第 8 卷◆第 11 期◆版本 1.0◆2025 年
文章类型：论文 | 刊号（ISSN）：2630-4686 /（中图刊号）：380GL020

数据挖掘分类算法在客户流失预测中的实践

朱少强
广东工业大学
DOI:10.12238/er.v8i11.6590

[摘　要] 客户流失已成为企业持续经营中的关键挑战，文章研究了数据挖掘中分类算法在客户流失预测任务中的应用流程，分析了数据预处理、特征构建、模型选取与评估策略，探讨了预测结果在高风险客户识别与干预机制中的落地路径，验证了模型在实际业务中的可部署性与预测效果，对提升客户管理精度与企业决策智能化水平具有重要的工程应用价值。
[关键词] 客户流失预测；分类算法；数据预处理；模型评估；应用实践
中图分类号：TP311 文献标识码：A

The Practice of Data Mining Classification Algorithm in Customer Churn Prediction

Shaoqiang Zhu
Guangdong University of Technology

Abstract: Customer churn has become a key challenge in the continuous operation of enterprises. This paper studies the application process of classification algorithm in data mining in customer churn prediction task, analyzes data preprocessing, feature construction, model selection and evaluation strategy, discusses the landing path of prediction results in high-risk customer identification and intervention mechanism, and verifies the deployability and prediction effect of the model in actual business, which has important engineering application value for improving customer management accuracy and enterprise decision-making intelligence level.
Keywords: customer churn prediction ; classification algorithm ; data preprocessing ; model evaluation ; application Practice

引言

在用户生命周期管理日益精细化的背景下，客户流失问题对企业运营稳定性提出更高要求，准确预测潜在流失客户已成为数据驱动决策的重要方向。分类算法因其在非线性建模、样本识别与泛化能力上的表现，被广泛用于流失倾向判断任务中。研究中需解决特征变量选择、数据预处理规范、模型构建路径及评估体系的关键技术问题，明确模型输出在客户分层与挽留干预中的具体应用机制。文章围绕分类算法在客户流失预测中的实践展开系统分析，从数据准备到模型训练，再到预测结果在运营中的落地应用，构建一套可实施的技术路径体系，为企业提升客户留存率与响应策略智能化水平提供方法支持。

1 客户流失预测的数据准备

1.1 客户信息的特征提取方法

客户流失预测模型的有效性依赖于特征变量对用户状态的刻画精度。特征提取需涵盖静态属性与动态行为两个维度。静态属性包括年龄、性别、注册渠道、所在区域等固定信息，动态行为侧重于客户在指定周期内的活跃频次、登录时长、访问路径、交易行为、服务响应记录等交互特征。可构建如近 30 日活跃天数、近三月消费次数、平均客单价、客服联系频次、投诉响应延迟等变量[1]。针对交易记录，可引入时间窗分组聚合特征，如月消费波动率、行为周期性指标。部分行为特征可基于时间序列变化率生成衍生变量，以捕捉客户活跃度和黏性趋势的边界特征，提升模型对潜在流失客户的识别能力。

1.2 训练数据的预处理流程

客户行为数据中存在缺失、离群、类别不平衡等问题，需在建模前进行系统性处理。缺失值根据变量属性选择均值填充、中位数插补、同类客户聚类均值填充等策略，并记录填充状态作为新特征以保留异常信息。异常值检测采用 IQR 方法或 Z-score 标准化方法识别行为指标的极端样本，对明显失真记录进行剔除或压缩。变量类型差异较大，需对类别型变量采用 One-Hot 或目标编码，对连续型变量实施 Z-score 或 Min-Max 归一化处理。样本不平衡问题常见于流失客户占比较低情形，可结合 SMOTE 过采样与 Tomek Link 清除边界样本策略提升样本分布合理性。所有处理环节应构建数据流水线以保障处理一致性与模型训练的稳定性。

2 分类算法的选取与模型构建

2.1 常见模型的适配性分析

客户流失预测任务面临数据特征维度多、样本不平衡和流失行为非线性演化等问题，不同分类模型的适配性直接影响预测效果。逻辑回归模型结构简单，可解释性强，适合线性可分问题，对特征维度较少且变量间关系稳定的数据集表现稳定，但难以捕捉复杂的非线性特征交互[2]。决策树模型具备自动特征选择能力，能处理非线性关系且对异常值不敏感，结构可视化有助于业务理解，但容易过拟合且泛化能力受限。随机森林作为集成学习方法，集成多个弱分类器构建强模型，对特征噪声与数据稀疏情况具有较强鲁棒性，能在不平衡数据上通过类别权重调整优化预测表现，但其模型输出缺乏直观解释性。针对客户流失问题，可优先选取随机森林作为主模型，对比逻辑回归模型在概率输出与可解释性方面的表现，综合评估其在具体业务场景下的适应度。

