2020 DIGIX

全球校园AI算法精英大赛路演

赛道A：广告CTR预测

风犹惊入萧独夜队

团队介绍

团队成员：李昱、吕玲玲、唐高中
华南理工大学电子与信息学院二年级硕士

本次比赛成绩：

选拔赛：广告CTR预估A/B榜rank1（8.24-9.30）
搜索相关性预测A/B榜rank1（8.25-9.30）

精英赛：广告CTR预估A榜rank2，B榜rank1

答辩将介绍我们在广告CTR预估的方案

目录 CONTENTS

01 基础方案

1.1 赛题理解

本次广告CTR预估题目较为传统，给定前7天内每条曝光的点击行为，预测将来某一天内曝光的点击率，评价指标AUC

选拔赛：训练集1-7天，A榜测试集第8天，B榜测试集第9天

精英赛：训练集1-7天，A榜测试集第8天，B榜测试集第10天

1.1 赛题理解

用户基础特征

age:用户的年龄
city：用户的常驻城市
city rank：用户常驻城市的等级
device name：用户使用手机型号
career：用户的职业
gender：用户的性别
residence：用户的常驻省份

广告点击标签

label：0未点击，1点击
uid：匿名化处理后的用户唯一标识

用户基础特征

task_id：广告任务唯一标识
creat_type_cd：素材的创意类型id
dev_id：广告任务对应的开发者id
spread_app_id：投放广告任务对应的应用id
app_first_class:广告任务对应的应用的一级分类
app_second_class:广告任务对应的应用的二级分类
app_score:app得分

adv_id：广告任务对应的素材id
adv_prim_id：广告任务对应的广告主id
inter_type_cd：广告任务对应的素材的交互类型
tags：投放广告任务对应的应用id
his_app_size:app存储尺寸
his_on_shelf_time:上架时间
indu_name:广告行业信息

设备基础特征

device size：用户使用手机尺寸
emui_dev：emui版本号
list_time：上市时间
device_price：设备价格

场景特征

slot_id：广告位id
net_type：行为发生的网络状态

用户画像标签

communication_onlinerate：手机在线时段
membership_life_duration：会员用户生命时长
communication_avgonline_30d：手机日在线时长

