生存分析报告

12210356 袁可帆

本报告使用 IBM 电信客户流失数据集，通过生存分析方法研究客户流失行为。生存分析是研究事件发生时间分布的统计方法，在本案例中用于预测客户流失时间并计算客户生命周期价值 (CLV)。

1. 数据准备

数据来源于 IBM 提供的电信客户流失数据集，包含客户的人口统计信息、服务计划、媒体使用情况和订阅状态。关键变量包括客户在公司停留的月数（tenure）和客户是否仍在订阅（churn）。数据预处理部分，数据被存储为 Delta Lake 的青铜表和银表。银表对原始数据进行了清洗和筛选，例如过滤出月度合同的互联网用户。

```
[language=Python, basicstyle=\small\ttfamily]
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, when
from pyspark.sql.types import StructType, StructField, StringType, DoubleType

# 创建 SparkSession
spark = SparkSession.builder.appName("TelcoChurnAnalysis").getOrCreate()

# 定义数据结构
schema = StructType([
    StructField('customerID', StringType()),
    StructField('gender', StringType()),
    StructField('seniorCitizen', DoubleType()),
    StructField('partner', StringType()),
    StructField('dependents', StringType()),
    StructField('tenure', DoubleType()),
    StructField('phoneService', StringType()),
    StructField('multipleLines', StringType()),
```

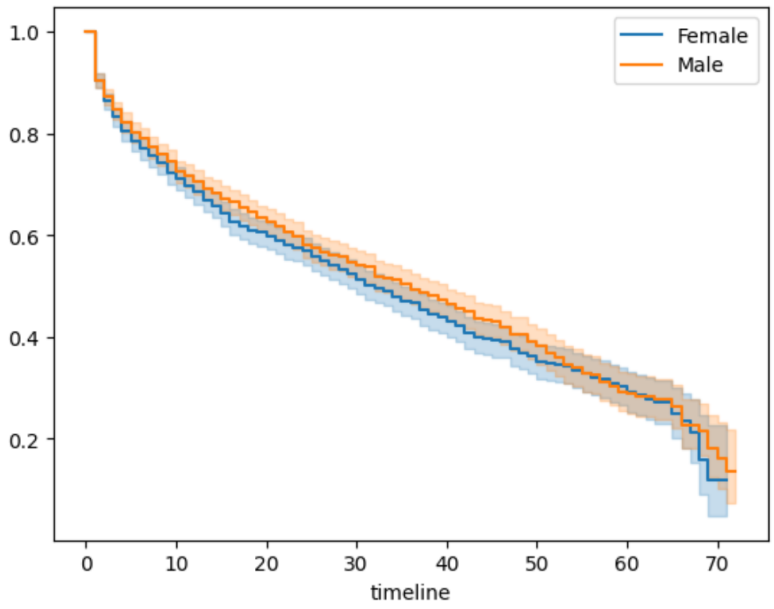
```
    StructField('internetService', StringType()),
    StructField('onlineSecurity', StringType()),
    StructField('onlineBackup', StringType()),
    StructField('deviceProtection', StringType()),
    StructField('techSupport', StringType()),
    StructField('streamingTV', StringType()),
    StructField('streamingMovies', StringType()),
    StructField('contract', StringType()),
    StructField('paperlessBilling', StringType()),
    StructField('paymentMethod', StringType()),
    StructField('monthlyCharges', DoubleType()),
    StructField('totalCharges', DoubleType()),
    StructField('churn', StringType())
])

# 读取数据
bronze_df = spark.read.format('csv').schema(schema).option('header', 'true').
    load('Telco-Customer-Churn.csv')

# 数据预处理
silver_df = bronze_df.withColumn('churn', when(col('churn') == 'Yes', 1).when(
    col('churn') == 'No', 0).otherwise('Unknown'))
silver_df = silver_df.drop('churnString')
silver_df = silver_df.filter(col('contract') == 'Month-to-month')
silver_df = silver_df.filter(col('internetService') != 'No')
```

