2018中国高校计算机大赛——大数据挑战赛

队伍名称：搬砖

清华大学 Tsinghua University | 快手

1 赛题回顾

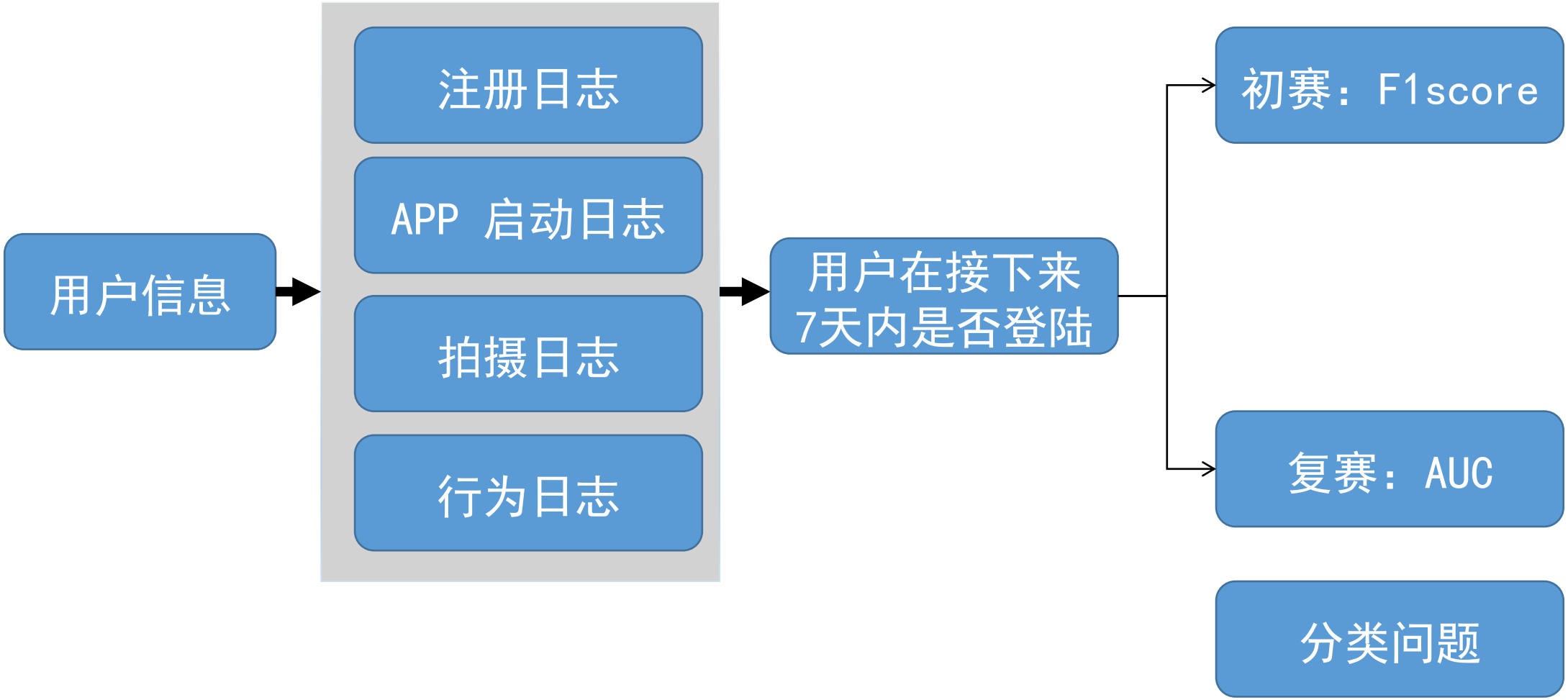
2 解题思路与算法

3 结果与分析

4 收获与思考

赛题回顾

赛题回顾


用户信息
注册日志
APP 启动日志
拍摄日志
行为日志
用户在接下来
7天内是否登陆
初赛：F1score
复赛：AUC
分类问题

解题思路与算法

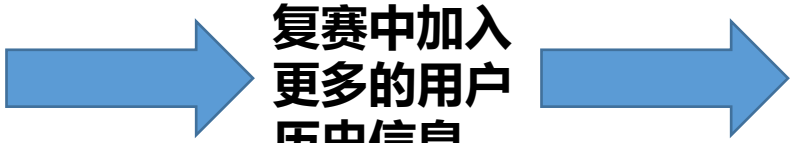
数据分析

初赛
训练集构造:
用户信息1-16 label 17-24
用户信息8-23 label 24-30

测试集构造:
