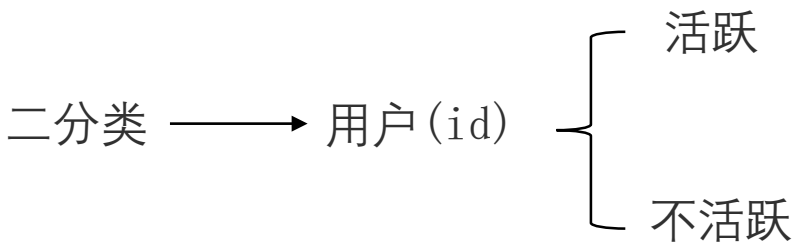
大数据挑战赛解决方案分享

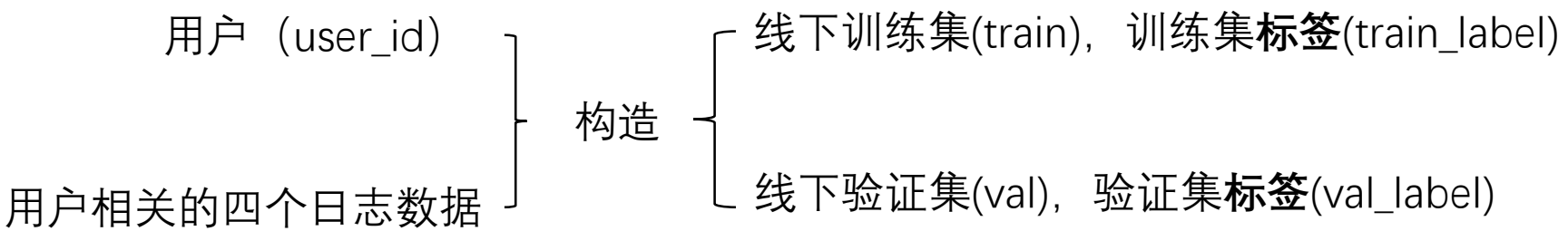
队伍：小小蚂蚁

赛题分析：

题目描述：给定1-30天时间内的四个日志数据，通过这些数据来预测未来一段时间（即31-37天）活跃（即出现在以上四个日志中任意一个）的用户。

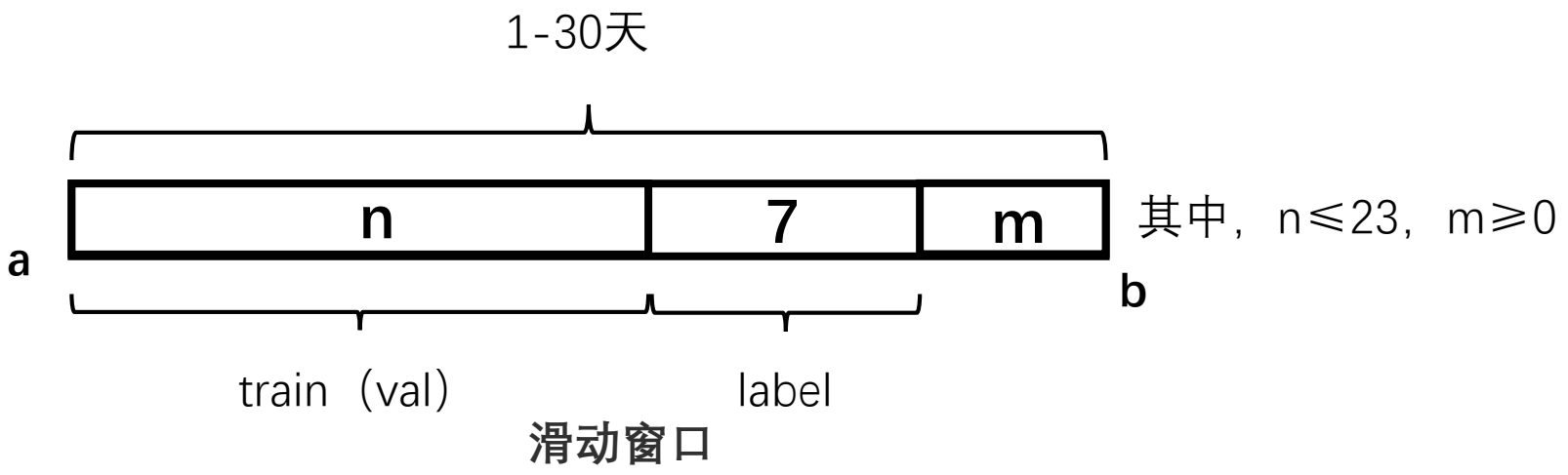


难点与挑战？



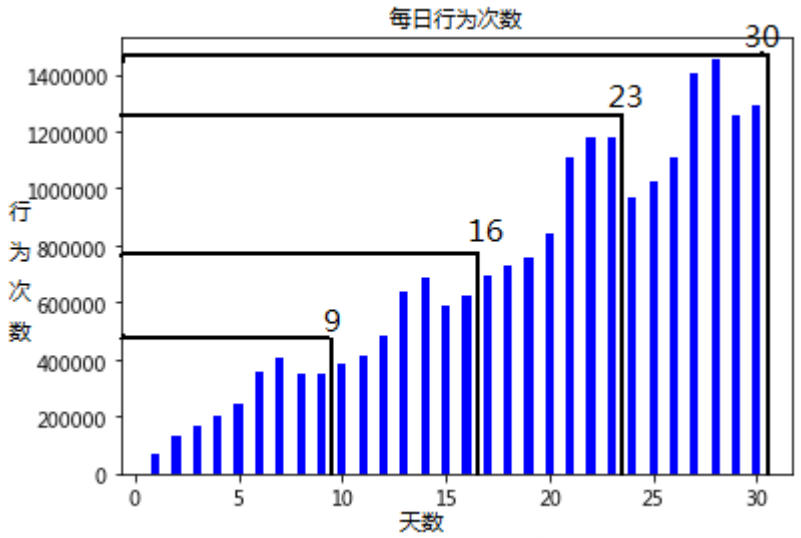
数据探索：

图一：



其中，n≤23，m≥0

窗口选择：（1）线上线下窗口间隔一致（2）满足数据分布的周期性



