快手活跃用户预测
Legend94rz 队解决方案

- 目录

- 赛题解读

赛题描述

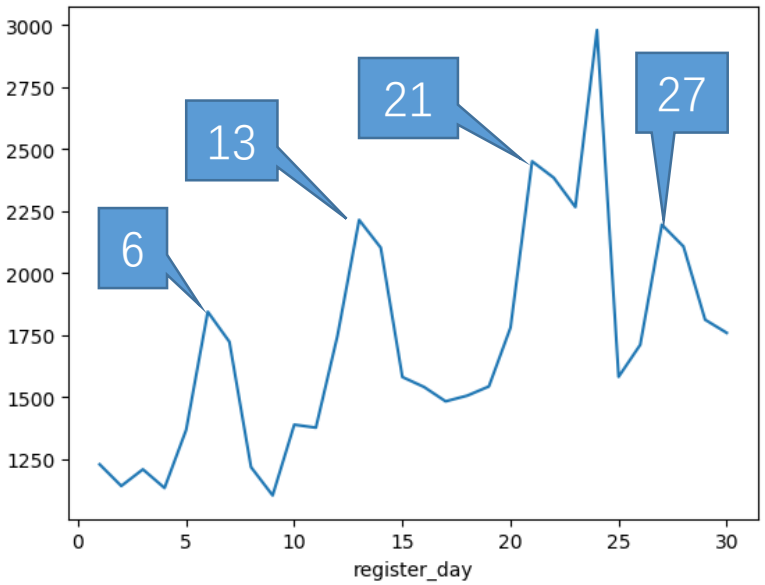
给定用户的基本信息及30天的各种行为记录，预测该用户在接下来的7天内是否有任何活动。

