DOI: 10.17323/2587-814X.2024.4.7.24

Neural network technologies in supply chain management: Consumer selection technique*

Valeriya A. Nazarkina
E-mail: valeria71@bk.ru

Vladislav Yu. Shchekoldin
E-mail: schekoldin@corp.nstu.ru

Novosibirsk State Technical University (NSTU), Novosibirsk, Russia

Abstract

The supply chain management's effectiveness depends, among other things, on the selection and coordinated interaction with product consumers. This article is devoted to the development of a method for selecting a consumer in the regional wholesale and retail fuel market. The methodological basis of the study is the theory of statistical analysis and neural networks. The main tool for developing the methodology was neural network technologies, with the help of which it is most likely possible to correctly estimate the boundaries for indicators' values that characterize consumers and reflect their history of purchasing behavior, to select potential clients and the possibility of further cooperation with existing ones. The information base for the work is the data on consumers of a given company's products, data from the 2GIS electronic directory, as well as the results of the primary statistical analysis and forecasts made based on neural networks of various topologies. The author presents his methodology for selecting a consumer. It has the potential for development and implementation for solving a number of other management problems. As part of the testing, the best configuration (topology) of the neural network was determined, and standard values of entry barriers when consumer choice accomplished were assessed. The methodology we developed

* The article is published with the support of the HSE University Partnership Programme

was tested using the example of a company operating in the wholesale and retail fuel market in Novosibirsk and the Novosibirsk region. When verifying the neural network model, the quality of client classification was compared based on logistic regression, decision tree and random forest models and we found that the neural network approach provides the best results for assessing the degree of client suitability. As a result of testing the methodology, recommendations for improving neural network models were developed, including expanding the set of factors that determine the characteristics of consumers, as well as optimizing the internal structure of neural networks.

Keywords: neural networks, supply chain management, distribution logistics, consumer selection, neural network training

Citation: Nazarkina V.A., Shchekoldin V.Yu. (2024) Neural network technologies in supply chain management: Consumer selection technique. *Business Informatics*, vol. 18, no. 4, pp. 7–24. DOI: 10.17323/2587-814X.2024.4.7.24

Introduction

The unpredictability of the influence of external environmental factors poses challenges to the management of organizations associated with the risk of choosing an incorrect management decision in various areas of the company's activities. Despite the desire to attract as many customers as possible and gain their loyalty, the organization faces problems of interaction complexity, transaction inefficiency and default on contractual obligations by counterparties. A sufficient number of methods for selecting a reliable resource supplier from theoretical and practical points of view have been developed, but scientists and business representatives do not deal with the issues of choosing a consumer in detail. The decision on the choice of consumers and further cooperation with them is often determined by trial and error, which ended up affecting the effectiveness of management decisions in the field of supply chain management in general, and interactions with consumers in particular.

Nowadays one of the topical methods for solving a wide variety of economic and management problems is the implementation of neural network technologies. From the point of view of logistics, significant objects for the application of neural networks are processes of supply chain management, because these methods can be used to evaluate, model and forecast the development with predetermined accuracy.

The purpose of this study is to develop a method for selecting consumers based on the use of neural networks and to test it using the example of a company operating in the regional wholesale and retail fuel market.

1. Theoretical basis of the research

Neural networks are a certain type of artificial intelligence models based on the structure, dynamics and functions of the human brain. Neural networks in the most general form consist of many interconnected nodes (neurons) and are used to process, model and analyze complex heterogeneous processes. Neural networks have attracted wide attention and are being intensively used in various fields due to their ability not only to analyze data, but also to independently build logical conclusions, form value judgments and develop predictive solutions of a wide variety of types.

Historically, the first works related to the application of theories describing the functioning of the brain appeared in the 1940–1950's, when it became necessary to build concepts of artificial neurons to determine their potential in the development of complex intelligent systems. One of the first successful representations of the nervous system was the perceptron model proposed by Rosenblatt in 1957 [1], which turned out to be capable of solving some recognition problems. Despite the initial successes of the perceptron model, it had certain shortcomings that did not allow it to be used to solve certain types of problems. In the 1970–80's, the backpropagation algorithm was developed [2], which made it possible to train more complex networks. Among the extensive research devoted to the issues under consideration, it is worth noting the significant contribution to the development of the theory of neural networks and pattern recognition by the Soviet and Russian scientist Galushkin [3].

The 21st century has also contributed to the development and dissemination of neural network technologies, with a rapid growth in research related to so-called deep learning. Pioneering research in this area is associated with the name of Hinton, who made a significant contribution to the development of modern deep neural network training algorithms [4]. Deep learning involves training neural networks with several hidden layers, including those with a dynamically varying structure, the presence of external and internal topologies, etc.

It should be noted that in addition to neural network models, other approaches are also encountered in practice that allow solving various classification problems. Such methods include, for example, the construction of logistic regression models, decision trees, random forest models, etc. [5], and each of these approaches has its own advantages and disadvantages. Let us consider some of them.

The advantages of logistic regression models include [6]: high result interpretability, high efficiency of statistical procedures used to estimate the model parameters; and flexibility in solving binary and multifactor classification problems. Logistic regression models also have a number of disadvantages: dependence of the results obtained on the structure of input factors; high sensitivity to outliers in the initial data; and frequent manifestation of the multicollinearity.

The application of decision tree models ensures the construction of an easy and convenient interpretation of the solutions obtained; universality of the computational scheme for any type of data; robustness of the classification results, i.e., independence on outliers in the initial data. The disadvantages of this approach are as follows: the resulting decision trees often have a complex and confusing structure (so-called tendency to overfitting); classification instability to changes in the initial data; and the appearance of locally optimal solutions, which are determined by the heterogeneity of the initial data or the processes being studied, and so on [7].

When using random forest models, the researcher obtains the following advantages: the ability to scale and conveniently parallelize the basic algorithm depending on the properties of the problem being solved; the ability to rank independent factors included in the model; and high efficiency of classifications for large-scale problems. However, such models have these disadvantages: the complex and ambiguous structure of the model due to a wide variety of averaging options; instability to fluctuations in the initial data; a large number of empirically determined algorithm parameters; and high requirements for the characteristics of computing equipment (memory, speed, etc.) [8, 9].

