天猫用户复购行为预测分析报告

**1. 引言**

**1.1 分析目标**

本项目的主要分析目标是预测在“双十一”当天对特定商家首次发生购买行为的新用户，在未来6个月内是否会再次在该商家进行购买。这是一个典型的二分类问题，旨在通过分析用户历史行为数据和用户基本信息，构建预测模型，识别出具有高复购潜力的用户群体，从而为商家制定精准营销策略和提升用户忠诚度提供数据支持。

**1.2 问题说明**

本案例来源于“天池”学习赛中的天猫用户重复购买预测问题 。其核心挑战在于，训练数据是基于“双十一”活动前6个月的用户行为，而预测目标是“双十一”当天产生首次购买的新用户在未来6个月的复购情况 。这其中涉及到新用户观察期不足6个月的问题，需要进行妥善处理。此外，数据提供的特征较为零散，需要我们针对其做出处理以准确完整地衍生出对训练模型有意义的特征。并且在原始数据中存在分布不平衡、包含缺失值等问题，需要我们在数据预处理阶段进行处理。

**2. 数据集描述与集成**

**2.1 数据集概览**

项目使用的数据主要包含以下四个文件：

user\_log\_format1.csv (用户行为日志)

内容：记录了用户的历史行为数据。

字段：user\_id (用户ID), item\_id (商品ID), cat\_id (商品品类ID), seller\_id (商家ID，为对齐字段在脚本中已重命名为 merchant\_id), brand\_id (品牌ID), time\_stamp (行为发生时间，格式为mmdd), action\_type (行为类型：0表示点击，1表示加入购物车，2表示购买，3表示收藏)。

初步理解：这是最核心的数据，包含了用户的详细行为序列。brand\_id 存在缺失值，已在脚本中用0填充。time\_stamp 已被转换为更易于分析的年内天数 (abs\_day) 和月份 (month)。此表可能存在同一用户在同一天对同一店铺的同一品牌商品执行相同动作的记录，这被理解为正常的用户重复操作，而非冗余。

user\_info\_format1.csv (用户信息表)

内容：提供用户的基本画像信息。

字段：user\_id (用户ID), age\_range (年龄段), gender (性别)。

初步理解：age\_range 和 gender 存在缺失值。脚本中将 age\_range 的缺失值（包括NaN和0）用0填充，并将原来的7和8（均表示>=50岁）统一为6。gender 的缺失值（包括NaN和2）用2填充。这些字段本质上是分类变量，但在原始数据中可能被识别为数值型。脚本中已检查并移除了基于 user\_id 的重复行。

train\_format1.csv (训练集)

内容：提供了用户在特定商家是否复购的标签信息，用于模型训练。

字段：user\_id (用户ID), merchant\_id (商家ID), label (标签：1表示复购，0表示未复购)。

初步理解：该数据集的 label 是基于“双十一”前6个月的行为统计。一个关键问题是，这个 label 的观察期对于新注册用户可能不足6个月，甚至可能只反映了“双十一”当天的行为，这与预测“双十一”后6个月的复购目标存在差异。训练集中的用户-商家对是唯一的。

test\_format1.csv (测试集)

内容：提供需要预测复购概率的用户-商家对。

字段：user\_id (用户ID), merchant\_id (商家ID), prob (预测的复购概率，初始全为NaN)。

初步理解：测试集中的 user\_id 与训练集中的 user\_id 没有重复，但 merchant\_id 可能有重复。这意味着我们无法直接获取测试集用户在目标商家历史上的复购信息，但可以利用他们在其他商家或目标商家的“双十一”当天行为信息。并且如果一个商家同时存在于训练集与测试集中，那我们就可以从训练集中了解这个商家的历史表现，将其作为商家的特征，帮助预测新用户在该商家是否会复购。

