Determinants of Personal Loan Default and Performance of the Proportional Hazards Model with that ofaRandom Survival Forests Models

Loku Pathiranage Himali (PhD Candidate)

Department of Statistics, School of Mathematics, Northwest University,
Xi'an, Shaanxi, China.

IJASR 2020
VOLUME 3
ISSUE 4 JULY – AUGUST **ISSN: 2581-7876**

Abstract – Loan Default is the failure of an applicant to fulfill his/her obligation with respect to the repayment of loans. The purpose of this study was to investigate the determinants of personal loan default and performance of the proportional hazards model with that of a random survival forests models. Primary data collected through questionnaires from 1500 customers who take the personal loan from a major Sri Lankan financial institution. The binary logistic regression model, proportional hazards model and random survival forest models was used as major analytical tools. Study found that, customer-related factors highly affected to the personal loan default such as occupation, monthly income and purpose of loan. The Random Survival Forest model considered monthly Income, occupation, the purpose of loan, and the amount of loan are important. The Cox Proportional Hazard model additionally considered other liabilities and frequency paid as important.There is need for the government to reduce the strains to the general economy in order not only to facilitate economic growth but also to enhance the minimization of the customer-related factors that precipitate loan default.

Keywords: Binary logistic regression model, Cox Proportional Hazard model, Loan Default, Random Survival Forest model

Introduction

Every business has a credit relationship with a financial institution, especially banks. Some rely on periodic short term loans to finance temporary working capital needs. Others primarily use long-term loans to finance capital expenditure, new acquisitions or permanent increases in capital. Regardless of the type of loan, all credit request mandate a systematic analysis of the borrower's ability to repay as at when due. Since most bank loans are important source of external finance.

The issue of loan default has become an issue in the financial circle all over the world. Financial expert is still researching various ways of addressing this problem. The consensus among these experts is that no one method stands out, the choices independent on other factors such as economic stability and effectiveness and dependability of the national data base.

The idea of using survival analysis techniques for constructing credit risk models is not new. Narain was one the first authors who used survival analysis methods for credit scoring. He analysed a data set of 1242 applicants accepted for a 24-month loan between mid-1986 and mid-1988. The data was analysed using the Kaplan-Meier method and by fitting exponential regression models. It was shown that the results obtained are encouraging and reasonable.

Many authors followed the example of Narain (1992) and started to use more advanced methods as compared to the parametric accelerated failure time survival methods used in this first work.

Stepanova and Thomas perform behavioural scoring using proportional hazards analysismodels. The authors conclude by saying that the proportional hazards analysis scores are useful as indicators of both risk and profit.

Survival analysis mainlyconcentrates on two majorsections of information: that is whether or not a participant suffers the event of interest during the study period and the follow up time for each individual being followed. Many researchers consider survival data analysis to be merely the application of two conventional statistical methods to a special type of problem: parametric if the distribution of survival times is known to be normal and nonparametric if the distribution is unknown. This assumption would be true if the survival times of all the subjects were exact and known; however, some survival times are not. Further, the survival distribution is often skewed, or far from being normal. Thus, there is a need for new statistical techniques.

Under the semi-parametric category, Cox model is the most commonly used regression analysis approach for survival data and it differs significantly from other methods since it is built on the proportional hazards assumption and employs partial likelihoodfor parameter estimation. In addition, several useful variants of the basic Cox model, such as penalized Cox models, CoxBoost algorithm and Time-Dependent Cox model are also proposed in the literature.

One of the most important developments is due to a special feature of survival data in the life sciences that occurs when some subjects in the study have not experienced the event of interest at the end of the study or time of analysis. Recently many machine learning algorithms are adapted to effectively handle survival data and tackle other challenging problems that arise in real-world data.

Machine learning techniques as a set of methods that canautomatically detect patterns in data, and then use the uncovered patterns to predict futuredata, or to perform other kinds of decision making under uncertainty. Machine learning algorithms, such as survival trees, Bayesian methods, neural networks andsupport vector machines, which have become more popular in the recent years.

