Predicting Academic Performance and Identifying Contributing Factors through a Web Application

Dr. Emma Rodriguez, Dr. Liam Chen, and Dr. Ava Patel

Dr. Emma Rodriguez and Dr. Liam Chen, Department of Data Science, University of California, Los Angeles (UCLA), Los Angeles, CA, USA;

ABSTRACT

Data-driven approaches have received a lot of attention recently from higher education researchers and policy-makers as well. At the Budapest University of Technology and Economics, we aim to extract insight from big data stored in the educational administration system in the framework of a project carried out in the cooperation of Central Academic Office and Institute of Mathematics. Among many other questions, we studied curriculum prerequisite networks with a student flow approach, the effect of mathematical remediation, the impact of living on-campus on academic achievement, the connection between grade inflation and student evaluation of teaching and efficient visualization of student flows. However, one of the most burning problems in higher STEM education all over the world is dropping out. In this paper, we present a predictive analytical approach for early detection of students at-risk of academic failure. We achieved relatively high accuracy, compared to the results of related works, which makes it suitable to deploy as a decision support system. We also present a web application that is able to identify at-risk students mainly based on their high school grades and matura scores using machine learning algorithms (e.g. XGBoost). The application can also be used to recommend tutoring sessions and remedial courses for at-risk students. Based on SHAP values, the application is also capable of making suggestions for students which skills to improve in order to succeed in their university studies. Besides Hungarian education system, our proposed methodology is also applicable to other educational environments all over the world.

1 INTRODUCTION

Data-driven approaches have been extensively used in a number of scientific fields, including educational research. The big data stored in educational administrative systems hold great potential for data-driven educational research. Due to this, new scientific fields have emerged such as educational data mining and learning analytics, for systematic reviews we refer to [1, 2].

At the Budapest University of Technology and Economics, we also initiated a project in the cooperation of the Central Academic Office and the Institute of Mathematics with the objective to extract knowledge from the massive educational data of the university. The fruitful cooperation has resulted in a number of publications and the development of decision support applications to help policy-makers and other stakeholders.

We have introduced a data-driven probabilistic student flow approach to characterize curriculum prerequisite networks which can be used to identify courses that have a

huge impact on the graduation time [3]. Our approach is also capable of simulating the effects of policy changes and modifications of the prerequisite network [4]. We also developed an efficient visualization tool to analyze student flow patterns by alluvial and Sankey diagrams that allows decision-makers to gain a better insight on how students are processing and it makes easier to understand the effects of policy changes on retention and graduation rates [5]. We introduced a novel approach for ranking secondary schools based on their students' later university performance [6]. Furthermore, in [7] we measured the direct and longer term effects of mathematical remediation on academic achievement using a regression discontinuity approach. The impact of living on-campus on academic performance was also investigated [8]. Moreover, we studied the connection between grade inflation and student evaluation of teaching. As a first study of this nature from Central Europe, in accordance with other studies, we found that increasing the grade of a student by 1, will lead to approximately 0.25 higher evaluations for the instructor [9, 10]. Furthermore, we analyzed the predictive power of the Hungarian nationally standardized admission point score and its variants on academic performance [11].

Predicting students' academic performance is a challenging task of great importance. In particular, predicting dropouts and early detection of at-risk students have attracted a lot of research interest [12], since dropping out is associated with considerable personal and social costs [13] and it is regarded as one of the most burning problems in higher STEM education all over the world. Machine learning algorithms have been applied in many studies to predict dropout risk and academic achievement measures and to discover the important factors affecting student performance [14, 15]. Early detection of at-risk students allows institutions to offer more proactive personal guidance, remedial courses and tutoring sessions in order to mitigate academic failure.

The first two authors of the present study also employed several machine learning algorithms (e.g. neural networks and gradient boosting trees) to predict student dropout mainly based on secondary school performance [16]. The present study extends [16] in several directions, most importantly by extracting insights from the machine learning models, namely identifying the most important features and analyzing the effect of each feature on the prediction. A key issue and one of the hottest topics in machine learning nowadays is model interpretability, especially if there are implications associated with the model's prediction [17]. Interpreting the results also highly assists students, policy-makers, and other stakeholders since it sheds light on factors affecting academic performance and being an "at-risk student". For model interpretation, we use cutting-edge techniques such as permutation importance [18] and SHAP (SHapley Additive exPlanations) values [19]. Another contribution of this work is that we developed a novel web-application by adding several new features to enrich user experience.

