Predicting the Debtor Paths in a Multi-state Process Using a Hazard Rate Model

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Quantitative Finance

Author:

Lotte Harrijvan

Student number:

415083

Supervisor:

Dr. H.J.W.G. Kole

Second Assessor:

Dr. A.J. Koning

April 11, 2020

Abstract

In this thesis, we aim to give more insights into the debt collecting process of a collection agency. We perform an in-depth analysis of the path prediction of debtors in this process and provide monthly predictions of their corresponding payments over the next five years. The transitions in this multi-state process are governed by a Markov process and are modeled using a combination of time-to-event models. First, we fit separate intensities to all permitted transitions according to the semi-parametric Cox proportional hazards regression. The impact of fixed individual covariates on the hazard ratio, such as the age of debt and the outstanding amount, are discussed in the duration-independent model. Additionally, the duration-dependent model also discusses the impact of a variable that accumulates the durations in previous states. It turns out that this duration-dependent variable improves the model performance, and we use this model to obtain the transition intensities. Secondly, the individual transition probabilities are estimated with the Aalen-Johansen estimator and put together in monthly transition matrices. These are then used to obtain predictions of the debtor-state variables, indicating the active state of the debtor in each month of our 5-year forecast horizon. Finally, to predict the payments, we use the debtor-state variables as explanatory variables in the linear and logistic regression. It appears that the logistic regression slightly outperforms the linear regression. However, both regressions suggest that the debtor-state variables are very informative predictors. The results show that most payments will be collected within the first year of our 5-year forecast horizon with an average payment rate of 0.11.

Keywords— Debt collecting process, Multi-state model, Survival analysis, Cox proportional hazards regression, Competing risks model, Aalen-Johansen estimator

Acknowledgement

I would like to take this opportunity to thank several people who have supported me during my time of writing this thesis. To start with, I want to thank my thesis supervisor Dr. Erik Kole for his effort, interesting insights, patience, and constructive feedback. Next, I would like to thank Notilyze for the opportunity to expand my skills by writing my thesis at their company. I would also like to thank my company coaches at DirectPay, Colin Nugteren and Youri Koomen, for our brainstorming sessions as well as their sincere interest and enthusiasm for my research. Finally, I want to thank my family, friends, and especially my boyfriend for their continuous support throughout my whole studies, but specifically during this thesis. It is comforting to know that I have such caring people surrounding me.

Contents

1 Introduction

This thesis performs an in-depth analysis of the debt collecting process implemented by a collection agency. The aim is to map the debtor paths in this process and elaborate on this knowledge by using it in the monthly prediction of the payments to the agency. The company in question is DirectPay, part of the Credit Exchange Group, which provides credit management solutions across several industries. DirectPay acts as a middleman; it collects the outstanding debt of customers, called debtors, and renders it back to the original creditor. The original creditors are companies that provide a service or item on invoice. Such a service entails a risk of not receiving the amount due. In such an event, the creditor can choose to retrieve the amount due from the customer themselves or let an external party, such as DirectPay, take over. DirectPay is specialized in collecting outstanding debts of smaller amounts that belong to energy, internet, and mobile providers. Such a provider pools their customers together that do not pay their obligations in time and sell it as a claim to DirectPay. At that moment, DirectPay has the right to pursue the debtors and collect the outstanding debt. They profit from this exchange by buying the claim for a lower amount than the total debt is worth. However, it is not ensured that all debt will be paid back, and therefore this is a risky exchange. In order to get more insights into this debt collecting process, we obtain path predictions and monthly payment predictions over a 5-year forecast horizon.

DirectPay currently does not know the future paths of their debtors in this process. In addition, they also face difficulties in predicting their incoming payment stream. These incoming payments depend on the debt collecting process. Both the payment predictions and the future paths of the debtors are of great value to DirectPay in order to secure funding and support decision making. This includes assessing the effectiveness of the process. Is it worth to summon a debtor if the model indicates no payments? It also allows them to evaluate the resources needed for a batch of debtors or to assess the impact of a change in the debt collecting process. For instance, what happens if the court closes and they cannot summon the debtor?

In the existing literature, there are numerous models for predicting financial time series that we can use for the monthly payment predictions. However, the cash inflow of DirectPay does not feature a clear pattern other than some small seasonal effects corresponding to holiday payment in May and the end of the year bonus in December. Despite that, this pattern is entirely dependent on the payment behavior of the debtors. In addition, DirectPay has information about each individual debtor that describes their well being, age, living situation, level of education, and occupation.

They also have monthly information about the financial transactions involving the debtor and a dummy variable indicating the active state for each month in the process. Hence, we deal with panel data instead of a classical time series because we have observations over time for each individual debtor. In addition to the payment predictions, the objective of this thesis is also to obtain path predictions of the individual debtors. Therefore, we use a time-to-event type of model to investigate the risk of transition in this process and use this information to predict the monthly payments.

We construct a main model that models the moment of payment, which depends on the future path, and the paid sum separately. Before reaching final payment, we can distinguish a number of different states in the debt collecting process. These states represent the actions that DirectPay takes to enforce payment and remind the debtor of his outstanding debt, ranging from friendly letters to legal actions. DirectPay proceeds to the next state if the previous state does not result in a final payment. All these states together constitute a Markov chain in which the probability of transferring to the next state only depends on the current state. In this Markov chain, we model the transitions according to a hazard rate model. This is a time-to-event type of model originating from survival analysis. The events of interest are the transitions in the debt collecting process.

In order to meet certain transition restrictions between the states in this process, the hazard rate model is built out of four competing risks models and a simple two-state survival model. For each possible transition in this system, we fit separate hazards using Cox proportional hazards regressions. We introduce different fixed covariates into these Cox regressions, using the individual data of DirectPay. In addition, we also construct a hazard rate model that includes a duration-dependent variable into the Cox regressions. This duration-dependent variable is equal to the sum of the durations in all previous states. For this reason, we investigate the hazard rate of transition in the debt collecting process using a duration-independent model as well as a duration-dependent model.

The hazard rates are the transition intensities between the states. They are used in the Aalen-Johansen estimator of the transition probabilities. For each debtor individually, we construct monthly transition matrices from these estimated probabilities. With these transition matrices, we can predict the future states of the debtor and construct a debtor-state variable over time. This debtor-state variable indicates the active state of the debtor in each month of our 5-year forecast horizon. By combining all the individual debtor-state variables, we get a monthly overview of the dispersion of the debtors amongst the different states. This provides much insight into the debt collecting process. Additionally, this debtor-state variable is used in the final step of our main

model to predict the monthly payments to DirectPay. We predict the monthly payments with a linear and logistic regression using the individual debtor-state variable as an explanatory variable.

This thesis focuses on modeling the future states of the debtors in the debt collection process and uses this information to predict the payments to DirectPay. This is done with a three-step approach using a hazard rate model to obtain the transition matrices, followed by obtaining the debtor-state variables and, finally, the prediction of the payments itself. The most challenging part of this approach is the estimation of the transition matrices after that linear equations are used to obtain the debtor-state variables and, subsequently, the payments. Firstly, we find that predicting the transition intensities with duration-dependent Cox regressions improves the model performance. Hence, the covariate that accumulates the durations in all previous states is a significant predictor in the Cox regressions. The average hazard ratio of this covariate is equal to 0.952. So an extra month spent in the previous states is associated with a 4.8% decreased risk of transitioning to the next state. Next, the debtor-state variables that indicate the active state, are very informative predictors in the regressions of the payment predictions. They all have a negative effect on the payment rate. This effect is the strongest for state 1 and decreases as we advance through the process. So a debtor in state 1 is less likely to pay than a debtor is state 5. Finally, the payment predictions indicate that most payments will be collected within the first year of our forecast horizon with an average payment rate of 0.11.

Modeling this problem as a time-to-event type of model adds value since it provides us with a monthly overview of the dispersion of the debtors amongst the different states as well as with monthly predictions of the incoming payments. Using Cox regressions to model the transitions in the debt collection process allows us to analyze each individual debtor separately. In this way, we exploit all the available information and obtain adequate predictions on an individual level. The aforementioned is also academically interesting because most research using hazard rate models is done in the medical field in which death is the event of interest. In our knowledge, using a hazard rate model in the context of predicting financial time series has never been done before. Additionally, this paper extends the research into the hazard rate of transition in the debt collecting process by also providing explicit expressions for both the non-parametric and semi-parametric estimation of the transition probabilities.

The remainder of this thesis consists of Section 2, presenting a more detailed explanation about credit management and DirectPay as a business. Followed by an overview of the existing literature in Section 3. Next, the methodology behind survival analysis, together with the explanation of

the hazard rate model and the payment predictions, are discussed in Section 4. Furthermore, an elaborate discussion of the data is presented in Section 5. Then the results are discussed in Section 6. Finally, we conclude in Section 7 along with suggestions for further research.

2 Credit Management and DirectPay

DirectPay is a pioneer in purchasing and managing claims in the Netherlands. It provides services to companies that sell products or subscriptions on invoice, which is a common payment option nowadays. However, selling such a product on invoice entails the risk of not receiving the payment in time. That is where DirectPay comes in to ensure their clients of their payments by billing their debtors on behalf of them. They mainly focus on invoices between 50 and 2000 euros that belong to energy, internet, and mobile service providers. Such invoices are commonly frequent, and the majority of the invoices are less than 500 euros.

In the Netherlands, the consumer is obliged to pay the invoice within 14 days. After these 14 days, the customer is in default, and the company can call in the experts, like DirectPay. DirectPay takes over those unpaid invoices, called a claim, such that the company can continue to focus on its core activities instead of pursuing the unpaid debts. Then DirectPay will start the debt collecting process. The debt collecting process can be divided into three phases: the debt management phase, the summons phase, and the confiscation phase.

The first phase, debt management, consists of friendly reminders to collect the outstanding debt without taking legal actions. The first reminder gives the debtor 14 additional days to pay the outstanding debt without any extra costs. In the unfortunate event that the debtor failed to pay his obligation again, another reminder is sent, and DirectPay is legally allowed to charge the debtor for the extra costs made to collect the outstanding debt. This reminder initiates another 14 days for the debtor to fulfill the claim, giving him approximately 60 days to pay the invoice. When DirectPay does not succeed in receiving the payments within these 60 days, it moves on to the summons phase. In this phase, DirectPay takes legal action and summons the debtor to appear in court. The debtor is informed by a bailiff who delivers the summons document in person. This document contains all the relevant information about the claim, the defendant, and court proceedings. After the judge pronounces a verdict in favor of DirectPay, the confiscation phase starts. In this phase, the debtor can voluntarily pay off the claim, or otherwise, the bailiffs have a whole range of possibilities to collect the outstanding debt. This consists of confiscation on either

tax returns, wages, or personal belongings like cars. The verdict is 20 years valid, giving DirectPay 20 years to redeem the outstanding debt. Going through the debt collecting process is expensive, and the costs are attributed to the debtor, increasing the amount of outstanding debt significantly.

In addition, the three above mentioned phases of the debt collecting process are subdivided into five intermediate states in which an account can be situated in, and are shown in Table 1 together with the final payment state:

Table 1: The debt collecting process.

Debt management phase	Summons phase	Confiscation phase	
1) Amicable	3) Summon	5) Confiscation	6) Final payment
2) Profitletter	5) Verdict		

The three phases of the debt collecting process of DirectPay, divided into five separate states plus the final payment state.

These six states are more informative about the exact position of the debtor in the debt collecting process. DirectPay tracks each individual debtor in this debt collecting process, indicating an active state with 1 and non-active states with 0. The summon, verdict, and final payment states are evident. The amicable state consists of friendly reminders (maximum of 3 letters) to pay the outstanding debt. The profitletter is an extra letter send to debtors that moved to another address to make sure that before they receive a summon, they have been informed friendly. In addition, there are several ways to confiscate the outstanding debt, including tax confiscation, wage confiscation, and a Debtscan, which allows DirectPay to confiscate somebody's car.

During this process, debtors can pay a lump sum of the debt all at once, when economic times are better, or they can enter into a payment arrangement with DirectPay to payoff $x\%$ of the outstanding debt each month These arrangement payments are not considered as a separate state since the debtors remain in their current state until they reach their final payment. Figure 1 displays the debt collecting process together with the possible transitions. This figure is the foundation of our problem and is used to model the transitions as a Markov process. It shows that the final payment state can be reached from every other state, while the other states increasingly follow each other, except for state 2. Only debtors that have moved houses enter this state before they move to state 3.

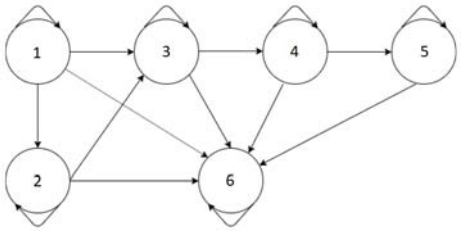


Figure 1: The debt collecting process of DirectPay, with the corresponding permitted transitions.

3 Literature Review

Forecasting financial time series is usually done by implementing time series analysis techniques. Commonly known time series models are the Auto-Regressive (AR) models. They have extensively been used to predict returns assuming that the time series can be modeled with a linear and time-dependent model. However, instead of using this classical time series approach to predict the payments, we use another type of modeling known as time-to-event modeling. Time-to-event modeling is done in survival analysis, which is mostly used in the (bio)medical or epidemiological research field, with the event of interest usually being the death of a particular disease. The basic statistical theory of survival analysis is explained in the book of Cameron and Trivedi (2005).

In addition, DirectPay has a considerable interest in the future paths of the debtors in the debt collecting process. Modeling our problem with a time-to-event model is therefore also relevant, because it provides insights into this dynamic process (Andersen et al., 2002). It produces a transition intensity matrix that contains the hazard rates, or intensities, of transition between states. Alternatively, the survival function can be obtained. To obtain the survival function, non-parametric and (semi)-parametric methods can be used. One of those non-parametric methods is the Kaplan and Meier (1958) estimator for the survival function. However, a more conventional method nowadays is the popular semi-parametric model in survival data, the Cox proportional hazards regression model (Cox, 1972). This model is able to include extra information regarding individual debtors that might influence the transition between the states. This also applies to the parametric estimation of the hazard rates, such as the exponential regression and the Weibull regression. These are the most common parametric approaches according to Rodriguez (2010).

Moreover, Cox proportional hazards models are used in the analysis of univariate survival data. In recent years more research into multivariate survival models has been conducted. These models can deal with more complex survival data, such as clustered survival data. An example of

such a model are the frailty models explained in detail in Duchateau and Janssen (2007). Frailty models extend the Cox proportional hazards model with a multiplicative frailty factor. This factor represents the existence of a random effect and conditional on this random effect the survival functions in the multivariate model are independent. Frailty models provide an excellent way to capture and to describe the dependence of observations within and between clusters. However, fitting a frailty model is way more cumbersome than the Cox proportional hazards model.

In economics the Cox regression models are used to model unemployment duration. For instance, Kavkler et al. (2009) study the impact of gender, education level, region, and the time-dependent variable age on the hazard ratio. They find that the longer the unemployment spell lasts, the less pronounce the differences between various age groups are. This research was done for countries in Eastern Europe. A similar study into the impact of age on unemployment has been conducted for North European countries in the paper of D'Agostino and Mealli (2000). Their results show a negative association between age and reemployment for employees in Denmark, France, and Portugal. Whereas in Italy, the UK and Spain not only the old but also the young have a hard time being reemployed.

Furthermore, in the papers of Andersen (1988); Andersen et al. (2000); Putter et al. (2007); Shu and Klein (2005) they use multi-state Markov models to preserve the underlying relations between the states instead of fitting separate analyses per process step using marginal survival functions. For instance, in Andersen et al. (2000) they compare multi-state models and models for the marginal survival distribution for bleeding episodes and mortality in liver cirrhosis. Their main finding is that a multi-state model is preferable over a model for the marginal survival distribution, because the precision of the survival probability tend to be better. However, there are more assumptions made in the multi-state model, referring to the assumption that the process being modeled is a (semi)-Markov process. These kinds of assumptions are not necessary for the marginal analysis. In addition, Andersen (1988) also compares a multi-state model with a two-state marginal survival model in his study of nephropathy and mortality in diabetes. He concludes that multi-state models provide a flexible framework to study the effects of covariates, and important insights into the dynamic process may be gained.

