Establishment of Business Loan Default Prediction Model by Integrating Survival Analysis with Logistic Regression

Yung-Chia Chang [a], Kuei-Hu Chang [b, c, *] and Yi-Xin Lin [a]

[a] Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu 300, Taiwan

[b] Department of Management Sciences, R.O.C. Military Academy, Kaohsiung 830, Taiwan

[c] Institute of Innovation and Circular Economy, Asia University, Taichung 413, Taiwan

* Corresponding author.
E-mail address: evenken2002@yahoo.com.tw (Kuei-Hu Chang)
School Address: No.1, Weiwu Rd., Fengshan City, Kaohsiung County 83059, Taiwan
Tel:+886-7-7403060

Abstract

An insufficient amount of capital conservation buffer would cause a financial institution to be unable to withstand fluctuations in the economic cycle; while an excessive amount would reduce the financial institution's available funds, which would lead to a loss of the capital available for investment. In order to address this issue in an effective manner, the business loan default prediction model is established in this study by integrating survival analysis with logistic regression. In the section of case validation, the reliability of the proposed approach is validated with the information of businesses that have been granted loans by financial institutions in Taiwan, and the proposed approach was also compared with the Cox proportional hazards model approach, which is frequently applied by financial institutions. The empirical results demonstrate that the approach proposed in this study could predict a business loan default state closer to the actual default trend, and provide prediction results superior to that of the Cox proportional hazards model, thus, providing financial institutions with effective and reliable information for reference, which will allow them to prepare an appropriate amount of capital conservation buffer, and improve the capital flexibility of the financial institution.

Keywords: Business loan default prediction; Small and medium-sized enterprises; Basel capital accord; Survival analysis; Logistic regression

1. Introduction

The correctness of business loan default prediction results is a key issue that will affect the stability and security of the financial system [1]. Therefore, the issue of returns to investors and the level of default are crucial for ensuring the continuity of market borrowing [2]. All financial institutions set their respective capital demands against the default risk mainly in accordance with the Basel capital accord, as developed by the Basel committee on banking supervision, the Bank for International Settlement. Basel III is mainly for highlighting a reduction in the extent of the risk jointly borne by financial institutions, in addition to improving the capability of financial institutions to adapt to changes in financial environments. The new Basel III aims at improving the capital requirements for financial institutions, and its capital adequacy ratio (CAR) is still maintained at 8%, as set out in Basel II, but with the addition of a 2.5% capital conservation buffer [3].

The key to assessing credit trends lies in assessing and predicting potential default trends and the total number of default cases among outstanding business loan cases of financial institutions in the future, which will allow financial institutions to prepare an appropriate amount of capital conservation buffer, and thus, mitigate the impact on it in case of default in the future. Outstanding business loan default predictions are generally determined by means of data mining, where all historical data are used as training data to establish the prediction model, and then, prediction is conducted regarding the research objects. For example, Noh et al. [4] applied the survival analysis approach to discuss credit card default risk; the survival analysis approach was applied because such approach could predict the default "time" and "probability", in order to facilitate whether or not relevant financial institutions will approve the credit application of these consumers and grant credit to them. Gepp and Kumar [5] applied a combination of the Cox proportional hazards model and a non-parametric classification and regression trees (CART) in financial prediction, in an attempt to predict whether businesses will be subject to bankruptcy due to financial distress in the future. The research results demonstrate that the combination and application of a decision tree and the Cox proportional hazards model could improve prediction accuracy. Gupta et al. [6] respectively applied logistic regression in the discrete time hazards model and the continuous time Cox proportional hazards model to predict the bankruptcy and financial crisis of medium-sized enterprises in America. The results demonstrated that logistic regression would produce a prediction result that superior to the continuous time Cox proportional hazards model. Many authors applied the Cox proportional hazards model exploring issues related to financial institutions [7-12].

Many researches used logistic regression analysis to solve the related issues of categorical data. For example, Yap et al. [13] researched the effects of applications with different data mining algorithms in the credit rating model for the purpose of improving misinterpretation, as caused by the subjective judgment of financial institution personnel in the process of credit rating. Liu et al. [14] stated that logistic regression is an effective classifier for text analysis. In this study, a new classification method is proposed by combining logistic regression with decision-theoretic rough set (DTRS). Research has demonstrated that logistic regression can be applied to validate the rationality and effectiveness of an approach. Croux et al. [2] used logistic regression analysis to survey the default determinants of Fintech loans through the LendingClub consumer platform during the period of 2007-2018. They used the contractual loan characteristics, borrower characteristics, and macroeconomic variables as independent variables, and used the lasso selection methods to reduce the number of independent variables from 100 to 58 variables. The simulation results indicated that Fintech lenders should use alternative data without violating fair lending rules from traditional subprime loan pools. At present, many scholars have implemented research on data mining by applying logistic regression (such as [2, 15-25]).

