University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Janika Smirnov

Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques

Financial and Actuarial Mathematics Curriculum
Master's thesis (30 ECP)

Supervisors:
Imbi Traat, PhD
Peep Küngas, PhD

Tartu 2016

Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques

Master's thesis

Janika Smirnov

Abstract. The aim of this thesis is to explore possibilities of modelling late payment times of invoices in business-to-business sales process using real data of sales ledgers. Survival analysis and a novel ensemble method of Random Survival Forests is applied to the right-censored data of late invoices. A theoretical overview of Random Survival Forests is given and concordance index as a performance measure for survival models is explained. A comprehensive overview of data preprocessing and deriving payment times from sales ledgers is presented. We propose two separate models, for first-time debtors and for repeated debtors, and explore the effect of different predictors in a model. Random Survival Forests prove to have advantages over Cox Proportional Hazards model as there are no underlying assumptions that need to be taken into consideration. Overall, it is concluded that Random Survival Forests model which additionally uses historical payment behaviour of debtors, performs the best in ranking payment times of late invoices.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics.

Keywords: Survival Analysis, Machine Learning, Random Survival Forests, Late Invoices, Sales Ledger, Censoring.

Ülетähtaegsete arvete tasumisaegade modelleerimine elukestusanalüüsi ja juhuslike metsade meetoditega

Magistritöö

Janika Smirnov

Lühikokkuvõte. Käesoleva magistritöö eesmärgiks on uurida võimalusi ettevõtetevahelises müügiprotsessis ületähtaegsete arvete makseaegade modelleerimiseks kasutades müügireskontrote reaalseid andmeid. Paremalt tsenseeritud andmetele rakendatakse elukestusanalüüsi ja uudset juhuslike eluskestusmetsade meetodit. Tutvustatakse juhuslike elukestusmetsade teooriat ja samasuunalisuse (*concordance*) näitajat kui headuse mõõdikut. Antakse kõikehõlmav ülevaade müügireskontrote töötlemisest ja arvete makseaegade tuletamisest. Magistritöö käigus luuakse kaks mudelit – üks esmakordsete võlgnike makseaegade modelleerimiseks ja teine korduvate võlgnike jaoks ning uuritakse erineva sisuga tunnuste mõju mudeli ennustustäpsusele. Juhuslike elukestusmetsade meetod osutub eelistatumaks, sest vastupidiselt Coxi võrdeliste riskide mudelile ei nõua see meetod täiendavaid eeldusi, mis peavad olema täidetud. Lõpptulemusena järeldatakse, et parima täpsusega on juhuslike elukestusmetsade mudel, mis võtab arvesse ka võlglaste ajaloolist maksekäitumist.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.
Märksõnad: Elukestusanalüüs, masinõpe, juhuslikud elukestusmetsad, ületähtaegsed arved, müügireskontro, tsenseerimine.

Contents

Introduction

Account receivables collection is an important part of business management in every firm. Poorly managed cash flow process could lead to significant liquidity problems. This, in particular, is critical for smaller firms that are more dependent on the trade credit [1]. Ineffective collection of account receivables may lead to cash shortage in a firm and that, in turn, could lead to problems meeting liabilities to its suppliers.

Therefore, liquidity problems and inability to pay its debts in one firm may have a snowball effect. In worst case scenario, creditor that often receives late payments may need to take short-term loans to overcome liquidity problems ([2], p. 7). Ability to reliably forecast the cash flow and make wise decisions whether to credit sales invoices to some debtors or not, is an essential part of the business process.

Credit management firms provide assistance to companies in their account receivables collection and give recommendations whether it is advised to grant further credit to some debtors. The decision is based on the creditworthiness of the debtor at the current moment and the decision making process is manual, time consuming and therefore expensive in terms of labor costs.

The purpose of this thesis is to perform experimental analysis to model the payment behaviour of the debtors and find solutions to provide predictive analysis that could be used in the decision making of account receivables collection. It is of interest to reduce the need for manual assessment of creditworthiness of the debtors.

The subject of our work is conditional analysis to model the payment behaviour of companies that are already overdue on their invoice payments. More precisely, the thesis focuses on modelling late invoice payment times in business-to-business sales process using survival analysis and random forests techniques, particularly, quite recent random survival forests method. Survival analysis allows us to model the ranking of payment times and evaluate which characteristics of a company have an impact on the payment time. Data for our analysis is provided by Register OÜ and Kreedix OÜ.

So far, the main methods when predicting the bankruptcy of a company were using financial information of the companies. But there are many restrictions to using financial information from the annual reports. Most importantly, annual reports have to be submitted 6 months after the year end, thus the information we get from the reports alone is

not up-to-date enough to predict short-term payment behaviour of a company. Secondly, accessibility to entities' financial information is restricted and reliability of the information provided is questionable. Therefore it is necessary to find some alternative data and methods that could be used in modelling the payment behaviour of a debtor. Our goal is to use additional variables that are updated more frequently in order to predict payment behaviour of a debtor on a timely basis and also assess if the financial data is needed in a model.

The thesis is structured into 4 chapters. Chapter 1 provides the knowledge of important terminology, introduces context of the problem and reviews of literature of previous related research. Chapter 2 gives a theoretical overview of the methods used in the analysis part of the thesis. Data collection, preprocessing and descriptive analysis is described in Chapter 3. Modelling, predictive analysis and comparisons of the models is presented in Chapter 4. Conclusions and proposals for future work summarizes the thesis. Data processing and analysis is performed with statistical computing software R. The R Code used to perform the analysis is provided separately as R files on the webpage https://drive.google.com/open?id=0B703_4DFiD2dQXVENmhFeXJFVVE. The link is accessible until 01.09.2016, after this date it is possible to inquire access to the R Codes from the author of this thesis.

1 Background and Related Work

Understanding the problem of this thesis in addition to statistical knowledge requires some basic understanding of business processes and accounting. Therefore, in the next sections a small overview of terminology and introduction to the sales process is given. In the last section of this chapter previous literature and research is reviewed.

1.1 Definitions

Creditor - A creditor is a person, bank, or other enterprise that has lent money or extended credit to another party [3]. In business-to-business sales process, creditor is the company that issues a sales invoice to debtor.

Debtor - A debtor is a person or enterprise that owes money to another party [3]. In business-to-business sales process, debitor is the company that receives a sales invoice from creditor. Debitor becomes a debtor once it has not paid for an invoice before due date.

Credit risk is the risk of loss of principal or loss of a financial reward originating from a borrower's failure to repay a loan or otherwise meet a contractual obligation [3].

Invoice-to-Cash Process is the process from the moment the invoice is created until the moment the customer's debt (payment) is settled/reconciled [4].

Due Date of an Invoice - payment term provided by Creditor to the Debitor; the date when the invoice should be paid for.

Sales Ledger is a detailed itemization of sales made and not yet paid for. The report includes both - invoices that are due and invoices that are not yet due [5].

Accounts Receivable (AR) - refers to money owed by customers (individuals or corporations) to another entity in exchange for goods or services that have been delivered or used, but not yet paid for. Receivables usually come in the form of operating lines of credit and are usually due within a relatively short time period, ranging from a few days to a year [3].

1.2 Introduction to the late payments problem and evaluating creditworthiness of debtors

The credit risk evaluation problem is to make a classification of good or bad for a certain customer using the attribute characteristics of the customer ([6], p. 8). So the problem of late payments and credit risk are closely related. When analysing late payments we also need to determine which debtors are likely to cover their debts early and which debtors will have very long overdue days or, even worse, never pay their debts.

The credit risk modelling can be roughly categorized into two: consumer credit risk and corporate credit risk. This thesis focuses on the corporate credit risk through business-to-business (B2B) sales process (more precisely, invoice-to-cash process). See the typical workflow of invoice-to-cash process in the Figure 1.1 [7]. Collection management highlighted in the figure is what we are interested in.

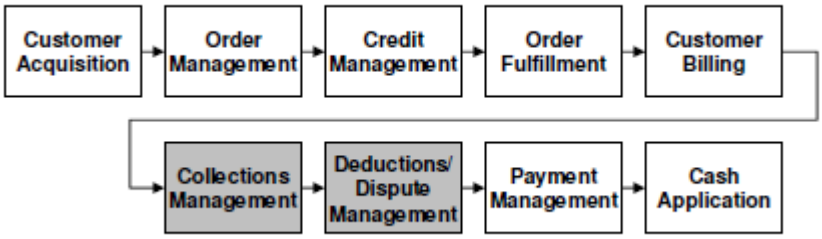

Figure 1.1: Typical invoice-to-cash process [7].

In this thesis we focus specifically on the payment behaviour of debtors. But due to the limited amount of research and literature of modelling late invoice payments, in Section 1.4 we also review relevant literature and research of predicting business failure, loan defaults, credit risk modelling of loans for both, corporate and consumer loans.

The debt collection plays an important role in businesses cash flow management. Receiving late payments from invoices can cause problems in company's liquidity and cash flow predictions. Being able to predict cash-flows accurately is essential for all businesses to achieve financial stability. In fact, according to the European Payment Report 2015, a survey carried out by Intrum Justitia [8], 40% of companies claimed that their customers inability to pay on time is hindering growth and 31% see late payments as a threat to their long-term survival. The respondents of the survey see 3.1% of their yearly revenues being written off. This certainly implies that proper management of trade credit risk is crucial to eliminate the losses.

Invoice-to-cash process is therefore an important part of the overall sales processes. At the same time it is also slow, expensive and inaccurate as the collection activity steps are processed manually [7]. There are also many factors to be considered in the collection activities – typically all customers are contacted at fixed intervals even if they have always paid on time. Also note, that it is generally true that the later a customer is contacted the less likely the invoices will get paid on time. But then again, repeated contacting of "good" customers may lead to lower customer satisfaction.

The seriousness of debt collection problems can also be illustrated by the fact that trade credit insurance started evolving already in 1926 when the first conference on trade credit insurance was held in London [9]. The need for a credit insurance for companies indicates that there are problems in the payment behaviour of the debtors. In Estonia, there is one state-owned insurance company that deals with trade credit insurance - KredEx Krediidikindlustus AS [10].

Payment behaviour in Estonia. It follows from the Intrum Justitia survey [8], where multiple choice answers were possible, that the main reasons for late payments in Estonia are debtors that are in financial difficulties (74%), intentional late payments (66%) and administrative inefficiencies of the customers (53%). Less respondents (18%) considered disputes regarding goods and services delivered as a reason for late payments. According to the survey, 78% of Estonian companies have been asked longer payment terms than they feel comfortable with. The average business-to-business (B2B) payment term allowed to customers in Estonia is 15 days, whereas the average time to actual payment is 20 days.

1.3 Thesis motivation and objectives

The main purpose of this thesis is to model the late payment days of invoices that are already overdue in business-to-business sales process. The desired output is a measure that could assess the late payment time of an invoice. The information provided by the model is of interest for a credit management firm to direct further actions in debt collection of their clients and provide the client with some suggestions about the creditworthiness of those debtors.

The questions of interest of this thesis are:

- Identify characteristics that have an impact in the payment behaviour of companies that are already overdue on their payments.
- Without making many adjustments to the data, do the automated procedures of modelling result in a trustworthy model?

- Does the ranking of payment times improve when using historical data of debtor payment behaviour?

More specifically, the objectives of this thesis can be divided into two:

- for first-time debtors (new debtors in credit management firm database) create a model that uses external information (publicly available information) as predictors;
- for repeated debtors create a model that additionally uses information about historical payment behaviour of this debtor as predictors.

It is specifically of interest if using the historical data of previous invoice payments (which we can define from the same data of sales ledgers) results in a more accurate model. This would allow us to use up-to-date information about the debtor's current financial situation. A detailed description of data provided for analysis and preprocessing of raw data is given in Chapter 3.

1.4 Review of literature and previous research

Since the late 1960s, numerous studies were devoted to predict business failure using publicly available data and combining it with statistical classification techniques. Pioneering work and one of the first attempts to perform modern statistical failure analysis was done by Tamari [11].

There was a steady growth in the number of articles published related to credit risk since year 2000. A few possible reasons can be associated with the increased interest in credit risk modelling - rapid development of some new data mining techniques, the availability of more open credit datasets, the growth of credit products and credit markets ([6], p. 7).

When analysing the literature and research of problems that are similar to the question of interest in this thesis, we can review material that in it's essence is close to late invoices payment problem - credit risk of companies, predicting failed firms (bankruptcies), modelling defaulted loans (corporate and consumer) and prediction models of invoice payments in collection activities. We also review both - classification and regression problems. The reason for the wide range of topics is that there is very limited literature of prediction models of the invoice-to-cash process. But late payments of invoices is also a problem of making difference between better debtors (late payments with less overdue days, e.g. up to 30 days), bad and worst debtors (e.g. when no payment is received).

Credit Risk of Companies and Failures of Firms

The techniques for credit risk modelling can be roughly categorized into the following groups ([6], p. 8):

- Statistical models: linear discriminant analysis, logistic regression, probit regression, k-nearest neighbour, classification tree, etc.
- Mathematical programming methods: linear programming, quadratic programming, integer programming, etc.
- Artificial intelligence techniques: artificial neural networks, support vector machines, genetic algorithm and genetic programming, rough set, etc.
- Hybrid approaches: artificial neural network and fuzzy system, rough set and artificial neural network, fuzzy system and support vector machines etc.
- Ensemble or combined methods: neural network ensemble, support vector machine ensemble, hybrid ensemble etc.

Often it has been found out that hybrid and ensemble approaches usually achieve better classification performance than individual models ([6], p. 23). As later described in Section 2.2, we also implement an ensemble method in modelling late payment days of debtors and we use a statistical model for comparison.

Defaulted loans

The problem of defaulted loans is similar to credit risk (or credit-scoring) evaluations as credit-scoring systems aid the decision of whether to grant credit to an applicant or not. Traditionally this is done by estimating the probability that an applicant will default [12].

In the industry of credit scoring the standard method with a long history is using logistic regression making a decision between good and bad applicants. As a more recent modelling technique in credit scoring and modelling defaults, it has been shown in [13] and [14] that survival analysis can be applied in credit scoring.

Using survival analysis in predicting loan early repayment and predicting loan defaults, has been discussed in [12]. It was concluded that survival analysis models are competitive with the standard approach of logistic regression in credit scoring industry when used for classifying loan applicants into two groups.

Invoice debt collection

There is rather small amount of literature dealing with the problem of invoice debt collection and predictive analysis of payment times. An article [7] and a Master's thesis [15] both covered the problem of classifying whether an invoice will be late (overdue) or not, and if an invoice is overdue then how many days it will be overdue until payment. The problem was solved as a multiple classification problem using machine learning techniques.

In [15] invoice data and historical data were used as the predictors in the analysis and predictive classification of overdue days of late payments. It was concluded that adding historical data as predictors results in a more accurate model. From all the methods applied, random forests performed the best when predicting overdue days classification. Predictive analysis was performed based on one firm's invoices to debtors.

In [7], invoices created by 4 firms were under observation. Predictive analysis was performed for both, first-time invoices and returning invoices. For all 4 firms it was concluded that using historical data as predictors improves the accuracy of invoice payment time classification. Models were built separately on each firm but also a unified model that could be used for all the firms was implemented. The unified model performed better than the individual models (classification accuracies for unified model were 77%–96% and for individual models 66%–93%) which suggests that the unified model can find common patterns in these four firms that are overlooked by individual models.

Conclusions

The problem of this thesis is not a classification problem in nature (although it can be solved as a classification problem as previously described). When analysing invoice debt payments we are more interested in when the event occurs rather than whether it occurs or not. This means we are interested in analysing time to an event and therefore survival analysis approach is reasonable. The problem of invoice debt payments is similar to loan early repayments which, as previously described, can be solved using survival analysis. Interest in using survival analysis for loan defaults modelling and credit risk estimation has shown some increase, several articles ([12], [16], [17], [18]) deal with the use of survival analysis.

As our initiative is to view the problem of invoice late payments as a regression problem rather than a classification problem, survival analysis fits our purpose. In ([19], p. 55) it was concluded that in modelling credit scoring the differences between survival and logistic models for a fixed time period are nearly indistinguishable. We present some more specific reasons in Chapter 2 why survival analysis is a better fit than other methods.

2 Overview of Methods

The question of interest of this thesis is to analyze payment times of overdue invoices and we have chosen survival analysis for this purpose.

A possible solution when modelling payment time of an invoice could also be linear regression as a function of a set of predictor variables. However, linear regression is not the best choice in terms of our data as we do not have the precise payment time (minimum and maximum payment times will be derived from data). Also, invoices that have been long time overdue, are more likely to never be paid for and removing them from analysis would result in too optimistic prognosis. In addition, time to payment is a positive number and ordinary linear regression may not be the best choice unless event times are first transformed in a way that removes this restriction [20]. Most importantly, as described above, ordinary linear regression cannot effectively handle censoring of observations (invoices that have been overdue for some time without any information about the actual payment time).

Therefore, in terms of our data, survival analysis is an appropriate method to analyze and model time to payment of an overdue invoice as it takes censoring into account (in the Figure 3.5 it can be seen that when removing data without known payment time, we would underestimate the actual payment times).

2.1 Survival analysis

Survival analysis is used in statistics to model and analyze the expected duration of time to a certain event. Generally such events are defined as *'failures'*. Survival analysis is mostly used for medical data and clinical trials where the event of failure is usually defined as death or recurrence of a disease. In this section we follow [21], pp. 1-17, if not stated otherwise.

In survival analysis, subjects are usually followed over a specified time period and the focus is on time at which the event of interest occurs. We can see that it is possible that a *'failure'* time will not be observed in time period of observation due to deliberate design of the experiment or random censoring.

In context of the problem of this thesis, time to event is the time from due date of an invoice to the payment date of an invoice. So, objects under observation are invoices that are overdue and time $t = 0$ is due date of the invoice. Therefore, instead of having

a *'failure'* as an event, we have a positive event of *'payment'* under observation. The negative event of failure (time point when the invoice would never be paid) would be undefinable. From now on our events of *'death'* and *'payment'* are equivalent terms. In the context of survival analysis *'alive'* refers to unpaid invoice (invoice is *'alive'* when it has not been paid for yet).

Let T denote a non-negative random variable representing the time to the event of interest (in our case time to payment). The probability that the invoice is not paid before time t is given by the survival function

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx, \tag{2.1}$$

where $f(\cdot)$ is the density and $F(\cdot)$ the distribution function of T.

The survival function is considered to meet the following conditions:

- $S(t)$ is a monotone decreasing function,
- $S(0) = 1$,
- $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

The last condition might not be fully applicable for the defaulted loans problem and invoice-to-cash process but it can be assumed that in an infinite time period the debt of invoice is either paid by the debtor, the claims are collected with the use of encashment firm or in case of debtor's bankruptcy, claims are covered with the assets of debtor firm. So, in this thesis we assume that the third condition is satisfied.

Note that these three conditions are theoretical properties of survival curve and in practice, when using actual data, we usually obtain graphs that are step functions, rather than smooth curves. The theoretical survival curve $S(t)$ and the estimated survival function $\hat{S}(t)$ are depicted in the Figure 2.1.

Density function of T can be defined through survival function (2.1),

$$f(t) = \lim_{\Delta t\to 0+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \tag{2.2}$$

Hazard function $h(t)$ specifies the rate of success at time $T = t$ given that the invoice has not been paid for up to time t. Using (2.1) and (2.2), the hazard function can be defined as

$$h(t) = \lim_{\Delta t\to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t\to 0+} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t) \cdot \Delta t} = \frac{f(t)}{S(t)}.$$

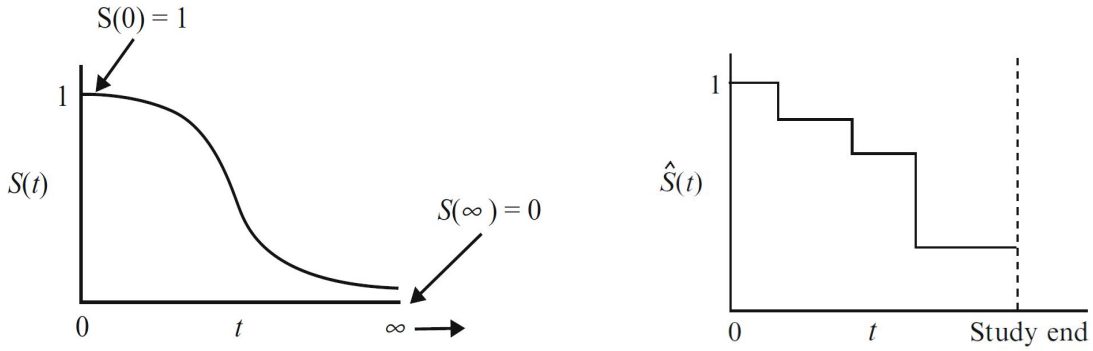

Figure 2.1: The theoretical survival curve S(t) on the left and the estimated survival function $\hat{S}(t)$ on the right ([22], p. 10).

