

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](#_Toc108392488)

[数据库对接、搭建工作 5](#_Toc108392489)

[搭建合适的数据库 6](#_Toc108392490)

[概念模型 6](#_Toc108392491)

[逻辑模型 6](#_Toc108392492)

[物理模型 6](#_Toc108392493)

[ER图 6](#_Toc108392494)

[配合后端完成数据库的对接工作 6](#_Toc108392495)

[数据库配置对接 6](#_Toc108392496)

[气象预测算法寻找数据 5](#_Toc108392497)

[气象预测算法数据预处理 5](#_Toc108392498)

[计算均值填充缺少的数值 6](#_Toc108392499)

[气象预测算法的实现 5](#_Toc108392500)

[使用GetData.py编写相关的爬虫程序爬取数据 6](#_Toc108392501)

[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 6](#_Toc108392502)

[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。 6](#_Toc108392503)

[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](#_Toc108392504)

[航班延迟预测算法寻找数据 5](#_Toc108392505)

[寻找到合适的数据集 6](#_Toc108392506)

[航班延迟预测算法数据预处理 5](#_Toc108392507)

[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](#_Toc108392508)

[删除取消的航班 6](#_Toc108392509)

[将时间戳转化为普通时间 6](#_Toc108392510)

[延迟时间计算 6](#_Toc108392511)

[通过两地经纬度计算距离 6](#_Toc108392512)

[删除多余的列 6](#_Toc108392513)

[将计划出发日期转化为年，月，日，起飞时间的列 6](#_Toc108392514)

[将出发延迟转化为分类预测对象 6](#_Toc108392515)

[爬取所有地点的2015-2017年每天的气象数据 6](#_Toc108392516)

[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 6](#_Toc108392517)

[航班延迟预测算法的实现 5](#_Toc108392518)

[数据分析 6](#_Toc108392519)

[延迟程度频次统计图 6](#_Toc108392520)

[相关性分析 6](#_Toc108392521)

[使用基本的预测算法，成功实现基础预测功能 6](#_Toc108392522)

[将字符型数据转化为字典映射编码 6](#_Toc108392523)

[划分数据集 6](#_Toc108392524)

[标准化处理 6](#_Toc108392525)

[建模 6](#_Toc108392526)

[对算法进行一定改进，使其更加合理，并给出对比结果 6](#_Toc108392527)

[可使用的参考文献 5](#_Toc108392528)

