e-commerce is becoming one of the most important areas in retail activity. Despite the fact that Ukraine is far behind the global pace of e-commerce market development, in recent years' Ukrainian online sales growth is even faster than in Europe. Nowadays, e-commerce companies have to refer to clients' wants and need in the decision-making process to meet the requirements of today's economy. At the same time, everyday contact with thousands of customers makes it difficult to consider each of them. The solution to this problem is facilitated by the development of a clear segmentation of the client base, which can be done based on the mathematical modeling methods.

Thus, modeling consumer behavior is an actual problem, which solution will not only improve the efficiency of e-commerce but also contribute to the development of the whole economy and better fulfilling of consumers' needs. In particular, the important task of e-marketing is to classify online store consumers by the level of their purchasing activity. The peculiarities of this task are a large amount of available data and their constant updating and accumulation, which requires the use of Data Science techniques, including machine learning methods.

The goal of this work is to classify online store customers by the level of their purchasing activity based on Data Science techniques, including machine learning methods.

2 Literature review

The majority of Ukrainian scientists' researches is based on the analysis of the client base, which is supported only by a personal understanding of the process [1]. Also, the overwhelming amount of scientific work is based on the socio-demographic statistics of an individual company or the whole country [2]. E-commerce customer activity data is almost not investigated. Recently, the first publications with examples of the application of machine learning methods in marketing have started to appear in the Ukrainian scientific space [3]. But most of them do not take full advantage of these technologies. For example, if cluster analysis methods are used, then the number of clusters is selected based on their own expert judgment [3]. As the analysis of Ukrainian scientific works shows, machine learning algorithms, the RFM model, and the process of automated decision-making are hardly used in them. Even if these technologies are used, they are quite limited. For example, the RFM model is used, but the segmentation is performed manually [4, 5].

Foreign scientific literature has many studies that reveal the peculiarities of the usage of intelligent systems in marketing [6]. Most papers describe the complete process of building an automated customer analysis system which includes: calculating RFM activity metrics, customer clustering using the machine learning methods (e.g. K-means), developing an individual approach for each segment [7,8,9]. However, the methodology for selecting the number of clusters to which necessary to divide the input data is hardly addressed. As a rule, only one method is used - “average silhouette width”, which usually does not allow to solve the problem correctly [10]. It is advisable to decide the required number of clusters, based on the value of 26 additional criteria that can be obtained using the NbClust data analysis package in RStudio [11].

Also, the question of the further efficiency of the built algorithms is almost not solved. Most of the research in this area details the methodology for clusterization's existing customer base, but they do not take into account that new clients are coming every day and current clients tend to change their behavior over time. It is considered appropriate to consider clustering as the first stage of data analysis, which only allows us to understand which customer groups are active, while it is important to have a system for automatically assigning a segment to customers. In scientific research, there are two approaches to solving this problem - fuzzy logic methods [12] or classification algorithms [13]. The analysis of different approaches allows us to give preference to the classification model.

3 Proposed methodology and experiments

The approach proposed in the paper is implemented in 2 stages: the first stage involves customer segmentation by cluster analysis methods; the second stage involves the development of a client classification algorithm that would allow continuously update current clients segment and assign a segment to new customers.

This research is aimed to reduce the human impact in strategic decisions making. Therefore, particular attention is paid to the accuracy and relevance of the proposed methods and algorithms. The number of clusters is selected based on 26 different criteria and indices. For classification task were applied 5 different models with tenfold cross-validations. After that, the most accurate and appropriate algorithm was chosen for implementation.

Note that all calculations are performed using the R-Studio software with R programming techniques.

3.1 Execution of RFM Analysis

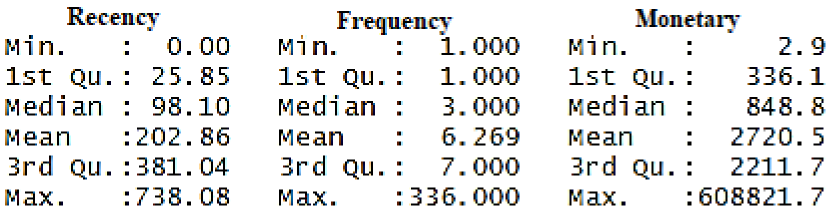
The study was performed on the sample of data from one of the online stores [14]. Data include 1 067 371 transactions of purchase and return of goods during the period from 01.09.2009 to 09.12.2011.

