Survival Analysis Report

Shanli Ouyang

April 2025

In this quesion, I will explain the use of those five survival analysis process, and the result will be as well provided.

1.1 Intro

In "01_intro.ipynb", through relevant operations on the columns of raw data, the "silver_data.csv" has been processed for subsequent functionality, as shown in Figure 1.

	customerID	gender	seniorCitizen	partner	dependents
1	7590-VHVEG	Female	0.0	Yes	No
2	3668-QPYBK	Male	0.0	No	No
3	9237-HQITU	Female	0.0	No	No
4	9305-CDSKC	Female	0.0	No	No
5	1452-KIOVK	Male	0.0	No	Yes
6	6713-OKOMC	Female	0.0	No	No
7	7892-POOKP	Female	0.0	Yes	No
8	9763-GRSKD	Male	0.0	Yes	Yes
9	0280-XJGEX	Male	0.0	No	No
10	5129-JLPIS	Male	0.0	No	No
11	4190-MFLUW	Female	0.0	Yes	Yes
12	4183-MYFRB	Female	0.0	No	No
13	8779-QRDMV	Male	1.0	No	No
14	6322-HRPFA	Male	0.0	Yes	Yes
15	6865-JZNKO	Female	0.0	No	No
16	6467-CHFZW	Male	0.0	Yes	Yes
17	8665-UTDHZ	Male	0.0	Yes	Yes
18	8773-HHUOZ	Female	0.0	No	Yes
19	4929-XIHVW	Male	1.0	Yes	No

Figure 1: silver data

1.2 Kaplan Meier

In "02_kaplan_meier.ipynb", the procedure mainly conducts a survival analysis of "telecom customer churn" data, revealing the impact of various features (e.g., gender, internet service type, senior citizen status) on customer churn.

Through visualized survival curves and statistical tests, it provides an in-depth understanding of customer churn behavior.

After short data loading and preprocessing, a Kaplan-Meier model (KaplanMeierFitter) is initialized and fitted using the "tenure" and "churn" data. An overall survival curve(Figure 2) is plotted to show the probability of survival at different time points for the entire customer population.

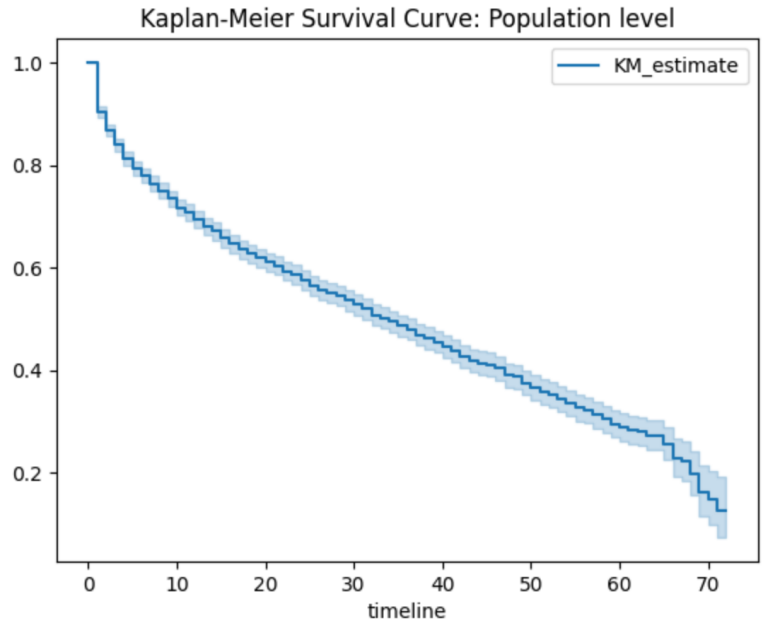


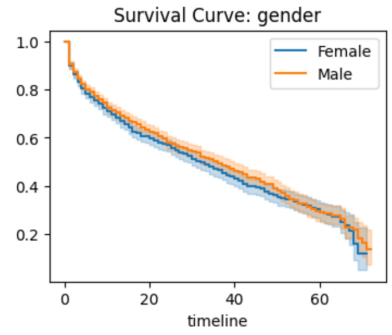
Figure 2: Survival Curve of KaplanMeierFitter

Then,A function "plot_km(col)" is defined to group customers by a specified column (e.g., gender, internetService) and plot the survival curves for each group. Also, a function "print_logrank(col)" is defined to perform a Log-rank test for a specified column, determining whether there are significant differences in survival curves between groups.

