DataFunTalk

分享主题

京东科技推荐算法探索与实践

李欣如 高级算法工程师

目录

CONTENTS

DataFunTalk

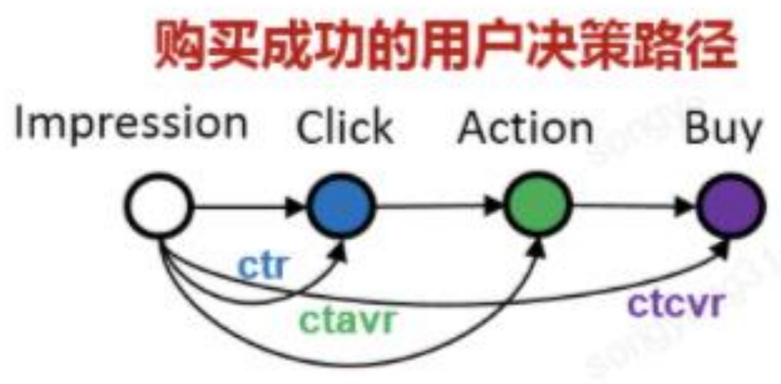
业务介绍

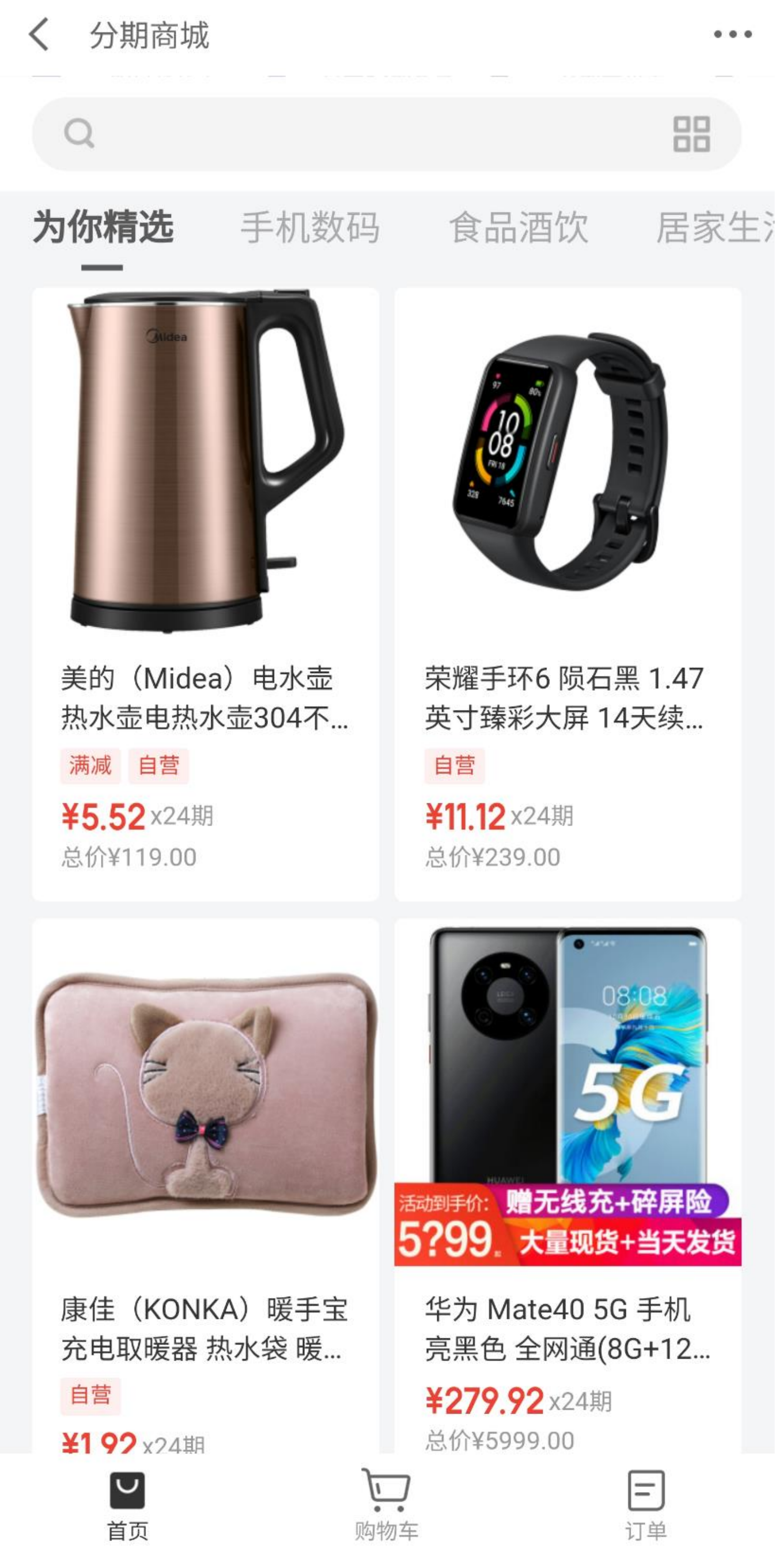
- **业务场景**

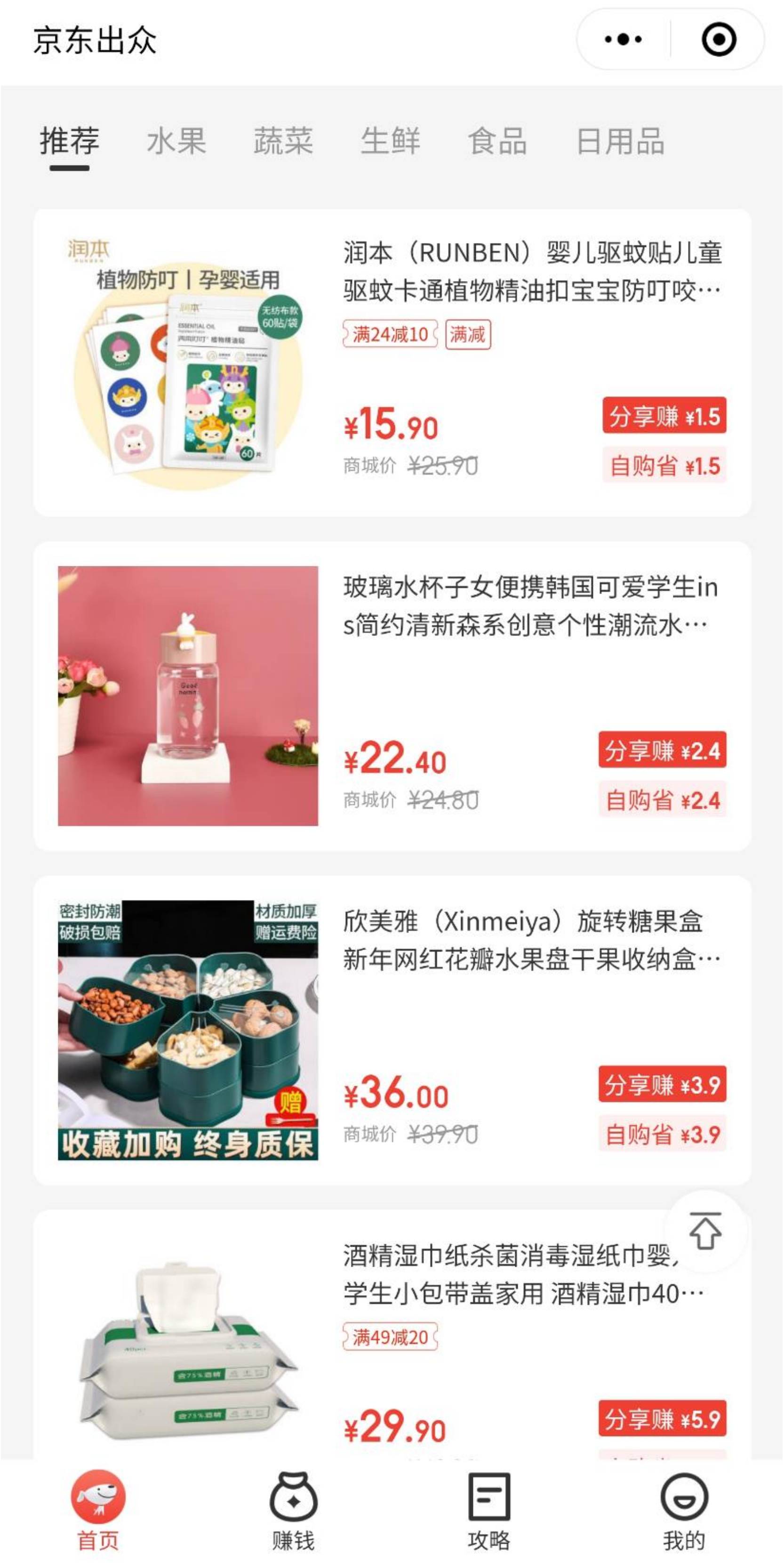
  分期商城页
  楼层推荐页
  京东出众页
  营销推荐页

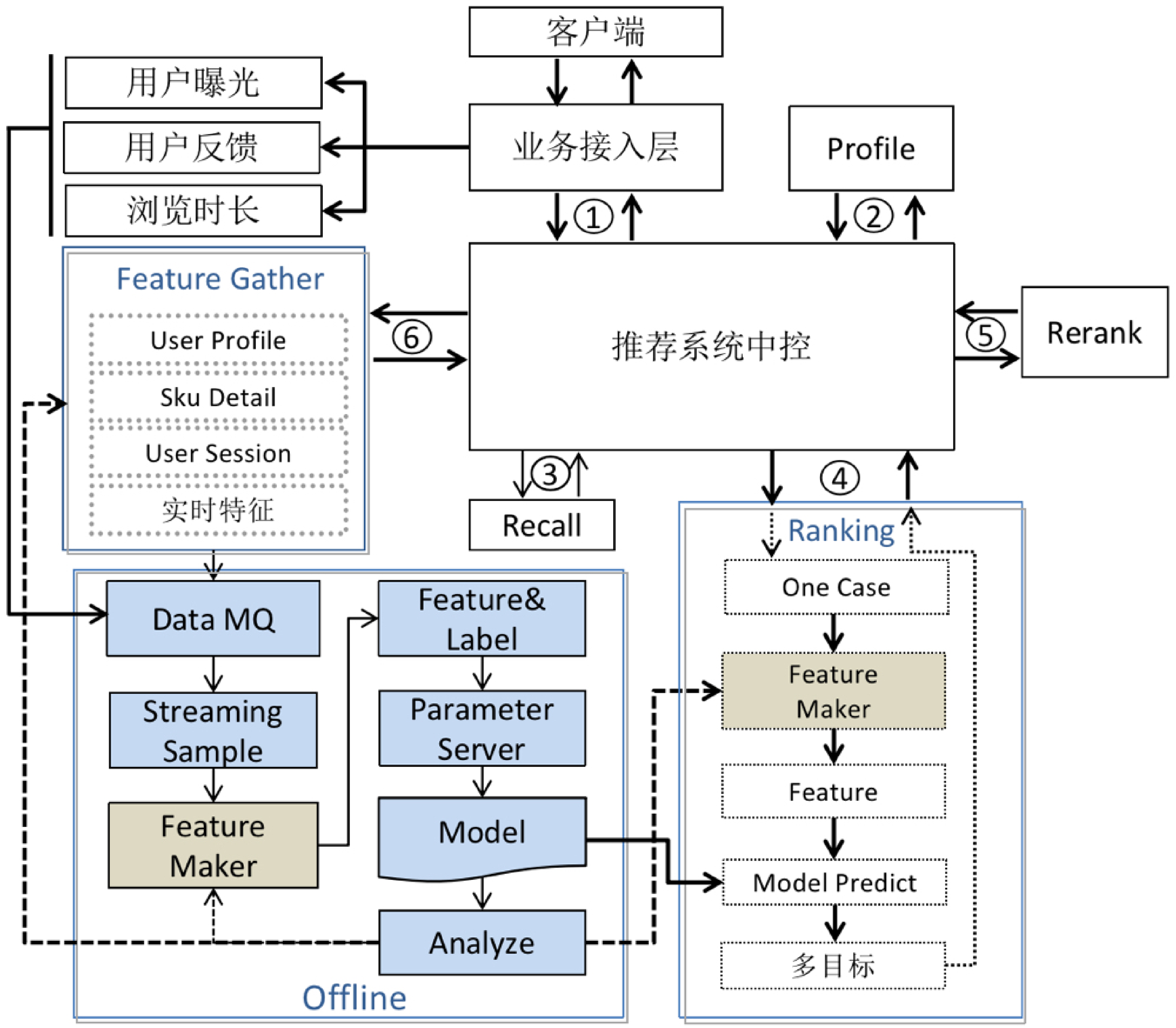
- **业务目标**

  CTR、GMV
  ARPU、单量
  加购、收藏
  浏览时长、分享









DataFunTalk

画像召回-优化

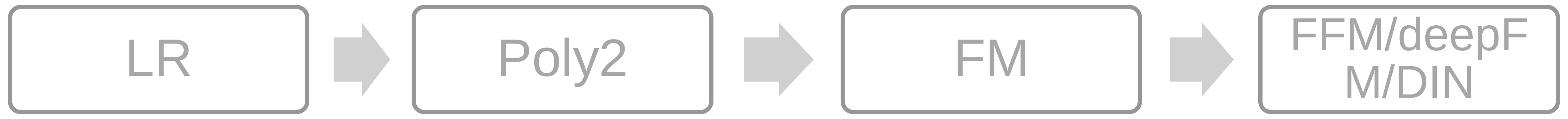
➢ 常规流程

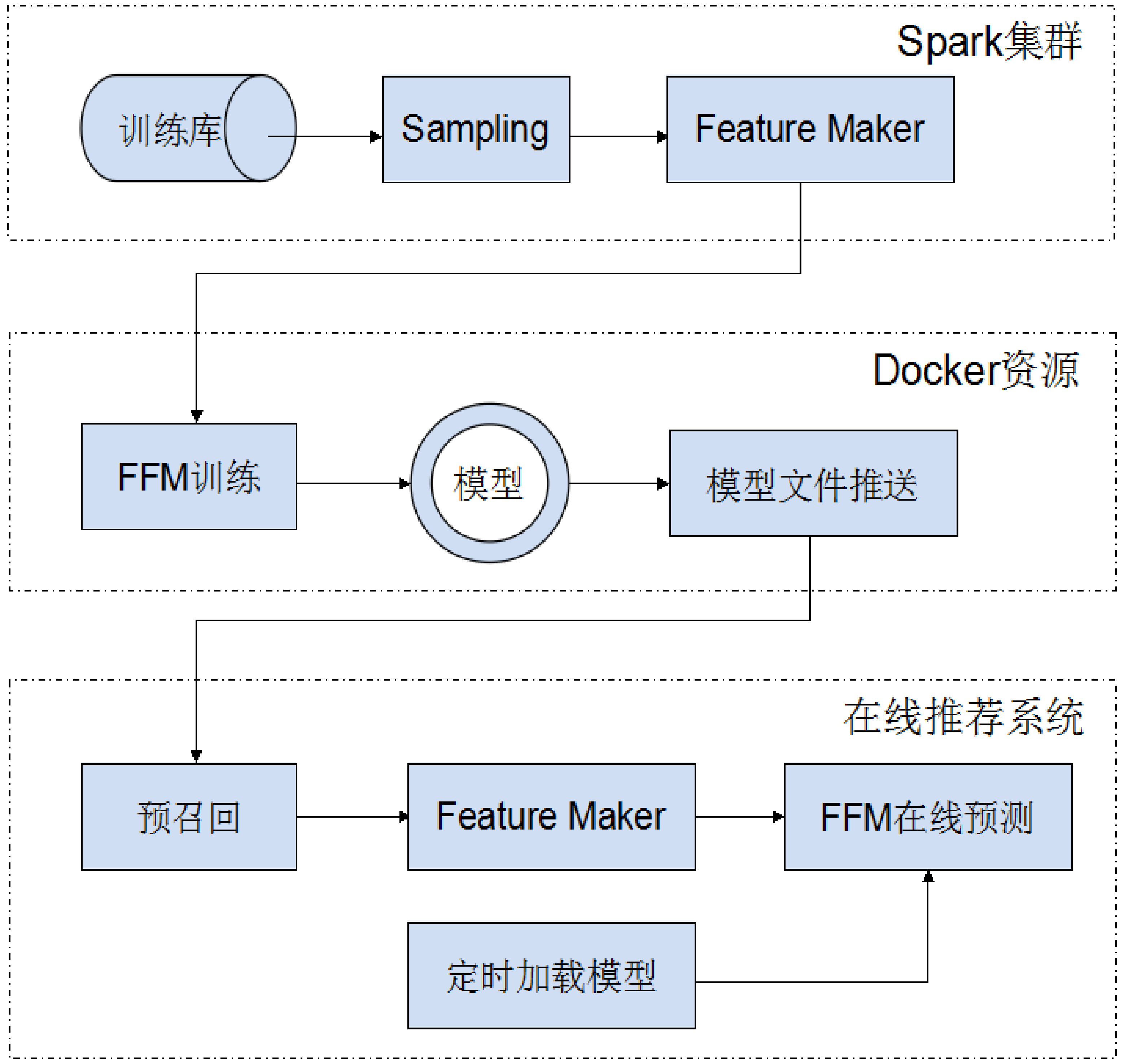
➢ 存在问题

1、画像类别多，召回量较大
2、使用规则排序效果差，ctr等指标较ICF低
3、人工规则难以针对性优化，无学习能力

➢ 优化难点

