IBM Telco Customer Churn Prediction with Survival Analysis

Hafiz Rahman[1], Ridho Pandhu Afrianto[1], Farisi Mohammad[1], Dhia Alif Tajriyaani Azhar[1], Ratih Ardiati Ningrum[1]

[1] Data Science Technology, Airlangga University, Surabaya, Indonesia
ratih.an@ftmm.unair.ac.id

Abstract. Customer churn or attrition occurs when customers stop engaging with a company or product. A high churn rate can be a serious threat to a company's sustainability. In the competitive telecommunications industry, churn analysis is key to developing customer retention strategies. This research will conduct churn analysis on the IBM Telco public dataset using the survival analysis method with two competing risks in the dataset, namely internal company deficiencies and external factors. The results of the subscription duration analysis show that customers who stop tend to have shorter subscription durations and spend less. Kaplan-Meier revealed significant differences in probability distributions for variables such as contract type, additional services, and demographic factors. The stratified Cox method confirms the significant impact of variables such as the presence of a partner, type of contract, multiple lines, and additional services on the risk of churn. The Goodness of Fit test validates the ability of the survival analysis model to differentiate churn cases overall, with a concordance value reaching 0.867. The CIF Plot shows that certain factors, such as gender and telephone service, are not significant in differentiating between events caused by internal company deficiencies or better offerings from other providers. The research results show that churn analysis using the survival analysis method can provide a basis for companies to optimize business strategies in facing competitive challenges so that they can make better decisions in the future.

Keywords: Survival analysis, Churn, Business, Customer behavior.

1 Background

Customer churn, also called customer attrition or customer turnover, is a phenomenon in which customers stop doing business with a company or stop using its products or services. It is a metric that a company can use to measure its ability to retain customers. A high churn rate indicates that a large portion of customers are dropping out of business, which can be a concern for a company's long-term sustainability and growth. On the other hand, a low churn rate indicates that a business is able to maintain its customer base well. Reducing the churn rate can increase a company's revenue and give it an advantage over other companies [1]. The purpose of churn analysis is to identify the reasons for customer churn so that the company can fix existing problems with its

T. Amrillah et al. (eds.), *Proceedings of the International Conference on Advanced Technology and Multidiscipline (ICATAM 2024)*, Advances in Engineering Research 245,
https://doi.org/10.2991/978-94-6463-566-9_25

reasons for customer churn so that the company can fix existing problems with its products or services [2]. Given the fierce competition in the telecommunications industry, churn analysis can be performed on customer data to help companies develop effective customer retention strategies [3].

Churn analysis has been researched extensively using various methods. Research [4] recommends the Cox model with SCAD to identify the reasons for vehicle insurance customer churn. Meanwhile, the results of research analysis [2] show that the use of the random forest algorithm can predict churn in the telecommunications sector with 99% accuracy. Research [5] analyzed car sharing program transaction data in a year using the extended Cox model. The results of the study show that the probability of customer retention decreases over time and the number of discount coupons can reduce the customer churn rate over 130 days.

This research will use the survival analysis method to analyze and predict churn on the IBM Telco dataset. The methods used include Kaplan-Meier, stratified Cox models, and Fine-Gray to analyze two competing risks in the dataset, namely internal company deficiencies and external factors.

2 Literature Review

The Kaplan-Meier estimator is a non-parametric statistical method that can calculate the survival function without distribution assumptions [6]. Calculations using the Kaplan-Meier method involve calculating the probability of an event at a specific point in time. The probability is then multiplied successively by the probability of the previous time point to obtain the final estimate [7]. The results of the Kaplan-Meier estimates can be compared to see if there are significant differences in survival between two or more groups. In addition to simple visual comparisons, a log-rank test can be performed to calculate the statistical significance of differences in curves [7, 8].

The stratified Cox model is a modified Cox model that can be used to handle covariates that do not meet the PH assumption. Covariates that did not meet the assumptions were included in the strata. The influence of covariates in the stratum can be observed through the Cox PH model created for each stratum [9, 10]. In addition to being used to deal with covariates that violate PH assumptions, stratified models can also aid in analyzing dealing with subsets of patients.

When there is more than one outcome or event that may occur in survival data, the situation is called competing risks [11]. For example, deaths from other causes can prevent deaths from cancer, which is the event we want to observe. To handle this situation, the Fine-Gray model [12] can be used, which models CIF (Cumulative Incidence Function) with covariates. The sub-distribution function, a CIF curve, has similarities to the Cox model. The difference between the two is that the sub distribution function models the hazard function (hazard sub distribution), which comes from the CIF.

As an industry with a high level of competition, customer churn in the telecommunications industry has been widely studied from various points of view. Research [1, 13, 14, 15] applies survival analysis methods in analyzing customer churn of telecom-

munications companies. In research [1], topic modeling was carried out with LDA (Latent Dirichlet Allocation) to identify topics of customer complaints to customer service and used them as variables in the model. The results of research [13] show that the use of the Cox proportional hazards model can provide simple results and help in the process of optimizing customer retention policies. Research [14] examined the effect of promotional campaigns, cell phone tariffs, age, and use of the auto payment feature on survival time and hazard using the Cox and Kaplan-Meier models. Research [15] found that the performance of the Cox PH model can be outperformed by SVM (Support Vector Machine) for predicting customer churn for Internet and IPTV packages in the Jabodetabek area. This result may be obtained because the data does not meet the PH assumption.

As the field of machine learning develops, many studies [2, 16, 17, 18] have analyzed telecommunications customer churn using various machine learning methods. Research [2] obtained an accuracy rate of 99% when predicting churn in the telecommunications sector. Research [16] proposed a customer churn prediction model based on daily behavior because monthly prediction models were deemed unable to capture dynamic changes in customer behavior. Monthly models are also too slow to detect churn compared to daily models. The results obtained show that the performance of the daily model far exceeds the performance of the monthly model. Research [17] applied XGBoost to customer churn data from a large Nepalese telecommunications company and obtained an accuracy of 97% and an f1 value of 88%. Research [18] proposes an adaptive learning approach that can adapt and respond quickly to changes in customer habits or choices. Using a Naive Bayes classifier with genetic algorithm-based feature weighting, the adaptive model achieved accuracy of 95%, 97%, and 98% on the BigML Telco churn, IBM Telco, and Cell2Cell datasets. Findings from research [19] show that not all factors assumed to have a significant influence on churn affect genuine customers. Call quality, billing, and brand image influence customer loyalty and churn in Turkey. However, other factors, such as loyalty programs, length of subscription, and previous churn experience, unexpectedly do not have a significant effect on customer loyalty.

3 Methodology

This research uses data sourced from Kaggle titled Telco Customer Churn: IBM Dataset. This data is fictional and provides 7043 observations and 33 variables for an internet service product. The data used already includes customer demographics, customer locations, population for each area, types of services, and the status of each customer.

Before starting the survival analysis on customer churn, it is important to conduct data exploration and visualization. This data exploration is carried out with the hope of providing insights regarding the data used and its distribution. Survival analysis begins with the application of the Kaplan-Meier method for each variable deemed relevant in the dataset. This method is used to estimate the customer lifetime function over time and provides an overview of how long customers are likely to remain loyal or switch.

The result of this step will provide a foundation for understanding churn behavior in general.

Next, to understand the factors contributing to the risk of churn simultaneously, the stratified Cox Proportional Hazard method was applied. This involves determining significant categorical variables and dividing the population into homogeneous groups. This method is expected to provide insights into the relationship between variables and the risk of churn. The next step is the Goodness of Fit test, which aims to evaluate how well the generated model fits the existing data. This test is crucial to ensure that the statistical model used can accurately and consistently represent the data.

To gain a deeper understanding of churn dynamics, a competing risk analysis was conducted. This analysis considers the factors that may cause customers to switch to different churn categories. The results are expected to provide a more detailed picture of the triggers for churn. Finally, to provide a clear visualization related to the churn risk levels based on certain factors, a CIF Plot Based on Factor is created. This plot will be very useful for identifying and understanding the impact of specific variables on customer survival rates.

4 Results and Discussion

4.1 Churn Label Distribution

A pie chart is shown comparing the number of customers with the label 'no,' meaning did not switch to another provider, and 'yes,' meaning switched to another provider.

