UDC 004.852
https://doi.org/10.47533/2023.1606-146X.38

ZH. E. TEMIRBEKOVA*, Z. M. ABDIAKHMETOVA , B. IMANBEK, G. TURKEN

Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: temyrbekovazhanerke2@gmail.com, zukhra.abdiakhmetova@gmail.com, baks_teen@mail.ru, turken.gulzat@gmail.com

COMPARATIVE ANALYSIS OF DATA CLASSIFICATION METHODS FOR PREDICTION OF TRADE-IN AUTO PRICES

In the presented article, machine learning algorithms are used to predict the price of cars. Forecasting the price of cars is one of the most important issues in modern times, because the number of car users is increasing year by year worldwide. Therefore, it may be interesting for many car owners to know the approximate price of cars in advance. The vehicle's build date, mileage, size and other parameters are essential data for the machine learning process. Based on these parameters, the data needed to predict the price of any car was classified and a training dataset was created. Based on this data set, some interesting predictions were made. The purpose of the article is to consider the pre-processing of data, to determine and analyze what achievements the direction of artificial intelligence has achieved on the basis of predicting the price of cars. Hybrid forecasting methods using statistical analysis and machine learning methods were used in the research.

Keywords*: machine learning, classification problems, logistic regression, random forest, decision tree, k-nearest neighbor, REST API.*

Introduction. Machine learning (ML) has become one of the most exciting and breakthrough technologies of our time [1, 2]. Such large companies as Google, Apple, Microsoft, Amazon, and others invest significant capital in the development of methods and applications in this field of research, opening the way to new opportunities, becoming an integral part of life every day. For example, the Kaspi bank application makes a decision on loan approval or when Netflix recommends a movie that you might like, conversations with speech assistants on a smartphone are done using machine learning algorithms.

Working in the field of sales of new passenger cars and light commercial vehicles, we are faced with a global problem as a decline in production and new problems in logistics related to the disruption of supply chains. The key problem for the automotive industry since the summer of 2020 remains the shortage of electronic components, which is why car factories are forced to reduce the production of cars and go into downtime. This led to a shortage of cars and an increase in prices for new passenger cars. Compared to October 2020, sales fell by 18.1% in 2021. Analysts say that the automotive industry still has a long way to go to overcome the current crisis. Therefore, the management of R-Motors LADA LLC decided to compensate for the decline in sales of new cars by buying out a secondary car for further resale.

If the pricing in the primary car market includes logistics, taxes, the desired profit of the dealer, and the salary of the chain of its employees, then the factors of price formation for trade-in cars are much more extensive. Therefore, it is important to assess the condition of

* E-mail корреспондирующего автора: temyrbekovazhanerke2@gmail.com

the machine as objectively as possible and, in accordance with this, set the cost, taking into account the indicators as: year of manufacture of the car; technical condition and condition of the body; mileage; features of the configuration; the time of sale (even the season in which the car is put up for sale has an impact on demand and, accordingly, the cost); the demand for the model in the market; service history.

The traditional approach to pricing is based entirely on the word of an expert who makes a decision based only on his experience.

Machine learning uses complex algorithms to take into account many factors and set the right prices for thousands of products in almost seconds [4]. Pricing models based on machine learning determine the patterns of the data obtained, which makes it possible to determine prices taking into account factors that the buyout manager might not even have guessed.

In practice, it is always recommended to compare the quality of at least several different learning algorithms in order to choose the best model for a particular task, since the most experienced data processing and analysis specialists will not be able to tell which algorithm is more efficient [3]. Algorithms may differ in the number of features or samples, the noise level in the data set, and whether classes are linearly separable or not. Within the framework of this article, classification methods such as logistic regression, random forest, decision tree, k nearest neighbors for predicting prices for supported cars using machine learning technology will be considered [4].