When a multi-state model satisfies the Markov assumption, the probability of a transition to the next state only depends on the current state and not on the process history. According to Putter et al. (2007) only 'clock forward' models can be Markov models. In a 'clock forward' model time t refers to the running time since entering the initial state. Alternatively, the 'clock reset' models

refer to semi-Markov models, in which time t resets every time a debtor moves to a new state. Hence, the time scale depends on the history, and thus this model does not satisfy the Markov assumption.

Furthermore, Andersen et al. (2002) presents an overview of different multi-state models in survival analysis, representing the most common types such as the two-state survival model, the competing risks model, and the illness-death model. According to Putter et al. (2007), competing risks models always satisfy the Markov assumption since there is no event history, and additionally, the same holds for the two-state survival model.

In many multi-state studies for modeling the process of a particular disease, the transition intensities are modeled by the Cox proportional hazards Markov regression model (Andersen, 1988; Andersen et al., 2000; Klein et al., 1993; Shu and Klein, 2005). Alternatively, the Aalen additive hazards regression model has also been used in the literature of semi-parametric estimation (Andersen et al., 1991; Shu and Klein, 2005; Aalen et al., 2001). Then, the estimated cumulative transition intensities are used in the Aalen and Johansen (1978) estimator of the transition probabilities. In their paper, they developed the asymptotic properties of this estimator using stochastic integrals and martingales. The solution is represented by the product integral, of which its properties are explained in Slavík (2007). The estimates of the transition probabilities usually have explicit expressions, and the derivation of these expressions in this paper are based on the work of Andersen et al. (2002), Andersen et al. (1993) and Shu and Klein (2005).

The proposed research contributes to the literature by implementing a hazard rate model in the prediction of a financial time series, which has mostly been used in (bio)medical research. It is an innovative approach of predicting the payments, and it also provides us with the future states of the debtors in the debt collecting process. It supports the decision making on whether they should or should not summon a debtor based on the expected payments or use the information to adapt their way of collecting the outstanding debt. Besides, this thesis also provides a clear overview of the explicit expressions for both the non-parametric and semi-parametric estimation of the transition probabilities in a competing risks model and two-state survival model.

4 Methodology

The debt collecting process is a multi-state Markov model with finite number of states. DirectPay tracks the transitions of the individual debtors in this process, indicating an active state with 1

and non-active states with 0. They are interested in a model that can predict the future paths of the debtors together with the monthly payments to DirectPay for the next five years. To predict the future path of a debtor, we need to know the probabilities of transitioning from one state to another. We can model these probabilities with a time-to-event type of model using the theory of survival analysis. Survival analysis data consists of individual data observed over time with respect to the occurrence of a particular event. The event of interests are the transitions between the states in the debt collecting process. We model these individual transitions with a hazard rate model and construct monthly transition matrices from the estimated transition probabilities. These transition matrices are used to predict the time-dependent debtor-state variable indicating the active state of each month in our 5-year forecast horizon. These predicted debtor-state variables are, in turn, necessary to predict the monthly payments. Hence, we introduce the three-step approach:

1. Model the monthly transition matrices according to the theory of survival analysis.
2. Predict the monthly debtor-state variables with the corresponding transition matrices.
3. Predict the monthly payments based on the debtor-state variables.

This approach is performed for all debtors individually.

The following sections elaborate extensively on this three-step approach and the theory behind survival analysis. Section 4.1 starts with the general definitions of the survival and hazard functions. Next, we model the transition matrix by a hazard rate model explained in 4.2. Followed by Section 4.3, which discusses the prediction of the debtor-state variables. Subsequently, Section 4.4 elaborates on the prediction of the monthly payments.

4.1 The hazard and survival function

First, we explain some basic statistical concepts of survival analysis according to Cameron and Trivedi (2005), in order to understand the estimation of the transition matrix. Survival models are time-to-event models, denoting T as the survival time starting from a given time origin (cession date of the debt purchase) until the occurrence of a certain event. Since survival analysis has mostly been used in the medical world, the event of interest is usually death of a certain disease. However, in this specific case, the event of interest is a transition to another state in the debt collecting process. Assume that T is a non-negative continuous random variable with probability

density function $f(t)$ and cumulative distribution function:

$$F(t) = \Pr[T \leq t] = \int_0^t f(x)dx \tag{1}$$

This represents the probability that the event has occurred by time t. Naturally, the survival function is the probability that the event of interest has not occurred by time t and is given by:

$$S(t) = \Pr[T > t] = 1 - F(t) \tag{2}$$

Along with this survival function, we are also interested in the rate at which the event takes place. This is denoted by the hazard function, or the instantaneous risk of the event happening at time t given that the event did not happen prior to time t. Therefore, the hazard function is given by:

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr[t \leq T < t + dt | T \geq t]}{dt} \tag{3}$$

where the numerator is the conditional probability that an event will happen in a small time interval dt given that the event has not happened up till time t. Dividing by the length of this interval will give us the rate of event occurrence per unit of time. Hence, note that the hazard rate is a measure of risk and not a probability because it can exceed 1.

A convenient way of writing the expression of the hazard function is in terms of the pdf and the survival function:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{4}$$

From Equation 2 we can see that the derivative of $S(t)$ is equal to $-f(t)$. Therefore we can write Equation 4 as the change in log-survival function:

$$\lambda(t) = -\frac{d}{dt} \log S(t) \tag{5}$$

Integrating the hazard function and using the fact that the survival probability at time 0 is equal to 1, the survival function can be written as:

$$S(t) = \exp\left(-\int_0^t \lambda(x)dx\right) \tag{6}$$

The integral in this expression is the final related function in survival analysis and is known as the

cumulative hazard function:

$$\Lambda(t) = \int_0^t \lambda(x)dx = -\log S(t) \tag{7}$$

4.1.1 Discrete data

In survival analysis it is very common to deal with discrete data and therefore we also illustrate the discrete-time hazard models. Discrete data arises when the underlying process generating transitions is discrete or when the data are observed discretely. The risk of transition at discrete time t_j, $j = 1, 2, \ldots,$ given we survived up till time t_j is equal to the discrete-time hazard function that is defined as:

$$\begin{aligned} \lambda_j &= \Pr[T = t_j | T \geq t_j] \\ &= \frac{f^d(t_j)}{S^d(t_{j-})} \end{aligned} \tag{8}$$

where the subscript d indicates discrete. Additionally, the discrete-time survivor function is equal to:

$$\begin{aligned} S^d(t) &= \Pr[T \geq t] \\ &= \prod_{j|t_j \leq t} (1 - \lambda_j) \end{aligned} \tag{9}$$

Finally, the discrete-time cumulative hazard function is:

$$\Lambda^d(t) = \sum_{j|t_j \leq t} \lambda_j \tag{10}$$

4.1.2 Censoring

Survival data usually suffers from incomplete measurements, because the event of interest does not occur in the observation time or people are lost to follow-up. This common problem of incomplete data in survival analysis is called censoring. Even when we do not know when the event will happen, we still include the data of this specific individual to estimate the survival probability. If we would exclude this information, underestimation can become a problem. Including censored data is the power of survival analysis.

In our analysis, we only deal with right censoring, which means that the transition to the next state has not occurred in the observation time. However, for survival analysis to be valid, the

censoring mechanism should be non-informative or independent. Non-informative or independent censoring implies that the risk of future events is the same for censored and uncensored individuals. This means that censored observations should not be informative about the parameters of the survival function. If this condition is met, non-informative censoring results in the same likelihood function. In other words, observations observed after independent right-censoring are also representative of the group of individuals without censoring (Andersen et al., 2002). To obtain the likelihood for individual i (observed for time t_j) we analyze the contribution of both uncensored and censored observations to the likelihood. In the case of uncensored observations, the event has occurred at observation time t_j. Therefore, the contribution to the likelihood is the density at time t_j:

$$L_i = f(t_j) = S(t_j)\lambda(t_j) \tag{11}$$

However, for the uncensored case, the individual has not experienced the event yet at time t_j. If we assume independent censoring, the only information we have about these censored observations is that the survival time, T, exceeds t_j. The contribution of this censored observation to the likelihood is the survival function itself:

$$L_i = f(t_j) = S(t_j) \tag{12}$$

Equation 11 and 12 differ in the multiplication of the hazard function since this indicates whether the event has occurred or not. These two contributions can be combined into one expression for the total likelihood function:

$$L = \prod_{i=1}^{n} L_i = \prod_{i} \lambda(t_j)^{d_i} S(t_j) \tag{13}$$

where d_i is a dummy variable indicating the occurrence of an event.

4.1.3 Non-parametric estimation

In this section we introduce the non-parametric estimators of the hazard, cumulative hazard and survival functions for the discrete case. In this non-parametric estimation of survival functions no regressors are included. Let N be the number of individuals and $0 < t_1 < t_2 < \cdots < t_j < \cdots < t_N$ are the discrete failure times, then the estimator of the hazard function is equal to:

$$\hat{\lambda}_j = \frac{d_j}{n_j} \tag{14}$$

where n_j is the number of individuals at risk (not experienced the event yet) just prior to time t_j and d_j is the number of events occurred at time t_j. The discrete-time cumulative hazard function defined in 10 can be estimated with the sample equivalent known as the Nelson-Aalen estimator:

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \hat{\lambda}_j = \sum_{j:t_j \leq t} \frac{d_j}{n_j} \tag{15}$$

Additionally, the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function is the sample equivalent of the discrete-time survival function in 9:

$$\hat{S}(t) = \prod_{j:t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} \tag{16}$$

This non-parametric estimator estimates the survival function of a homogeneous sample of individuals with a counting process. A counting process is an integer-valued stochastic process counting the events at associated event times.

4.1.4 Semi-parametric estimation

DirectPay has information about the individual debtors that might impact the transitions from one state to another. However, the Kaplan-Meier estimator is not able to incorporate multiple predictors. Hence, we want to include this individual information by using a regression model for the hazard function that relates these several risk factors to survival time. We do so by using the most popular regression technique in survival analysis: the semi-parametric Cox proportional hazards regression model (Cox, 1972). This model allows the prediction of the hazard rate tailored to the individual debtor. The Cox proportional hazards regression can be written as follows:

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}) \tag{17}$$

where $\lambda(t|\boldsymbol{X})$ are the conditional hazard rates, $\lambda_0(t)$ is the baseline hazard function corresponding to the individual risk when all covariates are equal to 0, $\boldsymbol{X} = (X_1, \ldots, X_P)$ are the individual fixed covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\exp(\boldsymbol{\beta}^T\boldsymbol{X})$ is the relative risk. This approach is semi-parametric because we make no assumption about the shape of the baseline hazard function. The representation in 17 shows the multiplicative effect of the covariates on the baseline hazard function. Besides, the expression is separated in a clear time effect captured in the baseline hazard and a constant covariates effect. The hazard rates of different individuals can be compared using

the hazard ratio:

$$HR = \frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T \boldsymbol{X_i})}{\lambda_0(t)\exp(\boldsymbol{\beta}^T \boldsymbol{X_j})} = \exp(\boldsymbol{\beta}^T(\boldsymbol{X_i} - \boldsymbol{X_j})) \tag{18}$$

This ratio is a constant proportion that does not depend on time. Hence, the Cox regression model is a proportional hazards model. Additionally, we want to allow the baseline hazard to be different across strata, $h = 1, \ldots, m$, for which the Cox regression is defined as:

$$\lambda_h(t|\boldsymbol{X}) = \lambda_{h,0}(t)\exp(\boldsymbol{\beta}^T \boldsymbol{X}) \tag{19}$$

An advantage of this semi-parametric model is that we do not have to specify the baseline hazard. Specifically, Cox (1972) derived an estimation method for $\boldsymbol{\beta}$, where the specification of this baseline hazard is not necessary. The solution for $\boldsymbol{\beta}$ in this stratified Cox model is obtained by maximizing the partial likelihood per stratum

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{h=1}^{m} \mathrm{L}_h(\boldsymbol{\beta}) \tag{20}$$

with

$$\mathrm{L}_h(\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_j)}{\sum_{l \in R_{hj}} \exp(\boldsymbol{\beta^T} \boldsymbol{X}_l)}$$

where R_{hj} includes all the individuals at risk in stratum h at time t_j. For each event time in stratum h, the fraction in this expression compares the hazard of the individual experiencing the event at time t_j to the hazard of all the individuals at risk at time t_j. From this equation, we can see that the baseline hazard function drops out of the expression. The parameter estimate $\hat{\boldsymbol{\beta}}$ is used in the estimation of the baseline cumulative hazard, known as the Breslow estimator:

$$\hat{\Lambda}_0(t) = \sum_{j:t_j \leq t} \frac{1}{\sum_{l \in R_{hj}} \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{X_l})} \tag{21}$$

However, by imposing assumptions about the parametric form of the baseline hazard, we can obtain parametric regressions models such as the exponential and Weibull regression proportional hazard models (Rodriguez, 2010). When the assumption about the form of the distribution is correct, we can make adequate inference about the parameters. However, when the distributional assumptions cannot be verified, researchers often resort to the semi-parametric Cox regression model to prevent inconsistent parameter estimates.

An extension of the Cox model can be achieved by relaxing the fixed covariates assumption:

$$\lambda_h(t|\boldsymbol{X}(t)) = \lambda_{h,0}(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}(t)) \tag{22}$$

where $\boldsymbol{X}(t)$ is a vector of time-varying covariates. Incorporating time-varying covariates complicates the estimation of the Cox regression since the simple separation of the time effect and covariates effect is lost. Therefore, it is difficult to determine the effects of the covariates on survival time since they could also be highly correlated with time. Define $\boldsymbol{X_t}$ as the history generated by the covariates in time interval [0,t). To satisfy the Markov assumption, only the current value of $\boldsymbol{X}(t)$, instead of the entire process history of $\boldsymbol{X_t}$, should be considered in Equation 22. Then the estimation of the Cox regression is similar to the estimation with fixed covariates, according to Cameron and Trivedi (2005).

4.1.5 Discrete-time proportional hazards

In this section, the discrete-time variant of the proportional hazards model is presented, assuming we observe the failure times at aggregated time intervals. The discrete-time hazard function is denoted as:

$$\lambda^d(t_a|\boldsymbol{x}) = \Pr(t_{a-1} \leq T < t_a | T \geq t_{a-1}, \boldsymbol{x}(t_{a-1})), \quad a = 1, \cdots, A \tag{23}$$

where T is a discrete random variable with grouping points $t_a, a = 1, \cdots, A$. The corresponding discrete-time survival function is defined as:

$$S^d(t_a|\boldsymbol{x}) = \Pr(T \geq t_{a-1}|\boldsymbol{x}) = \prod_{s=1}^{a-1}(1 - \lambda^d(t_s|\boldsymbol{x}(t_s))) \tag{24}$$

The discrete-time hazard can be written as a function of the survival probability because it is the probability that the event occurs within time interval $[t_{a-1}, t_a)$ divided by the survival probability up till time t_{a-1}:

$$\lambda^d(t_a|\boldsymbol{x}) = \frac{S(t_{a-1}|\boldsymbol{x}) - S(t_a|\boldsymbol{x})}{S(t_{a-1}|\boldsymbol{x})} = 1 - \exp\left(-\int_{t_{a-1}}^{t_a}\lambda(s)ds\right) \tag{25}$$

where we use the continuous-time definition of the survival function (Eq. 6) in the last step of the derivation.

