American Journal of Theoretical and Applied Statistics
2015; 4(4): 247-253
Published online June 1, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150404.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Modelling of Credit Risk: Random Forests versus Cox Proportional Hazard Regression

Dyana Kwamboka Mageto, Samuel Musili Mwalili, Anthony Gichuhi Waititu

Jomo Kenyatta University of Agriculture and Technology, Department of Statistics and Actuarial Science, Nairobi, Kenya

Email address:
dkmageto@gmail.com (D. K. Mageto)

To cite this article:
Dyana Kwamboka Mageto, Samuel Musili Mwalili, Anthony Gichuhi Waititu. Modelling of Credit Risk: Random Forests versus Cox Proportional Hazard Regression. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 4, 2015, pp. 247-253.
doi: 10.11648/j.ajtas.20150404.13

Abstract: In survival analysis several regression modeling strategies can be applied to predict the risk of future events. Often, however, the default choice of analysis tends to rely on Cox regression modeling due to its convenience. Extensions of the random forest approach to survival analysis provide an alternative way to build a risk prediction model. This paper discusses the two approaches in reference to credit management and compares the impact and results of both methods. The Cox Proportional Hazard model displayed a better performance than that of Random Survival Forest when estimating credit risk.

Keywords: Credit Risk, Random Forests, Survival Models

1. Introduction

Credit scoring is one of the most important aspects in a business. It is a system that aids the decision maker on whether to grant a loan to an applicant or not (Thomas et al. 1999). Credit risk refers to the probability that a borrower will default on any type of debt by failing to make required payments (Basel, 2000). Traditionally this was done by using subjective judgment to assess the credit worth of corporate borrower. However, development of such a system was found to be very time-consuming, cumbersome and expensive.

In past records a great number of the world's largest banks have developed sophisticated systems to try and model the credit risk arising from a business (Wekesa, 2012). However despite the increase in knowledge some institutions fail to make full use of the information at hand. In Zimbabwe between the periods 2003 to 2004 a number of banks were forced to close down in what was termed the Zimbabwean Banking Crisis and the main cause being poor credit risk management (Njanike, 2009). The US 2008 Financial crisis was a very clear and painful illustration of the effects of an inadequate risk management system. In fact, at the end of 2008, the federal government pledged more money to bail out the financial industry than it spent on the korean war, the race to the moon, the vietnam war, Operation iraqi freedom and NASA's lifetime budget combined (Politico, 2008). Africa was also not spared as the rapid growth she had for long harbored was interrupted in 2009 by the crisis. In the beginning, many economists underestimated its likely impact in Africa. However by early 2009 it became evident that the crisis had profound effect throughout the continent. South Africa for one experienced “Sudden stops” of capital flows already in 2008.

The 2008 global financial crisis did not spare Kenya as well. Its impact was both direct and indirect. The indirect effect included the slowdown of tourism industry. Exports as well greatly reduced, which in turn had an effect on the foreign exchange earnings.

The 2008 Financial crisis was a wakeup call to all if not most Micro-Financial Institutions (MFIs). It is thus very important that they put measures in place to curb the credit crisis.

Credit Risk has thus become a subject of considerable research interest in banking and finance, and has also recently drawn attention to statistical researchers (Zhang, 2009).

A lot has been done in developing default models to deal with credit risk. In most circumstances the default choice of analysis tends to rely on Cox regression modeling due to its convenience.

The main aim of this paper is to introduce Random Survival Forests (RSF) as an alternative approach for modeling credit risk, and to compare it with that of Cox Proportional Hazard regression.

2. Review of Previous Research

A lot has been done in developing default models to deal with credit risk. In most circumstances the default choice of analysis tends to rely on Cox regression modeling due to its convenience.

The use of survival analysis for building time to default models was first introduced by Narain (1992) and was further developed by Thomas *et al*. (1999). In which Narain (1992) applied the accelerated life exponential model to a 24 months of loan data. He illustrated that the proposed model estimated the number of failures at each failure time. The author then built a scorecard using multiple regressions, showing that a better credit-granting decision could be made if the score was supported by the estimated survival times. Thomas et al. (1999) on the other hand compared performance of exponential, Weibull and Cox's nonparametric models with logistic regression and concluded that survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach.

