20250627-多目标级联延迟反馈AIR双周会

🚩 **Highlights**

1. **学术数据集构建：**按**用户购买数差异化采样**数据集产出完成，30d样本数4100w，转化率9.1%、复购率2.3%、退款率4.9%，符合预期；**仅保留购买用户数据** TODO
2. **Base代码库：**开发torch版base，已集成ED-DFM, Defuse, Defer, FNW, Oracle等经典base，构造伪数据上跑通训练，待对齐现有数据集特征。
3. **净成交建模：**推导多目标重要性采样loss
4. **ECAD模型分析：**退款数据稀疏导致退款模型预测过度自信，从而引发净转化PCOC过低。

TODO:

数据集构建： 2种数据集构建 @李欣钰(晚漾)

可联网开发机： @产张明(明言)

公开数据集方法（criteo + 自己数据）： @骆明轩(洛谨)

1. 现有延迟反馈方法（不考虑假正样本）
2. 现有延迟反馈方法（考虑假正样本）
3. 进一步的方法探索

GMV模型：尝试现有延迟反馈方法 @李欣钰(晚漾)

学术数据集构建@李欣钰(晚漾)

主表逻辑修正：

- 问题：一个adgroup_id下有多个不同entity_id，以及存在entity_id='0'的数据
- 修正：添加筛选条件 entity_source = 'item' and entity_id != '0'

在sample数据前会将以下数据筛除：

ad_item_id IS NULL

ad_item_id = '-'

ad_item_id != entity_id

u_id筛选

- **方案一：**

  根据20250501-20250525期间，用户购买数量差异化采样

  - 活跃用户：购买数量>5，随机取100万个
  - 中等活跃用户：1< 点击数量 < 5，随机取60万个
  - 不活跃用户: 点击数量 = 0，随机取40万个

  每天数据只取筛选出的这200万个用户的数据。

- **筛选结果：**

  总量（20250501-20250525）：

总记录数	购买数	复购数	退款数	购买率	复购率	退款率
41037005	3742731	946973	2015540	9.1204%	2.3076%	4.9115%

  每天数据量级：

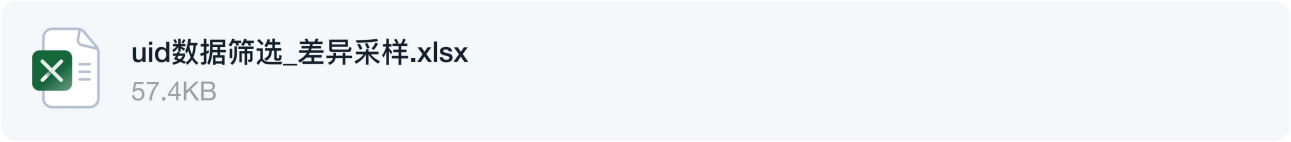
  记录数：一百万

  用户数：50万

  购买数：10万（10%）

  退款数：4万~10万

  复购数：2万~5万



- **方案二：**

从20250401-20250430数据中找到有购买的用户

每天的数据从有购买的用户中sample

- **筛选结果：**

20250501-20250525有购买的用户数：79133648

ds	total_records	total_user	pay_cnt	mul_pay_cnt	refund_cnt	pay_ratio	mul_pay_ratio	refund_ratio
20250501	19696224	9597147	823688	95213	295110	4.1820%	0.4834%	1.4983%
20250502	19979002	9727099	849673	99300	303685	4.2528%	0.4970%	1.5200%
20250503	20276612	9912526	871347	102579	309494	4.2973%	0.5059%	1.5264%
20250504	21662878	10483774	940230	111962	336787	4.3403%	0.5168%	1.5547%
20250505	23853515	11441772	1096409	133416	393265	4.5964%	0.5593%	1.6487%

每天数据量在千万级，还需再采样

再采样方式一：

- 根据用户购买数量，取前100万个用户的数据

筛选结果(20250501-20250525)：

总记录数	购买数	复购数	退款数	购买率	复购率	退款率
45650219	2612713	682160	1580986	5.7233%	1.4943%	3.4633%

