UDC 004.4, 004.6
DOI: 10.15587/2706-5448.2022.262562
Article type «Reports on Research Projects»

Olha Bulhakova,
Yuliia Ulianovska,
Victoria Kostenko,
Tatyana Rudyanova

CONSIDERATION OF THE POSSIBILITIES OF APPLYING MACHINE LEARNING METHODS FOR DATA ANALYSIS WHEN PROMOTING SERVICES TO BANK'S CLIENTS

The object of the research is modern online services and machine learning libraries for predicting the probability of the bank client's consent to the provision of the proposed services. One of the most problematic areas is the high unpredictability of the result in the field of banking marketing using the most common technique of introducing new services for clients – the so-called cold calling. Therefore, the question of assessing the probability and predicting the behavior of a potential client when promoting new banking services and services using cold calling is particularly relevant.

In the course of the study, libraries of machine learning methods and data analysis of the Python programming language were used. A program was developed to build a model for predicting the behavior of bank customers using data processing methods using gradient boosting, regularization of gradient boosting, random forest algorithm and recurrent neural networks. Analogous models were built using cloud machine learning services Azure ML, BigML and the Auto-sklearn library.

Data analysis and prediction models built using Python language libraries have a fairly high quality – an average of 94.5 %. Using the Azure ML cloud service, a predictive model with an accuracy of 88.6 % was built. The BigML machine learning service made it possible to build a model with an accuracy of 88.8 %. Machine learning methods from the Auto-sklearn library made it possible to obtain a model with a higher quality – 94.9 %. This is due to the fact that the proposed libraries of the Python programming language allow better customization of data processing methods and machine learning to obtain more accurate models than free cloud services that do not provide such capabilities.

Thanks to this, it is possible to obtain a predictive model of the behavior of bank customers with a fairly high degree of accuracy. It is worth noting that in order to make a prediction (forecast), it is necessary to study the context of the task, process the data, build various machine learning algorithms, evaluate the quality of the models and choose the best of them.

Keywords: *artificial intelligence, machine learning methods, banking services, credit scoring, credit risk.*

Received date: 19.05.2022
Accepted date: 02.08.2022
Published date: 11.08.2022

How to cite

Bulhakova, O., Ulianovska, Y., Kostenko, V., Rudyanova, T. (2022). Consideration of the possibilities of applying machine learning methods for data analysis when promoting services to bank's clients. Technology Audit and Production Reserves, 4 (2 (66)), 14–18. doi: http://doi.org/10.15587/2706-5448.2022.262562

1. Introduction

Despite the constant and intensive development of banking marketing, the most common technique for introducing new services to customers still remains the so-called cold calling. Its essence is to make phone calls with offers of various products to potential bank customers. The main goal of cold calling is extensive, that is, quantitative, expansion and increase in the client base. Therefore, in the field of banking marketing, the issue of assessing the probability is especially relevant due to the high unpredictability of the result of cold calling.

Currently, many banks benefit for themselves and their customers by processing and analyzing the data they store, while applying new methods of data analysis, including machine learning methods. However, it is rather difficult to compare the task with other similar banking tasks, since the information necessary to solve it is often kept secret.

But, despite corporate secrets, it can be said that there are projects related to the prediction of customer behavior, for example, in credit scoring (a system for assessing the creditworthiness (credit risks) of a person based on numerical statistical methods) [1, 2]. There is also the task of classifying clients by the possibility of default, by fraud, by the interest of various banking products. Most of these problems are still far from being solved, but machine learning offers many methods, some of which are discussed in this paper.

Real published cases on the application of new methods are, for example, in PrivatBank and Oschadbank [3, 4]. Among them: identification of patterns based on card transactions, optimization of the sales funnel (funnel), modeling the probability of default for small businesses and others. However, machine learning does not play a key role here, but only an advisory one. Therefore, the next step for big business, in particular, for the banking system, is the ability to predict certain events. Since a bank usually has a huge amount of information about its customers that it can work with, the task has really great potential in practice. Recent studies [5, 6] highlight that banks are looking for more ways to leverage product trends, market dynamics, customer behavior, and new technologies through rich data analytics.

Thus, *the object of research* is modern online services and machine learning libraries for predicting the probability of a bank client's consent to the provision of the offered services.

The aim of research is to analyze the possibilities of using online machine learning services and libraries for analysis, modeling and forecasting of the Python programming language in order to predict the client's response to the provision of banking services based on human data.

