密级：公开　　中图分类号：F062.4

☑全日制　☐非全日制

浙江工商大學

硕士学位论文

（专业学位）

论文题目：短租异质性房源定价问题研究——基于Airbnb平台不同房东类型的数据分析

作者姓名：葛皓天

专业学位类别：应用统计

专业学位领域：应用统计

研究方向：数据挖掘

指导教师：徐静

提交日期：2022年 1月

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Zhejiang Gongshang University Master's Degree Thesis in Applied Statistics

Research on the Pricing Problem of Heterogeneous Short-term Rental Housing——Based on the data analysis of different types of landlord housing on the Airbnb platform

Author： Haotian Ge

Major： Applied Statistics

Supervisor： Jing Xu

School of Statistics and Mathematics

Zhejiang Gongshang University

Hangzhou, 310018, P. R. China

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

摘要

根据国家信息中心发布的《中国共享经济发展年度报告（2020）》显示，2019 年我国共享住宿市场交易额达到 225 亿，相比于 2011 年初期，市场规模扩张了数千倍。但由于疫情的影响，最新数据显示 2020 年市场交易额降为 158 亿元，同比下降 30%，同时近两年来，共享住宿市场的直接融资规模骤减，从 2018 年的 33 亿到 2020 年的 1 亿，这就更加考验了现存共享短租平台的经营和生存能力。Airbnb 平台属于单纯的 C2C 模式，房东拥有更多的定价自主权，不同类型房东的定价模式差异化更明显。本文以房东的房源拥有量作为判断条件，将房东分为专业房东和非专业房东来展开差异化定价问题研究。如何为不同类型房东分别建立差异化定价模型，来帮助房东进行更加合理的定价，是本文研究的主要问题。

1、本文以 Airbnb 平台提供的北京市的短租房源真实数据作为研究对象，原始变量共有 74 个。通过对过往学者大量文献的归纳梳理，并结合大量的爬虫技术来量化房源的地理位置优势以及利用文本挖掘方法来量化房源的主题丰富性，最后将房源价格影响因素主要归结为 4 个方面，分别是房源属性、房东属性、区位优势、在线评价，总计 45 个二级变量。

2、基于 45 个变量分别对专业房东和非专业房东建立逐步回归模型，分析不同影响因素是如何影响两种房东定价的，以及所对应的经济意义。比较专业房东和非专业房东逐步回归模型中显著性变量的差异，并通过分位数回归模型探究不同价格分位数点下，各个因素对于房源的影响程度。

3、从机器学习树模型的角度出发，构建逐步回归变量选择、Lasso 变量选择和随机森林、XGBoost、CatBoost 的组合模型。将原有数据分为训练集和测试集，利用训练集来训练模型，再分别比较各组合模型在测试集上的预测效果。通过网格搜索和交叉验证方法为专业房东和非专业房东选择最优定价模型，最后对黑盒定价模型进行模型可解释性分析。

研究发现：逐步回归模型结果显示专业房东定价时共考虑 32 个显著变量，非专业房东定价时共考虑 23 个显著因素。两类房东均对房间类型、房源的空间（床位数、浴室数）及基础设施、房东的回复情况、房源的区位优势、房源历史评论数、综合评分等因素十分看重。除此以外，专业房东对于房源的闪订政策、主题丰富性、房源整洁

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性评分、便利性评分、沟通交流评分更加看重。专业房东定价体系更加成熟精准，非专业房东定价体系尚未成熟，通过比较，Lasso-XGBoost 模型对于专业房东房源价格的预测精度最高，RMSE 值最小且 R 方达到了 84.95%；Lasso-CatBoost 模型对非专业房东房源价格的拟合性能最佳，预测误差最小且 R 方达到了 80.73%。

关键词：逐步回归；分位数回归；lasso；随机森林；XGBoost；Catboost

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ABSTRACT

According to the "Annual Report on China's Sharing Economy Development (2020)" released by the National Information Center, the transaction volume of my country's shared accommodation market reached 22.5 billion in 2019. Compared with the beginning of 2011, the market scale has expanded thousands of times. However, due to the impact of the epidemic, the latest data shows that the market transaction volume in 2020 has already dropped to 15.8 billion yuan, a year-on-year decrease of 30%. At the same time, the scale of direct financing in the shared accommodation market has plummeted in the past two years, from 3.3 billion in 2018 to 100 million in 2020, which further tests the operation and viability of existing shared short-term rental platforms. The Airbnb platform is a C2C model. Landlords have more pricing autonomy, and the pricing models of different types of landlords are more differentiated. In this paper, the landlord's housing ownership is used as the judgment condition, and the landlord is divided into professional landlords and non-professional landlords to carry out the research on the differential pricing problem. How to establish differentiated pricing models for different types of landlords to help landlords make more reasonable pricing is the main issue studied in this article.

1. This article takes the real data of short-term rental housing in Beijing provided by the Airbnb platform as the research object. There are 74 original variables in total. By summarizing and sorting out a large number of literatures from past scholars, combining a large number of crawling techniques to quantify the geographical advantages of listings, and using text mining methods to quantify the richness of listings, finally the main factors affecting the prices of listings are summarized into four terms of housing properties, landlord properties, location advantages and online evaluation-a total of 45 secondary variables.

2. Establish a stepwise regression model for professional landlords and non-professional landlords based on 45 variables, analyzing different influencing factors' affection and economic significance to the pricing of different types of landlords. Comparing the differences of significant variables in the stepwise regression model between professional landlords and non-

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

professional landlords, and use quantile regression models to explore the impact of various factors on housing availability at different price quantile points.

3. From the perspective of the machine learning tree model, constructing a combined model of stepwise regression variable selection, Lasso variable selection and Random Forest, XGBoost and CatBoost. Divid the original data into a training set and a test set, use the training set to train the model, and then compare the prediction effects of each combined model on the test set. The optimal pricing model is selected for professional landlords and non-professional landlords through grid search and cross-validation methods, and finally the black box pricing model is analyzed for model interpretability.

The study finds that: stepwise regression model shows that professional landlords consider a total of 32 significant variables when pricing, and non-professional landlords consider a total of 23 significant factors. Both types of landlords attach great importance to factors such as room type, room space (number of beds, number of bathrooms), infrastructure, landlord's response, location advantages of the room, the number of historical reviews of the room, and comprehensive ratings. In addition, professional landlords pay more attention to the flash booking policy, theme richness, house cleanliness score, convenience score, and communication score. Professional landlord pricing system is more mature and accurate, and non-professional landlord pricing system is not yet mature. Through comparison, the Lasso-XGBoost model has the highest prediction accuracy for professional landlord housing prices, with the smallest RMSE value and an R-square of 84.95%; the Lasso-CatBoost model is best fit to the non-professional landlord’s housing prices, the prediction error is the smallest, and the R-square reaches 80.73%.

Keywords: stepwise regression; quantile regression; lasso; random forest; XGBoost; Catboost

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目录

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 1 章 引言

1.1 研究背景

1.1.1 共享经济的发展

共享经济是指利用互联网等现代信息技术，以使用权分享为主要特征，整合海量、分散化资源，满足多样化需求的经济活动总和。近年来，随着大数据、云计算等高科技产业和传统产业的深度融合，共享经济在全球范围内快速发展起来。自 2016 年起，我国政府工作报告中已经连续 5 年提及共享经济，这几年来国内资本大量涌入共享经济领域，滋养出了众多领域的共享经济企业，例如滴滴出行、小猪短租等。根据国家信息中心发布的《中国共享经济发展年度报告（2021）》[1]显示，我国的共享经济规模在 2017 至 2019 年之间迅速增长，即便在受到突发疫情的影响下，2020 年的共享经济市场交易额依然增长了 2.9%，达到了 33773 亿元。但也可以发现，国内的共享经济自 2019 年起，增长速度进入了拐点，由原来的高歌猛进逐渐进入了调整阶段，同时从 2015-2020 共享经济的直接融资状况来看，2018 和 2019 年的负增长率在逐年变大。2020 年共享办公和共享医疗在得益于疫情导致的人们只能隔离在家的环境下，得到资本的迅速青睐。另外在疫情的冲击下，共享经济中的知识技能、生活服务、生产能力等方面也重新成为了融资的风口。共享经济这一新经济形式，正在影响着国民生活的方方面面，通过和 5G、人工智能和物联网等技术的加速融合，共享性服务和共享性消费新业态正在迅速扩张，在保障居民日常生活需要、稳定就业市场的不确定性、促进经济企稳回升方面发挥了重要的作用。

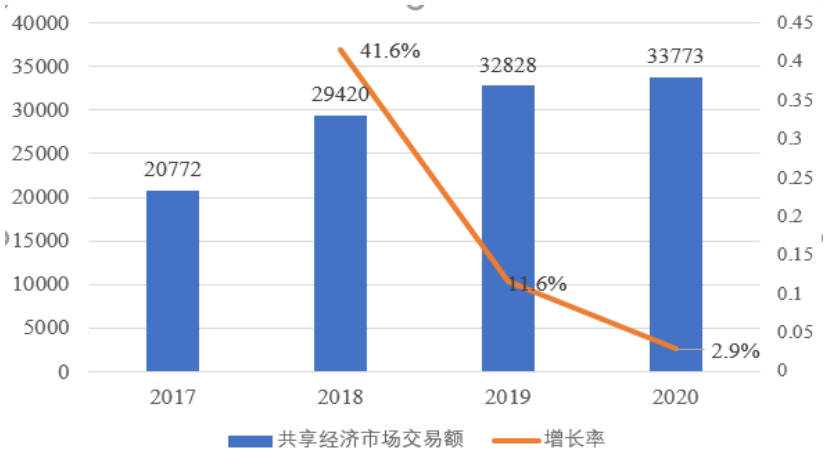


图 1.1 2017 年-2020 年我国共享经济市场交易规模

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

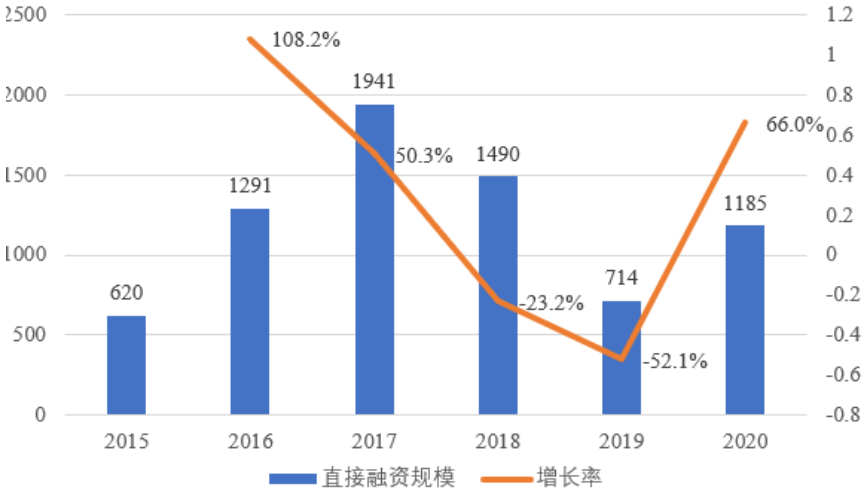


图 1.2 2015 年-2020 年我国共享经济直接融资状况

1.1.2 共享住宿平台的发展

随着经济的发展和居民消费水平的提高，消费者在考虑在外住宿时，已经不满足于单一形式的连锁以及高档酒店。同时近年来，由于房地产行业的快速发展，可居住住房量在不断上升，越来越多的房主希望拿出闲置的房源以共享住宿的形式来获取额外的收益，共享住宿房源的种类也越来越多，例如公寓、别墅以及各种有特色装修风格的民宿，恰好能满足不同消费者的个性化体验。互联网经济下的在线短租能够充分利用共享平台，以低成本、高匹配度的优势实现房源需求双方高效匹配与交易，既满足了不同消费者租房需求，也盘活了闲置房产资源。

根据国家信息中心发布的《中国共享经济发展年度报告（2021）》[1]显示，2019 年我国的共享住宿市场交易额已经达到了 225 亿，同比增长 36.4%，相比于 2011 年的 8357 万元，已经翻了无数倍，同时 2019 年的共享住宿人数约为 2 亿，同比增长 53.8%，这无疑证明了我国共享短租市场的潜力极大。但由于受到疫情的影响，最新数据显示 2020 年的共享住宿市场交易额为 158 亿元，同比下降 30%，同时也能发现，近两年共享住宿的直接融资规模骤减，从 2018 年的 33 亿元，到 2019 年的 1.5 亿和 2020 年的 1 亿，这就更加考验了现存共享短租平台的经营和生存能力。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

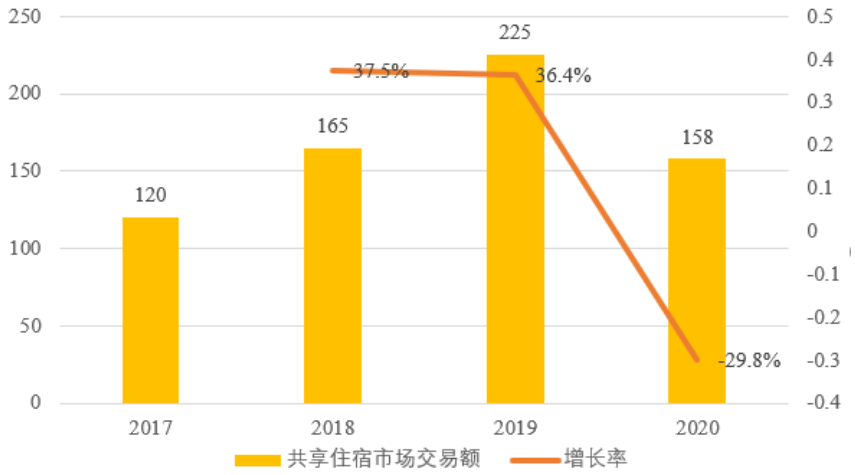


图 1.3 2017-2020 我国共享住宿市场交易规模

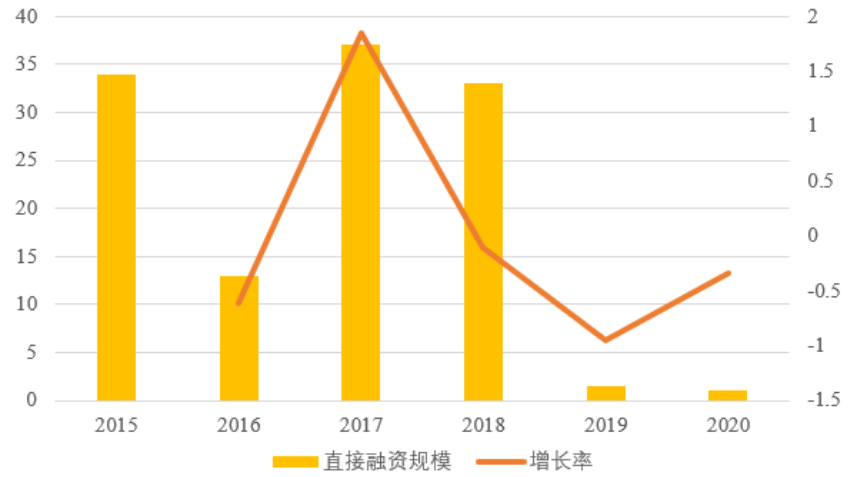


图 1.4 2015-2020 我国共享住宿直接融资规模

目前国内在线短租平台进入蓄力发展期，各家企业竞争特色都非常明显，行业布局日趋完善，共享短租平台的产业已经达到了一定的规模。国内的短租平台主要有两种运营模式，一种是以途家为代表 B2C+C2C 并行模式，途家既有通过租赁、收购来收集分散房源，再按照统一标准进行装修并统一管理的模式，也有通过平台连接房主和房客，一方面通过吸引房东，将优秀房源入驻，另一方面通过平台完善订房交易流程来吸引更多的房客的模式。另一种是单纯的 C2C 模式，平台作为中介鼓励房东，给予房东更多的自主性，来打造人情味，主要的代表是 Airbnb、小猪短租和蚂蚁短租等。

短租房因为性价比、类型多样以及独具的特色受到越来越多旅游者的青睐，但也正是因为短租房源的独特性，其价格也参差不齐，相似性房源呈现出巨大的价格差的现象比比皆是，对于消费者来说，价格的差异可能会严重影响到入住体验，进而影响

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

到共享住宿平台的发展。同时除房源的选址、风格、宽敞度等基本信息外，房东之间不同特性的差异定价同样影响着双边市场中住宿供需关系。

综上所述，目前共享经济的市场发展规模增速已经进入了拐点，同时共享短租市场由于受疫情冲击较大，在直接融资没有增加的环境下，共享短租市场的复苏进程更加曲折。但随着国内疫情防控取得的显著成效，以及中央提出要“加快形成以国内大循环为主体、国际国内双循环相互促进的新发展格局”的政策偏向来看，人们的旅游需求即将迸发，短租平台应该利用更加精细化的经营以及更加合理的定价策略，才能抓住这波机会，完成疫情后的大翻盘。

1.2 研究意义

1.2.1 理论意义

消费者在进行消费时更多的是受到心理的影响，消费者心理对价格的敏感度，指的是消费者在线预订房源时，愿意接受短租房价格的上限和下限，在一定的价格承受范围内，选择哪些房源。消费者不会预定同类房源价格过高的短租房，并且对价格过低同质性房源时也不会考虑。由于在线短租平台信息的多样化以及产品的透明化，对短租房进行合理定价并给出重要定价因素可以使消费者在预定房源时能够获取充分的信息，也能助力短租平台不断完善和长远发展。