The discrete-time proportional hazards model for time interval $[t_{a-1}, t_a)$ is defined as:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t_{a-1})) \tag{26}$$

Notice that the covariates are fixed within the interval, but can vary across intervals. The baseline hazard is still a function of time and can, therefore, vary within the interval. By implementing Equation 26 into the discrete-time hazard function of 25 we obtain:

$$\begin{aligned}\lambda^d(t_a|\boldsymbol{x}) &= 1 - \exp\left(-\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t_{a-1})) \cdot \int_{t_{a-1}}^{t_a} \lambda_0(s)ds\right) \\ &= 1 - \exp\left(-\lambda_{0a}\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t_{a-1}))\right) \\ &= 1 - \exp\left(-\exp(\ln\lambda_{0a} + \boldsymbol{\beta}^T\boldsymbol{x}(t_{a-1}))\right)\end{aligned} \tag{27}$$

where $\lambda_{0a} = \int_{t_{a-1}}^{t_a} \lambda_0(s)ds$. The discrete-time survival function becomes:

$$S^d(t_a|\boldsymbol{x}) = \prod_{s=1}^{a-1} \exp\left(-\exp(\ln\lambda_{0s} + \boldsymbol{\beta}^T\boldsymbol{x}(t_{s-1}))\right) \tag{28}$$

Since the density function of individual i is the product of the survival function in each period multiplied by the hazard function in case the event has occurred, the likelihood function boils down to a combination of Equation 27 and Equation 28:

$$\begin{aligned}L(\boldsymbol{\beta}, \lambda_{01}, \cdots, \lambda_{0A}) = \prod_{i=1}^{N} &\left(\prod_{s=1}^{a_i-1} \exp\left(-\exp(\ln\lambda_{0s} + \boldsymbol{\beta}^T\boldsymbol{x}_i(t_{s-1}))\right)\right) \\ &\times \left(1 - \exp\left(-\exp(\ln\lambda_{0a_i} + \boldsymbol{\beta}^T\boldsymbol{x}_i(t_{a-1}))\right)\right)\end{aligned} \tag{29}$$

4.2 Modeling the transition matrix using a hazard rate model

In the following section, we model the individual transitions of the debtor through various states in the debt collecting process with a hazard rate model. The debt collecting process is a stochastic multi-state process in which the individual debtors move independently between the six states. Define the multi-state process as $\Gamma(t)$ for $t \geq 0$, taking the values $S = 1, \ldots, 6$. A multi-state process generates a history $\mathcal{X}_{t-}$ consisting of the observation of the process in the interval $[0, t)$. It describes the evolution of the process over time t, including the previously visited states and

the time of transitions. In addition to the process history, the history of the covariates may also influence the transition probabilities. We extend the history for fixed covariates X according to $\mathcal{F}_t = \mathcal{X}_t \vee X_0$ and for time-dependent covariates X_t according to $\mathcal{F}_t = \mathcal{X}_t \vee X_t$, where X_t is the history generated by the covariates in time interval $[0, t)$. Relative to the (extended) history, the transition probabilities are defined by $p_{hj}(s,t) = \Pr(\Gamma(t) = j | \Gamma(s) = h, \mathcal{F}_{s-})$. However, we assume that our multi-state process satisfies the Markov assumption that future evolution of the process only depends on the current state at time t and not on the process history. In other words, the current state at time t summarizes the history of the process. Therefore, the transition probabilities in this multi-state model are independent of $\mathcal{F}_{s-}$ and are derived according to:

$$p_{hj}(s,t) = \Pr(\Gamma(t) = j | \Gamma(s) = h), \quad s \leq t \tag{30}$$

We can predict the future states of our multi-state model by varying time t, given the present at time s. For Markov models, these probabilities will only depend on the current state at time s (Putter et al., 2007).

The corresponding transition intensities are denoted as:

$$\lambda_{hj}(t) = \lim_{dt \to 0} \frac{p_{hj}(t, t+dt)}{dt} \tag{31}$$

Transition intensities provide the instantaneous hazard of progression to state j conditionally on current state h. In other words, the transition intensities in multi-state models are equal to the hazard rate (Eqution 3). The intensities are not constant but rather a function of time. Hence, this is a non-homogeneous Markov process, for which we track the debtor's migration with a transition matrix such as the one below:

$$\mathbf{P}(t) = \begin{bmatrix} 1 - p_{12} - p_{13} - p_{16} & p_{12} & p_{13} & 0 & 0 & p_{16} \\ 0 & 1 - p_{23} - p_{26} & p_{23} & 0 & 0 & p_{26} \\ 0 & 0 & 1 - p_{34} - p_{36} & p_{34} & 0 & p_{36} \\ 0 & 0 & 0 & 1 - p_{45} - p_{46} & p_{45} & p_{46} \\ 0 & 0 & 0 & 0 & 1 - p_{56} & p_{56} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where each row sums up to one. This transition matrix gives an overview of the possible transitions

in the debt collecting process corresponding to the transitions in Figure 1. It clearly indicates that only transitions to the following states are allowed, together with the direct transitions to state 6. State 6 corresponds to the final payment state and can be reached from any other state. Besides, a direct transition from state 1 to state 3 is also possible, because not every debtor will enter state 2. Only debtors that have changed address do so.

In order to satisfy the imposed transitions, we can subdivide our main multi-state model into five smaller multi-state models. Each multi-state model corresponds to one of the first five rows in $\mathbf{P}(t)$. For simplicity, we define module A as the multi-state model corresponding to row 1; module B to row 2; module C to row 3; module D to row 4; and module E to row 5. These multi-state modules are independent of each other since we condition on the modules in the previous row(s). We model these multi-state modules with the theory of survival analysis, using two different time-to-event models. Modules A-D have two or three mutually exclusive exists conditional on the transitions in the previous module(s). Hence, we model each of these four modules separately according to a competing risks model (CRM). In addition, row number five has only one possible exit, and therefore module E is modeled using a simple two-state survival model. State 6 is the final payment state, and once reached, the debtor leaves the debt collecting process. It is not possible to transfer from state 6, and therefore it is an absorbing state with $p_{66} = 1$.

In both the CRMs and the survival model, we use the Cox-like proportional hazards function to model the hazard rate of transition. Specifically, modules A-E all correspond to a unique Cox regression in which separate hazards are fitted to all permitted transitions. These hazard rates are the transition intensities for each transition represented at every point in time. Then the predicted intensities are used to estimate the transition probabilities, based on their dependence through the Kolmogorov forward differential equations. The solution is expressed as the matrix product integral (Gill and Johansen, 1990).

In the following sections, we elaborate on the theory of CRMs (Section 4.2.1) and derive explicit expressions of the transition probabilities in the CRM and two-state survival model (Section 4.2.2).

4.2.1 Competing risks model

In this section, we discuss the theoretical background of the CRM in a continuous-time framework. The CRM consists of an initial state and multiple absorbing states. These absorbing states are called the competing states and are mutually exclusive, implying that a transition to one of these states excludes the risk of transitioning to the other states. Hence, there is only one possible

transition out of the initial state. For this reason, CRMs always satisfy the Markov assumption since there is no event history.

The CRM models the time to the event together with the type of *first* event. Hence, the main interest of a CRM is to provide a joint distribution of the duration denoted as τ and the type of exit r. The competition between the states in the CRM results in some sort of censoring since we only observe one of the transitions. If we would have m competing states, each event provides one complete duration and $m-1$ censored durations (Cameron and Trivedi, 2005). The total number of different exits is m, implying that r will be equal to an integer value in the set $J = (1, 2, \cdots, m)$. Since there is only one transition made from the initial state, we only observe the first transition or the shortest duration, and the rest is censored:

$$\begin{aligned} \tau &= \min(t_1, \ldots, t_m) \\ &= \min_j(t_j), \quad t_j > 0 \end{aligned} \tag{32}$$

The joint survival function is then defined as:

$$\begin{aligned} S_\tau(t) &= \Pr(\tau > t) = \Pr(t_1 > t, \ldots, t_m > t) \\ &= \Pr(t_1 > t) \cdot \Pr(t_2 > t) \cdot \cdots \cdot \Pr(t_m > t) \end{aligned} \tag{33}$$

where the last derivation results from the assumption of independent risks. The associated exit route is given by:

$$r = \arg\min_{j \in J}(t_j) \tag{34}$$

The hazard function, under the independent risks assumption, for transition $j \in J$ is denoted as:

$$\lambda_j(t_j|\boldsymbol{x}_j) = \lim_{dt_j \to 0} \frac{\Pr(t_j \leq T \leq t_j + dt | T \geq t_j, \boldsymbol{x}_j)}{dt_j} \tag{35}$$

where $\boldsymbol{x}_j$ are transition specific covariates that may affect the hazard rate. The associated cumulative hazard function for the jth transition is:

$$\Lambda_j(t_j|\boldsymbol{x}_j) = \int_0^{t_j} \lambda_j(s|\boldsymbol{x}_j) ds \tag{36}$$

The density of the duration is the hazard function multiplied with the survival function, using the

relation in 4, 6 and 7, this results in:

$$\begin{aligned} f_j(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}_j) &= \lambda_j(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}_j) S_j(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}_j) \\ &= \lambda_j(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}_j) \exp(-\Lambda_j(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}_j)) \end{aligned} \tag{37}$$

To obtain the joint density of τ and r, define $\boldsymbol{x} = (x_1, \ldots, x_m)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$ and use the law of conditional probabilities:

$$\begin{aligned} f_j(\tau, r|\boldsymbol{x}, \boldsymbol{\beta}) &= f_r(\tau|\boldsymbol{x}_r, \boldsymbol{\beta}_r) \prod_{j \neq r} \exp(-\Lambda_j(\tau|\boldsymbol{x}_j, \boldsymbol{\beta}_j)) \\ &= \lambda_r(\tau|\boldsymbol{x}_r, \boldsymbol{\beta}_r) \exp(-\Lambda_r(\tau|\boldsymbol{x}_r, \boldsymbol{\beta}_r)) \prod_{j \neq r} \exp(-\Lambda_j(\tau|\boldsymbol{x}_j, \boldsymbol{\beta}_j)) \\ &= \lambda_r(\tau|\boldsymbol{x}_r, \boldsymbol{\beta}_r) \prod_{j=1}^{m} \exp(-\Lambda_j(\tau|\boldsymbol{x}_j, \boldsymbol{\beta}_j)) \end{aligned} \tag{38}$$

The product in the first line represents the multiplication of all the survival probabilities of the alternative exits and results from the independence of risk assumption. The hazard rates of these alternative exists are not included since we only transfer to exit r, and therefore we only include the hazard rate corresponding to exit r in this expression.

In Section 4.2, we mentioned that the hazard rates in the CRM are modeled with a Cox regression in order to include covariates. This results in the Cox CRM regression defined as:

$$\lambda_j(t|\boldsymbol{X}(t)) = \lambda_{0j}(t) \exp(\boldsymbol{\beta}_j^T \boldsymbol{X}(t)) \quad j = 1, \ldots, m \tag{39}$$

specific for type j hazard and $t_{j1} < \cdots < t_{jk_j}$ where k_j is the number of individuals who transitioned to state $j \in J$. Given this hazard function, the likelihood function of the Cox CRM is equal to:

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = \prod_{j=1}^{m} \prod_{i=1}^{k_j} \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_{ji}(t_{ji}))}{\sum_{l \in R(t_{ji})} \exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_l(t_{ji}))} = \prod_{j=1}^{m} L_j(\boldsymbol{\beta}_j) \tag{40}$$

because of independent risks, the joint maximization of this expression is equal to maximizing each $L_j(\boldsymbol{\beta}_j)$.

4.2.2 From transition intensities to transition probabilities

The transition intensity represents the expected number of events per time unit and is, therefore, not a probability because it can exceed 1. For this reason, we need to transform these intensities of our multi-state Markov model into transition probabilities. These probabilities are dependent on the transition intensities through the solution of the Kolmogorov differential equations. This solution represents the powerful relationship between the two as the product integral, resulting in a matrix of transition probabilities $p_{hj}(s,t)$ given by:

$$P(s,t) = \prod_{(s,t]} \{I + d\Lambda(u)\} \tag{41}$$

where $\prod_{(s,t]}$ is the product integral (see Appendix A.1), I is the identity matrix and $\Lambda(u) = \{\Lambda_{hj}(u)\}$ is the cumulative transition intensity matrix. Aalen and Johansen (1978) suggested to substitute the Nelson-Aalen estimators of the cumulative transition intensities into formula 41, to obtain the matrix of estimated transition probabilities. This results in the Aalen-Johansen estimator:

$$\hat{P}(s,t) = \prod_{(s,t]} \{I + d\hat{\Lambda}(u)\} \tag{42}$$

where $\hat{\Lambda}(u) = \{\hat{\Lambda}_{hj}(u)\}$ is the matrix of Nelson-Aalen estimators with diagonal elements $\hat{\Lambda}_{hh}(u) = -\sum_{j \neq h} \hat{\Lambda}_{hj}(u)$. For our multi-state problem the estimated cumulative transition intensity matrix takes the form:

$$\hat{\mathbf{\Lambda}}(t) = \begin{bmatrix} -\hat{\Lambda}_{12} - \hat{\Lambda}_{13} - \hat{\Lambda}_{16} & \hat{\Lambda}_{12} & \hat{\Lambda}_{13} & 0 & 0 & \hat{\Lambda}_{16} \\ 0 & -\hat{\Lambda}_{23} - \hat{\Lambda}_{26} & \hat{\Lambda}_{23} & 0 & 0 & \hat{\Lambda}_{26} \\ 0 & 0 & -\hat{\Lambda}_{34} - \hat{\Lambda}_{36} & \hat{\Lambda}_{34} & 0 & \hat{\Lambda}_{36} \\ 0 & 0 & 0 & -\hat{\Lambda}_{45} - \hat{\Lambda}_{46} & \hat{\Lambda}_{45} & \hat{\Lambda}_{46} \\ 0 & 0 & 0 & 0 & -\hat{\Lambda}_{56} & \hat{\Lambda}_{56} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The product integral in 42 reduces to the regular product in the discrete case (Equation 9) and to the exponential of the integral in the continuous case (Equation 6). For discrete-time transition intensities the product integral in 42, therefore, changes into a product. Let $s < T_1 < T_2 < \cdots < T_m \leq t$ be the times of observed transitions between s and t. Then we obtain the estimated

transition matrix according to:

$$\hat{P}(s,t) = \prod_{i=1}^{m}(I + d\hat{\Lambda}(T_i)) \tag{43}$$

This non-parametric estimator is a matrix version of the Kaplan-Meier estimator used for multi-state models. It does not consider individual covariates that may influence the transitions and, therefore, assumes that all individuals have the same transition intensities, $\lambda_{hji} = \lambda_{hj}$. Hence, the transition probabilities can be estimated with a multivariate counting process (Aalen, 1978; Aalen and Johansen, 1978). The multivariate counting process is defined as $N(t) = \{N_{hj}(t); h \neq j\}$, with $N_{hj}(t)$ counting the number of observed direct transitions from state h to state j in [0, t]. This counting process has intensity process $\lambda(t) = \{\lambda_{hj}(t); h \neq j\}$.

