队名：levy

Contents

目录

赛题理解

评测指标/赛题建模分析

大赛赛题：广告日曝光预估

01

1）准确性指标

$$\mathbf{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(F_t + A_t)/2}$$

02

2）出价单调相关性指标

$$\text{score} = \frac{1}{n}\sum_{k=1}^{n}\frac{(imp_0 - imp_k)(bid_0 - bid_k)}{|(imp_0 - imp_k)(bid_0 - bid_k)|}$$

03

3）最终得分

$$\text{TotalScore} = w_1 * \left(1 - \frac{SMPAE}{2}\right) + w_2 * \frac{MonoScore + 1}{2}$$

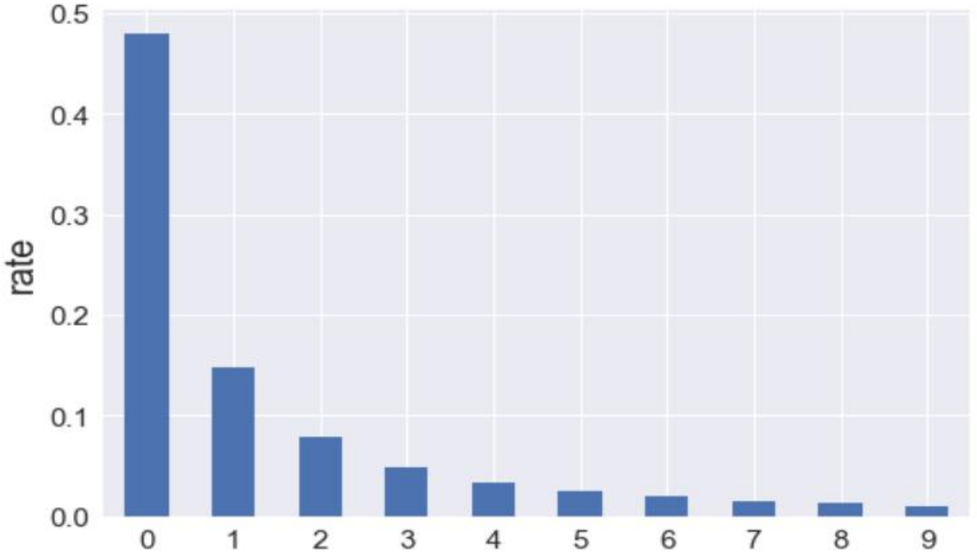
曝光量为非负整数

小曝光量样本占绝大多数

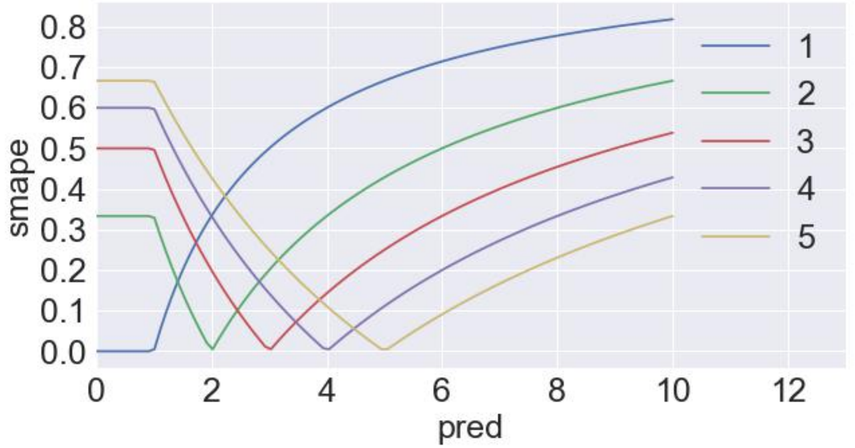
线上对0曝光样本的平滑处理

$$pred \leftarrow \max(pred, 1)$$

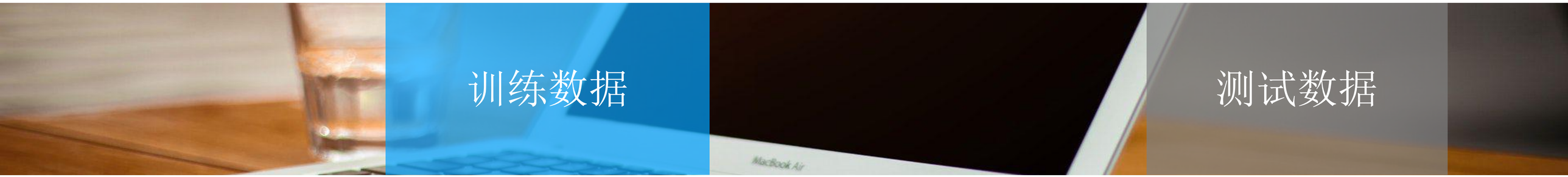
$$label \leftarrow \max(label, 1)$$



不同曝光量样本的训练集占比（长尾）



不同曝光量下smape与预测日曝光量的关系



线下验证集构造

4/10 4/11 4/12 ..... 4/15 4/16 4/17

4/10 - 4/15的竞价日志中竞价数>35的广告

4/16- 4/17的竞价日志(按竞价数分布采样挖去5000+广告的竞价日志做广告采样)