The development of neural network technologies has had a great impact on the field of logistics. Many formulations of logistics problems and methods for solving them have undergone significant changes in light of neural network theory. We will mention some of them.

In the research [10], the authors apply the idea of constructing a multilayer neural network to analyze order data, including the number of visits to the company's website, the time of visit taking into account working, weekdays, and holidays. The article [11] discusses the implementation of the Hopfield neural network to solve the problem of constructing a dynami-

cally optimal route in a telecommunications network and proposes a heuristic rule for stopping the neural network training process, all of which allows for effective limitation of the training time.

Neural networks can be used to optimize warehouse operations. In the article [12] the authors prove that networks with special neural activation functions such as *Traingdx* are most effective in warehouse management when using three-layer neural networks with a 6-8-1 topology.

Another area of application of neural networks is the analysis and forecasting of risks in logistics. The article [13] considers the problem of safe flight around obstacles by manned and unmanned aerial vehicles, where the author proposes to use a multi-layer network of sequential error propagation with three layers.

It should be noted that among the large number of works devoted to solving logistics problems in which neural network technologies would be used, there is practically no research in which consumers of goods and services were studied. At the same time, studying the history of consumer behavior and constructing consumer classifications is given a lot of attention when solving problems outside the field of logistics [14–17], since understanding the structure of the customer base and the dynamics of its changes gives the company additional tools for improving the efficiency of interactions with consumers.

In today's rapidly changing and often complicated to predict economic environment, companies have to deal with situations where it is important to determine priorities when cooperating with certain customers. Thus, at the stage of concluding contracts, it is sometimes necessary to identify promptly whether a particular client is financially independent and solvent, whether there will be difficulties with fulfilling orders due to geographical remoteness, whether it is possible to use jointly warehouse capacities, etc. Undoubtedly, by analyzing the history of consumer behavior, it is possible to determine which of the already concluded and fulfilled contracts appeared to be profitable for the company, which did not lead to the achievement of the set goals, and which turned out to be a complete failure in terms of revenue, resource costs, and employee time, causing damage to the company's image, etc. This information can be used to determine in advance the degree of profitability of cooperation with the next potential client based on its contemporary characteristics. The study presented here is aimed at solving this problem.

2. Research design

The necessity in solving optimization problems of the logistics processes stipulated the possibility of using neural network technology in terms of determining the boundaries of certain factors' values for selecting potential consumers, and its significance for the company in terms of profitability and fulfillment of contractual obligations. This study was structured based on empirical data collected from secondary and primary sources. To solve the research problems, we used the method of primary statistical analysis, correlation and regression analysis, the method of back propagation of error and data mining methods [2, 6, 18, 19].

The content and expected results of the application of the consumer selection methodology are presented in *Table 1*.

At *the first stage* of the methodology, it is necessary to describe the essence of the problems in terms of interaction with the company's consumers. We analyze the stages of the logistics cycle typical for working with consumers. Problems arising in the process of product distribution may concern issues of information interaction between the company and consumers, quality control and transportation, documentation, etc. As a result of identifying the problem, the organization's management makes a decision to find the best methods for selecting consumers.

At *the second stage*, it is required to select the indicators that the organization is guided by when making a decision on interaction with the consumer. The factors typical for choosing a consumer include: distance to the delivery location; number of types of products subject to simultaneous sale; equipment capacity; rating; sales volume for a certain period, etc.

Table 1.

The main stages of the consumer selection methodology using a neural network

	Stage	Content	Results
1	Problem definition	Analysis of the logistics cycle stages, problem's identification	Determining the need to find relevant methods for obtaining information about consumers
2	Identifying factors influencing the consumer selection process	Composition of a set of factors subject to quantitative assessment by company specialists, experts, and research consumers	List of factors characterizing consumer properties that are used to build a neural network
3	Creating a database of current consumers of the organization	Determining the values of selected factors for consumer evaluation	A database of up-to-date data on consumers and their transactions for training the neural network
4	Primary statistical analysis	Calculation and interpretation of statistical characteristics of factors	Preliminary conclusions about the properties of observed objects (consumers)
5	Defining the roles of factors	Building a simple neural network	Ranked list of factors for making a consumer choice decision
6	Building a neural network of complex structure	Building neural networks of different structures and comparing them with results obtained on the basis of other classification models	Choosing the best configuration (topology) of a neural network
7	Using the "best" neural network to identify prospective clients	Evaluating the range of factor values that determine the consumer's status	Determining standard values of entry barriers for consumer selection

At *the third stage*, it is necessary to develop and fill a database for training the neural network. For all potential consumers of the company, the values of the factors selected at the second stage are determined. Some factors are assessed by direct measurements of the relevant indicators, while others require obtaining and using secondary information. In addition, each of the potential consumers must be assessed by a logistics specialist for the priority of choosing him as a real client.

The fourth stage involves calculating the main statistical characteristics determined by the values of the factors of the consumer database constructed in the third stage. To ensure the correctness of the statistical analysis, in particular, to determine homogeneous groups within which one can speak of a certain identity of the analyzed objects (consumers), the data must be distributed (grouped) by homogeneity classes, the number of which k is determined, for example, by the Sturges formula [20]:

$$k = [1 + 3{,}32\lg(N)], \quad (1)$$

where

N – total number of data (the sample size, or the database volume);

lg(.) – decimal logarithm;

[.] – operation of calculating the integer part of a number.

As a result, a statistical analysis of the values of the measures of central tendency, variation and shape is carried out, and preliminary conclusions are made about the properties of consumers and the history of their purchasing behavior.

At *the fifth stage*, the factors are ranked according to their degree of influence on the decision to interact with potential clients. For this purpose, a simple neural network is built, consisting of one neuron and input variables corresponding to the factors selected at the third stage. The degree and nature of the influence of the factors will be determined by the values of the network weights. As a rule, the efficiency of such a network is quite low, which does not allow it to be used to predict the priority of consumers. However, it allows us to rank the input factors, which makes it possible to build a high-quality interpretation of the decision-making process for choosing customers.