Therefore, with respect to our problem, $h(t)\Delta t$ can be interpreted as the approximate probability of payment of the invoice in time range $[t, t + \Delta t]$, given that there was no payment up to time t. We can view $h(t)$ as a measure of intensity at time t or a measure of the potential success at time t. It is important to notice that the hazard is a rate, rather than a probability (it can assume values in $[0, \infty)$).

Since

$$h(t) = -\frac{dS(t)/dt}{S(t)} = -\frac{d\log(S(t))}{dt}, \tag{2.3}$$

it is verified that $h(t)$, similarly to $S(t)$, specifies the distribution of T.

Cumulative hazard function $H(t)$ is gained by integrating $h(u)$ over $(0, t)$ – which due to (2.3) results in the expression

$$H(t) = \int_0^t h(u)du = -\log(S(t)). \tag{2.4}$$

And therefore also

$$S(t) = \exp[-H(t)]. \tag{2.5}$$

So, each function $S(t)$, $H(t)$ and $h(t)$ can be derived from any of the other functions. The distribution of T is described by any of these three functions ([23], p. 405). Once the distribution of T is known, many useful characteristics can be found.

Median life length (0.5-quantile) is the time (measured in days) for which the probability that invoice is not paid (and equally the probability that invoice is paid), $S(t)$, is 0.5. In practice it means that the median life length (0.5-quantile) is the time by which half invoices will not be paid for (and equally half invoices will be paid for). The median life length is obtained by setting $S(t) = 0.5$ and finding respective t, denoted by $T_{0.5}$:

$$T_{0.5} = S^{-1}(0.5).$$

More generally, we can find the life length $T_\alpha = S(^{-1}(\alpha)$ for which the probability is α that invoice is not paid.

Survival probabilities are other quantities of interest. The company might be interested in the probability that the invoice is still not paid before $T = a$, i.e. interested in $S(a)$, or the probability that invoice is paid between $a < T < b$, i.e. in $S(a) - S(b)$.

2.1.1 Censoring

Censoring is a form of missing data problem which is common in survival analysis. There are 3 censoring types: right-, interval- and left-censoring ([23], p. 401). The most common type of censoring, and also relevant in the context of this thesis, is right-censoring. In clinical trials, where survival analysis is most commonly used, patients typically enter a study at different times ([21], p. 12). In the context of this thesis the same applies - new invoices are generated throughout the year, so new invoice observations enter the study at different times.

We want to observe payment time of an invoice but censoring can occur in the following ways ([22], p. 6) when adjusted to our data:

1. Loss to follow-up. At some time due to some reason the Creditor has not any more provided sales ledgers (information about its claims). All invoices that occur in the last sales ledger of the Creditor are censored. We know that the payment time of the invoice is greater that the overdue days of the invoice at the balance date of last sales ledger provided.

2. Partial loss to follow-up. At some periods of time Creditor has not provided Sales Ledgers periodically and payment time of an invoice cannot be calculated reliably. We know that payment time of invoice is greater than the overdue days of invoice at the balance date of the previous sales ledger provided.

3. Termination of study. The last date of the study is 31.12.2015. All the invoices that occur in the sales ledger at 31.12.2015 (meaning that the invoices are not paid) are censored as we do not have further information of payments.

Random variable T denotes a random failure (payment) time from the survival distribution $S(t)$. We need additional notation for the response of the j-th invoice that contains both censoring and payment events. Note, that the invoice may be censored if the invoice is not followed long enough (the study ends at 31.12.2015) or due to loss to follow-up or partial loss to follow-up. In fact, even the paid invoice may have a censoring time (the end of study) and the censored invoice may have a payment time (some time after censoring).

The response of the j-th invoice is defined to be either payment time T_j or censoring time C_j. Let us define the event (payment) indicator δ_j of the j-th invoice as

$$\delta_j = \begin{cases} 1 & \text{if the event was observed} \quad (T_j \leq C_j), \\ 0 & \text{if the event was censored} \quad (T_j > C_j). \end{cases} \tag{2.6}$$

The observed response is

$$Y_j = \min\{T_j, C_j\}, \tag{2.7}$$

which is the time that occurred first: the payment time or the censoring time. Hence we observe K iid random pairs (Y_j, δ_j) that contain all the response information needed ([23], p. 406).

2.1.2 Kaplan-Meier method

Kaplan-Meier method is the best known and simplest non-parametric method for projection of survival curve. It can be used to estimate the survival function, also called the product limit estimator. This estimator incorporates all the information available (uncensored and censored) by considering survival to any point in time as series of steps defined by the observed survival and censored times ([24], p. 28).

The steps are intervals defined by a rank ordering the survival times: each interval begins at an observed time and ends just before the next ordered time. So, say we have observations of payment times of invoices and for the i-th invoice, time to payment is t_i. Let the number n_i represent the survivors (number "at risk", in our case number of unpaid invoices) up to time t_i, d_i the number of payments at the time t_i, and c_i the number of censored observations at time t_i.

The Kaplan-Meier estimator of $S(t)$ is the product of the form

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

where $n_i = n_{i-1} - d_{i-1} - c_{i-1}$ and has a convention that $\hat{S}(t) = 1$ if $t < t_1$ ([24], p. 34).

For example, if we have the following 8 time observations (3 censored observations marked with '+' and 5 payments):

time	7	7	11+	15	15+	30	92	92+

We can calculate the Kaplan-Meier survival estimates as given in the Table 2.1.

Table 2.1: Example data of invoice payments and the calculation of survival estimate.

i	t_i	n_i	d_i	c_i	$\hat{S}(t_i)$
0	0	8	0	0	1
1	7	8	2	0	$1 \cdot \frac{8-2}{8} = 0.75$
2	11	6	0	1	$0.75 \cdot \frac{6-0}{6} = 0.75$
3	15	5	1	1	$0.75 \cdot \frac{5-1}{5} = 0.6$
4	30	3	1	0	$0.6 \cdot \frac{3-1}{3} = 0.4$
5	92	2	1	1	$0.4 \cdot \frac{2-1}{2} = 0.2$

The Kaplan-Meier curve is a right continuous step function which steps down only at an uncensored observation ([21], p. 30). The plot of Kaplan-Meier curve for the data in Table 2.1 is shown in the Figure 2.2.

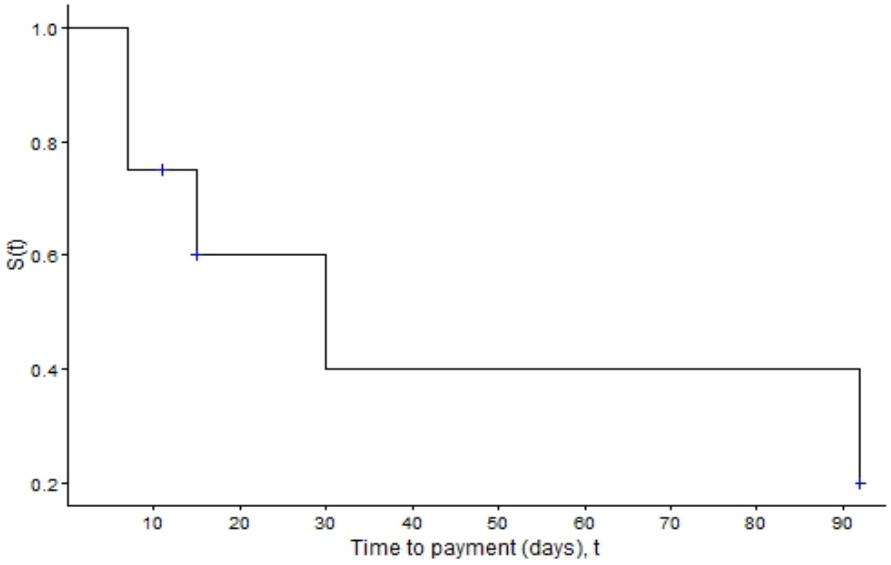

Figure 2.2: The Kaplan-Meier curve for the example data in Table 2.1. The blue crosses represent the censored observations.

Note that the Kaplan-Meier curve in the Figure 2.2 does not go to zero as the largest survival time 92+ is censored. We cannot estimate $S(t)$ beyond $t = 92$ days. Some refer to $\hat{S}(t)$ as a defective survival function. Alternatively, $\hat{F}(t) = 1 - \hat{S}(t)$ is called a sub-distribution function as the total probability is less than 1 ([21], p. 30).

2.1.3 Cox Proportional Hazards model

Cox Proportional Hazards (PH) model is a semi-parametric method of survival analysis – it is assumed that the baseline hazard function is the same for all the objects under the observation. The following theoretical overview of Cox PH model is provided on the basis of [23], pp. 427–428.

Suppose that on each observation there are p predictor variables $X = (X^1, ..., X^p)$ available. The survival model is generalized from a hazard function $h(t)$ for the failure time T to a hazard function given the predictors X:

$$h(t|X) = h_0(t) \exp(X\beta), \tag{2.8}$$

where $X\beta = \beta_0 + \beta_1 X^1 + \cdots + \beta_p X^p$.

The regression formulation (2.8) is called the proportional hazards model. The $h_0(t)$ part of the $h(t|X)$ is called an underlying (or baseline) hazard function or a hazard function for a standard subject, which is a subject with $X\beta = 0$.

Depending on whether the underlying hazard function $h_0(t)$ has a constant scale parameter, $X\beta$ may or may not include an intercept β_0. The term $\exp(X\beta)$ can be called a relative hazard function and in many cases it is the function of primary interest as it describes the (relative) effects of the predictors.

Based on (2.4) and (2.5), the PH model can also be written in terms of the cumulative hazard and survival functions:

$$\begin{aligned} H(t|X) &= H_0(t) \exp(X\beta), \\ S(t|X) &= \exp(-H_0(t) \exp(X\beta)) = \exp(-H_0(t))^{\exp(X\beta)}, \end{aligned}$$

where $H_0(t)$ is an underlying cumulative hazard function. Note, that $S(t|X)$, the probability of surviving (not paying) past time t given the values of predictors X, can also be written as

$$S(t|X) = S_0(t)^{\exp(X\beta)},$$

where $S_0(t)$ is the underlying survival distribution, $S_0(t) = \exp(-H_0(t))$.

Proportionality assumption

The way in which the predictors affect the distribution of the response should be by multiplying the hazard or cumulative hazard by $\exp(X\beta)$ or equivalently by adding $X\beta$ to the log hazard or log cumulative hazard at each t. The effect of the predictors is assumed

to be the same at all values of t since $\log(h_0(t))$ can be separated from $X\beta$. In other words, the PH assumption implies no t-by-predictor interaction ([23], p. 428).

Suppose we have two specifications of the predictors X, defined as X_* and X. From the Cox PH model (2.8) we can obtain general formula for estimating a hazard ratio that compares $X_* = (X_*^1, ..., X_*^p)$ and $X = (X^1, ..., X^p)$:

$$\text{HR} = \frac{h(t|X_*)}{h(t, X)} = \frac{h_0(t)\exp(\beta X_*)}{h_0(t)\exp(\beta X)} = \exp(\beta(X_* - X)) = \theta,$$

where θ is a constant, i.e. not depending on t.

The Cox PH model assumes that the hazard ratio comparing any two specifications of predictors is constant over time t. Equivalently, this means that the hazard for one invoice is proportional to the hazard for any other invoice, where the proportionality constant is independent of time ([22], p. 165).

Partial likelihood function and estimation of β

The likelihood construction is based on [23], pp. 476–477. Let $t_1 < t_2 < ... < t_N$ represent the unique ordered failure (payment) times. Assume for now that there are no tied payment times (tied censoring times are allowed). Consider a set of invoices at risk of payment an instant before event (payment) time t_i. Let this risk set of invoices at time t_i be denoted by n_i. So n_i is the set of invoices j such that the subjects had not failed (not paid) or been censored by time t_i. Therefore n_i includes invoices with payment or censoring time $Y_j \geq t_i$, where Y_j is defined as in (2.7).

The conditional probability that invoice k is the one that was paid at t_i, given that the invoices in the set n_i are at risk of being paid, and given further that exactly one payment occurs at t_i, is by the rules of conditional probability

$$\text{P("invoice } k \text{ is paid at } t_i\text{"}|n_i, \text{"one paid invoice at } t_i\text{"}) = \frac{\text{P("invoice } k \text{ is paid at } t_i\text{"}|n_i)}{\text{P("one paid invoice at } t_i\text{"}|n_i)}.$$

This conditional probability equals

$$\frac{h_0(t_i)\exp(X_k\beta)}{\sum_{j\in n_i} h_0(t_i)\exp(X_j\beta)} = \frac{\exp(X_k\beta)}{\sum_{j\in n_i}\exp(X_j\beta)} = \frac{\exp(X_k\beta)}{\sum_{Y_j\geq t_i}\exp(X_j\beta)}. \tag{2.9}$$

Note, that (2.9) is independent of $h_0(t)$. In order to understand this likelihood, let us consider a special case of $\beta = 0$, meaning that the predictors have no effect. Then $\exp(X_k\beta) = \exp(X_j\beta) = 1$ and P("invoice k is paid at t_i"$|n_i$, "one payment at t_i") equals

$1/n_i$. As before, n_i is the number of invoices at risk (of payment) at time t_i.

A total likelihood can be computed by multiplying individual likelihoods over all failure (payment) times as these conditional probabilities are themselves independent across the different payment times. A partial likelihood for β is:

$$\mathrm{L}(\beta) = \prod_{(Y_k, \delta_k = 1)} \frac{\exp(X_k \beta)}{\sum_{Y_j \geq Y_k} \exp(X_j \beta)}. \tag{2.10}$$

The log partial likelihood can be directly derived from (2.10):

$$\log \mathrm{L}(\beta) = \sum_{(Y_k, \delta_k = 1)} \left[(X_k \beta) - \log \left(\sum_{Y_j \geq Y_k} \exp(X_j \beta) \right) \right]. \tag{2.11}$$

The maximum partial likelihood estimator can be derived by maximizing (2.10) or (2.11) with respect to β ([24], p. 96).

Stepwise Cox model and the AIC procedure

It would be ideal to perform variable selection by trying out a lot of different models that contain a different subset of the predictors in order to decide on the best model. Based on the number of predictors p, there are 2^p models containing subsets of predictors. When p is large, automated approaches are more efficient and preferred. There are three classical approaches for the predictor selection ([25], pp. 78-79):

- Forward selection. We start with a *null model* (a model with no predictors, only an intercept). The predictor that improves the model the best, is added. The procedure is performed as long as there is no statistically significant predictor to be added.
- Backward selection. We start with a *full model* (a model that contains all predictors). The least statistically significant predictor is removed. This selection is applied multiple times until there is no predictor that is statistically insignificant.
- Mixed selection. A combination of forward and backward selection. We start with the null model and forward selection is performed. At each step it is considered if any of the previously added predictors have become insignificant and if needed, a backward selection is performed. Forward and backward steps are performed until all predictors in the model have a sufficiently low p-value (and all predictors that are not in the model would have a large p-value if added to the model).

As the mixed selection involves both – forward and backward selection – we implement mixed selection when fitting the Cox model.

To evaluate whether adding or removing a predictor improves the accuracy of the model we use the Akaike's information criterion (AIC). The statistic used to make comparisons between possible models is

$$\text{AIC} = -2 \cdot \log(L) + 2 \cdot m,$$

where m is the number of β coefficients under consideration and L is the likelihood of the model (the likelihood is replaced by the partial likelihood), refer to (2.10) and (2.11). The decision rule for choosing the best model is that the smaller the AIC value, the better the model is ([21], pp. 123–124).

2.2 Decision Tree Methods

Decision trees can be applied to both regression and classification problems ([6], p. 303). A tree-based method *Random Forests* has recently also been extended for the analysis of right censored survival data [27].

In Random Forests, a number of decision trees is built on bootstrapped training samples. When building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. So, the split is allowed to use only those m predictors, meaning that the algorithm is not allowed to even consider a majority of the available predictors. At each split a fresh sample of m predictors is taken. Typically $m \approx \sqrt{p}$ is chosen ([25], p. 319).

As brought out in Section 1.4, ensemble methods often outperform individual models. In the following we will give an overview of an ensemble tree method Random Survival Forests.

2.2.1 Random Survival Forests

In survival analysis many different regression modelling strategies can be applied to predict the risk of future events. Often, however, the default choice of analysis relies on Cox regression modelling due to its convenience. Extensions of the random forest approach to survival analysis provide an alternative way to build a risk prediction model [28].

The main advantage of Random Survival Forests is that it is highly data adaptable and virtually model assumption free. In survival analysis we often need to rely on some restrictive assumptions (e.g. the assumption of proportional hazards). There is always the concern whether associations between predictors and hazards have been modelled appropriately, and whether or not non-linear effects or higher order interactions for predictors should be

included. In contrast, such problems are handled seamlessly and automatically within a Random Forests approach [29].

The algorithm for Random Survival Forests is as follows:

1. Draw B bootstrap samples from the original data.
2. Grow a tree for each bootstrap sample. At each node of the tree randomly select m predictors (covariates) for splitting on. Split on a predictor using a survival splitting criterion. A node is split on that predictor which maximizes survival differences across daughter nodes.
3. Grow the tree to full size under the constraint that a terminal node should have no less than a specified number of unique deaths (payments).
4. Calculate an ensemble cumulative hazard estimate by combining information from the n trees. One estimate for each individual (invoice) in the data is calculated.
5. Compute an out-of-bag (OOB) error rate for the ensemble derived using the first b trees, where $b = 1, \ldots, B$.

We provide an illustrative example of Random Survival Forest tree in the Figure 2.3. The top internal node corresponds to splitting factor variable representing submission of last annual report and has two levels: SUBMITTED and UNSUBMITTED. Left-hand node consists of invoices for which the factor variable value is SUBMITTED. The right-hand consists of the remaining observations.

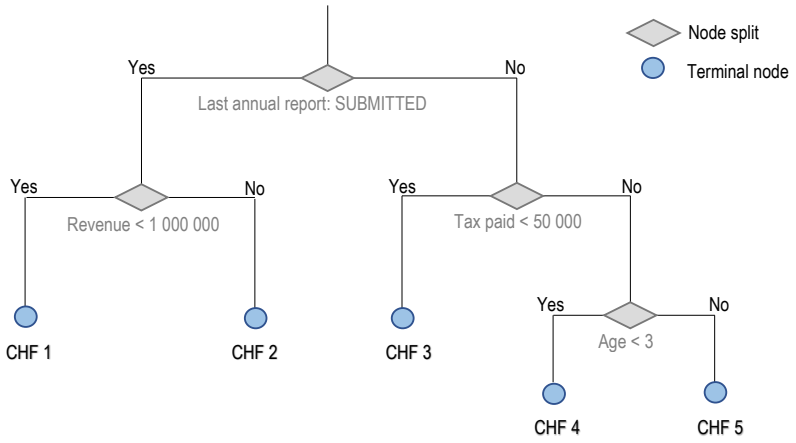

Figure 2.3: Example of a tree in Random Survival Forests.

An example of a continuous variable splitting in the Figure 2.3 can be seen for Revenue. The tree is grown such that when all invoices are dropped down the tree at least the specified number of unique payment events are in each terminal node. A cumulative hazard function (CHF) is constructed for each terminal node ([30], p. 21). A comprehensive description of splitting process based on [29] follows.

Splitting rules

The node splits are essential and most important parts of the algorithm. The available options for splitting rules are a log-rank splitting rule (the default splitting rule), a conservation of events splitting rule, a log-rank score rule, and a fast approximation to the log-rank splitting rule. As the log-rank splitting will be used in the purpose of our analysis, other methods will not be covered in this section.

Assume that during the growing process of a tree we seek to split node h into two daughter nodes. Let's assume that there are K_h observations $\{(Y_1, \delta_1), ..., (Y_{K_h}, \delta_{K_h})\}$ (as defined in (2.7)) within h. As before, assume we have predictors $X = (X^1, ..., X^p)$. A proposed split at node h on a given predictor X^u ($u \in \{1, ..., p\}$) is always of the form $X^u \leq c$ and $X^u > c$. The split forms two daughter nodes and two new sets of survival data. A good split maximizes survival differences across the two sets of data.