1.2.2 实际意义

对于房东而言，非专业房东由于缺乏专业的管理知识和系统的管理体制，本身不具备优势，这时合理的定价机制能够帮助他们实现高效的动态定价，而专业房东可以结合定价模型以及依靠运营和市场数据制定更加精细的定价决策，对自己的不同类型房源实行特色定价来收获更高的市场占有率，充分发挥多房源的优势；对平台而言：挖掘房东背后的定价机制，有利于短租平台进一步了解房东特性，满足个性化房东和消费者的需求；对消费者而言，定价模型可借助短租平台信息的多样化及透明化，将房源的优势清晰化地展现给消费者，降低交易风险并保障用户利益，提高房源与消费者需求的匹配度进而提升入住率，间接提升入住体验感，实现短租平台、房东、消费者三方的收益最大化。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.3 研究内容与初步方案

1.3.1 研究内容

消费者心理行为研究表明，任何时代，价格永远是消费者争论的主题，而促成一笔交易成功的因素主要靠价格来做决定，任何电商平台的发展都是为了追求最大订单量和销售额，而目前国内外很少有学者对在线短租平台数据来做差异化定价研究，本文在对在线短租行业和平台经营模式有了一定的了解后，基于在线短租平台上的真实房源数据，通过结合过往学者研究以及短租房源的特性来选取短租房价格的影响因素，并利用机器学习方法对短租房做更为精准的定价预测。本文的研究内容如下:

本文第二章是文献综述。主要分为两部分，一部分是对共享住宿价格影响因素的研究，另一部分是有关共享住宿价格预测方法的研究。由于目前对共享住宿价格影响因素的研究过少，发展还还未成熟，本文借鉴了传统酒店以及住房行业中涉及的价格因素的文献，来完善对共享住宿价格理论的理解。其次在价格预测模型方面，本文也同时借鉴了酒店及住房价格预测的文献，更加全面地了解了各种价格预测模型之间的差异和优劣性，对后续分别为专业房东和非专业房东建立定价模型奠定了基础。

第三章主要基于过往学者的研究，并结合目前已有的原始数据以及官方平台的可获取数据，将影响短租房源定价的一级因素归纳为以下四个方面，分别是房源属性、房东属性、区域优势、在线评价，并且将四个方面的影响因素都量化为指标，最终保留了 45 个二级变量，为后续异质性房源的定价研究打下基础和铺垫。

第四章包含异质性房源的定义以及数据的收集和预处理。在过往学者研究的基础上，根据房东的房源拥有量来判断专业和非专业房东类型。本文所选的定价影响因素主要包含房源属性、房东属性、区域优势以及在线评价这四个方面。其中房源的区位优势因素通过 Python 调取高德 API 爬取地铁站、公交站、景点、旅馆和酒店所在的经纬度进而来量化房源的区位优势变量。对于房源描述丰富度方面的因素，本文通过文本挖掘的方法来量化房源描述的内容丰富度和主题丰富度。此外，数据预处理是数据挖掘中的一项重要的任务，数据的质量会直接影响模型的效果。本文基于在线短租真实房源数据进行了缺失值的处理、异常房源的剔除、特征变量的转化等操作，对数据进行完整的清洗，为后面的定价问题研究做好铺垫。

第五章主要是对不同房东类型的房源数据进行差异化定价的实证研究。第一部分主要结合方差膨胀因子和逐步回归的方法来去除多重共线性，筛选出影响专业房东和

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

非专业房东各自定价模型的重要影响变量。第二部分主要是根据逐步回归变量选择后的结果分别对专业房东和非专业房东建立回归模型和分位数回归模型，来进一步阐释不同房东之间定价时考虑因素的差异性。

第六章是共享短租异质性房源定价模型的构建。主要是从机器学习模型的角度出发，通过逐步回归变量选择、Lasso 变量选择与集成学习构建组合模型，分别为专业房东与非专业房东建立定价模型，运用网格搜索和十折交叉验证的方法来选择模型的最适参数，通过比较模型的预测精度为两类房东选定最佳模型。并利用 PDP 图对专业房东和非专业房东各自的最佳黑盒模型进行模型可解释性分析。

1.3.2 研究方法与技术路线

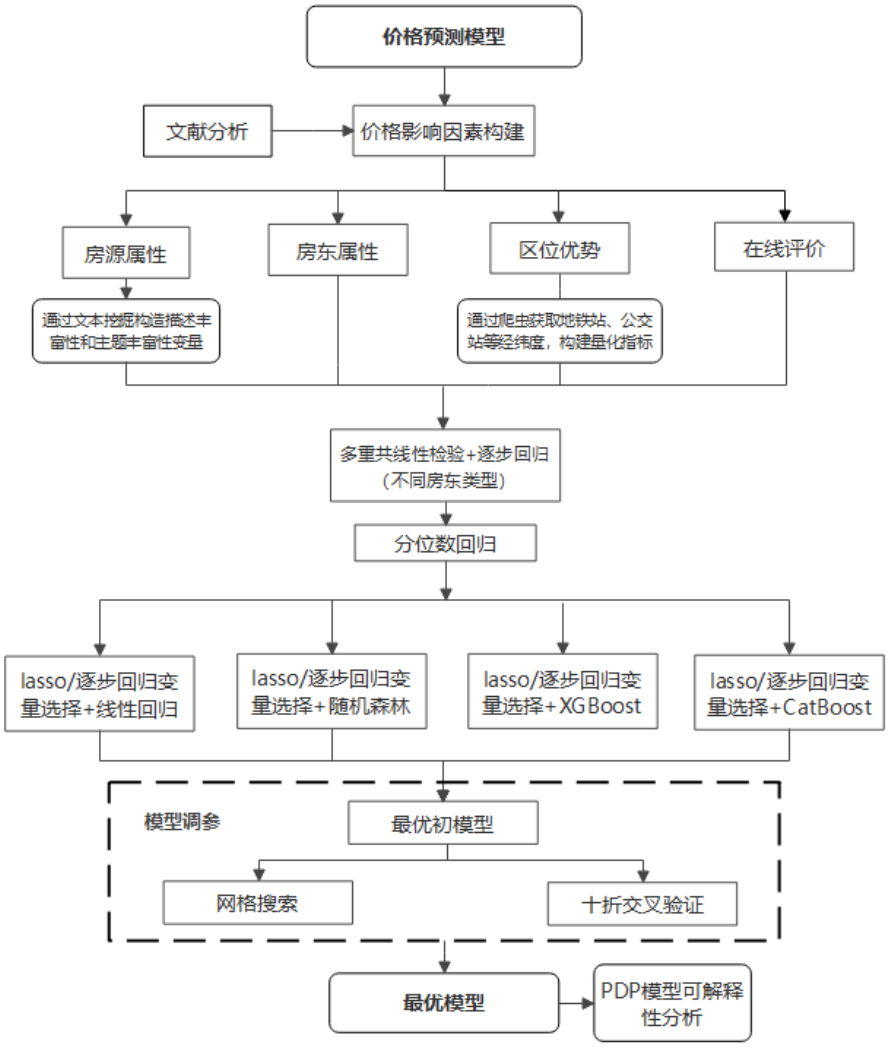


图 1.5 文章研究技术路线图

本文的技术路线如图 1.5 所示。研究方法主要包含以下几点：一、首先根据过往专家学者对短租房源价格影响因素的梳理，从房源属性、房东属性、区位优势、在线评

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

价这四个方面来构建短租房源定价影响因素模型；二、结合短租房源的特性，通过爬虫和文本挖掘的方法，构建房源的区位优势变量以及房源描述丰富性变量，进一步完善短租房源的定价影响因素模型；三、根据过往学者的研究，定义专业房东和非专业房东，并对两类房东分别构建逐步回归和分位数回归模型，通过对比模型保留的显著性变量以及各变量对房源价格的影响效应，来分析专业房东和非专业房东定价时所考虑因素之间的差异；四、基于逐步回归和 lasso 变量选择对两类房东定价模型所选择的变量，分别建立线性回归、随机森林、XGBoost、CatBoost 模型，通过比较模型的预测精度来选择专业房东和非专业房东各自的基模型，并运用网格搜索和十折交叉验证的方法对模型进行调优，来得到两类房东最优定价模型，最后再进行模型的可解释性分析。

1.3.3 创新点

一、对于短租房源的定价影响因素较为全面和准确。本文以 Airbnb 平台公布的北京市的真实房源数据为基础，结合过往学者的研究，将短租房源定价影响因素主要归结为房源属性、房东属性、区位优势、在线评价四个方面共 45 个二级影响因素，其中在房源属性中结合短租房源的特点，构建了房源描述的丰富度变量实词数以及房源主题丰富度变量主题熵，针对于区位优势中，本文通过爬虫结合 haversine 公式计算经纬度直线距离构建出短租房源 1km 内的地铁站数、1km 内的公交站数、1km 内的景点数、1km 内的旅馆和酒店数、距离最近的地铁站距离、距离最近的公交站距离、距离最近的景点距离、距离最近的旅馆和酒店距离来细节量化地段优势。

二、在过往研究中，学者对不同房东类型的房源差异化定价的研究较少，并且没有给出差异化的定价模型。本文根据房源拥有量对房东进行分类，从专业房东和非专业房东出发，基于定价影响因素分别构建逐步回归和分位数回归模型，分析两类房东定价时所考虑因素的差异，以及各类因素对房源价格的影响情况并在最后给出了合理的差异化定价模型。本文基于真实数据的实证研究，进一步完善了该领域的研究。

三、本文从机器学习非线性回归的角度出发，构建 lasso 与随机森林、XGBoost、CatBoost 的组合模型，相比于过往学者的线性回归模型有更好的预测精度和拟合优度，并对黑盒模型进行了模型的可解释性分析，为异质性房源的定价模型提供了一种较为新颖的思路，也为北京市不同类型房东的准确定价提供了相关的理论参考。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 2 章 文献综述

2.1 共享住宿价格影响因素的研究

Wang 和 Nicolau 利用普通最小二乘法和分位数回归对来自 33 个城市中的 18 万个房源研究发现，在 4 个房东属性因素，10 个房源属性因素，5 个房源设施因素、4 个租赁规则因素以及在线评论和在线评分因素中，24 个变量都对价格的影响都是显著的，并且发现大部分显著影响的因素在酒店价格影响因素的分析中也同样重要，例如房源的区域位置、顾客的评论等，并且大多数的正负号影响性都相同，例如可容纳人数、更多的浴室和卧室、提供更多的设施和服务等因素均对共享住宿和酒店的价格呈正向影响[2]。王春英和陈宏民基于小猪短租平台上 10 个重点旅游城市的 12527 个短租房源，利用最小二乘回归和分位数回归探究了房源类型、宜住人数、与城市旅游中心的距离、是否有无线网络、是否可以做饭、房东房源数量、总订单数、总评论数以及总得分这 9 个因素对短租房价的影响，得出价格与离市中心的距离、评论数呈负向关系，与可容纳人数、是否有厨房、Wifi 等呈正相关[3]。Zhang 等基于田纳西州纳什维尔地区的 Airbnb 数据利用 GLM 和 GWR 空间模型研究了影响 Airbnb 房源价格的因素发现，会议中心的距离与房源价格呈显著的负相关，并且离中央地区的距离越近，影响越显著。同时房源的上架日期、评论数量、评级分数也是房源价格的显著影响因素[4]。Hong 等基于纽约和洛杉矶的 Airbnb 数据利用最小二乘回归发现可居住人数、床位数、浴室数、超赞房东等因素对共享住宿价格呈现正向显著影响，其中共享房间、评论数以及是否闪订对价格呈现负向显著影响，此外作者利用 MGWR 模型发现变量的解释能力会随着地点的变化而变化，不同城市应该采用不同的定价模式[5]。

Gutt 等基于纽约市 14000 多个 Airbnb 房源数据进行了定量分析，发现被评为明星级的房东，其房源的价格平均比其他房东高出 2.69€[6]。吕姝基于 2016 年 12 月的 264 家 Airbnb 上海地区的房源数据，构建了用户评价和成交价格的联立方程模型，利用广义矩估计法进行了模型的参数估计，发现用户评价的提高将会导致房源价格的提高，并且老用户的评论增加所造成的粘度效应，会对潜在消费者具有一定的诱导效应[7]。

另外由于短租不同于一般的酒店，他是住客和房东之间的双向选择，需要双方建立一定的信任机制，才能促使订单的完成。正式因为这种特殊属性，用户在考虑短租房的时候考虑的往往不止是一些房源的基本属性和价格，还会重点考虑房东的披露信息和未披露信息。王璐瑶基于 2018 年 11 月 Airbnb 和小猪短租在上海的成交数据，在

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同一个平台下，房东披露的个人信息不同对最终房源的成交价格的影响不同[8]。吕姝的相关研究认为 Airbnb 房源的成交价格与房东身份信息的披露具有显著关系，提供信息披露的房东比未提供的有能力指定更高的价格[7]。赵寒的相关研究证明，房东供给个人介绍的图片的面部表情例如视线方向对房屋的出租具有显著影响[9]。Tussyadiah 对美国 14 个城市的 Airbnb 房东自我介绍信息进行文本挖掘发现房东提交的自我描述信息对消费者最后的决策起着显著性的影响[10]。

除了房东的社交度外，不同类型房东的房源成交价格也存在显著差异。Xie 和 Mao 研究发现，共享短租的房东中很少接触过专业的培训缺乏经验，造成了专业房东和非专业房东在市场的运营掌控、获得的收益、房源的调控和占有率上有显著的差异[11]。Li 等通过 Airbnb 数据的实证研究发现，专业房东和非专业房东在定价行为上存在着显著的差异，由于非专业房东缺乏经验，导致其价格效率低下，使得专业房东房源的市场占有率越来越高[12]。Martin 等认为专业房东与非专业房东在个人信誉度、所在地区、自我介绍等因素的差异会直接影响到房东和租客双边市场中房屋供需关系，进而影响短租房源的成交价格[13]。陈子燕和邓丽基于 2017 年 2 越美国纽约地区 Airbnb 数据，利用分位数回归分别对专业房东和非专业房东的定价机制进行了差异性研究，发现专业房东更重视房间的容纳量、浴室和卧室数，以及在线评分中的干净整洁、位置便利与性价比这几个变量，非专业房东更关注房间是否闪订、房源的总评分、位置评分以及性价比，且房源的可容纳人数和位置便利评分变量的权重均高于专业房东[3]。牛阮霞和何砚在对蚂蚁短租平台 31 个城市的房源数据的研究中，将专业房东和非专业房东作为房东性质变量放入房源价格预测的最小二乘法回归和分位数回归中，发现在一线城市下，房东性质变量在 OLS 回归和分位数回归中的都是显著的，专业房东的经营对房源产生的溢价性在一线城市表现尤为突出[14]。吴晓隽基于 Airbnb 网站中获取的上海市短租房源数据，研究发现比起业余房东来说，专业房东对市场需求更加敏感，更懂得扩大收入，业余房东比较分散，缺乏经验，他们更被共享住宿的社交模式所吸引，更接近共享住宿的本质[15]。

2.2 共享住宿价格预测模型的研究

通过对影响共享住宿的价格因素的文献梳理，发现目前的学者主要采用最小二乘法回归和分位数回归来判定各类因素对房源价格的影响程度是否显著，以及与房源价格之间的正负向关系，但如果仅从最小二乘法回归出发来拟合共享短租的价格预测模

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

型，效果并不理想。并且目前学者主要研究了共享短租价格的影响因素，而对共享短租价格预测模型的研究还尚且较少。王春英和陈宏民基于小猪短租平台的真实房源数据建立了最小二乘法回归模型，自变量只能解释 54.77%的因变量的变动，分位数回归对因变量的解释能力仅仅达到了 38%[3]。蒋钰洁基于 2018 年 9 月爬取的途家 5183 条房源数据，从机器学习的角度出发，对短租房源的价格预测建立了非线性回归模型，包含回归树、bagging、boosting、随机森林和神经网络模型，发现随机森林模型效果最好，模型的拟合优度远高于最小二乘法回归[16]。曹睿等以美国洛杉矶市 2020 年最新的 37048 条 Airbnb 数据作为研究对象，利用 XGBoost 建立了短租房源的价格预测模型并利用 SHAP 模型对黑箱模型进行了解释，通过比较 Adaboost、KNN、决策树、最小二乘回归等模型的 R 方发现，XGBoost 的预测效果最好[17]。因为共享短租和酒店以及二手住房价格的变动具有一定的相关性，因此以往关于酒店和二手住房的价格预测模型的研究也可以作为参考。

张利君从宏观经济角度出发，基于 2000 至 2015 年的 11 个商品房销售价格的影响因素，利用主成分分析进行降维后的数据建立部分线性模型，发现部分线性模型的结果优于线性回归模型[18]。李圆圆基于北京市 2016 年 6 月到 2017 年 6 月 12000 条二手房成交数据，将房价影响因素主要归因于房屋属性硬指标和房屋交易软指标，通过灰色关联度分析来观测变量与价格之间的关系，最后构建 BP 神经网络对价格进行预测，R 方达到了 97%[19]。麻顺顺基于郑州市金水区二手房数据，在时间序列的基础上将空间分布的差异也考虑了进去，建立了空间序列+Attention 机制+LSTM 神经网络的预测模型，模型预测效果较好[20]。赵晨阳基于“链家”平台 2019 年上海市浦东区成交的二手房源数据，以房屋属性、行为属性、交易属性及区位属性作为影响二手房价格的主要因素，发现基于 Bagging 集成方法的随机森林模型优于基于 Boosting 集成算法的 XGBoost 模型，将两个模型进行融合后，融合模型的预测误差要明显小于挂牌价格与成交价格的差价[21]。慕钢基于爬取的 2018 年链家网上海市二手房房价数据，分别建立了线性回归、KNN、神经网络、XGBoost、LightGBM 二手房房价预测模型，通过比较模型的 R 方，发现 LightGBM 模型的预测效果最好，并且模型具有较强的鲁棒性[22]。张家棋从住房类型、住房楼层、地理位置等各种因素角度出发构建房价模型，最后选择了其 中部分较为基础且对房价波动有着较大影响的 80 个特征，分别构建了多元回归模型 、前馈神经网络、XGBoost 算法，通过 RMSE 的比较，发现 XGBoost 模型的拟合