At *the sixth stage*, neural networks of various topologies are developed and trained to ensure the best degree of predictability of the customer's "utility." In this case, it is necessary to consider several network options that differ from each other in the number of hidden layers, their interrelations and the number of neurons in each layer [21]. The result of this stage will be the selection of a neural network configuration that ensures the lowest level of error in predicting the reliability of consumers.

To ensure the adequacy of the obtained results, it is necessary to compare the quality of the customers classification obtained based on neural networks and classifications constructed by other methods of data mining. The simplest way to compare different classifications is to use the contingency table method [22]. We will assume that the best classification will be the one that provides the least number of forecasting errors.

At *the seventh stage*, the best of the neural networks built at the sixth stage is used to determine the ranges of factor values at which the status of the current client is maintained. This will also allow us to determine the standard (base) values of the input factors to simplify the procedure for selecting a new customer.

3. Practical aspects of the research: testing the methodology

The methodology was tested based on information provided by a company specializing in the sale of liquefied gas to organizations and individuals in Novosibirsk and the Novosibirsk Region. The company also sells related products and services, carries out design work, provides maintenance services and organizes technical maintenance of natural gas pipelines.

At the *first stage*, the operations of the logistics cycle were considered. The consumers were defined as LPG filling stations (hereinafter referred to as LPGFS) located in the city of Novosibirsk and the Novosibirsk region. The study of the logistics process in the company revealed the need for regular verification of information on the characteristics of consumers, since there are no clear boundaries of the values of factors that make it possible to identify consumers with properties that allow them to be considered suitable for cooperation. In this regard, errors arise, the causes of which are, for example, the influence of external factors, limited resources, errors of specialists, etc., which can lead to incorrect identification of customers and, as a consequence, to apply incorrect decisions on working with them.

The organization's management decided to look for ways to optimize interaction with consumers in terms of determining standard values of factors for both current consumers and for selecting and interacting with potential consumers.

At the *second stage*, a pool of factors was compiled that determine the characteristics of consumers

taken into account when making decisions. This pool included the following indicators:

- distance from the company to the consumer (km);
- range of products (number of types of liquefied gas purchased by the consumer);
- production capacity (number of dispensers at LPGFS);
- gas station rating;
- total volume (capacity) of the main and additional tanks at LPGFS;
- average volume of fuel sales per day (thousand liters).

At the *third stage*, a database structure was developed that included information on the company's transactions. The database contained the company's operating results for the period September–November 2023.

A company specialist who held the position of head of the logistics department assessed the profitability of completed transactions, as a result of which all database records were marked with the values of a binary variable that determined the "reliability" of clients: the designation "1" was used for "suitable" clients with whom it was profitable for the company to continue to cooperate, and "0" for "unsuitable" ones.

At the *fourth stage*, a primary statistical analysis was conducted for each of the factors that characterize the activity of LPGFS. The values of the measures of central tendency (mean, median), variation measures (standard deviation, lower and upper limits), measures of shape (skewness and kurtosis), as well as extreme values (minimum and maximum) were calculated. The results are summarized in *Table 2*.

Table 2.

Values of statistical characteristics of factors

Statistical characteristics	Factors and their designations					
	Distance, km	Assortment width, units	Production capacity, units	Rating, conv. units	Tank capacity, thousand liters	Sales volume, thousand liters
	x_1	x_2	x_3	x_4	x_5	x_6
Mean	118.469	2.563	4.547	2.469	19.983	5.610
Median	51.500	2.000	4.000	2.300	17.000	4.375
Standard deviation	136.189	1.344	2.949	0.920	14.351	3.749
Lower limit	0.000	1.219	1.598	1.549	5.632	1.861
Upper limit	254.658	3.907	7.496	3.389	34.334	9.359
Skewness	1.687	0.531	2.09	0.633	1.986	1.428
Kurtosis	2.535	1.035	8.054	0.141	5.649	2.335
Minimum	11	1	1	1	3.3	0.8
Maximum	628	5	18	5	85	18

Note that the lower and upper boundaries in *Table 2* correspond to the so-called "one sigma" interval and are defined as the difference and sum of the mean value and standard deviation, respectively. The interval between these boundaries contains the most probable values of the random variable being analyzed, which in the case of a normal distribution include about 70% of the sample observations [7, 23].

To ensure the correctness of the analysis, the data were distributed into homogeneity classes, the number of which was determined by the Sturges formula (1) and was found to be seven. The histogram, which is a graphical interpretation of the frequency distribution of LPGFS by distance to customers, is shown in *Fig. 1*.

Based on *Fig. 1*, we can assume about the exponential distribution of the distance to customers, which is explained by a significant increase in their number when approaching the city. 41 gas stations out of 64 studied (64%) are located within 100 km from the company. The average distance to customers is 118 km. Half of all values fit into the interval from 11 km to 51.5 km, which indicates a strong predominance of low values of this factor, which is also confirmed by the positive value of the sample's skewness. The most probable values of this indicator are in the interval (0–254), 79% of all consumers fall into it. The kurtosis coefficient is greater than zero, which indicates a good predictability of the indicator relative to the normal distribution, since a large number of gas stations are grouped in one class.

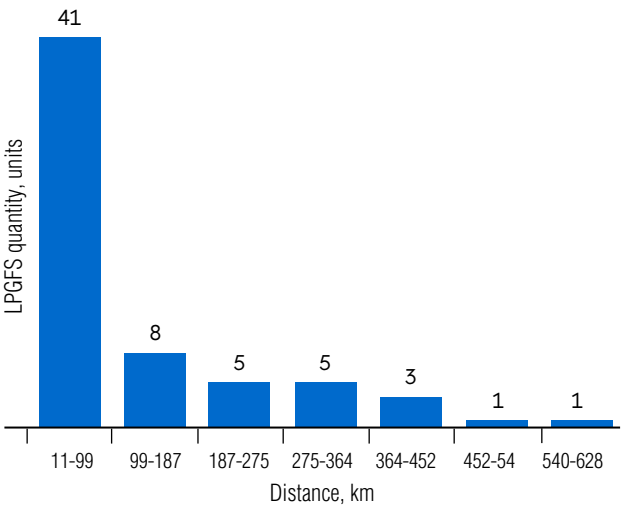

Fig. 1. Distribution of consumers by the factor "distance from the company to the client".

