

**计算机科学技术学院**

**本科生毕业论文**

**基于回归算法的公交线路客流预测**

**班 级： 计算机科学与技术（非师范）**

**学 号： 20161102910**

**姓 名： 白璐**

**指导教师： 李艳玲**

**2020 年 4 月 22 日**

**计算机科学技术学院制**

毕 业 论 文 目 录

[1 引言 - 1 -](#_Toc38655327)

[1.1 背景介绍 - 1 -](#_Toc38655328)

[1.2 研究内容 - 1 -](#_Toc38655329)

[2 开发环境及工具 - 2 -](#_Toc38655330)

[2.1 Python - 2 -](#_Toc38655331)

[2.2 Spyder - 2 -](#_Toc38655332)

[3 需求分析 - 2 -](#_Toc38655333)

[4 特征数据分析 - 3 -](#_Toc38655334)

[4.1 数据分析 - 3 -](#_Toc38655335)

[4.2 数据预处理 - 3 -](#_Toc38655336)

[4.3 特征分析 - 4 -](#_Toc38655337)

[4.4 特征提取 - 13 -](#_Toc38655338)

[5 回归算法 - 14 -](#_Toc38655339)

[5.1 线性回归模型的一般形式 - 15 -](#_Toc38655340)

[5.2 拟合 - 15 -](#_Toc38655341)

[5.3 python实现线性回归 - 15 -](#_Toc38655342)

[5.4 回归评价指标 - 16 -](#_Toc38655343)

[6 实验 - 17 -](#_Toc38655344)

[7 运行结果显示 - 18 -](#_Toc38655345)

[8 总结 - 19 -](#_Toc38655346)

[9 致谢 - 20 -](#_Toc38655347)

[参考文献 - 21 -](#_Toc38655348)

**全文共** 24  **页** 5642 **字**

基于回归算法的公交线路客流预测

计算机科学技术学院 2016级 计算机科学与技术（非师范）白璐 20161102910

指导老师 李艳玲 副教授

摘要 本次毕业设计主要使用python编程语言，通过使用Spyder开发环境，选择合适的特征并处理后，构建并训练模型，实现基于回归算法的公交线路客流预测。目的是精准预测未来某天的公交客流，以指导公交公司进行合理的车辆调度。

关键词Python；Spyder；回归算法

# 1 引言

## 1.1 背景介绍

在我国经济高速发展，国家日益强大，社会结构急剧变化，城市化进程不断加快的情况下，居民对未来生活质量提出了更多新的要求，解决日益严重的交通拥挤问题已经迫在眉睫。推行公交优先策略可以有效缓解城市交通拥堵问题，而推行公交优先的关键在于公交车辆的及时调度和公交线路的合理规划,并且公交调度和线路规划都取决于对公交线路客流量细致的剖析和精准的预测[1]。公交线路客流量是规划公交运营的基础数据 [2]，单从IC卡获取的客流数据不能代表公交客流整体情况，天气情况也会影响市民的出行，所以需要综合分析多种因素。

本次毕业设计题目来自于“阿里云天池大赛”，目的是通过大量的数据分析，挖掘市民在公共交通中的行为模式，指导公交公司利用有限的车辆资源制定科学合理的车辆调度方案，提高城市公交的运营效率和服务水平。

## 1.2 研究内容

本次毕业设计通过分析2014年8月1日到2014年12月25日广州市公交线路的历史公交刷卡数据和天气状况数据，挖掘城市居民在公共交通中的行为模式，建立公交线路客流预测模型，并预测未来一周（20141225-20141231）的公交客流量。

# 2 开发环境及工具

## 2.1 Python3.7

Python是一种跨平台的计算机程序设计语言，也是一种解释型的脚本语言，它简洁的语法和对动态输入的支持，特别适用于快速的应用程序开发[3]，使得越来越多的人选择使用，是现在公认的大数据处理分析语言之一。如今python既可以做前台网页，也可以做后台程序，还可以进行数据分析展示，以及编写网络爬虫等。

## 2.2 Spyder

Spyder是Anaconda中一个简单版本的集成开发环境。Spyder可以配置虚拟环境，同时对于包的管理十分方便，便于处理数据分析问题，它的优势在于可以很方便地观察和修改数组的值。

