

机器学习

标题：基于机器学习算法的杭州市房价预测

|  |  |
| --- | --- |
| 姓 名： | 周洋 |
| 学 号： | 22020040174 |
| 专 业： | 应用统计 |
| 学 院： | 统计与数学学院 |

2023 年 4月 20 日

**基于机器学习算法的杭州市房价预测**

# 摘要

近年来，杭州市经济的飞速发展不断带动各区二手房房价直线上升，根据不同特征预测二手房房价成为重点难题。本文根据特征价格理论，将房价影响因素分为三类：建筑特征、邻里特征与区位特征。本次数据来源于链家网，通过Python爬虫获取，并调用API搜集邻里特征数据与区位特征数据，经过数据清洗，最终得到了25043条样本与21个特征。数据集分布充分展示了杭州市高度上涨的房价水平与相对集中的房价分布。在模型上选择了多元线性回归模型、岭回归模型、支持向量机模型与XGBoost模型进行对比，利用网格搜索与交叉验证进行参数优化，实验结果显示XGBoost模型表现最好，模型评价指标MAPE仅为11.17%，拟合优度R2达到0.91，综合表现非常优异，这充分展示了XGBoost模型在处理高维数据集时极强的拟合能力与预测优势。

**关键词：** 二手房；Python；特征价格理论；多元他线性回归；岭回归；支持向量机；XGBoost

目录

[摘要 2](#_Toc432823886)

[1 引言 4](#_Toc943939613)

[1.1 研究背景与意义 4](#_Toc1331375302)

[1.2 国内外研究综述 4](#_Toc1792582621)

[1.3 主要研究内容及方法 5](#_Toc888027384)

[2 数据描述性分析及预处理 6](#_Toc64896238)

[2.1 描述性分析 6](#_Toc1936863037)

[2.2 相关性分析 9](#_Toc1299941633)

[2.3 数据预处理 10](#_Toc1767884900)

[2.3.1 缺失值处理 10](#_Toc257774408)

[2.3.2 异常值处理 11](#_Toc939959257)

[2.3.3 One-hot编码 11](#_Toc1005525067)

[3 基于机器学习算法的二手房价格预测模型 13](#_Toc1310982826)

[3.1 多元线性回归模型 13](#_Toc506138362)

[3.2 岭回归模型 14](#_Toc484724367)

[3.3 支持向量机模型 15](#_Toc1356963098)

[3.4 XGBoost模型 16](#_Toc202456946)

[3.4.1 基本原理 16](#_Toc1079794574)

[3.4.2 目标函数 16](#_Toc1870588068)

[3.4.3 树的分裂 16](#_Toc1960550443)

[3.4.4 树的生成 17](#_Toc2129699580)

[3.4.5 网格搜索与交叉验证 17](#_Toc1750896511)

[3.5 模型的对比 18](#_Toc349245536)

[4 总结 20](#_Toc696916301)

[4.1 本文主要结论 20](#_Toc696460169)

[参考文献 21](#_Toc1620184233)

# 引言

## 研究背景与意义

自颁布房改制度以来，中国房地产市场发展进入黄金期，整个房地产产业出现了空前盛况，开始逐步掌控国民经济的发展命脉。但随着各类资产的大量涌入，市场隐含的矛盾与危机逐渐浮出水面。由于房源供应量远远小于居民的住房需求，供需不平衡引起房价失控增长，再加上投资者一系列的疯狂炒作，引起了全国性房地产投机现象，严重影响国民生活水平和消费水平，国民经济发展的不平衡也再一次导致市场产业结构无法得到有序调整，整个社会经济随时面临崩塌和瓦解。作为新一线城市，杭州市一直以卓越的经济发展水平领跑全国，同时由于地处沿海地区，房地产市场在稳定的条件下迅猛发展，然而在资本的冲击下，其房地产市场也往往存在一些隐性危机。2008年，金融危机的到来，使得本就充满“泡沫”的房地产市场面临随时崩盘的风险。因此在杭州市政府紧急颁布各项政策来控制房价波动，2010年国务院颁发《关于坚决遏制部分城市房价过快上涨的通知》，杭州市引来史上最严格的限购令，在政府的一系列引导下，杭州市的房地产价格体系得到重塑，2010年房地产销售指数在经历了2008年的下降后重新回到了高位，杭州市城镇居民人均消费支出达到了20218元，比2008年环比增长了20.9%，其中用于居住服务的支出为1853.18元，较2009年环比下降了9.4%。而在之后至2016年之间，杭州市房价依然受经济与政策影响不断波荡起伏，这实际上是一场时间很长的博弈过程，只有不停的向房地产市场进行抗衡，才能保证其健康的发展趋势。

目前，国内外学者已经在房价预测领域进行了大量的研究，搭建了很完善的预测模型体系，而本文将以杭州市二手房市场为研究对象，定性分析杭州城镇居民的二手房价格的影响因素，通过应用机器学习模型中的支持向量机、XGBoost方法建立房价预测模型，并与传统的多元线性回归模型、岭回归模型进行对比，建立起适合杭州市二手房住宅价格的预测模型。

## 国内外研究综述

国外学者对于房地产市场的研究相对深厚，从房价影响因素的研究到对房价预测模型的建立，有着较为完整的理论体系与丰富的科研成果。Shashi Bhushan Jha等人通过收集弗罗里达州的房地产数据集，并检索各个年份，包括国内生产总值、居民消费价格指数以及生产者价格指数等社会经济因素，通过逻辑回归、随机森林回归、Voting Classifier和XGBoost回归，结果表明XGBoost模型的精准度和鲁棒性都远远优于其它模型。Mansi Jain等人利用Python建立不同机器学习模型探索房屋价格的影响因素并预测房价，使用简单叠加方法比较各种模型的输出效果。

国内学者针对房地产市场的研究，相较于国外学者会稍晚一些。武秀丽与张锋对广州市房价采用时间序列分析，通过三次移动平均消除时间数列中隐含的季节变动、循环变动等影响，从而建立模型，最终根据残差分析图表现出了时间序列模型在房价预测方面的合理性。续云丰等对杭州市房价数据建立多元线性回归模型，辅之以逐步回归方法来剔除不符合的变量，从而建立岭回归模型，模型结果表现了岭回归模型在面对具有多重共线性变量时具有一定的稳定性。龚洪亮通过XGBoost模型对武汉市的二手房房价进行建模，利用GridSearch算法对模型进行参数调优，并与Lasso模型进行对比，模型结果表现了XGBoost模型的预测精度与效率更为优异。王冬雪和郭秀娟对数据进行缺失值处理与归一化处理后，选取影响程度比较大的特征建立XGBoost模型，并选取均方根对数误差来评价模型，结果表现了XGBoost模型运算效率非常高，并且能更好适应不平衡的数据，反映了该类模型泛化能力很好。

