基于机器学习的5G潜在用户识别研究

宋晨旭
合肥工业大学　安徽合肥　230000

摘　要：随着5G技术的全面商用与数字经济的深度融合，精准识别5G潜在用户成为运营商优化资源配置、提升市场竞争力的关键。本文以用户行为数据、消费特征与网络使用偏好为核心，构建基于机器学习的5G潜在用户识别体系。首先，通过数据预处理技术完成特征工程，筛选出用户流量消耗、套餐类型、终端设备、消费频次等关键指标；其次，对比逻辑回归、随机森林、支持向量机（SVM）及梯度提升决策树（GBDT）四种经典机器学习算法的识别性能；最后，基于某运营商真实用户数据集开展实证分析。实验结果表明，GBDT算法在准确率、召回率与F1分数上均表现最优，分别达到89.7%、88.3%与89.0%，显著优于其他算法。研究成果可为运营商制定差异化营销战略、精准触达潜在用户提供技术支撑与决策参考，同时为通信行业用户识别研究提供新的思路与方法。

关键词：机器学习；5G；潜在用户识别；特征工程；算法对比；用户行为分析

引言

5G技术作为新一代信息通信技术的核心载体，凭借其高速率、低时延、广连接的技术优势，已广泛应用于工业互联网、智慧医疗、车联网等多个领域，推动社会生产生活方式的深刻变革。截至目前，我国5G基站数量已突破300万个，5G移动电话用户数超8亿户，但仍有大量4G用户及传统通信用户尚未迁移至5G网络，存在巨大的市场增量空间。对于通信运营商而言，传统的“广撒网”式营销模式不仅成本高昂，且用户转化率较低，难以适应5G市场的竞争需求。因此，如何利用大数据与机器学习技术，从海量用户中精准挖掘潜在5G用户，成为当前运营商亟待解决的关键问题。本文基于用户多维度特征数据，构建机器学习识别模型，通过算法优化与实证验证，实现潜在5G用户的高效识别，为运营商的市场拓展与资源优化提供科学依据。据某通信行业报告显示，传统营销模式下5G用户转化率不足5%，而精准营销模式可将转化率提升至15%以上，成本降低30%左右[1]。

一、相关理论与研究现状

（一）5G用户特征与需求分析

5G用户的核心需求与4G用户存在显著差异，主要体现在三个维度：一是高速率需求，5G峰值速率可达10Gbps，是4G的数十倍，能够满足超高清视频（8K/120fps）、云游戏（实时渲染）、VR/AR（沉浸式交互）等大流量应用的极致体验需求，此类用户通常具有较高的流量消耗水平，月均流量普遍在20GB以上，部分重度用户甚至超过50GB，对网络带宽的稳定性要求严苛；二是场景化需求，不同行业用户的需求呈现细分特征：工业互联网用户关注低时延（<10ms）与高可靠性（99.999%），以支撑远程设备控制、机器视觉质检等实时应用；智慧家庭用户注重多设备连接（支持超百台物联网设备）与场景化服务（如智能安防、能源管理）；个人用户则偏向移动办公（云文档实时协作）与娱乐体验（车载AR导航）的深度融合；三是消费能力特征，5G套餐价格相对4G高出30%–50%，潜在用户通常具有稳定的中高收入来源（月收入8000元以上占比超60%），消费意愿较强，对云存储、视频会员、智能家居控制等增值服务的接受度较高（调研显示接受率达75%）。此外，终端设备也是重要特征之一，使用5G兼容终端的用户转化为5G用户的概率显著高于4G终端用户，某运营商数据显示，5G终端用户的转化率超60%，而4G终端用户不足10%，且终端更新周期越短（如1–2年更换一次），转化意愿越强。

（二）潜在用户识别技术研究现状

潜在用户识别本质上属于二分类问题，目前相关研究主要集中在传统统计方法与机器学习方法两个方向。传统统计方法以Logit模型、Probit模型为代表，通过分析用户特征与行为的相关性进行预测，但此类方法对特征的线性假设较强，难以处理复杂的非线性关系。随着大数据技术的发展，机器学习方法凭借其强大的特征学习与拟合能力，已成为用户识别的主流技术。现有研究中，随机森林、SVM、神经网络等算法被广泛应用于通信行业的用户识别场景：例如，李等人（2022）基于用

户消费数据与网络行为数据，采用随机森林算法构建5G潜在用户识别模型，准确率达到82.5%；王等人（2023）结合梯度提升树与特征选择技术，有效提升了模型的泛化能力，但该研究未考虑用户场景化需求特征。总体而言，现有研究在特征维度的完整性与算法优化方面仍有提升空间，亟需构建更全面的特征体系与更高效的识别模型[2]。

