# 基于\*\*的客户流失预测模型研究

网络金融部/创新应用科 王许一 李飏

**摘要：**

本文针对竞赛数据给出的电信客户相关信息，使用\*\*方法，进行客户流失预测建模研究，通过\*\*\*过程，取得\*\*成果，同时结合数据分析下探，根据客户在公司停留的生命周期阶段，按事前干预、事中运营和事后挽回，给出三种层次的分类营销建议。

## 背景介绍

目前各行各业对客户竞争越来越激烈，获取新用户的成本越来越高，如何在维护现有用户情况下进行拓新，成为通信公司维持用户数、保证收入的一个重要因素。

本次课题为电信客户流失预测，数据集共计6043条记录，每个样本包含21维特征。该数据集主要包含以下几个方面内容：

**客户属性**：客户ID、年龄、性别、是否有亲属、是否有伴侣；

**订阅服务**：电话服务、多线服务、因特网服务、在线安全、在线备份、设备保护、技术支持、流媒体电视、流媒体电影；

**账户信息**：在网时长、合同方式、无纸化账单、付款方式、月消费、总消费；

**流失特征**：是否流失；

详细字段如表1所示：

表1 特征含义

| **类别** | **字段名** | **数据类型** | **字段描述** |
| --- | --- | --- | --- |
| 个人属性 | customerID | String | 顾客ID |
| gender | String | 无论客户是男性还是女性 |
| SeniorCitizen | Integer | 客户是否为老年人（1,0） |
| Partner | String | 客户是否有合作伙伴（是，否） |
| Dependents | String | 客户是否有家属（是，否） |
| 订阅服务 | PhoneService | String | 客户是否有电话服务（是，否） |
| MultipleLines | String | 客户是否有多条线路（是，否，没有电话服务） |
| InternetService | String | 客户的互联网服务提供商（DSL，光纤，否） |
| OnlineSecurity | String | 客户是否具有在线安全性（是，否，没有互联网服务） |
| OnlineBackup | String | 客户是否有在线备份（是，否，没有互联网服务） |
| DeviceProtection | String | 客户是否有设备保护（是，否，没有互联网服务） |
| TechSupport | String | 客户是否有技术支持（是，否，没有互联网服务） |
| StreamingTV | String | 客户是否有流媒体电视（是，否，没有互联网服务） |
| StreamingMovies | String | 客户是否有流媒体电影（是，否，没有互联网服务） |
| 账户信息 | Tenure | Integer | 客户在公司停留的月数 |
| Contract | String | 客户的合同期限（每月，一年，两年） |
| PaperlessBilling | String | 客户是否有无纸化账单（是，否） |
| PaymentMethod | String | 客户的付款方式（电子支票，邮寄支票，银行转帐（自动），信用卡（自动）） |
| MonthlyCharges | Integer | 每月向客户收取的金额 |
| TotalCharges | Integer | 向客户收取的总金额 |
| 流失特征 | Churn | String | 客户是否流失（是或否） |

本文将通过上述客户属性、行为数据，通过数据处理和可视化分析，理解客户行为，采用机器学习方法，建立客户流失流失模型，及时发现有流失倾向的客户，通过特征分析，寻找相对应的策略来挽留流失客户，实现企业更大利润。

### 研究目标

本文根据竞赛课题要求进行解析，旨在分析出一套最大化覆盖流失客户的模型方案，再结合相关数据分析，按客户在公司的生命周期时间轴，以事前干预、事中运营和事后挽回三个层次，提出一定的营销建议。

## 研究思路和方法

### 1、建模目标

本文的建模分析思路主要是使用大数据模型方法，从现象发现结论，实现流失客户名单，并根据数据分析结果给出挽留及促活方案。

### 2、建模流程

本文的数据分析思路是从已有标签的数据结论出发，总结分析相关的特征，辅助相关的营销分析。

2.1 数据预处理

通过数据分析探索发现，总消费金额类型为object，将其转为数值型，继而进行缺失值统计，发现特征总消费金额有9个缺失值，同时在网时长均为0。根据业务意义将缺失值填充为月消费额，同时将对应在网时长转为1。

