Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
4(349) 2020

DOI: http://dx.doi.org/10.18778/0208-6018.349.05

Dominik Kubacki
University of Łódź, Faculty of Economics and Sociology, Department of Banking
dominik.kubacki@uni.lodz.pl

Robert Kubacki
robertkubacki@o2.pl

Examining Selected Theoretical Distributions of Life Expectancy to Analyse Customer Loyalty Durability The Case of a European Retail Bank

Abstract: One of the key elements related to calculating Customer Lifetime Value is to estimate the duration of a client's relationship with a bank in the future. This can be done using survival analysis. The aim of the article is to examine which of the known distributions used in survival analysis (Weibull, Exponential, Gamma, Log-normal) best describes the churn phenomenon of a bank's clients. If the aim is to estimate the distribution according to which certain units (bank customers) survive and the factors that cause this are not so important, then parametric models can be used. Estimation of survival function parameters is faster than estimating a full Cox model with a properly selected set of explanatory variables. The authors used censored data from a retail bank for the study. The article also draws attention to the most common problems related to preparing data for survival analysis.

Keywords: survival analysis, customer lifetime value, banking, parametric models, Kaplan–Meier estimator

JEL: C34, M31, G21

1. Introduction

Nowadays, there is an increasing need to measure the effectiveness of marketing activities. One of the indicators that synthetically describes a client's value for a company is CLTV (Customer Lifetime Value). This ratio, apart from the revenues and costs incurred so far, also includes future cash flows. It differs from NPV (Net Present Value) in that it also takes into account the probability of customers who will leave (Jeffery, 2010: 167).

$$CLTV = -AC + \sum_{n=1}^{N} \frac{\left(M_n - C_n\right) p^n}{\left(1+r\right)^n}, \tag{1}$$

where:
AC – cost of customer acquisition,
M_n – margin achieved on transactions with clients in period n,
C_n – the cost of marketing and customer service activities in period n,
p – the probability that the client will not cease cooperation within the next year[1],
N – total number of years or other periods.

Estimating the survival probability of the client population is crucial in calculating a client's value over time. Survival analysis methods can be used for this purpose.

Survival time data measure the time to a certain event, such as failure, death, response, relapse, parole, divorce, or the development of a disease. These times are subject to random variations, and like any random variables, they form a distribution (Balicki, 2006: 17).

Let T denote the survival time. The distribution of T can be characterised by three equivalent functions: the survival function, the cumulative survival function, and the cumulative hazard function. The survival function, denoted by $S(t)$, is defined as the probability that an individual will survive longer than t:

$$S\left(t\right) = P\left(T > t\right), 0 < t < \infty. \tag{2}$$

Here, $S(t)$ is a nonincreasing function of time t. The probability of surviving at time zero is 1, while the probability of surviving up to infinity is 0. The cumulative distribution function $F(t)$ is defined as the probability that an individual will fail before t:

$$F\left(t\right) = P\left(T \le t\right), 0 < t < \infty. \tag{3}$$

1 Probability may be proportional to the duration of the relationship with the bank or it may vary depending on the client's seniority.

The hazard function (t) of survival time T gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual will fail within a very short interval, $t + \Delta t$, given that the individual has survived till time T:

$$h(t) = \lim_{\Delta t \to 0}[\frac{P(t \le T < (t + \Delta t) / T \ge t}{\Delta t}] = \frac{f(t)}{S(t)}. \tag{4}$$

The cumulative hazard function is defined as:

$$H(t) = -\log(S(t)) = \int_0^t h(u)\,du. \tag{5}$$

Given any one of them, the other two can be derived:

$$S(t) = 1 - F(t) = \exp(-H(t)). \tag{6}$$

A parametric survival model is one in which survival time, thus the outcome, is assumed to follow a known distribution. By reviewing the literature about modelling survival data, it can be seen that the Exponential, Gamma, Log-normal, and Weibull probability distribution functions are commonly used in survival analysis. The $f(t)$ probability density function, $S(t)$ survival function, and mean lifetime, denoted by the $E(t)$ form of these distribution models, can be summarised below (Erişoğlu, Erişoğlu, Erol, 2011: 545):

Exponential Distribution:

$$f_{exp}(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}},\ t>0,\ \lambda>0, \tag{7}$$

$$S_{exp}(t) = 1 - e^{-\frac{t}{\lambda}}, \tag{8}$$

$$E_{exp}(t) = \lambda. \tag{9}$$

The exponential model is a parametric model. It assumes that the baseline hazard is constant over time. The probability of surviving another time unit does not depend on how long the object has lived so far.