## 主要研究内容及方法

本文主要围绕杭州市二手房房价的影响因素，利用Python爬虫挖掘数据，从而构建XGBoost模型，并与其它模型进行比较。

1. 引言。介绍了本文研究的杭州市二手房市场的背景以及该选题的意义。
2. 数据描述性分析及预处理。介绍了数据收集、清洗与预处理的过程，通过可视化作图来呈现数据的分布状况与综合情况。
3. 分析。建立机器学习模型，使用网格搜索和交叉验证进行参数调参，通过评级指标对比机器学习模型之间的优劣，得出相应结论。
4. 总结。利用训练模型进行模型应用，根据近几年杭州市二手房的市场背景，提供相应建议，总结全文仍需要改进的地方。

# 数据描述性分析及预处理

## 描述性分析

本文共有25043条二手房数据样本，每条数据均有19条有效的特征属性。其中由于经纬度代表的是该房源的地理位置，其对房价的影响已经包含在地段等级对房价的影响中，再加上经纬度本身数值过于密集，将其带入模型中会导致模型效果大幅度变差，因此将经纬度特征进行剔除。而房屋户型作为分类型变量，由于变量值类型太多，不适合直接进行One-hot编码，因此将其拆分成4个数值型变量。

因此在建模中，一共有14个数值型变量与7个分类型变量。

数值型变量分布情况如下图：

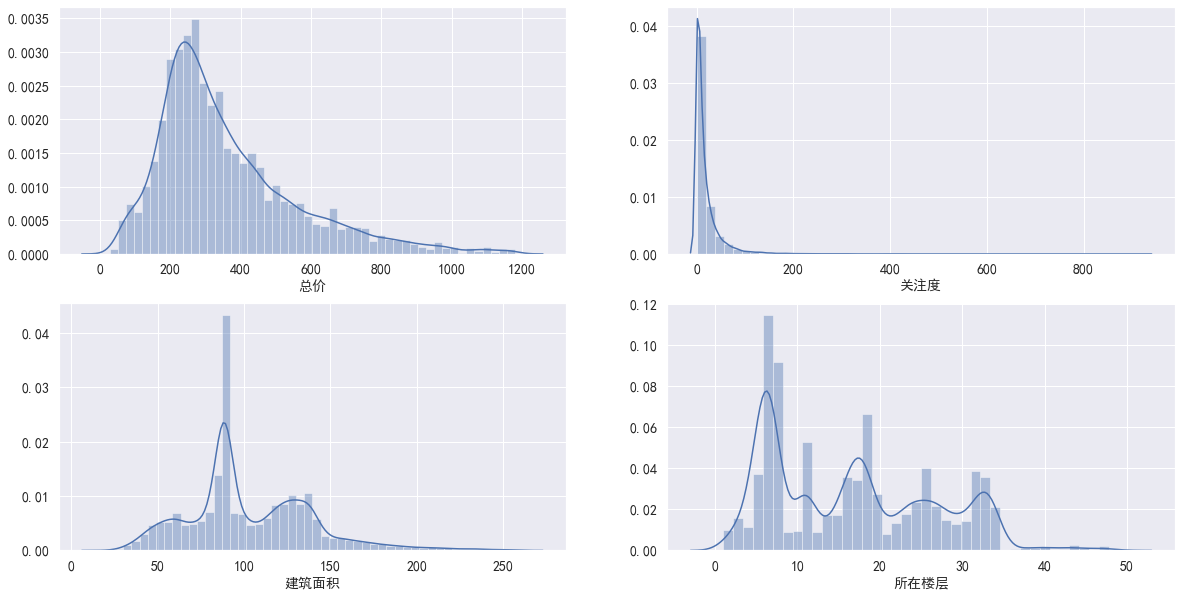


图 ‑3 数值型建筑特征数据变量频率分布图

从图中我们可以看到，每套二手房的关注度在50-100之间，也就说明了每套二手房的曝光程度相差不大，居民查找房源信息时可以较为方便地找到自己想要的内容。在“建筑面积”图中，我们可以看到建筑面积分布基本呈正态分布，而80平方米的二手房较多，其次是120-150平方米；在“楼层数”图中，各个楼层数几乎都有占比，其中第7层的二手房最多，在第1到第3的低楼层上交易量较少，基本都聚集在中高楼层上，这也可能说明了中高楼层更容易获得人们的青睐，而低楼层由于潮湿、易受蚊虫叮咬等原因不受居民欢迎；

