**《数据分析原理》**

**课程设计报告**

|  |  |  |
| --- | --- | --- |
| **组长** | 姓名 | 王瑞 |
| 学号 | 5120194818 |
| **组员** | 姓名 | 黄婧宇 |
| 学号 | 5120196267 |
| **组员** | 姓名 | 王梦林 |
| 学号 | 5120194745 |
| **组员** | 姓名 | 周子琳 |
| 学号 | 5120194441 |

**指导老师: 骆蓉**

**2021年12月**

**基于多模型比较的二手房价格评估研究**

**摘要：** 随着经济的发展,我国伴随房地产的相关经济活动越来越频繁,由于城市可供开发的土地越来越少,二手房的交易也变得更加活跃,对房地产估价的需求也随之增大。无论是从市场参与者的角度,还是从国家开征税费的角度看,二手房价格的精确衡量都是个永恒的话题。目前使用的市场比较法、成本法和收益法三种传统评估方法在实际评估运用中过多依赖于评估者的经验,对数理模型运用较少,在实际应用中成本较高。同时，成都，四川省省会，也是西南地区经济、文化、交通中心之一，自古有“天府之都”的美誉。近年来，成都在经济、交通等多个方面都发展快速。同时，成都也以其包容、开放的特点吸引着来自全球各地的游客。由于成都地理位置优越，且气候适宜，越来越多的人选择在这座城市工作、生活，这也就导致了住房问题的涌现。在利用特征价格理论构建二手房评估模型过程中,多使用多元线性回归方法进行回归预测,函数形式的选择对评估的效果影响较大,却常依赖于人为的线性假定,容易造成较大的误差。本文引入随机森林方法、决策树方法、K近邻回归等对特征价格模型进行回归预测,建立二手房价格评估模型，对爬取的成都市片区的近2万5千多条特征价格数据,进行建模与预测。最后采用了常用的多元线性回归方法、决策树方法、K近邻回归与随机森林回归方法等方法建立的预测模型进行对比研究,表明基于随机森林方法的二手房评估模型的预测结果有着较高的准确性,值得推广应用。

**关键词：**二手房 房价预测 随机森林 决策树 K近邻回归

目录

[一、数据分析目标与任务 1](#_Toc19104)

[（一） 背景介绍 1](#_Toc16083)

[（二） 课程设计目标与任务 1](#_Toc2600)

[（三） 研究方法与技术路线 2](#_Toc19401)

[二、数据预处理 2](#_Toc2489)

[（一） 数据说明 2](#_Toc15973)

[（二） 数据清洗 3](#_Toc12315)

[（三） 数据处理 5](#_Toc23101)

[三、数据探索分析 6](#_Toc4790)

[（一） 数据探索性分析 6](#_Toc26269)

[（二） 可视化呈现结果 9](#_Toc19750)

[（三） 结论 18](#_Toc9282)

[四、数据分析模型 19](#_Toc10872)

[（一）模型建立 19](#_Toc12838)

[（二）模型评估 28](#_Toc21852)

[（三）可视化结果 29](#_Toc24942)

[五、方案评估 31](#_Toc11976)

[附录 3](#_Toc14838)6

# **一、数据分析目标与任务**

## 背景介绍

1. 业务背景

由于随着社会的不断发展，人民生活水平的不断提高，导致房价一直呈上升趋势。在当代社会，越来越多的人反应买不起房的问题。

链家网是集房源信息搜索、产品研发、大数据处理、服务标准建立为一体的国内领先且重度垂直的全产业链房产服务平台。作为全国二手房市场的领头羊，链家网拥有全国最大的真房源数据库，收录了全国20多个重点城市6000万套房源信息，使用户能够在线精准定位，找到满意的房屋，并快速匹配经纪人完成看房、交易等环节，真正帮客户做到省时省力的安心服务。

1. 研究意义

通过爬取链家网中成都地区二手房的相关信息，对不同朝向，不同地区等的二手房的房价进行分析并建立模型，从而实现对二手房房价的预测。通过该模型，也能够使用户更好地了解满足自己需求的二手房的价格区间，使选房更加方便。

## 课程设计目标与任务

1. 目标
2. 利用可视化图像进行数据分析，得出影响房价的因素等信息。
3. 利用数据分析模型对二手房房价进行预测。
4. 任务
5. 从链家网中爬取成都各区的二手房的数据。
6. 对爬取的数据进行数据清洗，得到需要的数据。
7. 利用数据绘制图像，通过可视化形式展示数据，并对数据进行分析。
8. 通过数据分析模型对二手房价进行预测。