Let $t_1 < t_2 < \cdots < t_N$ be the distinct payment times in the parent node h, and let $d_{i,j}$, and $n_{i,j}$ equal the number of payments and individuals at risk at time t_i in the daughter nodes $j = 1, 2$. Note that $n_{i,j}$ is the number of invoices in daughter j that are alive (not paid) at time t_i, or who have an event (payment) at time t_i. More precisely,

$$n_{i,1} = \#\{T_l \geq t_i, X_l^u \leq c\}, \qquad n_{i,2} = \#\{T_l \geq t_i, X_l^u > c\},$$

where X_l^u is the value of X^u for observation $l = 1, \ldots, K$.

Finally, define $n_i = n_{i,1} + n_{i,2}$ and $d_i = d_{i,1} + d_{i,2}$. Let n_j be the total number of observations in daughter j. Thus, $n_j = n_1 + n_2$. Note that $n_1 = \#\{l : X_l^u \leq c\}$ and $n_2 = \#\{l : X_l^u > c\}$.

Log-rank splitting

The log-rank statistic for a split at the value c for predictor X^u is

$$L(X^u, c) = \frac{\sum_{i=1}^{N} \left(d_{i,1} - n_{i,1} \frac{d_i}{n_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{n_{i,1}}{n_i} \left(1 - \frac{n_{i,1}}{n_i} \right) \left(\frac{n_i - d_i}{n_i - 1} \right) d_i}}.$$

The value $|L(X^u, c)|$ is the measure of node separation: the larger the value for $|L(X^u, c)|$, the greater the difference between the two groups, and the better the split is. The best split at node h is determined by finding the predictor X^v and c^* such that $|L(X^u, c)| \leq |L(X^v, c^*)|$ for all X^u and c.

Ensemble estimation

Random Survival Forests produce an ensemble estimate for the cumulative hazard function which is also the basis for calculating the model performance (see C-index in section 2.3.1).

The cumulative hazard function is estimated for each tree grown from a bootstrap data set by grouping hazard estimates by terminal nodes. Consider a node h, let $t_{i,h}$ be the distinct payment times in h and let $d_{i,h}$ and $n_{i,h}$ equal the number of payments and invoices at risk at time $t_{i,h}$. The cumulative hazard estimate for node h and fixed t is defined as

$$\hat{H}_h(t) = \sum_{t_{i,h} \leq t} \frac{d_{i,h}}{n_{i,h}}.$$

A sequence of such estimates, $\hat{H}_h(t)$, is provided by each tree. Suppose there are M terminal nodes in the tree, then there are M such estimates. Consider individual (invoice) j, and denote its predictor vector as $X_j = (X_j^1, ..., X_j^p)$. In order to compute $\hat{H}(t|X_j)$ for an individual j, one has to drop X_j down the tree and find the terminal node for respective j. If the terminal node is h then the estimate is

$$\hat{H}(t|X_j) = \hat{H}_h(t), \quad X_j \in h. \tag{2.12}$$

The cumulative hazard function (2.12) is computed for all invoices j in the data and for $t \in \{t_1, ..., t_N\}$.

Also note that the estimate (2.12) is based on one tree. In order to produce an ensemble, (2.12) is averaged over all B trees. Let $\hat{H}_b(t|X)$ denote the cumulative hazard estimate (2.12) for tree $b = 1, \ldots, B$. The out-of-bag (OOB) ensemble cumulative hazard estimate for j is

$$\hat{H}_e^*(t|X_j) = \frac{\sum_{b=1}^{B} I_{i,b} \hat{H}_b(t|X_j)}{\sum_{b=1}^{B} I_{j,b}}, \tag{2.13}$$

where $I_{j,b} = 1$ if j is an OOB point for b, otherwise $I_{j,b} = 0$.

Note, that estimator (2.13) is obtained by averaging over bootstrap samples in which j is excluded, meaning datasets in which j is an OOB value. The ensemble cumulative hazard estimator uses all bootstrap samples:

$$\hat{H}_e(t|X_j) = \frac{1}{B}\sum_{b=1}^{B}\hat{H}_b(t|X_j). \tag{2.14}$$

2.2.2 RandomForestSRC package

To be able to reliably compare Cox Model and Random Survival Forests we divide data into two: training set and test set. The results of models are validated on the same test set.

In the RandomForestSRC package [31] it is advised to experiment with different node sizes, as it also determines the number of splits and a bad choice of node size may result in underfitting or overfitting.

We give an overview of important parameters used in RandomForestSRC:

- **formula** – a symbolic description of a model to be fit;
- **ntree** – number of trees grown in a forest;
- **nodesize** – minimum terminal node size (default is 3 deaths (payments));
- **splitrule** – splitting rule used to grow a tree (default is logrank).

To plot the results and evaluate which variables were used and are important in the tree growing process, we use Variable Importance (VIMP) from ggRandomForests package [34]. VIMP measure was originally defined in CART package using a measure involving surrogate variables [32] and it also used for Random Survival Forests. A description of VIMP follows in Section 2.3.2.

2.3 Performance measures

In this section we introduce concordance index as a performance for survival models and explain variable importance in Random Survival Forests models.

2.3.1 C-index

To estimate prediction error for both, Cox model and Random Survival Forests, we use concordance index (C-index) which is one of the most commonly used performance measures of survival models ([23], pp. 256–258). It estimates the probability that, in a randomly selected pair of cases, the case that is paid first had a worst predicted outcome. In

its essence it calculates the fraction of times that for a pair of invoices the invoice that was predicted to be paid later was actually paid sooner, i.e., the prediction ranked the payment times incorrectly. Unlike other measures of survival performance, the C-index does not depend on a single fixed time for evaluation. The C-index also specifically accounts for censoring.

Calculating C-index is based on [27]. It requires a predicted outcome which we define using OOB ensemble cumulative hazard function (CHF) similar to (2.13) to define a predicted outcome. This value is derived from OOB data and thus it can be used to obtain an OOB estimate for C.
Let $t_1^0, ..., t_m^0$ denote pre-chosen unique time points (we use the unique event (payment) times $t_1, ..., t_N$). To rank two invoices k and j, we say k has a worse predicted outcome than j if

$$\sum_{i=1}^{m} \hat{H}_e^*(t_i^0|X_k) > \sum_{i=1}^{m} \hat{H}_e^*(t_i^0|X_j). \tag{2.15}$$

C-index calculation:

1. Form all possible pairs $\{(Y_k, \delta_k), (Y_j, \delta_j)\}$ of cases over the data.

2. Omit pairs (Y_k, δ_k) and (Y_j, δ_j) where shorter survival time is censored. Omit pairs (Y_k, δ_k) and (Y_j, δ_j) if $Y_k = Y_j$ unless at least one is a death (payment). Let Permissible denote the total number of permissible (remained) pairs.

3. For each permissible pair where $Y_k \neq Y_j$, count 1 if the shorter survival time has worse predicted outcome; count 0.5 if predicted outcomes are tied. For each permissible pair, where $Y_k = Y_j$ and both are deaths (payments), count 1 if predicted outcomes are tied; otherwise (one is censored), count 0.5. For each permissible pair where $Y_k = Y_j$ but not both are deaths (payments), count 1 if the death (payment) has worse predicted outcome; otherwise (one is censored), count 0.5. Let Concordance denote the counts over all permissible pairs.

4. The C-index, C, is defined by

$$C = \frac{\text{Concordance}}{\text{Permissible}}.$$

Error is defined Error $= 1 - C$. The error value of 0.5 indicates that the model has no predictability (it is equivalent to tossing a coin). The error value of 0 shows that the model has a perfect predictability. If Error > 0.5, it indicates that the predictors of the model predict the opposite direction [29].

Concordance index for Cox model is calculated using function *survConcordance* in the survival package. This function calculates the C-index based on survival times ([33], pp. 94–95).

2.3.2 Variable importance

Random Forests typically result in improved accuracy over the prediction that uses one tree (e.g. classical decision tree). Unfortunately it is difficult to interpret the resulting model since there is a large number of trees. Thus it is not clear which variables are most important in the procedure ([25], p. 319).

Variable importance (VIMP) is a measure that shows how worse would the prediction be if that variable were not available. To calculate VIMP for a variable X^u, drop OOB cases down their in-bag survival tree. Whenever a split for X^u is encountered, assign a daughter node randomly (both nodes have equal probability). The cumulative hazard function from each such tree is calculated and averaged. The VIMP for X^u is the prediction error for the original ensemble subtracted from the prediction error for the new ensemble obtained using randomizing X^u assignments [32].

As the prediction error in Random Survival Forests is characterized by C-index, VIMP can be interpreted in terms of misclassification. As described in Section 2.3.1, C-index estimates the probability of correctly classifying (ranking) two cases. Thus, VIMP for X^u measures the increase (or decrease) in misclassification error on the OOB (or test) data if X^u were not available.

A large VIMP value indicates that excluding the variable reduces the predictive accuracy in the forest. VIMP close to zero indicates the variable contributes nothing to predictive accuracy, and negative values indicate the predictive accuracy improves when the variable is excluded [34].

Note that it is incorrect to think VIMP estimates change in prediction error for a forest grown with and without a variable. For example, if two variables are highly correlated and both predictive, each can have large VIMP values. Removing one variable and regrowing the forest may affect the VIMP for the other variable (its value might get larger), but prediction error will likely remain unchanged [32].

3 Data Description and Processing

Data about invoices (more precisely, sales ledgers of creditors) for this thesis is provided by Kreedix OÜ that is a credit management service provider in Estonia. Data about all Estonian companies is provided by Register OÜ, a firm that intends to provide procedures to small and medium companies to optimize their invoice-to-cash processes. Refer to the Figure 3.1 to get an understanding of the data and subjects of our analysis.

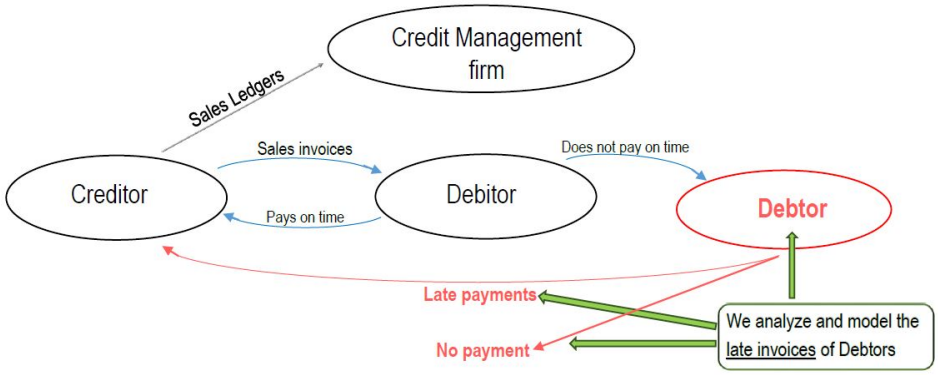

Figure 3.1: Flowchart of the process of debt collection and credit management which is source of sales ledgers data for our analysis.

The database available for our analysis is all sales ledgers reports provided by the clients (in Figure 3.1 referred to as *Creditor*) to the credit management bureau Kreedix OÜ. Sales ledgers (SL) are from time period January 2015 to December 2015 (balance dates (printout dates) of SL reports). For simplicity purposes we do not distinguish between the terms Debitor and Debtor: from now on we only use term Debtor to name the firm an invoice is addressed to.

For the invoice payment analysis we have the following sources of data:

- DATA 1: Sales ledgers data – information about invoices that were overdue and invoices that were not overdue;
- DATA 2: Database of all Estonian companies: registry code, company name, number of employees, date of registry, age, registry status (active, bankrupt, deleted, liquidation);
- DATA 3: Data from annual reports 2012–2014 (financial data is available for all companies that have yearly sales in an amount that is more than 50 000 euros);
- DATA 4: Tax debt data from Estonian Tax and Customs Board (e.g. sum of tax

paid, paid tax change quarter over quarter, number of unsubmitted tax declarations).

Note that none of the databases reflect the date of payment of an invoice. Payment time can be derived from the balance date times of sales ledgers. For example, if an invoice 'InvID1' with due date 30.01.2015 is in the sales ledger at 02.02.2015, it reflects that the invoice is overdue 3 days as of 02.02.2015. If the same invoice 'InvID1' is in the sales ledger at 09.02.2015, it means the invoice has not been paid for and it is overdue 10 days. Now, when the invoice 'InvID1' is not in the sales ledger at 16.02.2015, it means the invoice has been paid in the time interval of 10.02.2015 (included) to 16.02.2015 (included). Time $t = 0$ is the due date (30.01.2015) of invoice 'InvID1', the minimum payment date of 'InvID1' is 10.02.2015 ($t = 11$) and the maximum payment date of 'InvID1' is 16.02.2015 ($t = 17$).

3.1 Terminology and variable definition

The variables available to us from the sales ledgers (DATA 1) database are the following:

- **Creditor** – name of the firm issuing a sales invoice;
- **Creditor_reg_code** – registry code of the firm issuing an sales invoice;
- **Debtor** – name of the firm (the buyer who has to pay) the invoice is addressed to;
- **Invoice** – the number / ID of invoice;
- **Invoice_Date** – the date of issuing the invoice (this might be NA). This date is not actually not important information for us);
- **Due_Date** – the due date of the invoice;
- **Sum** – the sum that has not been paid for the invoice;
- **Balance_Date** – the date of the sales ledger report.

We have generated the following variables to the database to be later able to preprocess DATA 1 for payment time analysis:

- **Debtor_reg_code** – the registry code of the *Debtor* is added to the data from Register OÜ database (where the debtor's registry code was missing originally, we added registry code from DATA 2, using Debtor (name of debtor). Where debtor name had mistakes in DATA 1, names were corrected and registry code was added);
- **InvID** – Unique invoice ID (defined as *Creditor_Invoice*) to avoid having same invoice numbers from different creditors;
- **minBalDate** – for the *InvID* the first date the invoice appeared in a sales ledger;

- **maxBalDate** – for the *InvID* the last date the invoice appeared in a sales ledger;
- **OverdueDays** = *Balance_Date* – *Due_Date*.
  - if $x = OverdueDays > 0 \rightarrow$ the invoice is overdue x days at the *Balance_Date*
  - if $x = OverdueDays \leq 0 \rightarrow$ the invoice is not overdue at the *Balance_Date*
- **minPaymentDate** – the minimum possible payment date of the invoice, *see also "Definition of limits for payment date" for details*;
- **maxPaymentDate** – the maximum possible payment date of the invoice, *see also "Definition of limits for payment date" for details*.

Definition of limits for payment date. The invoice is in the sales ledger as long as it hasn't been paid for, therefore:

- if the *maxBalDate* of the invoice is the last date of the sales ledgers from that creditor $\rightarrow$ we do not know the payment date $\rightarrow$ we define
  - **minPaymentDate** $\leftarrow NA$;
  - **maxPaymentDate** $\leftarrow NA$;
- otherwise
  - **minPaymentDate** $\leftarrow$ *maxBalDate* $+ 1$;
  - **maxPaymentDate** $\leftarrow$ next *Balance_Date* of the creditor following to the *maxBal_Date* of that creditor.

3.2 Data preprocessing

From the sales ledgers database all data rows that had $Sum < 0$ were removed (as all data with $Sum < 0$ indicates credit invoices or prepayments that don't need to be taken into account when analysing debt data). The database consisted of 195 108 rows before aggregation and data filtering.

In the initial database one row of data is an entry in one sales ledger, the database needed to be aggregated to a invoice level so that we would have one row per one invoice. As our observations in the analysis will be invoices (more precisely time to payment or censoring time of an invoice), we aggregated the initial database to a invoice level to have the data in the survival analysis context. After aggregation we have 84 635 invoices (both overdue and not overdue invoices).

The data of sales ledgers is a database that has many different sources (the accounting systems of Kreedix clients differ) and therefore the data is subject to some quality problems that need to be taken into account and identified. The original sales ledger reports provided by the clients have been manually processed by the employees of Register OÜ. In the final database of sales ledgers provided by Register OÜ we have identified two types of sources for potential errors:

1. Problems in the initial sales ledgers provided by Kreedix OÜ clients.
    - The sales ledgers are reports from different accounting systems of Kreedix OÜ clients (mostly in Excel) and may cause problems in the preprocessing phase;
    - Manual mistakes in debtor name caused by typing errors in accounting system, not up-to-date debtor registry code;
    - During the analysis we have identified data rows that do not seem logical (e.g. due date of the invoice is 365 days in the future) - this might be caused by long-term liabilities (such as loans).
2. Manual processing of the sales ledgers into one database. As the reports of sales ledgers were in many different forms, it is easy to make mistakes like taking a wrong column or calculating the dates incorrectly.

We have taken into account these types of mistakes and implemented procedures to eliminate these shortcomings.

3.2.1 Data cleaning - faulty data corrections, outlier detection and exclusion from database

There are some specific characteristics or restrictions we can define for an invoice of interest, we predefined the conditions that our invoice in the database should meet:

- Invoice_Date $\leq$ Due Date - *The due date of the invoice should not be earlier than the date of the invoice;*
- $|\text{Due\_Date} - \text{Balance\_Date}| \leq 90$;
- $|\text{Due\_Date} - \text{Invoice\_Date}| \leq 90$;

The first condition was set to identify manual processing mistakes. Invoices, for which the condition $Invoice\_Date \leq Due\_Date$ was false, were first examined separately and compared to original raw data. Where a manual preprocessing mistake was found, the database was adjusted to be in accordance with the initial raw data (sales ledger provided

by the Creditor).

Other conditions needed to be set because the database of sales ledger included some long-term claims (such as loans, etc) that are not related to invoice-to-cash process.

When examining the data of invoices, a limit of 90 days was set as the maximum possible days of payment term. In 2013 there was a change in the Law of Obligations Act (*Võlaõigusseadus*) and 60 days was declared to be a maximum payment term in business-to-business sales (*Võlaõigusseadus,* §82[1], section (2)). It is still possible to have special agreements and contracts to have longer payment terms and therefore as in our data it could be seen that 90 days is also a plausible payment term, a limit of 90 days was set. The limit is needed to exclude possible loan borrowings in the sales ledgers.

The initial database did not have Invoice Date for all the invoices – 12 912 invoices had a missing invoice date. As the invoice date is not relevant to our modelling process, we keep the data where the invoice date is missing. The invoice date is only needed to clean the data of some of the faulty entries.

All the data entries that did not meet our conditions might have been either a faulty entry (a manual mistake in the data processing phase or a long-term debt, such as a loan). After filtering the data with conditions set for an invoice, 1.2% of data was discarded (83 620 rows of invoices remained).

3.2.2 Data corrections

To eliminate the faulty data due to manual preprocessing and the mistakes in the original sales ledgers provided by the clients, some detective analysis was performed:

- The original database had debtor registry codes in some cases. To assure that the registry codes were correct, the debtor company name was compared to the Register OÜ database (DATA 2) using registry code. Where there were inconsistencies found between the company name in the DATA 2 and the name in DATA 1, *Debtor_reg_code* was changed for the registry code that is in compliance with the company name in the sales ledger. It is due to the fact that in the accounting system the registry code might be outdated/mistyped/missing but the name of the company is correct. All the missing registry codes were added manually (registry code of the debtor is the ID for merging the data of predictors).
  Invoices that did not have *Debtor_reg_code*, registry code was added from DATA 2 using *Debtor* as an ID. Where NA values were generated, debtor names were corrected (manual typing mistakes by accountants or informal names of companies that

did not match the name in DATA 2) and *Debtor_reg_code* was manually added.

- The data of foreign debtors was removed from the database (registry code containing letters or the registry code not of length of 8 numbers).
- *Balance Date* analysis was performed to identify manual processing mistakes – inconsistent *Balance Dates* were corrected as in the original sales ledgers.

3.2.3 Data selection process

As the purpose of this thesis is to analyse data in the survival analysis context, we needed to define the time period of the study (time of observing). In our case the study time is year 2015 – from the database the overdue invoices that had a due date in 2015 were selected.

Note, that a separate database of all invoices (both overdue and not yet overdue - DATA 1) was kept to later define historical payment behaviour of the debtors.

We also eliminated data of the creditors that had provided us with only 1 sales ledger. Data of one sales ledger would not provide us with any payment information and would generate censored data.

After all the preprocessing and data filtering we have 28 510 overdue invoices that had a due date in 2015.

3.2.4 Defining payment time and censoring time

The difference between maximum and minimum payment time for our final data is characterized by the histogram on Figure 3.2. Note that 3 465 invoices had NA min and max payment dates and therefore these invoices do not reflect in the histogram.

