基于随机森林的房地产估价模型研究
——以上海市二手房市场为例

摘要

随着中国经济的不断发展，人民生活越来越富裕，住宅房地产的居住性需求和投资性需求都在不断高涨，推动房价不断上涨。然而城市土地供给有限，使得人们的目光逐渐转向二手房市场。在一线城市，二手房交易由于受到宏观政策的制约少于一手房，更能体现出市场特性。二手房的需求增大，房地产估价的需求也不断增大，无论从房地产转让、租赁、抵押，还是从房地产保险、房地产征税、征地和房屋征收拆迁补偿来看，房地产估价都是一个重要的环节。但是房地产估价的传统三大方法成本法、收益还原法和市场比较法存在着许多主观因素，实施过程中过于依赖估价师的经验，可能导致估算价格与实际价格偏离，并且这些估价方法费时费力，浪费了资源。随着计算机科学和机器学习的兴起，众多学者逐渐将定量的方法引入房地产估价，取得了良好的效果。

本文以特征价格模型为理论框架，从建筑因素、邻里因素和区位因素方面对二手房价的影响因素进行分析，并在此基础上选择了构建估价模型的特征变量指标体系。本文运用网络爬虫技术爬取安居客网站内上海市二手住宅挂牌数据，利用文本挖掘中文分词技术提取文字描述性变量中隐藏的信息，扩充了区位因素的变量。然后通过对原始数据的探索性分析对数据进行整理和清洗，获得适合建模的结构化数据。

在建模阶段，本文运用了随机森林方法构建了房地产估价模型，并调整模型的超参数不断优化模型的预测能力。此外，本文还使用传统的多元线性回归构建模型与随机森林模型进行对比，实证研究的结果证明随机森林能够显著提高估价的准确性。最后本文提出了研究中存在的几个局限，并针对不足之处提出了未来的研究展望。

关键词：随机森林；房地产估价；特征价格理论；网络爬虫；文本挖掘

Research on Real Estate Appraisal Model Based on Random Forest Algorithm
——A Case Study of Shanghai Second-hand House Market

Abstract

With the continuous development of China's economy and the increasing affluence of the people, the residential and investment needs of residential real estate are constantly rising, driving up housing prices tremendously. However, the supply of urban land is limited, resulting in people gradually turning their attention to the second-hand housing market. In first-tier cities, the second-hand housing transactions are less restricted by the government than the first-hand housing transactions, therefore better characterizing the market. The rising demand of second-hand housing is consistent with the demand of the real estate appraisal. No matter from the perspective of real estate transfer, lease, mortgage, or in terms of the property insurance, real estate tax, land acquisition and house demolition compensation, real estate appraisal is an indispensable process. However, many subjective factors exist in the three traditional methods of real estate valuation: Cost Approach, Income Approach and Market Comparison Approach. This implementation process over-relies on the appraisers' experience, which is very likely to cause the deviation between the estimated price and the actual price. Furthermore, it is hugely time-consuming and costs a lot of manpower. With the emergence of computer science and machine learning, many scholars have gradually introduced quantitative methods to real estate appraisal and achieved good results.

Taking the Hedonic Pricing Theory as the theoretical framework, this paper analyzes the influential factors of the second-hand housing price from the aspects of Structure, Neighborhoods and Location. Based on this, the paper establishes the characteristic variable index system to build the appraisal model. The web crawler is used to crawl the Shanghai second-hand residential listing data on Anjuke website, and the text mining technique is used to extract the hidden information in descriptive variables, which expands the variables in location factors. Through exploratory analysis on the original data and cleaning of the data, I make them structured data suitable for modeling.

In the modeling stage, the paper applies the Random Forest algorithm to construct the real estate appraisal model, and then adjusts the model's hyper-parameters to optimize the model's prediction ability. In addition, this paper also compares the random forest model with the traditional multiple linear regression model. Empirical results show that random forest model can significantly improve the accuracy of the appraisal result. In the end, the paper presents some limitations of this research and puts forward the prospect for future study.

Keywords: Random Forest, Real Estate Appraisal, Hedonic Pricing Theory, Web Scrawler, Text Mining

目录

第一章 绪论

1.1 研究背景

习近平主席在十九大中强调了“坚持房子是用来住的、不是用来炒的，加快建立多主体供给、多渠道保障、租购并举的住房制度，让全体人民住有所居。”十九大后，各地尤其是一二线城市纷纷落实新一轮的调控力度，确保房价维持稳定。在这样的宏观政治和经济环境之下，中国房地产一级市场和二级市场正在不断成熟，因此房地产估价的作用越来越明显。房地产估价就是专业房地产估价人员根据特定的估价目的，遵循公认的估价原则，按照严谨的估价程序，运用科学的估价方法，在对影响估价对象价值的因素进行综合分析的基础上，对估价对象在估价时点的价值进行估算和判定的活动[1]。关于房地产估价方法的研究最早可追述到 1884 年，距今已经有一百多年的历史，学者们从不同的角度提出了许多不同的方法，至今为止最为典型的有三种方法，即成本法、收益还原法和市场比较法，此外还有假设开发法、基准地价修正法、路线价法和长期趋势法。这些估价方法各有其适用的特定对象，针对不同的估价目的可以选择不同的估价方法。

然而，房地产估价的传统方法存在着许多主观人为因素导致估算价格与实际价格偏离，其中王克强、王红梅和姚玲珍指出市场比较法存在搜集选择什么样的交易案例作为可比交易实例以及如何对待估对象与可比交易实例的因素差异进行量化修正的问题，修正和调整一般难以采用数学公式或数学模型来量化，而主要依靠估价人员根据其掌握的扎实的估价理论知识、积累的丰富的估价经验和对可比实例、估价对象所在地房地产市场行情、交易习惯等的深入调查了解做出判断。如果估价人员不具有扎实的估价理论知识，没有丰富的估价实践经验，对可比实例、估价对象所在地的房地产市场行情和交易习惯等不够熟悉，就很难运用市场法得出正确的估价结果[1]。

随着计算机科学和机器学习算法的发展，研究学者们针对市场比较法、成本法和收益还原法存在的这些客观问题，借助计算机技术和机器学习研究问题、解决问题的方法进行多方面的探索与研究，取得了较好的成绩，并将房地产估价逐步推向定量化研究。

1.2 文献综述

1.2.1 国外研究

随机森林（Random Forest，RF）是由 Breiman 提出的一种基于决策树算法的组合分类算法[12]。该算法具有快速处理大样本数据，不容易产生过拟合，抗噪音能力强，能评价变量重要性等优点。Elena B.Pokryshevskaya 和 Evgeny A. Antipov 首次将随机森林算法与房地产价格评估相结合，通过算法原理阐述了随机森林抗噪声能力，同时与多元线性回归和神经网络的方法进行对比[13]。

1.2.2 国内研究

杨沐晞(2012)是国内第一位将随机森林模型运用于房地产估价中的学者。基于特征价格模型，其对广州市天河区的 49 个小区住宅进行随机森林建模，然后利用随机森林对变量的重要性进行排序，为投资者购买该区域二手房提供了评估体系。在构建模型的过程中，杨沐晞从其搜集而来的住宅数据中随机抽取了 298 套房源，利用随机森林进行建模，并与线性回归模型结果对比，从而证明了随机森林在估价上的优越性[2]。陈奕佳(2015)使用随机森林构建了北京市二手房的批量评估模型，并采用交叉验证方法对随机森林、Bagging、回归树、人工神经网络和多元线性回归 5 种机器学习算

法构建的模型进作比较，实证结果表明随机森林取得的效果最好。其还利用随机森林对各变量的重要性计算了重要性评分[3]。庞枫(2016)提出了基于随机森林的房地产估价方法，利用随机森林对特征指标进行排序，筛选出 12 个最重要的指标，然后将重庆市 6000 条二手房成交价用神经网络分成 10 个类别，用训练好的模型预测待估案例所属类别，在该类中用随机森林训练，最后对待估案例进行预测[4]。庞枫(2017)进一步细分评估区域以提高模型对待估二手房的预测精确度，其以市场比较法为基础，首先利用支持向量机分类算法搜寻与待估楼盘最相似的三个楼盘，增加了训练数据的可比性，其选取了 9 个特征变量，使用 R 语言编写随机森林批量评估模型[5]。时文静(2017)通过网络爬虫从链家网站收集了 3 万多条北京市二手房信息，选取了包括建筑特征、区位特征、小区环境等影响二手房价格的 38 个因素。首先用特征选择和 Lasso 回归两种方法进行初步的特征选择，剔除对评估模型影响不大的因素，减小模型的复杂度，共筛选出 33 个变量进行下一步的建模。而后进行模型的对比，二手房估价模型主要构建了 Lasso 模型、回归树、Boosting、 Bagging 以及随机森林。用五折交叉验证法对比 5 种模型的预测精度，结果显示随机森林模型误差最小，拟合效果最好。最后对随机森林模型进行了参数调整和模型优化，并对测试集数据进行预测，经检验模型拟合效果较好，预测结果有着较高的准确性[6]。

1.2.3 文献评述

从国内外关于基于随机森林的房地产估价的研究来看，该领域才刚刚兴起不到 10 年，目前研究的人并不多。此外，先前的学者所用的研究方法智能程度并不高，研究的区域偏小，部分变量数据的获取仍然依靠实地调研获得，这导致研究数据量偏少。先前学者为了尽可能挖掘出影响房价的特征因素引入了至少 20 个变量，由于机器学习算法普遍比较复杂，过少的观测值和过多的变量可能会导致模型出现过分拟合的现象，使模型的泛化性能降低。哪怕模型没有出现过拟合，模型也只适用于所研究的小区域，推广性和实用性不够。时文静的研究利用网络抓取技术爬取了大量数据，有效解决了模型泛化性和实用性的问题，并且引入特征工程技术筛选主要影响因子，极大程度提高了建模的智能化程度。

1.3 研究目的、意义和内容

1.3.1 研究目的和意义

传统的房地产估价需要较多依赖评估者的经验,而且估价过程比较复杂，评估结果也不好判别是否准确。为了改善这些问题，诸多学者将机器学习模型应用到房地产估价领域获得显著的效果。当前房地产价格与其影响因素之间仍无具体的函数表达式能够完美地阐述他们之间的关系。随机森林模型不需事先假设自变量与因变量之间的数学关系式，通过对训练数据进行自主学习和拟合，内部自动构建特征与价格之间的关系，这种关系可以是任意复杂的关系，因此对给定的数据特征能够预测出比较准确的结果。

