Manual for GLOBAL AI CHALLENGE (Preliminary)

Proposal 3: CTR prediction through cross-domain data from ads and news feeds

1. Introduction

Ad recommendation models are usually built based on historical ad impressions, clicks, and other user behavior data. If only data from the ads domain is used, user behavior data will be sparse, and the user behavior types that can be identified will be limited. However, if a user's behavior data in other domains from the same app is explored, the user's interests and behavior characteristics can be better identified. Of course, introducing user behavior data from other apps can also help enrich the data of user behavior characteristics and ad performance.
You are expected to enhance ads click-through rate (CTR) prediction accuracy by leveraging ad logs, user profiles, and cross-domain data. With ads as the target domain and news feeds as the source domain, you should build user interest models through impressions, clicks, and other user behavior data obtained from the news feeds domain, thus improving the CTR prediction performance of the ads domain.

2. Proposal Description

This proposal provides you with 7-day data for training and 1-day data for testing. The data includes (a) user behavior logs, user profiles, and ad material information in the target domain (ads domain) and (b) user behavior data and basic information about news items in the source domain (news feeds domain). You are required to identify and create representations of user behavior characteristics that map user interests in the source domain and can be applied in the target domain, and build joint models for the source and target domains with reference to user behavior sequences, in order to predict ads CTR. The given data is anonymized to ensure data security.

3. Data Description

The provided data includes data from the target domain (such as user behavior logs, user profiles, and ad material information) and that from the source domain (such as user behavior data and basic information about news items).

3.1 Data from the Target Domain

No.	Field	Description	Can Be Empty	Type	Example Value
1	label	Tag indicating whether a user clicks the ad. The	No	Int	0, 1

No.	Field	Description	Can Be Empty	Type	Example Value
		value can be **0** (no) or **1** (yes).			
2	user_id	User ID.	No	String	1, 2...
3	age	Age.	Yes	String	1, 2, 3...
4	gender	Gender.	Yes	String	1, 2...
5	residence	Permanent residence (province).	Yes	String	1, 2...
6	city	Permanent residence (city ID).	Yes	String	1, 2...
7	city_rank	Permanent residence (city level).	Yes	String	1, 2...
8	series_dev	Device series.	Yes	String	1, 2...
9	series_group	Device series group.	Yes	String	1, 2...
10	emui_dev	EMUI version.	Yes	String	1, 2...
11	device_name	Model of the device used by a user.	Yes	String	1, 2...
12	device_size	Size of the device used by a user.	Yes	String	1, 2...
13	net_type	Network under use when a behavior occurs.	Yes	String	1, 2...
14	task_id	Unique ID of an ad task.	Yes	String	1, 2...
15	adv_id	ID of the material used by an ad task.	Yes	String	1, 2...
16	creat_type_cd	Creative type ID corresponding to a material.	Yes	String	1, 2...
17	adv_prim_id	ID of the advertiser who creates the ad task.	Yes	String	1, 2...
18	inter_type_cd	Material display form of an ad task.	Yes	String	1, 2...
19	slot_id	Ad slot ID.	Yes	String	1, 2...
20	site_id	ID of the media app.	Yes	String	1, 2...
21	spread_app_id	ID of the advertised app.	Yes	String	1, 2...
22	Tags	App tag of an ad task.	Yes	String	1, 2...
23	app_second_class	Level-2 category of the advertised app.	Yes	String	1, 2...
24	app_score	App rating score.	Yes	Int	4
25	ad_click_list_001	ID list of ad tasks clicked by a user.	Yes	[String,]	[1^2...]
26	ad_click_list_002	ID list of advertisers	Yes	[String,]	[1^2...]

