题　　目：杭州二手房价格预测系统

课程设计评分细则

序号	评 分 内 容			分值	得分
论文质量 (60%)	(30)目标一：问题分析是否深入详细、合理，获得有效结论。				
	(30)目标二：解决方案或数据分析方案合理性；论文结构严谨性，逻辑性，论述层次清晰，语言准确，文字流畅，符合规范化要求。				
答辩质量 (40%)	(20)目标三：是否能用非常流畅的文字阐述和分析问题并得出有效结论。				
	(20)目标四：是否查阅大量高质量国内外文献资料；分析整理各类信息、从中获取新知识的能力。				
综　合评语		最终成绩		教师签名	

目 录

摘 要

随着经济高速发展，人民的生活水平提高，房地产市场占据重要地位，我国房地产市场高速升温，住宅出售价格居高不下。然而市场的资源配置介入手段却并没有取得良好效果。在市场失灵的情况下，政府介入房地产行业进行宏观调控和干预。本文章对二手房的价格进行研究，建立模型实现对二手房房价的预测并依托该模型开发出一个二手房信息平台，旨在打破买卖双方信息差，让二手房交易市场变得更加透明，促进二手房市场更好的发展。

对二手房的交易价格进行研究，可以增加二手房市场价格的透明度，帮助买卖双方基于更准确的信息进行交易决策，减少信息不对称导致的不公平交易，合理的评估体系能够确保二手房的定价更加公正、合理，反映房屋的真实价值，保护消费者免受高价低质或欺诈行为的影响，同时也保障了卖方的合法权益。

本通过查阅现有的二手房价格预测相关的研究文献，发现现有的研究对数据集的研究不够深入，大多仅限于简单的数据预处理，并且大多数研究都只用到了单一模型，少量使用到多模型的也仅于简单的对比分析以及加权融合。本文吸收了现有研究的经验并在其基础上进行创新，对现有研究所使用到的模型进行训练并对比分析选择出 XGBoost、以及随机森林(RF)作为基学习器，并使用 stacking 的方式进行模型集成。在模型的训练上，采取五折交叉，利用其中 4 份进行训练，另外 1 份进行验证，并将 5 次验证结果合并作为元学习器的输入。在总结现有的经验后，本文对数据集进行特征工程，发掘数据集中潜在的关系并使用嵌入法进行特征选择，在参数调节环节，本文利用贝叶斯优化进行调参，该方法可以在较短的时间内消耗较少的算力资源获得相对较好的超参组合。

最后本文建议消费者理性购买二手房,在充分考虑个人实际需求的前提下,还要考虑到未来的房价趋势,最适合自己的房屋;建议二手房销售平台方透明化二手房定价,并全面考虑二手房的各种信息,建立一套更为科学的定价体系,共同促进二手房市场的平稳发展。

关键词：二手房价格预测；XGBoost；随机森林；Stacking；贝叶斯优化

Abstract

With the rapid development of the economy and the improvement of people's living standards, the real estate market occupies an important position, China's real estate market is heating up at a high speed, and the price of residential properties for sale remains high. However, the market's intervention in resource allocation has not achieved good results. In the case of market failure, the government intervenes in the real estate industry for macroeconomic regulation and intervention. This paper studies the price of second-hand housing, establishes a model to predict the price of second-hand housing, and develops a second-hand housing information platform based on the model, aiming to break the information gap between buyers and sellers, make the second-hand housing transaction market more transparent, and promote the better development of the second-hand housing market.

The study of the transaction price of second-hand housing can increase the transparency of the second-hand housing market price, help buyers and sellers make transaction decisions based on more accurate information, reduce unfair transactions caused by information asymmetry, and a reasonable evaluation system can ensure that the pricing of second-hand housing is more fair and reasonable, reflecting the true value of the house, protecting consumers from the impact of high price, low quality or fraud, and also protecting the legitimate rights and interests of sellers.

By reviewing the existing research literature related to second-hand housing price forecasting, it is found that the existing studies are not in-depth enough on the dataset, and most of them are limited to simple data preprocessing, and most of the studies only use a single model, and a small number of multiple models are only used for simple comparative analysis and weighted fusion. This paper absorbs the experience of existing research and innovates on the basis of it, trains the models used in existing research, compares and analyzes XGBoost, and selects XGBoost and Random Forest (RF) as base learners, and uses stacking to integrate the models. In the training of the model, a five-fold cross was adopted, 4 of which were used for training and the other 1 for verification, and the results of the five verifications were merged as the input of the meta-learner. After summarizing the existing experience, this paper performs feature engineering on the dataset, explores the potential relationships in the dataset and uses the embedding method to select features,

and uses Bayesian optimization to adjust parameters in the parameter tuning link, which can consume less computing resources in a short time to obtain relatively good hyperparameter combinations.

Finally, this paper suggests that consumers buy second-hand houses rationally, and on the premise of fully considering their actual personal needs, they should also take into account the future housing price trend to be the most suitable for their own houses. It is suggested that the second-hand housing sales platform should be transparent about the pricing of second-hand housing, and comprehensively consider all kinds of information on second-hand housing, establish a more scientific pricing system, and jointly promote the stable development of the second-hand housing market.

Keywords: Second-hand home price prediction, XGBoost, Stacking, Bayesian optimization

第 1 章 绪论

1.1 项目背景

随着经济高速发展，人民的生活水平提高，房地产市场占据重要地位，我国房地产市场高速升温，住宅出售价格居高不下。然而市场的资源配置介入手段却并没有取得良好效果。在市场失灵的情况下，政府介入房地产行业进行宏观调控和干预，2020 年十二月，央行对于银行的房贷比例进行了限定，从放贷额度上，将与之对应的房贷业务款项额度进行约束。早在之前也划定了房地产三条红线，严格控制超过红线的房地产企业融资，降低房地产企业的杠杆率。除此之外，还出台了老旧小区改造，人才落户，补办房产证明，优化国土空间配置，财产登记，等一系列与房地产市场息息相关的新政策,。但从中国住宅近年的实际价格涨幅看中国楼市却进入了“越调控，价越高”的怪圈。表现出来的“强政府”印象与调控失灵是中国房地产市场治理过程中一个凸显的矛盾[1]。

房价的走高与我国社会经济金融发展息息相关。资料显示我国房地产企业的资产负债率极高。各大银行贷款总额的 39%流向了房地产市场，还有大量基金、地方政府融资，金融衍生品等进入房地产市场。可以说，房地产是我国化解和防范金融风险的最大挑战。同时涌现了越来越多的房地产市场现象无法使用经济学原理进行合理的解释，从 2003 年到 2017 年，房价保持了连续 14 年上涨。单边市场、只涨不跌，这在股市等其他市场中是不可想象的事情。大量的实例都证明了中国房地产市场某些现象一定程度上不符合经济学假设[2]

2023 年 1 到 11 月，全国二手房成交量和新建商品房销售面积比 2022 年同期增长 6.9%，二手房成交量占新房和二手房交易总量的比重近 40%，较上年同期提高约 10 个百分点，部分主要城市二手房成交量占比甚至超过 50%。房价的高低成为了与人民生活息息相关的问题，因此预测房价变化趋势的问题备受关注。而随着经济社会的发展，生活品质成为了越来越多的人的追求，对于居住问题来说不再是局限在单纯的“住”。小区环境，配套的教育，医疗，交通，购物等等都成为了人们在购房时的主要考虑因素。而房地产开发商在进行定价宣传时也主要是通过当地的主流媒体发布大版面的购房价目表，突出表现出的除了价格，还包括具体的楼层，朝向，户型，房号，面积，单价、总价。而在宣传时则着重宣传楼盘所属学区，距离各个地铁站口距离，距离三甲医院以及附近配套商业核心区的距离。

所以在对房价的变化趋势进行分析时要更多考虑不同的维度，如小区方圆一千米的地铁站(公交站)个数，医院距离，学校知名度，小区的绿化面积，层高等。

随着信息化技术进入生产生活的一切活动，生活的每一个节点，关键因素都被自动化的计算机记录了下来。同时电子通信与移动计算机的结合，使得人们的互联互通变的迅捷简便。使得人类活动的各方面的表现都具有了信息活动的特征，信息成为了一切活动的参与者。此时信息的过载成为了信息化时代的大问题。而机器学习正是将这些信息进行挖掘创造价值的统计学手段，是数据分析领域的重要方法。未来各个行业的生产力和核心竞争力的突破都依赖于将数据转化为技术和信息的能力和速度。随着各行各业的发展，数据量级的极大程度提高，对数据处理和分析的效率和精度有了更高的要求，一系列的机器学习算法应运而生。机器学习算法是通过挖掘数据背后的隐藏信息，运用数理统计的手段求解损失函数最优化的过程，不同结构的数据集要用不同的机器学习算法才能达到具有应用意义的效果[3]。在具有上述特点的时代背景下，房价问题已经与我国经济社会发展，民生福祉紧密相连。使用机器学习方法和统计学理论知识对二手房数据的研究具有现实意义。本文从二手房微观层面入手，使用杭州市二手房数据进行机器学习算法训练，构建杭州市二手房售价预测模型进行研究，并依托该预测模型开发系统，支持历史房屋交易数据的查询、房间行情概览、小区数据查询以及房价评估预测等功能。

1.2 研究意义

房屋作为遮风避雨的地方，对于每一个人来说几乎都是不可或缺的，随着经济的发展和人们生活水平的提高，不少家庭、不少人名下有着不只一套房产。因此二手房交易市场是一个非常活跃的市场。二手房的价格受到多种因素的影响，如房龄、地理位置、小区、朝向、楼层、装修情况、房屋格局、建筑结构等。因此，预测二手房的交易价格成为了一个具有挑战性的问题。

对二手房交易价格进行研究，可以增加二房市场价格的透明度，帮助买卖双方基于更准确的信息进行交易决策，减少信息不对称导致的不公平交易，合理的评估体系能够确保二手房的定价更加公正、合理，反映房屋的真实价值，保护消费者免受高价低质或欺诈行为的影响，同时也保障了卖方的合法权益。对于金融机构和保险公司而言，准确的二手房估值有助于他们在提供贷款、保险服务时更好地评估风险，制定合理的费率，减少因房屋价值评估不当造成的经济损失。对于政府和监管机构可以根据价格研究结果来制定或调整

相关政策，如税收政策、环保鼓励措施等，以引导市场健康发展。

总之，对二手房辆价格的研究有助于推动行业标准和规范的建立，比如评估标准、报告格式等，这对于规范市场秩序、打击非法交易具有重要意义，不仅对直接参与交易的双方有利，还对整个二手房市场乃至宏观经济都有着深远的影响。

1.3 研究方法

本研究实验的数据来源与阿里天池大数据集，为链家二手房交易平台在杭州市的交易数据，数据集包含约乎 30000 条交易数据，共 17 个字段，如小区名称，所在楼层，建筑面积等。具体字段情况详见表 1-1。

对于得到的数据集，我们首先将其格式进行统一，保证单位一致，然后数据集的状况进行统计分析，包括数据集各字段的数据类型、缺失值、异常值，对于数值型变量，我们还对其平均值、极差、标准差进行统计，对于类别型的变量，我们对其众数进行统计。统计完成后我们对数据集的缺失值以及异常值进行处理，处理完成后我们对数据的分布特征进行分析，并绘制出其分布图。

将数据集预处理完成后，进入特征工程，进行新特征的构建。从任务目标来看，很显然这是一个回归问题。本文章的大致研究思路为使用随机森林、XGBoost、以及线性回归等算法进行预测，对模型的效果进行评估，并选择出其中预测效果最好的算法。

基于模型建立系统，配合数据库使用 python 作为后端语言，vue 框架构建前端，实现房价信息概览、小区信息查询、历史交易数据查询以及房价评估预测等功能。

表 1 - 1 字段说明

Field	**Description**
小区名称	房屋所在小区名称
所在楼层	房屋所在楼层
建筑面积	建筑面积
户型结构	平层、复式等
建筑类型	板楼、塔楼等
房屋朝向	房屋主要朝向
建筑结构	钢混结构、框架结构等
装修情况	精装、简装等

表 1-1 字段说明（续）

配备电梯	是否配有电梯
梯户比例	电梯配置比例
交易权属	商品房、回迁房等
房屋用途	普通住宅、商住两用或者别墅
房屋年限	房屋完工到迄今为止的时间
产权所属	共有、非共有
所在城市	房屋所在城市
房屋户型	房屋为几室几厅
抵押情况	房屋是否有抵押以及抵押详情