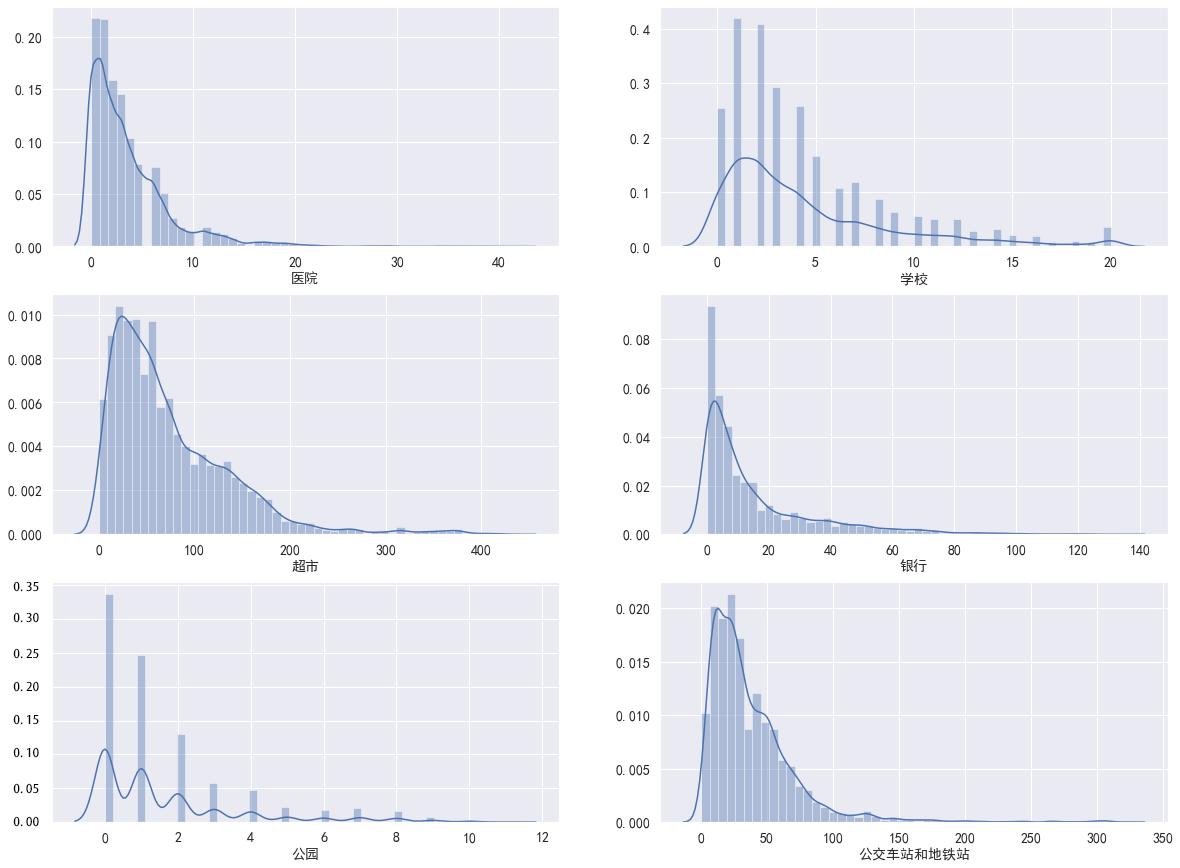


图 3‑4 数值型邻里特征数据变量频率分布图

在医院、学校、超市、银行、公园与公交车站与地铁站分布图上，图表走势基本比较统一，而周边医院、公园与学校为0个的二手房超过25%，但银行，地铁站和车站数与超市数为0的均低于1%，这一方面是由于占地面积不同所导致的，前三者占地面积往往远远大于后三者的占地面积，辐射范围更广，因此数量会相对较少，但这并不代表人们对于前三者的需求下降，而是其本身数量就较少；另一方面也说明了当前房地产市场往往会将周边配套设施的重心放在交通与购物方面。其中仍有一需要注意：超市数过多，甚至会出现1公里内有100多家的情况，这是因为超市数包含了一些居民自营的小超市，占地面积小，辐射范围窄，因此数量会很多。但对居民而言，超市不仅是购物的地方，还应包含各类娱乐设施，能够为居民提供丰富的生活服务，因此在此处超市数仍有改进的空间。

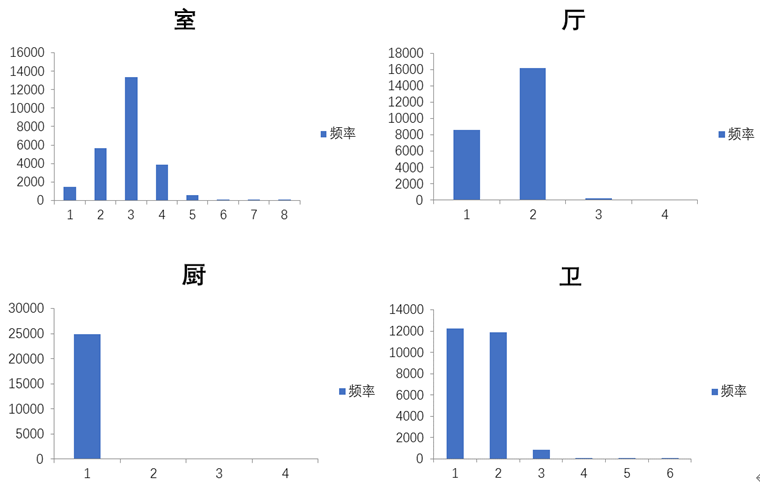


图 3‑5 房屋户型分布图

根据房屋户型分布图可以看出，目前主流房型为3室2厅1厨1卫或2卫。室的分布基本上集中在2~4室之间，属于常见房型。而其中厅包含了餐厅和客厅，越来越多的年轻人考虑到很少请客以及吃饭时很喜欢看电视等因素，会减少餐厅布局，将餐厅与客厅融为一体，扩大客厅的使用面积。但厨房一般都会有1个，很少会不带厨房的户型。而卫一般为1卫或2卫，一般来说小户型就一个卫生间，但90m2以上的户型一般会有主卫和客卫之分。

分类型变量分布如下：

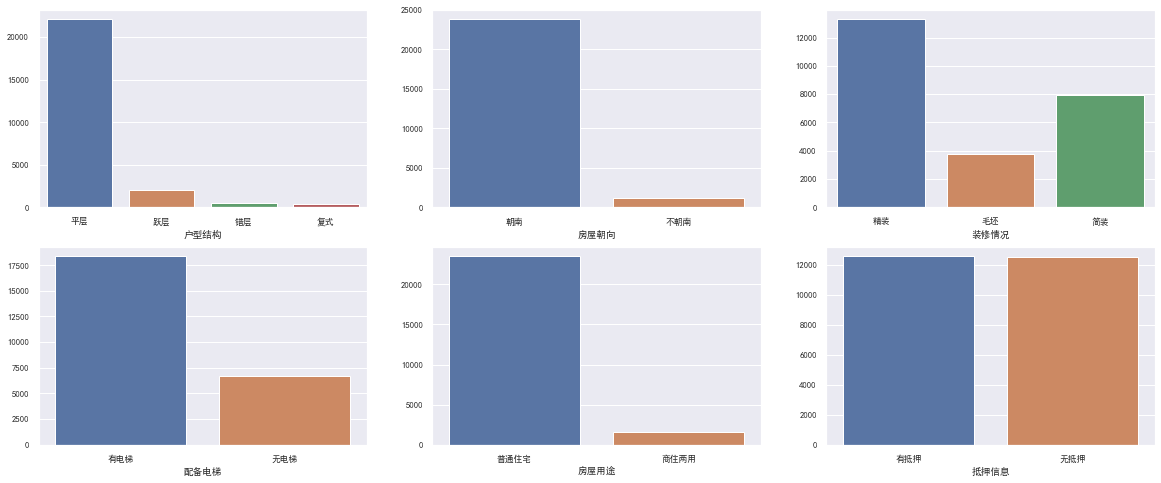


图 3‑6 分类型变量分布柱状图

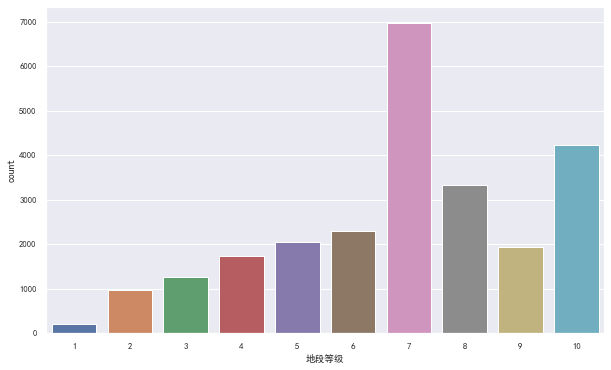


图 3‑7 地段等级分布柱状图

通过柱形图，我们可以明显地看出来，在房屋户型、房屋朝向、房屋用途三个变量中，基本上集中在“平层”、“朝南”、“普通住宅”这三个变量值上，这是因为目前住宅的趋势都是平层化，朝南也成了买房的硬性要求，且大部分居民购买房屋都是为了当作普通住宅使用，而不是商业化，这也是符合现实依据的。同时，二手房普遍是住过一段时间再行出售的，再加上目前开发商对住宅都会采取精装修的方式获取更多利润，因此二手房一般都是装修过的，毛坯房会相对较少。从地段等级分布图可以看出，房屋的分布与面积成正比，因此第七等级的分布数量最多，正是由于第七等级所属区域最大所造成的。同时，处于第10等级的房屋较多，是因为我们将所有不属于前9个等级的房屋全部归属于第十等级，因此在第十等级里面包含了一定的误差量，之后可以改进的。

## 相关性分析

为了判断各个变量之间的相关性，我们对全部数值型变量以及分类型变量中的“地段等级”计算皮尔森相关系数，使用seaborn.heatmap()函数绘制相关性热力图。以此来判断总价与各个变量之间的相关程度。

