UDC 004.6
DOI: 10.15587/2706-5448.2021.244003
Article type «Reports on Research Projects»

Eduard Kinshakov,
Yuliia Parfenenko,
Vira Shendryk

COMPARATIVE ANALYSIS OF METHODS FOR PREDICTION CONTINUOUS NUMERICAL FEATURES ON BIG DATASETS

The object of research is the process of choosing a method for predicting continuous numerical features on big datasets. The importance of the study is due to the fact that today in various subject areas it is necessary to solve the problem of predicting performance indicators based on data collected from different sources and presented in different formats, which is the task of big data analysis. To solve the problem, the methods of statistical analysis were considered, namely multiple linear regression, decision trees and a random forest. An array of extensive data was built without specifying the subject area, its preliminary processing, analysis was carried out to establish the correlation between the features. The processing of the big data array was carried out using the technology of parallel computing by means of the Dask library of the Python language. Since working with big data requires significant computing resources, this approach does not require the use of powerful computer technology. Prediction models were built using multiple linear regression methods, decision trees and a random forest, visualization of the prediction results and analysis of the reliability of the constructed models. Based on the results of calculating the prediction error, it was found that the greatest prediction accuracy among the considered methods is the random forest method. When applying this method, the prediction accuracy for a dataset of numerical features was approximately 97 %, which indicates a high reliability of the constructed model. Thus, it is possible to conclude that the random forest method is suitable for solving prediction problems using large data sets, it can be used for datasets with a large number of features and is not sensitive to data scaling. The developed software application in Python can be used to predict numerical features from different subject areas, the prediction results are imported into a text file.

Keywords: *machine learning, data analysis, big data, linear regression, decision tree, random forest.*

Received date: 20.05.2021
Accepted date: 02.08.2021
Published date: 07.12.2021

© The Author(s) 2021
This is an open access article under the Creative Commons CC BY license

How to cite
Kinshakov, E., Parfenenko, Y., Shendryk, V. (2021). Comparative analysis of methods for prediction continuous numerical features on big datasets. Technology Audit and Production Reserves, 6 (2 (62)), 15–17. doi: http://doi.org/10.15587/2706-5448.2021.244003

1. Introduction

Today, when solving problems of searching for patterns and predicting the values of variables for future periods of time, there is an increasing need to process big amounts of data, which, as a rule, are heterogeneous and unstructured. In a general sense, the term big data refers to a growing set of data presented in different formats, characterized by volume, speed, variety and reliability [1]. Given the significant volumes of data, working with them is not limited to database management systems, which are not capable of processing large amounts of heterogeneous data, but requires the development of appropriate technologies for processing big data. Big data changes over time, requires fast processing and appropriate management decisions based on these changes. Managing big data is necessary to prepare it for the data analysis task. It includes extracting data from different sources, storing it in a form suitable for fast mining, cleaning data from gaps and errors, converting data into a single format. Given the large volume, the analysis of big data should occur automatically and give the result to the decision-maker in such a way that he/she could unambiguously interpret the result [2, 3].

The task of analyzing big data arrays is relevant for various subject areas, in particular economics, finance, medicine, energy, requiring the processing of information from different sources – websites, social networks, automated control systems, etc. The analysis of big data is aimed at improving the efficiency of management decisions, solving the problems of strategic planning. Among the tasks of analyzing big data, there are descriptive analysis, prediction and prescriptive analytics [4]. This work is devoted to the analysis of the effectiveness of prediction methods using big data.

Thus, the object of research is the process of choosing a prediction method with the uncertainty of the subject area and the nature of features on big data samples. *The aim of research* is to analyze methods for predicting continuous numerical features using big datasets.

ISSN 2664-9969

2. Methods of research

Prediction, as a kind of data analysis tasks, is used to find patterns and establish relationships between data. Methods for predicting big datasets can be divided into methods of statistical analysis and machine learning [5]. In this work on prediction continuous numerical features, methods of linear regression [6], decision trees [7] and random forest [8] are considered.

Multiple linear regression is used to establish a linear relationship for the original variable y_i on n independent input variables x_{ij}. The hypothesis function for multiple linear regression can be represented as [9]:

$$y \approx h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_j x_j, \quad (1)$$

where $\theta = \{\theta_0, \theta_1, ...\theta_j ...\theta_n\}$ – vector of parameters, unknown; n – the number of independent variables; $x = \{x_1, x_2, ...x_n\}$ – independent variables-regressors.

Decision tree method refers to supervised machine learning algorithms. A decision tree is a branched structure representing the hierarchy of all possible prediction results. To build a decision tree, it is necessary to select the attribute by which the splitting in the node will be performed, select the method for cutting off branches, and determine the criterion for stopping learning.

