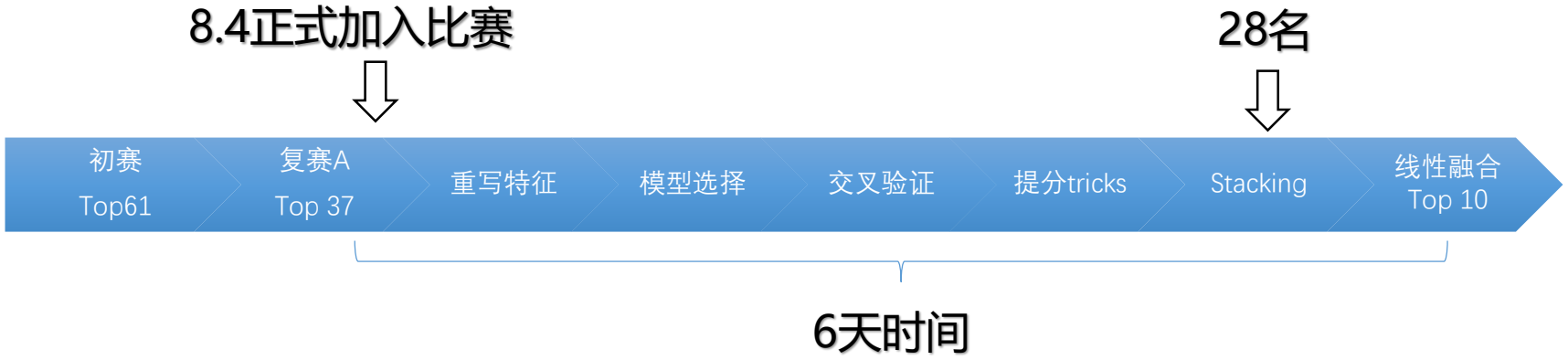

清華大學
Tsinghua University
快手
2018年
中国高校计算机大赛
——大数据挑战赛
团队:今我来思

1. 团队介绍
2. 问题描述
3. 框架设计
4. 数据分析
5. 特征工程
6. 算法及模型融合
7. 总结

01 团队介绍

上分历程-破釜沉舟


8.4正式加入比赛
28名
初赛
Top61
复赛A
Top 37
重写特征
模型选择
交叉验证
提分tricks
Stacking
线性融合
Top 10
6天时间