如今，房子已不是单单满足居住的需求，慢慢的成为一种投资的方式。近年来房地产的投资属性致使房子的价格逐年攀升，价格是影响市场房源量和购房意愿的非常重要的因素之一。因此，房价的波动及走势以及众多的购房政策成为了市场参与者非常重视及关心的问题。房地产估价是房地产交易中非常受到交易双方重视的环节，也是现在众多房地产交易中介一项核心服务，对交易双方都有非常重要的参考价值。本文通过分析影响房产价格的指标体系，为二手房价格评估提供新的可操作方法。希望在住宅价格预测的实践中能提供新的思路。

1.3.2 研究内容

基于当前房地产估价行业现状和之前学者研究中所存在的问题，本文采用网络爬虫技术获取海量数据，通过文本挖掘技术进一步提取价格特征变量。最后利用随机森林构建房地产估价模型，并对模型参数进行调整以提高模型性能，与传统的多元线性回归进行对比，证明随机森林模型的优越性。

第一章绪论，介绍了本文的研究背景，通过对国内外相关文献的简单解读回溯了随机森林在房地产估价中运用的发展历程。然后介绍了本研究的目的、内容和意义。最后介绍了本文对比先前研究的创新点，并给出了本文的技术路线图。

第二章理论基础，介绍了本报告所用到的理论和技术，其中会着重介绍房地产特征价格理论，以及所用到的网络爬虫技术和随机森林算法。

第三章指标体系构建和数据爬取，讲解备选指标的选取和数据的获取，数据的分布状况以及数据的预处理方法，通过数据可视化分析上海市二手房价格的现状。

第四章基于随机森林的上海市二手房估价模型，讨论上海市二手房的估价模型构建，通过拟合优度、均方根误差、平均绝对误差、平均相对误差 4 个指标分析模型的预测能力，并调整模型参数以优化模型。之后利用随机森林对变量的重要性排序，并分析各个变量重要性产生的背后原因和现实意义。

第五章随机森林模型与传统回归模型的对比，利用传统多元线性回归构建估价模型，与随机森林模型进行对比。

第六章结论与展望，总结研究的成果，指出研究存在的不足并以此对未来研究提出展望。

1.4 创新点与技术路线

1.4.1 论文主要创新点

本研究与之前学者的研究的创新之处在于：1、利用网络爬虫技术从安居客网站爬取海量数据，去除数据获取过程中的主观人为因素，实现数据的全自动获取，因此大大增加了训练模型的样本数，能够有效避免模型泛化性能和实用性能不足的问题；2、首次尝试将文本挖掘技术利用到房地产价格特征选择和房地产估价模型的构建；3、在同一训练数据集上构建多元线性回归模型，并通过 4 项评价指标比较两个模型的准确度，从而证明随机森林的优越性。

1.4.1 技术路线图

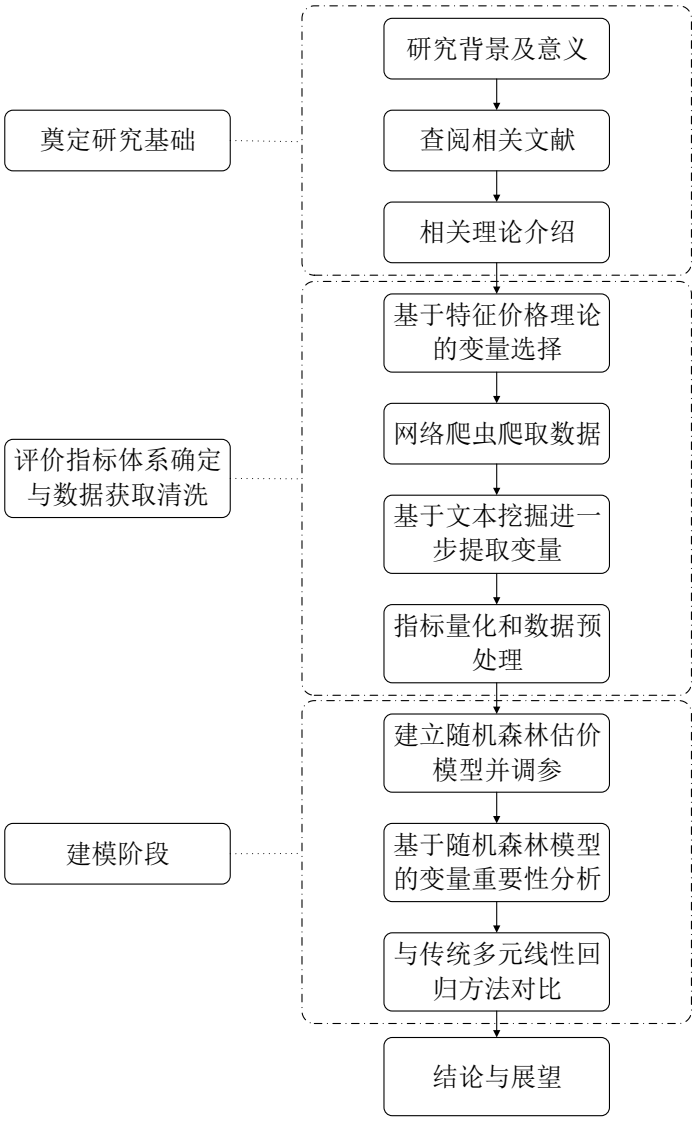

图 1-1 技术路线图

第二章 理论基础

2.1 网络爬虫

2.1.1 网络爬虫定义

网络爬虫是指能按照预先设定的规则自动搜集网页信息的程序或者脚本。它通过不断访问网页上的链接地址来寻找相关的网页，从一个或几个初试链接开始，访问链接地址，抓取网页内容，并在网页中寻找需要的链接地址，然后不断地访问需要的链接地址，抓取网页内容，直到把规则设定要抓取的网页内容全部抓取完为止[7]。

2.1.2 基于 R 语言 rvest 包和 stringr 包的网页信息抓取

虽然 R 语言是一门专长数据分析的语言，但是在网络爬虫领域也有其独特的优势，R 语言编写的爬虫代码相对简单整洁，对操作者的要求也相对较低，不需要具备很多 web 编程和网页设计的知识，许多初学者能够很快地利用 R 语言上手爬取网页数据。除此之外，R 语言可以轻松地处理百万数量级以下的数据， 由于其本身就是一个适用于统计分析和数据可视化的功能强大的工具，通过 R 语言爬取数据之后可以直接进行统计分析和数据挖掘，省去了数据的导入导出的步骤，更加直接方便。

因此本文运用R语言中的rvest包和stringr包来实现抓取网页信息数据。运用该包中read_html()、html_nodes()、html_attr()和 html_text()四个函数与 SelectorGadget 工具相配合。SelectorGadget 是谷歌浏览器的一个扩展程序，可以通过点选网页模块的方式输出待提取内容的 CSS，输出结果为 html_nodes()函数的一个重要参数。首先使用 read_html()函数抓取整个网页的原始 HTML 程式码；再使用 html_nodes()函数从整个网页的元素中选出由 SelectorGadget 获取 CSS 路径的信息；使用 html_attr()函数可以读取节点容器中包含的 href 属性，读取超链接网址；最后使用 html_text()函数将 HTML 程式码中的文字资料提取出来得到我们所需要的数据。处理大量网页源码相似的网页时，可以根据网页 URL 的规则，利用 for 循环函数实现多网页的信息抓取工作[8]。

stringr 包是 R 语言中一个简单易用的处理字符串的程序包。使用该包可以非常方便地对搜集来的字符串信息进行截取、提取、拼接、替换、分割等操作，在该包提供函数基础上，利用正则表达式可以自定义出我们想要达到的数据效果，写出来的代码不仅实用而且美观简洁。

2.2 随机森林

2.2.1 随机森林的概念

随机森林(Random Forest)是一种从基于 Bagging 算法改进而来的一种集成算法，该算法由许多棵完全分支的决策树组合而成。其基本思想是通过自助法（bootstrap）和重采样(resampling)，从数量为 N 的原始数据集中有放回地随机抽取 N 个样本组成新的训练数据集，重复这一过程生成的多个新训练集分别生成许多棵决策树组成“森林”，用于测试集数据的分类或回归。随机森林与一般决策树存在两点差异：第一，随机森林使每一棵分类回归树充分生长，不再对其进行剪枝。第二，在决策树分枝节点选择进行分枝的变量时，不再是将所有可能的划分条件一一进行比较，而是在所有变量中随机选取一定数量的自变量，根据节点不纯度最小原则选择最优变量进行节点的分枝。

2.2.2 随机森林回归算法实现原理

第一，从数据量为 N 的数据集中，用 bootstrap 方法从总体中有放回地随机抽取 N 个样本组成训练集，对训练集每个样本建立回归树，根据概率统计训练集的样本约占总体的 68%。测试集由剩余 32%的样本构成，这些未被抽中的数据被称为袋外数据（out-of-bag，OOB）。
第二，设原始数据集中有 M 个特征因素，对每棵树的每个节点处都随机抽取 m 个变量（R 中默认 m=M/3）作为备选分枝变量，按照节点不纯度最小原则，从这随机选取的 n 个特征里选择一个使该节点不纯度达到最小的特征作为该节点的分枝特征，不对该树进行剪枝操作使其充分生长。
第三，利用生成的多棵决策树对待估对象进行回归预测，将每一棵决策树的预测结果进行平均，得到待估向量最终的预测结果。
第四，用每次随机抽样所生成的袋外数据对相应的决策树进行预测，汇总平均各棵决策树的预测结果，可以得出整个随机森林的预测准确率。模型的预测效果可以用 OOB 均方误差和拟合优度表现：

$$\mathrm{MSE}_{OOB} = \frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k} \tag{2-1}$$

$$R_{RF}^2 = 1 - \frac{\mathrm{MSE}_{OOB}}{\hat{\sigma}^2} \tag{2-2}$$

上式中 k 代表每棵树袋外数据的样本数量，y_i代表袋外数据中因变量的实际值，$\hat{y}_i$代表用随机森林回归模型得到的预测结果，$\hat{\sigma}^2$代表袋外数据预测值$\hat{y}_i$的方差[5]。