The database contains the following information: Invoice - unique operation code; StockCode - unique product code; Description - the name of the product; Quantity - the quantity of purchased/returned products; InvoiceDate - date of operation; Price - the price of the goods; Customer ID - unique customer code; Country - a country of the operation.

The first task to be addressed in the research process is the selection of criteria for evaluating the level of customer purchasing activity. We will take a classic approach to measure purchasing activity - RMF-model (Recency - Frequency - Monetary) [15].

Recency for each individual customer is calculated as the difference between the actual date in the database and the date of the customer's last purchase. In our case, the metric is measured in days. The frequency of purchases (Frequency) for each individual customer is determined by the number of transactions performed by the client during his client life. Monetary for each individual customer is defined as the total return on all customer transactions during his or her client life. In our case, the metric is measured in dollars.

In the previous research phase, these customer activity metrics were calculated for each customer in the sample. After that, the characteristics of the statistical distributions of Recency, Frequency, Monetary were calculated, namely: average, minimum and maximum values of indicators, as well as 1, 2 and 3 quartiles. The values of these characteristics are shown in Fig. 1.



```
       Recency              Frequency               Monetary
Min.    :  0.00      Min.    :  1.000      Min.    :      2.9
1st Qu.: 25.85      1st Qu.:  1.000      1st Qu.:    336.1
Median : 98.10      Median :  3.000      Median :    848.8
Mean   :202.86      Mean    :  6.269      Mean    :   2720.5
3rd Qu.:381.04      3rd Qu.:  7.000      3rd Qu.:   2211.7
Max.   :738.08      Max.    :336.000      Max.    :608821.7
```

Fig. 1. Distribution of purchasing activity metrics

As you can see, the average customer of our online store had the last purchase 202 days ago. On average the customers buy 6 goods during the client's life while spending $ 2 720.

Further, we will use these indicators as the main metrics.

3.2 Customer segmentation

The process of clustering an online store's customer base relates to Unsupervised Learning algorithms where algorithms do not receive any clues as to the desired result, but rather generate new results based on the data. Unsupervised learning technologies are commonly used at the beginning of the study. The main result of the implementation of these algorithms is to find certain patterns in the available data and to characterize their structure.

The most efficient and simple algorithm for cluster analysis is k-means. This method is very common in economic research, but its practical application for clustering e-commerce customers has some difficulties.

Firstly, the final results are sensitive to the initial random selection of group centers. To solve this problem, a procedure involving multiple executions of an algorithm with different random assignment of initial centroids was applied. An iteration with a minimum value of W_{total} is selected as the final clustering option. Within Cluster Sum of Squares (W_{total}) measures the squared average distance of all the points within a cluster to the cluster centroid [16].

$$W_{total} = \sum_{l=1}^{k} \frac{\rho(C^l)}{n^l}, \tag{1}$$

where $\rho(C^l)$ is the sum of Euclidean distances between points within the cluster l; n^l- number of points in cluster l; k is the number of clusters.

The sum of Euclidean distances between points within cluster l is calculated by the formula:

$$\rho(C^l) = \sum_{i=1}^{n} \rho(X^i, C^l), \quad (2)$$

where n is the number of points in cluster l; C^l is the cent of the weight of cluster l [17].

The second problem is the need to prioritize a fixed number of clusters for partitioning, which is certainly not always chosen to be optimal. Therefore, one of the main tasks of cluster analysis is to select the optimal value of k.

There are several versions of the solution:

- quantity is determined by business needs. This approach is commonly used if there is exist a proven customer classification system in the enterprise segment. An example would be the distribution of customers by their purchasing activity level (Low, Below Average, Medium, High);
- quantity is selected using machine learning algorithms. This approach is used when the decision-maker has no understanding of the typology of their clients. Machine learning algorithms help you to select customer classes based on the level of similarity of their behavior;
- a mixture of the first and second approaches. The most common approach is when a decision is made based both on business understanding and the results of mathematical modeling.

The basic methods of machine learning that help to solve the problem of choosing the number of clusters are the methods of "elbow" and "medium silhouette". The elbow method explores the nature of the W_{total} (1) variation spread with an increasing number of groups k. Combining all n observations into one group, we have the largest intra-cluster variance, which will decrease to 0 as k → n [16].

