腾讯社交广告
Tencent Social Ads

高校算法大赛

团队介绍 – SYSU_至九

队长：
郭达雅
中山大学本科生，计算机科学与技术
特征工程、数据探索

队员：
张俊鸿
中山大学本科生，计算机科学与技术
模型设计与融合

刘昕
中山大学本科生，计算机科学与技术
数据清洗、算法实现

目录

赛题分析

赛题分析

竞赛题目：

竞赛题目以移动App广告为研究对象，对给定广告、用户和上下文情况下，预测App广告点击后被激活的概率。

赛题分析:

对于CVR问题来说，App广告是否被激活的主导因素是用户，其次是广告信息。

特征工程

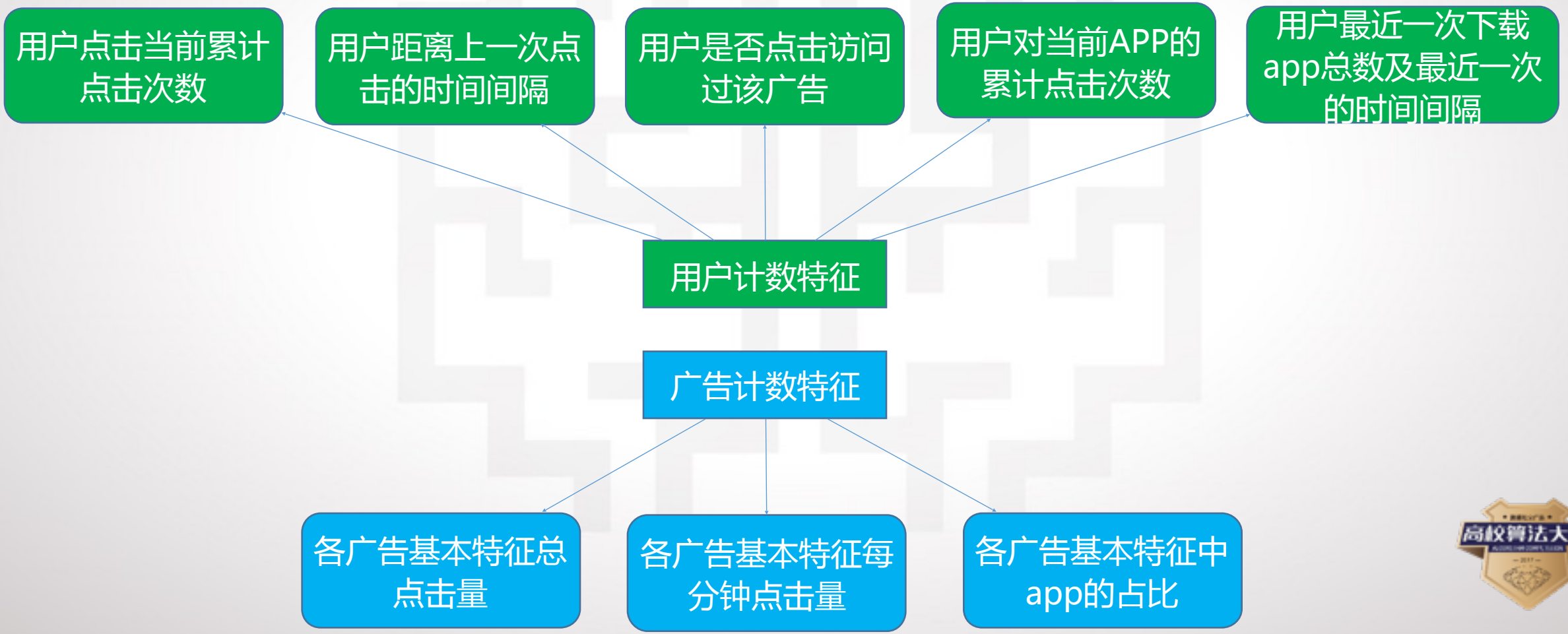
特征工程

除了基本特征，我们还生成了以下特征:

