E-Commerce Letters 电子商务评论, 2025, 14(11), 1189-1202
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14113551

Hans汉斯

# 基于HGBT的电商客户流失预测与精细化营销策略

**高江军，李　超，刘冰洋**

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2025年9月29日；录用日期：2025年10月17日；发布日期：2025年11月18日

## 摘　要

**在电商市场竞争日益激烈的当下，提升用户留存对电商平台的稳健发展至关重要，预测用户流失并制定针对性营销策略也因此极具现实意义。本文以Retail Rocket零售电商平台数据为基础，构建了“模型预测－特征解释－策略落地”一体化融合框架。通过特征集重要性与边际增益分析解释流失驱动因素：新近度是核心特征，贡献占比达57.22%，频次次之，占比32.98%；不同风险层级下的主要特征存在明显的非线性阈值效应，例如高风险层“最小购买间隔”阈值为10.94天，同时表明“会话新近度 × 历史购买次数均值”等关键特征的交互关系。基于此，本文构建了“风险－特征－干预”分层策略，形成从流失预警到精准营销的完整闭环，论证了HGBT模型在处理电商行为序列数据时，应对非线性与交互效应相关挑战的独特优势，可为电商平台提供高效、可靠的决策支撑。**

## 关键词

**HGBT，非线性阈值，交互，精准营销**

# E-Commerce Customer Churn Prediction and Refined Marketing Strategies Using HGBT

**Jiangjun Gao, Chao Li, Bingyang Liu**

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: September 29, 2025; accepted: October 17, 2025; published: November 18, 2025

## Abstract

**In the current context of increasingly fierce competition in the e-commerce market, improving user**

文章引用: 高江军, 李超, 刘冰洋. 基于 HGBT 的电商客户流失预测与精细化营销策略[J]. 电子商务评论, 2025, 14(11): 1189-1202. DOI: 10.12677/ecl.2025.14113551

**retention is crucial for the stable development of e-commerce platforms, and thus predicting user churn and formulating targeted marketing strategies is of great practical significance. Based on the data from the Retail Rocket retail e-commerce platform, this paper constructs an integrated framework of "model prediction - feature interpretation - strategy implementation". By analyzing feature set importance and marginal gain, the drivers of user churn are explained: recency is the core feature, accounting for 57.22% of the contribution, followed by frequency, accounting for 32.98%; the main features at different risk levels have obvious nonlinear threshold effects, for example, the threshold of "minimum purchase interval" in the high-risk layer is 10.94 days, and it also shows the interaction relationship of key features such as "session recency × average historical purchase times". Based on this, this paper constructs a "risk-feature-intervention" hierarchical strategy, forming a complete closed loop from churn early warning to precise marketing, demonstrating the unique advantages of the HGBT model in dealing with challenges related to nonlinearity and interaction effects when processing e-commerce behavior sequence data, and can provide efficient and reliable decision support for e-commerce platforms.**

## Keywords

**HGBT, Non-Linear Thresholds, Interactions, Precision Marketing**



## 1. 引言

随着电商行业的飞速发展，行业竞争加剧，获取客户的成本显著增加，客户流失问题已成为影响电商企业短期盈利和长久发展的重要因素之一。因此，客户流失预警与客户留存策略的实施是当前电商行业的关键问题。

客户流失预测是当前电商领域的研究热点之一，基于传统 RFM 模型[1]的方法受限于人工设定规则、固定的数据范围，并不适用于复杂、稀疏、高维及有较强时序相关性的数据；而机器学习方法可直接从数据中学习到特征间的非线性关系从而提高预测准确性，但其可解释性及效率低下的问题仍是一大难点。

本文构建了“特征解释–策略干预”的融合框架，以应对电商行为序列数据特性以及客户流失挽留。通过 HGBT 模型将模型输出的阈值、分层概率和重要特征值映射到具体的干预措施。形成完整闭环。

## 2. 文献综述

在客户流失分析中，已有许多研究围绕客户价值与行为规律展开。Fader [2]等人将经典的 RFM (近度、频率、金额)与客户生命周期价值(CLV)相联系，借助“等价值”曲线呈现 RFM 指标与 CLV 之间的主要相互作用关系，同时揭示了仅靠观测数据难以发现的非线性关联。针对在线团购情景，Wu [3]等人采用固定效应面板模型以及扩展模型研究了阈值引发的效应，发现阈值前后的正相关关系存在差异。

随着电商行为数据规模与复杂度的增大，模型效率成为新的研究热点。Ke [4]等人提出的 LightGBM 算法，通过直方图分析、梯度采样等技术，显著提升了梯度提升树在大规模稀疏行为特征上的分裂稳定性与速度，可高效处理海量电商行为数据。在此基础上，Guryanov [5]等人使用一种高效的基于直方图的算法构建树的梯度提升集成模型，值得注意的是该算法不受单个特征线性变换。推动了梯度提升类模型的应用。

与此同时，模型可解释性成为干预措施决策的前置条件。基于树的机器学习模型[6] [7]，如随机森林和梯度提升树等，是广泛应用的非线性预测模型。Lundberg [8]等人将 SHAP 值扩展到交互效应分析，可解释局部特征交互效应。通过结合每个预测的众多局部解释来理解全局模型结构。此外，Peng 将 GA-XGBoost [9] [10]结合 SHAP 揭示了特征交互的具体贡献以及单个样本特征的影响。

当前，模型可解释性是复杂算法落地的重要前提，尤其是电商行为序列数据中包含了大量多特征交互关系。而 HGBT 不仅承继梯度提升模型的高效性，还适配电商序列数据动态特征、捕捉非线性、复杂交互。为此，本文聚焦电商行为序列数据的特定挑战，针对性论证 HGBT 在非线性阈值识别与多维度特征交互建模上的独特优势，并将结果映射到干预措施中。

## 3. 研究方法与实施

### 3.1. 数据与特征工程

客户流失(customer churn)指在给定时间窗口内与企业终止业务关系或不再发生价值性交互的客户集合[11]。本文使用了电商零售数据集 Retail Rocket，属于非契约的电商场景，我们将数据中的 target_event 列中值为 1 的用户视为流失用户，值为 0 的用户视为未流失用户。我们提取了 2019 年 6 月至 9 月期间的用户行为与交易数据，建立多维度特征集合。总计 8355 个样本量，整体客户流失率为 36.11%，每个月份详细流失情况如表 1 所示。

**Table 1.** Statistics of churn rate from June to September 2019
**表 1.** 2019 年 6~9 月流失率统计

| Month | number | churn rate |
|---|---|---|
| 6 | 214 | 29.08% |
| 7 | 2282 | 35.08% |
| 8 | 2798 | 35.46% |
| 9 | 3061 | 37.96% |