In this study, from the perspective of a financial institution, and in consideration of the different time characteristics of different outstanding business loan cases, the data meeting the time characteristics of each outstanding business loan case to be predicted were identified through data screening in two phases, thus, ensuring that all modeling data can practically meet the time characteristics of each outstanding business loan case to be predicted. Finally, logistic regression was applied to establish an exclusive complete prediction model for each outstanding business loan case, which has the purpose of providing financial institutions with reliable information for reference, preparing an appropriate amount of capital conservation buffer, and improving the capital flexibility of financial institutions.

Subsequent sections of this study are set out, as follows: Section 2 offers the Literature Review, which presents a brief introduction of survival analysis and logistic regression; Section 3 presents the Research Method, which presents an introduction of the overall process and approach for establishing the prediction model in this study; Section 4 is Empirical Analysis, which presents case validation, as based on the loan data provided by a financial institution in Taiwan; Section 5 offers the Conclusion and Suggestions, which presents the conclusions of this study and provides an outlook for prospective researches.

2. Literature Review

2.1 Survival analysis

Survival analysis is mainly for implementing in-depth discussions on the correlation between the survival time of a sample group and each variable. Therefore, general survival analysis relates to research regarding the probability of the occurrence of a specified event within a certain period after each individual sample has gone through a certain event. In other words, “survival analysis relates to research based on a group’s life time trend”.

We assume the life time period of a sample as *T*. Generally, such numerical value commences from the start of an event till the time point of the occurrence of a specified event, and *T* was considered as a random variable; for example, from the time when a patient receives treatment till the time of death of the patient. The specified event mentioned herein will be subject to different interpretations according to cases in different fields; for example, in medicine, it is referred to as the survival rate; in manufacturing engineering, it is referred to as reliability; in a financial institution, it is referred to as default, etc.

Regarding survival analysis, the important survival function is denoted with *S*(*t*), as shown in Eq. (1):

$$S(t)= \ P(T> \qquad (1)$$

where, *t* denotes a certain time value of this individual sample; the survival function in Eq. (1) can be used to describe different time points, as well as the survival probability of this individual sample [26].

2.2 Logistic regression

Berkson [27] proposed that Logistic Regression is a kind of linear regression, and is frequently applied in addressing categorical data related issues. The most significant difference between general linear regression and logistic regression lies in the “processible data attributes”; logistic regression could be applied for processing binary data and predicting its odds ratio for the occurrence of an event, regardless of whether the predictor variable is a categorical variable or continuous variable.

When the dependent variable in logistic regression is replaced by a binary variable, *X* is assumed as a predictor independent variable, and *Y* is assumed as a binary target dependent variable, indicating that there are only two results, namely, failure (*Y=0*) or success (*Y=1*). Therefore, the general linear regression equation is established based on *Y*, as shown in Eq. (2):

$$f(x) = Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \,, \tag{2}$$

where, X_k denotes the independent variable of the k^{th} item, while coefficient β_k is the increment in dependent variable *Y* per each unit increased in X_k. *P* is assumed as a failure rate, while 1-*P* is assumed as a success rate; the odds ratio can be obtained by dividing the failure rate by the success rate $\frac{P}{1-P}$, while $\ln\frac{P}{1-P}$ is referred to as log odds; the log odds can be used for linearizing *Y*. Therefore, the logistic regression equation is shown in Eq. (3):

$$\ln\frac{P}{1-P} = f(x) = Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \,, \tag{3}$$

Let *x* denote the failure rate, which is a logistic distribution, $\bar{x} = \left(x_1, x_2, \ldots, x_k\right)$, and its equation is shown in Eq. (4):

$$p(\bar{x}) = p\left(Y = 1 \middle| \bar{x}\right) = F\left(g(\bar{x})\right) = \frac{1}{1 + e^{-g(\bar{x})}} \tag{4}$$

For logistic regression, the *k+1*th unknown parameter β_i in Eq. (3) must be estimated principally; while *i = 0, 1, 2, ..., k*; the unknown parameter β_i in this model can be estimated by the application of the maximum likelihood method.

In estimating parameter β_i, all observed values are regarded as a Bernoulli test, and the probability distribution function of random variable Y_i is shown in Eq. (5):

$$P(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \,, \quad i = 1, 2, \ldots, n \;\; \text{and} \;\; y_i = 0,1 \tag{5}$$

As observed values are independent of each other, they are subject to the marginal distribution of their product in Eq. (5), namely, joint distribution, which is also called the Likelihood Function, as shown in Eq. (6):

$$L(\theta)=\prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}=\prod_{i=1}^{n}(\frac{\exp(\beta' X_i)}{1+\exp(\beta' X_i)})^{Y_i}(\frac{1}{1+\exp(\beta' X_i)})^{1-Y_i}$$

$$=\frac{\exp(\beta')\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n}\left[1+\exp(\beta' X_i)\right]} \tag{6}$$