The frequency distributions of the number of LPGFS for the remaining factors were interpreted in a similar manner.

When analyzing the number of types of fuel sold, it turned out that the most common LPGFS are those with two or four types of liquefied gas (40%). For the "production capacity" indicator, it was noted that the total number of gas stations has no more than six dispensers. Analyzing the data on consumer ratings, one can notice some unpleasant statistics: most often, gas stations have a low rating (no higher than 3 points), while the average rating is around 2.5. The distribution of tank volume values is similar to exponential. The overwhelming majority of consumers (more than 80%) have installed tanks with a total volume of no more than 34 thousand liters. When analyzing the volume of daily sales, it turned out that the largest number of LPGFS is characterized by low sales relative to the rest – less than 6 thousand liters of gas per day. Only six filling stations out of 64 have average daily sales of more than 10 thousand liters (less than 10% of their total number).

After analyzing the initial data for the selected indicators, it is important to understand how significantly these factors affect the final result (the client's "suitability" for cooperation). For this purpose, we will use the idea of constructing the Rosenblatt perceptron [1, 2].

At the fifth stage, a neural network consisting of one neuron was constructed. As the activation function of the neuron (as for all other variants of neural networks considered in the work), the logistic function was taken in the form of

$$\sigma(x) = \frac{1}{1+e^{-x}}. \quad (2)$$

The choice of the logistic function is due to its continuity, which ensures smoothness in the transition region. The *ESS* value, the residual sum of squares [7] between the specialist's assessment (Y) and the assess-

ment issued by the neural network ($\hat{Y}$), was used as a functional determining the correctness of the neural network operation:

$$ESS = \sum_{k=1}^{N}(Y_k - \hat{Y}_k)^2 \to \min. \quad (3)$$

In (3) the summation is carried out over all database records; N is the database size (number of records). For input factors, the notations x_1–x_6 are used in the order of listing (*Table 2*).

The calculation of the estimates determined by the neural network was carried out based on the value of the activation function (2) from the linear combinations of the values of the input variables of the model x_1–x_6, as well as any variables of the internal layers of the neural network, determined by the topology of the neural network. For example, for a neural network consisting of one neuron and having six input variables, the estimated value of the probability that the client will be recognized as "suitable" for cooperation will be determined as

$$\hat{Y} = \sigma(w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6) = \sigma(\Sigma), \quad (4)$$

where w_i, $i = 1, \ldots, 6$ are the weight coefficients of each of the input factors, determined by solving problem (3); Σ is the value of the linear combination of input factors, called the adder.

Minimization of the functional (3) is carried out by changing the unknown weight coefficients w_i in the adder. Since the problem (3) has no solution in analytical form, the weight coefficient estimates were found numerically [24]. *Figure 2* shows the diagram of the implemented model from one neuron.

The model of the simplest network includes the values of the input factors (circles on the left in *Fig. 2*), the adder Σ from formula (4), and the value of the activation function (2). The output of the neural network is determined by the proportion of correct forecasts – the ratio of the number of correct responses of the neural network to the total volume of the database N (in this case, $N = 64$). The degree of darkening of the arrows reflects the strength of the influence of the corresponding input factor on the result of assessing the "suitability" of the consumer: the darker the corresponding arrow, the stronger the influence of the factor.

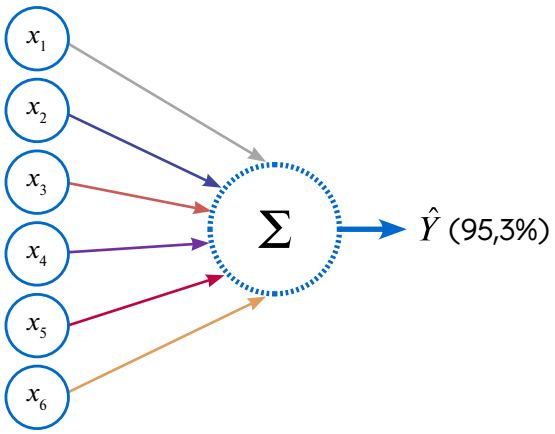

Fig. 2. Representation of a single neuron model.

By minimizing the sum of squared deviations of the probabilities of making a decision on cooperation with a consumer, estimated by the company's expert, from its values predicted by the neural network, estimates of the values of weight coefficients were obtained. *Table 3* presents the values of the coefficient estimates and the results of ranking the factors analyzed by the absolute values of the weight coefficients.

The obtained values of weight coefficients can be interpreted. For example, the sign of the coefficient indicates the direction of the dependence. From *Table 3* it is evident that only the coefficient w_1 turned out to be less than zero, which means the presence of a negative dependence, i.e., with an increase in the distance to the consumer, the probability of making a decision on cooperation decreases on average. The remaining factors have a positive dependence on the probability of making a decision on cooperation.

The absolute values of the coefficients mean the "power of influence" of the corresponding factors on the decision to cooperate with a particular client. Let us analyze the results obtained.

Table 3.

Neuron weighting coefficients' estimators and factor ranking

Weighting coefficient	w_1	w_2	w_3	w_4	w_5	w_6
Coefficient estimator	–50.000	11.101	0.671	11.639	8.600	3.744
Degree of importance, %	58.305	12.944	0.783	13.573	10.029	4.366
Factor rank	1	3	6	2	4	5

1. The distance to LPGFS (factor x_1) has the strongest influence; its importance relative to all other factors is almost 60%. At the same time, the assessment of the corresponding weighting coefficient (w_1) rests on the limit of acceptable values (±50), which means its complete dominance in decision-making. It is quite logical that it makes no sense for the company to organize deliveries over long distances.

2. The gas station rating (factor x_4) is in second place, accounting for just over 13% of importance. Of course, the higher the reputation and consumer ratings of the company, the more reliable it is.

3. The variety of fuel types at LPGFS (factor x_2) is in third place, slightly inferior to the rating (significance less than 13%). The significance of this factor is due to the higher financial stability of the enterprise in case of need for a wide range of fuel for consumers.