Another, popular method of assessing the quality of the model is the "Average silhouette width". The value of the silhouette shows how similar the object is to its cluster compared to other clusters.

Suppose that the data were clustered into k clusters. For the point $X^i \in C^l$ (the point X^i is in the cluster C^l), let:

$$a(X^i) = \frac{1}{|C^l|-1} \sum_{X^j \in C^l, X^i \neq X^j} \rho(X^i, X^j), \quad (3)$$

where $a(X^i)$ is the average distance from $X^i \in C^l$ to other objects in the cluster C^l; | $|C^l|$ is the number of objects in C^l clusters.

We can interpret $a(X^i)$ as a measure of how well X^i is assigned to its cluster (the smaller the value, the better the destination).

Then we determine the average dissimilarity of the point X^i to some cluster C^k as the average distance from X^i to all points C^k (where $C^k \neq C^l$). For each data point $X^i \in C^l$, we now define:

$$b(X^i) =_{k\neq l}^{min} \frac{1}{|C^k|} \sum_{X^j \in C^k} \rho(X^i, X^j), \quad (4)$$

where $b(X^i)$ is the smallest average distance X^i to all points of any other cluster, where X^i is not a member.

A cluster with this smallest mean difference is considered a "neighboring cluster" to C^l, since it is the next cluster best suited for the point X^i. Now let's define the silhouette of one data point X^i:

$$s(X^i) = \frac{b(X^i) - a(X^i)}{\max(b(X^i), a(X^i))}, \quad (5)$$

where $a(X^i)$ is the average distance from $X^i \in C^l$ to other objects in the cluster C^l; $b(X^i)$ is the smallest average distance X^i to all points of any other cluster.

The silhouette varies from -1 to +1, where a high value indicates that the object is well-matched to its own cluster. If most objects are of high value, then the clustering configuration is appropriate [18].

By the formula (1) for each number of clusters (from 1 to 10), the level of variance explained by clustering was determined (Fig. 2).

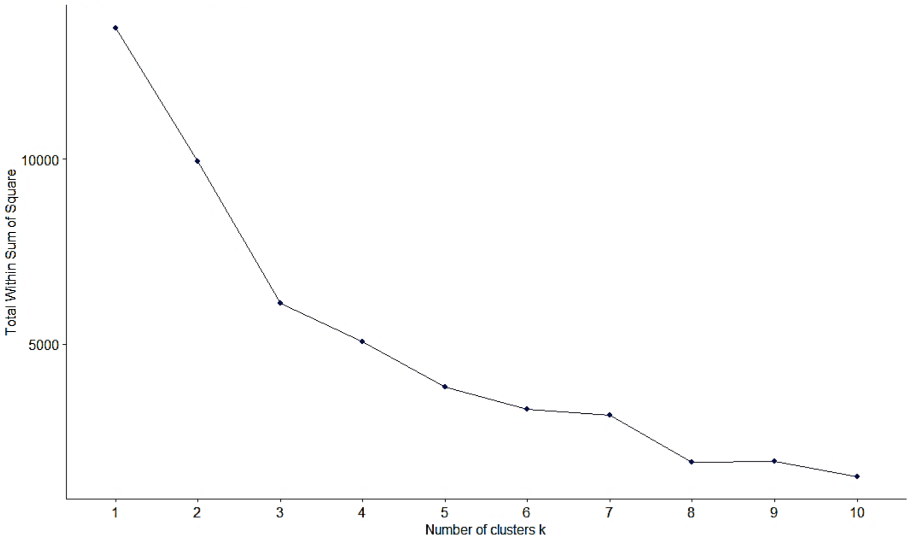


Fig. 2. Graphical implementation of the elbow method

In Fig. 2 we need to identify the breaking point where the drop starts to slow. Points 3, 6 and 8 look most similar to the hacking point, but decisions made on one approach alone are in most cases not accurate.

Therefore, the next step will be a silhouette check. Using formulas (3) - (5), we calculate the value of "silhouette" for each variant of the number of clusters (from 1 to 10). A graphical representation of the calculation results is shown in Fig. 3.