基本思路

转化为时间序列型的二分类问题。

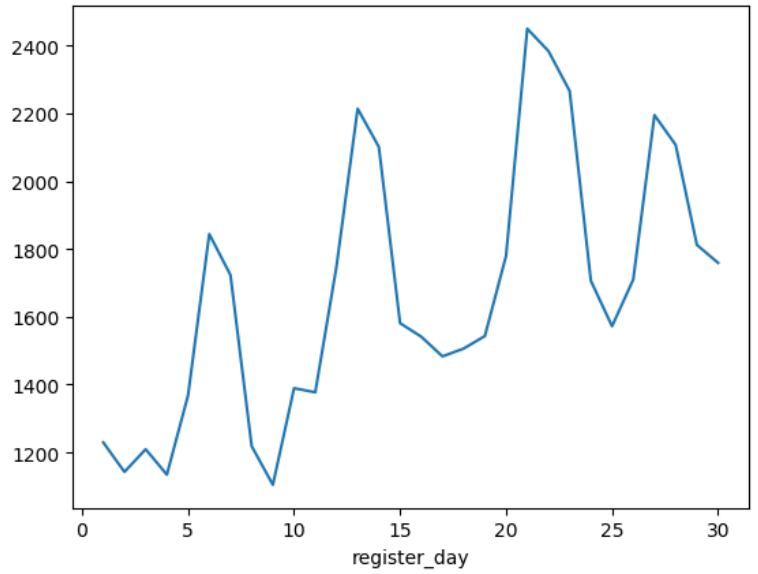
EDA

日注册量



删掉注册类型为3 且 设备类型为1的用户

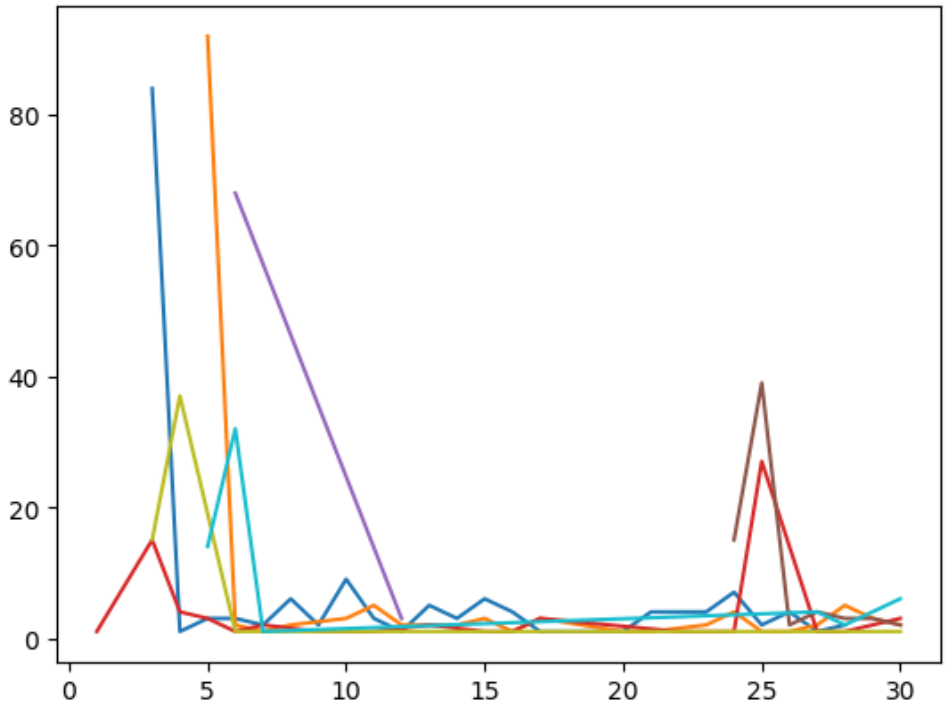
日注册量



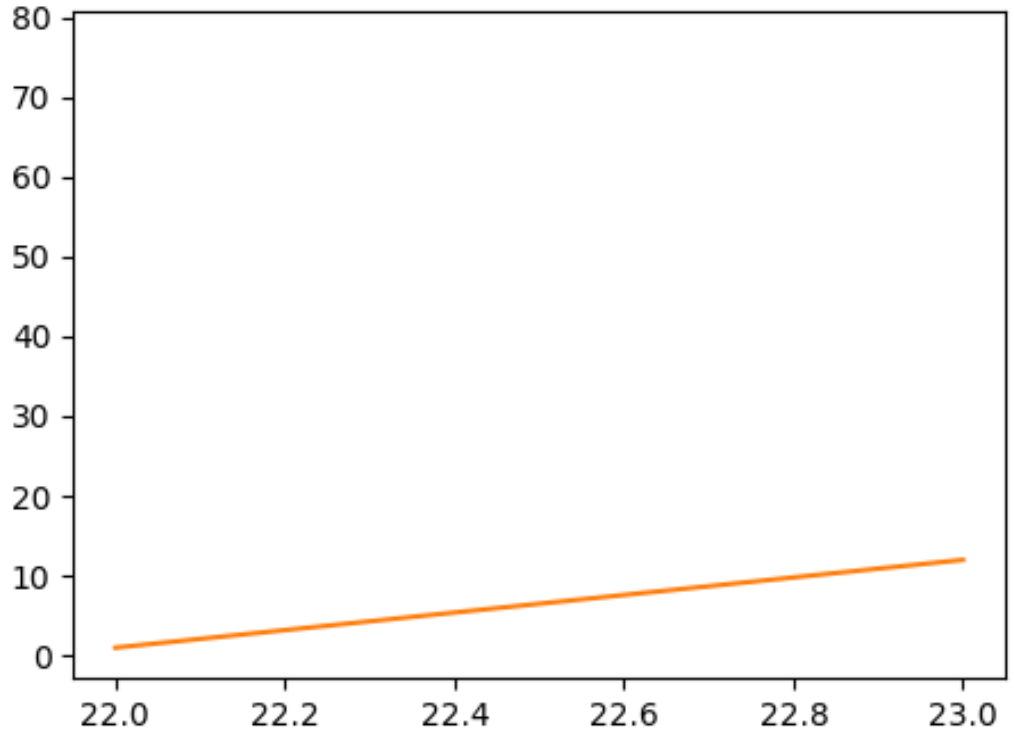
6,7
13,14
21,22,23
27,28
具有明显周期性。
推测为节假日/周末，这里可以构造出一批特征来。

