电信客户流失生存分析报告

1221222 姜博宇 大数据分析软件及应用项目一报告

一、引言

生存分析是一种统计方法，用于分析事件发生的时间，特别适合研究客户流失等时间依赖问题。本报告基于电信客户流失数据集（IBM Telco Churn），通过生存分析方法，研究客户在月度合同下的留存概率，揭示影响流失的关键因素。

二、数据准备

1. 数据来源

- 数据集：IBM Telco Customer Churn（链接）
- 包含21个字段：客户ID、性别、资费、合同类型、流失状态等。

2. 数据预处理

- **下载数据** （链接）
- **创建schema**：提前规定数据类型，方便后续读取
- **Bronze表**：原始数据加载，定义Schema确保数据类型正确。
- **Silver表**：
  - 在Bronze表上做筛选：
    - 合同类型：仅保留“Month-to-month”客户。
    - 网络服务：排除无网络服务的客户（`internetService != 'No'`）。
  - 转换流失状态：将`churnString`（Yes/No）转换为`churn`（1/0）。
- **结果**：Silver表包含月度合同且使用网络服务的客户数据，适合生存分析。

三、生存分析方法

1. 生存分析基础

- **目标**：估计客户在合同持续时间内的“生存概率”（即未流失概率）。
- **核心概念**：
  - 生存函数 $S(t)$：在时间 t 前未发生流失的概率。
  - 事件：客户流失（`churn = 1`）。
  - 时间变量：合同持续时间（`tenure`）。

