Exploring the Importance of Factors Contributing to Dropouts in Higher Education Over Time

Hasan Tanvir, Irene-Angelica Chounta
University of Tartu, Estonia
{hasan.mohammed.tanvir, chounta}@ut.ee

ABSTRACT

The aim of this work is to provide data-driven insights regarding the factors behind dropouts in Higher Education and their impact over time. To this end, we analyzed students' data collected by a Higher Education Institute over the last 11 years and we explored how socio-economic and academic changes may have impacted student dropouts and how these changes may have been reflected or captured by students' data. To analyze the data, we engineered features that may predict student dropouts on three dimensions: academic background, students' performance and students' effort. Then we carried out a correlation analysis to investigate the potential relationship between these features and dropouts, we performed a multivariate analysis of variance (MANOVA) to investigate whether the engineered features change significantly among student cohorts with different admission year and, finally, we carried out a regression analysis to confirm that the engineered features' impact on predicting dropouts changes over the years. The results suggest that the importance of features regarding the academic background of students (such as the students' prior experience with the academic institution), and the effort students make (for example, the number of days students spend on academic leave) may change over time. On the contrary, performance-based features (such as credit points and grades) do interact with time suggesting that performance measures are stable predictors of dropouts over time. On the basis of the findings, we argue that the performance of prediction models for assessing students at risk of dropping out of their studies can be affected by the age of data and we outline the possibility of including a forgetting factor for non-recent data in order to leverage their impact on prediction performance.

Keywords

dropouts, feature engineering, predictive modeling, higher education

1. INTRODUCTION

Student retention is pivotal for success of an educational institute. To understand the reasons behind dropouts, individual cases of students had to be analyzed on one by one basis. The advent of information technology, the use of digital technologies by educational institutes and the collection of rich data regarding students' background, performance and effort offer the possibility of using advanced analytical approaches, such as machine-learning in order to identify trends and patterns that may indicate students at risk of dropping out from their studies [13].

To ensure quality education, Higher Education Institutes (HEIs) typically offer analytical solutions - such as, learning dashboards - to inform stakeholders (for example, program directors, academic specialists and instructors) with respect to student dropouts [2]. To do so, machine-learning models are typically employed to analyze data collected by Study Information Systems (SISs) and Learning Management Systems (LMSs) and to predict whether a student faces a risk to drop out from their studies [1, 4]. This is a well-established practice but little research has been carried out with respect to the temporal aspects of data, such as the age of data used to train predictive models for assessing dropouts. One may argue that – in terms of predictive performance – the more training data, the better. However, our hypothesis is that the factors affecting dropouts in Higher Education (HE) change significantly over time due to socio-economic conditions [16] and to such an extent that data age may affect the computational model's predictive accuracy.

The goal of this research is to analyze the data collected over 11 years, 2010 to 2020 from the SIS of a national European HEI. The objective is to engineer and identify the important log-based features behind dropouts, how these features may change over years, and to explore their impact on predicting dropouts. The contribution of this work is twofold:

- to provide insights regarding log-based features that may relate to student dropouts in HE;
- to explore the relationship between the aforementioned features and time regarding their impact on dropout prediction.

In the following section we provide a short overview of related research, then we present our methodological approach and we follow up with the results of our analysis. We conclude with a contextualized discussion on our findings, the

Hasan Tanvir and Irene-Angelica Chounta "Exploring the Importance of Factors Contributing to Dropouts in Higher Education Over Time". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 502-509. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

practical implications of this work and limitations as well as potential future directions.

2. RELATED WORK

As dropouts in Higher Education, we identify the cases of students who do not successfully complete their studies for reasons that indicate lack of motivation and willingness to pursue an academic degree. Dropouts in HE is a prominent issue with negative impacts for students, and institutions that also affects national and international policies[1].