up_life_duration：华为账号用户生命时长
consume_purchase：付费用户
up_membership_grade：服务会员级别

本题给出的每次曝光特征，可根据内容划分为**用户、广告、媒体**三方面特征，也可根据数据类型划分为ID特征和连续特征
关注重点在于**用户特征和ID特征**

1.2 特征工程

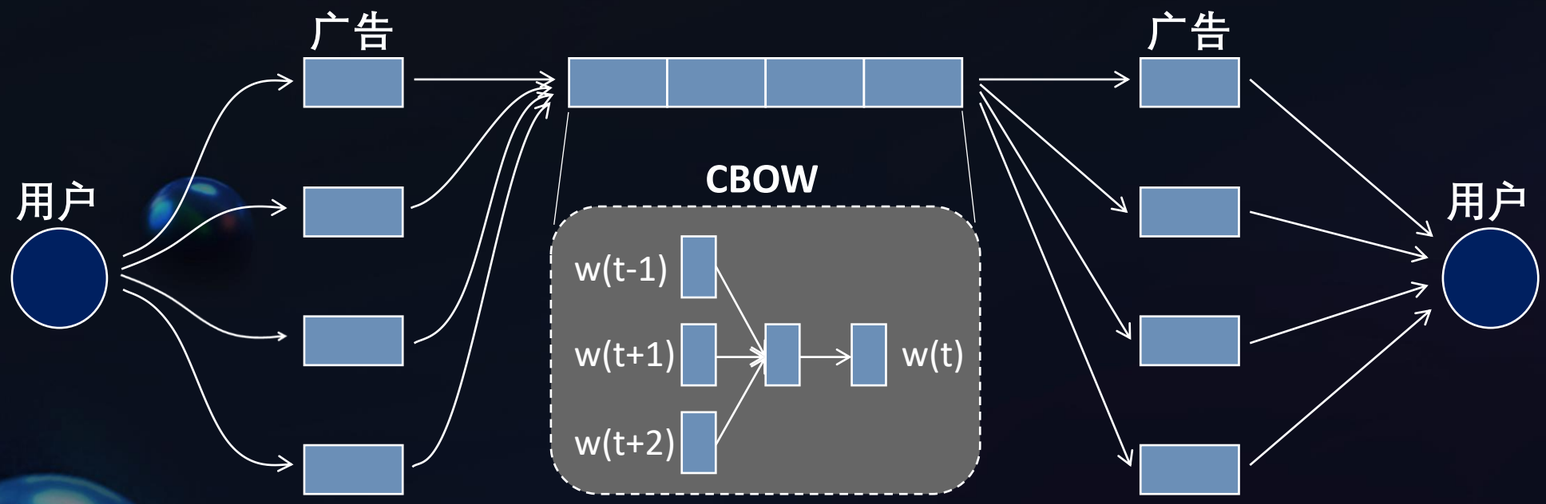
曝光特征：统计所有ID类特征在8天内的曝光次数

交叉特征：统计用户ID与所有广告侧ID、广告ID与所有用户侧ID的类别交叉，如某个用户ID曝光过多少不同的广告ID

CTR特征：统计所有ID类特征前所有天的历史点击率

embedding特征：构建广告曝光序列，训练word2vector得到广告表征，平均广告表征得到用户表征

可以用滑窗方式来统计特征，B榜阶段加入A榜测试集辅助统计

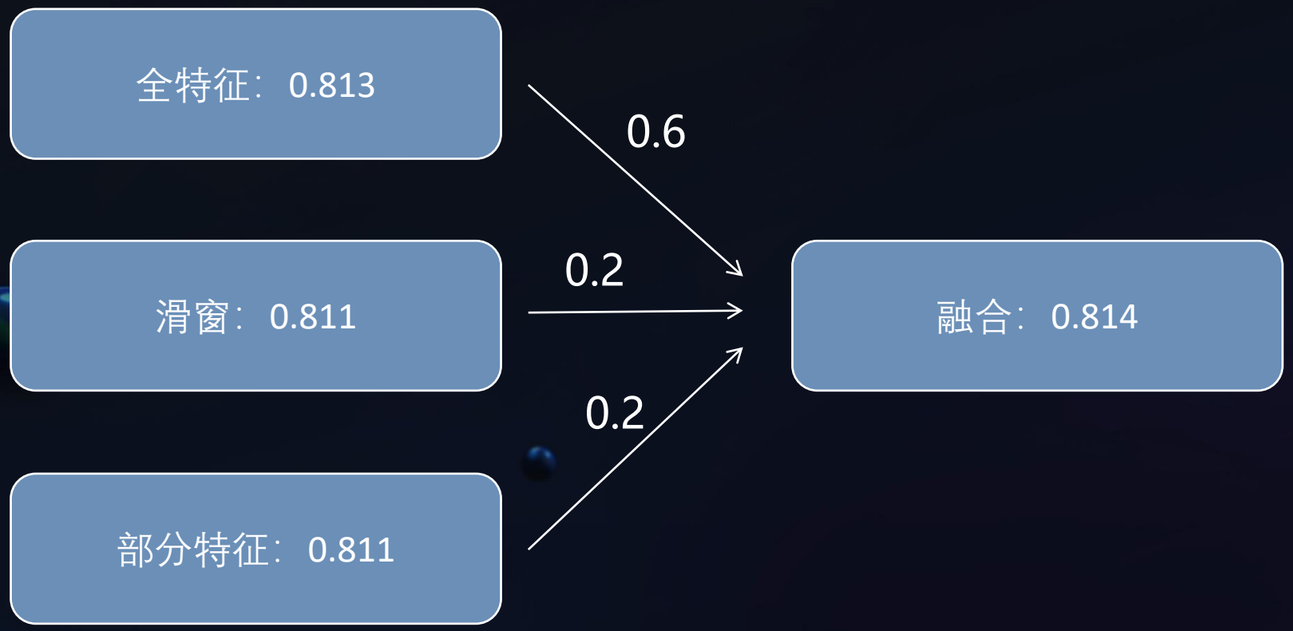


1.3 算法实现

使用lightgbm作为我们的模型，时间上越靠近测试集的样本给予更大的权重

通过滑窗和去除冗余特征构造差异性模型来进行融合

在选拔赛A榜阶段一直使用单模，B榜阶段双模，精英赛阶段三模



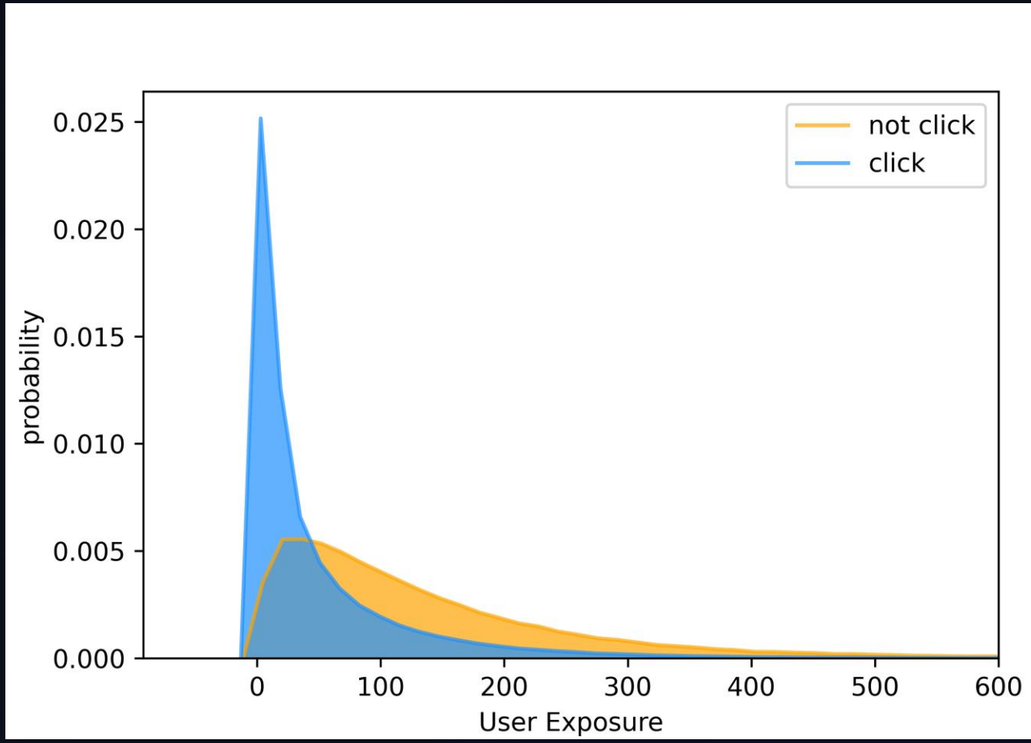
是否觉得方案平平无奇?

