分 类 号：TP391
密 级：公 开
单位代码：10878
学 号：20193305036

安徽建筑大学
ANHUI JIANZHU UNIVERSITY

硕士专业学位论文

（全日制）

论 文 题 目：使用多分类器的房产价格预测方法
专业学位类别：工 程
专业学位领域：计算机技术
研 究 方 向：数据挖掘
作 者 姓 名：李西洋
导 师 姓 名：史东辉
完 成 时 间：2022 年 4 月

使用多分类器的房产价格预测方法

A method of real estate price prediction using multiple classifiers

专业学位类别：工 程

专业学位领域：计算机技术

研 究 方 向：数据挖掘

作 者 姓 名：李西洋

导 师 姓 名：史东辉

完 成 时 间：2022 年 4 月

安徽建筑大学

本论文经答辩委员会全体委员审查，确认符合安徽建筑大学硕士学位论文质量要求。

答辩委员会签名

主席：

委员：

导师：

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得安徽建筑大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：李雨萍 签字日期：2022年5月22日

导师签名：史东辉 签字日期：2022年5月22日

学位论文版权使用授权书

本学位论文作者完全了解 安徽建筑大学 有保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于 安徽建筑大学 。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 安徽建筑大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。（保密的学位论文在解密后适用本授权书）

学位论文作者签名：李雨萍 签字日期：2022年5月22日

导师签名：史东辉 签字日期：2022年5月22日

摘　要

近两年来，国家提倡房住不炒，各大银行纷纷上调房贷利率，房价上涨势头有所缓解。在此背景下，房地产市场的未来走势越来越成为全民关注的焦点。对于购房刚需族，如何高效准确地掌握各类商品房的真实价格成为了困扰他们的头号难题。本文在分析了实验所用房产数据的真实性和合理性之后，着手对数据样本进行了一定的分类研究，并从实用性的角度出发，对商品房价格进行了一定的预测和分析，从而反映出不同价位商品房的价格走势，对于普通购房者具备一定的参考意义。

有关房价的预测分析，国内外也有不少学者进行了相关的研究，不过基本都是选择传统神经网络或者机器学习方法，结合部分房屋价格影响因素直接构建房价预测模型，对于不同价值的房屋进行预测时可能会出现较大误差。本文将多种分类器与回归算法结合应用于房屋价格预测这一热点民生问题上，大幅度提高了当前房价预测方法的精确度。同时为了更加准确地把握数据规律，本文对数据进行了特征重要性分析，去除了数据中对房价影响较小的因素及无关属性，达到数据降维的目的，有效避免了过拟合现象。

本次研究所使用的房产数据集来源于 Kaggle 网站，数据信息公开透明，真实可靠。实验过程中，先使用贝叶斯神经网络， 梯度提升决策树，K 近邻和逻辑回归四种分类器对数据进行分类，将不同销售记录的商品房分为低、中、高三种价值。根据分类结果，对四种分类器的分类性能进行初步比较，最终发现贝叶斯神经网络分类器处理实验数据时的分类性能最优。针对不同分类器分出的房产数据，再依次使用多元线性回归、决策树、随机森林、长短期神经网络以及门控循环单元五种模型进行房价预测。根据预测结果，分别计算出五种模型依次处理不同分类器分出的三类数据集时所得评价指标的相应加权平均值，并将其作为五种回归模型处理不同分类器分出的数据集时的评价指标。最后，将各种模型依次处理不同分类器分出的数据集时所得评价指标分别与该种模型处理未分类数据集时的评价指标对比，发现各种模型处理任意一种分类器分类后的数据集时效果都比该种模型处理未分类数据集的效果更好。同时，通过将同种模型依次处理不同分类器分出的数据集时所得评价指标进行两两对比，从而得出各种模型的最佳预测效果。最后通过对比五种回归模型各自取得最佳预测效果时的各项评价指标，发现贝叶斯神经网络和随机森林的组合模型预测效果最佳，三种误差(MAE、RMSE、MAPE)分别为 1408.33 元、2138.48 元、8.46%，相比较直接使用多元线性回归模型(处理

未分类数据集时效果最佳)进行预测，实验误差 MAE 减少了 669.75 元，RMSE 减少了 567.05 元，MAPE 减少了 6.25%，预测效果显著提升。

图[19]表[14]参[60]
关键词：多分类器；房产价格；预测
分类号：TP391

Abstract

In recent two years, the central government has advocated that houses are for living in and not for speculative investment. In addition, major banks have raised mortgage interest rates one after another, and the rising trend of house price has eased. In this context, people are increasingly concerned about the future trend of the real estate market. For those people who really need houses to live in, it has become the most important issue that how to effectively and accurately grasp the real price of all kinds of commercial houses in the city. After analyzing the authenticity and rationality of the real estate data used in the experiment, this paper starts to study the classification of house data sets and then predicts the house price for the house sample data. By analyzing the prediction results, we can grasp the real prices of different types of commercial houses, which has some reference for ordinary buyers.

In the field of house price prediction and analysis, domestic and foreign scholars have done some related research, but they usually choose the traditional neural network or machine learning methods and directly to build house price prediction models in combination with some influencing factors. There may be large errors in the prediction of houses with different values. In this study, the combination of a variety of classifiers and regression models is proposed to predict house prices, which greatly improves prediction results. Before classifying real estate data, this study analyzes the characteristic importance for the real estate data set, removes irrelevant attributes that have little impact on house prices, achieves the goal of data dimensionality reduction, and avoids the phenomenon of over fitting effectively.

The real estate data set used in this study is obtained from the website www.kaggle.com. The data set can be accessed freely and is transparent, true and reliable. During the experiment, four classifiers: Bayesian Neural Network(BNN), Gradient Boosting Decision Tree(GBDT), K-Nearest Neighbors(KNN) and Logistic Regression(LR), were used to classify the real estate data. A commercial house in a sale record is classified to low value, medium value and high value. According to the classification results, the classification performances of the four classifiers was preliminarily compared. Finally, it was found that BNN classifier has the best classification performance among the four classifiers. Combined the real estate data sets with different value types, which were separated by different classifiers, Multivariable Linear Regression(MLR), Decision Tree(DT), Random Forest(RF), Long Short-Term Memory(LSTM) and Gated Recurrent

Unit(GRU) are used to build prediction model to predict house prices in turn. We calculated the weighted average values for MAE, RMSE and MAPE for the five regression models respectively, which are used on the three data sets classified by the each classifier. Finally, the evaluation indexes, the weighted average values are compared with those for the five regression models which are used for the unclassified data set in turn. It is found that the effect of each model dealing with the data sets classified by each classifier is better than that of these models dealing with the unclassified data set. At the same time, by comparing the evaluation indexes for the same model processes for the data sets separated by different classifiers in turn, we can get the optimal evaluation indexes for various models.

Finally, by comparing the optimal evaluation indexes of various models, it is found that the combined prediction model of BNN and RF has the best prediction result, and the three values (MAE, RMSE and MAPE) are 1408.33 yuan, 2138.48 yuan and 8.46% respectively. Compared with the direct use of MLR model (obtain the best prediction result when dealing with the unclassified data set), the experimental error MAE is reduced by 669.75 yuan, RMSE is reduced by 567.05 yuan, MAPE is reduced by 6.25%, and the prediction result is significantly improved.

Figure[19] Table[14] Reference[60]

KeyWords： multiple classifiers，real estate price，prediction

Chinese books catalog: TP391

目录

摘　要.....I
Abstract.....III
插图清单.....XI
表格清单.....XII
第一章 绪论.....1
1.1 研究背景及意义.....1
1.1.1 研究背景.....1
1.1.2 研究意义.....2
1.2 国内外研究现状.....3
1.2.1 国内研究现状.....3
1.2.2 国外研究现状.....4
1.2.3 缺点分析.....5
1.3 主要内容.....6
1.4 论文结构.....7
1.5 本章小结.....8
第二章 总体方案设计及方法介绍.....9
2.1 总体方案.....9
2.2 分类器.....11
2.2.1 贝叶斯神经网络.....11
2.2.2 梯度提升决策树.....12
2.2.3 K 近邻.....12
2.2.4 逻辑回归.....12
2.3 回归算法.....13
2.3.1 多元线性回归.....13
2.3.2 决策树.....13
2.3.3 随机森林.....14
2.3.4 长短期神经网络.....14

2.3.5 门控循环单元 15
2.4 本章小结 16
第三章 数据处理及评价指标选择 17
3.1 数据预处理 17
3.1.1 数据获取 17
3.1.2 数据清洗 17
3.1.3 数据初步分析 18
3.2 数据降维 20
3.3 评价指标选择 22
3.3.1 分类评价指标 22
3.3.2 预测评价指标 23
3.4 本章小结 24
第四章 房产价格预测实现 25
4.1 实验环境 25
4.2 实验步骤 25
4.3 实验过程 27
4.3.1 基于不同分类器的分类实验 27
4.3.2 基于不同分类数据集的预测实验 28
4.3.3 基于未分类数据集的预测实验 31
4.4 本章小结 32
第五章 结果对比分析 33
5.1 分类实验结果对比 33
5.2 预测实验结果对比 34
5.2.1 各种回归模型处理未分类数据集的实验结果对比 34
5.2.2 各种回归模型处理 BNN 分出的三种数据集的实验结果对比 35
5.2.3 各种回归模型处理不同分类器分出数据集的实验结果对比 36
5.2.4 各种回归模型的最佳实验结果对比 40
5.3 本章小结 41
第六章 总结与展望 43

6.1 总结 43
6.2 展望 44
参考文献 45
致谢 50
作者简介及读研期间主要科研成果 51

Contents

Abstract III

Illustration list XI

Form list XII

Chapter 1 Introduction 1

1.1 Research background and research purpose 1

1.1.1 Research background 1

1.1.2 Research purpose 2

1.2 Domestic and foreign research 3

1.2.1 Domestic research 3

1.2.2 Foreign research 4

1.2.3 Defect analysis 5

1.3 Main content 6

1.4 Paper structure 7

1.5 Chapter summary 8

Chapter 2 Overall Scheme Design And Methods Introduction 9

2.1 Overall scheme 9

2.2 Classifiers 11

2.2.1 Bayesian neural network 11

2.2.2 Gradient boosting decision tree 12

2.2.3 K-nearest neighbor 12

2.2.4 Logistic regression 12

2.3 Regression algorithms 13

2.3.1 Multivariable linear regression 13

2.3.2 Decision tree 13

2.3.3 Random forest 14

2.3.4 Long short-term memory 14

2.3.5 Gated recurrent unit 15

2.4 Chapter summary 16
Chapter 3 Data Processing and Selection of Evaluation Indexes 17
3.1 Data processing 17
3.1.1 Data acquisition 17
3.1.2 Data cleaning 17
3.1.3 Preliminary data analysis 18
3.2 Data dimensionality reduction 20
3.3 Selection of Evaluation Indexes 22
3.3.1 Evaluation indexes of classification 22
3.3.2 Evaluation indexes of prediction 23
3.4 Chapter summary 24
Chapter 4 Realization of real estate price prediction 25
4.1 Experimental environment 25
4.2 Experimental procedures 25
4.3 Experimental process 27
4.3.1 Classification experiments based on real estate data set 27
4.3.2 Prediction experiments based on different classified data sets 28
4.3.3 Prediction experiments based on unclassified data set 31
4.4 Chapter summary 32
Chapter 5 Comparation and Analysis of Results 33
5.1 Comparison of classification experiment results 33
5.2 Comparison of prediction experiment results 34
5.2.1 Comparation and analysis of experimental results of various models dealing with unclassified data set 34
5.2.2 Comparation and analysis of experimental results of various models dealing with three value data sets based on BNN 35
5.2.3 Comparation and analysis of experimental results of various models dealing with different data sets 36

5.2.4 Comparison of best experimental results of various models dealing with data sets based on different classifiers 40
5.3 Chapter summary 41
Chapter 6 Summary And Outlook 43
6.1 Work summary 43
6.2 Outlook 44
References 45
Acknowledgement 50
About the author and the research results in the research process 51

插图清单

图 2-1 实验整体架构 10
图 2-2 传统神经网络与贝叶斯神经网络结构对比图 11
图 2-3 随机森林分类原理图 14
图 2-4 LSTM 结构图 15
图 2-5 GRU 结构图 16
图 3-1 数据清洗原理 18
图 3-2 不同面积的房屋成交量占比 19
图 3-3 不同楼层的房屋成交情况 20
图 3-4 房产数据各特征重要性 21
图 4-1 实验步骤图 25
图 4-2 贝叶斯神经网络分类散点图 27
图 4-3 五种模型对未分类数据集的预测效果图 32
图 5-1 四种分类器的分类性能对比图 34
图 5-2 五种回归模型处理未分类数据集的各项评价指标 35
图 5-3 五种模型处理各种数据集的 MAE 38
图 5-4 五种模型处理各种数据集的 RMSE 38
图 5-5 五种模型处理不同数据集的 MAPE(%) 39
图 5-6 五种回归模型各项评价指标减少值 40
图 5-7 五种模型预测效果最佳时的各项评价指标对比图 41

表格清单

表 3.1 部分房产数据特征的具体信息 19
表 3.2 多分类类别 22
表 4.1 多元线性回归模型参数 29
表 4.2 决策树模型参数 29
表 4.3 随机森林模型参数表 30
表 4.4 长短期神经网络模型参数表 30
表 4.5 门控循环单元模型参数表 31
表 5.1 四种分类器对各类别的预测概率 33
表 5.2 四种分类器分类性能评价指标 33
表 5.3 基于未分类数据集的预测实验各项评价指标 35
表 5.4 五种回归模型处理贝叶斯神经网络分出的三种数据集时的各项评价指标 36
表 5.5 五种模型依次处理四种分类器分出数据集时的各项评价指标 37
表 5.6 五种模型各项评价指标减少值 39
表 5.7 五种回归模型各自预测效果最佳时的各项评价指标 40

第一章 绪论

1.1 研究背景及意义

1.1.1 研究背景

住房是人类生存之本，房地产是与人民生活密切相关的产业。从宏观经济角度来看，房地产经济是不可或缺的一环，它能拉动内需市场，带动相关产业发展，促进中国经济内循环，为我国经济持续增长提供了强劲动力。

过去二十年，中国楼市暴涨，具体原因[1]可以分为以下三个方面：一是由于政府的有力推动，比如中央在 2008 年次贷危机之后不仅将大量资金注入楼市，同时采取了包括放宽信贷在内的一系列挽救楼市措施；又比如部分城市由于没有核心产业，地方政府的财政收入严重依赖土地拍卖，间接推动了房产价格的上涨。二是由于国人的购房意愿一直很强烈，经济良好运行时，百姓有钱买房，经济下行时，百姓选择买房规避风险。在国家城市化加速进程中，民间流动性资金的不断涌入维持了楼市基本面的持续向好。三是国家对于住宅保障监管制度并不完善，在各地炒房客大肆囤积城市核心地段的房源和众多第三方房产中介公司哄抬房价的共同作用下，房产的金融属性被过度放大，破坏了市场买卖平衡，加剧了房价的飙升。

过去两年，伴随着全球新冠肺炎疫情肆虐[2]，国际贸易严重受阻[3]，我国外贸市场遭受较大冲击。一方面，为了实现充分就业，政府通过扩张性财政政策刺激消费和投资，拉动内需[4]，加速推动国民经济高速增长到高质量发展的转型；另一方面，为了避免民间资本涌入楼市导致处在高位的房价过热，中央多次重申全面落实“稳地价稳房价稳预期”的长效管理调控机制，不仅发布了一轮又一轮严格的楼市调控政策，还多次出手干预部分地区的房贷利率。随着限购政策的收紧，不少房企纷纷遭遇了资金链危机，其中不乏诸如恒大这样的龙头企业。它们中的多数只能通过低价抛售楼盘来快速回笼资金，这种做法直接拉低了当地房价。在此背景下，不少城市通过发布房价“限跌令”来应对房地产市场潜在的连锁崩盘危机，稳住市场对楼市的信心，实现中国经济的软着陆。与此同时，民间对于楼市未来发展态势的看法已经明显从以前的盲目乐观开始走向褒贬不一，有人依旧对楼市充满信心，认为房价在后疫情时代将会迎来一轮大涨，也有人认为当前中国楼市泡沫化严重，人口红利即将消失，房价已经

基本见顶，待房产税政策发布之后，房价必然从高点开始回落，趋于合理区间。

时至今日，住房问题早已不仅仅是民生问题，同样也是经济问题，中国现有行业中与房地产市场直接相关的不下二十个，房贷余额在银行所有贷款余额中比重接近三成，房地产行业已经是推动国民经济增长的重要引擎。楼市的未来走势不仅时刻牵动着无数购房者的心，同样也影响着中国金融业、制造业和其它行业的发展，更关系着国家和社会的稳定。

房价对人们的生活水平和国民经济发展都有着很大的影响，针对房价走势的研究和预测一直是众多国内外专家学者讨论的焦点。本文在分析了实验所用房地产数据的真实性和合理性之后，着手对我国商品房数据进行了一定的实验研究，并从实用性的角度出发，结合多种机器学习和深度学习算法对国内部分城市的商品房价格进行了一定的预测和分析，从而反映出不同价值商品房的真实价格，对于普通购房者以及相关监管部门具备一定的参考意义。

1.1.2 研究意义

房价的走势能直观地反映经济运转的好坏，除了普通购房者，房地产开发商和政府也都密切关注着房价的波动。对于普通百姓而言，贷款买房已成为普遍现象，房贷已经是绝大多数家庭的最大负债之一，众多家庭的消费水平以及生活幸福指数时刻受到房价波动的影响。对于房地产开发商而言，房价的波动更是直接影响着房企的利润高低，左右着众多中小型房企的生存空间。对于政府而言，房地产行业捆绑了包括金融行业在内的众多社会行业[5]，为数以万计的从业者直接或者间接地提供了工作岗位，政府有关部门需要根据楼市最新态势发布相对应的调控政策。可以看出，房价无论是暴涨还是暴跌都会对经济发展造成较大冲击，进而引发一系列问题，甚至会造成社会动荡。通过深入研究房产数据集，能确定对房价影响较大的因素，再通过构建相关算法模型就能对未来房价进行预测。这不仅能帮助人们更真实地了解房产价格走势，也能帮助政府有关部门及时对房地产市场进行价格调控，对稳定金融市场、改善民情民生都有着重要意义。

