Project1-生存分析案例报告

一、数据导入

首先，我导入了一份客户相关的数据，里面包括每位客户的 ID、开始和结束日期、是否流失的标记（churn）和持续时间（tenure）等。在导入数据的过程中要进行基础清洗，比如：处理特殊空格、进行类型转换、根据指定条件进行初步筛选。

然后，进行针对生存分析任务的预处理，形成 silver data，这个过程包括：保留关键特征、处理缺失值以及保证数据有效性，举例而言，时间列不能出现负值。

最后，生成一个数据状态简报：

```
数据状态简报:
样本量          3875
事件发生率       42.7%
平均持续时间     18.0月
特征维度          17
dtype: object

数据结构示例:
   tenure  Churn  gender Partner Dependents
0       1      0  Female     Yes         No
2       2      1    Male      No         No
4       2      1  Female      No         No
```

二、Kaplan-Meier 生存曲线

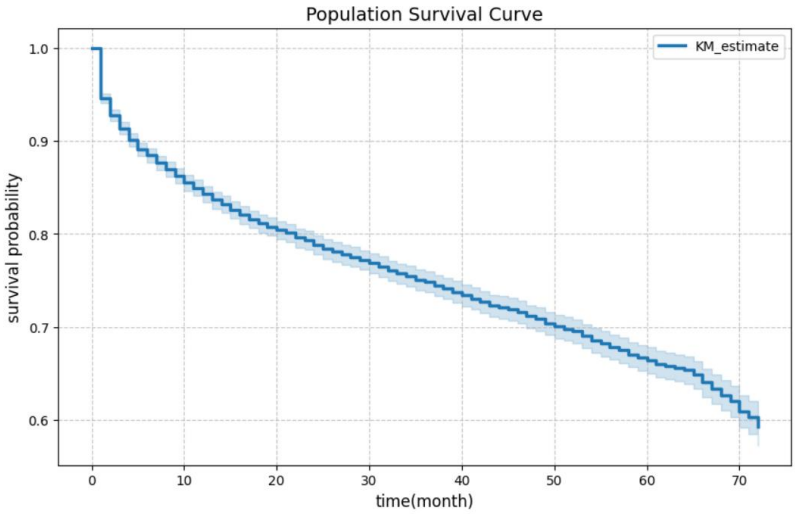
Kaplan-Meier 是一种很经典的生存分析方法，它用来估计在每个时间点，客户还“活着”（即还在使用服务、没有流失）的比例。

我用 lifelines 库来做了拟合，并画出了生存曲线：

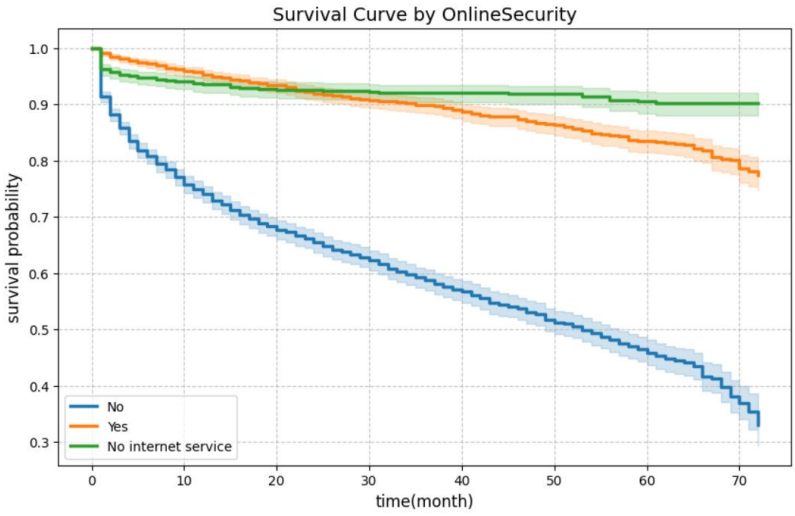
kmf = KaplanMeierFitter()