原始数据集共包含 247 个特征，包含了用户行为序列和变异系数等多种统计量。我们对该特征集进行了以下数据清洗与特征筛选，经过处理后，我们最终保留 45 个特征用于后续任务。

(1) 缺失值处理：

对于数值型特征中的缺失值，采用中位数填充法。中位数填充能够有效减弱长尾分布和极值对数据的影响。

(2) 低方差特征过滤：

采用 VarianceThreshold 算法，设置方差阈值为 0.01，剔除方差低于该阈值的特征。低方差特征通常表现为在大多数样本中取值相对恒定，没有显著的变化，为冗余信息，效区分不同用户流失风险的关键信息。

(3) 高相关性特征筛选：

基于皮尔逊相关系数计算特征之间的相关性，去除相关系数高于 0.90 的特征对，仅保留其中之一。高度相关的特征往往会导致多重共线性问题，影响模型的稳定性与解释能力。去除后可以有效减小过拟合的风险。

### 3.2. 模型原理

HGBT (基于直方图的梯度提升树)是梯度提升的优化版本，通过直方图算法加速训练并应对大规模

数据，能捕获数据中的复杂模式，适合电商行为数据分析。该模型通过迭代添加弱学习器(决策树)来最小化损失函数。具体地，在第 m 轮预测中，模型更新为：

$$F_m(x) = F_{m-1}(x) + vh_m(x) \tag{1}$$

其中 $h_m(x)$ 是新决策树，$v$ 是学习率控制步长。直方图算法将连续特征分桶离散化，减少计算开销。训练中使用梯度下降近似，每轮拟合残差。正则化(如树深度、叶子节点数)和早停机制防止过拟合。

### 3.3. 指标建立

本文从模型预测性能指标(ACC、F1 和 AUC)以及运行效率(训练时间与预测时间)来进行综合评价。此外，我们还进行特征重要性分析，提升模型的可解释性。具体如下：

(1) 准确率(Accuracy, ACC)

表示衡量模型整体分类的正确比例，定义为：

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

其中 TP、TN、FP、FN 分别表示被模型预测为正类的正样本、被模型预测为负类的负样本、被模型预测为正类的负样本、被模型预测为正类的正样本。ACC 直观易懂，但在客户流失预测的类别不平衡场景下，可能被多数类(未流失用户)主导，因此仅作整体参考。

(2) F1 值

表示为精确率(Precision)和召回率(Recall)的加权平均：

$$\mathrm{F1} = \frac{2\mathrm{Prediction} \cdot \mathrm{Recall}}{\mathrm{Prediction}+\mathrm{Recall}} \tag{3}$$

$$\mathrm{Precisio} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

F1 能够综合反映模型对少数类(流失用户)的识别能力，避免单纯追求召回或精度的不平衡，在流失预测中更具实际意义。

(3) AUC (Area Under the ROC Curve)

AUC 反映的是二分类模型的性能，范围是[0, 1]，公式为：

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\mathrm{dFPR} \tag{6}$$

表示在所有可能的分类阈值下，模型的真正例率(TPR)与假正例率(FPR)的综合表现。由于 ACC 与 F1，AUC 不依赖某一个阈值，相对于 ACC 和 F1 更为稳定，同时也可以较好地规避类别不平衡的问题，在客户流失等二分类场景中经常会用到 AUC 指标来进行评价。

## 4. 实验结果

### 4.1. 性能基准对比

表 2 展示了 HGBT 模型在准确率(ACC)、F1 分数和 AUC 值等指标上相较于传统的逻辑回归(LR)和决策树模型具有明显的优势。具体来说，HGBT 模型在分类性能上表现最佳，AUC 值达 0.9067，较 LR

模型提升了约 2.96%，且具有显著的训练效率优势。这表明 HGBT 能够更好地捕捉电商用户行为序列中的复杂关系。此外，HGBT 模型在可解释性方面也展现了重要优势。HGBT 通过其基于树的结构提供了更为透明的决策过程。通过 SHAP 值(Shapley Additive Explanations)等解释方法，可以深入分析各个特征对最终预测结果的贡献，帮助业务团队理解影响客户流失的关键因素。

**Table 2.** Performance comparison of different models
**表 2.** 不同模型的性能对比

| 算法 | ACC (95% CI) | F1 (95% CI) | AUC (95% CI) | Training time [s] | Prediction time [s] |
|---|---|---|---|---|---|
| LR | 0.8351 (±0.0026) | 0.7376 (±0.0039) | 0.8806 (±0.0022) | 105.66 (±13.97) | 0.39 (±0.01) |
| 决策树 | 0.8455 (±0.0023) | 0.7827 (±0.0035) | 0.8913 (±0.0021) | 42.37 (±1.26) | 0.40 (±0.04) |
| HGBT | 0.8558 (±0.0022) | 0.7691 (±0.0033) | 0.9067 (±0.0019) | 19.23 (±0.64) | 0.42 (±0.01) |

## 4.2. 特征集贡献分析

为验证各特征组在整体与不同风险层人群中的相对贡献，我们对 20 折 OOF 验证样本进行 TreeSHAP 解释，并将同组内各特征的|SHAP|求和得到组层级贡献；随后计算均值与 95%置信区间，并计算得到贡献占比。

(1) 特征集贡献

在全体样本上，Recency 组的平均|SHAP|占比为 57.22%，显著高于其他组；Frequency 占 32.98%，居第二位；Monetary (5.01%)、Date & Time (3.77%)与 Preference (1.02%)贡献较小(见表 3)。这一结果表明，Recency (用户新近度与活动间隔相关的统计量)对流失风险的解释力最强，而行为频次提供次级但稳定的补充信息。

**Table 3.** SHAP values of different feature sets
**表 3.** 不同特征集的 SHAP 值

| 组别 | SHAP 值(95% CI) | 贡献占比 |
|---|---|---|
| Recency | 1.842 ± 0.0062 | 57.22% |
| Frequency | 1.062 ± 0.0016 | 32.98% |
| Monetary | 0.161 ± 0.0004 | 5.01% |
| Date&Time | 0.121 ± 0.0007 | 3.77% |
| Preference | 0.033 ± 0.0002 | 1.02% |

(2) 不同风险层

为更精细化地揭示特征组对不同流失风险用户的影响差异，本小节用 20 折 OOF (out-of-fold)预测概率对全量样本打分，按分位数划分为 High (Top10%)/Mid/Low 三类不同风险层。OOF 分数由未见过该样本的模型产生，避免信息泄漏。

