|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| C:\Users\Administrator\Desktop\78D172AE6E29983F939CC0FCDA3_12AD0ED6_2FE0.gif78D172AE6E29983F939CC0FCDA3_12AD0ED6_2FE0 | | | |
| 学士学位论文 | | | |
| 美国金县二手房价的分析与预测 | | | |
|  | | | |
|  | 姓 名 | 陈博文 |  |
|  | 学 号 | 18854004 |  |
|  | 学 院 | 计算机与信息技术学院 |  |
|  | 专 业 | 电子商务 |  |
|  | 指导教师 | 周子程 |  |
|  | 职 称 | 讲师 |  |
|  | | | |
| 二零二二 年 六 月 一 日 | | | |

学位论文原创性声明

本人所提交的学位论文《美国金县二手房价的分析与预测》，是在导师的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

论文作者（签名）： C:\Users\86157\AppData\Local\Temp\WeChat Files\341298396258774299.jpg 指导教师确认（签名）：

2022年 6 月 1 日 2022 年 6 月 1 日

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学海滨学院有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅。本人授权北京交通大学海滨学院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、缩印或其它复制手段保存、汇编学位论文。

论文作者（签名）：C:\Users\86157\AppData\Local\Temp\WeChat Files\341298396258774299.jpg 指导教师（签名）：

2022年 6 月 1 日 2022 年 6 月 1 日

摘 要

住房一直以来都是人民群众最为关心的问题，是关系民生的大事。但是过快上涨的商品房房价始终困扰着对住房与需求的人群，尤其是收入较低的人群，这已经发展成为了人民日常生活中的敏感问题，因此研究房地产价格的影响因素并对其进行预测有助于帮助对住房有需求的人民群众选择适当的买房时机，也有助于帮助政府相关人员了解房地产价格的走势并对其进行调控。

基于房价过快上涨的情况，本文通过贝叶斯算法对美国金县二手房价进行了分析和预测，主要工作内容是对从网站获取出的数据再进一步的进行信息的处理。并通过贝叶斯算法系统对二手房价信息数据进行过滤并对其过滤后的数据进行分析及预测，对户型、朝向、面积、楼层高度、有无电梯配备、装修程度，进行了可视化分析，更加直观的了解了市场。本文的核心内容是对房价进行精准预测。

关键词：二手房价的价格预测；贝叶斯线性回归分析；大数据分析；供求分析

ABSTRACT

Selling housing has always been the most concerned issue of the people and a major event related to the people's livelihood. However, the rapid rise in commercial housing prices has always plagued people with housing and demand, especially people with low incomes, which has developed into a sensitive issue in people's daily lives, so studying the influencing factors of real estate prices and predicting them will help people with housing demand to choose the appropriate time to buy a house, and also help relevant government personnel understand the trend of real estate prices and regulate them.

Based on the rapid rise in house prices, this paper analyzes and predicts the second-hand house prices in King County, USA through bayesian algorithms, and the main work content is to further process the data obtained from the website. Through the Bayesian algorithm system, the second-hand house price information data is filtered and the data after its filtering is analyzed and predicted, and the visual analysis of the apartment type, orientation, area, floor height, elevator equipment, and decoration degree is carried out, and the market is more intuitively understood. The core of this article is to make accurate predictions about house prices.

**Keywords:** Price prediction for second-hand houses; Bayesian linear regression analysis；Big data analytics; Supply and demand analysis

目 录

[第1章 概述 1](#_Toc104652183)

[1.1 研究背景 1](#_Toc104652184)

[1.2 目的及意义 1](#_Toc104652185)

[1.3 研究目标与内容 2](#_Toc104652186)

[1.4 国内外研究现状 2](#_Toc104652187)

[1.5 本文研究思路 4](#_Toc104652188)

[第2章 相关理论介绍 5](#_Toc104652189)

[2.1 贝叶斯算法理论 5](#_Toc104652190)

[2.1.1 贝叶斯定理 5](#_Toc104652191)

[2.1.2 贝叶斯算法的优缺点 6](#_Toc104652192)

[2.2 特征提取 6](#_Toc104652193)

[2.3 本章小结 8](#_Toc104652194)

[第3章 数据来源及数据处理 9](#_Toc104652195)

[3.1 数据来源 9](#_Toc104652196)

[3.2 数据属性 9](#_Toc104652197)

[3.3 数据预处理 10](#_Toc104652198)

[3.4 本章小结 11](#_Toc104652199)

[第4章 二手房价信息的可视化 12](#_Toc104652200)

[4.1 二手房价信息的描述性可视化分析 12](#_Toc104652201)

[4.2 二手房价信息的特征挖掘 14](#_Toc104652202)

[4.3 本章小结 14](#_Toc104652203)

[第5章 二手房价信息的数据分析 15](#_Toc104652204)

[5.1 二手房价信息的数据模型分析 15](#_Toc104652205)

[5.1.1 模型训练结果 15](#_Toc104652206)

[5.1.2 模型检验 15](#_Toc104652207)

[5.2 实验结果及性能分析 16](#_Toc104652208)

[5.2.1 系统实验结果的质量评价指标 16](#_Toc104652209)

[5.2.2 阈值的确定及对过滤精度的影响 17](#_Toc104652210)

[5.2.3 实验结果分析 17](#_Toc104652211)

[5.3 本章小结 19](#_Toc104652212)

[第6章 基于网页的房价预测模块设计与实现 20](#_Toc104652213)

[6.1 模块设计 20](#_Toc104652214)

[6.1.1 环境配置 20](#_Toc104652215)

[6.1.2 页面展示效果 20](#_Toc104652216)

[6.1.3 功能与性能测试 22](#_Toc104652217)

[6.2 本章小结 26](#_Toc104652218)

[第7章 总结与展望 27](#_Toc104652219)

[7.1 本文工作总结 27](#_Toc104652220)

[7.2 展望 27](#_Toc104652221)

[参考文献 28](#_Toc104652222)

[致 谢 29](#_Toc104652223)

# 第1章 概述

本章介绍了二手房价的研究背景、目的及意义，对于国内外的研究现状进行简要概括，在本章最后介绍了本文的研究思路。

## 1.1 研究背景

影响房价的因素，考虑到短期的货币政策、中期的土地供应和长期的人口统计影响商品房价格的主要宏观因素是：国家宏观政策、区域经济发展水平和监管政策市之间的房价。与宏观环境因素相比，中观和微观环境因素，这些因素对房价有更直接的影响。中观因素，例如与地铁站的距离，与公园的距离，如果房子位于学区。微观因素，如房子的年龄，房子是否有电梯，房子的布局等，都会影响房子的价格。土地的价值在于其位置。例如，有生活资源、商业资源和景观资源。生物资源是一个基本的前提，围绕应该至少有一个吃饭的地方，一个超市，一个餐馆和一辆地铁巴士。商业资源比生物资源高一个层次，景观资源是最高层次。然而，如果豪华别墅位于远离城市的地方，也是没有价值,近几年来计算机网络技术高速发展，网络信息量呈指数级增长。二手房价的分析和预测可以形成一种个性化网络信息采集与处理系统。将该系统应用微博用户兴趣群体发现信息，能有效提二手房价的分析和预测数据可视化服务系统的满意程度，实为一举两得。

## 1.2 目的及意义

衣食住行是人们生活的必需品，当房地产价格过高时，会影响人们的生活。此外，房地产行业在全球经济发展中发挥着重要作用。因此，深入研究影响商业地产价格的因素，预测商业地产价格的趋势，不仅有助于提高国家的竞争力，也有助于提高国家的竞争力。这不仅有助于国家调控商品房价格，使其在合理范围内波动，还能让人们在更合适的时间购买房屋。商业地产价格商业地产的价格受到许多因素的影响，如行政、经济、社会、位置等。这些因素的任何变化都会导致商业地产的价格变化。这些因素与商业房地产价格之间的关系是复杂和不确定的，因此有必要建立适当的模型来研究这种关系并预测商业房地产的价格。这些因素与商业地产价格之间的关系是复杂和不确定的，需要开发适当的模型来研究这种关系并预测价格趋势。

在对政府的实际影响方面，可以根据本文的结论制定一些房价政策。另一方面对开发商来说，它具有重要的经济意义。这是因为一个房地产项目从开始到结束涵盖了一个很长的时期，房价预测模型可以用来反映这一点。该项目实施时间较长，使用商品价格预测模型来捕捉一段时间内的价格变化，然后根据这些变化选择适当的投资方式和营销工具，这种模式对房地产开发商来说也非常有价值，因为它可以在一定程度上避免风险，获得相对更大的经济利益；对于那些希望购买或出售房地产的人来说，预测是非常有用的。对于那些想买房或卖房的人来说，房价预测模型可以预测商品房的价格，以便明智地计划买房或卖房，在一定程度上，它还可以消除风险，减少损失。