别急，下面才是我们工作的重点

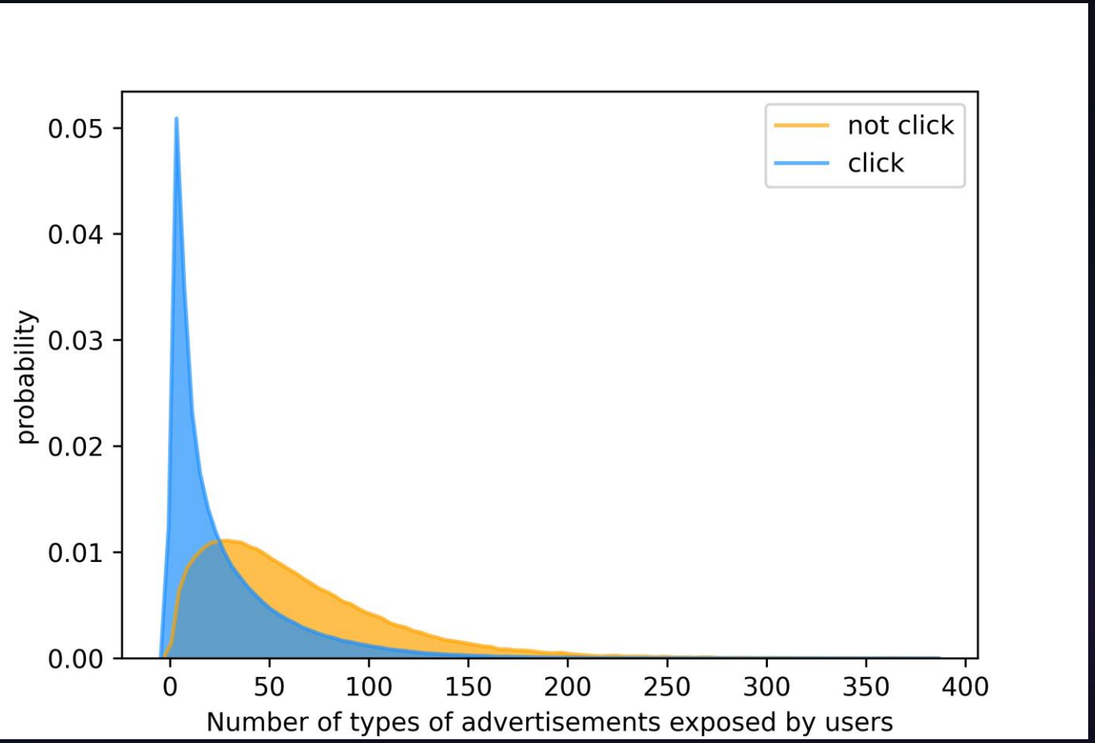
02 冷启动探索

2.1 数据分析

用户特征与点击行为强相关



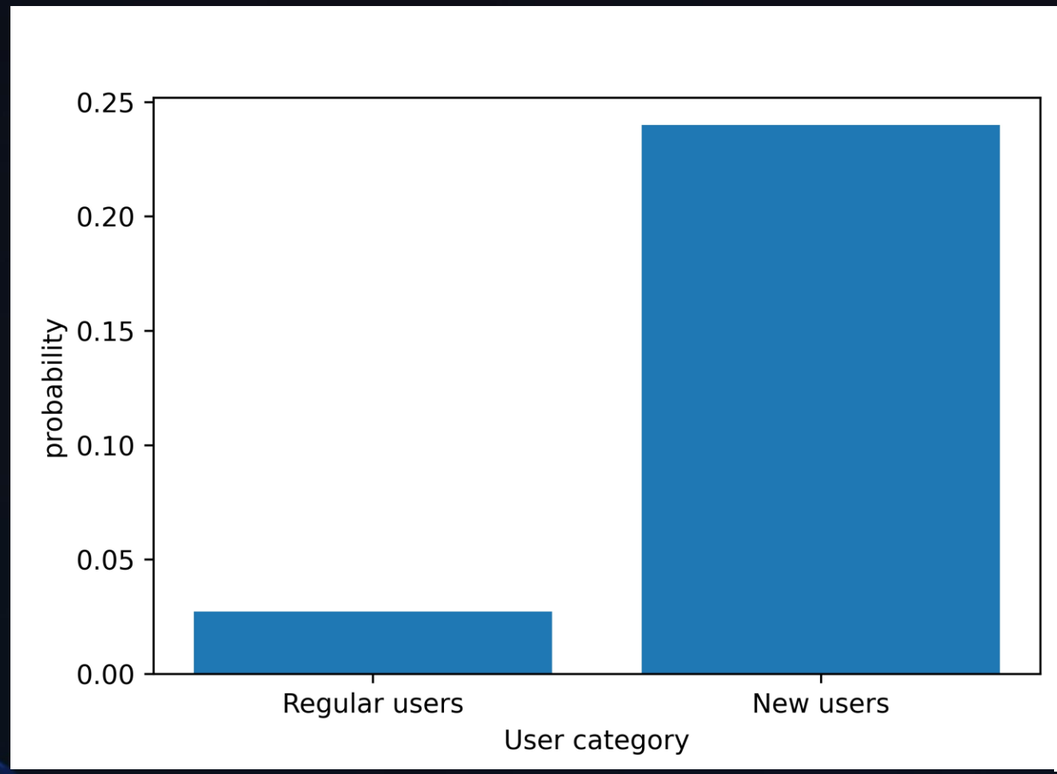
用户曝光特征的概率密度图



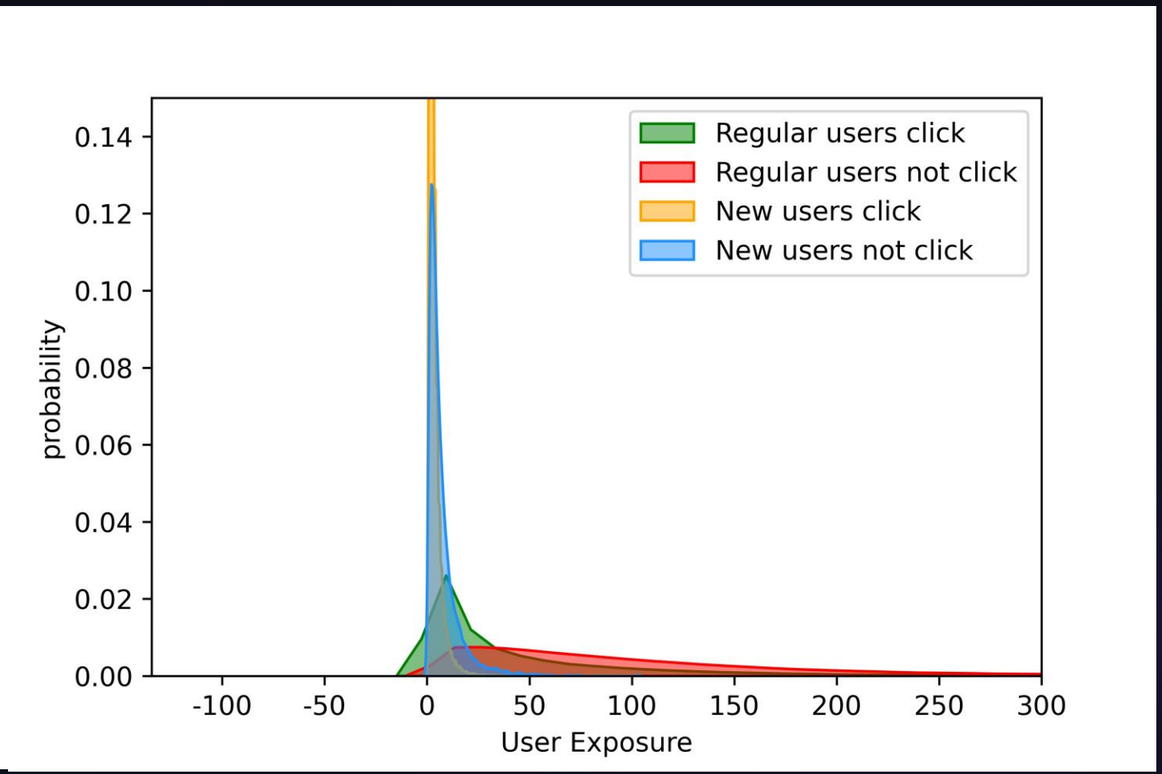
用户交叉特征的概率密度图

2.1 数据分析

冷启动用户点击率与特征分布呈现巨大的差异性



老用户与新用户的点击率



老用户与新用户的曝光特征概率密度图

2.1 数据分析

测试集用户冷启动比例暴涨

相对于前所有天统计

	冷启动用户数量	冷启动用户比例
第2天	157577	0.2478
第3天	90516	0.1425
第4天	63367	0.1005
第5天	46649	0.0743
第6天	41381	0.0652
第7天	30796	0.0491
第8天	90420	**0.2204**

相对于前一天统计

	冷启动用户数量	冷启动用户比例
第2天	157577	0.2478
第3天	164510	0.2590
第4天	162943	0.2583
第5天	162405	0.2585
第6天	171007	0.2694
第7天	164211	0.2617
第8天	141257	**0.3443**