图 1 展示了不同风险层中每个特征组的贡献占比。高风险层显著区别于中低风险层。在中低风险层中，中风险层中 Recency 占比仅仅比低风险层的高 3.7%，低风险组的 Date&Time 特征组高于中风险组，其余特征组相差不大。而在高风险层中，Recency 的贡献由 52%上升到 79.9%，而 Frequency 的相对权重明显下降(由 36.5%下降到 16.6%)。这意味着当用户进入高风险区间时，模型的解释结构由“频次–新近度并重”转变为“新近度主导”。在相对占比上发生系统性的偏移。

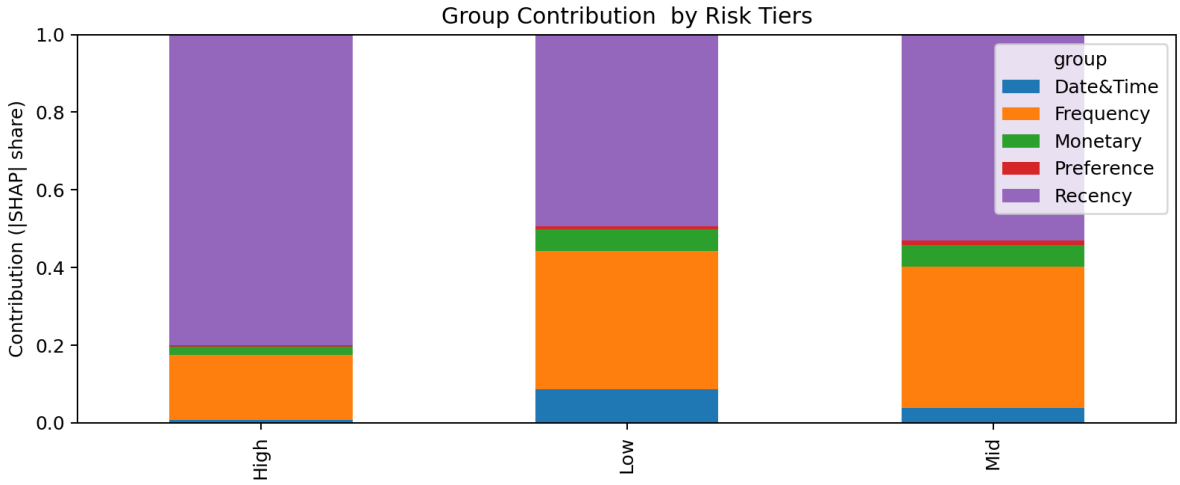



**Figure 1.** Feature contribution by different risk tiers
**图 1.** 不同风险层的特征贡献

### 4.3. 特征集边际增益

**Table 4.** Marginal gains and F1 performance of different feature groups
**表 4.** 不同特征组的边际增益与 F1 表现

| 特征组 | 去除该组 AUC | Δauc (边际增益) | f1@0.54 |
|---|---|---|---|
| Without-Recency | 0.8793 | 0.0264 | 0.7485 |
| Without-Frequency | 0.8955 | 0.0101 | 0.7202 |
| Without-Date&Time | 0.9054 | 0.0016 | 0.6583 |
| Without-Monetary | 0.9040 | 0.0002 | 0.4447 |
| Without-Preference | 0.9056 | 0.00008 | 0.4844 |

**Table 5.** Model performance comparison of single feature groups
**表 5.** 单一特征组的模型性能对比

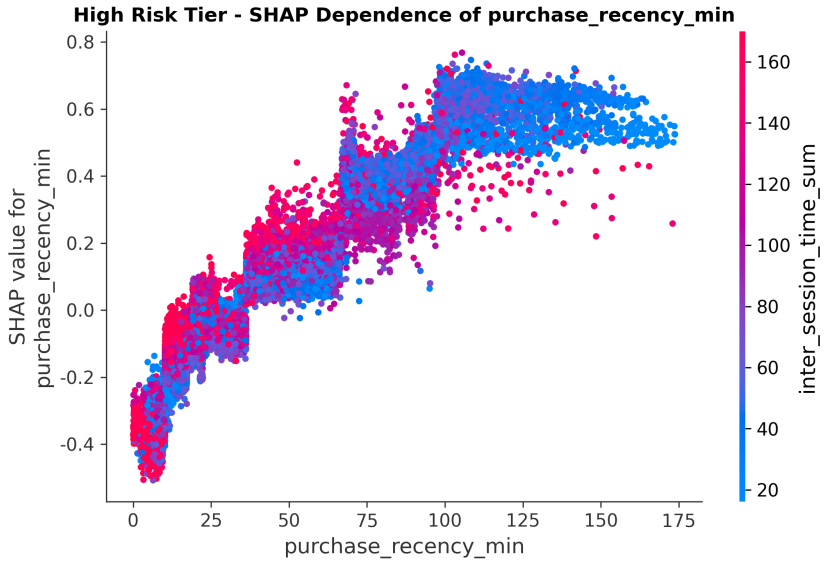
| 特征组 | AUC | f1@0.54 |
|---|---|---|
| 0nly-Recency | 0.8933 | 0.7485 |
| 0nly-Frequency | 0.8732 | 0.7202 |
| 0nly-Date&Time | 0.5673 | 0.6583 |
| 0nly-Monetary | 0.5632 | 0.4447 |
| 0nly-Preference | 0.5514 | 0.4844 |

由表 4 可知，仅使用 Recency 特征组后 AUC 从 0.9057 降至 0.8793，其边际增益为 0.0264；仅使用 Frequency 组，其 AUC 略有下降，边际增益 0.0101。其余特征组(Date&Time, Monetary, Preference)对整体模型的边际提升极小($\Delta AUC \leq 0.0016$)。这说明在电商行为序列的流失判别中，Recency 是首要信号、Frequency 次之，其余组信息更多被前二者吸收或可由其近似替代。固定阈值的分类表现与此一致：在 F1@0.54 下，Without-Recency = 0.7230、Without-Frequency = 0.7502 的下滑最明显，显示两组对可执行判

