# 第 3 章 推荐系统评价体系

从前两章我们知道，推荐系统是一个复杂的系统工程。而对于推荐系统的优化通常是为不同业务的北极星指标服务的。因此，在推荐系统的优化过程中，建立科学有效的评价体系是至关重要的。推荐系统的评价体系主要包括离线评价指标、在线评价指标和用户体验指标三个层面。离线评价指标主要用于模型开发阶段的效果评估，在线评价指标则用于模型上线后的 AB 测试评估，而用户体验指标则关注推荐结果对用户真实满意度和长期忠诚度（即平台对用户粘性）的影响。

## 3.1 离线评价指标

离线评价体系是推荐系统优化过程中的核心组成部分。模型训练本质上是在不断优化某种目标函数，而评价指标则决定了模型优化的方向。一个设计合理的评价体系能够帮助算法工程师快速判断模型优劣，提高模型迭代效率，并降低在线实验的试错成本。

在工业界，一个完整的推荐模型通常需要经历数据准备、模型训练、离线评估、在线实验以及最终上线等多个阶段。其中，**离线评估（Offline Evaluation）**承担着模型筛选和效果验证的重要职责。相比成本较高且具有一定业务风险的在线 A/B 实验，离线评估能够在不影响真实用户体验的前提下快速验证模型效果，因此成为推荐算法优化过程中最常使用的评估手段。

对于大多数模型预估任务（如点击率、转化率、点赞率、评论率等），本质上都可以建模为二分类问题，即**预测用户是否会发生某种特定行为**。传统机器学习中的分类评估指标同样适用于推荐模型的效果评估。然而，推荐系统最终服务于排序场景，仅仅关注分类正确率是不够的，还需要进一步衡量模型的排序能力以及概率预估能力。因此，工业界通常从分类指标、排序指标和概率校准指标三个维度来对模型进行综合评估。

### 3.1.1 分类指标

分类指标主要用于衡量模型对于正负样本的区分能力。

- **Accuracy（准确率）**

  Accuracy 表示模型所有预测结果中预测正确的比例，其计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

  其中，$TP$、$TN$、$FP$ 和 $FN$ 分别表示真正例、真负例、假正例和假负例。

  Accuracy 是机器学习中最基础的评价指标之一，但在推荐系统中往往参考价值有限。原因在于推荐场景通常具有极度不平衡的数据分布。例如，短视频场景的点击率仅为 1%，即使模型将所有样本全部预测为未点击，其 Accuracy 仍然能够达到 99%。因此，高 Accuracy 并不意味着模型具备良好的推荐能力。在实际工业场景中，Accuracy 通常不会作为模型选择依据，而更多作为辅助观察指标。

- **Precision（精确率）**

  Precision 表示模型预测为正样本的结果中，真实为正样本的比例：

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

  Precision 反映了模型预测结果的准确程度。Precision 越高，说明模型产生的误报（False Positive）越少。在推荐场景中，可以理解为模型推荐给用户的内容中，有多少真正会被用户点击或转化。

- **Recall（召回率）**

  Recall 表示所有真实正样本中被模型成功识别出来的比例：

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Recall 反映了模型发现正样本的能力。Recall 越高，说明模型遗漏的正样本越少。在召回阶段或候选集生成阶段，Recall 通常是重点关注指标，因为过低的 Recall 会导致大量潜在感兴趣内容在后续排序阶段之前就被过滤掉。

- **F1 Score**

  F1 Score 是 Precision 与 Recall 的调和平均值，其计算公式为：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.4}$$

F1 能够综合衡量模型的查准率和查全率。当 Precision 和 Recall 存在明显冲突时，F1 能够给出更加均衡的评价结果。在推荐系统中，F1 虽然不像 AUC 那样被广泛使用，但在召回模型评估、冷启动任务以及样本极度不均衡的场景中仍具有一定参考价值。