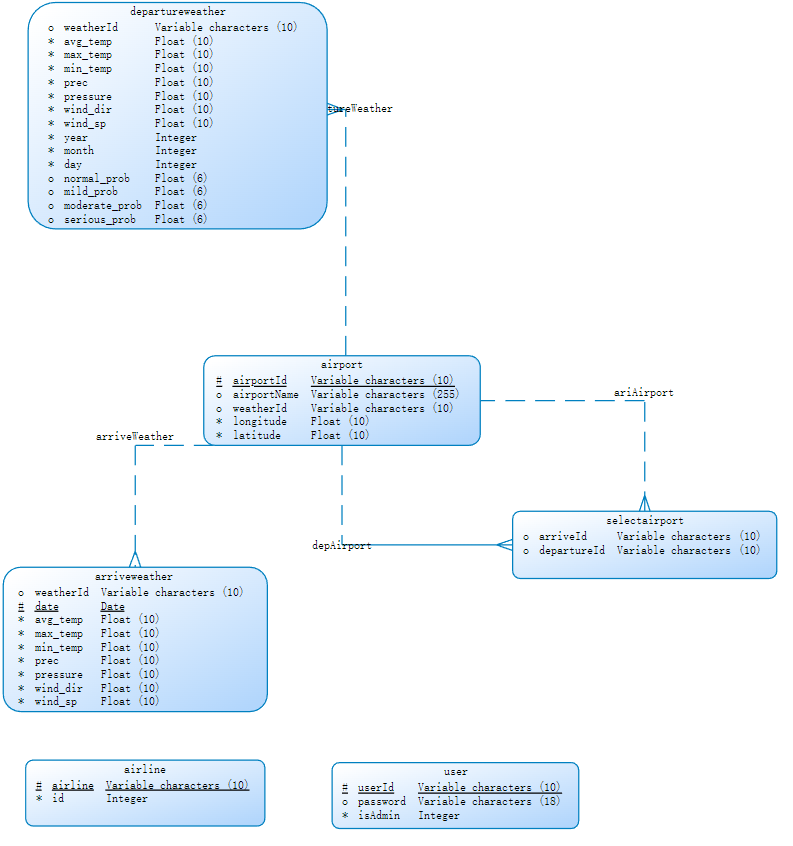
数据/算法文档

# 数据库对接、搭建工作

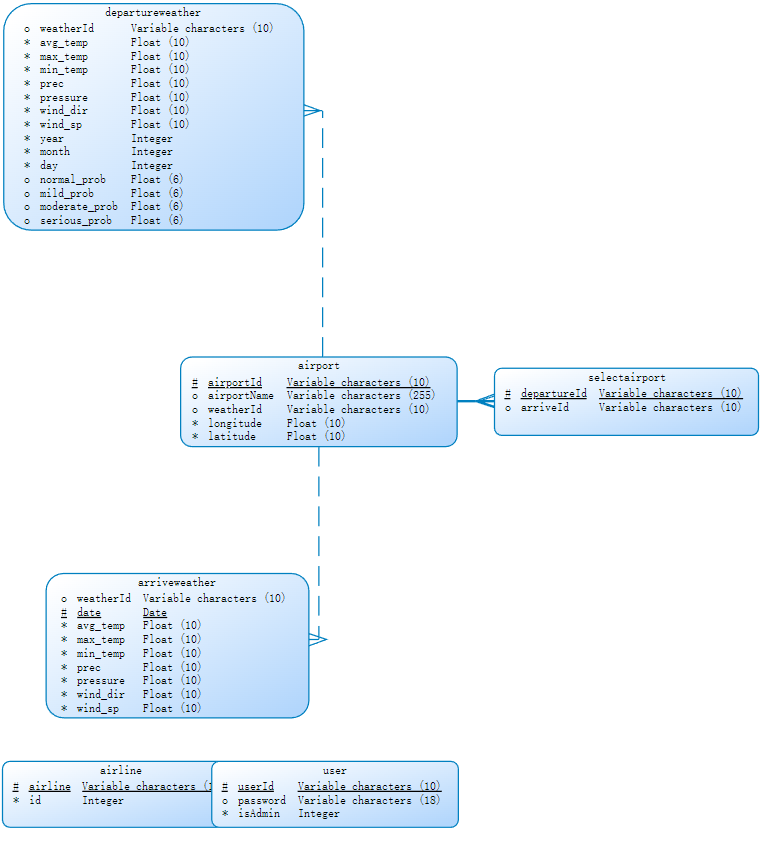
## 搭建合适的数据库

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