**2.2 数据集成**

为了构建统一的分析数据集，进行了如下的数据集成工作：

合并用户信息：将 train\_df\_orig 和 test\_df\_orig 分别与 user\_info\_df 左连接（left merge），将用户的年龄和性别信息添加到主数据集中，形成包含用户画像的 all\_df。

聚合用户行为日志形成路径特征：首先，对 user\_log\_df 按 user\_id 和 time\_stamp 排序。然后，按 user\_id 分组，将每个用户的 item\_id, cat\_id, merchant\_id (由seller\_id重命名得到), brand\_id, time\_stamp, action\_type 序列合并成空格分隔的字符串，形成 item\_path, cat\_path 等路径特征。

合并聚合后的用户行为路径：将上一步得到的 user\_log\_path (包含了每个用户的行为序列路径) 与 all\_df (或其早期版本) 基于 user\_id进行合并。

通过这些集成工作，目标是为每个用户-商家对构建一个全面的特征集，包含了用户基本信息、用户整体行为摘要以及用户在该商家的具体互动行为摘要。

2.3 初步数据探索发现

脚本对原始数据表进行了重复性检查。例如，检查了 user\_info\_df 中 user\_id 的重复情况，并在发现重复时予以移除，保留第一个出现的记录。训练集的标签分布也进行了初步查看，以了解正负样本的比例。这些初步探索有助于了解数据质量，为后续的数据清洗和预处理提供依据。

**3. 数据预处理 (Data Preprocessing)**

**3.1 内存优化**

由于用户行为日志 (user\_log\_format1.csv) 数据量较大，为提高数据处理效率和避免内存不足问题，脚本中定义并使用了 reduce\_mem\_usage 函数。该函数通过检测数值型特征列（整型和浮点型）的实际取值范围，将其转换为能够容纳该范围的最小数据类型（如将 int64 转换为 int32 或 int8，float64 转换为 float32 或 float16），从而显著减少了数据在内存中的占用。此函数被应用于所有加载的原始数据表：user\_log\_df, user\_info\_df, train\_df\_orig, test\_df\_orig。

**3.2 基础数据清洗**

**3.2.1 用户行为日志 (user\_log\_df) 清洗**

针对用户行为日志数据，主要进行了以下清洗操作：

**字段重命名**：将 user\_log\_df 中的 seller\_id 列重命名为 merchant\_id，以确保与训练集和测试集中的商家ID字段名保持一致，方便后续数据合并与关联。

**缺失值填充**：brand\_id (品牌ID) 列存在缺失值，脚本中使用 0 进行填充。这通常意味着将缺失品牌视为一个特殊的“未知品牌”类别。在后续计算如 user\_brand\_nunique （用户交互过的不同品牌数量）时，脚本中明确排除了值为0的品牌ID (lambda x: x[x != 0].nunique())，这表明 0 在这里被视为一个特殊标记，不计入有效品牌计数。

**时间戳转换**：原始的 time\_stamp 字段为 mmdd 格式（如511代表5月11日）。为便于进行时间相关的计算和分析，脚本将其转换为两个新特征：

abs\_day：年内绝对天数，通过自定义函数 mmdd\_to\_day\_number 实现，该函数将 mmdd 转换为相对于某个起始点（如5月1日为第0天或第1天）的具体天数。

month：行为发生的月份，通过 time\_stamp // 100 计算得到。

**无效时间记录处理**：在 time\_stamp 转换为 abs\_day 后，若因为原始时间戳无效导致 abs\_day 产生 NaN 值，这些包含无效时间记录的行将被从 user\_log\_df 中移除，以保证后续分析的准确性。

**3.2.2 用户信息表 (user\_info\_df) 清洗**

针对用户信息表，主要进行了以下清洗：

**缺失值填充**：

age\_range (年龄段)：原始数据中的缺失值（包括NaN和0）统一用 0 填充。这是一种将缺失值视为一个独立类别的策略，或者认为 0 是一个合理的默认值/未知类别。后续还会将7和8（均表示>=50岁）替换为6，进一步规整年龄段的表示。

gender (性别)：原始数据中的缺失值（包括NaN和2）统一用 2 填充。将 NaN 填充为 2 是为了将缺失值明确标记为一个独立的“未知性别”类别。这样做可以让模型区分出性别已知和性别未知用户可能存在的行为差异。

