BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1433–1440
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1433-1440

A COMPARISON OF COX PROPORTIONAL HAZARD AND RANDOM SURVIVAL FOREST MODELS IN PREDICTING CHURN OF THE TELECOMMUNICATION INDUSTRY CUSTOMER

Sitti Nurhaliza[1] , Kusman Sadik[2*], Asep Saefuddin[3]

[1,2,3] Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Jl. Raya Dramaga, Bogor City, 16680, Indonesia

Corresponding author's e-mail: [2]kusmans@apps.ipb.ac.id*

Abstract *The Cox Proportional Hazard model is a popular method to analyze right-censored survival data. This method is efficient to use if the proportional hazard assumption is fulfilled. This method does not provide an accurate conclusion if these assumptions are not fulfilled. The new innovative method with a non-parametric approach is now developing to predict the time until an event occurs based on machine learning techniques that can solve the limitation of CPH. The method is Random Survival Forest, which analyzes right-censored survival data without regard to any assumptions. This paper aims to compare the predictive quality of the two methods using the C-index value in predicting right-censored survival data on churn data of the telecommunication industry customers with 2P packages consisting of Internet and TV, which are taken from all customer databases in the Jabodetabek area. The results show that the median value of the C-index of the RSF model is 0.769, greater than the median C-index value of the CPH model of 0.689. So the prediction quality of the RSF model is better than the CPH model in predicting the churn of the telecommunications industry customer.*

Keywords: *C-index, churn, cox proportional hazard, random survival forest, right-censored, survival analysis.*

Article info:

Submitted: 9th August 2022 *Accepted: 3rd November 2022*

How to cite this article:
S. Nurhaliza, K. Sadik and A. Saefuddin, "A COMPARISON OF COX PROPORTIONAL HAZARD AND RANDOM SURVIVAL FOREST MODELS IN PREDICTING CHURN OF THE TELECOMMUNICATION INDUSTRY CUSTOMER", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1433-1440, Dec., 2022.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Copyright © 2022 Author(s)

1. INTRODUCTION

Survival analysis is a statistical procedure for analyzing data with a response variable: the time until an event occurs (time-to-event). The time of the incident is in the form of years, months, weeks, or days from the beginning of the observation until an event occurs. An event is every individual experience that is observed from an observation that is possible to occur [1]. In survival analysis, the term censored time is known, namely events or failures that occur when the individual is not being observed [2]. Research's most frequently censored time is the right time [3]. Right-censored time occurs at the end of the study the individual has not experienced an event, the individual withdraws from the study and cannot be observed further, or the individual experiences another event that causes the observation to be discontinued. Therefore, the survival data is incomplete when there is right-censored data due to random factors for each individual.

The Cox (Proportional Hazard CPH) model is the most popular method for analyzing right-censored survival. This method is a semi-parametric model used to identify predictor variables that significantly affect the response time until an event occurs [4]. This method is efficient if the assumption of Proportional Hazard (PH) is fulfilled. If this assumption is not fulfilled, then this method does not provide an accurate conclusion.

A new innovative method with a non-parametric model approach is now developing to predict the time until an event occurs based on machine learning techniques creating opportunities to overcome the limitations of the CPH model [5]. One of the machine learning methods, Random Survival Forest (RSF), is an ensemble method for analyzing right-censored survival data. The basic principle of RSF is to build several survival trees with the basic idea of recursively partitioning predictor variables as binary separators to form the same subject group based on time-to-event response variables. The advantages of using RSF are without considering any assumptions. Besides that, RSF is strong against outliers in predictor variables and is a useful tool in exploratory survival analysis with limited information [6].

Researchers in the survival analysis have widely used the RSF method. Ruyssinck et al. compared the RSF with the basic approach model and the Random Forest in predicting the number of patient beds based on the length of stay (LOS) in the ICU [7]. The results show that the RSF performance is better than the basic approach and random forest based on the absolute mean error. Nasejje et al. compared the CIF model with RSF on the right-censored simulation data [8]. The results show that CIF's performance is better than RSF's based on the Integrated Brier Score (IBS).