Materials and methods of research. The classification task is a subcategory of machine learning methods with a teacher, the purpose of which is to determine categorical class labels for the following instances based on historical observations [5]. Here, the definition of "with a teacher" refers to a collection of samples in which the necessary class membership labels are already known. When teaching with a teacher, a model is extracted, based on classification algorithms and from labeled training data, which allows making predictions about previously unknown or future data [6]. Another subcategory of teaching methods with a teacher is regression, where the result is a continuous value. Labels in the classification can have a binary nature, for example, filtering mail for spam and not spam. A typical example of multiclass classification is handwritten character recognition. There are many classification methods with different approaches to implementation. Each algorithm has its own characteristics and is based on certain assumptions. Ultimately, the quality of the classifier, the percentage of prediction accuracy depends on the training of the algorithm. During algorithm training, steps such as feature selection, selection of qualitative metrics, selection of classifier and optimization algorithms, evaluation of model quality, fine-tuning of the algorithm are involved. Classifiers based on the Decision Trees algorithm (DT) [7] are a hierarchical tree-like structure (subsets) that were formed by making decisions based on the formulation of a number of questions [8].

Step one: exclusion of features that do not carry a semantic load for incoming analysis. In this dataset, this is the id, vin code, links to the site where you can see the car in detail, coordinates.

Step two: imputation and deletion of data [9]. This is the process of replacing missing, incorrect values with other values.

Step three: Correlation analysis. It is the basis of statistical data analysis, the purpose of which is to determine the presence of any significant relationships, patterns or trends. The

result of this analysis is the correlation coefficient, which shows how strong the relationship between two variables in the data set is [10].

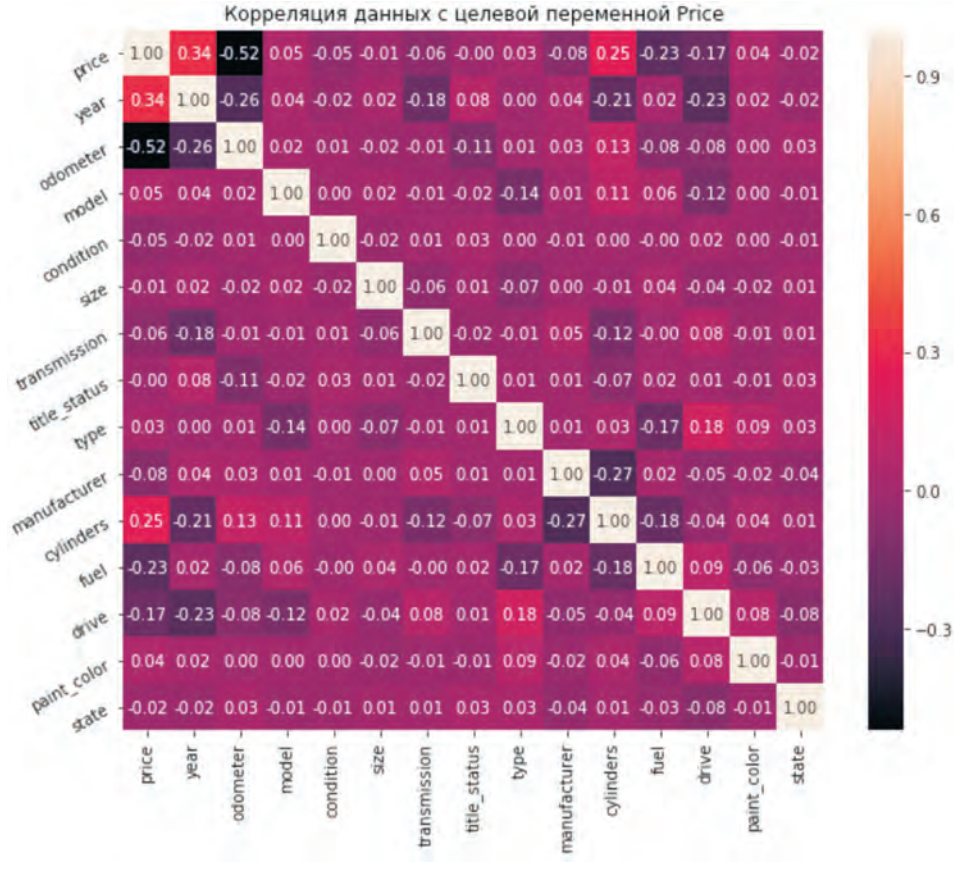

Figure 1 – Visualization of correlation analysis

In Figure 1 the correlation coefficient with the objective function is low, which may lead to a less accurate prediction. A heat map will be more effective in presenting data if redundant data is removed, which acts as distracting noise to data analysis. Step four: Get rid of emissions. Outliers are very different from other data sets, due to variability in measurements or during a data entry error [11-17]. If possible, outliers should be excluded from the dataset. However, detecting these abnormal instances can be difficult and not always possible.