线上训练数据集

4/10 4/11 4/12 ..... 4/22 4/23 4/24

4/10 - 4/22的竞价日志中竞价数>35的广告

4/24参与的竞价队列的广告当日竞价数>35

建模设想1：Point-wise rank

Basic_ecpm = cpm_bid
= 1000 * cpc_bid * pctr
= 1000 * cpa_bid * pctr * pcvr

Quality_ecpm ≈ 20*pctr

Total_ecpm = Basic_ecpm + Quality_ecpm

曝光

广告id	Rank of Total_ecpm	是否过滤
广告a	1	True
广告b	2	True
广告c	3	True
广告d	4	False
广告e	5	False
广告f	6	False

建模设想1：Point-wise rank

存在的问题：

1. pctr, pcvr的预测偏差

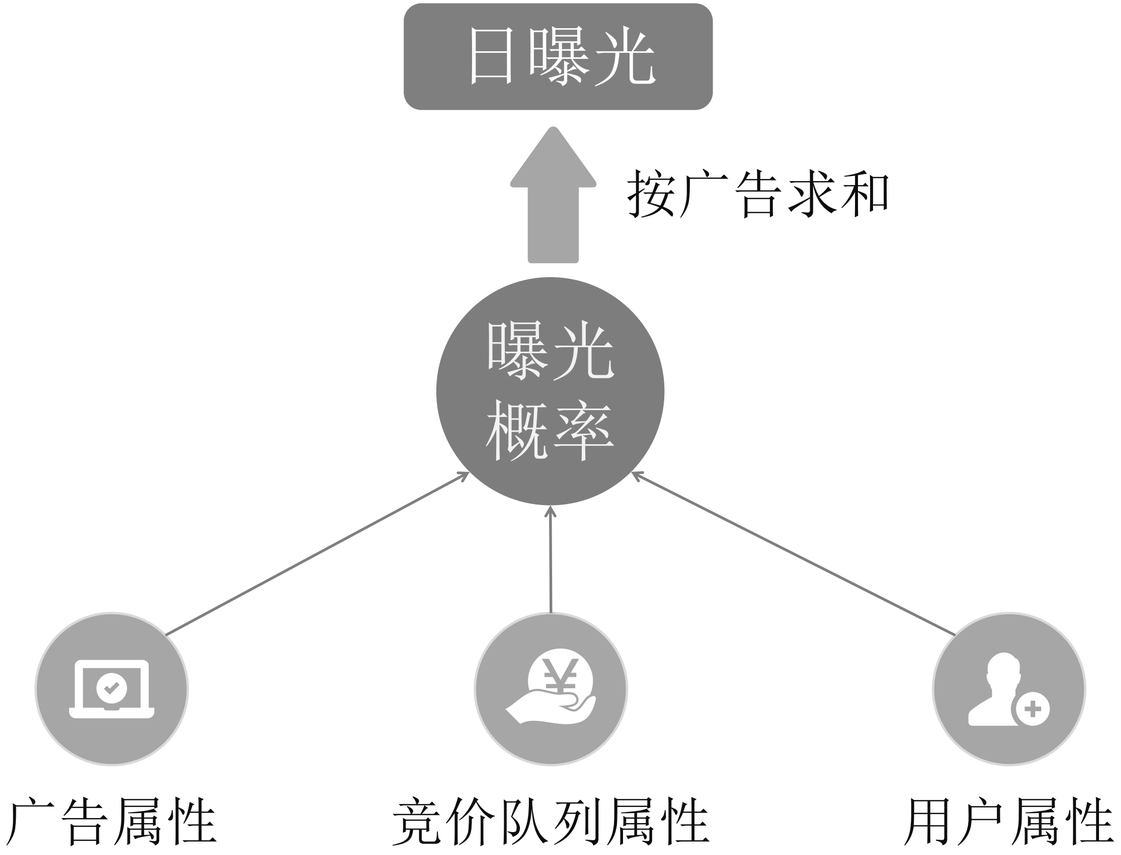
2. 每条测试广告的真实出价隐藏在多条不同的虚拟出价中

广告id	Rank of Total_ecpm	是否过滤
广告a	1	True
广告b	2	True
广告c	3	True
测试广告1	4	True
测试广告2	4	False
测试广告3	4	False
广告d	5	False
广告e	6	False
广告f	7	False