商品房价格的精准预测影响深远，其意义主要分为以下几个方面：对于潜在购房者而言，他们能够根据房价的走势挑选价格最低时进行买房，力争个人资产收益最大化；对于二手房出售者而言，他们能够根据当前房屋的真实价格进行合理定价，防止由于信息不对称导致要价过高，进而造成不必要的损失；对于房地产商而言，他们可以根据房价的未来走势，提前制定相应的营销策略，

提升房产成交量，从而维护企业的正常运转；对于地方政府而言，它们可以根据房价的变化实时采取针对性强且效益高的举措，让房价趋向合理，让房地产市场的发展更加稳定，并引导银行等金融机构根据政府政策方向做出合理决策，避免出现银行坏账，降低金融风险，让中国经济持续充满活力。

由于同一城市不同地区的房屋价值差距比较大，直接使用机器学习模型进行统一房价预测时，可能会出现较大误差。因此，本文在进行房价预测工作之前，先使用贝叶斯神经网络、梯度提升决策树、K 近邻和逻辑回归四种分类方法依次对房产数据进行价值分类，将商品房分成较低、中等、较高三种价值，再使用不同的回归模型对分类所得到的不同类别房产数据集进行房价预测工作，有效降低了实验误差，从而为不同价位的商品房提供了更准确更真实的参考价格。

综上所述，本次研究在理论联系实际的基础上进行了有价值的探索，采用分类与回归的组合模型开展房屋价格预测研究。不仅解决了当前居民购房决策困难、二手房挂牌价格虚高和调控政策发布不及时等问题，也为商品房市场的长期良性运行提供了一定的参考。

1.2 国内外研究现状

1.2.1 国内研究现状

房价的走势在国内一直都是人们的热点议论话题，国内学者们针对房价的预测进行了大量的探索，使用的研究方法以及选取的实验指标各有不同。由于房价预测是一个回归问题，许多学者选择使用机器学习算法建立房价预测模型。国内学者关于房价预测方面的研究具体如下：

丁飞和江铭炎[6]结合改进后的狮群算法对 BP 神经网络的权值和偏置进行了一定的优化，根据房屋的户型、面积等相关指标对青岛二手房价格进行有效预测。实验结果表明，狮群与 BP 结合模型相较于传统的 BP 神经网络在房价预测问题上处理的效果更好。

申瑞娜，曹昶，樊重俊[7]选取上海房产数据作为实验数据集，将主成分分析与支持向量机相结合构建模型对房价进行预测，结果证明该模型对于小样本数据的预测具有较好的泛化能力和预测精度。但是文章中支持向量机在预测过程中交叉验证的参数没有进行调优，对于房价影响因素的选择，也有待进一步的优化与改进。

高玉明和张仁津[8]使用贵阳市 1998 年至 2011 年间的部分房产数据作为实验数据集，结合改进后的 BP 神经网络对其进行训练并开展房价预测工作。结果对比发现，优化后的 BP 神经网络模型在收敛速度方面和预测精度方面，都好于传统的 BP 神经网络模型。

麻顺顺[9]在郑州市金水区的部分二手房数据作为实验数据集，使用 Haversine 算法对数据进行空间序列排序，基于 LSTM 神经网络构建预测模型，并设计了 Attention 机制层针对数据序列特性进行改进，最后结合相关指标对房价进行预测。结果表明，该模型的预测效果相较传统 LSTM 模型更精准。

张智鹏，郑大庆[10]以北京市部分居民住房数据作为研究对象，应用梯度提升决策树模型(GBDT)对房价影响因素进行分析，选出其中影响较大的因素，并设计房价预测界面，对短时间内的区域房价进行预测，尽可能避免了人为因素的干扰，为房产提供了客观合理的估值，也为房地产行业的良性发展提供了参考。

罗博炜，洪智勇，王劲屹[11]选用 2019 年美国波士顿地区的 6028 条房产数据作为实验数据集，去除掉部分异常值和无效值之后对数据进行多元回归分析，进而构建了多元线性回归模型。实验结果表明修正后的模型相较于多元线性回归统计模型，实验误差进一步缩小。不过研究使用的数据集样本数量过小，不具代表性，实际应用性不强。

郑永坤，刘春[12]选用 2013 年至 2018 年间广州和深圳的二手房数据作为实验数据集，构建基于时间序列的 ARIMA 模型对房屋进行训练和房价预测，使用滚动预测的方法进行持续性预测，相较于直接使用模型预测，对于预测精度有进一步提升，为房屋买卖者提供了一定的参考。

1.2.2 国外研究现状

国外学者针对房价预测问题也进行了一系列的研究工作，除了常用的 ARIMA 序列模型和神经网络之外，国外学者还使用了多元线性回归模型、集成学习、深度信念网络以及多种神经网络的组合等，具体研究内容如下：

Khamis[13]选取了纽约市 1047 套房屋样本数据作为实验数据集，其中房价影响因素包括室内面积、房屋总面积、浴室个数、房间个数以及房屋建成年份等。首先使用神经网络模型对实验数据进行训练并进行预测，再使用多元线性回归模型进行房价预测对照组实验。结果表明，神经网络模型的房价预测效果

更好一些。但是实验使用的数据集中二手房价格不是真实成交价格，只是房屋估值，因此研究具有一定的局限性。

Yang[14]获取了部分加利福尼亚州的房产数据作为实验数据集，将各种预测因子与集成学习(Ensembel Learning)算法相结合，构建模型并进行房价预测。同时分别开展基于随机森林和梯度提升决策树算法的房价预测对照组实验。结果表明，采用集成学习的房价预测模型相对于其它两种模型表现出更高的预测精度和更好的稳定性，同时该模型可以降低数据噪声，有效避免了模型的过拟合。

Wang[15]等选取美国波士顿等几个城市的房产数据样本作为实验数据集，在神经网络中加入记忆电阻之后可以对实验数据进行自动训练，再通过调整忆阻器的权重来同步构建回归模型并对房价进行预测。结果表明，训练和预测同步进行的模式提升了房价预测精度，使得预测值更接近于真实值。

L Hao[16]选取美国爱荷华州埃姆斯市房地产市场的 1461 套房产交易数据作为实验数据集，其中包括 41 个分类变量和 38 个连续数值变量，共计 79 个数据特征。仔细分析数据属性之后，基于深度信念网络算法构建了深度学习预测模型并开展房价预测研究。与此同时，基于偏最小二乘法、人工神经网络和支持向量机三种回归模型，分别开展房价预测对比实验，结果表明，基于深度信念网络模型的房价预测效果好于其它三种模型的预测效果。

Serrano[17]收集了房地产、股票和债券市场的价格相关时序序列数据作为实验数据集，结合各数据之间的相关性，使用循环神经网络(RNN)构建预测模型并逐一进行价格预测，实验结果表明，基于 RNN 的预测模型能够对不同领域的价格走势做出准确预测，从而得出最佳投资组合。

Zona Kostic[18]采集部分与房屋属性有关的图像，之后从不同的视角对这些图像的合理性进行论证，再使用这些图像量化房屋的内部特征并构建房地产价格预测模型对房价进行直接预测，预测过程中对于可见特征的描述真实有效，无需更换专家，避免了人为因素的干扰。但是实验使用的数据集中样本数量过少且特征属性有限，涉及房产区域较为单一，忽略了不同城市之间的差异性，容易造成模型的过拟合。

1.2.3 缺点分析

通过对国内外学者关于房屋价格的研究成果的梳理，我们发现大部分学者基本都是选择传统方法，稍加改进和优化之后直接开展房价预测工作，存在着以下缺点：

(1) 实验模型选取存在不足。房产价值是由多种因素共同决定的，多因素差异导致不同价值类别的房产价格悬殊，使用相同模型进行价格预测时会出现较大误差。先前学者大多只研究了对房价影响较大的因素，未将房产数据进行价值分类，因此难以准确预测房价。

(2) 拟合优度欠佳，部分学者在房产数据进行价格预测时，对于数据处理不完全，实验数据特征维度过高，后期无法增加实验样本数量，导致数据的数量级小于模型的复杂度，进而出现过拟合。

(3) 实用性不强，对于一些较为高级的预测方法，虽然有部分实验说明有效，但是多数偏理论并未实际应用于我国的城市，实践意义不大。

本文的数据来源是国内某城市的真实房产交易数据，在数据特征选取[19]上不仅选取了建筑固有属性，如房屋所在位置、所在楼层、房间数量、公摊面积、房屋面积等微观因素，也添加了包括房屋周边公园数量、商场数量、医院数量、公交站数量、地铁口数量在内的诸多便民特征信息[20]。数据预处理之后，对数据进行了降维处理。对数据特征分析之后，选取出特征重要性最大的前十项数据特征，避免了模型训练不充分而导致的欠拟合或者过拟合现象。对于实验方法，本次研究进行创新，在开展房价预测工作之前，先使用分类算法对数据集进行多分类处理，即将数据集进行分段[21]，价值类别相同的房产数据被划分到同一个子数据集中。接着，使用回归模型分别对不同价值的数据集进行房价预测，避免了不同价值的房屋预测的价格出现较大范围误差波动，有效提高了模型的预测精度。

1.3 主要内容

由于房屋的价格影响因素较多且较为复杂，不同类型的房屋又存在一定的价值差异性，使用单一的机器学习算法直接对其进行价格预测会存在较大的误差。本文从实际情况出发，将多种分类器与回归算法结合，构建高效且实用的预测模型对房屋进行价值评估。结合房价主要影响因素，本次研究首先使用多种分类器对每条房产数据进行价值分类，针对分类之后的数据集，再使用回归模型进行房产价格预测，预测效果相较于传统预测模型有大幅度提高。研究成果不仅成为消费者购房时的重要参考指标，对于相关监管部门的宏观调控工作也具有一定的指导意义。本文的主要工作有：获取真实房屋交易数据，导入至数据库中并对数据集进行相关清洗，并对数据进行特征重要性分析，选出对房价影响较大的特征维度作为分类实验的数据特征，分类实验完成之后生成的低

中高三种价值的房产数据集用于房价预测工作。最后参考获取到的数据基础，对未来房价进行预测。具体研究内容如下：

(1) 首先获取数据集，对采集到的数据集进行相关清洗工作，再对数据进行简单分析，找出每条数据的房屋面积以及所在楼层属性，从而验证原始数据的合理性。最后对数据样本进行特征重要性分析，提取出对房价影响最大的十项特征，去除对房价影响较小因素。

(2) 深入研究贝叶斯神经网络、梯度提升决策树、K 近邻和逻辑回归四种分类器的原理，掌握其在房产数据领域的使用方法。接下来对随机森林模型、决策树模型、多元线性回归模型、门控循环单元和长短期神经网络模型的内部结构以及组成原理进行详细的了解，做好充分的理论准备。

(3) 详细介绍数据预处理及评价指标选择，根据房屋的最近一次成交价从低到高，依次将数据样本分为较低价值、中等价值和较高价值三种类别。并选择确定了分类器的各种性能评价指标，其中包括准确率、精确率、F1 值、召回率；回归算法的性能评价指标主要包括平均绝对误差、均方根误差和平均百分比误差。至此，所有实验相关的评价指标均已选取。

(4) 建模实验，先使用贝叶斯神经网络、梯度提升决策树、K 近邻和逻辑回归四种分类器分别对实验数据集进行价值分类处理，得到 12 个不同的子数据集。接着使用多元线性回归、决策树、随机森林、长短期神经网络和门控循环单元五种回归模型依次对分类得到的 12 个子数据集进行房价预测。最后，使用五种回归模型对未分类的数据集进行房价预测。详细记录各实验结果。

1.4 论文结构

本文主要研究基于多种分类器的房价格预测工作，首先对本次研究的相关背景和国内外研究现状及发展趋势做出一定的了解，接着对本次研究过程中使用到的机器学习算法进行深入了解，获取数据集分析数据并分离出房价影响较大的因素，再选择合适的实验评价指标，然后进行实验并对比分析各组实验结果，最后将本次研究做出一个总结。具体章节内容如下：

第一章绪论部分，本章深入了解本次研究涉及的时代背景和意义，掌握国内外当前研究现状，并分析其优点与不足。通过理论研究，将多种分类器与房产价格预测相结合，为本次研究提供了研究内容和实验思路。

第二章总体方案设计及方法介绍部分，本章对总体实验方案进行设计，并着重介绍了实验过程中需要用到的多种分类器和回归算法，熟悉了各种算法原

理及使用方法，做好实验准备工作。

第三章数据处理及相关指标选择部分，本章首先介绍了数据预处理过程，对数据进行特征重要性分析后进行降维操作。然后选取实验相关评价指标，对于分类实验，选取分类性能(准确率、精确率、召回率、F1 值)；对于回归实验，选取平均绝对值误差(MAE)、均方根误差(RMSE)和平均百分比误差(MAPE)。

第四章实验过程部分，本章首先详细介绍了实验所需要的环境和实验步骤，然后按照步骤分别完成各组实验，认真观察实验数据，并详细记录实验结果。

第五章实验结果对比分析部分，本章汇总实验结果并进行对比分析，首先对比了各种回归模型处理未分类数据集时的各项评价指标，得出哪种回归模型处理未分类数据集时的预测效果最好。接着仔细对比五种回归模型依次处理贝叶斯神经网络分出的三种不同数据集时的预测效果，观察五种模型处理哪种价值的数据集时，预测效果相对较好。然后将每种回归模型处理不同分类器分出数据集时的各项误差指标分别与该种回归模型处理未分类数据集时的各项评价指标进行对比，并将五种模型依次处理不同分类器分出数据集时的误差指标相互对比，得出预测效果最佳的回归模型。最后，将各种回归模型预测效果最佳时的各项评价指标进行比较，确定哪种分类器和哪种回归模型的预测组合效果最佳，并计算出各项预测实验误差指标的减少值。

第六章总结与展望部分，本章主要是对本次研究进行总结，并分析研究过程中存在的一些不足与展望，后期可以在本次研究的基础上进一步提升。

1.5 本章小结

本章首先描述了房地产行业的发展背景，依次从宏观和微观角度分析了房地产价格预测研究的意义，并分析指出了国内外相关研究的缺点与不足，突出了本次研究的优势所在，确定了先分类后预测的实验路线，并简要概括了本文的研究内容和论文结构。

第二章 总体方案设计及方法介绍

2.1 总体方案

本次研究首先从专业学术网站 Kaggle 上采集房屋交易数据，并进行数据清洗。根据房屋的最近交易价格将数据样本划分成较低价值、中等价值和较高价值三种类别，所有数据样本均添加价值标签。之后根据部分数据特征之间的关系，对数据样本开展初步分析工作。再使用随机森林分类器对数据各项特征进行重要性分析，去除掉对房屋价值影响较小的特征属性，保留下对房屋价值影响最大的前十项特征，从而达到数据降维的目的。

接着开展实验部分，首先选取好各种实验评价指标，然后结合房产数据的前十项特征，使用贝叶斯神经网络、梯度提升决策树、K 近邻和逻辑回归四种分类器依次对房产数据集进行价值分类，得到 12 个不同的数据集，每种分类器分出的每种价值的数据集都分别使用多元线性回归、决策树、随机森林、长短期神经网络以及门控循环单元五种回归模型进行预测。每一种回归模型处理每一种分类器分出的三种价值数据集时，都可以得到三组评价指标(MAE、RMSE、MAPE)。结合每种价值的数据集在测试集中的权重占比，将三个 MAE、三个 RMSE、三个 MAPE 分别进行加权平均计算，所得结果分别作为该种回归模型处理该分类器分出数据集时的 MAE、RMSE、MAPE。所有预测实验结束之后，得到五种回归模型依次处理不同分类器分出数据集时的评价指标(MAE、RMSE、MAPE)。同时使用各种回归模型依次处理未分类数据集，作为对照组实验。

最后是各实验结果的对比分析部分，先将各种回归模型处理未分类数据集时所得误差指标进行比较，得出哪种回归模型处理未分类数据集时的预测效果最好。接着仔细对比五种回归模型依次处理贝叶斯神经网络分出的三种不同数据集时的预测效果，观察五种模型处理哪种价值的数据集时，预测效果相对较好。然后将每种回归模型处理不同分类器分出数据集时的误差指标与该种模型直接处理未分类数据集时的误差指标逐一对比，可以直观地看出分类器对于房价预测实验效果的有效提升。再通过对比每种回归模型处理不同分类器分出数据集时的误差指标，得出该种模型的评价指标最小值。最后通过对比每种回归模型处理不同分类器分出数据集时的误差指标最小值，确定哪种分类器与哪种回归模型的组合模型误差指标最小，预测效果最佳，并分析主要原因。同时，计算出分类器与回归模型的最佳组合相对于直接预测时效果最佳的模型各项误差指标的减少值。本次研究的

