JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Comparison of credit duration distribution based on some factors

Muça Markela[1], Puka Llukan[2], Ina Hila[3]

[1,2] Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana Albania

[3] National Agency of Employment and Skills, Albania

Received: 14 April 2020 Revised and Accepted: 8 August 2020

ABSTRACT: Various methods and models have been proposed for estimating not only the risk of a loan (M. Kabir Hasan et al, 2018) but also for the duration of a loan. Survival analysis is a general methodology especially for data-time. It offers great flexibility in duration modeling and statistics calculation, which are valid for evaluations and decision-making. The method can be used on data-time which contains not only information from completed objects but also from ongoing objects.

In this paper we will present a general method for a duration analysis and illustrate it with an application on dataGerman_credit (ftp.ics.uci.edu/pub/machine-learning-databases/statlog/). Through the analysis of survival, based on one or some factors, can be identified the period where an applicant is considered problematic. Also, through this analysis it is required to discover some factors related to the time variable which realize or not differences of impact on the reliability of borrowers. Differences between groups will be illustrated through survival curve obtained by the Kaplan Meier. Log-rank, Gehan Wilcoxon and Torone Ware tests are using to compare survival curves which are part of the SPPS statistical package.

Keywords: Duration time, credit, survival analysis.

INTRODUCTION

Banks produce financial products and services to customers, while managing many risks that are related to liquidity, capital adequacy, credit and interest. The scale of a bank's activities, its relatively low profit margins often combined with a high financial leverage, makes the field of risk assessment and control a vital function for any bank.

Credit is defined as a belief in a person's ability and desire to repay the money offered to them at a later time. Risk management and that of benefit are very closely related to each other. In order to regulate effective management to credit risk exposure, a bank nowadays needs a sophisticated system based on analytical tools for calculating, monitoring and managing risk control. Building models and adapting them into groups and dividing weights that are thought to be dominant through various statistically predictive methods is very tangible today and has an impact on a country's economy.

Several methods have been developed and applied in the banking sector to assess credit risk. Regression analysis (logit analysis) is identified as the most widely used method of CS (Credit Scoring) methods in the banking sector (Valbona Çinaj, 2017). Survival analysis is a general methodology with spaces for traditional methods formulated specifically for data-time, especially in the banking sector (Ricardo Cao et al, 2009). The terminology used for it is not unique.

Statistical models have been proposed in financial risk management as alternative tools to model the interval risk parameter, according to the development improvement required by the New Basel Capital Agreement (Basel II). They are considered a development compared to traditional methods of credit risk analysis, which are based on a distinct response variable (good or bad behavior.

Given that the duration of repayment of a loan is a time variable (measured in months or years, is considered as a dependent variable while the factors that affect the duration can be categorical or continuous; latter are called covariates. Factors and covariates constitute the set of independent variables.), we can perform a duration analysis based on a set of independent variables. The duration of loan can then be considered the time that were initiated on or after the study starting date until before the termination date.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

The duration of a loan as a time variable has two features: First, it is characterized by two critical events, the beginning and the end of a loan for which dates are marked respectively. Second, the duration can be measured as a distance from any point to the initial point.

Survival analysis can be used for two reasons: first, we can construct models for the duration utilizing all the available information, not only from completed objects but also from objects that have not finished yet (M. Kabir Hasan et al., 2018). The last one is important, because even in the most developed studies, there will be subjects who decided to leave during the evaluation, who no longer be followed or there will simply be no event before the end of the study period (censored observations). Censored observations (Efron B., 1977) are necessary because they enable researchers to analyze incomplete data due to delayed entry or exit from the study. Second, these models can predict not only the factors that contribute to default but also the time when the default is more likely to occur.

Credit duration analysis and some examples

In this paper, the duration time is the time from the approval of a loan to the repayment of a loan. The study is focused on estimating the probability of the duration, in comparisons of different loan distributions and in identifying predictive factors related to duration.

In general, we can distinguish two types of approaches in studying the distribution of duration: Parametric: the data distribution is assumed to be known and their parameters are required to be evaluated. These distributions are used to describe the time of occurrence of the event of interest. Non-parametric: the theoretical distribution of data is unknown.

However, there are some situations in the duration data that make it impossible to use some methods, e.g., the duration of some items in the portfolio may be unknown. Generally, such cases in a data set are known as censored time. This situation can also occur in the cases where there are objects that have been lost, the object is sold or the loan is repaid ahead of schedule. Ongoing work on these types of data we will use the term completed loan.

Survival Analysis estimates the survival function throughout the data period. Therefore, it is not necessary to exclude from the study objects that are sold (missing or removed) from the portfolio. Otherwise, the number of objects in the data set would be significance reduced.

The duration of some loans considered to be the value of a random variable Tϵ R . The duration distribution can be described from three functions: probability density function f(t), survival function S(t) and hazard function h(t). It is seen below that these three functions are mathematically equivalent (see Figure 1), in the sense that if one of them is known the other two are derived from a mathematical formula. However, their interpretation is different and can be used in different way to describe the data. At t=0, S(t)=1, H(t)=0 and at t=∞, S(t)=0, H(t)=∞.

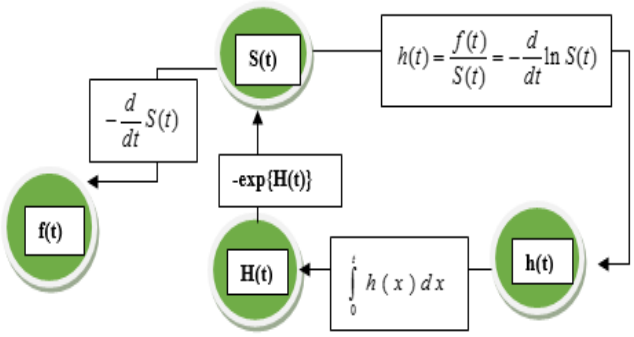


Figure 1. Duration distribution description scheme

In general, f(t) it is a continuous function with positive slope, with a long extension right. Continuous distribution, such as exponential distribution or Weibull distribution are good models for such situations.