This paper investigates theperformance of the proportional hazards model with that of a random survival forests models by studying when customers tend to default or pay off their loan early.

Literature Review

The objective of this section is to gather existing knowledge relating to the research topic and to identify the gap in the literature to which this research addresses.

In accordance with the words of Osayemeh, (1998)" Lending decision" very much like investment decisions are generally full of risk but the ability to banks to thoroughly assess and analyse such risk will lead to qualitative and more pragmatic decision. Lack of adequate knowledge of the loan sector could result to worst debts and moreover, the success of most loan depend on the perfect assessment of the customer's character.

Personal Loan:

This loan is normally for durable customer goods, repair and decoration, rents, school fees, house renovation, purchases of car or other personal consumption or investment items. Such loan is usually for small sums and is often referred to as petty lending. The banker must find out the purpose of the loan through discussion with the loan seeker, how much is needed by the customers and whether repayment cold be made without much stress on the

Customers or the bank either. The net monthly income is computed while the monthly expenses are deducted to assess the available disposal income from which the loan can be repaid. The reason for calling the "personal loans" are because they are normally granted enable individuals to purchase consumer durable goods. Interest is charged 'up front', that is, when the loan is granted. Thus, interest is added to total amount which is repayable. This means that the interest is charged on full amount over the total period. As a result, the customer is repaying the capital in regular instalments, which means that each month, the debt due to the bank is being reduced, yet the interest charged is on the full amount for the whole period which makers the interest charged penal. Once interest is fixed, the customer can be at a disadvantage of interest rates are reduced rationally. Alternatively, if during the life of the loan, interest rates are increases.

Problems of Loan Default

Loans are classified as problem credits when they cannot be repaid. Problem loans and losses essentially reflects the difficult risk inherent in a borrower's ability and willingness to repay all obligations. The lending process by its nature is imperfect. Credit analysis may be incomplete or based on faulty data. If management concentrates solely on minimizing losses, a bank will make virtually no loans: profit will shrink and the legitimate credit needs of customers will not be met. Lenders cannot completely eliminate risks, so more loan losses are expected. The objective is to manage losses well so that the bank can meet its risk and returns targets (Orjih, 2002). Loan default arises as a result of debts due to creditors but for some inherent weakness, the full or partial recovery is considered impossible. From banker's point of view, loan defaults are components of accounts receivable to the customers of the bank.

Causes of Loan Default

Incidences of bad doubtful debts in banks may be as result of inability to monitor and recover their loans due to lack of swell-articulated and professionally knowledgeable to appraise projects properly. Consequently, most loans granted by them becomes irrecoverable.Changes in economic policies may affect the operations of a borrowing firm and consequently the ability to honour loan obligations.Customer failure to disclose vital information during the application process leads to occurrence of loan default (Brownbridge, 1998).

Diversion of funds occurs when the funds borrowed have to be used only for a particular or the purpose it was intended (Ashiq, 2003). Nevertheless, more often than not, such funds are not used for the primary purpose they were intended for and as such, many projects become halfway done, in such case the funds were meant for income generating project but the borrower decides to divert into a different project thus leading to loan overdue.

Roche (2003) explains that loss of unemployment can be brought about by different circumstances such as retrenchment, retirement, sacking or firing due to events such as strikes and many others factors. When such a thing takes place and the individual affected had taken a loan from bank, then such a lending institution is likely to lose money hence leads to loan default.

Under certain circumstances of illness, the borrower instead of repaying the loan he/she uses the funds for medical expenses instead of the intended purpose. This can be seen in the illness such as HIV Aids, Cancer which can be very expensive to treat. Due to illness, such borrowers will find it difficult in honouring their loan obligation (Lehman and Neuberger, 1998). In many cases where the borrower is terminally ill or dies, the borrower may end up not repaying the loan in good time or not even repay at all. This kind of problem could be more pronounced in cases where the borrower is either an individual or a principal partner of a company. The general health of the borrower should be taken into consideration and health of a close relative (Nzambi, 2010).