EDA

用户A

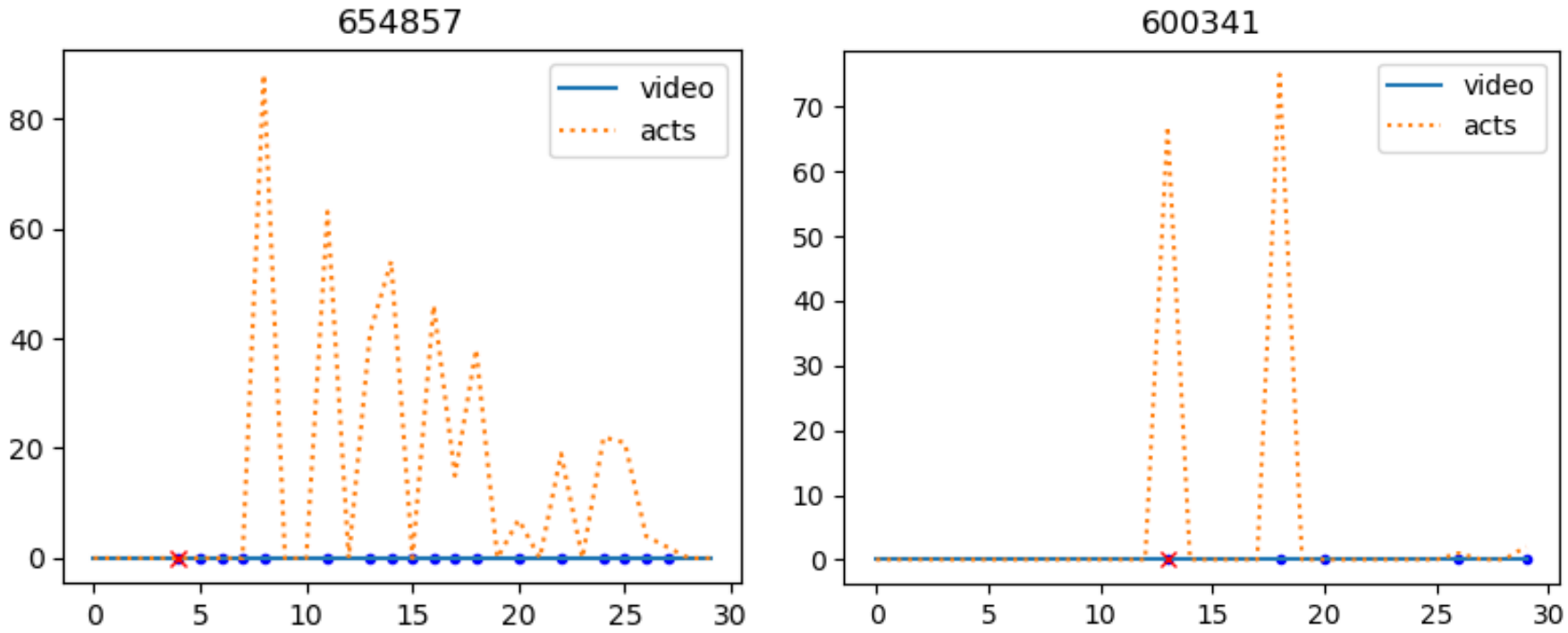


用户B



用户每天的与其他作者交互的数量分布

EDA



用户30天内活动记录

- 模型设计

针对时间序列问题，常用滑窗法来构造线下训练集与线上测试集：

训练集构造

测试集构造

记为[-6, 1, 8]