Gamma Distribution:

$$f_{gm}(t) = t^{\alpha_1 - 1} \frac{e^{-t/\beta_1}}{\beta_1^{\alpha_1 \Gamma(\alpha_1)}},\ t \text{ and } \alpha_1, \beta_1 > 0, \tag{10}$$

$$S_{gm}(t) = 1 - \frac{\Gamma_x(\alpha_1)}{\Gamma(\alpha_1)}, \tag{11}$$

$$E_{gm}(t) = \alpha_1 \beta_1, \tag{12}$$

$$\Gamma_x(\alpha_1) = \int_0^x t^{\alpha_1 - 1} e^{-t} dt, \tag{13}$$

$$\Gamma(t) = P(T \le t), 0 < t < \infty. \tag{14}$$

Weibull Distribution:

$$f_{wbi}(t) = \frac{\beta_2}{\alpha_2} \left(\frac{t}{\alpha_2} \right)^{\beta_2 - 1} e^{-\left(\frac{t}{\alpha_2}\right)^{\beta_2}},\ t \text{ and } \alpha_2, \beta_2 > 0, \tag{15}$$

$$S_{wbi}(t) = e^{-\left(\frac{t}{\beta_2}\right)^{\alpha_2}}, \tag{16}$$

$$E_{wbi} = \beta_2 \Gamma\left(1 + \frac{1}{\alpha_2}\right). \tag{17}$$

The Weibull distribution can also be viewed as a generalisation of the exponential distribution. It reduces to the exponential distribution when the shape parameter $\beta_2 = 1$. When the shape parameter is greater than 1, the hazard function increases; otherwise, it decreases.

Log-normal Distribution:

$$f(t) = \frac{\exp\left(-\frac{1}{2}\left(\frac{lnt - \mu}{\sigma}\right)^2\right)}{t\sigma\sqrt{2\pi}}, t > 0, \mu, \sigma > 0, \tag{18}$$

$$S(t) = 1 - \phi \frac{lnt - \mu}{\sigma}, \tag{19}$$

$$E(T) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{20}$$

where Φ is the cumulative distribution function of the standard normal distribution function and is defined by $\phi\left(\frac{lnt-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{lnt-\mu}{\sigma}} \exp\left(-\frac{u^2}{2}\right) du$ (Balicki, 2006: 131).

In order to select the appropriate distribution of the variable that characterises the survival curve, two assessment criteria can be used for the estimated models. The first criterion is the Akaike Information Criterion (Akaike, 1974: 716–723), and the other is the logLik or Maximised Log-likelihood (Jackson, 2016: 1–33).

2. Applications in retail banking

The study was conducted on a random sample of 100,000 retail clients in a bank located in Europe. The characteristics of the dataset are as follows:

1) individual customers aged 18–75,
2) right-censored data (date of last observation: 1.09.2018),
3) without clients with a planned termination agreement,
4) returning customers are treated as a continuous relationship if the interval does not exceed 12 months,
5) with a relationship with the bank longer than one month,
6) primary owners of the product,
7) response variable – duration in months of the customer's relationship with the bank between opening the first product and closing the last product.

The calculations and graphs were made using R and R Studio software. The packages used for the calculations included survival, flexsurv, and e1071[2]. One of the most important steps associated with preparinga survival analysis is properly preparing the data.

The first challenge is to determine what is considered to be the beginning of the relationship with the customer, whether it is the date of opening the first product or the date of establishing the general customer agreement. If the client had a relationship with the bank that handles him/her from the beginning, then these two dates should be the same. If the client was migrated to the bank as a result of a merger or takeover, then the date of establishing the customer file is usually the date of the operational merger of the two banks.