Some borrowers don't clearly understand the purpose of bank loans. They sometimes regard bank loans as windfall receipts and very much unwilling to refund them.

Theoretical Review

This part of study includes some important theories that explain the effects of bank-and borrower's specific factors on personal loan default.

Adverse selection theory

Pagano and Jappelli (1993) shows that information sharing reduces adverse selection by improving banks information on credit applicants. The theory of asymmetric information states that it may be difficult to distinguish creditworthy from bad borrowers (Richard, 2011) which may result into adverse selection and moral hazards problems. The theory explains that in the market, the party that possesses more information on a specific item to be transacted (in this case the borrower) is in a position to negotiate optimal terms for the transaction than the other party (in this case, the lender) (Richard, 2011). The party that knows less about the same specific item to be transacted is therefore in a position of making either right or wrong decision concerning the transaction. Adverse selection and moral hazards have led to significant accumulation of loans default in banks (Bester, 1994; Bofondi and Gobbi, 2003).

Moral hazard theory

The moral hazard problem implies that a borrower has the incentive to default unless there are consequences for his future applications for credit. This result from the difficulty lenders have in assessing the level of wealth borrowers will have accumulated by the date on which the debt must be repaid, and not at the moment of application. If lenders cannot assess the borrowers' wealth, the latter will be tempted to default on the borrowing. Forestalling this, lenders will increase rates, leading eventually to the breakdown of the market (Alary &Goller ,2001).

Methods

The main question of interest is to identify the determinants of personal loan default. To examine that binary logistic regression analysis is applied in this study. Given the

Dichotomous attribute of the dependent variable, binary logistic regression is suitable over other regression

$$Logit(0,1) = \beta_0 + \beta_i^{Age} + \beta_j^{Gender} + \beta_k^{MaritalStatus} + \beta_l^{Education} + \beta_m^{Income} + \beta_n^{Occupation} + \beta_p^{Purpose} + \beta_q^{Amount} + \beta_r^{Liabilities} + \beta_s^{Dependents} + \beta_t^{Frequencypaid} + \varepsilon i$$

Techniques. The logistic regression model can be defined as,

Where:

- Logit (0, 1): Is the dependent variable, 1 is for loan defaultand 0 otherwise.
- β_0 : Is the constant
- εi : The residual error.

Logistic regression model is a form of regression where the outcome variable is binary or dichotomous and the independents are continuous variables, categorical variables, or both. (Rodrguez (2007). Thus, the purpose of Logistic Regression is to ascertain whether a certain phenomenon happens or not. To do this, it defines a dependent variable which will adopt the value "I" if the event happens and "0" if it does not.The value "0" represents by the repayment performance - default and value "I" represent the repayment performance - on time repayment. Further according to Tranmer and Elliot, proportions and probabilities are different from continuous variables in a number of ways. They are bounded by 0 and 1, whereas in theory continuous variables can take any value between plus or minus infinity. This means that it cannot assume normality for a proportion, and must recognize that proportions have a binomial distribution.

Proportional hazard model and random survival forests models investigate the performance. A fundamental characteristic of survival data is that survival times are frequently censored. Right-censored cases have survival times that are greater than some defined time point. The data used in this study contains right-censored survival times.

The Proportional Hazards Model

The Cox proportional-hazards model (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables.

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening at a particular point in time. This rate is commonly referred as

The hazard rate. Predictor variables are usually termed covariates in the survival-analysis literature.

The Cox model is expressed by the hazard function denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$h(t)= h_0(t) \times \exp(b_1x_1+b_2x_2+...+b_px_p)$

here,

- *t* represents the survival time.
- h(t) is the hazard function determined by a set of p covariates (x_1, x_2, ..., x_p)
- the coefficients (b_1, b_2, ..., bp) measure the impact of covariates.
- the term h_0 is called the baseline hazard. It corresponds to the value of the hazard if all the x_i is equal to zero. The ‘t’ in h(t) reminds us that the hazard may vary over time.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables ‘’x_i”, with the baseline hazard being an ‘intercept’ term that varies with time.