建模设想2：预测某广告的某次竞价是否曝光（2分类）

存在的问题：

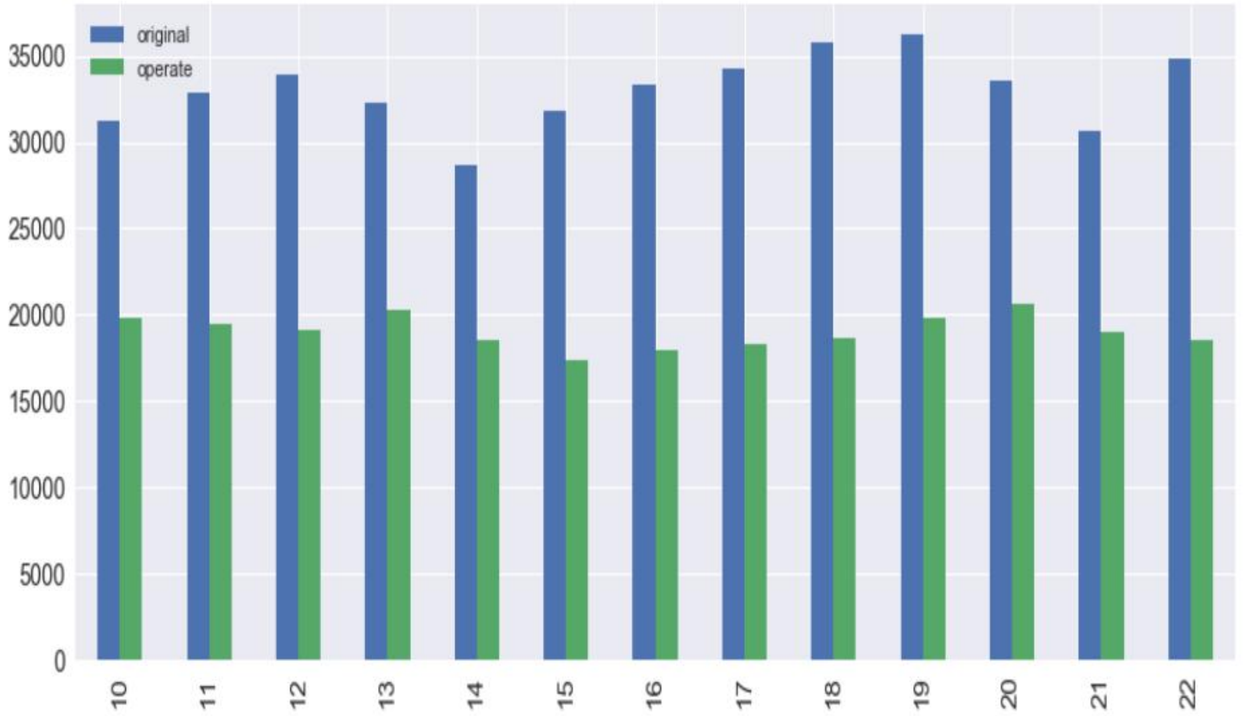
1. 正负样本极度不均衡
2. 预测的概率和真实的曝光期望之间存在偏差
3. 未充分考虑广告过滤的影响
4. 数据量很大



