Sustainable Development 可持续发展, 2026, 16(2), 80-89
Published Online February 2026 in Hans. https://www.hanspub.org/journal/sd
https://doi.org/10.12677/sd.2026.162060

Hans汉斯

基于关联规则与RFM聚类的社交电商用户购买行为分析

刘　巧*，李　明#

成都信息工程大学统计学院，四川 成都

收稿日期：2026年1月2日；录用日期：2026年2月2日；发布日期：2026年2月9日

摘　要

本文基于公开数据集网站天池提供的社交电商平台的真实数据，从用户特征、内容特征、社交特征及行为序列特征等多个维度对用户行为进行系统分析。首先，采用Apriori关联规则算法挖掘用户关键行为之间的关联关系，结果表明加购行为是用户购买决策中强烈的前置信号。然后基于RFM模型并结合K-means聚类算法对用户进行客户价值分层，在肘部法与轮廓系数法的支持下将客户划分为四类，并进一步通过AHP层次分析法与熵权法对客户价值进行综合评估。最后，在客户价值分层基础上，对不同用户群体的购买行为特征进行对比分析，发现价格与折扣因素对各类客户均具有普遍的正向影响，而视频内容与社交互动特征在不同价值客户群体中的作用存在差异。研究结果为社交电商平台实施精细化客户运营与差异化营销策略提供了数据支持与实践参考。

关键词

电子商务，Apriori，RFM，用户行为分析

Analysis of Social E-Commerce Users' Purchasing Behavior Based on Association Rules and RFM Clustering

Qiao Liu*, Ming Li#

School of Statistics, Chengdu University of Information Technology, Chengdu Sichuan

Received: January 2, 2026; accepted: February 2, 2026; published: February 9, 2026

*第一作者。
#通讯作者。

文章引用: 刘巧, 李明. 基于关联规则与 RFM 聚类的社交电商用户购买行为分析[J]. 可持续发展, 2026, 16(2): 80-89.
DOI: 10.12677/sd.2026.162060

Abstract

This study is based on real-world data from a social e-commerce platform provided by the public dataset website Tianchi. User behavior is systematically analyzed from multiple dimensions, including user characteristics, content characteristics, social characteristics, and behavioral sequence characteristics. First, the Apriori association rule algorithm is applied to mine the relationships among key user behaviors. The results indicate that the add-to-cart behavior serves as a strong antecedent signal in users' purchase decision-making process. Then, a customer value segmentation is performed using the RFM model combined with the K-means clustering algorithm. Supported by the elbow method and silhouette coefficient analysis, users are divided into four customer segments, and their customer value is further comprehensively evaluated using the Analytic Hierarchy Process (AHP) and the entropy weight method. Finally, based on customer value segmentation, comparative analyses of purchasing behavior across different user groups are conducted. The results show that price and discount factors have a universal positive impact on purchase behavior across all customer segments, while the effects of video content and social interaction features vary among customers with different value levels. The findings provide data-driven support and practical insights for implementing refined customer management and differentiated marketing strategies on social e-commerce platforms.

Keywords

E-Commerce, Apriori, RFM, User Behavior Analysis

1. 引言

随着移动互联网与社交媒体的快速发展，电子商务正逐步从以交易为核心的传统模式，向以内容与社交关系驱动的社交电商模式转型。以小红书、抖音等为代表的社交电商平台，通过内容分享、用户互动和社群传播，将商品信息嵌入用户的日常社交行为中，使用户的购买决策不再仅由价格或功能属性决定，而是受到内容质量、社交反馈以及他人行为示范等多重因素的共同影响。

李海杰等(2023) [1]使用改进的 Apriori 算法对旅客购票行为特征进行关联规则挖掘分析建立旅客购票行为关键特征体系并以此提出部分客运营销策略。Xiaoying Zhang (2023)等[2]利用 Apriori 算法对图书馆用户的借阅行为进行模式挖掘分析。该研究使用 Apriori 关联规则算法提取频繁项集和强规则，揭示了图书借阅之间的潜在关联关系，为图书推荐与馆藏优化提供数据支持。

Xian Z (2023)等[3]提出一种基于 Apriori 关联规则的电子商务潜在客户数据挖掘方法，通过优化 Apriori 算法并结合多维树结构和相似度计算，对电商用户交易行为进行分析和分类。该方法对潜在客户行为进行挖掘，为电商平台调整营销策略、识别高价值客户提供了有效的数据支持。

