基于人工智能与大数据的电信用户流失预测研究

李　涵　张旭峰
（国家开放大学）
DOI：10.65285/JSJYRGZN.V1I1.001

摘　要：针对电信行业市场饱和、获客成本远高于留存成本的现实困境，用户流失预测已成为运营商精细化运营的核心课题。本文以一份电信用户数据为研究对象，构建了"数据清洗—特征工程—模型训练—评估—业务测算"的完整分析流程，依次训练逻辑回归、随机森林与梯度提升三种机器学习模型，并从准确率、精确率、召回率、F1 值与 AUC 等维度进行系统对比，进而识别影响流失的关键因素并开展业务价值测算。研究验证了人工智能与大数据技术在解决真实商业问题中的有效性，并给出了分层挽留的落地建议。
关键词：人工智能；大数据；用户流失预测；机器学习；精准营销

Research on Telecom Customer Churn Prediction Based on Artificial Intelligence and Big Data

Han Li, Xufeng Zhang
The Open University of China

Abstract: Against the backdrop of a saturated telecom market and much higher customer acquisition cost than user retention cost, customer churn prediction has become a core task for refined operation of telecom operators. Taking real telecom user data as the research sample, this paper constructs a complete analytical framework covering data cleaning, feature engineering, model training, model evaluation and business measurement. Three machine learning models including logistic regression, random forest and gradient boosting are trained and compared comprehensively in terms of accuracy, precision, recall, F1-score and AUC. On this basis, key influencing factors of user churn are identified, and corresponding business value assessment is carried out. This study verifies the effectiveness of artificial intelligence and big data technology in solving practical industrial problems, and puts forward implementable strategies for hierarchical user retention.
Keywords: artificial intelligence; big data; customer churn prediction; machine learning; precision marketing

前言

随着 5G 网络的全面铺开与移动通信市场趋于饱和，我国电信行业已从“增量竞争”全面转入“存量竞争”阶段。工业和信息化部公布的行业数据显示，移动电话用户普及率长期维持在高位，新增用户空间十分有限。在这一背景下，用户流失（Customer Churn）问题被提到前所未有的高度：业界普遍认为，获取一名新用户的成本约为维系一名老用户成本的 5 至 7 倍，而将客户流失率降低 5% 往往能够带来可观的利润增长。因此，如何提前识别出具有流失倾向的用户并实施有针对性的挽留，直接关系到运营商的收入稳定与长期竞争力。

传统的流失管理主要依赖业务人员的经验判断或简单的规则筛选，例如“连续三个月未产生流量的用户”。这类方法阈值僵化、维度单一，既难以刻画用户行为的复杂关联，也无法对流失概率进行量化排序，导致挽留资源投放粗放、命中率低。大数据与人工智能技术的成熟为破解这一难题提供了新路径：一方面，运营商积累了海量的用户属性、消费账单与业务使用数据，构成了典型的大数据资产；另一方面，机器学习算法能够自动从高维数据中挖掘非线性规律，输出可解释、可排序的流失概率，从而支撑精准营销决策。

本文立足于一个真实可复现的电信用户流失预测场景，尝试回答三个问题：其一，哪些用户特征对流失具有显著影响；其二，不同机器学习模型在该问题上的表现差异如何；其三，模型预测结果如何转化为可量化的商业价值。围绕上述问题，本文构建了从数据预处理到业务落地的完整分析链条，所有数据分析与建模均通过 Python 实现，力求为运营商的智能化用户运营提供可操作的方法参考。

一、相关理论与技术基础

（一）大数据驱动的用户流失分析

用户流失预测本质上是一个二分类问题：以用户在观测期内的多维特征为输入，预测其在未来一段时间内是否会终止服务。大数据技术在其中的价值体现在数据的“规模性”与“多样性”上——运营商可整合用户的人口属性、合约信息、消费金额、业务订购和网络使用等结构化数据，形成高维特征空间。相比于抽样统计，全量数据能够更完整地刻画长尾用户群体的行为模式，使模型对少数高风险用户的识别更加稳健。

