
# Recovery process optimization using survival regression

Jiří Witzany

Anastasiia Kozina

**FFA Working Paper 4/2020**


**FACULTY OF FINANCE AND ACCOUNTING**

**Bibliographic information:**

Witzany J., Kozina A. (2020). *Recovery process optimization using survival regression*. FFA Working Paper 4/2020, FFA, University of Economics, Prague.

This paper can be downloaded at: wp.ffu.vse.cz

Contact e-mail: ffawp@vse.cz

# Recovery process optimization using survival regression

## Authors

Jiří Witzany[1]
Anastasiia Kozina[2]

## Abstract

The goal of this paper is to propose, empirically test and compare different logistic and survival analysis techniques in order to optimize the debt collection process. This process uses various actions, such as phone calls, mails, visits, or legal steps to recover past due loans. We focus on the soft collection part, where the question is whether and when to call a past-due debtor with regard to the expected financial return of such an action. We propose using the survival analysis technique, in which the phone call can be compared to a medical treatment, and repayment to the recovery of a patient. We show on a real banking dataset that, unlike ordinary logistic regression, this model provides the expected results and can be efficiently used to optimize the soft collection process.

**AMS/JEL classification:** G21, G28, C14

**Keywords:** credit risk modelling, survival analysis, scoring, receivables, debt recovery, collection, retail banking, credit risk

## 1. Introduction

The recovery process has become an important part of the banking business model. Its main task is to manage overdue receivables through various enforcement tools, with the goal of maximizing the final recovery. At present, due to growing portfolios and in order to streamline all the activities performed, in particular those related to the retail segments, banks are trying to make most of the daily recurring processes as automated and efficient as possible. The recovery process is, in this respect, no exception, and, therefore, modifications and improvements are constantly being developed.

The aim of this study is to use logistic regression or survival analysis to develop and propose a system that streamlines the process of debt recovery and creates more effective soft collection strategies through telephone communication with the client.

There is relatively limited research on the subject of the optimization of the recovery process. One of the first papers by De Almeida Filho et al. (2010) proposes building a dynamic programming model to optimize collections. The dynamical programming approach has been followed by several other papers (e.g. van de Geer et al., 2018, or So et al., 2019 using Bayesian dynamic programming). Chehrazi at al. (2015, 2019) model repayments as a self-exciting stochastic process and propose using stochastic

[1] Corresponding author: University of Economics in Prague, Faculty of Finance and Accounting, W. Churchill Sq. 4, 130 67, Prague, Czech Republic, +420 224 095 174, e-mail: jiri.witzany@vse.cz.
[2] University of Economics in Prague, Faculty of Finance and Accounting.

This research has been supported by the Czech Science Foundation Grant 18-05244S "Innovation Approaches to Credit Risk Management" and by the VSE institutional grant IP 100040.

optimization approaches. He et al. (2015) and Liu et al. (2019) model state transitions of loan accounts using Markov transitions matrices and determine the optimal action conditional on the state and time.

Surprisingly, there are not many applications of classical regression methods such as logistic regression or survival analysis. Murgia and Sbrilli (2012) test logistic regression against other methods in a collection decision support system. Thomas et al. (2016) develop a Markov chain model and a hazard rate model in order to study the impacts of different write-off strategies. Besides loan lending, Thomas et al. (2017) mention other fields where scoring card techniques can be used, such as pre-screening, preapproval, fraud prevention, and also debt recovery.

The key problem in the recovery process modelling can be formulated as a classical binary classification problem: Is the marginal effect of calling a debtor about the probability of the repayment of past due exposure positive or negative? Is the administrative and personal cost of the call offset by the increased recovery return? The marginal effect can be estimated using logistic regression or another binary classification technique. However, the binary classification set-up neglects the fact that, due to operational reasons, the calling rarely takes place immediately the exposure becomes past due, or it takes place with some fixed delay, and, in addition, there is the question of optimal timing. We propose to apply the survival analysis approach to solve the question of timing, and, at the same time, to handle the data when the calls historically took place at different times and the outcome of the recovery is often unknown (censored).

Survival analysis is a common statistical method from the medical and healthcare sectors (see Marubini & Valsecchi, 1995 or Collet, 2003), but has also found widespread use in credit risk (Witzany et al., 2012 or Witzany, 2017). One of the first uses of survival analysis in banking can be attributed to Narain (1992). Recently, a significant amount of research has been conducted using survival analysis in the area of credit risk modelling in banking (Cao et al., 2009, Hosmer et al., 2008, or Thomas et al., 2016).

We will focus on the survival analysis methods to estimate the probability of recovery repayments conditional on calling or not calling at a certain point in time. We will also briefly report the results of the logistic regression approach and discuss its main problems. The remainder of the paper is divided into the following parts: Section 2 describes the debt recovery process; Section 3 includes data description and data pre-processing and also discusses the logistic regression results; Section 4 examines the survival analysis methodology; Section 5 reports and discusses the models' results; Section 6 formulates and illustrates a soft collection optimization model in a case study; and finally, Section 7 provides conclusions.

## 2. Debt recovery process

Banks approach the recovery process on an individual basis, depending on their capabilities and experiences, but there are also many common features. Generally, there are two phases of the debt collection process: Early Collection and Late Collection. Different methods of recovery can be used in each of these two phases, such as phone communication, sending SMS and email messages, and even handing over the debt portfolio management to an external firm. Due to the still relatively high probability of repayments and the lower accrued penalty interest, the Early Collection phase is the more appropriate phase to use for scoring in the recovery process. The application of scoring to decide on the timing and types of actions in the Early Collection process could increase recovery process efficiency, reduce the extent of write-offs, and decrease the workload and staff costs.

In this analysis, we focus primarily on streamlining the recovery process by increasing the efficiency of telephone reminders. Although telephone reminders are a more expensive way of recovering compared to SMS and e-mail messages, they often have a much greater effect on the repayment rate of overdue receivables. The costs of telephone recovery include mainly the wage costs paid to employees, ICT costs and other expenses. If the average costs per one phone call are estimated, then the simple condition for making a call is that the expected marginal amount to be recovered should always be greater than the phone call costs.

## 3. Data description and logistic regression results

This analysis is based on partial empirical results from Kozina (2020) using a real dataset provided by a Czech bank for the period 2017 to 2019. The input data set contains 42,382 observations with more than 30 variables. The data represent historical records of past due retail exposures and debtors, including personal information, the characteristics of the products with which the debtors have become past due and entered the recovery process, and also whether and when telephone communication took place. It should be emphasized that the decisions to call were based on a relatively simple set of rules and their timing often depended on call center capacity. The dataset also contains the information whether and when the recovery process was successful, i.e. whether full recovery took place or not. Tables 1-3 list the variables from the data set used for the analysis.

***Table 1: Product information - explanatory variables***

| Variable name | Variable type | Variable description |
|---|---|---|
| prod_type_1 | Char(3) | Product classification by type (BU – current account small debits, BYV - mortgages, IU – investment loans, KK – credit cards, OVD – overdrafts, SU – consumer loans, TOD – current account unauthorized debits) |
| prod_type_2 | Char(2) | Product classification by type including segment classification |
| ovd_amount | Float | Overdue amount on the day of entry into collection process |
| branch_prod | Char(4) | Branch on which the product was based |
| dt_open | Date | Date of product creation |
| limit | Float | Product credit frame |
| card_risk_code | Char(1) | Risk group for credit cards |
| currency | Char(3) | Product currency |
| loan_status | Char(1) | Product status |
| int_rate | Float | Current interest rate |

***Table 2: Client personal information – explanatory variables***

| Variable name | Variable type | Variable description |
|---|---|---|
| dt_open_client | Date | Date of client's entry into the bank's portfolio. It is used to calculate the derived variable *exist_time* |
| employ_flag | Boolean | Job information |
| legal_form | Char(3) | Legal form of the client |
| main_owner_state | Char(2) | Owner's seat (only for legal entities) |
| country_code | Char(2) | The code of the country of residence of the client |
| resident_flag | Boolean | Resident information |
| risk_group | Double | Client risk group |
| age | Integer | Client age at the date of entry into overdue. |
| exist_time | Integer | Information about the client's total existence in the bank |
| cnb_class | Char(1) | Client classification. Applies only to credit products |
| boi_class | Char(1) | Client classification. Applies only to credit products |
| branch | Char(4) | Client's branch |
| segment_1 | Char(3) | Client segmentation |
| segment_2 | Char(2) | Client segmentation |

| | | |
|---|---|---|
| client_type | Char(2) | Client classification by type |

*Table 3: Recovery process information – explanatory and the target variables*

| Variable name | Variable type | Variable description |
|---|---|---|
| tel | Boolean | Information whether a telephone communication occurred (1 meaning a call, 0 meaning that there was no call) |
| dt_tel | Date | The date of the telephone reminder |
| tel_dpd | Integer | The number of days in overdue on the day of telephone communication |
| dluzi_od | Date | Date of entry into debt collection |
| max_dt | Date | Date of exit from debt collection due to repayment or going to the Late Collection phase (used to define the time variable in survival analysis) |
| **repaid** | Boolean | Information whether overdue receivables have been paid (**the target variable**) |

The standard logistic regression (Witzany, 2017) can simply be set up with the binary target variable "repaid" and with a selection of the explanatory variables including the binary variable *tel* indicating whether the collection call was made or not. We will show that it is more appropriate to use the survival analysis with the time dependent *tel* variable, while all the other variables remain constant.

In order to apply any of the techniques, it is important to preprocess the data, which includes, among other things, selection of only the statistically significant variables and elimination of those with low informative value. This can be achieved following the standard logistic scoring function development process (see e.g. Witzany, 2017) on the basis of a univariate Gini coefficient values, Weight of Evidence and variable Information Values (for categorical and binned numerical variables).

