Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study

Chapter · March 2021

DOI: 10.1007/978-3-030-68201-9_70

3 authors:

Natalja Maksimova
Tallinn University of Technology
3 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Avar Pentel
Tallinn University of Technology
16 PUBLICATIONS **140** CITATIONS
SEE PROFILE

Olga Dunajeva
Tallinn University of Technology
8 PUBLICATIONS **41** CITATIONS
SEE PROFILE

Predicting First-Year Computer Science Students Drop-Out with Machine Learning Methods: A Case Study

Natalja Maksimova[1], Avar Pentel[1][0000-0002-3789-2263] and Olga Dunajeva[1]

[1] Virumaa College of Tallinn University of Technology, Järveküla tee 75, 30322 Kohtla-Järve, Estonia
natalja.maksimova@taltech.ee
avar.pentel@taltech.ee
olga.dunajeva@taltech.ee

Abstract. In this paper, we describe the results of the educational machine learning case study with the aim to predict first-year computer science students' drop-out in the Virumaa College of Tallinn University of Technology and determine factors that influence dropout rates. In this study two different datasets are used: (1) data obtained from the TalTech study information system; (2) students' history and study results collected in Virumaa College. To build predictive models, the following machine learning algorithms are applied: Naïve Bayes, decision trees, Logistic Regression, Support Vector Machines and Neural Networks. As a result of this study were evaluated how the dropout prediction accuracies change from the moment of the students' admission to the end of the first semester. We found, that data that were available about students before enrollment allowed to predict dropout with 70% of accuracy. Using data that obtained from first semester allowed to rise prediction accuracy to 90%. Besides, the factors were determined that are related with drop-out and that are not. Any higher education institution can conduct a similar study, since it is conducted on publicly available data from the official academic information environment.

Keywords: Students' Dropout, Machine Learning, Prediction.

1 Introduction

From 2012 to 2019 Virumaa College admits 412 students in Computer Science curriculum. During 2012-2016 231 students studied in college in Applied Computer Science curricula. 162 out of 231 students dropped out and on average 43% of the admitted students dropped out in the first academic year. In 2017 new modernized curriculum – Telematics and Intelligent Systems - entered into force, in which 181 students were enrolled during 2017-2019. Considering that Virumaa College does not require attendees to have State Examination results, in particular level, and it is enough when attendees have graduated on an upper secondary level with GPA at least 3.5, it gives opportunities for many people to start studying at higher education level. However, in

this case, quantity and quality are not always compatible, and after experiencing difficulties on first year courses, many students give up. That partly explains the high dropout rate. In the new curricula, high dropout rate during 2012-2016 was taken into account and 1st semester courses are rearranged to 2nd and 3rd semester, and partly integrated into other courses. This way the student's workload on semester was reduced without losing in quality in order to decrease first-year students' dropout. Unfortunately, the dropout rate remained at about the same level. One of the main goals of the college is to reduce student dropout rates by carrying out introductory interviews with student candidates and executing a mentoring program.

First year drop-out prediction model helps to achieve these goals. Current model that is created using Machine Learning predicts whether student drops-out during the first year and shows what is the probability of such event. This model does not discover the causes of drop-out neither gives the solution. If prognosis will be positive, mentor support will be executed. Mentor then will advise student in order to help to solve student problems and to introduce him/her to the students' support group.

The process of creation prediction models also helps to formulate questions for interviewing student candidates, in order to assess their maturity for particular curricula and higher education in general.

2 Background and Related Work

Student dropouts can happen at every level of education and in every curriculum. The problem is investigated, but there is no universal solution [1]. First substantial explanation for university dropouts was given by Vincent Tinto in 1975, who also proposed student-university integration model. After the advent of digital information systems, universities accumulate huge amount of data, that allows researchers to find latent trends in data using data mining techniques. Many universities in the United States and Europe started to analyze learning data, including creation of prediction models. Higher Education Commission of Tennessee conducted an analysis of a successful first year student, who started in one university and then moved to another [2]. Karlsruhe Institute of Technology in Germany carried out Computer Science student's dropout analysis [3]. First year student dropouts are investigated also in the University of Washington, in United States [4] and in the Technical University of Denmark [5]. Students drop-out is a general problem and the most researched are student dropouts in Computer Science and Technology curricula. While Computer Science and student's graduation rate in nominal time is less than in other curricula. A study conducted in Estonia found that demographics, student income, motivation, performance in the university, student's psychological condition, institutional characteristics and year of studies influence student dropout [6]. Another study that was done in Estonia did not confirm widespread anecdotal evidence as if most of the dropouts are caused by the wrong choice of specialty [7].

