Research Article

Journal of Applied Science and Emerging Technology (JASET)
Published by Faculty of Applied Science, KMUTNB, Thailand
doi: 10.14416/JASET.KMUTNB.2024.01.003
Vol. 23, No. 1 [2024]: e253276

Estimation of Loan Repayment Events in Microfinance Bank Using L1 - Lasso Penalized Cox Proportional Hazards Approach

Usman Mohammed[1*], Doguwa Sani[1], Alhaji Bukar[2], Dikko Hussaini[1] and Anuwat Tangtha[3]

[1]Department of Statistics, Ahmadu Bello University Zaria, Nigeria
[2]Department of Mathematics, Nigerian Defence Academy Kaduna, Nigeria
[3]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok Thailand

ABSTRACT

This study applied L1- Lasso estimation for Cox proportional hazards model to select variables that are relevant to credit repayment rates of loan in Microfinance bank. Records of 186 borrowers in Federal College of Education Zaria Nigeria Microfinance bank who took loans from 2017 to 2022 were used to identify the most predictive variables for repayment rates of loan. The findings of this study reveal that age of the borrowers, loan amount, occupation, type of collateral, residence and amount of loan influence the repayment rates of loan. Finally, the variables selected by the model can be used in granting loan to borrowers in Microfinance banking. R Programming Language was used for the analysis.

KEYWORDS: Bank Credit, L1 - Lasso, Cox Model, Penalization and Predictive Variables

***Corresponding Author: Author:ummohammed67@yahoo.com**

Received: 27/06/2023; **Revised:** 30/11/2023; **Accepted:** 19/01/2024

1. INTRODUCTION

Basically, banks and other financial institutions engage in the following essential activities. These are:

1) accepting and safeguarding customer deposits including other assets.
2) making payments on behalf of customers.
3) granting credit to customers and
4) making investment.

The primary activity amongst these have been granting credit, and sometimes customers fail to repay the debt in time which they contractually owe in form of principal, interest and other fees, thereby defaulting on their obligation. This exposes such institutions to credit risk (Thackham, 2021). The term bank credit refers to the amount of credit available to a business or individual from a banking institution in the form of loans. It is the sum of money a person or business can borrow from a bank or other financial institutions (Twin, 2020). Risk Management is an important facet of Bank's policies. According to (Gunduz, 2020) credit risk is arguably the most significant, especially for commercial banks. In (Apostolik et al., 2009) defined credit risk as "the potential loss a bank would suffer if a borrower fails to meets its obligations". Socio-political and economic development of any country mostly rely on the ability of its banking institutions to give loans to their customers. It is one of the major economic functions of banks to finance investment activities by government, business and individuals. Granting credit to customers supports the growth of new businesses and jobs which promote economic activities. Banks earn most of their revenues from loan accounts. The repayment behavior of such loans is associated with many factors in the banking system (Li et al., 2022). When a bank experiences a financial problem, it may be as a result of bad loans. Banks are interested in ascertaining factors that are mostly influential in predicting loan repayment rates of the customers.

Survival analysis methods possess the ability to model the loan repayment rates of borrowers. It is a statistical procedure for data analysis for the estimation of time until an event occurs. In this paper the event of interest is the loan repayment. In survival analysis, the estimated time for loan repayment can be the number of months (or days) from the day this loan is granted to the day the loan is repaid.

An event in survival analysis can be referred to as different outcomes obtained in a study. In medicine, the event can be the onset of a disease or death. In engineering, event can be a failure of a machine or an equipment system. In social sciences, event can be a change in social status of an individual. Survival analysis can also be applied in credit modelling, where the time to repayment of loan can be modelled. Survival data can be divided into complete and censored. Complete data refers to an observation that contains the beginning and end date which is determined by the event time. On the other hand, censored data is incomplete and it occurs when the required information is not available from the beginning to the end of the study. The three types of censored observations can be collected in survival analysis studies. These are:

1) right-censored observations
2) left-censored observations and
3) interval censored observations

Right-censored observations occur most in survival analysis studies because, the actual survival times of the individual which is "incomplete" (i.e censored) at the right-hand side of the follow-up period of the study giving the observations that is short of the actual survival time (Kleinbaum & Mitchel, 2005). Right-censoring occurs when an individual did not experience the event before the study end, when an individual is lost to follow up and when an individual withdraws from the study.

The traditional linear regression models cannot be applied to survival data or time-to-event data. This may result in biased estimation and the results may be misleading (Zhang et al., 2021).