当房地产价格上涨时，人们不应惊慌失措，当房地产价格下跌时，也不应急于买房，而应根据实际情况正确、合理地买卖房屋。

## 1.3 研究目标与内容

本文的研究以贝叶斯算法为基础，对二手房价格进行分析和预测，对相关算法进行分析、比较、验证和检验，最后基于改进的贝叶斯算法对二手房价格进行分析和预测。相关的算法也将被分析、比较、验证和测试，最后将实现基于贝叶斯算法的房价分析和预测。

本文的主要工作是研究和调查一种基于数据分析的房屋价格预测方法，分析贝叶斯算法的原理和实现及其改进算法的原理与方法，并研究一个原型过滤系统的设计和实施。主要工作内容如下：

对现有过滤技术的发展状况，基本原理，方法和过程，进行分析和研究，主要过滤技术的优点和缺点比较。

根据分类后的二手房价格单位样本，对二手房价格信息进行分类检查，对每一类的二手房价格信息进行提取关键词，是二手房价格中的敏感符号，是表达该类别文本内容的类别名称，也就是分类向量。计算每种情况的出现的概率，就可以的到特征向量。

木文采用互信息的方法进行特征选取，计算每个关键词的RMI值，利用RMI值确定特征向量的分量词条。

解决基于向量空间模型的二手房价信息表示、数字分类中的特征向量的选择等问题，在核心的过滤实现技术中，重点是设计和实现一个基于贝叶斯算法的高级过滤系统，该算法基于一个简单的贝叶斯方法，完成二手房价信息分析及预测程。

## 1.4 国内外研究现状

针对房价问题，中国许多知名研究者从不同角度分析了影响房价的主要因素。 邱启荣和于婷首先对影响房价的因素进行了主成分分析。通过进行主成分分析，将每个指标组合成一个新的综合指标，然后用BP神经网络进行建模，其优点是非线性的预测能力十分强大，可以对房价进行预测。赵丽丽和焦继文从需求和供给的角度选择了影响商品房价格的变量，然后。然后他们使用灰色关联分析来确定每个变量的相对重要性。武秀丽、张锋以广州市几个有代表性的以行政区的房价数据为分析对象，结果显示，预测值与实际观测值基本一致，时间序列模型顺利通过了残差分析等检验，且误差较小，达到了预测的目的。钟丽艳和高淑兰分析了南宁市人均地区生产总值、商品房销售面积、房地产投资、税收收入、城镇居民人均可支配收入和城镇人口对房价的影响，并建立了多元线性回归模型，采用逐步回归的方法分析和预测房价走势。胡流星和吴杰飞应用灰度理论作为理论基础，分别构建了GM(1,1)模型和纳入灰度理论的单变量线性回归模型来预测房价。GeorgeMin用实证分析表明生产成本的波动对房价波动的强大作用。曾军、李伯煌和张伟以住房价格为研究对象，建立了一个基于数据挖掘的住房价格预测模型。他们运用灰色理论深入分析，确定了影响房价的主要因素及其相应的权重，经过分析，最终确定了影响房价的四个主要因素。曹锐和周峰用多项式回归模型找到了影响房价的主要因素，从而为解释现实中的房价波动提供了依据，也为政府调控房价波动提供了依据。此外，他们还利用多项式回归模型找到影响房价的主要因素，从而为解释现实中的房价波动提供依据，为政府调控房价波动提供指导。侯普光和乔泽群将小波分析理论与ARMA模型相结合，从而实现美国金县二手房价的预测。影响二手房价格的主要因素有：

（1）地段和交通因素

小区位置比较繁华，娱乐、交通、教育都比较便利的房子价格自然就会比较高。城区中心黄金地段属于稀缺资源，谁拥有它意味着谁就可享受便捷的交通资源、完善的商业配套资源、成熟生活配套资源等，房价则不言而喻。一般情况下，市中心区域的二手 房与郊区二手房价格相比能产生好几百甚至上千元一平米的落差。交通状况对房价的影响也不容忽视，越来越多的上班族选择有良好交通连接的住宅，主要是为了方便出行。其主要原因是交通便利，但交通便利并不一定意味着房产价格无可匹敌。还需要评估交通噪音以及周边地区的卫生和空气质量等因素，应考虑到房产的环境影响。

（2）楼层因素

高层在视野、环境、噪音污染相比多层来说比较有优势。但是对于物业费来说多层的比较有优势。所以对于不同的人群，喜好不同，心理价位也会有变化。2009年之前一般来说高层的房子会比多层的价格要高一些。但是现在大多都是高层，由于物以稀为贵的思想，反而多层的房子价位比较高。

（3）建成年代因素

现在的房屋无论是在设计的风格、结构还是装修，绿化设计都比较符合现在人的审美，从而获得的评价比较高。一般来说，距离现在越近的二手房价位越高。处于同一地段的二手房与商品房，在价格方面不受地理位置及交通因素的影响，其主要区别则是由房屋的房龄及户型等决定。房龄在五年以上的二手房，与同样楼层、朝向、面积的商品房相比，由于房型及小区环境存在一定差距而造成房价比商品房低，一般房价差距约在1000元/平米左右；房龄在五年以内的二手房，与同样情况的商品房相比，在房型及小区环境上并不落后，有时由于五年以内的二手房价具有装修时尚、社区成熟等优势其价格与商品房相比差距也并不大。

（4）装修程度因素

二手房屋的买卖通常都会把装修成都算到价格里面，但是至于装修值多少钱，不但只看房主装修时花费了多少，也会看买家对这个装修风格是否满意，是否喜欢，不喜欢的话等于装修没有，甚至还会进行装修的拆除，这样又会造成许多建筑垃圾，而且还会需要一部分费用来进行拆除。并且装修还具有折旧原则，一年以上折旧30%以上，两年以上折旧60%以上，三年以上房屋装修成本计入总价的基本为零。

（5）其他因素

例如三小房屋（即卧室小、客厅小、卫生间比较小的）房价会比较低。我们速算的思路就是选取一个比较基准点的价格，找出这些影响因素的比例，也就是计量学上的β系数。然后下次估算的时候直接运用*β*系数就可以快速的计算出相关地区的房价。

Elaine Worzala、Margarita Lenk和Ana Silva（1995年）使用连接模型收集了科罗拉多州柯林斯堡的288个房地产交易数据，构建并预测了一个房屋估价模型，并将预测结果与简单的多元线性回归的结果进行比较。CrippsA1（2001）使用一个连接模型和一个简单的多元线性回归模型来构建一个使用不同数量的数据点预测房价的模型，并据此估计房价。Jim和Chen（2009）在分析房价时使用了一个特征价格模型。他们关注的是周围环境这一因素与房价变化之间的关系。EAAntipov和ElenaB、Pokryshevskaya等人（2012）试图使用随机森林模型来估计房价，在此基础上，运用多元回归、随机森林、神经网络、决策树等方法对实际住宅进行了定价预测。最后的预测结果表明，在不同的预测方法中，随机森林法的预测效果要好于其他方法。

## 1.5 本文研究思路

本文根据美国金县二手房价数据集中包含的8个字段信息。如价格、占地面积、楼层、装修程度、卧室数和建成年代等信息，使用Pycharm、Python这些工具，对数据集内容进行分析和建模，具体内容如下：

（1）利用PyCharm进行数据预处理。

（2）利用PyCharm软件分别对各个特征进行描述性统计分析，分析各特征的分布。

（3）利用PyCharm软件建立模型预测。

（4）利用PyCharm软件，构建数据挖掘相关模型，分析客户美国金县二手房价的影响因素。

（5）对二手房价的预测进行页面展示。

（6）基于前面的研究成果归纳总结，为对住房有需求的人群给出了一些参考的建议。

# 第2章 相关理论介绍

本章主要二手房价分析和预测的相关方法及理论的介绍，包括对贝叶斯算法的定理和贝叶斯算法优缺点的介绍，以及对特征提取方法的介绍。

## 2.1 贝叶斯算法理论

贝叶斯是机器学习中的一个重要分支。经过多年的研究，该方法已被广泛地应用于各个领域，特别是在专家系统方面，获得了很好的效果。由于其在理论上能最大限度地减少错误，因而在分类问题中得到了越来越多的应用，因而又一次成为了数据挖掘的热门话题。

### 2.1.1 贝叶斯定理

贝叶斯推理是以概率为基础的一种推理方式，其理论依据是与概率有关的理论。有条件概率是一种在日常生活中经常使用的方法。比如，既然知道了今天要下雨，那明天还有什么可能？这个问题涉及到了有条件的可能性。一般情况下，如果A事件已发生，则B的可能性有多大？我们把它叫做B事件的有条件的可能性。其定义如下：假定A、B为两个事件，且知道 A发生的可能性P(A)>0，则B事件的发生几率为公式（2-1）：

|  |  |
| --- | --- |
|  | （2-1） |

在某些情况下，我们必须考虑两个事件同时出现的可能性，也就是A与B的合并几率是P（AB）。它的公式（2-2）为：

|  |  |
| --- | --- |
|  | （2-2） |

也就是当A事件发生时，B事件也发生的概率，即两个事件同时发生的概率。假设影响事件A的事件有,…并且满足下面的条件：，，，则有如下全概率公式（2-3）：

|  |  |
| --- | --- |
|  | （2-3） |

虽然我们并不清楚其中的一些可能性，但我们可以通过我们所掌握的资料和自己的经历来估算每个事件发生的机率。这个可能性叫做“先验”。若从已有资料中得到的先验机率，则此机率即为先验机率；如果用我们的经验来判断，这个可能性被称为一个先验的可能性。