2.2 新用户异常

测试集用户冷启动曝光比例异常（均采样到总曝光量一致）

相对于前所有天统计

	老用户日曝光	冷启动用户日曝光
第2天	3.1581	1.7704
第3天	3.0466	1.6316
第4天	2.9012	1.5452
第5天	2.8454	1.5070
第6天	2.8517	1.5274
第7天	2.8578	1.4983
第8天	2.5307	**2.1773**

相对于前一天统计

	老用户日曝光	冷启动用户日曝光
第2天	3.1581	1.7704
第3天	3.3022	1.8363
第4天	3.2297	1.8062
第5天	3.2405	1.8146
第6天	3.2824	1.8598
第7天	3.3251	1.8471
第8天	2.9030	**2.0186**

2.2 新用户异常

测试集用户冷启动异常
线下线上分数差异巨大

- 猜测1：特定的采样方式——>与线上分数不匹配
- 猜测2：广告系统的更新——>无法验证
- 猜测3：部分用户ID丢失——>线上分数匹配

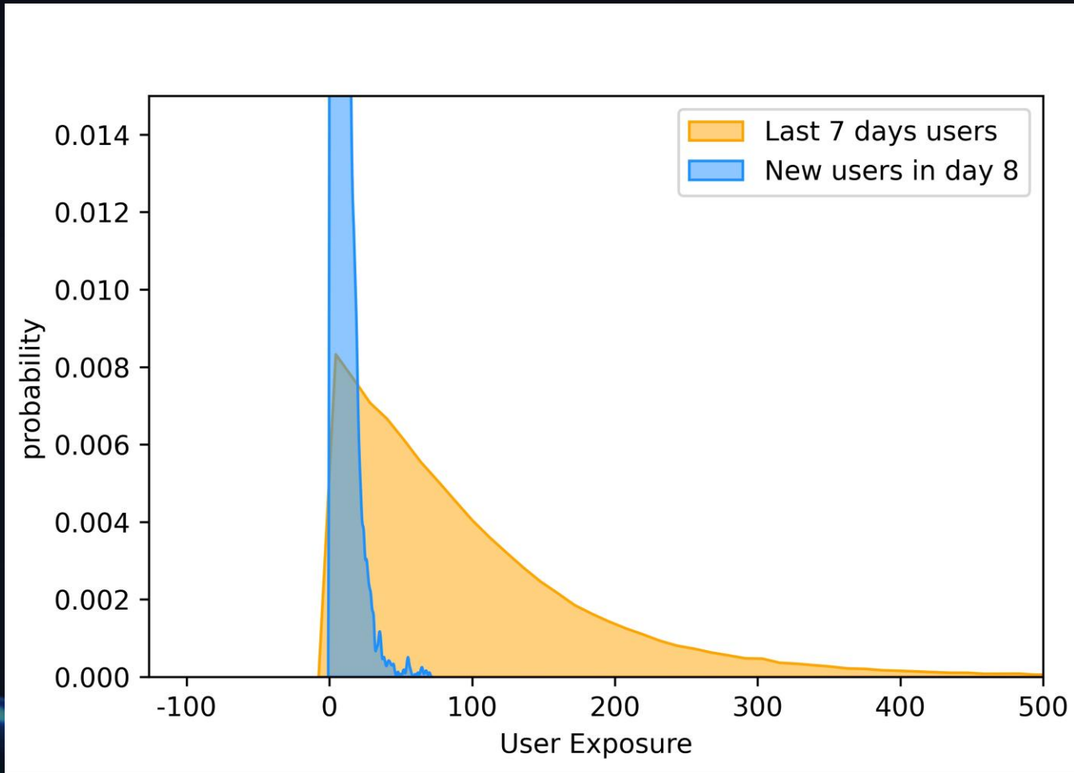
根据猜测三，测试集的用户冷启动中，包含部分ID丢失的老用户， 拥有着新用户的特征分布，却是老用户的行为模式