2 BACKGROUND ON HUNGARIAN HIGHER EDUCATION

In this section, we summarize some characteristics of the Hungarian education system with an emphasis on the admission procedure to higher education institutions,

a more detailed overview can be found in [16]. Hungarian education consists of 8 years of primary education, 4 (or sometimes 5) years of secondary education and optionally of higher education standardized by the Bologna Process. A five-point grading scale is used where (5) corresponds to excellent and (1) is the failing grade.

The nationally standardized higher education admission system defines the Admission Point Score (APS) that is mostly based on the secondary school performance and on the score of the centralized exit exam, called matura. Undergraduate programs rank students according to their APS, while students also have a preferential ranking of the desired programs. The admission procedure is governed by the student-optimal matching algorithm of Gale and Shapely.

Due to the admission system, universities have a lot of data regarding prior academic performance of incoming students such as grades in the core subjects (mathematics, Hungarian language and literature, history, a chosen foreign language, a chosen science subject) from the last two academic years, the level (normal or advanced) and scores of matura exams, foreign language certificates etc. The aforementioned results indicate a broad spectrum of knowledge and skills. In this study, we investigate how future academic performance, particularly dropping out can be predicted based on the prior performance and what the most important contributing factors are.

3 DATA DESCRIPTION

The present study is based on the data of 6,774 students enrolled in one of the undergraduate programs of Budapest University of Technology and Economics between 2013 and 2018. The available attributes are presented in *Table 1*.

Table 1. Overview of the data fields

Feature class	Feature name	Type
University program related	Student ID	Nominal
	Program ID	Nominal
	Financing source	Binary (state-funded or self-funded)
	Re-enrolling	Binary (True or False)
Label	Final status	Binary (Graduated or dropped out)
High school performance related		
Matura exam scores	Elective subject	Numeric
	Foreign language	Numeric
	History	Numeric
	Hungarian language and literature	Numeric

	Mathematics	Numeric
Average of grades	Foreign language	Numeric
	History	Numeric
	Hungarian language	Numeric
	Hungarian literature	Numeric
	Mathematics	Numeric
	Science subject	Numeric
	Score	
Certificate of foreign language(s)	(based on the number and level of certificates)	Numeric
Personal details		
	Gender	Binary (female or male)
	Years between the dates of matura and enrollment	Numeric
	Location of high school	Categorical (capital city, city with county rights, foreign city and other)

After the data were collected in an anonymized way from the administrative educational system, various data preparation steps were conducted. The data preprocessing and cleaning tasks include pivoting and merging the data sheets, handling missing data, attribute transformation and dimension reduction. For handling missing data, we used Multiple Imputation by Chained Equations (MICE) with Bayesian Ridge Regression.

In order to reduce the high complexity of study related attributes, we carried out several data transformation steps such as merging and combining attributes. For example, by combining the point (0-100) and the level (normal or advanced) of the matura exams into one attribute (by multiplying the score with 1.5 for advanced level exams), we reduced the number of matura-related attributes by a factor of 2. We also merged the results of several extremely similar subjects into a smaller set of more meaningful attributes.

For the sake of simplicity, we define the final status to be binary by omitting the students who are still active in their studies and we consider students as being graduated if they completed the required studies even if they did not obtain a degree certificate (e.g. due to the lack of the required language certificates). It is also important to note that as opposed to other similar studies, we do not face the problem of imbalanced class distribution due to the high dropout rate at our university and due to

the fact that students who have started their university studies after 2016 could not graduate yet, meaning that from this student cohort we have more data of dropped out students. Overall, in this study we analyze the data of 3,003 graduated and 3,411 dropped out students.

There are university programs that were launched or terminated during the examined period, we also omit these programs from the data. The aforementioned steps both reduce computational complexity and improve the interpretability of the results. After data pre-processing, we obtained 19 attributes of 6,414 students.