1、使用模型优化， LR、Ploy2、FM模型无法较好的挖掘特征交叉信息，需较多人工特征组合
2 、样本选择问题



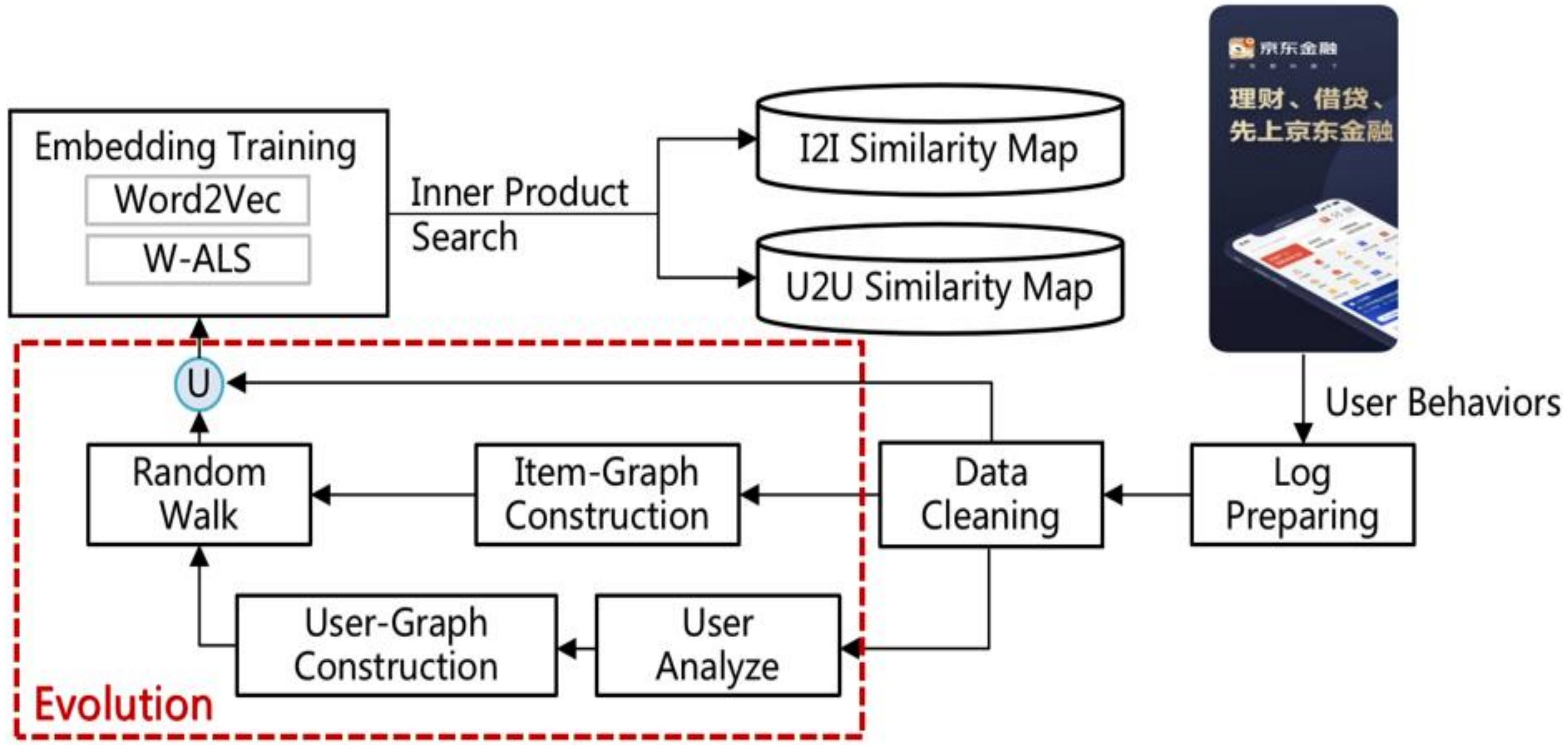


相似召回&评测

➢ **解决痛点**：较难挖掘item之间的高阶相似性

Embedding类的大概演化过程

MF/ALS → word2vec → tag2vec → Graph embedding



◆上线评测：点击率等指标

◆离线评测：模型评测、F1值

召回的物品集记作$\mathcal{P}_u(|\mathcal{P}_u| = M)$　真实的物品集记作$\mathcal{G}_u$。

$$Precision@M(u) = \frac{|\mathcal{P}_u \cap \mathcal{G}_u|}{M} \qquad Recall@M(u) = \frac{|\mathcal{P}_u \cap \mathcal{G}_u|}{|\mathcal{G}_u|}$$

$$F - Measure@M(u) = \frac{2 \cdot Precision@M(u) \cdot Recall@M(u)}{Precision@M(u) + Recall@M(u)}$$

◆评测成本较高，其他方式?

