

大数据统计方法在电子商务精准营销中的应用分析

陈铭佳

湖北汽车工业学院数字经济学院, 湖北 十堰

收稿日期: 2026年7月7日; 录用日期: 2026年7月21日; 发布日期: 2026年9月1日

摘要

面向电商高频互动与多源数据并行场景, 本文构建面向精准营销的大数据统计方法组合, 包括K-means用户细分、基于协同过滤与线性回归的候选重排, 以及ARIMA滚动预测, 用以联动获客、转化与留存。以某平台为例, 方法针对数据口径不一、跨端识别断点与触达节奏失配等痛点落地, 并配套统一的多维评估体系与工程化迭代机制。线上验证表明, 在预算不变的前提下, 转化率由3.2%提升至5.8%, ROI由1.17提升至1.43, 30天留存率提高10个百分点, 复购率提升3.2个百分点, 流量获取成本下降9.6%。研究进一步提出联邦学习与数据清洁室的隐私优先方案及MLOps治理路径, 为电商平台在复杂触点环境下实现策略闭环提供了可复用范式。

关键词

大数据统计方法, 精准营销, K-Means聚类, 协同过滤, ARIMA

Application Analysis of Big Data Statistical Methods in Precision E-Commerce Marketing

Mingjia Chen

School of Digital Economics, Hubei University of Automotive Technology, Shiyan Hubei

Received: July 7, 2026; accepted: July 21, 2026; published: September 1, 2026

Abstract

Against the backdrop of high-frequency interaction and multi-source data parallelism in e-

文章引用: 陈铭佳. 大数据统计方法在电子商务精准营销中的应用分析[J]. 电子商务评论, 2026, 15(9): 54-60.

DOI: 10.12677/ecl.2026.159968

commerce scenarios, this paper constructs a combined framework of big data statistical methods for precision marketing. The framework integrates K-means user segmentation, candidate re-ranking based on collaborative filtering and linear regression, and ARIMA rolling prediction, so as to coordinate customer acquisition, conversion and retention. Taking an e-commerce platform (Yigou) as a case study, the proposed methods are implemented to address pain points such as inconsistent data calibers, cross-terminal identification breakpoints and mismatched contact rhythms, supported by a unified multi-dimensional evaluation system and engineering iteration mechanism. Online verification results show that under the same budget, the conversion rate increases from 3.2% to 5.8%, ROI rises from 1.17 to 1.43, the 30-day retention rate improves by 10 percentage points, the repurchase rate increases by 3.2 percentage points, and the customer acquisition cost decreases by 9.6%. The study further proposes privacy-first schemes of federated learning and data clean rooms, as well as an MLOps governance path, providing a reusable paradigm for e-commerce platforms to achieve strategy closed-loop in complex contact environments.

Keywords

Big Data Statistical Methods, Precision Marketing, K-Means Clustering, Collaborative Filtering, ARIMA

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

广告竞价走高与公域流量分散使电商触达成本持续上扬, 移动内容平台的分流延长了转化链路, 跨端标识缺口与归因失真加剧预算浪费。在此背景下, 数据驱动的精准营销从单点优化转向全链路协同, 亟需将用户细分、推荐排序与需求预测纳入同一决策体系, 并以统一的数据治理与评估口径承载策略迭代。本文以 4C 视角与用户生命周期为理论支撑, 将细分、响应建模与时序预测分别对位于获客激活、转化留存、挽回唤醒环节, 构建方法与目标相一致的闭环框架。以某平台为对象, 围绕数据采集与特征工程、候选生成与重排、节奏预测与资源编排进行系统化实践, 旨在在不增加预算前提下提高内容匹配、触达与时机三方面的协同效率, 并为行业在复杂触点环境中实现稳健增长提供可操作路径[1]。

