



腾讯社交广告
Tencent Social Ads



高校算法大赛

nju_newbie 团队

2017-07-06

freeze()

for

<=Date()

9

+

8

11

~

3

11

Your

var

if

time

目录

contents

1

团队介绍

2

解题思路与算法

3

关键问题与解决方案



腾讯社交广告
Tencent Social Ads

团队介绍

`<=Date()`

`freeze()`

`for`



`+`

`9`

`if`

2017-07-06



`your`

`var`

`time`

`if`



`if`

团队介绍

南京大学计算机科学与技术系
计算机软件新技术国家重点实验室
机器人智能与神经计算研究组

指导老师：申富饶教授

队 长：杨 毅（研二）

队 员：梁 雨（研二）
沈少峰（研三）

实验室主页：
<http://cs.nju.edu.cn/rinc>





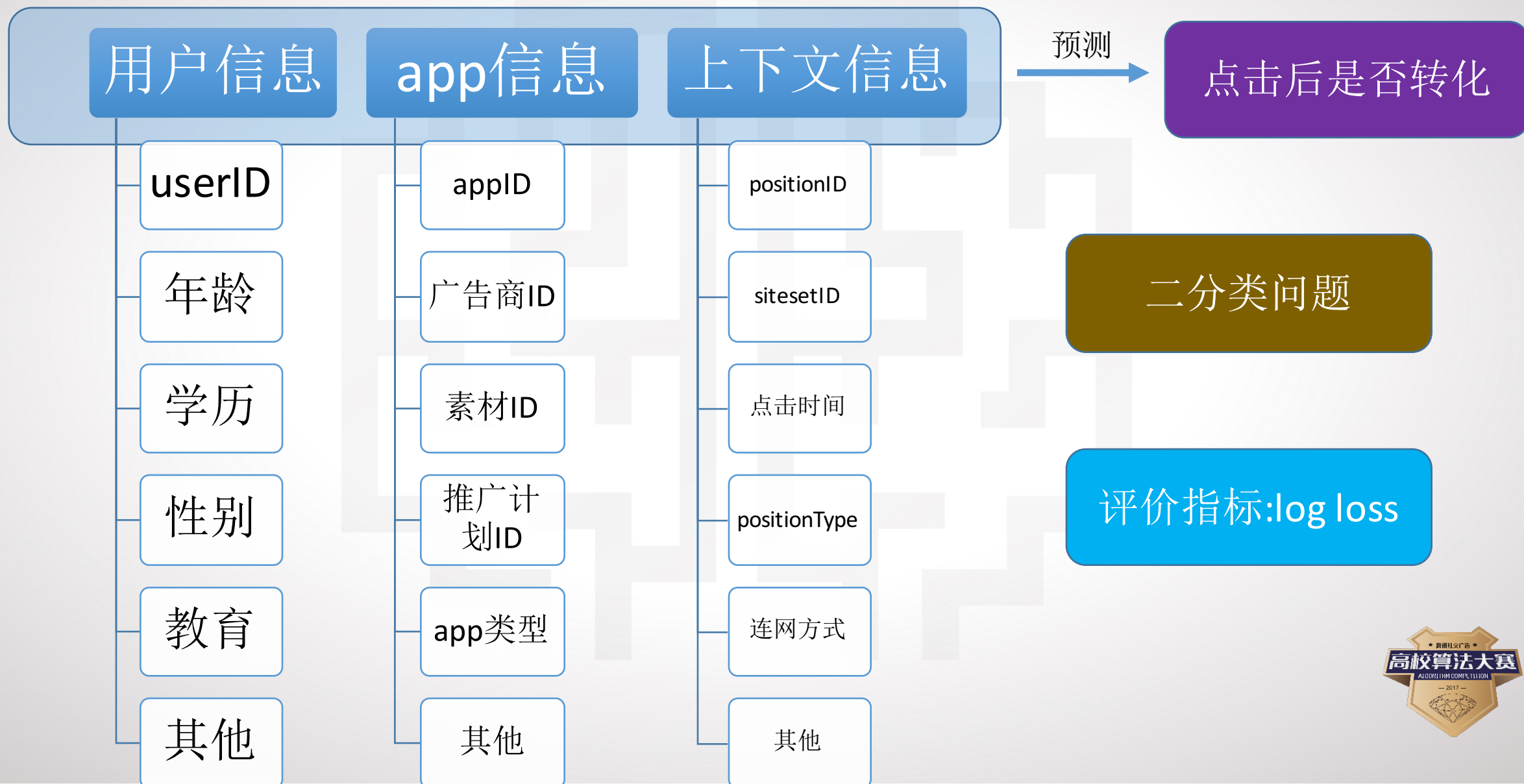
腾讯社交广告
Tencent Social Ads

解题思路与算法

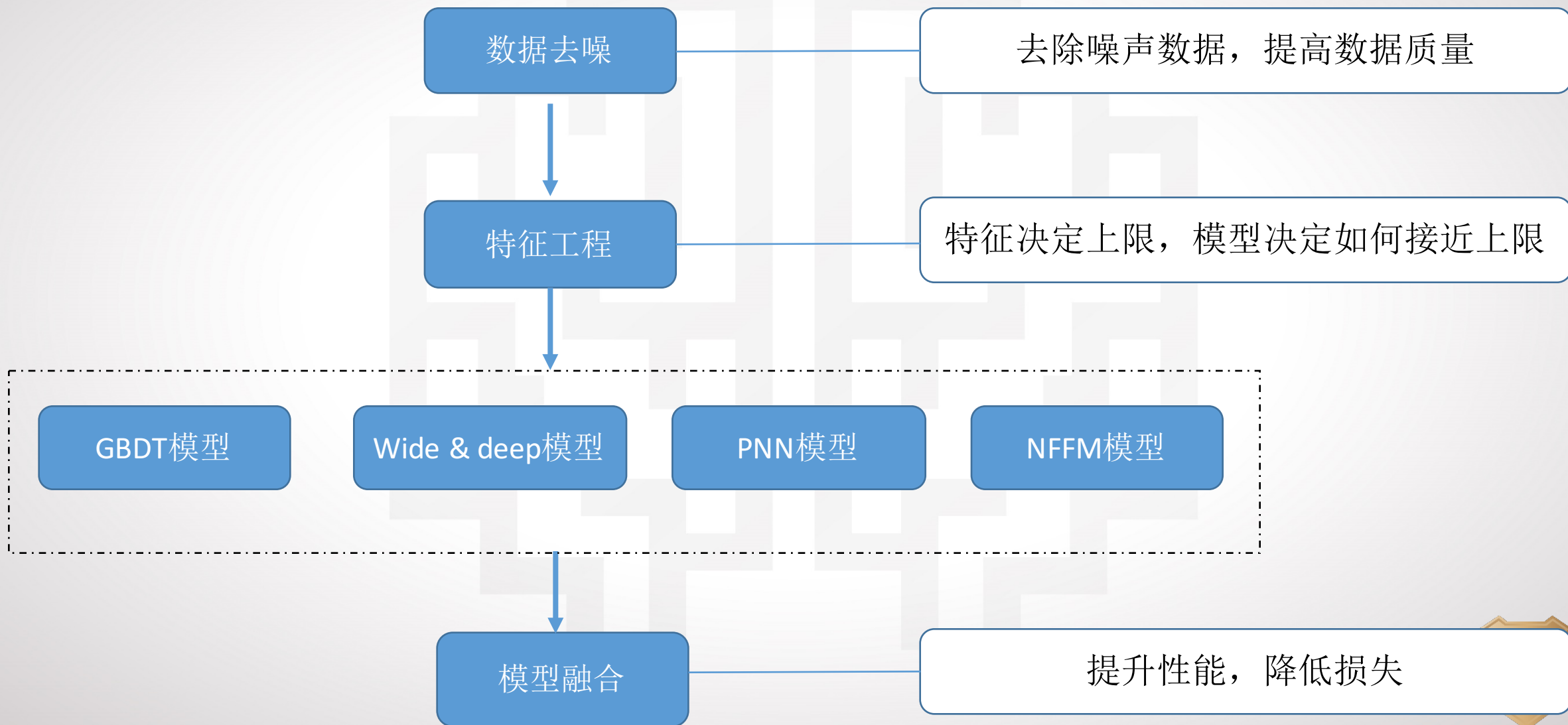
2017-07-06



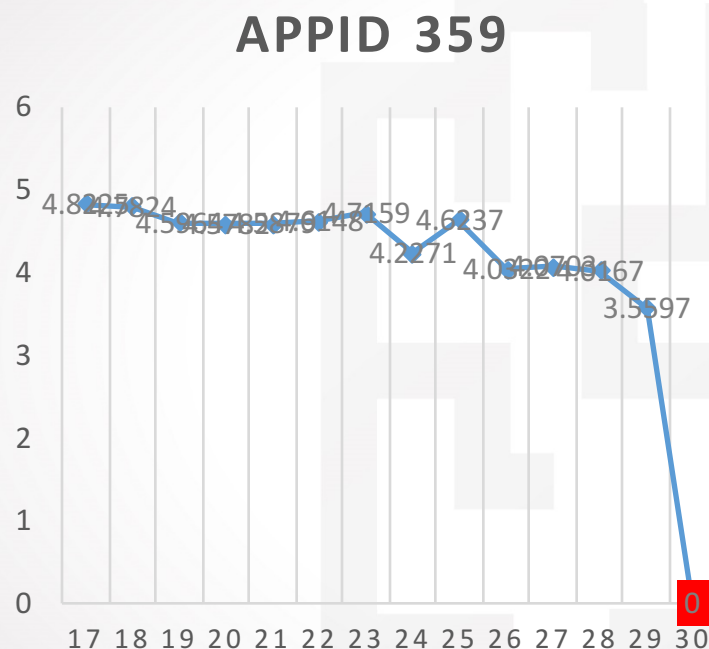
赛题简介



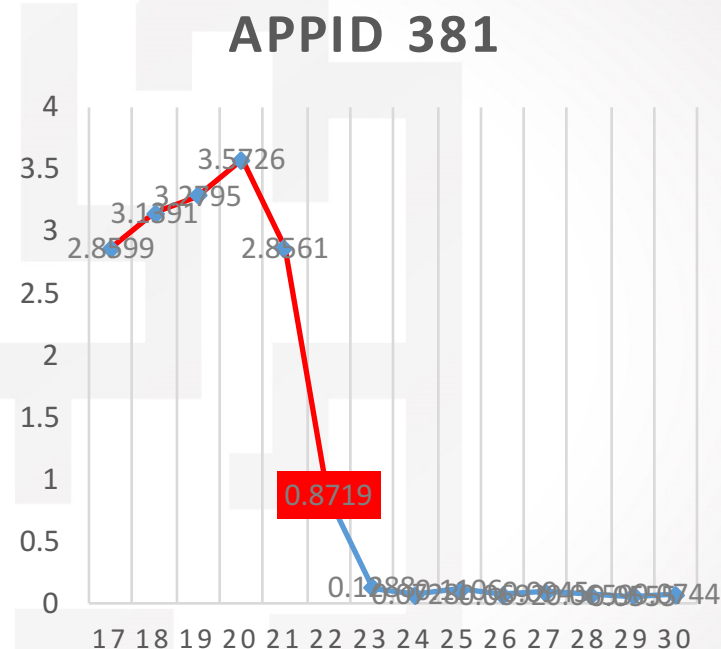
解题思路



数据去噪：去除噪声数据，提高数据质量



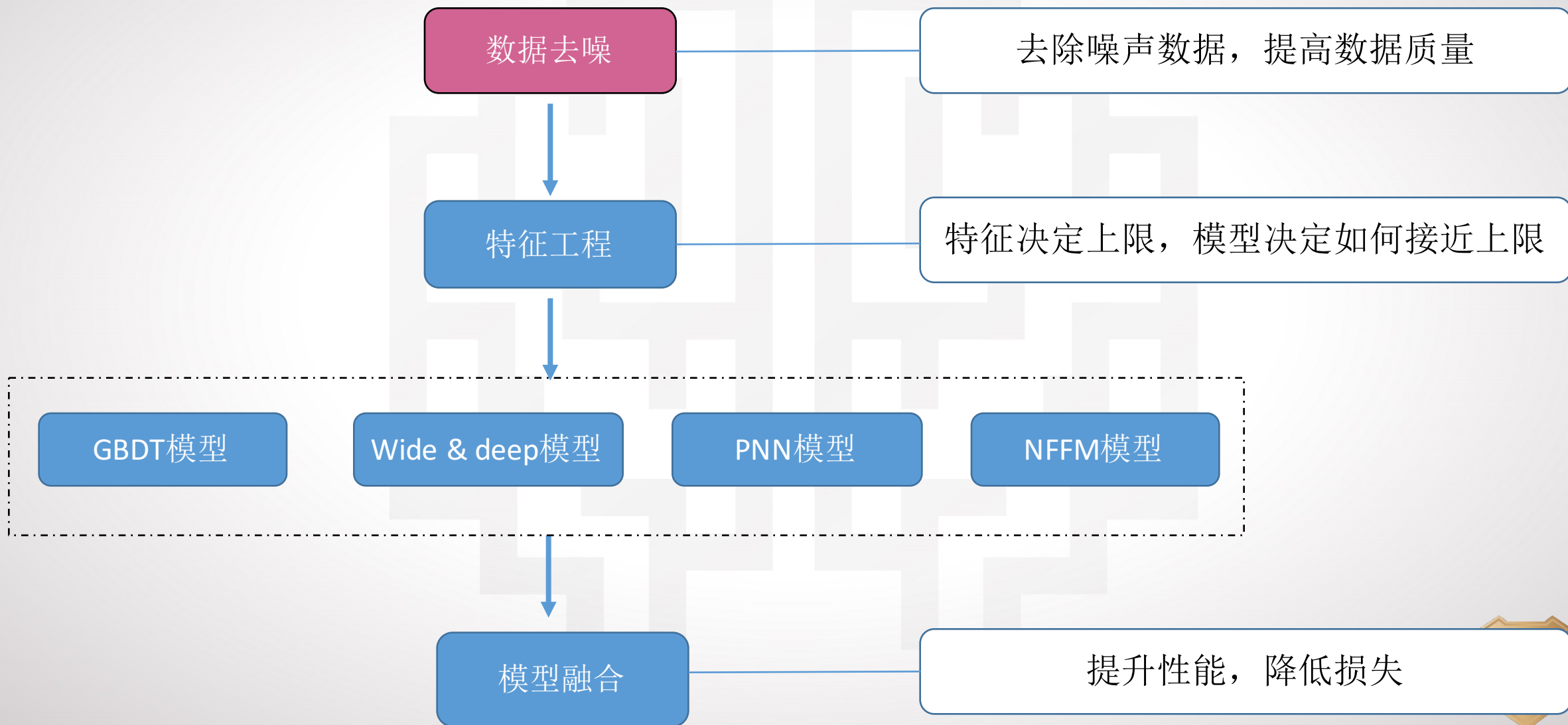
最后一天数据异常(报送延迟)
删除此类app最后一天的数据



转化率突变(app更新)
删除突变之前的数据



解题思路



特征工程

样本构成:

