基于生存分析的电信行业案例分析

12212361 刘凡哲

目录

生存分析（Survival Analysis）是一组统计方法，用于研究和预测直到某个感兴趣事件发生的时间。这种分析形式起源于医疗保健领域，最初主要关注死亡时间。此后，生存分析已成功应用于全球几乎所有行业的各种场景。接下来的数据与问题分析均以电信行业使用案例为例，通过生存分析可以研究以下问题：

- Customer Retention(客户留存)：一般来说，留存客户的成本低于获取新客户的成本。通过对服务取消事件进行生存分析，电信公司可以更有效地管理客户流失情况，从而更好地预测特定客户在何时可能面临流失风险。
- Hardware Failures(硬件故障): 客户对产品和服务的体验质量在其决定续约或取消服务时起着关键作用。网络本身是这种体验的核心。将硬件故障时间作为感兴趣事件，生存分析可用于预测硬件何时需要维修或更换。
- Device and Data Plan Upgrades(设备和数据套餐升级): 在客户生命周期中，存在一些关键节点会发生套餐变更。使用生存分析探究套餐变更事件，可用于预测这种变更何时会发生，进而采取行动积极影响客户对产品或服务的选择。

1 数据加载

下载 CSV 格式的数据集后，使用分析数据集表头的含义如表1所示（由于数据的表头大小写书写方式并不统一，采用小驼峰命名法便于后续操作）,churnString 在表中对应列名为 churn

1.1 数据集的转换与精炼

将未经过过滤与处理的数据表命名为 bronze, 并按照以下条件处理与筛选后得到 silver：

- 根据 churnString 进行处理：Yes 记为 1 代表已流失客户，No 记为 0 代表未流失客户，其他情况标注为 Unkbown 异常值，记为新的一列 churn
- 删除原始的字符串类型的 churnString 列
- 依照 contract 进行筛选: 通过 contract 列筛选合约类型为 Month-to-Month 的客户，以便于分析短期合约客户的流失
- 依照 internetService 筛选：排除没有使用互联网服务的客户

bronze 表的示例数据如图1所示，silver 表的示例数据如图所示

1.2 保存数据至 PARQUET 格式

原链接中数据保存至 Delta Lake 中，查询得知 Delta Lake 是一个开源存储层，它为数据湖（Data Lake）提供可靠性、安全性和高性能。它构建在 Apache Spark 之上，为大数据工作流带来 ACID 事务、数据版本控制和 Schema 管理等关键功能。

在 Jupyter 环境中，不易配置参数且无法达到性能优化的效果，因此采用了 Parquet 格式进行替换

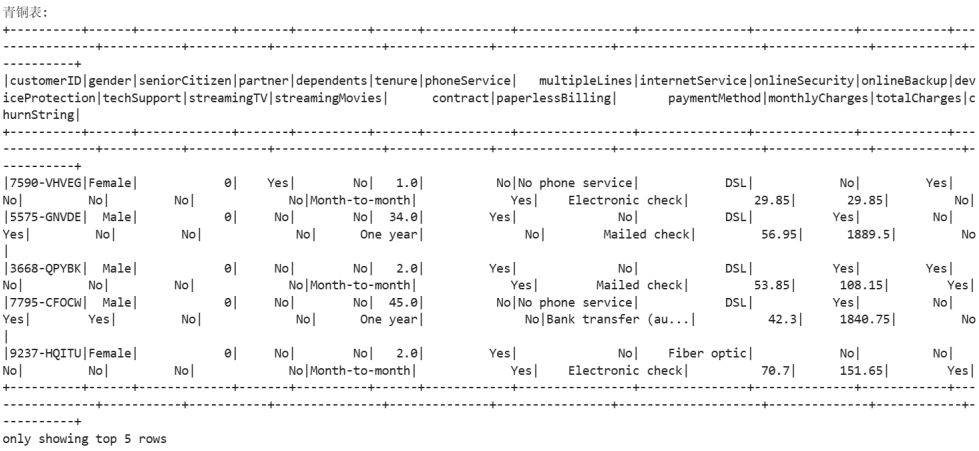
2 Kaplan-Meier 生存概率曲线绘制与分析

Kaplan-Meier 主要是分析单一因素对生存期的影响，在此数据处理中用于探究对于客户流失因素的分析

列名	意义	可能取值
customerID	唯一标识符，用于区分不同的客户	-
gender	客户的性别	Male, Female
seniorCitizen	是否为老年人	0, 1
partner	是否有配偶	Yes, No
dependents	是否有经济上的依赖人	Yes, No
tenure	在公司的使用时长，表示客户成为公司用户的月数	整数
phoneService	是否使用电话服务	Yes, No
multipleLines	是否使用多条线路服务	Yes, No, No phone service
internetService	使用的互联网服务类型	DSL, Fiber optic, No
onlineSecurity	是否使用在线安全服务	Yes, No
onlineBackup	是否使用在线备份服务	Yes, No
deviceProtection	是否使用设备保护服务	Yes, No
techSupport	是否使用技术支持服务	Yes, No
streamingTV	是否使用电视流媒体服务	Yes, No
streamingMovies	是否使用电影流媒体服务	Yes, No
contract	合同类型	One year 等
paperlessBilling	是否使用电子账单	Yes, No
paymentMethod	付款方式	Electronic check 等
monthlyCharges	每月的费用支出	数值
totalCharges	总费用支出，	数值
churnString	是否流失	Yes, No

表 1: 数据集表头含义说明 (已全部更改为小驼峰命名法)

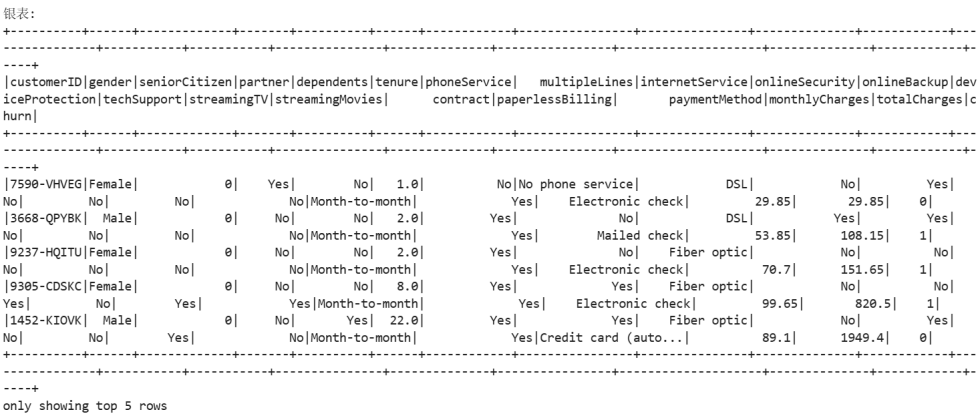
青铜表：



```
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----------+
|customerID|gender|seniorCitizen|partner|dependents|tenure|phoneService|   multipleLines|internetService|onlineSecurity|onlineBackup|dev
iceProtection|techSupport|streamingTV|streamingMovies|      contract|paperlessBilling|       paymentMethod|monthlyCharges|totalCharges|c
hurnString|
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----------+
|7590-VHVEG|Female|            0|    Yes|        No|   1.0|          No|No phone service|            DSL|           No|         Yes|
No|         No|         No|             No|Month-to-month|             Yes|    Electronic check|         29.85|       29.85|          No|
|5575-GNVDE|  Male|            0|     No|        No|  34.0|         Yes|              No|            DSL|          Yes|          No|
Yes|         No|         No|             No|      One year|              No|        Mailed check|         56.95|      1889.5|          No
|
|3668-QPYBK|  Male|            0|     No|        No|   2.0|         Yes|              No|            DSL|          Yes|         Yes|
No|         No|         No|             No|Month-to-month|             Yes|        Mailed check|         53.85|      108.15|         Yes|
|7795-CFOCW|  Male|            0|     No|        No|  45.0|          No|No phone service|            DSL|          Yes|          No|
Yes|        Yes|         No|             No|      One year|              No|Bank transfer (au...|         42.3|     1840.75|          No
|
|9237-HQITU|Female|            0|     No|        No|   2.0|         Yes|              No|    Fiber optic|           No|          No|
No|         No|         No|             No|Month-to-month|             Yes|    Electronic check|          70.7|      151.65|         Yes|
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----------+
only showing top 5 rows
```