Substitute $t^*=\sum_{i=1}^{n} X_i Y_i$ into Eq. (6), and then, obtain its natural logarithm to obtain Eq. (7):

$$\ln(L)=\beta' t^* - \sum_{i=1}^{n}\ln\left[1+\exp(\beta' X_i)\right] \tag{7}$$

When we set $S(\beta)=\frac{\partial \ln(L)}{\partial \beta}=0$, we can obtain β_{ML}, and the β_{ML} equation is shown in Eq. (8):

$$S(\beta)=\frac{\partial \ln(L)}{\partial \beta}=-\sum_{i=1}^{n}\frac{\exp(\beta' X_i)}{1+\exp(\beta' X_i)}X_i+t^*=0 \tag{8}$$

where, $S(\beta)$ is the nonlinear function of β; therefore, this computational process must be conducted repeatedly, till β is converged to the maximum likelihood function. After obtaining coefficient β_k, success rate (*P*) can be obtained from Eq. (9):

$$P=f\,\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} \tag{9}$$

Logistic regression can be applied for class prediction, and its basis of prediction is based on the *P* calculated by the application of logistic regression. First, set a threshold value, which is generally 0.5; when the *P* calculated by the application of logistic regression is above the threshold value, it should be determined as a success (1); otherwise, it should be determined as a failure (0).

3. Research Method

3.1 Establishment of model data

Based on a certain time point, the data collected in this study could be simply divided into two categories: first, “outstanding cases”; such data refers to cases with an outstanding loan not in a defined default state by a certain time point, which will be predicted in this study; second, “historical data”; such data refers to cases in a defined state by a certain time point, for example, default, prepayment, or normal repayment; such data were used as the training model data in this study.

The main purpose of this study is to research and predict the default trend and the total number of default cases among outstanding business loan cases in the future. Therefore, historical cases in a defined state in the research data are used as training model data in this study, while outstanding cases in an undefined state were set for prediction. Modeling data in this study are processed in two phases.

(1) Phase 1: Regarding timeliness-based data screening, as the timeliness of data was considered in few cases of modeling and prediction through data mining, the data of all historical cases can be directly incorporated for modeling. However, the timeliness of modeling data must be considered; therefore, data incorporated for modeling in this study were initially screened based on timeliness, in order to identify historical cases meeting the timeliness of outstanding cases, and then, appropriate historical data were used for modeling.

In researches regarding whether there is default for outstanding cases in the future, survival analysis has been frequently applied for discussion, which is mainly attributable to the inclusion of the "time" factor in analysis. Therefore, this study attempts to simulate the time characteristics of survival analysis for timeliness-based screening of modeling data.

Each of the outstanding case and historical case have time characteristics, and the two most important time variables include "survival time", representing the time from the day when loan was granted by the financial institution to the business to date; and "period", representing the repayment period, as agreed with the financial institution in this loan case. Where repayment is failed upon expiry of this period, it will be deemed as default. However, in this study, it was predicted whether repayment will be made on time or overdue within the period from the time point of survival till the period, provided that the loan case of a business has survived for a certain period.

While outstanding cases are predicted in this study, the time factor must still be considered. If historical cases with time characteristics inconsistent with that of outstanding cases were incorporated for modeling, the established model may be unable to produce a result truly close to the actual state, which may cause significant errors in the prediction results. The following three scenarios are presented in respect of inconsistent time characteristics:

Scenario 1: The survival time of a historical case is less than that of an outstanding case, indicating that the outstanding case is still under way, while the historical case has been closed. In such a case, the historical case may be unable to fully explain the potential default state of an outstanding case in the future.

Scenario2: The survival time of a historical case is greater than that of an outstanding case, indicating that the outstanding case has been closed, while the historical case is still under way. Therefore, the historical case is unable to provide a defined case state within the research time of the outstanding case, and thus, is not suitable for modeling the outstanding case.

Scenario 3: The period of a historical case is greater than that of an outstanding case. In such a case, we must further observe the state of survival time. If the survival time of the historical case is less than that of an outstanding case, it indicates that the historical case meets the outstanding case; if the survival time of the historical case is greater than that of an outstanding case, as mentioned in Scenario 2 above, it indicates that the historical case is not suitable for modeling the outstanding case.

(2) Phase 2: Sample size-based data screening is discussed in this phase. Subsequent to the timeliness-based screening of historical data, this study considers that the expected effect may not be achieved, as the established prediction model was not representative enough due to insufficient modeling data, thus, sample size as a result of timeliness-based screening had to be considered. In the case of an insufficient training sample size, an appropriate treatment must be made.

In order to avoid adverse impact on prediction results due to a non-representative model, which is a result of insufficient modeling data, sample size-based data screening is proposed in Phase 2 of this study. This phase is based on the results in Phase 1. The number of modeling data screened out in Phase 1 was confirmed in this phase, and the minimum sample size was set. Regarding how to select an appropriate sample size, many scholars have proposed “empirical law” to simplify complicated methods when selecting sample size. The research viewpoint proposed in Stevens [28] was adopted in this study. This scholar researched the relation between variables and sample size in the application of regression, and found a certain correlation between the minimal sample size and independent variables. Through demonstration, the scholar recommended at least 15 samples for one independent variable.