	线下训练集		线下验证集	线上训练集			线上测试集
变长窗口	1-9	1-16	1-23	1-9	1-16	1-23	1-30
定长窗口	1-16	8-23 (4/5)	8-23 (1/5)	1-16	8-23		15-30

数据探索：

窗口确定：

	线下训练集		线下验证集	线上训练集			线上测试集
变长窗口	1-9	1-16	1-23	1-9	1-16	1-23	1-30
定长窗口	1-16	8-23 (4/5)	8-23 (1/5)	1-16	8-23		15-30

原因分析：

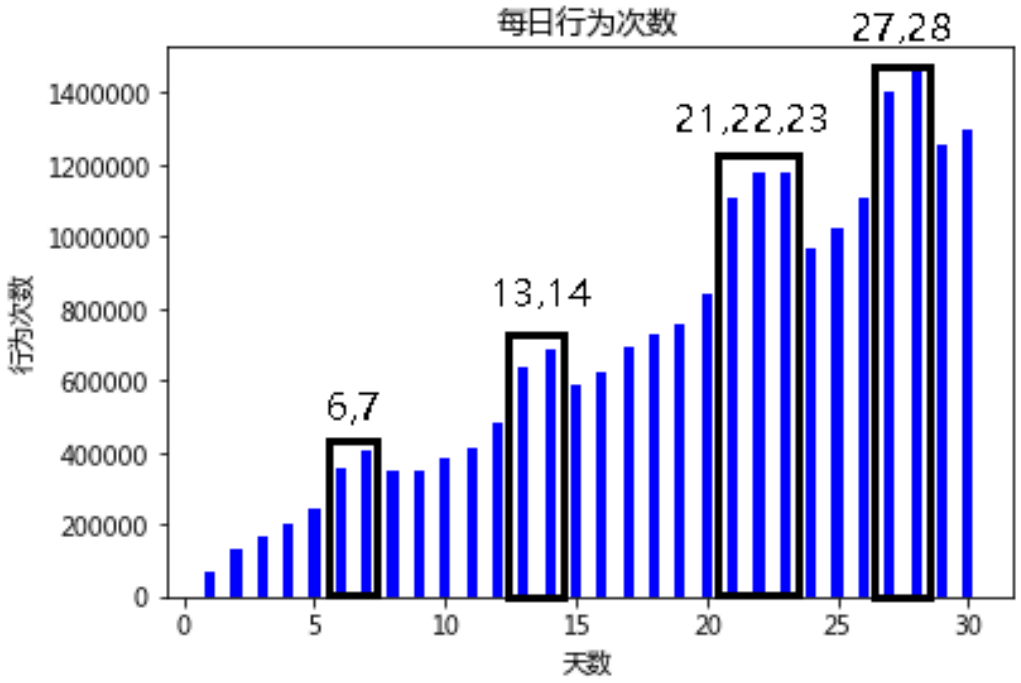
(1) 变长窗口可以预测全范围的用户活跃概率，而定长窗口只能预测是15-30时间段内的所有用户。