图 1: bronze 示例数据

银表：



```
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----+
|customerID|gender|seniorCitizen|partner|dependents|tenure|phoneService|   multipleLines|internetService|onlineSecurity|onlineBackup|dev
iceProtection|techSupport|streamingTV|streamingMovies|      contract|paperlessBilling|       paymentMethod|monthlyCharges|totalCharges|c
hurn|
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----+
|7590-VHVEG|Female|            0|    Yes|        No|   1.0|          No|No phone service|            DSL|           No|         Yes|
No|         No|         No|             No|Month-to-month|             Yes|    Electronic check|         29.85|       29.85|    0|
|3668-QPYBK|  Male|            0|     No|        No|   2.0|         Yes|              No|            DSL|          Yes|         Yes|
No|         No|         No|             No|Month-to-month|             Yes|        Mailed check|         53.85|      108.15|    1|
|9237-HQITU|Female|            0|     No|        No|   2.0|         Yes|              No|    Fiber optic|           No|          No|
No|         No|         No|             No|Month-to-month|             Yes|    Electronic check|          70.7|      151.65|    1|
|9305-CDSKC|Female|            0|     No|        No|   8.0|         Yes|             Yes|    Fiber optic|           No|          No|
Yes|         No|        Yes|            Yes|Month-to-month|              Yes|    Electronic check|         99.65|       820.5|    1|
|1452-KIOVK|  Male|            0|     No|       Yes|  22.0|         Yes|             Yes|    Fiber optic|           No|         Yes|
No|         No|        Yes|             No|Month-to-month|             Yes|Credit card (auto...|         89.1|      1949.4|    0|
+----------+------+-------------+-------+----------+------+------------+---------------+---------------+-------------+------------+---
-------------+-----------+-----------+---------------+--------------+----------------+------------------+-------------+------------+-
----+
only showing top 5 rows
```

图 2: silver 示例数据

2.1 Kaplan-Meier 曲线原理

Kaplan-Meier 包含以下三个关键指标

- 事件（失效时间）：是二分类结局事件，例如，客户是否流失
- 生存时间：从检测开始到事件发生所经过的时间，例如，从用户被记录（时间为 0）到用户流失发生所经历的时间
- 删失（截尾）：研究对象在观察时间内没有发生事件。一种情况是研究对象在中途失访或退出；另一种情况是超过了最长的随访时间事件仍未发生。

Kaplan-Meier 曲线的绘制分为以下步骤

- 在每个发生死亡事件的时间点上，进行生存率的计算
- 在每个发生删失的时间点上，画小竖线标记删失样本
- 根据观察/对照条件分组作图

2.1.1 生存率的点估计与区间估计

设 n_{i-1}, n_i, d_i 分别表示活过时间 t_{i-1} 且未在 t_{i-1} 截尾的对象数、期初例数、死亡数，则时间 t_i 处的生存率估计为：

$$S(t) = \left(1 - \frac{d_1}{n_0}\right)\left(1 - \frac{d_2}{n_1}\right)\cdots\left(1 - \frac{d_i}{n_{i-1}}\right), \quad i = 1, 2, \cdots, k.$$

生存率标准误差的近似计算公式为：

$$SE[S(t_i)] = S(t_i)\sqrt{\sum_{j=1}^{i} \frac{d_j}{n_j(n_j - d_j)}}$$

假定生存率近似服从正态分布，则总体生存率的 $(1-\alpha)$ 置信区间为：

$$S(t_i) \pm z_{\alpha/2} \cdot SE[S(t_i)]$$

2.2 生存率的组间检验

Log rank 检验是指一种用于比较两组或多组生存数据的非参数统计方法。当零假设（即两组的生存函数相同）成立时，各时间点的理论死亡数和实际死亡数的差异应较小。如果差异显著，则拒绝零假设，表明两组的生存曲线存在差异。

Log rank 检验的统计量计算公式为

$$\sum \frac{(O-E)^2}{V}$$

，其中 O 是实际死亡数，E 是理论死亡数，V 是 O 的方差估计。

在等比例风险假设成立时，Log rank 检验与 Cox 比例风险模型关于组别变量的似然比检验结果相似。

2.3 电信行业案例的 Kaplan-Meier 曲线绘制与分析

2.3.1 生存率整体分析

使用 Lifelines 库，实现 Kaplan - Meier 估计方法

- 事件：表示客户是否流失
- 生存时间（T）：数据表中”tenure” 列，表示客户在电信公司的使用时长
- 删失（C）：数据表中”churn” 列，1 表示发生了感兴趣事件（客户流失），0 表示未发生（在观察期结束时客户仍未流失）

在拟合过程中 KaplanMeierFitter 的拟合函数会进行以下操作

- 初始化与数据整理：对输入的生存时间和删失数据进行整理，将数据按照生存时间从小到大排序，以便后续逐步计算
- 计算生存概率：从生存时间最小值开始，在每个时间点上，计算该时间点的生存概率。在每个时间点，计算仍处于“存活”（未流失）状态的个体比例。对于未删失个体，会直接参与计算生存概率的更新；对于删失个体，在其删失时间点上，认为其在该时间点之前是“存活”的，不影响之前时间点的生存概率计算，但之后不再参与生存概率更新。

- 构建生存函数：随着时间推进，不断更新生存概率，最终构建出一条完整的生存函数曲线。

对上述的 T 和 C 进行拟合，得到图3，由图可知（0，1.0）时代表客户至少存活 0 个月的概率是 100%，随着

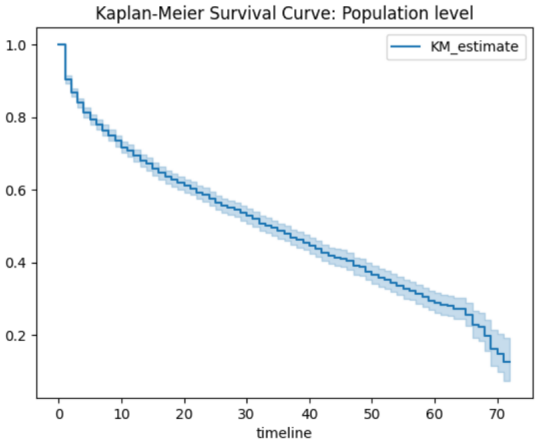


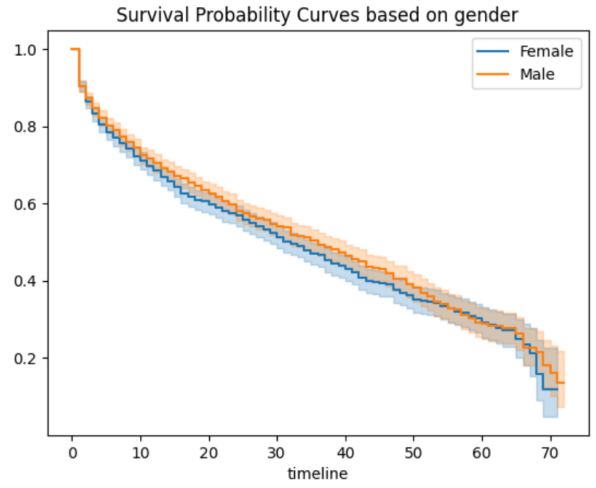
图 3: 整体的 Kaplan-Meier 曲线

时间推移，存在概率显著降低。围绕生存概率曲线的浅蓝色边框表示置信区间。区间越宽，置信度越低。如图所示，随着时间线的延长，对估计值的信心逐渐减弱

通过 KaplanMeierFitter 对象的 median_survival_time_ 函数获得中位生存时间为 34，即客户存活 34 个月的概率是 50%

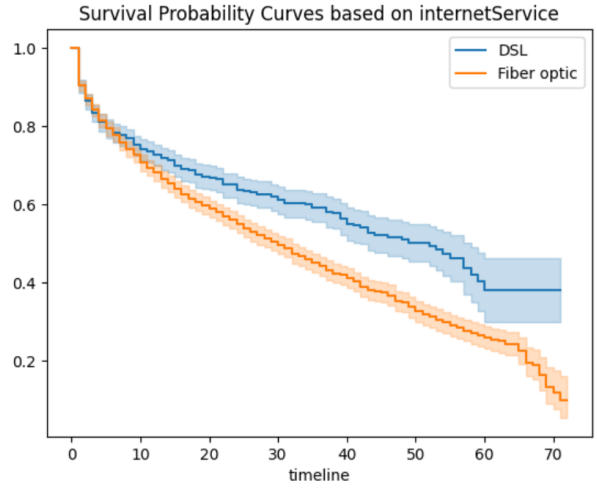
2.3.2 基于协变量评估生存概率

定义两个函数分别绘制基于特定一列数据的生存概率曲线，并对不同组进行 Log-Rank 检验，以检验不同组生存曲线之间是否存在显著差异。对所有协变量分析生存概率及其组间差异，得到图



