Time to Default in Credit Scoring Using Survival Analysis

Ana Maria SANDICA[1]
Monica DUDIAN[2]

ABSTRACT
Credit risk assesment has been dominated by logistic and probit regression techniques. As the use of credit scoring has expanded over the past 20 years, concerns have been raised about whether its use may unfairly affect minorities. The aim of this paper is to investigate the behavioral of gender variable in different credit scoring models and what are the benefits of using this variable. The first result shows that females have a probability of surviving 22 months of 75%; conversely, for the male group, the probability of surviving the same time is slightly more than 75%. Adding education as variable, we observe that male with university degree recorded a survival probability of 90% after 20 months, male with high school 75% while female with high school only 50%. The hazard for female is in average 1.12 times the hazard for males. The results show that the single and divorced females survive more than males in the same marital status.

KEYWORDS: *credit scoring, survival analysis, proportional hazards, default.*

JEL CLASSIFICATION: *C34, C35, D61, D81, G21, J16*

1. INTRODUCTION

Survival analysis, seen as a continuous process of monitoring risk over time, was a new proposal technique in risk management as alternative to model the default probability. These new models were considered a new step comparing with more traditional methods used in credit scoring. The main feature is that the survival models predict not only if a client defaults but also the time when a default occurs. The aim of this paper is to investigate the specific predictive factors, or "credit variables," used in the models that generate credit scores and the question of whether the use of individual credit characteristics may have a disparate impact.

The paper is organized as follows: the second section provides a review of literature on credit scoring models. The next section describes the methodology used for logistic regression, probit and survival analysis. Data and empirical results show the main output of the analysis and the last section concludes.

2. LITERATURE REVIEW

Myers and Forgy (1963) compared discrimination analysis with regression in credit scoring application. Altman (1968) introduced variables in a multivariate discriminant analysis and obtained a function depending on several financial ratios. Orgler (1970), based on the research of Altman, investigated the use of the credit scoring technique on commercial loans. The purpose was to review periodically the quality of the loans already disbursed. He found that

[1] The Bucharest University of Economic Studies, Romania, anamaria.sandica@gmail.com
[2] The Bucharest University of Economic Studies, Romania, monica.dudian@economie.ase.ro, corresponding author

business borrowers in general do not belong to large homogeneous populations as do customers for consumer credit.

Sexton (1975) investigates which variables have predictive power for high-income and low-income households and he concluded that different techniques are necessary for these two categories. Eisenbeis (1978) took into consideration the credit scoring systems developed by academics, mostly with discriminant analysis and presented the problems related with the named technique.

Ang, Chua and Bowling (1979) built a non-parametric credit scoring system based on a decision tree technique and arrived to the conclusion that linear credit scoring models are not the best solution, given that the relationship between some variables are nonlinear. For the financial institution it is important not only if but also when the creditor defaults. The moment of default led to some series of researches in this domain. Narain (1992) and Banasik et al. (1999) were the first researches that used survival analysis in estimation of time of default.

The first studies that used Cox PH regression in credit risk area was in 1991 for modeling time to bankruptcy of institutions in the U.S. financial market. Narain (1992) used survival analysis to predict time to default in a loan bank portfolio. Witzany et al. (2010) made a comparison between survival analysis and logistic regression as a method to model LGD.

Their results indicate that survival analysis had a better performance, this main conclusion was also presented in Banasik et al. (1999) and Stepanova and Thomas (2002). An important classification of these models is related to the distribution of the model, incidence model component by modelling a binary distribution, and the component parametric or semi parametric time to event distribution, latency model component. This theory was initially proposed by Berkson and Gage (1952).

For latency model component, Kuk and Chen (1992) extended the model by using Cox PH regression for conditional survival function. This model is known as the Cox PH mixture cure model. Tong et al. (2012) compared the Cox PH mixture cure model performance to the Cox PH regression model and standard logistic regression. Their results indicated that for both survival methods, there were good performances presented by the marginal survival function against Kaplan-Meier estimates stratified by the covariate levels available in the research.

Dirick, Claeskens and Baesens (2015) investigated "the performance of various survival analysis techniques applied to ten actual credit data sets from Belgian and UK financial institutions". Their main conclusion is that "spline-based methods and the single event mixture cure model perform well in the credit risk context ".

The first important step for the classification between good and bad credit is to use an appropriate classification technique. As a second step is to have as many characteristics available such as age, gender, marital status, education level, occupation time, filed of activity, seniority, time at present address, postal code or wither if the client uses internet banking or not. Regarding the usage of these variables, we choose to observe which of them might be used together in order to capture more information in the survival analysis context and to answer the following questions: Are woman more credit constraint? Is education an important part of the survival function?

3. METHODOLOGY

In a logistic regression, if p_i is the probability that applicant *i* has defaulted, the purpose is to find w^* that best approximate

$$p_i = w_0 + x_{1i}w_1 + x_{i2}w_2 + \cdots + x_{ip}w_p \quad (1)$$

The purpose is to find a function of p_i which could take values between 0 and 1 and one such function is the log of probability odds.

The log likelihood function then is:

$$LL = \sum_{i=1}^{N} y_i \log(P(y_i = 1|x_i)) + (1 - y_i)\log(1 - P(y_i = 1|x_i)) \quad (2)$$

In probit analysis if *N(x)* is the cumulative normal distribution function so that:

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy \quad (3)$$

Then main objective is to estimate $N^{-1}(p_i)$ as a linear function of the characteristics of the applicant, so that:

$$N^{-1}(p_i) = w_0 + w_1 x_{1i} + w_2 x_{2\,i} + \cdots w_p x_{pi} = w \cdot {x_i}^T \quad (4)$$

Regarding survival analysis, suppose Y is a random variable with probability density function f. Then Y is a survival random variable if an observed outcome, y of Y, always lies in the interval $[0, \infty)$. The cumulative density function F for this random variable is:

$$F(y) = P(Y \leq y) = \int_0^y f(u)du \quad (5)$$

The *survival function* and the *hazard function* are the two functions most often used to describe survival data. The *survival function* is defined as $S(y) = P(Y > y) = 1 - F(y)$ and therefore the *survival function* defined in terms of the probability density function $f(u)$ becomes:

$$S(y) = P(Y > y) = \int_y^{\infty} f(u)du \quad (6)$$

In this respect, given S, the probability density function is:

$$f(u) = -\frac{d}{du}S(u) \quad (7)$$

The hazard function could be represented as the instantaneous rate of failure at any time, y, given that the individual has survived up to that time,

$$h(y) = \lim_{\Delta y \to 0} \frac{P(y < Y < y + |\Delta y| Y \geq y)}{\Delta y} \quad (8)$$

The hazard function could be defined in terms of the survival function, as follows:

$$h(y) = \frac{f(y)}{S(y)}, \qquad y > 0 \tag{9}$$

In a similar way, the survival function could be expressed in terms of the hazard function,

$$S(y) = e^{-\int_0^{\infty} h(u)du} \tag{10}$$

The only constraint of the hazard function is that must be positive and one common form of this function is known as "bathtub" curve proposed by Collett (1994). The cycle of living could be detailed as following: in the first year, the mortality is high and as the time goes by, the risk of death decreases. After constant failure rate during adult life, at the end of life the risk increases once more. The survival distribution could be derived by the cumulative hazard, $H(y)$:

$$H(y) = \int_0^y h(u)du = lnS(y) \tag{11}$$

The failure time of the subject might not be observed and this is a characteristic of lifetime data, is that it is often subject to censoring. There are three types of censoring, right, left and interval censoring. The right censoring is when the event of interest is only known to sometime after the censoring point. Left censoring occurs when the failure is known to have occurred before a certain time while the interval censoring occurs within a particular time interval. In credit scoring, right censoring is most frequent used while some studies Smith (2002) reveals that left-censoring might be transformed in right censoring by changing axis. In order to detect those observations that are censored, a way is to construct a censoring indicator variable. In order to do that, the indicator variable δ is defined as follows:

$$\delta_i = \begin{cases} 1 \text{ if } Y \leq y \text{ (uncensored)} \\ 0 \text{ if } Y \geq t \quad \text{(censored)} \end{cases} \tag{12}$$