2.2 数据探索

本文将从客户属性、订阅服务、账户信息三个方面进行数据探索。

2.2.1总体流失情况

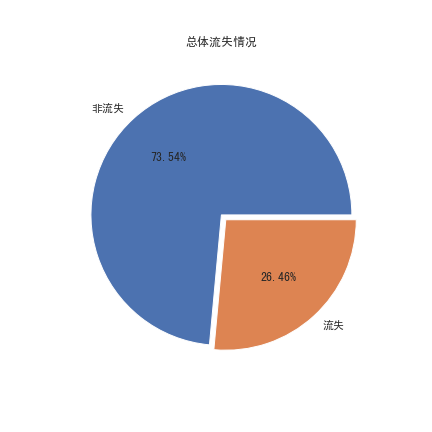
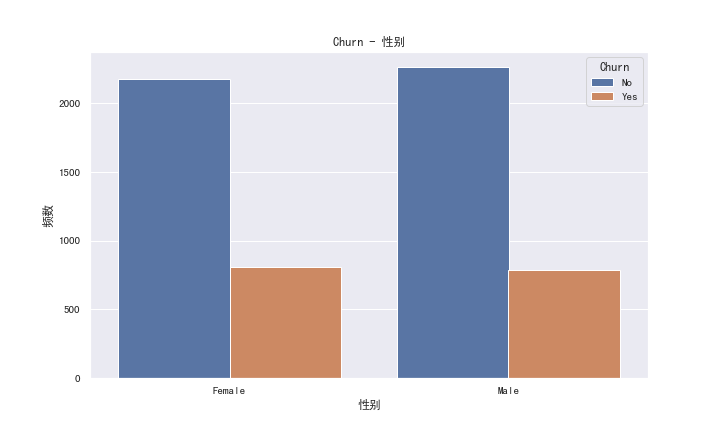
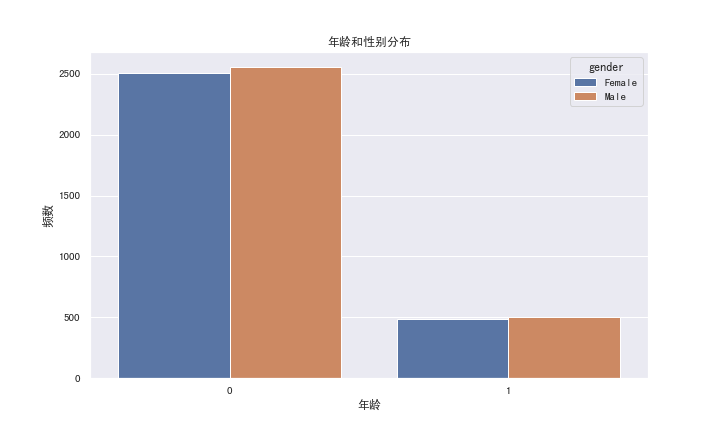
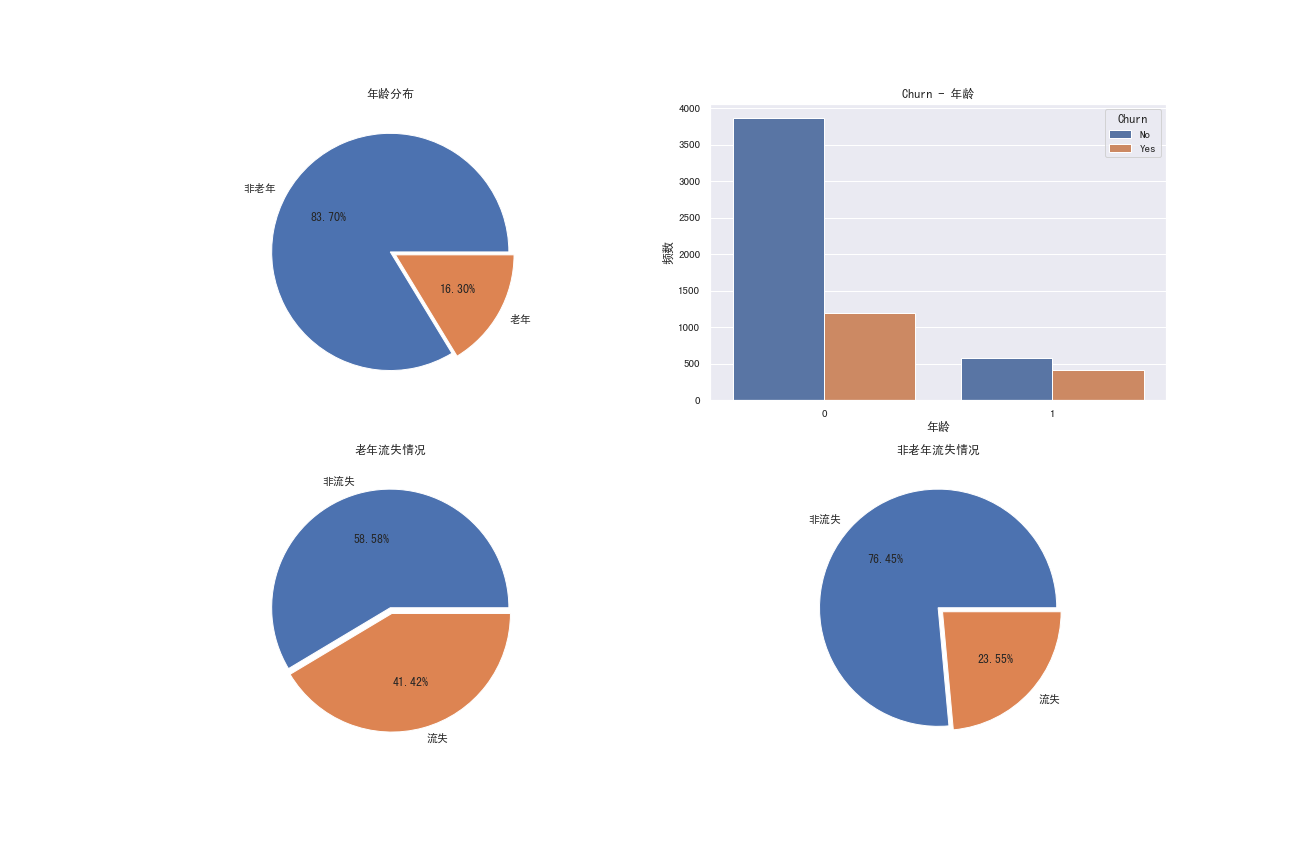