4. Tank capacity (factor x_5) is in fourth place. The influence of this factor is confirmed by the "convenience" of the selected gas station for cooperation. To some extent, this indicator can be considered as the volume of a warehouse for a trading company: it must be sufficient (as well as the stocks in it) so that the company can carry out its activities without hourly delivery of products. The situation is similar for LPGFS. By optimization, it is possible to minimize the required capacity of tanks, but delivery of products using just-in-time systems [25] can be quite complicated due to the specifics of the goods being transported, so it will be important to have spare gas storage tanks.

5. Daily gas sales volume (factor x_6) is in the penultimate place (less than 5% importance), which is somewhat surprising. The influence of this factor should be obvious, because the higher the company's turnover, the higher its stability. However, it is important to understand that all companies (as well as LPGFS) are in different situations. Thus, for powerful stations with six or more dispensers located in the city center, a high, at first glance, turnover may actually be quite low compared to other consumers located nearby.

6. The production capacity of LPGFS (factor x_3) is the last (importance less than 1%). Like the previous indicator, it does not have clear limits, so it is not a strong factor. In general, a larger number of gas dispensers is an opportunity for the company to develop, which can play a role in the long term.

It should be noted that the estimates obtained using the neural network correctly reflect the thought process of a specialist who does this manually. Analyzing the results obtained, it should be noted that out of 64 records in the company's client database, the single neuron network made a mistake in only three, which is 4.7% of errors. To understand the causes of these errors, it is necessary to examine the results and identify the features of these gas stations. In all three cases, the probability assessment of the client's "reliability" issued by the neural network was more than 0.9, i.e. it was more than confident in their "suitability" for cooperation.

The first LPGFS has low values for almost all indicators, but it has a high rating, which is most likely the reason for such an assessment. Obviously, the ratings

provided by special services are the most unreliable indicator, since in many cases they are either unrepresentative due to the small number of averaged assessments, or incorrect due to the use of certain schemes for "winding up" the necessary values.

The second gas station has good factor values, but the company's specialist noted that they do not cooperate with this station, since the other branch is closer to the client, and interaction is carried out through them.

The third gas station was rated positively, but the specialist rated the experience of working with this company negatively because the station is in the process of launching; the data on it is contradictory, and more time is needed to sign contracts. This situation is not an error, but in light of the increased recognition capabilities of the neural network, it is recommended to add a check for the operation time of LPGFS on the market.

At the *sixth stage*, to eliminate erroneous triggering of the neural network, it was decided to consider more complex options of network structure by adding internal layers with different numbers of neurons; the most suitable option turned out to be a topology with seven neurons 4-2-1 (*Fig. 3*).

The selected topology allowed us to successfully describe the interaction process of the factors considered, while no discrepancies were found between the expert's assessments and the results of the neural network. The darker arrows of the neuron connections in the network correspond to the third neuron of the first layer and the sixth neuron of the second (*Fig. 3*). The difference in the influence of the values of the neurons of the second layer on the network output is only 18.4%. It should also be noted that the values of the fifth neuron have a negative effect on the result, and the sixth – a positive one. Unfortunately, this fact cannot be used to interpret the results (as it was for the single neuron network), since the values of each of the neurons of the internal layers are added up under the influence of the previous layers in a nonlinear manner due to the selected activation function in the form of a logistic function (2).

In order to ensure the suitability of the results obtained, in addition to the neural network model, consumer classifications were constructed using logistic regression, decision trees and random forest. The freely distributed Orange Data Mining software package was used to develop and identify the corresponding models [19].

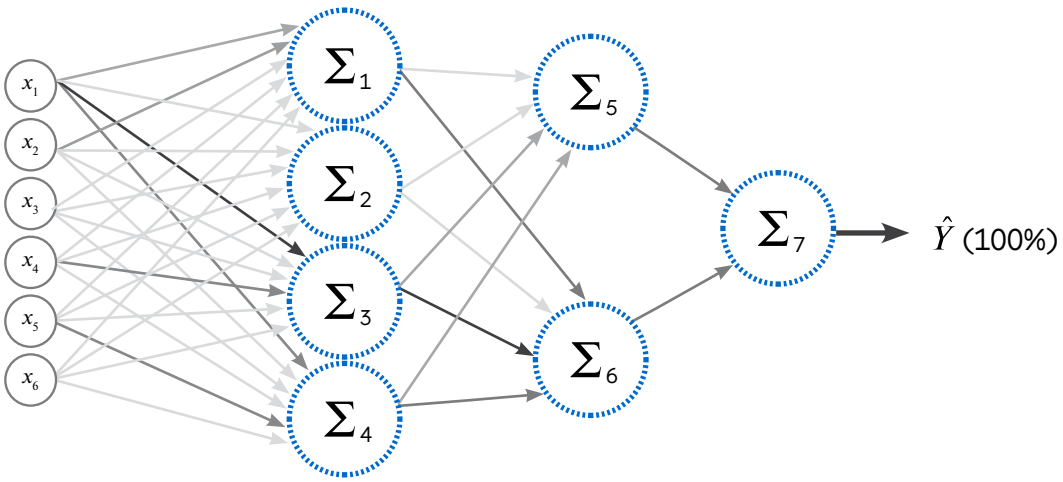

Fig. 3. Representation of the neural network model according to the 4-2-1 topology.

To construct the best logistic regression model, the maximum likelihood method was chosen as the most frequently used in such situations. Within the framework of this study, a standard version of regression model linear in both parameters and input factors was implemented.

As for the construction of the model using the decision tree and random forest methods, various tree options were considered. The quality of the classifications built on their basis was determined using a standard indicator – the F_1-measure [4, 19, 22]. In this case, to ensure the construction of effective classification trees, trees with different parameters were considered. The results are presented in *Table 4* and *Table 5*. In addition, *Table 6* shows the results of the comparison of classifications based on the best results of the models obtained.

Table 4 presents the results of assessing the quality of models constructed using the decision tree method. The following notations are used:

a is the minimum number of elements in one leaf of the tree;

b is the number of elements in a leaf at which no further splitting is performed;

F_1 is the value of the quality measure of classifications.

In *Tables 4* and *5*, the classification variants that are the best in terms of the maximum value of the quality measure are highlighted in bold. *Table 5* presents the results of assessing the quality of models constructed using the random forest method. The number of trees of the corresponding random forest is indicated in brackets after the F_1-measure.

