



腾讯社交广告
Tencent Social Ads



`<=Date()`

`freeze()`

`for`

高校算法大赛

9

2017-07-06

8

`Your`

`var`

`if`

`time`

- **队伍**

- 永不理解

- **队员**

- 符汉杰，孙冠东，万美含

- **学校**

- 复旦大学，硕士

- **研究方向**

- 数据挖掘





赛题分析

特征工程

模型训练

模型融合

比赛总结



01

数据分布线上线下分布不一致

- 某些app和用户的记录比较少
- 数据的时效性要求较高

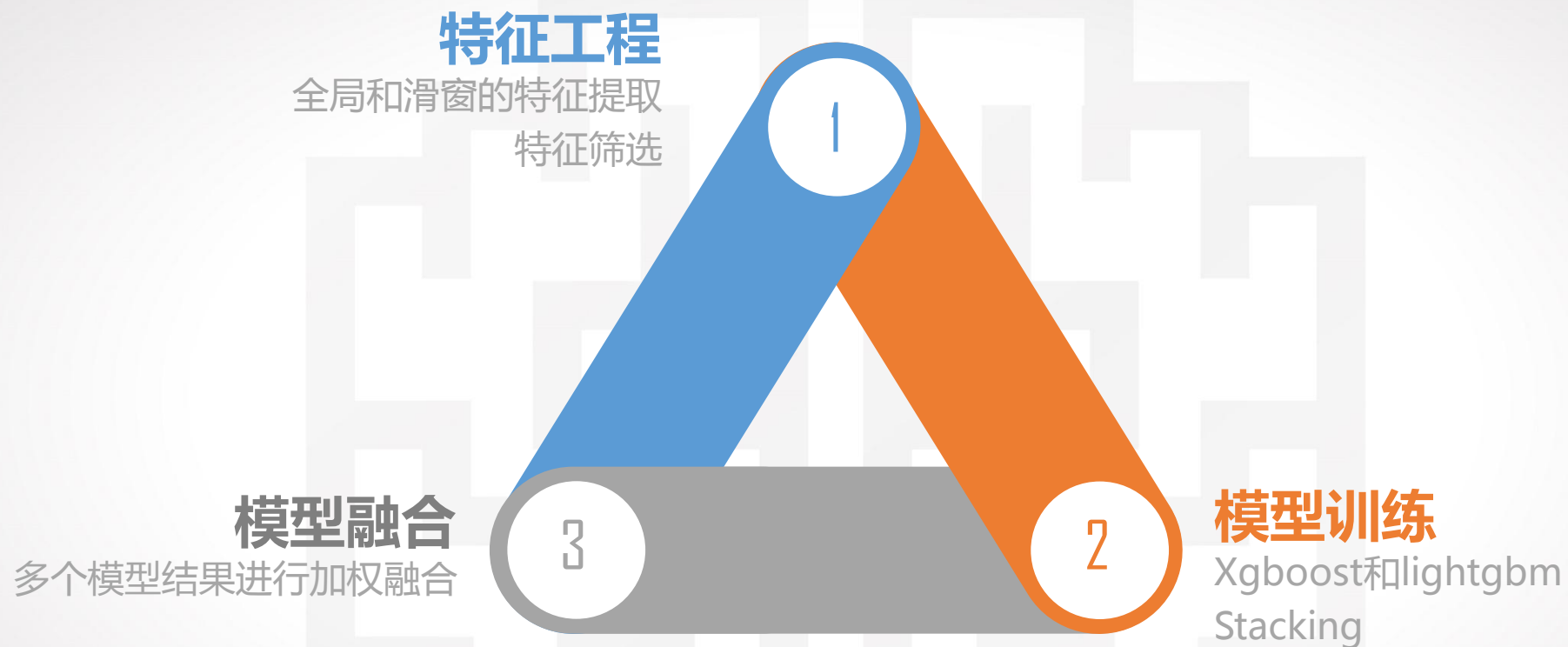
02


决赛数据量大

- 相比初赛，决赛数据量大，对代码特征机器的要求提高
- 部分工作需要需要重做来考虑在决赛的效果

1

方案流程





赛题分析

特征工程

模型训练

模型融合

比赛总结





基础特征

- 转化率
- 计数特征
- 比例特征
-



实时特征

- 当天用户计数
- 当天app计数
- 用户点击行为
-



用户行为挖掘特征

- Word2vec计算用户当前操作与历史行为的联系，如app与历史app的vec相似度

01

全集统计

基于全部数据统计生成特征

- 优点：效果稳定
 - 缺点：容易信息泄露，难以反映时序信息
- cv统计能很好的防止信息泄露

02

滑窗统计

基于邻近几天的数据统计生成特征

- 优点：能反映时序信息，不会有信息泄露
- 缺点：特征数量多，线上线下的特征分布差异大，特征工程的工作量大

2

特征选择

01

初赛


- 删除线上线下均值差异30%以上的特征
- 删除特征重要性较低的特征 (xgboost)
- Wrapper方法选择特征
但决赛的数据量大，这方面的工作会比较耗时。

