International Journal of Artificial Intelligence in Education (2024) 34:274–300
https://doi.org/10.1007/s40593-023-00331-8

ARTICLE

Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention

Marcell Nagy[1] · Roland Molontay[1,2]

Accepted: 7 February 2023 / Published online: 14 March 2023

Abstract

Student drop-out is one of the most burning issues in STEM higher education, which induces considerable social and economic costs. Using machine learning tools for the early identification of students at risk of dropping out has gained a lot of interest recently. However, there has been little discussion on dropout prediction using interpretable machine learning (IML) and explainable artificial intelligence (XAI) tools. In this work, using the data of a large public Hungarian university, we demonstrate how IML and XAI tools can support educational stakeholders in dropout prediction. We show that complex machine learning models – such as the CatBoost classifier – can efficiently identify at-risk students relying solely on pre-enrollment achievement measures, however, they lack interpretability. Applying IML tools, such as permutation importance (PI), partial dependence plot (PDP), LIME, and SHAP values, we demonstrate how the predictions can be explained both globally and locally. Explaining individual predictions opens up great opportunities for personalized intervention, for example by offering the right remedial courses or tutoring sessions. Finally, we present the results of a user study that evaluates whether higher education stakeholders find these tools interpretable and useful.

Keywords Dropout prediction · Higher education · Interpretable machine learning · Educational data science · Explainable AI

Abbreviations

AUC	Area Under the ROC Curve
BME	Budapest University of Technology and Economics
GPA	Grade Point Average

✉ Marcell Nagy
marcessz@math.bme.hu

Roland Molontay
molontay@math.bme.hu

[1] Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics, Műegyetem Rkp. 3., Budapest 1111, Hungary

[2] ELKH-BME Stochastics Research Group, Műegyetem Rkp. 3., 1111 Budapest, Hungary

HSGPA	High School Grade Point Average
ICRM	Interpretable Classification Rule Mining
IML	Interpretable Machine Learning
LIME	Local Interpretable Model-agnostic Explanations
LLM	Logit Leaf Model
MOOC	Massive Open Online Courses
PI	Permutation Importance
SHAP	SHapley Additive exPlanations
STEM	Science, Technology, Engineering, and Mathematics
UES	University Entrance Score
XAI	EXplainable Artificial Intelligence

Introduction

Innovative and efficiently operating higher education is the basis of successful and thriving knowledge and technology-based economy, where services and production rely on knowledge-intensive and intellectual activities (Powell & Snellman, 2004). However, student drop-out and delayed completion are serious issues in higher education all over the world, especially in Science, Technology, Engineering, and Mathematics (STEM) programs, inducing both personal and social costs (Latif et al., 2015).

AI-based approaches have been extensively used in a number of scientific fields. The big data stored in educational administrative systems hold great potential for data-driven educational research. Artificial intelligence in educational settings can have several roles, for example, the recent developments in natural language processing make it possible to automatize grading of free text answers (Schneider et al., 2022) and even essays (Kumar & Boulanger, 2021). Additionally, AI can customize learning paths (Yu et al., 2017), or notify the instructors about which students are likely to drop out of a massive open online course (MOOC) (He et al., 2015). Applying the tools of AI to identify students at risk of dropping out and the factors affecting university success has also attracted a lot of research interest recently (Alyahyan & Düştegör, 2020; Helal et al., 2019; Márquez-Vera et al., 2016; Rovira et al., 2017; Varga & Sátán, 2021). To help higher educational stakeholders, AI-based intelligent tutoring systems and decision support tools have also been proposed throughout the last few years. For systematic reviews, we refer to Avella et al (2016); Dutt et al (2017); Zawacki-Richter et al (2019); and Rastrollo-Guerrero et al (2020).

While the vast majority of dropout prediction studies use machine learning algorithms to classify future university dropouts and graduates as accurately as possible, the interpretation of the trained models and the explanations of their predictions are usually neglected. However, besides having a machine learning model with a high classification performance, a well-functioning decision support system also helps in explaining why a certain prediction was output and what intervention should be performed to help the individual. The latter makes it possible to provide personalized guidance, remedial courses, and tutoring sessions for the students. Moreover, explaining the predictions and interpreting the results also help in making the model more transparent and establishing the trust of decision-makers which are key steps towards deploying these solutions in real education environments. These goals can be achieved by using the tools of a recently emerged field of machine learning, called interpretable machine learning (IML) or also referred to as explainable artificial intelligence (XAI), which involves methods that make black-box models transparent. For a comprehensive overview of the tools of IML, we refer to the book of Molnar (2020).

According to the taxonomy of Molnar (2020), interpretability methods can be classified as (1) intrinsic/post hoc, (2) model-specific/model-agnostic, and (3) local/global. Intrinsic methods involve low-complexity white-box models that are interpretable without using any IML tool, such as linear regression and decision rules. On the other hand, post hoc methods analyze the model after it has been trained. Model-specific interpretation means that one has to know the mechanism of the model to be able to interpret its prediction: the interpretation of intrinsically interpretable models is always model-specific. Model-agnostic tools can be used on any model after they have been trained, i.e., model-agnostic tools are post hoc. Finally, we say that the interpretation is local if it explains an individual prediction and global if it describes the behavior of the whole model.

Several papers use intrinsic methods in educational data science, such as Márquez-Vera et al (2016) proposed a modification of the so-called Interpretable Classification Rule Mining (ICRM) algorithm to identify students at risk of failing a course as early as possible. Similarly, Zhang et al (2019) also emphasize the importance of interpretability and they apply the ICRM algorithm to predict the performance of online learners. Cano and Leonard (2019) focus on underrepresented and underperforming students and they introduce an interpretable early warning system relying on multi-view genetic programming, which generates similar rules to the rule mining algorithm. To predict the risk of dropping out in online learning systems, Coussement et al (2020) use the logit leaf model (LLM), which combines a decision tree and logistic regression, and they argue that LLM has a good balance between predictive performance and interpretability. Moreover, the authors also propose a multilevel informative visualization for the LLM that not only shows the dropout probability but also helps understand the local prediction of the model.

There are only a few studies that apply post hoc model-agnostic interpretable machine learning tools (e.g., LIME and SHAP) in an educational setting. The local interpretable model-agnostic explanations (LIME) method provides interpretable explanations of individual predictions by locally approximating the black-box model's predictions with an intrinsically interpretable model such as linear regression (Ribeiro et al., 2016). The SHapley Additive exPlanations (SHAP) method is based on a game-theoretical concept, called Shapley value, and it can be used to explain both individual predictions and the effects of the features on the model output (Lundberg & Lee, 2017). Nagrecha et al (2017) predicted MOOC dropout and used LIME to make the individual predictions of black-box machine learning models (Random Forest and Gradient Boosting Trees) interpretable, moreover, they compared it to intrinsically interpretable models (Logistic regression and Decision Tree). Recently, Vultureanu-Albişi and Bădică (2021) used tree-based ensemble classifiers to predict students' performance in courses based on publicly available small data sets, moreover, similarly to Nagrecha et al (2017), they applied LIME to interpret the local predictions of the model. Sargsyan et al (2020) performed a cluster analysis to find groups of students with similar characteristics, by first predicting the students' GPA scores and then clustering the students using the weights from the local prediction explanations given by LIME. The authors argue that their proposed approach finds groups of students that have similar academic attainment indicators.

Most of the related works use LIME to explain local predictions, however, there are a few recent works that utilize the game theory-based approach: the Shapley Additive explanation (SHAP) that was introduced by Lundberg and Lee (2017). Mingyu et al (2021) predicted the weighted grade point average of university students using a CatBoost regressor, moreover, they applied SHAP to globally interpret the model and estimate the importance of the features. Similarly, Karlos et al (2020) predicted the final grades of undergraduate students in an online course and used SHAP values to study the global contribution of the features. Moreover, in a closely related work, Smith et al (2021) did not only investigate the global importance and general effect of features, but they used both SHAP and LIME to locally explain the predictions of a model that was designed to identify students at risk of course failing. While LIME seems to be more frequently used in related works, Smith et al (2021) and Molnar (2020) suggest using SHAP for local prediction explanations instead of LIME because SHAP is found to be more stable than LIME. In this work, we mainly focus on SHAP's explanations of the individual predictions, but to provide a comparison, we also demonstrate how LIME can be used for local explanations.