Sample size-based screening in Phase 2 was based on the number of modeling data screened out in Phase 1. If the number of modeling data screened out in Phase 1 was greater than the minimal sample size for modeling, no treatment was required, and the modeling data screened out could be used for modeling.

On the contrary, if the number of modeling data screened out in Phase 1 was less than the minimal sample size for modeling, the prediction result of the outstanding case would be subject to the majority rule.

3.2 Establishment of the prediction model

Appropriate modeling data for each outstanding case could be immediately used to establish a prediction model by applying logistic regression. In this study, historical cases meeting the time characteristics of each outstanding case were acquired for modeling, in order to establish an exclusive prediction model for the outstanding case. Then, the said prediction model was applied to predict the default probability and state of the outstanding case. That is to say, an exclusive model was customized for each of the outstanding cases to predict their respective default probability and state.

4. Empirical Analysis

In order to validate whether the process and method applied in this study could practically predict the default state in an effective and accurate manner, case validation was implemented by relevant programming and applying R language software with the information of middle and small-sized enterprises, which were granted loans by a financial institution in Taiwan.

4.1 Research data and variables

Research data are the information of middle and small-sized enterprises, which were granted loans by financial institutions in Taiwan. The time period for loans was from November 2001 to December 2015, with a total of 21,239 cases, where non-default cases accounted for approximately 93.69% of the total cases and default cases accounted for approximately 6.93% of the total cases, for a total of 51 variables for research data. In this study, the variables for research data are classified by nature into three categories, namely, response variable, time variable, and financial variable (as shown in Table 1).

Insert Table 1 Explanations of variables by nature

4.2 Establishment of a prediction model by applying the approach proposed in this study

Research data used in this study were divided into historical cases and outstanding cases. Historical cases were modeling data; outstanding cases indicate the accuracy that the prediction target validates the prediction model. Data were divided based on "whether the case had been closed by December 2015". There were a total of 19,330 historical cases after data division, including 18,090 non-default cases, accounting for 93.59% of the total historical cases; and 1,240 default cases, accounting for 6.41% of the total historical cases. The class imbalance ratio was approximately 14.59.

In addition, there were 1,909 outstanding cases, including 1,809 non-default cases, accounting for 94.76% of the total outstanding cases; and 100 default cases, accounting for 5.24% of the total outstanding cases. The class imbalance ratio was approximately 18.09.

4.2.1 Screening of modeling data

In this study, all historical data were screened in two phases according to the time characteristics of each outstanding case, in order to select modeling data that met the time characteristics of each outstanding case. Regarding data screening, the data of the outstanding cases was initially imported into R software according to their respective item numbers from the first case till the i^{th} outstanding case, which assumed that a total of 1,909 outstanding cases had to be imported. That is to say, $i = 1, 2, ..., 1909$. Upon the import of the i^{th} outstanding case, the survival time and period of the i^{th} outstanding case can be obtained. In this study, a_i was assumed as the survival time of the i^{th} outstanding case; and $b_i$was assumed as the period of the i^{th} outstanding case.

First, in the time characteristic-based data screening in Phase 1, the relevant information of historical cases was obtained by importing the cases into the software from the first case till the j^{th} case, and it was assumed that a total of 19,330 cases were imported. That is to say, $j = 1, 2, ..., 19330$. In this study, S_j was assumed as the survival time of the j^{th} historical case, and D_j was assumed as the period of the j^{th} historical case. Time-based screening was subject to the condition that "the survival time of a historical case is equal to or greater than that of the i^{th} outstanding case, and the period of the historical case is equal to or less than that of the i^{th} outstanding case", i.e., "$S_j \geq a_i$ and $D_j \leq b_i$". The historical case would be determined as data suitable for modeling of the i^{th} outstanding case upon satisfaction of the said condition, and would be determined as data not suitable for modeling of the i^{th} outstanding case in case of dissatisfaction of the said condition. It was assumed in this study that there were a total of N_i sets of modeling data identified for the i^{th} outstanding case.

After the modeling data meeting the time characteristics were screened out in Phase 1, this study also considered the indirect impact of the sufficient or excessive modeling data on the accuracy of the prediction model. Therefore, sample size-based data screening was conducted in Phase 2. A certain correlation between the minimal sample size and the independent variable was identified by Stevens [28], which recommended at least 15 samples for one independent variable. In this study, after deducting the response variables, the research data consists of a total of 50 independent variables; therefore, at least 750 samples were required for this study. If N_i sets of sampling data screened out in Phase 1 were greater than 750, such modeling data could be used immediately for modeling; if N_i sets of screened out sampling data in Phase 1 were less than 750, such modeling data were not suitable for modeling; however, the majority rule was applied instead for predicting the i^{th} outstanding case.

4.2.2 Establishment of the prediction model

After screening the historical cases in above two phases, the modeling data for each of the outstanding cases, which met their respective time characteristics, were acquired. In this study, the prediction model was established by applying logistic regression for modeling data satisfying the minimal sample size. For modeling data below the minimum sample size, the majority rule was applied for prediction.