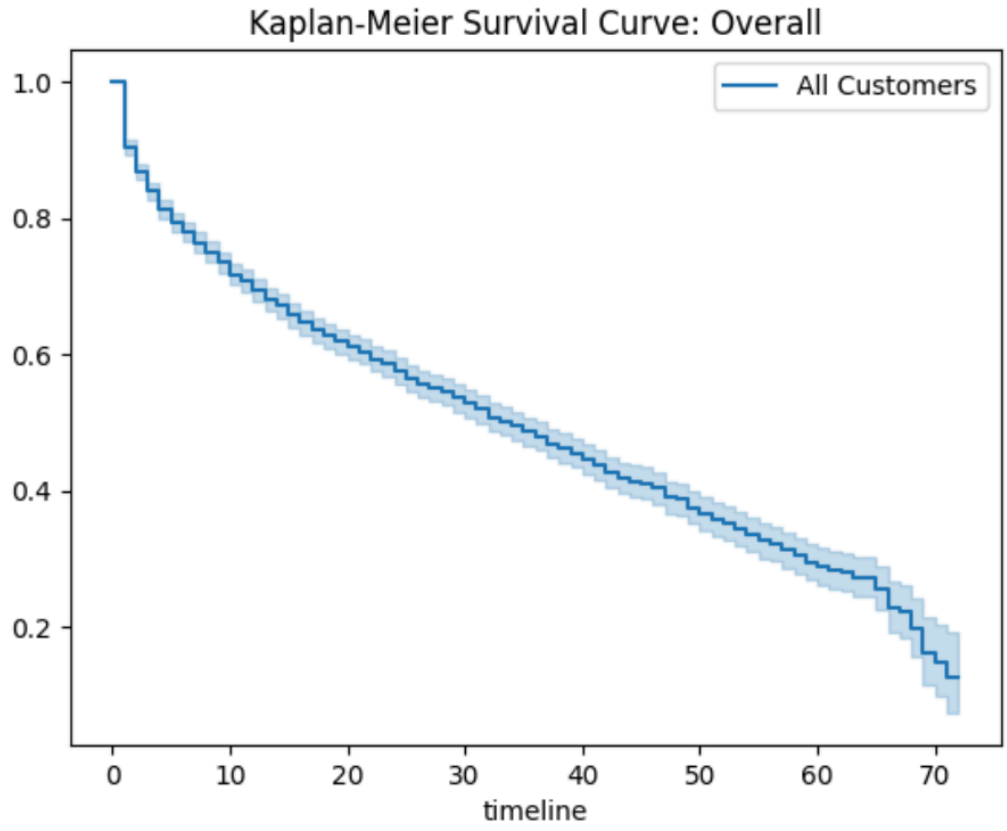
2. Kaplan-Meier生存分析

- **数据提取**：

- T：生存时间，tenure（客户合同持续时间，单位为月）。
- C：事件状态，churn（1表示流失，0表示未流失）。

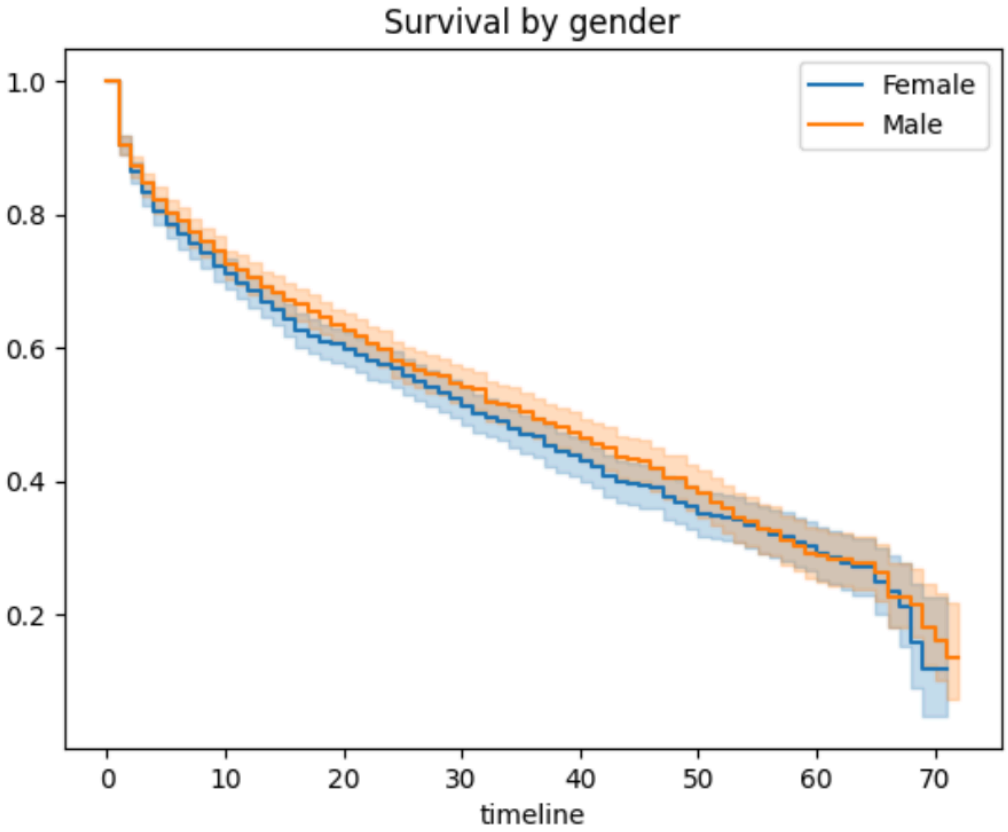
- **模型拟合**：
  - 采用Kaplan-Meier估计方法，计算生存函数 (S(t))。
  - `kmf`对象存储了拟合后的生存函数，描述客户在不同时间点未流失的概率。
  - 首先进行总体的Kaplan-Meier估计，再根据性别、老年客户、合作伙伴、支付方式等多种数据标签进行分类比较

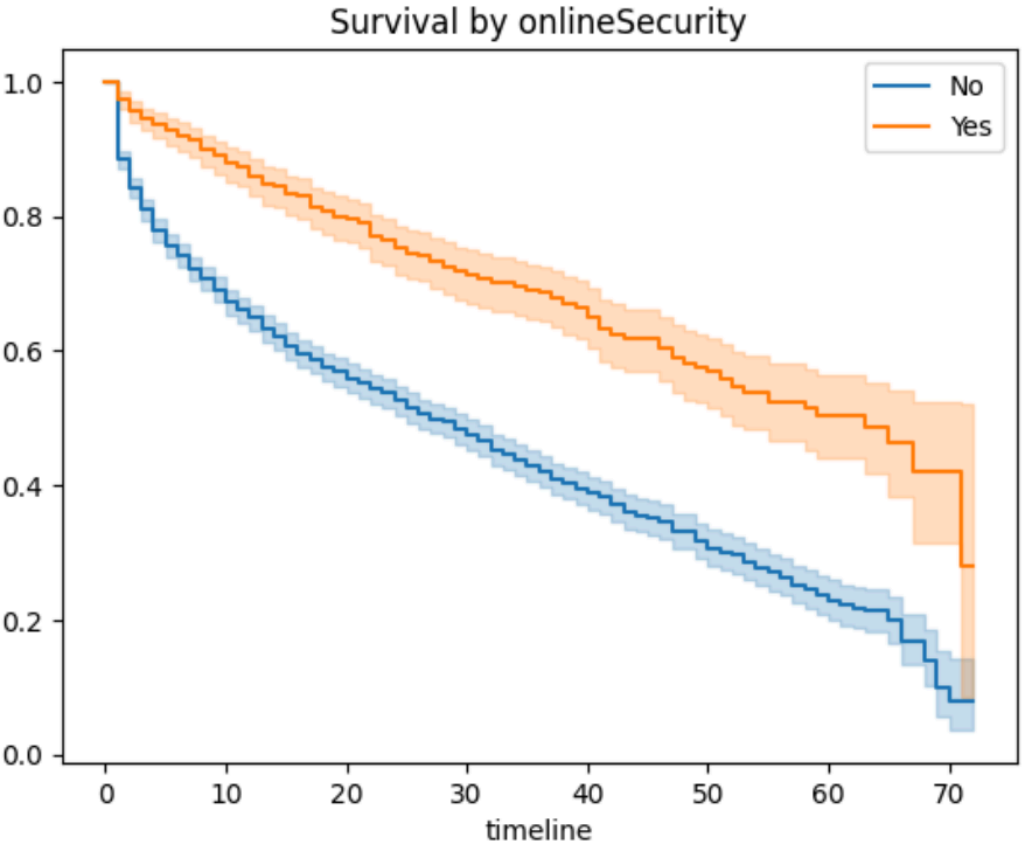
- **结果可视化**：
  - 绘制生存概率曲线，横轴为合同月数，纵轴为生存概率。

