Predicting Dropout Using High School and First-semester Academic Achievement Measures

Botond Kiss*, Marcell Nagy†, Roland Molontay‡ and Bálint Csabay§
*†‡Department of Stochastics, Budapest University of Technology and Economics, Budapest, Hungary
†‡Faculty of Informatics, University of Debrecen, Debrecen, Hungary
‡MTA–BME Stochastics Research Group, Budapest, Hungary
§Central Academic Office, Budapest University of Technology and Economics, Budapest, Hungary
*bkiss@math.bme.hu, †marcessz@math.bme.hu, ‡molontay@math.bme.hu, §csabay.balint@kth.bme.hu

***Abstract*—Due to the big data accumulated in educational administrative systems and due to the advance of machine learning techniques, a new scientific discipline has emerged in the last few years, namely educational data science. An important research objective of this field is to predict dropout and improve graduation rates, in particular in STEM higher education.**

The goal of this study is to identify students at risk of dropping out at a large Hungarian technical university using predictive analytical tools. We use data of 10,196 students who finished their undergraduate studies (either by graduation or dropping out) between 2013 and 2018. We analyze dropout predictability in two main scenarios: first using data available at the time of enrollment, e.g. pre-enrollment achievement measures, and personal details, then in another scenario we supplement this feature set with first-semester performance indicators and use this richer set of attributes for further analysis. We apply artificial neural networks and boosting algorithms for prediction, and examine how the predictive power can be improved by the additional information. In other words, we study the incremental predictive validity of the early university performance indicators on graduation over the pre-enrollment achievement measures and vice versa.

***Index Terms*—higher education, dropout prediction, incremental predictive power, academic performance, at-risk students**

I. Introduction

Preventing student dropout in tertiary education has become of great interest across the globe due to economic and personal costs associated with dropping out and delayed completion. The issue of study success is on the policy agenda in most of the European countries [1]. In particular, in Hungary where the proportion of graduates is one of the lowest among OECD[1] countries [2].

The prompt development of artificial intelligence, machine learning, and data-driven methods, moreover the educational data accumulated by universities have made it possible to help students and decision-makers address the problem of early school leaving. In the past years, a new scientific discipline has emerged, called educational data mining (EDM), with the objective of tackling issues related to student attrition. Most common research goals include predicting course grade and GPA [3], [4].

According to a recent comprehensive review [4], which investigates 357 EDM-related publications, the most frequently analyzed educational data are that of students studying in STEM fields (Science, Technology, Engineering, and Mathematics), mainly because in these fields student drop-out rates are higher compared to other disciplines. More systematic reviews have been published on research trends and the most frequently used algorithms and methodologies in EDM [5]–[7], one specialized solely in summarizing the use of cluster analysis [8].

Some research intend to tackle non-conventional problems using high volume data, such as investigating the direct and long-term effects of remediation courses [9], determining the best secondary schools based on their students' later performance in higher education [10], and answering other educationally relevant questions, such as how the financial reward of student evaluation of university instructors affects grades [11]. A technically interesting work [12] managed to outperform conventional methods in predicting course grades using matrix factorization on data stored in learning management systems and e-learning data.

Since Budapest University of Technology and Economics (BME) plays a central role in providing STEM human resources to the labor market in Hungary, early detection and avoidance of student attrition via predictive analytic tools are of great importance. Due to the structure of the Hungarian admission system, rich secondary education-related data are available for exploitation at higher education institutes.

In this study, we apply cutting-edge predictive analytical tools to predict the final academic status of students, using the data of 10,196 undergraduate students, who finished their studies at BME between 2013 and 2018 (either by graduation or dropping out). First, we use secondary school related and personal data for predictions, then we examine to what extent additional first-semester performance indicators affect predictive power, and whether pre-enrollment achievement measures have incremental predictive power given first-semester performance data. We also monitor what the key factors are that affect university success.

