Evaluation of students drop out reasons in Information Technologies study programs

Liga Paura, Irina Arhipova

Department of Control System, Faculty of Information Technologies, Latvia University of Agriculture, Liela str. 2, Jelgava, LV-3001, Latvia
Liga.paura@llu.lv

Abstract: *The present study analyse the first study course students' dropout rates in higher education institutions, using the real data of Information Technology study program in Latvia University of Agriculture. The problem is to identify, what are the reasons, why only 50% of students completed university and obtained a bachelor's degree. Particular attention was paid to the fact that 30-40 % of students drop out during the first study year. In this research evaluation of the probability of completing University studies was made. Using Survival analysis Proportional hazard model the factors that allow identifying students who are in drop out risk group were described. The following factors were evaluated: students' study duration (month), age, gender, secondary school marks, priority to study in the program (first, second, third) and finance source (budget, private). The results of this study have allowed defining the necessary decision solutions for the 1st study year students' dropout rate decreasing and students motivation increasing to study in information technologies field.*

Keywords: Survival analysis, students' dropout.

Introduction

Survival analysis is a class of statistical methods for studying the occurrence and timing of events and the methodology has been developed over several decades bay Cox, 1972, Kaplan and Meier, 1958. Survival analysis is so named because the method is most often applied to the study of deaths. Although survival analysis was originally developed to analyse cancer data, its use was later extended to study a variety of events (cited by Min et al., 2011)

The analysis of survival experiments is complicated by issues of censoring, where an individual's life length is known to occur only in a certain period of time, and by truncation, where individuals enter the study only if they survive a sufficient length of time or individuals are included in the study only if the event has occurred by a given date. The use of counting process methodology has allowed for substantial advances in the statistical theory to account for censoring and truncation in survival experiments (Klein and Moeschberger, 2005).

Survival analysis is the name for a collection of statistical techniques used to describe and quantify time to event data. There are many different methods used to conduct survival analyses: Life Tables, Kaplan-Meier estimators, Exponential regression, Log-normal regression, Cox proportional-hazards regression.

In 1958, Product-Limit (P-L) method was introduced by Kaplan and Meier (K-M). In the Journal of the American Statistical Association, Kaplan and Meier proposed a way to nonparametrically estimate $S(t)$, even in the presence of censoring (Kaplan and Meier, 1958). The method is based on the ideas of conditional probability and survival function is defined us $S(t) = \Pr(T \geq t)$. $S(t)$ is calculated by Kaplan and Meier estimator:

$$S(t)=\prod(1-(d_t/n_t)), \quad (1)$$

where $t_1, \ldots t_K$ – the set of K distinct death times observed in the sample
d_t –is the number of deaths at time t
n_t – the number of individuals at time t

Kaplan-Meier is technic to analyse survival-time data and to compare two treatment groups on their survival time. Another aim of Survival Analysis is to compare two or more group's survival curves, which usually are made by the Log-rank test (Hosmer and Lemeshow, 1999). The null hypothesis for Kaplan-Meier and Log-rank test is: no difference between the population survival curves (i.e. the probability of an event occurring at the time point t_i is the same for each group).

Cox proportional-hazards regression is useful to identify the risk factors and their risk contributions, selecting efficiently a subset of significant variables, upon which the hazard function depends. Cox's proportional hazards model is analogous to a multiple regression model and enables the difference between survival times of particular groups of respondent to be tested while allowing for other factors (Bewick et al., 2004).

The uses in the survival analysis of today vary quite a bit. Applications now include time until onset of disease, time until stockmarket crash, time until equipment failure, time until earthquake, and so on (Smith and Smith, 2001).

The aim of our study is to evaluate the factors which affect the 1st study course students' dropout rate.

Materials and methods

The data set include 91 full-time students from Faculty of Information technology, enrolled in 2011-2010 academic year at the Latvia University of Agriculture.
Information about students' study duration (month), gender, secondary school marks, priority to study in the program (first, second, third and lower) and finance source (budget, private) were included in the data set.
According the Latvia enrolment rules all potential students may choose several programs at the same time during application process. Students must indicate priority to study in the each program separately (first, second, third etc.) depending on financing source (budget or self-finance). In our case all data have defined by 3 groups: 1st, 2nd, 3rd and lover priority.
Students' dropout was defined, as a student who registered for a study programme, but leaves the University during the first 15 study month. As we can see from the Fig. 1 – 47 students leave the faculty and their study time was between 2 and 12 month.