2. 大数据统计方法与电子商务精准营销的理论基础及现状分析

2.1. 相关概念界定与理论支撑

在电商平台高频互动与多源数据并行的情境下, 大数据统计方法被用来刻画用户行为与市场响应的结构规律。聚类分析依靠相似性度量把多维样本聚合为同质群体, 用来暴露偏好差异与细分边界; 回归分析把促销强度、价格、曝光等自变量与点击或购买结果建立函数关系, 用来提取边际效应; 时间序列分析关注随时间变化的趋势、季节与滞后关联, 用来描述动态演化。电子商务精准营销指在细分人群与供给要素匹配的前提下, 借助数据驱动策略去提高内容、触达与时机的匹配度, 其特性表现为强数据依赖、策略迭代加快以及跨渠道协同。把 4C 理论关于消费者诉求、成本感受、获取便利和沟通互动的视角嵌入模型选择, 可把聚类结果转译为差异化价值主张, 把回归系数量化为成本敏感度, 把时序结构用于安排触达节律。结合用户生命周期理论对获客、激活、留存和挽回的阶段划分, 可形成方法与目标的对应: 在获客与激活阶段侧重细分与相似度衡量, 在留存阶段侧重因果识别与响应建模, 在挽回阶段侧重

周期性预判与趋势识别，从而把数据分析与营销决策关联为闭环[2]。

2.2. 电子商务精准营销的现状与痛点分析——以某平台为例

在广告竞价持续走高与公域流量不断分散的背景下，电子商务平台在获取有效触达时呈现边际成本上扬，移动端内容平台的分流效应进一步把用户转化环节拉长，导致交易闭环形成更加迟滞[3]。从行业运行机制来看，数据采集的断点与跨端 ID 匹配的缺口，使得用户识别链条出现不连贯，归因模型无法对复杂触点进行稳定刻画。结合此行业情境，某平台在 2022~2023 年呈现出典型特征：流量获取成本同比上升 15%，用户转化率在 2023 年仅为 3.2%，广告投放强度加大而 ROI 承压，私域与公域协同不足。需重点关注的是，历史订单、站内行为与短视频外链等多源数据被分散存放在广告投放系统与站内 CDP 之间，标签粒度偏粗，难以把潜在高价值用户与价格敏感用户区分，从而把推送节奏与优惠策略的细分一致性打散[4]。由此推导出两类核心痛点，一是数据治理层面的口径不一与埋点不完整，二是用户定位与触达策略缺乏对行为序列的动态刻画，导致对中长链路商品的引导效率偏低。相关量化表现见表 1，用以呈现成本、转化与留存的结构化变化。

Table 1. Comparison of core marketing data of a platform from 2022 to 2023

表 1. 某平台 2022~2023 年营销核心数据对比表

指标	2022	2023	同比变化
流量获取成本(CNY/新客)	28.0	32.2	+15.0%
广告投放费用(CNY)	120,000,000	138,000,000	+15.0%
用户转化率(%)	3.6	3.2	-11.1%
ROI	1.28	1.17	-8.6%
月活用户(人)	5,200,000	5,600,000	+7.7%
平均访问时长(min)	4.6	4.3	-6.5%
客单价(CNY/订单)	168.0	171.0	+1.8%
复购率(%)	22.0	20.0	-9.1%
自然流量占比(%)	42.0	38.0	-9.5%

3. 大数据统计方法在电子商务精准营销中的具体应用策略

3.1. 用户画像构建中的 K-Means 聚类方法应用

选用 K-means 聚类的原因：电商用户行为数据具备高维度、样本量大以及类别不平衡的特性，并且 K-means 的计算复杂度较低，可解释性较强，且易于开展并行化工作，适宜用来进行初步的用户群体划分。相较于层次聚类和 BSCAN，K-means 在千万级样本中还拥有线性扩展性，并且簇中心能够直接用来定义业务标签[5]。

特征选择以及标准化的过程：针对某平台，所选取的特征包括会话平均浏览时长(min)、近 30 d 购买频次、近 7 d 加购次数、近 90 d 最近访问间隔(d)客单价以及品类多样性也就是购买类目数。鉴于各个特征的量纲存在很大程度上的差异，如浏览时长是以分钟为单位，而购买频次是个位数，所以采用 z 分数标准化：

$$z = \frac{x - \mu}{\sigma}$$

其中将超出均值 ± 3 倍标准差的值进行截尾处理,来消除极端促销活动所产生的噪声。

K 值确定的依据:运用肘部法计算簇内误差平方(SSE),并结合品类运营的实际需求来开展,当 K 从2增加到8的时候,SSE下降的速度在 $K=5$ 时明显变缓了;经与运营团队确认,5个簇对应“高频购买型”“潜在兴趣型”“价格敏感型”“流失风险型”“沉默型”五类,最便于后续策略匹配。最终确定 $K=5$ 。