线下采用5折交叉验证，评估指标选择logloss与auc。如果这两个指标的5折平均都上升/下降，线上一般也上升/下降。

- 模型设计

采用LGB模型。用到的全部特征有：

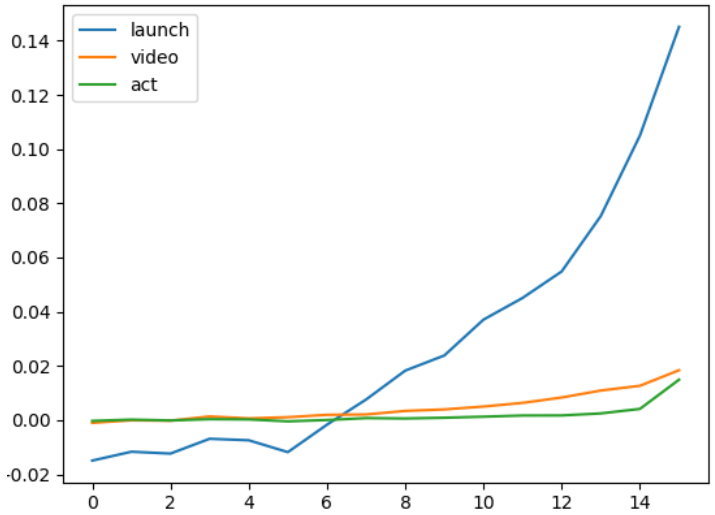
```
feature_filter = 'register_type|device_type|register_day_off|device_count'\
                 +'|[AVL]ori1[2345]|[AVL]W?sum|[AVL]W?avr|[AVL]mx0|[AVL]mx1|[AVL]dist[01]|[AV]rng|[AVL]inHoliday|[AVL]ratInHoliday'\
                 +'|[AV]max|[AV]min|[AVL]median|[AVL]ske|[AVL]kur|[AVL]std|[AVL]o(sum|avr|std)|Ldiff2'\
                 +'|Aac0[0-5]|Aacr0[0-5]|Atotbeac'\
                 +'|Apg0[0-4]|Apgr0[0-4]'\
                 +'|Amxaut1[2345]|Adbeact'\
                 +'|Arctavr|Arctsum|Arctday|Arctcnt'\
                 +'|Alngsum|Alngcnt|Aratlngcnt|Aestibeact|yhat'\
                 +'|mean_active_launch_ratio|mean_launch_gap'
```

register_type: 注册类型；
device_type：设备类型；
device_count：30天的设备统计数量；
LWsum：带权重的登录次数之和；
AinHoliday：在节假日内的行为数量；
AratInHoliday：在节假日内的行为数量占比；

Apg0[0-4]: 分页面类型，各行为的数量；
Alngsum：用户与关注时间最久的作者交互的总次数；
Aratlngcnt：用户与关注时间最久的作者交互的天数占比；
mean_active_ratio：登录之后的平均行为数量；
mean_launch_gap：登录的平均间隔；