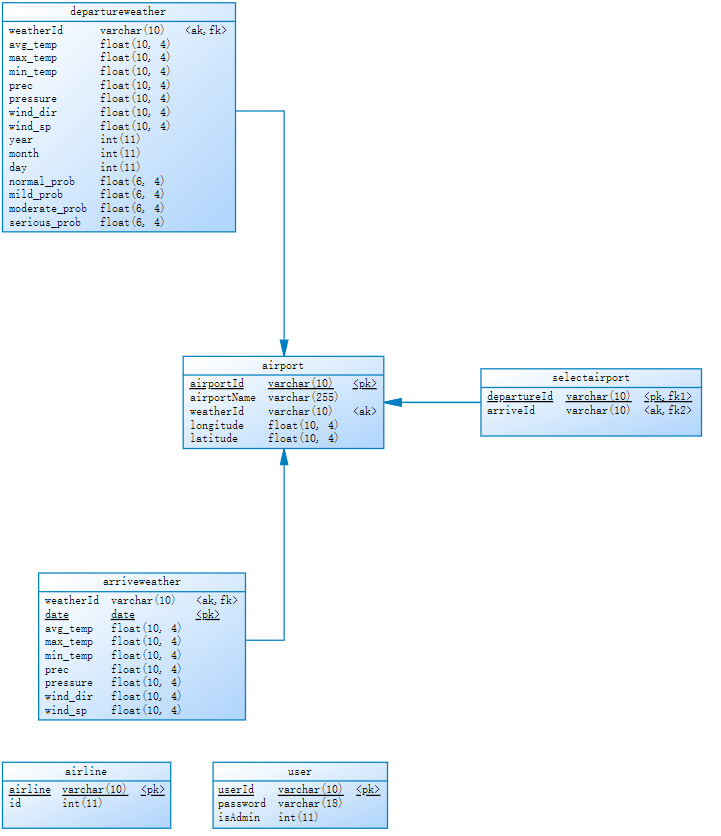
### 概念模型



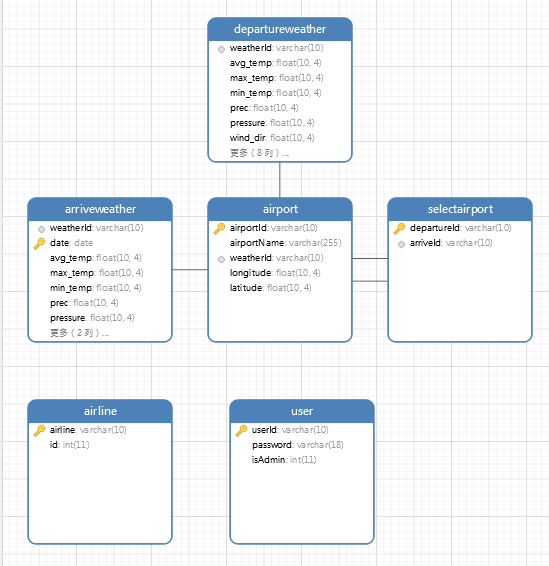
### 逻辑模型



### 物理模型

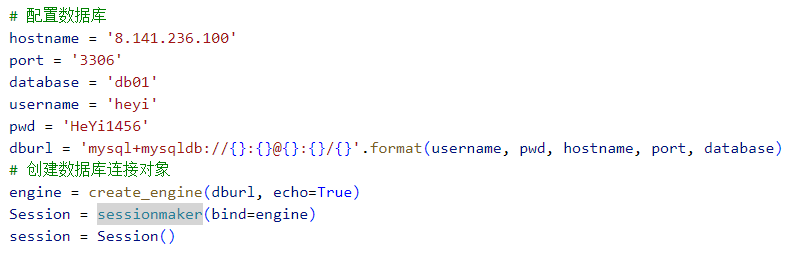


### ER图



## 配合后端完成数据库的对接工作

### 数据库配置对接