The cumulative distribution function F(t) and the survival function S(t) of the random variable T are determined by equations (1) and (2) respectively. The latter gives the probability that an individual's survival time is greater than or equal to t.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

$$F(t) = P(T < t) = \int_0^t f(x)dx \quad (1)$$

$$S(t) = P(T \geq t) = \int_t^\infty f(x)dx \quad (2)$$

Hazard function h(t) (3) is defined as the probability that the individual will fail in time t, assuming that the loan is still active by the beginnig of the interval.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t<T\leq t+\Delta t|T>t)}{\Delta t} \quad (3)$$

From equation

$$P(t \leq T < t + \Delta t / T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)}$$

We have

$$h(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}$$

and $f(t) = \frac{dF(t)}{dt}$

Thus the Hazard rate or Hazard function $h(t)$ is given by (4)

$$h(t) = \frac{f(t)}{S(t)} \quad (4)$$

From equation (2) it is noticed that the derivative of $S(t)$ is equal to $-f(t)$. Then equation (4) is written as follow:

$$h(t) = -\frac{d}{dt} lnS(t) \quad (5)$$

If we integrate equation (5) side by side from 0 to t and as initial condition is taken S(0)=1 (since the event {that did not happen at t=0} is safe event) then equation (6) will be obtained.

$$S(t) = \exp\{-\int_0^t h(x)dx\} = \exp\{-H(t)\} \quad (6)$$

Where H(t) is cumulative conditional rate (cumulative Hazard rate) at the beginning of the interval (from 0 to t). We notice that h(t) has inverse meanings of H(t). Thus, large value of H(t) correspond to small value of h(t) and inversely.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Function S(t) also known as the cumulative survival rate. To describe the survival rate, the graph of the function S(t), known as the survival curve, is constructed. A sloping survival curve represents a low survival rate or a short survival time (figure 2a). A gradual or flat survival turn represents a higher survival rate or longer survival (figure 2b). Also, this type of graph can be used to calculate the percentage (quartile) of survival time (e.g. 25%, 50% and 75%) and to compare the survival distributions of two or more groups. In survival distributions the mean is often a better than the arithmetic mean. For more information see (Lee E.T. & Wang J. W. 2003; Collett David, 2004)).

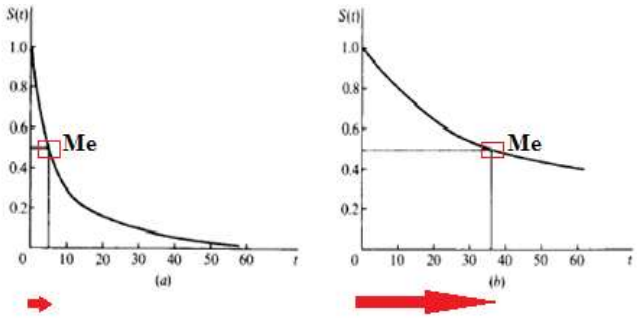


Figure 2. Two examples of survival curves

In Figure 2a and 2b, it is seen that mean values of the survival time are approximately in the time unit 5 and 36 respectively.

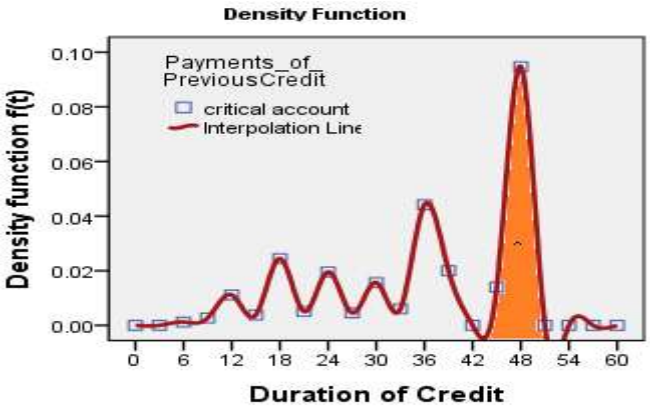
In practice, if there are no censored observations, probability density function, f(t) survival function S(t) and hazard function h(t) are estimated respectively by:

$$\hat{f}(t) = \frac{\text{number of loans featured in the interval beginning at time } t}{(\text{total number of loans}) \times (\text{interval width})}$$

$$\hat{S}(t) = \frac{\text{number of loans with duration} > t}{\text{total number of loans}}$$

$$\hat{h}(t) = \frac{\text{Number of loans featured in the interval beginning at time } t/(\text{interval width})}{(\text{Number of loans active at } t) - (\text{Number of loans featured in the interval})/2}$$

Figure 3 shows an example of a density curve. In Figure 3, the peak of the curve corresponds to the highest value of the failure frequency which occurs on time t=48 months or in the 16th quarter. The fraction of events that fail between the 15th quarter and the 17th quarter time units is equal the areas in red. So, the applicant is more likely to fail in the 16th quarter and less likely in the 5th quarter. The risk function may increase, decrease, remain constant or indicate a mare complicated process. For more information on (Lee and Wang 2003) you can see graphs of several types of risk functions for different problems.



ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Figure 3. One example of density curves for the variable Payments of Previous Credit

Life Table, is a descriptive procedure for examining the distribution of time to event variables. We also can compare the distribution by levels of a factor (e.g.,variable *Payments of Previous Credit* by some categories). The basic idea of life tables is subdivide the period of observation into smaller time intervals with equal width. Then, the probability from each of the intervals are estimated. This method also assumes that the failure rate at one interval is the same for all subjects and is independent of the probability of survival at other time periods. Consider the ordered range of survival time values t_1, t_2,...,t_k. We subdivide the period time into smaller time intervals with equal width $[t_j, t_{j+1}[$.

Notice

n_j-number entering interval on time t_j, w_j -number of terminal events (censored observations) on time t_j, $d_j=n_j-n_{j+1}-w_j$ is the number withdrawing (failed) during the interval $[t_j, t_{j+1}[$ and $r_j=n_j-d_j/2$ is the number exposed to risk during the interval $[t_j, t_{j+1}[$.

The proportion terminating and the proportion surviving (the conditional probability of duration) is estimated respectively by:

$$q_j = w_j / [n_j - (d_j / 2)] ,$$

$$S(t) = P(T > t_{j+1} / T \geq t_j) = 1 - w_j / [n_j - (d$$

The Kaplan-Meier procedure (K-M), is a method for estimating models for time variables according to the interest event in case there is censored data. This is also a descriptive procedure for examining the distribution of time variable. Also, we can compare distributions according levels of one or more factors. So, the differences between groups can be illustrated by the graph of survival curves obtained from the K-M method (Lee E. T. & Wang J. W., 2003), but this gives us a visual comparison and does not reveal whether the differences are statistically significant or not (Kim J. S. & Dailey R. J, 2008). This procedure is based on individual survival times and assumes that censorship is independent of survival time (that is, the reason that an observation is censored is not related to the cause of the failure).

The Kaplan-Meier estimator (Kaplan E. L & Meier P., 1958) of the survival function on time t, S(t), can be defined as

$$KM = \hat{S}(t) = \prod_{j:t_j \leq t} \frac{r_j - d_j}{r_j} \quad \text{for } 0 \leq t \leq t^+$$

Where t_j, j=1,2,….,k is the total set of recorded failure times, d_j is the number withdrawing (failed) during the interval $[t_j, t_{j+1}[$ and r_j is the number exposed to risk during interval $[t_j, t_{j+1}[$.

Survival comparison test

We are often interested in comparing how in group has survived compared to another (defined by categorical covariates). The fact that after training two different groups we obtain two different graphs of survival functions (two survival curves), this helps us to create a clearer picture to identify which are the factors that affect more in survival. It gives bad performance when the two survival functions are crossing over.

Let's $t_1 < t_2 < ... < t_k$ be the event time in two groups. At any time t_j there are d_{1j}, d_{2j} individuals respectively of groups 1 and 2, which fail. At t_j there are n_{1j}, n_{2j} individuals respectively of groups 1 and 2, which are at risk of failure on time t_j. Thus, on time t_j there are $d_j = d_{1j} + d_{2j}$ total failures from

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

$n_j = n_{1j} + n_{2j}$ individuals who are at risk. So, a contingency table is constructed. Then, the general test statistics is calculated (Leton E. & Zuluanga P., 2005) by

$$T\_statistic = \frac{(\sum_{j=1}^{k} w_j(d_{ij} - d_j E(T_{1j})))^2}{\sum_{j=1}^{k} w_j^2 D(T_{1j})}, \quad E(T_{1j}) = d_j \frac{n_{1j}}{n_j} \text{ and } D(T_{1j}) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

where w_j is the weight at time t_j.

Long rank (Mantel, 1966) tests equality of survival functions by weighting all time points the same, $w_j=1$. It gives bad performance when the two survival functions are crossing over.

Gehan Generalized Wilcoxon or Breslow (Gehen E., 1965; Breslow A., 1970) tests equality of survival functions by weighting all time points by the number of cases at risk at each time, $w_j=n_j$

Tarone-Ware (Tarone, R.E. & Ware, J.,1977) tests equality of survival functions by weighting all time points by the square root of the number of cases at risk at each time point, $w_j = \sqrt{n_j}$.

For additional details about the survival comparison test, see (Pinar Gunel Karadeniz & IIker Ercan, 2017).

The proportional hazard model, proposed by Cox in 1972 (Cox David R.,1992) is the most popular regression technique in survival analysis, which is used to link several risk factors, considered simultaneously with survival time. In this model, the measure of effect is the degree of risk, which is the risk of failure (i.e. the risk or the probability of suffering the event of interest), given that the subject has survived up to a certain time. Risk represents the expected number of events per unit of time, as a result the risk in a group may exceed 1. The Cox regression model is expressed by the risk function denoted by h(t). Briefly, the risk function can be interpreted as the risk of failure on time t. It can be estimated as follows

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... \beta_p x_p)$$

where *t* represents survival time, *h(t)* is the risk function defined by a group with p-variables $(x_1, x_2, ..., x_p)$ and *p* corresponding coefficients $(\beta_1, \beta_2, ..., \beta_p)$, which measures the impact of the dependent variables of time t, the term h_0 is called baseline hazard ratio, which is the value when all the $\beta's$ equal zero.

The Cox model can be written as a multiple linear regression of the logarithmic risk function in the variables x_i, with the baseline risk being the term changing over time. Quantities $\mathrm{Exp}(\beta_i)$ are called H*azard Ratio, HR).*

- A value of β_i -s greater than zero, or equivalent a risk ratio greater than one ($\mathrm{Exp}(\beta_i) > 1$), indicates that as the value of the i-th variable increases, the risk of the event increases. Thus, the length of survival decreases. A risk ratio above 1 indicates a variable that is positively related to the probability of events. Consequently, is negatively related to the length of survival. So, when the value of HR=1 then we say has no effect, when the value of H<1 we say there is a reduction in risk and when the value of HR>1 we say there is an increase risk.