The reasons behind students dropouts can vary between personal (for example, students feeling isolated or homesick [8]), academic (such as students' lack of background knowledge or study skills [9]) and socio-economical (for example, financial difficulties and cultural adaptation [16, 12]). At the same time, factors that relate to the academic institution rather than the students themselves (such as the quality of studies and resources that the institution offers [3]) can also affect student dropouts. Tinto's theoretical model of students' dropouts from college [15] identified two dimensions as crucial in terms of academic success: student's characteristics (such as family background and goals) and student's experience with the academic system (such as student's performance and relationship with mentors and colleagues). Crosling et al. [7] attributed student dropouts to the services the academic institutes offer to students, such as information regarding the admission process, quality of the teaching, and assessment and [11] investigated the relationship between the socio-economic status of a country and students dropouts. Other work [10] argued that student dropout is often related to a combination of reasons that include individual and curriculum-level factors, for example, inefficient study skills and inefficient academic or social environment.

In this work, we examine the case of an Estonian HEI. Estonia, being a relatively new member of European Union, is going through social, structural and economic changes in many sectors, including higher education. We argue that these socio-economic changes that arguably affect student dropouts, may also affect the performance of predictive algorithms that model student dropouts if temporal aspects of students' data (such as, the age of data as depicted for example by students' admission year) are not taken into account.

3. METHODOLOGY

This research was carried out in an Estonian Higher Education Institute (HEI). Recently, the HEI launched an initiative aiming to support students in successfully completing their studies. To do so, the HEI designed a learning analytics (LA) dashboard that provided information to academic stakeholders (in this case, program directors and academic specialists) regarding potential reasons that may contribute to dropouts in their programs and suggestions concerning appropriate feedback and support that they could offer to students-at-risk. To provide this information, the LA dashboard used students' data collected by the SIS of the HEI – with the students' informed consent – throughout the students' academic career [5].

To identify students at risk from dropping out from their studies, the LA dashboard used a predictive model (described in [5] that assessed dropout risk on three dimensions: academic background of the student, student's performance, and effort. The separation of the dimensions would help the institute to link dropout factors directly to students' cohorts. Each dimension was defined based on pre-selected engineered features from the SIS database. In this work, we used data collected for students on the bachelor level from 2010 to 2020) to explore whether the predictive features used by the model change over time, to what extent, and what is the impact of this change on dropout prediction.

3.1 Method of Study

Our hypothesis was that the performance of dropout predictive models that were trained with students data collected over various admission years, will not be consistent over time; the reason for that being that the predictive features change significantly over time. For the purpose of our research, we followed a three-step approach:

- we performed a correlation analysis to explore indications of potential relationships between log-based, engineered student features and dropouts per admission year;
- we carried out a MANOVA to establish that the log-based features retrieved from the correlation analysis vary significantly over student cohorts of different admission years;
- we performed a regression analysis with interaction terms to investigate the effect of the log-based features – retrieved from correlation analysis and MANOVA – on dropout prediction for student cohorts admitted on different years.

As a proof of concept, we trained a regression model as a binary classifier to predict student dropouts using the engineered features that we acquired from the aforementioned process. Then, we tested the performance of the classifier on unseen data. An overview of the method of study is presented in Figure 1.

3.2 Description of data

In this work, we used data of bachelor-level students that the HEI collected using the Study Information System (SIS) over a period of 11 years (from 2010 to 2020). The data was originally organized in 4 tables containing information regarding students' academic background, demographics, study place, and study info data.

In the SIS database, each student and each study place (or else, curriculum enrollment) have different unique identifiers (”person ID” and ”study place ID”, respectively). This consequently means that the relationship between students and curriculum enrollments is 1 to N - that is, one student may be enrolled in multiple curricula at the same time. In order to create one working dataset, we merged the four database

[1] http://publications.europa.eu/resource/cellar/d9de3b17-0dcf-11e6-ba9a-01aa75ed71a1.0001.01/DOC_1

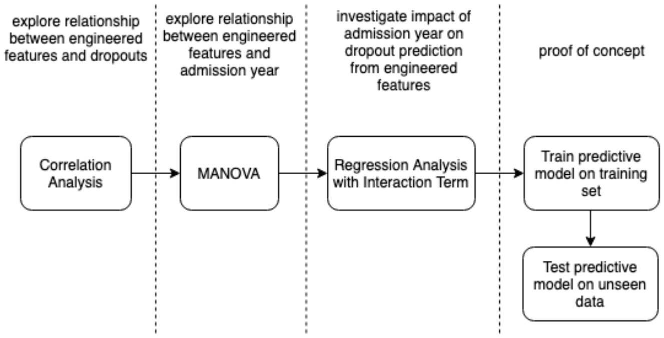
explore relationship between engineered features and dropouts
explore relationship between engineered features and admission year
investigate impact of admission year on dropout prediction from engineered features
proof of concept
Correlation Analysis
MANOVA
Regression Analysis with Interaction Term
Train predictive model on training set
Test predictive model on unseen data