While the related works predict students' grade point average and performance in specific courses, in this paper, the output variable is the final academic performance of an undergraduate student, i.e., we aim to distinguish between students expected to graduate and students at risk of dropping out. This work can be considered as an extension of our previous conference papers (Baranyi et al., 2020; Nagy et al., 2019). Earlier, we investigated how to utilize deep learning algorithms for globally interpretable dropout prediction (Baranyi et al., 2020). Moreover, we have also presented a web application for interpretable dropout prediction (Nagy et al., 2019). The main contributions of the present work are that we thoroughly demonstrate how explainable artificial intelligence tools can be used for interpretable dropout prediction, and more importantly to provide personalized feedback for the students by highlighting which skills are needed to be improved to increase the chances of graduation. Furthermore, we compare the output and readability of different XAI tools, and finally, we present a user study where we surveyed students and higher education decision-makers about the applied tools. We assessed the participants' data visualization literacy, moreover, asked their opinion about the interpretability, utility, and design of the charts.

We interpret our black-box machine learning model both globally and locally. First, we identify the most influential features and study their impact on the model's output, moreover, we compare it to the intrinsically interpretable logistic regression. In contrast to the majority of the aforecited related works, this work also concentrates on the local explanations of predictions besides the global interpretation of the model. Namely, using the power of SHAP values and LIME, we study the reasons behind individual predictions and compare these two methods. In addition, we also evaluated the interpretability of the two methods using user experience research, which involved gathering feedback from stakeholders. Note that in contrast to related works, we also adjust the predictions of our machine learning model so that the outputs can be interpreted as the probability of graduation.

Data description

This study is based on data from the Budapest University of Technology and Economics (BME), a large public university in Hungary. At BME, programs are offered in the following fields of study: engineering & technology, economics & social sciences, and natural sciences. The data set contains students who enrolled between 2013 and 2017 and finished their undergraduate studies either by graduation or dropping out. First, we performed some data preparation steps: We excluded the incomplete rows, and we considered only the "first try" of the students, i.e., if a student dropped out and then re-enrolled again, we only considered the outcome of the first program that they started. We also excluded those students whose reason for dropout is changing majors. The prepared data set contains 8,508 records.

The features of our data set are mostly those pre-enrollment achievement measures that are used to calculate the composite university entrance score (UES) on what the student's admission is based on (Nagy & Molontay, 2021). More precisely, we have data on their high school grades and their scores on the high school leaving exam, called *matura*. High school performance is measured by the high school grade point average (HSGPA) of the following subjects: history, mathematics, Hungarian language and literature, a foreign language, and a science subject of the student's choice. The matura consists of four core subjects and at least one elected subject. The core subjects are history, Hungarian language and literature, a foreign language, and mathematics.

Originally matura exam scores range between 0 and 100, however, students can take matura exams at a normal and advanced level. To avoid the usage of additional variables indicating the level of the exam, in the case of an advanced level exam, we multiplied the score by 1.35. This rewarding scheme was found to have the highest predictive power on student success (Molontay & Nagy, 2022). Thus, a matura exam score above 100 necessarily means that the student took the advanced level exam.

The data set contains a variable called *Lang. cert.*, which measures the quantity and level of the foreign language certificates that the student has. A complex language exam at level B2 is worth a 1-point score, and if the student only has "half" of the exam (only written/oral exam), then it is worth a 0.5-point score. The scores for level C1 exams are twice as much as the points for the B2 level.

In addition to pre-enrollment achievement measures, the data set also contains some personal data, such as gender. We also defined a generated feature called *Years between*, which measures how many years elapsed between high school graduation and enrollment in the university.

Finally, our target variable, the *Final status* is defined as a binary variable having a value of one if the student graduated and zero if the student is a dropout.

To sum up, our data set contains the following features: *Mathematics* (score in math exam), *Chosen subject* (exam score in the chosen subject), *Hungarian* (exam score in Hungarian language & literature), *History* (exam score in history), *Foreign lang.* (exam score in a foreign language), *HSGPA* (high school grade point average), *Lang. cert.*, *Years between*, *Male* (1 for males and 0 for females), *Eng. and Tech.* (1 if the field of study is engineering and technology, 0 otherwise), *Nat. Sci.* (1 if the field of study is natural sciences, 0 otherwise). If both *Eng. and Tech.* and *Nat. Sci.* are zero, it means that the field of study is economics and social sciences. Finally, the target variable, called *Final status*, denotes the final academic status: 1 for graduates and 0 for dropouts.

Methodology

To avoid data leakage – in contrast to the vast majority of related works – we do not evaluate our machine learning models by using a random train-test split or cross-validation, but we split the data along the enrollment year. Namely, we train the model on students who enrolled between 2013 and 2016 (6,398 observations) and evaluate the model's performance on the student cohort enrolled in 2017 (2,110 observations). On the first hand, this procedure makes the model's utility more informative when it is run in a real environment because during deployment the model is trained on historical data and applied to current students. On the second hand, the correlation of the variables within the same student cohort (year) might be higher than across cohorts, hence if there were students from the last cohort (the year 2017) in the training set, it would yield a higher model performance, but it would only be an artifact of the within-cohort correlations. For the same reasons, Yu et al (2021) also suggest using a cohort-based train-test split. The proportion of graduates in the training and test sets is 65% and 63%, respectively.

A central element of this work is model interpretability and explainability, which are crucial in machine learning, especially in high-risk environments where predictions have implications (Adadi & Berrada, 2018; Gunning et al., 2019; Molnar, 2020). Interpreting the results of dropout predictions assists students, policy-makers, and other stakeholders by shedding light on factors affecting academic performance and being an at-risk student. Moreover, it also makes it possible to implement a personalized intervention.

Complex, hardly interpretable machine learning models like gradient-boosted trees and neural networks have better performance than simple, interpretable models like logistic regression and shallow decision trees. However, with the help of post-hoc model-agnostic IML tools, the trade-off between performance and interpretability no longer poses a problem. These tools make black-box models interpretable and transparent.

For model selection, we test several machine learning algorithms (their performance is detailed in Table 1) and to make predictions, we use the best-performing model, the CatBoost algorithm, which achieves state-of-the-art results on several tabular benchmark data sets (Prokhorenkova et al., 2018), where the most recent deep learning models are under-performing compared to tree-based models

Table 1 Performance (AUC and average precision) of the machine learning models on the test data set. The models are ordered by their AUC score, and their rank according to the average precision is written in parenthesis

Model	AUC	Average precision
CatBoost	0.774	0.847 (1)
Natural Gradient Boosting Classifier (NGBoost)	0.767	0.838 (2)
Explainable Boosting Classifier	0.760	0.833 (3)
Logistic Regression	0.734	0.799 (5)
Extreme Gradient Boosting (XGBoost)	0.732	0.813 (4)
Gradient Boosting Classifier	0.728	0.714 (7)
Linear Discriminant Analysis	0.722	0.710 (10)
Ada Boost Classifier	0.717	0.710 (9)
Light Gradient Boosting Machine	0.708	0.714 (6)
Quadratic Discriminant Analysis	0.707	0.711 (8)

(Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022). As a baseline intrinsically interpretable model, we use logistic regression.

For model interpretation, we use modern techniques such as permutation importance (Fisher et al., 2019), two-dimensional partial dependence plots (Greenwell et al., 2018), Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), and SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017), that is based on the game-theoretical concept of Shapley value.

For global interpretation, we calculate the feature's importance and study their partial dependence plots. In particular, we employ two methods to estimate feature importance in our model, permutation importance and the built-in method in CatBoost. Permutation importance (PI) works by shuffling the values of a feature and observing the effect on the model's performance (Fisher et al., 2019). The more the performance decreases due to shuffling, the more important the feature is. To "cancel out" the randomness of the method, we average the importance measures over 100 repetitions. On the other hand, the built-in method in CatBoost is based on the model's internal workings, namely, it works by computing the average reduction in the model's loss caused by each feature during the training process (Prokhorenkova et al 2018).

To further investigate the relationship between features, we use two-dimensional partial dependence plots (PDP). These plots show the effect of two features on the model's output while holding other features constant (Greenwell et al., 2018).

Feature importance metrics and PDP helps the global interpretation of the model. For local interpretations, i.e., a more detailed understanding of individual predictions, we use SHAP values, which provide a measure of the contribution of each feature to a specific prediction. In particular, the contribution of the features is compared to the average prediction of the model, which is called the base value. Hence, it explains an individual prediction by showing how and to what extent the features are moving the model's prediction from the base value (Lundberg & Lee, 2017). By aggregating these values, we can also obtain global importance for each feature. Additionally, we also use LIME which is a popular tool for local interpretability, it creates an interpretable model locally around the prediction using linear approximations of the complex model (Ribeiro et al., 2016).

Note that one has to be careful with using the tools of explainable artificial intelligence. For example, both PI and PDP assume that the features are statistically independent, hence in the case of correlated features they can lead to misleading interpretations (Molnar et al., 2020). That is why we also study the correlation of the variables and interpret the results with caution.