**值规整**：

age\_range：原始数据中的年龄段7和8（均表示>=50岁）被统一替换为6，以合并等效类别，减少冗余。

**重复记录处理**：

检查 user\_info\_df 中是否存在基于 user\_id 的重复记录。如果存在，则保留首次出现的记录，移除后续的重复项，以确保每个用户ID对应唯一的用户信息。

**3.3 缺失值处理策略与原因**

脚本中对多个阶段产生的不同特征都进行了缺失值填充工作。选择何种填充方法通常取决于特征的类型、缺失原因、缺失比例以及后续模型的特性。

**原始特征的缺失值填充**

**特征工程后 all\_df 中特征的缺失值填充** (主要在脚本“3.5 处理特征工程产生的缺失值”部分)：

**计数、比例、熵、近期行为等特征** (如 user\_log\_count, user\_action\_type\_0\_ratio, u\_logs\_last\_7d)：这类特征的缺失通常意味着用户没有发生相应的行为，或者没有足够的数据来计算这些统计量。例如：

如果一个用户没有任何日志记录，那么他的 user\_log\_count 自然是 0。

如果用户没有购买行为，那么与购买相关的比率（如 user\_addcart\_to\_buy\_ratio）的分子为0，或者因为分母（加购行为）也为0而导致整个比率为NaN。此时用0填充表示没有发生这种转化，或者转化率为0。

对于近期行为（如 u\_logs\_last\_7d），如果近期没有活动，计数为0是合理的。

因此，用 0 填充在这些情况下是符合业务逻辑的，即“无行为”或“无数据”等同于数值上的“零”。

**时间差及购买间隔相关特征** (如 u\_mean\_buy\_interval, u\_mean\_time\_diff)：首先尝试使用该列的**中位数**填充，因为中位数对极端值不敏感，不像均值那样容易受到极端值的影响，能较好维持数据分布；若中位数本身为 NaN，则用 0 作为备选填充，需要注意的是这里需要注意，0本身也可能是一个有效的时间差（例如同一天内相隔很近的两次行为）。

**首次/末次活跃日期特征** (如 user\_first\_abs\_day)：这些特征表示用户首次或末次活动的绝对日期（年内天数）。如果一个用户在 user\_log\_df 中完全没有行为记录，那么这些值计算出来就是 NaN。用 0 填充可以将这些无活动用户标记为一个非常早的日期（第0天），或者作为一个特殊的“无活动历史”标记。这使得后续基于这些日期的计算（如活动时长）能够进行，或者模型能够将这些“第0天”的用户区分开。

**其他所有未被覆盖的数值型特征**：统一使用 0 作为默认填充策略。

all\_df['age\_range'] 和 all\_df['gender'] 的再次填充：确保合并后若产生新的缺失，仍按原规则填充为“未知”类别。

**特征变换/选择过程中的临时填充**：

**分箱前**：在计算分箱边界时，通常需要完整的数值数据。使用中位数填充是一个常见的预处理步骤，以确保分箱器能够正确处理所有样本，同时尽量减少对原始数据分布的扭曲。

**SelectKBest/RFE前**：对用于特征选择的临时数据，缺失值（包括inf替换产生的NaN）用**中位数**填充（RFE在之后还可能用0填充剩余NaN），以满足算法对无缺失数据的要求。

**3.4 类型转换**

脚本在多个阶段进行了类型转换，以确保数据格式的正确性、优化内存使用并满足模型输入要求。

**时间戳转换**：time\_stamp (mmdd格式整数) 被转换为数值型的 abs\_day (年内天数) 和整数型的 month (月份)。

**内存优化 (reduce\_mem\_usage)**：如3.1节所述，该函数将数值型（整型和浮点型）列转换为占用内存更小的数据类型。

**特定列的数值转换**：

user\_first\_abs\_day 在处理观察期差异时，通过 pd.to\_numeric 转换为 user\_first\_abs\_day\_numeric，无效值转为 NaN。