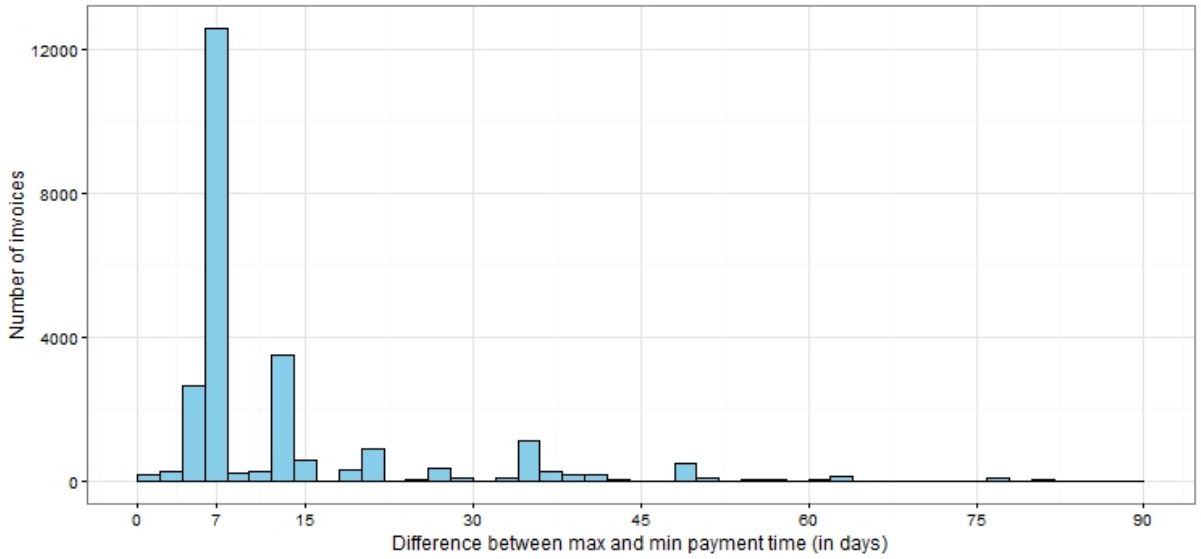

Figure 3.2: Histogram of the difference between maximum and minimum payment time.

Since most of the observations had up to 8 days difference between maximum and minimum payment time (see the histogram), it was decided to define the fixed payment time for all of those invoices that had the difference of payment times less than 8 days. The payment time was defined as a mean of the difference. In this way the defined payment time differs from the actual (unknown) payment time ±4 days (it is the maximum difference). All the other invoices for which the difference between maximum and minimum payment time is more than 8 days or for which the payment times are NA, we treat as censored observations. The censoring time is the last date the invoice is known to be unpaid.

3.3 Example data in context of survival analysis

In this section we present an example data for our analysis. Table 3.1 demonstrates how the payment date of an invoice is derived and how the time is calculated in the context of survival analysis. Status indicates whether an observation is censored or not (*Status* 1 indicates that the *Time* is the payment time of the invoice and *Status* 0 indicates that the *Time* is the time of censoring).

Table 3.1: Example database of sales ledger.

Creditor	Debtor	Invoice	Due Date	Balance Date	Payment Date	Time	Status
A	E	1	13/01/2015	28/01/2015	31/01/2015	18	1
A	B	2	13/03/2015	27/05/2015	NA	75	0
B	F	3	01/02/2015	05/04/2015	NA	63	0
C	G	4	25/04/2015	20/05/2015	23/05/2015	28	1

The example data of the Table 3.1 as time-to-event data is demonstrated in the Figure 3.3. Note, that for survival analysis the time $t = 0$ is the *Due_Date* of *Invoice*.

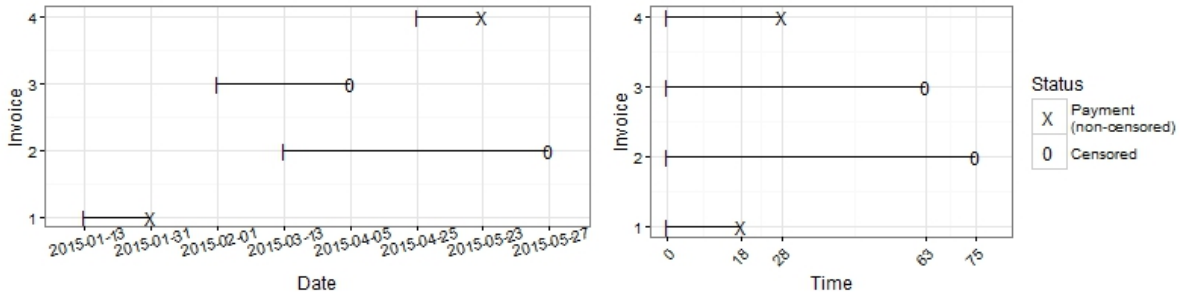

Figure 3.3: Example data in its original format (left) and in survival analysis context (right).

3.4 Possible predictors, preprocessing of predictor values and missing data treatment

A list of all the possible predictors is presented in the Appendix A. In this section we present information about more complex predictors that need some further explanation. Note that all the variables are added respectively to the due date of an invoice.

3.4.1 Initial variables available

1. Invoice data

Sum. The base sum of the invoice. The sum of the invoice is added as a predictor because some companies (especially smaller companies) that have liquidity problems, might be able to pay smaller invoices first and larger invoices are more likely to be more overdue.

2. Tax data

Tax debt. Tax debt data is updated at the 1st and 15th date of each month. In the database we have as predictors tax debt at the current month (*tax_debt0*), 1 month before (*tax_debt1m*) and 2 months before (*tax_debt2m*). So, if the invoice has a due date at the 8th day of the month, *tax_debt0* is dated from the 1st day of the month. If the due date of the invoice is at 18th day of the month, the tax debt 0 is dated from the 15th day of the month.

Tax paid. Tax paid is quarterly paid taxes (total sum of turnover and social tax). The variable is added to the invoice data by due date - taxes paid by the debtor in previous quarter before the due date of an invoice.

For 89 invoices the tax paid was a negative sum in the database. This could be caused by the fact that the original data from Estonian Tax and Customs Board (ETCB) is cumulative tax paid. In case of tax corrections (which is common in accounting) the previous period data might not be updated in the ETCB database and when calculating the non-cumulative tax paid sums for quarters, there might be some negative sums. All the negative sums of tax paid were treated as 0 in our analysis.

Tax qoq. The tax change quarter over quarter (this is ratio of the change in tax paid). Infinite values were encountered as in the calculation of the ratio some *Inf* and *-Inf* values had been generated. For the purposes of our analysis, the *Inf* values were replaced with the maximum value of other *Tax qoq* and the *-Inf* values were replaced with the minimum value of other *Tax qoq*.

Unsubmitted declarations. Number of unsubmitted tax declarations (cumulative number). For the purposes of analysis this variable will be transformed into binary variable: 1 – the debtor has had unsubmitted tax declarations in the past and 0 - the debtor hasn't had any unsubmitted tax declarations.

3. Company specific data

Business type. The business type of the company. It could be a measure of company size different business types have different requirements on minimum equity.

Business type	Osaühing	Aktsiaselts	Other
Number of invoices	22 549	5 227	734

In the table 'Other' consists of business types such as FIE, Mittetulundusühing, Euroopa Äriühing, Riigi- ja kohaliku omavalitsuse asutus, Sihtasutus, Täisühing, Tulundusühistu, Usaldusühing, Välismaaäriühingu filiaal. As discussed later, we will not use business type as a predictor because we believe financial data from annual report reflects better the size of a company.

EMTAK letter. EMTAK letter represents the field of activity of the debtor. For example, it is usually considered that there are more problems with late invoices in construction. A more specific table of EMTAK letters in our database is presented in the Appendix B. We also present a possibility to classify the EMTAK letter into three groups (refer to Appendix B). To have balanced groups, we finally combine the EMTAK letters into two larger groups as presented in the Table 3.2.

Table 3.2: EMTAK letters combined into two groups of Manufacturing and Service.

Group	EMTAK letters	No. of debtors	No. of invoices	Percentage of invoices
EMTAK1 (Manufacturing)	A, B, C, D, E, F	2 650	14 951	52%
EMTAK2 (Service)	G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U	2 981	13 559	48%

Age. Debtor's age in years starting from the registry time of the company to the due date of invoice. In the database there were some invoices with a negative age of -1 which could be caused by the rounding of difference in time. All the negative sums for age were replaced with 0.

Claims current. Debtor's other liabilities (in euros) at the due date of a specific invoice.

4. Annual reports and financial data

The financial information from year 2014 is gathered from the annual reports of the companies (DATA 3). We added the relevant financial data to the invoice data for each debtor.

Total sales. Total sales or revenue is the amount of money that a company actually receives for goods or services sold. Investors will often consider a company's revenue and net income separately to determine the health of a business. It is possible for net income to grow while revenue remains stagnant, as a result of cost-cutting [3]. Therefore, as income is more volatile and more of a subject to be biased by company's management and strategical decisions in a year, we use sales as a baseline in our model. Net income will not be added as a predictor.

Equity. Shareholders' equity is equal to a firm's total assets less total liabilities and is one of the most common financial metrics employed by analysts to determine the financial health of a company. Shareholders' equity represents the net value of a company, or the amount that would be returned to shareholders if all the company's assets were liquidated and all its debts repaid [3]. Equity also comprises the retained earnings of a company and therefore it represents the information of income from the beginning of the company. A low value of equity shows that a company's liabilities are large in comparison to its assets (meaning that the financial health of the company is not so good), large value of equity shows that the liabilities of company are well covered by its assets. Note that in Estonia it is not allowed to have a negative equity.

Total Assets. Total assets is a balance sheet item representing what a firm owns. Assets are bought to increase the value of a firm or benefit firm's operations. Asset can be thought of as something that can generate cash flow [3]. We have added total assets as a predictor indicating the size of a company.

Current Assets. Current assets are assets that the company could convert into cash within a year in the normal course of business. Current assets include cash, accounts receivable, inventory, marketable securities, prepaid expenses and anything else that can easily be turned into cash [3].

Current Liabilities. Current liabilities are a company's debts or obligations that are due within one year, including short term debt, accounts payable, accrued liabilities and other debts. Normally, companies withdraw or cash current assets in order to pay their current liabilities [3].

Some missing data of financial information occurred. There are two different reasons for the missing data of annual reports.

Sources of missing data:

- The debtor has not submitted annual report 2014;
- Register OÜ has annual report data of 2014 for companies that had total sales greater than 50 000 euros.

Solutions for treating missing data (step-by-step):

1. Missing data of 2014 was replaced with annual report data from 2013 (where possible);
2. It was checked if remained missing observations could be replaced with data from 2012 but none of the debtors with missing data occurred in the 2012 annual report database;
3. The remaining missing data (4 554 invoices, 1 402 debtors) was treated as follows:
    - Total sales - sales information is missing for companies that have total sales less than 50 000 euros, the missing data of total sales was replaced with a mean (25 000 euros).
    - Equity - the minimum requirements of equity are established by the Äriseadustik. There are different requirements for AS and OÜ, therefore the missing data was replaced with the minimum requirement as follows (in Estonia the equity is not allowed to be negative):

  - AS: The minimum requirement of equity is 25 000 euros ([35], §222);
  - OÜ: The minimum requirement of equity is 2 500 euros ([35], §136);
  - Other: Business type other consists multiple types of businesses and there is no minimum equity requirement established. In this case missing data (for 572 invoices) was replaced with 0.
- Total Assets - in the accounting the basic rule of the balance sheet is that the Total Assets = Equity + Total liabilities. Therefore, as we have no baseline to evaluate total liabilities, we assume the liabilities to be 0 euros. Missing data of Total Assets is replaced with the value of equity.
- Current Assets - we have no baseline for assessing current assets, therefore missing data is replaced with 0.
- Current Liabilities - we have no baseline for assessing current liabilities, therefore missing data is replaced with 0.

3.4.2 Additionally added predictors of ratios

In model fitting different methods will be tested. For example, for financial data from annual reports we will try to fit total sales, current assets and equity straight into the model. But, also as it is suggested in predictive modelling literature ([36], p. 27), using combinations of predictors can sometimes be more efficient than using the individual values (using ratios of predictors may be more effective than using two independent predictors). Therefore, some ratios of financial data that we intend to use in modelling are introduced.

We guess that the use of ratios may be more useful for the Cox PH model than for Random Survival Forests since the latter is a decision tree and in essence takes the interactions of predictors into account with its multiple splits. But as ratios combine multiple variables into one, they may be a beneficial form for random survival forests as well. This will be tested in the experiments performed in Chapter 4.

1. Invoice data

Sum ratio. The invoice base sum depends on the debtor - we assume that companies with a greater turnover also have larger purchasing amounts. For this reason we include a *Sum ratio* that is the invoice base sum divided by the debtor's weekly sales (total sales divided by 52). In the calculation, 8 infinite values were created. Those infinite values were replaced with the maximum of finite ratio values of the database.

2. Annual report data

Asset turnover. Asset turnover ratio shows company's sales value relative to its assets. Turnover ratio can often be used as an indicator of the efficiency with which a company is deploying its assets in generating revenue. Asset turnover also indicates the sector of company - for example, retail is a sector that most often yields the highest asset turnover ratios. Conversely, firms in sectors like utilities and telecommunications, which have large asset bases, will have lower asset turnover [3].

Current ratio. Current ratio is a liquidity ratio and it measures company's ability to cover short-term obligations. Current ratio is defined as current assets divided by current liabilities. The higher the current ratio is, the more capable the company is of paying its obligations. Having a current ratio below 1 shows that the company is not in good financial health and suggests that it is questionable if the company is able to pay off its obligations. Although it is a measure on the balance sheet date of the annual report, it still indicates company's financial situation [3].

Infinite ratios that are generated when calculating the ratios, are replaced with the maximum value of finite ratios. Values of NA (indicating division of 0/0) are replaced with a 0.

3.4.3 Historical payment behaviour data as variables

Defining the predictors representing the historical payment behaviour of debtor is presented in the Appendix A. Overall we defined multiple measures that could indicate historical payment. To be used in the analysis as predictors, we have selected three measures of historical payment behaviour (all measures take into account only the invoices of the debtor for the specific invoice under consideration):

Average days late. Average number of overdue days for paid late invoices in the time period of 30 days before due date of the invoice to the due date of invoice.

Ratio of OS late. Defined as 'Sum of late invoices base sums outstanding (OS)' relative to 'Sum of all invoices base sums outstanding (OS)' at the date of 30 days before due date of specific invoice.

Ratio of paid late. Defined as 'Sum of late invoice base sums that were paid in the time period of 30 days before due date of the invoice to the due date of invoice' relative to 'Sum of all invoice base sums that were paid in the time period of 30 days before due date of the invoice to the due date of invoice'.

Note, that we have defined those predictors for 30 days before due date on an invoice. It would be of interest to define additional measures for longer time periods as well to see if the accuracy of model is improved. In the context of this thesis it is out of scope.

We used 30 days of historical data as previous payment behaviour definition because it should be representative of debtor's current financial situation.

3.5 Initial data analysis

In this section we first illustrate our data – the payment behaviour in general and in groups.

3.5.1 Descriptive analysis

The overdue days of invoices are shown in the Figure 3.4. The density functions are plotted separately for invoices with defined payment times (time to payment) and for censored observations (time to censoring - last date when the invoice is known to be unpaid) separately. The plot is cut at 120 days as there where few observations beyond 120 days (see Table 3.3).

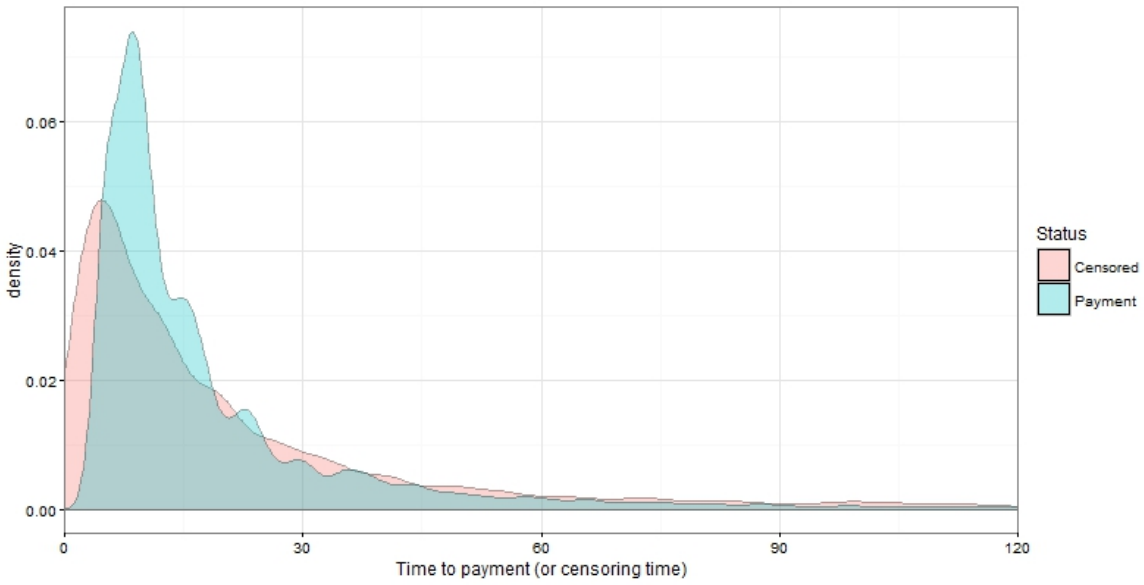

Figure 3.4: The density functions of payment and censoring times for overdue invoices

It can be seen from the Figure 3.4 that the censored and uncensored invoices have approximately the same density. As the proportion of censored and uncensored invoices is difficult to see in the graph, we will later illustrate the comparison in a table.

The Kaplan-Meier survival curves of the invoice data are in the Figure 3.5. As could be seen from the Figure 3.4, there is little invoice data after 90 days, therefore we have plotted the Kaplan-Meier survival curves up to 90 days.

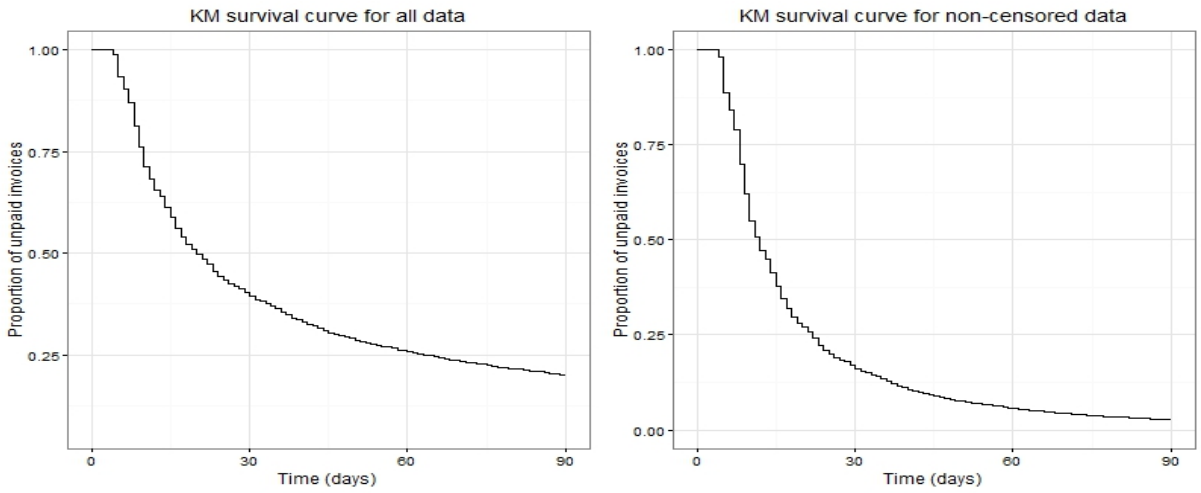

Figure 3.5: Kaplan-Meier survival curves of all overdue invoice data (left) and for non-censored data (right).

We can again see from the Figure 3.5 that leaving out the censored data (the data we have no or insufficient information about payment time) would result in underestimating the time to payment. It can be seen from the graph that when taking censoring into account, the median payment time of all invoices is more than 15 days. If we would discard the censored invoices, the median payment time would be less than 15 days. And the proportion of unpaid invoices at 90 days is significantly higher when taking censored observations into account. That is the fundamental reason why censored times need to be added in our analysis.