### 3.1.2 排序指标

推荐系统最终向用户展示的是一个排序列表，因此相比分类指标，排序指标更加符合推荐系统的实际优化目标，也是工业界最常使用的离线评估指标。

- **AUC（Area Under ROC Curve）**

  AUC 用于衡量模型将正样本排序在负样本之前的能力。从概率角度来看，可以表示为：

$$AUC = P(s_{pos} > s_{neg}) \tag{3.5}$$

其中，$s_{pos}$ 和 $s_{neg}$ 分别表示正样本和负样本的预测分数。AUC 的取值范围为 $[0, 1]$。

  - 当 AUC=0.5 时，模型与随机排序等价；
  - 当 AUC 接近 1 时，说明模型具有较强的排序能力；
  - 当 AUC 接近 0 时，说明模型排序结果几乎完全错误。

  从直观角度来看，AUC 可以理解为随机抽取一个正样本和一个负样本时，模型将正样本排在负样本之前的概率。AUC 最大的优点在于其不依赖具体阈值，并且对样本不平衡问题具有较强鲁棒性，因此长期以来一直是 CTR 预估模型最重要的离线评价指标之一。然而，AUC 关注的是全局排序关系，并不直接反映用户最终看到的 Top-K 推荐结果质量，因此通常需要结合 NDCG 等指标共同评估。

- **UAUC（User-wise AUC）**

  传统 AUC 会将所有样本统一计算，容易受到高活跃用户的影响。为了更加真实地反映用户维度的排序效果，工业界提出了 UAUC 指标。其首先对每个用户单独计算 AUC，然后再进行平均：

$$UAUC = \frac{1}{N} \sum_{u=1}^{N} AUC_u \tag{3.6}$$

其中，$AUC_u$ 表示用户 $u$ 对应的 AUC 值。UAUC 能够有效避免少数高活跃用户的大量行为数据主导整体评估结果，从而更加公平地衡量不同用户群体的体验。

- **GAUC（Group-wise AUC）**

  GAUC 是在 UAUC 基础上的进一步改进，其根据用户样本量进行加权平均：

$$GAUC = \frac{\sum_{u=1}^{N} n_u AUC_u}{\sum_{u=1}^{N} n_u} \tag{3.7}$$

其中，$n_u$ 表示用户 $u$ 对应的样本数量。相比 UAUC，GAUC 既保留了用户维度评估思想，又能够反映真实流量分布，因此更符合工业生产环境。目前在 CTR 预估领域，GAUC 已经成为应用最广泛的离线评价指标

之一。在工业界实践中，GAUC 有时候也被称之为 **WUAUC（Weighted UAUC）**，虽然具体定义可能存在细微差异，但本质上都是基于用户样本维度的加权 AUC 指标。

- **NDCG（Normalized Discounted Cumulative Gain）**

  AUC 关注的是全局排序能力，而 NDCG 更加关注推荐列表头部位置的排序质量。在实际推荐场景中，用户通常只会浏览推荐列表前若干条内容。因此，越靠前的位置越重要。

  首先定义 DCG 指标：

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \tag{3.8}$$

  其中，$rel_i$ 表示位置 $i$ 处内容的相关性得分。可以看到，DCG 通过对数衰减项对排名位置进行加权，使得前部位置具有更高权重。进一步利用理想排序下的 DCG 进行归一化：

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{3.9}$$

  NDCG 取值范围为 $[0, 1]$，越接近 1 表示排序结果越接近理想排序。由于 NDCG 指标建模了推荐中存在的位置偏置 (Position Bias)，因此在搜索排序、首页推荐以及重排阶段，NDCG 通常比 AUC 更能反映真实用户体验，因此被广泛应用于 Top-K 排序质量评估。

这里需要为读者强调一下 AUC、UAUC 和 GAUC 之间的区别。AUC 通常是将所有用户的样本同等对待进行计算，但这样容易受到少数高活跃用户的影响。假设：普通用户每天可能只产生 10 次曝光；而高活跃的用户每天能产生 1000 次曝光。如果直接计算全局 AUC，则高活跃用户贡献的大量样本将会主导最终的 AUC 计算结果，使得模型评价更多反映少数活跃用户的行为特征，而无法真实反映整体用户体验。UAUC 和 GAUC 的引入缓解了这一问题。其中，GAUC 先对每个用户单独计算 AUC，然后再根据用户在推荐系统中的实际交互样本量进行加权平均，这样能够更公平地衡量不同用户群体的体验。UAUC 强调用户公平性；GAUC 强调流量真实性。因此，在工业界实践中，GAUC 已经成为应用最广泛的离线评价指标之一。