4 METHODOLOGY

The aim of this paper is to predict whether an incoming student will graduate or drop out based on data available at the time of enrollment (prior academic performance, personal details). Since we can rely on historical data where the actual final status is known, this problem is a supervised learning task. More precisely, it is a binary classification task where the label is the final status and the explanatory variables are detailed in *Table 1*.

Various techniques have been proposed for solving binary classification tasks, in our earlier work we also compared several methods for dropout prediction and gradient boosting trees (GBT) turned out to be the best performing model for a restricted number of attributes [16]. In the present study, we also use an advanced implementation of gradient boosted decision trees, called eXtreme Gradient Boosting (XGBoost) [20] that is considered to be the state-of-the-art machine learning algorithm for structured data with many advantageous characteristics such as parallelized tree building, regularization for avoiding overfitting and efficient handling of missing data. For fine-tuning the hyperparameters of the model, we used grid optimization with cross-validation.

For the final evaluation of the model we use ROC/AUC (receiver operating characteristic/area under ROC curve) analysis together with the accuracy, precision and recall measures on stratified test samples (random samples, such that the class distribution in the subsets is the same as in the whole dataset).

For model interpretation, we use two additional techniques: permutation importance for global and SHAP (SHapley Additive exPlanations) for local interpretation. Permutation importance measures how the accuracy of the prediction decreases if a single feature of the validation data is randomly shuffled (mimicking the effect of removing that feature) [18]. SHAP provides local explanations for the output of any machine learning models based on game theoretical concepts [19]. For a particular prediction, it measures how each feature contributes to push the model output from the base value (the average output over the training data) to the output value, thus it directly shows the strengths and weaknesses of a particular student and helps to identify his/her possible skill gap.

5 WEB APPLICATION

The web application was developed in Python 3 using Dash web framework. Based on the user input (the input fields are shown in Fig. 1), the web application returns a prediction whether the student will graduate or drop out together with a probability score of graduation. Moreover, the application also helps the user to interpret the results by showing what contribution each feature has on the prediction, i.e., by the visualization of SHAP values, see *Fig. 2*.