We can take the natural logarithm of both sides, to give what

is presented below, which connects the relative risk logarithm

to a linear predictor function

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

$$\ln \frac{h(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

The *Wald test* will be applied to select the variable that best explain the risk function. Hypothesis $H_0 : \beta = 0$ can be controlled by calculating the value of statistics $T = \tilde{\beta} / \sigma(\tilde{\beta})$. In general, point estimator for each of the parameters β_i in the proportional risk model, there are not all independent of each other, therefore care should be taken in their choice because if they are not statistically significant in the specific step but may be important when other variables are added to the model **(**Collett David, 2004).

To compare alternative models, a statistics is needed which shows that one model is more appropriate than another. Since the likelihood function summarizes information those data contains about unknown parameters in a given model, an appropriate statistics is the value of the likelihood function when the parameters are replaced by their maximum likelihood estimators.

For a data set, the higher the value of the maximum likelihood function, the stronger the relationship between the model and the observation. The maximum likelihood function $\tilde{L}$ is the product of some conditional probabilities. So, the value of statistic is less than 1. Consequently, $-2\log\tilde{L}$ will always be positive. For a data set, the smaller the value of $-2\log\tilde{L}$ the better the model. $-2\log\tilde{L}$ statistics can no longer be used as a model of degree of suitability because the value of $\tilde{L}$ and $-2\log\tilde{L}$ is dependent on a number of observations in the data set. Also, the value of $-2\log\tilde{L}$ is valid only to make comparisons for models that fit the same data set (Collett David, 2004).

Application to real data

We will use a sample with real data, known German credit, the currency is the German Mark (Deutsche Mark, *DM*), to establish a credit rating rule that can be used to determine whether a new applicant is trustworthy or not, based on the values of one or more predictor variables. The set of data in German credit provides information for n=1000 loan applicants. It contains two dependent variables: credibility and duration of credit, and 16 predictor variables. The dependent variable, credibility, indicates whether an applicant has been approved credit (is a reliable person) depending on 16 independent variables. The variables of the set are in Table 1 reported.

Classification of loans according to credit risk classes, is realized mainly on the basis of the delays the client shows towards the payment of obligations. Loans with more than 90 days delay (3 months) in repayment are considered NPL problem loans.

First, the lifetime is calculated (see Table 2) for the variable *Payments of Previous Credit* (PPC), whereas time dependent variable has *Duration of Credit*, status is credibility, so who

reliable is he to the bank and event interest is {applicant is not reliable} = 0 to then calculate the probability of an event of interest during each time interval that is under study.**Table 1.** Description of the variables in the data set.

Nr.	**Variables**	**Categories**	**Units of Measurement**
1	**Creditability**	0=No , 1=Yes	Binare
2	Account_Balance	1={<0 DM}, 2={0<=...<200 DM}, 3={>=200} 4={no checking account}	DM
3	**Duration_of_Credit**	Scale {4:72 months}	Month
4	Payments_of_Previous_Credit	0={no credits taken}, 1={all credits at this bank paid back duly}, 2={existing credits paid back duly till now}, 3={delay in paying off in the past}, 4={critical account}	Categorical

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

5	Credit_Amount	Scale	DM
6	Value_Savings	1={< 100 DM}, 2={100<= ... < 500 DM}, 3={500<= ... < 1000 DM}, 4={=>1000 DM}, 5={unknown/ no savings account}	DM
7	Length_of_current_ Employment	1={unemployed}, 2=< 1 year, 3={1 <= ... < 4 years}, 4={4 <=... < 7 years}, 5={>= 7 years}	Years
8	Marital_Status	1={divorced}, 2={single}, 3={married},4={widower}	Categorical
9	Guarantors	1={has co-applicant}, 2={has a guarantor}, 3={has not a guarantor}	Categorical
10	Duration_in_Current_ Address	1={<= 1 year}, 2={1<…<=2 years}, 3={2<…<=3 years}, 4={3:>4years}	Years
11	Most_valuable_ available_asset	1={Applicant owns real estate}, 2={Applicant owns not real estate}, 3={Applicant owns no property}, 4={unknown}	Categorical
12	Age	Scale	Years
13	Concurrent_Credits	1={No}, 2={Yes}, 3={Unknown}	Categorical
14	Type_of_apartment	1={Applicant rents}, 2={owns residence}, 3={No information}	Categorical
15	Credits_at_this_Bank	Scale	Number
16	Occupation	1={unemployed/ unskilled -non-resident}, 2={unskilled - resident}, 3={skilled employee / official}, 4={nt/ self-employed/highly qualified employee/ officer}	Categorical
17	No_of_dependents	Scale	Number
18	Foreign_worker	1={No}, 2={Yes}	Categorical
19	Purpose	0={New house},1={Business},3={Health}, 4={New car},5={Old car}, 6={Furniture},7={TV},8={Education}.9={Retraining}, 10={Overall}	Categorical

From Table 2, it is seen that for the first categories of applicants (no credits taken and all credits at this bank paid back duly) based on *payments of previous credit*, the applicant identifies unreliable in the *15-th month*. This group of applicants should be closely monitored during this period, to increase security against them. Regarding other categories

related to the *Payments of Previous Credit* of applicants, monitoring should be done for a longer period. Here we can disjoint the group of applicants with critical account, which consists of a much larger number of unreliable cases. For this

category of applicants, this is an expected result while these have the largest fluctuations in pay/non-pay.

