https://doi.org/10.36315/2022v1end043
Education and New Developments 2022

SURVIVAL MODELS FOR PREDICTING STUDENT DROPOUT AT UNIVERSITY ACROSS TIME

Chiara Masci[1], Mirko Giovio[1], & Paola Mussida[2]
[1]MOX- Modelling and Scientific Computing Laboratory. Department of Mathematics, Politecnico di Milano (Italy)
[2]DEIB – Department of electronics, information, and bioengineering, Politecnico di Milano (Italy)

Abstract

The aim of this study is to develop a tool to recognize major responsible factors of student dropout through time, both in terms of student characteristics and type of degree courses, and to accurately predict the student time to dropout, if any. From a predictive point of view, we aim at developing an early warning system to early predict the status of a student career, identifying the risky timings in terms of dropout, as a supporting tool for early interventions policies. To this end, we follow a Survival Analysis approach, applying time-dependent COX frailty models, in which the target variable is the time to dropout of students within the first three years after the enrolment. Student careers are tracked over time, collecting time-dependent information. Results show that first year information is already powerfully predictive of the time to dropout and that dropout trends differ across degree courses and student profiles.

Keywords: *Student dropout, higher education, survival analysis, frailty COX models, time-dependent covariates.*

1. Introduction

The Italian Higher Education (HE) system measures a high level of dropout, with many students abandoning their studies during the Bachelor (Cannistrà et al., 2021; Pellagatti et al., 2021). Data from the Annual activity report of Eurostat 2020 (EUROSTAT, 2020) show that the educational attainment at the overall tertiary level is very low compared to most of the rest of EU countries. In this study, we analyze data from Politecnico di Milano (PoliMI), focusing on the careers of students enrolled in a Bachelor of Science in Engineering between 2010 and 2017. PoliMI dropout rate in engineering is around 30% and student dropout mainly occurs during the first three years after the enrolment. Since the dropout occurrences are distributed across time and that their drivers and determinants might be potentially heterogeneous, we do not focus only on the dropout event per se, but rather on time to dropout.

Dropout has been broadly studied in the literature (Aljohani, 2016; Contini, 2018; Tinto, 1975), with various sources of data and methodological approaches, mainly focused on the classification of dropout (Aljohani, 2016; Cannistrà et al., 2021; Larrabee Sønderlund et al., 2019; Viberg et al., 2018). In the last decades, some researchers have moved to a survival analysis approach (Kleinbaum & Klein, 2004), taking account and modelling the dropout timing (Ameri et al., 2016), being the student dropout the event of interest.

In this work, we contribute to this literature by proposing a Cox survival model (Cox, 1972) to analyze PoliMi dropout phenomenon. Our aim is twofold: to identify responsible factors of student dropout through time, both in terms of student characteristics and degree courses, and to develop a tool to accurately predict the student time to dropout, as soon as possible. To this end, we rely on time-dependent Cox frailty models (Therneau & Grambsch, 2000). The Cox model allows to investigate the association between the survival time of students and more predictor variables. The inclusion of time-varying covariates allows to inform the model by updating student-level covariates semester by semester, tracking the student career through time. Lastly, being the students enrolled in 16 different degree courses, the inclusion of the frailty allows to investigate and quantify the dropout phenomenon heterogeneity present at the degree courses level, by considering the nested structure of students within degree courses. To the best of our knowledge, this is the first time that time-dependent Cox frailty models are applied to educational data. We further provide a prediction analysis comparing models performance when student career information is added stepwise.

Results show that first year information is already powerfully predictive of the time to dropout and that dropout trends differ across degree courses and student profiles.

2. PoliMi dataset

We consider data coming from careers of bachelor students enrolled in an engineering faculty at PoliMi between 2010 and 2021. We focus on the first three years of their career, excluding from the analysis all dropouts occurred during the first semester[1]. Data come from two different sources. The first dataset contains student information at the enrolment (each record refers to a different student and contains his/her personal information). The features we consider from this dataset are the student gender and origins, the age at the enrolment, the PoliMi admission score, the high school grade, the type of previous studies, family income and degree course. The student career duration is indicated by the variable *CareerDuration3y*, which is computed as the difference between the day of the start of the career and the day of the end of the career, using the number of semesters as unit of measure. Students whose career is still active at the end of the third year are censored and their career duration is fixed at 6 semesters. The variable *Status3y* indicates whether the student drop out or not during the follow up time, with a high concentration of dropout events in the early semesters of the student career. The second dataset, instead, collects the student academic exams track semester per semester. In this table each, observation describes the student performance during a specific semester (exam session). The variables we consider from this dataset are *CFUP*, indicating student number of credits gained, and *Average*, indicating student weighted average grade during the specific exam session. In the time-dependent framework, the two datasets are merged to include both student personal information and student career progression results, in which the number of gained credits and the weighted average grade are computed progressively through the career, as shown in the table reported in Figure 1. In this table, each observation represents a specific interval of time, corresponding to university semesters. The time interval is defined by the variables Start and End, while *EventDrop* describes whether the student drops or not during each specific semester. The dataset collects information about 49,501 students, enrolled within 16 engineering degree courses.

