(Vol. 2, No. 3, p. 4). – Cambridge, MA: MIT press.

5. Wald, A. (1992). Sequential tests of statistical hypotheses. In Breakthroughs in statisti cs: Foundations and basic theory. – New York, NY: Springer New York. – P. 256–298.

УДК 004.021

COMPARATIVE STUDY OF BOSTON HOUSE PRICE PREDICTION MODELS BASED ON GAN DATA AUGMENTATION

He Runhai
Belarusian State University
e-mail: fpm.he@bsu.by

Summary. *This study aims to compare the performance of four machine learning models in the Boston housing price prediction task, and explore whether GAN data augmentation can improve the prediction performance of the model. Experimental results indicate that on the original data set, The CatBoost model demonstrated superior predictive performance, with an MSE of 4.76, RMSE of 2.18, MAE of 1.65, and R^2 of 0.94. However, no significant performance improvement was observed when GAN data augmentation was applied to these models.*

The real estate market is an important part of the national economy, and house price prediction has always been a hot topic of concern in academia and industry. The Boston house price dataset is a recognized classic benchmark dataset for house price prediction. The dataset comes from a real estate study in the 1970s [1]. In recent years, with the rapid development of machine learning technology, various advanced models have been widely used in house price prediction tasks and have achieved good results.

This study aims to compare the performance of four mainstream machine learning models (random forest, XGBoost, gradient boosting, CatBoost) on the Boston house price prediction task, and explore whether the introduction of GAN data augmentation technology on this basis can further improve the prediction accuracy of the model [2, 3, 4, 5]. GAN-based data augmentation generates synthetic samples that mimic the statistical characteristics of the original dataset, aiming to enhance model generalization capabilities [6].

The Boston house price dataset contains 506 samples, 13 features, and the target variable is the house price. During the model training phase, a grid search optimization is performed on the hyperparameters of four basic models and then, compared them with the version with GAN data augmentation.

The Figure 1 below shows the comparison of the predicted values and actual values of each basic model on the test set. As can be seen from the figure, the predicted curve of the CatBoost model is closest to the actual value curve and has the best prediction performance.

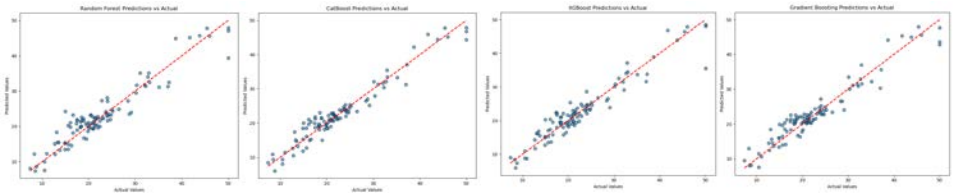

Figure 1 – Comparison of actual and predicted values across models

GAN data augmentation was then applied before training each model to evaluate its impact on prediction performance.

Table 1 – Prediction performance metrics comparison of all models

Model	**MSE**	**RMSE**	**MAE**	**R^2**
Random Forest	5.80	2.41	1.73	0.93
Random Forest + GAN	8.39	2.90	2.19	0.90
CatBoost	4.76	2.18	1.65	0.94
CatBoost + GAN	4.76	2.18	1.68	0.94
XGBoost	5.81	2.41	1.71	0.93
XGBoost + GAN	7.07	2.66	1.90	0.92
Gradient Boosting	5.02	2.24	1.77	0.94
Gradient Boosting + GAN	6.41	2.53	1.96	0.92

Optimal hyperparameters for the best performing model 'CatBoost': Depth: 4, Iterations:300, L2 Leaf Regularization:1, Learning Rate: 0.2

Table 1 summarizes the prediction performance indicators of all models. It can be seen that the CatBoost model has an MSE of 4.76, an RMSE of 2.18, a MAE of 1.65, and an R^2 of 0.94, which is the best among the four basic models. After GAN data augmentation, the performance indicators of the other four models also failed to surpass the original version of CatBoost.

In conclusion, this study assesses the performance of four machine learning models for predicting Boston housing prices, highlighting CatBoost's superior fit and generalization over random forest, XGBoost, and gradient boosting. Although GAN-based data augmentation was applied, it did not significantly enhance model performance.

Future research may consider more scenarios driving machine learning models and alternative data augmentation techniques to further improve prediction accuracy [7].

References

1. Harrison Jr, D., & Rubinfeld, D. L.(1978). Hedonic housing prices and the demand for clean air. – Journal of environmental economics and management. – 5(1). – P. 81–102.

2. Rigatti, S. J.(2017). Random forest. – Journal of Insurance Medicine. – 47(1). – P. 31–39.

3. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. – P. 785–794.

4. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

5. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Cat Boost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

6. Tanaka, F. H. K. D. S., & Aranha, C. (2019). Data augmentation using GANs. arXiv preprint arXiv:1904.09135.

7. Xia, S., Wu, Z., & Song, Q. (2024). Scenarios Driving Economic Forecasts: Choosing Econometrics or Machine Learning. Journal of Economic Theory and Business Management, 1(2). – P. 7–26.

УДК 377.1:004.9

ADAPTATION OF THE UNIVERSITY PROGRAM IN ALGORITHMS AND PROGRAMMING IN PYTHON FOR ADDITIONAL PRE-UNIVERSITY EDUCATION

Kastsiuk D., Markina A., Shulgan A.
Brest State Technical University, ООО «АйТи Скул»
e-mail: assyamarkina2@gmail.com

Summary. *The experience of modifying the basic course of algorithms and programming for schoolchildren is presented, based on the original course designed for the first level of higher education of the computer science and radio electronics profile. The presentation peculiarities for the basic constructions, data abstractions, and basic algorithms in Pyton language used for levelling of low indices of self-regulation and concentration are considered, as well as splitting the topics into sets of small time intervals for adapting to the ability of keeping the attention focus. The specifics of the used development tools distributed under free/libre licences within the framework of the course are considered.*

In recent years, there is a tendency to lower the age of entry into programming. In contrast to the former approach with teaching 12-year-old children block programming in Scratch, the age of learners starts now from 10 years and less, which imposes new requirements to teaching materials for teenagers. The experience of adapting the curriculum "Algorithmisation and Programming in Python" of Brest State Technical University for teenagers is discussed below as a final stage of learning a bundle of three programming languages: Scratch, Lua, and Python.

When adapting the course, we had to take into account that children of 13–16 years old have significantly lower speed of writing, reading and the volume of perceived learning material than students, not taking into account the speed of typing. The cognitive and emotional characteristics of the learners, associated with the fact that they have not yet fully formed prefrontal cortex responsible for planning, decision-making and social behaviour, require special pedagogical approaches [1]. In addition to adaptation for children, the teaching of this course required adaptation for their parents who want their children to receive quality knowledge, and at the same time, the majority of them either have a connection