However, in our main multi-state model the goal is to estimate the transition matrix with a semi-parametric approach in which we do include the covariates of the individuals. The covariates enter the model through the Cox proportional hazards regression for the individual transition intensities, λ_{hji}. The Aalen-Johansen estimator can be generalised for this semi-parametric estimation of the transition probabilities by replacing $\hat{\Lambda}(u)$ in Equation 42 with $\hat{\Lambda}(u; \boldsymbol{X}_0)$, where $\boldsymbol{X}_0$ is a vector of fixed covariates. This results in the semi-parametric estimation of the transition matrix denoted as:

$$\hat{P}(s,t;\boldsymbol{X}_0) = \prod_{(s,t]}\{I + d\hat{\Lambda}(u;\boldsymbol{X}_0)\} \tag{44}$$

where $\hat{\Lambda}_{hh}(u;\boldsymbol{X}_0) = -\sum_{j\neq h}\hat{\Lambda}_{hj}(u;\boldsymbol{X}_0)$ and $\hat{\Lambda}_{hj}(u;\boldsymbol{X}_0) = \hat{\Lambda}_{hj0}(t)\exp(\hat{\boldsymbol{\beta}}_{hj}^T\boldsymbol{X}_0)$ for $h \neq j$. This latter equation is the extension of the Cox proportional hazards regression to multi-state models, in which the transition intensities are defined as:

$$\lambda_{hj}(t|\boldsymbol{X}) = \lambda_{hj,0}(t)\exp(\boldsymbol{\beta}_{hj}^T\boldsymbol{X}) \tag{45}$$

where $\boldsymbol{X}$ is the vector of covariates, $\boldsymbol{\beta}_{hj}$ is the vector of regression coefficients for transition $h \to j$ and $\lambda_{hj,0}$ is the baseline hazard function of transition $h \to j$. This notation implies that separate baseline hazards are used for each transition, together with transition specific covariates (Fiocco et al., 2008). For simple multi-state models the continuous semi-parametric estimators of $\hat{p}_{hj}(s,t;\boldsymbol{X}_0)$ have explicit expressions in terms of λ_{hj} using integrals because the product integral in 44 reduces to the exponential of the integral.

We derive explicit expressions for both the non-parametric and semi-parametric estimates of

the transition probabilities in the simple two-state model and the CRMs. These derivations are based on the work of Andersen et al. (2002), Andersen et al. (1993) and Shu and Klein (2005).

Simple two-state survival model (module E)

Module E is just a simple two-state survival model and is depicted in Figure 2 together with its corresponding transition matrix. It is the most simple type of multi-state models, with one transient state (state 5) and one absorbing state (state 6). Hence, there is only one possible transition. This means that there is no event history and module E satisfies the Markov assumption. In this module, the probability of staying in state 5 is equal to the survival function $S(t)$. So naturally, the probability of transitioning to state 6 is equal to $1 - S(t) = F(t)$. This cumulative distribution is known as the cumulative incidence function in survival analysis. It originates from the medical world in which incidence denotes the occurrence of a disease. Regarding our problem, incidence indicates the occurrence of a transition, and therefore the cumulative incidence function represents the probability of transitioning from state 5 to state 6.

$$\mathbf{P} = \begin{array}{c} \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cc} 5 & 6 \end{array} \\ \begin{bmatrix} p_{55} & p_{56} \\ 0 & 1 \end{bmatrix} \end{array}$$

Figure 2: The simple two-state survival model with corresponding transition matrix.

For the two-state survival model, the expression between the brackets in 43 is given by:

$$I + d\hat{\Lambda}(T_i) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -d\hat{\Lambda}_{56} & d\hat{\Lambda}_{56} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 - d\hat{\Lambda}_{56} & d\hat{\Lambda}_{56} \\ 0 & 1 \end{bmatrix}$$

Therefore, the non-parametric transition probability estimates are found by plugging in the multi-state representation of the Nelson-Aalen estimators, $\hat{\Lambda}_{hj}(s,t)$, of Equation 15:

$$\hat{\Lambda}_{hj}(s,t) = \sum_{s \leq t} \frac{\Delta N_{hj}(s)}{Y_h(s)}, \quad h \neq j \tag{46}$$

where $\Delta N_{hj}(s)$ is the number of transitions made from state h to state j at time s and $Y_h(s)$ is the number of individuals at risk (in state h) just prior to time s. Hence, the non-parametric estimates are obtained by a counting process:

$$\begin{aligned}\hat{p}_{55}(s,t) &= \prod_{i=1}^{m}(1 - d\hat{\Lambda}_{56}) \\ &= \prod_{i=1}^{m}\left(1 - \frac{\Delta N_{56}(T_i)}{Y_5(T_i)}\right) \\ &= \hat{S}_0(s,t)\end{aligned} \tag{47}$$

This expression is equal to the Kaplan-Meier survivor estimator. The transition probability from state 5 to state 6 is estimated according to:

$$\hat{p}_{56}(s,t) = \sum_{i=1}^{m}\left\{\prod_{h=i+1}^{m}\left(1 - \frac{\Delta N_{56}(T_h)}{Y_5(T_h)}\right)\right\}\frac{\Delta N_{56}(T_i)}{Y_5(T_i)} \tag{48}$$

Additionally, for the semi-parametric estimation we know that $\Lambda_{56}(t) = \int_0^t \lambda_{56}(u)du$ and the probability to stay in the initial state is, again, equal to the survival function:

$$p_{55}(s,t) = \exp\left(-\Lambda_{56}(s,t)\right) = \exp\left(-\int_s^t \lambda_{56}(u)du\right) = S(t) \tag{49}$$

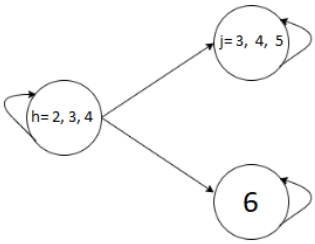
It is no surprise that the probability of staying in state 5 is equal to the survival probability because remaining in state 5 implies that the event of interest has not yet occurred at time t. Additionally, the probability of transitioning from state 5 to 6 is equal to the cumulative incidence function:

$$\begin{aligned}p_{56}(s,t) &= \int_s^t \exp\left\{-\int_u^t \lambda_{56}(v)dv\right\}\lambda_{56}(u)du \\ &= \int_s^t S(u-)\lambda_{56}(u)du \\ &= \int_s^t p_{55}(u,t)\lambda_{56}(u)du\end{aligned} \tag{50}$$

The competing risks multi-state model (modules A-D)

Modules A-D are independently modeled according to a CRM. The CRM is a special case of a multi-state model that extends the two-state survival model with several extra exits. In our main multi-state model, $\Gamma(t)$, with transition matrix $\mathbf{P}(t)$ there are various transition restrictions imposed. Hence, using a CRM for each of these modules is a convenient way to force specific transition

probabilities to be zero. For each of these CRMs, there is one transient state and two or three competing states or exits. Between these states, there is competition to determine the path of a debtor in the debt collecting process. Only one transition is made from the initial state. This means that there is no event history and modules A-D all satisfies the Markov assumption. Additionally, we assume independence of risk between the competing states. We also require different baseline hazards for each transition within a module, and therefore we fit separate intensities to all permitted transitions. Modules B-D each have two competing states; either there is a transition to the following state in the debt collecting process, or there is a direct transition to state 6. The module with corresponding transition matrix is illustrated in Figure 3.



$$\mathbf{P} = \begin{array}{c} \\ \text{h} \\ \text{j} \\ 6 \end{array} \begin{array}{c} \begin{array}{ccc} \text{h} & \text{j} & 6 \end{array} \\ \begin{bmatrix} p_{hh} & p_{hj} & p_{h6} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

Figure 3: The competing risks multi-state model for module B-D with corresponding transition matrix.

Let us define $H = \{2, 3, 4\}$ and $L = \{3, 4, 5\}$, then $J = \{l \in L, 6\}$ represents the competing states in the CRMs for module B-D. The permitted transitions per module are $\Omega = \{(23, 26), (34, 36), (45, 46)\}$ and therefore the expression between the brackets in 43 is given by:

$$I + d\hat{\Lambda}(T_i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -d\hat{\Lambda}_{hj} - d\hat{\Lambda}_{h6} & d\hat{\Lambda}_{hj} & d\hat{\Lambda}_{h6} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 - d\hat{\Lambda}_{hj} - d\hat{\Lambda}_{h6} & d\hat{\Lambda}_{hj} & d\hat{\Lambda}_{h6} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Hence, the discrete non-parametric estimation of the transition probabilities results in:

$$\begin{aligned} \hat{p}_{hh}(s,t) &= \prod_{i=1}^{m} (1 - d\hat{\Lambda}_{hj} - d\hat{\Lambda}_{h6}) \\ &= \prod_{i=1}^{m} \left(1 - \frac{\Delta N_{hj}(T_i)}{Y_h(T_i)} - \frac{\Delta N_{h6}(T_i)}{Y_h(T_i)}\right) \\ &= \hat{S}_0(s,t) \end{aligned} \tag{51}$$

$$\hat{p}_{hj}(s,t) = \sum_{i=1}^{m} \left\{ \prod_{h<i} \left(1 - \frac{\Delta N_{hj}(T_h)}{Y_h(T_h)} - \frac{\Delta N_{h6}(T_h)}{Y_h(T_h)} \right) \right\} \frac{\Delta N_{hj}(T_i)}{Y_h(T_i)} \tag{52}$$

and

$$\hat{p}_{h6}(s,t) = \sum_{i=1}^{m} \left\{ \prod_{h<i} \left(1 - \frac{\Delta N_{hj}(T_h)}{Y_h(T_h)} - \frac{\Delta N_{h6}(T_h)}{Y_h(T_h)} \right) \right\} \frac{\Delta N_{h6}(T_i)}{Y_h(T_i)} \tag{53}$$

Alternatively, the continuous semi-parametric estimation of the transition probabilities is equal to:

$$p_{hh}(s,t) = \exp\left(-(\Lambda_{hj}(s,t) + \Lambda_{h6}(s,t))\right) = \exp\left(- \int_s^t (\lambda_{hj}(u) + \lambda_{h6}(u)) du \right) = S(t) \tag{54}$$

$$p_{hj}(s,t) = \int_s^t S(u-)\lambda_{hj}(u)du = \int_s^t p_{hh}(u,t)\lambda_{hj}(u)du \tag{55}$$

$$p_{h6}(s,t) = \int_s^t S(u-)\lambda_{h6}(u)du = \int_s^t p_{hh}(u,t)\lambda_{h6}(u)du \tag{56}$$

Note, that similar to the two-state survival model, the probability of staying in the initial state is equal to the survival function, and the probabilities of transitioning to one of the competing states are equal to the cumulative incidence function.

The CRM in module A has three competing states. A debtor can either go from state 1 to state 2, from state 1 directly to state 3 when there is no address change or from state 1 straight to state 6. Let us define $J = \{2, 3, 6\}$ as the competing states with permitted transitions $\Omega = \{(12, 13, 16))\}$, then the semi-parametric transition probabilities are defined as:

$$\begin{aligned} p_{11}(s,t) &= \exp\left(-(\Lambda_{12}(s,t) + \Lambda_{13}(s,t) + \Lambda_{16}(s,t))\right) \\ &= \exp\left(- \int_s^t (\lambda_{12}(u) + \lambda_{13}(u) + \lambda_{16}(u)) du \right) = S(t) \end{aligned} \tag{57}$$

$$p_{12}(s,t) = \int_s^t S(u-)\lambda_{12}(u)du = \int_s^t p_{11}(u,t)\lambda_{12}(u)du \tag{58}$$

$$p_{13}(s,t) = \int_s^t S(u-)\lambda_{13}(u)du = \int_s^t p_{11}(u,t)\lambda_{13}(u)du \tag{59}$$

$$p_{16}(s,t) = \int_s^t S(u-)\lambda_{16}(u)du = \int_s^t p_{11}(u,t)\lambda_{16}(u)du \tag{60}$$

4.3 Prediction of the debtor-state variables

In this section, we discuss the monthly prediction of the debtor-state variable. This debtor-state variable is a dummy variable indicating the active state of an individual debtor with 1. The prediction is rather straightforward because it is just a simple multiplication between the current

debtor-state variable and the estimated transition matrix of the next month. The transition matrix is not constant over time since the probabilities in this matrix are obtained from the intensities modeled by a Cox regression. This contains the baseline hazard, which is a function of time. Hence, the transition matrix is also a function of time.

To predict the debtor-state variable for the next five years with $M = 1, \cdots, 60$, we need to find the distribution over the states. This can be written as a stochastic row vector x with the relation:

$$\begin{aligned} x_{t+1} &= x_t \cdot \hat{P}_{t+1} \\ x_{t+2} &= x_{t+1} \cdot \hat{P}_{t+2} \\ &\vdots \\ x_{t+m} &= x_{t+m-1} \cdot \hat{P}_{t+m} \end{aligned} \tag{61}$$

where $\hat{P}_{t+m}$ is the estimated transition matrix corresponding to time $t + m$. At each point in time, this results in a vector of probabilities. One way to determine the debtor's location is to assign 1 to the state with the highest probability and 0 to the others. However, this prediction is somewhat deterministic because the model will probably assign the highest probability to the nearest state since this is observed in the data (Table 4). Such a prediction resembles a "winner takes it all" prediction. Hence, we calculate the cumulative sum of these probabilities. To add some randomness, we sample a random number between $(0, 1)$ and insert the value in the sequence of partial sums. The first state that has a larger cumulative value than this random number is the state the debtor transfers to.

For example, lets say that at time $t = 3$ a debtor is situated in state 1. We calculate the debtor-state variable for time $t = 4$ according to $x_4 = x_3 \cdot \hat{P}_4$ and obtain the probability vector $[0.63, 0.15, 0.09, 0.00, 0.00, 0.13]$. The cumulative probability vector is then equal to $[0.63, 0.78, 0.87, 0.87, 0.87, 1.00]$. Now, insert the sampled random number, for example 0.82, in this sequence of partial sums. The first highest cumulative value is observed for state 3, indicating that the debtor transfers from state 1 to state 3. Figure 4 gives a visual representation of this procedure.

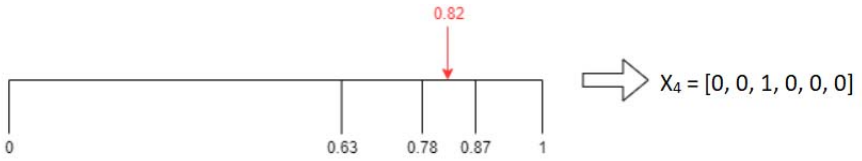


Figure 4: The procedure of determining the debtor-state variable by generating a random number and inserting this into the cumulative probability vector.

4.4 Predicting the payments

The next step is to predict the monthly payments of each debtor. The dependent variable in this prediction is the percentage of payments relative to the original claim amount of the individual's outstanding debt. We predict the payment rate with a simple linear regression as well as with a logistic regression. This latter regression method is used since we are predicting a percentage.

We expect that the monthly predicted debtor-state variable in the previous step is very informative, and we include this as explanatory variable in both regressions. We also include the covariates used in the Cox regressions as explanatory variables. The linear regression is given by:

$$y_t = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}_t + \epsilon_t \tag{62}$$

Alternatively, the logistic regression assumes a linear relationship between the log-odds and the explanatory variables. The logistic function maps the real numbers to the $(0, 1)$ interval:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{63}$$

However, our dependent variable, the payment rate, is not bounded between $(0, 1)$, because some of the costs made in the debt collecting process are charged on the debtor. Therefore, the payment rate can exceed 1. In order to do the logistic regression, we have to transform our dependent variable to make sure it is bounded between $(0, 1)$. This transformation is done according to $\frac{y-a}{b-a}$, where a is the lower bound and b the upper bound of the payment rate. The logistic regression is then given by:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}_t = z \tag{64}$$

resulting in

$$p = \frac{e^z}{1 + e^z} \tag{65}$$

Afterward, we transform the dependent variable back to its original form.

5 Data

DirectPay has different types of clients for which they confiscate the outstanding debt of its customers. These clients belong to different branches, as shown in Figure 5. Most of their clients belong to the telecom branch, followed by the internet, and the energy branch. We focus on implementing the multi-state model for the telecom branch.

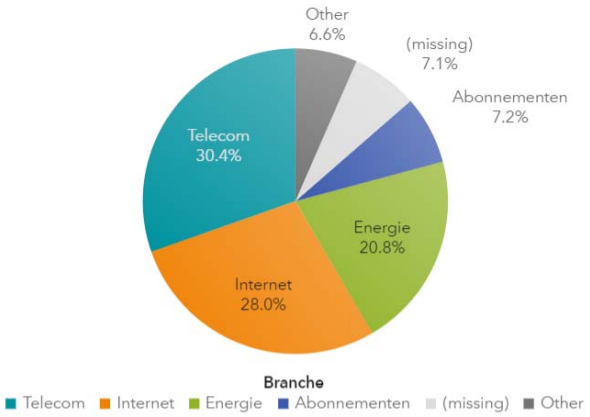


Figure 5: The different branches in which DirectPay operates.