2 The Comprehensive R Archive Network, https://cran.r-project.org (accessed: 23.03.2019).

The case could get even more complicated if the customer had been served by both banks. For this study, the principle was adopted that we take into account the date of opening the first product, irrespective of the bank in which the relationship was initiated. Another solution would be to prepare separate survival curves for clients coming from the home bank, migrated clients (but new ones for the bank), and shared clients.

The second problem in preparing data for a survival analysis may be the client's return to the bank and the related opening of a new product when the last product under the previous relationship was closed. An estimation for this particular survival curve can be made. In this analysis, a business assumption was made that it was an existing relationship if the gap between the closing of the last product and the opening of a new product after returning to the bank does not exceed 12 months. This assumption can be accepted if customers use products that are characterised by a short time period, and they regularly buy a product with similar parameters after repaying the products. This may apply to banks that focus both on short-term deposits and cash loans. It is necessary to simplify the modelling of the phenomenon because such gaps may result from system limitations or the duration of setting up the product, not because of a customer actually leaving the bank.

The third problem that occurs is the large skewness of the data we work on. One of the ways to deal with this is to transform the variables, which will bring the distribution of the variable being analysed to a more symmetrical distribution. One of the most commonly used transformations of variables is the logarithmic transformation. When a log transformation is performed, adding a constant solves the problem of the legitimisation of zero. In the case of survival analysis, this condition is always met (Jajuga, Walesiak, 1999: 105–112).

3. The results of the empirical analyses conducted

The results of the non-parametric estimation of the survival function using the Kaplan–Meier (Kaplan, Meier, 1958: 457–481) estimator are presented in Figure 1. The curve is relatively regular from the 25th month[3]. In the initial period, i.e., around the 10th and the 20th month, there is a gradual decline in the survival function.

Using R software, theauthors estimated parametric models for the survival banking dataset. Four distributions were compared, and the best estimates for each distribution are presented in Table 1.

3 Months – means the time for which clients have maintained relations with the bank.

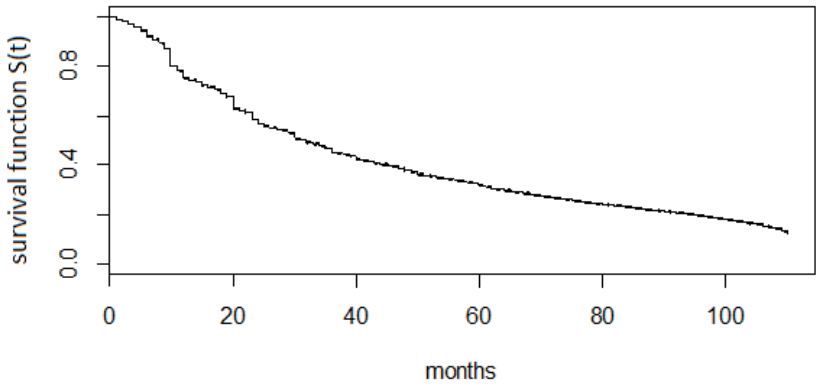

Figure 1. Survival function for months with the Kaplan–Meier estimator
Source: banking survival dataset

In Figure 2, the authors present the cumulative events for a number of months.