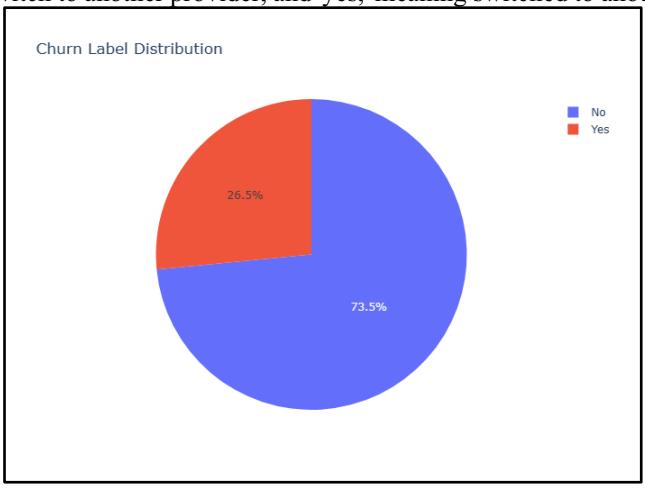


Fig. 1. Churn Label Distribution

It is evident from the pie chart that the data is dominated by the churn label 'no,' which accounts for 73.5%, indicating that there is a large amount of censored data in this study.

4.2 Churn Reason Distribution

The distribution of reasons for customers leaving or switching to another provider is shown. The lighter the color, the higher the frequency of customers leaving for that reason.

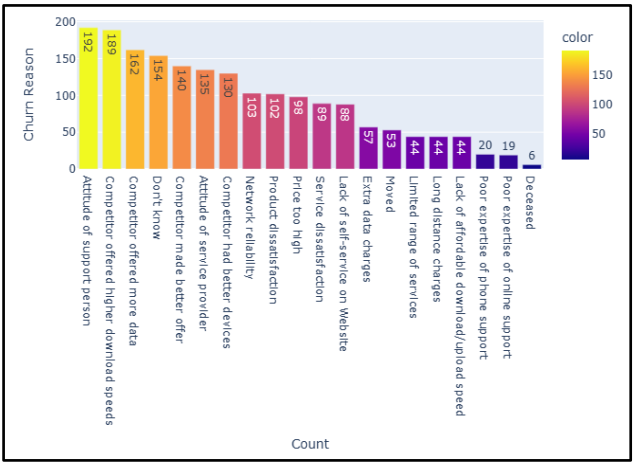


Fig. 2. Churn Reason Distribution

The plot shows 20 reasons why a customer chooses to switch to another provider. Based on the distribution, the reasons for switching can be divided into two categories: internal and external. Internal reasons are those influenced by the company's policies, such as product quality, service quality, and pricing. External factors include things beyond the company's control, such as competitor offers and network reliability. In this analysis, we will also analyze competing risks, where events caused by provider shortcomings/internal factors will be labeled as 1, and external or other factors will be labeled as 2.

4.3 Churn Total Based on City

The distribution of the number of customers who switched, based on the city, is shown below. The plot shows that the city with the highest churn cases is Los Angeles, with 305 cases. This is followed by San Diego with 150 cases, and San Jose with 112. Meanwhile, Modesto has the lowest number of churn cases.

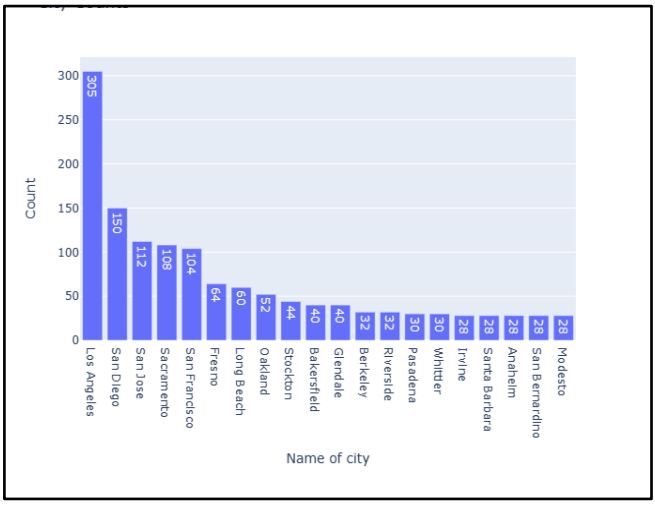
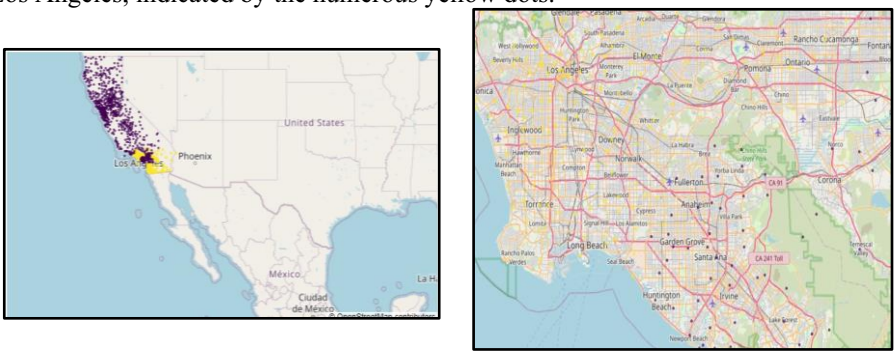


Fig. 3. Churn Total Based On City

4.4 Churn Plot on Real Map

It is shown in detail the locations where many customers are experiencing churn. This visualization shows the detailed locations where churn events occurred. It is evident from the visualization that most churn events are concentrated in downtown Los Angeles, indicated by the numerous yellow dots.



(a) (b)

Fig. 4. Churn Plot on Real Map: (a) Abstract Map and (b) Detailed Map

4.5 Tenure Months based on Churn Label

Show the distribution of customer usage duration based on the churn label, which are 'yes' and 'no'. It can be seen that customers who experience churn (“yes”) tend to have a much shorter service time than customers who do not experience churn (“no”). This can be indicated if customers who tend to switch (churn) have an unstable nature

or are not loyal to the provider, as evidenced by the service time usage which tends to be very small.

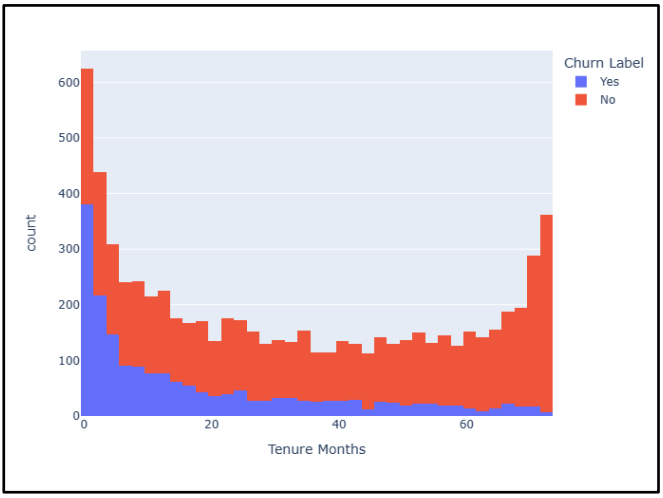


Fig. 5. Tenure Months Distribution based on Churn Label

4.6 Total Charges based on Churn Label

It shows a marginal box that indicates the median, quartiles, and outliers. It appears that the total spending of customers who have churned tends to be significantly lower compared to customers who have not switched. This is in line with the previous analysis of usage time (tenure months).

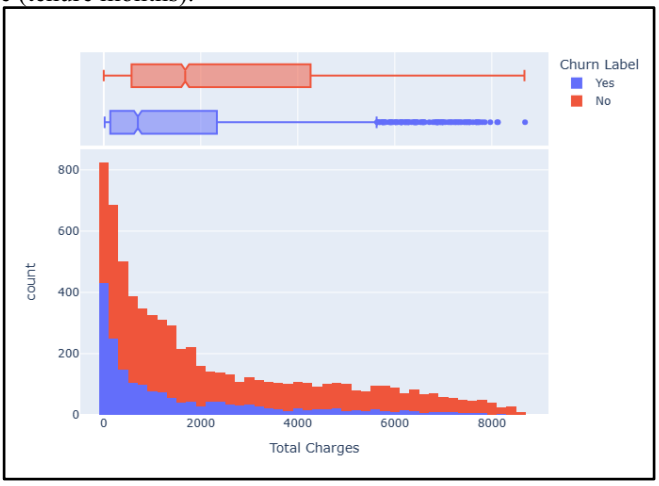


Fig. 6. Marginal Box of Total Charges based on Churn Label

4.7 Kaplan Meier Analysis

In this analysis, an event consisting of only one occurrence is used (0 = No Churn, 1 = Churn) because the Kaplan-Meier method is specifically employed to estimate the survival function for single data or data that only has one type of event or risk.