We have classified the overdue days of invoices into traditional groups of overdue days of invoices (also referred to as overdue classes of invoices). The grouping of invoices by overdue days is usual when doing impairments on claims and in the process of cash flow predictions. See Table 3.3 to see number of different creditors and debtors per overdue classes of invoices.

Table 3.3: Table of overdue payment and censoring times of invoices that had a due date in 2015.

Time in days	**1–7**	**8–15**	**16–30**	**31–60**	**61–90**	**91–120**	**121–...**
# of paid invoices	3 303	6 393	3 394	1 647	470	187	222
# of censored invoices	4 288	2 957	2 498	1 621	547	341	642
# of all invoices	**7 591**	**9 350**	**5 892**	**3 268**	**1 017**	**528**	**864**
# of Creditors	53	64	73	74	65	59	54
# of Debtors	2 494	3 034	2 344	1 358	529	280	351

It follows from the Table 3.3 that most of the observed times for invoices are censored after 60 days (in total 1 530 invoices, which is 64% of all invoices). In the first overdue class of 1–7 days, a large proportion (56%) of censored invoices is caused by loss to follow-up and termination of the study (refer to Section 2.1.1).

3.5.2 Kaplan-Meier curves for debtor payment behaviour analysis

In the following we represent some analysis of debtors with the use of Kaplan-Meier curves. We group the debtors to our best knowledge just to see if there are some different payment behaviour indicators for some predictors and to see if our hypothesis about the payment behaviour of debtors is in correspondence with the data.

To see if the payment behaviour of the debtors differs in age groups, the age (in years) of the debtor company was classified into three groups: 0–2, 3–7 and more than 8 years. Note, that as roughly half of the data is censored, we have not depicted censoring times in the graphs for readability reasons. The dashed lines represent confidence intervals.

As it follows from Figure 3.6, the Kaplan-Meier survival curves do indicate a difference but it can also be seen that the curves cross after 240 days.

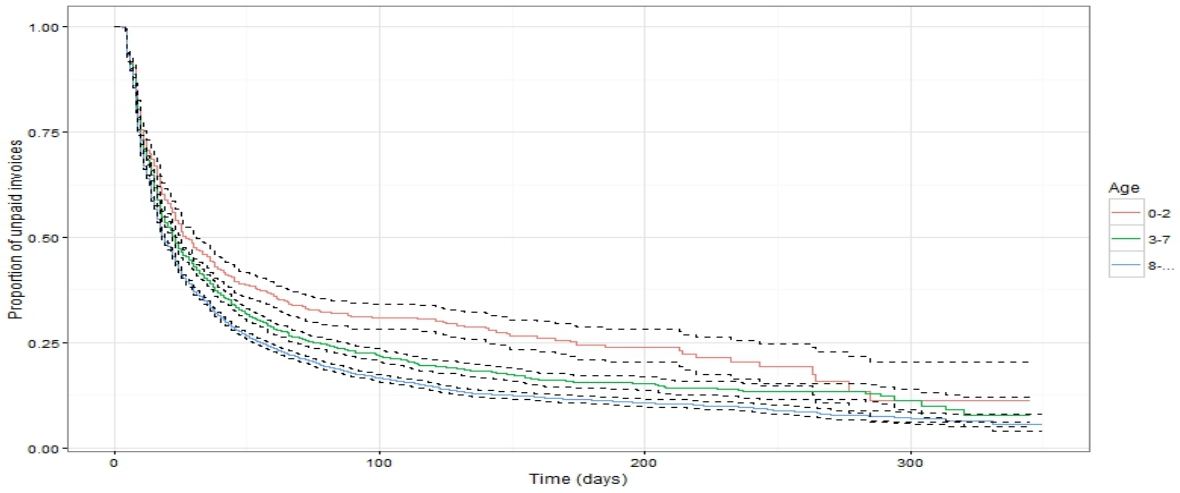

Figure 3.6: Debtors' payment behaviour by age groups (with confidence intervals).

Submission of last annual report (year 2014) indicates a huge difference in payment behaviour. The Figure 3.7 shows that the invoices of debtors that have not submitted the annual report of 2014, take longer time to pay for the invoices (half of the debtors have paid for the invoice 30 days after the due date). The invoices of the debtors that have submitted last annual report, are paid in less time (half of the debtors have paid for the invoice 15 days after the due date). We can also see that in day 100 the proportion of unpaid invoices is twice bigger for the debtors that have not submitted the annual report.

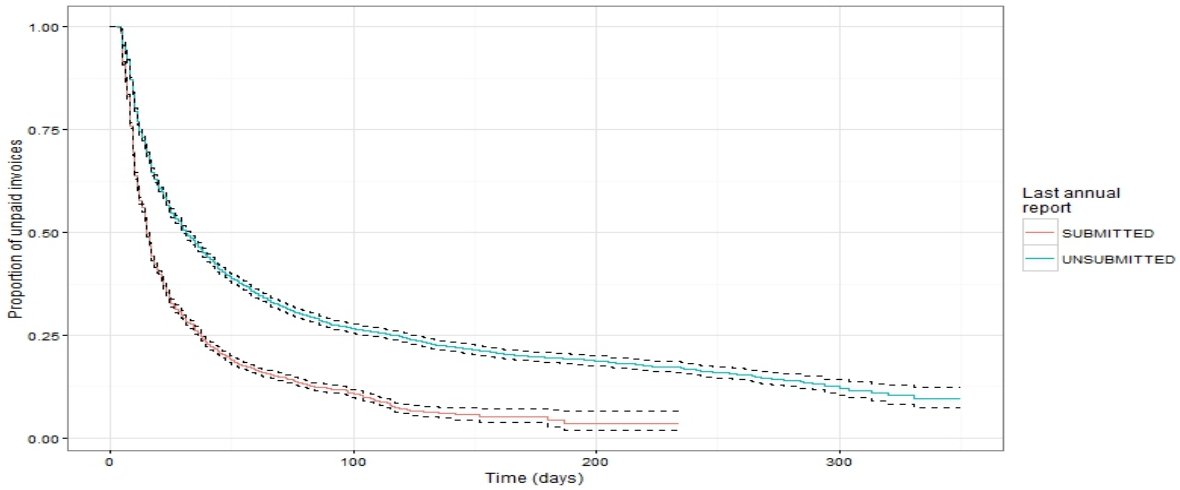

Figure 3.7: Debtors' payment behaviour by submitting annual report of 2014.

We also hypothesized that the business type could indicate a difference in payment behaviour as a measure of company size (due to the minimum requirements set for the equity). The Figure 3.8 does not show that there might be such a relationship between the business type and payment behaviour. We can see that the Kaplan-Meier curves overlap indicating that there isn't a significant difference between the groups. We also have financial

information of debtors and hence the business type of a company becomes unnecessary.

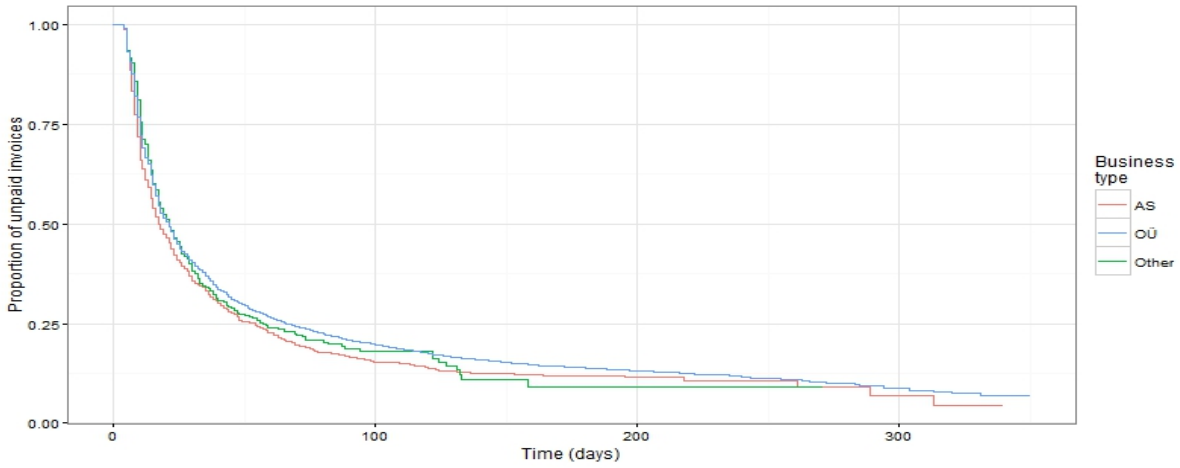

Figure 3.8: Debtors' payment behaviour by business type.

EMTAK letter represents the field of activity of the company. The Figure 3.9 shows that the grouping of EMTAK letters into two groups as described in Appendix B does indicate some difference in the payment behaviour. In can be seen from the graph that debtors classified as service (by EMTAK letter) tend to pay for the late invoices sooner (median around 10 days) than the debtors classified as manufacturing (median around 25 days).

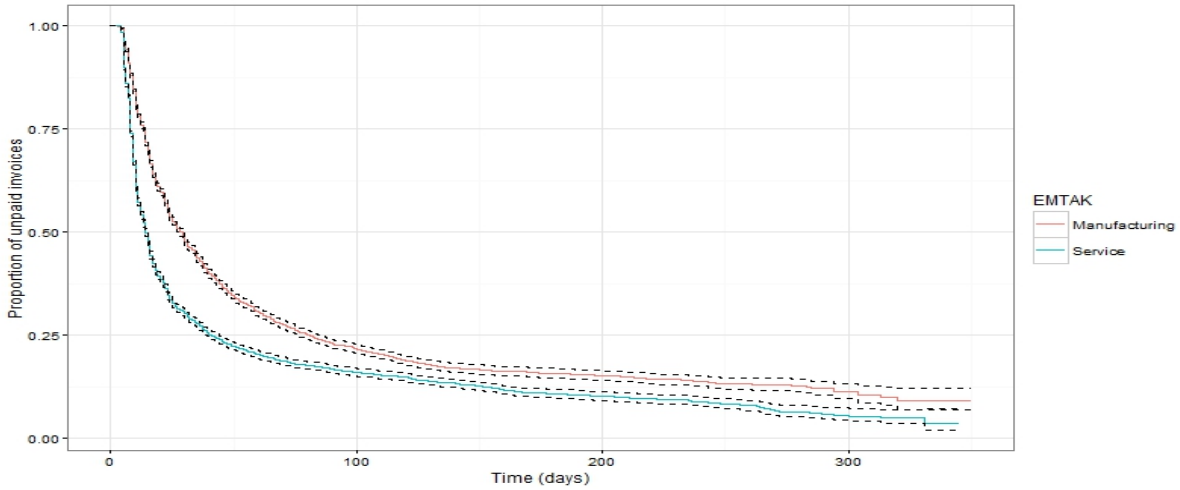

Figure 3.9: Debtors' payment behaviour by EMTAK letter group (see also Appendix B).

4 Models, Experimental Results and Analysis

In this chapter we present the experimental results of fitting two models – for first-time debtors and repeated debtors. We hypothesize that adding historical payment behaviour of debtors as predictors would improve the predictive accuracy of the model and allow us to reduce the number of other predictors.

4.1 Models without historical payment information

Model that uses the overall and external (publicly available) information about the debtor (invoice, company measures, financial data and ratios – see Appendix A and Appendix C.1) is applied to explain the payment time of a first-time debtor (a debtor for which no historical payment behaviour is available).

In Table 4.1 our train and test data is described. We acknowledge that a random selection of train and test sets may result in using future information in the training phase of model building. That appears due to the fact that one debtor may have multiple invoices in one month and with random selection there might be one invoice in the training set and second in the test set. Thus, the variables describing the debtor would mostly be the same in training and test sets.

On the other hand, if we select the training and test sets according to time sequence (by due date), the C-index calculation for the test set may be penalized. In Table 4.1 division by time means that train set contains all invoices with a due date before 1.11.2015 and test set contains all invoices with a due date after 1.11.2015 (included). Due to the construction of the study all invoices are censored at 31.12.2015, the latest. Therefore, division by time results in C-index penalization: in the test set it is more difficult to rank the payment times correctly (the maximum time of payment is 60 days) than in the training set (the maximum time of payment is 365 days).

Random division into training and test sets was performed with no restrictions. We randomly allocate 80% of invoices to train set and 20% of to test set. To be able to compare models, we use the same training and test sets in all of our experiments (applies to both, Cox PH model and Random Survival Forests).

For these reasons described we experiment with two different division types as shown in Table 4.1. It can be seen that in both cases the train and tests set are similarly balanced.

Table 4.1: Division into train and test sets randomly and taking time into account.

	Random division			**Divison by time**		
	Train	Test	**Total**	Train	Test	**Total**
No. of invoices with payment time	12 532	3 084	**15 616**	12 575	3 041	**15 616**
No. of invoices with censored time	10 276	2 618	**12 894**	9 997	2 897	**12 894**
Total	**22 808**	**5 702**	**28 510**	**22 572**	**5 938**	**28 510**

4.1.1 Cox Proportional Hazards model

Cox PH model was fit to the data using Akaike criterion rule for selecting variables (mixed selection method). Concordance index was used to evaluate how well the model performs when ranking the payment times. Error (rate) shown in model comparison tables is $1-Concordance$ (in percentages).

Experiments

We experiment with three different sets of predictors in a full model to evaluate the importance of different types of predictors (e.g. the effect of financial data from annual reports). In Table 4.2 we present the final results of fitting Cox PH model using mixed selection with AIC rule.

Table 4.2: Comparison of models with three different sets of predictors ('#of pred' shows the number of predictors in the full model, 'Pred' shown the number of predictors in the final model).

Model	**# of pred.**	**Random division**			**Division by time**		
		Pred	Error (train)	Error (test)	Pred	Error (train)	Error (test)
M1: Invoice + Company measures + Fin data + Ratios	16	14	35.53%	35.33%	15	35.35%	40.85%
M2: Invoice + Company measures + Fin data	13	11	35.55%	35.35%	13	35.34%	40.73%
M3: Company measures + Ratios	12	10	35.63%	35.51%	11	35.51%	40.98%

We can see that giving different sets of predictors has little effect on the ranking ability of model: the change in error rate between models is 0.26% the most. From the table it also follows as stated earlier, the division of train and test sets by time sequence may penalize

Concordance index calculation as the test set error is significantly larger than for random selection. At the same time, it is difficult to guess how big is the effect of penalization for division by time sequence and what is the effect of providing future information in the training phase when using random division.

The ranking ability of payment times is around 65% for all models using random division of train and test set and 59-65% for models using division by time. Recall that if the error rate is 50%, the model is no better than a coin toss. Hence, these models have some predictability but we conclude that the ranking ability of the model is not very good.

We would actually prefer the model with less predictors (model **M3**) as there is not a significant change in error rate between the three models. However, in this section we will analyze the model with all predictors (model **M1**) in more depth to have an overview of all the predictors. In all model interpretations and performance assessments we use the training and test set division by time sequence.

Model interpretation

When testing the proportional hazards assumption for model **M1**, the PH assumption is fails for 9 variables (see Appendix C.2.1, p-value < 0.05). However, the p-values for some variables are not very below 0.05 (e.g. for the age of debtor). Also, when plotting the Schoenfeld residuals, we do not notice a violation (see Figure 4.1 with residuals for the variable Sum).

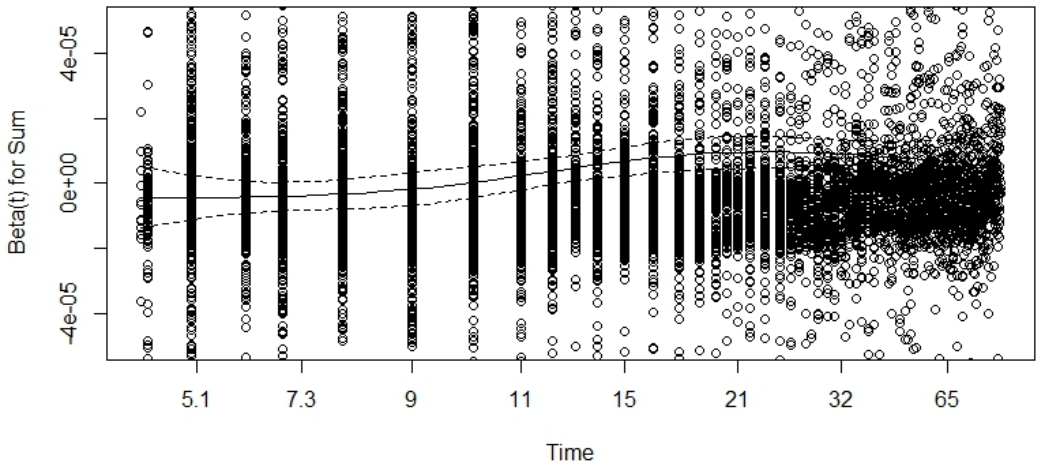

Figure 4.1: Schoenfeld residuals for the variable Sum.

In the plots of Schoenfeld residuals a non-zero slope is evidence against proportionality. A smoothing spline is shown on graph by a solid line. Systematic departures from a horizontal line are indicative of non-proportional hazards. Visual evaluation of all the graphs of predictors does not capture a significant violation which also may be caused by the wide limits of values. However, note that even 'significant' nonproportionality may make no difference to the interpretation, particularly for large sample sizes ([37], pp. 127-145), therefore we will continue to further explore the Cox PH model.

When analyzing variables in the model **M1**, the only variable that was dropped from the full set, was ratio of invoice sum (see Appendix C.2.1).

In this model **M1**, for example invoices with an EMTAK group EMTAK2 (Service) have 66% increased rate of payment when compared to group EMTAK1 (Manufacturing) under the assumption that all the other variables do not change. Invoices of debtors that have not submitted tax declarations have 52% (1/0.66=1.52) decreased rate of payment when compared to invoices of debtors that have no unsubmitted tax declarations (see Appendix C.2.1).

In conclusion, from Table 4.2 it follows that using only Company measures + Ratios as predictors (model **M3**) approximately results in the same performance as a full model **M1** where financial information and invoice sum is added. In practice we would prefer a model with less predictors (a more robust model), i.e. model **M3**.

4.1.2 Random Survival Forests (RSF)

To begin with Random Survival Forests, we need to decide on the number of trees to grow when training the model. Therefore we firstly fitted all predictors from the data to the train set and averaged over 1 000 trees. The error rate corresponding to the number of trees fit is shown in the Figure 4.2. It follows that there is not a significant change in the error rate (1–*Concordance*) after 250 trees is grown. Therefore, in our analysis we grow 250 trees in each experiment.

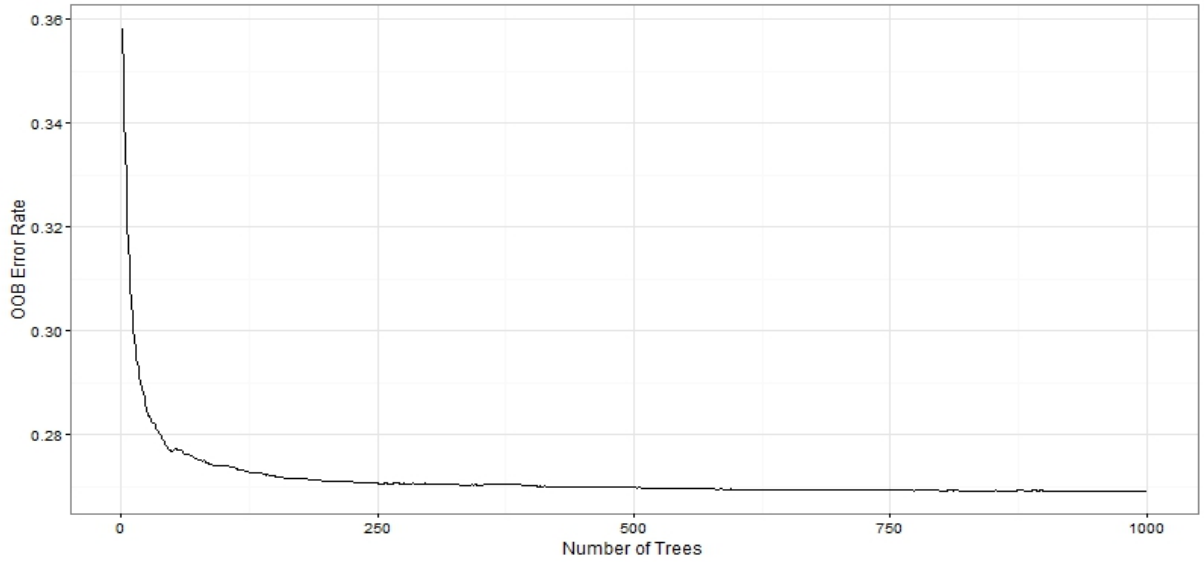

Figure 4.2: Out-of-bag error rate change respective to the number of trees grown.