### 3.1.3 概率校准指标

优秀的推荐模型不仅需要具备良好的排序能力，还需要输出具有真实概率意义的预测值。在推荐系统中，模型输出的预测概率往往会参与后续的流量分配、多目标融合以及广告竞价计算。因此，仅关注排序能力是不够的，还需要评估模型的概率校准能力（Calibration）。

- **PCOC（Predicted Click over Observed Click）**

  PCOC 用于衡量模型预测概率与真实概率之间的一致性，其计算公式为：

$$PCOC = \frac{\sum_i \hat{y}_i}{\sum_i y_i} \tag{3.10}$$

  其中，$\hat{y}_i$ 表示模型预测 CTR；$y_i$ 表示真实点击标签。

理想情况下，$PCOC = 1$，当 $PCOC > 1$ 时，说明模型整体高估了点击概率；当 $PCOC < 1$ 时，说明模型整体低估了点击概率。例如，模型预测总点击数为 500 次，而实际点击数为 400 次，则：$PCOC = \frac{500}{400} = 1.25$ 说明模型整体高估了 25% 的点击概率。

需要注意的是，高 AUC 并不一定意味着良好的概率校准能力。假设某模型对 3 个候选物品的预测值为：

$$[0.50,\ 0.25,\ 0.05] \tag{3.11}$$

而真实的 CTR 分别为：

$$[0.10,\ 0.05,\ 0.01] \tag{3.12}$$

虽然预测顺序完全正确，因此 AUC 接近 1，但所有预测值都被系统性放大了 5 倍，此时 PCOC 将远离 1。这说明模型虽然具备优秀的排序能力，但概率预估能力较差。

在广告推荐系统中，广告排序模型预测出的点击率 (pCTR) 通常直接参与竞价排序：

$$eCPM = Bid \times pCTR \tag{3.13}$$

如果 pCTR 存在系统性偏差，将直接影响广告排序结果以及平台收益。因此，工业界通常会同时监控 AUC（或 GAUC）与 PCOC，以分别评估模型的排序能力和概率校准能力。此外，部分公司和论文也将类似指标称为 **COPC（Calibration Of Predicted CTR）**。虽然具体定义可能存在细微差异，但本质上都用于衡量模型预测概率与真实概率之间的匹配程度。

总结起来，不同的离线评价指标关注模型性能的不同侧面。AUC、GAUC 以及 NDCG 主要衡量模型的相对排序能力，而 PCOC 则衡量模型的绝对概率预估能力。从工业实践角度来看：

- GAUC 通常用于衡量 CTR 模型整体排序能力；
- NDCG 用于评估 Top-K 推荐质量；
- PCOC 用于监控模型概率校准效果；
- Precision、Recall 以及 F1 更多用于召回模型评估。

一个优秀的推荐模型不仅需要将用户感兴趣的内容排在前面，还需要保证预测概率能够准确反映真实用户行为。因此，工业界通常采用多指标联合评估策略，从排序能力、用户体验以及业务收益等多个不同的维度去综合评价模型的质量。

**笔记**

需要特别强调的是，离线评估指标的提升并不必然意味着线上业务指标一定能够获得收益。

在实际工业场景中，我们经常会观察到这样的现象：某个新模型相比 Base 模型离线 AUC、GAUC 甚至 LogLoss 均有明显提升，但上线 A/B 实验后，CTR、时长、GMV 等核心业务指标却没有显著增长，甚至可能出现下降。这种现象通常被称为离线效果与线上效果之间的 **Gap（Offline-Online Gap）**。

造成这一问题的首要原因在于离线评估数据与线上推理数据之间存在天然的数据分布差异。离线测试集通常来源于历史日志，仅包含已经被推荐系统曝光过且获得用户反馈的样本。而在线推荐服务执行时，需要面对的是整个候选库中的海量内容，其中绝大部分内容从未被曝光过，也没有任何真实反馈数据。因此，离线评估所覆盖的样本空间只是线上候选空间的一个子集，两者往往存在明显的分布偏移（**Distribution Shift**）。当模型能力提升主要体现在历史曝光样本上，而无法泛化到更广泛的候选内容时，即使离线 GAUC 有所提升，线上业务指标也未必能够同步增长。