4.7.1. Kaplan Meier with Gender Factor

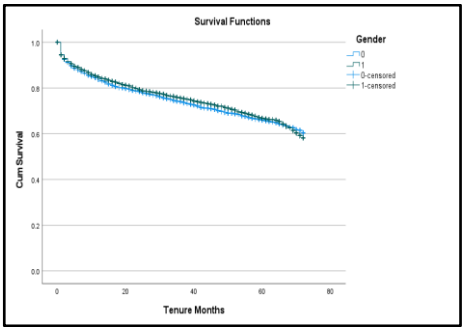
The results of the Log-Rank test, Survival Plot, and Hazard plot based on the Gender Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is no significant difference in the distribution of churn probabilities between the male (1) and female groups (0).

Overall Comparisons

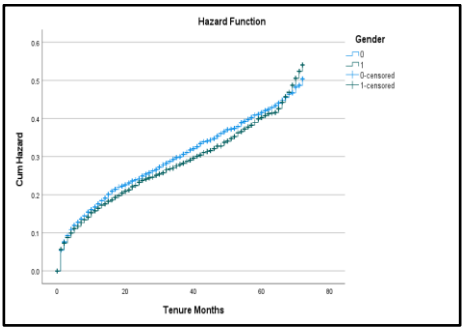
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.526	1	.468

Test of equality of survival distributions for the different levels of Gender.

(a)



(b)



(c)

Fig. 7. Kaplan Meier On Gender Factor: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.2. Kaplan Meier with Senior Citizen Factor

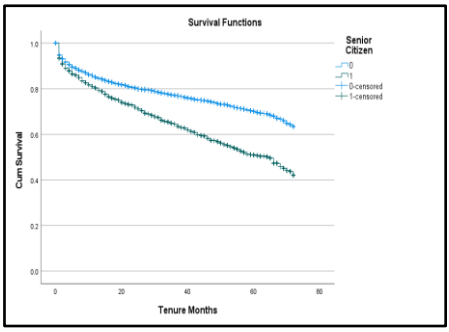
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Senior Citizen Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the senior citizen group (1) and those who are not (0).

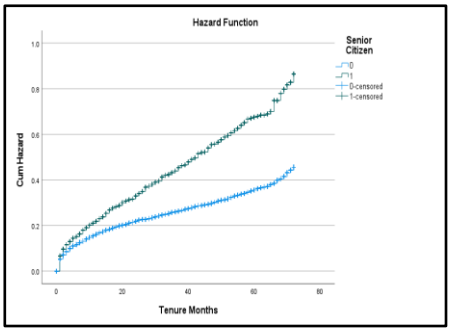
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	109.490	1	<,001

Test of equality of survival distributions for the different levels of Senior Citizen.

(a)





(b) (c)

Fig. 8. Kaplan Meier On Senior Citizen: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.3. Kaplan Meier with Partner Factor

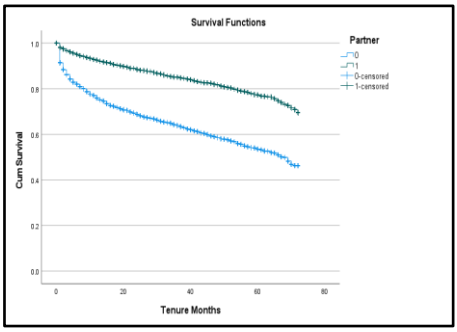
It shows the results of the Log-Rank test, Survival Plot, and Hazard Plot based on the With Partner factor. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the group with partners (1) and those without (0).

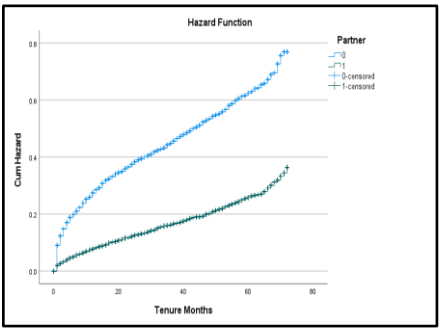
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	423.543	1	<,001

Test of equality of survival distributions for the different levels of Partner.

(a)





(b) (c)

Fig. 9. Kaplan Meier On Partner: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.4. Kaplan Meier with Phone Service Factor

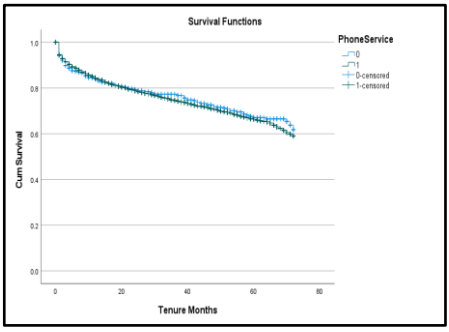
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Phone Service Factor are presented. The results of this test indicate that at the specified significance level (α = 0.05), there is no significant difference in the distribution of churn probabilities between the group that uses phone services (1) and the group that does not (0).

Overall Comparisons

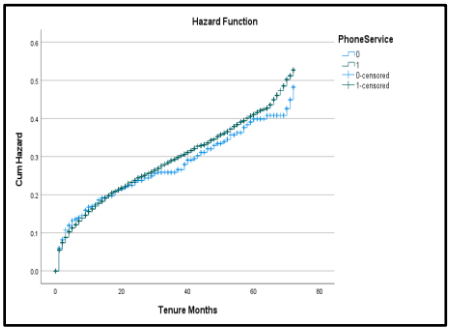
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.431	1	.512

Test of equality of survival distributions for the different levels of PhoneService.

(a)



(b)



(c)

Fig. 10. Kaplan Meier On Phone Service: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.5. Kaplan Meier with Multiple Lines Factor

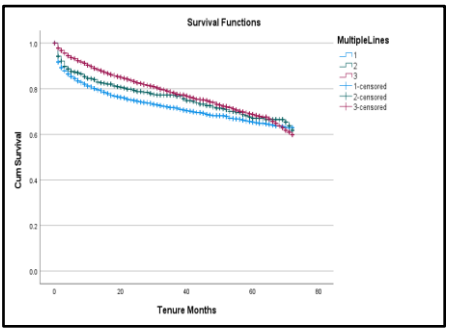
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on Multiple Lines Factor are presented. The results of this test indicate that at the specified significance level (α = 0.05), there is a significant difference in the distribution of churn probabilities between the group that has multiple telephone lines (3), those without telephone service (2), and those without multiple telephone lines (1).

Overall Comparisons

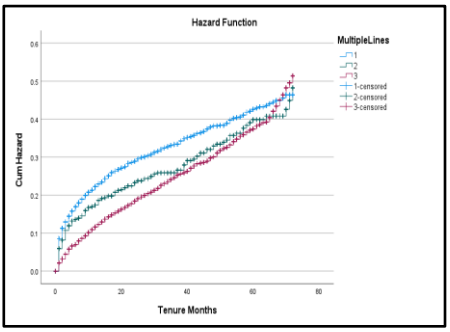
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	30.969	2	<,001

Test of equality of survival distributions for the different levels of MultipleLines.

(a)



(b)



(c)

Fig. 11. Kaplan Meier On Multiple Lines: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.6. Kaplan Meier with Internet Service Factor

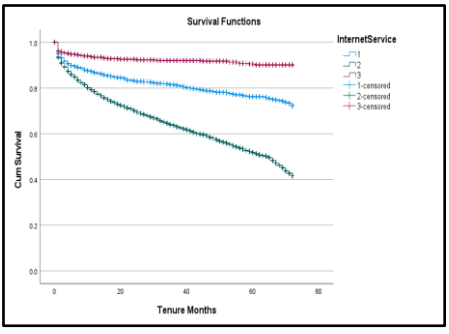
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on Internet Service Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities among the groups that use DSL-internet service (1), Fiber Optic (2), and those who do not use internetservices (3).

