ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

C
CONSUMER BEHAVIOR CLUSTERING OF FOOD RETAIL CHAINS BY MACHINE LEARNING ALGORITHMS

Olena Liashenko[1*1], Tetyana Kravets[2], Matvii Prokopenko[3]

[1, 2,3] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

e-mails: [1]lyashenko@univ.kiev.ua, [2]tankravets@univ.kiev.ua, [3]matveyy58@gmail.com

Received: 10 July 2021 ; Accepted: 04 August 2021 ; Online Published: 08 August 2021

ABSTRACT

Analysis of the behavior of an economic agent is one of the central themes of microeconomics. Right now, with the comprehensive increase in the amount of data and the expansion of the computing capabilities of personal computers, there is a need to implement methods of behavioral economics in the study of human behavior. In the course of this study, a survey was created aimed at identification of patterns of behavior of the modern consumer according to his selection criteria stores and reactions to questions based on Behavioral Economics theorems. Clustering the obtained results were performed using machine learning algorithms, after which the Random Forest classification algorithm was trained. According to the results of Silhouette analysis, K-means clusters were selected as the main ones for further modeling. T-SNE algorithms, hierarchical and spectral analysis were used for additional visual representation. This study offers a tool for classifying customer preferences and analyzing current industry trends. A tool has been created to classify consumers of food retail chains in order to improve their "buyer's journey" and better understand their needs. The created tool for clustering and classification by machine learning methods can be used in business processes. To improve the result, it is necessary to choose a more representative sample, because used in this study consists of an average rationally thinking and knowledgeable individuals, which cannot be said of the average consumer not only among the older generation but also among the younger. Therefore, the next directions in the study may be to identify new ones behavioral trends in other industries; deepening understanding of food retail; use of geodata to improve analysis, etc. Potentially attractive the direction may be to add an assessment of the impact of network advertising on behavior consumers through semantics analysis and image recognition.

Keywords: clustering, machine learning algorithms, food retail, behavioural economics, consumer behavior

JEL classification: C83, C890, D120, D910

Citation:
Liashenko, O., Kravets, T., Prokopenko, M. (2021). Consumer behavior clustering of food retail chains by machine learning algorithms. *Access to science, business, innovation in digital economy*, ACCESS Press, 2(3): 234-251. https://doi.org/10.46656/access.2021.2.3(3)

INTRODUCTION

Currently, players in the food retail industry are rapidly recovering from the pandemic, testing new formats and approaches. It is important to identify fundamental changes in consumer needs for better business adaptation. The food retail industry, both worldwide and in Ukraine, was hit hard during the pandemic, but that allowed it to transform it into a new and better version of itself. First, the e-commerce segment has grown extremely rapidly - networks have had to build entire infrastructure in a few weeks to improve the consumer experience. Leading chains have launched their own delivery services (Silpo) or developed an ecosystem of

[1] Corresponding author Olena Liashenko – lyashenko@univ.kiev.ua

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

stores (Auchan Pick-Up Point). Due to the increasing role of location, the segment of "home shops" has grown significantly, which led to the opening of mini-analogues of supermarkets - Varus-To-Go, Novus Mi and the above-mentioned Pick-Up Point from Auchan).

According to a study by Deloitte (2021), Ukrainians began to spend less on alcohol and more on medicines. The segment of online shopping and the total cost of products on the Internet has grown, and most purchases are now paid by card or gadget (Fig. 1). It is also noted that 73% of Ukrainians use delivery services, but their own experience and the results of the study suggest that this figure is much lower, especially in the regions of the country. However, one fact remains the same - the niche of delivery services in Ukraine is developing rapidly with the entry of new players and the establishment of internal logistics networks of leading stores.

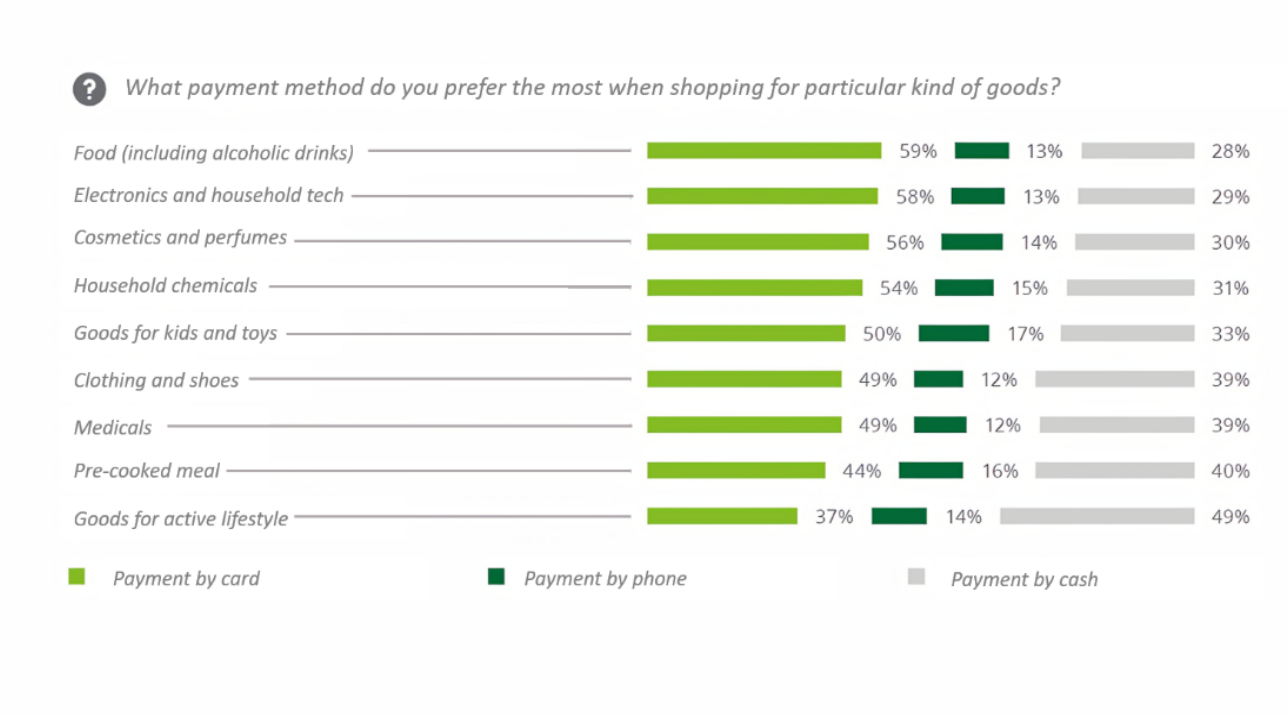


Figure 1. Contactless payment by Deloitte (2021)

Among the changes in the industry in 2021, McKinsey (2020) highlighted the development of clearly differentiated formats (as opposed to conventional supermarkets); higher requirements for budget products; trend towards healthy eating and eco-awareness. And, of course, the powerful development of analytics and new methods of collecting and processing consumer information. For example, specially selected promotions in loyalty applications based on purchases, or the use of customer order patterns through the internal delivery service for notifications at certain times (such information is not widely disclosed by networks, so these applications are theoretical assumptions, although quite likely). In addition, respondents often noted that they choose stores by price and convenience of location (Fig. 2).