## 研究方法与技术路线

1. 研究分析方法

首先使用爬虫对网站数据进行爬取，再对爬起的数据进行数据清洗，其中包括去除重复值，缺失值处理，异常值处理；再将处理后的数据进行可视化展示，探究装修情况、所在楼层、不同楼层直接是否配备电梯、建筑类别、房屋用途、建筑结构、各个区域等多条件和二手房的关系以及每个区单位面积关系，房价分析、单价、数量、总价和房屋所属市辖区之间的关系 ，房屋面积和总价关系、房屋面积和总价、房屋朝向和总价的关系等，并对并对其进行分析得到探索性数据分析的结果，最后，使用数据分析模型进行数据结果预测包括对应论文实现相关随机森林算法预测二手房房价，包含随机森林、决策树、K邻近模型构建、训练、预测等操作，并对给出的买房数据要求给出合适的预测价格。

1. 研究编程语言

Python 3.8

1. 开发环境
2. VSCode 1.63.2
3. Pycharm Community Edition 2020.1.2
4. Jupyter

# 二、数据预处理

## 数据说明

|  |
| --- |
| **（一）数据说明** |
| 1. 数据规模   通过爬虫技术爬取了成都的成华区、龙泉驿区等20个区的二手房数据，合计23850条数据。在数据中保留了单价、房屋所属市辖区、房屋地址（街道）等15个属性。   1. 数据文件和字段   将所有地区的所有属性数据写入data.csv文件，在对数据进行数据清洗时仅读取单价, 房屋所属市辖区， 房屋地址（街道），房屋户型，所在楼层，建筑面积，户型结构，建筑类型，房屋朝向，建筑结构，装修情况，配备电梯，挂牌时间，交易权属，房屋用途的相关列。其中单价的单位为百元，建筑面积的单位为平方米。 |
| **（二）数据样本展示** |
| 图 1 数据样本展示 |

## 数据清洗

|  |
| --- |
| **（一）重复值处理** |
| 利用drop\_duplicates对数据中的重复值进行删除，并保留最后一个重复行（即设置keep值为“last”），在原数据集上作出修改。数据中包含2417条重复数据，通过输出去除重复行前的数据维度和去除重复行后的数据维度，对比看出是否成功删除重复行。  去除重复行前：    图 2 去除重复行前  去除重复行后：    图 3 去掉重复行后 |
| **（二）缺失值处理** |
| 首先利用isnull查看数据每列是否含有空数据。通过函数将数据集中标明“暂无数据”的信息置为空，查看空数据数量为6595条，再使用dropna对数据中的所有空数据进行去除。通过查看去除重复值后的数据中是否还有列包含空值，判断是否成功去除缺失值。去除缺失值后数据维度为（14838,15）。  去除缺失值前：    图 4 去除缺失值前  去除缺失值后：    图 5 去除缺失值后  去除缺失值后的数据维度：    图 6 去除缺失值后的数据维度 |
| **（三）异常值处理** |
| 定义error\_data函数将数据中单价和建筑面积小于零的数据置为空，并对其进行去除。由于该数据集中无单价和建筑面积小于零的异常值，所以处理异常值前后，数据的维度未发生改变。    图 7 异常值处理完毕 |

## 数据处理

|  |
| --- |
| **（一）特定数据提取** |
| 数据只保留了单价, 房屋所属市辖区， 房屋地址（街道），房屋户型，所在楼层，建筑面积，户型结构，建筑类型，房屋朝向，建筑结构，装修情况，配备电梯，挂牌时间，交易权属，房屋用途的相关列。 |
| **（二）数据变换** |
| 1. 由于原始数据中的单价的单位为百元，与常用价格单位不同，所以使用apply将单价的单位修改为元。     图 8 修改元单位   1. 原始数据中建筑面积列中包含了单位，不利于后续数据的使用，利用apply保留建筑面积的数字部分。     图 9 修改数字部分 |
| **（三）难点** |
| 在对各地区数据进行合并后，读取总数据。由于数据中某些列中包含了“，”，而csv文件中的间隔符也为“，”，所以导致读取数据时，无法正确读取。当设置sep为“\t”时，读出的数据每一行为完整字符串，不便于后续操作。所以在读取数据时，使用usecols属性设置所需要读取的列，避免无关信息对数据的读取造成影响。    图 10 读取数据 |