In order to be able to interpret the output of the model as a probability, it is important to consider any potential bias in the model. One way to correct for bias is by applying calibration methods such as Platt's calibration (Platt et al., 1999) and isotonic regression (Niculescu-Mizil & Caruana, 2005). Platt's calibration, which is based on logistic regression, adjusts the output of the model to better reflect the true probabilities of the classes (Platt et al., 1999). On the other hand, the isotonic regression method is a non-parametric approach that finds the optimal piecewise constant function that maps the model's output to the true probabilities of the classes (Niculescu-Mizil & Caruana, 2005).

As a tool for evaluating explainable machine learning methods, we conducted a user study. The questionnaire for the students collected information about the clarity, interpretability, utility, and aesthetics of the local explanations provided by SHAP and the ease of interpreting force plots and bar plots. The form for decision-makers had additional questions about the global effect of features and interaction plots. To assess the participants' understanding of the visualizations, we asked test questions and used a 5-point Likert scale to gather their opinions on interpretability, usefulness, and appeal. Participants were also able to leave free-text comments about the visualizations.

Results and discussion

In this section, using the data set of BME we present how explainable artificial intelligence tools can help discover the effect of the most influential factors in dropout prediction, and how the output of these tools can be used for personalized intervention and feedback provision.

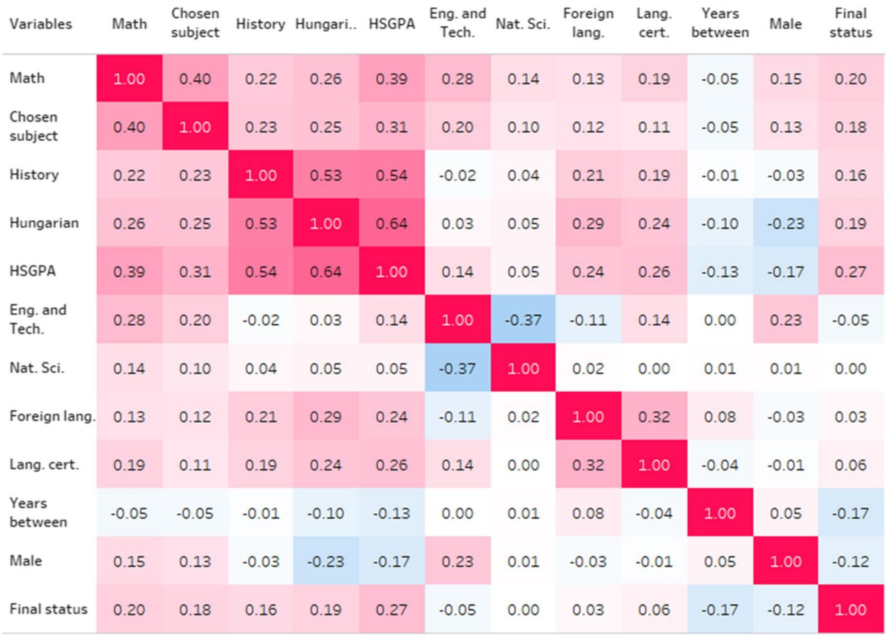
First, we focus on the global interpretation of our model by identifying the importance of the features. Using SHAP values, we analyze how the features generally affect the probability of graduation, moreover, we also study the interaction of the features using 2D partial dependence plots. Then in the subsequent subsection, we demonstrate how individual feedback can be provided using SHAP and LIME techniques. Finally, we address the problem of the interpretation of the model output, i.e., whether it indeed has a probabilistic meaning.

Global interpretation

First, before fitting any machine learning models, we present the correlation of the variables, including the target variable. Figure 1 shows the correlation heatmap of the features of our data set. We can observe that the matura exam score in mathematics and HSGPA variables correlate most strongly with the final status (point-biserial correlation).

In what follows, we apply XAI tools to interpret the machine learning model that we trained to predict the final academic status. We have tested several machine learning models, including XGBoost, Linear Discriminant Analysis, CatBoost, AdaBoost, NGBoost, Explainable Boosting Machine, etc., and we have found that CatBoost has the best performance on our data set (see Table 1). On the test data set, the CatBoost, optimized with Optuna (Akiba et al., 2019), achieved an average precision of 0.847 and AUC of 0.774. The achieved performance is in alignment with the related works that perform dropout prediction on a large heterogeneous data set, e.g., Behr et al (2020) achieved an AUC of 0.77, moreover, the average precision of the best-performing model of Lee and Chung (2019) is 0.898. In this work, we also apply logistic regression to compare both the predictive performance of the model and to demonstrate interpretable dropout prediction using an intrinsically interpretable model. The average precision and AUC scores of the logistic regression are 0.799 and 0.734, respectively.

Table 2 shows the results of the logistic regression, namely the coefficients of the features and the corresponding standard error and significance (p-value). The



Variables	Math	Chosen subject	History	Hungari..	HSGPA	Eng. and Tech.	Nat. Sci.	Foreign lang.	Lang. cert.	Years between	Male	Final status
Math	1.00	0.40	0.22	0.26	0.39	0.28	0.14	0.13	0.19	-0.05	0.15	0.20
Chosen subject	0.40	1.00	0.23	0.25	0.31	0.20	0.10	0.12	0.11	-0.05	0.13	0.18
History	0.22	0.23	1.00	0.53	0.54	-0.02	0.04	0.21	0.19	-0.01	-0.03	0.16
Hungarian	0.26	0.25	0.53	1.00	0.64	0.03	0.05	0.29	0.24	-0.10	-0.23	0.19
HSGPA	0.39	0.31	0.54	0.64	1.00	0.14	0.05	0.24	0.26	-0.13	-0.17	0.27
Eng. and Tech.	0.28	0.20	-0.02	0.03	0.14	1.00	-0.37	-0.11	0.14	0.00	0.23	-0.05
Nat. Sci.	0.14	0.10	0.04	0.05	0.05	-0.37	1.00	0.02	0.00	0.01	0.01	0.00
Foreign lang.	0.13	0.12	0.21	0.29	0.24	-0.11	0.02	1.00	0.32	0.08	-0.03	0.03
Lang. cert.	0.19	0.11	0.19	0.24	0.26	0.14	0.00	0.32	1.00	-0.04	-0.01	0.06
Years between	-0.05	-0.05	-0.01	-0.10	-0.13	0.00	0.01	0.08	-0.04	1.00	0.05	-0.17
Male	0.15	0.13	-0.03	-0.23	-0.17	0.23	0.01	-0.03	-0.01	0.05	1.00	-0.12
Final status	0.20	0.18	0.16	0.19	0.27	-0.05	0.00	0.03	0.06	-0.17	-0.12	1.00

Fig. 1 The correlation heatmap of the variables of our data set

Table 2 Results of the logistic regression on the training data set. The variables are ordered by their *p*-value. Pseudo R-squared: 0.091

Variable	Coefficient	Std err	*P*-value
Years between	-0.164	0.014	0.000
Eng. and Tech	-1.124	0.103	0.000
Math	0.018	0.002	0.000
Male	-0.680	0.072	0.000
Foreign lang	-0.018	0.002	0.000
Nat. Sci	-1.489	0.198	0.000
Chosen subject	0.011	0.002	0.000
HSGPA	0.335	0.073	0.000
Lang. Cert	0.155	0.040	0.000
History	-0.002	0.003	0.362
Hungarian	-0.001	0.003	0.644

coefficients show, for example, that the higher the value of the *Years between* variable is, the less likely the graduation is. On the other hand, a better HSGPA or higher scores in the matura exam in mathematics and in the chosen subject significantly increase the probability of graduation.

According to the logistic regression, the coefficient of the matura exam scores in Hungarian and history are not statistically significant, however, that is due to their high correlation with the HSGPA variable (multicollinearity), see Fig. 1. If we remove any two of these three highly-correlated variables, then the remainder feature will have a statistically significant positive coefficient. The coefficients and standard errors of these variables after the removal of the others are as follows: HSGPA: 0.277 (0.052), *History*: 0.005 (0.002), *Hungarian*: 0.006 (0.002).

We also calculated the CatBoost's built-in feature importance and permutation importance of the features, the results are shown in Table 3. According to both importance metrics, the most important features are HSGPA, mathematics, and

Table 3 The importance of the features according to CatBoost and PI. The permutation importance is calculated on the test data set with respect to the AUC score. The permutations were repeated 100 times and the mean importance is shown. The features are ordered by the CatBoost importance, however, their rank according to the PI is written in parenthesis

Variable	CatBoost feature importance	Permutation importance
HSGPA	18.321	0.055 (1)
Math	18.185	0.044 (2)
Years between	17.766	0.018 (6)
Chosen subject	11.618	0.026 (3)
Male	8.859	0.019 (5)
Eng. and Tech	7.323	0.023 (4)
Foreign lang	5.462	0.003 (8)
History	4.301	0.003 (9)
Hungarian	3.353	0.001 (11)
Lang. Cert	2.893	0.003 (10)
Nat. Sci	1.920	0.004 (7)

the chosen subject. Note that the high correlation between the features can also confuse the feature and permutation importance metrics (Molnar et al., 2020), probably that is why the importance of humanities (foreign language, history, Hungarian language & literature) is low especially according to the PI.