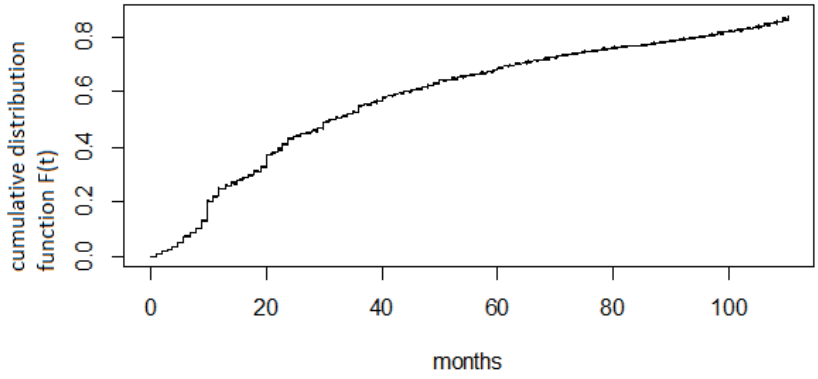

Figure 2. Cumulative events for months
Source: banking survival dataset

Table 1. Estimated parameters for the parametric survival models

Model	Parameter Estimation
Exponential	$\hat{\lambda} = 52.0833$
Gamma	$\hat{\alpha}_1 = 1.1278 \quad \hat{\beta}_1 = 44.8491$
Weibull	$\hat{\alpha}_2 = 51.9572 \quad \hat{\beta}_2 = 1.0417$
Log-normal	$\hat{\mu} = 3.488 \quad \hat{\sigma} = 1.2027$

Source: own calculation

In corresponding Table 2, the authors present the values of logLik and AIC to choose the best distribution out of the four competitors. The lowest AIC value is calculated for a log-normal distribution.

Researchers should not always focus only on the lowest AIC or logLik values. Sometimes it is better to choose a distribution with fewer parameters. This makes it easier to explain the phenomenon to business owners, as not all of them have deep statistical knowledge to interpret empirical results. Statistical significance tests can be used to check the hypothesis that the observed values do not differ from theoretical distributions.

Table 2. Values of logLik and AIC which correspond to the best-fitted distributions (variable months)

Distribution	LogLik	AIC
Exponential	–329 984	659 971
Weibull	–329 903	659 810
Gamma	–329 664	659 333
Log-normal	–326 655	653 314

Source: own calculation

Finally, in Figure 3, the authors present how a log-normal distribution fits the observed dataset. The curve fits the observed dataset. Only in regions mentioned at the beginning of the article (10th and 20th month), does the red line not fit the data.

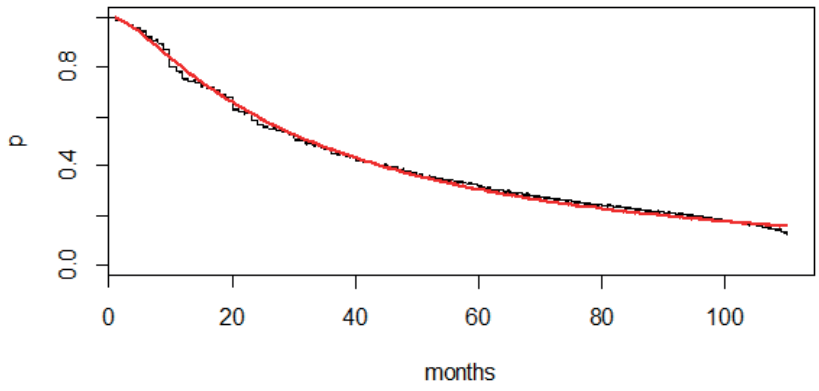

Figure 3. Log-normal survival curve for the banking dataset

Source: banking survival dataset

In Figure 4, the probability plots for the predicted and theoretical log-normal distribution are presented.

To have a good comparison between available solutions, it is sometimes worth checking other possibilities. In this study, the authors also checked how distributions for log-transformed variable months performed.

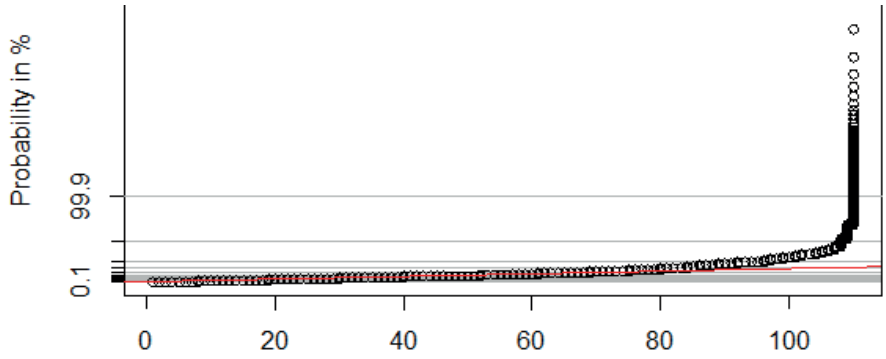

Figure 4. Probability plots for a Log-normal distribution

Source: banking survival dataset

The results and estimates obtained for these models are presented in Table 3.

