电信客户流失的生存分析报告

12211151 李嘉兴

2025 年 4 月 12 日

摘要

本文对电信客户流失数据（`Telco-Customer-Churn.csv`）进行了生存分析，研究客户流失的时间规律及影响因素。采用 Kaplan-Meier 估计、Log-Rank 检验、Cox 比例风险模型和加速失效时间模型（AFT）等方法，分析了客户未流失概率、不同合同类型客户的流失差异，以及月费、合同类型、网络服务类型和支付方式等特征对流失风险的影响。结果表明，长期合同和网络服务类型显著降低客户流失风险，而月费和支付方式影响较小。最后通过测试样例验证了分析过程的正确性。

1 引言

生存分析是一种统计方法，用于研究事件发生的时间分布及其影响因素。在电信行业中，客户流失是影响企业收益的关键问题。通过生存分析，可以估计客户未流失的概率，并识别导致流失的主要因素。本文基于 `Telco-Customer-Churn.csv` 数据，使用 PySpark 和 `lifelines` 库，完成了以下分析：

- Kaplan-Meier 估计：估计客户未流失的概率随时间的变化，并按合同类型分组分析。
- Log-Rank 检验：比较不同合同类型客户的流失模式差异。
- Cox 比例风险模型：量化特征对流失风险的影响。
- 加速失效时间模型（AFT）：进一步分析特征对流失时间的影响。

2 数据与方法

2.1 数据描述

数据集 `Telco-Customer-Churn.csv` 包含 7043 条客户记录，主要字段包括：

- `tenure`：客户使用服务的时间（月），作为生存时间（`duration`）。
- `Churn`：客户是否流失（"Yes" 或 "No"），转换为事件列（`event`，1 表示流失，0 表示未流失）。
- `MonthlyCharges`：月费，数值型特征。
- `Contract`：合同类型（"Month-to-month", "One year", "Two year"），分类特征。
- `InternetService`：网络服务类型（"DSL", "Fiber optic", "No"），分类特征。
- `PaymentMethod`：支付方式（"Electronic check", "Mailed check" 等），分类特征。

2.2 生存分析流程

生存分析的流程如图 1 所示。

2.3 方法

1. **Kaplan-Meier 估计**：Kaplan-Meier 估计是一种非参数方法，用于估计生存函数 $S(t)$，即在时间 t 之前未发生事件的概率：

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

其中，t_i 是事件发生的时间点，d_i 是时间 t_i 处的事件数量（流失客户数），n_i 是时间 t_i 前的风险集（未流失且未删失的客户数）。

2. **Log-Rank 检验**：Log-Rank 检验用于比较两组或多组的生存分布差异。零假设 H_0 为两组生存分布相同，检验统计量基于各时间点的事件发生率差异。

3. **Cox 比例风险模型**：Cox 模型是一种半参数模型，假设风险函数 $h(t, X)$ 为：

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

其中，$h_0(t)$ 是基准风险函数，β_i 是特征 X_i 的系数，$\exp(\beta_i)$ 表示特征 X_i 对风险的倍数影响。

4. **加速失效时间模型（AFT）**：AFT 模型假设事件发生时间 T 的对数与特征线性相关：

$$\log(T) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

其中，ϵ 服从特定分布（本文使用 Weibull 分布），$\exp(\beta_i)$ 表示特征 X_i 对生存时间的倍数影响。

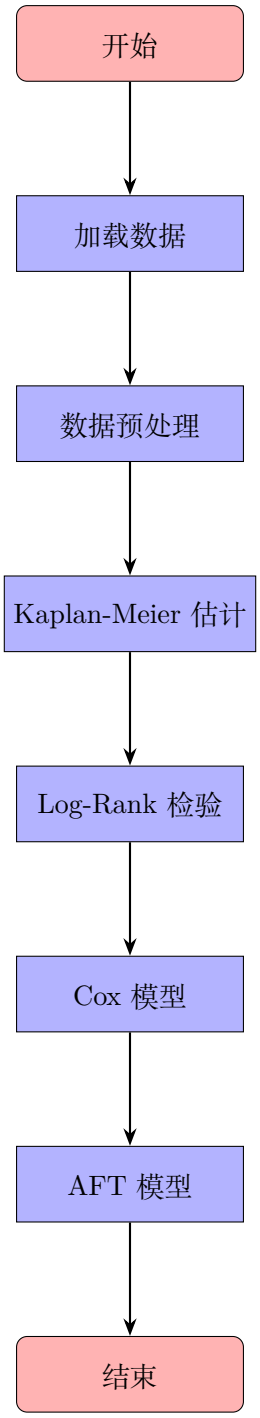

图 1: 生存分析流程图

3

3.1 数据预处理

1. 加载 `Telco-Customer-Churn.csv` 数据，包含 7043 条记录。
2. 处理 `TotalCharges` 列的空字符串，转换为 `null`，并删除缺失值，最终保留 7032 条记录。
3. 转换 `Churn` 为 `event` 列：`Yes` $\rightarrow$ 1，`No` $\rightarrow$ 0。
4. 使用 `tenure` 作为生存时间 `duration`。
5. 编码分类变量:`Contract` $\rightarrow$ `contract_indexed`,`InternetService` $\rightarrow$ `internet_service_indexed`, `PaymentMethod` $\rightarrow$ `payment_method_indexed`。