Spyder的界面由许多窗格组成，窗格包括“Editor”、“Object inspector”、“Variable explorer”、“File explorer”、“Console”、“History log”。用户可以自行在View菜单中设置是否显示这些窗格以及调整它们的相对位置和大小范围。

# 3 需求分析

通过对原始数据的处理，选择合适的特征并提取，对数据进行独热编码、标准化等处理后，选择合适的回归算法，构建并训练预测模型，运行输出预测的拟合曲线、错误率和均方根误差RMSE。

# 4 特征数据分析

## 4.1 数据分析

乘车刷卡数据表（gd\_train\_data）如表1所示

表1 乘车刷卡数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 数据类型 | 字段含义 | 数据示例 |
| use\_city | object | 使用地 | 广州 |
| line\_name | int64 | 线路名称 | 281 |
| terminal\_id | object | 刷卡终端ID | 4589bb610f9be53a43a7bc26bb40e44d |
| card\_id | object | 卡片ID | 8ce79e0b647053f191d20c5552eb49f0 |
| create\_city | object | 发卡地 | 佛山 |
| deal\_time | int64 | 交易时间 | 2014091008 |
| card\_type | object | 卡类型 | 学生卡 |

天气状况数据表（gd\_weather\_report）如表2所示

表2 天气状况数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 数据类型 | 字段含义 | 数据示例 |
| data | object | 日期 | 2014/8/12 |
| weather | object | 天气状况（白天/夜间） | 大雨/中雨 |
| temperature | object | 气温（最高/最低） | 36℃/26℃ |
| wind\_direction\_force | object | 风向风力（白天/夜间） | 北风3-4级/无持续风向≤3级 |

## 4.2 数据处理

将原始数据进行预处理后可以对数据分析更加方便，有利于模型的建立，提高预测准确率[4]。

查看原始数据发现，线路共有两条（线路281和线路565），有200多万用户刷卡，刷卡信息是约800万条左右，将数据按2条线路各个时间段的客流整理之后，约7200条左右。

异常数据处理：异常值是指数据集中存在的个别值与其它数值明显相差距离过大，其存在影响正常的数据分析 [5]。通过统计每小时客流量，画出箱形图，计算出每小时客流量的上下限，根据上下限去掉异常数据。

## 4.3 特征分析

（1）天气状况：分别画出白天和夜间的天气状况与出现次数的关系柱状图，如图1所示，x轴是白天的天气状况，y轴是天气状况出现的次数，如图2所示，x轴是夜间的天气状况，y轴是天气状况出现的次数，可以发现不论是白天还是夜间，天气状况为多云和晴时出现的次数较多，天气状况为雨天和霾时出现的次数较少。

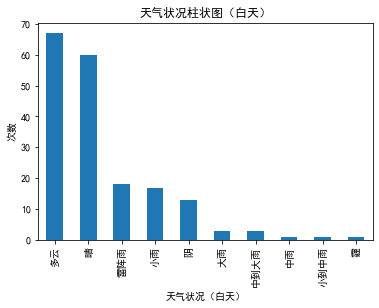


图1 天气状况柱状图（白天）

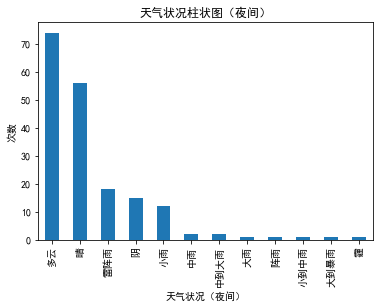


图2 天气状况柱状图（夜间）

通过分析，将天气状况划分为3个等级，并对天气状况进行编码，方便后续预测分析：多云、晴、阴和霾取值为0，雷阵雨取值为1，小雨、小到中雨、中雨、中到大雨和大雨取值为2。

计算单日白天和夜间天气变化的差值，并画出单日天气变化与出现次数的关系柱状图，如图3所示，x轴是单日天气变化的差值，y轴是出现次数，可以发现变化差值为0时出现的次数较多，而变化差值为1和2时出现的次数较少。