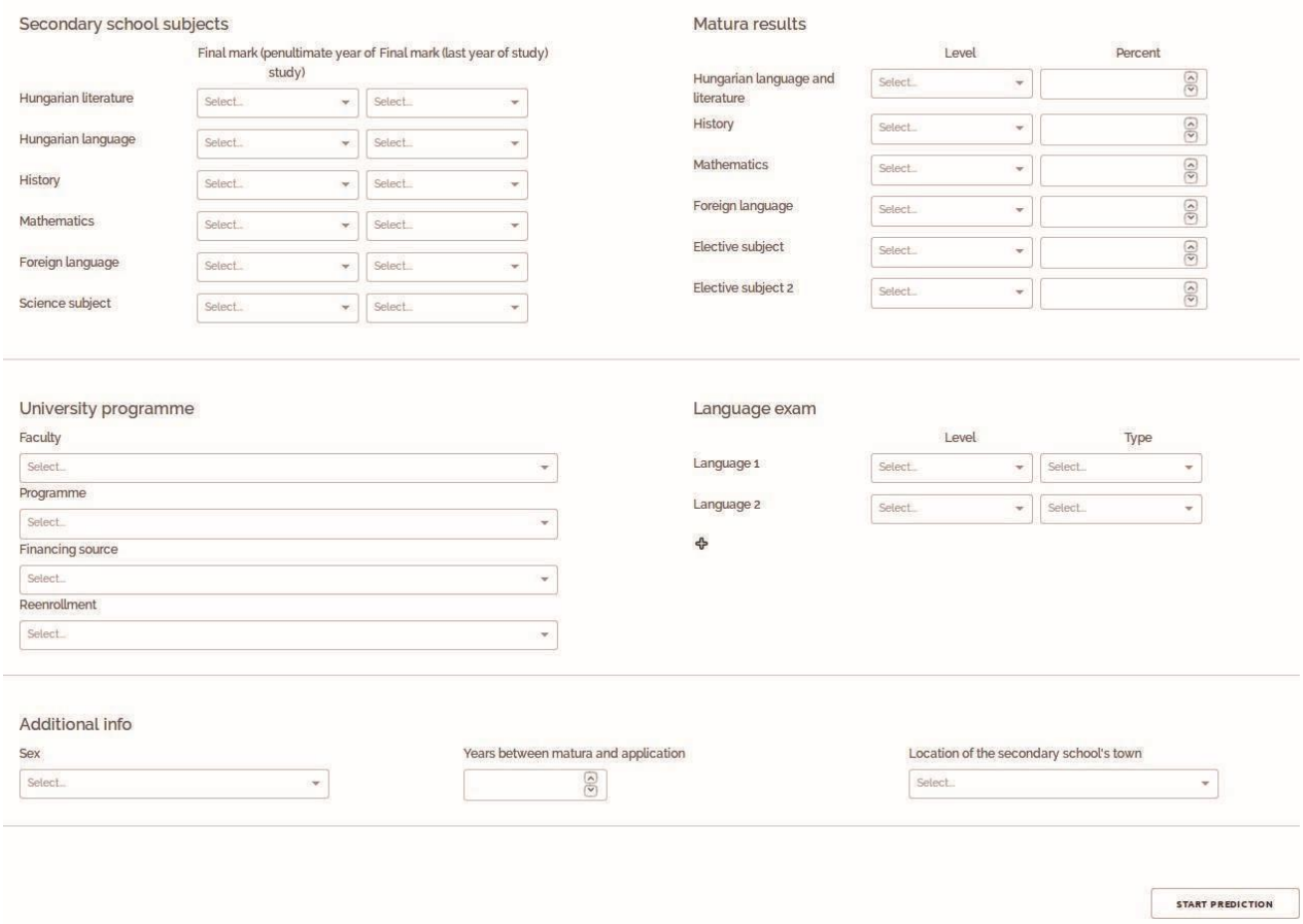


Fig. 1. The user input interface of the web application

In *Fig. 2* we can observe a strong student (bottom row), a borderline student (middle row) and an at-risk student (top row). The output values (probability of graduation) is highlighted in each case together with the interpretation of the prediction. Regarding the at-risk student, we can observe that the most endangering factor is the fact that on the mathematics matura exam (s)he obtained just 62%, followed by the fact the average grades in his/her elected science subject is 3.5 and the average grades in history is just 2. On the other hand, the fact that (s)he enrolled in university right after the matura exam contributes positively to the prediction.

As for the borderline student, his/her good matura scores push the prediction higher, while the fact that two years passed between the matura exam and the university enrollment together with his/her satisfactory (3) high school grades in Hungarian language push the prediction lower. Finally, regarding the strong student, his/her

excellent matura scores in mathematics and in his/her elected subject contribute the most to the positive prediction.

The presented results may help the students to understand their strengths and weaknesses and to choose the right action plan. Moreover, it is also advantageous for university policy-makers to understand what characteristics make a student more likely to drop out.