Thanks to the above, food retail, like almost all industries, is beginning to take wide strides towards the total personalization of "consumer travel". This term means the general experience of the client from the moment of the beginning of use of any service till the end of the current session (in this case it is stay in shop

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

and use of applications of loyalty programs). All these changes make the tool proposed in this study relevant and possible for use in business processes.

Analysis of the behavior of an economic agent is one of the central themes of classical microeconomics. However, a strong focus on standardized examples did not allow researchers to draw real conclusions. The emergence of economic psychology in the works of Gabriel Tarde, George Cato, Laszlo Garai allowed the modification of purely theoretical models. The real impetus for the emergence of behavioral economics was the work of Daniel Kahneman and Amos Tversky - they developed alternative models of behavior of economic agents, taking into account the inconsistency of their adherence to the principle of maximizing utility.

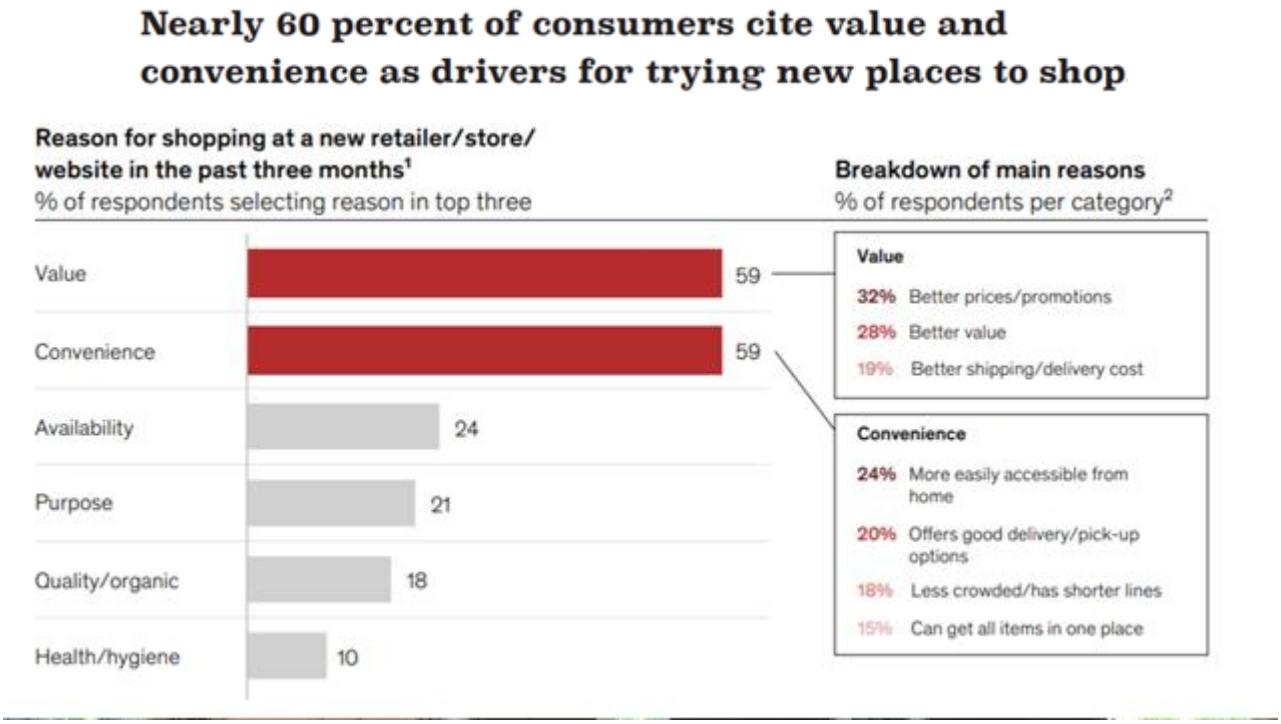


Figure 2. Factors for choosing a new place to shop by McKinsey (2020)

Right now, with the comprehensive increase in the amount of data and the expansion of the computing capabilities of personal computers, there is a need to implement methods of behavioral economics in the study of human behavior. This is also evidenced by the work of recent years - for example, (Kolumbus&Noti, 2019).

Consumer habits are the most important external factor influencing retail, which must be taken into account when formulating a strategy. The process of consumer decision-making to purchase goods is influenced by various factors: economic, cultural, social, personal, psychological and situational factors. Such factors include the global pandemic, which has globally changed the lives of consumers.

This study offers a tool for classifying customer preferences and analyzing current industry trends. Determining current industry trends through the eyes of consumers is done using clustering tools that are applied to respondents' responses to an online survey based on the principles and theorems of behavioral

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

economics. The survey took place in December 2020 during Late Covid. According to the results of the survey, the classification was carried out using machine learning algorithms. And also, the Random Forest in-depth learning method was used to classify consumer preferences.

ANALYSIS OF RECENT RESEARCH

Behavioral economics is a branch of economic theory that is based on the psychological characteristics (cognitive, emotional, social aspects) of human actions, decisions and perceptions in various economic situations (Della Vigna, 2018). For some time, thanks to classical economic theory, it was believed that most actors always choose the optimal course of events. However, there are now studies that emphasize that 80% do not follow such a strategy (İnaç, 2019).

In 1979, D. Kahneman and A. Tversky proposed the theory of perspectives, which explained the irrational decisions of the individual. It has been shown that with a medium level of risk, people tend to maintain their current financial position rather than increase it. And even a paradox was revealed: individuals tend to take greater risks to avoid losses. At the same time, most will not try to benefit more by taking on an additional level of risk. It is these scientists who have identified the most common standard types of human behavior (heuristics), namely: similarity (based on self-confidence, the illusion of control, bias); availability (decision-making based on recently received information or on that which is most vividly reflected in memory) and anchors (the foundation is the basic assessments, perceptions, experience) (Lekovic, 2019).

Behavioral economics argues that man is able to act "limited by the rationalist method", while being influenced by many other factors: the behavior of others; level of economic, social and psychological satisfaction, etc. As the famous economist Richard Thaler proved in 2017, even the simplest thing, such as the "fly in the toilet" can dramatically change the type of behavior of the individual (de Arruda et al., 2015).

To create the survey, the effects of behavioral economics were selected, which clearly describe the specific characteristics of the individual's behavior. The first is Risk Aversion (Reyes et al., 2019). It describes the behavior of individuals who, when faced with uncertainty, try to reduce its degree. In practice, this is manifested when a person chooses an option with smaller achievements, but one that gives confidence in obtaining them, avoiding a riskier, but one that promises greater achievements. The opposite effect is "Risk-Seeking". It describes the actions of an individual who, on the contrary, in most cases will choose the option with a higher degree of uncertainty.

The second effect is Disposition Effect (Reyes et al., 2019). A classic example is an investor who sells stocks that have risen in price but hold those that have lost some of their value. Thus, this effect shows that the individual is much less satisfied with additional income than suffers from excessive losses.