The methods of survival analysis can be applied to credit risk management due to the need of accurate credit risk calculations. Credit risk analysis is important in financial risk management, especially, in practical applications (Assef & Steiner, 2020). In (Doris et al., 2022) observed that, survival analysis studies allow for the prediction of time to default of loan obtained by considering the length of time taken, between the origin of loan and its default. Credit risk in finance and banking has drawn a considerable research attention (Mungasi & Odhiambo, 2019). Repayment of loan performance refers to the total sum of loans, a customer paid on time which was agreed between the parties. When a loan is not repaid, it may be due to inability of the customer to repay or he/she is not willing to repay (Nawai & Shariff, 2013). Ganiyu states that in the study of perspective on Nigerian financial safety net with qualitative analysis procedures, most banks had poor credit policies. Large amount of loans was granted without due regard of the ability of the customers to repay. In (Okpara, 2009) noted that the high rates of bank failure in Nigeria may be as a result of poor credit policies in place. According to (Xia et al., 2021) with recent advancement, accurate survival models are applied in the assessment of credit risk.

2. MATERIALS AND METHODS

A total of 186 customers who took loans from Federal College of Education Microfinance Bank Zaria, Nigeria were included in the study. The study aims to identify the factors that are associated with customers' repayments rates of loan in Microfinance bank. The dataset contains client descriptive variables which can influence the loan repayment rates. These are personal information about the client. It includes age (x_1), gender (x_2), marital status (x_3), educational background (x_4), residence (urban/rural) (x_5), occupation (civil servant/business) (x_6), purpose of loan (x_7), type of collateral (x_8), loan amount (x_9), repayment time (x_{10}) and repayment periods (y). Cox with L1-Lasso models were applied to perform variable selection, examine the regression coefficients path and build a model that can predict time-to-event of repayment rates of loan.

The following are some of the functions that are studied in survival analysis methods. These are survival function, hazard function, probability density function and cumulative hazard function. Let 'T' be a nonnegative random variable of survival time of a bank borrower from the time the loan is granted to the time the loan is repaid representing time until some events of interest occurs (such as repayment of loan). The response (dependent) variable has two parts – survival function and hazard function. The survival function is the complement of the hazards function and survival function describes the probability of loan repayment up to a specified duration of time. The hazard function on the other hand, describes the probability of not paying the loan in a specified period of time given T > t (cond. prob) (Sangeetha & Chitra, 2021).

Given a distribution of a random variable '*T*' the survival function denoted as S(t) is given as:

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

where *F(t)* is the cumulative distribution function of '*T*' representing the cumulative probability of a customer chosen at random to have a survival time '*T*' less than some stated value '*t*', and is given as:

$$F(t) \ = \ P(T \le t) \ = \ \int_0^t f(u)du \qquad (2)$$

where '*f*' is the probability density function of '*T*'

The hazard function denoted as $h(t)$ and is given as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \qquad (3)$$

Here, the hazard function $h(t)$ is the conditional probability that the loan repayment occurs in time interval $(t, \ t \ + \ \Delta t)$ given that the loan repayment has not occurred before time *t*. (3) can be written as:

$$h(t) \ = \ \frac{f(t)}{1 \ - \ F(t)} \qquad (4)$$

Probability density function f(t) of survival time '*T*', refers to the limit of probability in which a customer repays the loan in a small time period '*t*' to (*t* + Δ *t*) per its time interval of length Δt and is given as:

$$f(t) \ = \ \lim_{\Delta t \to 0} \frac{P(t < T \le t \ + \ \Delta t)}{\Delta t} \qquad (5)$$

Cumulative hazard explains the accumulated hazards to time '*t*'. Thus, the cumulative hazard function, denoted as H(t) is obtained by taking the integral of hazard function given as:

$$H(t) \ = \ \int_0^t h(u)du. \qquad (6)$$

The cumulative hazard describes the cumulation of risks of a customer when the time passes from 0 to *t*.

2.1 Least Absolute and Shrinkage Selection Operator (Lasso)

In this study, we focus on identifying the predictor variables (or exposure variables) that have the strongest impact on the repayment rates of loan by a customer. The study aims to use those predictor variables (x's) to predict the repayment rates of loan. The study will utilize the L1- Lasso and Cox proportional hazards approaches to the dataset. The Cox proportional hazards model is used to investigates the impact of different predictor variables on the probability of loan repayment while the L1- Lasso penalized model, select the most predictive variables and avoid overfitting. L1 - Lasso method proposed by (Tibshirani, 1996), regularizes linear regression method which shrinks the coefficients closer to zero and other coefficients to exactly zero for a sparse solution and therefore, improving the interpretation of the model.

Nowadays, we collect more and more data. Sometimes with many and poorly described different kind of variables. In some researches, data are collected with more variables than the observations. Lasso is a procedure designed to scout through the data and extract few variables that have the ability to predict outcomes with accuracy. The main purpose of Lasso is to perform variable selection and regularization in order to enhance simplicity and accuracy of the model. This is achieved by adding penalty term to the linear regression. Lasso shrink unimportant variables to zero and the nonzero variables are selected to be utilized in the model. Lasso is a refined procedure that minimize the prediction errors encountered in statistical modeling.