Ptak-Chmielewska & Matuszyk compared RSF and CPH in predicting the bankruptcy of Small and Medium Enterprises (SMEs) in Poland using the prediction accuracy of the *C-index* [9]. The results showed that RSF provided not only better results but also more stable ones than the semiparametric CPH model. Mageto got contradictory results in comparing CPH and RSF, the results showed the CPH model displayed a better performance than that of RSF when estimating credit risk [10]. The purpose of this study is to compare CPH and RSF in predicting the churn time of telecommunication industry customers based on the median C-index value. The median C-index value was obtained from the results of 20 repetitions. The model's prediction performance generated from 20 repetitions is expected to provide more consistent accuracy.

2. RESEARCH METHODS

This study will compare the CPH and RSF methods in predicting customer *churn* in the telecommunications industry based on the *C-index value* and identify the important variables in the best model. The procedure of analysis : (1) Pre-processing data, (2) Dividing data into 70% training data and 30% testing data, (3) Performing CPH model analysis, (4) Testing PH assumptions, (5) Performing RSF analysis, (6) evaluating the model by comparing CPH and RSF based on the C-index value with 20 repetitions, (7) Selecting the important variables based on positive Variable importance (VIMP) value.

2.1 Data

The data used in this research is customer data of the telecommunications industry, namely 2P packages consisting of Internet and TV, which are taken from all customer databases in the Jabodetabek area. When the customer was first registered is defined as the start time, and the last observation date (31 December 2019) is the end. During the observation period, customer time churn is recorded. Then, the predictor variable data used are gender (X1), age (X2), internet speed (X3), the total monthly bill (X4), internet data usage (X5), duration of watching TV (X6), and the number of TV channels watched (X7).

2.2 Cox Proportional Hazard Model

Cox Proportional Hazard Model (CPH), introduced by Cox (1972), is a semi-parametric model approach that is most commonly used in evaluating predictors with survival time [11]. The CPH mathematical model is

$$\boldsymbol{h(t|x) = h_0(t)\exp\,(\beta' x)} \tag{1}$$

where t is time, x is a vector of predictor variables, β is a vector of regression coefficient of predictor variables x, and $h_0(t)$ is a baseline hazard function, which is a hazard function of the response if the subject xi s equal to 0. The model describes the effect of the predictor on the response to the occurrence of a hazard. An important assumption on the CPH regression is that it has a constant hazard function proportion for each time. The Hypothesis of Proportional Hazard assumption is as follows:

H0: The Assumption of Proportional Hazard is fulfilled
H1: the assumption of proportional Hazard is not fulfilled

with rejection test H $_0$ if *the p-value* <0.05.

2.3 Random Survival Forest Algorithm

Random survival forest (RSF) was introduced by Ishwaran et al. as one of the ensemble tree methods for analyzing right-censored [12]. This method is a development method of the Random Forest method introduced by Breiman (2001). The implementation of the RSF algorithm is as follows [8], [13].

1. Taking as many as B bootstrap samples from the data by performing a return. Each bootstrap sample is used to form a survival tree. Each bootstrap sample is selected for about 37% of the data, which is called out-of-bag (OOB data).
2. For each terminal node on the tree, m predictor variables are randomly selected to be used as splitting variables.
3. Splitting a predictor variable by using the splitting log-rank rule. A node will be split based on a predictor variable that produces the greatest difference between the two survival functions of the derived node.
4. Repeating steps 2 and 3 until multiple trees are obtained, and the stopping rule criteria is each terminal node has at least $d_0 > 0$ a unique failure data.
5. Calculating the CHF value for each terminal node on each tree using the Nelson-Alaen estimator.

$$\boldsymbol{\widehat{H}_h(t) = \sum_{t_{l,h}<t} \frac{d_{l,h}}{r_{l,h}}} \tag{2}$$

where $t_{l,h}$is the occurrence time 1 of the sample on the derived node h, $d_{l,h}$is the number of occurrences on $t_{l,h}$, and $r_{l,h}$is the number of individuals risk on $t_{l,h}$.

6. Finding the CHF ensemble value by calculating the average of all trees in the *survival forest* to get the CHF bootstrap ensemble.

$$\boldsymbol{H_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*\,(t|x_i)} \tag{3}$$

where $H_b^*(t|x_i)$is CHF on the bootstrap node b.

