Application of Survival Model to Analyse Default Rates of Personal Bank Loans: The Case of a bank in Ghana

Gershon Kwame Mantey
Department of Mathematics Education, Faculty of Science Education
University of Education
Winneba-Ghana

Samuel Ewusi Dadzie
Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Isaac Gyasi
Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Francis Tabi Oduro
Department of Mathematics Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Correspondence:
mygersh2020@gmail.com
+233-50-136-5304

https://dx.doi.org/10.4314/ajmr.v28i2.5

Abstract:
The high levels of non-performing loans in Ghana over the past few years reduced the profitability of the banking industry which have caused bank failures that have adversely affected economic development. The study identifies the predictors for the risk of default of personal bank loans using data from a rural bank in Ghana. A sample of 196 personal loan borrowers was examined. The number of dependants, educational level, type of employer, gender, age, and marital status were noted. The Cox Proportional Hazard model was fitted using the sample data. Educational level, gender, age, and marital status were found to be non-significant predictors of default. However, the number of dependents and employer type were significant predictors of hazard. It was observed that hazard increased by 21.025% for an additional dependant a borrower takes on. The risk of default is 84.118% higher for a borrower whose employer is not government as compared to a government employee.

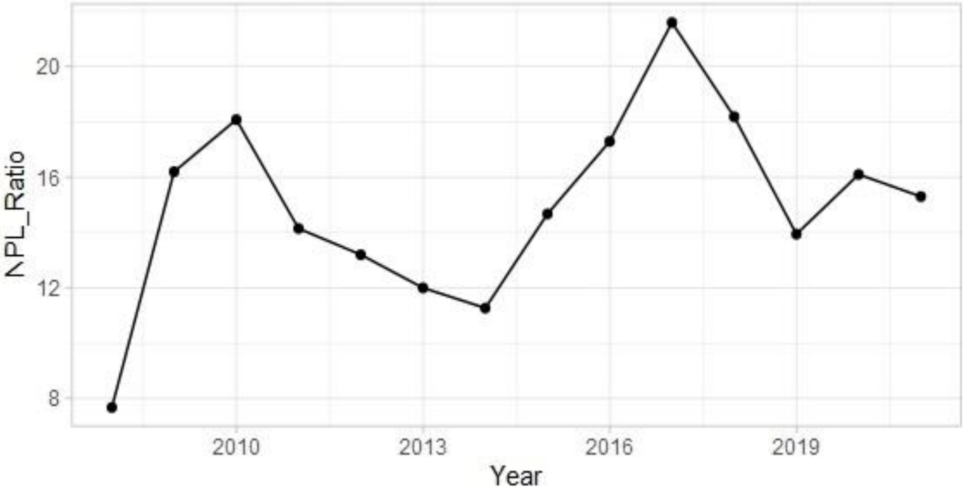
Key words: bank failures, Cox proportional hazard model, non-performing loan, risk of default, survival probabilities

Introduction

Loans are a significant part of the assets of banks. However, the rate of loan default in Ghana is on the increase due to poor analysis of the background of borrowers (Boachie, 2016). Records from the World Bank, and Census and Economic Information Center (CEIC) indicate that the non-performing loan (NPL) ratio of the country, which is a measure of the ratio of non-performing loans to total gross loans, reached a peak of 23.4% in April 2018. It was 22.7% in 2002, 18.08% in 2010, and in 2016, it was 17.29% (World

Bank, 2021; CEIC, 2021).

Figure 1: Non-performing Loan rate of Ghana



These high levels of non-performing loans (NPLs) are a cause for concern as they hurt the country's economy and the livelihood of the citizenry. Im et al. (2012) have argued that loan default has negative implications for the lender and the borrower. The borrower's credit standing is severely damaged making it difficult to secure a subsequent affordable loan. Banks on the other hand suffer because of loan default and in some cases, could result in insolvency due to the inability of their customers to pay back loans (Asantey and Tengey, 2014). The 2017 - 2019 banking sector crisis in Ghana saw some prominent banks, several rural banks, and Micro financial institutions collapsing, leaving thousands of people jobless. Other people had their monies lost or locked up. Key among the reasons given by the Bank of Ghana (BoG) for the collapse of these banks was insolvency due to NPLs on the banks' books mainly due to interrelated lending, loan approval without the necessary process, lending to risky borrowers, breaching the single obligor limit and general low compliance to common Corporate Governance practices (Aboagye, 2020; Osei et al., 2019; Torku and Laryea, 2021). It is undeniable that a major part of the income of banks is the interest earned on loans (Boachie, 2016; Whited et al., 2021). However, poor statistical analysis of the demographic backgrounds of borrowers before loans are issued due to numerous and competitive quick loan facilities in the banking sector has resulted in huge NPLs (Boachie, 2016; Nikolopoulos & Tsalas, 2017). Boateng and Oduro (2018) found demographic characteristics such as borrowers' educational level and the number of dependents, among other factors, have significant effect on default when they used the logistic regression model to analyse data of loan beneficiaries in Northern Ghana. Similarly, Agbemava et al. (2016) found marital status, and high dependency ratio to be statistically significant determinants in the prediction of loan default payment with a predicted default rate of 86.67%. An earlier study by Volkwein and Szelest

(1995) also found that earning good grades, persisting in completing a college degree, being married, and not having many dependent children substantially increase the likelihood of repayment and lower the likelihood of default. Fhatuwani and Karabo (2013) likewise found several sociodemographic variables including age, occupation, marital status, gender, number of dependents, residential area, education level, unemployment rate, and date since employment, to influence credit default. This study seeks to use survival analysis to estimate the hazard function for loan applicants of a rural bank in Ghana and use it to assess the relationship between predictor variables and the risk of default (hazard).

Concept of Loan Default and NPL

In the financial market, loan default is claimed to be the oldest and most significant form of hazard (Adams & Mehran, 2003; Agbemava et al. 2016; Arku 2013). A loan is defined to be delinquent when it is late in its payment. When the chances of recovery of a delinquent loan are minimal, it is said to have defaulted (Arku, 2013). A loan default occurs when a debtor fails to oblige to the legal obligations of a loan contract. This could be due to failure to make scheduled payments or violation of the terms and conditions of the loan. Murray (2011), and Pearson and Greef (2006) asserted that a loan default occurs when a borrower fails to pay back a debt or does not comply with the loan term agreement. Per the concept of Alton and Hazen (2001), a loan becomes non-performing if the full payment of principal and interest is unmet on the maturity date nor is it anticipated in the future. Their concept is in agreement with the International Monetary Fund (IMF, 2005) which went further to categorised NPL into three subsections namely;

- When debtors have not paid interest or principal payments in at least 90 days or more.
- When interest payments equal to 90 days or more have been capitalized, refinanced or delayed by agreement.
- When payments have been delayed by less than 90 days, but come with high uncertainty or no certainty the debtor will make payments in the future.

