Churn Analysis in Telecommunication Industry Customers Using Semiparametric and Non Parametric Survival Method

Churn Analysis in Telecommunication Industry Customers Using Semiparametric and Non Parametric Survival Method

Ade Vreyyuning Monika*, Indahwati, and Muhammad Nur Aidi

Department of Statistics, IPB University, Indonesia

*E-mail : ade_monika@apps.ipb.ac.id

Abstract. One of the challenges faced by Customer Relationship Management (CRM) of Telecommunication Company is customer retention efforts. The success rate of customer retention can be seen by the customer switching or churn process. Using the background and usage data related to the loyalty of the customers, it can be known the probability that customers will churn at a certain time. To overcome the problem, survival analysis will be carried out with parametric model using Cox Proportional Hazard and non-parametric using Support Vector Machine (SVM) for customer data. The data used in this study are customer data from telecommunication company who subscribe bundling package Internet and IPTV which are taken from 1000 customers in the Jabodetabek area. The time a customer is registered for the first time is defined as the start time and the last observation date is the end time. During the observation period, customer churn time is recorded. Based on the Kaplan Meier curve and log rank test, it is shows that there is a significant difference curve between customers with six different age categories. The results of the analysis found that SVM survival is able to compensate for Cox Proportional Hazard. Based on the result of the concordance index, the performance of the SVM survival has a better performance than Cox Proportional Hazard. One possibility that survival of SVM give a better performance is because the assumption of Cox Proportional Hazard is not met.

1. Introduction

One of the Telecommunication Companies, as a provider of telecommunication services in Indonesia, presents a product as home services product based on subscriber. To increase the market share, made the Customer Relationship Management (CRM) has played an important part in this company. One of the challenges faced by Customer Relationship Management (CRM) of telecommunication companies is customer retention effort. The success rate of customer retention can be seen in the process of switching customers or churn. High retention is equivalent to low switch [1]. Therefore, to maintain profitability and market share, the industry should not only focus on acquisition program but also on retaining customers and reducing the churn rate. Also, keeping an existing customer is five times cheaper than one attracting a new one [2]. Telecommunication companies have a lot of data about customers, such as demographics, finances, usage behaviour and call logs which present an opportunity to make this data actionable using analytical techniques [3]. With this data, it can also be known the probability that customers will churn at a certain time.

Survival analysis is a statistical procedure for data analysis with the output of variables in the form of time to an event. Time of occurrence can be measured in days, weeks, months, years, when observing the occurrence of events. While the event what is meant is every experience from observation

that is likely to occur [4]. The two most fundamental things of the survival analysis are the survival function and the hazard function. The survival function is the probability of an object surviving at a certain time. Hazard function represents the failure rate of an object [5].

The survival analysis using cox proportional hazard for telecommunication industry customers was conducted by Suhartono et al. [6]. However, the research was conducted for cellular telecommunications so that the variables used are customer profiles and cellular data usage. Based on this study provide effective results to distinguish the relative risk between churn and non-churned customers, and also differentiate between loyal (longer stay) and shorter customers with a significant factor.

The development of survival analysis with non-parametric methods was carried out which apply survival analysis with the Support Vector Machine (SVM) approach [7][8]. Van Belle et al. [8] used survival SVM to compare the performance of methods using the Cox Proportional Hazard, Accelerated Failure Time Model (AFT model), cSVM (SVM with linear kernel), and cSVM-Gaussian Radial Basis (SVM with RBF kernel). By using two data sets, it was concluded that the results using SVM survival were better than using Cox Proportional Hazard or AFT for the data in these cases. Besides that, Khotimah et al. [9][10] also compared Cox Proportional Hazard and Survival SVM. The results showed that the SVM Survival method was better than the Cox Proportional Hazard for cervical cancer case data.

