基于机器学习的消费者行为预测模型构建与应用研究

任怡彤

西安外国语大学, 陕西 西安 710100

摘要: 在数字营销加速迭代的背景下, 精准预测消费者行为成为企业提升决策效率的关键路径。机器学习因其对非线性、多维度变量的高效建模能力, 在行为预测领域展现出广泛潜力。本文基于真实平台数据构建改进型 LightGBM 模型, 结合特征工程与交叉验证, 提升对消费者购买意图的预测准确度。实验表明, 该模型在 AUC、F1 值、召回率等关键指标上显著优于 SVM、随机森林等常用算法, 尤其在不平衡数据结构下仍保持良好的泛化性能。研究还通过对模型预测结果的分类可视化, 揭示了用户忠诚度与行为特征之间的隐含关系。结果显示, 该模型具备高效预测潜力, 可为企业实现智能推荐、动态再营销与个性化服务提供技术支撑与策略参考。

关键词: 消费者行为; 机器学习; LightGBM; 预测模型

DOI: 10.63887/fem.2025.1.4.18

消费者的购买决策不再是单一的线性反应, 而是信息过载、触点多元、反馈加速环境下的复杂博弈。面对这一变局, 传统营销逻辑中的用户细分、静态标签体系难以捕捉个体行为的真实轨迹。相比之下, 机器学习能够通过算法自适应识别消费信号、行为模式与触发机制, 实现动态的精准预测。尤其是在数据维度高、样本结构不平衡的实际场景中, 轻量级模型如 LightGBM, 因其训练效率与预测精度的平衡性, 正在成为商业智能中的关键工具。如何构建兼具性能优势、又适应复杂消费情境的行为预测模型, 已成为人工智能赋能数字营销的现实命题。

1 消费者行为预测的理论基础与技术演进

1.1 消费者行为的数字画像构建逻辑

在数字经济环境下, 消费者的行为已不再是模糊不可量化的心理过程, 而是通过平台交互留下的大量可追踪行为数据所构成的连续序列。这些行为数据包括页面浏览、停留时长、点击频率、加购动作、支付行为、评价反馈、社交转发等, 均构成了用户在不同触点上的偏好与决策倾向。通过对这些数据的提取与归类, 可以识别出典型的行为特征模式, 并据此对用户进行有效划分, 如忠诚用户、高意向潜在客户、犹豫型用户与潜在流失用户等。与传统依赖问卷或访谈的用户研究不同, 数字画像构建强调以客观数据轨迹为基

础, 以行为连续性与反馈结果为分析依据, 从而形成动态的用户标签系统。这类标签并非人为赋予, 而是通过数据驱动自动识别行为之间的关联逻辑, 反映出用户在真实消费场景中的偏好转移与意图演化过程。数字画像不仅是静态属性的集合, 更是对个体消费路径、兴趣变化与心理节奏的结构化还原, 为后续建模提供了扎实的输入变量支撑^[1]。

1.2 消费者行为预测模型研究现状

消费者行为预测起初主要依赖于线性统计模型, 如逻辑回归 (Logistic Regression)、判别分析 (Discriminant Analysis) 等方法。这类模型在解释变量关系和模型可读性方面具有一定优势, 但在面对高维度特征、变量间非线性关联以及数据不平衡问题时, 往往表现出预测精度不足和泛化能力弱的局限。随着数据体量的剧增与用户行为复杂性的上升, 机器学习方法逐渐成为行为预测的核心工具。常用模型包括支持向量机 (SVM)、随机森林 (RF)、梯度提升树 (GBDT) 等, 它们通过多变量联合建模、非线性拟合和特征自动选择机制, 有效提升了行为预测的覆盖度与敏感性。在此基础上, LightGBM 作为微软提出的一种新型梯度提升框架, 采用基于直方图的决策树算法, 不仅训练速度快、内存占用低, 还在处理类别特征、高维稀疏数据和不平衡样本时展现出显著性能优势。尤其是在