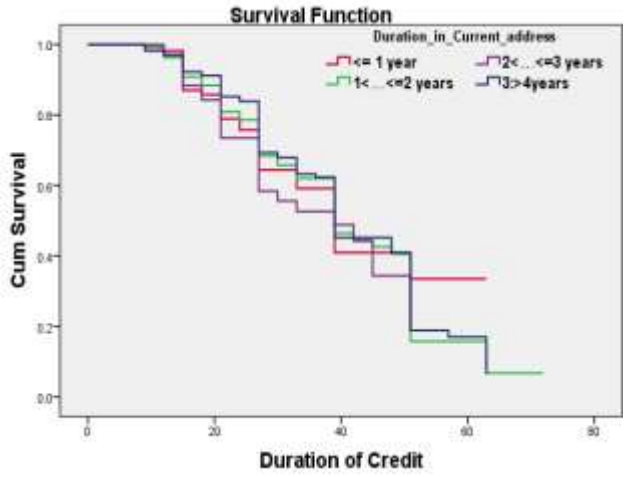
ISSN- 2394-5125VOL 7, ISSUE 15, 2020

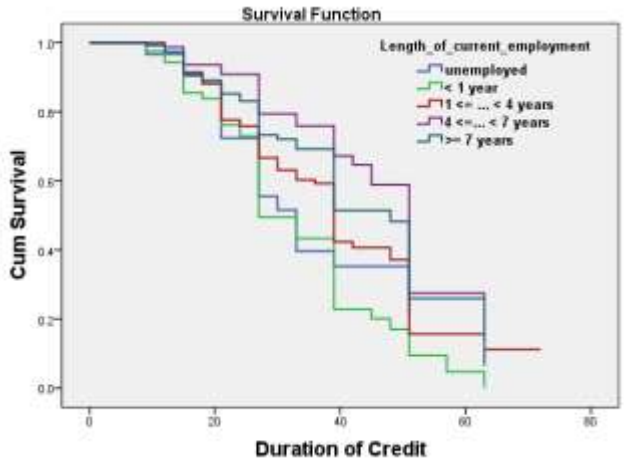
Table 2. Life Table for the variable *Payments of Previous Credit.*

Payments_of_Previous Credit	Interval Start Time t_{j+1}-$t_{j=3}$	Number Entering Interval n_j	Number withdrawing during Interval $d_j=n_j-n_{j+1}-w_j$	Number Exposed to Risk $r_j=n_j-d_j/2$	Number of Terminal Events w_j
no credits taken	0	40	0=40-40-0	40.000=40-0/2	**0**
	3	40	0=40-40-0	40.000=40-0/2	**0**
	6	40	0=40-38-2	40.000=40-0/2	**2**
	9	38	1	37.500=38-1/2	**0**
	12	37	4	35.000	**2**
	15	31	2	30.000	**0**
	…	…	…	…	…
	54	2	1	1.500	**1**
all credits at this bank paid back duly	0	49	0	49.000	**0**
	3	49	0	49.000	**0**
	6	49	3	47.500	**2**
	9	44	2	43.000	**1**
	12	41	4	39.000	**3**
	15	34	2	33.000	**1**
	…	…	…	…	…
	60	1	1	.500	**0**
existing credits paid back duly till now	0	530	0	530.000	**0**
	3	530	0	530.000	**3**
	6	527	5	524.500	**37**
	9	485	12	479.000	**42**
	12	431	34	414.000	**73**
	15	318	6	315.000	**29**
	18	283	23	271.500	**39**
	21	221	3	219.500	**14**
	…	…	3	101.500	**3**
	…		…	…	…
	72	1	1	.500	**0**
delay in paying off in the past	0	88	0	88.000	**0**
	3	88	0	88.000	**0**
	…				

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

	18	69	3	67.500	**7**
	21	59	2	58.000	**3**
	24	54	7	50.500	**13**
	27 …	34 …	0 …	34.000 …	**3** …
	60	4	1	3.500	**3**
critical account	0	293	0	293.000	**0**
	3	293	0	293.000	**4**
	6	289	1	288.500	**30**
	9	258	2	257.000	**25**
	12	231	7	227.500	**48**
	15	176	2	175.000	**19**
	18	155	11	149.500	**27**
	21 …	117 …	2 …	116.000 …	**6** …
	60	1	0	1.000	**1**





ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Figure 4: Curve of survival function (a) Duration in current address, (b) Length of current employmentIn another analysis, the survival curve which gives a visual representation of the life table, see Fig 4, confirms conclusions that have results from the life table for two variables*; duration in current address** and *length of current employment**. In such graphical representations, reductions in the survival curve occur whenever a factor (e.g., *length of current employment**) enters into force. Any point on the survival curve shows the probability that an applicant of a given length of current address* or length of current employment* category will remain an applicant past that time. Also, in the presentation of this graph is seen the effect of each factor by the category. Thus, for the variable length of current address* we can not say which category of applicants has the biggest/the smallest survival effect as there are many overlaps. On the other hand, for the variable length of current employment* changes in the survival curve are observed. The survival curve changes higher turns for the groups of applicants with 4 to 7 years of work, while lower for those

with less than one year of work. In terms of survival time,

applicants with 1 to 4 years of work have the longest time in loan repayment. In other words, these are considered more stable loan payers.

Wilcoxon test is used to compare survival distribution among groups (categorical variables), with the test statistics based on differences in groups mean scores.

Table 7 presents results given by SPSS for each factor and covariates, where status is the dependent variable *credibility* and the event interest is {applicant is not reliable} =0. It is noticed that, at least two survival curves are different for the variable's length of current employment* (5-categories 5-curves), where significance value for the factor length of current employment is equal $0 < 0.05$, while for the factor *length of current address* where the significance value is equal to $0.282 > 0.05$, there are no distinct survival curves between length of current address categories. This mean that there are no differences in the impact on applicants' reliability from their duration in a given length of current addrss.