L1 – Lasso model forms a diamond shape in the plot for its constrained region, as shown in Figure 1. The diamond shape includes corners, and the proximity of the first point to the corner shows that the model comes with one coefficient, which is equal to zero. Thus, shrinking the regression coefficient of the variable to zero to perform variable selection. The closeness of the first point to the corner of the diamond shape shows that L1 – Lasso model comes with one coefficient of a variable, which equal to zero. Thus, shrinking the coefficient in the model to zero to select variable. In Figure 1, x-axis represents the coefficient (β_1) and y-axis represents the coefficient (β_2). The red ellipses contour represents the Residual Sum of Squares (RSS).

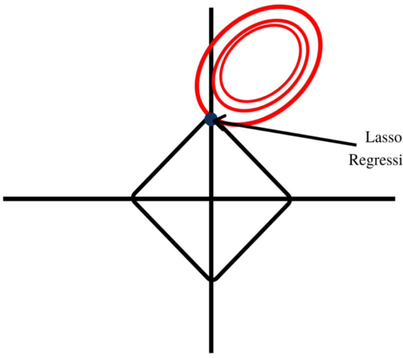

Figure 1 Lasso Geometry (https://corporate financeinstitute.com/data-science lasso)

L1 – Lasso model adds a constraint (or penalty) ($\lambda \| \beta \|_1$) to obtain equation (8). The consequence of applying this constraint is to reduce the regression coefficient towards zero, so that less contributive variables will have a regression coefficient close to zero or exactly equal to zero. In penalized regression, the aim is to reduce the impact of multicollinearity since predictor variables in the study may be highly related to one another (Abhinaya et al., 2021). L1 – Lasso selects variable and estimate the coefficient simultaneously by constraining the log-likelihood function of variable coefficients. Given a linear regression model:

$$Y = X^T \beta + \varepsilon \tag{7}$$

where **Y** is an $n \times 1$ column vector of dependent variable, **X** is an $n \times p$ matrix of predictor variables and β is a $p \times 1$ column vector of parameters. The last $n \times 1$ column vector is a vector of error terms. Also, $b_0, b_1, b_2 ..., b_p$ are the estimates of unknown parameters $\beta_1, \beta_2, ..., \beta_p$. In L1 – Lasso method, the coefficient β are estimated by minimizing

$$\hat{\beta}_{Lasso} = \underset{\beta}{\arg\min} \| y - x\beta \|_2^2 + \lambda \| \beta \|_1 \tag{8}$$

$$\hat{\beta}_{Lasso} = \underset{\beta}{\arg\min} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} | \beta_j | \right] \tag{9}$$

where

n is the number of observations

p is the number of predictor variables

$\lambda > 0$ is the regularization parameter.

$$\hat{\beta} = \text{argmin} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{10}$$

where $\sum_{j=1}^{p} | \beta_j | \le t$

Lambda (λ) is the regularization parameter which controls the shrinkage in estimating the coefficients of L1 – Lasso model and $\lambda \geq 0$. If the lambda (λ) value is large enough more variables of estimated coefficient β become zero, and the nonzero coefficient of variables will be shrunken toward zero. If lambda (λ) is small, it implies less regularization by the model. Cross-validation is a method applied to estimate the lambda (λ) parameter. When a small value of lambda (λ) is estimated, it may result in over fitting of the model. On the other hands, a large value of lambda (λ) would lead to under fitting, because the procedure may not be able to capture the relationship in the model (Thevaraja et al., 2019). A 'one-standard-error rule' method in cross-validation, will be applied to select the best lambda (λ). For each $MSE(\lambda_S)$ the standard error of the mean is obtained, and the largest λ_S is selected for which $MSE(\lambda_S)$ is within one standard error of the minimum MSE value. Thus, we obtain a regularized regression model while the MSE is increased by one standard error.

2.2 L1 - Lasso with Cox Proportional Hazards Model

Cox proportional hazards model is a regression model proposed by (Cox, 1972). It is an effective approach in survival analysis studies. The model is mostly used for multivariate regression analysis to analyze survival data. Cox model investigates the relationship between an event occurrence and a set of predictor variables (or covariates). The hazard function is the probability or the chance that an individual or subject will be affected by an event within an interval of time given that the individual or subject has survived up to the beginning of that interval of time. The response variable or outcome in Cox proportional hazards model is the hazard function at a given time. If a number of variables are involved, then the hazard or risk of an event', can be modeled by:

$$h(t,\mathrm{X}) = h_0(t)\exp(\beta^T \mathrm{X}) \qquad (11)$$

where $h_0(t)$ is the unspecified baseline hazard function which is the probability of an event when all the predictor variables (X) equal to 0.

$\boldsymbol{\beta}$ is a vector of parameters

X is a matrix of predictive variables.