(2) 变长窗口有利于线下的验证集构造，因为数据分布基本一致，保持线上线下一致，而相比定长窗口的切分窗口数据好。

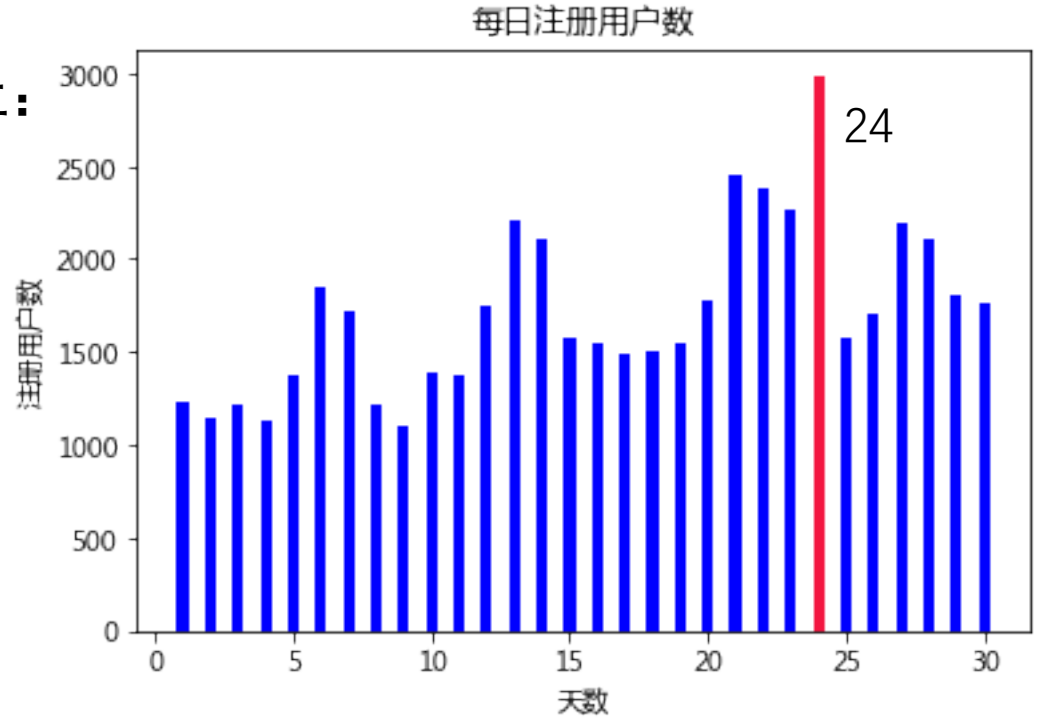
(3) 变长窗口构造的训练集相比较定长窗口多。

数据探索：对用户进行是否活跃预测

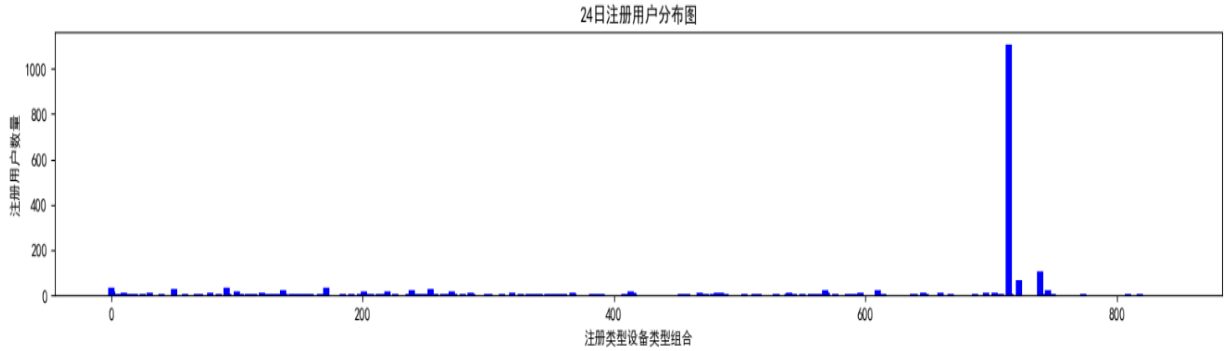
图一：



图二：



图三：



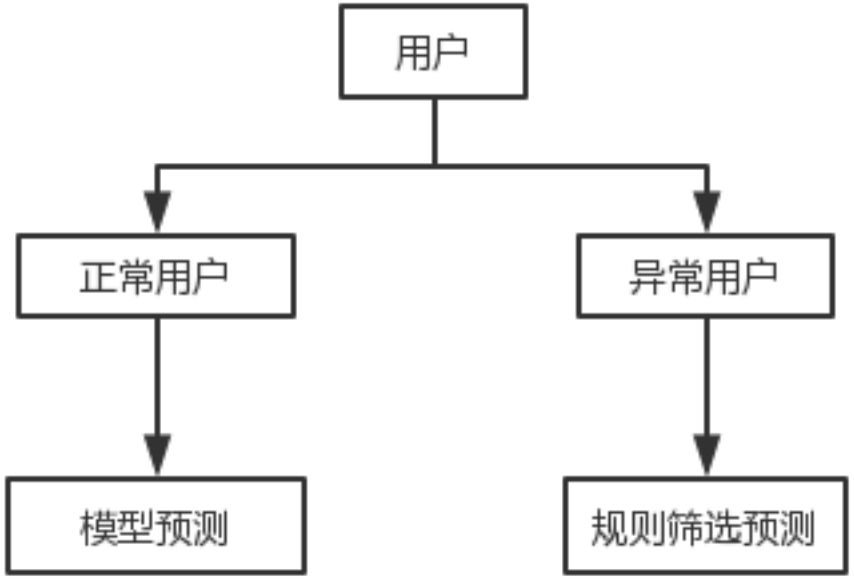
异常用户：

（1）组合相同，一天内注册量突增

（2）在随后的所有天数内都没有活跃，只在当天活跃。

数据探索：

分析结果（对用户群的分类）：

