SHAPLEY VALUES TO EXPLAIN MACHINE LEARNING MODELS OF SCHOOL STUDENT'S ACADEMIC PERFORMANCE DURING COVID-19

Yunusov Valentin
Kazan Federal University, Kazan, (Russian Federation).
E-mail: valentin.yunusov@gmail.com

Gafarov Fail
Kazan Federal University, Kazan, (Russian Federation).

Ustin Pavel
Kazan Federal University, Kazan, (Russian Federation).

Reception: 26/10/2022 **Acceptance**: 10/11/2022 **Publication**: 29/12/2022

Suggested citation:

Valentin, Y., Fail, G., y Pavel, U. (2022). Shapley values to explain machine learning models of school student's academic performance during COVID-19. *3C TIC. Cuadernos de desarrollo aplicados a las TIC, 11*(2), 136-144. https://doi.org/10.17993/3ctic.2022.112.136-144

ABSTRACT

In this work we perform an analysis of distance learning format influence, caused by COVID-19 pandemic on school students' academic performance. This study is based on a large dataset consisting of school students grades for 2020 academic year taken from "Electronic education in Tatarstan Republic" system. The analysis is based on the use of machine learning methods and feature importance technique realized by using Python programming language. One of the priorities of this work is to identify the academic factors causing the most sensitive impact on school students' performance. In this work we used the Shapley values method for solving this task. This method is widely used for the feature importance estimation task and can evaluate impact of every studied feature on the output of machine learning models. The study-related conditional factors include characteristics of teachers, types and kinds of educational organization, area of their location and subjects for which marks were obtained.

KEYWORDS

Data Science, Python, education, Machine Learning, Feature Importance.

https://doi.org/10.17993/3ctic.2022.112.136-144

1. INTRODUCTION

Failure to achieve educational goals negatively affects society as a whole and is a serious problem. This problem can manifest itself most significantly during periods of drastic changes, one of which was the introduction of distance learning during the COVID-19 pandemic. To quantify the influence of this event on educational system, a variety of quantitative models based on modern statistical methods in combination with Big Data approaches can be used, as has shown in Li et al. [2021].

Machine learning (ML) is one of the new and actively developing methods of analysis, combining approaches that can "learn" based on the received data, which allows to perform a wide range of different tasks. ML can be used to solve problems of detection, recognition, prediction, prediction, diagnostics, and optimization.

A large number of huge datasets has been accumulated recently in educational system, which can be used to analyze and then improve educational process, as was demonstrated by Park [2020]. For example, Livieris et al. [2019] analyze a dataset consisting of performance of 3716 students in course of Mathematics of the first 5 years of secondary school. They develop two semisupervised machine learning algorithms to predict students' performance in the final examinations and then evaluate methods' accuracy. Authors compare these two methods with supervised machine learning method and as a result, these approaches outperform it, and the final accuracy exceeds 80%.

Jeslet et al. [2021] used well-known algorithms of machine learning Logistic Regression and Support Vector Machine to predict whether student is eligible to acquire a degree or not. Authors analyzed dataset of 1460 students' final year's results and obtained a model trained to 99.27% and 99.72% accuracy. Also, Nuanmeeseri et al. [2022] analyzed dataset of 1650 university students' academic performance. As a result, after adjusting model's parameters, authors achieved accuracy of 96.98%, so their model outperformed other considered machine learning methods and can be effectively used to evaluate significant academic performance factors in drastically changing period.

In our work, we study changes of academic performance of whole school grades in the framework of a variety of machine learning methods with the following feature importance analysis to identify significant parameters that affect academic performance the most after the introduction of distance learning format due to the COVID-19 pandemic.

2. MACHINE LEARNING METHODS AND FEATURE IMPORTANCE

2.1 MACHINE LEARNING TECHNIQUES

Hastie et al. [2009] introduce Machine learning as a set of mathematical techniques that give computer algorithms an ability to learn. This methodology is based on the input and required output of the algorithms and can automate the way how humans are able to carry out the task, as stated by Mnih et al. [2015].