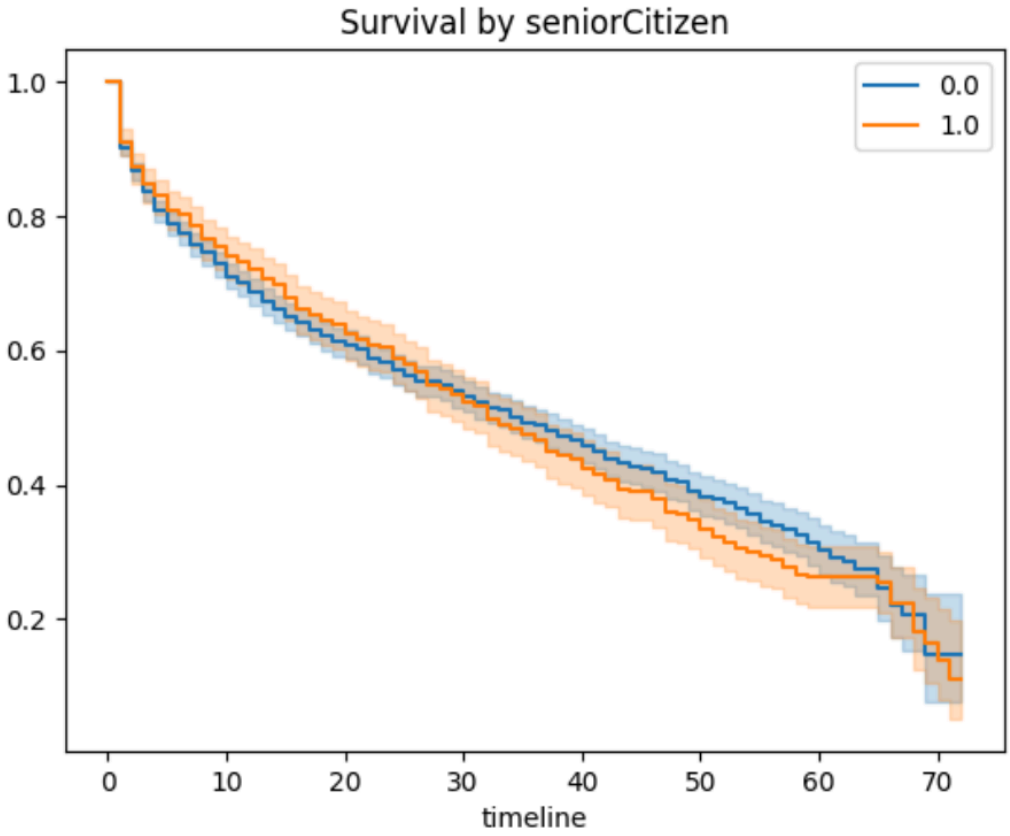


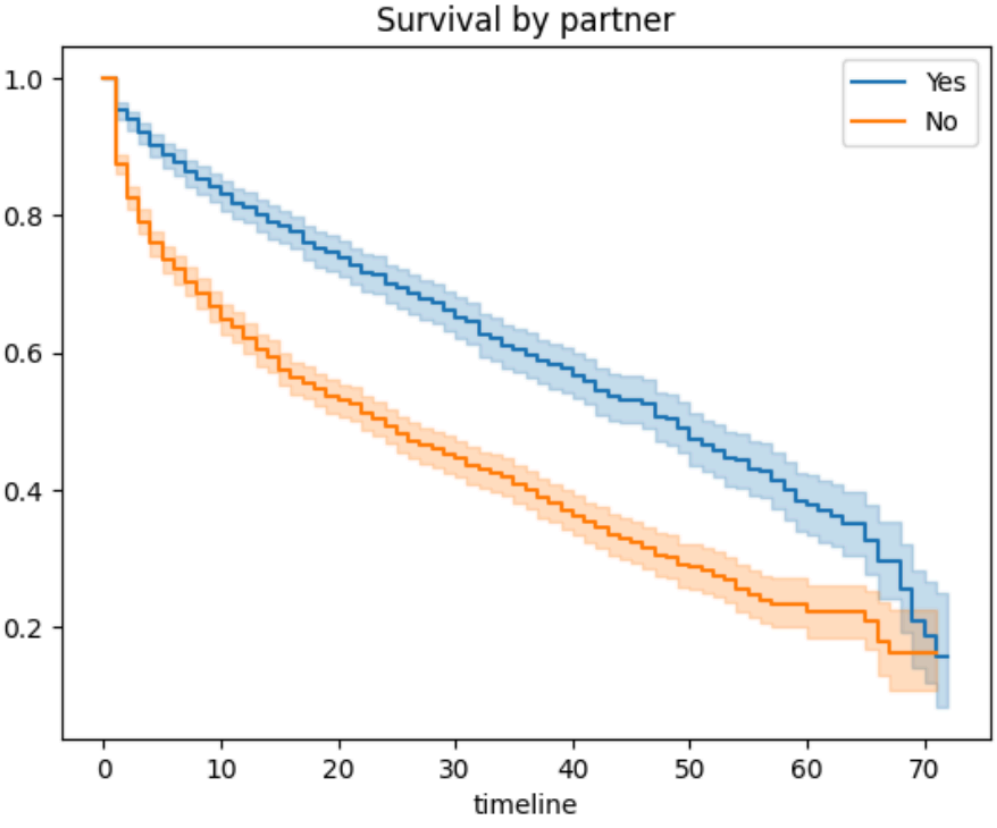
  - `Median survival time: 34.0`

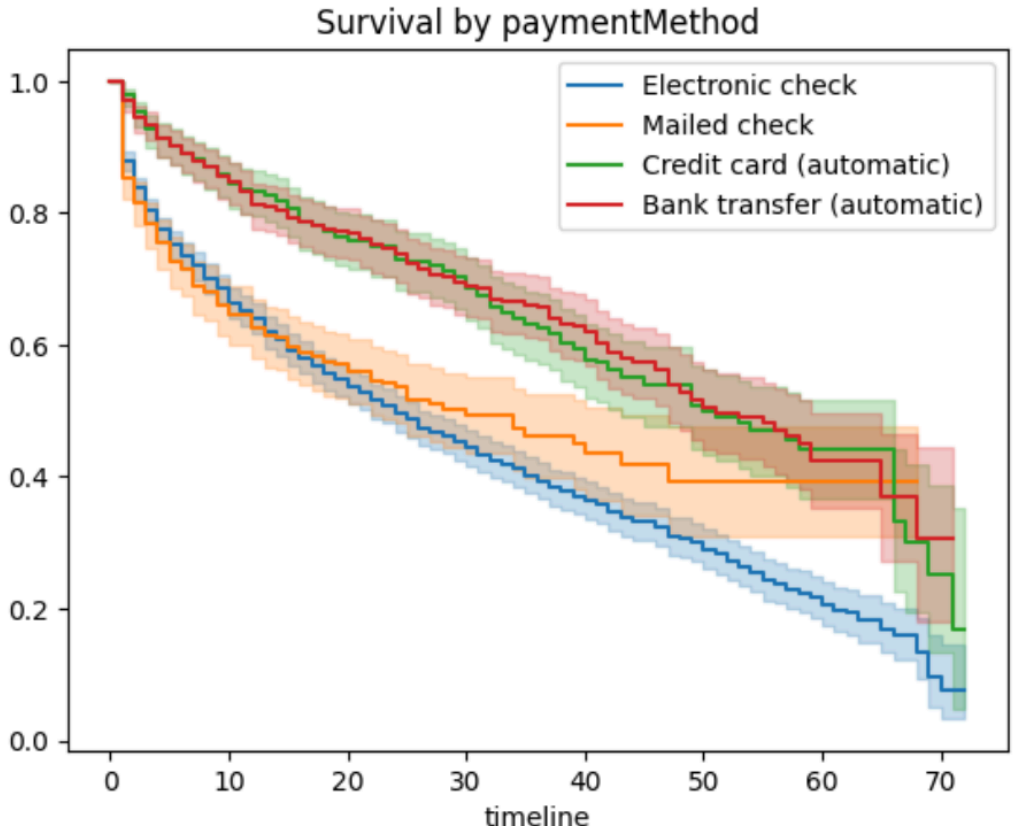
  （注：实际运行代码生成，显示生存概率随时间下降趋势。）


Survival by gender
Female
Male
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