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

效果最好[23]。张望舒等通过对二手房价格建立 Lasso-灰色预测-随机森林组合模型，研究发现模型预测效果很好，拟合优度很高[24]。

综上所述，目前学者基于共享短租的研究主要基于 Airbnb 平台，有少数国内学者基于以 B2C 为运营模式的途家以及与 Airbnb 相似运营模式的小猪短租和蚂蚁短租来展开研究。本文研究目的是探究不同房东类型的共享短租房源的定价之间的差异性以及为两种房东类型给定合理的定价模型，所以选择以 C2C 为主要模式、能够给予房东更多主观定价权的 Airbnb 平台作为研究对象更为合适。通过对共享短租、酒店及住房价格影响因素方面的文献的梳理，将共享短租的房源价格的影响因素主要归纳为房源基本属性、房东属性、区位优势、在线评价四个方面。且过往学者对国内共享短租数据的实证研究较少，对不同房东类型共享短租的定价差异的研究也较缺乏。本文从不同房东类型出发，研究不同类型房东的定价差异，并为专业房东和非专业房东分别建立更为精确的定价模型。此外，在过往学者在房价价格预测模型的研究中，大部分研究成果都证明，机器学习中的一些非线性回归模型能更好地拟合涉及众多变量的房价类的预测问题，虽然深度学习模型在模型预测精度上有一定的优势，但可解释性不强，无法反映各定价影响因素对价格的具体影响效应，因此本文后续也会对利用线性回归构建的短租房源定价模型和基于机器学习非线性回归构建的短租房源定价模型进行综合比较，为两类房东构建预测精度最好、拟合优度更佳的定价模型。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 3 章 短租房源定价影响因素构建

3.1 数据来源

本文数据来源于 Airbnb 公开的北京短租房源的数据(http://insideairbnb. com/get-the-date.html)。这里选取 2021.4.30 披露的数据，累计获得 3 万个原始样本，共计 74 个变量。原始房源数据中包含房源基本信息、房东基本信息、房源在线评价三个方面的数据，其中房源的基本信息包含房源 id、房源名称、房源上传时间、房源描述文本、经纬度、房源属性类别、房间类别、可容纳人数、浴室数、卧室数、房源设施情况、房源预订规则等变量，房东基本信息包含房东 id、房东地址、房东姓名、房东注册时间、房东介绍、房东回复时间、房东回复率、房东接受率、是否超赞房东、房源拥有总量等变量，在线评价数据主要包含房源的总评论数、房源的综合评分、以及对于房源描述的准确性、清洁性、位置便利性、入住便利性、性价比等多个方面的评分。

本文所使用的其他官方数据，包含北京市地铁站、北京市公交站、北京市旅店和酒店的名称或地址信息均来自于北京市政务数据资源网。北京市热度 Top100 景点来自于去哪儿网 APP 所提供的排名。通过所提供重要地点的的名称和地址，调取高德地图 API 来获取对应地点的经纬度，并根据 haversine 公式量化出每个房源与地铁站、公交站、旅店和酒店、景点之间的距离来作为区位优势变量。

3.2 房源属性

3.2.1 房源显示特征

房源属性包含房源的显示属性和隐式属性。就显示属性来说，主要指的是房源的硬性条件，包含房源房间类型、可容纳人数、浴室数、卧室数、床位数、其他基本设施数等。其中房间类型包括整租房源、独立房间、共享房间三种类型。整租房源指独享房源的所有空间，不需要和别人共用，独立房间指的是有自己的独立的卧室，与他人共享其它空间，包括客厅、卫生间等，共享房间指的是与他人共享卧室以及其他空间。王春英基于小猪短租平台的数据建立对数回归定价模型，房间类型对短租房源定价影响显著，整套出租的房源价格比独立单间出租的房源住宿价格高 12%[25]；三种房源类型的价格差异显著，用户可以根据自己的需求和经济实力选择自己喜欢的房源。可容纳人数、浴室数、卧室数、床位数、床的类型代表的是房源的空间大小和实用性，王春英认为传统酒店的住宿定价主要受到其建造时或租借时的成本、房间标准和

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

酒店星级等因素的影响，而短租房源与传统酒店的定价不同，由于共享短租平台上的房源大多是房东的闲置资源。因此，房东在自主定价时，不会以当时的建造成本作为主要的定价参考，而是会考虑房源的房间类型、床铺数量、可容纳人数及配套设施等因素[1]。吴晓隽，通过回归分析也证明房源设施与房源价格呈现正相关[7]。房源的其他设施代表房源所提供的一些满足的其他的生活设施，一般包括是否有无线网络、厨房、吹风机、空调、洗衣机、衣架、电梯等，除此之外显示特征还包括房源的租赁规则，主要包含是否可以闪定、未来 30 天内可以入住的天数、未来 60 天内可以入住的天数、未来 90 天可以入住的天数以及未来一年可以入住的天数。未来多天内可以入住的天数代表着该房源的热度和受欢迎的程度，房源的热度也会影响到房东对于房源价格的及时设定，来提高自己的总收益。Beatriz 提出是否能即时预订信息与其商品在平台中的需求量，以及与商品最后的定价均呈现显著相关[26]。

3.2.2 房源隐式特征

房源的隐式特征包含短租房源描述的丰富性和主题的丰富性。短租房源相比于酒店的最大区别，就是它可以满足用户多元化的需求，一个短租房源要成为高热度房源，就必须拥有自己的特色或者主题。例如目前市场上比较流行的主题 Loft、电竞主题房、超大投屏房、别墅轰趴等；张乐利用 LDA 主题模型来挖掘房东描述中潜在的主题，并提出使用主题熵来衡量房源主题的丰富性，利用实词数来量化房源内容丰富性，来探究房东的信任影响因素[27]。对于房源活动内容的丰富性和房源的主题性、特色性，本文主要从房源的描述出发来刻画相关的指标，通过计算房源描述中的实词数以及主题熵，来量化房源的隐式特征。

3.2.3 隐式特征 LDA 主题模型

分词是在进行自然语言处理的第一步，也是很重要的一步。目前 jieba 分析是大多数学者广泛使用的，它不仅开源，并且分词的效果也较好。它包含三种分词模式。第一种是全模式，代表着尽量要把所有词切分出来。全模式虽然可以将词分的细致，但这也会导致出现比较多的冗余信息和额外浪费的计算力。第二种就是精确模式，视图进行准确的分词，在默认情况下就是使用的精确模式。第三种就是搜索引擎模式，就是在精确模式基础上，会对长词进行再次切分。本文首先调取了百度翻译 API 对房源描述为英文的进行翻译，然后使用正则化技术的手段，只保留描述中的中文、数字等

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

内容，在经过文本数据清洗后，使用 jieba 分词的精确模式进行分词，来为后续的主题分析做好数据铺垫。

LDA 是一种三层贝叶斯网络结构模型，Blei 认为一篇文本是没有固定顺序且由不分前后关系的一组词构成的，因此可以将文本中的词汇聚类成的几个主题，以此达到由“文本—词汇”的高位空间映射到“文本—主题”和“主题—词汇”的低维空间[48]。对于给定文本 LDA 主题模型计算公式如下：

$$P(\text{词汇}|\text{文本})=\sum_{i=1}^{K} P(\text{主题}|\text{文本})\cdot P(\text{词汇}|\text{主题}) \tag{2-1}$$

其中，$P(\text{词汇}|\text{文本})$ 表示文本中词汇的分布，可被直接观测。LDA 模型以动态链式的训练模式，通过已知 $P(\text{词汇}|\text{文本})$ 间接训练出文本聚成类别的主题数 K、文本中主题的分布 $P(\text{主题}|\text{文本})$ 及主题中词汇的分布 $P(\text{词汇}|\text{主题})$ 。

在文本挖掘领域，常用困惑度 $perplexity(D)$ 衡量 LDA 主题模型中概率模型性能的优劣，如果一个分布的困惑度指标得分较低，说明文本中主题出现的概率较均衡，模型较好的性能来预测新样本[28-29]。计算困惑度公式为：

$$perplexity(D)=\exp(-\frac{\sum \log\ p(\text{主题}|\text{文本})}{\sum_{d=1}^{M} N_d}) \tag{2-2}$$

$\sum_{d=1}^{M} N_d$ 为 M 篇测试数据中所有词汇的总和，在本文中指 20017 条“房间介绍”的文本数据通过 Python 的 jieba 库精确分词后根据词性来最后筛选得到最终的实词总数。计算出 $perplexity(D)$ 数值与聚类主题数关系如下图 3.1 所示，因此确定“房源介绍”的文本主题数为 6，即 LDA 主题模型性能最佳。

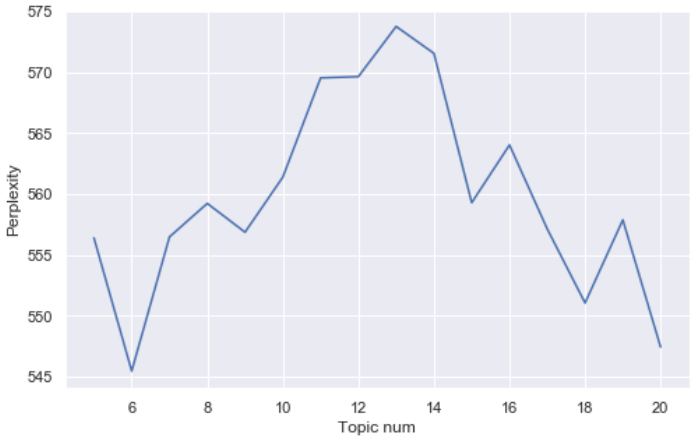


图 3.1 困惑度数值与聚类主题数关系图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

将“房间介绍”的文本数据中识别出的六个主题分别命名为“交通”、“环境”、“位置”、“体验”、“景点”和“设施”。表 3.1 中列举出了每个主题对应出现频率最高的五个词汇及该词汇属于对应主题的概率。

表 3.1 LDA 模型识别的主题、词汇概率汇总表

P(词汇|主题)	**主题 1**	**主题 2**	**主题 3**	**主题 4**	**主题 5**	**主题 6**
	交通	**环境**	**位置**	**体验**	**景点**	**设施**
词汇 1	分钟（0.017）	胡同（0.016）	五道口（0.010）	房间（0.014）	北京（0.020）	厨房（0.009）
词汇 2	地铁（0.016）	文化（0.009）	中关村（0.009）	请（0.010）	南锣鼓巷（0.008）	卧室（0.008）
词汇 3	距离（0.015）	院子（0.007）	大学（0.008）	提供（0.005）	步行（0.009）	入住（0.007）
词汇 4	号线（0.015）	民宿（0.007）	医院（0.007）	温馨（0.005）	故宫（0.008）	客厅（0.006）
词汇 5	公里（0.011）	四合院（0.006）	小区（0.006）	舒适（0.005）	雍和宫（0.006）	卫生间（0.006）

通过计算每条 “房源介绍”所包含六大类主题的信息总量，最终将文本数据转化为“主题熵”的数值型变量，用于衡量文本中潜在主题的多样性，即房源介绍文本中的内容丰富程度，主题熵的计算公式如式 2.3 所示，$E(d)$ 代表文本 d 的主题熵。

$$E(d)=-\sum_{i=1}^{K} p(\text{主题}\,|\,\text{文本})\log(p(\text{主题}\,|\,\text{文本})) \tag{2-3}$$

3.2.4 房源属性变量汇总

表 3.2 房源属性变量汇总

变量名称	属性解释
是否整租房源	是否整套房源出租
是否单人间	房间单人享用，其余区域共享
可容纳人数	房源可以居住的最多房客数
卧室数	房源里的卧室数量
浴室数	房源里的浴室数量
床数	房源里的床数量
其他设施数量	房源里的其他设施数量
未来 30 天可定天数	房源未来 30 天内可以预定的天数
未来 60 天可定天数	房源未来 60 天内可以预定的天数

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3.2 房源属性变量汇总（续）

未来 90 天可定天数	房源未来 90 天内可以预定的天数
未来 365 天可定天数	房源未来 365 天内可以预定的天数
最短预订天数	租户每次预订最少天数
最长预订天数	租户每次预订最多天数
是否闪订房源	房源是否可以立即预定
房源描述实词数	通过 jieba 分词后，每个房源描述中的实词数
房源描述主题熵	通过 LDA 主题模型确定的房源的主题熵

3.3 房东属性

短租房源与酒店相比，具有其产品的独特性。用户在选择酒店时，在选择好自己感兴趣的酒店房源后，可以经过平台进行直接预定。而租户在选择好自己感兴趣的短租房源后，往往需要与房东进行充分地沟通，因为租户和房东之间需要建立互相的信任关系，租户在预定后可能也需要和房东同住一个房源。房东能够给予到租户的信任度，在一笔短租房源订单中起着非常重要的作用。目前 Airbnb 平台中已有的房东的信息主要包含：是否超赞房东、房东注册天数、房东总房源数、房东是否认证、房东认证方式数、房东回复率、房东回复时长、房东接受率。其中超赞房东，是平台内部通过一定标准评选的，规则如下 1、至少完成 10 次行程或完成 3 次预订，接待住宿的总晚数要达到 100 晚。2、回复率不低于 90%。3、取消率不高 1%(即每 100 笔预订只有 1 笔取消)。4、总体评分保持在 4.8 (此评分依据一年内的评价而定)。结合数据实际的情况，主要涉及的变量如下表 3.3：

表 3.3 房东属性变量汇总

变量名称	属性解释
是否超赞房东	根据平台一定规则判定的标签，1 代表超赞房东
是否专业房东	房东房源拥有数量大于 1 为专业
房东注册天数	房东自注册账号起距今天数
房东总房源数	房东拥有的总的短租房源数量
房东是否认证	房东是否有通过平台的身份认证
房东认证方式数	房东在平台内的认证方式数
房东回复率	房东回复租户问题的概率
房东回复时长	房东回复租户的平均时长
房东接受率	房东接受预定的概率

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.4 区位优势

房源的区位优势受到多种因素的影响，在本文中会基于第 2 章的文献梳理，通过 Python 调取高德地图 API 来获取北京市热度 Top100 的景区、地铁站、公交站、旅馆和酒店的经纬度，最后通过 haversine 公式来计算短租房源与交通枢纽、超市、景区的直线距离。北京市不同区的房源价格方差较大，不同地区的二手房平均价格很大程度上会影响到短租房源的价格，所以房源是否在主城区对于房源的价格影响较大。

王春英通过研究发现各区二手房平均价格、与城市旅游中心距离对短租房源的价格呈显著性影响[25]。对于旅客来说，便捷的交通环境可以帮他们节省更多的时间来实现旅游效益的最大化，本文中选择用短租房源距离公交站、地铁站最近的距离，以及 1 公里以内的公交站数和地铁站数来量化短租房源的交通优势。短租房的需求更多的是接待外来的旅客，旅客最关心的就是距离目的景区的距离，旅客在选择短租房源的时候，更多是根据距离景区的距离来进行排序选择，这里通过计算了短租房源 1km 内的北京热度前 100 的景点数。

对于生活消费类这一种可选择的经济模式，周边竞争对手数量也会对房源价格产生影响，Balaguer 等在探讨酒店房间价格影响因素时，纳入了与竞争者的距离这个外部因素进行讨论，认为其对价格的影响不是简单线性正或负相关[53]。ZERVAS 通过研究结果发现，Airbnb 上房源数量每增加 10%，会导致酒店收入下降 0.37%[30]，这里通过计算 1 公里内的旅馆或酒店数目来量化竞争对手带来的影响。

表 3.4 区位优势变量汇总

变量名称	属性解释
该区二手房均价（万元/㎡）	北京每个区二手房平均价格
1km 内的地铁站数	1km 直线距离内的地铁站数
1km 内的公交站数	1km 直线距离内的公交站数
1km 内的景点数	1km 直线距离内的景点数
1km 内的旅馆或酒店数	1km 直线距离内的旅馆或酒店数
最近的地铁站距离（km）	距离房源最近的地铁站距离
最近的公交站距离（km）	距离房源最近的公交站距离
最近的景点距离（km）	距离房源最近的景点距离
最近的旅馆或酒店距离（km）	距离房源最近的旅馆或酒店距离

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.5 在线评价

过往学者对于在线评论对短租房源价格的影响争议较大，其中 Zhang 对田纳西州纳什维尔地区的房源研究发现评论数量、评论分数与短租房源价格之间存在显著的负相关[4]而吕姝[7]以及 Gutt [6]等人通过研究发现租客过往的评论数量越大，短租房源的价格越高。Ert 等[31]通过研究发现，短租房源价格与其过往评分并无显著性的关系。本文选择从房客历史评分数据以及房客历史评论数据来刻画该房源的过往的信誉特征，其中房客历史评分数据包括综合评分、描述准确性评分、干净整洁评分、入住顺利评分、沟通交流评分、位置便利评分、性价比评分，房客历史评论数据包括总评论数、月平均评论数、首次评论时间、最新评论时间。

表 3.5 在线评论变量汇总

变量名称	属性解释
综合评分	评价综合评分
描述准确性评分	对房源描述真实性评分
干净整洁评分	对房源的整洁性评分
入住顺利评分	对房东的入住引导评分
沟通交流评分	对房东的服务评分
位置便利评分	对房源位置的评分
性价比评分	对房源价格体验的评分
总评论数	房源自上线以来的总评论数
月平均评论数	近一年的月均评论数
首次评论时间	首次评论距离数据截止时间的时间间隔（天）
最新评论时间	末次评论距离数据截止时间的时间间隔（天）