To understand the joint impact of features on the prediction, we turn to two-dimensional partial dependence plots. The top left plot of Fig. 2 shows predictions for any combination of mathematics and history matura exam scores. The CatBoost model's output increases with higher scores in mathematics, which is in alignment with the correlation analysis (Fig. 1) and the results of the logistic regression (Table 2). Moreover, the highest probability of graduation can be achieved if the score in history is also relatively high. On the other hand, if one has an outstanding score in mathematics and the score in history is too high, the predicted probability of graduation slightly decreases.

Similarly, the top right contour plot of Fig. 2 shows the interaction between the Hungarian language & literature and mathematics matura exam scores. Since the data of a technical university is being investigated, mathematical skills strongly influence the model's prediction, however, the model returns the highest probability of graduation when the score in the Hungarian exam is also high.

The bottom left plot of Fig. 2 shows the interaction of the HSGPA and the score in the mathematics exam. The figure clearly illustrates that these features are equally important, and students are more likely to graduate if they achieve high scores in mathematics and have excellent grades in high school.

The bottom right figure suggests that if senior high school students do not start their university studies right after graduation, then their chances of obtaining a degree are decreasing over the years. Moreover, the diagonal contour lines indicate that HSGPA and *Years between* both have a high influence on the model output.

Figure 3 shows the SHAP summary plot, which helps to understand how features affect the probability of graduation in general. It gives an overview of the overall effect of all features, their importance, and also the distribution of the students along the features.

Figure 3 illustrates the impact of the features on the model's output according to the SHAP technique. Recall that the SHAP values quantify the marginal contribution of the features to the model's local prediction. The plot indicates, that the most important variable is the HSGPA, which is in alignment with the results of permutation importance (Table 3). In our earlier work, we have also shown that among the components of the university entrance score, HSGPA has the highest predictive power on the final academic success (Nagy & Molontay, 2021). The figure clearly shows that the higher the HSGPA is, the more likely the graduation is, since a high feature value pushes the prediction higher, and a low feature value pulls the model prediction lower.

Since the examined university is a technical university, it is not so surprising that math skill is a great predictor of final academic success. It is a well-known global problem that mathematics is the main reason for dropout and delayed completion in STEM education (Baranyi & Molontay, 2021). There is a similar relationship between the chosen subject and the probability of graduation. That is probably due to the fact that by the design of the admission procedure, the chosen subject of the

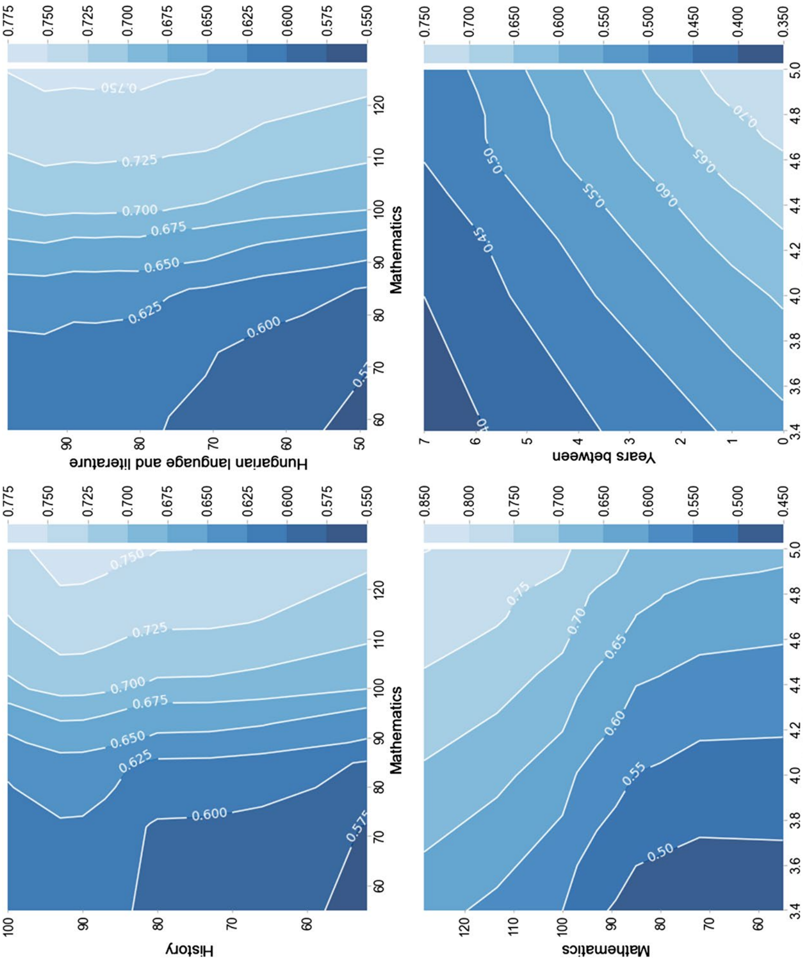


Fig. 2 Two-dimensional partial dependence plot of mathematics with *Hungarian* and *History* (top figures) and the interaction of the *HSGPA* with *Mathematics* and *Years between* (bottom figures)

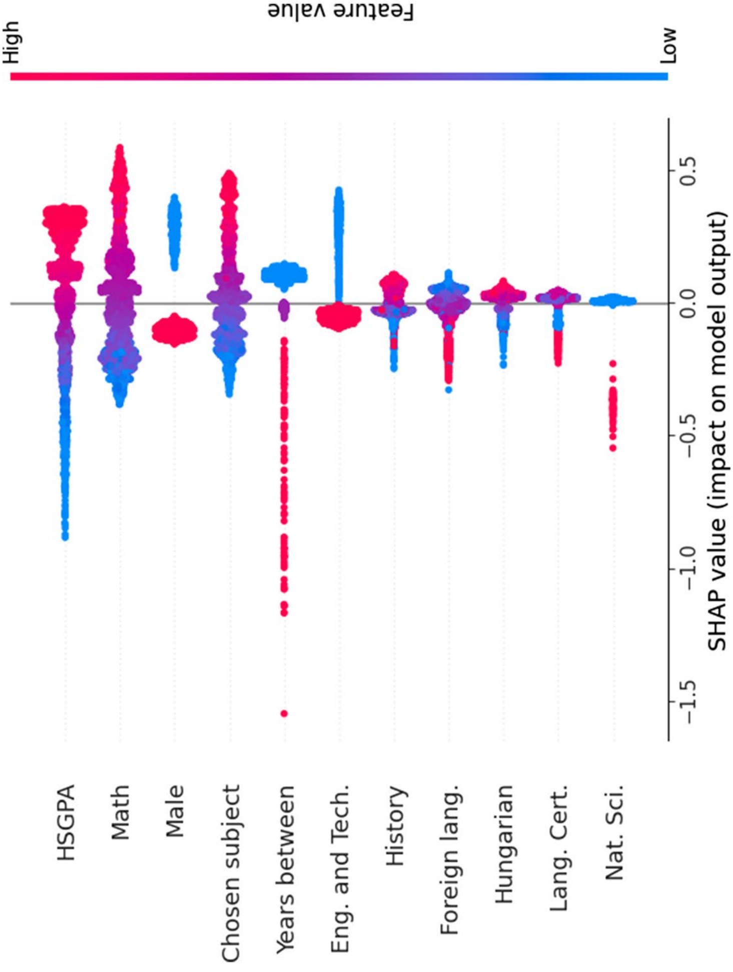


Fig. 3 SHAP summary plot. Each student is represented with one point in each row. The points are colored by the value of the corresponding feature and the *x*-axis shows the impact on the model's prediction. The features are ordered by their importance (average absolute impact on the model output)

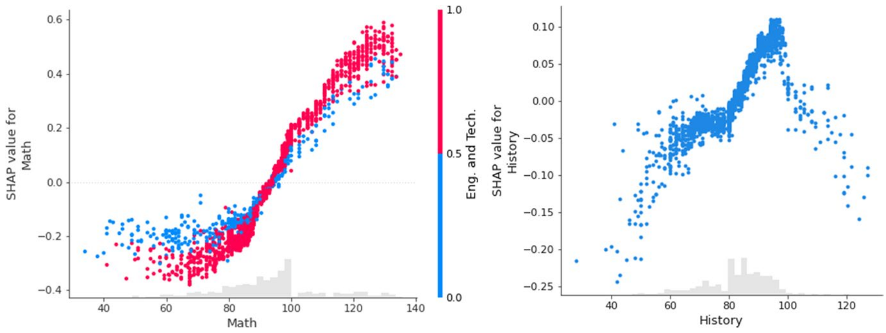


Fig. 4 Scatterplot of the value of the mathematics (left) and history (right) matura exam scores on the x-axis and the SHAP value of these features on the y-axis. The density of the data is shown with the gray histogram along the x-axis

matura exam in most cases has to be related to the field of study (the set of acceptable subjects is defined by the undergraduate programs).