后验概率是指在调查中利用贝叶斯公式修正前验概率，最后获得更为精确的概率，也就是后验概率。

### 2.1.2 贝叶斯算法的优缺点

朴素贝叶斯模型是从经典数学中得到的，其分类效率是稳定的。对于小型数据，它具有很好的性能，可以进行多种类型的工作，适用于增量式的训练，在数据量大于内存的情况下，可以一次一次地进行增量训练。该方法对丢失的数据不敏感，且算法简单，是文本分类中的常用方法。但从理论上讲，相对于其它的分类方法，单纯贝叶斯模型的错误率最小，而事实并非如此，因为当一个朴素贝叶斯模型给出了一个输出类型时，属性间的关系是互相独立的，这种假定经常适用于实际，当属性数量过多，或属性间的关联度很高时，分类效果就会很差；当属性相关度很低时，朴素贝叶斯的表现是最好的。

## 2.2 特征提取

二手房价格信息筛选方法的特点选取与特征抽取是利用特征评价函数将测量空间的数据映射到特征空间中，获得特征空间的数值，再对其进行评估，从而使特征选取成为具有较高价值的几个特征。

目前，通常采用七种特征选择算法：文档频率、信息增益、相互信息、交叉引用期望值、文木测试法、词强度、开放适应性测试，本文采用相互信息的计算方法，以下是简单介绍。

文件频度（DF）是一个最简单的评价功能，它的数值是在一个训练集合中出现该字的文字数目，它的理论假定，如果稀少的字没有任何有用的信息，或者很少，不能对分类造成影响，可以删除。该方法具有计算量少的优点，但由于某些类型的文字数量较少，且含有较多的重要信息，因此准确率较低。我们经常使用此方法来判断其他评价功能。

这个方法有一个问题，即当一个词在I类中出现，而在II类中没有被发现时，它的作用很大，但当它被每一吨的重量加起来时，它就会变成0。该问题的解决办法主要有两种：一是将每个bg的数值都设为绝对值，另一种是略减数值小于0的情形。预期交叉爛（CE）。

其计算公式(2-4)为：

|  |  |
| --- | --- |
|  | （2-4） |

其中，P(CilT)是文字T出现在文本中时具有Ci的文本数量的概率，P(Ci)是类别Ci出现的概率。交叉爛反映了文本类别T字的交叉爛越高，对文本类别分布的影响就越大。一个T字的跨度越大，对文本类别分布的影响就越大。

文本证据权（WE）将类别的发生几率与类别发生的条件概率Z之间的差异进行了比较，当词条与文字的相关性大于且对应的类别出现的可能性较低时，则表示该条目对分类的影响较大，则WE（T）的数值也会更大。其计算公式(2-5)为：

|  |  |
| --- | --- |
|  | （2-5） |

词条强度最早是威尔伯和西罗特金提出的。该方法的主要目的是评价一个条目在相关文件中的普遍程度。在TS算法中，首先要有一组可用于计算文档相似程度的文件。假定有n个文件，全部由dtn个有效的条目组成,如果在文档i中出现了t个布尔向量，如果在文档if中出现了TERMk(lWkWt)，那么DOCi第k个单位TERMik=l，否则TERMik=0于是可以得到一个有0，1构成的矩阵表2-1所示。

表2-1 词条出现状态表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 文件 | 状态 | 状态 | 状态 | 状态 |
|  | TERM1 | TERM2 | ••• | TERHt |
| D0C1 | TERM11 | TERM12 | ••• | TERMIt |
| D0C2 | TERM21 | TERM22 | ••• | TERM21 |
| ••• | ••• | ••• | ••• | ••• |
| DOCn | TERMnl | TERMn2 | ••• | TERMnt |

通过中文文献的多次试验，发现相似概率与夹角余弦值呈单调增加的趋势，且随角度余弦的增大而增大。在角度余弦小于0.4的情况下，相似概率几乎为0。设置阈值 threshold，从试验数据的统计数据来看，这个阈值在0.6左右比较合理。词条的强度是估计一个词条在两个段落中同时发生的一个条件的可能性。词条强度法与之前提到的方法的本质区别在于：它假定了一个完整的词汇集，所有的文字都是从一个词条集合中聚集起来的，而且越是相似的文档，就会拥有更多的条目，所以，当一个单词被更多的文档共享时，这个单词就会被更多的文件分类。开放式拟合试验（CHI）。

在用MI度量特征项目重要性的时候，仅将正相关因素对其重要性的影响因素进行了分析。当属性值t与Ci值呈相反关系时，则表示包含属性值t的文件更可能不是Ci，这将有助于判定一份文件是否不属于Ci分类。为了解决此问题，CHI采用以下方程式来求特性值t与类Ci之间的关联度。

其公式(2-6)如下：

|  |  |
| --- | --- |
|  | （2-6） |

其中：A是t与Ci的共存时间；B表示t的出现，但Ci不存在。C表示Ci的出现，但t不存在；D表示t和Ci不能同时存在的时间。N是文件的数量。与MI相似，当t与Ci无关时，它的数值是0。与MI一样，若存在k个类别，则各t具有k个值，取其平均值即可获得该属性选择所需要的线性序列，选择较大的平均特征的概率较大。

互信息(RMI)是本文采用的特征选取方法。视觉上的意思就是，每一个条目在各个类别中的出现，所占的比例，就是它们在各个类别中所起的作用。但是，该方法忽视了分类中的文本数量对各分类中词条的比例的影响。为此引入类别文本量占整个文本集的比率来表示。

关键词wi的互信息计算公式(2-7)为：

|  |  |
| --- | --- |
|  | （2-7） |

N(i)表示类别G中岀现的总的文本量，其中R(i)=N(i)/N表示类C；的文本量在整个二手房价信息中的文本量的比率；P(WilCj)则是Wi在Cj屮出现的概率，P(w)是在整个二手房价信息集合屮出现的概率。

## 2.3 本章小结

本章主要介绍了二手房价分析和预测的相关方法及理论，通过一些简单的公式对贝叶斯算法进行介绍，并对贝叶斯算法的定理及其优缺点的进行了介绍，本文还对文档频率、信息增益等常用的特征抽取技术进行了分析。

# 第3章 数据来源及数据处理

在本章节，主要介绍美国金县二手房价信息的数据来源、数据属性，并对数据集进行数据清洗，提出接下来需要处理的问题。

## 3.1 数据来源

美国金县二手房价的数据集选自datafountain网站，共包含21,614个二手房价信息，都是美国金县一些真实的二手房价信息。这个站点的宗旨是为学术界和有关行业的爱好者提供一个组织机器学习竞赛，数据库编写和代码分享的平台。在平台上对问题进行说明，并为对应的问题寻找有竞争力的解决办法。参与者可以在平台上下载相关的资料，利用机器学习、深度学习等相关知识，构建相应的数学模型，并对问题进行求解，并给出结论。

数据集下载网址为https://www.datafountain.cn/datasets/67。

## 3.2 数据属性

本研究所用的资料是 EMSCAD，可从网页上下载。下载的资料集档案采用 CSV档，将资料全部插入一栏。使用MS-Excel对资料进行分类。该资料集包含21614条线，每条线都表示美国金县的二手房价格资讯，每一栏包含了以下表格3-1所列出的具体资料。数据集结合了各种数据类型。本次整理的数据共有8个类别，分别为价格、卫生间数、卧室数、售卖日期、占地面积、楼层高度、总楼层数、建造年代。

表3-1 属性展示

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段名 | 数据类型 | 字段描述 |
| 1 | id | Integer | 房屋编码 |
| 2 | date | String | 售卖日期 |
| 3 | price | Integer | 价格 |
| 4 | bedrooms | Integer | 卧室数 |
| 5 | bathrooms | Integer | 卫生间数 |
| 6 | sqft\_living | Integer | 居住面积 |
| 7 | sqft\_lot | Integer | 地点面积 |
| 8 | yr\_built | Integer | 建成年份 |

在这项研究工作中使用的数据集是EMSCAD，可以从网站下载获得。下载的数据集文件是CSV文件格式，所有数据都插入到一列中。利用MS-Excel将数据分为一列。数据集包括21614行，每行代表一个美国金县二手房价信息，每列包括上述表3-1中的特定相关信息。数据集结合了各种数据类型。本次整理的数据共有8个类别，分别为价格、卫生间数、卧室数、售卖日期、占地面积、楼层高度、总楼层数、建造年代。