2.2.3 基于随机森林的重要性排序

传统线性回归模型中对于特征变量的重要性度量主要是看对应于特征变量的回归系数的大小，回归系数越大，该自变量对因变量的影响就越大。随机森林本质上属于非线性模型，训练过程中不产生固定的变量系数。对随机森林模型中变量重要性的评估主要依赖于在对众多特征中的某一个特征加入噪声，对比随机森林预测结果的显著性是否有变化，主要通过均方误差平均减小值(%Inc MSE)来反映自变量的重要性程度，计算过程为：
（1）在一个随机森林中，利用每棵树的袋外数据对每一棵树计算均方误差，得到的每棵决策树对应的均方误差记为$MSE_1, MSE_2, \ldots, MSE_n$。
（2）在每一个基分类器决策树对应的袋外数据中人为加入噪声干扰，通过特征变量的随机置换形成 n 个新的袋外数据样本集，用原先建立的随机森林模型对新生成的测试集再次进行预测，得到均方误差矩阵，记为：

$$\begin{bmatrix} MSE_{11} & \cdots & MSE_{1n} \\ \vdots & \ddots & \vdots \\ MSE_{N1} & \cdots & MSE_{Nn} \end{bmatrix} \tag{2-3}$$

（3）将进行置换后的袋外数据均方误差与初始袋外数据均方误差进行相减，平均并除以标准误差就得到了特征变量X_i的精度平均减小值，即变量的重要性评分。计算公式如下：

$$\%\mathrm{IncMSE}(X_i) = \left(\frac{1}{n}\sum_{j=1}^{n}\left(MSE_{ij} - MSE_j\right)\right)/S_E, (1 \le n \le \mathrm{k}) \tag{2-4}$$

其中，特征变量受噪声干扰后 MSE 增加越多，说明该变量对因变量的解释能力越高[2]。

2.3 特征价格理论

2.3.1 特征价格理论的概念

特征价格理论（Hedonic Pricing Theory）的核心理念是假设商品价格可以通过公式 P=f(x)决定，其中 x 为一系列可以影响商品定价的因素。理论上来说任何一个参数的改变都会影响商品的价格，此时 x 要求囊括客观现实中所有影响价格的因素，且所有消费者的偏好和收入水平相近。

特征价格理论建立在以下四条假设之上的：（1）市场是完全竞争的；（2）市场上有大量的差异性产品；（3）产品的价值是由其属性对效用的贡献大小决定的，且可以量化这些属性；（4）供求双方追求利润最大化，需求者寻求效用最大化[3]。

1967 年，Ridker 首次将特征价格理论引入房地产市场的研究。近年来房地产行业的飞速发展促使特征价格理论成为了房地产领域的热点研究话题。Ridker 将房地产看做一种商品，其效用来源于房地产自身的一些客观特征，房子的所有特征所带来的效用总和即反映为房地产的价格。

美国学者 Butler 提出了影响房地产价格的三大特征因素：Location（区位特征）、Structure（建筑特征）以及 Neighborhoods（邻里环境）。区位特征指的是住宅小区位于哪个区域，以及该区域具备的相关特征，包括：所在区域、交通便捷度、地铁、到市中心距离。建筑特征指的是房地产本身的客观性质，包括：户型、面积、房龄、建筑结构、装修、所在楼层、卧室个数。邻里环境指房屋所在小区的自然环境、人文环境和治安管理等方面状况，包括：绿化率、物业费、停车位。

依据 Butler 的理论，住宅价格用公式可以表示为：

$$P = f(L,S,N) \tag{2-5}$$

在该公式中，L、S、N 分别代表区位特征、建筑特征、邻里环境，通过回归分析可以获得模型的估计参数，进而获得房地产期望价格的函数表达式。当满足特征价格理论的四条假设并且将所有特征纳入模型中时，总效用即预期价格将会无限趋向于实际房价。

2.3.2 特征价格理论的优点

相比于三大房地产估价方法，基于特征价格理论的房地产估价方法有着诸多优点。第一，相比于收益还原法，该理论无需对收益性房地产的未来收益进行预估和折现，避免了利率风险的影响。第二，该理论相比于成本法，该理论直接采用房屋年龄替代了折旧处理过程，使得结果更准确。第三，该理论是对市场比较法的有效改进。特征价格估价模型则弥补了市场比较法在考虑影响房价因素方面缺乏系统全面的分析这一缺陷，它通过运用大量的样本和加入足够多的特征因素，能够更精准地评估价格。第四，该理论支持房地产的批量评估，省去了市场比较法的比准价格修正工作，节约了时间和人力成本[11]。

第三章 指标体系构建和数据爬取

选取房地产估价模型的指标是非常重要的一步，因为特征价格理论假设房地产价格是由一系列影响因素决定的，即由一系列自变量通过映射关系得出。因此，发掘出解释性强的解释因子能够显著提高模型的预测性能，遗漏了重要的因子则会极大降低模型准确率。可以说，指标的选取决定了模型性能的上限。

3.1 因变量的选取

由于部分城市推行一手房限价措施，政府对新房进行限价和限售，导致部分一线城市出现一二手房价格倒挂现象，新房价格低于周边二手房的价格。如图 3-1 所示，上海市在 2018 年 9 月前新房价格显著低于二手房价格。这一违反经济学原理的价格倒挂现象是由于政府的管控造成的，不符合特征价格理论的假设。本文选用二手房的挂牌价格作为因变量正是基于这一考虑。相对新房，二手房市场更加接近完全竞争市场，更加符合供给需求决定价格的机制。

此外，由于二手房成交价格属于商业机密，很难获取，而挂牌价格公开透明容易获取，两者有显著的正相关关系，因此本文选用二手房中介平台的挂牌价格而不是实际成交价格。

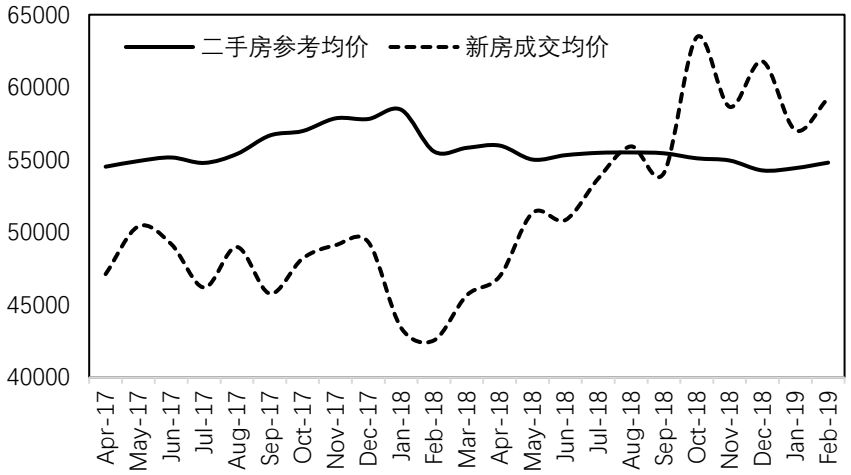

图 3-1 上海市二手房与新房价格

3.2 二手房价格候选指标的选取和量化

本文参考学者之前的文献以及链家、安居客、搜房网等房产中介网站后，根据房地产特征价格的框架，提出了一套获得性强、范围较广的候选指标体系，如表 3-1 所示，包含三大类共 21 个指标。通过比较了链家、安居客和搜房网的公布信息和网页结构，笔者认为安居客网站的数据最为全面，网页结构最为规律，因此本文选择爬取安居客网站公布的数据进行数据建模。

表 3-1 价格评估指标体系

大类	指标		描述	数据来源
	变量名	中文含义		
	unitprice	房屋单价	房屋挂牌单价(元/㎡)	安居客网站获取
建筑因素	area	房屋面积	房屋建筑面积(㎡)	安居客网站获取
	year	房龄	截至 2019 年的房屋年龄	安居客网站数据修正
	direction	朝向	朝向包含南-1，其他-0	安居客网站数据量化

	decoration	装修程度	豪华装修-4 精装修-3 简单装修-2 毛坯-1	安居客网站数据量化
	bedroom	卧室数量	房屋包含卧室数量	安居客网站获取
	hall	厅数量	房屋包含厅数量	安居客网站获取
	lavatory	厕所数量	房屋包含厕所数量	安居客网站获取
	storey	所在楼层	分为低层、中层、高层	安居客网站获取
	elevator	电梯	有电梯-1 无电梯-0	安居客网站数据量化
	property	产权性质	商品房-1 其他类型-0	安居客网站数据量化
邻里因素	totalstorey	总楼层数	房屋所在楼总层数	安居客网站获取
	far	容积率	小区容积率	安居客网站获取
	parkinglot	停车位	小区停车位数量	安居客网站获取
	greening	绿化率	小区绿化率	安居客网站获取
	fee	物业费	小区绿化率(元/㎡)	安居客网站获取
区位因素	admdist	行政区	所在行政区	安居客网站获取
	secdist	细分区域	所在细分区域二手房均价(元/㎡)	安居客网站数据计算
	subway	地铁	周边是否有地铁，有-1，无-0	文本挖掘获取
	school	学区房	周边是否有学校，有-1 无-0	文本挖掘获取
	business	商业设施	周边是否有商业设施，有-1 无-0	文本挖掘获取
	amenity	公园医院	周边是否有公园或者医院，有-1 无-0	文本挖掘获取

安居客并没有在其网站上挂出所有的二手房源，按行政区搜索每个区最多能够获得该区 3000 条二手房房源，而且相同的房源可能会在不同的页码重复出现，因此可能被多次爬取。笔者运用 R 语言编写了网络爬虫代码，获取标题、网址、房屋单价、房屋面积、房龄、朝向、装修程度、卧室数量、厅数量、厕所数量、所在楼层、总楼层数、电梯、产权性质、容积率、停车位、绿化率、物业费、行政区、细分区域、地址、核心卖点、专家点评 23 个特征，在 2019 年 3 月 17 日爬取了当日安居客网站上 41261 条上海市二手房源截面数据，去掉重复的房源后一共有 41261 条。

其中，房屋单价、房屋面积、房龄、卧室数量、厅数量、厕所数量、总楼层数、容积率、停车位数量、绿化率、物业费这些指标原本就是数值型变量，无需量化。朝向、装修程度、所在楼层、电梯、产权性质、行政区、细分区域这些指标是类别变量。在这之中，（1）朝向，中国人更倾向于购买面朝南的房子，因此朝向带有南（南，西南，东南）转化为 1，其他朝向为 0；（2）装修程度，装修程度可以按照由复杂到简单的顺序分别将豪华装修、精装修、简单装修和毛坯转化为 4、3、2、1；（3）电梯，有电梯的房子能够省去爬楼梯的劳累，所以有电梯转化为 1，无电梯为 0；（4）产权性质，商品房的产权最完整，而商住两用房、经济适用房、公房等等其他类别的房子都有产权的残缺或限制，因此商品房转换为 1，其他产权类别转化为 0；（5）细分区域，安居客中提供了房源的细分区域信息，细分区域可以反映出详细的地段信息，价值很高，但是由于细分区域属于类别变量，类别数量过多导致计算难度激增，因此本文用该细分区域的平均二手房价来加以量化。原始数据中所在楼层分为低层、中层和高层，但是由于人们对于高中低层的偏好是否与楼层呈现相关的关系不得而知，本文将这一变量以类别变量保留，同时本文将保留行政区指标不变。随机森林算基于的决策树能够轻而易举地处理这些类别变量。最后，本文将通过标题、核心卖点和专家点评 3 个文字表属性的特征中运用文本挖掘技术，以获得地铁、学区房、商业配套设施和公园医院 4 个是否型指标。