We can observe, that according to Fig. 3 being male has a negative effect on the model output. That is in alignment with our earlier work (Nagy & Molontay, 2021), where we have shown that female students are more likely to graduate at BME than males.

Quite surprisingly both logistic regression (see Table 2) and the SHAP values (see Fig. 3) suggest that foreign language matura exam score has a significant negative effect on the probability of graduation.

In the case of Hungarian and history matura exams, a higher score typically implies a higher probability of graduation, however, the impact of these features on the model output is smaller compared to the exam scores of mathematics and the chosen subject.

If the field of study is engineering & technology or natural sciences, then it negatively affects the probability of graduation. However, when both of these indicators are zero – meaning that the field of study is economics & social sciences – has a positive effect on the model output. That is because the graduation rate varies across disciplines, and it is the highest in the case of economics & social sciences at BME.

Figure 4 shows how the model depends on the given feature, which can be considered as a SHAP value-based alternative to the one-dimensional partial dependence plot. The left side of Fig. 4 shows the effect of the score in mathematics and its interaction with the binary Eng. and Tech. variable. From the figure, it is clear that the higher the score in mathematics is the higher the predicted probability of graduation is. Moreover, the figure also suggests, that mathematics has a slightly higher impact on university success in the case of engineering students (red points) because the range of the SHAP values is larger in the case of the red points, which means that the absolute impact is larger.

The right side of the figure shows a very interesting phenomenon, namely, that the relationship between the score in history and the probability of graduation is quadratic. In the lower half of the range, the higher the score in history is, the

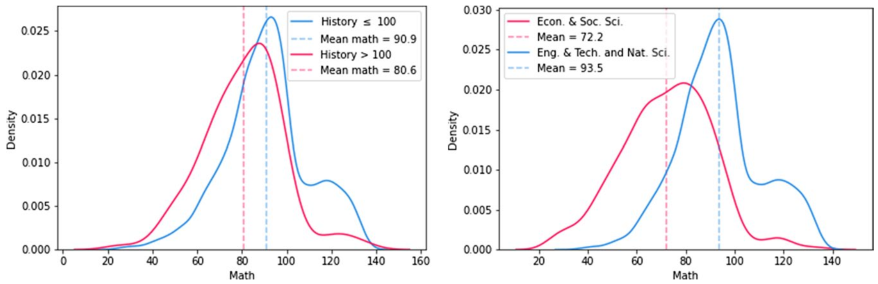


Fig. 5 The mean and the kernel density estimate (KDE) plot of the scores in mathematics matura exam in different student groups

higher the probability of graduation is, but if the score is above 100, an increase in the score decreases the model output. Overall, when the score is between 80 and 100, then it pushes the output higher, while outside of this region, the score contributes negatively to the prediction. To find out the reason for this quadratic effect of the history matura exam score requires further investigation. However, we argue that this is presumably because students have a finite amount of time to prepare for the high school exit exams, and if they focus more on a subject to achieve a high score, then they can allocate less time for other subjects necessarily. Potentially, they might have less time for subjects that are more important at this technical university. Furthermore, when the score is above 100, it indicates that the exam level is advanced, and typically students only take one advanced-level exam, hence their mathematical or program-specific knowledge might not be strong enough. This explanation is supported by Fig. 5, where the left side of the figure shows that those students who achieve more than 100 points in history have a lower score in mathematics. On the other hand, those having less than 100 points in history, are more likely to take an advanced-level math exam (since the blue line is way above the red line for high math exam scores).

The right side of Fig. 5 is related to the left side of Fig. 4 and it shows that not so surprisingly, economics & social sciences students typically take mathematics matura exam at a normal level and on average their mathematics skill is lower compared to STEM students. The fact that math score has a slightly smaller impact on the model output in the case of economics & social sciences is in alignment with the finding of Nagy and Molontay (2021), namely that in economics & social sciences pre-enrollment achievement measures have smaller predictive validity on university success than in STEM fields.

Explanation of individual predictions

In this section, we show how the local interpretation of black-box models can help provide personalized feedback for university students regarding their strengths and weaknesses. In other words, we demonstrate how SHAP and LIME can be used to explain individual predictions.

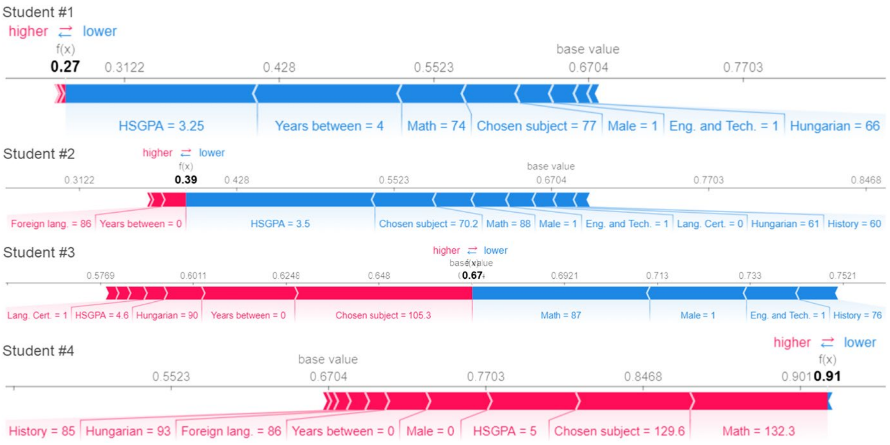


Fig. 6 Force plots of explanations of the individual predictions for four students using the SHAP values. The base value – the average prediction – is 0.67. The features pushing higher the model output from the base value are shown in red, and the features decreasing the prediction are in blue. The length of the bars corresponds to the amount of contribution of the feature

Figure 6 shows the so-called force plots of four examples for the individual explanations of predictions using SHAP values. Factors that negatively affect the prediction are shown in blue and those increasing the probability of graduation are in red. The output of the machine learning model (denoted by $f(x)$) is written in boldface.

In the case of the first student, the model output (the estimated probability of graduation) is 0.27, which means that this student is at high risk of dropout. However, the SHAP values also tell us why this student is at risk of dropout. There is nothing that would push the prediction higher, but the fact that 4 years have elapsed between high school and university, and that the mathematics score and the HSGPA are relatively low significantly decreases the model's prediction. The low HSGPA and the low exam scores in mathematics and in the chosen subject indicate that the student should improve both in general and program-specific knowledge, moreover, the fact that he started the university after a 4-year gap period suggests that both remedial courses and student success courses would be highly beneficial for this student at high risk of dropping out.

In the second example of Fig. 6 the estimated probability of graduation is somewhat higher, due to the fact that this student started their university studies right after high school and he achieved a good score on the foreign language exam. However, a similar action plan would be desirable for this student as well.

The third example shows a borderline student, whose prediction is the same as the base value (average prediction). The student's score in the chosen subject is high, moreover, he also started university after high school, which has a large positive impact on the prediction. On the other hand, relatively low math and history scores push the prediction lower and they "cancel out" the effect of positive factors. Moreover, since male students are less likely to graduate at BME than females, gender also negatively affects the prediction. Since the mathematics score has the most

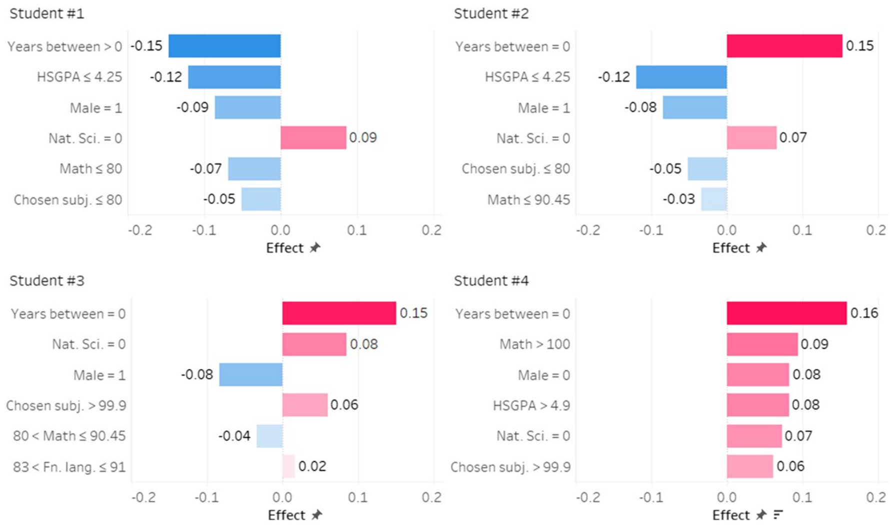


Fig. 7 LIME explanations for the four instances that are shown in Fig. 6. The *x*-axis shows the feature effect and the explanations are created with six features

significant negative effect on the prediction of this student, as a personalized intervention plan, mathematical remediation should be offered for him.