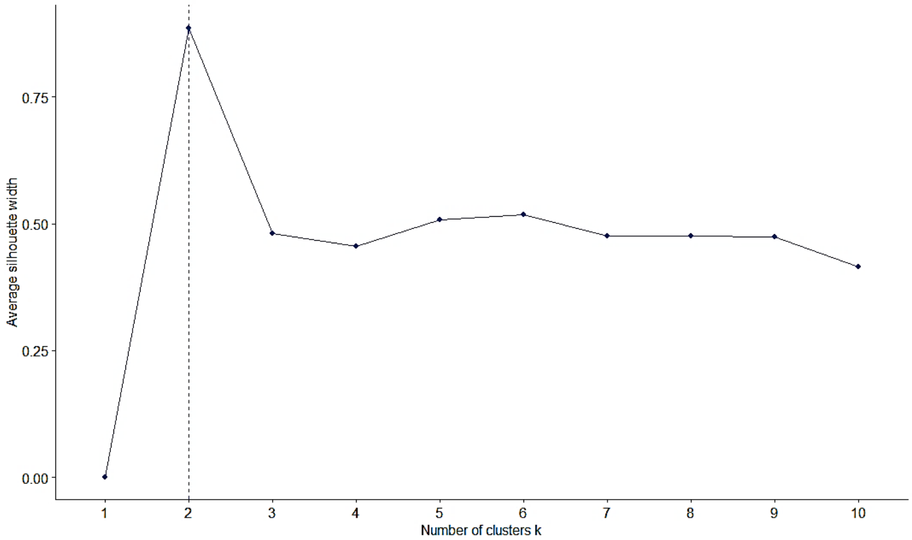


Fig. 3. Graphic implementation of the " average silhouette" method

By this method, we look for the highest value of this indicator. As we can see in Fig. 3, the optimal number of clusters is 2 (6 in second place).

NbClust analytical package was used to refine the results, which allows us to calculate 26 additional criteria.

Table 1. Characterization of indices in the NbClust analytical package [11]

Full name	Short name	Selection criterion
Hubert index. Hubert and Arabie 1985	Hubert	Graphical method
Dindex. Lebart et al. (2000)	Dindex	Graphical method
KL index. Krzanowski and Lai (1988)	KL	Maximum value of the index
CH index. Calinski and Harabasz (1974)	CH	Maximum value of the index
Hartigan index. Hartigan (1975)	Hartigan	Maximum difference between hierarchy levels of the index
Cubic Clustering Criterion (CCC). Sarle (1983)	CCC	Maximum value of the index
Scott index. Scott and Symons (1971)	Scott	Maximum difference between hierarchy levels of the index
Marriot index. Marriot (1971)	Marriot	Max. value of second differences between levels of the index

TraceCovW index. Milligan and Cooper (1985)	TrCovW	Maximum difference between hierarchy levels of the index
TraceW index. Milligan and Cooper (1985)	TraceW	Maximum value of absolute second differences between levels of the index
Friedman index. Friedman and Rubin (1967)	Friedman	Maximum difference between hierarchy levels of the index
Silhouette index. Kaufman and Rousseeuw (1990)	Silhouette	Maximum value of the index
Ratkowsky index. Ratkowsky and Lance (1978)	Ratkowsky	Maximum value of the index
Ball index. Ball and Hall (1965)	Ball	Maximum difference between hierarchy levels of the index
PtBiserial index. Examined by Milligan (1980,1981)	Ptbiserial	Maximum value of the index
Dunn index. Dunn(1974)	Dunn	Maximum value of the index
Rubin index. Friedman and Rubin (1967)	Rubin	Minimum value of second differences between levels of the index
C-index. Hubert and Levin (1976)	Cindex	Minimum value of the index
DB index. Davies and Bouldin (1979	DB	Minimum value of the index
Duda index. Duda and Hart (1973)	Duda	Smallest number of clusters such that index > criticalValue
Pseudot2 index. Duda and Hart (1973) (тільки ієрархічний)	Pseudot2	Smallest number of clusters such that index < criticalValue
Beale index. Beale (1969)	Beale	number of clusters such that critical value of the index >= alpha
Frey index. Frey and Van Groenewoud (1972)	Frey	the cluster level before that index value < 1.00
Mcclain index. McClain and Rao (1975)	McClain	Minimum value of the index
SDindex. Halkidi et al.(2000)	SDindex	Minimum value of the index
SDbw. Halkidi et al.(2001)	SDbw	Minimum value of the index

The results of calculations for each of the indices (Table 1) are presented in the Table. 2. The optimum values for each index are in bold.

