PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 381

Financial Investments and Insurance – Global Trends and the Polish Market

edited by
Krzysztof Jajuga
Wanda Ronka-Chmielowiec

Publishing House of Wrocław University of Economics
Wrocław 2015

Copy-editing: Agnieszka Flasińska

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

Information on submitting and reviewing papers is available on
the Publishing House's website
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
e-ISSN 2392-0041

ISBN 978-83-7695-463-9

The original version: printed

Publication may be ordered in Publishing House
tel./fax 71 36-80-602; e-mail: econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Printing: TOTEM

Contents

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 381 • 2015

Financial Investment and Insurance –
Global Trends and the Polish Market

ISSN 1899-3192
e-ISSN 2392-0041

Ewa Wycinka
University of Gdańsk
e-mail: ewa.wycinka@ug.edu.pl

TIME TO DEFAULT ANALYSIS IN PERSONAL CREDIT SCORING

Summary: Credit scoring is a technique mainly used in making consumer credit decisions. Traditional credit scoring systems aim to estimate the probability that an applicant will default. However, for the financial institution, not only if but also when the creditor defaults is important. The aim of this paper is to analyse the usefulness of survival methods, especially the Kaplan-Meier estimator, in the process of making decisions of credit granting, as well as in monitoring credit portfolios. The focus is put on the differences in the influence of particular characteristics used in scoring models on the probability of default and the probability of early repayment during credit life-span. The survival analysis approaches to credit scoring were tested on 60-month personal loan data from one of the Polish financial institutions. All loans were observed until 24 months after inception, or until default or earlier repayment if one of these happened earlier. In the first part of analysis, early full repayments qualified as censored data, and in the second as failures.

Keywords: Credit scoring, survival analysis, Kaplan-Meier estimator.

DOI: 10.15611/pn.2015.381.38`

1. Introduction

Credit scoring is a technique mainly used in making consumer credit decisions. Traditional credit scoring systems aim to estimate the probability that an applicant will default. However, for the financial institution it is important to consider not only if but also when the creditor defaults. With an increasing number of repaid instalments the loss of the lender decreases. If the time of default is long, the acquired interest will compensate or even exceed the value of credit [Stepanova, Thomas 2002]. Most creditors fully repay the loan, so default will never occur for them.

Such an approach to credit risk has a lot in common with survival analysis. In survival analysis, interest centres on a group of individuals for each of whom there is a defined point, an event called failure, occurring after a length of time, which is called the failure time. Failure can occur once at most for any individual [Cox, Oakes 1984]. Narain [1992] first introduced survival analysis methods to credit scoring. He applied a proportional hazard model to loan data. Since then, several research studies have been carried out on this topic. Banasik, Crook and Thomas [1999] applied three types of proportional hazard models and accelerated life models to loan data and compared the results with regression scorecard approaches. They also considered competing risk approaches to loans, assuming that there are two reasons for repayment schedule interruption: defaults and early repayments. Stepanova and Thomas [2001] explored different applications of proportional hazard models to behavioural scoring. They used survival probability profiles of customers to calculate the expected profit from a loan. In an article published in 2002, these authors focused on competing risk issues. They proposed the method of coarse classifying variables with the use of Cox's proportional hazard models with binary variables. They proved that this method is more appropriate when using survival analysis modelling than traditional approaches. They also underlined the necessity of separate splits for every type of failure considered [Stepanova, Thomas 2002]. Mavri et al. [2008] suggested a two-stage dynamic credit scoring model; while Cao, Vilar and Devia [2009] applied GLM under censoring and a nonparametric kernel method to estimate the default probability. A review of the improvements in credit modelling techniques was offered in [Thomas, Oliver, Hand 2005] and [Marques, Garcia, Sanchez 2013].

The aim of this paper is to analyse the usefulness of the survival method, especially the Kaplan-Meier estimator of survival function, in the process of making decisions about credit granting, as well as in monitoring of credit portfolios. The focus is on the differences in the influence of particular characteristics used in scoring models on the probability of default and early repayment during credit life-span.