3 Methods

3.1 Datasets description

Our datasets are based on student's data from two Computer Science curricula's. Two different information systems were used - SAIS (Study Admission Information System, www.sais.ee), which is used by 38 Estonian educational institutions from different educational levels and Study Information System ÕIS that is used by Tallinn University of Technology (www.ois2.ttu.ee). Datasets include following information: study information, personal data and data about previous educational institution described in table 1, where calculated attributes are marked with (c).

Table 1. Attributes.

	Attribute	Description	Value
Data available at enrollment	Gender	student's gender	female; male
	Age	age at enrollment, in years (c)	17-64
	Month_ birth	student's month of birth	1-12
	Citizenship	student's citizenship	2 levels
	County	county in which the student resides (c)	15 counties
	Ida_Viru	student is from Ida-Virumaa (c)	1=yes; 0=no
	Years_btwn	years between entering college and graduating from a previous school (c)	0-43
	Schl_lang	language of school	Estonian; Russian; other
	Math_state_ex	mathematics state exam score	0-100
	School_GPA	school cumulative grade point average	0-5
	School_level	level of education of graduated school	3 levels
	Vocat_schl	student graduated from a vocational school (c)	1=yes; 0=no
	School	secondary school student graduated from	>50 schools
	Year_enrol	year of enrollment	2012-2019
	Study_form	form of study	daytime; session-based
	Free_of_charge	indicator of first-year free of charge study	1=yes; 0=no
First year study	SGPA _1_sem	first semester weighted average grade based on all exams	0-5
	ECTS_1_sem	in first semester accumulated credit points	0-47
	Mat_anal	student's grade in Mathematical Analysis	0-5
	Math _refresh	student's grade in Refresher Course in Mathematics	0-5
	Iinform_grade	student's grade in Informatics	A (pass); M (fail)
	English_grade	student's grade in English language	0-5

General students' academic performance data (table 2) were used in the preliminary analysis and in the calculation of the predicted class attribute Y.

Table 2. Students total academic performance.

Attribute	Description	Value
ECTS_total	total accumulated credit points	1-287
SGPA_total	total cumulative SGPA for all study	0-5
NSS	number of semesters studied	0-15
Exm_reason	reasons for exmatriculation	5 reasons
Status	indicator of study status	graduate; drop-out; enrolled
Dropout_sem	dropout semester	0-7
Y	indicator of first-year drop-out	1=yes; 0=no

In each year there were 30 to 70 candidates who want to study in Computer Science curriculum. From 2012 to 2019 average dropout percent was 43% by the end of first year, therefore, we can say that our dataset is balanced in relation to the predicted class attribute. Dropout reasons are presented in following graph (Fig.1).

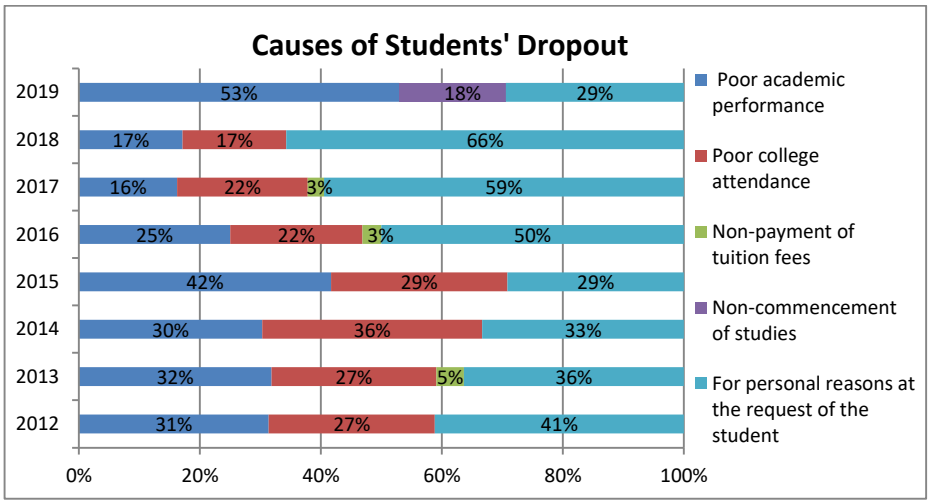

Fig. 1. Causes of students' drop-out by year of enrollment.