2. Kaplan-Meier 方法

Kaplan-Meier 方法用于构建生存概率曲线，评估客户流失的时间分布。以下是基于不同变量的生存曲线分析结果：

2.1 性别（Gender）

以性别为例进行分析。生存曲线显示，男性和女性客户的流失概率随时间变化的趋势相似。Log-rank 检验的 p 值为 0.204476，表明性别对客户流失没有显著影响。



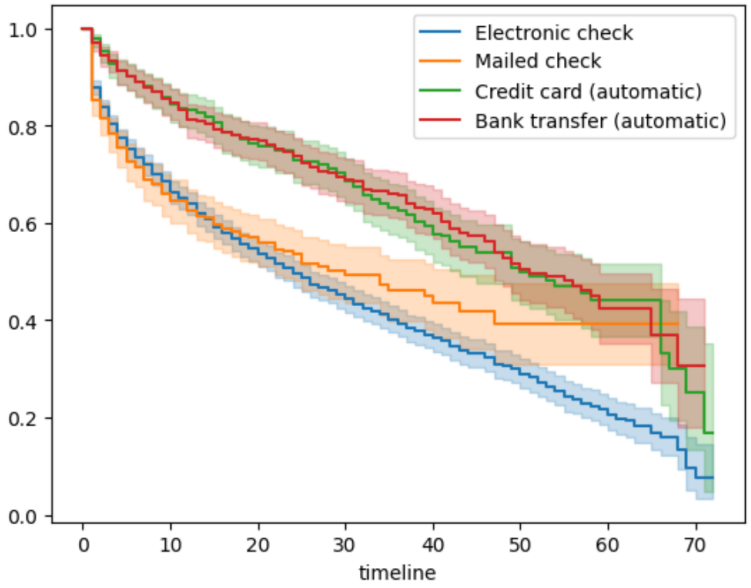
2.2 其他变量分析

其他变量如高级公民状态、伴侣状态、是否有受抚养人等对客户流失的影响也进行了分析，图表不再列出。具体结果如下表所示：

其中支付方式有多个变量，图表和表格分析如下，可知部分支付方式对客户流失有显著影响：

变量	Log-rank p 值	是否显著影响
高级公民	2.523624e-12	是
伴侣	5.063437e-58	是
受抚养人	0.000251	是
电话服务	0.377599	否
多条线路	3.190116e-11	是
互联网服务	2.369872e-20	是
流媒体电视	1.813974e-32	是
流媒体电影	6.484901e-39	是
在线安全	3.138886e-18	是
在线备份	2.620909e-67	是
设备保护	7.904692e-39	是
技术支持	3.468692e-07	是
无纸化账单	5.000937e-07	是
支付方式	-	是

表 1: Kaplan-Meier 方法分析结果



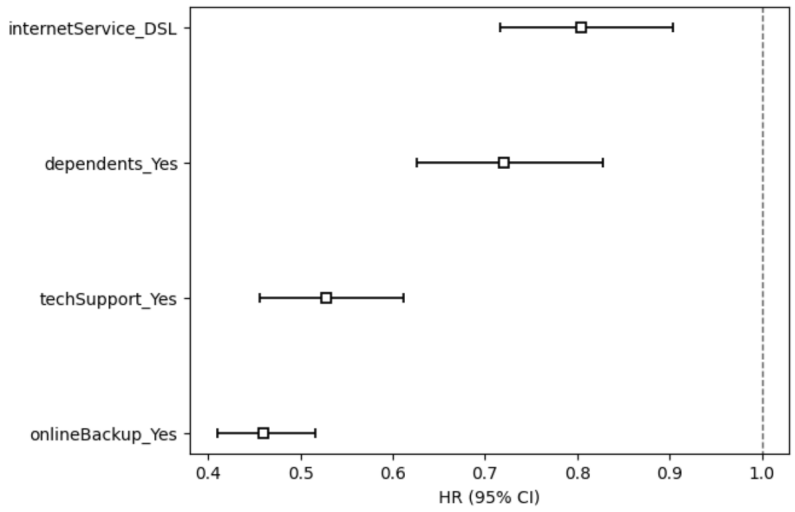
		test_statistic	p	-log2(p)
Bank transfer (automatic)	**Credit card (automatic)**	0.153545	6.951703e-01	0.524562
	Electronic check	55.164654	1.108442e-13	43.036532
	Mailed check	190.000457	3.178566e-43	141.174532
Credit card (automatic)	**Electronic check**	45.167592	1.808736e-11	35.686227
	Mailed check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