Step five: Processing categorical data. The prediction result of algorithms such as decision tree can be obtained directly from categorical data without data transformation, when algorithms like KNN cannot work with categorical data directly. They require all input and output variables to be numeric. Therefore, to encode class labels, the LabelEncoder method of the scikit-learn library was used, which encodes dummy variables for categorical data once. Then you can apply a dictionary of correspondences to convert class labels to integers (Table 1).

Table 1 – Dictionary of correspondence after assigning labels

№	Drive	Fuel	Color	Sign
1	FWD	Gas	Red	0
2	RWD	Diesel	White	1
3	AWD	Petrol	Black	0
4	4WD	Electric	Gray	1

Results and discussions. The training data was taken from the Haraba service [18]. This is a database of ads for supported cars from all over Russia since 2017. Data exchange with the Haraba service is carried out using the REST API architecture [19]. For this purpose, the Windows Service [20] has been written, the task of which is to send a request to Haraba every 10 minutes to receive new ads.

By requesting all historical data from the service, we get a dataset that has passed 5 stages of data processing, as described above. During the experiment, the proportion of 80:20 was used, thus dividing the set into training and test data. The task is to train the model to analyze each factor that affects pricing, and choose the most optimal among the 4 algorithms under consideration. Let's use the error matrix to visually represent the result of predicting the classifier of k nearest neighbors [6]. The values of the matrix give a summary of correct and incorrect forecasts, broken down by each category. The matrix shows 0 + 2175 = 2175 correct forecasts and 257 + 5 = 262 incorrect forecasts (Figure 2).

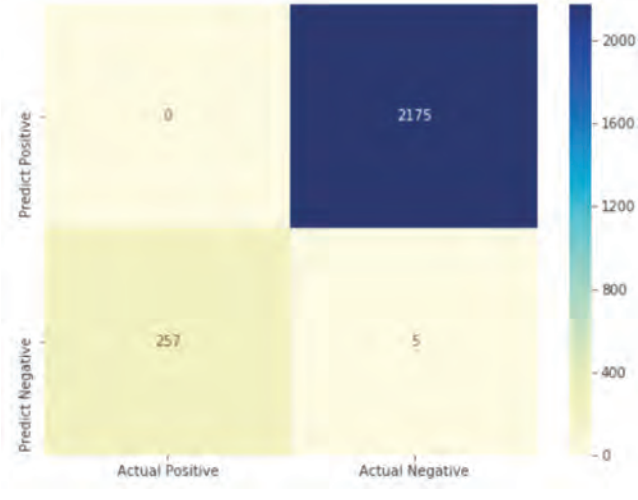

Figure 2 – Evaluation of the classifier of k nearest neighbors using the error matrix

Classifying data using the k nearest neighbors algorithm shows the accuracy of the model was 0.86 at k = 5. During the experiment, the following questions were considered 2, 3, 4, 5, 6, 7 neighbors in the KNN model. With five or more neighbors, the boundaries of the solution showed smoother boundaries, assuming an optimal balance between over-training and under-training. Since the number of votes in the implementation of the KNN algorithm

between 5 and 6 neighbors is the same, it is preferable to choose the neighbors with the smallest distance to the sample. The average time to train the classifier took 1115.65 ms. In logistic regression, we use the default value C = 1 (inversion force of regularization). This provides good performance with 0.89 accuracy for both training and test suite. The result given using the error matrix shows 1188 + 51 = 1249 correct predictions and 38 + 0 = 38 incorrect predictions (Figure 3). We will also give the result using a probability histogram (Figure 4). Among the classifiers under consideration, logistic regression turned out to be the fastest, showing a result of 215.15 msec.