The third effect is Loss Aversion (Reyes et al., 2019). It notes that the individual will make more effort to avoid losses than to receive equivalent additional income. This differs from previous effects in that the expected utility depends on events that have already occurred or are expected in the future.

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The fourth effect is Mental Accounting (Reyes et al., 2019). It is that individuals perceive the value of the same amount of money differently, depending on the situation and subjective opinion. Therefore actions are often based not on the present optimality, and on subjective.

The fifth effect is Subjective Probability (Reyes et al., 2019). It consists in the fact that the individual determines the possibility of the occurrence of an event based on his own empirical experience or ephemeral instinct (the so-called "sixth sense"). Thus, no calculations take place, but a person's confidence in his own rightness remains high.

The sixth and final effect is "Big Fish Little Pond" (Marsh, 2005). It is that when they are in an environment where they are the best at something (or have the most pronounced trait), individuals feel better in the short term. However, in the opposite situation (the worst in a particular case, the least pronounced trait), individuals receive greater benefits in the long-term game.

METHODOLOGY

In this work we used machine learning algorithms that learn "without a teacher" - k-means, spectral clustering, hierarchical clustering, t-sne. Additionally, algorithms such as principal component analysis (PCA) and MinMaxScaler were used for data reprocessing. The Random Forest Classifier algorithm, which works on the basis of a combination of decision trees, was used to classify consumers. All algorithms were used from the library of the scikit-learn (sklearn) Python programming language (Ahuja et al., 2020).

The basis of clustering algorithms is minimization in different ways the distances between groups of datapoints in order to highlight the groups as close as possible to each other. Most often, the data are divided into training and test samples to verify the accuracy of models in a completely unfamiliar situation and to avoid retraining (giving increased level of importance to characteristics that are clearly manifested only in training data).

K-means selects K centroids randomly or from those suggested by the user. After that, the algorithm calculates the distance from each point in the dataset to the centroids, updates the position of the centroids, putting them in the position of the arithmetic mean of the points of the cluster that they describe. In the following steps, the algorithm lists the points belonging to the clusters and again updates the position of the centroids according to the formula $V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \gamma_i)^2$, where де k is the number of clusters, S_i are the obtained clusters, $i \in \{1; k\}$, γ_i are the centers of mass of all vectors x from the S_i cluster.

The stage when the position of the latter remains unchanged is called the final - it becomes the basis for clusters, which gives the model output. To improve the operation of the algorithm, it is recommended to use methods to reduce the dimensionality of the data, such as Principal component analysis (PCA) (Roweis, 1998).

PCA is used to display a multidimensional data array on a two-dimensional plane or to improve the performance of clustering algorithms. It minimizes the sum of the squares of the deviations of the vectors x_i

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

from the linear spaces L_k by the formula $\sum_{i=1}^{m} dist^2(x_i, L_k)$, $dist^2$ is Euclidean distance. The method requires a predetermined number of principal components. PC1 absorbs the vector of the largest variance of dataset characteristics, PC2 - the second largest, and so on. Thus, according to the Pareto principle, the first two components must absorb most of the data information - but they are not informative in themselves (Von Luxburg, 2007).

Spectral clustering is an algorithm that, unlike K-means, works better with non-standard data (such as a spiral). It uses a pairwise similarity matrix to pre-reduce the dimension, such as a symmetric normalized Kirchhoff matrix given by the formula $L^{norm} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is diagonal matrix $D_{ii} = \sum_j A_{ij}$. This is followed by a process of recalculating the matrix until it is possible to increase the similarity of the two points without reducing their similarity with others [(Murtagh&Contreras, 2012).

Hierarchical clustering works iteratively, where in the first step each point is its own cluster. At each subsequent step, the algorithm merges the two nearest clusters until one cluster remains. At the same time the method of singular communication is used $R_{\min}(U,V) = \min_{u \in U, v \in V} \rho(u,v)$, where $R(U,V)$ is distance between clusters, $\rho(u,v)$ is the distance between the points of these clusters. At the output you can get a graphical interpretation of all steps, which potentially allows the researcher to identify hidden common features of different clusters and analyze how many groups to choose (at which step to set the stop signal) (Roweis, 1998).

T-SNE - map clustering algorithm and dimension reduction. At each step, the Gaussian distribution of neighbors is modeled for each point. The task of the algorithm is to find a position in two-dimensional space that minimizes the distance between the t-distribution built on it and the previous Gaussian distribution at all their points based on the Shannon entropy. Like PCA, the algorithm is often used to visualize multidimensional datasets (Wattenberg et al., 2016).

To improve the performance of clustering algorithms, the MinMaxScaler procedure was used, which normalizes the data by minimum/maximum. This avoids further algorithms to reduce the dimensionality of the data giving more weight to the characteristics with greater variance.

The next step was to use Principal Component Analysis to be able to visualize a multidimensional dataset on a two-dimensional graph and simplify further work with it. The elbow method and the silhouette coefficient were used to determine the optimal number of clusters (Yadav, 2017).

The "elbow method" is to select the number of clusters that correspond to the largest positive change in mathematically explained oscillations (explained variation). That is, it is the choice of the point where an additional explanation of the model of oscillations in the data is not worth highlighting additional clusters. This is because the more clusters, the easier it is to explain all the connections in the dataset - but at some stage the model begins to relearn, highlighting purely specific to the training sample characteristics. Alternatively, instead of explained variation, the inertia index can be used - the sum of the quadratic distances of points to their nearest cluster centers (Yadav, 2017).

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

Silhouette score (SI) assesses how the points collected in one group are identical in characteristics. It is calculated by the formula: $\frac{b-a}{\max(a,b)}$, if the number of clusters is greater than 1, where a is the average intracluster distance; b is the average near-cluster distance (average of the distances from the point to all others in the nearest cluster). If the cluster is only 1, then SI is 0. The final value varies within (-1;1), where 1 means the most clearly demarcated clusters; 0 - clusters with points that are very close to the boundaries of neighboring clusters; and a negative value indicates that an error has crept into the process of assigning observations to the clusters (Yadav, 2017).

The modern method of solving classification and regression problems is a set of decision trees, the so-called Random Forest (Liaw&Wiener, 2002). In contrast to the previously described algorithms, this is an example of "guided learning" or "learning with a teacher". This means that in addition to the mandatory division of the sample into training and test, it is necessary to label educational data.

The basis of any variation of the random forest algorithm is an ordinary decision tree. A random forest is a collection of N decision trees, each of which receives an incomplete random set of characteristics and builds its own forecast or classification. The final option is one that has been chosen by most trees (Liaw&Wiener, 2017).

All committee trees are built independently of each other according to this procedure:

- A random subsample with a repetition of size N from the training sample is generated. (Thus, some examples will get into it several times, and about $N/3$ examples will not be included in it at all).

- A decision tree is constructed that classifies examples of such a subsample, and during the creation of the next node of the tree, the feature on the basis of which the partition is performed is selected not from all M features, but only from m randomly selected. The choice of the best of these m features can be done in different ways (for example, the criterion of information growth is used).