- 计数特征
- 转化率的贝叶斯平滑
- 用户历史点击
- Word embedding

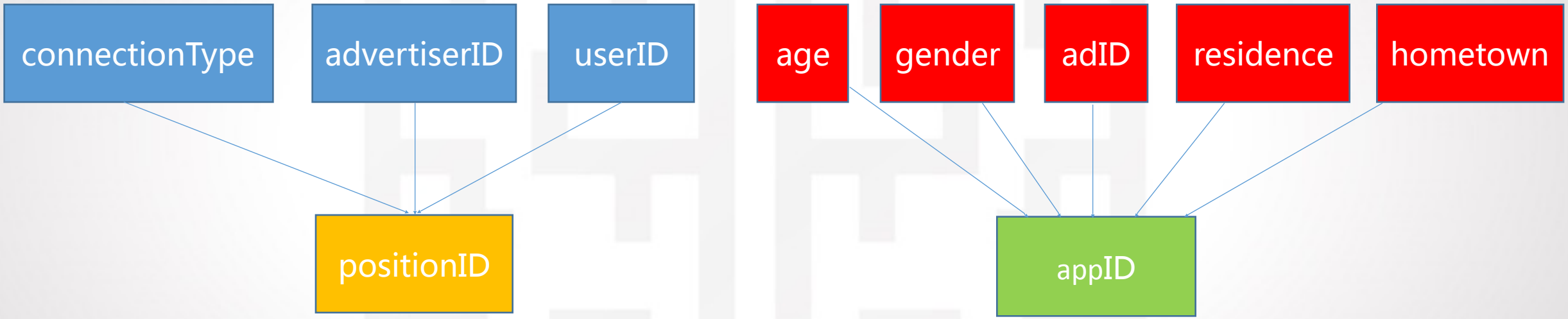
计数特征

单特征提取


用户点击当前累计点击次数
用户距离上一次点击的时间间隔
用户是否点击访问过该广告
用户对当前APP的累计点击次数
用户最近一次下载app总数及最近一次的时间间隔
用户计数特征
广告计数特征
各广告基本特征总点击量
各广告基本特征每分钟点击量
各广告基本特征中app的占比

计数特征

交叉特征的总点击量



不同广告特征在
广告位置的热度

不同用户特征在
appID的热度

转化率的贝叶斯平滑

由于数据稀疏性的原因，直接观测到的CVR与真实的CVR之间的误差较大。因此利用贝叶斯平滑对CVR预估进行优化

·对于某广告,C表示回流次数，I表示点击次数

·用平滑转化率r作为特征

$$r_i = \frac{C_i + \alpha}{I_i + \alpha + \beta}$$

参考于文献:Click-Through Rate Estimation for Rare Events In Online Adversting

用户历史点击

使用01串表示用户历史点击，例如:

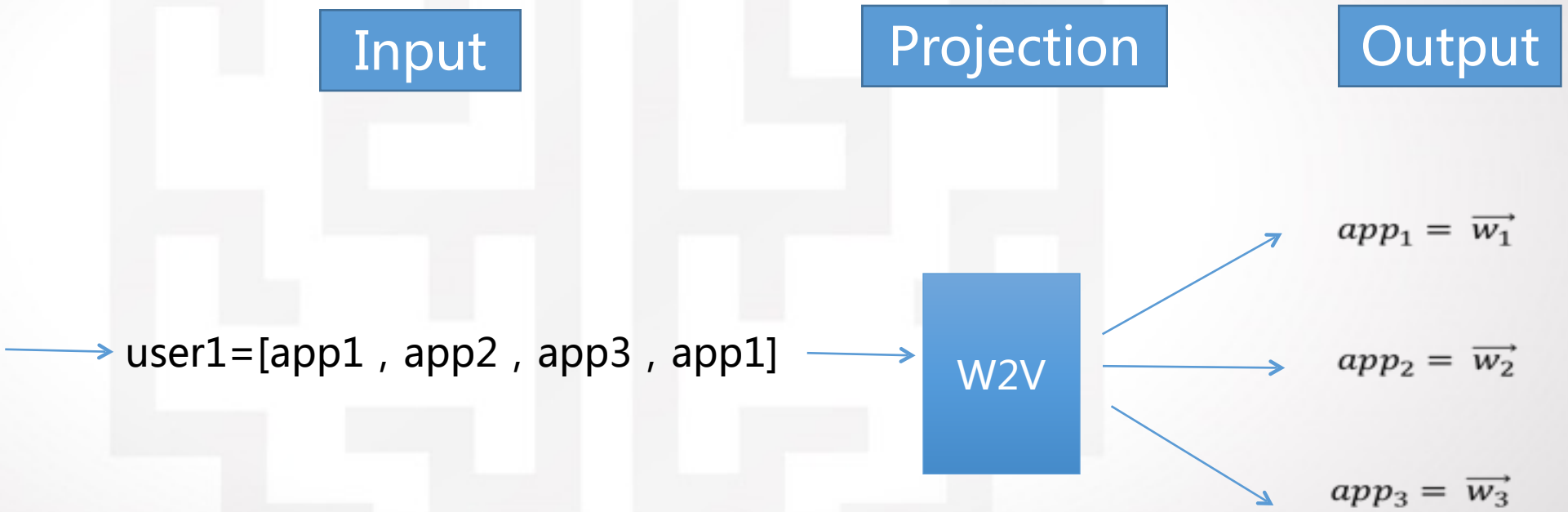
Label	user	history
0	*user1*	
1	*user1*	*0*
1	*user1*	*01*
0	*user1*	*011*
1	*user1*	*0110*

