Jurnal Ilmiah Matematika dan Terapan
Volume 21 Number 2 December 2024 (Pages 182 - 192)
doi: https://doi.org/10.22487/2540766X.2024.v21.i2.17498

JIMT
Jurnal Ilmiah Matematika dan Terapan
ISSN : 2540 – 766X

COMPARISON OF RANDOM SURVIVAL FOREST AND FUZZY RANDOM SURVIVAL FOREST MODELS IN TELECOMMUNICATIONS INDUSTRY CUSTOMER DATA

Sitti Nurhaliza[1*], Andi Harismahyanti[2] and Alimatun Najiha[3]

[1,2,3]Study Program of Statistics, Department of Mathematics,
Faculty of Mathematics and Natural Sciences, Tadulako University
[1]sittinurhaliza1218@gmail.com
*(*corresponding Author)*

ABSTRACT

The telecommunications data is facing increasing competition, and customer churn is still a major challenge despite the implementation of advanced promotions and high-quality services. Churn refers to the discontinuation of services by customers, influenced by several factors that can be found through data modeling. This study compares two predictive models, Random Survival Forest (RSF) and Fuzzy Random Survival Forest (FRSF), for telecommunications data in predicting of customer churn time. The median value of the C-index is used to evaluate both models obtained from 20 iterations, ensuring more consistent and reliable results. RSF, a widely used survival analysis method, has shown strong predictive power, with studies reporting up to 99% accuracy in churn prediction. However, FRSF, a modified version that incorporates fuzzy logic, has proved superior performance, particularly in handling imprecise or uncertain data. The results show that FRSF achieves a lower error rate of 0.1739, compared to RSF's error rate of 0.1906. The median C-index value for the FRSF model is 0.78, which is higher than the median C-index value for the RSF model at 0.77. These findings suggest that FRSF outperforms RSF in churn prediction, making it a more reliable and righter model for finding at-risk customers. The study concludes that the model of FRSF is the preferred choice in telecommunications data for predicting churn, offering better predictive quality and consistency in handling uncertain data.

Keyword: Survival analysis, Right-censored, Random Survival Forest, Fuzzy Random Survival Forest, C-index

I. INTRODUCTION

Survival analysis is a statistical methodology employed to model the time until a significant event occurs, such as customer churn. Customer churn refers to the phenomenon where customers discontinue the use of a company's products or services, potentially leading to substantial financial losses [13]. One of the primary reasons survival analysis is chosen for predicting churn is its ability to estimate the time until the churn event, offering deeper insights than merely predicting whether a customer will churn [7]. In the business environment, particularly within the telecommunications sector, it is crucial for companies to predict when and why customers will cease their subscriptions, enabling the design of more effective retention strategies [1]. By modeling the timing of churn, companies can develop more timely and targeted interventions [9]. Churn prediction based on survival analysis provides valuable insights into both the risk of churn and the timing of such events [7]. Furthermore, survival analysis is capable of handling time-sensitive data, such as customer subscription duration [2]. Two essential functions in survival analysis are the survival function and the hazard function [8].

The survival function describes the probability that a customer will remain subscribed until a certain point in time without experiencing a churn event [5]. This function offers an estimate of the likelihood that a customer will continue their subscription at a given time [2]. Typically, the survival function decreases over time, as the longer the duration, the greater the likelihood of churn [4]. Conversely, the hazard function measures the rate of churn events occurring at a particular time [14]. This function provides valuable information about the risk of customer churn at any given point in time [16].

In the context of modeling survival and hazard functions for churn prediction, a popular method is Random Survival Forest (RSF). RSF is a machine learning approach that extends the Random Forest algorithm to survival analysis [5]. RSF constructs multiple decision trees to model the relationship between predictor variables and the time to churn [4]. Each individual tree generates a prediction of the survival function, which are then aggregated to produce more robust results [6]. Moreover, RSF can also be used to estimate the hazard function, which indicates the risk level of churn at a specific time [13].

