Corvinus University of Budapest

THESIS COLLECTION

Prediction modelling for early identification of student dropout based on data available in higher education institutions

Author: Balint Duraczky

Supervisor: Dr. Gergely Rosta, associate professor

Budapest

2022

Doctoral School of Sociology and Communication Sciences

THESIS COLLECTION

Supervisor: Dr. Gergely Rosta, associate professor

© Balint Duráczky 2022

Table of contents

1. Background and the relevance of the research

Research on explanatory factors for student dropout has gained a particularly strong boost in the second half of the last decade. Studies and volumes of studies have been published in Hungary as well, which, catching up with the international mainstream, seek to explore the entire structure of the causal background of the dropout. Despite the great momentum, national and institutional statistical indicators demonstrate that research results contributing to the understanding of the phenomenon of dropout have not yet been utilised to such an extent that there be no drastic discrepancy between the number of people enrolled in higher education and those graduated in the given class. While according to OECD data, 47% did not reach absolutory in one of the classes started in the early 2000s (OECD, 2013), according to an analysis produced by the Education Office (hereinafter: OH) in 2020, this was still of similar magnitude a decade later (OH, 2020).

There are several reasons why the decrease does not reach the desired level. Among these, the vast majority of research examines the problem with cross-sectional or at best retrospective logic, so we know that the risk of dropout threatens students of which characteristics even at the moment of enrollment, but we do not have a system-wide solution to identify them. Without this, protection against students' unintended abandonment will be less effective. For this, due to the nature of dropout, the earliest possible detection and initial prevention based on it are of key importance, since most students with unsuccessful finish leave the initiated course during the first two years (Keller, 2020).

Furthermore, it is also a problem that theories explaining the process of dropout (Tinto 1975; Bean 1980; Cabrera et al. 1993; Bennett 2003) typically compile causal models using such complex contextual and interaction factors, for the measurement of which there is no widely applicable, systematic practice. The operationalization of concepts in well-developed theoretical models is difficult to operationalize, they are difficult to reconcile with the indicators available in administrative databases that give the most credible picture of dropout. Therefore, based on the processing of literature, it is striking that works focused on the theoretical and empirical side capture dropout in different ways. The former concentrates on the interaction of the individual and the institution, while the latter is able to capture individual attributes and use them to explain the phenomenon that is harmful from a social, institutional and individual point of view. This divergence of empiry and theory is not ideal

for the development of effective interventions, nor does it provide an opportunity to measure back the effectiveness of interventions.
Finally, it should be noted that empirical results are partly based on data that are not available in higher education institutions. In case of soft indicators used in researches of qualitative nature this may be a self-explanatory problem but the findings based on quantitative data analytics also use such variables to which – without primary data collection – not even similar ones are available in the databases managed by the higher education institution.

With regard to these reasons I base my dissertation to the premise that the application of the scientific evidences with social benefit and thus the effective containment of the dropout can be realized in case of the fulfillment of three conditions:

1. The forecasting of dropout is achieved on datasets that are available in higher education institutions.
2. Data for predicting dropout are available early, already in the first academic year.
3. The estimation method thus created is bound to the social science theory which on one hand supports the institutional use, especially with regard to the planning of the intervention, on the other hand provides guidance in connection with the development of models.

Based on this, in my doctoral dissertation I seek to establish, building on theoretical and empiry-based research results, a survival analysis based prediction model which can be applied at any domestic higher education institution, using the dataset available there, for the purpose of early warning of student dropout. In addition to the high level of estimation effectiveness, the survival model fits well with the objectives of the dissertation due to its strong interpretability.

To verify the workability of the methodology based on theoretical foundations, I use the data of a higher education institution in Hungary, which contributes to the implementation of the research and helps to carry out the work by producing the database. The higher education institution provides courses in all of Hungary's most popular specialties. On the basis of the data available in the first period of the course, I test on the database whether it is possible to effectively segmentate the student population according to the probability of future involvement in dropout using the statistical procedure of survival analysis and whether the variables involved in segmentation can be linked to theoretical models defining the field.

During the writing of the dissertation, I did not set hypotheses due to the applied nature of the objective, in line with the proposals made in the workshop discussion, since the theoretical background supports the foundation and development of methodological innovation. Below I present the four research questions of the thesis, on the basis of which I will summarize my results as well.

1) Do the indicators available in the electronic study system provide sufficient information to produce reliable, individual-level dropout predictions for institutions with heterogeneous course palettes?