1.4 本组织架构

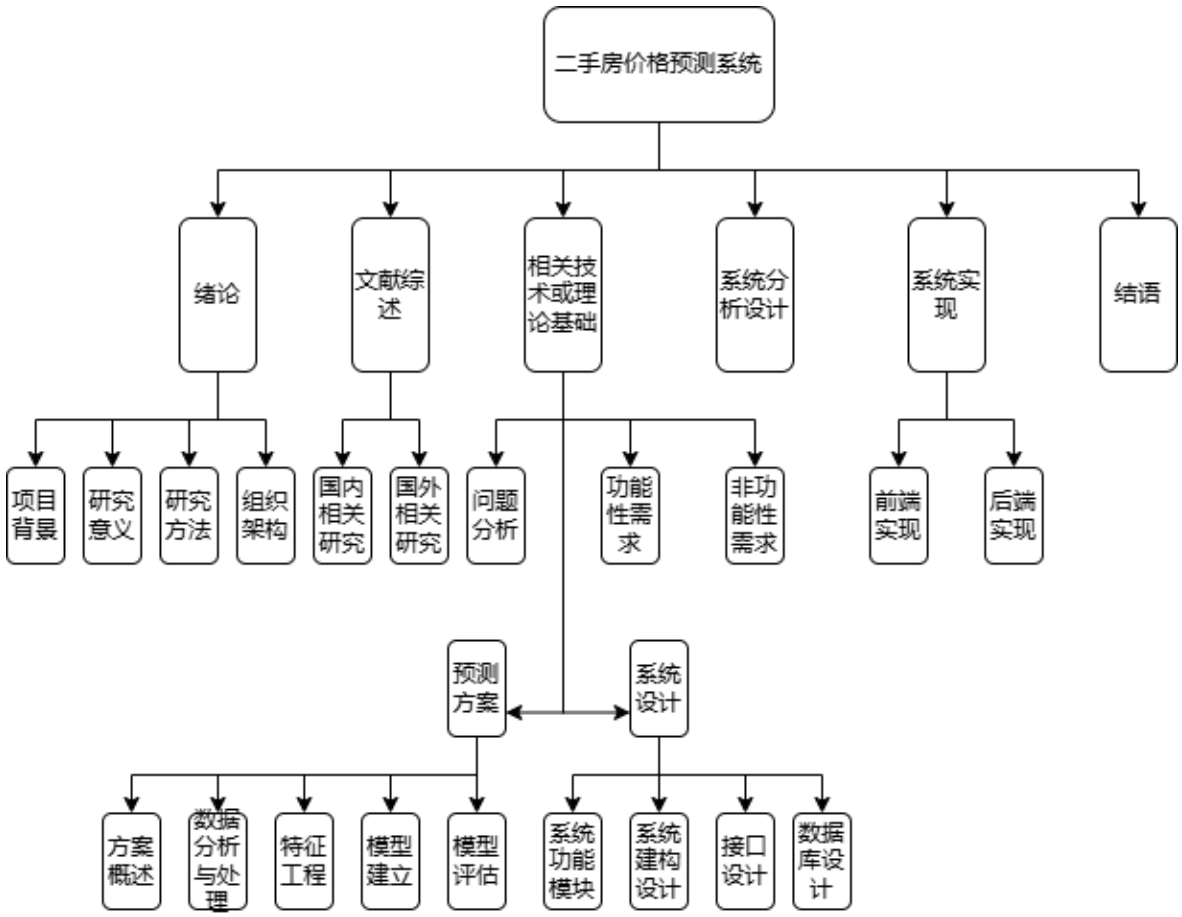

图 1.1 论文组织架构

第 2 章 文献综述

2.1 国内相关研究

考虑宏观因素对于房屋价格的影响。在政策对于房屋价格的影响方面，张梦和施同兵[4]在 2020 年研究了土地政策对于的房屋价格的影响，研究发现作为房屋成本组成部分的土地政策，会通过影响土地价格来对房屋价格产生较大影响；韦金洪和刘佳[5]在 2012 年探究了货币政策对于房屋价格的影响，建立 VAR 模型证实利率与货币供应量对房价有较大影响，其中货币供应量与房地产价格呈正相关。而利率对房地产价格的影响在短期和长期中是不一样的，短期会表现出正相关，而长期则表示为负相关；陈将浩[6]在 2014 年探究了消费者信心指数对于房屋价格的影响，通过研究发现，消费者信心指数对房价有正向影响，即房价随着消费者信息指数的增高而增高；汪雅倩和陈依萍[7]在 2016 年探究了 GDP 对于房屋价格的影响，利用最小二乘回归，分析出地区 GDP 与房屋平均销售价格成正相关关系的结论；徐建炜等人[8]在 2012 年探究了人口结构对于房屋价格的影响，研究发现老年人口比例的增加对房价呈负向影响，少年人口比例的增加对房价呈正向影响。吴齐林[9]在 2007 年探究了供需关系对于房屋价格的影响，结果显示房价的快速上涨是由于需求方的无序竞争与供给方的寡头垄断造成的；汪慧颖等人[10]在 2007 年同样探究了需求与供给方面对于房屋价格的影响，从供求关系角度得出武汉市二手住房价格的变化原因，并且预测武汉市房价的未来走势；谭刚[11]在 2001 年探究了经济发展对于房屋价格的影响，发现市场经济的发展速度会促进房地产行业的发展，即两者是正相关的关系。

考虑微观因素对于房屋价格的影响。刘冰等人[12]在 2017 年使用多元回归模型，基于南京八个地区的二手房数据来分析所选的八个微观变量对于房价的影响程度，研究发现对于二手房价格影响程度较高的两个变量是房屋所在区域以及住房有无电梯这两个变量；邝文竹[13]等人在 2012 年使用多元回归模型对青秀区的房价展开研究，研究发现楼龄和楼层对于房价的影响显著，但房价随着楼龄的增长呈现出不规则的变化；邹高禄等人在 2005 年通过建立城市住房 Hedonic 价格模型来分析住房特征和区位这两个变量对于房价的影响。发现住房面积对二手房价格的影响显著，并且当住房面积从小往大增加时，住房价格增加的边际效用呈现明显的递减趋势，周边环境特征也是住房价格评估中一个重要因素，而住房年龄、装修状况以及区位变化对于二手房价格的影响不显著；2017 年李晓童等人在

研究中选择了学校、地铁、经纬度、面积以及卧室数量等九个可能影响二手房房价的特征，建立随机森林模型得到这 9 个变量对于房价的影响程度，进而发现对北京市二手房价格影响最大的因素是什么；宋丹华在 2015 年的研究中发现房间数量、绿化程度以及交通状况等特征对于房价有正向影响，而建筑层数与医疗状况等特征则对房价有负面影响；王鹤在 2012 年探究了全国范围及东部、中部、西部地区的房屋价格，基于空间面板模型的分析结果，发现不同区域的房价影响因素各不相同；范莉丽也在 2015 年使用泰尔指数来证实了中国房价具有明显的区域性差异，并且区域间的差异在逐渐增大。

在分析房价问题所采用的研究方法上。郝丹璐在 2014 年对房价与影响因素之间的关联程度研究中，使用了灰色关联度分析和因子分析；王振宇在 2016 年使用逐步回归法和 ARIMA 模型对青岛房价的影响因素进行探究；甘霖等人在 2016 年分析北京市房价与地价之间的关系时选用了结构方程模型；许基伟和马欣在 2020 年使用 Hedonic 模型，分析建筑特征、邻里特征以及区位特征中的 10 个变量对于房价的影响情况。

除此之外周洪伟在 2016 年从时间序列角度出发，利用南京地区的 2013 年 2 月-2015 年 2 月的二手房房价数据，建立了二手房价格预测的随机模型，给出了二手房价格的预测公式，并作了统计的显著性检验，以期给交易双方提供参考。

庞枫在 2017 年以二手房价格的批量评估模型为出发点，首先从公共、房屋、区位、其他四个方面确定模型的特征变量，然后通过划定某一价格范围的比较楼盘池，使用 SVM 模型从中来定位出接近目标楼盘中的三个比较楼盘，使其组成一个高同质化的评估区域，最后以评估区域为单位使用随机森林来预测其中的待估楼盘的价格。结果显示该方法可以显著降低误差。

陈嘉彤、左剑凯和陈铖在 2018 年在特征选取方面从房产供求关系、社会因素和经济因素三个角度中挑选了七个特征变量，建立 BP 神经网络和 NAR 神经网络 9 对海口市 2007 年至 2017 年的数据进行建模，然后通过设置权值将两个模型组合起来，结果显示与单一模型相比，组合模型的误差更小，波动幅度也更加稳定。

郑永坤和刘春在 2018 年通过爬取广深两地的二手房月度均价数据，然后建立 9 NAR 神经网络表示非线性自回归神经网络，在处理具有非线性与时变性的时间序列具有优势。

何飞在 2018 年借鉴国内外文献研究，分析并选取了一些影响上海市二手房价格的影响因素，其中包括 2011~2018 年的上海市房地产开发投资额、常住人口数等统计年鉴季度数据。然后，构建单隐藏层的 BP 神经网络模型，按照 8:2 的比例划分出训练集与测试集。最后根据在测试集上的误差，来证明模型的实用性，并对未来的房价进行季度预测。

龚洪亮在 2018 年更多地关注二手房价格的影响因素及其重要性。首先通过两个途径收集原始数据，一是武汉市链家网上的二手房屋属性，由爬虫获取；二是二手房的区位因素，由百度 API 来查询二手房周边配套的餐饮交通数据。随后建立了极端梯度提升树模型，并且和 LASSO 模型的效果作比较，结果显示极端梯度提升树模型具有较为明显的优势。

刘洋在 2018 年主要利用特征价格理论，通过爬取上海市近 3 万条二手房数据，分别构建多元线性回归模型和随机森林模型并进行预测，结果显示随机森林模型的预测精度优于多元线性回归模型。之后又选取 2004-2017 年上海市二手房价格月度指数，利用时间序列 R/S 分析法 10 对其进行研究，验证了上海市二手房价格指数具有长期记忆性并计算得出其平均循环周期，最后为消费者购房、户主卖房以及政府决策提出意见。

黄明宇和夏典在 2019 年通过合肥市链家网收集 2017 年 3 月至 2018 年 3 月的二手房交易数据，经过预处理后共 9185 条数据。然后以房屋价格为因变量，区域均价、装修程度、房屋朝向、户型等为自变量建立多元线性回归模型，并经过十折交叉验证评估模型的泛化能力，最终得出高拟合度的二手房价格预测模型。

徐阳阳在 2019 年使用济南市 2018 年 1 月到 2018 年 12 月已成交二手房数据，进行一房一价的预测。首先利用灰色关联分析从原始数据中选择合适的预测指标，随后建立三层 BP 神经网络模型，其模型的拟合优度为 96.27%。

张学新和吴凯泽在 2019 年使用武汉市二手房产中介安居客发布的文本结构数据，从中提取特征关键词以及对变量进行数值化编码，然后根据分类变量较多的情形构造出一种新线性回归模型，并使用统计检验来衡量不同特征对房价预测的重要性，得出前四位重要特征，即所在区域、装修等级、房龄和房屋类型，最后从样本中得出一些置信度大的推理规则来简略地判断房价区间。

2.2 国外相关研究

在宏观因素对于房屋价格的影响方面。Apergis[14]从货币政策的角度出发，研究发现货币供给量与其他宏观因素的共同作用会对房地产价格有重要影响；Elbourne[15]同样从货币政策的角度出发，研究发现货币供给对于房价的影响很大；Abraham 和 Hendershott[16]在 2004 年从就业率对房价影响的角度展开研究，研究表明就业率以及居民收入都与房价变动密切相关，并且利率与房价呈现负相关的变动趋势。Kau 和 Keenan[17]在 1995 年的研究中发现住宅需求与利率之间是反向关系；在通货膨胀对于房屋价格的影响方面，

Brueggeman 等人[18]在 1984 年通过建立房地产资产定价模型，发现物价指数对房地产报酬率有正向影响，因此认为投资房产能够有效消除通货膨胀导致的货币贬值风险；Miller[19]使用夏威夷的数据进行汇率对房价的影响，结果表明随着日元兑美元升值，夏威夷的住宅市场价格明显上涨。

在微观因素对于房屋价格的影响方面。Bitter 等人[20]在 2007 年将空间自相关引入特征价格模型以及空间误差模型，分析结果表明空间位置对住房特征的边际价格有显著影响。Hendershott[21]在 1996 年的研究中发现居民的税后收入、就业率以及房屋建造成本对于房价有着显著的影响。Thibodeau[22]在 2003 年对房屋规模及所在位置对于房价的影响展开研究，特征价格方程的结果表明这两者对于房价均有显著影响。Oates[23]在 1969 年对新泽西东北部城镇的房价展开研究，结果显示地方公共服务是购房者的首要考虑因素，并且地方公共支出对于房价呈正向影响。

在研究房价问题所选用的模型方面。Shim[24]等人在 2004 年提出了半参数空间效应核最小二乘误差模型和半参数空间效应最小二乘支持向量机来估计特征价格函数，并将研究结果应用于住房价格预测。Gao 等人[25]在 2007 年的研究与 Kim 等人[26]在 2015 年的研究中，均使用了基于 Box-Cox 的回归方法对房屋价格进行预测。Liu 等人在 2011 年使用支持向量模型来分析房价影响因素，该模型分析效果较好，并且在方法与结论上都为之后房价问题的研究提供了新思路。在 2012 年，Antipov 等人首次尝试使用随机森林模型对房价进行预测研究，并将其结果与多元线性回归、决策树以及人工神经网络的预测效果进行对比，发现随机森林模型的预测效果最好。

Evgeny A and Elena B 在 2012 年考虑到随机森林模型的稳定性，将其与住宅价格评估研究相结合。经实证分析，随机森林的模型误差显著低于多元回归、人工 10 R/S 分析法主要通过定量比较的方法来分析时间序列的分形特征和长期记忆过程。

Vincenza C, Leonardo C and Mario M 在 2014 年将基于人工神经网络的房地产估价模型应用于房地产估价，并进行敏感性分析以确定最重要的输入变量，其中强调了环境质量对房屋价格的作用。

Fotheringham A S, Crespo R and Yao J 在 2015 年认为传统的 Hedonic 价格预测模型忽略了空间效应，提出将将空间经济学、统计学和地理信息科学相结合，使用一种局部空间建模技术—地理加权回归（GWR），应用于伦敦 1980-1998 年房价数据集，以探索伦敦房价的时空分布及其决定因素。

Park B and Bae J K 在 2015 年研究分析了弗吉尼亚州费尔法克斯县的 5359 栋联排别墅

的住房数据，并基于 C4.5、Ripper11、朴素贝叶斯和 AdaBoost 等机器学习算法开发了房价预测模型，并用十折交叉验证比较了其分类准确度性能。实验证明，基于准确性的 RIPPER 算法在住房价格预测方面始终优于其他模型。

Lowrance R E 在 2015 年意图建立洛杉矶县 2003 年至 2009 年住宅房地产价格的最佳线性模型。特别的是，该模型添加了一些地理因素和时间因素特征，如经纬度、销售日期、所处地段。然后通过测试确定该模型的最佳形式以及产生最小误差的训练周期，特征和正则化器。实证分析显示，最佳线性模型的精度显著高于随机森林。

