

项目文档

软件学院

数据处理人员

作者：贺思超



目录

[数据/算法文档 5](#_Toc108421495)

[1.数据库对接、搭建工作 5](#_Toc108421496)

[1.1搭建合适的数据库 5](#_Toc108421497)

[1.1.1概念模型 6](#_Toc108421498)

[1.1.2逻辑模型 6](#_Toc108421499)

[1.1.3物理模型 6](#_Toc108421500)

[1.1.4ER图 6](#_Toc108421501)

[1.2配合后端完成数据库的对接工作 5](#_Toc108421502)

[1.2.1数据库配置对接 6](#_Toc108421503)

[2.气象预测算法寻找数据 5](#_Toc108421504)

[3.气象预测算法数据预处理 5](#_Toc108421505)

[3.1计算均值填充缺少的数值 6](#_Toc108421506)

[4.气象预测算法的实现 5](#_Toc108421507)

[4.1使用GetData.py编写相关的爬虫程序爬取数据 5](#_Toc108421508)

[4.2通过Write.py进行爬取数据的分析、清洗；之后写到创建好的csv文件中进行下一步的数据预处理 5](#_Toc108421509)

[4.3通过ProcessData.py调用Write中的方法write进行数据的读取与csv文件的写入，并通过这些文件进行数据集的处理与划分。 5](#_Toc108421510)

[4.4最后模型构建在GetModel.py中，使用随机森林模型，填入处理好的数据集，进行模型的训练与保存 6](#_Toc108421511)

[5.航班延迟预测算法寻找数据 6](#_Toc108421512)

[5.1寻找到合适的数据集 6](#_Toc108421513)

[6.航班延迟预测算法数据预处理 6](#_Toc108421514)

[6.1缩减数据集，删除不是选中机场的出发机场和到达机场的行 6](#_Toc108421515)

[6.2删除取消的航班 6](#_Toc108421516)

[6.3将时间戳转化为普通时间 6](#_Toc108421517)

[6.4延迟时间计算 6](#_Toc108421518)

[6.5通过两地经纬度计算距离 6](#_Toc108421519)

[6.6删除多余的列 6](#_Toc108421520)

[6.7将计划出发日期转化为年，月，日，起飞时间的列 7](#_Toc108421521)

[6.8将出发延迟转化为分类预测对象 7](#_Toc108421522)

[6.9爬取所有地点的2015-2017年每天的气象数据 7](#_Toc108421523)

[6.10将气象数据整合到处理好的航班延误数据集中，合并为最终的完整数据集 7](#_Toc108421524)

[7.航班延迟预测算法的实现 7](#_Toc108421525)

[7.1数据分析 7](#_Toc108421526)

[7.1.1延迟程度频次统计图 7](#_Toc108421527)

[7.1.2相关性分析 7](#_Toc108421528)

[7.2使用基本的预测算法，成功实现基础预测功能 7](#_Toc108421529)

[7.2.1将字符型数据转化为字典映射编码 8](#_Toc108421530)

[7.2.2划分数据集 8](#_Toc108421531)

[7.2.3标准化处理 8](#_Toc108421532)

[7.2.4建模 8](#_Toc108421533)

[7.3对算法进行一定改进，使其更加合理，并给出对比结果 7](#_Toc108421534)

[8.可使用的参考文献 7](#_Toc108421535)