实验整体架构如图 2-1。

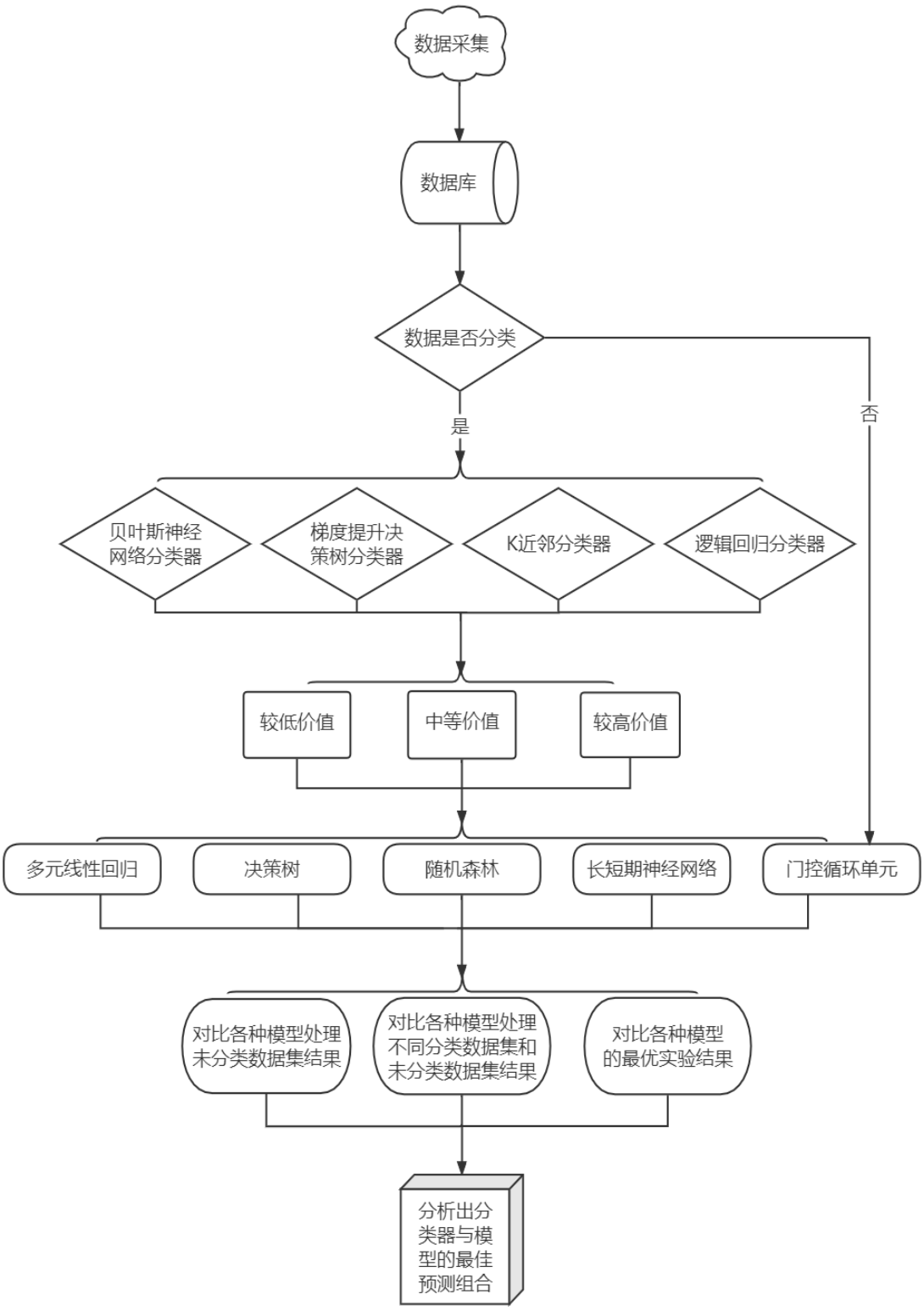


图 2-1 实验整体架构

2.2 分类器

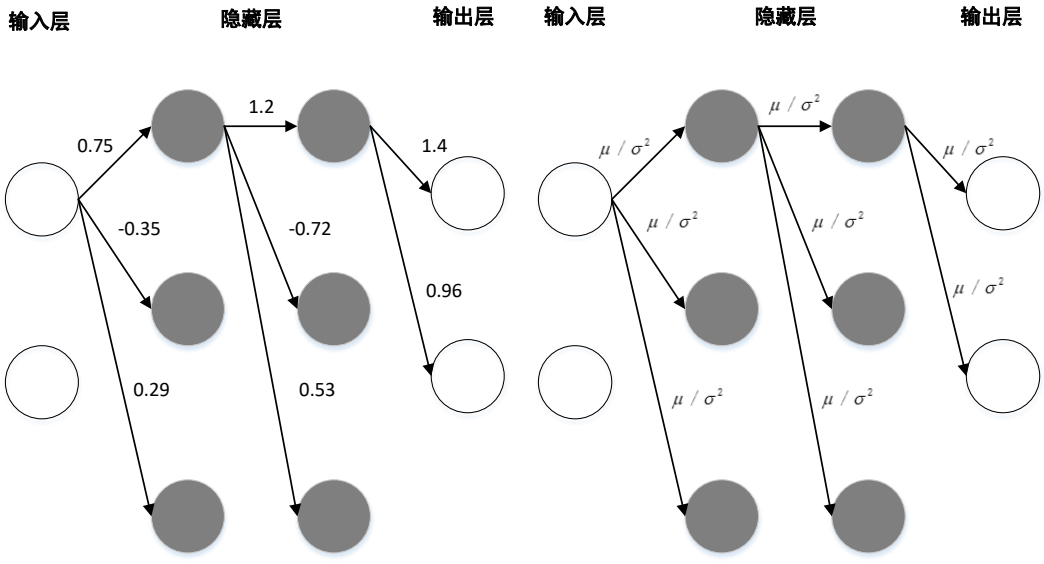
2.2.1 贝叶斯神经网络

贝叶斯神经网络(Bayesian Neural Network，BNN)[22,23]是基于贝叶斯定理的一种特殊神经网络。传统神经网络每次访问神经元时得到的权重参数都是固定值，一般通过交叉熵或者损失函数来拟合标签值。而贝叶斯神经网络则使用概率分布的参数来代替固定值作为权重参数，再通过变分推断法来计算后验分布，有效降低了模型的过拟合，两者的结构差异对比如图 2-2。神经网络的初始化具有随机性，一般将各节点都设置成均值 μ 为 0，方差为 σ2 的正态分布。

传统神经网络凭借强大的非线性拟合能力在各领域应用广泛，其核心工作是根据训练集的数据，得到各层的模型参数，从而最小化 Loss 值。对于贝叶斯神经网络，核心工作就是最小化 KL 散度，KL 散度的计算方式如式 2-1。

$$KL(q \| p)) = E_q[\log q(W) - \log p(W) - \log p(Y| X,W) + \log p(Y| x)] \quad (2\text{-}1)$$

贝叶斯神经网络的优点[24]是能使用较少的数据得到较好的模型，而且得到的是各层参数的分布(一般假设各层参数 w_i，b_i 服从高斯分布，根据训练集数据计算得出 w_i，b_i 的均值和方差)，从而得到 $p(W| X,Y)$，除了能对结果进行预测，还可以对预测结果不确定性进行量化，具有非常强的鲁棒性。



(a) 传统神经网络 (b) 贝叶斯神经网络

图 2-2 传统神经网络与贝叶斯神经网络结构对比图

2.2.2 梯度提升决策树

梯度提升决策树(Gradient Boosting Decision Tree，GBDT)是一种以决策树为基学习器的分类方法[25,26]。它通过梯度提升的方法集成多个决策树，降低了单棵决策树的复杂性，有效避免了过拟合现象。该分类算法具有便于处理多类型数据、计算速度快、对异常值的鲁棒性强等特点。梯度提升决策树分类算法的学习过程中，每迭代一次就会通过贪心策略[27]生成一棵新的决策树，计算出树中每个叶子节点的预测值之后，根据公式 2-2 将新决策树添加到模型中。

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t\left(x_i\right) \tag{2-2}$$

上式中的 $f_t\left(x_i\right)$ 表示新生成的决策树，$\hat{y}_i^{t-1}$ 表示加入新决策树之前的模型，$\hat{y}_i^t$ 表示加入新决策树之后的模型。

2.2.3 K 近邻

K 近邻(K-Nearest Neighbors，KNN)是一种理论成熟，思想简单的机器学习分类方法[28,29]，被学者广泛运用于多分类问题。该分类算法的核心思想为多数表决法，即先输入测试数据，将测试数据中的所有特征属性与训练集中数据的所有特征属性和已知标签进行一一比较，找到两者最相似的 K 组数据样本，最后 K 组数据样本中出现最多的类别就是该测试数据样本的对应类别。分类过程中，一般采用交叉验证的方式来确定最合适的数值作为 K 值，采用欧氏距离[30]的方式进行距离度量，计算方法如式 2-3。该分类算法具有准确性高、对异常点不敏感、适用于非线性分类和大样本量分类等优点。

$$D(x,y) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \tag{2-3}$$

上式中 n 表示空间维度，计算结果表示 n 维空间中两点的真实距离，x_i 和 y_i 分别表示两个点在某个空间内的坐标。

2.2.4 逻辑回归

逻辑回归(Logistic Regression，LR)是一种原理简单的分类方法[31-33]，常用于处理二分类及多分类问题。该分类算法的特点是将样本的发生概率与样本数据中

各特征属性相关联，输出部分为离散值。逻辑回归具有简明易懂、计算速度快、存储资源低、便于实现及适用于多重共线性问题等优点。分类过程中的核心问题是代价函数(Sigmoid)的构建，再通过改进方法对模型的参数进行迭代调优，最后对求解出的模型最优参数进行测试验证，以确保其合理性与有效性。Sigmoid 函数的表达式如式 2-4。

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-4}$$

2.3 回归算法

2.3.1 多元线性回归

多元线性回归(Multivariable Linear Regression，MLR)[34]是一种常见的回归预测方法。该算法将多种因素相关联，计算出最佳因素组合作为预测中的自变量，而所求标签作为预测中的因变量，通过建模即可得到其数值。实际生产生活中，绝大多数的现象都受到多个不同因素的影响，因此多元线性回归相较于一元线性回归实用性更强，被普遍运用于预测领域[35]。多元线性回归算法具有适用性广、可解释性强以及机制原理简单等优点。房价的走势受多个相互关联因素的综合影响，而本次研究的目的就是使用具有时序性的房产数据对房价进行预测，符合多元线性回归的算法特性，实验过程中根据房产数据各影响因素建立预测模型并分析实验误差。

2.3.2 决策树

决策树(Decision Tree，DT)是一种呈树形结构的基本算法，经常用于处理回归预测问题[36,37]。决策树模型的构建过程包括决策属性的选择、生成决策树和对决策树进行一定的剪枝。剪枝的目的是为了避免训练过程中出现过拟合现象。决策树的结构特性决定了其只能有一项输出，复数的输出只能通过生成多棵决策树来实现。决策树模型具有可读性强、适用于离散属性以及运行速度快等优点。本次实验过程中对于房价的预测属于单一变量输出，且使用的数据集中各属性之间呈非线性关系，因此使用决策树这种简单实用算法进行处理较为合适。

2.3.3 随机森林

随机森林(Random Forest，RF)[38]是一种性能优越的集成学习算法，其原理较为简单，被广泛应用于价格预测方面[39-41]。该算法在决策树的基础上引入了随机属性选择，同时将多棵树进行合并。不仅大幅提升了分类能力，同时有效避免了单棵决策树的过拟合现象。

随机森林算法的基本思想[42]是通过 Bootstrap 抽样[43]从原始数据集中随机抽取样本，并有放回地重复 n 次，根据选取出的数据样本的特征属性构建一棵决策树并对决策树进行训练，然后重复上述步骤直到生成包含 m 棵决策树的随机森林。当新的数据样本出现时，随机森林对其类别的判断方法就是利用所有决策树共同投票，投票结果中的最高分对应的类别即为新样本数据的类别。随机森林的分类原理如图 2-3。该算法具有性能稳定、抗噪能力良好、计算开销小、准确率高、适合处理高维数据以及一定程度避免过拟合等优点。

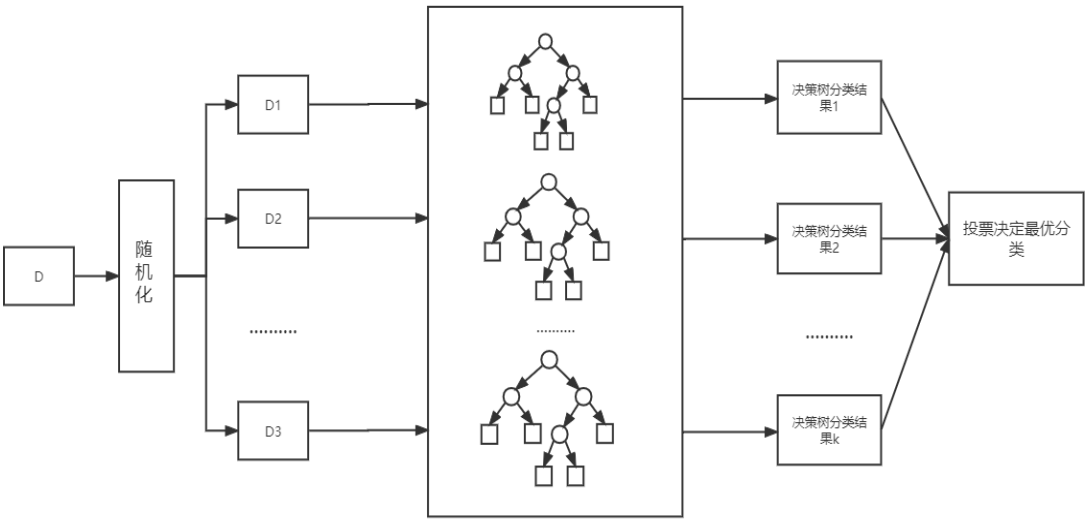


图 2-3 随机森林分类原理图

2.3.4 长短期神经网络

长短期神经网络(Long Short-Term Memory，LSTM)[44]是一种基于 RNN 模型的改进算法，有效解决了 RNN 中的长期依赖问题，在价格预测领域[45,46]应用广泛。由于 LSTM 是一种时间循环神经网络，而本次研究使用的房产数据为时序数据，因此使用该算法进行建模处理时表现优异。该模型不仅具有链状结构，同时有着独特的重复模块。组成该重复模块的四层神经网络之间能够相互作用，有效

提高了该模型的回归预测效果[47]。LSTM 神经网络的基本结构如图 2-4。

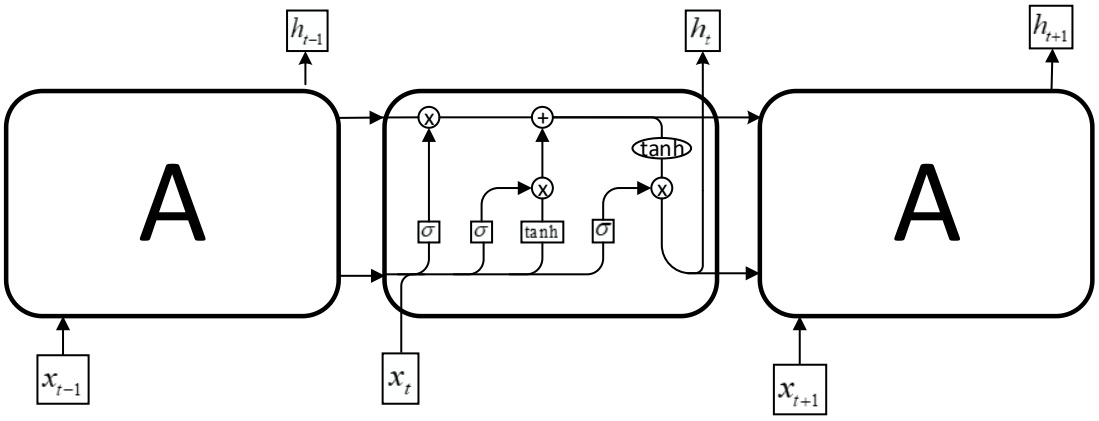


图 2-4 LSTM 结构图

图中的 X 和+分别表示的是缩放信息和添加信息，б 表示 LSTM 中的 Sigmoid 激活函数层，tanh 表示 LSTM 中的 tanh 激活函数层，h(t-1)表示上一个 LSTM 单元的输出，h(t)表示当前 LSTM 单元的输出，h(t+1)表示下一个 LSTM 单元的输出。X(t-1)表示上一个单元的输入，X(t)表示当前输入，X(t+1)表示下一个单元的输入。所有传输单元状态相互连接，贯穿整个重复结构，共同决定 LSTM 网络。此外，LSTM 可以对单元状态中传输的信息进行添加或者删除，几个结构共同组成门限来传输信息并选择性让其通过。通过分析 LSTM 的结构原理，可以看出 LSTM 适合处理时序性数据，而房屋价格随着时间变化不断波动，因此房产数据集适合采用该模型。

2.3.5 门控循环单元

门控循环单元(Gated Recurrent Unit，GRU)是由 Cho、van Merrienboer、Bahdanau 和 Bengio 在 2014 年提出的，是 LSTM 网络的一种改进算法，相较于 LSTM 神经网络，GRU 的结构更为简单，大幅度提升了模型的训练效率，使得效果更理想，因此被广泛应用于价格预测领域[48-50]。相对于 LSTM 中的输入门、输出门和遗忘门三个门函数，而 GRU 改为更新门和重置门两个门函数。GRU 模型的结构如图 2-5。

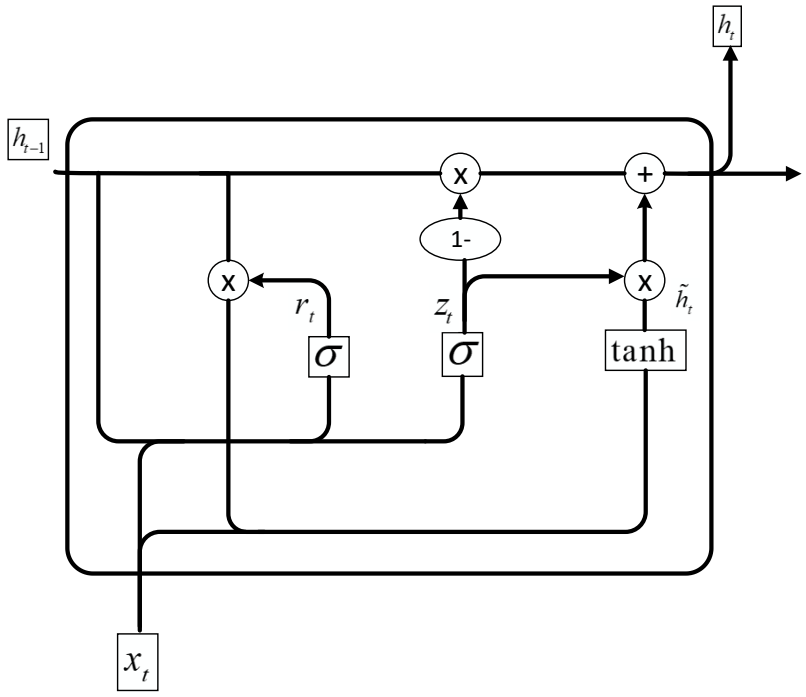


图 2-5 GRU 结构图

图中的 z_t 表示更新门，r_t 表示重置门，$\tilde{h}_t$ 表示当前状态的候选集。其它符号与 LSTM 结构中的含义相同，更新门决定前一时刻的状态信息被带入到当前状态中的程度，其值越大，说明前一时刻的状态信息带入越多。重置门则决定前一状态带入的信息有多少将被写入 $\tilde{h}_t$ 中，其值越小，写入 $\tilde{h}_t$ 中的上一时刻状态信息就越少。由于 GRU 相对于 LSTM 少了一个结构，参数更少，在处理大量数据时速度更快，适用于本次研究中的数据样本。

2.4 本章小结

本章首先交代了本次研究的总体设计方案，并详细介绍了实验流程以及对实验结果的对比方法。结合本次研究的内容，深入学习了实验过程中需要使用的各种分类算法和回归模型的相关原理知识，并对各种算法的优点进行了一定的分析，为实验的顺利进行打好理论基础。

