腾讯社交广告
Tencent Social Ads

高校算法大赛

我很难受

演讲人 李 强

腾讯社交广告
Tencent Social Ads

高校算法大赛

团队介绍

框架设计

特征工程

模型融合

腾讯社交广告
Tencent Social Ads

高校算法大赛

01 团队介绍

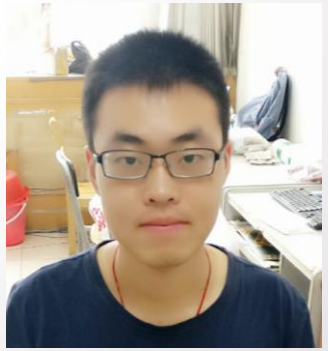
李　强
吉林大学

研一在读
计算机专业

李　智
北京航空航天大学
研二在读
电子科学与技术

我很难受

李博
北京邮电大学
研一在读
自动化专业



腾讯社交广告
Tencent Social Ads

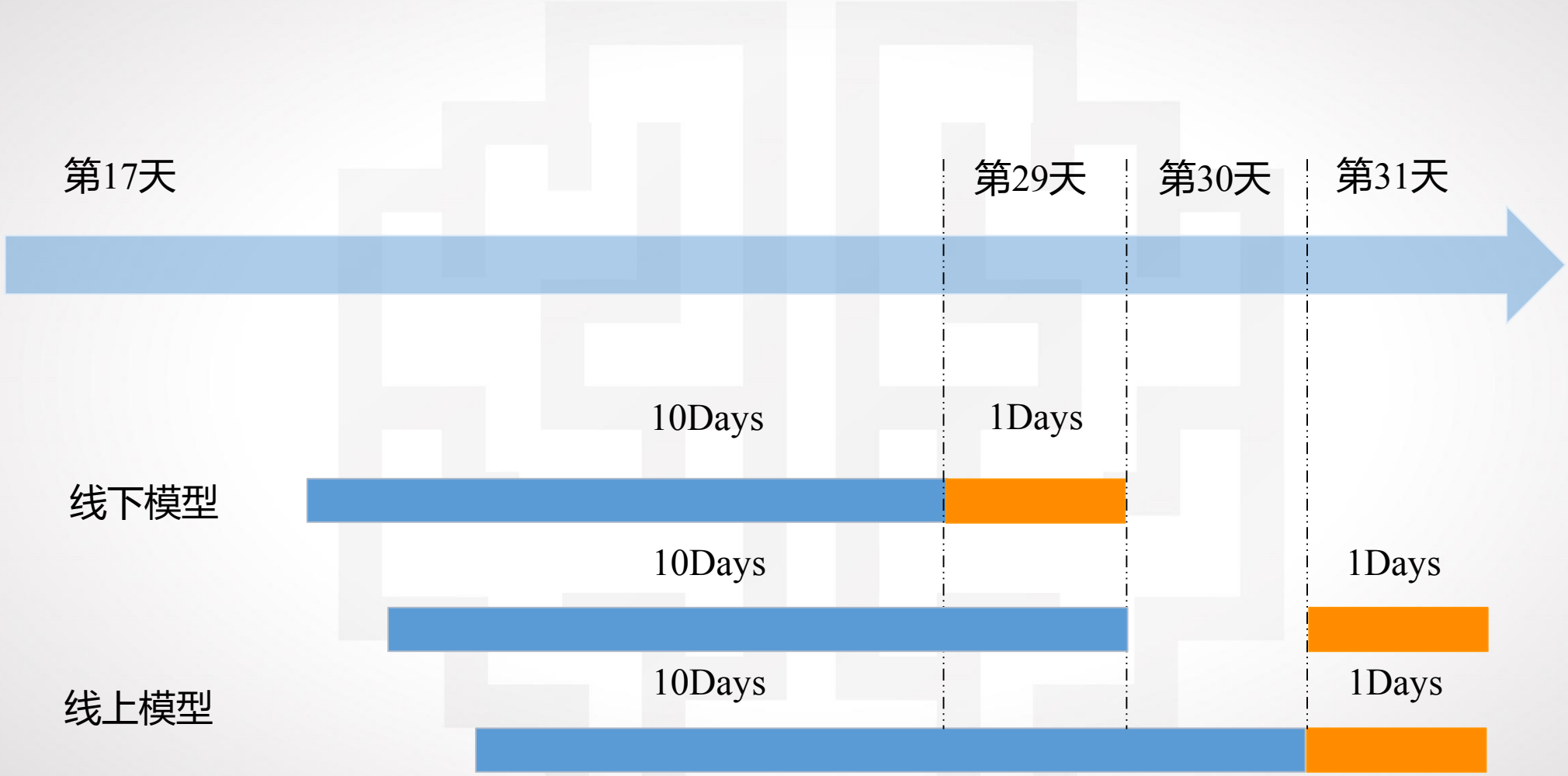
高校算法大赛

02 框架设计

给定：用户信息表，app种类表，广告信息表，广告位信息表，第一天之前用户安装的app列表，第1天到第30天的app安装列表，第17天到30天的训练表。

预测：用户在第31天点击某个广告后会发生转化的概率。

评价： $$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i\, log(p_i) + (1-y_i)log(1-p_i))$$