Table 3. The estimated parameters for the banking survival dataset with log-transformed variable months

Model	Parameter Estimation
Exponential	$\hat{\lambda} = 4.74$
Gamma	$\hat{\alpha}_1 = 7.5347 \quad \hat{\beta}_1 = 0.4812$
Weibull	$\hat{\alpha}_2 = 3.9248 \quad \hat{\beta}_2 = 0.2865$
Log-normal	$\hat{\mu} = 1.2353 \quad \hat{\sigma} = 0.4074$

Source: own calculation

The corresponding values of logLik and AIC are presented in Table 4. The Weibull model has the lowest LogLik value. The log transformation of the data changed the winner to the best-fitted distribution.

Table 4. The values of the log-likelihood function and AIC that correspond to the best-fitted distributions (with LOG_MONTHS)

Model	LogLik	AIC
Exponential	–170 300	340 601
Weibull	–125 582	251 169
Gamma	–126 285	252 575
Log-normal	–128 409	256 822

Source: own calculation

Figure 5 presents the Weibull survival curve for the banking dataset, which fits the observed data better than the other curves using thelogLik and AIC criteria. For regions located at 2.5 and 3, there is an abrupt lowering of the survival curve.

From a business perspective, it is very interesting to investigate what type of clients end their relationship with the bank. It might be a starting point for a deeper analysis of what factors cause a customer to leave abank.

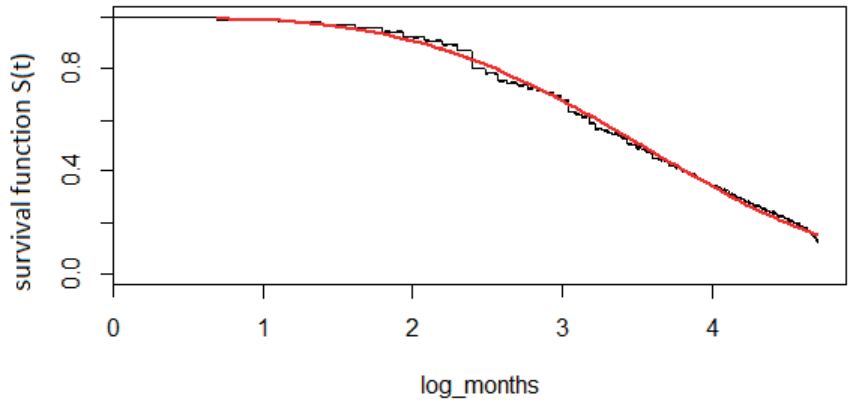

Figure 5. The Weibull survival curve for the banking dataset with a log-transformed variable
Source: banking survival dataset

Figure 6 presents the probability plots for the predicted and theoretical Weibull distribution.

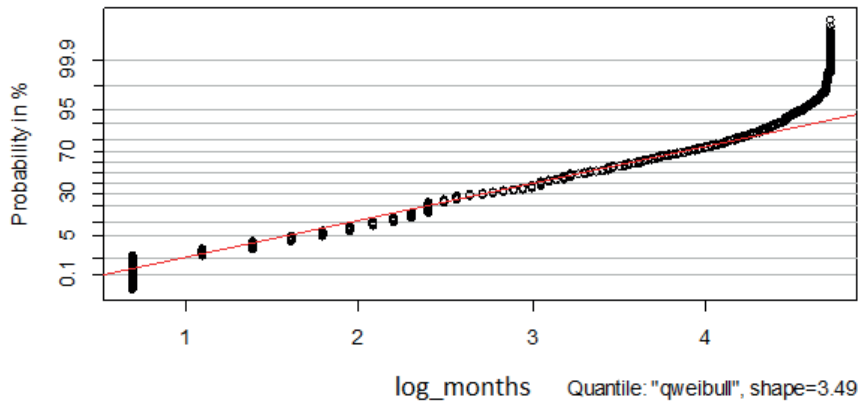

Figure 6. Probability plot for a Weibull distribution
Source: banking survival dataset