2. Theory of analysis of lifetime data

Let T be the random variable representing time until failure. The cause of failure could be default or early repayment. In the first part of the analysis, time is estimated until default, assuming time until early repayment is censored. In the second part, time until early repayment is a failure, while time until default is censored. Therefore, survival analysis will be performed separately on both types of failures.

Time-to-event is described by a survival function, the probability of an entity surviving beyond time t. It is defined as

$$S(t) = P(T > t).$$

When T is a continuous random variable, the survival function is a continuous, strictly decreasing function [Klein, Moeschberger 1997, p. 22].

The standard estimator of the survival function is the product limit estimator proposed by Kaplan and Meier [1958]. It is defined as

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \prod_{t_i \le t} (1 - \frac{d_i}{l_i}) & t > t_1 \end{cases},$$

where: d_i – a number of events at time t_i; l_i – a number at risk at time t_i [Klein, Moeschberger 1997, p. 84].

The graphical presentation of the KM estimator is a step curve that starts with a horizontal line at a survival probability of 1, and then steps down to the other survival probabilities to follow ordered failure times. The KM estimator is based on an assumption of non-informative censoring, which means that knowledge of a censoring time for an individual provides no further information about this entity's survival at a future time, should the individual continue the study [Klein, Moeschberger 1997, p. 91].

To determine whether there is a significant difference between two or more survival curves, one can test the hypothesis. The hypothesis tests on the equality of Kaplan-Meier curves can be conducted using one of the several available statistical tests designed for this purpose, but the most commonly used is a log-rank test [Suciu, Lemeshow, Moeschberger 2004, p. 252]. The null hypothesis is that all survival curves are the same. If the number of groups (k) being compared is more than two, the log-rank statistic has asymptotic chi-square distribution with $k - 1$ degrees of freedom. The mathematical formula of the test statistic can be found in [Kleinbaum, Klein 2005, p. 82]. The computer package used in empirical analysis uses Mantel's procedure to calculate test statistics. First, a score is assigned to each survival time; next, a chi-square value is computed based on the sums (for each group) of this score. If only two groups are specified, then this test is equivalent to Gehan's generalized Wilcoxon test. The characteristics of this test can be found *inter alia* in [Jurkiewicz, Wycinka 2011, p. 107–114].

3. Data set

The survival analysis approaches to credit scoring were tested on 60-month personal loan data from one of the Polish financial institutions. The data consisted of application information for 5000 loans accepted in the period of the following six months. All loans were observed for 24 months or until failure. The failure was default or earlier repayment. The default was defined as 90 days' lateness in payment of instalments.

The initial characteristics included information about the creditor such as, age, marital status and residency type, as well as about the loan: the purpose of the loan and its capacity. Altogether, 15 characteristics have been taken into account. They have been coded by letters: X_1, …, X_{15}.[1] In the first step, the application

[1] Due to the know-how of the financial institution that shared data for the purposes of this research, detailed information about the creditors and characteristics used in analysis was encrypted.

characteristics were split into attributes.[2] The continuous variables were subdivided into fractiles. Then K-M curves were plotted for each fractile. Groups with the most similar curves were connected. In categorical variables, attributes with few units were joined to similar ones. Then, on the basis of the shape of survival curves, the number of groups was reduced by putting together those with similar curves. An example of such a procedure for variable X_{12} is presented in Figures 1 and 2. Figure 1 shows K-M curves for 10 groups of values created after connecting small attributes with similar ones.