It is crucial to know however that the definition of NPL differs across various banking systems (Cucinelli, 2015). NPL globally is a measure of the financial health of the banking sector of economies which necessitates policymakers to consider its relevance for the stability of the macroeconomy (Asafo, 2018).

Survival Analysis vs The Generalised Logistic model

In their study, Madormo et al. (2013) suggested that survival analysis modelling for credit risk models leads to more robust conclusions when they compared the results of the survival model to classical models like generalised linear models based on logistic regression and non-parametric techniques based on classification trees. Banasik et al. (1999) likewise studied when borrowers will default. They postulated that significant results could be achieved by applying the survival analysis techniques to credit scoring when they compared the survival analysis approach to logistic regression in their estimation of which credits will be paid off ahead of schedule within the initial year and which clients still reimbursing after a year will pay off right on time within the following year. They observed that the PH model is at par with the logit regression approach in the identification of the people who default the primary year. They also demonstrated that the Cox PH model is better than the logistic model for observing who will pay off right in the primary year. Stepanova and Thomas (2001) studied the use of Cox Proportional Hazard (Cox PH) regression in building behavioural scoring models. They found the PH analysis of behavioural scores is highly competitive with the conventional logit modelling scores, particularly after about two (2) years into the credit. They argue that the movement of lenders from scoring techniques has made it important to use Cox PH regression since it could approximate the 'survival' likelihood of the credit over the long haul which is the likelihood of getting each month's reimbursements.

Whalen (1991) also argued that the Proportional Hazards (PH) model could be used to build an efficient early warning tool against bank failure. Similarly, Alves et al. (2009) explain the main financial ratios of private bank failure in Brazil using survival analysis. Lariviere and Van den Poel (2004) successfully used the PH model and Multinomial Probit Analysis in their study on how product features affect savings and investment customers of a large Belgian financial service provider.

Several early studies including Annesi et al. (1989), D'Agostino et al. (1990), Green and Symons (1983), and Ingram and Kleinman (1989) found the coefficients of regression of the logistic regression to closely approximate those of the Cox PH model. However, with a long follow-up time, outputs from the logistic regression become biased and not reliable making the survival analysis approach a more suitable, robust, and reliable model for analysing time-to-event data.

Materials and Methods

Survival Analysis Methodology

Event

An event in survival analysis means death, disease incidence, recovery, relapse, or an experience of interest that might occur to a person. It is additionally typically alluded to as a failure because it is generally some bad individual experience. However, in a situation where survival time denotes time to find employment after being unemployed for some time, failure is a positive event (Kleinbaum and Klein, 2010). In the context of this study, the event of interest is the default on personal bank loans.

Survival/Failure Time (T)

The outcome variable of interest in survival analysis is the survival/failure time. This is the time from the beginning of a study or when an individual entered the study up to:

i. The desired event happening, or

ii. End of the study, or

iii. Loss of contact with the individual or withdrawal of the individual from the study.

We define the failure time random variable, T, as a nonnegative ($T \geq 0$) which may be discrete or continuous (Rahardja and Wu, 2018). In defining a random variable for the failure time, there has to be:

1. An unequivocal time origin. That is, the start and end periods of the study must

be very specific.

2. A defined duration (e.g., days, weeks, months, quarters, years). The period must be very specific.

3. Clearly defined event (e.g., loan default, death, relapse of disease, etc.)

Censoring

This is the phenomenon where during the period of observation, the event which defines the survival time may not be experienced by some of the people in the study. This becomes what is known as censored information, which is, the time to event for those people who have not encountered the event under study is censored (before the study ends).

Censoring may as well occur when an individual during the study period is parted with and cannot follow through with the study due to death or other situations. Finally, it occurs when an individual pulls out of the study automatically as a result of death (i.e., when the event of interest is not death) or any other factors such as a negative reaction toward a drug (Kleinbaum and Klein, 2010). There are several types of censoring but the data used in this study is right-censored. In right censoring, the true unobserved event is to the right of the censoring time. The complete survival time interval is unknown. Thus, all that is known is that the event has not happened at the end of the follow-up or study.

Survival Function, S(t)

This is a nonincreasing function that specifies the probability that an individual survives
beyond some stated time t. It is a measure of the probability that the
survival/failure time is far ahead of some specified time t. The S(t) takes on the value of 1 at the start of the study (where t = 0) and 0 as t approaches infinity
( ). Theoretically, toward the beginning of a study, since nobody has encountered the event at this point, the likelihood of surviving beyond time zero is one. Also, if the study period increased without limit, eventually, everyone would experience the event, so the survivor curve must eventually fall to zero. Let T be a nonnegative random variable representing the waiting time until a customer defaults (Kleinbaum and Klein, 2010). The survival function, S(t) is given by:

$$\begin{aligned} S(T) &= P(T \geq t) \\ &= 1 - F(t) \end{aligned} \qquad (1)$$

where F(t) is the survival up to time t and t is any specific time value T

The Hazard Function, λ(t)

This is a nonnegative function that is greater or equal to zero (λ(t) - 0) without an upper limit. It provides the immediate potential for each time unit for the occurrence of an event, considering that an individual has subsisted up to a certain time t. The hazard function unlike the survivor function, which centers around nonfailure, centers around failure, which is the occurrence of the event of interest (Kleinbaum and Klein, 2010). The hazard function is expressed mathematically as;