Wekesa (2012) reviewed modeling of credit risk for personal loans using Product-Limit Estimator. The results demonstrated that there is no significant difference between male and female applicants in terms of their survival times and hazard rates. Creamer (2012) however took a different approach and compared Random Forests and Logistic regression while comparing their predictive ability on Latin American Banks. Where RSF model approach indicated that the most important variables that affected banks were size, number of efficient systems and number of deposits. The analysis also revealed that RSF approach had better predictive capacity in comparison to logistic regression. Zhou and Wang (2012) Used RSF approach on Loan data. They improved the original random forests approach by allocating weights to decision trees. The experiments finally concluded that the weighted approach in tree aggregation improve the overall accuracy and performance of the model.

It is evident that most researchers have result to Cox PH and Artificial Neural Networks (ANN) as a form of analysis not only on loan but also on other survival data. Very little has been done on usage of Random forest Approach. The main aim of this paper is to introduce Random Survival Forests (RFS) as an alternative approach for modeling credit risk, and to compare it with that of Cox Proportional Hazard regression

3. Methodology

3.1. Random Forests

A Random Forest (RF) is basically a non-parametric machine learning method that can be applied in survival prediction models. In survival settings, the predictor is an ensemble formed by combining the results of many survival trees Ulla, Hemant and Thomas (2012). According to Leo Breiman (1999) it is an ensemble method that uses random selection of variables and bootstrap samples.

3.1.1. Bootstrapping in Random Survival Forest

Randomization in RSF is brought about in 2 cases. In the first circumstance, a randomly selected bootstrap sample (approximately 67% of the original data) is used for growing the tree called the "in-bag data". Each sample excludes 37% of the data called Out-Of-Bag data (OOB). This selected sample can be viewed as the root of the tree. Secondly, the root is split into 2 daughter nodes by using a splitting rule on a randomly selected co-variant. The split is the best when survival difference between the daughter nodes is maximized as much as possible. Eventually, as the number of tree nodes increases with every split, and dissimilar cases become separated, each node in the tree becomes homogeneous and is populated by cases with similar survival. The tree reaches a saturation point when a terminal node (the most extreme node in a saturated tree) has at least 1 death with unique survival times.

3.1.2.Developing the Random Survival Forest Model

Firstly the conditional cumulative hazard function is estimated using the Nelson-Aalen estimator. For those subjects that are in the bootstrap sample or rather the "in-bag" data. For us to illustrate the risk prediction for the Random forests we will denote the bth survival $\tau_b(x)$ as the terminal node of subjects in the bth bootstrap sample where a subject with predictor values x ends up. It is vital to note that when the bootstrap samples are drawn with replacement some subjects from the original data set may occur a number of times. Therefore we denote c_{ib} as the number of times i occurs. In a case where the ith subject is not in the bootstrap sample then $c_{ib} = 0$

We also introduce a counting notation Andersen, Borgan, Gill, and Keiding (1993).

$$N_i(s) = (T_i \leq s, \Delta_i = 1) \tag{1}$$

$$Y_i(s) = (T_i > s) \tag{2}$$

$$N_b^*(s,x) = \sum_{i=1}^{N} c_{ib}\big(X_i \epsilon \tau_b(x)\big) N_i(s) \tag{3}$$

$$Y_b^*(s,x) = \sum_{i=1}^{N} c_{ib}\big(X_i \epsilon \tau_b(x)\big) Y_i(s) \tag{4}$$

In RSF the ensemble is then constructed by aggregating tree based Nelson-Aalen estimators. In other words in each terminal node the CHF is estimated using the subjects that are in the bootstrap sample while using the Nelson-Aalen estimators Ishwaran (2008).