Finally, the last example in Fig. 6 shows a smart student whose perfect HSGPA and excellent matura scores in mathematics and the chosen subject contribute the most to the positive prediction. In addition, the aforementioned fact that females are more likely to graduate pushes the probability of graduation higher.

LIME is one of the most frequently used tools in related works to explain individual predictions. In what follows, we demonstrate the output of LIME and compare it to the SHAP explanations. LIME approximates the prediction of the model locally around the student of interest using an interpretable linear model, in particular, LIME returns weights (coefficients) of the binned versions of the features. Similarly to SHAP, the output of LIME can also be used for planning a personalized intervention. Figure 7 shows the explanations of the predictions for the same four students who are displayed in Fig. 6. Similarly, the negative features are colored blue and the positive features are shown in red. The two methods agree, i.e., both SHAP and LIME find the same factors influential and more importantly, they also agree on how these features affect the probability of graduation. On the other hand, in all these four cases the *Years between* seems to have the highest impact on the prediction, yet the scores in mathematics and chosen subject seem to have a slightly smaller impact, compared to the SHAP explanations.

The reason why SHAP and LIME use different representations for the explanations of the local predictions is that LIME simply shows the coefficients of the local linear model, while SHAP shows “forces”, i.e., SHAP shows the relative impact of the features with respect to the base value. However, it is also possible to visualize SHAP values on bar charts, which can be interpreted as signed feature importance (see Fig. 9).

It is important to note, that based on the local explanations of SHAP and LIME presented in Figs. 6 and 7, it is not clear how the change of the value of a feature would affect the prediction of the model. For example, in the third example, both SHAP and LIME show that 87 points in mathematics have a negative effect on the model output, but it does not tell us if it is because this value is too low or too high. Of course, we know that the higher the mathematics score is, the higher the probability of graduation is, as Figs. 3 and 4 show, but this information is not present in the local explanations. In general, these explanations only tell the contribution of a factor but do not reveal whether the value of a feature should be increased or decreased to push the model prediction higher. However, it is also possible to extract this information from LIME as follows. By default, LIME discretizes the continuous variables – as can be seen in Fig. 7 – and performs a Lasso regression with dummy variables, but it is also possible to use the continuous variables without discretization. Unfortunately, as Molnar (2020) also points out, if the continuous features are not categorized, then the output of LIME is less interpretable, and that is why the implementation discretizes these features by default.

While the results given by SHAP and LIME are consistent, we suggest using the SHAP values for local explanations because it also shows the model output (the probability of graduation), the discretization of the features by LIME can be confusing, and finally, because it has proven to be more stable than LIME (Molnar, 2020; Smith et al., 2021).

To be able to use SHAP or LIME as an efficient decision-support tool, we believe that not only the contributions of the factors should be communicated, but also the relationship of the features with the model prediction. In other words, not only the explanation of a prediction is important, but also the identification of how the model prediction could be improved. One simple solution could be the communication of the interaction of the feature and the model prediction, which are shown in Fig. 4. There might be another solution, called counterfactual explanation, which is a recent research area of interpretable machine learning (Karimi et al., 2020; Looveren & Klaise, 2021; Mothilal et al., 2020). The goal of a counterfactual explanation is to find how to change the feature values to modify the prediction of the model to a predefined output (Molnar, 2020). For example, what a student should improve to change the model's prediction from “dropped out” to “graduated”. However, the problem with counterfactual explanations is that for one instance one can usually find several counterfactual explanations, and then it is not clear which explanation is the “best” (Molnar, 2020).

Interpretation of the model output

Although the output of the underlying black-box model is said to be the “probability” of graduation, usually the scores of binary classifiers cannot be directly interpreted as actual probabilities, which is also illustrated on the right side of Fig. 8. The used machine learning model, CatBoost, usually overpredicts the students' performance, i.e., it predicts a higher “probability” of graduation than what the actual

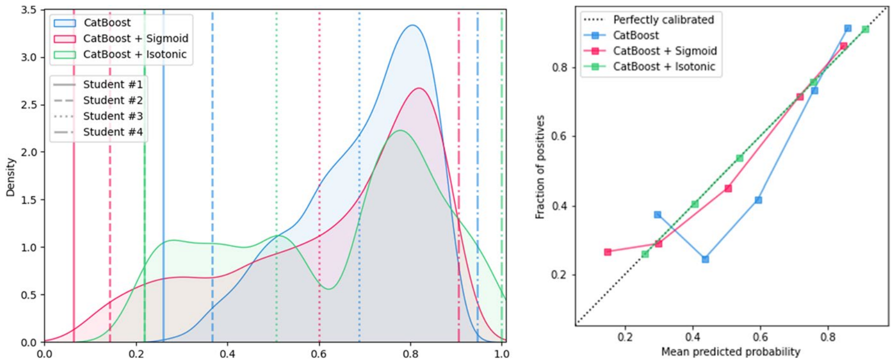


Fig. 8 Left: the distribution of the model's predictions on the test data set. The blue vertical lines indicate the positions of the four individual predictions that are shown in Fig. 6. Moreover, the red and green lines correspond to the sigmoid and isotonic regression corrected predictions, respectively. Right: the probability calibration curve of CatBoost and its calibrated versions

fraction of graduated students is. Hence, the direct usage of the model's output might be misleading. Another important step towards interpretability is to adjust for this bias. In this work, we make corrections using Platt's calibration (Platt et al., 1999) and the isotonic regression method (Niculescu-Mizil & Caruana, 2005). The left side of Fig. 8 shows the distribution of the model's output values in blue and the corresponding blue lines show the predictions for the four students investigated in Fig. 6. The red lines and red density function correspond to the corrected predictions using the sigmoid method, and the green lines and green density function correspond to the isotonic method. The solid line corresponds to the first, the dashed to the second, the dotted to the third, and the dash-dotted to the fourth student. The bimodal distribution of the predictions made by the isotonic calibration significantly differs from the other two distributions. However, the isotonic calibrated distribution seems to be the most natural, since the adjusted prediction distinguishes the graduates and dropouts better because the two distinct peaks correspond to the graduate and dropout student groups. Moreover, according to Niculescu-Mizil and Caruana (2005), the isotonic-regression-based calibration outperforms the logistic regression method when the size of the data set is large enough.

User experience research

One of the goals of IML/XAI tools is to produce outputs that are easily interpretable by the end user and to establish the trust of decision-makers and potential stakeholders. While several IML methods have been introduced recently, there is less research on the usability and clarity of these tools, or whether they are indeed interpretable on the application level. A short while ago Jin et al (2022) argued that XAI research ignores non-technical users, and they conducted a user study, where they assessed the XAI literacy of 32 adult laypersons. While the authors involved multiple XAI tools (e.g., feature importance, partial dependence plots, counterfactual examples),

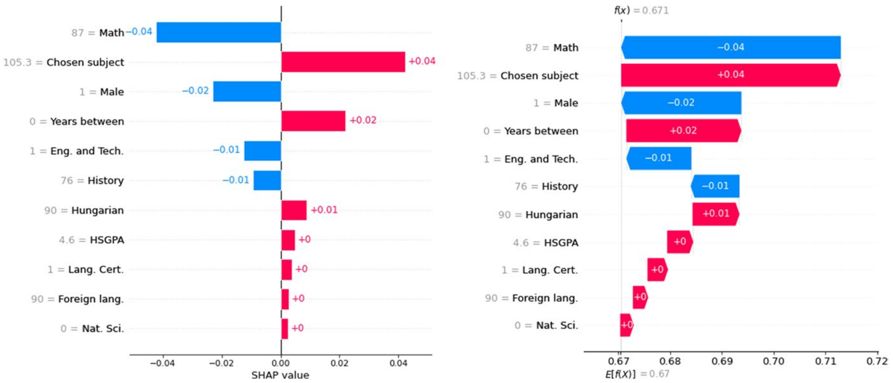


Fig. 9 Alternative visualizations of a local SHAP explanation. The left chart is called a (horizontal) bar plot, and the chart on the right is referred to as a waterfall plot. Both of the charts correspond to the same student (3rd student in Figs. 6 and 7)

Table 4 The results of the survey about individual explanations generated by SHAP. A five-point Likert scale was used and the table shows the mean responses

	Students ($n=42$)	Decision-makers ($n=10$)
Avg. % of correct answers	89%	95%
Easily interpretable	3.67	3.30
Useful	4.02	3.60
Appealing/stunning	4.21	4.30
Bar plot is easier to interpret than the force plot	3.95	4.30

they did not assess the interpretability of the visualizations of SHAP and LIME explanations.