二、5G潜在用户识别模型构建

（一）数据来源与预处理

本文数据来源于某大型通信运营商的用户数据库，选取2023年1–6月的用户行为数据作为研究样本，共包含10万条用户记录。数据维度涵盖用户基本信息（年龄、性别、收入水平）、消费特征（月均消费金额、套餐类型、增值服务订阅情况）、网络行为特征（月均流量消耗、通话时长、上网时段分布）、终端设备特征（终端类型、使用年限、是否支持5G）及服务反馈特征（投诉次数、满意度评分）。数据预处理过程主要包括三个步骤：一是数据清洗，剔除缺失值占比超过30%的样本，采用均值填充法处理数值型缺失值，众数填充法处理分类变量缺失值；二是数据标准化，对流量消耗、消费金额等连续型特征采用Z–score标准化，消除量纲差异对模型的影响；三是特征编码，对套餐类型、终端类型等分类变量采用独热编码（One–Hot Encoding）转换为数值型特征，最终得到28个有效特征变量。为保证模型训练的有效性，采用分层抽样法将数据集按7：3的比例划分为训练集与测试集，其中训练集用于模型构建，测试集用于性能验证[3]。

（二）特征工程设计

特征工程是提升模型识别性能的关键，本文从核心特征筛选与特征衍生两个方面进行设计。核心特征筛选采用递归特征消除（RFE）算法，结合特征重要性评分，最终筛选出15个关键特征，包括：月均流量消耗、月均消费金额、套餐类型（5G兼容套餐/4G套餐）、终端是否支持5G、上网时段集中度（高峰时段上网占比）、增值服务订阅数量、年龄、收入水平、通话时长、家庭宽带订阅情况、近3个月流量增长率、投诉次数、满意度评分、终端使用年限、是否使用云服务。特征衍生主要基于用户行为的关联性构建新特征，例如："流量消费比"（月均流量消耗/月均消费金额），用于反映用户对流量的重视程度；"5G适配度"（终端5G支持情况 × 套餐5G兼容情况），综合体现用户的硬件与套餐基础；"行为活跃度"（上网频次 × 增值服务使用频次），反映用户对通信服务的依赖程度。通过特征工程，有效提升了特征与目标变量（是否为潜在5G用户）的相关性。

（三）机器学习算法选择与模型构建

本文选取四种经典机器学习算法构建识别模型，并进行性能对比：

逻辑回归（LR）：作为线性分类算法的典型代表，该算法不仅模型结构简洁明了，而且其内在逻辑和运作机制具有极高的可解释性，这使得它在初步的分类预测任务中表现出色。具体而言，该算法通过巧妙地运用Sigmoid函数，将原始的预测结果进行有效转换，使其能够被映射到[0，1]的连续区间内。这一映射过程不仅使得结果更加直观和易于理解，还为进一步的概率判断提供了坚实基础。基于此，我们可以精准地判断每一位用户成为潜在5G用户的概率大小，从而为后续的市场策略和用户服务提供有力的数据支持。

支持向量机（SVM）：支持向量机（SVM）是一种强大的机器学习算法，它通过寻找最优分离超平面来实现数据的分类。这种算法不仅可以处理线性可分的数据，还可以通过采用径向基函数（RBF）核函数来处理非线性特征，这使得SVM能够有效地应对高维数据。在SVM中，最优分离超平面的寻找是通过最大化不同类别数据点之间的间隔来实现的，这样可以确保模型具有良好的泛化能力。

随机森林（RF）：基于多棵决策树构建的集成学习算法，通过采用Bootstrap抽样技术以及特征的随机选择策略，有效降低了模型过拟合的风险。具体而言，Bootstrap抽样能够在每次构建决策树时从原始数据集中随机抽取样本，形成多个不同的训练子集，从而增加模型的泛化能力。同时，特征的随机选择则确保了每棵决策树在分裂节点时仅考虑部分特征，进一步减少了模型对特定特征的过度依赖。此外，该算法还利用袋外数据（Out–of–Bag，OOB）进行特征重要性的评估，袋外数据是指在Bootstrap抽样过程中未被选中的样本，这些样本可以用来对模型进行无偏估计，从而更准确地衡量各个特征对模型预测结果的贡献程度。综合以上机制，该集成学习模型的稳定性得到了显著增强，能够在复杂多变的数据环境中保持较高的预测准确性和鲁棒性[4]。

梯度提升决策树（GBDT）：采用迭代提升（Boosting）策略，每棵决策树都以前一棵决策树产生的残差作为训练基础，通过逐步迭代的方式，利用梯度下降算法不断优化损失函数，从而逐步提升模型的预测精度。这种方法不仅能够有效地减少误差，还具有较强的非线性拟合能力，能够深入捕捉特征之间的复杂交互关系，从而在处理高维数据和复杂问题时表现出色，显著提升模型的泛化能力和预测效果。