部分数据展示如图3-1所示。

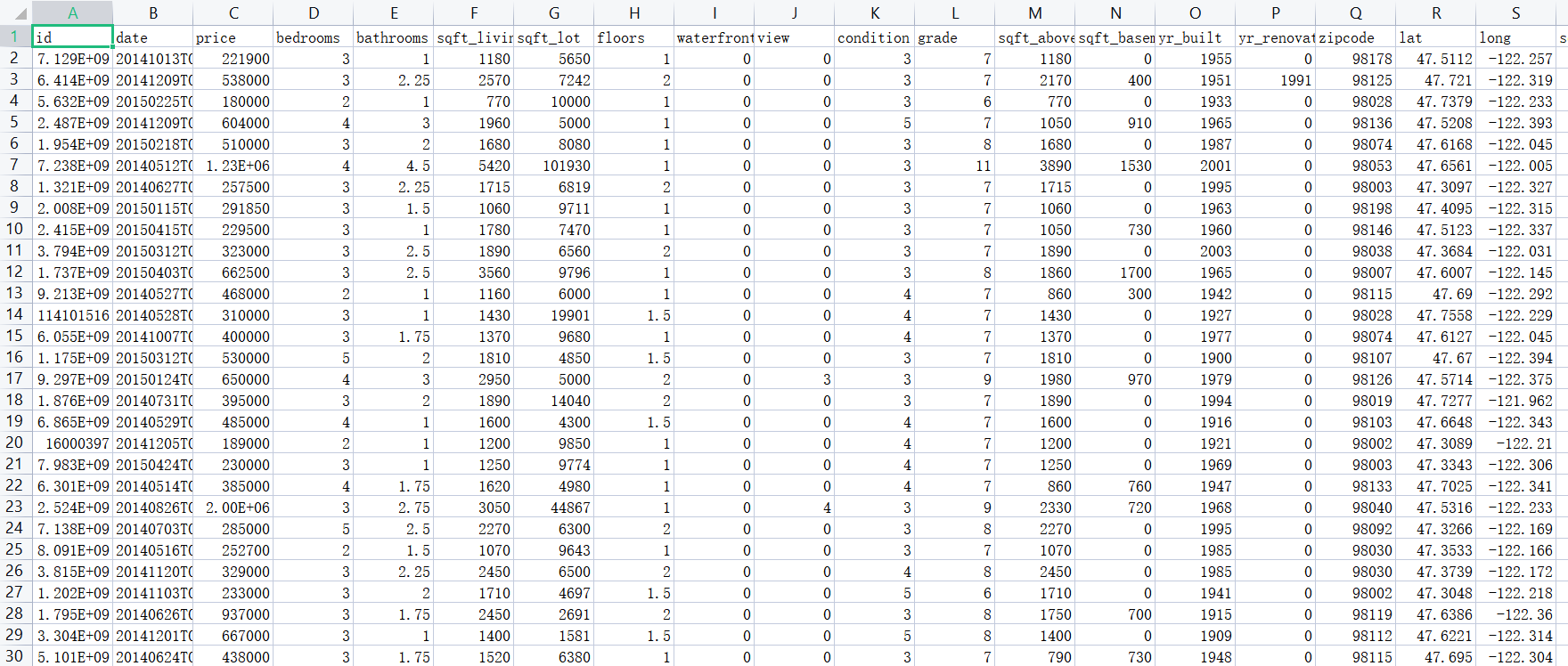


图3-1 数据展示图

3.3 数据预处理

数据预处理就是在对数据进行归类和归类之前，需要进行诸如审核、筛选、排序等的过程。目前有多种数据预处理的方法。包含数据清洗，数据合并，数据转换等。所谓的数据整理，就是指一些我们并不想要的数据，或者是缺少了一些数据，这些数据都会影响到数据的完整性，当某个属性的缺失数值太多的时候，这些数据就会被删除。如果缺损所占比例较低，可用适当的缺损处理方法对其进行补足。调整错误的资料格式，达到相应的资料规范，排除重复资料及异常资料，以纠正错误。代码如下：

importpandasaspd

file=pd.read\_csv('chengdu.csv',encoding='utf-8')

file=file.drop(['Unnamed:0','WebPage','Info','VR','followInfo','Title'],axis=1)

print(file.head())

first=file['Tag'].str.split('|',expand=True)

first.rename(columns={0:'室厅数',1:'面积（平米）',2:'orientation',3:'Style',4:'楼层',5:'建筑时间',6:'Type'},inplace=True)

df=pd.concat([first,file],axis=1)

df=df.drop(['室厅数','面积（平米）','楼层','Tag','建筑时间'],axis=1)

\#df.describe()

df.dropna(axis=0,how='any',inplace=True)

\#fillna(value=None,method=None,axis=None,inplace=False,limit=None,downcast=None,\*\*kwargs)

\#print(df['Year'].value\_counts())

df=df[df['Year']!=0]

df=df.reset\_index(drop=True)

foriinrange(len(df)):

ifdf['Updateday'][i]=='>365':

df['Updateday'][i]=str(df['Updateday'][i])

df['Updateday'][i]=df['Updateday'][i].replace('>365','365')

foriinrange(len(df)):

df['Updateday'][i]=float(df['Updateday'][i])

print("--------")

print(df.head())

3.4 本章小结

本章首先介绍了本文所用的数据库来源，并且介绍了数据属性，对部分数据进行了介绍，并且对二手房价的数据信息进行了数据预处理，调整不正确的数据格式，满足对应的数据标准，将重复数据和异常数据进行了处理和修正。

# 第4章 二手房价信息的可视化

本章主要二手房价的信息进行可视化和特征提取等研究分析进而进行系统设计，从而实现二手房价的预测。

## 4.1 二手房价信息的描述性可视化分析

首先进行数据集的可视化分析，了解数据集各个变量的统计状况和分布状况。在分析的过程中使用Pycharm进行部分数据集的展示。

首先是观察美国金县经过简单清洗后的实际房价的实际分布。以下是核心代码参考示例：

floorLevel=['低','中','高']

forlevelinfloorLevel:

iflevelinrow[5]:

row[5]=floorLevel.index(level)

break

如图4-1所示，由该图可见，在此选择的情况下。大部分都是高层，二十二层以下的，只有一小部分。

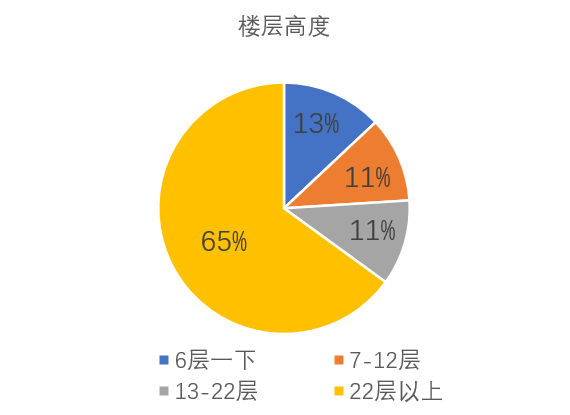


图4-1 楼层高度

decorationMap=['毛坯','精装修','简装修']

row[6]=decorationMap.index(row[6])

得到如图4-2所示的图，从图中可以看出，本次所选的案例中。装修程度为精装占据绝大部分，毛胚装修占比其次，简装占比最少为13%。

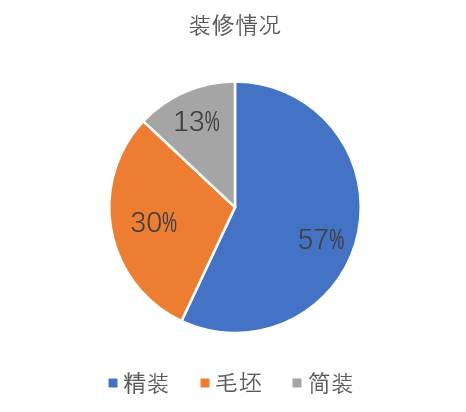


图4-2 装修情况图

leftMap=['有','无']

row[8]=leftMap.index(row[8])

得到了饼状图，由图4-3可见，在这个选择的情况下。大多数房屋都有安装电梯，而没有安装电梯的房屋则占很少。

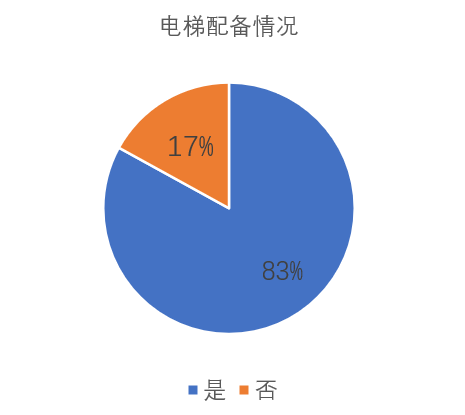


图4-3 电梯配备情况图

accessMap=['个人产权','商品房','商品房(免税)','经济适用房']

row[9]=accessMap.index(row[9])

buildTypeMap=['普通住宅','酒店式公寓','公寓','商住楼']

row[10]=buildTypeMap.index(row[10])

structMap=['平层','复式','跃层']

row[11]=structMap.index(row[11])

## 4.2 二手房价信息的特征挖掘

二手房价信息过滤种的特征选择和特征提取是通过一个特征评价函数，把测量空间的数据映射到特征空间中，得到特征空间的值，然后对特征向量进行评估，特征选择就成了选择值最高的若干个特征。核心代码如下：

defLinerPredict():

data=pd.read\_csv('4.csv')