在用户细分的技术路径中, K -means把样本空间划为 K 个互斥簇,依靠样本分配与质心更新的迭代来最小化簇内平方和,从而把相似用户汇聚到同一簇内。鉴于电商特征量纲差异与偏态分布,研究把数值特征进行 z 分数标准化并进行异常值截尾,选用 k -means++初始化以降低初值敏感性,停止准则设为目标函数下降幅度与迭代上限的联合条件。围绕画像构建链路,数据采集、预处理、聚类分析与画像输出形成闭环[6]。

针对某平台,数据项覆盖会话平均浏览时长 $\min$ 、近30 d购买频次、近7 d加购次数、近90 d最近访问间隔 d 、客单价与品类多样性。预处理中把缺失值进行插补,把类别变量进行独热编码,并把极端值进行截尾处理,以肘部法结合品类运营判断来确定 K 。训练完成后对簇中心进行业务命名,归纳为高频购买型、潜在兴趣型与流失风险型,并把策略嵌入CRM与推荐引擎,高频购买型倾向组合优惠,潜在兴趣型倾向价格提醒,流失风险型倾向唤醒激励,使细分结果落到可执行触达。

$$d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

其中, d 表示样本间的欧氏距离, x 与 y 为两个用户的特征向量, p 为特征维度数, x_i 与 y_i 为在第 i 个维度上的取值。

3.2. 精准推荐系统中的协同过滤与线性回归分析

选用基于用户的协同过滤,主要原因是平台商品更新速度快,基于物品的协同过滤容易受到新品冷启动的影响;而用户行为模式相对稳定,该算法可以输出直观的“购买过该商品的用户也购买了……”解释文案,这有助于提高用户信任。线性回归则作为排序层模型,其参数可解释性较强,便于分析折扣幅度、曝光位权重等特征对于购买概率的边际贡献,适合与业务约束(毛利、库存)联合优化[7]。

协同过滤的关键细节:对隐式评分矩阵进行赋值,即点击赋1分、加购赋3分、购买赋5分进行加权,未交互项不填充。运用皮尔逊相关系数来计算用户相似度时,只保留至少共同交互过5件商品的用户对。借助行为密度筛选,每个用户开展至少10次有效交互,否则归入冷启动组使用均值填充,时间衰减因子采用半衰期为7天的指数衰减函数,近期行为权重更高。

鉴于某平台用户规模与商品种类持续扩张,基于用户的协同过滤被把群体行为相似性当作兴趣近邻来使用,从而在候选生成环节形成可解释的推荐信号。具体做法是把点击、加购、购买等交互按权重转换为隐式评分,依靠皮尔逊相关系数在共同交互集合上计算目标用户与候选近邻的相似度;再把相似度作为权重,对近邻在未互动商品上的偏好偏移进行加权求和,得到用户对该商品的偏好分数,并且结合时间衰减以及品类流行度因子进行校正,用于生成Top-N候选集合。为降低噪声干扰,需要对评分矩阵进行中心化与缺失值掩蔽处理,同时按行为密度筛选 K 个稳定近邻,以维持邻域结构的可学习性[8]。

$$r_{u,v} = \frac{\sum_{i \in I_{u,v}} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_{u,v}} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{i \in I_{u,v}} (R_{v,i} - \bar{R}_v)^2}}$$

其中, $r_{u,v}$ 表示用户 u 与用户 v 的相似度, $I_{u,v}$ 为二者共同交互过的商品集合, $R_{u,i}$ 为用户 u 对商品 i 的评

分或隐式偏好强度, $\bar{R}_u$ 为用户 u 在 $I_{u,v}$ 上的平均评分, $\bar{R}_v$ 为用户 v 在 $I_{u,v}$ 上的平均评分。