To fill this gap, we have conducted two surveys, one with 10 higher education decision-makers, and another one with 42 students (involving both college and high school students). The questionnaire for the students focuses on the local explanations provided by SHAP, and whether force plots (Fig. 6) or bar plots are easier to interpret (left chart of Fig. 9). The form for decision-makers includes the same figures and questions, plus additional questions about the features' global effect (Fig. 3) and interaction plots (Fig. 2). We asked test questions and used a Likert scale to gauge participants' understanding and perception of the visualizations, and also allowed for free-text comments.

Table 4 shows the results of the survey about the interpretability of the individual explanations (Fig. 6). Overall, both students and decision-makers performed relatively well on the test questions about the individual visualizations, since the mean percentage of correct answers is relatively high (89% for students and 95% for decision-makers). On average, students find the visualizations more easily interpretable than the decision-makers, but both groups find these local explanations only moderately interpretable. Students also find these plots more useful, however, this might be because decision-makers find them harder to interpret. On the other hand, both students and decision-makers find the visualizations undoubtedly appealing.

On average both students and decision makers agree that the bar plots (Fig. 7 and left side of Fig. 9) are easier to interpret than the force plots (Fig. 6), however, decision-makers find the force plots significantly more complicated than the students do. On the other hand, the participants noted that a drawback of the bar plots is that the final result (predicted probability of graduation) is not shown, moreover in the case of multiple features the force plots are less transparent. One of the decision-makers said that "*The two types of plots (force and bar) have different advantages, the force plot shows the effect and its magnitude and the model output, and a bar chart may be more transparent about how much each factor affects the outcome if more (10–20) features are used.*". Similarly, some students also remarked that "*With many variables, sometimes the name of a column doesn't fit in the figure and it was confusing for me because I couldn't understand what was there without a name.*", moreover, "*The bar chart is better, but I couldn't read the probability of graduation from it.*". Some others said that "*Although I found the bar chart a little easier to understand, it is definitely a plus for the force plots that the 'final result' is clearly visible.*", likewise, another student said that "*Although the bar chart is easier to interpret, it is not enough to get the full picture, for which the force plot is much better. However, there should be some proactive explanatory part at the beginning, because at first glance it was not very clear to me.*".

Another student remarked that an advantage of the bar charts is that they could see the exact magnitude of the effects "*On the bar chart, I think it's very good to see specific numbers of which variables have what effect, it can be less obvious on the force plot, so it's definitely good supplementary information on the bar chart.*". Finally, according to a student which visualization is better depends on what the goal is: "*I can't choose between the two visualizations, as I'm not sure which one is better, as I'm not clear about the purpose. If the point is to see how at risk someone is of dropping out, the force plot is a better visualization. And if the goal is to see what the student needs to improve in order to be less likely to drop out, the bar chart is the purposeful visualization.*".

Finally, multiple students and decision-makers said that the colors are confusing since the color red usually denotes something bad, and in these examples, it denoted the positive effects: "*For positive influence, I wouldn't necessarily choose the color red because it's usually associated with something negative, and at first I was confused and had to read the instructions several times to make sure I understood the diagram correctly.*" another student said that "*For reasons of color symbolism, I would change the color of the figures to a light green–red pair, where light green would indicate a positive effect and red a negative one.*"

A solution for the problems mentioned by the respondents could be the usage of a bar plot together with a separate visualization showing the model output, or a mixture of the force and bar charts that shows the effects (forces) on a horizontal bar chart. This type of chart is referred to as a waterfall plot, which is illustrated on the right side of Fig. 9. Regarding the colors, an orange-blue color palette might be less confusing, since orange does not have an associated meaning, and it is colorblind-friendly unlike the combination of red and green (Shaffer, 2016).

According to the decision-makers, the SHAP summary plot (Fig. 3) is the most difficult to understand. The average score of the interpretability on a five-point Likert scale is only 2.4 (that of the utility is 3.6, and for the aesthetics is 4.1). A decision-maker said that "*In the SHAP summary plot, I found it difficult to make sense of all the information, such as the hue, the distribution, the order, and the side scale. But maybe it's just a matter of getting used to it, and after an example, interpretation will be easy for everyone.*", moreover, another decision-maker said that "*Neither a teacher nor a decision-maker will have time to figure out the complicated diagrams (SHAP summary and force plots). It is very important to keep the representations as simple as possible.".*

Overall, it seems that the SHAP summary plot compresses so much information in a chart that it makes it hard to understand at the first sight. It might be better to show the scatterplots for the features that are shown in Fig. 4, which also show the density of the data on histograms. The partial dependence plots (interaction plots shown in Fig. 2) are definitely the easiest to understand according to the decision-makers. The average scores for interpretability, utility, and aesthetics are 4.2, 4.4, and 4.8 respectively. One of the stakeholders remarked that "*The interaction diagram can be interpreted without further explanation. The others were not clear to me. An example description of how to interpret a specific figure would have helped*". One student also remarked that an example "walk-through" of the interpretation would have been necessary: "*The force plot would also be easier to understand if the description were more precise. It wasn't clearly explained, it would be nice to have an example situation with a detailed explanation.*".

Summary and conclusion

This work provides a comprehensive demonstration of applying explainable artificial intelligence tools for interpretable dropout prediction. Since in this paper, we collected the applicable XAI tools, highlighted their pitfalls, and compared them, this work can serve as a tutorial. Using the modern tools of interpretable machine learning, we showed how the predictions of a black-box model can be interpreted both locally and globally.

Global interpretations can shed light on the factors that have the highest impact on academic performance. Correlation analysis (Fig. 1), permutation importance (Table 3), and SHAP importance (Fig. 3) all showed that the feature with the highest predictive power on the final academic status is the high school grade point average (HSGPA), which measures general knowledge since it takes into account the grades in history, mathematics, Hungarian language & literature, a foreign language, and a natural science subject. However, the fact that mathematics and the chosen subject are also among the top important variables suggests that program-specific knowledge is not negligible and it complements general knowledge. We have also found that students are more likely to drop out if they do not start their university studies right after graduation from high school. Using a partial dependence plot, we showed that scores in humanities have incremental predictive power even though this analysis is built on the data of a technical university.

The novelty of our approach is that we not only estimate the probability of graduation and provide global interpretation, but we also explain the individual predictions. That makes this approach applicable in various domains, in other words, it can serve as a useful decision-support tool for all stakeholders. For example, the predictions may assist high school students to choose the appropriate major in university and provide career guidance by predicting in which field they can be the most successful based on their high school performance indicators. Moreover, the local explanation of the output can also help both high school and university students in identifying the skills that need to be improved to succeed in their university studies. The presented tools may support secondary school policy-makers as well to make the transition from high school to university smoother. Furthermore, it is also helpful for higher education decision-makers to choose the right action plan in terms of offering personalized tutoring and remedial courses for at-risk students. For example, if a student is identified as someone who is at risk of dropping out, the tutor can also take a look at the local interpretations provided by SHAP. After identifying the factors that contribute negatively to the prediction (e.g. lack of math skills), the tutor (or an automatized software) can act accordingly by recommending mathematical remediation.

Finally, an important contribution of our work, is that we conveyed a survey with high school students, university students, and higher education decision-makers and managers to assess and evaluate whether they indeed find the applied XAI tools interpretable and useful. The results of the survey show that the SHAP summary plot (Fig. 3) compresses so much information in one single chart that it makes it difficult to understand it at first glance, and it loses its purpose. On the other hand, the interaction plots were the easiest to understand, and they do not require further explanations. In the case of the local explanation, both the students and the decision-makers agree that while the bar charts (left side of Fig. 9 and similarly the LIME plots in Fig. 7) are easier to interpret, the force plots have their advantages as well, namely, that they show the final result, i.e., how the effects of the features accumulate and what is the predicted probability of graduation. We propose alternative visualizations to these problems, such as separate feature effects instead of the summary plot, and the usage of waterfall plots instead of the force plots, however, we believe that finding the best solutions requires further research and more user studies.