The referenced literature reveals that in the dropout, the academic results, degree of commitment, the social integration have a major preventive role, but besides these, the characteristics of the individual and the institution and their match are extremely decisive. It was possible already when formulating the question to foresee that the range of factors deemed to be decisive by the theoretical models cannot be fully covered by the data available from the study system. Therefore, this research question actually asks whether the available variables alone have sufficient explanatory power to estimate dropout, either by taking over the effect of some latent factors with no clearly linkable indicators by an available indicator. While for STEM courses we have evidence of this (Kiss et al. 2019), in the case of popular courses among human-minded students there is none.

2) Do the direction and extent of the impact of variables included in the model correspond to the correlations described in the dropout theories?

In case of effective answer to the first research question, that is, creating a model suitable for estimation, it requires further investigation whether the value of the coefficients of the variables that are part of the resulting models corresponds to the theories presented in the dissertation and supports the causal relationships described therein. Linking theory with EDM based estimation confirms the likelihood of institutional implementation. A good basis for the development of the interventions implemented by the institution is the research area that accumulates a very broad knowledge, theoretical and empirical results, examining dropout from a social scientific point of view, to which specialists who develop and implement anti-dropout interventions can connect.

3) Is it possible to estimate the probability of student dropout using a survival model without data on academic performance in the higher education institution, i.e. only on the basis of the initial data forming by the end of the first month of the term time? How much does it increase the accuracy of prediction if the results of study at the end of the first semester also play a role in modeling?

The dissertation emphasizes an important literature finding related to the research question. The earlier intervention is essential in order to prevent dropout, since research examining the temporal course of dropout shows that a significant part of the dropout occurs practically after the first and second active semester.
According to this the question raises whether we may be able to identify at-risk individuals outside STEM majors without knowledge of academic achievement at university level. If this also requires academic results in higher education, then the end of the first semester will be the first time when an estimate of the expected success rate of students can be made. If a model based on enrollment data and not including academic results in higher education is suitable for estimation, it is necessary to examine whether a new estimate should be made every six months in order to achieve greater efficiency. There is a great chance for it because all theories highlight that the student's academic achievement is — obviously — the strongest predictor factor.

4) Is it worthwhile to create sub-samples within an institution to achieve the best fit estimator models?

The dissertation discusses in detail the issue of institutional impact on the social composition of the student community. I point out that both the institution and the major have a significant influence on the composition of the forming student community. From the latent variables through the social background to the most obvious demographic variables, the institution can be decisive for a number of characteristics. However, modelling should also take into account the fact that individual institutions, especially universities with a large number of faculties, cannot be considered homogeneous. Areas of courses and course locations can act as separate, independent communities within the institution. It may be possible to increase the effectiveness of estimates by looking at student dropout chances in some breakdown, such as course location. This is particularly

likely in view of the fact that estimates for STEM majors are more accurate according to literature information.

Several models will be created to answer the research question and their estimation effectiveness will provide a clear answer to Research Question 4.

2. Methods

The student sample for model building includes students in the full-time work schedule of a Hungarian higher education institution, which started in autumn 2016/17, at the undergraduate level. Of the students, only those who were admitted to the course due to their results obtained in the central admissions system were included in the sample. The test sample for validating the model from the above differs from the above only in that it includes students of the training courses who will start in the subsequent academic year, i.e. in the autumn of 2017/18. Accordingly, both samples, apart from some technical narrowing, are complete (N_t=1860; N_v=1935).

The database contains data on students who are enrolled in the first semester of the 2016/17 academic year, as well as those who started their studies in September of the subsequent year. I treat these as two separate samples, actually simulating the situation where, building on the dropout data of a known grade, an institution would be able to estimate the probability of dropping out of members of another grade. Simulation differs from the possibilities available in reality only in that it would not be possible in reality to use successive grades for modeling and to make an estimate for the next grade. This is possible from the perspective of so many years, accordingly, the estimate is made on the basis of the data of the 2016/17 grade, and the applicability of the estimates is tested on the sample of the 2017/18 year.

I used the data of the student sample cases available from the Neptun system, the Freshmen database and the Student Government for model building.

The final estimates were made using Cox's proportional-hazards model. Using Cox regression, I attempt to create ten models. In addition to the institutional main sample, the ten models are created on the basis of four subsamples and two estimates based on different variable sets. The breakdown of the institutional sample of the 2016/17 grade into subsamples is carried out by fields of courses. Thus, a model is being formed that examines dropout (institutional model) regardless of the field of course, while the four others are broken down by field of course groups for MTMI, human sciences, social and economic sciences, and sports and health sciences.

The difference between the two variable sets is that in the first one I use the information that was available to the institution until 15th October 2016, close to the start of the semester, and the second variable set builds on the data available up to the end of the first semester, i.e. 15th March 2017. Thus, in the latter, for example, the academic results of the first semester will be shown. The results calculated on the basis of ten models ultimately yield four types of estimates that I test on a sample of students starting in the first semester of 2017/18.