3.3 基于文本挖掘的区位特征提取

图 3-2 房源文字描述信息词云

由于样本量巨大，网页中又没有系统性地给出房源区位配套的信息，于是本文将创新性地首次运用文本挖掘技术批量地从每个房源的标题、核心卖点和专家点评 3 个文字表述型属性中提取地铁、学区、商业配套、公园医院 4 个变量。地铁、学区、周边商业和公园医院是人们考虑买房的重要因素，关系着居住的方便程度，安居客在文字性描述中大概率会提及这些信息，以便吸引购房者的目光。于是本文运用 R 语言的 jiebaR 包进行中文分词，提取与这 3 个变量相关的话术。根据分词结果画出的词云结果如图 3-2 所示，字越大说明该词语更频繁地被使用。从词云中可以得知以安居客为代表的房地产中介在宣传房源时经常强调小区环境、房屋户型、周边配套、交通、学校、朝向采光等方面的信息，从侧面反映了这些因素往往是人们选择评估房子的重要因素。根据分词和词云的结果，笔者提取了表 3-2 所示的信号词语。标题、核心卖点和专家点评三部分只要出现了这些信号词语，便可判断为是，如果都没出现则判断为否。通过文本挖掘的中文分词技术，我们将地铁、学区房、商业设施和公园医院 4 个变量补充进数据集中。

表 3-2 变量提取信号词

地铁	号线、轨道交通、地铁、地铁口、双轨、轨交、轻轨
学区	小学、上学、学校、中学、学区、附小、名校、读书、教育、幼儿园、校区、上外、复旦、交大、就读、同济
商业配套	万达、商圈、商业、菜市场、菜场、新天地、家乐福、百联、联华、超市、商场、华润、购物
公园医院	公园、广场、医院、滨江、长兴岛、苏州河、佘山、河畔

3.5 数据清洗和预处理

3.5.1 数据探索性分析

笔者首先对安居客网站爬取的数据进行描述性统计分析。图 3-3 展示了上海市各个行政区的平均房价，其中最高的黄浦区二手房均价超过了 90000 元每平方米，最低的崇明区均价则不足 15000 元每平方米。各行政区的二手房均价基本反映了上海市越靠近市中心房价越高的现象。

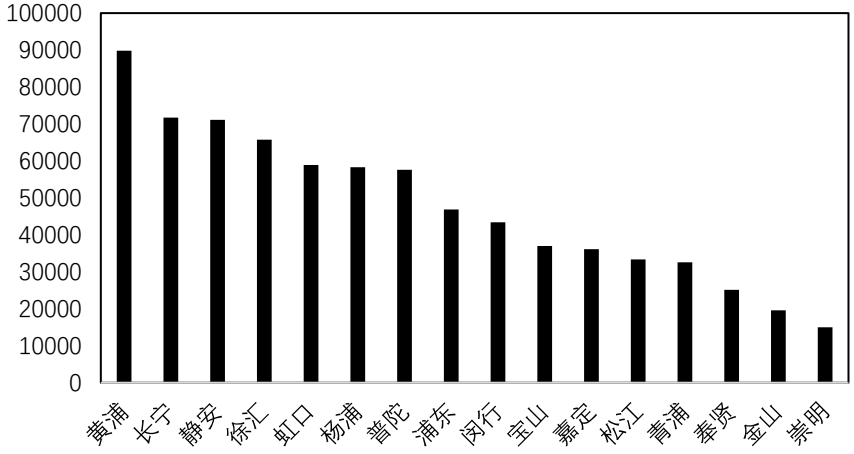

图 3-3 上海市各行政区二手房均价

图 3-4 展示了上海市房价最高的 20 个次级区域，有 7 个次级区域的二手房均价超过了 10 万元每平米，这些区域都位于静安区、黄浦区、徐汇区的老牌核心商圈地带。再往下则出现了一些普陀区、长宁区和浦东新区的核心商圈地带，这些“富人区”地带二手房均价都已经超过了 80000 元每平方米。

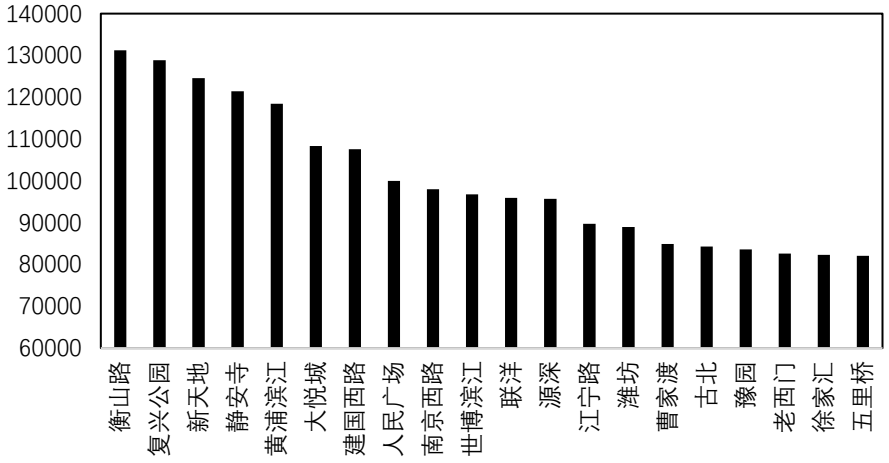

图 3-4 上海市细分区域二手房均价前十

图 3-5 展示了上海市二手房房价的整体分布情况直方图，蓝色曲线描绘了分布的概率密度曲线。从整体上看上海市二手房价分布接近正态分布形态，平均价为 48254 元/㎡，但是略微正偏，偏度为 1.25。数据说明二手房市场低于平均价的房价更加集中，而高于平均价的房价更加离散，最高房价甚至超过了 28 万元每平方米。数据的峰度为 6.81，大于正态分布的 3，说明二手房房价非常集中地分布在峰值 35000-50000 内。房价分布情况与收入分布情况基本相匹配，最多的中等收入人群能够对应最多的中等房价房源。少部分非常富有的人能够负担的起非常昂贵的住房，这导致了高房价的房源离散程度明显更高一些。

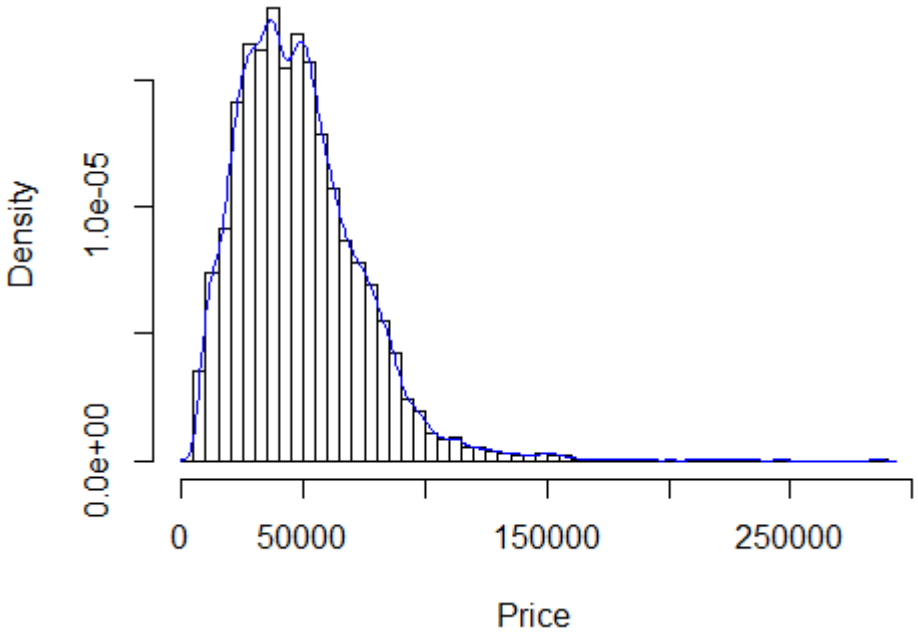

图 3-5 上海市二手房房价分布情况

3.5.2 缺失值处理

由于部分老小区面临着数据缺失的情况或者安居客没有现成数据的情况，数据集存在着不少缺失值。图 3-6 展示了安居客网站获取数据的缺失情况（深黑色为缺失值），缺失值大约占了全部数据量的 4%。缺失值主要分布在容积率、停车位、绿化率、房屋所在楼层、小区物业费、房龄和房屋产权这几个指标中。在本文中考虑到原始数据集数据量已经非常大而缺失值并不多这一条件，不进行缺失值补差，直接将含有缺失值的二手房删除。经过这一处理后得到了 16935 条无缺失值的二手房数据。

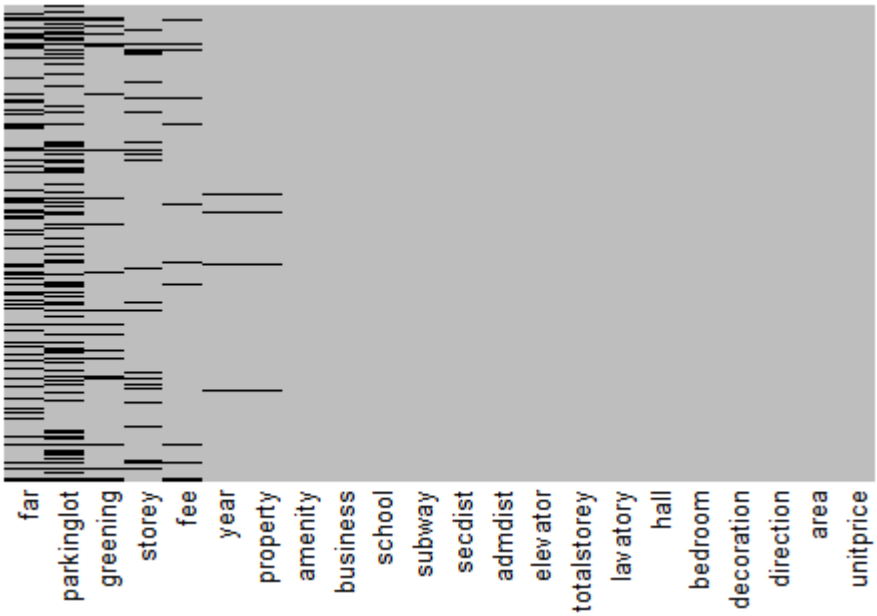

图 3-6 原始数据缺失情况

第四章 基于随机森林的上海二手房估价模型

4.1 研究思路

本章利用上一章提取的上海市二手房数据集和确定的指标体系建立随机森林模型，并评价模型的预测能力。本文将 16935 条二手房观测值按照 3：1 的比例随机划分为训练数据集和测试数据集，测试数据集含有 12703 条观测值，测试数据集含有 4232 条观测值。测试集用于训练模型，测试集用于判断模型对未知数据的预测能力。本文将用拟合优度、均方根误差、平均绝对误差和平均相对误差作为模型预测精确能力的测度指标，四个指标的计算公式如下[10]：

$$\text{拟合优度 } R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{nVar(Y_i)} \tag{4-1}$$

$$\text{均方根误差 RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{4-2}$$

$$\text{平均绝对误差 MAE} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{4-3}$$

$$\text{平均相对误差} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|/Y_i \tag{4-4}$$