Cox proportional hazards model, computes hazard ratio (HR), which measures the effect of predictor variables on the hazard of event. Estimate of HR of two individuals with different predictor variables X and X * is given as:

$$\hat{HR} = \frac{h_0(t)exp(\hat{\beta}' X)}{h_0(t)exp(\hat{\beta}' X^*)} = \exp\left[\sum \hat{\beta}' (X - X^*)\right] \qquad (12)$$

The HR does not depend on time. This is the reason why the model is refers to as proportional hazards model. In other words, proportional hazard is a required assumption in Cox regression. It means the relative hazard or risk of event which is the value of the coefficient $\boldsymbol{\beta}$ in the model is constant over time 't'.

Cox model can also be expressed by taking the natural logarithm of equation (11) and divide both sides by the baseline hazard function. $h_1(t)$ and $h_2(t)$ are the baseline hazards functions of the two individuals whose hazards of event are compared.

$$\log\left\{\frac{h_2(t)}{h_1(t)}\right\} = \log(\exp(\beta^T X)) \qquad (13)$$

$$\log_e\left\{\frac{h(t,X)}{h_0(t)}\right\} = \beta^T X \qquad (14)$$

In Cox model, there is no assumption made on the probability distribution of the hazards i.e baseline hazards function. Cox model assumes that the ratio of the hazard function of two individuals is constant over survival time and that there is log-linear relationship between predictor variables and hazard function. This assumption makes Cox proportional hazards model to be a semiparametric. The results of the analysis of Cox model can be interpreted as, for a unit increase in variable (X_i) the hazard function is multiplied by the term e^{β_i}. With this, the predictor variables have multiplicative effect with hazard function. Taking a unit increase in one variable for an individual with the hazard function in Cox model:

$$h_1(t) = h_0(t)e^{\beta x_1} \qquad (15)$$

Then for one unit increase

$$h_2(t) = h_0(t)e^{\beta(x_1+1)} \qquad (16)$$

$$\frac{h_2(t)}{h_1(t)} = \frac{h_0(t)e^{\beta(x_1+1)}}{h_0(t)e^{\beta x_1}} \qquad (17)$$

$$\frac{h_2(t)}{h_1(t)} = e^{\beta(x_1+1-x_1)} \qquad (18)$$

$$\frac{h_2(t)}{h_1(t)} = e^{\beta} \qquad (19)$$

then taking the logarithm of both sides we have:

$$\log\left\{\frac{h_2(t)}{h_1(t)}\right\} = \log e^{\beta} \qquad (20)$$

$$\log\left\{\frac{h_2(t)}{h_1(t)}\right\} = \beta \qquad (21)$$

In equation (21), the coefficient β is the logarithm of hazard ratio for a single unit increase in x_p. But when the variable increases by a single unit, the hazard of event happening (called hazard ratio) will increase by e^{β} unit. The β - parameter in equation (11) is estimated by maximizing the partial likelihood method.

Researchers are interested in the associations between each of the risk factors (X_1, X_2, ..., X_p) and the results or outcome. The associations are determined by the coefficients in the model (b_1, b_2, ..., b_p). The estimated coefficients in the Cox regression model say b_1, is the change in the expected log of the hazard ratio relative to a one unit change in predictor variable X_1, holding all other predictors constant.

Given a vector (**t**, δ , **x**) that consist of three items ***t*** is the length of time taken until an event occur or not occurring (censoring)

δ is the censoring indicator, 0 = censored, 1 = event. Here. **x** is a matrix of predictor variables.

Let '*n*' be the number of observed individuals in a study, '*r*' of them are affected by the event, and n − *r* individuals become right censored observations.

If $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ be an ordered event times and we let $\mathbf{X}_{(i)}$ be the vector of predictor variables with individual whose survival time is $t_{(i)}$. We define $R(t_{(i)})$, to be the risk of a set at $t_{(i)}$ as the set of individuals who are still in the study, the time earlier to $t_{(i)}$, then the probability or chance, that the individuals with predictor variable $X_{(i)}$ experience the event at $t_{(i)}$ given that one individual from $R(t_{(i)})$ experience the event at $t_{(i)}$ is given as:

$$\frac{h(t_{(i)}, X_{(i)})}{\sum_{j\in R(t_{(i)})} h(t_{(i)}, X_{(i)})} \qquad \text{by equation (11)}$$

The probability can be written in terms of the baseline hazard function and relative risk as:

$$\frac{h_0(t_{(i)})\exp(\beta^T X_{j(i)})}{\sum_{j\in R(t_{(i)})} h_0(t_{(i)})\exp(\beta^T X_j)}$$

The probability now, is given as:

$$\frac{\exp(\beta^T X_i)}{\sum_{j\in R(t_{(i)})} \exp(\beta^T X_j)}$$

It does not depends on the baseline hazard function since the baseline hazard function $h_0(t)$ cancel out. Cox (1972) made the assumption that if there is no tied event meaning that no two or more events occur at the same time, then parameter β can be estimated by the method of partial likelihood function. The probabilities are multiplied together over all distinct event times and the resulting product become conditional likelihood since it is a product of conditional probabilities.