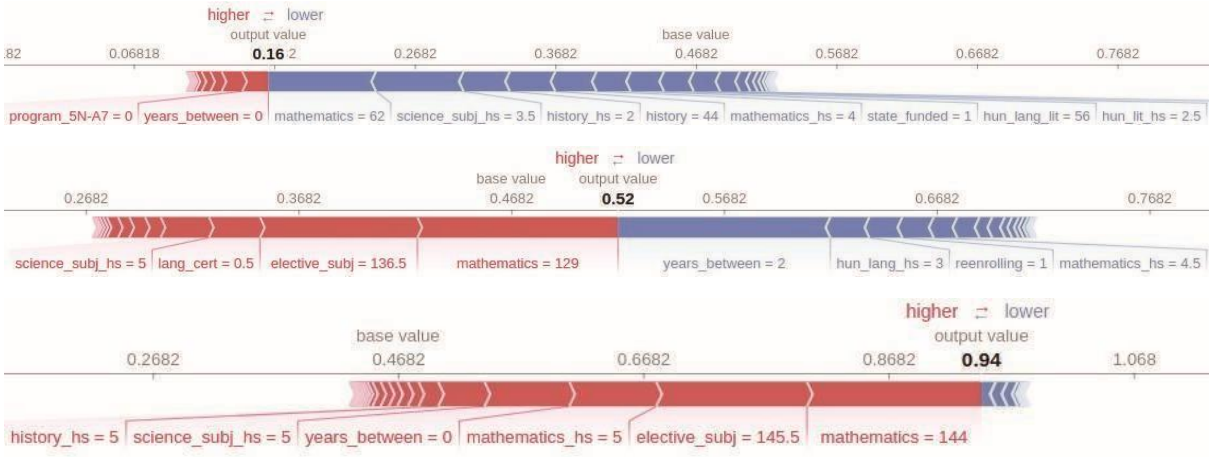


Fig. 2. Prediction explanation for three students of the test dataset. Features pushing the graduation prediction higher are shown in red and those pushing the prediction lower are in blue.

For assessing the utility and clarity of the web application, we presented it for a body of university management and we have already started to test the tool with high school students as well. The overwhelming majority approved the application and found it useful and easy to understand with high potential for further development as it answers the demands of the university to better understand the characteristics of at-risk students and it also assists students in identifying their possible skill gaps.

6 RESULTS

In this section, we present the results of the predictive model together with the global importance of the features and the local interpretation of the prediction.

6.1. Model performance

The (ROC) curves of the models are presented in *Fig. 3*. One can observe that XGBoost algorithm on the original (incomplete, without imputation) data set has the best performance with AUC=0.772, XGBoost and Random Forest algorithms also perform quite well on the imputed data set. The results suggest that the internal treatment of missing values in XGBoost is more efficient than the used imputation method (MICE). XGBoost on the original data set has accuracy of 71.4%, precision of 71.8%, recall of 76.2% and F-score of 73.9%. To compare our results to other related studies, this is a remarkable performance considering the fact that the data are heterogeneous and the prediction only relies on data available at the time of enrollment

[16]. Moreover, this performance makes it possible to build a relatively reliable decision support system on this model.

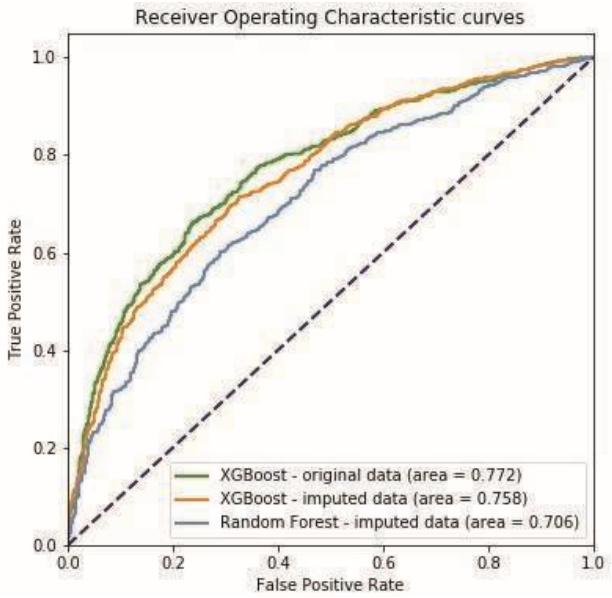
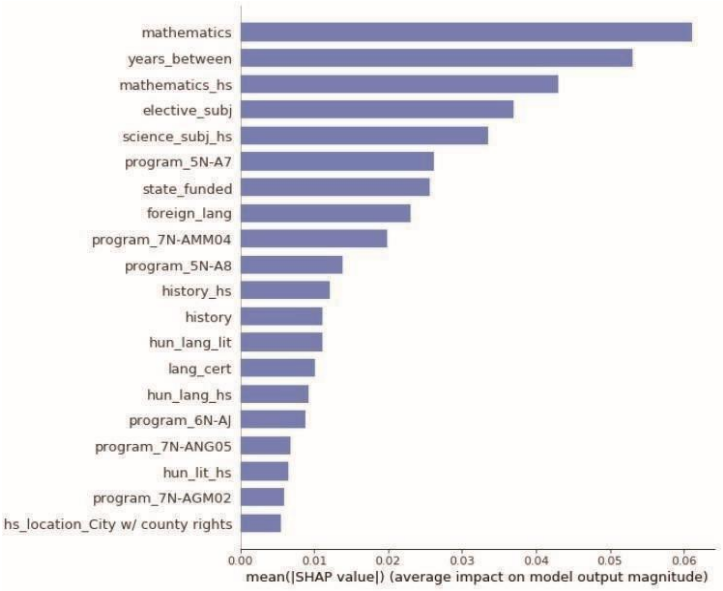