除此之外，现代推荐系统通常采用多阶段架构。从召回、粗排、精排到重排，再到最终的策略后处理，模型输出的预测分数往往只是整个决策链路中的一个环节。在精排模型之后，系统还可能引入多样性打散、作者去重、内容生态调控、商业化约束、风险控制以及运营干预等大量后置策略。即使新模型能够为某些内容赋予更高的预测分数，这些内容最终也未必能够获得更多曝光机会。例如，模型认为最优的内容可能因为多样性约束被降权，也可能因为商业化目标或生态目标而被后续模块重新排序。此时，模型能力的提升会在复杂的后链路中被部分削弱甚至完全抵消，从而导致线上实验结果不明显。

从更本质的角度来看，离线指标衡量的是模型在既定数据集上的拟合能力，而线上指标衡量的则是模型对于真实用户行为和业务目标的最终影响。两者虽然相关，但并不等价。因此，在推荐系统研发过程中，离线评估的价值更多体现在模型筛选和快速迭代阶段，用于判断模型是否具备进入线上实验的潜力；而模型是否真正有效，最终仍需要通过在线 A/B 实验进行验证。推荐算法本质上是一门实验科学（Experimental Science），任何离线指标都无法完全替代真实线上反馈。

换句话说，离线指标决定了模型是否有资格进入赛场，而在线实验结果才决定了模型能否真正赢得比赛。

## 3.2 在线评价指标

上一节介绍了推荐系统中常用的离线评价指标，如 AUC、GAUC、NDCG 以及 PCOC 等。这些指标能够帮助算法工程师快速评估模型的排序能力与概率预估能力，是模型迭代过程中不可或缺的重要工具。然而，正如前文所讨论的那样，离线指标的提升并不一定能够直接转化为线上业务收益。由于离线数据与线上流量之间存

在分布差异（Distribution Shift），同时推荐系统内部还存在召回、粗排、精排、重排以及策略后处理等复杂链路；因此，一个离线效果更优的模型最终未必能够为平台创造更高的业务价值。

从业务角度来看，推荐系统存在的根本目标并不是提升 AUC 或者 GAUC，而是提升用户价值与平台价值。例如：

- 内容平台希望提升用户消费时长（Watch Time）；
- 社区产品希望提升互动率（点赞、评论、分享等）；
- 电商平台希望提升成交额（GMV）；
- 广告平台希望提升广告收益（Revenue）。

因此，对于推荐系统而言，真正决定模型价值的并非离线指标，而是线上评价指标（Online Metrics）。在线评价指标直接来源于真实用户行为反馈，能够更加准确地反映模型对于用户体验和业务目标的实际影响。因此，在工业界推荐系统研发流程中，离线评估负责筛选候选模型，而在线 A/B 实验则负责验证模型的最终价值。

在介绍在线评价指标之前，我们首先需要了解推荐系统中最常见的一类概念，即 **pxtr（Predictive X Through Rate）**。在工业级推荐系统中，推荐模型本质上是为业务目标服务的。然而，实际业务场景中的优化目标往往并非单一指标，而是多个目标之间的综合权衡与协同优化。例如，平台既希望提升用户消费时长，也希望提高互动率、转化率以及商业化收益等指标。与此同时，用户在 APP 中的行为反馈也是多维度的，包括曝光、点击、观看、点赞、评论、分享、关注、购买等不同层次的行为。

业务侧的多目标需求与用户侧的多行为反馈共同决定了现代推荐模型采用**多任务学习（Multi-Task Learning，MTL）**的建模范式，即通过统一模型同时预测多个用户行为目标。因此，在推荐系统的各个排序阶段，模型往往会输出大量不同类型的预估值，用于刻画用户发生某种行为的概率或对应收益。业界通常将这些预估值统称为 pxtr。其中，$X$ 表示具体的用户行为类型。例如，点击对应 CTR（Click-Through Rate）、点赞对应 LTR（Like Through Rate）、评论对应 CMR（Comment Rate）、转化对应 CVR（Conversion Rate）等。

需要注意的是，pxtr 本质上是模型输出的概率预估值，而 CTR、CVR、Watch Time 等在线评价指标则是真实线上流量产生的统计结果。前者属于**先验的模型预测结果**，后者属于**后验的业务观测结果**。推荐系统的核心目标之一，就是通过不断提升 pxtr 预估精度，最终带动对应线上业务指标的增长。