Table 2. The value of additional quality criteria for dividing objects into clusters for different number of clusters

Index	Number of clusters					
	2	3	4	5	6	7
KL	0.721	**54.5**	0.037	2.63	0.963	0.148
CH	1645	2755	2525	2857	**2894**	2565
Hartigan	2835	930	1439	862	220	**3241**
CCC	-34.5	-15.7	-17.7	-0.845	**5.32**	-1.10
Scott	5085	8886	**13805**	16936	19114	20893
Marriot	6.7E+10	6.5E+10	**3.9E+10**	3.0E+10	2.7E+10	2.5E+10
TrCovW	1.08E+6	**4.04E+6**	2.99E+6	2.03E+6	1.73E+6	1.68E+6
TraceW	9933	**6102**	5060	3837	3221	3072
Friedman	2.29	4.16	**10.1**	12.7	14.5	19.4
Rubin	1.36	2.22	2.68	**3.53**	4.21	4.41
Cindex	0.0231	0.0215	0.0174	0.0135	0.0122	**0.0116**
DB	1.512	0.971	0.981	**0.899**	0.982	0.956
Silhouette	0.460	0.481	0.455	0.508	**0.518**	0.475
Duda	0.454	**0.974**	1.03	0.852	1.01	1.16
Pseudot2	2651	**72.3**	-53.6	312	-20.5	-241
Beale	**2.05**	0.045	-0.045	0.277	-0.017	-0.197
Ratkowsky	0.308	**0.427**	0.392	0.377	0.355	0.331
Ball	4967	**2034**	1265	767	537	439
Ptbiserial	0.251	0.325	0.327	**0.375**	0.374	0.352
Frey	**-0.902**	0.732	-0.205	0.515	2.215	0.468
McClain	0.541	0.451	**0.645**	0.556	0.574	0.669
Dunn	4.0E-04	**3.0E-04**	1.0E-04	2.0E-04	2.0E-04	2.0E-04
Hubert	1.0E-04	1.0E-04	2.0E-04	2.0E-04	2.0E-04	2.0E-04
SDindex	4.51	**53.8**	41.4	46.4	74.3	65.1
Dindex	0.799	0.741	0.575	0.498	0.466	0.430
SDbw	2.12	**14.6**	11.2	12.4	19.6	17.0

In Fig. 4 presents the number of criteria that supported the corresponding number of clusters. As we can see, the number of clusters in size 3 showed itself best (eight criteria selected this number). The next best option is to have 2 and 6 clusters. We can immediately discard option "2" as it will not bring us any value in future calculations.

Therefore, the main options are 3 and 6 clusters. The next step will be the practical implementation of the k - means method and validation of the results on business logic. The clients were divided into 3 and 6 clusters. Tables 3 and 4 show the average values of RFM metrics for the case of clusters 3 and 6, respectively.

The analysis of Table 3 allows us to give the following interpretation of clusters:

- Cluster 1. This includes customers who, on average, make small purchases every 2 months.

- Cluster 2. This includes wholesale buyers who, on average, purchase a large number of goods once a month for a considerable amount.
- Cluster 3. Here, retail customers make an average purchase once a year.

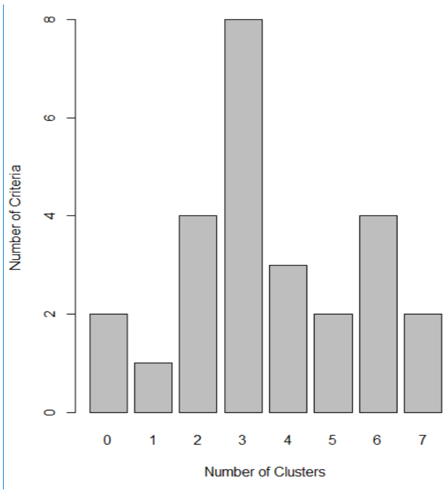


Fig. 4. The number of criteria that support the corresponding number of clusters

Table 3. Results of division into 3 clusters

Cluster	Recency	Frequency	Monetary
1	58	7.17	2 775
2	30	84.1	78 231
3	291	2.13	704

Table 4. Results of division into 6 clusters

Cluster	Recency	Frequency	Monetary	Number of clients
1	424	1.54	498	723
2	26	57.8	46 400	34
3	218	2.71	889	1 615
4	32	19.8	8 528	345
5	54	4.65	1 603	1 794
6	33	145	135 430	7

As can be seen from Table 4, the resulting clusters characterize the following types of clients:

- Cluster 1. Lost clients - Has made less than 2 purchases, the last of which was over a year ago.
- Cluster 2. New wholesale buyer - high average check and activity, but it's been a while since the first purchase. Efforts must be made to increase customer loyalty to the business.