02

问题描述

02

问题描述

用户

- 设备类型
- 注册类型
- 注册天
- Launch
- Create video
- Activity

video

- 作者
- 被点击信息
- 被点击时间

预测用户未来七天是否有任何一种行为

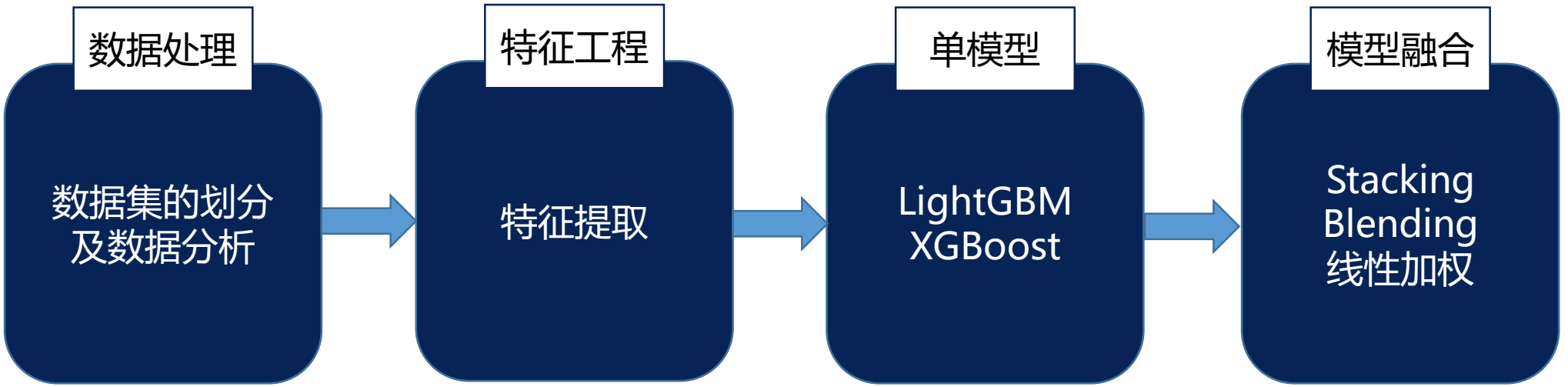
二分类问题

评价指标: AUC

03 解题思路

03

解题思路



数据去噪，提高数据质量

数据决定上限
模型决定如何逼近上限

降偏差、降方差
提高模型稳定性


04
数据分析

04

数据分析

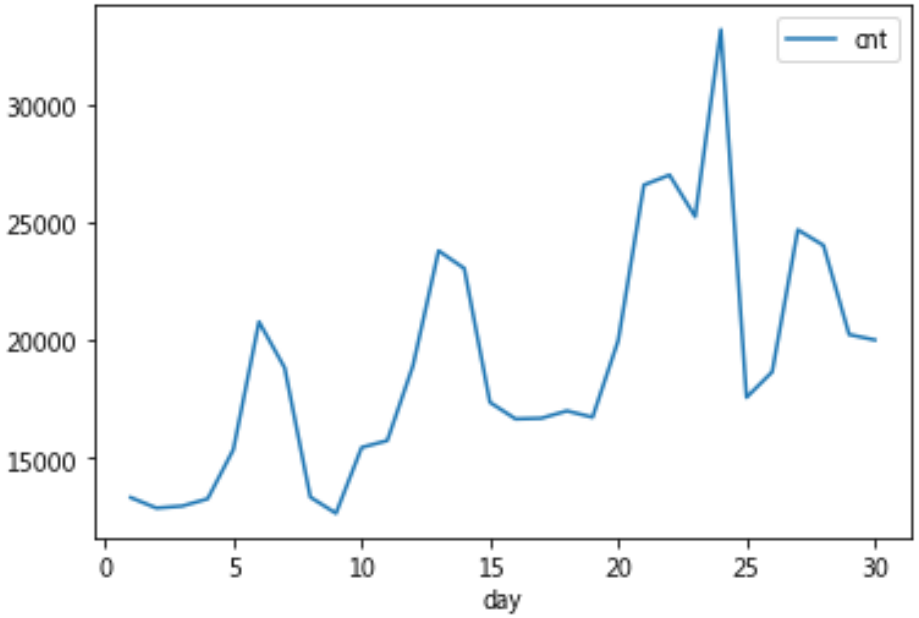
4.1 数据清洗

- **有activity,必须有登录，用activity表补全launch**