Where t is the censor time and the number of censored observations have a δ of 0 and uncensored observations have a δ equal to 1. Therefore, the lifetime can be represented in terms of the pairs (Z_i, δ_i) where for each i, $Z_i = \min(Y_i, t_i)$

The first step to estimate the survival function represents the construction of life *tables* (also named actuarial tables). To build a life table we have to split up the survival lifetime into intervals. Then, for each interval one counts the number of subjects entering the interval alive, the number that didn't survive within the interval, and the number of subjects that are censored within the interval. The Product-Limit Estimator (Kaplan &Meier, 1958) is a non-parametric method used to estimate the survival function from a right censored data. This estimator is seen as an extension of life table method, since it is a life table where each interval contains one un-censored observation.

For n subjects, let $(Z_{(1)} < Z_{(2)} < Z_{(3)} \ldots Z_{(n)})$ be the ordered observations with corresponding censoring indicators $(\delta_{(1)}, \delta_{(2)}, \delta_{(3)}, \ldots, \delta_{(n)})$. For data with no ties the PL-estimate $\hat{S}$ is defined as follows:

$$\hat{S}(u) = \prod_{j:Z_{(n)} \le u} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} \quad (13)$$

Efron (1967) proposes the redistribute-to-the right (RTR) algorithm as a different approach way to calculate the PL estimator. The first step is to assign equal mass to each observation, and then the mass of the first censored observation is equally redistributed among the other observations to the right. The second step is repeated for all remaining censored observations with the exception for the largest one. Having a regression model where a vector of p explanatory, $x^T = (x_1, x_2, \dots, x_p)$ used to model the behavior of Y_x, there are two categories of techniques that could be used: the accelerated failure lifetime models and proportional hazards regression.

One of the main features of accelerated failure time models (AFT) is that the explanatory variables apparently either speed up or slow down the rate of failure. Considering g_2, a positive function of x and S_0 is the baseline function, then the AFT model could be expressed as:

$$S_x(y) = S_0(y g_2(x)) \quad (14)$$

Where the failure rate is slowed in case $g_2(x) > 1$.

The associated hazard function is calculated by differentiating the eq. [14]

$$h_x(y) = h_0[y g_2(x)] g_2(x) \quad (15)$$

In order to have the model in log linear form, considering $g_2(x) = e^{\beta^T x}$, we have,

$$\log_e Y_x = \mu_0 + \beta^T x + \sigma Z \quad (16)$$

where Z is a random variable having zero mean and unit variance.

Cox regression is one of the models from category of proportional hazards models, having the following form:

$$h_x(y) = h_0(y) g_1(x) \quad (17)$$

where $h_0(t)$ is the baseline hazard and $g_1(x)$ is a positive function of x, where $x^T = (x_1, x_2, \dots, x_p)$. In order to calculate a proportional hazard through maximum likelihood, it is required to specify the forms of the functions from Eq. [17]. The proposal made by Cox is to estimate β, when $h_0(y)$ is still arbitrary and $g_1(x) = e^{\beta^T x}$ where $\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$. Therefore the eq. [17] can be expressed as:

$$h_x(y) = h_0(y) e^{\beta^T x} \quad (18)$$

Where $h_0(y)$ is the baseline hazard which occurs then $\beta^T x = 0$.

The estimation of β involves maximizing a function called the partial likelihood (also called the conditional likelihood). The approach assumes that the observations are ordered according to $y_{(1)} < y_{(2)} < y_{(3)} \dots < y_{(k)}$. The risk set, R_i, at the time $y_{(i)}$ is the set of subjects alive and under observation at time $\overline{y_{(1)}}$, just prior to $y_{(1)}$. Based on the eq. [8], the hazard function, h_x for an individual with explanatory variables, x, is written as:

$$h_x(y) \approx \frac{P(y < Y_x < y + \Delta y | Y > y)}{\Delta y} \quad (19)$$

Approximating,

$$h_x(y)\Delta y \approx P(y < Y_x < y + \Delta y | Y > y\ ) \approx P(Die\ at\ y | Alive\ up\ to\ y) \quad (20)$$

And defining $P_{y_{(j)}}$ as $P_{y_{(j)}} = P(\text{The individual from } R_j \text{ eith covariate } x_j \text{ dies at } y_j | \text{A member of } R_j \text{ dies at } y_{(j)})$, therefore $P_{y_{(j)}} = \frac{e^{\beta^T x_{(j)}}}{\sum_{l \in R_j} e^{\beta^T x_l}}$. The Cox partial likelihood is calculated by taking the product of the probabilities for each individual as:

$$L_c(\beta) = \prod_{j=1}^{k} P_{y_{(j)}} \quad (21)$$

Which then gives the expression for the Cox partial likelihood,

$$L_c(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T x_{(j)}}}{\sum_{l \in R_j} e^{\beta^T x_l}} \quad (22)$$

Cox (1972) indicated that the partial likelihood can be treated in the same way as other likelihood method and the estimation of the parameters could be realized by maximizing the partial likelihood from eq. [11]. One feature of the proportional hazards model is that the baseline function is not related to the process of estimating the parameters. The baseline hazard disappears from the derivation of the partial likelihood which can be seen in the equation [22] that does not contain the baseline hazard. As a first step, one can determine the weights and as a second step to derive the baseline hazard. To provide predictions of lifetimes, one must then calculate the baseline hazard, usually via maximum likelihood techniques, as explained in Kalbfleisch & Prentice (1980). Cox regression is related to a semi rather than non-parametric regression because while the baseline hazard is not a specified shape, the covariates are forced to enter the model in the form of $\beta^T x$. The first assumption is that relative hazards in Cox regression must be proportional. This is often called the *proportionality assumption*. For an example, take two experimental subjects with associated covariate vectors x_1 and x_2. Then the ratio between the respective hazards remains constant and is independent of *y*:

$$\frac{h_{x1}(y)}{h_{x2}(y)} = \frac{h_0(y)g_1(x_1)}{h_0(y)g_1(x_2)} = \frac{g_1(x_1)}{g_1(x_2)} \quad (23)$$

The assumption of proportional hazards has consequences for analysis because a subject considered riskier than another will always be considered riskier independent of the time passed. Proportional hazards and accelerated failure models are equivalent under certain conditions if the survival distribution is Weibull. By taking the logarithm of both sides, the equation can be written as $\log h_x(y) = \alpha(y) + \beta^T X$, where $\alpha(y) = \log h_0(y)$. When $\alpha(y) = \alpha$ is constant, the equation corresponds to an exponential AFT model and when $\alpha(y) = \alpha \log(y)$ then the same equation is a Weibull model. Credit scoring data is generally reported monthly, and as a consequence event times are often tied.