# 三、数据探索分析

## 数据探索性分析

|  |
| --- |
| **（一）数据方面** |
| 1. 对各个不同的管辖区的二手房的单位面积的价格与所属管辖城区进行单变量分析 2. 对各个不同的管辖区的二手房的单位面积的价格均值、总价均值、房屋数量进行单变量分析 3. 对二手房的建筑面积和房屋总价进行单变量分析 4. 对二手房的房屋户型和各个户型的房屋单价的均值进行单变量分析 5. 对二手房的房屋朝向和各个朝向的房屋总价的均值进行单变量分析 6. 探究装修情况二手房的关系   按照装修情况对数据进行分组，之后分别对其数目求和并求均值后对其进行排序，得到我们每种装修情况的数据，以及各种装修情况与房间单价的均值。      图 11 数据格式   1. 探究所在楼层和房间单价的关系   按照所在楼层对数据进行分组，分为低楼层、中楼层、高楼层以及地下室四种类别，之后通过得到各个楼层种类的平均值并对其进行排序。    图 12 所在楼层平均值   1. 探究不同楼层间是否配备电梯和二手房的关系   对中低楼层和高楼层进行分类，之后按照是否配备电梯对数据进行分组，分为有电梯和无电梯两种类别，分别统计中低楼层和高楼层是否有电梯。    图 13 电梯数据   1. 探究建筑类型与二手房价格的关系   按照建筑类型对数据进行分组，分为板塔结合、板楼、塔楼和平房四种类别，之后得到各个建筑类型的中位数并对其进行排序。    图 14建筑类别   1. 探究房屋用途与二手房价格的关系   按照建筑类型对数据进行分组，分为商业办公类和普通住宅两种类别，之后得到各个建筑类型的平均值。    图 15 房屋用途与二手房价格的关系   1. 探究建筑结构与二手房价格的关系   按照建筑结构对数据进行分组，分为钢混结构、钢结构、框架结构、混合结构 、砖混结构、未知结构和砖木结构7种类别，之后得到各个建筑结构单价的平均值并对其进行排序。    图 16 建筑价格数据分析   1. 探究各个区域与二手房数量的关系   分别对每个地区的二手房数量进行统计，得到画图的数据。      图 17 地区与二手房数量数据分析 |
| **（二）可视化方面** |
| 1. 箱型图：根据各个不同的管辖区的二手房的单位面积的价格绘制箱线图进行分析。从中大致了解到各个区域的房屋单价的分布情况 2. 折线图：根据二手房的单位面积价格、总价、房屋数量的均值数据分别与管辖区绘制柱状图和折线图进行分析。可以从中得到具体某个区的房屋数量、房屋单价、总价的价格最大值，从而进行对比得出结果。 3. 散点图：根据二手房的建筑面积和房屋总价绘制散点图进行分析。可以从图中得到建筑面积和房屋总价之间的关系 4. 雷达图：根据二手房的房屋朝向和房屋总价绘制雷达图进行分析。可以从图中得到朝向与房屋总价的关系即分析出某一个朝向的房屋价格较高，哪一个朝向较低。 5. 柱状图：通过柱状图来反应数据属性与二手房单价之间的关系。根据二手房的户型和房屋单价绘制柱状图进行分析。可以从图中得到户型与房屋价格的大致关系即分析出某一种户型的房屋价格较高，某一个户型较低。 6. 饼图（南丁格尔玫瑰图）：通过饼图（南丁格尔玫瑰图）来反应各个数据属性与二手房数量之间的相关关系。 7. 漏斗图：通过漏斗图来反应建筑结构与二手房单价之间的关系。 8. 热力图：通过热力图来反应不同楼层配备电梯的数量关系。 9. 地图：通过地图来反应各个地区的二手房数量，查看哪个地区的二手房数量更多。 |