Ensemble methods are groups of algorithms that use several machine learning methods at once and makes correction of each other's errors. Bostanabad et al. [2016] define supervised learning as a type of algorithms where the method is supplied with example inputs along with the required output, which then allows it to learn a rule that maps inputs to outputs. Bengio et al. [2013] state that in unsupervised learning, on the contrary, only the inputs are supplied, and the learning algorithm is required to determine the structure of the input and perform according to unknown characteristics [10].

In this work we use supervised machine learning methods: Decision Tree, Gradient Boosting, K-nearest neighbors (KNN) Regressor, Lasso Regression, Linear Regression and MultiLayer Perceptron neural networks, Support Vector Regressor; and ensemble method: Random Forest.

https://doi.org/10.17993/3ctic.2022.112.136-144

In our study, we solved the regression task to predict Cohen's effect size, defined by Cohen [1988], based on subsets of school grades' marks in February and March, and April and May. Cohen's effect size measures the difference between mean values of two variables Cohen [1988].

2.2 SHAP FEATURE IMPORTANCE IMPLEMENTATION

Usually, machine learning models are difficult to interpret and it's hard to identify which features affect the output of the models the most. SHAP method (Shapley additive explanations) is one of the techniques used to solve this problem. This method is based on cooperative game theory, explained by Shapley [1953], and is used to increase transparency and interpretability of machine learning models. Absolute SHAP value shows us how much a single feature affected the prediction. SHAP values can represent the local importance of features and how it changes with lower and higher values, as shown by Sahakyan et al. [2021].

3. EXPERIMENTAL DATA DESCRIPTION

In this work, we study the influence of COVID-19 pandemic on school students' academic performance by analyzing a large dataset consisting of data from all schools in Tatarstan Republic, introduced by Ustin et al. [2022]. The dataset includes marks of entire grades of school students for main subjects for grades from 2 to 11.

During the preprocessing of original data, for the following analysis by machine learning methods, the initial dataset was modified into a new dataset consisting of features describing different parameters. These parameters included teachers' characteristics (age, sex, and educational category), mean mark of grade for February and March of 2020, school characteristics (location in or out of town, region of location, organization kind and type, subject). Data was filtered to consider school grades with at least 60 school grades in certain time periods (February and March, April and May 2020). For every row in dataset, Cohen's effect size was calculated. Figure 1 shows histograms for certain grades that represent whole dataset. It should be noted that most parameter values are positive, i.e., after the introduction of distance learning format, grades have generally increased.