3. Cox 比例风险模型

Cox 比例风险模型用于进行多变量分析，评估不同变量对客户流失风险的影响。模型的一致性指数为 0.64，表明模型具有中等的预测能力。显著变量包括 dependents_Yes、internetService_DSL、onlineBackup_Yes 和 techSupport_Yes。然而，internetService_DSL、onlineBackup_Yes 和 techSupport_Yes 违反了比例风险假设。

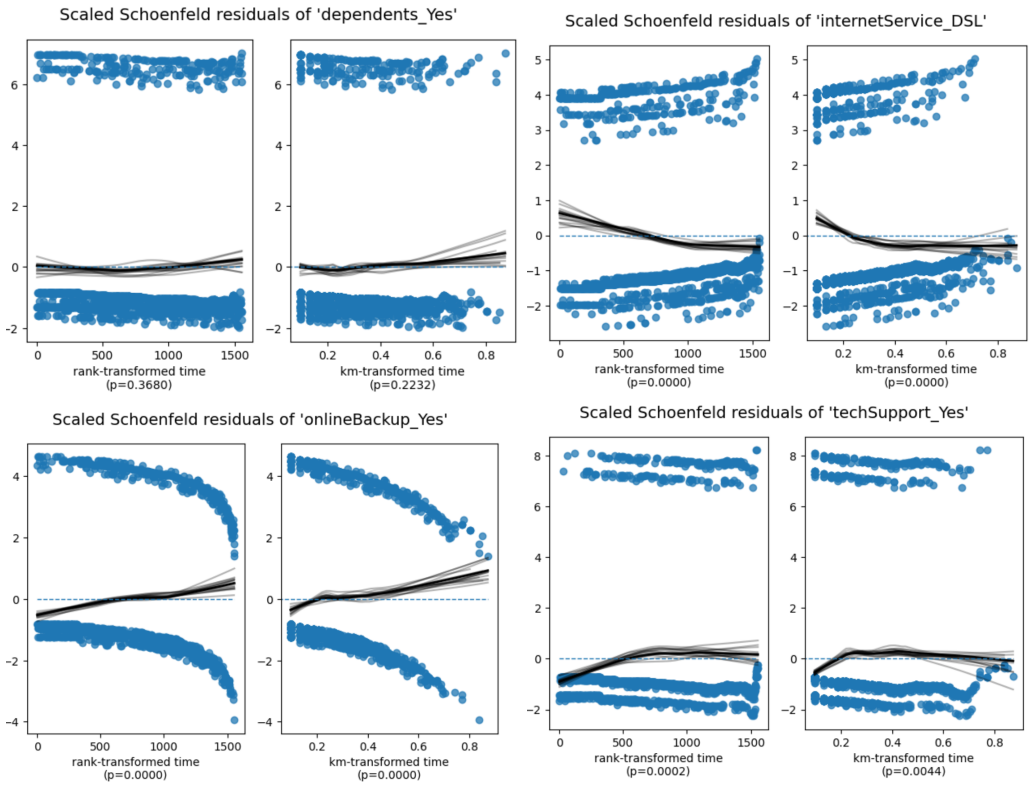
```
[language=Python, basicstyle=\small\ttfamily]
from lifelines import CoxPHFitter

# 独热编码
encode_cols = ['dependents', 'internetService', 'onlineBackup', 'techSupport',
    'paperlessBilling']
encoded_pd = pd.get_dummies(telco_pd, columns=encode_cols, prefix=encode_cols,
    drop_first=False)

# 拟合 Cox 比例风险模型
cph = CoxPHFitter(alpha=0.05)
cph.fit(encoded_pd, duration_col='tenure', event_col='churn')
cph.print_summary()
```