## 可视化呈现结果

|  |
| --- |
| **（一）每个区单位面积房价分析** |
| 1. **数据说明**   使用数据集中的单价数据进行分析。首先根据房屋所属市辖区进行分组，分组完成之后根据单价数据进行绘制箱线图。采用箱线图可以得到数据的最大值、最小值、中位数、四分之一分位数、四分之三分位的大小，从而能够准确稳定地描绘出数据的离散分布情况。   1. **可视化界面（箱线图）**     图 18 箱线图   1. **结果分析**   从上图可以看出，20个城区中高新区和天府新区房价的最大值在所有数据中是最大的，高新区超过5万每平方米，锦江区、武侯区、青羊区、天府新区四个区域的房屋单价超过4万每平方米，其余的城区都在1万到3万每平方米。所有城区的最小值都在1万每平方米。高新区和天府新区的房屋单价的中位数值在2万到3万每平方米，金堂区和彭州区的房屋单价中位数在1万以下，其余的城区都在1万到2万之间。利用箱线图的特点判断单价数据的偏态和尾重，可以得到彭州区、金堂区、青白江区、简阳区的单价分布相对集中变化幅度不大，在1万到1.5万之间，而锦江区、武侯区、高新区、青羊区、天府新区的单价分布发散价格变化幅度大，从1万到3万之间。这些价格变化可能是受到房屋的地理位置、大小、户型、朝向等的影响变化而变化。仅根据单价分析可以大致看出高新区和天府新区的房价较高，金堂区和彭州区房价较低。 |
| **（二）探究单价、数量、总价和房屋所属市辖区之间的关系** |
| 1. **数据说明**   本次涉及到的数据是房屋所属市辖区和单价。首先根据房屋所属市辖区进行数据分组，对各个城区的房屋单价进行求解单价平均值、求解各个地区的房屋数量以及各城区内各个房屋的总体房价，然后在此基础上进行绘图分析单价、数量、总价分别和房屋所属市辖区的关系。采用的是柱状图   1. **可视化界面（柱状图+折线图）**     图 19 单价和地区    图 20 总价和地区    图 21 数量和地区   1. **结果分析**    1. 由上面箱线图可以得到大致单价均值最高是天府新区和高新区，均值最低的是金堂区、彭州区、青白江区。现在可以从图2可以准确得到高新区房屋单价均值最高为23663元/平方米，其次是天府新区为23268元/平方米；单价均值最低的后三位分别是彭州区8097元/平方米、青白江区7465元/平方米和金堂区6534元/平方米。可以从图中可以看出我们分析的成都的20个城区中，大部分的房价均值都在1万到2万每平方米。    2. 由图3可以得到，天府新区的房屋总价均值最高，其次是高新区，价格都在三百万以上，属于高价格地区；彭州区、金堂区、青白江区三个地区的房屋总价均值都在一百万以下，属于低价格地区。其余地区都在一百万到三百万范围内，属于中等偏高价格地区。    3. 由图4可以看出，金牛区房屋总数最多，都江堰区房屋总数最少。可以看到单价以及总价较高的地区如天府新区，它现存出售二手房数量较少；而成华区、金牛区、锦江区、高新区现存出售二手房较多，可能是因为所在地段的地理位置以及商业发展等因素导致的。 |
| **（三）探究房屋面积和总价的关系** |
| 1. **数据说明**   本次涉及到数据是房屋的建筑面积以及总价。首先根据房屋的单价和建筑面积算出房屋总价的数据，然后根据房屋总体建筑面积和总价进行绘图分析，采用的是散点图。   1. **可视化界面（散点图）**     图 22 面积和总价\_散点图   1. **结果分析**   由图5可以看出，横坐标为建筑面积，纵坐标为房屋总价。可以看出房屋面积都集中在300平方米以下，极少数在300平米以上更有甚者超过700平米；房屋价格大多在四百万以下，极少数大于四百万但有的高达八百多万。故面积和房价大致上呈正相关的关系，当面积越大时，总价越高，建筑面积较小时，房屋总价相对低。 |
| **（四）探究户型和单价的联系** |