从排序建模角度来看, 线性回归被选用来把协同过滤信号与上下文要素进行联合建模, 以估计用户对候选商品的购买概率。以某平台的推荐场景为例, 特征向量包含相似度加权分数、最近 7 d 浏览频次、折扣幅度、类目匹配度、价格区间、曝光位权重、品牌活跃度以及用户近 30 d 成交频次等; 训练阶段借助 L2 正则与分层交叉验证来抑制过拟合, 并且把时序窗口切片用于避免信息泄露; 预测输出经最小最大归一化映射到 0~1 区间, 在线服务把该概率与库存安全阈值以及毛利权重相乘得到最终排序分值, 支撑推荐列表生成与业务约束协同。

3.3. 营销效果预测中的 ARIMA 时间序列分析

选用 ARIMA 的原因: 电商日销量序列通常会有趋势、周期性以及非平稳性等特性, ARIMA 通过差分可有效去除趋势, 且比深度学习模型(如 LSTM)所需要的训练数据要少, 在滚动预测场景中, 计算开销较低, 且可解释性强, 这样便于运营团队理解[9]。

定阶过程: 以某平台 2023 年 3 月 1 日至 6 月 18 日的日销量序列为例子, 首先开展 ADF 检验(即 $p = 0.32$), 接受非平稳原假设; 进行一阶差分后, 开展 ADF 检验($p = 0.01$), 这表明序列是平稳的。根据自相关图(ACF)在一阶差分后 1 阶截尾、偏自相关图(PACF)拖尾, 初步判定 $q = 1$; 再结合 PACF 在滞后 2 阶显著, 尝试 $p = 2$ 。使用 AIC 准则比较 ARIMA (2, 1, 1)与 ARIMA (1, 1, 1)及 ARIMA (2, 1, 0), 最终选择 AIC 最小的 ARIMA (2, 1, 1)。

残差检验结果: 在模型完成拟合之后, 对残差的 Ljung-Box Q 进行检验, 滞后 10 阶的 p 值为 0.23, 没办法拒绝残差属于白噪声这一原假设; 无法拒绝残差为白噪声的原假设; 残差 Q-Q 图基本落在 45°线附近, 表明正态性可接受。上述检验通过, 说明模型已充分提取序列信息[10]。

面向某平台在 618 期间呈现的高峰性与突发性需求波动, 本研究把自回归差分滑动平均模型当作核心预测工具来使用。该模型以自回归项刻画销售序列的惯性, 以差分运算去除趋势并使序列近似平稳, 以滑动平均项吸收随机扰动, 进而在有限历史信息下给出短期滚动预测。参数阶数 p 、 d 、 q 由自相关与偏自相关图谱联合判别, 结合信息准则进行筛选, 在极大似然框架下完成估计和残差诊断, 以降低模型误设带来的漂移风险。

结合某 618 的运营节奏, 构建以日为粒度的销量序列, 优先完成节假日与营销事件标记, 随后对 7 日周期波动选用季节差分, 对异方差序列采用对数变换。继而开展滚动起点的多步预测, 把预热期、活动日与退潮期纳入统一窗口, 并设置层级一致性约束以使单品与品类、总量的预测相互协调。为把误差维持在 5%以内, 模型在训练集上运用交叉时间切片验证, 在线阶段以阈值告警触发参数重估; 所得预测被用于指导仓配前置备货、广告投放节奏与客服排班, 把有限资源在峰值时点进行更高效的配置[11]。

从数据处理角度来看, 为减少营销脉冲对序列结构的扭曲, 选用稳健滤波与异常点回填策略, 并把大额满减、直播上架、站外导流等事件作为干预项编码进入流程。在落地层面, 通过按品类分层建模与需求相干性聚合, 把不同仓网的区域销量同步推断, 配合日内 2 次滚动重算机制, 使运营团队在高峰来临前优化广告出价、转移安全库存以及调整配送时段。

4. 大数据统计方法应用的效果评估与优化建议

4.1. 精准营销效果评估指标体系构建

鉴于电商精准营销以数据驱动策略迭代为核心, 本研究把效果评估体系构建为兼顾交易效率、投入产出与用户稳定度的多维框架。转化率被界定为有效交易相对于有效触达的比例, 口径上把去重用户与有效会话作为匹配统计单元, 并在渠道与活动窗口两个维度开展分层, 用来揭示落地链路的阻滞环节与