Let us notice that for some combinations of random forest' parameters, the resulting classifications are the same, which is reflected in the F_1-measure values. In addition, note that further increase in the number of trees in the forest does not lead to changes in the values of the classification quality measure.

The classification results by the methods mentioned are presented in *Table 6*, where the designation "1" was used for "suitable" clients with whom it is profitable for the company to cooperate, and "0" for "unsuitable". In this case, the "observed values" correspond to the specialist's assessments, and the "predicted" ones correspond to the classification results by the methods under consideration.

Table 6 shows that the use of logistic regression resulted in six errors (9.4% of errors), three of which occurred for "suitable" clients, and the other three for "unsuitable" clients. When using the decision tree method, the number of errors for "suitable" clients was two, and for "unsuitable" clients – five (a total of 10.9% of errors). The random forest method resulted in four errors for "unsuitable" clients (a total of 7.8% of errors). In general, it can be noted that methods alternative to the neural network did not provide a complete match between the expert's assessments and the classification results. This allows us to conclude that the neural network is the most suitable for solving the problem under consideration for assessing the quality of clients.

It is important to understand that to check the correctness of the neural network, it is not enough to use only the data set that is currently available. It is important to ensure that its predictive properties are preserved for new portions of data, which from a statistical point of view means ensuring the adequacy of the model [6].

Table 4.

Parameters of the decision tree model and the quality of the resulting classifications

a	2			4		6	8	10
b	2, 4, 6	8	10	4, 6, 8	10	6, 8, 10	8, 10	10
F_1	0.876	**0.891**	0.857	0.889	0.857	0.842	0.842	0.844

Table 5.

Parameters of random forest models and the quality of the resulting classifications

a	2	2	2	2	2	4	4	4	4	6	6	6	8	8	10
b	2	4	6	8	10	4	6	8	10	6	8	10	8	10	10
$F_1(10)$	0.840	0.840	0.855	0.855	0.870	0.874	**0.889**	0.859	0.874	**0.889**	0.859	0.874	0.859	0.874	0.874
$F_1(50)$	**0.904**	**0.904**	**0.904**	0.889	0.889	0.874	**0.904**	0.874	0.874	0.889	0.874	0.874	0.874	0.874	0.874
$F_1(100)$	0.887	0.887	0.887	0.887	0.887	0.874	**0.889**	**0.889**	0.874	0.874	**0.889**	0.874	**0.889**	0.874	0.871

Therefore, after completing the training of the neural network, ten new LPGFS were taken to check its performance. Contracts with that consumers were concluded during the study and, naturally, were not included in the original database. The results of the neural network (in the form of estimates of the probability of the client's "suitability" and the final assessment) were compared with the specialist's assessments (*Table 7*).

The neural network correctly assessed nine positions out of ten. The error was made with the seventh station. The error in the probability assessment is less than one tenth (0.411 against 0.500, which would be enough to recognize the LPGFS as suitable for cooperation). This station has a high rating (4), sells five types of fuel and has an average capacity (four dispensers). However, it is 149 km from the company and has installed tanks of relatively low capacity (about 16 thousand liters). The volume of gas sales is less than 6 thousand liters, i.e. this LPGFS is quite average, but ambiguous in characteristics.

The neural network has assessed that cooperation with such a consumer will not be profitable, but the company's specialist has decided that it is suitable. This contradiction may be a sign that this LPGFS may need to conduct a more thorough study of its capabilities to ensure successful cooperation with the company. When a certain number of such "controversial" situations accumulate, a decision may be made to retrain the neural network on a larger database, or, if the results continue to be inconsistent with reality, a more radical action may be required – changing the topology of the neural network and its complete retraining.

Table 6.

Classification methods' quality estimation

		Predicted values by logistic regression		**Predicted values by decision tree**		**Predicted values by random forest method**		
		0	1	0	1	0	1	∑
Observed values	0	20	3	18	5	19	4	**23**
	1	3	38	2	39	0	41	**41**
	∑	**23**	**41**	**20**	**44**	**19**	**45**	**64**

Table 7.

Comparison of neural network and expert assessments for new customers

No. of LPGFS	1	2	3	4	5	6	7	8	9	10
Expert assessments	1	0	1	1	1	1	1	1	1	1
Probability assessment	0.954	0.121	0.851	0.884	0.732	0.999	0.411	0.970	0.804	0.925
Network assessment	1	0	1	1	1	1	0	1	1	1

At the *seventh stage*, for independent analysis of the consumer for compliance with the company's requirements, one can use a list of threshold values for each factor, which will help you quickly assess how clients with factor values close to average are suitable for further cooperation.

To build such a list, the neural network we developed was applied. The ranges of factor values were determined one by one, at which the client's "suitability" is maintained, while the values of all other characteristics were fixed at average levels (*Table 2*). Based on the weighting coefficients, the neural network independently calculates the values at which the probability assessment (assessment of the "client's reliability") will be in the range from 0.5 to 1, which corresponds to the decision that the client is suitable for cooperation. Thus, we will compile a table of threshold values (*Table 8*), which can be used as a hint for a specialist when assessing the degree of customer suitability.

Comparing the data in *Table 2* and *Table 8*, we can conclude that the modeling results obtained during the study allow us to speak about the correctness of using neural networks for the tasks of selecting consumers in the wholesale and retail trade of oil and gas products.

4. Discussion

In the process of testing the methodology we developed, it became necessary to construct an additional interpretation of the results, not only in terms of obtaining analytical and statistical conclusions, but also for developing visual and specific recommendations for interaction with the company's customers. For example, it is of interest to construct a geographic interpretation of the results obtained by depicting the approximate coverage area of the company on a map and analyzing it. The radius of this area according to *Table 8* is 156 km.

The distance threshold can be called the most important characteristic, since it is based on the factor that has the greatest weight in the neural network. It determines how far away potential consumers are and, particularly, how convenient it is for them to access LPGFS. This value plays a key role for the company, since it directly affects the cost of transporting gas to consumers, which, in turn, can significantly affect their own costs.

Table 8.

