机器学习与统计融合的客户行为分析

冯好昌

天津杰纳医药科技发展有限公司，天津市河东区，300252；

摘要：客户行为具有高度异质性、动态性和随机性，传统统计模型难以完整刻画其潜在结构。本文基于概率论与数理统计的逻辑框架，融合机器学习的表达能力与泛化性能，提出一套“统计-机器学习融合分析体系”，以实现对客户行为的系统建模与策略支撑。在方法设计上，强调变量构造的可解释性、模型估计的稳定性及结果应用的可执行性。在实践层面，选取零售电商场景进行案例验证，涵盖数据整理、特征抽取、模型构建与策略输出全过程。该融合方法具备良好的适应性与推广潜力，能有效提升行为识别精度与业务响应效率，为数据驱动下的客户经营提供可靠支撑。

关键词：客户行为建模；概率论；机器学习；融合分析；生命周期价值

DOI：10.64216/3104-9672.25.02.023

引言

客户行为分析是指对客户在与某品牌相关的所有自营线上渠道接触点上产生的全渠道、全场景行为进行埋点、收集、存储与分析，通过深度还原客户使用场景、浏览细节及操作路径等来掌握客户线上轨迹旅程中客户与产品或服务之间产生的全部联系并进行量化分析，从而能更好地制定业务决策、实现精细化客户运营、指导业务增长[1]。在技术驱动与业务融合日益紧密的背景下，单一分析手段难以完整揭示客户行为的生成机制与演化路径。为提升行为建模的准确性与策略输出的操作性，亟需构建融合统计推断逻辑与机器学习算法的协同框架，实现从行为表征到决策干预的系统化建模。本文围绕“机器学习与统计融合的客户行为分析”主题，提出一套可复用、可落地的集成化方法体系，兼顾预测精度、因果识别与业务适配能力。

1 理论基础：概率论与数理统计视角

1.1 概率模型与客户行为建模

客户行为具有随机性与不确定性，适合以概率模型予以刻画。购买频次可拟合泊松或负二项分布，行为转移可映射为马尔可夫链，生命周期价值则可建模为随机过程。在实际应用中，用户点击、浏览、下单等行为往往呈现异质性，传统模型难以穷尽其变化。概率论提供了一套严谨的逻辑框架，有助于构建符合现实分布特征的行为模型，同时为后续推断与预测打下基础。不同模型选取应依据行为属性、时间依赖性与变量之间的结构关系具体判定。

1.2 参数估计与假设检验在客户行为分析中的作用

在客户行为模型中，参数的合理估计直接决定了模型的解释力与应用效能。采用最大似然估计、贝叶斯估计等手段，不仅可以得出行为指标的数值描述，还能在模型稳定性与可靠性层面提供理论支撑。通过假设检验判断变量间的行为差异，如不同地区客户的响应行为是否存在统计显著性。这类推断过程对于精准营销、人群细分等策略具有指导价值，也强化了模型在实际场景中的应用可信度。

1.3 机器学习算法与统计视角的衔接

机器学习算法强调预测精度，而统计方法关注建模逻辑与推断依据。两者虽路径不同，却在客户行为分析中可形成互补。许多分类与回归算法本质上可转化为概率模型的近似表示，例如逻辑回归对应伯努利分布的条件概率建模[2]。从统计角度解析机器学习结果，有助于识别模型偏差、控制误差传播、评估样本代表性。统计推理能力的引入，使机器学习从“有效”走向“可信”，从而提升整体模型系统的解释力。

1.4 融合分析的必要性与逻辑链条

客户行为数据呈现高维、动态与非线性特征，单一建模技术难以兼顾预测精度与结构解释。融合分析以统计推断为基底，引入机器学习进行特征构造与模式识别，在模型建构层面实现优势互补。整体流程包括数据刻画、变量抽取、模型训练、结果检验与策略输出。每一环节

均嵌入概率逻辑与估计原则，保障分析的严密性。融合方法不仅提供了高效的建模工具，更在决策支持中体现出灵活性与可控性，成为客户行为研究的主流方向[3]。

2 融合框架设计：机器学习和统计方法协同分析客户行为

2.1 框架总览

为提升客户行为分析的系统性与可执行性，构建“统计-机器学习融合框架”（Stat-ML Fusion Framework, SMF），聚焦预测性能、因果解释与运营落地的协同统一。该框架由五个环节构成：数据治理、特征工程、模型构建、结果解释和策略反馈。前两者确保输入规范且具备统计表达力；模型阶段兼顾误差控制和估计稳定性；解释机制连接算法输出与人群洞察；策略环节则基于预测信息生成行动路径。整个流程嵌入风险控制、漂移检测与隐私合规机制，并辅以 AUC、F1、Lift、ROI 等双重指标体系评价。建议配套建设特征仓与实验平台，统一特征口径，提升策略测试与迭代效率。