2.2 分类模型的构建流程

构建客户流失预测模型需设定完整的训练流程以确保数据质量、特征价值和模型稳定性。数据输入阶段基于清洗后的数据集建立统一格式，设定目标变量为客户是否流失，输入变量包括静态信息、行为指标与交易频率类变量。特征选择通过卡方检验、信息增益、L1 正则筛选或基于模型的特征重要性评估方法进行维度压缩，剔除低相关性或冗余变量以提升训练效率。模型训练阶段根据前期模型适配性分析结果构建逻辑回归、随机森林等模型，采用网格搜索或贝叶斯优化算法调整超参数，提升模型泛化能力。在交叉验证中采用 K 折交叉策略，保障样本间分布一致性，记录各轮准确率、AUC 值与 F1 分数，对比不同模型在各项数据上的表现，判断模型在不同样本组合下的稳定程度，为后续模型集成与部署提供性能基准。

2.3 模型性能的评估方法

评估客户流失预测模型需结合多维指标体系以刻画模型在实际应用场景中的表现。混淆矩阵用于直观展示预测结果中 TP、FP、TN、FN 的分布情况，可计算准确率、召回率、精确率等基本指标，捕捉模型在正负样本判别中的偏向性。F1 分数作为精确率与召回率的调和均值，适用于处理样本不均衡场景，衡量模型对少数类流失客户的识别效果。AUC 值用于度量模型在不同阈值下的整体区分能力，反映预测排序质量，特别适用于概率输出型模型。进一步可引入 KS 值分析客户分层效果，确定风险分段的稳定性。为保障模型在上线部署中的业务适应性，建议在验证集和测试集上分别记录指标变化趋势，并结合模型复杂度与响应速度等非精度性指标，完成针对不同业务场景的多模型评估与筛选。

3 预测结果的应用实践

3.1 高风险客户的识别流程

在分类模型输出结果中，每个客户样本会被赋予一个属于“流失”类别的概率值，该概率值可用于后续客户流失风险等级划分[3]。为提升识别精度与业务匹配度，需结合实际历史流失样本的分布特征对概率阈值进行优化设定，常采用 ROC 曲线下的 Youden 指数或最大 KS 值作为最优划分依据。在金融、电信等行业中，常将模型输出概率划分为多段，用于控制干预资源投放范围。阈值设置过程需与业务团队协同，确保模型的判断结果与实际运营目标一致。高风险识别结果可形成客户列表，结合客户 ID 与基础属性编码，供后续策略系统或外呼平台直接调用使用。

在实际部署中，模型输出需嵌入评分引擎或推荐系统中作为实时或定期运行模块，对客户进行周期性扫描。为避免模型输出波动影响运营策略，应对风险评分结果引入滑动窗口平滑机制或分布监控逻辑，防止因模型漂移导致风险分段失真。对模型打分后进入高风险区间的客户，可进一步聚合行为标签与历史转化记录，增强风险判断的上下文解释力。例如近 30 天登录频率降低、服务响应频次上升、投诉词频高发等特征叠加，具备较强流失前兆特征。通过构建风险触发条件与规则引擎联动机制，使识别逻辑具备稳定性与实时性，为后续精准干预奠定基础[4]。

3.2 预测结果的分级应用策略

针对模型输出的流失概率结果，可设定多档风险分层机制，按概率区间将客户划分为高风险、中风险、低风险等等级，每一级风险层对应差异化响应策略。高风险客户优先纳入一对一外呼干预或高频互动渠道，中风险客户进入周期性触达策略池，低风险客户维持常规推送频率。风险分层标准除模型得分外，可叠加客户生命周期阶段、历史价值贡献、营销敏感度等多维指标构建综合评分函数，以实现更精细的策略匹配。策略执行系统需具备策略流控与执行频率监测能力，确保干预手段的资源消耗与客户体验之间取得平衡。