While RSF is highly effective in modeling customer churn, it has limitations in handling uncertainty and ambiguity in data [14]. Ambiguous or uncertain variables often arise in customer data, such as unclear categories or inaccurate measurements [10]. To address these challenges, Fuzzy Random Survival Forest (FRSF) was developed by integrating fuzzy logic into survival analysis [12]. FRSF utilizes fuzzy membership values to handle uncertainty and provides a more robust model, particularly when the data exhibits high variability or noise [1]. By incorporating fuzzy logic, FRSF is able to generate more flexible and reliable predictions, especially when the data is inherently uncertain or noisy [11].

The data used in this study is customer data from the telecommunications industry. The telecommunications industry is highly dynamic and competitive, where customer churn represents a

critical issue. Given the intense competition, telecommunications companies face considerable challenges in retaining customers, as various external factors such as price, service quality, or competing offers can influence customers' decisions to switch providers [15]. Therefore, churn prediction within the telecommunications industry is crucial for identifying customers at high risk of leaving and for developing more effective retention strategies [3].

One of the reasons for selecting telecommunications data is the variety of customer-related information it contains, such as service usage, monthly charges, subscription duration, and more, providing deeper insights into the factors influencing customer churn [15]. Survival analysis in this context is well-suited to handle not only numerical data but also categorical data, such as gender, subscription status, and package type, which are common in churn prediction models. In this regard, models based on survival and hazard functions provide a more accurate representation of when customers are most likely to discontinue their subscriptions.

Given the above considerations, this study compares the Random Survival Forest (RSF) and Fuzzy Random Survival Forest (FRSF) methods for predicting customer churn in the telecommunications industry, where data often contains high levels of uncertainty and ambiguity.

II. METHODS

In this study, we compare two machine learning models for predicting customer churn in telecommunications industry such as Random Survival Forest (RSF) and Fuzzy Random Survival Forest (FRSF). Both models are designed to handle survival data, which is data includes “censored” observations. The goal of this research is to determine which model is more effective in predicting customer churn over time.

2.1. Data

This study utilizes customer data from the telecommunications data, specifically focusing on the 2P package, which includes Internet and TV services, derived from the entire customer database in the Jabodetabek region. The starting point is defined as the time of the customer's initial registration, with the final observation date being December 31, 2019. During this period, the times at which customers churned were recorded. The predictor variables considered include age, gender, internet data usage, internet speed, total monthly bill, duration of TV viewing, and the number of TV channels viewed. The data contains both categorical variables (such as gender) and numerical variables (such as age and internet usage). In this research, fuzzy logic was specifically applied to the age variable to address uncertainty and imprecision in data categorization. The age variable was transformed into fuzzy linguistic categories such as young, adult and elderly. The data was split into training and testing sets. We used an 80-20 split, which means 80% of the data was used for training the models, while the remaining 20% was held back for testing.

2.2. Algorithm Random Survival Forest

Random Survival Forest (RSF) is an advanced ensemble tree method introduced by Ishwaran et al. to analyze survival data with right-censoring [6]. This method builds on the Random Forest algorithm developed by Breiman (2001) extending its application to accommodate the specific challenges of survival analysis. By integrating survival-specific adaptations, RSF effectively models time-to-event outcomes while handling the complexities of censored data and nonlinear relationships among variables. The algorithm was implemented using the survival package in R. The RSF algorithm is implemented in the following manner [6].

1. Take B bootstrap samples from the dataset by repeatedly sampling with replacement. The survival tree is constructed using each bootstrap sample. About 37% of the data is left out of each bootstrap sample, known as OOB, out-of-bag data.
2. Random selection of m predictor variables is performed the survival tree is constructed for each terminal node of the tree to be utilized as splitting criteria.
3. The variable of the predictor is split using the log-rank splitting rule. Based on a given predictor variable, a node is partitioned to produce the maximum possible difference in the survival functions between its descendant nodes.
4. Repeat steps 2 and 3 until a large tree is formed and stops when each terminal node has a minimum of unique failure data points.
5. In each tree, each terminal node is searched for the CHF value with the Nelson-Alaen estimator.