As a criterion by which the correctness of the choice of an attribute for splitting a decision tree into branches can be determined, a statistical approach can be used, which consists in determining the Gini index, which makes it possible to assess the uneven distribution of the studied feature [10]:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2, \quad (2)$$

where T – current node; n – the number of classes; p_i – probability of the i-th class at the node T.

The decision tree allows to formulate clear regression rules, but the task of constructing a decision tree of the optimal structure is difficult.

The random forest method refers to ensemble prediction methods that combine multiple machine learning models. A random forest is an ensemble of classification trees that have been generated from subsamples of data using bagging [11].

The final classifier averages the individual classification algorithms as follows:

$$a(x) = \frac{1}{M} \sum_{i=1}^{M} a_i(x), \quad (3)$$

where $a_i(x)$ – classifier for the i-th sample; M – the number of data samples.

The random forest method has a number of advantages over statistical methods and decision trees for big data, since it is practically insensitive to feature scaling, data outliers, and is capable of analyzing data with a large number of features.

To study the effectiveness of the prediction methods presented above, a dataset was created based on anonymized data obtained from the Upwork online resource. The generated dataset initially contained the numeric values of 16 input variables and 2 target variables with a size of 1,815,696 rows.

3. Research results and discussion

The analysis of methods for predicting numerical features using a set of numerical data was carried out according to the algorithm shown in Fig. 1. For data processing, the Dask library of parallel computations of the Python language was used, which allows processing big amounts of data on a computer with up to 4 GB of RAM and a multi-core processor. Data in Dask is stored as a multi-dimensional array, which is a collection of NumPy arrays.

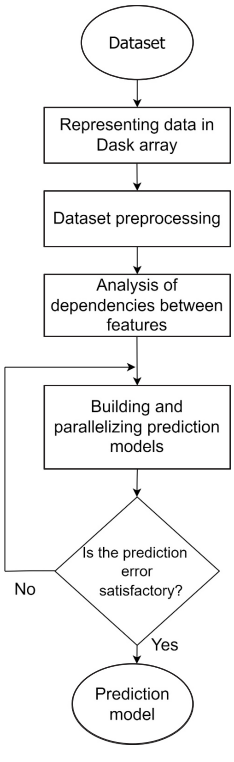

Fig. 1. Algorithm for the study of prediction methods

First, the preliminary processing of the dataset was carried out: the spaces were removed, the data was checked for the presence of text characters, duplicates, and the search for possible data outliers. After that, the analysis of the dependencies between the input features was carried out and it was found that linear stochastic and nonlinear dependencies are not traced. A correlation map was built and uninformative features were found that did not affect the original variables that were removed from the dataset.

Since the dataset contained both integer data and floating point numbers, the data was normalized to cast them to the same type.

Machine learning models were built using linear regression methods, decision trees and a random forest using Python libraries. The Google Colab service was used to test the models. The results of prediction the values of target variables Target 1 and Target 2 for the random forest model are shown in Fig. 2, 3, which shows the correlation between actual and predicted values for a test sample of 363,140 values.

The accuracy of prediction methods was assessed using the criteria of MSE mean square error, MAE mean absolute error, and MAPE mean absolute percentage error. The data were divided into training and test samples in the ratio of 1452556 to 363140. The results of the calculated prediction errors are presented in Table 1.

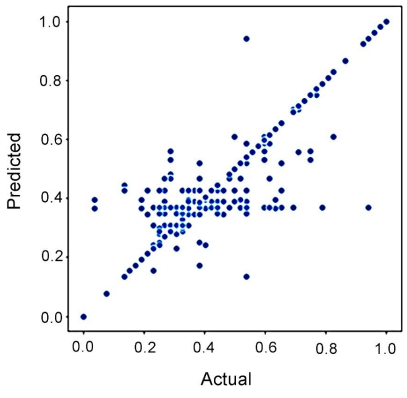

Fig. 2. The result of prediction for the Random forest method (Target 1)

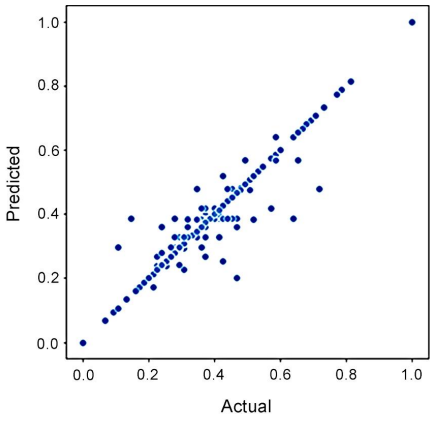

Fig. 3. The result of prediction for the Random forest method (Target 2)