All events during a month are reported to have occurred at the same time at the end of the month. There three ways of handling ties, two of the methods are approximations, Breslow (1974) and Efron (1977), while the third way is to calculate the exact partial likelihood. The so-called exact method becomes computationally intensive when there are many ties because every possible ordering needs to be taken into account. The technique is presented by Kalbfleisch and Prentice (1980). The Breslow approximation assumes that tied data occur

sequentially. The form of the partial likelihood in this approximation is $L_c(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T x_{(j)}}}{\left[\sum_{l \epsilon R_j} e^{\beta^T x_l}\right]^{d_j}}$, where d_j is the observed number of failures at time y_j and $s_j = \sum_{l=1}^{d_j} d_{(j)t}$. Efron (1977) proposed an approximation closer to exact partial likelihood and in practice they give similar results for credit data (Stapanova, 2001),$L_c(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T s_j}}{\prod_{r=1}^{d_j}\left[\sum_{l \epsilon R_j} e^{\beta^T x_{(j)} - (r-1) d_j^{-1} \sum_{l \epsilon D_j} e^{\beta^T x_{(j)}}}\right]}$, where D_j is the number of failures at time j.

4. DATA AND EMPIRICAL RESULTS

The first step in a credit scoring model development is to define the default event. In the Basel II Capital Accord, the Basel Committee of Banking Supervision gave a reference definition of the default event and announced that banks should use this regulatory reference definition to estimate their model internal rating based. According to this, a default is considered to have occurred with regard to particular obligor when either or both of the two following events took place, the bank considers that the obligor is unlikely to pay or the obligor is past due more than 90 days on any material credit obligation to the banking group. Taking into consideration that the data sample used contains customers with approval date of the credit between 2007 and 2008, an observation period for 24 months was considered. For example, if the client has been approved on January 2007 then for two years it has been observed to see it he meets the definition of default if this thing happened then a status of 1 have been recorded, otherwise a status of non-defaulter,0.

The available variables are split into two different categories: socio-demographical variables and financial information such as “Monthly Income” or “Financial Expenses”. These have proven to be of great importance in defining the profile of a default person. For instance, “Education” represents valuable information whereas persons with a higher degree of education tend to be more responsible. “Industry” is also very relevant especially during times like these affected by financial crisis when some fields (i.e. real estate, commerce, constructions etc.) have reached an unemployment rate higher than others.

“Marital status” and “Gender” have also shown significance in the rating process. For instance, married men are considered to be better payers than single ones who tend to be less responsible. Financial variables have considerable predictive power. They reveal the capacity of paying monthly instalments taking into consideration the wages of the applicants and their monthly expenses too. The database consists of 29159 observations, provided from a portfolio of personal loans and the ratio of default clients reaches the level of 14.81% on our database.

The first step of the analysis was to estimate parameters using logistic regression and Z-statistic indicates that all variable’s coefficients are significantly different by zero. The odd ratio for Gender and Expenses are higher than 1, while the other variables are having the odd ratio lower than the unit. Therefore, the default probability for women and higher expenditures are increasing (Table 1).

The likelihood ratio chi-square of 8326.99 with a *p-value* of 0.000 indicates that the probit model as a whole is statistically significant, that is, it fits significantly better than a model with no predictors. The coefficient of gender is 0.0899, this means that for female (gender=1)

increases the predicted probability of default. The other social-demographic parameters are negative, meaning that for instance a higher degree in education is negatively correlated with increases in PD. (Table 2).

Table 1. Estimated parameters for logistic regression

Variable	Odd ratio	Std. Error	z	P>|z|	[95% Conf. Interval]	
Repayment	0.0472	0.0033	-43.20	0.0000	0.0411	0.0542
Gender	1.1655	0.0496	3.60	0.0000	1.0723	1.2669
Age	0.9859	0.0029	-4.86	0.0000	0.9803	0.9916
Marital Status	0.8471	0.0137	-10.29	0.0000	0.8207	0.8743
Education	0.3770	0.0143	-25.77	0.0000	0.3500	0.4060
Profession	0.7723	0.0313	-6.37	0.0000	0.7133	0.8362
Seniority	0.7382	0.0106	-21.21	0.0000	0.7178	0.7592
Industry	0.9439	0.0037	-14.81	0.0000	0.9367	0.9511
Residence	0.5118	0.0139	-24.65	0.0000	0.4853	0.5398
Income	0.9999	0.0000	-8.20	0.0000	0.9999	1.0000
Expenses	1.0005	0.0000	12.98	0.0000	1.0004	1.0006
Log likelihood	-8587.5233					
LR chi2(11)	8533.5					
Prob >chi2	0.0000					

Source: authors' calculation

Table 2. Estimated parameters for probit regression

Variable	Coeff	Std. Error	z	P>|z|	[95% Conf. Interval]	
Repayment	-1.5597	0.0346	-45.10	0.0000	-1.6275	-1.4919
Gender	0.0899	0.0233	3.86	0.0000	0.0443	0.1356
Age	-0.0067	0.0016	-4.30	0.0000	-0.0098	-0.0037
Marital Status	-0.0935	0.0089	-10.52	0.0000	-0.1109	-0.0761
Education	-0.5165	0.0207	-25.00	0.0000	-0.5570	-0.4760
Profession	-0.1295	0.0222	-5.82	0.0000	-0.1731	-0.0859
Seniority	-0.1639	0.0078	-21.02	0.0000	-0.1792	-0.1487
Industry	-0.0305	0.0021	-14.57	0.0000	-0.0346	-0.0264
Residence	-0.3726	0.0152	-24.48	0.0000	-0.4024	-0.3428
Income	0.0000	0.0000	-7.45	0.0000	0.0000	0.0000
Expenses	0.0002	0.0000	15.27	0.0000	0.0002	0.0003
Log likelihood	-8690.7763					
LR chi2(11)	8326.99					
Prob >chi2	0.0000					

Source: authors' calculation

In order to choose which model fits better the data, we took into account the log likelihood and AIC values; based on them the first conclusion is that the logistic regression fits better the data. The analysis continues with switching the parameters by considering the time to default and the aim is to choose a model that explains better and what might be the role of gender in this framework. The life of the account is measured from the month it was opened until the

account becomes 'bad' or it is closed or until the end of observation. Therefore, we have tested the equality of survival functions by using the long rank test. The test of equality across strata for the predictor gender has a *p-value* of 0.0001, thus gender will be included a potential candidate for the final model (Table 3).

Table 3. Log-rank test for equality survival function-gender

	Events observed	**Events expected**		
0	3182	3302.27	chi2(1)	15.08
1	1505	1384.73	Pr>chi2	0.0001
Total	4687	4687		

Source: authors' calculation

The survival estimate of the portfolio indicates that there is a probability of 75% surviving after 20 months of observation (Figure 1).