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad (2)$$

is the pdf while $S(t)$ is the survival function. Between only the survival function and hazard function, a relationship may be obtained as follows:

$$S(t) = P(T \geq t)$$
$$= 1 - F(t)$$
$$\frac{d}{dt}S(t) = \frac{d}{dt}\left(1 - F(t)\right)$$
$$S'(t) = -f(t)$$
$$\lambda(t) = -\frac{S'(t)}{S(t)}$$
$$\lambda(t) = -\frac{d}{dt}lnS(t)$$

$$\int_0^t \lambda(\mu)d\mu = -\int_0^t \frac{d}{d\mu}lnS(\mu)d\mu$$

$$\int_0^t \lambda(\mu)d\mu = -lnS(\mu)\Big|_0^t$$
$$\text{-}\ln S(t) + lnS(0) = \int_0^t \lambda(u)du$$
$$lnS(t) = -\int_0^t \lambda(u)du$$
$$S(t) = e^{\left(-\int_0^t \lambda(u)du\right)} \qquad (3)$$

$\lambda(u)$ is the *hazard function* while $S(t)$ is the *survival function.*

Hazard Ratio

The hazard ratio (HR) is the ratio of the hazard for one person and the hazard for another. The two people who are analyzed can be recognised by their qualities for the collection of explanatory variables, (Kleinbaum and Klein, 2010).

$$Hazard\ Ratio = \frac{\lambda(t, X^*)}{\lambda(t, X)} \qquad (4)$$

Here, X^* is the explanatory variable set for the first person, while X represents that of the other person. On account of the Cox PH model, this simplifies to:

$$HR = e^{\sum_{i=1}^{P}\beta_i\left(X_i^* - X_i\right)} \qquad (5)$$

Cox Proportional Hazard Regression

This is a semiparametric method for investigating the effect of one or some predictors on the time a specified event takes to happen. Per the assumption of the Cox PH model, the Hazard Ratio associating two particulars of predictor variables is continuous over the long run (Kleinbaum and Klein, 2010). The model is usually given by:

$$\lambda(t, X) = \lambda_0(t)e^{\sum_{i=1}^{P}\beta_i X_i} \qquad (6)$$

Model (6) provides an equation for the hazard rate $\lambda(t, X)$, at a specified time for a person with a given specification of a set of indicators signified by X. Implying that, X denotes a vector of predictors being modelled to estimate the hazard function of a person. The Cox PH model stipulates that the hazard at any time t is the product of two quantities. The first is the *baseline hazard function*, $\lambda_0(t)$. The other is the exponential function *e* which is raised to the power of a linear summation $\beta_i X_i$. The summation is carried out on the p predictors *(X)* and *p* parameters *(β)* of the predictors. A significant aspect of this expression concerning the assumption of the Cox PH model states that the baseline hazard is a function of time t, however, *X* is not included.

This *X* may either be a time-independent variable or vice versa. This postulation of proportional hazards in the Cox proportional hazard model is broken when the model includes time-dependent covariates which result in what is known as the extended version of the Cox model.

The Cox model's baseline hazard $\lambda_0(t),$ is an unspecified function, the very property making it semiparametric. The Cox PH is a robust model such that the findings from using it will closely approach that of the actual parametric model. The exponential part of the Cox PH model ensures that the fitted model will always give estimated hazards that are nonnegative. Hence, using minimum assumptions, the fundamental information expected from survival analysis, which is *a hazard ratio* and *a survival curve* could be obtained with the Cox PH model.

Method of estimation

Estimation of the parameters of the Cox PH model is done by partial likelihood estimation where the partial likelihood function is maximised (Cox, 1975). Partial likelihood takes into account probabilities for those people who come up short and does not unequivocally factor in probabilities for censored persons. Notwithstanding, survival time data before censoring is utilized for the censored persons. Meaning the one censored later than the i^{th} time of failure is essential for the hazard collection for figuring out the i^{th} probability although this subject is cut out later. Let $R\left(t_i\right)$ denote the number of counts at risk of defaulting loan at a time t_i , which represents the risk set. It implies that the probability that a j^{th} case will default at a certain time T_i is expressed as:

$$P\left(t_j = T \mid R\left(t_i\right)\right) = \frac{e^{\beta X_i^{'}}}{\sum_{j \in R(t_i)} e^{\beta X_j^{'}}} \qquad (7)$$

The summation sign above in the denominator sums over every individual in the risk set. When we take the product of the conditional probabilities in equation (7), we obtain the partial likelihood function as:

$$L_p = \prod_{i=1}^{n} \left[\frac{e^{\beta X_i^{'}}}{\sum_{j \in R(t_i)} e^{\beta X_j^{'}}} \right]^{\delta_i} \qquad (8)$$

where *n* represents distinct failure times while δ_i is the failure time indicator, which is 0 if the case is censored, and 1 if the event of interest occurred. The log-likelihood function therefore becomes;

$$\log L_p = \sum_{i=1}^{n} \delta_i \left[\beta X_i^{'} - \log \sum_{j \in R(t_i)} e^{\beta X_j^{'}} \right]$$

(9)

When equation (9) is maximised, we obtain the β estimates.

The loan data

The data for the study was obtained from the headquarters of a limited liability rural bank in Ghana. The dataset for the study consists of 196 successful personal loan applicants of the bank from 1st March 2016 to 30th July 2020. Nine (9) variables are measured on each applicant.

Table 1. Variable names, labels, and values

Name	Label	Values
Default	Default Status	0 = “No” 1 = “Yes”
Dep	Number of Dependants	None
Edu	Educational Level	1 = “Illiterate” 2 = “Basic” 3 = “Post Basic”
EmpType	Type of Employer	1 = “Government” 2= “non-Government”
Gen	Gender of the loan applicant	1 = “Male” 2 = “Female”
Age	Age of loan applicant	None
MStatus	Marital Status	1 = “Single” 2 = “Married”
Iss_Date	Loan Issue Date	None
Rep_Date	Last Repayment Date	None
Months	Months between Iss_Date and Rep_Date	None

Results

Model Building

This study used six (6) predictors to build a predictive model for this study. Two of the predictors namely Dep and Age are quantitative variables. The other four: Edu (with 3 levels), EmpType (with 2 levels); Gen (with 2 levels) and MStatus (with 2 levels) are categorical. The Cox PH model was fitted as displayed in equation (10) based on (6)

Equation (10)

$$\lambda\left(t,X\right)=\lambda_0(t)\exp\left(\begin{array}{l}\beta_{Dep}X_{Dep}+\beta_{Basic}X_{Basic}+\beta_{Post\ \text{Basic}}X_{Post\ \text{Basic}}+\beta_{Age}X_{Age}\\+\beta_{Married}X_{Married}+\beta_{Non\ \text{Government}}X_{Non\ \text{Government}}+\beta_{Female}X_{Female}\end{array}\right)$$

