

高校算法大赛

2017-07-06

freeze()

for

<=Date()

1+

9

if

time

var

Your

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

1

赛题分析

2

特征工程

3

模型设计

4

总结回顾



赛题分析

1



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

2



特征工程

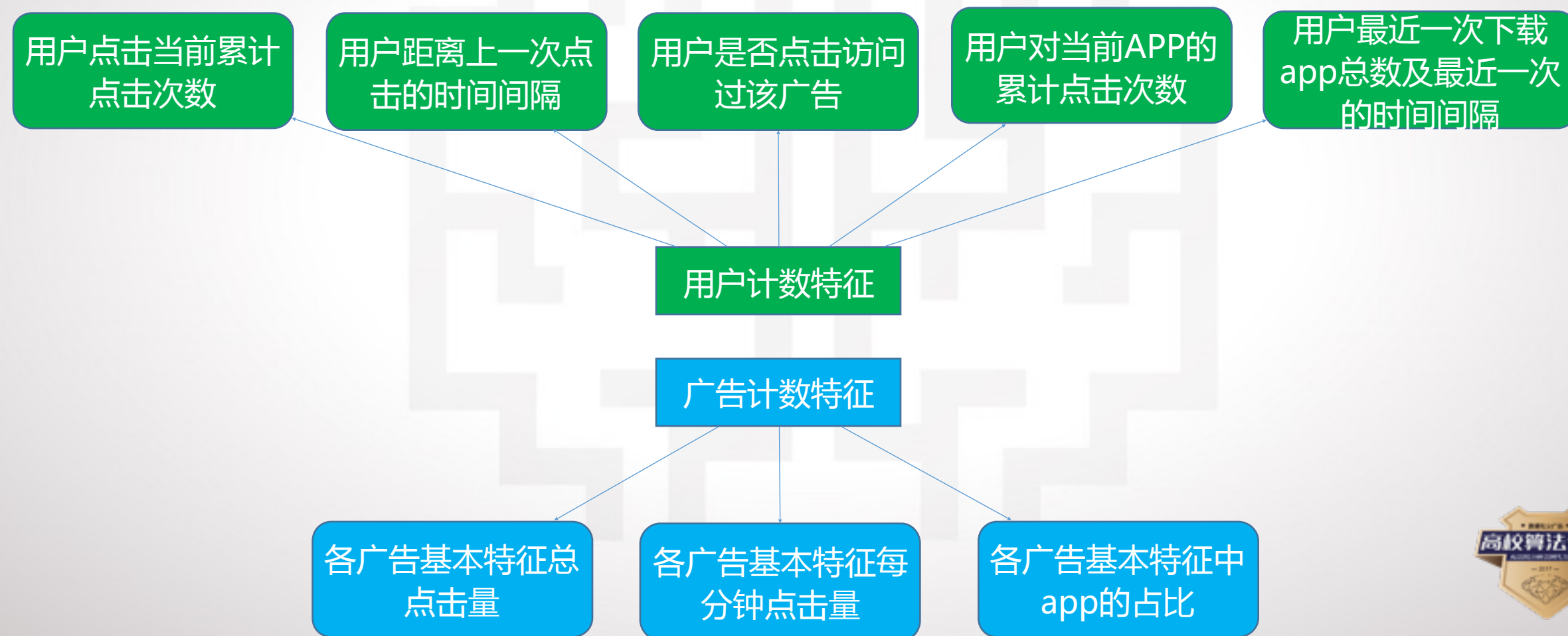
除了基本特征，我们还生成了以下特征：

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



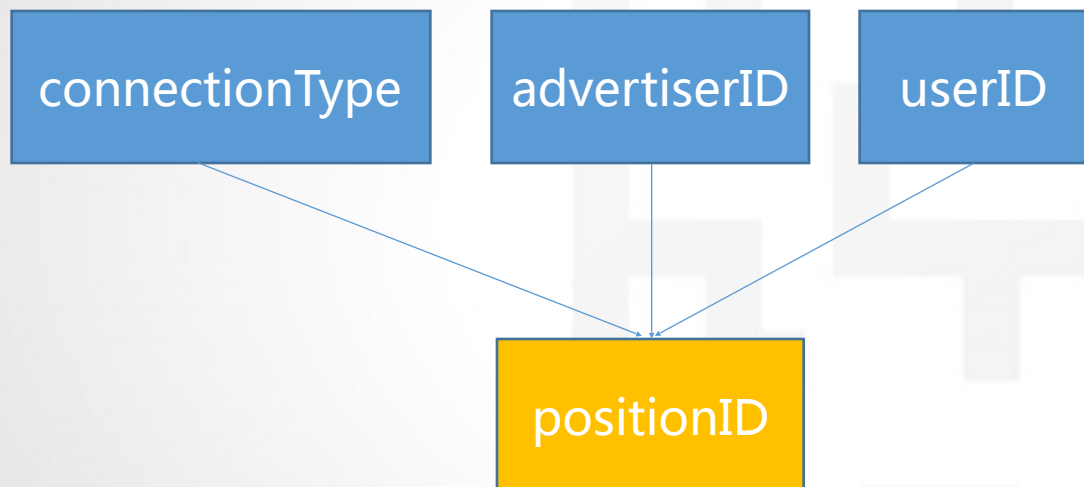
计数特征