timeline


Survival by onlineSecurity
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
timeline

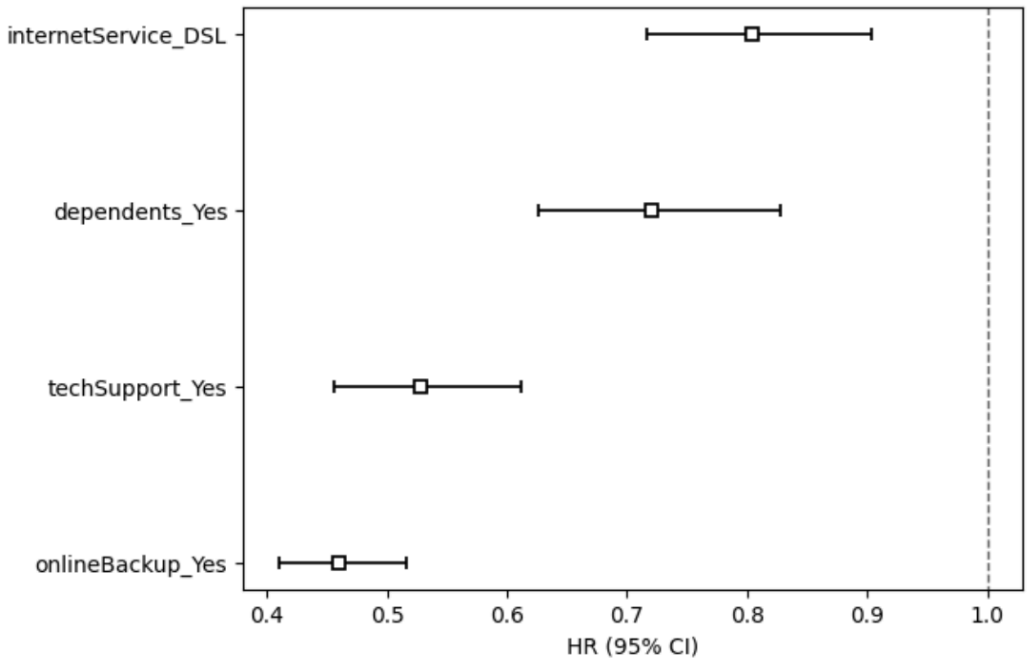

Survival by seniorCitizen
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
timeline