List of threshold values for customer characteristics

Factor	Admissible values	
	min	max
Distance, km	–	156
Assortment width, units	2	–
Production capacity, units	1	–
LPGFS rating, conv. units	1.1	–
Tank capacity, thousand liters	20	–
Sales volume, thousand liters	5.6	–

To clarify the factor of "the distance to the client", one can add a new characteristic that takes into account the area of the client's location relative to the company's location. However, it is important to understand that the weight coefficient of the new factor will "take away" some of the strength of the old one since they based on similar customer characteristics.

Another solution to this problem is to replace the existing factor with four new ones, each of which is a distance in a certain direction (for instance, north-south-west-east). Then the distance to each LPGFS will be taken into account by one or two separate factors (for example, city gas stations will be predominantly located in the west, north or south direction due to the specific geography of Novosibirsk). This approach will allow us to determine more accurate boundary values of distances and will help in a more accurate assessment of the coverage area, which will presumably be stretched in different directions. For example, the maximum possible distance to the east, towards Novosibirsk, will be noticeably greater than to the west, since most of the city, and, accordingly, the company's potential consumers are located on the eastern bank of the Ob River.

It is important to understand that determining the transition distance value is a complex task and may depend on many factors (transport infrastructure, population density, terrain topography, etc.). Therefore, when deciding on cooperation with a gas station, it is necessary to take these factors into account and conduct a detailed analysis of the market and the relevant infrastructure.

The threshold number of fuel types is important for LPGFS because it affects customer service and the efficiency of fuel management. This indicator may depend on various factors, such as the location of the gas station and the needs of local residents. Typically, remote stations offer no more than two or three types of fuel for cars, which corresponds to the threshold value found at two types of fuel. If the station carries out gas refueling, it is also sold in cylinders (pressure tanks), which can be convenient for car owners who use gas both as a fuel and for household needs.

The lower threshold value of the capacity of the LPGFS was found to be equal to one, which is most likely due to its low degree of importance (0.783%, *Table 3*). However, if the station has too few dispensers, it may be insufficient to meet the needs of all consumers. On the other hand, too high capacity may lead to excessive costs for the construction and maintenance of the gas station. Therefore, before selecting the capacity of the LPGFS, it is recommended to analyze the needs of consumers, the availability of gas pipelines and other factors in order to select the optimal capacity that will meet the requirements of all stakeholders and ensure maximum efficiency and cost-effectiveness of the LPGFS.

The minimum possible rating of the gas station was 1.1 points on a five-point scale, which is very low and is a consequence of the low level of customer satisfaction expressed in reviews of the LPGFS. To obtain a more objective picture of the quality of service and the level of customer satisfaction, it is necessary to study reviews from 2GIS and other sources, using the average weighted assessment, taking into account the reputation and reliability of the sources, as well as their quantity.

The threshold value of the volume of tanks installed at the LPGFS is 20 thousand liters, which is quite consistent with the volumes of average gas filling stations (*Table 2*, the average is 19.983 thousand liters). However, if the costs significantly exceed expectations, then the installed tanks may be sufficient for no more than a day, which will entail frequent refueling several times a day, and, therefore, may require the installation of additional tanks to increase the total capacity.

Minimum daily sales at LPGFS is an important factor that determines the minimum volume of fuel that must be sold (and, respectively, be available at the time of sale) for the station to remain profitable. For an average LPGFS the minimum daily sales are 5.600 liters. However, this value is quite low for city stations that service a large number of cars. At the same time, for LPGFS on the outskirts of the city or in sparsely populated areas, where the volume of fuel sales is lower, this value will be more than sufficient.

In further modifications of the neural network, it could be reasonable to replace this factor with a more objective one, such as the ratio of sales to the population of the area where the gas station is located, to take into account the potential demand for fuel in a particu-

lar area. In addition, it is possible to calculate separate factors for the city and region, based on the characteristics of the regional fuel market.

Of additional interest is the study using a neural network of how the threshold values can change when not only one, but also two (three, etc.) other factors change simultaneously. An example of such a calculation, carried out using regression analysis methods to construct the corresponding dependencies [6], is shown in *Fig. 4*.

Analyzing *Fig. 4*, we can say that the nature of the change in the maximum possible distance varies significantly: for consumers selling four types of fuel ($x_2 = 4$), with an increase in rating, the distance decreases, while with three or four types of fuel, it first increases and then decreases. Moreover, we can determine the maximum permissible distance, which for consumers with $x_2 = 2$ is 193 km, and for customers with $x_2 = 3$ is 138 km.

The threshold values so obtained can be applied for fast assessment of consumers when concluding a contract. During further development of the neural network, the factors can change both quantitatively and qualitatively; therefore timely updating of the pool of input factors and the contents of the database will allow us to obtain more correct assessments of consumers.

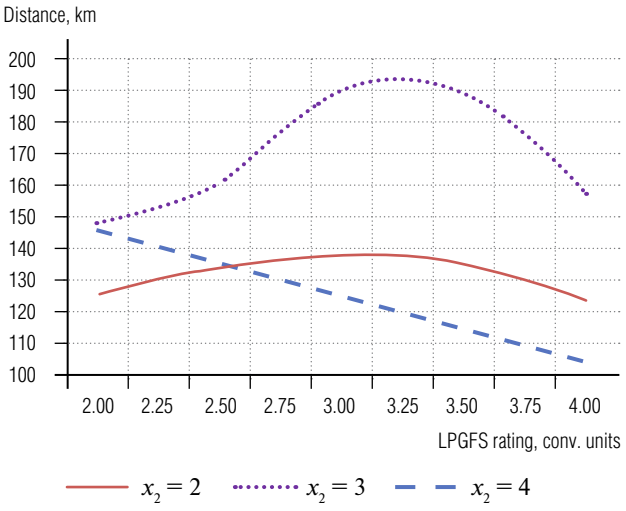

Fig. 4. Comparison of upper distance thresholds depending on rating and number of fuel types.

Conclusion

The result of using the author's methodology is the construction of a decision-making model for choosing consumers with whom the company plans further mutually beneficial cooperation. For this purpose, a pool of factors characterizing consumers and their purchasing behavior was formed. Analyzing the database of the company's clients and expert assessments determining their reliability, neural networks were developed that allow us to assess the prospects of cooperation with clients. With their help, the problem of ranking factors characterizing consumers was solved in relation to the degree and nature of their influence on the decision-making process on the reliability of a particular consumer.