No.	Field	Description	Can Be Empty	Type	Example Value
		whose ads are clicked by a user.			
27	ad_click_list_003	List of advertised apps clicked by a user.	Yes	[String,]	[1^2...]
28	ad_close_list_001	List of ad tasks closed by a user.	Yes	[String,]	[1^2...]
29	ad_close_list_002	List of advertisers whose ad tasks are closed by a user.	Yes	[String,]	[1^2...]
30	ad_close_list_003	List of advertised apps closed by a user.	Yes	[String,]	[1^2...]
31	pt_d	Timestamp.	No	String	202205221430
32	log_id	Sample ID.	No	Int	12345678

3.2 Data from the Source Domain

No.	Field	Description	Can Be Empty	Type	Example Value
1	u_userId	User ID.	No	String	0001
2	u_phonePrice	Price of a user's device.	Yes	String	13
3	u_browserLifeCycle	User engagement on Browser.	Yes	String	10
4	u_browserMode	Browser service type.	Yes	String	11
5	u_feedLifeCycle	User engagement on news feeds.	Yes	String	12
6	u_refreshTimes	Average number of valid news feeds updates per day.	Yes	String	16
7	u_newsCatInterests	Liked news feeds categories based on the click behavior of a user.	Yes	[String,]	[1^2...]
8	u_newsCatDislike	Disliked news feeds categories based on negative comments.	Yes	[String,]	[1^2...]
9	u_newsCatInterestsST	Liked news feeds categories based on a user's short-term interests.	Yes	[String,]	[1^2...]
10	u_click_ca2_news	Click sequence of the	Yes	[String,]	[1^2...]

No.	Field	Description	Can Be Empty	Type	Example Value
		categories of images and texts.			
11	i_docId	Article ID.	Yes	String	0001
12	i_s_sourceId	Article source ID.	Yes	String	0001
13	i_regionEntity	Geographic word ID in an article.	Yes	String	0001
14	i_cat	Article type ID.	Yes	String	0001
15	i_entities	Named entity word ID in an article.	Yes	[String,]	[1^2...]
16	i_dislikeTimes	Number of negative comments for an article.	Yes	String	60
17	i_upTimes	Likes on an article.	Yes	String	22
18	I_dtype	Article display mode.	Yes	String	20
19	e_ch	Channel.	Yes	String	1, 2...
20	e_m	Device model on which the event occurs.	Yes	String	1, 2...
21	e_po	Position.	Yes	String	9
22	e_pl	Visited location.	Yes	String	1, 2...
23	e_rn	Sequence number of updates.	Yes	String	1
24	e_section	News feeds scenario type.	Yes	String	13
25	e_et	Timestamp.	No	String	202205221430
26	label	Tag indicating whether a user clicks the news feed. The value can be **–1** (no) or **1** (yes).	No	String	1
27	cilLabel	Tag indicating whether a user likes the news feed. The value can be **–1** (no) or **1** (yes).	No	String	1
28	pro	Article browsing progress.	No	String	1, 2...

4. How We Score

Scoring method: Collect the predicted ads CTR values of the samples in the ads domain, and calculate the GAUCs and AUCs (AUC, area under the ROC curve).

Scoring indicator: The sum of weighted GAUC and AUC will serve as the scoring

indicator. The formula is as follows:
xAUC = α x GAUC + β x AUC
A higher xAUC means a better result, and thus a higher ranking.
AUC in the formula is the sum of the AUCs of all samples, and GAUC refers to the weighted sum of group AUCs. The samples are grouped by user. The group weight is the ad impressions in a group divided by the total impressions.

$$\text{GAUC} = \frac{\sum_{k=i}^{n} AUC_i * Impression_i}{\sum_{k=i}^{n} Impression_i}$$

Weights used for the preliminary round: α is 0.7; $\beta\ is\ 0.3$.

5. How to Submit

Submit a **submission.csv** file encoded in UTF-8 without BOM. The file content contains **log_id** and **pctr** in a format, which refer to the log ID of the corresponding test sample, and the predicted CTR of the test sample calculated by your model, respectively. The value of **pctr** shall contain six decimal places.
The file format example is as follows:

```
log_id, pctr
  1, 0.002345
  2, 0.010456
  ...
```