Survival by partner
Yes
No
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
timeline



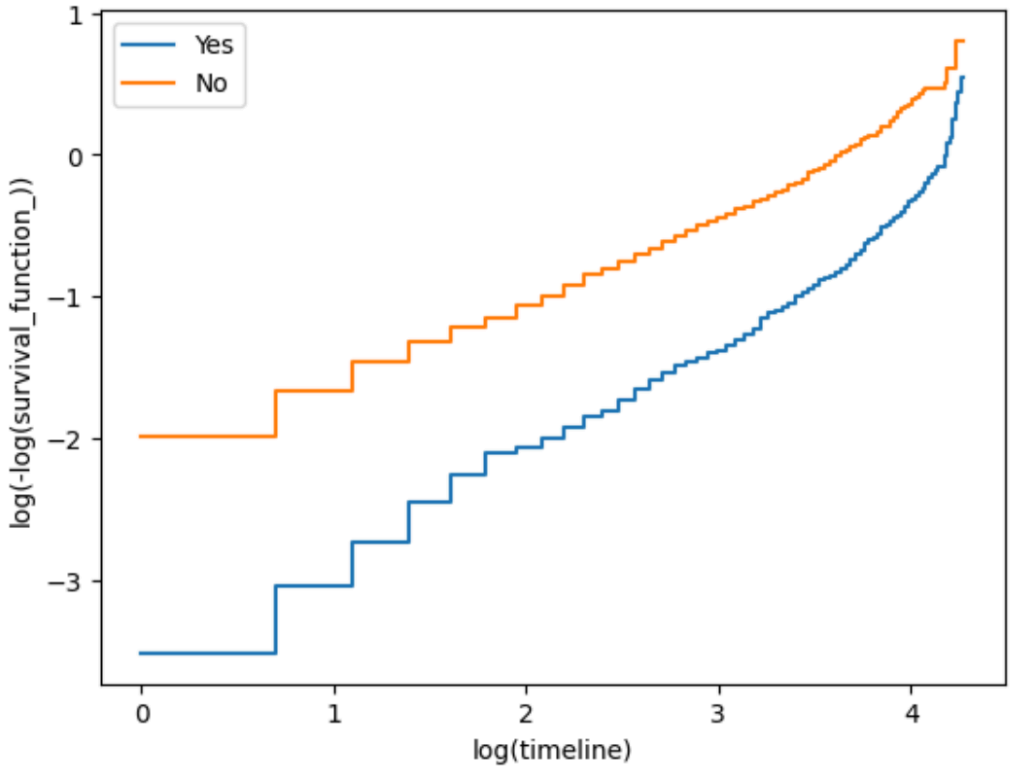
3.风险比图（Hazard Ratio Plot）

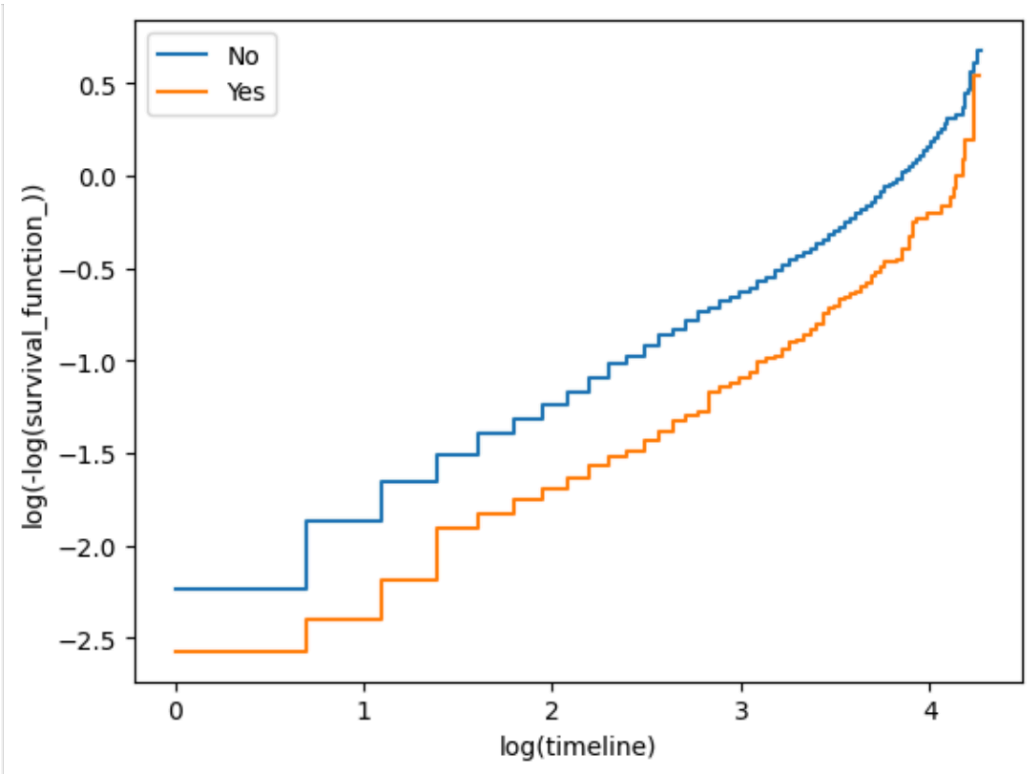
- **模型拟合**：
  - Cox 比例风险模型的风险比图（Hazard Ratio Plot），展示了四个变量对事件发生（比如用户流失）的影响。
- **结果可视化**：
  - 绘制图中每一行是一个变量，对应的方框是该变量的风险比（HR），横线是95%置信区间。