During the preliminary data analysis, we found that females had slightly higher drop-out rate. Often, they leave on their own initiative. The dropout was not dependent on form of study (stationary or session-based). Dropout does depend on where the student obtained the upper secondary level education. Also, it was found that most of dropouts occurred with students who belong to age group 28-32.

3.2 Data pre-processing

Preprocessing procedure included substitution of missing values and finding outliers. Missing value replacement is presented in table 3.

Table 3. Missing values replacement.

Attribute	Number of missing values	Replacement value
ECTS_1_sem	117	0
ECTS_total	117	0
SGPA _1_sem	146	0
Math_state_ex	165	based math state exam score
Mat_anal	65	0
Years_btwn	2	Age-18
County	20	'unknown'
Exm_reason	174	'enrolled' or 'graduate'
Iinform_grade	76	M (fail)
Math _refresh	57	M (fail)

Missing values were mostly substituted with zeroes, because they occurred when a student missed the test, not declared a mandatory course, etc. Missing mathematics state exam scores for 2014-2016 enrolled students were replaced using the corresponding year mathematics state exam scores by the formula: mean - standard deviation considering the type of school that students graduated, i.e. upper-secondary or vocational school.

Attribute Age had outliers: 75% of students are in age up to 29 years, 25% are older and only 9 (2% of all) are older than 42 years. Among those 9 outliers 5 dropped out on the first year. In this study outliers were left in the data set.

3.2 Data Analysis and Experiments

Freeware data analysis package Weka was used for generating models [8]. The following machine learning algorithms is used in our experiments to generate prediction models: decision trees, Naïve Bayes, Neural Networks, Support Vector Machines (SVM) and Logistic Regression.

Experiments carried out on the datasets, where 2012-2018 students data from old and new Computer Science curriculum are combined. We excluded students' data, who came in 2019, because their first year is not over yet. These datasets contain data about 367 students.

Each dataset was split into 80/20 - training and testing data accordingly. For validation we used 10 times repeated 5-fold cross validation in order to obtain more stable model quality assessments. We compared our prediction results using accuracy. Experiments were divided into two stages. On the first stage, for the purpose of early dropout prediction, we generated models on data that was available before admission of students. In the second stage, we included data that were obtained from first semester.

In order to determine important factors that are related to dropout, and for taking these into account in designing further college admission interviews, over 33 different sets of attributes were used for model generation in both stages. We began with the minimal possible set of attributes, Math_state_ex and School_GPA, data always available at admission. Then we generated models with the maximal set of all possible attributes in order to study their effect on increasing model's accuracy. Further, attribute sets were composed using Infogain Attribute evaluation method and dependent attributes removing.

4 Results

4.1 Prediction accuracies with different attribute sets

The best prediction models accuracies and used attribute sets in first and second stage presented in table 4. As our datasets were balanced, the baseline prediction accuracy that can be achieved by chance is 50%.

Table 4. Sets of attributes and best models' accuracies for stage 1.

Stage	Attributes	Best models	Accuracy, %
Stage 1	Math_state_ex, School_GPA	Logistic Regression	**70.00**
		SVM	**68.46**
		Neural Networks	**67.34**
		Naïve Bayes	**67.24**
	Citizenship, Gender, Ida-Viru, Math_state_ex, School_GPA, Schl_lang, Study_form, Vocat_schl	Logistic Regression	**66.22**
		SVM	**65.52**
		Neural Networks	**62.41**
	Citizenship, County, Free_of_charge, Gender, Math_state_ex, Schl_lang, School_GPA, School_level, Study_form	Logistic Regression	**63.10**
		SVM	**62.76**
		Decision Tree	**62.24**
Stage 2	ECTS_1_sem, SGPA _1_sem	Naïve Bayes	**90.05**
		Neural Networks	**89.68**
	ECTS_1_sem, Free_of_charge, SGPA _1_sem, Study_form, Vocat_schl	Naïve Bayes	**90.12**
		SVM	**89.70**
	Gender, County, ECTS_1_sem, Free_of_charge, SGPA _1_sem, Study_form, Vocat_schl	Naïve Bayes	**89.95**
		SVM	**89.78**

As we can see from Table 4, using data that is available before the study (Stage 1) - namely State Math examination result and previous school GPA - were enough to predict student's dropout with 70% of accuracy. Adding data from first semester (Stage 2) improved prediction result up to 90%. In the 2nd stage, the sum on ECTS and student's

GPA were sufficient attributes to make prediction over 90% of accuracy. However, some additional attributes as Study form, whether the student was studying free of charge, and whether the previous school was vocational school, had some role to play in improving prediction result. In the next section we took a closer look on different attributes and how they are related to dropout.

