广告 CTR 预估大作业

一、简介：

CTR(clickthroughrate/点击通过率)是计算广告领域衡量广告效果的重要指标。计算方法为某广告的实际点击次数除以广告的展现量，即$CTR = \frac{Click}{Showcontent}$。预测用户对某个广告的点击率是广告平台的常见需求。

常见的点击率预测模型使用用户在该平台下的行为特征和交互记录进行建模，预测用户对某个广告可能的点击概率。

但是单个平台下的用户特征数据（目标域数据）相对稀疏、用户行为相对单一，预测效果有限。因此我们可以引入相同用户在其他平台下的行为特征数据（源域数据）进行联合建模，深度挖掘用户的兴趣偏好，丰富用户行为特征，提高模型的预测性能。

二、任务要求：

本次任务训练集包括**目标域数据（用户-广告交互记录和目标域用户行为特征数据）**和**源域数据（用户在另一个平台的行为特征数据）**，训练集包含用户 6 天的交互数据。

测试集为 1 天的用户行为特征数据（源域特征和目标域特征）。

本任务要求同学们结合源域数据和目标域数据进行联合建模，预测测试集（final_test.csv）中每条广告的点击概率。

三、数据集：

1) 目标域用户行为数据 train.csv

序号	字段名称	字段含义	是否可为空	字段类型	取值样例
1	label	0：未点击，1：点击	否	int	0，1
2	user_id	用户 id	否	int	1，2…
3	age	年龄	是	int	1，2,3…
4	sex	性别	是	int	1，2…
5	residence	常住地-省份	是	int	1，2…
6	city	常住地-市-编号	是	int	1，2…
7	city_rank	常住地-市-等级	是	int	1，2…
8	series_dev	设备系列	是	int	1，2…
9	series_group	设备系列分组	是	int	1，2…
10	device_dev	系统版本号	是	int	1，2…
11	device_name	用户使用的手机机型	是	int	1，2…
12	device_size	用户使用手机的尺寸	是	int	1，2…

13	net_type	行为发生的网络状态	是	int	1，2…
14	task_id	广告任务唯一标识	是	int	1，2…
15	adv_id	广告任务对应的素材 id	是	int	1，2…
16	creat_type_cd	素材的创意类型 id	是	int	1，2…
17	adv_prim_id	广告任务对应的广告主 id	是	int	1，2…
18	inter_type_cd	广告任务对应的素材的交互类型	是	int	1，2…
19	slot_id	广告位 id	是	int	1，2…
20	site_id	媒体 id	是	int	1，2…
21	spread_app_id	投放广告任务对应的应用 id	是	int	1，2…
22	Tags	广告任务对应的应用的标签	是	int	1，2…
23	app_second_class	广告任务对应的应用的二级分类	是	int	1，2…
24	app_score	app 得分	是	float	4
25	ad_click_list_001	用户点击广告任务 id 列表	是	[string,]	[1^2…]
26	ad_click_list_002	用户点击广告对应广告主 id 列表	是	[string,]	[1^2…]
27	ad_click_list_003	用户点击广告推荐应用列表	是	[string,]	[1^2…]
28	ad_close_list_001	用户关闭广告任务列表	是	[string,]	[1^2…]
29	ad_close_list_002	用户关闭广告对应广告主列表	是	[string,]	[1^2…]
30	ad_close_list_003	用户关闭广告推荐应用列表	是	[string,]	[1^2…]
31	pt_d	时间戳	否	int	20220522 1430
32	log_id	样本 id	否	Int	12345678
33	u_newsCatInterestsST	用户短时兴趣分类偏好	是	[string,]	[1^2…]
34	u_refreshTimes	信息流日均有效刷新次数	是	int	16
35	u_feedLifeCycle	信息流用户活跃度	是	int	12

2) 源域用户行为特征数据 train_feeds.csv

序号	字段名称	字段含义	是否可为空	字段类型	取值样例
1	u_userId	用户标识	否	int	0001
2	u_phonePrice	用户手机价格	是	int	13

3	u_browserLifeCycle	浏览器用户活跃度	是	int	10
4	u_browserMode	浏览器业务类型	是	int	11
5	u_feedLifeCycle	信息流用户活跃度	是	int	12
6	u_refreshTimes	信息流日均有效刷新次数	是	int	16
7	u_newsCatInterests	信息流图文点击分类偏好	是	[string,]	[1^2…]
8	u_newsCatDislike	信息流图文负反馈分类偏好	是	[string,]	[1^2…]
9	u_newsCatInterestsST	用户短时兴趣分类偏好	是	[string,]	[1^2…]
10	u_click_ca2_news	用户图文类别点击序列	是	[string,]	[1^2…]
11	i_docId	文章 docid	是	string	0001
12	i_s_sourceId	文章来源的 sourceid	是	string	0001
13	i_regionEntity	文章地域词 id	是	int	0001
14	i_cat	文章类别 id	是	int	0001
15	i_entities	文章实体词 id	是	[string,]	[1^2…]
16	i_dislikeTimes	文章负反馈量	是	int	60
17	i_upTimes	文章点赞量	是	int	22
18	I_dtype	文章展现形式	是	int	20
19	e_ch	频道	是	int	1,2…
20	e_m	事件来源设备机型	是	int	1,2…
21	e_po	第几位	是	int	9
22	e_pl	拜访地	是	int	1,2…
23	e_rn	第几刷	是	int	1
24	e_section	信息流场景类型	是	int	13
25	e_et	时间戳	否	int	20220522 1430
26	label	是否点击，-1：否，1：是	否	int	1
27	cilLabel	是否点赞，-1：否，1：是	否	int	1
28	pro	文章浏览进度	否	int	1,2…

四、评估方法

统计广告域（目标域）的样本 CTR 预估值，计算 AUC 指标。

五、提交方式

提交结果为一个 csv 文件，该 csv 文件包含两列，分别为 id 和 pctr（点击率预测值），其中 id 为测试集中广告样本序号，顺序与测试集中顺序一致。提交示例如下：

id	pctr
0	0.001413
1	0.028096
…	…