建模设想3：回归日曝光/出价

优势：
利用了出价信息，模型输出即已满足出价单调性

缺点：
可用样本数减少近一半



建模设想4：直接回归日曝光（<u>最终方案</u>）

缺点

- 上限较低
- 难以充分利用用户信息和竞争队列信息
- 输入label的处理和loss function的选择对结果影响较大

优势

- 模型简单
- 易于快速迭代
- 鲁棒性好
- 能充分利用训练集样本

特征工程

特征介绍/特征有效性

特征工程

广告属性特征

- 广告静态属性（广告主、素材尺寸、广告领域、……）
- 广告动态属性（计费类型、转化类型）

竞价日志特征

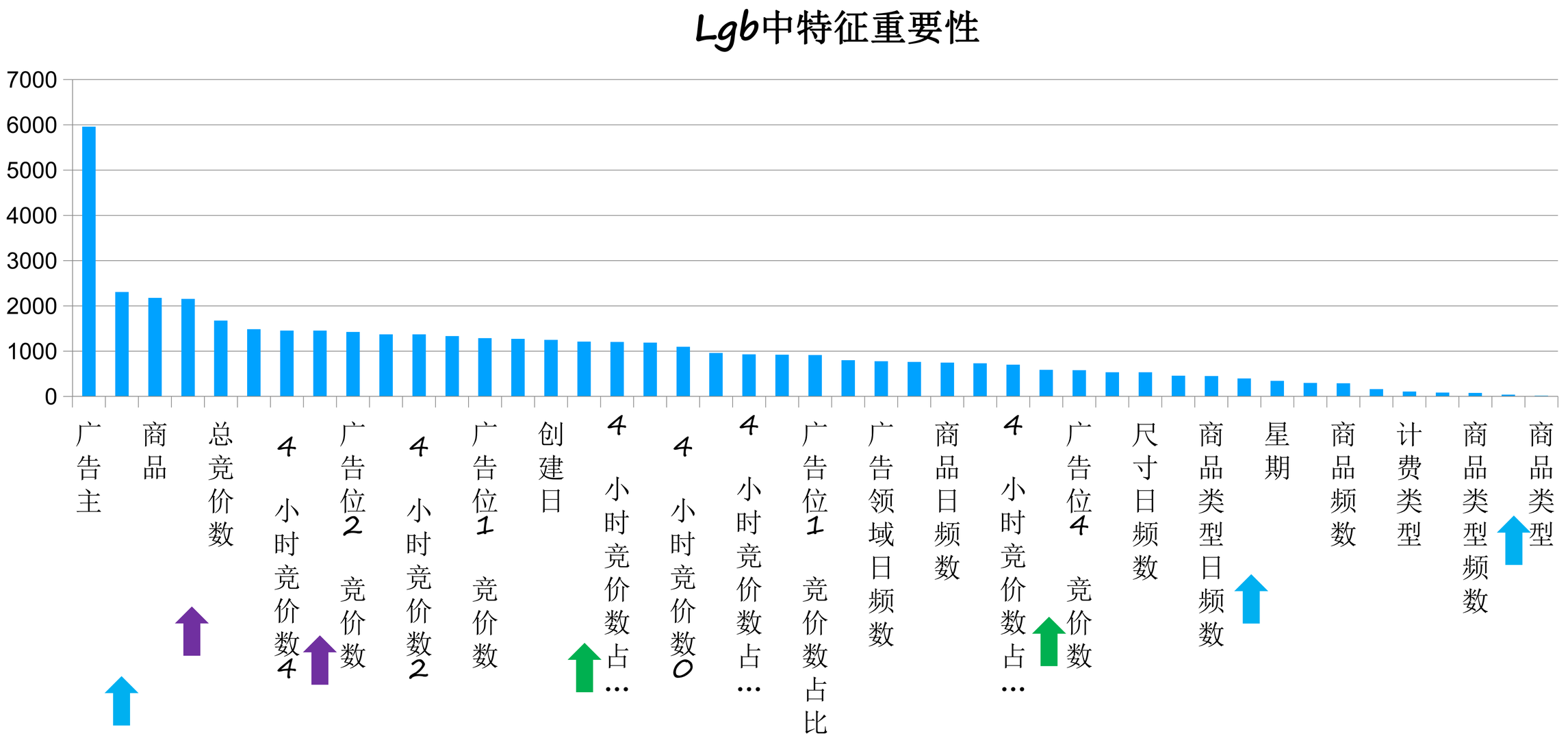
- 竞价数及其时空信息（不同广告位和时间段的竞价数分布）
- 用户重复曝光情况
- 针对旧广告，统计对应的历史n天的曝光量以及竞价次数

统计特征

- 广告属性频数（总体）
- 广告属性频数（当天）

时间特征

- 预测日
- 星期
- 广告生存周期（预测日-创建日）


Lgb中特征重要性
7000
6000
5000
4000
3000
2000
1000
0
广告主
商品
总竞价数
4小时竞价数4
广告位2竞价数
4小时竞价数2
广告位1竞价数
创建日
4小时竞价数占…
4小时竞价数0
4小时竞价数占…
广告位1竞价数占比
广告领域日频数
商品日频数
4小时竞价数占…
广告位4竞价数
尺寸日频数
商品类型日频数
星期
商品频数
计费类型
商品类型频数
商品类型

模型介绍

通用处理/旧广告/新广告

STEP1

Label处理

•用0-1之间的给定值REPLACE_ZERO替换0曝光样本的标签并取log

Why not log1p?

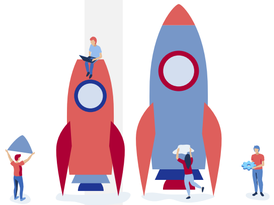
$$\left\|\log\frac{e_1}{e_0}\right\|^2 \qquad \left\|\log\left(\frac{1+e_1}{1+e_0}\right)\right\|^2$$