02

决赛

- 加入一部分的特征，通过线上的成绩来选择特征去留





赛题分析

特征工程

模型训练

模型融合

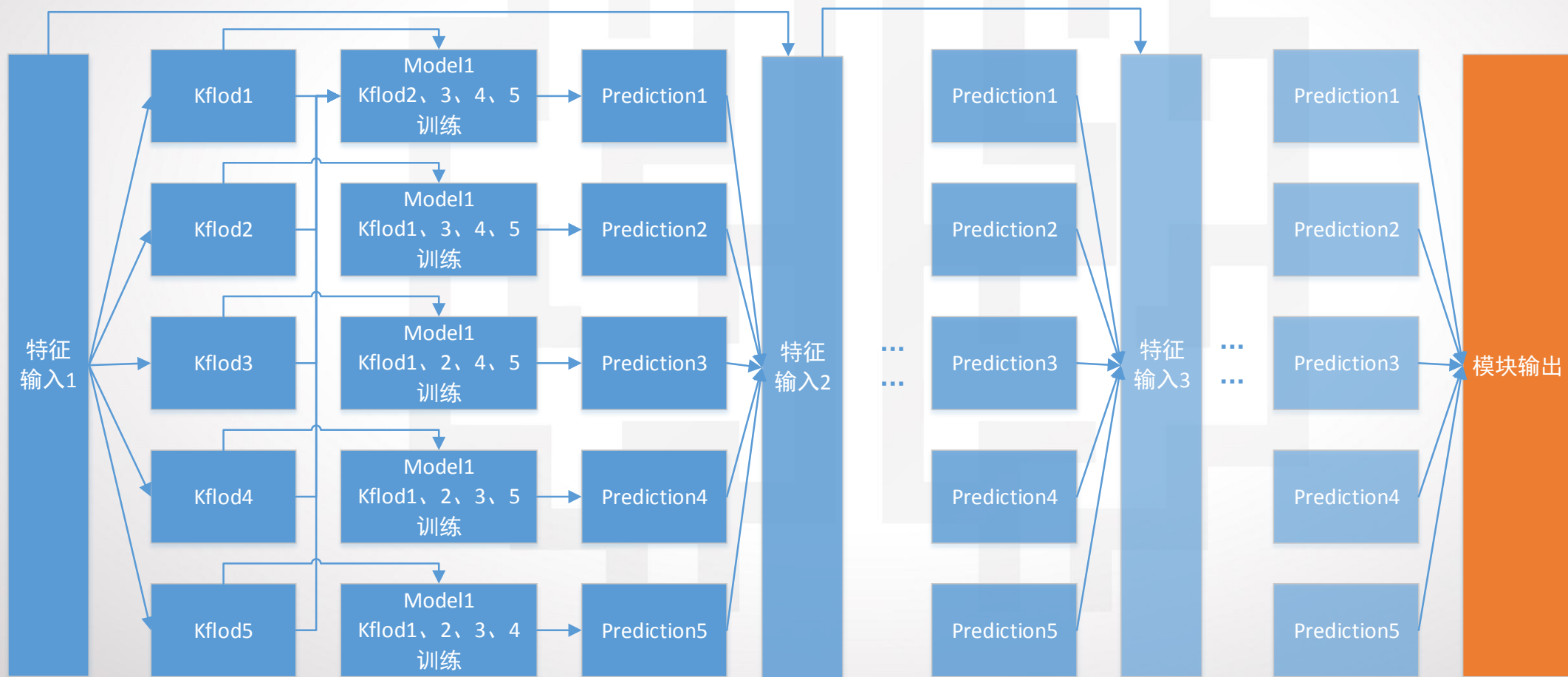
比赛总结



0

模型模块

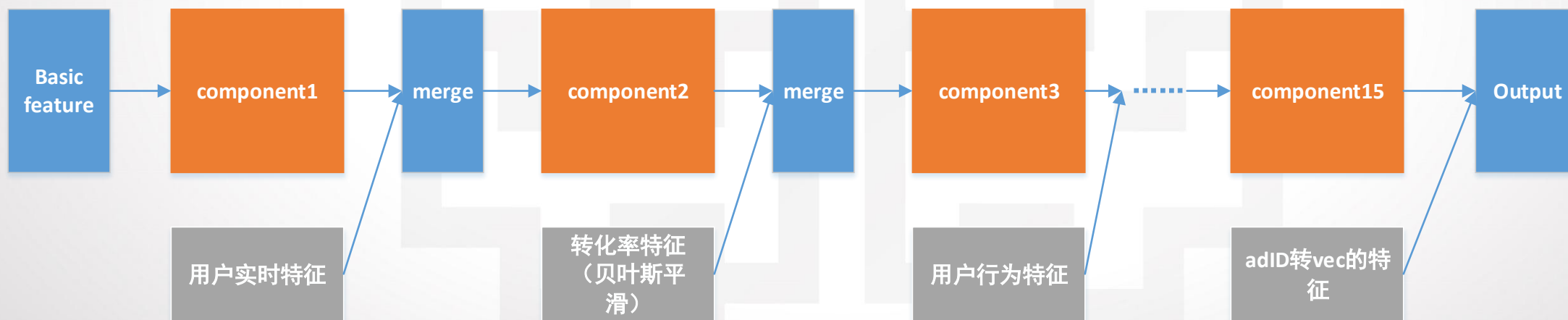
- 模型使用xgboost和lightgbm
- Cross validation , 增强模型鲁棒性
- Stacking , 保证模型的精度