用户特征

app特征

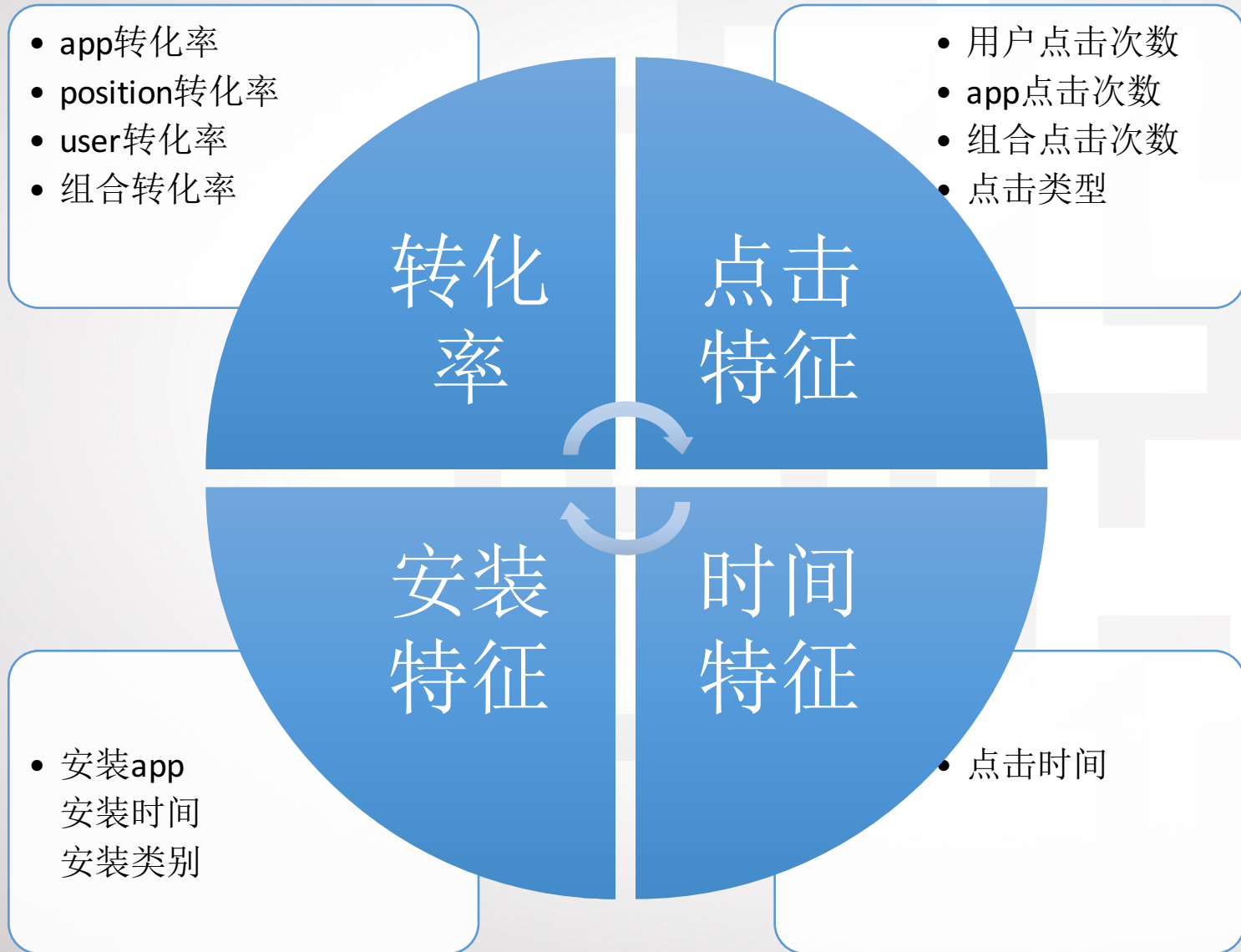
上下文特征

转化率特征

点击特征

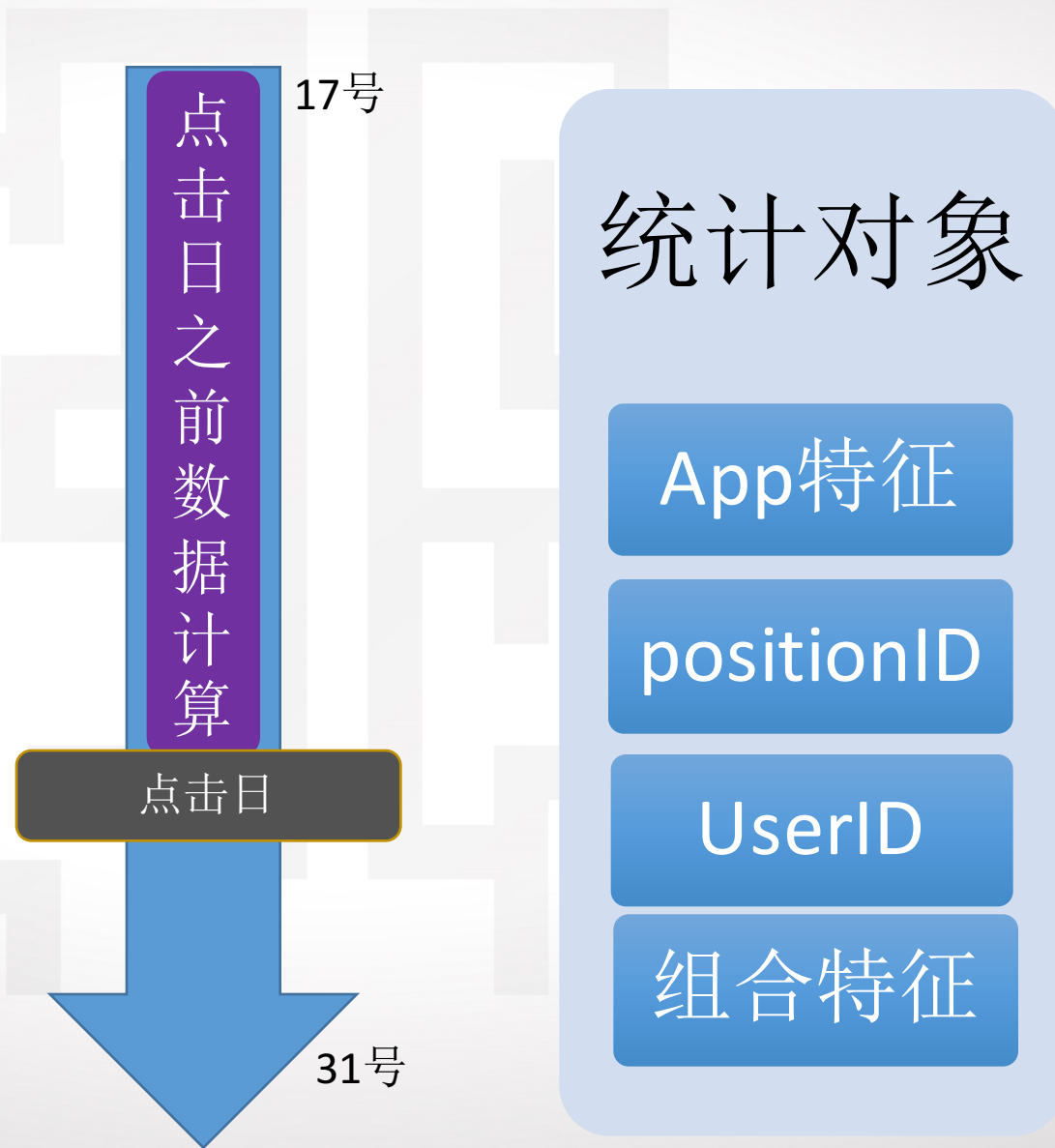
安装特征

时间特征

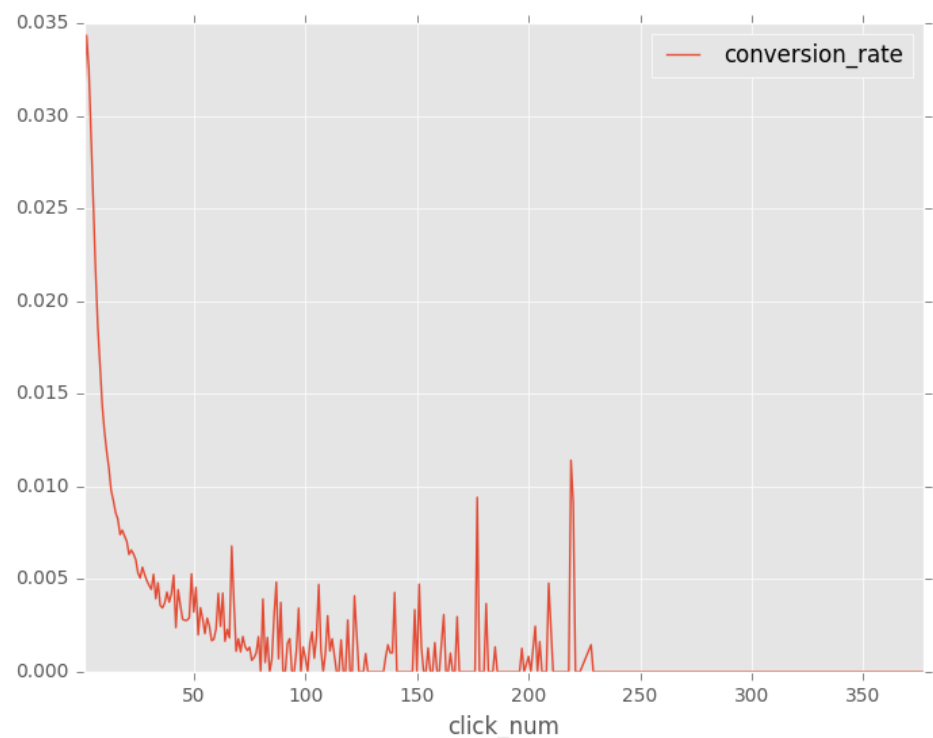


特征工程——转化率(挖掘历史转化信息)