Figure 1: A graphical representation of the method of study, including the three-step analytical approach and the proof-of-concept example.

tables using a combination of the unique keys "person ID" and "study place ID". Using this dataset, we engineered a set of features that can potentially describe student's academic profile on three dimensions – that is, student's academic background, student's academic performance, and student's academic effort – and may provide insights regarding students who may be at risk of dropping out from their studies. Following the recommendation of the ethics committee of the HEI, we excluded information that could be linked to students identity and demographic background to avoid potential discrimination, gender or racial bias.

As dropouts, we identified students who terminated their studies due to reasons (as recorded by the SIS of the HEI) that may indicate lack of motivation, or unwillingness to pursue an academic degree. Students can dropout at any point during the academic year, but the HEI records students' "exmatriculation" in the beginning of every semester. In total, the dataset consisted of 9623 students who are enrolled in the bachelor programs offered by the HEI. Out of these students, 3428 students dropped out at some point during their studies before they acquire an academic degree. Figure 2 shows the distribution of the dropout ratio – that is, the number of students who dropped out over the whole bachelor-level student population per admission year, over 11 years. For Year 2020 we only obtained data for the first academic semester (February to June).

3.3 Features Engineering

For each dimension of a student's academic career, we engineered a set of features from data recorded from the SIS of the HEI. In brief:

- Academic Background: The SIS records information regarding students' earlier academic background when students enroll to a study program offered by the HEI. We engineered features related to students earlier academic degrees, the admission score, admission special conditions (for example, good results in Olympiads, high scores in the academic aptitude test) and the number of previous enrollments to study programs offered by the same HEI.
- Performance: Here, we engineered features related to students' performance as depicted by grades and awarded credits throughout the study program. Performance-related features include credits earned, grades, and cumulative positive and negative study results (that is, numbers of passed and failed courses).
- Effort: Here, we considered features that can represent a student's overall effort during their studies. Some of these features are, the number of days a student spends on academic leave, the number of credits the student cancelled throughout the semester, the number of the registered courses during a semester and information about student's allowances and achievement stipends.

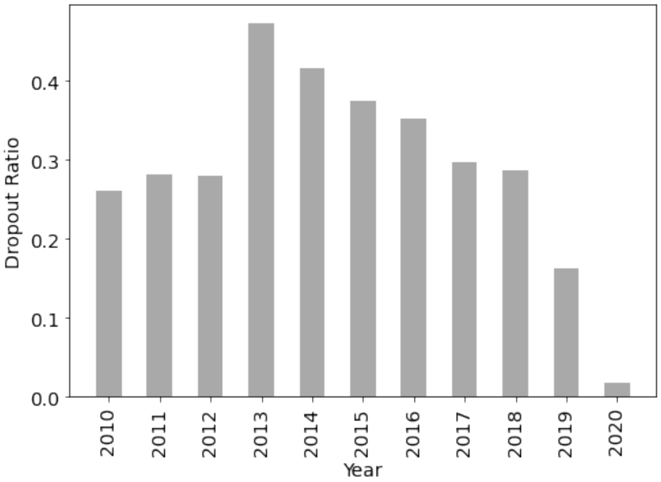

Figure 2: The dropout ratio, that is the ratio of the students who dropped out over the whole bachelor-level student population per admission year, from 2010 to 2020

The complete set of features per dimension along with a short description for each is presented in the appendix (Table 4).

4. RESULTS

The results are presented per each step of the analytical process: the correlation analysis, the MANOVA and the regression analysis. For simplicity, we only report statistically significant findings at the $p < 0.05$ level. Then, we report our exploratory findings from the prediction example as proof-of-concept.