- The tree is built until the subsample is completely exhausted and will no longer undergo the procedure of pruning.

The classification of objects is done by voting: each tree of the committee assigns the object to be classified to one of the classes, and wins the class for which the largest number of trees voted.

RESULTS

An online survey based on the principles of behavioral economics was created. A total of 141 respondents passed it. The works of D. Kahneman and A. Tversky, in particular (Kahneman&Tversky, 1979) and (Kahneman, 2011), were taken as the theoretical basis for creating this survey. In addition, materials from the online course Behavioral Finance from Duke University on the Coursera platform, podcasts “Freakonomics Radio” and “No Stupid Questions” and articles (Roweis, 1998), (Prince, 2018), (Beerbaum&Puaschunder, 2018), (Hrnjic&Tomczak, 2019), (Najdenovska et al., 2018), (Dyer&Kolic, 2020), (March, 2019), (Bechara, 2002) were used.

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The final version of the survey consisted of 26 questions, where questions 1-5 gave general information about the respondent, questions 6-15 focused on behavioral drivers (Kahneman, 2003), (Kahneman, 2011), questions 16-26 - on food retail trends and consumer habits.

The first and second questions characterize riskiness or prudence. This is an example of the Risk Seeking effect or Risk Aversion (Stefansson&Bradley, 2017). In the question, it is suggested to play a game where you can either take a guaranteed amount of money (option A), or toss a coin, and with a probability of 50% get a bigger prize or nothing (option B). In total, out of 141 respondents in the first question, 66% chose option A. In the answers to the second question, in which the amounts were 10 times larger, the effect of Risk Seeking is even more noticeable - 82.3%.

The third and fourth questions are examples of Disposition Effect (Barberis&Xiong, 2009). In the third question, the respondent is given 7,000 hryvnias and is offered to either receive 3,000 hryvnias guaranteed (A), or toss a coin and with a probability of 50% receive 6,000 hryvnias (B). In the next question, the initial rate is higher, because 1000 hryvnias is added to the capital left over from the previous question, and the options are as follows: losing 6,000 hryvnias is guaranteed (A), or tossing a coin and losing 12,000 hryvnias with a 50% probability (B).

Most respondents choose 3-A (3,000 hryvnias guaranteed) and 4-B (lose 12,000 hryvnias or nothing). In this case, if you choose 3-A (get 3000 hryvnias guaranteed) and 4-A (lose 6000 hryvnias guaranteed), then in both cases the prize will be 12 000 hryvnias. Also, if the respondent chooses 3-B and 4-B, the win is ranked from the negative (loss of 4,000 hryvnias) to 14,000 hryvnias. The effect is that most people tend to become "risk-free" when it comes to being able to gain something, and "risky" when it comes to losing something (Johnson et al., 2006) The percentage distribution in the questions is as follows: 61.7% (A) in the third and 50.4% (B) in the fourth.

The following questions demonstrate such effects as Mental Accounting (Tversky, 1986), Subjective Probability (Machina&Schneider, 1992) and Big Fish Little Pond (Roweis, 1998), (Marsh et al., 2008). The first shows how different individuals give different subjective weight to the same amount of money. Two situations of buying a ticket to the cinema were presented in the question. In one individual buys a ticket in advance and loses it. In this situation, 51.1% of respondents said they would not buy a ticket again. In another, the individual loses money on the ticket amount. In this case, 53.9% of respondents said they would buy a ticket. Thus, in both cases the same amount was lost, but the perception of this loss is different.

The Subjective Probability effect is based on the fact that the individual, assessing the probability of a particular event, often does not make calculations, but refers to their own feelings or experiences. One historical example is the mass phobia of aircraft after 9/11, although according to statistics, driving a car is a much more dangerous choice. When asked where poker players with the same experience but different winnings over the last few games are represented, 81.6% of respondents correctly stated that everyone has an equal chance of winning the next game.

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The Big Fish Little Pond effect is based on a constant choice of environment. Being the smartest / sexiest can be quite a tempting option, but in the long run it is more profitable to be a "little fish", because it gives more opportunities for development. In answering the question with mental injection, 78.7% of respondents chose the option of becoming 10% smarter (but everyone around them will be smarter than them) instead of becoming 5% smarter (and smarter than everyone else). When it comes to earnings, 82.7% of respondents chose to earn 35,000 hryvnias (all others earn at least 40,000 hryvnias) instead of earning 30,000 hryvnias (all others earn no more than 25,000 hryvnias). Finally, when it comes to using an experimental drug to increase sexuality, 75.2% of respondents chose the assessment of their expected condition after the procedure as 9 out of 11 (surrounding - 11 out of 11) instead of 7 out of 11 (but surrounding 5 out of 11).

Despite the fact that the questions from the block of determining behavioral characteristics are quite abstract, they nevertheless demonstrate that the surveyed respondents are more knowledgeable than the average individual.

The next block of questions was aimed at identifying Late-Covid trends in food retail in Kyiv, as almost all respondents live in the capital. According to the results of the survey, 61.7% of respondents buy products from Silpo, 54.6% - from ATB, 41.1% - from Novus, 36.9% - from Auchan, 19.9% - from Varus, less - in other networks (respondents had the opportunity to choose up to 3 options).

When choosing a store, most respondents preferred the proximity of the location, range of products and prices of products. More than half of the respondents (55.3%) stressed that the proximity of the store location became the most important criterion for choosing a place of purchase after the pandemic, which is confirmed by the rapid development of the segment "at home" in 2020 (Fig. 3).

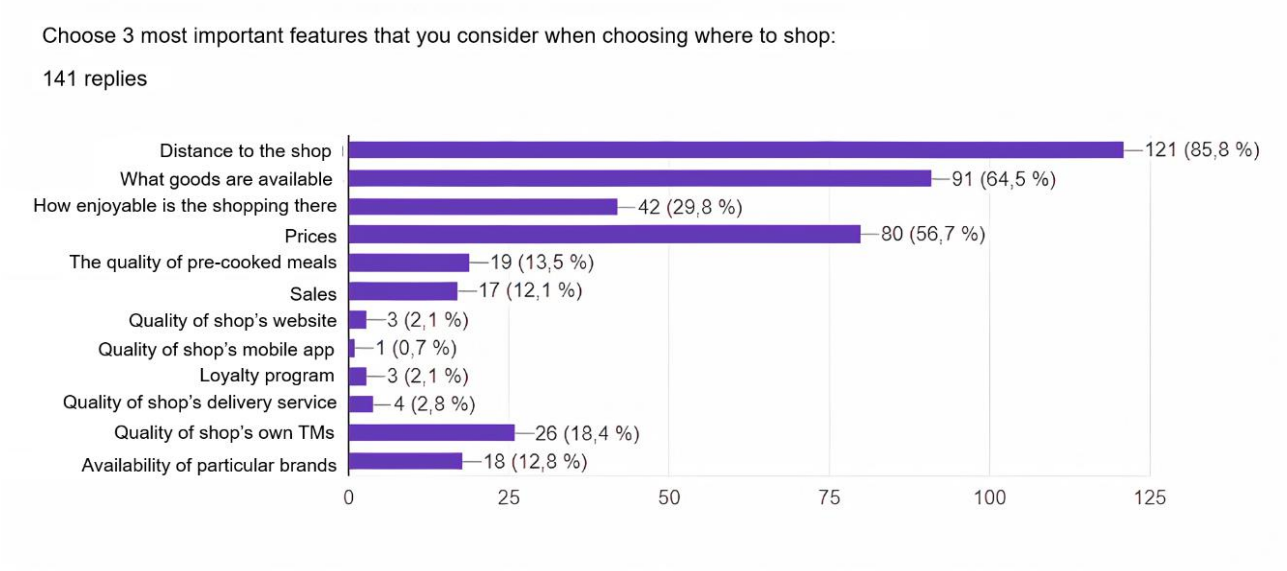


Figure 3. Distribution of the sample according to the criterion of choosing the place of purchase

