[2018年高校校园大数据竞赛](https://bigdata.bupt.edu.cn/about.html)

bupt\_try团队解决方案代码说明文档

一、代码文件及运行说明

1. 运行环境

Ubuntu16.04、16G内存、安装jupyter notebook

2. 运行依赖

python3.6、numpy、pandas、sklearn、lightgbm、scipy、datetime、matplotlib、contextlib等

3. 代码文件说明

0\_analysisi.ipynb: 为可视化分析的部分，分析规则与模型预测结果和9,10,11月的趋势，以此来指导模型融合的部分

weather.ipynb: 从网站 http://tianqi.2345.com/ 上爬取的天气信息，保存为天气编码格式和天气未编码的格式

1\_process.ipynb: 为预处理部分，只用9,10,11月数据，根据3σ原则，剔除部分异常值，并且去掉十一假期等异常日期的数据

2\_model.ipynb: 模型预测部分，使用lgb，rf，gbdt三个模型来预测12月的人流量

3\_rule.ipynb: 规则预测部分，仅使用11月数据，添加日期衰减系数，和天气修正系数，得到12月的预测结果

4\_blend.ipynb: 融合规则和模型的预测结果，根据0\_analysisi.ipynb的可视化结果，确定最终的融合方案

4. 执行步骤

要得到预测结果，运行步骤为1-4，依次运行以下文件:

1\_process.ipynb 2\_model.ipynb 3\_rule.ipynb 4\_blend.ipynb

二、代码内容精简介绍

1. **数据预处理**

1. 按时间和位置进行分组，求出每个loc\_id在每个时间点上的phone\_id的数量
2. 从时间戳time\_stamp中提取出month、day、hour、dayofweek信息
3. 求出星期小时的均值μ和标准差σ，用（μ-2.4σ, μ+2.4σ）原则筛除样本异常值约0.5%
4. 取9月4号-11月30号的样本作为训练集，去除十一假期

2. **模型**

特征：

**位置特征**：

loc\_id：地点位置，有33个地点

loc\_type：对33个地点按照食堂、教学楼和公寓进行简单分类，提取地点类别。

**时间特征：**

month：月份特征，即每天记录所在的月份

day：天数特征，即每天记录在本月的第几天

hour：小时特征，每条记录在当天的第几个小时

day\_cut：月份分区，根据天数特征day将月份分为上、中、下三旬

hour\_cut：小时分区， 根据小时特征将一天划为三个时段

dayofweek：星期特征，即计算记录的时间是星期几

isweekday和isweekend：即将星期特征划分为两类，工作日和周末

**天气特征：**

maxTem：一天中的最高气温

minTem：一天中的最低气温

wea：当天的天气状况，晴、多云、多云~晴等

pm：当天的空气质量状况，优、良、中度污染等

rain：当天是否下雨

**历史流量特征：**

loc\_hour\_mean：该地点当前小时的历史流量均值

loc\_week\_hour\_mean：该地点当前星期-小时的历史流量均值

模型：

1. LightGBM模型
2. RandomForestRegressor模型
3. GradientBoostingRegressor模型
4. 规则模型，在历史数据中求出每个地点当前星期-小时的流量平均值作为预测值，设置衰减权重。
5. 得到各个模型对验证集的预测结果，用minimize进行多变量的最小化训练每个模型的权重，进行加权融合

3. **规则**

1. 只使用11月份数据，用3σ原则去除异常值
2. 获取每个位置不同星期小时的分布权重
3. 获取距离预测日期12月1号的天数，按指数衰减得到距离衰减权重
4. 遍历搜索最优值，乘星期小时权重和衰减权重，使与11月记录的rmse最小
5. 最终预测结果乘以天气的修正系数
6. 用历史均值作为预测值填充缺失值

4. **融合**

1. 规则预测偏小的结果保留
2. 规则预测偏大的结果与模型预测结果加权融合
3. 对所有预测结果乘以修正系数

注：代码的具体细节和实现，见ipynb内的注释