In this study, a survival analysis will be carried out using a parametric model using Cox Proportional Hazard and non-parametric using Survival SVM for customer data of telecommunications company. This is because there is a possibility that the model that has been built does not meet the assumptions. The data used in this research is data of telecommunication company subscribers who subscribe bundling package Internet and IPTV, with data taken based on customer background and customer behavior.

2. Research Methods

2.1. Data

The data used in this study are customer data from telecommunication company who subscribe bundling package Internet and IPTV which are taken from 1000 customers in the Jabodetabek area from 2017 to 2019. The time a customer is registered for the first time is defined as the start time and the last observation date is the end time. During this observation period, customer churn time is recorded.

The variables used in this study include customer background data, such as: age, gender, average total monthly bill (Rupiah), internet speed (Mbps), and customer behavior, for example: internet data usage (Gb), number of TV channels watched, and TV watching duration (hours). Survival time is used to indicate the time the customer churn occurred, or for censored cases, the last time the customer was observed, both measured from the initial time (survival time = 0). And the status variable is used to distinguish censored cases from observed cases (churn). Status = 0 for customer churn cases and status = 1 for censored cases in this study.

2.2. *Data Analysis Procedure*

Data analysis in this study use R software. The step of the analysis are as follows:

1. Processing data; such as defining the framework, classifying some variables into categories, and standardizing customer behavior variables.
2. Exploring data of each categorical variable to provide an initial description of the survival function for each group and see whether the group is proportional using Kaplan Meier curve and Log Rank test.

   The general equation for the probability of the Kaplan Meier is

$$\hat{S}(\mathrm{t}_{(j)}) = \hat{S}(\mathrm{t}_{(j-1)})\hat{P}(T > t_{(j)} \mid T \geq t_{(j)})$$

$$\hat{S}(\mathrm{t}_{(j-1)}) = \prod_{i=1}^{j-1} \hat{P}(T > t_{(j)} \mid T \geq t_{(j)})$$

The Log Rank test is a test used to compare survival curves in different groups [4]. The hypothesis of the Log Rank test for two or more groups is as follows

H_0: there was no difference of survival curves in different groups

H_1: there was at least one difference of survival curves in different groups

The test statistic used is

$$\chi^2 = \sum_{h=1}^{G} \frac{(O_h - E_h)^2}{E_h}$$

Test criteria used are rejected H_0 if the test statistic value $\chi^2 > \chi^2_{\alpha,G-1}$

3. Analyzing the semi-parametric model, namely the Cox Proportional Hazard from customers to evaluate the factors that have a significant influence. The Cox Proportional Hazard Regression Model first introduced by D.R. Cox [11]. This model is a semi-parametric model because the baseline hazard expressed as $h_0(t)$ does not require following a certain distribution. The regression equation of Cox Proportional Hazard is as follows.

$$h(t, \boldsymbol{x}) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x})$$

with $\boldsymbol{x}$ is a vector of predictor variables $(x_1, x_2, \ldots, x_k)^T$ while $\boldsymbol{\beta}$ is a coefficient parameter for predictor variable $\boldsymbol{x}$ [4].

4. Analyzing non-parametric models using the Survival SVM method. Model output is a prognostic function also referred to as utility function and more specifically in medical research it is called prognostic index or health function where $u: \mathbb{R}^D \rightarrow \mathbb{R}$ is defined as follows

$$u(\boldsymbol{x}) = \boldsymbol{w}^T \varphi(\boldsymbol{x}),$$

where w is the unknown parameter vector and $\varphi(x)$ is the transformation of the covariate x.

5. Obtaining the best survival analysis method based on the goodness of the model, using the concordance index.

3. Result and Discussion

3.1. Kaplan Meier Curve and Log Rank Test

Kaplan Meier Curve and Log Rank Test were conducted before modeling with Cox Proportional Hazard and Survival SVM. Kaplan Meier Curve and Log Rank Test were conducted to determine the difference in survival curves between categorical variables used in this study. The results of the Kaplan Meier Curve and the Log Rank Test are presented in Figure 1.