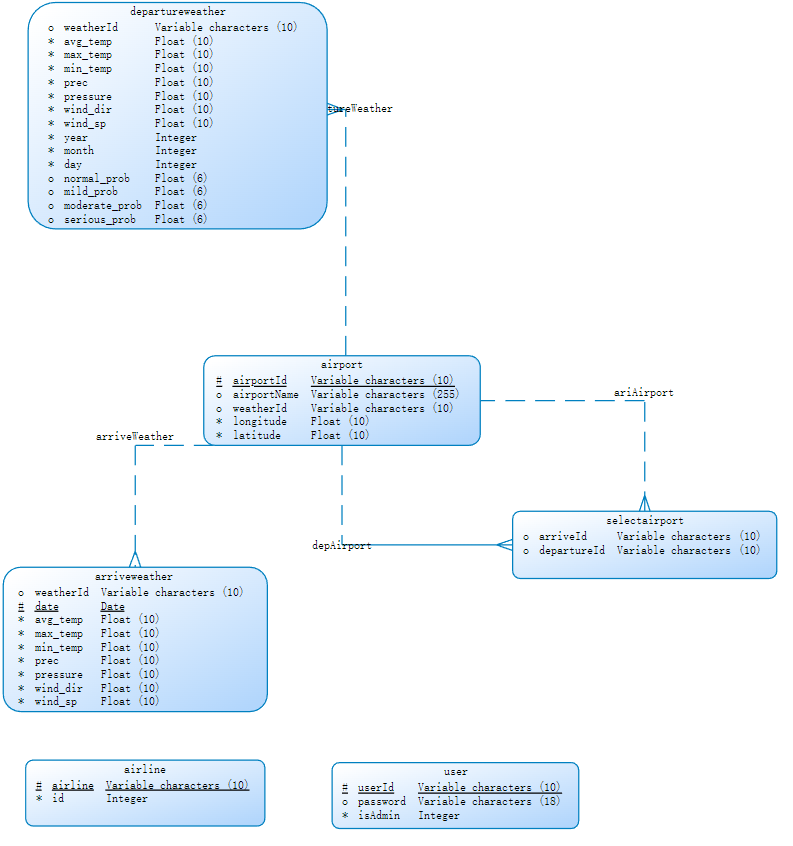
数据/算法文档

# 1.数据库对接、搭建工作

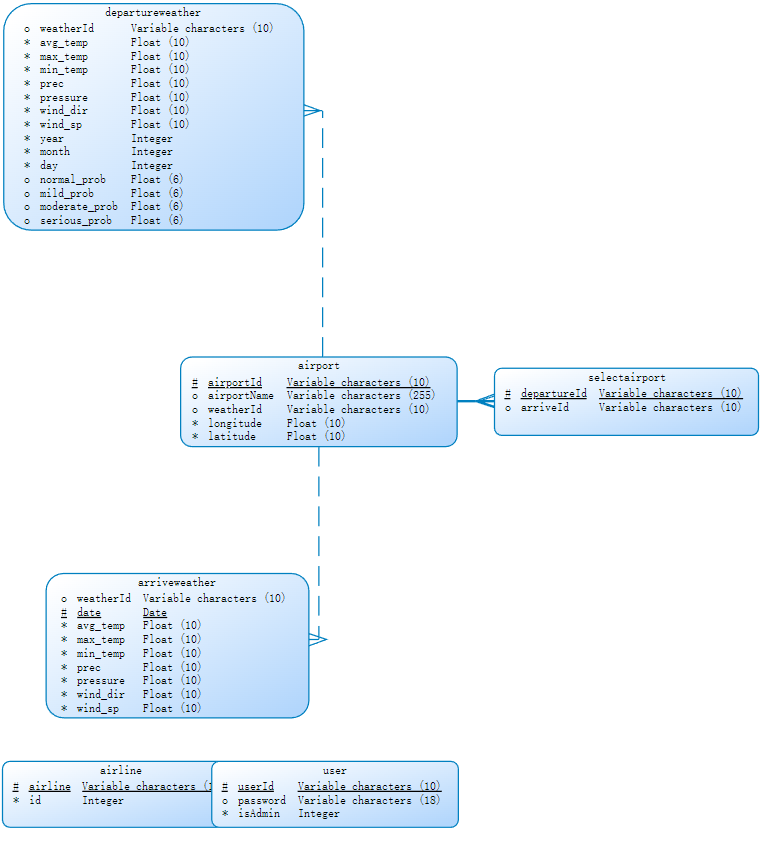
## 1.1搭建合适的数据库

使用mysql数据库，powerdesigner进行概念模型与物理模型的初步搭建

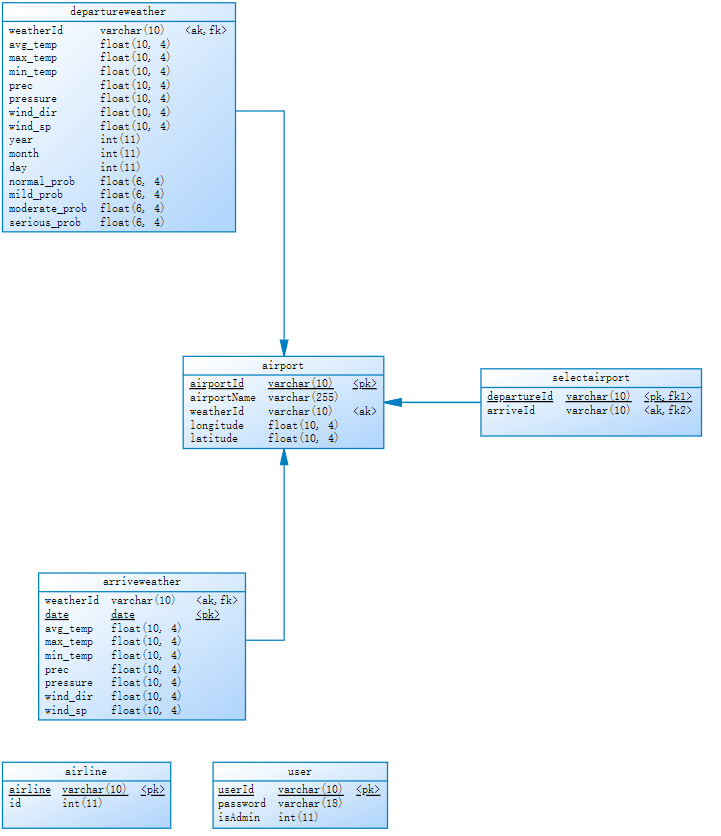
### 1.1.1概念模型



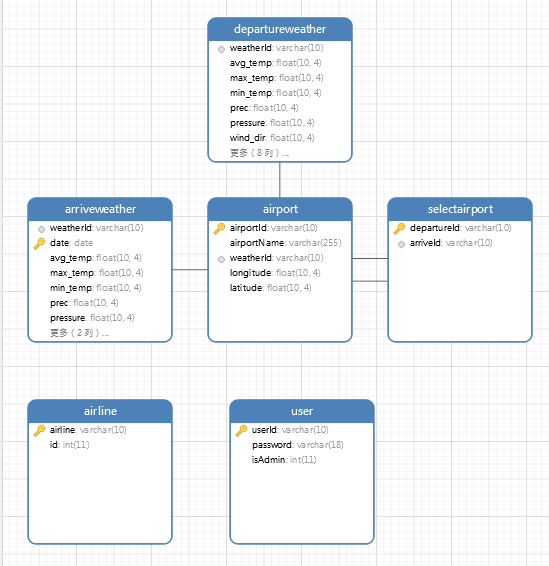