3.2 Kaplan-Meier 估计

使用 `lifelines.KaplanMeierFitter` 估计生存函数，并按 `Contract` 类型分组。

3.3 Log-Rank 检验

比较 ”Month-to-month” 和 ”Two year” 客户的生存曲线差异。

3.4 Cox 模型

分析 `MonthlyCharges`、`contract_indexed`、`internet_service_indexed` 和 `payment_method_indexed` 对流失风险的影响。

3.5 AFT 模型

使用 Weibull AFT 模型进一步验证特征对流失时间的影响。

4 结果

4.1 Kaplan-Meier 估计

整体生存曲线如图 2 所示。生存函数估计值（部分）如下:

时间（月）	未流失概率
0.0	1.000000
1.0	0.945961
2.0	0.927835
3.0	0.913725
4.0	0.901945
⋮	⋮
72.0	0.592790

表 1: 生存函数估计值（整体）

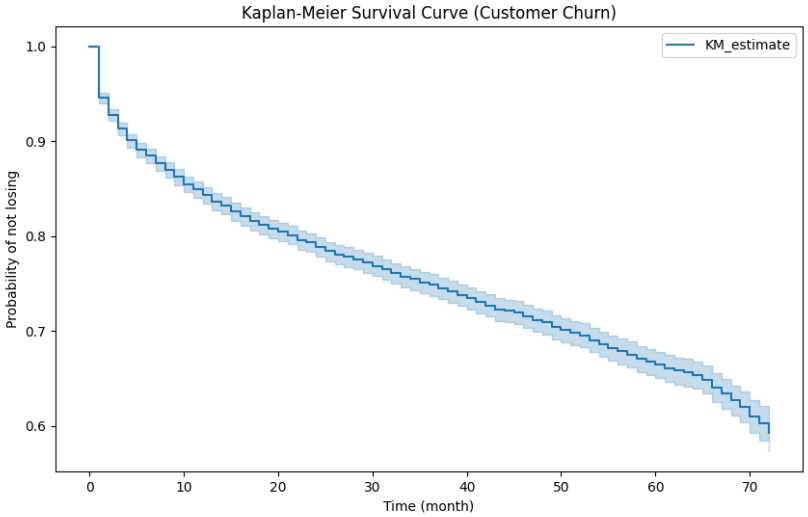

图 2: Kaplan-Meier 生存曲线（整体）

按合同类型分组的生存曲线如图 3 所示：

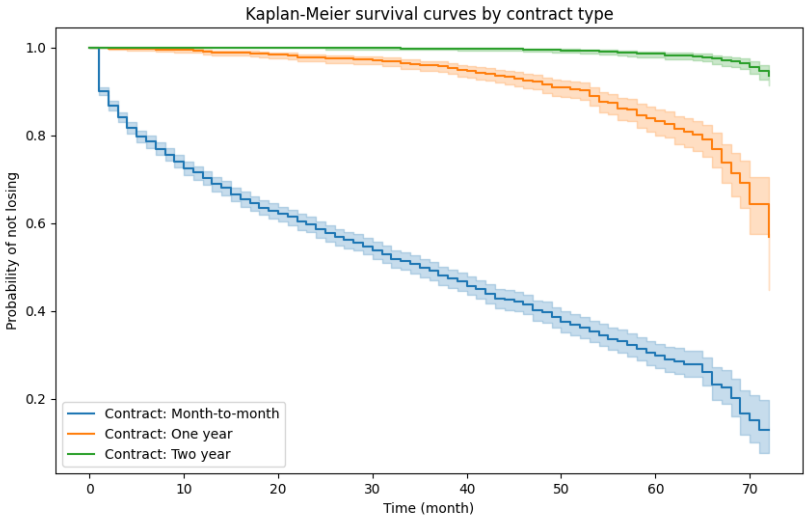

图 3: 按合同类型分组的 Kaplan-Meier 生存曲线

从图 3 可以看出，”Month-to-month” 客户的流失速度最快，生存概率下降最明显；”One year” 客户次之；而 ”Two year” 客户的流失速度最慢，生存概率下降最平缓。