Overall Comparisons

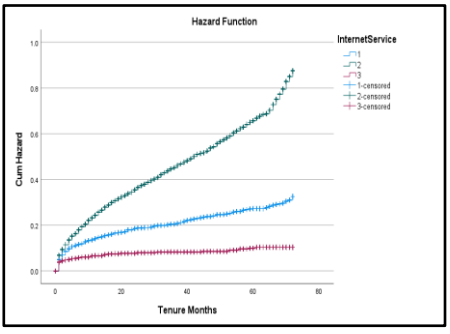
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	520.121	2	<,001

Test of equality of survival distributions for the different levels of InternetService.

(a)



(b)



(c)

Fig. 12. Kaplan Meier On Internet Service: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.7. Kaplan Meier with Online Security Factor

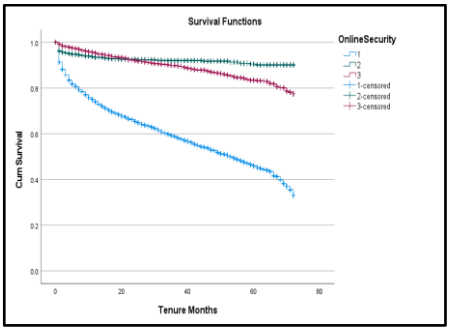
It shows the results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Online Security Factor. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the group without online security (1), the group without internetservice (2), and the group with online security (3).

Overall Comparisons

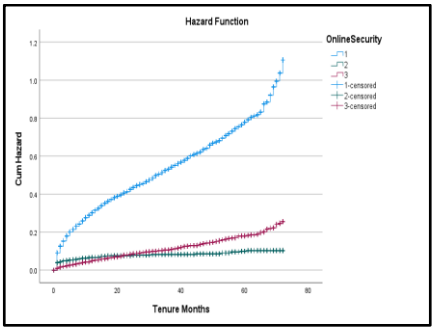
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1013.865	2	<,001

Test of equality of survival distributions for the different levels of OnlineSecurity.

(a)



(b)



(c)

Fig. 13. Kaplan Meier On Online Security: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.8. Kaplan Meier with Online Backup Factor

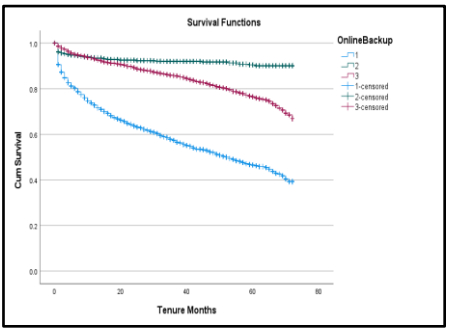
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Online Backup Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the group without online backup (1), the group without internetservice (2), and the group with online backup (3).

Overall Comparisons

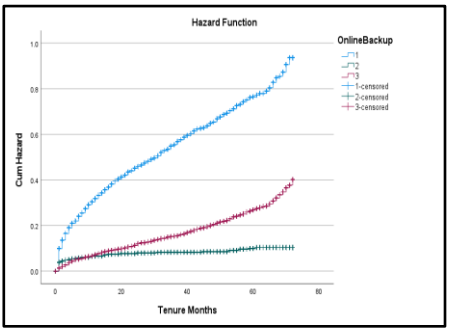
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	821.339	2	<,001

Test of equality of survival distributions for the different levels of OnlineBackup.

(a)



(b)



(c)

Fig. 14. Kaplan Meier On Online Backup: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.9. Kaplan Meier with Device Protection Factor

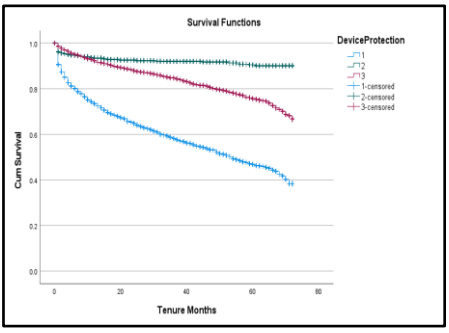
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Device Protection Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the group without device protection (1), the group without internetservice (2), and the group with device protection (3).

Overall Comparisons

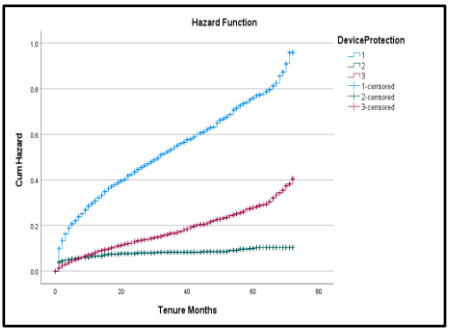
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	763.506	2	<,001

Test of equality of survival distributions for the different levels of DeviceProtection.

(a)



(b)



(c)

Fig. 15. Kaplan Meier On Device Protection: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.10. Kaplan Meier with Tech Support Factor

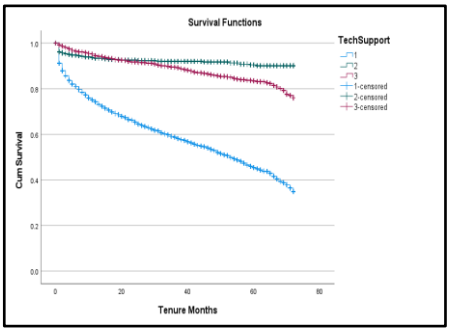
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Tech Support Factor are presented. The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Tech Support Factor are presented.

Overall Comparisons

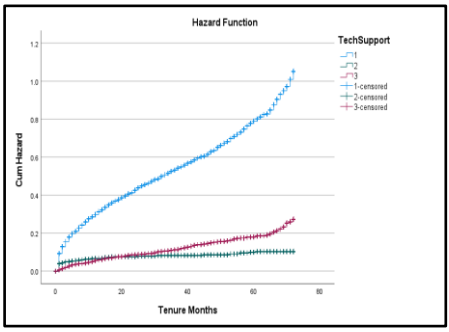
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	989.560	2	<,001

Test of equality of survival distributions for the different levels of TechSupport.

(a)



(b)



(c)

Fig. 16. Kaplan Meier On Tech Support: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.11. Kaplan Meier with Streaming TV Factor

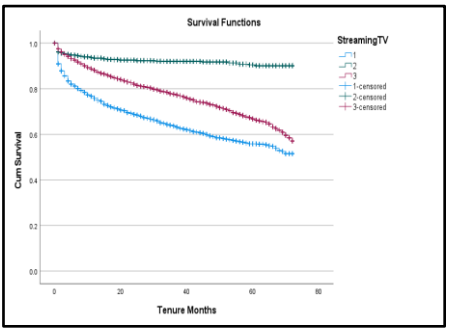
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Streaming TV Factor are presented. The results of this test indicate that at the specified significance level (α = 0.05), there is a significant difference in the distribution of churn probabilities between the group without tv-streaming (1), the group without internetservice (2), and the group with TV streaming (3).

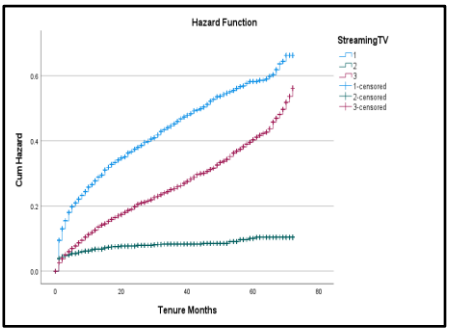
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	368.307	2	<,001

Test of equality of survival distributions for the different levels of StreamingTV.

(a)





(b) (c)

Fig. 17. Kaplan Meier On Streaming TV: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.12. Kaplan Meier with Streaming Movie Factor

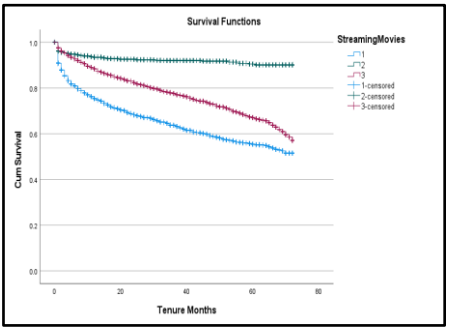
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Streaming Movie Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities between the group without movie streaming (1), the group without internetservice (2), and the group with movie streaming (3).

