

高校算法大赛

闹猫

2017-07-06

freeze()

for

<=Date()

9

+

8

if

Your

var

time



目录

1/ 赛题描述

2/ workflow

3/ 工程实践——大数据

4/ 思考总结



1

赛题描述



App转化率预估

与pCTR不同点：

回流数据采集困难

回流时间长

数据稀疏

与kaggle经典比赛不同点：

本比赛数据丰富，质量高

以3 Idiots架构为主，丰富特征工程

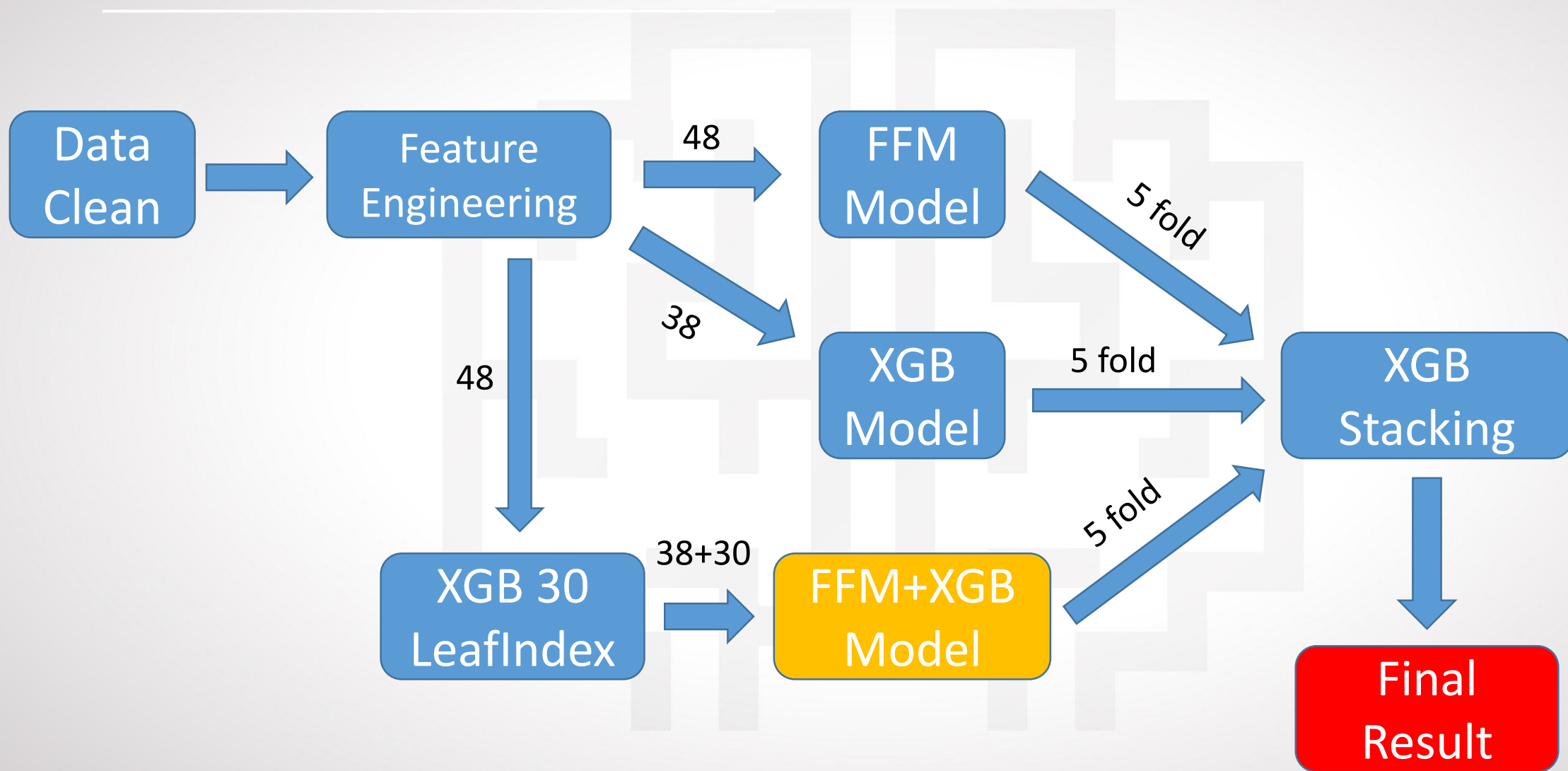


2

workflows






Workflow



理解业务流程

移动APP转化跟踪技术？

	方案简介	广告主选择
后台上报方案1	1. 广告点击后，广点通记录点击设备id，并同步给广告主； 2. App激活后，广告主接收激活设备ID； 3. 广告主对比有点击的设备号，计算激活量；	 (广告主自己掌握主动权)
后台上报方案2	1. 广告点击后，广点通记录点击设备id； 2. App激活后，广告主接收激活设备ID，并同步给广点通； 3. 广点通对比有点击的设备号，计算激活量；	 (接入难度小，较简单)
SDK方案	1. 广告主在创建广告前需要为应用集成广点通的mta-SDK； 2. 通过SDK来统计移动App的激活量；	 (广告主疑虑大)