STEP2

梯度修正

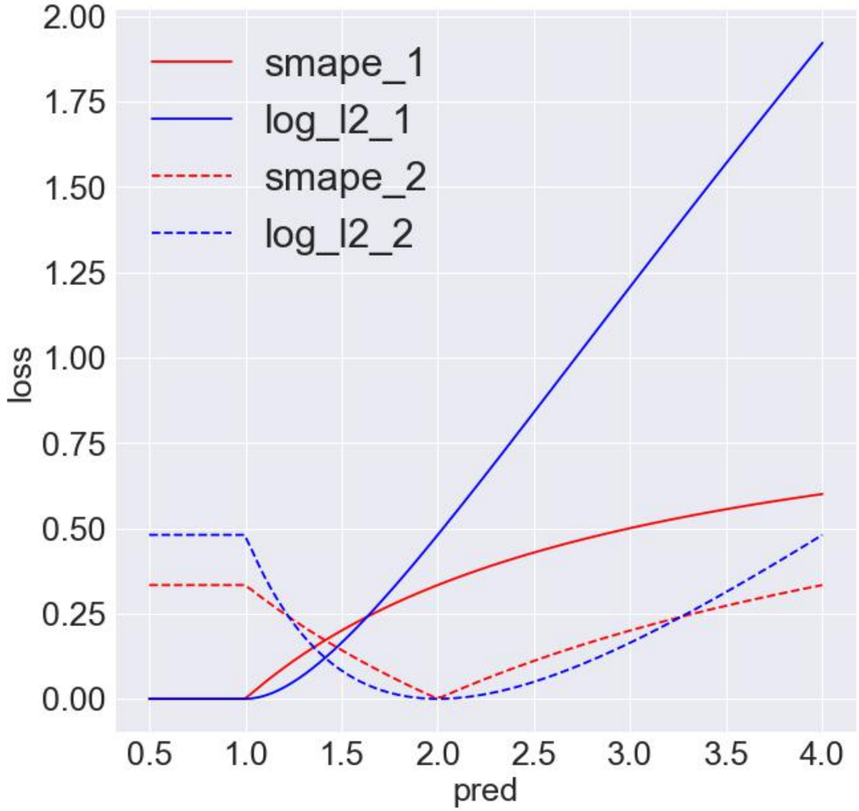
•对于预测值偏小的0&1曝光样本不反传梯度

STEP3



组合loss

•L2-loss与smape的加权组合



旧广告预测

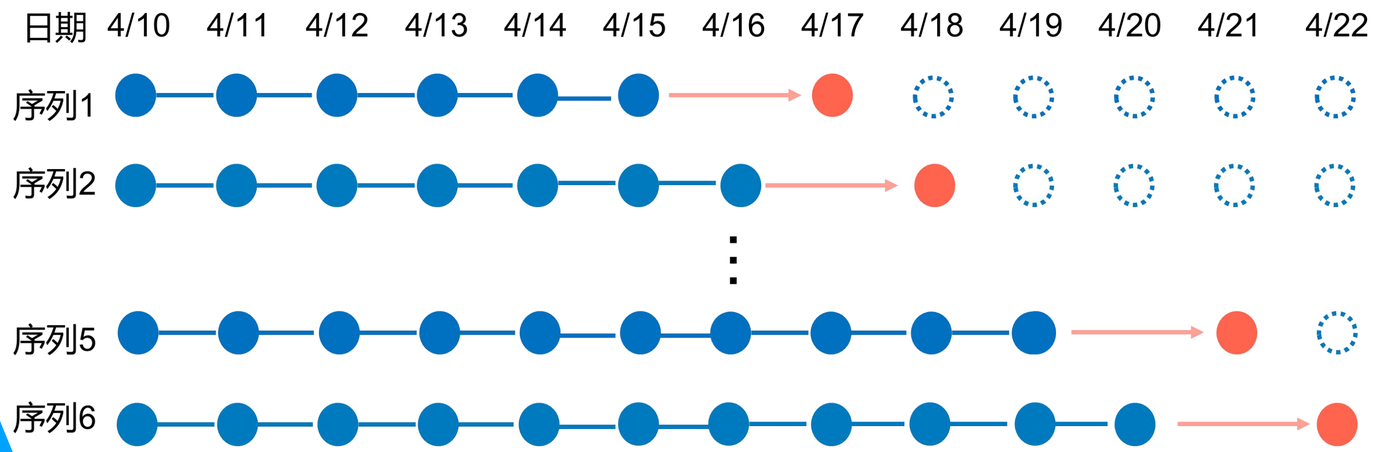
未利用预测日竞价数信息

历史日曝光？

加权历史日曝光率

历史日曝光率？

剥离地利用不同天的曝光率信息



$$e_{pred} = c_{test} \times \frac{\sum w_i e_i}{\sum w_i c_i}$$

优势：