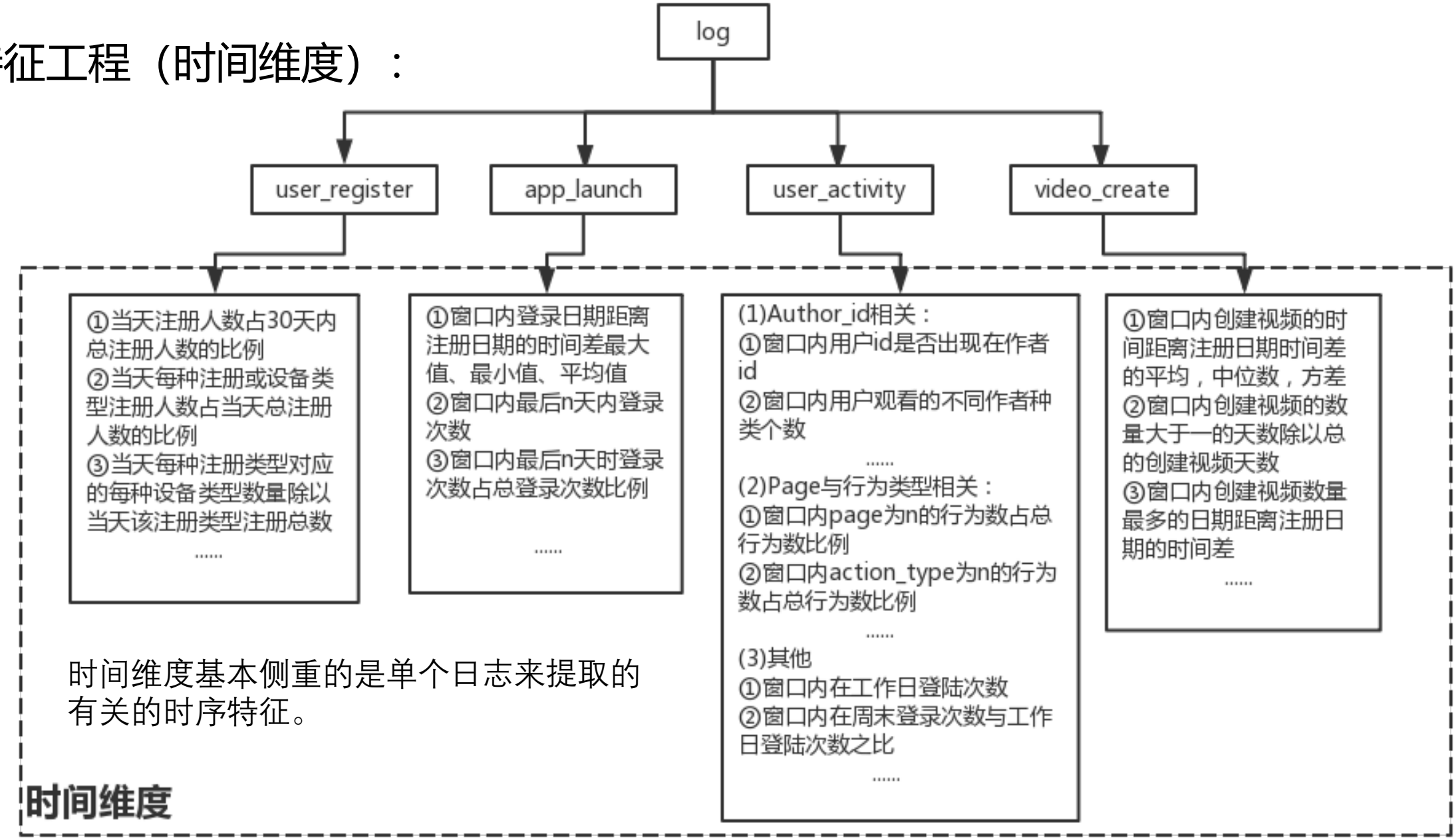


异常用户和正常用户分开的**优点**：

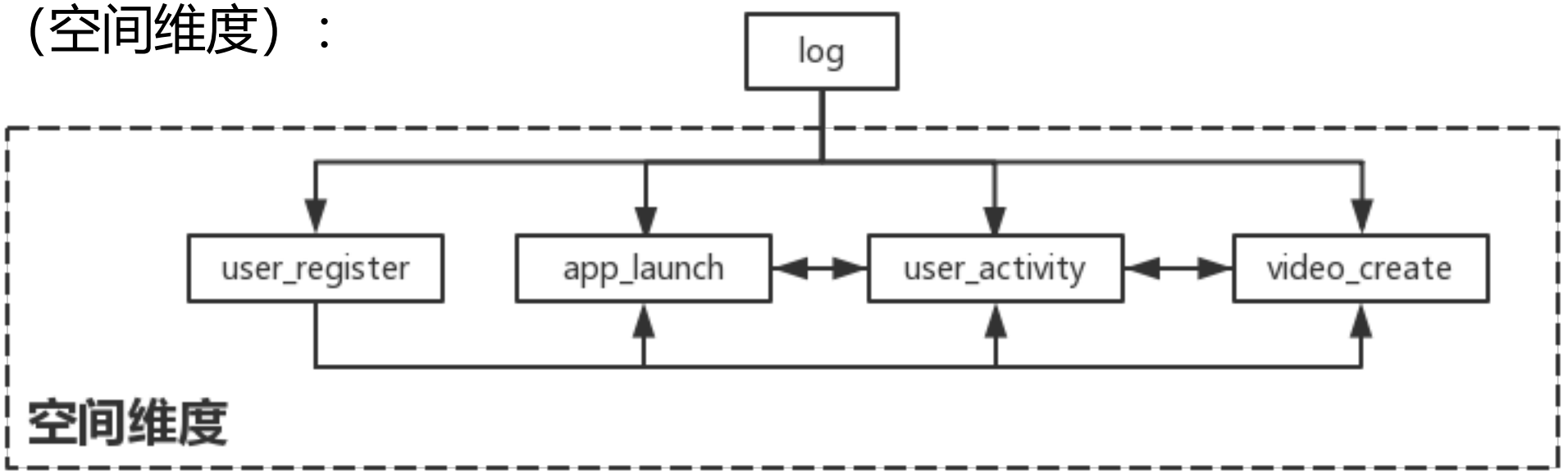
（1）这样预测异常用户更加准确

（2）剔除之后，减少训练集噪声，增加模型预测准确率

特征工程（时间维度）：


log
user_register
app_launch
user_activity
video_create
①当天注册人数占30天内总注册人数的比例
②当天每种注册或设备类型注册人数占当天总注册人数的比例
③当天每种注册类型对应的每种设备类型数量除以当天该注册类型注册总数
.....
①窗口内登录日期距离注册日期的时间差最大值、最小值、平均值
②窗口内最后n天内登录次数