*Figure 1: Calculated univariate Gini coefficient values*

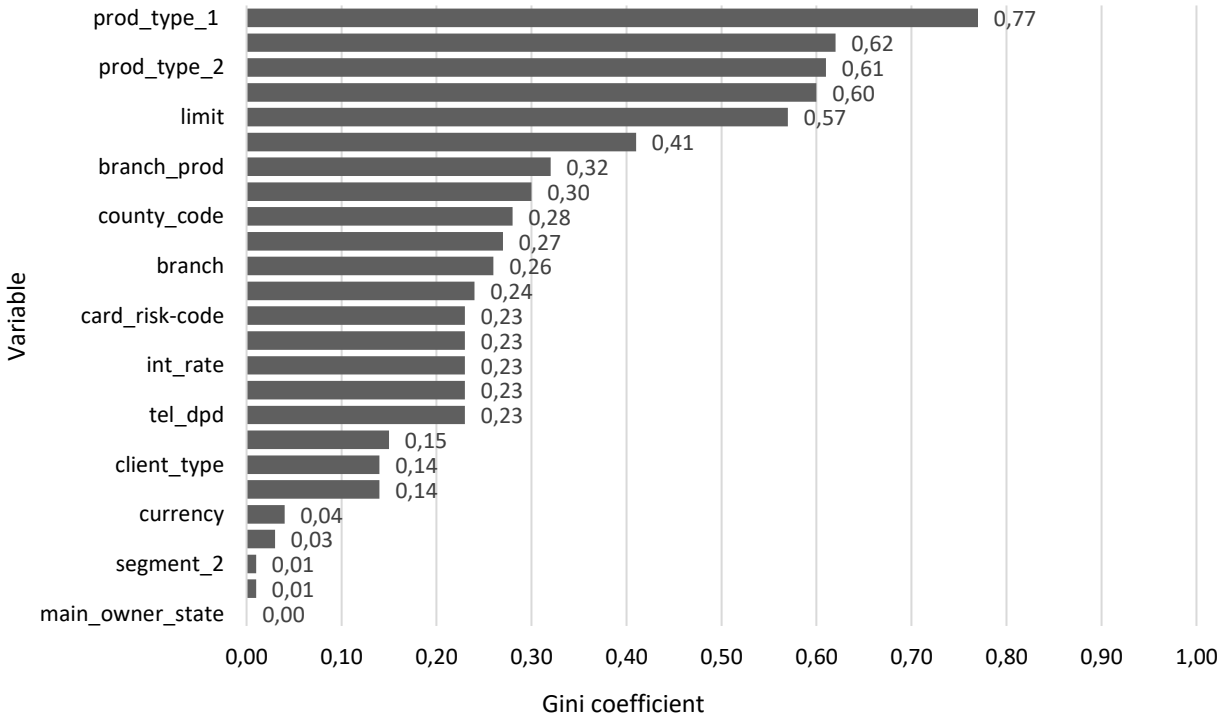



Figure 1 shows the univariate Gini coefficient values calculated for all relevant variables. Typically, the Gini coefficient should be greater than 10%, so we consider excluding such variables as

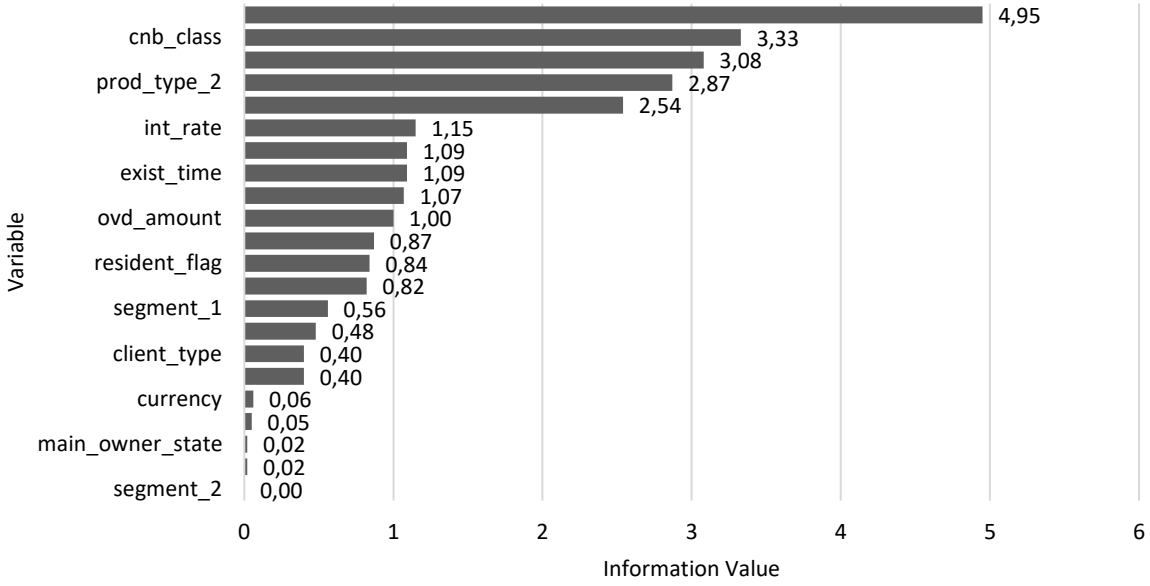
*main_owner_state*, *segment_2*, *employ_flag, risk_group* and *currency*. Note that the univariate Gini coefficient of the key variable *tel* is still above the 10% threshold, but relatively low. As we explain below, this measure underestimates the real effect of calling due to the issue of timing. The calls are made with a certain delay after the past due exposures enter the soft collection process, and so the debtors that repay very early are not usually called. In other words, the fact that a debtor is called already indicates that the initial probability of repayment (independently on the effect of the call itself) is lower.

Figure **2** also shows the calculated Information Values (IV) for all variables, with the numerical variables' quantities being divided into intervals and then converted into categorical values. In practice, only variables with $IV > 4\%$ are generally pre-selected, and so the variables *segment_2*, *risk_group*, *main_owner_state* and *employ_flag* can, indeed, be excluded from the model due to the low values of the Information Values. The same comment as above applies to the relatively low Information Value of the variable *tel*.

Based on the high Information Value and Gini coefficient of the variable *prod_type_1*, it appears obvious that the different types of products might have very different repayment behavior, which means that the scoring models should, rather, be developed for each product group separately. The difference in the repayment rates of the overdue amounts for specific banking products, depending on the *tel* variable, can also be seen in Figure 3.

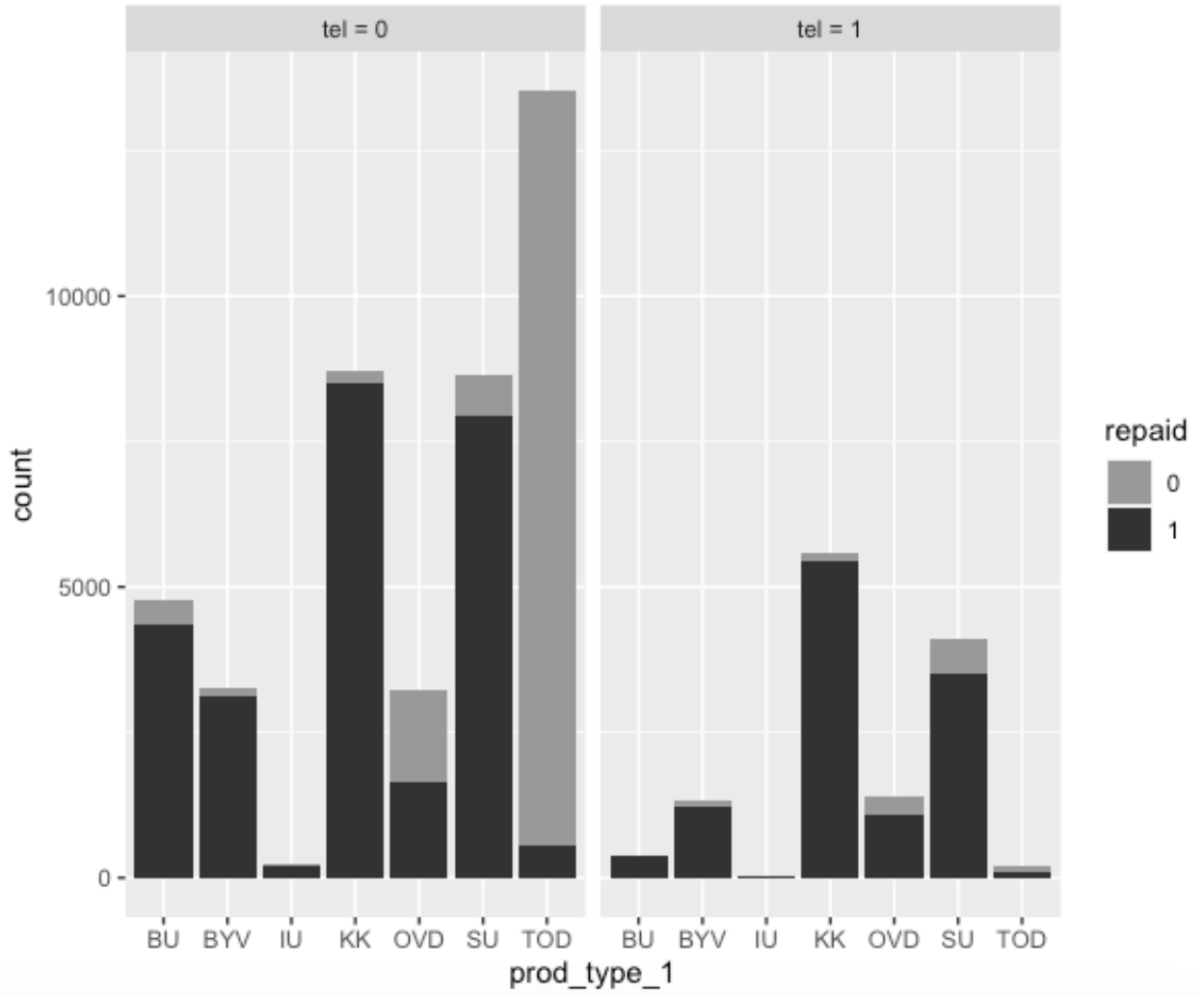
Next, we transform all the categorical variables (with the exception of *tel*) and bin the numerical variables back to continuous variables based on the Weight of Evidence (WoE) values (see Witzany, 2017). These continuous variables can be used to check for correlations in the subsequent regressions too, reducing the number of parameters to be estimated. It can be observed in **Chyba! Nenalezen zdroj odkazů.** (Annex) that there is a high correlation between the variables *client_type* and *legal_form*, between *prod_type_2* and *prod_type_1*, as well as between the variables *country_code* and *resident_flag*. The highly correlated variables should be eliminated in the variable selection process, so that at most one from each correlated group remains.