2. Research methodology

For this research, a set of real data was used, collected by the National Bank of Portugal during a campaign to attract new customers to its own banking products in the amount of 45 thousand records [7].

In the study sample, the following factors are available for analysis:

- age – client's age;
- job – employment type;
- martial – marital status;
- education – level of education;
- housing – whether the client has a mortgage;
- loan – availability of a credit card;
- campaign – the number of contacts with the client during the advertising campaign;
- poutcome – the result of a past marketing campaign with this client;
- other less important factors.

The data used is raw, unprocessed data recorded by the bank's marketing department, which makes it impossible to use it without pre-processing, which would lead it to a form suitable for machine learning models. Therefore, as part of the mandatory data processing, an analysis of all data and parameters was carried out, taking into account the importance of existing factors and the type of their values. Also, to improve the quality of the analysis, the data was cleaned from outliers and the missing features were filled in with average values for the feature.

The following approaches were used to encode features:

- grouping of features;
- binarization of categorical features using the one-hot coding method;
- analysis of correlated features (removal at $corr > 0.9$);
- noise removal;
- class balancing;
- pass processing.

In the course of the study, one of the software tools in which the original problem was solved was IPython, an interactive shell for the Python programming language.

In the course of solving the problem, the following data processing methods were used [8]:

1. Classification:

- Gradient Boosting is a machine learning technique for classification and regression problems that builds a prediction model in the form of an ensemble of weak models, usually decision trees.
- XGBoost is a library used in machine learning and provides functionality for solving problems related to regularization of gradient boosting.
- Random Forest is a random forest algorithm (a universal machine learning algorithm, the essence of which is to use an ensemble of decision trees. The decision tree itself gives an extremely low classification quality, but due to the large number, the result improves significantly).
- Recurrent Neural Network is the recurrent neural networks (a type of neural networks where connections between elements form a directed sequence).

2. Clustering:

- K-means – *k*-means algorithm (machine learning algorithm that solves the clustering problem. It splits the set of elements of the vector space into a predetermined number of clusters *k*. The action of the algorithm is such that it seeks to minimize the standard deviation at the points of each cluster. This algorithm has become very popular due to its simplicity, clarity of implementation and sufficiently high quality of work).

When solving the problem of marketing foresight, the following libraries were used:

- Pandas (data processing and analysis);
- Numpy (work with multidimensional arrays);
- Math (work with numbers);
- Sklearn (machine learning algorithms);
- XGboost (adaptive boosting algorithms);
- Seaborn, ggplot, matplotlib (charts);
- Pybrain (neural networks).

3. Research results and discussion

In the course of the research, data was processed (binarization, one-hot-encoding, standardization), correlation matrices were built, and the sample was divided into training and test. For each model, the scales were adjusted and the required parameter values were set, after which the model was trained. The result of the models was evaluated by the ROC-curve and AUC (Area Under Cover). The results of measurements are shown in Table 1.

Table 1

Model evaluation indicators according to the ROC curve and AUC (Area Under Cover)

Data processing methods / Model evaluation results	XG Boost	Gradient Boosting	Random Forest	Neural Networks
AUC-ROC	0.948	0.945	0.941	0.955

As it is possible to see, neural networks have the best result.

As part of the study, a number of models were created using online services that provide data analysis services as a service (MLaaS). Among the many available solutions, the following services were chosen: BigML and Azure ML due to their availability and free.

Azure ML is a cloud service built on top of Microsoft technologies and services. Data analysis is built by creating a simple scheme in the visual editor of the service [9]. As with manual analysis, the chain of actions consists of loading the data into the tool, processing it into a form suitable for the model, and finally training the model. Thus, the scheme is shown in Fig. 1.

After running the developed scheme based on the two-class SVM model, the following results were obtained (Fig. 2, 3).

BigML is a cloud-based machine learning service [10] developed by BigML, Inc. of the same name. Unlike the service from Microsoft, this service does not have any parameters to customize the learning process of the prediction model. The user only needs to upload the file, after which the service itself decides all data analysis issues. The model learned by the service coincided in quality with that in Azure ML, which indicates the similarity of the approaches used in these services (Fig. 4, 5).

Auto-sklearn is a library whose methods consist in unsupervised model learning [11]. The name «auto» itself indicates that the work in it is automatic: the model independently selects the best parameters. As the main hyperparameters, they usually set the time limit for searching for the optimal model (time_left_for_this_task) and the time limit for running individual machine learning models (per_run_time_limit).