The quantities exp (b_i) are called hazard ratios. A value of "b_i" greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases.

Random Survival Forests

Random survival forests, an ensemble tree method for analysis of right-censored survival data. As is well known, constructing ensembles from base learners, such as trees, can substantially improve prediction performance. Recently it has been shown by Breiman (2001) that ensemble learning can be improved further by injecting randomization into the base learning process, an approach called random forests. Random survival forests methodology extends Breiman's random forests method.

Random Survival Forests Algorithm

Begin with a high-level description of the algorithm. Specific details follow:

1. Draw B bootstrap samples from the original data. Each bootstrap sample excludes on average 37% of the data, called out-of-bag data.

2. Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split using the candidate variable that maximizes survival difference between daughter nodes.

3. Grow the tree to full size under the constraint that a terminal node should have no less than $d_0 > 0$ unique deaths.

4. Calculate a Cumulative Hazard Function for each tree. Average to obtain the ensemble Cumulative Hazard Function.

5. using out of bag data, calculate prediction error for the ensemble Cumulative Hazard Function.

Variable selection in Survival Models

Among possible variable section methods, use Cox Regression and random survival forest, respectively. Firstly, using Cox regression within a backward stepwise method, the variables in each step are selected using the Akaike information criteria.

Secondly, using random survival forest, the out of bag based C-index is obtained by dropping out of bag cases down them in-bag survival tree and then assigning a daughter node randomly as soon as a split for the predictor variable is encountered. For each of the predictor variables, out of bag prediction error are obtained by subtracting the out of bag based C-index from one. A variable importance measure is then defined as the difference between the original out of bag prediction error and the new out of bag prediction error. Predictor variables having a large variable importance measure are considered to have greater prognostic capacities, whereas the variables with zero or negative variable importance measure can be dropped from the original model, as they add nothing to its predictive ability.

Descriptive Analysis

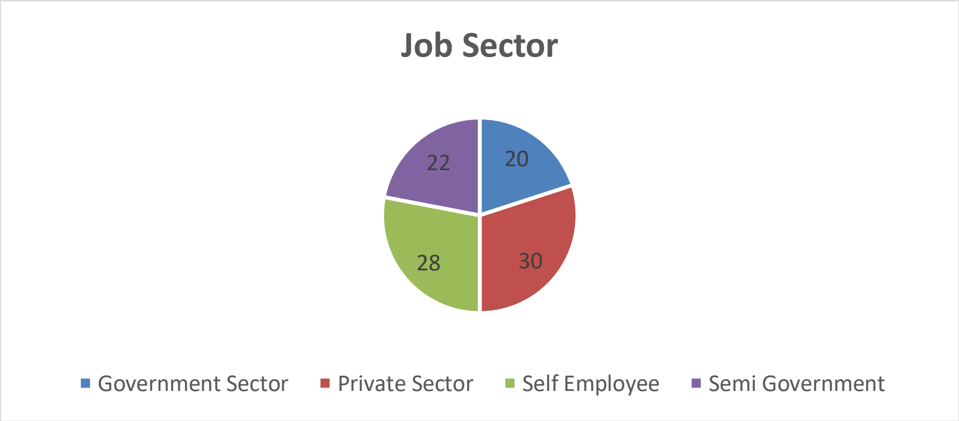

Figure 01: Job SectorWiseSample Composition

According to figure 01, It was clearly visible that the majority of the customers engage in private sector and self-employers. Least number of employers engages in government sector.