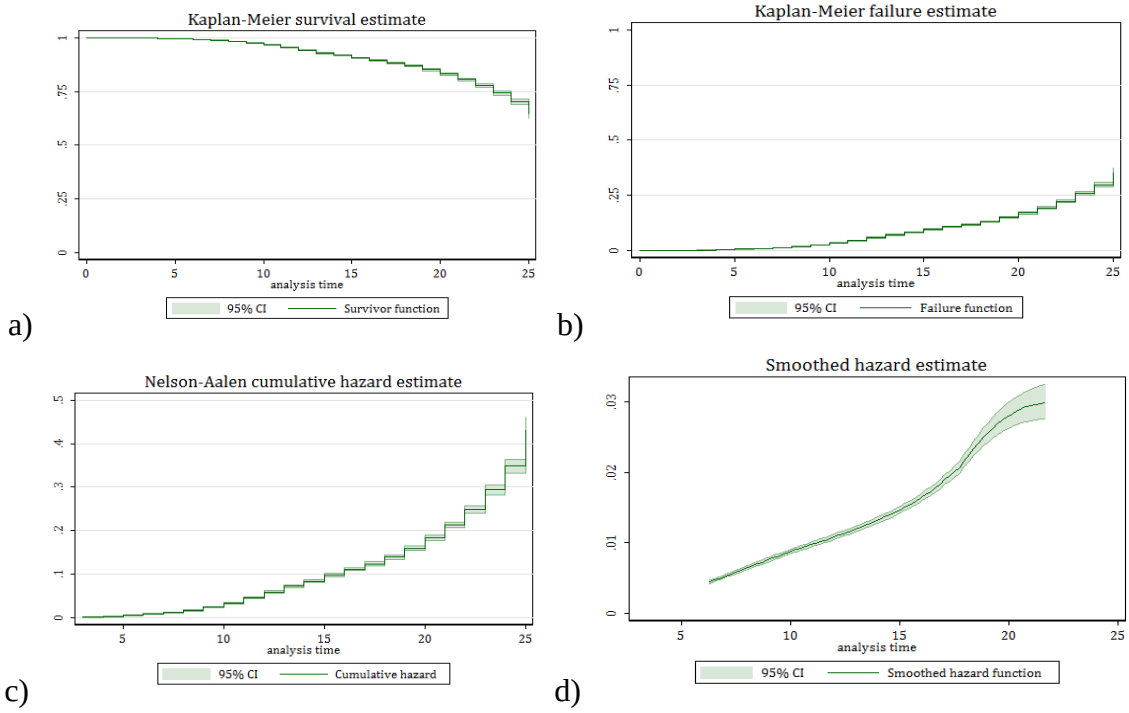


Figure 1. Kaplan-Meier survival/failure estimate

Source: authors' conception

But what happens if we split the data on gender? From the Figure 2 (a) and b), we see that the survival function for each group of gender are not perfectly parallel but separate except at the very beginning and at the very end.

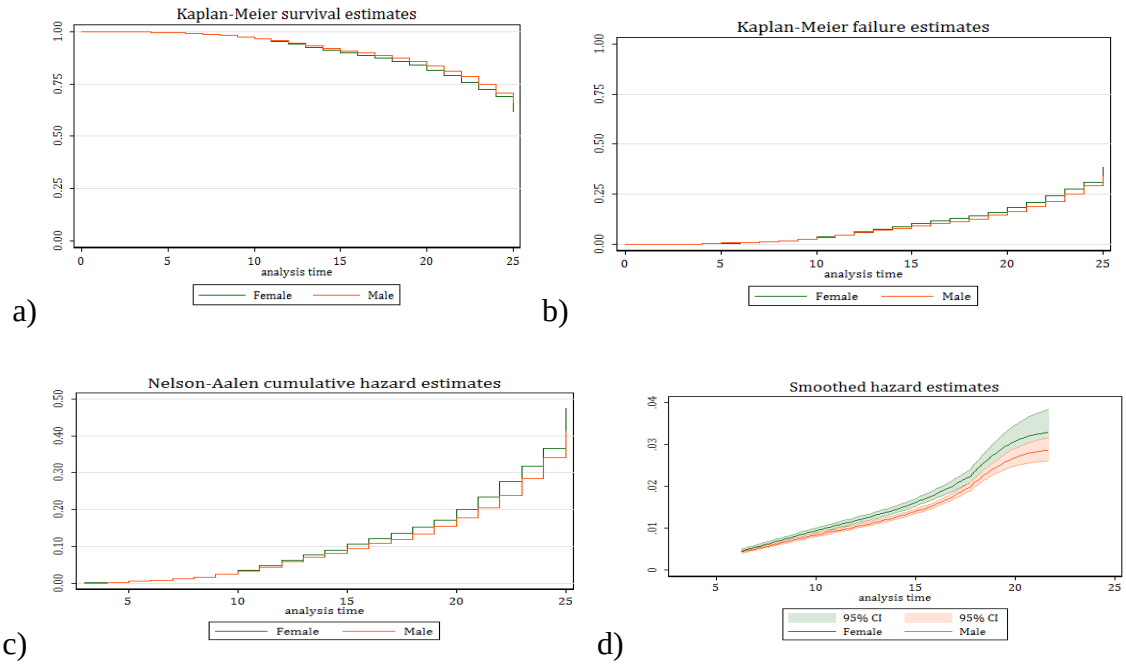


Figure 2. Kaplan-Meier survival/failure estimate split on gender
Source: authors' conception

The overlap at the very end should not cause too much concern because it is determined by only a very few number of censored subjects out of a sample. For the female group, the probability of surviving 22 months is 75%; conversely, for the male group probability of surviving the same time is slightly more than 75%. Adding education as variable, we observe that male with university degree recorded a survival probability of 90% after 20 months, male with high school 75% while female with high school only 50 %. (Figure 3).

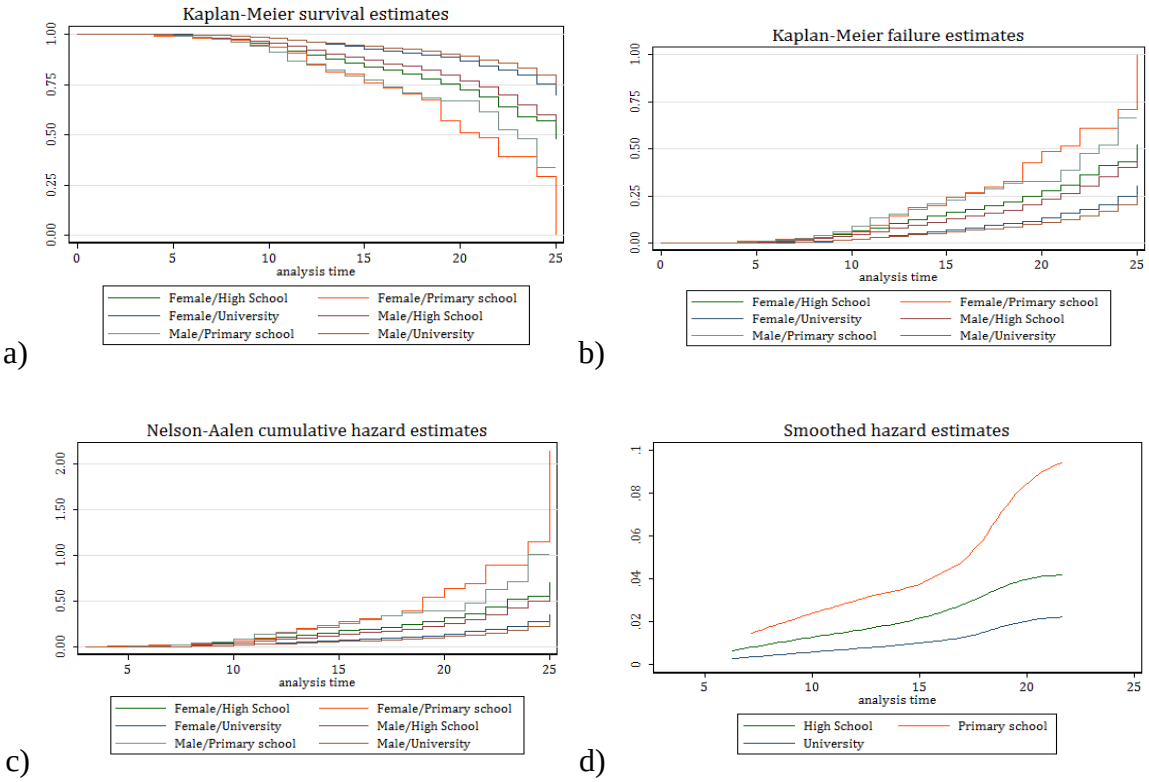


Figure 3. Kaplan-Meier survival/failure estimate split on gender and education
Source: authors' conception

The smoothed hazard estimates (d) in the same figure reveals that the hazard after 10 months is almost 5% for university while for high school is almost 20%. The smoothed hazard estimates for marital status indicates those that are more likely to survive are married people, widowed, single and last the divorced ones. An interesting point is after 14 months, there is a change in the hazard curve between single and divorces people. The explanation is that up to 1 year, single persons are more likely to survive while once the time goes by; the single persons (getting older) are likely to default more than divorced persons. This is observed also in the survival estimates where the single females have a probability of surviving higher than single men (Figure 4).

