Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Lauri Viitanen

Identifying At-Risk Students at Metropolia UAS

Estimating Graduation Probability with Survival Models and Statistical Classifiers

Master's Thesis
Espoo, May 19, 2016

Supervisor: Professor Samuel Kaski, Aalto University
Advisor: Doc. ICS Pekka Marttinen Ph.D. (Stats.), Aalto University

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Lauri Viitanen

Title: Identifying At-Risk Students at Metropolia UAS: Estimating Graduation Probability with Survival Models and Statistical Classifiers

Date:	May 19, 2016	**Pages:**	vii + 77
Major:	Information and Computer Science	**Code:**	SCI021Z
Supervisor:	Professor Samuel Kaski, Aalto University		
Advisor:	Doc. ICS Pekka Marttinen Ph.D. (Stats.), Aalto University		

Since the legislation changed in 2014 universities in Finland have been competing against each other for a larger share of the total government funding in higher education. The single most significant measure of the amount of funding is the university's share of all graduated students in a given year. Thus there is a need for sophisticated tools that help to increase the probability of a student to graduate.

In this thesis we apply survival analysis to the student data of Metropolia UAS. The dataset contains over 100,000 study rights of which one third are used to find variables that separate graduates from dropouts and estimate the time remaining until graduation. We analyze the effect of non-linear interaction between the variables and time on the proportionality assumption in the Cox PH model and on the model's accuracy. We also classify students to the two categories using naive Bayes, generalized linear model, support vector machine and Gaussian process classification.

The findings are in line with previous research: GPA and gender are significant predictors of graduation probability. We also find that gaining more credit points (3% of degree extent) during the first year increases the chances of graduation more than an increase of one unit in GPA (e.g. from 2 to 3 on scale 1–5). The constant variable model is preferred to the interaction-enabled model, having 0.73 concordance and 0.87 years RMSE for time to graduation. Support vector machine was found to be the best performing classifier with accuracy of 74%, which is a 14% improvement over classifying everyone into the larger category.

Keywords:	Cox, hazard, survival analysis, graduation, prediction, duration, Gaussian process, university, identification, classification, risk, Machine learning, Metropolia
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Lauri Viitanen

Työn nimi: Korkean keskeyttämisriskin opiskelijoiden tunnistaminen Metropolia AMK:ssa: Valmistumistodennäköisyyden arviointi elinaikamalleilla ja tilastollisilla luokittelijoilla

Päiväys:	19. toukokuuta 2016	**Sivumäärä:**	vii + 77
Pääaine:	Tietojenkäsittelytiede	**Koodi:**	SCI021Z
Valvoja:	Professori Samuel Kaski, Aalto University		
Ohjaaja:	Dosentti Pekka Marttinen, Aalto Yliopisto		

Lainsäädännön muututtua 2014 suomalaiset yliopistot ovat päätyneet kilpailemaan osuudestaan valtion korkeakoulutukseen suuntaamasta rahoituksesta. Merkittävin yksittäinen rahoituksen mittari on yliopiston osuus kaikista valmistuneista opiskelijoista kunakin vuonna. Yliopistoilla on siis tarve kehittyneille työkaluille, joilla voitaisiin korottaa valmistumisastetta entisestään.

Tässä lopputyössä sovellamme elinaikamalleja Metropolia AMK:n opintorekisterin tietoihin. Rekisteri sisältää yli 100 000 opiskeluoikeutta, joista kolmasosaa käytetään elinaikamallin sovittamiseen sellaisten muuttujien löytämiseksi, jotka parhaiten erottelevat valmistuvat opiskelijat keskeyttävistä ja tarkimmin ennakoivat jäljellä olevaa opintoaikaa. Analysoimme epälineaarisen muunnoksen vaikutuksia käyttämämme Coxin PH mallin suhteellisuusoletukseen sekä ennustetarkkuuteen. Lajittelemme opiskelijat em. luokkiin myös käyttämällä naiivia Bayesilaista luokittelijaa, yleistettyä lineaarista mallia, tukivektorikoneluokittelijaa ja Gaussista prosessia.

Tulokset ovat linjassa aiempien tutkimusten kanssa: arvosanojen keskiarvo ja opiskelijan sukupuoli ovat merkitseviä muuttujia valmistumistodennäköisyyden arvioinnissa. Havaitsemme myös, että ylimääräisten opintopisteiden (3 % tutkinnon laajuudesta) suorittaminen 1. vuonna kasvattaa todennäköisyyttä valmistua enemmän kuin yhden numeron parannus keskiarvossa (asteikolla 1–5). Vakiomuuttujamalli havaitaan soveltuvammaksi kuin epälineaarisen muunnoksen salliva malli. Sen konkordanssi on 0,73 ja valmistumisajankohdan ennusteen keskivirhe 0,87 vuotta. Tukivektorikone havaittiin parhaaksi valmistumista luokittelevaksi menetelmäksi. Sen tarkkuus on 74 % eli n. 14 % parempi kuin luokittelemalla kaikki enemmistön mukaan.

Asiasanat:	Cox, hasardi, elinaika-analyysi, valmistuminen, ennakointi, kesto, Gaussinen prosessi, yliopisto, tunnistaminen, luokittelu, riski, koneoppiminen, Metropolia
Kieli:	Englanti

Acknowledgements

I wish to thank the excellent specialists at Metropolia University of Applied Sciences, whose help and insights have made it possible to construct the dataset used in this thesis and to understand the results. Especially helpful were the insights of O. Troberg. Special thanks go also to Aki Vehtari at Aalto University for his clever ideas on the technical details of the modeling phase, Tomi Peltola for clarifications, Tuomas Ollikainen for helpful comments, and of course to my thesis instructor Pekka Marttinen.

Espoo, May 19, 2016

Lauri Viitanen

Abbreviations and Acronyms

CGPA	Cumulative GPA is the GPA of attainments from several consequtive time periods (e.g. years 1-3), whereas GPA includes only the attainments of a single time period (e.g. only 2nd year's).
ID	Identifier, usually a number e.g. "1234" or a string e.g. "12.34.567"
GLM	Generalized linear model, a machine learning method
GP	Gaussian process, a machine learning method
GPA	Grade point average is the weighted mean of grades received during a time period where the weight of the grade of an attainment equals the number of credit points gained from that attainment.
PH	Proportional hazards survival model
SVM	Support vector machine, a machine learning method
UAS	University of Applied Sciences
XDW	X Data Warehouse: a data model for Finnish universities. It is defined with the Unified Modeling Language (UML).

Contents

Chapter 1

Introduction

Applying for or attending to studies that lead to a Finnish university degree costs nothing to the (Finnish) applicant or student [1, 2]. In comparison, the tuition for a first-year MBA student for the nine-month academic year of 2015–2016 in Stanford Graduate School of Business is $64,050 without and up to $129,415 with a study trip to fulfill the global experience requirement [3]. Instead of tuitions, Finnish universities get their funding from the government and prior to 2015 universities of applied sciences (UASes) received funding also from the local municipalities [4–6].

Previously the amount of funding was proportional to the weighted sum of the number of students and the number of attained degrees in the UAS with weights 0.7 and 0.3 respectively [7, 8]. As of January 1st 2014 this changed into a weighted sum of 11 criteria, including e.g. the number of attained degrees (with 46% weight), the number of students who attained 55 credit points during an academic year (24%), the number of attained Master's degrees in the UAS (4%), the number of students in exchange (3%) and course feedback given by students (3%) [9–12]. Perhaps the most significant change in the criteria is that the number of students no longer affects the funding directly, but only via the number of attained degrees and credit points. Thus a UAS should pay more attention to the way the students perform and progress in their studies.

When the proportion of the overall funding given to a UAS depended mainly on the number of students, a UAS with a larger pool of students got more funding than a smaller one. In the new funding scheme the funding is proportional to university's capability to produce degrees, students who gain enough credit points etc. [10]. Thus, a UAS that is more efficient to fulfill the criteria gets more funding than a less efficient one of similar size [13, 14]. This leads to universities paying additional attention to ensure that every student gains at least 55 credit points per year and that their graduation is not

delayed. Because every university aims to improve their graduation rate and other performance measures as much as possible, only those universities that are able to improve it relatively more than other universities get increased funding. Improvement itself no longer guarantees sufficient funding for the next year.

1.1 Problem statement

To ensure sufficient funding a university must be able to utilize their available resources such as premises, computers, and teachers efficiently and also prevent the students from dropping out or delaying their studies. However, it is challenging to efficiently find the students that need additional guidance with planning their studies. Additionally, they need to be found before the next semester begins so that changes can be implemented in the person's studies. Preferably the personnel of a university would have effective tools to pre-emptively find the students who could gain most from the additional guidance. If they had an accurate estimate of the time remaining to a person's graduation, they could concentrate on only those students whose studies are going to be prolonged and help them to remove the obstacles from timely graduation. On the other hand, if they had predictive profiles of each student, they could concentrate on students with the highest risk of dropping out or unfavorably changing their field of studies later on.

In this thesis, we apply survival analysis for both graduation time prediction and drop out risk profiling. We also apply four other machine learning methods to classifying students to those who will graduate and those who will not. The survival analysis uses the students' background information such as gender, field of studies and the progress of studies as explanatory variables for the graduation rate and drop out risk. The most significant variables identified by the survival models are used for classification. The analysis is performed with the data of Metropolia University of Applied Sciences, the largest UAS in Finland. We focus on predicting the future graduation of students who have already studied exactly one year. This way the actual predictive capabilities of the model are much better revealed than if students on the verge of graduation would be included as well. In addition, numerous earlier studies conclude that grades obtained during the first year, i.e. student performance and GPA in first and second semesters may well be the single best predictors of student retention, which is a prerequisite of graduation [15–18].

1.2 Structure of the Thesis

This thesis is organized as follows: in Chapter 2 we review earlier work and present the mathematical foundations of survival models and briefly introduce the other methods applied to the problem. In Chapter 3 we describe in detail the data used in the thesis and the preprocessing steps taken to make it into a single numerical matrix. The quality of the dataset is discussed and in Section 3.1.5 we compare it with officially reported numbers. In Chapter 4 we present the implementations of the models that are going to be used and the variables that we select from the full dataset as covariates of the survival model(s). In Chapter 5 we evaluate the performance of the predictions, assess the models' fit to data and the possible violations of any assumptions inherent to the selected survival models. In Section 5.3 we present our approach to optimizing the set of covariates for the survival models and in Section 5.4 report the results. Finally, the performance of the classifiers is discussed in Section 5.5. We then draw conclusions of the results in Chapter 6 and make suggestions for future improvements. Chapter 7 concludes the thesis.

Chapter 2

Background

The reasons behind student attrition and retention have been a subject of scientific inquiry at least since the 1920's. Summerskill pointed out in his 1962 paper summarizing 40 years of research on student attrition that the term "attrition" itself was not defined consistently across the studies [19]. One of the earliest papers where statistical methods were used to study the parameters that affect student attrition or retention is from 1968 when Bayer applied multiple regression and correlation analysis to student data [20]. Around these times appeared the popular and often cited theoretical models of student attrition and retention by Spady, Tinto and Bean [21–24]. In addition to theoretical models, Spady and Bean published also results using statistical methods.

According to the listing by Nandeshwar et al., Spady applied multiple regression to student data in 1971 and Bean in 1980 [25]. The listing contains 21 studies, the most recent from 2008. Nine of them applied either multiple or logistic regression. The studies made in 2006 or later all applied decision trees, neural networks or both. They also list a study by Murtaugh et al. applying survival analysis in 1999.

A similar listing was made by Ribeiro [26]. He lists 54 studies and groups them by the algorithms used in the research. Neural networks and decision trees cover 26 studies. Logistic regression is listed as the third most popular algorithm among the studies with eight occurrences. He does not list any papers using survival analysis. In addition to the algorithms used in earlier studies he also lists factors found to affect student attrition in a set of 23 studies between 1983 and 2012. The studies found 16 different factors. Among the most often found factors were Grade Point Average, Scholastic Aptitude Test score, place of residence, and ethnicity.

A listing of how demographics and academic performance affect student retention was given by Weng [27]. It contains 13 studies from 1996 to 2005

and claims the most significant attributes affecting student persistence to be the secondary school and first-year academic experience, entrance test score, backgrounds, and socioeconomic status/financial situation of the students. In his own study of the retention of IT students the most significant factor in identifying at-risk students was the second-semester grade. Study major, gender, age, second-semester credits, loan status, and residency (with or away from their family) were also found to have a significant impact on student retention. He concluded that first-year students encounter difficulty when first living away from home, while higher year level students tend to experience financial problems in the longer term.

Applying survival analysis to analyze student attrition or retention has been done at least since the 1990's. Willet and Singer were among the first ones to publish a study in 1991 where survival analysis was used for studying student dropout [28]. Desjardins et al. applied discrete time hazard model three years later to predict student stop out [29]. Time step was defined as a semester (or term) and the variation of the effect of covariates in time was examined. Murtaugh et al. applied survival analysis to model the retention of undergraduate students [30]. They noticed that retention decreased with age and correlated with high school GPA and first-quarter GPA. Non-residents had lower retention rates than resident students and attending to the Freshman Orientation Course appeared to reduce the risk of dropping out. They did not find gender to be a statistically significant predictor of student retention.

An interesting and useful modification to survival models was made in 1999, when DesJardins et al. applied it to the examination of the temporal dimensions of student departure [31]. The study documents that even though the values that the variables take may not change over time, the effects they have on student departure may vary over time. They account for the unobserved heterogeneity in the model by including an independently distributed random variable θ to their model. This is basically a *frailty survival model* which assumes that the hazard for individual i is multiplied by an unobserved random effect θ_i [32, 33]. When the unobserved heterogeneity was controlled, the gender did not have effect anymore. Higher ACT scores, being from out-of-state, and having transfer credits from earlier studies all contributed to lower risk of dropping out. Being older than a typical first year student had the opposite effect. Having higher GPA reduced the risk of retention, but the effect was seen to wane as time passes.

A wider range of different survival models was examined by Calgano et al., when they compared the educational outcomes of older and traditional-age students with a single-risk discrete-time logistic hazard model, a non-proportional model with interaction between the covariates and time, and a

discrete-time competing-risk model that was implemented by estimating a multinomial logistic regression with four possible outcomes: dropout, transfer, certificate or associate degree completion, and no outcome for the comparison group [34]. They find that after controlling for mathematics test score, older students have a 1.24 times higher conditional probability of completing a degree or certificate in the observed event period than younger students. The finding confirms the hypothesis that older students enrolled in community colleges graduate less, but not simply because they are older, are differentially affected by environmental factors, and are more likely to enroll part-time, but rather because they have been away from the formal education system for some time and need to refresh their math skills. They also find that women are more likely to graduate each period.

Another rarely seen approach was taken by Guillory, who investigated college student retention using a multilevel discrete time hazard model with baseline logit-hazard curve [35]. The model was a combination of survival analysis and *hierarchical* linear modeling. They found that in the discrete multi-level model gender and ethnicity were not significant. The time-indicator variables were found to be significant and with many variables the proportional error assumption was violated.

A fully hierarchical survival model was later applied to student data by Lamote et al., who predicted student dropout using a multilevel discrete-time model with logistic hazard and exploring the effect of different school classifications [36]. They noticed that schools with higher mean socioeconomic composition had far lower dropout rates, similarly to schools where the relationships between students and teachers are good. At the student level, gender was found to be the most commonly observed student characteristics predicting dropout. First year student performance was also was one of the major predictors of dropping out, along with student mobility, which approximately doubles the risk of dropping out. In addition, older students were found to have four times higher risk of dropping out than students entering the institution at the typical age.

Finally, Bates applied event history analysis to understand the temporal dimensions of graduation and factors that affect whether students succeed or fail [37]. She used pre-enrollment, enrollment, and financial aid variables to model the timing of graduation for three cohorts of first-time, full-time, and degree-seeking undergraduate students for a six year period. The dataset was restructured from person-level to person-period level, changing the unit of analysis from the individual to the individual's semesters of enrollment. She argues that "*time is not only significant but the fundamental predictor of time to degree completion*". The study found strong relationship between the longitudinal effects of academic performance while in college (as measured

by cumulative GPA) and graduation. Females were also found to be 1.28 times more likely than males to graduate.

Survival analysis is the primary method used in this thesis for investigating the effects of different explanatory variables on student graduation. In the next section, we lay out the details of this approach and discuss its benefits and restrictions. Finally, a brief overview of the secondary machine learning methods used to predict student graduation is also given.

2.1 Survival Analysis

Survival analysis (also known as hazard analysis, event history analysis, reliability analysis, failure time analysis, duration analysis, and transition analysis) is a regression method for estimating the impact of explanatory variables on the rate of a predefined event occurring over time [35, 36, 38]. The time can be continuous or discrete. For example, if one year is split into two semesters and time is counted semester-wise $(0, 1, 2, \ldots$ semesters), then the time is discrete. At every point in time the rate is the *risk* or *hazard* of the event occurring, assuming that it has not occurred yet. The function that describes this *risk* over time is called the *hazard function*, often denoted by $\lambda(t)$ or $h(t)$ [30]. More formally, hazard function is a non-negative function of time t and it tells the rate of the event occurring at the given time *instant*, given that is has not occurred before [39]:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.1}$$

Here, T is a non-negative random variable giving the waiting time until the occurrence of the event. Thus $Pr(T \geq t) = 1 - Pr(T < t)$ is the probability that the event has *not* occurred by time t. This is also called the *survival* function $S(t)$, because it tells the probability of *surviving* until time t without experiencing the event [40, 41]. Integrating over $\lambda(t)$ we get the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(x)\, dx, \tag{2.2}$$

which describes the sum of risks one faces when time passes from 0 to t. Intuitively, the smaller the exposure to the risk or hazard, the better chances are that one can survive until some time t without experiencing the event. To see how the hazard and survival functions are related, we can rewrite the hazard function as

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T < t + \Delta t \ \cap \ T \geq t)}{Pr(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T < t + \Delta t)}{Pr(T \geq t)} \\ &= \frac{f(t)}{S(t)}.\end{aligned} \tag{2.3}$$

where $f(t)$ is the probability density function of T, i.e. probability of the event occurring at time t. From the definition of the survival function

$$S(t) = Pr(T \geq t) = 1 - Pr(T < t) = 1 - \int_0^t f(x)\, dx \tag{2.4}$$

we notice that

$$\frac{d}{dt}\, S(t) = -\frac{d}{dt} \int_0^t f(x)\, dx = -f(t) \tag{2.5}$$

and thus

$$\begin{aligned}\lambda(t) = \frac{-\frac{d}{dt}\, S(t)}{S(t)} = \frac{-S(t) \frac{d}{dt} \log S(t)}{S(t)} &= -\frac{d}{dt} \log S(t) \\ -\lambda(t) &= \frac{d}{dt} \log S(t) \\ \int_0^t -\lambda(x)\, dx &= \log S(t) \\ -\Lambda(t) &= \log S(t) \\ \exp\left(-\Lambda(t)\right) &= S(t).\end{aligned} \tag{2.6}$$

Here we have assume $S(0) = 1$, i.e. the event has not occurred before any time has passed. We also assume $S(\infty) = 0$, i.e. the event cannot be avoided indefinitely. The above result shows that the survival and hazard functions both describe $f(t)$, the distribution of T. Thus to attain the mean μ of T we can calculate it using e.g. survival function. Integration by parts gives

$$\mu = \int_0^\infty t f(t)\, dt = t(-S(t))\Big|_0^\infty - \int_0^\infty -S(t)\, dt = \int_0^\infty S(t)\, dt \tag{2.7}$$

which means that the expected survival time is the integral of the survival function [41].

2.1.1 Types of Survival Models

In most survival analysis settings however the interesting question is not so much the expected survival time as it is the way different covariates affect the risk of event occurring. For example, survival analysis could be used to study how much a new drug increases the survival probability of the patient over time. The study could use e.g. dose, age, and gender of the patient as explanatory variables denoted by $\mathbf{X}$. Assuming a log-linear model for the survival time T_i of observation i,

$$\begin{aligned} \log(T_i) &= \mathbf{X}\beta + \epsilon_i \\ T_i &= \exp\{\mathbf{X}\beta\} T_{0,i}, \end{aligned} \tag{2.8}$$

where $T_{0,i}$ is the exponentiated error term ϵ_i, we can arrive to different types of linear survival models via the specification of the error term. The models can be *parametric*, e.g. the exponential and Weibull survival models, or non-parametric, as the Cox proportional hazards model introduced by David R. Cox in 1972, where the shape of the baseline hazard function is left completely unspecified [42]. The *proportional hazards* property means that the shape of the *baseline hazard function* $\lambda_0(t)$ is the same for all observations, but its magnitude depends on the covariates $\exp\{\mathbf{X_i}\beta\}$ s.t. for individual i at time t the hazard is

$$\lambda_i(t|\mathbf{X}_i) = \lambda_0(t) \exp\{\mathbf{X}_i\beta\}. \tag{2.9}$$

Figure 2.1 illustrates the hazard functions of a model where the only covariate is gender. This explanatory variable increases or decreases the baseline hazard by a multiplier that is *constant through time*, realizing the identically shaped gender specific hazard functions. This is the proportionality assumption in the Cox proportional models.

