

# 生存分析在电信客户流失预测中的应用报告

## 1. 概述

生存分析（Survival Analysis）是一类用于预测“事件发生时间”的统计方法，最初应用于医疗领域（如患者生存时间预测），现已被广泛应用于各行业。其核心目标是通过历史数据建模，估计个体在特定时间点发生目标事件的概率（如客户流失、设备故障等）。

在电信行业中，生存分析可帮助企业优化客户管理、硬件维护及产品升级策略，从而降低运营成本并提升客户价值。

## 2. 电信行业应用案例

### 2.1 客户保留

- 问题：客户流失成本远高于新客户获取成本。
- 分析目标：预测客户在何时可能流失（事件定义为“取消服务”）。
- 应用价值：提前识别高风险客户并制定干预策略（如定向优惠）。

### 2.2 硬件故障预测

- 问题：网络设备故障直接影响客户体验。
- 分析目标：预测硬件设备失效时间（事件定义为“设备故障”）。
- 应用价值：制定预防性维护计划，降低服务中断风险。

### 2.3 套餐升级预测

- 问题：客户生命周期中存在关键决策节点（如更换套餐）。
- 分析目标：预测客户升级设备或套餐的时间点。
- 应用价值：在适当时机推荐个性化产品，提升客户价值。

## 3. 分析方法与技术实现

### 3.1 数据准备

- 数据集：IBM提供的电信客户流失数据集，包含客户人口统计、服务计划、使用行为及流失状态（Tenure表示客户在网时长，Churn表示是否流失）。

- 数据处理：
  - 将Churn列转换为布尔类型。
  - 筛选“按月签约”的互联网用户作为分析对象。
  - 使用Delta Lake存储数据（Bronze表存储原始数据，Silver表存储清洗后数据）。

## 3.2 Kaplan-Meier模型

- 原理：非参数方法，估计整体或分组生存概率曲线。
- 实现步骤：
  1. 拟合模型：输入Tenure（在网时长）和Churn（流失状态）。
  2. 绘制生存曲线：横轴为时间，纵轴为生存概率。
  3. 分组分析：通过Log-Rank检验比较组间差异（如“是否开通在线安全服务”显著影响生存概率，而性别分组无显著差异）。
- 输出：提取生存概率用于后续客户生命周期价值（CLV）计算。

## 3.3 Cox比例风险模型

- 原理：半参数模型，估计协变量对风险比（Hazard Ratio）的影响。
- 关键假设：风险比随时间保持恒定（比例风险假设）。
- 实现步骤：
  1. 对分类变量（如网络服务类型、技术支持）进行独热编码，避免多重共线性。
  2. 拟合模型，输出风险比及置信区间。
  3. 验证假设：
    - 统计检验：部分变量（如onlineBackup\_Yes）的p值<0.05，违反比例风险假设。
    - Schoenfeld残差图：残差与时间存在相关性（如techSupport\_Yes随时间呈现趋势）。
    - Log-log图：曲线交叉表明假设不成立。
- 优化方案：分层建模、引入时间依赖变量或改用参数模型（如加速失效时间模型）。

## 3.4 加速失效时间模型（AFT）

- 原理：参数模型，假设失效时间服从特定分布（如对数逻辑分布）。
  - 实现步骤：
    1. 拟合Log-Logistic AFT模型，输出加速因子（如光纤用户流失时间比DSL用户快1.47倍）。
    2. 验证假设：
      - Log-log图：曲线是否平行（验证比例优势假设）？
      - 分布拟合：曲线是否接近直线（验证分布假设）？
  - 结果：模型部分符合假设，但并行性不足，需进一步优化。
-

## 4. 结果应用与可视化

### 4.1 客户生命周期价值 (CLV) 看板

- 关键指标：
  - 生存概率（从模型预测中提取）。
  - 月度利润（假设为固定值，如30美元）。
  - 净现值 (NPV)：考虑资金时间价值，公式为：
$$NPV = (\text{月度利润} \times \text{生存概率}) / (1 + \text{内部收益率})^{\text{合约月份}}$$
- 可视化：
  - 累积NPV曲线：展示客户生命周期内总价值，辅助制定获客预算。
  - 生存概率曲线：动态展示不同套餐或用户分组的生存趋势。

## 5. 结论与建议

- 方法对比：
  - Kaplan-Meier：适用于单变量探索性分析，直观但无法处理多变量。
  - Cox模型：灵活性高，但需严格验证比例风险假设。
  - AFT模型：参数假设明确，适合分布已知的场景。
- 业务建议：
  - 优先使用Cox模型进行多变量预测，并结合分层或时间依赖变量优化假设违反问题。
  - 对高风险客户（如无在线安全服务用户）实施定向保留策略。
  - 基于CLV看板动态调整营销预算，最大化投资回报率。
- 未来方向：
  - 探索更复杂的模型扩展（如使用立方样条增强Cox模型灵活性）。
  - 结合实时数据更新生存预测，提升策略时效性。

报告总结：生存分析为电信企业提供了从数据到决策的关键工具，通过精准预测客户行为与设备状态，企业可显著优化资源分配并提升客户价值。