In order to obtain our data set, three original data sets of DirectPay are merged based on debtor ID. One of the data sets contains all forward flow purchases and all related transactions about collections and expenses. It contains the most important financial transactions involving a debtor, including the purchase amount of the claim, the costs regarding court fees and summons, and the payments obtained from the debtors. The other two data sets contain the debtor-state variable indicating the active state of the debtors in the debt collecting process and individual information regarding the debtors. In total, we have 157,400 debtors, observed over a time period from January 2008 until November 2019. Around 37,180 of these debtors have already reached the final payment state. For the other 120,220 we will predict their path in the debt collecting process.

The data is grouped by debtor ID, of which the observations are ordered by time using a long data format. For each debtor, these observations consist of some fixed covariates, the time-dependent debtor-state variable indicating the active state, and a duration-dependent variable that is equal to the sum of durations in all previous states. This duration-dependent variable is fixed within each module, but differs amongst the modules. Considering we estimate the Cox regressions

for each module separately, the fixed covariates and the "fixed" duration-dependent variable are both used in the Cox regressions to determine the transition intensities. The fixed covariates are the original claim amount of the outstanding debt (*"Claim amount"*), the debtor's age at time of acquiring (*"Age debtor"*), the age of the outstanding debt at time of acquiring (*"Age debt"*), a credit score (*"Credit score"*), a WOZ-score [1] (*"WOZ score"*) and an income score (*"Income score"*). The latter three covariates are all based on their own scoring system within DirectPay. The credit score is between 0 and 1, with a higher credit score indicating a better valuation of the debtor. The WOZ score is categorized into thirteen groups, where a high group implies a higher house value. Additionally, the income score is split into five groups based on the average income in the Netherlands. Again, a higher score indicates a higher income. A more detailed explanation of the respective scores is represented in Table 14 and Table 15 in the Appendix A.3.

The summary statistics of the covariates used in the Cox regressions are displayed in Table 2. The claims that are bought by DirectPay range from €10 - €13,300, and are on average, approximately €500. The average age of the debtor at time of acquiring is equal to 37. Additionally, the debtors that have reached final payment spent, on average, 17 months in the whole process. The maximum stay of debtors in the process is almost equal to 11 years.

Table 2: Summary statistics of the covariates.

	Mean	St. dev.	Median	Min	Max
Claim amount (in euros)	507	473	370	10	13,300
Age debtor (in years)	37	13	35	18	100
Age debt (in months)	7	6	5	1	123
Credit score	0.77	0.13	0.78	0.12	0.97
WOZ score	4	2.66	4	1	13
Income score	2	1.31	2	1	5
Duration*	17	21	10	1	134

Summary statistics of the covariates used in the Cox regressions, based on 157,400 debtors over the period $2008 - 2019$.
*Summary statistics based on debtors (N = 37,180) that have gone through the whole debt collecting process and thus reached final payment.

The summary statistics of the covariates WOZ score and the income score are difficult to interpret because these are ordinal variables indicating the social-economic status of the debtors. Ordinal data is categorical data with a natural ordering. The mean or median is difficult to interpret, and instead, we display the distributions of the score groups in Figures 6 to 8. These figures are

[1] WOZ = Waardering Onroerende Zaken, the Dutch real-estate valuation.

more informative and present an overview of how the debtors are distributed over the different score groups. It should be noted that a large amount of the debtors fall into the lower groups, and this confirms our expectation that most of the debtors are coming from a population with low social-economic status. Specifically, most debtors have an income score equal to 1, indicating that the average debtor has an income lower than the average income in the Netherlands. Besides, most debtors have a WOZ score equal to 3, and this suggests that the house value of the average debtor is between €125,000 - €150,000.

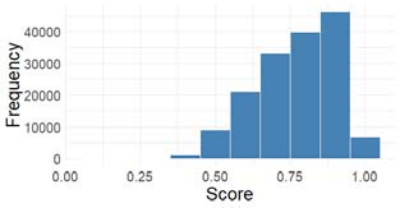


Figure 6: Score distribution

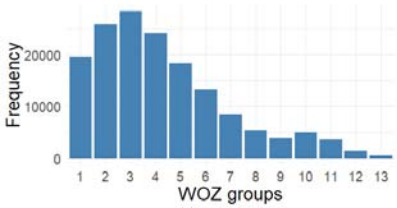


Figure 7: WOZ distribution

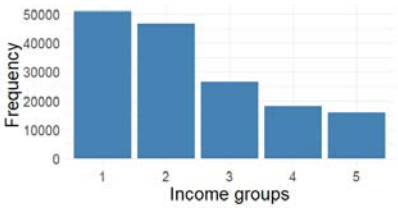


Figure 8: Income distribution

In Table 3, we present the correlations between the score groups in order to find out their dependency. The score groups are all positively correlated, with the highest correlation of 0.67 between the WOZ score and the income score. This confirms our expectation that these two score groups are strongly related to each other because both score groups give an indication of the social-economic status of the debtors.

Table 3: Correlation matrix.

	Credit score	WOZ score	Income score
Credit score	1		
WOZ score	0.63	1	
Income score	0.54	0.67	1

Correlation matrix of the score group variables.

Data preparations

In order to work with the data set provided by DirectPay, we need to make some alterations. Table 12 in the Appendix A.2 gives a clear overview of the data preparation process. The first step in prepping the data is to exclude all debtors that have an original claim amount smaller or equal to €10.00 since these debtors will (almost) not provide any income for DirectPay. Next, we have to add a time variable for each of the debtors. This time variable keeps track of the number of months

that have passed since entering the system. It allows us to calculate the time spent in previous states, which is used as a duration-dependent covariate in the Cox regressions.

A useful way to summarize multi-state data is by a frequency table counting the number of times each pair of states were observed in successive observation times. For our data, this results in Table 13 in the Appendix. This table suggests that backward transitions are also possible. However, in our multi-state problem, we assume that once a state is reached, the debtor cannot go back to previous states. The same holds for some direct transitions that are not allowed according to our formulation of the multi-state problem in Section 4.2. The total contribution of these non-valid transitions in our data set is 5%. Hence, we remove these transitions and therefore change for example a $1 \to 3 \to 2 \to 6$ path into a $1 \to 3 \to 6$ path. In addition, we delete debtors with entering state different than state 1, because this is not representative.

The semi-parametric estimation of individual transition intensities can not deal with missing values. Therefore, we also eliminate the debtors that have missing values for some of the covariates. After this last step in the data preparation process, we have a data set containing approximately 157,400 debtors with corresponding frequency table presented in Table 4

Table 4: Frequency table of the transitions in the original data set.

	1	2	3	4	5	6	no event	total entering
1	0	31,548	19,817	0	0	22,656	0	74,021
2	0	0	13,242	0	0	3,209	0	16,451
3	0	0	0	23,333	0	3,172	0	26,505
4	0	0	0	0	16,047	1,548	0	17,622
5	0	0	0	0	0	4,371	0	4,371
6	0	0	0	0	0	0	0	0

(a) The frequency numbers

	1	2	3	4	5	6	no event
1	0	0.426	0.268	0	0	0.306	0
2	0	0	0.805	0	0	0.195	0
3	0	0	0	0.880	0	0.120	0
4	0	0	0	0	0.912	0.088	0
5	0	0	0	0	0	1.000	0
6	0	0	0	0	0	0	1.000

(b) The frequency fractions

Frequency table of the total number of transitions for each pair of states in successive observation times, based on the period from January 2008 to November 2019. Subtable (b) presents the respective frequency fractions.

The high frequency fractions corresponding to the successive states indicate that in our original data set most transitions are made to the following state. Besides, a relative large proportion (30.6%) of the debtors situated in state 1 make a direct transition to state 6. In addition, Table 5 summarizes the possible paths in the debt collecting process together with a bar plot in Figure 9 counting the number of times a particular path occurs in our final data set. This figure illustrates that most debtors are still situated in state 1 and that the most common path is $1 \rightarrow 6$.

	Paths
1	1
2	$1 \rightarrow 2$
3	$1 \rightarrow 2 \rightarrow 3$
4	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4$
5	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$
6	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$
7	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 6$
8	$1 \rightarrow 2 \rightarrow 3 \rightarrow 6$
9	$1 \rightarrow 2 \rightarrow 6$
10	$1 \rightarrow 3$
11	$1 \rightarrow 3 \rightarrow 4$
12	$1 \rightarrow 3 \rightarrow 4 \rightarrow 5$
13	$1 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$
14	$1 \rightarrow 3 \rightarrow 4 \rightarrow 6$
15	$1 \rightarrow 3 \rightarrow 6$
16	$1 \rightarrow 6$

Table 5: All possible paths in the debt collecting process.

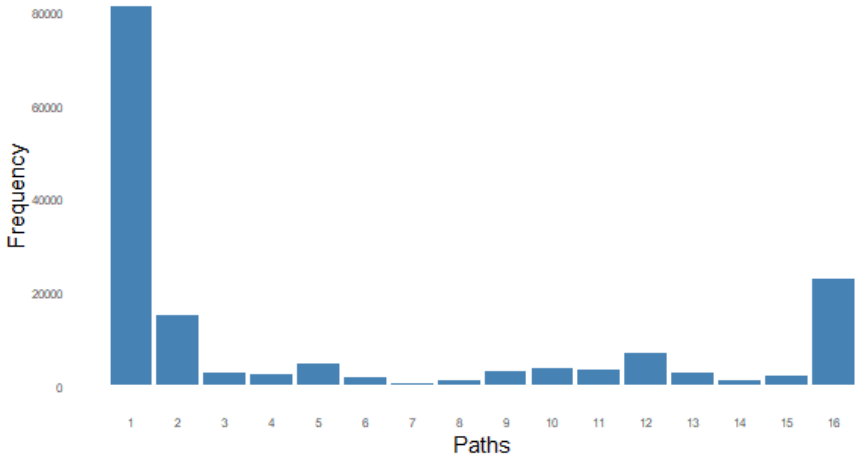


Figure 9: The number of times a particular path occurs in the original data set.

6 Results

In this section, we discuss the results of our analysis. First, we discuss the estimation results of two different Cox regression models. Secondly, we asses the performance of these two models and select the best performing one. Next, we resume discussing the results of the best performing model. Finally, we discuss the results of the predicted payments. Again, the estimation results of the two different regression models are discussed first before we elaborate on the results of the best performing model.

6.1 Estimation results of the Cox regressions

In order to estimate the transition probabilities in our main multi-state model, we used the separate modules A-E to predict the hazard rates of the transitions. These five modules are combined to obtain one transition matrix at every point in time for each individual debtor. Debtors that are currently situated in state 1 will have to go through all five modules. In contrast, debtors currently situated in state 2 will only have to go through modules B-E, and debtors in state 3 only need to go through modules C-E, etcetera. Each module corresponds with a unique Cox regression and combining them results in the main model. In our analysis, we investigate two different main models, the duration-independent model, and the duration-dependent model. Whether the model is duration-independent or duration-dependent is determined by the covariates used in the Cox regressions. In particular, the duration-dependent model includes the duration-dependent covariate that accumulates the durations in all previous states.

DirectPay has all kinds of variables that are used to describe the individual debtor based on their well being, living situation, level of education, occupation, and age. Selecting the final covariates used in the model was done in co-operation with DirectPay and are shown in Table 2. These covariates sum up all the critical information about an individual debtor. At first, only the fixed covariates are used in our Cox regressions to predict the transition intensities resulting in the duration-independent model. Next, we also add the duration-dependent covariate into our Cox regressions. Adding this variable was done to incorporate the belief that when a debtor already spent years in the process, it is less likely that it will pay back the debt. While the probability of a debtor paying back the debt that just entered the model is higher. In this duration-dependent model, we also included the same fixed covariates as in the duration-independent model.

First, we discuss the results of the Cox regressions of each of these two models. They are presented in Table 6 and 7. For regression purposes, we add transition specific covariates to see the effect of certain covariates on different transitions in more detail. Table 16 in the Appendix gives an overview of the ten possible transitions indicated with a transition number and we use them in our discussion of the results. In the Cox proportional hazards regression model, the estimated coefficients represent the change in the expected log of the hazard ratio relative to a one unit change in the specific covariate, holding all other covariates constant. In our model, all the covariates are continuous variables, therefore to interpret the results, it is useful to exponentiate the parameter estimate to get the hazard ratio for transitioning to another state relative to a one unit change in

the corresponding covariate. Table 6 shows the estimated coefficients for each Cox regression in the duration-independent model, together with their corresponding hazard ratio ($\exp(\beta)$). A hazard ratio of approximately one implies no effect of the covariate on the hazard. In contrast, a hazard ratio smaller than one implies a reduction in hazard, and a hazard ratio larger than one implies an increase in hazard.

First of all, we want to highlight the overall significance of each Cox regression reflected in the high values of the Wald test statistics. The Wald test evaluates the null hypothesis that all beta's in the regression are equal to zero. Considering the highly significant Wald test statistics, we reject this null hypothesis at the 95% significance level. The second feature to note in the Cox model results is the sign of the regression coefficients. A positive sign means that the hazard of transition to another state increases with a higher estimate. A negative sign, on the other hand, indicates a decreased risk of transitioning. In the duration-independent model, we can see that the sign of the coefficients for a certain covariate differs amongst the transitions. In general, each covariate is significant for at least one of the transitions in every module, except for the income score in module B and C and for the age of debtor, credit score and WOZ score in module E. Another result is that the coefficients in module B are generally smaller than the coefficients in the other modules. This could be due to the fact that the debtors in module B are a different kind of debtors. These debtors have changed addresses and probably just forgot to pay because they did not receive the friendly reminders at first.

Furthermore, discussing the estimation results in more detail, the sign of the WOZ score and income score are usually the same within the modules. This is no surprise, considering that these two variables are positively correlated with a relatively high correlation equal to 0.67 (Table 3). Additionally, the age of the debtor has a hazard ratio of approximately one for each transition. This indicates that the age of the debtor barely has any influence on the hazard of transition. Moreover, the credit score seems to have the most substantial effect in each module and is significant for transition 1, 3, 4, 6, 7, 8, and 9. Generally, when the credit score gets better, the transition to state 6 becomes more likely compared to the other transition(s) in each module. This is reflected in high hazard ratios for the direct transitions to state 6. More specifically, in module A the credit score has a significant negative effect on transition $1 \rightarrow 2$ with a parameter estimate of -0.309 relative to a one unit change in credit score. The exponent of the parameter estimates is equal to 0.734. So a better credit score is associated with a 26.6% decrease in expected hazard. On the contrary, in that same module, the credit score has a significant positive effect on transition $1 \rightarrow 6$, with a

hazard ratio equal to 2.129. Hence, when the credit score gets better, the transition to the final payment state 6 is almost three times ($2.129/0.734 \approx 3$) more likely to happen than the transition to state 2.

The coefficients of the original claim amount for direct transitions to state 6 are always significant and negative. The most negative coefficient belongs to transition 3 and is equal to -0.149. The corresponding hazard ratio is 0.862, meaning that a one unit increase in claim amount is associated with a 13.8% decreased risk for transitioning from state 1 to state 6. In other words, the higher the original claim amount of a debtor, the less likely he directly transits to state 6. This decreased risk becomes smaller as we advance through the process. Out of all the covariates, the covariate age of debt at time of acquiring is most often significant. More precisely, the coefficient is significant for eight out of the ten transitions. All these eight significant coefficients are positive, meaning that a one unit increase in the age of debt will increase the hazard rate of transition to the next state. The highest significant coefficient is related to transition 9, with a hazard ratio equal to 1.026. Therefore, a one unit increase in the age of debt is associated with a 2.6% increase in the expected hazard.