Word embedding

用户的点击记录作为文本,使用word2vec进行Word embedding，例如:

history

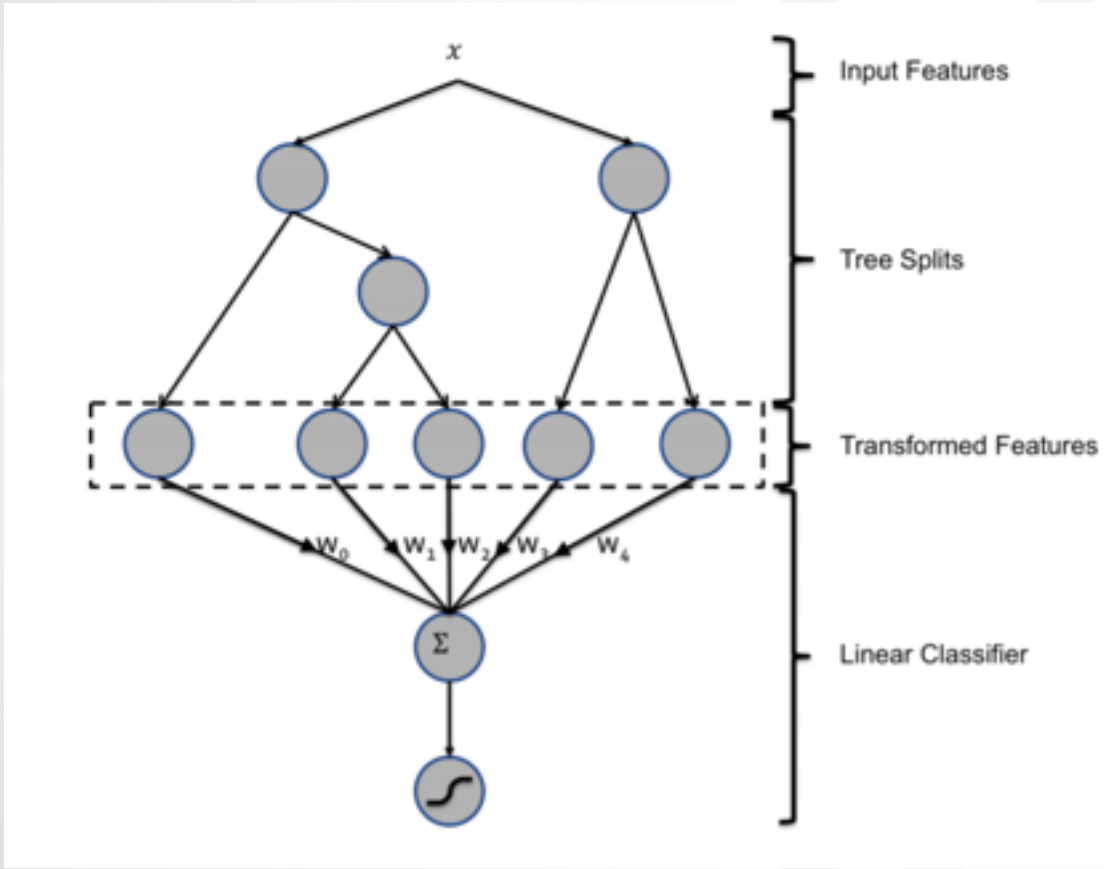
userID	appID
user1	*app1*
user1	*app2*
user1	*app3*
user1	*app1*