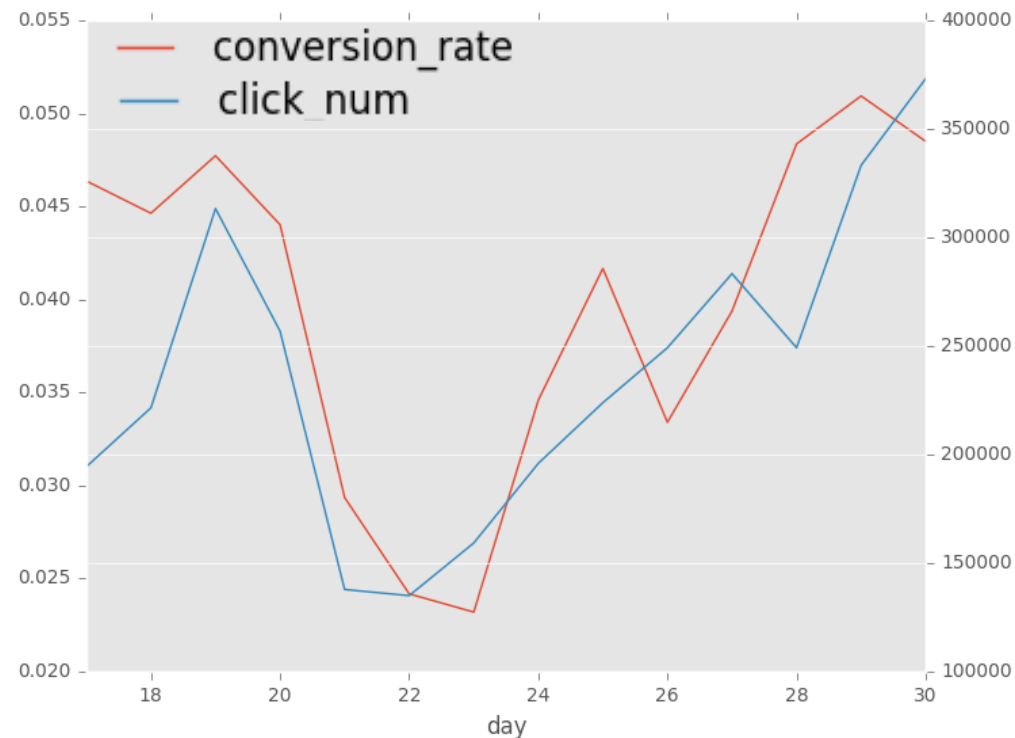
$$\text{转化率} = \frac{\text{转化次数}}{\text{点击次数}}$$



特征工程——点击特征（挖掘重复点击与app热度）



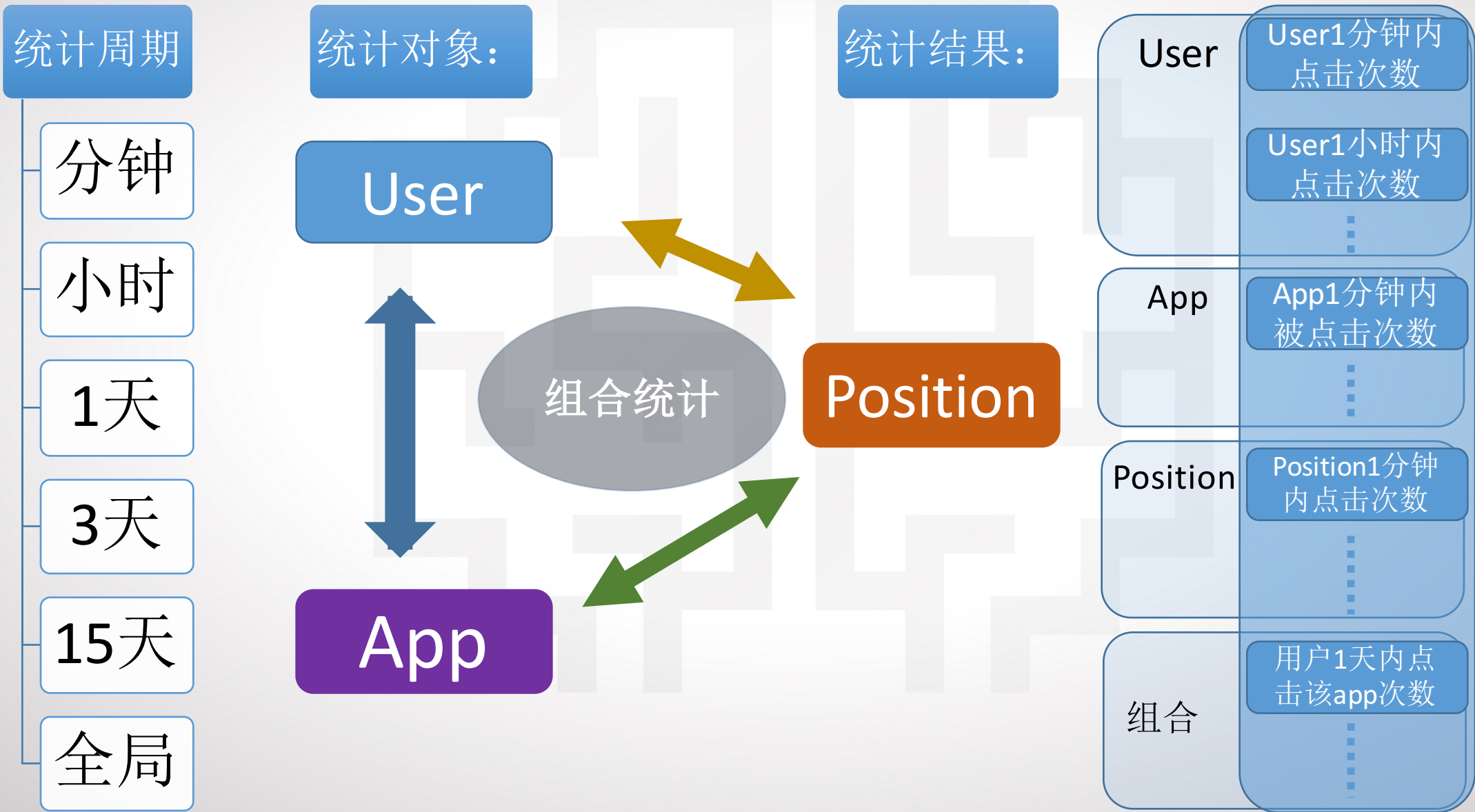
用户点击次数越多，他的转化率越低



app每天的转化率与当天点击次数正相关



特征工程——点击特征（挖掘重复点击与app热度）



特征工程——安装特征(挖掘用户安装偏好)