   $$\widehat{\boldsymbol{H}}_{\boldsymbol{h}}(\boldsymbol{t}) = \sum_{\boldsymbol{t}_{l,h}<\boldsymbol{t}} \frac{\boldsymbol{d}_{l,h}}{\boldsymbol{r}_{l,h}} \quad (1)$$

   with $t_{l,h}$ is event time 1 of examples at descendant nodes *h*, $d_{l,h}$ is the event number on $t_{l,h}$, and$r_{l,h}$ is the individual's number at risk on $t_{l,h}$.
6. Calculate the ensemble bootstrap of CHF by finding the value of CHF ensemble, using the all of trees in the survival forest average value.

   $$\boldsymbol{H}_{\boldsymbol{e}}^{*}(\boldsymbol{t}|\boldsymbol{x}_{\boldsymbol{i}}) = \frac{\mathbf{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \boldsymbol{H}_{\boldsymbol{b}}^{*}(\boldsymbol{t}|\boldsymbol{x}_{\boldsymbol{i}}) \quad (2)$$

   With $H_b^*(t|x_i)$ is the CHF at the b-th bootstrap node.
7. From the CHF ensemble, the prediction error is calculated using OOB data.

   $$\boldsymbol{H}_{\boldsymbol{e}}^{**}(\boldsymbol{t}|\boldsymbol{x}_{\boldsymbol{i}}) = \frac{\sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \boldsymbol{I}_{\boldsymbol{i},\boldsymbol{b}} \boldsymbol{H}_{\boldsymbol{b}}^{*}(\boldsymbol{t}|\boldsymbol{x}_{\boldsymbol{i}})}{\sum_{\boldsymbol{b}=\mathbf{1}}^{\boldsymbol{B}} \boldsymbol{I}_{\boldsymbol{i},\boldsymbol{b}}} \quad (3)$$

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

The following is the flowchart of the RSF algorithm.

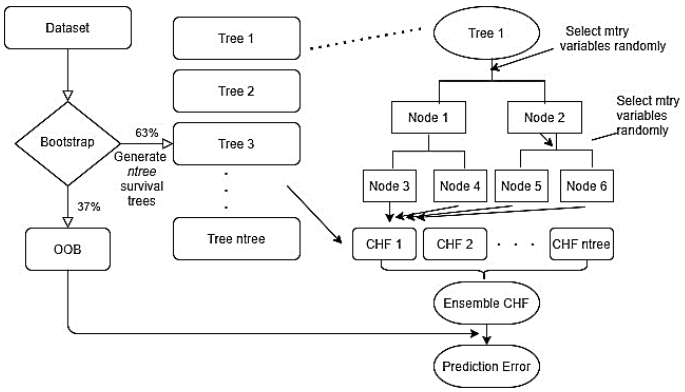


Figure 1 : Flowchart of Random Survival Forest

2.3. Algorithm Fuzzy Random Survival Forest

The Algorithm was implemented using the randomForestSRC package in R. The cumulative hazard function (CHF) serves as a measure of portfolio risk assessment in survival analysis, it is calculated as $\Lambda_t = \sum_m \frac{\alpha_{tm}}{|E_{tm}|}$, where α_{tm} is the lapses produced and $|E_{t_m}|$ is the policyholders in force at the time t_m. Time is discrete is assumed to have a value of $t \epsilon t_1, t_2, t_3, \ldots, t_e$. There are three main stages in the implementation of fuzzy systems:

1) Fuzzification transforms precise input data into fuzzy values. In the fuzzy set representation, I $|E|$ refers to the universe of all policyholders in a dataset. A fuzzy set $M \subset |E|$ is defined by a membership function $\mu_M : |E| \rightarrow [0,1]$ that associates each policyholder I of $|E|$ with a number $\mu_M(i)$ in the interval [0,1] representing the degree of membership of policyholder I to data subset M. Within a tree, fuzzy sets and partitions are established at each test node, using a linear membership function.
2) The process of inference includes the fuzzy rule base (FRB) and the inference mechanism:
   a. FRB consists of a set of fuzzy rules and a fuzzy database containing fuzzy sets of information. The fuzzy rule can be expressed by

$$if(x_1 bis A_1 b) and (x_2 bis A_2 b) and \ldots and (x_j bis A_j b) then (CHF bis \Lambda b)$$

   Where A serves as the explanatory variable and bb signifies the quantity of rules.
   b. The inference mechanism produces the system's output through fuzzy reasoning, mapping inputs to outputs using the FRB. This step is carried out to estimate the process prior to defuzzification and to define the child nodes.