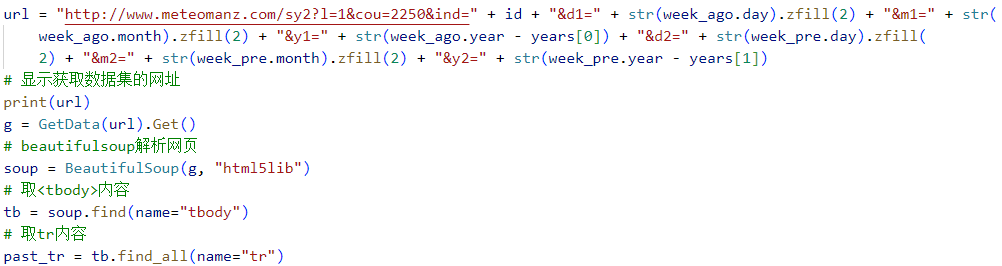


使用session便可以使用sql语句对数据库进行操作

# 气象预测算法寻找数据

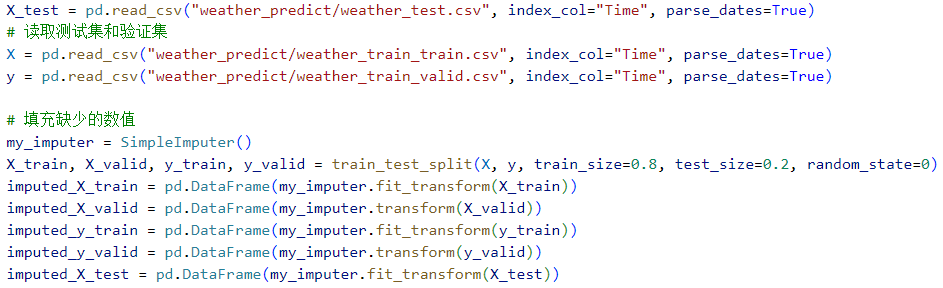
训练数据来源于[http://www.meteomanz.com/](http://www.meteomanz.com/" \t "dlt)