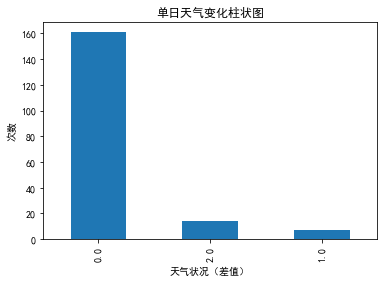


图3 单日天气变化柱状图

通过对天气状况的分析，选择白天天气状况、夜间天气状况和天气状况变化差值作为预测特征值。

（2）温度：分别画出单日最高温度和最低温度与出现次数的关系折线图，如图4所示，x轴是单日最高温度，y轴是最高温度出现次数，可以发现最高温度在30度-35度时出现的次数较多，在小于15度或高于35度时出现的次数较少，如图5所示，x轴是单日最低温度，y轴是最低温度出现次数，可以发现最低温度在23度-26度时出现的次数较多，在低于5度时出现的次数较少。

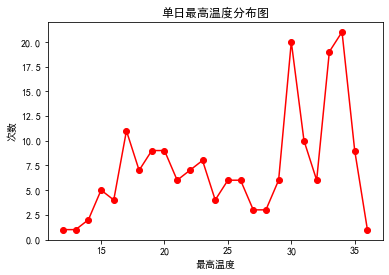


图4 单日最高温度分布图

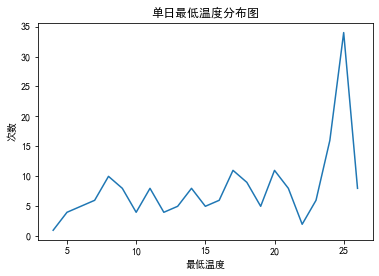


图5 单日最低温度分布图

计算单日最高温度和最低温度的平均值，并画出单日平均温度与出现次数的关系柱状图，如图6所示，可以发现平均温度在28度-30度时出现的次数较多。

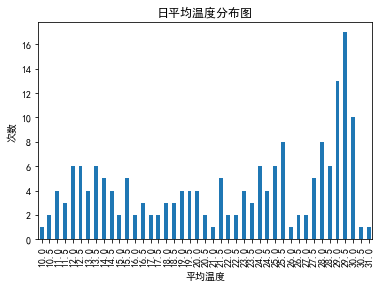


图6 日平均温度分布图

通过对单日温度的分析，选择最高温度、最低温度和平均温度作为预测特征值。

（3）风向风力：分别画出白天和夜间的风向风力与出现次数的关系柱状图，如图7所示，x轴是白天的风向风力，y轴是出现的次数，如图8所示，x轴是夜间的风向风力，y轴是出现的次数，可以发现不论是白天还是夜间，风向风力为无持续风向≤3级时出现的次数较多，而其他风向风力则出现的次数较少。

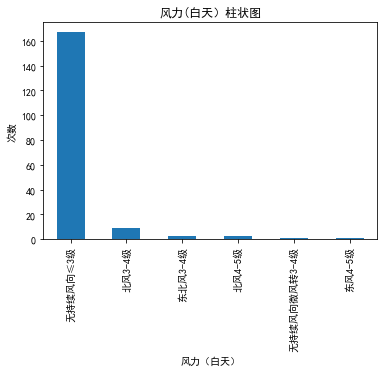


图7 风力（白天）柱状图

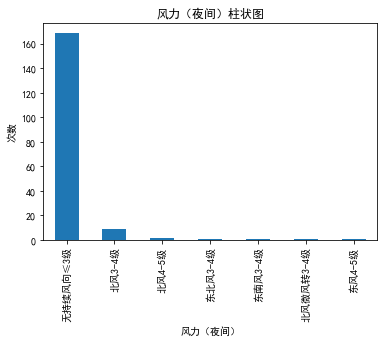


图8 风力（夜间）柱状图

通过分析，将风向风力划分为3个等级，并对天气状况进行编码，方便后续预测分析：无持续风向≤3级取值为0，无持续风向微风转3-4级、北风微风转3-4级、北风3-4级、东南风3-4级和东北风3-4级取值为1，东风4-5级和北风4-5级为取值2。