*Figure 2: Calculated Information values*




Finally, Table 4 reports the logistic regression coefficients estimated for the selected significant variables, with *repaid* being the target variable for the overall portfolio and individual product classes. All variables have been transformed using the WoE values (based on the full dataset) with the exception of the binary *tel* variable that takes only the values 0 (no call) and 1 (call). Hence, the estimated coefficient of the *tel* variable can be interpreted as the marginal effect of calling on the log-odds probability of repayment, and so it is expected to be positive. The coefficient is significant and positive for the logistic model with all products, as it is also for current account unauthorized debits (TOD), for overdrafts (OVD), credit cards (KK), and current account small debits (BU). For example, the coefficient $3.347$ estimated for TOD means that the multiplicative effect of calling on the repayment probability is quite high $\exp(3.347) \cong 28.4$. However, it is strange that the effect of calling is much smaller on the overall portfolio, and most importantly, the coefficient is significant and negative in the two important product classes: mortgages (BYV) and consumer loans merged with investment loans (SU+IU, the two products we merged due to a low number of observations in IU). That could be interpreted as a negative effect of calling. However, there is the problem of opposite causality – due to a delay in the decision to call, there are no calls to debtors who repay shortly after becoming past due. This problem does not directly apply to the other variables, whose values do not depend on the time in the soft collection process. To interpret correctly the coefficients of the other variables, we need to combine them with the WoE values (see **Chyba! Nenalezen zdroj odkazů.** in the annex).

*Figure 3: Reimbursement of overdue amounts depending on product types*




We have tried to resolve the problem of the incorrect *tel* coefficient by restricting the dataset only to cases where the call took place during a limited time interval. However, this leads to a significant reduction of observations where a call was made, and it is still not clear how to eliminate the bias for cases where a call was not made at all. The problem could be ideally resolved by designing an experiment with two groups of exposures having similar characteristics, where one group would be treated by being called in a defined time interval, and the other one would not be. Since we do not have this type of data, and it is generally expensive to get (in terms of the costs of such an experiment), we further focus on the survival modeling approach, which, in our opinion, provides a good solution to the problem.

*Table 4: Logistic regression coefficients – full dataset and selected products*

| Variables / Est. Coef. | All products | Current account unauthorized debits | Overdrafts | Mortgages | Credit cards | Consumer and Investment loans | Current account small debits |
|---|---|---|---|---|---|---|---|
| Intercept | 0.419*** (0.025) | -3.295*** (0.101) | -1.733*** (0.109) | 5.164*** (1.331) | 1.298*** (0.180) | 1.286*** (0.132) | 2.301*** (0.054) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Tel | 0.499*** (0.050) | 3.347*** (0.164) | 2.891*** (0.110) | -1.559*** (0.311) | 0.786*** (0.149) | -1.385*** (0.128) | 2.383*** (0.583) |
| prod_type_1_woe | 0.847*** (0.011) | - | - | - | - | - | - |
| exist_time_woe | 0.575*** (0.029) | 0.168* (0.066) | 1.409*** (0.083) | 0.615*** (0.165) | - | 4.210*** (1.242) | 0.215** (0.068) |
| ovd_amount_woe | -0.136*** (0.029) | -0.445*** (0.170) | 0.350*** (0.091) | -0.349* (0.155) | -0.470*** (0.099) | -0.312*** (0.066) | - |
| resident_flag_woe | 0.344*** (0.023) | 0.734*** (0.057) | 0.450*** (0.067) | - | - | - | 0.470*** (0.046) |
| age_woe | 0.101** (0.035) | - | 0.898*** (0.107) | - | - | 0.255** (0.093) | - |
| cnb_class_woe | 0.513*** (0.020) | - | - | 0.806*** (0.079) | 0.869*** (0.042) | 0.998*** (0.038) | - |
| limit_woe | -0.518*** (0.028) | - | - | -1.066* (0.494) | 0.196** (0.068) | - | - |
| segment_1_woe | 0.361*** (0.033) | 0.695*** (0.019) | - | - | - | 0.201* (0.096) | 0.353*** (0.062) |
| int_rate_woe | -0.497*** (0.029) | - | - | - | - | - | - |
| **Gini** | **0.930** | **0.598** | **0.784** | **0.750** | **0.600** | **0.763** | **0.350** |
| **AIC** | **18,119** | **3,711.7** | **2,710.7** | **746.61** | **1,675.1** | **3,439.7** | **2,701.5** |
| **Log-likelihood** | **-9,048.48** | **-1,849.86** | **-1,349.36** | **-367.30** | **-832.56** | **-1,713.85** | **-1,345.75** |

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

# 4. Survival analysis methodology

The goal of survival analysis is to estimate the probability distribution of the time of exit of an object conditional on a set of explanatory variables (see Hosmer et al., 2008). The distribution can be specified by the cumulative distribution function $F(t) = F(t|\mathbf{x})$ , or by the density function $f(t)$, or the survival function $S(t) = 1 - F(t)$, or by the hazard function $h(t) = f(t)/S(t)$. The classical examples of exit are the death of a patient or the breakage of a machine part, the default of an exposure, etc. In our case, it will be the event of repayment, with the time measured (in days) from the exposure entry in the soft collection process.

A dataset to estimate the model contains observations with explanatory variables, and not just with binary outcomes, but also with the time of exit outcomes. Another advantage of this class of models is that we can also use observations where the time of exit is censored. Here, we only know that the object has survived until a time limited by our observation window, which is the case of the soft collection repayment dataset (see

Table 3). In addition, we can also work with left censored observations, which can be useful for dealing with explanatory variables that change their value during the life of an object. A survival dataset can be used to calculate Kaplan-Meier empirical hazard and survival functions simply by counting the number of exits over a period (e.g. on daily basis) out of all cases alive.

There are two broad classes of survival analysis models: parametric models, where the hazard function or survival function have a parametric form with coefficients to be estimated, and semi-parametric models, where the shape of the hazard function is not specified (e.g. the Cox model). The simplest parametric model is the exponential model where the hazard $h(t) = \lambda$ is constant conditional on the explanatory variables, typically in the form $\lambda = \exp(\mathbf{x}'\boldsymbol{\beta})$. We will also test Accelerated Failure Time (AFT) models, which are characterized by a specific distribution of the log-time of exit, e.g. normal for the lognormal model, or logistic for the loglogistic model, etc.

The vector of coefficients $\boldsymbol{\theta}$ of a parametric model is estimated by maximizing the total log-likelihood function

$$LL(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln L(T_i, c_i, \mathbf{x}_i, \boldsymbol{\theta}), \tag{1}$$

where $L(T_i, c_i, \mathbf{x}_i, \boldsymbol{\theta})$ is the likelihood of the observation $i = 1, \dots, n$ with the time $T_i$ of exit or censoring indicated by $c_i \in \{0,1\}$ and with covariates $\mathbf{x}_i$. In the case of exit, the likelihood is $L = f(T_i; \mathbf{x}_i, \boldsymbol{\theta}) = h(T_i; \mathbf{x}_i, \boldsymbol{\theta})S(T_i; \mathbf{x}_i, \boldsymbol{\theta})$, while in the case of censoring it is just the survival probability $L = S(T_i; \mathbf{x}_i, \boldsymbol{\theta})$. If an observation is left censored from the time $T_{i,0}$ (which needs to be indicated by a left censoring variable), then the likelihood, defined as above, is simply divided by $S(T_{i,0}; \mathbf{x}_i, \boldsymbol{\theta})$.

We will start by estimating the Cox semi-parametric model where the hazard function shape is given a nonparametric baseline function $h_0(t)$ and $h(t) = h_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})$. This model belongs to the class of proportional hazard, where the coefficients $\boldsymbol{\beta}$ can be estimated by maximizing the partial log-likelihood function (assuming no ties)

$$PLL(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum A_{ij} \exp(\mathbf{x}_j'\boldsymbol{\beta})}\right), \tag{2}$$

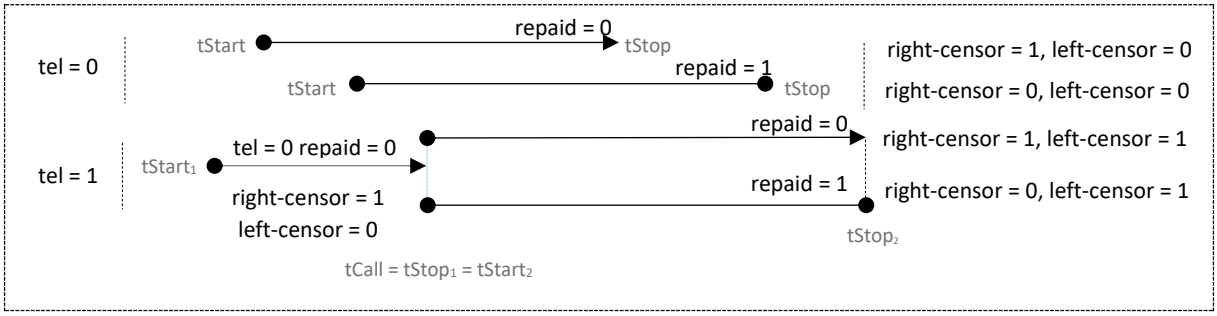
where $A_{ij}$ $(j = 1, \dots, n)$ is an indicator which takes the value 1 if the object $j$ is still at risk (alive) at the beginning of period $t_i$ and 0 otherwise. The baseline hazard function can then be directly calculated by the Breslow-Crowley estimator. Notice that the partial log-likelihood given by (2) is not the same as the one given by (1), even if the hazard functions are the same. Therefore, if want to compare the Cox model with a parametric one using the log-likelihood or Akaike criterion (AIC) consistently, we have to calculate the log-likelihood of the Cox model according to (1).

