STUDY OF MODELLING THE TIME TO LOAN DEFAULT USING MULTI-STATE MODELS

Filip Habarta – Lubomír Štěpánek – Stanislav Kováč – Ivana Malá

Abstract

In the credit market, one of the biggest parts of the risk for the creditors is the default of provided loan and as there is typically a huge competition on the credit market the ability to precisely estimate such risks connected to each loan application is crucial for success. In this paper, we focus on the modelling of the time to default of loans on individual level using multi-state models to quantify the risk of loan default for creditors. We define a basic multi-state model with possible states (regular payments, payments delay, default) and transitions in the loan repayment process. Further, we show how to estimate the corresponding transition hazards using Cox proportional hazard model. The defined model is then estimated on the portfolio of loans obtained from the peer-to-peer lending platform Bondora. Presented results show the dependency of time to default on the gender, age, and the borrower's known history of behaviour.

Key words: Time to loan default, Multi-state-model, Cox proportional hazard model

JEL Code: C14, C15

Introduction

Problematics of credit markets can be viewed from a variety of perspectives. In this paper, we have decided to look closer at the perspective of loan default as it takes a huge part in the correct assessment of risks connected to borrowers. Credit risk is important for correct pricing and is needed to be quantified and reported to the financial authorities. As the data about borrowers are becoming more detailed, we can also focus on higher detail in risk modelling. One of the processes closely connected to the overall risk of the loan is the repayment process that leads to the default of the loan that is most likely dependent on the factors given at the beginning of the loan, like borrower's characteristics, but also on the changes in the time as the loan is being repaid.

We will further present a model to estimate client behaviour in time before the loan is defaulted and present its application on the real market data from peer to peer investing platform Bondora (Bondora, 2020).

In measuring the time to loan repayment, researchers focus mainly on the methods from the area of survival analysis that are often taken over from the field of biostatistics. The problematics of modelling time to loan repayment using methods of survival analysis is for example presented in papers (Drick, 2017) and (Stepanova, 2002). Furthermore, the multi-state model approach to modelling time to debt events is for example described in papers (Chamboko, 2020) and (Balibek, 2010). Generally, introduction to problematics of multi-state models in survival analysis that are estimated using R programming language (R Core Team, 2020) is presented in the following papers (Wreed 2011), (Wreed, 2010) and (Putter, 2007).

1 Modelling approach

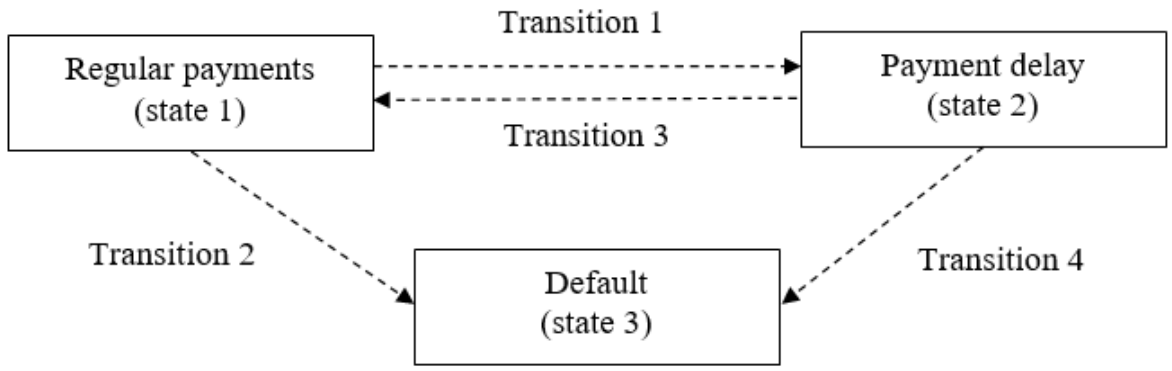
Our goal is to follow the process of loan repayment that ends in the inability to repay the loan value by the borrower. Modelling time to event data is the domain of survival analysis and to add the possibility to study the whole process we use multi-state models. A difference between the classical survival approach, where we focus only on modelling time to default and multi-state models which allows for the modelling of time to default in all specified states including time-dependent covariates and state-specific covariates, e.g. clients history is mainly in the detail which is intended to be studied using these models. In this regard, multi-state models perform better, but often the decision between these approaches is taken based on the available data, since if there is a low number of observations at our disposal the estimates from complex multi-state models will yield higher variability.