The proportionality assumption is satisfied if each predictor in the model has an identical effect at every point in time. This is a very strong assumption and Singer and Willett argue that "*violations of the proportionality assumption are the rule, not the exception*" [28]. If the assumption is violated, the model can still be used, but the results will contain some amount of bias that is dependent on time. The magnitude of violation can be reduced by incorporating a non-linear interaction between time and the covariates. In the case of a piece-wise constant survival models, the assumption holds only inside a single time interval, i.e. each interval has its own baseline hazard.

Other assumptions of the Cox proportional hazards model that also should be tested for possible violations are the linearity assumption and the assumption of no unobserved heterogeneity. The linear additivity assumption is satisfied if the covariates are independent of each other, i.e. their effect does

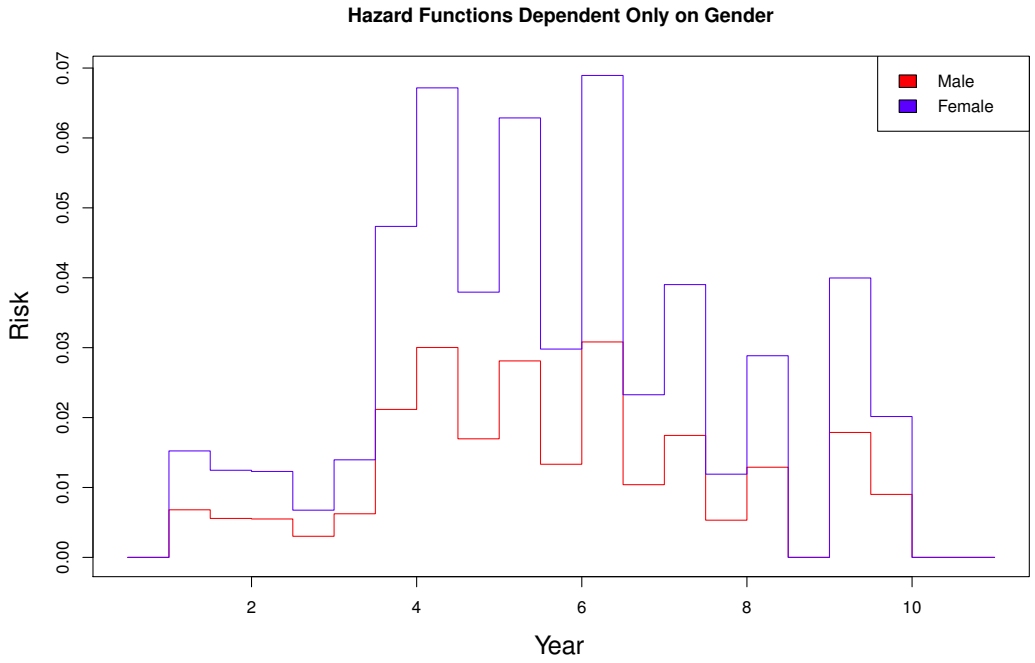

Figure 2.1: Hazard functions for male and female students in a Cox PH model where gender is the only covariate.

not depend on the value of another predictor in the model. The assumption of no unobserved heterogeneity means that the model is assumed to contain *all* significant predictors, i.e. none of the significant predictors have been omitted from the model. Singer and Willett instruct that unobserved heterogeneity has a consistent effect on the time variables and will lead to decreasing hazard functions [43]. [36]

Despite these restrictions (some of which do not apply to more general forms of survival models), there are still several benefits of using survival analysis over traditional methods. In addition to providing a tool for examining the change of risk *in time* and the possibility to include constant and time-dependent covariates, a very important feature is that survival analysis provides a systematic way to deal with censored observations, i.e. observations for which the event of interest does not occur during the time frame of the study [39, 44]. In such cases we either know that the event has occurred before the beginning of the study or that it will occur after the study has ended, but in either case the time of occurrence is unknown. If the event has not occurred by the time the study ends, the observation is called right censored.

Traditional methods have two approaches to this problem. First, the censored observations can be removed from the dataset, but this causes the amount of hazard to be overestimated. Second, the events can be assigned a time of occurrence e.g. by guessing, but this will obviously distort the results. In survival analysis, however, the contribution of a censored observation i in the estimation of the model parameters equals its survival function $S(t_i)$ at the end of the study at time t_i, while the non-censored observation also considers the effect of the hazard function. This makes sense, since if the event is not observed, the only thing we can assume is that the unit (e.g. patient) survived past the end of the study. If the event is observed, the model should be such that it maximizes the probability of the event, which in turn depends on both the survival function and the hazard function s.t. $f(t) = S(t)\lambda(t)$ as shown earlier.

In this thesis we apply a discrete time Cox proportional hazards model with non-linear interaction between time and the model's coefficients to examine how different variables affect the probability of a student to attain a degree. A single discrete time step corresponds to one semester (two semesters per year). In addition we compare the performance of four different methods – Naive Bayes, generalized linear models, support vector machines and Gaussian processes – to predict whether or not the student graduates at all. We briefly introduce these methods in the following sections after which in Chapter 3 we discuss the data used in the analysis.

2.2 Naive Bayes Classifiers

Naive Bayes classifier is a statistical classification method based on the Bayes rule, where the prior distribution and likelihood of the data are normalized into a posterior distribution. In the case of classification, we have the prior probabilities of classes and distributions that describe the probability of data given a class. The posterior then tells the probability of a class given data,

$$Pr(C_k|\mathbf{X}) = \frac{Pr(C_k)Pr(\mathbf{X}|C_k)}{Pr(\mathbf{X})}, \tag{2.10}$$

where C_k is class k, $\mathbf{X}$ is the data and $Pr(\mathbf{X}) = \int_{C_k} Pr(\mathbf{X}|C_k)$ is the probability of data integrated over all possible classes, i.e. the overall probability of data. The model assumes that one dimension or feature in the data does not affect the probability of any other dimension or feature given a class. In other words, it assumes that the dimensions of the data are independent.

The accuracy of the Naive Bayes classifier depends greatly on the probabilities used to model the data. Using Gaussian distributions is a typical

choice, although sometimes kernel density functions can be preferred for increased flexibility. In the end the classification is done by calculating the posterior probabilities of each class given the data and selecting the class with the highest probability as the class of the given data.

2.3 Generalized Linear Models

Linear classifier is a linear combination of explanatory features that maps them into the output score, usually assumed normally distributed. In a two-dimensional problem the space is split by a line into two segments; when dimensions increase, the principle stays the same but line is changed into a hyperplane. The model is fit using an optimization routine to minimize both the prediction error and the weights of the model in order to avoid overfitting.

The generalized linear models allow the response variable to be distributed differently, i.e. it does not need to be normal distributed. For example, if the output should be between $[0, 1]$, a suitable model would be log-odds model, there the logarithm of the odds of an outcome is a linear combination of the covariates.

2.4 Support Vector Machines

Support vector machines (SVMs) are non-linear binary classifiers that aim to maximize the margin between the classification boundary and the data points closest to it in both classes. These closest data points are called support vectors and they define the shape of the classification boundary.

The main idea behind SVMs is to map the data into a higher-dimensional space by non-linear transformations and then fit a hyperplane in that space to separate the two classes as well as possible. The separating hyperplane is then projected back to the original space, where its shape is usually highly non-linear. The accuracy of an SVM depends on the transformation of the data and in non-separable case also on the amount of cost caused by samples that are misclassified. The parameters are usually tuned by trial-and-error.

2.5 Gaussian Processes

A Gaussian process is a generalization of the finite Gaussian distribution to infinite Gaussian process, which defines a distribution over functions rather than over the parameters of a certain form of functions. It is specified by its

mean and covariance function, where the covariance expresses the covariance between the values of the function at points $\mathbf{x}$ and $\mathbf{x}'$. [45]

For Gaussian process classification we squash the output of the process through a function that transforms it into the interval of $[0, 1]$. One possible function for that is the logistic function $\lambda(z) = 1/(1 + e^{-z})$. This enables us to fit a Gaussian process to the probability landscape of the data and find the class of a data point by using the $p(\mathbf{x}) = 0.5$ contour arches as the decision boundary. [45]

Chapter 3

Environment

The survival analysis is performed with student data of Metropolia UAS. Data from multiple institutions would allow comparing school specific attributes such as school size and reputation as done by e.g. Lamote et al. and Guillory [35, 36]. We concentrate on a single organization, because it is prohibitively time consuming to learn the peculiarities and inconsistencies (e.g. why the same attainments have been classified differently at different times) of the data of several institutions, even if the matter experts can be consulted. Secondly, the data of an UAS is in some ways simpler than that of a traditional university e.g. due to the lack of doctoral programs, much stronger separation of candidate and Master's programs and more rigid study plans.

Metropolia UAS is a relatively young institution. Despite being formed in 2008 as a result of a merge between EVTEK UAS and Stadia UAS the student data of Metropolia spans over two decades. This is due to the much longer history of EVTEK and Stadia. The former dates back to 1985 and the latter to 1996, when a merge between ten vocational schools in the capital area formed the Stadia UAS. [46, 47]

These merges have led to the harmonization of the conventions and data to some degree. However, during two decades neither the staff managing the student registers nor the laws concerning universities have remained unaltered. This has led to the data in the student registries becoming incommensurate. For example, before the academic year 2005–2006 students did not have any restrictions in the maximum duration of their studies [48]. Moreover, the study plans have changed and the varying employment rates in different fields of industry has affected the desire of students to drop out and enter the working life instead. Most significantly, the varying conventions among the staff and possible lack of enforcement of common best practices led to many kinds of errors and inconsistencies in the recorded data.

Thus, instead of analyzing the raw data in the institution's study registers, we rely on the data in the higher education achievement register of the National data warehouse for higher education, owned by the Ministry of Education and Culture. The register, entitled Virta, contains information of the attainments, grades, semester enrollments, degrees, study rights, and other relevant records of 1.3 million students of 38 Finnish higher education institutions (universities and UASes). The institutions transfer their data electronically to the register in a format that is defined by the data model for higher education (XDW-model) [49]. Thus the data in the register has identical structure for all institutions, i.e. the institutions have had to make their data commensurable for the Virta register. This has forced them to review the data, fix errors and contradictions, and filter away trash data e.g. dummy students or attainments created for the purposes of testing or output formatting. As a consequence the data quality has improved making it better suitable for reporting and analysis. [50]

Not only the technical restrictions result in higher data quality but also the fact that the national statistics bureau, Statistics Finland, and the Ministry of Education and Culture use Virta as their data source [51, 52]. Because the national statistics and the ministry's funding are based on the data in the Virta register, the institutions have a strong incentive to get their data into Virta as correct and complete as possible. Additionally, a model applicable to one institution's Virta data is applicable to any other institution's Virta data with negligible amount of additional work. Thus the solution presented in the following chapters could be applied potentially nationwide. Before that, however, it is necessary to first describe how the data was transformed into a format suitable for the intended statistical analysis.

3.1 Preprocessing the Dataset

The data in the Virta higher education achievement register consists of several comma separated value (CSV) files. There is one CSV file containing every student's personal information such as gender and birthdate, one for student's semesters, one for the attainment category etc. The data is hierarchical, i.e. it has one-to-many references. At the top of the hierarchy is a person. One person has zero or more study rights. Each study right has zero or more related institutions, statuses, semesters, attainments etc. Each attainment has zero or more related institutions, classifications etc.

To simplify the analysis step, the data is aggregated into one numerical two-dimensional array. Obviously a lot of the data available for analysis is lost, because there is no way to retain e.g. the information of all institu-

tions and their roles of a single attainment after they have been aggregated (summarized) over a semester or a study right. However, everything that is considered to be potentially valuable information for the survival analysis is either included as is or transformed into an appropriate summary statistic.

The data preprocessing involves several steps that are described in detail below. Section 3.1.1 discusses the problem of deciding the most appropriate level of the data hierarchy to build the final dataset from. Section 3.1.2 describes the contents of the resulting dataset and Section 3.1.3 explains the transformations used to construct the dataset from the separate files. Finally, Section 3.1.5 compares the contents of the resulting dataset against summaries reported to the ministry during the previous years and Section 3.2 discusses how the dataset was refined to better suit the intended analysis.

3.1.1 Deciding the Highest Level of Aggregation

The first decision to do is to choose the highest aggregation level, i.e. should the dataset contain summaries over a person, over a study right or something else. Because the goal is to estimate student graduation times and dropout risks, it would make sense to have the student at the highest aggregation level. However, it is quite common for students to change the topic they are studying. For example, someone studying electronics might change into mechanics or someone studying information technology might change into healthcare.

These changes have two consequences. First, it is likely that the student cannot transfer all attainments from the previous field of studies into the new one. Thus the cumulative credit point sum will have sharp jumps down, which do not normally occur (the extent of a course is always positive). This either complicates the predictive model, because it has to account for what appears as negative credit points, or reduces the model's accuracy, because the non-noisy credit point sum must now be treated as noisy. Second, the requirements for progressing in studies and graduation vary between fields of study. There are fields where no restrictions apply and fields where students must e.g. perform some set of studies before they can take on an internship, which in turn is obligatory for progressing any further in their studies. Additionally, the number of credits required before graduating varies between fields of study. Thus the potential graduation date depends on the study field and the full histories of students are not comparable. We conclude that using student as the highest level of aggregation is not a recommendable approach.

Using study right as the highest level of aggregation is problematic as well, because it will lose information of the person (e.g. gender) the study

right belongs to and most importantly, it loses the history of the person's studies if they have more than one study right. The former problem is easily solved by copying the person's information to every study right (s)he has. The history of studies, most importantly the history of study rights, must also be included.

However, this is not possible if the student is from another institution. In such a case it is known only that the student transferred from some other institution, along with any credits transferred into the new institution. These credits are usually summed into modules or degrees, losing information of how the studies progressed. If the student changes fields of study within one institution all data gathered before the transfer is still available after the transfer. Now, at the study right level both cases describe a life span of a single study right, although the amount of background information differs.

This leaves two options for using study rights as the aggregation level: either every study right is treated similarly ignoring the transfer status, in which case the study right history is limited by the amount of information we have on transfer students. The other option is to have separate models for transfer and non-transfer students, where the latter group can use all information available from the previous study rights, including grade point mean, credit point accumulation rate, whether or not the student was a tutor at some point etc. Obviously the latter option allows better fit for the models and is the recommended approach.

There is also the option of aggregating data at the semester or attainment level. The semester level has the same benefits as the study right level described above, but as semesters are comparable, there is no information gained from treating every semester independently. On the contrary, it has the drawback that the amount of data is multiplied as the information of the student and study rights must be copied for every semester. Because the analyzed students are those who are still rolled in after the first year, the model starts to count time from the beginning of the second year rendering the first two semesters' GPA and other attributes time invariant. It should be noted that the person-semester dataset used commonly in regression models are in fact records where the data is aggregated semester-wise [35, 37, 39]. In this thesis we use a similar dataset when the non-linear interaction between time and the constant covariates is included to the model.

Finally, aggregating the data at the attainment level is equal to not aggregating at all. It causes significant repetition of data for a person, study right and semesters while also allowing the time step to be a day instead of a semester. However, this level of accuracy is unnecessarily detailed for both the personnel who follow students' progress and those who estimate future graduation amounts. Therefore, it seems most appropriate to use study

right as the aggregation level when the interaction between time and covariates is not included, and semester level aggregation or *person-period* dataset otherwise.

3.1.2 Contents of the Dataset

As explained in Section 3.1.1, the hierarchical Virta data is aggregated into the study right level. This means that in the resulting dataset each row will contain information related to a single study right. A student with multiple study rights will also have multiple rows in the dataset. Each row will contain personal information of the student, followed by study right specific variables. The student to whom the study right belongs to is described by the following variables:

1. date of birth as days since January first of year 0
2. gender (male/female)
3. mother tongue
4. nationality
5. municipality of residence.

These fields stay constant through all the study rights of a person. The age of a student has been observed to affect the graduation probability by e.g. Lamote et al. and Murtaugh et al. [30, 36]. Bates and Alexander et al. noticed that gender has a significant effect to the duration of studies s.t. female students graduated with higher probability and faster than the male students [37, 53]. However, some studies have found gender to be an uninformative variable [30, 31]. Mother tongue and nationality are included to study the effect of studying in a foreign language or being a foreigner. The former makes studying more troublesome and the latter might weaken the feeling of being a part of the student body or cause anxiety in the student. Municipality of residence is included to point out students who live far away from the campus. Spending a lot of time traveling might affect the study performance in some way, e.g. via the level of social integration with the student body, which has been observed to correlate with student retention [21, 22, 24]. Bates found that out-of-state students graduate earlier than their in-state and international peers. The catch is that they are charged almost double the tuition than in-state residents, creating a "*strong financial incentive for out-of-state students to finish their undergraduate degree within 4 years*" [37]. On the contrary, Murtaugh et al. found that non-residents

were more likely to drop out than were residents and international students [30].

Next, the dataset contains variables describing the past study rights of the student:

6. number of study rights the person had before the current one
7. whether the last study right was in the field of educational studies
8. whether the last study right was in the field of cultural studies
9. whether the last study right was in the field of commercial studies
10. whether the last study right was in the field of technological studies
11. whether the last study right was in the field of healthcare studies
12. whether the last study right was in the field of traveling and services studies
13. whether the last study right was in an unknown study field
14. the number of semesters used during the previous study rights with the same study fields as the current one
15. the number of attained credit points during the previous study rights with the same study fields as the current one
16. whether any of the previous study rights was adult education
17. whether the student was transferred in any earlier study rights.

There are six variables summarizing the past study rights of a student. One of them is expanded into seven true/false variables, totaling 12 variables. The first one tells the number of previous study rights the student has. If this is the student's first known study right, the value is zero. This variable is a direct indicator of the determination and commitment of the student so far. It might also correlate with the probability of graduating during this study right, because attainments from the past study rights can be transferred to the current one. The next seven variables tell the field of studies of the last study right. Combined with the current study right's field of studies it is possible to identify how the change of topic affects graduation. It can e.g. indicate increased motivation to study the new topic or increased risk of the person changing the topic again. Bates found that students who transferred at least once were half as likely to graduate as their peers who did not transfer between academic schools [37]. Similarly, DesJardins et al. found that students who enter the institution with previous college experience have increased risk of stopping out in year one [31]. The number of used

semesters and attained credits during the past study rights can be useful in estimating the student's future rate of progress. For example, it can be used to estimate the amount of time and effort the student has already used in previous studies. Using a lot of time but gaining very few credits creates different expectations of the future progress than the gaining many credits in a short time. The information of whether the student has been an adult student or a transfer student is included to identify whether the study right in question has the adult or transfer student for the first time or not. If a student has once been an adult student, she will be so in all the following study rights as well. Changing from young student to adult student might signify a major change in the life of the student, e.g. the beginning of a demanding day job. The transfer student status is per study right, and thus the knowledge of the previous transfer student statuses is not that useful.

Next, the dataset contains information of the current study right. The related variables are listed below:

18. identification number of the study right
19. identification number of the student the study right belongs to
20. date of the beginning of the right to study
21. date of the end of the right to study
22. whether or not a degree was attained with the study right
23. date when the degree was attained (or 0 if no degree exists).
24. latest state of the study right (active/terminated)
25. study right type
26. type of funding the study right has
27. ID of the institution granting the study right
28. role of the granting institution
29. language of studies
30. municipality of the institution granting the study right
31. the field of study of the study right
32. classification code of the education
33. whether or not the study right belongs to an adult student
34. date of transfer of the study right (0 if not transferred)
35. number of semesters spent at the institution the student was transferred from

36. amount of credit points required for graduating.

The dataset contains the identification numbers of student and study right to enable a closer examination of a specific student or study right. These fields are not used in by the survival model. The start and end dates of the study right allow filtering study rights by time, e.g. filter out all study rights that begun before year 2000. The graduation date is the date when the degree was attained or zero if the student was not observed to attain a degree. If the latest state of the study right indicates that the student is no longer present in the institution, the lack of a degree means the student dropped out. On the contrary, if the student has a degree and no longer studies in the institution, we observe that the "event of graduation" occurred for this study right. As such, these variables are not used as explanatory variables for the remaining duration of studies, but as filters and output values.

The type of the study right defines the level of education (e.g. high school or college) the student is attending. It would be more useful if we were comparing the graduation times of different institutions, but at present it is a dummy variable. This is also the case for the funding type of the study right, granting institution's ID and role. Although they could have different values, in practice all study rights are granted by Metropolia and funded by the typical government funding given to educational institutions. The dataset includes these variables to make it easier in the future to include data from several institutions and compare the effect of the institution itself to the graduation rate and duration of studies.

The language of studies and municipality of the granting organization are included to examine the differences between English and Finnish teaching and between campuses in different municipalities. Some study fields such as information technology can have large portions of teaching in English whereas e.g. healthcare is almost fully taught in Finnish. The English information technology studies does not take place in all campuses evenly, but is focused to certain campuses. Thus these variables have some degree of interference. There is no *a priori* assumptions on how the language of teaching or campus location might affect the duration of studies after the effect of other variables is removed. This is one question the survival analysis might give clarity to.