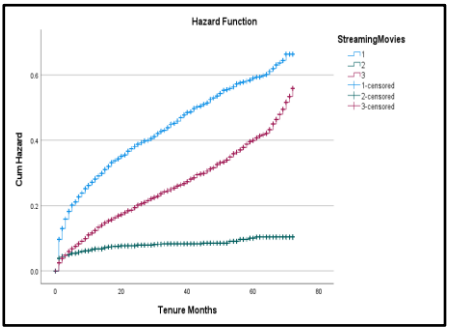
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	378.426	2	<,001

Test of equality of survival distributions for the different levels of StreamingMovies.

(a)





(b) (c)

Fig. 18. Kaplan Meier On StreamingMovies: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.13. Kaplan Meier with Paperless Billing Factor

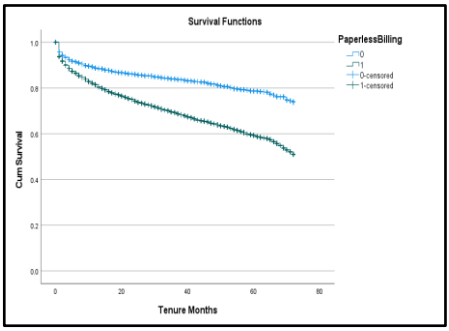
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Paperless Billing Factor are presented. The results of this test indicate that at the specified significance level (α = 0.05), there is a significant difference in the distribution of churn probabilities between the group without paperless billing (0) and the group with paperless billing (1).

Overall Comparisons

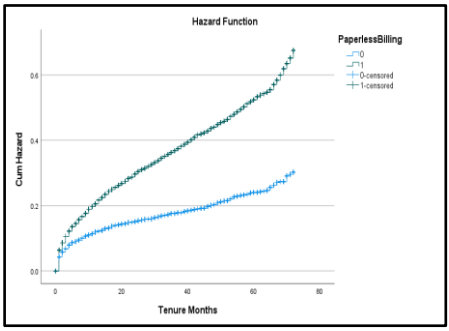
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	189.511	1	<,001

Test of equality of survival distributions for the different levels of PaperlessBilling.

(a)



(b)



(c)

Fig. 19. Kaplan Meier On Paperless billing: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.14. Kaplan Meier with Payment Method Factor

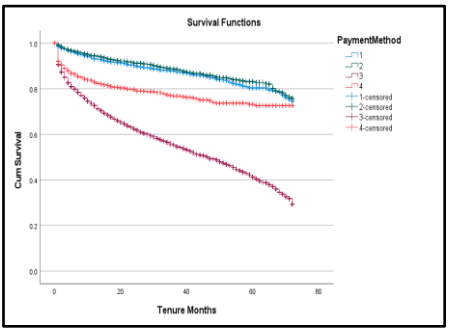
It shows the results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Payment Method Factor. The results of this test indicate that at the specified significance level (α = 0.05), there is a significant difference in the distribution of churn probabilities among the groups that pay by bank transfer (1), creditcard (2), electronic check (3), and email check (4).

Overall Comparisons

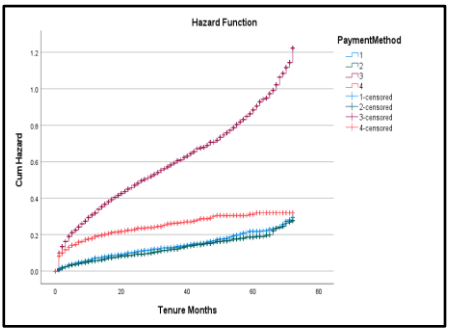
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	865.239	3	<,001

Test of equality of survival distributions for the different levels of PaymentMethod.

(a)



(b)



(c)

Fig. 20. Kaplan Meier On Payment method: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.7.15. Kaplan Meier with Contract Factor

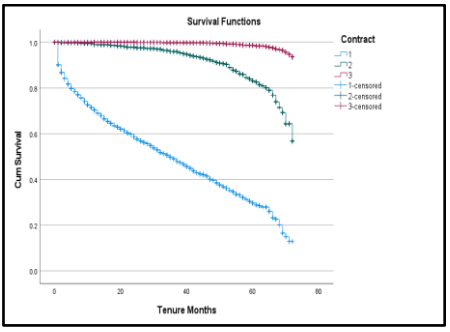
The results of the Log-Rank test, Survival Plot, and Hazard Plot based on the Contract Factor are presented. The results of this test indicate that at the specified significance level ($\alpha = 0.05$), there is a significant difference in the distribution of churn probabilities among the groups with monthly contracts (1), one-year contracts (2), and two-year contracts (3).

Overall Comparisons

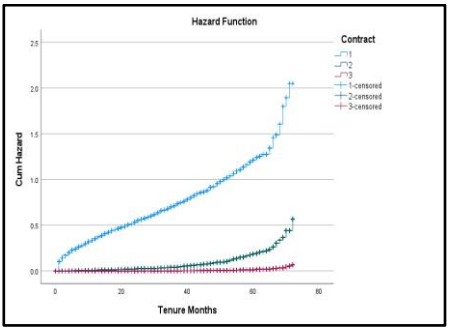
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	2352.873	2	.000

Test of equality of survival distributions for the different levels of Contract.

(a)



(b)



(c)

Fig. 21. Kaplan Meier On Contract: (a) Overall Comparison, (b) Survival Plot and (c) Hazard Plot

4.8 Stratified Cox Proportional Hazard

4.8.1. Coefficients and Stratified Cox PH Equation

Since there are several variables that are not significant to the survival and hazard functions, modeling will be conducted using Stratified Cox PH.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
`Senior Citizen`1	-0.06206	0.93983	0.05573	-1.113	0.26550	
Partner1	-0.54400	0.58042	0.05065	-10.741	< 2e-16	***
Contract2	-1.61162	0.19956	0.08837	-18.237	< 2e-16	***
Contract3	-3.22976	0.03957	0.16617	-19.436	< 2e-16	***
MultipleLines2	NA	NA	0.00000	NA	NA	
MultipleLines3	-0.46816	0.62615	0.05350	-8.750	< 2e-16	***
InternetService2	0.38543	1.47025	0.07244	5.320	1.04e-07	***
InternetService3	-1.10107	0.33251	0.12076	-9.118	< 2e-16	***
OnlineSecurity2	NA	NA	0.00000	NA	NA	
OnlineSecurity3	-0.65559	0.51913	0.06638	-9.876	< 2e-16	***
OnlineBackup2	NA	NA	0.00000	NA	NA	
OnlineBackup3	-0.65838	0.51769	0.05484	-12.006	< 2e-16	***
DeviceProtection2	NA	NA	0.00000	NA	NA	
DeviceProtection3	-0.32012	0.72607	0.05486	-5.836	5.36e-09	***
TechSupport2	NA	NA	0.00000	NA	NA	
TechSupport3	-0.40926	0.66414	0.06547	-6.251	4.07e-10	***
StreamingTV2	NA	NA	0.00000	NA	NA	
StreamingTV3	-0.04105	0.95978	0.05455	-0.752	0.45177	
StreamingMovies2	NA	NA	0.00000	NA	NA	
StreamingMovies3	-0.12807	0.87979	0.05488	-2.334	0.01961	*
PaperlessBilling1	0.18405	1.20208	0.05677	3.242	0.00119	**
PaymentMethod2	-0.08891	0.91493	0.09085	-0.979	0.32778	
PaymentMethod3	0.58425	1.79364	0.07141	8.182	2.80e-16	***
PaymentMethod4	0.55344	1.73923	0.08837	6.263	3.79e-10	***

Fig. 22. Stratified Cox PH Modelling

In this modeling, strata were used for the variables *gender* and *phone services* because they did not meet the Kaplan-Meier test criteria. The resulting model equation is as follows:

$$
\begin{aligned}
h_s\ (t, X)\ =\ h_{0s}\ (t)\ exp\ (&-0.544\ Partner\ -\ 1.61162\ Contract\ one\ year \\
&-\ 3.22976\ Contract\ Two\ Year\ -\ 0.46816\ Multiple\ Lines \\
&+\ 0.38543\ Fiber\ Optic\ Internet\ Service \\
&-\ 1.10107\ No\ Internet\ Service\ -\ 0.65559\ Online\ Security \\
&-\ 0.65838\ Online\ Backup\ -\ 0.32012\ Device\ Protection \\
&-\ 0.40926\ Tech\ Support\ -\ 0.04105\ Streaming\ TV \\
&-\ 0.12807\ Streaming\ Movies\ +\ 0.18405\ Paperless\ Billing \\
&-\ 0.08891\ Credit\ Card\ Payment\ Method \\
&+\ 0.58425\ Electronic\ Check\ Payment\ Method \\
&+\ 0.55344\ Mailed\ Check\ Payment\ Method)
\end{aligned}
$$

From the results, we can interpret the following:

- A person with a partner has a 0.58042 times lower chance of churning compared to someone without a partner.
- A person with a one-year contract has a 0.19956 times lower chance of churning compared to someone with a monthly contract.
- A person with a two-year contract has a 0.03957 times lower chance of churning compared to someone with a monthly contract.
- A person with multiple lines has a 0.62615 times lower chance of churning compared to someone without multiple lines.