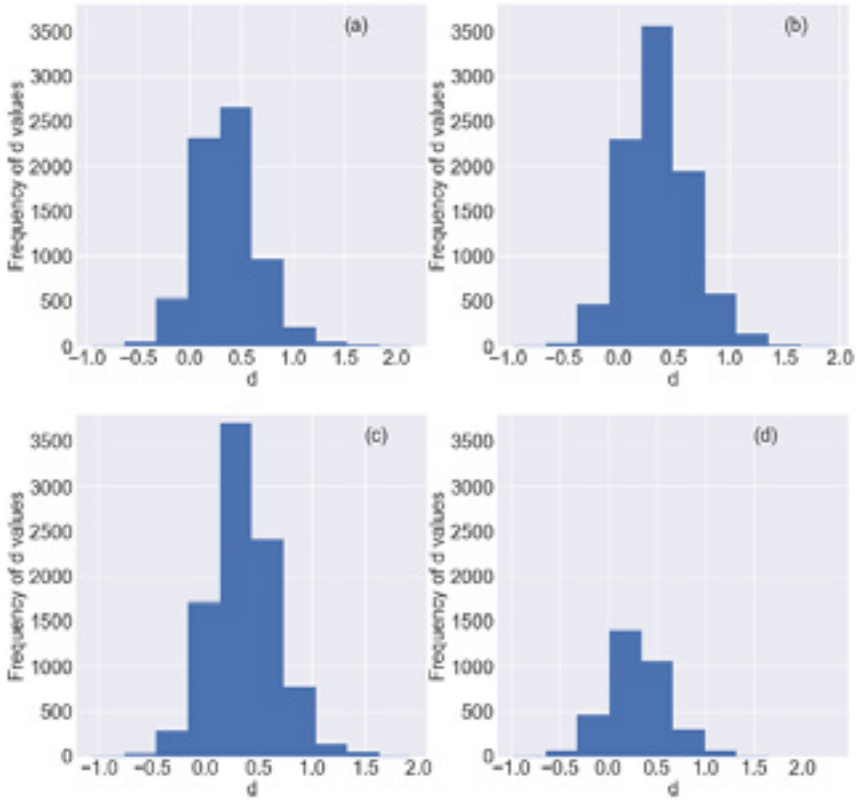


Fig. 1. Histograms of parameter d for: (a) 5th grade; (b) 7th grade; (c) 8th grade; (d) 11th grade.

https://doi.org/10.17993/3ctic.2022.112.136-144

4. APPLICATION OF MACHINE LEARNING METHODS IN THE ANALYSIS OF THE SCHOOL STUDENTS' ACADEMIC PERFORMANCE

The analysis was performed in two stages. At the first stage, we implemented a variety of machine learning methods for a comparative analysis of machine learning methods in the regression problem of predicting the values of the Cohen's effect size based on a large set of features. At the second stage, we performed evaluation of the importance of explanatory variables in the predictive model.

4.1 MACHINE LEARNING METHODS IMPLEMENTATION

In our work we applied machine learning techniques realized in PyTorch and scikit-learn frameworks in Python. Among the applied methods: one-layer Linear regression and MultiLayer Perceptron realized in Pytorch; Decision Tree, Gradient Boosting, K-nearest neighbors algorithm, Lasso regression, Random Forest and Support Vector Regression realized in scikit-learn framework.

MultiLayer Perceptron consisted of the input layer, two hidden layers with 64 neurons and output layer with 1 neuron. We used ReLU as activation function, Adam as optimizer with learning rate equal to 0.00005 and Mean Squared Error (MSE) as loss function. Figure 2 shows the learning curve of one-layer Linear regression and MultiLayer Perceptron.

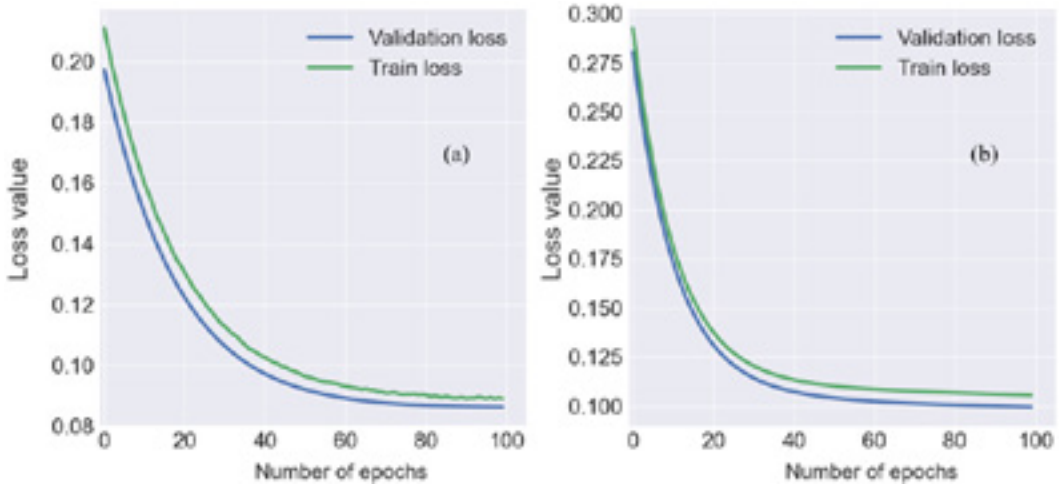


Fig. 2. Learning curve of: (a) MultiLayer Perceptron for 6th grade; (b) one-layer Linear regression for 6th grade.