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 4 章 短租房源定价影响因素预处理

4.1 异质性房源定义

通过梳理文献，过往学者的研究成果中已经证明不同类型房东的短租房源定价存在显著性的差异。牛阮霞等在基于蚂蚁短租平台进行共享住宿平台房源价格影响因素研究时，根据房东在短租平台上的房源数量是否大于 1 将房东区分专业房东和非专业房东[14]。陈子燕等在研究不同房东类型的定价机制时，也将房东在平台上的房源拥有量来作为判断房东专业性的规则，拥有量大于 1 即为专业房东[32]。吴晓隽在研究 Airbnb 房源价格影响因素的时候，将房源数大于 5 的房东归纳为专业房东，否则定义为非专业房东[15]。

为了后续对不同类型房东的房源进行差异化定价，经过汇总过往学者对于专业房东和非专业房东的定义，本文选择将拥有的房源数量大于等于 2 的房东作为本次研究的专业房东，拥有房源数量为 1 的房东作为本次研究的非专业房东。

4.2 数据预处理

由于短租房源类型的多样性，所以的价格方差较大，波动较大，本文通过使用房源的价格比上可容纳人数来构建房源人均价格，来作为本文的研究对象。

鉴于价值低的信息会降低数据的可信度，因此本文在数据分析前对短租数据进行预处理，目的在于减少噪音数据的干扰。在本文中，暂且不考虑节假日波动对于价格的影响，所以首先对 Airbnb 平台房源近几个月数据的房源价格进行均值处理。核查其他房源信息基本不发生变动的变量，截取 2021 年 4 月 30 日的单日数据为准。随后，在通过以下步骤的数据预处理后共计得到有效房源样本数据 20017 条，其中专业房东房源数据 15398 条，非专业房东房源数据 4619 条。

在 Airbnb 平台中租客承担的房价不仅仅是单日房租，还可能承担一定的清洁费，且对房东而言清洁费用也算是房东出租的一部分营收。因此本文结合实际情况将单日房租与清洁费加总合成房源的“实际房价”，后文所分析和预测的房源价格都是“实际房价”意义上的价格。

1、房源真实性处理

第一，删除不真实房源。对“房源名称”数据进行查找，删除房源名称包含“测试房源”、“下架房源”、“不能租”等相关内容的房源，并删除重复房源。前类房源均已失效，后类多是为了提高平台曝光度的重复房源。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二，删除无效房源。删除“实际房价”为 0、“未来 365 天内可预定天数”为 0 和“最短预定天数”数值>365 的无效房源。“未来 365 天内可预定天数”是指在之后的一年中未被用户预定，剩余可在线预定的总天数，而“最短预定天数”是指预定房源时用户单次预定的最短时长，以上几类房源在平台上已不具备短租的条件，大概率已下架。

第三，查找可疑房源。依次查看“实际房价”、“浴室数”等变量的数值分布情况，处理明显不合常理的异常房源，根据箱线图结果，依据异常识别准则剔除小于 QL-1.5IQR 和大于 QU+1.5IQR 的房价异常房源。对其他异常变量的房源做如下处理：

表 4.1 不真实房源的部分数据列举表

序号	房源 ID	房东拥有房源总数	房源类型	可住人数	浴室个数	卧室个数	床个数	单日房价	处理方式
1	10272873	2	整租	4	101.5	2	2	429.00	删除
2	28292530	4	单间	2	16.5	1	1	329.00	删除
3	31244706	20	单间	2	20	20	30	168.00	修改
4	33619980	1	单间	16	50	1	50	242.00	删除
5	28795201	8	床位	1	2	1	8	87.00	修改

上表中的 5 个房源中，1、2、4 号房源的浴室数量明显与“可住人数”、“卧室个数”及“单日房价”等数据不匹配，可能是房东在输入信息时登记错误但目前无法查证，因此直接删除。3 号房源经查实是一家拥有 20 间客房的旅店式民宿，其数据的真实含义是旅店式民宿的客房有标间、大床两种规格，每间客房可容纳 2 人，配备 1 间浴室、1 间卧室、定价为 168 元/晚。因此将其“浴室个数”、“卧室个数”、“床个数”分别修改为 2、2、1.5，保留该样本数据。5 号房源经核实是一家位于北二环附近的八人间共享房源，主要为出门旅游的青年大学生们提供一个廉洁的住宿服务，其位置优越、交通方便、价位低廉是青年旅行者们出游时的热门选择。该房源数据的真实含义是青旅式民宿配有 1 间八床的合住卧室、2 间共享浴室，单次可预定一个床位入住 1 人，定价为 78 元/晚。因此按照均摊资源的原则，将其“浴室个数”、“卧室个数”、“床个数”分别修改为 0.25、0.125、1，同样保留该样本数据。以此类推处理以上两类可疑房源。

2.文本数据预处理

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（1）数字化赋值

首先，将是否为专业房东、所属区县、房源类型、能否闪订、是否为超赞房东、是否实名认证、大型超市地址、地铁站地址八个文本数据数字化，具体如表 4.2 所示：

表 4.2 变量及变量转换对应表

变量	变量转换描述
是否为专业房东	以房东拥有房源数的多少作为区分专业房东和非专业房东的根本标准，将拥有 1 套房源的房东定义为非专业房源并赋值为 0，将拥有多套房源的房东定义为专业房东并赋值为 1。
能否闪订	提供闪订服务记为 1，不提供闪订服务记为 0。
是否为超赞房东	超赞房东记为 1，不是的记为 0。
是否实名认证	实名认证记为 1，未实名认证记为 0。
地铁站地址	通过爬虫爬取地铁站对应经纬度，根据 haversine 公式计算出短租房源与周边地铁站的距离，转化为“短租房源 1km 内地铁站数”
公交站地址	通过爬虫爬取公交站对应经纬度，根据 haversine 公式计算出短租房源与周边公交站的距离，转化为“短租房源 1km 内公交站数”
旅店或酒店地址	通过爬虫爬取旅店或酒馆对应经纬度，根据 haversine 公式计算出短租房源与周边旅店或酒店的距离，转化为“短租房源 1km 内旅店或酒店数”
景点地址	通过爬虫爬取北京市热度前 100 的景点对应经纬度，根据 haversine 公式计算出短租房源与景点的距离，转化为“短租房源 1km 内景点数”

（2）中英文转化

对于房源描述中存在英文、日文等其他语言的房源，本文中利用 Python 调取百度翻译 API 来进行外文的转化，对经过判断存在其他语言的房源描述进行遍历翻译，为后续主题建模做好数据准备。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3 短租房源描述性分析

4.3.1 房源分布

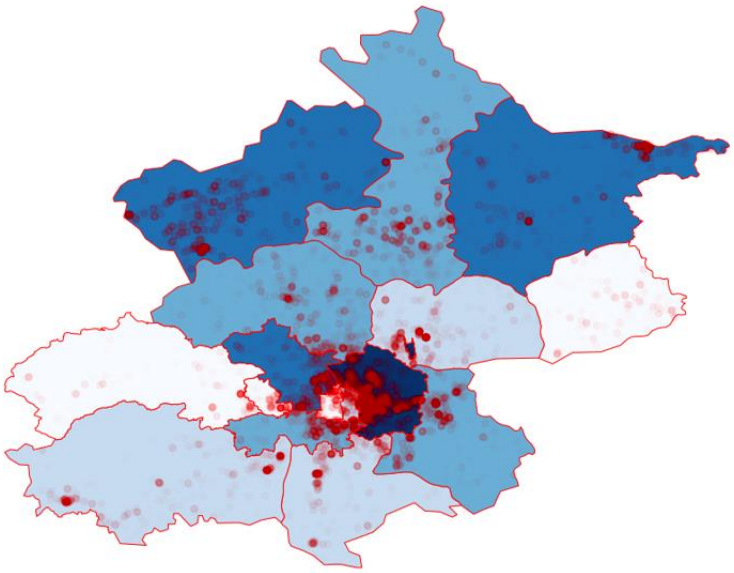


图 4.1 北京市短租房源密集度分布

北京市共划分为十六个区，包括“市中心”：东城区、西城区、崇文区、宣武区；“近郊区”：朝阳区、丰台区、石景山区、海淀区；“远郊区”：延庆区、怀柔区、密云区、昌平区、顺义区、平谷区、通州区、大兴区、房山区、门头沟区。图 4.1 中通过底图颜色的深浅表示各区块房源的数量多少，由图可知房源分布在区块有明显的差别，北三环内的朝阳区、东城区、海淀区占地面积虽小但包含了北京市绝大多数的短租房源，丰台区和西城区次之，门头沟区和平谷区最少。

4.3.2 房源周边景点分布

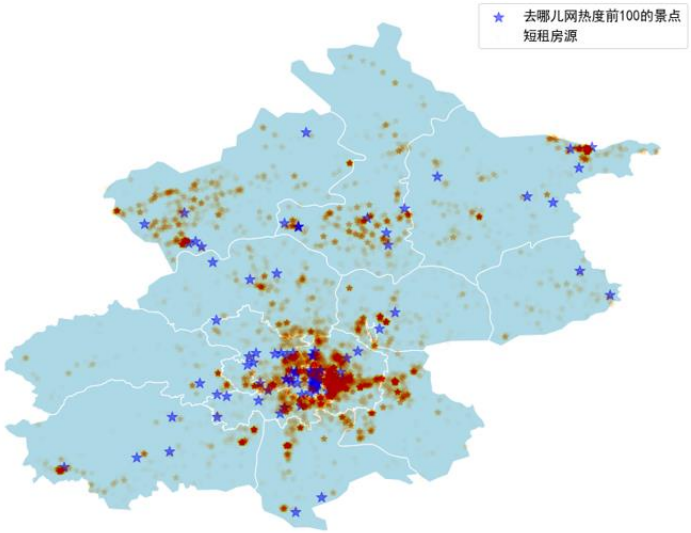


图 4.2 北京市短租房源与排名前 100 景点分布

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

通过房源经纬度将房源转换为点数据，图 4.2 中房源分布通过黄点标识。图中房源分布密度大致以北三环聚集地为中心，向四周由高密度区向低密度区散开。同时，本文根据“去哪儿”网站上北京景点的人气排名，通过调用高德地图 API 获取到北京市排名前 100 热门景点的经纬度，将景点数据也转化为点数据，在图 4.2 中景点分布通过蓝星标识。可以看出房源高密度区的地方多为景区汇集处，低密度区的房源分布虽然比较分散但大多也分布在景点周围。

4.3.3 房源周边交通站点分布

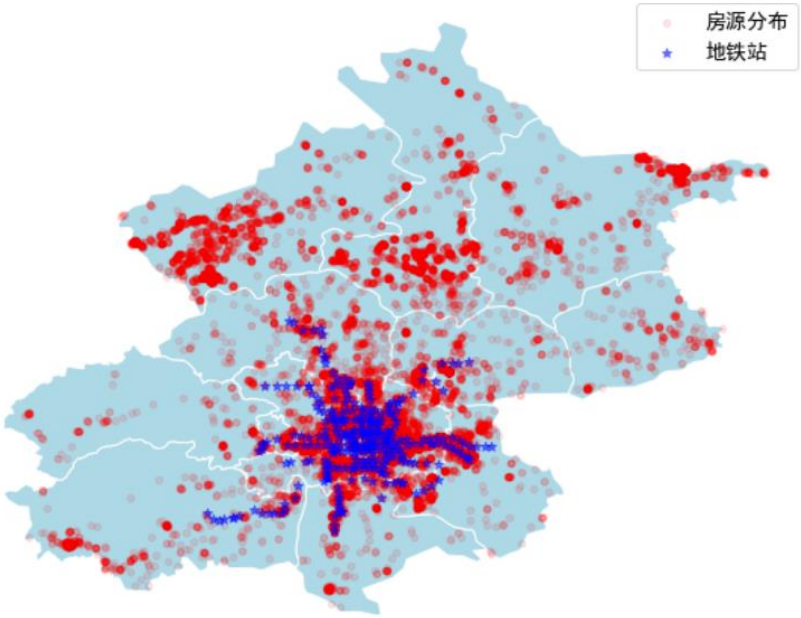


图 4.3 北京市短租房源与周边地铁站分布

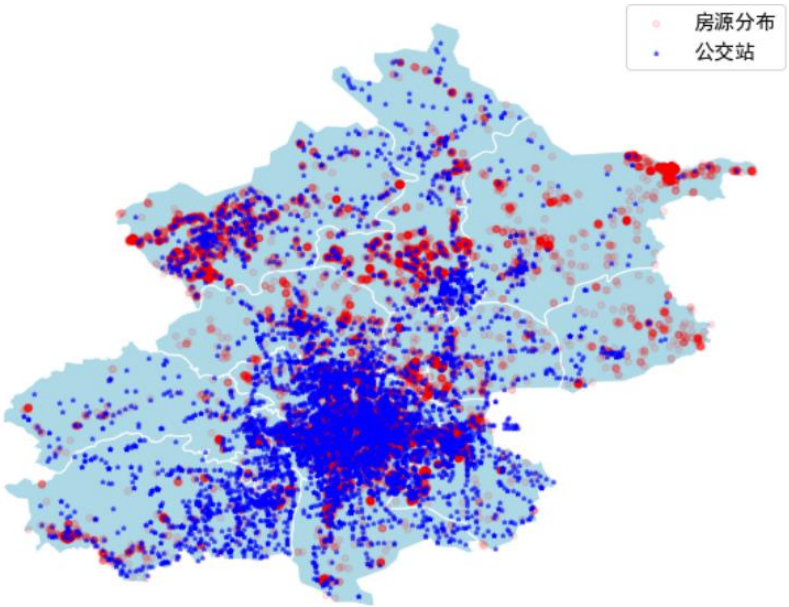


图 4.4 北京市短租房源与公交站分布

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据北京政务数据资源网上提供的北京市地铁站名称和公交站名称，通过调用高德地图 API 获取到它们对应的经纬度，将公交站和地铁站转换为地图上的蓝色星星标识。通过图 4.3 和图 4.4 可得，短租房源多数分布在公交站点密集处，且公交站越密集的区域短租房源的分布也越密集，同时和地铁站之间的分布规律也是如此，地铁站点主要分布于海淀区、朝阳区、丰台区、昌平区、大兴区、顺义区，在该几个区域内的短租房源最为密集。

4.3.4 房源周边旅馆分布

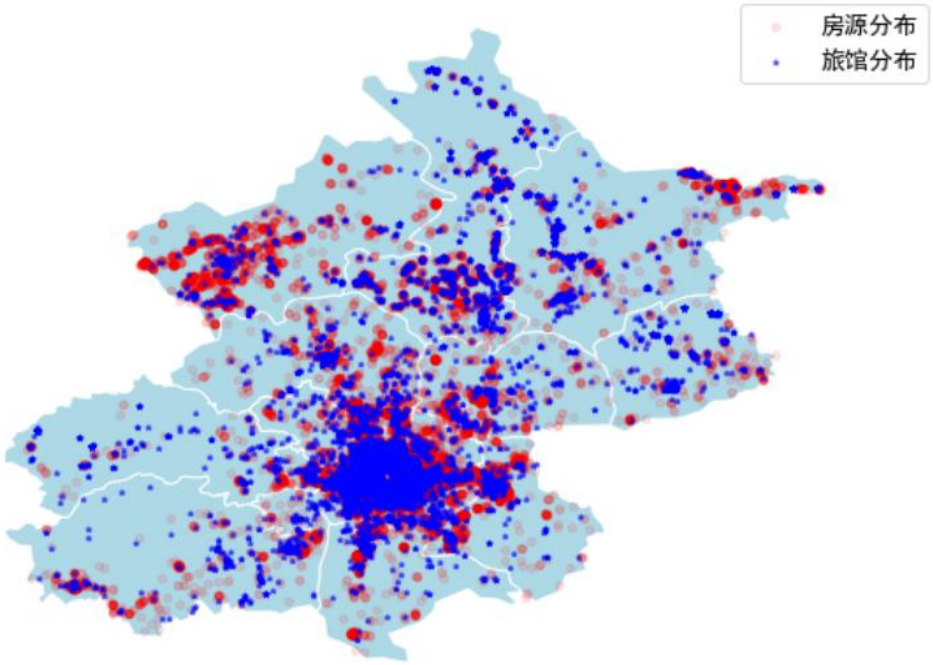


图 4.5 北京市短租房源与周边旅馆酒店分布

根据北京政务数据资源网上提供的北京市旅馆名称及地址，通过调用高德地图 Api 获取到它们对应的经纬度，将旅馆蓝色星星标识。通过图 4.5 可得，短租房源和旅馆分布具有极为相似的特征，两者主要都密集分布在海淀区、朝阳区、丰台区、昌平区、怀柔区，两者与景点的分布也是极为密切的关系，存在很大的竞争关系。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.5 各区房源价格分布

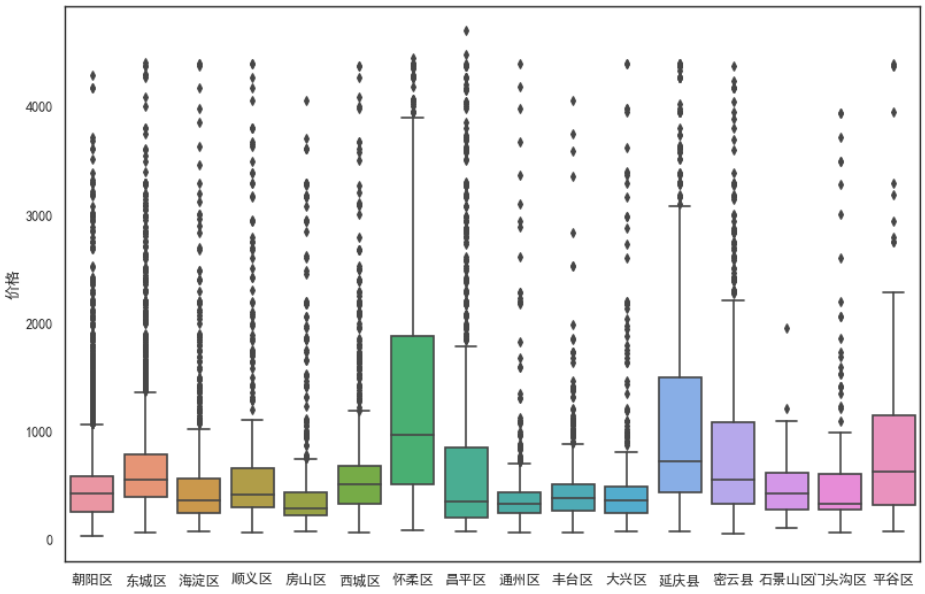


图 4.6 北京市各区房源价格箱线图

从图 4.6 中可得，怀柔区短租房源价格总体水平最高，且高端房源数量较多，可能是该区块房源地段较好，特色民宿占比较高等。短租房源价格水平排名前四的分别有怀柔区、延庆区、平谷区及密云县，短租房源价格水平排名倒数前四的分别为山头沟区、房山区、通州区及昌平区。而在前文分析得出的房源分布较多的朝阳区、东城区、海淀区属性价比较高的区位，短租房源价位在 300-500 内的居多，且地处都市繁华的北三环附近，满足多数短租者对价位和地理位置的期望，因此广受欢迎。