X=pd.DataFrame(data,columns=['建筑面积','朝向','楼层','装修','建筑年代','有无电梯','产权性质',

'住宅类别','建筑结构','经度','纬度','房间数','客厅数','卫生间数'])

Y=pd.DataFrame(data,columns=['单价'])

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y)

linerRegression=LinearRegression(fit\_intercept=True,normalize=True,copy\_X=True,n\_jobs=None)

linerRegression.fit(x\_train,y\_train)

y\_pred=linerRegression.predict(x\_test)

MSE=metrics.mean\_squared\_error(y\_test,y\_pred)

RMSE=np.sqrt(MSE)

print('MSE:',MSE)

print('RMSE:',RMSE)

print(len(y\_test),len(y\_pred))

plt.figure(figsize=(15,5))

plt.plot(range(len(y\_test)),y\_test,'r',label='Real')

plt.plot(range(len(y\_pred)),y\_pred,'b',label='Pred')

plt.show()

plt.scatter(y\_test,y\_pred)

## 4.3 本章小结

本章主要针对数据分析进行重点阐述，对数据进行了可视化分析，也对数据进行了特征提取，同时也对数据分析的逻辑讲述。

# 第5章 二手房价信息的数据分析

本章主要综合前面数据分析结合数据挖掘中的算法对该数据集进行建模分析。

## 5.1 二手房价信息的数据模型分析

本节主要综合前面数据分析结合贝叶斯算法对该数据集进行模型分析。

### 5.1.1 模型训练结果

通过该模型，对128个试验集合的数据进行了预测，图5-1为预测房价和实际成交价格的关系，绿色表示实际成交价格，蓝色为预测房价。由图表可知，预测结果与实际值之间存在较小的偏差，并且波动幅度较小，说明了预报的有效性。

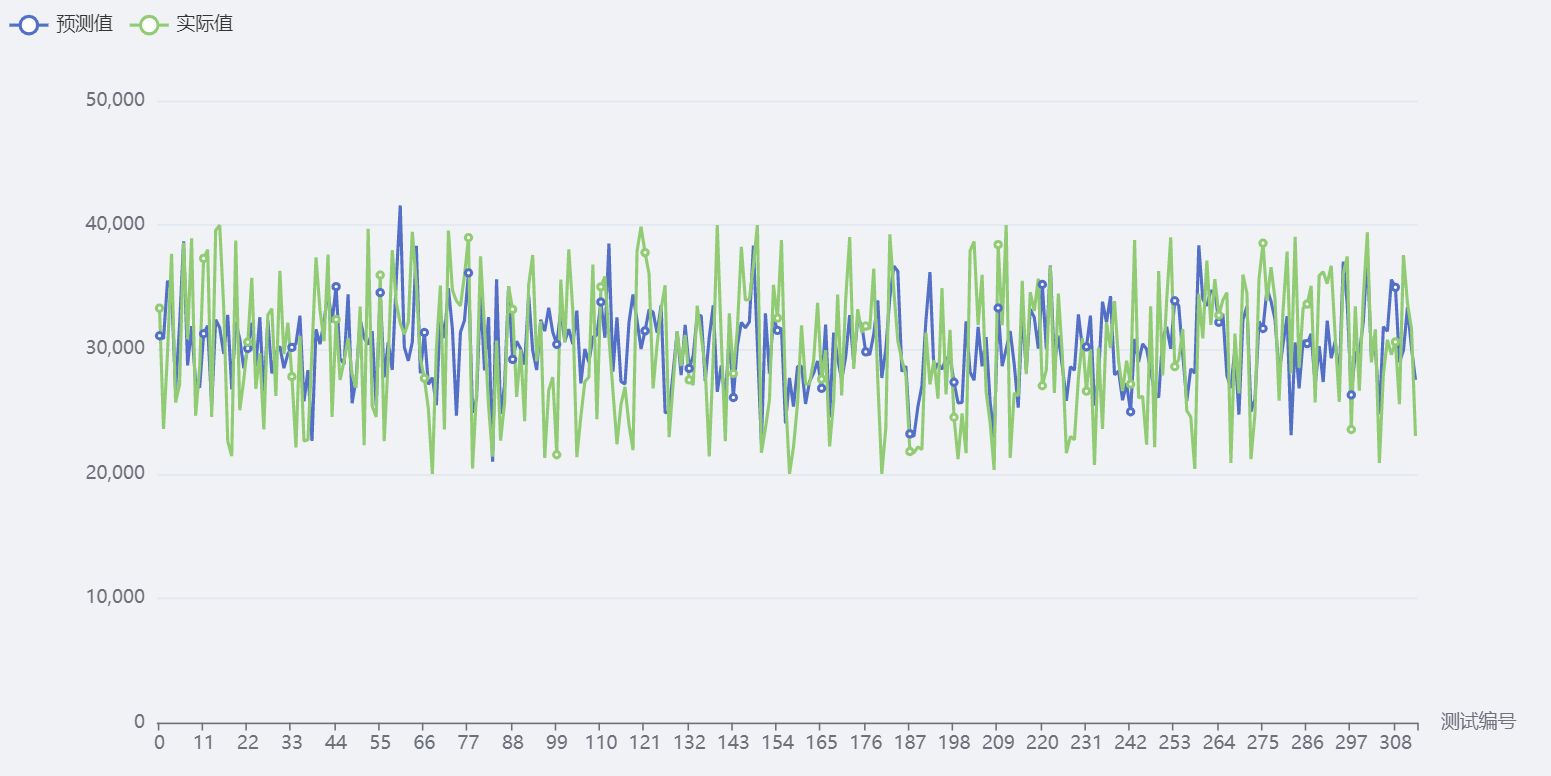


图5-1 预测结果图

### 5.1.2 模型检验

用实际值作横坐标，用预测值作纵坐标，绘制其散点图，若资料接近y=x线，则表示仿真结果良好，由图5-2可知，大多数资料接近y=x线，表明此模式的仿真结果是有效的。用模拟数据和实际数据进行相关性分析，得到的结果是0.98，数值与实际值相近，可见该模型的有效性。

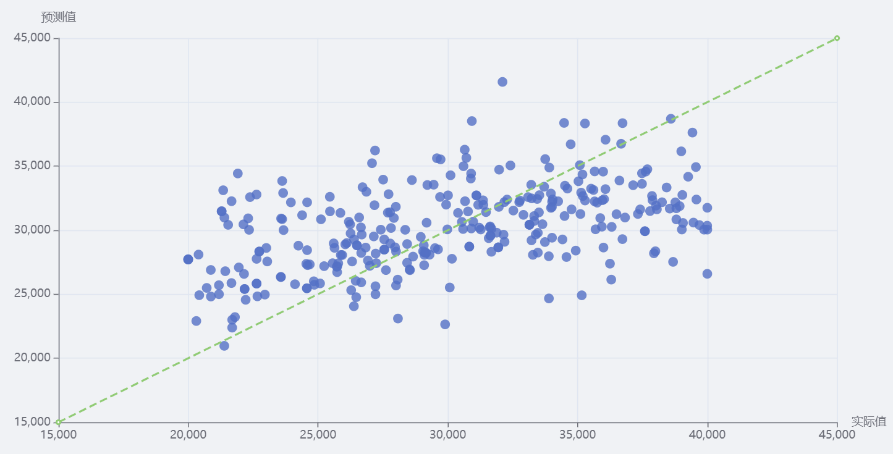


图5-2 预测价格散点图

## 5.2 实验结果及性能分析

根据所建立的二手房价格资讯过滤模型，利用朴素贝叶斯算法建立二手价格资讯过滤器，进行二手价格资讯训练及二手价格资讯辨识试验，并根据不同训练集规模、不同阈值等因素，得出不同的判别结果，不同的检出率及准确度，经过不断的修正，最终取得满意的准确率与准确度。

5.2.1 系统实验结果的质量评价指标

二手房价信息过滤的性能评价通常指文本分类的相关指标。假设待测试的二手房价信息集合共有N处房产，一个二手房价信息过滤系统的判定结果如下表5-1所示：

表5-1 二手房价信息系统判定标准

|  |  |  |
| --- | --- | --- |
|  | 实际为垃圾信息 | 实际为合法信息 |
| 系统判定为垃圾信息 | A | B |
| 系统判定为合法信息 | C | D |

设N为二手房价总数，则N=A+B+C+D,其中A+C为二手房价信息的数量，B+D为合法二手房价信息的数量，木文采用了下面几个标准。

查全率：指的是由人工和系统判定的二手价格信息数量与系统判定的正常价格信息的比率，它反映了一个过滤系统对二手价格的信息进行筛选，查全率越高，则说明该系统可能漏掉的二手房价信息就越少。

正确率：指的是由人工和系统判定为正常的二手价格信息数与人为判断的正常价格信息的比率，准确率越高，错误判断为二手价格的信息就越少。

精确率是指对所有二手房屋价格的判断。一种计算方法是否准确：一是要对真实的二手价格进行准确的筛选，提高对二手价格的筛选精度；二是减少把止常二手房价信息给过滤掉，即减少正常二手房价信息的误判率。我们这里通过两个指标来衡量算法的准确性。二手房价信息查全率和正确率二手房价信息误判率。