表3.1汇总了推荐系统中常见的 pxtr 指标及其业务含义。从表3.1可以看出，不同的 pxtr 指标实际上对应着用户价值链路中的不同阶段。从用户行为发生顺序来看，这些目标通常具有天然的漏斗结构（Funnel Structure）：

$$Impression \rightarrow Click \rightarrow Watch \rightarrow Interact \rightarrow Coversion \rightarrow Pay \tag{3.14}$$

其中：

- CTR 反映用户是否愿意进入内容；
- EVR、LVR、Watch Time 反映用户是否真正消费内容；
- LTR、CMR、WTR 等互动指标反映用户参与程度；
- CVR 反映用户是否完成目标转化；
- GTR 则进一步反映用户付费意愿和商业价值。

从推荐系统演进历史来看，行业对于优化目标的关注重点也经历了明显变化。在早期资讯推荐阶段，由于用户行为相对简单，CTR 几乎是唯一核心优化目标。随着短视频平台兴起，行业逐渐发现高 CTR 并不一定意味着用户真正喜欢内容。例如某些“标题党”内容虽然能够获得较高点击率，却无法带来持续观看行为。因此，Watch Time、EVR、LVR 等消费深度指标逐渐成为新的核心优化目标。

进一步地，随着社区生态和商业化需求的发展，点赞率、评论率、分享率、关注率、转化率以及 GMV 等指标的重要性也不断提升。现代推荐系统已经很少只优化单一目标，而是采用多目标优化（Multi-Objective Optimization）的方式，在用户体验、内容生态和商业收益之间寻找最优平衡点。

因此，当前工业界主流推荐模型往往会同时预测多个 pxtr 目标，并通过加权融合、价值模型（Value Model）或者长期价值建模（Long-Term Value Modeling）等方式构建最终排序分数。

表 3.1: 各类用户行为预估（pxtr）的定义及计算公式

| 缩写 | 定义及计算公式 |
|---|---|
| CTR | 点击率（Click-Through Rate）：用户点击行为占曝光次数的比例。<br>计算公式：$\text{CTR} = \frac{\text{点击次数（Clicks）}}{\text{曝光次数（Impressions）}} \times 100\%$ |
| LTR | 点赞率（Like Rate）：用户点赞行为占曝光次数的比例。<br>计算公式：$\text{LTR} = \frac{\text{点赞次数（Likes）}}{\text{曝光次数（Impressions）}} \times 100\%$ |
| CMR | 评论率（Comment Rate）：用户评论行为占曝光次数的比例。<br>计算公式：$\text{CMR} = \frac{\text{评论次数（Comments）}}{\text{曝光次数（Impressions）}} \times 100\%$ |
| WTR | 关注率（Follow Rate）：用户关注行为占曝光次数的比例。<br>计算公式：$\text{WTR} = \frac{\text{关注次数（Follows）}}{\text{曝光次数（Impressions）}} \times 100\%$ |
| EVR | 有效播放率（Effective View Rate）：有效播放次数占总播放次数的比例（通常以播放时长达到阈值为有效标准）。<br>计算公式：$\text{EVR} = \frac{\text{有效播放次数（Effective Plays）}}{\text{总播放次数（Total Plays）}} \times 100\%$ |
| LVR | 长播率（Long View Rate）：长时播放次数占总播放次数的比例（通常以播放完成为长播标准）。<br>计算公式：$\text{LVR} = \frac{\text{长时播放次数（Long Plays）}}{\text{总播放次数（Total Plays）}} \times 100\%$ |
| VTR | 播放时长率（Video Time Rate）：平均播放时长占内容总时长的比例。<br>计算公式：$\text{VTR} = \frac{\text{总播放时长（Total Watch Time）}}{\text{内容总时长（Content Duration）} \times \text{播放次数（Plays）}} \times 100\%$ |
| CVR | 转化率（Conversion Rate）：用户完成目标转化行为（如购买、下载）占点击次数的比例。<br>计算公式：$\text{CVR} = \frac{\text{转化次数（Conversions）}}{\text{点击次数（Clicks）}} \times 100\%$ |
| GTR | 打赏率（Gratuity Rate）：用户打赏行为占曝光次数或播放次数的比例。<br>计算公式：$\text{GTR} = \frac{\text{打赏次数（Gratuities）}}{\text{曝光次数（Impressions）或播放次数（Plays）}} \times 100\%$ |