③窗口内最后n天时登录次数占总登录次数比例
.....
(1)Author_id相关：
①窗口内用户id是否出现在作者id
②窗口内用户观看的不同作者种类个数
.....
(2)Page与行为类型相关：
①窗口内page为n的行为数占总行为数比例
②窗口内action_type为n的行为数占总行为数比例
.....
(3)其他
①窗口内在工作日登陆次数
②窗口内在周末登录次数与工作日登陆次数之比
.....
①窗口内创建视频的时间距离注册日期时间差的平均，中位数，方差
②窗口内创建视频的数量大于一的天数除以总的创建视频天数
③窗口内创建视频数量最多的日期距离注册日期的时间差
.....
时间维度基本侧重的是单个日志来提取的有关的时序特征。
时间维度

特征工程（空间维度）：



交叉特征列举：

1.窗口起始日期到窗口结束日期前n天创建视频次数或登录次数与行为次数做运算

2.注册日期至窗口结束日期创建视频次数或登录次数与行为次数做运算

3.距离待预测日期前n天创建视频次数或登录次数与行为次数做运算

4.距离待预测日期前n天创建视频次数或登录次数平均值与行为次数平均值做运算

5.距离待预测日期前n天创建视频次数或登录次数方差与行为次数方差做运算

……

特征工程（创新型特征）：

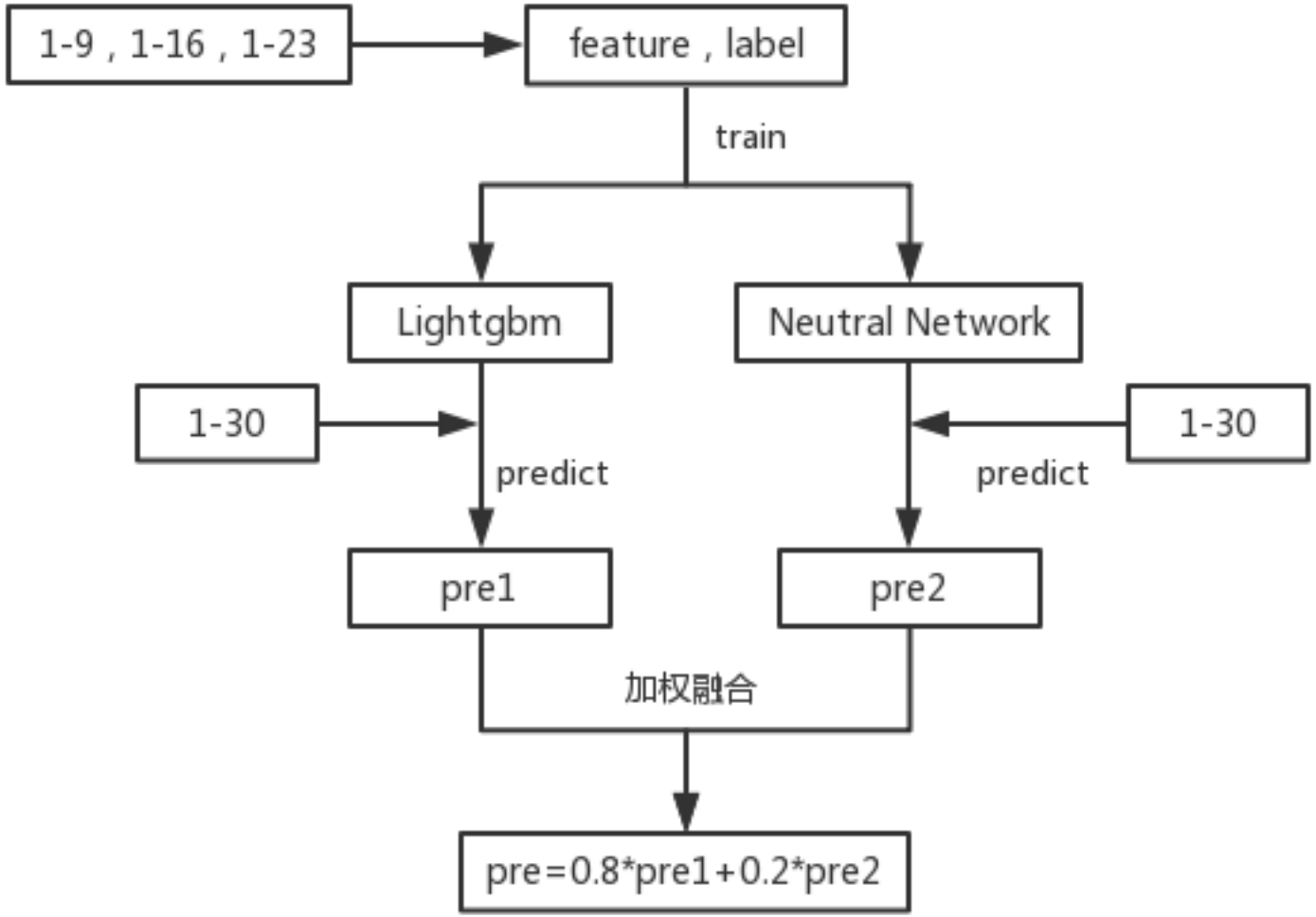
TF-IDF（term frequency–inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。

