Master Thesis Data Science and Marketing Analytics

Predicting the factors influencing student dropout for the program Business IT and Management Using Survival Analysis

Supervisor: Prof. dr. Patrick J.F. Groenen
Co-reader: dr. Radek Karpienko
Student: Hassan Alaoui (366875)

May, 2021
Erasmus University Rotterdam
Erasmus School of Economics

Abstract

This study aims at identifying the factors influencing school dropout for the most vulnerable students who are at the beginning of their study. The consequences of student dropout are also significant for universities and for the whole society. Therefore, it is crucial to identify ‘at-risk’ students at an early stage to be able to address this problem effectively. This study is conducted in the context of an internship by the Business IT & Management (BIM) program by using five student cohorts from 2015 to 2020.

The statistical technique used in this study is called survival analysis (SA). The SA method used is the semi-parametric Cox Proportional Hazard Model. The advantage of this method in comparison to the classical methods like linear regression and logistic regression is the possibility of modeling survival times or time to event and censoring data. By using the survival data extracted from OSIRIS which is a student tracking system used by the most universities in the Netherlands, we build different models including demographic variables and the grades of the courses obtained during the first year.

After fitting the Cox Proportional Hazard Model to these data using six models, we evaluate how well each model performs and the variables affecting the dropout. The results showed that the model estimated after applying the stepwise method based on the Akaike Information Criterion on the demographic variables combined with the grades of the first quarter is the best model according to the concordance index. The selected model shows that to be able to tackle the dropout rates of the program Business IT & Management, the university should address the effort to the younger students born outside Rotterdam city and obtaining a lower grades for the courses of the first quarter.

Keywords: Student dropout, longitudinal data, Survival Analysis, Cox Proportional Hazards Model, event prediction, regression.

Acknowledgements

This thesis has been written as part of the study program 'Economics and Business' with a specialization in 'Data Science and Marketing Analytics' at the Erasmus University of Rotterdam. The internship has been done at the Business IT & Management department of the Rotterdam University of Applied Sciences.

I would like to thank my colleagues at the Rotterdam University of Applied Sciences, Phyllis Rietdijk, Mohamed Ziani, Esther Heemskerk, Lex Freund, Dominique Booms for providing me a lot of guidance and advice and giving me all information that I need to finish my Master's degree. They always took the time to share their experience with me and to help me shape this Thesis. I also would like to thank my supervisor from the Erasmus University of Rotterdam, Patrick Groenen for his patience and understanding of my personal situation. His way of supervision and feedback offers a lot of flexibility to accomplish the task. Lastly, I would like to thank my wife, my children and my parents for their unconditional support.

Contents

1 Introduction

High student dropout rates is becoming one of the major problems in higher education. Cohen (2017) argues that the challenge for most institutes of higher education and universities is to develop a broader vision on how the retention of the students can be increased. Further, he proves that there is a considerable need to understand how resources provided by the government can be used effectively and efficiently to improve the student learning experience and to increase their academic performance. "Prevention is better than cure" is one of the major solutions that can be used to support the student at risk. In addition, there are many factors such as economic, demographic, social-cultural and the background of the students that may influence their school performance. Unfortunately, these factors fall outside the scope of universities, and they cannot be manipulated.

However, school-related factors can be directly affected by the universities like the design of educational curricula, the quality of education or the development of Learning Management Systems (LMS). In addition, as the cost of storage and processing power is decreasing, the storage of the data has become affordable and more accessible. As a result, that the educational data collected is immense, and the volume is growing exponentially (Decker & Lenz, 2007). Thus, the technological revolution makes it possible for universities to collect a large volume of data about the learning behavior of their students and their performance during their studies. However, other authors argue that the availability of the data does not mean that the universities are allowed to use all this data. Rubel & Jones (2016) goes even further and argues that the legislation about stakeholder privacy can be seen as a barrier which can complicate the task. The law protects the privacy of the individuals; and this is the reason why universities are limited in their possibilities of using the data.

The internship of this Master's thesis took place on the program Business IT & Management (BIM), a part of the Hogeschool Rotterdam Business School (HRBS) which is a part of the Rotterdam University of Applied sciences (RUAS). This university consists of nine institutes, five knowledge centers, two expertise centers and four staff departments. The institutes are responsible for providing education to almost 40,000 students in the region of Rotterdam, and the number of employees, including lecturers, is more than 3,000. The other departments facilitate the institutes to achieve their goals. RUAS is a dynamic knowledge enterprise for higher professional education which offers 90 degree programs, spread over various locations in the city of Rotterdam, in the fields of Economy, Management, Education, Behavior and Society, Healthcare, Technology, Media & ICT and Art. Education is offered in a practical, thoughtful and flexible way and the students work on real and innovative projects where they can develop a result-oriented approach and their ability to think critically.

Despite the fact that many social and economic changes are taking place in society, there have been a few educational innovations in recent years. This is one of the reasons why RUAS is still suffering from a high dropout rates of students. Another reason is the rapid increase in the number of students of some programs which did not have time to manage the dropouts. Most curricula are outdated, not digitalized and do not take care of the needs of the individual students.

There are many students who, due to the current industrialized education system, have little or no chance of being adequately supervised. Some students leave the program voluntarily because they find out late that the program does not suit them. Another group of students cannot achieve the binding norm of 48 credits. The first year is considered as a selection phase. But if the percentage of students failing reaches a percentage of more than 30%, and more than 45 percent by the Business IT & Management degree, then something must be done to address this problem. Fortunately, institutes of higher education have realized in recent years that they have to do something to reduce the dropout rate of students and to encourage study success. Dropping out in the primary phase can occur because there is no match between the student's expectations and the study reality or because there is no substantive match between the student and the study program. However, these students fall out the scope of this research.

The study success committee at RUAS conducted last year qualitative research to advise the institutes on the measures to be taken. The research of this Master's thesis will be conducted at a micro level and will advise the individual students about their chance to succeed in the Business IT & Management degree, whether they will survive the first year and when they are at risk of dropping out.

The central question of this research that will be answered is as follows:

"How can freshman student dropout rates of the Business IT & Management degree be modelled using Survival Analysis?"

This research aims to gain insight into the profile of Business IT & Management students who are at high risk of dropping out during the first year of the Bachelor's degree. By conducting this research, a target group can be reached and receive help from dedicated coaches or tutors. The expectation is that this study can be applied to all programs of the RUAS in order to reduce dropout rates.

The remainder of this thesis is structured as follows. The next chapter analyzes the available literature on the student dropout. Chapter 3 explains the data and the different variables provided by the university administration, gives descriptive statistics and explains the data cleaning process. Chapter 4 describes the Survival Analysis. Additionally, the Cox Proportional Hazard model is elaborated. Chapter 5 presents the results of the analysis. finally, Chapter 6 gives a conclusion and discusses the limitations of this study.

2 Literature review

In times of scarce resources, both educational institutions and students are interested in an adequate solution for high dropout rates (Baars & Arnold, 2014a). Higher education needs to be engaged to help students obtain their degree without study delays and to not leave their study before receiving a diploma. In addition, Students must enter in a discipline and an academic level where their chances of success are highest, and future career prospects are guaranteed (Baars & Arnold, 2014a). The following review of literature explains the student dropout problem in higher education from a theoretical and empirical point of views, discusses specific and general solutions, and concludes that specific initiatives are needed to reduce the number of dropouts both now and in the future.

Student dropout rates have long been a serious problem in higher education that deserves serious attention, and the extent of this problem has changed over time (Ben-Tsur, 2007; Horstschräer & Sprietsma, 2013). The current industrialized educational system is the product of the increasing number of students who are participating in academic courses and the interpersonal instructor-student interactions. The increasing numbers of students has made it very difficult for educators to detect high-risk students in the early stage. Most universities discover during the final evaluation of the courses that students drop out, which makes it impossible to take some measures to help this group of students (Cohen, 2017).

Cohen (2017) argues that several strategies must be developed to make it possible for educators and decision-makers to identify at-risk students at the right moment and to support their learning process before they drop out. This way of looking at the problem serves as a basis for the creation of an appropriate system; which will enable those students to profit from the right assistance from their tutors and lecturers and to be able to complete the courses successfully.

In addition, the development of the learning management systems (LMS), make it possible to collect a large amount of data related to the learning processes and to study the behaviour of the students. This amount of data can be generated automatically by the LMS and can be used directly by the lecturer to have an idea about how students are learning and which of them need some extra help to pass the course (Johnson, Becker, Estrada, & Freeman, 2014). Further, a prediction process can be automated and integrated into the LMS in the form of dashboards, similar to those used in business intelligence to make a decision at a company. These dashboards can display a different kind of information about the needs of students which can be customized according to the specific needs of each student. Furthermore, it can also be used to predict potential failures and dropouts (Macfadyen & Dawson, 2010).

Cohen (2017) explains that a detailed understanding of the reasons why students may leave a specific course is the first part of the story, the more important part is allowing the lecturers to gain more insight into the efficiency of their course material and the way in which they are teaching. Furthermore, lecturers will be able to discover some hidden patterns which can be improved in their didactic approach, teaching skills and their curriculum.