$$H_b(t|x_i) = \int_0^t \frac{N_b^*(ds,x)}{Y_b^*(s,x)} \tag{5}$$

The survival prediction from the random survival forest at x is then obtained as;

$$\hat{S}^{RSF}(t|x) = exp\left(-\frac{1}{B}\sum_{b=1}^{B} \widehat{H}_b(t|x_i)\right) \tag{6}$$

3.2. Cox Proportional Hazard Regression

The Cox PH model is the most generally used regression model this is due to the fact that it is not based on any assumptions concerning the nature or shape of the particular

survival distribution. In Cox Regression the CHF is dependent on the vector of predictor variables.

$$X_i = (X_i^1, \dots, X_i^K) \quad (7)$$

The Cox model can then be written as:

$$\Lambda(t|X_i) = \Lambda_0(t)\exp(\beta^T X_i) \quad (8)$$

Here Λ_0 describes the baseline hazard function, in our case the risk of a client defaulting payment. While the parameter $\beta = (\beta_1, \beta_2, \dots, \beta_K)$ is the vector of regression coefficients. They describe how the hazard varies in response to the models co-variants. The survival Predictor values of x are then obtained by:

$$\hat{S}^{cox}(t|x) = \exp\left(\hat{\Lambda}_0(t)\exp[\hat{\beta}^T X]\right) \quad (9)$$

In this study the Cox model will be built in R statistical package (Version 3.1.2). We will use the model to check which co-variants are significant in Credit default analysis.

3.3. Performance Measure

The two models under studied will be compared on basis of their predictive ability. In this study error will be measured by Harrell's concordance index (Harrell et al., 1982). Unlike other measures of survival performance, Harrell's C-index does not depend on choosing a fixed time for evaluation of the model and specifically takes into account censoring of individuals (May et al., 2004). According to Kattan et al. (1998) the method has quickly become quite popular in the literature as a means for assessing prediction performance in survival analysis.

The error rate is Error = 1 − C. Note that 0≤Error≤1 and that Error = 0.5 corresponds to a procedure doing no better than random guessing, whereas Error = 0 indicates perfect accuracy.

4. Data Exploration

4.1. Data Structure

The data used in this experiment was secondary data. It was obtained from leading commercial banks in Kenya. The loan applicants in the study were randomly picked from the banks database comprising of 70 branches. The Sample obtained was based on a portfolio of personal loans whose maturity was 45 months. The study thus included loans taken from the month of January, 2004 to September 2008. The sample obtained included 250 male applicants and 250 female applicants.

4.2. Variable Description

The variables in the account are to be measured from the month it was opened until the account becomes 'bad' implying it is closed or until the end of observation. The account is considered bad if payment is not made for two consecutive months in accordance to the industry practice. If the account is does not miss two payments and is closed or survives beyond the observation period, it is considered to be censored. The study will also assume that those who made early payment or settlement were censored.

The variables under study are enlisted below,

Table 1. *Variables Used.*

Variable	Measurement
Marital Status	Married, Not Married
Gender	Male, Female
Age	Varied
Status	Default, Non Default
Time of Payment	Varied
Employment	Employed, Unemployed
Homeownership	With Home, Without Home
Education Level	Secondary and above, Below secondary

5. Results and Discussion

5.1. Data Presentation

The dominant Characters in this study were, the married, the Unemployed, those without homes and also not having studied beyond secondary school. As for status most of the applicants Do not default. This can be illustrated in the Table below.

Table 2. *Summary of the Data.*

Marital Status	Sex	Employment	Home Ownership	Education Level
Married: 300	Male: 250	Employed: 201	Home: 92	Post Secondary: 48
Unmarried: 200	Female: 250	Unemployed: 299	No Home: 408	Secondary or Below:352

As for age the youngest applicant was 22yrs while the oldest was 55. The shortest dated loan payment was 12 months and the highest 36 months.

5.2. Random Forest Model

The random Survival Forest package used in this study produces an ensemble estimate for the cumulative hazard function. This is a machine learning algorithm consisting of many trees used in classification and analysis. In our study we will only focus on applications of this model that are relevant for our analysis.

