超密网络下业务流量预测算法使用说明书

# 毕设代码功能概述

毕设代码主体包含以下几个部分功能：

（1）基于多元历史信息和MLT-Boosting算法的整体区域模型；

（2）基于空时影响特性结合残差拟合机制的登机口区域模型；

（3）整体区域模型和登机口区域模型的模型融合；

算法架构示意图如下：



图1.1 算法架构示意图

基于MLT-Boosting算法的系统架构如下所示：

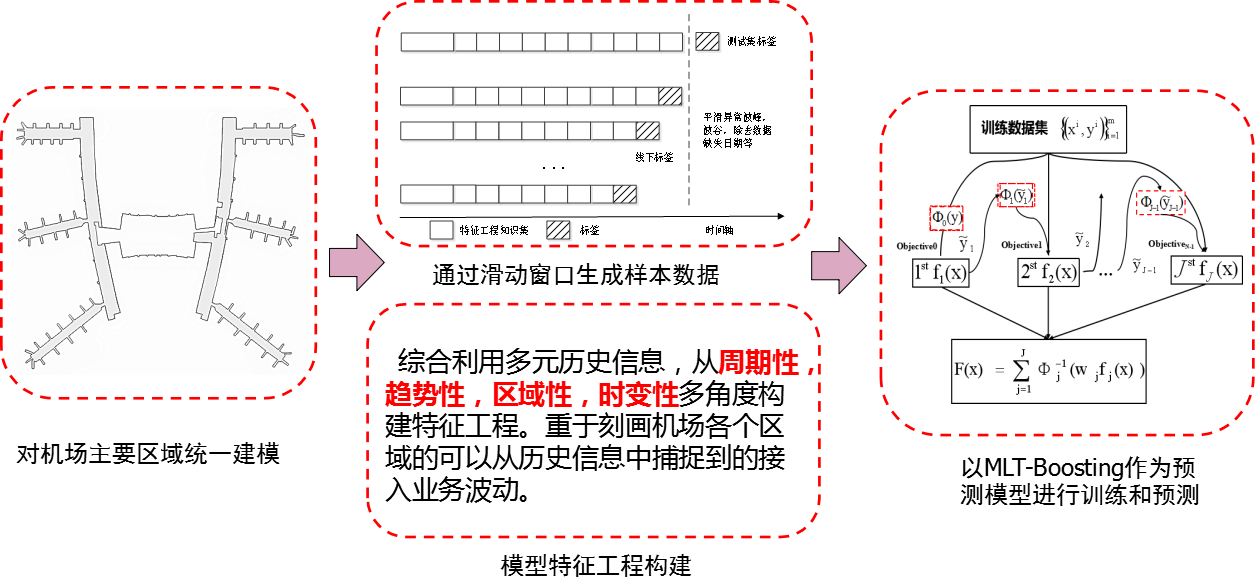


图1.2 整体区域模型架构示意图

基于空时影响因子的登机口区域模型示意图如下所示：

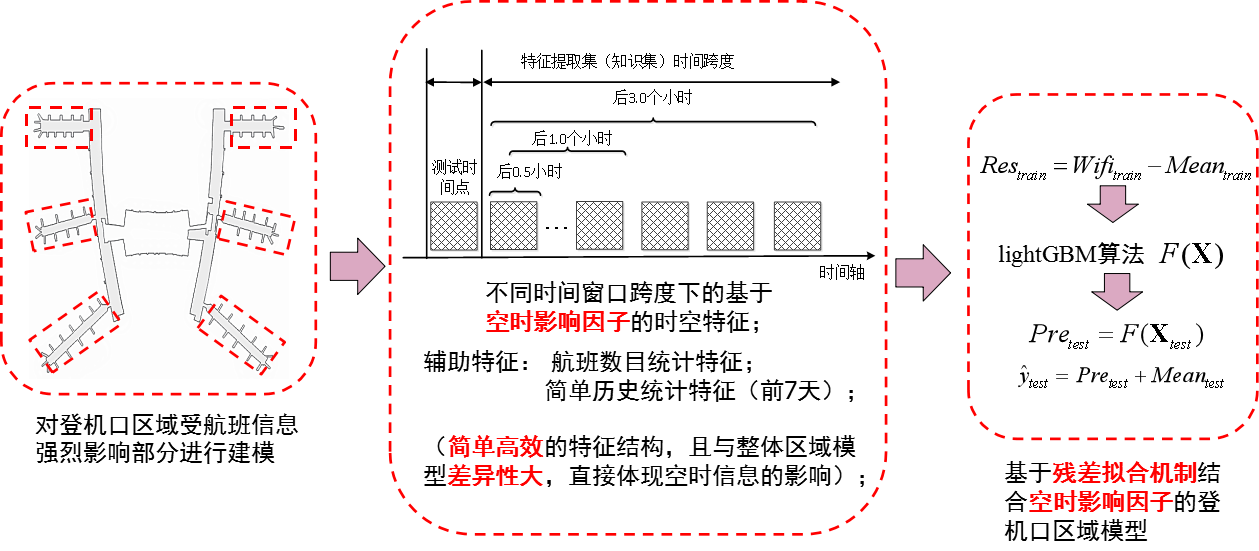


图1.3 登机口区域模型架构示意图

# 二、仿真运行

2.1首先打开Anconda3软件,python版本号为python3.6软件界面如下：



图2.1 软件启动界面

算法运行所需软件包及参考版本型号说明如下：

Numpy, 1.13.3