## 3.3 A/B 实验体系

在前两节中我们已经讨论过，离线评价指标（如 AUC、GAUC、NDCG 等）的提升并不必然意味着线上业务指标一定能够获得收益。造成这一现象的原因在于离线数据与线上流量之间存在天然的数据分布差异，同时推荐系统内部还存在召回、粗排、精排、重排以及策略后处理等复杂链路，任何一个环节都可能影响最终用户体验。

因此，对于推荐系统而言，真正需要回答的问题并不是“模型是否更准确”，而是：

笔记 新的推荐模型或策略是否能够在真实用户环境下创造更高的业务价值？

而回答这一问题最可靠的方法，便是在线 A/B 实验（A/B Testing）。在工业界推荐系统的算法迭代与产品优化过程中，无论是推荐模型升级、特征工程优化、召回策略调整，还是产品交互改版，最终都需要通过 A/B 实验进行验证。因此，A/B 实验被广泛视为推荐系统效果评估的黄金标准（Golden Standard）。

相比离线评估或回放测试（Backtest），A/B 实验直接面向真实用户和真实流量，能够观察用户在自然环境下产生的行为反馈，从而有效降低数据泄露、样本偏差以及模型过拟合等问题带来的影响，为产品决策提供更加可信的因果证据（Causal Evidence）。

### 3.3.1 A/B 实验基本原理

A/B 实验的核心思想非常简单：

**定义 3.1**

对于两组统计特征相同的用户，仅改变其中一组所接受的推荐策略，然后比较两组用户行为指标的差异。如果实验组的关键指标显著优于对照组，则可以认为新的推荐策略带来了正向收益。

一个标准的推荐系统 A/B 实验通常包含以下几个核心组成部分：

- **Base 组（Control Group，对照组）**
  运行当前线上稳定版本的推荐策略。Base 组代表系统当前最优实践（Best Practice），其指标表现作为所有实验的比较基准。
- **Exp 组（Experiment Group，实验组）**
  部署待验证的新策略，例如：新增一路召回源；模型新增特征实验；新排序策略；新的多目标优化方案等。
- **分流机制（Traffic Allocation）**
  系统需要将用户随机分配到不同实验组中。工业界通常基于 User ID，Device ID，Request ID 等进行哈希取模（Hash Bucketing）完成随机分流。例如：$hash(user_id) \bmod 100$ 若结果位于 $[0, 94]$，则进入 Base 组；反之，则进入 Exp 组。这样就形成了 $95\% : 5\%$ 的流量划分。随机分流的目标是确保实验组与对照组在统计意义上具有一致的数据分布，从而保证实验结果具有可比性。
- **实验指标（Online Metrics）**
  实验指标用于衡量策略对业务目标的影响。通常可分为：核心指标（North Star Metrics）和诊断指标（Guardrail Metrics）。核心指标直接反映业务价值，例如：CTR，Watch Time，Retention，GMV，Revenue 这些指标。而诊断指标则用于监控及观测作用，例如：内容多样性及聚集度；冷启动内容曝光率；负反馈 hate 率；系统延迟；客户端崩溃率等。

### 3.3.2 A/B 实验平台

随着推荐系统规模不断扩大，工业界通常会构建统一实验平台（Experiment Platform）来管理实验流量。一个典型实验平台包含：实验配置管理，流量分桶系统，指标计算系统，实验分析系统，风险控制系统。为了提高流量利用率，许多公司还会采用分层实验（Layered Experimentation）架构。例如：顶层 Holdout 实验，各业务 Combo 实验，普通世界实验等。不同层之间互不干扰，从而实现同一批用户流量被多个实验同时复用。