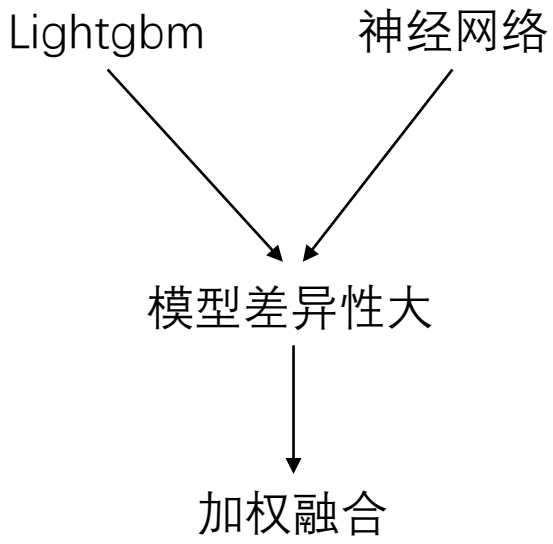
Author id	与该作者有交互记录的用户数（千）	IDF	TF-IDF
117366	62.3	0.603	0.0121
179845	0.484	2.713	0.0543
289303	0.973	2.410	0.0482

合理性：这个特征就相当于对每一个用户在窗口内进行了编码，表征着这个用户与author_id的亲密程度，用于区分特定的用户群体。

例如：一个人气高的作者一般是经常活跃的，用户与该作者交互次数越多，此用户对应该作者的tf-idf的值也越大，那么该用户活跃的概率也越大。

模型结构：


1-9 , 1-16 , 1-23
feature , label
train
Lightgbm
Neutral Network
1-30
predict
1-30
predict
pre1
pre2
加权融合
pre=0.8*pre1+0.2*pre2


Lightgbm
神经网络
模型差异性大
加权融合

不足与总结：

NN模型的改进与提高

Lightgbm单模型线下的auc比nn单模型高两个千分点，因此nn模型的进一步提升的空间较大。

01

02

融合方法的进一步尝试rank算法

我们线上单模型的分数较高，但是融合提升的分数只有不到一个万分点，可以进一步尝试对提高auc指标很有帮助的rank算法。

总结：

对于用户是否活跃的问题更应该关注具体的业务流程，结合用户使用快手app的实际场景，从而对用户的特性有更深的了解以及用户群有更深更细的划分，使得特征的构造更加合理与有效。


清華大學
Tsinghua University
快手
谢谢聆听