$$L_p(\beta) = \prod_{i=1}^{n}\left[\frac{\exp(\beta^T x_{(i)})}{\sum_{j\in R(t_{(i)})} \exp(\beta^T x_j)}\right] \qquad (22)$$

In equation (22), *n* is the number of individuals who have experienced the event- repayment of loan at time 't', $x_{(i)} = (x_{(i)1}, x_{(i)2}, \ldots, x_{(i)p})$ are the predictor variables for the individual that experienced the event at the i^{th} ordered time $t_{(i)}$ and R_i is the set of subjects that are at risk just before time $t_{(i)}$.. Taking the logarithm of both sides of Cox partial likelihood of equation (22), we have:

$$\log(\mathrm{L_p}(\beta) \;=\; \sum_{i=1}^{n}\left[\log\frac{\exp(\beta^T x_{(i)})}{\sum_{j\in R(\mathrm{ti})} \exp(\beta^T x_j)}\right] \qquad (23)$$

$$\log(L_p(\beta)) = \sum_{i=1}^{n}\log(\exp(\beta^T x_{(i)})) - \sum_{i=1}^{n}\log\left[\sum_{j\in R(t_i)} \exp(\beta^T x_j)\right] \qquad (24)$$

$$\log(L_p(\beta)) = \sum_{i=1}^{n}\beta^T x_{(i)} - \sum_{i=1}^{n}\log\left[\sum_{j\in R(t_i)} \exp(\beta^T x_j)\right] \qquad (25)$$

And taking the partial derivatives of equation (25) with respect to each parameter β_h, h =*i*,....., *p*

$$U_h(\beta) = \frac{\partial}{\partial \beta_h}\log(L_p(\beta)) = \sum_{i=1}^{n} x_{(i)h} - \sum_{i=1}^{n}\frac{\sum_{\mathrm{j}\in \mathrm{R}(\mathrm{t}_i)} x_{(j)h}\exp(\beta^{\mathrm{T}} x_j)}{\sum_{\mathrm{j}\in \mathrm{R}(\mathrm{t}_i)}\exp(\beta^{\mathrm{T}} x_j)} \tag{26}$$

Equation (26) refers to as the scores and the estimates of the model are obtained by solving the equations i.e setting $U_h(\beta)$ = 0 according to (Ekman, 2017). Numerical method can be used to estimate the parameter β, for example Newton Raphson method. In Cox model the baseline hazard function is measured non-parametrically and therefore, the survival times are not assumed to follow a particular probability distribution at time 't' and the Cox model indicates that the hazard function or hazard rate may change over time. Estimates of the coefficient β in

Cox model with L1-Lasso are found by (27)

$$\hat{\beta} = \arg\min_{\beta}\left(-\sum_{i\in m}\log\left[\frac{\exp(\beta^T x_{(i)})}{\sum_{j\in R_i}\exp(\beta^T x_j)}\right] + \lambda\,||\,\beta\,||_1 \right) \tag{27}$$

$$\hat{\beta} = \arg\min_{\beta}\left(-\sum_{i\in m}\left[\beta^T x_{(i)} - \log\sum_{j\in R_i}\exp(\beta^T x_j)\right] + \lambda\,||\,\beta\,||_1 \right) \tag{28}$$

where $\lambda\,||\,\beta\,||_1$ is the penalty term of
L1 - Lasso model
The first term in equation (28)
is the log of the partial
likelihood of Cox model.

L1-Lasso performs variable selection. The regularized parameter lambda (λ) *is* chosen by *k*-fold cross-validation method, and *k* takes value between 5 and 10 (Hastie et al., 2015).

3. RESULTS AND DISCUSSION

This study utilized bank loans data of 186 customers obtained from Microfinance bank. The dataset consists of actual observations and censored. The censoring observation indicator is 0 for defaulting i.e non- payment of loan within the agreed period of time. and 1 for non-defaulting/event occurrence i.e payment of loan on time. From Table 1, the total number of borrowers is 186. 35 (18.8%) of them defaulted (unpaid loan as at when due) and 151 (81.2%) of them repaid their loan on time.

Table 1 Status of Repayment of Loan

	N	Percent (%)
Event	151	81.2
Censored	35	18.8
Total	186	100

In regression analysis, presence of multicollinearity in a dataset is a violation of one of the assumptions required by regression model. Multicollinearity is a situation whereby some variables in the regression model are related. A small bit of multicollinearity can cause huge problem in regression analysis. Therefore, detection of multicollinearity in a dataset is very important. The impact of multicollinearity can affect the precision of the estimated regression coefficients negatively on the power of a model. The variance inflation factor (VIF) is a method that quantifies the extent of correlation between one predictor and the others in a regression model. The VIF estimates how much the coefficient of a variable is inflated or influenced as a result of the predictor variables in the analysis. Higher values of VIF indicates that it is difficult or impossible to accurately assess the contribution of predictor variable in a model. The VIF value of one means that the predictor variable is not related with other variables. The VIF values greater than five (5) indicates the presence of multicollinearity.