4.3.6 不同房源类型价格分布

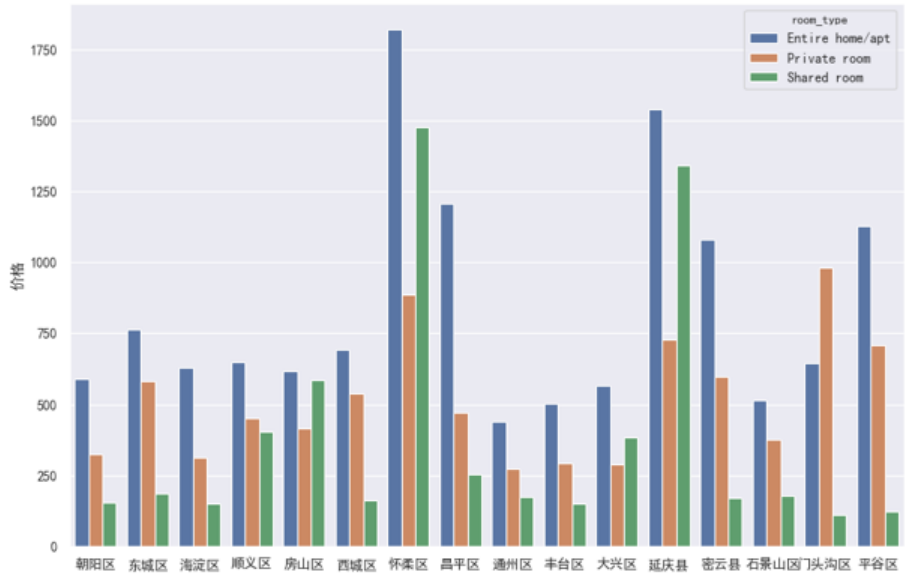


图 4.7 北京市不同房间类型的短租房源价格分布图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

从房间类型来看，存在显著性的差异，在 16 个区中，整租房源的价格明显最高。除了在房山区、怀柔区、大兴区、延庆县之外，其余区县的拥有私人房间的房源价格均高于共享房间的房源的价格。

4.3.7 房东属性差异

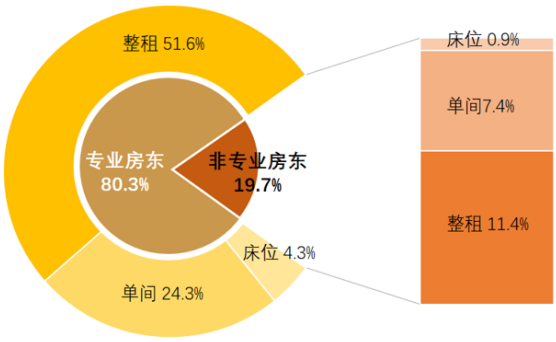


图 4.8 专业/非专业房东各类房源占比图

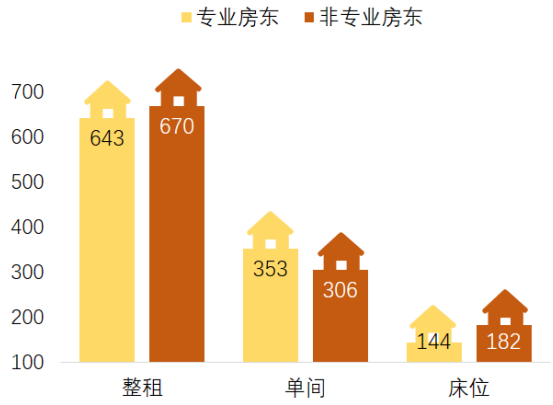


图 4.9 专业/非专业房东的各类房源定价均值图

Airbnb 在北京市的房源大部分由专业房东持有运营，且专业房东特别热衷经营整租房源，而非专业房东对单间房源的经营意愿比专业房东高。由图 4.8 可知，专业房东持有房源占比 80.3%，整租房源约占专业房东所有房源的 64.3%；非专业房东持有家房源占比 19.7%，单间房源约占非专业房东房源的 37.6%，同比高于专业房东的 30.3%。在专业和非专业房东运营的所有整租、单间、床位三种房源类型中，整租房源都是数量最多占比最高的，可见现行短租市场对于整租房源的需求也是最高的。其次，具体来看两类房东对不同房源的定价也略有不同。如图 4.9 可知，非专业房东的整租房源和床位房源的定价相对略高，专业房东单间房源的定价相对略高。

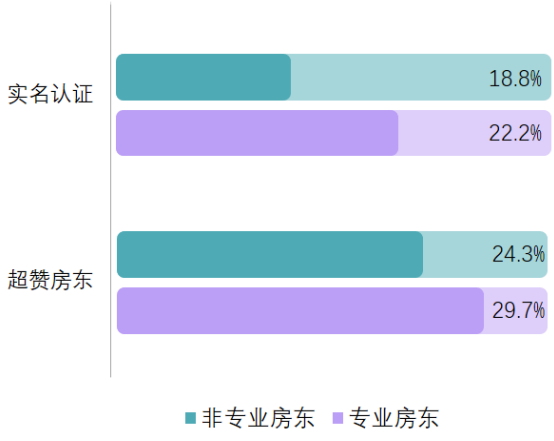


图 4.10 专业/非专业房东实名认证与超赞房东占比图

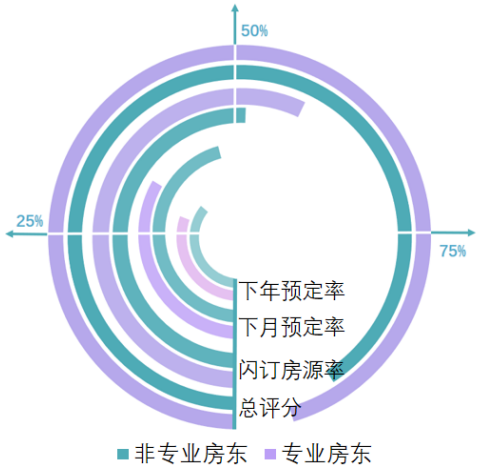


图 4.11 专业/非专业房东房源属性对比图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

房东的“超赞房东”身份是短租平台综合房东的订单接待量、信息回复率、订单取消率及总体评分四个维度衡量的，实名认证确保了房东的真实性和信用程度。总体来看(见上图 4.10、图 4.11)，虽然两类房东的实名认证率和超赞房东率都不是很高，但是专业房东拥有更丰富的经营管理经验，在平台建设、身份核实、订单成功率、闪订房源率等方面还是普遍优于非专业房东。另外，非专业房东的房源预定率更高，不管是未来一年内的预定率或是未来一月的预订率都略高于专业房东。

4.3.8 短租房源词云图

图 4.12 中英文房源介绍词云图

对“房间介绍”的海量文本数据进行词频统计分析，并以出现频率 top100 的中文实词、英文实词生成词云图，词汇的字体越大表示该实词被提及的频率越高。综合来看（图 4.12），房源的中英文介绍关键词较为一致，两张词云图中的高频词主要围绕着房客的房源配置和入住体验，如房源类别（公寓 apartment、房间 room、住宅 home）房源位置（位置 location、地铁站 subway、交通 traffic、 景点 Sights）、便利程度（超市 supermarket、餐馆 restaurant、便利 convenient）、服务承诺（舒服 comfortable、温馨 kind、热情 friendly）、房源环境（干净 clean、安静 quiet、优越 helpful）等。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 5 章 短租房源定价影响因素实证分析

5.1 逐步回归模型及 VIF 值

5.1.1 逐步回归

当回归模型中使用两个以上变量时，这些变量往往会提供一些多余的信息，也就是自变量之间存在一定的相关性，也就产生了多重共线性的现象。多重共线性可能会对参数估计值的正负号产生影响，也就是会影响参数的实际经济意义，进而把分析引入歧途。针对于多重共线性，统计学中常常通过向前选择、向后剔除、逐步回归的方式进行解决。

其中，向前选择步骤通常如下：1、对 k 个自变量（$x_1, x_2, ...x_k$），分别与因变量 y 拟合建立一元线性回归模型，在这些模型中找出 F 统计量值最大的模型以及对应的自变量 x_i，并将该变量首先引入模型。2、在第一步引入 x_i 的基础上，在分别你和引入其他 $k-1$ 自变量（$x_1, x_2, ...x_k$）的线性回归模型，形成 $k-1$ 个双变量的变量组合，分别考察这些模型，挑选出 F 统计量值最大的模型，然后将最大的那个自变量 x_j 引入模型。

3、如此反复进行，直至模型外的自变量均无统计显著为止。

向后剔除法与向前选择法相反，主要是先将所有的变量都假如模型内，然后建立 k 个模型，每个模型都剔除其中的一个自变量，选择使得模型 SSE 值减小最少的自变量从模型中剔除，直至剔除一个自变量不会使 SSE 值显著减小为止。

逐步回归法，是将向前和向后两种方法结合起来筛选自变量的方法。前面两步的步骤与向前选择法相同。但是在增加一个变量之后，它会对模型中所有的变量都进行探察。分析是否有在后续变量加入后，前面加入的自变量对模型的贡献变得不显著，如果不显著则会删除。因此逐步回归法就是，按照这样的方式不停地增加变量并考虑剔除之前加入变量的可能性，直到再加入一个变量的时候，SSE 值不再显著减小为止。

5.1.2 方差膨胀因子

方差膨胀因子是容忍度的倒数。某个自变量的容忍度等于 1 减去该自变量作为因变量而其他 $k-1$ 变量作为自变量时构建的线性回归方程的拟合优度 R_i^2，即

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$VIF = \frac{1}{1-R_i^2} \tag{5-1}$$

VIF 值越大，说明模型存在的多重共线性越严重，当 VIF 值>10 时，认为存在严重的多重共线性。

5.2 专业房东逐步回归模型

5.2.1 专业房东 VIF 值检验

由于逐步回归可能会由于因变量之间存在的多重共线性而出现误删的情形，所以本文选择先计算所有变量的方差膨胀因子，当方差膨胀因子大于 10 时进行手动剔除，并且当出现多个相似变量导致的 VIF 大于 10 的情形，选择将 VIF 更大的变量进行剔除，直到剩下的变量的 VIF 值均不大于 10 为止。

表 5.1 所有变量的 VIF 值

变量	VIF	变量	VIF
未来 60 天可定天数	67.03	位置便利评分	2.24
未来 90 天可定天数	34.04	每月评论数	2.08
最新评论时间	29.30	最近公交站距离	1.85
首次评论时间	26.78	1km 内的地铁站数	1.75
未来 30 天可定天数	13.17	最近的景点距离	1.70
是否是整租房	6.91	1km 内的旅馆酒店数目	1.50
是否是单人间	6.77	房东接受率	1.49
综合评分	5.74	未来 365 天可定天数	1.48
性价比评分	4.54	房东注册天数	1.43
可容纳人数	4.49	是否闪订房源	1.41
描述准确性评分	4.43	实词数	1.38
入住顺利评分	3.94	最近旅馆距离	1.36
卧室数	3.49	是否超赞房东	1.33
沟通交流评分	3.48	房源是设施数	1.33
干净整洁评分	3.45	房东认证方式数量	1.31
最近的地铁站距离	2.95	1km 内的景点数	1.29
卧室数	2.93	主题熵	1.24
房东回复率	2.89	房东总房源数	1.15
房东回复时长	2.74	1km 内的公交站数目	1.13

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.1 所有变量的 VIF 值（续）

该区二手房均价（万元/㎡）	2.54	最短预定天数	1.03
浴室数	2.53	最长预定天数	1.02
总评论数	2.50	房东是否认证	1.02

如表 5.1，对专业房东所有变量的 VIF 值，发现 60 天内可预订天数、90 天内可预订天数、上一次评论距今时间、首次评论距今时间的 VIF 均明显大于 10。在选择将 60 天内可预订天数、90 天内可预订天数、首次评论距今时间进行剔除后，得到的方差膨胀系数如下表 5.2，目前 VIF 值均已经小于 10。

表 5.2 专业房东剔除 VIF>10 的变量后的变量 VIF 值

变量	VIF	变量	VIF
是否是整租房	6.91	1km 内的地铁站数	1.78
是否是单人间	6.76	最近的景点距离	1.50
综合评分	5.73	1km 内的旅馆酒店数目	1.49
性价比评分	4.53	房东接受率	1.47
可容纳人数	4.49	最新评论时间	1.47
描述准确性评分	4.43	房东注册天数	1.40
入住顺利评分	3.94	是否闪订房源	1.38
卧室数	3.49	实词数	1.35
沟通交流评分	3.48	最近旅馆距离	1.33
干净整洁评分	3.45	是否超赞房东	1.33
最近的地铁站距离	2.94	房源是设施数	1.30
卧室数	2.93	房东认证方式数量	1.28
房东回复率	2.88	1km 内的景点数	1.25
房东回复时长	2.73	主题熵	1.14
该区二手房均价（万元/㎡）	2.54	房东总房源数	1.14
浴室数	2.53	未来 30 天可定天数	1.12
位置便利评分	2.24	1km 内的公交站数目	1.03
每月评论数	1.98	最短预定天数	1.02
总评论数	1.87	最长预定天数	1.02
最近公交站距离	1.86	房东是否认证	1.01

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5.2.2 专业房东逐步回归

表 5.3 专业房东逐步回归结果

	非标准化系数		标准化系数	T	显著性
截距项	B	标准误差	Beta		
变量	-62.840	14.11		-4.45	0.00
最近的地铁站距离	1.155	0.00	0.23	18.37	0.00
实词数	0.235	0.02	0.11	13.47	0.00
1km 内的景点数	14.972	1.41	0.09	10.63	0.00
是否是整租房	97.777	4.19	0.45	23.36	0.00
是否是单人间	74.863	4.25	0.33	17.63	0.00
1km 内的旅馆酒店数目	0.349	0.04	0.08	9.47	0.00
未来 30 天可定天数	1.102	0.08	0.10	13.64	0.00
浴室数	24.006	1.13	0.24	21.22	0.00
可容纳人数	-7.167	0.50	-0.22	-14.42	0.00
最新评论时间	0.013	0.00	0.11	13.02	0.00
房源设施数	1.188	0.11	0.09	10.72	0.00
该区二手房均价(万元/㎡)	2.441	0.43	0.07	5.64	0.00
最近旅馆距离 Km	13.17	0.00	0.08	9.76	0.00
是否闪订房源	9.184	1.78	0.04	5.16	0.00
床位数	-3.233	0.51	-0.08	-6.35	0.00
总评论数	-0.510	0.07	-0.08	-7.55	0.00
最近的景点距离	1.084	0.00	0.05	5.96	0.00
最短预定天数	-0.203	0.04	-0.04	-5.42	0.00
房东注册天数	0.006	0.00	0.03	3.48	0.00
1km 内的地铁站数	4.079	0.91	0.04	4.48	0.00
干净整洁评分	7.420	1.68	0.06	4.41	0.00
性价比评分	-15.399	1.83	-0.13	-8.44	0.00
综合评分	1.032	0.20	0.09	5.13	0.00
位置便利评分	7.007	1.58	0.05	4.44	0.00
卧室数	3.381	1.03	0.05	3.30	0.00
沟通交流评分	-5.948	1.80	-0.04	-3.31	0.00
房东回复率	0.209	0.04	0.06	4.80	0.00
房东回复时长	-5.627	1.50	-0.05	-3.75	0.00

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.3 专业房东逐步回归结果（续）

每月评论数	3.683	1.32	0.03	2.80	0.01
主题熵	7.630	2.76	0.02	2.76	0.01
房东认证方式数量	0.842	0.33	0.02	2.56	0.01
最近公交站距离 km	2.378	0.00	0.02	1.98	0.05

利用 SPSS 软件对于 15398 个专业房东的短租房源进行逐步回归，对已经剔除大 VIF 值的所有变量进行逐步回归的操作，最后筛选出的模型的变量如上表 5.3，根据逐步回归得到的结果，计算得到模型的 R 方达到了 62%，逐步回归模型为：

房源人均价格 = -62.84+1.155*最近地铁站距离+0.235*实词数+14.972*1km 内景点数+97.777*是否整租房+74.863*是否单人间+0.349*1km 内的旅店酒馆数+1.102*未来 30 天可定天数+24.006*浴室数-7.167*可容纳人数+0.013*最新评论时间+1.188*房源设施数+2.441*该区二手房房价+0.013*最近旅馆距离+9.184*是否闪订房源-3.233*床位数-0.510*总评论数+1.084*最近的景点距离-0.203*最短预定天数+0.006*房东注册天数+4.079*1km 内的地铁站数+7.42*干净整洁得分-15.4*性价比得分+1.032*综合评分+7.007*位置便利评分+3.381*卧室数-5.948*沟通交流评分+0.209*房东回复率-5.627*房东回复时长+3.683*每月评论数+7.63*主题熵+0.842*房东认证方式数-0.002*最近公交站距离