模型训练过程中，采用5折交叉验证法避免过拟合，通过网格搜索（Grid Search）优化各算法的超参数：例如，RF的决策树数量设置为100，最大树深度为10；GBDT的学习率设置为0.1，决策树数量为150。以准确率（Accuracy）、召回率（Recall）、精确率（Precision）与F1分数作为模型性能评价指标，其中F1分数综合考虑精确率与召回率，计算公式为：F1=2×（Precision×Recall）/（Precision+Recall）。

三、实证分析与结果讨论

（一）数据样本描述性统计

本次研究的10万条样本中，已转化为5G用户的样本有2.3万条，潜在5G用户（未转化但具有转化潜力）样本通过专家标注确定为1.8万条，非潜在用户样本5.9万条。样本的描述性统计结果显示：潜在5G用户的月均流量消耗均值为28.6GB，显著高于非潜在用户的12.3GB；68.2%的潜在用户使用5G兼容终端，而非潜在用户中这一比例仅为15.7%；潜在用户的月均消费金额均值为128元，高于非潜在用户的85元；年龄分布上，25–45岁的潜在用户占比达到72.3%，该群体是5G服务的核心目标客群。

（二）模型性能对比与分析

基于测试集数据，四种算法的性能评价结果如下表所示：

表1　不同机器学习算法的5G潜在用户识别性能对比

准确率（%）	精确率（%）	召回率（%）	F1分数（%）	准确率（%）
81.2	79.5	78.3	78.9	81.2
85.6	83.2	82.1	82.6	85.6
87.8	86.4	85.7	86.0	87.8
89.7	88.9	88.3	89.0	89.7

从表中结果可以看出，GBDT算法的各项性能指标均最优，准确率达到89.7%，F1分数为89.0%，显著优于其他三种算法。这是因为GBDT能够有效捕捉用户特征间的非线性关系与交互效应，例如“5G适配度”与“流量消费比”的组合特征对潜在用户识别的影响；随机森林作为集成算法，性能次之，其优势在于模型的稳定性与抗过拟合能力；支持向量机在处理高维特征时表现较好，但对参数调节较为敏感；逻辑回归由于受线性假设限制，性能相对较差，但模型解释性较强，可用于初步筛选关键影响因素。

（三）模型应用价值与优化方向

本文构建的GBDT识别模型具有较高的实用价值，能够为运营商提供精准的潜在用户名单，降低营销成本：例如，基于模型识别结果，运营商可针对高潜力用户推出定制化5G套餐，结合终端以旧换新补贴、增值服务赠送等营销手段，提升转化效率；对于中等潜力用户，可通过推送5G应用体验活动、网络提速服务等方式，培养用户的使用习惯。此外，模型筛选出的关键特征可为运营商的产品设计提供参考，例如增加大流量套餐的供给、优化5G终端的推广策略等。模型的优化方向主要包括三个方面：一是特征维度拓展，未来可纳入用户的社交关系数据、地理位置数据等，进一步提升模型的识别精度；二是算法融合，考虑将GBDT与神经网络结合，构建混合模型，充分发挥深度学习在特征提取方面的优势；三是实时更新机制，随着5G技术的普及与用户行为的变化，需建立模型实时更新机制，通过增量学习适应数据分布的动态变化[5]。

结语

本文围绕5G潜在用户识别问题，构建了基于多维度特征与机器学习算法的识别体系。通过数据预处理、特征工程与算法优化，最终确定GBDT算法为最优识别模型，其准确率与F1分数分别达到89.7%与89.0%，能够有效实现潜在5G用户的精准挖掘。研究结果表明，用户的流量消耗、终端设备、套餐类型、消费能力等特征是影响5G转化的关键因素，而机器学习技术能够高效捕捉这些特征与用户转化意愿的内在关联。本文的创新点在于：一是构建了涵盖用户基本信息、消费特征、网络行为、终端设备与服务反馈的多维度特征体系，并通过特征衍生提升了模型的识别能力；二是系统对比了四种经典机器学习算法的性能，验证了GBDT算法在5G潜在用户识别中的优越性。研究成果不仅为运营商的市场拓展提供了技术支撑，也为通信行业的用户识别研究提供了新的思路。

参考文献

[1]洪晓晴，潘珈，栾瑶瑶，等.基于机器学习对5G潜在客户分析与挖掘[J].数据挖掘，2023，13（2）：173-184.

[2]慎于蓝，盛小宁，谢瑞阳，等.基于大小模型和PCC的精准营销智能推广产品研究[J].广东通信技术，2025，45（10）：6-10.

[3]张溶芳，许丹丹，王元光，等.机器学习在物联网虚假用户识别中的运用[J].电信科学，2019，35（7）：136-144.

[4]朱壮军，秦晓飞.一种5G潜在客户识别模型的构建方法[J].电子世界，2020，（14）：81-82.

[5]李汀，徐子恒，李飞.基于量子机器学习的无线宽带信号检测方案[J].信号处理，2023，39（7）：1299-1308.