The topic of studies is identified on a crude level by the field of study and on a detailed level by the classification code of the education. The education code can be used to separate Master's students from Bachelor's students, field of study, study major (or discipline), and study minor (or specialization). Because the requirements of graduation vary between study fields, these variables have significant effect on the distribution of the duration of studies. For example, students of nursing have strict structure and

order in their studies, whereas students of IT and technology have almost no structure or regulations. The classification code defines the total amount of credit points required for graduating, but in some cases this limit has been decreased by considering some parts of it either already attained or not required from this student. Thus the credit point limit needs to be recorded separately. The classification code of the education cannot separate adult students from young students and thus the separation is indicated by a specific field in the dataset. Separating adult students is necessary due to the different ways of studying. Whereas young students usually study during days on campus, adult students have their lessons in the evenings after work and on Saturdays. They progress slower and have more opportunities for distant learning. Lastly, the study right records whether the student has transferred from another institution just before receiving the current study right. If so, the number of semesters spent at the previous institution tells how long that study right lasted. As discussed earlier, these values are useful in estimating the rate of progress of the student's future studies. Next, we discuss the number of credit points transferred from the last study right along with other summaries of the attainments acquired and semesters spent during the study right. These variables are listed below:

38. number of attainments credited in the local institution
39. number of attainments credited in some other institution
40. number of transferred attainments
41. number of attainments in total
42. credit points of attainments credited in the local institution
43. credit points of attainments credited in some other institution
44. credit points of transferred attainments
45. credit points of attainments in total
46. mean of graded attainments
47. credits of graded attainments
48. standard deviation of graded attainments
49. GPA at the end of the study right
50. whether or not one of the attainments is the final thesis
51. whether or not one of the attainments is an internship
52. whether or not one of the attainments is graded abroad
53. number of semesters spent during the study right

54. number of semesters the student was present during the study right
55. number of gaps between study rights (missing semester data).

For each study right the number of attainments, the number of credit points, their ordinary mean, standard deviation, and GPA are attached. The quantity of attainments in number and extent (number of credit points) tells how much work the student has done during the study right. Divided by the duration of the study right we attain an estimate of the student's past efficiency. Mean, standard deviation and GPA are calculated from the graded attainments, which excludes all attainments with no numerical grade. They describe the quality of the work done for the attainments. Especially the importance of GPA in predicting student retention and duration of studies has been recognized in numerous studies [24, 26, 30, 30, 31, 37, 54–61]. For comparison, we include also the ordinary mean which is not weighted by attainment credit point. The standard deviation is included as an indicator of how uniform the student's effort and learning rate have been. Large standard deviation may signal of disturbances in the personal life of the student that affect the motivation or capability to study.

The number of attainments and the amount of credit points are split into three categories: those credited in the local institution, those credited in some other institution while studying in Metropolia UAS and the transferred attainments. This way the amount of work the student has done during the study right and the estimate of the amount of work done prior to the study right in question can be separated. For example, if any attainments are transferred, it must be because the student has been in exchange, and the dataset is able to tell the amount of studying done there. The study right is associated also with flags describing the presence of specific types of studies. Credits from a final thesis describe different kind of work load than those from internship or normal lecture based studies. There is a trivial correlation between the presence of a final thesis and a degree, but as some study rights have thesis without a degree and *vice versa*, both variables are included in the dataset. The semester summaries are used for testing dataset integrity and validating the model. The number of spent semesters is the primary output variable we want to extract from the model. The number of active semesters is included to allow for an option where some culture specific phenomena can be removed in the modeling, e.g. maternal leave which might create a bias for female students or the conscription for male students [62]. The number of gaps between study rights is an error indicator. If the value is greater than zero, the raw data did not contain all semesters it should have, leaving gaps between the start and end dates of the study right's semesters. Such students should be pruned from the dataset before fitting a model to it. Note

that the remaining dataset will still contain students who have their study right in the subject institution even if they are absent for any period of time.

Finally, the dataset contains information of the student's performance on each semester spent during the study right. This includes the following fields:

- whether the student was present or absent during the semester
- mean of the graded attainments received during the semester
- number of graded credit points gained during the semester. A graded attainment has a numerical grade.
- number of all credit points gained during the semester. This includes all grades, e.g. pass/fail.
- GPA of the attainments received during the semester
- cumulative GPA of the attainments received during this and all previous semesters of the current study right.
- the date of the beginning of the semester.

For every semester, the date of the semester and the student's status are recorded. The date is used for filtering e.g. when comparing a particular year's credit points between the dataset and historical reports. The student's performance is measured with five different methods. The simple mean, weighted mean (GPA), and cumulative GPA (CGPA) are included to compare the effect of weighting on the duration modeling. For example Bates reported that increase in the cumulative GPA also increases the probability of graduation whereas DesJardins et al. drew similar conclusions using non-cumulative GPA [31, 37]. The GPA is the attainment extent weighted mean of a single semester's attainments whereas the calculation of CGPA includes all the previous semesters' attainments as well. It is included separately, because one cannot calculate the CGPA by simply summing over consecutive GPAs. The graded credit point sum is included to enable the examination of the $mean \times credits$ and cumulative simple mean variables on the duration model. It seems intuitive that the amount of credit points attained affects the duration of studies more than just the mean grade. In fact, Weng found that the amount of earned credits in the first semester had a significant influence on dropout risk [27]. Multiplying mean grade and number of attained credit points yields a performance weighted credit point count, which might explain well the remaining duration until graduation.

To create a numerical array, the semester values are stored into vectors of constant length such that the first element represents the semester when the student's study right begun, the second element is the next immediate

semester etc. The constant length is set to 15 and all semesters exceeding this count are summed into the last (16th) semester. In other words, the 16th semester contains the number of credit points, mean etc. of all attainments received after the 15th semester. The first 15 semesters correspond to a 7.5 years long time span.

Students who begun their university studies in the academic year 2005–2006 or later are subject to a limited study right duration by a modification of the Finnish university law on August 1st 2005 [48]. The law defined the recommended duration of studies in different study fields and limited the study right duration to at most two years longer than the recommended duration [63]. Currently, the longest study right length is eight years (6+2) for those studying a Master's degree in medicine. However, a student can be absent up to four semesters without reducing their remaining study right duration. Additionally the (compulsory) military service and parental leave do not reduce the study right duration. If the student still cannot graduate and wants to continue their studies, they have to apply for an extension to their study right at the university. This extension is granted on case-by-case basis.

The law for universities of applied sciences states that a degree should be attainable in at most four years excluding some special cases [64]. Thus it is reasonable to assume that a Bachelor's degree in e.g. engineering is attained in seven years (4+2 years plus one from the military service). Indeed, 99.72% of the students who started their studies no sooner than the fall of 2005 finished them within seven years. If all students are included, 99.74% of the study rights ended within eight years. Thus limiting the dataset time series length to 16 semesters will contain all typical cases and truncate only the few exceptionally long study right durations.

This construction enables the analysis of time-dependent variables without restrictions, but as we discuss in later sections, this thesis uses only the first two semesters of the time series as time-independent covariates. The focus is on analyzing the factors affecting the student's graduation after the first year as well as trying to estimate the remaining duration of studies as soon as possible, i.e. right after the first year has passed.

3.1.3 From CSV Files to Arrays

As stated in the beginning of Section 3.1, the Virta register comprises a hierarchical structure of CSV files. Thus the dataset is also produced hierarchically starting from the lowest level and gradually approaching the information at the student or person level. First we read the data from the file system. In the case of attainments we read in the attainment data,

the transferred attainments, the attainment classifications, and associated institution information. Then, we create a single array that combines the contents of these files. The transfer date, if exists, is attached to the attainment, as well as the classification number. However, one attainment can have several classification IDs. In such a case, they "cover" some proportion of the attainment. To simplify the relationship from one-to-many to one-to-one, the attainment's classification ID with largest proportion is selected and the others are discarded. If all portions are equal, one of them is selected randomly.

Attainments are also associated with one or more organizations with different roles. We test whether the source institution of the attainment is the local institution (Metropolia UAS) or some institution abroad. The result (true/false) of this test is attached to the attainment as well. There can be a case where the attainment's source is neither local nor abroad. In that case the only possibility is another national institution from which the attainment was transferred and the attainment should have a transfer date. The attainments in Virta include also degrees, but as they are not attainments *per se*, only the link to the correct study right and the attainment date are retained in their case.

After the attainment array is constructed, the crediting dates of the attainments are altered to compensate for the humane delay in the crediting process. For example, an attainment of a course that ended in December might be credited in January. This effectively transfers the attainment from one semester to another. Thus four weeks is reduced from every attainment's crediting date. This way attainments most likely fall into the time span of the semester they were actually attained instead of the semester during which they happened to be registered by the faculty staff.

Next, we read the semester information from the file system. Semesters are the simplest part of the student data and need no further processing. We move on to read the study rights from the file system. This involves the study rights themselves, the states of the study rights and their time ranges, the institutions associated with the study rights, and the degree field codes.

The municipality where the home campus of the study right is located, the principal language of the studies, an auxiliary student ID, funding type, the education code of the studies and the associated institution along with its role are attached to a study right. In the presence of multiple institutions, the one with the largest portion is picked as in the case of attainments.

Finally we load the student information from the file system. The personal student information consists of the student's ID number, date of birth, gender, mother tongue and municipality of residence. The auxiliary student ID and nationality are attached to the student. If the birth date is missing,

the student is removed from the dataset. In case there is several nationalities associated with a person, one is picked at random. If none exists, the nation is interpreted as "unknown".

At this point it is necessary to fix the study right associations of the attainments. As implied by the national data model for higher education, the Virta register contains only a snapshot of the institution's student register [49]. It has no access to the history of the information, i.e. to how a study right's set of attainments has evolved through time. Due to certain maintenance conventions and limitations in the student register used in Metropolia, attainments are sometimes moved from one study right to another. This happens e.g. when a student in information technology studies for six years, drops out, applies again and is accepted, gets a new study right, and wants to retrieve the attainments from his/her previous study right. The faculty then copies the attainments from the previous study right to the new one and deletes them from the previous study right. This way, when the student finally graduates from the new study right, they correctly get their attainments for the transcript of academic records. However, this is not always the case. Sometimes the old attainments are left intact and sometimes they are accredited or validated in the new study right. Of these three approaches only the last one correctly preserves the history of the studies.

Thus before the separate data arrays are combined, the attainments of each student are reassigned to the study right where they were credited for the first time. This is possible, because the moved attainments retain their original crediting date. If an attainment is reassigned to its correct original study right, the newer study right will retain a copy of the attainment as an accredited or validated attainment. If the crediting date of the attainment is not included in any of the student's study rights' time span, it is assigned to the nearest one. If it is included in more than one study right's time span, it is assigned to the study right with the longest span. The rationale here is that study rights leading to a degree often last longer than e.g. open UAS study rights. This process restores the actual credit point accumulation history of the student, allowing one to estimate the actual amount of studying done during each study right.

Having available the matrices of attainments, semesters, study rights, and students, we aggregate them into one numerical table, where every row is a study right, as described in Section 3.1.1. The approach chosen here is to first sum up the attainments of a semester, then combine the semesters into a time series at the study right level and finally attach the student's study right history and personal information to the study right. This process is discussed in more detail in the next section.

3.1.4 Combining the Separate Data Arrays

The aggregation of the separate data arrays into one numerical array starts by extending the semesters with attainment information. However, many study rights are missing some or all of their semesters due to chosen registry maintenance conventions. For example, a student in open UAS never enroll in for a semester. Thus, semesters are generated from the start to the end of a study right. Existing semesters are left intact and the generated ones will be associated with the same institution as the majority of the existing semesters (of all study rights of every student). A student is always present during a generated semester. It is possible to assume that if student got credit points during a semester, (s)he was present, but if they did not attain credit points, it is not valid to assume that the student was absent. Additionally, because in most cases the study right exists only because the student is going to attain credit points without aiming for a degree (e.g. the open UAS or an exchange student), then by default in such study rights the student should be considered present.

Next the attainments are attached to the semesters. The attainments and semesters of a study right are matched according to the attainment's crediting date. If any attainment is dated earlier than the beginning of the earliest semester, the attainment's date is set to the semester's starting date. Similarly is done for attainments credited later than the end date of the latest semester. In case an attainment falls into a gap between two semesters, it is assigned to the nearest one. The semesters are then sorted by their starting time and the attainments are summarized per semester. This includes calculating the semester's attainment count and the number of credit points of locally credited, elsewhere credited and transferred attainments and the grand total (see Section 3.1.2 for discussion on the different attainment types). In addition, each semester contains variables describing whether a final thesis, an internship or an attainment gained abroad was evaluated during its time span. Finally, the grade mean, number of graded credit points, standard deviation of the grades, the GPA, and the CGPA of the attainments gained during the semester are calculated.

Having aggregated the attainments into semesters, the semesters of a study right are then concatenated into time series as described in Section 3.1.2. The study rights are extended also with information of the presence of a degree and its date. The attainments are aggregated over the entire study right as well as each semester. The grade mean, credits point count, standard deviation of the grades, the GPAn and the CGPA of the entire study right are calculated using attainments associated with it. These sums reveal e.g. the overall study performance during the study right and the

amount of studies done prior to the beginning of this study right in the form of transferred credit points. They also provide verification sums for the semester attainments.

Finally, the student data and summaries of the preceding study rights are attached to the extended study rights. The study rights of a student are sorted by their start time and for each study right after the first one the number of preceding study rights, the field of studies of the last study right and the number of spent semesters and gained credit points during the preceding study rights are calculated. Also, each study right tracks whether any of their predecessors was adult education or whether the student has been a transfer student.

This finishes the preparation of the dataset used in this paper. For every study right there are 55 variables and seven time series, each having 16 samples. In total the dataset contains 104,021 study rights of 88,722 students. In the next chapter we examine the quality of the constructed dataset by calculating some measures from it and comparing them against the published statistics. We then proceed to explore the dataset and form initial hypotheses of the roles of different variables.

3.1.5 Comparing the Dataset with Historical Reports

The contents of the dataset described in the previous sections are compared against the reported number of new study rights, degrees and credit points in the national statistics service Vipunen. It contains statistics and indicators of education, research in universities, the sosioeconomical background of students and the placement of graduated students in the working life [65].

As discussed in the beginning of the chapter, the student registry contains many kinds of errors and inconsistencies. As shown in Sections 3.1.3 and 3.1.4, the data quality in Virta is not perfect either. The statistics supplied to the officials in the past were commonly edited manually to match the known reality better. In other words, someone for example knew that a student graduated in 2007, although according to the student register the graduation was in 2008. Likewise, someone from the faculty involved in gathering the statistics might know that e.g. some attainments credited at the beginning of the fall semester should be included in the spring semester's statistics, contrary to the crediting date of the attainment. Finally, the statistics are just snapshots of the approximate contents of the student registry at each point in time. After the statistics have been gathered and recorded, the student registry continues undergoing modifications: attainments are moved from one study right to another or deleted altogether, graduation dates are altered etc. Thus it is not realistic to assume a perfect match between the

statistics gathered from the dataset and the statistics in Vipunen.

We begin the comparison between the dataset and Vipunen by examining the number of new students per study field and study right type between 1997 and 2013. As we see in Figure 3.1, when summed over the entire time span, the proportions of young, adult and Master's students is very accurate, as are the proportions between different fields of study. Indeed, the overall number of new students in the dataset is approximately 2% greater than in Vipunen. However, when the number of new students is summed over the fields of study, the errors are revealed. Figure 3.2 shows how the number of new students per study right type differs from what has been recorded in Vipunen in different years. The error is intolerably large for students in the adult education program and in the number of new young students it increases quite steadily before peaking in 2005. Finally, the number of new Master's students match perfectly until 2008 when Metropolia begun. The first Master's students graduated in 2003. These errors might be due to changes in the student registry and thus also in the dataset or due to mistakes done by the officials publishing the statistics.

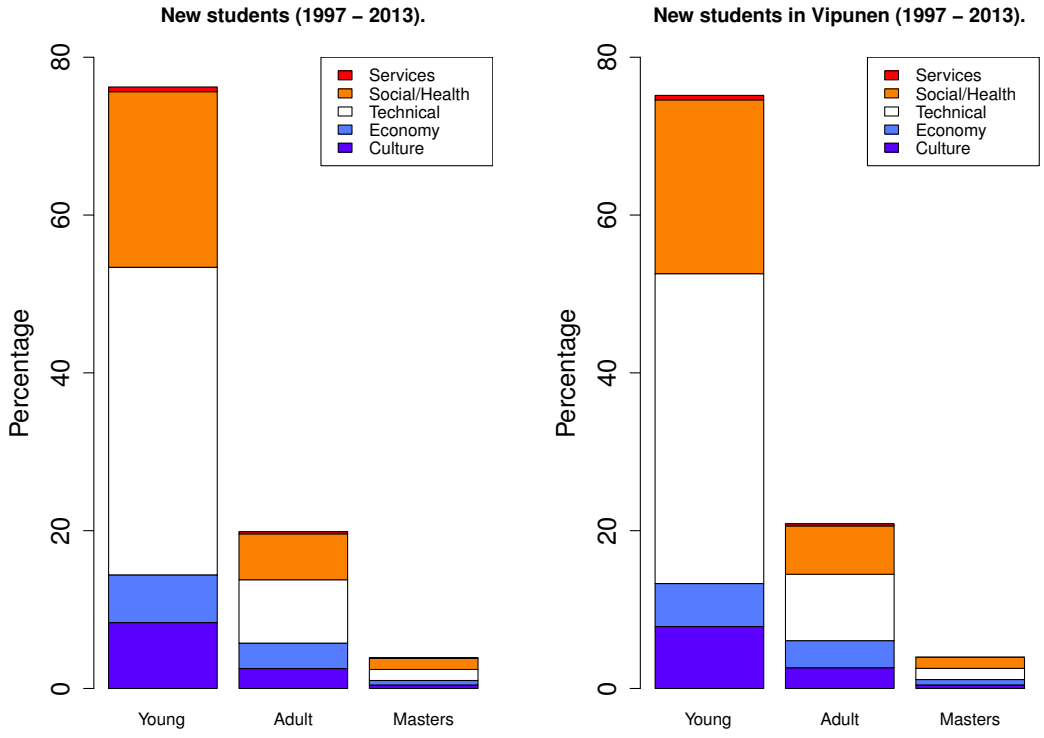

Figure 3.1: Comparison of new study rights between the constructed dataset and Vipunen by study field.

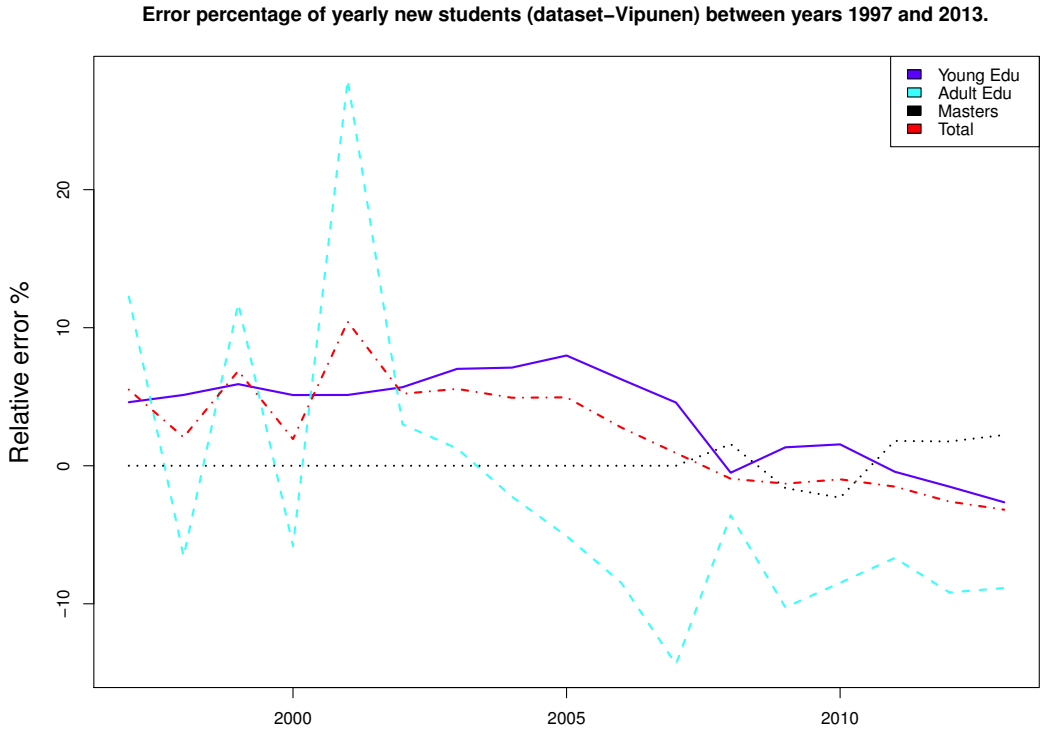

Figure 3.2: Comparison of new study rights between the constructed dataset and Vipunen in yearly basis.