Table 6: Coefficients of the Cox regressions in the duration-independent model.

	Module A					
Transition	$1 \to 2$		$1 \to 3$		$1 \to 6$	
	coef	exp(coef)	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	0.004* (0.001)	1.004*	-0.022*(0.002)	0.978*	-0.149* (0.002)	0.862*
$\beta_{\text{AgeDebtor}}$	-0.002* (0.000)	0.998*	0.000 (0.001)	1.000	0.001* (0.001)	1.001*
β_{AgeDebt}	0.018* (0.001)	1.018*	0.016* (0.001)	1.016*	0.012* (0.001)	1.012*
$\beta_{\text{CreditScore}}$	-0.309* (0.057)	0.734	0.004 (0.072)	1.004	0.755* (0.070)	2.129*
β_{WOZscore}	0.007* (0.003)	1.007*	-0.019* (0.004)	0.982*	0.016* (0.004)	1.016*
$\beta_{\text{IncomeScore}}$	0.025* (0.006)	1.026*	-0.031* (0.008)	0.969*	0.033* (0.007)	1.034*
Wald statistic	5678*					
Concordance	0.596					

(a) The parameter estimates of the Cox regression in module A, regarding transition 1, 2, and 3.

Table 6: Coefficients of the Cox regressions in the duration-independent model.

	Module B			
Transition	$2 \rightarrow 3$		$2 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	0.002 (0.002)	1.002	-0.081* (0.005)	0.922*
$\beta_{\text{AgeDebtor}}$	-0.004* (0.001)	0.996*	0.000 (0.001)	1.000
β_{AgeDebt}	0.008* (0.001)	1.008*	0.004 (0.003)	1.004
$\beta_{\text{CreditScore}}$	-0.231* (0.085)	0.794*	0.125 (0.177)	1.133
β_{WOZscore}	0.004 (0.005)	1.004	0.026* (0.010)	1.026*
$\beta_{\text{IncomeScore}}$	0.005 (0.009)	1.005	0.016 (0.018)	1.016
Wald statistic	340.600 *			
Concordance	0.552			

(b) The parameter estimates of the Cox regression in module B, regarding transition 4 and 5.

	Module C			
Transition	$63 \rightarrow 4$		$3 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	-0.002 (0.002)	0.998	-0.113* (0.006)	0.893*
$\beta_{\text{AgeDebtor}}$	-0.002* (0.001)	0.998*	0.000 (0.001)	1.000
β_{AgeDebt}	-0.001 (0.001)	0.999*	0.012* (0.003)	1.012*
$\beta_{\text{CreditScore}}$	0.305* (0.067)	1.357*	1.193* (0.187)	3.296*
β_{WOZscore}	-0.017* (0.004)	0.984*	-0.003 (0.010)	0.997
$\beta_{\text{IncomeScore}}$	-0.009 (0.007)	0.991	0.029 (0.019)	1.029
Wald statistic	529.600*			
Concordance	0.543			

(c) The parameter estimates of the Cox regression in module C, regarding transition 6 and 7.

	Module D			
Transition	$4 \rightarrow 5$		$4 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	0.003** (0.003)	1.003 **	-0.061* (0.008)	0.941*
$\beta_{\text{AgeDebtor}}$	-0.002* (0.001)	0.998*	0.007* (0.002)	1.007*
β_{AgeDebt}	0.014* (0.003)	1.014*	0.026* (0.004)	1.026*
$\beta_{\text{CreditScore}}$	-0.207* (0.080)	0.813*	0.808* (0.270)	2.243*
β_{WOZscore}	-0.001 (0.005)	0.999	0.034* (0.015)	1.035*
$\beta_{\text{IncomeScore}}$	0.027* (0.008)	1.028*	0.063* (0.027)	1.065*
Wald statistic	363.900*			
Concordance	0.544			

(d) The parameter estimates of the Cox regression in module D, regarding transition 8 and 9.

Table 6: Coefficients of the Cox regressions in the duration-independent model.

	Module E	
Transition	$5 \rightarrow 6$	
	coef	exp(coef)
$\beta_{\text{OriginalClaimAmount}}$	-0.034* (0.004)	0.967*
$\beta_{\text{AgeDebtor}}$	0.001 (0.001)	1.001
β_{AgeDebt}	0.025* (0.002)	1.025*
$\beta_{\text{CreditScore}}$	0.237 (0.161)	1.268
β_{WOZscore}	0.003 (0.009)	1.003
$\beta_{\text{IncomeScore}}$	0.045* (0.016)	1.046*
Wald statistic	202.200*	
Concordance	0.570	

(e) The parameter estimates of the Cox regression in module E, regarding transition 10.

The parameter estimates of the covariates in each Cox regression in module A-E, with their standard errors between brackets.
*p <0.05, ** p <0.10

Next, we evaluate the estimated coefficients of each Cox regression in the duration-dependent model (Table 7). Again, we notice the overall significance of each Cox regression, based on the high values of the Wald statistics. Additionally, the predictive ability of the duration-dependent model is better than the duration-independent model because the concordance statistics of the modules are higher. The concordance statistic is the fraction of concordant pairs, where a pair of observations is concordant if the prediction and the data go in the same direction. It is the most used goodness-of-fit measure in survival models (Harrell Jr et al., 1996). The difference between these Cox regressions and the ones in the duration-independent model is the addition of the *"DurationPreviousStates"* covariate in modules B to E. Module A does not include this variable because there are no previous states. Hence, the Cox regression of module A is identical to the Cox regression of module A in the duration-independent model (and not mentioned in Table 7). The covariate for the total duration in the previous states is significant for each transition in all four modules. For all these corresponding transitions, the coefficient is negative, meaning that a one unit increase in total duration in the previous states is associated with a decreased risk of transitioning. In other words, the more time a debtor has spent in the previous states, the less likely he will transfer to the next state. The average coefficient is equal to -0.049, which corresponds to an average hazard ratio equal to $\exp(-0.049) = 0.952$. Hence, if the debtor spends an extra month in one of the previous states his expected hazard to transfer to the next state decreases with 4.8%.

In the duration-dependent model, the signs of the coefficients for a certain covariate still differs amongst the transitions. In general, each covariate is significant for at least one of the transitions in every module except for the income score in module B, and for the age of debtor, credit score, WOZ score and income score in module E. This almost coincide with the duration-independent model. Furthermore, the credit score has, again, the largest effect in each module and is significant for transition 1, 3, 4, 6, 7, 8, and 9. These are the exact same significant credit score coefficients as in the duration-independent model. Again, a better credit score ensures that direct transitions to state 6 become more likely compared to the other transition(s) in each module. This is reflected in higher hazard ratios for the transition to state 6 than the other hazard ratios. For instance, in module C, when the debtor has a better credit score, the transition to final payment state 6 is almost two times ($2.914/1.324 \approx 2$) more likely than the transition to state 4.

Again, the age of the debtor has almost no influence on the hazard because the hazard ratios are approximately one for all the transitions. In addition, the coefficients of the original claim amount for direct transitions to state 6 are always significant and negative. So the higher the original claim amount of a debtor, the less likely the direct transition to state 6. This decreased risk becomes smaller as we advance through the process. Moreover, the most often significant covariate is the age of debt at time of acquiring. This corresponds with the most often significant covariate in the previous model. The coefficient of the age of debt is significant for nine out of the ten transitions of which the five positive values originate from module A and D. The most significant negative coefficient corresponds to transition 10, with an estimate equal to -0.029 and a hazard ratio of 0.971. Therefore, a one unit increase in the age of debt is associated with a 2.9% decrease in the expected hazard.

Table 7: Coefficients of the Cox regressions in the duration-dependent model.

	Module B			
Transition	$2 \rightarrow 3$		$2 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	0.008* (0.002)	1.008*	-0.077* (0.005)	0.926*
$\beta_{\text{AgeDebtor}}$	-0.002* (0.001)	0.998*	0.001 (0.001)	1.001
β_{AgeDebt}	-0.011* (0.002)	0.989*	-0.023* (0.003)	0.977*
$\beta_{\text{CreditScore}}$	-0.253* (0.085)	0.777*	0.068 (0.177)	1.070
β_{WOZscore}	0.004 (0.005)	1.004	0.022* (0.010)	1.022*
$\beta_{\text{IncomeScore}}$	-0.010 (0.009)	0.990	0.003 (0.018)	1.003
$\beta_{\text{DurationPreviousStates}}$	-0.050* (0.001)	0.951*	-0.054* (0.002)	0.948*
Wald statistic	3739.000*			
Concordance	0.698			

(a) The parameter estimates of the Cox regression in module B, regarding transition 4 and 5.

	Module C			
Transition	$3 \rightarrow 4$		$3 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	-0.001 (0.002)	0.999	-0.111* (0.006)	0.895*
$\beta_{\text{AgeDebtor}}$	-0.001* (0.001)	0.999*	0.000 (0.001)	1.000
β_{AgeDebt}	-0.008* (0.001)	0.992*	0.000 (0.003)	1.000
$\beta_{\text{CreditScore}}$	0.281* (0.067)	1.324*	1.069* (0.188)	2.914*
β_{WOZscore}	-0.015* (0.004)	0.986*	0.002 (0.010)	1.002
$\beta_{\text{IncomeScore}}$	-0.015* (0.007)	0.985*	0.024 (0.019)	1.024
$\beta_{\text{DurationPreviousStates}}$	-0.027* (0.001)	0.973*	-0.055* (0.002)	0.946*
Wald statistic	2326.000*			
Concordance	0.632			

(b) The parameter estimates of the Cox regression in module C, regarding transition 6 and 7.

	Module D			
Transition	$4 \rightarrow 5$		$4 \rightarrow 6$	
	coef	exp(coef)	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	0.003 (0.002)	1.003	-0.062* (0.008)	0.940*
$\beta_{\text{AgeDebtor}}$	-0.001* (0.001)	0.999*	0.007* (0.002)	1.007*
β_{AgeDebt}	0.003* (0.001)	1.003*	0.011* (0.004)	1.011*
$\beta_{\text{CreditScore}}$	-0.268* (0.080)	0.765*	0.824* (0.269)	2.279*
β_{WOZscore}	-0.001 (0.005)	0.999	0.033* (0.015)	1.034*
$\beta_{\text{IncomeScore}}$	0.020* (0.008)	1.020*	0.050** (0.027)	1.051**
$\beta_{\text{DurationPreviousStates}}$	-0.046* (0.001)	0.955*	-0.044* (0.003)	0.957*
Wald statistic	2808.000*			
Concordance	0.678			

(c) The parameter estimates of the Cox regression in module D, regarding transition 8 and 9.

Table 7: Coefficients of the Cox regressions in the duration-independent model.

	Module E	
Transition	$5 \rightarrow 6$	
	coef	exp(coef)
$\beta_{\text{ClaimAmount}}$	-0.016* (0.004)	0.984*
$\beta_{\text{AgeDebtor}}$	-0.001 (0.001)	0.999
β_{AgeDebt}	-0.029* (0.003)	0.971*
$\beta_{\text{CreditScore}}$	0.057 (0.161)	1.059
β_{WOZscore}	0.005 (0.009)	1.005
$\beta_{\text{IncomeScore}}$	0.011 (0.016)	1.011
$\beta_{\text{DurationPreviousStates}}$	-0.065* (0.001)	0.937*
Wald statistic	2838.000*	
Concordance	0.810	

(d) The parameter estimates of the Cox regression in module E, regarding transition 10.

The parameter estimates of the covariates in each Cox regression in module A-E, with their standard errors between brackets.
*p <0.05, ** p <0.10

6.2 Model performance

The predictive performances of the two models are assessed with the root mean squared prediction error (RMSPE). Both models predict the paths for all the debtors in our test data set, and we compare it with their actual path. This is done by computing the RMSPE based on the predicted number of months until a specific transition compared to the actual number of months till that specific transition. We compute the RMSPE based on an 80%-20% split in the training data, using only 80% of the training data to estimate the model and the other 20% as test data to evaluate the model's performance. Results are shown in Table 8. The results show that the duration-dependent model outperforms the duration-independent model in predicting the debtor's path. Including the duration-dependent variable into the Cox regressions thus improves the performance. Therefore we select the duration-dependent model as the best model and hence discuss its results extensively.

Table 8: RMSPE estimates of the two main models.

Model	RMSPE
duration-independent	10.51
duration-dependent	10.10

Comparing the duration-independent model and the duration-dependent using the RMSPE estimates. The RMSPE is based on an $80\% - 20\%$ split in train and test data.

The next results to discuss for the duration-dependent model are the total number of transitions for each pair of states in successive observation times predicted for the test set. These results are obtained by the model and shown in Table 9. They also indicate whether our model is able to predict correctly since we can compare the frequency fractions computed for the test set with the frequency fractions in our original data set (Table 4b). As we can see, these fractions are reasonably close to each other. However, there are some slight differences. Please take a look at the first row, what we can see here is that our model assigns more direct transitions to state 6 than is present in the original data set. More specific, module A assigns 40.5% of the transitions direct to state 6, which is approximately 10% more than in the original data set where the percentage of direct transitions to state 6 is equal to 30.6%. However, the other direct transitions to state 6 in module B-D have a somewhat smaller frequency fraction than in the actual data. These differences are between the $2.7\% - 5.7\%$. Subsequently, modules B-D usually assign the highest probability to the first competing state, resulting in more transitions to the nearest state than direct transitions to state 6. This phenomenon is supported in the data, as shown in table 4.

Table 9: Frequency table of the predicted transitions with the duration-dependent model for the test data set.

	1	2	3	4	5	6	no event	total entering
1	0	5,882	2,935	0	0	5,988	0	14,805
2	0	0	7,853	0	0	1,253	0	9,106
3	0	0	0	14,558	0	1,493	1	16,052
4	0	0	0	0	16,956	1,089	1	18,046
5	0	0	0	0	0	17,740	32	17,772
6	0	0	0	0	0	0	0	0

(a) The frequency numbers

	1	2	3	4	5	6	no event
1	0	0.397	0.198	0	0	0.405	0
2	0	0	0.862	0	0	0.138	0
3	0	0	0	0.907	0	0.093	0
4	0	0	0	0	0.940	0.060	0.000
5	0	0	0	0	0	0.998	0.002
6	0	0	0	0	0	0	0

(b) The frequency fractions

Frequency table of the total number of transitions for each pair of states in successive observation times, predicted for the test data set with the duration-dependent model. Subtable (b) presents the respective frequency fractions.

6.3 Duration-dependent model

In this section, we discuss some other results of the duration-dependent model in more detail, together with the payment predictions. We start with the estimated cumulative baseline hazards for each of the transitions in the Cox regressions illustrated in Figure 10. The baseline hazard represents the hazard when all of the covariates are equal to 0. The hazard, in our case, is the risk of transitioning to another state. Overall this figure depicts that all cumulative baseline hazards related to the direct transitions to state 6 (all red lines) increase exponentially. Furthermore, the baseline hazards of the transitions to the following states usually start higher than the transition to state 6 but flatten out after some time. In Figure 10a, we can see that in the first years after the purchase of the debt, the debtor is most likely to transfer to state 2. However, after approximately ten years, the baseline hazard of transition $1 \rightarrow 6$ exceeds the baseline hazard of transition $1 \rightarrow 2$. The baseline hazard of transition $1 \rightarrow 3$ is always beneath the baseline hazard of transition $1 \rightarrow 2$. Hence, without any additional covariates the $1 \rightarrow 3$ transitions is not likely to happen. Figure 10b shows that in the first years, more transitions from state 2 to state 3 will happen, while after approximately three years, transitions from state 2 to state 6 are more likely. Same holds for transitions $3 \rightarrow 4$ and $3 \rightarrow 6$, in which the latter becomes more likely after 2 years (Figure 10c). The $4 \rightarrow 6$ transition becomes more likely than the $4 \rightarrow 5$ transition after approximately seven years (Figure 10d).

