QG 工作室暑期实习生两日结

姓名： 许继元　　组别： 数据挖掘　年级： 19 级　　日期： 2020.08.15

生活小记	8.14 日，上午在训练模型的时候发现出租车流量数据有点问题，有一个经纬点在后期一直重复，导致模型拟合度很差，这显然是出 bug 了，从早上找到中午十一点多才找到 bug，原来是在分析统计流量的代码里出现的 bug，其同时也影响了流量数据，此时移动组也反馈有一个点持续出现，修复完 bug 之后，重新生成了数据，模型也就正常了，对模型进行训练和评估之后，发现模型的好坏程度为如下顺序：随机森林回归、XGBoost、LightGBM、GBDT。于是选取了随机森林回归作为流量预测模型，预测出了一周的出租车流量数据。下午练了半个多小时的吉他，接着寻找项目相关论文资料，用于后续挖掘出新颖功能模块；晚上和其他组开了个会议沟通需求区域功能模块的细节，重新理顺了思路和细节处理，并对异常情况模块的功能进行了修改和拓展，最后我也给出了自己的拓展思路，画了个简陋的设计图供设计组参考。 8.15 日，上午准备做行政区流量统计的模块，过程中遇到了困难，由于增加了行政区判断的步骤以及坐标体系转换步骤，计算量急剧上升，程序跑起来就超级慢。下午重构了一下代码结构，优化了步骤之后，虽然降低了复杂度，但是还是需要花费很多时间，于是只能减少抽样量，后期可以通过数学手段来拟合出原始的流量。晚上前端组的师兄开了技术交流会，了解到了一些开发框架的知识，师兄的故事也很有趣。康乐活动就没有打游戏了，练了一下吉他，接下来继续看看还有什么方式处理上午的那个困难，最后采用多进程的手段，同时运行七个进程来统计，测试了一下，大约还是需要 12 小时的运行时间，于是我准备睡觉前让程序开始运行，当然运行结果也要通过数学手段来处理，这个明天再做。睡觉前，我把笔记本电脑的后座垫高，使其方便散热，然后运行程序，等待明天中午的结果。
学习 开发 比赛 概要	学习上，学习了 sklearn 的 MLP、随机森林回归、GBDT 模块的参数及其使用以及 XGBoost、LightGBM 等框架的参数及其调用。了解到 Python 的多进程之间是默认不共享变量的。 开发上，修复了分析统计流量数据中有一个点持续出现的 bug。选取了随机森林回归作为流量预测模型，预测出了一周的出租车流量数据。给出了自己的拓展思路，画出简易的异常情况模块设计图。优化了行政区流量统计的模块架构，降低了复杂度。采用多进程手段进行行政区流量统计。

感想收获	这两天数据挖掘组陆续完成了自己负责的功能模块，接下来就是其他各组的进度跟进以及系统完善了。同时我们也在阅读一些项目相关的论文，看看是否能够添加一些新颖的、有价值的功能模块。
存在问题（备注）	