4.1 Correlation Analysis

We carried out a correlation analysis (Spearman's rank-order correlation) to explore the potential relationship between the engineered features and student dropouts. We only report statistically significant correlations at the $p < 0.05$ significance level with medium and strong correlation coefficients ($\rho \geq |0.3|$) (Table 1). The correlation analysis suggests that features representing student performance and effort, such as the number of credits a student earns or the number of courses they register, may relate negatively with the probability of dropping out from their studies (that is, the more courses they register, the less likely to dropout). One interesting finding was that the student's economic support was negatively correlated with student dropouts from 2010 to

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Performance Features											
nr.of.courses.with.any.grade	-0.58	-0.65	-0.71	-0.59	-0.66	-0.72	-0.78	-0.71	-0.70	-0.35	
credits.earned	-0.72	-0.74	-0.74	-0.76	-0.79	-0.80	-0.80	-0.74	-0.73	-0.47	
extracurricular.credits.earned	-0.55	-0.56	-0.56	-0.53	-0.58	-0.54	-0.59	-0.43	-0.48		
all.results	-0.49	-0.59	-0.62	-0.53	-0.60	-0.66	-0.75	-0.68	-0.63	-0.36	
negative.results		0.32	0.38	0.37		0.30				0.31	
grade.A	-0.34	-0.42	-0.46	-0.46	-0.43	-0.45	-0.38	-0.34	-0.35		
grade.B	-0.47	-0.48	-0.50	-0.54	-0.56	-0.55	-0.53	-0.38	-0.41		
grade.C	-0.33	-0.35	-0.30	-0.40	-0.43	-0.46	-0.48	-0.35	-0.36		
grade.F	0.31										
passed	-0.66	-0.69	-0.65	-0.60	-0.58	-0.64	-0.65	-0.52	-0.35		
not.present				0.31							
Effort Features											
days.on.academic.leave			0.31								
days.studying.abroad	-0.30										
credits.cancelled		-0.34	-0.44								
nr.of.courses.registered	-0.58	-0.65	-0.71	-0.59	-0.66	-0.72	-0.78	-0.71	-0.74	-0.43	
credits.registered	-0.67	-0.71	-0.73	-0.57	-0.67	-0.71	-0.77	-0.73	-0.74	-0.47	
total_economic_support	-0.48	-0.58	-0.52	-0.31					-0.47		
study_period_in_years	-0.40	-0.63	-0.40	-0.40	-0.50	-0.59	-0.83	-0.72	-0.96	-0.87	

Table 1: Spearman's Rank Correlation for the engineered features and student dropouts per admission year. Here we present correlations where $\rho \geq |0.3|$ and p < 0.05. The features for the dimension of Academic Background did not appear to correlate strongly with dropouts over the admission years.

2013 but no correlation appears for the past few years (with the exception of 2018). This may suggest that presently students are in a better financial situation and can therefore afford studying until they complete their degrees. Alternatively, it may indicate a change in the state's or the university's policy regarding tuition fees.

The correlation analysis did not reveal any strong and significant relationship between dropouts and features of the student's academic background. However, correlations only suggest the potential existence of relationships. Therefore additional analysis is necessary to establish whether the importance of the engineered features on dropouts may change over time.

4.2 Multivariate Analysis of Variance

Next, we performed a one-way MANOVA to investigate whether the engineered features vary significantly for student cohorts admitted over different academic years. The engineered features were the dependent variables and the admission year was the independent variable for each of the dimensions. The results of the MANOVA are presented in Table 2. For the academic background dimension, all the features appear to be significantly different among the independent groups ($p < 0.05$) which may indicate that the academic background features are year-dependent. For both the performance and effort dimensions, the majority of features vary significantly among student cohorts of different admission years with $p < 0.05$.

Based on the MANOVA results we assume that the engineered features appear to be significantly different for student cohorts based on the admission year. This may consequently signify that the impact of the log-based features on dropout can be time-dependent.