江丽桃等(2024) [4]构建含客户特征、当前及潜在价值三维度的分类指标体系，通过 BP 神经网络实现客户四分类，为客户管理提供依据。杨晓梅(2025) [5]以京东为对象，基于 RFM 模型识别重点客户，指出其在客户标签、服务响应及个性化方面的问题，提出差异化服务与智能化升级方案。徐文杰(2025) [6]改进传统 RFM 模型为 RFME 模型，新增 App 交互度维度，结合机器学习实现用户细分与休眠预测。

在此背景下，深入挖掘社交电商场景下用户行为特征，具有重要意义。基于此，本文以社交电商平台业务数据为研究对象，从用户行为模式挖掘与客户价值分类两个层面，对用户购买决策进行系统分析。

2. 理论基础

2.1. Apriori 算法

Apriori 算法最早由 R. Agrawal 等人提出，它通过逐层搜索的方式迭代寻找数据库中所有项集的关系[7]。将原始数据转化为由若干事务组成的事务数据库，其中每个事务包含用户在平台上的一组行为或事件项。然后通过扫描事务数据库计算各项集的支持度，筛选出满足最小支持度阈值的频繁 1 项集。随后，基于频繁项集的所有非空子集必然也是频繁的，由频繁 k 项集生成候选 k + 1 项集，并再次计算其支持度，剔除不满足阈值的项集。通过不断迭代，直至无法生成新的频繁项集为止。最后，在得到的频繁项集基础上，进一步计算规则的置信度和提升度，筛选出满足最小置信度要求的关联规则，用以刻画变量之间的关联关系。

2.2. RFM 模型

客户价值指标主要包括当前价值和潜在价值两个方面。对于当前价值的衡量，较多采用利润率和购买量作为评价指标；对于潜在价值的衡量，更多的是客户与企业联系的紧密度和未来预期的利润[8]。

本质上，RFM 模型主要是通过三个指标来分别刻画客户关系：最近一次消费时间 R (Recency)反映客户所处的生命周期阶段；消费频率 F (Frequency)体现客户行为的稳定性与黏性；消费金额 M (Monetary)反映客户对企业的经济价值水平，三者共同构成对客户关系状态和价值特征的综合刻画。传统的 RFM 认为 R 为最近一次消费时间与截止时间的间隔；F 为某段时间内的购买频率；M 为某段时间内的消费总额。

在传统 RFM 模型中，R (Recency)通常以最近一次购买时间间隔来衡量客户行为的活跃程度。然而，在社交电商场景下，用户的购买行为具有低频性，而浏览、点击等交互行为更加频繁，且对后续购买决策具有重要影响。鉴于数据集中未包含明确的最近一次购买时间，本文采用“距上次点击小时数”作为 R 指标的替代变量，用以刻画用户最近一次与平台发生交互的时间间隔。数值越小表示用户近期活跃度越高。因此本文 RFM 指标构造如下：

R (Recency)：距上次点击小时数。

F (Frequency)：近 30 天购买次数。

M (Money)：累计购买金额。

2.3. K-Means 算法

K-means 算法是聚类算法的典型代表，其基本原理是通过计算样本距离聚类中心的距离大小判断样本类别，是一种迭代求解的聚类分析算法[9]，K-means 算法的具体步骤如下。首先，根据预先设定的聚类数 K，在样本空间中随机选择 K 个样本点作为初始聚类中心。其次，计算每个样本点与各聚类中心之间的距离，并将样本划分到距离最近的聚类中心所在的簇中。然后，对每一个簇内的样本重新计算其均值向量，并将该均值作为新的聚类中心。上述“样本分配–中心更新”过程不断迭代，直至聚类中心不再发生明显变化或达到预设的迭代次数为止。最终得到 K 个内部相似度较高、簇间差异较大的聚类结果。

3. 实验分析与研究

实验数据来自公开数据库天池网站，数据包含了小红书、抖音等典型社交电商平台的真实业务场景，具有较强的现实代表性。数据集涵盖多个维度的用户行为信息：包括用户特征(如年龄、性别、用户等级等，共 10 个变量)、内容特征(如商品价格、折扣率、商品类目等，共 7 个变量)、社交特征(如点赞数、评