To model loan defaults, we have decided to use multi-state model in the form of clock forward markovian model. To be able to fully define such a model it is needed to define states and transitions based on the problematics of the loan repayment process and available data. Furthermore, it is needed to get the transition specific survival times, usually estimated either by Kaplan Meier estimator (Meier, 1958) or Cox proportional hazard model (Cox, 1984). In our case, we are going to use Cox proportional hazard model as it allows for simple incorporation of influence of various covariates.

1.1 Multi-state model definition

Definition of the estimated model follows a basic form of process that leads to the default of clients borrowing money. The multi-state model captures 4 possible transitions between 3 states (i) representing the phases of loan repayment process – regular payments (i = 1), late repayments (i = 2) and default (i = 3) which is in every case considered as absorbing state. The initial state of the model is represented by the client and the time when he made the first scheduled repayment of the loan. The definition of the model allows for a client to recover from payment delays by making regular payments again. A multi-state model diagram showing all states and possible transitions between them is visualized in Figure 1.

Fig. 1: Multi-state model diagram showing possible states and transitions



Source: author

We assume that clients always meet the first scheduled loan repayment. Furthermore, we assume that underlying data for modelling consists only of loans for which we observed default event-absorbing state. Note that because of these assumptions, there is no censoring present in the data as all transition times are fully known and recorded in the data. States of regular payments and payment delays are recurrent, but since we expect non-zero transitional probabilities to the state of default, the whole model is aperiodic.

Key to the successful modelling of the defined model scheme is to be able to define and estimate the desired transition matrix of the multi-state model. The first step is always to adjust the data structure so that it allows for modelling for specified transitions which is further described in the following chapter. A transformed model in the form of a transition matrix is shown in (1), where *NA* represents an impossible transition between states

$$transition\ matrix\ = \begin{pmatrix} NA & 1 & 2 \\ 3 & NA & 4 \\ NA & NA & NA \end{pmatrix}. \quad (1)$$

1.2 Transition hazards

As a method to obtain transition hazards, we decided to use Cox's proportional hazards model (Cox, 1984) and (Therneau, 2000). A Cox model was explicitly designed to be able to estimate the hazard ratios without having to estimate the baseline hazard function. This is exactly what we need to get the generalised possibility of predicting outcomes of the multi-state model. The Cox model hazard ratios are an exponential function (*exp*) of an arbitrary baseline hazard λ_0.

Basic model notation is taken from (Wreed, 2010). Assume that we have a model with N states and transitions are always from state $i = 1, 2, \ldots, N$ to state $j = 1, 2, \ldots, N$ and specific covariate vector regarding this transition is denoted by $\boldsymbol{Z}$. Then the hazard λ_{ij} for the Cox proportional hazard model is for the transition from state i to state j in time $t < T$ given by (2), where T represents the time of reaching absorbing state in our specific model

$$\lambda_{ij}(t \mid \mathbf{Z}) = \lambda_{ij,0}(t)\exp\left(\boldsymbol{\beta}_{ij}^{\top}\mathbf{Z}\right). \tag{2}$$

In the previous equation (2), $\boldsymbol{\beta}_{ij}^{T}$ represents the regression coefficients and $\lambda_{ij,0}(t)$ is the corresponding baseline hazard. If we assume that event times are distinct, we can search for the coefficients $\boldsymbol{\beta}$ by maximising the partial likelihood (3)

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}^{\top}\mathbf{Z}_j)}{\sum_{l\in R_j}\exp\left(\boldsymbol{\beta}^{\top}\mathbf{Z}_l\right)}. \tag{3}$$

Which is a hazard of the events observed in concrete transition in comparison to all the individuals at risk in a given transition. The estimate, called "Breslow's" (Breslow, 1975) for baseline cumulative hazard is (4)

$$\widehat{\Lambda}_0(t) = \sum_{j:t_j\leqslant t} \frac{1}{\sum_{l\in R_j}\exp\left(\widehat{\boldsymbol{\beta}}^{\top}\mathbf{Z}_l\right)}. \tag{4}$$

In regard to multi-state modelling, it leads to modelling and estimating a stratified Cox model, which suggests only that there will be as many Cox models estimated as there are transitions in the model.

1.3 Multi-state model predictions

In order to get predictions for the multi-state model that captures the history of individuals, we need to estimate the conditional probabilities of state transitions given already known information. The estimates of these probabilities are based on the results obtained from the Cox model on the transition hazards between the states.