Note that while an AI-based decision support system can be an effective tool to increase the graduation rate in higher education, it is essential to note that applying predictive analytics to human beings can cause ethical dilemmas. In our educational setting, considering final academic success prediction, the ethical issue lies in the usage and communication of the output of the algorithm, i.e., the “probability” of graduation. Clearly, we cannot inform the freshman students about their predicted final academic performance, since telling someone that they are going to drop out with a probability of 0.6 at the beginning of their studies would be incredibly harsh and could backfire by leading to self-fulfilling prophecies and hence increase the dropout rate. Moreover, we also drew attention to the fact that the output of the machine learning models usually cannot be interpreted as probabilities, and while there are some ways to calibrate the outputs, it is still an open question of how to communicate the predicted university success. Therefore, we rather suggest focusing on the outputs of the local explanations as feedback which indicates what skills have to be improved.

A limitation of this study is that it solely builds upon the data of the Budapest University of Technology and Economics, where most of the students are enrolled in STEM programs, hence our future research plan is to carry out our analyses on the data of other higher education institutes. Moreover, it is important to note that while student drop-out is a thousand-factor problem, we only have a few, mostly high school performance-related measurements. Other variables have also been associated with student success such as socio-economic status (Freitas et al., 2020; Zwick & Himelfarb, 2011), on-campus living (Zeleny et al., 2021) and various psychological factors (Séllei et al., 2021).

Although we present the interpretable dropout prediction in the Hungarian context, all of the approaches and methodologies proposed in this paper can be applied in any other educational environment all over the world if pre-enrollment achievement measures and other features are available for dropout prediction.

Acknowledgements We are extremely grateful to Bálint Csabay and Mihály Szabó, the colleagues of the Central Academic Office of the Budapest University of Technology and Economics, who helped us with data collection and data understanding.

Authors contributions The authors confirm their contribution to the paper as follows: study conception and design: R. Molontay, M. Nagy; data preparation: M. Nagy; data analysis and visualization: M. Nagy; interpretation of the results: R. Molontay, M. Nagy; draft manuscript preparation: M. Nagy; revision of the manuscript: R. Molontay. All authors approved the final version of the manuscript.

Funding Open access funding provided by Budapest University of Technology and Economics. The research presented in this paper, carried out at BME, was supported by the European Union project RRF-2.3.1–21-2022–00004 within the framework of the Artificial Intelligence National Laboratory. Marcell Nagy has been also supported by the doctoral student scholarship program of the Cooperative Doctoral Programme of the National Research, Development, and Innovation Fund. Roland Molontay is also supported by the National Research, Development, and Innovation Fund through the OTKA Grant PD-142585.

Data availability The data used in this study is sensitive, hence it is not allowed to make them publicly available.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1), 1–21.

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: a systematic literature review. *Online Learning, 20*(2), 13–29.

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. *Proceedings of the 21st Annual Conference on Information Technology Education* (pp. 13–19)

Baranyi, M., & Molontay, R. (2021). Comparing the effectiveness of two remedial mathematics courses using modern regression discontinuity techniques. *Interactive Learning Environments, 29*(2), 247–269.

Behr, A., Giese, M., Theune, K., et al. (2020). Early prediction of university dropouts – a random forest approach. *Jahrbücher Für Nationalökonomie Und Statistik, 240*(6), 743–789.

Cano, A., & Leonard, J. D. (2019). Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies, 12*(2), 198–211.

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: the beneficial impact of the logit leaf model. *Decision Support Systems, 135*, 113325. https://doi.org/10.1016/j.dss.2020.113325

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access, 5*, 15991–16005.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

Freitas, FAd. S., Vasconcelos, F. F., Peixoto, S. A., Hassan, M. M., Dewan, M., Albuquerque, V. HCd., et al. (2020). IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics, 9*(10), 1613.

Greenwell, B. M., Boehmke, B. C., McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. arXiv preprint arXiv:180504755.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. https://doi.org/10.48550/arXiv.2207.08815

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120.

He, J., Bailey, J., Rubinstein, B., & Zhang, R. (2015). Identifying at-risk students in massive open online courses. *Proceedings of the AAAI Conference on Artificial Intelligence, vol 29 no 1.*

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., & Murray, D. J. (2019). Identifying key factors of student academic performance by subgroup discovery. *Int J Data Sci Anal, 7*(3), 227–245.

Jin, W., Fan, J., Gromala, D., Pasquier, P., Li, X., Hamarneh, G. (2022). Transcending XAI algorithm boundaries through end-user-inspired design. arXiv preprint arXiv:220808739.

Karimi, A. H., Barthe, G., Balle, B., & Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. *International Conference on Artificial Intelligence and Statistics, PMLR* (pp. 895–905)

Karlos, S., Kostopoulos, G., & Kotsiantis, S. (2020). Predicting and interpreting students' grades in distance higher education through a semi-regression method. *Applied Sciences, 10*(23), 8413.

Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: how are rubric scores determined? *International Journal of Artificial Intelligence in Education, 31*(3), 538–584.

Latif, A., Choudhary, A. I., & Hammayun, A. A. (2015). Economic effects of student dropouts: a comparative study. *Journal of Global Economics, 3*(137), 2.

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 3093.

Looveren, A. V., & Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 650–665). Springer.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates Inc.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33*(1), 107–124.

Mingyu, Z., Sutong, W., Yanzhang, W., & Dujuan, W. (2021). An interpretable prediction method for university student academic crisis warning. *Complex & Intelligent Systems, 8*, 1–14.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... & Bischl, B. (2022, April). General pitfalls of model-agnostic interpretation methods for machine learning models. In: AI-Beyond Explainable AI: *International Workshop, Held in Conjunction with ICML 2020, July 18, 2020* (pp. 39–68). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-04083-2_4

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/.

Molontay, R., & Nagy, M. (2022). How to improve the predictive validity of a composite admission score? a case study from hungary. *Assessment & Evaluation in Higher Education* (pp. 1–19)

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617)

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017). MOOC Dropout Prediction: Lessons learned from making pipelines interpretable. *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 351–359)

Nagy, M., Molontay, R., & Szabó, M. (2019). A web application for predicting academic performance and identifying the contributing factors. *47th Annual Conference of SEFI* (pp. 1794–1806)

Nagy, M., & Molontay, R. (2021). Comprehensive analysis of the predictive validity of the university entrance score in Hungary. *Assessment & Evaluation in Higher Education* (pp. 1–19)

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning* (pp. 625–632)

Platt, J., et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61–74.

Powell, W. W., & Snellman, K. (2004). The knowledge economy. *The Annual Review of Sociology, 30*, 199–220.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS'18)* (pp. 1–11)

Rastrollo-Guerrero, J. L., Gomez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences, 10*(3), 1042.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016* (pp. 1135–1144)

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS One, 12*(2), e0171207.

Sargsyan, A., Karapetyan, A., Woon, W. L., & Alshamsi, A. (2020). Explainable AI as a Social Microscope: A Case Study on Academic Performance. *International Conference on Machine Learning, Optimization, and Data Science* (pp. 257–268). Springer.

Schneider, J., Richner, R., & Riser, M. (2022). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education, 33*, 88–118.

Séllei, B., Stumphauser, N., & Molontay, R. (2021). Traits versus grades—the incremental predictive power of positive psychological factors over pre-enrollment achievement measures on academic performance. *Applied Sciences, 11*(4), 1744.

Shaffer, J. (2016). *5 tips on designing colorblind-friendly visualizations*. https://www.tableau.com/about/blog/examining-data-viz-rules-dont-use-red-green-together. accessed: 2022–09–27.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Inform Fusion, 81*, 84–90.

Smith, B. I., Chimedza, C., & Bührmann, J. H. (2021). Individualized help for at-risk students using model-agnostic and counterfactual explanations. *Education and Information Technologies, 27*, 1539–1558.

Varga, E. B., & Sátán, Á. (2021). Detecting at-risk students on computer science bachelor programs based on pre-enrollment characteristics. *Hungarian Educational Research Journal, 3*(11), 297–310.

Vultureanu- Albişi, A., & Bădică, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. *IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE* (pp. 215–220)

Yu, H., Miao, C., Leung, C., & White, T. J. (2017). Towards AI-powered personalization in MOOC learning. *npj Science of Learning, 2*(1), 1–5.

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? *Proceedings of the eighth ACM conference on learning@ scale* (pp. 91–100)

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27.

Zeleny, K., Molontay, R., & Szabó, M. (2021). A kollégiumi lét egyetemi teljesítményre gyakorolt hatásának vizsgálata. *Statisztikai Szemle, 99*(1), 46–79.

Zhang, W., Zhou, Y., & Yi, B. (2019). An interpretable online learner's performance prediction model based on learning analytics. *Proceedings of the 2019 11th International Conference on Education Technology and Computers* (pp. 148–154)

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of sat scores and high school grade-point average. *Journal of Educational Measurement, 48*(2), 101–121.