In the following, we shortly present recent studies that are related to ours. Aulck et al. using a high volume data containing the achievements of 66,060 students show that demographic and pre-admission data have little impact on graduation compared to early university performance indica-

[1] Organisation for Economic Co-operation and Development

978-1-7281-4967-7/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

tors [13]. Radunzel et al. focus on the use of pre-enrollment achievement measures, such as the Admission College Test (ACT) score and high school GPA (HSGPA), and predict academic success (e.g. first-year GPA and degree completion) based on the data of 190,000 students [14]. Furthermore, similarly to the approach of this work, Radunzel et al. supplement HSGPA and ACT scores with the first-year GPA in order to predict long-term university success. They find that the effects of HSGPA and ACT scores on degree completion are small compared to first-year GPA. Marquez et al. use a small set of imbalanced data in order to predict high-school dropout using some special features e.g. that measure personal motivation and smoking habits besides conventional performance measures [15]. In a recent work, Vulperhost et al. analyze the data of two Dutch universities and find that the final university GPA is best predicted by a combination of first-year GPA and high school GPA [16].

Although a great amount of research has been conducted worldwide, in Central Europe, especially in Hungary, fewer EDM-related papers were published. This study closely relates to and builds on two Hungarian predictive analytic studies, which attempt to tackle dropping out [17], [18] at BME. In the aforementioned papers, several machine learning algorithms are trained and tested in order to predict the final status. Most influential factors of early university leaving are investigated as well. These works find that the variables that may measure industriousness (such as high school performance in humanities) have a considerable impact on graduation, even though BME is a technical university. In addition, in [18] a web application is also presented, which predicts final status, and helps in interpreting the most contributing factors of the predictions.

Compared to the aforecited works, the contribution of the present study is that we do not only build our models on the data available at the time of enrollment, but we use students' first semester university performance as well, furthermore, using state-of-the-art machine learning techniques, we also investigate whether pre-enrollment achievement measures and personal details have incremental predictive power on dropout over first semester student achievements and vice versa.

II. Hungarian higher education admission system

We present concisely how the Hungarian higher education admission system works, in order to make it easier to follow the rest of this paper. There are three different types of scores that can be counted towards higher education admission.

1) Matura exam scores:
Secondary education ends with a centralized exit exam called matura exam, which consists of the following subjects: Hungarian language and literature, mathematics, history, chosen foreign language and at least one arbitrary chosen subject, which has been studied by the examinee for at least 2 years. Exams can be taken at normal and advanced level. Each higher education institute indicates two required exam subjects (with level), as a prerequisite for admission to a given undergraduate program. A maximum of 200 matura exam points can be gained.

2) Study scores:
It consists of the average percentage of the five matura exams (100 points) together with two times the sum of high school grades of the four mandatory matura exam subjects plus a chosen natural science subject regarding the last two academic years in which the subjects were studied. This means that at most 200 study points[2] can be earned.

3) Extra points:
Students can earn some extra points e.g. with advanced matura exams (50 points, if the subject is an admission subject), intermediate (28 points) or advanced level (40 points) foreign language certificate, competition achievements and certain disabilities provide extra submission points, which cannot exceed 100 points altogether.

There are two possible methods for summing up these points that yield to the admission point score (APS), which is also referred to as the university entrance score). The central admission system automatically chooses the calculation method that is more advantageous for the candidates applying to a particular program. One is the *doubling* method, which doubles the matura exam scores and adds extra points (in this case study scores do not count), the other is the *summation* method which is simply the sum of the three scores. Minimal admission score for each undergraduate program is declared by the higher education institutions relying on the student-optimal matching algorithm of Gale and Shapely [19] (the minimum score cannot be lower than a state defined minimum), and those who reach this limit gain admission. A more detailed description of the Hungarian admission system can be found in one of the related previous works [17].

III. Data description and preparation

Although high volume data are available, provided by the Central Academic Office, we face the absence of a large amount of data for various and particular reasons. After merging several data frames into one single table, in total 10,196 students' data were chosen with no or moderate amount of missing fields. On the other hand, in contrast to other works [15], [20], in this study, we do not have to deal with imbalanced class distribution, since on the first hand dropout rate is relatively high at BME, but more importantly, students who enrolled after 2015 could not graduate until 2018, meaning that from 2015 to 2018 we only have data about dropped out students.