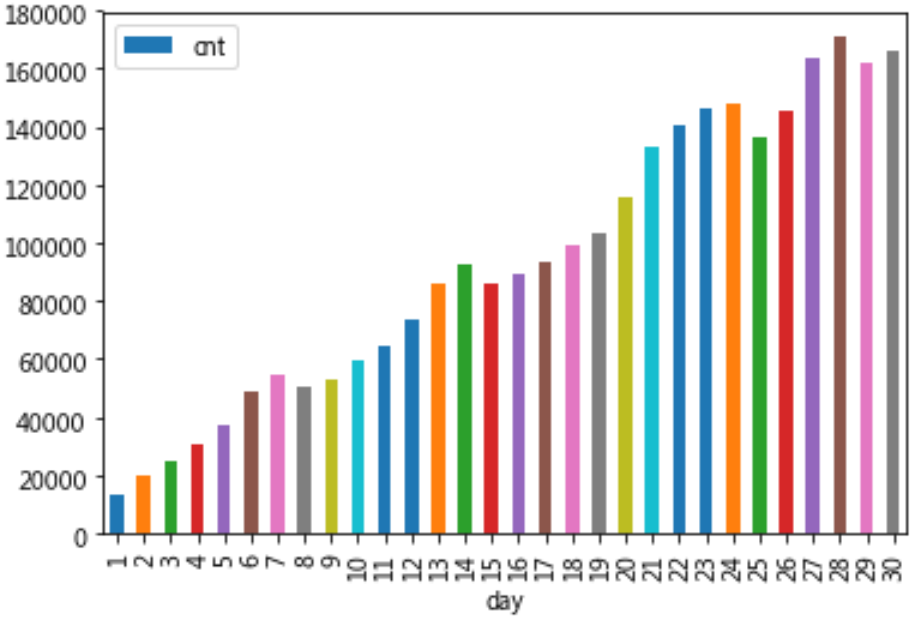
  修复了0.022%的数据

- **Activity 数据去重**
- **异常值处理**

数据分析



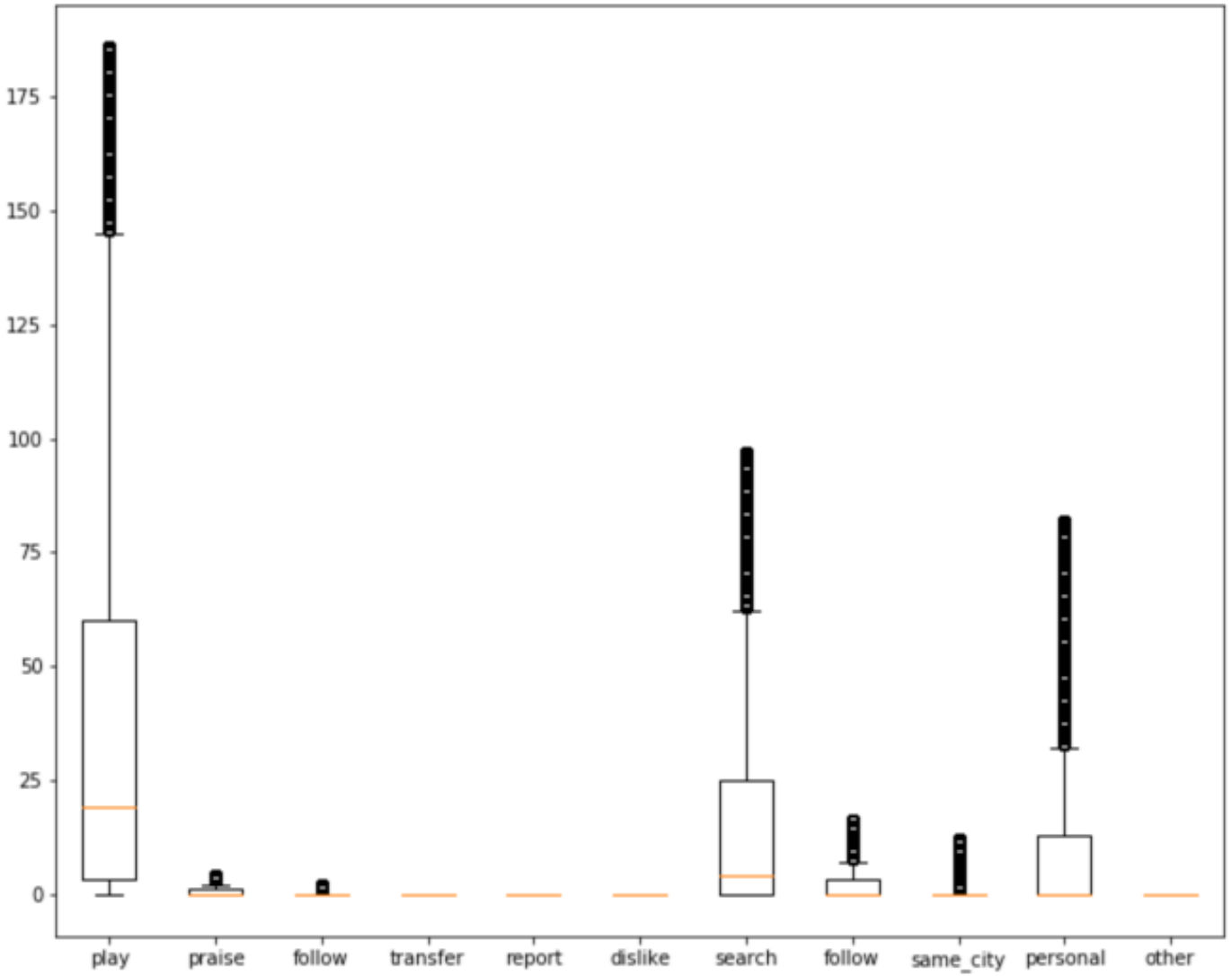
注册次数周期性



用户活动量统计图

04 数据分析

activity	Cover_rate
play	0.8618
praise	0.3500
follow	0.2384
transfer	0.0668
report	0.0003
dislike	0.0019
search	0.6502
follow	0.4715
Same_city	0.3118
personal	0.5368
other	0.0838



