Project1 report

Lingli WANG

April 2025

1 Introduction of Survival Analysis

Survival analysis is a branch of statistics that deals with analyzing the time until the occurrence of an event of interest. This event is often referred to as a "failure" or "death," but it can represent any endpoint, such as machine breakdown, customer churn, or recovery from a disease.

What makes survival analysis unique is its ability to handle censored data—situations where the exact time of the event is unknown for some subjects. For example, if a study ends before a patient dies, we only know that the patient survived up to a certain point, but not how much longer they lived.

2 Detailed Procedures and Methods

2.1 Load and Process Data

Here there are many ways to deal with data. The tutorial used Databrikcs package, which is quite hard for me to use, so I chose an ordinary way – DataFrame! And it's also convenient to process the data.

2.2 Fit models

In this part, we'll divide the models into two kinds who estimate different numbers of variables.

2.2.1 Kplan-Meier Method

Single Variable Analysis: Kaplan-Meier: The method describes the probability of survival at time t. It requires two main parameters,time parameter —— means the survival time of each individual and event parameter —— means if something happened. And the probability formula is displayed as below: And

在某个时间点 t_i 处，设：

- d_i：第 i 个时间点发生事件的人数
- n_i：还未发生事件、仍在“风险中”的人数（即还活着、没流失）

那么：

$$P(\text{在 } t_i \text{ 之后仍然存活}) = 1 - \frac{d_i}{n_i}$$

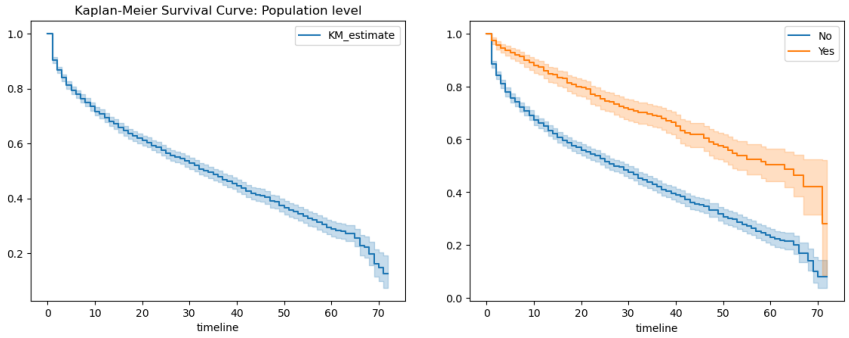
Kaplan-Meier 的估计是这些存活率的连续乘积：

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

这个乘积就是 Kaplan-Meier 曲线的“**台阶**”。

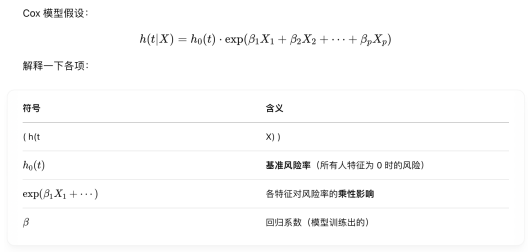
Figure 1: Caption

some examples of figures:



2.2.2 Cox Proportional Hazards

Multiple Variable Analysis: Cox Proportional Hazards: The method describes what features influence the probability of survival by estimating hazard function instead of survival time. An example on certain five features is dis-



Cox 模型假设：

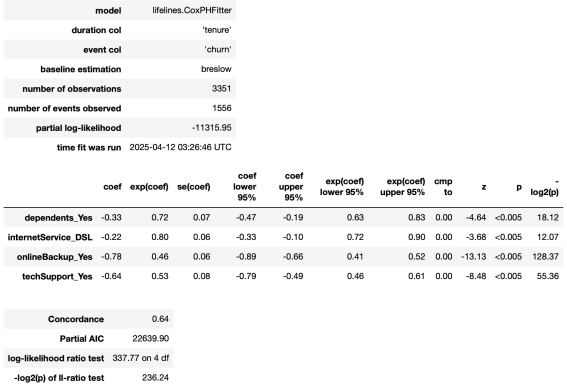
$$h(t|X) = h_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

解释一下各项：

符号	含义
(h(t	X))
$h_0(t)$	**基准风险率**（所有人特征为 0 时的风险）
$\exp(\beta_1 X_1 + \cdots)$	各特征对风险率的**乘性影响**
β	回归系数（模型训练出的）

Figure 2: Caption

played below:



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	2025-04-12 03:26:46 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

Figure 3: Caption

2.2.3 Accelerated Failure Time

Multiple Variable Analysis: Accelerated Failure Time: The method suppose that the covariables influence the result at multiplying level, so some features might make our costumer live longer or die sonner.