Table 3: Pairwise comparisons for variables (a) *Duration in current address*, (b) *Length of current employment*

Duration_in_ current_address (a)		**Statistics Wilcoxon**	**Df**	**Sig.**	***Length_of_current _employment (b)***		**Statistics Wilcoxon**	**Df**	**Sig**
1= {<=1 yera}	2	0.249	1	0.617	1= {unemployed}	2	0.240	1	0.624
	3	0.179	1	0.672		3	1.124	1	0.289
	4	1.338	1	0.247		4	9.320	1	0.002
2= {1-2 years}	1	0.249	1	0.617		5	3.450	1	0.063
	3	1.102	1	0.294	2= {<1 year}	1	0.240	1	0.624
	4	0.666	1	0.414		3	5.717	1	0.017
3= {2-3 years}	1	0.179	1	0.672		4	20.252	1	0.000
	2	1.102	1	0.294		5	10.841	1	0.001
	4	3.526	1	0.060	3= {1-4 years}	1	1.124	1	0.289
4= {>=4 years}	1	1.338	1	0.247		2	5.717	1	0.017
	2	0.666	1	0.414		4	8.104	1	0.004
	3	3.526	1	0.060		5	1.473	1	0.225
					4= {4-7 years}	1	9.320	1	0.002
						2	20.252	1	0.000
						3	8.104	1	0.004
						5	2.548	1	0.110
					5=	1	3.450	1	0.063

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

	{>7 years}	2	10.841	1	0.001
		3	1.473	1	0.225
		4	2.548	1	0.110

Pairwise comparisons show which two groups are significantly different in survival curves.

The pairwise comparisons test showed that the survival curves for the group of applicants with 4 to 7 years of work are statistically similar only for the group of employed over 7 years of work. The latter is statistically different only for the

group of applicants with less than 1 year of work.

Similarly, we can reason for the variable d*uration in current address*. It is noticed that, the survival curves according to a certain level (years in a certain settlement) are not statistically distinct with each of the other levels.

Table 4. Mean of survival time for two variables; *Duration in current address* (a) *Duration in current employment* (b)

***Duration_in_Current_address* (a)**			***Duration_in_Current_employment* (b)**		
First order control		Mean time	First order control		Mean time
Duration_in_current_address	<1	39.06	***Duration_in_current_employment***	Unemployed	31.07
	1<...<=2	40.31		<1	29.73
	2<...<=3	38.45		1<=...<4	39.24
	>=4	40.30		4<=...<7	49.67
				>7	44.33

From Table 4, it is observed the unemployed applicants and applicants with less than 1 year of work have a lower average survival time compared to the applicants of the other three groups. On the other hand, for the variable duration in current address changes in mean survival time values are less noticeable.

Applicants who live in the same residence (*duration in current address)* from 2 to 3 years have lowest survival time.

This means that the latter the applicants with less than 1 year of work have a lower level of reliability compared to the applicants of other groups.

The Table 5 shows that, the confidence interval of the mean and median of survival time respectively, overlap. This means that there are unlikely to be many changes in the mean of survival time.

Table 5. Mean and median of survival time for the variable *value savings*

Value_ Savings DM	***Mean*[a]**				***Median***			
	Estimate	**Std. Error**	**95% CI**		**Estimate**	**Std. Error**	**95 % CI**	
			Lower bound	**Upper bound**			**Lower Bound**	**Upper bound**
<100	35.083	.901	33.318	36.848	36.000	1.205	33.638	38.362
100<=...<500	42.270	3.085	36.224	48.317	45.000	4.514	36.152	53.848
500<=…<1000	38.840	2.614	33.717	43.963	48.000	13.377	21.780	74.220
>=1000	43.128	2.250	38.718	47.538	48.000	0.000		-
Unknown/no	47.697	1.852	44.068	51.326	-	-	-	-

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

savings account								
Overall	40.324	1.047	38.272	42.377	36.000	1.624	32.817	39.183

a. Estimation is limited to the largest survival time if it is censored.

From Table 6, we see that *account balance*, *payments of previous credit*, *value savings*, *length of current employment*, *marital status*, *age*, *type of apartment*, *occupation* and *purpose* are nine of the sixteen factors that most influences the difference whether a customer is reliable or not.

The 9 most important variables were identified from the KM procedure. The Cox Regression procedure will be applied. The backward linear regression method, BRL, will be used to select the important independent variables. The goodness of the model is valued through statistics $-2\log\tilde{L}$.

We will construct some model.

The first is obtained starting from the complete model (with all the important variables that resulted from the K-M procedure) to the best model (with the least number of variables). The values of -2LL will be saved in a table (Table 10) for each variable of the corresponded models. The smaller the value of -2logL the more the variables of the model effect the risk function.

Table 6. Comparison for control variables with Log-rang test, Breslow test and Tarone-Ware test**Variables**	**Log Rank (Mantel-Cox)**			**Breslow** (Generalized Wilcoxon)			**Tarone-Ware**		
	Chi-Square	df	Sig.	Chi-Square	Df	Sig.	Chi-Square	Df	Sig.
Account_Balance	77.686	3	0.000	52.14	3	0.000	63.641	3	0.000
Payments_of_Previous_Credit	24.481	4	0.000	33.319	4	0.000	31.011	4	0.000
Value_Savings	35.603	4	0.000	17.845	4	0.001	24.067	4	0.000
Length_of_current_employment	34.007	4	0.000	24.526	4	0.000	29.893	4	0.000
Marital_Status	23.779	3	0.000	18.229	3	0.000	21.835	3	0.000
Age	81.046	52	0.006	68.171	52	0.066	73.372	52	0.027
Type_of_apartment	19.319	2	0.000	8.761	2	0.013	13.014	3	0.001
Occupation	13.032	3	0.005	12.077	3	0.007	13.641	3	0.003
Purpose	56.146	9	0.000	61.320	9	0.000	64.187	9	0.000
Guarantors	1.928	2	0.381	2.668	2	0.263	2.586	2	0.274
Duration_in_Current_address	1.863	3	0.601	3.877	3	0.275	3.242	3	0.365
Most_valuable_available_asset	5.824	3	0.121	6.421	3	0.093	6.858	3	0.077
Concurrent_Credits	3.904	2	0.142	2.691	2	0.260	3.423	2	0.182
Credits_at_this_Bank	1.169	3	0.655	5.173	3	0.160	3.641	3	0.303
No_of_dependents	0.000	1	0.993	0.326	1	0.568	0.211	1	0.646
Foreign_Worker	0.002	1	0.968	0.253	2	0.615	0.101	1	0.751