Figure 3 shows the resulting plot of minimal MSE values for each method of machine learning for every studied school grade. It should be noted that the most precise algorithms are Random Forest, Lasso Regression, K-nearest neighbors and Support Vector Regression. Decision Tree and Gradient Boosting, on the other hand, have high values of error function. Also we obtained that for 8th and 10th grade values of MSE are increased significantly of the methods, and hence the Cohen's effect size values are more difficult to predict, while for 4th and 9th these values are decreased. Therefore, marks of students in 8th and 10th grade after the introduction of distant learning format due to the COVID-19 pandemic changed more randomly than the marks of students in other grades, especially in 4th and 9th grades.

https://doi.org/10.17993/3ctic.2022.112.136-144

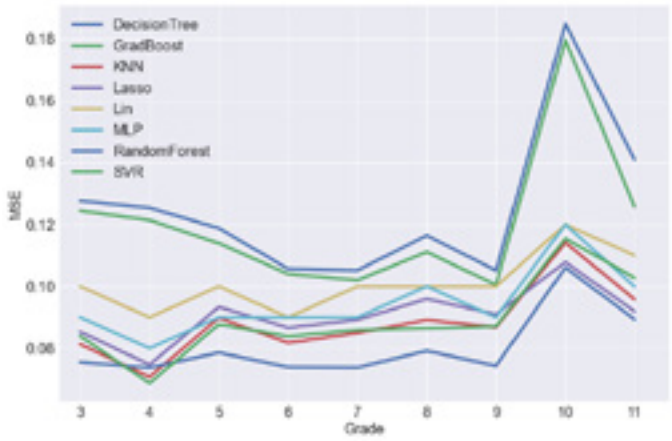
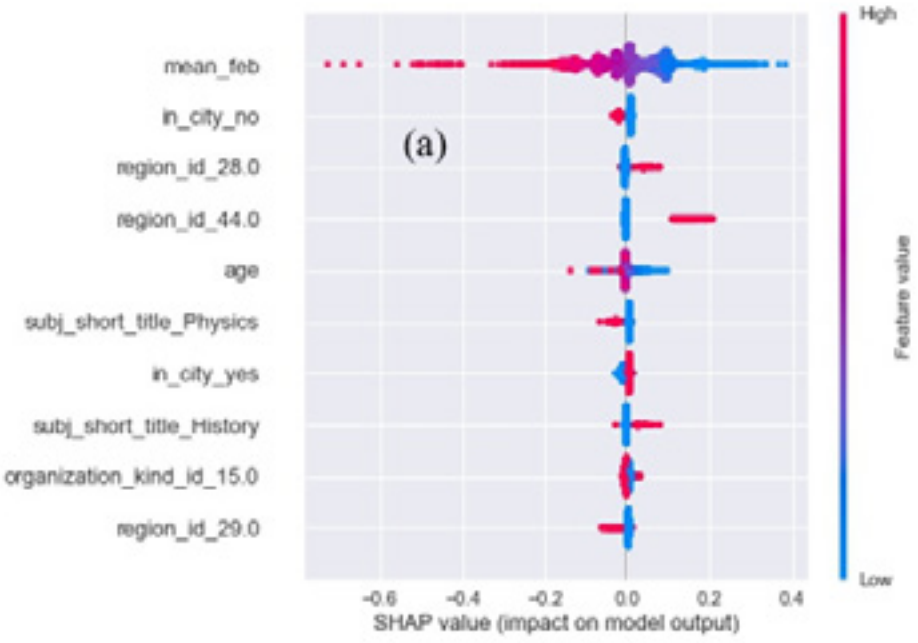


Fig. 3. The values of MSE for machine learning algorithms for each studied school grade.