Pandas, 0.20.3

Lightgbm, 0.10

Scikit-learn 0.19.0

Scipy ..............0.19.1

2.2数据预处理阶段代码运行，处理原始数据生成日期，小时，分钟等字段：

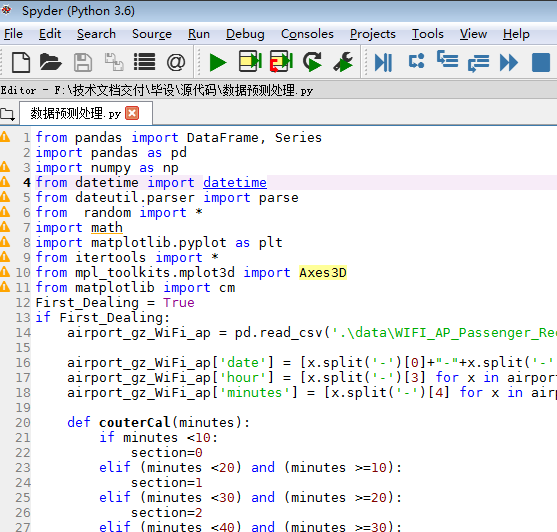


图2.2数据预处理示意图

其中原始数据的存储位置设定见如下路径，命名为：

'.\data\WIFI\_AP\_Passenger\_Records\_chusai\_2ndround.csv'

预处理后的数据存储位置及命名为：

r'.\data\airport\_gz\_WiFi\_apAfter.csv'

2.3生成整体区域模型的训练和测试数据，构建整体区域模型特征工程，运行代码文件示意如下：

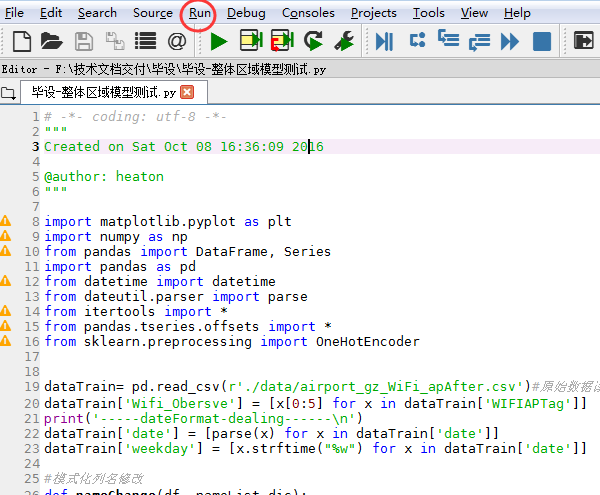
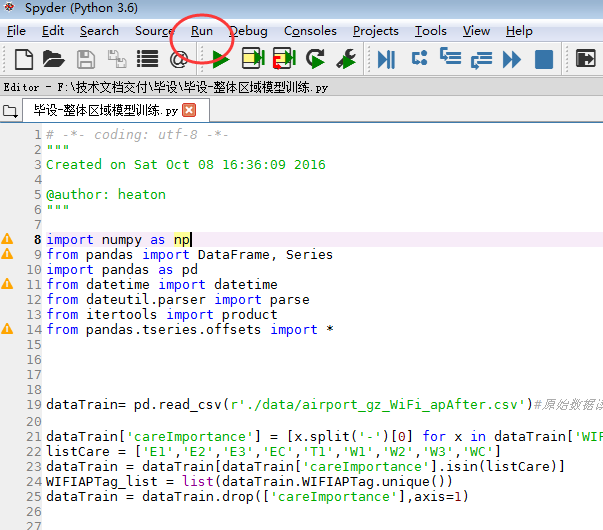