Table 1

Assessment of the accuracy of prediction methods

Method	Error	Target 1		Target 2	
		train	test	train	test
Linear regression	MSE	0.02	0.04	0.01	0.02
	MAE	0.13	0.17	0.9	0.11
	MAPE	14.20	16.67	14.28	16.41
Decision tree	MSE	0.01	0.02	0.0	0.01
	MAE	0.06	0.11	0.03	0.07
	MAPE	7.49	10.63	3.15	6.55
Random forest	MSE	0.02	0.04	0.01	0.03
	MAE	0.01	0.03	0.02	0.04
	MAPE	2.54	3.46	1.0	1.96

Based on the results of the assessment of the developed prediction models, it was found that among the three considered methods, the highest prediction accuracy for big datasets can be achieved using the random forest method.

A prediction software application has been developed that can be used for another dataset, which must first be loaded into the root directory. The prediction result is imported into a .xls file.

It should also be noted that Big Data tools and technologies were not used for such a big dataset. The technology used can be used with limited capabilities of the computer hardware. In order to implement this approach, it is necessary to use the Dask library.

Further exploration is loading the Dask library on even bigger datasets and investigating data processing speed and learning algorithms for big data. This will make it possible to establish whether this technology is capable of completely replacing Big Data technology for this kind of tasks.

4. Conclusions

The paper evaluates the efficiency of prediction numerical big datasets using linear regression methods, decision tree and random forest. The task was performed on the created and previously prepared set of anonymized data. To work with the dataset, the method of parallel computations was applied, which is implemented in the Dask library of the Python language. The prediction accuracy for the test dataset is about 83 % for linear regression, 89 % for decision trees and 97 % for a random forest. It was found that the random forest method among the considered methods is the most accurate when used on big datasets.

References

1. Rahmani, A. M., Azhir, E., Ali, S., Mohammadi, M., Ahmed, O. H., Yassin Ghafour, M. et. al. (2021). Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ Computer Science, 7,* e488. doi: http://doi.org/10.7717/peerj-cs.488
2. Labrinidis, A., Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5 (12),* 2032–2033. doi: http://doi.org/10.14778/2367502.2367572
3. Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University – Computer and Information Sciences, 30 (4),* 431–448. doi: http://doi.org/10.1016/j.jksuci.2017.06.001
4. Joseph, R. C., Johnson, N. A. (2013). Big Data and Transformational Government. *IT Professional, 15 (6),* 43–48. doi: http://doi.org/10.1109/mitp.2013.61
5. Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35 (2),* 137–144. doi: http://doi.org/10.1016/j.ijinfomgt.2014.10.007
6. Khine, K. L. L., Nyunt, T. T. S.; Zin, T., Lin, J. W. (Eds.) (2019) Predictive Big Data Analytics Using Multiple Linear Regression Model. *Big Data Analysis and Deep Learning Applications.* ICBDL, 9–19. doi: http://doi.org/10.1007/978-981-13-0869-7_2
7. Song, Y.-Y., Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27.* doi: http://doi.org/10.11919/j.issn.1002-0829.215044
8. Islam, S., Amin, S. H. (2020). Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. *Journal of Big Data, 7 (1).* doi: http://doi.org/10.1186/s40537-020-00345-2
9. Zrazhevskyi, O. H. (2010). Metody pobudovy modelei dlia dovhostrokovoho prohnozuvannia finansovykh chasovykh riadiv. *Systemni doslidzhennnia ta informatsiini tekhnolohii, 1,* 123–142.
10. Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *International Journal of Advanced Computer Science and Applications, 11 (2),* 612–619. doi: http://doi.org/10.14569/ijacsa.2020.0110277
11. Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5–32. doi: http://doi.org/10.1023/a:1010933404324

Eduard Kinshakov, *Postgraduate Student, Department of Information Technology, Sumy State University, Sumy, Ukraine, ORCID: https://orcid.org/0000-0001-7116-7244*

✉ ***Yuliia Parfenenko,*** *PhD, Associate Professor, Department of Information Technology, Sumy State University, Sumy, Ukraine, e-mail: yuliya_p@cs.sumdu.edu.ua, ORCID: http://orcid.org/0000-0003-4377-5132*

Vira Shendryk, *PhD, Associate Professor, Department of Information Technology, Sumy State University, Sumy, Ukraine, ORCID: https://orcid.org/0000-0001-8325-3115*

✉ *Corresponding author*