The logistic regression analysis model was established by applying R software with screened out modeling data (N_i), in order to satisfy the minimum sample size, and the prediction model for the i^{th} outstanding case was obtained accordingly. Next, the default probability of the i^{th} outstanding case could be obtained by substituting the case into the model. The default probability was interpreted against the established point of tangency. It was determined that the outstanding case would be subject to default in the future (denoted with 1) if the default probability is greater than the point of tangency, and it was determined that the outstanding case would not be subject to default in the future (denoted with 0) if the default probability is less than the point of tangency. Determination results were saved in the database, and it was determined whether the i^{th} outstanding case was the last case (the 1,909th case). If the case was determined as the last outstanding case, it indicated that all outstanding cases have been predicted for default, and the predicting process could be ended. If the case was not determined as the last outstanding case, it would repeat the first step in screening, and repeat the above predicting process from the $(i+1)^{\mathrm{th}}$ outstanding case, till the last outstanding case.

Errors in prediction may be caused if a model is directly established using modeling data that failed to satisfy the minimum sample size; therefore, for such minority cases with insufficient samples, the majority rule was applied in this study for prediction. When it was determined that the N_i sets of sampling data screened out for the i^{th} outstanding case were less than the minimal sample size (750), the number of default cases and non-default cases in the N_i sets of sampling data would be calculated, and prediction result for the outstanding case would be determined by following the majority rule. That is to say, it would be determined that the case would be subject to default in the future (denoted with 1) if the number of default cases was greater than the number of non-default cases; and it would be determined that the case would not be subject to default in the future (denoted with 0) if the number of default cases was less than the number of non-default cases. The process flow of the prediction model proposed in this study is shown in Figure 1.

Insert Figure 1. Process flow of the prediction model proposed in this study

4.3 Establishment of the Cox proportional hazards model

In practice, many financial institutions would apply the Cox proportional hazards model [29] to predict whether outstanding cases would be subject to default in the future. The research data for establishing the Cox proportional hazards model were the same data of cases used for the approach proposed in this study. Like the research data division principle mentioned in section 4.2, such data were also divided into two types, namely, “historical cases” (19,330 cases in total) and “outstanding cases”

(1,909 cases in total).

Regarding the process for establishing the Cox proportional hazards model, first, the 19,330 historical cases and 1,909 outstanding cases were imported into R software, respectively; second, the Cox proportional hazards model was established by using all historical cases and applying the programming language of R software, and the Cox proportional hazards prediction model was acquired accordingly. Then, all outstanding cases were substituted into the Cox proportional hazards prediction model to obtain their respective default probability. Determination of default was conducted by using the default probability against the established point of tangency. For fair comparison with the prediction results, as obtained by applying the approach proposed in this study, the established point of tangency for the Cox proportional hazards model was the same as that of the approach proposed in this study. The basis of determination for the point of tangency was: It was determined that the outstanding case would be subject to default in the future (1) if the default probability is greater than the point of tangency, and it was determined that the outstanding case would not be subject to default in the future (0) if the default probability is less than the point of tangency.

4.4 Comparison of prediction results of the approach proposed in this study and the Cox proportional hazards model

For facilitating practical applications by financial institutions, prediction results were validated by comparing them with the actual default state of the research cases, including the total number of default cases, the number of default cases for different starting years, and the default trend. Regarding the outstanding cases, cases that were not closed by December 2015, but were closed by September 2017 with an actual state, were selected. The 1,909 outstanding cases consist of a total of 100 default cases, and the number of actual defaults in each year is shown in Table 2. As concluded by professionals in financial institutions, as based on their experience, generally, cases in the early stage are in an extremely unstable state; therefore, default cases have been frequently observed. However, with the passage of time, cases with a longer term are in a more stable state, default cases can be significantly reduced accordingly, and repayment can be made on time on the whole.

According to the prediction result by applying the Cox proportional hazards model, there were 13 default cases among the outstanding cases. This result is quite different from the actual number of default cases (100 cases). It can be seen from the total number of default cases that the number of default cases was significantly underestimated by 87 cases when the Cox proportional hazards prediction model was applied. The prediction results by applying the Cox proportional hazards prediction model and the actual number of defaults are shown in Table 2.

For the purpose of the approach proposed in this study, historical cases were screened in two phases, and the prediction model was established by applying logistic regression. According to the prediction result, there were a total of 108 default cases among the outstanding cases, which is close to the actual number of default cases (100 cases). It can be seen from the total number of default cases that the prediction result by applying the approach proposed in this study is closer to the actual default state. The prediction results by applying the approach proposed in this study and actual number of default cases is shown in Table 2.