7. Using OOB data to calculate the prediction error of the CHF ensemble.

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^{B} I_{i,b}} \quad (4)$$

$I_{i,b} = 1$if i is OOB for b,$I_{i,b} = 0$

The following is a flowchart of the RSF algorithm.

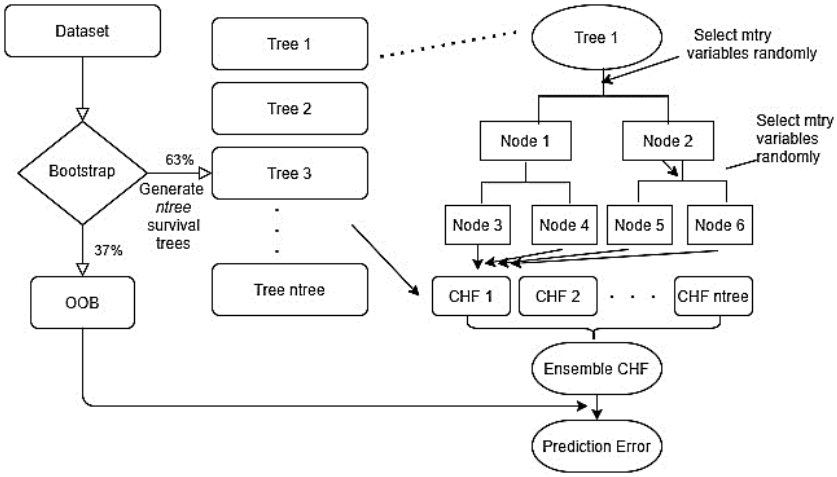

Figure 1 . Flowchart of RSF

2.4 C-Index

One of the metrics used in determining the model's predictive quality is Harrell's Concordance index (C-index). The C-index is related to the area under the ROC curve. The C-index predicts the probability that in selected pair of examples randomly, the failed sample (status=1) has the worst predictive outcome. The error rate is calculated as 1-(C-index). The error rate value is between 0 and 1, with 0.5 based on the procedure that is not better than a random guess, and a value of 0 represents perfect accuracy [14].

2.5 Variables Importance

The important variable selection aims to identify the importance of different variables in the probability of survival. VIMP is the difference in OOB prediction error before and after permutation [15], so a sizeable VIMP value indicates that specification error reduces prediction accuracy in the forest. VIMP close to zero indicates the variable does not contribute any value to prediction accuracy, and negative indicates prediction accuracy when there is an error in determining the variable. VIMP calculations were carried out across random forests. VIMP is calculated by comparing OOB data with a previously constructed survival tree. Therefore, VIMP measures the change in prediction error when the desired change is not available when building a new forest [6], [16].

3. RESULTS AND DISCUSSION

3.1. CPH Estimation Results and PH Assumption Test

The estimation results of the CPH model showed that the Gender variable, Age (30-39), Age (40-49), Age (50-59), Age (≥60), Speed 50 Mbps, Internet Data Usage, TV Watching Time, and The number of channels watched has a significant effect on the churn time of telecommunication industry customers. While the Age variable (20-29), Speed 20 Mbps, Speed 30 Mbps, Speed 40 Mbps, and Speed 100 Mbps had no significant effect. CPH Modeling has a weakness, namely the assumption of PH that must be fulfilled. The results of testing the PH assumption showed that the variables for the number of monthly bills (P-value=0.000) and TV viewing duration (P-value=0.000) do not fulfill the assumptions (P-value <0.05). CPH Model does not fulfill the assumptions, so other alternatives are needed to overcome it. Another alternative method used in this research is RSF.

3.2. RSF Model Results

Table 1 shows that from 713 samples (70% training data), 83 customers experienced churn during the research period. Tree formation was carried out with as many as 1000 survival trees on the RSF model by involving all predictor variables by bootstrapping. The goal is to get the RSF model by estimating all the original data. The result of Bootstrap data is known as out-of-bag or OOB. Based on the RSF model formed, an estimated error rate of 25.79% is obtained.