第三章 数据处理及评价指标选择

3.1 数据预处理

3.1.1 数据获取

本次研究是从专业学术网站 Kaggle 上获取到国内某市的真实房屋交易数据，使用 Navicat 工具将数据集导入 Oracle 数据库[51]中，发现数据集中一共包含 21562 条交易记录，房屋的特征属性具体包括：房产交易完成天数，房产上次交易总价，房屋所属小区均价，所属开发商，所属物业，房屋面积，房间数量，公摊面积，公交车路线数量，所在楼层类型，总楼层，房屋建成年份，装修情况，梯户数量，是否回迁房，是否有车位，房屋所属行政区，周边工厂数量，周边商场数量，周边学校数量，周边公园数量，周边医院数量，周边人口数量，有无地铁口，是否靠近政府以及所属行政区在 2020、2021 年的 GDP 值。

3.1.2 数据清洗

数据清洗[52,53]就是根据一定的数据清洗策略与规则，利用相关技术将不规范或不完整的数据转化为满足指定要求的可用数据集，具体原理图如图 3-1。原始数据中常见的问题一般包括同一字段的不同表示、相同数据重复记录、存在噪声数据和无效数据、存在空值和缺失值。因此，我们首先要检查数据是否合乎要求，对于超出正常范围以及逻辑不合理的数据，给予适当的处理。本文中的数据清洗过程主要分成数据删除、缺失数据修复[54]以及价值分类三步。

(1) 数据删除

本文获取到的初始房产数据不理想，数据中存在一定的重复样本以及数据类型错误的无效特征值。比如“Floors”代表总楼层数，数据类型为整型，部分样本的该值为 3.65，则属于无效值。删除掉数据集中无效值对应的样本和重复样本后，保证了数据集的准确性和有效性。

(2) 缺失数据修复

对于一些特征值的缺失，如果缺失率较高选择删除该特征，某些样本如果同时缺失多个特征值则删除这些数据样本。部分特征缺失率较低，可以采取中值填充的方式对其进行修复，尽可能地保留了原始数据集的特征。例如“Stall_Area”代

表房屋公摊面积(缺失率高达 76%)，我们选择将该特征删除。而特征“Num_Factories”代表周边工厂数量(缺失率不到 1%)，针对这种情况，可以人为地把“Num_Factories”的缺失值设置为该特征的中值。

(3) 价值分类

由于数据集中不同房屋的价值存在着一定的差异，为了降低在房价预测工作中出现的误差大小，本次研究对房产数据集进行进一步价值分类处理，后续实验中会结合房产数据的各属性，使用分类器对各条房产数据的价值种类进行预测训练，之后再使用多种回归模型分别进行房价预测工作，能有效减少误差。本文按照房屋最近成交价对数据集进行排序,将房屋价值从低到高依次划分成较低价值、中等价值和较高价值三类，分别用标签值 0、1、2 表示，再将数据集中的样本数据打乱。最后数据集中每一条数据样本均增添了一列标签属性，而标签列的内容即表示每一条数据的真实价值。

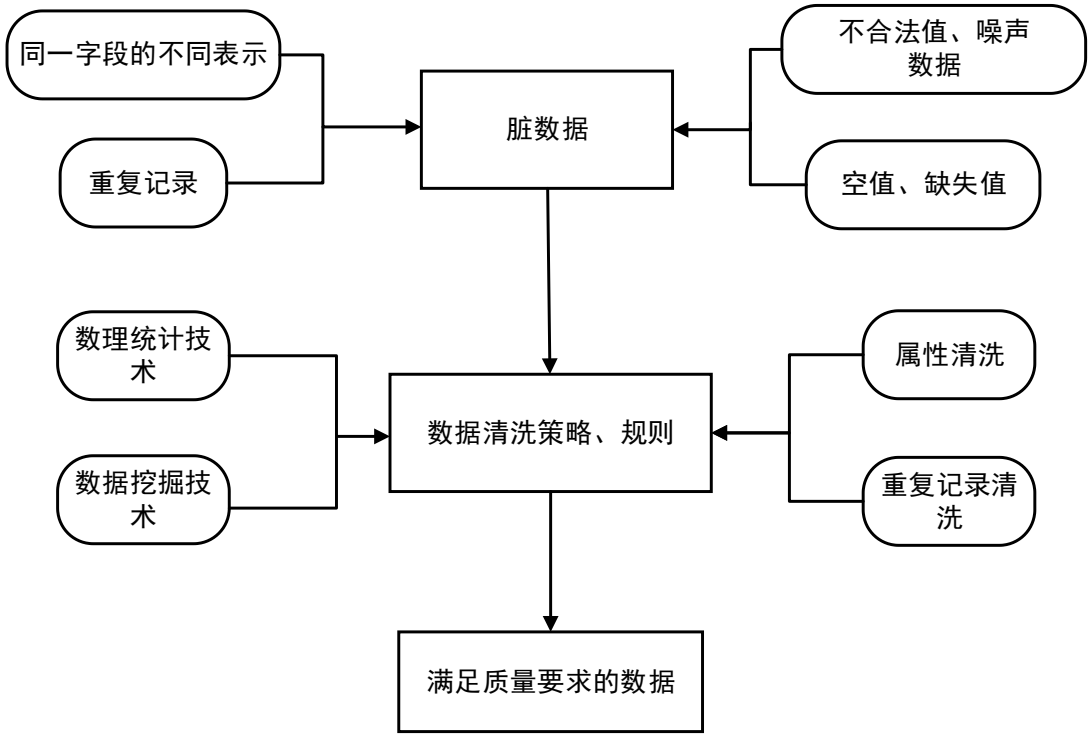


图 3-1 数据清洗原理

3.1.3 数据初步分析

数据分析指的是利用一定的工具手段对数据中各因素间的关系进行概括总结，从而挖掘更多有用信息和有价值的结论。数据清洗是开展数据分析工作的必要前提，文中数据清洗工作结束之后，数据集中还剩下 18362 条有效数据样本，

去除房屋公摊面积、小区公交车路线数量等属性之后，数据样本包括 21 项数据特征，部分数据特征的极大值、极小值、均值和标准差等具体信息如表 3.1。

表 3.1 部分房产数据特征的具体信息

特征名称	极大值	极小值	均值	标准差
上次交易总价	47830.0	5024.0	16452.5	756.5
交易完成天数	1204.0	13.0	363.5	299.5
房屋面积	739.700	17.990	61.415	39.115
房间数量	16	1	5	2
梯户数量	20.000	0.030	0.265	0.235

本文选择房产数据集中的房屋面积与成交量关系以及房屋所在楼层与成交价关系，对清洗过的数据进行初步分析，并以数据可视化[55]的方式将其展示出来。

(1) 房屋面积

房屋面积是人们购房时考虑的重点因素之一，房屋面积的大小直接决定了居住者的生活品质及环境体验。在进行社会调研后发现，生活中普通购房者最常选择的房屋面积在 80 平米至 120 平米之间。因此，本文进一步分析房产数据样本中房屋面积 80 平米以下、80 至 120 平米之间和 120 平米以上三种类型房屋的成交量以及在所有数据样本中所占的比例，如图 3-2。

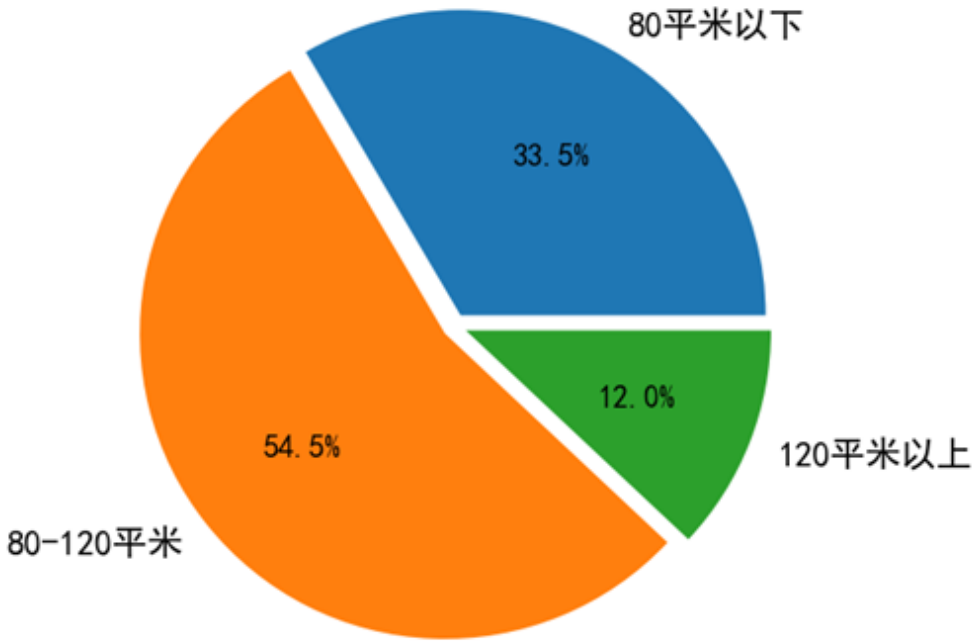


图 3-2 不同面积的房屋成交量占比

根据图 3-2 可以看出，所有成交的房屋数据中，房屋面积 80 平米以下的占 33.5%，房屋面积在 80 平米到 120 平米之间的占 54.5%，房屋面积 120 平米以上

的占 12%，该分类结果符合实际生活中大部分人的购房选择，间接证明了数据集中房产数据的合理性和真实性。

(2) 所在楼层

商品房开发过程中，同一楼盘中不同楼层的房屋价格也有所区别。在查阅相关文献[56]以及实地考察之后，总结出低楼层房屋通常伴有采光差和易潮湿问题；高楼层房屋容易出现由于水压不足引起的供水问题，此外等待电梯时间较长，严重影响出行效率；中间楼层的房屋相对宜居，但是数量相对较少，对应的价格也更高。本文针对各种楼层的房屋，计算出各种楼层房屋的成交价中值以及各种楼层房屋在总交易记录中所占的比例，具体情况如图 3-3。

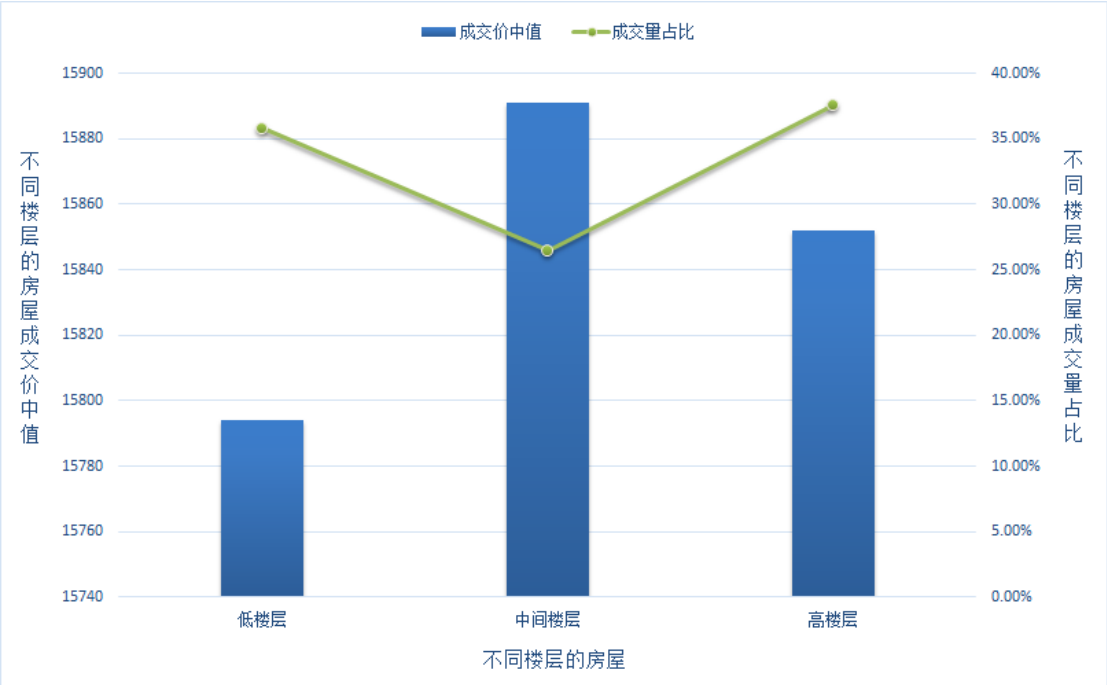


图 3-3 不同楼层的房屋成交情况

根据图 3-3 可以看出，中间楼层的房屋成交价中值偏高一些，且成交量相对较少，而低楼层和高楼层的房屋成交价中值偏低一些，对于普通购房者有更强的吸引力，因此房屋成交量占比相应更高一些。综上所述，数据整体表现符合实际情况，满足学术研究基本要求。

3.2 数据降维

数据降维[57]就是去除高维特征数据中的噪声以及不重要属性，将重要的数据

特征保留下来，从而降低算法计算开销，达到加速数据处理的目的。经过数据清洗之后，去掉了无效和重复的数据，发现数据集中的特征维度较多较复杂，部分无关属性与和冗余特征在高维数据处理中容易造成模型训练不充分，导致过拟合或者欠拟合现象，从而影响了到价值分类的效果。本次研究在进行分类实验之前，预先对数据进行特征重要性分析，通过特征选择去除数据中的无关属性，保留对房价走势影响较大的数据特征，达到了降维的目的。实验过程中使用随机森林分类方法[58]计算出房产数据除价值标签属性之外的各项特征属性的重要性大小，效果如图 3-4。

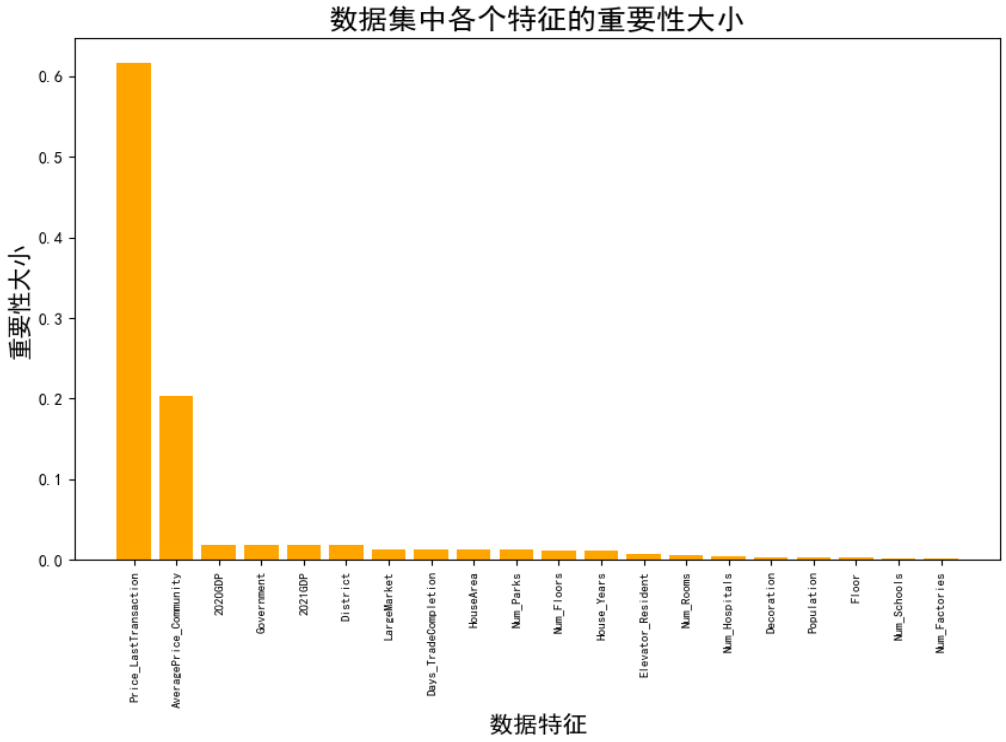


图 3-4 房产数据各特征重要性

根据图 3-4 可以看出，对房屋价值影响最大的前十项特征从高到低分别为房屋上次交易总价、所属小区均价、所属行政区在 2020 年的 GDP、是否离市政府较近、所属行政区在 2021 年的 GDP、房屋所属行政区、周边商场数量、房产交易完成天数、房屋面积、周边公园数量。分类实验部分会对房产数据进行价值类别预测处理，好的特征选择能够提升分类器的性能，更能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。因此，我们选择房产数据的前十项特征作为下文中房屋价值分类实验的依据。

3.3 评价指标选择

3.3.1 分类评价指标

本文对房产数据进行多类别分类研究[59]，将房产价值划分为三种，分别为：较低价值(0)、中等价值(1)和较高价值(2)，采用的分类评估标准为：准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 值。多分类类别如表 3.2。

表 3. 2 多分类类别

混淆矩阵		预测标签		
		0	1	2
真实标签	0	a	b	c
	1	d	e	f
	2	g	h	i

在此模型中，预测的样本被区分为九种类型，分别标记为 a、b、c、d、e、f、g、h、i。其中，a 表示预测结果和实际结果一致且均为较低价值，b 表示预测结果为中等价值而实际结果为较低价值，c 表示预测结果为较高价值而实际结果为较低价值，d 表示预测结果为较低价值而实际结果为中等价值，e 表示预测结果和实际结果一致且均为中等价值，f 表示预测结果为较高价值而实际结果为中等价值，g 表示预测结果为较低价值而实际结果为较高价值，h 表示预测结果为中等价值而实际结果为较高价值，i 表示预测结果为较高价值而实际结果为较高价值。

在本次房产价值分类研究中，a、e、i 的值越高，说明分类模型的分类效果越好。相反，b、c、f、g、h 越低，模型越能准确地分辨出房产的价值类别，分类性能越好。本次分类研究的混淆矩阵是一个 3×3 的矩阵，从矩阵的对角线角度来看，主对角线上面的值均为分类正确，其它值均为分类错误。具体评估方法[60]如下：

(1) 准确率(Accuracy)，表示预测结果与实际结果相同的样本数量与总样本数的比值，表达式如式 3-1。

$$Accuracy = \frac{a+e+i}{a+b+c+d+e+f+g+h+i} \tag{3-1}$$

(2) 精确率(Precision)，要分别计算各个类别的精确率，若将 0 作为正类，则表示样本中实际结果和预测结果均为较低价值的样本与所有预测结果为较低价值的样本的比值；若将 1 作为正类，则表示样本中实际结果和预测结果均为中等价