Insert Table 2. Prediction results by applying different approaches and the actual number of default cases

It can be seen from Table 2 that the default trend, as predicted by applying the approach proposed in this study, for each year is consistent with the actual state, i.e., the largest number of default cases has been observed in cases performed from 2015, while cases performed from 2014 show a significant declining default trend. In addition, specifically, the number of default cases predicted by applying the approach proposed in this study for each starting year is quite close to the actual number of default cases. This indicates that an accurate prediction result consistent with the actual default state can be provided by applying the approach proposed in this study, in order to allow financial institutions to prepare an appropriate amount of capital conservation buffer.

5. Conclusion and Suggestions

In most literature, financial institution loan default prediction models were established by using all the data, without the consideration of the respective time characteristics of different cases. In this study, in consideration of the different time characteristics of outstanding business loan cases, data were screened in two phases to select the appropriate modelling data for each of the outstanding business loan cases in a customized manner, and a sound prediction model was established by applying logistic regression. In addition, the difference between the prediction results by applying the Cox proportional hazards model, as commonly applied by financial institutions, and the approach proposed in this study, was discussed based on actual cases. Empirical results show that the approach proposed in this study could provide a better prediction result closer to the actual default trend, which could provide financial institutions with effective decision-making information, help businesses accurately understand the default trend to prepare a sufficient amount of capital conservation buffer, resist fluctuations in economic cycle as a whole, and maintain the stability of financial institutions. Future researches may attempt modelling by applying different algorithms, such as random forest, neural network, and decision tree, to discuss whether different algorithms could improve the prediction results and accuracy of the model.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research under Contract No. MOST 109-2410-H-145-002, MOST 110-2410-H-145-001, MOST 109-2221-E-009-070, and MOST 110-2221-E-A49-120.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Yung-Chia Chang https://orcid.org/0000-0001-6272-3560

Kuei-Hu Chang https://orcid.org/0000-0002-9630-7386

References

1. Chang, Y.C., Chang, K.H., and Hsiao, C.W. "A novel credit risk assessment model using a granular computing technique", *Journal of Testing and Evaluation*, **42(6)**, pp. 1427-1437 (2014).
2. Croux, C., Jagtiani, J., Korivi, T., Vulanovic, M. "Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform", *Journal of Economic Behavior and Organization*, **173**, pp. 270-296 (2020).
3. Basel Committee on Banking Supervision, "Basel III: The Liquidity Coverage Ratio and liquidity risk monitoring tools", *Basel Committee on Banking Supervision* (2013).
4. Noh, H.J., Roh, T.H., Han, I. "Prognostic personal credit risk model considering censored information", *Expert Systems with Applications*, **28(4)**, pp. 753-762 (2005).
5. Gepp, A., Kumar, K. "Predicting financial distress: A comparison of survival analysis and decision tree techniques", *Procedia Computer Science*, **54**, pp. 396-404 (2015).
6. Gupta, J., Gregoriou, A., Ebrahimi, T. "Empirical comparison of hazard models in predicting SMEs failure", *Quantitative Finance*, **18(3)**, pp. 437-466 (2018).
7. Ng, G.S., Quek, C., Jiang, H. "FCMAC-EWS: A bank failure early warning system based on a novel localized pattern learning and semantically associative fuzzy neural network", *Expert Systems with Applications*, **34(2)**, pp. 989-1003 (2008).
8. Chang, Y.C., Chang, K.H., Chu, H.H. Tong, L.I., "Establishing decision tree-based short-term default credit risk assessment models", *Communications in Statistics - Theory and Methods*, **45(23)**, pp. 6803-6815 (2016).
9. Dirick, L., Claeskens, G., Baesens, B. "Time to default in credit scoring using survival analysis: a benchmark study", *Journal of the Operational Research Society*, **68(6)**, pp. 652-665 (2017).
10. Han, J.T., Choi, J.S., Kim, M.J., Jeong, J. "Developing a risk group predictive model for Korean students falling into bad debt", *Asian Economic Journal,*, **32(1)**, pp. 3-14 (2018).
11. Lippi, A., Barbieri, L., Poli, F. "Money transfer between banks Evidence regarding the factors affecting speed of portfolio transfer when advisors migrate", *International Journal of Bank Marketing*, **38(2)**, pp. 283-295 (2020).
12. Kocenda, E., Iwasaki, I. "Bank survival in Central and Eastern Europe", *International Review of Economics & Finance*, **69**, pp. 860-878 (2020).

13. Yap, B.W., Ong, S.H., Husain, N.H.M. “Using data mining to improve assessment of credit worthiness via credit scoring models”, *Expert Systems with Applications*, **38(10)**, pp. 13274-13283 (2011).

14. Liu, D., Li, T.R., Liang, D.C. “Incorporating logistic regression to decision-theoretic rough sets for classifications”, *International Journal of Approximate Reasoning*, **55(1)**, pp. 197-210 (2014).

15. Zhang, S.Y., Tjortjis, C., Zeng, X.J, Qiao, H., Buchan, I., Keane, J. “Comparing data mining methods with logistic regression in childhood obesity prediction”, *Information Systems Frontiers*, **11(4)**, pp. 449-460 (2009).