Eman A and Mohamed M 在 2016 年认为房价的最终价格是由买家进行视觉评估从而估算出的。因此提出从房屋照片中提取视觉特征，并将其与房屋的文本信息相结合，然后构建多层人工神经网络模型来训练数据，将房价作为其单输出进行估计。实验表明，与纯文本特征相比，添加视觉特征使 R2 增加了 3 倍。

Liu J, Yang Y, Xu S, et al 在 2016 年认为非欧几里德距离度量可以改善地理加权回归（GWR）模型 12 的拟合，探索出一种考虑时空非平稳性的地理时间加权回归（GTWR）方法来估计基于出行时距度量的房价。结果表明，GTWR 模型具有较高的拟合优度和较强的时空解释能力，可为房地产管理制定更为有效的政策提供参考。

Yusupova A, Pavlidis N G and Pavlidis E G 在 2016 年在动态模型平均模型的基础上，采用随机优化算法，顺序更新每个动态线性模型的遗忘因子，并使用最新非参数模型组合算法，开发出具有自适应性的动态平均模型（ADMA）。在实证部分，使用 1982 年第一季度至 2017 年第四季度英国 13 个区域性住房市场季节性调整后的区域房价指数。研究结果表明，ADMA 模型与 DMA 模型相比，总体上可在有限时间内提供更准确的预测。

第 3 章 相关技术或基础理论

3.1 预测基础理论

通过机器学习的相关技术来做预测。它的核心思想是通过分析和理解数据中的模式和规律，从而使计算机能够根据以往的经验来做出预测。

3.2 XGBOOST 算法

GBDT 算法的运行往往要生成一定数量的树才能达到令人满意的准确率。当面对的数据集结果较为庞大且复杂时，可能需要进行上千次的迭代运算，还会造成一定的计算瓶颈，并增加了计算空间的消耗。华盛顿大学的陈天奇博士研发出了 XGBoost（eXtreme Gradient Boosting）解决了这一技术难点，此算法基于 Gradient Boosting Machine 框架，并使用 C++ 实现，从而极大地提升了模型训练速度和预测精度。

XGBoost 是一个优化的分布式梯度增强库，作为 GBDT 模型的升级版，集高效性，灵活性和便携性等特点于一身[27] 。利用 XGBoost 模型可以在较短的周期内解决数据科学问题，往往能够得到较高精度的实验结果。 利用 XGBoost 算法构建的数学模型单台机上运行速度比现在经常使用的数学模型的训练速度快十倍以上，并且在分布式模式或内存设置需要限制时还是可以取得较为准确的实验结果。

XGBOOST 第 t 棵树的目标函数为：

$$obj(\theta) = \sum_{i}^{n} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.1}$$

其中 为实际结果， 为模型的输出结果，右边加上了表示模型复杂度的正则化项，目的在于惩罚复杂的模型，通过减小树的深度和单个叶子结点的权重值以减缓过拟合。

根据 Boosting 的原理，第 t 棵树对样本 i 的预测值=前 t-1 棵预测树的预测值+第 t 棵树的预测值。

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i), 0 < \eta \leq 1 \tag{3.2}$$

模型每次走一小步逐渐逼近结果的效果，要比每次迈一大步很快逼近结果的方式更容易避免过拟合。就是说它不完全信任每一个棵残差树，它认为每棵树只学到了真理的一小部分，累加的时候只累加一小部分，通过多学几棵树弥补不足。即给每棵数的输出结果乘

上一个步长η。

将公式（3.2）带入公式（3.1）并根据二阶泰勒展开公式进行展开得到如下：

$$obj^{(t)} \approx \sum_{i=1}^{n}\left[L(y_i,\hat{y}_i^{(t-1)} + g_i \cdot f_t(x_i) + \frac{1}{2}h_i \cdot f_t(x_i)^2)\right] + \Omega(f_t) + Cons\,tan\,t \tag{3.3}$$

其中

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)})$$

g_i和h_i分别表示预测误差对当前模型的一阶导与二阶导，$l(y_i,\hat{y}_i^{(t-1)})$为前$t-1$棵树组成的学习模型的预测误差，$L(y_i,\hat{y}_i^{(t-1)})$为常数，忽略常数项对公式（3.3）进行化简得到ru下：

$$bj^{(t)} = \sum_{i=1}^{n}\left[g_i \cdot w_{q(x_i)} + \frac{1}{2}h_i \cdot w^2_{q(x_i)}\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} {w_j}^2 \tag{3.4}$$

其第一部分为对所有训练集样本进行累加，此时所有样本均为映射为树的叶子节点。因此从叶子节点出发，对所有的叶子节点进行累加，得：

$$obj^{(t)} = \sum_{i=1}^{n}\left[G_j \cdot w_j + \frac{1}{2}(H_j + \lambda)\cdot w_j^2\right] + \gamma T$$

$$\sum_{i\in I_j} g_j = G_j, \sum_{i\in I_j} h_j = H_j \tag{3.5}$$

G_j表示映射为叶子节点j的所有输入样本一阶导之和，同理，H_j表示二阶导之和。最后得到最终目标函数：

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{3.6}$$

XGBoost在构建回归树时采用贪心法进行树节点的分裂，首先从树深为0开始对每个叶子节点尝试进行分裂；其次在每次分裂后，原来的一个叶子节点继续分裂为左右两个子叶子节点，原叶子节点中的样本集将根据该节点的判断规则分散到左右两个叶子节点中；最后新分裂一个节点后，我们需要检测则此分裂是否会给损失函数带来增益，增益的定义如下：

$$Gain = Obj_{L+R} - (Obj_L + Obj_R) = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{3.7}$$

如果增益 Gain>0，即分裂为两个叶子节点后，目标函数下降了，那么我们会考虑此次分裂的结果。

XGBoost 树的分裂停止情况如下：

情况一：根据推导得到的打分函数是衡量树结构好坏的标准，因此，可用打分函数来选择最佳切分点。后根据增益判断是否分裂，若增益为负就停止。

情况二：当树达到最大深度是，停止建立，因为树的深度太深容易出现过拟合，这里需要设置超参数 max_depth。

情况三：当引入以此分裂后，重新计算新生成的左右两个叶子节点的权重和。如果任何一个叶子节点的样本权重低于某个阈值，也就放弃此次分裂。这涉及到一个超参数最小样本权重和，是指一个叶子节点包含的样本数量太少也会放弃分裂，防止树分得太细，这也是防止过拟合的一种措施。

XGBoost 对比传统的 GBDT 其精确度更高，并且其灵活性也更强，除了决策树以外，XGBoost 还支持线性分类器，支持自定义损失函数，由于加入了正则化以及缩减的思想，即不完全信任每一棵树，而是一点点的累加去逼近结果，因此 XGBoost 能够很好的避免过拟合，但与此同时，XGBoost 计算的时间以及空间复杂度都相当的高。

3.3 模型融合

模型融合是通过组合多个机器学习模型的预测结果，以提高整体性能，常见的模型融合方法有加权平均、投票法、堆叠法以及模型组合等，其中投票法适用于分类任务，平均法更适合于回归任务，加权平均法进其权重占比公式如下：

$$w_j = \frac{MAE_j^{-1}}{\sum_{k=1}^{n} MAE_k^{-1}} \tag{3.8}$$

其中，w_j表示第j个模型的权重，MAE_j^{-1}为第j个模型 MAE 的倒数，这样效果好的模型（即 MAE 小的模型）会提供更多的贡献，MAE 为绝对平均误差，其计算公式如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3.9}$$

其中，y_i为第i个样本的真实值，$\hat{y}_i$为对该样本的预测值，N为样本众数。最后模型融合

的公式如下：

$$\hat{y}_{ensemble} = \sum_{j=1}^{n} w_j * \hat{y}_j \tag{3.10}$$

$\hat{y}_{ensemble}$为融合后的预测值，w_j是第j个模型的权重，$\hat{y}_j$是第j个模型的预测值。

本文拟采用 Stacking 方式进行模型融合，Stacking（叠加）是一种集成学习方法，通过将多个基础模型的预测结果结合起来，以提高整体预测的准确性。其基本原理如下：

数据集划分：首先，将训练数据集划分为多个子集，通常是两个或更多个。这些子集可以按照不同的方式划分，例如随机划分或使用交叉验证。

(1) 基础模型训练和预测：对于每个子集，使用不同的基础模型进行训练和预测。这些基础模型可以是任何机器学习算法，如决策树、支持向量机、神经网络等。每个基础模型都会对子集进行训练，并生成对未知数据的预测结果。

(2) 特征组合：将每个基础模型的预测结果作为新的特征，组合成一个新的训练数据集。这些预测结果可以作为原始特征的补充，提供更多的信息来训练元模型。

(3) 元模型训练：使用新的训练数据集来训练一个元模型。元模型可以是任何机器学习算法，如逻辑回归、随机森林、梯度提升等。元模型的目标是学习如何结合基础模型的预测结果，以最大化整体模型的准确性。

(4) 预测：最后，使用训练好的元模型来对测试数据进行预测。元模型会根据基础模型的预测结果进行组合和加权，生成最终的预测结果。

通过 Stacking，不同基础模型的优势和特点可以得到充分的发挥，从而提高整体模型的性能和泛化能力。然而，Stacking 也需要进行适当的调参和验证，以避免过拟合和提高模型的稳定性。Stacking 作为一种集成学习方法，具有以下优点和缺点：

优点：

(1) 提高预测性能：通过结合多个基础模型的预测结果，Stacking 可以显著提高整体模型的准确性和泛化能力。它可以充分利用不同模型的优势和特点，从而提供更准确的预测结果。

(2) 灵活性：Stacking 可以集成各种类型的模型，包括线性模型、非线性模型、树模型等。这使得 Stacking 非常灵活，可以适应不同类型的数据和问题。

(3) 可解释性：与其他集成方法相比，Stacking 在生成最终预测时使用了多个模型的预测结果。这可以提供更多的信息来解释预测结果，使得模型的解释性更强。

缺点：

(1) 计算资源和时间消耗：由于 Stacking 涉及到多个模型的训练和预测，因此需要更多的计算资源和时间。这可能限制了 Stacking 在大规模数据集和实时预测任务中的应用。

(2) 风险过拟合：如果不适当地使用 Stacking，可能会导致过拟合问题。当基础模型过于复杂或训练数据集过小时，Stacking 容易过度依赖训练数据，导致在测试数据上的性能下降。

(3) 超参数选择：Stacking 需要选择合适的模型和超参数配置，以获得最佳的性能。这可能需要进行大量的实验和调参，增加了模型的复杂性和训练的难度。

Stacking 作为一种集成学习方法，在提高预测性能和灵活性方面具有明显的优势。然而，它也存在一些限制，如计算资源和时间消耗、风险过拟合以及超参数选择的挑战。因此，在应用 Stacking 时需要权衡其优缺点，并根据具体问题和数据情况进行适当的调整和优化。

3.4 VUE.js 框架

Vue.js 是一套构建用户界面的渐进式框架。与其他重量级框架不同的是，Vue 采用自底向上增量开发的设计。Vue 的核心库只关注视图层，并且非常容易学习，非常容易与其它库或已有项目整合。另一方面，Vue 完全有能力驱动采用单文件组件和 Vue 生态系统支持的库开发的复杂单页应用。

从历史的潮流来说，人们从之前的:原生 JS -> Jquery 之类的类库 -> 前端模板引擎 ,他们都有一个共同的特点需要我们去操作 dom 元素。近几年来，得益于手机设备的普及和性能的提升，移动端的 web 需求大量增加，产生了一种叫 webapp 的东西，也就是移动端的网页应用。为了更好满足当前移动 webapp 项目的开发需求，MVVM 框架诞生。

Vue 具有响应式编程和组件化两大特点，Vue 是一个轻量级的前端框架，具有简单易学、双向数据绑定、组件化、数据和结构的分离、虚拟 DOM、运行速度快的特点。vue 是单页面应用，使页面局部刷新，不用每次跳转页面都要请求所有数据和 dom，这样大大加

快了访问速度和提升用户体验。而且他的第三方 ui 库很多节省开发时间。

(1) 响应式编程:而是指 vue.js 会自动对页面中某些数据的变化做出响应，通过 MVVM 思想实现数据的双向绑定，让开发者不用再操作 dom 对象，有更多的时间去思考业务逻辑。

(2) 组件化: Vue 将组成一个页面的 HTML，CSS 和 JS 合并到一个组件中，可以被其他组件或页面引入而重复利用。通常每个.Vue 文件作为一个组件导出，组件可以作为基础组件（如按钮）或一个页面（如登录页面）。组件化很好的将一个庞大复杂的前端工程拆分为一个个组件，重复利用的性质也大大提高了开发的效率。

(3) 响应式虚拟 DOM: 对于 DOM 来说，当 HTML 的一个元素（如 div）需要响应数据更改时，会刷新整个页面，导致效率堪忧。对于虚拟 DOM，浏览器会将 HTML 文件转换为 JS 文件并复制一个额外使用（虚拟）。对于任何更改，虚拟 DOM 都将复制的 JS 与原始 JS 进行比较，只重新加载更改的部分，局部修改到真实 DOM 上。在 Vue 中，每个绑定 data 属性的组件都有一个 Watcher 检测 data 属性的变化。一旦检测到改变，则重新渲染该组件，这就是响应式。

(4) 生命周期: 每个 Vue 组件都有生命周期，过程为创建 -> 挂载 -> 更新 -> 销毁。开发者可以通过钩子函数（如 mounted）在组件生命周期中的不同时刻进行操作。

3.5 Flask 框架

Flask 是一个使用 Python 编写的轻量级 Web 应用框架，它简洁而灵活，适用于开发小型至中型的 Web 应用。本文将介绍 Flask 框架的基本概念、特点以及如何使用 Flask 来快速搭建 Web 应用。Flask 是基于 Werkzeug 和 Jinja2 库构建的，它遵循了 MVC（模型-视图-控制器）的设计模式。Flask 的核心思想是保持简洁和易用，它提供了一些核心功能，但也允许开发者通过扩展来添加更多功能。