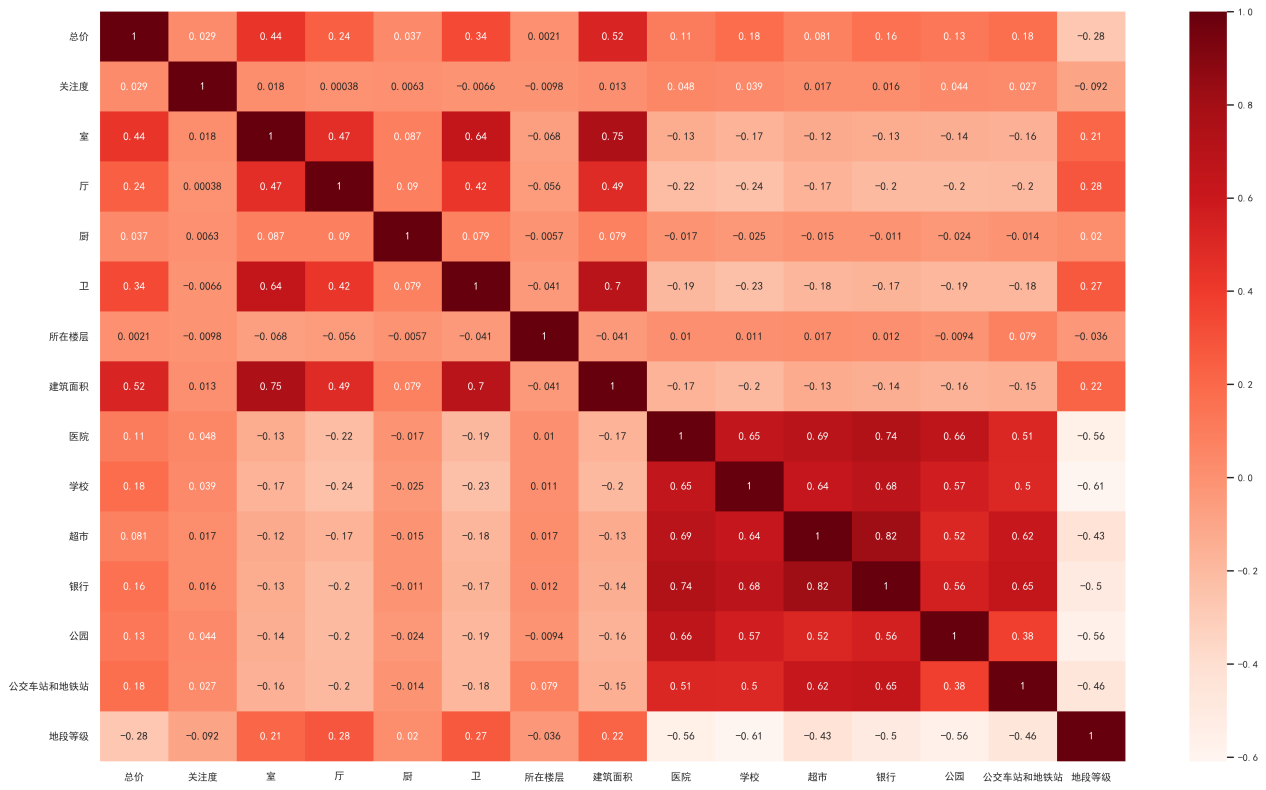


图 3‑8 相关性热力图

通过相关性热力图，我们可以明显看出除了地段等级以外，其它特征均与总价呈正相关，其中建筑面积相关性最大，其次是室、卫、地段等级、厅、学校、公交车站和地铁站、银行、公园、医院、超市、厨、关注度、所在楼层。建筑面积相关性最大是因为目前单价已经基本上超出1万/，可以说是“寸土寸金”。而房屋户型与总价息息相关，一部分也是由于房屋户型往往越复杂，建筑面积越大，另一部分则是由于优异的户型往往总价越高。其中，周边配套设施的相关性也相对较高，这是由于其数值越大，代表其所在的区域越发达，一般来说，经济越发达的地方，周边配套设施越完善，故而周边配套设施的变量的相关性会较高。地段等级呈负相关仅与其取值有关，此处采用官方所定义的地段等级，数值从1到10，1代表地段最发达的区域，10代表虽不发达但很有潜力的区域。

## 数据预处理

经过数据清洗后的数据仍会存在缺失值与异常值，因此我们需要进行数据预处理。但由于XGBoost模型是决策树模型，不支持梯度下降，所以我们的数据无需进行对数化、归一化处理，只需进行简单的缺失值、异常值处理与One-hot编码。

### 缺失值处理

在数据搜集过程中，往往会因为各种情况造成数值缺失，比如官方未搜集到数据、数据格式与数据搜集所写入的格式不一致等，同时往往二手房在上传至链家二手房时，卖家不愿暴露过多的信息或者希望能够进行线下交易以避免不必要的损失，因此经常会空缺部分数据以引导买家进行讯问或者保护个人隐私，因此这个时候就需要我们进行人为处理缺失值。

本文通过数据清洗后的数据已经剔除了全为空值的行与列，故缺失值会相较于原始数据大大减少，但仍会有关键列存在缺失值。其中，“关注度”，“户型结构”，“建筑类型”，“配备电梯”均存在1000~2000条数据存在空缺，对“关注度”、“配备电梯”可采取众数填充的方式进行处理，而“建筑类型”与“户型结构”由于种类较多，故直接删除样本。

### 异常值处理

由于样本数据中存在系统性误差与偶然性误差，这些误差都会直接或间接地导致随机抽取的样本无法代表总体。因此我们需要对样本数据中存在的异常值进行处理，以尽量控制其带来的影响。而针对异常值，本文通过设定边界值来判断数值是否大幅度越界，如果越界过大，则采取直接剔除的策略。判定依据如下：



其中W一般取1.5，U为上界，L为下届。

通过处理缺失值和异常值，剔除了683条数据，总共留下了25043条有效数据，这种数据量仍然适合用机器学习模型建模。

### One-hot编码

One-hot编码也称独热编码。它的具体方法为，使用N个位置存储器来存储某个变量的N种变量值，即N种状态，每个存储器内有且仅有一种状态。例如：

形状

中度可信度描述已自动生成

图 3‑13 One-hot编码示意图

在本文中，由于“户型结构”、“房屋朝向”、“装修情况”、“配备电梯”、“房屋用途”、“抵押信息”这六类分类型变量的变量值类型并不多，均在4个以内，故而可以对这6个分类型变量进行One-hot编码，而不会建模产生过多的影响，而对“地段等级”不进行任何处理。

表 3‑6 One-hot编码值

|  |  |
| --- | --- |
| 变量名称 | One-hot编码后数值 |
| 户型结构 | 户型结构\_复式、户型结构\_平层、户型结构\_跃层、户型结构\_错层 |
| 房屋朝向 | 房屋朝向\_朝南、房屋朝向\_不朝南 |
| 装修情况 | 装修情况\_毛坯、装修情况\_简装、装修情况\_精装 |
| 配备电梯 | 配备电梯\_无、配备电梯\_有 |
| 房屋用途 | 房屋用途\_商住两用、房屋用途\_普通住宅 |
| 抵押信息 | 抵押信息\_无、抵押信息\_有 |