| 1. **数据说明**   本次涉及到的数据是房屋户型和总价。首先根据房屋户型进行分组，得到相应的房屋户型组及其单价，然后根据分组后的单价数据进行求解平均值，最后根据所得数据进行绘图分析。本次采用柱状图。   1. **可视化界面（柱状图）**     图 23 户型和单价\_柱状图   1. **结果分析**   由图6可以看到横坐标为房屋的户型分为X室X厅X厨X卫，纵坐标为房屋的平均单价。可以看出户型为5室2厅1厨6卫、5室2厅1厨8卫的单价最高分别为142000元、110800元，都超过10万每平米。户型为2室0厅1厨2卫的单价最低，为3600元。可以从图中那个看到户型中厨房和卫生间的数量都较小时或者卫生间数量为0或为1时房屋的单价较低，一般为2万以下。单价在5万及以上的户型的卫生间都比较多，一般为4个及以上。 |
| **(五)探究房屋朝向和总价的关系** |
| 1. **数据说明**   本次涉及到数据是房屋朝向和总价。首先根据数据集中的房屋朝向数据进行处理因为存在“南北”、“东西”朝向不符合现实情况，然后根据房屋朝向进行分组，最后对分组后的数据进行求解总价均值继而进行绘图分析房屋朝向和总价的关系。本次采用的是雷达图。   1. **可视化界面（雷达图）**     图 24 朝向和总价\_雷达图   1. **结果分析**   由图7可以看出八个方位的房屋朝向的房价均值相差不大。其中价格最高为房屋朝南1916450.36元，价格最低为朝向西边1549731.8元。其他朝向都在160万到190万之间。由上图可以得到房屋朝向为南的房价是最高的，其次是朝向为北。 |
| **（六）探究装修情况和二手房的关系** |
| 1. 探究装修情况和二手房的关系     图 25 装修情况数目与二手房的关系（南丁格尔图）    图 26 装修情况与房屋单价的关系  分析：从图表中可以看出，在我们的南丁格尔图中，表示了各个种类的房屋所占的比例，可以发现，有接近一般的房屋均为精装房。而在各种装修情况的房屋中，可以发现房屋的单价相差并不是特别大，但是仍然可以发现精装房的房间单价是比其余种类的房间的价格更高的。 |
| **（七）探究所在楼层与二手房的关系** |
| 图 27 所在楼层与二手房的关系  分析2：从图表中可以发现，所在楼层的数量分布较为均匀，大多集中在30%多，相比之下，售卖地下室的商户几乎不存在，在中高低楼层中，最多的是高楼层的房屋，其次是中楼层最后是低楼层的房屋。从各楼层的价格也可以看出，地下室的价格普遍比较便宜，而中高低楼层的价格大多集中在1万4千元左右。 |
| **（八）探究不同楼层直接是否配备电梯与二手房的关系** |
| 图 28 不同楼层是否配备电梯与二手房的关系  分析3：从图表中可以发现，中低层是否配有电梯的比例相差不大，而相比之下，高楼层配有电梯的数目较多。 |
| **（九）探究建筑类别与二手房价格的关系** |
| 图 29 建筑类别与单价的关系  分析4：通过图表可以发现，在各个建筑类别中，平方的单价是最低的，可能平方的建筑成本较低或是买房的顾客不太青睐平房这样的建筑，其余三个建筑类别的价格相差不大，但板塔结合的价格是最高的，板楼其次，塔楼最后。 |
| **（十）探究房屋用途与二手房单价的关系** |
| 图 30 房屋用途与二手房单价的关系  分析5：通过图表可以发现，所售卖的房屋几乎都是用来商业办公或者作为普通住宅的，而相比之下普通住宅的房屋单价贵了很多，是商业办公类房屋的3倍不止。 |
| **（十一）探究建筑结构与二手房单价的关系** |
| 图 31 建筑结构与二手房单价的漏斗图  分析6：通过图表可以发现，不同的建筑价格对二手房的单价影响较大，其中钢材相关的结构价格相对较高，而砖混等结构相比之下价格要低上很多。 |
| **（十二）探究各个区域与二手房数量的关系** |
| [MY[YL]KQ[4@~6Y0LW(FCT3  图 32 地图  分析7：通过图表可以发现，出售二手房数量较多的锦江区、金牛区、武侯区等区县，这些都是成都的中心城区。 |