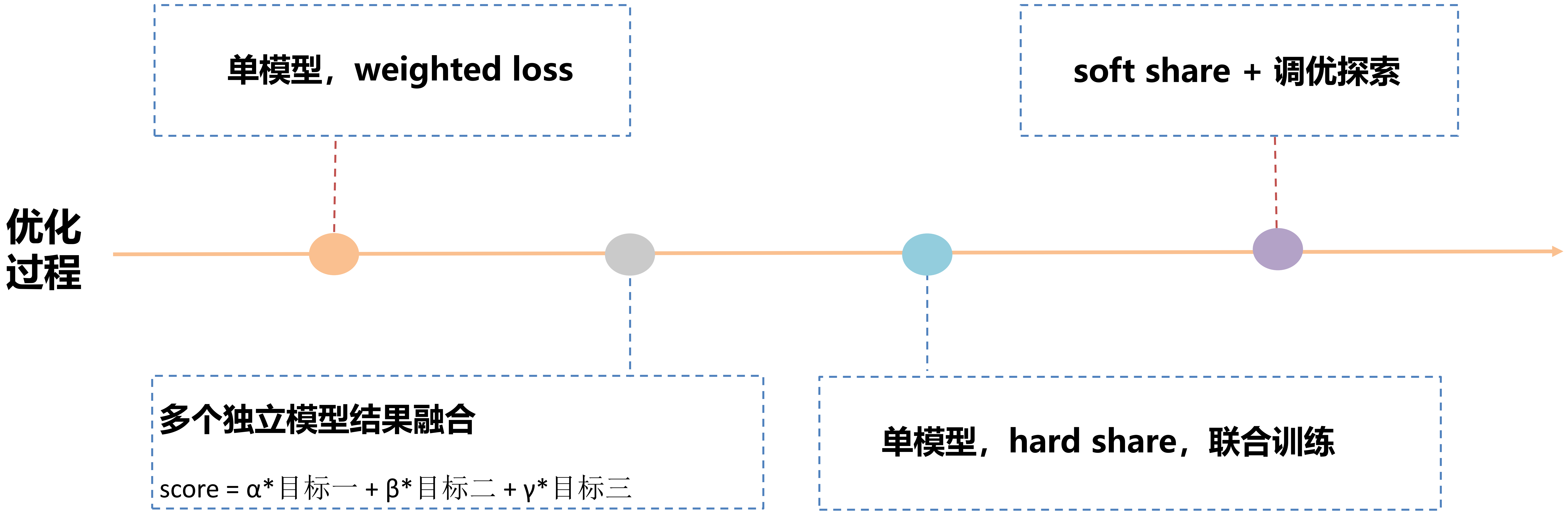
ICF数据 → 构造正负样本对 → 构造交叉特征 → 有监督学习

排序优化-多任务迭代

➢ **问题&挑战**

1、CTR、单量、GMV多个目标导向

2、转化数据稀疏，建模困难等



排序优化-样本调权

➢ 原理

点击率模型考虑转化因素，根据label区分点击、加购、收藏、支付

根据订单金额、点击到转化时间间隔调权

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} reweight \cdot y^i log\left(h_\theta\left(x^i\right)\right) + \left(1-y^i\right) log\left(1-h_\theta\left(x^i\right)\right)$$

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\gamma} log y', & y=1 \\ -(1-\alpha)y'^{\gamma} \log(1-y'), & y=0 \end{cases}$$