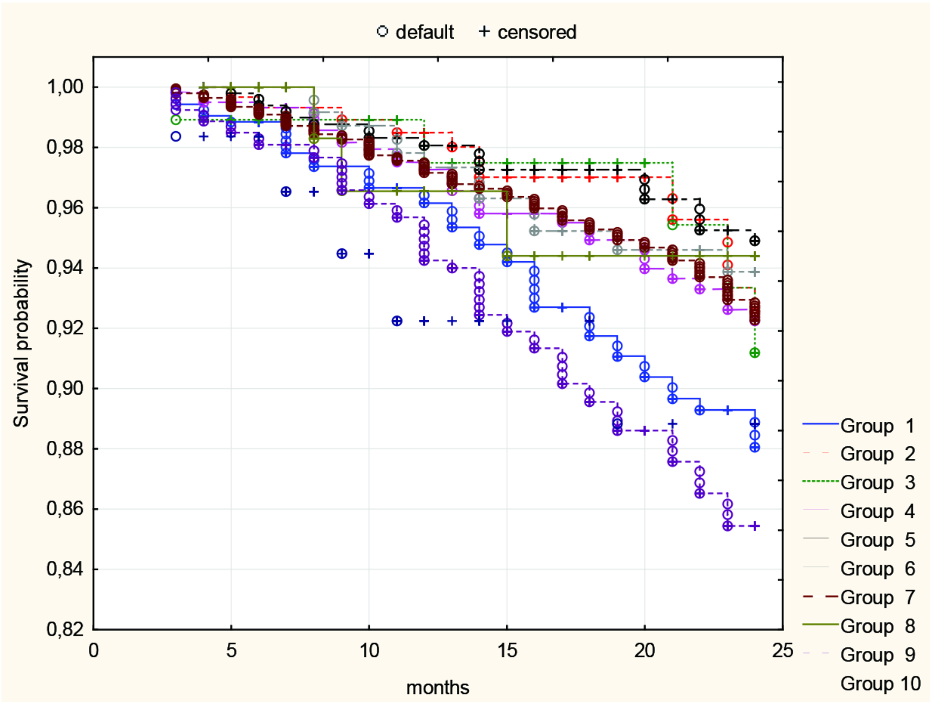

Figure 1. Kaplan-Meier curves for initial groups of X_{12} variable

Source: own elaboration.

On the basis of the shape of the survival curves in Figure 1, groups 1, 9 and 10 were connected into one new group, and the rest of the groups were connected to the second group. Finally, there were two groups with significantly different survival curves (Figure 2). The second group is at high risk of default and the first group is at low risk.

[2] Classic methods of coarse classification of the characteristics are described in [Thomas, Edelman, Crook 2002]. In this article, the author proposed their own method of classification based on K-M estimators.

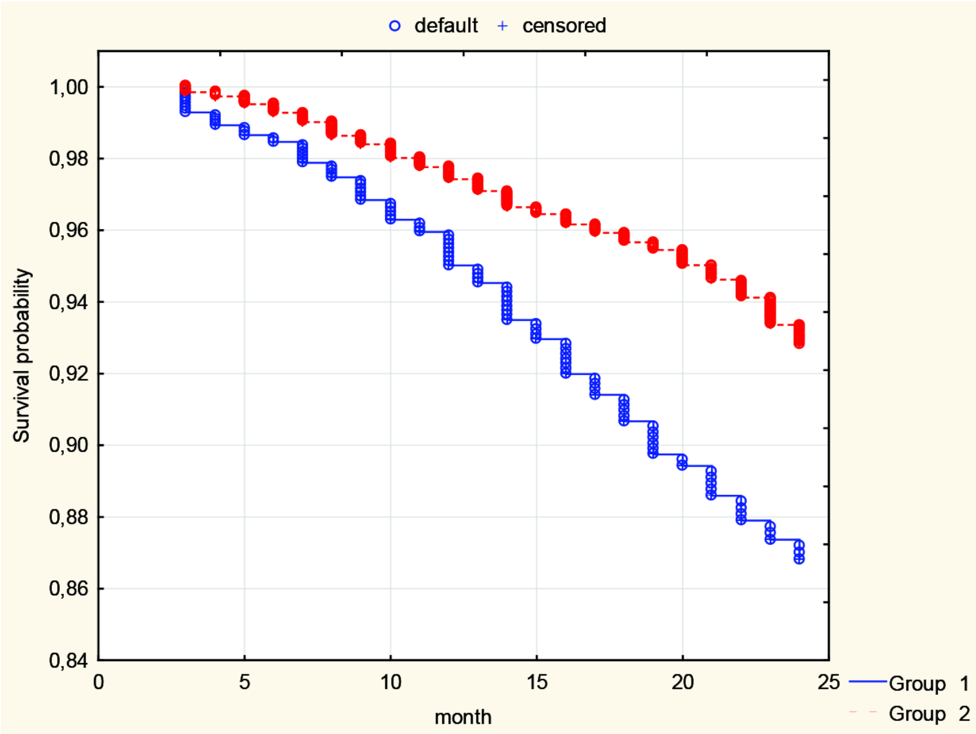

Figure 2. Kaplan-Meier curves for ultimate groups of X_{12} variable

Source: own elaboration.