- A person with fiber optic internet service has a 1.47025 times higher chance of churning compared to someone with DSL internet.
- A person without internet service has a 0.33251 times lower chance of churning compared to someone with DSL internet.
- A person with online security has a 0.51913 times lower chance of churning compared to someone without online security.
- A person with online backup has a 0.51769 times lower chance of churning compared to someone without online backup.
- A person with device protection has a 0.72607 times lower chance of churning compared to someone without device protection.
- A person with tech support has a 0.66414 times lower chance of churning compared to someone without tech support.
- A person with streaming TV has a 0.95978 times lower chance of churning compared to someone without streaming TV.
- A person with streaming movie services has a 0.87979 times lower chance of churning compared to someone without streaming movies.
- A person using paperless billing has a 1.20208 times higher chance of churning compared to someone who does not use paperless billing.
- A person who pays by credit card has a 0.91493 times lower chance of churning compared to someone who pays via bank transfer.
- A person who pays by electronic check has a 1.79364 times higher chance of churning compared to someone who pays via bank transfer.
- A person who pays by mailed check has a 1.73923 times higher chance of churning compared to someone who pays via bank transfer.

4.8.2. Goodness of Fit

Concordance is a measure of how well the model can correctly distinguish between individuals who churn and those who do not. The concordance value ranges from 0 to 1, with 1 indicating a model that perfectly distinguishes between these pairs. In this case, the concordance value is 0.867, suggesting that the model has a good ability to differentiate between churn cases.

```
Concordance= 0.867  (se = 0.003 )
Likelihood ratio test= 3530  on 17 df,   p=<2e-16
Wald test            = 2054  on 17 df,   p=<2e-16
Score (logrank) test = 3211  on 17 df,   p=<2e-16
```

Fig. 23. Goodness of Fit Result

The likelihood ratio test assesses whether the overall model provides a significant explanation of the data compared to the null model (without predictor variables). A very low p-value ($p < 2e\text{-}16$) indicates that the model is highly statistically significant overall.

The Wald test evaluates the significance of each predictor variable's coefficient in the model. A very low p-value ($p < 2e\text{-}16$) indicates that at least one of the predictor variables has a significant effect on the hazard.

The log-rank test is used to compare the survival functions between groups or categories of variables. A very low p-value ($p < 2e\text{-}16$) shows significant differences in survival functions between groups, and the model can generally distinguish between these groups.

4.9 Competing Risk

In this competing risk analysis, we specified and categorized the events based on existing churn reasons. Internal factors include 'Attitude of service provider', 'Attitude of support person', 'Extra data charges', 'Lack of affordable download/upload speed', 'Lack of self-service on Website', 'Limited range of services', 'Long distance charges', 'Moved', 'Network reliability', 'Poor expertise of online support', 'Poor expertise of phone support', 'Price too high', 'Product dissatisfaction', 'Service dissatisfaction'.

External factors include 'Competitor had better devices', 'Competitor made better offer', 'Competitor offered higher download speeds', 'Competitor offered more data', 'Deceased', 'Don't know'. For events caused by internal factors, the churn value is 1, while for events caused by external factors, the churn value is 2.

4.9.1. Estimates and variances for the cumulative incidence function (CIF) at specific time points

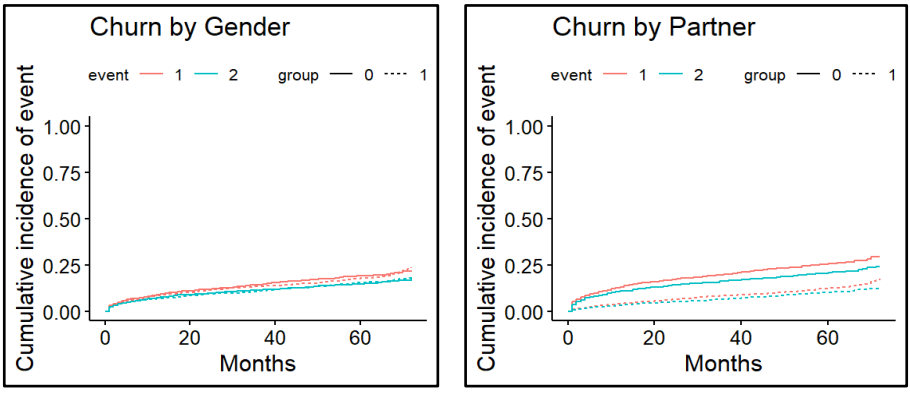
```
Estimates and Variances:
$est
            20        40        60
1 1 0.1073444 0.1472046 0.1845265
1 2 0.0881357 0.1182981 0.1510696

$var
                20           40           60
1 1 1.513680e-05 2.231468e-05 3.267774e-05
1 2 1.272513e-05 1.846176e-05 2.778967e-05
```

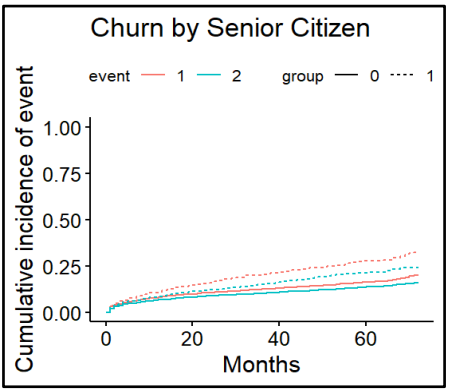
Fig. 24. Estimation and Variance from Competing Risk

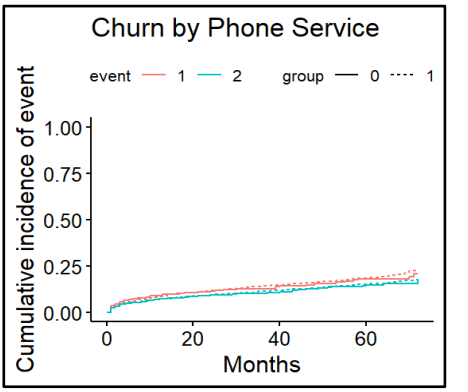
The results indicate estimates and variances for the cumulative incidence function (CIF) at specific time points. For example, at time 20 for event 1, the variance is 1.513680e-05.

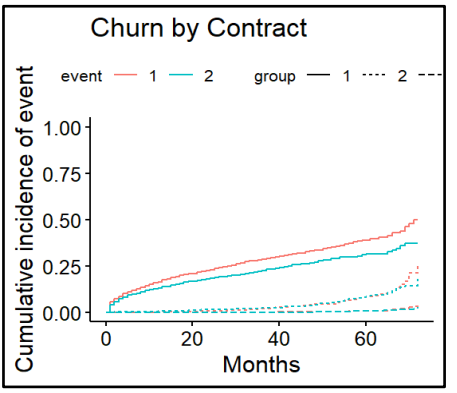
4.9.2. CIF Plot based on Factor

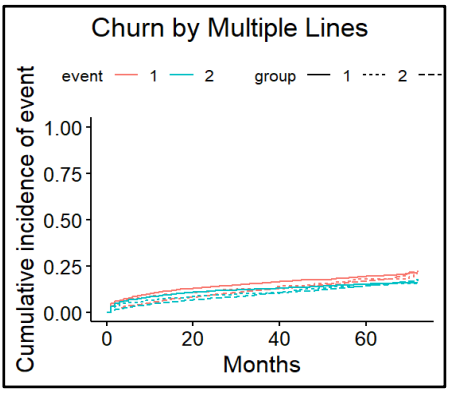
The CIF plot indicates that factors such as gender and phone service do not have a significant influence or distinguish between events caused by poor provider service or better offers from other providers.