4.3 Regression Analysis

To further explore whether the performance of a predictive model depends on temporal aspects of training data, we carried out a (logistic) regression analysis with the variable "*dropout*" as the dependent variable, the predictive features as the independent variables and admission year as the interaction term. Table 2 presents the features that interacted with admission year. Regarding students' academic background, we found that the students' previous experience with the HEI is dependent on admission year while the normalized admission score is significant in terms of regression analysis but marginal ($p = 0.07$) in terms of interaction with admission year. Time-dependency of previous experience with the HEI may reflect structural or policy changes of the academic institution that affect students' experience. Regarding the admission special conditions, we did not find any interaction with admission year or dropout (also evident from the correlation analysis). Furthermore, the results suggested that the students' previous study level – in case of master's level studies – may be important for dropout prediction and interact with admission year. A potential explanation could be that there is a confounding effect between the feature indicating previous studies in the same institution and the feature indicating previous study level.

Concerning students' performance, the results suggest that features such as the credits a student earns or the grades they are awarded can be used to indicate dropout risk. However, performance-based features do not appear to interact with admission year. In other words, their impact on predicting student dropout does not depend on the year a student was admitted in the academic institution. Regarding the importance of features that denote effort, such as the time a student spends on an academic leave, and the number of registered credits, they seem to have a different effect on student dropouts depending on the year of admission. This means that the features' weight on dropout prediction changes when coupled with the interaction term. This may indicate that the impact of these features on student dropouts is not consistent over time. We did not find any indications that effort-related features such as the credits a student cancels over the semester or the duration of studies (study period in years) depend on admission year.

4.4 Proof of Concept

To further explore the impact of time on modeling student dropouts using log-based student features, we split the data in two sets: a training and a test set. For the training set, we included all records of students admitted from 2010 to 2020, except those who were admitted on 2011 and on 2018 (both points representing instances close to the chronological beginning and the ending of our data collection). We used the training set to train a regression model, and we used the trained model as a binary classifier to predict student dropouts on the test set of unseen data. For both the training and testing, there are notable differences when examining the confusion matrices for the binary classifiers (Table 3). The binary classifier for performance and effort performed differently for student cohorts that were admitted on 2011 and on 2018 performed in terms of accuracy, precision and recall while the results were similar for the academic dimension. The model performed better on the 2011 dataset while in terms of recall the model performed better on the 2018 dataset. Recall is important here as the objective is to determine the students who are likely to dropout (positive class) and reduce the false negative outcomes (that is, students who were predicted as not at risk of dropping out but actually dropped out). Higher precision in 2011 test set indicates the models' dropout prediction inability as the model seem to retain less relevance with older data, (like, academic year 2011), resulting in lower false positives and increasing the overall precision value. On the contrary, higher recall in 2018 dataset indicates the model's better fit with recent data, thus contributing in lower false negatives. However, we acknowledge that 2018 is fairly recent and some students enrolled in that year might not have dropped out yet, leading to inaccurate results. As for accuracy, there are 36.05% (3273 out of 9078) instances are dropout (positive label) in the training dataset resulting in an imbalanced label distribution. The proof of concept analysis supports the hypothesis of the paper that age of the data affects the models' performance, therefore models' trained on newer data is important to increase the performance. We argue that this finding may suggest that the age of the data is pivotal to training predictive models. For the academic dimension, the overall performance was poor for both student cohorts.

5. CONCLUSION

In this paper, we explored the impact of log-based, engineered features that can be extracted from recorded student data on predicting student dropouts in Higher Education. In particular, we focused on investigating potential interactions between the engineered features and time – as represented by students' admission year – that may affect the performance of student dropout predictive models. We argued that the age of the data we use to train machine-learning models for predicting dropouts will impact the models' performance since socio-economic and cultural conditions, that arguably affect student retention, can change over time. For the purpose of this research, we engineered three sets of student features from data collected in the SIS of the HEI: one set describing the academic background of students, one set describing the performance of students and one describing the effort students put in their studies. To explore relationship between dropouts and features, and relationship between features and admission year we combined correlation analysis, MANOVA and regression analysis with admission year as the interaction term.