（二）人工智能分类方法

本文选取三种代表性的监督学习算法进行对比。逻辑

回归（Logistic Regression）通过 Sigmoid 函数将特征线性组合映射为概率，模型简洁、可解释性强，回归系数可直接反映各因素对流失的方向与强度，是工业界流失建模的常用基线。随机森林（Random Forest）以决策树为基学习器，通过对样本和特征的双重随机抽样构建多棵树并投票集成，能够捕捉特征间的非线性交互，对噪声和过拟合具有较好的鲁棒性。梯度提升（Gradient Boosting）则采用前向分步的加法模型，逐轮拟合前一轮的残差，将大量弱学习器串联为强学习器，通常在结构化数据的分类任务上具有领先的精度表现。三者在可解释性与拟合能力上各有侧重，适合用于横向比较。

二、数据来源与预处理

（一）数据集说明

本文使用的数据集共包含 7043 条用户记录，涵盖用户的基础属性、合约与消费信息，以及是否流失的标签。经统计，数据集整体流失率为 26.82%，属于典型的类别不平衡数据。主要字段及含义如表 1 所示。

表 1　数据集主要字段说明

字段名称	类型	含义说明
在网月数	数值	用户在网时长（0—72 个月）
月费	数值	当月应缴费用（元）
总消费	数值	累计消费金额（元），含少量缺失
合约类型	类别	按月 / 一年 / 两年
网络类型	类别	光纤 / DSL / 无
支付方式	类别	电子支票 / 邮寄支票 / 自动转账 / 信用卡
技术支持	类别	是否订购技术支持服务
老年用户	类别	是否为老年用户（0/1）
是否流失	标签	目标变量，1 表示流失

（二）数据清洗与特征工程

原始数据存在两类需要处理的问题。其一是缺失值：“总消费”字段约有 1.5%（105 条）的记录缺失，考虑到该字段分布右偏，本文采用中位数填充以降低极端值的影响。其二是类别变量的编码：合约类型、网络类型、支付方式与技术支持均为文本型类别特征，无法直接输入模型，需通过独热编码（One-Hot Encoding）转换为数值向量，并删除首列以避免多重共线性。此外，由于逻辑回归对特征量纲敏感，需对数值特征进行标准化处理。核心预处理代码如下：

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

# 缺失值填充（中位数）
df['总消费'] = df['总消费'].fillna(df['总消费'].median())

# 类别变量独热编码
df = pd.get_dummies(df, columns=['合约类型','网络类型',
                '支付方式','技术支持'], drop_first=True)

X = df.drop(columns=['是否流失']); y = df['是否流失']
X_tr, X_te, y_tr, y_te = train_test_split(
    X, y, test_size=0.25, stratify=y, random_state=42)
scaler = StandardScaler()
X_tr_s = scaler.fit_transform(X_tr)
X_te_s = scaler.transform(X_te)
```

其中，train_test_split 按 25% 的比例划分测试集，并通过 stratify 参数保持训练集与测试集的流失比例一致，确保评估的公平性。

三、探索性数据分析

在建模之前，本文先对各特征与流失之间的关系进行探索性分析，以形成业务直觉并验证特征的有效性。图 1 展示了不同合约类型的流失率以及流失率随在网时长的变化趋势。

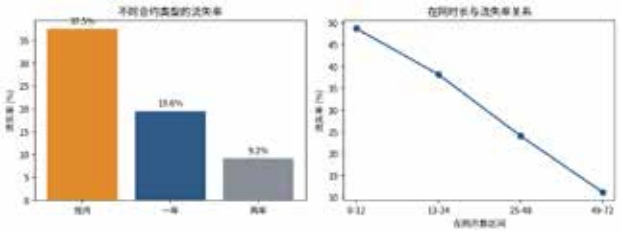

图 1　合约类型与在网时长对流失率的影响

分析结果揭示出清晰的规律。从合约维度看，按月合约用户的流失率高达 37.5%，而一年、两年合约用户的流失率分别降至 19.6% 和 9.2%，长期合约对用户具有显著的锁定作用。从在网时长看，流失率随时长增加单调下降：入网 0—12 个月的新用户流失率高达 48.7%，13—24 个月降至 38.2%，而在网 49 个月以上的老用户流失率仅为 11.2%，说明用户生命周期早期是流失的高危阶段。此外，进一步统计显示，流失用户的平均月费为 76.0 元，明显高于未流失用户的 64.9 元；使用电子支票支付的用户流失率（33.8%）也高于其他支付方式。这些发现为后续的特征重要性分析提供了业务佐证。

四、模型构建与实验

（一）建模流程与评价指标

本文对三种模型采用统一的训练与评估流程。针对类别不平衡问题，逻辑回归与随机森林均启用 class_weight='balanced' 参数，对少数类（流失）赋予更高权重，以提升模型对流失用户的识别能力。建模核心代码如下：

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import (RandomForestClassifier,
                GradientBoostingClassifier)
```

```
from sklearn.metrics import roc_auc_score, f1_score

models = {
  ' 逻辑回归 ': LogisticRegression(max_iter=1000,
          class_weight='balanced', random_state=42),
  ' 随机森林 ': RandomForestClassifier(n_estimators=300,
          max_depth=9, class_weight='balanced'),
  ' 梯度提升 ': GradientBoostingClassifier(n_estimators=250,
          max_depth=3, learning_rate=0.08),
}
for name, m in models.items():
    m.fit(X_tr, y_tr)
    proba = m.predict_proba(X_te)[:, 1]
    print(name, round(roc_auc_score(y_te, proba), 4))
```