The errors in the number of new students are not clearly reflected in the number of graduated students, shown in Figure 3.3. The grouping here is identical to that in Figure 3.1; however, in this case we can see large deviations between proportions of different fields of study. The services and management field has shrunk to almost nonexistent while the commercial field has swollen much larger than in Vipunen. This applies through all types of study rights. On the other hand, the error in the total number of degrees is less than 0.3%. Observing Figure 3.4 we see that the error in the study field classification is not strongly correlated with the annual error. In fact, the largest annual error between the dataset and Vipunen is less than 2%, which occurs in 2007. This happens to be the last year of operation for Stadia and EVTEK [46, 47]. Most likely the student register underwent large alterations in 2008 causing the observed spike. Another similar error occurs at the beginning of the time series in 2002 and 2003.

Finally we compare the number of credit points attained per study right type and field of study between 2010 and 2014. Figure 3.5 shows the number of attained credit points summed over time. As in the case of new students,

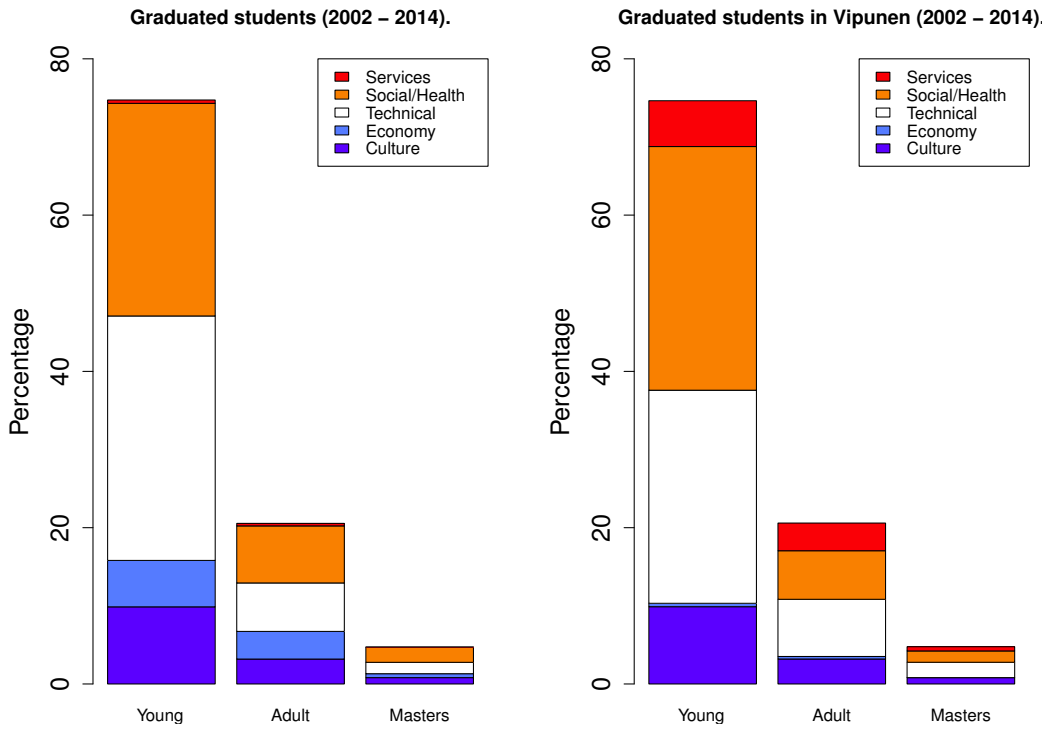

Figure 3.3: Comparison of the number of graduated students between the constructed dataset and Vipunen by study field.

here the figures of the dataset and Vipunen seem to match almost perfectly. There is no visible error between the proportions of study fields either. Figure 3.6 shows large deviations between the dataset and Vipunen when compared year by year for different study right types. Some study right types, like immigrant preparatory education and national student exchange, have errors exceeding 30%. The overall error is not very large on early basis, because the typical degree student's credit points have been quite well recorded in Virta and thus in the dataset.

We conclude that while the official numbers are often not identical with the summaries made from the dataset presented in the previous chapters, the errors are not too large to invalidate the usage of the dataset. However, it is reasonable to assume that the errors are not just noise but may in some cases be systematic. This can lead to biases in the predictions, but as stated above, the differences are in many cases so small that they will probably have only negligible effect on the accuracy of the model. Nevertheless, this must be kept in mind when fitting the model and analyzing the results.

So far the dataset contains data of all available study rights. This allows

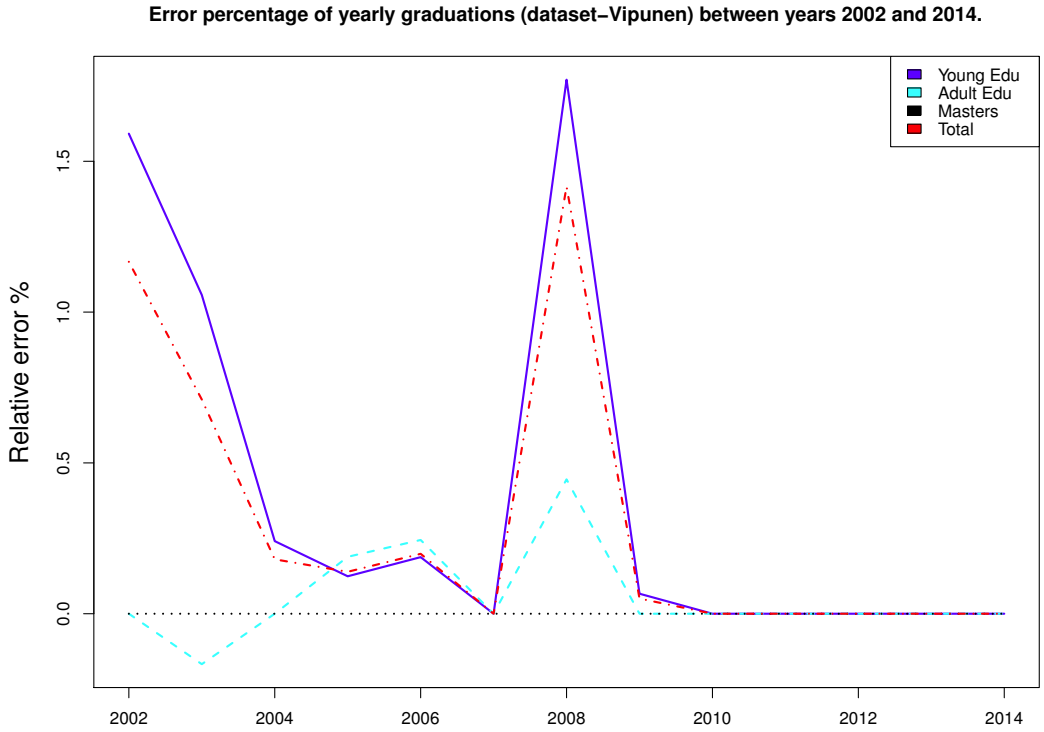

Figure 3.4: Comparison of the number of graduated students between the constructed dataset and Vipunen in yearly basis.

us to take different points of view to the data and enables the comparisons above. In this thesis the focus is on the factors that affect the probability and the amount of time required to attain a degree. Many of the study rights in the dataset do not aim for a degree and should thus be ignored. In addition, some variables' contents are too detailed and others are redundant. In the following section we recognize and remove redundant variables, transform others to better suit the intended form of analysis and eliminate observations that are not in the focus of this thesis. Following this, we present some frequency tables of the key variables.

3.2 Refining the Dataset

The dataset used to compare against the numbers in Vipunen in section 3.1.5 contained 104,021 study rights of 88,722 students. Of those study rights 23,368 do not lead to a degree and are thus left out of the dataset. Attaining a degree is the measure of graduation. In addition the dataset contains some

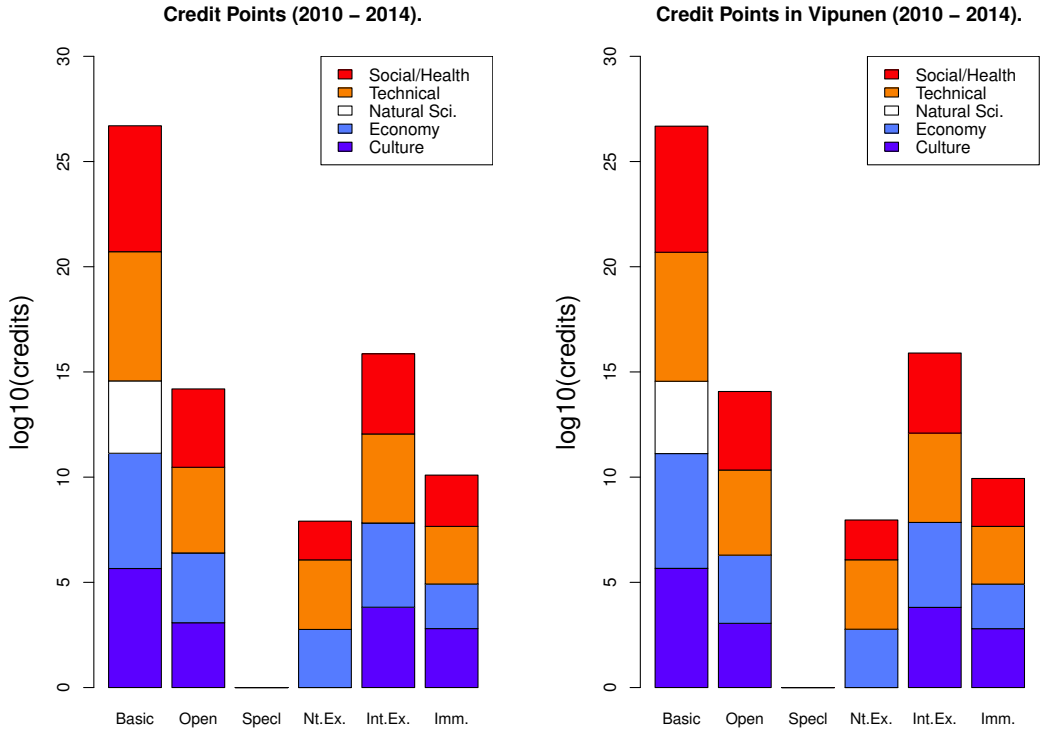

Figure 3.5: Comparison of the number of attained credit points between the constructed dataset and Vipunen by study field. The study right types are from left to right normal/basic students, open university students, all forms of special permission study rights, national transfer students, international exchange students and preparatory education for immigrants.

study rights with gaps between semesters. Such cases are called stop out students in some publications. As discussed in Section 3.1.2 above, they are removed from the dataset. There are also study rights requiring less than 60 credit points at minimum to graduate. Because we know for certain that every degree requires at least one year of full time studies (equal to 60 credit points), such cases must be errors in the data and thus need to be removed from the dataset as well. Finally, there are some students with unknown gender. Even if they are not errors in the dataset, a careful investigation of this minority falls outside the main focus of this thesis. Thus they are removed out as well.

As stated in Section 1.1, we focus on students (or more precisely, study rights) who have exactly one year of studies done. Thus we exclude from the dataset study rights that have either never been present or have no semesters following the first semester they were present. We also exclude all students

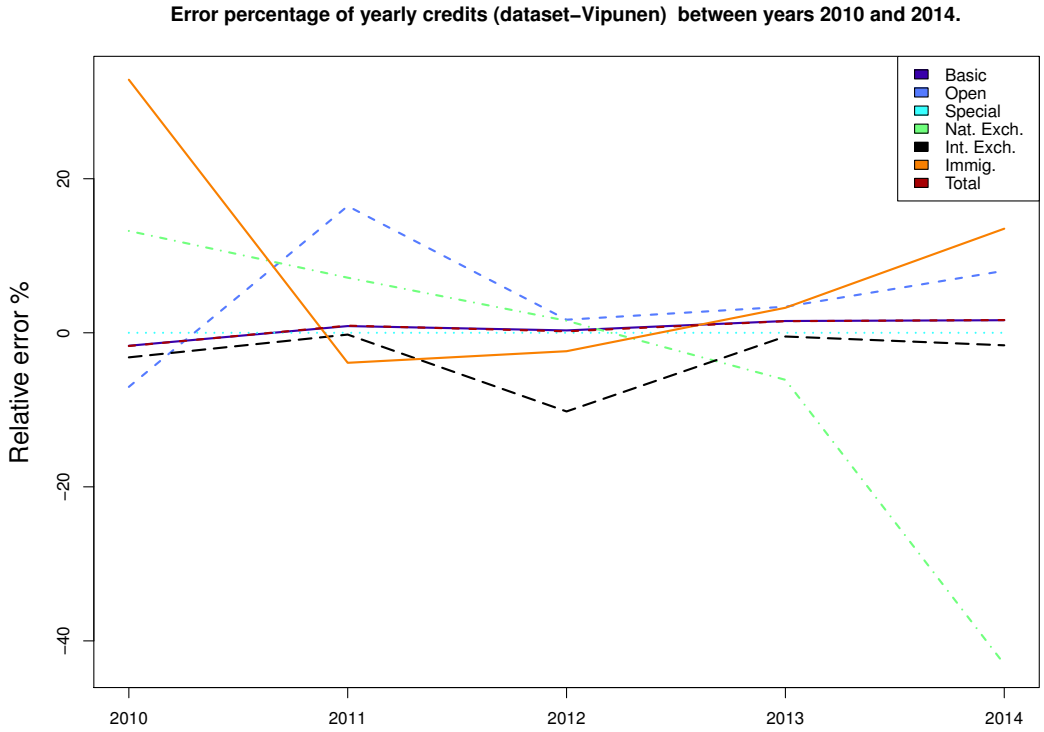

Figure 3.6: Comparison of the number of attained credit points between the constructed dataset and Vipunen in different years.

who have unlimited time to study (i.e. those who began their studies earlier than August 1st, 2005). We do this for two reasons: comparisons are not commensurable if some students have a limit on their remaining time and others do not. The older curricula might have differences compared to their modern versions which increase the mismatch between old and new study rights. Because the aim is to apply the model to future data, we choose to concentrate on the data from the more recent years.

We also decide to discard all transfer students. This is done for two reasons. Firstly, the rate of studies for transfer students is very different from that of a typical student. They come in with well-progressed studies without gaps or pauses and continue more or less from where they got at the previous institution. However, their numbers are small so that some education classification codes have only one or two transfer students. Thus statistical analysis cannot produce informative results with such a small sample size. For this reason, we discard this marginal set of students from the dataset, which reduces the number of study rights available for later analysis by few more than 1,200.

Next, we investigate the variables of the dataset critically. The number of gaps between study rights (variable 55) is zero for all remaining study rights due to the previous steps. Thus the variable can be discarded. Similarly it seems that neither adult nor transfer students have more than one study right per student ID (variables 16 and 17). Thus no study right has previous study rights that were adult education or belonged to a transfer student. These constant fields are also removed. Due to the removal of transfer students from the dataset, we also discard the variable of transfer date (variable 34) as unnecessary. We observe the latest state of the study right, institution and its role (variables 24, 27 and 28, respectively) are constant throughout the dataset and thus we discard them. Finally, we remove the date of degree and the end date of the study right (variables 23 and 21) from the dataset, because they are not needed in the modeling phase.

We make some transformations to the variables in the dataset. First, we replace study right language (variable 29) with a flag indicating whether or not the language is English. Then, we replace the mother tongue (variable 3) with a flag indicating whether or not the student's language differs from the study right's language. For example, a Russian student studies in English, but if the teaching is in Finnish, the flag gets a value of one indicating potential problems in the future progress due to a language barrier. The funding type of the study right (variable 26) is also simplified such that there is only the "base funding" (value zero) and all others (value one). The other type of funding includes funding from the European Social Fund, the Finnish Employment and Economic Development Administration and organizations who want private training or teaching. This type of funding is rare and it is difficult to reason why funding type would be a significant predictor. The municipality of the institution is expanded into three Boolean variables. They indicate the study right's home campus as either Espoo, Helsinki or Vantaa.

The student's age (variable 1) is available at day's accuracy, which is unnecessarily detailed. Instead, we transform the variable to age in years at the beginning of the studies. Similarly, the time of the beginning of studies (variable 20) is in days. We convert it to one variable containing the year of the beginning of studies and one variable indicating whether or not the studies begun during the spring or fall term. Finally, we convert the student's municipality (variable 5) to an indicator of whether or not the student lives more than 50 km away from the campus and simplify the student's nationality (variable 4) to tell the difference between Finnish (zero) and all others (one).

The last phase involves removing redundant variables from the dataset. One such is the municipality of the study right's granting institution, i.e. campus. There are three options, Espoo, Helsinki, and Vantaa. Because the

value is always exactly one of these three, just two indicator variables are enough to explain perfectly the value of the third variable. Thus we can discard the Vantaa indicator variable. The education classification code of the study right has a great descriptive power among the variables. It defines the extent of the study right, the study field, and the study right's type (variables 36, 31 and 25). Thus it is necessary only to have the education code included while the three latter variables can be discarded without loss of information.

These steps leave us with a dataset with 38,876 study rights spanning 11 years from 2005 to 2015. Many of the variables left into the dataset will not be utilized in the modeling phase, because they contain information spanning the entire study right (e.g. standard deviation of all grades). Additionally, because the analysis is constrained on students who have exactly one year of studies behind them, we only need the first two semesters of the time series of the study rights. In the next section, we examine the distributions of retained variables by plotting empirical survival curves.

3.3 Numerical Summaries of the Dataset

To get an understanding of the dataset's contents, we summarize some ratios and distributions of the most interesting variables. Observing the overall percentages we find that 45% of the study rights reach a degree. We observe that 52% of students are male, 4.8% are foreigners, 20% are adult students and 4.3% are Master's students. The number of foreign students is rather low, which might affect the results for them. The number of Master's students is similarly very low. This is typical for universities of applied sciences in Finland. Most of the Bachelor's and Master's students have their campus in Helsinki (62–68%) while foreign students are evenly distributed among all the campuses.

The most interesting summaries are those made per field of study. While overall the percentage of males is only slightly larger, we notice that typically within one field of studies there is either 80% or 30% of males. In fact, only among students of technology (logistics, electronics, IT, wood and metal etc.) are the male students a majority with 80–87% proportion. On the contrary, healthcare (middle wives, therapists, paramedics etc.) has only 12% of men and services (consisting only of cosmetics and beauty services) is 98% female dominated field of studies. For other types of studies, the percentage of males varies in the range of 24–40%.

Greatest proportion of foreign students is in the commercial and technological fields of studies. There are 3–5 times more foreign students in these

two fields than in any other field. Adult students are present in business studies as well, but their proportion is largest in the healthcare studies. Perhaps there is a larger national trend where people who have spent some time in working life want to start doing something completely different.

The worst graduation ratio is among the students of technology, 32–34%. The best graduation ratio is among the Master's students of teaching and pedagogy, over 82%. The Bachelor's students of teaching and pedagogy graduate with 56% probability, which is only a few percentage points above the median graduation ratio. The mean is drawn down by the large amount of technology students, which is roughly 22 times larger than the number of students of teaching and pedagogy.

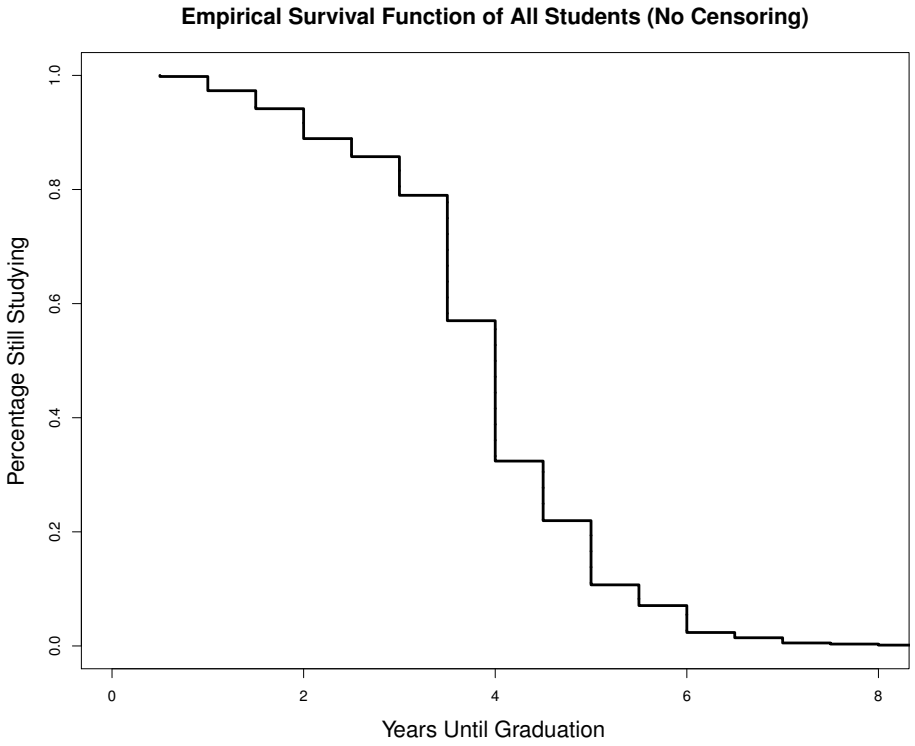

Figure 3.7: Empirical survival function over the entire dataset. Students with all lengths of studies are included.