## 4.1 Survival dataset modification

The censor variable $c_i$ of an observation $i$ from our dataset described by Tables 1-3 is simply set to 0 if the exposure is repaid, and to 1 otherwise. The survival models can generally be estimated with time varying covariates. However, the standard implemented functions (e.g. the R functions `coxph` and `flexsurvreg`, which we will use) usually assume that the covariates are constant. In our case, for observations where the phone call took place, we are given just one time of the call $t_{call}$ and so we can set $tel = 1$ , starting on that date, and $tel = 0$ before $t_{call}$. In other words, we need to split the observation into two; one right-censored at $t_{call}$ , and the second left censored at $t_{call}$ . The time of exit and the censor of the second observation are the same as the original ones (see Figure 4 for an

illustration). Due to the splitting effect, the number of observations (row) in the survival dataset increases to 55 388.

***Figure 4: Left/right censoring definitions***

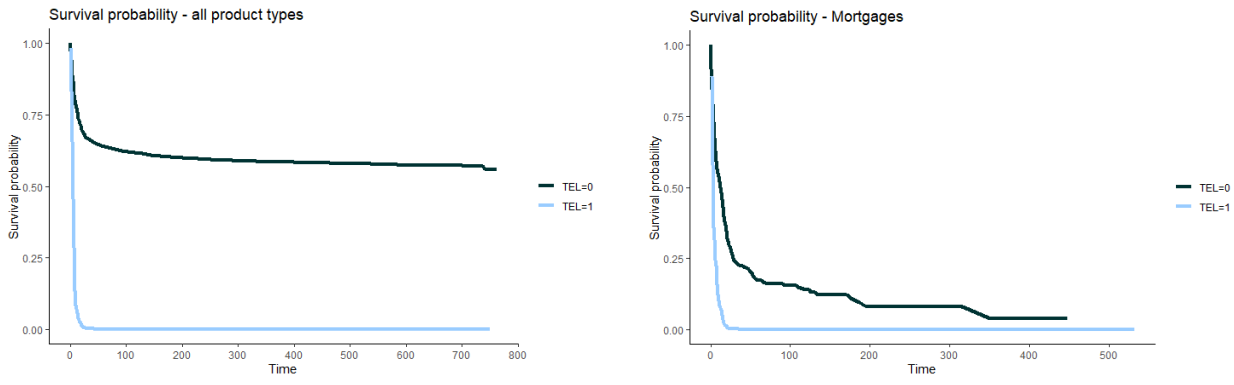



# 5. Empirical results

We are firstly going to estimate the Cox model and then several alternative parametric models on the full dataset (for all products) and separately for the specific product classes, as in the case of logistic regression. In order to select the best model, we will calculate and report the AIC for all models based on the log-likelihood (1), which differs from the partial log-likelihood given by (2). The coefficients of the Cox model, expressing the linear effects of covariates on the log-hazard independent of time, are easy to interpret, but the proportional hazard assumptions also need to be tested in order to decide between the Cox and the parametric models.

Figure 5 shows, within the framework of a preliminary inspection, the Kaplan-Meier empirical estimates of the survival, cumulative hazard, and hazard functions for the portfolio of all products and, separately, for mortgages conditional only on calling or not calling at the beginning of the collection process. It is obvious that the effect of calling on the survival probability, i.e. on the probability of repayment over time, is substantial. It seems to be larger in the case of the all-product portfolio than in the case of mortgages. The empirical cumulative hazard and hazard functions show that the effect of calling is larger in the days immediately following the call and diminishes over time, but more slowly in the case of mortgages.

***Figure 5: Kaplan-Meier estimates of the survival, cumulative hazard, and hazard functions conditional on calling or not calling at the beginning of the collection process for the dataset of all products and for Mortgages***



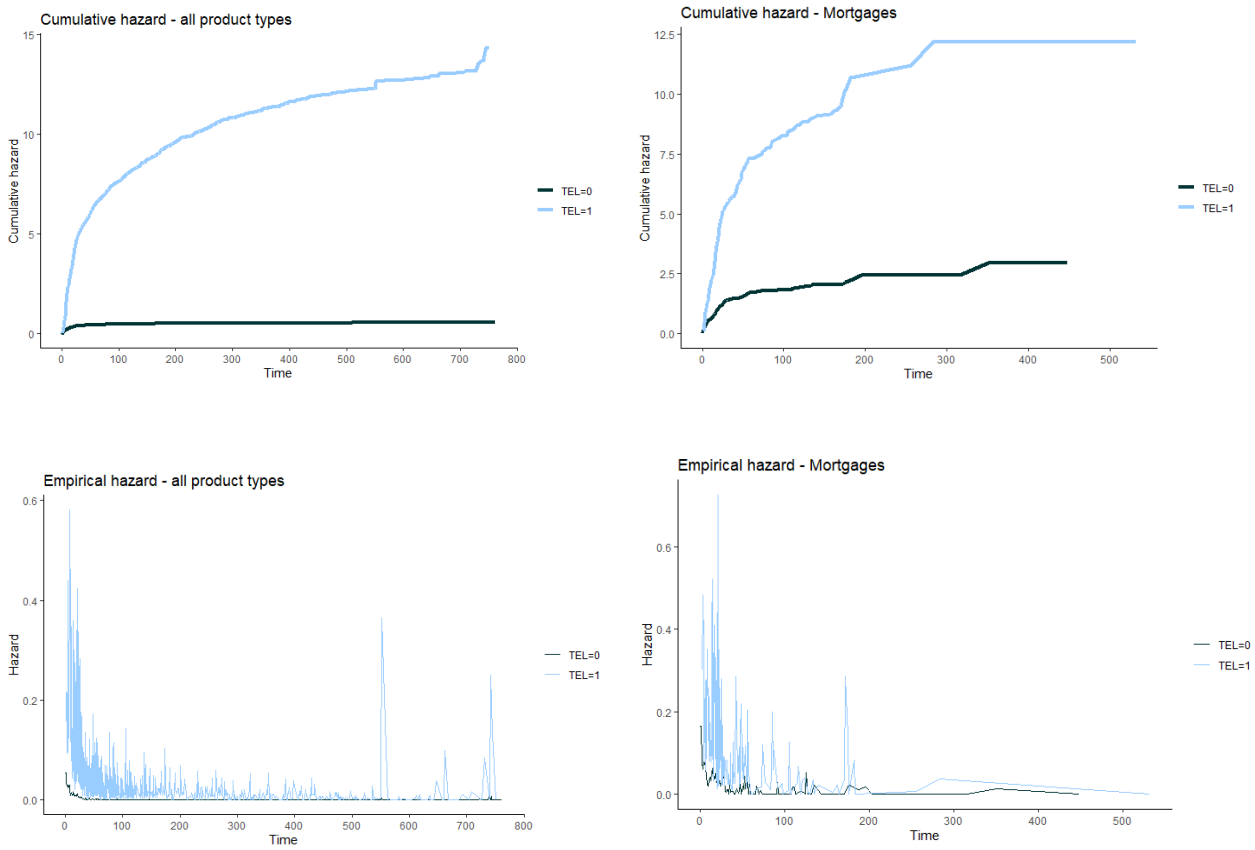



Table 5 shows the Cox model output for the entire input data set and individual products, and the impact of the explanatory variables on the hazard level. Notice that the coefficient of *tel* is now significant positive for all models and does not vary much. The overall multiplicative effect of calling, i.e. $tel = 1$, on the hazard function (for the all-product portfolio) is $\exp(2.223) \cong 9.2$. Practically, this means that the daily probability of repayment increases more then 9-times due to the call. Figure 6 shows the survival curves for an "average" exposure, i.e. with the WoE variables set to their mean values, in the overall and mortgage portfolios. Note that the curves cannot be directly compared to the Kaplan-Meier estimates in Figure 5, which are not conditional on the explanatory variables with the exception of *tel*. Nevertheless, the shapes and the effect of calling correspond to our expectations.

Regarding the impact of other variables, as in the case of logistic regression, to correctly interpret the output, we have to combine the estimated coefficients and the WoE values of the individual categories (**Chyba! Nenalezen zdroj odkazů.** – Annex). Most estimated coefficient are positive, in line with the expected impact of the WoE transformed variables, but in some cases the signs depend on the model. For example, the greater age of a debtor has a significantly positive impact on the probability of repayment in the case of overdrafts and mortgages, but the opposite effect in the case of consumer and investment loans.