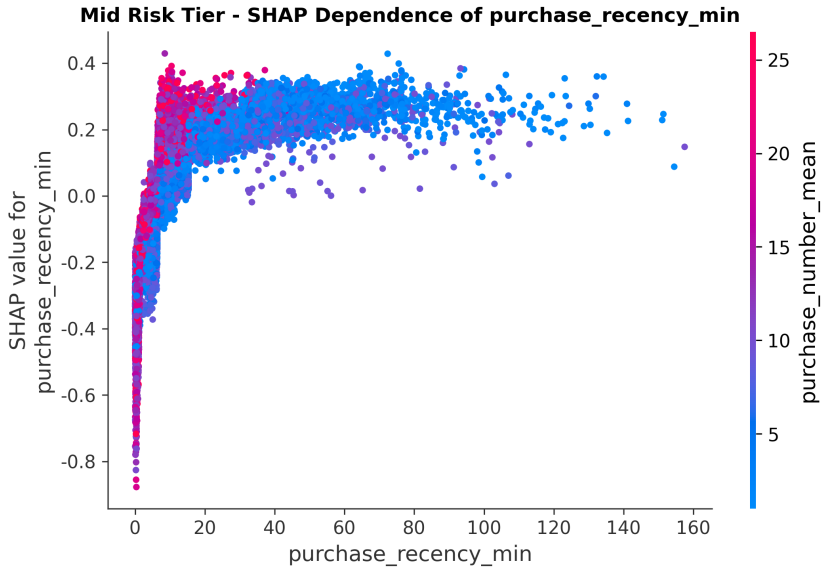
别能力贡献最大。与之相对，单一特征组结果(见表 5)表明 Only-Recency (AUC = 0.8933)与 Only-Frequency (AUC = 0.8732)单独即可形成强基线，而其余特征组的独立可用性显著偏弱，进一步巩固了上述结论。

### 4.4. 交互效应与阈值

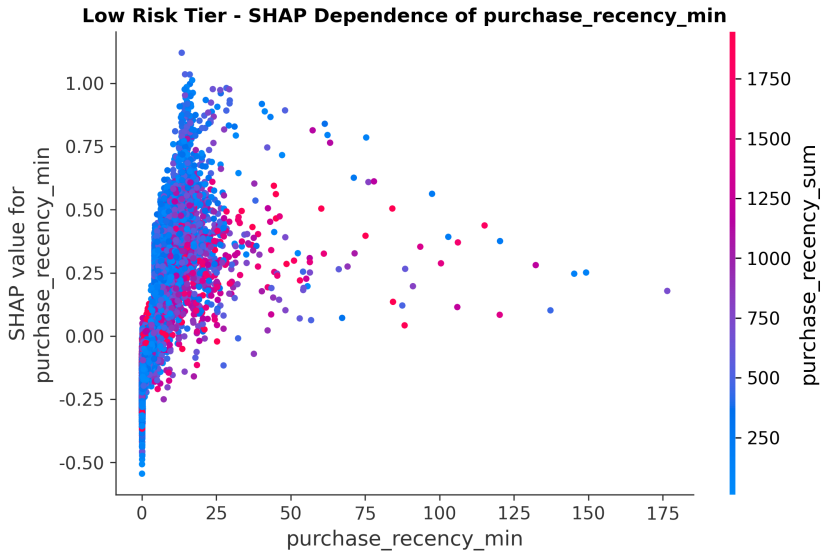
为进一步探究风险分层后核心特征的作用差异，我们基于 SHAP 方法绘制了不同风险层下核心特征“purchase_recency_min”(购买间隔最小值)对客户流失风险的依赖关系图，图 2 展示了不同风险层下，主特征与其他特征的交互关系。横轴为 purchase_recency_min (最近一次“购买”距今的天数)，Y 轴为 SHAP 值，SHAP 值为正时，代表该特征会增加客户流失概率；为负时，代表会降低流失概率。图中右侧颜色条为各风险层对应交互特征，高风险层为“inter_session_time_sum”、中风险层为“purchase_number_mean”、低风险层为“purchase_revenue_sum”。




(a)




(b)



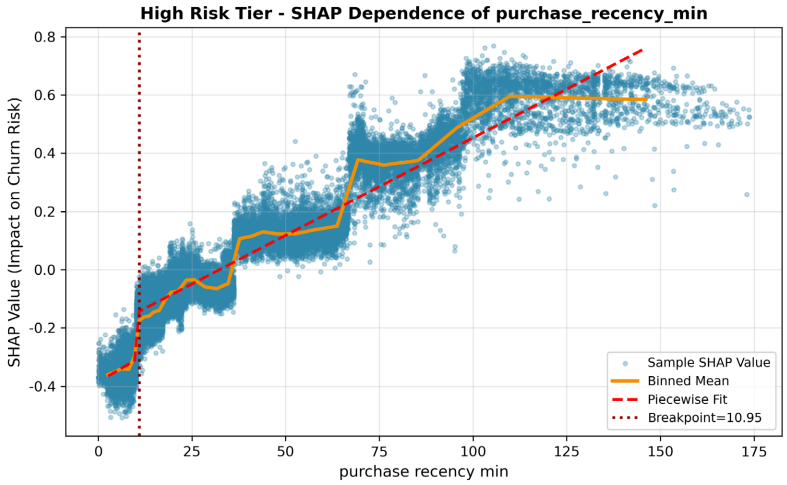

(c)

**Figure 2.** SHAP dependence plot between two features
**图 2.** 两特征之间的 SHAP 依赖图

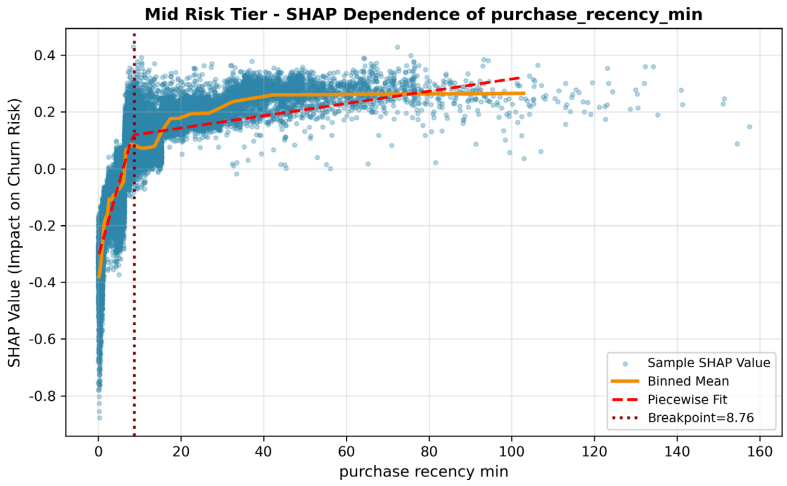
整体上看，高风险层(见图 2(a))中的 SHAP 值(正表示增加流失风险)整体上升，购买间隔越长，流失风险越高，且会话间隔总和较大时，该特征对流失风险的提升作用更明显。不同的是，中低风险层(见图 2(b))中，只有极小区域为负值(降低流失风险)，随后在中风险层中，“purchase_recency_min”超过一定阈值后由负转正且上升即购买间隔超过一定阈值会增加流失风险，同时平均购买次数不同会改变该特征的影响模式，趋近于饱和；而在低风险层(见图 2(c))，大量样本点集中在“purchase_recency_min”数值在 25 左右，且为正值，购买收入总和较高时，其对降低流失风险的作用更突出。进一步说明低风险层客户在“高购买收入 + 合理购买间隔”的特征组合下，流失风险被有效控制。可见，“purchase_recency_min”对不同风险层级客户的流失风险影响模式存在差异，结合各风险层辅助特征，能更清晰地看出不同特征组合的综合影响。