- Cluster 3. Customers whose we will lose soon. They showed typical activity, however, a long time has passed since the last purchase. We should pay attention to these customers and try to persuade them to do more frequent operations.
- Cluster 4. Active retail buyer - high activity, buys for a long period, average check. The most valuable and loyal type of customer for the business.
- Cluster 5. New retail customers - high activity, but during a short period, average check. Efforts should be made to turn them into regular customers.
- Cluster 6. Active wholesale buyer - high average check and activity, buys over a long period. The most profitable type of customers.

Given the business logic, it was decided that the division into 6 customer groups is more acceptable and better characterizes the current situation of the functioning of the online store. The number of clients in each cluster is shown in Table 5.

3.3 Classification of online store customers based on machine learning methods

The next step after customer base segmentation is to build classification models for the distribution of e-commerce customers by these segments. The classification is the task of dividing the set of observations or objects by the values of certain attributes into a priori given groups called classes. Within each of these groups, objects are considered to be similar to each other [19].

The most common machine learning methods for classification are Linear discriminant analysis (LDA); Support vector machine (SVM), Classification and regression trees (CART), k - nearest neighbors (KNN), Random forests (RF).

Discriminant Analysis is a kind of multidimensional data analysis designed to solve random pattern recognition problems. It is used to decide what factors divide (“discriminate") certain data sets (so-called "groups").

SVM (support vector machine) is a set of similar supervised learning algorithms used for classification and regression analysis tasks. A feature of the reference vector method is the constant reduction of the empirical classification error and the intention to increase the distance, so this method is also known as the maximum distance classification method [20].

The decision tree develops solutions with the help of a tree model. The algorithm splits the sample into two or more homogeneous sets (branches) based on the most significant differentiators of the input variables. To select a differentiator (predictor), the algorithm takes into account all the features and makes a binary partition. He then selects the lowest cost option (the highest precision) and repeats recursively until the successful partitioning of the data across all branches (or reaches the maximum depth).

The Classification and Regression Tree (CART) is one of the implementations of the decision tree. Periodic nodes of trees of classification and regression are root and internal nodes - branches. The end nodes are leaf nodes. Each periodic node represents one input variable (x) and a splitting point on that variable; leaf nodes represent the output variable (y). The model is used to predict the following algorithm: it is necessary to go through all the splits of a tree in order to reach the node "leaves" and deduce the value present in it.

Random Forest (RF) is an ensemble model that builds several trees and classifies objects on a "vote" basis. That is, the object belongs to the class that has the majority of votes from all the trees. The algorithm trains several decision trees on different subsamples of data and uses the average to improve model prediction accuracy.

The K-Nearest Neighborhood Classification (KNN) algorithm assumes that objects are divided into different classes so that they can be classified based on their similarity. The distance between the objects may be a measure of similarity. KNN does not need a training phase, it is trained in the sense that it begins to classify data points at once, based on the majority of votes of its neighbors. The object is assigned the class that is most common among its k nearest neighbors. [21]

The data set was divided into training and test samples (75% and 25% respectively). As a result, the training sample includes data about 3388 clients, and the test sample - 1130 clients. "Accuracy" was used to evaluate the quality of the simulation, which is the ratio of correctly distributed customers to their total. The clients are classified according to the 5 methods presented above (LDA, CART, KNN, SVM, RF). Tenfold cross-validation was applied during the implementation of the customer classification algorithm. It is necessary to test whether the simulation results are dependent on a particular dataset. A ten-fold test involves splitting the sample into ten randomly selected sets (test and training samples) and testing the model built on them.

Table 5 shows the characteristics of the Accuracy distributions (minimum, maximum and average values, as well as 1, 2, and 3 quartiles) obtained from the training sample for each method.