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The most popular products were the segment "fresh" (fresh vegetables and fruits), dairy products and cereals / pasta. In the publications of public food retail companies, these types of products are also highlighted as the main ones. This fact confirms the current trend towards healthy eating.

Interestingly, 39.7% of respondents admitted that they make impulsive purchases during almost every visit to the store, and 36.9% do so several times a month. At the same time, only 9.9% check the list of components of each product, but for 38.3% this action is only as an exception to the rules.

As for environmental initiatives of food retail, 72.3% of respondents consider them interesting, but in no case as reasons to buy in the stores that conduct such initiatives.

Even after the pandemic, 66.7% of respondents either do not buy food online or make less than 1 purchase out of 10 online purchases. But only 9.2% of respondents use paper money to pay. This shows that the pandemic has greatly increased the use of electronic funds. However, in Ukraine there is still distrust in ordering food online, even among young people (Fig. 4).

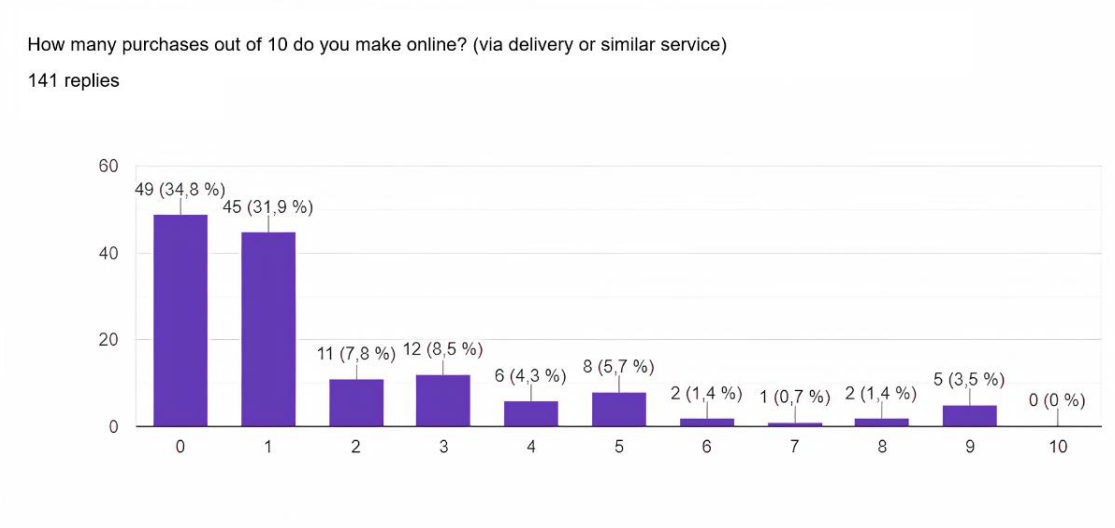


Figure 4. Number of food orders online

The next step in the study was to identify clusters of consumers of food retail chains in order to improve their "buyer's journey" and better understand their needs.

When clustering with K-means, three clusters were identified. SI was 0.48, which shows the average distance of the clusters. Fig. 5 presents a visualization of the results of clustering, relative to the first (PCA1) and second (PCA2) main component. PCA1 lies on the abscissa axis. It is worth noting that there were more women than men in the sample, so the sex parameter is not significant in the clusters.

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

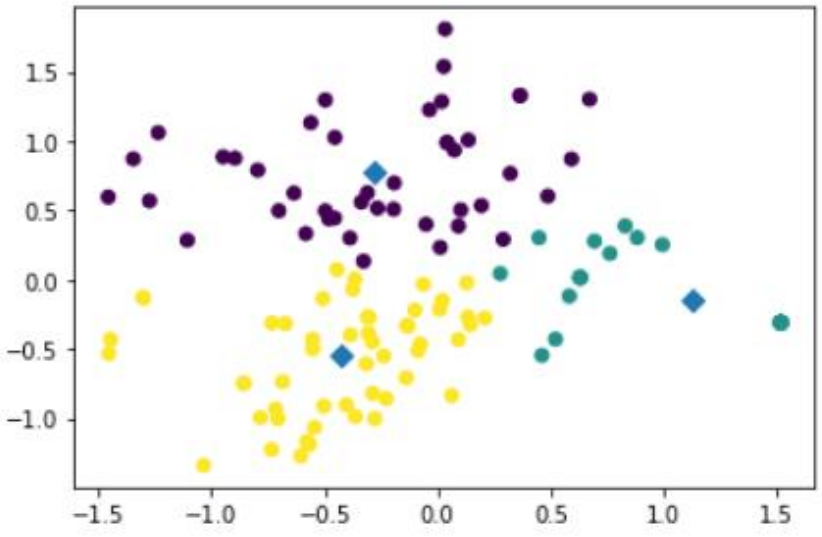
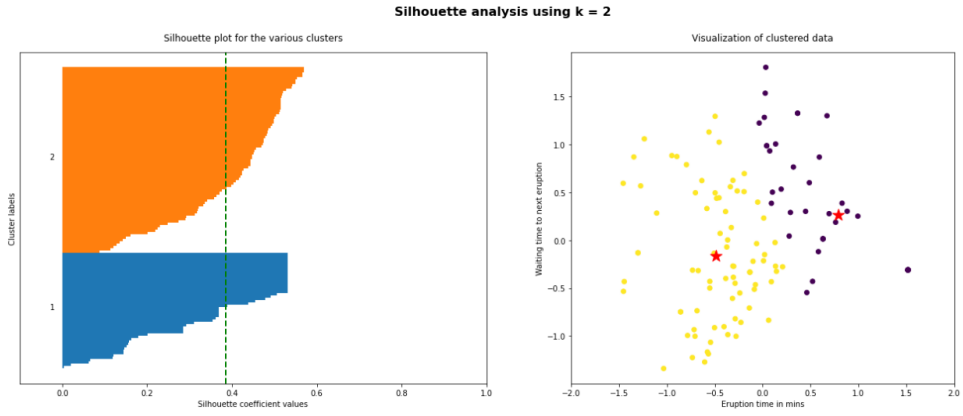


Figure 5. Graphical interpretation of K-means clusters

PCA1 and PCA2 explain 57% and 43% of the scatter, respectively. This means that the PCA-analysis passed normally and gives significant results. Each characteristic has its own level of impact on the main components. The larger the Scores in absolute terms, the more important the characteristic for the corresponding component. Using the "elbow method", 3 clusters were identified.