图2.3 整体区域模型训练集测试集生成

其中整体区域模型训练集及测试集存储位置设定见如下路径，分别命名为：dataFeatureOfflineTrain，dataFeatureOfflineTest。

训练集：'./data/dataFeatureOfflineTrain.csv'

测试集：'./data/dataFeatureOfflineTest.csv'

2.4生成登机口区域模型的训练和测试数据，构建登机口区域模型特征工程，运行代码文件示意如下：

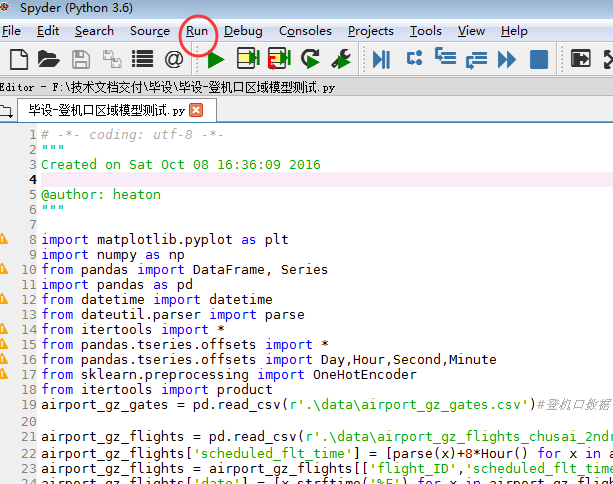
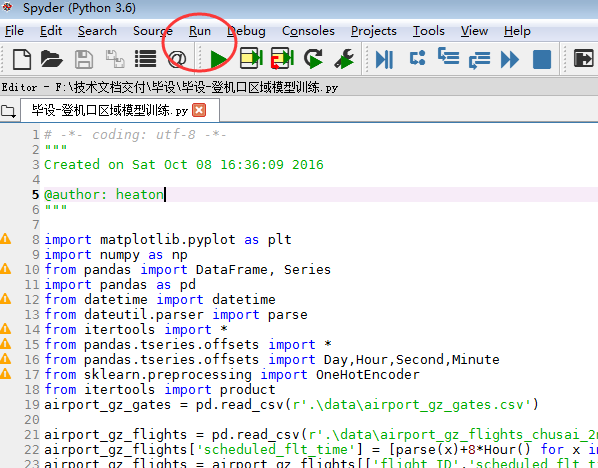


图2.4 登机口区域模型训练集测试集生成

其中登机口区域模型训练集存储位置设定见如下路径，分别命名为：dataFeatureOfflineTrainSpace，dataFeatureOfflineTestSpace。

训练集：'./data/dataFeatureOfflineTrainSpace.csv'

测试集：'./data/dataFeatureOfflineTestSpace.csv'

2.5运行基于mlt-boosting的整体区域模型训练预测代码：

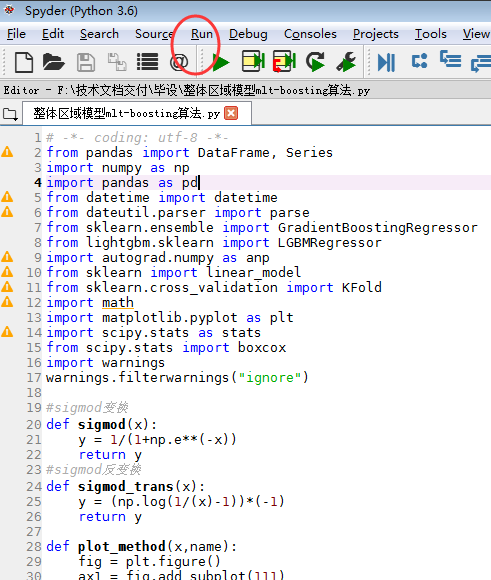


图2.5 基于mlt-boosting的整体区域模型预测代码

其中整体区域模型的预测结果存储路径及命名如下：

'./data/dataTestboosting.csv'