Figure 1. Complete sample observations for a random student.

Stud.ID	Gender	Income	CFUPprog	Averageprog	Start	End	EventDrop	Status3y	C.Duration3y
333858	M	High	0	0.0	0.0	1.0	0	L	5.7
333858	M	High	30	23.3	1.0	2.0	0	L	5.7
333858	M	High	60	24.3	2.0	3.0	0	L	5.7
333858	M	High	90	25.2	3.0	4.0	0	L	5.7
333858	M	High	120	25.5	4.0	5.0	0	L	5.7
333858	M	High	160	24.9	5.0	5.7	0	L	5.7

3. Methods

The analysis is composed by three main parts: we start by conducting an explorative univariate analysis in which survival curves of different profiles of students are measured by means of Kaplan-Meyer estimator and compared, standing on their gender, age, family income, etc. We then apply Cox models both with time-invariant covariates, measured at the baseline (end of first semester), and with time-varying covariates, measured until the end of student careers. Lastly, we include the frailty in the Cox models to take account of the nested structure of students within degree courses and to model the induced heterogeneity.

We denote by T the nonnegative random variable representing the time to the event, or *survival time*, and by t any specific value of interest for the random variable T. We consider a student as survived at time t if he/she did not experience the event of dropout until time t. The *survival function* $S(t)=P(T>t)$ indicates the probability of an individual to survive longer than a specific time t, while the *hazard function* $h(t)$ gives the instantaneous risk for the event to occur, given that the individual has survived up to time t. The Kaplan-Meyer estimator is a nonparametric statistic that estimates $S(t)$ and allows to compare the estimated survival curves of group of students, according to the membership to a specific category. The Cox model estimates the hazard function $h(t)$ for each student as the product of a common baseline hazard function, estimated nonparametrically, and the exponential of a linear predictor composed by student-level covariates. The frailty Cox model includes a degree course-specific multiplicative factor, a Gamma distributed random variable, to the baseline hazard exploiting the heterogeneity at the degree courses level.

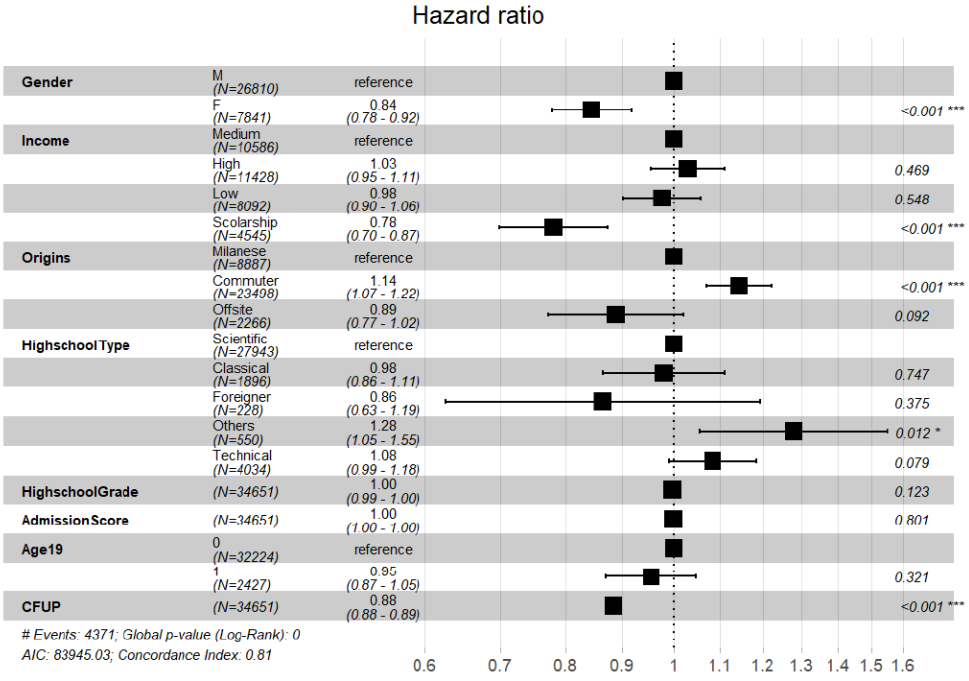
[1]During the first semester, and especially during the first two months, we observe a high number of dropouts, mainly given to unpredictable external factors like the acceptance in other universities.