Table 1. The basic composition of the RSF Model

Sample size	713
Number of deaths	83
Number of trees	1000
Forest terminal node size	15
Average no. of terminal nodes	38,455
No. of variables tried at each split	3
Total no. of variables	15
Resampling used to grow trees	Swr
Resample size used to grow trees	713
Analysis	RSF
Family	Surv
Splitting rule	Logrank *random*
Number of random split points	4
Error rate	25,79%

Figure 2 shows the estimated value of the OOB error rate based on 1000 trees formed in the RSF model. The number of trees is less than 100 trees, and the estimated value of the OOB error rate is quite high and unstable. The number of trees from more than 100 to less than 500 trees, the estimated OOB error rate value tends to be stable and begins to decline when compared to the number of trees with less than 100 survival trees. The OOB error rate value tends to be stable and low at a minimum number of trees of 500 survival trees.

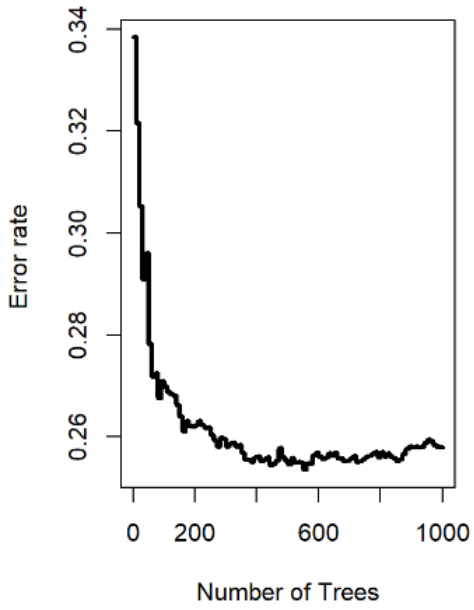

Figure 2. Error Rate Value

3.3. Prediction Quality Comparison

Figure 3 shows a boxplot of the results of the comparison of the C-index values of the CPH and RSF models. C -*index* median value of The RSF model of 0.769 is greater than the median C-index of the CPH model of 0.689. It is concluded that the predictive quality of the RSF model is better than the CPH model, so the RSF model is the best model for predicting customer churn in the telecommunications industry.

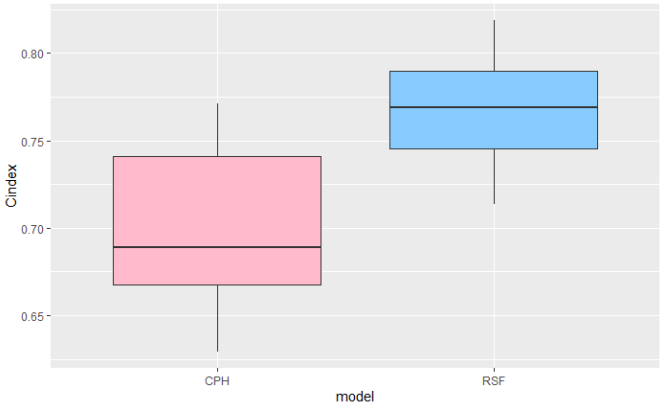

Figure 3. Comparison of C-index values

3.4. Selection of Important Variables

Figure 4 shows the predictor variable with the value of the positive VIMP. The most important are the TV Watching Duration, Number of Channels Watched, Internet Data Usage, Monthly Billing Amount, Age 20-29 years, Speed of 20 Mbps, and Age >60 years variables that create the RSF model.