论数、分享数等，共 6 个变量)、行为序列特征(如是否加购、是否使用优惠券、浏览行为等，共 5 个变量)，用户群体以年轻用户为主，平均年龄约为 27 岁，其中女性用户占比约为 63.8%，符合当前社交电商平台的典型用户画像特征。

3.1. 客户行为模式挖掘——基于 Apriori 关联规则算法

本文采用 Apriori 关联规则算法对用户行为模式进行挖掘。得到如表 1、表 2 所示关联规则与反向关联规则，以用户为分析单位，将用户在平台上的交互行为、购买行为及部分内容特征转化为离散事件项，构建用户事务数据集。通过设置最小支持度和置信度阈值，挖掘用户在平台上的关键行为之间是否存在协同关系，从而揭示社交电商用户的典型行为组合特征。

Table 1. Correlation analysis table of user behaviors
表 1. 用户行为关联分析表

前项	后项	支持度	置信度	提升度
(加购_是)	(购买_是)	15.18%	64.83%	1.7774
(加购_是)	(购买_是，关注作者_否)	13.94%	59.56%	1.8033
(加购_是)	(购买_是，领券_否)	8.51%	60.41%	1.3429
(关注作者_是)	(购买_是)	11.96%	59.48%	1.3223
(领券_是)	(购买_是)	15.18%	64.83%	1.7774

Table 2. Reverse correlation analysis table
表 2. 反向关联分析表

前项	后项	支持度	置信度	提升度
(加购_否，领券_否，关注作者_否)	(购买_否)	37.66%	71.56%	1.30
(加购_否，领券_否)	(购买_否)	42.19%	68.85%	1.25
(加购_否，关注作者_否)	(购买_否)	44.46%	67.61%	1.23
(加购_否)	(购买_否)	49.70%	64.89%	1.18
(领券_否)	(购买_否)	46.87%	58.66%	1.07

在从表 1 可以看出，只要用户加购，64.83%会购买，所以加购是购买的最强前置信号；在加购的用户中，即使不关注作者，也有 59.56%会购买，从这里我们可以知道在已发生加购行为的用户中，是否关注作者对购买概率的提升作用不明显；加购的用户中，即使不领券，也有 60.41%会购买，所以领券不是加购用户的购买必要条件；关注作者的用户中 59.48%会购买，但支持度仅 11.96%，价值远低于加购；领券用户中 64.83%会购买，但支持度仅 15.18%，且提升度远低于加购。

在从表 2 的反向关联分析表可以看出，不加购不领券也不关注作者的三无用户 71.56%不购买，只要不加购的客户，64.89%概率不购买，这也再次验证加购的是购买前很强烈的信号，其余动作对于客户几乎不影响他们的购买决策。

3.2. 用户行为特征——基于客户价值分类

在进行用户行为分析前先要进行特征选择，为避免多重共线性的干扰。本文对客户的基本信息以及社交信息特征进行方差膨胀因子(VIF)的分析，VIF 大于 10 的变量如表 3 所示。从表 3 中可以看出点赞数、评论数、分享数、收藏数成高度正相关。从业务理解来看客户的点赞数、评论数、分享数、收藏数代表客户的社交互动力，因此我们可以选择点赞数来代表社交互动力。

Table 3. Collinearity analysis
表 3. 共线性分析

变量	VIF	容差
like_num	29.2215	0.0342
collect_num	12.3910	0.0807
share_num	10.9401	0.0914
comment_num	10.1515	0.0985

3.2.1. 基于 RFM 模型的 K-Means 算法聚类

本文首先对 *R*、*F*、*M* 进行归一化处理，以消除不同指标量纲差异对聚类结果的影响。在此基础上，采用 K-means 聚类算法对客户进行分类。为合理确定聚类数量，本文结合肘部法和轮廓系数法对不同聚类数下的聚类效果进行评估。得到的肘部法以及轮廓系数法计算过程如图 1、图 2 所示。结果表明，当聚类数取 4 时，类内 SSE 随聚类数增加的下降幅度出现明显拐点；同时，在轮廓系数分析中，聚类数为 4 时对应的轮廓系数达到最大，这也表明该聚类方案具有较好的类内紧密性和类间分离度。因此，综合考虑聚类效果与模型稳定性，本文最终将客户划分为 4 个聚类，并以 4 个簇心作为最终的聚类结果。