➢ 优点

1、复用单模型pipline，线上无需改动

2、实现简单，效果较好

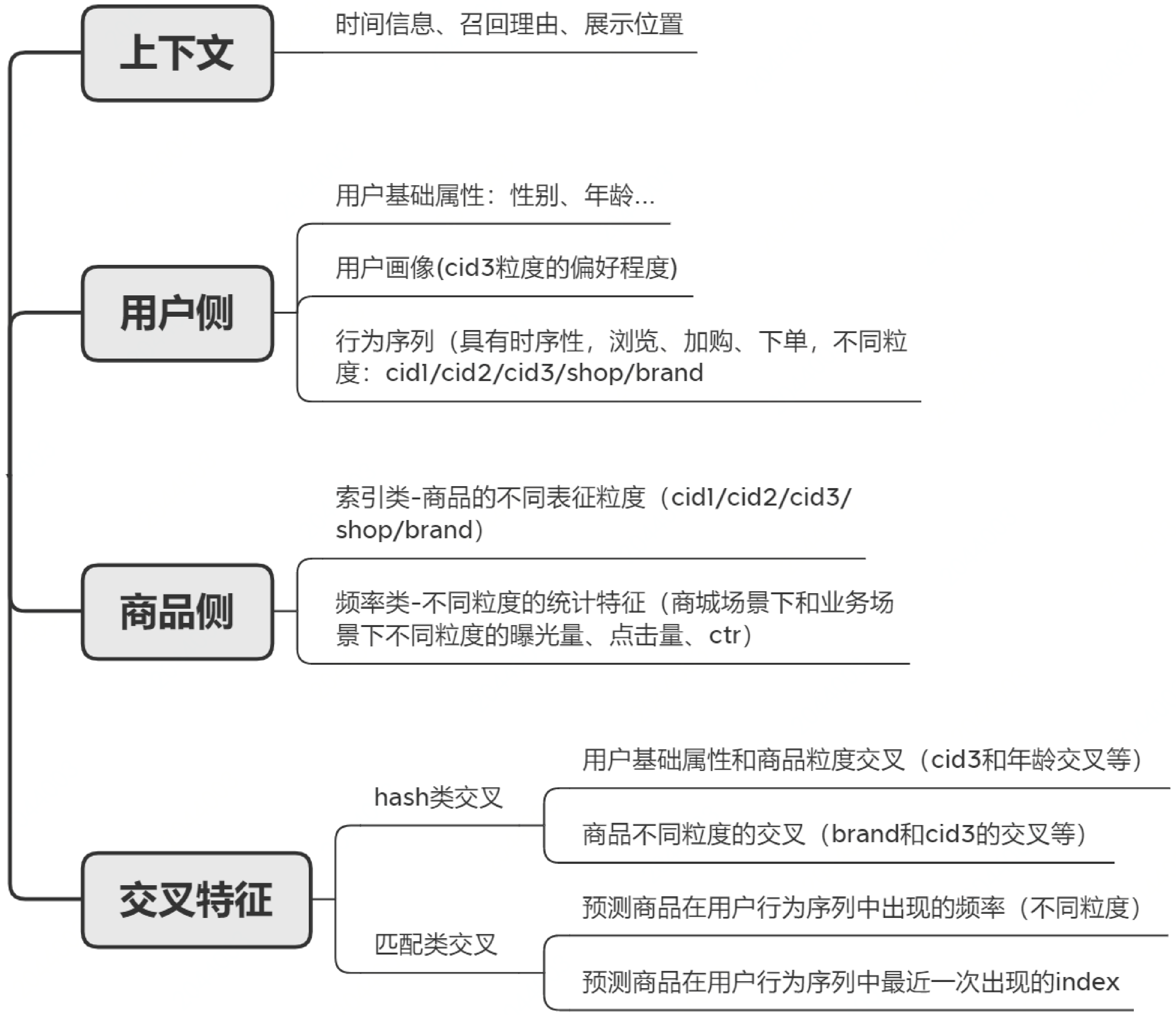
➢ 缺点

1、AB成本较高，线上AB需多个模型（不同weight训练得到）

2、推断分无直观含义，只是序关系，难以达到全局最优

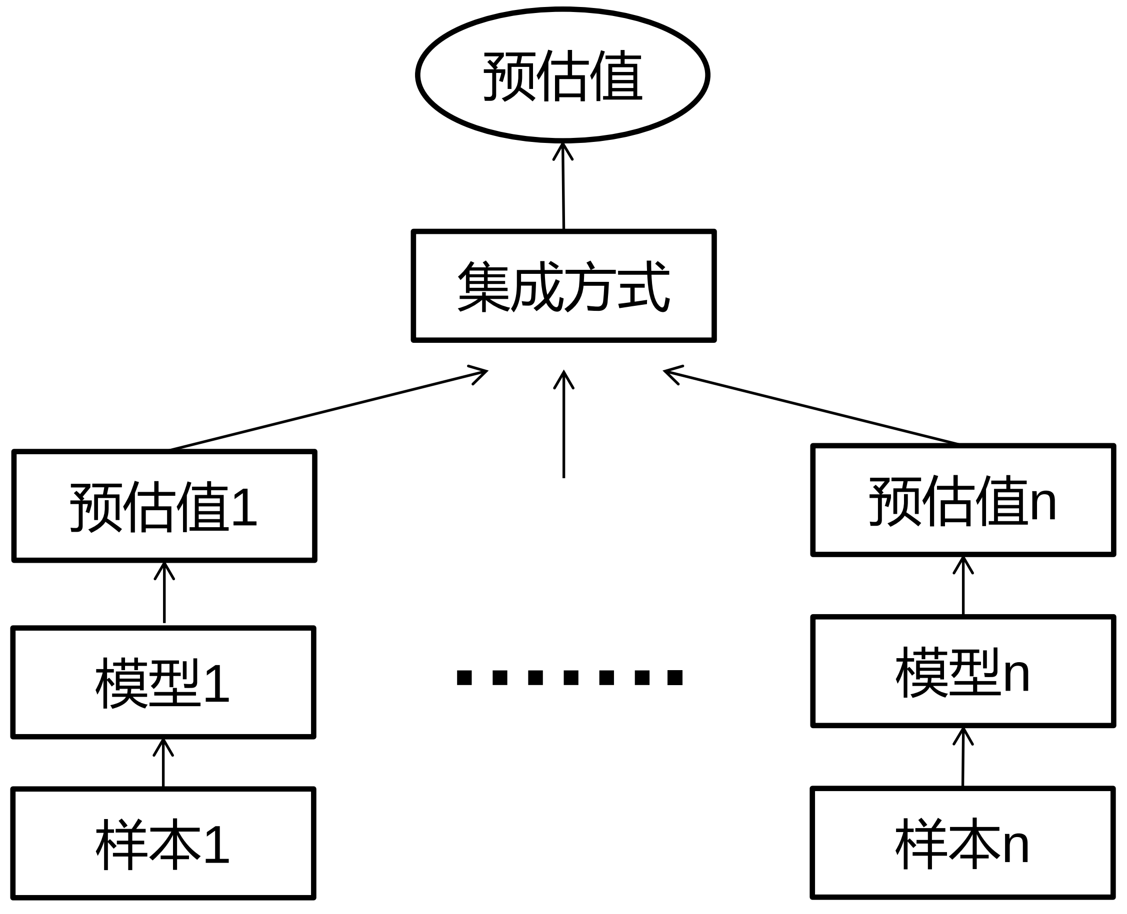
➢ 线上效果

加购率提升3%，转化率提升1.8%，ARPU提升3.1%，点击率负向0.6%



DataFunTalk

排序优化-多模型集成



E_ORDER = pCTR*pCVR

E_GMV = pCTR*pCVR*PRICE

优点

- 任务解耦，模型独立，可多人迭代/维护

缺点

- 多套模型，离线训练/线上成本高
- 参数共享较困难
- 样本选择偏差/稀疏性问题

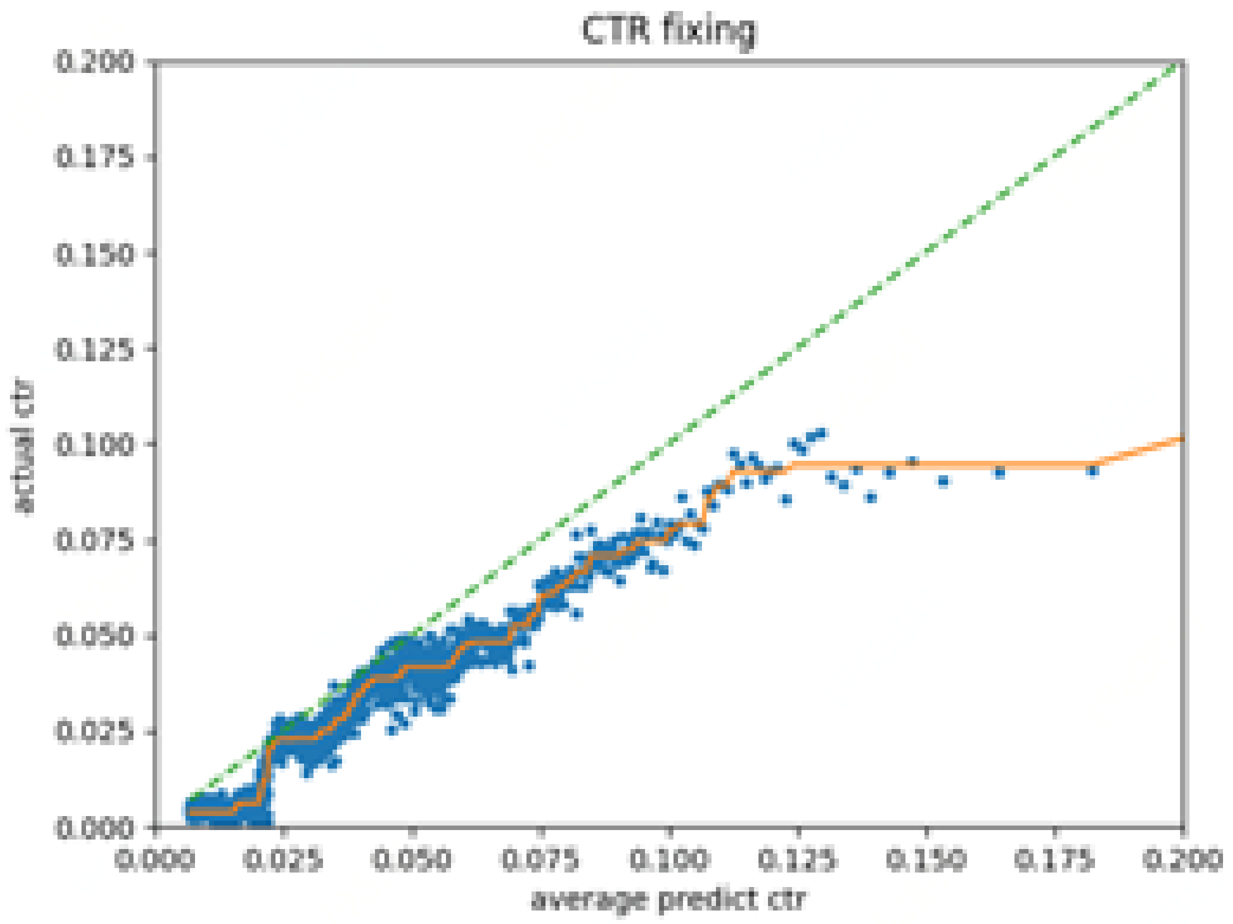
抽样/加权误差

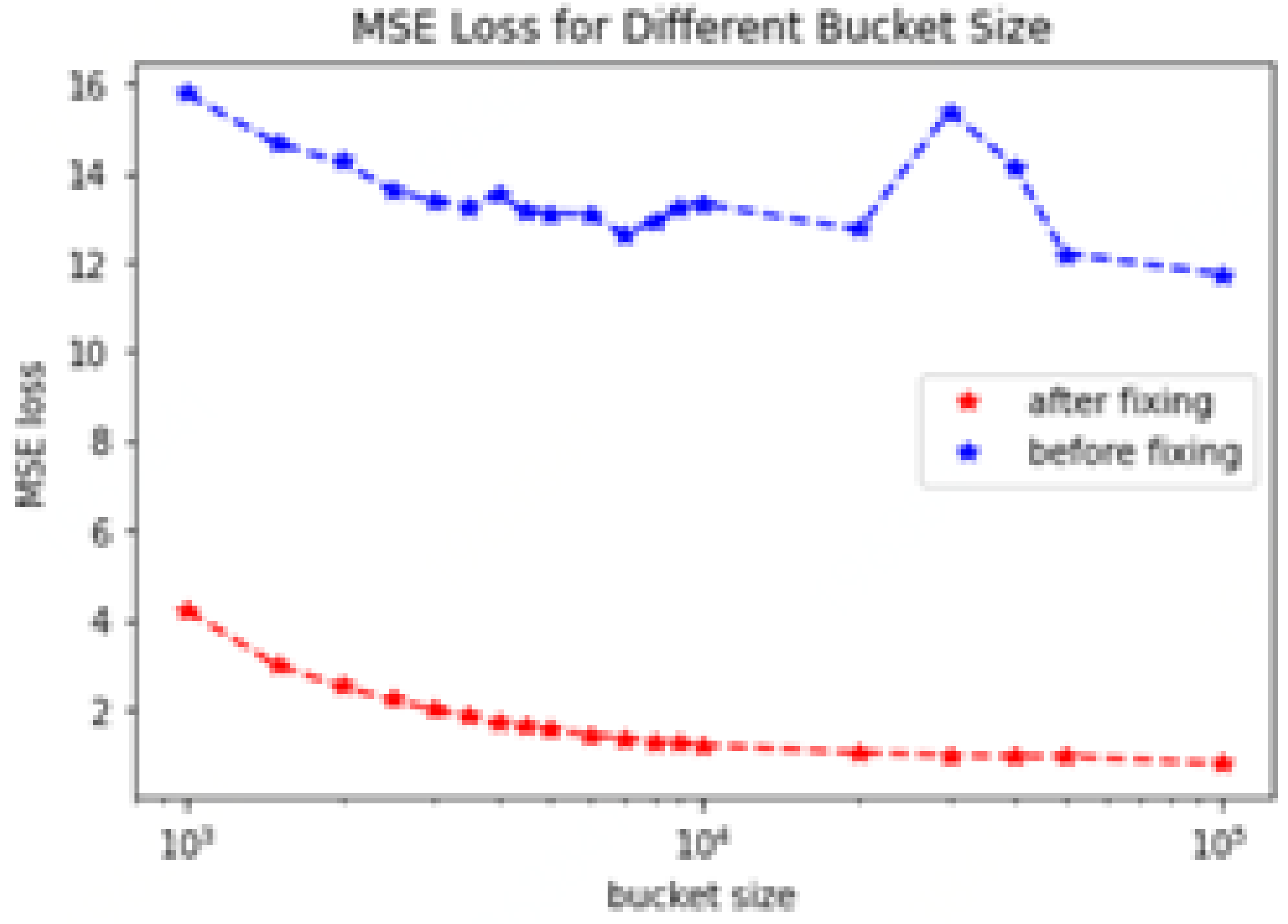
$$p = \frac{1}{1 + e^{-(wx + ln(r))}}$$

预估误差

$$\min \sum_{i=1}^{N} w_i (\boldsymbol{X}_i - y_i)^2$$

$$\text{s.t.} \quad \boldsymbol{X}_1 \leq \cdots \leq \boldsymbol{X}_N, \quad \boldsymbol{w} = \{w_1 \ldots, w_N\} > 0$$