安装时间

上次 安装
app时间

上次安装此
次点击app
时间差

安装个数

最近安装
app个数

最近安装同
类别app个
数

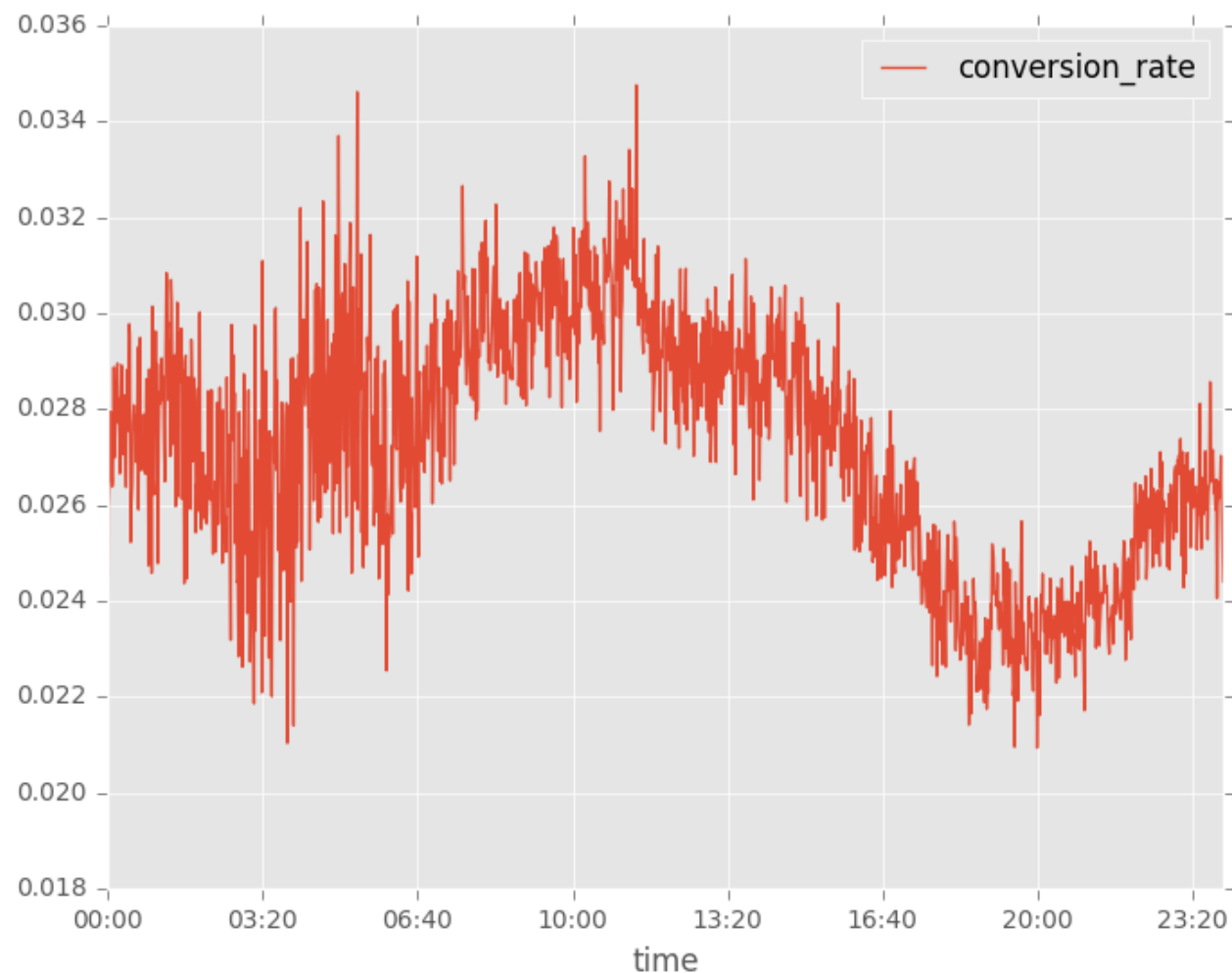
安装app

上次安装
appID

前两次安装
appID组合



特征工程——时间特征（挖掘时间偏好）

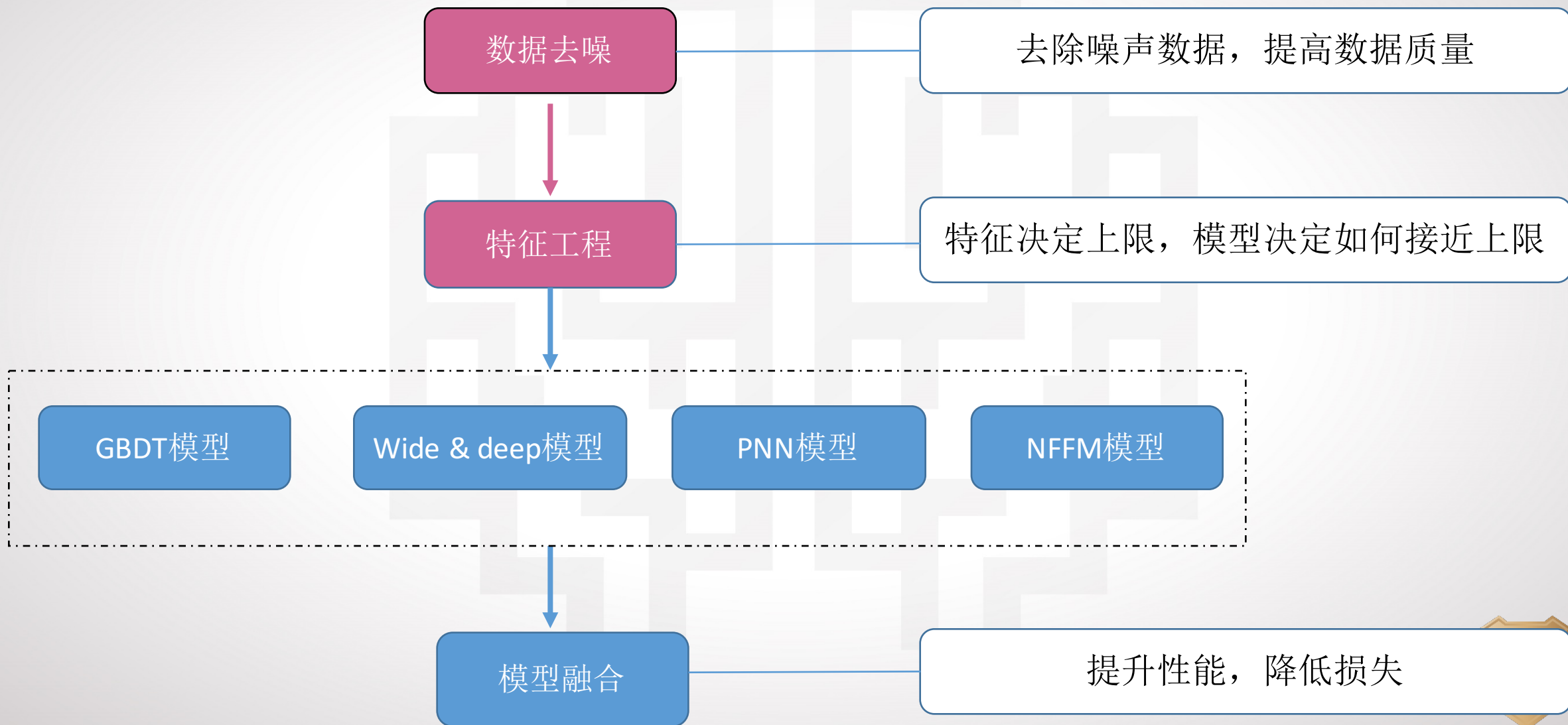


一天24小时分成48个半小时，