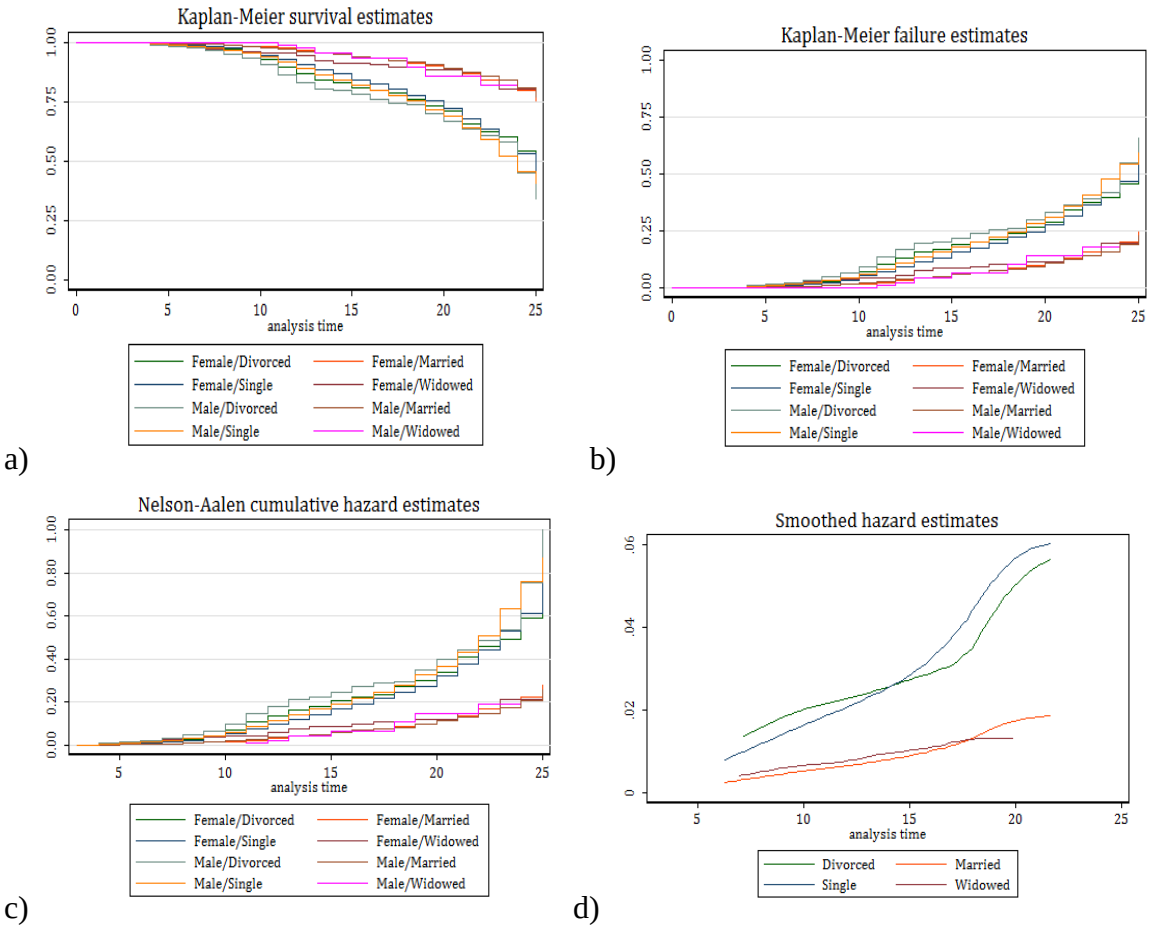


Figure 4. Kaplan-Meier survival/failure estimate split on gender and marital status
Source: authors' conception

The next step was not just to estimate if a client default or not and how the difference between genders affects it but to see if the time to default is also dependent. In this respect, we have considered different proportional Hazard and accelerated failure time models. The Weibull and exponential models are parameterized as both PH and AFT models. The Weibull distribution is suitable for modeling data with monotone hazard rates that either increase or decrease exponentially with time, whereas the exponential distribution is suitable for modeling data with constant hazard.

The Gompertz regression is parameterized only as a PH model. This distribution is suitable for modeling data with monotone hazard rates that either increase or decrease exponentially with time. The lognormal and log logistic models are implemented only in the AFT form. These two distributions are similar and tend to produce comparable results. For the lognormal

distribution, the natural logarithm of time follows a normal distribution; for the log logistic distribution, the natural logarithm of time follows a logistic distribution.

Unlike the exponential, Weibull, and Gompertz distributions, the lognormal and the log logistic distributions are indicated for data exhibiting no monotonic hazard rates, specifically initially increasing and then decreasing rates. The hazard function of the generalized gamma distribution is extremely flexible, allowing for many possible shapes, including as special cases the Weibull distribution when kappa = 1, the exponential when kappa = 1 and sigma = 1, and the lognormal distribution when kappa = 0. The generalized gamma model is, therefore, commonly used for evaluating and selecting an appropriate parametric model for the data.