- 模型设计

权重的设计：逻辑回归模型，给定三个日志，每个日志16维特征，共48维学得。权重可视化如图所示：

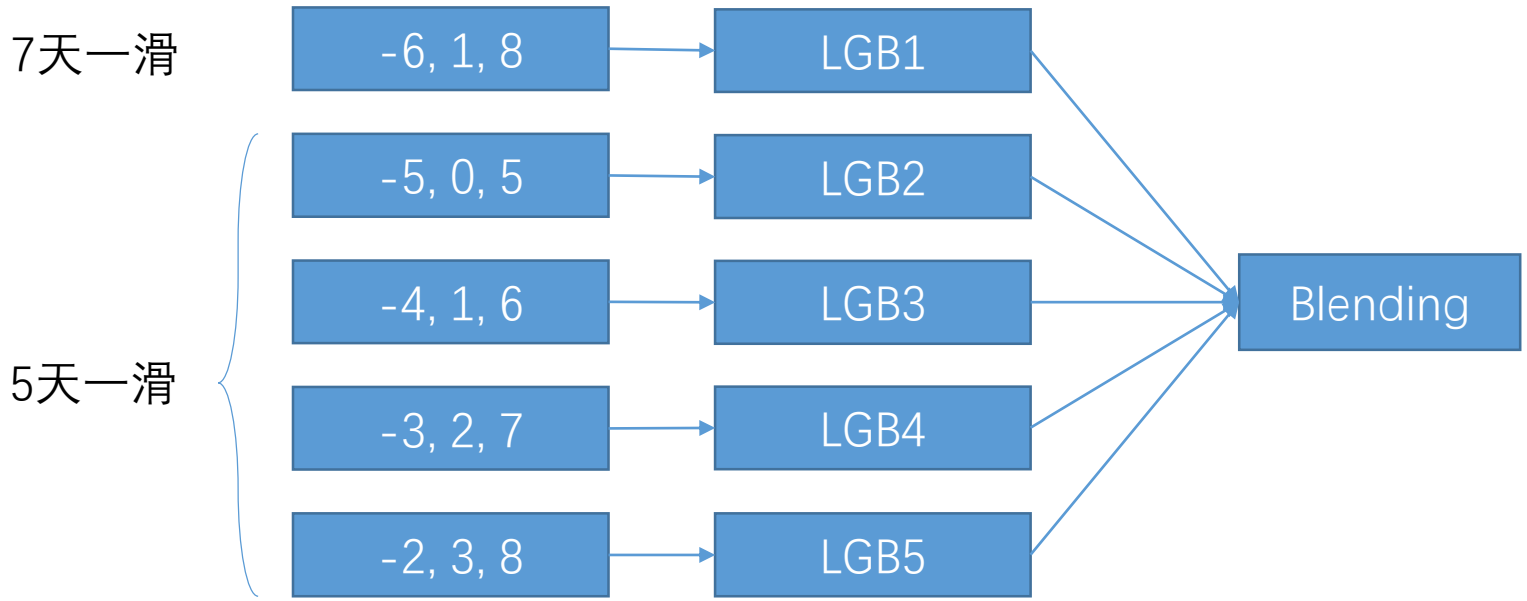


特征选择：我们的特征比较少，并且由于采用正则表达式，所以手工加减特征更加方便。删掉一些分裂次数少的特征往往可以带来分数提高。

B榜单模成绩：0.91263513

- 模型融合

本次比赛在划分数据上有很大灵活性，最充分利用数据的方法是一天一滑，但是这样会导致严重的过拟合，我们经过线下的测试，发现5天一滑既可以构造出大量的数据，同时也不会像1天一滑那样有严重的过拟合。因此我们构造了多组5天一滑的训练集。



这样5个LGB融合可得到0.91309408的B榜成绩，比7天一滑单模提高约4个万分点。

- 模型融合

我们还尝试了把LGB换成CatBoost模型，类似的方式再构造4天一滑的特征，共14个模型，加权融合到结果里，有微弱提升，最终线上B榜0.91310950，比用5个模型高约1个十万位。

Trick

由EDA的分析，我们发现注册类型为3 且 设备类型为1的用户，绝大部分都在24/25日注册，在注册当天仅有一次登录而之外没有任何活动，所以我们怀疑这些用户属于异常注册，且今后不会活跃。把这部分用户的活跃概率置0，可以提高约1个万分点。

- 总结与反思

  - 做得比较好的

1. 生成到保存特征文件采用了一种类似二级缓存的方式，速度快且占用磁盘少。一级缓存一般不需要反复重新生成，二级缓存只需要30分钟左右即可。
2. 线上几乎线下同增同减。

  - 存在不足/未实现的

1. 尝试更多的滑窗划分方式。如长度为9天的窗口或滑动距离为6天。
2. 其他更好的模型融合方法。如Stacking。
3. 其他模型，如RNN，XGB。RNN我们尝试过，即把用户16天的活动记录作为输入，先经过CNN提取特征，然后接RNN，效果不好，期待其他选手分享思路。XGB则是由于十分慢，复赛放弃了这种模型。
4. 用其他的提升树类型（LGB参数），如dart。这个线下一部分训练子集可以提高约一个万分点，但是很遗憾特别慢，只跑了一部分，没有提交。


清華大學
Tsinghua University
快手
谢谢聆听