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

(a)

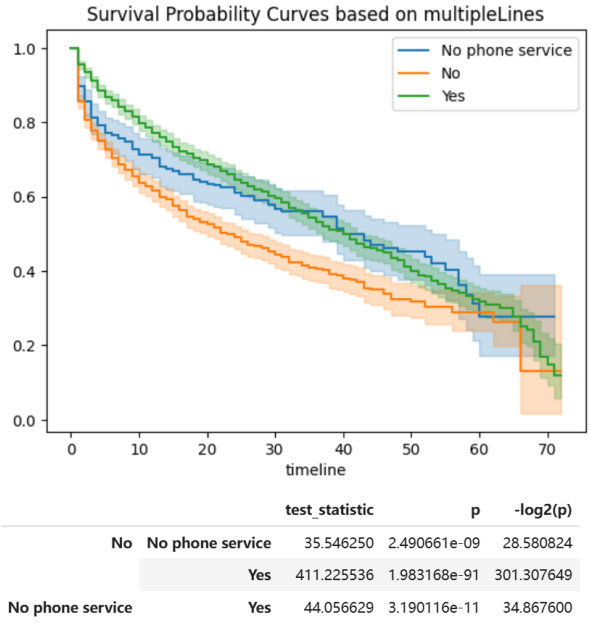


		test_statistic	p	-log2(p)
DSL	Fiber optic	85.455399	2.369872e-20	65.193753

(b)

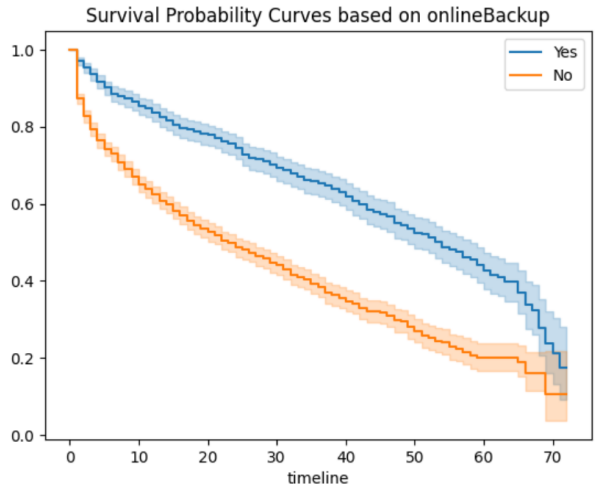
图 4: gender，internetService 各自组间差异

分析基于不同协变量的生存概率曲线及其统计量对应 p 值，可在置信水平为 95% 的条件下得到如下结论



		test_statistic	p	-log2(p)
No	No phone service	35.546250	2.490661e-09	28.580824
	Yes	411.225536	1.983168e-91	301.307649
No phone service	Yes	44.056629	3.190116e-11	34.867600

(a)

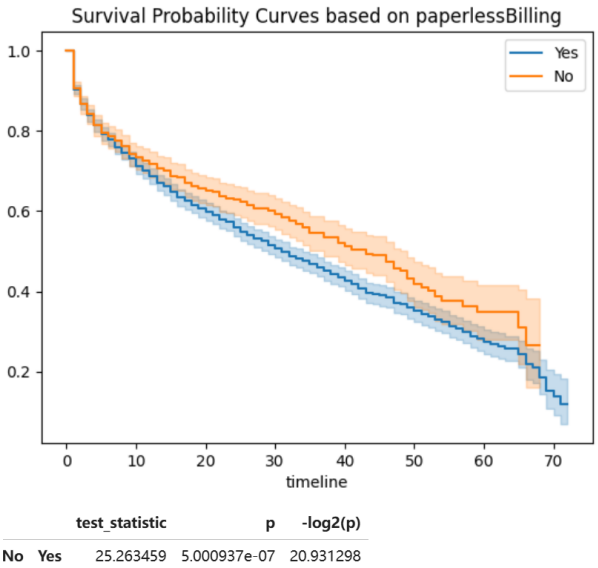


		test_statistic	p	-log2(p)
No	Yes	300.455875	2.620909e-67	221.179115

(b)

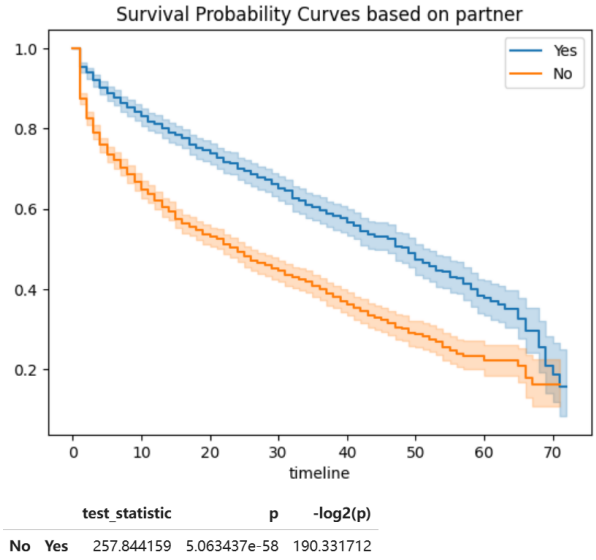
图 5: multipleLines，onlineBackup 各自组间差异

- dependents: 客户存留事件小于 70 月时，有经济依赖的客户存活率高于没有经济依赖的客户，70 个月后，有经济依赖的客户存活率突然低于没有经济依赖的客户，即有经济依赖的客户总体来说更不易流失（后续对图分析简化），且两组之间的差异是显著的
- deviceProtection: 使用设备保护服务的客户更不容易流失，是否使用设备保护服务的组间差异是显著的
- gender: 性别不是影响客户流失的显著因素
- internetService: 使用 DSL 互联网服务的客户相较于使用 Fiber optic 互联网服务的客户更不易流失，组间差异显著
- multipleLines: 使用多线路服务和没有手机服务的客户更不易损失，三种不同服务状态下，生存函数上存在显著差异
- onlineBackup: 使用在线备份服务的客户更不易流失，且与未使用的客户组差异显著
- paperlessBilling: 不适用电子账单的客户更不易流失，且组间的差异是显著的
- partner: 有配偶的用户更不易流失，且有无配偶的组间差异显著
- paymentMethod：电子支票和邮寄支票下降相对较快，自动信用卡支付和自动银行转账下降相对较慢，但电子支票与邮寄支票的生存概率差异并不显著，剩余的组间差异均显著
- streamingMovies: 使用电影流媒体服务的客户更不易流失，且组间差异显著
- streamingTV: 使用电视流媒体流媒体服务的客户更不易流失，且组间差异显著
- techSupport: 使用技术支持服务的客户更不易流失，且组间差异显著
- onlineSecurity: 使用在线安全服务的客户更不易流失，且组间差异显著



		test_statistic	p	-log2(p)
No	Yes	25.263459	5.000937e-07	20.931298

(a)



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

(b)

图 6: paperlessBilling，partner 各自组间差异

2.3.3 生存概率精确计算

我们想要探寻生存概率在每个阶段的准确数值，将整个时间线分为均匀分成 10 个时间节点（0-9），假设探究互联网服务类型为 DSL 时，生存概率的精确值如图12所示