The dropout phenomenon can be explained from different angles. The most useful perspective is the cost-benefit one which examines the profit-loss balance from the institutional-social point of view. In this case, the investments made to educate the students who drop out can be seen as loss when the students leave the university without obtaining any degree. From a psychological perspective, there are many effects on the self-image of the students and their social status (Levin, Barak, & Yaar, 1979). Furthermore, there are many forms of dropout, students can leave a specific course, a whole degree, but they can also leave their study program in a certain year. In this research, the first-year will receive the most focus because the dropout rate is mostly very high and the higher the academic year, the lower the dropout rate.

Different theoretical models have been developed to explain the factors causing students dropout (Cabrera, Nora, & eda, 1992; Dey & Astin, 1993). In 1975, Tinto proposed one of the most influential models describing all aspects that can cause student's dropout from higher education. In his conceptual model, he divides these factors into three categories: the first ones are related to the educational system, the second to the students' personal characteristics and the last to all the elements of the outside community such as friends, family and the economic context. Furthermore, other attempts have been made to achieve more in-depth understanding of the phenomenon (Hovdhaugen, 2011; Lee & Choi, 2010; Ortiz & Dehon, 2013) including essential factors such as the characteristics of students who are academically at risk as potential dropouts (Breier, 2010). Cheng (2013) explains that there are characteristics related to higher education system that encourage dropout. He argues that the problem in higher education is related to the implementation of strategies to increase the students' retention. However, fewer initiatives were conducted to identify a dropout student on early-stage (Grau-Valldosera & Minguillón, 2014).

In addition to the theoretical models developed to understand the dropout phenomenon, different empirical studies using regression analysis has been widely elaborated in this area (Dey & Astin, 1993) and to calculate the probability of dropout, Logistic regression has also frequently been used in this domain (Baars & Arnold, 2014b; DeBerard, Spielmans, & Julka, 2004). Luna (2000) used regression tree, logistic regression and discriminant analysis to address this issue. Other authors have developed models based on logistic regression to recognize the freshman at risk of attrition (Glynn, Sauer, & Miller, 2011). However, the models mentioned before are not able to integrate longitudinal information and results produced are suboptimal (Ameri, Fard, Chinnam, & Reddy, 2016). Predictive analytics have been used in different industries for several years, but higher education can be seen as a late adopter of these techniques as a tool to support decision making (Barneveld, Arnold, & Campbell, 2012).

In the last ten years, different research studies have been conducted using machine learning algorithms to understand the parameters influencing the performance of students during their study. A decision tree algorithm is used by Khan (2019) to develop a model that allows students to predict their final grades in a programming course. Other researchers used the data provided by Moodle, an open-source Learning management system, to develop a personalized multiple linear regression model that can help to predict student performance at the University of Minnesota (Elbadrawy, Studham, & Karypis, 2014). Further research conducted to predict the academic performance have used decision tree models based on the exam scores (Hamsa, Indiradevi, & Kizhakkettottam, 2015) or applied neural networks to estimate the final grades based on the first year's grade (Arsad, Buniyamin, & Manan, 2012). Compared to the traditional statistical methods shown in the previous paragraph, support vector machine and boosting methods shows higher accuracy in detecting dropout (Zhang, Oussena, Clark, & Kim, 2010).

The call to action was needed by researchers who began to study the survival analysis, a slightly complicated but relevant technique which aims to study longitudinal data where the outcome variable is a combination between time and another variable. For example, Radcliffe et al.(2006) conducted a study to identify a student at risk utilizing survival analysis to study student-athlete attrition. The same technique was used by Ishitani (2006), who studied the attrition and degree of completion behaviour among first-generation college students in the United States. Increasing use of technology in combination with machine learning techniques, must stimulate lecturers, managers and educators, in general, to make use of those techniques to help their students in their studying process and to deal with the transition from secondary to higher education.

In addition to using machine learning techniques, different additional measures can be taken to help students choose a suitable degree, by organizing different transition programs from secondary to higher education (Thornsberry, 2010), creating learning communities, providing mentors that can be available to support students at the right moment (Pagan & Edwards-Wilson, 2002). A combination of these measures and a reform of educational curricula may lead to positive effects. Kristin (2016) adds that students have a higher chance of success at universities and universities of applied sciences when they do well in their first year of education. Previous studies has shown that there is a positive correlation between student achievement in the first months of attendance and the following achievements (Murtaugh, Burns, & Schuster, 1999).

Because this study will be conducted in a Dutch context, it is important to mention that there are some specific variables that will play a significant role in the decision of dropping out. The first one is related to the fact that students are obliged to obtain a certain number of European Credit Transfer and Accumulation System (ECTS), a minimum number of points that allows them to continue their study beyond year one. The second one is related to the study grants; students are allowed to transfer to a different program in the first months without losing their grants. Taking into account the aforementioned restrictions, it appears evident that the predictions on whether students have a good chance of completing the program need to be made at the beginning of the first academic year.

This research aims to identify which students will drop out in the very first months of the program, when they will drop out and also on which likelihood of dropping out. Early identification would be very beneficial in the sense that students can receive advice whether the program is suitable for them or perhaps it is preferable to switch to a different programs. The students who are highly motivated can then be offered a short remedial support program (Baars & Arnold, 2014b).

3 Data description

The current thesis used five datasets extracted from OSIRIS which is a student tracking system used by most universities in the Netherlands. In this study, a dataset was compiled by tracking 930 students enrolled at the bachelor Business IT & Management (BIM) starting from the school year 2015-2016 until 2019-2020. Among those, there were respectively 49%, 53%, 40%, 45% and 45% dropout rates at the end of every first academic year. The freshman will get high attention in this study because the dropout rates are very high during the first year. There is a different kind of dropouts, students who think that they have a lower chance of obtaining sufficient credits to continue, seems to leave their study before the first semester, other students have hope to increase their performance during the second semester. However, they did not obtain credits higher than 48 ECTS to be admitted in the second year. They must, unfortunately, leave the bachelor according to the university policy. Thus, the dropout is defined in this study when a student does not register in the remaining courses of the second year. However, there is a part of students who obtain lower than 48 credits but continue their study due to personal circumstances. This group is not relevant because it is too negligible and will not affect the final results of this research. At the time of the registration and during the study, several personal information, grades and the number of ECTS obtained, is collected about the students. This information has been pseudonymized according to the privacy legislation in such a way that it is no longer possible to trace the individual students.

The next section will be dedicated to the source of the data. Section 3.2 explains the data cleaning process. In Section 3.3, the description of the variables will be presented, and the distribution of the essential variables will be visualized in graphs. Section 3.4 gives a brief explanation about dealing with missing data.Lastly, section 3.5. will explain the right censoring.

3.1 Osiris dataset

The Osiris data is extracted in the form of an Excel sheet. Every cohort is printed separately and the rows contain the observations which in this case represents the students. The columns show different variables about the students. The initial data consists of 38 variables, which could be categorized into four different clusters: demographic, grades, number of ECTS obtained and the date of the last obtained results. After removing all variables used to manipulate the data like the dates and the variables containing no variation. The number of variables that will be used in the analysis is 30.

3.2 Data cleaning process

From the 930 students selected in the initial data composed of 5 cohorts, 162 observations are removed because they were empty. See figure 1 for the data cleaning process. Some students make the wrong choice of bachelor degree and stay registered in the record system of the university. Some observations contains a negative value of the variable “Time” which makes it impossible to run the analysis; these three observations are deleted. Hence the dataset that will be used contains 765 observations.

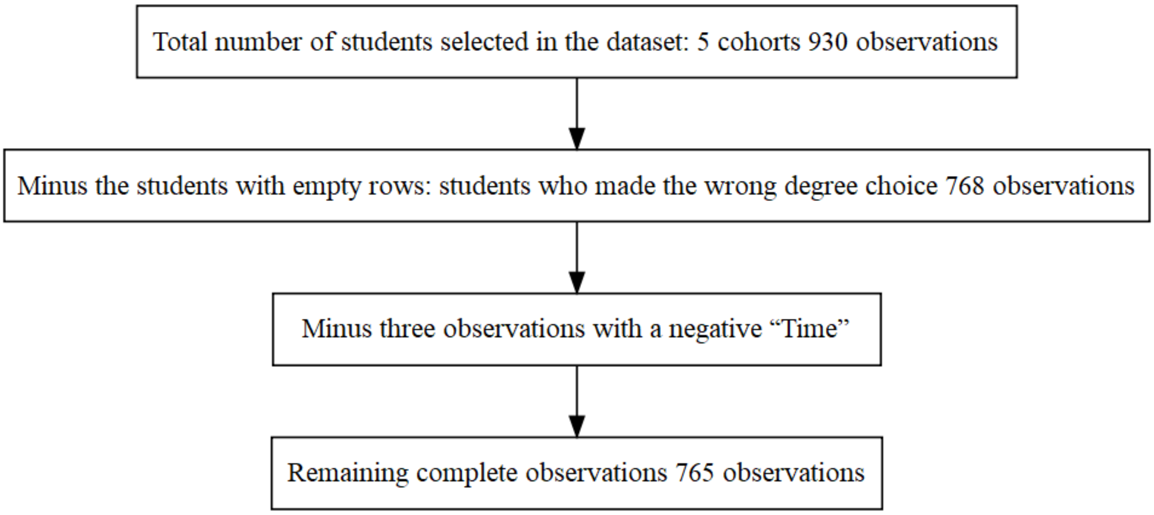

Total number of students selected in the dataset: 5 cohorts 930 observations
Minus the students with empty rows: students who made the wrong degree choice 768 observations
Minus three observations with a negative “Time”
Remaining complete observations 765 observations

Figure 1: Data cleaning process

3.3 Data characteristics

The dataset used in this research is composed from continuous and categorical variables. As mentioned earlier, 930 students of five consecutive cohorts are the observations of this study. A small part of the observations are women (13.5%) because it seems to be challenging to attract women to an IT-program. The average age is 22.3 years and the average number of credits obtained during the first year are 38.4 ECTS. The grades are measured at a scale between 0 to 10. A higher grade than 5.5, means that the student has passed the course successfully and can get the corresponding ECTS. The information concerning demographics is available in the variables “Country_of_birth,” “Birthplace” and “City.” It was not possible to leave additional information concerning demographics for privacy reasons. Most students are living outside Rotterdam and are born in the Netherlands, and 11.7 % of the students are born abroad, and the common ethnicities are Surinamese, Turkish, and Moroccan. Regarding the dropout students, 51.6 % of the studied population left the bachelor in the first year spread over the year.