Next, we examine the dataset via empirical survival functions. Figure 3.7 displays the survival function of all students in the dataset. One can see the curve to start decreasing right after the first year and accelerate around the fourth year. This survival function contains Master's students, whose estimated duration of studies ranges from one to two years, and engineering

students, who are assumed to study for four years. In addition, there are students who have credits from earlier studies and can thus graduate faster than normal. Figure 3.8 contains only Bachelor's students so that one survival function corresponds to one field of studies. Here the slope varies. The initial slope is flat longest for students of technology. For healthcare students the slope around three and half years is the steepest. They tend to study all the way until graduation at rather uniform pace.

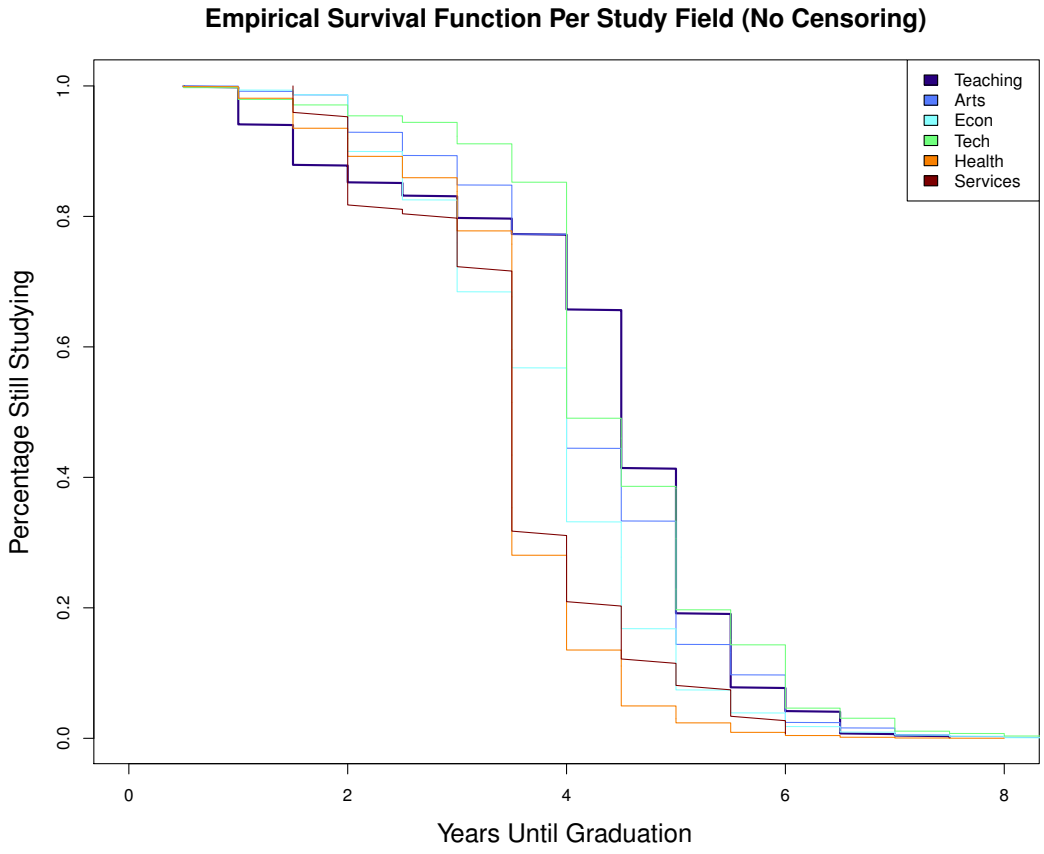

Figure 3.8: Empirical survival function over the entire dataset. The studies are separated by study field.

These initial data exploration and visualization steps prepare us for the next chapter, where we describe in detail the techniques used for data analysis. We pay attention to the details of the survival model, the Cox PH model, and briefly discuss other methods used to predict the graduation status of a student. Finally, in Chapter 5 we discuss the implementation and the observations made from the analysis.

Chapter 4

Modeling the Student Graduation

As explained in Chapter 2, to investigate the factors that affect the duration of studies until graduation, we implement a Cox proportional hazards model with non-linear interaction between time and the model's coefficients. We also apply four machine learning methods to examine the predictability of whether a student eventually graduates or not. In this chapter we discuss how the models are implemented and in Chapter 5 we discuss the results of the models.

We implement two versions of the Cox proportional hazards model. The first one has no time-dependent transformations applied to the data, i.e. all variables have their values fixed before the start of the follow up time. This model uses the length of the studies as the follow up duration, where study rights that do not lead to a degree, i.e. graduation are right censored. This approach assumes that everybody will graduate eventually, but the right censored cases graduate after their follow up period has ended.

The second model assumes non-linear interactions between time and the covariates. The non-linearity is implemented by explicitly defining covariates as functions of time. In this thesis, the covariates change over time according to a fifth order polynomial of time. The polynomial transformation is chosen for simplicity and the fifth order polynomial has the best fit to the mean of the coefficients of the constant variable model. This approach is almost identical to what was done by Guillory [35].

Including non-linear interactions between the covariates and time allows us to conform to the proportionality assumption of the model better, as discussed in Chapter 2. In this model, every semester is a separate, disjoint interval. The censoring is not done for the left side of the interval, because we do not allow for study rights to have a degree before the beginning of the

time interval. Thus the model uses a counting process approach, where we "count" the number of graduations interval by interval. [66]

The variables of the survival model are stratified by education code and adult student status. Stratification means that the population is divided into subpopulations or strata. These subpopulations do not overlap and an individual belongs to only one stratum. According to Lohr, stratified sampling can result in more precise estimates of the population [67]. In the context of the Cox proportional hazards model stratification allows study rights with different education code to have different shape of hazard functions, i.e. the hazards do not need to the proportional in any way. The $-\log(-\log(S))$ curves of the Kaplan–Meier survival functions in Figures 4.1 and 4.2 confirm this. When the proportionality assumption is satisfied, the curves should not overlap. This does not happen, so the different values of the parameter being tested have differently shaped hazard functions and should be stratified.

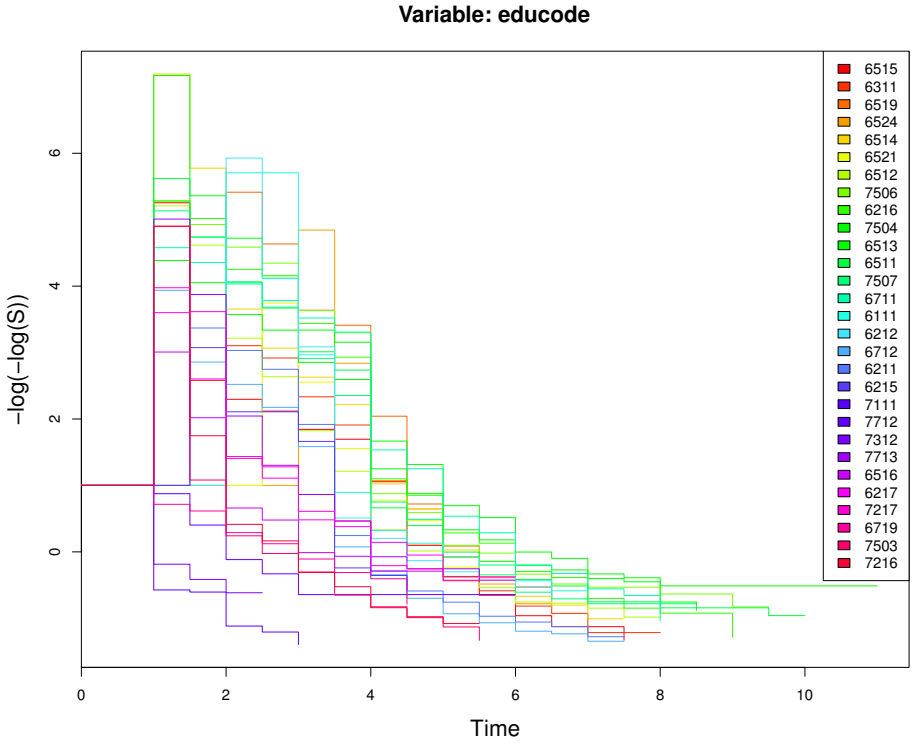

Figure 4.1: The $-\log(-\log(S))$ curves of the Kaplan–Meier survival functions of different education codes. The first digit of a code is 6 for Bachelor's students and 7 for Master's students. The second digit corresponds to the field of studies and the last two digits correspond to the discipline within that field.

Stratifying by education code and adult student status make intuitive sense. Students of healthcare have different requirements and limitations on their studies, e.g. they cannot start working in the field before graduation, whereas students of technology have more studies to take and can start working even before graduating. The studies are very different for young and adult students as well. Young students study by day while adults are more likely to have day jobs and family to care for while studying. They study in the evenings and progress sometimes slower than young students.

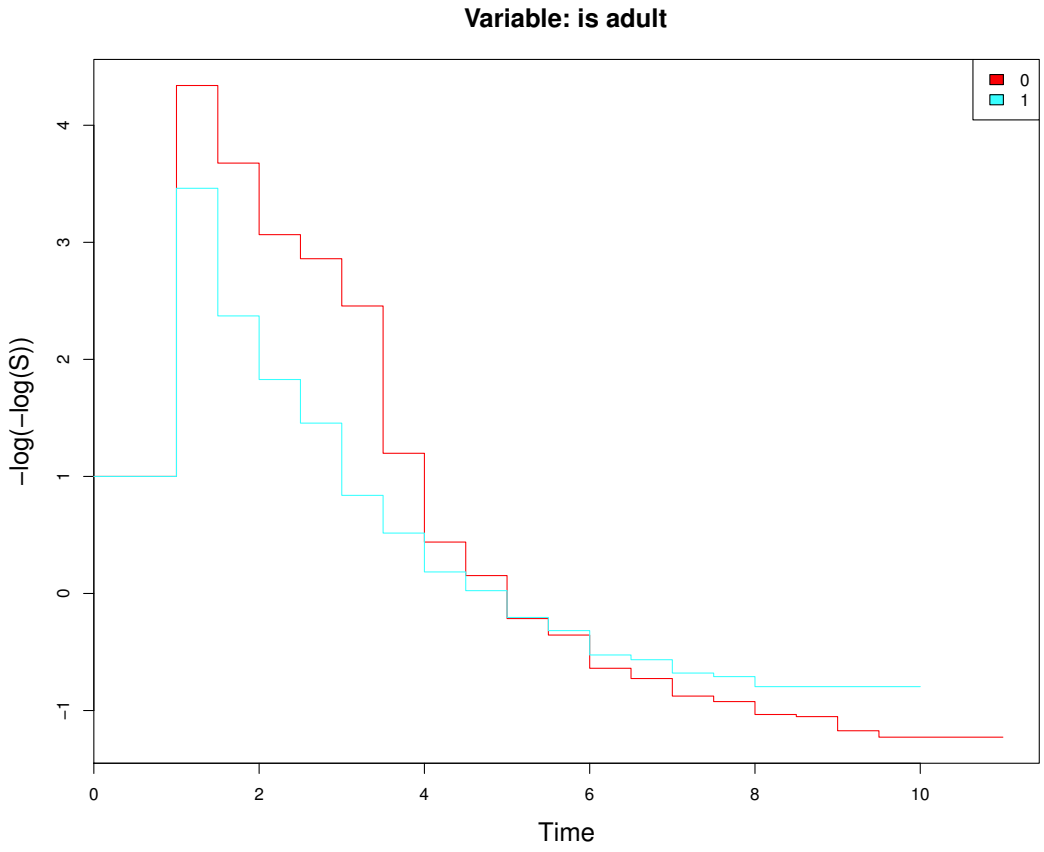

Figure 4.2: The $-\log(-\log(S))$ curves of the Kaplan–Meier survival functions of young (0) and adult (1) students.

We could also stratify by other variables like English speaking students and those whose previous study right was technical. This quickly leads to too small sample sizes as the dataset is split into smaller and smaller partitions by the stratified parameters. When we want to test the model's performance with out-of-sample data, we easily find that the test data has strata levels not found in the training data. Sometimes this is unavoidable, i.e. there's

only one sample with some certain value in a variable in a certain stratum. Thus, although there are more variables that we could stratify according to the proportionality test used above, we leave the variables intact and accept the violation of the proportionality assumption.

The study rights are clustered by student ID. In other words, the model samples the data by clusters instead of individual study rights. This way the weights for students are equal, i.e. a student with several study rights does not become selected more often than a student with only one study right. Taking into account the clusters makes the standard errors of the model more robust.

The model supports categorical variables via the factor property. When a parameter contains separate values each indicating a category, making the factors creates one variable for each separate value of the variable. For example, if the year of the beginning of the studies is factored, the model is then able to estimate the effect on hazard for each year separately instead of the average effect of an increase of the year of the beginning of studies by one. The dataset used in this thesis was constructed so that the multivariate variables are already split into binomial variables and thus the factor property is ignored.

Table 4.1 lists the variables used in the Cox proportional hazards model. The second column describes the way the variable is included in the model and the third column shows which variables have non-linear interaction with time in the second version of the Cox model. Basically all other variables can have interaction except stratified and clustering variables.

From the original 55 variables and 16 time series we have come down to 29 time invariant variables for all students who are still continuing studies after their first year. In the model where non-linear interaction with time is enabled, 26 new variables are included, i.e. there are the time invariant variables as the "baseline" variables and also the time-dependent variables.

To further reduce the number of variables, we optimize the survival models with backward stepwise model selection by AIC and penalized maximum likelihood. The stepwise model selection starts with the full model, dropping out the least significant variable. At every step, all but the most recently dropped variables are re-considered for re-introduction into the model. When adding or removing variables does not improve the model above a threshold value, the optimization routine stops. The penalized maximum likelihood model selection fits the elastic-net regularization path for Cox PH model using cyclical coordinate descent in a path-wise fashion as described by Friedman et al. [68]. These methods give us a statistical view on the significance of the variables explaining student graduation. The optimized models are fit to the training data and their ability to predict the remaining duration of

Variable	Modification	Interaction with Time
Age (in years)	None	Yes
Gender	Factor	Yes
Mother tongue is not Finnish	Factor	Yes
Home country is not Finland	Factor	Yes
Lives far away from campus	Factor	Yes
Number of prev. study rights	None	Yes
Prev. study field was education	Factor	Yes
Prev. study field was culture	Factor	Yes
Prev. s. f. was business	Factor	Yes
Prev. s. f. was technical	Factor	Yes
Prev. s. f. was healthcare	Factor	Yes
Prev. s. f. was services	Factor	Yes
Prev. study field was unknown	Factor	Yes
Sum of earlier semesters	None	Yes
Sum of earlier credit points	None	Yes
Student ID	Clustering	No
Started studies in spring term	Factor	Yes
Studies are not base funded	Factor	Yes
Studies are in English	Factor	Yes
Campus municipality is Espoo	Factor	Yes
Campus municipality is Helsinki	Factor	Yes
Education code	Stratified	No
Is adult student	Stratified	No
Sum of transferred credits	None	Yes
Credits gained during 1st term	None	Yes
Credits gained during 2nd term	None	Yes
GPA of 1st term grades	None	Yes
GPA of 2st term grades	None	Yes
Current semester is spring	Factor	Yes

Table 4.1: Variables in the Cox PH model.

studies is then measured against an out-of-sample test set. The optimized set of variables is then also fitted to a naive Bayes classifier, a generalized linear model, a support vector machine, and a Gaussian process model, and their ability to predict the graduation of study rights is tested. [69]

Using survival models to choose the set of parameters for other models is justified by the fact that the Cox PH model is extremely well suited for quantifying how much a change in a single variable affects the probability of the event occurring (see Chapter 2 for the benefits of using survival models over traditional methods). Weng used a similar approach for identifying the students at risk of dropping out [27]. Instead of survival models he used logistic regression and correlation tests to determine the significant predictors on attrition and then predicted the retention of students using a support vector machine.

The naive Bayes classifier will be implemented with no prior information about the values of the variables. The variable likelihoods are typically estimated by fitting Gaussian distributions to the data, but we estimate the densities using kernel density estimates. This allows for much more flexible distribution shapes which can be far from the Gaussian or other normal distributions. The model does not allow for interactions between variables and thus we need to exclude the interaction with time for the variables used in the naive Bayes model.

In the generalized linear model the distribution of the response variable is best modeled with the Bernoulli distribution, because the student either graduates or not. To do this, we use a binomial distribution to model the output variable with logit link that squashes the linear fit of the model to the range $[0, 1]$. The GML model supports interaction between variables and in that sense it has an advantage over the naive Bayes classifier. This approach was also used by Radcliffe et al.. They used survival analysis to identify factors that impact a student's ability to persist and graduate and then classified first year students as either graduates or non-graduates [70].

The support vector machine is used for C-classification meaning that the response variable is a factor and that the model is optimized using the cost of misclassification variable C. The kernel for transforming the variables into higher-dimensional representation is a radial basis kernel. It is more general than linear or polynomial model in that it is able to find isolated "islands" of occurrences of one class in the output space. The variables C and γ, the radial basis scale parameter, are optimized using a grid search. The implementation of the SVM model supports interaction between variables as well.

Finally, the Gaussian process model will use the logit link between the model and the output variable similar to the GLM. The covariance func-

tion for the data points (study rights) is squared exponential which assumes that similar data points should have similar output values. The covariance function is

$$k(\mathbf{x}, \mathbf{x}') = \sigma_{se}^2 \exp(\|\mathbf{x} - \mathbf{x}'\|^2 / 2l^2), \tag{4.1}$$

where σ_{se}^2 is the magnitude parameter and l is the length scale parameter. They are given priors of Student-t distribution with four degrees of freedom and a squared uniform distribution, respectively. The parameters are estimated using Laplace integration over the latent values and a maximum-a-posteriori estimate and finally optimized by the model. The interaction between the covariates and time is possible via the covariance matrix, if time is included as one input variable. [71]

Apart from the Gaussian process, all models are implemented using the R language. For the Cox PH model we use the `survival` package. The model optimization is performed using the `glmnet` and `MASS` packages. Naive Bayes classification uses the `klaR` package, which extends the implementation in the better known `e1071` package with kernel density estimation of the variable distributions and variable priors. The generalized linear model is implemented using the `stat` package and support vector machine uses the `e1071` package. Finally, we implement the Gaussian process with MATLAB and the `GPStuff` package.

In the next chapter, we present and discuss the results of these models. We examine the results of the optimization routines and compare the predictive accuracies of the models. In Chapter 6 we discuss how the data gathering, preparation and refining could be improved, how the selected models could be further developed and what additional models the future research could study. Finally, in Chapter 7 we present the final conclusions of what we observed from fitting the presented models to the data.

Chapter 5

Evaluation

To evaluate the fit of the survival models to the data with and without interaction with time, we verify the sensibility of the exponentiated coefficients, measure the statistical significance $Pr(Z > |z|)$ of the coefficients and calculate the concordance (C-index), R squared (R^2) and Wald statistics of the models themselves. We also test whether the proportionality assumption holds by examining the correlation coefficient between the (transformed) survival times and the scaled Schoenfeld residuals, the chi-squared (χ^2) statistic of the residuals, and the two-sided p-value of the χ^2 statistic, and also by visually inspecting the Kaplan–Meier survival curves of individual categorical variables and the scaled Schoenfeld residual plots with a Loess curve. Finally, we measure the models' accuracy of predicting the remaining time to graduation. The measurement is done separately for study rights that we know will end to graduation and study rights that are known to be right censored.

We then present the optimization results of the survival models. After removing the non-beneficial variables, we present the concordance, R squared, and predictive accuracies of the optimized models, and compare them briefly with the full models. The best set of variables (without interaction with time) is selected for graduation classifiers and their predictive accuracies are presented. For the GLM we also discuss the model's estimates of the significance of the coefficients for this task. First we introduce the methods we use for measuring the fitness of the models.

5.1 Measurement methods

The validity of the model and its fit to the data are measured in several ways, as described in the beginning of the chapter. We start by examining

the overall sensibility of the output. We then proceed to examine how well the underlying assumptions hold and what is the overall fitness of the model.

The sensibility of the survival model means that the exponentiated coefficients are not absurdly high or low. If they are, there probably is an error in the specification of the model or variables. Anything outside of the range $[0.1, 10]$ raises strong suspicion of a modeling error. The coefficients reported in earlier studies typically settle into the $[0.5, 2]$ range.

The statistical significance of the coefficients calculates the probability of observing at least as extreme Z statistic than what was observed, assuming the null hypothesis holds. The null hypothesis states that a coefficient of a variable is exactly one. If the data strongly supports some other value, then the probability approaches zero and the null hypothesis is discarded. In such a situation we have reached statistical confidence that the variable in fact has an effect on the hazard. Otherwise the variable has no significant effect and can be quite safely discarded.

The concordance or C-index is the proportion of pairs of cases in which the case with higher predicted risk had an event before the case with the lower predicted risk. It basically shows how well the model can "sort" the cases by their risk of having an event. It does not consider the actual values of the predicted times for the event, just their order.