The usability of the estimates was measured by two instruments. I use a ROC curve to examine the ranking ability of the models, the associated AUC value for which gives the accuracy of the ranking. The AUC value measures performance well from a statistical point of view, but to answer the question of suitability for application, I used the more applicable misrepresentation matrices.

3. Results

3.1. Structure of models

Into the institutional and field of course models developed on the basis of the data available on 15th October 2016, I incorporated twelve different variables: in addition to the different scores related to admission, dormitory housing, regular social support, number of credits and subject taken up, and in one case compensation by student loans. In addition, the institutional model also includes a variable that differentiates fields of course.

Each of the models significantly improved the explanatory capacity of the *baseline* model. This is indicated by the p-value for the Chi square test, which was 0,000 for all models.

		Institutional	MTMI	Human Sciences	Social and economic sciences	Sports and Health Sciences
Admission Total Score				x		
Transferred/Academic Score				x		
Earned/GCSE Score						x
Statutory score		x				
Excess score			x		x	
Base score			x		x	x
Dormitory Housing (cat.)		x	x	x		
1/2016/2017 Amount of regular social support		x	x	x		x
1/2016/2017 Number of Credits registered (cat.)		x	x		x	x
1/2016/2017 Number of subjects registered					x	x
Field of course (cat.)		x				
Student Loans (cat.)					x	
Chi-square goodness of fit		235.359	43.932	40.590	52.844	68.683
df		9	7	5	8	7
p-value		0.000	0.000	0.000	0.000	0.000
Cases that can be used for the model	Number of events examined	548 (29.5%)	175 (34.8%)	96 (26.7%)	87 (23.3%)	190 (30.4%)
	Censored cases	1312 (70.5%)	328 (65.2%)	263 (73.3%)	287 (76.7%)	434 (69.6%)
TOTAL		1860 (100%)	503 (100%)	359 (100%)	374 (100%)	624 (100%)

Summary table of institutional and field of course models based on known variables in the first semester (source: self-edit)

The difference between the variable set of models based on the data of 15th March 2017 is already much smaller. Instead of the twelve variables of the previous models, here I used only

ten variables. The decrease in the number of variables indicates the availability of clearer explanatory variables with being aware of the performance of the first half of the year.

The model that can be prepared at the beginning of the second semester also justifies the lesson learned on the theoretical history, that ultimately the higher education dropout is most determined by whether the student has managed to integrate academically. And the best indicator for this integration is the academic result itself.

The second semester models include five variables that describe all or part of the academic performance of the previous semester. We can also include the number of credits completed, the number of subjects not completed, the number of fails, the number of entries indicating completion, and partly the amount of the second semester study scholarship.

		Institutional	MTMI	Human Sciences	Social and economic sciences	Sports and health sciences
1/2016/2017 Number of credits completed		x	x		x	x
2016/2017/1 Number of non-completed subjects		x	x		x	x
1/2016/2017 Number of fails						x
1/2016/2017 Number of entries indicating completion					x	
2/2016/2017 Study Scholarship Amount		x	x	x		
2/2016/2017 Number of subjects registered		x	x	x		
2/2016/2017 Number of Credits registered						x
2/2016/2017 Amount of regular social scholarship			x			
Reclassification (cat.)		x		x		
Status change (cat.)		x		x	x	
Chi-square goodness of fit		543.062	212.184	93.781	96.089	178.327
df		10	5	5	6	4
p-value		0.000	0.000	0.000	0.000	0.000
Cases that can be used for the model	Investigated events	486 (26.1%)	160 (31.8%)	87 (24.2%)	83 (22.2%)	156 (25%)
	Censored cases	1298 (69.8%)	328 (65.2%)	257 (71.6%)	284 (75.9%)	429 (68.8%)
TOTAL		1784 (100%)	488 (100%)	344 (100%)	374 (100%)	585 (93.8%)

Summary table of institutional and field of course models based on known variables in the second semester (source: self-edit)

The models improved the explanatory capacity of the baseline model also in these cases (p-value = 0,000). Furthermore, it is true here that regarding students still having legal relationship in the second semester, the models were based on complete data.

As the background calculations of the thesis, ten estimator models have been created, of which only the two institutional versions are detailed in the thesis book. This is because due to the similarity in content of each variable, the negotiation of the coefficients of the two models provides sufficient information.

In the followings I will present the coefficients of the institutional model based on the data available at the beginning of the first semester. As expected, we can see that a better academic result proven during admission exam increases the chances of staying in (Statutory score — B: -0.005), just as the risk of dropout changes inversely proportionally to the increase in regular social scholarship (Amount of regular social scholarship — B: 0.015).

For low-level variables, the coefficients are to be interpreted in relation to the reference category. For example, the risk of dropping out of dormitory students is only 70% of that of non-dormitory students (Dormitory Housing — exp (b) :0.702).