用专业房东逐步回归的建模过程，最终保留了如下 32 个变量，其中包含房源属性变量：是否闪订房源、是否是整租房、是否单人间、可容纳人数、床位数、卧室数、浴室数，最短预定天数、未来 30 天可定天数、实词数、主题熵。其中可容纳人数、床位数对房源人均价格来说呈负向影响，其余变量均呈现正向影响。其中房源类型对房源人均价格起着非常重要的作用，从回归方程的经济含义来说，当短租房源其他属性不变的情况下，整租房源会比其他房源平均贵 98 元，独立房间会比其他短租房源平均贵 75 元。是否闪订房源对于房源人均价格也有很大的权重，闪订房源的价格比非闪订房源平均贵 9 元，另外一个房源如果可容纳人数、床位数过多，很有可能是多人合租间，空间相比来说会更加拥挤，房源的价格也理应更低，在专业房东房源的其他属性保持不变的情形下，可容纳人数每增加一个，房源人均价格平均下降 7.1 元，床位数每增加一个房源人均价格平均下降 3.23 元。卧室数与浴室数的数量越多，代表房源的空间越大，对于房源的人均价格起正向影响，每增加一个卧室数，价格平均增长 3.4 元，每增加一个浴室数，价格平均增长 24 元。房源设立的最短预定天数越大，就会错失一

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

些只住一晚的租户，从而对价格产生负向影响，未来 30 天内的可定天数越少，会减少一些想要常住的租户，从而对价格产生负向影响。实词数和主题熵，衡量了该房源描述中所包含的内容丰富性以及主题的丰富性，衡量着每一个房源的独特性，短租房源受到年轻一代的喜爱，正是因为短租房源的多样性和独特性，所以这两个变量对价格起着正向影响，其中主题熵每增加一个单位，房源人均价格平均增加 7.6 元。

房东属性中保留的变量有：房东的注册天数、房东的认证方式数量、房东回复率、房东的回复时长。其中除了房东的回复时长呈现负向影响之外，其余均对房源价格呈现正向影响。房东的注册天数越久，认证方式越多，说明房东以及房源真实性越强，更能赢得租客的信任。房东的回复率越高，说明房东对于租客更加看重，更加热情，由于短租房源的特性，房东对于租客的回复介绍是非常重要的。另外，房东回复时长占有较大的权重，回复时长每增加一个单位，房源人均价格平均降低 5.6 元，房东的回复时长越短，租客能够及时的获取到房源的信息，对于后续房源的预订起着积极作用。

在线评价属性中保留的变量有：综合评分、干净整洁评分、沟通交流评分、位置便利评分、性价比评分、总评论数、每月评论数、最新评论时间，其中除了性价比评分与房源人均价格呈现负向影响外，其余均为正向。性价比评分，是租客出于对房源价格的实惠型进行的打分，所以与房源的人均价格呈现负向。其余评分均是代表租客对于房源各个方面的认可程度，租客认可度越高，房东对于自己的房源也越自信，从而对定价产生影响，其中干净整洁评分和位置便利评分的正向影响权值最大，两者评分每增加一个单位，价格平均增加 7 元。其次，房源的总评论数、每月评论数代表着房源的历史成交情况，评论数越多代表房源在历史记录中的受欢迎程度，最新评论数反映的是房源近期受欢迎的程度，所以均对房源人均价格起着正向影响、

房源区位优势保留的变量有该区二手房均价（万元/㎡）、1km 内的景点数、1km 内的地铁站数、1km 内的旅馆酒店数、最近公交站距离、最近旅馆距离、最近的景点距离、最近的地铁站距离。房源区位优势的变量对房源的人均价格均有显著的正向影响，对于一个短租房源亦或是酒店来说，位置优势均是核心优势，在模型中，1km 内的地铁站数每增加一个单位，房源人均价格平均增加 4 元。短租房源的受众也有很大一部分来自于外来的旅客，而旅客在考虑酒店或是租房时，最关注的就是周边的交通

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

设施以及与目的地景点的距离，来最大化自己的实际游玩时间，减少在交通等其他方面上时间的耗费，1km 内的景点数每增加一个单位，房源人均价格平均上涨 15 元。

5.3 非专业房东逐步回归建模

5.3.1 非专业房东 VIF 检验

表 5.4 非专业房东所有变量 VIF 值

变量	VIF	变量	VIF
未来 60 天可定天数	59.97	房东回复时长	2.05
未来 90 天可定天数	28.50	1km 内的地铁站数	1.88
最新评论时间	26.11	最近公交站距离	1.80
首次评论时间	22.80	每月评论数	1.67
未来 30 天可定天数	14.60	1km 内的旅馆酒店数目	1.59
是否是整租房	7.01	实词数	1.55
是否是单人间	6.65	房东注册天数	1.54
描述准确性评分	6.39	最近的景点距离	1.53
综合评分	6.29	房源是设施数	1.50
可容纳人数	5.67	是否超赞房东	1.49
卧室数	4.85	房东认证方式数量	1.37
入住顺利评分	4.57	1km 内的景点数	1.37
性价比评分	4.51	是否闪订房源	1.33
沟通交流评分	4.07	最近旅馆距离	1.32
干净整洁评分	3.63	未来 365 天可定天数	1.30
卧室数	3.29	房东接受率	1.29
最近的地铁站距离	3.21	主题熵	1.26
浴室数	2.72	1km 内的公交站数目	1.07
该区二手房均价（万元/㎡）	2.66	最长预定天数	1.06
总评论数	2.63	房东是否认证	1.05
位置便利评分	2.33	最短预定天数	1.03
房东回复率	2.21		

如表 5.4，对于非专业房东的所有变量计算 VIF 值，发现 60 天内可预订天数、90 天内可预订天数、上一次评论距今时间、首次评论距今时间的 VIF 均明显大于 10。发现非专业房东变量的膨胀系数分布和专业房东相似，在选择将 60 天、90 天内可预订天

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

数、首次评论距今时间进行剔除后，得到的方差膨胀系数如下表 5.5，目前 VIF 值均已经小于 10。

表 5.5 非专业房东剔除 VIF>10 的变量后的变量 VIF 值

变量	VIF	变量	VIF
是否是整租房	7.00	1km 内的旅馆酒店数目	1.59
是否是单人间	6.65	每月评论数	1.58
描述准确性评分	6.38	实词数	1.55
综合评分	6.28	最近的景点距离	1.53
可容纳人数	5.66	房东注册天数	1.53
卧室数	4.85	房源是设施数	1.49
入住顺利评分	4.57	最新评论时间	1.49
性价比评分	4.50	是否超赞房东	1.47
沟通交流评分	4.07	房东认证方式数量	1.37
干净整洁评分	3.63	1km 内的景点数	1.36
卧室数	3.28	是否闪订房源	1.33
最近的地铁站距离	3.20	最近旅馆距离	1.32
浴室数	2.72	房东接受率	1.28
该区二手房均价（万元/㎡）	2.65	主题熵	1.26
位置便利评分	2.32	未来 30 天可定天数	1.21
房东回复率	2.21	1km 内的公交站数目	1.07
房东回复时长	2.04	最长预定天数	1.05
总评论数	1.95	房东是否认证	1.05
1km 内的地铁站数	1.87	最短预定天数	1.03
最近公交站距离	1.80	1km 内的旅馆酒店数目	1.59

5.3.2 非专业房东逐步回归

表 5.6 非专业房东逐步回归结果

	非标准化系数		标准化系数	T	显著性
	B	标准误差			
截距项	-124.436	22.26		-5.59	0.00
最近的地铁站距离	1.002	0.00	0.21	10.91	0.00
浴室数	23.923	1.55	0.32	15.45	0.00

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.6 非专业房东逐步回归结果（续）

卧室数	-2.890	0.77	-0.08	-3.76	0.00
是否是整租房	87.774	7.81	0.39	11.23	0.00
未来 30 天可定天数	1.675	0.15	0.16	11.19	0.00
最近的景点距离	1.345	0.00	0.07	4.54	0.00
房东回复率	0.330	0.05	0.10	6.67	0.00
是否是单人间	48.656	7.84	0.21	6.21	0.00
可容纳人数	-5.208	0.73	-0.20	-7.16	0.00
房源设施数	1.356	0.20	0.11	6.72	0.00
1km 内的酒店旅馆数目	0.241	0.08	0.05	3.08	0.00
总评论数	0.566	0.11	0.10	5.22	0.00
该区二手房均价(万元/㎡)	3.741	0.88	0.09	4.26	0.00
最短预定天数	-0.238	0.05	-0.07	-4.98	0.00
最新评论时间	0.008	0.00	0.07	4.43	0.00
实词数	0.134	0.04	0.06	3.76	0.00
综合评分	0.716	0.19	0.05	3.78	0.00
是否超赞房东	17.484	4.70	0.06	3.72	0.00
房东接受率	0.220	0.06	0.05	3.63	0.00
房东注册天数	0.009	0.00	0.05	3.49	0.00
1km 内的地铁站数	4.109	1.75	0.04	2.35	0.02
每月评论数	2.945	1.35	0.04	2.18	0.03
1km 内的公交站数目	0.038	0.02	0.03	2.06	0.04

利用 SPSS 软件对于 4619 个非专业房东的真实房源进行逐步回归，在剔除大 VIF 值的变量后对剩余变量建立逐步回归模型，最后筛选出的模型的变量如上表 5.6：根据逐步回归得到的结果，可以计算得到 R 方达到了 58%，根据上表的回归系数，可以得到非专业房东的短租房源的价格模型为：

房源单人价格=-124.434+0.001*最近的地铁站的距离+23.923*浴室数-2.89*床位数+87.774*是否是整租房+1.675*未来 30 天可定天数+0.001*最近的景点的距离+0.33*房东回复率+48.656*是否是单人间-5.208*可容纳人数+1.356*房源设施数+0.241*1km 内的旅馆酒店数目+0.566*总评论数+3.741*该区二手房均价-0.238*最短预定天数+0.008*最新评论时间+0.134*实词数+0.716*综合评分+17.484*是否超赞房东+0.220*房东接受率

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

+0.009*房东注册天数+4.109*1km 内的地铁站数+2.945*每月评论数+0.038*1km 内的公交站数目

对非专业房东的房源进行逐步回归，最终保留了 23 个变量。其中包含房源属性变量床位数、浴室数、房源设施数、可容纳人数、是否是整租房、是否是单人间、最短预定天数、未来 30 天可定天数、实词数。除了床位数、可容纳人数对价格呈现负向影响外，其余均呈现正向影响，正向影响最大的是是否是整租房、是否是单人间、浴室数，这三者每增加一个单位，房源人均价格平均增加 87.8 元、48.7 元、24 元。

房东属性中保留的变量有：是否超赞房东、房东注册天数、房东回复率、房东接受率。是否超赞房东是平台结合多方面的规则来对房东进行的评价，通过逐步回归系数可得，当非专业房东房源的其他属性不变的情况下，超赞房东的房源人均价格平均高 17.5 元。此外，房源接受率对房源人均价格也呈正向影响，租户在咨询房东后历史成交的可能性越大，说明房东与租户的交流更愉快，从而对于整体房源人均价格也会产生积极作用。

在线评价中保留的变量是总评论数、每月评论数、最新评论时间、综合评分。该四个因素对于非专业房东房源人均价格均呈现出正向影响。

区位优势中保留的变量有：该区二手房平均价格（万元/㎡）、最近的地铁站距离、最近的景点距离、1km 内的旅馆酒店数、1km 内的地铁站数、1km 内的公交站数。其中 1km 内地铁站数对价格的影响权值最大，1km 内的地铁站数目每增加一个单位，价格平均增加 4 元。

5.4 逐步回归差异性分析

对于专业房东来说，逐步回归后保留的变量明显多于非专业房东，说明专业房东在定价时会考虑到更多的因素，相比来说更加的成熟和全面。

针对于房源属性变量来说，专业房东和非专业房东逐步回归模型中均显著的变量包含床位数、浴室数、房源设施数、可容纳人数、是否整租房、是否是单人间、实词数、最短预定天数、未来 30 天可定天数，且对于房源人均价格的影响方向相同。其中专业房东模型中特有的显著变量是主题熵，对于专业房东来说，它们拥有多套房源，他们会更加注重房源之间的差异性，通过让自己多个房源更具特征性，来吸引更多的租户，实现自己的利润的最大化。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

针对于房东属性变量来说，专业房东和非专业房东逐步回归模型中均显著的变量包含房东的注册天数、房东的认证方式数量、房东回复率。无论是对于专业房东和非专业房东来说，租户对于房东的信任度对最后是否预定房源起着非常重要的作用，进而影响到房东的自身定价。非专业房东模型中，房东的接受率也是一个显著性变量，可能是由于非专业房东由于房源数量较少，对于软件的关注量过少，即便是遇到租户预定的情形，也有可能因为错过回复时间而导致订单失败。

针对于在线评价变量来说，专业房东和非专业房东逐步回归模型中均显著的变量包含总评论数、每月评论数、最新评论时间、综合评分，并且对于房源人均价格的影响方向相同。说明房源的历史成交情况以及近期房源的热度，对于专业房东和非专业房东来说都是很重要的参考因素。专业房东逐步回归模型中独有的显著性变量是干净整洁评分、沟通交流评分、位置便利评分、性价比评分。对于专业房东来说，他们对于自己房源的运营更加的专业，他们会更加注重自己房源的整洁性、便利性，以及对于租户的礼貌性，通过租户给予的反馈，来对自己的房源价格进行调整。

针对于区域位置优势。专业房东和非专业房东的显著性变量十分相似，相比较而言，专业房东考虑的会更加全面，地铁站、公交站、旅馆或酒店、周边景点四个方面的区域位置都为显著变量。

5.5 异质性房源分位数回归模型

分位数回归相比普通的最小二乘回归，能够更加精准地描述自变量 X 对因变量 Y 的变化范围，以及条件分布形状的影响。因此，本文采用了分位数回归方法，深入探讨两类房东在不同价格分位数处的异同。

5.5.1 分位数回归模型

用 $\hat{y}_{(\tau)t}$ 表示分位数回归的估计量，即房屋的价格，τ 表示模型中分位数的取值，以检查函数 $(checkfunction)\omega_{(\tau)}$ 为 y_t 对于任意 α 的加权差绝对值和 $\sum \omega_t |y_t - \alpha|$，只有在 $\hat{y}_{(\tau)t}$ 取得最小值：

$$\sum \omega_t |y_t - \alpha| = -\sum_{i:y_{i<\partial}}^{T} (1-\tau)(y_t - \partial) + \sum_{i:y_{i\geq\partial}}^{T} \tau(y_t - \partial) \tag{5-2}$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对于一般的回归模型 $y_t = X'\beta + \mu_t$，求第 τ 分位数回归方程系数的估计量 $\hat{\beta}_{(\tau)}$，即求解下列目标函数：

$$Q = -\sum_{\hat{u}_{(\tau)t<\partial}}^{T}(1-\tau)\hat{u}_{(\tau)t} + \sum_{\hat{u}_{(\tau)t\geq\partial}}^{T}\tau\hat{u}_{(\tau)t} = -\sum_{y_t<X'\hat{\beta}_{(\tau)}}^{T}(1-\tau)(y_t - X'\hat{\beta}_{(\tau)}) + \sum_{y_t\geq X'\hat{\beta}_{(\tau)}}^{T}\tau(y_t - X'\hat{\beta}_{(\tau)}) \quad (5\text{-}3)$$

其中，X 和 β 为列向量，$\hat{u}_{(\tau)t}$ 表示第 τ 分位数点上的残差值，第 τ 分位数的回归方程为：

$$y_{(\tau)t} = X'\beta_{(\tau)} \quad (5\text{-}4)$$

当得到相应的方程后，可计算该方程的残差值 $\hat{u}_{(\tau)t}$：

$$\hat{u}_{(\tau)t} = y_t - \hat{y}_{(\tau)t} = y_t - X'\hat{\beta}_{(\tau)} \quad (5\text{-}5)$$

对于一个样本而言，估计的分位数回归方程越多，对于解释变量 y_t 条件分布的理解就越充分。

5.5.2 分位数回归差异分析

表 5.7 专业房分位数回归结果

	0.10	0.25	0.50	0.75	0.90
截距项	-20.956**	-14.673	-71.264***	-80.264***	8.913
是否闪订房源	7.372***	4.616 ***	5.417***	8.197***	13.163***
是否是整租房	78.534***	79.809***	95.728***	113.195***	110.649***
是否是单人间	40.857 ***	38.239***	56.920***	98.577***	130.40***
可容纳人数	-7.271 ***	-9.710***	-10.599***	-10.391***	-9.312***
床位数	-3.113 ***	-3.187***	-2.081***	-2.275***	-2.876**
卧室数	2.997 **	6.942***	6.585***	6.811***	7.262***
浴室数	10.966 ***	19.134***	31.425***	38.115***	36.149***
房源是设施数	0.488***	0.775***	1.233***	1.310***	1.541***
最短预定天数	-0.352***	-0.273***	-0.257***	-0.204***	-0.077***
未来 30 天可定天数	0.566***	0.689***	0.978***	1.325***	1.681***
实词数	0.066***	0.114***	0.178***	0.249***	0.308***
主题熵	-3.163*	-3.422*	-0.809	16.023***	9.123***
房东回复率	0.099***	0.154***	0.208***	0.254***	0.109***
房东回复时长	0.470***	-0.100	-0.794	-8.771***	-12.995***

