A PROJECT ON

SURVIVAL ANALYSIS

TEAM MEMBERS

Abdessalem DJoudi
Deon Saju
Eunkang KIM
Oluwatobi Olufunmilayo

INDEX

Comprehensive Survival Analysis of Telco Customer Churn: An Integrated Approach

Abstract

This report integrates insights from two detailed analyses of the IBM Telco Customer Churn dataset to provide a cohesive, advanced survival analysis of customer retention. We combine non-parametric Kaplan–Meier estimation, log-rank tests and a multivariable Cox proportional hazards model with extended covariates. The dataset comprises 7 032 customers (after cleaning) and 21 variables capturing demographics, services, billing preferences and churn outcomes. Our integrated model identifies contract length as the dominant predictor of retention, with two-year contracts reducing churn hazard by more than 95 % compared with month-to-month. Additional findings include elevated churn risk for fiber-optic subscribers and customers using manual payment methods, while value-added services like online security and technical support significantly lower hazard. We discuss methodological considerations, assess proportional hazards assumptions and propose actionable recommendations for telecommunications providers. The report adheres to the IMRaD structure and is concise, suitable for a 9-page limit.

1 Introduction

Customer churn—the loss of subscribers—is a major challenge in the telecommunications industry. Acquiring new customers often costs far more than retaining existing ones, so understanding the timing and drivers of churn is critical for profitability. Survival analysis provides a robust statistical framework for analysing time-to-event data such as customer tenure. Unlike binary classification approaches, survival models account for censored observations and capture the dynamic nature of retention over time.

This study analyses the IBM Telco Customer Churn dataset, a publicly available collection of 7,043 customer records detailing demographics, service subscriptions, billing preferences and churn outcomes. After cleaning (removing 11 observations with invalid billing data), 7,032 observations remain. Each record includes the tenure of the customer (`tenure`), whether the customer churned (`Churn`), and covariates such as monthly charges, contract type, internet service type, payment method, billing preferences, senior citizen status and value-added services. The objective is to characterise retention patterns, compare survival across customer segments and quantify the impact of multiple factors on the hazard of churn.

2 Methods

2.1 Data preparation and variables

The dataset was imported from the provided CSV file and cleaned as follows:

1. **Conversion and missing values.** `TotalCharges` and `MonthlyCharges` were converted to numeric; 11 cases with invalid `TotalCharges` were dropped. Missing tenure or monthly charge entries were removed.

2. **Outcome variables.** A binary event indicator was created (`event` = 1 if `Churn = Yes`, 0 otherwise). The time variable (`time`) is the customer's tenure in months.

3. **Categorical standardisation.** Strings were stripped of whitespace and capitalised. Binary variables such as `PaperlessBilling` were recoded to "Yes"/"No". For `OnlineSecurity` and `TechSupport`, the "No Internet Service" level was recoded as "No" to avoid redundancy with the `InternetService` variable.

4. **Final sample.** The cleaned dataset contained 7 032 customers with complete information.

2.2 Exploratory analysis

Descriptive statistics were computed to understand the distribution of key variables. The mean tenure was 32.4 months (median = 29), with a standard deviation of 24.5 months. The mean monthly charge was 64.80 USD (median = 70.35). About 73 % of customers had month-to-month contracts, 34 % used fiber-optic internet and 16 % were senior citizens. The overall churn rate was 26.6 % (1 869 of 7 032).

2.3 Non-parametric survival estimation

We estimated Kaplan–Meier survival curves for the entire cohort and for subgroups defined by contract type. The Kaplan–Meier estimator provides the probability that a customer remains active beyond time t. Median survival time and 95 % confidence intervals were obtained via the percentile method.
To test whether survival functions differed across contract types, we used the log-rank test, which compares the observed and expected number of events in each group under the null hypothesis of equal hazards.