If we want to evaluate transition probabilities in future, we want to get the conditional probability of a certain event in time *t*, denoted as E_t, which represents individuals' ability to achieve that state representing such event from the initial state through the transition in our interest and is conditional on the individual's known characteristics Z and his observed history $\mathcal{H}_u$. Such conditional probability is (5)

$$\text{Prob}\,(E_t \mid \mathcal{H}_u, \mathbf{Z})\,. \tag{5}$$

Detailed description of the formula (5) way of calculation can be found for the example of basic three-state "illness dead" model in (Putter, 2007).

Data

In this paper, we leverage open access to the data from peer to peer lending platform Bondora (Bondora, 2020) that belongs among the biggest peer to peer lending platforms in Europe. Data were obtained from their public database.

From available datasets, we used two files, the first dataset containing general information about each loan and creditor and the second file that contains information about late payments, clients defaults and collection process. These data files were prepared for modelling using R programming language (R Core Team, 2020) and data preparation used the functions and syntax from the "tidyverse" package (Wickham, 2019).

To be able to analyse data using multi-state models it is crucial to transform data into so format where each row represent one transition between states of the multi-state model. The structure of the used model is shown in Figure 1.

Tab. 1: Example of data structure - long format

LoanId	from	to	trans	Tstart	Tstop	pr	Gender	Age	status
0003D…	1	2	1	0	1156	0	1	46	1
0003D…	2	1	3	1156	1158	0	1	46	1
0003D…	1	2	1	1158	1257	1	1	46	1
0003D…	2	3	4	1257	1319	2	1	46	1

Source: authors' calculations

The data structure of prepared data is shown in Table 1. Identifier, "LoanId", of given loan is in the first column, further information about multi-state model transitions and states are

in columns “from”, “to” and “trans”. Columns “Tstart” and “Tstop” represents time in days since the first repayment of the loan and together create the interval of the length of stay in the given state in days. Variable “pr” is constructed to represent the knowledge of debtor behaviour that was already observed and is known. We distinguish three possible situations, value 0 represents no previous knowledge of payment delay, value 1 represents observed one delay in loan repayment and finally, value 2 represent two or more observed delays in loan repayment.

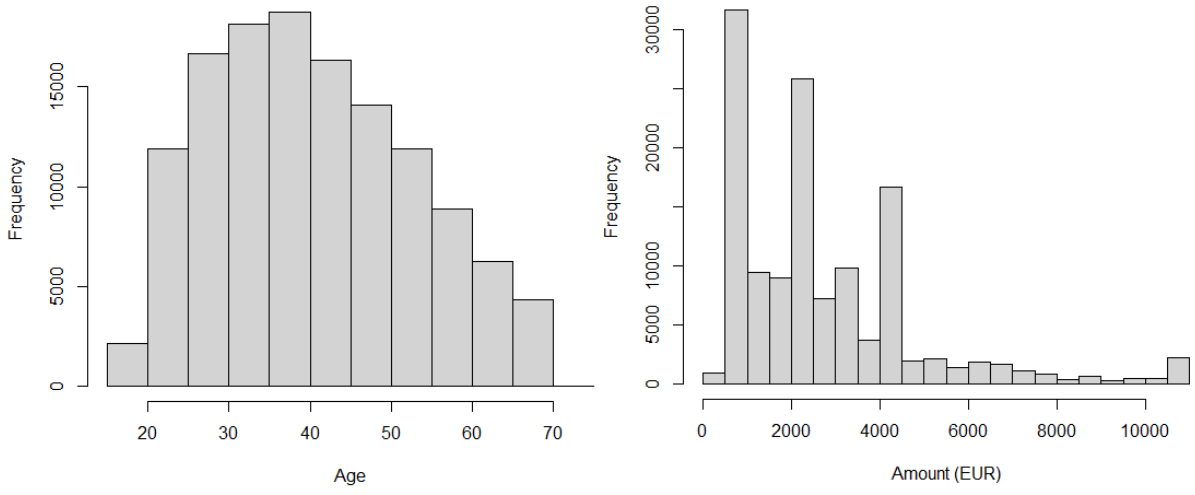
Tab. 2: Example of data structure – state specific covariates

LoanId	pr1.1	pr1.2	pr1.3	pr1.4	pr2.1	pr2.2	pr2.3	pr2.4
0003D…	0	0	0	0	0	0	0	0
0003D…	0	0	0	0	0	0	0	0
0003D…	1	0	0	0	0	0	0	0
0003D…	0	0	0	0	0	0	0	1

Source: authors’ calculations

Further covariates used for the analysis are information about gender and age. Histogram showing client age at the time of the first repayment of the loan (median is 40) is presented in Figure 2 on the left together with the histogram showing the amounts of loans in EUR at the beginning of repayment (median amount is 2125 EUR) on the right side. The proportion of males in the whole dataset is 65.4 %.