由于流失预测的目标是尽可能多地找出潜在流失用户，召回率（Recall）与综合了精确率和召回率的 F1 值比单纯的准确率更具业务意义；同时，AUC 衡量模型在不同阈值下区分正负样本的综合能力，不受阈值选择的影响，是模型排序能力的关键指标。

（二）实验结果对比

三种模型在测试集（1761 户）上的性能指标如表 2 所示。

表 2　三种模型性能指标对比

模型	准确率	精确率	召回率	F1 值	AUC
逻辑回归	0.753	0.526	0.795	0.633	0.839
随机森林	0.768	0.551	0.739	0.631	0.831
梯度提升	0.801	0.682	0.485	0.567	0.832

结果呈现出鲜明的“各有所长”格局。梯度提升模型的准确率（0.801）与精确率（0.682）最高，意味着它预测为流失的用户中确为流失的比例更高，但其召回率仅为 0.485，会漏掉过半的真实流失用户。相比之下，逻辑回归以 0.795 的召回率和 0.633 的 F1 值居首，能够捕获测试集中约 79% 的潜在流失用户，更契合“宁可错杀、不可放过”的挽留业务目标。图 2 的 ROC 曲线进一步显示，三种模型的 AUC 非常接近（0.831—0.839），排序能力相当，其中逻辑回归以 0.839 微弱领先。