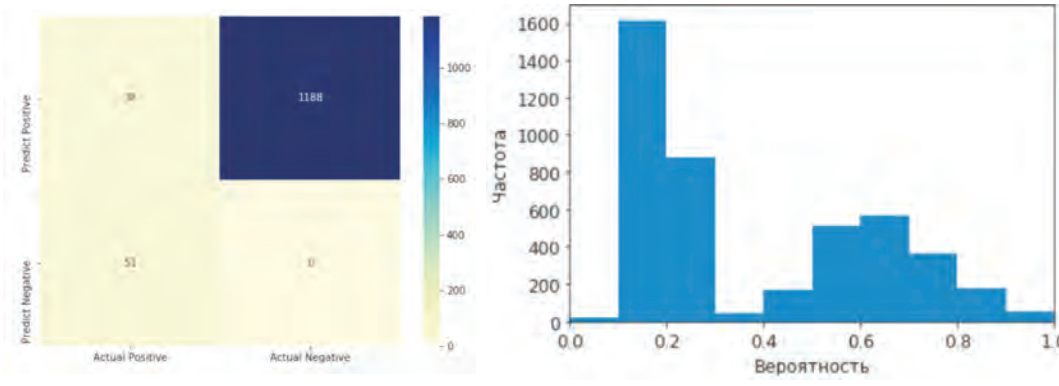

Figure 3 – a) Comparison of predictions with actual data using a logistic regression algorithm;
Figure 4 – b) Histogram of the probability of determining the price by logistic regression

In Figure 6, the histogram has a positive bias. The second column tells us that there are approximately 1600 observations with a probability from 0 to 0.2. There are a small number of observations with a probability greater than 0.5.

The result of forecasting using the decision tree method, with default parameters, also shows an excellent result with an accuracy of 0.93. 412.07 msec was spent on training, second in speed only to the model based on the logistic regression algorithm. The prediction result is presented using the error matrix (Figure 5).

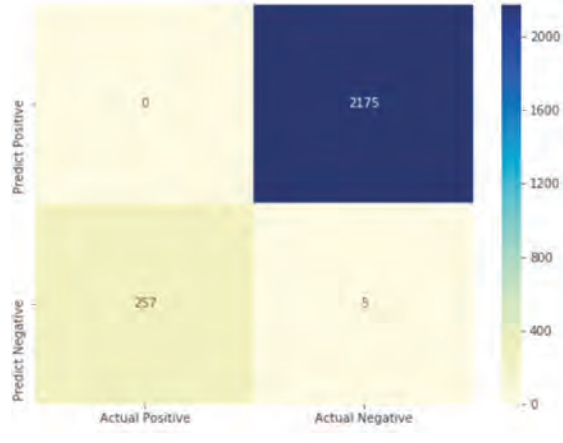

Figure 5 – Classifier evaluation decision tree using error matrix

And the last model considered in this article is a random forest. The model revealed more patterns in the data, showing the accuracy of the forecast at 94%, spending 541.03 msec on training. The result is presented as a ROC curve (Figure 6):

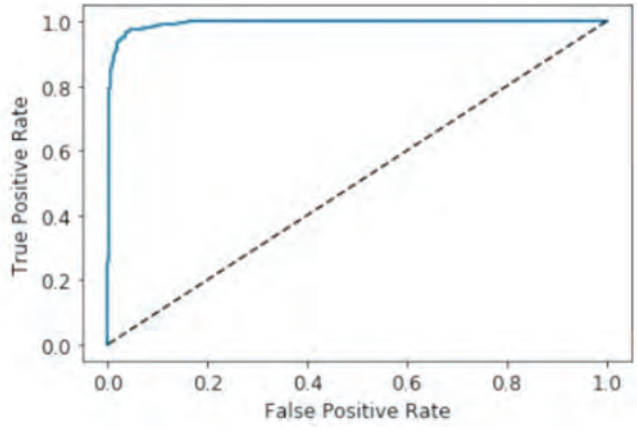

Figure 6 **–** ROC curve for a random forest

Since the classifier based on the random forest algorithm showed high results in predicting prices, it was decided to use the model to buy a secondary car.