Fig. 2: Histogram for the covariate of age and amount



Source: author

To be able to fully incorporate the covariates in the "clock-forward" Markov model, they need to be transformed to state-specific covariates (time and transition dependent covariates). Examples of the transformation for the "pr" covariates for the same loan that was presented in Table 1 are shown in Table 2. Transition is represented by the number following covariate name and value after the dot.

Model estimation

In this chapter, we will demonstrate the estimation of the Cox stratified proportional hazard model, followed up by an example of the multi-state model prediction. Model estimation was done in R programming language and using "mstate" package (Putter, 2007).

Observed transition frequencies in the data set that are presented in Table 3 shows how unbalanced the dataset is regarding the observed transitions. We do not handle this issue specifically here in this paper, but it is planned to address this issue in further research.

Tab. 3: Observed transitional frequencies

From \ To	Standard repayment (1)	Payment delay (2)	Default (3)
Standard repayment (1)	0	49 710	232
Payment delay (2)	1 928	0	47 954
Default (3)	0	0	48 186

Source: authors' calculations

We estimate the stratified Cox model where the covariates are gender and age of individuals receiving the loan. The stratification variable ("trans") is given by the given transitions. Time spent in each state is then further captured by incorporating the variables "Tstart" and "Tstop" in the model. Finally, as time-dependent covariate that represents knowledge about the client's history of repayment is taken the variable "pr" described in more detail in the previous chapter. In Figure 3 we present the values of estimated coefficients and respective standard errors.

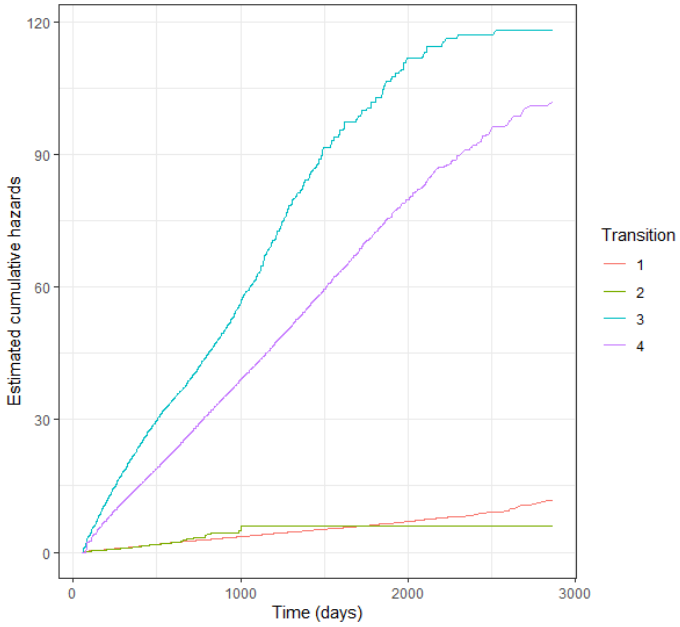
Fig. 2: Estimated stratified Cox model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
pr1.1	0.5934779	1.8102733	0.0215800	27.501	< 2e-16	***
pr1.3	0.1370644	1.1469021	0.1045728	1.311	0.189956	
pr1.4	-0.0197800	0.9804144	0.0260265	-0.760	0.447258	
pr2.1	0.6346399	1.8863427	0.0436653	14.534	< 2e-16	***
pr2.3	0.1144274	1.1212313	0.0566469	2.020	0.043382	*
pr2.4	-0.0747314	0.9279927	0.0194619	-3.840	0.000123	***
Gender1	-0.0327356	0.9677944	0.0057975	-5.647	1.64e-08	***
Age	-0.0015996	0.9984017	0.0002251	-7.107	1.19e-12	***

Source: author

To obtain prediction probabilities for the presented multi-state models, we need to further calculate baseline transition hazards for the specific loan covariates for each transition in the model. We can see the cumulative baseline hazards for each transition in Figure 4.