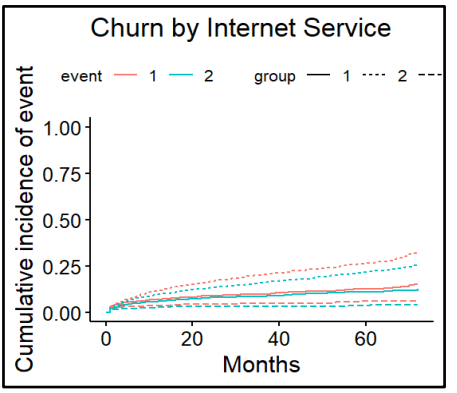


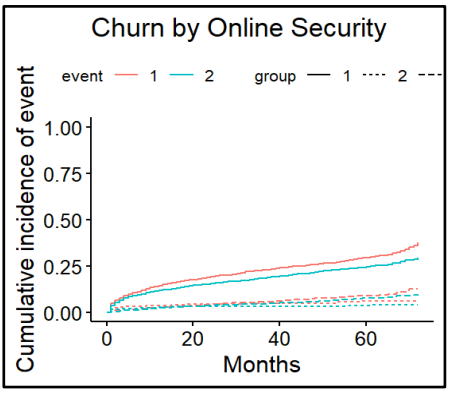

Churn by Senior Citizen
event 1 2 group 0 1
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

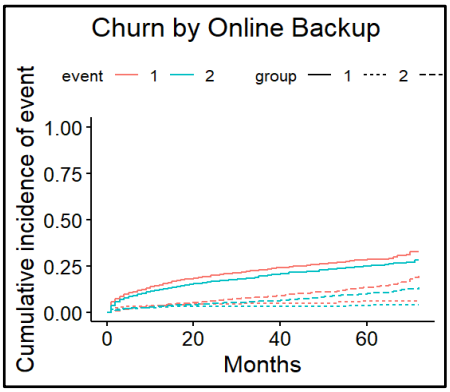

Churn by Phone Service
event 1 2 group 0 1
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

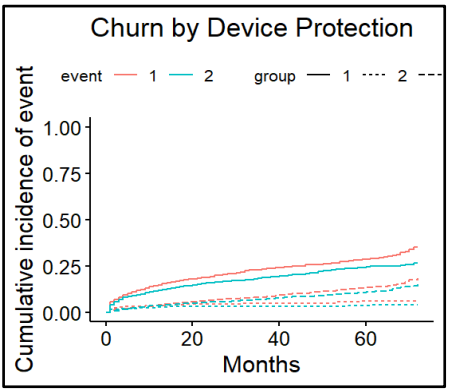

Churn by Contract
event 1 2 group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

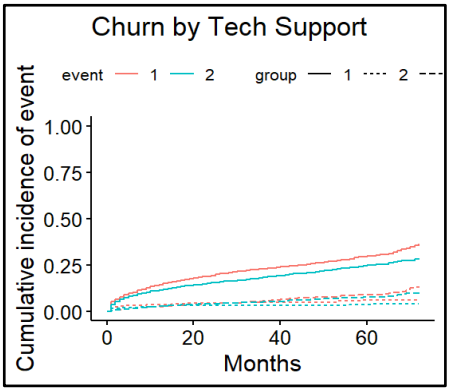

Churn by Multiple Lines
event 1 2 group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

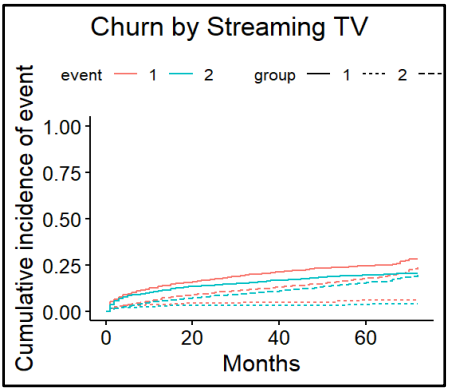

Churn by Internet Service
event 1 2 group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

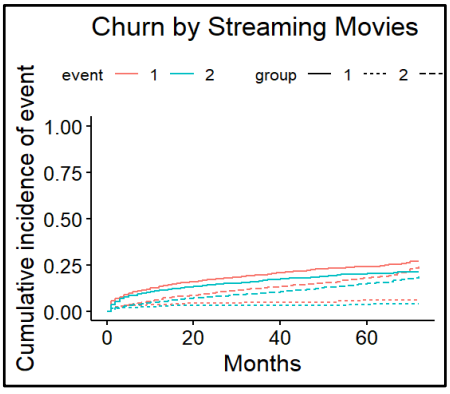

Churn by Online Security
event 1 2 group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

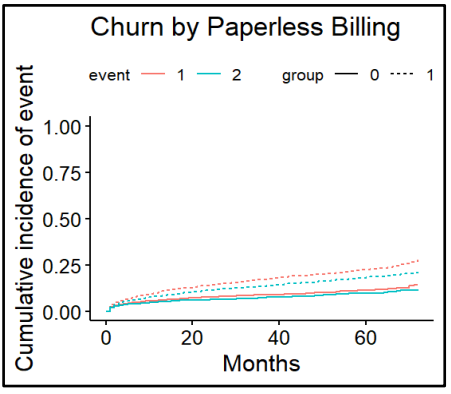

Churn by Online Backup
event 1 2
group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months


Churn by Device Protection
event 1 2
group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months


Churn by Tech Support
event 1 2
group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months


Churn by Streaming TV
event 1 2
group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months


Churn by Streaming Movies
event 1 2
group 1 2
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months


Churn by Paperless Billing
event 1 2
group 0 1
Cumulative incidence of event
1.00
0.75
0.50
0.25
0.00
0 20 40 60
Months

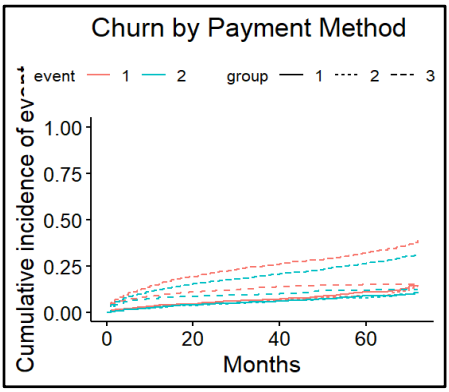


Fig. 25. CIF Plot By Factor

4.9.3. Competing Risk Regression Based on Event

	coef	exp(coef)	se(coef)	z	p-value
Gender	-0.1057	0.900	0.0634	-1.6667	9.6e-02
Senior Citizen	0.0169	1.017	0.0768	0.2199	8.3e-01
Partner	-0.4952	0.609	0.0688	-7.2014	6.0e-13
Contract	-1.7710	0.170	0.0840	-21.0740	0.0e+00
PhoneService	0.0083	1.008	0.1265	0.0656	9.5e-01
MultipleLines	-0.1869	0.830	0.0357	-5.2298	1.7e-07
InternetService	0.1642	1.178	0.0623	2.6370	8.4e-03
OnlineSecurity	-0.3618	0.696	0.0464	-7.8041	6.0e-15
OnlineBackup	-0.2934	0.746	0.0373	-7.8579	4.0e-15
DeviceProtection	-0.1734	0.841	0.0380	-4.5618	5.1e-06
TechSupport	-0.2403	0.786	0.0450	-5.3343	9.6e-08
StreamingTV	-0.0234	0.977	0.0374	-0.6253	5.3e-01
StreamingMovies	-0.0282	0.972	0.0372	-0.7569	4.5e-01
PaperlessBilling	0.2883	1.334	0.0781	3.6938	2.2e-04
PaymentMethod	0.2342	1.264	0.0354	6.6104	3.8e-11