Activity (include 0)

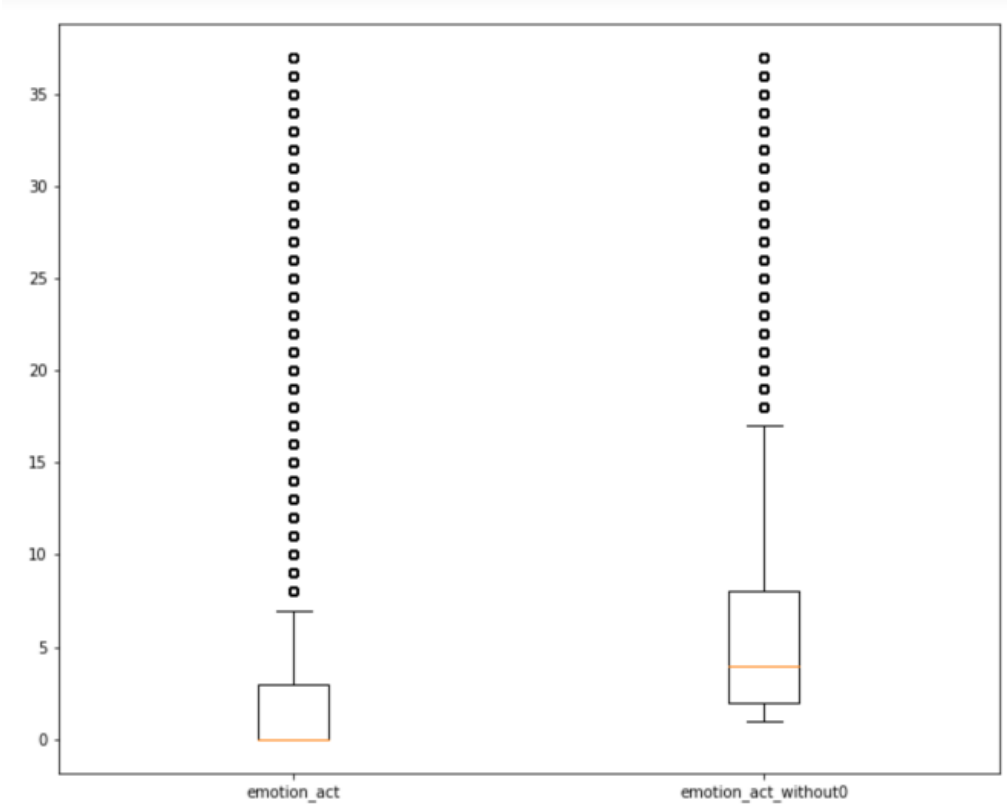
04

数据分析

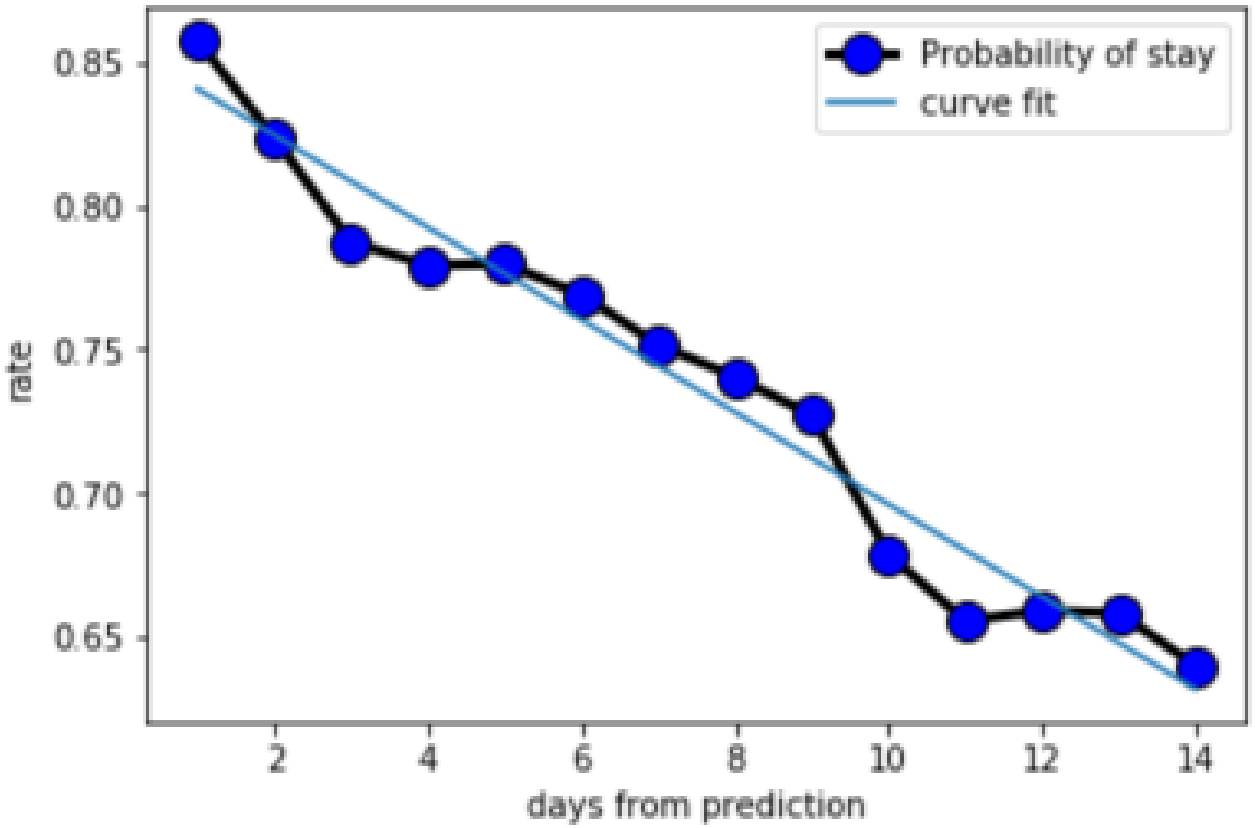
Dislike、Report行为很少，如何处理？

我们认为，除了播放行为，其他行为都是带有感情色彩的，因此可以直接相加。
从感情色彩的强度分析:
举报 > 转发 > 喜欢 == 不喜欢 > 播放