Variables	Coeff (B)	SE	Wald	df	p	exp (b)
Statutory score	-0,005	0,001	17,587	1	0,001	0,996
Dormitory Housing ref.: non-dormitory student	-0,353	0,097	13,316	1	0,000	0,702
1/2016/17 Regular social scholarship amount (thousand HUF)	-0,015	0,005	10,07	1	0,002	0,985
1/2016/2017 Number of Credits registered ref.: 0 credits registered			127,506	3	0,000	
25 credits registered or less	-0,872	0,170	26,247	1	0,000	0,418
Registered credits between 26 and 30	-1,361	0,131	108,192	1	0,000	0,257
31 or more credits registered	-1,472	0,150	96,670	1	0,000	0,229
Field of course ref.: MTMI			12,175	3	0,007	
Human sciences	-0,176	0,136	1,665	1	0,197	0,839
Social and economic sciences	-0,444	0,134	11,076	1	0,000	0,641
Sports and health sciences	-0,063	0,111	0,316	1	0,574	0,939

Coefficient table of institutional Cox regression model based on data from the beginning of the first semester (source: self-edit)

Cascading increasing levels of protection is provided against dropout by the number of credits registered. Those with maximum 25 credits registered reduce the risk of dropout to 41.8% of the reference group, those registering between 26 and 30 credits reduce it to 25,7%, while students with 31 or more credits registered reduce the risk of dropout to 22.9% compared to students with zero credits if no other variables are taken into account.

The field of course indicates that only social and economic sciences are characterised by a lower risk of dropout compared to the MTMI area (B: -0.444), while the two other fields of course also have a lower probability of dropout.

In the institutional model at the beginning of the second semester, variables indicating successful academic advancement and failures also appear. As expected, the number of credits completed increases (B: -0.045) the likelihood of "survival", while the risk of dropout increases directly proportionally to the increase in the number of subjects not completed (B: 0.077).

Variables	Coeff (B)	SE	Wald	df	p	exp (b)
1/2016/17 Number of Credits Completed	-0.045	0.009	27.960	1	0.000	0.956
1/2016/17 Number of subjects not completed	0.077	0.022	12.523	1	0.000	1.080
2/2016/17 Amount of academic scholarship received	-0.057	0.015	14.912	1	0.000	0.992
Reclassification ref.: remaining in state scholarship			26.189	3	0.000	
Became self-funded	0.449	0.431	1.082	1	0.298	1.567
Entered state scholarship	0.401	1.006	0.158	1	0.691	1.493
Remained self-funded	-0.561	0.117	23.051	1	0.000	0.571
Status change ref.: active in both semesters			15.240	3	0.002	
Active first, then passive	0.909	0.237	14.704	1	0.000	2.483
First passive, then active	0.235	0.737	0.101	1	0.75	1.231
Passive in both semesters	0.572	0.279	4.224	1	0.04	1.773

Coefficient table of institutional Cox regression model based on data from the beginning of the second semester (source: self-edit)

A positive feedback on the academic result is also the academic scholarship, which even assigns material interest to staying in. Accordingly, with the increase in scholarship, the probability of dropout decreases.

Reclassification data confirm that students left on self-cost course are more likely to stay in than those who started their first and second semester with government scholarship funding. Looking at the test sample, which can be considered complete, we see that both types of reclassification reduce the probability of a successful finishing.