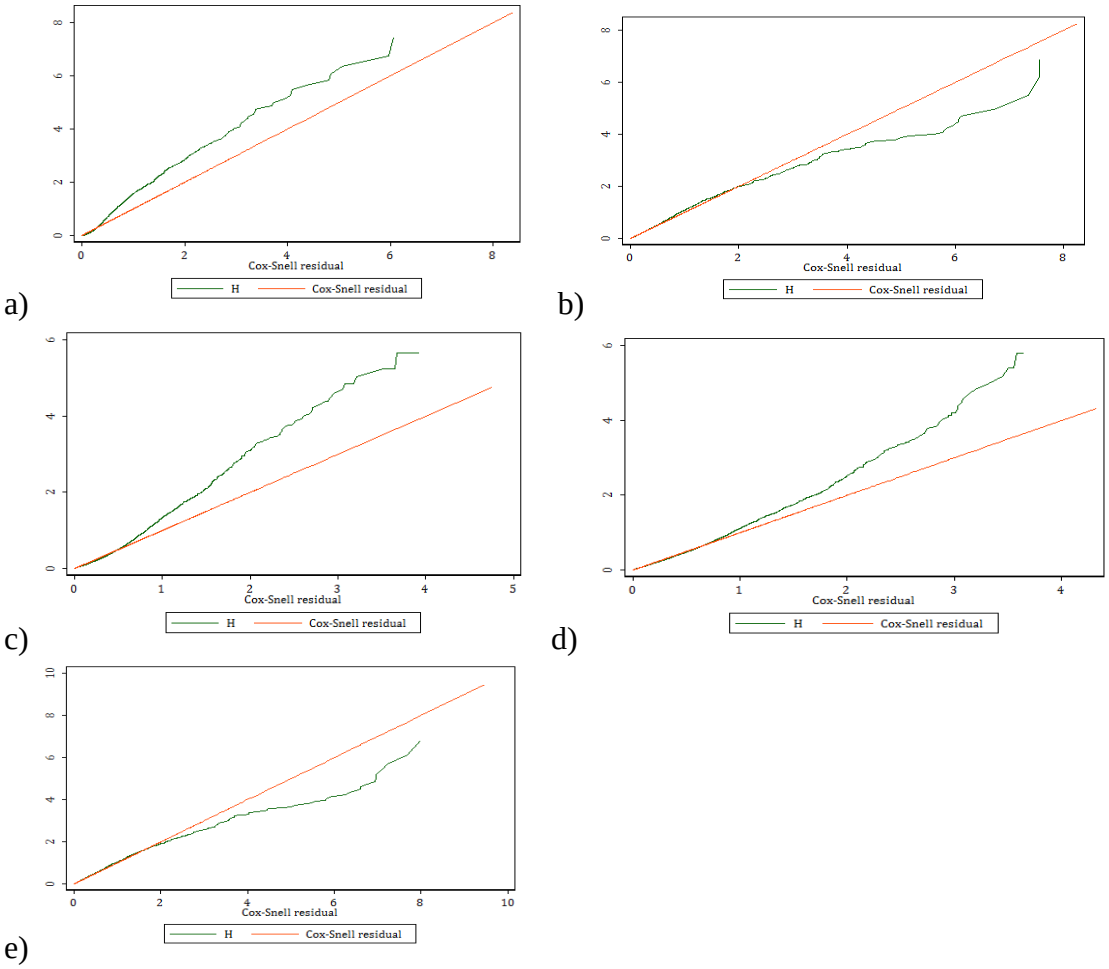
The results of the parametric models with our without accelerated failure time were realized for each type of distribution. The hazard ratio and standard error for each variable are displayed and the main conclusion is that all parameters are significantly different from zero. In terms of hazard ratio, it can be observed that hazard decreases with age since older people tend in earn more and the probability of default decreases.

Lower hazard ratio for age variable, was found also by Sarlija (2009) et al on a Croatian bank loans database. The gender variable has the same hazard ratio for all PH models from 1.1259, 1.1297 and 1.1363 meaning that hazard for female is in average 1.12 times the hazard for males. In AFT models, the sign of the coefficient indicates how a covariate affects the logged survival times. Thus, a positive coefficient increases the logged survival time and, hence, the expected duration. A negative coefficient decreases the logged survival time and, hence, the expected duration.

The statistical significance of the coefficient indicates whether these changes in the expected duration will be statistically significant or not. The results indicate that the gender coefficient is negative, meaning that for female, the logged survival time decreases. The highest coefficients are recorded for repayment, education and residence for all models analyzed. This means that the history of payments, the level of education and if the persons has an own house or not are more important in predicting the survival period that the gender.

In order to select which model fits better a common approach is to use the Akaike information criterion (AIC). Although the best-fitting model is the one with the largest log likelihood, the preferred model is the one with the smallest AIC value.

In terms of parametric models, we saw that the log-likelihood is higher, -8708 compared to -11235 for the Weibull comparison with exponential, so the model provides a better fit to the data. This model has also the smallest AIC from the parametric models but higher that the logit and probit model. This means that the model of fitting the good bad threshold, if the client default or not is better that fitting the survival period until default. The plots from Figure 5 indicate that the Weibull (b) and log logistic (d) models fit the data best and that the exponential model (a) fits poorly. These results are consistent with the previous results based on Akaike's information criterion.



a) Exponential; b) Weibull; c) Lognormal; d) Loglogistic e) Gompertz

Figure 5. Cox–Snell residuals to evaluate model fit of the regression models
Source: authors' conception

The next step is the comparison between Cox models with two different tied methods: Efron and Breslow. In the reported results (Table 4) we found that the females had a higher hazard—and therefore a smaller survivor time. The hazard ratios reported correspond to a one-unit change in the corresponding variable.

Table 4. Cox model-Breslow method

Variable	Hazard ratio	Std. Error	z	P>|z|	[95% Conf. Interval]	
Repayment	0.1794	0.0059	-51.87	0.0000	0.1682	0.1915
Gender	1.1348	0.0361	3.98	0.0000	1.0663	1.2077
Age	0.9896	0.0022	-4.64	0.0000	0.9852	0.9940
Marital Status	0.8813	0.0106	-10.49	0.0000	0.8607	0.9023
Education	0.5085	0.0140	-24.56	0.0000	0.4818	0.5367
Profession	0.8340	0.0254	-5.97	0.0000	0.7857	0.8852
Seniority	0.8109	0.0089	-19.18	0.0000	0.7937	0.8284
Industry	0.9568	0.0029	-14.47	0.0000	0.9511	0.9626
Residence	0.6793	0.0118	-22.26	0.0000	0.6566	0.7028
Income	1.0000	0.0000	-3.98	0.0000	1.0000	1.0000

Variable	Hazard ratio	Std. Error	z	P>|z|	[95% Conf. Interval]	
Expenses	1.0001	0.0000	8.12	0.0000	1.0001	1.0002
Log likelihood	-42270.924					
LR chi2(11)	7075.60					
Prob >chi2	0.0000					

Source: authors' calculation

There is no significant difference for gender variable if we use a different tied method but the goodness of fit results indicates that Efron method leads to a smaller AIC and higher log likelihood value (Table 5).

Table 5. Cox model-Efron method

Variable	Hazard ratio	Std. Error	z	P>|z|	[95% Conf. Interval]	
Repayment	0.1703	0.0057	-53.09	0.0000	0.1596	0.1819
Gender	1.1367	0.0361	4.03	0.0000	1.0681	1.2098
Age	0.9893	0.0022	-4.76	0.0000	0.9850	0.9937
Marital Status	0.8796	0.0106	-10.65	0.0000	0.8591	0.9006
Education	0.5017	0.0138	-25.05	0.0000	0.4753	0.5295
Profession	0.8311	0.0253	-6.09	0.0000	0.7831	0.8821
Seniority	0.8077	0.0088	-19.53	0.0000	0.7906	0.8252
Industry	0.9563	0.0029	-14.68	0.0000	0.9506	0.9620
Residence	0.6737	0.0117	-22.74	0.0000	0.6512	0.6971
Income	1.0000	0.0000	-4.15	0.0000	1.0000	1.0000
Expenses	1.0001	0.0000	8.26	0.0000	1.0001	1.0002
Log likelihood	-42115.351					
LR chi2(11)	7306.42					
Prob >chi2	0.0000					

Source: authors' calculation

One of the assumptions underlying of Cox PH model is the proportional hazards assumption. Recall that this assumption is the idea that covariates will have a proportional and constant effect that is invariant to time. Non-proportional hazards can arise if some covariate only affects survival up until time *t* or if the size of its effect changes over time.

Non-proportional hazards can result in biased estimates, incorrect standard errors, and faulty inferences about the effect of our covariates. Before we test this hypothesis we have analyzed –ln (-ln (survival)) curves for each category of a nominal or ordinal covariate versus ln (analysis time). These are often referred to as "log–log" plots. Optionally, these estimates can be adjusted for covariates. If the plotted lines are reasonably parallel, the proportional-hazards assumption has not been violated, and it would be appropriate to base the estimate for that variable on one baseline survivor function.