2.2 数据准备与描述性统计

客户行为数据分布于交易、日志、客服等系统，需构建跨系统主键与事件时间线，确保数据完整性。样本切分按时间滑窗推进，避免信息泄露。缺失值依其产生机制分类处理，对非随机缺失保留指示变量。初步统计聚焦分布形态与变量质量：浏览频次、下单次数等可尝试泊松或负二项分布拟合；金额型指标常呈对数正态；序列类行为则计算停留时间、路径深度等会话级指标。人群以渠道、地区、设备分组，标记行为差异明显的特征，形成基础切片结构，为后续建模提供分层依据[4]。

2.3 特征工程：统计与机器学习的协同构造

特征构造承载模型解释力与表达力双重任务。统计路径强调业务逻辑可解释性，如转化率、RFM 指标、生存概率、行为间隔、风险函数等；算法路径关注特征的非线性刻画与交互扩展，如图网络指标、序列嵌入、目标编码、自动交叉生成等。二者通过稳定性检验对接：利用交叉验证扰动评分观察变量在不同数据划分下的重要性一致性，结合 Bootstrap 或稳定选择方法提供置信区间。所有特征入库前需明确版本、生成逻辑与口径注释，避免口径漂移，提升复用能力。

2.4 模型构建与估计

客户行为建模任务包括分类（如流失预测）、回归（如 CLV 预估）和排序（如推荐排序）。在模型选择上，针对任务复杂性和数据规模选择梯度提升树、正则化广义线性模型（Generalized Linear Model, GLM）、Cox 比例风险模型或轻量神经网络。参数调优采用分层贝叶斯优化或网格-随机混合策略，评估维度涵盖 AUC、召回、校准误差、Lift 等。模型结果通过局部依赖图、SHAP 值和个体响应曲线增强可解释性，配合反事实检验识别变量因果贡献。上线前需完成鲁棒性测试与合规模块审核，构建模型生命周期闭环[5]。

2.5 行为洞察与策略支持

分析结果需回归业务，形成可执行策略。模型输出映射为干预名单与行动节点，再结合预算与资源进行优化分发。策略优先级根据预测风险、历史反应率及行为活跃度动态调整。干预效果评估设计双重路径：实验平台运行 A/B 测试，辅助以倾向得分调整法实现准因果推断。策略报告结构遵循三层逻辑：整体画像勾勒生命周期轨迹，驱动因子提取核心变量，执行剧本细化触达渠道、频率与优惠方案。投放后建立监控面板，实时追踪响应率与核心业务指标，触发异常即回滚或优化。

2.6 框架优势与适用条件

该框架优势在于：其一，融合预测与推断，确保模型具备前瞻性与可信度；其二，从变量到策略层层落地，增强分析的执行力；其三，配套工程治理体系，保障模型长期可维护性。适用条件包括：拥有完整的事件级链路、跨期数据结构、统一标签口径；具备稳定的算力资源与在线推理能力；企业内部建立跨部门口径审核与模型审查机制。在资源受限或数据初期阶段，可构建轻量版本，聚焦规则增强和统计分层。电商、通信、金融与出行等行业尤为适配，可先落地价值模型和流失模型，再逐步拓展至推荐与反欺诈模块，构建系统化客户经营底座。

3 案例说明与数据构思

3.1 案例背景

以某全国性零售电商 2022 至 2024 年间的客户经营为背景，探索“统计-机器学习融合”分析体系在真实业务中的落地可行性。企业目标聚焦三方面：遏制客户流失、提升客户生命周期价值（Customer Lifetime Value, CLV）、合理分配营销资源。数据覆盖交易、日志、

营销触达、客服及退换货等多个子系统，采用秒级事件标记，确保行为链精确追踪。在合规要求下，用户标识采取不可逆加密，跨表关联依赖安全主键拼接。该案例聚焦融合分析的实用性，采用结构对齐的合成样本进行展示，模拟常见数据形态和变量关系，便于分析过程说明。

3.2 数据准备与初步统计

样本构建以时间滑窗方式划分：训练集覆盖 2022Q1 至 2023Q4，验证集和测试集分别为 2024 年上、下半年，确保标签生成过程不穿越时间线。数据清洗遵循缺失机制分类，保留条件性与非随机缺失的结构信息；行为事件唯一性由用户标识、商品编码和时间戳联合定义。初步统计结果揭示行为变量的非对称特性：浏览与订单频次呈现长尾，金额类指标近似对数分布，行为序列聚合后形成明显的路径深浅差异。表 1 提供关键字段示例，用于展示数据维度与典型指标表现（仅为方法展示所用，非真实业务数据）。