通过BeautifulSoup类库进行爬取



# 气象预测算法数据预处理

## 计算均值填充缺少的数值

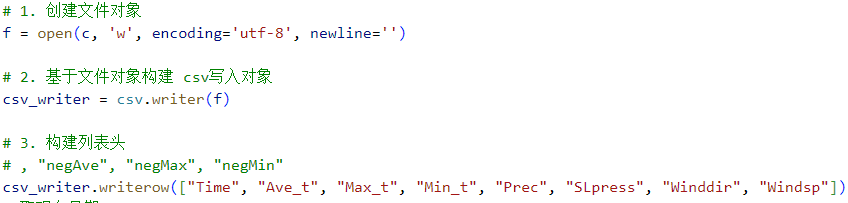


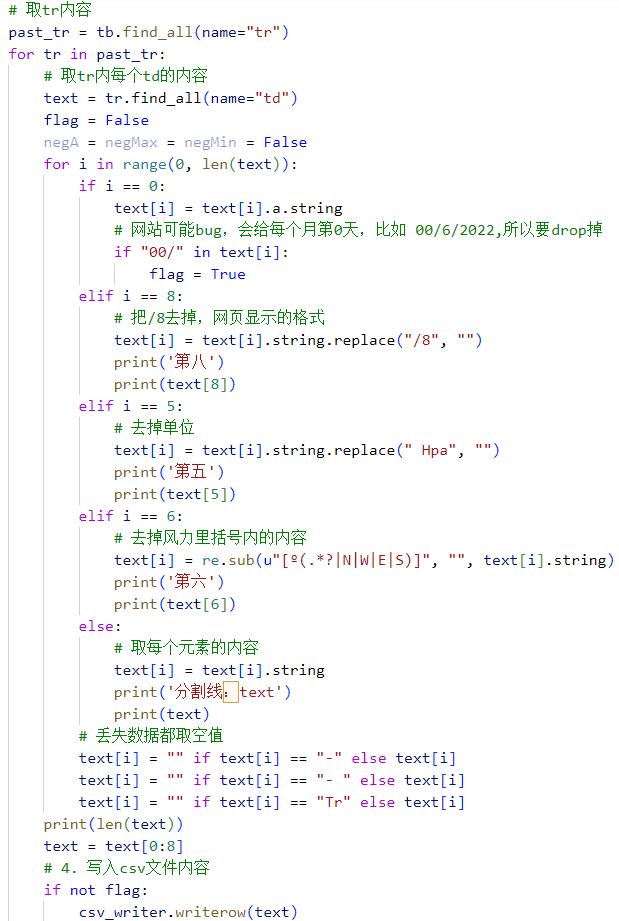
# 气象预测算法的实现

## 使用GetData.py编写相关的爬虫程序爬取数据



## 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





## 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。



## 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

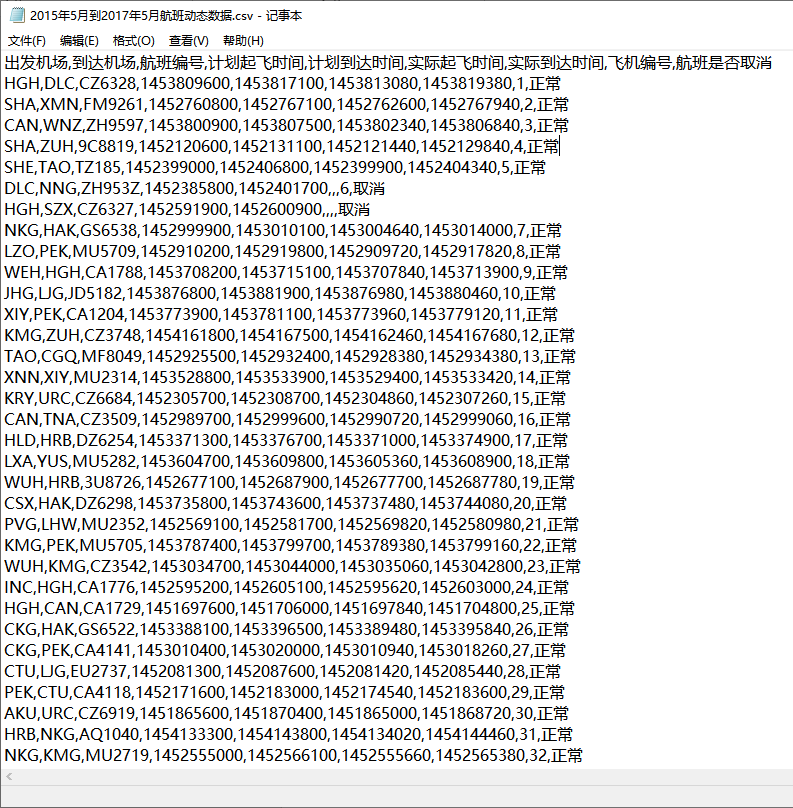


# 航班延迟预测算法寻找数据

## 寻找到合适的数据集

航班动态起降数据集 - Heywhale.com

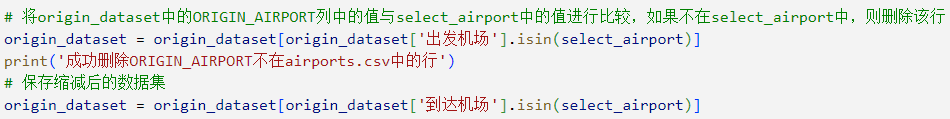
https://www.heywhale.com/mw/dataset/59793a5a0d84640e9b2fedd3



