Survival analysis in credit scoring

Oleg Naidovich
Belarusian State University
Minsk, Republic of Belarus
o.naidovich@gmail.com

Alexander Nedzved
Belarusian State University
Minsk, Republic of Belarus
nedzveda@gmail.com

Shiping Ye
Zhejiang Shuren University
Hangzhou, China
email: zjsruysp@163.com
ORCID: 0000-0002-9771-7168

***Abstract* — In the domain of credit risk assessment, innovative approaches have emerged to address the challenge of predicting loan default probabilities. This article explores Survival Analysis, a statistical method capable of predicting the timing of loan repayments and distinguishing between completed repayments and unpaid loans, treating them as censored events. By integrating Survival Analysis, financial institutions can enhance their ability to forecast repayment timelines, minimize losses from non-performing loans, optimize cash flow management, refine credit collection strategies. The primary goal of this article is to investigate the utility of survival models in estimating Probability of Default (PD) and developing credit scorecards.**

Keywords — Survival Analysis, credit risk modeling, probability of default, Cox proportional hazards model, Logistic Regression

I. Introduction

Credit risk assessment is vital in the financial sector, helping determine the likelihood of borrowers failing to meet their loan obligations. Traditional credit scoring models, such as logistic regression and decision trees, have been the standard for predicting the Probability of Default (PD) within specified timeframes [1]. However, these models encounter difficulties when handling censored and truncated data, a common issue due to a large number of borrowers who successfully repay their loans.

In addition to that, quantitative assessment of the PD value is one of the components of credit risk (PD, LGD, EAD, EL). The amount of Expected Losses is calculated using the following formula (1):

$$\text{EL} = \text{PD x LGD x EAD} \tag{1}$$

where EL - expected loss, PD - probability of default based on client characteristics, default statistics and market information, LGD - loss given default, EAD - exposure of default, explains what impact a customer default would have.

A. *Common Principles in Credit Scoring:*

All credit scoring models share fundamental principles. They start with categorizing a sample of previous customers based on their historical repayment performance as either good or bad borrowers. Subsequently, these models connect the characteristics (factors) of these borrowers to their default status. Various techniques are available for constructing these systems, including discriminant analysis, expert systems, and logistic regression [2]. Presently, logistic regression stands as the industry standard.

B. *Stages in Assessing Default Probability*

The assessment of default probability consists of several stages (Fig.1):

1. **Data Collection and Preparation**: This phase involves gathering and preparing data related to the characteristics and parameters of credit requests, encompassing credit risk factors, and calculating actual default values. As the main part of PD creation pipeline, it consists of 2 parts: univariate and multivariate analysis. The main processes that take place in these two parts are related to the transformation of features through WOE (weight of evident) transformation (2) and the removal of features to solve the multi-correlation problem.

$$WOE_i = \log\left(\frac{\dfrac{N_G(i)}{N_G}}{\dfrac{N_B(i)}{N_B}}\right) \tag{2}$$

where $N_G(i)$ and N_G – number of non-default observations in group i and in the entire sample, respectively, $N_B(i)$ and N_B – the number of default observations in group i and throughout the sample, respectively.

2. **Model Development**: Models are created to describe the relationship between actual default values and risk factors. The model is built in such a way that at the output we can easily interpret each factor and how much it influenced the formation of PD score.

3. **Model Calibration**: The developed models are calibrated using current data to ensure their accuracy. Calibration is an integral part of creating a PD model, since in the world of finance there are different business cycles, such as growth, crisis, stagnation.

4. **Application and Monitoring:** Since the task of assessing default is the most important task in a bank, then validation, accuracy assessment and constant updating of PD models is an essential part of the entire development and implementation cycle. Current models are regularly applied to observations within identified risk segments, typically on a monthly or quarterly basis. This allows for the reevaluation of default probability values based on up-to-date risk factor information and calibration.