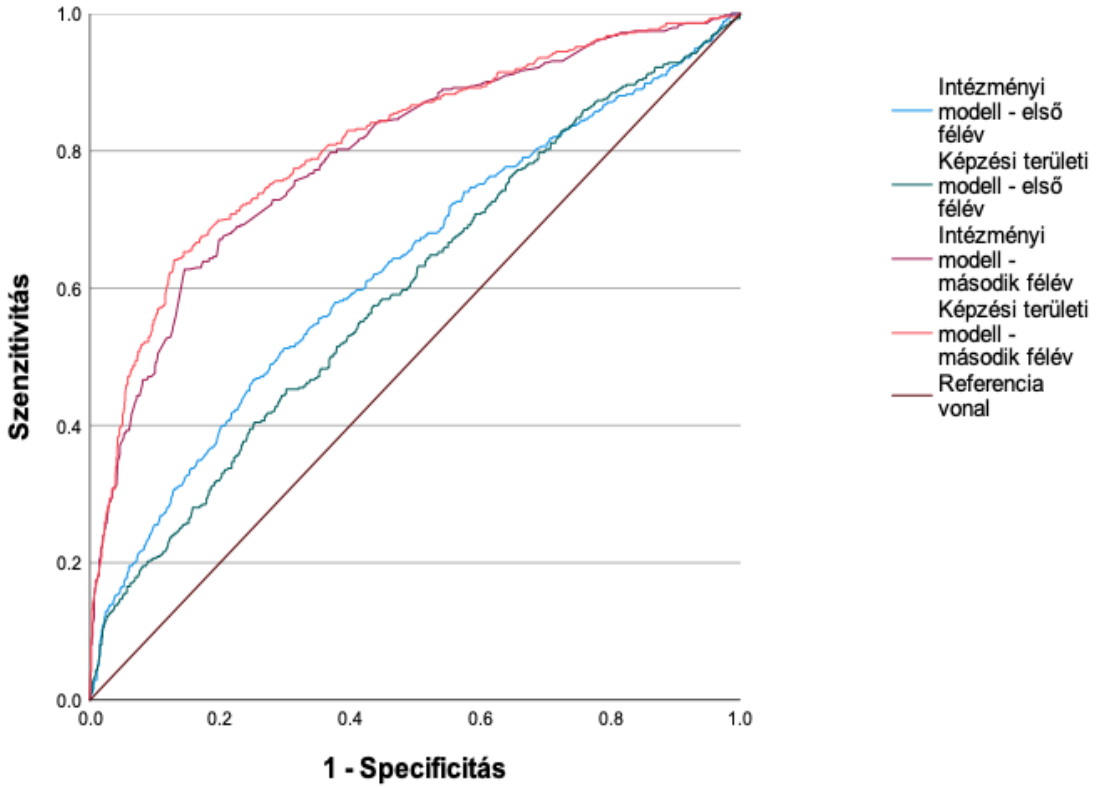
Finally, the change of status variable confirms the already stated finding that passivation worsens the chances of finishing, but this is particularly true if the student is not enrolled in the second semester. Compared to students with two active semesters, the probability of dropping out is 248% higher when a student switches to passive after an active semester, but the 177% of additional probability which threatens those passivating in both semesters is also not negligible. The group of students enrolling after a passive semester is characterized by a risk increase of 123%.

Taking into account the other 8 models not included in the thesis book, it can be said that the compilation of models, the diversity of variables and the differences between the first and second semesters provided the opportunity to link to the literature history at several points. As expected, the study results played a key role in both models, but the models of the second semester estimation were particularly determined by the indicators to be included.

3.2. Effectiveness of estimations

Below I only provide the data of the independent sample estimates, as these are the real key results of the thesis. I use ROC, AUC and error matrix data to evaluate the effectiveness.

In the curves of the figure below, we see that the efficiency of the first semester is described by a flatter line with more concave sections, which predestinates that the estimation efficiency is moderate. However, the curves of the second semester paint a more positive picture, approaching the upper left corner of the figure and have distinctly few concave sections. Furthermore, it is clear that the overall model developed on the institutional sample for the first semester data and the field of course model for the second semester provided better performance.



Estimated performance of institutional and field of course models on the test sample (starting in the first semester of 2017/2018) (source: self-edit) (legend- Intézményi modell – első félév = Institutional model first semester; Képzési területi modell – első félév = Field of course model – first semester; Intézményi modell – második félév = Institutional model seond secmester; Képzési területi modell – második szemeszter = Field of course modell second semester)

Information about the ROC curve is also confirmed by the AUC values. As a result of independent testing on the sample, the first semester models were around 0.6, but the second semester estimation reached an AUC value of around 0.8.

Model	AUC value
Institutional model — first semester	0.633
Field of course model — first semester	0.598
Institutional model — second semester	0.796
Field of course model — second semester	0.811

AUC value of estimator models applied to the test sample (those beginning in the first semester of 2017/2018) (source: own edit)

From ROC and AUC information, we could see how the estimation efficiency calculated on the test sample evolves for different models. This is partly confirmed on the basis of the data shown in the table below. According to the institutional estimation in the first semester, we are able to identify 44.1% of dropouts with a correct classification of 68%.

The performance of the estimation per field of course finds 47.6% of dropouts with 62% of right decisions, but with it the number of FP cases is almost double of that of the TP cases. Based on the dropout data, a randomly selected student being labeled a dropout, there is a 29.5% probability that we have judged correctly, while our decision yielded false-positive result with a probability of 70.5%. Models for the first semester improve this rate to an accuracy of 40-60%.

However, among the second semester models, the version compiled by field of course performs really well, and it was able to identify 70.7% of dropouts with a hit rate of 80.9%, practically surpassing the result obtained on the student sample. On the other hand, the institutional model was not able to reproduce its former effectiveness on the independent sample, but it could still be considered appropriate. For both second semester models, it is true that they result in more TP cases than FP or FN cases.