16. Cheng, C.J., Chiu, S.W., Cheng, C.B., Wu, J.Y. “Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan”, *Scientia Iranica*, **19(3)**, pp. 849-855 (2012).

17. Sheets, L., Petroski, G.F., Zhuang, Y., Phinney, M.A., Ge, B., Parker, J.C., Shyu, C.R. “Combining contrast mining with logistic regression to predict healthcare utilization in a managed care population”, *Applied Clinical Informatics*, **8(2)**, pp. 430-446 (2017).

18. Le, T.H.M., Tran, T.T., Huynh, L.K. “Identification of hindered internal rotational mode for complex chemical species: A data mining approach with multivariate logistic regression model”, *Chemometrics and Intelligent Laboratory Systems*, **172**, pp. 10-16 (2018).

19. Chen, W., Yan, X.S., Zhao, Z., Hong, H.Y., Bui, D.T., Pradhan, B. “Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China)”, *Bulletin of Engineering Geology and the Environment*, **78(1)**, pp. 247-266 (2019).

20. de Bem, P.P., de Carvalho, O.A., Matricardi, E.A.T., Guimaraes, R.F., Gomes, R.A.T. “Predicting wildfire vulnerability using logistic regression and artificial neural networks: a case study in Brazil's Federal District”, *International Journal of Wildland Fire*, **28(1)**, pp. 35-45 (2019).

21. Najafi-Ghobadi, S., Najafi-Ghobadi, K., Tapak, L., Aghaei, A. “Application of data mining techniques and logistic regression to model drug use transition to injection: a case study in drug use treatment centers in Kermanshah Province, Iran”, *Substance Abuse Treatment, Prevention, and Policy*, **14(1)**, Article Number: 55 (2019).

22. Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N., Mansoori, Z. “Type2 diabetes mellitus

prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches", *BMC Bioinformatics,* **21(1)**, Article Number: 372 (2020).

23. Nalic, J., Martinovic, G., Zagar, D. "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers", *Advanced Engineering Informatics*, **45**, Article Number: 101130 (2020).
24. Chang, Y.C., Chang, K.H., Wu. G.J. "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions", *Applied Soft Computing*, **73**, pp. 914-920 (2018).
25. Maleki, M.R., Amiri, A., Taheriyoun, A.R. "Identifying the time of step change and drift in phase II monitoring of autocorrelated logistic regression profiles", *Scientia Iranica*, **25(6)**, pp. 3654-3666 (2018).
26. Kleinbaum, D.G., Klein, M. "Survival analysis", *New York: Springer* (2010).
27. Berkson, J. "Application of the logistic function to bio-assay", *Journal of the American Statistical Association*, **39(227)**, pp. 357-365 (1944).
28. Stevens, J. "*Applied multivariate statistics for the social science*", New Jersey: Lawrence Erlbaum, (1996).
29. Cox, D.R. "Regression models and life-tables", *Journal of the Royal Statistical Society, Series B,* **34(2)**, pp. 187-222 (1972).

Biographies

Yung-Chia Chang is a professor in the Department of Industrial Engineering and Management at National Yang Ming Chiao Tung University, Taiwan. She received her Ph.D. in Industry and System Engineering from Texas A&M University, College Station, Texas, USA. Her research interests include supply chain and quality management, data analytics and smart manufacturing.

Kuei-Hu Chang is a professor at Department of Management Sciences in R.O.C. Military Academy, Taiwan. Dr. Chang received his Bachelor's degree in Mathematics from the R.O.C. Military Academy in 1996. He received his Master's degree in Resources Management at the National Defense Management College in 2000, and acquired his Ph.D. in the Department of Industrial Engineering and Management at National Chiao-Tung University in 2008. His main researches are centering around fuzzy logic, linguistic algorithms, supply chain, soft computing, and reliability.

Yi-Xin Lin is a Master's graduate student in the Department of Industrial Engineering and Management at National Yang Ming Chiao Tung University, Taiwan. She received her Bachelor's degree in Industrial Engineering from Chung Yuan Christian University, Taiwan. Her research interests include data analytics, data mining, and applied statistics.