2.4. C-index

The Concordance Index (C-index), commonly referred to as Harrell's C-index, functions as a measure for evaluation of the predictive accuracy of a model. This index is associated with the area

under the ROC curve. The probability estimates from this tree indicate that, in the random pair selection, the worst prediction will occur for examples with the event (status=1). The rate of error is calculated as 1−C-index, with values spanning from 0 to 1. An error rate of 0.5 suggests performance akin to random guessing, while an error rate of 0 denotes perfect accuracy [6].

III. RESULTS AND DISCUSSION

3.1. Model Result from Random Survival Forest (RSF)

Table 1 indicates that from 804 samples (80% of the training data), 84 customers experienced churn during the observation period. The survival trees were constructed using the RSF (Random Survival Forest) model, which involved 100 survival trees and included all predictor variables through the bootstrap method. The objective was to build the model of RSF by estimating it on all original data. The results from the bootstrap are referred to as out-of-bag (OOB) data. According to the resulting model of RSF, an estimated rate of error is 29.42% was obtained.

Table 1 : RSF Model

Size of Sample	**804**
Churns number	84
Trees number	100
Node size of terminal forest	15
Average count of terminal nodes	34,15
Count of variables evaluated at each split.	3
Number of variables	7
Resampling is applied to grow the tree	Swor
resampling size is employed for tree growth	507
Method	RSF
Type	Surv
Rule of splitting	Logrank *random*
Number of random partitioning points	10
Rate of error	29,42%

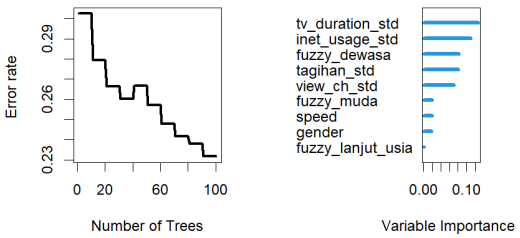
3.2. Model Result Fuzzy Random Survival Forest (FRSF)

Table 2 shows that out of 804 samples (representing 80% of the training data), During the observation period, 84 customers churned. Survival trees were developed utilizing the Flexible Random Survival Forest (FRSF) model, which entailed generating 100 survival trees using all predictor variables through the bootstrap technique. The objective was to build the FRSF model by making estimations on the entire original dataset. The out-of-bag (OOB) data from the bootstrap results indicated an estimated error rate of 23.15%, which is lower than the error rate of the previous RSF model.

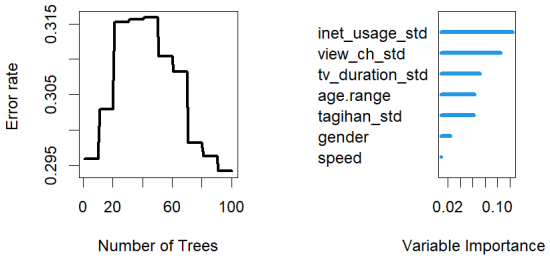
Table 2 : FRSF Model

Size of Sample	**804**
Churns number	84
Trees number	100
Node size of terminal forest	15
Average count of terminal nodes	34,98
Count of variables evaluated at each split.	3
Number of variables	9
Resampling is applied to grow the tree	Swor
Resampling size is employed for tree growth	507
Method	RSF
Type	Surv
Rule of splitting	Logrank *random*
Number of random partitioning points	10
Rate of error	23,15%

3.3. Model Comparison



(a) FRSF Model



(b) RSF Model

Figure 1 : Error Rate Comparison

Figure 1 (a) depicts the variation in error rate as trees number grows in the Fuzzy Random Survival Forest (FRSF) model, whereas Figure 1 (b) shows the related changes in the Random

Survival Forest (RSF) model. The analysis results demonstrate that the FRSF model exhibits superior performance compared to the RSF model in predicting customer churn for telecommunications data.