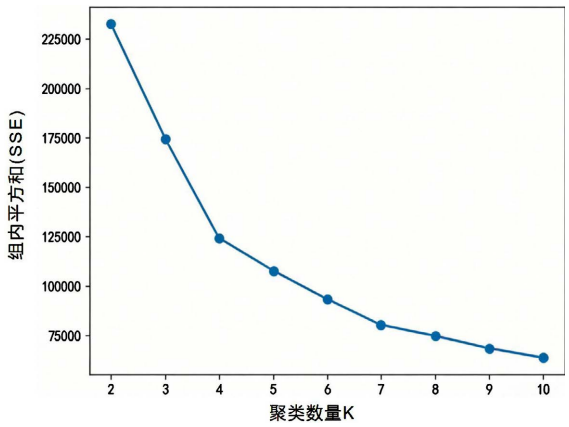


Figure 1. Plot of elbow method results
图 1. 肘部法计算结果图

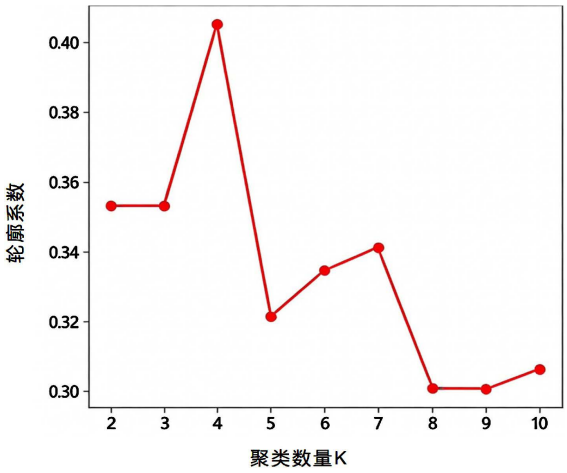


Figure 2. Plot of silhouette coefficient method results
图 2. 轮廓系数法计算结果图

最后得到聚类结果第一类：64,186，第二类 6657，第三类：13,296，第四类：15,861。聚类结果如表 4 所示。

Table 4. RFM clustering results
表 4. RFM 聚类结果

聚类类别	*R*	*F*	*M*	人数
1	0.043	0.161	0.040	64,186
2	0.061	0.203	0.219	6657
3	0.059	0.634	0.044	13,296
4	0.200	0.188	0.043	15,861

3.2.2. 基于 RFM 模型的 K-Means 算法聚类

常见的赋权方法有熵权法、AHP、变异系数法等等层次分析法，本文选择使用 AHP 层次分析法结合熵权法综合判断客户价值，主观赋权与客观赋权相结合的方法来确定权重[10]。

(1) AHP 层次分析法

结合相关文献以及专家打分的方法，由于本次采用数据来源为社交媒体兼网购的复合型平台，因此于 *F* 登录频率以及 *M* 消费金额相较于 *R* 最近一次点击时间间隔更重要，而电商平台首要聚焦于获利，所以 *M* 相比于 *F* 更为重要。基于以上分析，可以得到判断矩阵如表 5 所示。

Table 5. Judgment matrix
表 5. 判断矩阵

	R	*F*	*M*
R	1	1/4	1/5
F	4	1	2/3
M	5	3/2	1

对矩阵的每一列元素进行求和，得到列和向量 $S = [10, 2.75, 1.8667]$，归一化后得到 A'，最后得到权重 $R = 0993$，$F = 0.3742$，$M = 0.5265$。

为确保层次分析法计算出的权重结果具有合理性和可靠性，需要检测在构造判断矩阵时，两两比较的逻辑是否存在矛盾。通过计算得到一致性指标 CI = 0.0034，由于矩阵阶数为 3，得到平均随机一致性指标 RI = 0.58，一致性比率 CR = CI/RI = 0.0059，0.0059 < 0.1，因此一致性检验通过，权重结果可靠。为更直观展示数据，对于 RFM 分别乘以 100 再进行后续计算，根据权重结果，结合各类的指标数值可以得到各类用户的客户价值如表 6 所示：

Table 6. RFM clustering results based on AHP
表 6. 基于 AHP 的 RFM 聚类结果

聚类类别	*R*	*F*	*M*	总计
1	0.44	6.04	2.11	8.59
2	0.61	7.61	11.58	19.81
3	0.59	23.76	2.37	26.72
4	1.99	7.07	2.31	11.37