The Business IT & Management program aims to educate students to become a chief information officers. The knowledge and the skills teached during the the first academic year are divided into three categories. First, Management courses which emphasize the understanding and application of the theoretical models, for example, the students learn how to use the Porter’s five forces model to analyse the competition in certain market or how to apply the SWOT-analysis (strength, weaknesses, opportunities and threats) to formulate a strategy of a company. Second, the Information and Technology courses where students learn how to make a website, how to store data in the information systems and to select the data needed to develop a Business Intelligence tool. Third, the Skills Courses are intended to help students make a reliable and a valid research but also to develop the needed skills to be able to operate efficiently in their future job and to facilitate the transition between school and their future career. The knowledge that the students acquire during the theoretical and skills courses is integrated on the basis of the project courses.

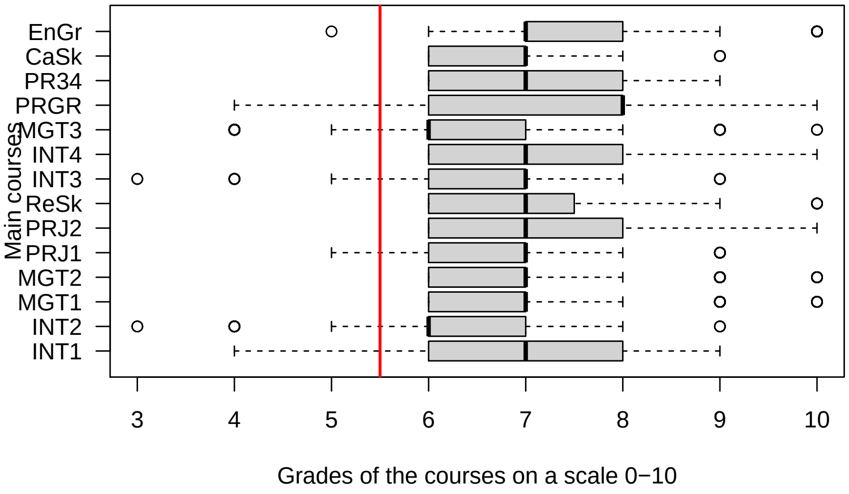


Figure 2: Boxplot showing the grades of all courses

Table 1 shows the description of the variables used in the analysis.

Table 1. Description of the variables (N=765)

Variables Description

1.**Pseu**: Student Pseudonym

2.**CoBi**: Country where the student is born; Netherlands=1 , outside the Netherlands=0

3.**BiPl**: City where the students is born; Rotterdam=1, outside Rotterdam=0

4.**Gend**: Sex; 1=male, 0=female

5.**City**: City where the student lives; Rotterdam=1, outside Rotterdam=0

6.**OCFY**: The number of credits (ECTS) obtained during the first year

7.**FCOb**: Obtaining 60 credits (ECTS) during the first year; Yes=1 , no=0

8.**OCOb**: Number of Credits (ECTS) obtained outside the program

9.**WAVG**: Weighted average for the grades of all courses

10.**NCFS**: Number of credits (ECTS) obtained during the first semester

11.**NCSS**: Number of credits (ECTS) obtained during the second semester

12.**INT1**: Informatiom Technologie course of the first quarter

13.**INT2**: Informatiom Technologie course of the second quarter

13.**INT3**: Informatiom Technologie course of the third quarter

14.**INT4**: Informatiom Technologie course of the Fourth quarter

15.**MGT1**: Management course of the first quarter

16.**MGT2**: Management course of the second quarter

17.**MGT3**: Management course of the third quarter

18.**PRJ1**: Project course of the first quarter

19.**PRJ2**: Project course of the second quarter

20.**PR34**: Project course of the third and fourth quarter

21.**ReSk**: Research skills course first semester

22.**PRGR**: Programming skills course third quarter

23.**CaSk**: Career skills course

24.**Elec**: Optional courses outside the program

25.**EnCr**: English course credits (ECTS) obtained; 1=obtained , 0=not obtained

26.**EnGr**: Grade of the english course

27.**Age**: Age of the student obtained from the birthdate recorded in years

28.**Time***: Number of days before dropping out obtained by substracting the date of the last result from the beginning of each academic year

29.**Status****: Dropped out=1 ; Censored=0

*Timing variable **Response variable

3.4 Missing values

The dataset is composed from many observations with missing values, it is ubiquitous in survival analysis to see this kind of data sets, these missing values are mostly caused by the dropout or when students are sick or absent, such as the highest one which is the variable “Career_Skills” (89.1% missing). The information and technology (IT) courses seem to suffer a lot from the missing values with the highest percentages of mostly 50 %. The management courses take second place with on average 40 % of the missing values. The projects of the first semester seem to have a low missing value namely (8.7%) for “Project1” and (17.1%) for “Project2.” In the second semester, the percentage of the missing values increases to 38.5 %. These values will be replaced by the value “0” which means that the students did not complete the course and get an insufficient grade.

3.5 Right censoring

A very important concept in survival analysis is called censoring which represents the observations where the exact time of the event is unobserved. There are three well-known types of censored observations: left-censored, interval-censored and right-censored. This last one can be explained by a lost in follow-up of a certain subject (Kleinbaum & Klein, 2012). In the case of our research, the students that will drop out after 325 days or after the first academic year are not considered into account. In other words, The time of event is not observed because it will take place after the academic year or outside the studied period. In the scheme below, Subject 22 is censored because it is not known if he dropped out after 325 days. However, subject 3 can be seen as dropout.

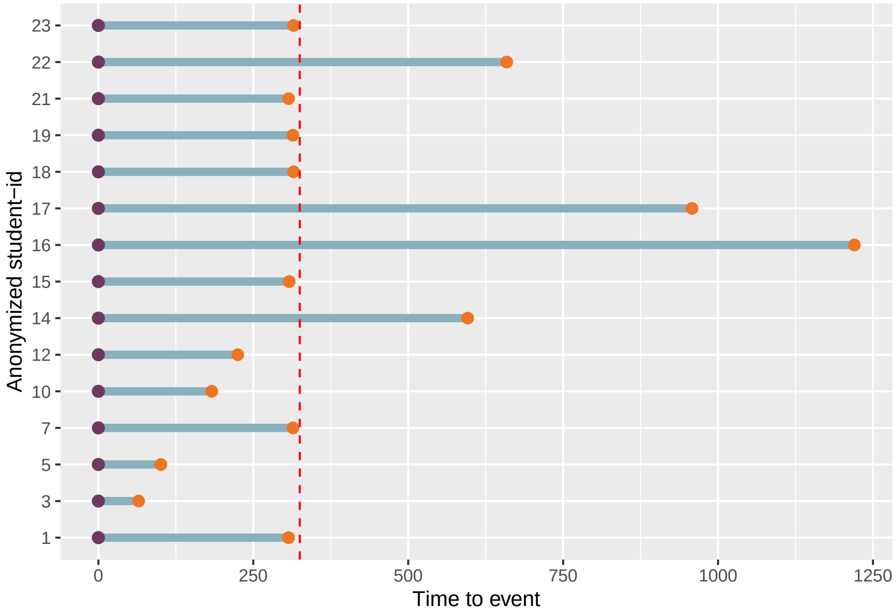


Figure 3: Experience of students with the same entry time and follow-up during the first academic year

In this research, students are followed during their first bachelor year and more precisely in the first 325 days of their study. The event in this case is the dropout. The students who dropped out will be assigned a value of “1” and the students who are still studying are assigned the value “0.” These observations are called censored and they are followed during the research but it is not known if they dropped out after the first year.

4 Methods

This section is concerned with the statistical techniques and methods applied in this thesis. Since the outcome variable is the “time to event” and there may be censored data. The data is analyzed with the survival analysis. A special case of Survival analysis is the Cox Proportional Hazard (CPH) model which accounts for the high number of covariates in the data. The assumptions of the CPH model are tested and the accuracy of the model will be evaluated. Lastly, the model performance will be assessed using the Akaike information criterion(AIC).