第17天
第29天
第30天
第31天
10Days
1Days
线下模型
10Days
1Days
线上模型
10Days
1Days

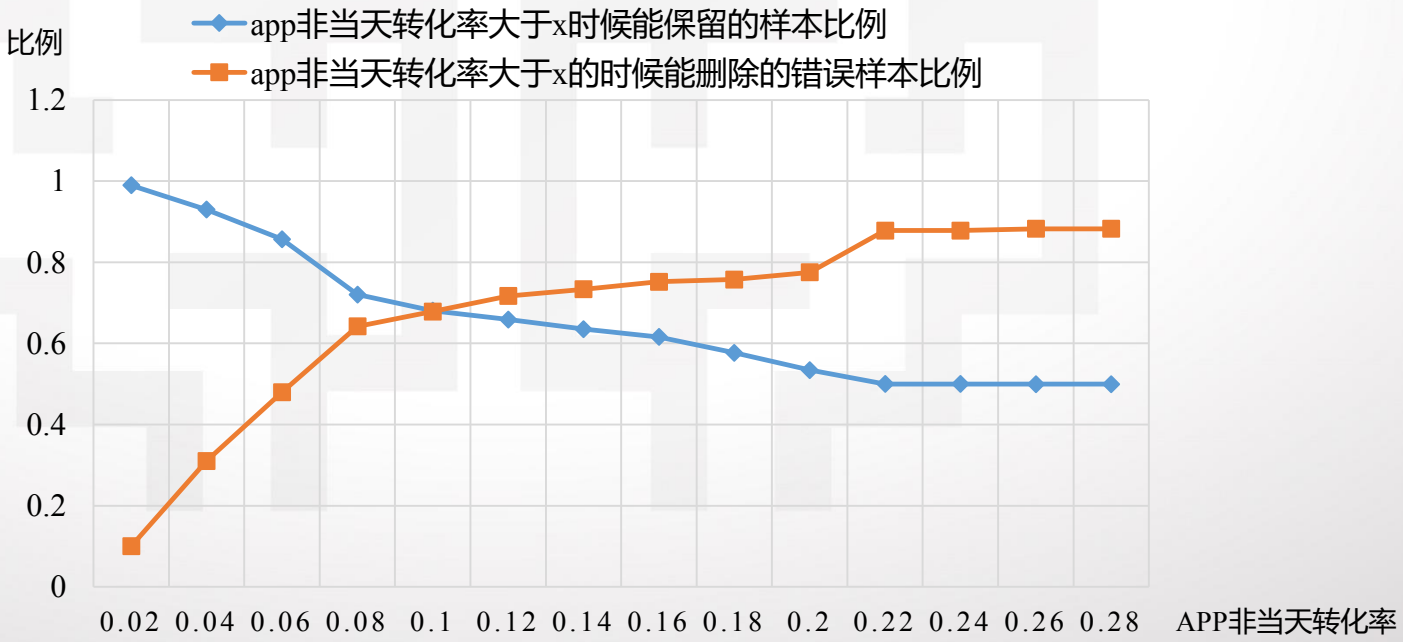
腾讯社交广告
Tencent Social Ads