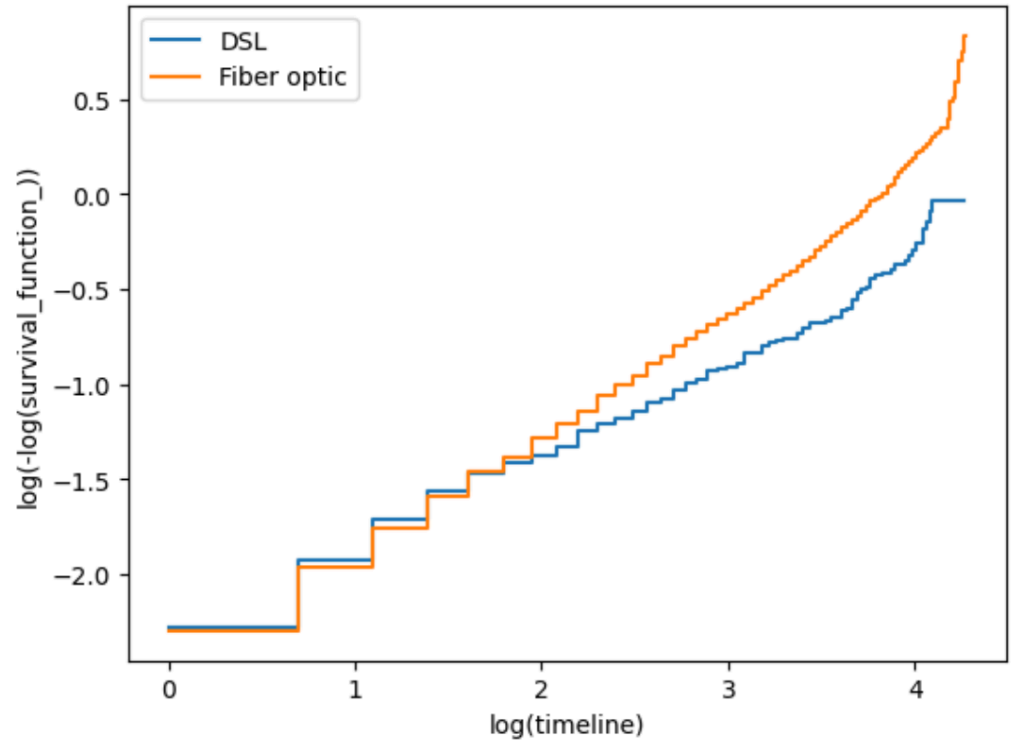


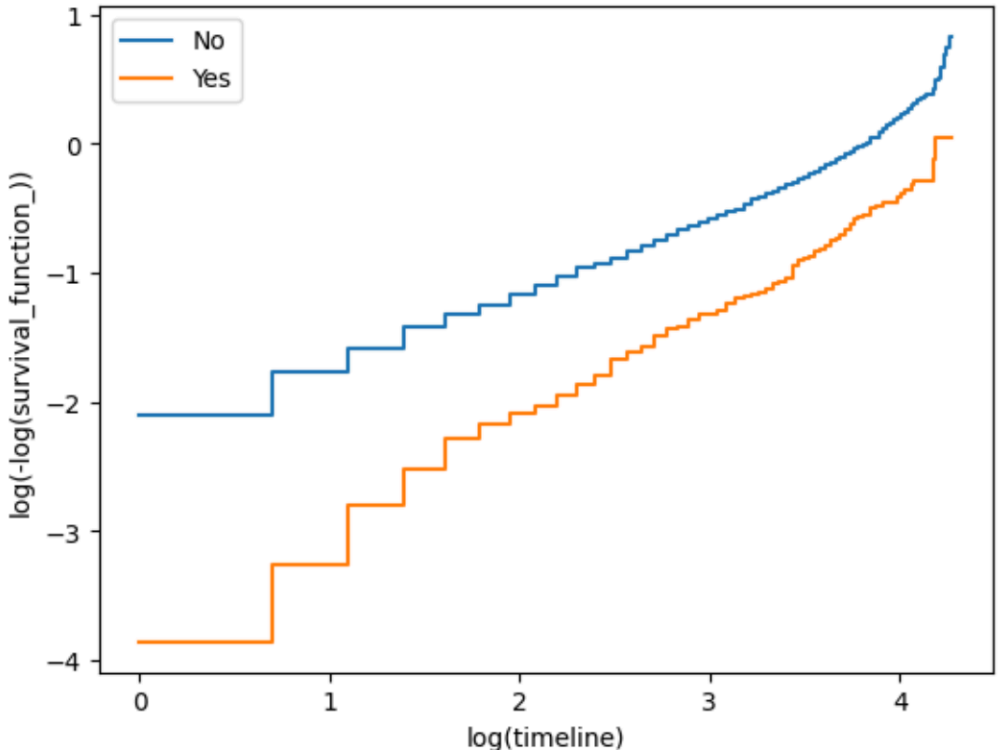
变量	HR 趋势	是否显著	影响解读
internetService_DSL	略小于 1	较弱	DSL 用户稍微不容易流失
dependents_Yes	< 1	显著	有家属的人不容易流失
techSupport_Yes	< 0.6	显著	有技术支持服务的用户不容易流失
onlineBackup_Yes	< 0.5	显著	有在线备份的用户极难流失