R squared measures the model's fit to data as $1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, where SS_{res} is the squared sum of residuals and SS_{tot} is the squared sum of actual values $\mathbf{y}$ minus the mean of the observed data $\bar{\mathbf{y}}$. If the model is able to explain all of the variance of the actual values from their mean, i.e. the model fits the data perfectly, the R squared gets a value of 1.0. In the presence of noise, outliers or a poorly specified model, the statistic will get values closer to zero. Thus a lower R^2 value does not necessarily invalidate the conclusions drawn of the importance of the variables.

Wald statistic tests the null hypothesis that none of the variables has an effect. In the case of survival models, this corresponds to all coefficients being simultaneously equal to one.

The proportionality assumption is tested visually by plotting the Schoenfeld residuals against time. The Schoenfeld residuals r_{ik} describe how much higher or lower the value of covariate k for individual i is than the average value of all individuals at risk. The time scale can be transformed. Typical transformations are identity, $\log(t)$, rank, and one less Kaplan–Meier survival function value. If the residuals show an increasing or decreasing pattern in the plot, then the proportionality assumption is not satisfied.

Another visual test of proportionality is done by plotting the $-\log(-\log(S))$ transformed Kaplan–Meier survival curves of individual categorical variables against time. If the curves intersect or are clearly non-proportional, we con-

clude that the variable violates the proportionality assumption.

Numerical tests for proportionality include the chi-squared test and its two-sided p-value. The χ^2 test is used to investigate whether the Schoenfeld residuals correlate with time or not. The null hypothesis is that the (scaled) residuals are independent of event times. If the χ^2 statistic is large, then it is unlikely that the residuals are uncorrelated with time. This is indicated by the p-value. In such a case the proportionality assumption must be discarded and the variable either stratified or allowed an interaction with time.

Finally we test the accuracy of predicting the remaining time to graduation. This is done by calculating the mean survival duration as described in Chapter 2. For individuals who did not experience the event before the follow up ended should theoretically have infinite mean survival time. In practice the numerical methods produce finite values. To remove this possibly infinite bias from the estimations of time to graduation, the measurement is done separately for study rights that we know will end to graduation and study rights that are known to be right censored.

The accuracy of the prediction is compared against a benchmark prediction derived directly from the Finnish law. The expected duration of a study right in years is the number of credit points required for a degree divided by 60. Thus, a degree worth 240 credit points should be attained in 4 years. In law, this is called the recommended duration of studies.

5.2 Full Survival Model

5.2.1 Constant Variable Survival Model

The full model with constant covariates is trained with 34,708 study rights of which 14,048 ended to graduation. As such, the dataset in this study is among the largest ones found in the literature. All hazard multipliers, i.e. the exponentiated coefficients of the model, lie within the $[0.58, 1.75]$ interval with at least 95% confidence. Thus no variable more than halves or doubles the baseline hazard function which is unique for each stratum. Statistically significant coefficients on at least 99% confidence are found for the following variables: age at the beginning of studies, gender, number of previously used semesters, whether or not a study right started on spring semester, number of transferred credits from earlier studies, whether or not the study right is base funded, credits and GPA of the first two semesters of studies.

The concordance for the model is 0.73 with the standard error of 0.014. The R squared statistic is very low, only 0.117. It seems that while the model is fairly good at ranking the study rights by their risk, it cannot explain the

variations found in the data, i.e. it cannot predict very well the actual length of studies for a specific student. The Wald statistic gives a value of 1716 on 25 degrees of freedom. This corresponds to approximately zero probability that a constant model can produce the observed data.

The proportionality test is violated by all variables except funding type and country of origin at 95% confidence level. Gender, the previous study right education code being services, and campus location being Helsinki also satisfy the proportionality assumption at 99% confidence level. These confidence levels should be interpreted with care, because the probability of violating the proportionality assumption depends on which study rights are included in the training set.

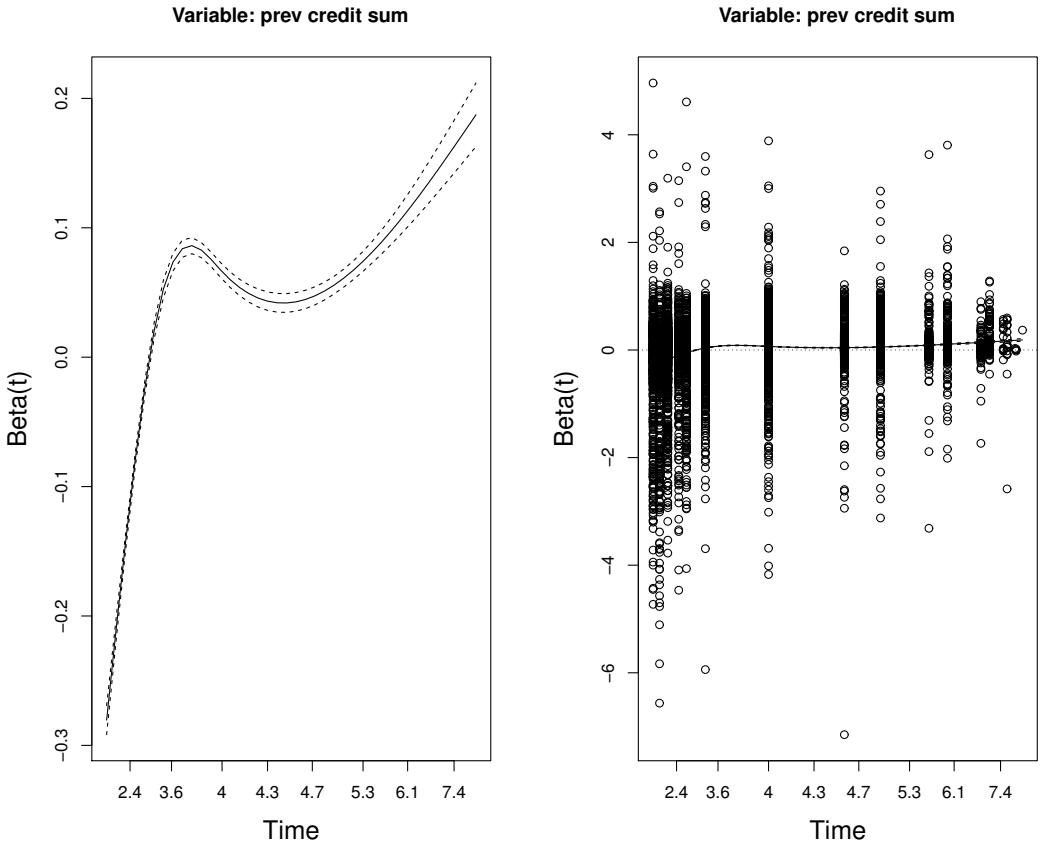

Figure 5.1: The Loess curve and the standard error bands (left) of the scaled Schoenfeld residuals (right). The curve looks like it could be approximated quite well by a fifth order polynomial.

Interpretation of the scaled Schoenfeld residual plots is not entirely straightforward. Figure 5.1 shows on the left the Loess and the standard error bands

of the residuals and on the right also the residuals themselves. The smooth mean residual curve is clearly non-proportional but in the context of the residual values themselves it seems to be perfectly flat. Similar behavior is seen for the variable of the study right's funding type in Figure 5.2. The main difference is that zero is included within the standard error bands more often. We conclude that the proportionality assumption is being violated, but not so severely that it would invalidate the use of the model. For all the rest of the variables, the plots are similar and, thus, are not discussed any further.

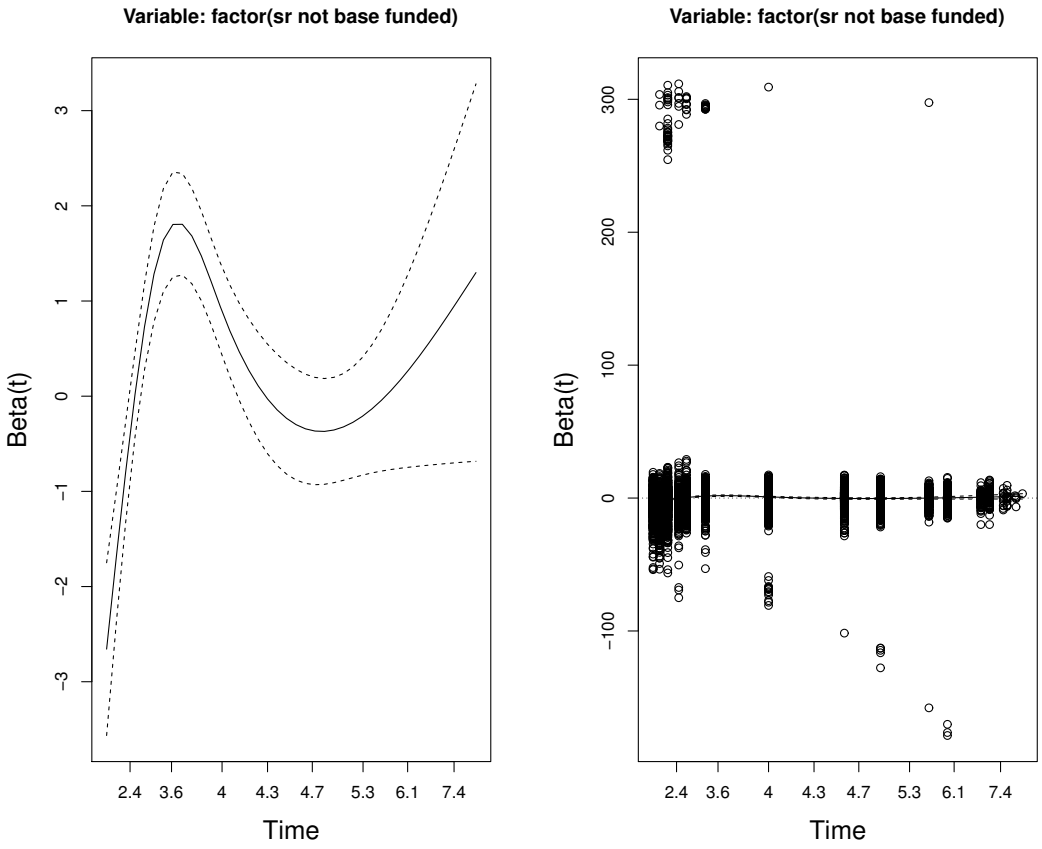

Figure 5.2: The Loess curve and the standard error bands (left) of the scaled Schoenfeld residuals (right).

To further test how well the proportionality is satisfied by the five variables listed above, we plot their $-\log(-\log(S))$ Kaplan–Meier survival curves in Figure 5.3. Most of the variables seem to be almost perfectly proportional, but funding type seems to violate it in the visual inspection. This might be an artefact created by the lack of study rights that are not base funded.

Similarly the number of study rights that were preceded by a study right in service studies is very low, causing the survival curves to look very different.

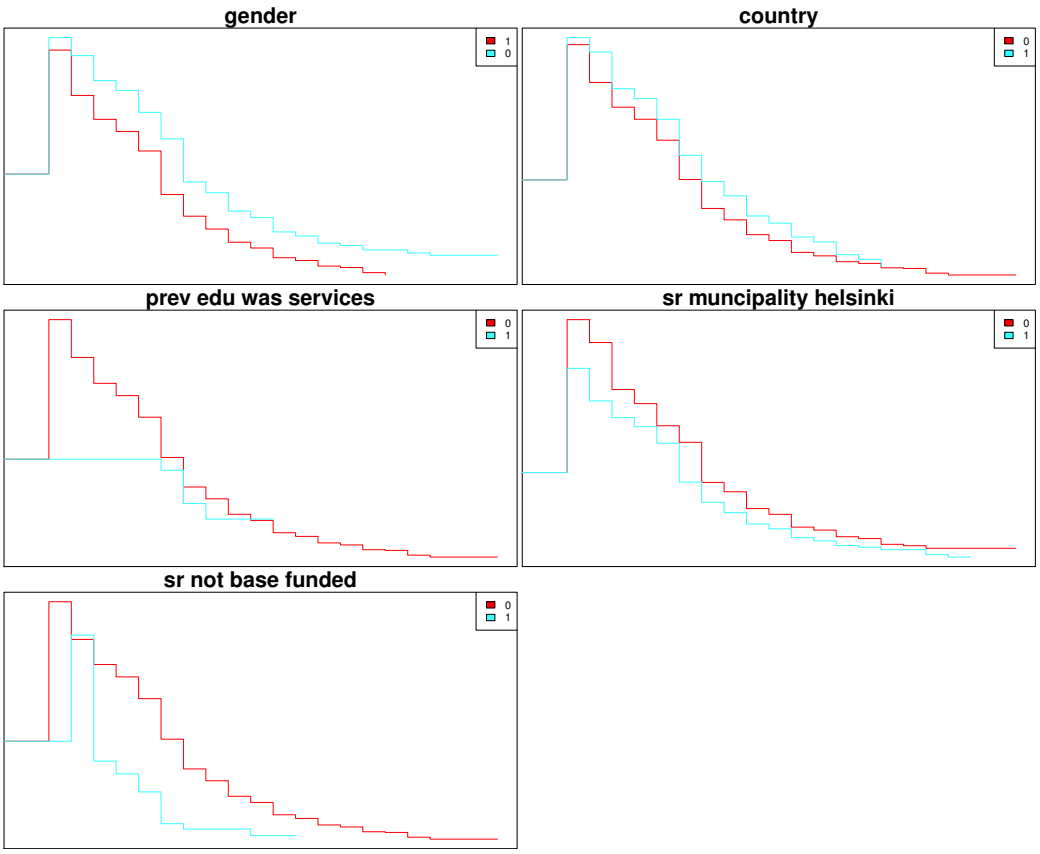

Figure 5.3: The $-\log(-\log(S))$ curves of the variables that satisfy the proportionality assumption with 99% confidence according to the numerical tests.

Figure 5.4 visualizes the Kaplan–Meier survival curves for the rest of the variables. It is clear that in most cases the proportionality assumption is violated, but some of them seem quite proportional, e.g. the variable of whether the study right started in spring and whether the municipality is Espoo. Nevertheless it is clear that the proportionality assumption is violated in the model by several variables. However, the evidence is relatively weak against using the survival model for the study right data.

Having the approximately proportional model at hand, we finally test the predictive accuracy. The mean squared error of the predicted time until graduation is 0.87 years and of the time until the end of the study right

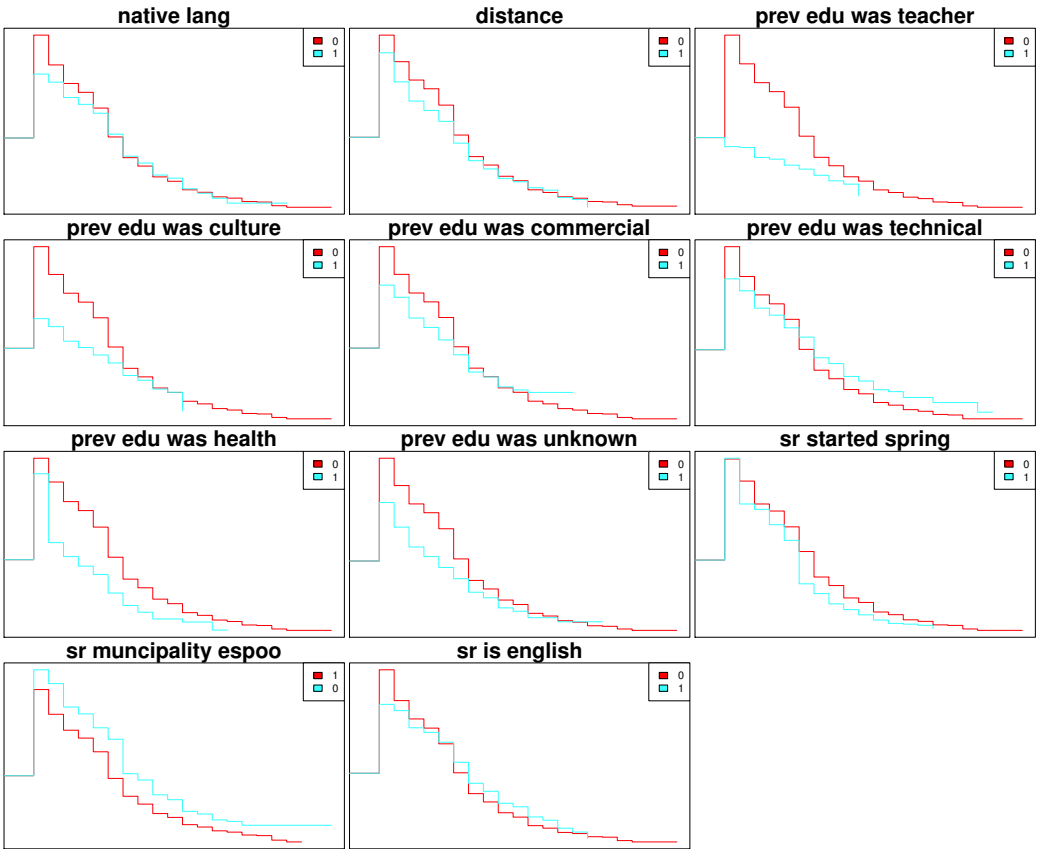

Figure 5.4: The $-\log(-\log(S))$ curves of the variables that violate the proportionality assumption according to the numerical tests.

without graduation is 3.2 years. This demonstrates clearly how the integral of the survival curve of the censored observation is theoretically infinite but in practice a finite large value. The benchmark method gives mean squared errors of 1.08 and 1.88, respectively. We see that by using the survival analysis it is possible to predict the duration remaining until graduation more accurately than with the benchmark method.

5.2.2 Time-Dependent Variables Model

When we allow non-linear interaction between the covariates and time, the number of observations is increased to 214,299 while the number of events stays constant. In the time-varying dataset there is one observation per semester instead of per study right. The exponentiated coefficients of the model lie within a much smaller range $[0.80, 1.50]$ with at least 95% confi-

dence.

The set of statistically significant variables differs somewhat from the constant variable model. Whether or not a student is Finnish and whether or not the previous study right was in healthcare studies are included in the time-varying model while the number of previously used semesters and whether or not study right started on spring semester are not statistically significant anymore. From the variables that interact with time, significant were whether or not the previous study right was cultural studies, whether or not the previous study right was technical studies, number of transferred credits from earlier studies, credits of the first two semesters of studies, and GPA of the second semester.

The concordance for the time-varying model is identical to that of the constant covariate model. The R squared statistic is a bit higher, 0.024. However, the maximum possible R^2 value is almost halved to 0.546. Thus, the two models' ability to explain the variance in the data is essentially identical. The Wald statistic gives a value of 2,235 on 52 degrees of freedom. In these respects the time-varying survival model does not differ from the constant covariate model.

The time-varying model seems to satisfy the proportionality assumption much better than the constant covariate model. Gender, whether the student is Finnish or not, distance from campus, whether or not the previous study right was in technical, healthcare or services studies, both municipality variables, funding type and first semester GPA satisfy the proportionality assumption. Ten out of 26 time-varying variables also seem to be proportional.

The effect of the interaction between the covariates and time is most apparent in the (scaled) Schoenfeld residuals in Figure 5.5. The residuals are more evenly around the zero line and the amount of non-proportionality is reduced by factor of 10 to 1,000 when compared to the constant covariate model. Therefore it is hard to claim that the non-proportionality is significant.