面向复杂消费场景的实际应用中, LightGBM 能够兼顾预测效率与准确率, 成为当前行为预测模型中具有高度适配性的算法工具^[2]。

2 数据获取与预处理机制设计

2.1 数据来源与样本构成说明

本研究所采用的数据集来自 UCI Machine Learning Repository 中的 Online Retail 数据集, 辅以某电商平台近一年用户交互日志作为补充, 涵盖网页浏览、商品点击、加购行为、支付记录、商品评价等完整行为链条。原始样本总量为 10000 条, 其中包含 1520 条购买行为 (正类) 与 8480 条非购买行为 (负类), 正负样本比例约为 1:5, 典型呈现出不平衡特征。每条记录对应一个独立会话, 字段类型包括连续变量 (如停留时间、加购数量)、类别变量 (如商品类型、用户等级)、时间序列变量 (如最后操作时间) 及行为标签 (是否完成购买)。此外, 部分数据记录存在缺失字段或异常值 (如停留时长为 0、支付状态与点击行为不一致等), 在建模前需进行统一校验与预处理。整体样本数据具备多维度、高复杂性与明显非线性特征, 适用于构建以分类为目标的消费者行为预测模型, 也为后续特征选择与模型优化提供了足够的输入维度与行为信息支撑。

2.2 特征工程与数据清洗策略

为确保模型训练数据的有效性与预测结果的稳定性, 必须对原始数据进行系统性清洗与特征工程处理^[3]。首先在清洗环节, 剔除缺失值比例超过 50% 的字段, 统一时间戳格式, 并删除行为逻辑冲突样本 (如支付时间早于点击时间)。连续变量中的异常值以 IQR 法 (四分位距) 筛查, 对偏离分布尾部的数据点进行修正或剔除; 类别变量统一采用标签编码 (Label Encoding), 保留语义信息, 避免 One-hot 编码带来的维度膨胀。其次, 构造阶段性衍生特征, 以增强模型对用户消费倾向的感知能力。例如, 从原始“浏览商品时长”与“点击次数”生成“平均点击停留时间”, 从“加购时间”和“支付时间”构造“决策延迟指标”, 这类时序变量可捕捉用户犹豫度与转化效率。为提升模型的训练效率与解释性, 使用信息增益和皮尔逊相关系数双重筛

选机制, 剔除冗余度高、贡献度低的特征字段。最终保留约 22 个核心特征作为输入变量, 涵盖用户行为强度、行为顺序、操作频率和转化路径等多个维度。

3 基于 LightGBM 的消费者行为预测模型构建

3.1 模型原理与算法结构优化

为有效应对消费者行为数据维度高、非线性强以及正负样本比例失衡等问题, 本研究选用 LightGBM 作为核心建模工具。该算法属于梯度提升树 (Gradient Boosting Decision Tree) 家族, 采用基于直方图的决策树构建策略, 不依赖精确数值划分, 而是将特征值离散化为区间后再进行分裂, 显著提升了训练效率与内存利用率。与传统 GBDT 按层分裂不同, LightGBM 基于“叶子优先”策略进行节点扩展, 可快速捕捉到信息增益较大的结构区域, 从而提高模型拟合能力^[4]。

针对原始 LightGBM 在不平衡数据下易发生预测偏倚的问题, 本研究进一步引入双层建模结构, 在第一层中构建初级 LightGBM 模型并输出其叶子索引作为新特征, 随后将这些特征输入至第二层 LightGBM 进行最终分类判断。该方式相当于对样本进行了非线性特征重构, 可提升模型对复杂行为路径的识别能力。同时借鉴随机森林的 Bagging 思想, 在第一层训练阶段引入子样本抽样与特征扰动机制, 增强模型的稳定性与鲁棒性。实验证明, 该优化结构在提升模型表达能力的同时, 有效缓解了训练过程中的过拟合风险, 为后续的精细化调参与预测评估奠定基础。

3.2 模型训练、调参与防过拟合机制

在模型训练阶段, 本研究将数据集按 7:3 比例划分为训练集与测试集, 并采用五折交叉验证以提高结果的稳定性与泛化能力。训练过程采用逻辑损失函数 (log loss) 作为优化目标, 基于 LightGBM 原生接口进行参数调优, 重点围绕模型复杂度控制、学习率平衡与不平衡样本调整三方面展开。