Table 2. Initial model output

n = 196, number of events = 97

Variable	Coef	Exp(Coef)	Se(Coef)	Z	Pr(>|Z|)	Lower 95% CI	Upper 95% CI
Dep	0.212	1.237	0.066	3.235	0.001**	1.087	1.407
EduBasic	-0.491	0.612	0.315	-1.557	0.119	0.330	1.135
EduPost Basic	-0.118	0.889	0.291	-0.406	0.685	0.330	1.135
EmpType Non-Government	0.640	1.897	0.258	2.483	0.013*	0.502	1.572
GenFemale	-0.321	0.275	0.215	-1.495	0.135	0.476	1.105
Age	-0.003	0.997	0.007	-0.373	0.709	0.984	1.011
MStatusMarried	0.028	1.028	0.216	0.128	0.898	0.673	1.570

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Concordance = 0.631 (se = 0.032)
Rsquare = 0.117 (max possible = 0.992)
Likelihood ratio test = 24.450 on 7 df, p = 0.001
Wald test = 22.380 on 7 df, p = 0.002
Score (logrank) test = 22.960 on 7 df, p = 0.002

Coef: represents the coefficients of the model predictors, beta, which in the Cox PH model, could be interpreted as the change in the log of the hazard function for each one-unit change in a predictor, holding other predictors constant.

Concordance: measures the predictive power of the Cox PH model. A model with a concordance greater than 0.5 has good prediction ability. Similarly, a model with A concordance less than 0.5 has poor predictive ability while the predictive ability of a model with a concordance of 0.5 is no better than random chance.

Rsquare: indicates the proportion of the variance in the data that is explained by the fitted model. A high Rsquare is preferred over a low one as high Rsquare indicate that the fitted model explains more of the variation in the data. However, the Rsquare in Cox model is highly sensitive to the proportion of censored values. The expected value of Rsquare decreases substantially as a function of the percent censored observations. At the latter part of the output are three tests: **Likelihood ratio test, Wald test and Score test** for testing the global null hypothesis, β = 0. These are asymptotically equivalent tests that test the hypothesis that a set of predictors have no effect.

They test the global statistical significance of the model by testing the null hypothesis that all of the coefficients of the predictors, βs, in a Cox model are zero (0).

The null hypothesis, H_0 , and the alternate hypothesis, H_1 , are given by:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

It should be noted that for a small sample size, the Wald, score and likelihood ratio tests may differ by a small degree, but they would give similar results for a large enough sample size. The likelihood ratio test has better behaviour for small sample sizes, so it is generally preferred and therefore attention was limited to the likelihood ratio test in this study.

Test for the Significance of the Predictors

A test for the significance of the model predictors would involve testing the significance of the coefficients of the model predictors. A hypothesis test below was employed.

$$Null\ hypothesis, H_0 : \beta_i = 0$$
$$Alternate\ hypothesis, H_1 : \beta_i \neq 0$$

Where i would take one of the predictors (Dep, Basic, Post Basic, Non-Government, Female, Age, Married) at a time.

Test for the significance of Dep at a confidence level of 95%

$$H_0 : \beta_{Dep} = 0$$
$$H_1 : \beta_{Dep} \neq 0$$

In Table 2, Dep has a coefficient of 0.212287 with a standard error of 0.065627. It has a hazard ratio of 1.236502 with a confidence interval of (1.0873, 1.406). The hazard ratio of 1.236502 means the predictor Dep would increase the risk of default by 23.6502%. The test statistic for testing the significance of this predictor is 3.235 with a p-value of 0.00122 which implies the null hypothesis should be rejected. Hence the predictor Dep is significant in the model. In Table 2, the variables Dep and Non-Government employer type were the only significant predictors at a 95% confidence level.

Test for the significance of the level Non-Government of EmpType at a confidence level of 95%

$$H_0 : \beta_{NonGovernment} = 0$$
$$H_1 : \beta_{NonGovernment} \neq 0$$

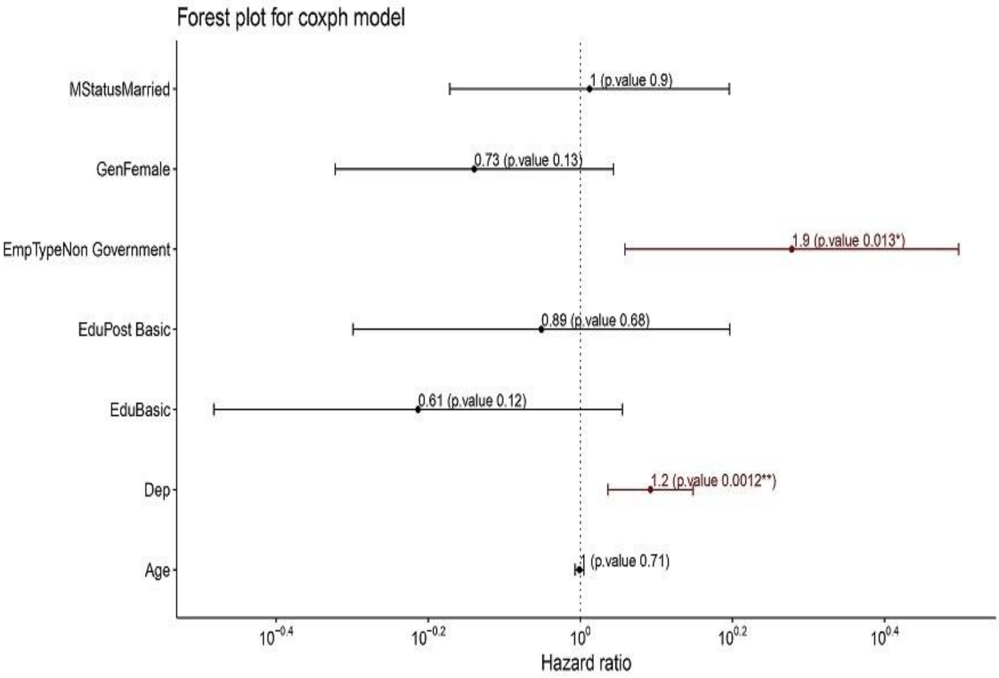
In Table 2, non-Government has a coefficient of 0.640040 with a standard error of 0.257778.