点击事件发生区间作为特征

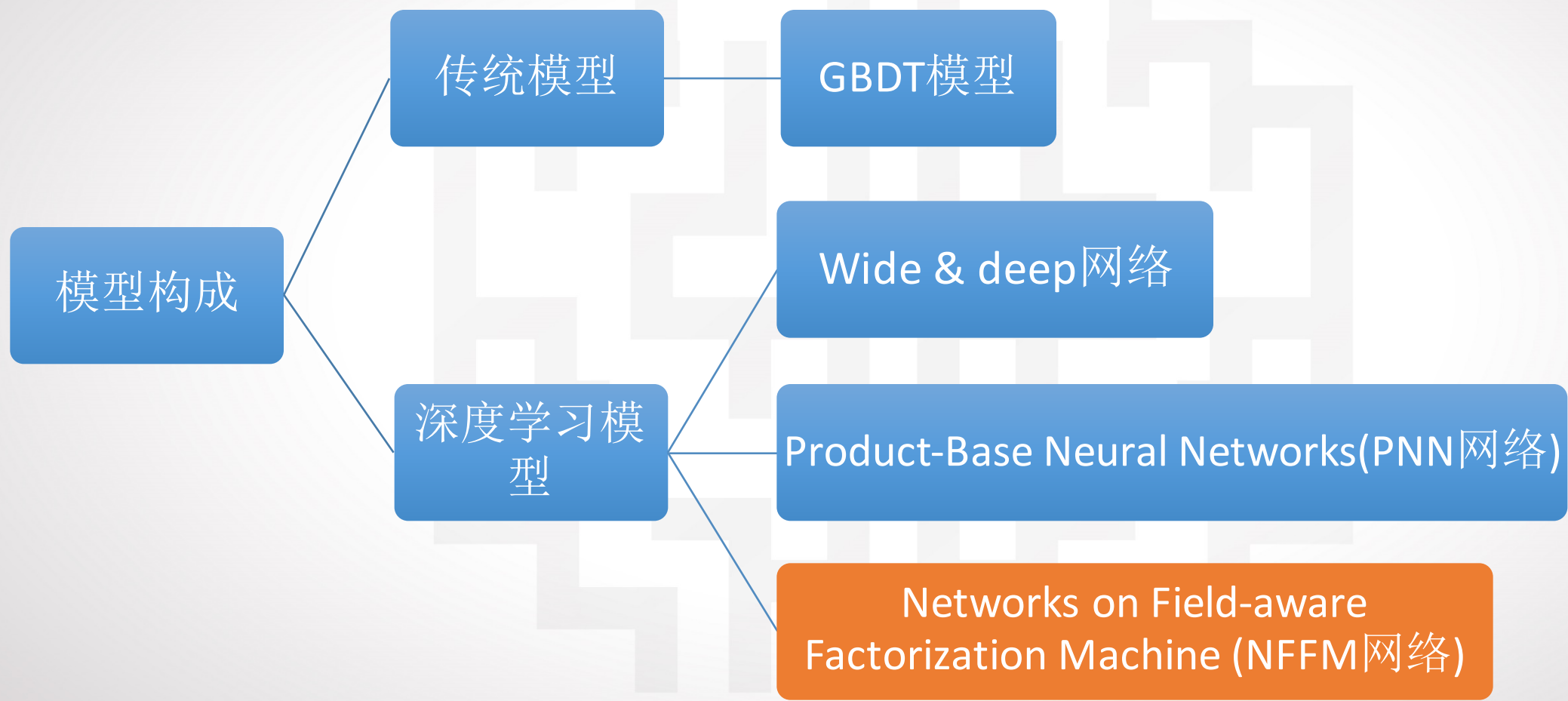
比如11点半点击的样本，它的
时间特征就是23（11点半是
一天中第23个半小时）



解题思路



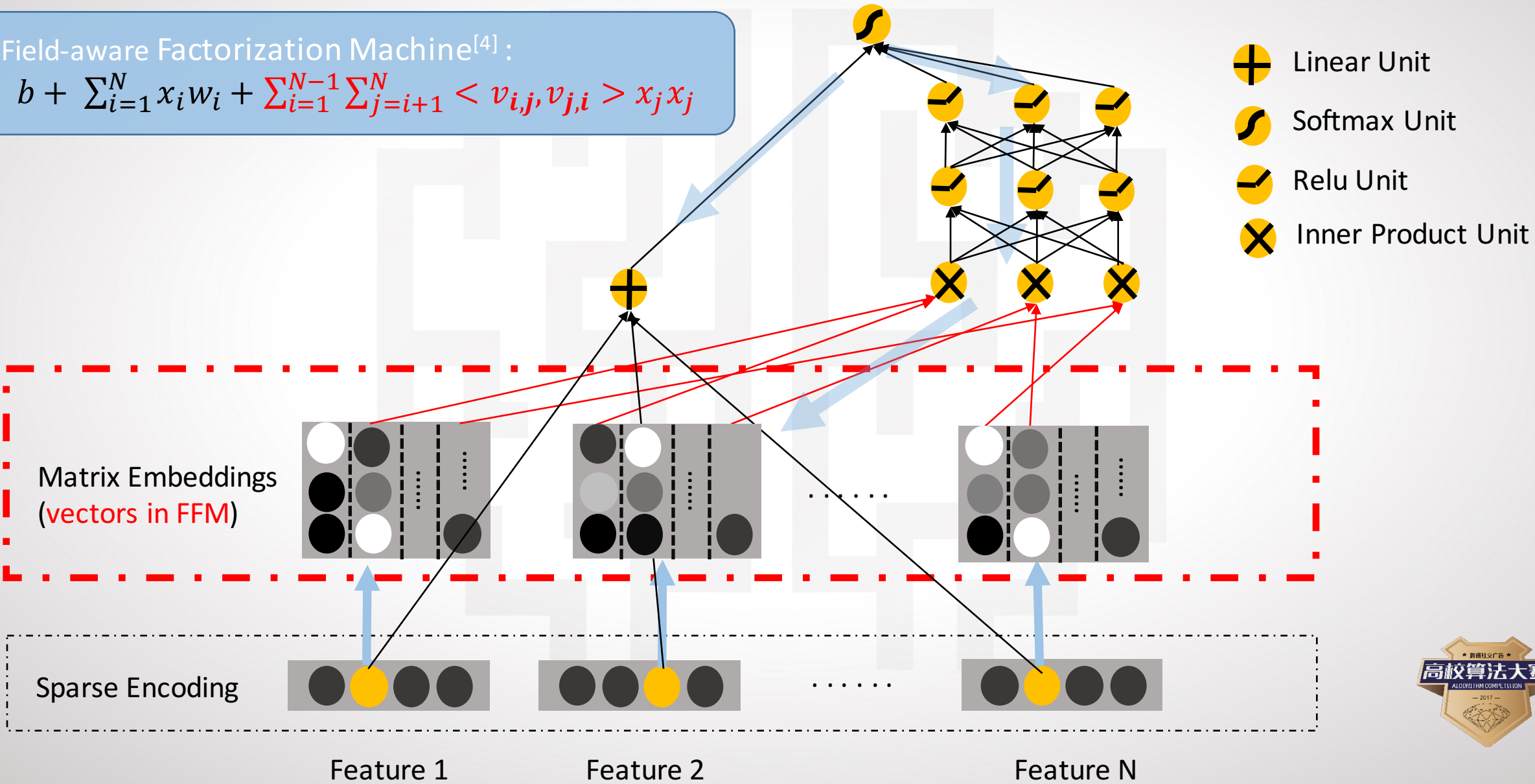
模型构成



模型创新——Networks on Field-aware Factorization Machine(NFFM)

