QG 工作室周记

姓名： 许继元　　组别： 数据挖掘　年级： 19 级　　周次： 第 5 周

生活随记	这周的主要任务是最终考核项目。 8.9 日晚上发布了最终考核项目的要求，开完会议后，数据挖掘组先组内开了会议，接着各组人员集体开一次会议，目的是吸取上次的经验，把需求分析和沟通好。 8.10 日，各组人员开会对最终考核项目再次进行分析和讨论，最后形成了一个初步的可行方案。 在最终考核之前我就想过担任项目总负责人，因此选择项目负责人的时候我毛遂自荐，最后也成为最终考核项目的总负责人以及数据挖掘组的负责人。希望通过这次经历能让我不仅在数据挖掘方面有大的进步，而且在项目管理、小组沟通、接口设计以及交互逻辑设计等方面有大的提升。 在数据挖掘组的周会，我对本阶段的学习成果进行了分享和展示，接着师兄询问了一些最终考核的问题并给出了解决方案，并针对下一个阶段的学习给出分析与建议，解决了我们的疑惑。知道我是项目总负责人，正婷师姐也给予了我鼓励，这让我有了更大的信心做好总负责人的工作。 在一次会议中我提出了合并历史行驶路径和异常情况模块，以实现数据的合理利用，这一决定减轻了大数据量的负担，同时也合理利用了数据价值。除此以外，我还学习了通过 Matplotlib 的 Path 模块来实现判断经纬点是否在地图区域内，且根据边界经纬点数据编写了一个判断经纬点是否处于行政区内的通用模块。感觉最近电脑内存不够用了，于是在网上买了块三星内存条。 完成了后台组分配的一些需求，获取广州市行政区内两个采样点的 geohash 码、获取广州市每个行政区的 5000 个经纬点。 对于流量分析模块，采用降低经纬度的精度来形成以所选取经纬点为期望的高斯分布这一方式来进行流量分析，数据用 json 格式进行传输。 GPS 采用 WGS-84 原始坐标体系。一般用国际标准的 GPS 记录仪记录下来的坐标，都是 GPS 的坐标。但在中国，任何一个地图产品都不允许使用 GPS 坐标，据说是为了保密。因此我们只能转化为 GCJ-02 坐标体系，其又称“火星坐标”。在中国，必须至少使用 GCJ-02 的坐标体系，于是找到了算法来进行坐标体系转换。 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。 抽出时间寻找了项目相关论文资料，用于后续挖掘出新颖功能模块。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。 项目开发期间抽出时间练习了吉他曲子的几个小节，希望暑假前能练完这首曲子。

学习开发比赛情况	学习上，学习了通过 Matplotlib 的 Path 模块来实现判断经纬点是否在地图区域内。学习了 pygeohash 模块对经纬点进行 geohash 编码。学会了 WGS-84 坐标体系转化为 GCJ-02 坐标体系的算法。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。 开发上，编写了判断经纬点是否处于行政区内的通用模块。解决了解决流量数据的经纬点偏移问题。完成了出租车流量数据的分析统计。选取了五个算法模型进行训练评估，最后选出最合适的流量预测模型。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。
一周总结	这周主要就是火力全开地完成最终考核的各大功能模块，由于吸取了中期考核的经验，这次的考核与其他各组及其负责人进行了多次会议，做了比较详细的沟通，所以各小组都有明确的目标和构想，一切开发进度都在计划中。
存在问题未来规划	存在问题： ①绿色计算大赛的赋能赛还未完成，要尽快。 ②每天应安排出时间多学习其他方面的知识。 未来规划： ①完成绿色计算大赛赋能赛。 ②阅读最终考核项目相关论文，增加新颖的功能。 ③完成数学建模国赛练习题。
导师评价	