从图中可以看出，随着时间的推移，客户流失逐渐增加，曲线也逐步下降。如果某一段时间曲线比较平坦，那就说明那个时间段流失较少，客户相对稳定。

我不仅做了整体的拟合，还做了单个协变量的拟合以及 logrank 检验，见下图：




```
<Figure size 640x480 with 0 Axes>
```




```
Log-rank检验结果 (OnlineSecurity):
                                      test_statistic             p  \
No                 No internet service      485.975805  1.070135e-107
                   Yes                      660.525069  1.148535e-145
No internet service Yes                      11.060731   8.817539e-04

                                       -log2(p)
No                 No internet service  355.348514
                   Yes                  481.479779
No internet service Yes                  10.147336
Log-rank检验结果 (gender):
             test_statistic         p  -log2(p)
Female Male        0.525707  0.468417  1.094134
Log-rank检验结果 (Partner):
         test_statistic             p    -log2(p)
No Yes       423.543082  4.132951e-94  310.214069
Log-rank检验结果 (Dependents):
         test_statistic             p   -log2(p)
No Yes       232.699042  1.537238e-52  172.11992
Log-rank检验结果 (OnlineSecurity):
                                      test_statistic             p  \
No                 No internet service      485.975805  1.070135e-107
                   Yes                      660.525069  1.148535e-145
No internet service Yes                      11.060731   8.817539e-04

                                       -log2(p)
No                 No internet service  355.348514
                   Yes                  481.479779
No internet service Yes                  10.147336
<Figure size 640x480 with 0 Axes>
<Figure size 640x480 with 0 Axes>
<Figure size 640x480 with 0 Axes>
<Figure size 640x480 with 0 Axes>
```

]:
```
# (3)cox_proportional_hazards
```

5]:
```
"""
包含：列名兼容性处理、数据验证增强、错误引导提示
"""
import os
```

三、Cox 比例风险模型

接下来，我们用到了 Cox 比例风险模型。这种模型可以帮我们判断，到底是哪些因素在影响客户是否流失。

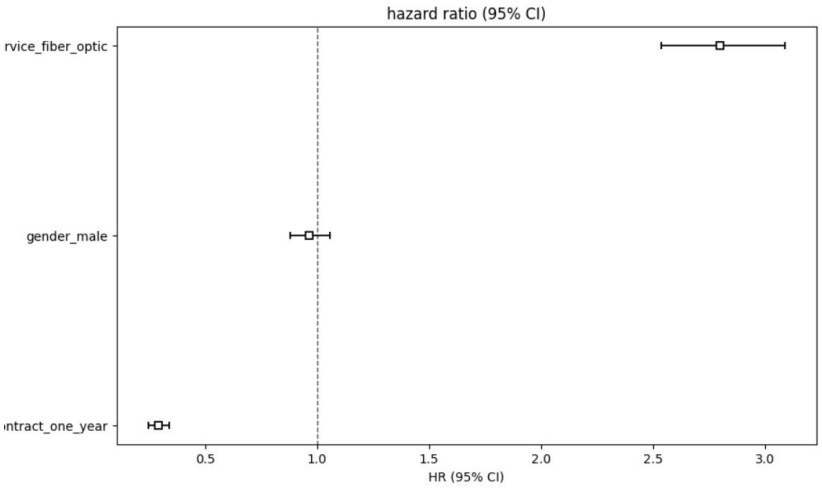
```
cph = CoxPHFitter()
cph.fit(data_cox, duration_col='duration', event_col='churn')
```

模型的输出会告诉我们每个变量对流失风险的影响程度：
- 如果某个变量的“风险比”大于 1，说明它会增加流失的可能；
- 小于 1，则是降低流失风险；
- 同时还能看到每个变量的显著性。

这一步非常关键，因为它能帮我们找出哪些客户特征值得关注，或者哪些行为可能预示着即将流失。

以下是分析示例：

	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)
internetservice_fiber_optic	1.03	2.80	0.05	0.93	1.13	2.54	3.09	0.00	20.50	<0.005	307.85
contract_one_year	-1.23	0.29	0.08	-1.39	-1.07	0.25	0.34	0.00	-15.12	<0.005	169.25
gender_male	-0.04	0.96	0.05	-0.13	0.05	0.88	1.06	0.00	-0.77	0.44	1.19