The resulting model gave the following results (Fig. 6):

– $MSE=0.0827$;
– $MAE=0.0859$;
– $ROC\ AUC\ score=0.9492$.

Auto-sklearn allowed to obtain a model with higher quality compared to the gradient boosting (Gradient Boosting XGBoost) and random forests (Random Forest) methods. However, it still turned out to be worse than the model obtained by using the neural network (Recurrent Neural Network).

Machine learning and data analysis methods can help banks predict customer behavior with a fairly high degree of accuracy.

It should be noted that in order to make a prediction (forecast), it is necessary to study the context of the problem, process the data, build various machine learning algorithms, evaluate the quality of the models and choose the best one.

The use of telephone marketing allows to achieve significant results in the sale of banking services. However, in order to effectively apply this method in practice, it is necessary to have a complete understanding of the interlocutor in order to interest the bank's services. On the basis of personal information already available in the customer database, in particular age, marital status, education level, information about the services provided, it is possible to formulate ideas about the interlocutor and the bank's products that may be of interest to it.

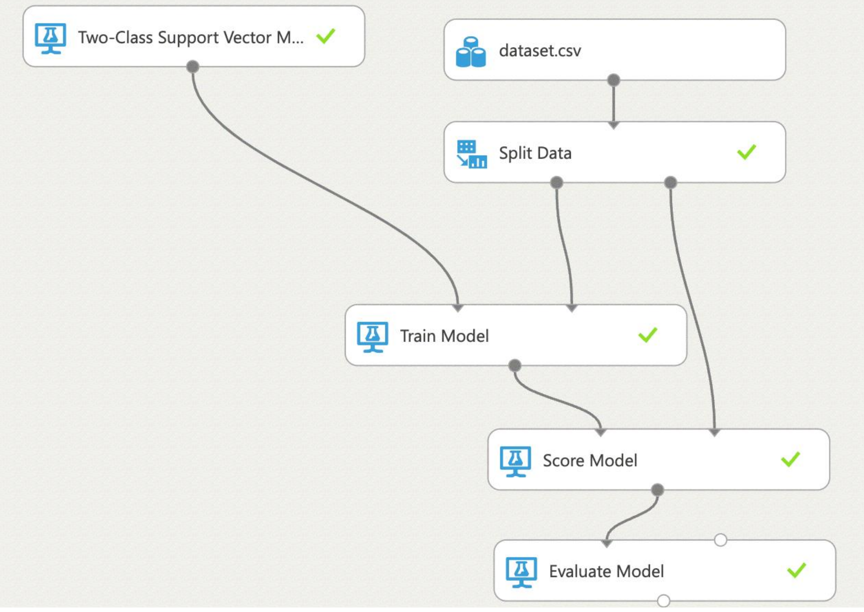


Fig. 1. Data analysis schema in Azure ML cloud service

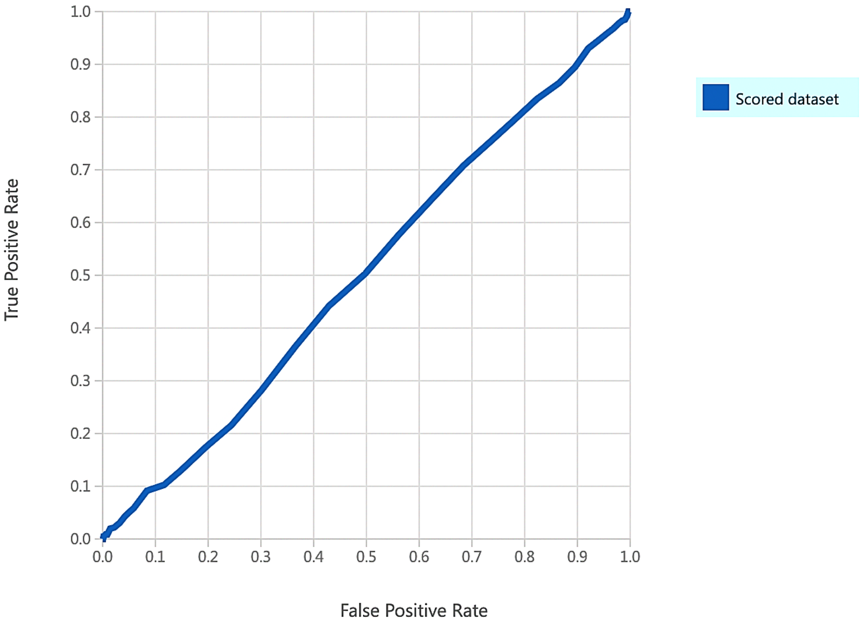
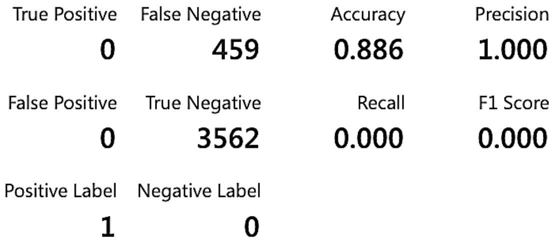
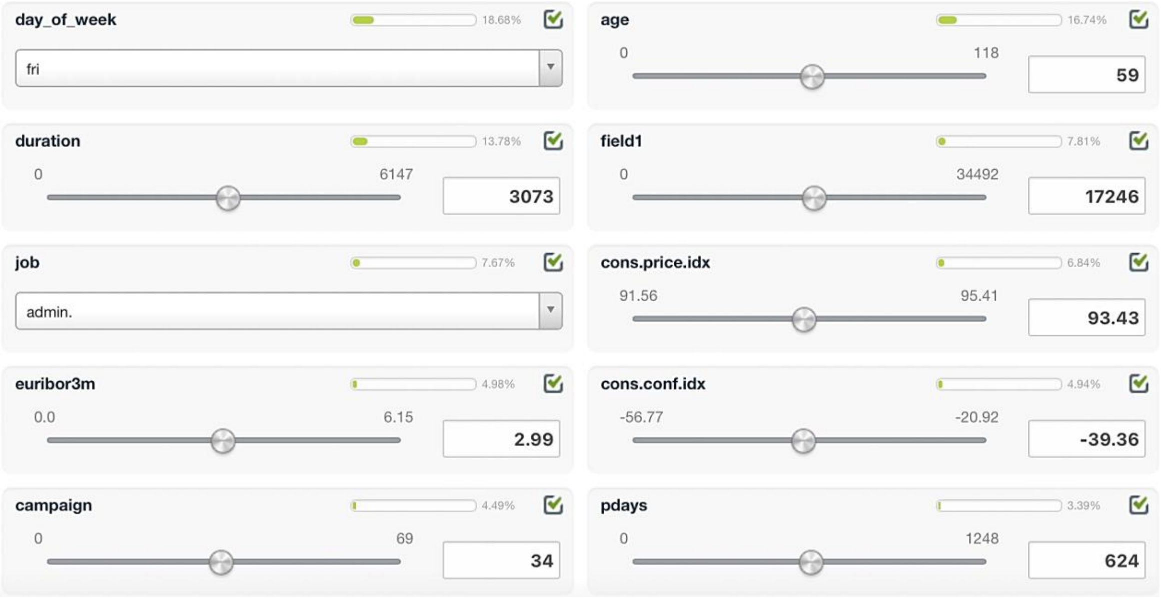


Fig. 2. TPR-FPR graph



True Positive	False Negative	Accuracy	Precision
0	459	0.886	1.000
False Positive	**True Negative**	**Recall**	**F1 Score**
0	3562	0.000	0.000
Positive Label	**Negative Label**		
1	0		

Fig. 3. The main metrics of the obtained SVM model



day_of_week 18.68% ☑
fri

age 16.74% ☑
0 118
59

duration 13.78% ☑
0 6147
3073

field1 7.81% ☑
0 34492
17246

job 7.67% ☑
admin.