The results of PH test are for both covariate-specific and global tests and we can see that there is evidence that the proportional-hazards assumption has been violated (Table 6) for both models. Although variables such as Gender, Marital Status and Education are statistically significant the entire model rejects the null hypothesis of zero slope, which is equivalent of testing that the log hazard-ratio function is constant over time.

Table 6. Test of proportional assumption-Cox-Efron/Breslow

Cox-Breslow					**Cox-Efron**			
Variable	**rho**	**chi2**	**df**	**Prob>chi2**	**rho**	**chi2**	**df**	**Prob>chi2**
Repayment	-0.0693	257.500	1	0.0000	-0.0869	410.600	1	0.0000
Gender	0.0095	0.4300	1	0.5121	0.0107	0.5400	1	0.4636
Age	-0.0754	282.500	1	0.0000	-0.0776	300.100	1	0.0000
Marital Status	0.0216	22.700	1	0.1317	0.0202	19.700	1	0.1602
Education	0.0111	0.5700	1	0.4517	0.0068	0.2200	1	0.6425
Profession	0.0321	47.800	1	0.0288	0.0312	45.100	1	0.0337
Seniority	-0.0349	55.700	1	0.0183	-0.0397	71.900	1	0.0073
Industry	0.0635	186.300	1	0.0000	0.0625	180.300	1	0.0000
Residence	0.0340	48.700	1	0.0273	0.0303	38.600	1	0.0493
Income	-0.0970	544.700	1	0.0000	-0.1017	603.200	1	0.0000
Expenses	0.0932	406.500	1	0.0000	0.0968	441.600	1	0.0000
Global test		186.85	11	0.0000		220.16	11	0.0000

Source: authors' calculation

Observing the fact that this assumption is not hold, we have modified the model in order to obtain feasible results. In this respect we have re-estimate only Cox model with Efron method and the results are presented in Table 7 and the PH test results indicates now that there is no evidence that the proportional hazard assumption is not respected (Table 8).

Table 7. Cox model modified-Efron method

Variable	Hazard ratio	Std. Error	z	P>|z|	[95% Conf. Interval]	
Gender	1.0704	0.0339	2.15	0.0320	1.0060	1.1389
Marital Status	0.6841	0.0069	-37.91	0.0000	0.6708	0.6976
Education	0.4554	0.0119	-30.08	0.0000	0.4326	0.4793
Log likelihood	-44696.74					
LR chi2(11)	2223.96					

Source: authors' calculation

Table 8. Test of proportional assumption-Cox-Efron modified

Variable	**rho**	**chi2**	**df**	**Prob>chi2**
Gender	0.0099	0.47	1	0.4948
Marital Status	-0.0241	2.71	1	0.0996
Education	0.0158	1.12	1	0.2899
Global test		4.88	3	0.1807

Source: authors' calculation

After each step of estimation of parameters, we have also included the Cox–Snell. If the Cox regression model fits the data, these residuals should have a standard censored exponential distribution with hazard ratio 1.

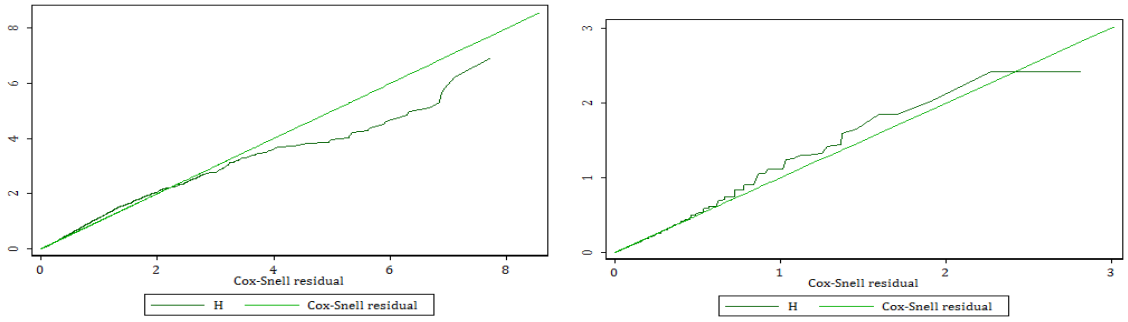


Figure 6-Cox. Snell residuals to evaluate model fit of the Cox model

Source: authors' calculation

We can verify the model's fit by calculating-based, for example, on the Kaplan–Meier estimated survivor function or the Nelson–Aalen estimator—an empirical estimate of the cumulative hazard function, using the Cox–Snell residuals as the time variable and the data's original censoring variable. If the model fits the data, the plot of the cumulative hazard versus residuals should approximate a straight line with slope 1. Comparing these graphs the first conclusion is that the model now fits better the data (Figure 6).

The next step was to calculate the concordance probability, which is defined as the probability that predictions and outcomes are concordant. This provides two measures of the concordance probability: Harrell's C and the Somers' D rank correlation.

The value of Harrell's C is 0.8281, which indicates that we can correctly order survival times for pairs of patients 82% of the time on the basis of measurement of all variables in the model (Table 9).

Comparing the results with the Cox modified model the Harrell's C decreased up to 68% and the Somer's D, which measures how concordant predicted hazards are with observed failure times, decreased up to 36%.

For the last model, we took into account the difference in the marital status and observed the hazard ratio for gender variable. What is interesting is that the hazard ratio is higher than 1 unit for married (1.1356) and widowed (1.0271) people while is less than 1 for single and divorced. T

hese values indicate the fact that the hazard increases for female in the first two categories while the survival increases for single and divorced females. This is highly correlated with the previous analysis of Keplan-Meier survival estimates, where it we have observed the slope for single. So, what a credit scoring model shall take into consideration? The higher the level of education will lead to higher income and indirectly the second income in the calculation of debt to income will help increasing the survival time. The gender, instead, is not significant as a single variable, therefore should be consider either together with marital status, either with age or education or both.

Table 9. Concordance probability

	Cox Efron	Cox Efron modified
Harrell's C = (E + T/2) / P	0.8281	0.6812
Somers' D	0.6561	0.3623

Source: authors' calculation

When comparing our results with literature review we took into account developed countries and emergent countries. Constangioara (2010) reported hazard ratios lower than 1 for age, marital status and education using a Hungarian database. Ganopoulou, Giapoutzi, Kosmidou and Moysiadis (2013) using a sample set of applications from a large Greek financial institution estimated a probit model. The main conclusion is that probability of default decreases for married people and older applicants while gender variable was not statistically significant for 2007 data sample. For 2009 data, instead, the results showed that only age was statically significant underling that the financial conditions became more difficult; the factors that influence the lending decision were related to the financial wealth of the respondent. Kočenda and Vojtek (2009) estimated a logistic regression using data on loans at the retail banking market from Czech Republic. Their main conclusion is that higher the education level leads to a lower probability of default and married clients are considered less risky from the credit perspective.