因此，通过数据预处理后，我们最终建模中所使用的变量有27个，在多元线性回归模型中会展示出所有变量，其中15个变量为通过0ne-hot编码后得到的变量，实际上只代表了6个分类型变量。

# 基于机器学习算法的二手房价格预测模型

## 多元线性回归模型

多元线性回归是回归方法中最常见的方法，拥有清晰的数学表达式来解释变量之间的相互关系，相较于其它模型，更容易理解与使用。因此该模型具有非常广泛的应用范围，大量应用于金融市场、房地产市场等领域，因此对于研究杭州市二手房房价具有一定的合理性与参照性。

多元线性回归模型的表达式如下：



其中， 为个回归系数，其中为回归常数；为个自变量，为因变量，为随机误差项，通常假定。

多元线性回归模型建立之后，不能直接进行应用，只有通过拟合优度检验和显著性检验之后，才能进行其它应用。一方面是计算模型的拟合优度，是否完全解释了因变量和自变量之间的关系；另一方面是检验模型各个参数的显著性，对不显著的参数进行处理或直接剔除。

（1）回归方程的拟合优度检验

判定系数是回归变差占总变差的比值，一般用来反映回归模型的拟合优度。在0-1之间，越大，说明回归方程的拟合程度越高，误差越小。的公式如下：



（2）回归系数的显著性检验

T检验是用来分别检验回归方程中各个回归系数的显著性检验，对通过T检验的回归系数保留回归系数，对不通过T检验的回归系数予以剔除。T检验的检验流程是根据t分布表与显著性水平计算变量的接受域与拒绝域，同时根据样本数据代入计算参数的检验值，从而判断各个参数是否显著的假设检验方法。T统计量的公式如下：



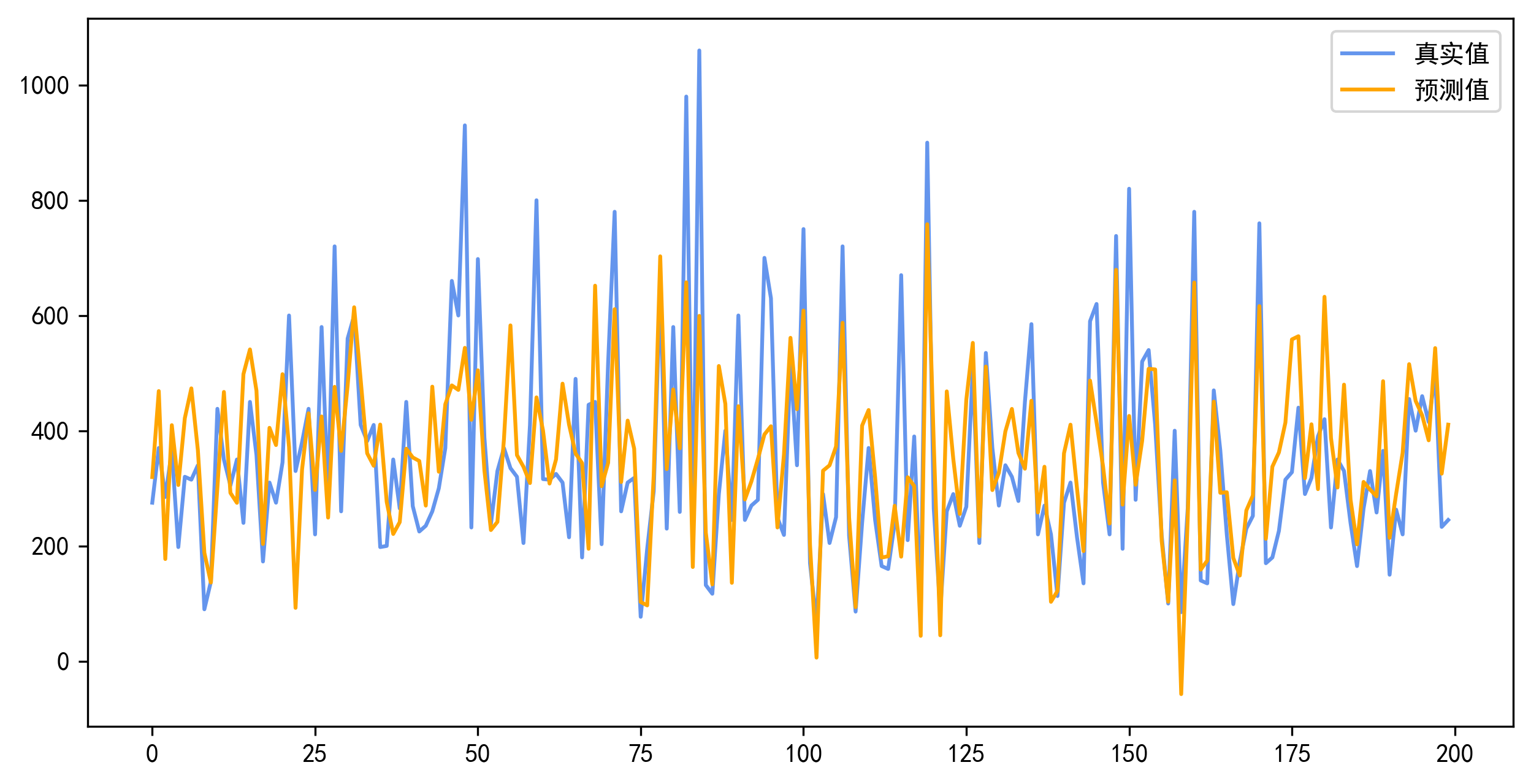
多元线性回归作为回归模型中最常用的模型，也是最基础的模型，我们往往将其作为对照模型，以检验其他模型的优良性。

在本文中，我们直接使用Python中的sklearn库的LinearRegression方法进行建模。最终得到回归方程式为



其中关注度，建筑面积、室、厅、厨、卫、户型结构\_复式、户型结构\_平层、户型结构\_跃层、户型结构\_错层、房屋朝向\_朝南、房屋朝向\_不朝南、装修情况\_毛坯、装修情况\_简装、装修情况\_精装、配备电梯\_无、配备电梯\_有、房屋用途\_商住两用、房屋用途\_普通住宅、抵押信息\_无、抵押信息\_有、楼层数、医院、学校、超市、银行、公园、公交车站和地铁站、地段等级。

预测结果如下：



## 岭回归模型

岭回归是对最小二乘法的改良，能够有效处理共线性的数据，属于有偏回归估计。它在模型建立过程中，不再考虑模型的无偏性，而是尽可能使回归系数更贴近实际，能更准确反映现实情况。这种回归模型主要用于拟合病态数据，防止多重共线性的发生。

在岭回归的目标函数中加入L2正则项，其结构式如下：



求解岭回归的过程就是选择正则项系数的过程，它是用来平衡模型的方差和偏差的系。目前，求解最优的值有两种方法，一种是通过可视化的方法绘制岭迹图，随着的增大，回归系数趋于稳定的点就是最优值。另一种是通过交叉验证的方式来获取最优值。本文就是选择第二种方式来选择最优值，因为这种方式更加快捷。