Based on Figure 1, it is found that the survival curves for male and female customers are different when seen visually. Male customers tend to have a greater chance of churning faster than women. Whereas the curve for the variable internet speed, shown by figure 1 (b), indicates that there is no different curves for each category. The survival curve on the variable age shows a difference in each category. Customers with age more than 59 years have the greatest chance to stay subscribed, while customers in the category of age less than 20 years tend to churn faster. However, when viewed based on the results of the log rank test, the p-values of the variable gender and internet speed are 0.068 and 1. Because the p-value is more than $\alpha = 0.05$, the decision is to not reject H_0. Therefore, there is no significant difference curves between male and female customers and there is no significant difference between customers with six different internet speed categories. Whereas for the variable age, the p-value in the log-rank test was 0.0024. Because the p-value is less than $\alpha = 0.05$, the decision is to reject H_0. It is concluded that there is a significant difference curve between customers with six different age categories.

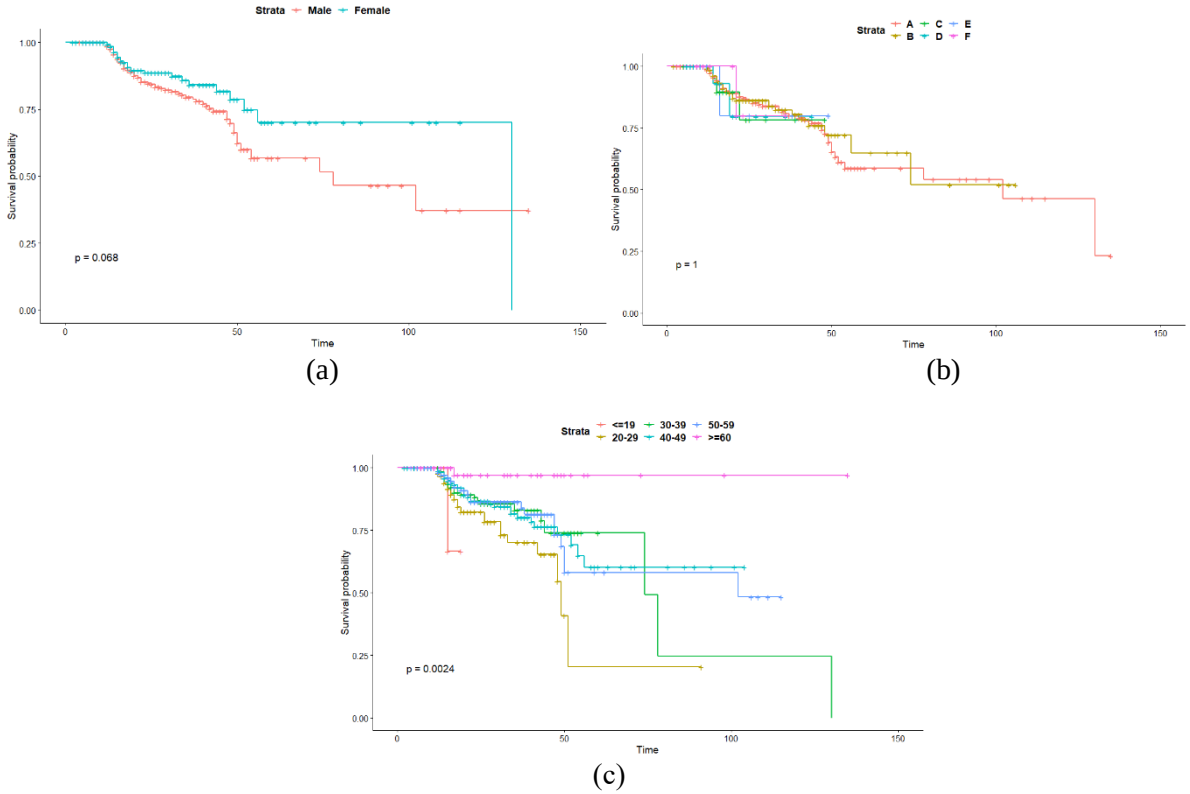

Figure 1 Kaplan Meier Curve and Log Rank Test of variable gender (a), internet speed (b), and age (c)