图 1 流失与非流失占比情况

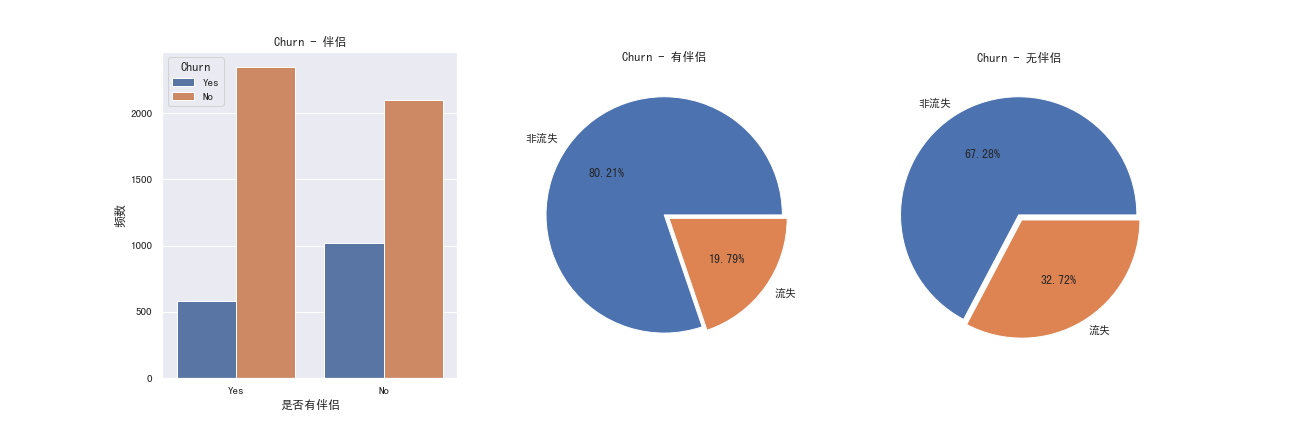
2.2.2 客户属性

a.性别流失

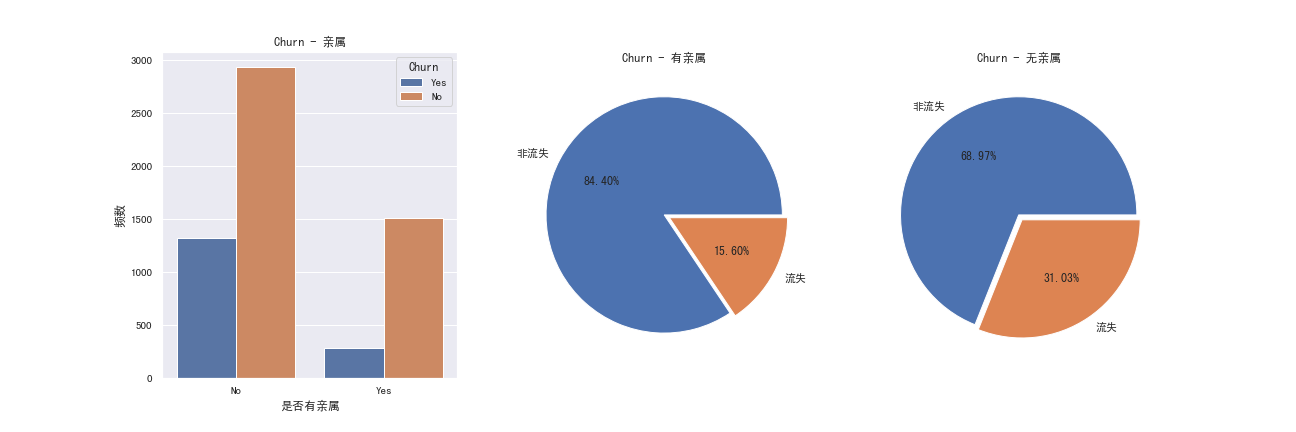


b. 年龄

c.伴侣

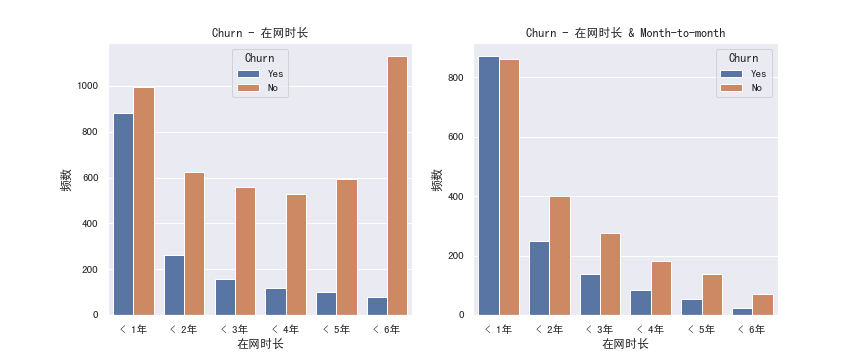
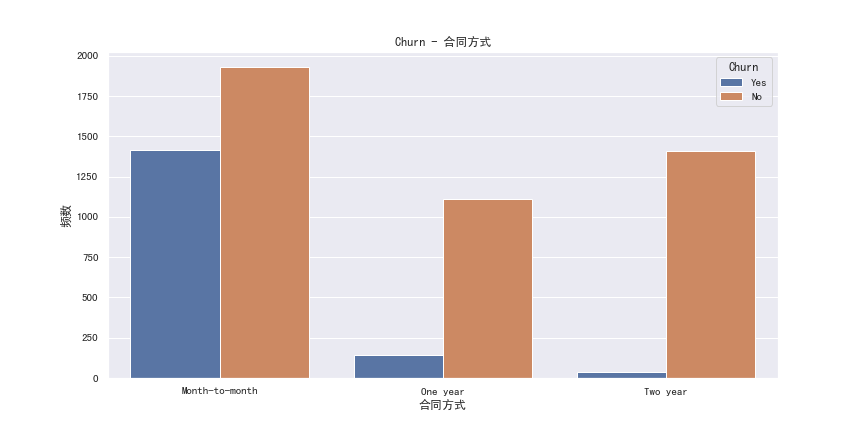