First of the basic composition of the model is illustrated in the table bellow

Table 3. *Random Forest Model results.*

Sample size	**500**
Number of deaths	108
Number of trees	2000
Minimum terminal node size	3
Average no. of terminal nodes	76.083
No. of variables tried at each split	3
Total no. of variables	7
Analysis	RSF
Family	Surv
Splitting rule	Logrank
Error rate	43.78%

From this we can observe that out of the 500 samples taken 108 defaulted payments. The family "surv" forest has built the model with 2000 trees with 3 variables ties at each split. In our study we use the default splitting criterion i.e. the logrank test statistic. The error rate on doing the performance evaluation the out-of-bag (OOB) estimates of the error rate was calculated. The "unbiased" estimates of error suggested that when the resulting model was applied the error was obtained as is smaller than 0.5 hence implying that we do not have enough evidence to conclude that the predictors are not important in predicting the probability of default. Hence suggesting it is fairly a good model.

5.2.1. Error Estimate Against Number of Trees

The figure below represents the OOB error estimates against the number of trees in the forest.

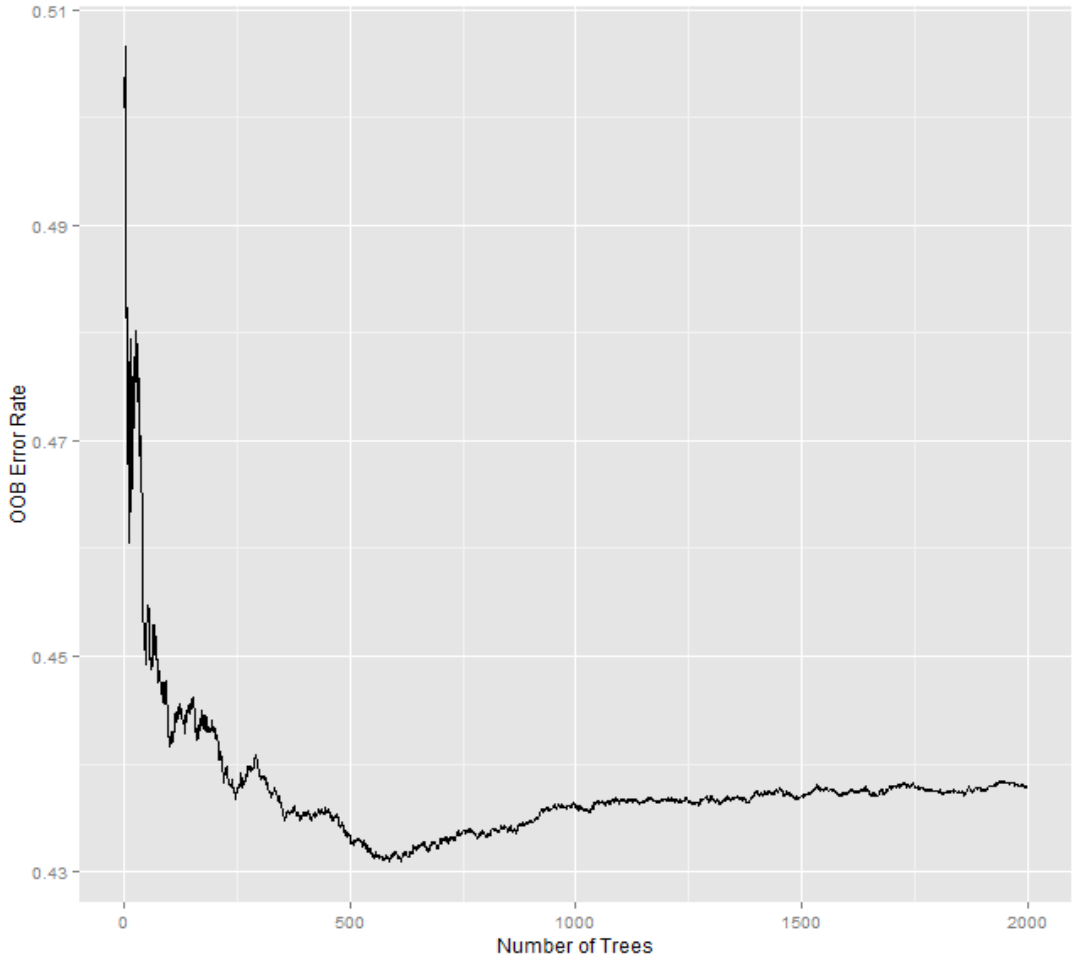

Figure 1. *Random Survival Forest OOB prediction error estimates against the number of trees.*