Conclusion. This article describes how the current realities of the shortage of electronic components have led dealerships to resell supported cars, and how machine learning methods help to identify whether the price is too high and allows you to find the optimal solution in this segment. As a result, we got a system that requests new ads from the Haraba service every 10 minutes, analyzes the data received and, finding patterns, predicts the price and future demand for cars. Thanks to the introduction of machine learning in pricing issues, the company has optimized operational efficiency, using algorithms for price recommendations and sales forecasts, allowing managers to focus on strategic tasks.

REFERENCES

1 Narender Kumar, Dharmender K. Machine Learning based Heart Disease Diagnosis using Non-Invasive Methods 2021 J. Phys.: Conf. Ser. 1950 012081.

2 Alarsan, F.I., Younes, M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. J Big Data 6, 81 (2019). https://doi.org/10.1186/s40537-019-0244-x.

3 S. Rashka. Python ı mashınnoe obucheneıe [Python and machine learning]. — Moskva: DMK press, 2017. — 265 p.

4 Akhmed-Zaki D.Zh., Mukhambetzhanov S.T., Nurmakhanova Zh.M. and Abdiakhmetova Z.M. Using Wavelet Transform and Machine Learning to Predict Heart Fibrillation Disease on ECG 2021 IEEE International Conference on Smart Information Systems and Technologies 28-30 April, 2021, Nur-Sultan DOI 10.1109/SIST50301.2021.9465990.

5 R. Bil'bro, T. Oheda, B. Bengfort. Language-Aware Data Products with Machine Learning. — O'Reilly Media, 2018. — 313 p.

6 Zheron O. Prikladnoe mashinnoe obuchenie s pomoshch'yu Scikit-Learn i TensorFlow [Applied Machine Learning with Scikit-Learn and TensorFlow]. — Dialektika Vil'yams Al'fa-kniga, 2018. — 465 p.

7 Dzh.V. Plas. Python dlya slozhnyh zadach nauka o dannyh: mashinnoe obuchenie [Python for Complex Problems Data Science: Machine Learning]. — Piter, 2018. — 265 p.

8 S. Boslav. Statistika dlya vsekh [Statistics for everyone]. — Moskva: DMK press, 2015. – 201 p.

9 B. Bengford, R.Bilbro. Prikladnoj analiz tekstovyh dannyh na Python. Mashinnoe obuchenie i sozdanie prilozhenij obrabotki estestvennogo yazyka [Applied analysis of text data in Python. Machine Learning and Building Natural Language Processing Applications]. Sankt-Peterburg: 2019. —181 p.

10 H. Brink, D. Richards, M. Feverolf. Mashinnoe obuchenie [Machine learning] / Biblioteka programmista. — Piter, 2019. —304 p.

11 F. Chollet. Deep learning with Python. — Manning, 2020. —169 p.

12 A. Burkov. Mashinnoe obuchenie bez lishnih slov [Machine learning without further ado]/ Biblioteka programmista. — Piter, 2020. — p. 60–69.

13 N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi. A deep learning approach tonetwork instrusion detec-tion// IEEETrans. Emerg. TopicsComput. Intell.–Vol. 2, no. 1.–Pp. 41–50.–Feb. 2018.

14 Chistyakov C. P. Sluchajnye lesa: obzor // Trudy Karel'skogo nauchnogo centra RAN. 2013. No 1.st. 117–136.

15 Grishanov K. M., Belov Yu. S. Metod klassifikacii K-NN i ego primenenie v raspoznavanii simvolov /Fundamental'nye problemy nauki. Sbornik statej Mezhdunarodnojnauchno-prakticheskoj konfer-encii 15 maya 2016 g. CH. 3. Tyumen': NIC Aeterna, 2016. S. 30-33

16 Jenhani I., Amor N. B., Eloued Z. Decision trees as possibilistic classifiers // International Journal of Approximate Reasoning. – no. 48 (2008).–pp. 786–801.–Nov. 2008. https://doi.org/10.1016/j.ijar.2007.12.002.