Flask 框架具有以下优势:

(1) 简单易学:Flask 的 API 设计简洁明了，学习和上手容易。开发者无需学习复杂的框架概念，只需了解几个核心概念即可开始开发。

(2) 轻量级灵活:Flask 没有过多的依赖，它的核心功能非常精简。开发者可以根据需要选择适合自己项目的扩展，使得框架更加灵活。

(3) 易于拓展: Flask 提供了丰富的扩展库，可以轻松集成常用的功能，如数据库访问、

表单验证、身份认证等。开发者可以根据需求选择适合自己项目的扩展，快速实现功能。

(4) 模板引擎支持:Flask 集成了 Jinja2 模板引擎，使得前后端分离更加方便。开发者可以通过模板引擎将逻辑和界面分离，提高代码的可维护性和可读性。

(5) 多种数据库支持:Flask 支持多种数据库，如 SQLite、MySQL、PostgreSQL 等。开发者可以根据项目需求选择适合的数据库，进行数据的存储和操作。

(6) 自动化测试:Flask 提供了测试客户端和测试工具，方便开发者进行自动化测试，保证代码的质量和可靠性。

Flask 是一个简单而灵活的 Python Web 框架，适用于快速开发小型至中型的 Web 应用。它的简洁的 API 设计和丰富的扩展库使得开发变得简单、灵活和高效。我们鼓励开发者们尝试使用 Flask 来构建自己的 Web 应用，体验其简单易用和强大的功能。

第 4 章 系统分析设计

4.1 问题分析

二手房价格预测评估是本系统的核心功能，对“二手房价格评估预测”这一需求进行分析，其主要涉及数据分析、机器学习和统计建模等领域，其核心目标是通过历史交易数据来预测二手房的未来交易价格[28]。在二手房市场中，房屋的交易价格受多种因素影响，包括但不限于所在城市、小区、房屋年限、楼层、装修情况、户型结构、建筑结构等。准确的价格预测不仅有助于买家避免过高支付，也能帮助卖家合理定价，促进市场的透明度和效率。此外，对于金融机构评估贷款风险、保险公司制定保费政策等方面也有重要价值。

数据集特点：

(1) 数据通常来源于各大交易平台的二手房交易记录，可能包括匿名变量以保护用户隐私。

(2) 数据量较大，为模型提供了丰富的训练资源。

(3) 包含房屋基本信息（如单价、面积、楼层、城市）、房龄、交易信息（如交易时间）、以及可能的市场环境因素等。部分变量会被匿名化处理以保护数据安全和隐私。

技术挑战：

(1) 如何从原始数据中提取出有意义的特征是关键，以及处理类别变量等。

(2) 由于数据来源多样，可能存在缺失值、异常值和噪声，需要进行有效处理。

(3) 选择合适的回归模型，如线性回归、决策树、随机森林、梯度提升树（XGBoost、LightGBM、CatBoost）或者深度学习模型。

(4) 如何进行模型调参，综合考虑时间成本和模型效果，在有限的计算资源内尽可能取得最好的参数组合

(5) 模型不仅要准确，还要能够给出价格预测的合理解释，便于用户理解和信任。

实践步骤：

(1) 理解数据，探索性数据分析（EDA）了解各变量分布，识别相关性。

(2) 数据预处理，处理缺失值、异常值，编码分类变量。

(3) 特征工程，创建新特征，选择对预测最有价值的特征。

(4) 模型构建与训练，选择合适的模型进行训练，并使用交叉验证进行调优。

(5) 模型评估，使用测试集评估模型性能，如使用 RMSE、MAE 等指标。

(6) 结果解释与应用，将模型预测结果转化为易于理解的形式，指导实际交易。

综合以上的分析，“二手房价格评估预测”该需求是对二手房的交易价格进行预测，二手房的交易价格是依据连续型的数据，所以，很显然该目标是一个回归任务。二手房交易价格预测是一个复杂的现实问题，它不仅要求技术人员具备扎实的数据处理和机器学习技能，还需要对二手房市场有一定的理解和洞察力。

除该需求外，本系统的其它需求主要为一些数据的搜索查询以及可视化，所用到的技术主要是 python、flask、JavaScript、html5 以及 css 等。

4.2 功能性需求

本系统为一个二手房房价信息以及房价评估预测，设计初衷其涵盖的功能有房价信息概览、小区信息查询、历史成交信息查询、房价评估预测，为防止系统服务器遭受到恶意攻击，其中小区信息查询、历史成交信息以及房价评估预测等功能只有当用户登录后才可以进行访问，从而移动程度上降低了服务器的压力避免系统瘫痪。系统的功能性需求主要包括以下:

(1) 数据输入/输出:用户能够输出特定类型的数据，系统能够正确的显示或输出数据，当用户不规范输入或错误输入信息是系统可以检测并给予提示。

(2) 数据存储与检索:能够存储用户数据和其它相关信息并快速准确的检索以存储信息

(3) 数据处理: 对输入的数据进行计算、分析或其他形式的处理, 根据业务规则对数据进行验证和过滤。

(4) 用户管理: 用户注册、登录和注销功能, 权限管理和角色分配。

(5) 搜索查询:允许用户通过关键字或其它标准搜索信息，提供高级查询功能以细化搜索结果。

(6) 安全性相关功能:登录和安全验证，防止未授权访问。

(7) 业务逻辑: 实现具体的业务流程和规则,支持复杂的决策过程，这里的具体需求包括能够查询获取到小区信息并进行可视化展示、能够查询历史交易数据以及提供房价评估预测服务等。

(8) 用户界面:易于使用的用户(GUI)界面，清晰的交互逻辑以及响应式的网页设计以适配不同的设备。

4.3 非功能性需求

非功能性需求描述了系统如何执行其功能，而不是系统具体要做什么。这些需求通常与系统的整体质量属性相关，包括但不限于性能、安全性、可靠性等。以下是本系统的非功能性需求分类及其具体内容:

(1) 性能需求:

- 响应时间:系统对用户请求的响应速度要在一定时间内
- 吞吐量:单位时间处理的数据量或事务数足够

(2) 可用性:

- 系统在绝大部分时间内都应当处于正常运行的状态
- 从故障发生到解决应当在短时间的处理完成
- 界面友好，用户学习和使用成本低

(3) 可靠性:

- 确保数据在存储和传输的过程中不会被损坏或者丢失
- 系统能够在部分组件失效的情况下继续提供服务

(4) 安全性:

- 保护系统避免未授权访问、攻击防止数据泄露
- 记录用户重要操作以备审查

(5) 兼容性:

- 兼容不同操作系统、浏览器
- 遵循行业标准和技术规范

(6) 可维护性:

- 代码结构和质量有保障，易于理解和修改
- 便于进行单元测试和集成测试

(7) 合规性:

- 遵循相关法律法规和行业标准
- 数据隐私法规遵守(如 GDPR、CCPA)

4.4 预测方案

4.4.1 方案概述

基于上述对“二手房交易价格预测”该题目的分析以及查阅国内外相关研究文献，对现有研究的分析、总结本文提出以下方案：对于取得的数据集首先进行探索性数据分析(EDA)，统计数据集中的缺失值、异常值以及数据分布，并识别其相关性，然后对数据集进行预处理，包括但不限于缺失值、异常值的处理，处理完成后对数据集进行特征工程构建更多有价值的特征并将构建好的训练集输入到算法模型中进行训练，最后得到结果并对模型进行评估，其流程图如下:

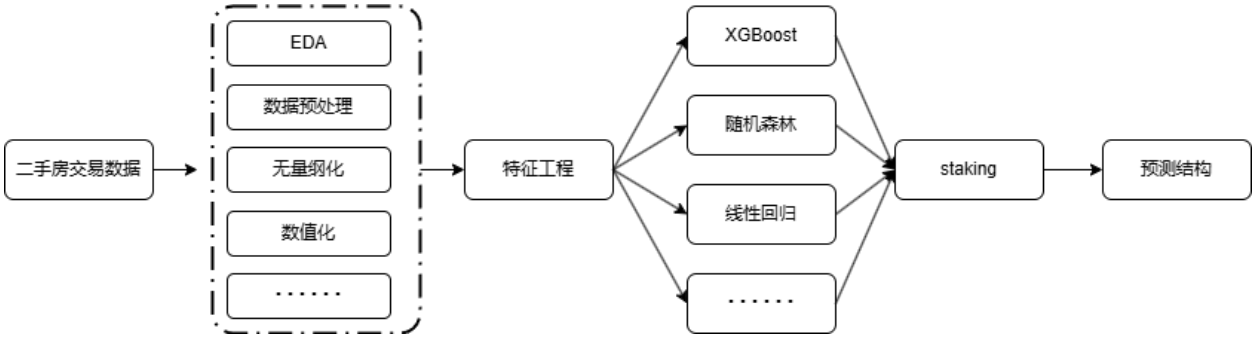

图 4.1 方案流程图

从流程图分析本方案主要有以下步骤：

(1) 数据预处理:首先对二手房交易数据进行预处理，包括探索性数据分析（EDA）、缺失值处理、异常值处理、无量纲化以及数值化等，因为在数据集中可能存在缺失值，部分算法是不支持缺失值的，并且由于数据来源于不同的平台，我们应该对其格式进行统一并做无量纲处理以保证数据的一致。

(2) 特征工程:在数据预处理的基础上，进一步进行特征工程，这可能涉及到特征选择、特征构造或降维等操作。

(3) 模型训练与集成:使用多种机器学习模型进行训练，包括但不限于 XGBoost、随机森林、决策树、线性回归等。对以上的结果进行对比分析，选择出其中效果最为理想的几个模型，并对其进行集成，在这里使用的 stacking 集成学习方式，其一般分为两层结构，第一层一般为 XGboost、CatBoost 以及等更复杂的学习器，第二层一般为决策树或者线性回归等比较简单的模型。

(4) 预测结果:最后，利用训练好的模型对新的二手房交易数据进行预测，并输出相应的结果。

4.4.2 数据分析与预处理

首先对数据集进行初步的探索性分析，了解数据集各字段的数据类型，数据集中有多少缺失值。对数据集有了基本的了解后对数据集进行预处理，处理其中的缺失值以及异常值，对于数值型的特征我们以三个标准差作为界定范围，考虑到数据总量较大，为了避免少量的极端值对模型造成影响，对超出三个标准差的数据进行剔除；对于数据集中缺失值分数值型和类别型进行处理，对于数值型的缺失使用均值进行填充，对于类别型的缺失使用众数替代。

数据预处理完成后对数据集进行大致的分析，绘制杭州市二手房单位价格直方图，如图 4.2 杭州市二手房单位价格直方图所示，绝大部分房屋的单价都在[20000,60000]的区间内，只有少部分的房屋单价低于 20000 或者高于 60000。

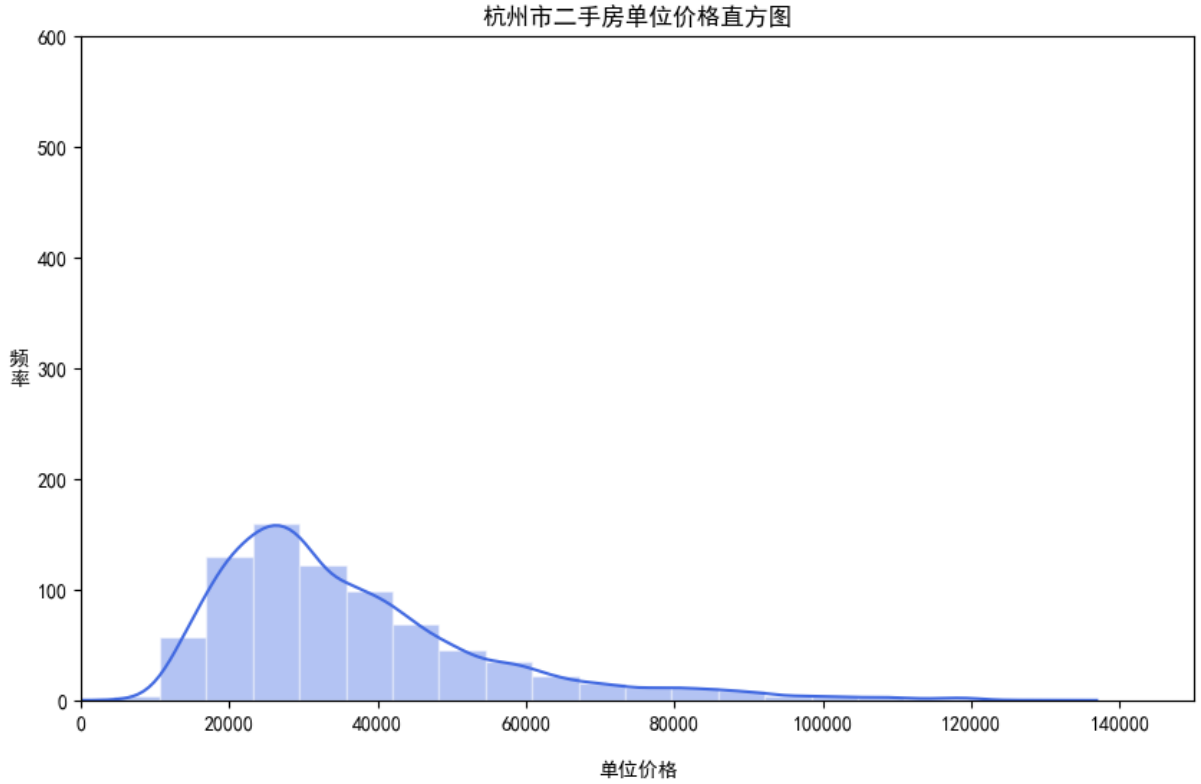

图 4.2 杭州市二手房单位价格直方图

杭州市房屋面积直方图如图 4.3 杭州市二手房建筑面积直方图所示，可以看出绝大部分房屋的面积都在[50，150]的区间内，只有少部分房屋的面积小于 50 平方米或者大于 150 平方米。