The two function can be used to analysis any of the columns, so I select two columns to give some examples: as shown in Figure 3,the survival curve of different groups are clearly shown on the Survival Prob over Timeline graphs. And the significance tests show that the p-value of gender group test is larger than 0.05, which means the survival curve difference between Female and Male is not significant; instead, in the partner group test, the p-value is very small, indicating a significant difference between having or not having a partner.

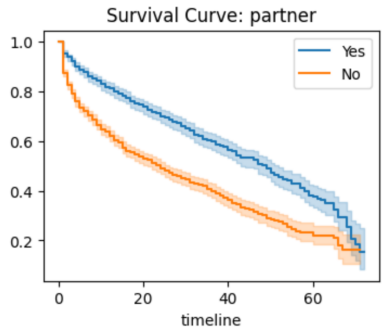
Those analysis of the two functions helps observe the differences in survival probabilities among different customer groups based on specific features, providing insights into the significance of features on customer churn.

At last, a function "get_survival_probs(col, val)" is defined to calculate the



		test_statistic	p	-log2(p)
Female	Male	1.61011	0.204476	2.289995

(a) Over Gender Groups



		test_statistic	p	-log2(p)
No	Yes	257.844159	5.063437e-58	190.331712

(b) Over Partners Groups

Figure 3: Survival Curves and Significance Test

survival probabilities for a specific group (e.g., internetService=DSL) at specified time points.The results are returned as a DataFrame, showing the survival probabilities for the group from time 0 to 9.

1.3 Cox Proportional Hazards

After data preprocessing, a Cox Proportional Hazards model (CoxPHFitter) is initialized, using tenure (customer tenure) as the time variable and churn (whether the customer churned) as the event variable to fit the model. The model summary is including coefficients, hazard ratios, and p-values for each feature, as shown in Figure 4.

model	lifelines.CoxPHFitter
duration col	'tenure'
event col	'churn'
baseline estimation	breslow
number of observations	3351
number of events observed	1556
partial log-likelihood	-11178.40
time fit was run	2025-04-11 09:59:50 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	-0.10	0.90	0.07	-0.24	0.03	0.79	1.03	0.00	-1.52	0.13	2.95
internetService_DSL	-0.04	0.96	0.06	-0.15	0.07	0.86	1.07	0.00	-0.78	0.44	1.19
onlineBackup_Yes	-0.35	0.71	0.06	-0.46	-0.23	0.63	0.79	0.00	-6.06	<0.005	29.49
techSupport_Yes	-0.21	0.81	0.07	-0.34	-0.08	0.71	0.92	0.00	-3.11	<0.005	9.06

Concordance	0.64
Partial AIC	22364.80
log-likelihood ratio test	57.57 on 4 df
-log2(p) of ll-ratio test	36.63

Figure 4: CPH Model Summary

A hazard ratio plot is generated to visually display the impact of each feature

on customer churn, as shown in Figure 5.

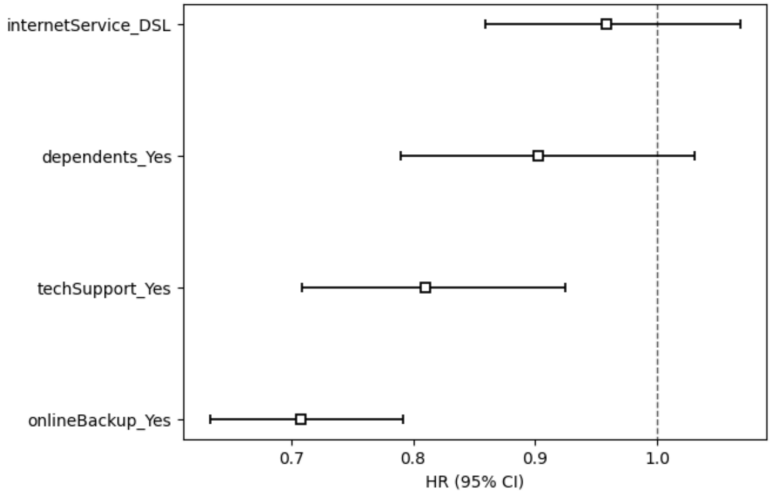


Figure 5: HR Plot of CPH Model

The proportional hazards assumption is tested using the "check_assumptions method", and relevant plots are generated to validate the assumption. The assumption result is shown in Figure 6, the p-value indicates whether the proportional hazards assumption holds.