4. Results

Results of the univariate explorative analysis by means of Kaplan-Meyer curves show that females have on average a 30% lower risk of dropout than males. Those students belonging to the LS ("legge stabilità") tax group tend to survive longer, encouraged by the low university taxes that they have to pay. Students coming from Milan belong to a higher risk category with respect to commuters or off-site students. Students who come from a Scientific high school have a higher survival probability through time, with respect to other types of previous studies. A significant difference is observed also in the numerical variables regarding the student admission score and the high school grade, with students with a high school grade >75 and a PoliMi admission score >71 being less at risk of dropout than the opposite categories, respectively. Regarding the early academic career results, students who gain at least 10 credits during the first semester are compared to those students who do not. The computed hazard ratio shows that students who do not pass at least 10 credits during the first semester are 7.87 times more at risk of dropout. All the analyzed features have been tested with the log-rank test, that confirms the heterogeneity of the survival probability in the different categories of students.

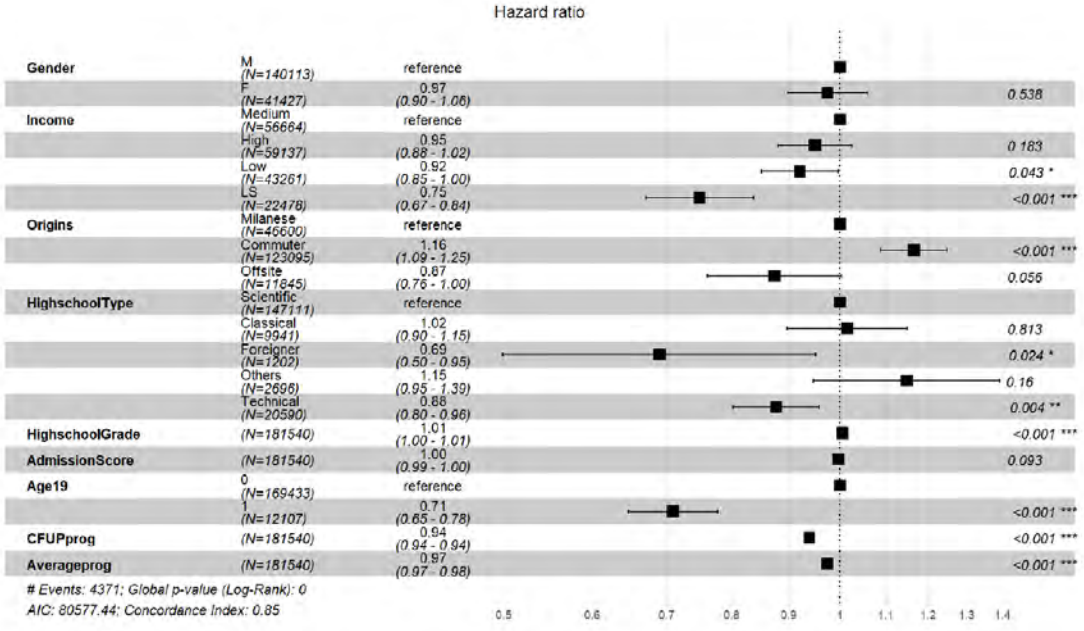
The first Cox Proportional Hazards (PH) model is fitted at the baseline, i.e., with the information at the end of first semester, and considers as predictors the categorical covariates *Gender*, *Income*, *Origins* and *HighschoolType*, the numerical covariates *HighschoolGrade* and *AdmissionScore*, while the variable *AdmissionAge* has been partitioned between those students who enroll until the 19th year of age and the ones who enroll later. Moreover, a variable indicating the number of credits passed during the first semester is added to the model. The response variable is the student career duration on a follow up time of 3 years, and the event of interest is the student dropout during this period. The model has been fitted on a training set containing the 70% of the data, obtaining the results displayed in Figure 2.

Figure 2. Result of the Cox PH model with time-invariant covariates, measured at the end of the first semester.