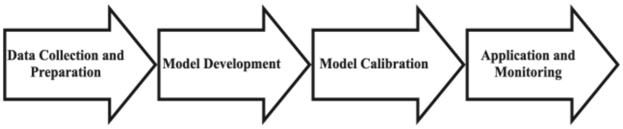

Fig. 1. Stages in Assessing Default Probability

II. Classic PD Models

A. *Contemporary Methods for Probability of Default (PD) Estimation*

Contemporary techniques for estimating the Probability of Default (PD) predominantly rely on the robust foundation of **logistic regression** (3). This approach involves classifying past customers into "positive" or "negative" categories based on their repayment history within a specified timeframe. Logistic regression serves a dual purpose in rigorously assessing credit risk and identifying the key variables that influence credit risk prediction.

$$PD(Y=1\,|\,X_1,...X_n)=\frac{1}{1+\exp\left(-(\beta_0+\beta_1X_1+...+\beta_nX_n)\right)} \quad (3)$$

where Y is a sign of default, $Y=1$ is a default event, $X_1, X_2, \ldots, X_n$ is a set of independent explanatory WOE-factors, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are logistic regression coefficients.

B. *Categorization of Portfolio Observations*

In this meticulous process, portfolio observations undergo thorough evaluation, leading to clear-cut categorization as "positive" or "negative." The primary criteria for this categorization primarily revolve around payment behavior, such as instances of payment delays exceeding 90 days. Observations displaying timely payments are categorized as "positive," while those that don't fit neatly into either category are labeled as "undefined" and thoughtfully excluded from the modeling process.

C. *Role of Logistic Regression in PD Estimation*

Logistic regression [3] distinguishes itself from traditional regression models by adeptly handling binary-dependent variables. It calculates the odds of a specific event occurring, enabling the approximation of the Probability of Default (PD) for a given facility. This essential measure of default odds adheres to a precise equation. The complex data preparation process involved in PD estimation through logistic regression ultimately determines whether a facility is categorized as "positive" or "negative."

D. *Limitations of Logistic Regression and Introduction of Survival Analysis*

Additionally, logistic regression inherently restricts PD estimation to a fixed one-year timeframe, leading to a resource-intensive data preparation process. To address these limitations, an emerging solution takes the form of survival analysis. This method excels in handling incomplete observations, classifying them as censored data instances where the event of interest remains unresolved within the study period. Survival analysis, with its primary focus on estimating the survival distribution, provides the critical advantage of assessing default risk across various future timeframes, thus elevating credit risk assessment.

III. Survival Analysis

A. *Introduction to Survival Analysis in Credit Scoring and key features*

Survival Analysis [4], a statistical technique widely used in fields like medicine and engineering, has found its way into credit risk assessment and scoring. This approach focuses on modeling event timing and is particularly well-suited for analyzing credit risk. Its distinctive strength lies in its ability to handle censored and truncated data, which is common in credit risk assessment due to borrowers not defaulting within the study period.

B. *Advantages of Survival Analysis in Credit Risk Assessment*

In the context of credit risk assessment, the crucial event of interest is borrower default. Survival analysis provides significant advantages over conventional credit scoring models by effectively incorporating censored and truncated data in the development sample. Conventional logistic regression models typically exclude such data. One prevalent form of censoring is right censoring, signifying that the event is not observed during the study period. In the context of credit risk, this primarily relates to non-defaulting customers, which makes up a substantial portion of the data.

C. *Math under Survival Analysis*

Let's view T as the time it takes for a facility to encounter a default event. The distribution function quantifies the likelihood that the time to the event T is less than or equal to a specific time t, and it is represented as follows (4):

$$F(t)=P(T\le t). \quad (4)$$

From this information, the survival function can be obtained, representing the probability that the time to the event T exceeds a specified time t. This function is expressed in this manner (5):

$$S(t)=P(T\ge t)=1-F(t). \quad (5)$$

The second method is through the density function. This function calculates the probability that the failure time precisely matches the time t, considering all potential times. It is represented as (6):

$$f(t)=\lim_{\Delta t\to 0}\frac{P(t\le T\le \Delta t+t)}{\Delta t}. \quad (6)$$

The hazard function quantifies the probability that, if a facility remains operational until time t, it will undergo the event in the immediate subsequent moment. It is defined as follows (7):