如何调整新用户的特征分布是解决问题的关键

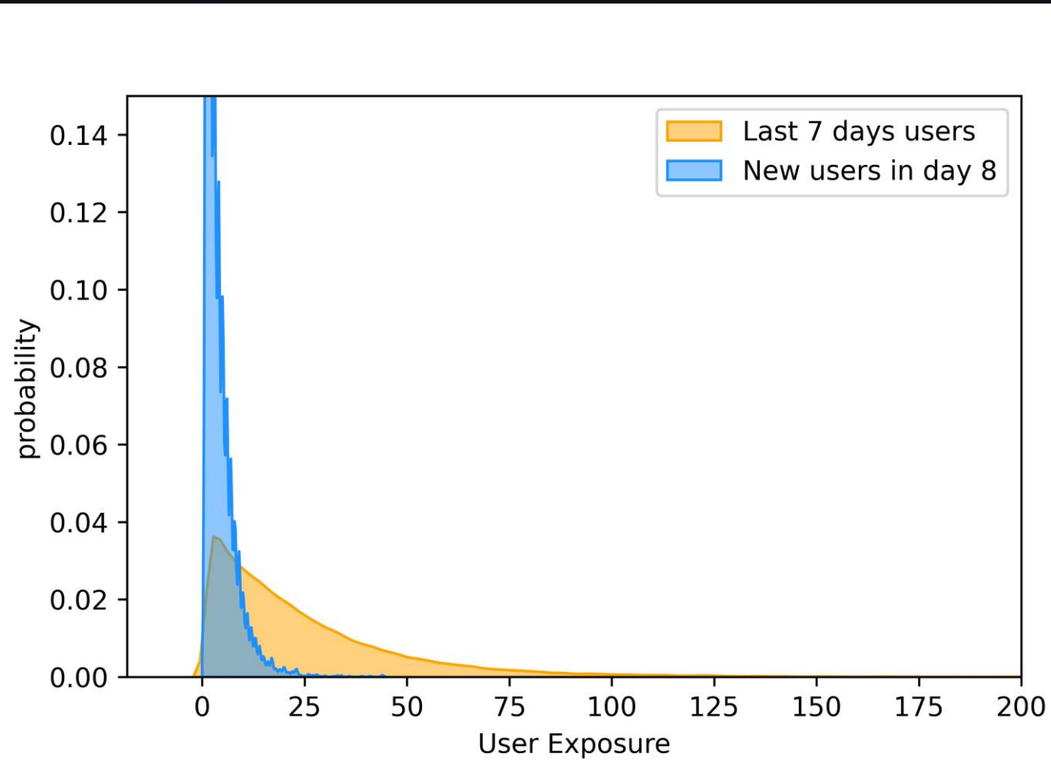
2.3 特征侧调整

方案一：通过采样来减少老用户和新用户特征分布的差异

仅通过负采样+多折验证的方式，选拔赛A榜就可以接近0.8



用户曝光特征的分布



采样后用户曝光特征的分布

2.3 特征侧调整

方案二：通过对特征直接调整使分布与老用户一致

针对曝光特征和交叉特征：**分布迁移**

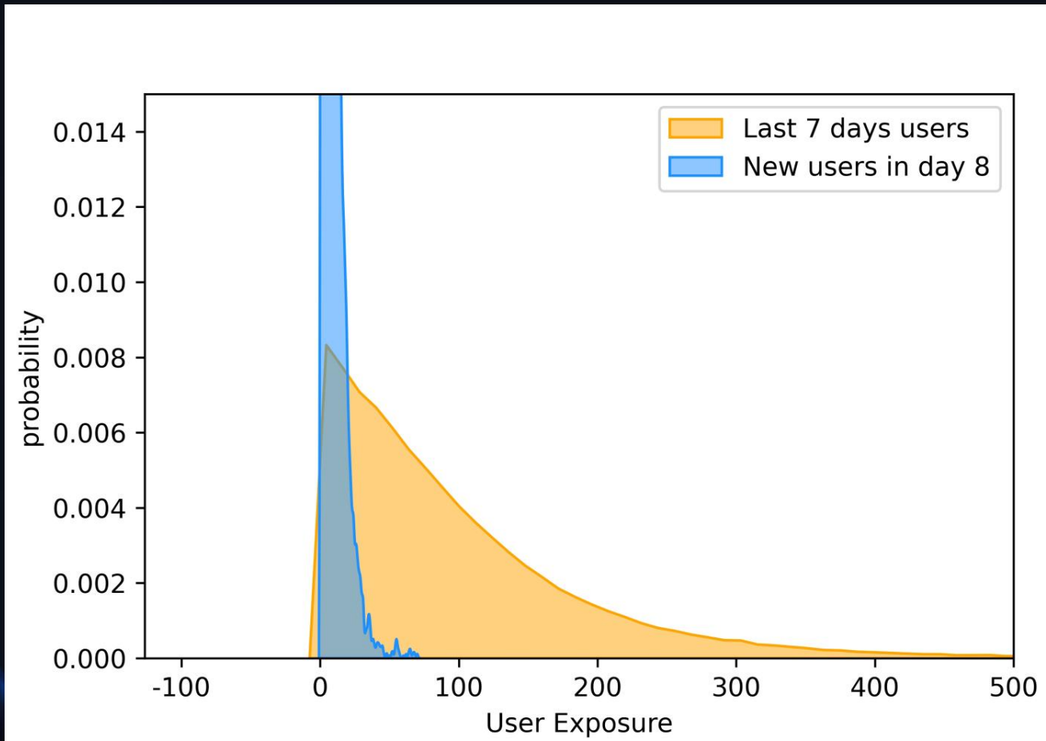
利用样本均值和修正样本标准差来拟合概率分布特性，将第8天冷启动的特征分布归一化到前7天上去

```
mean7 = df[df['pt_d'] < 8][feature].mean()
std7 = df[df['pt_d'] < 8][feature].std()
mean8 = df[(df['pt_d'] >= 8) & (df['coldu'] == 1)][feature].mean()
std8 = df[(df['pt_d'] >= 8) & (df['coldu'] == 1)][feature].std()
df.loc[(df['pt_d'] >= 8) & (df['coldu'] == 1), feature]= ((df[(df['pt_d'] >= 8) & (df['coldu'] == 1)][feature] - mean8) / std8 * std7 + mean7)
```