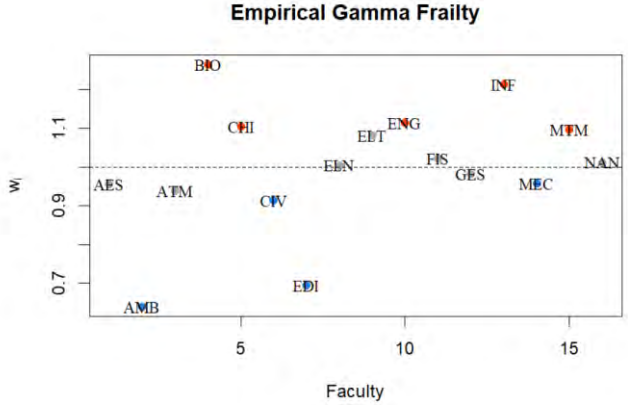
The values of the hazard ratios in Figure 2 confirm the results observed in the univariate analysis, with the most important feature being the variable related to first semester credits. We therefore extend our model to take into consideration the student academic progress through time, by means of the inclusion of time-varying covariates. In the extended Cox model with time-dependent covariates, the progressive weighted average grade and the total number of credits gained by the student across the semesters are introduced as predictors. Results of the model fitted under this new setting, in which each observation represents the student career in a specific semester, are shown in Figure 3. Some differences with respect to the time-invariant Cox model regard student gender, that loses significance in the time-dependent framework, and the admission age, which highlights a lower risk of dropout for those students who started their university career after the 19th years of age. The two time-dependent variables introduced are very significant, with students with lower grade point average and less CFU being more at risk of dropout.

Figure 3. Result of the Cox model with time-dependent covariates, updated semester per semester.



As a last step, we extend the model to include the nested structure of students within the 16 different engineering bachelor courses present at PoliMi. The resulting Frailty Cox model has been fitted both with only time-invariant and time-varying covariates. Results show that the estimated fixed-effects coefficients are similar to the ones shown in Figure 2 and 3, but it is interesting to observe how the estimated Gamma Frailty parameters differ one from another, with the highest risk faculty (BIO) having a risk to dropout that is the double with respect to the degree course with lowest risk (AMB), as showed in Figure 4.

Figure 4. Empirical Gamma frailties estimated in the Cox PH frailty model with time-invariant covariates, measured at the end of the first semester. Degree courses with a frailty higher/lower than 1 increase/decrease the dropout risk with respect to the average.



Finally, we compare the model performances across time from a predictive point of view. To do this, six different models have been implemented. Model 0 is the model computed at the enrolment, with a single observation for each student and no information about the student career result, while Model 5 is the most advised one, with up to six observations for each student and the information on 5 different exams session. The Concordance Indexes of these models, displayed in Figure 5, highlight the improvement of our predictive models when considering more observation for each student. In particular, the biggest improvement is obtained from Model 0 to Model 1, in which we observe that considering student academic results in the first semester leads to an increment in the C-Index from 0.672 to 0.810. A second great enhancement happens from Model 1 to Model 2, when considering the information until the second exams session. As we could expect, by adding information through time, the model accuracy increases, but, in the perspective of developing an *early warning system*, first and second semester information results to be already very informative.

Figure 5. Concordance Index computed on the test set, comparison between the 6 time-dependent Cox models.

Model	Concordance Index
Model 0	0.6720148
Model 1	0.8092473
Model 2	0.8428549
Model 3	0.8489186
Model 4	0.8521593
Model 5	0.8530412

5. Conclusions

This work investigates the potential of Cox regression models for describing the dropout phenomenon across time and predicting student time to dropout. The inclusion of time-varying covariates and of the frailty term constitutes a methodological novelty that allows to track students career over time and to estimate the effect of the degree courses on the student dropout risk. The exams related information is the most determining factor when analyzing the dropout phenomenon, as remarked by the increasing predictive accuracy across time. Nonetheless, a trade-off between model accuracy and the development of an early warning system arises: the model accuracy does not significantly improve after the second semester, suggesting the university to take preventive action on the early career of the student.

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.

Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016, October). Survival analysis based framework for early prediction of student dropouts. *In Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 903-912).

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T. & Paganoni, A.M. (2021) Early-predicting dropout of University students: an application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education,* 1-22.

Contini, D., Cugnata, F., & Scagni, A. (2018). Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy. *Higher Education*, 75(5), 785-808.

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

EUROSTAT (2020). Annual activity report 2020 *https://ec.europa.eu/info/publications/annual-activity-report-2020-eurostat_it*

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, 50(5), 2594-2618.

Kleinbaum, D. G., & Klein, M. (2004). *Survival analysis*. New York: Springer.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(3), 241-257.

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model* (pp. 39-77). Springer, New York, NY.

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

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98-110.