人工智能工程师直通车 项目 2: CTR 预估

1、任务简介：

广告点击率（Click-Through Rate Prediction, CTR）是互联网计算广告中的关键环节，预估准确性直接影响公司广告收入。机器学习技术可在计算广告中大展身手，Avazu 通过程序化广告技术进行效果营销。本项目我们对 Avazu 提供的 Kaggle 竞赛数据进行移动 CTR 预估，其 Kaggle 竞赛网页为：https://www.kaggle.com/c/avazu-ctr-prediction。

2、数据描述：

数据共包含 11 天的数据，其中 10 天为训练数据 train，1 天为测试数据 test。

（1） 文件说明

- train: 训练数据，包括 10 天的 CTR 数据，数据按时间顺序排列。被点击的样本和没被点击的样本根据不同的策略已经进行了下采样。
- test ：测试数据，1 天的广告数据，用于测试模型的预测。
- sampleSubmission.csv：提交结果文件样例。 corresponds to the All-0.5 Benchmark.

提交测试结果包含两个字段，分别为测试样本索引及其被点击的概率，格式如下：

```
id,click
60000000,0.384
63895816,0.5919
759281658,0.1934
895936184,0.9572
```

（2） 字段说明

- id: ad identifier （广告 ID）
- click: 0/1 for non-click/click （是否被点击，其中 0 为不被点击，1 为不被点击，此列为目标变量）
- hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC. （时间）
- C1 -- anonymized categorical variable （类别型变量）
- banner_pos （广告位置）
- site_id （站点 ID）

- site_domain （站点领域）
- site_category （站点类别）
- app_id （APP ID）
- app_domai
- C14-C21 -- anonymized categorical variables （类别型变量）

3、评价标准：

项目采用 logloss 评价模型性能。令y_i为第 i 个样本真正的标签，$\hat{y}_i$为模型对第 i 个的预测值，N 为样本数据，则 logloss 为

$$logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i)(1-y_i)\log(1-\hat{y}_i)$$ 。

4、提示：

(1) train 解压后文件有 5.6G，样本数目非常大（40M），项目参数调优时可考虑进行下采样。

可用 pandas：（选择 10M 样本）

```
train_df=pd.read_csv('train.csv',nrows=10000000)
```

或者用命令行：

```
Windows: get-content file | select-object -first 2000000

Unix (Mac/Linux): Type the following head command to display first 10 lines of the file :
head -10 train.csv

Use the pipe (">") command to save it to another file head -2000000 train.csv > subsetTrain.csv

gshuf -2000000 train.csv > new file.csv to extract random rows from the data and can download gshuf from home-brew
```

(2) 特征均为类别型特征，建议根据特征特点进行合适的特征工程。

(3) 模型建议考虑：FTRL、GBDT+FTRL、FFM