## 结论

|  |
| --- |
| **总结** |
| 1. 成都各区的二手房单位面积价格、总价最高的都是高新区和天府新区，价格较低是在彭州区、金堂区、青白江区 2. 成都各区的二手房数量最多的是金牛区、成华区、锦江区、高新区，数量最少的是都江堰区 3. 成都二手房建筑面积与房屋总价总体呈正相关关系，并且面积大多分布在200平米到300平米，价格在四百万以下 4. 成都的二手房的户型中卫生间和厨房数量都比较小的时候或者卫生间数量为0或者1的时候，房屋的单价都比较低，一般为2万每平方米以下，户型是4个及以上卧室和4个及以上卫生间的单价比较高，一般为5万每平方米及以上。 5. 成都的二手房的房屋朝向朝南时，总体价格均值最高，其次是朝北，价格最低的是朝西。但是各个朝向之间的价格差距不大，都在150万到190范围内。 6. 通过分析可得，装修情况、所在楼层以及建筑类别会对二手房的单价有一定的影响，但是相差并不是特别的大。相比之下精装房、低楼层的价格会高许多，这和房屋的成本有一定的关系，同时也和现代社会人群的生活习性有一定的关系。 7. 房屋用途以及建筑结构对二手房的影响较大，相比之下用于普通住宅的二手房价格较高，比用于商业用途的二手房贵了很多，说明大多数购房用户应该主要是购房用于安家的。而大多数的人家更加的青睐钢材的房屋，也许是和钢材的质量以及牢固程度更加好有关。 8. 通过地图可以发现，地图中颜色较深的是红橙色，而在地图中出现这样的颜色的有锦江区、成华区、金牛区等，这些都是成都的中心城区，说明大部分房屋都是属于成都的中心城区，人们也更加趋向于向发展更好的中心城区落户。 |

# 四、**数据分析模型**

## （一）模型建立

通过分析，我们主要采用了以下几种方式进行模型的建立：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Lasso线性回归 | RandomForestRegressor随机森林 | KNeighborsRegressorK近邻回归 | DecisionTreeRegressor决策树 |
| 训练集得分 | 0.77 | 0.97 | 0.74 | 0.89 |
| 测试集得分 | 0.76 | 0.86 | 0.7 | 0.66 |
| 平均绝对误差为 | 0.62 | 0.84 | 0.7 | 0.66 |

**选择的理由：**

1. **Lasso线性回归 ：**

Lasso回归的公式：



岭回归与Lasso回归最大的区别在于岭回归引入的是L2范数惩罚项，Lasso回归引入的是L1范数惩罚项，Lasso回归能够使得损失函数中的许多θ均变成0，这点要优于岭回归，因为岭回归是要所有的θ均存在的，这样计算量Lasso回归将远远小于岭回归。并且Lasso回归最终会趋于一条直线，原因就在于好多θ值已经均为0，而岭回归却有一定平滑度，因为所有的θ值均存在。

1. **RandomForestRegressor随机森林：**

RandomForestRegressor中文叫做随机森林回归模型，是一种常用的集成模型。并且通过验证，随机森林是多个决策树融合的集成模型，按照常理肯定会比单个决策树效果好。

使用随机森林回归模型可以取得较好的回归效果，比使用单个决策树回归模型提升了11%左右的准确率。

1. **KNeighborsRegressorK近邻回归：**

k近邻有分类也有回归，其实两者原理一样:定量输出是回归，进行预测比如明天的降水概率；定性输出是分类，需要定性的描述。kNN回归的原理：通过找出一个样本的k个最近邻居，将这些邻居的某个（些）属性的平均值赋给该样本，就可以得到该样本对应属性的值。所以也可以用来做预测。

1. **DecisionTreeRegressor决策树：**

基于树的学习算法被认为是最优秀的，也是使用最广泛的监督型学习方法。基于树的算法赋予预测模型高精度和易解释性。与线性模型不同，它们能很好地映射非线性关系。它们善于解决手头上的任何问题，包括分类和回归。例如Decision Trees，Random Forest，Gradient Boosting等方法，在各种数据科学问题中得到了广泛的应用。决策树是在分类问题中经常使用到的一种监督型学习算法。它适用于分类型以及连续性输入和输出变量。在这个方法中，我们根据输入变量中最显著的分裂点将总体或样本划分为两类或多类，同时决策树识别出最显著的变量以及取值，并给出效果最好的划分结果。

【结论】：通过不断的调参、实验，我们发现是【RandomForestRegressor随机森林】的预测效果是最好的，首先对于Lasso线性回归，得到结果的速度很慢，并且由于迭代次数大于1000之后，出结果就更慢了，误差也增大了；而对于KNeighborsRegressorK近邻回归来说，效果没有随机森林好，但基本得分相差不大；对于决策树回归来说，由于是树状结构，出结果很快，但是由于其树状结构，所以很容易不稳定，因此即使训练得分很高，但测试集得分很容易被max\_deep影响，并且平均绝对误差距离1还有一定距离。综上，我们选用随机森林的方式，齐次为Lasso线性回归和KNeighborsRegressorK近邻回归，最后为决策树回归模型。