Also, the result of the assumption can be intuitively seen in the plot. Two examples of two variables(group name) are provided in Figure 7.

In the "dependent_Yes" plot, p_value over rank is larger than 0.05, meaning we cannot reject the null hypothesis: the survival curve of dependent case or not are proportional, meaning that the difference is not significant. On the contrary, we can see that in "onlineBackup_Yes" plot, the p_values are all smaller than 0.05, so the difference of survival curve between groups are significant.

After that, a Kaplan-Meier model (KaplanMeierFitter) is initialized and fitted using tenure and churn data. A function "plot_km_loglog(col)" is defined to plot the log-log survival curves for specified columns (e.g., onlineBackup, dependents) to group customers by specific features. The log-log plots are used to observe whether the survival curves of different feature groups are parallel. If the curves are parallel, the proportional hazards assumption holds; if not, the assumption is violated. Some examples are shown in Figure 8.

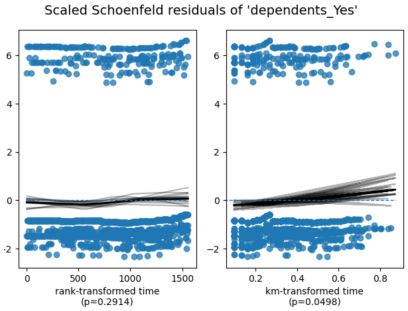
1.4 Accelerated Failure Time

After data preprocessing, a LogLogisticAFTFitter model is initialized to analyze time-dependent risk changes in customer "churn". The model is fitted using tenure (customer tenure) as the time variable and churn (whether the customer churned) as the event variable. The median survival time and parameter sum-

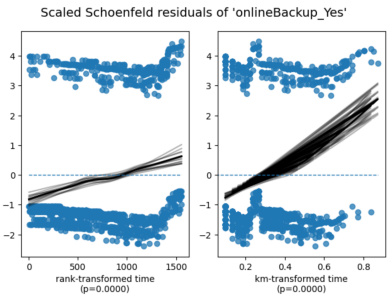
null_distribution	chi squared
degrees_of_freedom	1
model	<lifelines.CoxPHFitter: fitted with 3351 total...
test_name	proportional_hazard_test

		test_statistic	p	-log2(p)
dependents_Yes	**km**	3.85	0.05	4.33
	rank	1.11	0.29	1.78
internetService_DSL	**km**	48.21	<0.005	37.93
	rank	15.59	<0.005	13.63
onlineBackup_Yes	**km**	106.63	<0.005	80.63
	rank	47.80	<0.005	37.62
techSupport_Yes	**km**	13.68	<0.005	12.17
	rank	10.23	<0.005	9.50

Figure 6: Proportional Hazards Assumption Summary

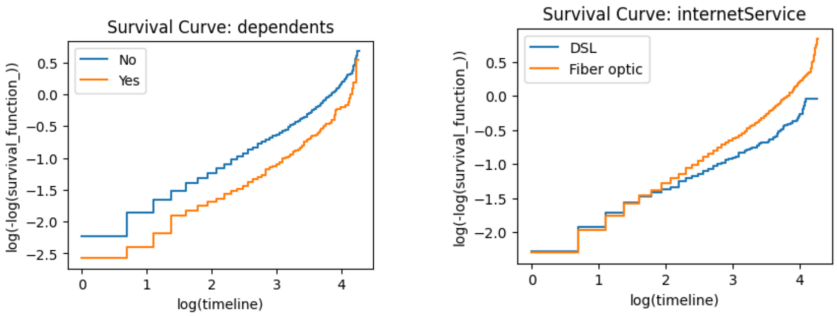


(a) Over Dependent_Yes



(b) Over OnlineBackup_Yes

Figure 7: Proportional Hazards Assumption Result



(a) Over dependent (b) Over internetService

Figure 8: Survival Curve of log(timeline)

mary are printed to evaluate the impact of different features on customer churn risk, as shown in Figure 8.