内容匹配度。ROI 被定义为营销可归因收益相对于投放成本的比值，核算时以毛利或贡献利润作为基底，剔除优惠补贴与退货冲减影响，同时引入多触点归因以缓解单触点偏差，从而刻画预算分配与投放强度的边际效应。用户留存率围绕注册或首单时点构建同批群组，在日、周、月不同时间标尺上统计继续发生访问或交易的占比，以描绘生命周期迁移并体现内容供给与服务体验的持续吸引力。复购率以首单后设定周期内再次下单为判定规则，结合品类层级与客单价区间进行分层输出，用来揭示产品组合与价格敏感度对持续购买的作用路径。需重点关注的是，上述指标需在统一数据治理规则下完成去重、对齐与异常处理，并把时间切片、渠道来源与人群细分作为三类必选分组，辅以漏斗转化、订单频次与客单价等辅助观测，把模型效果与经营目标对齐为可落地的闭环评估[12]。

4.2. 基于某平台的效果数据分析与验证

结合某平台 2023 年运营现状与方法落地路径，本研究把 K-means 细分接入 CRM，把协同过滤与线性回归构成的排序框架用于候选重排，把 ARIMA 的滚动预测嵌入投放与备货节奏，在获客、转化与留存三个环节构建联动机制。进一步观察发现，细分后的高频购买型人群在组合优惠与毛利权重联动下获得更高排序权重，潜在兴趣型人群在价格提醒与时序曝光控制下形成稳定触达频率，流失风险型人群在唤醒激励与内容收敛策略中降低无效曝光，因而把有限预算更集中地分配到高响应人群。需重点关注的是，归因口径统一到多触点窗口并把时间切片作为评估基准位，减少跨渠道噪声对指标的扰动，同时以 90 d 滚动窗口对新客与老客进行分层核算，以弱化季节与活动脉冲的偏移影响[13]。由此推导出核心改善路径，即在相同预算强度下，依靠候选生成与重排协同把成交概率抬升，依靠节奏预测把峰期资源有序错位投放，从而带来转化率提升至 5.8%，ROI 提升 22%，30 d 留存率提高 10 个百分点。见表 2，应用前后关键指标的对比呈现出方法组合在不同链路上的协同增益与结构性改善效应(数据来源：某平台 2023 年 A/B 测试及内部数据统计)。

Table 2. Comparison of marketing effects before and after the application of big data statistical methods on a certain platform
表 2. 某平台应用大数据统计方法前后营销效果对比表

指标	应用前	应用后	变化幅度
转化率%	3.2	5.8	+81.25%
ROI	1.17	1.43	+22.0%
30 d 留存率%	28.0	38.0	+10.0 个百分点
复购率%	20.0	23.2	+3.2 个百分点
自然流量占比%	38.0	41.5	+3.5 个百分点
客单价 CNY/订单	171.0	174.0	+1.8%
流量获取成本 CNY/新客	32.2	29.1	-9.6%

4.3. 应用过程中的问题与优化方向

鉴于某平台在多源数据并行的运营场景下推进精细化运营，进一步观察发现隐私治理与工程化迭代两端存在制约。隐私侧表现为原始日志在多部门间重复流转，跨端标识依赖可回溯标识符，授权粒度过粗，去标识化与数据生命周期未成体系，从而把合规风险与共享效率的矛盾放大。模型侧体现为特征工程依赖手工脚本，训练上线由临时任务拉动，漂移与标签延迟缺乏监测，策略对波动的响应滞后。需重点关注的是，这两类问题相互叠加，会把全链路协同效果削弱。由此推导出三条优化路径。本研究立足隐私优先的技术路线，选用联邦学习作为跨域联合建模机制，把用户侧特征留在端内或业务域内，只上