下面给出其余几个的实验，调参过程：

|  |
| --- |
| **RandomForestRegressor随机森林调参实验过程** |
| 1. **拆分数据集**     图 33 拆分数据集  我们将数据拆分成训练集和测试集，测试集大小为原始数据集大小的 1/4 ，同时由于而预测值为高次关系，我们可以使用PolynomialFeatures生成高次的特征去更好拟合预测值，再使用fit\_transform进行数据变换，最终分成才测试数据和验证数据。   1. **测试 RandomForestRegressor 的预测性能随 n\_estimators 参数的影响**   通过绘制对应对应参数变化趋势折线图，具体    图 34 参数变化趋势折线图  得到的结果为 ：  `[~M2M8{ZI[(K7SQ9ZDONER  图 35 n\_estimators对应折线图  所以我们可以看出，设置具体的n\_estimators为20即可。   1. **测试 RandomForestRegressor 的预测性能随 max\_depth 参数的影响**   同理，我们测试对应的max\_depth对RandomForestRegressor 的影响，绘制对应的参数变化曲线，具体代码展示如下:    图 36 max\_depth代码  JCRT]JH1Y9$QVQ1UN3$@3M2  图 37 max\_depth对应折线图  所以我们将其设置为15即可。   1. **测试 RandomForestRegressor 的预测性能随 max\_features 参数的影响**   同理，接下来我们判断 max\_features对RandomForestRegressor 的影响，之后在判断的过程中，绘制对应参数变化图形。具体的代码展示：    图 38 max\_features对应代码  得到的结果如下：  7(ZQMFO{TQ7)(XYJM]%ZLQT  图 39 max\_features 对应折线图  可以使用参数为0.4左右即可。  综上，我们配置参数为n\_estimators=20, max\_features=0.4, max\_depth=15，得到最终的训练集得分：0.97，测试集得分：0.86，平均绝对误差为: 0.84  最终可视化结果为 ：    图 40 RandomForestRegressor随机森林结果 |
| **KNeighborsRegressorK近邻调参实验过程** |
| 具体的划分数据同上，更具对应的测试集和训练接，我们的得到如下的代码：    图 41 KNeighborsRegressorK近邻绘图代码  对应拆分之后，我们绘制对应的验证结果：  W~T1$E5PVQ7ZS@3VMQR_2YS  图 42 KNeighborsRegressorK近邻绘图结果  并且最终得到训练集得分：0.74，测试集得分：0.7，平均绝对误差为: 0.6998880615930219，我们定为0.7 |
| **Lasso线性回归调参实验过程** |
| 1. 分隔数据集   同上的数据分隔方式，也是分成了对应的1/4的测试集   1. 训练对应的模型   训练过程中，我们进行查找了对应的alpha :正则项系数，初始值为1，数值越大，则对复杂模型的惩罚力度越大。所以在这里，我们分别使用了为0.1和0.8的正则化系数，同时一般我们不使用默认的1。训练的代码如下 ：    图 43 Lasso线性回归代码   1. 测试结果   最终我们经过多次调参，最终将其确定为如下：    图 44 Lasso线性回归预测结果  其中训练得分 ：0.77，测试得分为0.76，Lasso线性回归的平均绝对误差为: 0.61 |
| **DecisionTreeRegressor决策树回归调参实验过程** |
| 1. 分隔数据集   同上的数据分隔方式。   1. 训练对应的模型   这里训练的时候，我们主要的设置的参数就是max\_depth，由于是决策树模型，所以对于树状的模型出结果很快，但是从测试集得分来看，我们发现不是很问题，训练代码如下：    图 45 DecisionTreeRegressor决策树回归预测代码  训练结果在criterion='mse', max\_depth=9, max\_features='sqrt', min\_samples\_split=2,  min\_samples\_leaf=1, random\_state=0下拥有比较好的结果，训练集得分：0.89测试集得分：0.66，并且决策树回归的平均绝对误差为0.66，最终的可视化图形为：    图 46 DecisionTreeRegressor决策树回归预测结果 |
| **最终预测结果** |
| 4.1 现在买房要求为 :  1. 在天府新区购买(15)  2. 面积大概再105㎡左右（105）  3. 户型结构：平层（4）  4. 建筑类型：平房（1）  5. 房屋朝向：南（5）  6. 建筑结构：钢混结构（7）  7. 装修情况：精装（2）  8. 配备电梯：有（1）  9. 交易权属：商品房（5）  10.房屋用途：普通住宅（4）  11. 3室1厅1厨1卫（3、1、1、1）    图 47 总预测结果 |