In addition, a Silhouette analysis (SA) was performed to confirm the selection of the optimal number of clusters. In Fig. 6 you can see the result of SA, which is a calculation of SI for a different number of clusters. The choice is made on the most balanced result. For example, the graphs show that the 4th cluster is becoming too specific.



ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

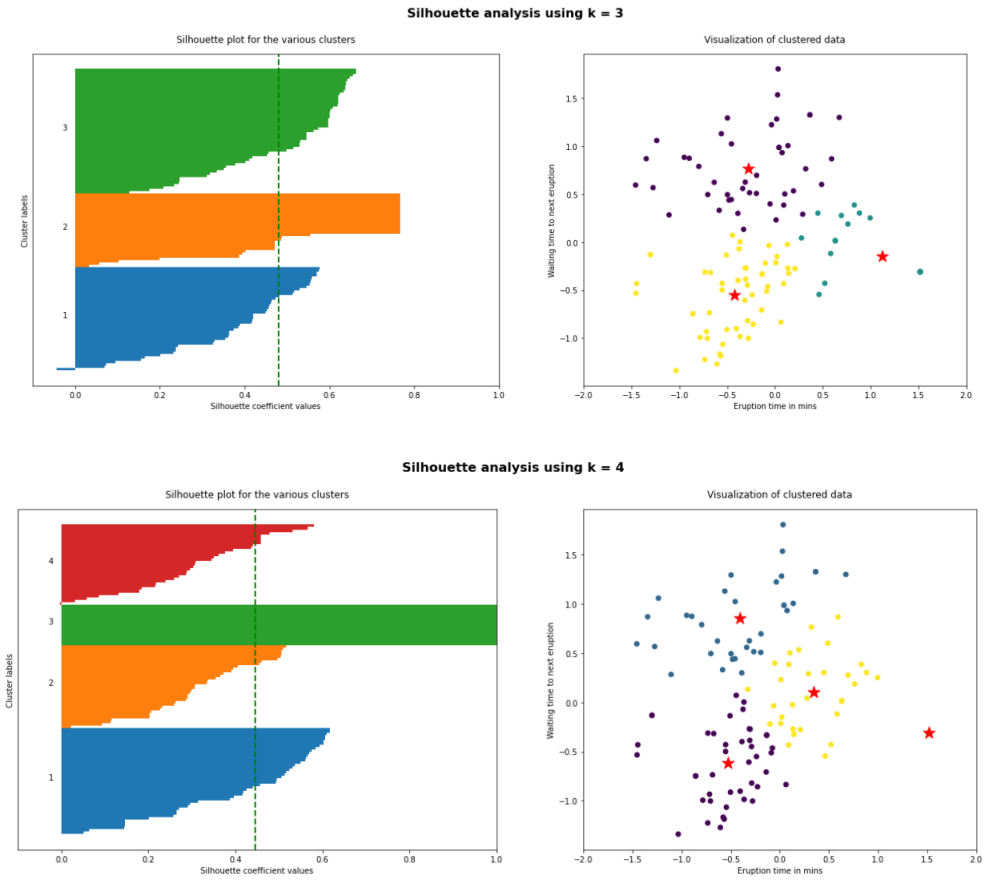


Figure 6. Silhouette analysis for k=2; 3; 4

For Spectral clustering, 3 clusters were also used as the optimal number. SI was 0.45, in addition, on the visual representation you can clearly see the problem of this method - the allocation of emissions into a separate cluster (regardless of the number of selected). Thus, in this case it is a less accurate representation of the internal groups of the dataset than K-means (Fig. 7).

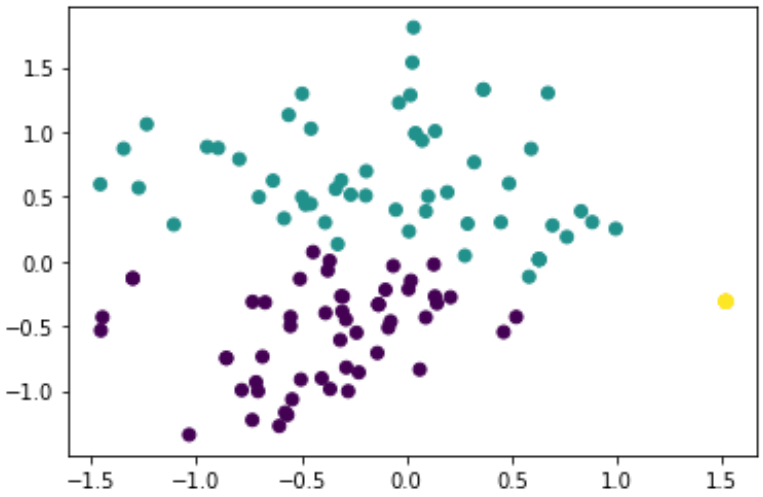


Figure 7. Spectral clustering results

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

The hierarchical method and T-SNE were used for additional visual representation. The first showed that in the data set (dataset) there are two rather large and full clusters and one consisting of a much smaller number of observations (Fig. 8).

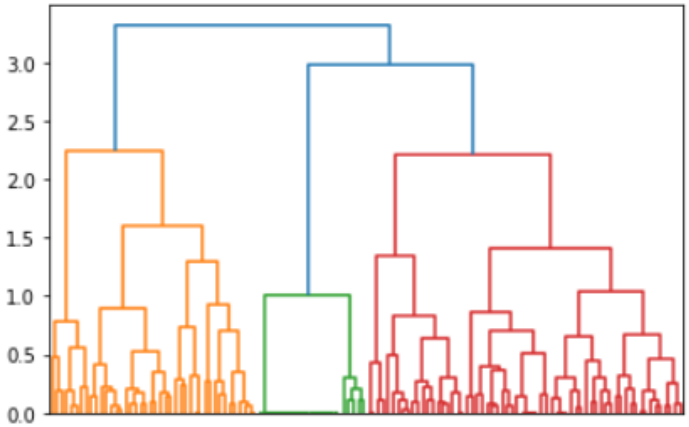


Figure 8. Hierarchical clustering results

The T-SNE method identified rather strongly superimposed clusters, so this clustering is not significant (Fig. 9).

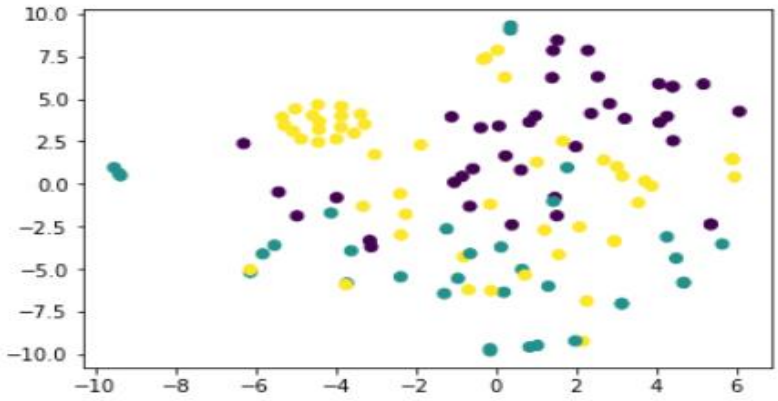


Figure 9. T-SNE clustering results

Therefore, clustering by the K-means method was the best, so clusters by K-means were used to analyze the dataset (table 1).