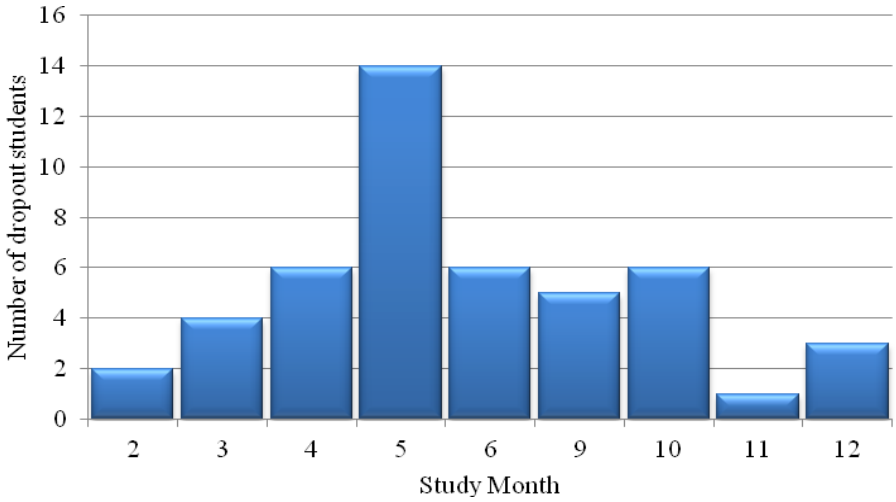

Fig. 1. **Distribution of drop out students study duration (n=47)**.

Study time is a quantitative variable, but the distribution of study time is not normal. Therefore we can't use parametric statistical methods, such as t-test, ANOVA or linear regression for data analysis. Survival analysis methods will be used for study time data and information about censored and uncensored data will be including in the analysis.
Information about censored and uncensored data is reported in Table 1.

Table1

Characteristic of the categorical variables for the whole sample based on number of individuals (n=91)

Variables	n	Study till now, Censored	Dropout, Uncensored
Total	91	44 (48.4%)	47 (51.6%)
Study program			
1 – Cs	61	29 (47.5%)	32 (52.5%)
2 – Pr	30	15 (50.0%)	15 (50.0%)
Gender			
1 – Male	76	34 (44.7%)	42 (55.3%)
2 – Female	15	10 (66.7%)	5 (33.3%)
Priority(n=85)			
1st	56	29 (51.8%)	27 (48.2%)
2nd	8	2 (25.0%)	6 (75.0%)
3rd and lover	21	12 (57.1%)	9 (42.9%)
Mark group			
≤ 25	12	3 (25.0%)	9 (75.0%)
26-35	58	24 (41.4%)	34 (58.6%)
≥ 36	21	17 (81.0%)	4 (19.0%)
Finance			
1 – budget	55	25 (45.5%)	30 (54.5%)
2 – self-finance	36	19 (52.8%)	17 (47.2%)

Study program: 1 – Cs (Computer Control and Computer Science); 2 – Pr (Programming)

Students with a higher probability of dropping out are those who started faculty with school mark 25 and lower (75%) and students with school mark in range 26-35 (58.6%), than students with higher mark. Male students have the highest rates of leaving the faculty.
51.8% and 57.1 % of the students with study priority 1st and 3rd and lover still study at the faculty. Students with the 2nd priority have the highest dropout rate and now only 25% of students from this group are study at the faculty.
For students drop out rate causes the following Survival analysis methods were used:

- Kaplan-Meier was used to compare two groups on their survival time;
- Log-rank test was used to compare two and more groups survival curves and
- Proportional hazard model to determine whether factors influence student dropout.

The proportional hazard model (Cox model) can be written as:

$$h_i(t)=[\,h_0(t)]\,e^{(b0+b1xi1+\,b2xi2+\,b3xi+\,b4x2+\,b5xi5)}, \quad (2)$$

where $h_i(t)$ – the hazard rate for the ith case at time t
$h_0(t)$ – the baseline hazard at time t
bj – the value of the jth regression coefficient
x_{i1} – gender (1 male, 2 female)
x_{i2} – study programme (1 Cs, 2 Pr)
x_{i3} – finance source (1 budget, 2 private)
x_{i4} – priority to study in the program (first, second, third and lower)
x_{i5} – secondary school marks (covariate)

Factor levels were compared using a levels of significance $\alpha=0.05$. Statistical analyses were carried out with the program IBM SPSS Statistics 20, IBM, New York, USA (Chan, 2004).