为提升分层策略在实际运营中的响应效果，应结合历史干预记录建立转化标签，追踪各风险层客户在不同策略下的行为变化。模型输出的概率值可转化为标准化评分与策略编号，写入客户主数据表，供 CRM 系统、外呼平台、短信平台统一调用。针对不同风险等级的客户，可制定标签化模板内容，结合客户兴趣画像进行多渠道触达，包括短信提醒、APP 推送、电话回访等方式。策略系统可建立策略－响应－结果的数据闭环，以实现对不同风险层策略执行后的效果评估，为模型反馈与策略更新提供数据依据[5]。

3.3 挽留方案的实施机制

挽留机制的设计应以客户流失原因的结构化分析为前

提，将模型识别出的高风险客户作为触发对象，通过特征画像与行为轨迹匹配对应的干预路径。系统中需建立干预规则与客户标签之间的映射逻辑，典型如消费金额下降、登录频率骤降、客服互动密度增加等特征组合同步触发干预模块。针对价格敏感型客户可制定优惠券推送、分期付款提醒等价格引导策略，对服务体验型客户则推送满意度调研邀请、专属客户经理回访等服务修复策略。所有挽留动作需具备执行权限控制、记录日志与客户反馈采集接口，保障策略过程可溯可控。干预模板应实现参数化配置，便于批量调用与个性化生成，提升内容贴合度与响应转化效率。

策略部署平台应具备多渠道联动能力，可根据客户偏好优先选择短信、电话、App 内推送、微信公众号、邮箱等触达方式，确保信息传达高可达、高点击。为避免策略冲突与资源浪费，需引入挽留窗口期控制机制，设定同一客户单位周期内仅允许接受一次高强度干预，并在系统中设定策略冷却时间与效果等待期。当客户对某一类干预方式存在屏蔽或负反馈行为时，系统可自动调整策略方案或降级处理，降低客户反感度。各类策略执行后，需在系统中实时记录响应行为，包括点击率、回访率、交易转化率、取消流失标签的时间节点等，构建完整的干预效果链条。策略执行节点可与外呼中心、CRM 系统、营销自动化平台打通，实现从识别、策略制定、执行、监控的全流程闭环[6]。

策略落地后的效果监控机制需嵌入挽留模块内部，可设置日度、周度与月度策略执行报表，监测不同干预路径在客户不同生命周期阶段的转化表现。系统可设定策略 AB 测试模块，自动分流客户进入不同干预路径，比较挽留效果，形成数据驱动的策略优化逻辑。策略输出结果需反哺建模层，更新客户行为数据与标签状态，供下一轮模型训练使用，实现预测与干预之间的持续迭代。在业务运营中，可设置“复流客户评分模型”，对已挽回客户的后续行为进行量化，判断干预是否具备长期价值。当挽留失败的客户进入沉默状态后，还可接入唤醒模型识别其二次激活机会，构建从流失识别、干预、评估到唤醒的多周期联动机制，提升整体客户生命周期管理能力[7]。

4 结语

研究围绕客户流失预测中的关键技术路径展开，构建了以特征提取、数据预处理、模型构建与评估为核心的分类算法应用流程，完成了高风险客户识别、分层响应策略制定与干预机制闭环的全流程实践验证，实证表明随机森林等模型具备较强的流失识别能力与业务适配性，配套的策略系统支持差异化触达与响应追踪，有效提升客户留存效率，未来可在实时预测部署与深度行为建模方向进一步拓展算法适应边界与平台集成能力。

[参考文献]

[1]张芸.基于复合 CatBoost 的银行客户流失预测模型[D].兰州大学,2021.

[2]高欣怡.基于大数据算法的 G 物流公司客户流失分析[D].东华大学,2019.

[3]刘叶.基于多算法融合的电子商务客户流失预测算法研究[D].昆明理工大学,2019.

[4]王圣节,张庆红. 基于可解释机器学习模型的电信行业客户流失预测研究[J].电信科学,2024,40(7):121-133.

[5]汪昱霖. 基于集成学习的广电客户流失预测模型研究[J].电视技术,2024,48(7):59-66.

[6]李龙戈,郑铿城. 基于集成森林元学习网络的客户流失预测[J].电信科学,2024,40(10):163-172.

[7]雷中锋,曹旭,霍振兴,等. 基于云计算数据库的运营商客户流失预测方法[J].电脑编程技巧与维护,2024,(5):104-106.

作者简介：

朱少强（1975.10-），男，汉族，湖北黄冈人，博士，讲师，研究方向为数据分析与决策。