The Gender variable was not considered in the model, as the information value for this variable was too low. Sarlija (2009) using a data set of 50,000 customer accounts (application data and transaction data) in Croatia over the period of 12 months estimated logistic regression and a survival based credit scoring model. One of their conclusions is that female clients are more likely to have an account in default compared to male clients. In terms of borrower's age, the older it is, the lower the risk. Regarding USA market Agarwal, Chomsisengphet and Liu (2009) estimated a Cox Proportional model using monthly panel data set of more than 170 000 credit cardholders. Their main conclusion was that married borrowers have lower risk. Marjo (2010) used a unique dataset of 14 595 observations from one of Finland's largest and well-known consumer credit companies who has over 150 000 customers. Estimating the model using logistic regression, one of his conclusions is that gender is a significant variable showing that female customers have much less difficulty in paying their debts and seem to default less than man. On the other hand, education variable wasn't statically significant and in terms of age, younger borrowers tend to default more often.

5. CONCLUSIONS

Duration analysis is an analytical tool for time-to-event data that has been borrowed from medicine and engineering to be applied by econometricians to investigate typical economic and finance problems. Lately this kind of analysis have been used in credit scoring models by calculating the survival period instead of detecting if the clients pays or not. The aim of this paper was to investigate the specific predictive factors, or "credit variables," used in the models that generate credit scores and the question of whether the use of individual credit characteristics may have a disparate impact.

The first analysis shows that females have a probability of surviving 22 months of 75%; conversely, for the male group, the probability of surviving the same time is slightly more than 75%. Adding education as variable, we observe that male with university degree recorded a survival probability of 90% after 20 months, male with high school 75% while female with high school only 50%. The gender variable has the same hazard ratio for all PH models from 1.1259, 1.1297 and 1.1363 meaning that hazard for female is in average 1.12

times the hazard for males. The plots of the Cox-Snell residuals indicate that the Weibull and log logistic models fit the data best and that the exponential model fits poorly. These results are consistent with the results based on Akaike's information criterion.

The next step was the comparison between Cox models with two different tied methods: Efron and Breslow. In the reported results, we have found that the females had a higher hazard – and therefore a smaller survivor time. There is no significant difference for gender variable if we use a different tied method but the goodness of fit results indicates that Efron method leads to a smaller AIC and higher log likelihood value. After observing that the proportional hazard hypothesis is violated, we have modified the Cox model by including only gender, marital status and education (those variables with p-value higher than 5%). Using the last model, we took into account the difference in the marital status and observed the hazard ratio for gender variable. The results show that single and divorced female survive more than males in the same marital status. The results from the model estimation with Romanian retail credit data provide the first evidence of variables reacting in opposite directions for gender depending on the marital status. After analyzing the behavioral of gender coefficient in two credit scoring categories models the main conclusion is that the gender as variable shouldn't be considered, but together with marital status or age or both of them. The unemployment rate split by gender shows that women were generally more affected by the crisis. The Eurostat figures for Romania indicates that between 2008 and 2010 the increase in unemployment rate for women is higher comparing with the increase for men. For instance, the unemployment rate for women between 50 to 64 years increased with 70%, while for men this increase was only of 18%. During the crisis, the increase in the size of the unemployment rate underlined the importance of this phenomenon. The composition of the identified class of out of work population between 2008 and 2011 reveals that the youth class increased after the crisis. Another important observation in the World Bank report is the increase in cluster of early retirements after 2008; moreover, the share of males in this group fell from 60% to 40% after the crisis and in terms of education, 73% of this cluster has high school completed. Regarding our results, that survival period for male with high school is 75% while female with high school only 50% is actually related to the movements in the unemployment rate behavior during crisis.

ACKNOWLEDGEMENTS

This work has been done through the Partnerships Program in Priority Areas - PNII, developed with the support of MEN-UEFISCDI, project no. 334/2014, project code PN-II-PT-PCCA-2013-4-0873, project title „A New Model for Corporate Risk Assessment: A Scientific Tool for Knowledge Based Management".

REFERENCES

Agarwal, S., Chomsisengphet, S. & Liu, C. (2009). *Consumer Bankruptcy and Default: The Role of Individual Social Capital*. Working Paper. Retrieved March 10, 2015, from http://ssrn.com/abstract=1408757.

Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance, 23,* 589-609.

Ang, J. S., Chua, J. H. & Bowling, C. H. (1979). The Profiles of Late-Paying Consumer Loan Borrowers: An Exploratory Study. *Journal of Money, Credit & Banking (Ohio State University Press) 11,* 222-226.

Banasik, J., Crook, J.N., & Thomas, L.C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society,50,* 1185-1190.

Berkson, J., & Gage, R.P. (1952). Survival curve for cancer patients following treatment. *American Statistical Association, 47*, 501-515.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics, 30*, 89-99.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Boca Raton, FL: Chapman & Hall/CRC

Constangioara, A. (2010). *Consumer Credit Scoring: An Application Using a Hungarian Dataset of Consumer Loans.* Saarbrücken: Lambert Academic Publishing.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, 34*, 187-220.

Cox, D.R. (1975). Partial likelihood. *Biometrika, 62*, 269-276.

Crook, J.N., & Banasik, J.(2004). Does reject inference really improve the performance of application scoring models?. *Journal of Banking & Finance, 28*, 857–874.

Dirick, L., Bellotti, T., Claeskens, G., & Baesens, B. (2015). *The prediction of time to default for personal loans using mixture cure models: including macro-economic factors*. Paper presented at the meeting of the Credit Scoring and Credit Control XIV conference. Edinburgh, Scotland.

Efron, B. (1967). *The Two Sample Problem with Censored Data*. Paper presented at the meeting of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *American Statistical Association, 72,* 557-565.

Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance, 2*, 205-219.

Ganopoulou, M., Giapoutzi, F., Kosmidou, K., & Moysiadis, T. (2013). Credit-scoring and bank lending policy in consumer loans. *International Journal Financial Engineering and Risk Management, 1,* 90-110.

Kalbfleisch, J. D., & Prentice, R.L. (1980).*The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Non-parametric estimation from incomplete observations. *American Statistical Association, 53*, 457-500.

Kočenda, E., & Vojtek, M. (2009). *Default Predictors and Credit Scoring Model.* (*CESifo* Working Paper No. 2862) Retrieved January 16, 2015 from file:///C:/Users/user/Downloads/cesifo1_wp2862.pdf

Kuk, A.Y.C., & Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika, 79*, 531–541.

Marjo, H. (2010). *The determinants of Default in consumer credit market.* (Master Thesis). Aalto University School of Economics

Myers, J. H., & Forgy, E. W. (1963). Development of Numerical Credit Evaluation Systems. *Journal of American Statistical Association, 50*, 797-806

Narain, B. (1992). *Survival analysis and the credit granting decision* (pp 109-121).UK: Oxford.

Orgler, Y. E. (1970). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit & Banking (Ohio State University Press) 2*, 435-445

Sarlija, N., Bensic, M., & Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications, 36,* 8778-8788.

Sexton Jr., D. E. (1975). Credit Riskiness of Low-Income Consumers. *Advances in Consumer Research, 2,* 197 -202.

Smith, P. J. (2002). *Analysis of Failure and Survival Data*. New York: Chapman & Hall.

Stepanova, M. (2001). *Using survival analysis methods to build credit scoring models*. University of Southampton.

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

Sundaram, R., Hoerning, U., Falcao, N., Millan, N., Tokman, C.,& Zini, M. (2014). *Portraits of Labor Market Exclusion*. The International Bank for Reconstruction and Development. The World Bank

Sy, J. P., & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics, 56,* 227-236.

Tong, E. N., Mues, C., & Thomas, L. C., (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218 (1),* 132-139.

Witzany, J., Rychnovský M., & Charamza P. (2010). *Survival Analysis in LGD Modeling* (Working Paper No 2). Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies.