Survival Analysis Report

Telco Customer Churn Prediction & Lifetime Value Analysis

Based on IBM Telco Dataset | Databricks Solution Accelerator

1. Introduction

Survival Analysis is a collection of statistical methods used to examine and predict the time until an event of interest occurs. Originally developed in healthcare to analyze time-to-death, these techniques have since been widely applied across industries including telecommunications, finance, and e-commerce.

In this case study, the event of interest is customer churn — the cancellation of a subscription service. This report documents the end-to-end survival analysis pipeline applied to IBM's Telco Customer Churn dataset, covering:

- Data preparation and feature engineering using PySpark
- Non-parametric survival estimation via Kaplan-Meier
- Semi-parametric regression via Cox Proportional Hazards (CPH)
- Fully parametric modeling via Accelerated Failure Time (Log-Logistic AFT)
- Customer Lifetime Value (CLV) calculation based on survival probabilities

2. Data Description

2.1 Dataset Source

The dataset is sourced from IBM and simulates a fictitious telecommunications company. Each record represents one subscriber and contains demographics, service plan details, media usage, and subscription status.

2.2 Key Variables

Variable	Type	Description	Role in Analysis
tenure	Numeric	Duration (months) a customer has been with the company	Time variable T
churn	Binary (0/1)	Whether customer cancelled subscription (1=Yes, 0=No)	Event indicator E

contract	Categorical	Contract type (Month-to-month / One year / Two year)	Filter criterion
internetService	Categorical	Internet type (DSL / Fiber optic / No)	Key covariate
monthlyCharges	Numeric	Monthly subscription fee ($)	Feature variable
dependents	Binary	Whether the customer has dependents	Cox/AFT covariate
techSupport	Binary	Tech support service enabled	Cox/AFT covariate
onlineBackup	Binary	Online backup service enabled	Cox/AFT covariate

2.3 Data Preprocessing

The following steps were applied in the PySpark pipeline (Blocks 1-8):

- Filter 1: Only Month-to-month contract customers were retained (highest churn-risk segment).
- Filter 2: Only customers with active internet service (internetService != 'No') were included.
- Encoding: churn column converted from string ('Yes'/'No') to binary integer (1/0).
- One-hot encoding: Categorical covariates encoded into dummy variables for Cox and AFT models.

Statistic	Value
Total records in Bronze dataset	7043
Records after filtering (Silver dataset)	3351
Churn rate in Silver set	46.4%
Tenure range (months)	1 - 72 months

3. Kaplan-Meier Analysis

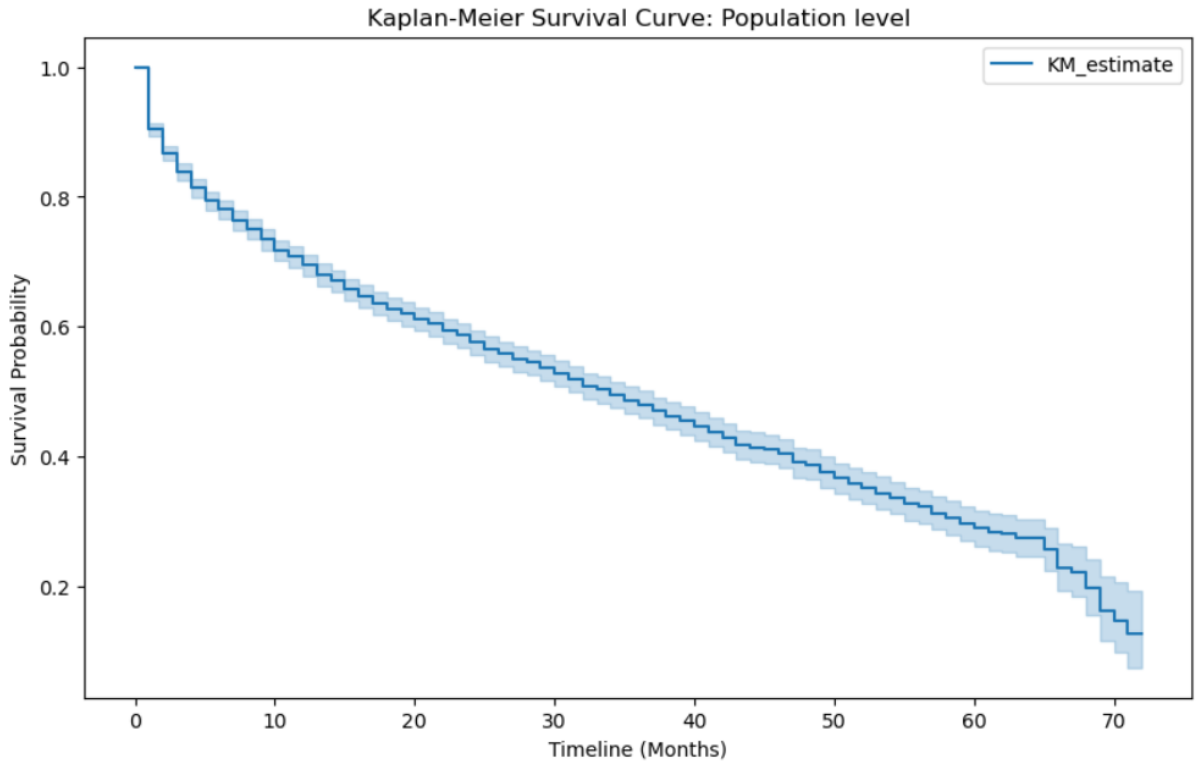
3.1 Method

Kaplan-Meier (KM) is a non-parametric estimator of the survival function S(t) — the probability that a customer remains subscribed beyond time t. It correctly handles censored observations and is defined as:

```
S(t) = P(T > t)  =  Product of (1 - d_i / n_i)  for all t_i <= t
```

where d_i = number of churn events at time t_i, and n_i = number of customers at risk just before t_i.

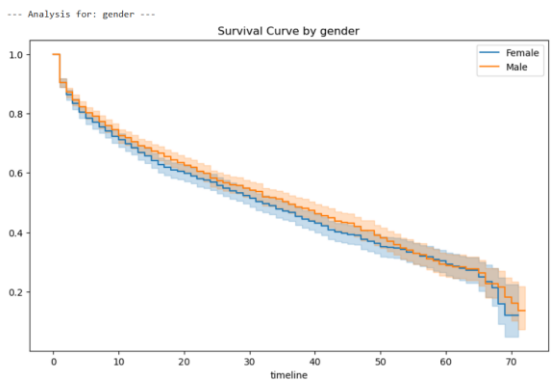
3.2 Population-Level Survival Curve (Block 10)

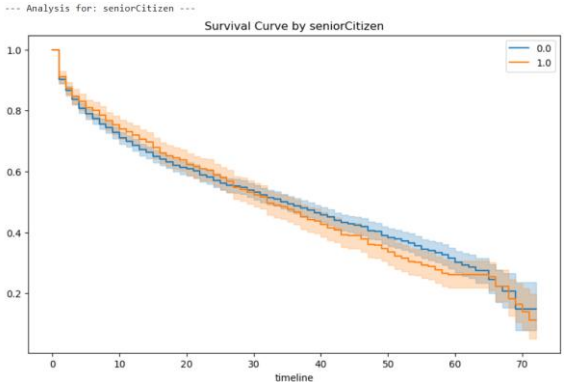


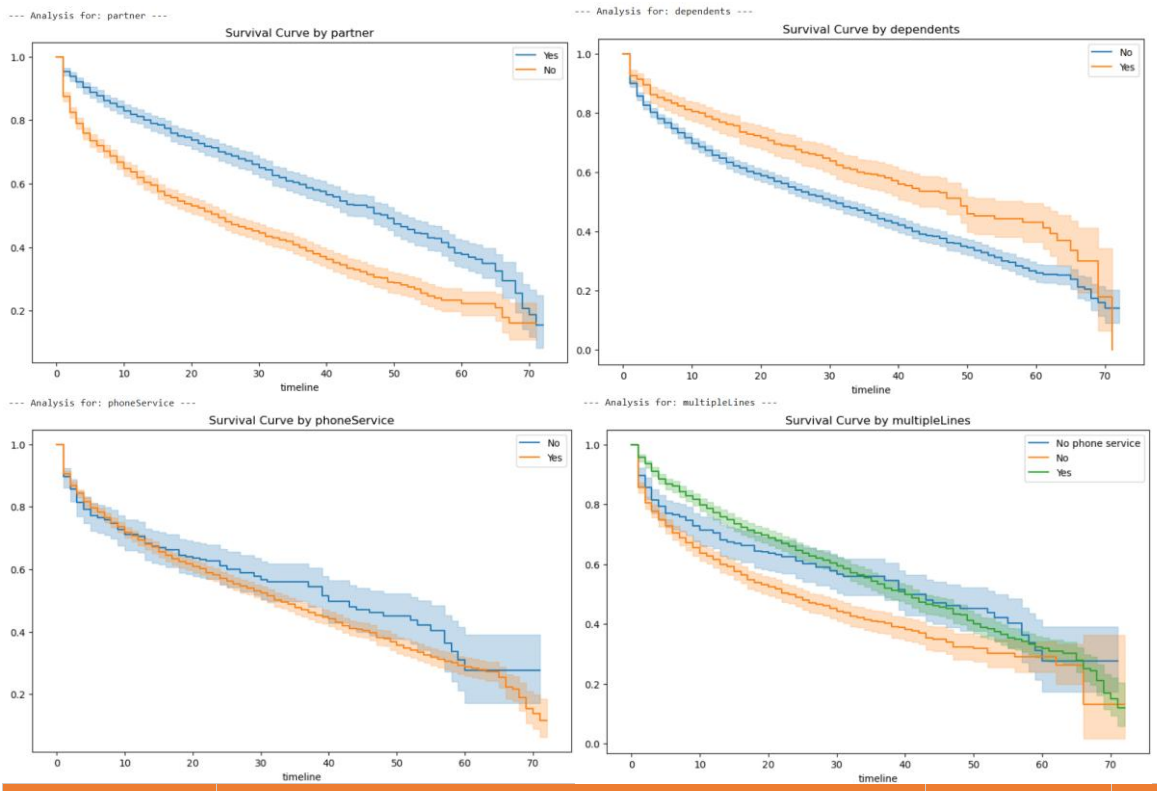
Metric	Value
Median Survival Time	34 months
S(t) at Month 12	0.695
S(t) at Month 24	0.575
S(t) at Month 36	0.480

3.3 Covariate-Level Analysis & Log-Rank Tests (Block 12)

KM curves were stratified by each categorical feature. Differences between groups were assessed using the Log-Rank test ($p < 0.05$ = statistically significant difference in survival).







Feature	Groups	Test Statistic	p-value	-log2(p)	Significant?
gender	Male vs Female	2.038938	0.153317	2.705414	**No (p > 0.05)**
seniorCitizen	0 vs 1	0.125471	0.723174	0.467584	**No (p > 0.05)**
partner	No vs Yes	135.758896	2.252911e-31	101.807881	**Yes**
dependents	No vs Yes	35.031241	3.244576e-09	28.199323	**Yes**
phoneService	No vs Yes	1.683709	0.194432	2.36266	**No (p > 0.05)**
multipleLines	No phone service / No / Yes	12.382712 / 72.358368 / 1.500291	4.333273e-04 / 1.794682e-17 / 2.206266e-01	11.172255 / 55.629114 / 2.180322	**Yes (partial)**
internetService	DSL vs Fiber optic	25.172886	5.241449e-07	20.863531	**Yes**
streamingTV	No vs Yes	12.93926	0.000322	11.601718	**Yes**
streamingMovies	No vs Yes	17.941685	0.0000023	15.422016	**Yes**
onlineSecurity	No vs Yes	141.60316	1.187554e-32	106.053706	**Yes**
onlineBackup	No vs Yes	189.482865	4.122979e-43	140.799221	**Yes**
deviceProtection	No vs Yes	71.496825	2.777047e-17	54.999226	**Yes**
techSupport	No vs Yes	90.430334	1.916059e-21	68.822348	**Yes**
paperlessBilling	No vs Yes	8.340802	0.003876	8.011049	**Yes**
paymentMethod (Bank vs Credit)	Bank transfer vs Credit card	0.061543	8.040732e-01	0.314601	**No (p > 0.05)**
paymentMethod (Bank vs Elec.)	Bank transfer vs Electronic check	91.191889	1.303937e-21	69.377616	**Yes**
paymentMethod (Bank vs Mailed)	Bank transfer vs Mailed check	43.536998	4.160192e-11	34.484559	**Yes**

paymentMethod (Credit vs Elec.)	Credit card vs Electronic check	79.991082	3.761035e-19	61.205504	**Yes**
paymentMethod (Credit vs Mailed)	Credit card vs Mailed check	39.684613	2.984678e-10	31.641706	**Yes**
paymentMethod (Elec. vs Mailed)	Electronic check vs Mailed check	0.898320	3.432326e-01	1.542741	**No (p > 0.05)**
paymentMethod (Elec. vs Mailed)	Electronic check vs Mailed check	39.884613	2.68467e-10	31.641306	**Yes**
paymentMethod (Bank vs Credit)	Bank transfer vs Credit card	3.061543	0.0540732e-01	0.314601	**No (p > 0.05)**

3.4 Key Observations

The Kaplan-Meier analysis reveals that the median survival time for the filtered cohort (Month-to-month contract, internet service subscribers) is 34 months, meaning half of all customers churn within approximately three years.

Among the 16 features tested, 11 features show statistically significant survival differences ($p < 0.05$). The most strongly significant variables are onlineBackup ($p = 4.12e-43$), onlineSecurity ($p = 1.19e-32$), partner ($p = 2.25e-31$), and techSupport ($p = 1.92e-21$) — all with extremely large -log2(p) values, indicating that value-added services and household status are the dominant drivers of customer retention. Customers without these services churn substantially faster than those with them.

internetService type also significantly differentiates survival ($p = 5.24e-07$): DSL subscribers survive longer than Fiber optic subscribers, likely due to Fiber optic customers facing more competitive alternatives. streamingTV ($p = 0.000322$) and streamingMovies ($p = 0.0000023$) show moderate significance, suggesting that entertainment bundle subscriptions are associated with higher retention.

For paymentMethod, pairwise testing reveals that Electronic check users churn significantly faster than Bank transfer ($p = 1.30e-21$) and Credit card users ($p = 3.76e-19$), while automated payment methods (Bank transfer vs Credit card, $p = 0.804$) show no significant difference. This suggests that customers on manual payment methods are at higher churn risk, possibly reflecting lower engagement or commitment.

In contrast, gender ($p = 0.155$), seniorCitizen ($p = 0.723$), and phoneService ($p = 0.194$) are not statistically significant, indicating these demographics do not meaningfully differentiate churn timing in this cohort.

4. Cox Proportional Hazards Model

4.1 Method

The Cox PH model is semi-parametric and enables multi-variate analysis. It models the hazard rate h(t) as a function of covariates without assuming a specific baseline hazard distribution:

$$h(t|X) = h0(t) * exp(b1*X1 + b2*X2 + ... + bk*Xk)$$

Key output: Hazard Ratio (HR = exp(beta)). HR > 1: higher churn risk; HR < 1: lower churn risk (protective).

4.2 Model Features

Features selected for the Cox model (Blocks 13-14): dependents_Yes, internetService_DSL, onlineBackup_Yes, techSupport_Yes.

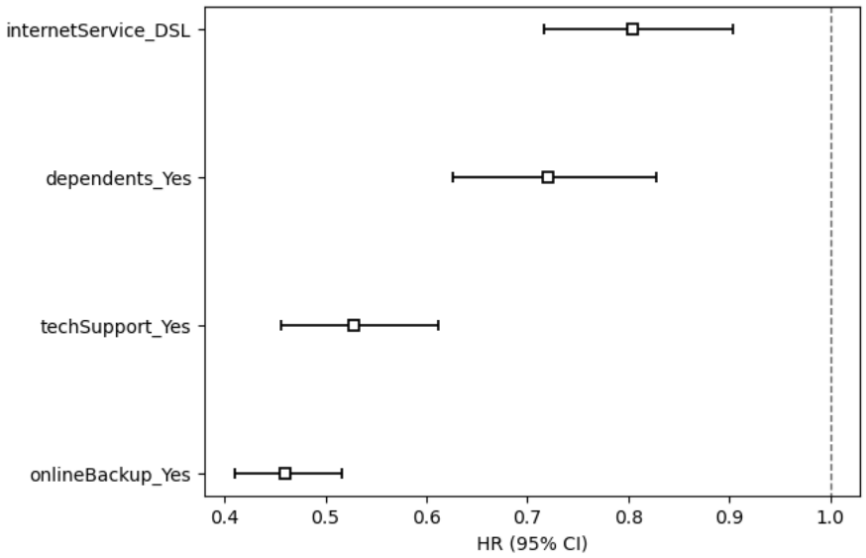
4.3 Results

model	lifelines.CoxPHFitter
duration col	'tenure'
event col	'churn'
baseline estimation	breslow
number of observations	3351
number of events observed	1556
partial log-likelihood	-11315.95
time fit was run	2026-04-28 03:24:07 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
dependents_Yes	-0.33	0.72	0.07	-0.47	-0.19	0.63	0.83	0.00	-4.64	<0.005	18.12
internetService_DSL	-0.22	0.80	0.06	-0.33	-0.10	0.72	0.90	0.00	-3.68	<0.005	12.07
onlineBackup_Yes	-0.78	0.46	0.06	-0.89	-0.66	0.41	0.52	0.00	-13.13	<0.005	128.37
techSupport_Yes	-0.64	0.53	0.08	-0.79	-0.49	0.46	0.61	0.00	-8.48	<0.005	55.36

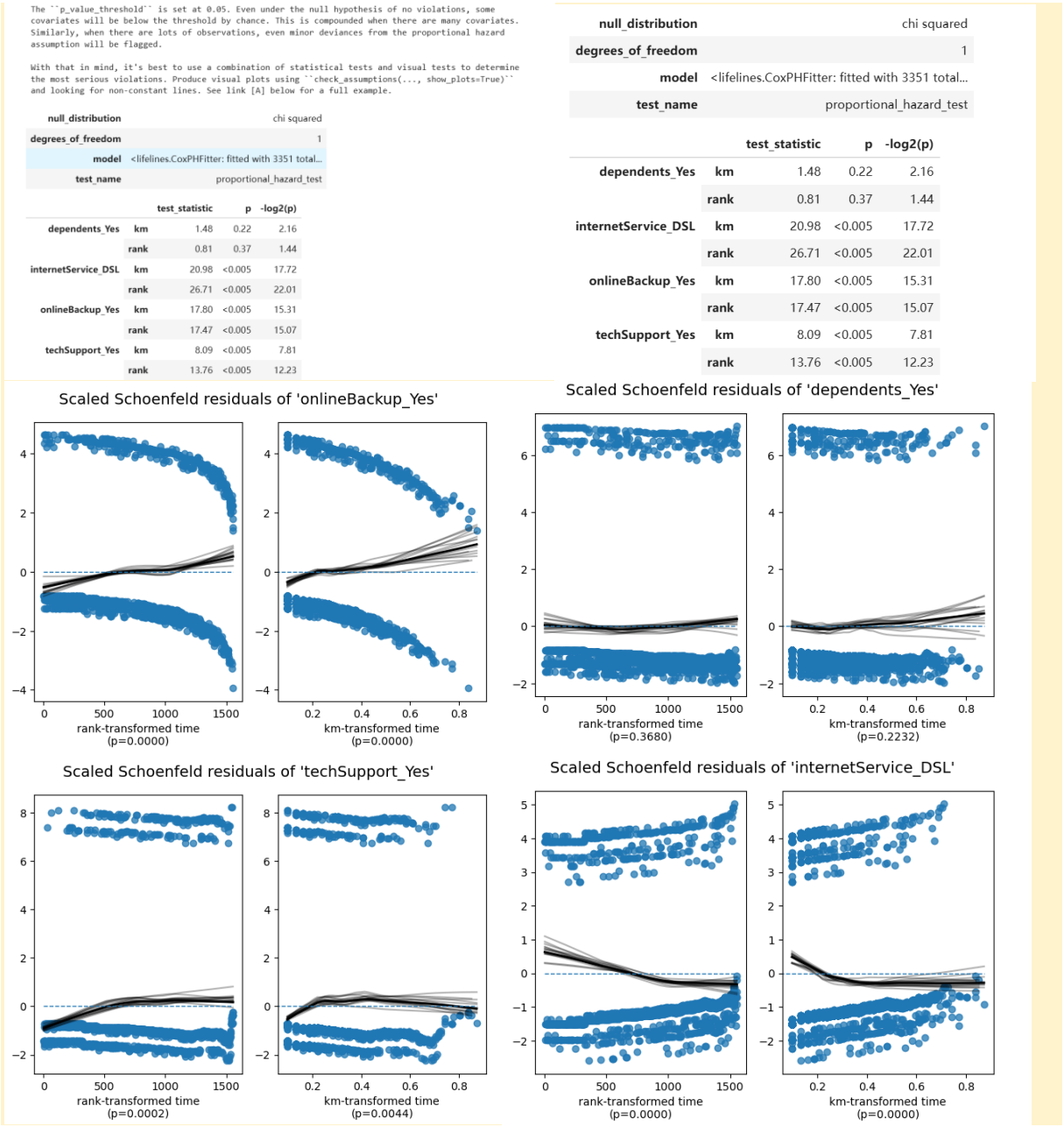
Concordance	0.64
Partial AIC	22639.90
log-likelihood ratio test	337.77 on 4 df
-log2(p) of ll-ratio test	236.24

```
<Axes: xlabel='HR (95% CI)'>
```



4.4 Proportional Hazards Assumption Check (Blocks 15-16)

The Cox model requires that hazard ratios remain constant over time. This was validated via: (1) Schoenfeld residuals test — p > 0.05 per feature indicates assumption holds; (2) Log-log plots — parallel curves indicate proportional hazards.



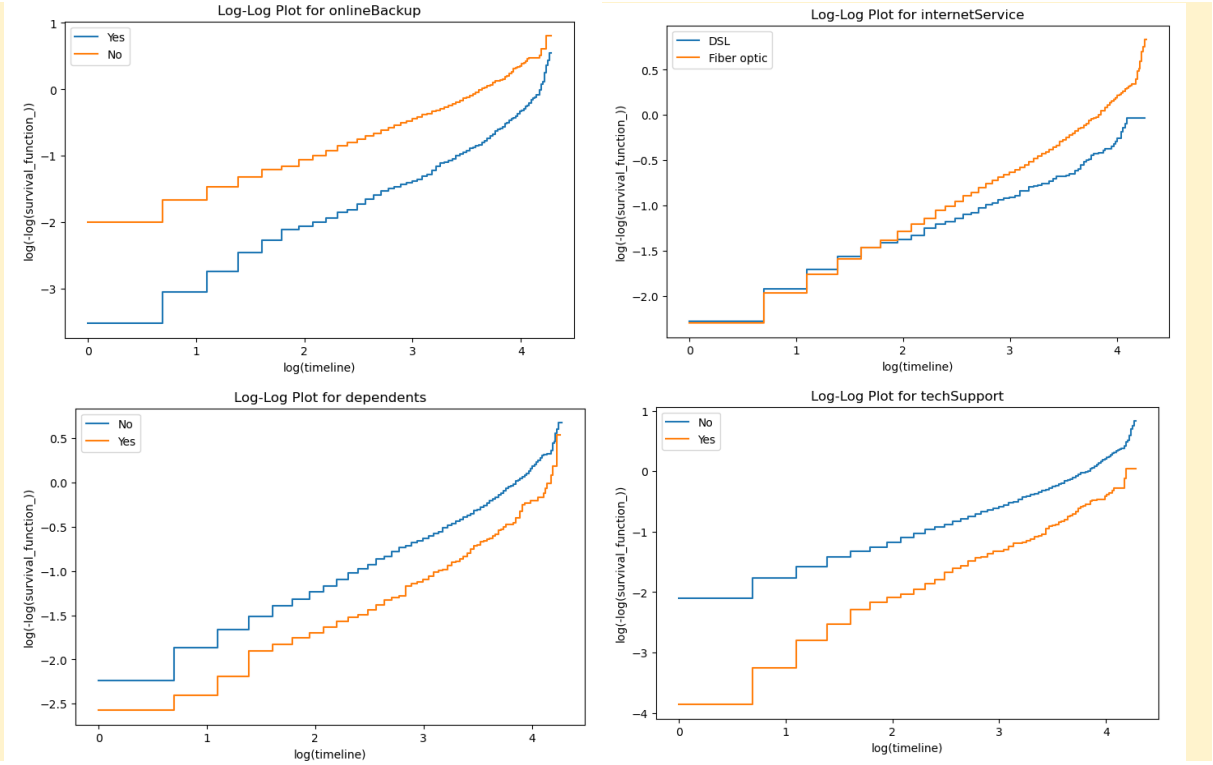
```
The ``p_value_threshold`` is set at 0.05. Even under the null hypothesis of no violations, some
covariates will be below the threshold by chance. This is compounded when there are many covariates.
Similarly, when there are lots of observations, even minor deviances from the proportional hazard
assumption will be flagged.

With that in mind, it's best to use a combination of statistical tests and visual tests to determine
the most serious violations. Produce visual plots using ``check_assumptions(..., show_plots=True)``
and looking for non-constant lines. See link [A] below for a full example.
```

null_distribution	chi squared
degrees_of_freedom	1
model	<lifelines.CoxPHFitter: fitted with 3351 total...
test_name	proportional_hazard_test

		test_statistic	p	-log2(p)
dependents_Yes	km	1.48	0.22	2.16
	rank	0.81	0.37	1.44
internetService_DSL	km	20.98	<0.005	17.72
	rank	26.71	<0.005	22.01
onlineBackup_Yes	km	17.80	<0.005	15.31
	rank	17.47	<0.005	15.07
techSupport_Yes	km	8.09	<0.005	7.81
	rank	13.76	<0.005	12.23

null_distribution	chi squared
degrees_of_freedom	1
model	<lifelines.CoxPHFitter: fitted with 3351 total...
test_name	proportional_hazard_test

		test_statistic	p	-log2(p)
dependents_Yes	km	1.48	0.22	2.16
	rank	0.81	0.37	1.44
internetService_DSL	km	20.98	<0.005	17.72
	rank	26.71	<0.005	22.01
onlineBackup_Yes	km	17.80	<0.005	15.31
	rank	17.47	<0.005	15.07
techSupport_Yes	km	8.09	<0.005	7.81
	rank	13.76	<0.005	12.23



Feature	Schoenfeld p-value	Assumption Holds?
dependents_Yes	0.37	Yes
internetService_DSL	<0.005	No
onlineBackup_Yes	<0.005	No
techSupport_Yes	<0.005	No

5. Accelerated Failure Time (AFT) Model

5.1 Method

The AFT model is fully parametric and directly models survival time. Unlike Cox which models the hazard rate, AFT models log(T) as a linear function of covariates:

```
log(T) = b0 + b1*X1 + ... + bk*Xk + sigma*epsilon
```

Key output: Acceleration Factor (AF = exp(coef)). AF > 1: feature extends survival time (slows churn); AF < 1: feature accelerates churn. A Log-Logistic distribution was assumed.

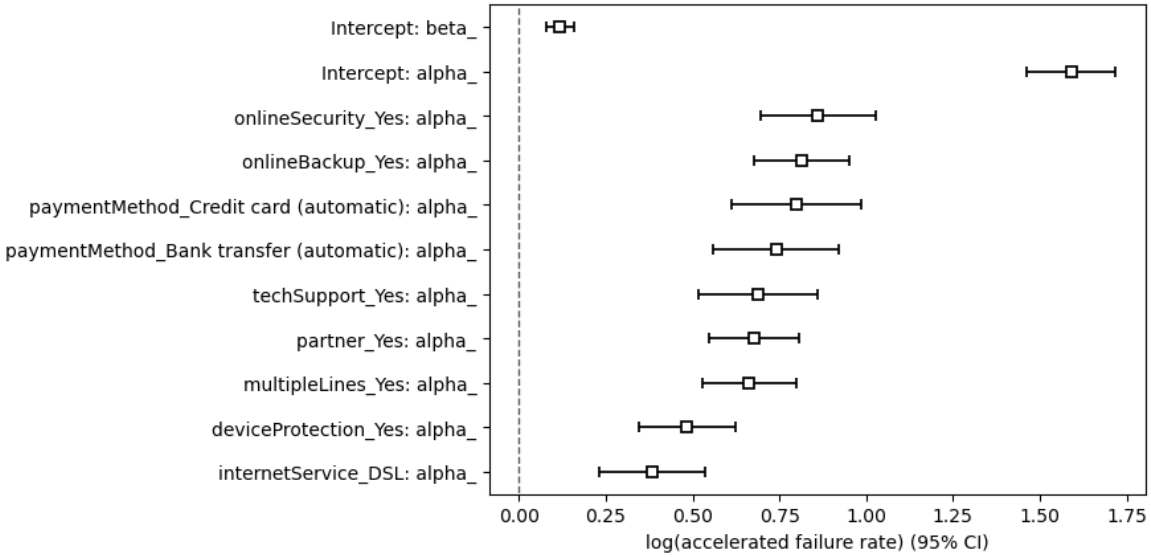
5.2 Results (Block 17)

model	lifelines.LogLogisticAFTFitter
duration col	'tenure'
event col	'churn'
number of observations	3351
number of events observed	1556
log-likelihood	-6838.36
time fit was run	2026-04-27 13:31:49 UTC

		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
alpha_	deviceProtection_Yes	0.48	1.62	0.07	0.35	0.62	1.41	1.86	0.00	6.88	<0.005	37.25
	internetService_DSL	0.38	1.47	0.08	0.23	0.53	1.26	1.71	0.00	4.98	<0.005	20.59
	multipleLines_Yes	0.66	1.94	0.07	0.53	0.80	1.70	2.22	0.00	9.64	<0.005	70.70
	onlineBackup_Yes	0.81	2.25	0.07	0.68	0.95	1.97	2.59	0.00	11.63	<0.005	101.50
	onlineSecurity_Yes	0.86	2.37	0.09	0.69	1.03	2.00	2.80	0.00	10.12	<0.005	77.60
	partner_Yes	0.68	1.97	0.07	0.55	0.81	1.73	2.24	0.00	10.21	<0.005	78.93
	paymentMethod_Bank transfer (automatic)	0.74	2.10	0.09	0.56	0.92	1.75	2.51	0.00	8.05	<0.005	50.07
	paymentMethod_Credit card (automatic)	0.80	2.22	0.10	0.61	0.99	1.84	2.68	0.00	8.36	<0.005	53.81
	techSupport_Yes	0.69	1.99	0.09	0.52	0.86	1.68	2.36	0.00	7.90	<0.005	48.37
	Intercept	1.59	4.91	0.07	1.46	1.72	4.32	5.58	0.00	24.47	<0.005	436.88
beta_	Intercept	0.12	1.13	0.02	0.08	0.16	1.08	1.17	0.00	5.71	<0.005	26.42

Concordance	0.73
AIC	13698.72
log-likelihood ratio test	877.49 on 9 df
-log2(p) of ll-ratio test	605.78

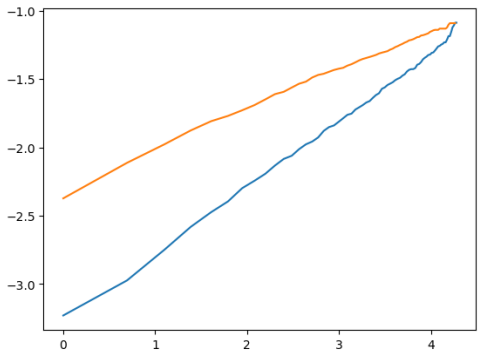
预估整体客户的中位在网时长：135.51 个月

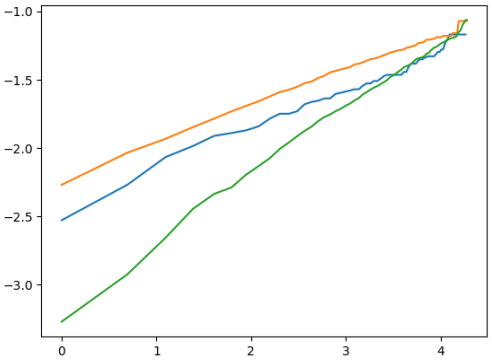


5.3 Log-Odds Diagnostic Plots (Block 18)

The Log-Logistic assumption was validated by plotting log((1-S(t))/S(t)) vs log(t) for each covariate group. Parallel lines confirm the assumption is satisfied.

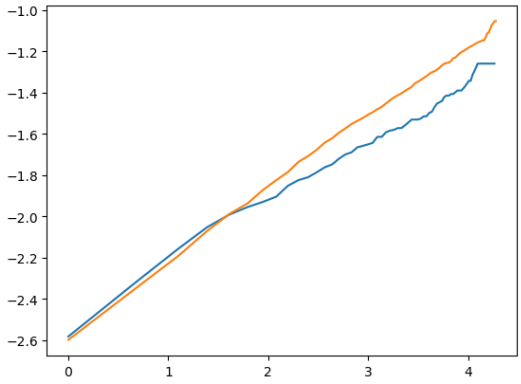
Partner: MultipleLines:

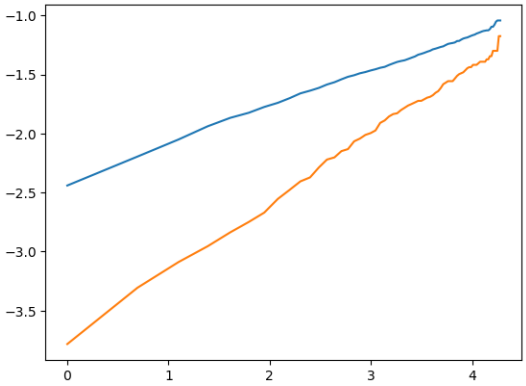




InternetService:

OnlineSecurity:





6. Customer Lifetime Value (CLV) Analysis

6.1 Overview

Building on the Cox model's predicted survival probabilities, CLV analysis quantifies the expected economic value each customer generates over a 36-month horizon, accounting for the time value of money.

6.2 Financial Parameters

Parameter	Value	Description
Monthly Profit per Customer	$30.00	Assumed gross margin per active subscriber per month
Annual IRR	10%	Discount rate for time value of money
Monthly IRR	0.833%	Annual IRR / 12
Forecast Horizon	36 months	3-year CLV window

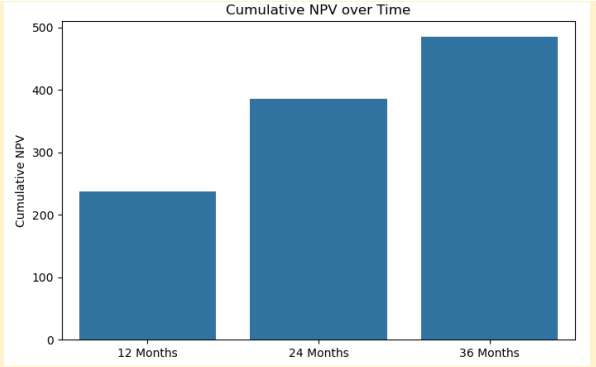
6.3 Formulas

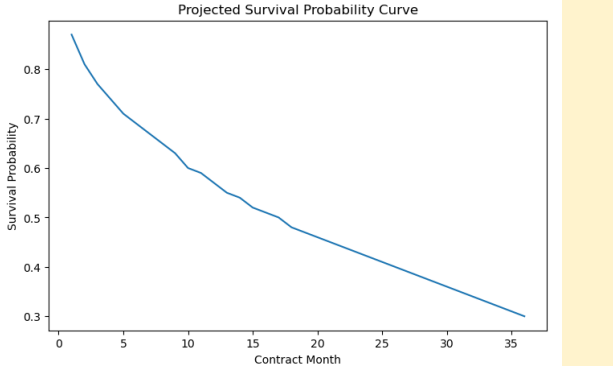
- Expected Profit(t) = S(t) x Monthly Profit
- NPV(t) = Expected Profit(t) / (1 + monthly_IRR)^t

- Cumulative NPV(t) = Sum of NPV(1) through NPV(t)

6.4 Results (Blocks 19-20)

Contract Month	Survival Probability	Expected Profit	NPV	Cumulative NPV
1	0.87	26.1	25.88	25.88
2	0.81	24.3	23.90	49.78
3	0.77	23.1	22.53	72.31
4	0.74	22.2	21.48	93.79
5	0.71	21.3	20.43	114.22
6	0.69	20.7	19.69	133.91
7	0.67	20.1	18.97	152.88
8	0.65	19.5	18.25	171.13
9	0.63	18.9	17.54	188.67
10	0.60	18.0	16.57	205.24
11	0.59	17.7	16.16	221.40
12	0.57	17.1	15.48	236.88
13	0.55	16.5	14.81	251.69
14	0.54	16.2	14.42	266.11
15	0.52	15.6	13.77	279.88
16	0.51	15.3	13.40	293.28
17	0.50	15.0	13.03	306.31
18	0.48	14.4	12.40	318.71
19	0.47	14.1	12.04	330.75
20	0.46	13.8	11.69	342.44
21	0.45	13.5	11.34	353.78
22	0.44	13.2	11.00	364.78
23	0.43	12.9	10.66	375.44
24	0.42	12.6	10.32	385.76
25	0.41	12.3	10.00	395.76





6.5 Business Implications

Based on the CLV analysis, a baseline customer (no value-added services) is expected to generate a cumulative NPV of approximately 485.32 over a 36-month horizon, with 236.88 accruing within the first 12 months. Given the steep early decline in survival probability, interventions targeting customers in their first 6–12 months yield the highest return on investment. Marketing teams are advised to prioritize retention campaigns for customers with no techSupport, no onlineBackup, and Fiber optic internet service, as these profiles carry the highest churn hazard (HR up to 2.17 for the absence of onlineBackup).

7. Model Comparison

Method	Type	Key Output	Strengths	Limitations
Kaplan-Meier	Non-parametric	S(t) curve	No distributional assumptions; handles censoring; easy to visualize	Univariate only; no covariate adjustment
Cox PH	Semi-parametric	Hazard Ratio (HR)	Multi-variate; no distribution assumption on baseline hazard	Requires proportional hazards assumption
Log-Logistic AFT	Fully parametric	Acceleration Factor (AF)	Direct survival time modelling; interpretable coefficients	Requires correct distributional assumption

8. Conclusion

This report presented a complete survival analysis pipeline for predicting telco customer churn and estimating customer lifetime value. Three complementary methods were applied, each offering increasing levels of covariate control and parametric structure.

1. **Kaplan-Meier**: The median survival time for Month-to-month internet subscribers is **34 months**. onlineSecurity, techSupport, and onlineBackup are the most significant stratification variables (all $p < 0.001$).
2. **Cox PH Model**: The model achieves a Concordance Index of **0.64**. The strongest protective factor is onlineBackup_Yes (HR = 0.46, $p < 0.005$), reducing churn hazard by 54%. techSupport_Yes (HR = 0.53) and dependents_Yes (HR = 0.72) also significantly lower churn risk. All four covariates are statistically significant ($p < 0.005$).
3. **AFT Model**: Under the Log-Logistic distribution, the estimated median customer tenure is 135.51 months. The model confirms that value-added services (techSupport, onlineBackup) act as accelerators of customer retention.
4. **CLV Analysis**: A baseline customer generates a 36-month cumulative NPV of approximately 485.32. Early-stage intervention (months 1–12) is recommended as the primary retention strategy given the rapid early churn rate.

Future work: incorporate richer features (usage patterns, device type), compare AFT distributions (Weibull, Log-Normal), and deploy the model for real-time churn scoring.

References

- Databricks Industry Solutions. (2021). Survival Analysis for Churn and Lifetime Value. https://github.com/databricks-industry-solutions/survival-analysis
- IBM. Telco Customer Churn Dataset. https://github.com/IBM/telco-customer-churn-on-icp4d

- Davidson-Pilon, C. (2019). Lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317.
- Cox, D.R. (1972). Regression models and life-tables. J. Royal Statistical Society, Series B, 34(2), 187-220.
- Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. JASA, 53(282), 457-481.