· 充分利用广告的所有历史竞价e_i和曝光c_i

· 通过权重w_i综合考虑竞价数和与预测日的间隔天数

简单NN预测权重

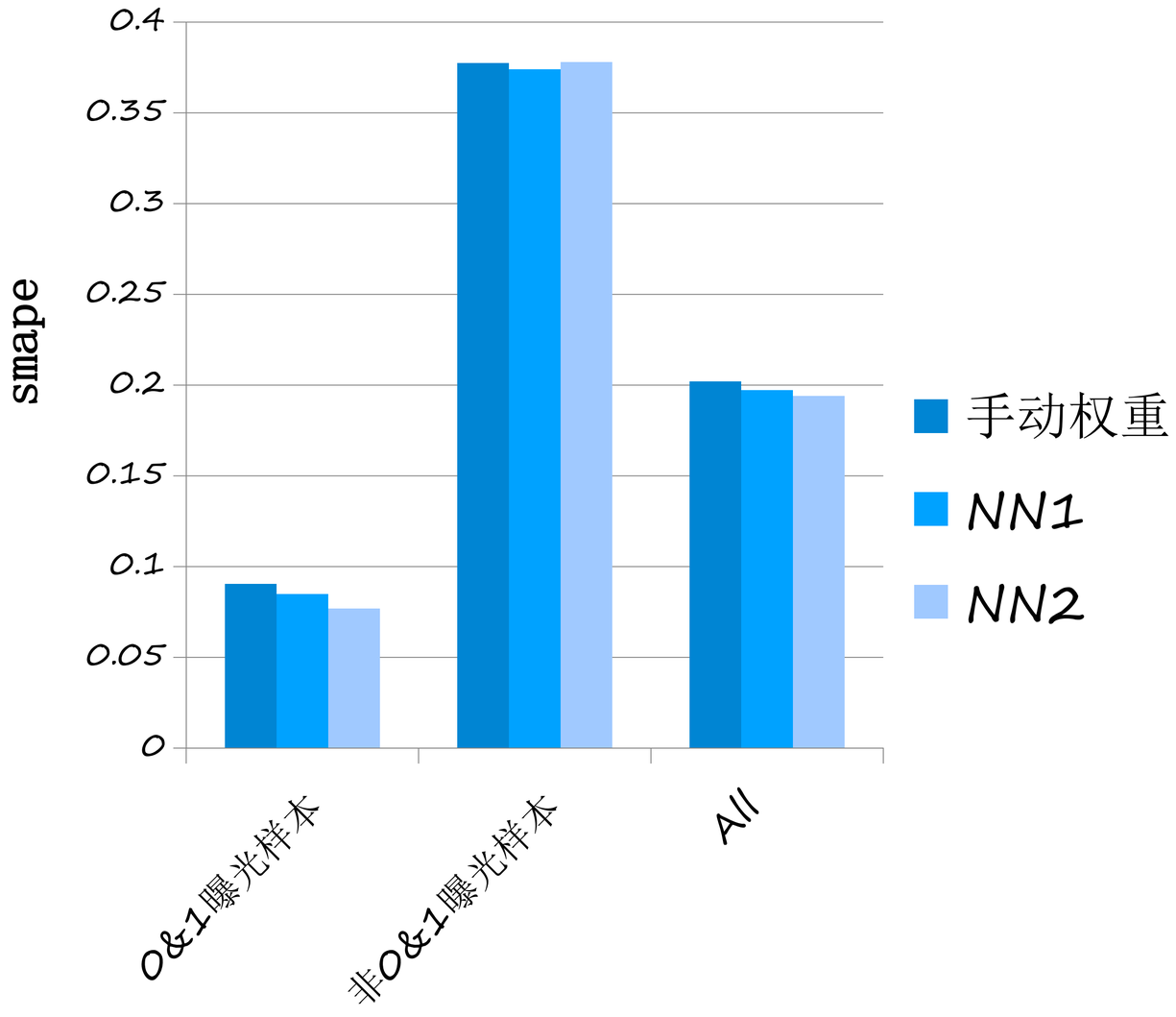
历史日	标识与预测日的天数差	标识预测日星期几	竞价数	曝光数	预测权重
	1	**1**	**21**	**4**	**0.08**
	1	**1**	**7**	**2**	**0.07**
4/10	1	1	12	1	0.1
4/11	1	1	5	1	0.15
4/12	1	1	0	0	0.17
4/13	1	1	0	0	0.23
4/14	1	1	33	4	0.46
4/15	1	1	61	16	0.85