我们进一步在验证集上，以“purchase_recency_min”为横轴、对应的 SHAP 值为纵轴，先做分位分箱平滑得到均值曲线；随后在横轴上扫描单一拐点，在拐点两侧分别做最小二乘直线拟合并以加权 MSE 最小为准确定出阈值 T*，图 3 同时显示了散点、分箱均值与两段线性拟合。这些带阈值估计的 SHAP 依赖图展示了核心特征“purchase_recency_min”(购买间隔最小值)在不同风险层级(高、中、低风险层)中对客户流失风险的临界节点。

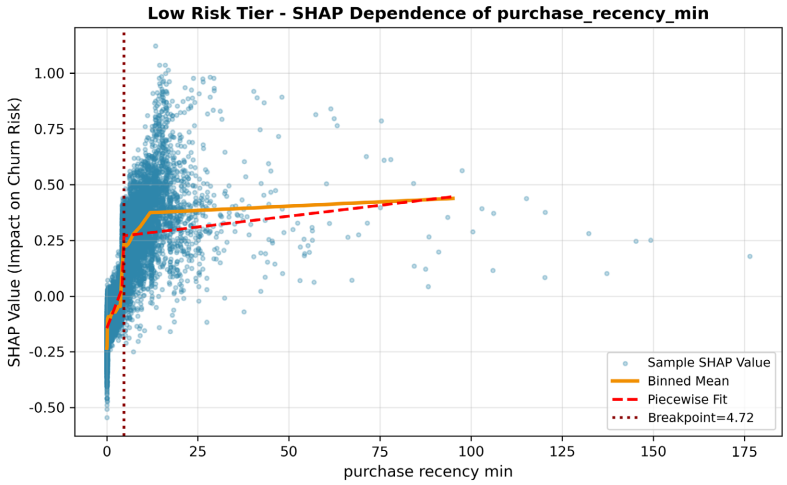
在高风险层(见图 3(a))，红色阈值线将其划分为两个区间——阈值(Breakpoint = 10.95)左侧 SHAP 值缓慢增长，右侧增长速率显著加快，表明当购买间隔超过该阈值后，对高风险客户流失风险的推动作用明显增强，与该层级客户本身体验脆弱、易受间隔延长影响的特征一致，并且右侧同样呈阶梯状。中风险层(见图 3(b))呈现更清晰的非线性关系：阈值(Breakpoint = 8.76)左侧 SHAP 值为负(特征降低流失风险)，右侧由负转正且持续上升，但是 SHAP 值逐渐趋于饱和，阈值成为影响方向的转折点，反映出该层级客户对购买间隔有较高的敏感度。低风险层(见图 3(c))则表现出稳定的风险抑制特征，绝大多数聚集在 SHAP 值为[0.0, 0.5]区间，没有较长的尾部。阈值(Breakpoint = 4.72)两侧的影响强度虽有差异但方向一致，并且存在于极小的“purchase_recency_min”区间，说明无论购买间隔如何变化，该特征始终发挥降低流失风险的作用，且不会因间隔延长而使风险性质发生逆转，这与低风险客户群体的稳定属性高度契合。三类风险层的阈值效应均与各自风险特征形成逻辑自洽，为分层干预策略提供了量化依据。




(a)

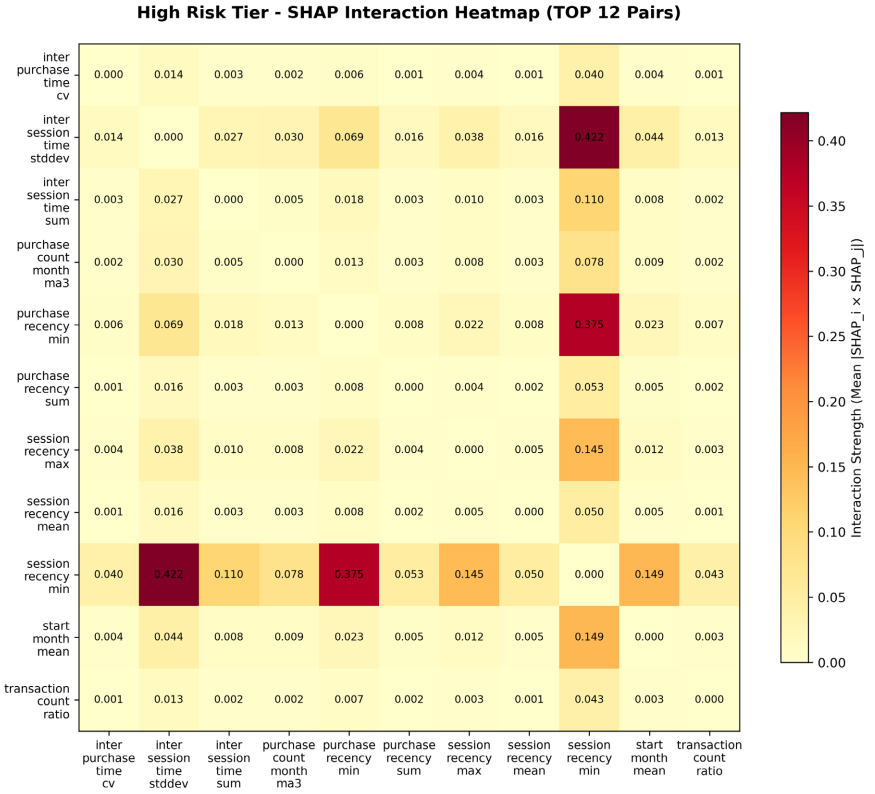



(b)