3.2. Cox Proportional Hazard

The analysis was continued by doing the Cox Proportional Hazard modeling. In the Cox Proportional Hazard modeling there are two things that are done, namely partial testing for each predictor variable and testing the Cox Proportional Hazard assumption. The results of parameter estimation and partial testing are presented in Table 1.

Table 1 Cox Proportional Hazard Result

Variable	Coefficient	Exp(Coefficient)	S.E(Coefficient)	Z	P-value
Gender	-0.447	0.639	0.217	-2.063	0.039
Age (20-29)	-0.767	0.465	1.024	-0.748	0.454
Age (30-39)	-1.576	0.207	1.031	-1.529	0.126
Age (40-49)	-1.510	0.221	1.029	-1.468	0.142
Age (50-59)	-1.408	0.245	1.041	-1.353	0.176
Age (>=60)	-3.859	0.021	1.442	-2.676	0.007
Inet speed (B)	-0.277	0.758	0.235	-1.181	0.237
Inet speed (C)	0.272	1.313	0.627	0.434	0.664
Inet speed (D)	0.281	1.325	0.747	0.376	0.707
Inet speed (E)	0.070	1.073	1.148	0.061	0.951
Inet speed (F)	-0.156	0.856	1.329	-0.117	0.907
Bill	0.104	1.110	0.091	1.144	0.253
Internet usage	-0.468	0.626	0.164	-2.852	0.004
TV watch duration	-0.539	0.584	0.138	-3.897	0.000
Number of channels	-0.400	0.670	0.150	-2.678	0.007

Based on Table 1, the variables Gender, Age (> = 60), Internet Usage, TV Watch Duration, and Number of Channels Watched have a significant effect on customer churn time in the telecommunication company. Meanwhile, other variables have no significant effect. The value of the odds ratio on a significant variable can be seen in the results of exp (coefficient). For example, in the variable Gender, the odds ratio is 0.639, which means that female customers are 0.639 times faster at churning than male customers. Furthermore, the Cox Proportional Hazard regression model can be written as follows

$$\begin{aligned} h(t,\boldsymbol{x})_{ep} = h_0(t)\exp\,(&-0.447\,Gender - 0.767\,Age(20-29) - 1.576\,Age(30-39) \\ &- 1.510\,Age\,(40-49) - 1.408\,Age\,(50-59) - 3.859\,Age(\geq 60) \\ &- 0.277\,Inet\ speed\ B + 0.272\,inet\ speed\ C + 0.281\,Inet\ speed\ D \\ &+ 0.070\,Inet\ speed\ E - 0.156\,Inet\ speed\ F + 0.104\,Bill - 0.468\,Inet\ usage \\ &- 0.539\,TV\ watch\ duration - 0.400\,Number\ of\ channels\ watched) \end{aligned}$$

Table 2 Cox Proportional Hazard Assumption Test

Variable	Correlation	P-value
Gender	1.474	0.225
Age (20-29)	0.536	0.464
Age (30-39)	0.007	0.932
Age (40-49)	0.287	0.592
Age (50-59)	0.043	0.836
Age (>=60)	1.140	0.286
Inet speed (B)	0.374	0.541
Inet speed (C)	0.078	0.781
Inet speed (D)	0.228	0.633
Inet speed (E)	0.349	0.555
Inet speed (F)	1.429	0.232
Bill	11.674	0.001
Internet usage	0.983	0.321
TV watch duration	16.676	0.000
Number of channels	3.697	0.055

However, the Cox Proportional Hazard method has a weakness that is the proportional hazard assumption that must be met. In Table 2, it can be seen that the variable Bill and variable TV Watch Duration do not meet the assumptions. Because the Cox Proportional Hazard model does not meet the assumptions, another model that does not pay attention to the assumptions is carried out.