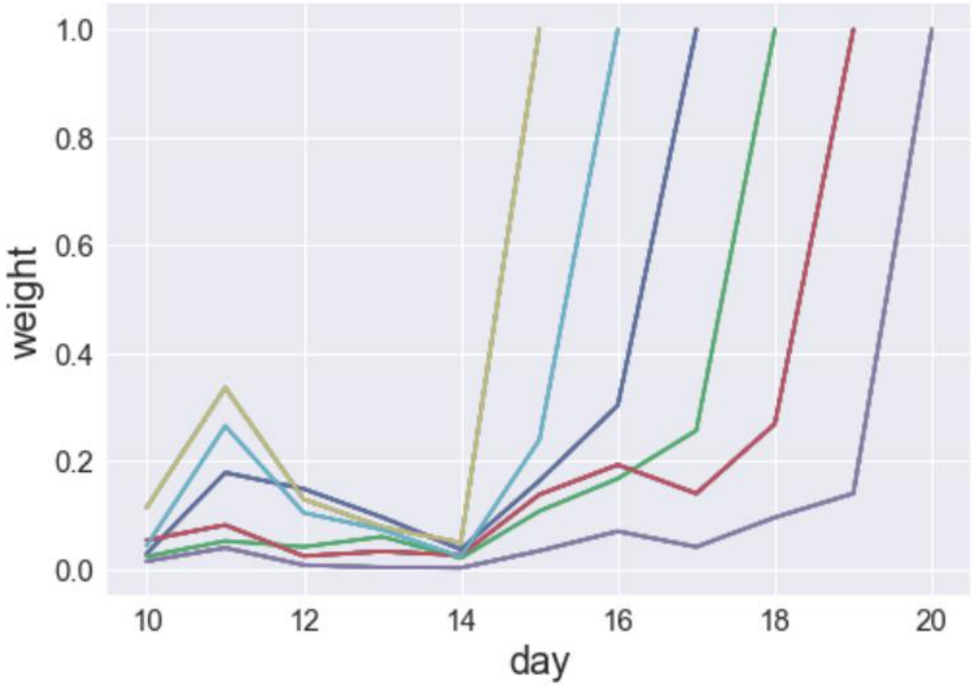
输入 → 两层神经网络 → 输出

NN1

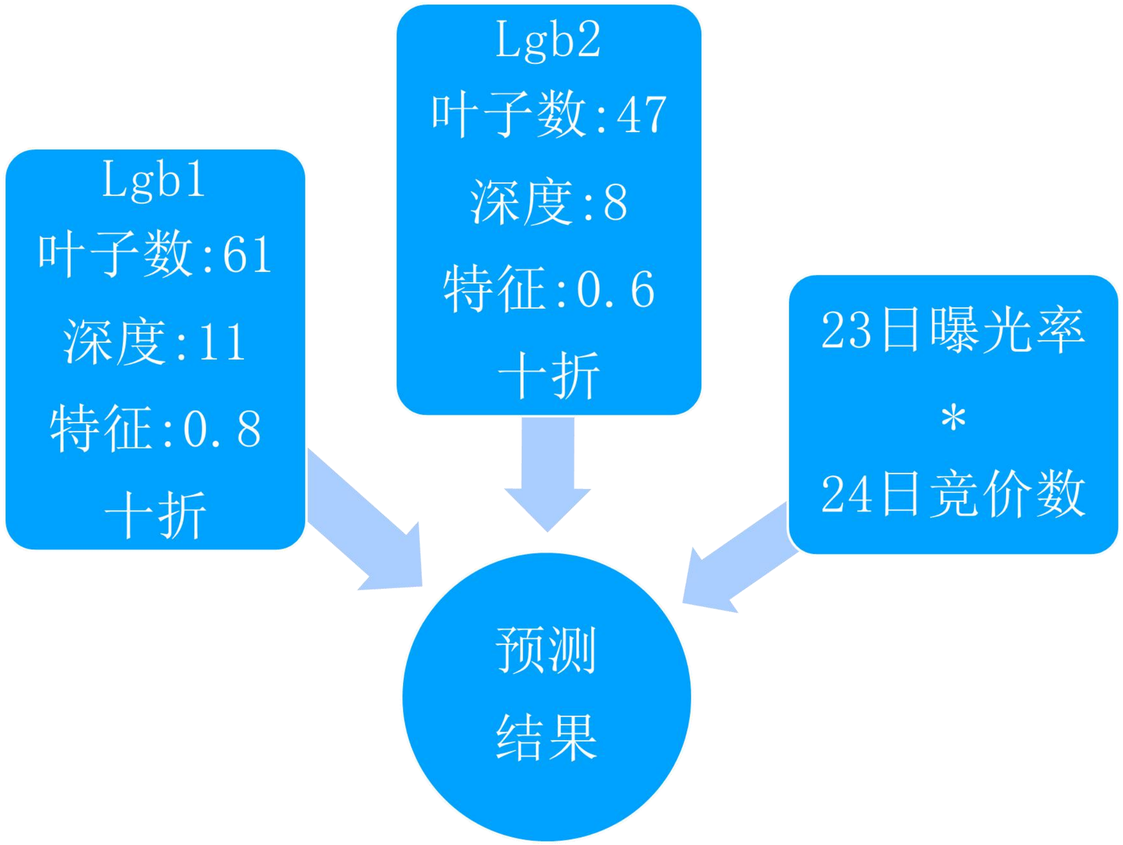
NN2

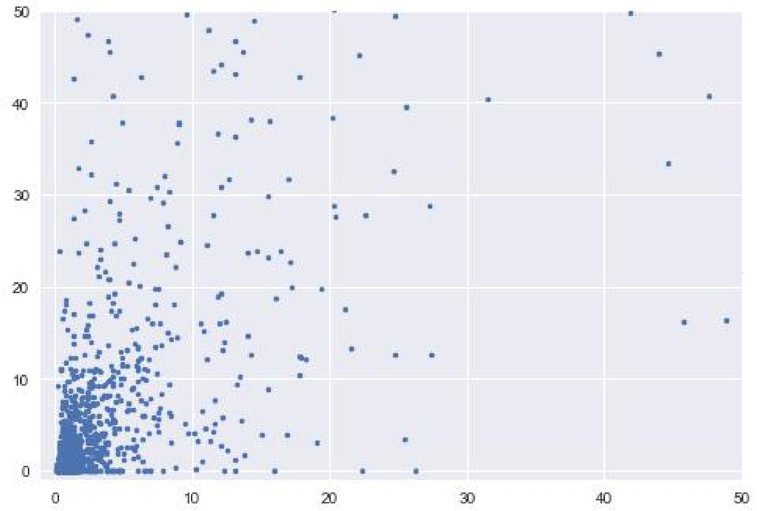
模型介绍


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
smape
0&1曝光样本
非0&1曝光样本
All
手动权重
NN1
NN2


1.0
0.8
0.6
0.4
0.2
0.0
weight
10
12
14
16
18
20
day

新广告预测


Lgb1
叶子数:61