When using random survival forests method it is advised to experiment with different node sizes [31]. Choosing the correct node size is important because it also reflects how many times a variable is used in the model. Recall that the terminal node size is defined as number of paid invoices. Alltogether we have 15 616 paid invoices (12 532 in train and 3 084 in the test set).

Table 4.3: Experiments with different terminal node sizes using all (16) predictors (model **M1**, random division of train and test set).

Minimum terminal node size	3	10	30	50	100	200	1 000
Average number of terminal nodes	6 015	2 531	1 007	623	315	162	29
Error (OOB)	28.11%	27.34%	27.12%	27.24%	28.08%	29.15%	32.43%
Error (test)	28.50%	27.49%	27.19%	27.24%	27.92%	28.78%	31.67%

It can be seen from the Table 4.3 that using a too small terminal node size may result in underfitting (OOB error can be decreased with bigger node size) and setting terminal node size too large may result in overfitting (OOB error starts to increase). As we have a separate validation set (test set), we will choose 50 as the terminal node size because the OOB and test error are the same with this node size. In fact, the OOB and test set errors are quite equal here which is caused by the nature of forming OOB and test sets.

Experiments

As before we experimented with different sets of predictors when fitting Cox model, we do the same when fitting RSF models.

Table 4.4: Comparison of models with different sets of predictors (terminal node size was set to 50 unique payments, 250 trees were grown).

Model	**Number of predictors**	**Random division**		**Division by time**	
		Error (train)	Error (test)	Error (train)	Error (test)
M1: Invoice + Company measures + Fin data + Ratios	16	27.46%	27.58%	26.51%	36.46%
M2: Invoice + Company measures + Fin data	13	27.81%	28.08%	26.88%	36.64%
M3: Company measures + Ratios	12	28.07%	28.22%	27.04%	37.09%

It can be seen from the Table 4.4 that error rate is the smallest when all possible predictors are fit into the model. However, when reducing the set of predictors to only Company measures and Ratios that we defined (model **M3**), the error increases only 0.60% roughly. We will look at the variable importance (VIMP) of the model **M1** with 16 predictors and model **M3** with 12 predictors in more depth to see which variables have the most effect (in all comparisons we view the models with training and test sets division by time sequence).

Best model

In the Figure 4.3 it can be seen that the predictors that most affect the error rate are EMTAK (classification into two: Service and Manufacturing) and submission of annual report in both models. When the financial data is added as predictors (sales, equity and current assets), they have a higher VIMP than ratios. This may be explained by the fact that random forests itself manages to account for interactions of different predictors due to its decision tree origin. In addition, as emphasized in Section 2.3.2, when variables are correlated, removing one variable and regrowing the forest may affect the VIMP for the other variable.

To illustrate this with an example, we can see from Figure 4.3 that when we removed financial data, the VIMP for current ratio (CR) in model **M3** increased and is larger than the VIMP for current claims ($claim\_current$). For comparison, in model **M1** the VIMP for current ratio is smaller than the VIMP for current claims. To evaluate the importance of financial data and ratios we may compare the error rates of the models. We can see

that the error rates for these three models to not change in a large scale. Hence, the additional predictors add some accuracy to the model but the difference in error rate is small.

No predictors have a negative effect on error rate (a negative VIMP would indicate that the predictive accuracy of the model would be improved when the predictor was removed). Note, that we have also fitted all three predictors of tax debt into the model because RSF might be able to capture the interaction of multiple past events, e.g. a debtor that had tax debts all three past months might take longer time to pay debts than a debtor that had no tax debts for past two months.

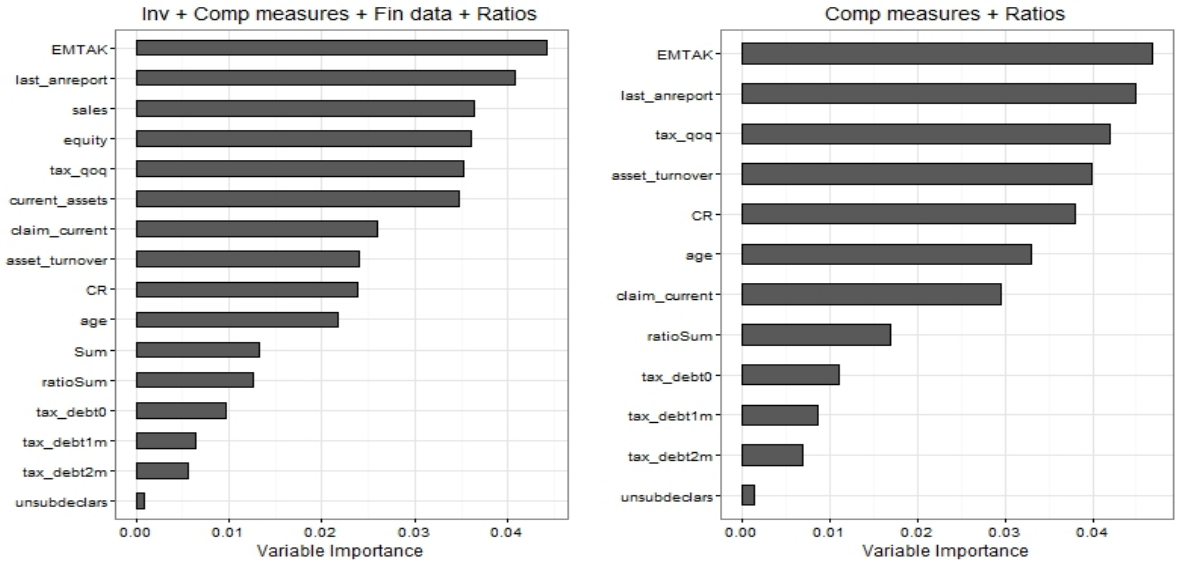

Figure 4.3: VIMP for model **M1** with all predictors (left) and for model **M3** with Company measures and Ratios as predictors (right).

4.1.3 Comparison of Cox and RSF models

If we compare the error rates of Cox model and RSF (Random Survival Forests) models, then in all cases the RSF model has smaller error rates than Cox model. However, we acknowledge that unfortunately the C-indexes used to calculate the error rates for Cox and RSF models are not directly comparable. To be more precise, the Cox model uses the predicted survival probability at the actual event time, whereas the RSF uses functions (2.15) described in Section 2.3.1. The same comparison method of C-index that uses different predicted outcome for each method, was presented in [27].

In addition, we point out that Cox model may improve if the continuous variables were transformed (e.g. log transformations or standardizing variables). In this thesis however, the main focus is to find automatized procedures without intensive human interference

and division into groups or transformation of continuous variables is not performed.

The fact that Random Survival Forests has no underlying assumptions that need to be fulfilled, is also an advantage over the Cox model.

We conclude that the Random Survival Forests has the best predictive accuracy with all predictors fit into the model. The error rate for training set is 27% and 36% for test set. It is said ([33], *survConcordance*) that the common result for survival analysis concordance is 60–70%, equivalent to 30–40% error rate. We can conclude that the model **M1** provides an acceptable predictive ability of ranking payment times.

4.2 Models with historical payment information

As the payment time of an invoice is a small time period and it depends on the current situation of the debtor, it is desired to have current and up-to-date information about the financial capability of debtor. Therefore, we define some predictors that would indicate the most recent payment behaviour (more precisely predictors for debtors at the invoice due date that would indicate problems of meeting its liabilities).

Model that uses historical payment behaviour of the debtor in addition to predictors used in Section 4.1 will be applied to data of debtors for which we have payment behaviour data from 1 month before the due date of an invoice. After defining historical variables (see Appendix A), we have a database of 16 181 invoices. The divisions into train and test sets are shown in the Table 4.5.

Table 4.5: Division into train and test sets randomly and taking time into account.

	Random division			**Division by time**		
	Train	Test	**Total**	Train	Test	**Total**
No. of invoices with payment time	7 789	1 945	**9 734**	7 492	1 719	**9 211**
No. of invoices with censored time	5 155	1 292	**6 447**	4 456	1 653	**6 109**
Total	**12 944**	**3 237**	**16 181**	**11 948**	**3 372**	**16 181**

As in Section 4.1, we again fit the models using different sets of predictors in a full model to see the effect and importance of different types of variables. To compare if adding historical data improves the models, we also fit models without predictors of 'History' to this data of 16 181 invoices (compared to Section 4.1 the number of invoices is reduced now).

4.2.1 Cox Proportional Hazards model

Again the Cox PH model was fit using Akaike criterion rule (mixed selection method) but now for the model with historical data.

Experiments

We experimented with different sets of predictors and the results are shown in Table 4.6. Adding historical data as predictors into the model reduces the error by 2% in all variations of predictors. Even the model **M9** with only Company Measures and History has a smaller error than the model with all predictors without 'History' (model **M1**).

Table 4.6: Comparison of models with different sets of predictors ('#of pred' shows the number of predictors in the full model, 'Pred' in the final model).

Model	**# of pred.**	**Random division**			**Division by time**		
		Pred.	C-index (train)	C-index (test)	Pred.	Error (train)	Error (test)
M1: Invoice + Company measures + Fin data + Ratios	16	13	34.80%	34.99%	13	34.14%	38.44%
M2: Invoice + Company measures + Fin data	13	11	34.92%	34.93%	12	34.21%	38.33%
M3: Company measures + Ratios	12	9	35.00%	35.21%	8	34.40%	38.73%
M4: Invoice + Company measures + Fin data + Ratios + History	19	17	32.02%	32.65%	15	32.05%	36.38%
M5: Invoice + Company measures + Fin data + History	16	15	32.14%	32.62%	12	32.10%	36.72%
M6: Company measures + Ratios + History	15	12	32.22%	32.65%	10	32.24%	36.57%
M7: Invoice + Company measures + Ratios + History	16	13	32.14%	32.78%	12	32.17%	36.49%
M8: Invoice + Company measures + History	13	12	32.25%	32.76%	10	32.19%	36.72%
M9: Company measures + History	12	10	32.31%	32.68%	9	32.25%	36.70%

Other variations of predictors did not have such an effect on error rate – adding and removing other types of predictors reduced or increased the error by less than 0.5%. When in Section 4.1 we analyzed the model with all predictors and smallest error rate, then in this section we will view the model that has less predictors but a competitive error rate when compared to other models.

Model interpretation

Although some models in Table 4.6 had a smaller error rate, model **M9** uses less variables and the error rate is not significantly different from other models also using historical payment behaviour as predictors. And therefore in practice we would prefer a more robust model that uses less predictors.

The output of the model **M9** is added in the Appendix C.2.2. The variables in the model are submission of tax declarations, tax debt 1 month ago, EMTAK letter, submission of 2014 annual report, other claims for the debtor currently, age (in years), average days of late invoices in the previous month, ratio of paid late invoice sums 1 month ago, ratio of outstanding late invoices 1 month ago. All these variables seem reasonable for modelling payment times of a debtor.

When analyzing the historical data predictors, we can see that increase of one overdue day of paid late invoices in the previous month will result in the decrease of payment rate 3% (1/0.97=1.03) under the assumption that all the other variables do not change. One unit increase in the ratio of paid late invoice sums in the previous month results in 86% increase of payment rate.

The proportional hazard assumption is again violated (see Appendix C.2.2) for 5 variables (p-value < 0.05). As before, when plotting Schoenfeld residuals, visual assessment does not capture a major violation for those variables. Even though we pointed out in Section 4.1.1 that a violation of proportional hazards may make no difference for large sample sizes, it is still one of the reasons why Random Survival Forests that has no assumptions to be taken into consideration may prove to be a preferred method.

4.2.2 Random Survival Forests (RSF)

As in Section 4.1.2, we experiment with different input variables to see the effect of variables. In addition, we perform experiments with historical behaviour variables to see if these predictors compensate the need of some types of predictors.

Experiments

We use the same sets of predictors as for the Cox model in Section 4.2.1. The results of RSF error rates for different model types are presented in Table 4.7. Similarly to Cox models in Section 4.2.1 it can be seen that model **M9** is comparable to **M1** in terms of error rates. This means that a model using less predictors with 'History', has approximately the same accuracy in ranking payment times like a full model without 'History'.

Table 4.7: Comparison of models with different sets of predictors (terminal node size was set to 50 unique payments, 250 trees were grown).

Model	**Number of predictors**	**Random Division**		**Division by time**	
		Error (train)	Error (test)	Error (train)	Error (test)
M1: Invoice + Company measures + Fin data + Ratios	16	28.14%	28.15%	26.92%	35.05%
M2: Invoice + Company measures + Fin data	13	28.41%	28.48%	27.39%	35.46%
M3: Company measures + Ratios	12	28.68%	28.60%	27.54%	35.89%
M4: Invoice + Company measures + Fin data + Ratios + History	19	26.34%	26.72%	25.06%	33.07%
M5: Invoice + Company measures + Fin data + History	16	26.58%	27.04%	25.40%	33.32%
M6: Company measures + Ratios + History	15	26.85%	26.99%	25.39%	33.35%
M7: Invoice + Company measures + Ratios + History	16	26.70%	27.03%	25.37%	33.24%
M8: Invoice + Company measures + History	13	28.03%	28.19%	26.71%	34.19%
M9: Company measures + History	12	28.08%	28.43%	26.79%	34.79%

We can overall conclude that similarly to experiments with Cox model in Section 4.2.1, it follows from the Table 4.7 that adding historical payment behaviour data of debtors into the model improves model accuracy in all variations of predictors.

When comparing the models we can see that a full model **M4** has the best performance for both train and test sets division types. As pointed out in 4.1.1, in model comparisons we use the division into train and test sets by time sequence.

Best model

We compare the variable importance of the full model without historical data (model **M1**) and the full model with historical data (model **M4**) in Figure 4.4. We can see that the predictors reflecting historical payment behaviour have a large variable importance, especially average overdue days of paid late invoices in the previous month.

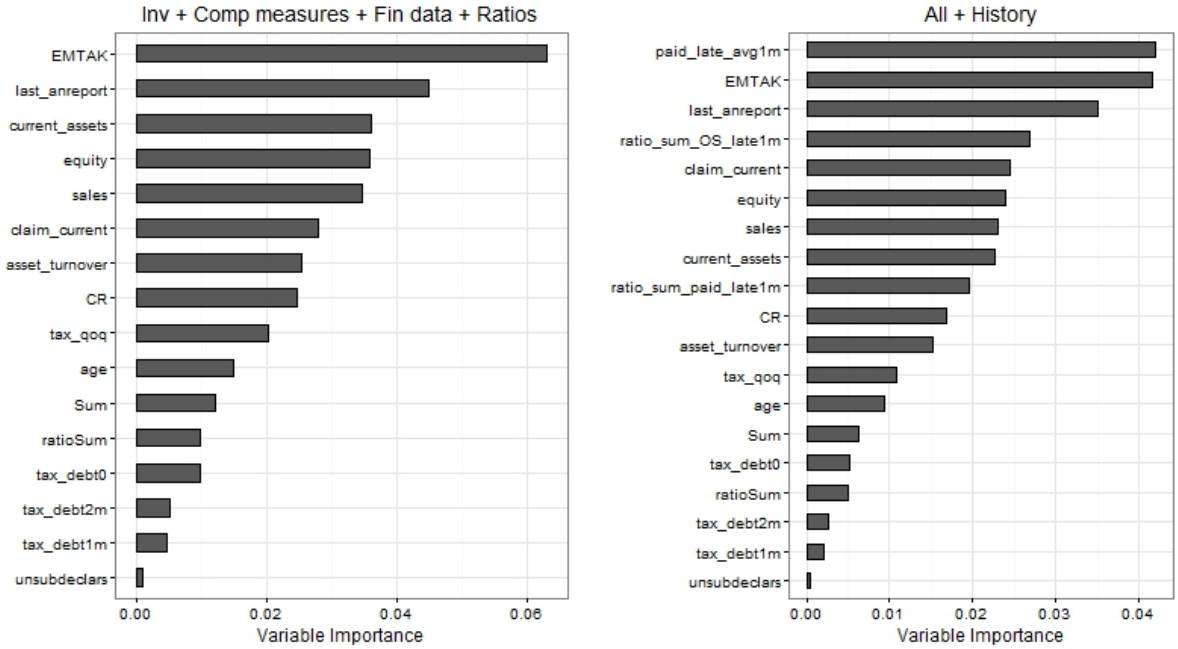

Figure 4.4: Comparison of VIMP for RSF model **M1** (left) and for RSF model **M4** (right).

The same applies here as described in Section 4.1.2 – when we remove some variables, the VIMP for other variables may change. Thus, to assess the actual effect of predictors we can view the error rates in Table 4.7 for different sets of predictors. For example, to see the effect of ratios and financial data, compare models **M5** and **M7**. We can see that the error rates are approximately the same which indicates that both variables have the similar impact to the performance of model.

4.2.3 Comparison of Cox and RSF models

When comparing the Cox and RSF, it follows that Random Survival Forests models have smaller error rate in all cases of predictor sets. However, as pointed out in Section 4.1.3, the C-indexes (and therefore error rates) of Cox and RSF models are not directly comparable. We can also see that adding historical data as predictors improved both models: Cox and RSF error rates were reduced roughly by 2%.

When compared to Cox PH model, the Random Survival Forests has multiple advantages: it has no underlying assumptions and since it makes splits on each node, there is no need for transformation of variables (e.g. log transformation or standardization of variables).

To summarize, we can conclude that for our data Random Survival Forests performed better in ranking payment times and is a preferred model since it has no underlying assumptions. When historical data is available for debtors, using previous payment behaviour as predictors improves the model accuracy of ranking payment times correctly.

4.3 Conclusions and applications in practice

In this section we present some examples of the results. We will restrict our examples to using only the Random Survival Forests model with historical data (model **M4**).

Let us compare three examples of invoices in the test set, see Figure 4.5. On the graphs we have presented the predicted (RSF model **M4**) survival curves for the corresponding invoices.

We can see that the invoice that was paid in less time (Invoice 1) has a steeper survival curve, going to zero faster. In contrast, invoice for which we know it hasn't been paid for at the 51st overdue day (Invoice 3), has a slowly sloping (downwards) survival curve. Invoices that are similar to Invoice 3 in their predictor variables, have probability of $\approx 37\%$ at the 60th overdue day to be unpaid. If we compare the median payment times to be then Invoice 1 has a median of 9 days, Invoice 2 has 23 days and Invoice 3 has 31 days. Median represents the overdue day by which the probability is 0.5 that the invoice with similar predictors has not been paid for (and 0.5 that it has have been paid for) and we can see that the median days also indicate the steepness of the survival curve.

In this example we can conclude that the ranking of invoices is appropriate and the survival curves are representative of the actual payment (or censor) times. Note, that the steepness of a survival curve also represents the payment rate due to the relation (2.5).

All the debtors of these three invoices are from the Manufacturing classification of EMTAK letter. The annual report of 2014 was submitted for Invoice 1 and 2, but not for invoice 3. According to the VIMP (Figure 4.4), paid average late days in previous month has a high importance in the model. For Invoice 1 the average was 0 days, for Invoice 2, it was 11 days and for Invoice 3 it was 39 days.