在岭回归模型中，最主要的工作是找一个合理的值来平衡模型。通过下图我们可以发现随着值的增大，模型的拟合优度下降，方差变大，整体泛化能力下降。

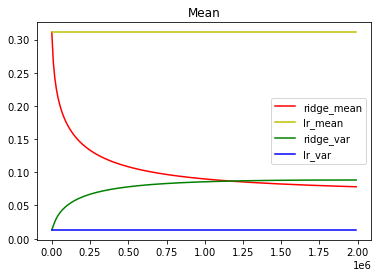
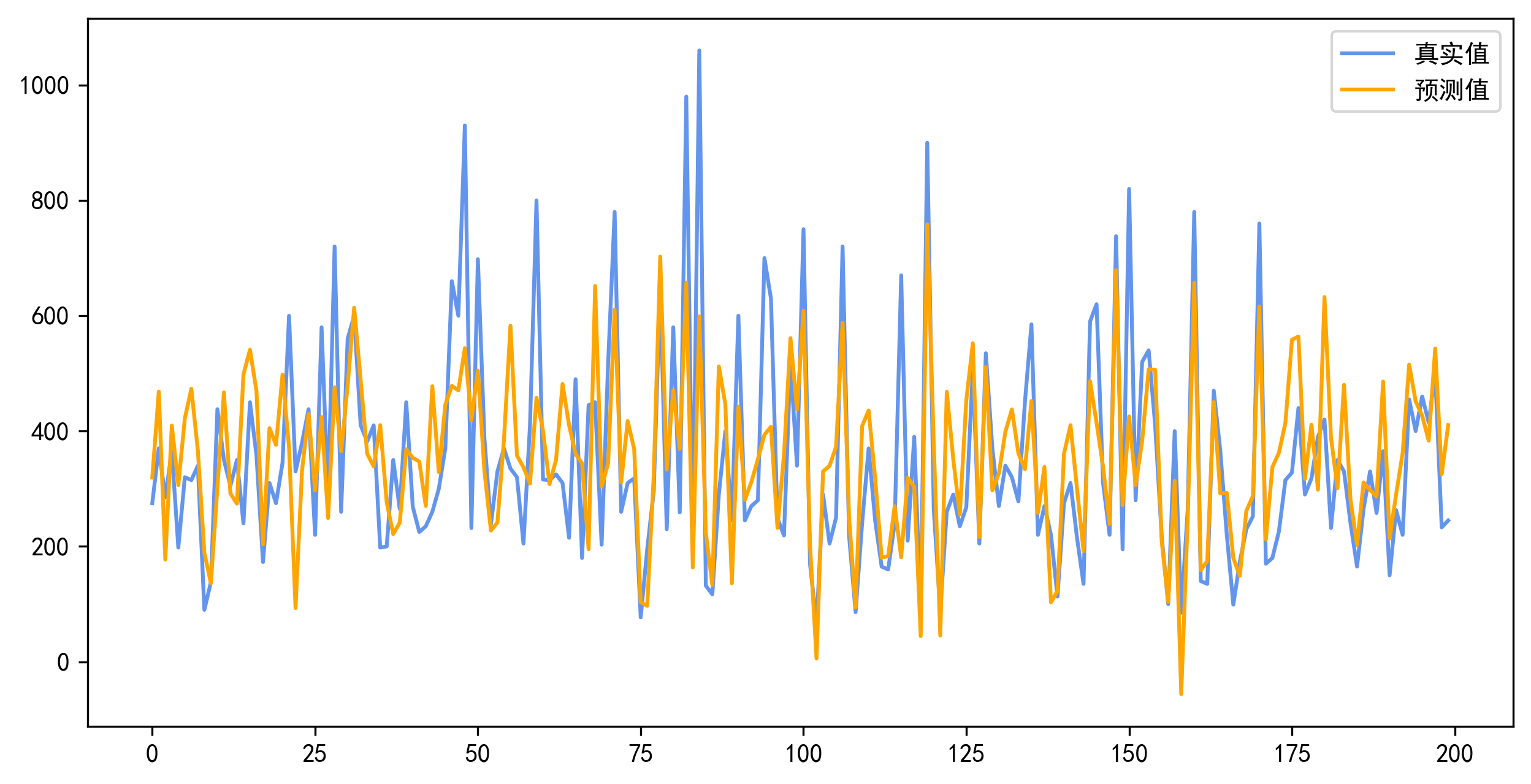


图 4‑7 的影响图

本文根据带有交叉验证的岭回归来判断岭回归的参数，使用RidgeCV()函数自动选择值。最终通过结果，我们得知当=1时，模型达到最优解，因此利用=1进行建模。

预测结果如下：



## 支持向量机模型

支持向量机SVM（support vector machine）是一类二分类模型，通过优化模型达到最小化错误率，属于监督学习算法。SVM算法的得到模型非常优异，在人像识别、文本分类问题中应用效果很好，其本质是在样本线性可分的前提下给出最优的将样本分类的唯一最大边距超平面。其最优超平面的示意图如图所示：

图表, 散点图

描述已自动生成

图 2‑1 示意图

当数据属于线性不可分时，支持向量机能够利用适合的核函数，将数据进行映射，一般将低维数据向高维空间进行映射。对于如何选择核函数，我们应当根据不同的情况选择不同的核函数，从而建立不同的SVM模型。目前多项式核函数、拉普拉斯核和Sigmoid核函数应用比较广泛，假设训练样本集为，那么核函数的具体表达式如下：









由于我们的数据集属于高维数据集，因此我们分别尝试线性核函数、径向核函数与高斯核函数来建模，通过对比评价指标值，最终我们选用均方误差最小的高斯核函数来建立支持向量机模型。

## XGBoost模型

XGBoost(e Xtreme Gradient Boosting)是陈天奇博士开发的一款机器学习模型，它是分布式梯度增强库的组成部分，大大提高了大数据科学问题的求解效率，在数学竞赛中非常受欢迎。

### 基本原理

XGBoost模型的诞生得益于Boosting算法，它在原有高度发达理论的基础上进行开发，底层基础模型依然是树模型，但是它是基于提升树的，通过建立新的树来训练原始树上表现不好的部分。XGBoost模型的基本原理就是通过汇聚多棵回归树进行拟合，在建立新树的过程中，通过选择特征最优分裂点来生长树，每建立一棵新的树就是去拟合前一棵树的残差值。最后通过得到训练模型，输入预测样本，每个样本都会在每棵树上落到对应的值，将值加总就能得到样本的预测值。

1. 目标函数

XGBoost的目标函数由两部分组成，一部分是训练损失函数，另一部分是正则化项表达式，目标函数定义如下：



其中代表第i个样本值，代表模型对第i个样本的预测值。代表训练损失函数，通过计算预测值和真实值之间的差异得到，常见的损失函数有平方损失函数与逻辑回归损失函数。代表树的复杂度，是正则项表达式，包括L1正则、L2正则，其值越小复杂度越低，泛化能力越强。

训练损失函数和正则化项在模型中负责的工作不同，训练损失函数用于计算、控制、优化模型的误差指标从而达到增加拟合优度的目的；正则化项方法注重于控制模型的复杂程度，权衡模型的利弊，主要对简单模型进行激励，避免过分复杂模型的出现，从而在训练误差与模型结构之间进行权衡后，寻求最优模型。