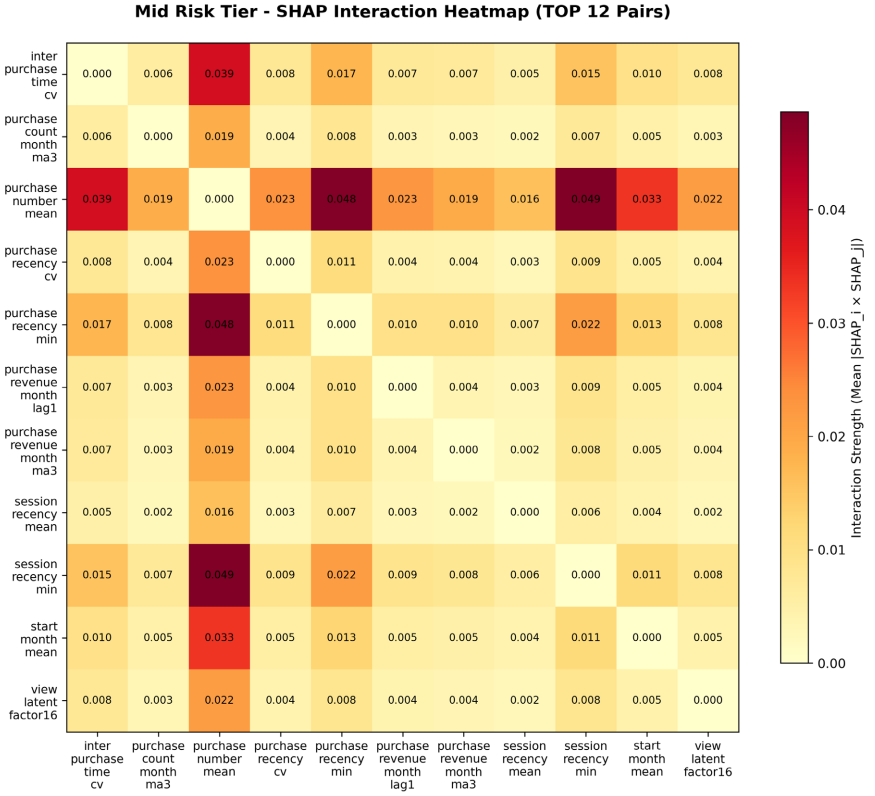



(c)

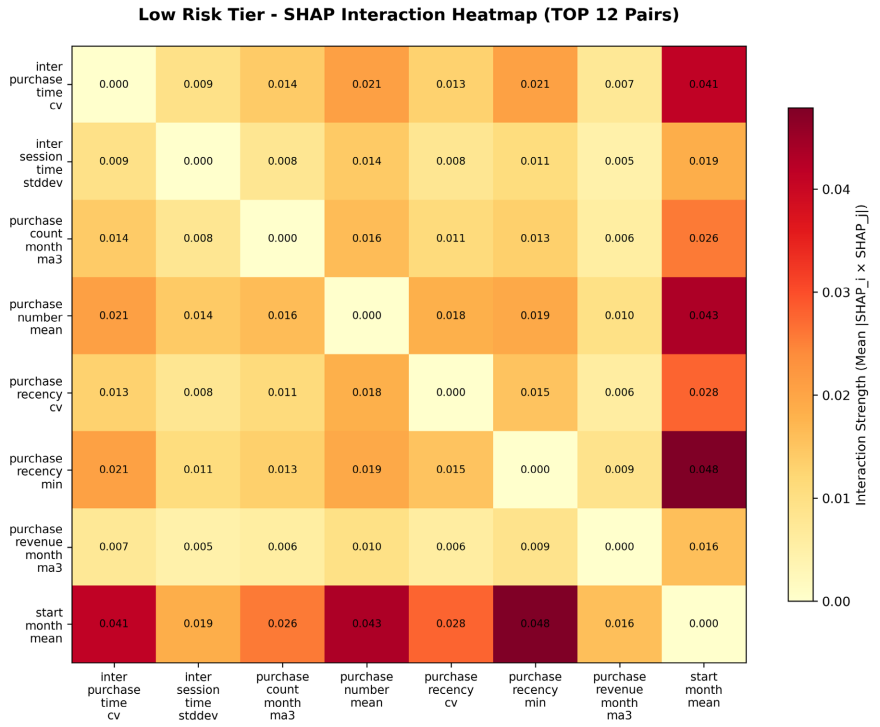
**Figure 3.** SHAP dependence plot with threshold estimation
**图 3.** 带阈值估计 SHAP 依赖图


High Risk Tier - SHAP Interaction Heatmap (TOP 12 Pairs)
inter purchase time cv 0.000 0.014 0.003 0.002 0.006 0.001 0.004 0.001 0.040 0.004 0.001
inter session time stddev 0.014 0.000 0.027 0.030 0.069 0.016 0.038 0.016 0.422 0.044 0.013
inter session time sum 0.003 0.027 0.000 0.005 0.018 0.003 0.010 0.003 0.110 0.008 0.002
purchase count month ma3 0.002 0.030 0.005 0.000 0.013 0.003 0.008 0.003 0.078 0.009 0.002
purchase recency min 0.006 0.069 0.018 0.013 0.000 0.008 0.022 0.008 0.375 0.023 0.007
purchase recency sum 0.001 0.016 0.003 0.003 0.008 0.000 0.004 0.002 0.053 0.005 0.002
session recency max 0.004 0.038 0.010 0.008 0.022 0.004 0.000 0.005 0.145 0.012 0.003
session recency mean 0.001 0.016 0.003 0.003 0.008 0.002 0.005 0.000 0.050 0.005 0.001
session recency min 0.040 0.422 0.110 0.078 0.375 0.053 0.145 0.050 0.000 0.149 0.043
start month mean 0.004 0.044 0.008 0.009 0.023 0.005 0.012 0.005 0.149 0.000 0.003
transaction count ratio 0.001 0.013 0.002 0.002 0.007 0.002 0.003 0.001 0.043 0.003 0.000
inter purchase time cv
inter session time stddev
inter session time sum
purchase count month ma3
purchase recency min
purchase recency sum
session recency max
session recency mean
session recency min
start month mean
transaction count ratio
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Interaction Strength (Mean |SHAP_i × SHAP_j|)

(a)