4.1 Application of the Survival Analysis

Survival analysis is a statistical technique where an outcome is modeled until an event. This technique is used in different domains. For example, in the medical sciences, Survival analysis is used to predict the time to death of cancer patient or to predict waiting time before a certain surgery. In different industries, survival analysis is used to investigate when machines needs maintenance. Furthermore, it is applied in the labor market to predict how long workers will be unemployed. The list of the domains where the survival analysis has been used is long (Moore, 2016). In the case of this research the event of the survival is defined as the dropout of the student.

4.2 Basic principles of Survival Analysis:

Survival models, discussed in details by Kleinbaum and Klein (2012), relate the time that passes, before some event occurs, to one or more covariates that may be associated with that quantity of time. Survival analysis methods depends on the survival distribution that can be specified with the survival function and the hazard function. The survival function represents the probability of surviving up to a specified point of time t. The Survival distribution has the following probability function,

$$S(t) = Pr(T > t), 0 < t < \infty.$$

This function decreases or remains constant over time and takes the value 1 at time 0, never drops below 0, and is also right continuous. The survival function $S(t)$ is often defined in terms of the hazard function, which is the instantaneous failure rate. The survival function is the probability that, given that a subject has survived up to time t, he or she fails in the next small interval of time, divided by the length of that interval. Formally, the relationship between the hazard function $h(t)$ and the survival function $S(t)$ may be expressed as:

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t | T \geq t) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr([t \leq T < t + \Delta t] \bigcap [T \geq t])}{Pr(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T < t + \Delta t)}{P(T \geq t)} \\
&= \frac{f(t)}{S(t)}.
\end{aligned}
$$

4.3 The regression model for survival data

After explaining the basic principles of the survival analysis, it is important to mention that there are different types of survival models that can be used to analyze survival data. In this research, the focus lies on the Cox proportional hazard model where the hazard rate is represented by

$$h(t|\mathbf{Z}) = h_0(t) \exp(\beta^T \mathbf{Z}),$$

and where $h_0(t)$ is a factor called the baseline hazard rate, this is the hazard rate, often in engineering referred to as the failure rate or instantaneous rate of occurrence, when all predictors Z_k are equal to zero. As long as $h_0(t)$ is non-negative, the second part of the function represented by the exponential ensures that the fitted model will always results in a non-negative estimated hazard. $\mathbf{Z} = (Z_1, Z_2, ..., Z_P)^T$ is the vector of predictors and $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_p)$ is the vector of unknown coefficients that need to be estimated.

Furthermore $\mathbf{Z}$ is a vector of covariates with p dimensions and include variables of different measurement levels: discrete like gender, continuous like age and grades and interactions like gender by age. The interaction effects can better be included when the separate variables are also included and they interact when the hazard of dropout depends on the combination of both variables. Categorical covariates are usually modeled as dummy variables and the reference group is not included in the model.

4.4 Advantages of the Cox proportional hazard model

The Cox Proportional Hazard (CPH) model dated from 1972 is the most commonly used model for survival data. The first advantage of the CPH model is the robustness, because the results from using the Cox model will closely approximate the results for the correct parametric model (Kleinbaum & Klein, 2012). Taking account of the number of covariates included in this research is the second advantage of CPH model, this is why when in doubt, the CPH model is a safe choice. In addition, the advantage of the Cox regression compared to logistic regression is that the Cox model uses survival times and censoring while the logistic regression uses a binary outcome (0,1) and ignores survival times and censoring.

4.5 Hazard Ratio

An additional part of this analysis that needs to be explained is the concept of the hazard ratio (HR). For any two sets of predictors, $\mathbf{Z}$ and $\mathbf{Z}^*$, the hazard ratio

$$\frac{h(t|Z^*)}{h(t|Z)} = \frac{h_0(t)\exp(\beta^T\mathbf{Z}^*)}{h_0(t)\exp(\beta^T\mathbf{Z})} = \exp(\beta^T(\mathbf{Z}^* - \mathbf{Z}))$$

is constant over time.

This is where the name proportional hazard comes from. The assumption of proportional hazard is the key assumption in the Cox model. If the only difference between $\mathbf{Z}$ and $\mathbf{Z}^*$ is that Z_k is increased by one unit, the result will be

$$\frac{h(t|(Z_1, ..., Z_k + 1, ..., Z_P))}{h(t|(Z_1, ..., Z_k, ..., Z_P))} = \exp(\beta_k),$$

where $\exp(\beta_k)$ represents the hazard ratio associated with one unit increase in Z_k. β_k represents the increase in log hazard ratio per unit difference in the k^{th} covariate taking into account that all other variables are hold constant. Furthermore, the interpretation for the covariates is as follows: when β is smaller than 0, the hazard ratio is less than 1, and the increase in the values of the covariates is associated with longer survival times and lower risk. In the case that β is higher than 0, the hazard ratio is higher than 1, it means that higher values of the predictors are associated with shorter survival times and higher risk. For instance, in the setting of our educational experiment, we are often interested in the binary variable indicating if a student is male or female. Normally the coding is $Z_k = 1$ if the students are male and $Z_k = 0$ if the students are female. There are three possible outcomes: the first one is when the hazard ratio is less than one. It means that the event of interest is happening slower for male than for female. The second outcome is when the hazard ratio is greater than one, it means that the event of interest (here drop out) has a higher probability for the male than for female. The third outcome is when the hazard ratio equals one, then it means that there is no gender difference between male and female. The same reasoning can be used for the categorical variables. It is important to mention that the hazard ratio can be used to compare groups and thus do not indicate how long time it will take for a certain subject to experience the event.

4.6 Estimation of the coefficients of the Cox Hazard Model

The corresponding estimates of the parameters β's of the general CPH model are called maximum likelihood (ML) estimates and can be denoted by $\hat{\beta}_k$. These estimates can be obtained for each variable used in the model with the R-package survival. The maximum likelihood estimates of the CPH model are derived by the maximization of the likelihood function denoted as L which is the mathematical function describing the joint probability of the observed data as a function of the unknown parameters β's (Kleinbaum & Klein, 2012).

Rather than using the word complete or maximum likelihood function, the formula for the CPH model likelihood is called a “partial” likelihood function. The reason for using the term “partial” is because the likelihood formula take into account the probabilities of the subjects who fail, and the subjects who are censored are not considered (Kleinbaum & Klein, 2012). Thus the use of a part of the probabilities of the subjects by the Cox model is the reason why the likelihood is called partial. The formula of the partial likelihood can be denoted by:

$$L = \prod_{i=1}^{n} L_i,$$

where L is the product of several likelihoods for each k failure times. Thus, at the i^{th} failure time, L_i represents the likelihood of failing at a particular time. For example, L_3 denotes the likelihood of failing at time 3, given survival up to this time and the set of subjects at risk at the third failure time is called “risk set,” when the failure time increases, this set becomes smaller in size. It is important to mention that despite the fact that the partial likelihood is calculated for the failed subjects, the survival time information obtained is used for censored subjects.

After determining the likelihood function for a given Cox model, the function can be maximized by maximizing the natural logarithm of L. This operation can be achieved by taking the derivatives of $\log L$ with respect to each parameter β in the model. The derivation equation is shown here:

$$\frac{\partial \log L}{\partial \beta_k} = 0,$$

where $k = 1, ..., p$ represents the number of parameters. The solution can be obtained by using the iteration in a stepwise manner.

4.7 Checking the validity of the regression parameters

Once we know how the parameters estimates are obtained, we are interested in the inferences of the regression coefficients. Hypothesis tests and confidence intervals can be used for this purpose. There are three statistical tests that will be displayed in the output of the Cox hazard model (Moore, 2016) in the results and that will be explained below:

- The Wald test: is mostly commonly used for multiple regression when testing the significance of a particular regression coefficients for an independant variable. For each covariate of interest, the null hypothesis is β_k is equal to zero and the alternative hypothesis is β_k is different from zero. The same reasoning is used for the Cox regression and the Wald test is calculated on the same manner. The test statistic can be calculated by the following formula:

  $$W = \frac{\hat{\beta}}{s.e.(\hat{\beta})},$$

  where W is the value of the Wald test, and s.e. stands for standard error and $\hat{\beta}$ is the value of β that maximizes the likelihood.

- Likelihood ratio test: or LR statistic, can be obtained by using the log likelihood statistic. The log likelihood ratio is, for instance, used to test the significance of the interaction term between a full model including the interaction term and a reduced model without interaction term. Suppose there are $p + q$ covariates measured for two models:
  - model 1 which contains p predictor variables:

$$h(t|\mathbf{Z}) = h_0(t)\exp(\beta_1 Z_1 + ... + \beta_p Z_p)$$

  - and model 2 which contains all predictor variables $p + q$

$$h_i(t|\mathbf{Z}) = h_0(t)\exp(\beta_1 Z_1 + ... + \beta_{p+q} Z_{p+q}).$$

  The likelihood ratio test the null hypothesis H_0:

$$\beta_{p+1} = ... = \beta_{p+q} = 0$$

  whether all estimated values of β are equal to zero, against the alternative hypothesis H_a assuming that alle estimated values of β are different from zero as:

$$\chi^2_{LR} = -2(\log L(M_1) - \log L(M_2)),$$

  where $L(M_1)$ and $L(M_2)$ denote the maximum partial likelihood under both models respectively. The likelihood ratio test is approximately distributed as χ^2 with q degrees of freedom.