Non-Government has a hazard ratio of 1.896557 with a confidence interval of (1.1443, 3.143). A hazard ratio of 1.896557 means an individual with a non-Government as the type of Employer would increase the risk of default by 89.6557% as compared to an individual who is a government worker. The test statistic for testing the significance of this predictor (non-Government) is 2.483 with a p-value of 0.01303 which implies the null hypothesis should be rejected. Hence the predictor non-Government is significant in the model.

Testing the significance of each one of the other predictors at a 95% confidence level indicated that the predictors MStatus, Gen, Edu and Age are not statistically significant since the p-value of each one of them is greater than 0.05.

Moreover, the test statistic for the likelihood ratio test is 24.45 with a p-value of 0.0009489. Therefore, the global null hypothesis is rejected at a 5% significance level indicating that at least one of the coefficients is not zero (0). This conclusion from the likelihood ratio test confirms the observations that were made on that the predictors Dep and EmpType as significant predictors of the hazard function. The non-significance of some of the predictors of the model above implies a rejection of the model and therefore a new model should be built. Figure 2 is a forest plot for the Cox regression model (10). Hazard ratioestimates along with confidence intervals and p-values are plotted for each predictor.

Figure 2: First forest plot



Building of a new Model

Discarding the non-significant predictors, we focused on building a new model with only the significant predictors. This model would be of the form:

Equation (11)

$$\lambda(t,X)=\lambda_0\left(t\right)\exp\left(\beta_{Dep}X_{Dep}+\beta_{Non\_Government}X_{Non\_Government}\right)$$

Fitting the new model in R gives the following output:

Table 3: New model output

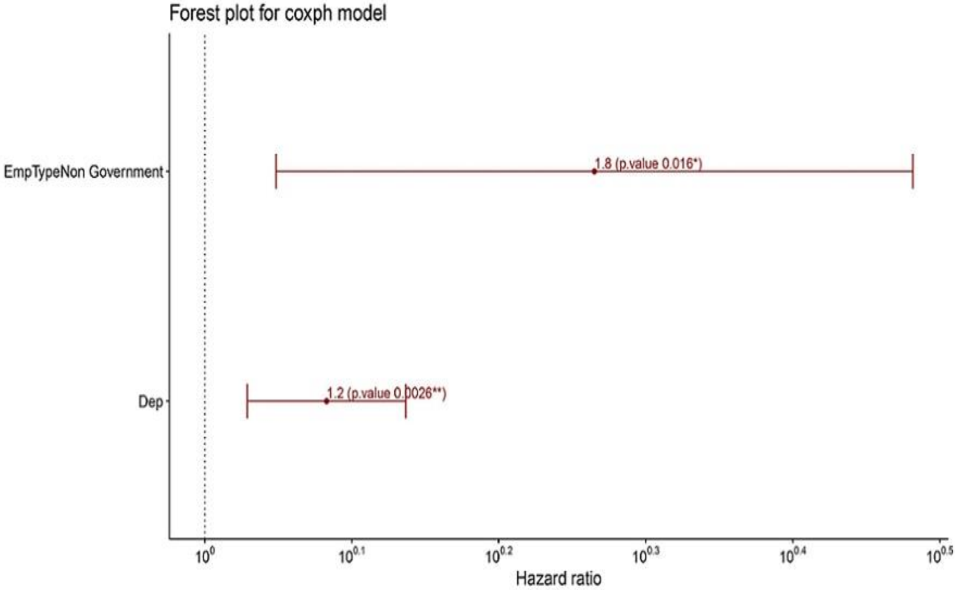
n = 196, number of events = 97

Variable	Coef	Exp(Coef)	Se(Coef)	Z	Pr(>|Z|)	Lower 95% CI	Upper 95% CI
Dep	0.191	1.210	0.063	3.012	0.003**	1.069	1.370
Non-Government	0.610	1.841	0.254	2.399	0.016*	1.118	3.032

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Concordance = 0.600 (se = 0.031)
Rsquare = 0.0900 (max possible = 0.992)
Likelihood ratio test = 18.390 on 2 df, p = 0.0001
Wald test = 22.380 on 7 df, p = 0.0002
Score (logrank) test = 22.960 on 7 df, p = 0.0001

The concordance of the model is 0.6 indicating that the model has good predictive power and hence could be used for making predictions. Although the model has a low Rsquare of 0.09, the Rsquare in the Cox model is affected by censored values. Thus, this low Rsquare value is expected because 49% of the observations are censored. Hence, even though the Rsquare value is low, the model still explains some of the variations in the data. Below, is a forest plot for the Cox regression model (11).

Figure 3: Second forest plot



Tests for the Validity of the new Model

Three tests, a test for proportionality of the hazard ratio, a test for influential observations, and a test for linearity of the parametric part of the cox model would be conducted to test the validity of the final model.

Table 4. Test for proportionality

	rho	chisq	p
Dep	0.153	2.063	0.151
ÈmpType non-Government	0.018	0.034	0.854
GLOBAL	NA	2.202	0.333

The null hypothesis, H_0, and alternate hypothesis, H_1, for testing the proportionality of Dep is given by:

H_0 : *The Hazard ratio of Dep is constant*

H_1 : *The Hazard ratio of Dep is not constant*

That of EmpType is given by:

H_0 : *The Hazard ratio of EmpType is constant*

H_1 : *The Hazard ratio of EmpType is not constant*

A global test of proportionality for the model is given by:

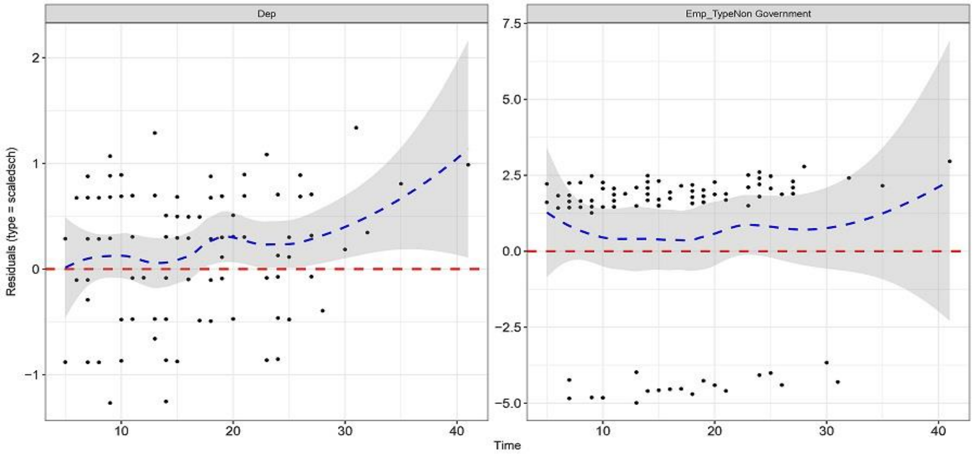
H_0 : *All the predictors of the model have constant hazard ratios*