3.3. Survival SVM

The SVM Survival method in this study is used to solve the problem of unfulfilled assumptions in the Cox Proportional Hazard model that has been carried out. There is a parameter γ used which is 0.1 and the kernel in Survival SVM uses the RBF kernel. The results of the descriptive statistics of the prognostic index produced by the SVM Survival are as follows in Table 3.

Table 3 Descriptive Statistics of Survival SVM Prognostic Index

Variable	**Mean**	**Median**	**Standard Deviation**
Prognostic Index	23.37	23.37	0.056

The prognostic index of the customers can be categorized into two categories, customers who have a great opportunity to subscribe to a longer period of time and customers who have little chance of subscribing longer or customers who tend to churn faster. Categorization was based on the mean of prognostic index score. If the prognostic index of a customer is below the mean, then the customer is in the category of customers who have a lower chance of churning and vice versa.

3.4. C-Index Comparison

After obtaining the prognostic value for each observation, the performance of the SVM and Cox Proportional Hazard method can be calculated. The results of the comparison of the C-Index values are presented in Table 4 below.

Table 4 C-Index of Cox PH and Survival SVM

Method	C-Index
Cox PH	0.743
Survival SVM	0.945

Based on the comparison of the C-Index values in Table 4, it is known that the SVM Survival method has a larger C-Index. So it can be concluded that the SVM Survival method has better performance when compared to the Cox Proportional Hazard method for customer data in telecommunications companies.

4. Conclusion

In the survival analysis modeling for Telecommunication company customers, the Survival SVM method provides better performance than the Cox Proportional Hazard method. In the Cox Proportional Hazard model, the result shows that the variables Gender, Age (≥60), Internet Usage, TV Watch Duration, and Number of Channels Watched have a significant effect on customer churn time. However, the Billing and TV viewing duration variables do not meet the Proportional Hazard assumption.

References

[1] Buttle F 2008 Customer Relationship Management: Concepts and Tools 255–290
[2] Gallo A 2014 The Value of Keeping the Right Customers Harvard Business Review 2–6
[3] Óskarsdóttir M, Bravo C, Verbeke W, Sarraute C, Baesens B, and Vanthienen J 2017 Social network analytics for churn prediction in telco: Model building evaluation and network architecture Expert Systems with Applications 85 204–220
[4] Kleinbaum D G and Klein M 2012 Survival Analysis: A Self Learning Text (Third ed) London: Springer
[5] Moore D F 2016 Applied Survival Analysis Using R London: Springer
[6] Suhartono D, Saefuddin A, and Sumertajaya I M 2013 Survival Analysis of Customer in Postpaid Telecommunication Industry Forum Statistika dan Komputasi: Indonesian Journal of Statistics Vol. 18 No.1 April 2013 p: 1-10
[7] Van Belle V, Pleckmans K, Suykens J A, and Van Huffel S 2007 Support Vector Machines for Survival Analysis Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), Plymouth (UK), 1-8
[8] Van Belle V, Pleckmans K, Suykens J A, and Van Huffel S 2008 Survival SVM: A Practical Scalable Algorithm Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN2008) Bruges (Belgium) 89-94
[9] Khotimah C, Purnami S W, and Prastyo D D 2017 Additive survival least square support vector machines: A simulation study and its application to cervical cancer prediction AIP Conference Proceedings, 1902, 050024
[10] Khotimah C, Purnami S W, and Prastyo D D 2018 Additive Survival Least Square Support Vector

Machines and Feature Selection on Health Data in Indonesia IEEE Xplore: International Conference on Information and Communications Technology (ICOIACT) 2018
[11] Cox DR and Oakes D 1984 Analysis of Survival Data Chapman and Hall, London