Fig. 26. Competing Risk Regression for Event 1

From the above results, we obtained the equation for event type 1 as follows:

$$h(t|X) = h_0(t) \times exp(-0.4952 \times Partner - 1.7710 \times Contract - 0.1869 \times MultipleLines + 0.1642 \times InternetService - 0.36180 \times OnlineSecurity - 0.2934 \times OnlineBackup - 0.1734 \times DeviceProtection - 0.2403 \times TechSupport + 0.2883 \times PaperlessBilling + 0.2342 \times PaymentMethod)$$

	coef	exp(coef)	se(coef)	z	p-value
Gender	-0.0583	0.943	0.0702	-0.830	4.1e-01
Senior Citizen	-0.0442	0.957	0.0859	-0.514	6.1e-01
Partner	-0.5123	0.599	0.0760	-6.743	1.6e-11
Contract	-1.6401	0.194	0.0866	-18.930	0.0e+00
PhoneService	0.1000	1.105	0.1380	0.725	4.7e-01
MultipleLines	-0.1621	0.850	0.0387	-4.193	2.8e-05
InternetService	0.0574	1.059	0.0669	0.859	3.9e-01
OnlineSecurity	-0.3876	0.679	0.0511	-7.589	3.2e-14
OnlineBackup	-0.3789	0.685	0.0434	-8.738	0.0e+00
DeviceProtection	-0.1689	0.845	0.0418	-4.045	5.2e-05
TechSupport	-0.2416	0.785	0.0502	-4.813	1.5e-06
StreamingTV	0.0363	1.037	0.0403	0.901	3.7e-01
StreamingMovies	-0.0506	0.951	0.0411	-1.231	2.2e-01
PaperlessBilling	0.2296	1.258	0.0839	2.737	6.2e-03
PaymentMethod	0.2359	1.266	0.0389	6.070	1.3e-09

Fig. 27. Competing Risk Regression for Event 2

Similarly, for event type 2, we obtained the following equation:

$$h(t|X) = h_0(t) \times exp(-0.5123 \times Partner - 1.6401 \times Contract - 0.1621 \times MultipleLines - 0.3876 \times OnlineSecurity - 0.3789 \times OnlineBackup - 0.1689 \times DeviceProtection - 0.2416 \times TechSupport + 0.2296 \times PaperlessBilling + 0.2359 \times PaymentMethod)$$

Next, we aim to determine the hazard ratio for these two equations. To find the hazard ratio, we need to sum the exponential coefficients of each equation and compare them, resulting in the following:

$$HR = \frac{0.609 + 0.170 + \ldots\ldots + 1.264}{0.599 + 0.194 + \ldots\ldots + 1.266}$$

Thus, a Hazard Ratio (HR) of 1.6 is obtained, meaning that the likelihood of churn for a customer is 1.6 times higher due to factors related to the serviceprovider (such as poor signal, etc.) compared to external factors (such as competitor advantages).

5 Conclusion

Based on the analysis conducted on the IBM dataset regarding customer churn in the telecommunications industry, several conclusions can be drawn. Initial exploration reveals that a significant portion of the data is censored, with the "No" label for churn accounting for 73.5%. The reasons for customer churn can be categorized into internal and external influences. The Competing Risk method was applied to understand events driven by both internal and external factors, providing deeper insights into the reasons customer's churn.

Subscription duration analysis shows that customers who churn tend to have shorter subscription periods and lower total expenditures. The Kaplan-Meier method also demonstrated significant differences in survival distributions across various variables, including contract type, additional services, and certain demographic factors. The Stratified Cox Proportional Hazard method confirmed that variables such as having a partner, contract type, multiple lines, and additional services have a significant impact on churn risk. The application of the Goodness of Fit test further validated that the survival analysis model has strong predictive capability in distinguishing churn cases, with a concordance value reaching 0.867. Additionally, the CIF (Cumulative Incidence Function) plot showed that certain factors, such as gender and phone service, do not significantly differentiate events driven by internal or external factors. Lastly, the competing risk analysis revealed that customers receiving poor service from the provider have a 1.6 times higher risk of churning compared to those who switch due to competition from other providers.

The implications of this study suggest that companies can more effectively manage churn by focusing on improving service quality, developing more flexible contract options, and creating personalized customer retention strategies. This research not only provides deep insights into customer behavior in the context of churn but also lays the groundwork for companies to optimize their business strategies to better address these challenges and improve decision-making in the future.

Disclosure of Interests. The authors declares that this paper has no competing interests.

References

1. D. Slof, F. Frasincar, and V. Matsiiako, "A competing risks model based on latent Dirichlet allocation for predicting churn reasons," *Decision Support Systems*, **146**, 113541 (2021). doi:10.1016/j.dss.2021.113541
2. S. K. Wagh *et al.*, "Customer churn prediction in telecom sector using Machine Learning Techniques," *Results in Control and Optimization*, **14**, 100342 (2024). doi:10.1016/j.rico.2023.100342
3. K. Dhangar, "A Review on Customer Churn Prediction Using Machine Learning Approach," *International Journal of Innovations in Engineering Research and Technology*, **8**(5), 193–201 (2021). doi:10.17605/OSF.IO/ACNKJ
4. Y. Chen, L. Zhang, Y. Zhao, and B. Xu, "Implementation of penalized survival models in churn prediction of vehicle insurance," *Journal of Business Research*, **153**, 162–171 (2022). doi:10.1016/j.jbusres.2022.07.015
5. S. Hu, P. Chen, and X. Chen, "Do personalized economic incentives work in promoting shared mobility? examining customer churn using a time-varying Cox Model," *Transportation Research Part C: Emerging Technologies*, **128**, 103224 (2021). doi:10.1016/j.trc.2021.103224
6. E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, **53**(282), 457–481 (1958). doi:10.1080/01621459.1958.10501452
7. J. Kishore, M. Goel, and P. Khanna, "Understanding survival analysis: Kaplan-Meier estimate," *International Journal of Ayurveda Research*, **1**(4), 274 (2010). doi:10.4103/0974-7788.76794
8. J. T. Rich *et al.*, "A practical guide to understanding kaplan-meier curves," *Otolaryngology–Head and Neck Surgery*, **143**(3), 331–336 (2010). doi:10.1016/j.otohns.2010.05.007
9. M. Mohamed Ahmed Abdelaal, "Modeling survival data by using cox regression model," *American Journal of Theoretical and Applied Statistics*, **4**(6), 504–512 (2015). doi:10.11648/j.ajtas.20150406.21
10. E. T. Lee and J. W. Wang, in *Statistical methods for survival data analysis*, New York: J. Wiley, 2003, pp. 348–352
11. Z. Zhang, "Survival analysis in the presence of competing risks," *Annals of Translational Medicine*, **5**(3), 47–47 (2017). doi:10.21037/atm.2016.08.62
12. J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, **94**(446), 496–509 (1999). doi:10.1080/01621459.1999.10474144

13. D. Suhartono, A. Saefuddin, and I. M. Sumertajaya, "Survival Analysis Of Customer In Postpaid Telecommunication Industry," *Indonesian Journal of Statistics and Its Applications*, **18**(1), 1–10 (2013).
14. M. Masarifoglu and A. Hakan Buyuklu, "Applying survival analysis to Telecom Churn Data," *American Journal of Theoretical and Applied Statistics*, **8**(6), 261 (2019). doi:10.11648/j.ajtas.20190806.18
15. A. V. Monika, Indahwati, and M. N. Aidi, "Churn analysis in telecommunication industry customers using semiparametric and non parametric survival method," *Journal of Physics: Conference Series*, **1863** (1), 012034 (2021). doi:10.1088/1742-6596/1863/1/012034
16. N. Alboukaey, A. Joukhadar, and N. Ghneim, "Dynamic behavior based churn prediction in Mobile Telecom," *Expert Systems with Applications*, **162**, 113779 (2020). doi:10.1016/j.eswa.2020.113779
17. S. M. Shrestha and A. Shakya, "A customer churn prediction model using XGBoost for the telecommunication industry in Nepal," *Procedia Computer Science*, **215**, 652–661 (2022). doi:10.1016/j.procs.2022.12.067
18. A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes," *Applied Soft Computing*, **137**, 110103 (2023). doi:10.1016/j.asoc.2023.110103
19. M. M. Uner, F. Guven, and S. T. Cavusgil, "Churn and loyalty behavior of Turkish digital natives: Empirical insights and managerial implications," *Telecommunications Policy*, **44** (4), 101901 (2020). doi:10.1016/j.telpol.2019.101901