单特征提取

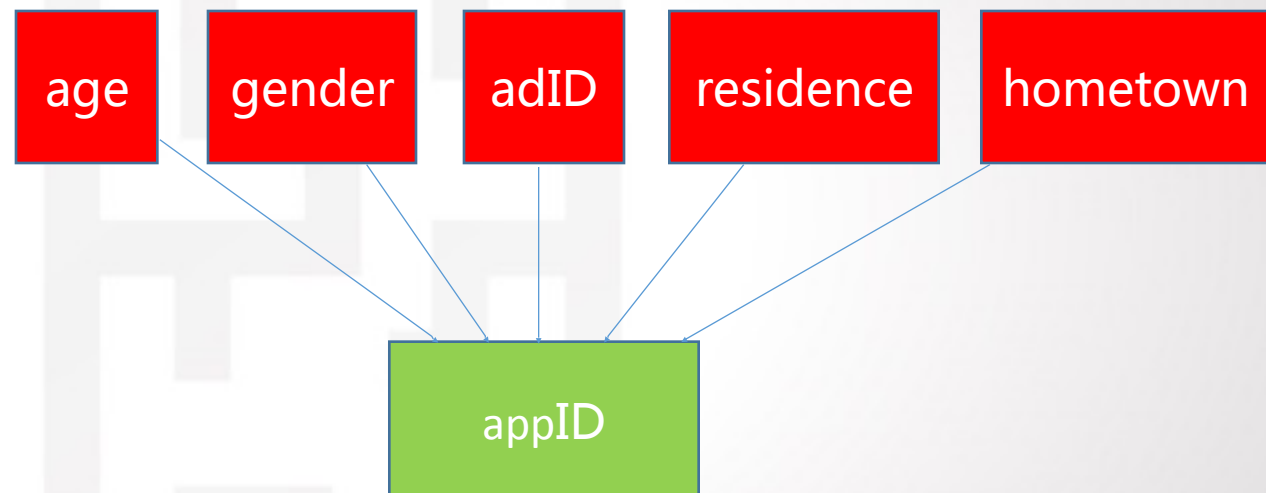


计数特征

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

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



用户历史点击

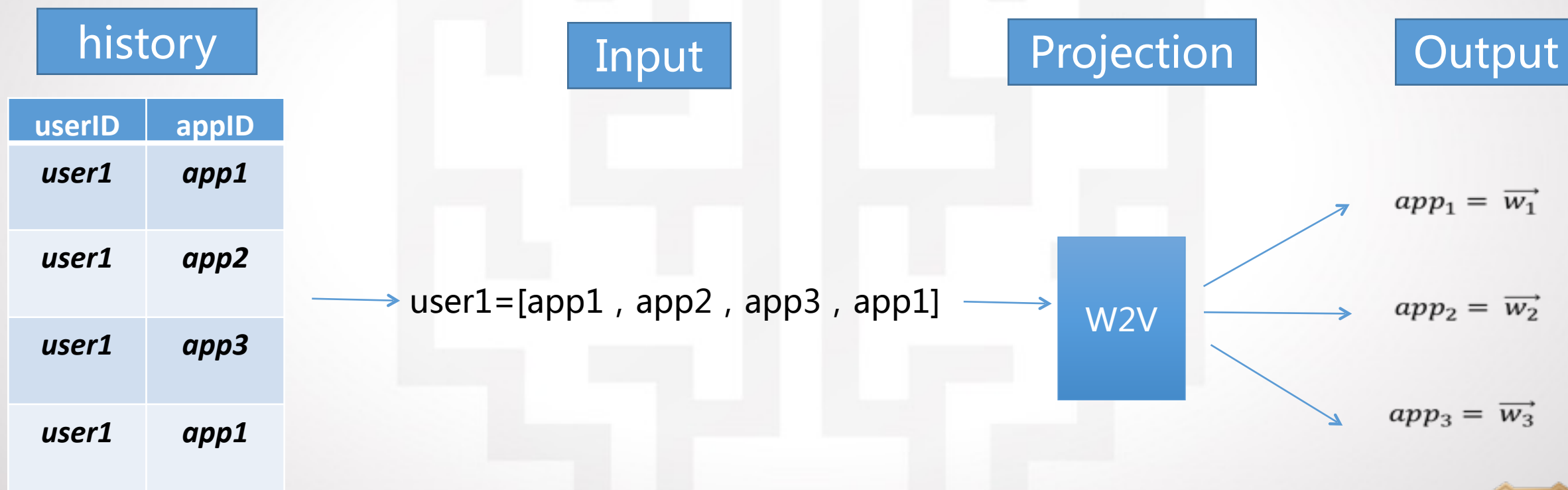
使用01串表示用户历史点击，例如：

Label	user	history
<i>0</i>	<i>user1</i>	
<i>1</i>	<i>user1</i>	<i>0</i>
<i>1</i>	<i>user1</i>	<i>01</i>
<i>0</i>	<i>user1</i>	<i>011</i>
<i>1</i>	<i>user1</i>	<i>0110</i>