3.1 风险比（Hazard Ratios）图



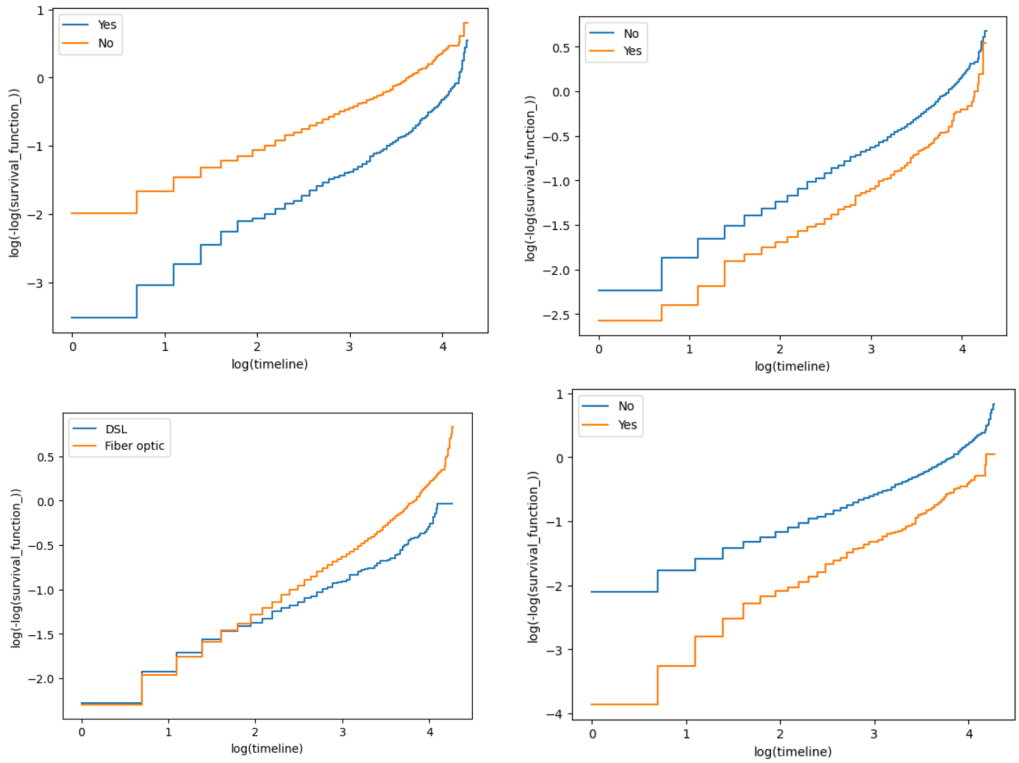
从图中可以看出，‘internetService_DSL‘、‘dependents_Yes‘、‘techSupport_Yes‘ 和 ‘onlineBackup_Yes‘ 的风险比均小于 1，表明这些变量与客户流失的较低风险相关。

3.2 Schoenfeld 残差图



从 Schoenfeld 残差图中可以看出，‘onlineBackup_Yes‘ 和 ‘techSupport_Yes‘ 的残差图显示出明显的趋势，表明这些变量可能违反了比例风险假设。

3.3 对数对数生存函数图



对数对数生存函数图进一步证实了 ‘onlineBackup_Yes‘ 和 ‘techSupport_Yes‘ 的曲线显示出明显的非平行趋势，表明这些变量可能违反了比例风险假设。

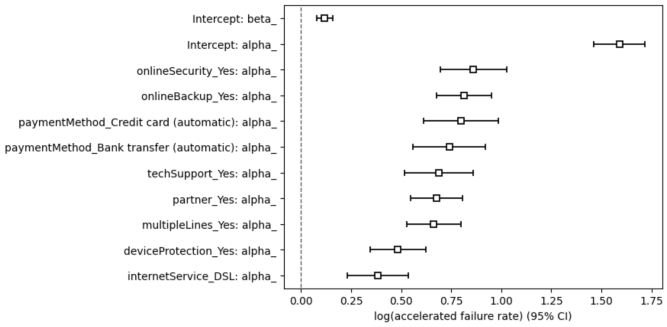
4. 加速失效时间模型（AFT）

加速失效时间模型通过参数化方法建模客户流失时间。模型的协变量一致性指数为 0.73，表明模型具有较好的预测能力。显著影响客户流失的变量包括设备保护、互联网服务、在线安全、在线备份、技术支持和支付方式。

```
[language=Python, basicstyle=\small\ttfamily]
from lifelines import LogLogisticAFTFitter

# 独热编码
encode_cols = ['partner', 'multipleLines', 'internetService', 'onlineSecurity',
    'onlineBackup', 'deviceProtection', 'techSupport', 'paymentMethod']
encoded_pd = pd.get_dummies(telco_pd, columns=encode_cols, prefix=encode_cols,
    drop_first=False)
```