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t\le T\le \Delta t+t\,|\,T>t)}{\Delta t}=\frac{f(t)}{S(t)}. \quad (7)$$

D. *Censoring*

Survival analysis often grapples with the challenge of censored data, also known as the missing data problem, where critical information such as end dates is absent. An ideal dataset would comprise both the start and end dates of all portfolio facilities to determine their lifetimes. However, when the end date is missing, it is categorized as right-censored data. When performing 1-year PD estimations with logistic regression, facilities need to endure at least a year. Estimating the survival function across the entire data period becomes tricky if observations exiting the portfolio are excluded, significantly reducing the size of dataset. Therefore, it is imperative for the model to encompass all observations, including censored ones not within the portfolio for the entire data duration. A crucial assumption in dealing with censored data is non-informative censoring, positing that censored facilities face the same risk of subsequent failure as their uncensored counterparts (fig.2).

E. *Non-parametric models*

To establish the distribution function for data, one can employ fundamental **Kaplan-Meier** methods (8). The KM estimator [5] calculates the median survival distribution function. The key benefit of these estimators lies in their

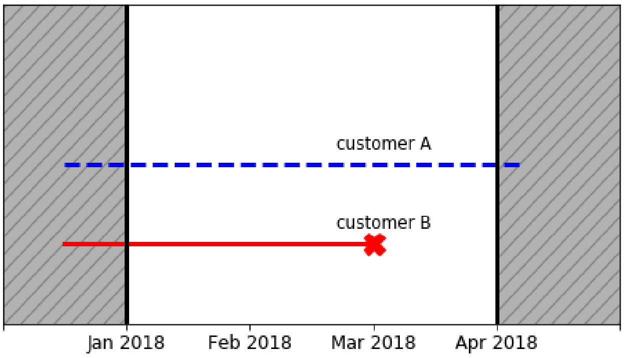

Fig. 2. Censoring missing data

ability to consider censored data, effectively representing the limit of the life-table estimator when intervals are reduced to a point where, in each interval, at most one unique observation is recorded.

$$S_{KM}(t) = \prod_{t_i < t} \left(1 - \frac{d_i(x)}{r_i(x)}\right), \tag{8}$$

where $S_{KM}(t)$ is a survival function of Kaplan-Meier method, d_i is the number of individuals experiencing an event at t_i, r_i is the number of individuals at risk within $[t_{i-1}, t_i)$ - those who have not been censored or experienced an event.

F. Fully parametric models

In case the hazard function or the Survival function are known to follow or closely approximate a known distribution, it is better to use Parametric models. Parametric models are better suited for forecasting and will return smooth functions of $h(t)$ or $S(t)$. The most common parametric models are: Exponential, Weibull, Log-Logistic, Lognormal (fig.3).

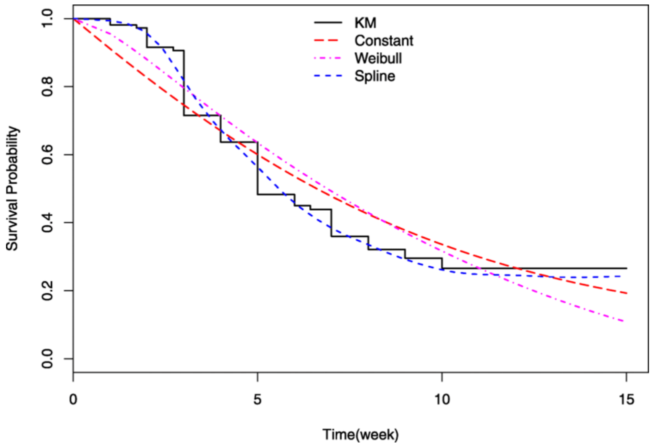

Fig. 3. Examples of estimating Survival Probability

G. Semi-parametric models

The **Cox Proportional Hazard** model [6], often referred to as the Cox PH model, is a leading semi-parametric model selected for several compelling reasons. First and foremost, it sets itself apart by not requiring any baseline assumptions. This attribute endows it with robustness, adaptability, and makes it a wise choice in diverse situations. Furthermore, it demonstrates versatility in handling both discrete and continuous event time measurements while also providing the capacity to integrate time-dependent covariates. This functionality enables adjustments for changes in covariate values during the observation periods. As a semi-parametric model, it primarily focuses on modeling the hazard function, denoted as $h(t|x_i)$. It does so by assuming that its time component, $\lambda_0(t)$, and its feature component, $\eta(x_i)$, are proportional like that (9):