计算白天和夜间的风力平均值和差值，画出平均风力和风力变化与出现次数的关系柱状图，如图9所示，x轴是单日平均风力，y轴是出现的次数，可以发现平均风力为0时出现的次数较多，其他平均风力则出现的次数较少，如图10所示，x轴是单日风力变化，y轴是出现的次数，可以发现风力变化为0时出现的次数较多，风力变化为1和2时出现的次数较少。

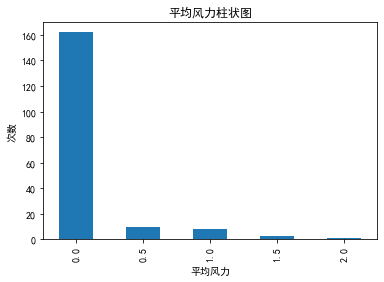


图9 平均风力柱状图

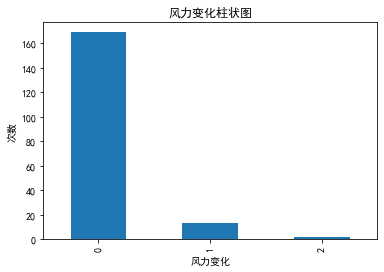


图10 风力变化柱状图

通过对风向风力的分析，选择白天风力、夜间风力、平均风力和风力变化差值作为预测特征值。

将数据中的时间划分为日期与小时两部分，方便分析数据。

（4）小时：画出小时与乘车人数的关系柱状图，如图11所示，可以发现在7点到10点，16点到18点乘车人数较多，而0点到5点，22点到23点乘车人数则很少，近乎没有，所以主要分析6点到21点的数据，选择小时作为预测特征值。

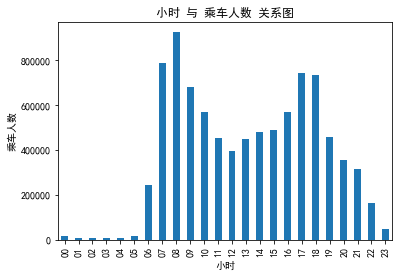


图11 小时与乘车人数关系图

（5）是否为节假日：画出日期与乘车人数的关系折线图，如图12所示，可以发现在国庆假期期间（20141001-20141007）乘车人数有明显减少，选择是否为节假日作为预测特征值。

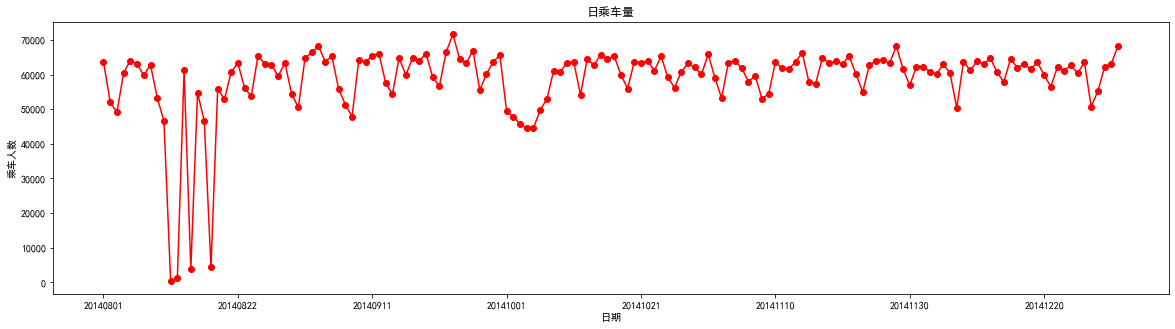


图12 日乘车量

（6）星期：通过对日期数据转换，添加星期字段（星期天是0），画出星期与乘车人数的关系柱状图，如图13所示，可以发现工作日乘车人数较多，休息日乘车人数较少，选择星期作为预测特征值。

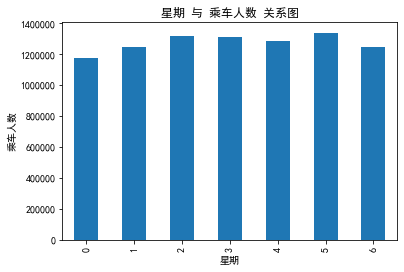


图13 星期与乘车人数关系图

预测模型不仅对数据进行总结概括，而且要通过分析得到客观规律，研究发现，在构建模型时，如果选择一些相关度不高的变量，模型的预测准确率会下降[6] 。一旦选择无关的特征变量，不仅干扰对变量之间关系的认识，还可能在后续的实际应用中浪费资源，造成损失。