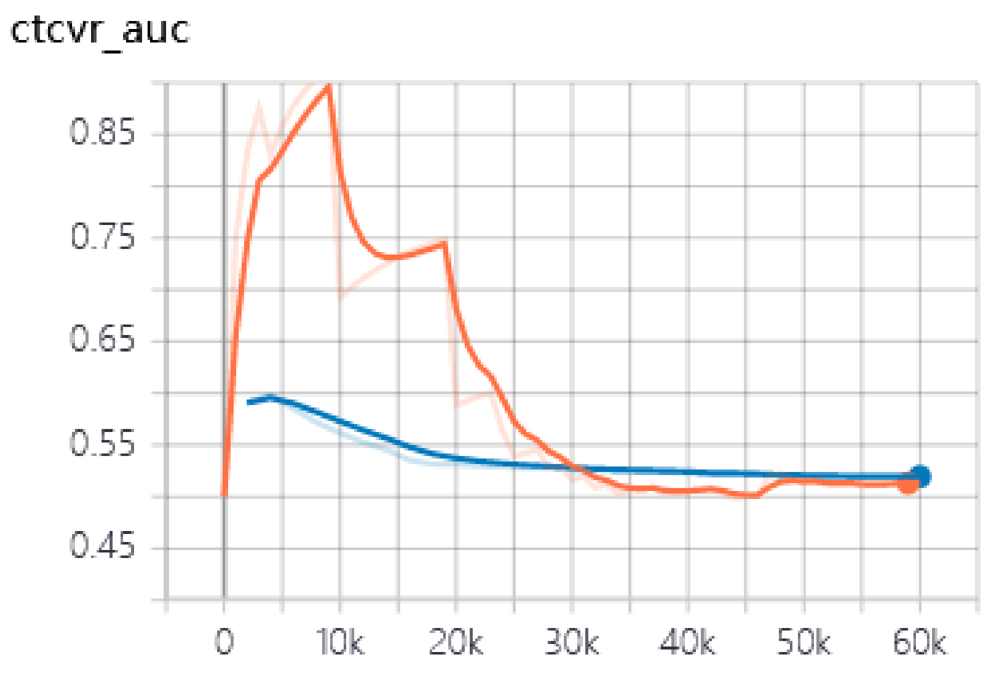


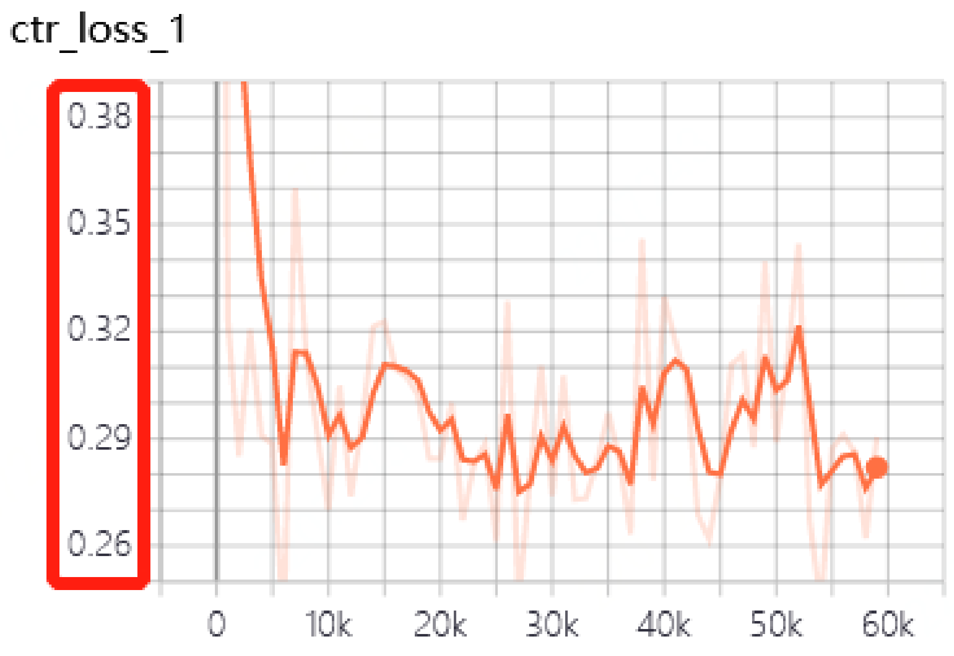


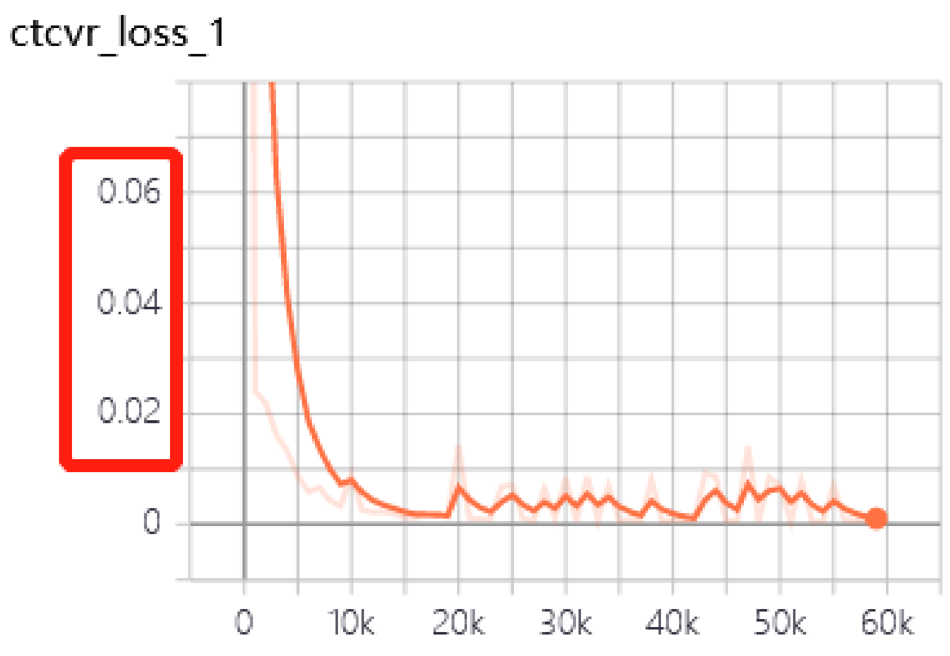
修正后，AUC保持不变；交叉熵损失，由0.1499降低为0.1243，降低了17%；
均方误差*1e+5由13.10降为1.53，修正后MSE Loss降低了88%；
上线后ARPU提升2.3%

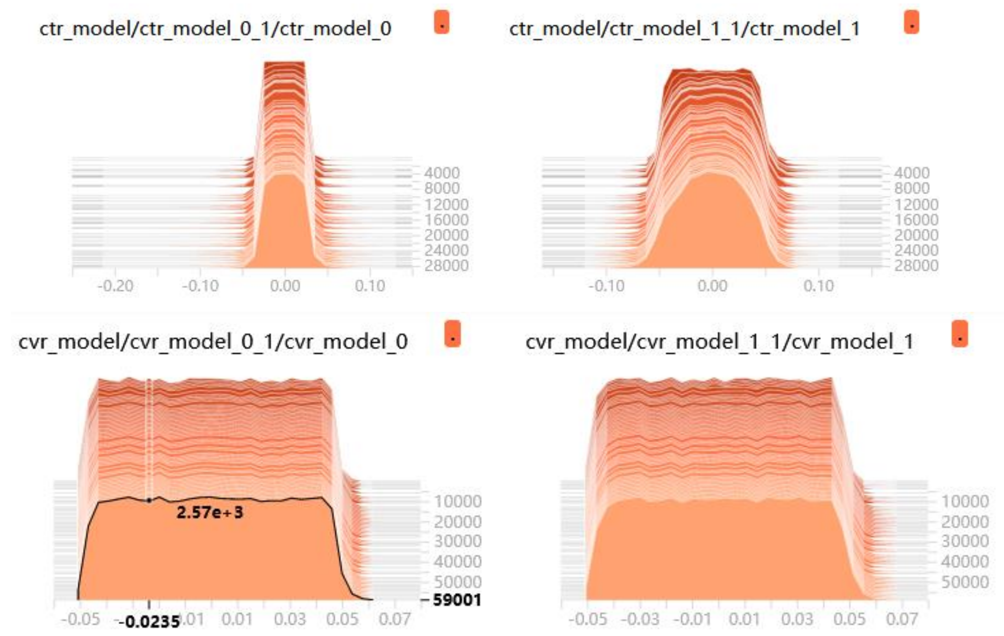
DataFunTalk

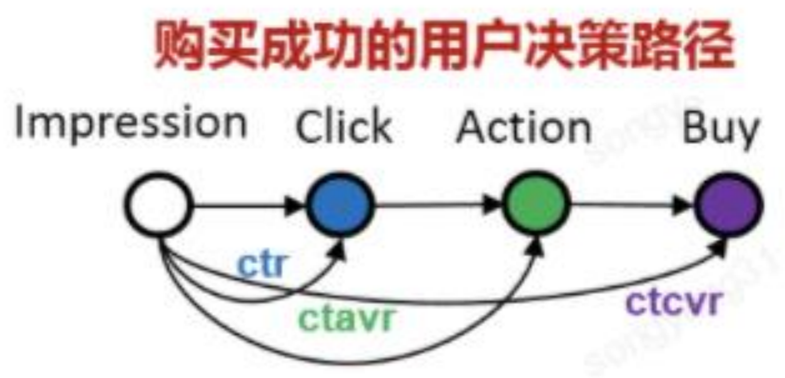
排序优化-多任务模型

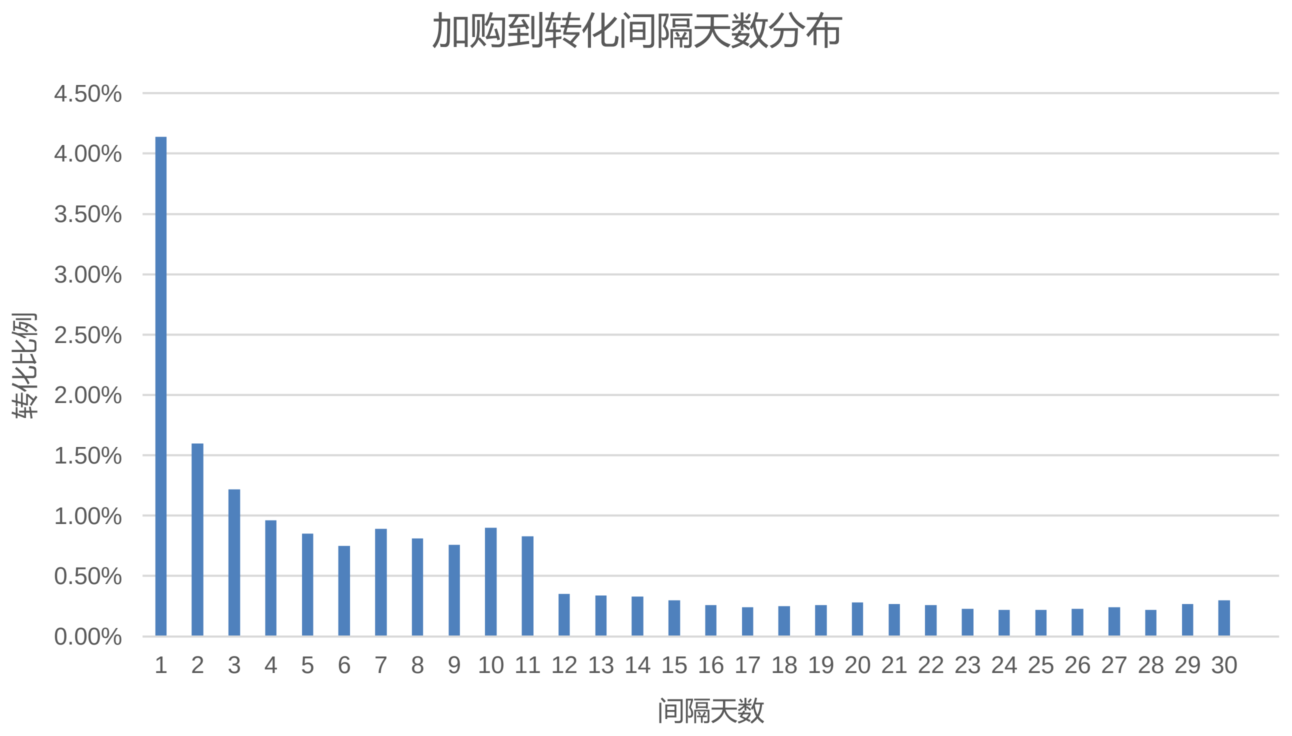












引入三个目标同时优化

E_GMV= pCTR*(pCVR + γ* pAVR)*PRICE

问题	改进
hard share 一侧未训练充分	ctcvr loss reweight(BASE)
	ctcvr 转化样本 reweight
	reweight +固定一侧参数
	reweight +梯度block
	reweight + swish激活函数
soft share	reweight + 转化序列
	reweight + 转化序列 +Attention
	reweight + 转化序列 +Attention +辅助label
action , position bias	增加action 任务
	position置0
	PAL塔预测曝光概率