数据清洗

后几天回流信息异常

删除 or 保留？

Tips：回流时间与广告主相关

例如：某广告主典型回流时间为120分钟，大于30215900的数据删除



特征工程

什么是强特征？

区分
度

覆盖
率

无泄
漏

如何发现强特征（Trick）？

假设->验证->解释->归纳

为何会有重复点击？
重复点击如何处理？

分钟 → 秒



其他特征

用户需求：

用户习惯
更换设备
应用商店
老年用户

分层贝叶斯：

平滑转化率
树形关系

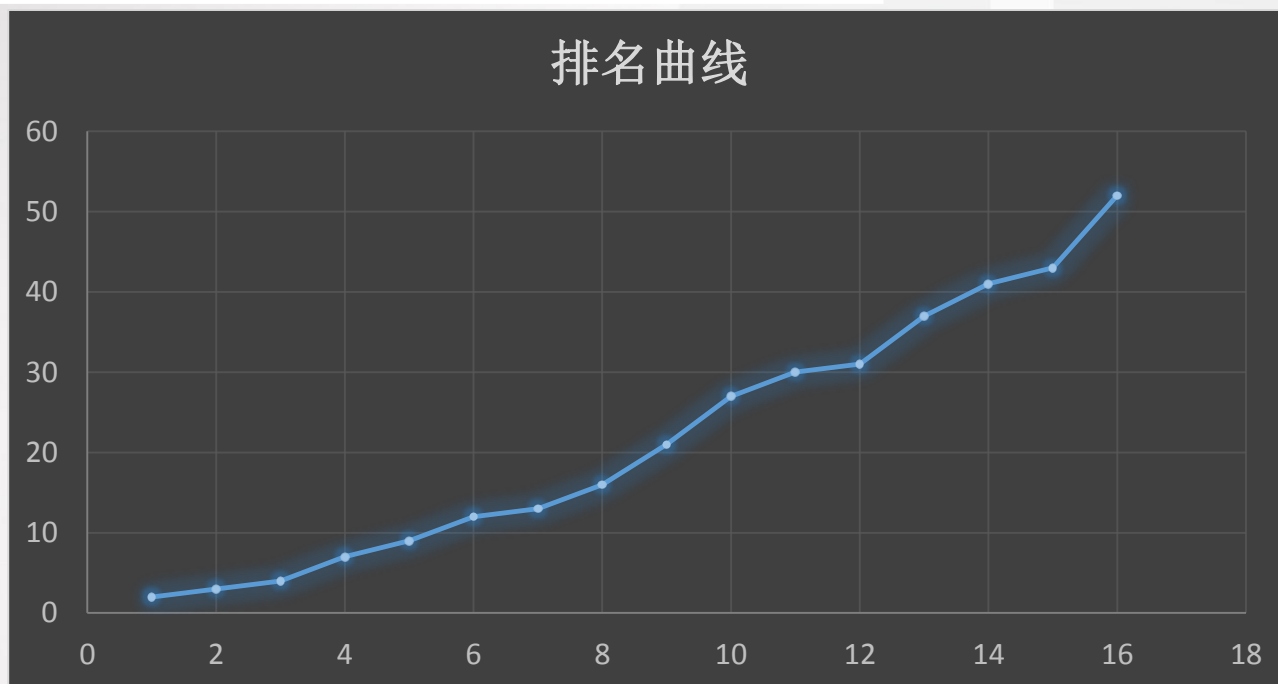
聚类：

LDA
TF-IDF

初赛21个特征 随机10% top2



决赛特征



毕业旅行期间
排名 2 → 52

噪音变大

初赛以appID随机
梳理App与广告主关系
作弊（异常）数据

冷启动

test中有大量新出现ID
404



3

工程实践 ——大数据



特征增量更新

存储已有特征，只提取新特征(15分钟)

Basic特征

clickTime
positionID
connectionType
advertiserID
campaignID
appID
appPlatform
appCategory
sitesetID
agegender

Count特征

userapp_conv
user_cnt
user_day_cnt
user_hour_cnt
user_app_day_cnt
user_app_hour_cnt
appCategoryConv
appConv
advertiserConv
creativeConv

Timeseq特征

history
successionType
sameTimeType
from_prev_click
to_next_click
.....

window特征

user_conv
cnt
user_appear_day
app_appear_day
ad_appear_day
creative_appear_day
position_appear_day
.....

action特征

user_day_uv
from_pre_action
action_cnt
.....

merge_feature



Encode并行计算

流式计算？

3 Idiots 为何使用Hash Trick？

age	gender
18	1
16	1
17	1
40	0
40	1
32	0

SPLIT

age	gender
18	1
16	1
17	1

age	gender
40	0
40	1
32	0

ENCODE

1:322:1	2:499:1
1:20:1	2:499:1
1:49:1	2:499:1
1:504:1	2:5:1
1:504:1	2:499:1
1:24:1	2:5:1

MERGE



运行效率——决赛全量

步骤	耗时
提取特征	10-30分钟
合并特征	3-5分钟
特征转码	15分钟
XGB训练	12分钟
FFM训练	120-160分钟
Stacking训练	25分钟



4

思考总结



思考总结

异常app与position

手机型号

深度学习

手机剩余空间、电量

iOS安装信息

社交



THANKS

freeze()

for

<=Date()

2017-07-06

Your

var

time

if