除此Z外，二手房价信息过滤中常常采用虚报率(Fallout)＞漏报率(Missrate)等指标。另外，我们在前面提到，在实际的二手房价信息过滤中为了表示不同情况下二手房价信息分析系统的代价。Androutsopoulos等人提出了代价因子的概念。他们把决策损失引入到评价体系中。

5.2.2 阈值的确定及对过滤精度的影响

在第四章中我们在利用贝叶斯过滤器对二手房价信息进行判断的一般过程是计算概率，将概率与阈值比较判断待检测二手房价信息是有用的二手房价信息。阈值在这里的作用就像一个“门槛”，概率阈值对贝叶斯过滤器对二手房价信息的漏判率和正常二手房价信息的谋判率有着直接的重要影响，因此如何选择一个合适的阈值是一个重要的问题。如阈值选取的比较大，可以减少有价值的二手房价信息被判为垃圾二手房价信息的可能，这两个都不是我们想要的。从这一点可以看出，阈值的设定是非常关键的，它直接关系到探测的精度。本文主要涉及到两个阈值的确定。

第一个步骤是计算分类中词条的发生概率，并对RMI进行RMI，并将RMI值作为特征矢量的成分，构造出此类特征矢量。

做法是取前*K*个词条作为该类别的特征，*K*就是阈值，这个阈值的选取是通过对样木本集的实验得出的。

最佳阈值与贝叶斯滤波器的最佳阈值有关，且每个使用者的最佳阈值也各不相同。贝叶斯滤波应该以具体的数据为基础，选择最佳的概率门限。不同的阈值下检测的结果的准确程度是不同的。

5.2.3 实验结果分析

通过对采样集进行了实验，并分析了阈值的选择对贝叶斯滤波效果的影响。

（1）阈值对数据结果的影响

木文对二手房价信息样本集进行二手房价信息分析的训练实验，不同特征项数目、不同阈值等因素对评价指标有不同的影响，不同的查正率和准确度也不同，通过不断地调节阈值直到获得满意的结果，并在不同的采样点选取不同阈值时对不同指标的影响。如图表5-2所示

表5-2 不同阈值对指标的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 概率阈值 | 特征阈值 | 查全率 | 正确率 | 精确率 |
| 0.41 | 22 | 97.7% | 89.7% | 95% |
| 0.49 | 95.6% | 90.5% | 93.5% |
| 0.51 | 93.4% | 88.7% | 92.9% |
| 0.42 | 15 | 95.8% | 94.7% | 95.3% |
| 0.48 | 94.1% | 92.1% | 94.2% |
| 0.49 | 90.6% | 90.8% | 92.3% |
| 0.45 | 10 | 93.6% | 93.2% | 93.7% |
| 0.48 | 92.4% | 92.8% | 92.4% |
| 0.50 | 84.9% | 81.9% | 89.8% |

由表5-3可知，选择的训练集合特征的大小会极大地影响到筛选的准确性和查全率。总体上，选择的特点太多，使用的关键词太多，导致准确率下降；若特征选择过少，则可提供一定的判别基础，且会遗漏许多对判别有影响的特征，从而影响最终的判别结果。

从表5-3可知，在选择不同的特征数目时，不同的概率门限具有不同的查全率、准确率和准确率。从表格中的数据看，当选择了22英寸的特征数，从0.41到0.51的概率阈值时，垃圾价格的查全率和普通的二手价格信息的误判率都会随 Z而改变，而查全率则会逐步降低，而准确率则会随著机率的增大而降低，最后以0.41作为最优的可能性门限；在选择了15个特征个数的条件下，从0.42到0.49的概率门限，垃圾二手价格的查全率和普通的二手价格信息的误判率出现了显著的改变，查全率、正确率和准确率都在下降，选择了15个特征个数时，最好的可能性门限是0.42；在选择了10个特征个数的条件下，从0.45到0.5的概率门限，查全率、正确率和准确率都在逐步降低，0.45是最好的选择概率门限。

（2）最优阈值的分析

贝叶斯滤波算法应该以具体的数据为基础，选择最佳的概率门限。不同的特性数目如表5-3所示。

表5-3 不同的概率阈值对指标的影响

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 特征阈值 | 概率阈值 | 查全率 | 正确率 | 精确率 |
| 22 | 0.41 | 97.7% | 89.7% | 95% |
| 15 | 0.42 | 95.8% | 94.7% | 95.3% |
| 10 | 0.58 | 93.6% | 93.3% | 93.7% |

表5-3展示了在使用不同的特征数目和不同的概率门限时，其查全率、正确率和准确率都有很大差异。在选取最优的概率门限时，特征数目的选择对过滤精度有很大的影响。在特征数22的情况下，房屋价格信息的查全率较高，但由于样本中的特征数太多，这些特征项之间存在较大的相关性，会造成错误判断的概率，进而影响到房屋价格的准确性；在特征数为15的情况下，其查全率、准确率和准确率均较高，且数值波动较小；当特征数为10时，由于样本中的特征数目较少，因此会造成二手价格信息的误判率，从而影响到垃圾邮件的查全率、准确率和准确率。结果表明，在选择了15个特征和0.42的概率阈值时，能够确保良好的查全率和准确率。这是本试验的最佳阈值。

最后我们对本次实验进行分析：在本章开始我们介绍了基于贝叶斯算法的垃圾信息过滤器的总体结构的设计，使我们更深一步认识到了过滤器的总体结构及对实验结果有关的两个阈值的取值问题。选择特征的数目对查全率的影响较为显著。当特征数目的减少时，其查全率将降低，特征数量如果选取过大，会导致关键词数量增加，从而加大对合法二手房价信息的误判率增大。概率阈值对过滤器影响非常大，选取合理的阈值对质量评价指标有很大的影响，经过反复实验，木文对二手房价信息的过滤取得了很好的效果。

## 5.3 本章小结

本章主要对二手房价信息进行了数据挖掘，建立了训练模型，对其进行了验证，并对其进行了质量评估，通过对阈值的确定以及对过滤精度的影响对结果进行了分析。

# 第6章 基于网页的房价预测模块设计与实现

## 6.1 模块设计

本课题侧重于二手房价的分析和预测的系统模块设计，对于所用系统的环境配置进行介绍，因此，前面的网页显示起来就比较容易了，主要就是用户可以通过关键词来查询相关的数据库，从而通过数据可以展示对于二手房价的预测，在对预测结果进行功能与性能测试。

### 6.1.1 环境配置

本系统需要对上万条数据进行预测分析，同时对相关数据进行可视化，因此对硬件环境配置有一定的依赖性，这篇文章采用了Windows10，它整合了一个强大的软件环境，可以开发出新的服务和改善的应用，从而极大地提升了进程的效率。通过本文的试验验证，Windows10具有良好的稳定性。

系统开发时需要的软件平台以及硬件环境如表6-1所示。

表6-1 系统配置环境表

|  |  |
| --- | --- |
| 名称 | 配置 |
| CPU | Intel(R)Core(TM)i7-7700K |
| 内存 | 32GB |
| 硬盘 | 2TB |
| 数据库 | Neo4j,MySQL |
| 开发环境 | PyCharm2019,Myeclipse2018 |
| 编程语言 | Python3.7,Java,JavaScript,HTML |
| 操作系统 | Windows1064bit |

### 6.1.2 页面展示效果

美国金县二手房价的分析和预测系统信息整体结构为左右结构，左面部分采用了树型层次结构罗列了系统的全部功能，其中包括主页、展板、数据、预测和切换主题等板块，其中展板中又包括趋势图和涟漪图两种类型，右边是一个客户区，它展示了一个用于当前操作的窗口。该系统首先为用户提供了一个可靠的操作平台，并通过图形化的方式来展示用户的操作权限，方便用户浏览和操作。如图6-1所示。