Table 5. «Accuracy» distribution performance values (training sample)

	Min	1st Qu	Median	Mean	3rd Qu	Max
LDA	0.95	0.97	0.97	0.97	0.97	0.98
CART	0.90	0.91	0.91	0.92	0.94	0.97
KNN	0.90	0.93	0.94	0.93	0.94	0.95
SVM	0.98	0.98	0.99	0.99	0.99	0.99
RF	0.99	0.99	0.99	0.99	0.99	1.00

As can be seen from Table 5, the RF model showed the smallest error in the training sample (mean 0.99). Using a built "random forest" model, we check it on a test sample. The results of the customers' distribution by classes are presented in Table 6.

On the test sample, this algorithm showed an accuracy of 99%, so RF was chosen to implement the classification process for the entire data sample (Table 7).

Each cluster characterizes a specific group of customers that are similar in purchasing activity. At the same time, clients have a significant difference between clusters. A graphical representation of the difference between the level of purchasing activity in different clusters is shown in Fig. 5-7.

Table 6. Random forest classification results

Segment		Real					
		1	2	3	4	5	6
Forecast	1	183	0	0	0	0	0
	2	0	3	0	0	0	0
	3	0	0	412	0	2	0
	4	0	1	1	85	0	0
	5	0	1	0	0	440	0
	6	0	0	0	0	0	2

Table 7. Dividing of all existing customers by random forest method

Cluster	Name	Number of clients
1	Lost clients	723
2	New wholesalers	32
3	Almost lost	1 616
4	Active retail	347
5	New retail	1 793
6	Active wholesalers	7

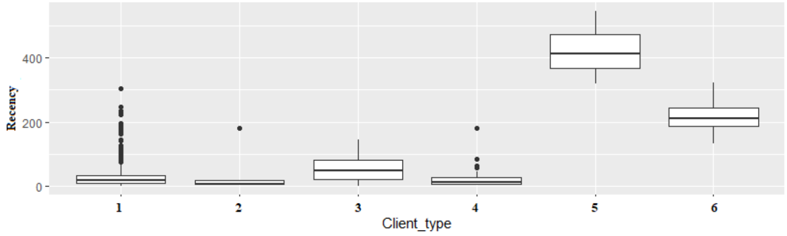


Fig. 5. Boxplot of Recency metric by client type

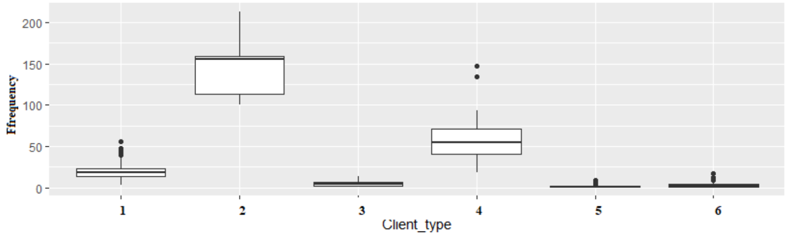


Fig. 6. Boxplot of Frequency metric by client type

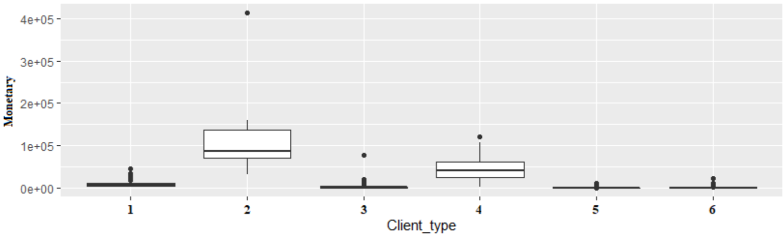


Fig. 7. Boxplot of Monetary metric by client type

4 Conclusion

The paper deals with the task of classifying online store customers by their purchasing activity based on Data Science techniques, including machine learning methods. The analysis of different approaches allowed us to propose the solution of this problem in two stages.

First, the customers of the online store were segmented according to the k-means method by RFM indicators, using algorithms for automated selection of the number of clusters and initial centroids.