推荐+广告业务

万川RTA业务

头部媒体
- 流量分配
- oCPX优化
- 广告竞价

代理服务商
- 渠道评估
- 投放计划
- 预算拆解
- 实时调控

广告主
- 投放预算
- 成本要求
- 效果反馈

- 投放引擎
- 算法预估
- 参竞控制

DSP业务

媒体1
媒体2
媒体3
媒体4

ADX 广告 交易平台

实时请求

竞价价格

DSP
- 作弊检测
- 人群定向
- 效果预估
- 预估修正
- 精准出价

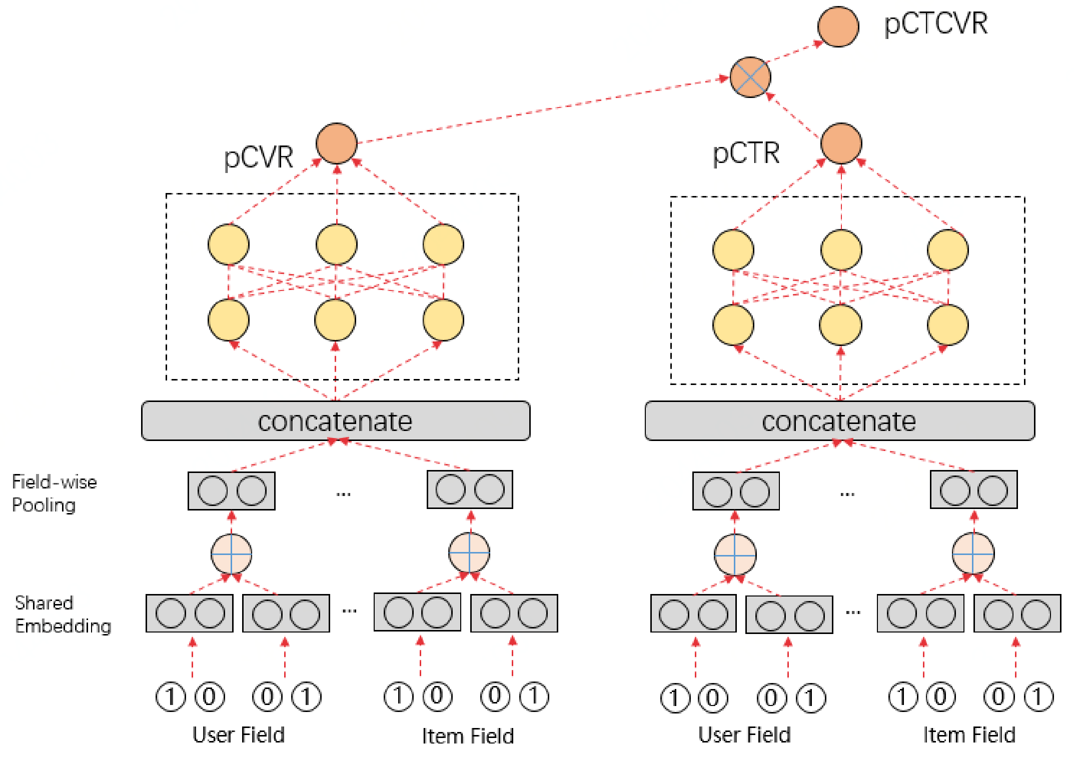
广告主1
广告主2
广告主3
广告主4

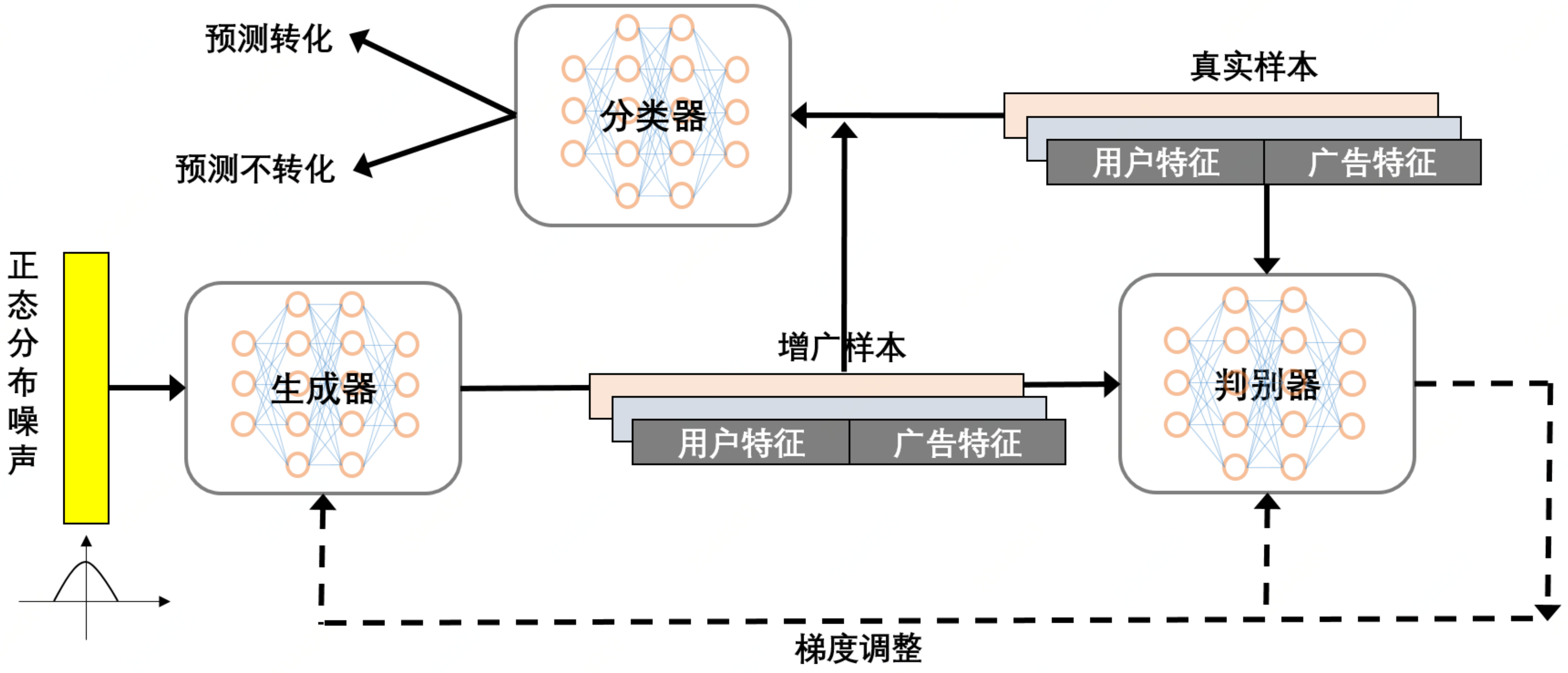
RTA广告-冷启动投放

➢ **问题&挑战**

1、转化稀疏，且无曝光数据

2、RTA无法决定出某个广告创意





	1	2	3
方案	**样本融合**	**SMOTE插值**	**GAN样本增广**
效果	样本扩充2倍，AUC提升0.4% 线上效果不显著	样本扩充1倍，AUC提升0.7% 线上转化率提升1%	样本扩充1.5倍，AUC提升0.5% 线上转化率提升0.3%

DSP广告-CPC控制

➢ **出价策略**

保量客户，CPM稳定

引流客户，CPC稳定，CTR稳定

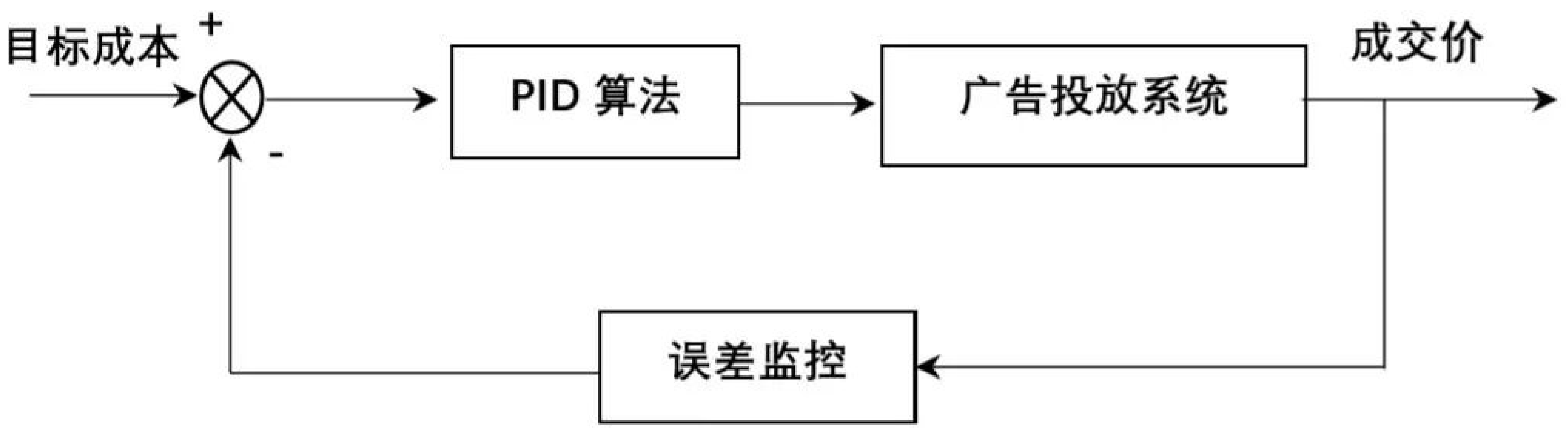
效果客户，ROI诉求，ROI稳定，ROI=E(CTR*CVR*PRICE) *1000/CPM