Table 2 reports the VIF value of each predictor variable in the dataset, and since all VIF values are less than five, this indicates that multicollinearity is not present in the dataset. In other words, no two or more predictor variables are related to each other and therefore, we can proceed with the analysis.

Table 3 gives the summary statistics for sample size of the customers included in the study. Out of 186, 140(75.3%) males customers participated in the study, while only 46(24.7%) females customers participated in the study. The mean and median of survival times (repayment periods) in months for male customers are 7.29 and 7.00 respectively, while the mean and median of survival times for female customers in months are 7.08 and 7.00 respectively.

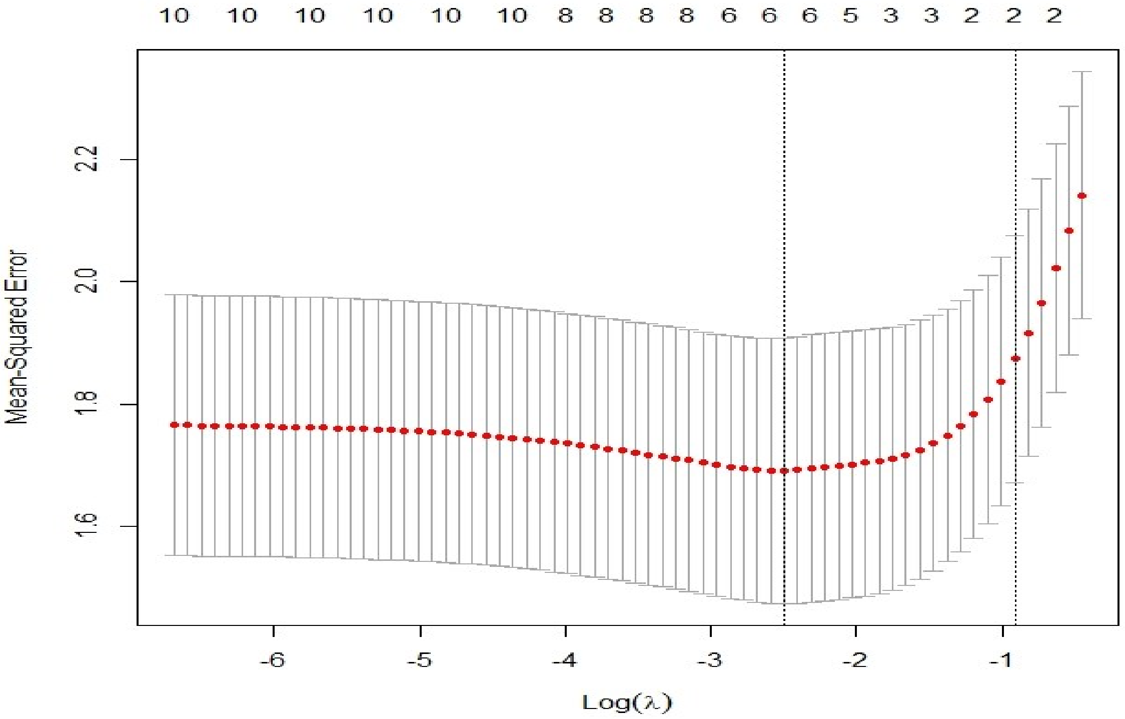

Figure 1 L1-Lasso Cross-validation Estimate for Mean Square Error

Table 2 VIF Values of the Ten (10) Variables

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1.38205	2.72735	0.32670	0.18461	1.59489	1.10752	2.38065	0.98787	1.10251	1.2354

Table 3 Summary Statistics for Sample Size

	N	Percent (%)
Male	140	75.3
Female	46	24.7
Total	186	100

The Cox model with L1-Lasso model were employed to perform variable selection and build a model that can predict for event occurrence i.e repayment rates of loan. L1-Lasso was applied to identified the important variables associated with the repayment rates of loan and Cox model est imates the hazards ratio (HR), obtains as the exponential of regression coefficient, and it gives the effect size of the important variables.

In Figure 1, the values at the top of the plot indicate the number of predictor variables in the model when lambda (λ) changes. Vertical dotted line on the left-hand side gives the lambda (λ) value for the minimum MSE and vertical dotted line on the right-hand side indicates the lambda (λ) that was chosen according to the model i.e MSE is within one standard error of the minimum MSE. Ten folds were used in cross-validation to obtained the optimal value of λ. From cross-validation results, the optimal value for lambda (λ) was found to be 0.06227052.