Mid Risk Tier - SHAP Interaction Heatmap (TOP 12 Pairs)
inter purchase time cv 0.000 0.006 0.039 0.008 0.017 0.007 0.007 0.005 0.015 0.010 0.008
purchase count month ma3 0.006 0.000 0.019 0.004 0.008 0.003 0.003 0.002 0.007 0.005 0.003
purchase number mean 0.039 0.019 0.000 0.023 0.048 0.023 0.019 0.016 0.049 0.033 0.022
purchase recency cv 0.008 0.004 0.023 0.000 0.011 0.004 0.004 0.003 0.009 0.005 0.004
purchase recency min 0.017 0.008 0.048 0.011 0.000 0.010 0.010 0.007 0.022 0.013 0.008
purchase revenue month lag1 0.007 0.003 0.023 0.004 0.010 0.000 0.004 0.003 0.009 0.005 0.004
purchase revenue month ma3 0.007 0.003 0.019 0.004 0.010 0.004 0.000 0.002 0.008 0.005 0.004
session recency mean 0.005 0.002 0.016 0.003 0.007 0.003 0.002 0.000 0.006 0.004 0.002
session recency min 0.015 0.007 0.049 0.009 0.022 0.009 0.008 0.006 0.000 0.011 0.008
start month mean 0.010 0.005 0.033 0.005 0.013 0.005 0.005 0.004 0.011 0.000 0.005
view latent factor16 0.008 0.003 0.022 0.004 0.008 0.004 0.004 0.002 0.008 0.005 0.000
inter purchase time cv
purchase count month ma3
purchase number mean
purchase recency cv
purchase recency min
purchase revenue month lag1
purchase revenue month ma3
session recency mean
session recency min
start month mean
view latent factor16
0.04
0.03
0.02
0.01
0.00
Interaction Strength (Mean |SHAP_i × SHAP_j|)

(b)




(c)

**Figure 4.** Heat map of SHAP interaction effects in different risk layers
**图 4.** 不同风险层的 SHAP 交互效应热力图

**Table 6.** Thresholds for main characteristics of different risk groups
**表 6.** 不通风险组主要特征的阈值

| 组别 | 特征 1 | 特征 2 | 特征 1 阈值 |
|---|---|---|---|
| High | session_recency_min | inter_session_time_stddev | 37.33 |
| | inter_session_time_stddev | session_recency_min | 2.22 |
| | purchase_recency_min | session_recency_min | 10.94 |
| Mid | session_recency_min | purchase_number_mean | 4.48 |
| | purchase_number_mean | session_recency_min | 10.75 |
| | purchase_recency_min | purchase_number_mean | 8.76 |
| Low | purchase_recency_min | start_month_mean | 4.71 |
| | start_month_mean | purchase_recency_min | 9.82 |
| | purchase_number_mean | start_month_mean | 10.25 |

图 4 为不同风险层下的特征中间 SHAP 交互效应热力图：对角线反映的是各特征的主效应强度，非对角线是为成对交互。从图中可以看到，除主效应强度外，最亮的区域出现在 session_recency_min × purchase_number_mean，说明“会话新近度”与“历史购买次数均值”的联合作用最强，而上述 purchase_recency_min × purchase_count_month_ma3 同样也有清晰但相对较弱的亮度——这表明“最近一次购买间隔”对流失风险的影响，会受到“近 3 个月购买频次”的调节，但这对交互并非全局最强。在少数关键对(尤其与“新近度/活跃度”相关)上存在显著的非线性耦合，表明 HGBT 模型能够有效捕捉电商行为序列中

“新近度/活跃度类特征(如 session_recency_min，purchase_recency_min)与购买频次/累计价值类”特征之间的复杂交互关系，为精准预测客户流失提供了更丰富的特征间联动信息。我们进一步计算了每个风险层中交互强度较强的前三对特征组，并计算了主特征的阈值(见表 6)。

## 5. 精细化研究策略

基于第四章实验分析结果，并结合 HGBT 模型输出的用户流失概率、主特征阈值及特征间的交互，本文构建了“风险－特征－干预”三层决策映射体系，将抽象的预测结果转化为可执行的差异化营销策略，实现营销资源精准分配与干预。实现降低流失率、增加用户留存和提升用户生命周期价值。本节包含决策规则构建与分层干预策略。

### 5.1. 决策规则构建

我们以 HGBT 模型的 OOF 预测概率(流失风险)为核心依据，结合 SHAP 交互热力图揭示的特征关联规律，筛选各风险层内具有较强影响的特征组合，再基于主要特征的阈值构建决策树策略触发规则集，最终生成 3 类核心规则(见表 6)，覆盖高、中、低全风险层级，确保每类用户群体的策略触发条件与特征影响逻辑高度匹配。决策树构建采用“风险分层－强交互特征组合－核心特征阈值分裂”的三级逻辑，突出特征交互优先性：

(1) 一级分裂：以 OOF 概率分位数为风险分层基础，划分为高风险层(概率 $P \geq 0.7$)、中风险层($0.3 < P < 0.7$)、低风险层($P \leq 0.3$)。

(2) 二级分裂：在各风险层内，基于 SHAP 交互热力图(见图 4)的 TOP3 交互对，识别强影响特征组合(见表 5)，确保捕捉特征间的协同影响。

(3) 三级分裂：在强特征组合框架下，以主要特征的阈值为关键分裂点，不同风险层重要特征具有不同的阈值以及与其他特征的交互作用，完成最终规则细分，既保留特征交互的主导作用，又通过阈值实现策略实践，提升业务解释力。

### 5.2. 分层干预策略

基于上述的决策规则触发结果，针对不同风险层强特征组合对应的用户痛点，设计“力度梯度化、渠道精准化”的干预策略，策略参数由上述特征交互与阈值分析确定，确保干预措施与用户流失驱动因素高度匹配。

所有策略参数均从 Retail Rocket 数据集的数据列中量化计算得出，例如折扣金额基于 purchase_revenue_sum (总购买收入)与 purchase_revenue_mean (平均购买收入)的分位数分层确定，触达渠道通过 start_hour_mean (平均会话小时)映射用户偏好，会话间隔阈值直接采用 session_recency_mean (平均会话间隔)的统计结果，具体计算依据与参数范围如下表 7 所示。

(1) 高风险层：紧急挽回策略

从特征交互上看，“最近会话间隔超 37.33 天与会话间隔波动高”时，用户流失风险达峰值，需最强干预，干预措施如下：

• 折扣力度最高：高消费用户折扣金额达 86~92 元(如 purchase_recency_min × session_recency_min 特征对中“满 430 减 86”)，折扣比例约 20%~25%，高于中低风险层，符合“高风险用户需高力度激励”的业务逻辑；

• 会话间隔最长：所有高风险组的平均会话间隔达 82.8~101.6 天(远超 7 天阈值)，因此干预时机均设置为“提前 1 天触达”，通过时间缓冲提升唤醒概率；