List of Figures

List of Tables

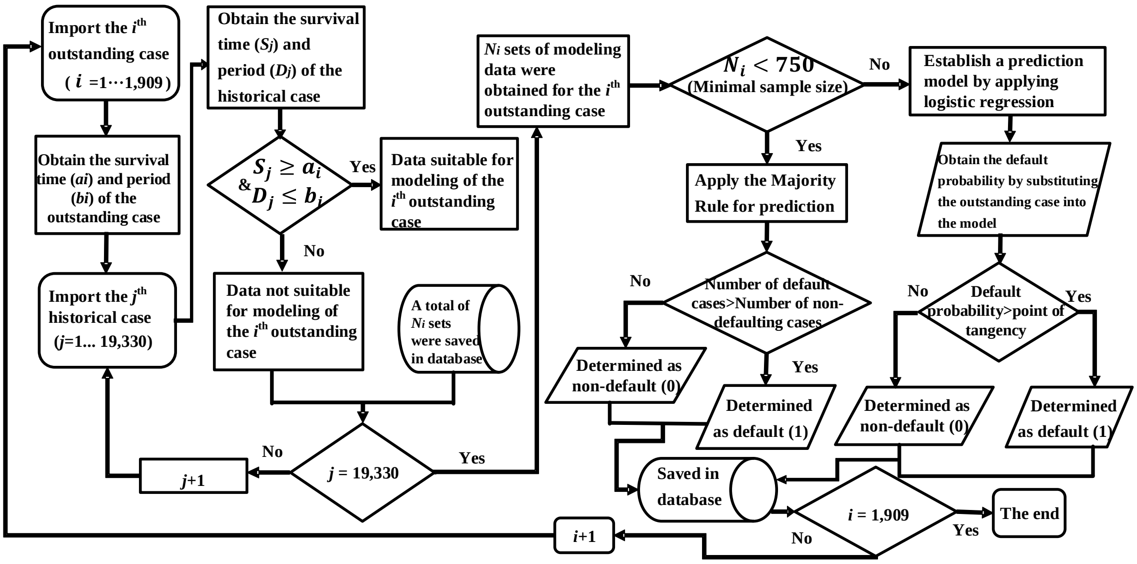
Import the i^{th} outstanding case (i =1⋯1,909)
Obtain the survival time (ai) and period (bi) of the outstanding case
Import the j^{th} historical case (j=1... 19,330)
Obtain the survival time (S_j) and period (D_j) of the historical case
$S_j \geq a_i$ & $D_j \leq b_i$
Yes
Data suitable for modeling of the i^{th} outstanding case
No
Data not suitable for modeling of the i^{th} outstanding case
A total of N_i sets were saved in database
j = 19,330
No
j+1
Yes
N_i sets of modeling data were obtained for the i^{th} outstanding case
$N_i < 750$ (Minimal sample size)
No
Establish a prediction model by applying logistic regression
Yes
Apply the Majority Rule for prediction
Obtain the default probability by substituting the outstanding case into the model
Number of default cases>Number of non-defaulting cases
No
Yes
Determined as non-default (0)
Determined as default (1)
Default probability>point of tangency
No
Yes
Determined as non-default (0)
Determined as default (1)
Saved in database
i = 1,909
Yes
The end
No
i+1

Figure 1. Process flow of the prediction model proposed in this study

Table 1 Explanations of variables by nature

No.	Nature of variable	Variable	Class	Explanation
1	Response variable	Default	Default and non-default	It refers to the final state of a case if the case has been closed by September 2017
2	Time variable	Period		It refers to the time of repayment, as agreed with the financial institution on the case; unit: month
3	Time variable	Starting year		It refers to the year when the loan was granted by the financial institution for the case
4	Time variable	Observation period		It refers to the remaining observation time up to the time of repayment, as agreed for the case; unit: month
5	Time variable	Survival time		If the loan in the case is still outstanding, it refers to the current repayment installment; unit: month; in case of default in the case, then installment when default was incurred; unit: month
6	Financial variable	Net loan amount (NTD',000)	Continuous number	Loan amount; unit: NTD 1,000
7	Financial variable	Financial strength	In classes from 0 to 4	In classes from 0 to 4
8	Financial variable	Reliability of financial report	In four classes	Truthfulness of financial statements
9	Financial variable	Proprietary capital ratio	In classes from 0 to 6	A company's shareholders' equity (capital + profit)
10	Financial variable	Gearing ratio	In classes from 0 to 4	Profit to be created by proprietary funds
11	Financial variable	Fixed ratio	In five classes	Flexibility of the application of capital
12	Financial variable	Liquidity ratio	In five classes	It refers to the ratio between a company's current assets and current liabilities and can reflect its short-term solvency
13	Financial variable	Quick ratio	In five classes	A measure of a company's capability to immediately cash out its current assets to repay current liabilities
14	Financial variable	DSR	In classes from 0 to 9	Debt service capability
15	Financial variable	Net turnover	In five classes	Average net turnovers
16	Financial variable	Accounts receivable days	In five classes	Average accounts receivable days
17	Financial variable	Days sales outstanding	In five classes	Average days sales outstanding
18	Financial variable	Gross profit	In four classes	To reflect the added value of a company's products
19	Financial variable	Net profit margin	In four classes	A company's profitability
20	Financial variable	Return on equity	In five classes	A company's rate of return on shareholders' investments
21	Financial variable	Net earnings per share growth rate	In five classes	Earnings per share growth rate
22	Financial variable	Turnover growth rate	In five classes	Turnover growth rate

Table 2. Prediction results by applying different approaches and the actual number of default cases

Starting year	2015	2014	2013	2012	2011	Total (case)
Actual number of default cases (case)	86	11	3	0	0	100
Cox proportional hazards model (case)	11	2	0	0	0	13
Approach proposed in this study (case)	97	8	4	0	0	109