2.6运行基于空时影响因子的登机口区域模型训练预测代码：

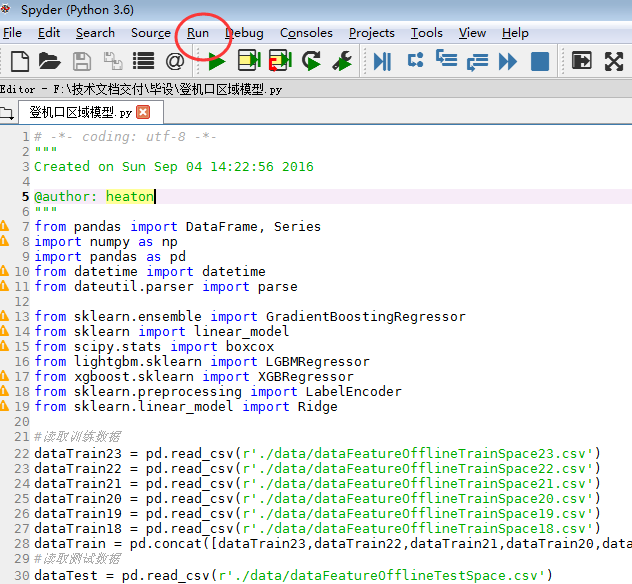


图2.6 登机口区域模型预测代码

其中登机口区域模型的预测结果存储路径及命名如下：

‘/data/dataTestAir.csv’

2.7 运行加权融合模型代码得到最终预测结果，融合整体区域模型预测结果和登机口区域模型预测结果，并对最终结果进行评分打印，运行代码示意如下：

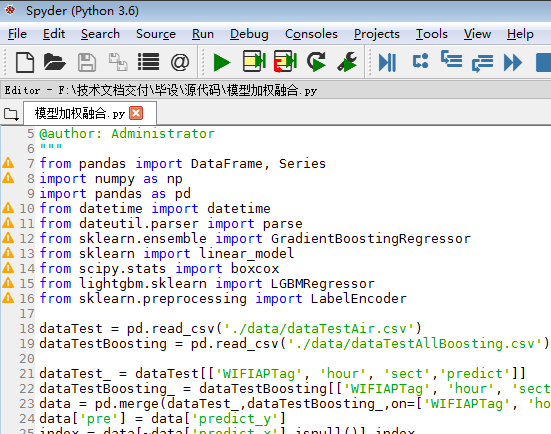


图2.7 模型融合预测代码

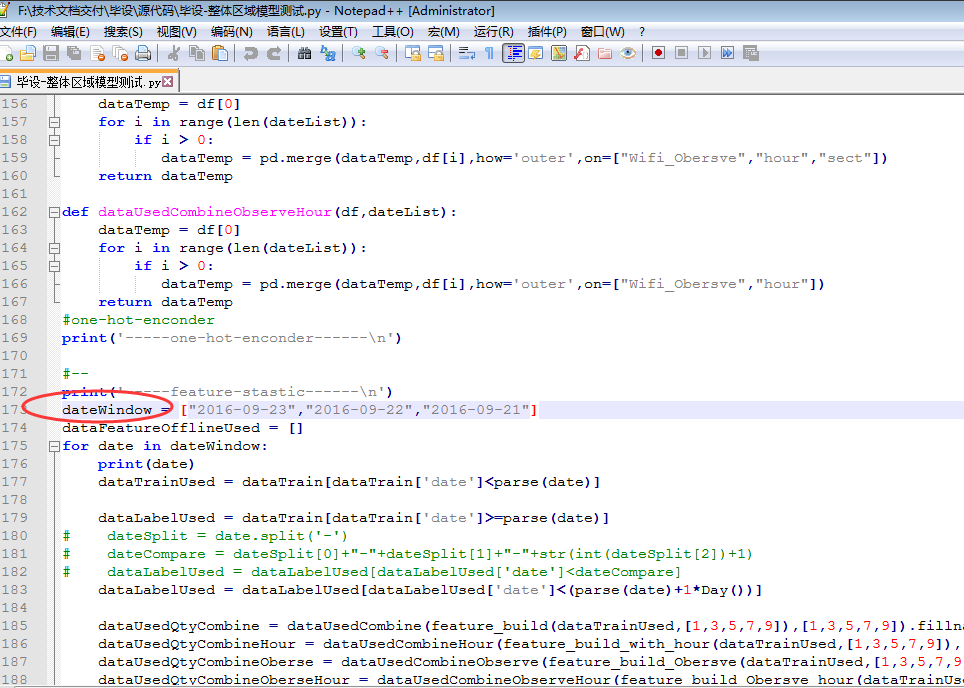
# 三、参数设置

1.整体区域模型设置训练集日期：调整dateWindow参数包含日期当前设置为：

"2016-09-23","2016-09-22","2016-09-21"