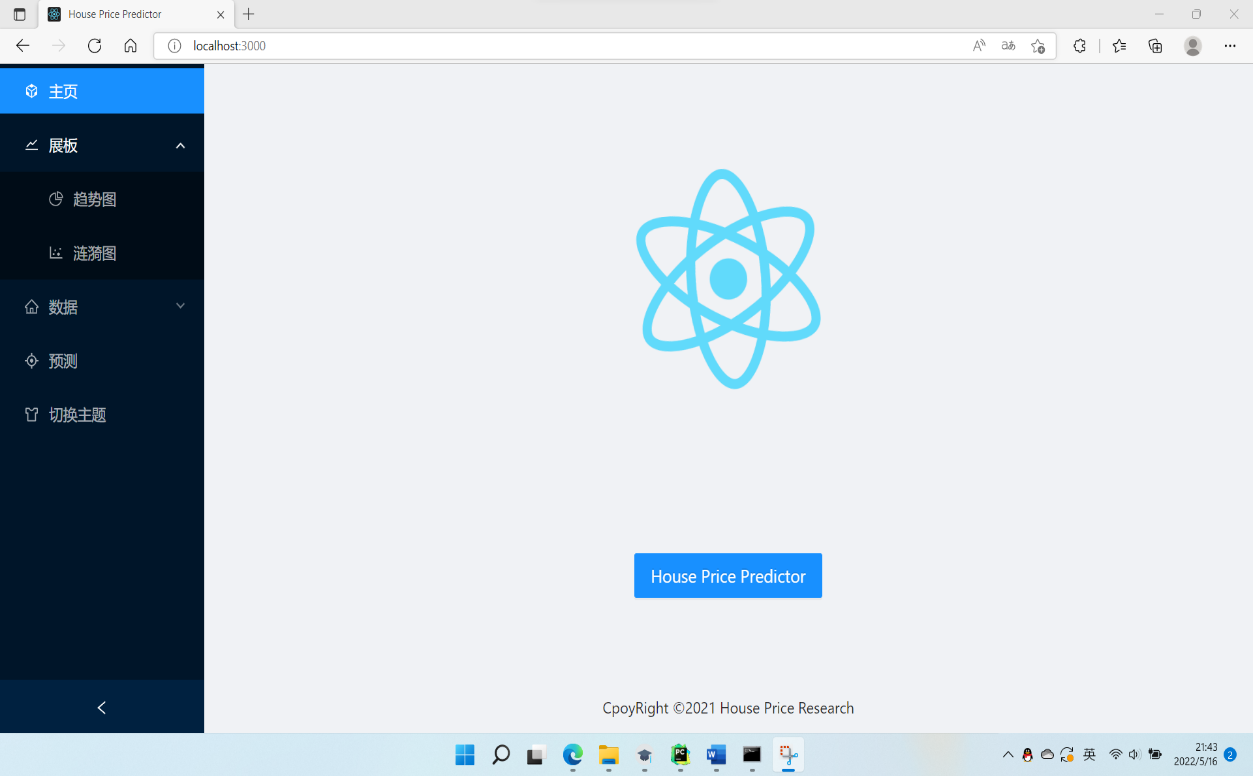


图6-1 用户操作界面图

用户操作界面的核心代码如下：

<Router>

<Layoutstyle={{height:"100%"}}>

<Sidercollapsiblecollapsed={this.state.SideCollapsed}onCollapse={this.onCollapse}theme={this.state.AppTheme}>

<divclassName="logo"/>

<Menutheme={this.state.AppTheme}mode="inline">

<Menu.Itemkey="1"icon={<CodeSandboxOutlined/>}>

<Linkto={"/"}>主页</Link>

</Menu.Item>

<SubMenukey="2"icon={<StockOutlined/>}title="展板">

<Menu.Itemkey="2-1"icon={<PieChartOutlined/>}>

<Linkto={"/demo-charts"}>趋势图</Link>

</Menu.Item>

<Menu.Itemkey="2-2"icon={<DotChartOutlined/>}>

<Linkto={"/dot-charts"}>涟漪图</Link>

</Menu.Item>

</SubMenu>

<SubMenukey={3}title="数据"icon={<HomeOutlined/>}>

<Menu.Itemkey="3-1"icon={<DragOutlined/>}>

<Linkto={"/data\_map"}>地图</Link>

</Menu.Item>

<Menu.Itemkey="3-2"icon={<ChromeOutlined/>}>

<Linkto={"/infos"}>查询</Link>

</Menu.Item>

</SubMenu>

<Menu.Itemkey="4"icon={<AimOutlined/>}>

<Linkto={'/predict'}>预测</Link>

</Menu.Item>

<Menu.Itemkey="5"icon={<SkinOutlined/>}onClick={this.handChangeTheme}>

切换主题

</Menu.Item>

</Menu>

### 6.1.3 功能与性能测试

本功能主要是由服务器管理的美国金县二手房价的分析和预测系统信息数据信息，从而进行二手房价预测，如图6-2所示。

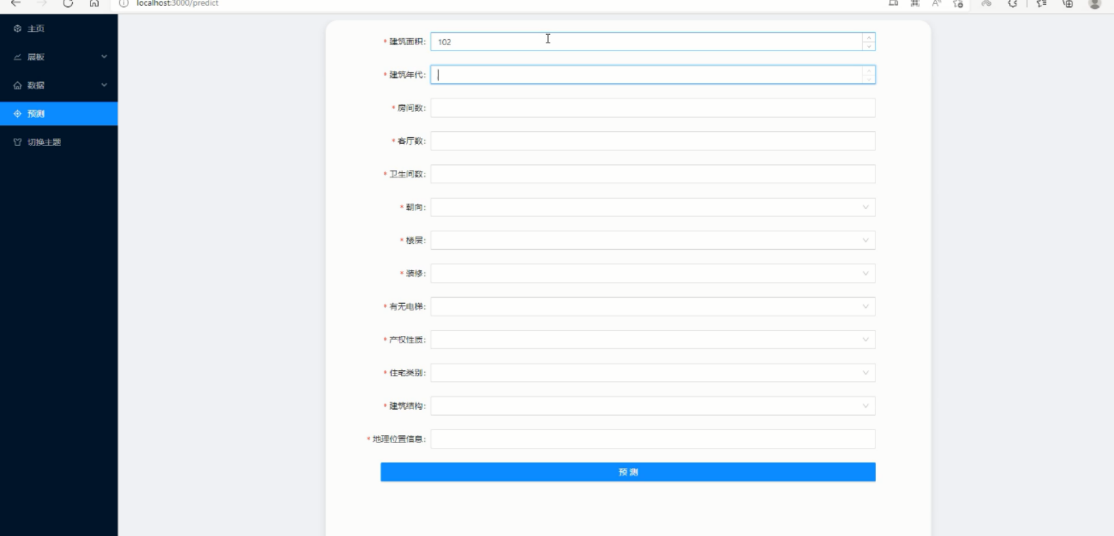


图6-2 二手房信息数据界面图

用户通过填写二手房属性，从而得到二手房价格，预测结果如图6-3所示。

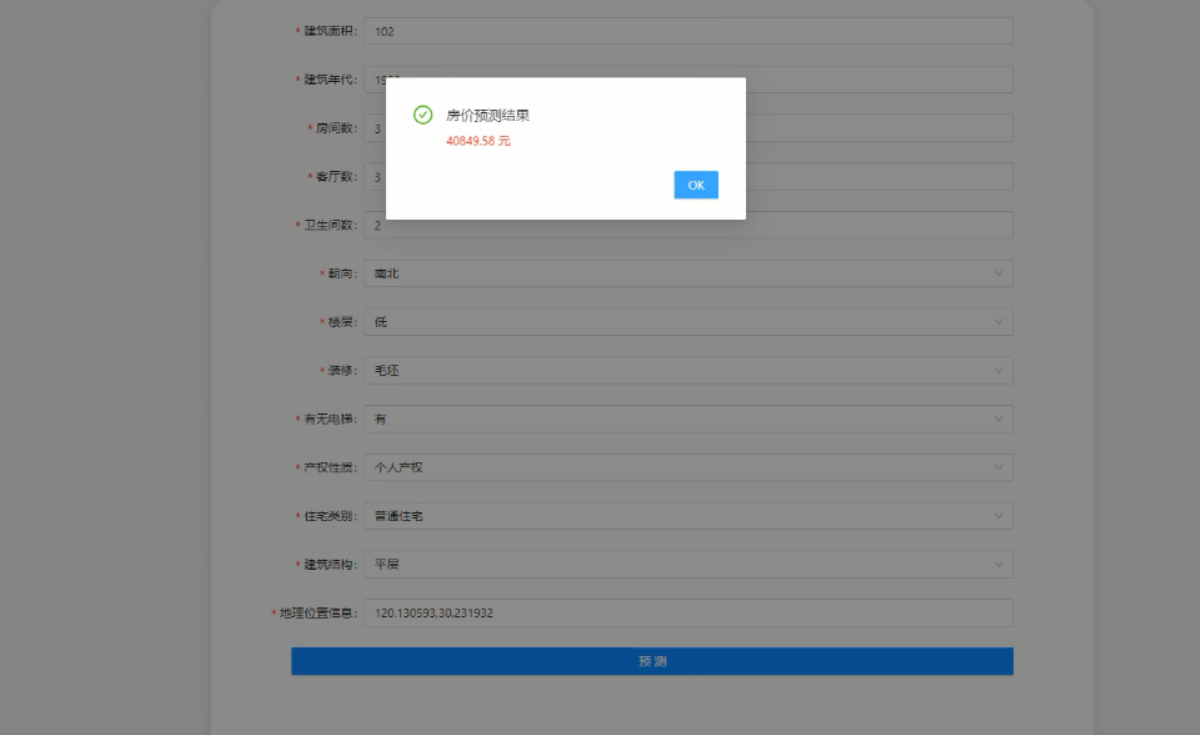


图6-3 二手房价预测结果图

用户预测二手房价的核心代码如下：

console.log('error')}}, [modalVisible])

const hideModal=()=>{

setModalVisible(false) }

const levelOptions = ['低', '中', '高']

const decoratonOptions = ['毛坯', '精装修', '简装修', '中装修', '豪华装修']

const liftOptions = ['有', '无']

const accessTypeOptions = ['个人产权', '商品房', '商品房(免税)', '经济适用']

const Submit=(values)=>{

axios.post('http://localhost:5000/predict', values).then(res=>{

const resp = res.data??{}

const result = resp.result??-1

value={index}>{text}</Select.Option>})}

</Select>

</Form.Item>

<Form.Item

label={'楼层'}

name={'level'}

rules={[{ required: true, message: 'Couldn\'t be Empty' }]}>

<Select style={propStyle}>

{

levelOptions.map((text, index)=>{

return <Select.Option value={index}>{text}</Select.Option>})}

</Select>

</Form.Item>

<Form.Item

label={'装修'}

name={'decoration'}

rules={[{ required: true, message: 'Couldn\'t be Empty' }]}>

<Select style={propStyle}>

{

decoratonOptions.map((text, index)=>{

return <Select.Option value={index}>{text}</Select.Option>})}

</Select>

</Form.Item>

<Form.Item

label={'有无电梯'}

name={'haveLift'}

rules={[{ required: true, message: 'Couldn\'t be Empty' }]}>

<Select style={propStyle}>

{

liftOptions.map((text, index)=>{

return <Select.Option

value={index}>{text}</Select.Option>

})