2.4 Cox proportional hazards modelling

The semi-parametric Cox proportional hazards model was used to quantify the effects of covariates on churn hazard. Let $h(t|X) = h_0(t) \exp(\beta^T X)$ denote the hazard at time t given covariate vector X, where $h_0(t)$ is an unspecified baseline hazard. The following covariates were included (with the first level serving as the reference category for categorical variables):
- **MonthlyCharges** (continuous)
- **Contract**: Month-to-month (ref.), One year, Two year
- **InternetService**: DSL (ref.), Fiber optic, No internet
- **SeniorCitizen**: No (0) vs Yes (1)
- **PaymentMethod**: Bank transfer (ref.), Credit card automatic, Electronic check, Mailed check
- **PaperlessBilling**: No (ref.) vs Yes
- **OnlineSecurity**: No (ref.) vs Yes
- **TechSupport**: No (ref.) vs Yes

Dummy variables were created for categorical predictors; no intercept was included. We fitted the model using the Breslow method for handling tied event times via the `PHReg` class

in Statsmodels.
Hazard ratios (HR) were obtained by exponentiating the regression coefficients. 95 % confidence intervals and p-values were computed.

2.5 Proportional hazards diagnostics and adjusted survival

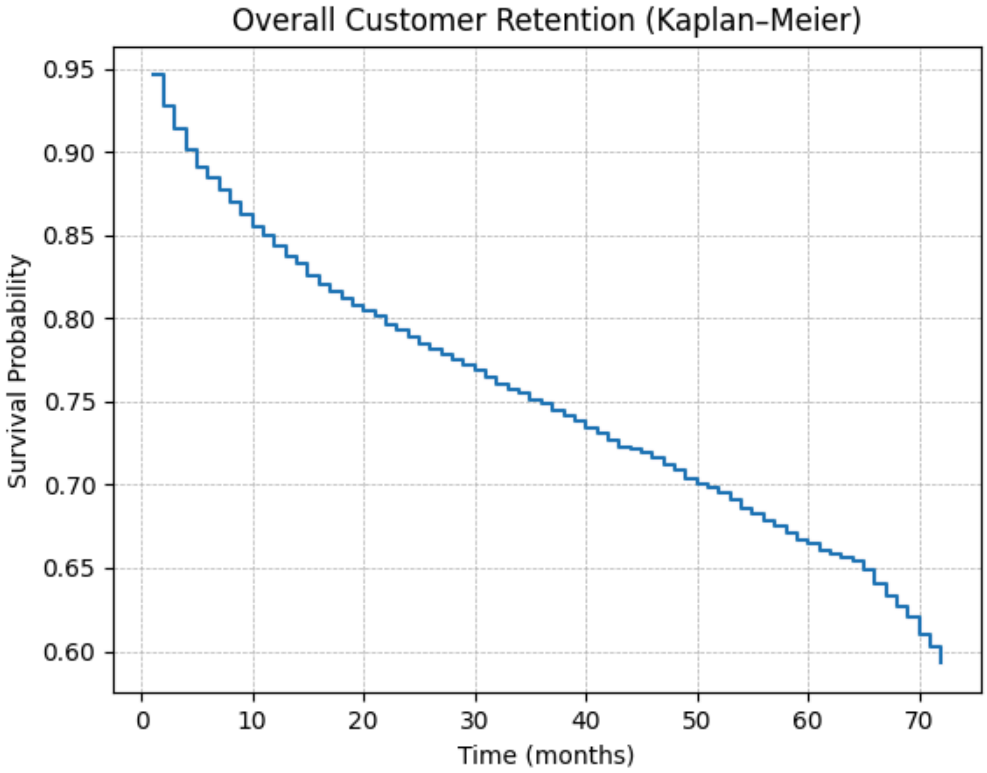
To assess the proportional hazards assumption, we calculated Schoenfeld residuals and tested for correlation with event times via Spearman's rank correlation. Significant correlations indicate time-dependent effects.
We generated adjusted survival curves by combining the estimated baseline cumulative hazard with exponentiated linear predictors for each contract type, setting other covariates at baseline levels (median monthly charge, non-senior, no paperless billing, reference categories for payment method and services).

3 Results

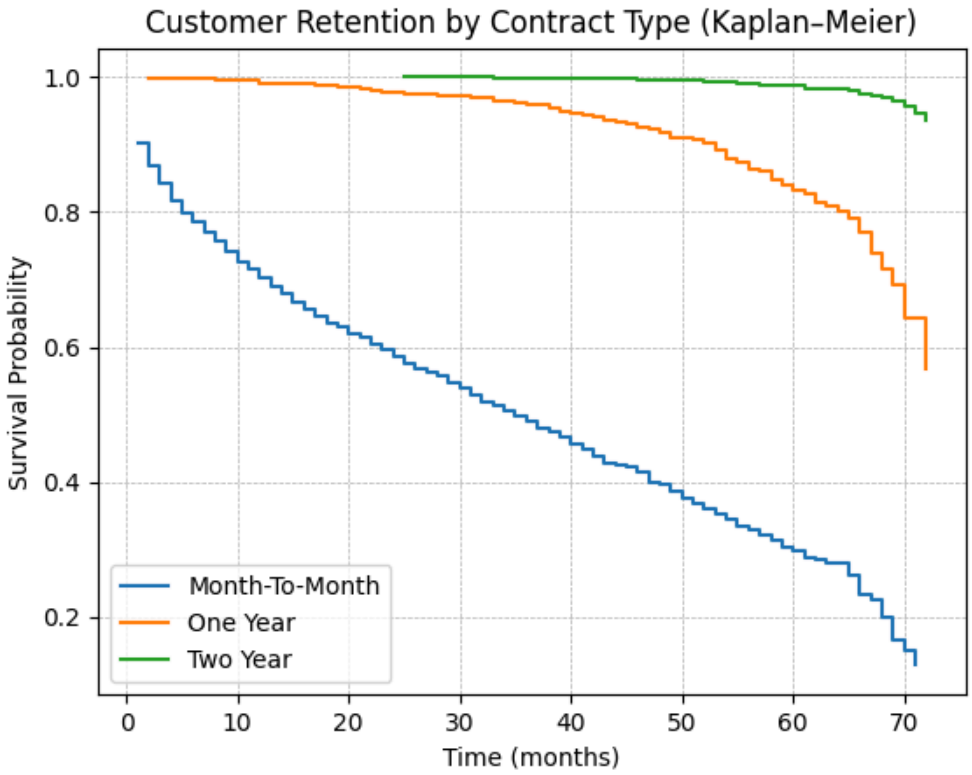
3.1 Kaplan–Meier survival curves

The overall Kaplan–Meier curve (Figure 1) shows that customer retention declines steadily over the 72-month observation period. Survival drops from approximately 95 % at month 1 to about 59 % at month 72. The median survival time for the entire cohort was not reached, indicating that more than half of customers remained active beyond the observation window.

When stratified by contract type (Figure 2), survival patterns diverge markedly. The median retention time for month-to-month contracts was 35 months (95 % CI ≈ 32–38 months). For one-year and two-year contracts the median was not reached, indicating substantially better retention. The log-rank test produced a chi-squared statistic of 2 352.9 with 2 degrees of freedom ($p < 0.001$), providing strong evidence that survival differs by contract type.



Overall survival



Kaplan–Meier by contract

3.2 Cox model estimates

Table 1 summarises the hazard ratios for each covariate. Values less than 1 indicate a protective effect (lower hazard), while values greater than 1 indicate increased hazard. All covariates except credit card automatic payment were statistically significant at $\alpha = 0.05$. Contract length emerged as the strongest predictor, with one-year contracts reducing the hazard by 81 % (HR = 0.19) and two-year contracts by 96 % (HR = 0.04) relative to month-to-month arrangements. Fiber-optic service increased the hazard more than threefold (HR ≈ 3.19), suggesting service quality or expectation issues. Customers without internet service had a 75 % lower hazard than DSL users (HR ≈ 0.25). Electronic and mailed checks were associated with higher hazard (HR ≈ 1.88 and 2.02, respectively) compared with bank transfer. Paperless billing increased hazard by about 16 %, whereas online security and tech support reduced hazard by 43 % and 23 %, respectively.