The attributes of our data set fall into four major categories (for more details see Table I), such as university program related data, high school performance metrics, university performance indicators, and personal information. Note that the reason why the value range of matura exam scores is $[0, 150]$ in Table I, is because we multiplied the advanced level matura exam percentage results by 1.5 in order to reduce the dimensionality of our data.

[2] In Hungarian education grading scales from 1 (insufficient) to 5 (excellent).

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

We intend to give more insights into our first-semester university performance data and the derived features. We also emphasize some findings from a previous related study on data of BME students [17]. Regarding university-related data fields, due to well-documented grades and results in Neptun educational administration system, we do not have to address the problem of missing data.

There are some first semester related attributes – such as living on-campus and mathematical assessment test related attributes – that are already known in the very first week of the semester, so we examine these features with special attention. The basic mathematical knowledge of incoming students (except for the architect students) is tested by a mathematical assessment test in the first week of the first semester. Histograms and kernel estimates of the mathematical assessment test scores of dropped out and graduated students show the potential predictive validity of this feature, see Fig. 1. From the figure, we can conclude that on average graduates achieve higher scores on the mathematics test than dropouts (in case of wrong answers negative score can be achieved).

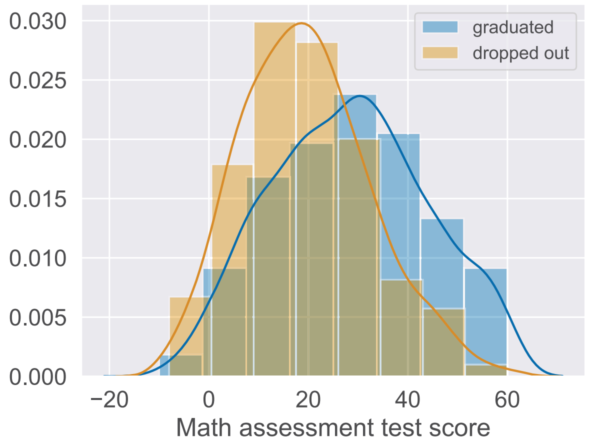

Fig. 1: Histograms and kernel estimates of mathematical assessment test scores of 3,135 students (enrolled in 2013) grouped by the final academic status.

We also use various university credit weighted[3] indices to assess the first semester performance of the students (see Table I).

In what follows, we investigate how student success depends on the university entrance score and on the calculation method for summing up high school performance points. Figure 2 suggests that those students, whose APS is at least 420 points and whose admission points are calculated via the summation method are more likely to graduate. Moreover, the figure also suggests that students whose APS is calculated via the doubling method are more likely to drop out than the others.

[3] Hungary is participating in the Bologna process and uses the European Credit Transfer and Accumulation System (ECTS)

TABLE I: Features of the data set.

Category	Name	Description
University program	Field of study	Categorical
	Final status	Graduated / dropped out
	Financing source	State-funded / fee-paying
	Re-enrolled	True / False
	Minimum admission point score (MAPS)	Varies across programs and years
High school performance	Study scores	In $[0, 200]$
	Matura scores	In $[0, 200]$
	Extra points	In $[0, 100]$
	Competition result	Place / none
	APS calculation	Doubling / summation
	Surplus score	(APS - MAPS)·MAPS
High school grades	Mathematics	All these features take values in $[2, 5]$
	Hungarian language and literature	
	History	
	Science subject	
Matura exam scores	Mathematics	All these features take values in $[0, 150]$
	Hungarian language and literature	
	History	
	Chosen subject	
Language certificate	Language	Categorical
	Level	Intermediate / advanced
	Number of certificates	Integer
First semester related	GPA	In $[2, 5]$
	Number of math assessment trials	Integer
	Credits earned	Integer
	Credits recognized	Integer
	Credit index (CI)	$\sum_i \frac{\text{grade}_i \cdot \text{credit}_i}{30}$, where sum is taken over completed courses
	Corrected credit index	$\text{CI} \cdot \frac{\text{credits earned}}{\text{credits taken}}$
Known in the first week	Credits taken	Integer
	First semester status	Active / passive
	Math assessment test score	In $[-15, 60]$
	Living on-campus	True / False
Personal information	Gender	Female / Male
	Age	At the time of enrollment
	Hungarian citizen	True / False
	Birth place	Hungary / Foreign country