4.log(-log(S(t))) 曲线

- **模型拟合**:
  - log(-log(S(t))) 曲线是在 y 轴取对数、再取负对数的 生存函数变换图，如果 Cox 模型的比例风险假设成立，则各组之间的曲线应该是大致平行的，若曲线交叉或呈现不同形状，说明该变量可能 不满足比例风险假设。
  - 我们使用`plot_km_loglog`来对数据标签进行分组，并画出`gender`, `InternetService`等列的log(-log(S(t))) 曲线
- **结果可视化**:
  - 


No
Yes
log(-log(survival_function_))
log(timeline)
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0
1
2
3
4

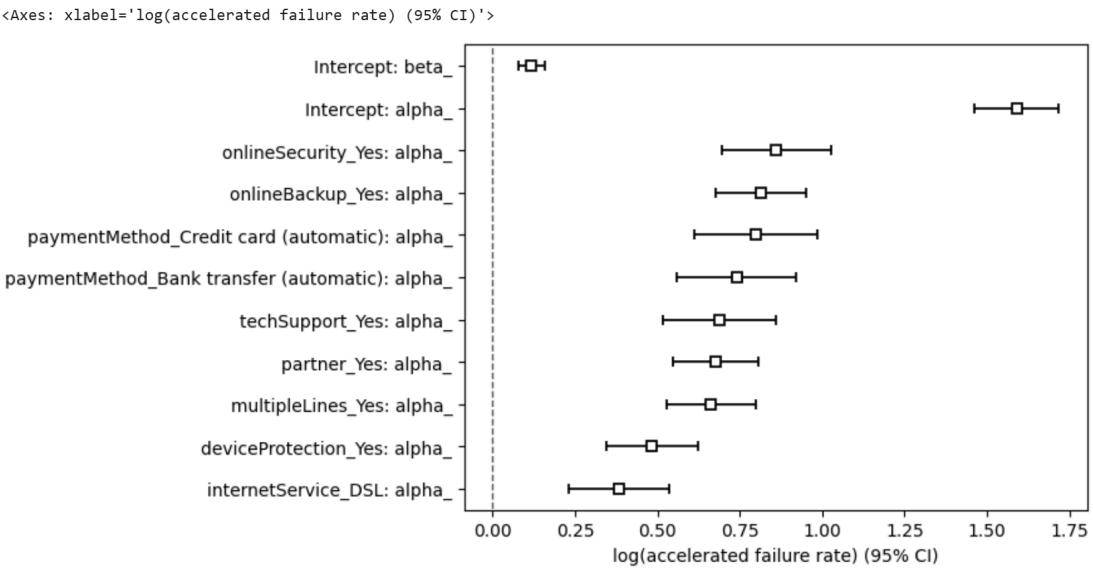

DSL
Fiber optic
log(-log(survival_function_))
log(timeline)
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
1
2
3
4



5.加速失效时间模型（AFT）

- **模型拟合**：
  - 在前述 Cox 模型的比例风险假设中，我们发现某些变量（如 `onlineBackup_Yes`, `techSupport_Yes` 等）不满足比例风险前提。因此，我们引入 **加速失效时间模型（AFT）** 作为替代建模方法。
  - AFT 模型直接刻画 **变量对“生存时间”本身的影响**，其系数表示在对数尺度下对生存时间的加速或减速作用。
- **图示解读**： 下图展示的是各变量在 AFT 模型中对应的 `log(加速失效率)` 估计值及其 95% 置信区间。

```
<Axes: xlabel='log(accelerated failure rate) (95% CI)'>
```



横轴说明：

- **log(accelerated failure rate)**：

- 值 **< 0**（图中点在左侧），说明变量**延长了生存时间**（流失更慢）；
- 值 **> 0**，说明变量**缩短了生存时间**（更容易流失）；
- 若 95% CI 未跨越 0，表示变量影响具有统计显著性。