可知在时间范围的中间时间点时，客户的流失率大概为 20%, 且从 0 时间点到中位时间点，客户流失率降低速度大于从中位时间点到时间最长所到时间的流失率

根据上述结果可知，有多个变量使得客户流失，因此使用 Cox 模型进行分析

3 Cox 比例风险模型

该模型的目的是同时评估几个因素对生存的影响，即允许我们检查特定因素如何影响特定时间点特定事件的发生率。该比率通常称为风险比率。预测变量（或因子）为协变量

3.1 Cox 比例风险模型原理

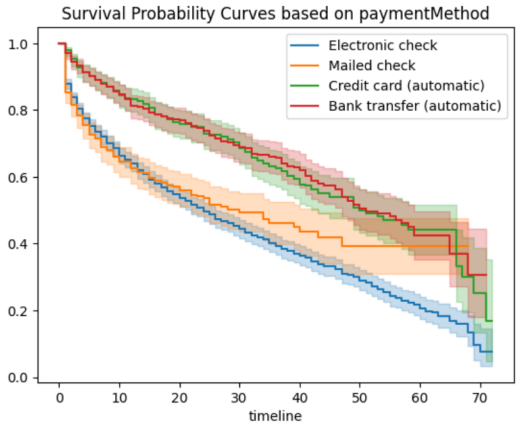
3.1.1 风险函数

用 $h(t)$ 表示风险函数，即在时间 t 死亡的风险。可以估计如下：

$$h(t) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + \ldots + b_px_p)$$

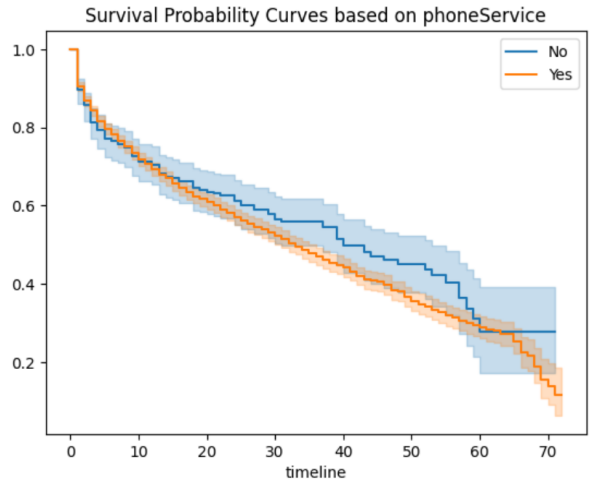
其中：

- t 表示生存时间
- $h(t)$ 是由一组 p 协变量（x_1, x_2, ..., x_p）确定的风险函数
- 系数（b_1, b_2, ..., b_p）衡量协变量的影响（即效应大小），可以理解为风险权重



		test_statistic	p	-log2(p)
Bank transfer (automatic)	Credit card (automatic)	0.153545	6.951703e-01	0.524562
	Electronic check	55.164654	1.108442e-13	43.036532
	Mailed check	190.000457	3.178566e-43	141.174532
Credit card (automatic)	Electronic check	45.167592	1.808736e-11	35.686227
	Mailed check	165.361074	7.628420e-38	123.301883
Electronic check	Mailed check	72.323100	1.826962e-17	55.603331

(a)



		test_statistic	p	-log2(p)
No	Yes	0.778505	0.377599	1.405074

(b)

图 7: paymentMethod，phoneService 各自组间差异

- 术语 h_0 称为基线危险。如果所有的系数等于零（既 exp(0) 等于 1），则对应于风险的值。$h(t)$ 中的 “t” 表示该数值可能随时间而变化。

Cox 模型可以被写为变量 $x(i)$ 的危险对数的多元线性回归，而基线危险是随时间变化的 “截距” 项。

系数 b_i 称为危险比率（HR, hazard ratio）。b_i 值大于零，或相当于风险比率大于 1，表明随着第 i 个协变量值的增加，事件风险增加，因此生存时间缩短。可以理解为，风险比大于 1 表示协变量与事件概率正相关，因此与存活时间负相关. 可以总结为

- $HR = 1$: 无影响
- $HR < 1$: 危害降低
- $HR > 1$: 危险增加

3.1.2 模型假设

以电信用户为例，假设两个 x 值不同的用户 k 和 k'。相应的风险函数可以简单地写成如下形式用户 k 的风险函数:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x_i}$$

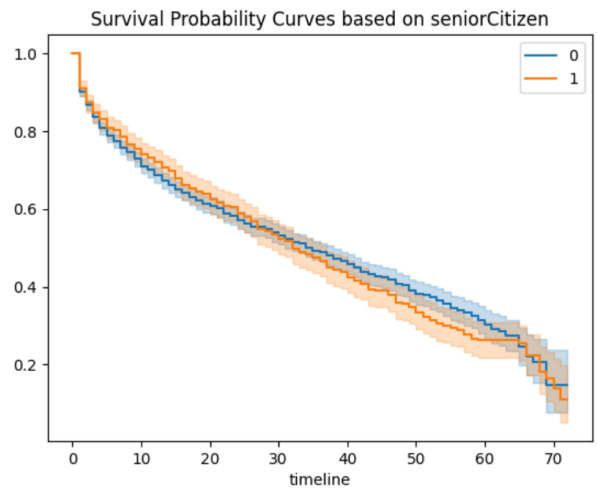
用户 k' 的风险函数:

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n}\beta x_i'}$$

这两名用户的危险比与时间 t 无关:

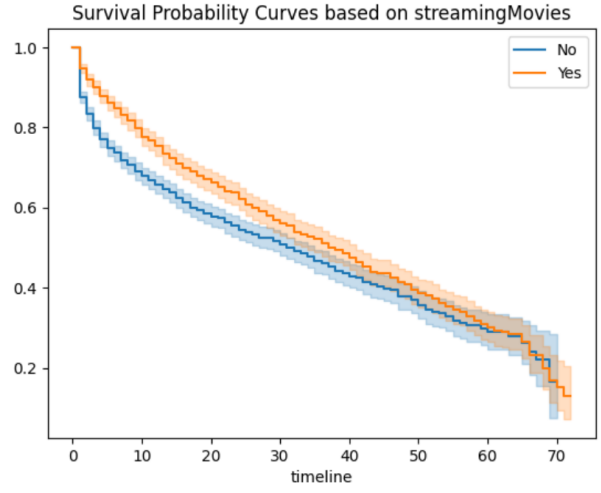
$$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^{n}\beta x_i}}{h_0(t)e^{\sum_{i=1}^{n}\beta x_i'}} = \frac{e^{\sum_{i=1}^{n}\beta x_i}}{e^{\sum_{i=1}^{n}\beta x_i'}}$$

因此，Cox 模型是一个比例风险模型：任何一组事件的风险都是其他任何一组事件风险的常数倍。这一假设意味着，各组的危险曲线应成比例，不能交叉



		test_statistic	p	-log2(p)
0	1	49.027784	2.523624e-12	38.52764

(a)



		test_statistic	p	-log2(p)
No	Yes	170.262183	6.484901e-39	126.858111

(b)

图 8: seniorCitizen，streamingMovies 各自组间差异

3.2 基于 Cox 模型的电信行业案例分析

首先分析数据集，发现许多列是分类变量变量，但存为 “YES” / “NO”，对所有分类变量进行独热编码，存为以 _Yes 和 _No 的列，变量取值为 0 或 1。当我们想要研究 dependents、internetService、onlineBackup、techSupport 四个协变量对于生存情况（客户流失）的影响时，选取 dependents_Yes、internetService_DSL、onlineBackup_Yes、techSupport_Yes、churn、tenure 做为一个新的 DataFrame。

3.2.1 模型拟合与结果分析

接着创建一个 CoxPHFitter 类的实例 cph，使用上述数据来拟合考克斯比例风险模型，其中 tenure 作为时间变量，churn 作为事件变量

拟合后的模型摘要如图13所示，结果解释如下

model `lifelines.CoxPHFitter`，表明使用 `lifelines` 库中的 `CoxPHFitter` 类来构建考克斯比例风险模型。

duration col `'tenure'`，即生存时间变量为客户在网时长。

event col `'churn'`，事件变量是客户流失情况。

baseline estimation `breslow`，采用 Breslow 方法估计基线风险。

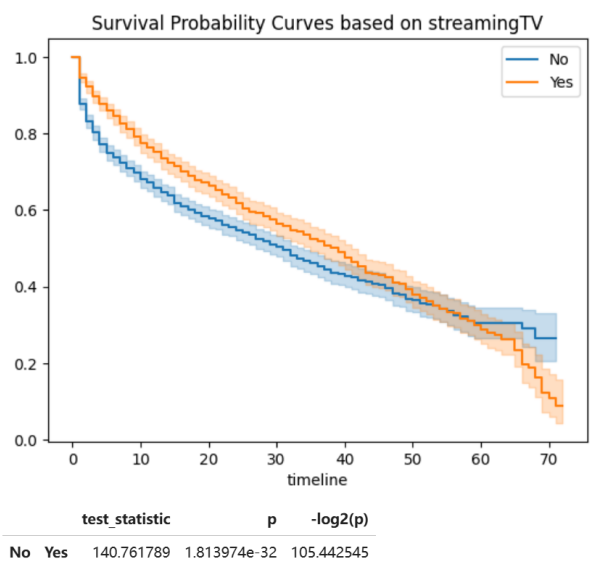
number of observations `3351`，样本数量为 3351 个。