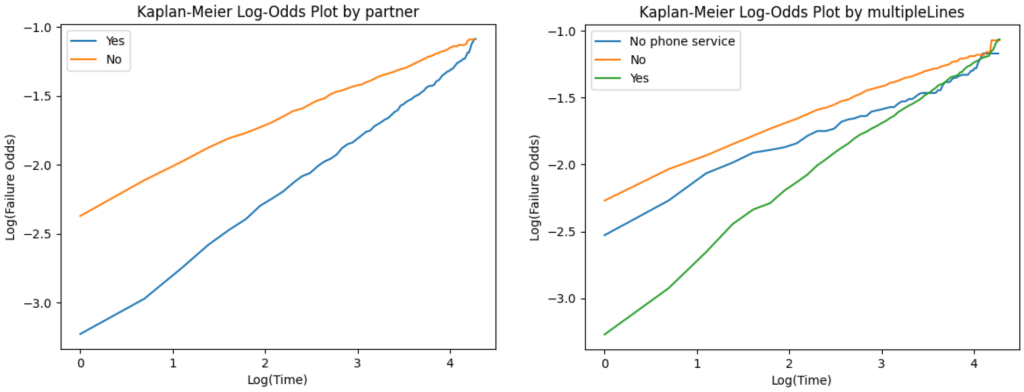
```
7
8  # 拟合 Log-Logistic AFT 模型
9  aft = LogLogisticAFTFitter()
10 aft.fit(encoded_pd, duration_col='tenure', event_col='churn')
11 aft.print_summary()
```