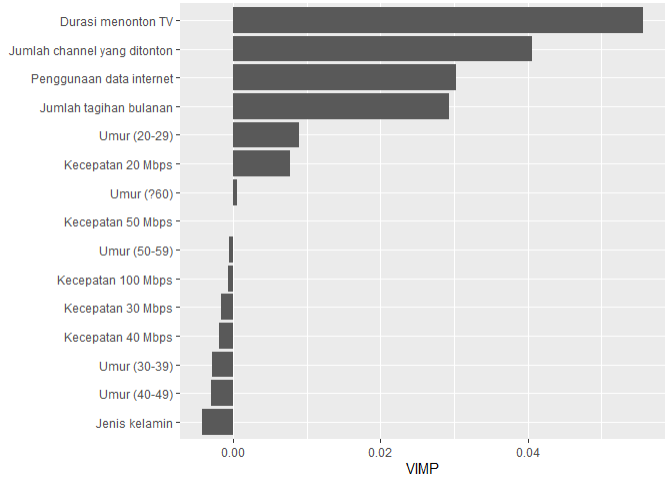

Figure 4. Important Variables

4. CONCLUSIONS

Survival analysis modeling on churn data of telecommunication industry customers by comparing CPH and RSF models. The assumption of PH on the CPH model has not been fulfilled, namely the variables of internet data usage and TV viewing duration, so an alternative is needed, namely the RSF model. The comparison of the prediction quality of the CPH and RSF models using the C-index obtained the median value of the C-index. The RSF model of 0.769 is greater than the CPH model's C-index median model of 0.689. It is concluded that the predictive quality of the RSF model is better than the CPH model, so the RSF model is the best model for predicting customer churn in the telecommunications industry. The important variables that build the RSF model are the duration of watching TV, the number of channels watched, the number of monthly bills, internet data usage, Age 20-29 years, speed of 20 Mbps, Age >= 60 years, and Age 50-59 years.

REFERENCES

[1] D. G. Kleinbaum and M. Klein, *Survival Analysis A Self-Learning Text*, Third Edit. 2012.

[2] J. Harlan, *Analsis Survival*. Depok: Gunadarma, 2017.

[3] H. Meiling and Y. Lin, "Nonparametric Inference for Right Censored Data Using Smoothing Splines Statistica Sinica Preprint No : SS-2017-0357 Title Nonparametric Inference for Right Censored Data Using Smoothing Splines Complete List of Authors Meiling Hao Yuanyuan Lin and Xingqi," no. February, 2019, doi: 10.5705/ss.202017.0357.

[4] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, "Evaluating Random Forests for Survival Analysis Using Prediction Error Curves," *J. Stat. Softw.*, vol. 50, no. 11, pp. 1–23, 2012, doi: 10.18637/jss.v050.i11.

[5] K. Afrin, G. Illangovan, S. S. Srivatsa, and S. T. S. Bukkapatnam, "Balanced Random Survival Forests for Extremely Unbalanced, Right Censored Data," no. April, 2018, [Online]. Available: http://arxiv.org/abs/1803.09177.

[6] D. K. Mageto, S. M. Mwalili, and A. G. Waitutu, "Modelling of Credit Risk: Random Forests versus Cox Proportional Hazard Regression," *Am. J. Theor. Appl. Stat.*, vol. 4, no. 4, p. 247, 2015, doi: 10.11648/j.ajtas.20150404.13.

[7] M. Saadati and A. Bagheri, "Comparison of Survival Forests in Analyzing First Birth Interval," *Jorjani Biomed. J.*, vol. 7, no. 3, pp. 11–23, 2019, doi: 10.29252/jorjanibiomedj.7.3.11.

[8] J. Ruyssinck *et al.*, "Random Survival Forests for Predicting the Bed Occupancy in the Intensive Care Unit," *Comput. Math. Methods Med.*, vol. 2016, 2016, doi: 10.1155/2016/7087053.

[9] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, "A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data," *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 1–18, 2017, doi: 10.1186/s12874-017-0383-8.

[10] A. Ptak-chmielewska and A. Matuszyk, "Application oF The Random Survival Forests Method in The Bankruptcy Prediction," vol. 1, no. 1, 2020, doi: 10.15611/aoe.2020.1.06.

[11] P. C. Austin, "Generating survival times to simulate Cox proportional hazards models with time-varying covariates," *Stat. Med.*, no. November 2011, 2012, doi: 10.1002/sim.5452.

[12] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008, doi: 10.1214/08-AOAS169.

[13] J. Wang, "Apply Machine Learning Approaches to Survival Data," Imperial College London, 2018.

[14] A. Schlossberg *et al.*, "Cox Proportional Hazard Regression," no. July, 2016.

[15] B. C. Jaeger, S. Welden, J. L. Speiser, K. Lenoir, and M. Segar, "A CCELERATED AND INTERPRETABLE OBLIQUE RANDOM," 2022.

[16] A. Hazewinkel, H. Gelderblom, and M. Fiocco, "Prediction models with survival data : a comparison between machine learning and the Cox proportional hazards model," no. Ml, 2022.