Word embedding

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

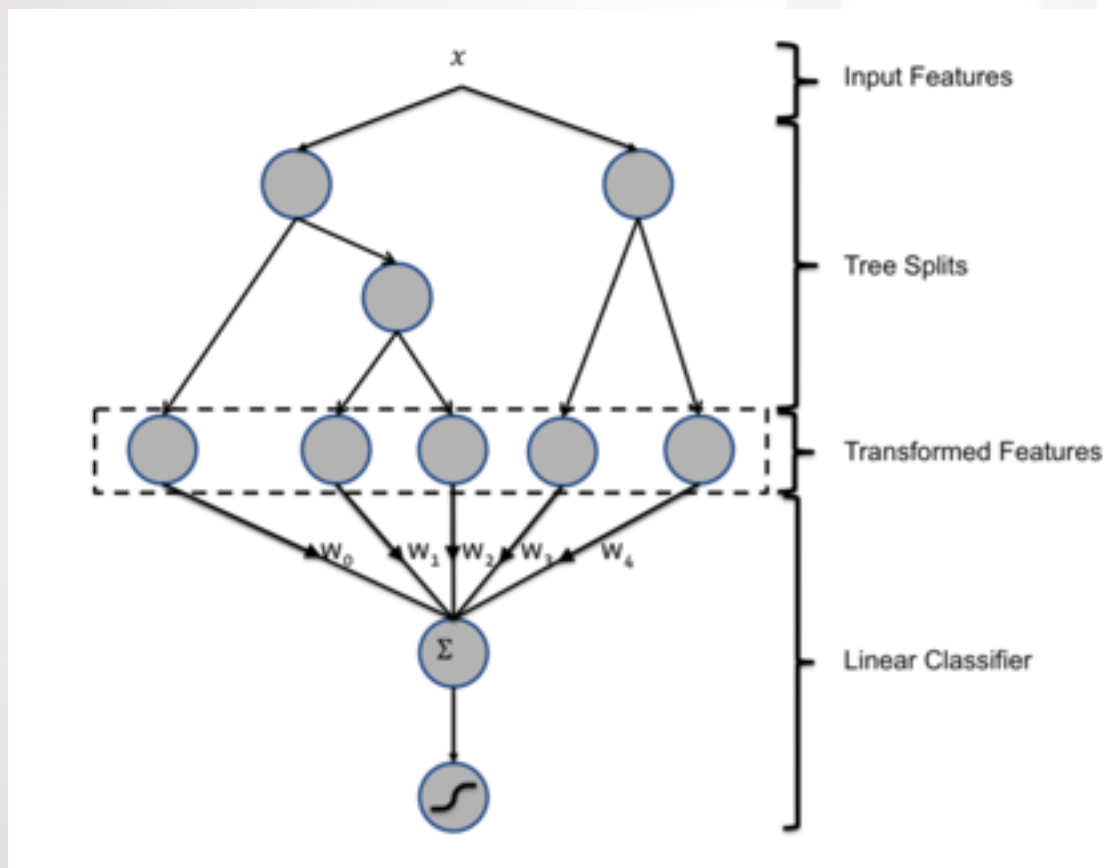


模型设计

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
<i>Xgboost</i> 单模型（28, 29天数据）	<i>0.0981</i>	<i>0.1023</i>
<i>Lightgbm</i> 单模型（28, 29天数据）	<i>0.0977</i>	<i>0.1019</i>
<i>Lightgbm</i> 单模型（28, 29, 30天数据）	<i>0.0919</i>	<i>0.1023</i>
<i>FTRL+gbdt</i> 特征（28, 29天数据）	<i>0.0986</i>	<i>0.1028</i>
<i>FFM+gbdt</i> 特征（28, 29天数据）	<i>0.0985</i>	<i>0.1027</i>