4.2 EVALUATION OF THE IMPORTANCE OF EXPLANATORY VARIABLES

At the second stage of our analysis, we evaluated importance of our explanatory features for predicting values of Cohen's effect size. Figure 4 shows the distribution of Shapley values, i.e., influence on the value of parameter exerted by the studied explanatory features. Analysis was performed for Gradient Boosting, Random Forest and MultiLayer Perceptron models with primary Explainer and Tree Explainer.



https://doi.org/10.17993/3ctic.2022.112.136-144

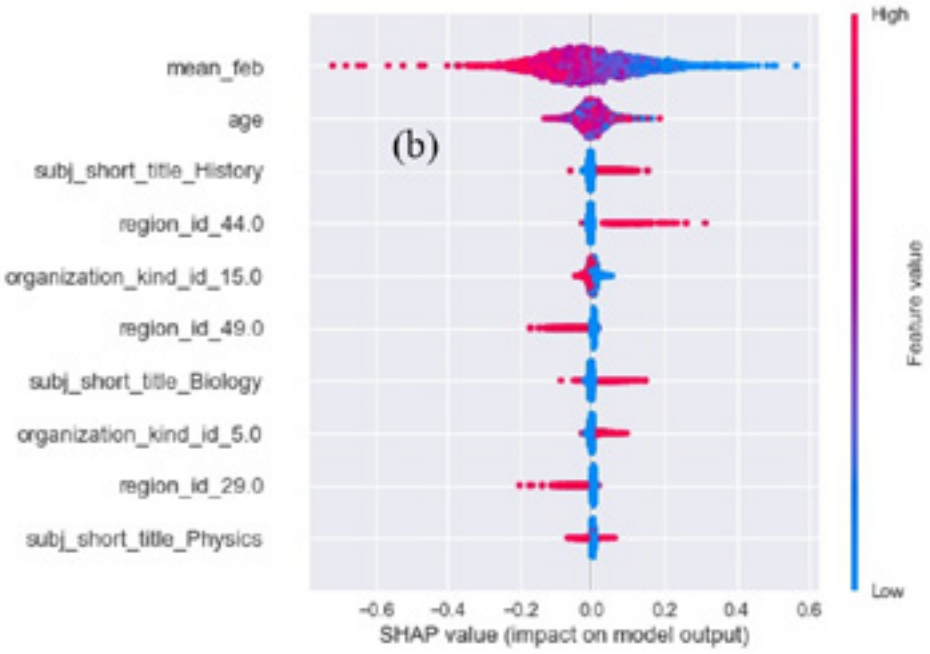


Fig. 4. The distribution of SHAP values (impact on parameter d) of explanatory features for predictive models: (a) for Gradient Boosting model with primary Explainer; (b) for Random Forest model with Tree Explainer. Cases with high values are shown in red, and those with low values are shown in blue. The variables are ranked in descending order. The horizontal location indicates whether the effect of that value is associated with a higher or lower prediction.

The main influence on the prediction of the Cohen's effect size value is exerted by the mean value of school grade in February and March, which obviously follows from the formula for the parameter . Also, significant contribution to the prediction of the parameter Cohen's effect size value is also made by the age of teachers: usually it is either not defined, or also negative (with an increase of age value, the value of the parameter decreases), which means that young teachers were more likely to give higher grades after introduction of distance learning format.

There exists also a significant improvement in school marks for the lessons of history, biology, while for such important subjects as physics, mathematics and Russian language, grades decreased after the introduction of distance learning. Besides that, location in certain regions: Naberezhnye Chelny, Kazan's Vakhitovsky, Novo-Savinovsky and Privolzhsky districts, also made significant positive contribution to the value of effect size . And in opposite, for schools located in Nizhnekamsk and Sovetsky district of Kazan, mean marks decreased significantly. Location of schools in the town also made positive contribution to the value of parameter d, while location outside of the town had a negative impact.

Besides that, different kinds of schools also played a special role as the used models features. The most significant influence was due to whether the educational organizations were secondary schools, lyceums, gymnasiums, or boarding schools. In case of lyceums, gymnasiums and boarding schools, the influence was strictly positive and increased the value of the Cohen's effect size , which means that after the introduction of distance learning into them, the marks of school grades increased. A different situation has developed in secondary schools: on average, the impact of the introduction of the distance learning format was mixed and did not affect academic performance in a certain way.