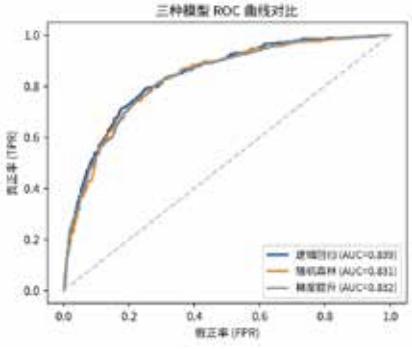

图 2　三种模型的 ROC 曲线对比

综合来看，在本场景下并非越复杂的模型越优。逻辑回归凭借更高的召回率、更强的可解释性以及更低的部署成本，成为兼顾业务价值与落地效率的首选模型。

五、结果分析与业务应用

（一）关键影响因素识别

为解释模型的决策依据，本文提取了梯度提升模型的特征重要性，结果如图 3 所示。

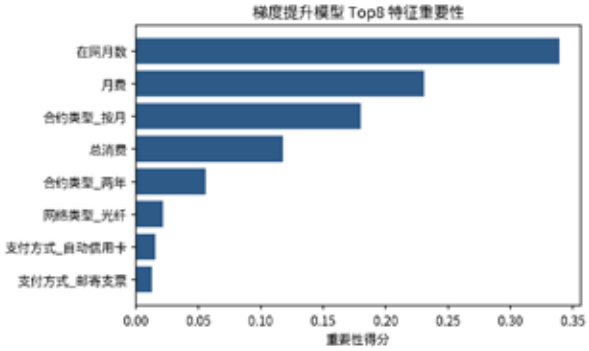

图 3　梯度提升模型的 Top8 特征重要性

特征重要性排序显示，合约类型（两年）、在网月数与月费位居前列，这与第 4 节的探索性分析结论高度一致，交叉印证了模型的合理性。具体而言，长期合约与较长的在网时长是抑制流失的“正向保护因素”，而较高的月费、光纤网络类型以及电子支票支付方式则是推高流失风险的“负向因素”。这一结论为业务干预提供了明确的抓手：运营商可通过合约升级激励、新用户关怀和资费优化三个方向系统性地降低流失。

（二）业务价值测算

模型的最终价值需回归到商业收益。本文以逻辑回归模型的预测结果为基础进行测算：假设挽留一名用户的营销成本为 50 元，一名用户流失造成的年收入损失为 800 元，而挽留活动的成功率为 30%。在 1761 户测试样本中，若不做任何干预，472 户流失用户将造成约 37.76 万元的预期损失；采用模型后，系统触达 713 户高风险用户，投入挽留成本约 3.57 万元，成功挽留可挽回约 9.00 万元收入，最终实现约 5.44 万元的净收益。测算表明，即便在保守的成功率假设下，基于人工智能的精准挽留仍显著优于“广撒网”或“不作为”策略。

（三）分层挽留落地建议

结合模型输出的流失概率，本文建议运营商实施分层挽留策略。对高概率流失用户（预测概率大于 0.7），应由客户经理主动外呼并配套专属优惠，实施重点挽留；对中等概率用户（0.4—0.7），可通过短信、App 推送等低成本触点推荐合约升级或增值权益；对低概率用户则维持常规服务、避免过度打扰。同时，鉴于新用户与按月合约用户是流失高发群体，运营商应在用户入网初期强化关怀，并设计阶梯式的长约激励，从源头上降低流失率。

结语：

本文以电信用户流失预测这一真实商业问题为切入点，

构建了融合大数据处理与人工智能建模的完整解决方案。研究表明：合约类型、在网时长与月费是决定用户流失的核心因素；在类别不平衡场景下，经权重调整的逻辑回归模型以 0.795 的召回率和 0.839 的 AUC 取得了兼顾业务价值与可解释性的最佳综合表现；基于预测结果的分层精准挽留策略可带来可观的净收益。这印证了人工智能与大数据技术在解决实际运营问题中的巨大潜力。

本研究仍存在可拓展之处。未来工作可引入用户通话详单、投诉记录等时序行为数据，采用 XGBoost、LightGBM 等更先进的集成算法或深度学习模型进一步提升精度；同时，可结合 SHAP 等可解释性方法对个体预测进行归因，并将模型嵌入实时营销系统，实现从“预测”到“自动化决策”的闭环，持续释放数据资产的商业价值。

参考文献：

[1] 陈志强 , 李明 . 基于机器学习的电信客户流失预测模型研究 [J]. 计算机应用研究 ,2022,39(4):1123-1128.

[2] 周涛 , 王芳 . 大数据背景下用户流失预警系统的设计与实现 [J]. 数据分析与知识发现 ,2021,5(9):45-54.

[3] 刘洋 , 张伟 . 类别不平衡数据下的客户流失预测方法比较 [J]. 统计与决策 ,2023,39(6):78-82.

[4]Verbeke W, Dejaeger K, Martens D, et al. New insights into churn prediction in the telecommunication sector[J]. European Journal of Operational Research, 2012, 218(1): 211-229.

[5]Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[6]Friedman J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[7]Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[8] 黄晓峰 , 赵磊 . 人工智能驱动的精准营销研究综述 [J]. 管理评论 ,2022,34(8):156-168.

作者简介：李涵（1993.10-）、女、汉族、江苏泰兴人、硕士、会计师、经济师、研究方向：财务大数据分析。