Data Science Hire Technical Challenge: Click-Through Rate Prediction

Actions

Updated Today All Users

Task Description

The goal of this challenge is to predict the user's click response in real-time bidding (RTB) display advertising (it is NOT essential to know RTB for this, but might worth checking the first few slides of our tutorial). Specifically, given the information of the incoming bid request, the bid agent should estimate the probability that the user will click on the ad if displayed.

For this purpose, you are asked to implement a model to predict the click-through rate (CTR).

The dataset used in this challenge is sampled and reformulated from iPinYou RTB contest dataset. This dataset is in a record-per-line, `\t` separated value (tsv) format. Each line contains the label, i.e., click (1) or no click (0), and the corresponding bid request and ad features. The detailed description of the fields is given with the dataset.

Data Fields

```
Column  Name
0       Click
1       Weekday
2       Hour
3       Timestamp
4       Log Type
5       User ID
6       User-Agent
7       IP
8       Region
9       City
10      Ad Exchange
11      Domain
12      URL
13      Anonymous URL ID
14      Ad slot ID
15      Ad slot width
16      Ad slot height
17      Ad slot visibility
18      Ad slot format
19      Ad slot floor price (RMB/CPM)
20      Creative ID
21      Key Page URL
22      Advertiser ID
```

```
23      User Tags
```

The `click` column is the label, which is in `train.txt` but missing in `test.txt`. Other fields may be used as features for the prediction.

Download the Dataset

train.txt.gz

test.txt.gz

sample_submission.csv

Submission and Evaluation

There are two parts of your submission: code (notebook) and predictions.

We encourage you to initialise your work in a git repo and record any changes. You could either zip and email the workspace along with `.git`, or upload it to Github. We encourage you to use Jupyter (iPython) notebook and keep everything in the same notebook, so 'Run all' command could easily validate your work, also reproduce outputs from cells. Given the scale of the dataset we don't ask you to run a cluster, however you could to demonstrate skills.

You also need to use the model to predict labels for `test.txt`. The AUC of your prediction will be discussed at the face-to-face interview.

Send you submission or any questions to Shuai Yuan. Good luck and have fun!

Last Author Tom
Subscribers ShuaiYuan, junwang