The predictive accuracy of the time-varying model is mixed when compared to the constant covariate model. The mean squared error of the predicted time to graduation is 0.94 (cf. 0.87) years and of the time until the end of the study right without graduation is 2.27 (cf. 3.2) years. The latter comparison is not an improvement, because if a student does not graduate, the time-to-graduation should be infinite, i.e. larger is better. Enabling interaction with time does not improve the concordance, R^2 nor the prediction accuracy of time remaining to graduation. Considering the significant increase in the complexity of the model, it seems that adding interaction with time is not worth the effort when predictions are concerned. This result is opposite to the findings of Calgano et al., who found that allowing interaction

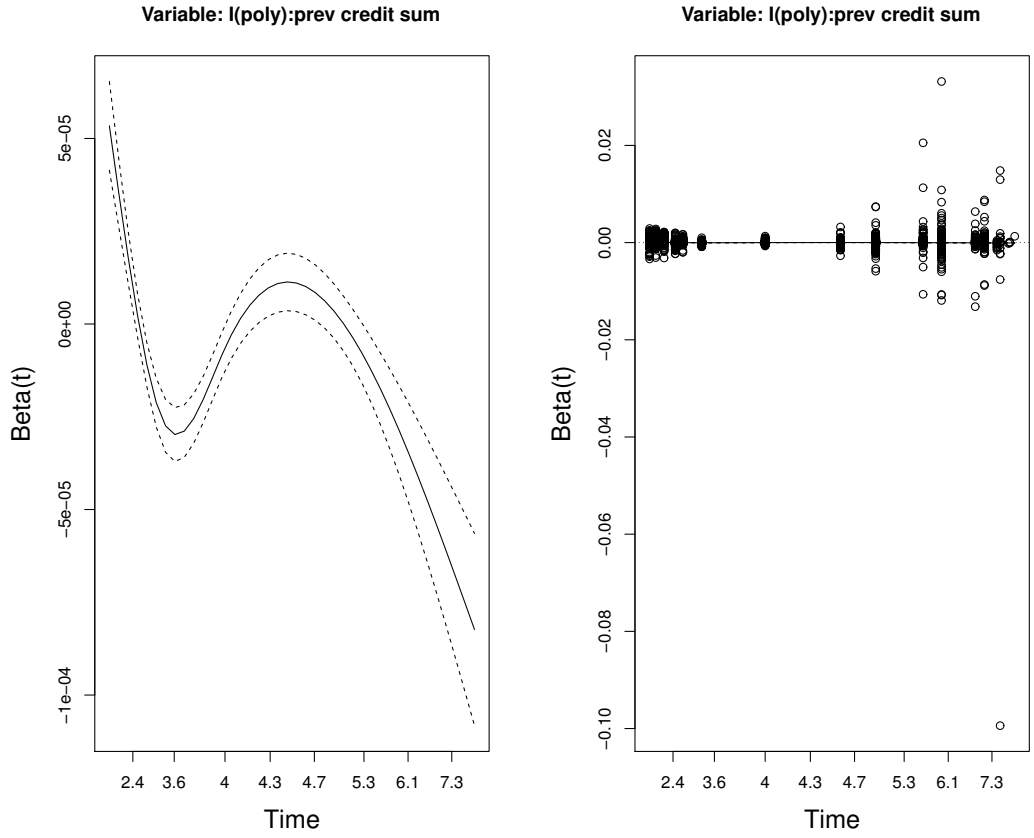

Figure 5.5: The Loess curve and the standard error bands (left) of the scaled Schoenfeld residuals (right) from the time dependent model. The shape and amplitude of the curve is similar to that in Figure 5.1.

with time increased the model's fit to data sufficiently to justify the larger number of parameters [34].

5.3 Survival Model Optimization

Optimizing the constant covariate survival model with the penalized maximum likelihood, elastic-net regularization, and tenfold cross validation preserves almost all of the covariates. Figure 5.6 illustrates this. The penalization weight λ gives the smallest partial log likelihood (i.e. best fit to data) when most or all parameters are included in the model. The parameter count is larger than the number of covariates in the model, because every value of the stratified education code variable is expanded into its own variable.

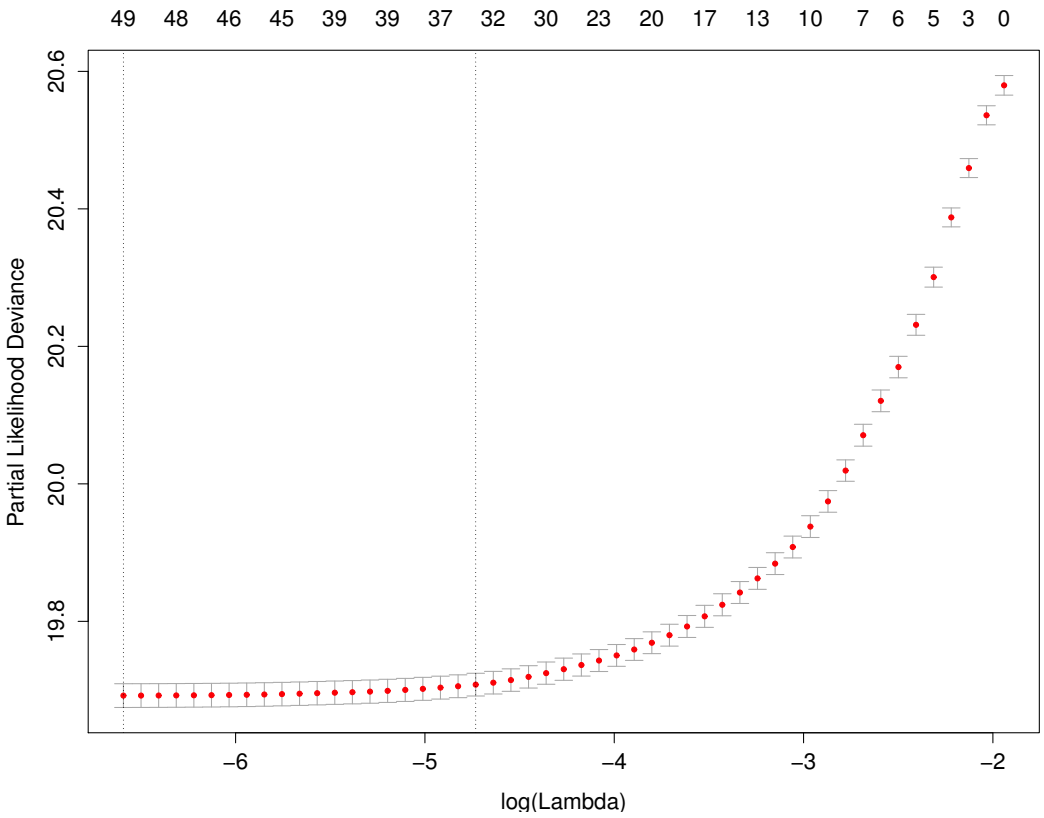

Figure 5.6: Likelihood of the model under different number of parameters. The penalized maximum likelihood optimization routine indicates that more parameters is better.

Only some of the education codes, previous study right count, whether the previous study right was in cultural or business studies and whether the studies were in English got discarded. Clustering was not observed to have much importance and student age was only weakly important. Figure 5.7 shows how the weights of the variables change as a function of the λ weighted regularization term. Even with the loosest regularization (on the right) the coefficients remain in range $[2, 0.5]$.

The stepwise AIC model selection produces somewhat different results. Only 14 variables are included in addition to the stratification and clustering variables. These are listed in Table 5.1. Note that the stratification and clustering variables are not listed, because they affect the result via the segmentation and clustering of data.

While the penalized maximum likelihood optimization dropped out the

Variable	**Significance**
Age (in years)	***
Gender	***
Home country is not Finland	*
Prev. study field was culture	.
Prev. s. f. was technical	***
Sum of earlier semesters	***
Started studies in spring term	**
Native language is not the teaching language	
Campus municipality is Espoo	.
Sum of transferred credits	***
Credits gained during 1st term	***
Credits gained during 2nd term	***
GPA of 1st term grades	***
GPA of 2nd term grades	***

Table 5.1: Variables in the stepwise AIC optimized Cox PH model. The stratifying and clustering variables are not listed, because they are used to identify the hazard function shape.

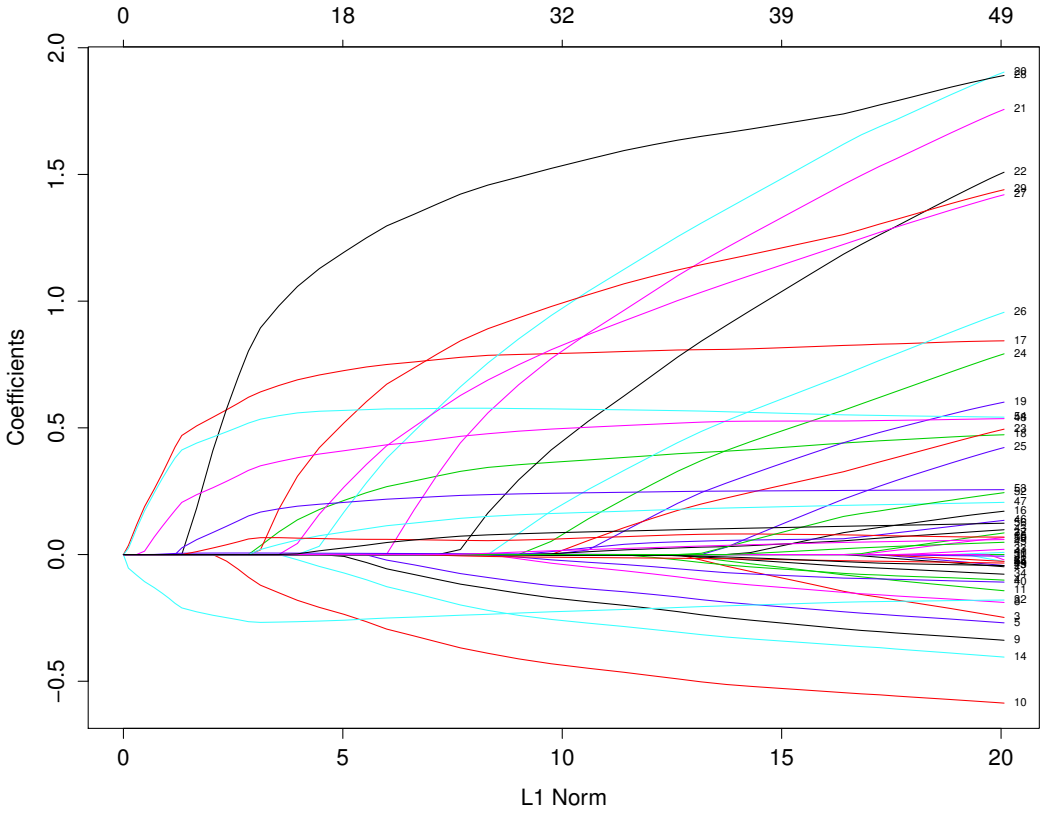

Figure 5.7: Importance of different covariates as the function of the regularization parameter. Even under the tightest regularization scheme (when L1 norm is near zero) three variables are significant.

variables of whether the previous study right was in cultural studies and whether the studies were in English, the stepwise AIC included these variables but with little certainty. In conclusion, the optimization methods produce comparable results given that the penalized maximum likelihood optimization includes more variables but with varying weights while the stepwise AIC optimization includes fewer variables and completely discards the rest.

5.4 Optimized Survival Model

We select the optimal survival model found by the stepwise AIC method. The penalized maximum likelihood method uses all variables with varying weights and is thus not suitable for building a survival model which either has or has not some covariates.

Change in Covariate	**Change in Risk (%)**
Age increases by one year	-0.99
Gender is male	-18
Home country is not Finland	-7.6
Prev. study field was culture	-16
Prev. s. f. was technical	-9.8
Increase of one in the number of previously used semesters	-2.8
Started studies in spring term	+6.7
Studies are in not in the student's mother tongue	-11
Campus municipality is Espoo	-6.8
Transferring degree extent / 100 credit points	+3.3
Gaining degree extent / 100 credit points during 1st term	+3.7
Gaining degree extent / 100 credit points during 2nd term	+5.1
Increase of one in the GPA of 1st term grades	+5.5
Increase of one in the GPA of 2nd term grades	+13

Table 5.2: Effect of optimized model's covariates on the probability of graduating (graduation risk or hazard).

The optimized model performance is as good as the full model with constant covariates. The range of values of the exponentiated coefficients of the model is slightly narrower while concordance, R^2 and predictive accuracies are identical.

The strongest multiplicative effect on the hazard function is the student's gender so that being male reduces it by 18%. The field of studies during the previous study right also has a strong effect. If the previous studies were cultural, the hazard function values are reduced by 16%. A more familiar covariate presented also in most of the earlier works is the GPA. Increase of one point in GPA (scale 1–5, zero is failed) on the second semester increases the hazard function values by 13%. This confirms the observation made by earlier research that second semester grade is a significant factor in identifying at-risk students [27]. Table 5.2 lists the coefficients of the optimized model.

For the most part the results are intuitive. Having better grades or attaining more credit points make graduation more probable. The amount of credit points required for the 3.7 or 5.1 percent increase in the graduation probability equals to the minimum number of credit points necessary for graduation (i.e. study right extent) divided by hundred. In other words, gaining credit points corresponding to one percent of the required extent (e.g. 2.4 points for a 240 credit point study right) qualifies for the reported

increase in the graduation probability. Being older, being male and being a non-native student are in line with the results from the previous studies reviewed in Chapter 2. The institution specific covariates are the previous study rights field and the campus location. It seems that students of culture and technology both graduate less probably than those that have not the history or studying in those fields. Perhaps surprisingly, having previous semesters reduces the probability of graduation while having previous credit points has the opposite effect. Similarly, studying in Espoo causes a reduction of almost seven percent in the graduation probability. This is most likely due to the facts that there are a lot of foreign students, mostly male, and the campus has heavy emphasis on technological studies. This is backed up by the fact that if the student speaking English attends courses held in Finnish, the chances of graduation drop by over ten percent. Finally, starting studies during the spring term increases chances to graduate. The school year starts in the fall and usually those who have spent the fall away (e.g. in military or traveling) start in the spring.

In the interpretation of the coefficients in Table 5.2 the shape of the hazard function itself must be considered. As described in Chapter 4, stratification enables different stratum to have different shape of hazard functions. Thus the most significant variables that affect the graduation probability are the education code and whether or not the student is an adult. The covariate coefficients are simply modifiers of the base hazard functions i.e. the probability to graduate for the first time.

These results are discussed in more depth in Chapter 6. For now we continue this chapter by presenting in the next section the results from the four classifiers – naive Bayes, GLM, SVM and Gaussian process – in the task of predicting whether a student will graduate or become a censored observation.

5.5 Classifier Results

We measure the classifier performance using an out-of-sample dataset. For each method we present the confusion matrix and accuracy, and discuss briefly the factors that might have affected the performance.

5.5.1 Naive Bayes Classification

The naive Bayes classifier is the weakest of the measured methods. With a test set where only 40% of the study rights end to the student graduating, the naive Bayes improves on this baseline result by only four percentage

points reaching an accuracy of 64%. Table 5.3 shows the confusion matrix for the naive Bayes classifier using kernel density estimation. Most errors are made when more than 50% of the non-graduating study rights are classified as to-be-graduated study rights.

		Expected	
		False	True
Observed	False	1,129 (0.29)	242 (0.06)
	True	1,147 (0.30)	1,338 (0.35)

Table 5.3: Confusion matrix of the naive Bayes classification using kernel density estimation for predicting whether a study right will end at the student graduating.

The more surprising observation is made when the kernel densities are replaced by Gaussian distributions for more throughout benchmarking reasons. The simpler distributions actually seem to work better on average attaining an accuracy of 66%. The false positives are reduced by almost the same amount as false negatives are increased (Table 5.4). Apparently the kernel density estimation overfits the data causing the apparent "optimism" in the results.

		Expected	
		False	True
Observed	False	1,453 (0.38)	501 (0.13)
	True	823 (0.21)	1,079 (0.28)

Table 5.4: Confusion matrix of the naive Bayes classification using Gaussian distributions for predicting whether a study right will end at the student graduating.

5.5.2 Generalized Linear Model Classification

The generalized linear model with the binomial distribution and logit link performs significantly better than the naive Bayes classifier with an accuracy of 70%. The number of false positives is almost half of that of the naive Bayes model while the number of false negatives is almost 30% larger (Table 5.5).

		Expected	
		False	True
Observed	False	1,779 (0.46)	673 (0.17)
	True	497 (0.13)	907 (0.24)

Table 5.5: Confusion matrix of the generalized linear model for predicting whether a study right will end at the student graduating.

5.5.3 Support Vector Machine Classification

Support vector machine classification with radial kernel function beats the generalized linear model by a small margin, having an accuracy of 74%. The performance is dependent on two parameters that need to be tuned for the model. Using grid search we find that there is a myriad of combinations that work well as illustrated in Figure 5.8. The optimization result is unexpected in the sense that with large enough value of the regularization parameter, all length scales converge to produce the same accuracy. Nevertheless, we choose a combination of values where neither parameter is too far from 1.0. For the slack cost parameter C we choose value 10 and for the radial basis function length scale λ we choose 0.1.

Table 5.6 shows the confusion matrix of the support vector machine. The values are a bit better than for GLM especially in for the true positives, i.e. for those who will graduate. However, the support vector machines require significantly more computational resources compared to generalized linear models and cannot be justified by these slight improvements.

		Expected	
		False	True
Observed	False	1,775 (0.46)	525 (0.14)
	True	485 (0.12)	1,071 (0.28)

Table 5.6: Confusion matrix of the support vector machine for predicting whether a study right will end at the student graduating.

5.5.4 Gaussian Process Classification

Gaussian processes seem to be the only method not previously applied to student retention or graduation. The most significant drawback of this approach is its exceptionally high computational complexity of $O(n^6)$ which

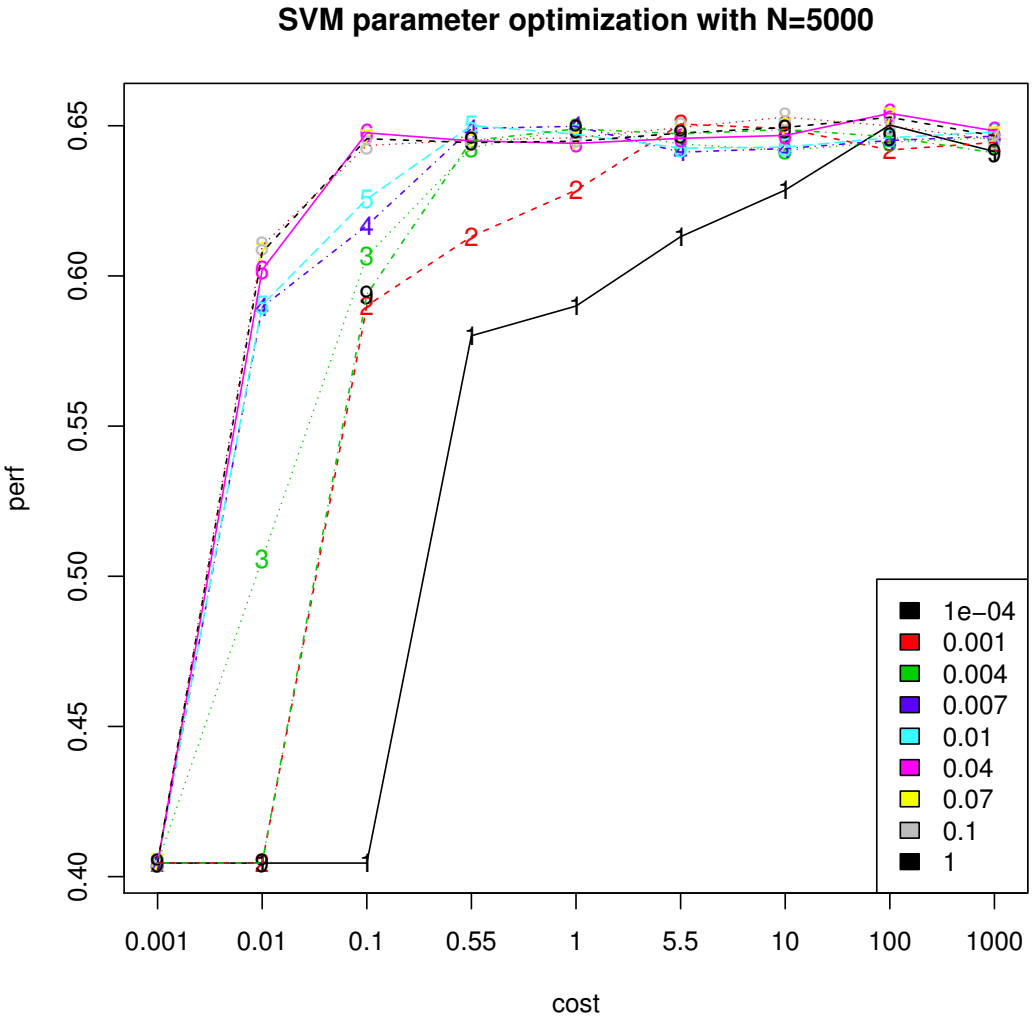

Figure 5.8: Performance of the SVM model with different combinations of the regularization and radial basis function length scale parameters.

limits maximum size of the dataset at hand to only few thousand students.

We use a squared exponential covariate function with length scale prior $\ell = 0.9$ and magnitude prior $\sigma_m^2 = 10$. The inference is approximated using Laplace approximation. Due to the computational limitations, we choose to use the study rights that begun during 2006 as the training set and study rights that begun during 2007 as the test set. Using any more recent data would make some students to start pre-Metropolia and others not. In 2008 the merger caused undesired disturbances in the studies and those students that begun their studies in 2010 have not for the most part graduated yet.

The Gaussian process seems to perform significantly better than the other models despite (or because of) the training set size being 90% smaller than with other methods. Table 5.7 shows the confusion matrix of the method. The portion of true negatives is almost ten percentage points larger than with other methods and the portion of false negatives is reduced to almost one third.

		Expected	
		False	True
Observed	False	1,950 (0.55)	223 (0.06)
	True	441 (0.13)	902 (0.26)

Table 5.7: Confusion matrix of the Gaussian process for predicting whether a study right will end at the student graduating.

With the overall accuracy of 81% Gaussian processes appear to be by far the best method for estimating whether a student will graduate or not. The approach adopted here did not try to find the optimal parameters for the covariance matrix function nor was the dataset stratified by education code and adult student status as with the other models. However, the results are not directly comparable due to the differences in the training set and the way the test set was "sampled". Indeed, support vector machines are able to achieve similar or better accuracy when trained with the same subset of data.