(2) 熵权法

由于量纲不同，为消灭其影响，首先利用离差标准化对 R、F、M 分别进行处理，但是由于 R 代表最近消费间隔是反向指标，数值越小，客户价值越高，因此 R、F、M 指标的转化函数依次如式(1)、(2)、(3)所示：

$$R' = \frac{R_{\max} - R}{R_{\max} - R_{\min}} \tag{1}$$

$$F' = \frac{F - F_{\min}}{F_{\max} - F_{\min}} \tag{2}$$

$$M' = \frac{M - M_{\min}}{M_{\max} - M_{\min}} \tag{3}$$

其中，R'、F'、M' 分别表示标准化处理后的数据值；R、F、M 分别代表客户消费数据的实际值；$R_{\max}$、$R_{\min}$、$F_{\max}$、$F_{\min}$、$M_{\max}$、$M_{\min}$ 分别代表所有客户中消费近度(R 指标)、频次(F 指标)、额度(M 指标)的最大值和最小值。

在经过离差标准化处理后，分别计算 R、F、M 三个指标的信息熵值。信息熵反映了指标数据的离散程度，熵值越小表明该指标包含的信息量越大。用 1 减去各指标的信息熵，得到对应的信息效用值。信息效用值代表了该指标在区分不同客户时的有效信息含量，效用值越大表明该指标的区分能力越强。经过计算得到最后权重 R：0.0045，F：0.4224，M：0.5731。处理过程同上，根据权重结果，结合各类的指标数值可以得到各类用户的客户价值如表 7 所示：

Table 7. RFM clustering results based on entropy weight method
表 7. 基于熵权法的 RFM 聚类结果

聚类类别	R	F	M	总计
1	0.0197	6.8184	2.3011	9.1392
2	0.0277	8.5956	12.6035	21.2268
3	0.0269	26.8182	2.5755	29.4206
4	0.0902	7.9808	2.5155	10.5866

AHP 以及熵权法最终得到的权重虽有一定差距，但是从客户最终价值的角度来看，对于客户分类的结果是相同的，分级结果如表 8 所示，等级 1 表示最高价值客户，以此类推，等级 4 表示最低价值客户。

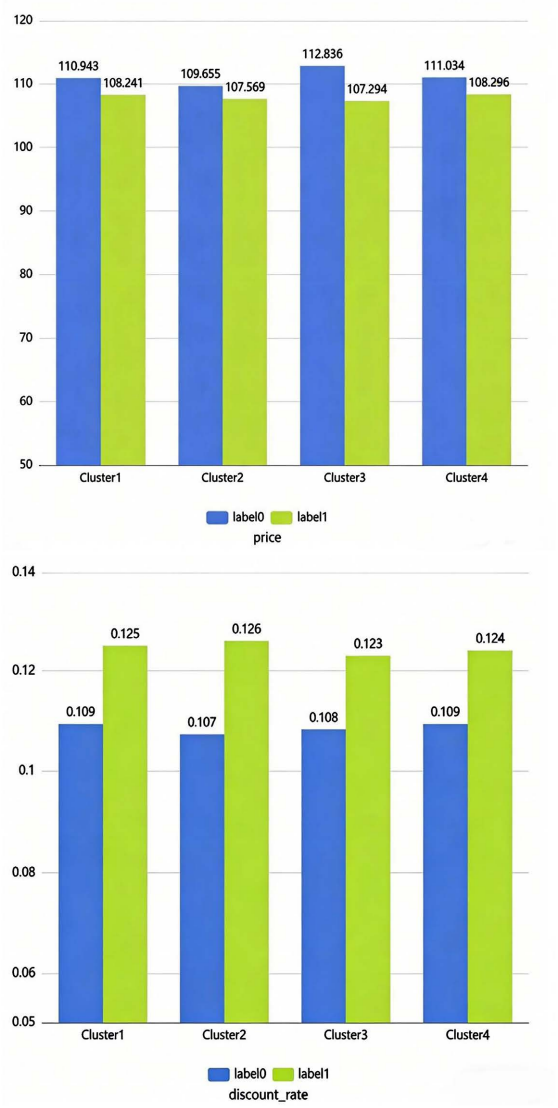
不同聚类类别在各维度上的表现存在明显差异，其中第 3 类客户的综合客户价值最高，主要受其在购买频次和消费金额维度上较高加权值的影响；第 2 类客户次之，其消费金额维度贡献较为突出；第 4 类客户综合价值与第 1 类客户综合价值均较低，在购买频次和消费金额维度上基本类似，第 1 类综合客户价值最低。整体来看，聚类结果能够有效区分不同客户群体的价值层级。