Results and discussion

Kaplan-Meier and Log-rank test

The students drop out rate can be affected by different factors. Each factor independently was analysed by long-rank tests. Summary results of Log-rank test show there are no differences between survival curves for factors study program, gender, priority and finance (Table 2).

Table2

Long-rank test statistics for equality of survival distributions for groups

Variables	Chi-Square	df	Significance
Study program	0.001	1	0.977
Gender	2.013	1	0.156
Priority	4.058	2	0.131
Mark group	11.891	2	0.003
Finance	0.841	1	0.359

Students' secondary school marks were in range 22-48 point. For this analysis the school marks were divided to three groups: 1st mark group ≤ 25 point, 2nd mark group 26-35 point and 3rd mark group ≥ 36 point. There are significant differences between survival curves for factor school mark group ($p<0.05$).

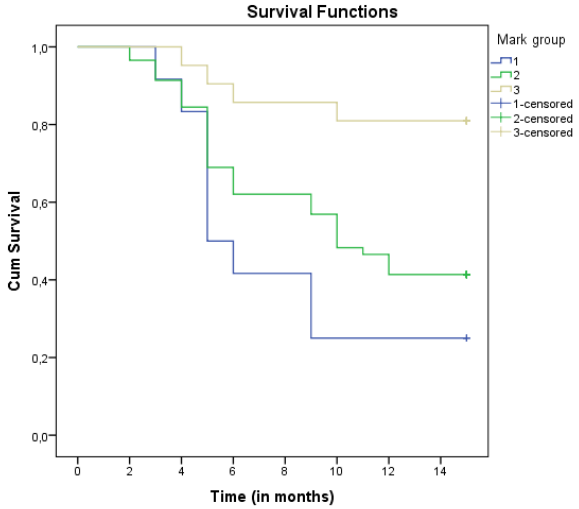

Fig. 2. **SPSS survival plot for factor school mark for three group comparison:** 1st, 2nd and 3rd mark group.

Students who complete their school study with the maximum mark (≥ 36 point), have the higher probability of completing their study, whereas those who have lower marks, between 26 and 35, have the lower probability of surviving, i.e. the lower rate of taking the degree and therefore the higher chance of leaving the faculty (Fig.2).

Proportional hazard model (Cox model)

For presentation the realistic situation when all factors are included to theoretical model the multivariate analysis is more preferable than one factor analysis. For this purpose all factors have investigated together. Main effect model with 5 factors were used for evaluation which factors are significant in students surviving (Table 3). Study program, gender, priority and finance factors were included in the Cox model as categorical covariates (qualitative factor) and school mark as covariate or quantitative factor. As the result finance and priority factors are statistically significant ($p<0.05$) despite to the Log-rank test results. By backward stepwise method the not significant factors study program were exclude from the model.

Table 3

Main effects model by enter and backward stepwise methods (n=91)

		B	SE	Wald	df	Sig.	Exp(B)
Enter method	Gender	0.900	0.561	2.572	1	0.109	2.460
	Study program	0.471	0.386	1.487	1	0.223	1.601
	Finance	1.067	0.374	8.121	1	0.004	2.907
	Priority			5.988	2	0.050	
	Priority (1-2)	1.142	0.477	5.747	1	0.017	3.134
	Priority (2-3)	0.913	0.588	2.409	1	0.121	2.492
	School Mark	-.175	0.047	14.027	1	0.000	0.839
Backward Stepwise method - Step 2	Gender	0.840	0.559	2.254	1	0.133	2.316
	Finance	0.999	0.376	7.084	1	0.008	2.717
	Priority			5.301	2	0.071	
	Priority (1-3)	1.115	0.484	5.293	1	0.021	3.048
	Priority (2-3)	0.650	0.552	1.390	1	0.238	1.916
	School Mark	-.178	0.048	13.864	1	0.000	0.837

Results in Table 3 (step2) show that there are significant differences for finance group ($p<0.05$), priority ($p<0.1$) and school mark ($p<0.001$). A positive sign of coefficient b means that the hazard rate or risk of student's dropout is higher for the first group to compare to the second group and prognosis for that group is worse. Male students, students with budget finance and higher priority are associated with poorer survival, whereas being female; students with self-finance and lower priority are associated with better survival.