值的样本与所有预测结果为中等价值的样本的比值；若将 2 作为正类，则表示样本中实际结果和预测结果均为较高价值的样本与所有预测结果为较高价值的样本的比值。本次研究的精确率将上述三种计算结果的平均值，表达式如式 3-2。

$$Precision=\frac{\frac{a}{a+d+g}+\frac{\mathrm{e}}{b+e+h}+\frac{i}{c+f+i}}{3} \tag{3-2}$$

(3) 召回率(Recall)，要分别计算各个类别的召回率，若将 0 作为正类，则表示样本中实际结果和预测结果均为较低价值的样本与所有实际结果为较低价值的样本的比值；若将 1 作为正类，则表示样本中实际结果和预测结果均为中等价值的样本与所有实际结果为中等价值的样本的比值；若将 2 作为正类，则表示样本中实际结果和预测结果均为较高价值的样本与所有实际结果为较高价值的样本的比值。本次研究的精确率将上述三种计算结果的平均值，表达式如式 3-3。

$$Recall=\frac{\frac{a}{a+b+c}+\frac{\mathrm{e}}{d+e+f}+\frac{i}{g+h+i}}{3} \tag{3-3}$$

(4) F1 值，表示的是精确率(Precision)与召回率(Recall)的加权调和平均值，表达式如式 3-4。

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{3-4}$$

3.3.2 预测评价指标

为了客观准确地衡量各预测模型的性能，本次研究首先选用平均绝对误差(Mean Absolute Error，MAE)、平均百分比误差(Mean Absolute Percentage Error，MAPE)以及均方根误差(Root Mean Square Error，RMSE)作为回归模型的预测指标，计算方式如式 3-5、式 3-6 及式 3-7。

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{3-5}$$

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{3-6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3-7}$$

上述三式中的 y_i 表示真实值，而 $\hat{y}_i$ 表示实验所得预测值，n 表示预测总数，上述三种指标用于表示实验的预测精度，它们的值越小，模型的预测精度越高。

每种回归模型依次处理每种分类器分出的三种不同数据集时都能得到三组 MAE、RMSE、MAPE 值，再根据每种价值的数据集在测试集中的权重占比，将三组 MAE、三组 RMSE、三组 MAPE 值各自进行加权平均，计算结果作为该种回归模型处理该分类器分出数据集的 MAE、RMSE 和 MAPE 值。综上所述，五种回归模型依次处理每种分类器分出数据集时，MAE 的计算方法如式 3-8，RMSE 的计算方法如式 3-9，MAPE 的计算方法如式 3-10。

$$MAE = \sum_{j=1}^{m} W_j * MAE_j \tag{3-8}$$

$$RMSE = \sum_{j=1}^{m} W_j * RMSE_j \tag{3-9}$$

$$MAPE = \sum_{j=1}^{m} W_j * MAPE_j \tag{3-10}$$

上述三式中，m 表示价值类别数，W_j 表示第 j 种价值的数据集在测试集中的权重，MAE_j 表示第 j 种价值数据集的平均绝对值误差，$RMSE_j$ 表示第 j 种价值数据集的均方根误差，$MAPE_j$ 表示第 j 种价值数据集的平均百分比误差。

3.4 本章小结

这一章节主要介绍了实验数据来源以及具体的数据处理方法和过程，并选取本次实验的评价指标。首先获取到大量的真实房屋交易数据作为实验数据集，同时对实验数据进行清洗。数据清洗结束之后，对数据开展初步分析工作，分别从房屋面积和房屋所在楼层两个方面，验证了研究所用数据集的合理性。然后，利用随机森林分类器计算各项数据特征的重要性大小，得出了对房价影响最大的前十项特征。最后结合各种机器学习算法和深度学习算法的特点以及数据特征，分析得出分类实验中的性能评价指标和后续预测实验中的结果评价指标。

第四章 房产价格预测实现

4.1 实验环境

本次实验使用的是一台处理器为英特尔酷睿 i5-5200u，主频 2.20GHz，内存16GB，操作系统为Windows10(64 位)的四核笔记本计算机。数据库使用的是Oracle 11.2.0 版本，程序设计语言使用的是 Python3.7。Python 除了高性能之外，凭借着 Numpy、SciPy 等极其强大的第三方库，在大数据领域被广泛应用。此外，本次研究使用的数据量较大，使用 Python 编程可以有效地节约开发成本和运行空间。编程软件使用的是 Pycharm2017.2.3，该开发工具是一种 Python IDE，带有一整套独特的辅助功能，诸如智能提示和代码跳转等。

4.2 实验步骤

本次研究将实验所需数据导入之后，先完成一组基于四种分类器的分类实验，再完成一组基于所有分类得到的数据集的预测实验，最后做一组基于未分类数据集的预测实验，并对比分析各组实验结果。具体实验步骤如图 4-1。

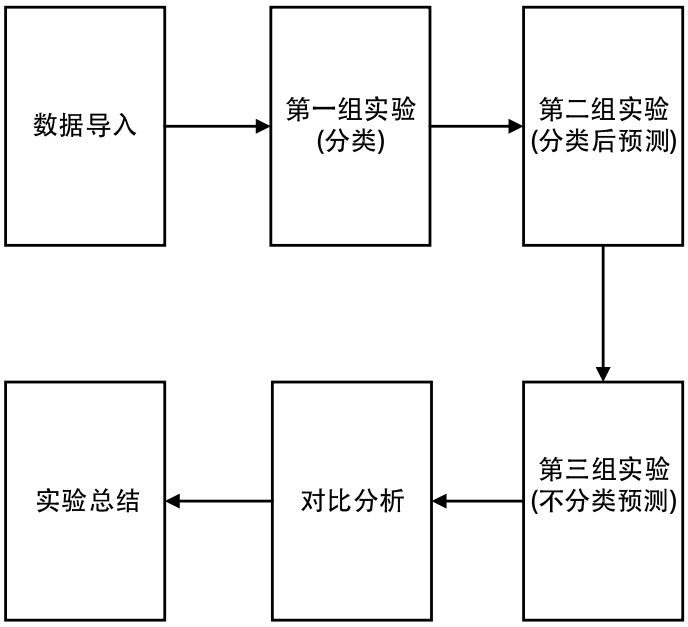


图 4-1 实验步骤图

(1) 首先使用 Navicat 工具将获取到的数据集导入到 Oracle 数据库中进行数据

清洗。清洗结束之后，对数据开展初步分析工作。之后将整理好的数据集进行特征重要性分析，计算出样本数据中各特征属性的重要性大小，确定对房价影响较大的前十项特征。

(2) 对房产数据集的分类实验。根据房产的属性特征，使用贝叶斯神经网络、提速度提升决策树、K 近邻和逻辑回归四种分类器依次对房产数据集进行价值分类，每种分类器分类后可以得到较低价值、中等价值、较高价值三个不同的数据集。 四次分类实验结束之后，生成了 12 个不同的数据集，并记录下各个数据集在测试集中的权重占比。详细记录各分类器的准确率、精确率、召回率和 F1 值。

(3) 对分类所得数据集的预测实验。使用多元线性回归、决策树、随机森林、长短期神经网络和门控循环单元五种回归模型依次对上述 12 个不同的数据集进行房价预测。每种回归模型处理每种分类器分出的三种价值数据集时，都可以得到三组评价指标(MAE、RMSE、MAPE)。结合每种价值的数据集在测试集中的权重占比，将三个 MAE、三个 RMSE、三个 MAPE 分别进行加权平均计算，所得结果分别作为该种回归模型处理该分类器分出数据集时的 MAE、RMSE、MAPE。所有预测实验结束之后，得到五种回归模型依次处理不同分类器分出数据集时的评价指标(MAE、RMSE、MAPE)，并进行详细记录。

(4) 对未分类数据集的预测实验。使用多元线性回归、决策树、随机森林、长短期神经网络和门控循环单元五种回归模型分别对未分类的房产数据集进行房价预测，作为上述分类后预测实验的对照组。实验结束后，得到各种回归模型处理未分类数据集时的评价指标(MAE、RMSE、MAPE)，并进行详细记录。

(5) 观察上述分类实验和预测实验的实验结果，汇总各种评价指标。首先将各种回归模型处理未分类数据集时所得误差指标进行比较，得出不使用分类器对数据集进行分类的情况下，预测效果最佳的回归模型。接着仔细对比了五种回归模型依次处理贝叶斯神经网络分出的三种不同数据集时的预测效果，观察五种回归模型处理哪种价值的数据集时，预测效果相对较好。然后将每一种回归模型处理不同分类器分出数据集时的误差指标与该种模型直接处理未分类数据集时的误差指标逐一对比，可以直观地看出分类器对于房价预测实验效果的有效提升。再通过对比每一种回归模型处理不同分类器分出数据集时的误差指标，得出该种模型的误差指标最小值。最后通过对比每一种回归模型处理不同分类器分出数据集时的误差指标最小值，确定哪种分类器配合哪种模型的组合预测模式误差指标最小，预测效果最有效，并分析主要原因。

(6) 对分类后预测实验和未分类预测实验的误差指标对比分析之后，得出本文的研究结论，并计算出分类器与回归模型的最佳预测组合相对于直接预测时的

最佳模型，各误差指标分别减少了多少。

4.3 实验过程

4.3.1 基于不同分类器的分类实验

(1) 贝叶斯神经网络分类

BNN 模型的构建基于 Pytorch 框架，读入房产数据集之后，按照 7.5:2.5 的比例划分训练集和测试集。使用 Sequential 方法构建一个 BNN 分类模型对房产数据集进行训练，训练过程中，Batch 中训练样本数量设置为 300，训练的轮数为 100。采用 KL 散度损失和交叉熵两种损失函数共同计算贝叶斯神经网络的损失值，将 KL 散度权重设置为 0.001，其中 KL 散度损失计算方式采用均值计算。通过反向传播最小化损失值，利用训练好的模型进行类别预测，得到最终的预测标签值。分类结束之后，房产数据的测试集被分成 BNN_low.csv，BNN_medium.csv 和 BNN_high.csv 三个不同价值的子数据集，依次计算出各个子数据集在总测试集中的权重占比。得到贝叶斯神经网络的分类效果如图 4-2。

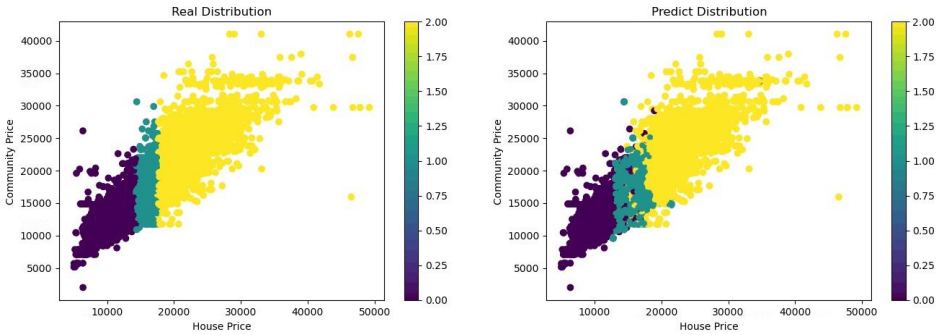


图 4-2 贝叶斯神经网络分类散点图

结合图 4-2 可以看出，贝叶斯神经网络对不同价值的房产数据进行类别预测时，中价值类别的房屋有一部分被判断成了低价值或者高价值，低价值房屋和高价值房屋均有一部分被判断成了中价值，极少数高价值房屋被判断成了低价值。

(2) 梯度提升决策树分类

在读入房产数据集后，使用 sklearn.ensemble.GradientBoostingClassifier 分类方法对房产数据集的前 75%进行训练。训练结束之后，利用训练好的梯度提升决策树分类器对房产数据集的后 25%进行测试，经过 10 折交叉验证之后得到最终的预测标签值。将预测标签值和真实标签值进行计算，分别得到梯度提升决策树

分类器的准确率、精确率、召回率和 F1 值。分类结束之后，房产数据的测试集被分成 GBDT_low.csv，GBDT_medium.csv 和 GBDT_high.csv 三个不同价值的子数据集，依次计算出各个子数据集在总测试集中的权重占比。

(3) K 近邻分类

首先读入房产数据集，按照 7.5:2.5 的比例划分训练集和测试集。数据特征维度为 10，输出标签即为房屋的价值类别。输出标签有三种可能值，分别为 0，1，2。使用 sklearn.neighbors.KNeighborsClassifier 分类方法对房产数据集进行训练，训练过程中邻近点的个数设置为 7。训练结束之后，利用训练好的 K 近邻分类器对房产数据集的进行测试，经过 10 折交叉验证之后得到最终的预测标签值。将预测标签值和真实标签值进行计算，分别得到 K 近邻分类器的准确率、精确率、召回率和 F1 值。分类结束之后，房产数据的测试集被分成 KNN_low.csv，KNN_medium.csv 和 KNN_high.csv 三个不同价值的子数据集，依次计算出各个子数据集在总测试集中的权重占比。

(4) 逻辑回归分类

在读入房产数据集后，将房产数据集的前 75%作为训练集，后 25%作为测试集。预测标签包含 0，1，2 三种类别。使用 sklearn.linear_model.LogisticRegression 分类方法对训练集进行训练，训练过程中采用 L2 正则化。然后使用训练好的逻辑回归分类器对测试集进行预测，经过 10 折交叉验证之后得到最终的预测标签值。将预测标签值和真实标签值进行计算，分别得到逻辑回归分类器的准确率、精确率、召回率和F1值。分类结束之后，房产数据的测试集被分成LogicR_low.csv，LogicR_medium.csv 和 LogicR_high.csv 三个不同价值的数据集，依次计算出各个子数据集在总测试集中的权重占比。

4.3.2 基于不同分类数据集的预测实验

(1) 多元线性回归预测

依次读取分类得到的十二个子数据集，数据集的特征维度为 10，输出标签为房屋价格。使用 L2 正则化的过程中，通过 10 折交叉验证来确定 alpha 正则化参数的最优值。使用 sklearn.linear_model 模型对每个数据集进行训练和测试，每个数据集的前 80%用于训练。训练结束之后，将数据集的后 20%用于测试。将预测所得标签值和真实值进行计算，得到多元线性回归模型依次处理十二个不同房产数据集时的各项评价指标(MAE、RMSE、MAPE)，并详细记录实验结果。多元线性回归模型的各项具体参数如表 4.1。

表 4.1 多元线性回归模型参数

参数	中文解释	设定值
Feature_dim	特征维度	10
Label_dim	标签维度	1
random_state	分割随机种子	1
train_size	训练集占比	0.8
alpha	正则化参数	10^{-3}，10^{-2}，10^{-1}，0，10，10^{2}
cv	N 折交叉验证	10

(2) 决策树预测

首先依次读入上文保存的十二个子数据集，按照 8:2 的比例分割数据集之后，使用 DictVectorizer 方法对数据集进行特征提取。对于树的最大深度这一参数，取值范围设置为从 3 到 10，再通过 10 折交叉验证的方式确定其最优值。构建 sklearn.tree.DecisionTreeRegressor 模型对每个子数据集的前 80%进行训练，训练完成之后，使用数据集的后 20%进行测试。将预测所得值和真实值进行计算，得到决策树模型依次处理十二个不同房产数据集时的各项评价指标(MAE、RMSE、MAPE)，并详细记录实验结果。决策树模型的各项具体参数如表 4.2。

表 4.2 决策树模型参数

参数	中文解释	设定值
Feature_dim	特征维度	10
Label_dim	标签维度	1
random_state	分割随机种子	1
train_size	训练集占比	0.8
max_depth	树的最大深度	3，4，5，6，7，8，9，10
cv	N 折交叉验证	10

(3) 随机森林预测

依次读入分类得到的十二个子数据集，数据集的特征维度为 10，输出标签为房屋价格。与决策树模型一样，按照 8:2 的比例对数据集进行分割之后。树的最大深度参数范围设置为从 3 到 10，树的数量参数范围设置为从 10 到 100。构建 sklearn.ensemble.RandomForestRegressor 模型对每个数据集的前 80%进行训练，训练过程中，随机森林会生成多棵决策树，在提高预测效果的同时，也严重增加了实验耗时。为了使两者达到平衡，采用 10 折交叉验证来确定森林中树的数量以及单棵树的最大深度这两个参数的最优值。训练完成之后，选择每个数据集的后 20%进行预测。将预测值和真实值进行计算，得到随机森林模型依次处理十二个不同房产数据集时的各项评价指标(MAE、RMSE、MAPE)，并详细记录实验结果。随

机森林模型的各项具体参数如表 4.3。

表 4. 3 随机森林模型参数表

参数	中文解释	设定值
Feature_dim	特征维度	10
Label_dim	标签维度	1
random_state	分割随机种子	1
train_size	训练集占比	0.8
max_depth	树的最大深度	3，4，5，6，7，8，9，10
n_estimators	树的数量	10，30，50，80，100
cv	N 折交叉验证	10

(4) 长短期神经网络预测

LSTM 模型是基于 Tensorflow 框架的，首先依次读入十二个子数据集，每个数据集中的数据都是 10 维，按照 8:2 的比例分割数据集。选择 Keras 模块进行神经网络实验，神经元节点的个数设置为 64，过拟合参数设置成 0.5。结合 keras.models 与 keras.layers 进行神经网络的搭建，再用 Sequential 方法自定义一个 LSTM 模型，用于训练每个数据集的前 80%。模型共有四层网络，第二层 LSTM 层使用的激活函数为 tanh，其它三层 Dense 层使用的激活函数均为 relu。训练过程中，Batch 中训练样本数量设置为 500，数据训练的轮数为 50,使用 model.compile 方法将优化器设定为 adam，MAE 为损失函数。最后使用训练好的 LSTM 模型对每个数据集的后 20%进行测试,将预测所得值和真实值进行计算，得到长短期神经网络模型依次处理十二个不同房产数据集时的 RMSE 和 MAPE，并详细记录实验结果。长短期神经网络模型的各项具体参数如表 4.4。

表 4. 4 长短期神经网络模型参数表

参数	中文解释	设定值
timesteps	时间窗口长度	50
data_dim	输入数据维度	10
Output_dim	输出数据维度	1
Train_size	训练集占比	0.8
Rnn_units	神经节点数目	64
Dropout	过拟合参数	0.5
Batch_size	Batch 中训练样本数量	500
Epoch	数据训练的轮数	50