**特征工程与模型准备中的转换**：

分箱特征 (\_binned)：KBinsDiscretizer 的输出被转换为整型 (.astype(int))。

标签 (label)：最终被转换为整型 (.astype(int))，确保为0/1二分类标签。

OOF预测结果 (oof\_binary\_preds\_final)：概率值经阈值判断后（布尔型）转换为整型（0或1）。

提交文件中的ID列 (user\_id, merchant\_id)：被转换为整型 (.astype(int))。

**3.5 定性变量数值化处理**

对于本质上是定性/类别型但以数字形式存储的变量，脚本采取了特定方式处理，以确保模型能正确解读：

**user\_id (用户ID)**：

主要用作**连接键** (如合并 user\_info\_df) 和**分组键** (如在 user\_log\_df 中聚合用户行为)。

在最终输入模型时，user\_id 通常会从特征矩阵中**移除**，因为它基数过高，直接使用易导致模型记住特定用户的行为而不是学习普适的模式，进而导致过拟合。其信息已通过聚合特征（如用户历史行为统计）间接反映。

**age\_range (年龄段)**：

经过初步清洗（缺失值填充为0，值7/8合并为6）后。

被包含在 categorical\_feats\_lgbm 列表中，在训练LightGBM模型时作为**类别特征**处理。模型会学习不同年龄段的模式，而非数值大小关系。

**gender (性别)**：

经过初步清洗（缺失值填充为2，代表未知）后。

同样被包含在 categorical\_feats\_lgbm 列表中，作为**类别特征**供LightGBM模型使用。

**3.6 观察期差异处理**

针对训练数据用户观察期可能不足6个月的问题，脚本中通过开关 PROCESS\_OBSERVATION\_PERIOD 控制相应处理逻辑：

当PROCESS\_OBSERVATION\_PERIOD在运行选项中设置为TRUE时，脚本才会执行处理观察期差异的逻辑。

之后脚本会先检查我们衍生的user\_first\_abs\_day（用户首次活跃的绝对日期）是否存在于数据中且user\_log\_df['abs\_day'] (用户日志中的绝对日期) 不为空。

**识别**：通过比较用户的首次活跃日期 (user\_first\_abs\_day) 与一个基于全体日志最早日期和30天宽限期计算的动态阈值，来识别那些在训练周期内相对较晚开始活跃的用户，并标记为 is\_short\_observation\_train\_user，表示他们的完整观察期可能不足6个月。

**处理选项** (由 FILTER\_SHORT\_OBSERVATION\_USERS 开关控制)：

若为 True：直接从训练集中**移除**这些被标记的观察期不足的用户

移除后，is\_short\_observation\_train\_user 这个指示特征列也会被删除 (因为它对于被筛选后的训练集意义不大)。

这种策略的思路是，为了保证训练数据与预测目标（观察6个月）的一致性，直接去掉那些观察期明显不足的用户。

若为 False：**保留**这些用户，并将 is\_short\_observation\_train\_user 作为一个二进制特征（指示是否观察期不足）加入模型训练，让模型学习其潜在影响。该指示特征也会被加入 categorical\_feats\_lgbm 列表。作为类别类特征传递给LightGBM模型，他提供了一个关于训练样本中“新旧程度“的信号。如果在模型训练过程中评估到其对区分复购用户和非复购用户的贡献较高，模型就可以用这个特征来辅助预测。

这种策略的思路是，让模型自己去学习观察期长短可能带来的影响，或者在建模时对这部分用户进行特殊处理（例如，不同的权重），其好处在于简单直接，但可能比较依赖模型捕捉这种信号的能力以及该信号与预测目标的相关性。