- Score (logrank) test: The first derivative of the partial log-likelihood is called the score function and the score test is equivalent to the log-rank test which can be carried out without calculating the maximum likelihood estimate $\hat{\beta}$. Under the the null hypothesis this test statistic is approximately chi-square distributed with p degrees of freedom.

4.8 Assumptions of the Cox Proportional Hazard Model

Similar to other regression models there are five assumptions that needs to be met before proceeding the interpretation of the results. The first assumption is that the students or the subjects and the events are independent of one another. For example, the fact that a certain student dropped out, will not cause an increase or a decrease of the likelihood of dropout of that student or other students experiencing the event. The second assumption is that censoring is not informative and this assumption is specific to the CPH regression compared to other regression models. This means that students who stayed in the study are not different from those who were lost to follow up. The third point assumes that the values of **Z** don't change overtime, for instance, if a student is living in Rotterdam at the beginning of the study, it is assumed that he will maintain the same status throughout the study. There are extensions of the CPH model that allows to account for variables that change overtime, such as time updated Cox models but they fall out of the scope of this study. The fourth assumption involves that the log hazard rate is a linear function of $\mathbf{Z}$. This assumption can be compared to the assumption of the logistic regression where the log odds is a linear function of the $\mathbf{Z}$'s. This can be checked with residual plots like linear regression models. A plot of the Schoenfeld's residuals against its failure times will be used to test this assumption, if this assumption holds for a particular covariate the plot should yield a pattern of points that are, independently of the failure time, centered at zero. To test this assumption formally, the residuals can be regressed against time, the slope of the regression line should not differ significantly from zero if the proportional hazard assumption holds. The Schoenfeld's test for proportional hazards is a test where a null hypothesis that the hazards are proportional or that the hazard ratio is constant overtime versus alternative hypothesis that the hazards are not proportional or that the hazard ratio is not constant overtime, doing this test is going to return a test for each of the individual variables each of the coefficients as well as a test for the overall model. The final and probably the most important one is the proportional hazards assumption which means that the hazard ratio is the same regardless of whether it is measured at different times. The hazard can vary. However, the relative difference between groups compared is constant overtime. To check this assumption visually the survival curves do not cross because the hazard ratio do not change overtime (Kleinbaum & Klein, 2012).

4.9 Choosing the best model

Choosing the best model in survival regression can be a little bit challenging, there are different statistics that can be used to choose the right model.

- The Concordance: the C-statistic or sometimes called C-index is a measure of goodness of fit for binary outcomes in a logistic regression model. The same statistic can be used for the Cox regression. The value of the concordance vary between 0 and 1. A value equals to 1 means that the predictions of the model are perfect. a values higher than 0.8 indicate a strong model and values higher than 0.7 are a sign of a good model. A values equal to or lower than 0.5 can be interpret as poor model [@RefWorks:doc:5fbcd252e4b00ce6e939e284].

- The Akaike's Information Criterion procedure for variable selection: when there is a large number of potential factors in the model, it is necessary to prune the model in order to select the most significant covariates. There are several methods to select the most appropriate model. One of the methods mostly used is ”stepwise” model selection with two versions: the forward version and the backward version. In the forward version, univariate models are fitted, one for each covariate. Then the covariate added to the base model are the covariates with the smallest p-value [@RefWorks:doc:5fbcd252e4b00ce6e939e284].

  Akaike's information criterion (AIC) is revisited in the context of the Cox PH regression model. Different models can be compared using:

$$AIC = -2\log(L(\beta)) + 2p,$$

where p is the number of β coefficients in each estimated model. Like mentioned before the maximum partial likelihood in the place of the maximum likelihood. The model with the smallest AIC is the best model. An automated model selection procedure can be applied using the function ”step()” in R.

5 Results

This section presents the results of this research. A Cox regression is implemented and the time to event variable taking into account whether the students are dropped out or censored is regressed on predictors such as Grades of the first academic year and variables measuring the demographic status. The first part investigates the model with the highest performance based on the standards explained in the methodology section. After selecting the best model, The Akaike’s Information Criterion will be applied to select the variables that influence the dropout of the students of the degree Business IT & Management. Then, the goodness of fit of the model will be analyzed and different assumptions of the Cox regression model will be discussed. Finally, a survival curve for different values of the variable Birthplace will be estimated.

5.1 Cox Regression fitted models

Several models will be fitted based on the predictor variables associated with each quarter in combination with the demographic covariates. For example, the first model will include only the demographic variables. The second model will include the demographic variables in combination with the grades of the first quarter. by each additional model that will be used, different grades associated with the quarter will be added to the model. The last model includes all predictor variables. The results of the analysis are shown in Table 2.

Table 2. Estimated coefficients and statistical significance of predictors in a multivariable Cox proportional hazards model fitted to the entire data for all models

Var	Model	D	Model	Q1	Model	Q2	Model	Q3	Model	Q4	Model	All
	HR	p-val	HR	p-val	HR	p-val	HR	p-val	HR	p-val	HR	p-val
CoBi	0.59	0.000	0.92	0.670	1.25	0.263	1.35	0.136	1.24	0.295	1.27	0.249
BiPl	0.98	0.887	0.71	0.053	0.60	0.012	0.73	0.112	0.76	0.175	0.82	0.347
Gend	1.65	0.028	1.39	0.162	1.38	0.166	1.21	0.425	1.20	0.465	1.20	0.459
City	0.99	0.946	0.96	0.812	1.09	0.629	0.95	0.755	0.89	0.504	0.83	0.310
Age	0.97	0.127	0.95	0.032	0.99	0.783	0.95	0.075	0.97	0.288	0.98	0.414
INT1			0.94	0.009	0.99	0.530	1.00	0.828	1.02	0.473	1.03	0.425
MGT1			0.78	0.000	0.86	0.000	0.89	0.000	0.89	0.000	0.92	0.055
PRJ1			0.86	0.000	0.99	0.672	0.94	0.072	0.94	0.075	0.96	0.298
CaSk			0.67	0.000	0.72	0.003	0.79	0.039	0.77	0.026	0.88	0.294
ReSk			0.81	0.000	0.90	0.000	0.99	0.600	0.98	0.521	1.04	0.305
INT2					0.83	0.000	0.93	0.017	0.93	0.029	0.97	0.516
MGT2					0.82	0.000	0.93	0.017	0.94	0.036	1.00	0.895
PRJ2					0.86	0.000	0.92	0.000	0.91	0.000	0.92	0.030
INT3							0.88	0.007	0.89	0.016	0.97	0.601
MGT3							0.79	0.000	0.87	0.013	0.97	0.531
PR34							0.78	0.000	0.81	0.000	0.98	0.685
PRGR							0.92	0.000	0.94	0.014	0.97	0.231
INT4									0.84	0.000	0.96	0.420
Elec									0.90	0.009	1.37	0.007
EnCr									0.57	0.120	1.56	0.287
EnGr									1.03	0.183	1.06	0.019
OCFY											0.66	0.000
FCOb											>2.0	0.992
OCOb											1.00	0.270
WAVG											1.24	0.012
NCFS											1.43	0.003
NCSS											1.32	0.022

Table 2 shows the variable names (var) used for all models, the hazard ratio (HR) and the p-values (p-val). The HR like mentioned in the methodology section is the effect of each variable adjusted for the other variables in the model. The p-values or the significance levels are presented in the second column after each hazard ratio. Model D includes the demographic variables Gender, Age, Country of birth, Birthplace and City where students live. The covariates with the highest predictive power for model D are Country of Birth and Gender. For the students born in the Netherlands, we can see that the hazard ratio is 0.59, the interpretation is as follows: at a given instant in time, a student born in the Netherlands has 41 % lower hazard rate as someone born outside the Netherlands, adjusting for the other variables in the same model. The interpretation for "adjusting for the other variables" means that the two students that will be compared need to have the same conditions regarding the rest of the variables in the model. In addition, we see that the hazard ratio for Gender is 1.65, it means that at a given instant in time, a male student is 1.65 times as likely to dropout as a female adjusting for other variables in the same model. If we subtract 1 from the hazard ratio, we can interpret it as a percentage change so that 1 minus 1.65 is 0.65 or 65%. At a given point a male student is 65% more likely to drop out than a female adjusting for the rest of the variables in the same model.

The rest of the models can be interpret on the same way. Model Q1 includes the demographic variables and the grades of the first period. The variables with a higher significance level are the courses: Research Skills, Career Skills, Management 1, Project 1, Information Technology 1 and Age, all these variables decrease the risk of dropout and the variable with the highest decreasing rate is Career Skills.

Model Q2 includes the demographic variables and the grades of the first and the second period. The variables with a higher significance level are: Birthplace, Management 1, Career Skills, Research Skills, Information Technology 2, Management 2 and Project 2, all these variables decrease the risk of dropout and the variable with the highest decreasing rate is BirthPlace.

Model Q3 contains the demographic variables and the grades of the first, the second and the third period. The variables with a higher significance level are Management 1, Career Skills and all the grades of the second and the third period, all these variables decrease the risk of dropout and the variable with the highest increasing rate is Project 34.

Model Q4 contains the demographic variables and all the grades obtained by the students during the first academic year. All grade variables are significant except Informationn Technology 1, Project 1 and Research Skills. The rest of the significant Variables decrease the risk of dropout.