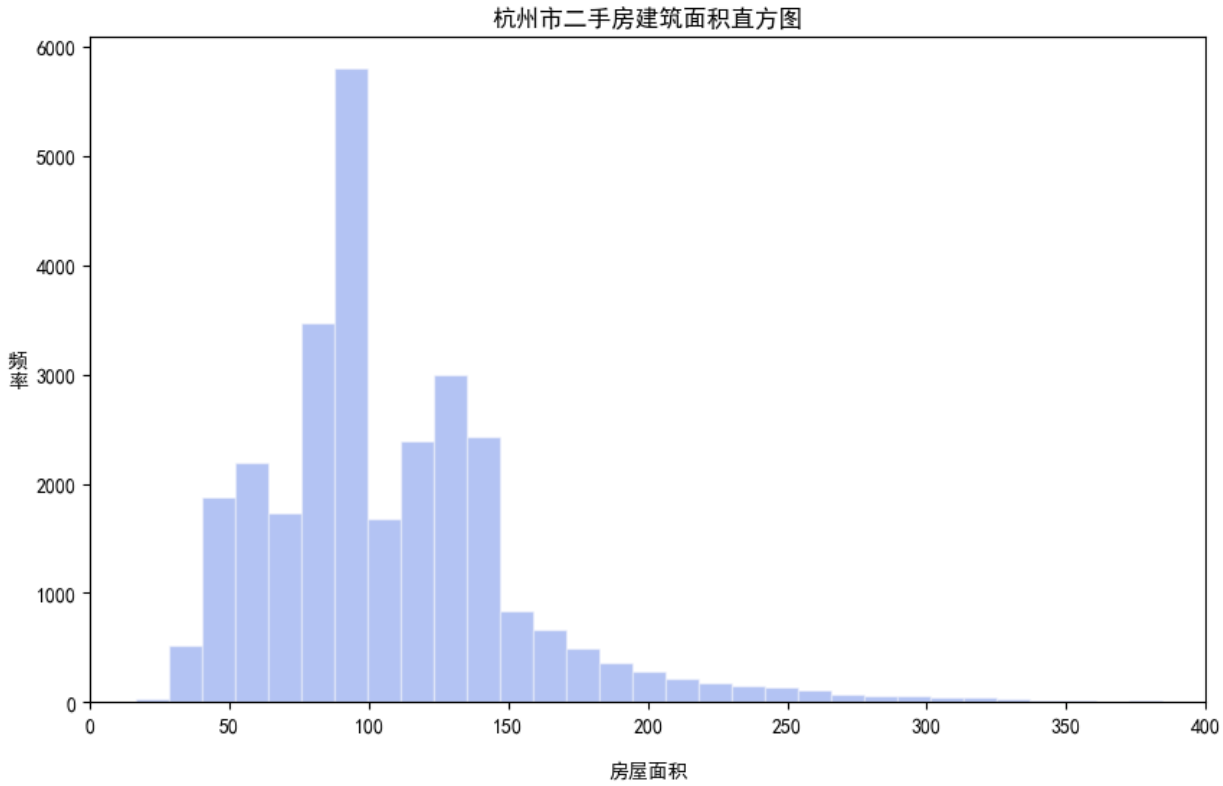

图 4.3 杭州市二手房建筑面积直方图

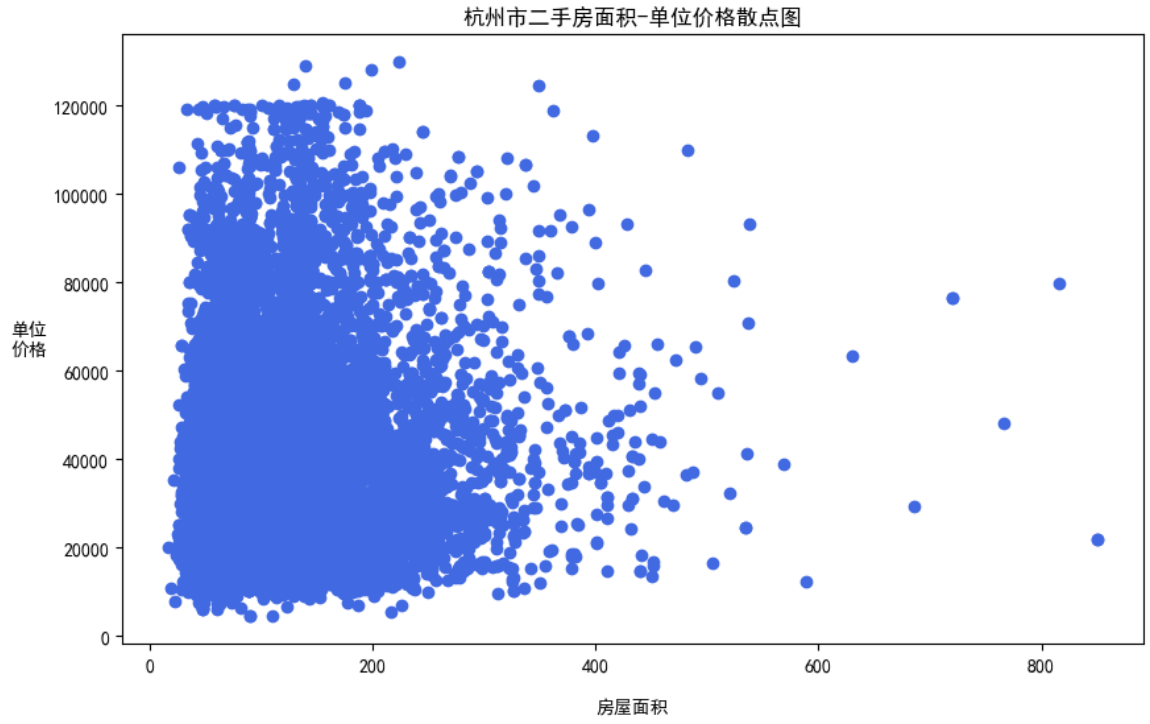

图 4.4 杭州市二手房面积—单位价格散点图

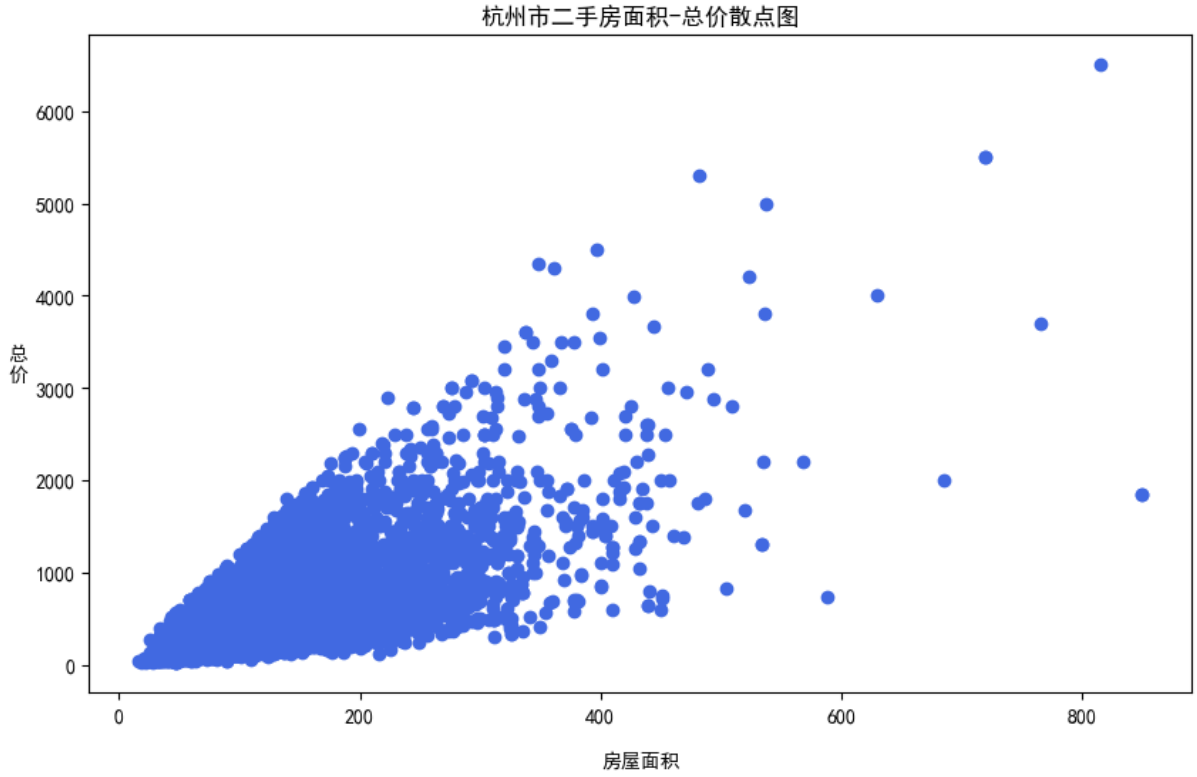

图 4.5 杭州市二手房面积—总价散点图

分别绘制杭州市二手房面积与单价和面积与总价的散点图，如图 4.4 杭州市二手房面积—单位价格散点图图 4.5 杭州市二手房面积—总价散点图所示，结合两张图可以看出，房屋的单价与面积关系不大，面积小的房屋也可能因为地理位置等其它原因导致单价更高，而房屋的总价基本与面积成正比，面积越大的房屋其总价一般情况下也越高。

4.3.3 特征工程

经过上述过程处理的数据已经基本达到规范，可以用于算法模型进行学习，但上述结果直接用于学习的效果并不理想，还需要进行特征工程操作。数据决定了机器学习的上限，而算法只是尽可能逼近这个上限”，此句话中数据的含义是对原始数据经过特征工程转换和处理等一系列操作得到的数据，而特征工程简单来说，就是把历史源数据转变为算法所需要的训练数据这么一个过程，进行这个过程的目的则是更好将原始数据转化为可以量化并且分类更加细致，有利于数据可视化分析，模型建模等后期处理。特征工程主要是由特征构建、特征提取、特征选择三个部分。数据中的特征对预测的模型和获得的结果有着直接的影响。可以这样认为，特征选择和准备越好，获得的结果也就越好。

并不是所有的属性都可以看做特征，区分它们的关键在于看这个属性对解决这个问题

有没有影响。可以认为特征是对于建模任务有用的属性。 表格式的数据是用行来表示一个实例，列来表示属性和变量。每一个属性可以是一个特征。特征与属性的不同之处在于，特征可以表达更多的跟问题上下文有关的内容。特征是一个对于问题建模有意义的属性，如果认为一个特征没有意义是不会被认为是特征的，如果一个特征对问题没有影响，那就不是这个问题的一部分。

特征重要性，可以被认为是一个选择特征重要的评价方法。特征可以被分配一个分值，然后按照这个分值排序，那些具有较高得分的特征可以被选出来包含在训练集中，同时剩余的就可以被忽略。特征重要性得分可以帮助我们抽取或者构建新的特征。挑选那些相似但是不同的特征作为有用的特征。 如果一个特征与因变量高度相关，那么这个特征可能很重要。相关系数是比较通用的评估方法。 另外还可以通过预测模型算法来对特征进行评分。这些预测模型内部有这样的特征选择机制，比如 MARS，随机森林，梯度提升机，也同样可以得出变量的重要性。

1.特征的构建

特征构建是指从原始数据中人工找出一些具有物理意义的特征。应该花费大量时间去仔细研究反复研究原始数据，寻找对应属性的潜在表现形式和数据结构及数据属性之间的联系。本文使用到以下的方法进行新特征的构建：

(1) 构建时间特征:原始数据集中的包含二手房的首次交易时间，猜测二手房的交易价格可能与二手房的房龄成反比，因此我们将二手房的首次交易时间作为房屋的完工时间，计算到当前时刻房屋的房龄，单位为年，数据类型为int型。

(2) 构建分组统计特征：基于房屋的地理位置以小区构建分组统计特征，不论是对于新房的购买，还是二手房的销售，房屋的地理位置都是对房屋售价的一个重要影响因素。为此，本文使用地区进行分类，统计不同地区二手房的“平均单价”，数据类型为int整型，统计完成有并根据城市字段与数据集进行连接。除房屋的地理位置之外，房屋位于哪个小区也是影响二手房销售价格的一大重要因素。与上述相同，因小区进行分类，统计该小区房屋的平均关注度以及平均单价，数据类型为int，并将统计结果根据小区名与数据集进行连接。

2.特征选择

特征选择也是特征工程中重要的一环,为什么要进行特征选择，而不是一股脑的将所有特征列加入到训练集中？因为特征个数越多，模型也会越复杂，其泛化能力会下降,分析特征、训练模型的所需的时间也会增加，特征选择能够明显的改善学习器的精度，减少模型

训练时间，有效的避免维灾难问题。维度灾难指当特征维度超过一定界限后，分类器的性能随着特征维度的增加反而下降（而且维度越高训练模型的时间开销也会越大）。导致分类器下降的原因往往是因为这些高纬度特征中含有无关特征和冗余特征，因此特征选择的主要目的是去除特征中的无关特征和冗余特征，特征筛选常用的形式主要可以分为三大类，分别为 Filter、Wrapper 以及 Embedded 下面对三种方式进行详细介绍：

Filter（过滤法）：按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择特征的个数对特征进行筛选。对于每个特征x_i，计算对于标签y的信息量$S(i)$得到 n 个结果并按大小排序，输出前 k 个特征，其中评价指标可以是 Pearson 相关系数、卡方验证、互信息和最大信息系数以及距离相关系数等。