A/B 实验本质上是一种因果推断方法。为了保证实验结论可信，需要满足若干关键假设。第一个假设就是**随机性（Randomization）**随机分流是实验成立的前提。实验组与对照组之间不应存在系统性差异，例如：用户年龄差异，地域差异，新老用户比例差异，手机机型差异等。否则实验结果可能受到混杂因素（Confounding Factors）影响。第二个假设是**一致性（Sticky Assignment）**。同一个用户在整个实验周期内应始终处于同一个实验组。如果用户频繁切换实验组，将导致策略影响相互污染，从而降低实验可信度。第三个假设是**无干扰假设（SUTVA）**。理想情况下，一个用户的实验结果不应受到其他用户实验分组的影响。然而在推荐场景中，这一假设往往容易被破坏。例如：社交传播行为，热门内容竞争，全局流量池调整，内容供给变化。这些因素都可能导致实验产生溢出效应（Spillover Effect）。因此，大规模推荐系统通常需要额外设计隔离机制来降低实验干扰。

实验结束后，需要判断观察到的指标变化是否具有统计显著性。常见方法包括：t-检验，Bayesian A/B Test 等。除了显著性检验之外，还需要关注：P-Value，Confidence Interval 等统计学指标。如果实验流量过小或实验周期过短，则可能导致真实收益无法被检测出来。此外，当同时观察多个指标时，还需要进行多重检验校正（Multiple Testing Correction），以降低假阳性（False Positive）风险。

### 3.3.3 日志体系与数据质量保障

实验结论的可靠性最终取决于数据质量。推荐系统通常维护两类核心日志：

- **曝光日志（Impression Log）**曝光日志主要记录：用户 ID，请求 ID，推荐结果列表，排序位置，用户侧行为历史，实验组信息等。
- **行为日志（Action Log）**行为日志记录：用户点击、播放、点赞、评论、分享、转化等反馈行为。

通过 $req\_id$ 关联两类日志，可以形成完整的“曝光—反馈”闭环。与此同时，工业界通常还会建立：分流比例监控，日志丢失率监控，指标漂移监控，实验异常告警等，以确保实验结果真实可信。

### 3.3.4 A/B 实验迭代流程

在实际工作中，一个新策略通常不会直接全量上线，而是经历逐步扩量的过程。典型流程如下：

$$5\% \rightarrow 10\% \rightarrow 20\% \rightarrow 50\% \rightarrow 100\% \tag{3.15}$$

首先通过 5% 左右的小流量验证策略安全性。若核心指标获得统计显著提升，且诊断指标无明显恶化，则逐步扩大实验流量。当实验在更大流量规模下依然保持稳定收益，并持续运行多个用户行为周期后，才会执行全量发布（Full Rollout）。与此同时，许多公司还会保留约 1% 的长期 Holdout 流量作为永久对照组，用于观察系统长期演化趋势以及进行跨实验横向比较。

总结起来，推荐系统是一门典型的数据驱动学科，而 A/B 实验则是连接算法优化与业务价值的最终桥梁。离线评估帮助我们筛选模型，在线实验帮助我们验证价值；离线指标决定模型是否具备进入实验的资格，而 A/B 实验则决定模型是否真正值得上线。因此，对于推荐算法工程师而言，理解 A/B 实验不仅仅是理解一种评估工具，更是在学习如何通过科学实验方法，将模型能力转化为真实的用户价值与业务增长。

## 3.4 推荐系统监控体系

离线评估指标帮助我们判断模型是否具备上线潜力，在线指标与 A/B 实验帮助我们验证模型是否真正创造业务价值。然而，当一个推荐策略正式上线并承担线上流量后，工作并没有结束。推荐系统作为一个复杂的实时决策系统，涉及数据采集、特征生成、召回、排序、重排、混排以及最终内容分发等多个环节，任何一个环节出现异常，都可能导致推荐效果下降甚至业务事故。因此，建立完善的监控体系（Monitoring System）是推荐系统稳定运行的重要保障。

从工程实践角度来看，推荐系统监控通常覆盖数据层、模型层、服务层以及业务层四个维度。

### 数据与索引监控

数据质量是推荐系统的基础。大量线上事故最终都可以追溯到数据异常。因此，工业界通常会重点监控：特征缺失率（Feature Missing Rate）；特征分布漂移（Feature Distribution Drift）；索引更新延迟（Index Delay）；索引覆盖率（Index Coverage）；用户行为日志完整率等。例如，当视频时长、作者标签或类目标签等关键索引字段出现异常时，召回系统可能无法正确检索目标内容，从而导致整体推荐效果下降。