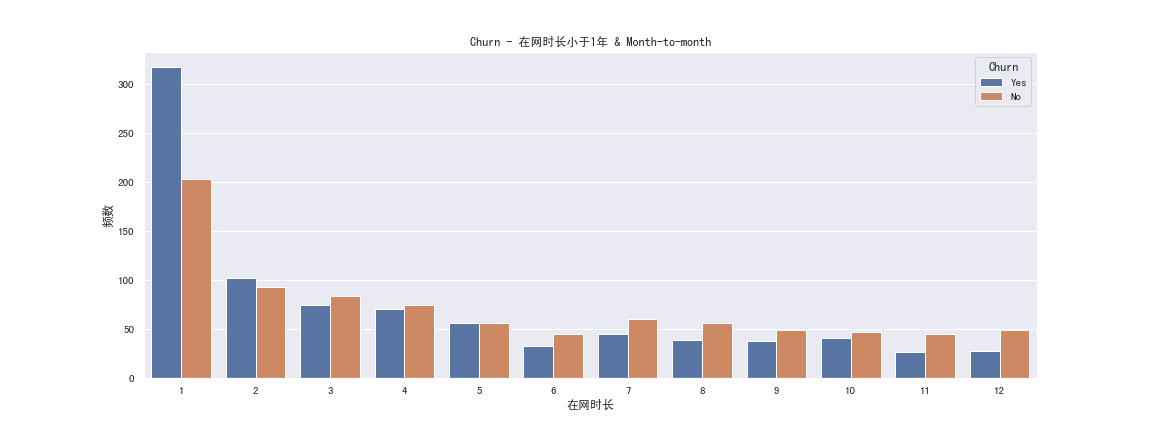
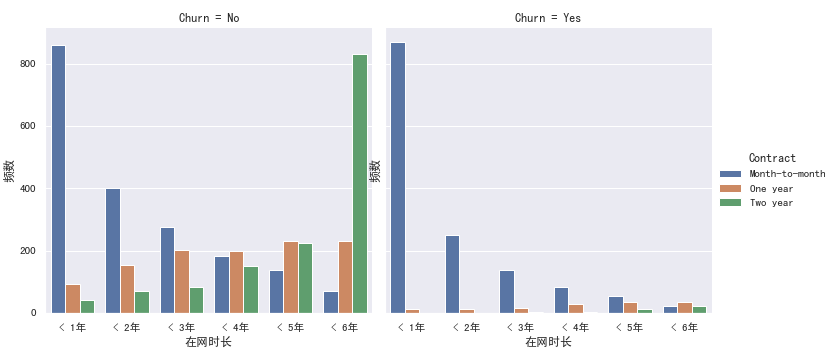
d. 亲属