Table 8. Classified scores of each indicator
表 8. 各指标分类得分

客户等级	聚类类别	人数	基于 AHP 总计	基于熵权法总计
1	3	13,296	26.72	29.42
2	2	6657	19.81	21.22
3	4	15,861	11.37	10.59
4	1	64,186	8.59	9.14

3.2.3. 基于客户价值分类的用户行为分析

为深入探究不同客户群体的用户行为差异，本文针对不同的用户特征、商品内容以及社交特征进行对比，存在差异的结果如图 3 所示，label0 代表未购买对应产品，label1 代表购买了对应产品，第一幅图总体而言客户在决定购买时会一定程度考虑价格因素，价格低的情况下，无论哪种价值类型的客户意愿都会更强。第二幅图可以看出折扣率对购买行为具有明显的正向影响，折扣力度越大，用户更容易产生购买行为，折扣对不同价值等级客户均有效，折扣率并未只对某一类客户“失效”或“过度敏感”。第三幅图，视频内容对购买行为的影响呈现出明显的客户价值分层特征。在最高价值客户群体中，视频展示对购买转化的促进作用并不显著，此类客户的购物行为可能更多依赖于品牌以及历史偏好或者商品本身的质量属性。对于中低价值的客户，已购买用户对应的视频比例显著高于未购买用户，商品主页的视频助于降低信息不对称，提高购买信心。第四幅图中，对于第 1、3、4 类用户，点赞数越高也就是社交互动力高的购买商品的可能性也越大，结合社交电商的特点，高的社交互动力意味着在该平台有着更高的依赖程度，在社交电商平台中，用户购买决策往往受到他人行为的影响。较高的社交互动水平能够通过社会认同和信任传递机制增强用户对商品的信心，同时丰富的信息交流也有效降低了购买过程中的不确定性。



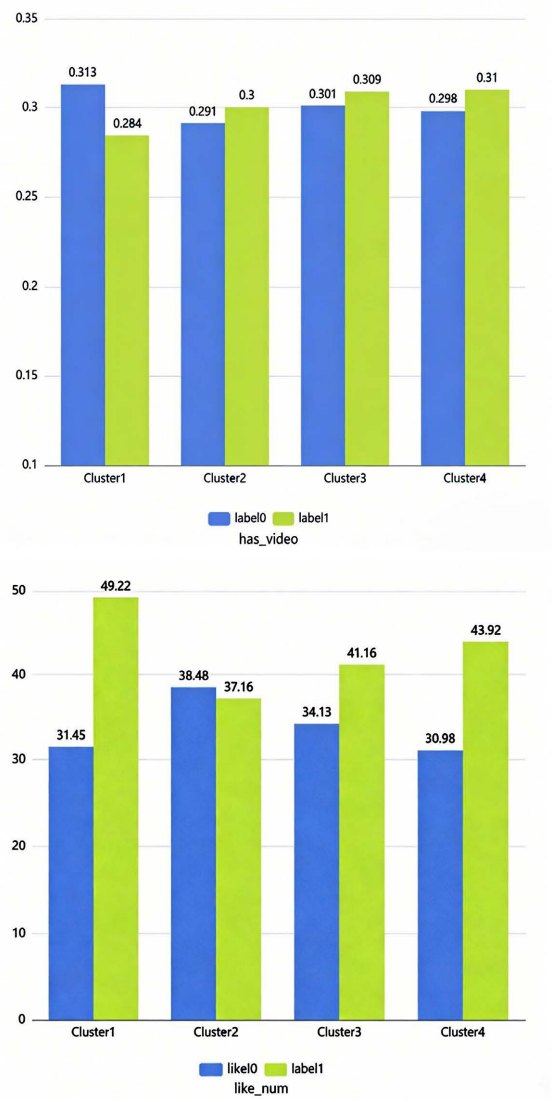


Figure 3. Mean value comparison plot for different customer groups
图 3. 各类客户的均值对比图