1

模型流水线

- 每次训练依赖上一个component的预测以及新加入的特征
- 所有特征分而治之，保证模型的效率
- 通过线上反馈来决定component的去留
- 累计加入>200维特征

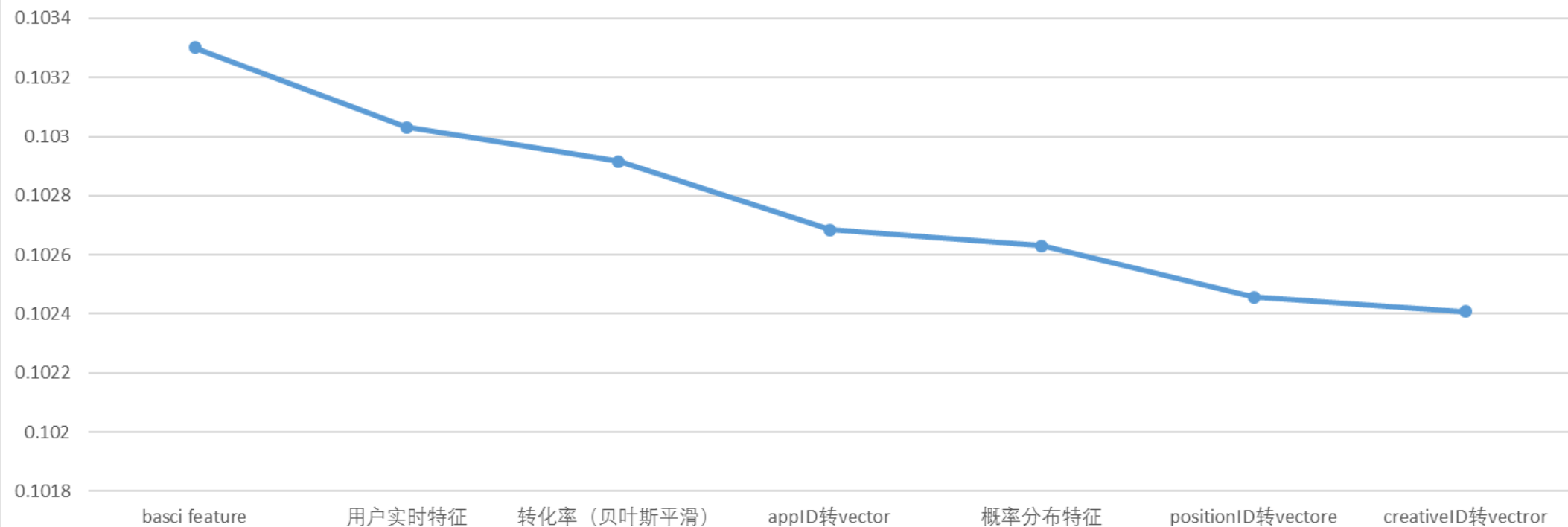



2

模型效果

- 实验中，随着加入的特征越多，模型效果变得更好

特征效果示意图





赛题分析

特征工程

模型训练

模型融合

比赛总结



01

加权融合

当模型效果差异较大时，对结果进行加权平均；
线上效果好的权重相对大些，而线上效果差的权重相对小些。

02

Logistic平均

当模型的效果差异比较小时，采用以下公式进行融合：
$$p = f\left(\frac{\sum f^{-1}(p_i)}{n}\right)$$


其中f函数是logistic函数。

From : 4 idiots - avazu

- 全集特征的单模型
- 滑窗特征的单模型
- 全集特征Stacking中component输出
- 选择不同数据集构造模型

加权融合
Logistic平均

输出



赛题分析

特征工程

模型训练

模型融合

比赛总结



- 比赛中，针对数据分布、数据量大等问题，尝试过各种特征工程、模型调参、模型融合等方法，受益良多。
- 比较心酸的是决赛数据量增多，机器配置受限的原因，需要代码重构。
- 在最后尝试使用ffm模型，但精力有限最终并没有弄出一个很好的模型，融合的效果有限。

- 感谢**腾讯**主办这次比赛，让我们能够接触到真实的业务数据，在比赛不断探索的过程中得到了锻炼和展示。
- 感谢比赛以来帮助过我们的朋友，以及给我们解决问题的相关工作人员。



THANKS

freeze()

for

<=Date()

2017-07-06

Your

var

time

if