真值处理流程与验证结果分析

September 18, 2025

1 真值处理流程梳理

1.1 数据清洗与标准化

做了什么： 从 `movies.csv` 与 `reviews.csv` 读入数据，处理缺失与异常字符串，统一数值类型；将评分人数（`rating_num`）、打分（`score`）、点赞数（`approvals_num`）转为数值；使用 `nickname` 作为 `user_id`，使用 `imdb_id` 作为 `item_id`，保证用户–电影关系唯一。
原理： 输入标准化，避免脏数据影响统计特性与后续建模。
目的： 降低缺失/异常/格式不统一引起的统计偏差。

1.2 电影质量真值（Item Quality）

做了什么： 采用贝叶斯平滑对电影平均星级进行校准：

$$s_{\text{hat}} = \frac{mC + R \cdot N}{C + N},$$

其中 R 为电影原始平均星级（0–5），N 为评分人数，m 为全局均值，C 为先验强度（在小样本条件下相当于"背负"C 个全局均值打分）。在长尾电影（N 小）的情况下，s_{hat} 被适度拉回均值以抑制极端评分的影响。
原理： 加权平均思想——样本量越小越依赖全局先验，样本量越大越接近观测均值。
目的： 提供更稳健的电影质量指标，降低长尾噪声并提升对比公平性。

1.3 用户–电影交互真值（Interactions Ground Truth）

做了什么： 显式评分规则：≥ 4 星记为正样本（$y = 1$），≤ 2 星记为负样本（$y = 0$），3 星丢弃；将点赞数作为置信度权重：

$$w = 1 + \log\bigl(1 + \text{approvals}\bigr),$$

并做均值归一化（整体均值 ≈ 1），防止个别高赞样本权重过大。
原理： 显式打分反映用户态度；更高的点赞意味着该评分更具代表性与可信度。
目的： 形成带标签且带置信度的用户–电影交互真值，用于监督学习。

1.4 时序切分（Train/Val/Test Splits）

做了什么： 按用户、按时间排序的正样本进行留后分割：每个用户最后一条正样本→test，倒数第二条（若存在）→val，其余→train。
原理： 模拟真实线上场景：用过去行为预测"未来会喜欢的电影"，避免信息泄漏。
目的： 让离线评估与线上表现一致。

1.5 评测样本构造（Eval Samples）

做了什么： 对每个 test 的正样本，按

$$p(i) \propto \text{popularity}(i)^{0.5} \times \text{quality}(i)^{\beta}$$

随机采样 $K = 50$ 个负例，并严格过滤（不包含正例本身与用户已看过的电影）。
原理： 负采样需要贴近真实候选集（偏热门，同时考虑质量）。
目的： 构建用于 HR@K、NDCG@K 的离线评测集合。

1.6 总体目标

核心原理： 统计平滑 + 点赞权重 + 时间切分，构造稳健可信的真值。
最终目的： 提供无信息泄漏的训练–验证–测试框架；为模型提供更可信的监督信号；输出干净的评测 ground truth（正例 + 随机负例）。

2 真值验证结果分析

2.1 数据一致性

结论： `item_quality` 主键唯一，`reviews` 关键列无缺失；但 `movies` 中存在一定数量的 `imdb_id` 重复，提示源数据有冗余。整体一致性较好。

2.2 电影质量真值

原始评分与平滑评分的相关系数接近 1.0，两者几乎重合；由于长尾电影占比偏小，平滑对整体排序影响有限，但保持了全局均值稳定，未引入偏移。图 1 为两者的散点对比。

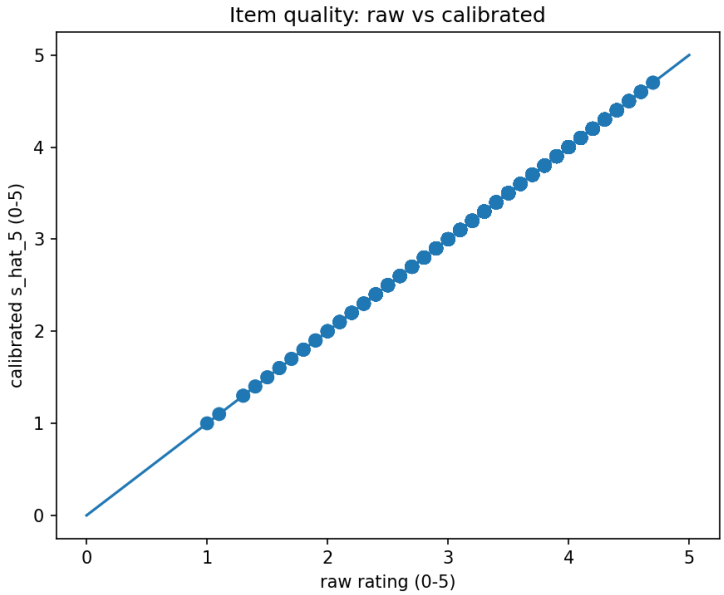

Figure 1: 原始评分 vs. 平滑评分散点图

2.3 交互真值

交互样本规模约 4.44×10^4，其中正例约占 83%、负例约占 17%；点赞权重均值 ≈ 0.98、95 百分位 ≈ 1.27、最大值 ≈ 1.41，分布平稳，无极端值。图 2 展示了权重分布。