4. Default and early repayment as failures

In the data set, there were 2274 creditors (45.5%) who repaid all 24 instalments and were active creditors; 297 creditors (5.9%) who defaulted during the first 24 months; and 2429 creditors (48.58%) who repaid the credit earlier during this time. The default and early repayment were treated as failures and analysed separately. In the first part of the analysis, the default was treated as a failure and the early repayments as censored data. In the second part, the early repayments were considered failures and the defaults were considered censored observations.

Time to default was treated as a non-negative continuous variable. Its distribution was described by a survival function. The product limit estimator (Kaplan-Meier estimator) was used to estimate the probability of not defaulting until a particular time (Figure 3).

Figure 3 shows that the risk of default increases in the last months of observed time (increasing steps of the curve). The survival function of the early repayment risk is shown in Figure 4.

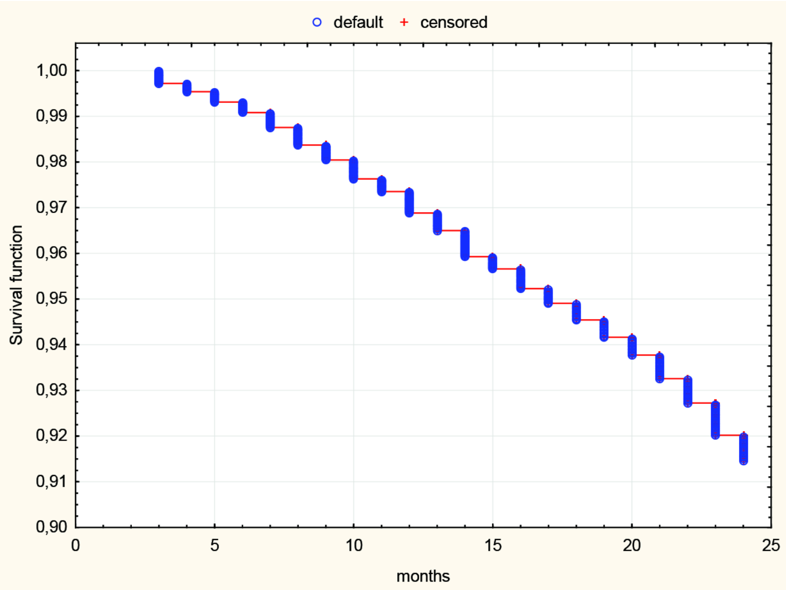

Figure 3. Kaplan-Meier curves for the risk of default

Source: own elaboration.

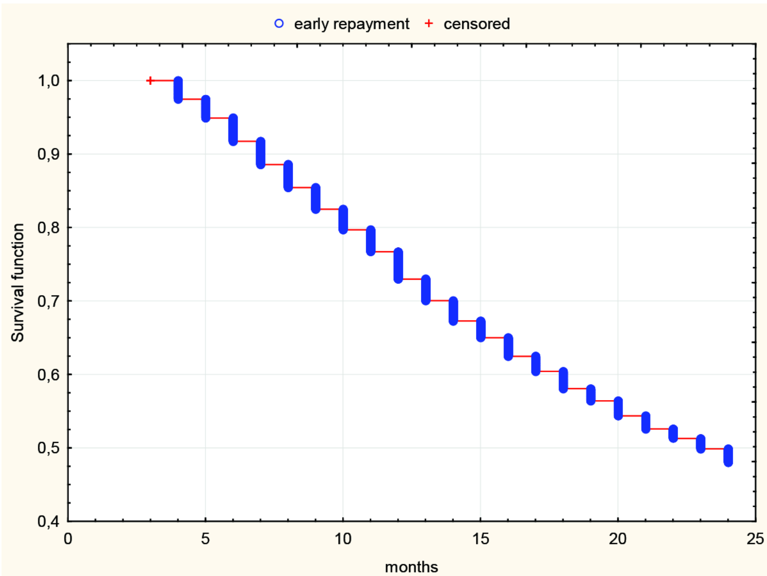

Figure 4. Kaplan-Meier curves for risk of early repayment

Source: own elaboration.