4.2 Log-Rank 检验

比较 ”Month-to-month” 和 ”Two year” 客户的生存曲线，Log-Rank 检验结果为：

$$p = 0.0 < 0.05$$

表明两组生存分布有显著差异，”Two year” 客户的流失风险显著低于 ”Month-to-month” 客户。

4.3 Cox 比例风险模型

Cox 模型结果如表 2 所示：

特征	系数	exp(coef)	标准误	p 值	置信区间
MonthlyCharges	-0.0386	0.9622	0.0020	< 0.005	[-0.0426, -0.0346]
contract_indexed	-1.2203	0.2951	0.0565	< 0.005	[-1.3311, -1.1096]
internet_service_indexed	-1.8225	0.1616	0.0862	< 0.005	[-1.9914, -1.6536]
payment_method_indexed	-0.3277	0.7206	0.0225	< 0.005	[-0.3717, -0.2836]

表 2: Cox 模型系数

- MonthlyCharges：系数为 -0.0386，exp(coef) = 0.9622，表明月费对流失风险的影响较小，每增加 1 单位月费，流失风险降低约 3.78%。
- contract_indexed：系数为 -1.2203，exp(coef) = 0.2951，表明合同类型对流失风险影响显著，长期合同客户的流失风险仅为短期合同客户的 29.51%。
- internet_service_indexed：系数为 -1.8225，exp(coef) = 0.1616，表明网络服务类型对流失风险影响显著，”Fiber optic” 和 ”DSL” 客户相比 ”No” 客户的流失风险更低。
- payment_method_indexed：系数为 -0.3277，exp(coef) = 0.7206，表明支付方式对流失风险有一定影响，”Electronic check” 客户的流失风险更高。

Cox 模型的系数图如图 4 所示：

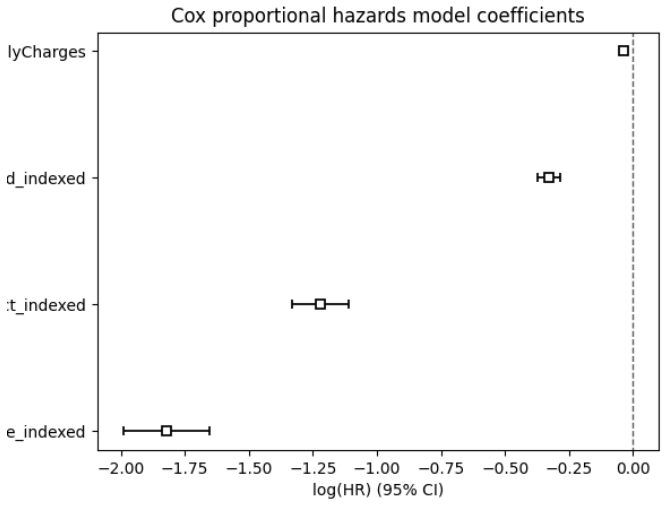

图 4: Cox 比例风险模型系数图

4.4 AFT 模型

AFT 模型（Weibull 分布）结果如表 3 所示：

特征	系数	exp(coef)	标准误	p 值	置信区间
MonthlyCharges	0.05	1.05	0.00	< 0.005	[0.04, 0.05]
contract_indexed	1.33	3.80	0.06	< 0.005	[1.21, 1.46]
internet_service_indexed	2.14	8.48	0.11	< 0.005	[1.93, 2.35]
payment_method_indexed	0.38	1.47	0.03	< 0.005	[0.33, 0.44]

表 3: AFT 模型系数（lambda 参数）

- contract_indexed：系数为 1.33，exp(coef) = 3.80，表明长期合同显著延长客户生存时间，生存时间约为短期合同客户的 3.8 倍。
- internet_service_indexed：系数为 2.14，exp(coef) = 8.48，表明网络服务类型对生存时间影响显著，”Fiber optic” 和 ”DSL” 客户的生存时间更长。
- MonthlyCharges 和 payment_method_indexed 的系数较小，影响有限。

AFT 模型的 Concordance 为 0.82，表明模型预测能力较好。

4.5 测试样例验证

1. **生存函数在 $t = 0$ 时的值**：预期为 1.0，实际值为 1.0，验证通过，表明 Kaplan-Meier 估计在初始时刻符合预期。
2. **Log-Rank 检验的 p 值**：预期 $p < 0.05$，实际 $p = 0.0$，验证通过，确认不同合同类型的流失模式存在显著差异。
3. **Cox 模型中 contract_indexed 的系数**：预期为负值，实际为 -1.2203，验证通过，表明长期合同降低流失风险的假设成立。

5 讨论与结论

通过生存分析，得出以下结论：

1. 客户流失主要发生在早期（前 20 个月），72 个月后约 59.3% 的客户未流失，表明客户忠诚度随时间增加而提升。
2. 合同类型对流失有显著影响，”Two year” 合同客户的流失风险仅为 ”Month-to-month” 客户的 29.51%，说明长期合同是降低流失的有效策略。
3. 网络服务类型对流失风险影响较大，”Fiber optic” 和 ”DSL” 客户的流失风险更低，可能是因为更好的服务质量提高了客户满意度。
4. 月费和支付方式对流失风险影响较小，月费每增加 1 单位仅降低 3.78% 的流失风险，而支付方式的影响也较为有限。