### 召回模块监控

召回阶段决定了推荐系统能够看到哪些候选内容，因此需要重点关注召回通道的供给情况。常见监控指标包括：各召回源召回数量；各召回源覆盖率；各召回源命中率；候选集 Duration 分布；不同内容类型占比；长尾内容覆盖率等。例如，当某一路协同过滤召回突然返回数量大幅下降时，可能意味着索引构建异常或特征计算出现问题。

### 排序模块监控

模型监控主要用于观察预测结果是否发生异常漂移。常见监控指标包括：CTR、CVR、LTR 等 pxtr 均值；pxtr 分位数分布（P50、P90、P99）；PCOC 校准指标；模型 AUC 是否掉到 0.5；模型是否出现 nan、inf 等异常值；模型输出值分布漂移；在线 GAUC 监控等。

在实践中，算法工程师往往会持续监控不同模型输出的 pxtr 分布。例如，若某次模型发布后 CTR 预测值整体提高 50%，而实际 CTR 并未同步提升，则可能意味着模型出现了严重的概率高估问题。

### 重排与混排模块监控

现代推荐系统通常采用多路内容混排架构，例如：图文内容；推荐内容；广告内容；电商内容；直播内容。因此需要重点监控：各内容池下发量；各内容池曝光流量占比；内容多样性指标；作者分布指标；类目分布指标；冷启动内容曝光占比等。这些指标能够帮助工程师判断混排策略是否符合产品预期，以及是否存在流量分配失衡的问题。

### 服务与稳定性监控

推荐系统通常需要在数十毫秒内完成在线推理，因此服务稳定性同样至关重要。常见监控包括：服务 QPS；平均响应时间（Latency）；P99 延迟；超时率；错误率；GPU 利用率等。当模型复杂度增加或流量突增时，这些指标能够帮助工程师及时发现系统瓶颈。

### 降级与兜底监控

推荐系统通常会设计多层容灾机制。例如：特征服务降级；模型降级；召回降级；热榜兜底；内容池兜底。因此需要持续监控：降级触发次数；降级流量占比；兜底内容占比；降级持续时长等。这些指标能够帮助团队快速发现潜在风险，避免线上服务出现大规模故障。

总体而言，监控体系的核心目标并非简单记录数据，而是帮助工程师及时发现问题、快速定位问题并持续优化系统。一个成熟的推荐系统往往拥有远比模型本身更加复杂的监控与告警体系，而这也是推荐系统能够长期稳定运行的重要保障。

## 3.5 本章小结

评价体系是推荐系统研发流程中的重要组成部分，也是连接模型优化与业务价值的关键桥梁。本章首先介绍了离线评价指标，包括 Accuracy、Precision、Recall、F1、AUC、GAUC、NDCG 以及 PCOC 等指标。其中，AUC 和 GAUC 主要用于衡量模型的排序能力，而 PCOC 则用于评估模型的概率校准能力。离线评估能够帮助算法工程师快速筛选模型方案，提高研发效率，但离线指标的提升并不一定能够直接转化为线上收益。

随后，本章介绍了在线评价指标及其对应的业务含义。现代推荐系统通常采用多任务学习架构，同时预测 CTR、CVR、Watch Time、点赞率、评论率等多种 pxtr 指标，并通过多目标优化实现用户价值、内容生态价值以及商业价值之间的平衡。

在此基础上，本章进一步讨论了 A/B 实验体系。A/B 实验通过随机分流、因果推断以及统计检验等方法，在真实线上环境中验证推荐策略对于业务指标的影响，是工业界公认的推荐系统效果评估黄金标准。对于推荐系统而言，离线评估决定模型是否具备进入实验的资格，而 A/B 实验则决定模型是否真正值得上线。

最后，本章介绍了推荐系统监控体系。一个成熟的推荐系统不仅需要具备准确的模型和科学的实验体系，还需要覆盖数据、召回、排序、混排、服务以及容灾等多个层面的监控能力。监控体系与离线评估、在线实验共同构成了推荐系统持续迭代和稳定运行的基础设施。

总体来看，推荐系统评价体系并非单一指标或单一工具，而是一套覆盖“离线评估—在线实验—线上监控”的完整闭环。只有将模型效果、用户体验和业务目标统一纳入评价框架，才能真正实现推荐系统的持续优化与长期价值增长。