	TP	TN	FP	FN	Correct decisions	Found dropouts
Institutional first semester	235 (12.0%)	1085 (56.1%)	322 (16.6%)	293 (15.2%)	68.1%	44.1%
Field of course first semester	252 (13.0%)	953 (49.2%)	454 (23.5%)	277 (14.3%)	62.2%	47.6%
Institutional second semester	278 (15.2%)	1137 (62.3%)	252 (13.8%)	157 (8.6%)	77.5%	63.9%
Field of course second semester	374 (19.3%)	1193 (61.6%)	214 (11.1%)	155 (8%)	80.9%	70.7%

Summary table of error matrices for estimating models generated on the test sample (starting in the first semester of 2017/2018) (source: self-edit)

3.3. Summary of results

I summarize the results of the dissertation in points related to the research questions.

1) Do the indicators available in the electronic study system provide sufficient information to produce reliable, individual-level dropout predictions for institutions with heterogeneous course palettes?

The cross-validation procedure carried out on the independent sample revealed that the models available at the beginning of the first semester are limited in terms of estimating dropout rates. The limited applicability could best be quantified by AUC values around 0,6. Data from the table of the error matrix associated with the analysis showed that the number of false positive and false negative cases is quite significant when identifying nearly 50% of future dropouts. In principle, estimation is also more effective in identifying dropouts, taking into account the applicability aspects, than if an accidental decision was made, but its applicability is limited in justifying interventions that rely on serious resources. This result confirms that recruitment data in fields of course outside the engineering and natural sciences do not count as a well-suited predictor.

On the other hand, models based on data available at the beginning of the second semester performed very well, regardless of the field of course. Both ROC and AUC and the error matrix data show that institutional interventions could be planned on their basis.

Regarding the first research question, it can be summarized that the methodology is suitable for estimation, as some of the survival models using administrative data are adequate, while some have limited efficiency in predicting dropout. In the latter, it is of paramount importance to examine the possibilities of development.

2) Do the direction and extent of the impact of variables included in the model correspond to the correlations described in the dropout theories?

Examining the institutional model built from the data available at the beginning of the first semester and the second semester, differences can be pointed out. At the beginning of the first semester, when higher education results are not yet available, variables primarily not covering academic performance will play a role in addition to the statutory score. Dormitory housing can represent community integration, the number of credits registered the commitment, the amount of regular social support can represent the disadvantageous situation in the model. This is accompanied by the field of course variable, which can be directly linked not to the findings of theoretical models, but to previous empirical results. It is true to the model of Spady, Tinto, and Cabrera and colleagues that low levels of commitment to the institution or study and insufficient academic performance lead directly to dropout. This is well returned by the early first semester model, visualizing the social status of the individual as well as community integration, which is typically included in the initial stages of the dropout

processes described by the models. Based on the strength of the variables built into the model, the number of credits registered stands out, which can be identified as an indicator of commitment. Commitment can play a greater role in the first half model because an indicator with sufficient explanatory power on academic result is not yet available, i.e. high school performance is not a good predictor of higher education achievement. This can be inferred at least from the second semester model, where variables of academic result in higher education dominate the model. In addition to the number of completed and non-completed subjects, the amount of academic scholarship can also be used to deduce the academic successes and failures of the first semester. The second semester model also includes reclassification, which at this point does not report changes in the reclassification period later than a semester later, but changes in reimbursement for other reasons. From this it can be seen that the invariability of reimbursement increases persistence. However, it is difficult to judge what theoretical element appears here. Perhaps we can conclude that the invariability of the original funding plan at enrollment could also indicate the existence of commitment. The fifth variable that plays a role in the model is the status change. We can now say more confidently that the importance of commitment can be seen appearing. If a student passivates any or even all of the first two semesters, then the probability of the is significantly reduced.

Looking at the two models as a whole, the theoretical statement that community integration, commitment and academic result have a direct impact on the decision on dropping out can be confirmed. Therefore, interventions should also be established and offered to students at risk in order to achieve appropriate academic results (e.g. catch-up courses; academic mentoring programme) and to increase integration (e.g. more accessible freshman camp; creation of school circles; community mentoring programme).

In addition to planning the interventions, it is also important to address the proposals for improving the model. The importance of academic performance, engagement and community integration seems to be reflected from the Neptun data. In the case of academic performance, the result of quarterly dissertations, possibly the results of the available placement tests, can be a data set that is already available in time, but which provides better prediction ability than GCSE. Similar to the commitment, the number of absences from contact lessons can serve as a data set that is available sooner in time, and user activity associated with online study systems can also be revealing data. In connection with community integration, the most obvious and data source which is manageable using administrative tools is freshman participation.

3) Is it possible to estimate the probability of student dropout using a survival model without data on academic performance in the higher education institution, i.e. only on the basis of the initial data forming by the end of the first month of the term time? How much does it increase the accuracy of prediction if the results of study at the end of the first semester also play a role in modeling?