Emotion_activity = 点赞+ 不喜欢 + 2 * 转发 + 3 * 举报

activity	Cover_rate
Emotion_activity	0.4436



数据分析



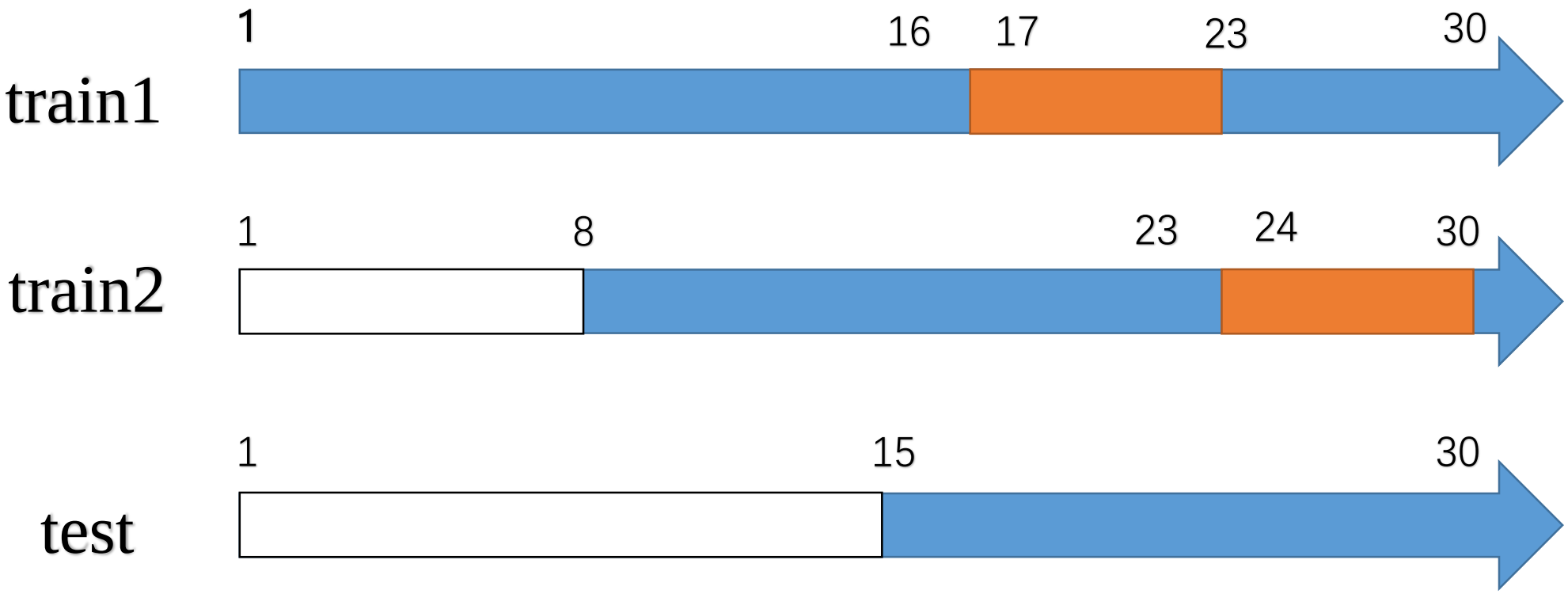
$$W(day) = -0.01606279 * \mathrm{DFP} + 0.85668832$$