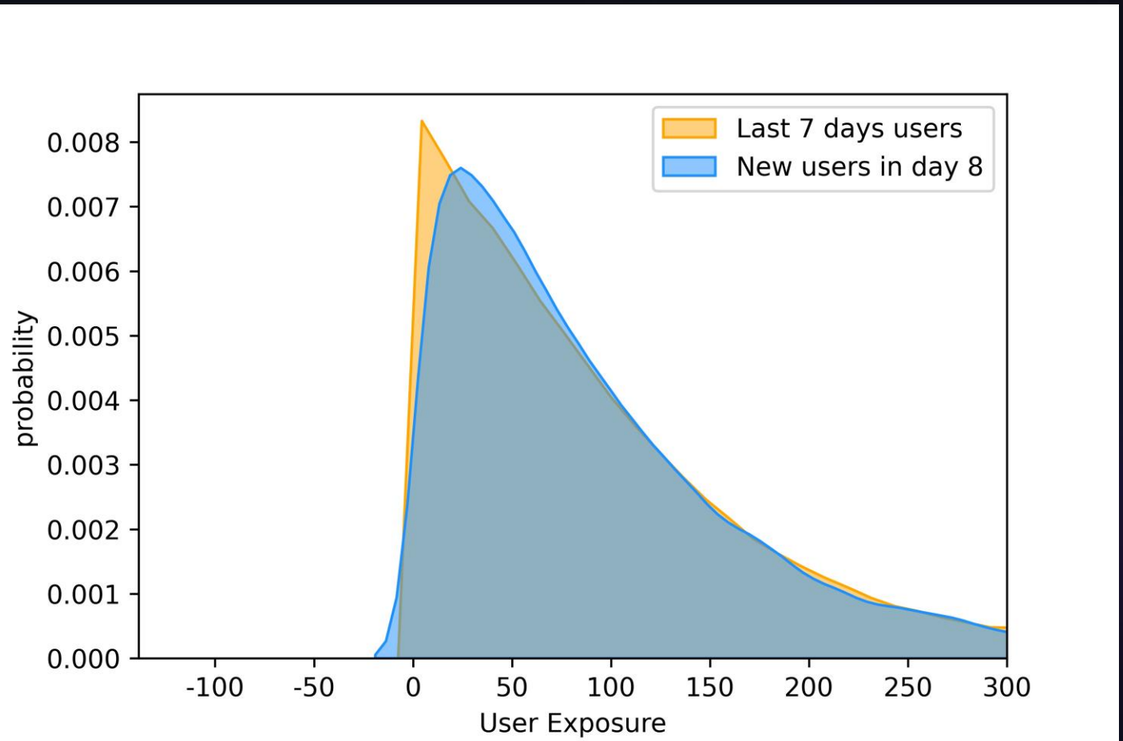
2.3 特征侧调整

方案二：通过对特征直接调整使分布与老用户一致

特征分布迁移到同一分布上，可带来百分位的提升



用户曝光特征的分布



分布迁移后用户曝光特征的分布

2.3 特征侧调整

方案二：通过对特征直接调整使分布与老用户一致

针对CTR特征：1. **特征映射** 2. **特征弱化**

1. **特征映射**：冷启动用所有用户均值填充——>根据当天曝光量映射为同一曝光量用户的均值

百分位提升，仅用该处理即可到选拔赛10名左右

由于训练中CTR特征顺位过高，导致推理时曝光特征和交叉特征得不到充分表达，无法与**分布迁移**兼容，因此放弃

2. **特征弱化**：反其道而行之，将第七天30%的老用户CTR改为均值填充，让模型学会在CTR较弱的时候更加依赖其他特征

千分位提升，可与**分布迁移**兼容，因此选用

2.3 特征侧调整

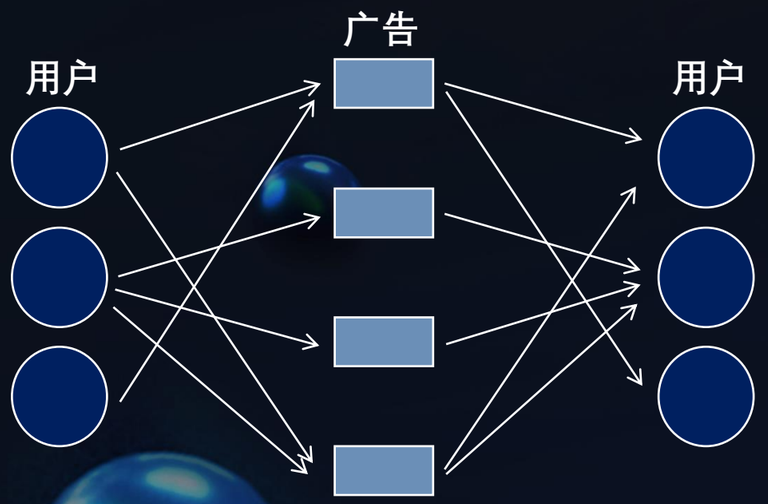