深度:11
特征:0.8
十折
Lgb2
叶子数:47
深度:8
特征:0.6
十折
23日曝光率
*
24日竞价数
预测
结果


50
40
30
20
10
0
0
10
20
30
40
50

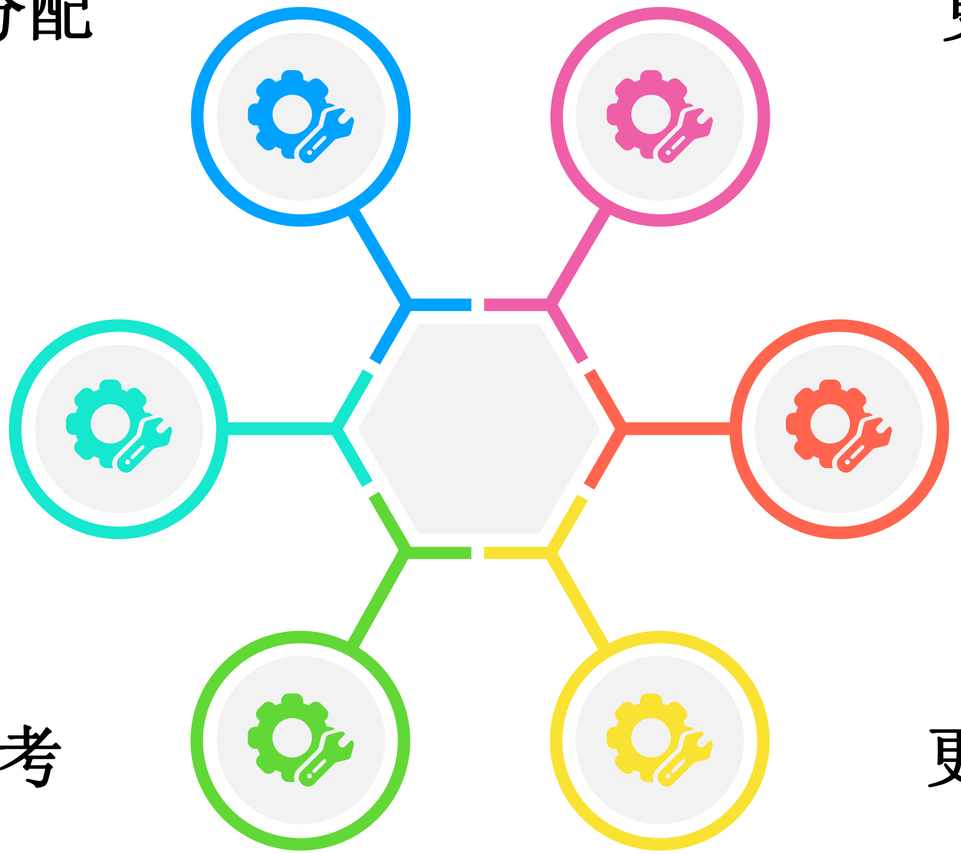

add
mul
0.210
0.205
0.200
0.195
0.190
0.185
0.180
0.0
0.2
0.4
0.6
0.8
1.0

总结与思考

思考/感谢

更合理的精力分配

更丰富的模型



更紧密的团队配合

更细致的特征工程

更深入的思考

更快速的迭代

感谢主办方的认真负责

感谢评委老师的倾听

感谢队友的一起拼搏

THANKS