四、AFT 模型（加速失效时间）

在 Cox 模型之后，我们还用了一种叫 AFT（Accelerated Failure Time）的模型，中文叫加速失效时间模型。

它和 Cox 模型不太一样，它更关注的是客户到底还能“活”多久，也就是预测他们还能保留多少时间。

我们用 Weibull 分布做了拟合：

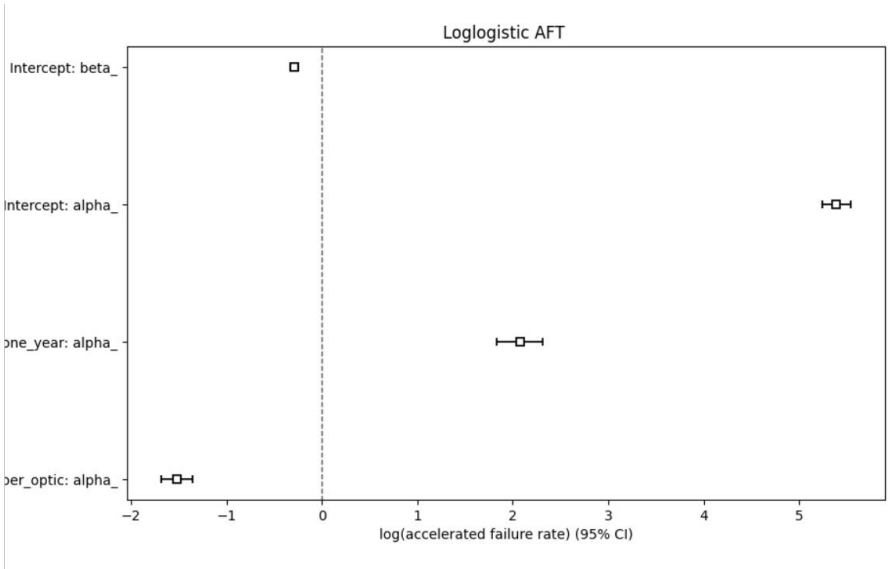
```
aft = WeibullAFTFitter()
aft.fit(data_aft, duration_col='duration', event_col='churn')
```

这个模型对预测客户剩余生命周期很有帮助，适合用在做个性化分析或未来时间预测上。

输出结果示例见下图

模型摘要：

		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)
param	covariate											
alpha_	contract_one_year	2.077421	7.983848	0.122979	1.836387	2.318454	6.273828	10.159959	0.0	16.892523	5.107164e-64	210.250876
	internetservice_fiber_optic	-1.518619	0.219014	0.082554	-1.680421	-1.356817	0.186295	0.257479	0.0	-18.395547	1.426151e-75	248.632480
	Intercept	5.392630	219.780680	0.075216	5.245210	5.540050	189.655686	254.690743	0.0	71.695622	0.000000e+00	inf
beta_	Intercept	-0.291916	0.746831	0.019827	-0.330777	-0.253055	0.718365	0.776425	0.0	-14.722879	4.596637e-49	160.573898



五、客户生命周期价值（CLV）预测

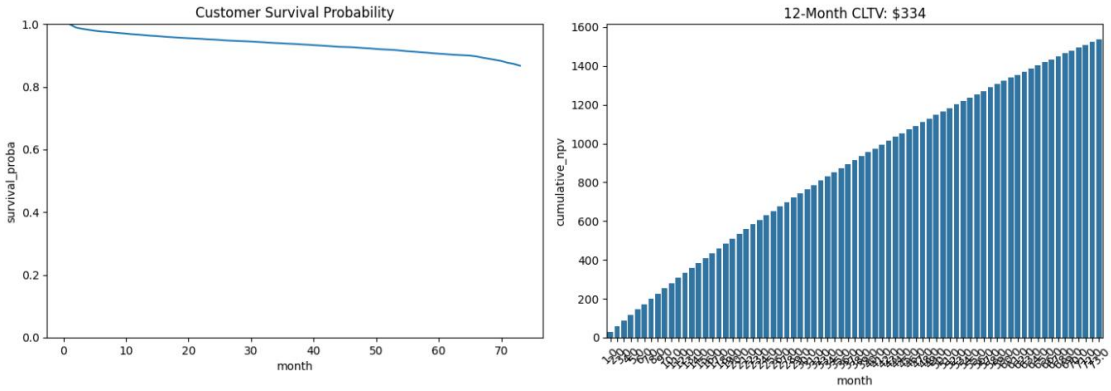
最后一步， CLV（Customer Lifetime Value），主要功能包括：

生存概率预测：评估客户在未来每个月的留存概率。

财务价值计算：结合利润和折现率，计算客户未来现金流的净现值（NPV）。

可视化仪表盘：生成生存曲线和累积价值的直观图表。

运行结果示例如下图：



CLV 是一个很实用的指标，它告诉我们：哪些客户值得长期维护；哪些客户看起来不活跃，但其实潜力很大。

总结

这次的分析，我们围绕客户流失问题，从多个角度进行了解读和预测，形成了一套完整的生存分析流程：

1. 数据准备：构建生命周期和流失信息
2. 生存曲线（KM）：观察整体流失趋势
3. Cox 模型：分析变量对流失风险的影响
4. AFT 模型：预测个体的生命周期
5. CLV 计算：评估每个客户带来的长期价值

这套方法具有复用性，可用于处理其他生存分析案例