The risk of the repayment seems to be higher in the first months of the credit span.

K-M curves were also drawn for creditors with different characteristics.[3] With the use of significance tests for survival curves, the differences in the influence of these characteristics on the probability of default in the following months of the credit span can be shown (Table 1).

Table 1. The application characteristics and the significance of their predictive power on the risk of default and early repayment

Variable	A number of atributes	Default as a failure				Early repayment as a failure			
		chi-square/z*	*p*-value	A group of the highest risk	A group of the lowest risk	chi-square/z*	*p*-value	A group of the highest risk	A group of the lowest risk
X_1	4	11.897	**0.0078**	**1**	**3**	4.413	0.2202	1, 2, 3	4
X_2	4	19.230	**0.0003**	**1**	**3**	11.509	**0.0093**	1	4
X_3	4	21.057	**0.0001**	**4**	**3**	20.355	**0.0001**	3	4
X_4	6	159.650	**< 0.0001**	**2**	**6**	29.221	**< 0.0001**	4	1
X_5	5	48.700	**< 0.0001**	**1**	**4**	14.325	**0.0063**	2	1
X_6	3	40.460	**< 0.0001**	**2**	**1**	15.157	**0.0005**	3	1
X_7	3	5.870	0.0531	1	2	6.311	**0.0426**	3	1
X_8	2	–1.855	0.0636	1	2	0.685	0.4930	2	1
X_9	4	18.343	**0.0004**	**2**	**4**	13.873	**0.0031**	4	2
X_{10}	4	17.476	**0.0002**	**1**	**3**	7.885	**0.0485**	1	3
X_{11}	2	–2.940	**0.0033**	**1**	**2**	3.477	**0.0005**	2	1
X_{12}	2	–5.560	**< 0.0001**	**1**	**2**	–0.691	0.4897	1	2
X_{13}	2	–4.258	**< 0.0001**	**1**	**2**	0.534	0.5937	2	1
X_{14}	2	2.156	**0.0311**	**1**	**2**	–0.289	0.7792	2	1
X_{15}	2	–1.639	0.1012	1	2	–3.018	**0.0025**	1	2

* Log-rank test (chi-square) for variables with more than two attributes and Gehan test (*z*) for dichotomous variables.

Source: own elaboration.

Twelve of the variables were significant predictors of default (at significance level 0.05). The numbers of attributes with the lowest and highest risks are given in Table 1. The right-hand part of the table includes significance tests for survival curves drawn for early repayment as a failure. The attributes of the variables were the same as in the first part of the analysis, which allowed the comparison of the results. Only 10 of the variables proved to be significant predictors of early repayment. It is worth mentioning that not all of these variables were significant predictors of default. For variables X_{10} and X_{12}, the same groups of attributes indicate, respectively, the highest and the lowest risk of default and early repayment. However, for the

[3] Due to limited space in this article, only selected curves were included. The values of significance tests for all of the characteristics are presented in Table 1.

variables X_3, X_5, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, the attributes that were the highest risk of default proved to be the lowest risk of early repayment. Moreover, as shown in Figure 5, variables should be split in different ways for predicting the risk of default and early repayment.