Results and discussions:

Table 7. Variables in the Equation

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Variables		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
STEP 3	Account_Balance **(AB)**	**-.318**	.057	30.755	1	.000	.727	.650	.814
	Payments_of_Previous_Credit **(PPC)**	**-.148**	.055	7.107	1	.008	.863	.774	.962
	Purpouse **(P)**	**-.066**	.021	10.272	1	.001	.936	.899	.975
	Value_Savings **(VS)**	**-.156**	.047	11.255	1	.001	.855	.781	.937
	Length_of_current_employment **(LCE)**	**-.131**	.050	6.891	1	.009	.878	.796	.967
	Type_of_apartment **(TA)**	**-.272**	.099	7.480	1	.006	.762	.627	.926
	Occupation **(O)**	**-.180**	.090	3.982	1	.046	.835	.700	.997

The proportional risk model is given by

$$\log\left(\frac{h(t)}{h_0(t)}\right) = -0.318AB - 0.148PPC - 0.066P - 0.156VS - 0.131LCE - 0.272TA - 0.18O \quad (7)$$

From the results presented in Table 7, it can be seen that seven are the variables that best explain the credibility of loan applications; *account balance* (AB), *payments of previous credit* (PPC), *purpose* (P), *value savings* (VS), *length of current employment* (LCE), *type of apartment* (TA) and *occupation* (O). In general, we divide these variables into two groups according to the impact they have on credit risk. Also, it is seen that the value of $Exp(\beta) < 1$ for each of the important variables. Consequently, these seven variables form a group. This group consists of seven variables that affect the reduction of credit risk.

The default risk, DR, is calculated for each of the significant variables according to the equation

$$DR = 100\% - \left(100\% \times \text{Exp}(\beta)\right)$$

This is a numerical indicator expressed as a percentage whose values are presented in Table 9 (first column). The value of $Exp(\beta)$ for the *purpose* variable indicates that the risk of default is reduced with 6.4%. This variable reduces the risk part less than do the other variables. For this reason, we have extracted this variable from the model. Then, we re-apply the Cox model to the remaining 6 variables. The results are presented in Table 8a and Table 9 (second column). Note that, when the purpose variable is extracted from the model, the occupation variable becomes insignificant with alpha=0.055>0.05. Then, we extract the *occupation* variable and re-apply the Cox model. The results are presented in Table 8b and Table 9 (third column).**Table 8a.** Variables in the Equation

Variables		B	SE	Wald	Df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
STEP 1	Account_Balance **(AB)**	**-.332**	.057	33.376	1	.000	.718	.641	.803
	Payments_of_Previous_Credit **(PPC)**	**-.114**	.053	4.577	1	.032	.892	.803	.990
	Value_Savings **(VS)**	**-.147**	.046	10.133	1	.001	.863	.788	.945
	Length_of_current_employment **(LCE)**	**-.134**	.049	7.426	1	.006	.874	.794	.963
	Type_of_apartment **(TA)**	**-.234**	.098	5.673	1	.017	.791	.653	.959
	Occupation (O)	**-.171**	.089	3.688	1	**.055**	.843	.707	1.004

The proportional risk model is given by

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

$$\log\left(\frac{h(t)}{h_0(t)}\right) = -0.332\,AB - 0.114\,PPC - 0.147\,VS - 0.134\,LCE - 0.234\,TA - 0.171\,O \quad (8)$$

Table 8b. Variables in the Equation

Variables		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	Account_Balance **(AB)**	-.343	.057	35.950	1	.000	.710	.635	.794
	Payments_of_Previous_Credit **(PPC)**	-.110	.054	4.222	1	.040	.896	.806	.995
	Value_Savings **(VS)**	-.143	.046	9.645	1	.002	.866	.791	.948
	Length_of_current_employment **(LCE)**	-.143	.049	8.580	1	.003	.867	.788	.954
	Type_of_apartment **(TA)**	-.255	.098	6.816	1	.009	.775	.640	.938

The proportional risk model is given by

$$\log\left(\frac{h(t)}{h_0(t)}\right) = -0.343AB - 0.11PPC - 0.143VS - 0.143LCE - 0.255TA \quad (9)$$

The value of $Exp(\beta)$ for the l*ength of current employment* variable indicates that the risk of default is reduced by 10.4% less than the values (risk of default) of the other 4 variables.

We extract this variable from the model and re-apply the Cox model to the remaining 4 variables. The results are presented in Table 8c and Table 9 (fourth column).