The first half of the question, already knowing the data from the analysis, examines that the correct hit rate of the estimator models exceeds the 70.5% retention rate known in the student sample. If everyone were considered remaining, then our prediction would be confirmed in 70.5% of the cases. The two first semester models perform slightly worse than this in terms of overall hit rates, but this is not due to the weakness of the model. As the analysis pointed out on several points, the true value is the increase in TP cases in addition to the decrease in FN cases. Therefore, the sensitivity of the estimator models has been adjusted to identify the highest number of dropouts, even at the cost of the increase of FP cases and thus reducing the percentage of correct hits by a few percent. However, the models would not exceed the 70.5% threshold without any problems without optimization, but this goal would be against aspects of practical application.

The answer to the second part of the question is obvious: The estimation accuracy of models based on data from the beginning of the second semester significantly exceeds the accuracy of models based on data available in the first semester.

It is also worth making a positive statement about the purposeful use of the survival model. The classification ability does not significantly lag behind the effectiveness of the most suitable estimator model for comparison (Kiss et al. 2019). The identification of the causes of the variation in estimation capacity would require further analysis. The differences may also result from differences in methodology arising from the two research approaches, in particular differences between the sampling procedure and the institutions on which the database is based. However, the clear virtue of the survival model is that it provides well-interpretable results, to which institutional experts who are not familiar with statistical solutions can relate more easily.

4) Is it worthwhile to create sub-samples within an institution to achieve the best fit estimator models?

Two of the four models were based on field of course subsamples and two on general institutional samples. When matching to the test sample, in case of first semester data the institutional model, in case of the second semester data the field of course model proved better. In my view, cross-validation based on one sample does not constitute sufficient justification to answer this question. It can certainly be seen that the covariates of field of course survival models sometimes differ significantly from each other, so I consider use of field of course subsamples justified in the future. However, there arose no irrefutable evidence for a sufficiently well-founded answer to this question.

4. Selected References

Aljohani, O. (2016). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6* (2), 1–18.

Állami Számvevőszék. (2021). *Felsőoktatás a változások tükrében-verseny, minőség, teljesítmény*. www.asz.hu

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

Bean, J. P. (1982). *The Synthesis of a Theoretical Model of Student Attrition.*

Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *The Review of Higher Education*, *6* (2), 129–148.

Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American Educational Research Journal*, *22* (1), 35–64.

Bennett, R. (2003). Determinants of undergraduate student drop out rates in a university business studies department. *Journal of Further and Higher Education*, *27* (2), 123–141.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education*, *64* (2), 123–139.

T. Cegledi (2012). Reziliens életutak, avagy a hátrányok ellenére sikeresen kibontakozó iskolai karrier. *Szociologiai Szemle*, *22*, 85–110.

T. Cegledi (2018). Ugródeszkán reziliencia és társadalmi egyenlőtlenségek a felsőoktatásban.

T. Cegledi (2019). Potyázók, anómiások, rituális perzisztensek és célorientált perzisztensek. A hallgatói lemorzsolódás szokatlan veszélyei. *Acta Medicinae et Sociologica*, *10* (28), 45–62.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34* (2), 187–202.

A. Czakó, L. Németh & K. Felvinczi (2019). A felsőfokú képzés befejezésére irányuló szándék előrejelzői. *Educatio*, *28* (4), 718–736. https://doi.org/10.1556/2063.28.2019.4.5

R. Csalódi & J. Abonyi (2021). Integrated survival analysis and frequent pattern mining for course failure-based prediction of Student dropout. *Mathematics*, *9* (5), 463.

Zsuzsanna Ódor Demcsákné (2016): A FIR adatok vizsgálata – Lemorzsolódási Vizsgálatok, Oktatási Hivatal, Budapest

A. Derényi (2015) Bizonyítékokra alapozott kormányzás és a kommunikáció képzés. Új jel-kép, Vol. 4. Klnsz. 1. pp. 21–34.

M. Dinyáné Szabó, G. Pusztai, & M. Szemersk (2019). Lemorzsolódási kockázat az orvostanhallgatók körében. *Orvosi Hetilap*, *160* (21), 829–834.

H. Fényes, M. Mohácsi, & K. Pallay (2021). Career consciousness and commitment to graduation among higher education students in Central and Eastern Europe. *Economics & Sociology*, *14*(1), 61-75.

T. Keller (2020). A hallgatói jellemzők feltárása érdekében adminisztratív adatösszekapcsoláson alapuló adatbázis elemzése. Budapest, Oktatási Hivatal

B. Kiss, M. Nagy, R. Molontay and B. Csabay (2019): Predicting Dropout Using High School and First-semester Academic Achievement Measures, 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)

K. Kovács, T. Ceglédi, C. Csók, Z. Demeter-Karászi, Á. R. Dusa, H. Fényes, ... & J. Váradi (2019). Lemorzsolódott hallgatók 2018.