The estimated hazard rate for male (coded 1) is exp(0.84) = 2.316 of that of the female; that is, a male dropout risk will be 2,3 times higher than for female after adjustment for the other explanatory variables in the model. However, the p-value = 0.133 is not statistically significant, no difference in survival.
Students with first and second priority are at higher risk than students with third and lower priority to be dropout (HR 3.048, p<0.021; HR 1.916, p<0.238).
A negative sign of coefficient b means that the hazard rate (risk of dropout) is reduced. Students with higher school mark are associated with better survival, whereas students with low mark.
School mark variable the regression coefficient refers to the decrease in hazard rate for an increase of 1 in the value of the school mark. The estimated hazard or risk of dropout decreases by 100%-(100%*0.837) = 16.3% for a one mark unit.
At IT faculty study programme curricula are included such topic us mathematic, physic and chemistry and it is influence the dropout among students. The reasons for students dropout is students' poor knowledge in Mathematics, Physics and Chemistry and poor pre-college academic qualification.
Fig. 3 and 4 displays the estimated survival function of a hypothetical student in interval 0-12 month of study according to different priority (1 to 3) and finance (1, 2). Students are at a very low risk of being dropout at the beginning of the study. The dropout risk slowly increases in the 2nd and 3rd month of study but becomes quite high before the session at the 5th and after session at the 6th month of study.

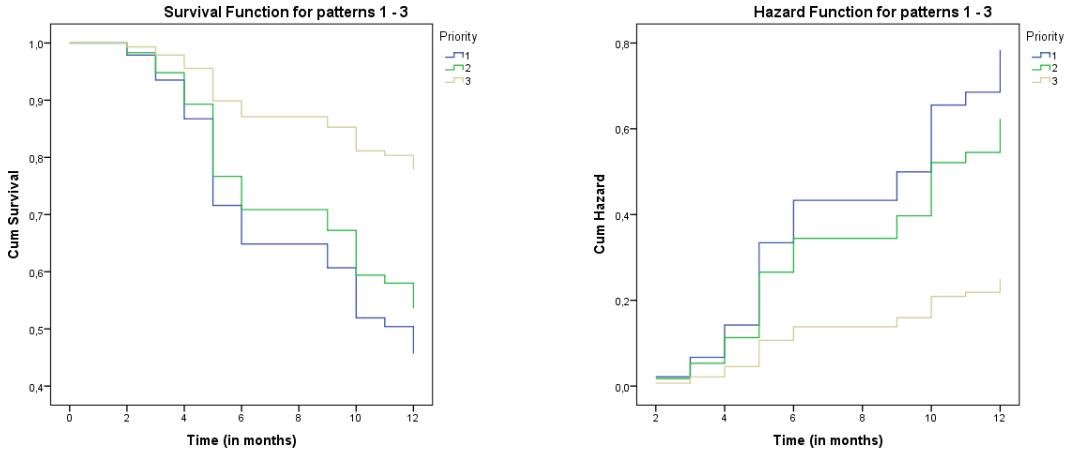

Fig. 3. **SPSS Survival and Hazard plots for factor priority for three group comparison:** 1st, 2nd, 3rd and lover priority.

Therefore, students who start faculty with 1st and second priority have a greater probability of taking the degree than those who decide to study at the faculty with 3rd and lover priority. 27 first priority students (48.2%), 6 second priority students (75.0%), and 9 third and lover priority students (42.9%) are dropping out from the faculty during the first study year. Second priority students have higher hazard curves because, as we have seen from their regression coefficients, they have a greater potential to dropout. During the first 6 month the higher risk to dropout have students with low notes (<25) and second priority.

Table4.

Average school mark in different priority groups
(Standard deviation in brackets)

Priority	Mark - study	Mark - don't study
First	37.5 (6.33)	31.6(3.46)
Second	28.3(0.85)	27.2(3.18)
3rd and lover	28.3(4.99)	27.1(4.47)
In total	34.5(7.22)	30.0(4.18)

The students with first priority have the higher school mark, than students with second and 3rd and lover priority; therefore they have higher risk to drop out. That results show the students with 1st priority and higher school mark leave the faculty (Table 4). The average study duration was 6.9, 4.7 and 6.3 months for the 1st, 2nd, 3rd and lower priority group, respectively.