## （二）模型评估

针对我们选择的随机森林模型，我绘制了对应的测试、验证数据相似性散点图观察其分布情况，同时，我们也是用对应的训练集得分、测试集得分做出具体的百分比展示，最后我们使用均方差根（mean\_squared\_error）、均方误差回归损失（r2\_score）、解释方差回归得分（explained\_variance\_score）来进行判断，

## （三）可视化结果

|  |
| --- |
| **预测数据、测试可视化相似性展示** |
| 图 48 可视化结果  由于我们使用数据量比较大，从图中我们可以明显看出对应的重合性较高，只有部分数据相关性不大，同时，我们能查看对应的测试集、训练集得分：    图 49 测试集、训练集得分  同时，我们开始在设置对应的随机森林的时候，对不同超参进行选择，对应的可视化图表如下：  `[~M2M8{ZI[(K7SQ9ZDONER  图 50 超参设置一  JCRT]JH1Y9$QVQ1UN3$@3M2  图 51 超参设置二 |
| 7(ZQMFO{TQ7)(XYJM]%ZLQT  图 52 超参设置三  最终选择我们的 |

# 五、方案评估

通过我们的分析过程，我们完成了所有的分析任务，达到预期目标；由于二手房房价预测在互联网行业中已经有了很多成熟的例子，所以在实现的过程中，我们有了更多的参考和学习的地方，同时，针对数据处理部分，我们充分运用此学期所学的python处理数据的方法，对数据进行爬虫，特征匹配，得到我们接近2万5千条数据，同时经过基本的数据清洗操作，针对得到的1万4千多条数据进行探索性数据分析，可视化结果，得到很多有意思的结论；同时我们不忘课程设计任务——对二手房数据进行预测，参考相关的论文，按照论文提出的随机森林方式做出合适的条件预测，并且与多元的线性回归、决策树、K近邻等多种算法相比较，得到我们最好的预测结果，并且可选方式更多，预测效果也更好。而在开发过程中，我们首先开小会，确定开发流程，撰写对应的开发文档，同时给每一个人安排对应的工作和任务，分工明确、合理有序，圆满的、按质按量的完成了课程设计任务。对于该课程设计的难度，我们是能够接受的。但是依旧有些须的问题，例如并没有过多在房子的房龄、挂牌时间等信息做出选择，或者是上一次房主的评价等等都有可能影响该二手房的放假，所以在后续也有待提升。

## 

## 附录

表 1 小组成员分工表

|  |  |  |  |
| --- | --- | --- | --- |
| **成员1** | **姓名** | **学号** | **专业班级** |
| 王瑞 | 5120194818 | 卓计1901 |
| **主要完成工作** | 建立对应仓库、撰写相关开发文档、安排分工、设置个人任务；  参照对应论文实现相关随机森林算法预测二手房房价，包括随机森林、决策树、K邻近模型构建、训练、预测等操作；同时结合给出的买房要求进行预测价格。  撰写课程设计报告，完善相关ppt； | | |
| **成员2** | **姓名** | **学号** | **专业班级** |
| 黄婧宇 | 5120196267 | 卓软1901 |
| **主要完成工作** | 完成可视化设计部分内容，包括探究单价、数量、总价之间的关系 、探究装修情况、所在楼层、不同楼层直接是否配备电梯、建筑类别、房屋用途、建筑结构、各个区域等多条件和二手房的关系；  完成数据爬虫；  撰写课程设计报告，完善相关ppt； | | |
| **成员3** | **姓名** | **学号** | **专业班级** |
| 王梦林 | 5120194745 | 卓计1901 |
| **主要完成工作** | 完成可视化设计部分内容，包括每个区单位面积关系，房价分析、单价、数量、总价和房屋所属市辖区之间的关系 、房屋面积和总价关系、房屋面积和总价、房屋朝向和总价的关系等多条件和二手房的关系；  撰写课程设计报告，完善相关ppt； | | |
| **成员4** | **姓名** | **学号** | **专业班级** |
| 周子琳 | 5120194441 | 卓计1901 |
| **主要完成工作** | 完成文档摘要、数据分析目标与任务、数据预处理等多方面的文档撰写并完善相关ppt；  完成数据预处理、数据聚合；  完成数据清洗，包括重复值处理、缺失值处理、异常值处理； | | |