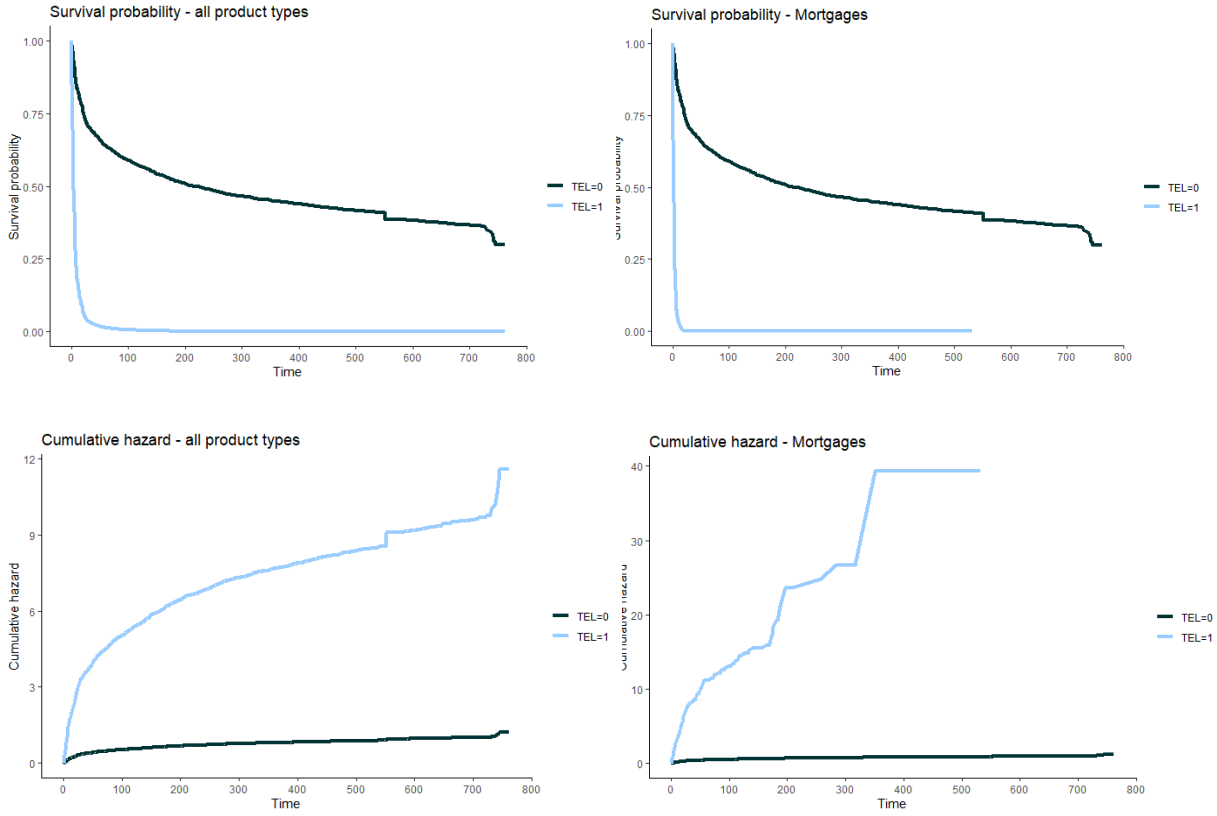
***Table 5: Cox model results***

| Variables / Est. Coef. | All products | Current account unauthorized debits | Overdrafts | Mortgages | Credit cards | Consumer and Investment loans | Current account small debits |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| tel | 2.223*** (0.016) | 4.209*** (0.036) | 3.865*** (0.059) | 1.758*** (0.050) | 2.694*** (0.028) | 1.752*** (0.031) | 1.273*** (0.056) |
| prod_type_1_woe | 0.516*** (0.007) | - | - | - | - | - | - |
| exist_time_woe | 0.119*** (0.008) | 0.205*** (0.058) | 0.620*** (0.037) | 0.182*** (0.028) | | -0.176*** (0.020) | 0.080*** (0.018) |
| ovd_amount_woe | -0.088*** (0.008) | -0.278* (0.148) | 0.132** (0.042) | -0.086*** (0.025) | -0.080*** (0.013) | -0.215*** (0.016) | - |
| resident_flag_woe | 0.150*** (0.013) | 0.714*** (0.056) | 0.218*** (0.043) | - | - | - | 0.182*** (0.019) |
| age_woe | - not.sign. | - | 0.355*** (0.067) | 0.200* (0.079) | | -0.118*** (0.027) | - |
| cnb_class_woe | 0.167*** (0.004) | - | - | 0.434*** (0.020) | 0.394*** (0.012) | 0.514*** (0.011) | - |
| limit_woe | -0.230*** (0.007) | - | - | - | 0.046*** (0.010) | - | - |
| segment_1_woe | 0.221*** (0.010) | 0.737*** (0.060) | - | 0.290*** (0.026) | - | 0.168*** (0.023) | 0.116*** (0.019) |
| int_rate_woe | -0.377*** (0.007) | - | - | - | - | - | - |
| **Likelihood ratio test** | **60, 656** | **1,222** | **5,663** | **2,043** | **13,184** | **7,077** | **586** |
| **Partial-log-likelihood** | **-237,557** | **-4,486** | **-9,673** | **-21,435** | **-62,218** | **-63,566** | **-33,085** |
| **Log-likelihood** | **-94,949** | **-4,071** | **-6,383** | **-9,094** | **-28,398** | **-27,012** | **-17,090** |
| **AIC (LL)** | **189,907** | **8,148** | **12,771** | **18,193** | **56,799** | **54,030** | **34,185** |

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

***Figure 6: Cox model survival and cumulative hazard curves for the all-product and mortgage portfolios conditional on tel = 0,1 and other explanatory variables set to the mean (WoE) values***




The survival data on which the Cox model is applied should also be tested for the proportional hazard assumptions. A simple visual way, for example, is to look at the proportion between the Kaplan-Meier

cumulative hazard function conditional on *tel* = 1 and the function conditional on *tel* = 0. If *tel* were the only relevant explanatory variable and the proportional hazard assumptions were valid, then the ratio would be constant. Figure 7 shows that this is weakly satisfied for the mortgage portfolio and hardly satisfied for the overall portfolio.

A more exact option for testing the proportional hazards assumption is to perform the Xue and Schifano, (2017) statistical test based on the Schoenfeld residuals. The test statistic is calculated for each covariate as well as for the entire model. The proportional hazard assumption is fulfilled when the relationships between Schoenfeld residuals and time are statistically insignificant. Table 6 shows the p-values based on the proportional-hazards statistics for each model calculated by the `cox.ph` function in R. The proportional hazard test in the case of *tel* is passed only for mortgages and the current account unauthorized debits portfolio. For other variables, the test is not passed in most cases.

***Figure 7: Proportions between Kaplan-Meier cumulative hazard functions conditional on tel = 1 and tel = 0 for the overall and mortgage portfolios***

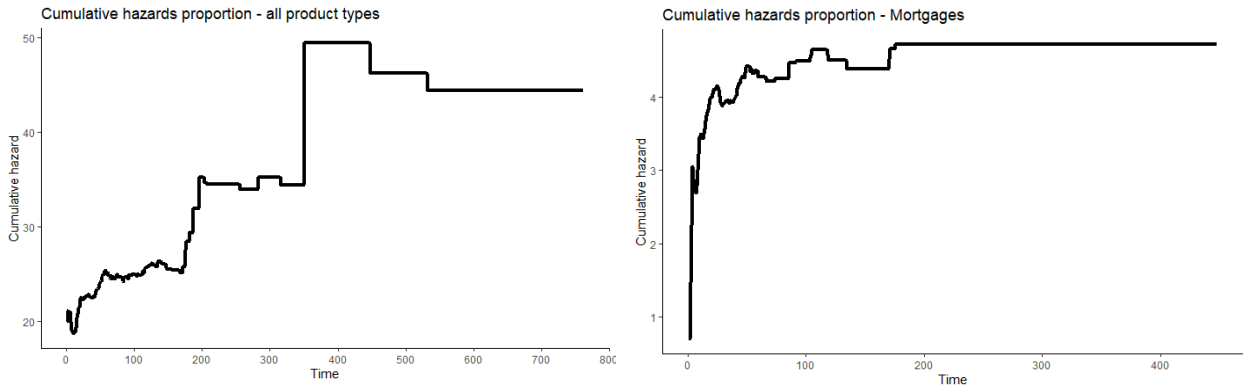



***Table 6: Test for the proportional-hazards assumption (p-values)***

| Variable / p-values | All products | Current account unauthorized debits | Overdrafts | Mortgages | Credit cards | Consumer and Investment loans | Current account small debits |
|---|---|---|---|---|---|---|---|
| tel | 2.7e-13 | 0.747 | < 2e-16 | 0.829 | 2.8e-16 | 1.6e-07 | 1.0e-11 |
| prod_type_1_woe | 2e-16 | - | - | - | - | - | - |
| exist_time_woe | 1.9e-11 | 0.095 | < 2e-16 | 0.002 | - | 0.042 | 0.002 |
| ovd_amount_woe | 0.0616 | 0.084 | < 2e-16 | 6.6e-06 | 1.1e-14 | 0.007 | |
| resident_flag_woe | 0.1821 | 0.323 | 0.49 | - | | - | 0.387 |
| age_woe | 7.5e-16 | - | 9.6e-10 | 0.095 | - | 0.602 | |
| cnb_class_woe | 0.0025 | - | - | 3.8e-07 | 1.5e-07 | < 2e-16 | |
| limit_woe | < 2e-16 | - | - | - | 0.49 | - | |
| segment_1_woe | 0.0121 | 0.511 | 8.5e-15 | 8.7e-10 | - | 1.5e-05 | 0.113 |
| int_rate_woe | < 2e-16 | - | - | - | - | - | - |
| **Global** | **< 2e-16** | **0.096** | **< 2e-16** | **< 2e-16** | **< 2e-16** | **< 2e-16** | **9.6e-13** |

Since the Cox model does not satisfy the proportional hazard tests very well, we shall estimate and compare several standard parametric survival models, namely the lognormal, Weibull, and loglogistic models. In order to select the best fitting model, we have calculated the Akaike criterion (AIC) shown in Table 7. The best AIC values are given by the lognormal model for all the product classes. In addition, the optimization algorithm did not converge for some of the products in the case of the Weibull and loglogistic models, and so the AIC value is not shown. It is interesting to note that the lognormal model AIC is substantially larger than the Cox model log-likelihood AIC reported in Table 5. However, the Cox model AIC calculation uses only the conventional number of degrees (number of parameters) and does not include a penalization term for the non-parametric baseline hazard function that might have an overfitting effect. Therefore, due to the proportional hazard test results, we would recommend selecting the lognormal model for most products with the possibility of the Cox model being applied to the current account unauthorized debits product class and to mortgages.

***Table 7: AIC values for the selected parametric models and product classes***

| Distribution / AIC value | All products | Current account unauthorized debits | Overdrafts | Mortgages | Credit cards | Consumer and Investment loans | Current account small debits |
|---|---|---|---|---|---|---|---|
| lognormal | 217,755 | 8,770 | 14,527 | 20,775 | 64,416 | 60,638 | 39,256 |
| Weibull | 223,386 | 8,825 | 15,148 | - | - | - | 40,526 |
| loglogistic | 222,995 | - | 15,278 | 21,240 | 67,480 | 61,383 | 39,325 |

The lognormal model coefficients estimated for the overall portfolio and individual product classes are reported in Table 8. In this case, the interpretation of the parameters is not as straightforward as in the case of the Cox model. In the lognormal model the log-time of exit is normally distributed with the mean $\mu(\mathbf{x}_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}$, where *meanlog* given in Table 8 is the intercept $\beta_0$, and with the standard deviation $\sigma$ given by *sdlog*. Therefore, the negative coefficients of $tel$ in all models indeed significantly reduce the expected time of repayment as expected. The effect of calling (at the beginning of the soft collection process) on the survival and hazard functions of products with mean covariate variables is illustrated in Figure 8. It is interesting to note that the coefficient of *tel* is (in absolute value) larger than the intercept *meanlog* for most products (with the exception of mortgage and consumer loans portfolios). This means that the action of calling has a very significant and fast effect on the intensity of repayments (hazard), possibly exaggerating the expectation if compared with the Kaplan-Meier curves (Figure 5).