# 航班延迟预测算法数据预处理

使用jupyter notebook进行数据集处理

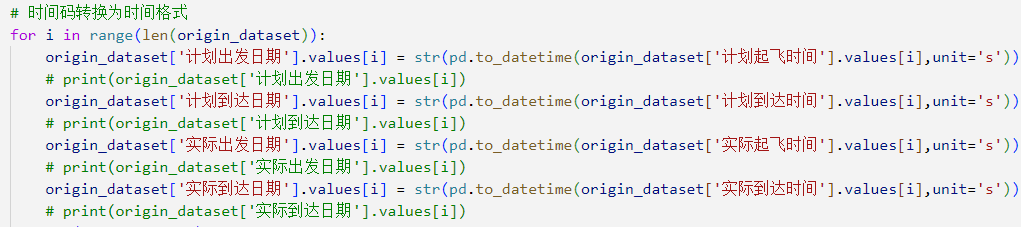
## 缩减数据集，删除不是选中机场的出发机场和到达机场的行



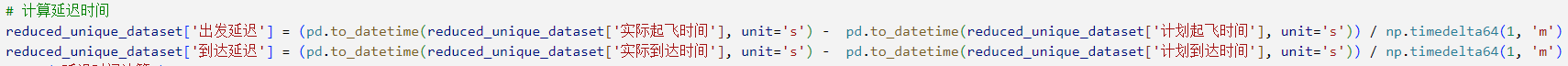
## 删除取消的航班



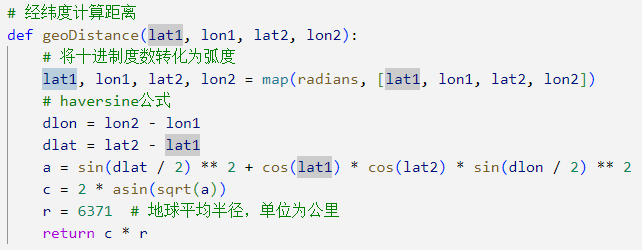
## 将时间戳转化为普通时间

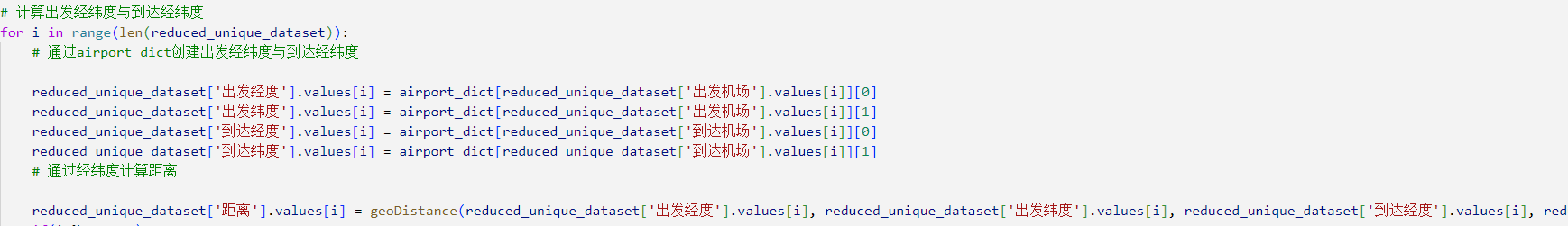


## 延迟时间计算

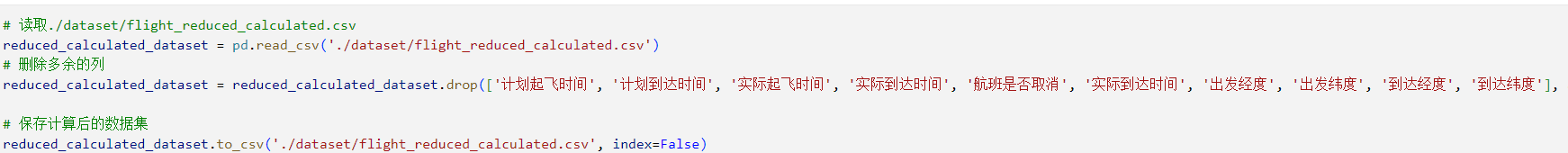


## 通过两地经纬度计算距离

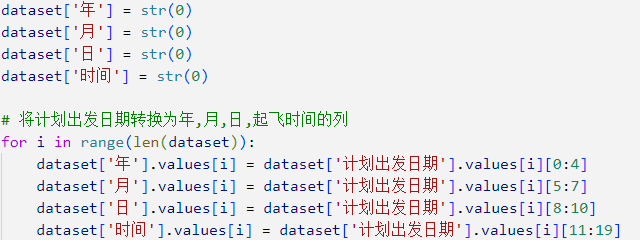