Model All includes the demographic variables and all variables included in the data. It seems that the combination of all variables in one model gives a counter-intuitive results because we suppose that obtaining a higher number of credits in the first and the second semester results in a lower hazard rate. In this case of Model All, the number of credits obtained seems to increase the hazard rate, the second problem of this model is that the most grades coefficient are not significant.

5.2 Comparison of the predictive power of all models

Table 3 shows a comparison of the predictive power between all models. The final model shows a high goodness of fit compared to the demographic model which is 0.55. In addition to that, the possibility to early predict the dropout in combination with the relatively small number of parameters and a high concordance make Model Q1 the most appropriate model. A very high value of the concordance can cause overfitting when trying to predict using another dataset.

Table 3. Comparison of the predictive power and the goodness of fit

Models	Concordance
Model D	0.55
Model Q1	0.90
Model Q2	0.93
Model Q3	0.96
Model Q4	0.96
Model All	0.90

5.3 Stepwise variable selection

The objective of this research is to find a model which can estimate the hazard ratio's at an early stage and because Model Q1 and Model Q2 have a concordance higher than 0.9, The choice is fallen on Model Q1 allowing in an early stage the prediction of the most influential dropout factors. We will apply the AIC on this model to select the covariates influencing the time to event. In order to build the final model, the stepwise selection method will be applied. Table 4 shows the model including the variables generated by the stepwise method.

Table 4. Estimated coefficients and statistical significance of predictors in a multivariable Cox proportional hazard model fitted to the entire data for all predictors after applying the stepwise selection method

Var	coef	exp(coef)	se(coef)	W	Pr(>$\|W\|$)
BiPl	-0.37	0.69	0.15	-2.458	0.014
Gend	0.32	1.38	0.23	1.371	0.170
Age	-0.05	0.95	0.02	-2.153	0.031
INT1	-0.06	0.94	0.02	-2.644	0.008
MGT1	-0.24	0.78	0.02	-9.853	0.000
PRJ1	-0.15	0.86	0.02	-5.422	0.000
CaSk	-0.40	0.67	0.11	-3.589	0.000
ReSk	-0.21	0.81	0.02	-8.719	0.000

The results of the model presented in Table 4 justify the necessity of a comprehensive examination of the dropout phenomenon by explaining the effects of the most significant covariates at 0.05 level. The variable Birthplace is significant. For the students born in Rotterdam, we can see that the hazard ratio is 0.69, the interpretation is as follows: at a given instant in time, being born in Rotterdam decrease the hazard ratio with 31 % comparing to the students born outside Rotterdam adjusting for the other variables in the same model. If we subtract 1 from the hazard ratio, we can interpret it as a percentage change so that 1 minus 0.69 results in 0.31 or 31 %. At a given point in time someone who's born outside Rotterdam is 31% more likely to dropout than someone born in Rotterdam adjusting for the other variables in the same model.

The variable Age is also significant, the hazard ratio comes out to be 0.95 with a 95% confidence interval. The interpretation is: at a given point in time the probability of dropping out for someone who is 1 year younger is 1 minus 0.95 or 5% higher than someone who is 1 year older adjusting for the other variables. So that when taking two individuals who have the same score for the rest of the variables. The student who's one year younger are expected that they have a 5 % higher chance of dropping out at any given point in time.

The courses of the first quarter are all significant and they seems to influence the dropout. For example, the hazard ratio of the course Career Skills is equal to 67 % with a 95% confidence interval. The interpretation is: at a given point in time the probability of dropping out for someone who obtains 1 point higher is 33% lower than student who obtain 1 point lower adjusting for the other variables. The same applies for the rest of the courses Research Skills, Information and Technology 1, Management 1 and Project 1.

The variable Gender is statistically not significant. Given the presence of many missing values in the data of students who did not attend the exams and to the relatively small size of the study, it is plausible that this variable is informative but not clear enough to attain greater statistical significance.

Figure 5 from Appendix A shows the 95% confidence interval for each covariate from the fitted model using the AIC. It can be concluded visually that the variables Career Skills and Birth Place have the highest predictive power for the dropout.

5.4 Checking the assumption of the Cox Proportional Hazard Model

The fitted model aims to estimate survival as a function of the variables mentioned in table 4. This section will focus on checking two of the assumptions mentioned in paragraph 4.8. The assumptions are the linearity as well as the proportional hazards assumption. When checking the linearity we assume that the relationship between any of the numeric variables and the log hazard is linear. This assumption can be checked in the same way we check linearity in other models like linear regression.

5.4.1 Testing non-linearity

Figure 4 shows the residuals on the X-axis and the predicted values on the Y-axis. After plotting the residuals, a horizontal line is added across the residuals equals 0 or when $y = 0$. after that we will add a smoother red line through the points, we can conclude that the red line is approximately linear and that there is no need to address the small non-linearity showed in the graph above.

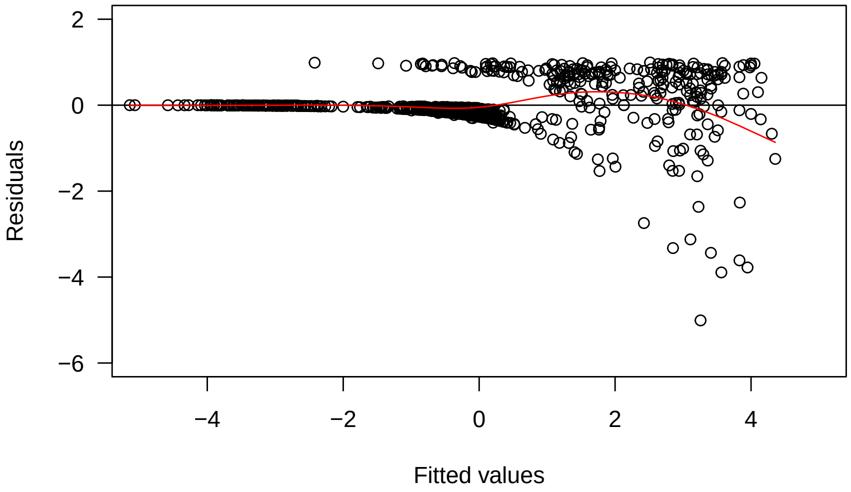
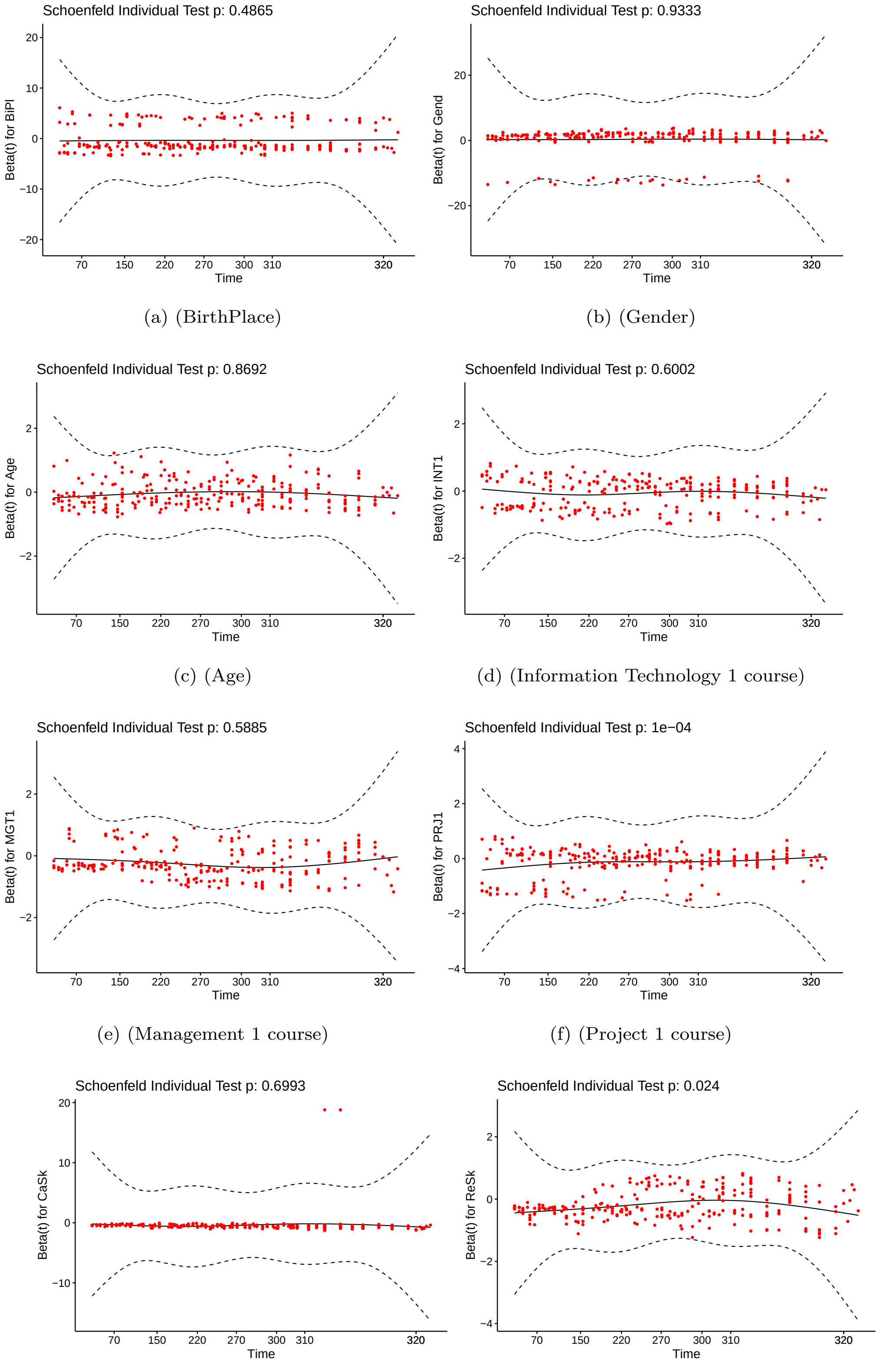


Figure 4: Plot of the fitted values on the X-as against the residuals on the Y-as

5.4.2 Testing proportional Hazards assumption

Figure 5 shows whether the coefficients or the hazard ratios change overtime for all variables included in the model. In the case that the coefficients does not change overtime we expect to see a change of zero. Since we have eight variables in this model, we will plot all of them, a line in the middle is looking if we allow the coefficient to change overtime which is essentially mean allowing the hazard ratio to change overtime. The dashed lines represents a 95% confidence interval. We can see that in mostly all graphs that the change seems to be zero or in other words that the black line on the middle is almost horizontal. As a reminder, the solution in the case that the proportional hazard assumption is not met is stratifying on the variables where the assumption is not met or using the time dependent coefficient models allowing the effect of that variable to interact with time.