4. 结论与建议

首先，关联规则分析结果表明，加购行为是用户购买行为最强烈的前置信号。无论用户是否关注作者或是否领取优惠券，只要发生加购行为，其后续产生购买行为的概率显著提高，且加购规则的支持度、置信度和提升度均明显高于其他行为变量。这说明，在社交电商平台中，加购行为已成为用户从兴趣向实际购买转化的关键节点。相对而言，关注作者和领券行为虽对购买具有一定促进作用，但其独立影响强度明显弱于加购行为。反向关联规则进一步验证了这一结论，即未发生加购行为的用户群体中，不购买的概率显著提高，表明加购在用户购买决策路径中具有不可替代的重要作用。

基于 RFM 模型的聚类结果能够有效区分不同客户群体的价值层级。各类客户在购买频次和消费金额维度上的差异显著，其中高价值客户主要依赖较高的购买频次和消费金额贡献，而低价值客户在两项指标上均表现较弱。在不同客户价值等级下，用户购买行为对商品价格、折扣率、内容形式及社交互动特征的响应存在明显差异。总体来看，价格与折扣因素对各类客户均具有普遍的正向影响，折扣力度越

大，用户产生购买行为的可能性越高，且该影响在不同价值等级客户中具有一致性，说明价格激励在社交电商场景下具有较强的普适性。相比之下，视频内容对购买行为的影响呈现出明显的客户分层特征：在最高价值客户群体中，视频展示对购买转化的促进作用并不显著，而在中低价值客户群体中，商品主页配置视频能够有效降低信息不对称，提高用户购买信心。

最后，社交互动特征在用户购买决策中发挥着重要作用。对于多数客户群体而言，点赞数等社交互动水平越高，用户购买商品的可能性也越大。这一结果表明，在社交电商平台中，社交互动不仅是用户参与度的体现，更通过社会认同和信任传递机制，对用户购买决策产生显著影响。

基于上述研究结论，本文从平台运营与营销实践角度提出以下建议。

第一，将加购行为作为购买转化的核心监测指标。平台应重点关注用户的加购行为，通过加购提醒、库存提示、价格变动通知等方式，加速用户从加购到购买的转化过程。同时，可针对已加购但尚未购买的用户实施更精准的推送策略，以提升转化效率。

第二，实施基于客户价值分层的差异化运营策略。对于高价值客户，应更加注重商品质量、品牌建设及专属权益设计，而非单纯依赖内容形式刺激；对于中低价值客户，则应通过优化商品展示内容，如增加视频介绍、强化场景化表达，降低其购买决策门槛，逐步提升其客户价值。

第三，合理运用价格与折扣激励机制。研究表明折扣对不同客户群体均具有显著的正向影响，平台可结合客户价值等级和商品特性，制定灵活的优惠策略，在避免过度价格竞争的前提下，提高促销活动的针对性和有效性。

第四，强化社交互动生态建设。平台应鼓励用户进行点赞、评论、分享等互动行为，通过优化内容推荐机制和互动激励机制，从而增强用户对商品的信任感和购买意愿。

参考文献

[1] 李海杰, 苗蕾, 聂磊, 等. 基于关联规则和主成分分析的高铁旅客购票行为特征研究[J]. 铁道科学与工程学报, 2023, 20(6): 2013-2025.

[2] Zhang, X. and Zhang, J. (2023) Analysis and Research on Library User Behavior Based on Apriori Algorithm. *Measurement: Sensors*, **27**, Article ID: 100802. https://doi.org/10.1016/j.measen.2023.100802

[3] Xian, Z. and Hai, H. (2023) Mining Potential Customer Behavior in E-Commerce Based on Apriori Association Rules. *Lecture Notes in Computer Science*, **13945**, 168-180.

[4] 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.

[5] 杨晓梅. 基于 RFM 模型的京东电商平台“重点客户”维系策略优化研究[D]: [硕士学位论文]. 长春: 吉林大学, 2025.

[6] 徐文杰. 基于改进 RFM 模型的 D 电商公司用户运营策略优化研究[D]: [硕士学位论文]. 上海: 华东师范大学, 2025.

[7] 张梦琦. 基于 Apriori 算法的关联规则分析[D]: [硕士学位论文]. 大连: 大连理工大学, 2021.

[8] 余方超, 徐雷, 张国锋, 等. 基于 RFMI 模型的家电产品定制服务的客户价值研究[J]. 制造业自动化, 2023, 45(1): 91-94, 100.

[9] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 202.

[10] 郭峰, 王靖一, 王芳, 等. 测度中国数字普惠金融发展: 指数编制与空间特征[J]. 经济学(季刊), 2020, 19(4): 1401-1418.