**3.7 异常值处理**

**分箱变换**：将连续特征分箱后，极端值会被归入特定箱子，模型学习的是箱子的类别而不是原始的精确值，降低其数值大小的直接影响。

**对数变换**：对长尾分布的特征进行对数变换，可以压缩大值的数值，减少其杠杆作用。

**缺失值填充策略**：对于某些特征，如时间差或购买间隔，脚本中使用中位数进行填充，中位数本身对异常值不敏感。

**基于树的模型**：LightGBM等树模型本身对异常值的敏感度相对较低。

**4. 特征工程 (Feature Engineering)**

在完成初步的数据清洗与集成后，特征工程是连接原始数据和机器学习模型的关键桥梁。本项目的特征工程旨在从用户的基本信息、用户在平台上的历史行为日志以及商家自身的信息中提取有效特征，以准确预测用户在“双十一”后对特定商家是否会发生复购行为。核心思路是构建能够反映用户购买习惯、兴趣偏好、活跃程度、用户与商家互动深度以及商家吸引力等多维度信息的特征体系。

**4.1 特征衍生总体思路**

根据“大作业说明”的要求，本次特征工程致力于建立明确的衍生思路和清晰的特征层次结构，以克服原案例中特征较为零散的问题。总体策略是多层次、多角度地从现有数据中挖掘信息，主要包括：

**用户维度特征**：基于用户的基本画像信息和其在整个平台上的历史行为，构建用户静态画像特征（如年龄、性别）和动态行为特征（如总体活跃度、购买力、行为类型偏好、品类偏好等）。

**商家维度特征**：基于商家在平台上的历史交易和互动数据，构建商家画像特征（如商家被访问次数、被购买次数、商品丰富度、用户覆盖面等）。

**用户-商家交互维度特征**：这是预测特定用户是否会在特定商家复购的核心。基于用户与目标商家之间的历史互动日志，构建能够反映两者关系强度和互动模式的特征（如用户在该商家的购买次数、浏览商品种类数、末次互动时间等）。

**时间维度特征**：考虑到用户行为具有时效性，近期行为对未来决策影响更大，因此引入时间窗口特征，捕捉用户近期的行为动态。

**高级统计与衍生特征**：不仅使用基础的计数和nunique特征，还计算转化率（如点击到加购）、时间间隔统计（如平均购买间隔）、用户兴趣熵等更深层次的统计量，以增强特征的表达能力。

**数据驱动的辅助特征**：针对数据本身的特性，如训练集中用户观察期可能存在的差异，构建指示性特征（如is\_short\_observation\_train\_user）供模型参考。

**4.2 特征的层次结构**

为了清晰地组织和理解衍生出的众多特征，我们将它们构建为一个具有层次性的结构。该结构从基础的用户和商家画像信息出发，逐步扩展到用户在平台上的整体行为，再聚焦于用户与特定商家的互动，并融入时间动态信息。

**层次一：基础画像与属性特征**

**A. 用户人口统计学特征 (User Demographic Features)**：

age\_range (年龄段)。

gender (性别)。

*衍生思路*：这些是用户固有的基本属性，直接从用户信息表 (user\_info\_df) 获取，用于基础的用户分群和行为模式关联。

**B. 商家整体聚合特征 (Merchant Aggregate Features)**：

商家受欢迎度与覆盖度：如 m\_user\_nunique (交互独立用户数), m\_log\_count (总日志数)。

商家商品丰富度：如 m\_item\_nunique (商品ID数), m\_cat\_nunique (品类ID数)。

商家销售表现：如 m\_buy\_count (总购买次数), m\_buyer\_conversion\_rate (购买用户转化率)。

*衍生思路*：基于所有用户与该商家的历史互动日志聚合而成，衡量商家的市场规模、商品多样性和吸引购买的能力。热门、商品丰富并且销售表现较好的商家可能更容易促成复购。