用户信息15-30

正负样本比例：1：1

评估指标：F1



复赛中加入
更多的用户
历史信息

复赛
训练集构造:
用户信息1-9 label 10-16
用户信息1-16 label 17-24
用户信息1-23 label 24-30

测试集构造:
用户信息1-30

正负比例：1：1

训练集：811728
测试集：572132

评估指标：AUC

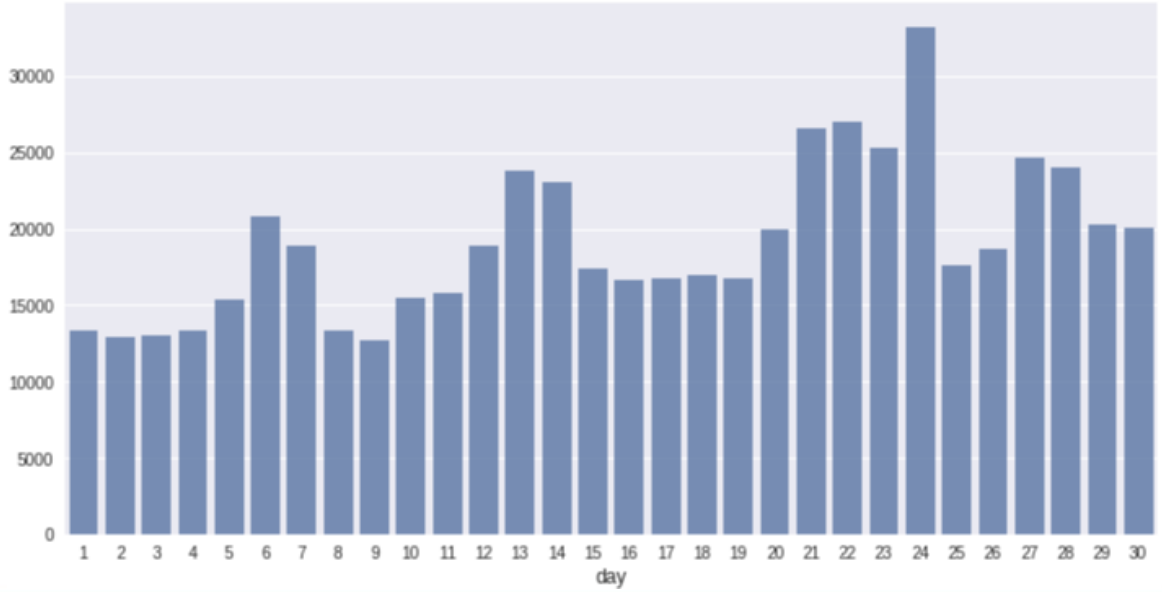
特征构造

注册日志：

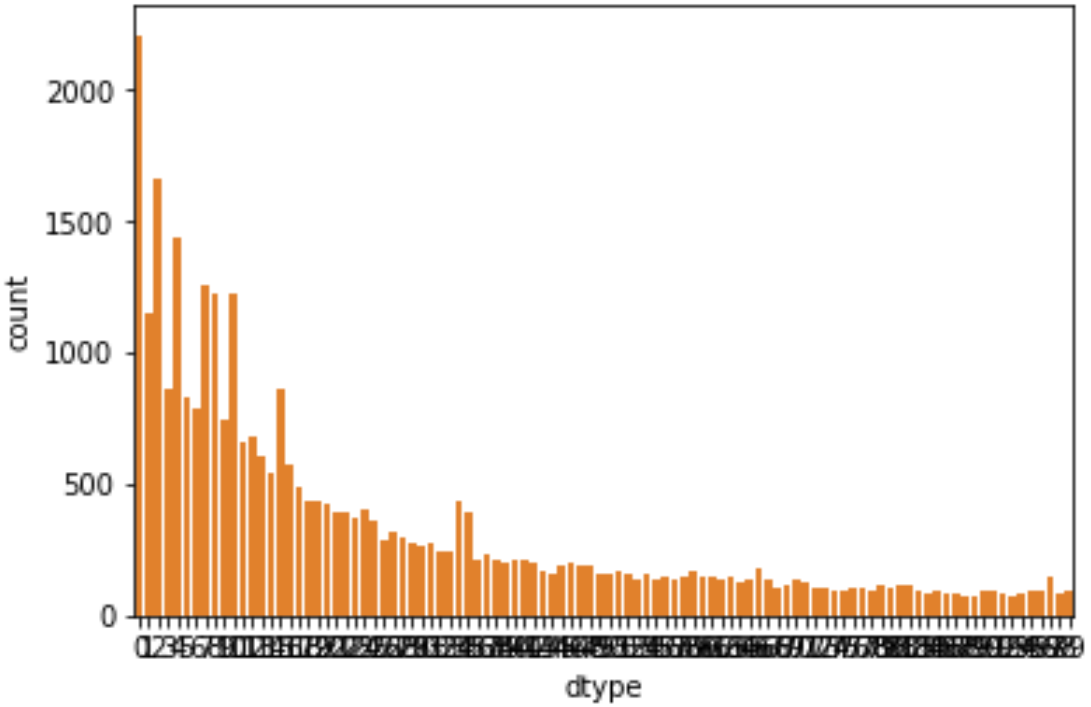
register_day	日期
register_type	来源渠道
device type	设备类型