Fig. 3. The ROC curves derived from the test data

6.2. Model interpretation

We have already demonstrated the explanatory ability that SHAP values have regarding particular predictions (see *Fig. 2*). Here we analyze the global importances of the features aggregated from the SHAP values of individual predictions.

Fig. 4 shows the global feature importances based on the SHAP values and on permutation importance method. We can observe that the results are consistent, both methods find mathematics matura score the most important contributing factor on separating at-risk students, followed by the number of years between the matura exam and the university enrollment.



Weight	Feature
0.0526 ± 0.0228	mathematics
0.0349 ± 0.0149	years_between
0.0171 ± 0.0047	elective_subj
0.0120 ± 0.0049	program_7N-AMM04
0.0116 ± 0.0095	state_funded
0.0086 ± 0.0066	foreign_lang
0.0085 ± 0.0095	program_5N-A7
0.0056 ± 0.0071	mathematics_hs
0.0034 ± 0.0019	program_3N-A0
0.0031 ± 0.0087	program_5N-A8
0.0030 ± 0.0025	program_4N-A5
0.0030 ± 0.0015	program_6N-AJ
0.0025 ± 0.0022	program_1N-AEM
0.0017 ± 0.0031	program_2N-AT0
0.0016 ± 0.0017	program_2N-AE0
0.0015 ± 0.0010	hs_location_Captial city

Fig. 4. Feature importances. On the left: the average absolute SHAP values of the attributes are shown. The figure on the right shows the outcomes of the permutation importance method.