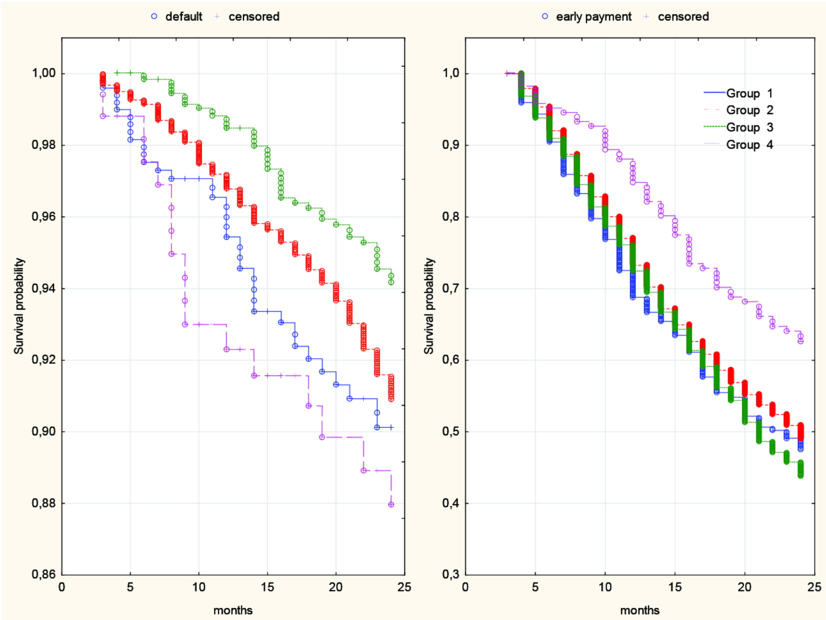

Figure 5. Kaplan-Meier curves for attributes of X_3 variable showing probability of not defaulting or repaying early, respectively

Source: own elaboration.

On the left-hand side of the graph, there are four groups with different curves showing the probability of survival until the time of default. For the risk of early repayment (the right-hand side of the graph), groups 1–3 do not differ significantly and should be aggregated. Finally, for the risk of early repayment, there should be only two different groups (attributes).

5. Conclusions and further research

Survival analysis techniques as approaches to credit scoring have many advantages in comparison with a classic approach. First of all, in the classic approach, units that cannot be interchangeably classified to the group of good or bad creditors, e.g., early repayments, have to be removed from the sample. In the survival methods, the information about such creditors is included in the analysis. Moreover, it is possible to evaluate the probability of early repayments in the following months and to identify the characteristics of the creditors more likely to repay credit earlier. In some credit portfolios, the share of such creditors is high, so it has an influence on the profit of the lender.

In the empirical analysis of loan data, it was shown that the group of early repaying creditors differs, according to the application characteristics, from the group of creditors who defaulted. The same conclusion was drawn by Stepanova and Thomas [2002] using Cox models in the coarse classification of the creditors.

The Kaplan-Meier estimator proved to be an effective tool both in classification of variables and in predicting the probability of not defaulting and repaying early. With the use of the K-M estimator, a lender can evaluate the number of defaults, and with knowledge of characteristics of new clients, he or she can evaluate the number of defaults and early repayments in the consecutive months of a credit span. It is, therefore, an important contribution to profit scoring.

Further analysis of these data should include the usage of regression models such as the Cox semi-parametric model and accelerated model. The specific character of the default modelling approach is that there is a high rate of censored data. In the analysed sample, only 5.9% of observations were completed. The robustness of the proposed techniques for such a structure of data should be analysed. Another important issue is the assumption of non-informative censoring. If this assumption is violated, the K-M estimator and abovementioned regression models can be biased. In such a situation, the techniques devoted to competing risks should be applied. The cumulative incidence of an event can better predict the probability of competing risks than the K-M estimator. These problems will be analysed in the author's further research.

References

Banasik J., Crook J.N., Thomas L.C., 1999, *Not If but When Borrowers Default*, The Journal of the Operational Research Society, vol. 50, no. 12, p. 1185–1190.

Cao R., Vilar M.V., Devia A., 2009, *Modelling Consumer Credit Risk via Survival Analysis*, Statistics and Operations Research Transactions, vol. 33, no. 1, p. 3–30.

Cox D.R., Oakes D., 1984, *Analysis of Survival Data*, Chapman and Hall, London.

Jurkiewicz T., Wycinka E., 2011, *Significance tests of differences between two crossing survival curves for small samples*, [in:] C. Domański, K. Zielińska-Sitkiewicz (eds)], *Methodological Aspects of Multivariate Statistical Analysis, Statistical Models and Applications*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, p. 107–114.

Kaplan E.L., Meier P., 1958, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, vol. 53, no. 282, p. 457–481.

Klein J.P., Moeschberger M.L., 1997, *Survival Analysis. Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Kleinbaum D.G., Klein M., 2005, *Survival Analysis: A Self-learning Text*, Springer, New York.