特征：

设备类型计数



注册日期统计

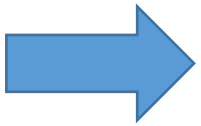


设备类型统计

特征构造

APP 启动日志：
day　　日期

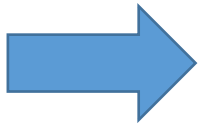
特征：
最大连续登录
最近登陆
总登陆次数
近15天的登陆日期
最近1天是否登陆
最近3天是否登陆
最近一周是否登陆



用户对软件的粘性

拍摄日志：
day　　拍摄日期

特征：
最大连续拍摄
一天最多拍摄次数
拍摄的总天数
最近拍摄
总拍摄次数
近15天拍摄日期
每天平均拍摄
最近1天是否拍摄
最近3天是否拍摄
最近一周是否拍摄



用户对拍摄视频的偏好

特征构造

行为日志：
day　　日期
Page　　行为发生的页面
video_id
author_id
action_type　　用户行为类型

特征：
最近行为时间
用户各种action_type分别计数
用户各种action_type的比例
用户在各种page上的action分别计数
用户在各种page上的action的比例
用户总行为数
用户每天平均行为数
用户有action的总天数
用户近15天每天操作数
用户拍摄视频的总观看人数
用户拍摄视频的平均观看人数
用户观看最多的作者计数
用户观看视频的平均热度
用户观看作者的平均热度