In addition, to substantiate our decision to add transition specific covariates, it is also interesting to perform a test to evaluate whether a certain covariate has a different effect on different transitions within a module. To test this, we use the Wald test that evaluates the null hypothesis of equality of the coefficients within a module. The results can be found in Table 17 in the Appendix. The coefficients of the original claim amount and credit score in module A are significantly different from each other for all three transitions, according to the highly significant test statistics. For the other covariates, at least two transitions are significantly different from each other. In module B, only the coefficients of the original claim amount, age of debtor, and age of debt are significantly different from each other. In other words, these three covariates have a different effect on transition 4 than on transition 5. In modules C and D, four out of the seven covariates have a different effect on the two transitions in the respective modules. All these significant Wald test statistics within the modules suggest that it is beneficial to use transition specific covariates in the Cox regressions. In the end, transition specific covariates also make the interpretation of the coefficients easier.

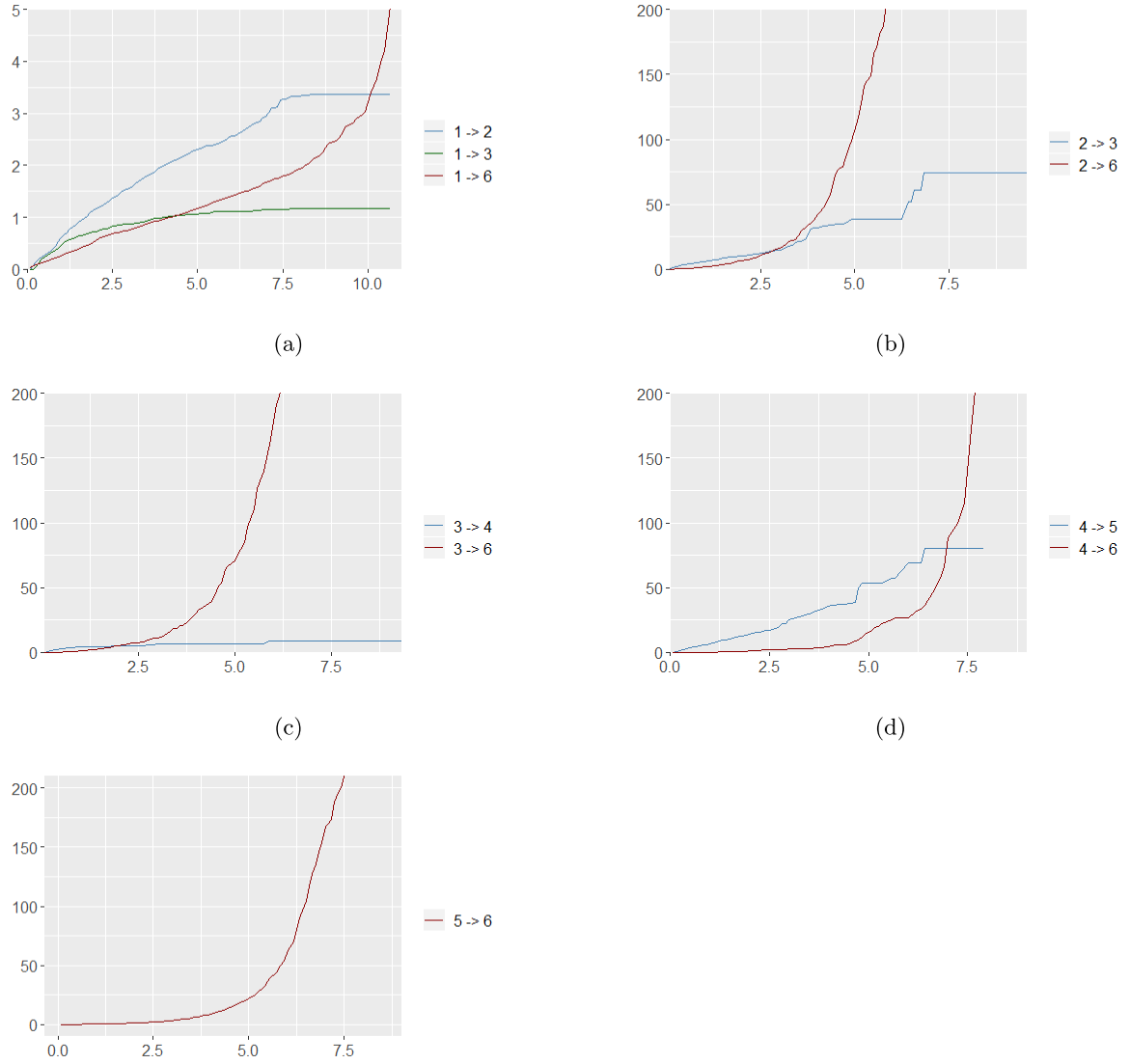


Figure 10: The cumulative baseline hazard curves for the duration-dependent model. Subfigure a shows the cumulative baseline hazard curves in module A; subfigure b shows the cumulative baseline hazard curves in module B etc. On the x-axis the years since purchase of the debt are shown and the y-axis presents the cumulative baseline hazard.

Next, after predicting the transition intensities with the Cox regressions, we construct the monthly transition matrices from the estimated transition probabilities over the next five years. We use them to obtain the individual debtor-state variables. By combining all the debtor-state variables we get more insights into the dispersion of the debtors amongst the states in the debt collecting process. Figure 11 presents a monthly overview of the debtors distributed over the different states for the next five years. It displays the percentage of total debtors per state. From this figure, we

can see that most debtors in our data set are currently located in state 1 (amicable) and we predict they remain in state 1 for the first half-year. Additionally, all debtors have made a transition to another state by the end of the first year of our prediction, with the highest percentage of debtors (48%) being in state 6. Specifically, the percentage of debtors in state 6 increases tremendously in the first year. However, this increase is quickly followed by a decrease of almost the same height in the following year. An interesting insight is that for each month, not more than approximately 10% of the total debtors is ever situated in either state 2 or state 3. This indicates that, in general, debtors move quickly through these states. Furthermore, we predict that after 1.5 years, most debtors are situated in state 5, followed by state 4. This continues until the end of our 5-year forecasting period. Hence, debtors that are in the process for multiple years most likely get stuck in the confiscation state, followed by the verdict state. In reality, confiscating the debt often does take up a few years.

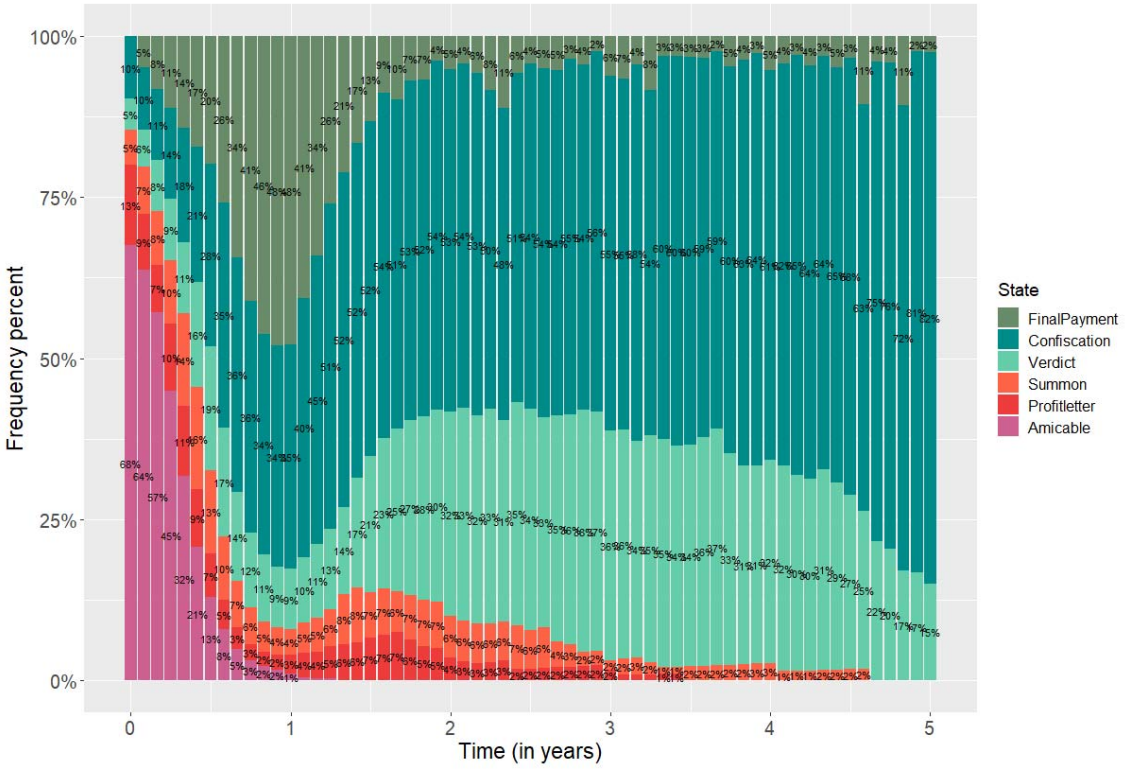


Figure 11: A monthly overview of the debtors distributed over the different states in the debt collecting process for the next five years.

6.3.1 Payment predictions

After the path predictions of all the debtors and constructing the debtor-state variables, we move on to the prediction of the payments as percentage of the original claim amount. The monthly average of this dependent variable in the train data set is equal to 0.10. This implies that, on average, the monthly payments made by the debtors are 10% of their original claim amount. Besides, the average payment rate when final payment is reached is equal to 1.01. Hence, the debtors eventually pay more than their original claim amount, which we can blame on the costs made in the process that are charged on the debtors. The estimation results of the linear and logistic regressions are presented in Table 10. The most interesting result is the significance of all debtor-state variables in both regressions. This corresponds with our expectation that the debtor-state variables are very informative predictors. The coefficients of the debtor-state variables have a negative sign, indicating that one extra month in one of these states has a negative effect on the payment rate. In both models, the most negative effect corresponds to state 1, and this negative effect decreases as we advance through the process. So a debtor in state 1 is less likely to pay than a debtor is state 5. The other explanatory variables are always positive, and most of them are significant in the linear regression. However, in the logistic regression, these remaining variables are not significant.

Table 10: Coefficients of the linear and logistic regressions.

	Linear	Logistic
$\beta_{\text{ClaimAmount}}$	0.000* (0.000)	0.000* (0.000)
$\beta_{\text{AgeDebtor}}$	0.000* (0.000)	0.000 (0.000)
β_{AgeDebt}	0.001* (0.000)	0.000 (0.001)
$\beta_{\text{CreditScore}}$	0.027* (0.005)	0.015 (0.035)
β_{WOZscore}	0.000 (0.000)	0.000 (0.002)
$\beta_{\text{IncomeScore}}$	0.002* (0.001)	0.001 (0.004)
$\beta_{\text{DurationPreviousStates}}$	-0.002* (0.004)	-0.001* (0.000)
β_{State1}	-0.992* (0.002)	-0.503* (0.013)
β_{State2}	-0.961* (0.003)	-0.486* (0.018)
β_{State3}	-0.932* (0.003)	-0.469* (0.018)
β_{State4}	-0.942 * (0.003)	-0.475* (0.017)
β_{State5}	-0.855* (0.003)	-0.424* (0.016)

The parameter estimates of the linear and logistic regressions, with their standard errors between brackets.
*p <0.05

.

Again, we assess the model performance of both models with the RMSPE. Table 11 indicates that the performance of both models is almost identical. However, the logistic regression slightly

outperforms the linear regression. Therefore, we discuss the results of the predicted payments based on the logistic regression in more detail.

Table 11: RMSPE estimates of the linear and logistic regressions.

Model	RMSPE
Linear regression	0.2998
Logit regression	0.2994

Comparing the linear regression and the logistic regression using the RMSPE estimates. The RMSPE is based on an $80\% - 20\%$ split in train and test data.

The predicted payments are depicted in Figure 12. This figure indicates that most payments will be collected within the first year of our forecast horizon. This coincides with the state distribution of state 6 in Figure 11, which shows a peak around the first year. Note that after reaching the final payment state, debtors leave the debt collecting process and do not contribute to the payments anymore. Furthermore, the average rate of payments relative to the original claim amount is approximately 0.11 in that first year.

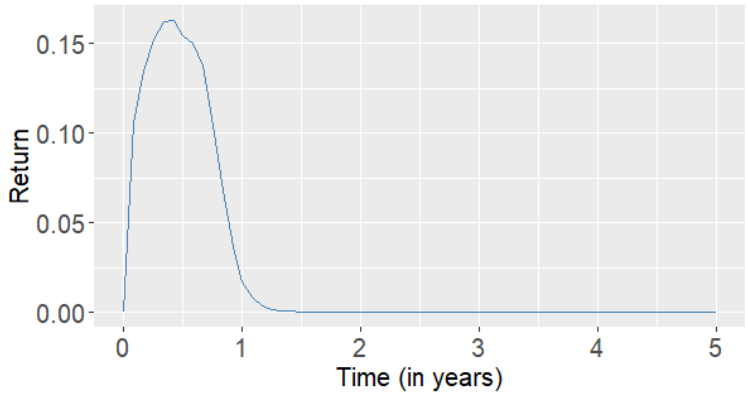


Figure 12: The predicted monthly payments for the next five years based on a logistic regression.

7 Conclusion

DirectPay is currently facing difficulties in predicting their incoming payment stream. In addition, it has a vast interest in predicting the future path of an individual debtor in order to get more insights into the debt collecting process. Therefore, this thesis performs an in-depth analysis of debtor paths in the debt collecting process and provides a prediction of the monthly payments. The process consists of six states in which transitions are governed by a Markov process. The

first step into modeling the path of the individual debtors is to predict the transition probabilities between these states over time. In order to do so, the main multi-state problem is split up into five smaller multi-state modules. In each module, the semi-parametric Cox regression is used to fit separate intensities to each transition. The modules are modeled with simple multi-state CRMs and a two-state survival model. This has the advantage that we can obtain explicit expressions for the Aalen-Johansen estimator of the transition probabilities.

In our analysis, we investigate two different versions of the main model. Firstly, the duration-independent model is built with fixed covariates in the Cox regressions. Then a duration-dependent model is constructed by adding a variable into the Cox regressions that accumulates the durations in previous states. The RMSPE implies that the duration-dependent model is the best performing model and is selected to predict the transition intensities. Hence, adding the duration-dependent variable improves the performance. This variable has a significant negative effect on all transitions, with an average coefficient equal to -0.049. This corresponds to an average hazard ratio equal to 0.952. So a one unit increase in total duration in the previous states is associated with a 4.8% decreased risk of transitioning. Hence, the more time a debtor has spent in the previous states, the less likely he will transfer to the next state.

After constructing the monthly transition matrices from the estimated transition probabilities, we obtain the individual debtor-state variables indicating the active state in each month over the next five years. These are, in turn, used in the linear and logistic regressions to predict the percentage of payments received relative to the original claim amount. According to the RMSPE, the logistic regression slightly outperforms the linear regression. The most interesting result in this logistic regression is the significance of all debtor-state variables. This corresponds with our expectation that the debtor-state variables are very informative predictors. The coefficients of the debtor-state variables have a negative sign, indicating that one extra month in one of these states has a negative effect on the payment rate. This negative effect is the strongest for state 1 and slightly decreases as we advance through the process. So a debtor in state 1 is less likely to pay than a debtor is state 5. A monthly overview of the debtors distributed over the different states is obtained by combining the future path predictions of all the individuals. Together with the results of the monthly payments predictions, they indicate that most payments will be collected within the first year of our forecast horizon. In that first year of our prediction, the average payment rate is equal to 0.11. Additionally, this overview also shows that most debtors get stuck in the confiscation state. However, in reality, confiscating the debt often takes up a few years.