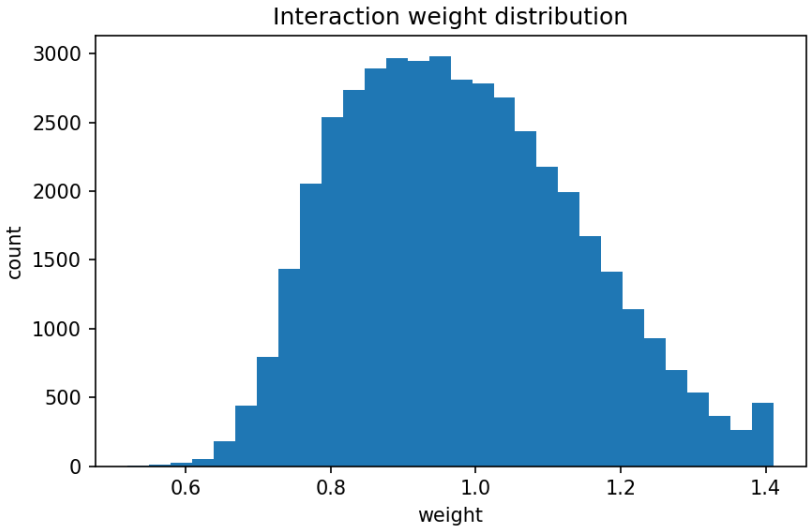

Figure 2: 交互权重分布直方图

2.4 时序切分

按用户留后一切分后，train/val/test 约为 83%/5%/12%；100% 的用户最后一条正样本位于 test，约有近一半用户缺少倒数第二条正样本导致 val 规模较小。图 3 展示了比例分布。

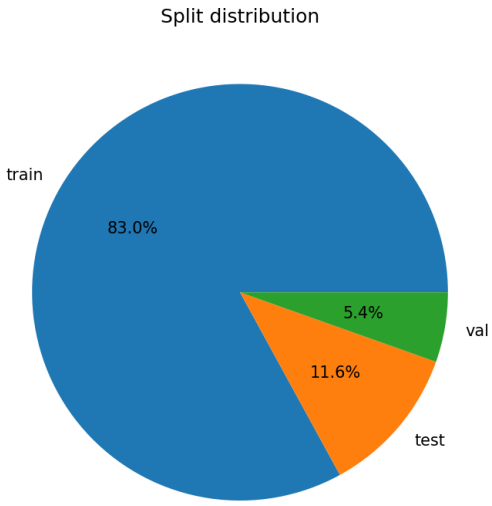

Figure 3: 时序切分比例分布

2.5 评测样本

评测样本覆盖率 100%，每个用例配足 $K = 50$ 个负例，且无违规（负例等于正例或落入用户已看集合）。负例的流行度分布与全体物品相似、略偏热门，符合推荐候选集特性。图 4 为负例与全体的流行度对比。

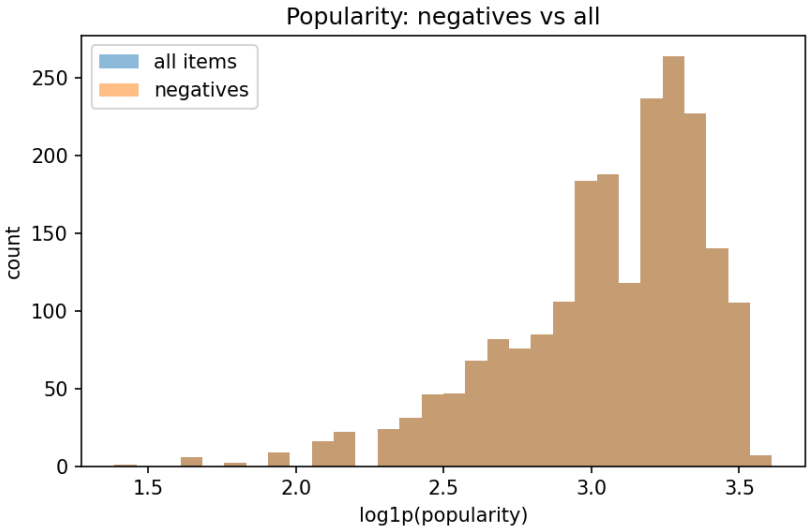

Figure 4: 负例与全体电影的流行度分布对比

2.6 总体结论

流水线干净、稳健、可复现：清洗规范、交互可信、切分合理、评测样本严格；主要风险在于源数据的 `imdb_id` 冗余与长尾样本偏少（使平滑优势未完全显现）。