$$Score = \sum_{i=1}^{t} W(i) * info(i)$$


05
特征工程

特征工程

5.1 数据集划分



特征工程

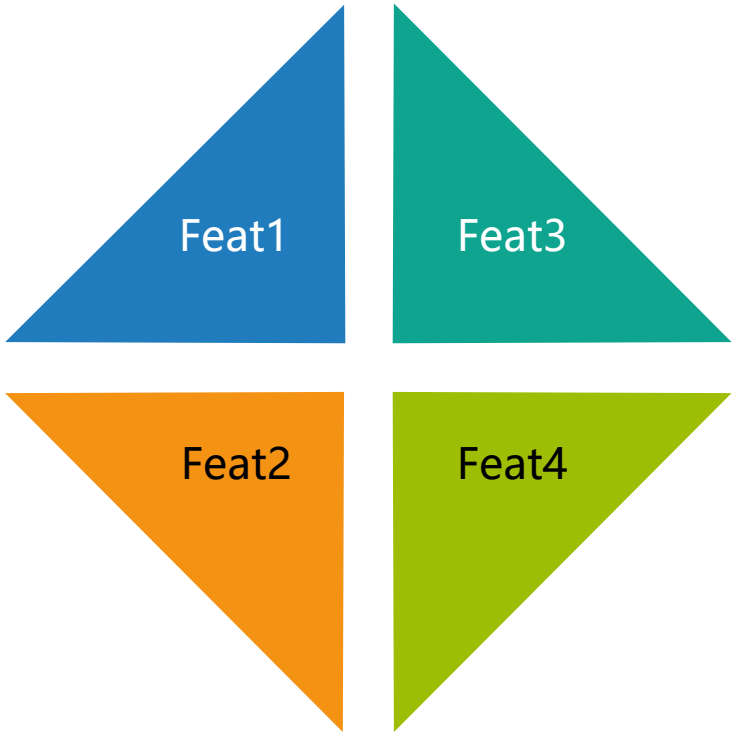
5.2 特征挖掘

全局数据特征群

全局角度
从全局不同角度描述用户习惯。包含用户信息特征、相关统计特征等。

行为序列特征群

用户行为角度
由用户出发，用细粒度的时间窗口描述行为的变化情况



时间衰减特征群

业务角度
考虑用户在不同时间上相同行为的不同含义，给予不同的权重表示。

特殊行为特征群

用户业务角度
考虑用户在特征行为，如粉丝属性，看某个author的连续天数、次数等。

05

特征工程

5.3 特征组成

注册　流行度　粉丝效应　NMF特征

登录　转化率　活跃度　生命周期

全局数据特征群

时间衰减特征群

$$w_i = \left(\frac{day_i}{\max(day_i)} \right)^2$$

统计对象

- Video特征
- Page特征
- Type特征
- 组合特征

统计周期

- 1天
- 3天
- 5天
- 7天
- 9天

行为对象

- User
- Activity

特征群

User	用户播放数
	.....
Activity	用户page点击数
	.....