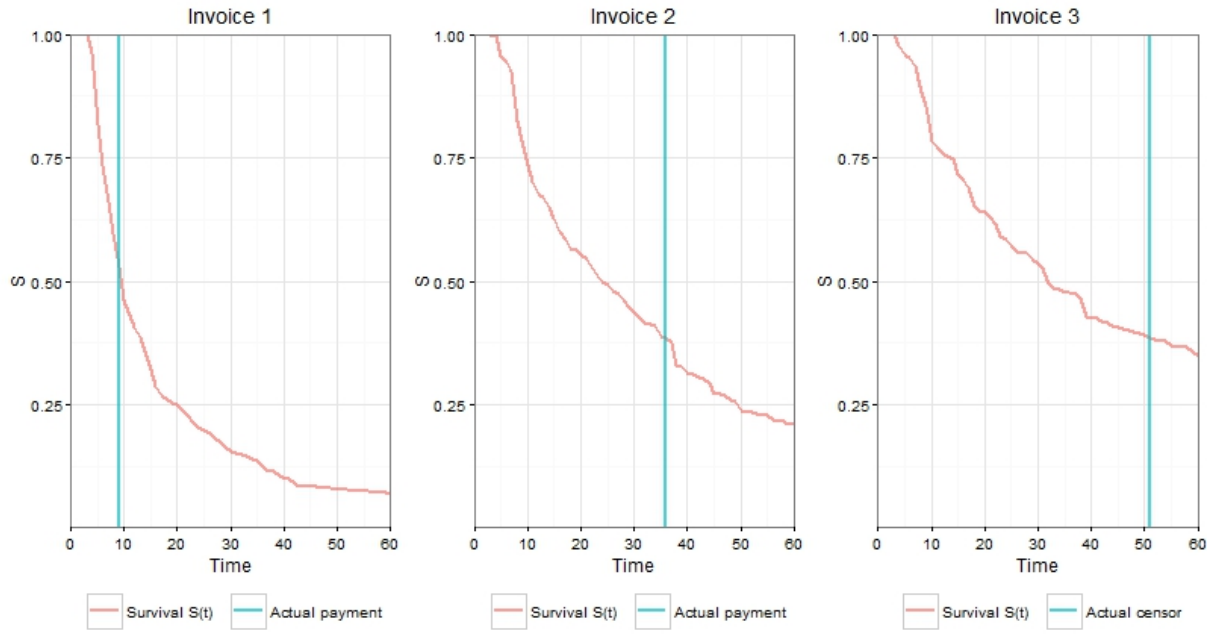

Figure 4.5: Comparison of predicted survival curves for three different invoices in test data (division by time). Invoice 1 has actual payment time of 9 days and Invoice 2 of 36 days after due date, Invoice 3 has censoring time of 51 days.

For the credit management firm it is of interest what is the probability that a late invoice is paid before a certain time. We illustrate the effects of predictor values to the probability of payment with an example. We choose a number of predictors (mostly based on VIMP shown in Table 4.4 for model **M4**). Other variables are fixed by the median values presented in Appendix A. Time period of interest in this case is 2 weeks (14 days). In the Table 4.8 we present the probability that a late invoice is paid before 14 overdue days (the probability is equal to $1 - \hat{S}(14)$).

Note that in the table we have fixed the initial values of predictors (Inv1) with median values (see Appendix A). For other invoices we have changed the variables one-by-one. As a reminder, EMTAK value 2 is Service sector and EMTAK value 1 is Manufacturing sector. Yes value for annual report (2014) submitting means that the report was submitted and No means the contrary. Other predictor values are continuous.

Table 4.8: Example of predicted probabilities that a late invoice is paid before 14 overdue days, conditional on predictor values.

	Inv1	**Inv2**	**Inv3**	**Inv4**	**Inv5**	**Inv6**	**Inv7**	**Inv8**	**Inv9**	**Inv10**
paid_late_avg1m (days)	4	50	4	50	4	4	4	4	4	4
EMTAK	2	2	1	1	2	1	2	1	2	2
last_anreport	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes
age (in years)	11	11	11	11	11	11	1	30	11	11
equity (in thousands)	89	89	89	89	89	89	89	89	1 000	89
tax_debt0 (in euros)	0	0	0	0	0	0	0	0	0	100 000
Prob. of payment (1–$\hat{S}(14)$)	0.89	0.25	0.50	0.25	0.50	0.43	0.58	0.51	0.84	0.78

We can see that for a median invoice (Inv1) where the debtor is from Service and the last annual report was submitted, the probability that the invoice is paid before 14 days, is 89%. Considering the overall distribution of payment times (refer to Figure 3.4), this seems plausible as the majority of invoices are paid before 15 days.

For invoice Inv2 the average days of paid late invoices in the previous month is set to 50 days. We can see that the previous payment behaviour has a huge effect – the probability of payment before 14 days is only 25% for this invoice. It can also be seen that changing the EMTAK classification has a huge effect (compare Inv1 and Inv3). But if we compare Inv2 and Inv4, the EMTAK classification has no effect, which indicates that the previous payment behaviour affects the probability more than EMTAK classification.

The effect of not submitting last annual report can be seen from Inv5 and Inv6. For Service sector the probability of payment decreased by 39% (compare Inv1 and Inv5), for Manufacturing sector the probability of payment decreased from 50% to 43% (compare Inv3 and Inv6). Age was changed for different EMTAK classifications and interestingly both – decreasing age and increasing age – resulted in increased probability of payment (compare Inv5 and Inv7; Inv6 and Inv8 correspondingly).

Setting equity to 1 million resulted in a small decrease of probability of payment (compare Inv1 and Inv9). It may seem odd at first sight but actually it may be caused by the fact that larger entities have specific times for payments and do not keep an eye on the exact due dates. Setting tax debt to 100 000 euros, decreased the probability of payment by 11% (compare Inv1 and Inv10) which is in correspondence with our hypothesis that companies with tax debt have a decreased payment probability.

In the Figure 4.6 we present the survival curves for 5 selected invoices. We can see that the predicted survival curve for the invoice with increased equity (Inv9) is similar to invoice Inv1. Invoices Inv3 and Inv8 (both from Manufacturing sector) are similar, invoice Inv8 has decreased probability of payment after 20 days when compared to invoice Inv3. That is because the debtor of invoice Inv8 has not submitted last annual report. Inv4 has the worst predicted payment probability compared to other 4 invoices due to the large overdue days of previous paid late invoices.

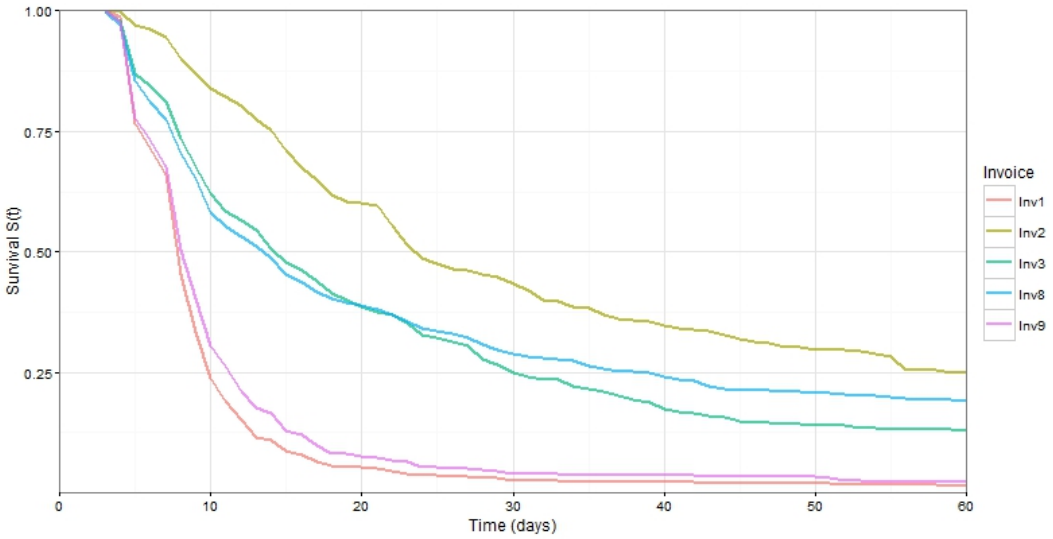

Figure 4.6: Survival curves for invoices Inv1, Inv2, Inv3, Inv8 and Inv9 from Table 4.8.

Finally, to have an intuitive comparison with the credit management firm's expert (manual) decisions and our predicted payment probabilities, we present box plots in the Figure 4.7. The credit management firm gives advice to the creditor whether to grant some extra credit (make additional sales) to the debtor or not. The advice is given in three categories: Yes, No or Wait.

These 4 box plots in the Figure 4.7 represent the predicted probabilities in the test set (November and December 2015) that the invoice is paid before 14, 30, 45 and 60 days correspondingly. We can see that for the majority of Yes decisions done by the experts, our predicted payment probability is much higher at all time points than for the No decisions. The predicted probabilities for the majority of Wait decisions is overlaying Yes and No decisions as expected. These box plots support the appropriateness of the times ranking ability of our model (RSF model **M4**).

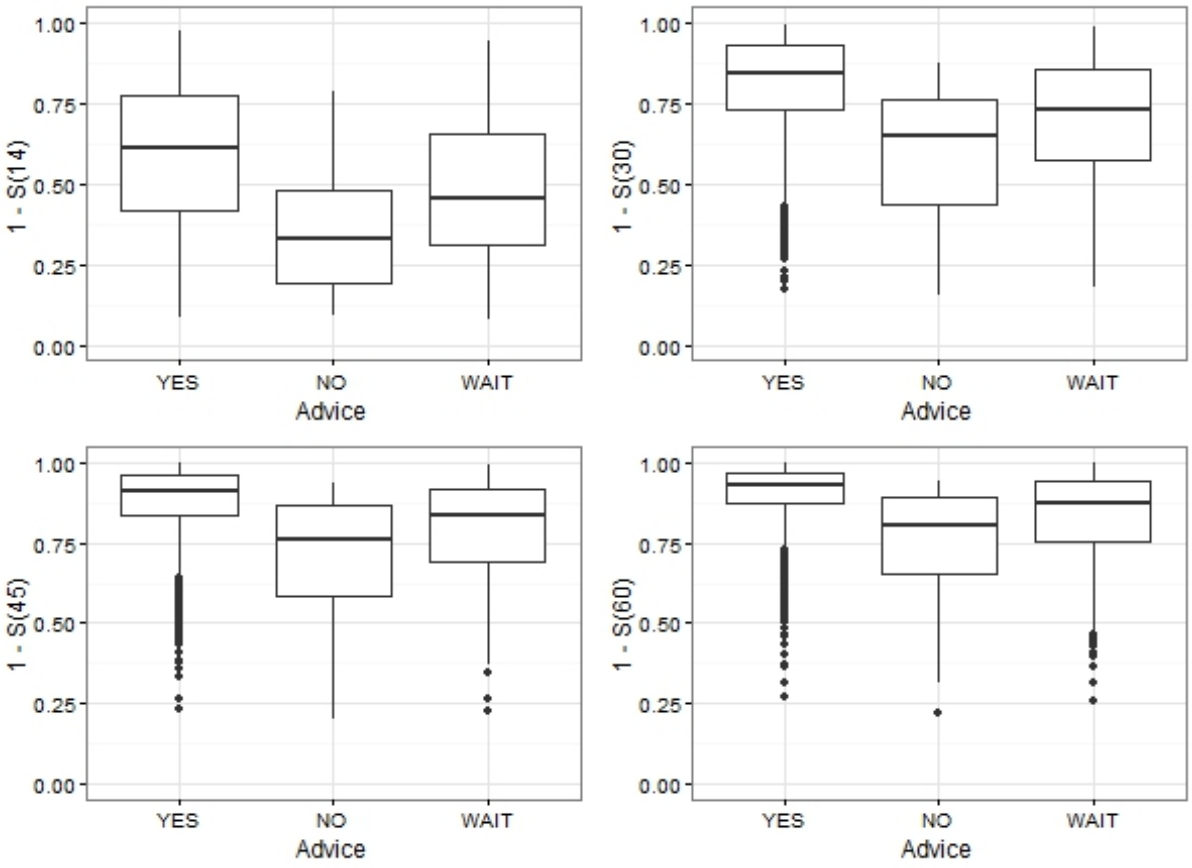

Figure 4.7: Predicted probabilities of payment (on the vertical axis) before 14 days (upper left), 30 days (upper right), 45 days (lower left) and 60 days (lower right) compared to the credit advice given by the credit management firm. The probabilities are predicted on test set (division by time) and the credit advice of the debtor is correspondingly taken from the period of November or December 2015.

Conclusions and Future Work

Predicting failure of companies and loan defaults has been a well-researched area in the past and it has a wide range of use in the financial sector. However, usually traditional methods that were developed in the past are applied and the models mostly based on annual reports of the companies which give multiple restrictions and are not up to date to reflect current situation of companies.

This thesis focused on invoice-to-cash process in business-to-business sales, more precisely modelling the payment behaviour of debtors with late invoices. Survival analysis that in recent years has been found to be a suitable method when predicting loan defaults was applied. Earlier research has also shown that the performance of survival analysis is competitive with logistic regression in credit scoring. In addition, survival analysis allows us to analyze right-censored data and was therefore a preferred method in this thesis. Cox Proportional Hazards model that is traditionally used with using survival analysis to model time-to-event data was compared to a recent and novel method of Random Survival Forests.

To use more up-to-date information in modelling the payment behaviour of debtors, monthly updated information of tax debts and changes in tax payments was used. It followed from the experiments that adding tax debt information into the models resulted in a more accurate model. Moreover, we experimented with the use of historical payment behaviour of debtors as predictors. It followed that historical information of even only 1 month back improved the models such that it allowed us to remove information of financial data without sacrificing the accuracy of the model.

Application of the recently developed ensemble method, Random Survival Forests, proved to be successful in ranking the payment times of late invoices. The main advantages of Random Survival Forests are that it has no underlying restrictions or assumptions for the data, it automatically manages to account for interactions of different predictors without much interference of the end-user.

The models presented in this thesis were the first experiments to explore methods and possibilities of partially automatizing the process of decision making in a credit management firm. We introduced a possibility to apply two models – one for first-time debtors and one for debtors that already have historical payment behaviour information available for the credit management firm.

To summarize, in this thesis we managed to use an alternative database from accounting systems, to derive approximate payment times of invoices and to build a model to rank those events. In conclusion two models that could be used in practice to evaluate the payment probability of a debtor were proposed. First model uses only external public information and is therefore applicable for all debtors to give initial evaluations, second model uses additionally previous payment history and is applicable for repeated debtors to give more precise evaluations.

Future work. As the data available for analysis in this thesis comprised of one year, it would be of interest to implement the results in 2016 to review the performance of models proposed. With the increase of the credit management firm's clients (in terms of creditors providing information about debtors), applying the models in 2016 could result in better performance as there might be more historical information available for the debtors. In this thesis we restricted the predictors of historical payment behaviour data to 1 month and 3 variables. It would be of interest to define some more variables to explore if the models could be improved.

Acknowledgements. I would like to thank my supervisors Imbi Traat and Peep Küngas for their continuous support and relevant advice on this topic.

Bibliography

[1] M.J. Peel, N. Wilson, C. Howorth. (2000). Late Payment and Credit Management in the Small Firm Sector: Some Empirical Evidence. *International Small Business Journal.* **18**(2), 17–37.

[2] W. Connell. (2014). Economic Impact of Late Payments. *Economic Papers 531.*

[3] Investopedia.
http://www.investopedia.com/ (last visited 09.05.2016).

[4] Techno Func.
http://www.technofunc.com/index.php/functional-skills2/order-to-cash/item/what-is-invoice-to-cash-process (last visited 10.02.2016).

[5] Accounting Tools.
http://www.accountingtools.com/questions-and-answers/what-is-a-sales-ledger.html (last visited 10.02.2016).

[6] L. Yu, S. Wang, K.K. Lai, L. Zhou. (2008). *Bio-inspired credit risk analysis: computational intelligence with support vector machines.* Springer.

[7] S. Zeng, I. Boier-Martin, P. Melville, C. Murphy, C.A. Lang. (2008). Using Predictive Analysis to Improve Invoice-to-Cash Collection. In *Proceedings of 14th Conference on Knowledge Discovery and Data Mining(KDD-08).*

[8] Intrum Justitia. (2015). *European Payment Report 2015.*

[9] International Credit Insurance & Surety Association.
http://www.icisa.org/history/1507/ (last visited 10.02.2016)

[10] KredEx Krediidikindlustus AS.
http://www.kredex.ee/en/kredex/kredex-krediidikindlustus-as/
(last visited 10.02.2016)

[11] M. Tamari. (1966). Financial Ratios as a Means of Forecasting Bankruptcy. *Management International Review.* **6**(4), 15–21.

[12] M. Stepanova, L. Thomas. (2002). Survival Analysis Methods for Personal Loan Data. *Operations Research.* **50**(2), 277–289.

[13] B. Narain. Survival analysis and the credit granting decision. In L.C. Thomas, J.N. Crook and E.D.B. Edelman. Credit Scoring and Credit Control. (1992). *Oxford University Press.* 109–122.

[14] J. Banasik, J.N. Crook, L.C. Thomas. (1999). Not if but When will Borrowers Default. *The Journal of the Operational Research Society.* **50**(12), 1185–1190.

[15] H. Peiguang. (2015). *Predicting and Improving Invoice-to-Cash Collection Through Machine Learning.* Massachusetts Institute of Technology.

[16] R.A. McDonald, A. Matuszyk, L.C. Thomas. (2010). Application of survival analysis to cash flow modelling for mortgage products. *Insight.* **23**(1), 1-–14.

[17] R. Cao, J.M. Vilar, A. Devia. (2008). Modelling consumer credit risk via survival analysis. *Universidade da Coruña*

[18] J.-K. Im, D.W. Apley, C. Qi, X. Shan. (2012). A time-dependent proportional hazards survival model for credit risk analysis. *Journal of the Operational Research Society.* **63**, 306-321.

[19] R. Man. (2014). *Survival analysis in credit scoring: A framework for PD estimation.* Quantitative Risk Analytics & University of Twente.

[20] S. Despa. What is Survival Analysis? *Cornell University, Cornell Statistical Consulting Unit, Newsletter.*
https://www.cscu.cornell.edu/news/statnews/stnews78.pdf (last visited 26.04.2016)

[21] M. Tableman, J.S. Kim. (2015). *Survival Analysis Using S: Analysis of Time-to-event-data.* Chapman & Hall/CRC.

[22] D.G. Kleinbaum, M. Klein. (2005). *Survival Analysis: A Self-Learning Text,* Third Edition. Springer.

[23] F.E. Harrell (Jr). (2001). *Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer.

[24] D.W. Hosmer (Jr.), S. Lemeshow. (1998). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Wiley.

[25] G. James, D. Witten, T. Hastie, R. Tibshirani. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

[26] W.N. Venables, B.D. Ripley. (2002). *Modern Applied Statistics with S.* Fourth Edition. Springer.

[27] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer. (2008) Random Survival Forests. *Ann. Appl. Statist.* **2**(3), 841–860.

[28] U.B. Mogensen, H. Ishwaran, T.A. Gerds. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software.* **50**(11).

[29] H. Ishwaran, U.B. Kogalur. (2007). Random Survival Forests for R. *R News.* **7**(2), 25–31.

[30] S. Sonderby. (2004). *Non-parametric survival analysis in breast cancer using clinical and genomic markers*. Technical University of Denmark.

[31] H. Ishwaran, U.B. Kogalur. (2016). *randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC).* R package version 2.1.0.

[32] H. Ishwaran. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics.* **1**, 519—537.

[33] T. Therneau. (2015). *A Package for Survival Analysis in S.* R package version 2.38, http://CRAN.R-project.org/package=survival.

[34] J. Ehrlinger. (2015). *ggRandomForests: Visually Exploring Random Forests*. R package version 1.2.1, http://cran.r-project.org/package=ggRandomForests.

[35] Äriseadustik. (RT I, 30.12.2015, 73).
https://www.riigiteataja.ee/akt/130122015073 (last visited 26.04.2016)

[36] M. Kuhn, K. Johnson. (2013). *Applied Predictive Modeling*. Springer.

[37] T.M. Therneau, P.M. Grambsch. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

[38] Euroopa Parlamendi ja Nõukogu määrus (EL) nr 549/2013.
Euroopa Liidus kasutatava Euroopa rahvamajanduse ja regionaalse arvepidamise süsteemi kohta.
http://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:32013R0549&qid=1460464412758&from=ET (last visited 05.04.2016)

A Predictors

A.1 Invoice data

Predictor	Description	Class	Values
Sum	Base sum of the invoice	Continuous	Min: 0 | Max: 390 105 Median: 154 | Mean: 1 005

A.2 Company measures

The following data of company measures and information about tax debt and paid taxes.