在参数设置上, num_leaves 用于控制树的最大叶子数, 本研究在 64 至 256 间进行网格搜索; max_depth 用于限制树深度, 以防模型对少数样本过拟合, 最终设置为 7; learning_rate 取值范围为 0.01 至 0.1 之间, 经多轮验证确定为 0.05; 同时将 min_data_in_leaf 设定

为 100，以避免过小叶子节点带来的高方差问题。在类别不平衡调整方面，引入了 `scale_pos_weight` 参数，按负样本与正样本数量比设置为 5.57，以强化正类预测

能力。此外，启用 `early_stopping_rounds` 策略，在验证集 AUC 值连续 30 轮无提升时提前停止训练，有效抑制过拟合。训练过程的主要参数调优结果如表 1 所示。

表 1 LightGBM 核心参数调优结果汇总

参数名称	取值范围	最终确定值	调优依据
<code>num_leaves</code>	64-256	128	平衡模型复杂度与训练速度
<code>max_depth</code>	5-10	7	控制树深，避免过拟合
<code>learning_rate</code>	0.01-0.1	0.05	提升精度同时稳定收敛速度
<code>min_data_in_leaf</code>	50-150	100	控制叶子节点最小样本量
<code>scale_pos_weight</code>	1-6	5.57	正负样本不平衡比
<code>early_stopping_rounds</code>	10-50	30	防止训练过度

3.3 模型评估指标体系构建

为全面衡量模型性能，本研究构建了包含 AUC 值、F1 值、召回率、精确率及 KS 值在内的多维评估体系，并将改进型 LightGBM 与 SVM、随机森林 (RF)、梯度提升树 (GBDT) 三类主流算法进行对比测试。在 AUC 指标上，LightGBM 模型取得 0.898 的得分，明显优于 GBDT (0.855)、RF (0.848) 和 SVM (0.808)，显示其在正负样本排序能力上的优势；F1 值为 0.764，在保持召回率 (0.808) 较高的同时维持合理的精确率 (0.725)，表明其对正类识别能力强，且误判率可控。在样本不平衡的测试集中，LightGBM 依然保持最高的 KS 值 (0.710)，反映出其在区分正负样本时的稳定性最强。相比之下，SVM 在多个指标上表现出明显劣势，尤其在处理高维稀疏数据时计算效率低、泛化能力弱；而 RF 和 GBDT 虽具有一定分类能力，但在预测延迟、内存占用方面仍不及 LightGBM。在运行效率方面，LightGBM 通过并行化处理显著缩短训练时间，仅为 3.1 秒，优于 RF (4.2 秒)、GBDT (7.7 秒) 与 SVM (71.5 秒)，同时其内存占用亦最低。综合各项指标结果，LightGBM 在本次消费者行为预测任务中展现出较

强的准确性、稳定性与部署效率，具备可推广性和工业应用价值。

4 实证研究与结果分析

4.1 多模型对比实验与指标表现

为验证所构建消费者行为预测模型的有效性，本研究选取改进型 LightGBM 与 SVM、随机森林 (RF)、梯度提升树 (GBDT) 三种主流机器学习算法进行对比实验。模型在统一训练集与测试集下，分别计算 AUC 值、F1 值、召回率、精确率和 KS 值等关键分类指标，结果如表 2 所示。整体来看，LightGBM 在各项性能指标上均表现优越，其中 AUC 值达到 0.898，明显高于 GBDT (0.855)、RF (0.848) 和 SVM (0.808)，表明其在正负样本排序准确性方面具有显著优势。F1 值为 0.764，在保持较高召回率 (0.808) 的同时维持了良好的精确率 (0.725)，显示其在处理不平衡样本结构下具备较强的分类适配能力。KS 值作为模型区分正负样本的能力衡量指标，LightGBM 同样取得最高值 (0.710)，优于其他模型的最大值 (GBDT 的 0.683)，反映出该模型在识别用户购买意愿方面具有更高的稳定性和可靠性。

表 2 四类模型性能评估指标对比表

模型名称	AUC 值	F1 值	召回率	精确率	KS 值	运行时间 (s)
LightGBM	0.898	0.764	0.808	0.725	0.710	3.1
GBDT	0.855	0.683	0.747	0.643	0.683	7.7
随机森林 RF	0.848	0.691	0.772	0.615	0.643	4.2
支持向量机 SVM	0.808	0.606	0.711	0.542	0.542	71.5