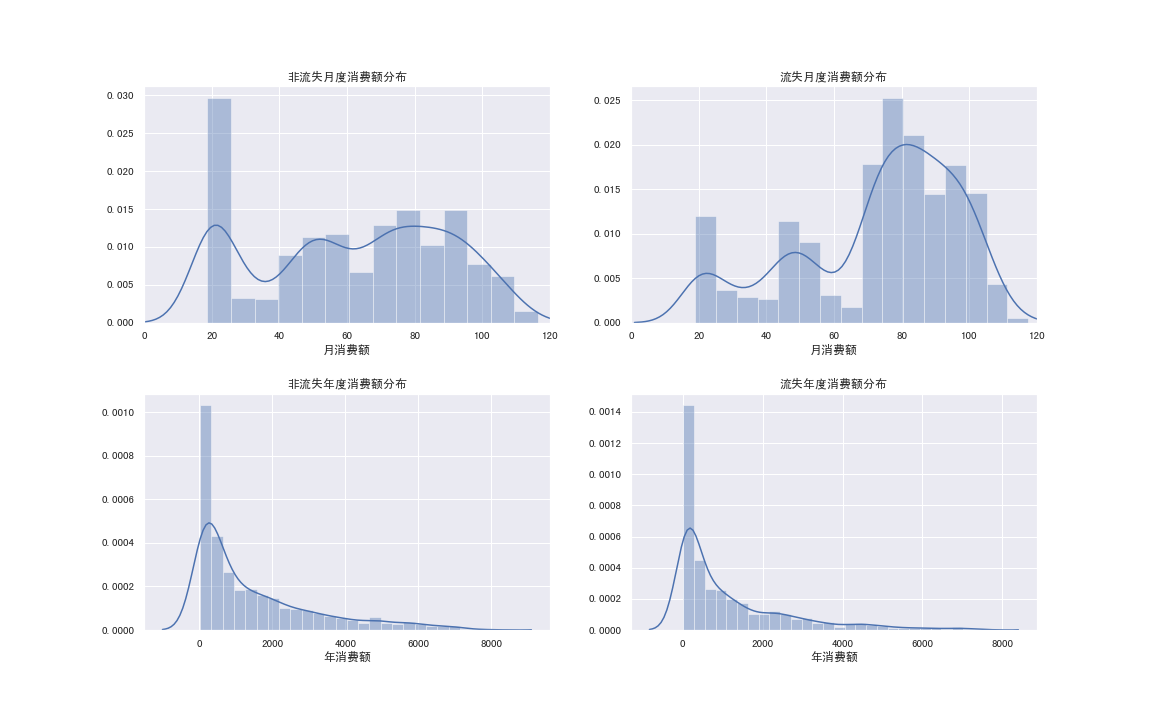


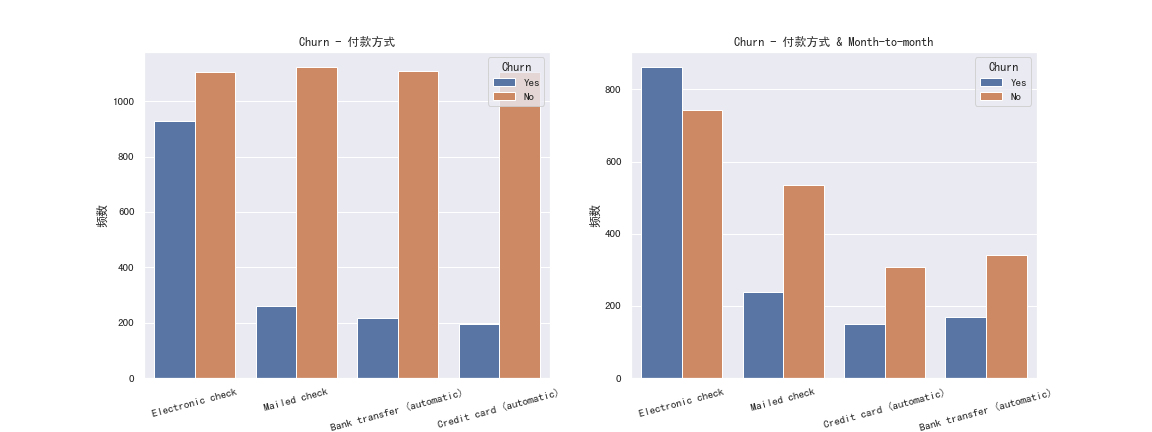
2.2.3 账户信息

a.签约合同

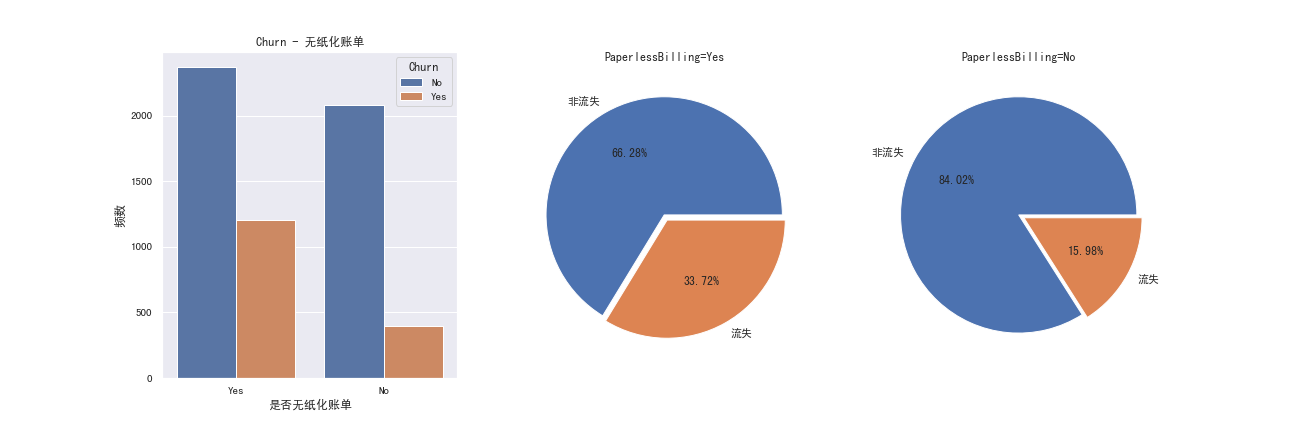




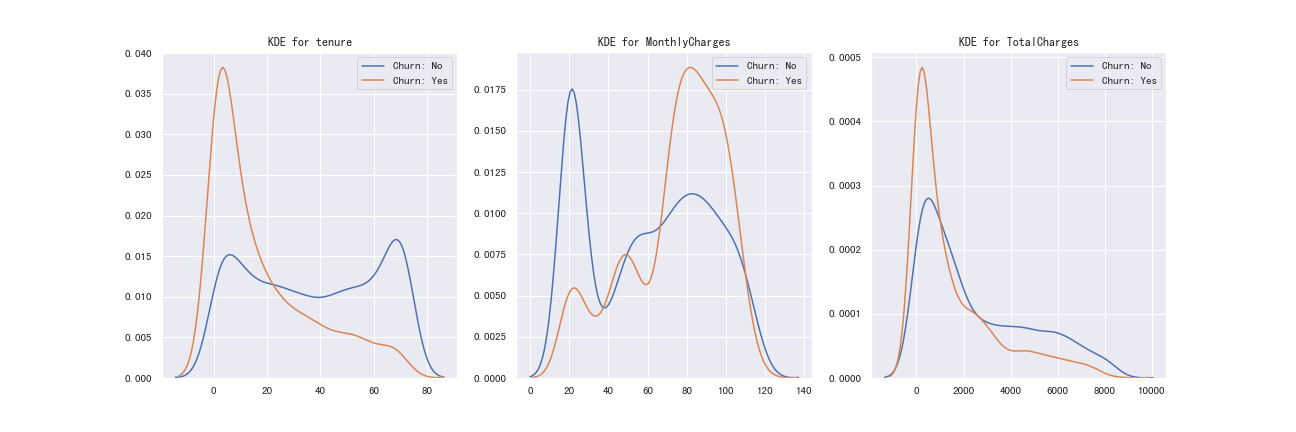




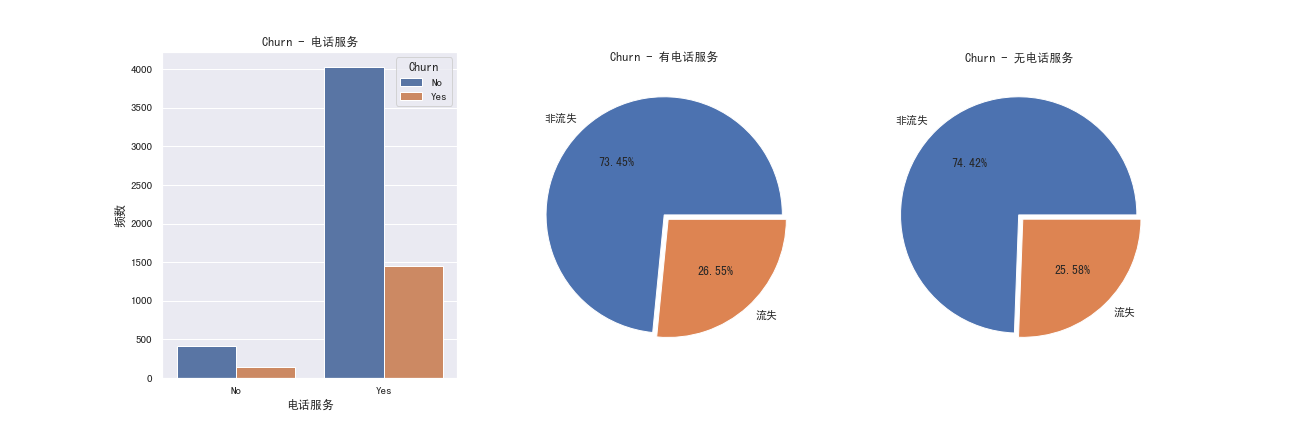
b. 无纸化



c. 数值型变量



2.2.4 签约服务



2.4 模型选择

经过前期探索，最终选用传统逻辑回归Logisticregression、集成模型XGBoost、支持向量机SVM作为主要预测模型。

数据探索中看出数据存在不均衡现象，结合商业目标，以减少客户流失为目标，最大限度识别流失客户，在进行建模过程中将对目标变量权重进行加权，提高流失客户权重，降低存留客户权重

选择常用机器学习

2.5 模型评估

通过验证集对上述模型进行最终评估，以F1-Score、AUC-Scrore、Recall-Score三种评价指标为依据，结果如下：

三、结论