4.2 Most distinctive attributes

As mentioned before, on the first stage, when we used the data available before student admission, the best model was generated using only two attributes – Math State examination result and average grade from previous school. The generated logistic regression prediction model was following:

$$dropout = \frac{1}{1+e^{-(3.42 + MathStateEx * -0.02 + SchoolGPA * -0.73)}} \quad (1)$$

While State examination results were in scale 0-100 and GPA in scale 1-5, the coefficients in this model do not reflect the real importance of an attribute. When we standardize attribute values, the according coefficient values will be for Math_state_ex -0.35 and for School_GPA -0.27, which mean that low Math state examination result is stronger dropout predictor than GPA from previous school.

In the second stage, when data from the first semester was available, the best prediction model used 5 attributes. Here we present (Fig. 2) decision tree that had an accuracy of 88.54% and used 2 attributes – ECTS in 1st semester, and whether student comes from vocational school.

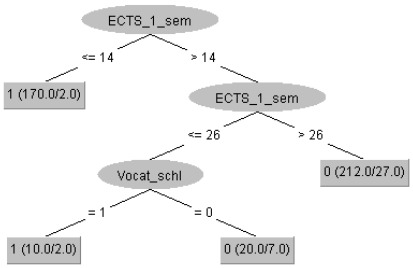

Fig. 2. Decision tree with semester data.

As we can see, the decision tree starts with most important attribute, which is the number of ECTS points student got during first semester. When it was less than 15, then the model predicts dropout, otherwise when a student got more than 26 ECTS, drop-out is not predicted. In all other cases the next important attribute is whether a student comes from a vocational school or not. Coming from vocational school predicts drop-out.

Conclusion

Our goal was to create models for students' first year drop-out prediction, and to find the factors that influencing drop-out. It was found that when using data that is available before the student's enrollment, student average grade from previous school and State examination result in mathematics, will predict dropout with 70% of accuracy. State Examination result in mathematics was found to be the most distinctive attribute with pre-enrollment data. Adding data available after the first semester raised prediction accuracy above to 90%, which is consistent with previous studies [3]. Most important attribute in this stage was amount of study in ECTS credit points on the first semester. We also found that some of demographic data are related to dropout, as for example the type of the previous school.

We also experienced during the process of data preparation that some of important data were incomplete, for instance, State examination results. So, we had to replace these missing values with somewhat questionable derived values. Requesting that information right before admission, could help to generate more accurate models based on pre-enrollment data. Also, we believe that, models can be improved, if we adjust the previous school GPA with some coefficient that reflects the level of that school. It can be done, for example, by comparing attendee's GPA with his/her state examination result.

Our study is limited to the data we had and therefore further study is needed to find out if some additional data could improve prediction models. At least, we got some insight about directions what might be important and what is not, and it helps to design interview questions and tests for student candidates, and to collect additional demographic details.

References

1. Olev Must, Aasa Must, "Kõrgkoolist väljalangevus ja üliõpilaste enesemäärtlus, (Students dropout and self-definition)" Tartu Ülikool, 2017.
2. Whitlock, Joshua Lee, "Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn" (2018). Electronic Theses and Dissertations. Paper 3356. https://dc.etsu.edu/etd/3356
3. Lorenz Kemper, Gerrit Vorhoff, Berthold U. Wigger, "Predicting Student Dropout: A Machine Learning Approach", Karlsruhe Institute of Technology, European Journal of Higher Education, 2020.
4. Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, Jevin West, "Predicting Student Dropout in Higher Education", University of Washington, 2017.
5. Ruta Gronskyte, "Methodologies of Early Detection of Student Dropouts," Technical University of Denmark, 2011.
6. Külli Kori, Margus Pedaste, Eno Tõnisson, Tauno Palts, Heilo Altin, Ramon Rantsus, Raivo Sell, Kristina Murtazin, Tiia Rüütmann, "First-year dropout in ICT studies." 2015 IEEE Global Engineering Education Conference (EDUCON). IEEE Digital Library, 2015.
7. "Tudengite õpingute katkestamise põhjused IKT erialadel. (Reasons of Computer Science student's dropout)", Eesti rakendusuuringute keskus Centar, 2015.

8. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA data mining software: an update”; SIGKDD Explorations, 2009, vol 11