H_1 : *At least one of the predictors of the* model *does not have a constant hazard ratio*

Results from Table 4 indicate that the test for proportionality of Dep returned a test statistic of 2.063 with a p-value of 0.151. The null hypothesis could not be rejected. Hence, Dep satisfies the proportional hazards assumption. Also, the test for proportionality of EmpType returned a test statistic of 0.034 with a p-value of 0.854. The null hypothesis could not be rejected. There is therefore strong evidence of proportional hazards for EmpType. Finally, the global test for proportionality returned a test statistic of 2.202 with a p-value of 0.333. Thus, the null hypothesis could not be rejected which goes to confirm that the two predictors satisfy the assumption of proportionality. A graphical assessment of the proportional hazards assumption to verify the results of the above test which involves plots of **scaled Schoenfeld** residuals against time for each predictor in the model would yield:

In a scaled Schoenfeld graph, systematic departures from a horizontal line are indicative of nonproportional hazards. That is, a non-zero slope is an indication of a violation of the proportional hazard assumption. Graphically, the assumption of proportional hazards appears to be supported by the two predictors (Dependents and EmpType) used in the model.

Figure 4: Scaled Schoenfeld graphs



This is because the slopes from their respective scaled Schoenfeld graphs are not significantly different from zero (0). This effect was also detected in the test reported above. Hence the rejection of the null hypothesis.

Test for influential observations

As part of assessing the validity of the model fit, checks are made to find the presence of influential observations. A plot of dfbetas would be used for this task. dfbetas is a measurement of how much effect every observed value has on a specific predictor. For a predictor and an observed value, the dfbetas is the difference between the coefficient of regression computed for the entire dataset and the one computed with the deleted observation, scaled by the standard error calculated with the deleted observation.

The threshold for *dfbetas* is $\left|\frac{2}{\sqrt{n}}\right|$, where n represents the total number of observations. In this study $n = 196$, which when evaluated gives a *dfbetas* cut-off of 0.1428571429. A plot of *dfbetas* to help with the analysis is as below:

Comparing the magnitudes of the dfbetas values to the dfbetas cut-off (0.1428571429) suggests that some of the observations are influential. Checks with the bank indicate that these observations are correct and so the presence of influential observations is ignored.

Test for linearity

In this section, the deviance residuals plot is examined to detect nonlinearity. This is done to ensure that the parametric part of the model is correctly specified. Deviance residuals are defined as a martingale residual and an event variable transformation. The martingale residual of an individual specifies excess failures beyond the expected baseline hazard. Deviance residuals are often symmetrically distributed around zero and have a standard deviation of 1.0. Non-linearity is not an issue for categorical variables (EmpType in this case) and so the focus of this analysis is on quantitative variables (Dep in this case). The deviance residuals plot obtained in R is displayed in Figure 6.

Figure 5: A plot of *dfbetas*

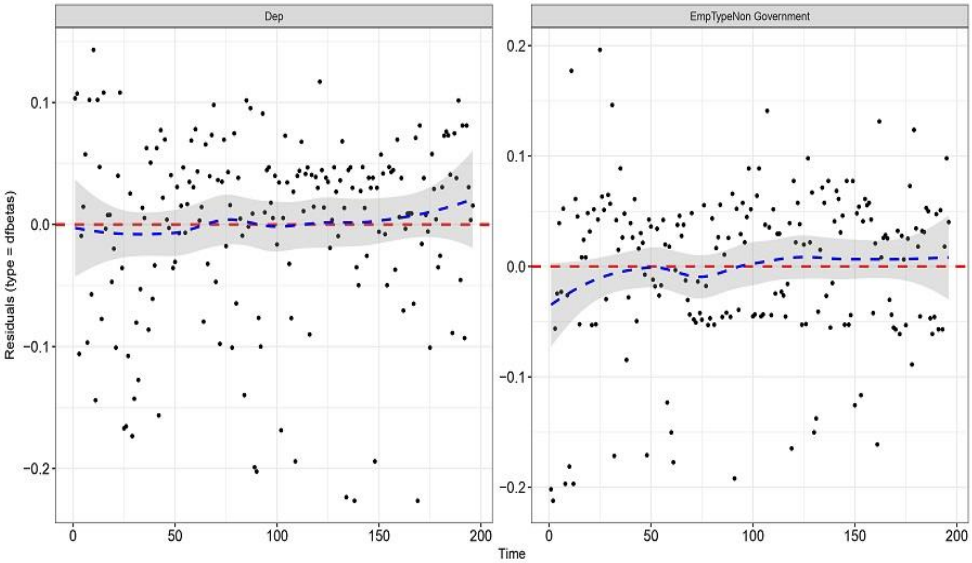
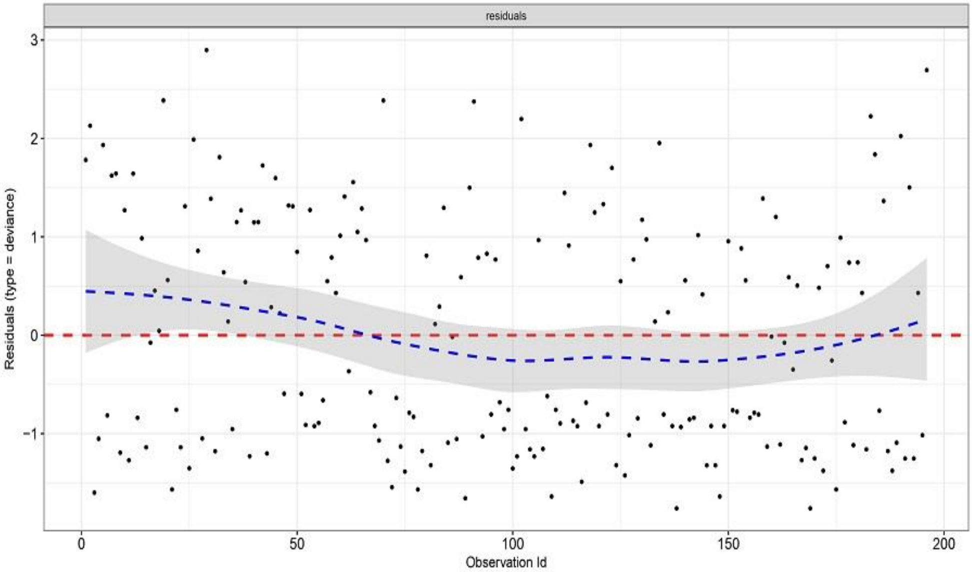