AFT 模型假设生存时间 T 满足如下形式：

$$\log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- T：生存时间
- $x_1, x_2, ..., x_p$：协变量（如年龄、是否有技术支持、套餐等）
- ϵ：误差项，符合某种分布（例如：对数正态、Weibull、极值分布等）

等价地，可以写作：

$$T = T_0 \cdot \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$

这里的 $\exp(\beta x)$ 就是加速因子（acceleration factor）：

- 如果 > 1：寿命变长，推迟事件
- 如果 < 1：寿命变短，加速事件发生

Figure 4: Caption

And after we fit the data and choose the features that we're interested in, we can get a summary on the influence of each feature.

model	lifelines.LogLogisticAFTFitter
duration col	'tenure'
event col	'churn'
number of observations	3351
number of events observed	1556
log-likelihood	-7095.29
time fit was run	2025-04-12 03:26:49 UTC

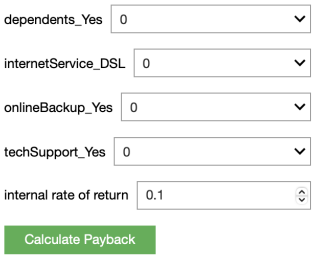
		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
alpha_	dependents_Yes	0.49	1.64	0.10	0.31	0.68	1.36	1.97	0.00	5.18	<0.005	22.13
	internetService_DSL	0.15	1.17	0.08	-0.00	0.31	1.00	1.37	0.00	1.93	0.05	4.22
	onlineBackup_Yes	1.12	3.07	0.08	0.97	1.27	2.63	3.58	0.00	14.36	<0.005	153.03
	techSupport_Yes	0.94	2.55	0.10	0.75	1.13	2.11	3.09	0.00	9.62	<0.005	70.37
	Intercept	2.63	13.87	0.05	2.53	2.73	12.52	15.37	0.00	50.38	<0.005	inf
beta_	Intercept	-0.02	0.98	0.02	-0.06	0.02	0.94	1.02	0.00	-1.13	0.26	1.94

Concordance	0.64
AIC	14202.58
log-likelihood ratio test	363.63 on 4 df
-log2(p) of ll-ratio test	254.79

Figure 5: Caption

2.3 Calculate Customer's lifetime

After fit cph, we can display the result by creating some widgets.



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	0.87	30	26.1	25.88	55.88
3	0.81	30	24.3	23.90	79.78
4	0.77	30	23.1	22.53	102.31
5	0.74	30	22.2	21.48	123.79
...	...	...	...	...	...
69	0.06	30	1.8	1.02	625.20
70	0.04	30	1.2	0.68	625.88
71	0.03	30	0.9	0.50	626.38
72	0.02	30	0.6	0.33	626.71
73	0.02	30	0.6	0.33	627.04

73 rows × 5 columns

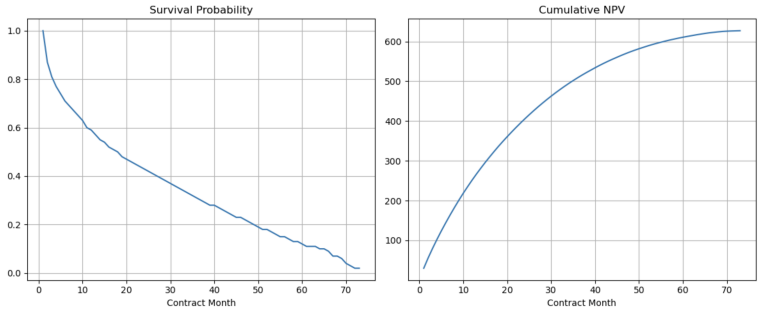


Figure 6: Caption

3 End

That's all of the survival analysis.