对用户近期行为的刻画，
对不同视频的关注度和兴趣程度

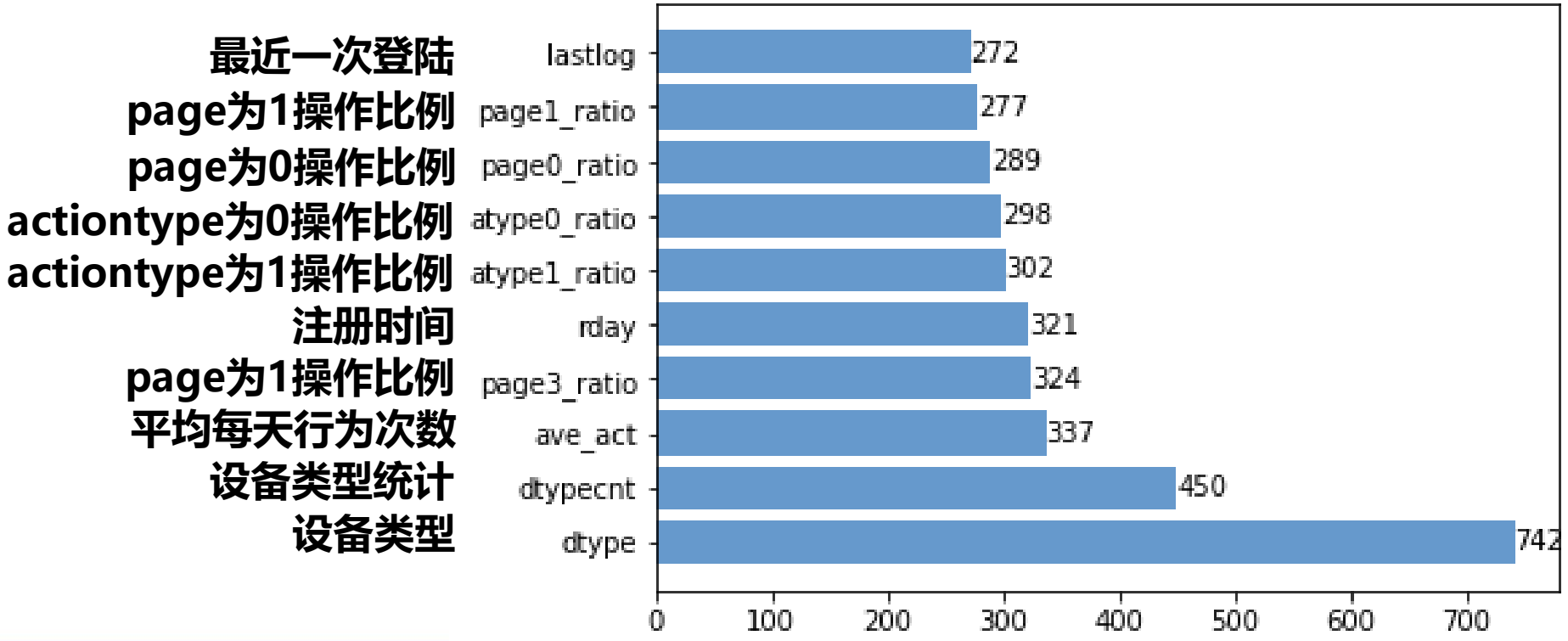
特征构造

- **基本特征：**
  - 单值特征如用户的来源渠道，用户的设备类型等
  - 多值特征如用户最近15天的登陆时间
- **计数特征：**
  - 如近期总登陆次数，最近总拍摄次数等
- **其他特征：**
  - 用户最后一次登陆时间，用户最后一次拍摄时间
  - 用户平均每天拍摄次数
  - 前30强特交叉相除后，取12个重要性较高的特征
- **总特征：106个**

模型选择

1.数据量不大
2.大部分为连续型特征

- LightGBM分类
- XGBoost分类
- LSTM
- LightGBM回归



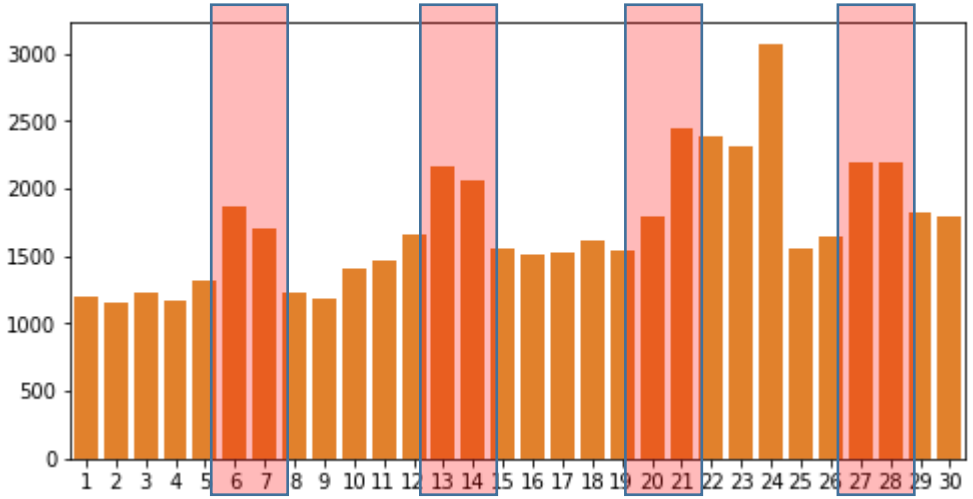
结果与分析

关键点

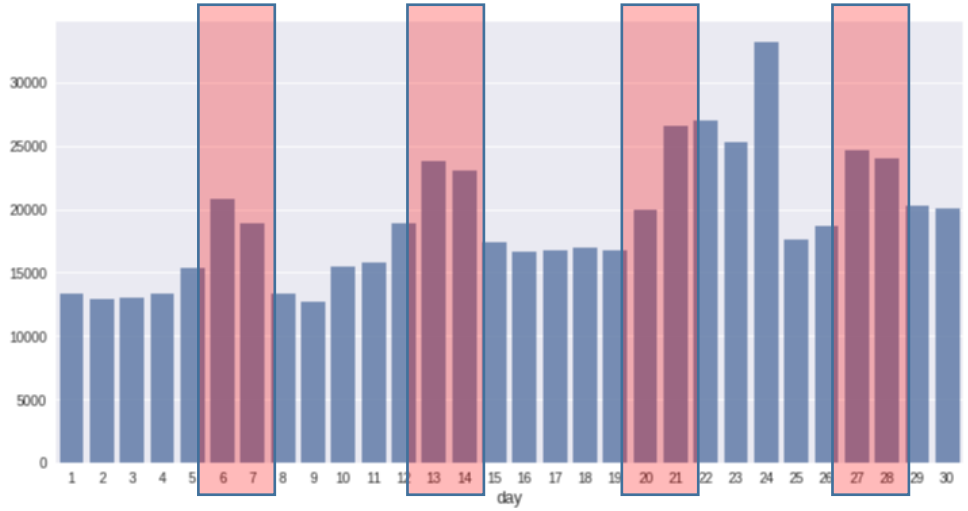
	线下	线上
device type处理	提升显著	提升显著
五折CV融合		提升显著且能保证线上线下同步
添加滑动窗口	提升显著	提升显著
加入更多历史信息	提升显著	效果不明显
回归模型	低于分类模型	融合后提升明显
强特交叉相除	提升显著	效果不明显
video_id相关特征	提升显著	反效果
author_id相关特征	提升显著	反效果