### 1.1.2逻辑模型



### 1.1.3物理模型

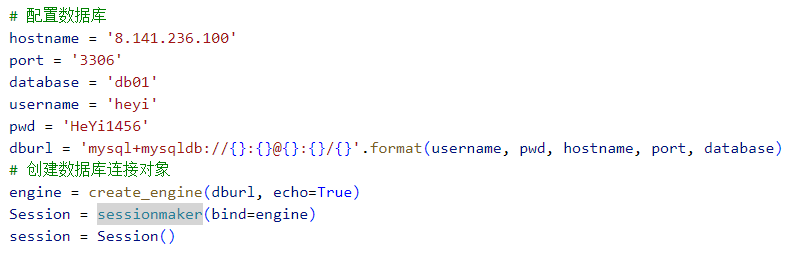


### 1.1.4ER图



## 1.2配合后端完成数据库的对接工作

### 1.2.1数据库配置对接

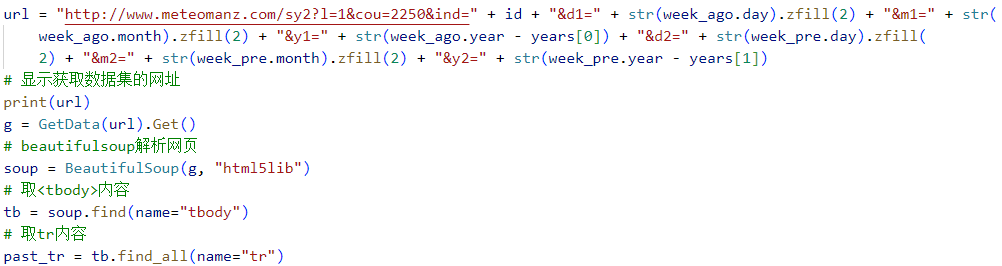


使用session便可以使用sql语句对数据库进行操作

# 2.气象预测算法寻找数据

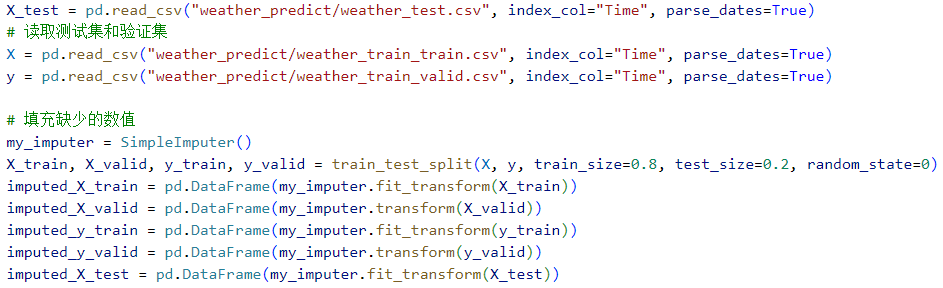
训练数据来源于[http://www.meteomanz.com/](http://www.meteomanz.com/" \t "dlt)