**层次二：用户全局行为特征 (User Global Behavior Features)** 这些特征从 user\_log\_df 中按 user\_id 聚合，描述用户在整个平台的历史行为模式，不针对特定商家。

**A. 整体活跃度与交互广度**：如 user\_log\_count (用户总日志数), user\_merchant\_nunique (交互商家数), user\_time\_stamp\_nunique (活跃天数)。

**B. 行为时间规律**：如 user\_first\_abs\_day (首次活跃日), user\_last\_abs\_day (末次活跃日), u\_mean\_time\_diff (平均活动间隔)。

**C. 行为类型偏好**：如 user\_action\_type\_0/1/2/3\_count (各行为类型计数), user\_action\_type\_0/1/2/3\_ratio (各行为类型占比)。

**D. 购买与转化习惯**：如 u\_mean\_buy\_interval (平均购买间隔)，以及不同行为间的转化率 (见4.3.6节)。

**E. 兴趣特征**：如 user\_cat\_entropy (品类兴趣熵)。

*衍生思路*：全面刻画用户的活跃程度、购物习惯（是否喜欢浏览多样商品/品类/店铺）、行为偏好（更倾向于点击还是直接购买）、购买力、购买规律性以及兴趣点的集中度。

**层次三：用户与目标商家互动特征 (User-Merchant Interaction Features)** 这些特征是预测的核心，基于 user\_log\_df 按 ['user\_id', 'merchant\_id'] 聚合，描述特定用户与特定商家之间的互动历史。

**A. 互动强度与广度**：如 um\_log\_count (用户在该商家的日志数), um\_item\_nunique (用户在该商家交互的商品数)。

**B. 互动时间规律**：如 um\_first\_abs\_day (在该商家首次活跃日), um\_last\_abs\_day (在该商家末次活跃日), um\_mean\_time\_diff (在该商家平均活动间隔)。

**C. 互动行为类型偏好**：如 um\_action\_type\_0/1/2/3\_count (在该商家各行为计数), um\_action\_type\_0/1/2/3\_ratio (在该商家各行为占比)。

**D. 相对互动强度**：如 um\_active\_days\_ratio\_in\_user (在该商家活跃天数占用户总活跃天数比例)。

*衍生思路*：精细刻画用户对特定商家的关注度和互动历史。用户与某商家互动越多、越深入，未来在该商家复购的可能性通常越高。

**层次四：时间动态与近期行为特征 (Temporal Dynamic Features)**

**A. 用户近期全局行为**：如 u\_logs\_last\_7d (用户最近7天日志数), u\_buys\_last\_30d (用户最近30天购买数) (如果 ADD\_TIME\_WINDOW\_FEATURES 为 True)。

**B. 用户在目标商家近期互动**：如 um\_logs\_last\_7d (用户在该商家最近7天日志数), um\_buys\_last\_30d (用户在该商家最近30天购买数) (如果 ADD\_TIME\_WINDOW\_FEATURES 为 True)。

*衍生思路*：通过限定时间窗口（如最近7、15、30天）内的行为日志进行聚合，捕捉用户行为的近期变化趋势，因为近期行为对未来决策的影响通常更。

**层次五：辅助及变换特征 (Auxiliary and Transformed Features)**

**A. 观察期指示特征**：如 is\_short\_observation\_train\_user，用于标记训练集中观察期可能不足的用户 (如果 PROCESS\_OBSERVATION\_PERIOD 且不 FILTER\_SHORT\_OBSERVATION\_USERS)。