公式中X_i代表预测房价，Y_i代表实际房价，n 代表二手房数量。调整森林模型的 mtry 参数和 ntree 参数优化模型，使得模型能够取得最高的精确度。通过随机森林模型可以计算各个变量的重要性，并分析原因。

4.2 模型构建和调节参数

由于随机森林模型构建需要巨大数量的计算，本文运用 R 软件的 randomForest 包构建随机森林模型。randomForest 包是由 Freiman 和 Cutler 创建的一个 R 包，适用于随机森林的分类或回归问题。我们利用训练数据集可以马上构建出一个随机森林模型，模型的决策树数量 ntree 默认为 500 棵，备选分枝变量数 mtry 默认为变量个数除以 3，在这里就是 7 个。该模型在训练数据集的表现如表 4-1 所示。

表 4-1 模型在训练集上的表现

拟合优度 R^2	99.02%
均方根误差(元/㎡) RMSE	2470.345
平均绝对误差(元/㎡) MAE	1562.698
平均相对误差	3.27%

模型的平均相对误差在训练集上为 3.27%，意味着预测值平均偏离实际值 3.27%，对应的平均绝对误差为 1563 元/㎡。表 4-2 展示了训练集前 10 套二手房预测房价和实际房价的数据。在这 10 套房中，第 1 套房几乎预测仅差了不到 1 元，其他绝大多数二手房相对误差也在 4%之内，精确度比较高。

表 4-2 训练集前 10 套房源预测信息

序号	挂牌价格(元/㎡)	预测价格(元/㎡)	绝对误差(元/㎡)	相对误差
1	60894	60893.17	0.8291	0.001%
2	84386	82211.67	2174.3266	2.577%
3	48864	48389.24	474.7558	0.972%
4	31947	33155.92	1208.9154	3.784%
5	45754	46150.52	396.5164	0.867%
6	67678	68729.86	1051.8595	1.554%
7	72941	67735.8	5205.1968	7.136%
8	32839	33760.29	921.2856	2.805%
9	75000	72584.58	2415.4248	3.221%
10	28289	31559.92	3270.9157	11.563%

我们将模型运用于测试数据集，测试数据集并没参与模型的训练，对于模型来说是前所未见的数据，这时的结果更能体现出模型在实际运用中能达到的精确程度，表 4-3 展示了模型在测试数据集的精确性测量。较之训练集随机森林模型在测试数据集上的各个指标均有所下降，但是仍然保持较高的准确率。当预测未知数据时，模型预测数据与实际数据的平均相对误差也仅仅为 6.83%，平均偏差 3306 元/㎡较之传统人工估价来说已经非常理想。

表 4-3 模型在测试集上的表现

拟合优度 R^2	95.58%
标准误差 RMSE	5313.136
平均绝对误差 MAE	3305.811
平均相对误差	6.83%

从个体数据上看，表 4-4 展示了测试集数据前十条的预测数据和实际数据，第二套房预测得十分精准，但是第 9 套房误差率却高达 32%，综合来看误差率有高有低，没有训练集理想。房地产评估可以接受的误差程度最高可以达到±20%的程度，一般最好不要超过±10%，所以测试集绝大部分二手房的价格预测能够被行业标准接受。

表 4-4 测试集前 10 套房源预测信息

序号	挂牌价格(元/㎡)	预测价格(元/㎡)	绝对误差(元/㎡)	相对误差
1	29897	36145.79	6248.78992	20.901%
2	35510	35467.88	42.11507	0.119%
3	45934	43339.79	2594.21363	5.648%
4	48024	49683.94	1659.93545	3.456%
5	53610	47782.71	5827.2876	10.870%
6	44667	47834.14	3167.13933	7.091%
7	88429	72608.7	15820.302	17.890%
8	37871	39113.55	1242.54617	3.281%
9	39655	52576.96	12921.95788	32.586%
10	35459	42607.41	7148.41431	20.160%

随机森林模型主要有两个超参数可以调节，一个是随机森林包含决策树的数量 ntree，另一个是决策树每次分枝备选变量的数量 mtry，因此笔者试图通过调节这两个参数进一步优化模型。在先前所述的模型中，笔者采用了 R 软件默认的 500 棵树和 7 个备选变量建立了第一个模型。图 4-1 展示了随着随机森林模型中树数量增加 OOB 均方误差的变化情况，当树的数量较少时，均方误差较大，随着决策树数量增加，均方误差也迅速下降，当 ntree 为 409 时，均方误差为 26659015，达到了 500 棵树中的最小，409 后均方误差略微上升但变化很微弱。由此可见当树的规模达到一定规模，

均方误差也会趋近极限值。

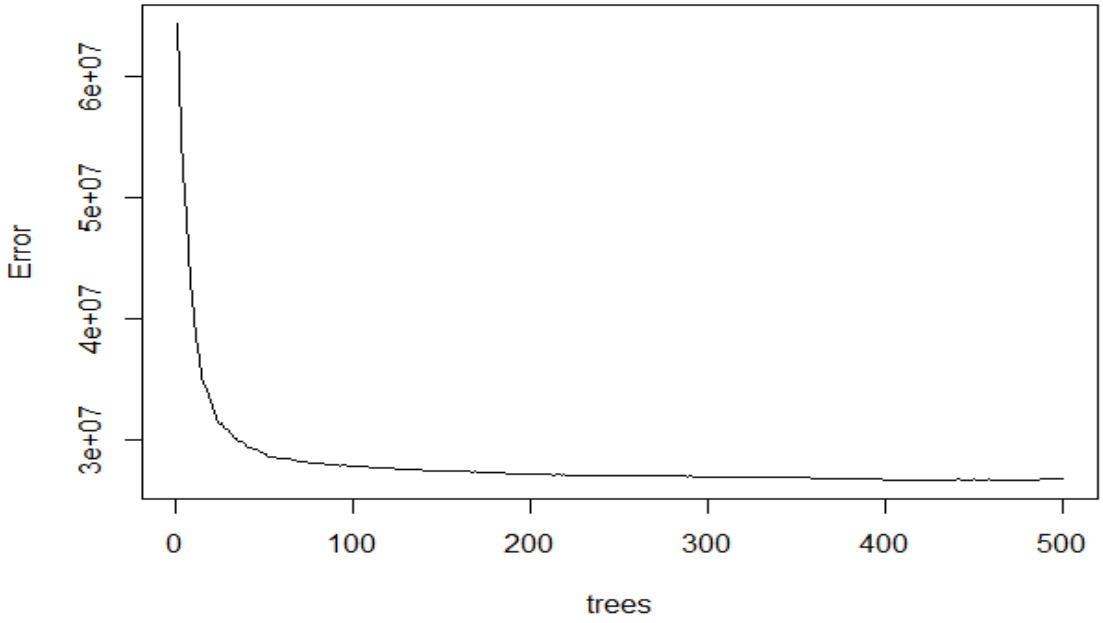

图 4-1 模型中树的数量与 OOB 均方误差关系图

随机森林中最为重要参数是 mtry 参数，即每一棵决策树分支时备选变量的的个数，不同的 mtry 值造成的模型差异相当可观。在 randomForest 包中 mtry 的默认值是自变量总数除以 3，在调整该参数时我们将决策树数量定为 400，从 2 到 20 选择最优 mtry 参数。如图 4-所示，横轴为 mtry 参数的值，左边纵轴为模型在训练集（实线）的平均相对误差，右边纵轴为模型在测试集（虚线）的平均相对误差。一般来说，平均相对误差越低拟合优度越高。由图 4-2 可知当 mtry 的数量增加时，模型在训练集和测试集上的拟合程度都呈现上升的趋势，在 mtry 值到达 8 之后误差率趋于平缓，当 mtry=18 时，模型在测试数据集上取得了最小的误差率。由于我们更加看重模型的泛化能力，因此本文选择 mtry=18 作为最优的每次分支备选变量数。

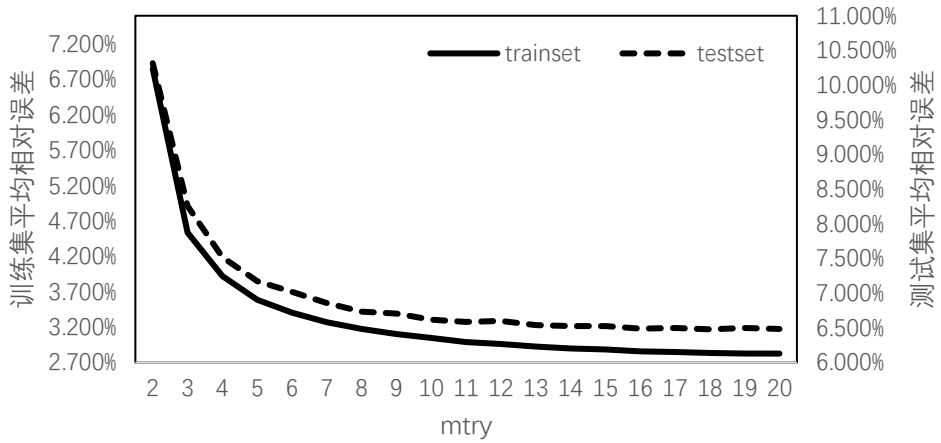

图 4-2 mtry 调参示意图

下面继续以平均相对误差率为指标在较宽的指标取值范围上选择 ntree 的值。图 4-3 为 ntree 值从 100 到 1000 时模型在训练集和测试集上的平均相对误差变化。图 4-3 的标注和图 4-2 相似。由图可知当树的数量增加时，模型在训练集和测试集上的拟合程度都呈现上升的趋势，当树的数目为 400 时测试集的平均相对误差达到了最小值，400 之后误差反而回升了一些，而在训练集上则继续下降，因此过高的决策树数量会造成轻微的“过度拟合”现象。由于我们更加看重模型的泛化能力，因