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.7 专业房分位数回归结果（续）

房东注册天数	0.004***	0.007***	0.008***	0.002	0.002
房东认证方式数量	0.152	0.023	0.093	1.391***	2.568***
总评论数	0.180***	0.361***	0.425***	0.575***	0.503***
每月评论数	0.694	1.922**	1.818	5.244***	5.225**
最新评论时间	0.001*	0.004***	0.010***	0.017***	0.022***
综合评分	0.302 **	0.469***	0.620***	1.267***	1.713***
干净整洁评分	2.972**	3.425***	6.380***	9.333***	11.836***
沟通交流评分	2.502*	3.667***	5.721***	7.255***	10.572***
位置便利评分	2.659**	3.971***	6.740***	8.517***	7.590***
性价比评分	-5.575***	-8.386***	-11.219***	-20.932***	-20.584***
该区二手房均价	1.998***	1.779***	2.585***	5.000***	4.525***
1km 内的景点数	5.467***	11.303***	14.704***	16.856***	16.579***
1km 内的地铁站数	3.665***	3.554***	4.071***	8.011***	7.248***
1km 内的旅馆酒店数	0.185***	0.333***	0.525***	0.394***	0.354***
最近公交站距离	-1.229***	-3.208***	-3.159***	-5.157***	-2.373***
最近旅馆距离	3.38***	6.333***	12.112***	15.503***	17.437***
最近的景点距离	-1.23***	-1.368***	-1.276***	-0.994***	-0.739***
最近的地铁站距离	-0.567***	-0.830***	-1.380***	-2.121***	-2.173***

表 5.8 非专业房分位数回归结果

	0.1	0.25	0.5	0.75	0.9
截距项	-40.47***	-73.68***	-122.93***	-161.03***	-162.12***
床位数	-2.377***	-2.103***	-2.341***	-1.651	-2.265
浴室数	10.518***	21.354***	29.923***	40.200***	38.502***
房源是设施数	0.521***	0.940***	1.266***	1.140***	2.098***
可容纳人数	-5.706***	-6.003***	-5.974***	-7.094***	-7.649***
是否是整租房	49.289***	60.567***	77.355***	99.400***	118.247***
是否是单人间	16.569***	19.764***	38.401***	60.513***	86.610***
最短预定天数	-0.171***	-0.207***	-0.250***	-0.242***	-0.245**
未来 30 天可定天数	0.631***	1.058***	1.458***	2.131***	2.375***
实词数	0.125***	0.143***	0.142***	0.125**	0.042***
房东回复率	0.092***	0.153***	0.277***	0.509***	0.562***

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.8 非专业房分位数回归结果（续）

房东接受率	0.056	0.126***	0.210***	0.295***	0.442***
是否超赞房东	3.085	9.258***	11.714***	20.800***	35.327***
房东注册天数	0.001	0.007***	0.007***	0.015***	0.014***
总评论数	0.219***	0.390***	0.355***	0.526***	0.766**
每月评论数	-0.207	1.716**	3.762***	3.342	3.634
最新评论时间	-0.001***	0.003*	0.008***	0.015***	0.019***
综合评分	0.470***	0.386***	0.576***	0.666***	0.994***
该区二手房均价	1.500***	3.027***	4.204***	5.354***	5.392***
最近的地铁站距离	-0.75***	-0.97***	-1.349***	-1.405***	-1.412***
最近的景点距离	-0.23***	-1.32***	-1.379***	-1.208***	-1.042*
1km 内的旅馆酒店数	0.100*	0.179***	0.161**	0.241**	0.094***
1km 内的地铁数	5.302***	4.699***	3.973**	5.356**	5.243***
1km 内的公交站数	0.035**	0.049***	0.035**	0.059**	0.052***

针对于房源属性来说，专业房东和非专业房东在相同的显著变量中表现相同，变量浴室数、是否整租房、是否单人间均是随着房源价格分位数点的增加，对房源人均价格的正向影响不断增强。可容纳人数，随着分位数点的增加，对房源人均价格的负向影响逐渐增加，而床位数随着分位数点的增加，负向影响基本维持不变。对于专业房东来说，主题熵变量随着分位数点增加，对房源人均价格的影响从负向变为显著的正向，说明了主题的丰富性和独特性对于专业房东的高价位房源正向影响显著。

针对于房东属性来说。专业房东，低价位房源房东的认证方式数量对于房源人均价格影响不显著，随着分位数点的增加，在 75 分位数点之后呈现显著。针对于房东注册天数变量，专业房东低价位房源中，该变量的影响显著，而高价位房源中不显著，想必是因为，针对于高价位的专业房东房源，租客对其的信任度更高，此时更加看重的是房源本身的特点及优势。而非专业房东中，该变量随着分位数点的增加，正向影响程度越来越大。非专业房东中，是否超赞房东对于低价位房源影响不显著，是因为该类租客由于经济原因，更加看重的是房源的价格及性价比，对于房东的信息披露没有那么关注。随着分位数点的增加，是否超赞房东，对于非专业房东房源人均价格的正向影响越来越显著。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

针对于在线评价属性，专业房东和非专业房东在共同显著变量下，两者的表现相似，都是随着房源人均价格分位数点的增加，正向影响越来也显著。其次对于专业房东来说，干净整洁评分、沟通交流评分、位置便利评分对于房源人均价格的正向影响随着分位数点的增加最明显。

针对于区域位置属性，在相同显著影响的变量下，专业房东和非专业房东两者的表现相似。随着房源人均价格分位数点的增加，1km 内的景点数、地铁站数、旅馆酒店数对房源人均价格的正向影响均会增大。对于最近的公交站距离、最近的景点距离、最近的地铁站距离，均随着房源人均价格分位数点的增加，负向影响呈现出波动上升的趋势。专业房东最近的旅馆距离则呈现的是正向影响，是因为短租房源和旅馆酒店两者是强竞争的关系，所以最近的旅馆酒店越远，短租房源的价格越有优势，且随着分位数点的增加，正向影响显著增大。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 6 章 短租房源定价模型构建

在第 5 章节中，本文从统计学线性模型的角度解释了短租房源定价的四个方面的影响因素，包括房源属性、房东属性、在线评价、区位优势各自对于房源价格的影响程度及趋势，并解释了各自在现实中所代表的经济含义。但目前也有很多学者发现，线性回归不能很好的满足于复杂模型，因为线性回归有着比较严格的假定，而在真实数据中，往往不能满足所有的假定条件。

机器学习的对象是数据，它从数据出发，通过提取数据的特征、在数据中进行训练，来抽象出数据模型，发现其中的规则和规律，再回到对新的一批数据的分析和预测中去。它对于数据的前提条件要求较低，通过不断地调整模型的参数，来进一步地拟合数据。目前机器学习算法已经在例如垃圾邮件检测、信用风险违约预测、商品预测等多个现实场景进行了应用。近几十年以来，由于集成学习能够高效地解决实际应用问题，所以在机器学习领域内备受关注，在数据挖掘、模型识别、文本分类、预测等方面有着广泛的应用，在各大竞赛平台，如 Kaggle、天池等大数据平台也广受欢迎，它通过对多个基模型的结果进行汇总来获得更高的预测精度。所以该章节将重点从机器学习中非线性回归的角度来搭建短租异质性房源的定价模型。

6.1 模型的构建与检验参数

模型构建与选择的操作框架如下图 6.1 所示，主要分为变量识别与模型选择、算法实现与模型结果两大块。第一，原始数据经过数据清洗与分类后，依据“房东拥有房源数”将房源数据分为专业房东的房源数据和非专业房东的房源数据两部分。第二，分别对两类房东的房源数据进行 Lasso 变量识别出差异化重要变量，同时利用第四章节通过逐步回归所选的变量来构建不同的机器学习非线性回归模型，一方面比较不同变量选择方法对于数据的拟合情况，另一方面也比较不同的预测模型的拟合精度。第三，对数据集按 4:1 的随机抽样比分为训练集与测试集，在模型结果输出后纵向对比两组专业房东定价预测模型误差，选择一个最优模型作为专业房东的房源最优定价模型，同理对非专业房东的房源进行相同操作，选择其中一个最优模型。第四，本文将对最终确定的两个模型（专业房东房源定价模型、非专业房东房源定价模型）的输出结果进行具体的分析与解释，给出两类房东的差异化定价规律。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

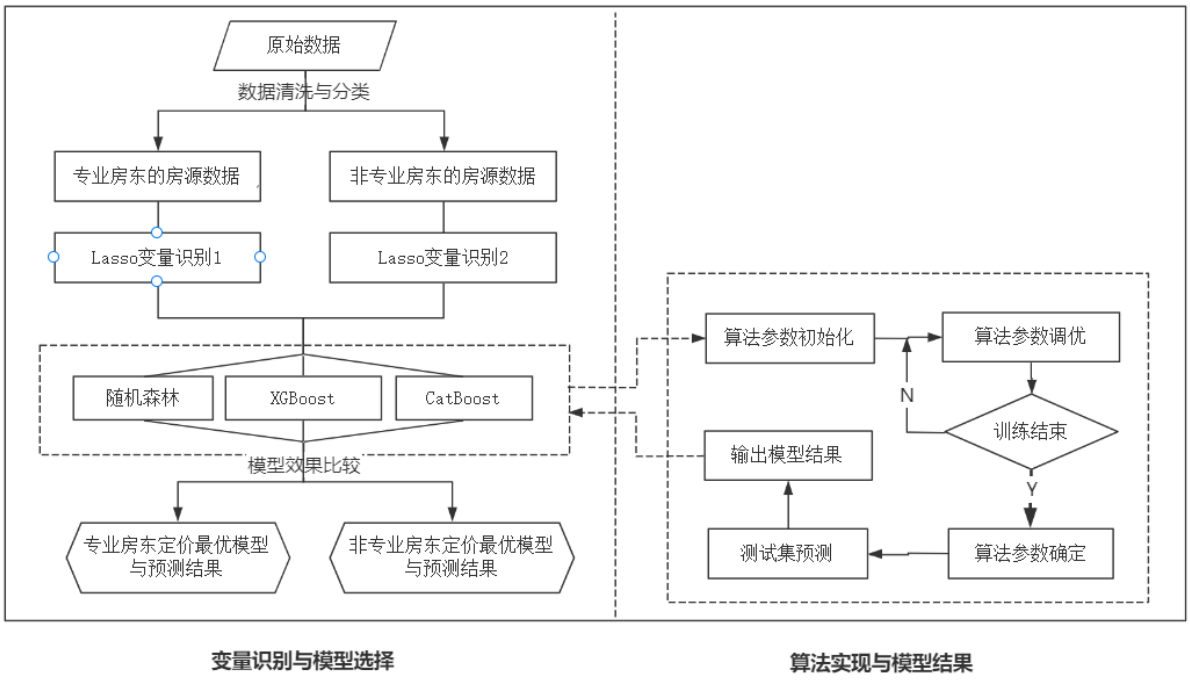


图 6.1 模型构建与选择操作流程图

在模型选择环节，为了考量短租房源定价模型预测效果，将三种模型与传统应用广泛的逐步回归线性模型、对数回归进行对比，对于回归预测结果，本文以平均绝对误差（MAE）、均方根误差（RMSE）和平均绝对百分比误差（MAPE）作为预测模型的评价函数，数值越小，表明预测的精度越高，其误差计算公式为：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y^*_i)^2} \quad (6\text{-}1)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-y^*_i\right| \quad (6\text{-}2)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y^*_i}{y_i}\right| \quad (6\text{-}3)$$

其中 y_i 为第 i 个房源的真实定价，y^*_i 为第 i 个房源的预测值，n 为预测样本总数。输出三种模型对专业房东和非专业房东房源的定价预测结果与误差，依据模型比较结果选择对应两类房东的最优模型。

PDP 图（*Partial Dependence Plots*）可以显示的展示机器学习黑盒模型中的自变量对因变量的边际效应，展现一个自变量是如何影响因变量的。部分依赖图可以显示目标与特征之间的关系是线性的，单调的还是更复杂的。其中 *PDP* 依赖函数定义

$$f_{x_s}(x_s)=E_{x_c}[\hat{f}(x_s,x_c)=\int\hat{f}(x_s,x_c)dR(x_c) \quad (6\text{-}4)$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中 X_S 描述的时部分相关函数需要绘制的特征，X_C 是模型中使用的其他特征。通常，S 集中只有一两个特征，也就是我们想知道的对预测的影响的特征。S 和 C 特征向量的综合构成了特征空间集 X。比分依赖通过对模型输出分布特性集边缘化，使我们得到一个只依赖于 S 中的特性并且与其他特性交互的函数：

$$\hat{f}_{x_S}(x_S)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S,x_C^{(i)}) \tag{6-5}$$

6.2 Lasso 变量识别

Lasso 变量识别从本质上讲是一种基于回归模型的数据降维方法，该方法不仅适用于线性情况，也适用于非线性情况。在求解回归模型时加入 L1 范式，即使用绝对数函数为惩罚项来约束模型的回归系数，使得回归系数在特定的约束条件下最小化残差平方和。数值试验中，通常使用 Lasso 回归从众多影响因素中选择出主要影响因素，这不仅保证了较高的预测精度，而且通过 Lasso 方法的因素选择整体上降低了模型训练过程中的计算量，简化了模型的复杂度，在高维度数据中有大量应用。Lasso 是基于惩罚方法对样本数据进行变量选择，通过对原本的系数进行压缩，假设只有有限个 X 的回归系数不为 0，但其余的都是 0。将原本很小的系数直接压缩至 0，从而将这部分系数所对应的变量视为非显著性变量，将不显著的变量直接舍弃，得到稀疏估计。对于线性归回模型 a 为常数项，$b_1,b_2,...,b_n$ 为线性回归系数，ε 为误差项，模型形式如下：

$$y=a+b_1x_1+b_2x_2+...+b_px_p+\varepsilon \tag{6-6}$$

设 $(x_{i1},x_{i2},...,x_{ip},y_i),i=1,2,...,n$ 是 n 个样本，记 $b=(b_1,b_2,...,b_p)^T$，则 Lasso 估计应满足：

$$(\hat{a},\hat{b})=\arg\min\{\sum_{i=1}^{n}[(y_i-a-\sum_{j=1}^{p}b_jx_{ij})^2+\lambda\|b\|_1]\} \tag{6-7}$$

这里 $\|b\|_1=\sum_{j=1}^{p}|b_j|,\lambda\geq 0$ 调和参数。相比较二次规划求解法，最小角回归算法能够更加高效地解出最优解。$(\hat{a},\hat{b})$ 的估计值与调和参数 λ 有关，当 λ 非零时，$\hat{b}$ 中的部分值会压缩为零，其对应的影响因素不会选入模型。影响因素的回归系数与压缩为零的个数的 λ 取值成正比关系。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

6.3 集成学习方法介绍

6.3.1 随机森林

随机森林是 2001 年由 Leo Breiman 和 Adele Cutler 基于 Bagging 集成学习的思想提出的，它由多个弱分类器决策树集合而成，这些弱分类器是相互独立的，最终的分类是由多数投票产生。其主要工作原理就是基于自助法(*bootstrap*)对原始训练数据集进行重采样，在原始训练数据集 N 中进行重复随机的有放回的抽取，对抽取的 K 个训练数据集分别建立决策树来组成随机森林，最后根据投票的原则对新的训练数据集进行分类，实现过程如下图：

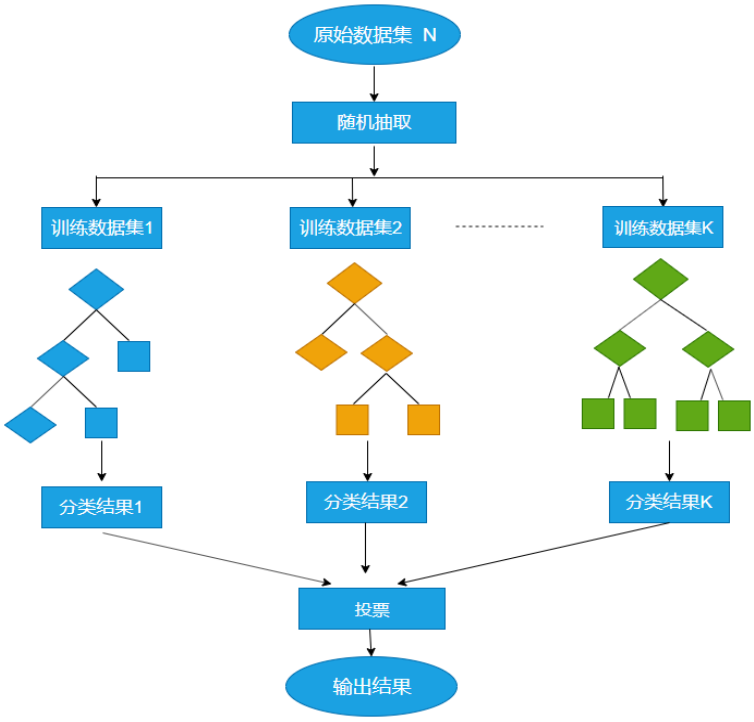


图 6.2 随机森林模型训练流程图

由图 6.2 可知，随机森林模型的构建可总结为以下几个步骤：

（1）确定抽样训练集。基于 bootstrap 方法在训练集中进行样本抽样，每一个样本就代表一颗决策树。因此合成抽取的 K 个样本即构建了包含 K 个决策树的随机森林。

（2）训练基决策树。若原始变量识别出的特征总数为 M 个，在其中随意抽取 $m(m \ll M)$个特征，依据决策树节点分布不纯度规则进行不剪枝训练，训练目的为充分生长每一棵树且使最小化每个枝节点的不纯度。

（3）集合随机森林。依据训练完成的多个基决策树对测试集样本分类，由 m 个基决策树投票产生最终的判断结果。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

6.3.2 GBDT

GBDT 由决策树算法和提升算法构建而成，是由 Friedma 提出的，它的主要思想就是以决策树作为基学习器，采用加法模型与前向分布算法的机器学习提升算法，他的特点是泛化能力较强。*GBDT* 相较于随机森林算法的优势在于，其模型训练时向减少残差的梯度方向生成新的决策树，通过一步步迭代最终实现模型精度提升的效果。因此，*GBDT* 较为广泛地应用在线性或非线性的分类预测中，同时也可用于重要变量提取的数据降维。