分析：购买率低可能是因为用户买的多，点的也多（用户购买率低）

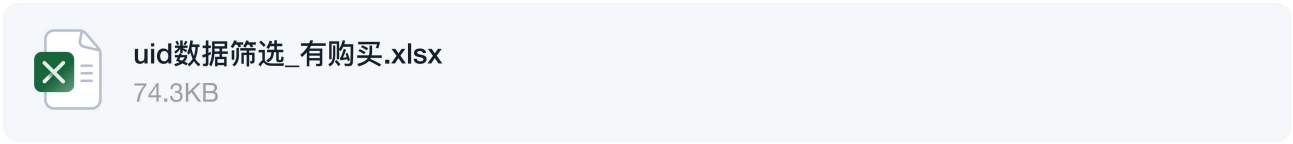
再采样方式二：

- 根据用户购买率（购买数/点击数）， 取前2500万个用户的数据

筛选结果(20250501-20250525)：

总记录数	购买数	复购数	退款数	购买率	复购率	退款率
41930460	8834022	1197282	2976438	21.0683%	2.8554%	7.0985%

方案二详细统计数据：



2.Base模型集成@骆明轩(洛谨)

已经用criteo数据构建伪造数据已经集成好torch版本的ED-DFM,Defuse,Defer,FNW,Oracle等典型的base，退款相关的净转化标签预测等也已经写好逻辑。