In terms of error rate, the FRSF model achieves a value of 0.23, which is lower than the RSF model's error rate of 0.285. The decline in the error rate for the FRSF model is consistent as the number of trees increases, reaching stability at 100 trees. In contrast, the RSF model exhibits a bell-shaped curve, with an optimal error rate observed between 50–60 trees, followed by an increase in the error rate as the number of trees continues to grow. This pattern suggests a potential overfitting issue in the RSF model, which is not observed in the FRSF model. Consequently, the FRSF model provides more stable results in capturing survival patterns, particularly in complex datasets with uncertainty or noise.

Table 3 further compares the error rates of the two models, showing that the FRSF model achieves an error rate of 0.1739, which is lower than the RSF model's error rate of 0.1906. These results indicate that the FRSF model surpasses the RSF model in predictive quality, establishing it as a more favored approach in predicting customer churn for telecommunications data.

Table 3 : Comparison of Error Rates in Prediction Models

Error rate FRSF	*Error rate* RSF
0.17395833	0.190625

Figure 2 presents a boxplot comparing the C-index values of the RSF and FRSF models over 20 iterations. The figure reveals that the median C-index value for the FRSF model is 0.78, which is higher than the median C-index value for the RSF model at 0.77. This indicates that the FRSF model performs better than the RSF model in distinguishing between different customer retention durations.

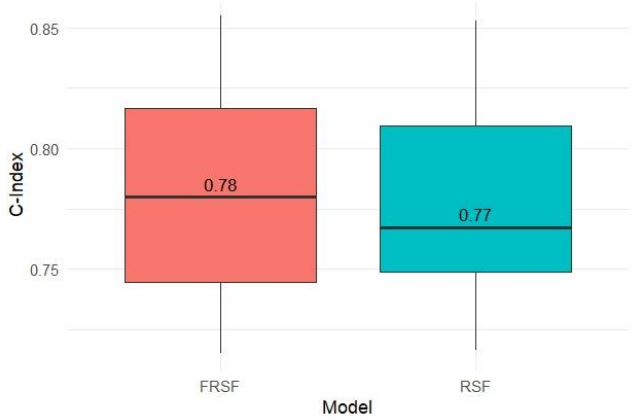


Figure 2 : Comparison of C-indeks value in prediction models

The results of this study indicate that the FRSF model outperforms the RSF model in predicting customer churn in the telecommunications industry. The FRSF model effectively addresses the uncertainty in categorizing customer age variables into young, adult, and elderly categories with

greater clarity. This distinction is crucial for telecommunications service providers because customers with different age characteristics may exhibit varying churn tendencies. In contrast, the RSF model employs a more traditional categorization of age variables: < 20 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, and over 60 years. While this approach is more straightforward, it may not be as flexible as the FRSF model in handling the variability and uncertainty of customer age data. In the context of the telecommunications industry, having a predictive model that can manage uncertainty and inaccurate data is essential. Customer data often varies and is not always precise. With the FRSF model, service providers can more accurately identify customers at risk of churn and implement more appropriate retention strategies, such as personalized offers or early interventions to prevent churn. The more explicit and flexible approach to handling age data by the FRSF model allows telecommunications service providers to be more responsive to their customers' needs, ultimately enhancing customer retention and reducing churn rates.

IV. CONCLUSION

In analyzing churn data for customers in the telecommunications data, this study compares RSF and FRSF models. The predictive comparison performance between the RSF and FRSF models was based on the value of C-index. The median C-index from 20 iterations for the FRSF model is 0.78, which is higher than the median C-index for the RSF model at 0.77.The error rate for the FRSF model was 0.1739, which is lower than the RSF model's error rate of 0.1906. This finding indicates that the model of FRSF offers superior predictive quality compared to the RSF model, making it the best choice in the telecommunications data for predicting customer churn. The key variables in the FRSF model include TV viewing duration, internet usage, adult age category, total monthly bill, The number of TV channels viewed, young age category, internet speed, gender, and elderly age category.