Government of Hungary. (2016). *Fokozatváltás a felsőoktatásban 2.0*.

P. Miskolczi, F. Bársony, & G. Király (2018). Student dropout in higher education: a summary of theoretical, explanatory paths and research findings. School Culture, 28 (3—4), 87—105. http://www.iskolakultura.hu/index.php/iskolakultura/article/view/22790

M. Nagy, & R. Molontay (2018). Predicting Dropout in Higher Education Based on Secondary School Performance. *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 389–394. https://doi.org/10.1109/INES.2018.8523888

M. Nagy, & R. Molontay (2021). Comprehensive analysis of the predictive validity of the university entrance score in Hungary, *Assessment & Evaluation in Higher Education*, 46:8, 1235-1253,

Oktatási Hivatal (2020): Lemorzsolódási vizsgálatok a felsőkotatásban. Összefoglaló Tanulmány. *Oktatási Hivatal, Budapest, 2020.*

G. F. Pusztai, F. Szigeti, & K. Pallay (2019). Dropped-out Students and the Decision to Drop-out in Hungary. *Central European Journal of Educational Research*, *1*, 31–40. https://doi.org/10.37441/CEJER/2019/1/1/3341

G. F. Pusztai, F. Szigeti (2018). Lemorzsolódás és perzisztencia a felsőoktatásban. *Debreceni Egyetemi Kiadó. Debrecen*.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1(1), 64-85.

M. Szemerszki (2018). Lemorzsolódási adatok és módszertani megfontolások. Lemorzsolódás és perzisztencia a felsőoktatásban, 15-27.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition.* ERIC.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *The Journal of Higher Education*, *59*(4), 438–455.

Tinto, V. (1993). Building community. *Liberal Education*, *79*(4), 16–21.

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, *8*(1), 1–19.

Tinto, V. (2012). *Completing college: Rethinking institutional action*. University of Chicago Press.

D. A. Toth, M. Semersk, T. Cegledi, & B. Mate-Szabo (2019). The different patterns of the dropout according to the level and the field of education. *Hungarian Educational Research Journal*, *9*(2), 257–269. https://doi.org/10.1556/063.9.2019.1.23

Ye, H. (2016). *Comparison of Cox regression and discrete time survival models*. Wayne State University.

5. Own publications on the topic

Cs. Bálity, B. Duráczky (2015). Nagycsaládban élő gyermekek iskolai teljesítménye és extrakurrikuláris tevékenységei. *Metszetek*, *4* (1).

B. Duráczky (2014). Fogyatékkal élő frissdiplomások munkaerőpiaci integrációja. Felsőoktatási Műhely 4. 55-65.

B. Duráczky, T. Gulyás, G. Maszlavér (2015). Az oktatói munka hallgatói véleményezésének intézményi felhasználása. Budapest, Educatio Társadalmi Szolgáltató Nonprofit Kft.

B. Duráczky, T. Dusek, B. B. Eisingerné, P. Fehérvári, Á. Frank, B. Filep, … L. Vasa (2020). Új paradigmák a vállalatokkal való egyetemi együttműködésben. Győr: Universitas-Győr Nonprofit Kft.

B. Duráczky, N.H. László, & N. Palkovits, (2017a). Amit nemzetközi mentorként tudnod kell, Budapest: Tempus Közalapítvány.

B. Duráczky, N.H. László, & N. Palkovits, (2017b). Mentorprogram nemzetközi hallgatók támogatására, Budapest: Tempus Közalapítvány.

HÖOK - Hallgatói Önkormányzatok Országos Konferenciája, & FETA - Felsőoktatási Tanácsadás Egyesület. (2016). *A hallgatói sikerességet akadályozó tényezők és azok intervenciói*. (Author: Bálint Duráczky and Orsolya Karner)

Sz. Nyüsti, B. Duráczky (2015). Mentorprogram kézikönyv. Budapest, HÖOK

A. Szabó, Cs. Balogi, S. Csuhai, B. Duráczky, T. Kovács, B. Nagy, D. Oross, V. Papházi, Á. Pakot, G. Pataki, E. Tóth, M. Ugrósdy. (2015). A Diplomás pályakövetés hazai és nemzetközi közegben. Budapest, Educatio Társadalmi Szolgáltató Nonprofit Kft.

Á. Szilágyi, B. Duráczky (2018). Application motivations, student composition. In Botond Feledy (ed.): Mit adtak nekünk a szakkollégiumok? I. A szakkollégisták motivációi, eredményei, hallgatói összetétele. Budapest: Társadalmi Reflexió Intézet.,