➢ **问题&挑战**

1、客户CPC投放，通过CPM采买，冷启动CPC较难控制

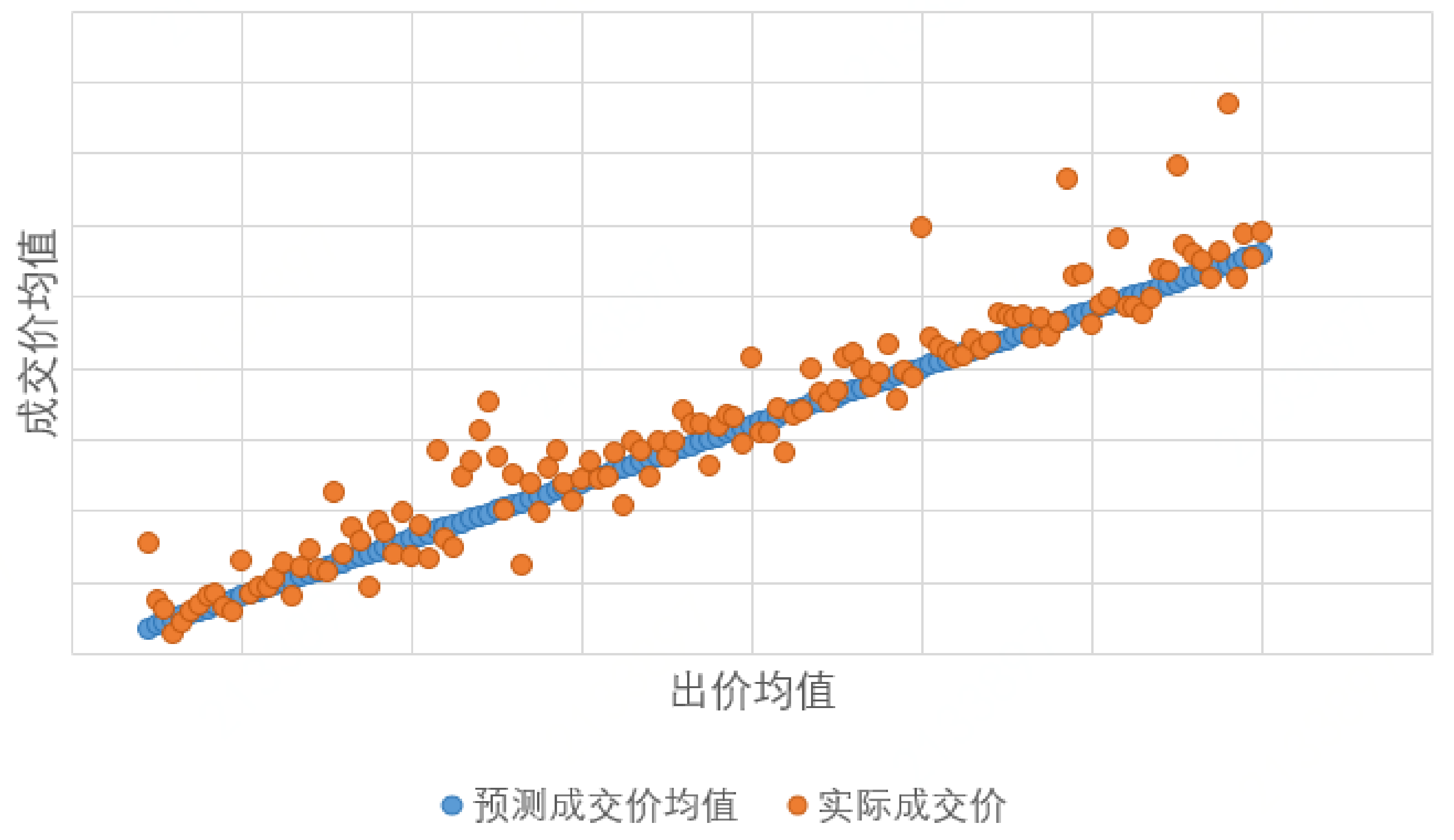
2、若前期CPC波动较大，影响客户的预算决策

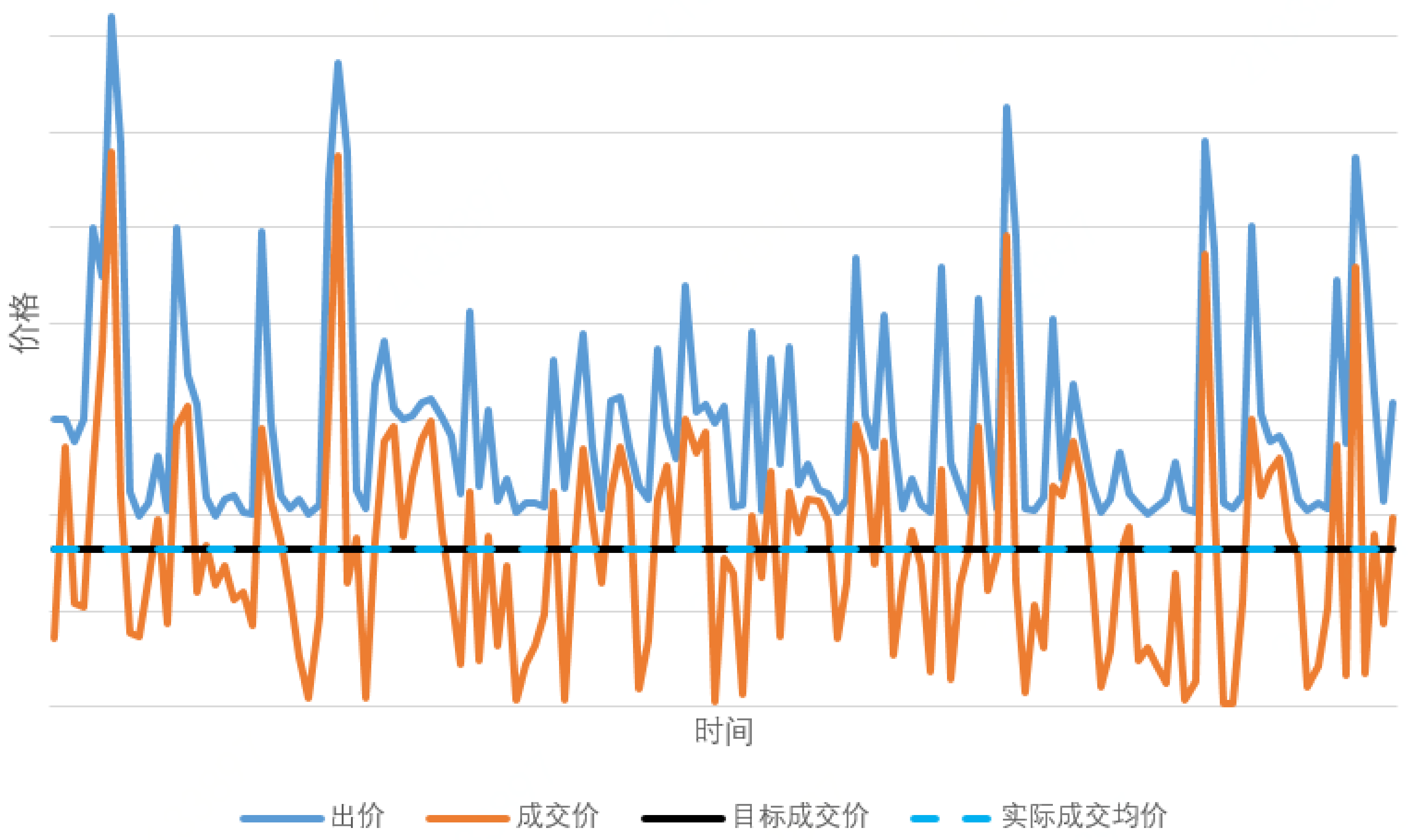
➢ **方案**



$$U(t) = K_p(err(t) + \frac{1}{T_i}\int err(t)dt + T_d\frac{derr(t)}{dt})$$

CPM出价 = pCTR * (目标CPC + *f(U(t))*) * 1000





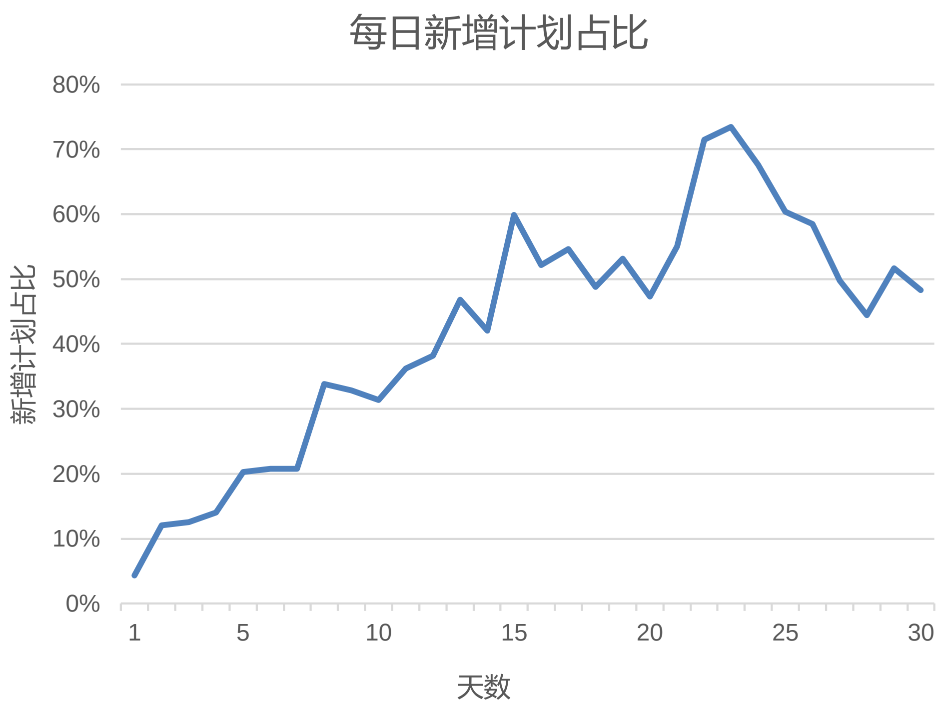
DataFunTalk

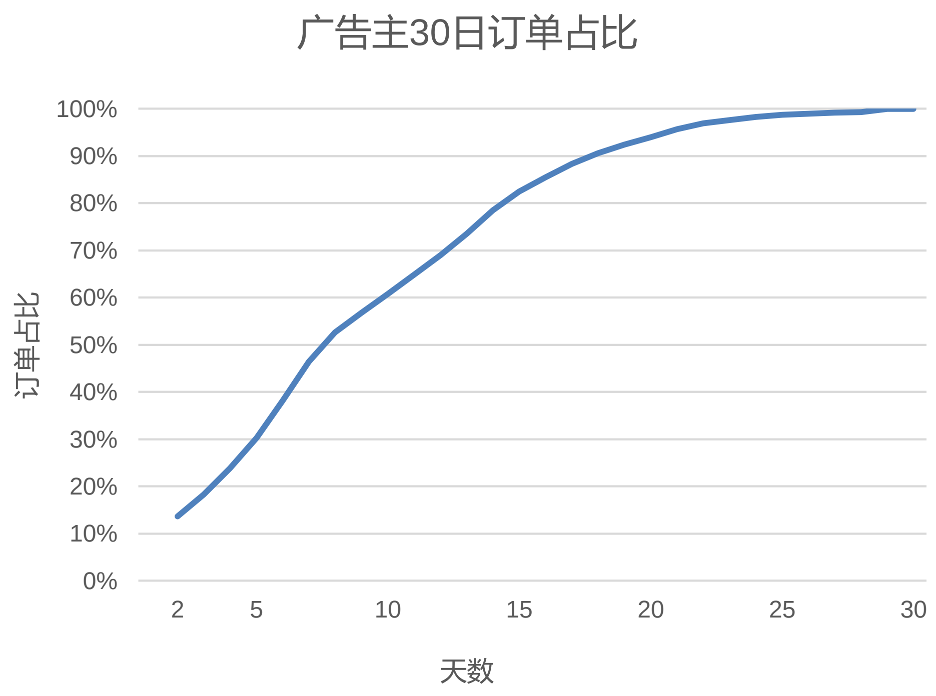
转化反馈延迟

➢ 问题&挑战

1、 用户下单前有决策时间

2、广告归因周期长，部分样本label未确定





Pr(C|X)，即建模是否会发生转化行为

Pr(D|X,C=1)，即当转化行为发生时，与点击行为的时间间隔

$$\Pr(C=1 \mid X=\mathbf{x}) = p(\mathbf{x}) \quad \text{with} \quad p(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}_c \cdot \mathbf{x})}$$

$$\Pr(D=d \mid X=\mathbf{x}, C=1) = \lambda(\mathbf{x}) \exp(-\lambda(\mathbf{x}) d)$$

$$\begin{aligned}
&Pr(Y=1, D=d_i|X=x_i, E=e_i)\\
&= Pr(C=1, D=d_i|X=x_i, E=e_i)\\
&= Pr(C=1, D=d_i|X=x_i)\\
&= Pr(D=d_i|X=x_i, C=1)Pr(C=1|X=x_i)\\
&= \lambda(x_i)\exp(-\lambda(x_i)d_i)p(x_i)
\end{aligned}$$

$$\begin{aligned}
&Pr(Y=0|X=x_i, E=e_i)\\
&= Pr(Y=0|C=0, X=x_i, E=e_i)Pr(C=0|X=x_i)\\
&+ Pr(Y=0|C=1, X=x_i, E=e_i)Pr(C=1|X=x_i)\\
&= 1-p(x_i)+p(x_i)\exp(-\lambda(x_i)e_i)\\
&\Uparrow\\
&Pr(Y=0|C=1, X=x_i, E=e_i) = Pr(D>E|C=1, X=x_i, E=e_i)\\
&= \int_{e_i}^{\infty}\lambda(x)\exp(-\lambda(x)t)dt = \exp(-\lambda(x)e_i)
\end{aligned}$$