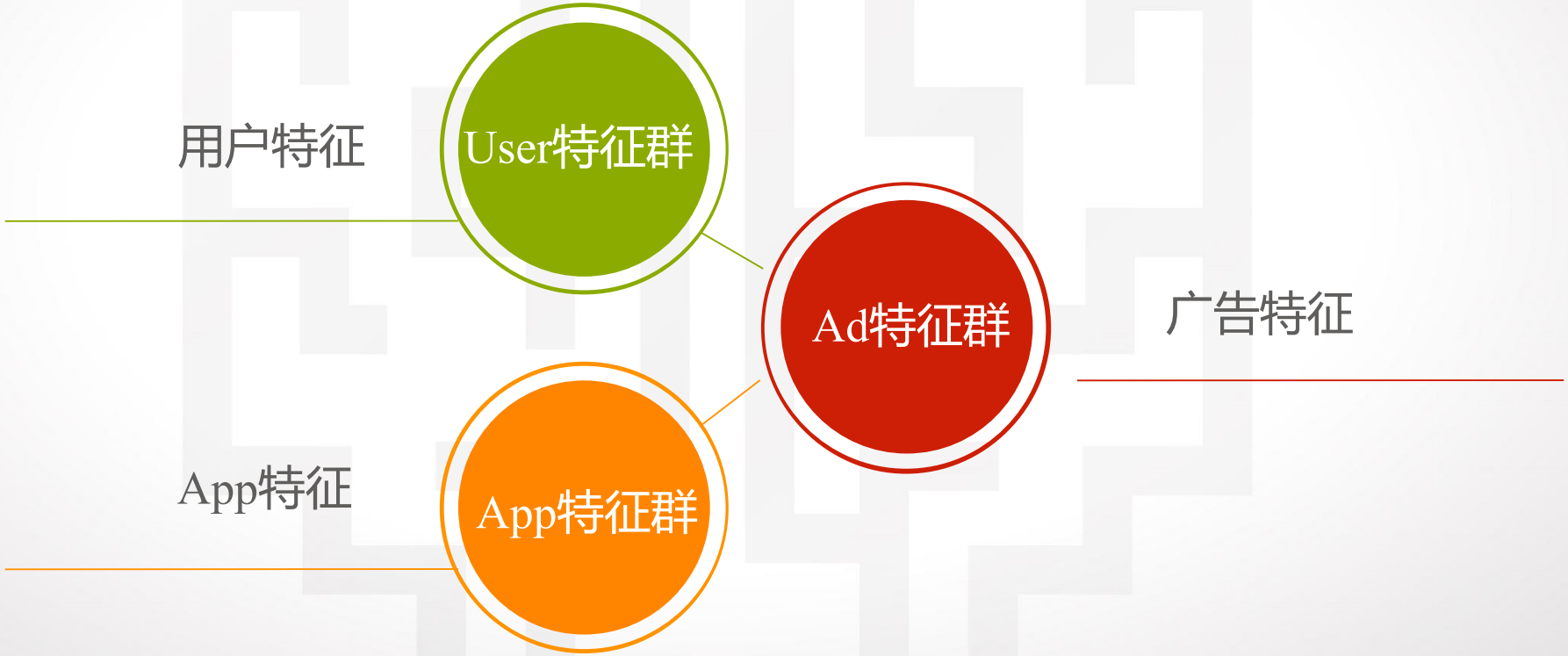
高校算法大赛

03 特征工程

训练数据的末尾几天由于转化时间的滞后性，存在错误标签的样本，越靠近第30天，错误样本的比例越大，线上成绩反馈表明，直接将第30天放入训练样本中训练，成绩将会大幅下降。为此我们定义了app非当天转化率来对数据做清洗：