***Table 8: Lognormal model results***

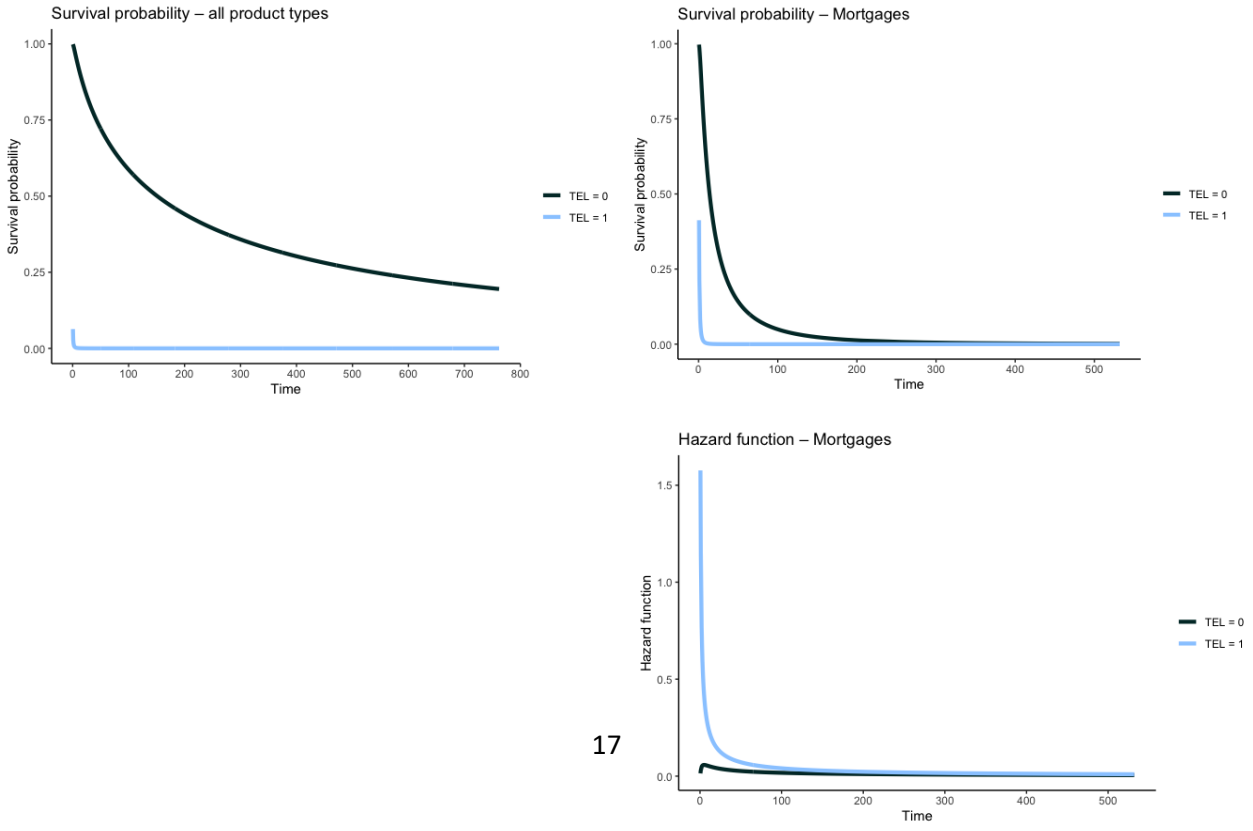
| Variable / Est. Coef. | All products | Current account unauthorized debits | Overdrafts | Mortgages | Credit cards | Consumer and Investment loans | Current account small debits |
|---|---|---|---|---|---|---|---|
| meanlog | 5.464*** (0.019) | 14.10*** (0.452) | 7.870*** (0.074) | 4.704 (0.255) | 6.793*** (0.092) | 5.564*** (0.053) | 2.982*** (0.025) |
| sdlog | 1.881*** (0.011) | 4.810*** (0.196) | 2.470*** (0.024) | 1.198*** (0.020) | 1.598*** (0.017) | 1.319*** (0.014) | 1.657*** (0.019) |
| tel | -8.585*** (0.128) | -24.2*** (2.23) | -15.6*** (0.071) | -3.586*** (0.172) | -8.108*** (0.186) | -3.251*** (0.098) | -3.233*** (0.223) |
| prod_type_1_woe | -0.871*** (0.009) | - | - | - | - | - | - |
| exist_time_woe | -0.354*** (0.016) | -0.538*** (0.150) | -2.37*** (0.011) | -0.293*** (0.034) | - | 0.208*** (0.026) | -0.147 (0.863) |
| ovd_amount_woe | 0.203*** (0.016) | 1.46*** (0.453) | - | 0.044*** (0.030) | - | 0.333*** (0.022) | - |

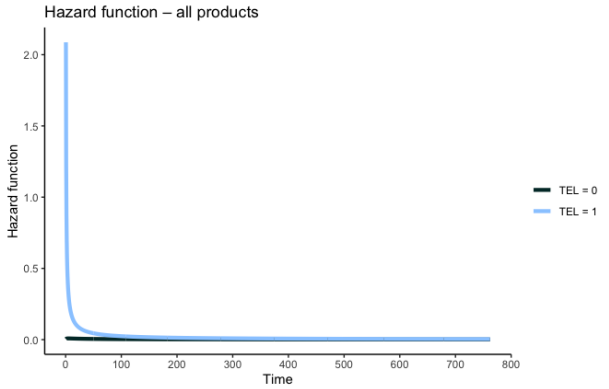
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| resident_flag_woe | -0.332*** (0.020) | -1.67*** (0.131) | -0.891*** (0.029) | - | - | - | -0.286*** (0.029) |
| age_woe | 0.083*** (0.025) | - | -1.99*** (0.029) | 0.460*** (0.093) | - | 0.330*** (0.036) | - |
| cnb_class_woe | -0.329*** (0.009) | - | - | -0.710*** (0.026) | -0.811*** (0.019) | -0.880*** (0.015) | - |
| limit_woe | 0.384*** (0.014) | - | - | 0.112*** (0.092) | -0.133*** (0.024) | - | - |
| segment_1_woe | -0.414*** (0.018) | -1.66*** (0.145) | - | -0.446*** (0.032) | - | -0.119*** (0.026) | -0.191*** (0.030) |
| int_rate_woe | 0.872*** (0.014) | - | - | - | - | - | - |
| **AIC** | **217,754.8** | **8,769.97** | **14,527.19** | **20,774.53** | **64,416.36** | **60,637.66** | **39,255.65** |
| **Log-likelihood** | **-108,865.4** | **-4,377.985** | **-7,257.594** | **-10,378.26** | **-32,203.18** | **-30,310.83** | **-19,621.83** |

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The other variables in terms of the effect on the probability of repayment can be split into two groups. The direction of the effect (see also **Chyba! Nenalezen zdroj odkazů.** in the annex) of *resident_flag_woe*, *cnb_class_woe*, *segment_1_woe*, and *ovd_amount_woe* is the same in all models where the variables are significant and in line with our expectation (being a resident, better CNB classification, affluent or larger company segment, or a smaller overdraft balance, all have positive effects on repayments). The direction of the effect of the other variables is mixed, depending on the product portfolio. For example, the effect of a longer time with the bank (*exist_time_woe*) is mostly positive, with the exception of the consumer and investment loans portfolio; the effect of greater age (*age_woe*) is also positive, again with the exception of the consumer and investment loans portfolio; on the other hand the effect of a larger credit limit (*limit_woe*) is positive in the case of credit cards but negative in the case of mortgages and on the overall portfolio. Finally, the product interest rate variable (*int_rate_woe*) turns out to be significant only on the overall portfolio, with the negative effect of higher interest rates on repayments, as expected. The same conclusions, in terms of the directional effects of the individual variables, can be drawn from the Cox model results (Table 5).

***Figure 8: Lognormal model survival and cumulative hazard curves for the all product and mortgage portfolios conditional on tel = 0,1 and other explanatory variables set to the mean (WoE) values***






## 6. Soft collection process optimization

In this section, we are firstly going to formulate a relatively simple theoretical approach towards the use of the estimated survival models in order to optimize the timing and the decision to call or not to call a debtor, which is part of the soft collection process. We will illustrate the method based on particular exposures and our estimated models.

Let us assume that we are given the survival functions

$$S^0(t) = S(t|\boldsymbol{x}_i, tel = 0 \; at \; time \; 0) \text{ and}$$

$$S^1(t) = S(t|\boldsymbol{x}_i, tel = 1 \; at \; time \; 0)$$

for a particular defaulted exposure $i$ with factors $\boldsymbol{x}_i$ and conditional on calling or not calling at time zero. Then we also know the cumulative distribution function $F(t) = 1 - S(t)$, the density function $f(t) = F'(t)$, and the hazard function $h(t) = f(t)/S(t)$ conditional on calling or not calling. Since generally $S(t) = -\int_0^t h(s)ds$, we can use the two survival functions to express the survival probability conditional on calling or not calling at time $t$,

$$S^0(t_1|tel = 0 \; at \; t) = -\int_t^{t_1} h^0(s)ds = \frac{S^0(t_1)}{S^0(t)},$$

and

$$S^1(t_1|tel = 1 \; at \; t) = -\int_t^{t_1} h^1(s)ds = \frac{S^1(t_1)}{S^1(t)}.$$

Assume that the soft collection process optimization problem has the following parameters:

- We want to make the decision to call or not to call for $t \in \{0, \dots, t_0\}$.
- We assume that the impact of calling is limited to $T_0$ days from the day of the call; after this period the exposure leaves soft collection.
- If the debtor does not repay and leaves the soft collection process, then the loss ratio is $l$, i.e. if $E$ is the exposure, the total loss in the case of unsuccessful soft collection is $l \times E$, and the complementary recovery rate $RR = 1 - l$.
- For simplicity, we do not take discounting into account, but the formulas could be easily modified to include it.

- The cost of making a call is $C_{TEL}$ in absolute terms (depending namely on the personal cost, the average time spent on calling, technical and communications costs, etc.)