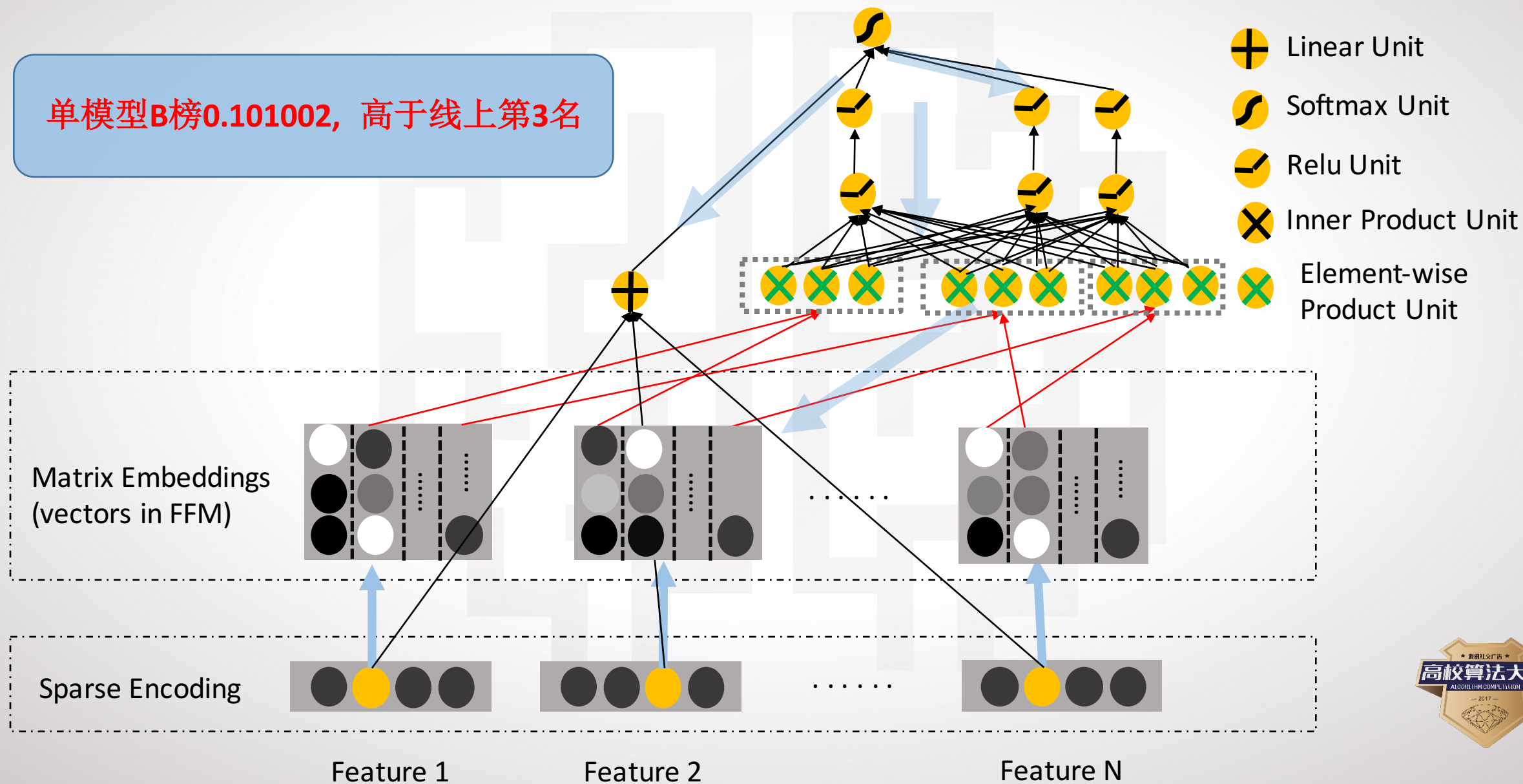
Field-aware Factorization Machine^[4] :

$$b + \sum_{i=1}^N x_i w_i + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \langle v_{i,j}, v_{j,i} \rangle x_j x_i$$

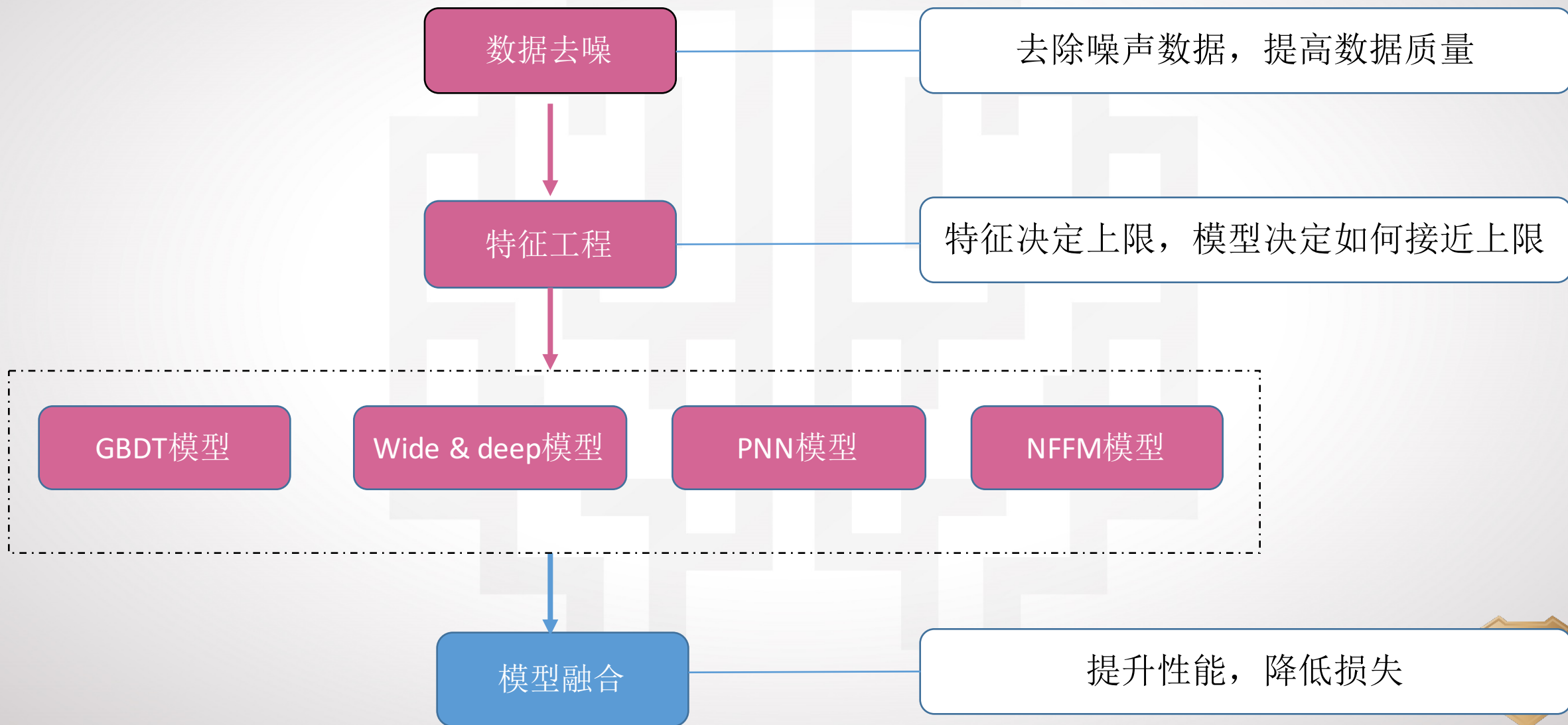


模型创新——NFFM with Element-wise product

单模型B榜0.101002, 高于线上第3名



解题思路



模型融合

两组特征

- 简单特征(39个特征),
复杂特征(49个特征)