通过BeautifulSoup类库进行爬取



# 3.气象预测算法数据预处理

## 3.1计算均值填充缺少的数值

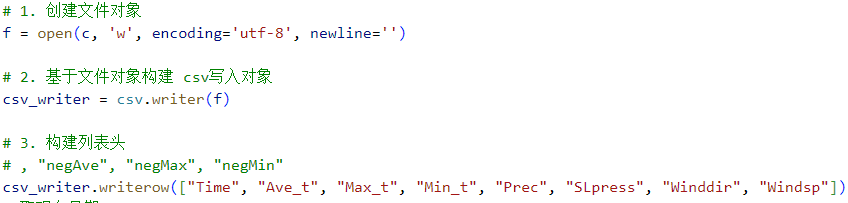


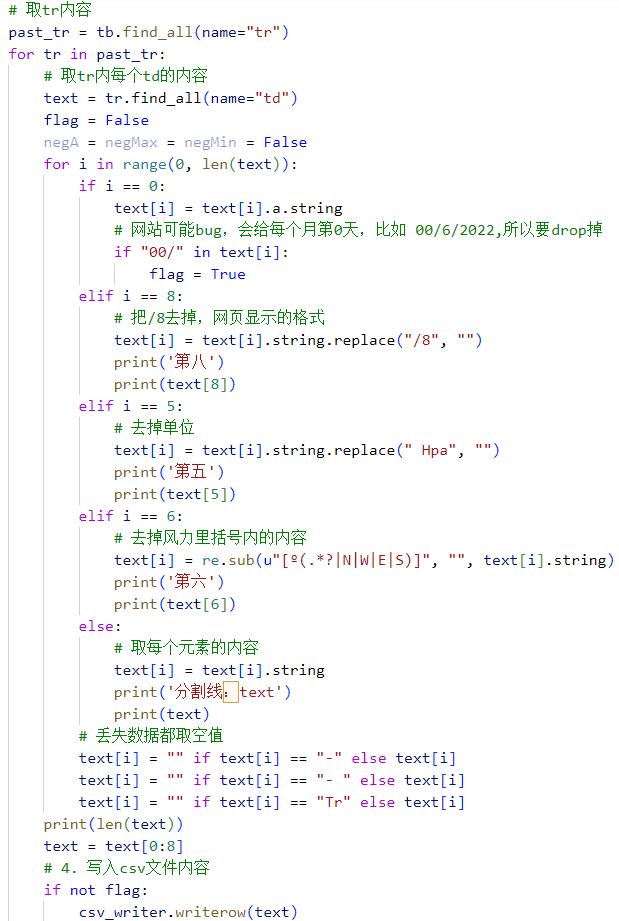
# 4.气象预测算法的实现

## 4.1使用GetData.py编写相关的爬虫程序爬取数据



## 4.2通过Write.py进行爬取数据的分析、清洗；之后写到创建好的csv文件中进行下一步的数据预处理





## 4.3通过ProcessData.py调用Write中的方法write进行数据的读取与csv文件的写入，并通过这些文件进行数据集的处理与划分。



## 4.4最后模型构建在GetModel.py中，使用随机森林模型，填入处理好的数据集，进行模型的训练与保存

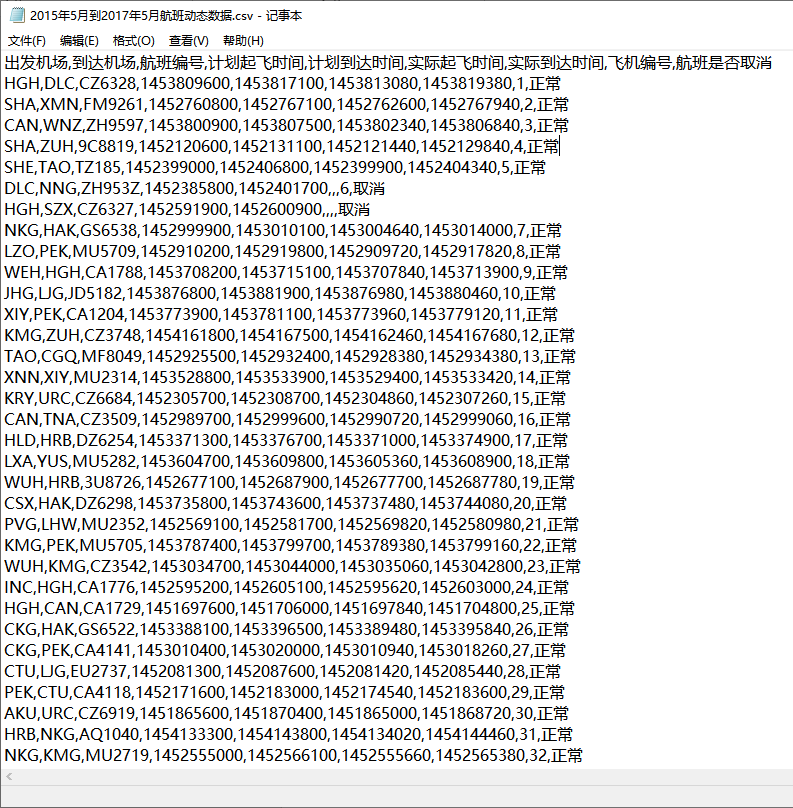


# 5.航班延迟预测算法寻找数据

## 5.1寻找到合适的数据集

航班动态起降数据集 - Heywhale.com

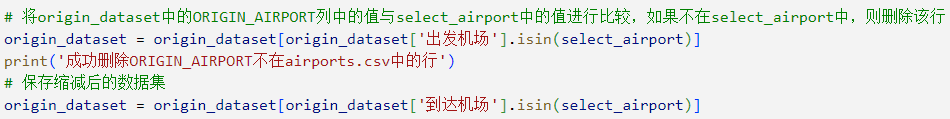
https://www.heywhale.com/mw/dataset/59793a5a0d84640e9b2fedd3