1. 树的分裂

生成XGBoost模型中最重要的步骤就是寻找最优分裂点，不同的分裂结构就代表不同的损失度，而我们要找的就是损失最小的最大增益点。目前主流的算法为贪心算法和近似算法。

（1）贪心算法

贪心算法就是将所有可能的结构进行评估，工作量最大，但精确度最高。其基工作原理是选择一个节点进行分类时，枚举出所有特征，并按照这些特征分别进行分裂树，计算损失函数，包括信息增益、信息增益率或基尼系数，选择损失函数最小的继续进行分裂，从而使每一步都是在上一次预测取得最优的基础上进行进一步分裂。

（2）近似算法

由于贪心算法在数据量很大的情况下会导致复杂度过高而难以运行，而近似算法弥补了不足。近似算法就是对各个分裂点设立候选集，对符合分裂点的连续变量数值放入候选集中，从中选取最优的分裂点。

1. 树的生成

树的生成并不可能会无限循环下去，每一次分裂整体的损失函数不一定增加，因此需要人为调整参数来阻断树的生长，防止过拟合。一般的调整参数有三种：

（1）当我们引入的分裂点所带的增益远远小于我们所设定的界限，我们可以选择删除这个分类点，相当于预剪枝。

（2）当特征数量过多容易过拟合时，可以调整参数max\_depth从而控制树的深度。

（3）当样本权重和小于设定阈值时，可以调整参数min\_child\_weight，从而阻断分裂。

### 网格搜索与交叉验证

网格搜索法是对指定参数值的穷举搜索方法，通过将待优化的参数通过交叉验证的方法进行优化，来找到最优参数组合的学习算法。交叉验证主要用于机器学习算法中，适合对训练模型进行调参及优化。交叉验证的基本思想是随机对样本进行分割处理，得到不同组的训练集和测试集，根据训练集训练模型，用测试集来进行评估。这里以10折交叉验证为例，将数据集分为10份，分别以其中一份作为测试集，剩下的作为训练集，就会得到10个模型，综合计算评级指标，选择最优模型进行输出。

在本文中，我们主要调整XGBoost模型的booster参数，通过网格搜索与交叉验证对参数进行调优，参数调优顺序为：eta，max\_depth，gamma，subsample、colsample\_bytree、colsample\_bylevel，lambda、alpha。下图为XGBoost模型的参数优化表。

表 4‑4 参数初始表

|  |  |  |
| --- | --- | --- |
| 参数 | 初始值 | 优化值 |
| eta | 0.3 | 0.08 |
| max\_depth | 6 | 5 |
| gamma | 0 | 0 |
| subsample | 1 | 0.9 |
| colsample\_bytree | 1 | 1 |
| colsample\_bylevel | 1 | 0.5 |
| lambda | 1 | 1 |
| alpha | 0 | 0 |