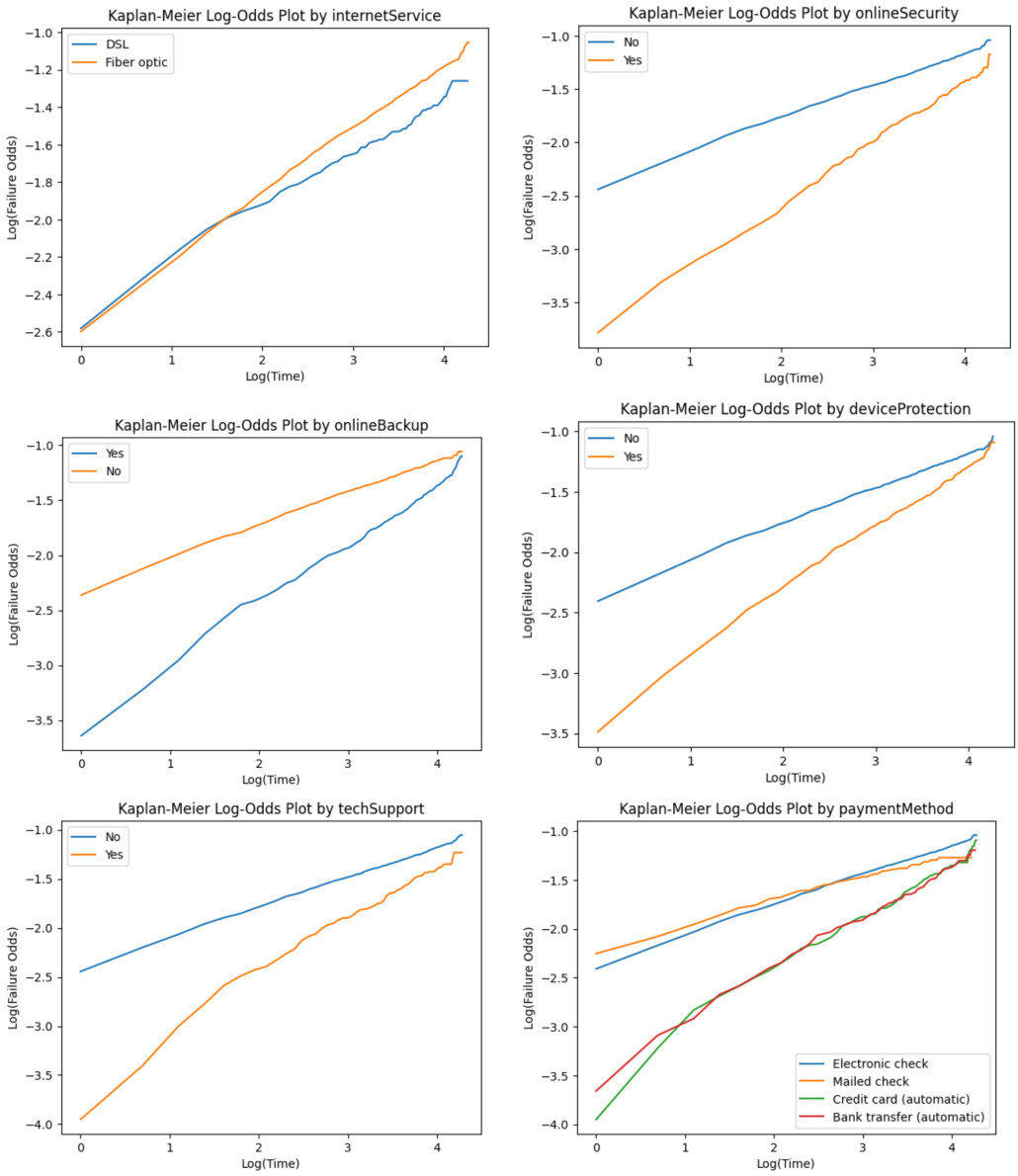
4.1 加速失效时间模型系数图



从图中可以看出，‘Intercept: alpha_‘的系数最大，表明其对客户流失时间的影响最大。其他变量如 ‘onlineSecurity_Yes: alpha_‘、‘onlineBackup_Yes: alpha_‘、‘techSupport_Yes: alpha_‘等也对客户流失时间有显著影响。

4.2 Kaplan-Meier Log-Odds Plot





从图中可以看出，对于 ‘partner‘、‘multipleLines‘、‘internetService‘、‘onlineSecurity‘、‘onlineBackup‘、‘deviceProtection‘和 ‘techSupport‘等变量，不同组别（如 Yes 和 No）的失败几率随时间的变化趋势不同，这表明这些变量对客户流失的影响显著。