The influence of all the above factors may be explained by the fact that, depending on the characteristics of teachers, subjects taught and geographical location, the approach and time of transition to a new, previously practically unused format of education varied in different schools.

5. CONCLUSIONS

In this paper, we performed analysis of variation of academic performance in a large set of schools in Tatarstan Republic in the period before and during distance learning caused by COVID-19 pandemic.

https://doi.org/10.17993/3ctic.2022.112.136-144

We used eight different machine learning methods to solve the regression task of forecasting value of Cohen's effect size . We determined the values of the error function corresponding to all applied algorithms and established school classes for which prediction is easier and the ones for which prediction is more difficult.

We discovered impact of age of teachers to the forecasting of parameter; lessons for which marks were more significant in the studied task and areas of Tatarstan Republic, location of school in which increased or decreased Cohen's effect size. Moreover, we discovered that the kind of educational organization also plays a special role in the forecasting task and identified the ones which had a significant impact on the value of Cohen's effect size. The impact of these study-related factors may indicate that different schools, school types and teacher had different periods of adaptation to a rapidly changing learning format, and these changes can be evaluated using feature importance method in combination with machine learning algorithms.

The results obtained during the research, after appropriate verification, may be used to evaluate the influence on academic performance of school students after introduction of distance learning.

ACKNOWLEDGMENTS

The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation, project № 22-28-00923, "Digital model for predicting the academic performance of school-children during school closings based on big data and neural networks".

REFERENCES

[1] BENGIO, Y., COURVILLE, A., AND VINCENT, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans Pattern Anal Mach Intell 35*, 1798–1828.

[2] BOSTANABAD, R., BUI, A., XIE, W., APLEY, D.W., AND CHEN, W. 2016. Stochastic microstructure characterization and reconstruction via supervised learning. *Acta Materialia 103*, 89–102.

[3] COHEN, J. 1988. *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Lawrence Erlbaum Associates, Hillsdale, NJ.

[4] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2009. *The Elements of Statistical Learning*. Springer, New York.

[5] JESLET, D.S., KOMARASAMY, D., AND HERMINA, J.J. 2021. Student Result Prediction in Covid-19 Lockdown using Machine Learning Techniques. *JPCS 1911*, 012008.

[6] LI, J., AND JIANG, Y. 2021. The Research Trend of Big Data in Education and the Impact of Teacher Psychology on Educational Development During COVID-19: A Systematic Review and Future Perspective. *Front. Psychol. 12*, 753388.

[7] LIVIERIS, I.E., DRAKOPOULOU, K., TAMPAKAS, V.T., MIKROPOULOS, T.A.M AND PINTELAS, P. 2019. Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. *J. Educ. Comput. Res. 57*, 448–470.

[8] MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A.A., VENESS, J., BELLEMARE, M.G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A.K., OSTROVSKI, G., PETERSON, S., BEATTIE, C., SADIK, A., ANTONOGLOU, I., KING H., KUMARAN, D., WIERSTRA, D., LEGG, S., AND HASSABIS, D. 2015. Human-level control through deep reinforcement learning. *Nature 518*. 529–533.

[9] NUANMEESERI, S., POOMHIRAN, L., CHOPYITAYAKUN, S., AND KADMATEEKARUN, P. 2022. Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network. *Int. J. Inf. Educ. Technol. 12*, 851–857.

[10] PARK, Y-E. 2020. Uncovering trend-based research insights on teaching and learning in big data. *J. Big Data 7*, 1–17.

[11] SAHAKYAN, M., AUNG, Z., AND RAHWAN, T. 2021. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access 9*, 135392.

https://doi.org/10.17993/3ctic.2022.112.136-144

[12] SHAPLEY, L.S. 1953. *Contributions to the Theory of Games vol 2*. Princeton University Press, Princeton.

[13] USTIN, P., SABIROVA, E., ALISHEV, T., AND GAFAROV, F. 2022. Key Factors of Teacher's Professional Success in the Digital Educational Environment. *ARPHA Proceedings 5*, 1747–1761.

https://doi.org/10.17993/3ctic.2022.112.136-144