Table 8c. Variables in the Equation

Variables		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	Account_Balance (AB)	**-.362**	.056	41.535	1	.000	.696	.624	.777
	Value_Savings (VS)	**-.141**	.046	9.361	1	.002	.868	.793	.950
	Length_of_current_employment (LCE)	**-.149**	.048	9.473	1	.002	.861	.783	.947
	Type_of_apartment (TA)	**-.268**	.098	7.443	1	.006	.765	.631	.927

The proportional risk model is given by

$$\log\left(\frac{h(t)}{h_0(t)}\right) = -0.362AB - 0.141VS - 0.149LCE - 0.268TA \quad (10)$$

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

Table 9. Non-payment risk for the 4 models according to each of the important variables.**Variables**	**Eq_7**	**Eq_8**	**Eq_9**	**Eq_10**
Account_Balance **(AB)**	0.273	0.282	0.290	0.304
Type_of_apartment **(TA)**	0.238	0.209	0.225	0.235
Occupation (O), alpha=0.055	0.165	**0.157**		
Value_Savings	0.145	0.137	0.134	0.132
Payments_of_Previous_Credit **(PPC),** the smallest value of the risk	0.137	0.108	**0.104**	
Length_of_current_employment **(LCE)**	0.122	0.126	0.133	0.139
Purpouse **(P)** the smallest value of the risk	**0.064**			

he results in Table 10 show that, *the value of statistics -2logL* decreases when insignificant variable is extracted in the model, while the value of statistics -2logL increases when in the model variable is extracts based on the part of the risk. In fact, model with fewer variable and the lowest statistical value -2logL are required. We choose equation (10) as the best model, since the value of -2logL does not increase significantly.

The value of $\mathrm{Exp}(\beta)$ for the variable *account balance*

indicates that the risk of default is reduced by 30.4% for the group of applicants with higher savings. The value of $\mathrm{Exp}(\beta)$ for the other three variables; *type of apartment* (TA), *length of current employment* (LCE), and *value savings* (VS) indicates that the risk of default is reduces by 23.5 %, 13.9% and 13.2% respectively for each group of applicants with owns residence, with more years of work in a certain place and with higher savings account.

Table 10. Comparison of models through statistics $-2\log L$

Model	**-2 Log Likelihood**	**Hi-square**
Null	3479.522	
[(a)]AB+PPC+P+VS+LCE+MS+A+TA+O	3356.792	118.861
[(b)] AB+PPC+P+VS+LCE+MS+ATA+O	3356.794	118.682
[(c)] AB+PPC+P+VS+LCE+MS+TA+O	3359.012	114.681
[(d)] AB+PPC+P+VS+LCE+TA+O, risk=6.4%	3362.917	103.137
[(e)] AB+PPC+VS+LCE+TA+O, alpha=0.055	3373.557	98.736
[(f)]AB+PPC+VS+LCE+TA, risk=10.4%	3377.768	95.673
AB+VS+LCE+TA		

Conclusions

With the life table procedure can be discovered whether the categories within each of the variables under consideration, are distinct from each other or not. The results of this procedure can be plotted with survival curves.

The analysis showed that, the group of applicants belonging to one of the firsts two categories; no credits taken and all credits at this bank paid back duly (based on credit history) should be monitored within the first 15 months while for the group of the other three categories applicants should be monitored for a longer period. Here we can disjoint the group of applicants with critical account, which consists of a much larger number of unreliable cases.

From the graph of the survival curve resulted that applicants with 1 to 4 years of work are considered to be the most stable payer of loans. It was also noted that there are no differences in the impact on applicants' reliability from their duration in a given residence. Applicants who live in the same residence (*duration in current address)* from 2 to 3 and those with less than 1 year of work have a lower level of reliability compared to the applicants of other groups.

ISSN- 2394-5125VOL 7, ISSUE 15, 2020

The Cox model results that 7 are the most important variables which affect the reduction of credit default risk. These variables can be divided into two groups:

- **Account balance**, **type of apartment**, **length of current employment**, and **value savings** are the four variables that affect the reduction of credit default risk.
- **Payments of previous credit, occupation** and **purpose** are the three variables which have the least impact on reducing the risk of loan default.

Results are described via a database from a German credit. We are considering a possible database with information from our country for similar analysis of practical interest as a future task.

REFERENCES

[1] Breslow A., 1970. Thickness, Cross-Sectional Areas and Depth of Invasion in the Prognosis of Cutaneous Melanoma. Annals of Surgery, 172, 902-908. https://doi.org/10.1097/00000658-197011000-00017
[2] Collett David, 2004. Modelling survival data in medical research (Chapman & Hall-CRC, 2004) (ISBN 1584883251)
[3] Cox, David R., 1992. "Regression models and life-tables." Breakthroughs in statistics. Springer New York, 1992. 527-541.
[4] Efron, B., 1977. The Efficiency of Cox's Likelihood Function for Censored Data. Journal of the American Statistical Association.
[5] IBM Corporation 1989, 2016. IBM SPSS statistics V24.0 documentation
[6] Gehan, E., 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika, 52, 203-223. https://doi.org/10.2307/2333825
[7] Kaplan E.L. & Meier P., 1958. Nonparametric Estimation from Incomplete Observations. JASA, 53, pp.457-481.
[8] Kim J. & S. Dailey, R. J., 2008. Biostatistics for oral healthcare, Blackwell Publishing Company, Iowa, pp. 287–291.
[9] Lee E. T. & Wang J. W., 2003. Statistical Methods for Survival Data Analysis. Third-Edition-Wiley-Series-in-Probability and statistics.pdf ISBN 0-471-36997-7
[10] Mantel (1966) Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemotherapy Reports, 50.
[11] M. Kabir Hasan, Jennifer Brodman, Blake Rayfield & Makeen Huda, 2018. Modelling Credit Risk in Credit Unions using Survival Analysis. International Journal pf Bank Marketing.DOI:10.1108/IJBM-05-2017-0091.
[12] https://www.researchgate.net/publications/323921643