Other features were also derived. One that measures how much the given student outperformed the minimal admission score with his/her admission score of the program that (s)he enrolled in, what we weighted by the quotient of admission score and the state-defined minimum (we call this attribute *surplus scores*). The *re-enrolled* feature indicates whether a certain student has attended to BME formerly, dropped out for some reason and then re-enrolled again. It was found that re-enrolled students have less chance to graduate [17].

IV. Methodology and Results

We face a balanced binary classification problem since our data set contains 10,196 labeled records with classes dropped out (5,064 instances) and graduated (5,132 instances). Note that since in this work dropping out is in the focus of interest, similarly to medical testing, during the classifications we consider dropping out as the *positive class* and graduation as the *negative class*.

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

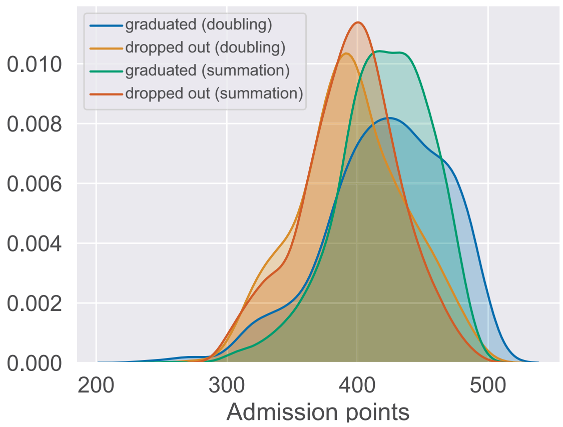

Fig. 2: Kernel density estimates of admission points based on 3,135 students data from a single year (2013) grouped by final status and admission point calculating method.

In the data preprocessing steps, first, we had to deal with missing data due to several one-side data frame joins (on a well-defined key) during data preparation. On average almost 3 out of 41 fields are missing in each record, mostly high school performance related variables, however, university performance related columns are almost fully complete. Tackling this problem, due to the large amount of missing data, we rather applied more sophisticated methods than simply excluding records with missing attributes, or inserting column means. In order to fully prepare data for modeling and imputation, one-hot encoding was used for converting categorical variables into numerical ones. Features were normalized (to 0-1 range) in order to maintain numerical stability and modeling performance, as well as decreasing training time. To evaluate models we apply ten-fold cross-validation using different partitions to prevent overfitting and to build generally well-performing predictors.

A. Addressing missing fields

Imputation techniques provide a state-of-the-art solution for the problem of incomplete data. In order to find the most efficient way to complete data, we applied several data imputation algorithms. The workflow of the evaluation of the data imputation is as follows:

1) Firstly missing data (using all available attributes) are filled using four different techniques (k-NN, MissForest, MICE and matrix factorization).
2) Gradient Boosted Tree (GBT) is trained on each imputed data set (not using first semester related features), to predict dropout.
3) At last, the trained models are compared on a validation set using binary classification metrics, and the best performing data imputation method for the problem is chosen.

In the following, we briefly summarize the applied imputation techniques. A conventional solution to the problem of missing data is applying the k-Nearest Neighbors imputer algorithm. However, there are several more advanced approaches, such as MissForest and MICE. MissForest is a Random Forest based imputation executed in an iterative fashion [21]. Moreover, currently the most frequently used technique is the Multivariate Imputation By Chained Equations (MICE) [22], which creates multiple multivariate imputations by a chain of univariate procedures and it can operate with several univariate imputers such as linear regression, decision tree, and random forest. A fourth possible solution is applying matrix factorization on the incomplete data matrix and filling missing fields by a gradient descent based method.