**B. 数据变换特征** (见第5章 数据变换)。

*衍生思路*：前者是为了处理数据本身的特定问题（如观察期差异），后者是为了改善特征分布以适应模型需求。

**4.3 主要特征类别详解**

**4.3.1 用户人口统计学特征**

基本人口统计学特征指的是描述个体或群体基本背景信息的特征，这些信息通常是相对稳定或变化缓慢的。在本项目中，主要包括：

age\_range (年龄段)：描述用户的年龄范围。

gender (性别)：描述用户的性别。 这些特征从 user\_info\_df 获取，是用户画像的基础组成部分。

**4.3.2 用户整体行为特征**

**4.3.3 用户-商家互动特征**

**4.3.4 时间窗口特征**

**4.3.5 商家自身特征**

**4.3.6 其他衍生统计量**

**不同行为间的转化率 (Conversion rates between different actions)**：

转化率指用户从一种行为状态转变为另一种行为状态的比例或概率。

脚本中计算了以下用户层面的转化率 (如果 ADD\_COMPLEX\_CONVERSION\_RATES 为 True)：

user\_click\_to\_addcart\_ratio (用户点击到加购转化率)：user\_action\_type\_1\_count / (user\_action\_type\_0\_count + 1e-6)。衡量用户点击后将其加入购物车的倾向性。

user\_addcart\_to\_buy\_ratio (用户加购到购买转化率)：user\_action\_type\_2\_count / (user\_action\_type\_1\_count + 1e-6)。衡量用户加购后最终完成购买的倾向性。

user\_fav\_to\_buy\_ratio (用户收藏到购买转化率)：user\_action\_type\_2\_count / (user\_action\_type\_3\_count + 1e-6)。衡量用户收藏后最终完成购买的倾向性。

这些转化率有助于反映用户意图强度、衡量营销/产品吸引力，并可能与复购行为相关。

通过上述层次化的特征衍生，我们期望能够从不同粒度和角度捕捉影响用户复购的关键因素，为后续的模型训练提供高质量的输入。

**5. 数据变换 (Data Transformation for Modeling)**

为了进一步优化特征，使其更适合模型学习，脚本中包含了两种主要的数据变换方法：特征分箱和对数变换。

**5.1 特征分箱 (Feature Binning)**

分箱变换是一种将连续的数值型特征转换为离散的类别型特征的数据预处理技术。即将一个连续变量的取值范围划分成若干个“箱子”（bins），然后用这些箱子（或区间的代号）来替代原始的连续值。

**目的**：

处理非线性关系：帮助模型捕捉连续特征与目标变量间的非线性模式。

减少异常值/极端值的影响：将极端值归入特定箱子，增强模型鲁棒性。

处理缺失值（可将缺失值单独作为一个箱子，但本脚本中分箱前已填充）。

提升模型可解释性。

**脚本实现**：

通过 APPLY\_FEATURE\_BINNING 开关控制。

目标特征：cols\_for\_binning 列表中的计数值特征，如 user\_log\_count。这些特征虽然是数值型，但其不同取值的数量可能非常多，即具有“高基数”特性（“高基数”指特征拥有大量不同取值）。

方法：使用 sklearn.preprocessing.KBinsDiscretizer，采用等频分箱 (strategy='quantile')，分为5个箱子 (n\_bins=5)，输出为序数编码 (encode='ordinal')。

新特征以 \_binned 为后缀，并被视为类别特征处理。

分箱前，目标列的缺失值用该列中位数填充。

**5.2 对数变换 (Logarithmic Transformation)**

对数变换主要用于处理具有正偏态（长尾）分布的特征。

**目的**：

压缩特征的取值范围，使其分布更接近正态分布，有助于某些模型的学习和稳定性。

减小极端大值在模型中的影响。

**脚本实现**：

通过 APPLY\_LOG\_TRANSFORM 开关控制。

目标特征：cols\_for\_log\_transform 列表中的计数值特征，如 user\_log\_count 及各行为类型计数。

方法：应用 np.log1p 函数 (计算 log(1+x)), 能够处理包含0的值。

新特征以 \_log1p 为后缀。

变换前，缺失值用0填充，并通过 .clip(lower=0) 确保所有值非负。

通过这些数据变换操作，旨在优化特征的分布和表达，从而提升后续模型的预测性能。