cons.price.idx 6.84% ☑
91.56 95.41
93.43

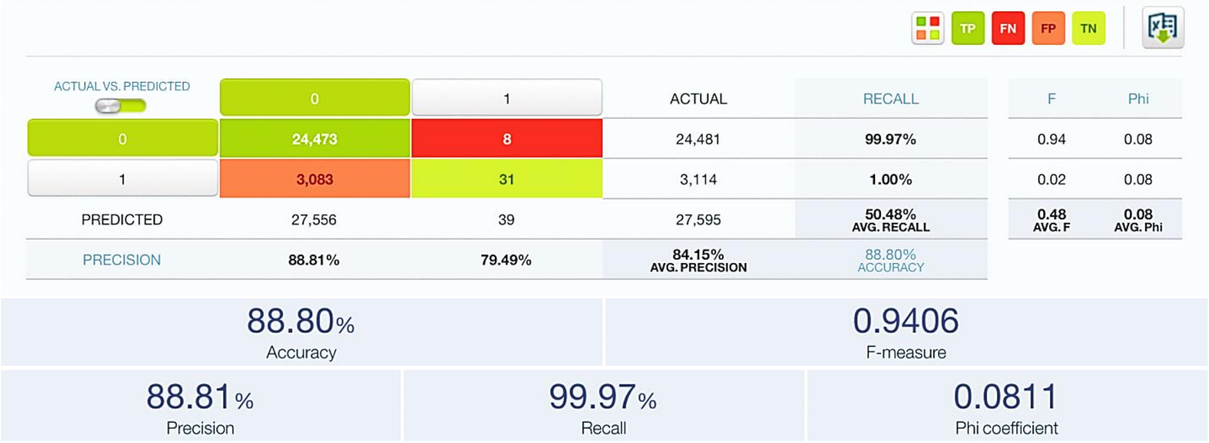
euribor3m 4.98% ☑
0.0 6.15
2.99

cons.conf.idx 4.94% ☑
-56.77 -20.92
-39.36

campaign 4.49% ☑
0 69
34

pdays 3.39% ☑
0 1248
624

Fig. 4. Description of the importance of parameters according to BigML



ACTUAL VS. PREDICTED	0	1	ACTUAL	RECALL	F	Phi
0	24,473	8	24,481	99.97%	0.94	0.08
1	3,083	31	3,114	1.00%	0.02	0.08
PREDICTED	27,556	39	27,595	50.48% AVG. RECALL	0.48 AVG. F	0.08 AVG. Phi
PRECISION	88.81%	79.49%	84.15% AVG. PRECISION	88.80% ACCURACY		

88.80% Accuracy

0.9406 F-measure

88.81% Precision

99.97% Recall

0.0811 Phi coefficient

Fig. 5. Evaluation of the quality of the constructed model using BigML

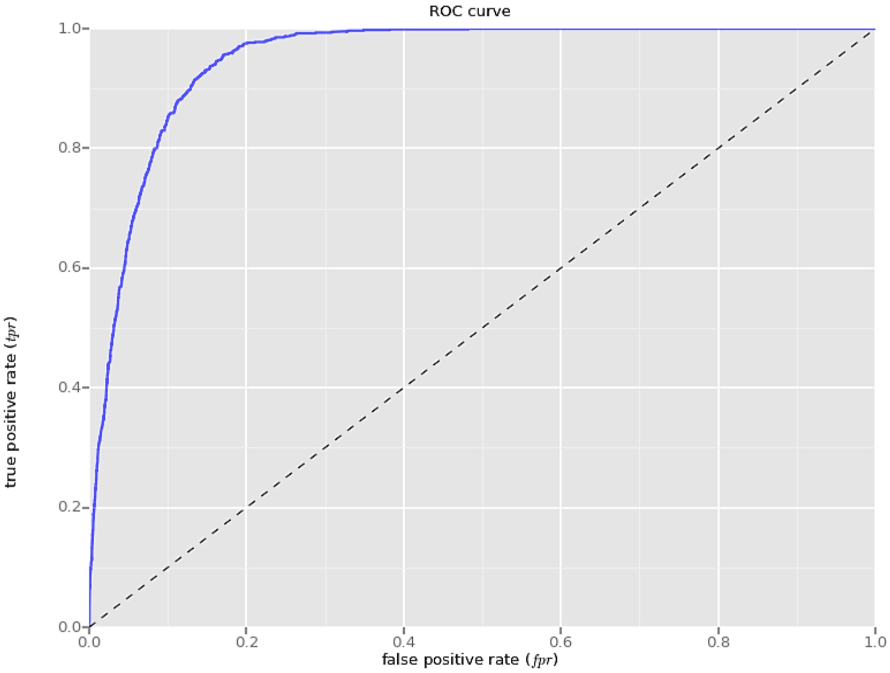


Fig. 6. Evaluation of the property of the constructed model by the Auto-sklearn library

The introduction of artificial intelligence technologies, machine learning methods and data analysis in the banking sector will effectively explore the customer database, facilitate the formation of recommendations, forecasts and personalized and relevant specialized financial offers for each client.