Chapter 6

Discussion

In this chapter, we discuss in depth the results and observations presented in the previous chapter and present ideas for future studies. We do not try to suggest how universities can improve the ratio of graduating students as it requires consideration of psychological, social, and economical processes that are out of the scope of this thesis.

6.1 Results and Observations

The experiment with non-linear interaction between the covariates and time did make the variables conform much better to the proportionality assumption of the Cox PH model, but surprisingly it did not improve the predictive accuracy. This suggests that the proportionality violation was not significant in the first place. The loss of predictive accuracy might be attributed to the change in the dataset: instead of being per study right it is now per semester and instead of being right censored it is now an interval censored counting process. It is important to make the distinction between the interaction-enabled model and the time-varying covariate model used in some of the earlier studies. In this case, we took the time-*invariant* covariates and transformed their value as a function of time to better compensate for the disproportional effect they have as time passes. Time-varying covariates would have meant taking e.g. GPA of the second year's semesters, third year's semesters etc.

The optimal set of 14 variables contains some mysteries as well. Being an older student decreases the chances to graduate by 0.99% per each year, agreeing with earlier research [31, 36]. Male students graduate much less probably in Metropolia than female students, similar to almost all other educational institutions around the world. Being a foreigner decreases the

chances of graduating by almost eight percent. Indeed, in the dataset only 30% of foreign students graduate whereas 40% of native students graduate.

The reasons behind the positive 6.7% effect of starting the studies during the spring term are more difficult to find. It might be due to more adult students starting their studies in the spring (50% more than during fall) or the fact that more than half of the studies starting during the spring term are in healthcare. The corresponding ratio during the fall term is only 19%. Healthcare students typically graduate more probably and more often within the recommended duration due to the legislation imposed limitation to work as a healthcare professional before graduating. There might be alternative explanations, but finding them requires more research.

Having more credits points, whether transferred or attained during the first year, increase the chances of graduating by at least 3.3% for every $extent/100$ transferred credit points to at most 5.1% for every $extent/100$ attained credit points during the second semester. Thus, if a foreign male student that's five years older than the typical student goes to study in Espoo, he should have at least 12% of the study right extent as transferred credit points to be on par with the native female students of the typical age studying at the 'typical' campus.

Espoo campus in Metropolia has a negative effect on graduation that is perhaps not due to the campus itself but other factors: 95% of the students in Espoo are in the field of technology; the ratio of foreign students is three times as high; and 83% of the students are male, whereas the average in all other campuses is 48%. These factors alone explain a lion's share of the negative effect. In future studies it might in fact be beneficial to examine whether the campus itself has any effect having all other variables controlled for.

Finally, there are the students that are spending their second study right, having the previous study right either in cultural or technical studies. In both cases it is statistically significant that their probability of graduation for the first time is reduced by 16% and 9.8% respectively. This effect in fact appears also in other types of studies with notable exceptions the business studies and the 'unknown' previous field of studies which respectively have a positive effect of 1.2% and 10% on graduation probability (see Table 6.1). The empirical frequencies show that nearly 90% of the previous students of technology continue in the same field, while the same number for previous students of culture is 70%. Here perhaps the small population size also affects the results: there are only 154 previous students of culture and they are segmented into tens of different strata. Similar problem might cause the results of the previous students of business although their number is over 350. In line with these findings is the observation that having previously

Change in Covariate	**Change in Risk (%)**
Prev. study field was education	-7.2
Prev. s. f. was culture	-16
Prev. s. f. was business	+1.2
Prev. s. f. was technical	-8.3
Prev. s. f. was healthcare	-8.2
Prev. s. f. was services	-1.4
Prev. s. f. was unknown	+10

Table 6.1: Effect of the study field of the previous study right on the probability of graduating. Depending on the data included in the training set and the convergence of the model, the values can change a few percentages.

spent semesters from the *same* field of studies decreases the probability of graduating by 2.8%. Indeed there are one third more graduations among students with no prior semesters from the *same* field of studies than those who do.

6.2 Future Work

Future research can expand on this thesis in both the data aspect and in the applied methodology. The studied dataset does not contain any variables of the students' financial status, family background, psychological test scores or social integration measures. This disables all efforts to study the theoretical models of student retention, e.g. Tinto's model or Holland's theory of person-environment fit, as done in Allen et al. [57]. In fact, Bates states that "*demographic variables generally account for only a small percentage of explained variance associated with student departure before degree completion*" [37]. One of the most interesting points to focus on would be finding a way to measure the amount of social integration of students in a mass scale via messaging services like IRC and Facebook and then include that data into a survival model. This line of investigation was in fact suggested earlier by Nandeshwar [72].

The dataset used in this thesis has also lost the effect of single courses to the graduation probability. Interesting questions like how do attending the freshmen's orientation courses or the grade of the first math course affect are left unanswered. Such analysis can give information on the critical points in the student's 'educational life span' and allow the faculty staff to focus their

extra effort to those points in time. In addition, it can enable the study of the effect of failing a course. Lamote et al. cite a study made by Rumberger in 1995 where it was observed that repeating a grade almost doubles the risk of dropping out [36].

The methods of analysis can be greatly improved as well. One interesting line of investigation would the application of frailty models to the student data. The approach used by DesJardins seems a lot like frailty models, although they are not referred to as such [31]. The idea of frailty models is to include an unobserved random variable that has a random multiplicative effect on the baseline hazard function of an individual and this way accounts for any unobserved covariates. Using this approach, DesJardins found that when accounting for the unobserved heterogeneity, gender does not anymore have effect on the stop out probability.

A much more promising subject to study is applying Gaussian processes to estimate the dropout risk and time remaining until graduation. There is in fact an implementation of Gaussian process Cox proportional hazards model by Vanhatalo et al. [71]. The requirement for immense computational power can be diluted if not completely removed by using monotonicity information and stochastic variational inference based methods as done by Tolvanen [73]. This approach would completely remove the limitations of the proportionality assumption and allow the calculation of individual hazard functions instead of stratum specific hazards. Considering the benefits of survival analysis over other methods, Gaussian process survival models are very likely to deliver the most accurate and flexible results in the field of educational data analysis.

Chapter 7

Conclusions

In this thesis we estimated the time until graduation and extracted the effects different covariates have on graduation probability or "risk" with survival analysis. We compared a constant Cox proportional hazards model against a non-linear Cox model and found that enabling fifth order polynomial interaction between time and the covariates did not improve the concordance or predictive accuracy enough to make up for the increased complexity.

Predicting the time until graduation was confirmed to be a very difficult task. At best the root mean squared error was 0.87 years or almost 11 months. This is on average only two months better than using the recommended duration of the studies, originating from the law.

Thus the main contribution of using survival models is in the hazard functions and the covariate coefficients. By optimizing the survival model's fit to data we extracted the most important variables affecting the graduation probability. They are age, gender, nationality (Finnish or not), whether or not the previous studies were cultural or technical, number of used semester, which term the studies begun (spring or fall), language, whether the student is at the Espoo campus or not, amount of transferred credit points, GPA, and credits gained during the first two semesters. Being male and having preceding study right in cultural studies decreased the probability to graduate by 30–40%. Starting studies during the spring term, having higher GPA and attaining more credit points all increased the chance of graduation.

These results are not free of the effects of hidden background variables. For example, studying at the Espoo campus most probably has no effect in itself, but it has more foreign, English speaking and technical students and thus primarily male students. All the parameters should ideally be decorrelated, but this was not realistic without severely reducing the interpretability of the results.

Observing that predicting the duration of studies accurately is very diffi-

cult, we also studied whether classifying students accurately to either graduating or non-graduating categories is any easier. We applied naive Bayes, generalized linear model, support vector machine, and Gaussian process classifiers and observed that the support vector machine is able to produce four percentage points better results than the other methods and 14% better than guessing the majority category. The Gaussian process could not be trained with the full dataset due to high computational complexity, so there is still room for improvement.

We conclude that survival models are ideal for estimating the effect of covariates on student graduation. A linear Cox proportional hazard model was observed to be too restricted for real world educational data and thus the two most important directions for future research are accounting for the unobserved heterogeneity with frailty models and removing the linearity and proportionality assumptions with non-linear models, e.g. with Gaussian process survival models.

Bibliography

[1] The Parliament of Finland, "Ammattikorkeakoululaki 2014, § 12: Opetuksen maksuttomuus," 2014.

[2] The Parliament of Finland, "Yliopistolaki 2009, § 8: Opetuksen maksuttomuus," 2009.

[3] Stanford Graduate School of Business, "Cost summary." WWW page, 2015. `https://www.gsb.stanford.edu/programs/mba/financial-aid/cost-summary`. Accessed 3 May 2016.

[4] The Parliament of Finland, "Ammattikorkeakoululaki 2014, §43: Valtion rahoituksen määräytymisperusteet," 2014.

[5] The Parliament of Finland, "Yliopistolaki 2009, 49§: Valtion rahoituksen määräytymisperusteet," 2009.

[6] The Parliament of Finland, "Ammattikorkeakoululaki 2003, 35§: Kuntien rahoitusosuus ammattikorkeakoulujen kustannuksista," 2003.

[7] The Parliament of Finland, "Laki opetus- ja kulttuuritoimen rahoituksesta 2009, 26§: Ammattikorkeakoulujen yksikköhinnat," 2009.

[8] The Parliament of Finland, "Valtioneuvoston asetus opetus- ja kulttuuritoimen rahoituksesta 2009, 15§: Ammattikorkeakoulujen yksikköhintojen laskeminen," 2009.

[9] The Parliament of Finland, "Laki opetus- ja kulttuuritoimen rahoituksesta annetun lain 26 ja 48 §:n muuttamisesta," 2013.

[10] The Parliament of Finland, "Hallituksen esitys eduskunnalle laiksi ammattikorkeakoululain muuttamisesta sekä eräiksi siihen liittyviksi laeiksi." Electronic, 2013.

[11] Finnish Ministry of Education and Culture, "Ehdotus ammattikorkeakoulujen rahoitusmalliksi vuodesta 2014 alkaen." Electronic, 2013. Online version: `http://www.minedu.fi/export/sites/default/OPM/Koulutus/ammattikorkeakoulutus/ammattikorkeakoulu_uudistus/aineistot/liitteet/amk_rahoitusmalli.pdf`. Accessed 3 May 2016.

[12] The Parliament of Finland, "Opetus- ja kulttuuriministeriön asetus 1457/2014 ammattikorkeakoulujen perusrahoituksen laskentakriteereistä." Electronic, December 2014.

[13] The Parliament of Finland, "Hallituksen esitys eduskunnalle laiksi lukiolaissa, ammatillisesta peruskoulutuksesta annetussa laissa ja ammatillisesta aikuiskoulutuksesta annetussa laissa tarkoitetun koulutuksen rahoituksesta ja laeiksi lukiolain, ammatillisesta peruskoulutuksesta annetun lain, ammatillisesta aikuiskoulutuksesta annetun lain 18 §:n ja opiskelijavalintarekisteristä, korkeakoulujen valtakunnallisesta tietovarannosta ja ylioppilastutkintorekisteristä annetun lain 7 ja 9 §:n muuttamisesta." Electronic, 2014.

[14] The Parliament of Finland, "Valtioneuvoston asetus ammattikorkeakouluista." Electronic, February 2014. Online version: `http://www.minedu.fi/export/sites/default/OPM/Koulutus/ammattikorkeakoulutus/ammattikorkeakoulu_uudistus/Liitteet/VNA_asetus_ammattikorkeakouluista_muistio.pdf`. Accessed 3 May 2016.

[15] P. T. T. Ernest T. Pascarella, *How college affects students: A third decade of research.* Jossey-Bass Higher & Adult Education, 2005.

[16] R. D. Reason, "Student variables that predict retention: Recent research and new developments," *NASPA Journal*, vol. 46, no. 4, pp. 172–191, 2003.

[17] J. Sidle, Meg Wright; McReynolds, "The freshman year experience: Student retention and student success.," *NASPA Journal*, vol. 36, pp. 288–300, Summer 1999.

[18] J. Heywood, *Engineering education: research and development in curriculum and instruction.* John Wiley and Sons, 2005.

[19] J. Summerskill, *Dropouts from College.* The American college: A psychological and social interpretation of the higher learning, New York: Wiley, 1962.

[20] A. Bayer, "The college drop-out: factors affecting senior college completion," *Sociology of Education*, vol. 41, no. 3, pp. 306–316, 1968.

[21] W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, no. 1, pp. 64–85, 1970.

[22] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, pp. 89–125, Jan 1975.

[23] J. P. Bean, *Dropouts and turnover: the synthesis of a causal model of student attrition*. PhD thesis, University of Iowa, 1978.

[24] J. P. Bean, "Dropouts and turnover: The synthesis and test of a causal model of student attrition," *Research in Higher Education*, vol. 12, no. 2, pp. 155–187, 1980.

[25] A. Nandeshwar, T. Menzies, and A. Nelson, "Learning patterns of university student retention," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14984 – 14996, 2011.

[26] P. S. Ribeiro, *Strategies to thwart students' attrition in higher education: a data mining approach*. Doctoral thesis proposal, Faculty of Engineering of the University of Porto, July 2013.

[27] F. Weng, *Modelling IT Student Retention at Taiwanese Higher Education Institutions*. Phd dissertation, School of Business Information Technology and Logistics College of Business, RMIT University, June 2010.

[28] J. B. Singer, J. D. & Willett, "Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events," *Psychological Bulletin*, vol. 110, no. 2, pp. 268–290, 1991.

[29] S. L. DesJardins, D. Ahlburg, and M. B., "Studying the determinants of student stopout: Identifying "true" from spurious time-varying effects.," June 1994. Paper presented at the Annual Meeting of the Association for Institutional Research.

[30] P. Murtaugh, L. Burns, and J. Schuster, "Predicting the retention of university students," *Research in Higher Education*, vol. 40, no. 3, pp. 355–371, 1999.

[31] S. DesJardins, D. Ahlburg, and B. McCall, "An event history model of student departure," *Economics of Education Review*, vol. 18, no. 3, pp. 375 – 390, 1999.

[32] T. Lancaster, "Econometric methods for the duration of unemployment," *Econometrica*, vol. 47, pp. 939–956, 1979.

[33] J. W. Vaupel, K. Manton, and E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, vol. 16, pp. 439–454, 1979.

[34] J. C. Calcagno, P. Crosta, T. Bailey, and D. Jenkins, "Does age of entrance affect community college completion probabilities? evidence from a discrete-time hazard model," *Educational Evaluation and Policy Analysis*, vol. 29, no. 3, pp. 218–235, 2007.

[35] C. W. Guillory, *A Multilevel Discrete-Time Hazard Model of Retention Data in Higher Education*. Dissertation, Lousiana State University, Educational Theory, Policy, & Practice, April 2008.

[36] C. Lamote, J. Van Damme, W. Van Den Noortgate, S. Speybroeck, T. Boonen, and J. de Bilde, "Dropout in secondary education: an application of a multilevel discrete-time hazard model accounting for school changes," *Quality & Quantity*, vol. 47, no. 5, pp. 2425–2446, 2013.

[37] L. R. J. Bates, *An event history analysis of time to degree completion*. Rutgers university electronic theses and dissertations, Rutgers University, 2012.

[38] P. D. Allison, *Survival analysis using SAS: A practical guide*. Cary, NC: SAS Press., 1995.

[39] J. D. Willet, J. & Singer, "Investigation onset, cessation, relapse, and recovery: Why you should, and how you can use discrete-time survival analysis to examine event occurrence," *Journal of Consulting and Clinical Psycology*, vol. 61, no. 6, pp. 952–965, 1993.

[40] L. A. E. William Q. Meeker, *Statistical methods for reliability data*. John Wiley and Sons, 1998.

[41] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[42] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society*, vol. 34, no. 2, pp. 187–220, 1972.

[43] J. D. Singer and J. B. Willett, *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.

[44] J. D. Willet, J. & Singer, "From where to when: new methods for studying student dropout and teacher attrition.," *Review of Educational Research*, vol. 61, no. 4, pp. 407–450, 1991.

[45] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[46] Metropolia Ammattikorkeakoulu, "Metropolia ammattikorkeakoulu vuosikertomus 2008." `http://www.metropolia.fi/fileadmin/user_upload/Yleiset/Vuosikertomukset/Metropolian_vuosikertomus_08.pdf`. Accessed 3 May 2016., 2008.

[47] Metropolia Ammattikorkeakoulu, "Metropolia ammattikorkeakoulun tarina." WWW page, 2015. `http://www.metropolia.fi/tietoa-metropoliasta/metropolian-tarina/`. Accessed 3 May 2016.

[48] The Parliament of Finland, "Laki yliopistolain muuttamisesta 2005, § 18 d: Alempien ja ylempien korkeakoulututkintojen tavoitteelliset suorittamisajat, § 18 e: Opiskeluoikeus ja § 18 f: Opiskeluoikeuden jatkaminen," 2005.

[49] CSC.fi, "Korkeakoulujen tietomalli 2.4," November 2015. WWW page: `http://kasitemalli.csc.fi/`. Accessed 3 May 2016.

[50] CSC.fi, "Virta-opintotietopalvelu," November 2015. WWW page: `https://confluence.csc.fi/display/VIRTA/VIRTA-Opintotietopalvelu`. Accessed 3 May 2016.

[51] Statistics Finland, "Syksyn 2015 opiskelijatiedonkeruut," October 2015. Online version: `https://www.tilastokeskus.fi/keruu/jaop/files/saate.pdf`. Accessed 3 May 2016.

[52] CSC.fi, "Tiedonkeruut tietovarannosta," April 2016. WWW page: `https://confluence.csc.fi/display/VIRTA/Tiedonkeruut+tietovarannosta`. Accessed 3 May 2016.

[53] K. L. Alexander, D. R. Entwisle, and N. S. Kabbani, "The dropout process in life course perspective: Early risk factors at home and school.," *Teachers College Record*, vol. 103, pp. 760–822, Oct 2001.

[54] G. S. May and D. E. Chubin, "A retrospective on undergraduate engineering success for underrepresented minority students," *Journal of Engineering Education*, vol. 92, no. 1, pp. 27–39, 2003.

[55] D. L. Tan, "Majors in science, technology, engineering, and mathematics: Gender and ethnic differences in persistence and graduation," *Norman, Okla: Department of Educational Leadership and Policy Studies*, 2002.

[56] R. Alkhasawneh, *Developing a hybrid model to predict student first year retention and academic success in STEM disciplines using neural networks*. Dissertation, Virginia Commonwealth University, School of Engineering, 2011.

[57] S. B. R. Jeff Allen, "Prediction of college major persistence based on vocational interests, academic preparation, and first-year academic performance," *Research in Higher Education*, vol. 49, pp. 62–79, February 2008.

[58] A. J. Bowers, "Grades and graduation: A longitudinal risk perspective to identify student dropouts," *The Journal of Educational Research*, vol. 103, no. 3, pp. 191–207, 2010.

[59] S. L. DesJardins, D. A. Ahlburg, and B. P. McCall, "A temporal investigation of factors related to timely degree completion," *Journal of Higher Education*, vol. 73, no. 5, pp. 555–581, 2002.

[60] S. Hu and E. P. S. John, "Student persistence in a public higher education system," *Journal of Higher Education*, vol. 72, no. 3, pp. 265–286, 2001.

[61] T. T. Ishitani and S. L. DesJardins, "A longitudinal investigation of dropout from college in the united states," *Journal of college student retention: research, theory & Practice*, vol. 4, no. 2, pp. 173–201, 2002.

[62] The Parliament of Finland, "Asevelvollisuuslaki," 2007.

[63] The Parliament of Finland, "Yliopistolaki 2009, § 40: Alempien ja ylempien korkeakoulututkintojen tavoitteelliset suorittamisajat, § 41: Opiskeluoikeus, § 42: Opiskeluoikeuden jatkaminen ja § 43: Opiskeluoikeuden menettäminen," 2009.

[64] The Parliament of Finland, "Ammattikorkeakoululaki 2003, § 19: Koulutusohjelmat ja opetussuunnitelmat," 2003.

[65] The Finnish National Board of Education and the Ministry of Education and Culture, "Vipunen - opetushallinnon tilastopalvelu." WWW page, April 2015. `https://vipunen.fi/fi-fi`. Accessed 3 May 2016.

[66] S. M. David W. Hosmer, Stanley Lemeshow, *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition*. John Wiley & Sons, April 2008.

[67] S. Lohr, *Sampling: Design and analysis.* New York: Duxbury Press., 1999.

[68] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[69] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. Statistics and Computing, Springer-Verlag New York, 4 ed., 2002.

[70] P. Radcliffe and J. Kellogg, "Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis," in *2006 AIRUM Conference, Bloomington, MN*, November 2006.

[71] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "Gpstuff: Bayesian modeling with gaussian processes," *Journal of Machine Learning Research*, vol. 14, pp. 1175–1179, 2013.

[72] A. R. Nandeshwar, *Longitudinal Study of First-Time Freshmen Using Data Mining*. Dissertations/theses - doctoral dissertations, West Virginia University, 2010.

[73] V. Tolvanen, "Gaussian processes with monotonicity constraint for big data," Master's thesis, Aalto university, school of electrical engineering, May 2014.