(5) 门控循环单元预测

与 LSTM 模型一样，GRU 也是基于 tensorflow 框架的深度学习方法。依次读入十二个子数据集后，按照 8:2 的比例对每个数据集进行分割。选用 Keras 模块中的 keras.models 与 keras.layers 来搭建神经网络，再用 Sequential 方法定义一个

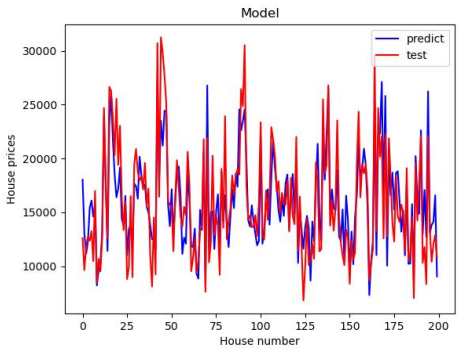
包含 64 个神经元节点且过拟合参数为 0.5 的 GRU 模型。该模型包含三层网络，其中 GRU 层的激活函数为 sigmoid。使用该模型对每个数据集的前 80%进行训练。训练过程中，Batch 中训练样本数量也设置成 500。优化器设定为 adam，损失函数为 MAE。训练结束之后，使用该模型对每个数据集的后 20%进行测试。将预测值和真实值进行计算，得到门控循环单元模型依次处理十二个不同房产数据集时的 RMSE 和 MAPE，并详细记录实验结果。门控循环单元模型的各项具体参数如表 4.5。

表 4. 5 门控循环单元模型参数表

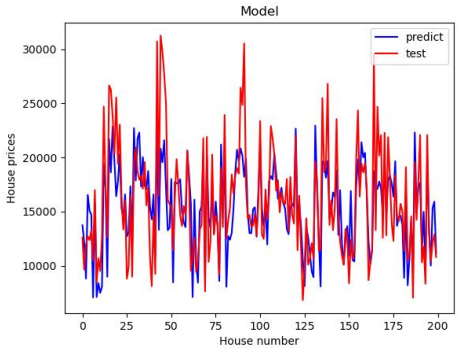
参数	中文解释	设定值
timesteps	时间窗口长度	50
data_dim	输入数据维度	10
Output_dim	输出数据维度	1
Train_size	训练集占比	0.8
Rnn_units	神经节点数目	64
Dropout	过拟合参数	0.5
Batch_size	Batch 中训练样本数量	500

4.3.3 基于未分类数据集的预测实验

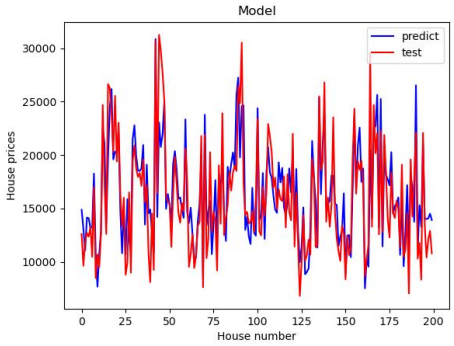
本组实验使用多元线性回归、决策树、随机森林、长短期神经网络和门控循环单元五种回归模型依次对未分类的房产数据集进行处理，数据集的特征维度为 10，输出标签为房屋价格。多元线性回归使用 L2 正则化，决策树和随机森林都采用 DictVectorizer 方法进行特征提取，通过 10 折交叉验证确定正则化参数 alpha、树的最大深度参数 n_estimators 以及树的数量参数 max_depth 三者各自的最优值。长短期神经网络和门控循环单元都是基于 Tensorflow 框架，使用 Sequential 方法进行建模，神经元节点个数设置为 64。数据集的前 80%用于训练，后 20%用于测试。将预测标签值和真实值进行计算，得到五种回归模型的各项评价指标(MAE、RMSE、MAPE)。五种回归模型的预测效果图如图 4-3。



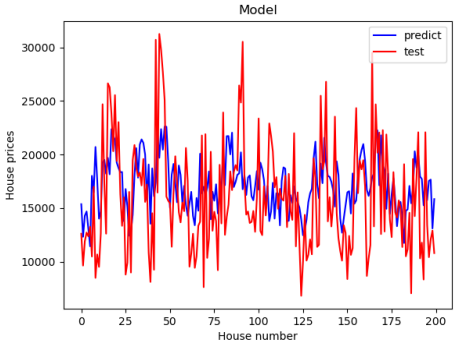
(a) 多元线性回归预测效果图



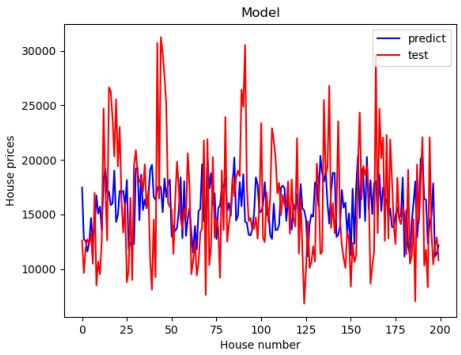
(b) 决策树预测效果图



(c) 随机森林预测效果图



(d) LSTM 预测效果图



(e) 门控循环单元预测效果图

图 4-3 五种模型对未分类数据集的预测效果图

4.4 本章小结

本章介绍了本次研究的所需要的实验环境，确定了具体的实验步骤，并按照实验步骤完成了四种分类实验以及基于分类后所得数据集的所有预测实验，最后完成了各种回归模型处理未分类数据集的实验，对各实验中使用到的模型参数进行了详细说明，并准确详细地记录了各实验的数据结果。

第五章 结果对比分析

5.1 分类实验结果对比

根据每种分类器的混淆矩阵，可以计算出每种分类器对于每种价值类别的预测概率，将上述四种分类器对于各种类别的预测概率汇总如表 5.1。

表 5.1 四种分类器对各类别的预测概率

预测类别 / 真实类别	贝叶斯神经网络			梯度提升决策树			K 近邻			逻辑回归		
	0	1	2	0	1	2	0	1	2	0	1	2
0	85%	14%	1%	85%	14%	1%	82%	17%	1%	87%	12%	1%
1	12%	77%	12%	14%	73%	13%	14%	73%	14%	25%	53%	22%
2	0	12%	88%	0	13%	86%	1%	14%	85%	1%	10%	90%

由表 5.1 可以看出，四种分类器对于价值较低以及价值较高的房屋类别预测更准，对于中等价值房屋类别预测较差一些。中等价值的房屋容易被错误判断为较低价值或者较高价值房屋，较低价值和较高价值的房屋都有一部分被错误判断为中等价值。但是较低价值的房屋基本不会被错误判断为较高价值，较高价值的房屋也很少被判断成较低价值。同时发现，贝叶斯神经网络错误分类的概率更小。

分类实验过程中，根据各种分类器的分类结果计算各种分类器对于每种价值类别的预测概率时，遵循四舍五入原则对计算结果的小数点后两位进行取舍。实验数据集较大时存在一定的精度损失，因此每种真实类别对应的三种预测类别概率相加后的结果，并不绝对等于 1。分类实验结束之后，四种分类器的各项分类性能评价指标汇总如表 5.2，四种分类器的分类性能对比如图 5-1。

表 5.2 四种分类器分类性能评价指标

评价指标 / 分类器	准确率	精确率	召回率	F1 值
贝叶斯神经网络	83.22%	82.67%	82.45%	82.74%
梯度提升决策树	81.86%	81.70%	81.67%	81.68%
K 近邻	80.32%	80.03%	79.83%	79.93%
逻辑回归	77.19%	76.11%	76.6%	76.35%

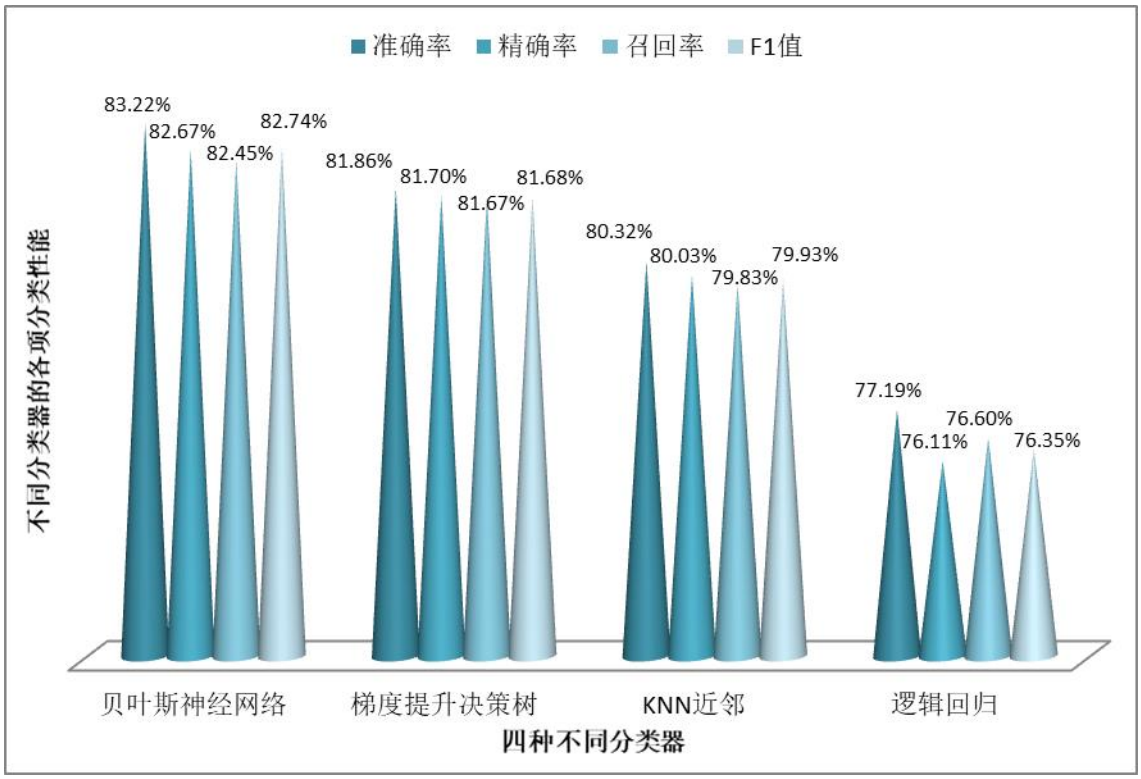


图 5-1 四种分类器的分类性能对比图

从图 5-1 可以明显看出，贝叶斯神经网络分类方法处理房产数据集时，得到的各项评价指标(准确率、精确率、召回率、F1 值)相较于其它三种分类器更优异，分类性能更出色，后续各种预测实验中，重点研究使用五种回归模型依次处理基于贝叶斯神经网络分类器的三种价值数据集时，得到的各项评价指标。贝叶斯神经网络的分类准确率达到 83.22%，精确率为 82.67%，召回率为 82.45%，F1 值为 82.74%。梯度提升决策树和 K 近邻两种分类器的各项评价指标(准确率、精确率、召回率、F1 值)稍小于贝叶斯神经网络，分类性能稍差一些，但分类准确率均超过八成，而逻辑回归分类器的分类性能较差。

5.2 预测实验结果对比

5.2.1 各种回归模型处理未分类数据集的实验结果对比

使用五种回归模型依次对未分类的房产数据集进行预测时，得到的各项评价指标(MAE、RMSE、MAPE)如表 5.3，五种模型的各项评价指标对比如图 5-2。

表 5. 3 基于未分类数据集的预测实验各项评价指标

实验指标 / 回归模型	MAE	RMSE	MAPE(%)
多元线性回归	2078.08	2705.53	14.71
决策树	2485.50	3475.16	15.93
随机森林	2299.46	3176.27	15.20
长短期神经网络	2895.36	3776.87	18.65
门控循环单元	2908.62	3849.13	18.29

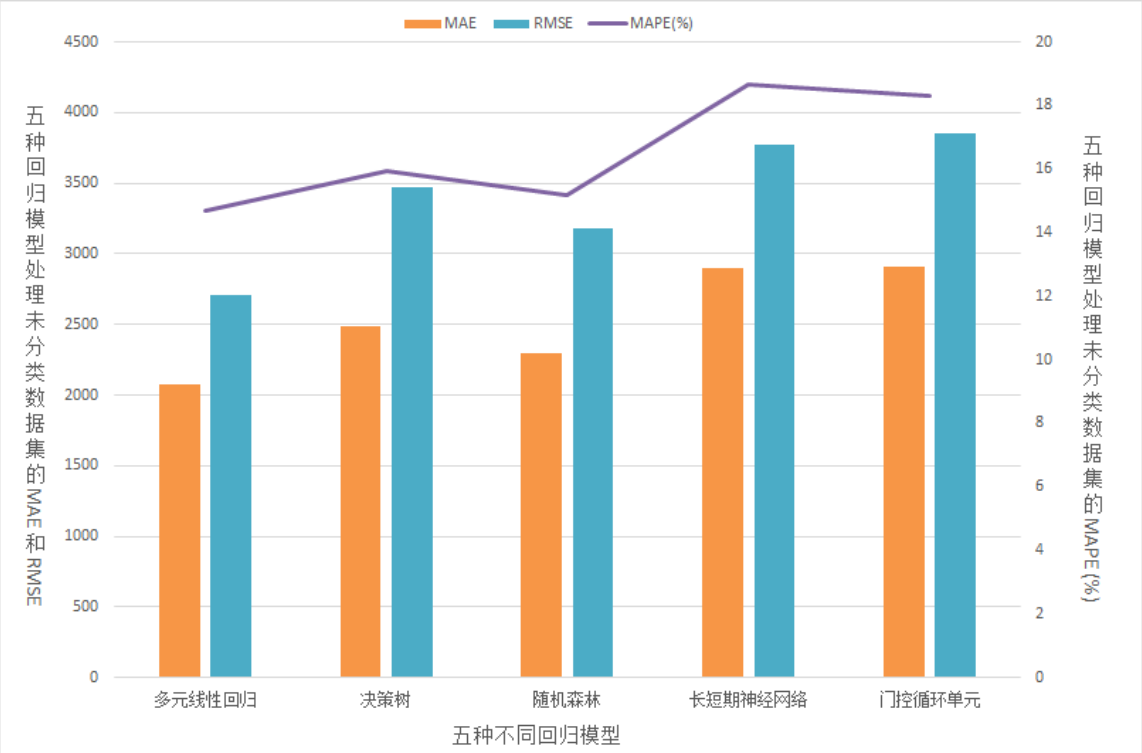


图 5-2 五种回归模型处理未分类数据集的各项评价指标

从图 5-2 中可以看出，对未分类的数据集进行价格预测时，多元线性回归模型的各项误差较小，效果最好；随机森林模型的实验误差稍大一些，效果次之，决策树模型的误差更大一些，长短期神经网络与门控循环单元模型的实验误差比较大，效果较差一些。

5.2.2 各种回归模型处理 BNN 分出的三种数据集的实验结果对比

基于贝叶斯神经网络的分类实验结束之后，测试数据样本集被分成了较低、

中等和较高三种不同价值的数据集。使用五种回归模型依次对这三种数据集进行处理时，得到的各项评价指标(MAE、RMSE、MAPE)具体数值如表 5.4。

表 5. 4 五种回归模型处理贝叶斯神经网络分出的三种数据集时的各项评价指标

价值类型和回归模型 \ 实验指标		MAE	RMSE	MAPE(%)
较低价值	多元线性回归	1561.09	1939.89	12.81
	决策树	1805.13	2182.49	15.65
	随机森林	1105.83	1537.96	10.24
	长短期神经网络	2009.72	2406.34	17.72
	门控循环单元	2216.94	2672.87	18.81
中等价值	多元线性回归	1532.39	1936.12	9.06
	决策树	1633.18	2001.79	10.35
	随机森林	1021.13	1238.35	6.19
	长短期神经网络	1745.75	2031.90	11.01
	门控循环单元	1810.71	2153.23	12.08
较高价值	多元线性回归	1946.29	2565.71	8.61
	决策树	2903.08	4125.88	12.60
	随机森林	2081.21	3601.91	8.93
	长短期神经网络	2513.23	3411.23	11.54
	门控循环单元	2849.22	3887.17	13.49

根据表 5.4 可以看出，五种回归模型依次处理贝叶斯神经网络分类器分出的三种价值的数据集时，较低价值以及中等价值的数据集相比较高价值的数据集，MAE 和 RMSE 值更小；中等价值和较高价值的数据集相比较低价值的数据集，MAPE 值更小一些。综合比较 MAE、RMSE、MAPE 三种评价指标后，发现五种回归模型在处理中等价值的数据集时，预测效果相对更好。

5.2.3 各种回归模型处理不同分类器分出数据集的实验结果对比

每种回归模型在处理贝叶斯神经网络分类器分出的三种价值数据集时，都可以得到三组评价指标(MAE、RMSE、MAPE)，结合每种价值数据集在测试集中的权重占比，将三个 MAE、三个 RMSE、三个 MAPE 分别进行加权平均计算，所得结果分别作为该种回归模型处理贝叶斯神经网络分类器分出数据集时的 MAE、RMSE 和 MAPE 值。根据上述方法，还可以依次求出五种回归模型处理其它三种

分类器分出数据集时的各项评价指标(MAE、RMSE、MAPE)。将五种模型依次处理四种分类器分出数据集时的各项评价指标(MAE、RMSE、MAPE)汇总如表 5.5。

表 5. 5 五种模型依次处理四种分类器分出数据集时的各项评价指标

分类器和回归模型 \ 评价指标		MAE	RMSE	MAPE(%)
贝	多元线性回归	1682.12	2150.62	10.15
叶	决策树	2120.45	2781.28	12.87
斯	随机森林	1408.33	2138.48	8.46
神	长短期神经网络	2093.44	2623.54	13.42
经网络	门控循环单元	2297.48	2913.24	14.79
梯	多元线性回归	1769.10	2161.94	11.41
度	决策树	2196.08	2866.61	13.36
提	随机森林	1510.24	2217.72	9.05
升	长短期神经网络	2104.86	2635.58	13.60
决策树	门控循环单元	2134.31	2703.78	13.69
K	多元线性回归	1823.91	2243.58	11.75
近	决策树	2105.71	2756.31	12.73
邻	随机森林	1704.16	2364.19	10.41
	长短期神经网络	2118.60	2708.58	13.53
	门控循环单元	2050.78	2698.03	13.23
逻	多元线性回归	1809.56	2242.84	11.30
辑	决策树	2240.19	2903.36	13.54
回	随机森林	1568.93	2216.11	9.74
归	长短期神经网络	2077.89	2611.40	13.26
	门控循环单元	2136.21	2842.90	13.83