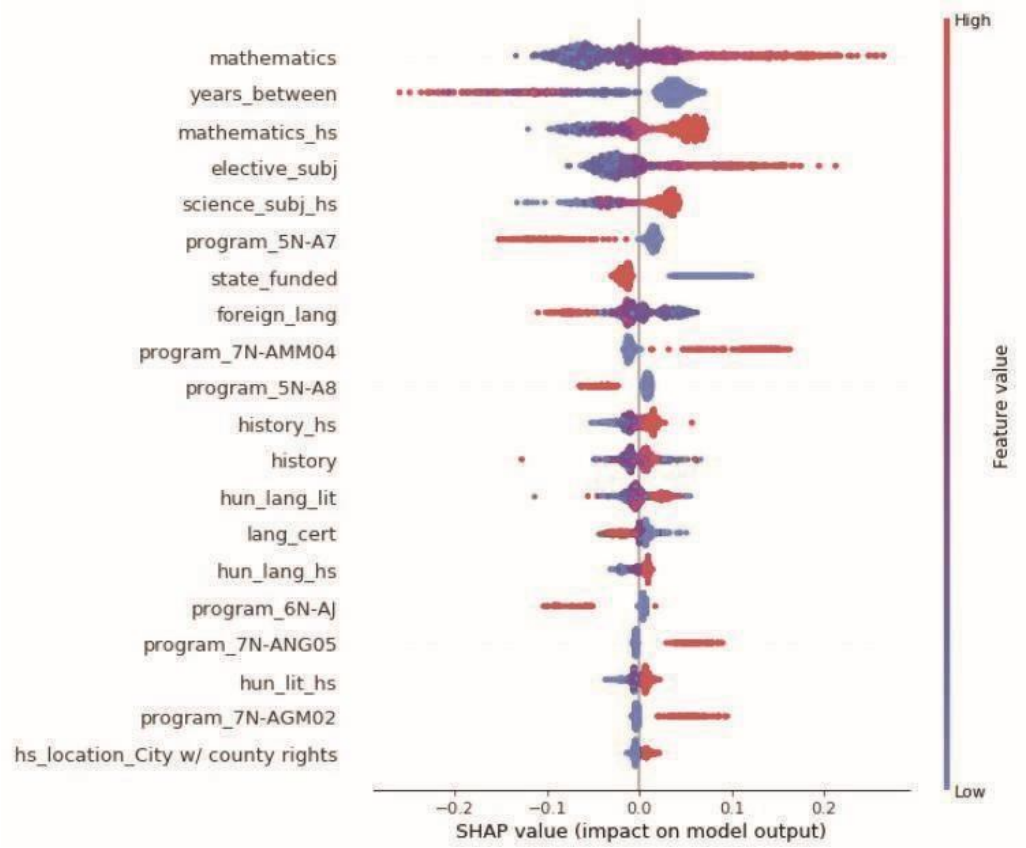


Fig. 5. Summary of the effects of the most important features. Every student (of the test data) has one dot in each row. The *x* position of the dot shows the impact of that feature on the prediction, and its color represents the value of that feature for the student. Dots that do not fit on the row pile up to show density.

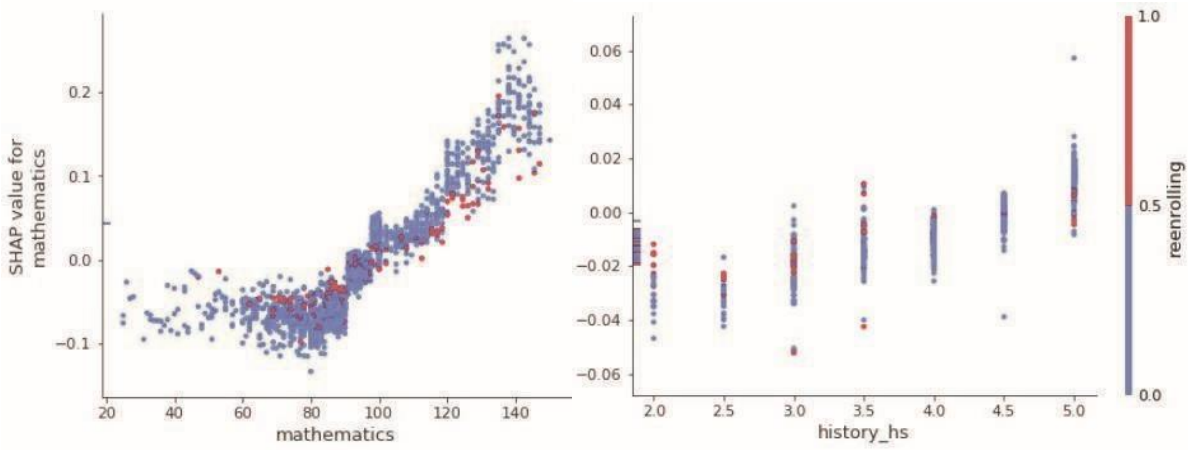


Fig. 6. SHAP dependence plots of mathematics exam scores and average high school grades in history. Each student is represented by a point on each figure, the *x* positions are the marks in the subjects, the colors indicate whether the student is re-enrolled and the *y*-axis shows how and to what extent the subjects influence the prediction.

Fig. 5 illustrates the distribution of the SHAP values for each feature. Here again we see that mathematics matura scores have the highest impact on the probability of graduation. We can also observe important outlier effects, e.g. the financing source (state_funded=0 if fee-paying and 1 otherwise) attribute is not so important globally, but it is an influential feature for the fee-paying students, since they graduate with higher probability. The reason behind the importance of university programs is the fact that graduation rates vary across the programs.

From *Fig. 6* we can observe that higher scores in mathematics or in history imply higher positive effect on a typical prediction. However, in the low range of mathematics scores, the effect seems to be constant.

7 SUMMARY

Predicting academic failure, early-detection of at-risk students and identifying the key contributing factors have ever-growing importance, in particular in STEM higher education where dropping out of students is a serious issue all over the globe. In this paper, we presented a machine-learning based decision support system to assist both students and decision makers. Our web application may assist secondary school students to choose the appropriate major in university based on their skill sets or help them identifying the skills that need to be improved in order to succeed in their university studies. The application is also helpful for higher-education decision-makers to choose the right action plan in terms of offering tutoring and remedial courses for at-risk students. Moreover, our findings may support secondary school policymakers as well to make the transfer from high school to university smoother.