表 1 各项指标示例

指标	口径说明	示例统计值
样本规模	去重活跃用户数（观察期内有任一行为）	200,000
观察期	事件时间窗	2022-01-01 至 2024-12-31
近 30 天客均浏览	页面曝光次数/活跃用户	18.6 次
近 30 天下单次数	订单笔数/活跃用户	0.85 次
客单价中位数	订单金额的中位数	168 元
平均购买间隔	相邻下单的日间隔均值	54.2 天
近 180 天活跃率	近 180 天内有下单或支付	61.3%
优惠券使用率	至少使用过 1 张券的占比	27.4%
营销触达覆盖	收到任一触达（短信/推送/邮件）占比	73.5%
退货率	含至少 1 笔退货的用户占比	8.9%
近 90 天流失率	90 天无下单定义为流失	24.7%
客服工单率	产生过工单的用户占比	3.1%
缺失比例	关键字段缺失占比（加权）	1.5%
异常订单占比	可疑重复/超大额/秒退单	0.7%

说明：合成样例仅用于变量展示，真实研究需接入企业自有数据，并按合规要求脱敏。

3.3 特征工程设计（统计＋机器学习）

特征构造以业务可解释性和模型表达能力为核心，构建双轨并行体系。统计路径提取 RFM 指标、行为间隔、生存概率、转化链条、品类覆盖度、路径长度及加购转化概率；机器学习路径生成目标编码、交叉特征、行为序列嵌入与图结构指标。特征评价基于扰动检验，记录其在不同交叉验证折次下的稳定程度，同时使用 Bootstrap 生成置信界限。对高漂移风险的特征设立阈值机制，监测建模稳定性。所有特征必须在入库前完成注释、版本锁定和脚本哈希注册，保障流程审计与代码复现。

3.4 模型构建与估计说明

根据任务类型，构建三类模型：流失预测（二分类）、CLV 回归、行为优先级排序。基准模型使用逻辑回归兼顾透明性与统计检验能力；主力建模采用 XGBoost 处理变量非线性关系；对时间相关任务，配合 Cox 比例风险模型描述行为发生概率的动态变化。模型评估不仅涵盖 AUC、召回率、精确率等传统指标，还包括校准曲线、分位提升率与投入产出比曲线（ROI 曲线）。解释层引入 SHAP 值与局部依赖分析（Partial Dependence），结合反事实扰动检验，强化变量与策略路径间的因果假设支撑。模型上线分阶段推送，控制预算消耗与误投风险，建构灰度流量＋规则闸门双轨保障机制。

3.5 行为洞察与策略建议

模型输出转化为三类洞察路径：其一，行为趋势类，如“最近访问频繁但未下单”的用户被标记为潜在流失边缘人群；其二，偏好型变量，如“低品类广度＋高折扣响应率”暗示促销敏感型客户；其三，触点价值判断，如小程序适用于低客单价回收，APP 更适合深度经营。策略建议围绕两套剧本设计：激活剧本以轻权益唤醒潜在流失客户，价值剧本针对高潜力客户定制差异化推荐与会员阶梯权益。效果评估引入 A/B 试验与倾向得分配重双通路验证，控制策略干预中的样本偏差，提升结论

外推力。

3.6 案例反思：融合方法的实施难点

从执行角度来看，融合框架在四个方面面临挑战：数据端多源整合不稳定，主键漂移影响标签精度；建模阶段变量交互复杂，需严格防控过拟合与冗余噪声；模型输出的可解释性依赖统计方法兜底，若缺乏统计审校机制，洞察无法转化为实际业务启发；运营层策略干预频繁更迭，若无标准化特征仓与实验体系支撑，将导致策略复现困难、模型失效周期缩短。应构建治理结构，包括设立口径审查机制、制定策略回滚预案，并建立模型风险清单，实现分析闭环的制度化与常态化。

4 结语

本文从理论与实践双重视角，系统论证了机器学习与统计融合在客户行为分析中的必要性与可行性。融合框架（SMF）以统计推断保证模型稳健，以算法学习提升预测表现，并在数据治理与策略执行层实现闭环。案例结果显示，该体系能有效识别高风险客户群体，优化资源配置，提高营销回报率。未来研究可在三个方向拓展：其一，探索因果推断与深度学习的结合；其二，强化模型可解释性在实时决策系统中的应用；其三，建立行业级特征仓与模型共享机制，推动企业间知识迁移与算法复用。融合视角的持续深化，将为客户智能运营提供更具科学性的分析范式与实践指南。

参考文献

[1]高尚.客户行为分析的框架、价值与应用——基于商业银行视角[J].银行家,2023(4):114-117.

[2]周艳聪,郝园媛.基于机器学习的运营商客户行为分析[J].科学技术与工程,2022,22(2):585-592.

[3]余康琪.基于客户行为挖掘的C2C电商用户复购率预测分析[D].上海财经大学,2021.

[4]孙同舟.基于机器学习的电商用户分类模型研究[D].桂林电子科技大学,2023.

[5]江丽桃,曾晶.数字经济下的电商客户行为分析研究[J].商业观察,2024,10(17):93-95+108.