Another way of addressing the problem discussed so far is using machine learning models that can handle missing data, since some of them have a built-in procedure to handle the missing fields e.g. XGBoost [23].

During the evaluation of imputation techniques, we found that MICE and MissForest provided the most efficient and equally well-performing solutions for the problem of missing data.

B. Modeling scenarios

The main goal of this study is to examine and compare the predictive power on long-term student success of two feature sets: the data that are available at the time of enrollment (e.g. mainly high school performance), and the data that are available at the end of the first academic semester.

In this work, we perform predictions in four different scenarios. These scenarios differ only in the set of attributes that are used for prediction (explanatory variables), and the binary target variable remains unchanged, which is the final academic status (graduated or dropped out):

- (**S1**) The explanatory variables are the features that are available at the time of enrollment. Namely all categories except *First semester related* features from Table I.
- (**S2**) The union of the explanatory variables of (S1) and the first-week attributes that are listed in Table I.
- (**S3**) The explanatory variables include all the attributes listed in Table I, i.e. including both high school and university performance measures.
- (**S4**) The explanatory variables are the first university semester related attributes, more precisely features that are not in the *High school performance* category of Table I.

In the first scenario (S1) we investigate how pre-enrollment performance measures can determine student attrition, in favor of early identification of at-risk students. As a spin-off scenario of (S1), in (S2) we study how the addition of first-week attributes improves the accuracy of the prediction.

In the third scenario (S3) we focus on giving the best prediction possible, using all data that is available at the end of the first semester. Here we analyze the whole feature set and we aim to investigate the incremental predictive validity of first-semester performance measures over the attributes of scenario (S1), namely over pre-enrollment achievement

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

measures. Since we aim to predict final academic status using data that is available at the end of the first semester, we exclude those students from the cohort who dropped out in the first semester. This operation decreased the number of records by 9% and the data remained heterogeneous (44.5% positive 55.5% negative label). The exclusion of these students does not affect the joint distribution of the variables much, making it possible to compare the performance of the models trained on these slightly different data sets.

We note that in Hungarian higher education one cannot be expelled from the university in the first semester due to study reasons, but the students can interrupt their studies on their own will. Dismissal policies such as minimum credit requirements take effects only in the second semester. Regarding scenarios (S3) and (S4), our data set contains students who managed to pass the first semester. We also mention that in the case of students who started their university studies with passive semester, we used their very first active semester related data in the analyses.

Finally, in the last scenario (S4), we examine the predictive power of the first-semester features, moreover investigate the incremental validity of high school related features over the academic performance indicators.

C. Evaluation and discussion

In the modeling scenarios, three supervised learning algorithms are applied for the classification task, which seem to be the most successful models in previous EDM-related studies [4], [6], [17], [18]. Namely, these models are the Gradient Boosted Trees (GBT), the state-of-the-art ensemble machine learning algorithm eXtreme Gradient Boosting (XGB) [23], and artificial neural networks (ANN). Considering the moderate complexity of our data set, to avoid overfitting, we used neural networks with only one fully connected hidden layer. Furthermore, dropout regularization, different architectures (number of layers and nodes) and various optimizers were also tested. In the final architecture no dropout regularization was applied, and RMSprop [24], an optimizer with adaptive learning rate control, has been selected for training the network. In the case of the decision tree based models (GBT and XGB), grid optimization with cross-validation was performed to tune the hyperparameters.

After the training process and hyper-parameter optimization (using imputed data and cross-validation), the predictors are evaluated via the following binary classification metrics: accuracy, precision, recall and AUC (area under the receiver operating characteristic curve) [25]. The average performance of the models regarding the four previously introduced scenarios can be found in Table II.

Observing the first section of the table, we can see that pre-enrollment achievement measures (S1) still have relatively high predictive power on final academic status. The best performing model is the artificial neural network with AUC = 0.815. We can also observe how the first-week attributes (S2) improve the performance of the prediction of final academic performance compared to the features of (S1). We can conclude that the features that are available in the first week of the first semester (e.g. score of the math assessment test, and the number of taken credits) moderately increase the accuracy of the predictions.