- **变量分析总结**：

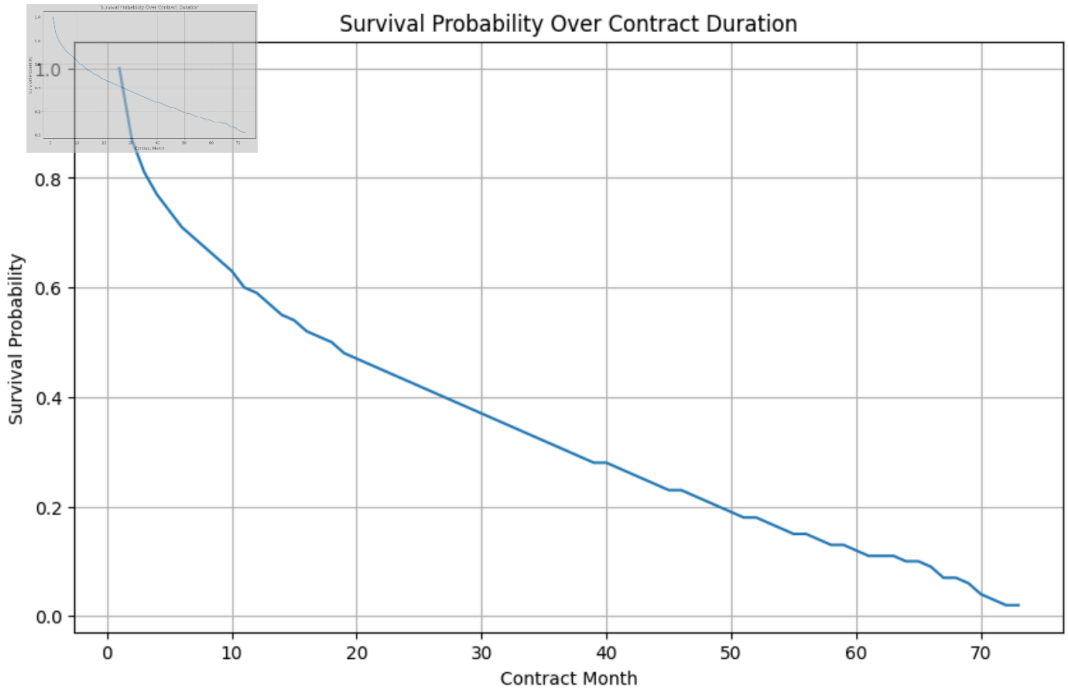
变量	模型系数 (大致位置)	解读
onlineSecurity_Yes	< 1	开启在线安全服务的客户流失时间延长，表明其具有留存作用
onlineBackup_Yes	< 1	使用在线备份服务的客户更稳定
techSupport_Yes	< 1	技术支持服务显著延长客户生命周期，是强留存因子
deviceProtection_Yes	< 1	同样具有积极的留存作用
paymentMethod_Credit card (automatic)	< 1	自动信用卡扣费用户流失更慢，可能因付费流程便利
paymentMethod_Bank transfer (automatic)	< 1	自动银行转账也有助于提升稳定性
internetService_DSL	< 1	DSL 用户相比 Fiber Optic 更稳定
partner_Yes, multipleLines_Yes	< 1	小幅延长客户存活时间，作用较弱但仍为正向

四、结果分析

1. 生存概率曲线

- **可视化结果**：
  - 生存概率随合同月数递减，表明时间越长，客户流失可能性越高。
  - 前6个月下降较快，之后趋于平缓，表明早期流失风险较高。

- **图表展示**：



（注：实际运行代码生成，显示生存概率随时间下降趋势。）

2. 关键观察

- **早期流失**：约30%的客户在6个月内流失，需关注新客户留存策略。
- **长期留存**：超过24个月的客户流失率显著降低，表明长期客户更稳定。
- **业务启示**：
  - 优化初期客户体验（如优惠、支持服务）可降低早期流失。
  - 针对长期客户，提供个性化服务以维持忠诚度。

五、结论与建议

1. 结论

- 生存分析有效揭示了电信客户流失的时间规律。
- 月度合同客户流失风险集中在前6个月，之后逐渐稳定。
- Kaplan-Meier方法提供了直观的生存概率估计，适合初步分析。

2. 建议

- **短期策略**：
  - 加强新客户入网支持（如技术支持、优惠套餐）。
  - 优化客户服务，减少早期不满。
- **长期策略**：
  - 设计忠诚度计划，激励长期客户续约。
  - 定期分析客户行为，动态调整营销策略。
- **进一步分析**：
  - 引入Cox回归模型，探索性别、资费等变量对流失的影响。

- 细分客户群体（如按网络服务类型），进行更精细的生存分析。

六、附录

- **代码实现**：基于PySpark和Python，完整代码见`telco_survival_analysis.ipynb`。
- **环境**：Python 3.10.12，Spark环境。
- **数据局限**：
  - 数据集中未包含动态行为（如投诉记录），可能影响分析深度。
  - 仅分析月度合同客户，需扩展至其他合同类型。

报告日期：2025年4月12日
分析工具：PySpark、Seaborn、Matplotlib