Register OÜ database

Predictor	Description	Class	Values
last_anreport	Submission of 2014 annual report - if the debtor has submitted the 2014 report or not at the due date of the invoice	Factor	Yes / No (15 038 / 13 472)
claim_current	All other current claims for the company	Continuous	Min: -96 805 | Max: 930 713 Median: 0 | Mean: 778
age	Age of the company in years	Positive Integer	Min: 0 | Max: 78 Median: 11 | Mean: 12
EMTAK	EMTAK letter (field of activity) of the company (see Appendix B)	Factor	EMTAK1 / EMTAK2 (15 096 / 13 414)

Data originating from Estonian Tax and Customs Board database

Predictor	Description	Class	Values
unsubdeclars	Unsubscribed declars (modified) – dummy variable if the debtor has had unsubmitted tax declarations	Factor	Yes / No (1 120 / 27 390)
tax_debt0	Tax debt current month	Continuous	Min: 0 | Max: 1 228 728 Median: 0 | Mean: 2 903
tax_debt1m	Tax debt 1 month ago	Continuous	Min: 0 | Max: 1 228 728 Median: 0 | Mean: 2 814
tax_debt2m	Tax debt 2 months ago	Continuous	Min: 0 | Max: 1 228 728 Median: 0 | Mean: 2 599
tax_qoq	Tax paid change ratio over tax paid in previous quarter	Continuous	Min: -3 000 | Max: 3 000 Median: 0 | Mean: 313

A.3 Financial data

Predictor	Description	Class	Values
sales	Total sales. Yearly sales of the debtor (2014)	Continuous	Min: 0 | Max: 1.4e+09 Med: 5.8e+05 | Mean: 6.3e+06
current_assets	Current assets in the balance sheet (2014)	Continuous	Min: 0 | Max: 1.2e+09 Med: 1.5e+05 | Mean: 2.1e+06
total_assets	Total assets in the balance sheet (2014)	Continuous	Min: 0 | Max: 1.5e+09 Med: 2.7e+05 | Mean: 5.2e+06
shortterm_ liabilities	Short-term liabilities (also referred to as current assets) in the balance sheet (2014)	Continuous	Min: 0 | Max: 6.1+e08 Med: 1.1e+05 | Mean: 1.8e+06
equity	Equity in the balance sheet (2014)	Continuous	Min: -7.5e+06 | Max: 1.5e+09 Med: 8.9e+04 | Mean: 2.3e+06

A.4 Ratios

Additional ratio variables were introduced to test if ratios provide more information than annual information.

Predictor	Description	Class	Values
ratioSum	Sum / (Total sales / 52). *Invoice base sum ratio over weekly turnover (weekly turnover = 2014 total sales / 52)*	Continuous	Min: 0 | Max: 350 Median: 0.02 | Mean: 0.33
asset_turnover	Total assets / Total sales	Continuous	Min: 0 | Max: 74 Median: 2.55 | Mean: 5.10
CR	Current ratio = Current assets / Current liabilities	Continuous	Min: 0 | Max: 417 Median: 1.07 | Mean: 3.88

A.5 Historical payment behaviour data

Historical payment behaviour data was introduced to test if it would result in a more accurate model. Historical information is available only for 16 181 invoices (see Section 4.2).

Predictor	Description	Class	Values
ratio_sum_ paid_late1m	Ratio paid late(sum). Invoice sums paid late divided by all invoices paid within 30 days before due date of the invoice under observation	Continuous	Min: 0 | Max: 1 Med: 0.72 | Mean: 0.76
ratio_sum_ OS_late1m	Ratio OS late (sum). Invoice sums that were late at 30 days before due date divided by all invoices sums outstanding 30 days before due date of the invoice under observation	Continuous	Min: 0 | Max: 1 Med: 1.00 | Mean: 0.76
paid_late_ avg1m	Average days late. Average overdue days of late invoices that were paid 30 days before due date of the invoice under observation	Continuous	Min: 0 | Max: 103 Med: 4.23 | Mean: 8.63

More specified calculation of the historic payment behaviour is described in the table below.

#	Predictor	Description	Class
1	No. of invoices OS	Number of invoices outstanding 1 month before	Positive integer
2	No. of invoices paid	Number of invoices paid 1 month before	Positive integer
3	No. of invoices paid late	Number of invoices paid late 1 month before	Positive integer
4	No. of invoices paid on time	Number of invoices paid on time 1 month before	Positive integer
5	No. of late invoices OS	Number of late invoices outstanding 1 month before	Positive integer
6	Sum of late invoices OS	Sum of late invoice sums outstanding 1 month before	Positive integer
7	Sum of invoices OS	Sum of invoice sums outstanding 1 month before	Positive integer
8	Sum of invoices paid	Sum of invoice sums paid 1 month before	Positive integer
9	Sum of invoices paid late	Sum of invoice sums paid late 1 month before	Positive integer
10	Average sum of invoices paid late	Average of invoice sums paid late 1 month before	Positive integer
11	Sum of invoices paid on time	Sum of invoice sums paid on time 1 month before	Positive integer
12	Ratio paid late (no.)	Ratio 3 over 2	Continuous
13	Ratio paid late (sum)	Ratio 9 over 8	Continuous
14	Ratio OS late (no.)	Ratio 5 over 1	Continuous
15	Ratio OS late (sum)	Ratio 6 over 7	Continuous
16	Average days late	Average overdue days of paid late invoices during past 30 days	Continuous

B Data description

Debtors by EMTAK letter

In Estonia, there are 21 EMTAK letter specifying the field of activity of a company. In our data, 19 EMTAK letters are represented. The following frequency table presents the invoice debtor's field of activity in our data.

EMTAK letter	Description	No. of debtors	No. of invoices	Percentage of invoices
A	Agriculture, forestry and fishing	244	536	2%
B	Mining and quarrying	18	149	1%
C	Manufacturing	1 186	7 198	25%
D	Electricity, gas, steam and air conditioning supply	21	62	0%
E	Water supply, Sewerage, waste management and remediation activities	34	182	1%
F	Construction	1 100	6 824	24%
G	Wholesale and retail trade; Repair of motor vehicles and motorcycles	1 511	8 441	30%
H	Transportation and storage	375	2 131	7%
I	Accommodation and food service activities	138	387	1%
J	Information and communication	72	138	0%
K	Financial and insurance activities	18	24	0%
L	Real estate activities	140	333	1%
M	Professional, scientific and technical activities	235	656	2%
N	Administrative and support service activities	179	772	3%
O	Public administration and defense; compulsory social security	29	75	0%
P	Education	36	72	0%
Q	Human health and social work activities	19	27	0%
R	Arts, entertainment and recreation	134	198	1%
S	Other service activities	95	160	1%
T	Activities of households as employers; undifferentiated goods and services producing activities for households for own use	0	0	0%
U	Activities of extraterritorial organisations and bodies	0	0	0%
NA	The field of activity is not specified	47	145	1%

The classification of EMTAK letters into three groups as suggested by the ESA 2010 ([38], p. 550) is presented in the following table.

Group	EMTAK letters	Description
G1	A	Agriculture, forestry and fishing
G2	B, C, D, E, F	Mining; Manufacturing; Electricity gas, steam and air conditioning supply; Sewerage, waste management and remediation activities; Construction
G3	G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U	Services

We have assigned the debtors that have not specified the EMTAK letter to the group G3 in the analysis part as the group G3 is more general and the reason for a company not to specify its field of activity might have been not finding a proper subgroup.

The group G1 only consists of 536 invoices while there are 14 415 in group G2 and 13 559 invoices in group G3. For this reason we add debtors with EMTAK letter A to the group G2 as Agriculture, forestry and fishing in its essence fits better to the Industrial activity that to the Service activities.

Finally, we have defined more robust groups for EMTAK letters as follows:

Group	EMTAK letters	No. of debtors	No. of invoices	Percentage of invoices
EMTAK1 (Manufacturing)	A, B, C, D, E, F	2 650	14 951	52%
EMTAK2 (Service)	G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U	2 981	13 559	48%

C Models

In experimental part of the thesis different types of models were fit to the data to be able to see the effect of different predictors. In this Appendix we introduce which predictors represent different models.

C.1 Model building - explanation of predictors fit into the models

Invoice

Data from the invoice (see Appendix A.1) - base sum of the invoice. In total 1 predictor.

Company measures

Data reflecting the tax debts, changes in paid tax and overall company (debtor) measures (all the variables introduced in Appendix A.2). In total 9 predictors.

Financial data

Data reflecting the company size and financial situation. Note that not all the predictors from Appendix A.3 were fit to the models. Predictors of financial data that were fit in models were Total sales, Current assets, Equity. In total 3 predictors.

Short-term liabilities was omitted as we already have *Claims current* in *Company measures*. We consider *Claims current* to be more representative of debtor's current financial situation.

Total assets was omitted as total assets contain *Current assets* as well and we consider *Current assets* to be more representative liquidity of the debtor.

Ratios

Data of the additionally defined ratios (see Appendix A.4). In total 3 predictors.

History

Data of the additionally defined historical payment behaviour (see Appendix A.5). Variables that were fit into the model, were *Ratio of paid late*, *Ratio of OS late*, *Average days late*. In total 3 predictors.

C.2 Examples of fitted models

C.2.1 Models without historical payment behaviour

Cox model

Train and test data division by time. Predictors for full model: Invoice + Company measures + Financial data + Ratios

```
> fitform.2.all <- Surv(TIME, STATUS) ~ Sum +
                   sales + equity + current_assets +
                   unsubdeclars + tax_debt0 + tax_debt1m + tax_debt2m +
                  EMTAK + last_anreport + claim_current + age + tax_qoq +
                   ratioSum + asset_turnover + CR
> fit.cox.2 <- coxph(fitform.2.all, data=trainTime)
> cox.train.2 <- stepAIC(fit.cox.2, direction='both', trace=TRUE)
> summary(cox.train.2)
Call:
coxph(formula = Surv(TIME, STATUS) ~ Sum + sales + equity + current_assets +
unsubdeclars + tax_debt0 + tax_debt1m + tax_debt2m + EMTAK +
last_anreport + claim_current + age + tax_qoq + asset_turnover +
CR, data = trainTime)

n= 22572, number of events= 12575

                              coef  exp(coef)   se(coef)      z Pr(>|z|)
Sum                       2.970e-06  1.000e+00  1.315e-06  2.259  0.02391 *
sales                     1.321e-09  1.000e+00  5.113e-10  2.583  0.00978 **
equity                    3.747e-09  1.000e+00  7.599e-10  4.932 8.16e-07 ***
current_assets           -3.160e-09  1.000e+00  1.564e-09 -2.021  0.04324 *
unsubdeclars1            -4.205e-01  6.568e-01  4.869e-02 -8.635  < 2e-16 ***
tax_debt0                -3.176e-06  1.000e+00  1.470e-06 -2.161  0.03067 *
tax_debt1m               -4.265e-06  1.000e+00  1.647e-06 -2.591  0.00958 **
tax_debt2m                2.539e-06  1.000e+00  1.416e-06  1.793  0.07303 .
EMTAKEMTAK2               5.048e-01  1.657e+00  1.825e-02 27.663  < 2e-16 ***
last_anreportUNSUBMITTED -4.354e-01  6.470e-01  1.969e-02 -22.109 < 2e-16 ***
claim_current            -1.502e-05  1.000e+00  1.504e-06 -9.992  < 2e-16 ***
age                       1.800e-02  1.018e+00  1.479e-03 12.169  < 2e-16 ***
tax_qoq                  -1.330e-04  9.999e-01  1.023e-05 -12.998 < 2e-16 ***
asset_turnover           -1.803e-03  9.982e-01  8.917e-04 -2.022  0.04321 *
CR                       -6.515e-04  9.993e-01  3.897e-04 -1.672  0.09454 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                          exp(coef) exp(-coef) lower .95 upper .95
Sum                          1.0000     1.0000    1.0000    1.0000
sales                        1.0000     1.0000    1.0000    1.0000
equity                       1.0000     1.0000    1.0000    1.0000
current_assets               1.0000     1.0000    1.0000    1.0000
unsubdeclars1                0.6568     1.5226    0.5970    0.7225
tax_debt0                    1.0000     1.0000    1.0000    1.0000
tax_debt1m                   1.0000     1.0000    1.0000    1.0000
tax_debt2m                   1.0000     1.0000    1.0000    1.0000
EMTAKEMTAK2                  1.6566     0.6036    1.5984    1.7169
last_anreportUNSUBMITTED     0.6470     1.5455    0.6225    0.6725
claim_current                1.0000     1.0000    1.0000    1.0000
age                          1.0182     0.9822    1.0152    1.0211
tax_qoq                      0.9999     1.0001    0.9998    0.9999
asset_turnover               0.9982     1.0018    0.9965    0.9999
CR                           0.9993     1.0007    0.9986    1.0001

Concordance= 0.647  (se = 0.003 )
Rsquare= 0.103   (max possible= 1 )
Likelihood ratio test= 2458  on 15 df,   p=0
Wald test            = 2272  on 15 df,   p=0
Score (logrank) test = 2392  on 15 df,   p=0
```

```
> survConcordance(Surv(TIME, STATUS) ~ predict(cox.train.2, trainTime),
                                                    trainTime)
Call:
survConcordance(formula = Surv(TIME, STATUS) ~ predict(cox.train.2,
trainTime), data = trainTime)

n= 22572
Concordance= 0.6465187 se= 0.003019216
concordant discordant  tied.risk  tied.time    std(c-d)
90955467.0 49729421.0       339.0  3175259.0   849518.2

> survConcordance(Surv(TIME, STATUS) ~ predict(cox.train.2, testTime),
                                                    testTime)
Call:
survConcordance(formula = Surv(TIME, STATUS) ~ predict(cox.train.2,
testTime), data = testTime)

n= 5938
Concordance= 0.5915388 se= 0.006208237
concordant discordant  tied.risk  tied.time    std(c-d)
5145818.0  3553216.0        23.0   295367.0   108011.6
```

Checking PH assumption

```
> cox.zph(cox.train.2)
                              rho    chisq        p
Sum                       0.03927   12.031 0.000523
sales                    -0.00309    0.316 0.573986
equity                   -0.01693    2.530 0.111702
current_assets            0.00239    0.220 0.638736
unsubdeclars1             0.02131    5.753 0.016462
tax_debt0                -0.00314    0.512 0.474175
tax_debt1m                0.01731    6.741 0.009420
tax_debt2m                0.01759    3.573 0.058734
EMTAKEMTAK2              -0.17180  353.590 0.000000
last_anreportUNSUBMITTED  0.03118   12.061 0.000515
claim_current             0.00305    0.125 0.723291
age                      -0.02054    4.880 0.027162
tax_qoq                  -0.02131    5.698 0.016985
asset_turnover            0.02296    6.824 0.008996
CR                       -0.03539   15.075 0.000103
GLOBAL                         NA  488.248 0.000000
```

Random Survival Forests

```
# 2. Invoice + Company measures + Annual Report + Ratios
> fitform.2.all <- Surv(TIME, STATUS) ~
                        Sum +
                        sales + equity + current_assets +
                        unsubdeclars + tax_debt0 + tax_debt1m + tax_debt2m +
                        EMTAK + last_anreport + claim_current +
                        age + tax_qoq +
                        ratioSum + asset_turnover + CR

> grow.time.2.all <- rfsrc(fitform.2.all, data=as.data.frame(trainTime),
                        forest=TRUE, ntree=250, coerce.factor=
                        c('last_anreport', 'EMTAK', 'unsubdeclars'),
                        nodesize=50, na.action='na.omit', splitrule =
                        'logrank')
> pred.time.2.all <- predict(grow.time.2.all, newdata =
                        as.data.frame(testTime), na.action='na.omit')
```

```
> grow.time.2.all
                         Sample size: 22572
                         Number of deaths: 12575
                         Number of trees: 250
                         Minimum terminal node size: 50
                         Average no. of terminal nodes: 625.608
                         No. of variables tried at each split: 4
                         Total no. of variables: 16
                         Analysis: RSF
                         Family: surv
                         Splitting rule: logrank
                         Error rate: 26.51%

> pred.time.2.all
                         Sample size of test (predict) data: 5938
                         Number of deaths in test data: 3041
                         Number of grow trees: 250
                         Average no. of grow terminal nodes: 625.608
                         Total no. of grow variables: 16
                         Analysis: RSF
                         Family: surv
                         Test set error rate: 36.46%
```

C.2.2 Models with historical payment behaviour

Cox model

```
> # 7. Company measures + History
fitform.7 <- Surv(TIME, STATUS) ~
                unsubdeclars + tax_debt0 + tax_debt1m + tax_debt2m + EMTAK +
                last_anreport + claim_current +
age + tax_qoq +
paid_late_avg1m + ratio_sum_paid_late1m + ratio_sum_OS_late1m
> fit.cox.7 <- coxph(fitform.7, data=trainData.hist)
> cox.train.7 <- stepAIC(fit.cox.7, direction='both', trace=TRUE)
> cox.pred.7 <- predict(cox.train.7, newdata=testData.hist)
> survConcordance(Surv(TIME, STATUS) ~ predict(cox.train.7, trainData.hist),
                                                    trainData.hist)
Call:
survConcordance(formula = Surv(TIME, STATUS) ~ predict(cox.train.7,
trainData.hist), data = trainData.hist)

n= 11948
Concordance= 0.6775722 se= 0.004017601
concordant discordant  tied.risk  tied.time   std(c-d)
29134621.0 13848398.0    59245.0  1531649.0   345853.3
> # Determine concordance
> survConcordance(Surv(TIME, STATUS) ~ predict(cox.train.7, testData.hist),
                                                    testData.hist)
Call:
survConcordance(formula = Surv(TIME, STATUS) ~ predict(cox.train.7,
testData.hist), data = testData.hist)

n= 4233
Concordance= 0.6330447 se= 0.007355821
concordant discordant  tied.risk  tied.time   std(c-d)
2781385.00 1611122.00    5501.00  179244.00   64701.92
> cox.train.7
Call:
coxph(formula = Surv(TIME, STATUS) ~ unsubdeclars + tax_debt1m +
EMTAK + last_anreport + claim_current + age + paid_late_avg1m +
```

```
ratio_sum_paid_late1m + ratio_sum_OS_late1m, data = trainData.hist)
```

	coef	exp(coef)	se(coef)	z	p
unsubdeclars1	−3.57e−01	7.00e−01	6.84e−02	−5.22	1.8e−07
tax_debt1m	−2.52e−06	1.00e+00	9.08e−07	−2.78	0.00542
EMTAKEMTAK2	5.77e−01	1.78e+00	2.48e−02	23.26	< 2e−16
last_anreportUNSUBMITTED	−4.12e−01	6.62e−01	2.42e−02	−17.01	< 2e−16
claim_current	−1.66e−05	1.00e+00	2.02e−06	−8.21	2.2e−16
age	6.78e−03	1.01e+00	1.95e−03	3.47	0.00052
paid_late_avg1m	−2.86e−02	9.72e−01	1.44e−03	−19.91	< 2e−16
ratio_sum_paid_late1m	6.23e−01	1.86e+00	3.40e−02	18.31	< 2e−16
ratio_sum_OS_late1m	−4.96e−02	9.52e−01	3.31e−02	−1.50	0.13324

```
Likelihood ratio test=2148  on 9 df, p=0
n= 11948, number of events= 7492
```

Checking PH assumption

```
> cox.zph(cox.train.7)
```

	rho	chisq	p
unsubdeclars1	0.00625	0.2981	0.585091
tax_debt1m	0.04171	12.5443	0.000397
EMTAKEMTAK2	−0.10804	79.2763	0.000000
last_anreportUNSUBMITTED	0.01995	2.8616	0.090716
claim_current	−0.00385	0.0819	0.774757
age	−0.03801	10.5030	0.001192
paid_late_avg1m	0.14004	191.2844	0.000000
ratio_sum_paid_late1m	−0.02214	4.1031	0.042804
ratio_sum_OS_late1m	−0.01507	1.5986	0.206108
GLOBAL	NA	462.8577	0.000000

Random Survival Forests

```
> grow.5.hist <- rfsrc(fitform.5, data=as.data.frame(trainData.hist),
                       forest=TRUE, ntree=250, coerce.factor=
                       c('last_anreport', 'EMTAK', 'unsubdeclars'),
                       nodesize=50, na.action='na.omit',
                       splitrule = 'logrank')
> pred.5.hist <- predict(grow.5.hist, newdata = as.data.frame(testData.hist),
                       na.action='na.omit')

> grow.5.hist
                       Sample size: 11948
                       Number of deaths: 7492
                       Number of trees: 250
                       Minimum terminal node size: 50
                       Average no. of terminal nodes: 330.924
                       No. of variables tried at each split: 5
                       Total no. of variables: 19
                       Analysis: RSF
                       Family: surv
                       Splitting rule: logrank
                       Error rate: 25.06%

> pred.5.hist
                       Sample size of test (predict) data: 4233
                       Number of deaths in test data: 2242
                       Number of grow trees: 250
                       Average no. of grow terminal nodes: 330.924
                       Total no. of grow variables: 19
                       Analysis: RSF
                       Family: surv
                       Test set error rate: 33.07%
```

D License

Non-exclusive licence to reproduce thesis and make thesis public

I, Janika Smirnov,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques,

   supervised by Imbi Traat and Peep Küngas,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 12.05.2016