$$h(t \mid x_i) = h_0(t)\eta(x_i), \tag{9}$$

where $h_0(t)$ is the baseline function, which is usually not specified, $\eta(x_i)$ is a risk function usually expressed via a linear representation (10):

$$\eta(x_i) = \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n). \tag{10}$$

IV. Experiment

A. Data

The "Default of Credit Card Clients Dataset," featured in a Kaggle competition, serves as a valuable resource for the article and investigation into credit scoring. This dataset provides rich information on borrowers' financial behaviors and demographics, contributing to the exploration of credit risk assessment.

B. Metrics

In the realm of evaluating two or more scoring algorithms, the **ROC curves** [7] serve as the established benchmark. To construct the ROC curve, scores for all facilities are initially arranged in ascending order from poor to excellent. The model's effectiveness can be numerically assessed through the area under the curve (AUC) value. A higher AUC value, approaching 1, signifies a superior quality of the model's estimation.

The **power statistic**, which shares a close connection with the Gini coefficient, serves as a tool for both visualizing and quantifying the predictive effectiveness of individual factors in a given context. The underlying concept is that the poorest factor values should align with the least favorable observations or outcomes.

$$\textit{Power Stats} = 2(\textit{Area under ROC curve} - 0.5) \tag{11}$$

C. Results

The effectiveness of the model is assessed across one, two, and three-year intervals. This approach aligns with survival analysis, which is designed to forecast event occurrence over the entire dataset duration, rather than within fixed time segments. When it comes to predicting the likelihood of default within a single year, the survival model demonstrates marginal or no advantages over the logistic model. However, as the prediction horizon extends beyond one year, survival analysis consistently outperforms the logistic model, delivering superior results.

V. Conclusion

The article has delved into the practical implementation of survival models for the estimation of Probability of Default (PD) and the creation of credit scorecards. Through the utilization of Survival Analysis, financial institutions can significantly augment their capacity to make well-informed decisions. This includes optimizing cash flow management, mitigating losses stemming from non-performing loans, fine-

TABLE I. RESULTS OF DEFAULT ON CREDIC CARDS DATA.

Metrics \ Methods	Cox PH	Logistic
1 year		
ROC-AUC	**0.83**	0.82
Power Stats	57.43%	**57.54%**
2 years		
ROC-AUC	**0.80**	0.75
Power Stats	**56.3%**	54.47%
3 years		
ROC-AUC	**0.76**	0.69
Power Stats	**54.91%**	53.78%

tuning credit collection strategies, and pinpointing high-value customers across an array of financial products. This contribution seeks to propel credit risk assessment methodologies forward and elevate the standards of risk management within the financial sector. In doing so, it sets the stage for more robust and resilient financial institutions in an ever-evolving economic landscape.

REFERENCES

[1] Stepanova, Maria & Thomas, Lyn. (2002). Survival Analysis Methods for Personal Loan Data. Operations Research. 50. 277-289. 10.1287/opre.50.2.277.426.

[2] Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008 Dec 16;3:17. doi: 10.1186/1751-0473-3-17. PMID: 19087314; PMCID: PMC2633005.

[3] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[4] Lawless, J. F. "Statistical Methods in Reliability." Technometrics, vol. 25, no. 4, 1983, pp. 305–16. JSTOR, https://doi.org/10.2307/1267846. Accessed 14 Sept. 2023.

[5] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53, 457-481. http://dx.doi.org/10.1080/01621459.1958.10501452

[6] Harrell, F.E. (2001). Cox Proportional Hazards Regression Model. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3462-1_19

[7] Andrew P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, Volume 30, Issue 7, 1997, Pages 1145-1159, ISSN 0031-3203, https://doi.org/10.1016/S0031-3203(96)00142-2.