Wrapper（包装法）：根据模型的预测效果评分，每次选择或者排除若干特征。对于每一个待选的特征子集，都在训练集上训练一遍模型，然后在测试集上根据误差大小选择出特征子集。

Embedded（嵌入法）：先使用某种机器学习模型进行训练，得到各个特征的权重系数，将系数从大到小进行排列然后筛选，类似于过滤法，但特征的权重系数是通过训练的结果得出。

在本实验中，选择 Embedded（嵌入法）进行特征选择，选取 LightGBM 作为训练模型，进行多次迭代，累计特征的重要性，最后取平均值作为该特征的权重。得到特征的权重后对其进行规范化，使所有特征的权重之和为 1，然后对特征的重要性进行降序并对重要性进行累计求和，按需求可以对单个重要性大于指定值的特征进行选取，也可以选取累计重要性大于指定阈值的所有特征列。对以上的操作重复进行 10 次，得到 10 组特征集合，对这 10 组特征进行做交集，得到的特征值作为最后选择出来的特征。

4.3.4 模型建立

系统的的需求是预测二手房房价，因此我们使用括线性回归、随机森林以及梯度提升回归，由于缺乏具体的参考，使用线性回归、随机森林、XGBoost 等模型进行初步训练，进行 5 折交叉并使用 MAE 平均绝对误差作为评价指标进行评价，MAE 的公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \tag{4.1}$$

模型以及评分见表 4 - 1 模型评分

表 4 - 1 模型评分

模型	MAE
线性回归	42916523.35
K 近邻	60.19
决策树	36.33
随机森林	30.08
XGBoost	29.41.，

从得到的结果来看线性回归的效果最差，随机森林和 XGBoost 的效果比较理想。对于大规模数据集，梯度提升树 XGBoost 通常表现得很好，因此选取随机森林和 XGBoost 作为预测算法。

图 4.1 方案流程图给出了方案的大致流程图，在这本节对模型的建立方案进行详细的阐述，确定所使用的基学习器为随机森林、XGBoost，对于每个算法建立模型，使用五折交叉和验证，将数据集随机划分为 5 等份，每次利用其中 4 份进行训练，利用另外 1 份进行验证，最后将得到的 5 份结果合并作为该模型的预测结果，其流程如，采用交叉验证可以很好的避免过拟合的问题，上述的模型均采取该方法进行训练。

模型建立后进行调参优化，特征工程决定模型的上限，而调参的目的就是取不断的逼近这个上限，特征工程即使做的再好，没有合适的参数也往往很难得到一个令人满意的结果。不同的超参数组合可能导致模型在相同任务上有不同的性能，超参数的选择直接影响模型的复杂度。通过调整超参数，可以防止模型过度拟合（对训练数据过于复杂）或过度简化（欠拟合）。超参数的合适选择可以提高模型在未见过的数据上的泛化能力。一个在训练集上表现良好但在测试集上表现差的模型可能是由于过拟合引起的。通过调参，可以找到最优的超参数，使模型在训练数据和新数据上表现更好。调参的方法有很多，可以根据经验手动调参，也可以进行自动调参[29]，常见的自动调参的方法就包括网格搜索、随机搜索、贝叶斯优化学习曲线分析[30]以及启发式搜索等，本实验使用网格搜索、随机搜索以及贝叶斯优化三种方式进行超参调节。

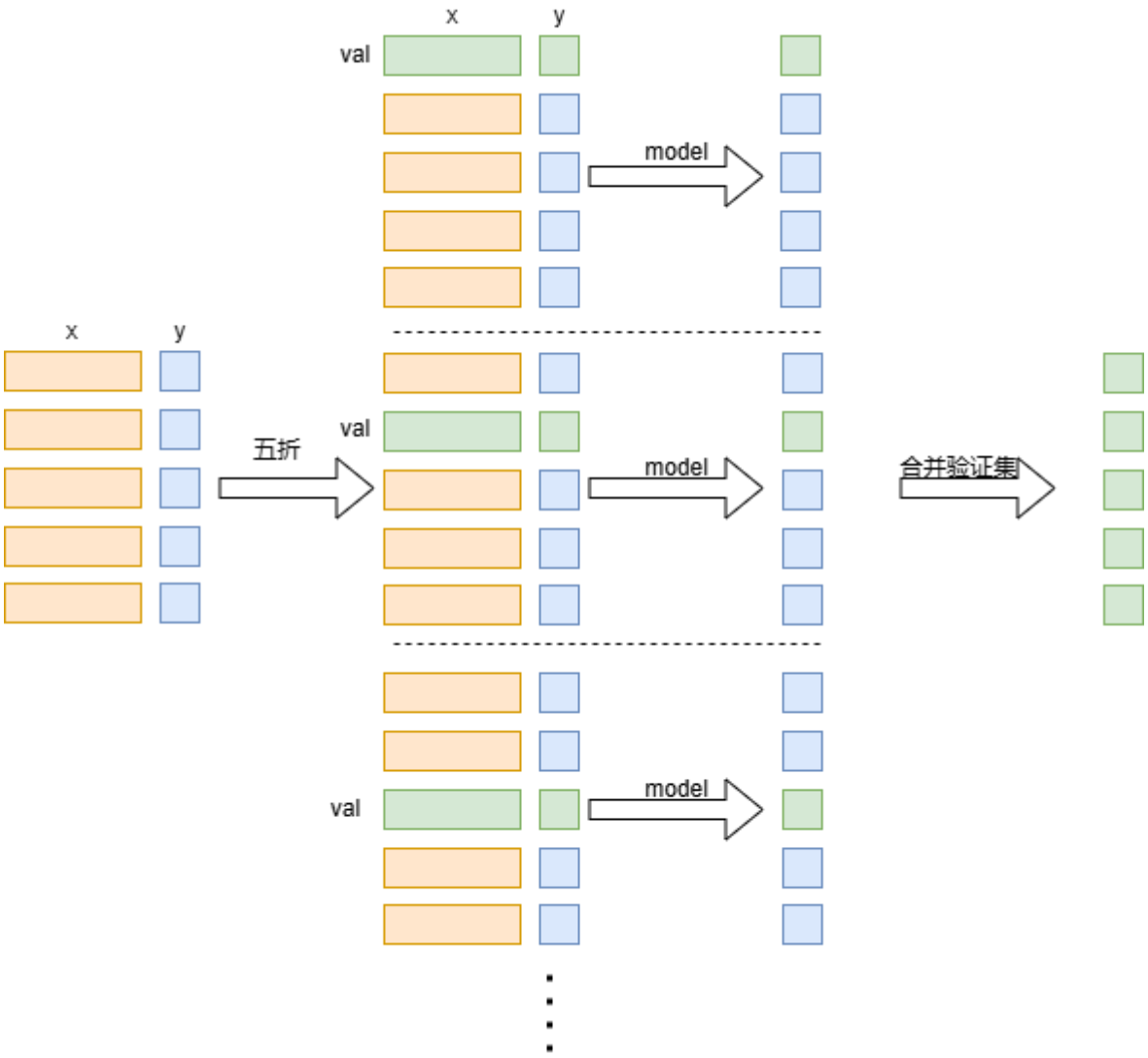

图 4.6 五折交叉

每一个模型采用上述五折交叉的方式进行训练和预测后分别得到随机森林、XGBoost 对于的结果后将 4 组预测结果进行合并，得到一组 n 行 4 列的数据集，其中每一列分别代表各模型的预测结果。得到该数据集后再利用元学习器进行训练，在 stacking 的过程中元学习器的任务是将基学习器(一般为强学习器)学习到的内容进行提取合并，因此元学习器一般选择比较简单的算法，比如线性回归、逻辑回归以及决策树等，元学习器的训练集为前面 n 各基学习器的预测结果，标签仍为数据集原始的标签，在这里本文使用线性回归作为元学习器，将数据输入到算法模型中并进行训练，其流程如图 4.7 stacking 过程所示。

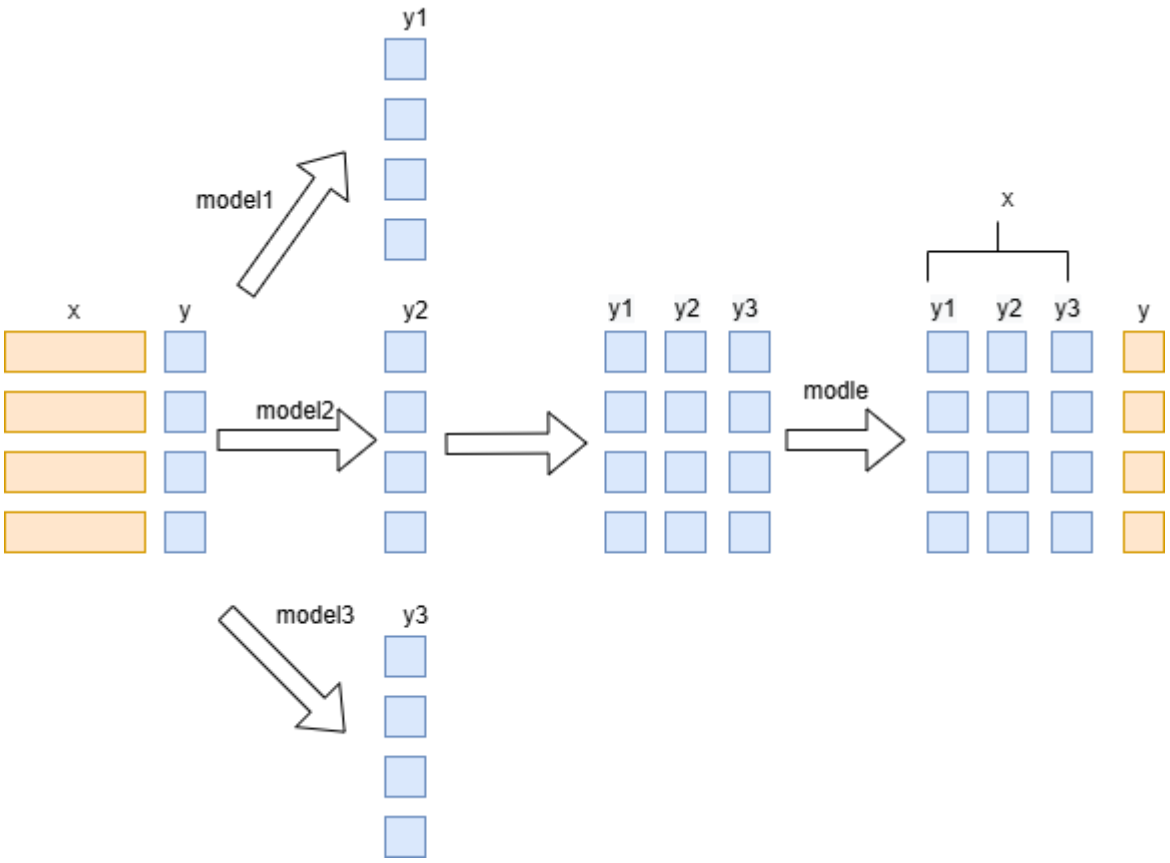

图 4. 7 stacking 过程

4.3.5 模型评估

本问题的目标是预测二手房的交易价格，其一般为连续型数值结果，因此 accuracy(预测准确率)并不适合作为评价指标，这里本文使用 MAE(平均绝对误差)来作为评价指标，其公式如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4.2}$$

其中，y_i为第i个样本的真实值，$\hat{y}_i$为对该样本的预测值，N为样本众数。

为了直观的展示模型的预测结构，取出部分数据绘制除真实值-预测值的对比曲线图，如图 4. 8 预测值—真实值对比曲线所示，可以观察到预测值曲线的整体趋势与真实值相同，只在部分高价格的房屋上预测有较大的偏差。

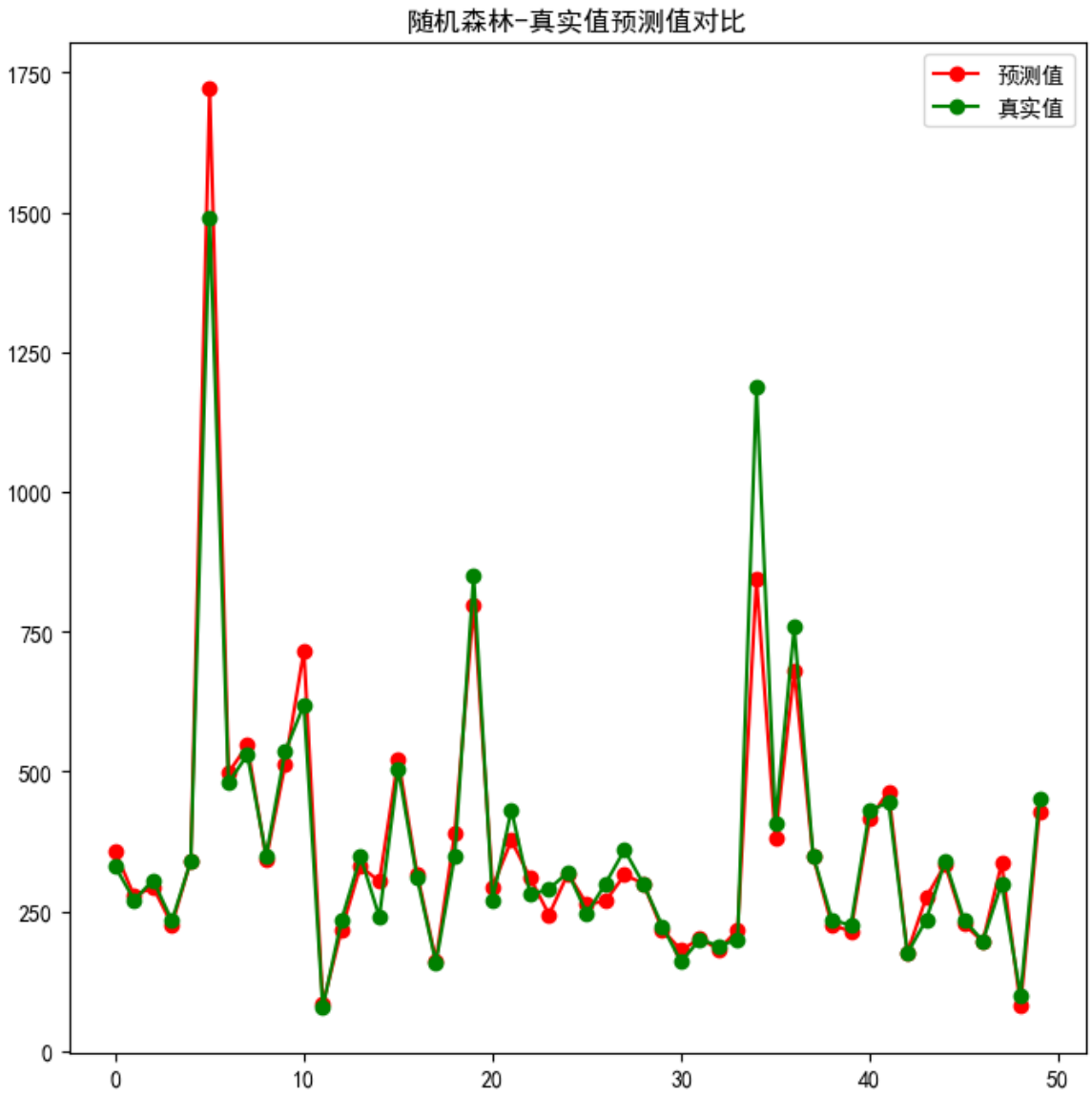

图 4.8 预测值—真实值对比曲线

4.5 系统设计

4.5.1 系统功能模块

本系统主要由登陆验证、用户信息管理、房价信息概览、小区信息查询、历史交易查询以及房价评估验证六大模块组成：

(1) 登陆验证模块:

- 功能: 用户注册、登录、权限控制。
- 关系: 所有其他模块都需要调用此模块来进行用户身份验证。

(2) 用户信息管理模块:

- 功能:用户通过该模块管理自己的信息，包括更新密码等

(3) 房价信息概况模块:

- 功能:展示整个地区的房价概览信息

(4) 小区信息查询模块:

- 功能:查询小区基本信息，包括地址、均价等。

(5) 历史交易查询模块:

- 功能:根据输入信息查询历史交易的详细数据

(6) 房价评估模块:

- 功能:本系统的核心模块、根据用户输入的房屋信息对房价进行评估被给出一个参考价格

系统功能模块的结构如图 4.9 系统模块结构所示，用户访问系统首先进行登录模块的验证，如果用户不授权登录就只能够访问房价信息概览等公开模块，如果访问房价评估预测模块将会被拦截回登录验证模块并提示用户进行登录授权。

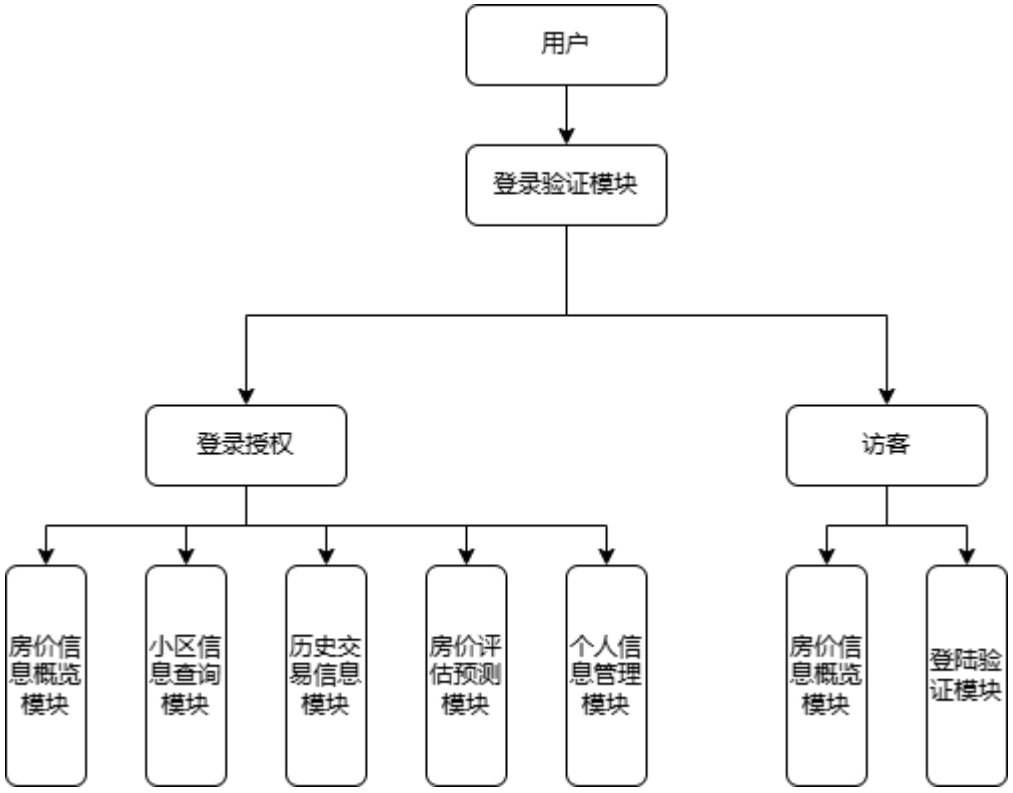

图 4.9 系统模块结构

4.5.2 系统架构设计

系统定位为 web 网页应用，用户界面成使用 Web 前端计算进行开发，在这里选用市面上最流行的前端框架之一 Vue.js；后端使用 flask 框架进行开发，本系统的体量较小，没有过多的功能，而 flask 作为一个轻量级的后端框架，语法简单，可以很好的胜任本系统的后端；数据存储服务使用 Mysql 进行，MySQL 是甲骨文旗下的开源数据库，其具有高性能、高可靠性以及简单易用；安全与认证则使用 JWT，JSON Web Tokens (JWT) 是一种开放标准，它定义了一种紧凑且自包含的方式，用于在各方之间安全地传输信息作为 JSON 对象。

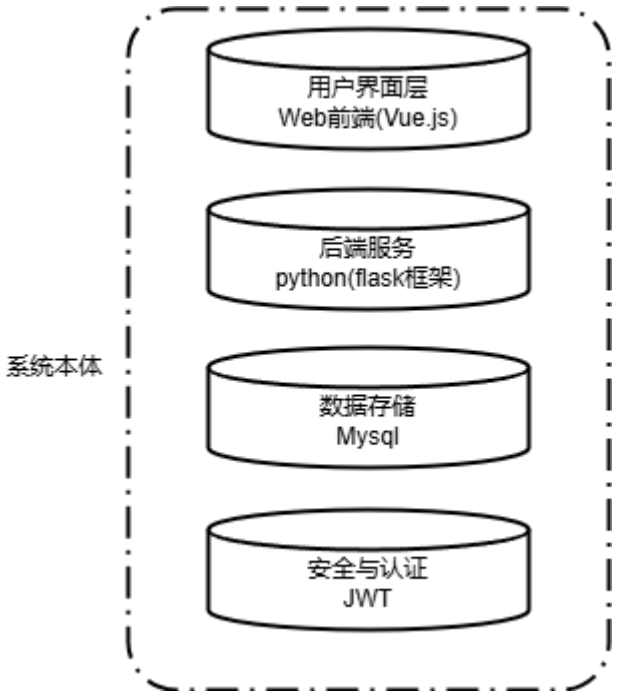

图 4.10 系统架构

(1) 用户界面层：web 前端(Vue.js)

- 用户登录注册
- 个人信息管理
- 房价信息概览
- 小区信息查询
- 历史交易数据
- 房价评估预测

(2) 后端服务：python(flask)

- 用户登录、注册、权限控制
- 读取并分析房屋交易数据

- 进行房价的评估预测

(3) 数据库:MySQL

- 用户信息表
- 小区信息表
- 房屋交易数据表

(4) 安全与认证：JWT

- 用户登录时生成 JWT
- 用户请求时验证 JWT
- 过期时间管理

4.5.3 接口设计

表 4 - 2 列出了系统所涉及的所有接口，其中包括接口的方法、接口路径，并简略描述了该接口的内容以及所负责的职能。表 4 - 2 接口列表

表 4 - 2 接口列表

方法	路径	描述
GET	/info/area	小区信息
GET	/info/area/areasec**tion**	小区房屋面积分布绘图数据
GET	/info/area/decorate	小区房屋装修情况绘图数据
GET	/info/area/floor	小区房屋楼层信息绘图数据
GET	/info/area/price	小区房屋单价信息绘图数据
GET	/info/area/history	历史房屋交易数据
GET	/index/register	用户注册
GET	/index/login	用户登录
GET	/index/avatar	用户信息更新
GET	/index/predict	房价预测

上述表格列出了系统全部接口的粗略信息，下面对其中比较重要的几个接口做详细的说明:

- 用户注册:

  方法:GET

路径:/index/register

描述:用户提交用户名和密码进行注册

请求参数:

Username:（string:required）

Password：（string:required）

响应:

成功:{result:succeed}

失败:{result:fail}

- 用户登录:

方法:GET

路径:/index/login

描述:用户提交用户名和密码进行登录

请求参数:

Username:（string:required）

Password：（string:required）

响应:

成功:{result:succeed，token:JKNSKDASNJDKN…,avatar:url}

失败:{result:fail}

4.5.4 数据库设计

本系统的核心功能为数据预测，故不涉及太多的数据读取，因此本系统的数据库仅包含 3 张表，分别为小区信息表，主键为 AreaId,交易记录表，主键为 Id，外键为 AreaId，小区信息表与交易记录表为一对多的对应关系，用户信息表，其主键为 UserId，数据库表结构设计图如图 4. 11 数据库表结构设计图所示。

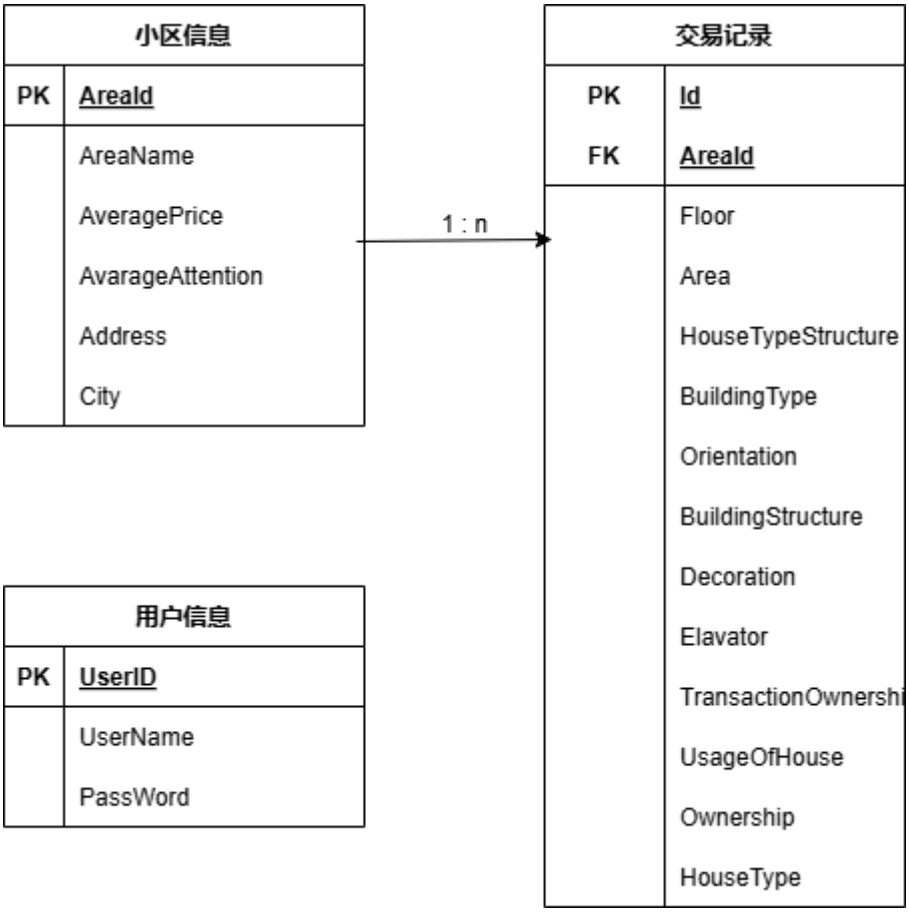

图 4.11 数据库表结构设计图

小区信息表:

AreaId：小区 ID，字符串，主键，自增

AreaName：小区名，字符串，不为空

AveragePrice：均价，整型，不为空

AverageAttention：平均关注度，整型，不为空

Address：地址，字符串，不为空

City：城市，字符串，不为空

用户信息表:

UserID:用户 ID，字符串，逐渐，自增

UserName:用户名，字符串，不为空

PassWord:密码，字符串，不为空

交易记录表:

ID:交易 ID，字符串，主键，自增

AreaId:小区 ID，字符串，外键

Floor:楼层，整型，不为空

Area:面积，浮点型，不为空

HouseTypeStructure:：房屋结构，字符串，不为空

BuildingType:建筑类型，字符串，不为空

Orientation:朝向，字符串，不为空

BuildingStructure:建筑结构，字符串，不为空

Decoration:装修情况，字符串，不为空

Elavator:电梯配备，字符串，不为空

TransactionOwnership:交易权属，字符串，不为空

UsageOfHouse:房屋用途，字符串，不为空

Ownership:权属，字符串，不为空

HouseType:户型，字符串，不为空

第 5 章 系统实现

5.1 前端实现

为了节省时间成本并保障系统能都在多端运行，本系统使用 Web 前端技术进行用户界面层的实现，主要技术栈为 JavaScript、css 以及 html5 经典前端三件套，为里简化页面开发并提高代码的复用性，使用 Vue.js 框架进行开发。Vue 为响应式开发，数据为双向绑定，修改数据后对应的元素会自动刷新，用户不需要再去关注 DOM 操作，从而可以将更多的时间花费在业务逻辑的实现上。得益于 Vue 的虚拟 DOM 机制，HTML 的一个元素（如 div）需要响应数据更改时，会刷新整个页面，导致效率堪忧，但对于虚拟 DOM，浏览器会将 HTML 文件转换为 JS 文件并复制一个额外使用（虚拟）。对于任何更改，虚拟 DOM 都将复制的 JS 与原始 JS 进行比较，只重新加载更改的部分，局部修改到真实 DOM 上，从而使页面效率大大提升。Vue 为组件化开发，将使用频率高的代码提炼为组件进行复用可以大大的提高开发效率。

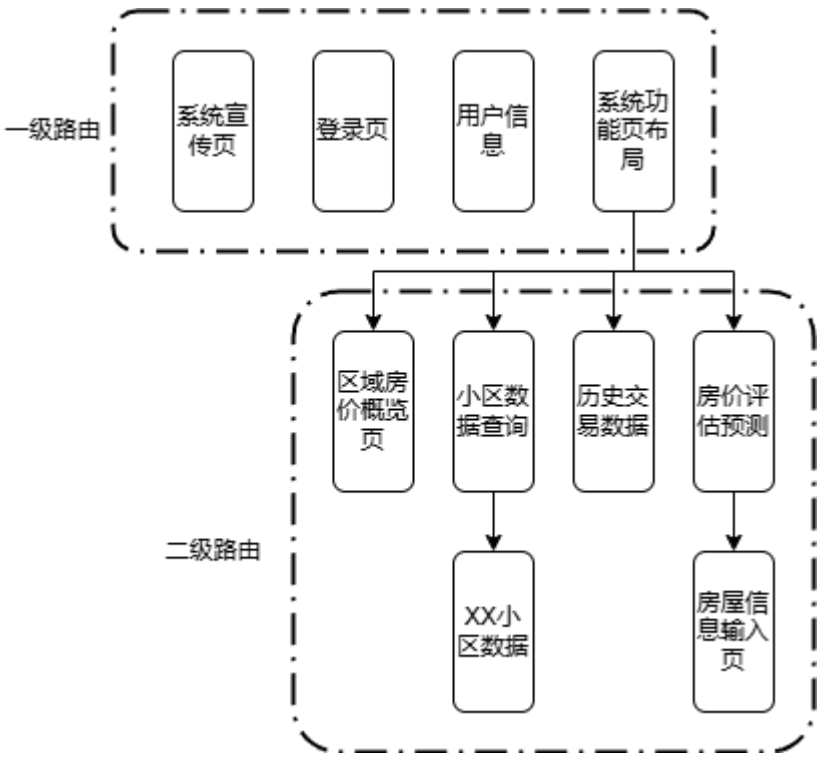

图 5.1 前端页面结构图

图 5.1 前端页面结构图给出了本系统的前端页面结构，其中系统宣传页、登录页、用

户信息、系统功能页布局为一级路由，其中系统功能页下包含区域房价概览页、小区数据查询、历史交易数据、房价评估预测、XX 小区数据以及房屋信息输入页等二级路由，其中小区数据查询页关联至 XX 小区数据，房价评估预测页关联至房屋信息输入页。当用户为进行登录注册时，会限制用户的访问，用户仅可浏览区域房价概览页面，如强制跳转将会被拦截回到登录页并提示登录；若用户进行登录授权后即可正常访问其它页面。

```
// 需要有权限的页面
const privtePage = ['/index/areaprice', '/index/areahistory', '/index/predict']
// 配置全局导航守卫
// (1)next()  直接放行
// (2)next(路径)  进行拦截_
router.beforeEach((to, from, next) => {
  // 非隐私页面直接可访问
  if (!privtePage.includes(to.path)) {
    next()
    return
  }
  // 隐私页面进行权限验证
  const token = store.state.user.userInfo.token
  if (token) {
    next()
  } else {
    next('/login')
  }
})
```

上述代码对前端访问权限进行限制，如果用户访问的非隐私界面则直接允许跳转，如果是隐私界面则授权进行登录验证，如果已登录则正常跳转，如果未登录则将用户拦截至登录页并提示登录。

5.2 后端实现

本系统的后端使用 python 进行开发，并使用 flask 作为后端开发框架，Flask 是一个使用 Python 编写的轻量级 Web 应用框架，它简洁而灵活，适用于开发小型至中型的 Web 应用。本文将介绍 Flask 框架的基本概念、特点以及如何使用 Flask 来快速搭建 Web 应用。Flask 是基于 Werkzeug 和 Jinja2 库构建的，它遵循了 MVC（模型-视图-控制器）的设计模式。Flask 的核心思想是保持简洁和易用，它提供了一些核心功能，但也允许开发者通过扩展来添加更多功能。

Flask 是一个简单而灵活的 Python Web 框架，适用于快速开发小型至中型的 Web 应用。它的简洁的 API 设计和丰富的扩展库使得开发变得简单、灵活和高效。它鼓励开发者们尝试使用 Flask 来构建自己的 Web 应用，体验其简单易用和强大的功能。

表 5 - 1 为基于 flask 框架为本系统开发的所有接口及其相关信息，包含接口的方法，接口路径以及接口负责何种功能。

表 5 - 1 接口列表

方法	路径	描述
GET	/info/area	小区信息
GET	/info/area/areasec**tion**	小区房屋面积分布绘图数据
GET	/info/area/decorate	小区房屋装修情况绘图数据
GET	/info/area/floor	小区房屋楼层信息绘图数据
GET	/info/area/price	小区房屋单价信息绘图数据
GET	/info/area/history	历史房屋交易数据
GET	/index/register	用户注册
GET	/index/login	用户登录
GET	/index/avatar	用户信息更新
GET	/index/predict	房价预测

```
data['房屋朝向'] = data.apply(lambda x: chaoxiang(x), axis=1)
one_hot_col_names = ['小区名称', '房屋朝向']
data_onehot = pd.get_dummies(data[one_hot_col_names])
data = pd.concat([data, data_onehot], axis=1)
```

```
# 按小区和城市连接平局关注度和单价
average_attention_area = pd.read_csv('../data/deal/average_attention_area.csv', index_col=0)
result = pd.merge(data, average_attention_area, on='小区名称', how='left')
average_price_area = pd.read_csv('../data/deal/average_price_area.csv', index_col=0)
result = pd.merge(result, average_price_area, on='小区名称', how='left')
average_price_city = pd.read_csv('../data/deal/average_price_city.csv', index_col=0)
result = pd.merge(result, average_price_city, on='城市', how='left')

# 删除不必要字段
result.drop(columns=['小区名称','房屋朝向'],inplace=True)

# 提取出待预测数据
waitPredict = result.tail(1)

# 加载模型
model = pickle.load(open('../data/model/randomForest.pkl', 'rb'))
predict_result = model.predict(waitPredict)
price = predict_result[0]
print(price)

responses = {
    'data': [{'result': price}]
}

return jsonify(responses)
```

上述给出了房价评估接口的关键代码部分，后端获取到用户输入的房屋信息后将其处理为所属格式后进行数据处理，将数据处理为预测模型所接受的格式，后端读出训练好的预测模型并进行预测，最后将预测的结构以 JSON 数据的形式通过接口发送给前端。

第 6 章 结语

著名经济学家 George Akerlof 曾于 1970 年发表的文章《柠檬市场:质量的不确定性和市场机制》中以二手车市场为例,指出了在市场交易中,信息的不对称会引发社会信任赤字,进一步导致劣质品充斥市场,优质品被逐出市场,从而制约二手车市场的发展。目前我国二手房的相关政策持续向好且二手房市场规模庞大,可以说二手房行业正迎来又一个发展风口。然而据调查,当前我国二手房市场对于二手房的定价机制并不完善,以普通线下二手房交易市场为例,几乎所有的二手房都是一口价,且这个价格完全由销售方一口咬定,外行人一般很难弄清楚二手房的具体行情,层出不穷的“黑心定价”新闻导致买卖双方产生信任危机。由此,利用二手房交易数据,制定一套科学合理的二手房价格预测体系就显得尤为必要。

对于二手房销售平台方而言，本研究提供了一套二手房价格预测系统的建立理论并进行实现，二手房销售平台方可以依据这些在本研究模型中显示出重要性的特征建立一套新的二手房价体系,透明化二手房定价机制,而不是如黑箱般直接给定出售价格。这不仅可以避免买卖双方因为信息不对称而导致的信任赤字,更有助于规范二手房交易市场的价格乱象,从而促进国内二手房市场的平稳发展。

本文基于数据集进行特征构建并筛选，利用随机森林、XGBoost 以及线性回归等模型进行训练并使用 stacking 的方式对模型进行融合集成并使用 MAE 作为评价指标，提供了一套相对完整的二手房交易价格预测的实验方案。本文所使用到的调参方式为贝叶斯优化，其可以在有限的时间和算力内得到较好的参数组合，若不考虑时间以及拥有充足的算力资源，可以考虑使用网格调参进行暴力搜索以获得最好的参数组合。

参考文献

[1] 王广中;.“强政府”印象与调控失灵:中国房地产业治理的悖论[J].理论与改革,2013(02).

[2] 瞿成元;.房地产市场的需求价格怪圈——基于现代经济学的解释[J].经营与管理,2017(07).

[3] 杨剑锋;乔佩蕊;李永梅;王宁;.机器学习分类问题及算法研究综述[J].统计与决策,2019(06).

[4] 张梦;施同兵;.基于 VAR 模型对房地产价格影响因素的实证研究[J].中国集体经济,2020(05).

[5] 韦金洪;刘佳;.货币政策变动对房地产价格的动态影响——基于 VAR 模型实证分析[J].商,2012(19).

[6] 陈将浩.房价影响因素及 R 语言实现[D].中国科学技术大学,2014(10).

[7] 汪雅倩;陈依萍;.我国房价的影响因素分析及政策建议[J].科技经济导刊,2016(17).

[8] 徐建炜;徐奇渊;何帆;.房价上涨背后的人口结构因素:国际经验与中国证据[J].世界经济,2012(01).

[9] 吴齐林;.关于房价飙升的理论思考[J].江苏广播电视大学学报,2007(04).

[10] 汪慧颖;.武汉市二手住房价格浅析与预计[J].消费导刊,2007(06).

[11] 谭刚著;.房地产周期波动[M].经济管理出版社,2001.

[12] 刘冰;金跃强;王书营;.南京市二手房房价影响因素的多元线性回归分析[J].南京工业职业技术学院学报,2017(01).

[13] 邝文竹;刘琳;.南宁市二手房价格影响因素分析及房价走势的预测[J].广西科学院学报,2012(02).

[14] Apergins N. Housing prices and macroeconomic factors: prospects within the European Monetary Union[J].International Real Estate Review, 2003, 6(1): 63-74.

[15] Elbourne A. The UK housing market and the monetary policy transmission mechanism: An SVAR approach[J].Journal of Housing Economics, 2008, 17: 65-87.

[16] Abraham B, Hendershott W. Bubblesinmetro Politanmarkets [J]. Journal of Housing Research, 2004, 12:36-52.

[17] Kau J B, Keenan D C, Muller III W J, et al. The valuation at origination of fixed-rate

mortgages with default and prepayment[J]. The Journal of Real Estate Finance and Economics, 1995, 11(1): 5-36.

[18] Brueggeman W B, Chen A H, Thibodeau T G. Real Estate Investment Funds: Performance and Portfolio Considerations, AREUEA Journal[J]. 1984, 12(3): 333-354.

[19] Miller N G, Sklarz M A, Real N O. Japanese purchases, exchange rates and speculation in residential real estate markets[J]. Journal of Real Estate Research, 1988, 3(3): 39-49.

[20] Bitter C, Mulligan G F, Sandy Dall'erba. Incorporating spatial variation in housing attribute prices:a comparison of geographically weighted regression and the spatial expansion method[J]. Journal of Geographical Systems, 2007, 9(1): 7-27.

[21] Hendershott P. Equilibrium Models in Real Estate Research: A Survey[J].Joural of Real Estate Literature, 1998, 6(1): 13-25.

[22] Thibodeau T G. Marking Single-Family Property Values to Market[J].Real Estate Economics, 1998, 44(2): 171-196.

[23] Oates W E. The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the tiebout Hypothesis[J].Journal of Political Economy, 1969, 77(6): 957-971.

[24] Shim J, Bin O, Hwang C. Semiparametric spatial ef fects kernel minimum squared error model for predicting housing sales prices[J]. Neurocomputing, 2014, 124(2): 81-88.

[25] Gao J, Zhou L P. Study on housing hedonic price theory based on Box-Cox transformation[J]. Journal of Hebei University of Science and Technology, 2007, 72(3): 33-45.

[26] Kim H G, Hung K C, Park S Y. Determinants of housing prices in Hong Kong: a Box-Cox quantile regression approach[J]. Journal of Real Estate Finance & Economics, 2015, 50(2): 270-287.

[27] Chen T , Guestrin C .XGBoost: A Scalable Tree Boosting System[J].ACM, 2016.DOI:10.1145/2939672.2939785.

[28] 赵艳伟,胡正祥,乔登攀,等.一种基于贝叶斯优化和 XGBoost 的膏体流变参数预测模型[J].有色金属(矿山部分),2024,76(05):118-128.

[29] 徐新越,蒋毅.一类最小二乘的自动调参问题的求解算法[J].四川师范大学学报(自然科学版),2024,47(06):812-817.

[30] 赵艳伟,胡正祥,乔登攀,等.一种基于贝叶斯优化和 XGBoost 的膏体流变参数预测模型

[J].有色金属(矿山部分),2024,76(05):118-128.