The results of this study can serve as a valuable consideration for the telecommunications industry in formulating customer retention policies. By employing the FRSF model, telecommunications companies can more accurately predict when customers will likely churn, enabling more targeted interventions.

This study does, however, have certain limitations. One such limitation is the application of fuzzy logic being confined to a single variable, namely age. Future research is advised to explore the application of fuzzy logic to other variables. Additionally, this study only compares the RSF and FRSF models, so it is recommended that future research expand by comparing the performance of these models with other methods, such as Survival SVM, to identify the most effective approach.

REFERENCES

[1]. Bhat, S. A., Khan, A. K., & Shamsher, M. (2020). Predicting customer churn in telecom industry using machine learning techniques. Journal of King Saud University-Computer and Information Sciences.

[2]. Chen, L., Yang, Q., & Ma, Y. (2019). Survival analysis of customer churn in subscription-based services: A case study in telecommunications. Journal of Data Science and Analytics, 4(2), 115-127.

[3]. Hossain, M. S., Bhuiyan, M. H., & Rahman, M. A. (2020). A machine learning approach for customer churn prediction in the telecommunications industry. Journal of Telecommunications and Digital Economy, 8(3), 295-306.

[4]. Hothorn, T., Hornik, K., & Zeileis, A. (2014). A Lego system for conditional inference. The American Statistician, 58(1), 34-47.

[5]. Ishwaran, H., Kogalur, U. B., & Blackstone, E. H. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841-860.

[6]. Ishwaran, H., Kogalur, U. B., & Lu, P. (2010). Random survival forests for R: Predictive accuracy and interpretation. Biostatistics, 11(3), 602-616.

[7]. Jin, S., Du, M., & Zhang, Y. (2020). A survey of churn prediction in telecommunications using survival analysis. Computers, Materials & Continua, 63(2), 1107-1118.

[8]. Kalbfleisch, J. D., & Prentice, R. L. (2011). The Statistical Analysis of Failure Time Data (2nd ed.). Wiley-Interscience.

[9]. Khan, I., Shah, F., & Younis, A. (2020). Predicting customer churn in telecom using machine learning techniques: A comparative study. International Journal of Computer Applications, 175(11), 5-15.

[10]. Kumar, P., Singh, V. P., & Kaur, A. (2017). A review on predictive analytics of customer churn in telecom industry. International Journal of Computer Science and Information Security, 15(12), 30-45.

[11]. Li, Y., Zhang, Z., & Zhang, S. (2019). Fuzzy random survival forest for churn prediction in the telecommunication industry. International Journal of Fuzzy Systems, 21(3), 803-815.

[12]. Liu, B., Wei, Y., & Zhang, H. (2020). Fuzzy random survival forest for the analysis of customer churn in telecom industry. Fuzzy Optimization and Decision Making, 19(3), 331-349.

[13]. Nurhaliza, S., Sadik, K., & Saefuddin, A. (2022). A comparison of Cox proportional hazard and random survival forest models in predicting churn of the telecommunication sector customer. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 16(4), 1433-1440. https://doi.org/10.30598/barekengvol16iss4pp1433-1440

[14]. Zhao, J., Wang, X., & Sun, Z. (2019). A comparative study of machine learning algorithms for churn prediction in the telecommunications industry. International Journal of Data Science and Analytics, 8(4), 387-396.

[15]. Zhao, Y., Li, Z., & Wang, H. (2020). Predicting churn in telecom using survival analysis techniques. Computers & Operations Research, 120, 104899.

[16]. Zhang, L., Lin, W., & Zhang, L. (2021). Customer churn prediction using survival analysis: A study of the telecommunications industry. Expert Systems with Applications, 176, 114862.