**Table 7.** Hierarchical statistical parameter quantity
**表 7.** 分层统计参数量

| 策略参数 | 数据计算依据列 | 高风险层参数范围 | 中风险层参数范围 | 低风险层参数范围 |
|---|---|---|---|---|
| 高消费用户满减门槛 | purchase_revenue_mean (平均购买收入) × 0.8 | 430~459 元 | 228~401 元 | 323~470 元 |
| 高消费用户折扣金额 | 满减门槛 × (20%~25%) | 86~92 元 | 56~80 元 | 65~94 元 |
| 平均会话间隔 | session_recency_mean 取均值 | 82.8~101.6 天 | 41.6~51.7 天 | 20.2~39.5 天 |
| APP 推送偏好占比 | start_hour_mean 映射后统计占比 | 98.9%~99.3% | 93.1%~99.3% | 95.8%~99.5% |
| 历史平均点击率 | click_count_mean/view_count_mean 均值 | 1.51~1.54 | 1.31~1.49 | 1.41~1.57 |

• 渠道集中度最高：APP 推送偏好用户占比 98.9%~99.3%，说明高风险用户虽长期沉默，但历史活跃时段仍以“APP 主动访问”为主(与 start_hour_mean = 14 (下午 2 点)的峰值小时匹配)，APP 推送触达精准度最优。

(2) 中风险层：预警干预策略

• 折扣力度适中：高消费用户折扣金额 56~80 元(如 purchase_number_mean × session_recency_min 特征对中“满 401 减 80”)，折扣比例 15%~20%，低于高风险层但高于低风险层，平衡“防流失”与“控成本”；

• 会话间隔中等：平均会话间隔 41.6~51.7 天(仍超 7 天阈值)，但较高速风险层缩短 50%以上，因此同样保留“提前 1 天触达”，但可考虑后续迭代中缩短提前时间(如 0.5 天)；

• 渠道占比略降：APP 推送偏好用户占比 93.1%~99.3%，其中 purchase_recency_min × purchase_number_mean 组占比最低(94.5%)，该组用户“平均购买次数低(purchase_number_mean < 8.76)”，可补充短信触达覆盖非 APP 活跃用户。

(3) 低风险层：留存维护策略

• 折扣力度最低：高消费用户折扣金额 65~94 元，但折扣比例 15%~20% (与中风险层相当)，但低消费用户折扣门槛更高(如 purchase_number_mean × start_month_mean 组“满 100 减 40”)，因低风险用户购买意愿强，无需低门槛刺激；

•会话间隔最短：平均会话间隔 20.2~39.5 天(虽超 7 天，但较中高风险层显著缩短)，其中 start_month_mean × purchase_recency_min 组仅 20.2 天，后续可迭代为“当天触达”，减少过度打扰；

•渠道集中度高：APP 推送偏好用户占比 95.8%~99.5%，且 purchase_number_mean × start_month_mean 组达 99.5% (与高风险层相当)，该组用户“平均购买次数高(purchase_number_mean > 10.25)”，APP 是其核心交互渠道，触达效果最优；

• 点击率最高：purchase_number_mean × start_month_mean 组的历史平均点击率 1.5795 (全风险层最高)，预期点击率 1.7374，说明低风险用户中“高购买频次 + 季节活跃”的群体对 APP 推送的响应度最高，可作为“核心用户留存”的重点对象。

## 6. 结论

本文以 Retail Rocket 电商数据集为研究基础，围绕电商客户流失预测与精细化营销开展研究。通过 HGBT 模型对具有复杂特征关联的数据进行特征集重要性分析，并进行非线性阈值计算得到特征的依赖关系。

特征分析显示 Recency (新近度)是影响流失的核心特征(贡献占比 57.22%)，Frequency (频次)次之

(32.98%)，且高风险用户中 Recency 贡献占比升至 79.9%，不同风险层核心特征存在显著阈值差异(如高风险层“purchase_recency_min”阈值 10.95 天)，HGBT 还能有效捕捉“会话新近度 × 历史购买次数均值”等关键特征交互；基于此构建的“风险–特征–干预”分层策略，可实现高风险层 86~92 元高折扣 + APP 推送紧急挽回、中风险层 56~80 元折扣 + APP + 短信平衡成本与效果、低风险层高门槛优惠减少打扰，最终形成“预测–解释–干预”闭环，为电商提升用户留存、优化营销资源分配提供可落地方案。

非线性阈值与特征交互关系对于电商行为数据分析尤为重要，本文通过特征重要性比较以及阈值计算论证了 HGBT 模型对于电商行为数据的优势。未来可引入深度学习与因果推断等技术，进一步提升对用户复杂行为模式的建模能力与干预策略的可解释性。

## 参考文献


[1] Fridrich, M. and Dostál, P. (2022) User Churn Model in E-Commerce Retail. *Scientific Papers of the University of Pardubice*, *Series D*: *Faculty of Economics and Administration*, **30**, Article 1478. https://doi.org/10.46585/sp30011478

[2] Fader, P.S., Hardie, B.G.S. and Lee, K.L. (2005) RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. *Journal of Marketing Research*, **42**, 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

[3] Wu, J., Shi, M. and Hu, M. (2015) Threshold Effects in Online Group Buying. *Management Science*, **61**, 2025-2040. https://doi.org/10.1287/mnsc.2014.2015

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y., *et al.* (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3149-3157.

[5] Guryanov, A. (2019) Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees. In: van der Aalst, W., *et al.*, Eds., *Analysis of Images*, *Social Networks and Texts*, Springer, 39-50. https://doi.org/10.1007/978-3-030-37334-4_4

[6] Suh, Y. (2023) Machine Learning Based Customer Churn Prediction in Home Appliance Rental Business. *Journal of Big Data*, **10**, Article No. 41. https://doi.org/10.1186/s40537-023-00721-8

[7] Li, X. and Li, Z. (2019) A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des systèmes d information*, **24**, 525-530. https://doi.org/10.18280/isi.240510

[8] Lundberg, S.M., Erion, G.G. and Lee, S.I. (2018) Consistent Individualized Feature Attribution for Tree Ensembles. arXiv: 1802.03888.

[9] Peng, K., Peng, Y. and Li, W. (2023) Research on Customer Churn Prediction and Model Interpretability Analysis. *PLOS ONE*, **18**, e0289724. https://doi.org/10.1371/journal.pone.0289724

[10] Zeng, F., Wang, J. and Zeng, C. (2025) An Optimized Machine Learning Framework for Predicting and Interpreting Corporate ESG Greenwashing Behavior. *PLOS ONE*, **20**, e0316287. https://doi.org/10.1371/journal.pone.0316287

[11] Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007) Computer Assisted Customer Churn Management: State-of-the-Art and Future Trends. *Computers & Operations Research*, **34**, 2902-2917. https://doi.org/10.1016/j.cor.2005.11.007