The results suggested that the admission year can play a critical role on the importance of the selected features for predicting dropouts. The importance of the features may change based on the socio-economic status of the state [11] which is subject to changes for multiple reasons, such as political functions, joining an economic trade or alliance, or even cultural changes and emergency situations, such as the COVID pandemic. For example, in our case, this is demonstrated by the importance of financial support provided by the state on dropout rates over the years that seems to be decreasing. One can argue that student dropouts in Higher Education is a complex topic that extends beyond the academic institution and the students themselves but it reflects socio-economic, cultural and political aspects of the society or the state. Thus, we would expect that the predictive power of engineered student features relating to societal or financial aspects – such as, the academic decisions students' make in terms of investing effort and financial support – are susceptible to change over time. On the other hand, performance-related features (such as grades and positive or negative exam results) do not appear to interact with admission year but instead their effect remains steady over student cohorts, confirming prior work [14]. Features that aim to represent the students' academic background may relate to some extent to student dropouts – as the regression analysis suggested – but their predictive power is limited and their dependency on admission year requires further investigation. To demonstrate the impact of time of predictive performance, we presented an example where we trained a binary classifier using time-sensitive features and we tested its performance on unseen data from two student cohorts that were admitted in the same HEI with a 7-year difference.

As a practical implication of this work, we envision establishing time-sensitive, predictive models for addressing student dropouts. Towards that direction, one approach would be to limit the datasets used for model training with respect to the chronology of the data, resulting in fewer older data as new data are received. However, this could lead to insufficient amount of data for training purposes. Another approach would be to incorporate ”forgetting” factors in order to minimize the impact of old, non-relevant data. In this case, forgetting could be implemented by applying weights to the training set in such a way so that temporally distant or temporally irrelevant data receive lower weights (and thus, have less impact on the training) than recent entries. Similarly, for random forest or decision trees models one could regulate the threshold limits for early stopping in tree growth as a means to include the forgetting factor.

In this research, we carried out our analysis on data collected during the past decade from the same institution. This does not allow us to generalize our findings across various tem-

	MANOVA		Regression Analysis			Regression with Interaction Term		
Feature Name	F value	$Pr(>F)$	coef	std err	$Pr(>\|z\|)$	coef	std err	$Pr(>\|z\|)$
Academic Background								
normalized_score	11.88	**5.9e-4**	-0.55	0.22	**0.01**	-3.5e+2	1.94e+2	0.07
admission.special.conditions	8.02	**4.7e-3**	0.03	0.29	0.26	1.7e+2	2.7e+2	0.52
prev.study.level_Masters	4.58	**0.03**	-0.57	0.28	**0.05**	2.7e-1	1.5e-1	0.072
nr.of.prev..studies.in.UT	8.40	**3.8e-3**	0.10	0.06	**0.079**	2.3e+2	6.7e+1	**7.4e-4**
Performance								
negative.results	56.787	**6.5e-14**	0.19	0.14	0.18	-7.1e-2	2.8e-1	0.80
grade.A	52.37	**5.9e-13**	-0.12	0.06	**0.05**	1.99e-3	4.3e-2	0.96
grade.B	225.32	**<2.2e-16**	-0.02	0.06	0.73	-5.5e-3	5.0e-2	0.91
grade.C	226.77	**< 2.2e-16**	-0.06	0.06	0.30	-1.6e-2	4.9e-2	0.75
grade.D	125.77	**< 2.2e-16**	-0.10	0.07	0.16	-2.2e-2	6.3e-2	0.73
grade.E	32.16	**1.6e-08**	-0.09	0.08	0.23	-7.3e-2	8.5e-2	0.39
grade.F	3.55	0.06	0.14	0.08	0.08	1.6e-3	5.96e-2	0.99
credits.earned	1685	**< 2.2e-16**	-0.01	0.01	**0.02**	1.6e-3	5.1e-3	0.75
extracurricular.credits.earned	76.613	**< 2.2e-16**	-2.6e-3	0.01	0.75	6.2e-3	7.5e-3	0.41
not.present	173.25	**< 2.2e-16**	-1.7e-3	0.01	0.87	-7.3e-3	8.1e-3	0.36
sum_passed_grade	1381.2	**< 2.2e-16**	0.18	0.14	0.19	-4.2e-2	2.9e-1	0.88
sum_failed_grade	108.05	**< 2.2e-16**	-0.01	0.02	0.66			
all.results	1455.1	**< 2.2e-16**	-0.13	0.13	0.31	4.94e-2	2.8e-1	0.86
Effort								
days.on.academic.leave	173.41	**<2.2e-16**	2.23	8.38e-2	**<2e-16**	7.5e-4	1.6e-4	**2.6e-6**
on.extended.study.period	253.53	**<2.2e-16**	2.16e-01	4.60e-02	**2.6e-6**	1.8e-2	4.1e-2	0.66
days.studying.abroad	240.55	**<2.2e-16**	-1.34e-02	1.37e-03	**<2e-16**	-1.2e-3	8.7e-4	0.17
days.as.visiting.student	6.8963	**8.7e-3**	-1.92e-3	1.6e-3	0.22			
credits.cancelled.during.2w	476.68	**<2.2e-16**	1.01e-02	1.11e-03	**<2e-16**	5.5e-4	7.1e-4	0.44
workload	6.9	**8.7e-3**	-1.1	8.5e-1	0.21			
nr.of.courses.registered	2501.5	**<2.2e-16**	-5.89e-01	2.81e-02	**<2e-16**	-1.8e-1	1.5e-2	**<2e-16**
credits.registered	2789.8	**<2.2e-16**	-3.36e-02	1.81e-03	**<2e-16**	1.1e-3	1.1e-3	0.32
nr.of.courses.with.any.grade	2774.9	**<2.2e-16**	5.44e-01	2.65e-02	**<2e-16**	1.7e-1	1.4e-2	**<2e-16**
nr.of.employment.contracts	20.657	**5.6e-6**	-6.1e-2	4.3e-2	0.16			
total_economic_support	17.14	**3.5e-5**	-3.02e-04	4.07e-05	**1.15e-13**	1.4e-4	2.4e-5	**1.8e-8**
study_period_in_years	3941.1	**<2.2e-16**	1.29	7.71e-2	**<2e-16**	-1.6e-2	4.6e-2	0.73