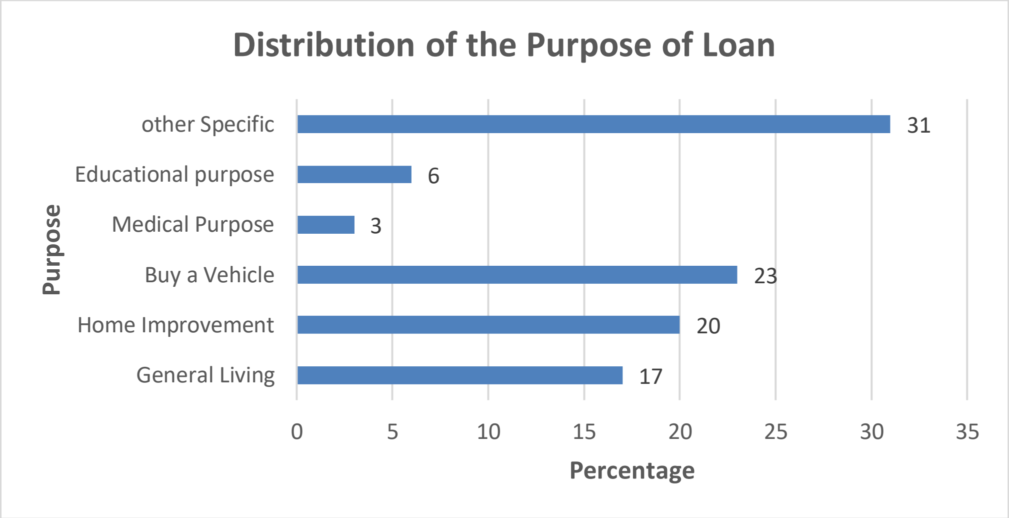

Figure 02: Distribution of the Purpose of Loan

Figure 02 illustrate that; majority of the people have got personal loan for otherspecific reasons; loans for weddings, pay other liabilities, vehicle repair, mixed purchases, kitchen units etc. least number customers got loans from the bank for medical purposes.

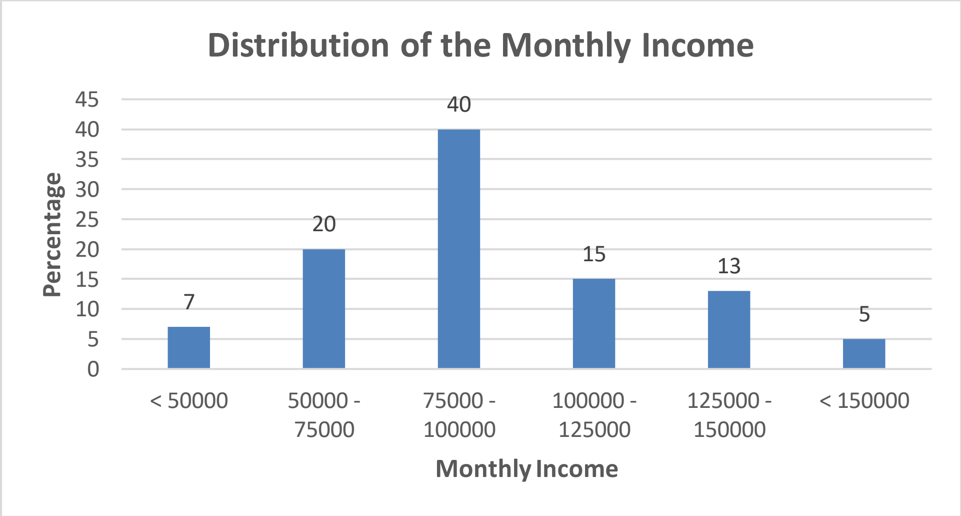

Figure 03: Income Distribution of the Respondents

Figure 3 clearly visible that the majority of the customers belongs to Rs.75000-100000 group. Minority belongs to > 150000 group. More than half of the customer's monthly income is less than or equal to 100000 rupees.

Empirical Evaluation

Test the Independence of the Variables

It is essential to test whether there is a relationship between the dependent variable and independent variables. For that chi-square test was used.

Hypothesis

H1a: Personal loan default is depending on the age of the customer.

H1b: Personal loan default is depending on the gender of the customer.

H1c: Personal loan default is depending on marital status of the customer.

H1d: Personal loan default is depending on the educational level of the customer.

H1e: Personal loan default is depending onmonthly Income of the customer.

H1f: Personal loan default is depending on theoccupation of the customer.

H1g: Personal loan default is depending on thepurpose of the loan.