app非当天转化率 = app非当天转化总数 / app转化总数




用户特征
User特征群
Ad特征群
广告特征
App特征
App特征群

基础id特征

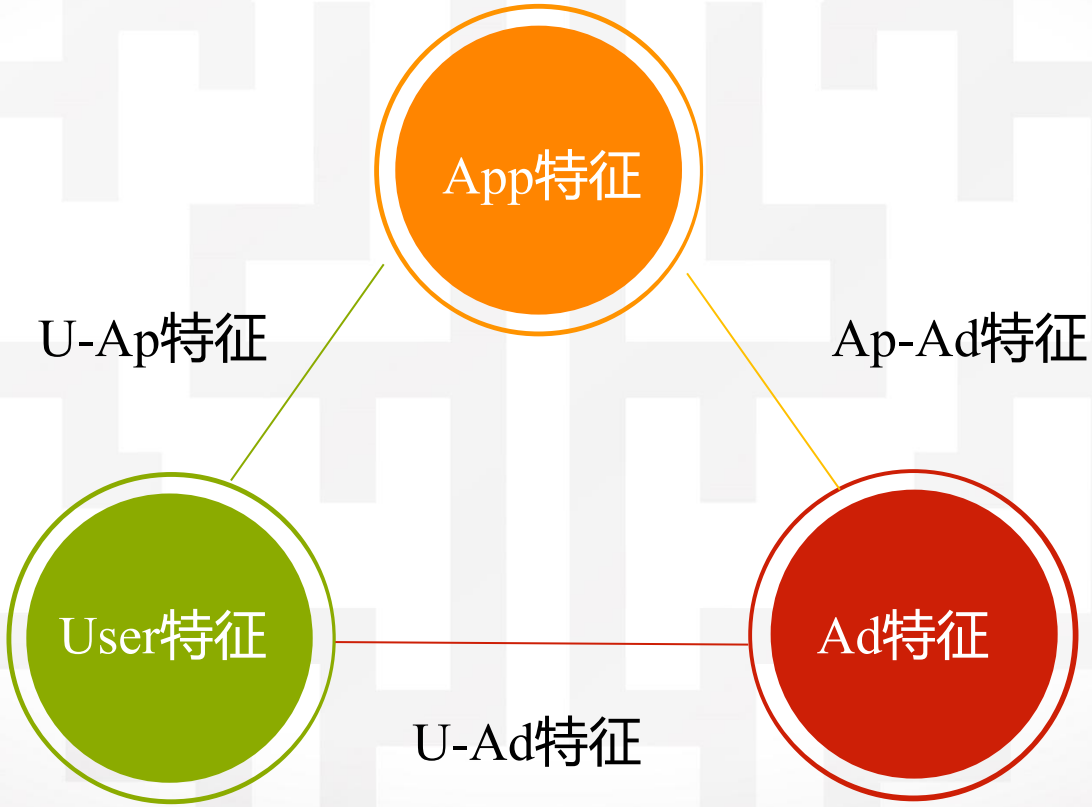
- 广告类id
- 广告位id
- app类id
- 用户id

统计特征

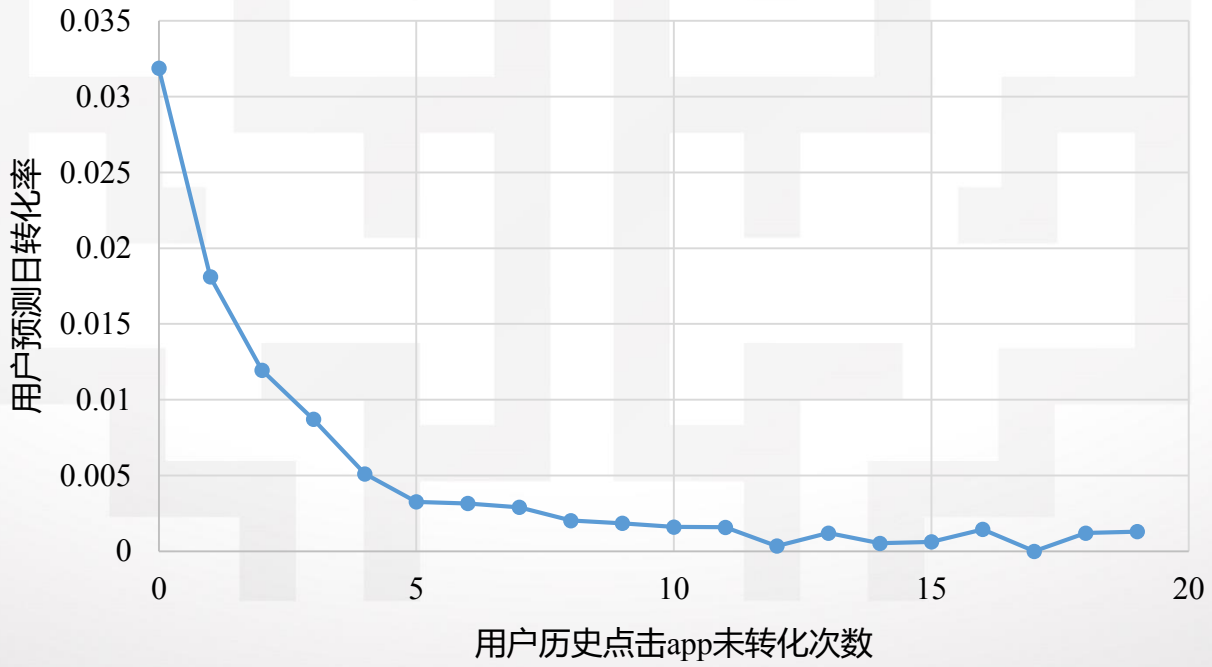
- 点击次数
- 转化次数
- 转化率

时间特征

- 时间间隔
- 用户活跃度
- 最早最晚时间


App特征
U-Ap特征
Ap-Ad特征
User特征
Ad特征
U-Ad特征

因为数据量庞大，盲目的通过模型测试特征变得十分浪费时间，因此我们在每次测试特征前都会对当前特征做分析，当特征在不同取值下转化率有较大差别的情况下才会用模型测试。



- 广告，广告平台，app的历史转化率作为特征能极大提升预估的准确性，但是当数据量比较少的时候，历史转化率与实际转化率相差很大，极端情况下，广告就被点击1次并被转化了，直接计算转化率为1与实际转化率相差很大.