此本文选择 ntree=400 为最优决策树数量。

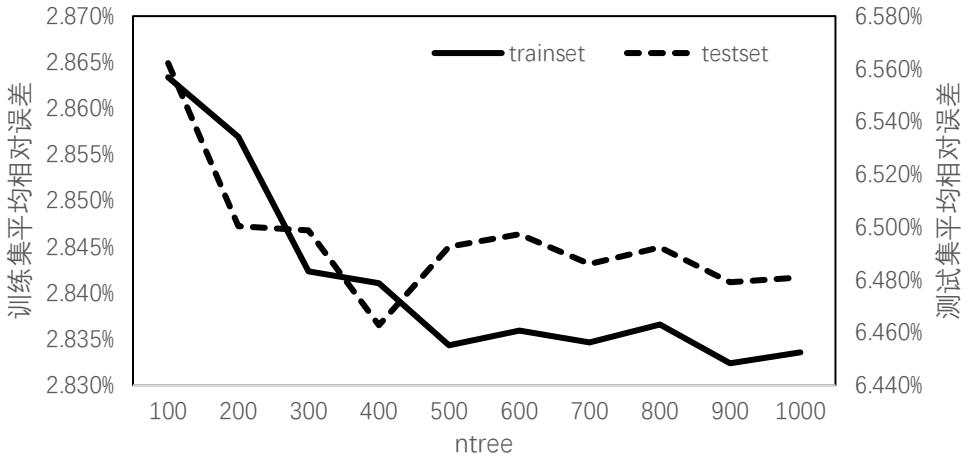

图 4-3 ntree 调参示意图

由图 4-2 和图 4-3 可知，当模型选用 mtry=18 和 ntree=400 这两个参数时模型在测试集展现了最好的预测性能，并且避免了过拟合的现象。因此本文选择 mtry=18 和 ntree=400 为最终使用的模型参数。最终模型的准确性测度如表 4-5 所示，通过调参工作我们将训练集平均相对误差从 3.27% 优化到了 2.84%，测试集平均相对误差从 6.83%优化到了 6.47%。

表 4-5 优化后模型的预测性能

	训练集	测试集
拟合优度 R^2	99.18%	95.68%
标准误差 RMSE	2252.309	5252.551
平均绝对误差 MAE	1375.444	3172.846
平均相对误差	2.84%	6.47%

4.3 变量重要性排序

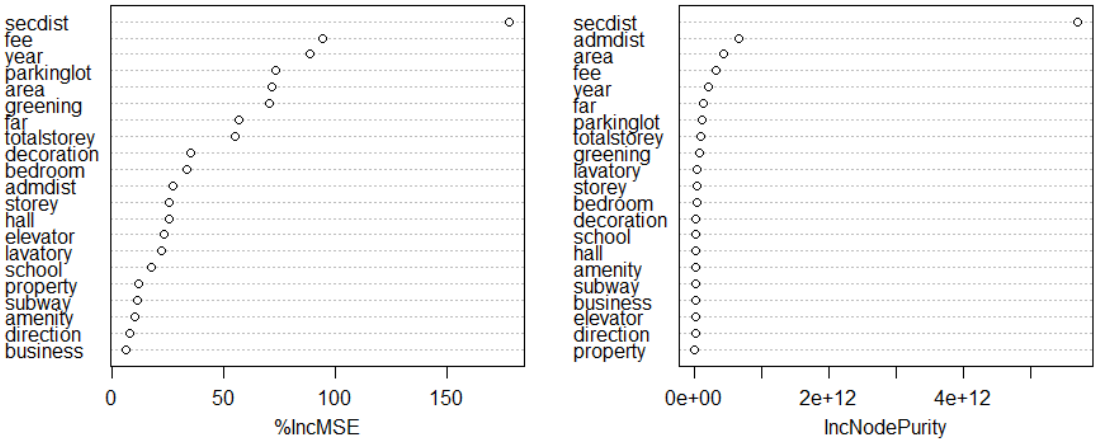

图 4-4 变量重要性排序

图 4-4 展示了影响房价的各个变量重要性排名。R 软件 randomForest 包内置了两种计算变量重要性的算法。%IncMSE 是对变量增加噪声干扰后均方误差的增加值，MSE 增加得越多，说明变量越重要。IncNodePurity 是节点纯度的增加值，也就是节点不纯度的减少值，通过决策树利用该变量分枝前后不纯度的降低平均而来的指标。通常来说 IncNodePurity 更倾向于赋予类别更多的变量更高的重要性，因此本文单就%IncMSE 的重要性排序进行讨论。

由图中的排名我们可以发现细分区域、物业费、房龄、停车场、建筑面积、绿化率、容积率、总楼层数、装修状况这几个因素最为重要。细分区域反映了房子的区位因素，根据市场有效理论来说，如果房子所处地段周边房子普遍较贵该房子价格也不可能显著的便宜。物业费往往体现着小区物业管理的质量，而物业服务质量又是购房者考虑的重要因素，会影响房价。房龄反映了房子的新旧程度，影响着购房者的居住舒适度，是房屋折旧程度的数值表现，同时房龄越高房屋剩余年限越少，可居住的时间也越少，因此房龄因素很大程度影响了房价。如今小汽车已经成为每个家庭的必备品，有些富裕的家庭甚至拥有多辆汽车，因此停车场的数量会影响到人们对房价的评估，购房者往往不愿意购买停车位紧张或停车不便的小区。中国的大城市包括上海深受雾霾困扰，居住地空气质量成为了购房者诉求之一，同时小区里的绿地为居住者提供了方便的散步去处，因此绿化率越高的小区能够售出更高的价格。房屋建筑面积越高的房子更可能是高档住宅，相反，面积越小的房子往往单价越低。容积率越高的小区平均每户分担的成本越低，同时容积率高的小区居住舒适度随之下降，所以价格往往越低。总楼层数与容积率等因素相关联，也会影响房价。房子所在行政区直接决定了房子的区位，间接反映了房子的交通通达度、配套设施、市中心距离等等方面，同样影响房价。

在这些变量中，最重要的是细分区域均价，也就是我们常说的地段，这是房地产不可移动性产生的独特之处，好的地段直接决定了房价的一般水平，这是其他任何要素都替代不了的。停车场、物业费、绿化率、容积率和总楼层数都属于邻里因素，这些因素合并起来也占据了很高比重。重要的建筑因素包括房龄和建筑面积，这两个因素也是购房者关注的重点。通过数据可以得出区位因素主导了房地产的价格，同时随着人们对环境有了更高的要求，邻里因素逐渐成为一项非常重要的考量标准，房龄和建筑面积两个建筑因素也会对房价造成一定的影响。

无论对于房地产估价师、购房者还是卖房者，房地产价格特征变量重要性排名都能提供重要的参考，当需要对房地产进行估价时，这些因素都是应该是要着重考虑的。

第五章 随机森林模型与传统回归模型的对比

本文采用了房地产估价中最为常用的多元线性回归建立模型作为对比，多元线性回归的基本形式为：

$$\hat{Y} = w_0 + \sum_{i=1}^{n} w_i X_i \qquad 5\text{-}1$$

本文运用 R 语言构建多元回归模型，利用向前向后逐步回归法剔除不显著的变量。R 软件以 AIC 值为指标计算每剔除或增加一个变量 AIC 值的变化，最终使得模型拥有最小的 AIC 值。AIC 值是赤池信息准则，是衡量统计模型拟合优良性的一种标准，在一般的情况下，AIC 可以表示为：AIC=2k+nln(SSR/n)，其中：k 是参数的数量，n 为观测值数量，SSR 为残差平方和[14]。AIC 鼓励数据更好地拟合但是尽量避免出现过度拟合的情况。所以优先考虑的模型应是使得 AIC 值最小的那一个。

经过计算，R 程序剔除了朝向、配套设施、地铁、厕所数量、产权类型和商业设施变量，进入最终模型构建的变量有 15 个。剔除的变量恰为最初回归模型中不显著的变量。表 5-1 展示了多元线性模型的参数和显著性检验，虽然模型非常显著，但是拟合优度仅为 86.57%，远不及随机森林在训练集取得的 99.02%拟合优度。表 5-2 展示了经过逐步回归后模型变量的回归系数以及显著性检验。除了所在楼层和所在区域这两个类别变量处理为哑变量后出现了个别哑变量不显著的情况，其他变量都至少在 90%的置信度下显著。

表 5-1 线性回归模型参数及检验

Multiple R-squared	0.8657
Adjusted R-squared	0.8654
F-statistic	2724
p-value	< 2.2e-16

表 5-2 线性回归的回归系数及检验

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1011	806.2	1.254	0.20971
area ***	101.5	3.042	33.358	< 2e-16
year ***	-153.6	16.38	-9.379	< 2e-16
decoration ***	1296	121.9	10.639	< 2e-16
bedroom ***	-2503	155.8	-16.063	< 2e-16
hall **	-601	209.6	-2.868	0.004141
storey 高层	-1.405	212.3	-0.007	0.994719
storey 中层 *	-412	199.2	-2.068	0.038644
totalstorey ***	152.4	14.97	10.178	< 2e-16
elevator ***	-1385	228.5	-6.065	1.36E-09
far .	-244.3	148	-1.651	0.098791
parkinglot ***	1.062	0.1842	5.764	8.39E-09
greening ***	6086	1309	4.651	3.34E-06
fee ***	1426	71.24	20.012	< 2e-16
admdist 崇明 ***	-3463	587.4	-5.894	3.86E-09
admdist 奉贤 ***	-2992	459	-6.519	7.37E-11
admdist 虹口 ***	7187	528.2	13.606	< 2e-16

admdist 黄浦 ***	5141	620.2	8.289	< 2e-16
admdist 嘉定	-448.2	432.2	-1.037	0.299836
admdist 金山 ***	-5753	466.7	-12.328	< 2e-16
admdist 静安 ***	4295	520.8	8.248	< 2e-16
admdist 闵行 ***	1434	423.1	3.389	0.000703
admdist 浦东 ***	2548	463.8	5.495	3.99E-08
admdist 普陀 ***	3500	455.9	7.676	1.76E-14
admdist 青浦 ***	-4038	455.2	-8.87	< 2e-16
admdist 松江 ***	-3043	419	-7.263	4.00E-13
admdist 徐汇 ***	7615	507.3	15.009	< 2e-16
admdist 杨浦 ***	6812	478.7	14.23	< 2e-16
admdist 长宁 ***	5320	516.7	10.295	< 2e-16
secdist ***	0.7572	0.009864	76.764	< 2e-16
schoolTRUE **	452.7	168.3	2.691	0.007142
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