方案二：通过对特征直接调整使分布与老用户一致

针对embedding特征：**GNN传递**

新用户仅有第8天广告信息，与前7天的广告分布有差异

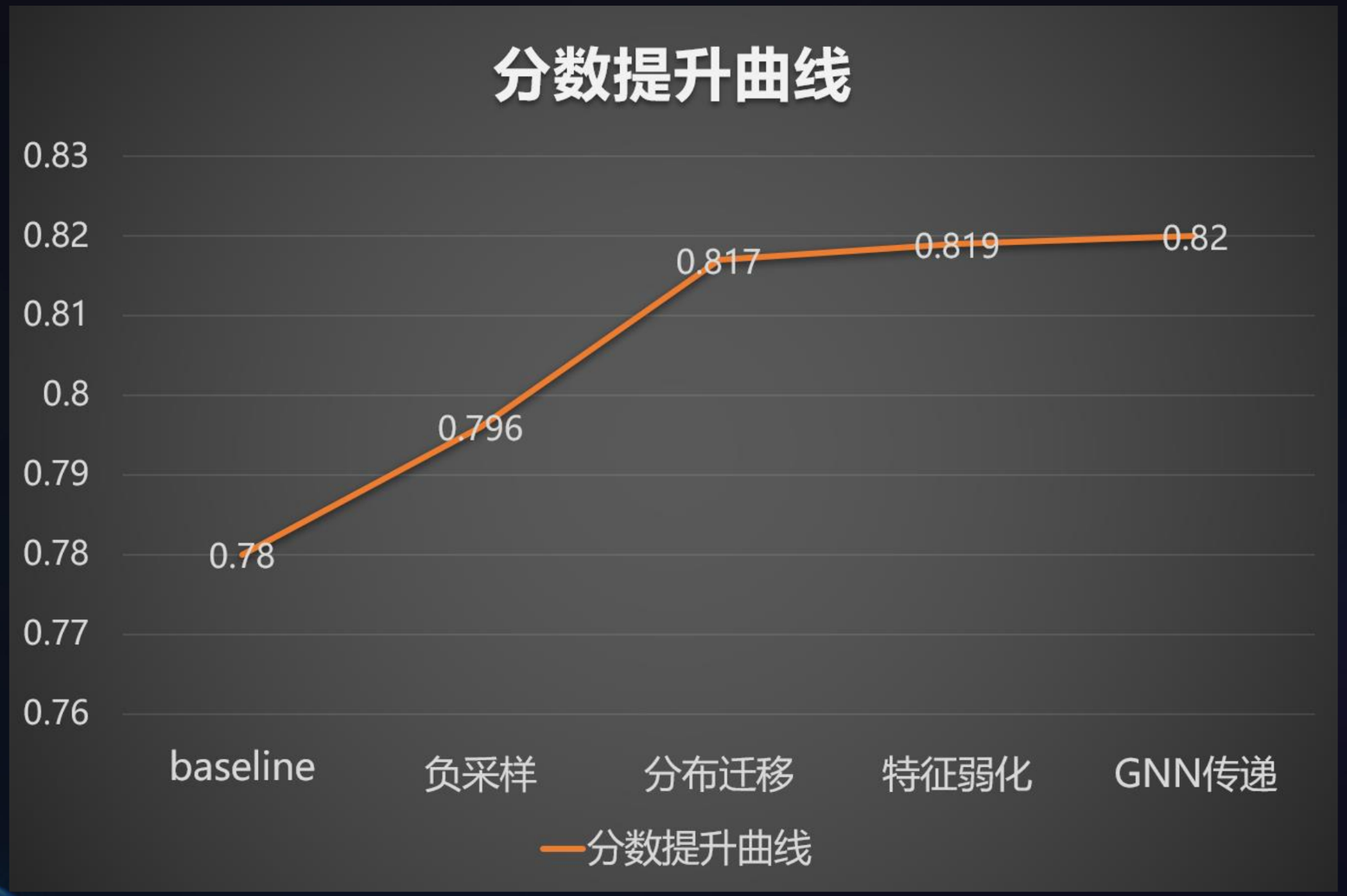
参考GNN中GraphSAGE的消息传递思路，将特征在用户与广告之间再传播一轮，从而联系上前7天的其他广告特征

由于GNN同时也导致了特征平滑，仅有万分位提升



2.3 特征侧调整

选拔赛A榜阶段，特征侧调整方案的效果



03 总结及不足

3.1 方案总结

四大特征

曝光特征

交叉特征

CTR特征

embedding特征

两大方案

负采样+多折

特征调整

三大调整

分布迁移

特征弱化

GNN传递

3.2 不足之处

1. 深度模型的使用
2. 差异化模型的构建
3. 如何针对老用户和新用户设计不同模型
4. 如何减少GNN的平滑性
5. 增量训练、点击率贝叶斯平滑、Graph embedding。。。

谢谢
Thank you.