Although our proposed methodology has been developed for the Hungarian education system, this approach is also applicable to other educational environments all over the world. Especially, if standardized pre-enrollment achievement measures are available (e.g. nationally standardized test scores such as the SAT and ACT) to predict college success.

REFERENCES

[1] Leitner, P., Khalil, M., Ebner, M. (2017), Learning analytics in higher education—a literature review, *Fundaments, applications, and trends*, pp. 1-23.

[2] Dutt, A., Ismail, M. A., Herawan, T. (2017), A systematic review on educational data mining, Fundaments, applications, and trends, *IEEE Access*, Vol. 5, pp. 15991-16005.

[3] Bergman, J., Molontay, R., Szabó, M., Szekrényes R. (2019), Kreditrendszerű képzések mintatanterveinek és előtanulmányi hálóinak elemzése a hazai matematika alapszakok példáján, *Alkalmazott Matematikai Lapok (to appear)*, in Hungarian.

[4] Bergman, J., Horváth N., Molontay, R., Szabó, M., Szekrényes R. (2019), Characterizing Curriculum Prerequisite Networks by a Student Flow Approach (under review).

[5] Horváth, D. M., Molontay, R., Szabó, M (2018), Visualizing student flows to track retention and graduation rate, Proc. of the 22nd IEEE International Conference of Information Visualisation (IV), Salerno, pp. 338-343.

[6] Horváth, N., Molontay, R., Szabó, M (2018), Who are the most important “suppliers” for universities? - Ranking secondary schools based on their students’ university performance, 2nd Danube Conference for Higher Education Management, Budapest

[7] Baranyi, M., Molontay, R. (2019), Effect of mathematics remediation on academic achievement -- a regression discontinuity approach, Proc. of the 5th IEEE International Symposium on Educational Technology (ISET), Hradec Kralove

[8] Zeleny, K. (2019), A kollégiumi lét hatásának vizsgálata az egyetemi teljesítményre, XXXIV. Országos Tudományos Diákköri Konferencia (OTDK), Gödöllő, in Hungarian

[9] Lukáts, G. D. (2019), The Effect of Grade Inflation on Student Evaluations of Teaching XXXIV. Országos Tudományos Diákköri Konferencia (OTDK), Pécs

[10] Berezvai, Z., Lukáts, G. D., Molontay, R. (2019), Hogyan hatnak a pénzügyi ösztönzők az egyetemi oktatók osztályozási gyakorlatára? Egy természetes kísérlet eredményei, *Közgazdasági Szemle (to appear)*, in Hungarian.

[11] Nagy M., Molontay, R. (2018), Predictive power of admission point score and its variants on academic performance, 2nd Danube Conference for Higher Education Management, Budapest

[12] 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*, Vol. 33, No. 1. pp. 107-124.

[13] Latif, A., Choudhary, A. I., Hammayun, A. A. (2015), Early dropout prediction using data mining: a case study with high school students, *Journal of Global Economics,* Vol 3

[14] Rovira, S., Puertas, E., Igual, L. (2017), Data-driven system to predict academic grades and dropout, *PLoS one*, Vol 12, No. 2, e0171207

[15] Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J. (2019), Identifying key factors of student academic performance by subgroup discovery, *International Journal of Data Science and Analytics*, Vol 7, No. 3, pp.
227-245.

[16] Nagy, M., Molontay, R. (2018), Predicting Dropout in Higher Education based on Secondary School Performance, Proc. of the 22nd IEEE International Conference on Intelligent Engineering Systems (INES), Las Palmas.

[17] Molnar, C. (2018). Interpretable machine learning: A guide for making black box models explainable. Christoph Molnar, Leanpub.

[18] Altmann, A., Toloşi, L., Sander, O., Lengauer, T. (2010), Permutation importance: a corrected feature importance measure, *Bioinformatics*, Vol 26, No. 10, pp. 1340-1347.

[19] Lundberg, S. M., Lee, S. I. (2017), A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, pp. 47654774.

[20] Chen, T., & Guestrin, C. (2016), XGBoost: A scalable tree boosting system, Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, pp. 785-794.