8个模型

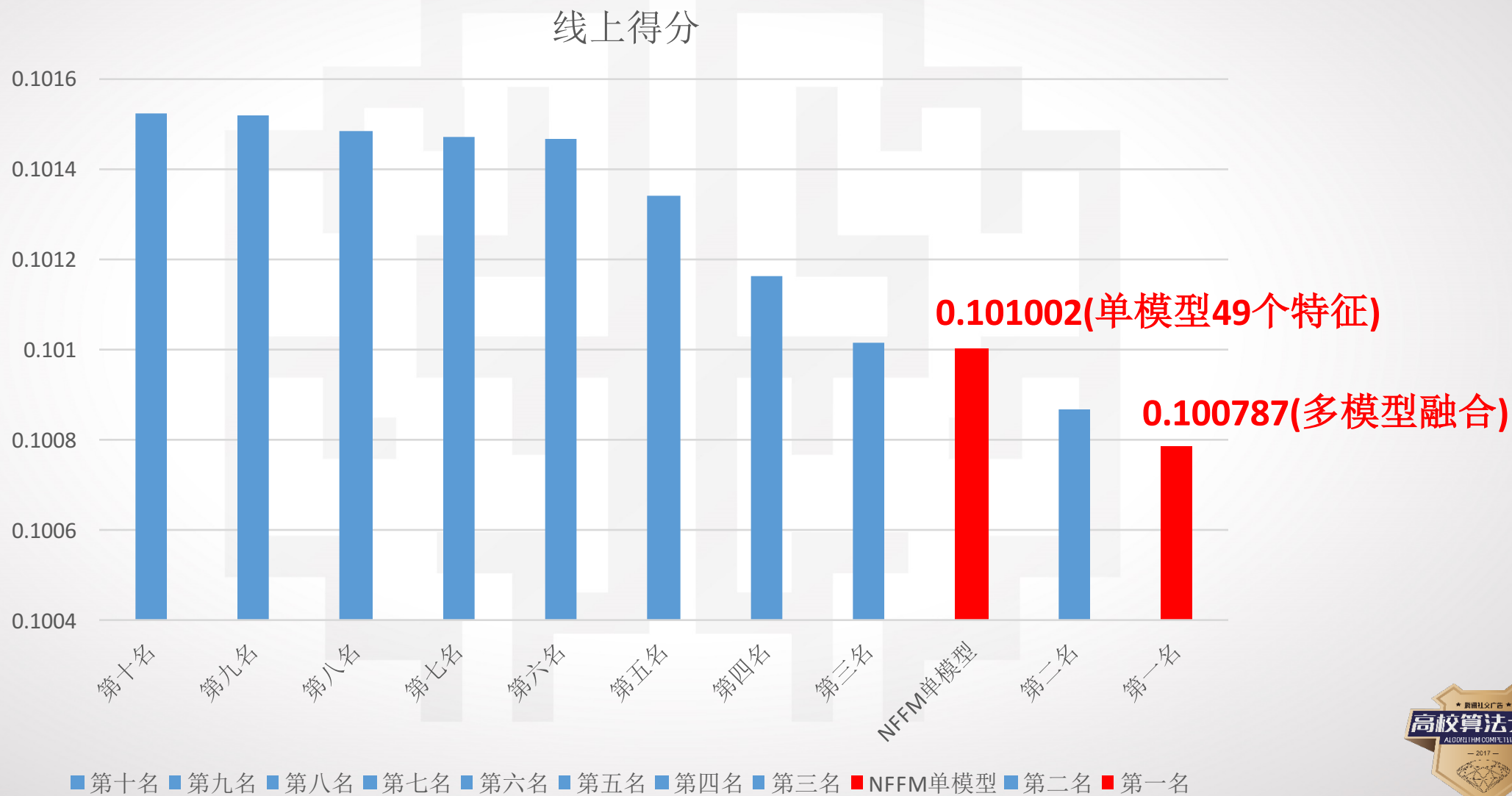
- $1 * \text{GBDT} + 1 * \text{wide\&deep} + 2 * \text{PNN} + 4 * \text{NFFM}$

加权平均

- logit逆变化后融合



模型结果

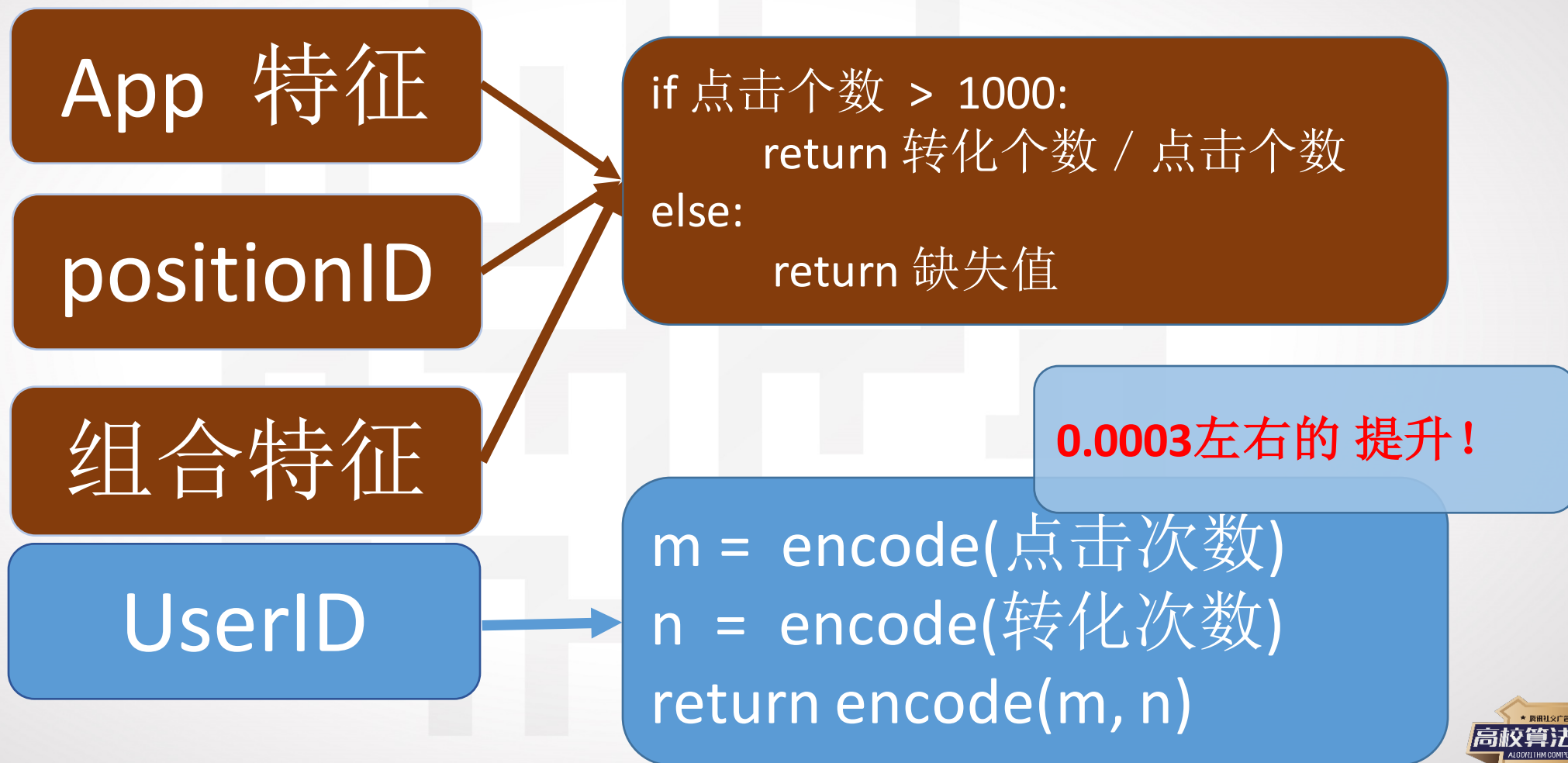


关键问题与解决办法

2017-07-06



问题一：点击次数过少的稀疏值
如何计算转化率？



问题二：过拟合问题

Failed

正则项

DropOut

Batch Normalization

同样的训练时间，提升**0.0006**左右！

Adam优化算法

Batch size: 2500 ~ 10000

3 epochs to **1 epoch**

log loss 降低 **0.0003**左右

随机顺序，训练3次取平均
log loss继续降低**0.0003**左右



问题三：NFFM训练速度问题

Embedding Vectors

VS

Embedding Matrix

训练速度提升三倍以上!



tensorflow的embedding_lookup操作非常耗时!



引文

- [1] Guo, Huifeng, et al. "DeepFM: A Factorization-Machine based Neural Network for CTR Prediction." arXiv preprint arXiv:1703.04247 (2017).
- [2] Qu, Yanru, et al. "Product-based neural networks for user response prediction." Data Mining (ICDM), 2016 IEEE 16th International Conference on. IEEE, 2016.
- [3] Cheng, Heng-Tze, et al. "Wide & deep learning for recommender systems." Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM, 2016.
- [4] Juan, Yuchin, et al. "Field-aware factorization machines for CTR prediction." Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.
- [5] Rendle, Steffen. "Factorization machines." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.



`<=Date()`

`freeze()`

`for`

THANKS

9

2017-07-06

8

`if`

`Your`

`var`

`time`