This figure illustrates that it takes about 1000 trees to construct the model. This plot is a good guide as to how many decision trees one requires when creating a random forest model. It is important to note that to ensure each variable is included in the forest it is better to create a large random survival forest tree.

5.2.2. Prediction of Survival Estimates

This is done by extracting the OOB estimates from the random forest. The figure below shows the predicted survival of our RSF model. Blue lines represent the observations who defaulted while the red lines represent those who did not default.

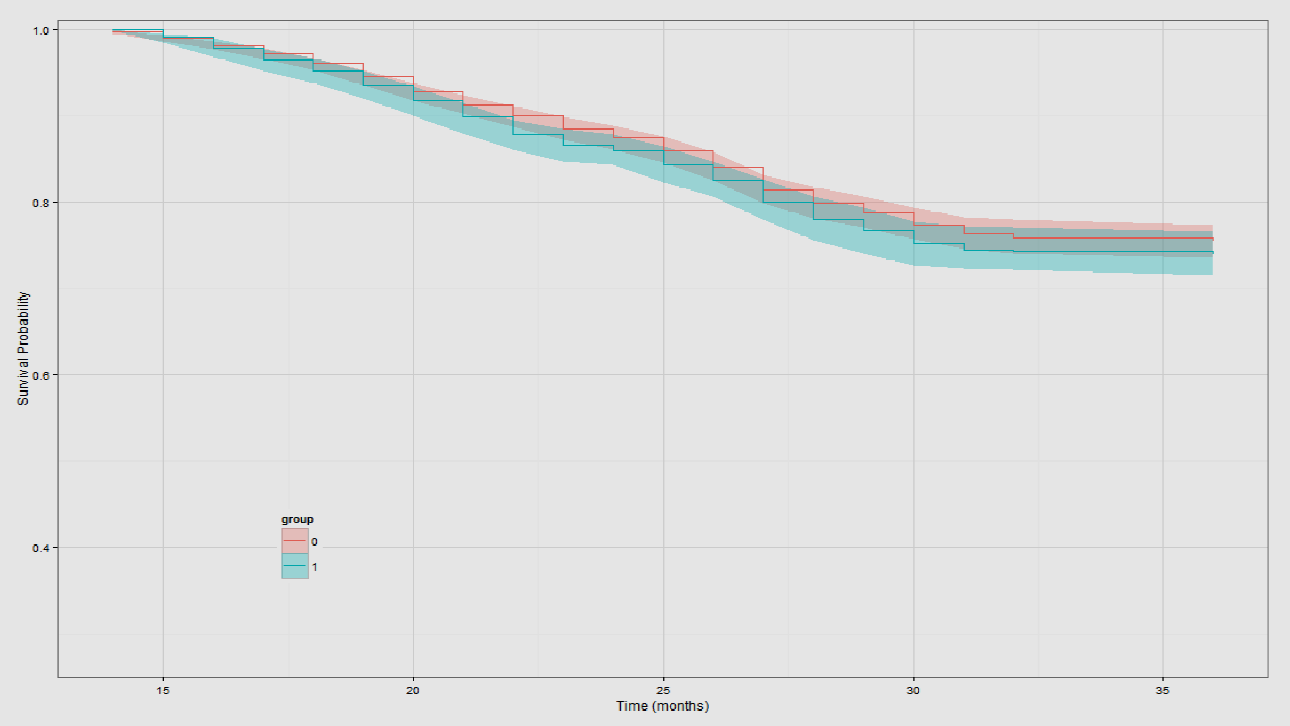

Figure 2. *Random forest predicted survival stratified by status.*

This figure also shows the median survival within a 95% confidence interval of status against time.