Table 2: The results of MANOVA with the engineered features as the dependent variables and the admission year as the independent variable, Regression Analysis without any interaction terms and with admission year as an interaction term. The features that interact with admission year on the level $p<0.05$ are presented in bold letters

Admission Year	Accuracy	Precision	Recall	F1
Academic Background				
2011	0.718	0.500	0.025	0.048
2018	0.709	0.400	0.027	0.050
Performance				
2011	0.912	0.937	0.738	0.825
2018	0.785	0.573	1.000	0.728
Effort				
2011	0.905	0.982	0.675	0.800
2018	0.889	0.725	0.987	0.836

Table 3: The perfromance metrics of the three models that predict student dropouts per dimension for two student cohorts: the cohort admitted on year 2011 and the cohort admitted on year 2018.

poral and spatial contexts. Additionally, in this work we only used basic information about students' background and study progress - excluding demographics. Further analysis on an extended dataset may reveal significant patterns on dropouts regarding cultural background or gender. However, it is important to ensure the safe and ethical use of sensitive and personal information of students and to establish that future use of the outcomes aims to support students and academic stakeholders in a fair and accountable context. In future work, we aim to design a predictive model for addressing dropouts in HE that will implement the forgetting factor based on data's recency. For triangulation, we will compare the forgetting factor's impact both for a regression model and for a random forest model and we will explore further the impact of the forgetting factor in terms of predictive accuracy, effectiveness and efficiency. Moreover, we will consider the possibility of analyzing gender segregated data to explore if the findings show gender bias [6].

6. ACKNOWLEDGEMENTS

This research was funded by the Estonian Research Council, grant number PSG286.

7. REFERENCES

[1] D. Azcona, I.-H. Hsiao, and A. F. Smeaton. Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction*, 29(4):759–788, 2019.

[2] D. Baneres, M. E. Rodríguez-Gonzalez, and M. Serra. An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, 12(2):249–263, 2019.

[3] J. P. Bean. Dropouts and turnover: The synthesis and test of a casual model of student attrition. *Research in Higher Education*, 12(2):155–187, June 1980.