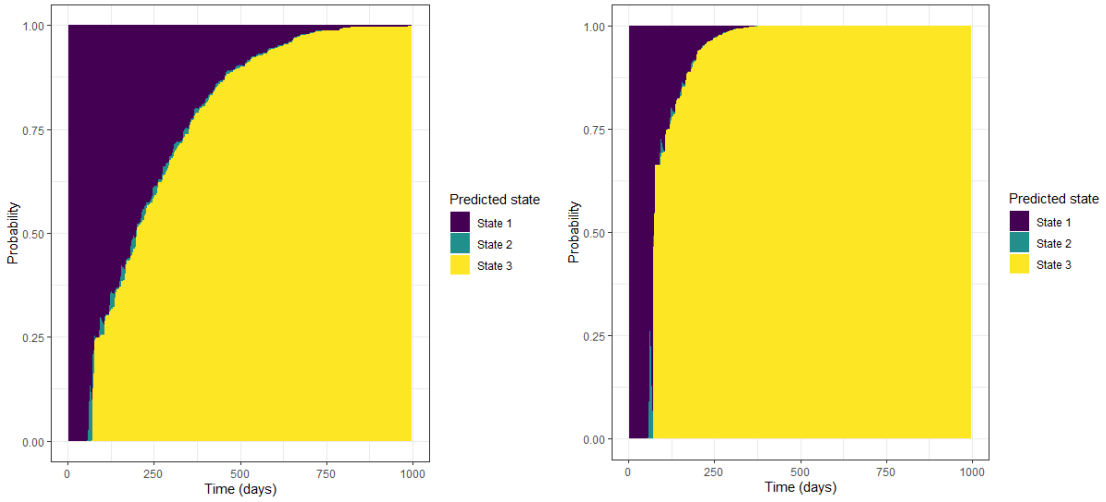
Fig. 4: Estimated cumulative baseline hazards



Source: author

The final step is to use the resulting loan-specific transition to obtain also, loan-specific, transition probabilities over time. For the male who is 40 years old, we can see the estimated probabilities of being in the modelled state over time if we assume that he has just currently made the first loan repayment in Figure 5 on the left. On the right side for comparison, we can see estimated probabilities of being in the modelled state overtime for a 20-year-old female, that has currently gone through the one payment delay.

Fig. 5: Estimated probability of being in given state at given time



Source: author

Conclusion

Area of modelling time to loan default allows for various ways of model construction. In this paper, we presented the multi-state model that further used Cox proportional hazard estimates dependent on the age, gender, and history of previous behaviour to calculate the state-specific transitional probabilities. Choice of this model was based on the desire to study the loan default process in higher detail and supported by the fact that we had data of a large portfolio at our disposal even though we were still dealing with the problem of observing quite rare events – defaults.

The theory behind the models was described and followed by the demonstration of the use of such models on the real data on the portfolio of loans. Study on the real data brought insides on the process of client's progress to default. We have presented on the real data from peer-to-peer lending platform how to get estimates of the time that it takes for an individual to get to default. As we expected there is a dependency between time to default and gender, age and borrower history.

This paper is focused on broadening the knowledge and getting a more detailed image of the loan default process. Further research should focus on the study of the influence of client-specific covariates like debtors' education or occupation and additionally the estimated models should be applied to the whole contract portfolio to study its development over time.

Acknowledgment

This paper is supported by the grant F4/45/2020 which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

Balibek, E., & Köksalan, M. (2010). A multi-objective multi-period stochastic programming model for public debt management. *European Journal of Operational Research*, *205*(1), 205-217.

Bondora (2021). *Bondora public datasets*. Bondora.com. https://www.bondora.com/en/public-reports.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 45-57.

Cox, D. R., & Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman & Hall. ISBN 978-0412244902

Chamboko, R., & Bravo, J. M. (2020). A multi-state approach to modelling intermediate events and multiple mortgage loan outcomes. *Risks*, *8*(2), 64.

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, *68*(6), 652-665.

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

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

R core team (2020). R: A language and environment for statistical computing, 2020, R Foundation for Statistical Computing, Vienna, Austria.

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, *50*(2), 277-289.

Therneau, T. M., & Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model* (pp. 39-77). Springer, New York, NY.

Wickham, et al. Welcome to the tidyverse. *Journal of Open Source Software*. 2019.

De Wreede, L. C., Fiocco, M., & Putter, H. (2011). mstate: an R package for the analysis of competing risks and multi-state models. *Journal of statistical software*, *38*(1), 1-30.

De Wreede, L. C., Fiocco, M., & Putter, H. (2010). The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine*, *99*(3), 261-274.

Contact

Ing. Filip Habarta
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
filip.habarta@vse.cz

Ing. MUDr. Lubomír Štěpánek
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
lubomir.stepanek@vse.cz

Bc. Stanislav Kováč
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
stanislav.kovac@vse.cz

doc. RNDr. Ivana Malá, CSc.
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague 3 – Žižkov, Czech Republic
malai@vse.cz