Based on the results of neural network training, a network with the "best" topology was selected, which ensured correct forecasting for all database records. Comparison of the results of this neural network's outputs with classifications built on the basis of other data mining methods allowed us to conclude that the neural network model is the best fit for solving the problem under consideration.

The neural network we constructed was used to determine the threshold values of the input factors of the model. This allowed us to develop recommendations for the company's employees on selecting proposals for cooperation with consumers.

Further improvement of the proposed methodology may consist in expanding the pool of input factors by involving new consumer characteristics, including those proposed in this paper, as well as by splitting the existing factors into several components, each of which characterizes a certain specificity of consumers. In addition, for neural networks with a large number of input variables, it makes sense to consider more complex topologies that can include not only additional internal layers, but also feedback.

Due to its universality, the methodology we developed can be recommended for solving various classification problems not only in the field of logistics, but also for a wide range of economic and managerial problems. ■

References

1. Rosenblatt F. (1965) *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Moscow: Mir (in Russian).

2. Rumelhart D.E., Hinton G.E., Williams R.J. (1986) Learning internal representations by error propagation. *Parallel Distributed Processing, Cambridge, MA, MIT Press*, vol. 1, pp. 318–362. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2

3. Galushkin A.I. (2010) *Neural networks: basic theory.* Moscow: Hotline –Telecom (in Russian).

4. Hinton G., Le Cun Y., Bengio Y. (2015) Deep learning. *Nature,* vol. 521, pp. 436–444. https://doi.org/10.1038/nature14539

5. Sevastyanov L.A., Shchetinin E.Yu. (2020) On methods for increasing the accuracy of multi-class classification on unbalanced data. *Informatics and its applications*, vol. 14, no. 1, pp. 63–70 (in Russian). https://doi.org/10.14357/19922264200109

6. Timofeev V.S., Faddeenkov A.V., Shchekoldin V.Yu. (2016) *Econometrics.* Moscow: YURAYT (in Russian).

7. Tsoi M.E., Shchekoldin V.Yu. (2021) *Marketing research: methods for analyzing marketing information.* Novosibirsk: Publishing house of NSTU (in Russian).

8. Breiman L. (2001) Random Forests. *Machine Learning,* vol. 45, no. 1, pp. 5–32. https://doi.org/10.1023/A:1010933404324

9. Chistyakov S.P. (2013) Random forests: a review. *Proceedings of the Karelian Scientific Center of the Russian Academy of Sciences*, no. 1, pp. 117–136 (in Russian).

10. Abbate R., Manco P., Caterino M., et al. (2022) Demand forecasting for delivery platforms by using neural network. *IFAC-Papers OnLine*, vol. 55, no. 10, pp. 607–612. https://doi.org/10.1016/j.ifacol.2022.09.465

11. Danilchenko M.N., Muravnik A.B. (2021) Neural network approach to route construction in a special-purpose automated control system. *High-tech technologies in space exploration of the Earth,* vol. 13, no. 1, pp. 58–66 (in Russian). https://doi.org/10.36724/2409-5419-2021-13-1-58-66

12. Sustrova T. (2016) A suitable artificial intelligence model for inventory level optimization. *Trends Economics and Management,* vol. 10(25), pp. 48–55. https://doi.org/10.13164/trends.2016.25.48

13. Mikhailin D.A. (2017) Neural network algorithm for safe flight around air obstacles and prohibited ground zones. *Scientific Bulletin of MSTU GA*, vol. 20, no. 4, pp. 18–24 (in Russian). https://doi.org/10.26467/2079-0619-2017-20-4-18-24

14. Hughes A. (1996) *Boosting Response with RFM*. New York: Marketing Tools.

15. Griffin J. (2002) *Customer loyalty: how to earn it, how to keep it.* San Francisco, CA: Jossey-Bass.

16. Guo Li. (2011) A research on influencing factors of consumer purchasing behaviors in cyberspace. *International Journal of Marketing Studies*, vol. 3, no. 3, pp. 182–188. https://doi.org/10.5539/ijms.v3n3p182

17. Tsoi M.E., Shchekoldin V.Yu., Lezhnina M.N. (2017) Construction of segmentation based on modified RFM analysis to increase consumer loyalty. *Russian Entrepreneurship*, vol. 18, no. 21, pp. 3113–3134 (in Russian). https://doi.org/10.18334/rp.18.21.38506

18. Saunders M., Lewis F., Thornhill E. (2006) *Methods of conducting economic research*. Moscow: EKSMO (in Russian).

19. Demsar J., Curk T., Erjavec A., et al. (2013) Orange: Data mining toolbox in Python. *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353.

20. Sturges H. (1926) The choice of a class-interval. *Journal of the American Statistical Association,* vol. 21, pp. 65–66. https://doi.org/10.1080/01621459.1926.10502161

21. Prieto A., Prieto B., Ortigosa E.M., et al. (2016) Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing,* vol. 214, pp. 242–268. https://doi.org/10.1016/j.neucom.2016.06.014

22. Upton G. (1982) *Analysis of contingency tables.* Moscow: Finance and Statistics (in Russian).

23. Cochran W. (1976) *Sampling methods.* Moscow: Statistics (in Russian).

24. Bakhvalov N.S., Zhidkov N.P., Kobelkov G.M. (2020) *Numerical methods.* Moscow: Knowledge Laboratory (in Russian).

25. Liker J. (2005) *Dao Toyota: 14 principles of management of the world's leading company.* Moscow: Alpina Business Books (in Russian).

About the authors

Valeriya A. Nazarkina

Cand. Sci. (Econ.), Assoc. Prof.;

Associate Professor, Marketing and Service Department, Novosibirsk State Technical University (NSTU), 20, Karl Marx Ave., Novosibirsk 630092, Russia;

E-mail: valeria71@bk.ru

ORCID: 0000-0002-2207-5228

Vladislav Yu. Shchekoldin

Cand. Sci. (Tech.);

Associate Professor, Marketing and Service Department, Novosibirsk State Technical University (NSTU), 20, Karl Marx Ave., Novosibirsk 630092, Russia;

E-mail: schekoldin@corp.nstu.ru

ORCID: 0000-0001-8016-1282