# 6.航班延迟预测算法数据预处理

使用jupyter notebook进行数据集处理

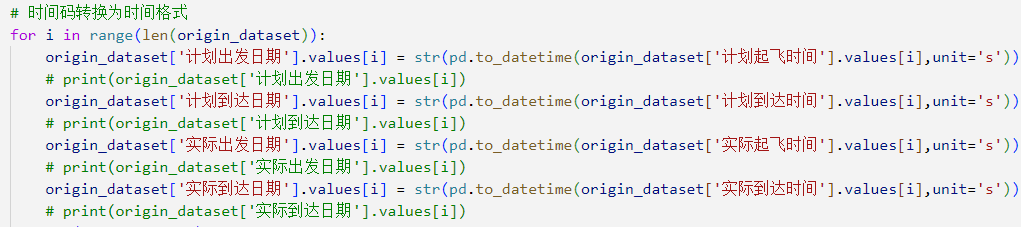
## 6.1缩减数据集，删除不是选中机场的出发机场和到达机场的行



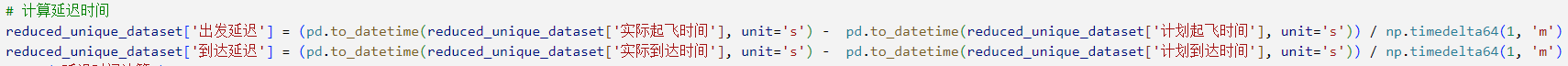
## 6.2删除取消的航班



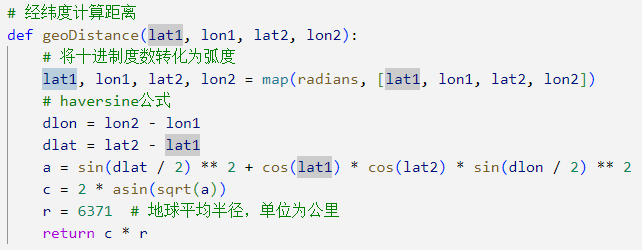
## 6.3将时间戳转化为普通时间

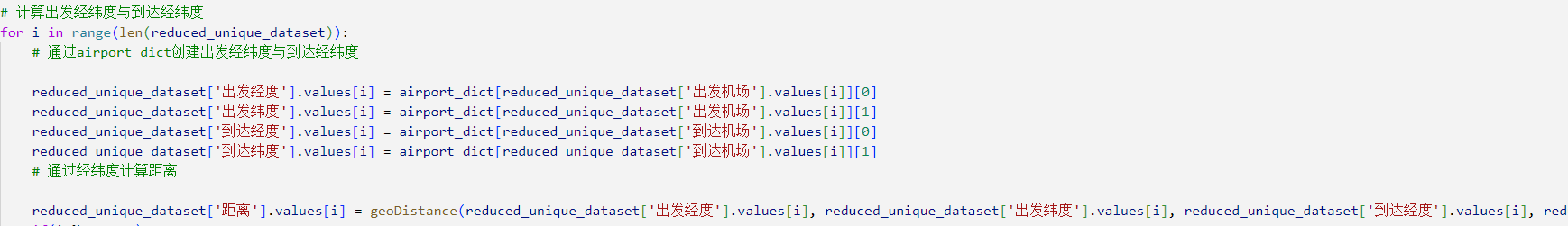


## 6.4延迟时间计算

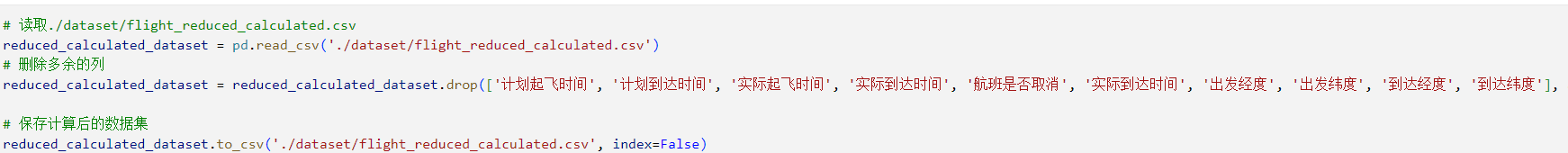


## 6.5通过两地经纬度计算距离

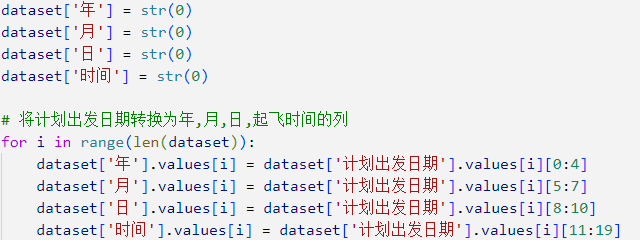




## 6.6删除多余的列



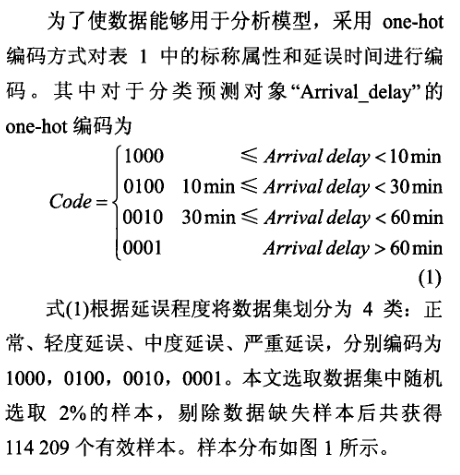
## 6.7将计划出发日期转化为年，月，日，起飞时间的列



## 6.8将出发延迟转化为分类预测对象

参考[1]唐红, 王栋, 宋博,等. 基于非线性赋权XGBoost算法的航班延误分类预测[J]. 系统仿真学报, 2021.