In order to decide whether to call or not to call at time $t$ the bank needs to look at the expected net cash flow conditional on $tel = 0,1$:

$$E[CF|tel = 0] = (F^0(t + T_0|t) + S^0(t + T_0|T) \times RR) \times E = F^0(t + T_0|t) \times l \times E + RR \times E,$$

and similarly

$$E[CF|tel = 1] = F^1(t + T_0|t) \times l \times E + RR \times E - C_{TEL}.$$

Therefore, the difference between the two expected net cash flows is

$$E[CF|tel = 1] - E[CF|TEL = 0] = \left(F^1(t + T_0|t) - F^0(t + T_0|t)\right) \times l \times E - C_{TEL}.$$

If the time $t$ was the only point at which to decide whether to call or not, then the condition to make the call is:

$E[CF|tel = 1] > E[CF|TEL = 0]$, i.e.

$\left(F^1(t + T_0|t) - F^0(t + T_0|t)\right) \times l \times E > C_{TEL}$.

If we want to find an optimal timing $t \in \{0, \dots, t_0\}$ from the perspective of time 0 then we should consider the possibility of repayment until $t$ without calling, i.e. we compare the expected cash flow when we do not call at all but wait till $t + T_0$:

$$E_0[CF|tel = 0] = (F^0(t) \times l + RR) \times E + S^0(t)(F^0(t + T_0|t) \times l + RR) \times E,$$

and the expected cash flow when we make the call at time $t$

$$E_0[CF|tel = 1\ at\ t] = (F^0(t) \times l + RR) \times E + S^0(t)\left(F^1(t + T_0|t) \times l + RR - \frac{C_{TEL}}{E}\right) \times E.$$

then we just need to find $t = t_{opt}$ maximizing

$$\begin{aligned} &\arg_{t_{opt}=t} \max(E_0[CF|tel = 1\ at\ t] - E_0[CF|tel = 0]) \\ &\qquad = \arg_{t_{opt}=t} \max\left[S_0(t)\left((F^1(t + T_0|t) - F^0(t + T_0|t) - \frac{C_{TEL}}{E \times l}\right)\right] \end{aligned} \tag{3}$$

If the maximum is negative, then no call is made. This is the case, for example, when $E \times l \leq C_{TEL}$, i.e. the exposure and/or the loss rate are relatively small compared to the cost of calling. The optimality condition also allows us to determine the minimum exposure from which it makes sense to consider calling the debtor: $E_{min} = \frac{C_{TEL}}{l \times \max_t(F^1(t+T_0|t) - F^0(t+T_0|t)}$, provided $\max_t(F^1(t + T_0|t) - F^0(t + T_0|t) > 0$.

Therefore, the optimal strategy that can be setup at the beginning of the soft collection process is as follows: if $E \leq E_{min}$ then do not call the debtor at all. If $E > E_{min}$ and if the debtor repays by day $t_{opt}$ no call is made. If the exposure is unpaid at $t_{opt}$ then the call is made. If the exposure has still not been repaid at $t_{opt} + T_0$ then it should be transferred into the next phase of the collection process.

## 6.1 Case study

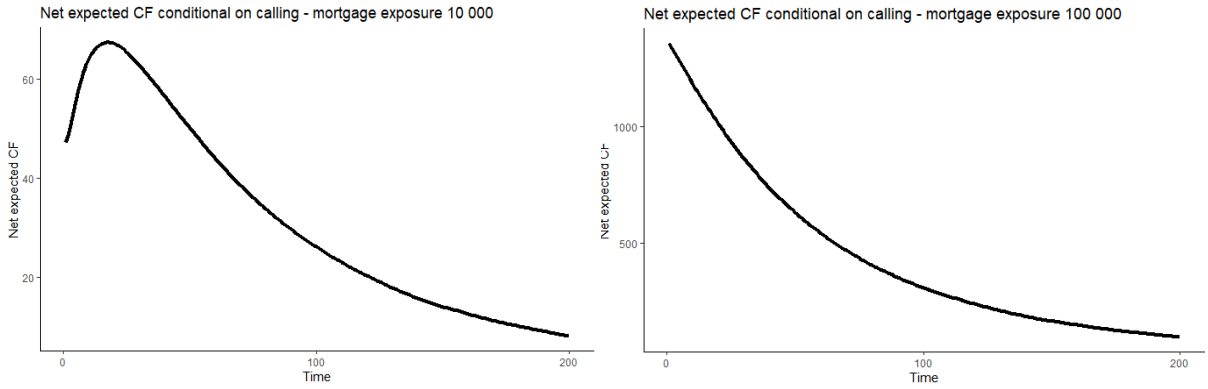
To illustrate the optimization methodology, we will consider three representative mortgage, consumer loan and current account unauthorized debit exposures with all covariates (besides *tel*) set to the mean

WoE values. We assume that the cost of calling is $C_{TEL} = 100$ and the collection time horizon $T_0 = 100$. For mortgages, we will use a lower loss rate $l = 30\%$, while for the consumer loans and unauthorized debits the loss rate will be $l = 70\%$. Using the estimated survival models, we want to make decisions on whether it makes sense to call, and, if so, what time to call the debtor, within the horizon of the next $t_0 = 150$ days.

In the case of the mortgage exposure, based on the estimated lognormal model and the outlined methodology, it turns out that it makes sense to make a call for exposures above 1400 ($E_{min} = 1316$). Nevertheless, the decision to call should be postponed for smaller exposures; for example, if $E = 2\,000$, then the optimal time of calling given by (3) is $t_{opt} = 101$, while if $E = 10\,000$ , then the optimal time of calling is $t_{opt} = 18$, and if the exposure $E = 100\,000$ , then the call should be made immediately after the exposure enters the collection process (see

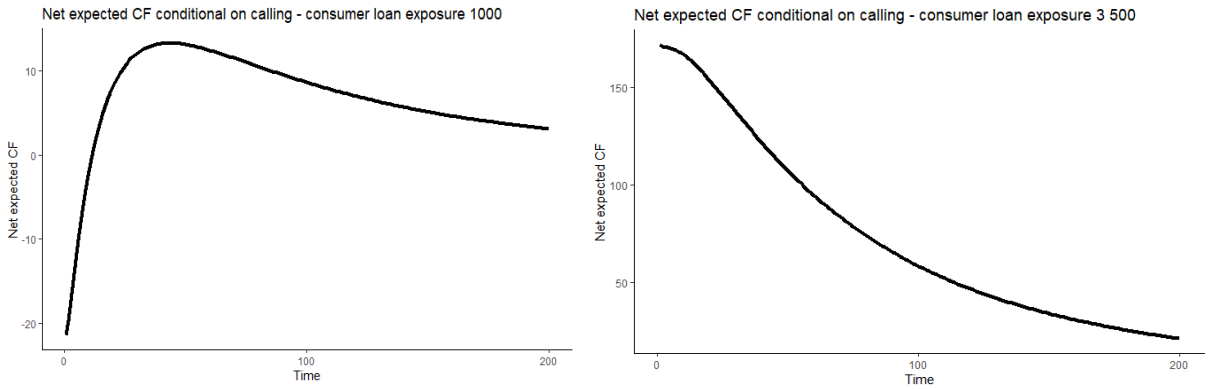
**Figure 9**).

***Figure 9: Net expected cashflow effect of a call given by (3) for two representative mortgage exposures***



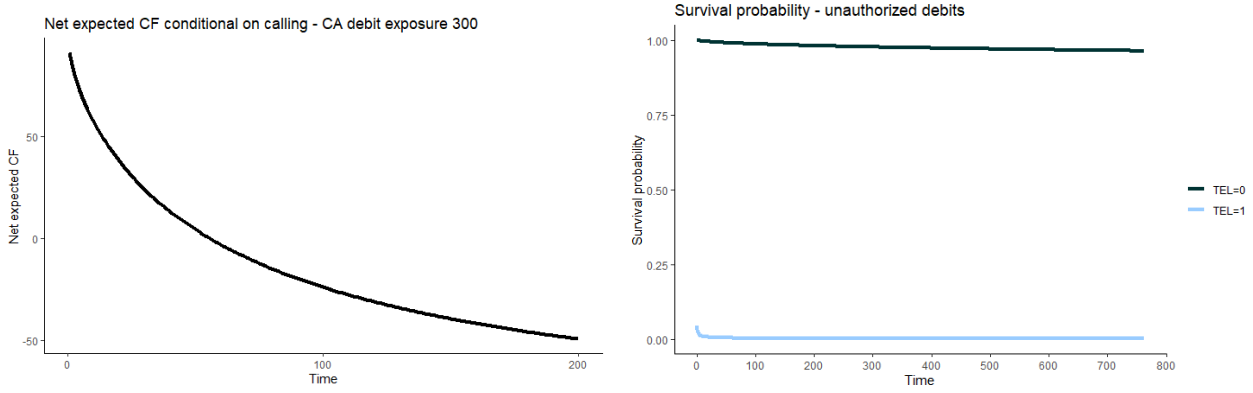

The second case will be a consumer loan, again with covariates set to the mean (WoE) values, and with different exposure amounts. Based on the lognormal model estimated on the portfolio of consumer and investment loans, we arrive at a conclusion similar to the one in the case of mortgages. It does not pay to make a call in the case of exposure roughly below 550 ($E_{min} = 555$). The call should be made for larger exposures, but for exposures below approximately 3000 it should not be made immediately. For example, if $E = 1000$ , then the optimal timing $t_{opt} = 44$ (see Figure 10).

*Figure 10: Net expected cashflow effect of a call for two representative consumer loan exposures*




Finally, the third case will be an unauthorized debit exposure, again with covariates set to the mean (WoE) values, and with different exposure amounts. In this case, the minimum exposure to call is $E_{min} = 157,$ and the call should always be made on the day the exposure first enters soft collection. This can be explained by the very low level of the repayment rates in the case of no calling, and so the advantage of postponing the call is relatively small (see Figure 11).

*Figure 11: Net expected cashflow of an unauthorized debit and the survival curves*




# 7. Conclusion

We have tested the logistic regression and several survival models in order to estimate the effect of calling a debtor in terms of the probability of the repayment of an exposure in the soft collection process. The ultimate goal was to propose and implement a methodology optimizing the decisions and timing of calls to debtors with exposures in the soft collection process. The models were estimated on a relatively large dataset of retail defaulted loans from the period 2017-2019. The dataset contained information about the dates of entry, calling (if any), and exit from the soft collection process, indicators of repayment and a number of demographic and behavior variables such as age, resident flag, existing time in the bank, etc. We have performed a univariate analysis and preselected the most important variables. The categorical values were replaced by the WoE evidence values, with the exception of the *tel* (call) indicator. The analysis has shown different repayment patterns for individual products and a need to estimate the models separately for individual products rather than for the overall portfolio.