Median Survival Time:135.51

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	2025-04-11 15:58:21 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

Figure 9: LogLogisticAFTFitter Model Summary

A log plot to show the 95% CI of different variables are shown in Figure 8.

After that, as the previous methods, a Kaplan-Meier model (KaplanMeierFitter) is initialized and fitted using tenure and churn data. A function is defined to plot the log-odds survival curves for specified columns (e.g., partner, multipleLines). The log-odds survival curve is generated by mathematically transforming the survival function, showing the change in customer churn risk over time for different feature groups. Two examples are provided in Figure 11.

The log-odds survival curves visually demonstrate whether there are signif-

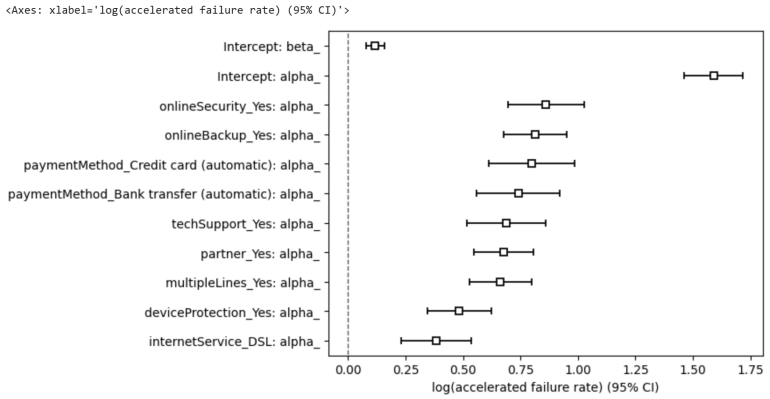
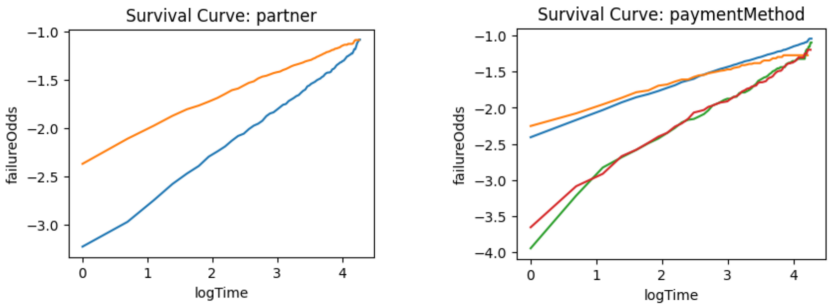


Figure 10: LogLogisticAFTFitter log(accelerated failure rate) (95% CI)



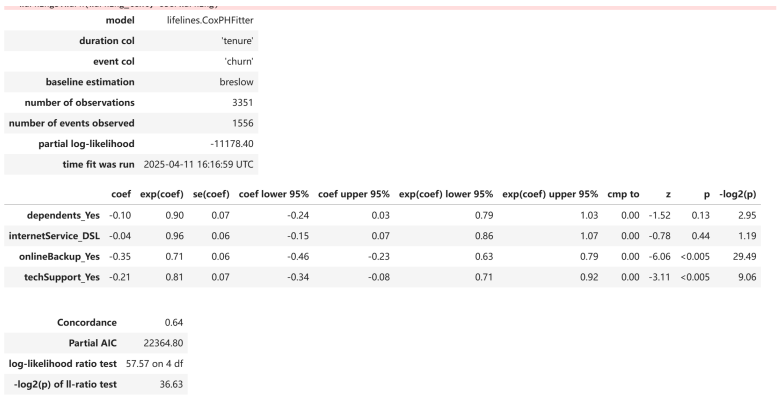
(a) Over partner (b) Over paymentMethod

Figure 11: Survival Curve of Timeline

icant differences in customer churn risk across different feature groups and how these risks change over time.