(a) (BirthPlace)

(b) (Gender)

(c) (Age)

(d) (Information Technology 1 course)

(e) (Management 1 course)

(f) (Project 1 course)

(g) (Career Skills course)

(h) (Research Skills course)

Figure 5: Schoenfeld Residuals for all variables generated by the stepwise method

5.5 Estimated survival curves for different values of selected variables

For each covariate in the model, we can estimate a survival curve for different values of the selected variable. For instance, consider the survival curves for different levels of the variables Birth Place in Figure 6. After fitting a Cox model to the data. There is a possibility to obtain a visualization precising the predicted survival proportion at any given point in time for a certain risk group of students. The estimation is done by considering the mean values of the covariates. The goal is to visualize how estimated survival depends upon the value of the predicted variable. Consider that we want to assess the influence of the variable Birthplace on the estimated survival probability. We see from the the figure below that the students born outside Rotterdam are more likely to dropout comparing to the students born in Rotterdam and that the survival probability decreases fastly during the fourth quarter. It seems that the program Business IT & Management do not give the possibility to the students who are not sure that the program suits them to leave the program during the first quarter or to switch to an appropriate one. In addition to that, the survival probability decreases from almost 80 to 35 % during the fourth quarter. It seems that the program do not pay any attention to this group of students.

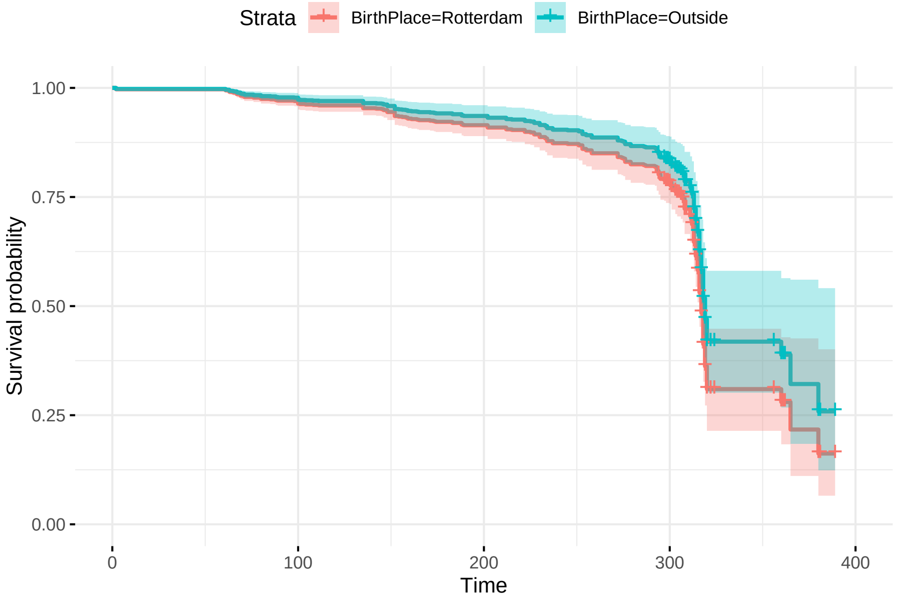


Figure 6: Estimated survival curves for different values of the variable Birthplace

6 Conclusion

In this study, we tried to answer the question: *"How can freshman student dropout rates of the Business IT & Management degree be tackled using Survival Analysis?"*. Therefore, we examined the structure of OSIRIS dataset, the tracking system of the Rotterdam University of Applied sciences. The program where this internship is done is called Business IT & Management. The data is cleaned and manipulated in order to apply the Survival Analysis method.

The OSIRIS dataset contained information on the demographic variables, the grades obtained by the students during the first year and some additional variables representing the number of credits obtained in different semesters. Due to the transformation of some variables in the dataset, new variables have been created like Time to event and a binary variable indicating whether a student is censored or not. Both variables were needed to compose the predicted variable in each Survival Analysis.

The reason of using the Survival Analysis to examine the student dropout, is the possibility of utilizing censored data in order to estimate the hazard ratios. The Rotterdam University of Applied Sciences suffers from a high dropout rates. Different initiatives inside the university are made to discover the factors affecting the student dropout. But these initiatives are expensive because they are based on qualitative research which takes a lot of time.

We developed six models based on the survival regression technique called the Cox Proportional Hazards regression. One model containing the demographic variables, four models containing the demographic variables and the grades of different quarters and one model containing all variables used in the dataset. After selecting the most appropriate model based on the goodness of fit criteria, we used the stepwise method based on the Akaike's Information Criterion (AIC) procedure for variable selection.

The selected model makes it possible to examine the predicting power of different factors causing the student dropout. This gives insight in variables which lower the risk of dropping out and the variables which increase the risk of dropping out. The best model is determined after measuring the performance of the models based on two standards, the concordance and log likelihood ratio.

Returning to the main research question, the model that scored high referring to the concordance and the log likelihood was extracted after applying the AIC procedure for variable selection on the model combining the demographic variables and the grades of the first quarter. The model shows that to be able to tackle the dropout rates of the program Business IT & Management. The university should address the effort to the younger students born outside Rotterdam city and obtaining a lower grades for the courses of the first quarter.

Therefore, the results of this thesis can be seen as an eye-opener to try to expand and enrich the Osiris dataset in the future with more variables with the aim of gaining more insight into, for example, motivation and study behavior of the students and also adding other exogenous variables such as socio-economic status to gain more understanding into the factors that can influence the dropout rate within the Rotterdam University of Applied Sciences.

7 Discussion and limitations

After presenting the conclusion of the analysis which provide guidance on the factors impacting the students dropout in the program Business IT & Management and suggest possible solutions for improvement. A balance between a good tutoring and a huge support during the first term of enrollment are therefore very critical to reduce dropout rates. The program business IT & management will try in response to this thesis to create an evaluation system for all freshman to help them at an early stage to intensify their studying strategy. Additional attention should also be paid to grading data in combination of the available variables in the Osiris System, given the predictability that can be provided and the possibilities that can be achieved to help students.

This thesis is written as part of a Data Science and Marketing Analytics master degree. The primary question of the reader will be, how can this thesis be used for marketing purposes? The answer is that the competition between universities worldwide has put them in a position to review their Key performance indicators. The priority in recent years has been school dropout because we know that school dropout causes even more inequalities in the society. In recent years, higher education has been seriously engaged in all kinds of initiatives to solve school dropout problems. Some universities are starting to see the student as a customer, although this is a different discussion, and one of the points used to attract more students is the low student dropout rates. If we place this thesis in the context of the universities as companies and the students as customers of the knowledge, then universities can use this thesis from a marketing point of view to attract more students. However, we know that there are many lecturers that will not accept the concept that the student can be seen as a customer, especially because the lecturers have to assess the students. This remains an ethical discussion for which we will not find an answer in this thesis.

This study had many limitations. First, there are many missing values in the dataset. The problem of these missing values can not be solved by replacing them by the mean or the median or by using the classical machine learning solutions. The students with an “NA” value are mostly students who did not attend the exam, we replace their missing data with the value zero assuming that in the case that they will make the exam that their results will be equal to zero. We know that the results of our research can be somewhat biased. However after interviewing the manager and the tutors of the program, they reassure us that the variables found in the results of this research are also expected by them. Second, Applying this method on a larger population is expected to enable the detection of more influential variables. In addition, The university is expecting next year to use a more advanced learning management system which will allow to add more variables to the used dataset. Third, this study can better be expanded to the whole university to identify more characteristics and relationships unique to the programs offered by the university, as well as to provide more information for improving educational system. Fourth, Another issue related to the analysis was whether the combination of demographic variables and the grades is justified. Last, including time-varying covariates will help to improve the results of the Cox Proportional Hazard Model.