<i>Lightgbm</i> （28, 29天数据, <i>Lightgbm</i> , <i>Xgboost</i> , <i>FTRL</i> , <i>FFM stacking</i> ）		<i>0.1020</i>
<i>Xgboost</i> （28, 29天数据, <i>Lightgbm</i> , <i>Xgboost</i> , <i>FTRL</i> , <i>FFM stacking</i> ）		<i>0.1018</i>



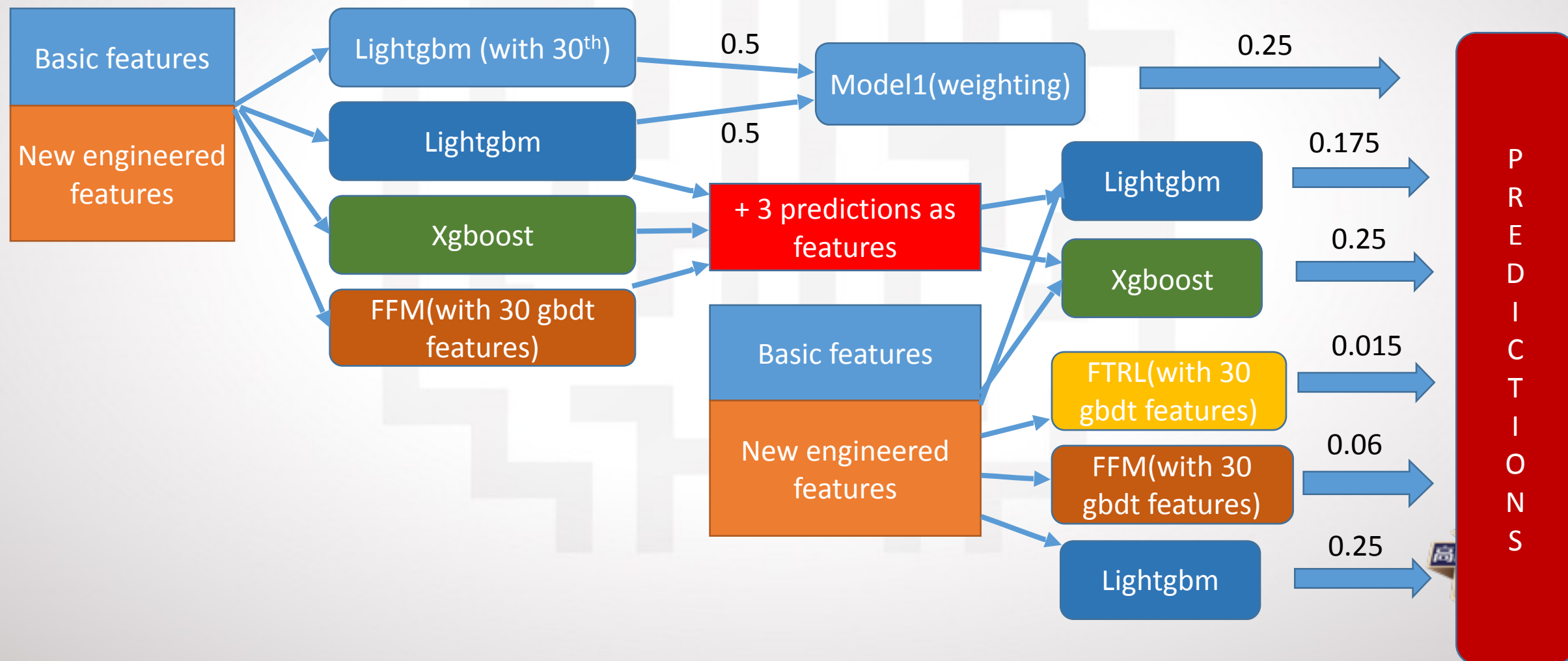
模型融合

Feature engineering

Level 1

Level 2

Level 3



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

4



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



THANKS

`<=Date()`

`freeze()`

`for`

2017-07-06

`Your`

`var`

`time`

`if`