通过分析，最后选择天气状况（单日白天天气状况，单日夜间天气状况，单日天气状况变化差值），温度状况（单日最高温度状况，单日最低温度状况，日平均温度状况），风力（单日白天风力状况，单日夜间风力状况，日平均风力状况，单日风力状况变化差值），小时，是否为节假日和星期共13个特征。

## 4.4 特征提取

将小时、星期独热编码（One-Hot Encoding）处理：独热编码，又称一位有效编码，编码后在不同时间的每一个特征值对应不同的0或1编码，且同一时间有且只有一个特征值可以有效。

如图14所示，是小时的独热编码实现过程

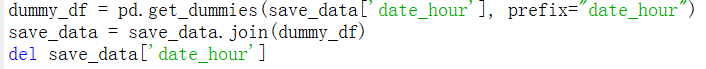


图14 小时的独热编码的实现代码

计算每个特征值和客流量的皮尔逊系数

如图15所示，是计算皮尔逊系数的实现代码



图15 计算皮尔逊系数的代码

汇总特征：通过左连接，将所有特征连接

如图16所示，是合并特征的实现代码



图16 合并特征的代码

# 5 回归算法

回归就是为了使预测的数据更加贴近实际情况，通过不断的实验分析而进行的预测工作。

线性回归(Linear Regression)是通过对一个或多个自变量和因变量的关系的分析，求出线性回归方程并建立模型的一种回归分析[7]，是通过分析大量的数据得到回归系数，构建比较符合实际情况的回归模型，对未来的发展情况进行预测。

公交客流量比较符合多元线性回归模型的特征提取简单，模型方便理解等特点，所以本次毕业设计选择线性回归算法进行预测。

## 5.1 线性回归模型的一般形式

y是预测值，x是特征值，w是回归系数，a是截距，b是误差

## 5.2 拟合

拟合指的是建立的模型能够较好的预测实际情况[8]。

1. 欠拟合：模型比较简单，没有充分分析样本数据，选择足够的特征，使得模型在训练集上的预测效果不是很好，训练集的预测值与真实值之间的误差较大。
2. 过拟合：模型比较复杂，对样本数据分析过度，选择的特征中存在特殊特征，使得模型在训练集上的预测效果非常好，但是在未知数据上的预测效果不好，训练集的预测值完全贴合真实值。

## 5.3 python实现线性回归

（1）初始化：回归系数初始化为等于0的矩阵（行数为1，列数为特征个数），截距初始化为0，初始化迭代次数和学习率。

1. 梯度下降法：求回归系数和截距，迭代找到参数的局部最小值

如图17所示，是梯度下降法的流程图

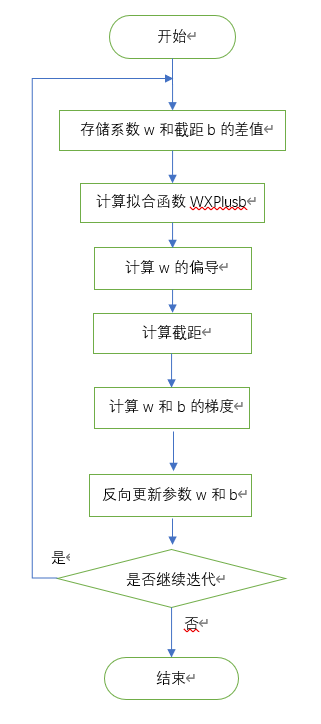


图17 梯度下降法的流程图

## 5.4 回归评价指标

（1）错误率：显示预测效果

如图18所示，是计算错误率的实现代码

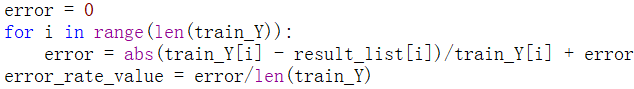


图18 计算错误率的实现代码

（2）均方根误差RMSE（Root Mean Square Error）：是预测值和真实值之间的样本标准差。可以显示样本的离散程度。

如图19所示，是计算RMSE的实现代码



图19 计算RMSE的实现代码

# 6 实验

线性回归预测模型是利用线性回归分析算法，通过分析原始数据构建回归模型，对未来进行预测[9]。

（1）按线路分割，构建两个模型：线路281模型和线路565模型。

（2）切割数据：为了训练并测试模型，将原始数据进行切割，2014年12月25日前的数据作为训练集，2014年12月25日到2014年12月31日这一周的数据作为测试集。