行为序列特征群

特征行为特征群

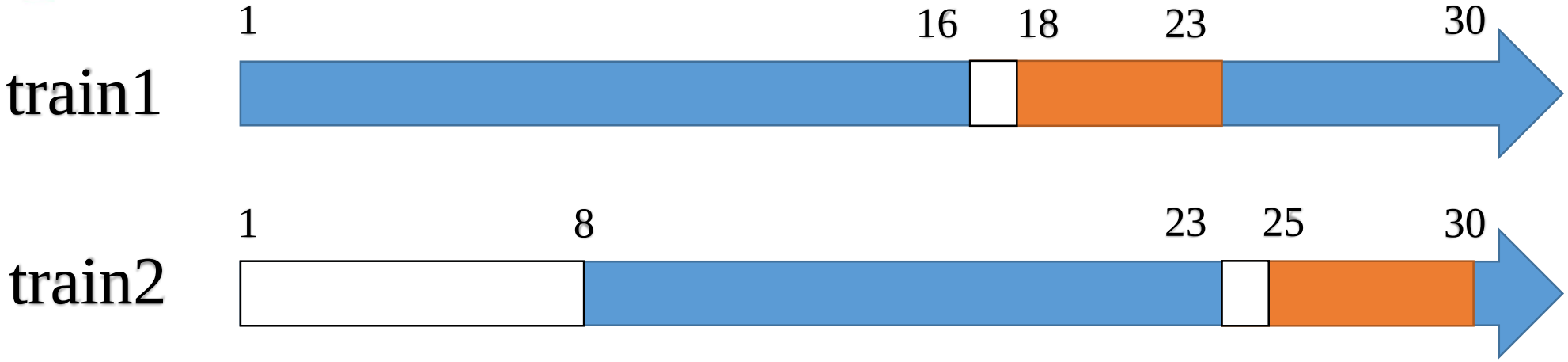
- 播放author连续N天的次数
- 最后一天pageN的次数
- 作者影响力
- 第一次有感情的行为
- 最后一次有感情的行为