利用调整好的参数带入模型中进行建模，并绘制其模型训练过程中的预测误差前后对比图。

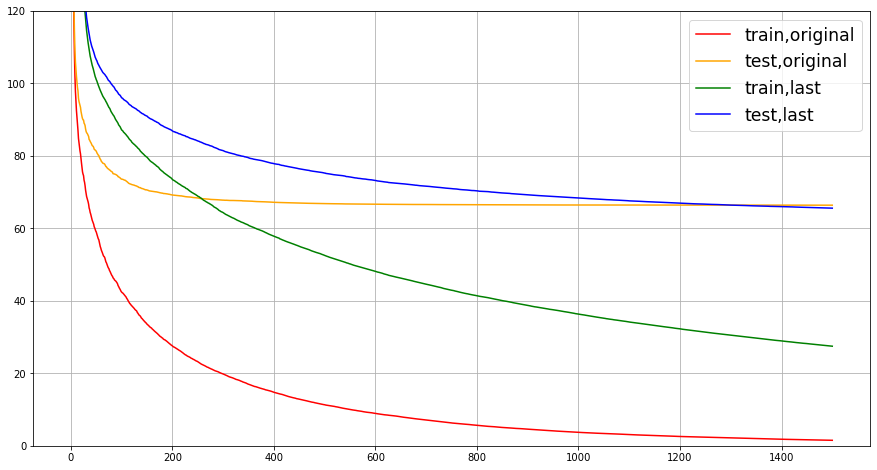


图 4‑4 XGBoost训练图

在上图中可以看出，随着训练次数的增加，训练模型的均方误差也随之减小。与初始默认参数相比，经过调参后的模型，泛化能力更好，且模型在训练集上的均方误差也变小了。

为更好表现预测的效果，这里挑选前200个房源，来绘制测试值与实际值的对比折线图。

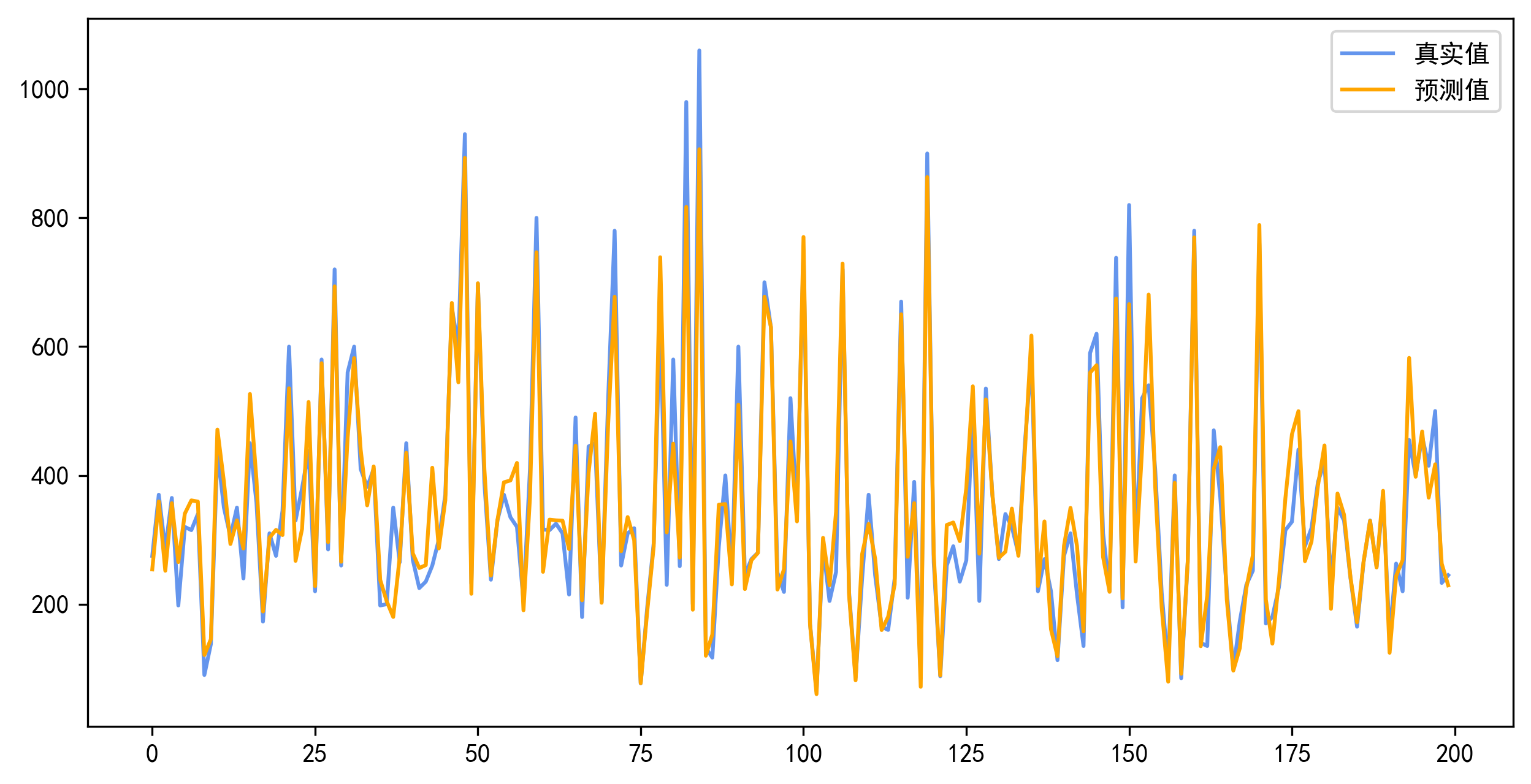


图 4‑6 XGBoost模型预测对比图

由上图测试图可以看出，预测值与实际值基本吻合，没有出现异常特别大的情况。

## 模型的对比

在评价指标的选择上，主要选择了均方误差(MSE)、平均绝对误差(MAE)、平均绝对百分比误差(MAPE)与拟合优度(R2)作为本次机器学习模型的评价指标。多元线性回归模型、岭回归模型、支持向量机模型与XGBoost模型在训练集上的评估效果表如下：

表 4‑7 训练集效果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 多元线性回归模型 | 岭回归模型 | 支持向量机模型 | XGBoost模型 |
| MSE | 18189.98 | 18189.53 | 10990.94 | 930.0986 |
| MAE | 98.86 | 98.85 | 65.17 | 21.2185 |
| MAPE | 30.72% | 30.72% | 24.32% | 6.769% |
| R2 | 0.55 | 0.54 | 0.72 | 0.9768 |

机器学习模型在测试集上的表现结果如下：

表 4‑8 测试集效果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 多元线性回归模型 | 岭回归模型 | 支持向量机模型 | XGBoost模型 |
| MSE | 18120.51 | 18110.53 | 12476.86 | 3722.44 |
| MAE | 97.9 | 97.89 | 71.64 | 37.98 |
| MAPE | 30.72% | 30.07% | 25.73% | 11.17% |
| R2 | 0.55 | 0.55 | 0.69 | 0.91 |

通过训练集效果表与测试集效果表，我们可以发现岭回归模型与多元线性回归模型的性能较差，而支持向量机模型虽然有很大的改进，但效果仍然较差，其中XGBoost模型的性能非常强大，

这是由于支持向量模型可以有效避免维数过大的优点，有很强的鲁棒性，但其性能仍然劣于XGBoost模型，这在一定程度上证明XGBoost模型具有更强的回归性能，且计算效率也远远强于其它模型，适合对此类房价数据进行建模。

# 总结

本文对杭州市二手房价格进行了探索性研究。根据特征价格理论，分析影响二手房价格的主要因素，为政府宏观调控提供决策支持，为居民购房提供数据支持，为开发商提供建议。本章主要做了以下几个方面的工作：

（1）本文通过描述性分析，来观察样本数据的分布情况。通过绘制数值型变量和分类型变量分布的柱状图，明显观察到杭州市房价水平基本集中在300~400万，也不乏1000万以上的房源，在全国属于较高房价水平，这对杭州居民带来了不小的购房压力。通过绘制数值型变量与地段等级的相关性热力图上，可以发现建筑面积对房价的影响水平遥遥领先，其次是房屋户型、地段等级、周边配套设施、关注度与所在楼层。

（2）本文通过建立机器学习模型，对杭州市二手房价格进行训练。模型参数分别进行了优化，根据实验结果与评价指标显示：多元线性回归模型和岭回归模型的效果较差，支持向量机模型效果较好，而XGBoost模型无论是在训练集上还是在测试集上，都表现了非常强大的优势，在训练集上MSE为930.0986，MAE为21.2185，MAPE为6.769%，R2为0.97，测试集上MSE为3722.44，MAE为37.98，MAPE仅为11.17%，R2为0.91，四类评价指标远远优于其它三个模型，这表明经过参数调优后的XGBoost模型具有极强的拟合能力和预测优势。

# 参考文献

1. 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要[N]. 人民日报,2021-03-13(001).
2. özsoy O, \_ahin H. Housing price determinants in Istanbul, Turkey[J]. International Journal of Housing Markets and Analysis. 2009, 2(2): 167-178.
3. Gibbons S, Mourato S, Resende G M. The Amenity Value of English Nature: A Hedonic Price Approach[J]. Environmental and Resource Economics. 2014, 57(2): 175-196.
4. Mohan S, Hutson A, Macdonald I, et al. Impact of macroeconomic indicators on housing prices[J]. International Journal of Housing Markets and Analysis. 2019, 12(6): 1055-1071.
5. Zhao Y, Chetty G, Tran D. Deep Learning with XGBoost for Real Estate Appraisal[C]. IEEE, 2019, pp.1396-1401.
6. Jha S B, Pandey V, Jha R K, et al. Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study[J]. 2020.
7. Jain M, Rajput H, Garg N, et al. Prediction of House Pricing using Machine Learning with Python[C]. IEEE, 2020, pp.571-574.
8. 续云丰，孔亚仙，徐仁旭. 统计模型和博弈论视角下的杭州房价影响因素分析[J]. 长沙大学学报. 2014, 28(02): 105-107.
9. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. ACM, 2016.pp.785-794
10. 龚洪亮. 基于XGBoost算法的武汉市二手房价格预测模型的实证研究[D]. 华中师范大学, 2018.
11. 王冬雪，郭秀娟. 基于XGBoost算法的房价预测模型[J]. 北方建筑. 2021, 6(03): 79-82.
12. 张志锋，崔亚东，崔霄. 基于XGBoost的二手房房价预测模型[J]. 数字技术与应用. 2019, 37(11): 178-180.
13. 梁佩. 基于机器学习的二手房估价模型[D]. 云南大学, 2019.
14. 张家棋，杜金. 基于XGBoost与多种机器学习方法的房价预测模型[J]. 现代信息科技. 2020, 4(10): 15-18.
15. 温海珍. 城市住宅的特征价格：理论分析与实证研究[D].浙江大学,2004.
16. 钟丽燕,高淑兰.多元线性回归模型在房价走势分析与预测中的应用[J].科技创业月刊,2017,30(09):94-96.
17. 陈绵旺. 基于RS-SVM的商品住宅价格预测研究[D].华东交通大学,2016.
18. 孙丽丽,方宏彬,朱星星,胡蕾明,齐龙武.基于网格搜索优化的XGBoost模型的股票预测[J].阜阳师范大学学报(自然科学版),2021,38(02):97-101.