number of events observed `1556`，观测到的事件（客户流失）数量为 1556 个。

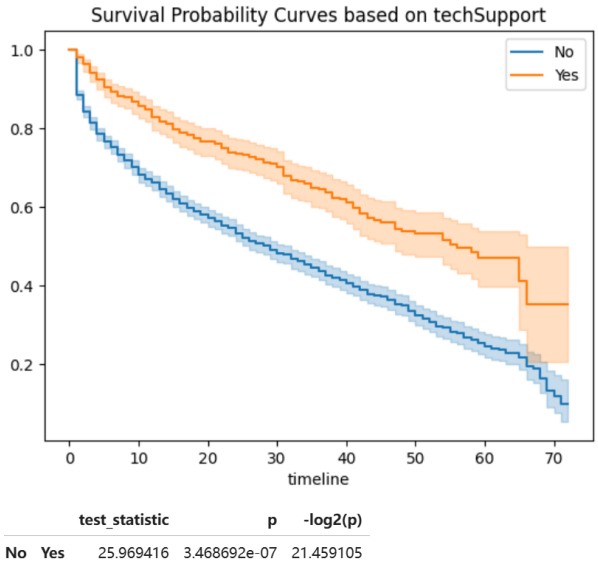
partial log - likelihood `-11315.95`，部分对数似然值，该值越大说明模型对数据的拟合程度相对越好，这里数值为负且绝对值较大，说明模型还有优化空间。

time fit was run `2025-04-12 05:49:43 UTC`，模型拟合的时间。

变量与相关统计量分析如下：



(a)



(b)

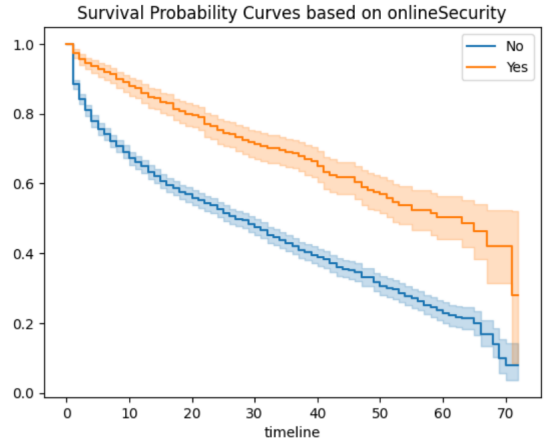
图 9: streamingTV&techSupport 各自组间差异

`dependents_Yes`
- **coef**：`-0.33`，回归系数为负，意味着在其他条件不变时，有家属（`dependents` 为 `Yes`）这一因素会降低客户流失风险。
- **exp(coef)**：`0.72`，即风险比（HR ）为 0.72 ，表示有家属的客户流失风险是无家属客户的 0.72 倍，进一步说明有家属降低了流失风险。
- **se(coef)**：`0.07`，系数的标准误差，衡量系数估计的精度。
- **coef lower 95%** 和 **coef upper 95%**：`-0.47` 和 `-0.19`，系数的 95% 置信区间，说明真实系数有 95% 的概率在这个区间内。
- **exp(coef) lower 95%** 和 **exp(coef) upper 95%**：`0.63` 和 `0.83`，风险比的 95% 置信区间。
- **z**：`-4.64`，z 值用于检验系数是否显著不为 0，绝对值越大越显著。
- **p**：`<0.005`，p 值极小，远小于常见显著性水平（如 0.05 ），表明该变量对客户流失风险的影响显著。
- **-log2(p)**：`18.12`，对 p 值进行 -log2 变换的值，越大表示越显著。

`internetService_DSL`
- **coef**：`-0.22`，系数为负，说明使用 DSL 互联网服务会降低客户流失风险。
- **exp(coef)**：`0.80`，风险比为 0.80 ，即使用 DSL 服务的客户流失风险是其他情况的 0.8 倍。
- 其他统计量类似上述变量，z 值为 `-3.68`，p 值 `<0.005`，表明该变量对流失风险的影响显著。

`onlineBackup_Yes`
- **coef**：`-0.78`，负系数表示使用在线备份服务会大幅降低客户流失风险。
- **exp(coef)**：`0.46`，风险比 0.46 ，即使用在线备份服务的客户流失风险仅为未使用客户的 0.46 倍。
- z 值 `-13.13`，p 值 `<0.005`，显著性极高。

`techSupport_Yes`
- **coef**：`-0.64`，意味着获得技术支持服务会降低客户流失风险。
- **exp(coef)**：`0.53`，风险比 0.53 ，获得技术支持服务的客户流失风险是未获得客户的 0.53 倍。
- z 值 `-8.48`，p 值 `<0.005`，影响显著。



		test_statistic	p	-log2(p)
No	Yes	75.800079	3.138886e-18	58.144453

图 10: onlineSecurity 的组间差异

	DSL
0	1.000000
1	0.902698
2	0.864380
3	0.834702
4	0.810522
5	0.794352
6	0.783900
7	0.776362
8	0.768486
9	0.750833

图 11: 互联网服务类型为 DSL 时生存概率的精确值

模型整体评估统计量:

Concordance `0.64`，也叫一致性指数（C - index ），衡量模型预测事件发生顺序的能力，取值范围 0 - 1 ，越接近 1 说明模型预测能力越强，0.64 表明模型有一定预测能力但仍有提升空间。

Partial AIC `22639.90`，部分 Akaike 信息准则，用于比较不同模型的拟合优度，在模型选择中，AIC 值越小越好，该值较大说明模型可能存在过度拟合或还需优化。

log - likelihood ratio test `337.77 on 4 df`，对数似然比检验统计量为 337.77 ，自由度为 4 。对数似然比检验用于比较包含不同变量的模型拟合优度。

-log2(p) of ll - ratio test `236.24`，对数似然比检验 p 值的 -log2 变换值，数值很大，表明对数似然比检验的 p 值极小，即模型中包含这些变量是显著合理的。

为了能更加直观展示各个变量对风险的影响情况，绘制风险比图如图14所示，发现四个变量的风险比均小于 1 ，且它们的 95% 置信区间都不包含 1 ，表明这些变量都显著降低了客户流失风险。从风险比数值来看，onlineBackup_Yes 降低客户流失风险的作用相对最明显，其次是 techSupport_Yes、dependents_Yes ，internetService_DSL 降低风险的作用相对较弱，但都有显著影响

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 05:49:43 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

图 12: Cox 模型拟合摘要

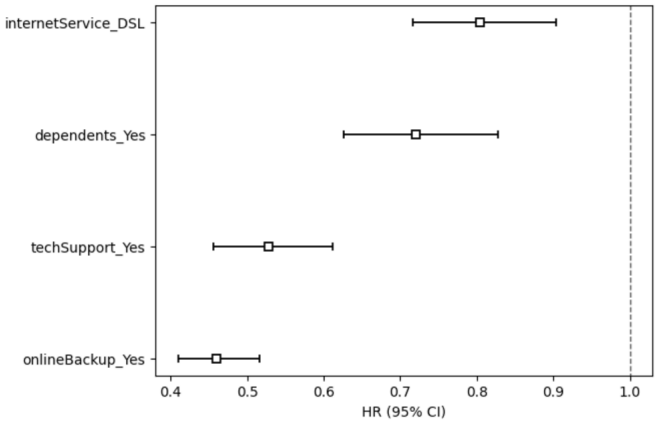


图 13: Cox 拟合模型的风险比图

3.2.2 Cox 模型假设检验

CoxPHFitter 类提供 check_assumptions 方法，专门用于对考克斯比例风险模型的比例风险假设进行检验, 设阈值为 0.05,得到图15至图19,变量“dependents_Yes”未违反比例风险假设,而“internetService_DSL”“onlineBackup_Yes” “techSupport_Yes” 显著违反该假设

由于模型多次违反比例风险假设。所以选择 Log-log Kaplan-Meier 来分析问题

3.2.3 Log-log Kaplan-Meier

顾名思义，该技术是在对数 - 对数尺度上绘制曲线，以 techSupport 为例，得到图20, 在 log(timeline) 取值为 1 到 3 之间时，两条曲线的平行程度相对较高，说明在这个时间段内，“No” 组和 “Yes” 组的风险比例相对稳定，比较符合比例风险假设, 当 log(timeline) 小于 1 或大于 3 时，两条曲线的平行程度变差，表明在这些时间段内，两组的风险比例可能不再保持恒定，存在违反比例风险假设的迹象