标签处理

标签最后七天的权重真的一样吗？

1 0 0 0 0 0 0

0 0 0 0 0 0 1




06
算法及融合模型

stacking

Xgb/Lgb/GBDT/RF/LR

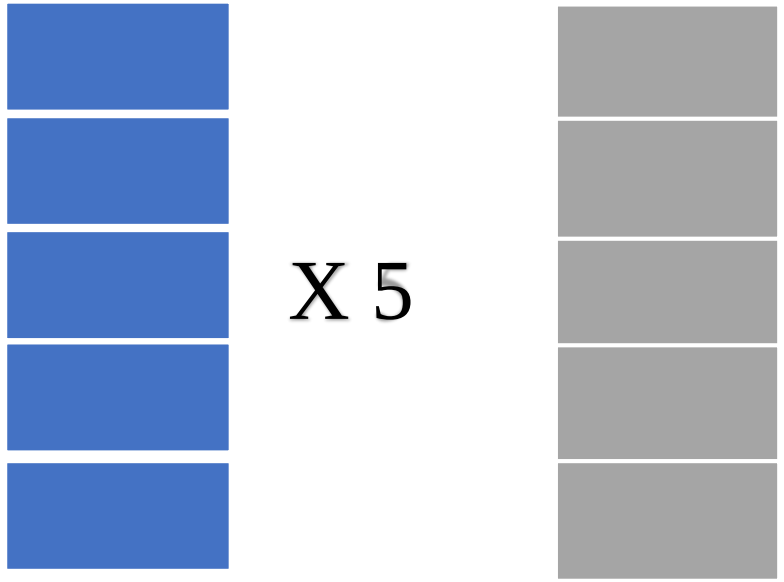
Train data
5 Flod

Build New Model

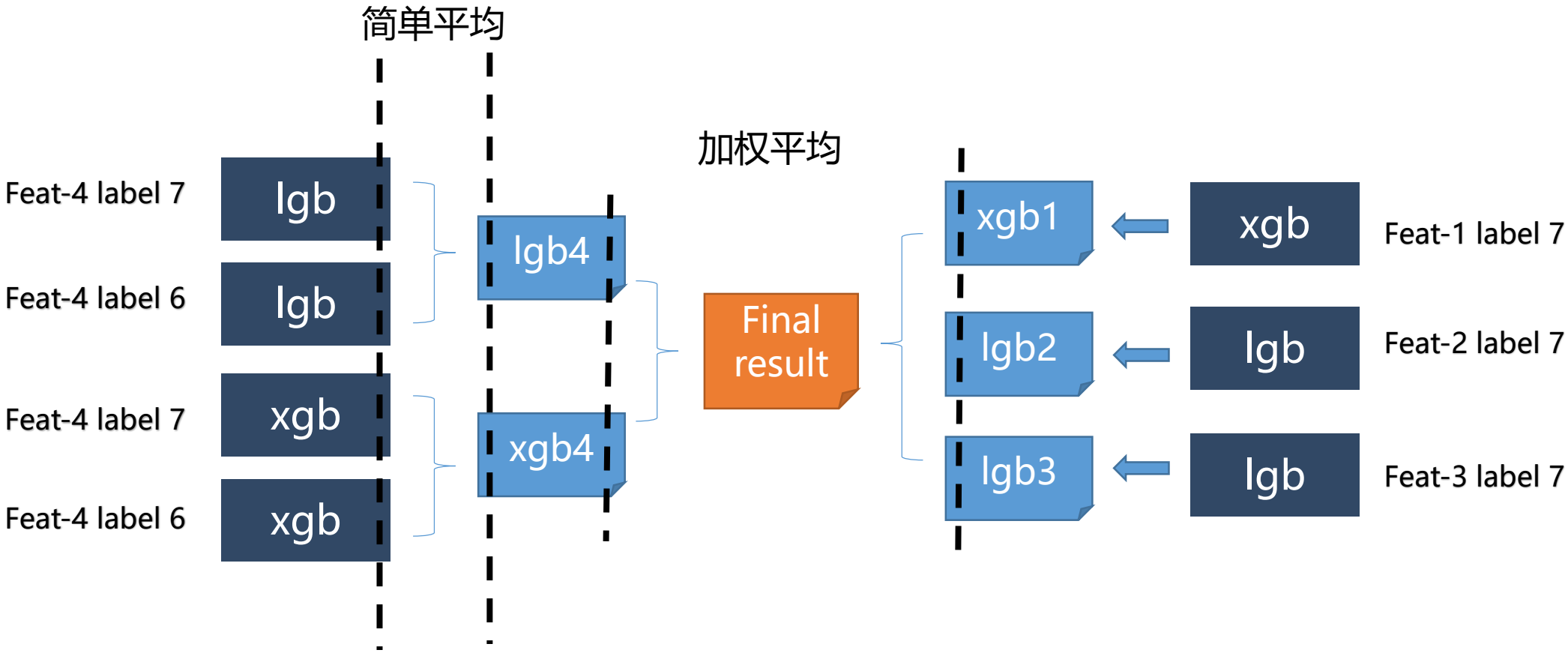
New train data

Label

X 5



06 算法及融合模型


简单平均
加权平均
Feat-4 label 7
lgb
Feat-4 label 6
lgb
lgb4
Feat-4 label 7
xgb
Feat-4 label 6
xgb
xgb4
Final result
xgb1
xgb
Feat-1 label 7
lgb2
lgb
Feat-2 label 7
lgb3
lgb
Feat-3 label 7

06 算法及融合模型

为什么线性融合效果好?

	xgb1	lgb2	xgb4
xgb1	1	0.9905	0.9933
lgb2	0.9905	1	0.9922
xgb4	0.9933	0.9922	1

<

xgb1 lgb1

0.9982

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$


07
总结

07

总结

- ✓ 对数据进行了充分的分析
- ✓ 模型多样性、特征多样性
- ✓ 融合方法适当

遗憾：

I. 融合方式还有很多没尝试。如rank融合，Blending等。
II. 没有充分挖掘用户行为 可以把user_id与authorid/videoid矩阵进行矩阵分解，获得用户兴趣爱好向量
III. 神经网络没有尝试 可以把用户每天都行为embedding，进而用LSTM/CNN进行训练。

07

总结

- 感谢所有参赛队伍；
- 感谢各周的周星星的分享；
- 感谢快手与清华一起举办这样成功的比赛，给了我们学习、锻炼和展示的机会。


清華大學
Tsinghua University
快手
谢谢聆听
THANKS!!