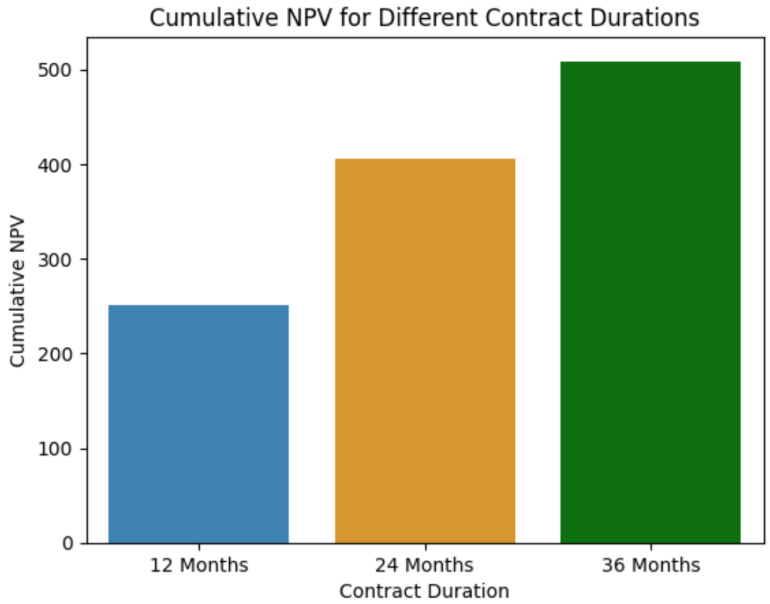
5. 客户生命周期价值（CLV）计算

客户生命周期价值（CLV）是指一个客户在其整个生命周期内为企业带来的预期净现值（NPV）。通过生存分析模型的输出，我们计算了客户的生存概率，并将其应用于客户生命周期价值的计算。假设每个客户的月度利润为 30 美元，内部收益率（IRR）设定为 10%（年化），用于计算净现值（NPV）。

```
[language=Python, basicstyle=\small\ttfamily]
# 定义计算 NPV 的函数
def calculate_npv(survival_prob, monthly_profit, irr):
    npv = 0
    for i, prob in enumerate(survival_prob):
        npv += monthly_profit * prob / ((1 + irr / 12) ** i)
    return npv

# 计算每个月的生存概率
survival_prob = cph.predict_survival_function(encoded_pd).mean(axis=1)

# 假设的月度利润和内部收益率
monthly_profit = 30
irr = 0.10  # 年化内部收益率

# 计算每个月的累计 NPV
cumulative_npv = [calculate_npv(survival_prob[:i+1], monthly_profit, irr) for i
    in range(len(survival_prob))]
```

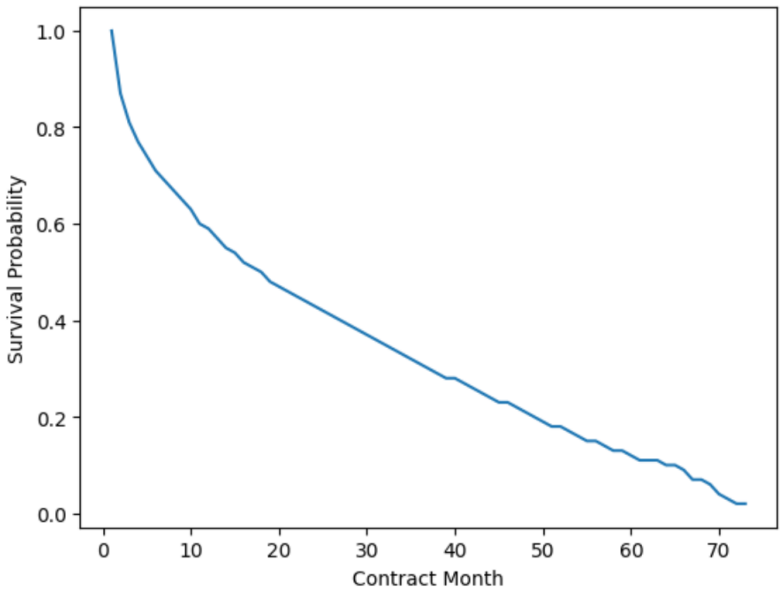
通过上述代码，我们计算了每个月的累计净现值（NPV）。以下是计算结果的可视化展示：

5.1 不同合同期限的累计净现值（NPV）



从图中可以看出，随着合同期限的延长，客户的累计 NPV 显著增加。这表明，长期合同能够为企业带来更多的经济收益。

5.2 生存概率图



从图中可以看出，生存概率随着合同月份的增加而逐渐下降，这表明随着时间的推移，客户流失的风险逐渐增加。

5.3 业务应用

通过生存分析模型计算客户生命周期价值，可以为企业提供以下业务应用建议：

- **优化客户保留策略**：通过提高客户保留率，特别是在早期阶段，可以显著提升客户的累计 NPV。
- **平衡客户获取成本（CAC）与客户生命周期价值（CLV）**：通过分析客户生命周期价值，企业可以更好地评估客户获取成本的合理性，从而提高整体业务盈利能力。

6. 结论

生存分析方法的选择需考虑模型假设和业务目标。Kaplan-Meier 方法适合单变量分析，Cox 模型适合多变量分析，AFT 模型适合参数化建模。生存分析的结果可以用于优化客户保留策略、预测客户流失时间以及计算客户生命周期价值。未来可以进一步探索 Cox 模型的改进方法，并将生存分析结果应用于更广泛的业务决策中。