5.2.3. Variable Importance According to RSF

The important variables according to RSF were Marital Status, Employment, Home Ownership and Education level. While the least were Sex and age.

5.3. Cox Proportional Hazard Model

On carrying out an analysis of the Cox-PH Model time and status were regressed against the other variables, the following results were obtained.

Table 4. *Cox-PH model results.*

Variable	Coef	exp(coef)	Lower .95	upper .95
Marital	1.111953	3.040292	1.3738	6.728
Sex	-0.26114	0.770175	0.5238	1.132
Age	0.003961	1.003924	0.918	1.098
Employment	0.43173	1.53992	1.0237	2.317
Home	0.729073	2.073159	1.1317	3.798
Education	0.072468	1.075158	0.6999	1.652

Table above gives a portion of the analysis done on the variables. It is evident that the "coef" column gives the coefficients corresponding to each variable. For instance holding other co-variants constant, an additional year of age reduces the hazard of Default by a factor $e^{b_3} = 0.003916$ on average. The exponential coefficients in the second column of the output are multiplicative effects of the hazard. While the lower.95 and upper.95 are basically confidence intervals for each specific variables.

5.3.1. Variable Importance According to Cox-Model

The co-variants marital status, employment and Home Ownership are significant at 99% confidence interval with marital status being the most significant. On the other hand Education Level and age are the least important variables.

5.3.2. Error Estimate

The R-square for this model is given as 0.924 which is very close to 1 indicating that the model predicts the probability of default very well.

The likelihood-ratio, Wald, and score chi-square statistics at the bottom of the output were asymptotically equivalent test $H_0: \beta = 0$ that is that the variables are not important. In this study the statistics are close in argument, and thus implying we reject the hypothesis concluding that the variables are significant in the model.

The results discussed are visible bellow.

Table 5. *Results for Cox-PH test statistics.*

Likelihood ratio test	Wald test	Score (logrank) test
30.96	28.56	29.83
on 9 df	on 9df	on 9 df
p = 0.0003005	p = 0.0007681	p = 0.0004691

5.3.3. Predicted Survival Probability

It is often of interest to examine the distribution of predicted survival times. Whereby there is a view of the survival

probability according to each time (months). This is illustrated bellow.