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

图 14: Cox 拟合模型假设检验 1

4 加速失效时间模型 (Accelerated Failure Time Model)

加速失效时间模型的核心思想是假设存在一些协变量（在数据中体现为某些特征列）会加速或延缓观测对象的失效时间。通过对包含生存时间和事件信息的数据进行拟合，模型可以估计出这些协变量对失效时间的影响系数等参数

AFT 模型与 Cox 模型的区别在于，它为基准失效时间指定了概率分布形式，从而确定了基准风险函数的形式，因此它是一个参数模型

4.1 AFT 原理

用 T 表示生存时间（survival time），$S(t) = P(T > t)$ 表示生存函数。

$F(t) = 1 - S(t) = P(T \le t)$ 表示 T 的累积分布函数，$f(t) = F'(t)$ 表示 T 的概率密度函数。

$h(t) = \lim_{\Delta t \to 0} \frac{P(t<T\le t+\Delta t|T\ge t)}{\Delta t} = \frac{f(t)}{S(t)}$ 表示风险函数。

AFT 模型假设某些与主体有关的因素会加速或延缓主体生存率到达某一水平的时间。与 Cox 模型类似，AFT 使用下式表示加速因子（acceleration factor）：

$$\eta_i = e^{X_i\beta}$$

这样，$S_i(t)$ 与 $S_0(t)$ 就存在如下关系：

$$S_i(t) = S_0(\eta_i t) = S_0(e^{X_i\beta} \cdot t)$$

对于基准风险函数，因为

$$h_0(t) = \frac{f_0(t)}{S_0(t)} = \frac{-S_0'(t)}{S_0(t)}$$

那么实际上主体 i 的风险函数 $h_i(t)$

$$\begin{aligned} h_i(t) &= \frac{-S_i'(t)}{S_i(t)} \\ &= \frac{-\eta_i S_0'(\eta_i t)}{S_0(\eta_i t)} \\ &= \eta_i h_0(\eta_i t) \\ &= h_0(e^{X_i\beta} \cdot t) \cdot e^{X_i\beta} \end{aligned}$$

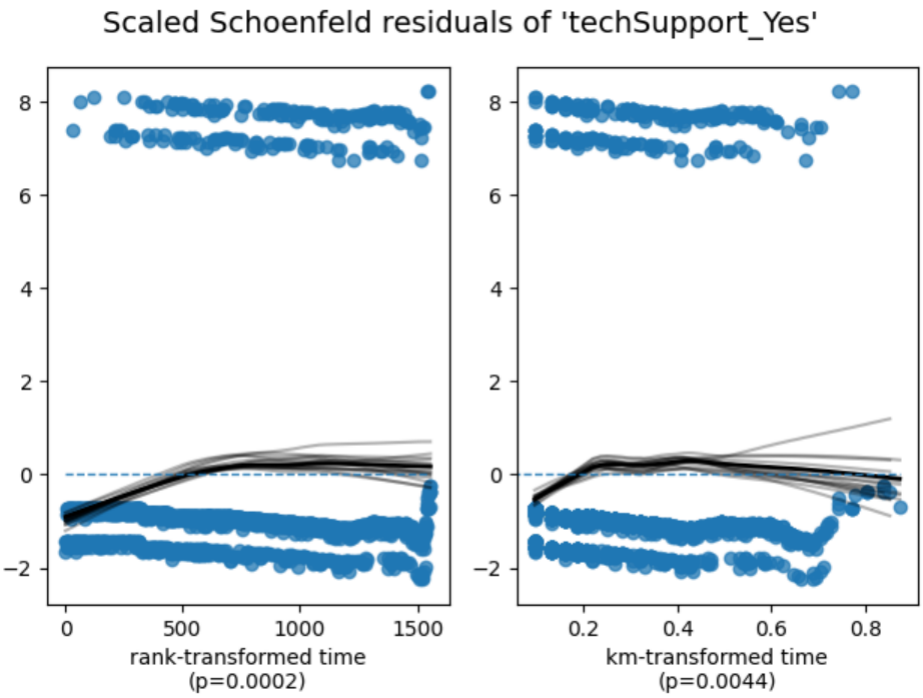


图 15: Cox 拟合模型假设检验 2

对失效时间 T_i 进行建模，有

$$T_i = \eta_i t_i$$

其中，t_i 表示在基准生存函数下的失效时间。两边取对数得

$$\ln T_i = X_i\beta + \ln t_i$$

t_i 是一个随机变量，服从一定的概率分布形式。对 $\ln t_i$ 进行标准化得

$$\epsilon_i = \frac{\ln t_i - \mu}{\sigma}$$

代入得

$$\ln T_i = \mu + X_i\beta + \sigma\epsilon_i$$

上式称为 AFT 模型的一般形式，σ 称作尺度参数

4.2 基于 AFT 模型的电信用户数据分析

以研究客户的一些服务选择情况（如是否有伴侣 partner_Yes 、是否有多条线路 multipleLines_Yes 等服务相关特征）以及支付方式（paymentMethod_Bank transfer (automatic) 、paymentMethod_Credit card (automatic) ）对客户生存时间（在网时长）和流失风险的影响为例，即通过构建加速失效时间模型等生存分析方法，分析这些因素如何加速或延缓客户流失这一事件的发生。

4.2.1 AFT 模型拟合

构造 LogLogisticAFTFitter 类的实例，用包含上述所有数据列的 DataFrame 进行 AFT 模型的拟合，共有 3351 条数据，其中客户流失事件有 1556 件，拟合模型的摘要信息如图21所示

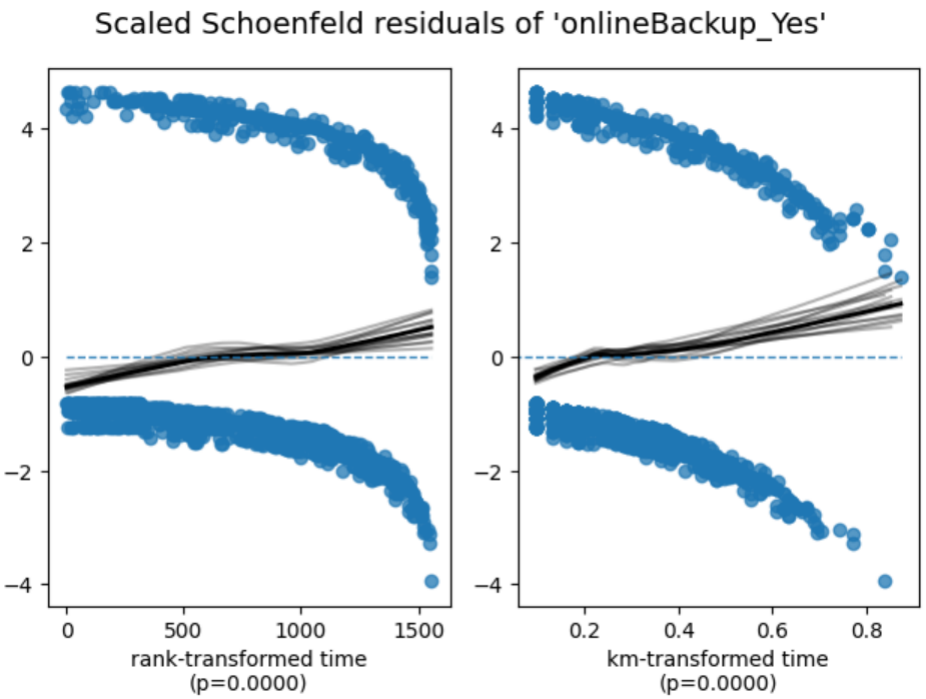


图 16: Cox 拟合模型假设检验 3

分析可知 1. deviceProtection_Yes（是否有设备保护服务）- 系数（coef）：0.48 ，表明在其他条件不变的情况下，拥有设备保护服务会使生存时间的对数增加 0.48 。从实际意义来说，即拥有该服务与生存时间延长存在正相关关系。- exp(coef)（风险比）：1.62 ，意味着拥有设备保护服务的对象，其失效风险是没有该服务对象的 1.62 倍。这看似与系数暗示的生存时间延长矛盾，实际是因为在生存分析中，风险比大于 1 表示风险增加，但结合系数可知，这里是在对数尺度下生存时间增加，只是风险相对也有所上升。- 置信区间：系数 95% 置信区间为 (0.35, 0.62) ，不包含 0；风险比 95% 置信区间为 (1.41, 1.86) ，不包含 1 ，说明该变量与生存时间及风险均存在显著关联。- z 值与 p 值：z 值为 6.88 ，p 值小于 0.005 ，强烈拒绝系数为 0 的原假设，即该变量对生存时间有显著影响。