TABLE II: Evaluation metrics using data available at the time of enrollment (S1), its supplementation with first-week features (S2) all available features (S3) and the first semester related features only (S4).

Scenario	**ML model**	**Performance**			
		Accuracy	Precision	Recall	AUC
(S1)	GBT	68.0%	67.0%	73.5%	0.729
	XGB	71.9%	71.6%	72.1%	0.787
	ANN	74.5%	73.0%	73.4%	0.815
(S2)	GBT	73.3%	72.0%	70.5%	0.789
	XGB	74.2%	72.6%	72.1%	0.816
	ANN	76.6%	75.9%	73.4%	0.823
(S3)	GBT	85.3%	85.5%	80.8%	0.920
	XGB	85.4%	86.1%	80.8%	0.920
	ANN	85.8%	86.3%	81.8%	0.916
(S4)	XGB	82.6%	82.0%	74.6%	0.892

By comparing the performance scores of the third scenario (S3) with the first (S1) (see Table II), one can observe the significant incremental predictive validity of university performance related data. Moreover, models managed to provide high precision, which means that the number of false dropout predictions is low, making these models favorable for decision-support systems, e.g. the one presented in [18]. Remarkable improvement was achieved by the first semester related measures, and ANN remained the best performing model in the (S3) modeling scenario. The AUC of the neural network model has grown from 0.815 to 0.920 after the addition of first semester related features. Comparing to another study that investigated graduation using rich, both pre-entry and university-related data [13], we have managed to achieve higher accuracy and AUC scores, since their best model's accuracy and AUC were 83.2% and 0.811 respectively.

Investigating solely the predictive power of first-semester achievements data may also provide important findings. In scenario (S4) the average accuracy and AUC of an XGBoost model using ten-fold cross-validation are 83% and 0.89 respectively, which means that given the first semester related information, pre-enrollment attributes have only little incremental predictive validity on final status over first-semester performance, which agrees with the findings of related studies [13], [14].

Testing the performance of the different data imputation techniques, we have found that if we use all the features for training GBT models, then we receive very similar accuracy and AUC scores regardless of the imputation method. This phenomenon may be due to the fact that the most important attributes are the university performance measures, and these variables were only slightly affected by incompleteness.

The five most important features in predicting student failure on the whole feature set according to XGBoost (based on 200 boosted trees) can be seen in Fig. 3, where F-score means that how many times did the algorithm use a given feature

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

to split data during the construction of decision trees. The two outstandingly important attributes are both first-semester performance related ones, namely corrected credit index and credits earned, and only one pre-enrollment achievement measurement occurs in the top five: the surplus score, which measures that how much a student over-performed the minimal admission point score.

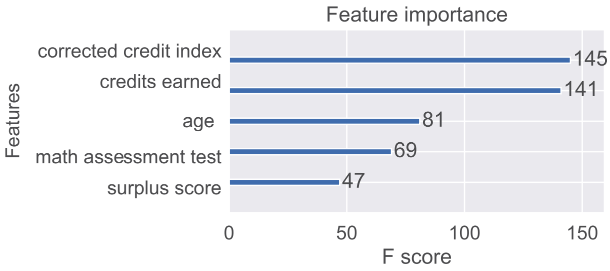

Fig. 3: Feature importance according to XGBoost based on 200 trees, learned on all the features (S3).

During the first two and the fourth prediction scenarios (S1), (S2), and (S4), models were only optimized on AUC and accuracy. In the third scenario (S3) we also optimize another binary classification metric, namely the decision threshold, that is tuned to keep precision high to prevent false positive labeling, thereby avoiding false dropout predictions that may affect students' attitude and motivation harmfully.