具体进度：1.利用criteo数据集构造伪数据，加入退款时间戳。

2.集成ED-DFM,Defuse,Bi-Defuse,Defer,FNW,Oracle等典型的baseline的pytorch版本。

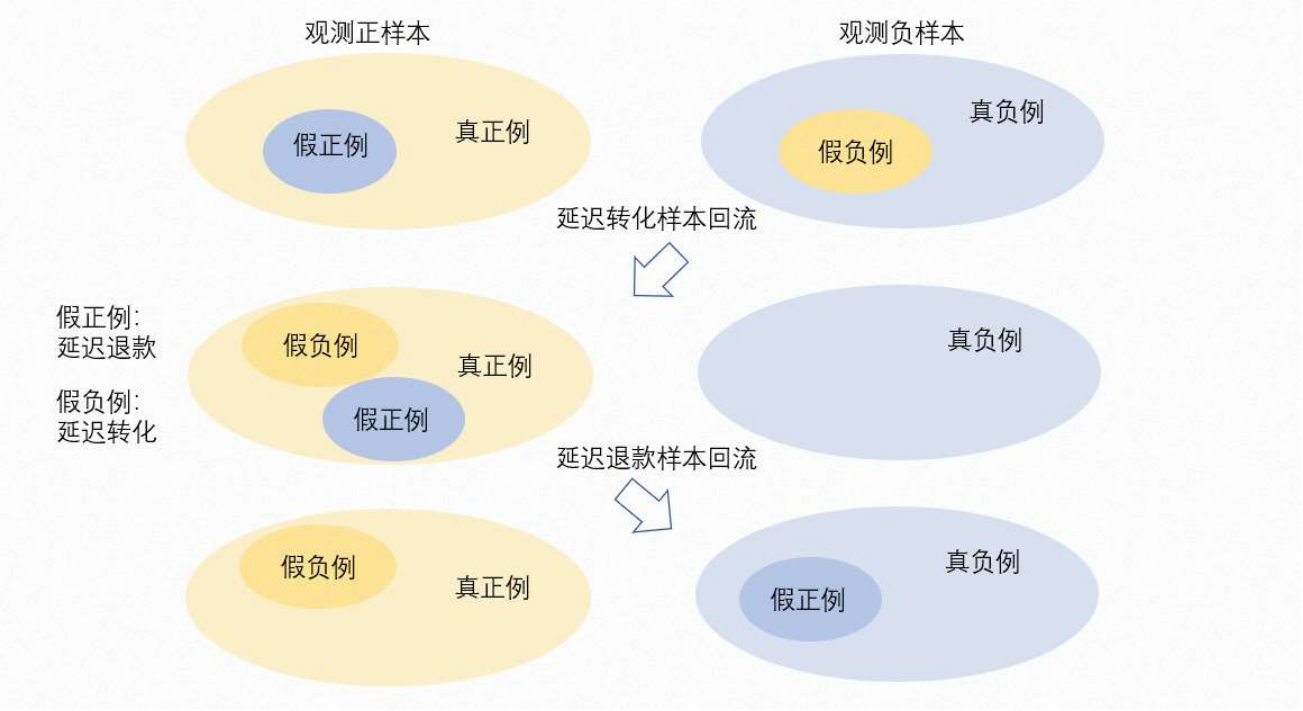
3.已经写好退款相关的净转化标签逻辑，差对齐现有的数据集特征

3.净转化相关建模

参考defer的逻辑去推理退款相关的样本回补

	观测正样本	观测负样本	
第一次	真正例 + 假正例	真负例 + 假负例	

第二次	真正例 + 假正例 + 假负例	真负例	延迟转化样本纠正
第三次	真正例 + 假负例	真负例 + 假正例	延迟退款样本纠正



$$q(x,y=1)=p(x,y=1)+\frac{2}{3}P_{dr}(x)-\frac{1}{3}P_{dp}(x)$$
$$q(x,y=0)=p(x,y=1)+\frac{2}{3}P_{dp}(x)-\frac{1}{3}P_{dr}(x)$$

其中y是考虑退款后的延迟标签,q()指的是观测分布，p()指的是真实分布。 $P_{dr}(x)$， $P_{dp}(x)$ 分别是预测是延迟退款的概率，以及是预测是延迟转化的概率

利用重要性采样

$$w^{+}=\frac{p(x,y=1)}{q(x,y=1)}=\frac{p(x,y=1)}{p(x,y=1)+\frac{2}{3}P_{dr}(x)-\frac{1}{3}P_{dp}(x)}$$
$$w^{-}=\frac{p(x,y=1)}{q(x,y=1)}=\frac{p(x,y=1)}{p(x,y=1)+\frac{2}{3}P_{dp}(x)-\frac{1}{3}P_{dr}(x)}$$

$$L=-\sum_{x,y}[y\cdot w^{+}\cdot log(f_{\theta}(x))+(1-y)\cdot w^{-}\cdot log(1-f_{\theta}(x))]$$

4.上周模型TODO

trace分析ECAD refund rate学习情况

退款数据相对太稀疏导致退款模型预测过度自信，从而引发净转化PCOC过低。

取20250424~20250429的数据去做训练，取20250430的数据去做分析

cvr_auc	netcvr_auc	rfr_auc
0.8616864608702576	0.8651729509981683	0.8259705892288873

cvr_pcoc	netcvr_pcoc	rfr_pcoc
1.3709798549552554	0.7235422805301877	5.8811459529308

PCOC 不一致：一个偏高，一个偏低，说明两个模型的预测方向不一致。
RFR 模型 PCOC 异常偏高（5.88），说明模型对实际行为严重高估，应该是数据稀疏导致。在训练时，模型看到的正样本太少，容易把少量的正样本“放大”，认为这些特征更可能触发行为。如果正样本非常少，模型可能会过度依赖这些样本中的某些噪声特征。
做了这五天的数据统计，退款数据/转化数据 = 0.14355
退款数据/净转化数据 = 0.12553，而净转化数据/转化数据 = 0.87446.
因此猜测原本ECAD效果好可能强关联数据集，原本ECAD数据集退款数据比较稠密。而在现有的退款数据稀疏的情况下，不如直接建模净转化标签。

5.TO Discuss

1. ECAD 这种的 BDL baseline 是否需要比?
2. 多窗口定义是什么？是等待窗口,归因窗口，还是实际的转化退款窗口
3. GMV baseline 及 多窗口定义。
4. 这周跟林老师开会遗留的问题
    i. 整合baseline的作用，后续是否准备出datasets文章以及benchmark文章
    ii. 哪些方法比较容易做出来，1.online模型的迭代速度。2.online模型迭代考虑是否对齐旧模型。3.online模型迭代不做全量，而是做部分层的微调。4. online模型的强化学习。5.对于退款样本稀疏问题做上采样或者构造伪退款样本。6.online 模型的deploy。