结合表 5.5，可以大致看出，随机森林和多元线性回归两种模型处理不同分类器对应数据集时，得到的各项评价指标(MAE、RMSE、MAPE)相对较小。接下来先将每一种回归模型依次处理不同分类器分出的数据集时所得各项评价指标相互对比，再将每一种模型依次处理不同分类器分出的数据集时得到的各项评价指标与该模型处理未分类数据集时的对应指标进行对比。五种回归模型依次处理不同分类器分出的各数据集以及未分类的数据集时，MAE 值对比如图 5-3，RMSE 值对比如图 5-4，MAPE 值对比如图 5-5。

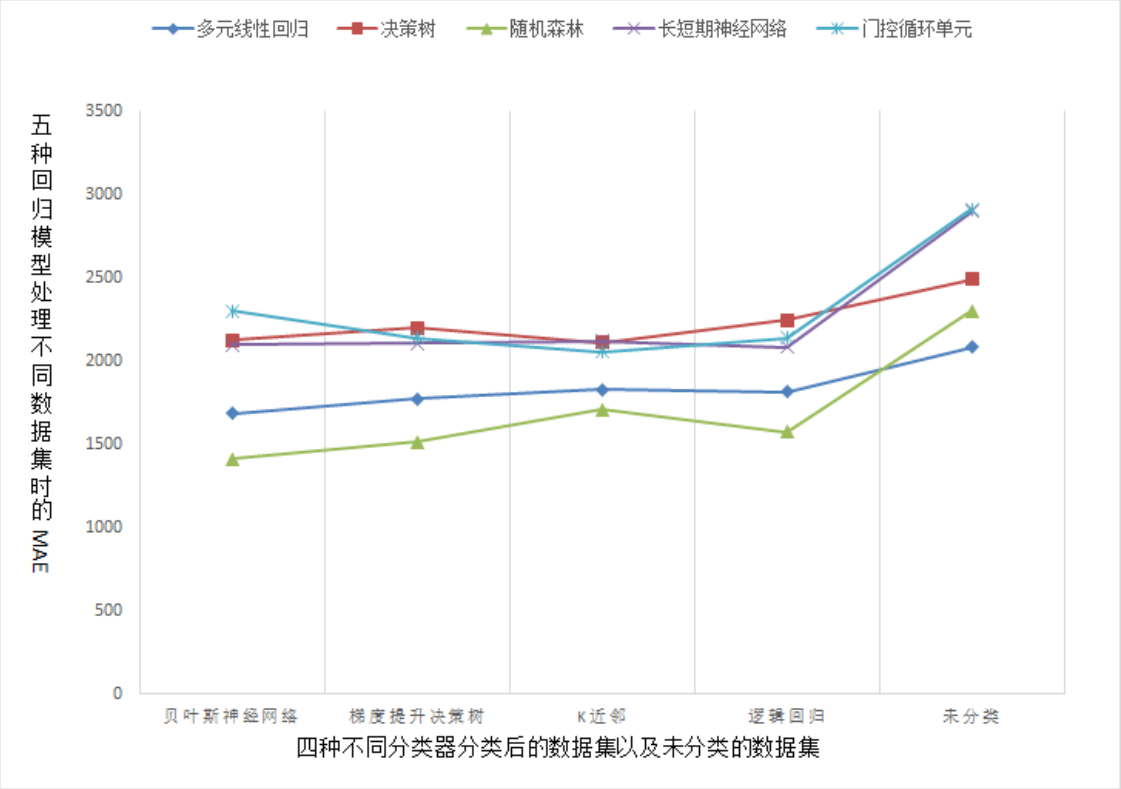


图 5-3 五种模型处理各种数据集的 MAE

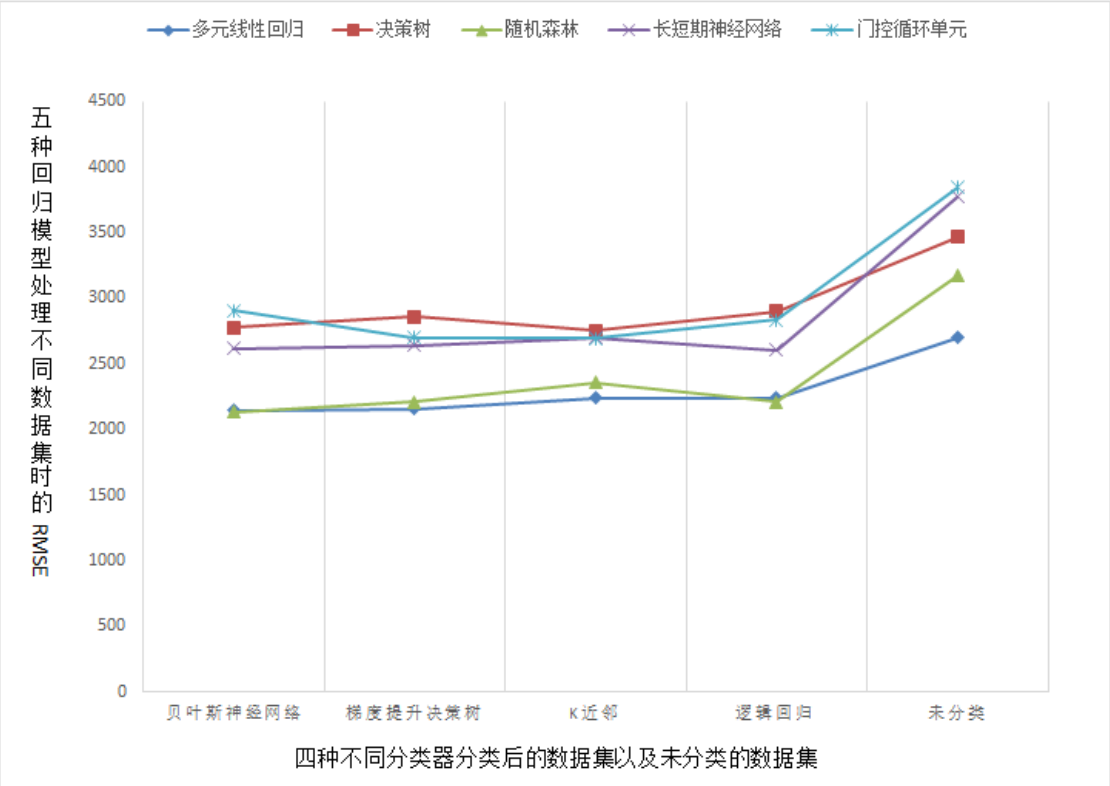


图 5-4 五种模型处理各种数据集的 RMSE

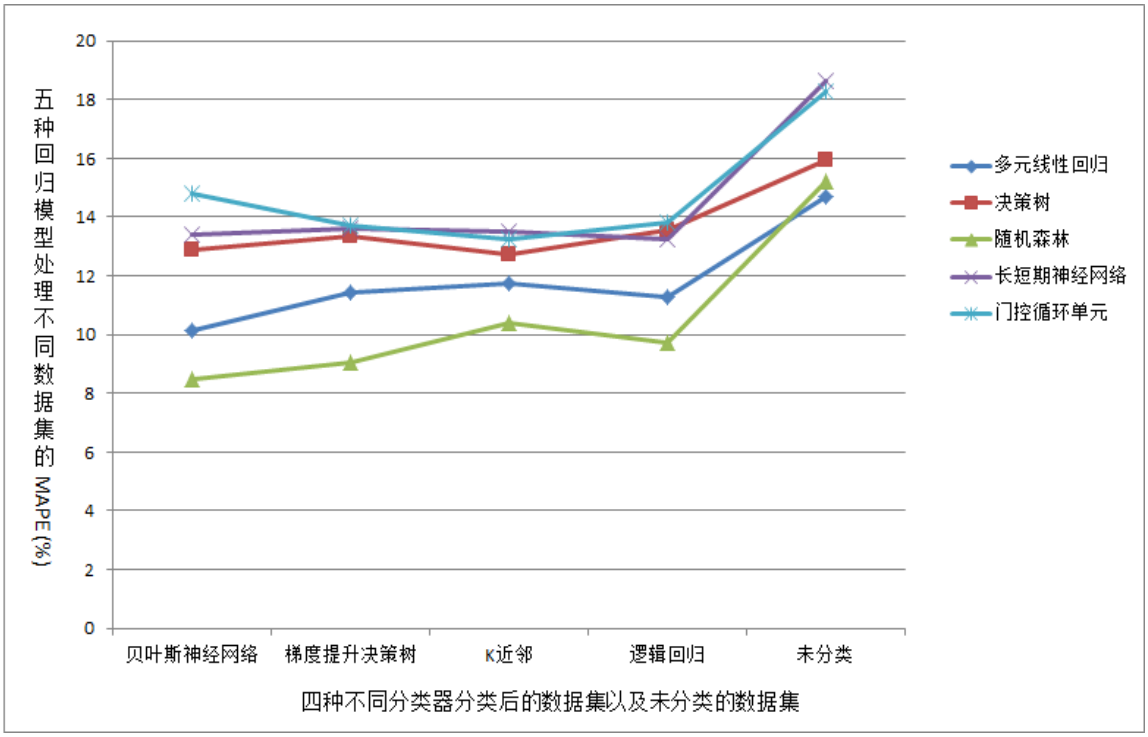


图 5-5 五种模型处理不同数据集的 MAPE (%)

结合图 5-3、图 5-4 和图 5-5 可以看出，使用每种回归模型依次处理不同分类器分出的数据集时，各项评价指标(MAE、RMSE、MAPE)均小于直接使用该种回归模型处理未分类数据集时的对应评价指标。通过比较每种模型在处理不同分类器分出的各数据集时的各项评价指标，可以得出多元线性回归和随机森林两种模型在处理贝叶斯神经网络分出的数据集时各项评价指标最小，效果最佳；决策树和门控循环单元两种模型在处理 K 近邻分出的数据集时各项评价指标最小，效果最佳；长短期神经网络模型在处理逻辑回归分出的数据集时各项评价指标最小。每种回归模型的各项评价指标最小时，相比较该种模型处理未分类数据集时，各项评价指标(MAE、RMSE、MAPE)具体减少值如表 5.6，各减少值对比如图 5-6。

表 5. 6 五种模型各项评价指标减少值

实验指标 / 模型	MAE	RMSE	MAPE(%)
多元线性回归	395.96	554.91	4.56
决策树	379.79	718.85	3.20
随机森林	891.13	1037.79	6.74
长短期神经网络	817.47	1165.47	5.39
门控循环单元	857.84	1151.07	5.06

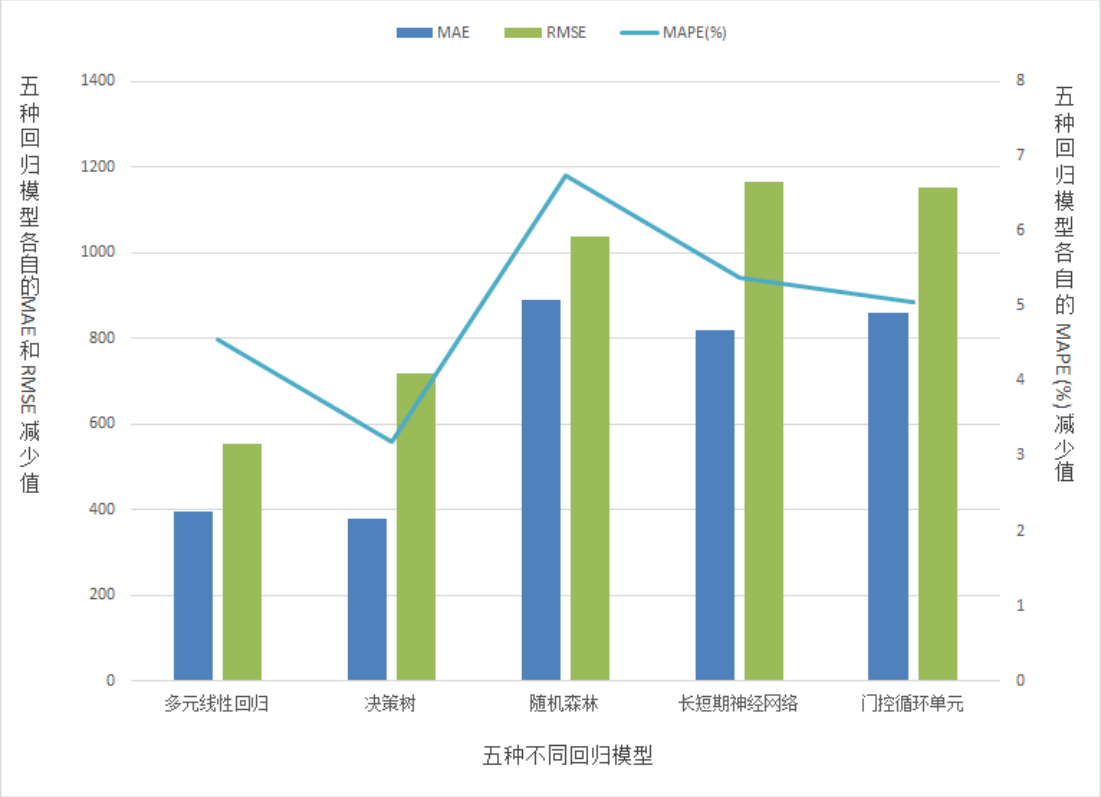


图 5-6 五种回归模型各项评价指标减少值

从图 5-6 中可以看出，随机森林模型的 MAE 值和 MAPE 值减少最多，长短期神经网络模型的 RMSE 值减少最多，且五种模型的三种评价指标(MAE、RMSE、MAPE)都出现了一定程度的减少，表明了分类实验中对于房产数据集的价值分类处理，可以有效减少回归模型对数据集进行处理时的各项误差指标。

5.2.4 各种回归模型的最佳实验结果对比

使用五种回归模型各自取得最佳预测效果时，各种回归模型的各项评价指标(MAE、RMSE、MAPE)如表 5.7，三种评价指标的对比如图 5-7。

表 5. 7 五种回归模型各自预测效果最佳时的各项评价指标

实验指标 / 模型	MAE	RMSE	MAPE(%)
多元线性回归	1682.12	2150.62	10.15
决策树	2105.71	2756.31	12.73
随机森林	1408.33	2138.48	8.46
长短期神经网络	2077.89	2611.40	13.26
门控循环单元	2050.78	2698.03	13.23

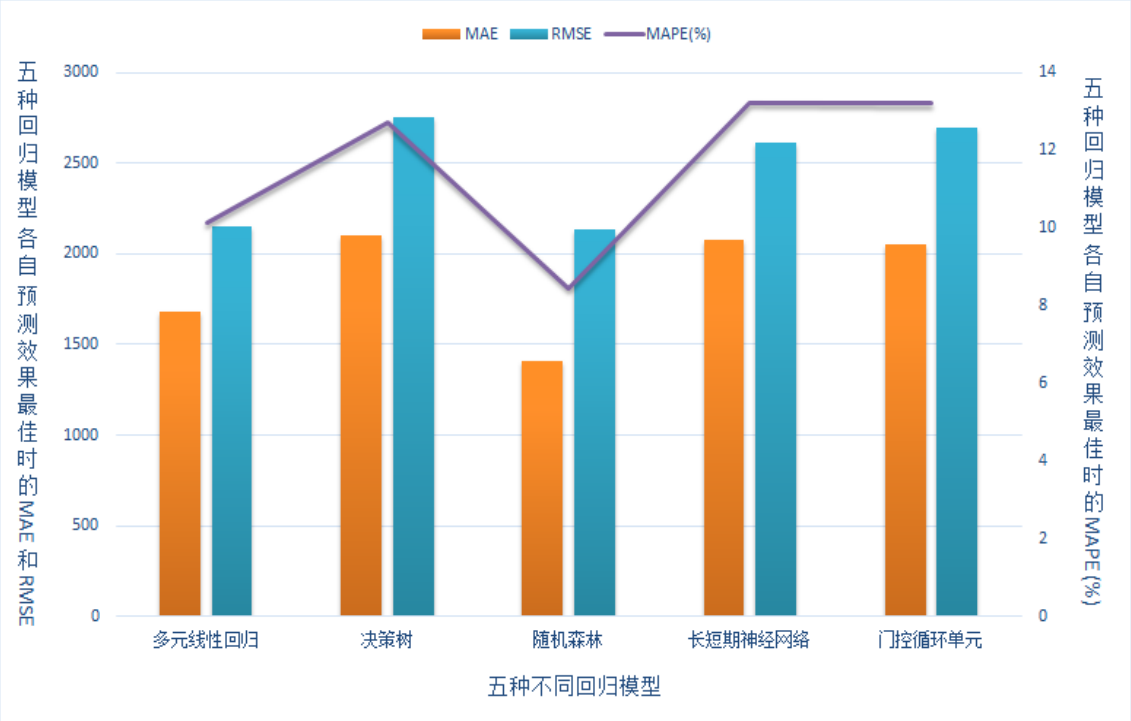


图 5-7 五种模型预测效果最佳时的各项评价指标对比图

根据图 5-7 可以看出，在各种回归模型各自效果最佳的情况下，随机森林的误差值最小，预测效果最好，多元线性回归模型效果次之，此时两种模型处理的数据集都是经过贝叶斯神经网络分类后得到的，说明贝叶斯神经网络对于数据集的分类处理确实有效提升了回归模型的预测精准度。决策树、长短期神经网络以及门控循环单元三种模型的实验效果更差一些，且三组实验的各项误差指标相近。此时，决策树和门控循环单元处理的数据集都是经过 K 近邻分类后得到的，而长短期神经网络处理的数据集是经过逻辑回归分类后得到的。

综上所述，贝叶斯神经网络和随机森林的组合模型预测效果最佳，三种误差指标(MAE、RMSE、MAPE)分别为 1408.33 元、2138.48 元、8.46%；相比较直接使用多元线性回归模型(处理未分类数据集时预测效果最佳)进行预测，实验误差 MAE 减少了 669.75 元，RMSE 减少了 567.05 元，MAPE 减少了 6.25%，预测效果显著提升。

5.3 本章小结

本章着重对上述各组实验结果进行对比分析，首先从分类实验的结果进行对比分析，得出贝叶斯神经网络分类器的分类性能最优，然后再着重从预测实验的结果进行对比分析，首先对比分析了五种模型处理未分类数据集的实验误差，得