  引入贝叶斯平滑对历史转化率做出修正：

$$r = (C + \alpha)/(I + \alpha + \beta)$$

- 通过划窗来增加样本势必会带来时间跨度不统一而带来的特征量级不同的问题，我们用日均点击量代替历史点击量，日均转化量代替历史转化量，使得特征量级归一化。

腾讯社交广告
Tencent Social Ads

高校算法大赛

04 模型融合

本次比赛，我们分别训练了三种模型，分别是xgboost，lightGBM以及FFM。

xgb

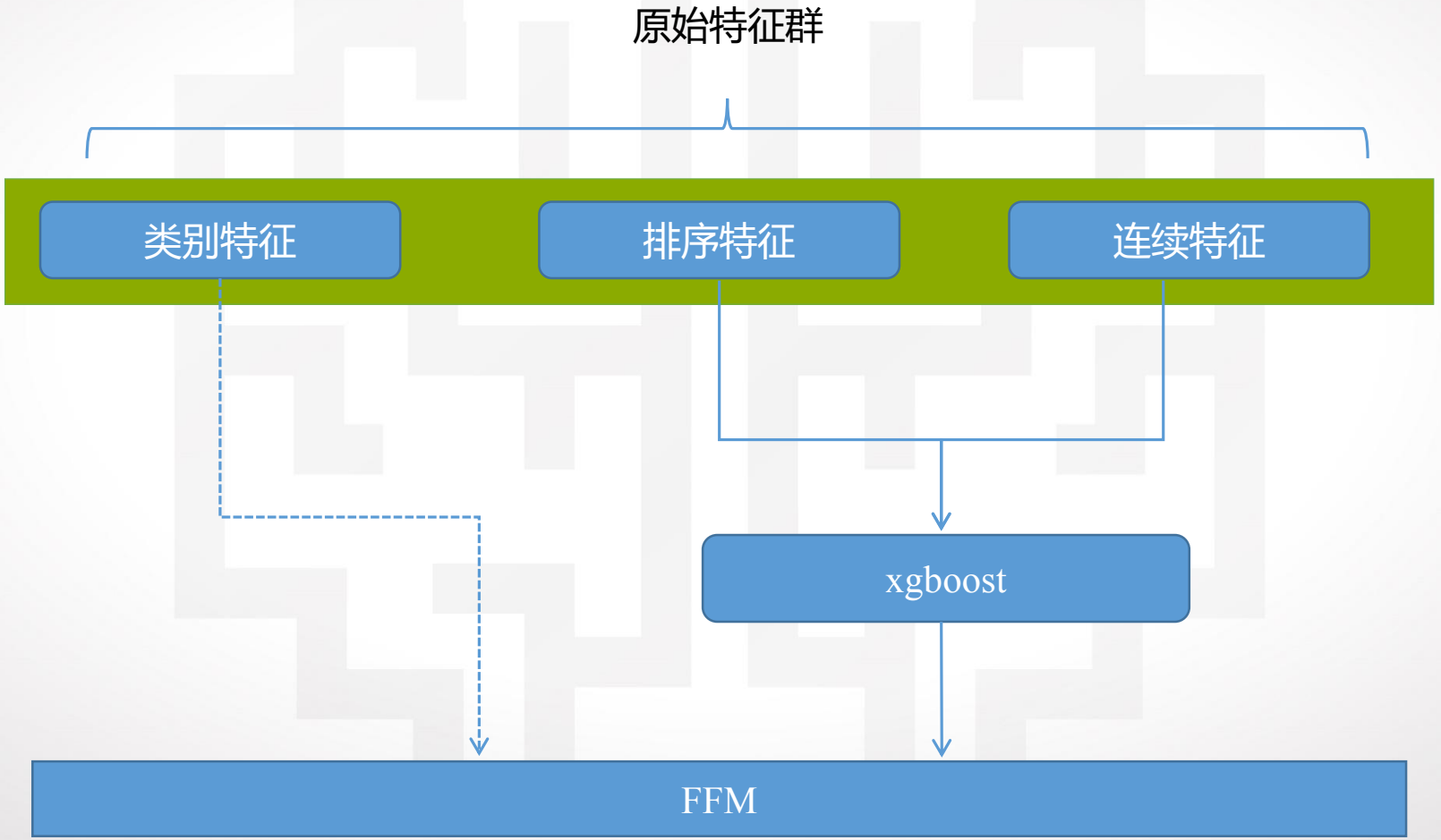
线上B榜0.101435

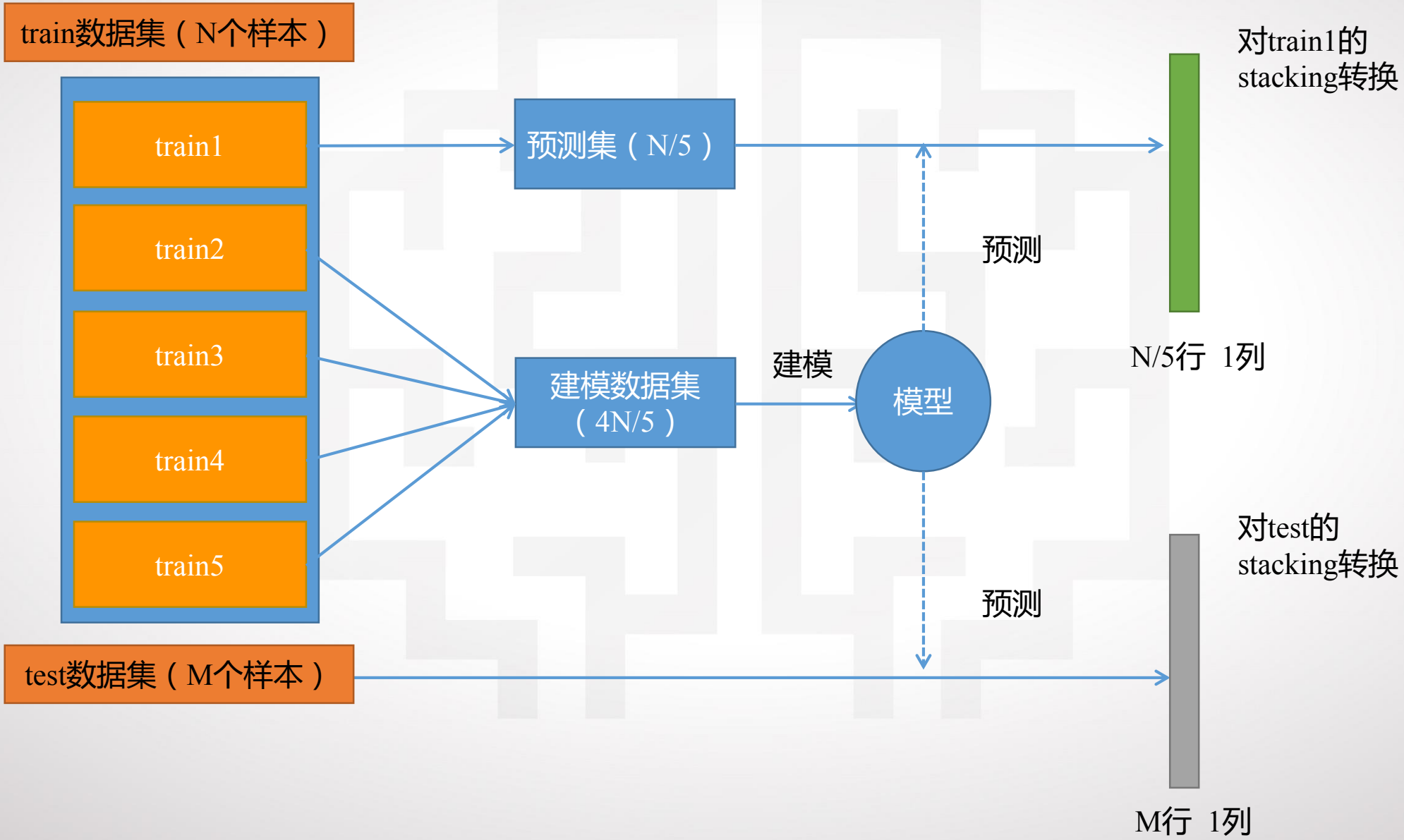
lgb

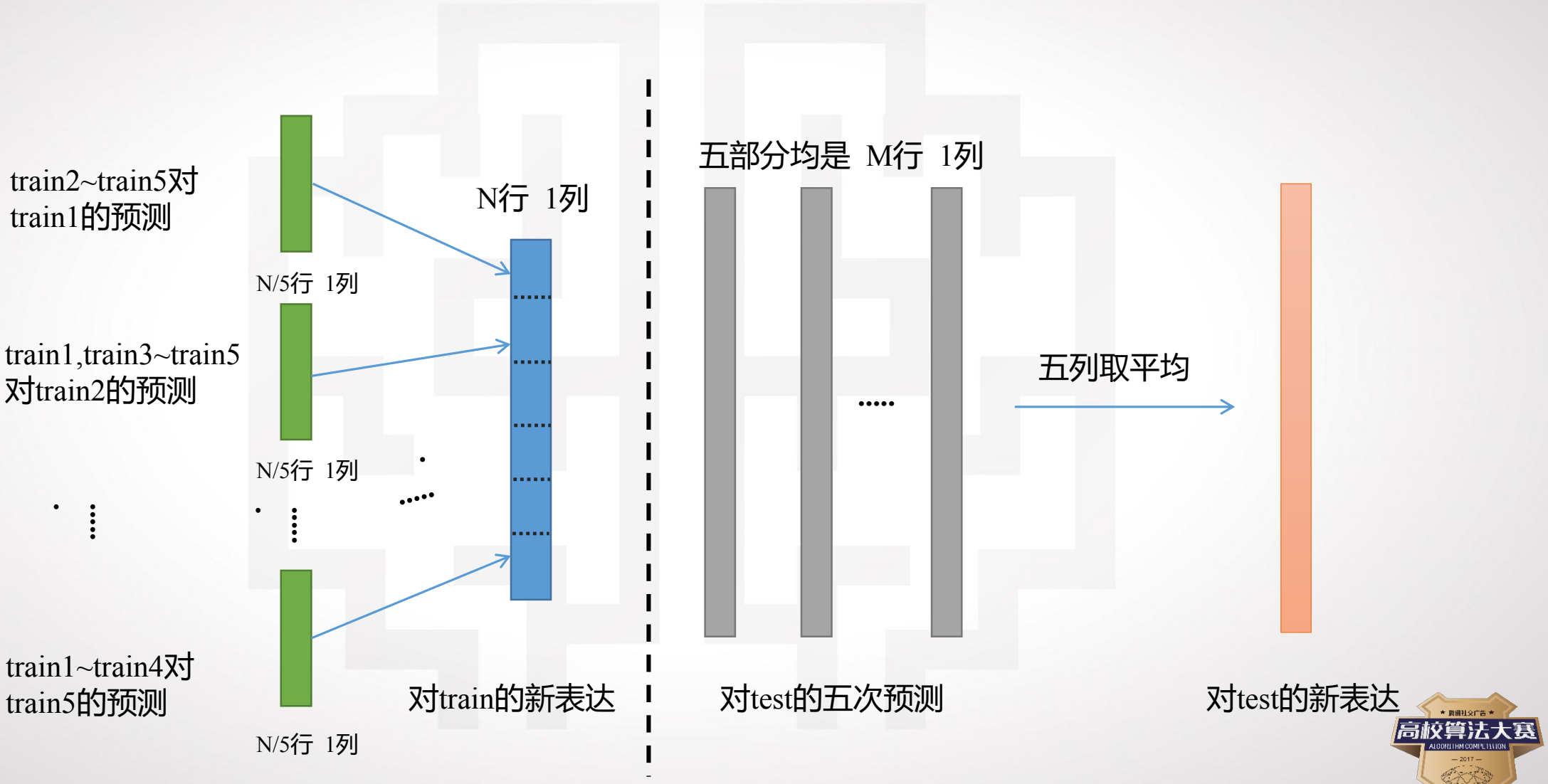
线上B榜0.101416

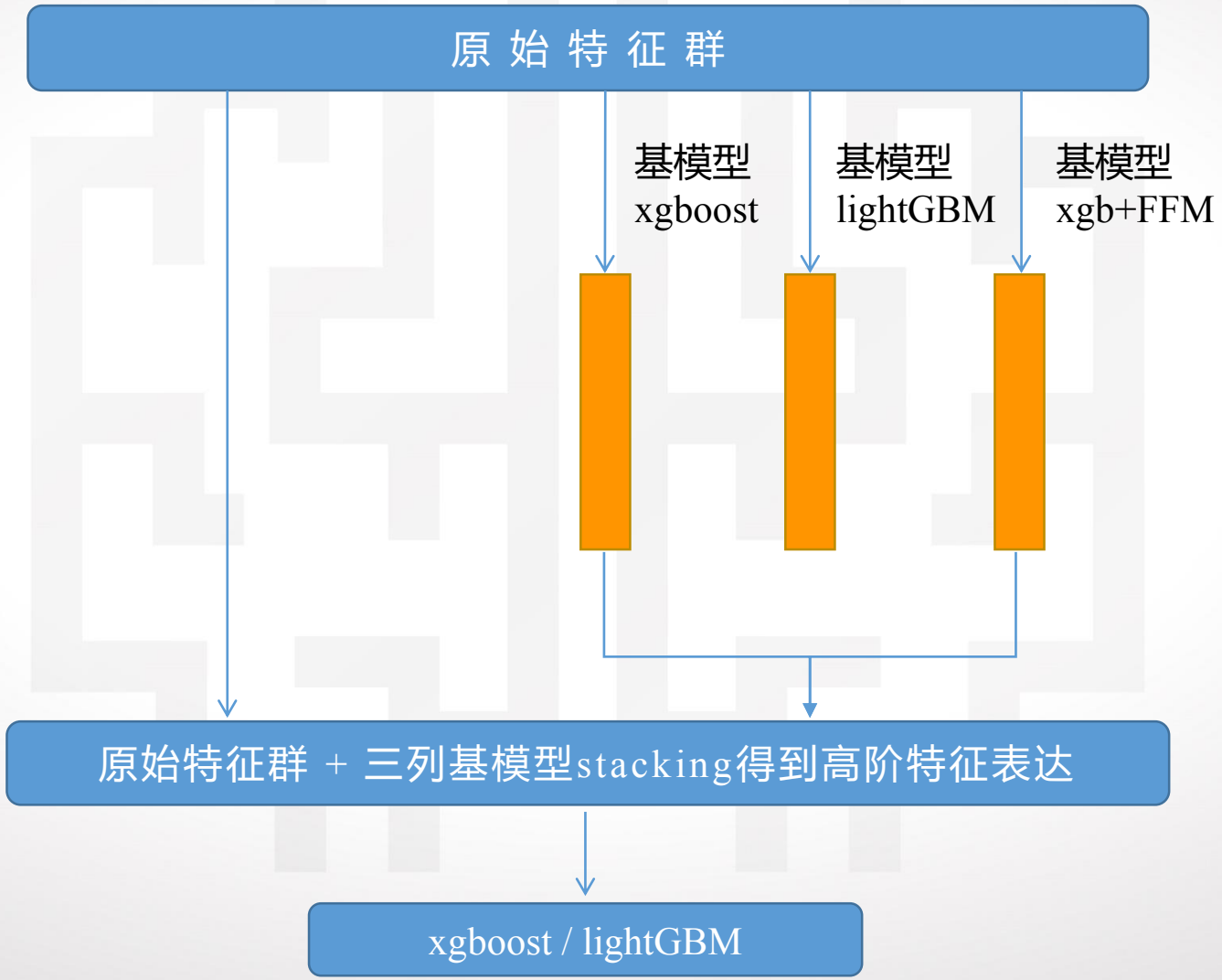
ffm

线上B榜0.1024左右


原始特征群
类别特征
排序特征
连续特征
xgboost
FFM


train数据集（N个样本）
train1
train2
train3
train4
train5
预测集（N/5）
建模数据集（4N/5）
建模
模型
预测
对train1的stacking转换
N/5行 1列
test数据集（M个样本）
预测
对test的stacking转换
M行 1列


train2~train5对
train1的预测
N/5行 1列
N行 1列
train1,train3~train5
对train2的预测
N/5行 1列
train1~train4对
train5的预测
N/5行 1列
对train的新表达
五部分均是 M行 1列
五列取平均
对test的五次预测
对test的新表达


原 始 特 征 群
基模型
xgboost
基模型
lightGBM
基模型
xgb+FFM
原始特征群 + 三列基模型stacking得到高阶特征表达
xgboost / lightGBM

腾讯社交广告
Tencent Social Ads