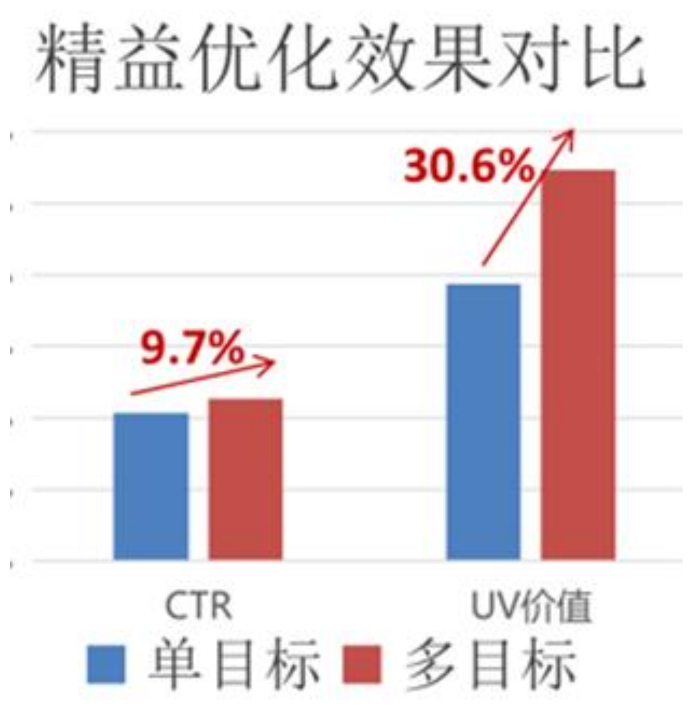
Loss Function:

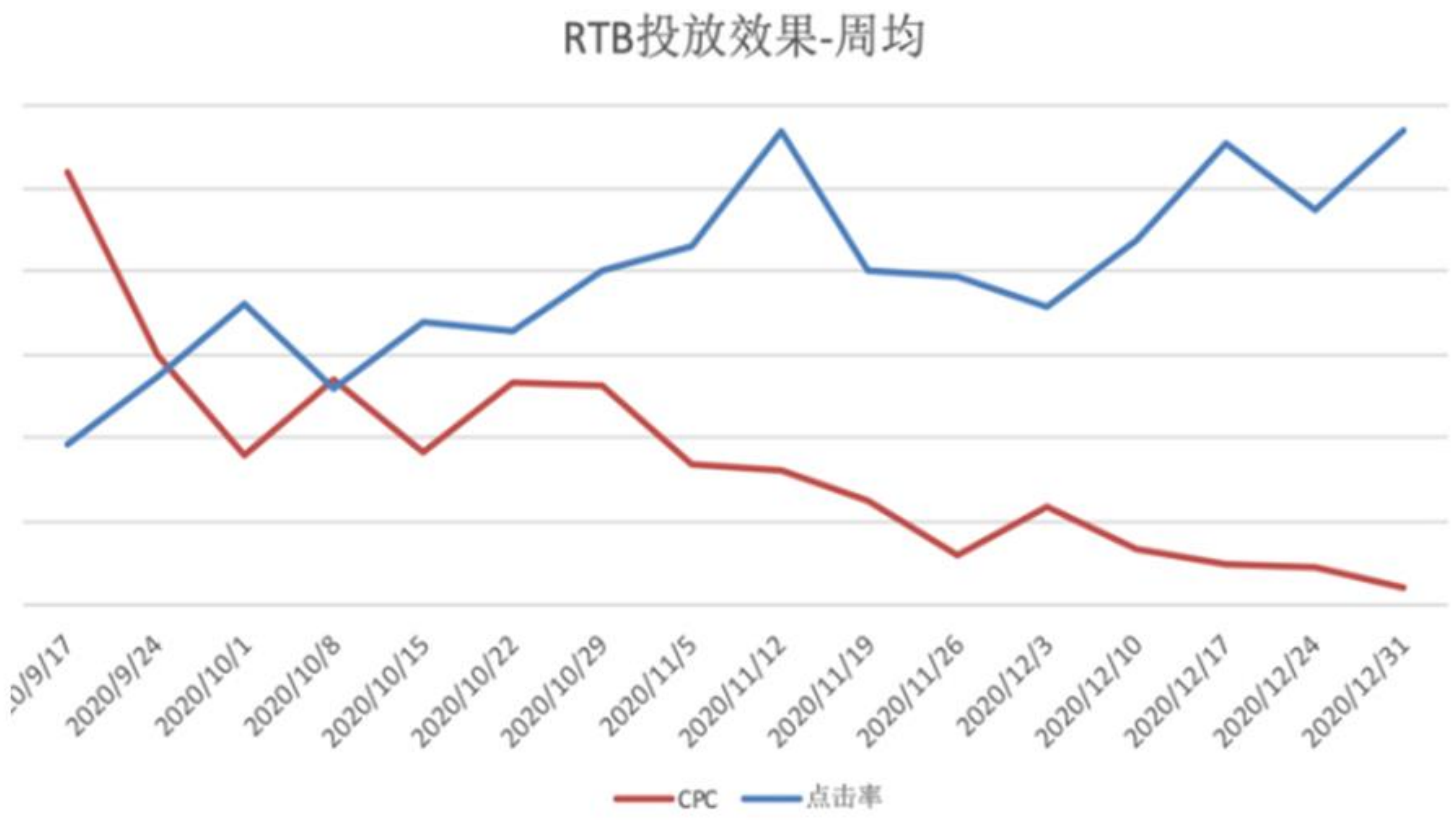
$$\arg\min_{w_c, w_d} L(w_c, w_d) + \frac{\mu}{2}(||w_c||_2^2 + ||w_d||_2^2)$$

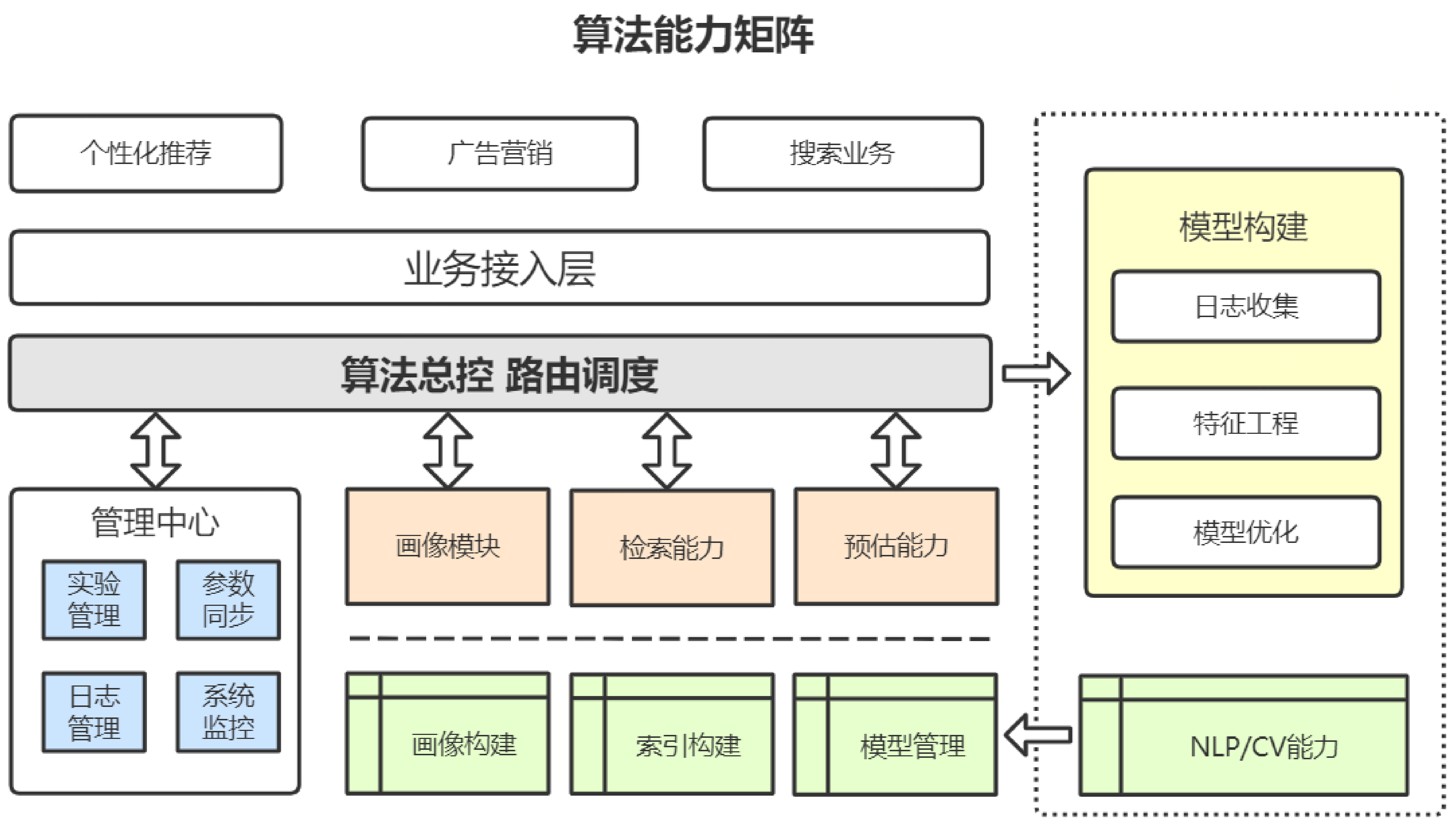
其中，μ是正则化参数，L是负对数似然：

$$\begin{aligned}
L(w_c, w_d) = &-\sum_{i, y_i=1} \log p(x_i) + \log\lambda(x_i) - \lambda(x_i)d_i\\
&-\sum_{i, y_i=0} \log[1-p(x_i)+p(x_i)\exp(-\lambda(x_i)e_i)]
\end{aligned}$$

成果与规划







攻克方向

- 增量模型、时延模型
- 多任务学习+ROI精准预估
- 参数自动寻优
- 行业特征挖掘

DataFunTalk

2021

DataFunTalk

THANKS!

今天的分享就到这里...

Ending