17 Rymarczyk T., Kozłowski E. Logistic Regression for Machine Learning in Process Tomography //MDPI. – no. 19(15). – pp. 206-208. –Aug. 2019. https://doi.org/10.3390/s19153400

18 https://haraba.ru.

19 Sanjay P. Pro RESTful APIs Design, Build and Integrate with REST, JSON, XML and JAX-RS Apress, Berkeley, CA, 2018. https://doi.org/10.1007/978-1-4842-2665-0

20 Stephen R.G. Fraser Windows Services. In: Pro Visual C++/CLI and the .NET 2.0 Platform. Apress. 2006. https://doi.org/10.1007/978-1-4302-0109-0_14.

Ж. Е. ТЕМИРБЕКОВА, З. М. АБДИАХМЕТОВА, Б. ИМАНБЕК, Г. ТУРКЕН

Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

TRADE-IN АВТО БАҒАЛАРЫН БОЛЖАУ КЕЗІНДЕ ЖІКТЕУ ӘДІСТЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ

Ұсынылып отырған мақалада автокөліктердің бағасын болжау мақсатында машиналық оқыту алгоритмдері қолданылады. Көліктердің бағасын болжау – қазіргі заманда өзекті мәселелердің бірі, оның себебі автокөліктер қоланушылардың саны жылдан жылға дүние жүзі бойынша геометриялық прогрессиямен артуда. Сондықтан бұл мәселеде көліктердің бағасын шамамен алдын ала біліп отыру көптеген автокөлік иелері үшін қызықты болуы мүмкін. Көліктің құрастырылған мерзімі, жүгірісі, көлемі және басқа да параметрлері машиналық оқыту үрдісі үшін қажетті мәлімет болып табылады. Осы параметрлер негізінде кез-келген автокөлік бағасына болжауға қажетті деректерді жіктеп, оқытуға қажетті датасетті құрылған болатын. Осы мәліметтер жиыны негізінде біршама қызықты болжамдар жасалды. Мақаланың мақсаты

– деректерді алдын-ала өңдеуді қарастырып, жасанды интеллект бағыты автокөліктердің бағасын болжау негізінде қандай жетістіктерге жеткенін айқындау және талдау. Зерттеуде статистикалық талдаулар мен машиналық оқыту әдістерін қолдана отырып гибридті болжау әдістері пайдаланылған.

***Түйін сөздер**: машиналық оқыту, жіктеу мәселелері, логистикалық регрессия, кездейсоқ орман, шешім ағашы, k-жақын көршілер, REST API.*

Ж. Е. ТЕМИРБЕКОВА[1], З. М. АБДИАХМЕТОВА,
Б. ИМАНБЕК , Г. ТУРКЕН

Казахский национальный университет им.аль-Фараби, г.Алматы, Казахстан

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ КЛАССИФИКАЦИИ ДАННЫХ ПРИ ПРОГНОЗИРОВАНИИ ЦЕН TRADE-IN АВТО

В представленной статье алгоритмы машинного обучения используются для прогнозирования цен автомобилей. Прогнозирование цен на автомобили является одним из наиболее важных вопросов в наше время, ведь количество пользователей автомобиля увеличивается с каждым годом во всем мире. Поэтому многим автовладельцам интересно заранее узнать приблизительную цену автомобилей. Дата сборки автомобиля, пробег, размер и другие параметры являются важными данными для процесса машинного обучения. На основе этих параметров были классифицированы данные, необходимые для прогнозирования цены любого автомобиля, и создан обучающий набор данных. На основе этого набора данных были сделаны некоторые интересные прогнозы. Цель статьи – рассмотреть предварительную обработку данных, определить и проанализировать, каких достижений добилось направление искусственного интеллекта на основе прогнозирования цены автомобилей. В исследовании использовались гибридные методы прогнозирования с использованием статистического анализа и методов машинного обучения.

***Ключевые слова**: машинное обучение, задача классификации, логистическая регрессия, случайный лес, дерево принятия решений, k-ближайших соседей, REST API.*