2. internetService_DSL（是否使用 DSL 互联网服务）- 系数（coef）：0.38 ，表示使用 DSL 互联网服务会使生存时间对数增加 0.38 ，暗示使用该服务可能延长生存时间。- exp(coef)（风险比）**：1.47 ，说明使用 DSL 互联网服务的对象，其失效风险是未使用对象的 1.47 倍。- 置信区间：系数置信区间 (0.23, 0.53) 、风险比置信区间 (1.26, 1.71) 均不包含 0 和 1 ，表明变量与生存时间和风险显著相关。- z 值与 p 值：z 值 4.98 ，p 值小于 0.005 ，说明该变量对生存时间影响显著。

3. multipleLines_Yes（是否有多条线路）- 系数（coef）：0.66 ，有多条线路会使生存时间对数增加 0.66 ，体现与生存时间的正相关。- exp(coef)（风险比）**：1.94 ，拥有多条线路的对象失效风险是没有多条线路对象的 1.94 倍。- 置信区间：系数 (0.53, 0.80) 、风险比 (1.70, 2.22) 置信区间不包含 0 和 1 ，显著相关。- z 值与 p 值：z 值 9.64 ，p 值小于 0.005 ，显著影响生存时间。

4. onlineBackup_Yes（是否有在线备份服务）- 系数（coef）：0.81 ，有在线备份服务使生存时间对数增加较多，为 0.81 。- exp(coef)（风险比）：2.25 ，拥有在线备份服务的对象失效风险相对较高，是无此服务对象的 2.25 倍。- 置信区间：系数 (0.68, 0.95) 、风险比 (1.97, 2.59) 不包含 0 和 1 ，关联显著。- z 值与 p 值：z 值 11.63 ，p 值小于 0.005 ，显著影响生存时间。

5. onlineSecurity_Yes（是否有在线安全服务）- 系数（coef）：0.86 ，表明在线安全服务与生存时间对数

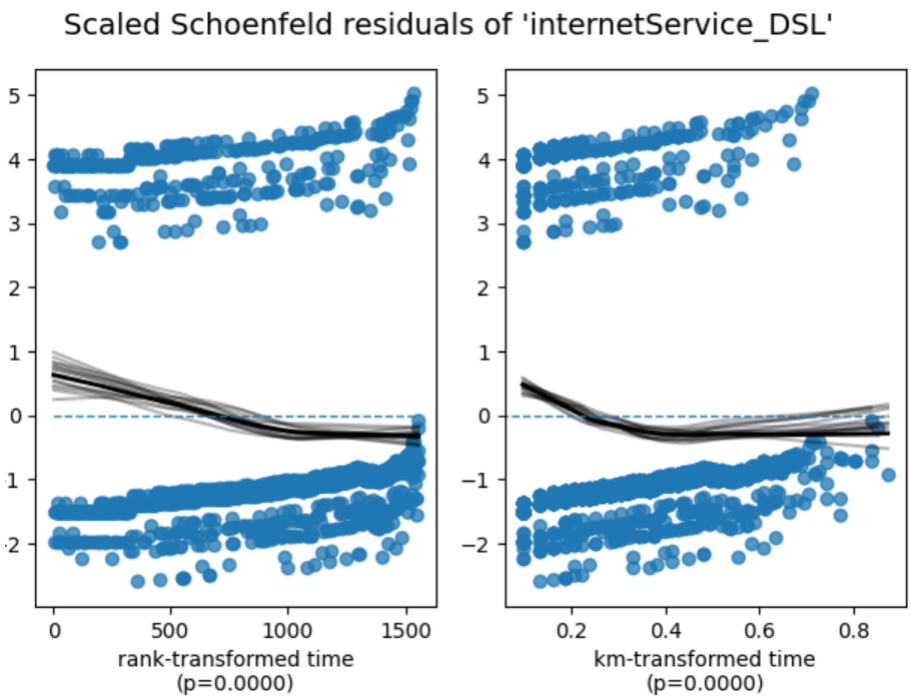


图 17: Cox 拟合模型假设检验 4

呈正相关，增加量为 0.86 。- exp(coef)（风险比）：2.37 ，有在线安全服务的对象失效风险是无此服务对象的 2.37 倍。- 置信区间：系数 (0.69, 1.03) 、风险比 (2.00, 2.80) 不包含 0 和 1 ，变量显著影响生存时间和风险。- z 值与 p 值：z 值 10.12 ，p 值小于 0.005 ，影响显著。

6. partner_Yes（是否有伴侣）- 系数（coef）：0.68 ，有伴侣会使生存时间对数增加 0.68 。- exp(coef)（风险比）：1.97 ，有伴侣的对象失效风险是无伴侣对象的 1.97 倍。- 置信区间：系数 (0.55, 0.81) 、风险比 (1.73, 2.24) 不包含 0 和 1 ，显著相关。- z 值与 p 值：z 值 10.21 ，p 值小于 0.005 ，对生存时间有显著影响。

7. paymentMethod_Bank transfer (automatic)（是否使用自动银行转账支付）- 系数（coef）：0.74 ，使用自动银行转账支付会使生存时间对数增加 0.74 。- exp(coef)（风险比）：2.10 ，采用该支付方式的对象失效风险是其他支付方式对象的 2.10 倍。- 置信区间：系数 (0.56, 0.92) 、风险比 (1.75, 2.51) 不包含 0 和 1 ，显著影响生存时间和风险。- z 值与 p 值：z 值 8.05 ，p 值小于 0.005 ，影响显著。

8. paymentMethod_Credit card (automatic)（是否使用自动信用卡支付）- 系数（coef）：0.80 ，自动信用卡支付与生存时间对数呈正相关，增加量为 0.80 。- exp(coef)（风险比）：2.22 ，使用该支付方式的对象失效风险是其他支付方式对象的 2.22 倍。- 置信区间：系数 (0.61, 0.99) 、风险比 (1.84, 2.68) 不包含 0 和 1 ，显著相关。- z 值与 p 值：z 值 8.36 ，p 值小于 0.005 ，显著影响生存时间。

9. techSupport_Yes（是否有技术支持服务）- 系数（coef）：0.69 ，有技术支持服务使生存时间对数增加 0.69 。- exp(coef)（风险比）：1.99 ，有技术支持服务的对象失效风险是无此服务对象的 1.99 倍。- 置信区间：系数 (0.52, 0.86) 、风险比 (1.68, 2.36) 不包含 0 和 1 ，显著影响生存时间和风险。- z 值与 p 值：z 值 7.90 ，p 值小于 0.005 ，对生存时间影响显著。

10. Intercept（截距）- alpha_ 部分：系数为 1.59 ，exp(coef) 为 4.91 ，表示在所有自变量都为 0 的情况下，生存时间对数的基础值为 1.59 ，对应的风险比为 4.91 。置信区间 (1.46, 1.72) 不包含 0 ，z 值 24.47 ，p 值小于 0.005 ，说明截距项显著。- 在 beta_ 部分：系数为 0.12 ，exp(coef) 为 1.13 ，同样表示基础值相关信息，置信区间 (0.08, 0.16) 不包含 0 ，z 值 5.71 ，p 值小于 0.005 ，截距项显著。截距反映了模型中未包含