Mavri M., Angelis V., Ioannou G., Gaki E., Koufodonti I., 2008, *A Two-stage Dynamic Credit Scoring Model, Based on Customers' Profile and Time Horizon*, Journal of Financial Services Marketing, vol. 13, no. 1, p. 17–27.

Marques A.I., Garcia V., Sanchez J.S., 2013, *A Literature Review on the Application of Evolutionary Computing to Credit Scoring*, The Journal of the Operational Research Society, vol. 64, no. 9, pp. 1384–1399.

Narain B., 1992, *Survival Analysis and the Credit Granting Decision*, [in:] L.C. Thomas, J.N. Crook, D.B. Edelman (eds.), *Credit Scoring and Credit Control*, Oxford University Press, Oxford, p. 109–121.
Stepanova M., Thomas L.C., 2001, *PHAB Scores: Proportional Hazards Analysis Behavioural Scores*, The Journal of the Operational Research Society, vol. 59, no. 9, p. 1007–1016.
Stepanova M., Thomas L.C., 2002, *Survival Analysis Methods for Personal Loan Data*, The Journal of the Operational Research Society, vol. 50, no. 2, p. 277–289.
Suciu G., Lemeshow S., Moeschberger M., 2004, *Statistical Tests of the Equality of Survival Curves: Reconsidering the Options*, [in:] N. Balakrishnan, C.R. Rao (eds.), *Advances in Survival Analysis*, Elsevier, Amsterdam, p. 251–262.
Thomas L.C., Edelman D.B., Crook J.N., 2002, *Credit Scoring and Its Applications*, SIAM, Philadelphia.
Thomas L.C., Oliver R.W., Hand D.J., 2005, *A Survey of the Issues in Consumer Credit Modelling Research*, The Journal of the Operational Research Society, vol. 56, no. 9, p. 1006–1015.

ZASTOSOWANIE ANALIZY HISTORII ZDARZEŃ W SKORINGU KREDYTÓW UDZIELANYCH OSOBOM FIZYCZNYM

Streszczenie: Skoring kredytowy jest metodą wykorzystywaną w procesie podejmowania decyzji o udzielenie kredytu. W tradycyjnym podejściu celem modelu skoringowego jest określenie prawdopodobieństwa, że kredytobiorca zaprzestanie spłat kredytu. Dla instytucji kredytowej ważna jest jednak nie tylko informacja, czy kredyt nie zostanie spłacony w całości, ale również to, ile rat zostanie wcześniej zapłaconych. Im później nastąpi przerwanie spłat, tym mniejszą stratę poniesie kredytodawca. Celem artykułu jest zbadanie użyteczności metod analizy przeżycia przy podejmowaniu decyzji o udzielaniu kredytu oraz przy monitorowaniu szkodowości portfela kredytów. Część empiryczną badania przeprowadzono na próbie pięciu tysięcy 60-miesięcznych pożyczek udzielonych przez jedną z polskich instytucji finansowych w ciągu kolejnych sześciu miesięcy. Każda pożyczka była obserwowana przez okres 24 miesięcy od jej udzielenia, chyba że jej spłaty zostały wcześniej przerwane. Jako zaprzestanie spłat przyjęto co najmniej 90-dniowe opóźnienie w spłacie wymagalnej raty. Całkowita wcześniejsza spłata była traktowana jako obserwacja cenzurowana. Czas do zaprzestania spłat kredytu jest nieujemną ciągłą zmienną losową, a jej rozkład został opisany za pomocą funkcji przeżycia. Do oszacowania tej funkcji wykorzystano estymator Kaplana-Meiera. Krzywe estymatorów zostały wyznaczone dla grup pożyczkobiorców wyodrębnionych na podstawie analizowanych charakterystyk. Wykorzystując testy zgodności krzywych przeżycia zidentyfikowano charakterystyki pożyczkobiorców, które istotnie różnicują prawdopodobieństwo spłaty pożyczki.

Słowa kluczowe: modelowanie ryzyka kredytowego, analiza przeżycia (analiza historii zdarzeń), czas do zaprzestania spłat kredytu.