紧接着笔者计算了多元线性模型在训练数据集和测试数据集的精确度，表 5-3 展示了线性回归模型和随机森林回归模型在训练集和测试集上的表现数据。无论在训练数据集还是测试数据集，随机森林从拟合优度、均方根误差、平均绝对误差和平均相对误差来看都能取得更好的结果。这一结果证明了随机森林在房地产估价中具有的优越性。此外，随机森林模型能够在训练模型的同时得到各个变量的重要性，这是包括线性回归在内的其他模型所没有的功能。变量重要性将为估价师和购房者提供重要参考。

表 5-3 线性回归模型与随机森林模型预测准确性对比

	线性回归		随机森林	
	训练集	测试集	训练集	测试集
拟合优度	86.57%	85.65%	99.18%	95.68%
均方根误差	9124.874	9575.03	2252.309	5252.551
平均绝对误差	6270.533	6419.805	1375.444	3172.846
平均相对误差	13.70%	13.77%	2.84%	6.47%

随机森林本质上是一种非线性的算法，可以根据训练数据拟合出一个非线性的模型。而线性回归只能拟合出线性的模型，当数据集是非线性时，线性回归的回归效果肯定低于随机森林。在现实中房地产价格评估中各个因素对房价的影响并非简单的线性关系，因此寻找一个更复杂的非线性模型能够取得更好的预测效果。随机森林模型不需要假定因变量与自变量是哪一种非线性关系，利用 CART 决策树和 bagging 集成的思想自动生成了判别模型，并且可以通过并行计算减少模型训练时间，同时回归性能在大多数情况下高于单棵决策树、k 近邻、支持向量机、神经网络等非线性的机器学习算法模型。

当然随机森林模型也有自身的缺陷，它不像多元线性回归能明确地给出房价与各特征之间的函数表达式，不能得到变量的边际变化对房价的影响。随机森林更像是一个“黑箱”，把自变量扔进这个“黑箱”，“黑箱”利用复杂的运算后给出预测值，人们很难明确指出计算机的判断过程。

第六章 结论与展望

6.1 研究结论

（1）本文首先归纳汇总了国内外关于将随机森林算法运用于房地产估价的文献，并揭示了先前学者的研究存在的不足之处。然后本文简要介绍了本研究使用到的特征价格理论、网络爬虫和随机森林算法。

（2）借助特征价格模型的框架，本文从建筑因素、邻里因素和区位因素三个大类根据前人经验和现实条件提取了总共 21 个自变量，并给出了变量的量化方案。在数据的获取过程中本文使用了网络爬虫技术，通过编写 R 语言代码从安居客网站中爬取了 41261 条上海市挂牌二手房信息。在变量提取和量化的过程中运用了文本挖掘技术，将标题、核心卖点和专家点评三个文字性描述的变量中进行中文分词处理，进而发掘出该二手房的地铁、学校、商业配套和公园医院配套的信息，进一步补充了这几个安居客网站上未提供但又比较重要的特征。

（3）基于网络爬虫爬取的数据集，本文对上海市二手房房价进行了简要的分析和可视化。由于部分变量数据存在缺失影响建模工作，本文将含有缺失值的二手房予以删除，最终得到 16935 条完整的二手房记录。

（4）在建模的部分本文将数据集划分为了训练数据集和测试数据集，在训练数据集上使用随机森林算法建立了房地产估价模型，并通过拟合优度、均方根误差、平均绝对误差、平均相对误差 4 个指标度量模型在测试集上的表现，反映其泛化能力。然后对模型参数 mtry 和 ntree 进行调参，进一步调优模型性能。

（5）基于随机森林算法本文获取了房地产价格特征变量的重要性排序，并从现实角度解释了变量与房价的关联，重要的变量将能为房地产估价相关人员提供了重要参考。

（6）最后本文运用传统多元线性回归模型在相同的训练数据集上构建模型，通过逐步回归剔除不显著的变量后与随机森林模型在训练集和测试集上进行对比，结果证明了随机森林能够极大提升传统线性回归的预测性能，非常值得推广。

6.2 研究不足和展望

受限于研究时间和作者的学术能力，本文还存在以下不足，针对这些不足还需要进一步改进与完善：

（1）本文在处理缺失值时采用了包含缺失值的二手房一律删除的策略，这导致了原始数据将近 2/3 被排除在了建模所用数据集之外，含有缺失值的房源有可能与数据完整的房源有显著的差别。例如位于老小区的房子更加可能缺失绿化率、容积率、物业费、停车位等等变量数据，这类房源在缺失值处理的过程中可能大多被删除了，这会造成建模数据集所用的样本并非所有上海市二手房随机抽样而来，建立的模型更加倾向于预测较新的住宅。如果能获取缺失的数据，或者采取有效的补差算方法则能够避免这种抽样偏差。

（2）本文在构建区位因素的地铁、学校、商业和公园医院 4 个变量时采用了文本挖掘中文分词的技术，然而构建出的变量仅仅为是否型，这 4 个变量用数值型代替能够取得更好的效果。例如地铁变量如果能够用附近 1km 地铁站数量，最近地铁站步行距离等指标代替能显著提高变量的重要性。目前很多编程语言已经开发出了连接百度地图 API 接口的程序包，通过地名、地址信息就能批量检索出距离和数量信息，通过爬虫程序就能够模仿人工操作提取想要的 GIS 数据。如果将这一技术运用在本研究能够极大提升区位因素变量的重要程度和预测性能。

（3）在对随机森林模型进行超参数调优时，本文先在 ntree 不变的情况下调整了 mtry 参数，然后在 mtry 不变的情况下调整 ntree 参数。然而这样的策略只能得到很小范围内的局部最优，实际的最优参数组合很有可能并非这样的简单组合。采取网格式的搜索策略遍历每种 ntree 和 mtry 的组合才能保证找到全局最优的参数组合。该种策略需要进行非常大量的计算，因此需要更为先进和快速的计算机设备才便于操作。

参考文献

[1] 王克强，刘红梅，姚玲珍.房地产估价[M].上海:上海财经大学出版社.2013:57,165.

[2] 杨沐晞.基于随机森林模型的二手房价格评估研究[D].中南大学，2012.

[3] 陈奕佳.基于随机森林理论的北京市二手房估价模型研究[D].北京交通大学.2015.

[4] 庞枫.大数据时代下房地产自动估价方法的探索[J].经贸实践.2016.(8):98

[5] 庞枫.随机森林模型在二手房批量评估中的应用研究[D].重庆交通大学.2017.

[6] 时文静.基于 Lasso 与数据挖掘方法的影响北京二手房价格的因素分析[D] .北京工业大学.2017.

[7] 曾伟辉.支持 AJAX 的网络爬虫设计系统与实现[D].中国科学技术大学.2009.

[8] 庄旭东,王志坚.基于 R 语言爬虫技术的网页信息抓取方法研究——以抓取二手房数据为例[J].科技风.2019.(6):54-56.

[9] 吴睿,张俊丽.基于 R 语言的网络爬虫技术研究[J].科技资讯.2016.14(34):35-36.

[10] 宋祖杰. 基于支持向量回归的二手房批量评估模型应用研究[D] .重庆大学.2016.

[11] 贾生华,温海珍.房地产特征价格模型的理论发展及其应用[J].外国经济与管理.2004.(5):4-46.

[12] Breiman L. Using Iterated Bagging to Debias Regressions[J]. Machine Learning, 2001, 45(3):261-277.

[13] Antipov E A, Pokryshevskaya E B. Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics[J]. Mpra Paper, 2010, 39(2):1772-1778.

[14] Deng H , Runger G , Tuv E . Bias of Importance Measures for Multi-valued Attributes and Solutions[J]. 2011.

[15] Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle[J]. 2nd Inter. Symp. on Information Theory, 1973, 1973.