1.5 Customer Life Time Value

In this procedure, a Cox Proportional Hazards model (CoxPHFitter) is initialized, using tenure (customer tenure) as the time variable and churn (whether the customer churned) as the event variable to fit the model. The model summary is printed, including coefficients, hazard ratios, and p-values for each feature, to evaluate the impact of features on customer churn. The summary is shown in Figure 12.



model	lifelines.CoxPHFitter
duration col	'tenure'
event col	'churn'
baseline estimation	breslow
number of observations	3351
number of events observed	1556
partial log-likelihood	-11178.40
time fit was run	2025-04-11 16:16:59 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	-0.10	0.90	0.07	-0.24	0.03	0.79	1.03	0.00	-1.52	0.13	2.95
internetService_DSL	-0.04	0.96	0.06	-0.15	0.07	0.86	1.07	0.00	-0.78	0.44	1.19
onlineBackup_Yes	-0.35	0.71	0.06	-0.46	-0.23	0.63	0.79	0.00	-6.06	<0.005	29.49
techSupport_Yes	-0.21	0.81	0.07	-0.34	-0.08	0.71	0.92	0.00	-3.11	<0.005	9.06

Concordance	0.64
Partial AIC	22364.80
log-likelihood ratio test	57.57 on 4 df
-log2(p) of ll-ratio test	36.63

Figure 12: Cox Proportional Hazards model Summary

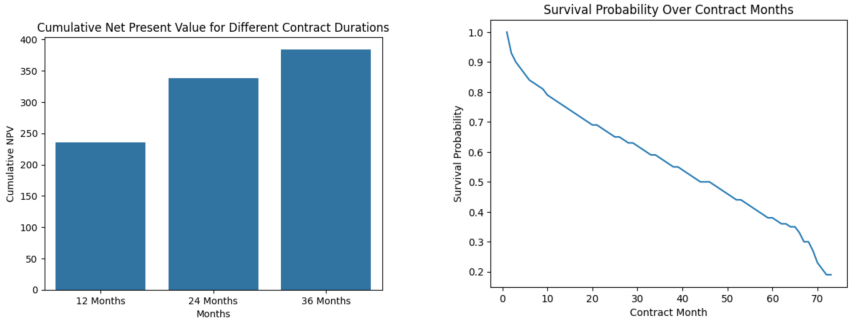
After that, a function "get_user_input()" is defined to collect user input for customer features (e.g., whether the customer has dependents, uses DSL, etc.). Then, a function "get_payback_df()" is defined to calculate the cumulative NPV and survival probability for different contract durations. The cumulative NPV is computed by discounting the expected monthly profit using the IRR and summing it over time. The survival probability is predicted by the Cox model and associated with the contract duration. And one of the result of prediction is shown in the Table 13.

```
dependents_Yes (0 or 1):  1
internetService_DSL (0 or 1):  0
onlineBackup_Yes (0 or 1):  1
techSupport_Yes (0 or 1):  0
partner_Yes (0 or 1):  1
internal rate of return (e.g., 0.10 for 10%):  0.66
```

Contract Month	Survival Probability	Monthly Profit for the Selected Plan	Avg Expected Monthly Profit	NPV of Avg Expected Monthly Profit	Cumulative NPV
1	1.00	30	30.0	30.00	30.00
2	0.93	30	27.9	26.45	56.45
3	0.90	30	27.0	24.26	80.71
4	0.88	30	26.4	22.48	103.19
5	0.86	30	25.8	20.83	124.02

Figure 13: Prediction example

Then, Seaborn and Matplotlib are used to visualize the cumulative NPV and survival probability over contract duration. A bar chart shows the cumulative NPV for different contract durations (e.g., 12 months, 24 months, 36 months), and a line chart shows the survival probability over contract months, which are shown in the Figure 14.



(a) cumulative NPV (b) Survival Probability

Figure 14: Visualization over contract duration

1.6 Conclusion

Above all, we can conclude that: If you are concerned with the time until an event occurs: Use the Accelerated Failure Time (AFT) Model.The AFT model directly models the time to the event and is suitable for analyzing how covariates influence the time until an event (e.g., customer churn).

If you are concerned with the impact of features on the risk of an event: Use the Cox Proportional Hazards (CPH) Model. The CPH model estimates the hazard ratio for each covariate, providing insights into how features affect the risk of an event occurring over time.

If you are only concerned with the distribution of survival probabilities: Use the Kaplan-Meier Estimator (KMF). The Kaplan-Meier method estimates the survival function over time, making it ideal for visualizing and analyzing the probability of survival without considering covariates.