[4] I. Chounta, M. Pedaste, and K. Saks. Behind the scenes: Designing a learning analytics platform for higher education. In *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge, Tempe, USA*, 2019.

[5] I.-A. Chounta, K. Uiboleht, K. Roosimäe, M. Pedaste, and A. Valk. Accuracy of a cross-program model for dropout prediction in higher education. In *Companion Proceedings of the 10th International Conference on Learning Analytics & Knowledge LAK20*, pages 750–755, 2020.

[6] C. Criado-Perez. *Invisible women : data bias in a world designed for men*. Abrams Press, New York, 2019.

[7] C. Glenda, M. Heagney, and L. Thomas. Improving student retention in higher education: Improving teaching and learning. *Australian Universities' Review*, 51(1):9–18, 2009.

[8] L. Hinton. Causes of attrition in first year students in science foundation courses and recommendations for intervention. *Studies in Learning, Evaluation, Innovation and Development*, 4(2):13–26, 2007.

[9] I. Johnson. Enrollment, persistence and graduation of in-site students at a public research university: Does high school matter? *Research in Higher Education*, 49:776–793, 2008.

[10] K. Kori, M. Pedaste, H. Altin, E. Tõnisson, and T. Palts. Factors that influence students' motivation to start and to continue studying information technology in estonia. *IEEE Transactions on Education*, 59(4):255–262, 2016.

[11] I. W. Li and D. R. Carroll. Factors influencing dropout and academic performance: an australian higher education equity perspective. *Journal of Higher Education Policy and Management*, 42(1):14–30, 2019.

[12] I. W. Li and D. R. Carroll. Factors influencing dropout and academic performance: an australian higher education equity perspective. *HIGHER EDUCATION POLICY AND MANAGEMENT*, 42(1):14–30, July 2020.

[13] X. Ochoa and A. Merceron. Quantitative and qualitative analysis of the learning analytics and knowledge conference 2018. *Journal of Learning Analytics*, 5(3):154–166, 2018.

[14] C. F. Rodriguez-Hernandez, E. Cascallar, and E. Kyndt. Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review*, 29:100305, 2020.

[15] V. Tinto. Through the eyes of students. *Journal of College Student Retention: Research, Theory & Practice*, 19(3):254–269, 2017.

[16] L. Willcoxson, J. Cotter, and S. Joy. Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse institutions. *Studies in Higher Education*, 36(3):331–352, February 2011.

APPENDIX

Feature	Description
Academic Background	
normalized_score	The admission score normalized by the min and max score
admission.special.conditions	Student’s admission subject to special conditions
prev.study.level	Student’s latest academic degree, such as high school graduate
nr.of.prev..studies.in.UT	Number of previous enrollments in the same HEI
Performance	
nr.of.courses.with.any.grade	Registered courses with any outcome (positive or negative)
credits.earned	Sum of credits the student earned
extracurricular.credits.earned	Credits for courses extra to student’s curricula
all.results	Number of all results cumulatively up to today
negative.results	Number of negative results up to today
pos.results	Number of positive results up to today
grade{A, B, C, D, E, F}	Number of all grades {A, B, C, D, E, F} up to today
passed	Number of passed, non-differentiated courses up to today
not.passed	Number of not passed, non-differentiated courses up to today
not.present	Number of non-taken exams due to absence up to today
Effort	
days.on.academic.leave	Days the student was on academic leave
on.extended.study.period	1 when student was on extended study period, 0 otherwise
days.studying.abroad	Days student was studying abroad (e.g. on an Erasmus exchange)
days.as.visiting.student	Number of days as visiting student to other Estonian universities
credits.cancelled	Number of credit points that the student cancelled
nr.of.courses.registered	Number of courses the student registered
credits.registered	Number of credits the student registered
credits.fulfilled	Ratio of credits earned vs. credits registered
nr.of.employment.contracts	Number of contracts the student has with the HEI
total_financing	Total amount of stipends and allowances
study.workload	Full time or part time student
study_period_in_years	Number of years a student has been studying

Table 4: The engineered features for each dimension. By ”*up to today*”, we mean the date of the data collection (19 Oct. 2020)