There were 6 customer groups found: cluster 1 lost clients - Has made less than 2 purchases, the last of which was over a year ago; cluster 2 new wholesale buyer - high average check and activity, but it's been a while since the first purchase; cluster 3 customers whose we will lose soon; cluster 4 active retail buyer - high activity, buys for a long period, average check; cluster 5 new retail customers - high activity, but during a short period, average check; cluster 6 active wholesale buyer - high average check and activity, buys over a long period.

The second step of the classification procedure, which is already directly carried out the classification of customers, due to the need to take into account the constant updating of the client base and the accumulation of new information. The analysis of calculations by 5 classification methods allowed us to give advantages of the "random forest" method.

References

1. Pursky O., Masokha D. Method of building a network of storefronts of online stores based on MVC architecture // Business Inform. - 2017. - №10. - P. 319–324. (in Ukrainian)
2. Kondruk N. "Using a longitudinal measure of similarity in clustering problems" Radio electronics, informatics, control, no. 3 (46), 2018, p. 98-105. doi: 10.15588 / 1607-3274-2018-3-11 (in Ukrainian)
3. Roskladka N., Roskladka A., Dzigman O. Cluster analysis of the client database of enterprises of the service industry. Economy and management of the national economy. International Economic Relations. No. 2 (35), 2019. p. 151-159 (in Ukrainian)
4. Matsuka V. Marketing technology of forming consumer loyalty in the tourist services market / V. Matsuka, A. Balabanyts // Bulletin of the Mariupol State University. Series:

Economics: Coll. of sciences. wash / goal ed. KV Balabanov. - Mariupol, 2017. - Issue. 14. P. 177–187. (in Ukrainian)

5. Shulgina L. "Methodical instructions on the application of analysis and quality assessment of tourist services" Business Inform, no. 3 (482), 2018, pp. 180-185. (in Ukrainian)
6. Kamthania, Deepali & Pahwa, Ashish & Madhavan, Srijit. (2018). Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business. Journal of Computing and Information Technology. 26. 57-68. 10.20532/cit.2018.1003863.
7. Chen, D., Sain, S. & Guo, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. J Database Mark Cust Strategy Manag 19, 197–208 (2012). https://doi.org/10.1057/dbm.2012.17
8. Dogan, Onur & Ayçin, Ejder & Bulut, Zeki. (2018). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. International Journal of Contemporary Economics and Administrative Sciences. 8. 1-19.
9. Ait daoud, Rachid. (2015). Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 9. 1795 - 1805.
10. P. Anitha and M. M. Patil, RFM model for customer purchase behavior using K-Means algorithm, Journal of King Saud University –Computer and Information Sciences,https://doi.org/10.1016/j.jksuci.2019.12.011
11. Charrad, Malika & Ghazzali, Nadia & Boiteau, Véronique & Niknafs, Azam. (2013). An examination of indices for determining the number of clusters: NbClust Package.
12. Ansari, Azarnoush & Riasi, Arash. (2016). Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms. International Journal of Business and Management. 11. 59. 10.5539/ijbm. v11n7p59.
13. Mathivanan, N.M.N. & Md.ghani, N.A. & Mohd Janor, Roziah. (2018). Improving classification accuracy using clustering technique. Bulletin of Electrical Engineering and Informatics. 7. 465-470. 10.11591/eei. v7i3.1272.
14. Online Retail II Data Set URL: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
15. Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). A review of the application of RFM model. African Journal of Business Management December Special Review. 4. 4199-4206.
16. Shitikov V., Mastitsky S. Classification, regression, Data Mining algorithms using R. URL: https://ranalytics.github.io/data-mining/101-Partitioning-Algos.html. (in Russian)
17. Bertagnolli N. Elbow Method and Finding the Right Number of Clusters. URL: http://www.nbertagnolli.com/jekyll/update/2015/12/10/Elbow.html.
18. Lengyel, Attila & Botta-Dukat, Zoltan. (2018). Silhouette width using generalized mean - a flexible method for assessing clustering efficiency. 10.1101/434100.
19. Lavrenyuk M. An overview of machine learning methods for the classification of large volumes of satellite data / [MS. Lavrenyuk, OM Novikov]; Systems research and information technology. 2018. №. 1. P. 52-71. (in Ukrainian)
20. Soofi, Aized & Awan, Arshad. (2017). Classification Techniques in Machine Learning: Applications and Issues. Journal of Basic & Applied Sciences. 13. 459-465. 10.6000/1927-5129.2017.13.76.
21. Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.