出多元线性回归的预测效果最好。其次仔细对比了五种回归模型依次处理贝叶斯神经网络分出的三种不同数据集时的预测效果，发现各回归模型处理中等价值的数据集时，预测效果相对较好。接着对比分析了每一种模型处理各种分类器分类后的数据集时的各项评价指标，进而得出各种模型分别处理各种分类器分类后数据集的最佳效果，并将各种模型获得最佳预测效果时的各项误差指标与对应模型处理未分类数据集时的各项误差指标相比较，发现前者预测效果均好于后者，进一步肯定了本次实验的可研究性。最后通过对比各种回归模型获得最佳预测效果时的各自误差指标，得出贝叶斯神经网络和随机森林的组合模型预测效果最佳，贝叶斯神经网络和多元线性回归的组合模型预测效果次之。两种组合模型相比较于直接使用线性回归模型(处理未分类数据集时预测效果最佳)，预测效果均有显著提升，充分肯定了贝叶斯神经网络算法良好的分类性能对于预测效果提升的重要性。原因是贝叶斯神经网络算法根据房产数据集中影响较大的因素来确定其概率分布的参数，并将这些参数作为神经网络中的权重，大大增加了房屋价值分类的准确率，从而有效减少了房价预测实验中的各项误差。

第六章 总结与展望

6.1 总结

本文从专业学术网站 Kaggle 上获取真实房产交易数据作为实验数据集，首先对数据集进行数据清洗，删除了原始数据集中的重复样本、无效值以及缺失率比较高的特征属性，对部分缺失值进行合理填充。接着，根据房屋的最近交易价格将数据样本划分成较低价值、中等价值和较高价值三种类别，再根据房屋面积与成交量关系以及房屋楼层与成交价和成交量关系，对数据开展初步分析工作。分析结束之后，使用随机森林分类器对数据各项特征进行重要性分析，去除掉对房屋价值影响较小的一些特征属性，保留对房价影响最大的前十项特征，从而达到数据降维的目的。

数据处理完成之后，对本次研究进行总体方案设计。首先使用贝叶斯神经网络、梯度提升决策树、K 近邻以及逻辑回归四种分类器依次对数据集进行价值分类，共得到 12 个不同的数据集。每种分类器分出的每种价值的房产数据集，都使用多元线性回归、决策树、随机森林、长短期神经网络以及门控循环单元五种回归模型进行房价预测。每一种回归模型处理每一种分类器分出的三种价值数据集时，都可以得到三组评价指标(MAE、RMSE、MAPE)。结合每种价值的数据集在测试集中的权重占比，将三个 MAE、三个 RMSE、三个 MAPE 分别进行加权平均计算，所得结果分别作为该种回归模型处理该分类器分出数据集时的 MAE、RMSE、MAPE。接着，使用多元线性回归、决策树、随机森林、长短期神经网络以及门控循环单元五种回归模型依次处理未分类数据集。所有预测实验结束之后，得到五种回归模型依次处理不同分类器分出数据集以及未分类数据集时的评价指标(MAE、RMSE、MAPE)。

汇总所有实验指标，开展对比分析工作。首先对比五种模型依次处理未分类数据集时的三种评价指标，发现多元线性回归模型预测效果整体最好。接着仔细对比了五种回归模型依次处理贝叶斯神经网络分出的三种不同数据集时的预测效果，发现五种模型在处理中等价值的数据集时，预测效果相对较好。然后将每种模型处理每种分类器分出数据集时的误差指标分别与该种模型处理未分类数据集时的误差指标进行对比，发现前者预测效果均好于后者。再通过对比每种模型依次处理不同分类器分出数据集时得到的三种误差指标，得出五种模型在处理不同分类器分出数据集时各自的最佳效果。最后通过对比不同模型取得最佳效果时的

各项误差指标，发现贝叶斯神经网络与随机森林的组合模型预测效果最佳，贝叶斯神经网络与多元线性回归的组合模型预测效果次之。两种组合模型相对于直接使用线性回归模型(处理未分类数据集时预测效果最佳)进行预测，效果均有显著提升，充分肯定了贝叶斯神经网络算法良好的分类性能对于预测效果提升的重要性。同时发现，虽然 K 近邻分类器的分类准确率较高，但是在使用回归模型对其分类后的数据集进行预测时效果却不理想。

本文的主要创新点：

(1) 将分类器与回归算法结合形成新的组合模型，并将其运用到房产价格预测领域，先前研究基本都是直接使用机器学习模型进行预测，少量学者使用聚类加回归的预测模式。本次研究使用的分类加回归的组合模型相对于直接使用回归模型，大幅减少了实验误差，有效提升了房价预测精度；

(2) 使用贝叶斯神经网络算法对房产数据集进行价值分类，良好的分类性能显著降低了房价预测实验中的各项误差指标，对于预测效果的提升有重要作用。

6.2 展望

本文运用贝叶斯神经网络、梯度提升决策树、K 近邻和逻辑回归四种分类器，将房产数据根据房产价值划分为三类，再结合多种回归算法对分类后的房产数据开展房价预测工作。通过对比分析得出，贝叶斯神经网络加随机森林的组合模式相较与其它分类器与不同模型的组合在房价预测上效果处于最佳。在仔细深入研究思考之后，发现本文也有一些有待提升之处，具体如下：

(1) 可以继续增加实验数据样本，根据一二三线城市不同的房产价位对房产数据进行更多类别的划分。

(2) 本次研究中分类实验部分对于数据集的价值分类属于有监督学习，可以在本文的基础上再使用基于无监督学习的聚类方法添加一组分类实验，所得不同价值房屋数据集再分别使用各种回归模型进行价格预测，通过对比分析各组预测实验的结果，得出具有研究价值的结论。

参考文献

[1] 李仲飞, 张浩. 成本推动、需求拉动——什么推动了中国房价上涨?[J]. 中国管理科学, 2015, 23(05): 143-150.

[2] Hsu L. How Did COVID-19 Affect the Chinese Real Estate Stock Market: An Empirical Research[J]. Financial Forum, 2020, 9(4): 244-249.

[3] 郭 艺. Take SARS as an Example to See the Effect of COVID-19 on Foreign Trade[J]. Emergence and Transfer of Wealth, 2020, 10(1): 1-7.

[4] 高凌云. 内需压力、经济规模与中国出口的可持续增长[J]. 经济与管理评论, 2018, 34(01): 31-44.

[5] 苗燕. 房地产经济对中国国民经济增长的作用[J]. 中外企业家, 2017(30): 236-237.

[6] 丁飞, 江铭炎. 基于改进狮群算法和 BP 神经网络模型的房价预测[J]. 山东大学学报(工学版), 2021, 51(04): 8-16.

[7] 申瑞娜, 曹昶, 樊重俊. 基于主成分分析的支持向量机模型对上海房价的预测研究[J]. 数学的实践与认识, 2013, 43(23): 11-16.

[8] 高玉明, 张仁津. 基于遗传算法和 BP 神经网络的房价预测分析[J]. 计算机工程, 2014, 40(04): 187-191.

[9] 麻顺顺. 基于 LSTM 的二手房价格预测模型研究及应用[D]. 郑州大学, 2020.

[10] 张智鹏, 郑大庆. 影响区域房价的客观因素挖掘分析[J]. 计算机应用与软件, 2019, 36(11): 32-38+85.

[11] 罗博炜, 洪智勇, 王劲屹. 多元线性回归统计模型在房价预测中的应用[J]. 计算机时代, 2020(06): 51-54.

[12] 郑永坤, 刘春. 基于 ARIMA 模型的二手房价格预测[J]. 计算机与现代化, 2018(04): 122-126.

[13] Khamis A B. Comparative Study On Estimate House Price Using Statistical And Neural Network Model[J], 2014, 3(12): 126-131.

[14] Yang B, Cao B. Research on Ensemble Learning-based Housing Price Prediction Model[J], 2018: 1-8.

[15] Wang J J, Hu S G, Zhan X T, et al. Handwritten-Digit Recognition by Hybrid Convolutional Neural Network based on HFO_2 Memristive Spiking-Neuron[J], 2018: 1-7.

[16] Hao L. A House Price Prediction Model Based on Deep Belief Network[J]. Tianjin Science & Technology, 2018, 45(10): 79-81.

[17] Serrano W. The random neural network in price predictions[J]. Neural Computing and Applications, 2021: 855-873.

[18] Kostic Z, Jevremovic A. What Image Features Boost Housing Market Predictions?[J], 2021, 22(7): 1-14.

[19] Zhao S, Li W, Zhao K, et al. Change Characteristics and Multilevel Influencing Factors of Real Estate Inventory—Case Studies from 35 Key Cities in China[J]. Land, 2021, 10(9): 1-29.

[20] 张晶. 苏州商品住房价格影响因素分析及房价预测[D]. 苏州大学, 2014.

[21] Shi D, Guan J, Zurada J, et al. An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction[J]. Journal of International Technology and Information Management 1543-5962, 2015, 24(1): 15-32.

[22] Roberto T, Davide B, Matteo C, et al. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers[J]. Journal of chemical information and modeling, 2015, 55(11): 1-34.

[23] Shi D, Zurada J, Guan J. A Bayesian Network Approach to Classifying Bad Debt in Hospitals[M]. 2016.

[24] Marcot B G, Penman T D. Advances in Bayesian network modelling: Integration of modelling technologies[J]. ENVIRONMENTAL MODELLING & SOFTWARE, 2019, 111: 386-393.

[25] 张岩. 基于梯度提升决策树分类器的进化计算动态性能研究[J]. 信息系统工程, 2020(03): 157-159.

[26] Ezhilraman S V, Srinivasan S, Suseendran G. Gaussian Light Gradient Boost Ensemble Decision Tree Classifier for Breast Cancer Detection[M]. Intelligent Computing and Innovation on Data Science, 2020.

[27] Long Y, Tang W, Yang B, et al. GTK: A Hybrid-Search Algorithm of Top-Rank-k Frequent Patterns Based on Greedy Strategy[J]. CMC-COMPUTERS MATERIALS & CONTINUA, 2020, 63(3): 1445-1469.

[28] Hassanat A B, Abbadi M A, Altarawneh G A, et al. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach[J]. Computer Science, 2014, 12(8): 33-39.

[29] Zhang S, Li X, Zong M, et al. Efficient kNN classification with different numbers of nearest

neighbors[J]. IEEE transactions on neural networks and learning systems, 2017, 29(5): 1774-1785.

[30] Puoya T, Ivan D, Martin V. Kinetic Euclidean Distance Matrices[J]. IEEE Transactions on Signal Processing, 2020, 68.

[31] Basheer S, Mathew R M, Devi M S. Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019, 8(12).

[32] 毛林, 陆全华, 程涛. 基于高维数据的集成逻辑回归分类算法的研究与应用[J]. 科技通报, 2013, 29(12): 64-66.

[33] Jin X B, Yu J W, Wang G C, et al. On Optimization of Multi-Class Logistic Regression Classifier[J]. Advanced Materials Research, 2013, 2385(694-697): 2746-2750.

[34] Qiang G, Zhe T, Yan D, et al. An improved office building cooling load prediction model based on multivariable linear regression[J]. Energy and Buildings, 2015, 107: 445-455.

[35] Tsoukalas V, Fragiadakis N. Prediction of occupational risk in the shipbuilding industry using multivariable linear regression and genetic algorithm analysis[J]. Safety Science, 2016, 83: 12-22.

[36] 岳根霞, 王剑, 刘金花. 决策树算法在诊断机械故障信息挖掘中的应用[J]. 机械设计与制造, 2022(01): 168-171+176.

[37] Rana A, Bhagat N K, Jadaun G P, et al. Predicting blast-induced ground vibrations in some Indian tunnels using decision tree[J]. Mining Engineering, 2020, 72(8): 1039-1053.

[38] Abdulsattar S Y, Hossam A R. Improved Random Forest Algorithm Performance For Big Data[J]. Journal of Physics: Conference Series, 2021, 1897(1): 1-13.

[39] 闫政旭, 秦超, 宋刚. 基于 Pearson 特征选择的随机森林模型股票价格预测[J]. 计算机工程与应用, 2021, 57(15): 286-296.

[40] Dan L, Zhi L. Gold Price Forecasting and Related Influence Factors Analysis Based on Random Forest[J]. Springer Singapore, 2017: 711-723.

[41] Sadeghi-Mobarakeh A, Kohansal M, Papalexakis E E, et al. Data Mining based on Random Forest Model to Predict the California ISO Day-ahead Market Prices[C]. Innovative Smart Grid Technologies (ISGT), 2017.

[42] 吕红燕, 冯倩. 随机森林算法研究综述[J]. 河北省科学院学报, 2019, 36(03): 37-41.

[43] Hansu K, Hee L T. A robust elastic net via bootstrap method under sampling uncertainty for significance analysis of high-dimensional design problems[J]. Knowledge-Based Systems, 2021, 225: 1-13.

[44] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[C]. Interspeech, 2012.

[45] 姚远, 张朝阳. 基于 HP-LSTM 模型的股指价格预测方法[J]. 计算机工程与应用, 2021, 57(24): 296-304.

[46] 张宁, 方靖雯, 赵雨宣. 基于 LSTM 混合模型的比特币价格预测[J]. 计算机科学, 2021, 48(11A): 39-45.

[47] Uwaisu A, Yahaya A S, Kamal S M, et al. A Hybrid Deep Stacked LSTM and GRU for Water Price Prediction[C]. 2nd International Conference on Computer and Information Sciences (ICCIS), 2020.

[48] Kong D, Liu S, Pan L. Amazon Spot Instance Price Prediction with GRU Network[C]. 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2021.

[49] Liu Y, Wang Z, Zheng B. Application of Regularized GRU-LSTM Model in Stock Price Prediction[C]. 2019 IEEE 5th International Conference on Computer and Communications (ICCC), 2020.

[50] 牛红丽, 赵亚枝. 利用 Bagging 算法和 GRU 模型预测股票价格指数[J]. 计算机工程与应用, 2021: 1-9.

[51] Klingerman S. Top Five Reasons to Use an Oracle Database[J]. Database Trends and Applications, 2020, 34(6): 34-35.

[52] 叶鸥, 张璟, 李军怀. 中文数据清洗研究综述[J]. 计算机工程与应用, 2012, 48(14): 121-129.

[53] P. C C. The Importance of Data Cleaning: Three Visualization Examples[J]. CHANCE, 2020, 33(1): 4-9.

[54] Sorokin A A, Dagaev A, Borodyansky I. Comparative Analysis of Missing Data Recovery Methods[J]. Izvestiya SFedU. Engineering Sciences, 2020: 93-107.

[55] Ying F, Zhang Z. Data Visualization Analysis of Big Data Recruitment Positions in Hangzhou Based on Python[J]. Review of Computer Engineering Studies, 2019, 6(4): 81-86.

[56] 孙平军, 冷红. 中国高层住宅楼层选择偏好研究——以石家庄市为例[J]. 哈尔滨工业大学学报(社会科学版), 2015, 17(06): 109-117.

[57] Reddy G T, Reddy M P K, Lakshmanna K, et al. Analysis of dimensionality reduction techniques on big data[J]. IEEE Access, 2020, 8: 54776-54788.

[58] Ramadhan A, Susetyo B. Classification Modelling of Random Forest to Identify the Important Factors in Improving the Quality of Education[J]. International Journal on Advanced Science, Engineering and Information Technology, 2021, 11(2): 501-507.

[59] 沈健, 蒋芸, 邹丽, et al. 基于节点选择优化的 DAG-SVM 多类别分类[J]. 计算机工程, 2015, 41(06): 143-146.

[60] Yu Y, Xiong Z Y, Xiong Y S, et al. Improved Logistic Regression Algorithm Based on Kernel Density Estimation for Multi-Classification with Non-Equilibrium Samples[J]. CMC-COMPUTERS MATERIALS & CONTINUA 2019, 61(1): 103-117.

致谢

三年的研究生学习生涯，匆匆而逝。带着自己积累的知识和经验，我即将离开校园成为一名职场人。离别之际，我想衷心地向我敬爱的导师、母校、父母亲人以及同学们致以真挚的谢意。

首先感谢我的导师史东辉教授，史老师是一位热爱学术、治学严谨、关心学生的人。这篇论文的完成离不开史老师的指导。史老师是我学术路上的指明灯，从选题到定稿，从内容到格式，老师都倾注了大量的心血，为我提供研究思路和方法，不断向我推送领域的前沿资料，每次上交论文后老师都会认真批注并监督修改，为我论文的完成保驾护航。除此之外，在读研的三年时间中，无论在学习上还是在生活上，史老师都给予我非常多的关心和帮助，让我受益良多，桃李不言，下自成蹊，跟随老师的步伐，我学到了对待学术的严谨态度以及对于新技术的敏锐洞察力，这些将会对我日后的工作产生积极的影响。史老师，谢谢您！

感谢陪伴我三年的母校——安徽建筑大学，感谢当初录取我，给了我一个学习和成长的机会。母校拥有宽阔的学习平台和浓厚的学习氛围，我在这里得到了不断地历练和全面地提升，母校的恩情似海深，我终身难忘！

感谢父母这么多年来无微不至的照顾，在自己当初选择读研时，给予我无限的理解与支持，让我能够毫无顾虑地从事学术科研。同时感谢我的女朋友，是你的陪伴，让我每一步走得都更坚实有力，是你的激励，让我不断努力，不断进步，成为了更好的自己。

感谢我的室友们，在同一个屋檐下的三年，我们一起创造了很多美好回忆。每当我遇到挫折时，是你们给了我继续前行的勇气，谢谢兄弟们！最后要感谢我的同门和师弟师妹们，是你们让我感受到了如同家一般的温暖，在这样的小集体中我们互相扶持，共赴难关，共同奔向更璀璨的明天。

作者简介及读研期间主要科研成果

李西洋，男，1997 年 9 月 23 日出生于安徽省铜陵市；2019 年 6 月在徐州工程学院获得计算机科学与技术专业学士学位；2019 年 9 月至今在安徽建筑大学攻读计算机技术硕士学位，主要研究数据挖掘，大数据方向。

参与项目与竞赛：

（1）参加安徽建筑大学校级科研项目本体驱动下基于概念的建筑安全事故风险评估方法研究（2019）

（2）参加建筑安全事故本体建立方法研究（2021）

（3）获得第七届中国国际‘2021 互联网+’大学生创新创业大赛校级三等奖

软件著作权：

基于大数据分析的智能旅游攻略定制系统 V1.0，证书号：软著登字第 7510163 号