模型设计

3

GBDT离散特征



Our features

XgbClassifier

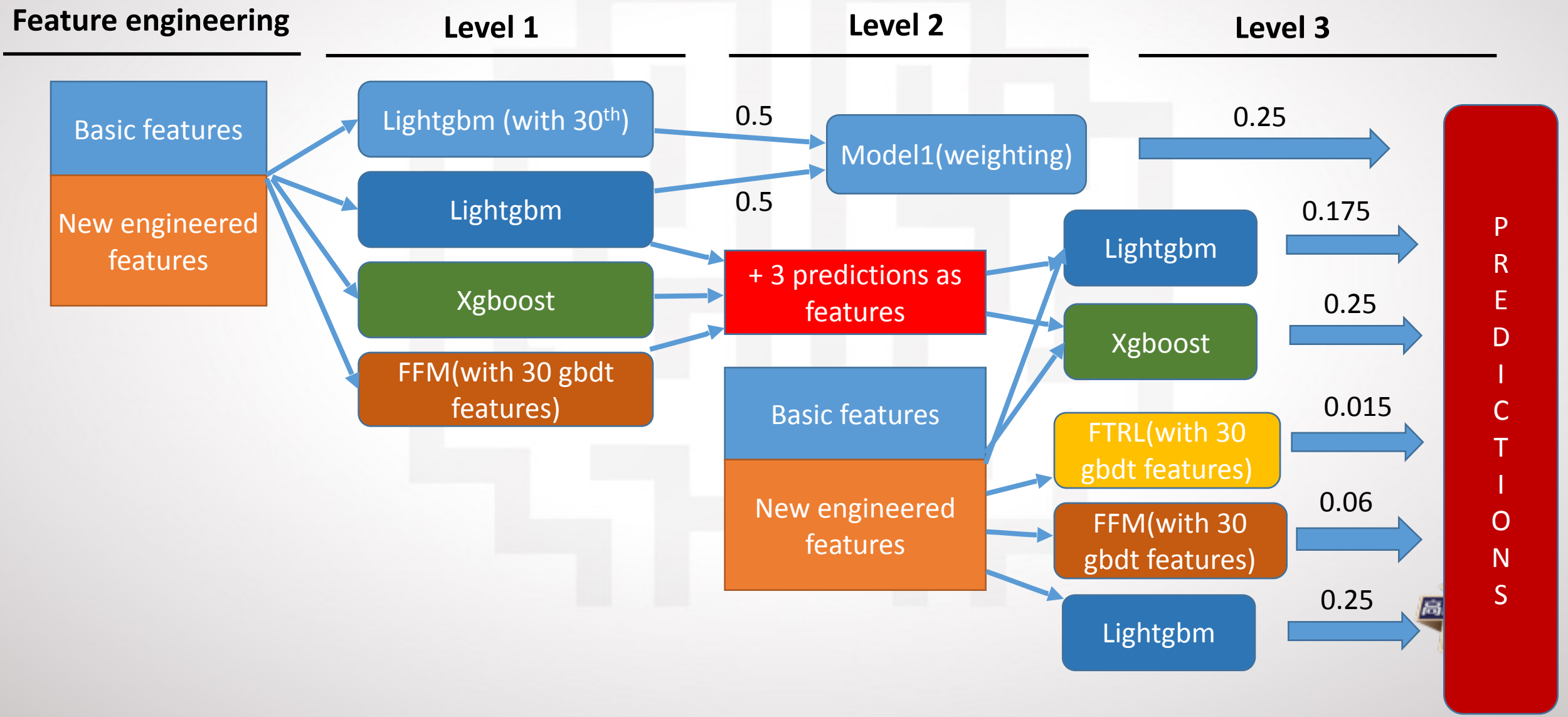
GBDT sparse features

Number of trees: 30
Learning rate: 0.1
Maximum tree depth: 8

模型融合

模型	线下成绩	线上成绩
*Xgboost*单模型（*28, 29*天数据）	*0.0981*	*0.1023*
*Lightgbm*单模型（*28*，*29*天数据）	*0.0977*	*0.1019*
*Lightgbm*单模型（*28*，*29*，*30*天数据）	*0.0919*	*0.1023*
*FTRL+gbdt*特征（*28*，*29*天数据）	*0.0986*	*0.1028*
*FFM+gbdt*特征（*28*，*29*天数据）	*0.0985*	*0.1027*
Lightgbm（*28*，*29*天数据，*Lightgbm*，*Xgboost*，*FTRL*，*FFM stacking*）		*0.1020*
Xgboost（*28*，*29*天数据，*Lightgbm*，*Xgboost*，*FTRL*，*FFM stacking*）		*0.1018*

模型融合


Feature engineering
Level 1
Level 2
Level 3
Basic features
New engineered features
Lightgbm (with 30th)
Lightgbm
Xgboost
FFM(with 30 gbdt features)
0.5
0.5
Model1(weighting)
+ 3 predictions as features
Basic features
New engineered features
0.25
Lightgbm
Xgboost
FTRL(with 30 gbdt features)
FFM(with 30 gbdt features)
Lightgbm
0.175
0.25
0.015
0.06
0.25
PREDICTIONS

均值调整

根据训练样本（28，29天数据，转化率为0.0268）推测31天的转化率大致为0.0270左右，使用如下公式对预测结果的均值进行调整：

$$final\_prediction = f(f^{-1}(x) + b)$$

其中：

$$f(x) = \frac{1}{1 + e^{-x}}$$

b值通过二分，逐步使得变换后的均值与目标均值误差小于1e-5

总结回顾

总结回顾

Q1，决赛数据量过大，有什么优化方法？
A1：特征构造可以使用流式统计方法，避免数据集读入内存中。同时，对数据进行一定的清洗，挑选能够处理而又有效的数据集。

Q2：Stacking方法在线下线上都得不到提高，如何调整？
A2：除了Stacking方法，还可以直接对不同模型进行线性组合。而Stacking方法带来的最重要的启发，其实更在于模型的多样性，模型多样化了，直接线性组合也能够提高非常大。

腾讯社交广告
Tencent Social Ads