Figure 2 displays the path of the coefficients (lasso path) for every variable when using L1- Lasso penalized Cox model. In Figure 2, the curves that are away from the center line represent the selected variables that can influenced the repayment rates of loan. Those variables selected with statistical association for repayment rates of loan were suitable for multivariate analysis in Cox model. The Cox model was used to investigate the effect of those variables that can affect the risks of repayment rates of loan. Table 4 gives the six selected predictor variables by L1-Lasso model that can affect the time to survival of repayment rates of loan.

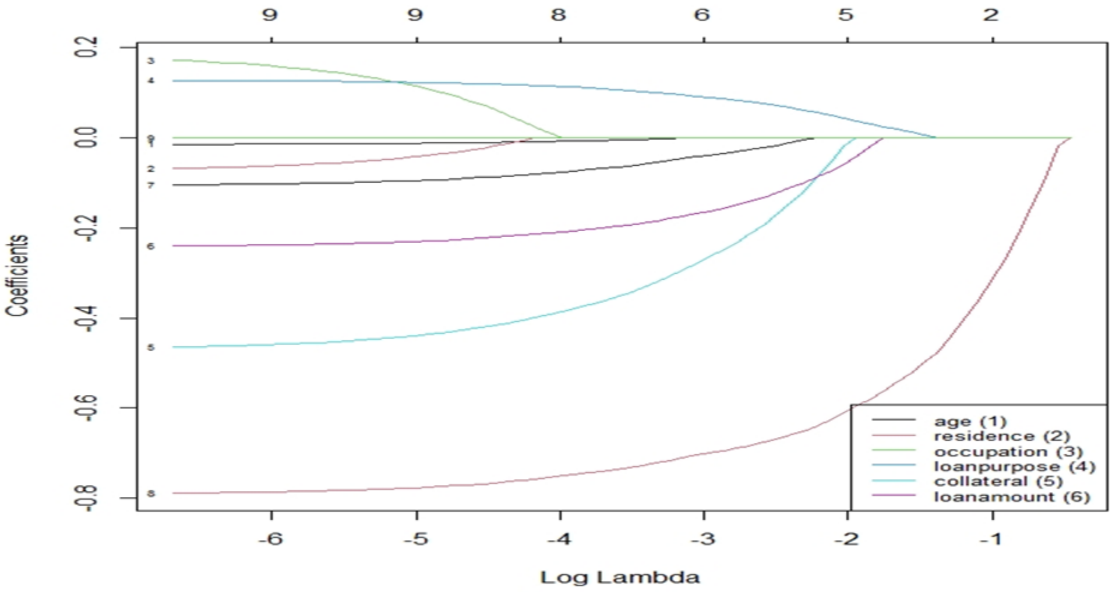

Figure 2 Coefficient Path (lasso path) for the Predictor Variables

Table 4 Multivariate Cox Proportional Hazard Results on the Time to the Repayment of Loans

Variables	coef (β)	exp(β) (HR)	se(β)	z-value	p-value
age 30-50 (**Ref.**) 51-70	-0.0377	0.9630	0.0498	-0.7590	0.4480
loanpurpose personal (**Ref.**) trading	0.0732	1.0760	0.3240	0.5540	0.5800
occupation business (**Ref.**) civil servant	0.1711	1.1870	0.1728	0.9900	0.3220
Collateral CFO (**Ref.**) salary acct	0.1469	1.1580	0.2113	0.6960	0.0121
residence rural (**Ref.**) urban	0.0925	0.0970	0.2083	0.4440	0.4870
loanamount	-0.0053	1.0000	0.0044	-1.1940	0.2320

CFO = certificate of occupancy

Ref. = reference category

3.1 Interpretation of the Predictor Variables in Cox PH Model

If the hazard ratio (HR) is greater than one, it indicates increase in the risk of the event i.e increase in the repayment rates of loan by a customer. On the other hand, a HR less than one indicates decrease in the event of interest i.e decrease in the risk of experiencing the event. Thus, decrease in repayment rates of loan by a customer. When the HR equals one it implies equal hazards of experiencing the event between the two categories of the customers.

i) Age

From Table 4, the HR of age is 0.9630 and the value is less than one. 100(1 - 0.9630) = 3.7%. This implies that a customer in age group 51 – 70 years is 3.7% times less likely to repay the loan within the agreed period of time compared to those customers in age group 30 – 50 years. In other words, granting loan to those customers in age group 51 – 70 years is a little bit risky.

ii) Loan purpose

The HR of loan purpose from Table 4, is 1.0760. The value is greater than one.