Figure 6: Deviance residuals plot



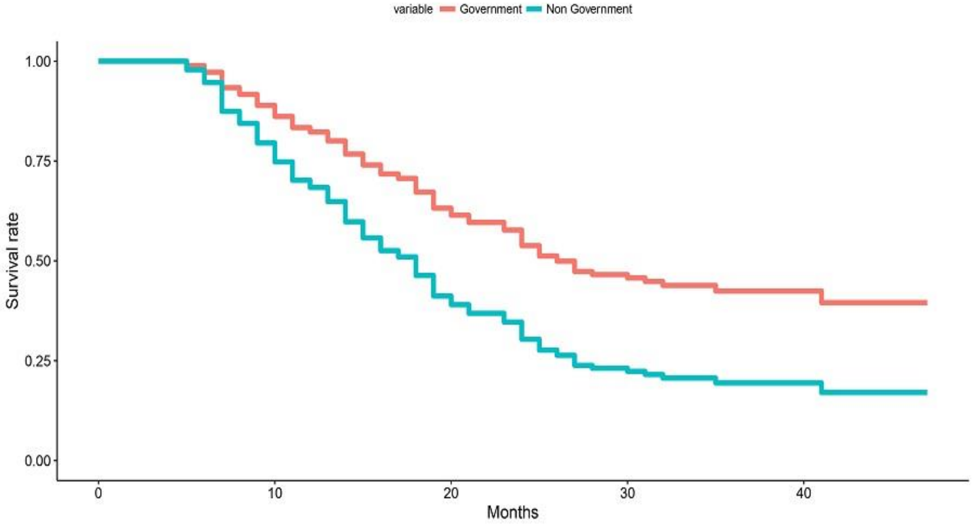
In a test of linearity, the more the deviance residuals plot is close to a zero line, the more the non-linearity can be excluded. From the plot, it appears nonlinearity, is only slight here. That is, the parametric part of the model has been correctly specified and fulfils the assumption of linearity.

Final Model of the Study
The final model of the study after going through all the necessary procedures of fitting
a Cox PH model is (11). Substituting the values of the coefficients would give:

$$\hat{\lambda}(t, X) = \hat{\lambda}_0(t)\exp(0.19083 X_{Dep} + 0.61040 X_{Non\_Government}) \qquad (12)$$

Adjusted Survival Curves

Having fit a Cox model to the data, the estimated distribution of survival times for each level of EmpType adjusted for the effect of Dep is examined. These survival curves show how having Government as one's employer or having an employer that is non-Government influences survival estimated from (12). The plot of the adjusted survival curves is displayed in Figure 7.



It is observed from Figure 7 that government employees consistently have higher survival probabilities than non-government employees after adjusting for Dep. Furthermore, the difference between government and non-government appears to widen over time.

Discussion

Much works has been done to find macroeconomic and bank specific determinants of NPLs. The current study applied the Cox PH model on microeconomic variables in establishing the link that exists between the risk of default and its predictors. Results show that level of education, gender, age as well as marital status of a loan applicant does not significantly affect their risk of defaulting. These current findings are in contrast to the results of Fhatuwani and Karabo (2013) who presented on the application of survival models to analyse default rates on bank loans at the Convention of Actuarial

Society of South Africa. They found all the insignificant variables noted here among others to be significant factors influencing default rate of loans. Boateng and Oduro (2018) likewise found educational level, type of loan, adequacy of the loan facility, duration for repayment of loan, number of years in business, cost of capital and period within the year the loan was given to have a significant effect on default of student loan. Earlier studies like Volkwein and Szelest (1995) also found that obtaining scientific and technological skills as well as getting and staying married thus being married, increases the likelihood of loan repayment. In this study, however, the predictors; number of dependants of an applicant, and the type of employer of the borrower, whether government or non-government, have a significant relationship with the risk of default. Fhatuwani and Karabo (2013), Agbemava et al. (2016), and Volkwein and Szelest (1995) also found the number of dependents significant factor of default. It was observed that holding employer type constant, the risk of default of a borrower would increase by 21.025% for any additional dependant the borrower takes on. A borrower with non-government employer type has 84.118% risk of default compared to a borrower with the government as the employer type when number of dependents was held constant.

Conclusion

Although Ghana's NPL ratio has not reached such crisis level as encountered by America during the Obama administration where they had to set up the Trouble Assets Recovery Program to buy NPLs to enable the banks to resume lending Herd-Clark and Murty (2013), nor that of Nigeria where the Central Bank had to set up the Asset Management Corporation of Nigeria (AMCON) which sought to buy NPLs on the books of the banks that were at risk Akpan (2013), it is important that policy makers and managers of banks adhere strictly to the laws regulating the financial and banking sector especially on the issues of NPLs. The government and monetary policy authorities must also collaborate actively with the financial sector players to monitor and improve the control of access to the limited investment funds and personal loans. The insight gained from the successful application of the Cox PH model to a small sample of personal loan data in this study makes the belief that statistical modelling is robust to help understand and correct processes and occurrences in the banking industry especially in the area of predicting which clients are more liable to default their loans based on their demographic characteristics. More research should be carried out in applying statistical modelling to banking processes. Further studies are recommended in comparing the robustness of the survival models with the Generalised Logistics Regression, a common model in predicting loan default in the Ghanaian banking sector.

REFERENCES

Aboagye, A. Q. (2020). Ghanaian banking crisis of 2017-2019 and related party transactions. African Journal of Management Research, 27(1), 2-19.

Adams, R. B. & Mehran, H. (2003). Is corporate governance different for bank holding companies?

Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal, 12(1).

Akpan, A. U. (2013). Relevance of the Asset Management Corporation of Nigeria (AMCON) to the non-performing loans of deposit money banks. International Journal of Economics, Business and Finance, 1(8), 249-261.

Alves, K. Kalatzis, A. & Matias, A. B. (2009). Survival analysis of private banks in

brazil. Technical report, EcoMod

Allen, L. N. & Rose, L. C. (2006). Financial survival analysis of defaulted debtors. Journal of the Operational Research Society, 57(6), 630-636.

Annesi, I., Moreau, T., & Lellouch, J. (1989). Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Statistics in medicine, 8(12), 1515-1521.

Arku, D. (2013). Delinquency and Default Risk Modeling of Microfinance in Ghana. Ph.D. thesis, University of Ghana

Asafo, S. (2018). The macro-economy and non-performing loans in Ghana: A BVAR approach. International Journal of Business and Economic Sciences Applied Research, 11(3).

Asantey, J. & Tengey, S. (2014). An empirical study on the effect of bad loans on

banks' lending potential and financial performance: The case of SMEs lending in

ghana. IMPACT: International Journal of Research in Business Management

(IMPACT: IJRBM), 2(11):1-12.

Banasik, J. Crook, J. N. & Thomas, L. C. (1999). Not if but when will borrowers

default. The Journal of the Operational Research Society, 50(12):1185-1190.

Boachie, C. (2016). Causes of non-performing loans in the banking industry. GRAPHIC ONLINE business news; Retrieved on February 02, 2022 from https://www.graphic.com.gh/business/business-news/causes-of-non-performing-loans-in-the-banking-industry.html

Boateng, E. Y., & Oduro, F. T. (2018). Predicting microfinance credit default: a study of Nsoatreman rural bank, Ghana. Journal of Advances in Mathematics and Computer Science, 26(1), 1-9.

CEIC DATA. (2021). Ghana Non-Performing Loans Ratio. Retrieved on October 20, 2021 from https://www.ceicdata.com/en/indicator/ghana/non-performing-loans-ratio#:~:text=in%20Feb%202021%3F-,Ghana%20Non%20Performing%20Loans%20Ratio%20stood%20at%2015.3%20%25%20in%20Feb,table%20below%20for%20more%20data

Cox, D. R. (1975). Partial likelihood. Biometrika, 62(2):269-276.

Cucinelli, D. (2015). The impact of non-performing loans on bank lending behavior: evidence from the Italian banking sector. Eurasian Journal of Business and Economics, 8(16), 59-71.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K., & Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. Statistics in medicine, 9(12), 1501-1515.

Fhatuwani, N. & Karabo, M. (2013). Application of survival models to analyse default Rates on bank loans. 2013 Convention of Actuarial Society of South Africa

Green, M. S., & Symons, M. J. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. Journal of chronic diseases, 36(10), 715-723.

Herd-Clark, D. & Murty, K. S. (2013). The Troubled Asset Relief Program (TARP): What has it accomplished in the Obama Era? Race, Gender & Class, 130-146.

Im, J.-K. Apley, D. W. Qi, C. & Shan, X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. Journal of the Operational Research Society, 63(3), 306-321.

Ingram, D. D., & Kleinman, J. C. (1989). Empirical comparisons of proportional hazards and logistic regression models. Statistics in medicine, 8(5), 525-538.

International Monetary Fund, IMF (2005). The Treatment of Nonperforming Loans. Eighteenth Meeting of the IMF Committee on Balance of Payments Statistics. Washington, D.C., June 27–July 1, 2005. Retrieved from https://www.imf.org/external/pubs/ft/bop/2005/18.htm

Nikolopoulos, K. I., & Tsalas, A. I. (2017). Non-performing loans: A review of the literature and the international experience. Non-performing loans and resolving private sector insolvency, 47-68.

Kleinbaum, D. G. & Klein, M. (2010). Survival analysis (Vol. 3). New York: Springer.

Larivi`ere, B. & Van den Poel, D. (2004). Investigating the role of product features in

preventing customer churn, by using survival analysis and choice modelling: The

case of financial services. Expert Systems with Applications, 27(2):277-285

Madormo, F. Mecatti, F. & Figini, S. (2013). Survival models for credit risk

estimation. In Advances in Latent Variables-Methods, Models and Applications.

Malik, M. & Thomas, L. C. (2010). Modelling credit risk of a portfolio of consumer

loans. The Journal of the Operational Research Society, 61(3):411-420

Murray, J. (2011). Default on a loan. United States Business Law and Taxes Guide.

Osei, A. A. Yusheng, K. Caesar, A. E. Tawiah, V. K. & Angelina, T. K. (2019). Collapse of big banks in Ghana: Lessons on its corporate governance. International Institute for Science, Technology and Education.

Pearson, R., & Greef, M. (2006). Causes of Default among Housing Micro Loan Clients . South Africa: FinMark

Trust Rural Housing Loan Fund,

National Housing Finance Corporation and Development Bank of Southern Africa

Rahardja, D. & Wu, H. (2018). Statistical methodological review for time-to-event data. Journal of Statistics and Management Systems, 21(1), 189-199.

Stepanova, M. & Thomas, L. C. (2001). PHAB scores: Proportional hazards analysis

behavioural scores. The Journal of the Operational Research Society,

52(9):1007-1016.

Torku, K. & Laryea, E. (2021). Corporate governance and bank failure: Ghana's 2018 banking sector crisis. Journal of Sustainable Finance & Investment, 1-21.

Volkwein, J. F., & Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. Research in higher education, 36(1), 41-72.

Whalen, G. (1991). A proportional hazards model of bank failure: an examination of its

usefulness as an early warning tool. Economic Review-Federal Reserve Bank of

Cleveland, 27(1):21.

Whited, T. M. Wu, Y. & Xiao, K. (2021). Low interest rates and risk incentives for banks with market power. Journal of Monetary Economics, 121, 155-174.

World Bank, International Monetary Fund, Financial Soundness Indicators (2021). Bank non-performing loans to total gross loans (%). Retrieved on August 21, 2021 from https://data.worldbank.org/indicator/FB.AST.NPER.ZS

Zopounidis, C. Mavri, M. & Ioannou, G. (2008). Customer switching behaviour in Greek banking services using survival analysis, Managerial Finance, 34(3):186-197.