将延迟时间转化为延迟程度





## 6.9爬取所有地点的2015-2017年每天的气象数据



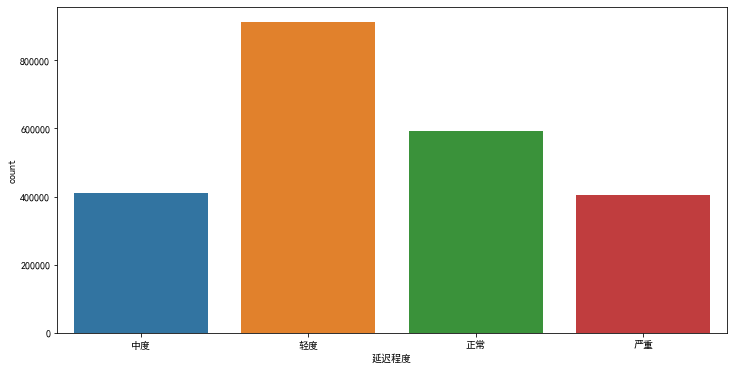
## 6.10将气象数据整合到处理好的航班延误数据集中，合并为最终的完整数据集



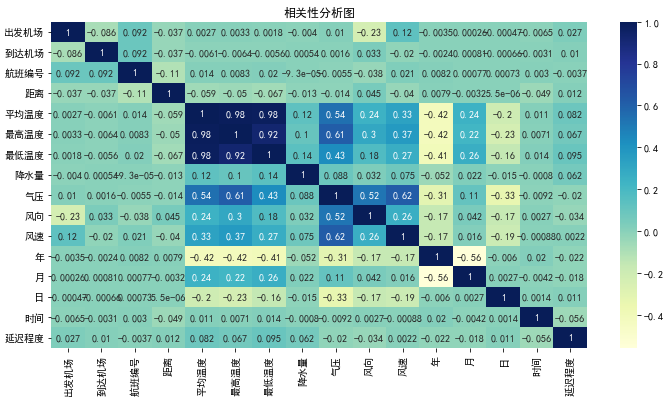
# 7.航班延迟预测算法的实现

## 7.1数据分析

### 7.1.1延迟程度频次统计图



### 7.1.2相关性分析

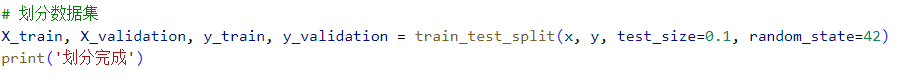


## 7.2使用基本的预测算法，成功实现基础预测功能

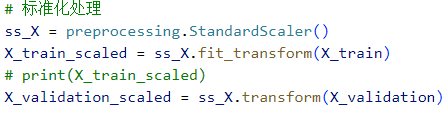
### 7.2.1将字符型数据转化为字典映射编码



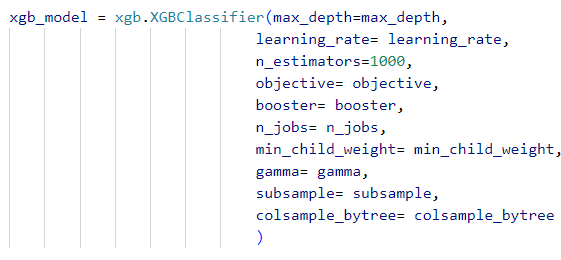
### 7.2.2划分数据集



### 7.2.3标准化处理

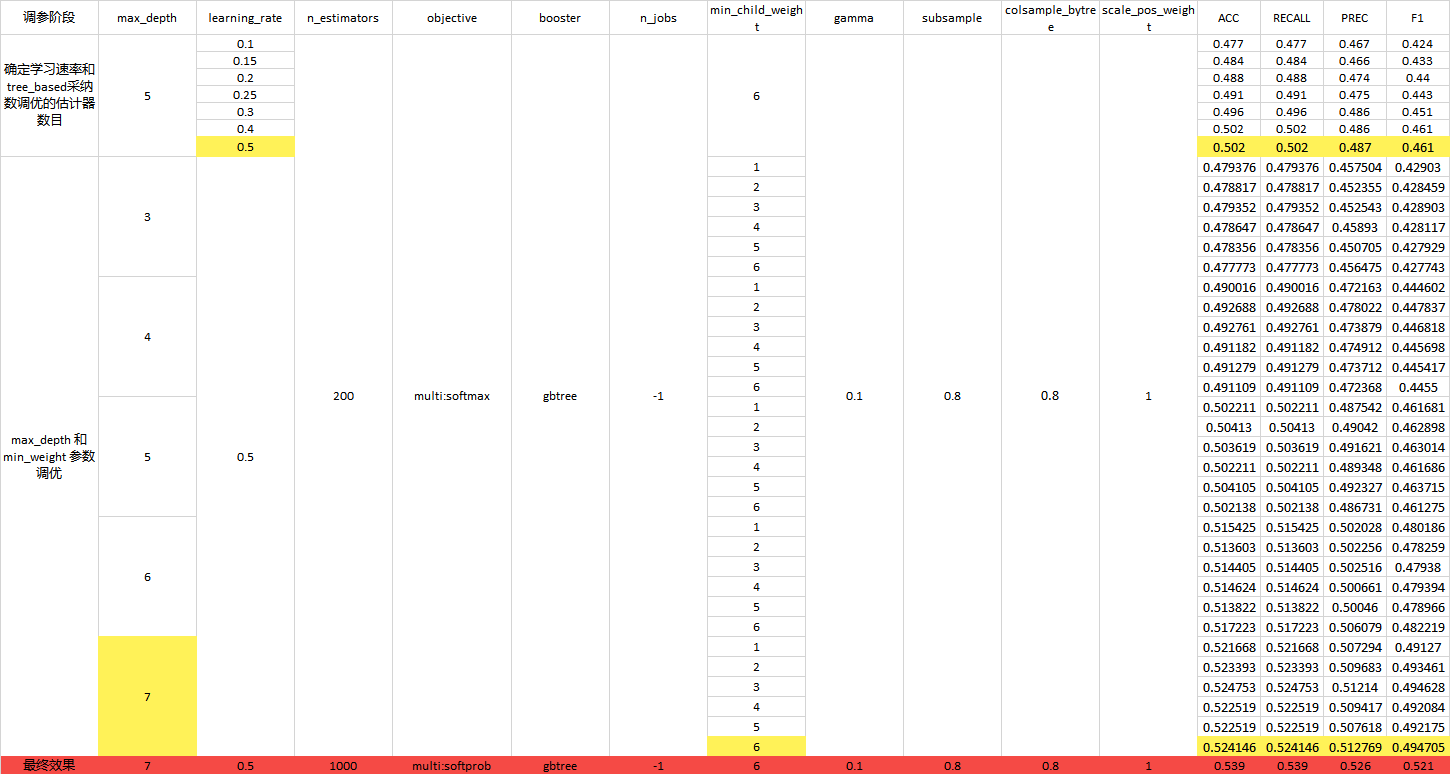