从运行效率来看,LightGBM 通过优化的直方图算法与并行计算框架,在训练时间上远优于其他模型,特别是在处理大样本时显著缩短训练周期,仅需 3.1 秒,相较于 SVM 的 71.5 秒、GBDT 的 7.7 秒具有更强的部署适应性。综上结果表明,改进型 LightGBM 不仅在预测准确性与样本区分性方面具有明显优势,同时兼具高效的运行性能,为后续的实时行为预测与在线系统部署提供了可行性保障^[5]。

4.2 预测结果分析与用户分类案例

在模型训练完成并经指标验证具备良好性能的基础上,本研究进一步对 LightGBM 模型的预测输出进

表 3 用户正负类预测概率与行为类型对照表

用户编号	实际标签	预测概率值	判断结果	行为类型
用户 1	正类	0.81	正确预测	忠诚用户
用户 2	正类	0.68	正确预测	潜在用户
用户 3	正类	0.55	正确预测	临界用户
用户 4	负类	0.52	预测偏差	临界用户
用户 5	负类	0.37	正确预测	非意向用户
用户 6	负类	0.32	正确预测	非意向用户

由表中数据可见,用户 1 预测为正类的概率为 0.81,且与其实际正类标签一致,模型置信度极高,说明该类用户在行为上通常表现为停留时间长、购买链路短、重复访问率高的典型忠诚型;用户 2 与用户 3 的预测概率处于 0.55 至 0.75 之间,归为潜在与临界型用户,其行为特征呈现为高频浏览但低转化、偏好对比但缺乏行动等模式。用户 4 虽为负类,却被预测为正类,概率值为 0.52,属于模型“模糊边界”预测范围,其行为或存在浏览引导性,但缺乏实际购买驱动。整体来看,LightGBM 不仅能准确预测用户是否有购买意愿,更具备对用户行为意向进行微观识别与分

行行为分层与用户分类分析。模型输出的预测概率可理解为用户被归为“有购买意愿”(正类)的置信度值。当该值高于 0.5 时,即被判断为正类用户,反之为负类。在测试集中,通过分析预测结果及实际标签的吻合情况,发现正类样本中准确率达到 81.2%,尤其在置信度值高于 0.7 的用户群体中,模型判断稳定性更高。为进一步细化用户类型识别,本研究将正类样本按预测概率分为三个层级:置信度 ≥ 0.75 定义为忠诚用户,介于 0.55 至 0.75 之间定义为潜在用户,0.5 至 0.55 之间则视为行为模糊、转化风险较高的临界用户。具体结果如表 3 所示。

类的能力,有助于后续精准运营的策略制定。

5 结语

消费者行为预测的核心在于对数据背后心理意图与行为链条的精准捕捉。通过引入改进型 LightGBM 模型,不仅提升了分类效率与预测精度,更在行为标签细分与用户类型识别上实现了显著突破。模型的高可解释性与运行效率,使其具备在实时系统中部署的可行性。未来,若进一步引入多模态数据源与在线学习机制,有望打通“认知—判断—干预”闭环路径,推动从静态预测走向动态运营,实现更深层次的用户智能洞察。

参考文献

- [1] 李宝库,孟思延. 社交媒体平台消费者行为预测研究——基于优化 LightGBM 算法[J]. 长春大学学报,2025,35(07):26-30.
- [2] 高子钦. 基于大数据分析的消费者行为研究——以电商平台为例[J]. 中国商论,2025,34(11):138-141.
- [3] 刘屹枫. 大数据时代市场营销中的消费者行为预测模型[J]. 数字经济,2024,(11):77-79.
- [4] 陶婉琼,王子卓,王健. 基于消费者介入理论的在线购买行为影响要素及预测研究[J]. 管理工程学

报, 2025, 39(02): 144-161.

[5] 汪蕾, 杨一恺, 郑杰慧, 等. 基于消费者神经科学视角预测消费者行为: 现状、挑战与未来[J]. 管理工程学报, 2020, 34(06): 1-12.