}

</Select>

</Form.Item>

<Form.Item

label={'地理位置信息'}

name={'location'}

rules={[{ required: true, message: 'Couldn\'t be Empty' }]}>

<Input style={propStyle}

本系统还可以将二手房的具体位置进行显示，如图6-4所示。



图6-4 二手房位置展示图

位置展示的核心代码如下：

defscatter():  
target=Resolve('./result/target.csv')  
answer=Resolve('./result/answer.csv')  
res=[]  
foriinrange(1,len(target)):  
res.append([int(target[i][1]),int(answer[i][1])])  
returnjsonify(res)  
defcurve():  
target=Resolve('./result/target.csv')  
answer=Resolve('./result/answer.csv')  
foriinrange(1,len(target)):  
answer[i]=int(answer[i][1])  
target[i]=int(target[i][1])  
returnjsonify({  
})  
defFindAround(xTrain,yTrain,lng,lat):  
xTrain["距离"]=(xTrain['经度']-lng)\*(xTrain['经度']-lng)+(xTrain['纬度']-lat)\*(xTrain['纬度']-lat)  
SIZE=xTrain.shape[0]  
index=xTrain.nsmallest(SIZE//10,'距离').index.values  
returnxTrain[xTrain.index.isin(index)],yTrain[yTrain.index.isin(index)]

功能测试的目的是观察和归纳出算法的功能是否能够满足使用者的要求。性能测试的重点是测试各功能模块在执行时所需的处理能力和网页的响应速度主要功能：Windows10，浏览器，WiFi；性能测试平台：Windows10,PyCharm2019,WiFi。

测试内容以及结果如表6-2所示。

表6-2 测试内容与结果表

|  |  |
| --- | --- |
| 测试内容 | 实测结果 |
| 测试目的 | 测试功能是否正常显示并且页面成功跳转。 |
| 配置条件 | Windows10、Chrome浏览器、WiFi |
| 测试步骤 | 1，启动算法服务，打开网页  2，点击左侧菜单栏功能框  3，查看所有链接是否正常跳转  4，查看是否正常出现数据回显 |
| 预测结果 | 功能运行正常、页面正常跳转 |
| 功能测试 | √通过，□部分通过，□不通过 |
| 性能测试 | 平均响应时间5秒 |

功能测试和性能测试的过程和结果见表6-2。在功能测试上，各子功能均能正确地显示功能，并可任意点选所需的结果，进行跳跃时，均可正常翻转。在性能上，采用了将算法服务端和客户端分离的方式，各自发挥各自的优点，从而提高了系统的性能响应速度。

## 6.2 本章小结

本章主要对基于网页的二手房价预测模块的设计进行了介绍，对环境配置进行了介绍，并对运行效果进行了展示，对功能和性能测试进行了介绍。

# 第7章 总结与展望

## 7.1 本文工作总结

这项工作的主要目的是评估美国金县的二手房价格，以获得对每套房屋价格的准确分析。首先，在回顾以往文献的基础上，拟选取影响二手房交易价格的九个变量，即面积、户型、层高、总层数、建筑面积、楼龄、装修和电梯设备。并利用训练集数据建立贝叶斯模型，用训练函数替换的测试集数据对模型进行测试，预测价格和实际价格之间的相对误差为[-6%，6%]，发现模型是有效的。本文提出的二手房房价预测，可以为购房者提供指导，减少不必要的经济损失，也可以为政府相关部门监控和完善二手房市场提供数据支持，从而在大数据时代实现智能化管理。

## 7.2 展望

由于科学能力和资源的限制，研究虽然已经有了一些结果，但是从下面的几个方面进行改进肯定会让这个方法更加完善，在实践中更加有效。

（1）数据方面，在建模时由于数据的缺乏只能选择有限年份的由于数据建模使得模型的训练样本变得很少，从而对模型的实际应用产生了一定的影响。应当从数据中获得更多数据。

（2）影响因素的选择方面，在对目标值影响因素的选择方面存在理论与实际的差距。本文从卧室数、卫生间数、楼层、装修、建成年代、等对已经完工的住宅投资、房屋成本等因素进行了综合考虑，但考虑的不充分。另外，区域因素、政府政策等因素对目标价值的影响也有一定的影响。

# 参考文献

|  |  |
| --- | --- |
| [1] | 李函谕，魏嘉银，卢友军．基于随机森林的深圳二手房价格预测与分析[J]．现代信息科技，2021（5）． |
| [2] | 汪静，罗维平，陈永恒．基于神经网络的房价预测与分析[J]．襄阳职业技术学院学报，2021(02)． |
| [3] | 张杰．支持向量回归机在房价预测中的分析与应用[D]．武汉理工大学，2017． |
| [4] | 张砚博．基于多元线性回归分析的西安市房价预测分析[J]．西部皮革，2020．42(10)：71． |
| [5] | 周怡君．重庆市商品房价格的影响因素分析与房价预测[D]．重庆大学，2017． |
| [6] | 白雪菲．基于模型选择和模型平均方法的青岛房价预测分析[D]．中央民族大学，2021． |
| [7] | 万得莉．青岛市房价分析及预测[D]．山东师范大学，2020． |
| [8] | 葛笑畅．上海市二手房价格影响因素分析[D]．上海师范大学，2020． |
| [9] | 代磊，李雪婷．基于多元线性回归模型的二手房价格影响因素分析——以成都市某区为例[J]．河南建材，2019(05)：80-82． |
| [10] | 沈孝玲．北京二手房市场的统计分析[D]．北京理工大学，2017． |
| [11] | 孙浩桐．河北省石家庄市二手房价格的影响因素分析[D]．天津财经大学，2019． |
| [12] | 傅行行．上海市二手房价格空间分布及其影响因素研究[J]．上海房地，2020(07)：11-15． |
| [13] | Huaping Zhao, Xu Wei. Spatial Changes of Urban Housing Prices: Analysis of Traffic CostsBasedonTaiyuan[J]. IOP Conference Series: Materials Science and Engineering. 2019, 688(5). |
| [14] | ChristouGuptaR, HassapisC.Does Economic Policy Uncertainty Forecast Real Housing Ret urns in a Panel of OECD Countries ?ABayesian Approach [J]. 2017. |
| [15] | Friskovec S, Janes A, tefan Bojnec. Analysis of Factors of the Second-Hand Housing Prices in the City and Vicinity of Ljubljana[J].Management,2020, 5:225-242. |

# 致 谢

时间如梭，大学四年马上就要结束了，在这四年的学习生活中，我学习到了很多知识，专业知识方面更是有了很大的提高，是我这一生当中最宝贵的财富，是不断超越自我的历程。在这最后的阶段，我衷心的向学校的各位老师和我的同学表示我最衷心的感谢，感谢他们在这四年当中对我的帮助和关心。很感谢学校能够提供我们这个锻炼自我的机会，给我们一个能够使大学四年所学到的知识加以实践的机会。

感谢从大一到大四所有的授课教师，他们孜孜不倦的教导我们，使我们真正的学习到了知识。感谢我的毕业设计辅导老师周子程老师，感谢这段时间对我的容错与指导，在他的帮助下我才能顺利的完成这次设计，他耐心的辅导我，对待我提出的任何问题都耐心的帮助作答，我真的很感谢指导老师。也感谢我的同组设计同学，在他们的大力合作和帮助之下，我们合理的分工，我们才能顺利有效的完成这次设计，没有他的努力，就没有我们的成果最后感谢帮助过我的每一个人，没有他们就没有我的今天。

最后，我要感谢我的父母，感谢父母培养了我的求知欲，让我深切地懂得知识的力量。感谢父母教给我坚强，在逆境中坚持，在艰难困苦中不放弃。感谢父母让我理解大爱的真谛，读懂关心他人、不图回报的内涵。