送加噪与加密后的梯度聚合结果,辅以内置的安全聚合与差分隐私策略,并配套哈希脱敏与双盲匹配的清洁室方案,用以在受控环境下完成跨端联结与归因校准[14]。围绕工程化迭代,构建覆盖特征存储与模型发布的 MLOps 流水线,即面向机器学习全生命周期的工程化体系,设置输入分布与预测输出的漂移告警,运用冠军挑战者、影子部署与灰度发布把更新风险控制在可接受区间,并在固定节律上进行小步重训与阶段性全量复盘。结合组织协同的现实需求,建立由数据治理、算法、商品与投放组成的联合工作组,以数据契约统一事件命名与取值口径,完善主数据与标识管理及数据血缘追踪,同时把仓配约束与毛利目标前置进入排序目标函数,从制度与流程层面把隐私合规、业务约束与模型优化联动起来[15]。

5. 结语

本文以细分、重排以及时序预测协同为主线,构建起电商精准营销数据方法的闭环,并在某平台验证其对转化、投入产出以及留存方面的综合提升效果。上述结果是基于特定平台 2023 年前后的运营环境取得的,模型效果可能会随着用户结构、竞争环境以及数据采集方式的变化而出现波动。因此在进行结论推广方面要谨慎,不适合直接把结论外推到所有电商场景中。

为了保障能够可持续地迭代,本文章提出了把隐私放在优先位置的数据联结方案以及 MLOps 体系。其中,联邦学习能够有针对性地处理当前研究中跨部门数据无法直接合并的合规瓶颈问题,平台的用户行为日志分属广告系统、推荐系统和 CRM 系统,直接合并涉及用户隐私授权模糊;运用联邦学习,各个系统仅仅上传加密梯度,在不暴露原始日志的状况下来完成联合建模。数据清洁室可以处理跨端 ID 匹配缺口:在受控环境中实现经过哈希脱敏之后的双盲匹配,防止明文 ID 向外泄露,MLOps 流水线可以解决特征工程依赖手工脚本、模型漂移无告警的工程化痛点,设置每日特征分布监控即 PSI 指标以及预测误差滑动窗口,如连续 3 天误差超过 5%,就会自动触发进行模型重训。

未来可以在因果识别和增量响应建模、面向毛利以及库存约束的多目标排序、流式特征和近实时决策等方面来深化工作,提高高波动环境中的稳健性以及可解释性。

参考文献

- [1] 马敬惠. 大数据驱动的精准营销在企业运营中的实践研究[J]. 营销界, 2025(13): 41-43.
- [2] 张钰. 大数据分析在网络营销精准获客中的实践应用与价值评估[J]. 营销界, 2025(13): 62-64.
- [3] 董波. 大数据与人工智能在电商运营中的应用研究[J]. 营销界, 2025(13): 16-18.
- [4] 杨慧丰. 大数据聚类分析算法在电信运营商精细化营销中的运用[J]. 中国信息界, 2025(6): 111-113.
- [5] 吴志玲. 大数据助力数字出版精准营销[J]. 中国报业, 2025(11): 70-71.
- [6] 程天祥, 任小静. 基于用户画像的电商内容营销策略演化研究[J]. 营销界, 2025(13): 13-15.
- [7] 陈国占. 大数据背景下企业市场营销的创新策略[J]. 营销界, 2025(13): 65-67.
- [8] 王晓艺. 基于大数据的旅游酒店精准营销策略研究[J]. 旅游纵览, 2025(11): 182-184.
- [9] 张子峰. 塑料产品在电商平台的精准营销策略[J]. 塑料工业, 2025, 53(4): 193.
- [10] 王娇. 数据驱动的企业个性化营销策略及其效果评估[J]. 商场现代化, 2025(14): 88-90.
- [11] 袁铁铮. 农村特色农产品电商精准营销模式与实践研究[J]. 营销界, 2025(13): 10-12.
- [12] 杨静, 林宗平. 大数据背景下商贸流通企业市场营销体系重建研究[J]. 商业经济研究, 2025(11): 62-64.
- [13] 陈迎琦, 侯亚芳, 朱俊阁. 大数据背景下高职院校网络营销课程体系优化探讨[J]. 信息与电脑, 2025, 37(13): 239-241.
- [14] 罗立, 杨泽慧, 朱红花. 大数据环境下高职“计算机基础”课程精准教学研究与实践[J]. 知识文库, 2025, 41(12): 175-178.
- [15] 赵芳芳. 大数据分析中概率论与数理统计运用分析[J]. 科技风, 2025(19): 159-161.