## 删除多余的列



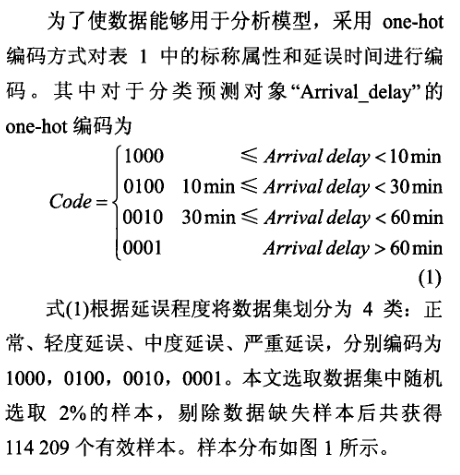
## 将计划出发日期转化为年，月，日，起飞时间的列



## 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





## 爬取所有地点的2015-2017年每天的气象数据



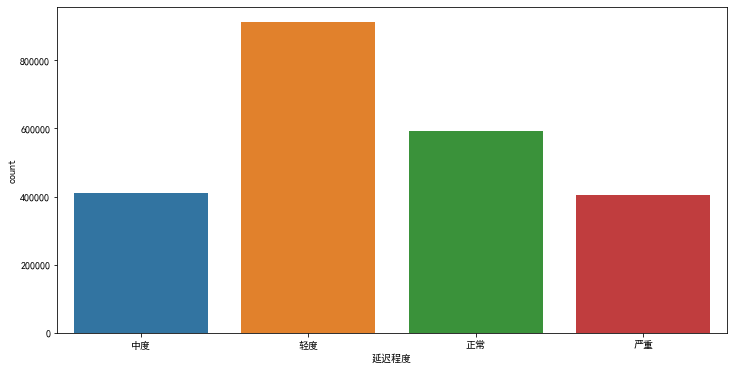
## 将气象数据整合到处理好的航班延误数据集中，合并为最终的完整数据集



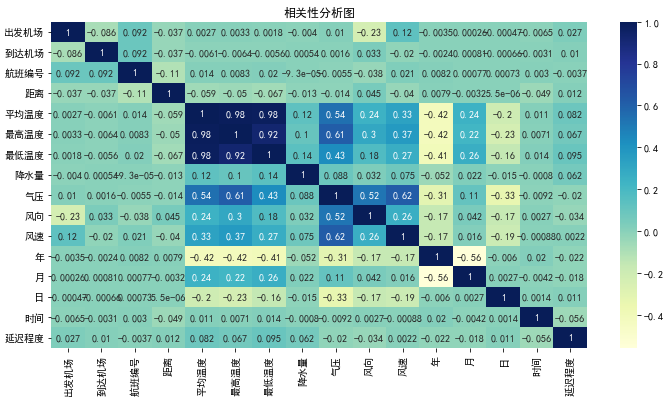
# 航班延迟预测算法的实现

## 数据分析

### 延迟程度频次统计图



### 相关性分析

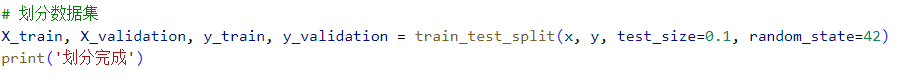