The results of the standard logistic regression demonstrated that this type of model is not appropriate, due to the issue of call timing. The calls were, in this case, made based on call center capacity and simple prioritization rules, and so smaller and quickly repaid exposures were not usually called. This interdependence provides an explanation for the unexpected signs of the *tel* indicator for some of the products, and an argument for applying a survival analysis, in which the call can be compared to a medical treatment. The estimated Cox and selected parametric models have shown the consistent directional effects of the main *tel* variable as well as of the other explanatory variables. Regarding the choice of model, the conclusions are mixed. Based on the log-likelihood Akaike Information Criterion (AIC) we would recommend the Cox model, but its violation of the proportional hazard assumptions for most of the products led us to prefer the lognormal model, which has the best AIC values among the various parametric models considered.

Finally, based on the estimated survival models, we have proposed a straightforward optimization methodology which enables decisions to be made on optimal timing and the minimum exposure to make a call. The methodology was illustrated in a case study on representative mortgage, consumer loans, and unauthorized debit with different exposures and loss rates. Other important parameters of the optimization exercise are the estimated overall cost of one call and the collection process time horizon. In line with our intuition, the results have shown that, for relatively small exposures, calling does not pay back its costs. For medium exposures a call should be made, but usually later, if the debtor does not pay without being called. For relatively large exposures the call should be made immediately the exposure enters the collection process.

# 8. Literature


Cao R, Vilar JM, Devia A (2009). Modelling consumer credit risk via survival analysis. SORT 33(1): 3–30.

Chehrazi, N., & Weber, T. A. (2015). Dynamic valuation of delinquent credit-card accounts. Management Science, 61(12), 3077-3096.

Chehrazi, N., Glynn, P. W., & Weber, T. A. (2019). Dynamic credit-collections optimization. Management Science, 65(6), 2737-2769.

Collet D. (2003). Modelling Survival Data in Medical Research, Chapman & Hall/CRC, London.

De Almeida Filho, A. T., Mues, C., & Thomas, L. C. (2010). Optimizing the collections process in consumer credit. Production and Operations Management, 19(6), 698-708.

He, P., Hua, Z., & Liu, Z. (2015). A quantification method for the collection effect on consumer term loans. Journal of Banking & Finance, 57, 17-26.

Hosmer DW, Lemeshow S, May S. (2008). Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition, John Wiley & Sons, Chichester, United Kingdom

Kozina A. (2020). Využití skóringu v procesu vymáhání pohledávek. Diploma Thesis, University of Economics, Prague

Liu, Z., He, P., & Chen, B. (2019). A Markov decision model for consumer term-loan collections. Review of Quantitative Finance and Accounting, 52(4), 1043-1064.

Marubini E, Valsecchi MG. (1995). Analysing Survival Data from Clinical Trials and Observational Studies, John Wiley & Sons, Chichester, United Kingdom.

Murgia, G., & Sbrilli, S. (2012). A decision support system for scoring distressed debts and planning their collection. In Methods for Decision Making in an Uncertain Environment (pp. 69-89).

Narain, B. (1992) Survival Analysis and the Credit Granting Decision. In: Thomas, L.C., Crook, J.N. and Edelman, D.B., Eds., Credit Scoring and Credit Control, OUP, Oxford, 109-121.

So, M. C., Mues, C., de Almeida Filho, A. T., & Thomas, L. C. (2019). Debtor level collection operations using Bayesian dynamic programming. Journal of the Operational Research Society, 70(8), 1332-1348.

Thomas, L., Crook, J., & Edelman, D. (2017). Credit scoring and its applications. SIAM.

Thomas, L. C., Matuszyk, A., So, M. C., Mues, C., & Moore, A. (2016). Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. European Journal of Operational Research, 249(2), 476-486.

Van de Geer, R., Wang, Q., & Bhulai, S. (2018). Data-driven consumer debt collection via machine learning and approximate dynamic programming. Available at SSRN 3250755.

Witzany, J. (2017). Credit Risk Management. Springer Books.

Witzany, J., Rychnovský, M., & Charamza, P. (2012). Survival Analysis in LGD Modeling. European Financial and Accounting Journal, 2012(1), 6-27.

Xue, Y., & Schifano, E. D. (2017). Diagnostics for the Cox model. Communications for statistical Applications and Methods, 24(6), 583-604.

# 9. Appendix

***Table 9: Correlation matrix***

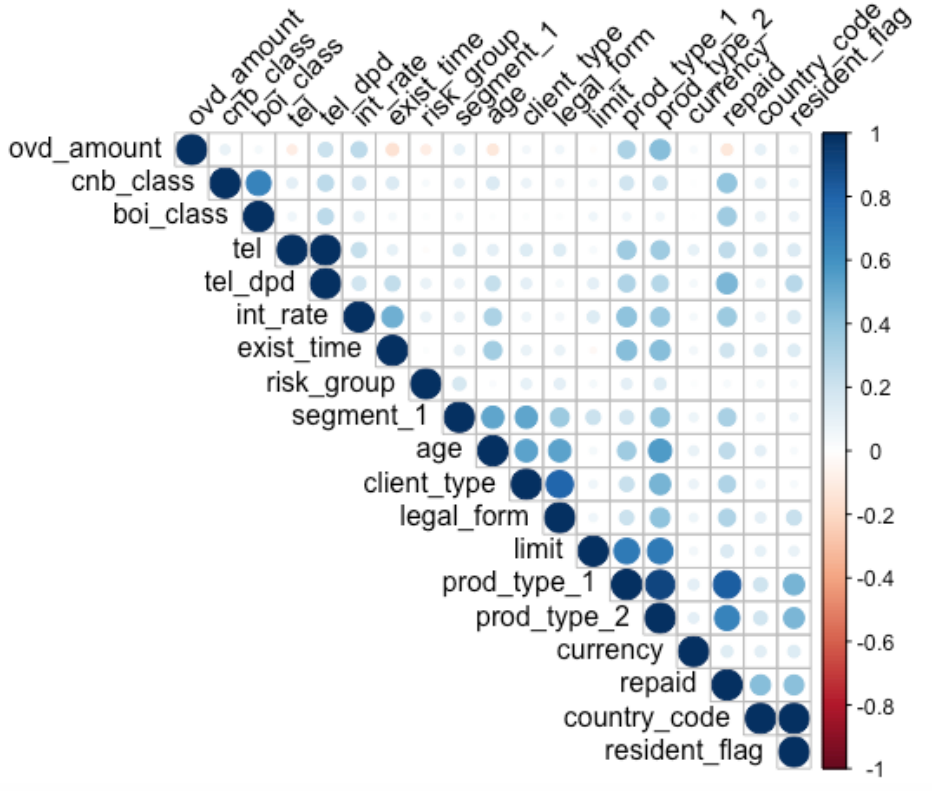


*Table 10: Weight of evidence and coarse classification overdue*

| Variables | Category | WOE | Impact |
|---|---|---|---|
| ***tel*** | *Yes - 1* | 1.000 | + |
| | *No - 0* | 0.000 | - |
| ***prod_type_1*** | *credit cards* | 3.097 | + |
| | *mortgages* | 2.830 | + |
| | *current accounts* | 1.823 | + |
| | *personal loans* | 1.953 | + |
| | *overdraft account* | -0.469 | - |
| | *current account unauthorized debits* | -3.663 | - |
| | *investment loans* | 2.145 | + |
| ***cnb_class*** | *1* | 3.919 | + |
| | *2* | 2.083 | + |
| | *3* | 0.878 | + |
| | *4* | 0.003 | + |
| | *5* | -0.235 | - |
| | *missing* | -1.325 | - |
| ***limit*** | *(-Inf; 50,000)* | -0.603 | - |
| | *[50,000;750,000)* | 2.068 | + |
| | *[750,000; Inf)* | 2.734 | + |
| ***exist_time*** | *(0; 5)* | -0.625 | - |
| | *[5; 10)* | 0.300 | + |
| | *[10; 15)* | 0.883 | + |
| | *[15; Inf)* | 1.704 | + |
| ***ovd_amount*** | *(-Inf; -12,000)* | 1.731 | + |
| | *[-12,000; -4,000)* | 0.454 | + |
| | *[-4,000; -2,000)* | 1.097 | + |
| | *[-2,000; 0)* | -0.555 | - |
| ***resident_flag*** | *Yes* | 0.408 | + |
| | *No* | -2.196 | - |
| ***age*** | *(0; 20)* | -2.018 | - |
| | *[20; 30)* | -0.299 | - |
| | *[30; 34)* | 0.038 | + |
| | *[34; Inf)* | 0.425 | + |
| ***segment_1*** | *Small enterprises low, Real estate clients, Small business core* | -2.127 | - |
| | *Freelancers, Mass market, Small enterprises high, Middle corporate clients, Retail other, Private low (3-10M), Private high (>10M)* | 0.051 | + |
| | *Affluent, Multinational corporate clients, Large corporate clients* | 1.780 | + |
| ***int_rate*** | *missing* | -0.375 | - |
| | *(-Inf,18.96)* | 2.578 | + |
| | *[18.96,23.4)* | 3.362 | + |
| | *[23.4, Inf)* | 3.429 | + |

## FFA Working Paper Series

2019

1. Milan Fičura: Forecasting Foreign Exchange Rate Movements with k-Nearest-Neighbor, Ridge Regression and Feed-Forward Neural Networks

2020

1. Jiří Witzany: Stressing of Migration Matrices for IFRS 9 and ICAAP Calculations
2. Matěj Maivald, Petr Teplý: The impact of low interest rates on banks' non-performing loans
3. Karel Janda, Binyi Zhang: The impact of renewable energy and technology innovation on Chinese carbon dioxide emissions
4. Jiří Witzany, Anastasiia Kozina: Recovery process optimization using survival regression

All papers can be downloaded at: wp.ffu.vse.cz