Analysis of the extensive data stored in the bank helps to better understand the needs and expectations of each client. Therefore, further research will be related to the analysis and selection of the best algorithms for clustering internal and external data points, which will help create a holistic picture of the desires of each client and form the right target audience for banking services. Also an interesting area for research is the analysis of data on the current device of the visitor, its location, history of visits and history of interaction with the bank in order to assess the personalized experience of the client, which, of course, will increase the level of satisfaction and expand the volume of services provided. The introduction of such technological solutions will not only help to understand the needs and expectations of each client, but will also help banks maintain the efficient functioning of their offices.

4. Conclusions

In the course of the study, in the interactive shell environment for the Python programming language, a program was developed to build a model of the behavior of bank customers using data processing methods using gradient boosting, gradient boosting regularization, random forest algorithm and recurrent neural networks. The manual analysis of the data made it possible to build models with high quality – 94.8 %, 94.5 %, 94.1 %, 95.5 %. As part of the study, based on the same initial data, a number of models were created using online services. Using the Azure ML cloud service, a predictive model with an accuracy of 88.6 % was built on the basis of a two-class SVM model. The BigML machine learning service made it possible to build a model with an accuracy of 88.8 %. Machine learning methods from the Auto-sklearn library made it possible to obtain a model with a higher quality – 94.9 %.

The study found that machine learning and data analysis methods can help banks predict the behavior of their customers with a fairly high degree of accuracy. It should be noted that in order to make a prediction (forecast), it is necessary to study the context of the problem, process the data, build various machine learning algorithms, evaluate the quality of the models and choose the best one. At the same time, the libraries of the Python programming language allow to better customize data processing and machine learning methods to obtain more accurate models than free cloud services that do not provide such capabilities.

The use of telephone marketing allows to achieve significant results in the sale of banking services. However, in order to effectively apply this method in practice, it is necessary to have a complete understanding of the interlocutor in order to interest the bank's services. So, the study shows that based on personal information already available in the customer database, in particular age, marital status, education level, information about the services provided, it is possible to build a sufficiently high-quality model using machine learning methods. And also to formulate an idea about the interlocutor and the products of the bank that may be of interest to it.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

References

1. Buchko, I. Ye. (2013). Skorynh yak metod znyzhennia kredytnoho ryzyku banku. *Visnyk Universytetu bankivskoi spravy Natsionalnoho banku Ukrainy, 2,* 178–182.
2. *Yak pratsiuie bankivskyi skorynh* (2021). Available at: https://finance.ua/ua/credits/kak-rabotaet-bankovskiy-skoring Last accessed: 20.04.2022
3. *NBU. Zvit pro finansovu stabilnist. Cherven 2019 roku.* Available at: https://bank.gov.ua/ua/news/all/zvit-pro-finansovu-stabilnist-cherven-2019-roku
4. Dunas, N., Bilokrynytska, M. (2019). Implementation of credit scoring system by ukrainian banks for consumer credit. *Pryazovskyi Economic Herald, 5 (16)*, 263–269. doi: http://doi.org/10.32840/2522-4263/2019-5-45
5. Maja, M. M., Letaba, P. (2022). Towards a data-driven technology roadmap for the bank of the future: Exploring big data analytics to support technology roadmapping. *Social Sciences & Humanities Open, 6 (1),* 100270. doi: http://doi.org/10.1016/j.ssaho.2022.100270
6. Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science, 2 (5)*. doi: http://doi.org/10.1007/s42979-021-00765-8
7. Gavrysh, V. (2019). *Bank marketing campaigns dataset*. Opening Deposit. Available at: https://www.kaggle.com/datasets/volodymyrgavrysh/bank-marketing-campaigns-dataset
8. Burns, S. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python*. Scikit-learn and Tensorflow, 176.
9. *Azure Machine Learning*. Available at: https://azure.microsoft.com/en-us/services/machine-learning/#product-overview
10. *BigML Tools*. Available at: https://bigml.com/tools/
11. *Auto-sklearn*. Available at: https://automl.github.io/auto-sklearn/master/#

✉***Olha Bulhakova,*** *Senior Lecturer, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, e-mail: olga.bulgakova.dp@gmail.com, ORCID: https://orcid.org/0000-0001-9834-2970*

Yuliia Ulianovska, *PhD, Associate Professor, Head of Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0001-5945-5251*

Victoria Kostenko, *Senior Lecturer, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: https://orcid.org/0000-0003-3847-2110*

Tatyana Rudyanova, *PhD, Associate Professor, Department of Computer Science and Software Engineering, University of Customs and Finance, Dnipro, Ukraine, ORCID: http://orcid.org/0000-0002-8685-4132*

✉*Corresponding author*