如图20所示，是线路281切割原始数据集的实现代码



图20 切割原始数据集的实现代码

1. 将星期、小时、温度、风力等特征规范化：z-score规范化：先计算均值和方差并保存，然后减其均值并除其方差。可以消除特征之间的差异性，使不同特征对参数的影响程度一致，即使得数据标准统一化，提高了数据可比性，削弱了数据解释性。

如图21所示，是线路281训练集特征规范化的实现代码

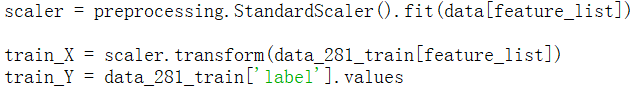


图21 训练集特征规范化的实现代码

1. 调用线性回归算法，创建模型

如图22所示，是建立模型的实现代码



图22 建立模型

1. 训练模型

如图23所示，是训练模型的实现代码



图23 训练模型的实现代码

1. 预测：通过对模型的训练，输出训练集训练后的拟合曲线、错误率和RMSE
2. 测试：通过对测试集的测试，输出测试后的拟合曲线、错误率和RMSE

# 7 运行结果显示

启动图形用户界面（Graphical User Interface）窗体，分别在线路281和线路565下输入预测的时间段，点击预测按钮，查看不同线路的拟合曲线和预测的错误率和RMSE。