Video和author特征过拟合原因分析
可能数据集中存在刷单用户，一天中大量观看同一个视频

结果分析



每天注册人数（初赛）



每天注册人数（复赛）

从大部分数据分析：

初赛与复赛数据十分相似
但是复赛无法上传初赛数据进行训练

6,7,13,14,20,21,27,28几天为周末
但是20,22,23,24几天注册人数反常

结果处理

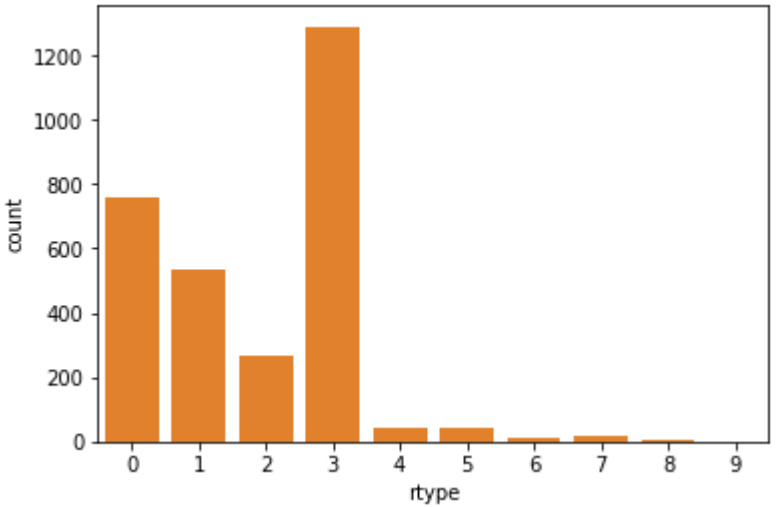
分析20,22,23,24几天注册人数反常原因

24号注册的人中设备类型为3明显出现异常

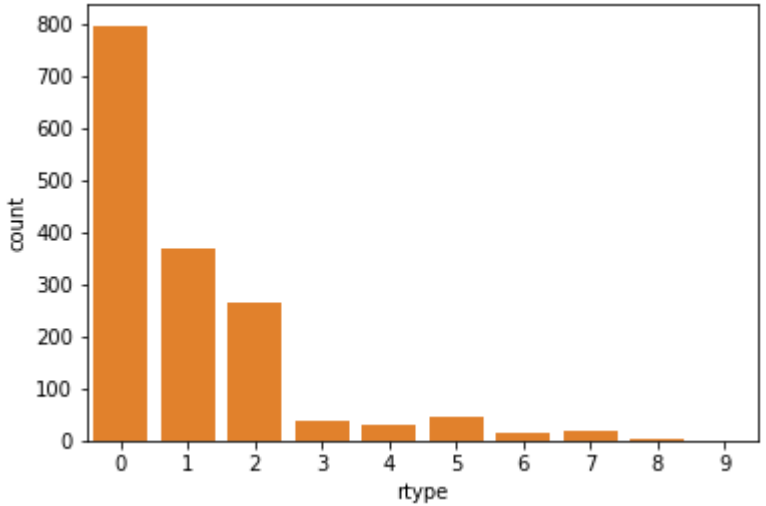
在24号注册的人中有1万多的人没有再次登陆过

直接将这些用户在结果中赋值为0

成绩提升4个十万分点



24号注册的设备类型



25号注册的设备类型


清華大學
Tsinghua University
快手
谢谢聆听