2.整体区域模型设置测试集日期：调整dateWindow参数包含日期当前设置为：

"2016-09-24"



3.1参数调整位置示意图1

1. 登机口区域模型调整训练集合及测试集时间跨度：

当前训练集合为：09-18,09-19,09-20,09-21,09-22,09-23号

dataTrain23 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace23.csv')

dataTrain22 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace22.csv')

dataTrain21 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace21.csv')

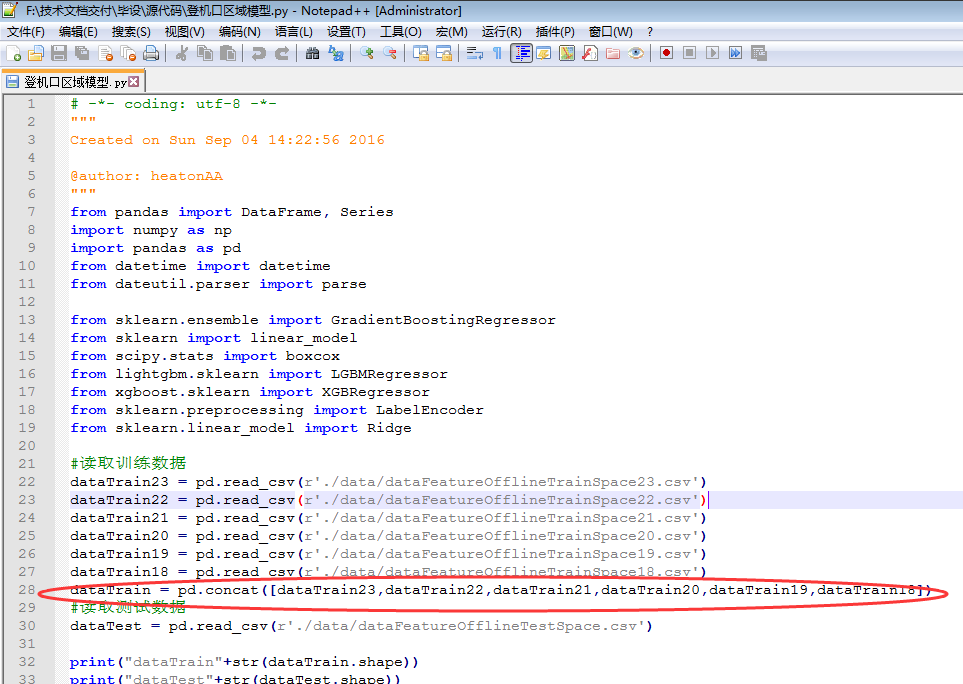
dataTrain20 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace20.csv')

dataTrain19 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace19.csv')

dataTrain18 = pd.read\_csv(r'./data/dataFeatureOfflineTrainSpace18.csv')

dataTrain = pd.concat([dataTrain23,dataTrain22,dataTrain21,dataTrain20,dataTrain19,dataTrain18])

当前测试集合为：09-24号



3.2参数调整位置示意图2

4.MLT-Boosting算法参数设置：

train#模型训练集

test#模型测试集

train\_label#模型训练集标签

First\_change#最外层是否引入变换

steps#最大变换次数

n\_estimators\_list#阶段迭代数列表

learning\_rate\_stage = 1#阶段权重

max\_depth=3#阶段内基模型深度

learning\_rate=0.02#阶段内学习速率

subsample=1#阶段内样本采样

colsample\_bytree=1#阶段内基模型样本采样

# 仿真数据说明

测试数据存储于“./源代码/data”中，本部分仿真所使用数据来自典型超密集网络场景大型机场，广州白云国际机场真实历史客流数据。大型机场作为典型的超密集网络架设场景，其拥有巨大的旅客吞吐量。且旅客的移动性导致用户业务分布呈现出动态和密度可变的特点，各个区域业务分布差异性较大。各种通信资源分配方案等都需要能够精准预测未来一段时间跨度下细粒度时间间隔的用户业务接入分布波动情况，并据此提前调配人力物力，进行通信资源的无缝调度，提升用户体验。具体数据介绍如下：