100(1.0760 – 1) = 7.6%. This means that a customer who is a trader is 7.6% times more likely to repay the loan in time compared to those customers who secured the loan for personal reasons.

iii) Occupation

Table 4, gives the HR of occupation to be 1.1870 and the value is greater than one. 100(1.1870 – 1) = 18.7%. This indicates that a customer who is a civil servant is 18.7% times more likely to repay the loan in time compared to those customers who obtained the loan for business.

iv) Collateral

The HR of collateral from Table 4, is 1.1580 and the value is greater than one. 100(1.1580 – 1) = 15.8%. This means that a customer whose salary acct was used as security against the loan by the bank is 15.8% times more likely to repay the loan in time compared to those customers whose CFO was collected by bank as security against the loan.

v) residence

Table 4, also reports the HR of residence to be 0.0970. The value is less than one and 100 (1 – 0.0970) = 90.3%. This means that a customer who lives in urban area is 90.3% times less likely to repay the loan in time compared to those customers who live in rural areas. In other words, loan advanced to customers in urban areas are more risky.

vi) loan amount

From Table 4, the HR of loan amount is one. This implies equal hazards or risks in the repayment rates of loan between the categories of the amount of loan granted to customers.

4. CONCLUSION

This study applied the L1-Lasso regularized Cox proportional hazards method to predict the event - repayment rates of loan of FCE Microfinance bank. With the number of selected variables that are truly informative, the method drops non-relevant variables. By discarding variables that are less important in L1-Lasso penalized Cox method, a parsimonious model was produced which can improve the interpretation of the model as compared to the classical statistical models. When a model is simple, its application and interpretation will be easier. This research identified the factors that affect the repayment rates of loan of Microfinance bank. From L1-Lasso Cox proportional hazards analysis, it is found that the repayment rates of loan is greatly influence by the predictor variable collateral with the p-value less than 0.05

Finally, the selected variables by the model can be used in issuing the loan in Microfinance banking. This study recommends that Microfinance banks should monitor the loans given to their customers in order to check any character change as it may affect the repayment rates of loan negatively.

REFERENCES

Abhinaya, D., Patil, S. G., Dheebakaran, G., Djanaguiraman, M., & Arockia, S. R. (2021). Use of statistical models in *predicting groundnut yield in relation to weather parameters*.

Apostolik, R., Donohue, C., & Went, P. (2009). *Foundations of banking risk: an overview of banking, banking risks, and risk-based banking regulation*. John Wiley & Sons. (pp. 16-22)

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

Ekman, A. (2017). *Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach* [Master thesis]. UMEÅ University University.

Fejza, D., Nace, D., & Kulla, O. (2022). The Credit Risk Problem-A Developing Country Case Study. *Risks*, *10*(8), 146.

Gunduz, V. (2020). *Risk Management in Banking Sector*. BahçeŞehir Cyprus University Banking and Finance. Publisher: Artikel Akademie. (pp. 121-135).

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, *143*(143), 8.

Kleinbaum, D. G., & Mitchel, S. (2005). *Survival Analysis: A Self-Learning Text* (3rd ed.). Springer. (pp. 98-164).

Li, H., Campbell, D., & Erdem, S. (2022). Measuring time preferences using stated credit repayment choices. *Journal of Quantitative Economics*, *20*(1), 43-67.

Mungasi, S., Odhiambo, C. (2019). Comparison of Survival Analysis Approaches to Credit Risks. .*American Journal of Theoretical and Applied Statistics*, 8(2), 39-46.

Nawai, N., & Shariff, M. N. M. (2013). Loan Repayment Problems in Microfinance Programs that use Individual Lending Approach: A Qualitative Analysis. *Journal of Transformative Entrepreneurship*, *1*(2), 93-99.

Ogunleye, G. A. (2010). *Perspectives on the Nigerian financial safety-net*. Nigeria Deposit Insurance Corporation.

Okpara, G. C. (2009). A synthesis of the critical factors affecting performance of the Nigerian banking system. *European Journal of Economics, Finance and Administrative Sciences*, *17*(17), 34-44.

Sangeetha, S., & Chitra, K. (2021). Solvency and Survival of Microfinance Institutions: An Indian Scenario-Policy Implications to Improve Endurance. *Indian Journal of Finance and Banking*, *5*(2), 130-140.

Thackham, M. (2021). *Survival Analysis: Applications to Credit Risk Default Modelling* [Doctoral thesis]. Macquarie University.

Thevaraja, M., Rahman, A., & Gabirial, M. (2019, April). Recent developments in data science: Comparing linear, ridge and lasso regressions techniques using wine data. In *Proceedings of the international conference on digital image & signal processing* (pp. 1-6).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Twin, A. (2020). The Investopedia Express Padcast. Retrieved 2020, from http:/www.investopedia.com/terms/b/bank-redit.asp

Xia, Y., He, L., Li, Y., Fu, Y., & Xu, Y. (2021). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy*, *27*(1), 96-119.

Zhang, D. F., Bhandari, B., & Black, D. (2020). Covariate Selection for Mortgage Default. Anareview approach and discussion. *Ingenieríae Investigación*, *40*(2), 50-71.