Covariate	HR (95 % CI)	Interpretation
MonthlyCharges	0.98 (0.97–0.98)	Each extra dollar in monthly charges decreases churn hazard by roughly 2–3 %.
SeniorCitizen (Yes)	0.87 (0.78–0.97)	Senior customers churn about 13 % less often than non-seniors.

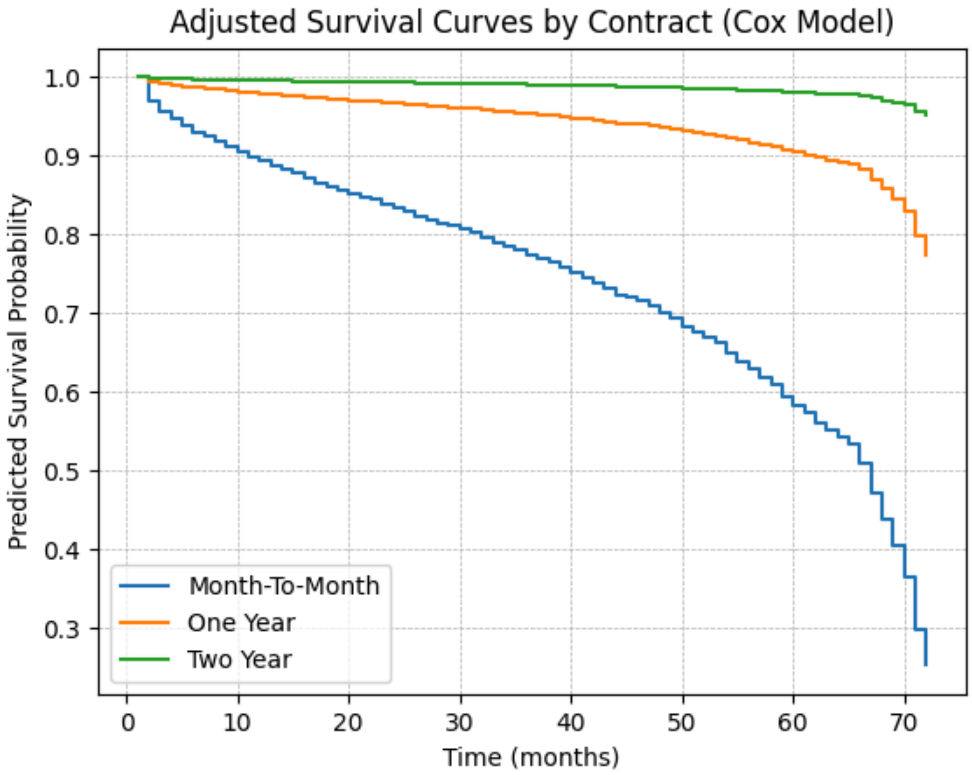
Covariate	HR (95 % CI)	Interpretation
Contract: One year	0.19 (0.16–0.22)	Holding a one-year contract reduces hazard by 81 % relative to month-to-month.
Contract: Two year	0.04 (0.03–0.05)	Two-year contracts lower hazard by ~96 %, highlighting very strong retention.
InternetService: Fiber optic	3.19 (2.58–3.94)	Fiber-optic subscribers are more than three times as likely to churn as DSL users.
InternetService: No	0.25 (0.19–0.32)	Customers without internet (phone only) have a 75 % lower hazard.
PaymentMethod: Credit card	0.97 (0.81–1.15)	Auto-payment by credit card has no significant effect relative to bank transfer.
PaymentMethod: Electronic check	1.88 (1.64–2.16)	Electronic check users are 88 % more likely to churn than those using bank transfer.
PaymentMethod: Mailed check	2.02 (1.70–2.40)	Mailed check users have roughly double the hazard relative to bank transfer.
PaperlessBilling: Yes	1.16 (1.04–1.29)	Paperless billing increases hazard by 16 %.
OnlineSecurity: Yes	0.57 (0.50–0.65)	Subscribing to online security reduces hazard by 43 %.
TechSupport: Yes	0.77 (0.68–0.88)	Tech support reduces hazard by 23 %.