Table 1. K-means cluster centroids

Cl/Q	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	1	0	0	1	1	3	2	1	1
2	1	1	1	0	0	1	3	2	1	1
3	1	1	1	1	0	0	3	2	1	1
Cl/Q	**sex**	**IOS/Andr**	**work**	**study**	**Silpo**	**ATB**	**Varus**	**Novus**	**Auchan**	**fin**

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

1	2	2	1	1	1	0	0	1	0	2
2	2	2	1	1	1	0	0	0	1	3
3	2	1	1	1	0	1	0	0	0	3

Source: Own calculations

It can be seen that the majority of respondents in all clusters are women (sex = 2), work or do internships (work = 1), study at university (study = 1) and at least partially provide for themselves (fin = 2 or 3).

Among the main defining characteristics of the first cluster are the following: respondents are more risky (in the first and third questions about tossing a coin or a fixed win, they choose a coin); buy a ticket in any case (the fifth and sixth questions); are consumers of the Novus and Silpo network and fully self-sufficient (fin = 2).

The second cluster is less risky (in the first and third questions a fixed win is chosen); ticket purchase occurs only in case of loss of funds (the fifth and sixth questions); respondents are consumers of Silpo and Auchan; they partially provide for themselves.

The third cluster has respondents who are risk-free even in case of loss of money (the fourth question); they are consumers of ATB stores and partially provide for themselves.

In order to confirm the practicality of using this survey, Silpo clients were classified by the Random Forest algorithm. The binary characteristic (1 - client, 0 - no) was taken as the dependent variable, which was explained by the answers to other questions. Initially, the sample was divided into training and test in the proportion of 80% by 20% using random sorting algorithm train_test_split. The next step was normalization by the MinMaxScaler algorithm.

After cyclic training of 30 models with random characteristics, a random forest was built, which included 100 decision trees. Depth of each tree was 10 (maximum number of division of branches); the algorithm uses bootstrap (each tree receives a random incomplete set of data from the total sample). The accuracy of the Random Forest Classifier was assessed by several metrics from the sklearn library.

Traditional methods of evaluating classifiers in machine learning use a confusion matrix, which shows how different the data classified by the model differs from the true data set. Matrix elements are denoted as TP - True Positive, FN - False Negative, TN - True Negative, FP - False Positive. Using an error matrix, you can calculate several metrics that will assess the effectiveness of classification algorithms (Murtagh&Contreras, 2012).

Precision is the share of correctly classified elements (True Positive) among all positives: $prec = \frac{TP}{TP + FP}$. Recall is the share of correctly classified elements among all relevant elements $rec = \frac{TP}{TP + FN}$. Accuracy is

ACCESS Journal:
Access to Science, Business, Innovation in Digital Economy
2021, 2(3), 234-251, https://doi.org/10.46656/access.2021.2.3(3)
ISSN 2683-1007 (Online)
https://journal.access-bg.org/

the share of correctly classified elements $acc = \frac{TP}{TP + FP + TN + FN}$. Balanced accuracy score, calculated as the average of all class corrections: $bacc = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$.

F1 measure (F1 score) is the average harmonic Precision and Recall: $F1 = 2 \cdot \frac{prec \cdot rec}{prec + rec} = \frac{2TP}{2TP + FP + FN}$.

The highest possible value of the F-measure is 1, which indicates ideal accuracy and sensitivity, and the lowest possible value is 0 if either accuracy or completeness is zero.

The quality of the constructed classifier is characterized by the following values of measures (Table 2). The Accuracy score metric, which determines the accuracy of the multiclass classification, became 0.73. The Balanced accuracy score metric, calculated as the average of all class corrections, showed a close value of 0.725. The values of these metrics do not match because the number of examples in each class is not the same. A balanced F-score or F1-measure gave a score of 0.77.

The average accuracy of the model based on the data of three metrics is 0.742, which is a rather mediocre result and requires further adjustment. One such option is to discard redundant characteristics by processing data by methods such as linear regression, lasso regression, recursive feature elimination (RFE), and so on.

Table 2. Effectiveness metrics of Random Forest classification

Metrics	Accuracy	Balanced Accuracy	F1	Average
Scoring	0.730	0.725	0.770	0.742

Source: Own calculations

CONCLUSION

In the course of this study, a survey was created aimed at identifying patterns of behavior of the modern consumer according to his criteria for choosing stores and reactions to questions based on the theorems of behavioral economics. The survey took place in December 2020 during Late Covid. Clustering using machine learning methods was applied to the obtained results. The next stage of the study was the classification based on the Random Forest algorithm.

Using Silhouette analysis, the optimal number of clusters was determined and the database was divided into three K-means clusters. T-SNE algorithms, hierarchical and spectral analysis were used for additional visual representation.

According to the results of the analysis, it was found that the respondents of the first cluster are at risk, self-sufficient and are clients of Novus and Silpo. The second cluster includes moderately risky respondents who are partially self-sufficient, consumers of Silpo and Auchan. The third cluster is formed by respondents who are risk-free, partially self-sufficient, and use the services of the ATB stores. It is recommended to take

ISSN 2683-1007 (Online)
https://journal.access-bg.org/

these characteristics into account when creating personalized consumer offers, for example, to offer a more "aggressive" one for customers from the first cluster.

The created survey-tool for clustering and classification of consumers by machine learning methods is suitable for use in business processes. One application is, for example, allocation customers of a food delivery company to clusters by customer behavior patterns.

However, to improve the result, it is necessary to have a more representative sample, because used in this study consists of on average rational and knowledgeable individuals. It should be borne in mind that the average typical consumer does not necessarily have such qualities, and this applies to both older and younger generations.

Machine learning algorithms have shown their effectiveness in studying consumer behavior. This technique can be successfully used as a tool of behavioral economics. Promising areas for further research are to identify new behavioral trends in other industries; deepening understanding of food retail; use of geodata to improve analysis, etc. A potentially attractive area would be to consider assessing the impact of advertising on consumer behavior through semantics analysis and image recognition.

Conflict of interests

The authors declare no conflict of interest.

References

Accuracy Score. Retrieved from: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score. (accessed: June 2021)

Ahuja, Ravinder, et al. (2020). Classification and clustering algorithms of machine learning with their applications. *Nature-Inspired Computation in Data Mining and Machine Learning*. Springer, Cham, 225-248.

Balanced Accuracy Score. Retrieved from: https://cutt.ly/ibZ3IgM. (accessed: June 2021)

Balanced Accuracy Score. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. (accessed: June 2021)

Barberis, N., Xiong, W. (2009) What drives the disposition effect? An analysis of a long-standing preference-based explanation, *The Journal of Finance*, LXIV(2).