如图24所示，分别是预测线路281线路565在20141225-20141231预测的拟合曲线、错误率和RMSE

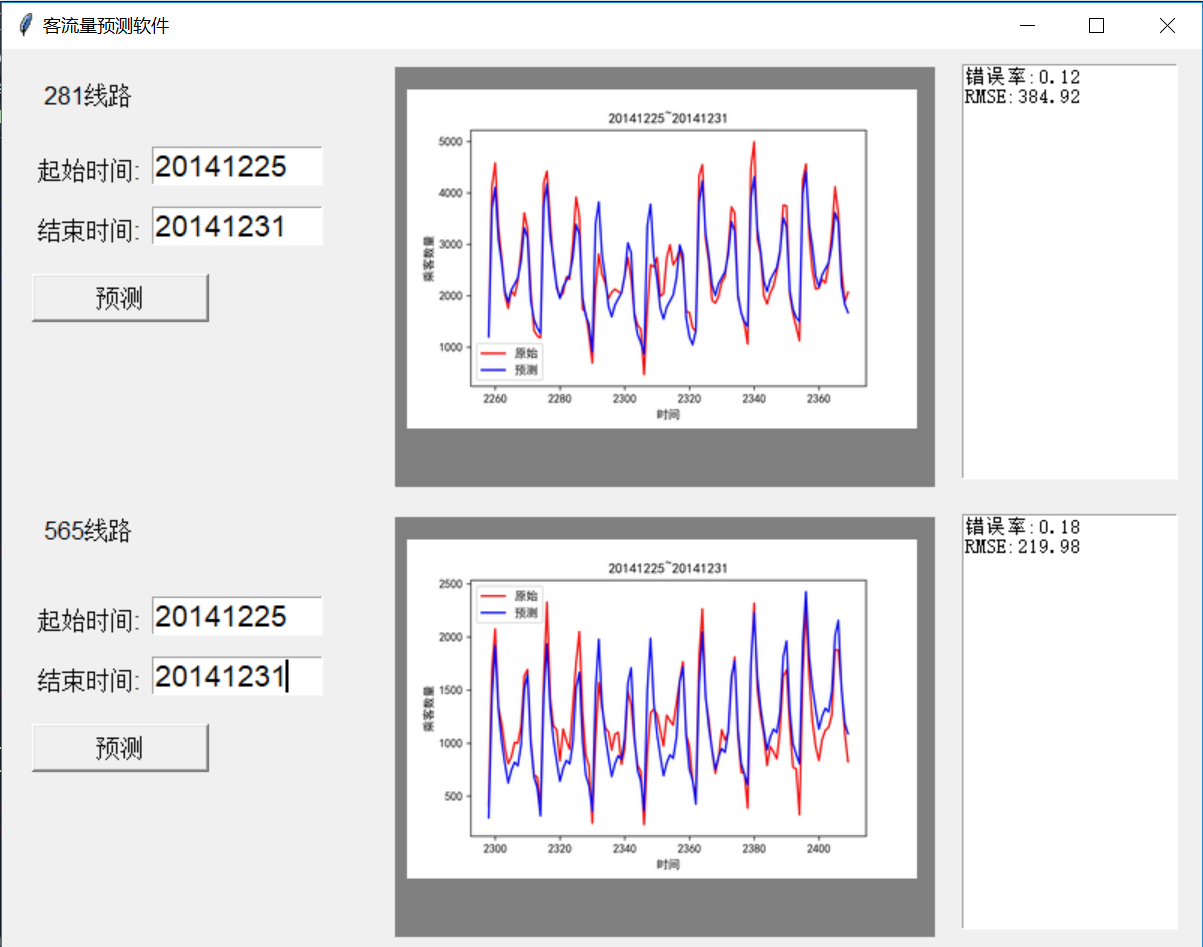


图24 预测结果

# 8 总结

在经过近三个月的设计开发后，学习并应用了python的各类功能，熟悉掌握并了运用Spyder开发环境，选择并处理了合适的特征，构建模型，对线性回归算法有了更深的认识和运用，完成了对公交客流量的预测。

本次毕业设计还有一些不足之处没有解决，如果特征的选择可以更加多样，可能对预测模型的影响有正有负。另外，特征的提取和处理还是可以更加细致。相信在今后的学习中，经过修改可以完善模型，提高预测的准确率。

# 9 致谢

在本次毕业设计中，不论是选择题目、查找资料，还是设计开发、编写论文，都非常感谢李艳玲老师对我的帮助和指导，为我提供了许多专业性的建设，并解决了许多专业性的问题，同时，也十分感谢给予了我帮助的我的同学们。本次毕业设计不但锻炼了我的编程能力，而且改变了我思考问题的思维方式，为我的能力带来了很大的提升，感谢曾经努力学习的大学四年时光。

# 参考文献

[1] 陈奇平.基于数据挖掘的公交线路客流预测研究[D].海南大学，2018

[2] 耿学贵.基于IC卡的公交客流时间规律研究[D].昆明理工大学，2012

[3] 钟同圣,卫丰,王鸷，等.Python语言和ABAQUS前处理二次开发[J].郑州大学学报(理学版)，2006(01) :65-69

[4] 罗强，何利力，王晓菲.数据仓库中数据清洗技术分析机[J].电脑编程技巧与维护，2015

[5] 郑家亨.统计大辞典[M]：中国统计出版社，1995

[6] 王大荣,张忠占.线性回归模型中变量选择方法综述[J].数理统计与管理，2010

[7] 叶锋.多元线性回归在经济技术产量预测中的应用[J].中外能源，2015，020(002)：45-48.

[8] 王征韬.深度神经网络压缩与优化研究[D].电子科技大学,2017.

[9] 薛付忠,王洁贞,马希兰.线性回归预测模型有效超前期的确定方法[J].中国卫生统计,1999,016(003):155-156

**Prediction of Bus Line Passenger Flow Based on Regression Algorithm**

College of Computer Science and Technology 2016 Computer science and Technology (non normal) Bai Lu 20161102910

Directed by Li Yanling Associate Professor

**Abstract** This graduation project mainly uses python programming language, through the use of PyCharm development environment, after selecting appropriate features and processing，build and train the model, and realize the bus line passenger flow prediction based on regression algorithm.The purpose is to accurately predict the passenger flow of public transport in the future, so as to guide the public transport company to make reasonable vehicle scheduling.

**Keywords** Python; PyCharm; Regression algorithm