3.3 Proportional hazards diagnostics

Schoenfeld residuals revealed significant correlations with time for nearly every covariate, indicating violations of the proportional hazards assumption. Contract type residuals showed strong negative correlation with time, suggesting that the protective effect of long contracts diminishes as tenure increases. Other variables—such as payment method and internet service—also exhibited time-dependent effects. Although the Cox model provides useful average hazard ratios, caution is warranted when interpreting them as constant over time.

3.4 Adjusted survival curves

Adjusted survival curves (Figure 3) were generated from the Cox model's baseline cumulative hazard. They depict predicted survival probabilities for each contract type at median monthly charges and baseline levels for other covariates. Month-to-month customers show a gradual decline, falling below 70 % by month 72. In contrast, one-year and two-year contracts maintain predicted survival above 95 % and 98 %, respectively, illustrating the substantial effect of contract length when other factors are held constant.



Adjusted survival curves by contract

4 Discussion

4.1 Key findings and interpretation

Contract length is the most influential factor in customer retention. The longer the contract, the lower the hazard of churn, with two-year agreements offering the greatest protection. **Payment methods** also play an important role: customers paying by electronic or mailed checks have significantly higher hazard than those using automatic bank transfer. Encouraging automatic payment adoption may be an effective retention strategy.

Service type impacts churn: fiber-optic subscribers are far more likely to leave than DSL users, perhaps due to unmet expectations or service reliability issues. Conversely, customers without internet service (voice-only plans) are markedly less likely to churn. **Value-added services** such as online security and tech support have protective effects, underscoring the value of bundling and customer support.

Demographic effects are modest: senior citizens churn slightly less often than younger customers. **Higher monthly charges** are associated with lower hazard, suggesting that customers paying more may be more committed or derive greater value from their plans.

4.2 Limitations

Several limitations merit consideration. First, the data represent a snapshot of a fictional telco and may not generalise to other markets. Second, the cross-sectional nature of the data precludes analysis of service changes over time; customers may switch plans, affecting churn risk. Third, significant violations of the proportional hazards assumption suggest that covariate effects vary with tenure. More flexible models—such as stratified Cox models, time-dependent covariates or parametric survival models—could better capture these dynamics. Fourth, unobserved factors such as customer satisfaction, marketing interactions or socioeconomic variables could confound the observed relationships. Finally, our analysis focused on a limited set of covariates; future work could incorporate additional variables such as contract tenure changes, product bundles and competition metrics.

4.3 Recommendations

Business strategies. Telecommunications providers should prioritise converting month-to-month customers into long-term contracts through incentives or bundled discounts. Proactive outreach to fiber-optic customers may help mitigate high churn risk; service quality audits and targeted support could address underlying issues. Encouraging automatic payment adoption—via discounts or convenience features—can reduce churn associated with manual payment methods. Expanding value-added services like online security and tech support and bundling them into plans may further enhance retention.
Model enhancements. Future research could employ stratified or extended Cox models to accommodate time-varying effects, incorporate additional covariates to capture unobserved heterogeneity and apply competing risks models to distinguish voluntary and involuntary churn. Gathering longitudinal data on service changes and customer interactions would enable dynamic models that more accurately reflect customer behaviour.

5 Conclusion

By integrating non-parametric and multivariable survival techniques, this report provides a comprehensive view of customer retention in the IBM Telco dataset. We confirm contract length as the dominant driver of churn, identify service and billing factors that meaningfully influence hazard and demonstrate the value of advanced survival methods for actionable business insights. While assumption violations highlight the need for more sophisticated models, the findings offer clear guidance for retention strategies: promote long-term contracts, encourage automatic payments, improve fiber-optic service quality and expand value-added services. These interventions could significantly reduce churn and improve profitability in competitive telecommunications markets.

Github Repository: https://github.com/djoudi92/Survival-Analysis