Bechara, A. (2002) The somatic marker hypothesis: a neural theory of economic decision. *Games and Economic Behavior*, 52(2), 336-372.

Beerbaum, D., Puaschunder, J.M. (2018). A behavioral economics approach to digitalisation – the case of a principles-based taxonomy. *Advances in Social Science, Education and Humanities Research*, 211, 45-53.

de Arruda, T.J., de Moraes, M.B., de Araujo Querido Oliveira, E.A. (2015). Behavioral finance: a study on investments decisions, *Business and Management Review Special Issue*, 4(7). Retrieved from: http://www.businessjournalz.org/bmr (accessed: June 2021)

Della Vigna, S. (2018) Structural behavioral economics, Handbook of Behavioral Economics, vol. 1 (eds. D. Bernheim, S. DellaVigna, and D. Laibson), Elsevier.

Deloitte. 2021. Consumer sentiment of Ukrainians in 2020. Industry group for retail and wholesale distribution. Retrieved from:https://www2.deloitte.com/ua/uk/pages/press-room/press-release/2021/2020-consumer-behavior-in-ukraine.html (accessed: June 2021)

Dyer, J., Kolic, B. (2020). Public risk perception and emotion on Twitter during the Covid 19 pandemic. *Applied Network Science*, 5(99).

Food retail trends by foodnavigator. Retrieved from: https://www.foodnavigator.com/Article/2021/01/06/Retail-predictions-2021-Experts-forecast-food-plastic-and-climate-smart-trends. (accessed: June 2021)

Google Trends in Ukraine (food retail), Retrieved from: https://rau.ua/novyni/5-golovnyh-tendentsij-v-marketyngu/. (accessed: June 2021)

Hrnjic, E., Tomczak, N. (2019) Machine learning and behavioral economics for personalized choice architecture, preprint, arXiv:1907.02100v1 [econ.GN], 2019.

İnaç, H. (2019) A Theoretical Perspective on Behavioral Finance with Lagrangian Approach, *Quantrade Journal of Complex Systems in Social Sciences*, 1(1), 22-27.

Jabnidze, N.; Tsetskhladze, L.; Meskhidze, I. (2021). Food Security Problems for Developing Countries in the Conditions of COVID-19: Case of Georgia. *Economics. Ecology. Socium* , *5*, 8-17.

Johnson, E.J., Gächter, S., Herrmann, A. (2006). Exploring the nature of loss aversion. IZA Discussion Papers, Institute for the Study of Labor, 2015.

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *The American Economic Review*, 93(5), 1449–1475.

Kahneman, D. (2011). Thinking, fast and slow, New York: Farrar, Straus and Giroux.

Kahneman, D., Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*. 47, 263-291.

Kolumbus, Y., Noti, G. (2019). Neural networks for predicting human interactions in repeated games. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Main track, 392-399.

Lekovic, M. (2019). Behavioral portfolio theory and behavioral asset pricing model as an alternative to standard finance concepts, *Economic Horizons*, 21(3), 255 – 271.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2.3, 18-22.

Machina, M., Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica*, 60(4), 745-780.

March, C. (2019) The behavioral economics of artificial intelligence: lessons from experiments with computer players, CESifo Working Paper, 7926, category 13: Behavioural Economics.

Marsh, H.W. (2005). Big Fish Little Pond Effect on Academic Self-concept: Cross-cultural and Cross-Disciplinary Generalizability, SELF Research Centre, University of Western Sydney.

Marsh, Herbert W., et al. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational psychology review* 20(3), 319-350.

McKinsey&Company, 2020. Perspectives on retail and consumer goods. Retrieved from: https://cutt.ly/ebZ1Ylb. (accessed: June 2021)

MOOC «Behavioral Finance». Retrieved from: https://www.coursera.org/learn/duke-behavioral-finance/home/welcome. (accessed: June 2021)

Murtagh, F., Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1, 86-97.

Najdenovska, I., Stojanovska, F., Gievska, S. (2018). Detecting emotions in tweets based on hybrid approach, Proceedings of the 15th Conference for Informatics and Information Technology: CIIT 2018, 20-22 Mavrovo, Macedonia, pp. 235-240. Retrieved from: https://www.researchgate.net/publication/331089014. (accessed: June 2021)

Podcast “Freakonomics Radio”. Retrieved from: https://freakonomics.com/podcast/. (accessed: June 2021)

Podcast “No Stupid Questions”. Retrieved from: https://freakonomics.com/podcast/. (accessed: June 2021)

Prince, E.T (2018). Risk management and behavioral finance. *Financial markets, institutions and risks*, 2(2), 5-21. DOI: 10.21272/fmir.2(2).5-21.2018.

Reyes, J.A.P., Miranda, M.R., Vera-Martinez, J. (2019). Capital structure construct: a new approach to behavioral finance. *Investment Management and Financial Innovations*, 16(4), 86-97.

Roster on Debit Cards in food retail. Retrieved from: https://rau.ua/novyni/novini-kompanij/53-vyruchky-restoratoriv-kartamy/. (accessed: June 2021)

Roweis, S. (1998). EM algorithms for PCA and SPCA. *Advances in neural information processing systems* 626-632. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. (accessed: June 2021)

Stefánsson, H.O., Bradley, R. (2017). What is risk aversion? *The British Journal for the Philosophy of Science*, 70(1), 77–102.

The Behavioral Economics Guide 2014. Introduction to behavioral economics. Retrieved from: https://www.behavioraleconomics.com/resources/introduction-behavioral-economics/. (accessed: June 2021)

Train_test_split. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. (accessed: June 2021)

Tversky, A. (1986). Rational choice and framing of decisions. *Journal of Business*, 59, 252–278.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing* 17.4, 395-416.

Wattenberg, M., Viégas, F., Johnson, I. (2016). How to use t-SNE effectively. *Distill*, 1(10).

Yadav, J. (2017). Selecting optimal number of clusters in KMeans Algorithm (Silhouette Score). *Medium*. Retrieved from: https://jyotiyadav99111.medium.com/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308. (accessed: June 2021)

About the authors

Olena LIASHENKO

Doctor of Economics, Professor of Taras Shevchenko National University of Kyiv (Ukraine). Head of Economic Cybernetics Department, Economics Faculty, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Research interests: nonlinear modeling in socio-economic and environmental systems, financial time series, modelling of economic processes by neuronetworks methods, fractal analysis, economic dynamics modelling.

ORCID ID: https://orcid.org/0000-0002-0197-4179

Tetyana KRAVETS

PhD (Physics and Mathematics), Associate Professor of Economic Cybernetics Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Research interests: modeling of economic processes by neuronetworks methods, clusterization, financial time series, wavelet analysis.

ORCID ID: http://orcid.org/0000-0003-4823-5143

Matvii PROKOPENKO, Student, Economic Cybernetics Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. Research interests: behavioral economics, modeling in economics, data science.

ORCID ID: https://orcid.org/0000-0003-1457-8762