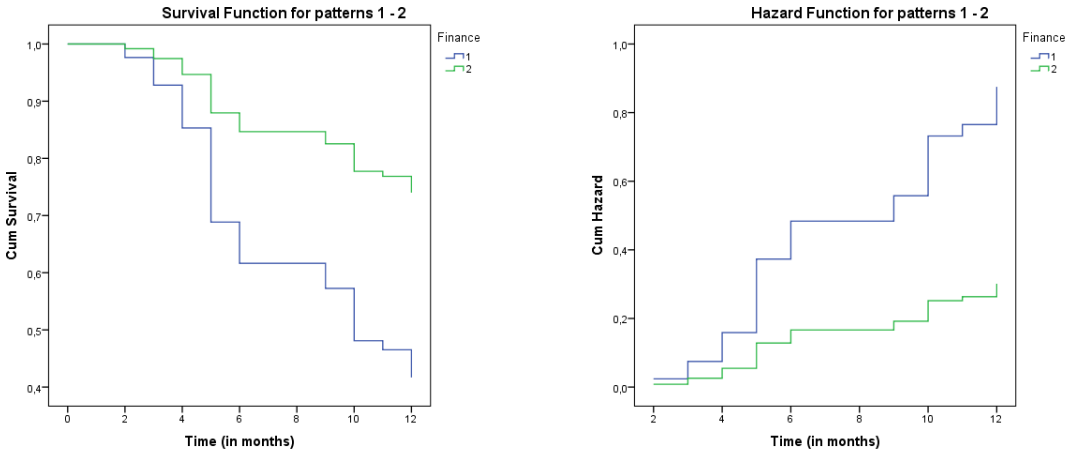

Fig. 4. **SPSS survival and Hazard plots for factor finance for two group comparison.**

According to survival plot can be noted the budget students have a lower survival rate that self-finance students. 30 budget students (54.5%) and 17 self-finance students (47.2%) are dropping out from the faculty during the first study year.

Table5

Average school mark in finance groups
(Standard deviation in brackets)

Finance	Mark - study	Mark - don't study
Budget	36.7 (7.49)	30.8 (4.05)
Self-finance	31.2 (5.78)	27.7 (4.71)
In total	34.3 (7.27)	29.7 (4.50)

The students their finance source gets according to their mark. From Table5 we can see the students, who leave the faculty from the budget group is 30.8 and higher than for self-finance students. The average study duration was for budget students 6.1 months and for self-finance student 6.3 months.

Not only finance factor, priority and school mark can affected the students' dropout rate there are other reason why student leave university.

In literature several factors are associated with student dropout in higher education institutions (Min et al., 2011; Murtaugh et al., 1999): factors associated with attributes or characteristics of the individual student, and factors associated with the institutional environment. When we have students who leave faculty with good mark and from budget, we should analysis the institution environment and topic which studied during the first study year.

Conclusion

1. The results of this study show finance group ($p<0.05$), priority ($p<0.1$) and school mark ($p<0.001$) factors are the main causes for students' dropout at the Faculty of Information Technology.
2. Kaplan-Meier, Log-rank test and Proportional hazard model can be used to evaluate of students' dropout causes.
3. All important factors should be included to the theoretical model for presentation the realistic situation and for this case the multivariate analysis is more preferable than one factor analysis.
4. Data from different study years are recommended to include for further investigations of students dropout rates.

Acknowledgements

Our grateful thanks to Sandra Sproge for her contribution in collecting the student dropout and school marks data to this publication.

References

Bewick, V., Cheek, L. and Ball, J., 2004. Statistics review 12: Survival analysis. *Critical Care*, 8, pp. 389–394. Available at: http://ccforum.com/content/8/5/389, 18.01.2013.

Chan, Y.H., 2004. Biostatistics 203. Survival analysis. *Singapore Medical Journal*, 45(6), pp. 249-256.

Hosmer, D.W. and Lemeshow, S., 1999. Applied Survival Analysis. *New York – John Wiley and Sons*. 400 p.

Kaplan, E.L. and Meier, P., 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), pp. 457–481.

Klein, J.P. and Moeschberger, M.L., 2005. Survival Analysis. Techniques for Censored and Truncated Data. *Springer*. 536 p.

Min, Y., Zhang, G., Long, R.A., Anderson, T.J. and Ohland, M.W., 2011. Nonparametric Survival Analysis of the Loss Rate of Undergraduate Engineering Students. *Journal of Engineering Education*, Vol. 100, No. 2, pp. 349–373 Available at: http://www.jee.org. 18.01.2013.

Murtaugh, P.A., Burns, L.D. and Schuster, J., 1999. Predicting the retention of university students. *Research in Higher Education*, 40(3), 355–371.

Smith, T. and Smith, B., 2001. Survival Analysis and the Application Of Cox's Proportional Hazards Modelling Using SAS. *Statistics. Data Analysis and Data Mining.* Paper 244-26. Available at: http://www2.sas.com/proceedings/sugi26/p244-26.pdf , 18.01.2013.