H1g: Personal loan default is depending on the amount of the loan.

H1g: Personal loan default is depending on the other liabilities.

H1g: Personal loan default is depending on the number of dependent children.

H1g: Personal loan default is depending on the frequency paid.

All the above hypotheses were tested using chi-square test statistic under 5% level of significant. Educational Level, monthly Income, occupation, purpose of loan, amount of loan, other liabilities, number of dependent children and frequency paid were significant.

Step Wise Selection Procedure

There are many predictors and the major interest of this report is to identify important predictors. That means, in this report it wishes to identify a small subset that relates significantly to the outcome. Therefore, stepwise selection procedure used to select a good model.

Binary Logistic Regression Analysis

In this study it had been consider about eight variables. Series of models had been fitted by adding and removing the variables in order to assess whether all eight variables are necessary for prediction or whether any could be dropped. Results are judged using AIC (Akaikes' Information Criterion), likelihood ratio test. Also, p-value (used 5% significant level) is used to find out whether the variables are significant or not.

As a first step, it considers about the null model. The fitted null model for this study is not significant. It concludes that the current model should be improved. After that all selected variables were added to the model.

In the initial model the variables; monthly Income, occupation, purpose of loan, other liabilities, amount of loan and frequency paidare most important determinant. Then a series of models fitted in order to assess whether all six variables are necessary for prediction or whether any could be dropped. For that, it should identify the most important determinant among these six significant determinants. The variable 'purpose of loan' is the most important one because that variable has the lowest AIC. Then the fitted model is,

$$Logit(\Pi_i) = \beta_0 + \beta_i^{purpose}$$

Keeping the variable purpose of loan in the model, the remaining five factors were added one by one to the model and checked the significance of additional variable. The results were judged using likelihood ratio test and Akaikes' information criterion. After the variable 'monthly Income' added to the current model it suggested that 'monthly Income' is needed to the model in addition to purpose of loan. Here, p-value of the variable purpose of loan is significant after adjusting for the variable monthly Income. The fitted model is shown in bellow.

$$Logit(\Pi_{ij}) = \beta_0 + \beta_i^{Purpose} + \beta_j^{Income}$$

When 'purpose of loan' and monthly Income' are already in the model, the significance of other variables was checked. Additional variable was checked using likelihood ratio and test whether those variables are necessary for prediction or any could be removed. Among significant models, variable 'amount of loan' showed the lowest AIC value when 'purpose of loan and monthly Income' were already included in the model.

$$Logit(\Pi_{ijk}) = \beta_0 + \beta_i^{Purpose} + \beta_j^{Income} + \beta_k^{Loanamount}$$

To investigate, further important variables which should be included in the fitted model and to exclude the unnecessary variables, it was tried to develop model when the variables 'purpose of loan, monthly Income', and 'amount of loan' exist already in the model. Then the variable 'frequency paid' was significant. The model with all those four variables is as follows.

$$Logit(\Pi_{ijkl}) = \beta_0 + \beta_i^{Purpose} + \beta_j^{Income} + \beta_k^{Loanamount} + \beta_l^{Frequencypaid}$$

All the main effects are considered and it can be concluding that, none of the models significant after the model with four variables. That means, the models with five variables, six variables, seven variables and eight variables are also not significant.

Then, check the significance of the models with interaction terms. According to those outcomes the models with interaction terms are not significant. Since, the associated P-values for each interaction terms are not significant. So, all interactions ignored from the model and considered the model except interactions.

As a final point, consider about the overall result, the final significant model can be shown as above and there are no any interaction terms with regarding to this model.

In this model there are 1500 binary observations and success is happening 995 times. The deviance of the resulting model is 0.9282. For that reason, it is possible to conclude that the model is good because, the value of the deviance is close to one. In addition to that, the P-values related to the parameters in the model are significant at 5% level.