## 使用基本的预测算法，成功实现基础预测功能

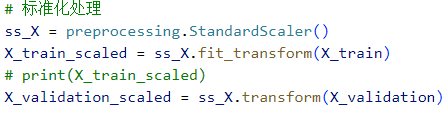
### 将字符型数据转化为字典映射编码



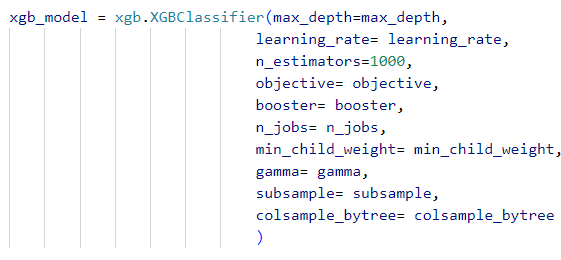
### 划分数据集



### 标准化处理

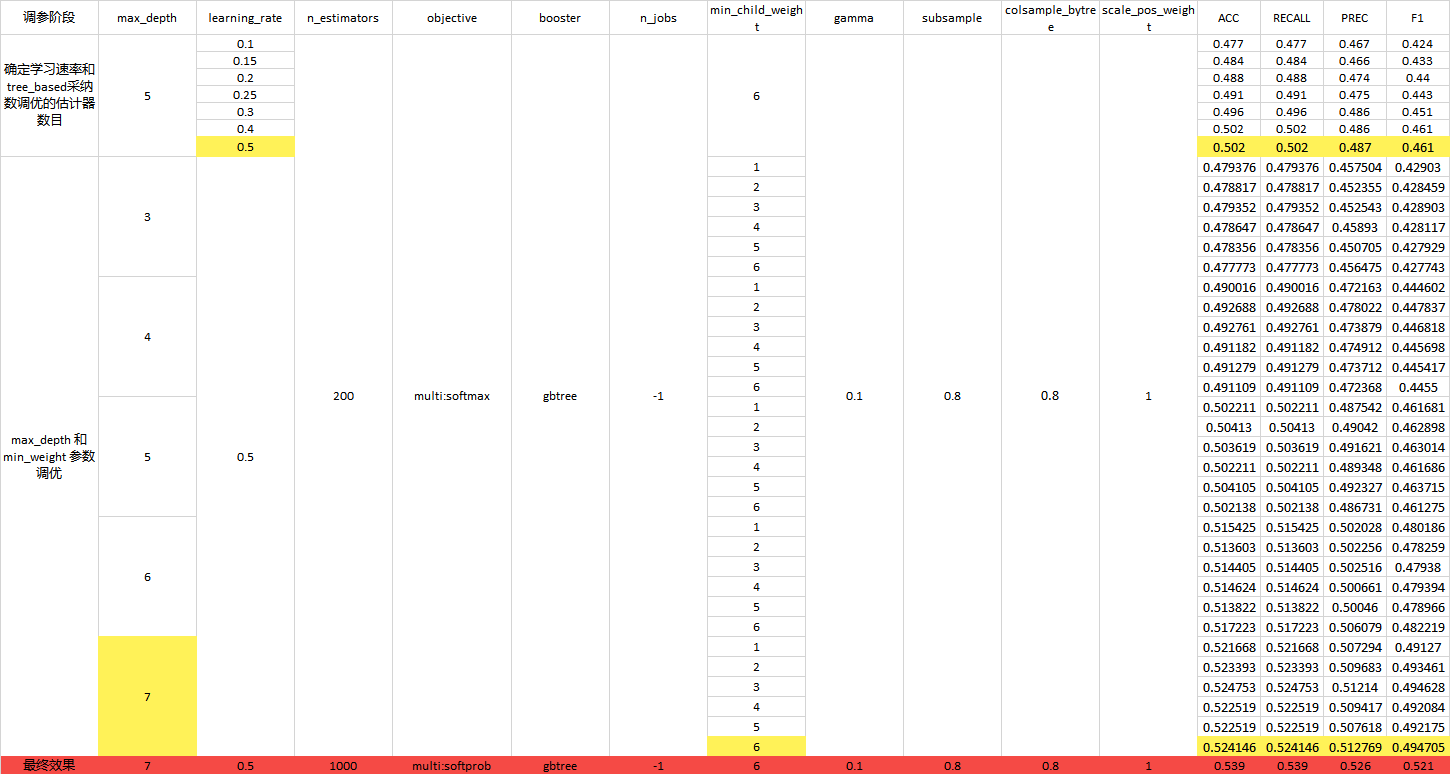


### 建模



## 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>



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。

最后模型准确率从最开始的0.446提升到0.539，F1 score从0.403提升到0.521。

# 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.