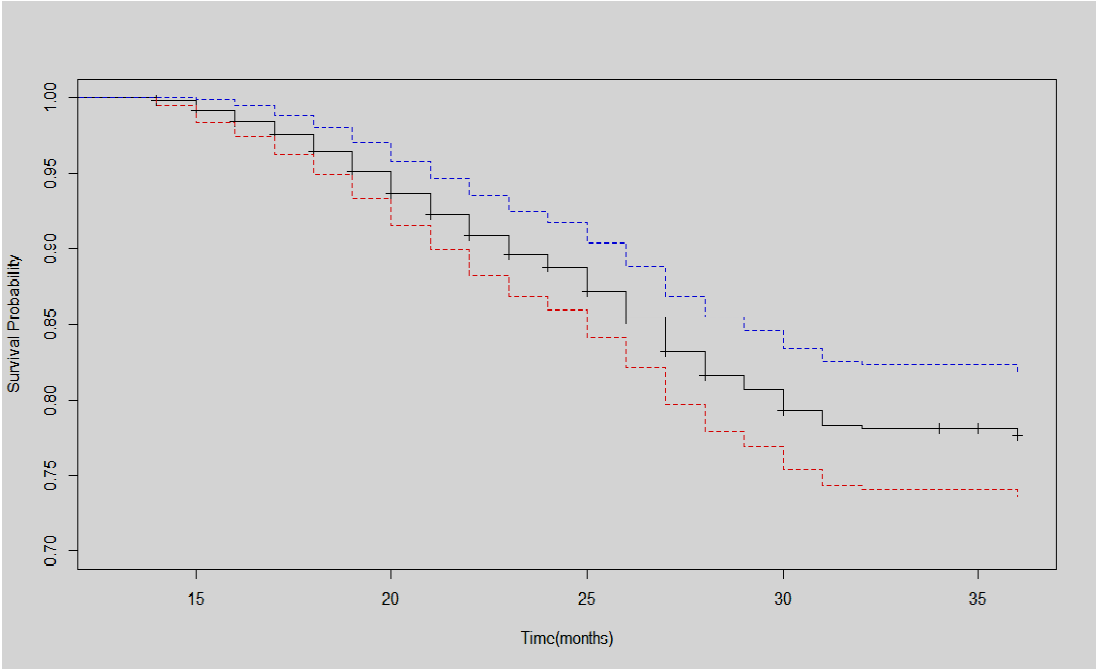

Figure 3. *Survival chart.*

6. Model Diagnostic

The two models used in this study were the Random Survival Forest model and the Cox Proportional Hazard Model. The section below looks at the performance evaluation of the two models. To measure the performance we used Harell's concordance index (C-index).

The C-index for RSF was obtained as 0.4378 while that of the Cox model obtained as 0.3376. From this it is evident to see that the Cox model has a lower C-index value than that of RSF. Hence according to Harell's concordance index the Cox model displays a better performance than that of RSF.

7. Conclusion and Recommendations

Cox-PH model was found to be a better model for predicting the probability of default as compared to RSF. In both models Marital status, Employment and Home ownership were found to be the common important variables. However the RSF model displayed Education Level as an important variable as well. It was also found that Sex, and Age do not affect were not important in predicting the probability of default.

We therefore recommend the use other methods to model credit risk like Accelerated failure-time models and Kaplan-Meier models to view how the models would behave.

References

[1] Anderssen, P. K. Borgann, Gill keiding N, (1993). Statistical Models Based on Counting Process. Springer series in statistics. New york.

[2] Basel (2000). Principles for the Management of Credit Risk, Basel Commettee. September 2000 1-30.

[3] Breiman, L. (1999). Using Adaptive Bagging to Debias Regression, Technocal report. 547, statistical Depertment UCB.

[4] Creamer, G. (2012). Using Random Forests and Logistic Regression for Performance Prediction of Latin American ADRS and Banks. Journal of Centrum Cathedra, 24-36

[5] Harell, F. E. Califf, R. M. Pryor, D. M. Lee, K. L. Rosati, R. A. Evaluating the Yield of Medical Tests. JAMA. 1982; 247: 2543-2587

[6] Katten, M. Hess, K. and Beck, J. (1998) Experiments to Determine whether Reccursive Partittioning (CART) or an Artificial Neural Network Overcomes Theoretical Limitations of Cox PH Regression, Computer Biomedical Research, 363-373. IBM, Commuter Survey (2014).

[7] May, M. Royston, P. Egger, M. Justice, A. C. and Sterne, J.A.C. (2004) Development and Validation of Prognosis Model for Survival Time Data: Application to Prognosis of HIV Positive Patients Treated with Anti-retroviral Therapy. Statistics Medicine, 23: 2373-2398.

[8] Narain, B. 1992. Survival analysis and the credit granting decision. L. C. Thomas, J. N. Crook, D. B. Edelman, eds.Credit Scoring and Credit Control. OUP, Oxford, U.K., 109–121.

[9] Njanike, K. (2009). The Impact of Effective Credit Risk Mangement on Bank Survival. Annals of the University of Petrosani Economics, 9(2), 173-184.

[10] Wekesa O. (2012), Modelling Credit Risk for Personal Loans Using Product-Limit Estimator. International Journal of Financial Research. (3) 22-32.

[11] Thomas, L. C. Banasik, J. N. Crook. (1999). Not if but when Loans Default. J.Operations Research Society. 50 1185-1190.

[12] Zhang, A. (2009). Statistical Methods in Credit Risk Modelling, University of Michigan, 3 4-27.

[13] Zhou, L. and Wang, H. (2012). Loan Default Prediction on Large Imbalanced Data Using Random Forests, Telkonmika Indonesian Journal of Electrical Engeneering. (10) 1519-1525.