1. 历史连接Wifi-AP（Access Point）的人数表，三个字段分别为：wifi\_ap\_tag，字符串，描述WiFi接入的AP点；passenger\_count, 整数，描述在某一时刻接入该Wifi-AP的设备数量；time\_stamp,字符串，描述该时刻，精确到秒。该表的实际数据下表所示。

|  |  |  |
| --- | --- | --- |
| wifi\_ap\_tag  Wifi-AP标识 | passenger\_count  接入用户量（人） | time\_stamp  时间戳 |
| E1-2C<E1-2C-01> | 20 | 2016-09-10-18-55-04 |
| E1-3A<E1-3-01> | 11 | 2016-09-10-18-55-04 |
| E1-3A<E1-3-02> | 24 | 2016-09-10-18-55-04 |

表4.1 Wifi-AP人数表

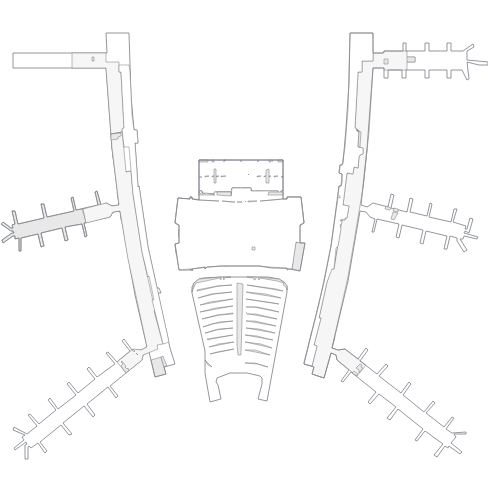


图4.2 广州白云机场平面图

本文关注的主要是机场的流聚集区域，区域代码为E1, E2, E3, EC, T1, W1, W2, W3, WC。其中，E和W分别代表机场的东侧和西侧登机区，各有三个小区；EC和WC是指连接安检区和登机区的走廊。机场平面图4.2所示。WiFi\_ap\_tag的前四位字符，代表了该Wifi-AP所在区域和楼层，例如名称的前四位“T1-1”意味着在T1航站楼的1楼。

1. 安检旅客过关人数表:真实旅客安检行为历史信息，下面的例子是这张表的实际数据，数据由三列构成分别为旅客ID，安检时间，及对应航班ID。

|  |  |  |
| --- | --- | --- |
| 旅客号passenger\_ID | 安检时间security\_time | 航班号flight\_ID |
| 1 | 2016/9/10 04:30 | 1001 |
| 2 | 2016/9/10 04:30 | 1002 |
| 3 | 2016/9/10 04:31 | 1003 |

表4.2 安检旅客过关人数表

1. 旅客进入－离开机场的行程表。

|  |  |  |  |
| --- | --- | --- | --- |
| 旅客ID  passenger\_ID2 | 航班ID  flight\_ID | 航班起飞时间  flight\_time | 旅客出票时间  checkin\_time |
| 25296480 | 1001 | 2016/9/11 13:50 | 2016/9/11 11:57 |
| 25296482 | 1002 | 2016/9/11 14:00 | 2016/9/11 11:55 |
| 25296512 | 1004 | 2016/9/11 13:05 | 2016/9/11 11:57 |

表4.3 旅客进入－离开机场的行程表

（4）航班排班表：每天的半夜两点，机场就会把接下来一整天的航班排班，实际上那时一整天的旅客在哪个登机口登机就已经确定了。所以，在执行预测任务的时候，是可以看到当天的航班排班表的。

而航班因素是大型机场场景下客流量及接入业务变化的一个重要影响因素。这张表里的字段为：flight\_ID、航班号、scheduled\_flt\_time、航班初始排班的预定起降时间、登机口号码、Arrival\_Departure（航班是离港航班起飞（D）或是到港航班降落（A））。scheduled\_flt\_time使用国际航空的标准时间：格林威治时间。在当前季节转换方法为：北京时间＝格林威治时间（GMT）＋8小时。

（5）机场登机口区域表：登机口在机场内哪个区域。BGATE\_ID代表登机口号，而BGATE\_AREA代表登机口在机场的区域代号，与Wifi-AP名称中的前两位相对应。

（6）机场Wifi-AP接入点坐标表：在机场CAD图上，标注了大部分Wifi-AP。位置坐标的单位采用毫米。

本文预测目标介绍如下：以广州白云机场真实的客流数据为基础，通过数据算法来构建旅客用户业务接入预测模型，预测未来一个整天以十分钟为粒度，密布在机场场景内700多个Wifi-AP的每十分钟的平均用户接入数目。结果以平方误差作为评价公式：



其中为待预测接入用户值，为实际Wifi-AP用户接入值。