Furthermore, this research is subject to some limitations. First of all, the data set is based on monthly observations, and only the last transition of the month is registered. Hence, some transitions are not captured by the data set. However, multiple transitions within a month are not often observed in reality. The second limitation of this research is that during this process, debtors can enter a payment arrangement with DirectPay to payoff $x\%$ of the outstanding debt each month. If a debtor is in such an agreement, it does not move states until he reaches his final payment. However, our model is not able to recognize this, so it might be interesting to investigate whether we can add a sub-state that describes this payment arrangement. In addition, DirectPay's ultimate goal is to have adequate predictions of the payments for the next ten years. In our model, the transition probabilities converge quickly and are therefore less adequate to use in a 10-year forecast.

Finally, future steps could be to implement the duration-dependent model for the other branches in which DirectPay operates. Alternatively, the customer branch can also be included as dummy variables and used as additional covariates in the Cox regressions. In addition, further research could examine the possibility of combining our analysis together with the analysis of a previous study done by Koomen (2018) commissioned by DirectPay. In his study, they use predictive modeling to increase the efficiency of the summons. They explore whether it is beneficial to summon specific debtors or that it might be worthless because the debtor is simply not able to pay off the debt. We gain from this knowledge because it indicates whether a debtor will advance in the debt collecting process and thus whether we could expect a final payment. We could include this information as another covariate in the Cox regressions. Lastly, another suggestion for further research is to implement a risk assessment of different scenarios using a bootstrap method. This simulation study can help us to analyze the effect of specific covariate values on the future path of the debtors and their respective payments.

References

Aalen, O. (1978). Non-parametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726.

Aalen, O., Borgan, Ø., and Fekjær, H. (2001). Covariate adjustment of event histories estimated from markov chains: the additive approach. *Biometrics*, 57(4):993–1001.

Aalen, O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150.

Andersen, P., , and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115.

Andersen, P. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine*, 7(6):661–670.

Andersen, P., Esbjerg, S., and Sørensen, T. (2000). Multi-state models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, 19(4):587–599.

Andersen, P., Hansen, L., and Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics*, 18(2):153–167.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. New York, Springer.

Cameron, A. C. and Trivedi, P. K. (2005). *Microecometrics: methods and applications*. New York, NY: Cambridge University Press.

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

D'Agostino, A. and Mealli, F. (2000). Modelling short unemployment in europe. Technical report, ISER Working Paper Series.

Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Fiocco, M., Putter, H., and van Houwelingen, H. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine*, 27(21):4340–4358.

Gill, R. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The annals of statistics*, 18(4):1501–1555.

Harrell Jr, F., Lee, K., and Mark, D. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kavkler, A., Danacica, D., Babucea, A. G., Bicanic, I., Bohm, B., Tevdovski, D., ..., and Borsic, D. (2009). Cox regression models for unemployment duration in romania, austria, slovenia, croatia and macedonia. *Romanian Journal of Economic Forecasting*, 10(2):81–104.

Klein, J., Keiding, N., and Copelan, E. (1993). Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine*, 12(24):2315–2332.

Koomen, d. Y. (2018). Increasing summons efficiency with predictive modelling. Master's thesis, Erasmus University Rotterdam.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Rodriguez, G. (2010). Parametric survival models. *Regres. Model. Strateg*, pages 1–14.

Shu, Y. and Klein, J. (2005). Additive hazards markov regression models illustrated with bone marrow transplant data. *Biometrika*, 92(2):283–301.

Slavík, A. (2007). *Product integration, its history and applications.* Matfyzpress Prague.

A Appendix

A.1 Product integral

The product integral correspond to taking the limit of a product. It is the continuous version of a discrete product. Therefore, the product integral reduces to the regular product in the discrete case (Equation 9) and to the exponential of the regular integral in the continuous case (Equation 6). Gill and Johansen (1990) discuss the key facts on product-integration in their paper. The most popular product integrals are the Volterra integral (Slavík, 2007):

$$\prod_a^b \{1 + f(x)dx\} = \lim_{\Delta x \to 0} \prod \{1 + f(x_i)\Delta x\} \tag{66}$$

Then the following relationship exists for scalar functions, $[a, b] \to \mathbb{R}$

$$\prod_a^b \{1 + f(x)dx\} = \exp\left(\int_a^b f(x)dx\right) \tag{67}$$

A.2 Data preparations.

Table 12: Data preparations

	Number of rows	Number of debtors
Original data set	10 million	219,443
Remove debtors with original claim amount $\leq$ €0.00	9.8 million	217,065
Selecting the months of transition	415,500	217,065
Removing non-valid transitions	403,100	217,065
Removing NA values	296,500	157,400

Overview of the data preparations in order to obtain the original data set.

Table 13: Frequency table of transitions in the original data set before preparing the data.

	1	2	3	4	5	6
1	0	43,927	25,214	2,082	31	29,949
2	1,806	0	18,184	1,314	30	4,361
3	1,296	782	0	30,563	577	4,393
4	92	57	582	0	23,239	2,311
5	4	115	1,222	31	0	6,270
6	0	0	0	0	0	0

Frequency table of the total number of transitions for each pair of states in successive observation times, in time period January 2008 - November 2019. (Including non allowed transitions)

A.3 Details scoring systems

Table 14: The WOZ score.

	Category
1	< €100,000
2	€100,000 - €125,000
3	€125,000 - €150,000
4	€150,000 - €175,000
5	€175,000 - €200,000
6	€200,000 - €225,000
7	€225,000 - €250,000
8	€250,000 - €300,000
9	€300,000 - €350,000
10	€350,000 - €400,000
11	€400,000 - €600,000
12	€600,000 - €1,000,000
13	> €1,000,000

The internal WOZ score that is used by DirectPay to categorize their clients.

Table 15: The income score.

	Category
1	lower than average income
2	average income
3	1.5x average income
4	2x average income
5	2.5x average income or higher

The internal income score that is used by DirectPay to categorize their clients.

A.4 Results

Table 16: Transitions in the debt collecting process.

Transition number	Transition
1	$1 \rightarrow 2$
2	$1 \rightarrow 3$
3	$1 \rightarrow 6$
4	$2 \rightarrow 3$
5	$2 \rightarrow 6$
6	$3 \rightarrow 4$
7	$3 \rightarrow 6$
8	$4 \rightarrow 5$
9	$4 \rightarrow 6$
10	$5 \rightarrow 6$

The transition numbers for each specific transition in the debt collecting process.

Table 17: Wald statistics for the duration-dependent model.

	Module A
$\beta_{\text{ClaimAmount.1}} = \beta_{\text{ClaimAmount.2}}$	205.14*
$\beta_{\text{ClaimAmount.1}} = \beta_{\text{ClaimAmount.3}}$	3673*
$\beta_{\text{ClaimAmount.2}} = \beta_{\text{ClaimAmount.3}}$	1788.9*
$\beta_{\text{AgeDebtor.1}} = \beta_{\text{AgeDebtor.2}}$	12.45*
$\beta_{\text{AgeDebtor.1}} = \beta_{\text{AgeDebtor.3}}$	33.18*
$\beta_{\text{AgeDebtor.2}} = \beta_{\text{AgeDebtor.3}}$	2.85
$\beta_{\text{AgeDebt.1}} = \beta_{\text{AgeDebt.2}}$	1.77
$\beta_{\text{AgeDebt.1}} = \beta_{\text{AgeDebt.3}}$	15.78*
$\beta_{\text{AgeDebt.2}} = \beta_{\text{AgeDebt.3}}$	5.23*
$\beta_{\text{CreditScore.1}} = \beta_{\text{CreditScore.2}}$	11.55*
$\beta_{\text{CreditScore.1}} = \beta_{\text{CreditScore.3}}$	138.91*
$\beta_{\text{CreditScore.2}} = \beta_{\text{CreditScore.3}}$	55.73*
$\beta_{\text{WOZscore.1}} = \beta_{\text{WOZscore.2}}$	22.83*
$\beta_{\text{WOZscore.1}} = \beta_{\text{WOZscore.3}}$	3.23
$\beta_{\text{WOZscore.2}} = \beta_{\text{WOZscore.3}}$	37.29*
$\beta_{\text{IncomeScore.1}} = \beta_{\text{IncomeScore.2}}$	34.41*
$\beta_{\text{IncomeScore.1}} = \beta_{\text{IncomeScore.3}}$	0.70
$\beta_{\text{IncomeScore.2}} = \beta_{\text{IncomeScore.3}}$	39.18*

(a) Covariates with .1 relate to the $1 \rightarrow 2$ transition; covariates with .2 relate to the $1 \rightarrow 3$ transition; covariates with .3 relate to the $1 \rightarrow 6$ transition.
* indicates significance on the 95% significance level.

Table 17: Wald statistics for the duration-dependent model.

	Module B
$\beta_{\text{ClaimAmount.4}} = \beta_{\text{ClaimAmount.5}}$	238.19*
$\beta_{\text{AgeDebtor.4}} = \beta_{\text{Agedebtor.5}}$	4.03*
$\beta_{\text{AgeDebt.4}} = \beta_{\text{Agedebt.5}}$	9.78*
$\beta_{\text{CreditScore.4}} = \beta_{\text{CreditScore.5}}$	2.67
$\beta_{\text{WOZscore.4}} = \beta_{\text{WOZscore.5}}$	2.73
$\beta_{\text{IncomeScore.4}} = \beta_{\text{IncomeScore.5}}$	0.363
$\beta_{\text{DurationPreviousStates.4}} = \beta_{\text{DurationPreviousStates.5}}$	3.27

(b) Covariates with .4 relate to the $2 \rightarrow 3$ transition; covariates with .5 relate to the $2 \rightarrow 6$ transition.
* indicates significance on the 95% significance level.

	Module C
$\beta_{\text{ClaimAmount.6}} = \beta_{\text{ClaimAmount.7}}$	313.71*
$\beta_{\text{AgeDebtor.6}} = \beta_{\text{Agedebtor.7}}$	0.86
$\beta_{\text{AgeDebt.6}} = \beta_{\text{Agedebt.7}}$	5.85*
$\beta_{\text{CreditScore.6}} = \beta_{\text{CreditScore.7}}$	15.73*
$\beta_{\text{WOZscore.6}} = \beta_{\text{WOZscore.7}}$	2.19
$\beta_{\text{IncomeScore.6}} = \beta_{\text{IncomeScore.7}}$	3.80
$\beta_{\text{DurationPreviousStates.6}} = \beta_{\text{DurationPreviousStates.7}}$	199.61*

(c) Covariates with .6 relate to the $3 \rightarrow 4$ transition; covariates with .7 relate to the $3 \rightarrow 6$ transition.
* indicates significance on the 95% significance level.

	Module D
$\beta_{\text{ClaimAmount.8}} = \beta_{\text{ClaimAmount.9}}$	63.38*
$\beta_{\text{AgeDebtor.8}} = \beta_{\text{Agedebtor.9}}$	13.77*
$\beta_{\text{AgeDebt.8}} = \beta_{\text{Agedebt.9}}$	3.53
$\beta_{\text{CreditScore.8}} = \beta_{\text{CreditScore.9}}$	15.17*
$\beta_{\text{WOZscore.8}} = \beta_{\text{WOZscore.9}}$	4.84*
$\beta_{\text{IncomeScore.8}} = \beta_{\text{IncomeScore.9}}$	1.13
$\beta_{\text{DurationPreviousStates.8}} = \beta_{\text{DurationPreviousStates.9}}$	0.50

(d) Covariates with .8 relate to the $4 \rightarrow 5$ transition; covariates with .9 relate to the $4 \rightarrow 6$ transition.
* indicates significance on the 95% significance level.

Wald statistics to determine whether the coefficients of the covariates in the duration-dependent model are significant differently from each other for the different transitions. Subtable a corresponds with the covariates in module A, Subtable b corresponds to the covariates in module B , Subtable c corresponds to the covariates in module C and Subtable d corresponds to the covariates in module D.

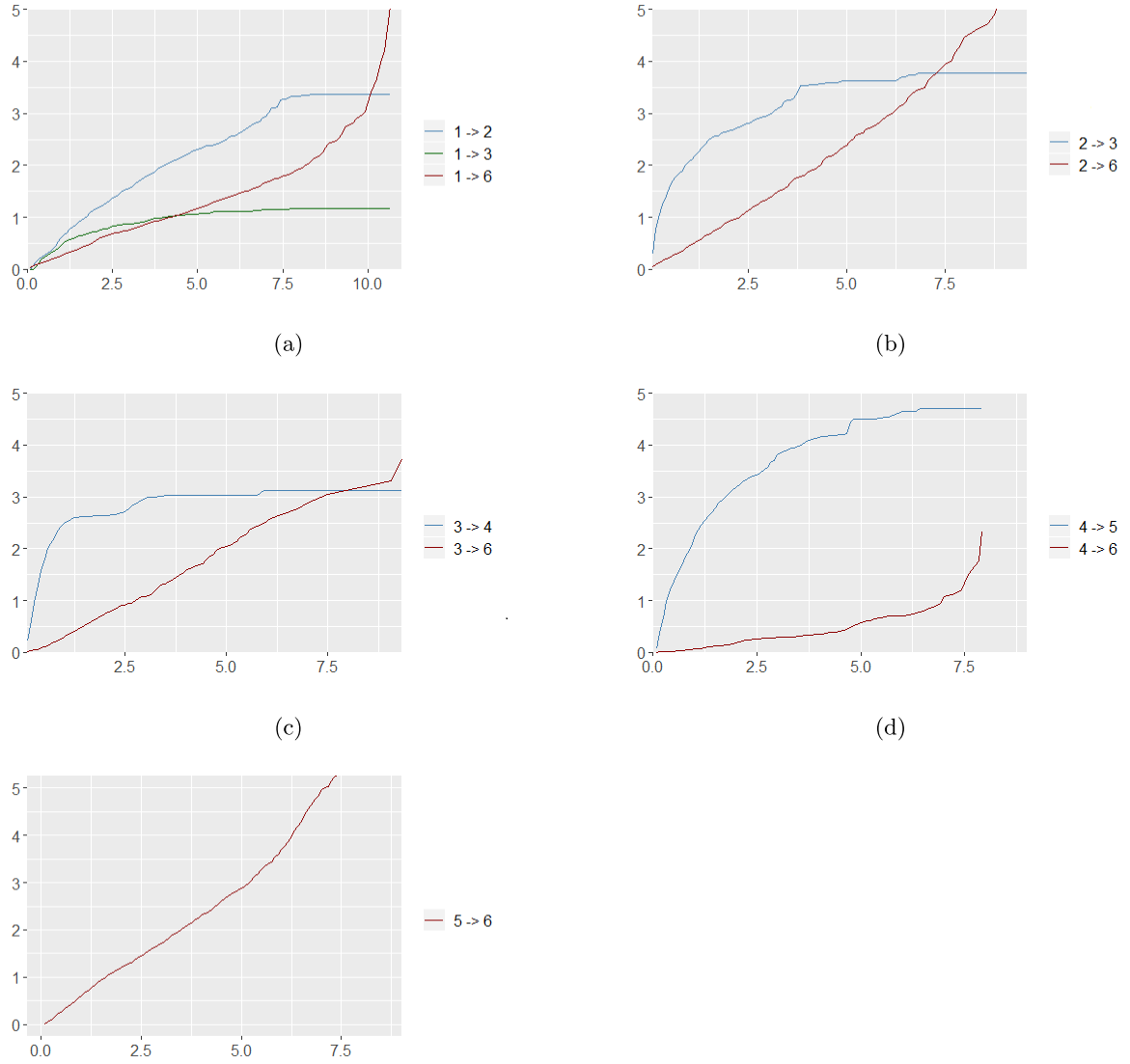


Figure 13: The cumulative baseline hazard curves for the duration-independent model. Subfigure a shows the cumulative baseline hazard curves in module A; subfigure b shows the cumulative baseline hazard curves in module B etc. On the x-axis the years since purchase of the debt are shown and the y-axis presents the cumulative baseline hazard.