4. Conclusions

In this paper, the authors compared how observed survival data fit different theoretical distributions, such as Exponential, Weibull, Gamma, and Log-normal. The estimation of the parameters and the calculation of statistics, such as AIC and logLik, have shown that the Log-normal and Weibull distributions are best for this particular sample of clients. The results obtained in the study confirm that parametric models are valuable sources of information on the duration of customer relationships with the bank, and the model parameters themselves provide valuable knowledge of whether increased extinction occurs at the beginning of the relationship or is proportional to the examined period. The estimated parameters of survival models can be used to compare subgroups of customers that may arise from bank mergers and acquisitions. Knowing which group of customers has a steeper survival curve enables better planning of retention activities. Estimating the parameters of the survival function is simpler than building a Cox model. Gathering and preparing explanatory variables requires additional time, and not all variables that could be used in the model are available in corporate databases.

However, the analyses presented in this paper are not sufficient to extend the results to the entire banking sector. Further research is needed in this field. It would be advisable to prepare and check mixed models (the sum of two or three distributions), especially in those areas where the observed data do not perfectly fit theoretical distributions.

References

Akaike H. (1974), *A New Look at the Statistical Model Identification*, "IEEE. Transactions on Automatic Control", vol. Ac–19, no. 6, pp. 716–723.

Balicki A. (2006), *Analiza przeżycia i tablice wymieralności*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Erişoğlu Ü., Erişoğlu M., Erol H. (2011), *A Mixture Model of Two Different Distributions Approach to the Analysis of Heterogeneous Survival Data*, "World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering", vol. 5, no. 6, pp. 544–548.

Jackson C. (2016), *flexsurv: A Platform for Parametric Survival Modeling in R*, "Journal of Statistical Software", vol. 70, no. 8, pp. 1–33.

Jajuga K., Walesiak M. (1999), *Standaridisation of data set under different measurement scales*, [in]: *Classification and Information Processing at the Turn of the Millennium: Proceedings of the 23rd Annual Conference of the Gesellschaftfür Klassifikatione.V.*, University of Bielefeld, Bielefeld, pp. 105–112.

Jeffery M. (2010), *Data-Driven Marketing. The 15 Metrics Everyone in Marketing Should Know*, John Wiley & Sons, Hoboken.

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

The Comprehensive R Archive Network, https://cran.r-project.org (accessed: 23.03.2019).

Ocena wybranych rozkładów teoretycznych trwania życia do analizy lojalności klientów na przykładzie europejskiego banku detalicznego

Streszczenie: Jednym z kluczowych elementów związanych z wyliczaniem wartości klienta w czasie (*Customer Life Time Value*) jest oszacowanie długości trwania relacji klienta z bankiem w przyszłości. Można ją oszacować z wykorzystaniem metod analizy przeżycia. Celem artykułu jest sprawdzenie, który ze znanych rozkładów wykorzystywanych w analizie przeżycia (Weibulla, wykładniczy, gamma, logarytmicznie normalny) najlepiej opisuje zjawisko odejść klientów z banku. Jeśli celem jest oszacowanie rozkładu, według którego „przeżywają" określone jednostki (klienci banku), a czynniki, które to powodują, nie są aż tak istotne, to modele parametryczne mogą być wykorzystane. Oszacowanie parametrów funkcji przeżycia jest szybsze niż oszacowanie pełnego modelu Coxa z odpowiednio dobranym zestawem zmiennych objaśniających. Do badania wykorzystano dane cenzurowane banku detalicznego. W artykule zwrócono uwagę na najczęstsze problemy związane z przygotowaniem danych do analizy przeżycia.

Słowa kluczowe: analiza przeżycia, wartość życiowa klienta, bankowość, modele parametryczne, estymator Kaplana–Meiera

JEL: C34, M31, G21

Received: 2016-01-01; verified: 2017-01-01. Accepted: 2017-11-24