附录 源代码

```
#1.网络爬虫代码
library(rvest)
library(stringr)

site1<-"https://shanghai.anjuke.com/sale/"
site2<-"/p"
site3<-"-t9/#filtersort"
district<-c('pudong','minhang','baoshan','xuhui','songjiang','jiading','jingan','putuo',
            'yangpu','hongkou','changning','huangpu','qingpu','fengxian','jinshan',
            'chongming')
dat<-data.frame()
convert1<-function(x){ if(identical(x,character(0)))  ' '  else x }
convert2<-function(x){ if(identical(x,character(0)))  NA_character_  else x }
for(d in 1:length(district))
  {site4<-paste0(site1,district[d],site2)
  for(page in 1:50)
    {site<-paste0(site4,page,site3)
    web<-read_html(site,encoding = 'utf-8')
    link<-html_nodes(web,'div.house-title a') %>% html_attr('href')
    for(house in 1:length(link))
    {
      web2<-read_html(link[house])
      #房源描述以及房价信息
      title<-html_nodes(web2,'.long-title') %>% html_text() %>% convert2() %>% str_trim()
      URL<-link[house]
      sellingpoint<-html_nodes(web2,'.js-house-explain span') %>% html_text()
      %>% convert2()
      opinion<-html_nodes(web2,'.good-character') %>% html_text() %>% convert1()
      unitprice<-html_nodes(web2,'.houseInfo-detail-item:nth-child(3) .houseInfo-content')
      %>% html_text() %>% convert2() %>% str_extract('\\d+') %>% as.numeric()

      #建筑因素
      area<-html_nodes(web2,'.houseInfo-detail-item:nth-child(5) .houseInfo-content')
            %>% html_text() %>% convert2() %>% str_sub(end=-4) %>% as.numeric()
      year<-html_nodes(web2,'.houseInfo-detail-item:nth-child(7) .houseInfo-content')
            %>% html_text() %>% convert2() %>% str_extract('\\d+') %>% as.numeric()
      year<-2019-year
      direction<-html_nodes(web2,'.houseInfo-detail-item:nth-child(8) .houseInfo-content')
                  %>% html_text() %>% convert2()

      decoration<-html_nodes(web2,'.houseInfo-detail-item:nth-child(12) .houseInfo-
content')
```

```
                %>% html_text() %>% convert2()

      category<-html_nodes(web2,'.houseInfo-detail-item:nth-child(10) .houseInfo-content')
                %>% html_text()%>% convert2()
      room<-html_nodes(web2,'.houseInfo-detail-item:nth-child(2) .houseInfo-content')
           %>% html_text() %>% convert2()

      bedroom<-str_sub(room,str_locate(room,"室")[1,1]-1,str_locate(room,"室")[1,1]-1)
              %>% as.numeric()
      hall<-str_sub(room,str_locate(room,"厅")[1,1]-1,str_locate(room,"厅")[1,1]-1)
           %>% as.numeric()
      lavatory<-str_sub(room,str_locate(room,"卫")[1,1]-1,str_locate(room,"卫")[1,1]-1)
                %>% as.numeric()
      storeyinfo<-html_nodes(web2,'.houseInfo-detail-item:nth-child(11) .houseInfo-
content')
                  %>% html_text() %>% convert2()

      storey<-str_extract(storeyinfo,'低层|中层|高层')
      totalstorey<-str_extract(storeyinfo,'\\d+') %>% as.numeric()
      elevator<-html_nodes(web2,'.houseInfo-detail-item:nth-child(14) .houseInfo-content')
                %>% html_text() %>% convert2()

      property<-html_nodes(web2,'.houseInfo-detail-item:nth-child(16) .houseInfo-content')
                %>% html_text()%>% convert2()


      #邻里因素
      community<-html_nodes(web2,'.houseInfo-content > a') %>% html_text() %>% convert2()
      far<-html_nodes(web2,'.commap-info-intro:nth-
child(4)') %>% html_text() %>% convert2()
           %>% str_extract('\\d\\.\\d') %>% as.numeric()
      parkinglot<-html_nodes(web2,'.commap-info-intro:nth-child(5)') %>% html_text()
                %>% convert2() %>% str_extract('\\d+')%>% as.numeric()

      greening<-html_nodes(web2,'.commap-info-intro:nth-child(6)') %>% html_text()
                %>% convert2() %>% str_extract('\\d+') %>% as.numeric()/100
      fee<-html_nodes(web2,'.no-border-rg') %>% html_text()%>%convert2()
           %>% str_extract('\\d+\\.[0-9][0-9]') %>% as.numeric()

      #区位因素
      admdist<-html_nodes(web2,'.loc-text a:nth-child(1)') %>% html_text() %>% convert2()
      secdist<-html_nodes(web2,'.loc-text a+ a') %>% html_text() %>% convert2()
      address<-html_nodes(web2,'.loc-
text') %>% html_text()%>%convert2() %>% str_split('\\n')
      address<-address[[1]][3] %>% str_trim()
```

```
84.
85.       dat<-rbind(dat,data.frame(title,URL,unitprice,area,year,direction,decoration,category,
86.                                 bedroom,hall,lavatory,storey,totalstorey,elevator,property,
87.                                 community,far,parkinglot,greening,fee,admdist,secdist,
88.                                 address,sellingpoint,opinion))
89.     }
90.   }
91.   write.csv(dat,paste0('C:\\Users\\Liu\\Desktop\\毕业论文\\',district[d],'.csv'),
92.             row.names = FALSE)
93.   dat<-data.frame()
94. }
95.
96. #2.数据预处理
97. library(moments)
98.
99. district<-c('pudong','minhang','baoshan','xuhui','songjiang','jiading','jingan','putuo',
100.            'yangpu','hongkou','changning','huangpu','qingpu','fengxian','jinshan',
101.            'chongming')
102.documents<-paste0(district,'.csv')
103.DAT<-data.frame()
104.for (i in 1:16)
105.{DAT<-rbind(DAT,read.csv(documents[i],header = T))}
106.which(is.na(DAT$title))->row
107.DAT[row,]->wrong
108.DAT[-row,]->DAT
109.unique(DAT[-2])->DAT
110.
111.#文本挖掘
112.library(jiebaR)
113.library(stringr)
114.library(wordcloud2)
115.
116.title<-paste(DAT$title)
117.sellingpoint<-paste(DAT$sellingpoint)
118.opinion<-paste(DAT$opinion)
119.info<-paste(title,sellingpoint,opinion)
120.cutter<-worker()
121.segwords<-segment(info,cutter)
122.segwords<-gsub("[0-9a-zA-Z]+?","",segwords)
123.segwords<-str_trim(segwords)
124.table(segwords)->wordfre
125.sort(wordfre,decreasing = TRUE)->wordfre
126.head(wordfre,200)->wordfre2
127.wordcloud2(wordfre2,size=4,shape='circle',fontFamily="楷体",color='black',
128.           backgroundColor='white')
```

```
129.
130.#属性补充
131.subway<-str_detect(title,'号线|地铁|轨')|str_detect(sellingpoint,'号线|地铁|轨')|
132.                   str_detect(opinion,'地铁|号线|轨')
133.school<-str_detect(title,'学|校|幼儿园|附小|附中|读')|
134.                   str_detect(sellingpoint,'学|校|幼儿园|附小|附中|读')|
135.                   str_detect(opinion,'学|校|幼儿园|附小|附中|读')
136.business<-str_detect(title,'商|市场|超市|菜场|万达|家乐福|百联|联华|购物')|
137.          str_detect(sellingpoint,'商|市场|超市|菜场|万达|家乐福|百联|联华|购物')|
138.          str_detect(opinion,'商|市场|超市|菜场|万达|家乐福|百联|联华|购物')
139.amenity<-str_detect(title,'公园|广场|医院|滨江|苏州河|佘山|河畔|滨江')|
140.         str_detect(sellingpoint,'公园|广场|医院|滨江|苏州河|佘山|河畔|滨江')|
141.         str_detect(opinion,'公园|广场|医院|滨江|苏州河|佘山|河畔|滨江')
142.
143.cbind(DAT[c(-1,-2,-8,-22,-23,-27)],subway,school,business,amenity)->dataset
144.Amelia::missmap(dataset,y.labels=F,y.at=F,main=NA,col = c("black", "grey"))
145.apply(is.na(dataset),2,sum)
146.apply(is.na(dataset),2,sum)/nrow(dataset)
147.write.csv(dataset,'C:\\Users\\Liu\\Desktop\\毕业论文\\上海二手住宅数据.csv',
148.          row.names = FALSE)
149.
150.#EDA
151.hist(dataset$unitprice,col='white',freq = F,breaks=seq(from=0,to=290000,by=5000),
152.     xlab='Price',ylab='Density')
153.lines(density(dataset$unitprice) ,col='blue',lwd=1.5)
154.mean(dataset$unitprice)
155.sd(dataset$unitprice)
156.skewness(dataset$unitprice)
157. kurtosis(dataset$unitprice)
158.
159. #3.随即森林建模
160.library(randomForest)
161.library(caret)
162.library(doParallel)
163.library(foreach)
164.
165.housedata<-read.csv('C:\\Users\\Liu\\Desktop\\毕业论文\\预处理后数据集.csv',header=TRUE)
166.names(housedata)
167.summary(housedata)
168.
169.set.seed(666)
170.createDataPartition(y=housedata$unitprice,p=0.75,list = FALSE)->inTrain
171.trainset<-housedata[inTrain,]
172.testset<-housedata[-inTrain,]
173.RF<-randomForest(unitprice~.,data=trainset,importance=TRUE)
```

```
print(RF)
predict(RF,trainset)->pred_train
1-sum((pred_train-trainset$unitprice)^2)/sum((trainset$unitprice-
   mean(trainset$unitprice))^2)
sqrt(mean((pred_train-trainset$unitprice)^2))
mean(abs(pred_train-trainset$unitprice))
mean(abs(pred_train-trainset$unitprice)/trainset$unitprice)
head(cbind(trainset$unitprice,pred_train,abs(pred_train-trainset$unitprice)),10)

predict(RF,testset)->pred_test
1-sum((pred_test-testset$unitprice)^2)/sum((testset$unitprice-mean(testset$unitprice))^2)
sqrt(mean((pred_test-testset$unitprice)^2))
mean(abs(pred_test-testset$unitprice))
mean(abs(pred_test-testset$unitprice)/testset$unitprice)
head(cbind(testset$unitprice,pred_test,abs(pred_test-testset$unitprice)),10)

RF$mse
plot(RF)

MRE_train<-c()
MRE_test<-c()
i=0
cl <- makeCluster(4)
registerDoParallel(cl)
for(t in c(2:20))
  {i=i+1
  RF1<- foreach(ntree=rep(100, 4),
              .combine=combine,
              .packages='randomForest') %dopar% randomForest(unitprice~.,
                                                             data=trainset,
                                                             ntree=ntree,
                                                             mtry=t)
  MRE_train[i]<-mean(abs(predict(RF1,trainset)-trainset$unitprice)/trainset$unitprice)
  MRE_test[i]<-mean(abs(predict(RF1,testset)-testset$unitprice)/testset$unitprice)}
stopCluster(cl)

MRE_train2<-c()
MRE_test2<-c()
i=0
cl <- makeCluster(4)
registerDoParallel(cl)
for(t in seq(from=100,to=1000,by=100))
  {i=i+1
  RF1<- foreach(ntree=rep(t/4, 4),
               .combine=combine,
```

```
                .packages='randomForest') %dopar% randomForest(unitprice~.,
                                                               data=trainset,
                                                               ntree=ntree,
                                                               mtry=18)
  MRE_train2[i]<-mean(abs(predict(RF1,trainset)-trainset$unitprice)/trainset$unitprice)
  MRE_test2[i]<-mean(abs(predict(RF1,testset)-testset$unitprice)/testset$unitprice)
  }
stopCluster(cl)

RF_final<-randomForest(unitprice~., data=trainset, ntree=400, mtry=18,
                       importance=TRUE,do.trace=TRUE)
predict(RF_final,trainset)->pred_train
predict(RF_final,testset)->pred_test

imp<-importance(RF_final)
varImpPlot(RF_final)

#线性回归
LM<-lm(unitprice~.,data=trainset)
summary(LM)
step(LM)

predict(LM,trainset)->pred_LM
1-sum((pred_LM-trainset$unitprice)^2)/sum((trainset$unitprice-mean(trainset$unitprice))^2)
sqrt(mean((pred_LM-trainset$unitprice)^2))
mean(abs(pred_LM-trainset$unitprice))
mean(abs(pred_LM-trainset$unitprice)/trainset$unitprice)

predict(LM,testset)->pred_LM2
1-sum((pred_LM2-testset$unitprice)^2)/sum((testset$unitprice-mean(testset$unitprice))^2)
sqrt(mean((pred_LM2-testset$unitprice)^2))
mean(abs(pred_LM2-testset$unitprice))
mean(abs(pred_LM2-testset$unitprice)/testset$unitprice)
```