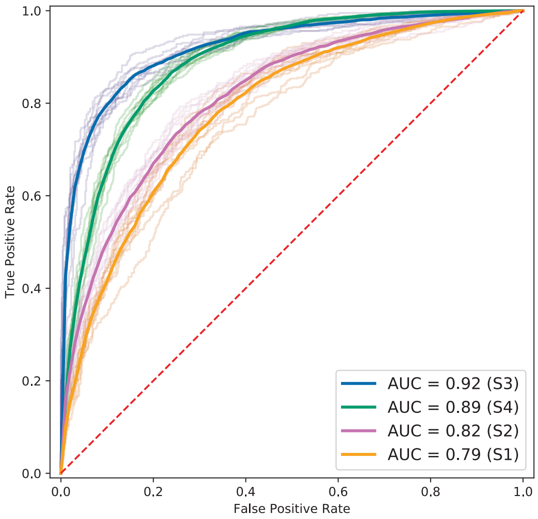

Fig. 4: ROC curves of 10-fold cross-validation of XGB predictors corresponding to four modeling scenarios. The average ROC curves of the cross-validations are shown in darker shades.

Figure 4 shows the receiver operating characteristic (ROC) curves (using 10-fold cross-validation) of XGBoost models: one that is trained solely on pre-enrollment features (S1), another that is supplemented with features available in the first week of the university studies (S2), a third one that uses only the first-semester features for prediction (S4), and a fourth one which is built on all variables (S3). We can conclude that the best predictive performance is achieved when both the pre-enrollment measures and first-semester performance indicators are used, which is in alignment with a recent study [16]. By comparing the orange (S1) and the blue (S3) ROC curves in the figure, one can observe that there is a significant improvement in predictive power when high school performance variables are supplemented with first academic semester related attributes, meaning that first semester academic performance indicators have significant incremental validity on final academic performance over pre-enrollment achievement measures. Moreover, the purple (S2) and orange (S1) curves show that even only the first week related features slightly increase the predictive power on university success. On the other hand, by comparing the green (S4) and blue (S3) ROC curves, we can conclude that pre-enrollment achievement measures (mostly high school performance) have only a slight incremental validity over first-semester features. Finally, again from the comparison of these two ROC curves, one can also conclude that first semester related measurements alone have significant predictive power on graduation.

V. Summary and Conclusion

The main objective of this study was to analyze the predictive power of high school and first-semester achievement variables on student dropout under different scenarios to gain a better understanding of the dropout phenomenon and identify at-risk students. At first, we carried out predictions that relied solely on data available at the time of enrollment (e.g. high school grades, matura exam scores, and university program related information), moreover we found that these predictions can be slightly improved by using features that are known in the first week of the first academic semester e.g. mathematics assessment scores.

In the second part of the analyses, by using state-of-the-art machine learning algorithms such as XGBoost and by supplementing the pre-enrollment achievement measures with first-semester performance indicators, we managed to provide powerful classifiers (AUC = 0.920) of final academic performance. Furthermore, we found that first-semester performance indicators have significant incremental predictive power over pre-enrollment achievement measures. On the other hand, given the first-semester performance related attributes, pre-enrollment measures have only little incremental predictive validity which is in alignment with related studies, i.e. the predictive power of high school achievements incorporate into first-semester performance measures.

As a possible application of this research, using a reliable predictive model, remediation courses could be recommended for at-risk students, and relevant information could be provided for university policy-makers. Moreover, the best performing

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.

model together with machine learning explainability techniques could be deployed as an AI-based feedback system that helps students find the skills they need to master to enter higher education and complete their university studies successfully.

Acknowledgment

The research reported in this paper was supported by the BME - Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/SC). The publication is also supported by the EFOP-3.6.2-16-2017-00015 project entitled ”Deepening the activities of HU-MATHS-IN, the Hungarian Service Network for Mathematics in Industry and Innovations”. The research of R. Molontay was supported by the ÚNKP-19-3-III New National Excellence Program of the Ministry of Innovation and Technology.

References

[1] J. Vossensteyn, A. Kottmann, B. Jongbloed, F. Kaiser, L. Cremonini, B. Stensaker, E. Hovdhaugen, and S. Wollscheid, *Dropout and completion in higher education in Europe: main report*. European Union, 2015.