References

Ameri, S., Fard, M., Chinnam, R., & Reddy, C. (2016). *Survival analysis based framework for early prediction of student dropouts.* 903–912. ACM. https://doi.org/10.1145/2983323.2983351

Arsad, P. M., Buniyamin, N., & Manan, J. A. (2012). *Neural network model to predict electrical students' academic performance.* 1–5. IEEE. https://doi.org/10.1109/ICEED.2012.6779270

Baars, G. J. A., & Arnold, I. J. M. (2014b). Early identification and characterization of students who drop out in the first year at university. *Journal of College Student Retention : Research, Theory & Practice*, *16*(1), 95–109. https://doi.org/10.2190/cs.16.1.e

Baars, G. J. A., & Arnold, I. J. M. (2014a). Early identification and characterization of students who drop out in the first year at university. *Journal of College Student Retention : Research, Theory & Practice*, *16*(1), 95–109. https://doi.org/10.2190/cs.16.1.e

Barneveld, A. V., Arnold, K. E., & Campbell, J. P. (2012). *Analytics in higher education: Establishing a common language.*

Ben-Tsur, D. (2007). Affairs of state and student retention: An exploratory study of the factors that impact student retention in a politically turbulent region. *British Journal of Sociology of Education*, *28*(3), 317–332. https://doi.org/10.1080/01425690701252382

Breier, M. (2010). From 'financial considerations' to 'poverty': Towards a reconceptualisation of the role of finances in higher education student drop out. *Higher Education*, *60*(6), 657–670. https://doi.org/10.1007/s10734-010-9343-5

Cabrera, A. F., Nora, A., & eda, M. B. C. (1992). The role of finances in the persistence process: A structural model. *Research in Higher Education*, *33*(5), 571–593. https://doi.org/10.1007/bf00973759

Cheng, J., Kulkarni, C., & Klemmer, S. (2013). *Tools for predicting drop-off in large online classes.* 121–124. ACM. https://doi.org/10.1145/2441955.2441987

Cohen, A. (2017). Analysis of student activity in web-supported courses as a tool for predicting dropout. *Educational Technology Research and Development*, *65*(5), 1285–1304. https://doi.org/10.1007/s11423-017-9524-3

DeBerard, M. S., Spielmans, G. I., & Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal*, *38*, 66+.

Decker, R., & Lenz, H.-J. (2007). *Advances in data analysis.* Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70981-7

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, *34*(5), 569–581. https://doi.org/10.1007/BF00991920

Elbadrawy, A., Studham, R. S., & Karypis, G. (2014). *Personalized multi-regression models for predict-*

ing students' performance in course activities. Retrieved from https://conservancy.umn.edu/handle/11299/215948

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2011). *A logistic regression model for the*.

Grau-Valldosera, J., & Minguillón, J. (2014). Rethinking dropout in online higher education: The case of the universitat oberta de catalunya. *International Review of Research in Open and Distance Learning*, *15*(1), 290–308. https://doi.org/10.19173/irrodl.v15i1.1628

Hamsa, H., Indiradevi, S., & Kizhakkettottam, J. (2015). ScienceDirect. *Clinical Microbiology Newsletter*, *37*(4), 33. https://doi.org/10.1016/j.clinmicnews.2015.01.008

Horstschräer, J., & Sprietsma, M. (2013). The effects of the introduction of bachelor degrees on college enrollment and dropout rates. *Education Economics*, *23*(3), 296–317. https://doi.org/10.1080/09645292.2013.823908

Hovdhaugen, E. (2011). Do structured study programmes lead to lower rates of dropout and student transfer from university? *Irish Educational Studies*, *30*(2), 237–251. https://doi.org/10.1080/03323315.2011.569143

Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the united states. *The Journal of Higher Education (Columbus)*, *77*(5), 861–885. https://doi.org/10.1080/00221546.2006.11778947

Johnson, L., Becker, S. A., Estrada, V., & Freeman, A. (2014). *The NMC horizon report 2014 k 12 edition examines emerging technologies for their potential*.

Khan, I., Sadiri, A. A., Ahmad, A. R., & Jabeur, N. (2019). *Tracking student performance in introductory programming by means of machine learning*. 1–6. Institute of Electrical; Electronics Engineers (IEEE). https://doi.org/10.1109/icbdsc.2019.8645608

Kleinbaum, D. G., & Klein, M. (2012). *Statistics for biology and health survival analysis a self-learning text third edition*.

Lee, Y., & Choi, J. (2010). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, *59*(5), 593–618. https://doi.org/10.1007/s11423-010-9177-y

Levin, J., Barak, A., & Yaar, E. (1979). College dropout and some of its correlates. *Megamot*, (4), 564–573. Retrieved from https://eur.on.worldcat.org/oclc/5792927011

Luna, J. (2000). *Predicting student retention and academic success at new mexico tech*.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Mah, D.-K. (2016). Learning analytics and digital badges: Potential impact on student retention in higher education. *Technology, Knowledge and Learning*, *21*(3), 285–305. https://doi.org/10.1007/s10758-016-9286-8

Moore, D. F. (2016). *Applied survival analysis using r*. Springer International Publishing. Retrieved from https://library.biblioboard.com/viewer/32e26cc1-c3d0-11ea-a9fb-0ae0aa0d175d

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355–371. https://doi.org/10.1023/A:1018755201899

Ortiz, E. A., & Dehon, C. (2013). Roads to success in the belgian french community's higher education system: Predictors of dropout and degree completion at the université libre de bruxelles. *Research in Higher Education*, *54*(6), 693–723. https://doi.org/10.1007/s11162-013-9290-y

Pagan, R., & Edwards-Wilson, R. (2002). A mentoring program for remedial students. *Journal of College Student Retention : Research, Theory & Practice*, *4*(3), 207–226. https://doi.org/10.2190/UFGM-8014-894V-CXFL

Radcliffe, P. M., Huesman, R. L. J., & Kellogg, J. P. (2006). *Identifying students at risk: Utilizing survival analysis to study student athlete attrition*. Retrieved from https://conservancy.umn.edu/handle/11299/159769

Rubel, A., & Jones, K. M. L. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, *32*(2), 143–159. https://doi.org/10.1080/01972243.2016.1130502

Thornsberry, J. L. (2010). *Freshman transition and its effectiveness on student success as measured by improved attendance, improved grades, decreased discipline referrals, and decreased dropout rate* (PhD thesis). Retrieved from https://search.proquest.com/docview/193523221

Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). *Use data mining to improve student retention in HE - a case study*. Retrieved from https://explore.openaire.eu/search/other?orpId=core_ac_uk___::f0d3bbaf5e7d0761028b95f66f7a283a

Appendix

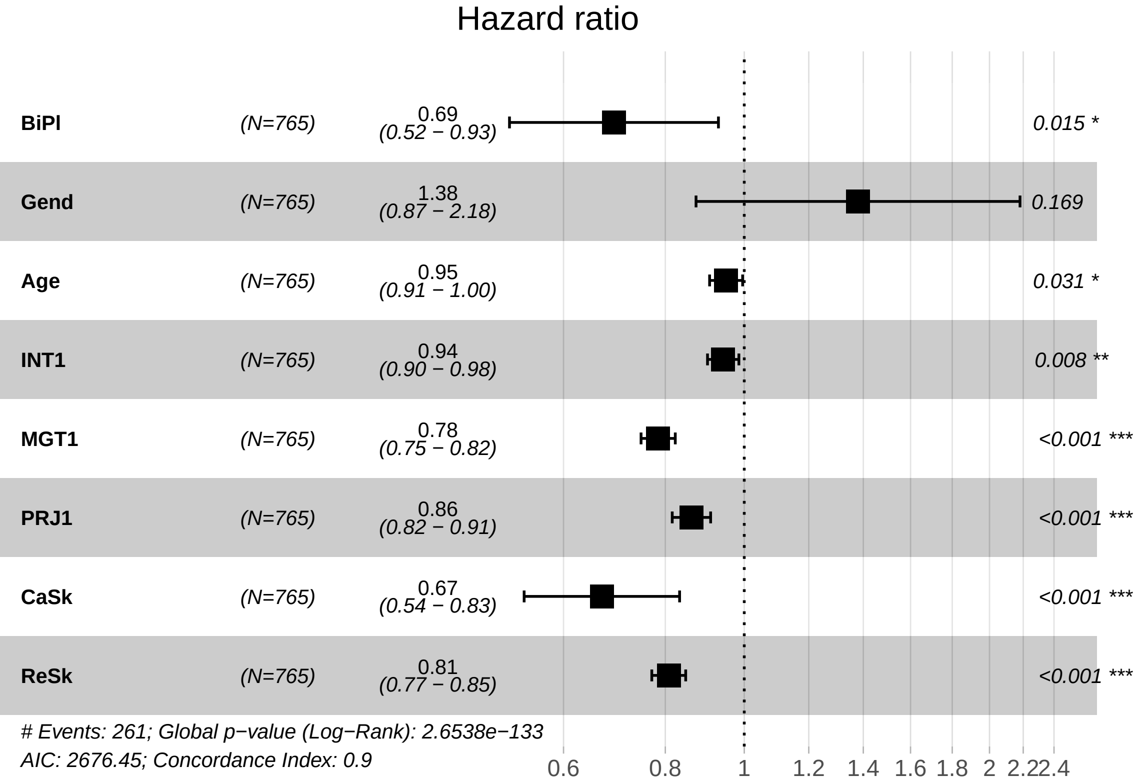


Figure 7: 95% Confidence interval for estimated regression coefficients