Check the Performance of the Proportional Hazards Model with that of a Random Survival Forests Models

In this section, to apply the Cox Proportional Hazard and Random Survival Forest models, selected random sample of 1500 customers who take the personal loan from a major Sri Lankan financial institution. All customers are Sri Lankan borrowers whohad applied to the bank for a loan. The data set consisted of the application information of1500 personal loans, together with the repayment status for each month of the observationperiod of 24 months. The status variable indicated which loans were bad, paid off to term, paid off early, or still open. Samples with missing data werenot incorporated in the study, to facilitate the demonstration of methods. Eleven covariates were considered in the analysis namely Age (in years), Gender (male, female), Marital Status, Educational Level, monthly Income, occupation, purpose of loan, amount of loan, other liabilities, number of dependent children and frequency paid.

All the covariates selected for the analysis are important in loan default studies. In this study all loans that are paid off early are considered as failures and all other loans as censored whereas in the latter case all defaulted loans are failures and the remaining ones censored. The censoring times were assumed to be independentof the failure times. For evaluating the performance of the two models, the datasetwas split randomly in the ratio 2:1 into a training set and a validation set.

On application of the cox proportional hazard model using the backward variable selection mechanism monthly Income, occupation, purpose of loan, other liabilities, amount of loan and frequency paid are chosen as the most important variables. However, the accuracy with which the model estimates the hazard ratios depends the proportional hazards assumption, which was violated by this dataset, thereby indicating the need for a more generalized model structure. Thus, applied random survival forest model using log-rank splitting and log-rank score splitting, which ranked its covariates by level of out of bag importance, based on 1000 trees. The four most important covariates in both the Random Survival Forests approaches are monthly Income, occupation, purpose of loan, and amount of loan with a slightly different ranking. The bottom four covariates based on importance values are ranked similarly for both the random survival forest models. The top four predictors having maximum variable importance values were also selected by the Cox model. However other liabilities andfrequency paid selected by the Cox model, have very low variable importance values for both the random survival forest models and are therefore considered unimportant for prediction purposes.

Conclusion

The purpose of this paper is identify the determinants of personal loan default and test the performance of the proportional hazards model with that of a random survival forests models by studying when customers tend to default or pay off their loan early. It reviews two modelling approaches and compares them in terms of variable selection methods.

This survey found out that the customer related factors highly affected to the personal loan default. According to the results of binary logistic regression model, the variables monthly Income, frequency paid, purpose of loan, and amount of loan are significantly affected to personal loan default.

To test the performance, applies proportional hazards model and random survival forests models to a real personal loan dataset. The covariates selected for analysis fail to follow the proportional hazards assumption of the cox proportional hazard model. Thus, random survival forest applies to this dataset to deal with its complex structure and attain increased accuracy in predicting survival times.

The Random Survival Forest model was chosen as a competing method to the Cox Proportional Hazard model, it is a decision tree structured black box model having high prediction accuracy and an efficient variable selection mechanism. Being black box models Random Survival Forest model lack interpretative capacities. It cannot directly quantify the risks presented by individual covariates to the overall hazard like the Cox Proportional Hazard model does in terms of hazard ratios.

The primary motivation in using the personal loan dataset was that in this research it was more interested in identifying a statistical model that predicts overall survival effectively based on a set of covariates.

The Random Survival Forest model considered monthly Income, occupation, purpose of loan, and amount of loan are important. The Cox Proportional Hazard model additionally considered other liabilities andfrequency paid as important. There are very few studies which have estimated the determinants of loan default using Cox Proportional Hazard model and machine learning models.

References

1. Breiman, L. (2001). Random forests. Machine Learning 45 5–32
2. Cox DR (1972). Regression models and life tables. J R Statist Soc B 34: 187–220
3. Kevin P. Murphy (2012) Machine Learning A Probabilistic Perspective.
4. Stepanova M and Thomas LC (2001) PHAB scores: proportional hazards analysis behavioural scores. J Opl Res Soc 52(9):1007-1016.
5. Stepanova M and Thomas LC (2002). Survival analysis methods for personal loan data. Opl Res 50(2):277-289.
6. Narain B (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN, and Edelman DB (eds). Credit Scoring and Credit Control. Oxford University Press: pp 109-121.