[2] OECD, *Education at a Glance 2013*, 2013.

[3] C. A. Del Río and J. A. P. Insuasti, “Predicting academic performance in traditional environments at higher-education institutions using data mining: A review,” *Ecos de la Academia*, vol. 2016, no. 7, 2016.

[4] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, “Predicting academic performance: a systematic literature review,” in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 2018, pp. 175–199.

[5] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in *Learning analytics*. Springer, 2014, pp. 61–75.

[6] A. M. Shahiri, W. Husain *et al.*, “A review on predicting student’s performance using data mining techniques,” *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[7] P. Leitner, M. Khalil, and M. Ebner, “Learning analytics in higher education—a literature review,” in *Learning analytics: Fundaments, applications, and trends*. Springer, 2017, pp. 1–23.

[8] A. Dutt, M. A. Ismail, and T. Herawan, “A systematic review on educational data mining,” *IEEE Access*, vol. 5, pp. 15 991–16 005, 2017.

[9] M. Baranyi and R. Molontay, “Effect of mathematics remediation on academic achievement–a regression discontinuity approach,” in *2019 International Symposium on Educational Technology (ISET)*. IEEE, 2019, pp. 29–33.

[10] N. Horváth, R. Molontay, and M. Szabó, “Who are the most important “suppliers” for universities? - ranking secondary schools based on their students’ university performance,” in *2nd Danube Conference for Higher Education Management: In search of excellence in higher education*, 2019, pp. 133–143.

[11] Z. Berezvai, G. D. Lukáts, R. Molontay *et al.*, “A pénzügyi ösztönzők hatása az egyetemi oktatók osztályozási gyakorlatára [how financially rewarding student evaluation may affect grading behaviour. evidence from a natural experiment],” *Közgazdasági Szemle (Economic Review-monthly of the Hungarian Academy of Sciences)*, vol. 66, no. 7, pp. 733–750, 2019.

[12] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, “Predicting student performance using personalized analytics,” *Computer*, vol. 49, no. 4, pp. 61–69, 2016.

[13] L. Aulck, D. Nambi, N. Velagapudi, J. Blumenstock, and J. West, “Mining university registrar records to predict first-year undergraduate attrition,” 2019.

[14] J. Radunzel and J. Noble, “Predicting long-term college success through degree completion using act [r] composite score, act benchmarks, and high school grade point average. act research report series, 2012 (5).” *ACT, Inc.*, 2012.

[15] C. Márquez-Vera, C. R. Morales, and S. V. Soto, “Predicting school failure and dropout by using data mining techniques,” *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7–14, 2013.

[16] J. Vulperhorst, C. Lutz, R. de Kleijn, and J. van Tartwijk, “Disentangling the predictive validity of high school grades for academic success in university,” *Assessment & Evaluation in Higher Education*, vol. 43, no. 3, pp. 399–414, 2018.

[17] M. Nagy and R. Molontay, “Predicting dropout in higher education based on secondary school performance,” in *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*. IEEE, 2018, pp. 389–394.

[18] M. Nagy, R. Molontay, and M. Szabó, “A web application for predicting academic performance and identifying the contributing factors,” in *47th Annual Conference of SEFI*, 2019.

[19] P. Biró, T. Fleiner, R. W. Irving, and D. F. Manlove, “The college admissions problem with lower and common quotas,” *Theoretical Computer Science*, vol. 411, no. 34-36, pp. 3136–3153, 2010.

[20] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, “A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition,” *Expert Systems with Applications*, vol. 41, no. 2, pp. 321–330, 2014.

[21] D. J. Stekhoven and P. Bühlmann, “MissForest—non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011.

[22] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software, Articles*, vol. 45, no. 3, pp. 1–67, 2011.

[23] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. ACM, 2016, pp. 785–794.

[24] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[25] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

Authorized licensed use limited to: Vanderbilt University Libraries. Downloaded on November 01,2022 at 18:26:48 UTC from IEEE Xplore. Restrictions apply.