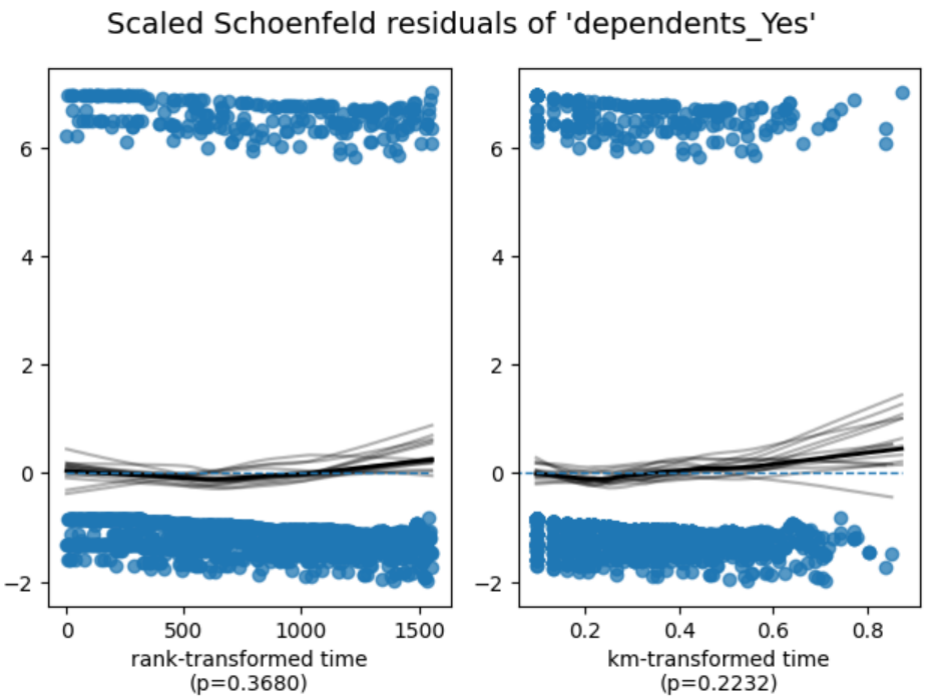


图 18: Cox 拟合模型假设检验 5

变量以及基础生存情况的综合影响。

生成 AFT 模型森林图如图22所示，每个变量对应一个方块和误差线，方块代表估计值，误差线代表 95% 置信区间。竖虚线对应 log (加速失败率) 为 0 处。若置信区间不与竖虚线相交，说明该变量对加速失败率有显著影响，得到与上述相同结论 中位生存时间的值为 135.51

4.2.2 假设检验

AFT 模型假设有两个

- 比例优势假设 (Proportional Odds)：判断模型是否符合该假设，要看对数 - 对数图中的线是否平行。若平行，则满足该假设
- 分布合理性假设：判断指定分布对模型是否合适，要看对数 - 对数图中的线是否呈直线。若呈直线，说明指定分布合适

定义一个接受协变量列名的函数，针对每个取值筛选数据，用筛选后的数据拟合 Kaplan - Meier 模型，然后计算对数赔率和对数时间，再绘制曲线，以 partner 为例绘制图23 两条折线既不是完全平行，也不是完全笔直的直线。在评估加速失效时间模型假设时，平行性用于判断是否符合比例优势假设，直线性用于判断指定分布是否合适。因此从这张图初步判断，可能在一定程度上违背了比例优势假设，且指定分布的合理性也并不可靠

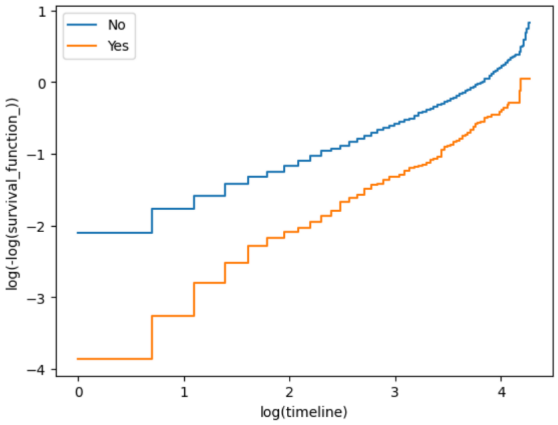


图 19: techSupport 的 Log-log Kaplan-Meier 图

		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

图 20: AFT 拟合模型摘要信息

5 评估客户终身价值

5.1 终身价值计算

继续研究上述问题定义一个函数来计算不同合同月份下客户所选套餐的月利润、平均预期月利润、平均预期月利润的净现值以及累计净现值等数据，定义如下

1. 生存概率：概率是通过使用 `predict_survival_function()` 从模型中提取的。
2. 所选套餐的月利润：目前为说明目的硬编码为 30 。在实际应用中，应使用相应的内部数据。
3. 平均预期月利润：一般来说，给定客户的预期月利润为生存概率 × 相应套餐的月利润。
4. 平均预期月利润的净现值：由于现在收到的一美元比几年后收到的一美元更值钱，因此通常使用净现值。使用的公式为：
   - 平均预期月利润 / ((1 + 内部收益率) ^ 合同月份)

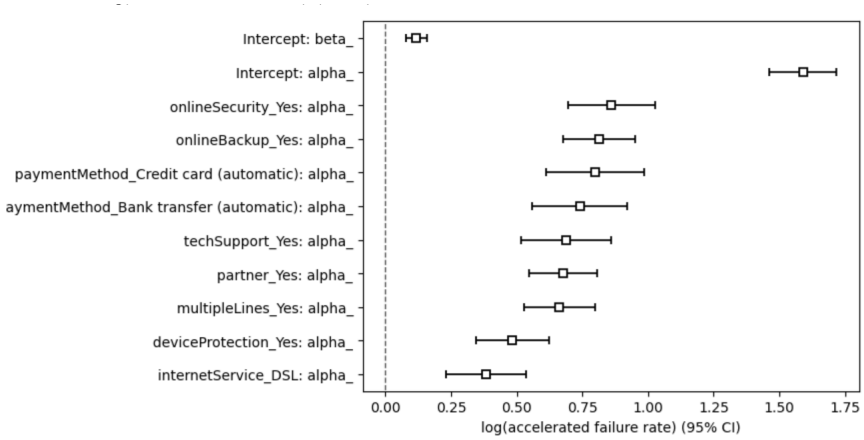


图 21: AFT 拟合模型森林图

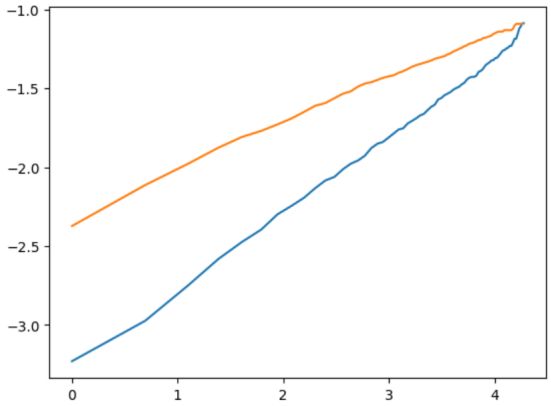


图 22: 基于 partner 的模型假设检验

- 内部收益率的默认值设置为 10%，但实际应使用所在业务领域认可的数值。

5. 累计净现值：平均预期月利润的净现值的累计和。

5.2 累积净现值可视化

展示不同时间点（12 个月、24 个月、36 个月）对应的累计净现值（Cumulative NPV ）情况如图24所示

绘制生存概率曲线图如图25所示 由图可知，随着合同月份增加，曲线呈明显下降趋势。说明随着时间推移，客户发生流失等失效事件的可能性逐渐增大，生存概率不断降低。在前期（大约前 10 个月）下降速度较快，之后下降速度有所减缓，但仍持续下降。

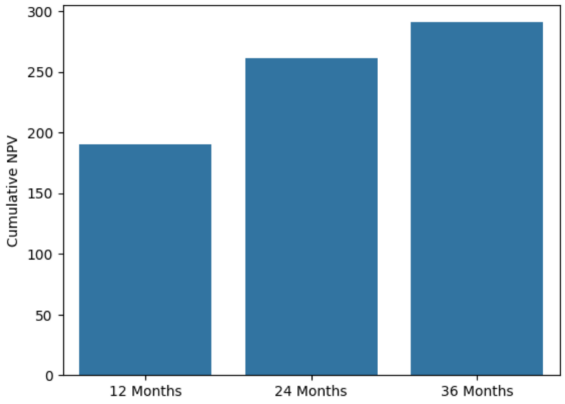


图 23: 累计净现值

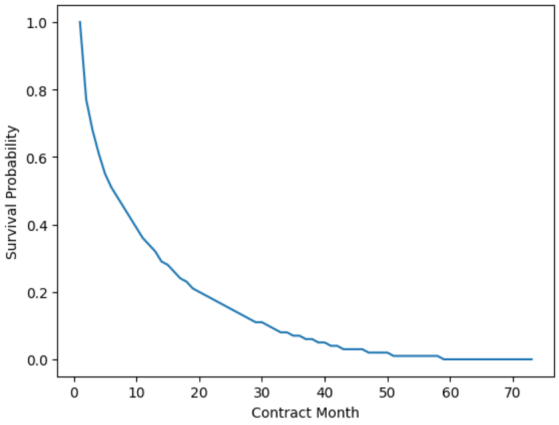


图 24: 生存概率曲线