### 7.2.4建模

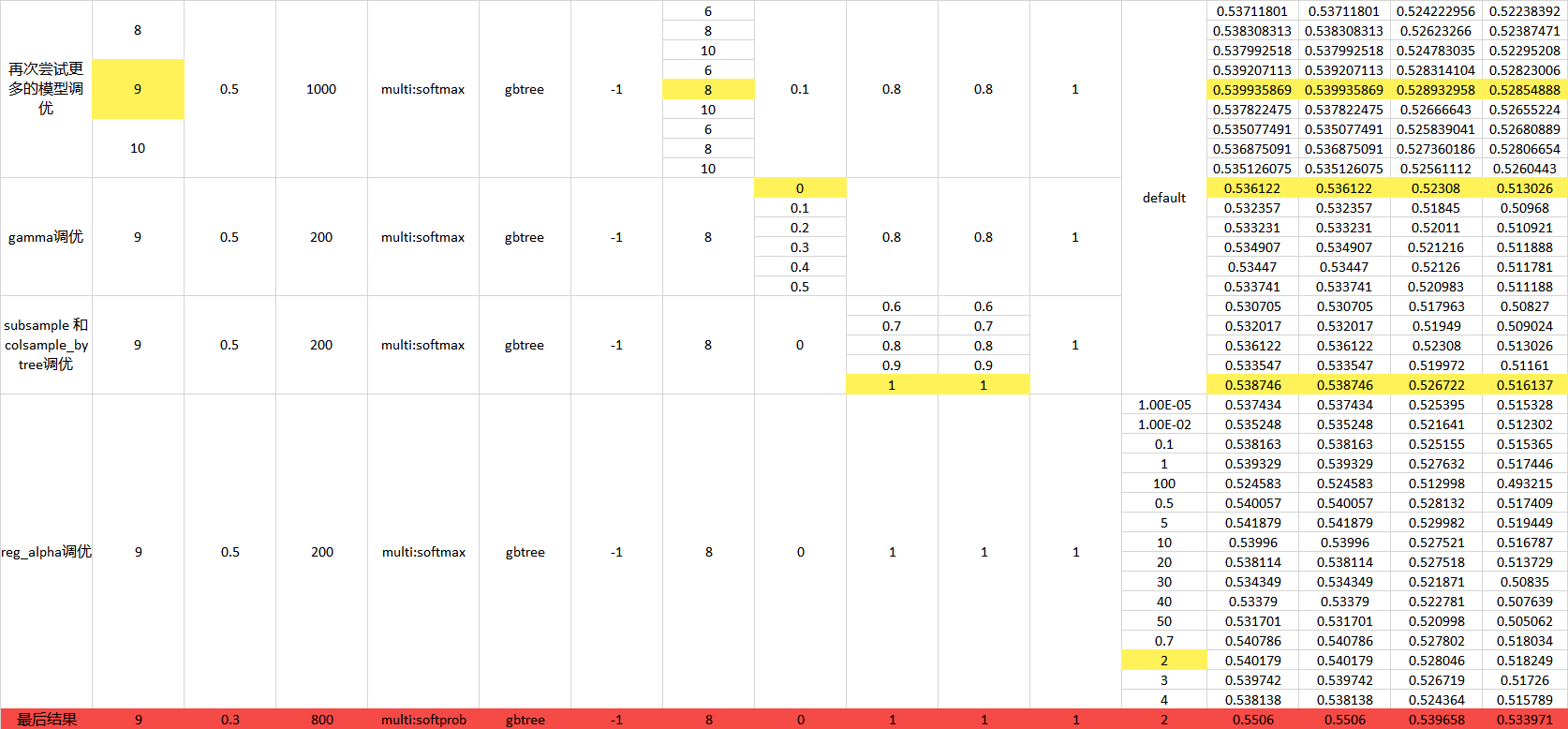


## 7.3对算法进行一定改进，使其更加合理，并给出对比结果

飞书表格链接地址：<https://lw547edakg.feishu.cn/sheets/shtcn46F5THkWBeZmcSjfCb7zLU>



首先对模型进行两个阶段的调参，第一阶段7次实验确定了学习率为0.5，第二阶段30次确定了最大深度为7，最小叶子节点权值为6。



然后继续又继续尝试了gamma, subsample, colsample\_bytree以及reg\_alpha的调优，将acc和f1尽可能地优化，通过75次的模型测试，花费近20个小时，最后达到acc为55.06%, f1score为53.397%。

与最开始的acc为47.7%，f1score为42.4%有着较为显著的提升。

# 8.可使用的参考文献

[1]中国民用航空局．2019年全国民航工作会议专题速递[EB/OL][2020-0616].http://www.caac.gov.cn/ZTZL/RDZT/2019Q GMHGZHY/.

[2]Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco:USA: ACM, 2016: 785-794

[3]Chen T Q, He T, Benesty M. Xgboost: Extreme Gradient Boosting[R], R Package Version 0.4-2, 2015: 1

[4]刘金元，丁勇，李涛．基于梯度提升决策树的航班延误分类预测[J]．数学的实践与认识，2018，48(4): 1-7.

[5]唐红, 王栋, 宋博,等. 基于非线性赋权XGBoost算法的航班延误分类预测[J]. 系统仿真学报, 2021.

[6]董念清. 中国航班延误的现状、原因及治理路径[J]. 北京航空航天大学学报(社会科学版), 2013, 26(6):25-32.

[7]唐国利, 任国玉. 近百年中国地表气温变化趋势的再分析[J]. 气候与环境研究, 2005, 10(4):8.