由于其拟合新树的方向选择特性，*GBDT* 的迭代过程主要由决策树和梯度提升两部分组成。决策树：首先基于初始数据获得第一颗决策树，利用测试集样本运行决策树时产生预测值，同时由于决策树迭代不够深入也将生成较大的预测残差。梯度提升：接着生成的决策树会经过不断学习，靠近残差降低方向再生决策树优化模型，直到残差达到指定的较小阈值，即可停止学习。最终将所有学习到的各决策树中用上的变量加权相加，就可以得到每个变量在该预测模型中的重要性，实现对变量的特征识别。在预测模型中每增加一个变量会提高一定的模型精度，但是当变量较多、能够体现数据较多信息时模型精度提升的速率逐级递减或直接导致模型精度下降。因此，当新增变量提升的模型精度比率低于某个阈值时，我们可以认为该变量在预测模型中并不重要，可以选择剔除。

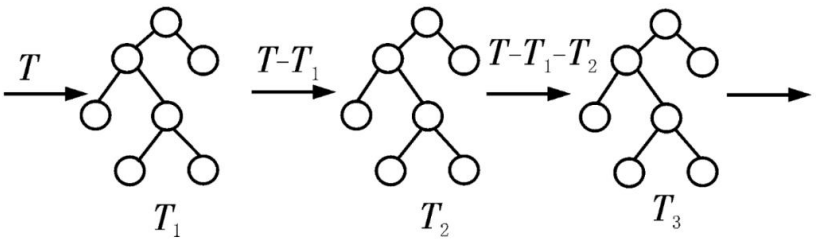


图 6.3 GBDT 模型训练流程图

上图是 *GBDT* 模型训练过程，第一颗决策树训练结果的残差项 $T-T_1$ 就是第二颗树修正模型的方向，以此类推指到模型残差值达到预期阈值，停止训练。最终模型的结果就是将每一棵树的预测结果累加，即 $\hat{T}=\sum_{i=1}^{N}T_i$。关于 *GBDT* 具体算法步骤可总结为：

（1）损失函数构建

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

将 X 和 Y 分别输入模型中，利用前向分布算法执行 m 次迭代，得到将要进行优化的损失函数为：

$$L(y, f(x_i)) = (y - f(x))^2 \tag{6-8}$$

（2）初始弱化分类器，在决策树只有一个根节点时估计的损失函数最小化的常数值：

$$f_o(x) = \arg\min \sum_{i=1}^{N} L(y_i, c) \tag{6-9}$$

（3）迭代 M 轮，产生决策树个数

*a.*对 $i = 1, 2, \cdots, N$（N 个样本）计算当前模型中损失函数的残差估计，用负梯度值衡量

$$r_{m_i} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{6-10}$$

*b.*生成针对优化 r_{m_i} 的新决策树，得到第 m 颗树叶节点区域 R_{m_j}, $j = 1, 2, \cdots J$

*c.*对 $j = 1, 2, \cdots J$ 计算 $c_{m_j} = \arg\min \sum_{x_i \in R_{m_j}} L(y_i, f_{m-1}(x_i) + c)$ 最小化损失函数。

*d.*更新决策树，$I(x)$ 为 *Indicator* 函数，判断 x 对应的叶节点，并累加上该叶节点的 c_{m_j}。

$$f_{m(x)} = f_{m-1}(x) + \sum_{j=1}^{J} c_{m_j} I(x \in R_{m_j}) \tag{6-11}$$

（4）输出最终 *GBDT* 模型式 4.5 所示：

$$\hat{f}(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{m_j} I(x \in R_{m_j}) \tag{6-12}$$

6.3.3 XGBoost

XGBoost 陈天奇在 2014 年初提出的，是在 Gradient Boosting 框架下实现机器学习算法，提供并行树提升，可以快速准确地解决许多数据科学问题。本文使用的 XGBoost 模型可以表示为：

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{6-13}$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中 K 表示树的总数，f_k表示第 k 棵树，$\hat{y}_i$表示样本x_i的预测结果，F 表示所有可能的回归树，f 表示一棵具体的回归树。目标函数包含两个部分，损失函数$l(\hat{y}_i, y_i)$和正则项$\Omega_{(f)}$。

$$Obj(\Theta) = \sum_{i}^{n} l(y_i + \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6-14}$$

采用如下泰勒二阶展开来定义一个近似的目标函数，进一步计算得：

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + cons\tan t \\ f(x+\Delta x) &\approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \\ g_i &= \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{t-1}), h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{t-1}) \\ obj^{(t)} &\approx \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + cons\tan t \end{aligned} \tag{6-15}$$

该目标函数仅依赖于每个数据点在损失函数上的一阶导数和二阶导数。对于一棵树的复杂度的定义如下。这个复杂度包含了一棵树里面节点的个数，以及每棵树叶子节点上面输出分数的 $L2$ 模平方。

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6-16}$$

在这种新的定义下，可以将目标函数进行如下改写，其中 I 被定义为每个叶子上面样本集合 $I_j = \{i \mid q(x_i) = j\}$

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ &= \sum_{i=1}^{n} [g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \end{aligned} \tag{6-17}$$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

通过不断枚举不同树的结构，利用打分结构来寻找一个最优结构树，加到模型中，重复这样的操作。对于具体的分割方案，可以通过下式计算：

$$Gain=\frac{1}{2}[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_R+H_L+\lambda}]-\gamma \tag{6-18}$$

6.3.4 Catboost

CatBoost（Categorical Boosting）算法是基于梯度提升树框架的一种新型实现算法，由 Yandex 公司于 2017 年提出，同目前 Boosting 族的两大主流算法 XGBoost、LightGBM 算法一样是基于 GBDT 的开源机器学习算法。CatBoost 算法主要提出了两种关键的方法，一个是用于处理分类特征的算法，另一个是排序提升算法—Ordered boosting。

（1）分类特征的算法（Categorical features）是对分类特征进行处理，常用的方法有标签编码、独热编码等，而 CatBoost 算法可以直接使用分类特征进行建模，它使用有关分类特征及分类和数字特征组合的各种统计数据,将分类特征值转换为数字,不需要任何明确的预处理就可以将类别转换为数字。记$\sigma =$（σ_1，$\sigma_2 \cdots \sigma_n$）为一个排列，则$x_{\sigma_p,k}$可以替代为：

$$\frac{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]Y_{\sigma_j}+a\cdot P}{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]Y_{\sigma_j}} \tag{6-19}$$

其中 P 为先验概率，a 为先验概率的权重（$a>0$），添加先验概率 P 有助于减小从低频类别获得的噪声。在本文中即使减少房价异常值对预测模型参数设定带来的影响。

（2）排序提升算法（Ordered boosting）相比传统的 GBDT，可以降低模型过拟合的概率。GBDT 在每次迭代时针对同一个数据集计算损失函数对目前模型的梯度，然后基于此梯度进行训练得到基学习器，但是这种方法会使逐点梯度产生估计偏差。Ordered boosting 方法改变传统算法中的梯度的无偏估计，降低估计偏差的影响，提高模型泛化能力，Ordered boosting 算法的流程如下所示：

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 6.1 Ordered boosting 算法流程

Ordered boosting
Input: *Training* **Set** $(X_k, Y_k)_{k-1}^n$, *Number of Iteration I*
1: $M_i \leftarrow 0$ for $i = 1...n$
2: *for iter* $\leftarrow 1$ to n do
3: *for* $i \leftarrow 1$ to n do
4: *for* $j \leftarrow 1$ to $i-1$ do
5: $g_j \leftarrow \frac{d}{da} Loss(y_i, a)\Big\|_{a-M_i(X_j)}$
6: *end for*
7: $M \leftarrow$ Learn$Model(X_i, g_j)$ for $j = 1 \cdots i-1$
8: $M \leftarrow M_i + M$
9: *end for*
10: *end for*
11: *return* $M_1 M_n; M_1(X_1), M_2(X_2) \cdots M_n(X_n)$

由上表 6.1 流程可知，CatBoost 算法是通过对每个样本 x_i 训练一个单独的模型 M_i，悬链模型 M_i 的数据 不包含 x_i 的训练集，然后使用模型 M_i 对样本的梯度进行估计，最后使用此梯度训练基学习器学得最终模型。

6.4 定价模型构建

6.4.1 专业房东 lasso 变量选择

由于逐步回归方法的特性，仅会保留显著性的特征，这样会导致很多补充性的变量或者是辅助变量的误删，但其实往往很多辅助性的变量也起着较为关键的作用。并且多元回归中当变量过多时，会因为变量维度过多，回归的有效性受到考验。贺平等人基于 41 个特征变量对上市公司的股票价格预测建立了 Lasso-logistic 组合算法模型，发现该模型相比于回归模型以及单一机器学习模型，表现得更好且更稳定[33]。所以针对于此，本文中也使用往模型中加正则化的方法，通过比较来选择最适合房东定价的变量选择模型。本文使用基于了 L1 范式的 Lasso 回归模型得到各个定价模型的重要影响因素，从而筛选出系数不为 0 所对应的影响因素参与最终的专业房东房价预测。将影响专业房东的 44 个个变量输入 Lasso 回归，识别出的重要变量及参数估计值汇总在表 6.2，剔除了房东是否认证、是否超赞房东、未来 30 天、60 天、90 天、365 天可定

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

天数、房东注册天数、首次评论距今天数、描述准确性得分、入住顺利得分、沟通交流评分这十一个变量。

表 6.2 专业房东定价模型的 Lasso 回归参数估计

变量	系数	变量	系数
房东注册天数	0.006	描述准确性评分	0.000
房东回复时长	-1.309	干净整洁评分	3.373
房东回复率	0.130	入住顺利评分	0.000
房东接受率	0.077	沟通交流评分	0.000
房东认证方式数量	0.669	位置便利评分	1.538
房东是否认证	0.000	性价比评分	-9.449
是否超赞房东	0.000	是否闪顶房源	3.755
房东总房源数	0.206	1km 内的地铁站数	2.780
该区二手房均价（万元/㎡）	1.707	1km 内的景点数	12.705
可容纳人数	-5.978	1km 内的旅馆酒店数目	0.400
浴室数	20.765	最近的地铁站距离	0.001
卧室数	3.906	最近的景点距离	-0.001
床位数	-4.134	最近旅馆距离	0.013
房源是设施数	1.355	1km 内的公交站数目	-0.002
最短预定天数	-0.183	最近公交站距离	-0.002
最长预定天数	-0.004	实词数	0.249
未来 30 天可定天数	1.010	主题熵	0.321
总评论数	-0.508	是否是整租房	25.315
每月评论数	2.526	未来 90 天可定天数	0.000
最新评论时间	0.012	未来 365 天可定天数	0.000
未来 60 天可定天数	0.000	首次评论距今天数	0.000
综合评分	0.760		

6.4.2 专业房东定价模型结果

为了更好地比较模型的预测精度，将专业房东 Lasso–随机森林、Lasso–XGBoost 模型、Lasso–CatBoost 模型与传统 Lasso-线性模型做比较研究。在经过各种模型的精细调参后，得到表 6.3 中各定价模型的误差结果。三个非线性回归定价模型预测精度相比 Lasso-线性模型而言，都有不同程度的提高，其中 Lasso–XGBoost 模型和 Lasso-CatBoost 模型的预测房价与真实房价的均方根误差、均方误差、平均绝对百分比误差

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

都相对较小。从模型本身来说，线性回归模型在对大样本进行预测时，无法避免受到样本中极大值和极小值的影响，短租房源因为不同房源类型的不同，别墅和普通合租房的价格有明显的差距，所以真实值和预测值的拟合效果明显不佳。而 Lasso-XGBoost 模型和 Lasso-CatBoost 模型在处理大样本数据时，从树模型的角度出发进行精确的预测，具有可靠性高、适用性强的特点。考虑到 Lasso-XGBoost 模型的均方根误差（RMSE）值为 41.23，低于 Lasso-CatBoost 模型的 42.02，且其他三个模型检验指标均最低，所以选择 Lasso–XGBoost 模型作为专业房东房源定价的最优模型。

表 6.3 专业房东定价模型结果表

定价模型	均方误差	平均绝对误差	平均绝对百分比误差	拟合优度 R 方
逐步回归模型	73.79	58.45	34.3	62.04%
逐步回归+随机森林	44.37	32.77	19.65	82.73%
逐步回归变量选择+XGBoost	42.75	30.45	18.12	83.25%
逐步回归变量选择+Catboost	43.34	31.27	19.54	83.13%
Lasso 变量选择+线性模型	72.89	56.46	34.1	64.21%
Lasso 变量选择+随机森林	43.30	30.87	19.17	84.34%
Lasso 变量选择+XGBoost	**41.23**	**30.11**	**17.12**	**84.95%**
Lasso 变量选择+CatBoost	42.02	30.54	18.07	84.45%

6.4.3 非专业房东 lasso 变量选择

基于非专业房东的房源数据进行 Lasso 回归来进行变量选择。Lasso 回归的对应重要变量及参数估计值汇总在表 6.4 中，剔除了未来 30 天、60 天、90 天、365 可定天数、描述准确性评分、入住顺利评分这六个变量，相比于专业房东来说，保留的变量明显更多，主要变量如下：

表 6.4 非专业房东定价模型的 Lasso 回归参数估计

变量	系数	变量	系数
房东注册天数	0.007	描述准确性评分	0.000
房东回复时长	-3.608	干净整洁评分	5.830
房东回复率	0.163	入住顺利评分	0.000
房东接受率	0.067	沟通交流评分	-2.591
房东认证方式数量	0.716	位置便利评分	4.212
房东是否认证	0.003	性价比评分	-12.916

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 6.4 非专业房东定价模型的 Lasso 回归参数估计（续）

是否超赞房东	0.004	是否闪顶房源	6.434
房东总房源数	0.209	1km 内的地铁站数	2.973
该区二手房均价（万元/㎡）	1.795	1km 内的景点数	14.181
可容纳人数	-6.536	1km 内的旅馆酒店数目	0.385
浴室数	21.935	最近的地铁站距离	0.001
卧室数	4.654	最近的景点距离	-0.001
床位数	-4.302	最近旅馆距离	0.013
房源是设施数	1.354	1km 内的公交站数目	-0.002
最短预定天数	-0.187	最近公交站距离	-0.002
最长预定天数	-0.004	实词数	0.243
未来 30 天可定天数	1.019	主题熵	1.705
总评论数	-0.536	是否是整租房	27.793
每月评论数	3.542	未来 90 天可定天数	0.000
最新评论时间	0.013	未来 365 天可定天数	0.000
综合评分	0.886	首次评论的时间	0.000
未来 60 天可定天数	0.000		

6.4.4 非专业房东定价模型结果

为了更好地比较模型的预测精度，将用两种不同的变量选择方法得到的非专业房东房源的重要性变量，分别应用到线性回归、随机森林、XGBoost、CatBoost 四种模型中，通过模型结果的比较来选择最适合非专业房东的定价模型。在经过模型调参后得到表 6.5 中的模型结果，可以发现非专业房东的三种非线性回归定价模型的预测精度相比线性模型由显著提高，Lasso–XGBoost 模型和 Lasso-CatBoost 模型的预测房价与真实房价的均方根误差、均方误差、平均绝对百分比误差同样相对较小。与专业房东不同的是，Lasso–CatBoost 模型的预测结果与真实房价的误差最小，均方根误差（RMSE）值为 47.62，且其余模型检验参数均最小。因此，最终选择 Lasso–CatBoost 模型作为非专业房东房源定价的最优模型。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 6.5 非专业房东各定价模型的误差结果

定价模型	均方误差	平均绝对误差	平均绝对百分比误差	拟合优度 R 方
逐步回归模型	80.46	67.91	49.31	53.12%
逐步回归+随机森林	52.41	41.39	30.54	76.21%
逐步回归变量选择+XGBoost	51.23	40.21	29.34	77.73%
逐步回归变量选择+Catboost	50.73	39.51	28.13	78.83%
Lasso 变量选择+线性模型	76.50	64.84	45.95	57.34%
Lasso 变量选择+随机森林	49.34	37.21	26.23	78.89%
Lasso 变量选择+XGBoost	48.85	36.61	25.11	79.21%
Lasso 变量选择+CatBoost	**47.62**	**35.51**	**24.65**	**80.73%**

6.5 专业房东定价模型评估

在经过各种模型的精细调参后得到表 6.3 中各定价模型的误差结果，通过对比决定结果可知 Lasso-XGBoost 模型的拟合效果最佳，因此专业房东选用 Lasso-XGBoost 模型对房源进行定价预测。对 Xgboost 的重要参数进行网格搜索调参后得到最佳的参数结果如表 6.6：

表 6.6 XGBoost 模型最终参数取值

参数	参数解释	最优取值
N_estimators	迭代次数，即决策树的数量	778
Max_depth	树的最大深度	9
learning_rate	收缩率，防止过拟合	0.03
subsample	训练模型的子样本占样本的比例	0.6
colsample_bytree	用来控制每颗树随机采样的列数的占比	0.7
Min_child_weight	决定最小叶子节点样本权重和	3
Gamma	节点分裂所需的最小损失函数下降值	0.04
reg_alpha	L1 正则化项的权重系数	0.1
reg_lambda	L2 正则化项的权重系数	0.6

在选定最佳参数后进一步选择最优迭代次数。由图 6-4 可知，当 N_estimators（迭代次数）达到 778 时，迭代次数的增加并没有使测试集上的误差减小，说明在考虑计算成本的情况下，最佳迭代次数 778。在选定重要参数下，拟合效果已经达到较好的水平，

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由图 6.5 得专业房东的 Lasso-XGBoost 模型测试集的预测值和真实值比较接近，拟合绝对误差值集中的 0 附近，基本稳定在 40 以内。

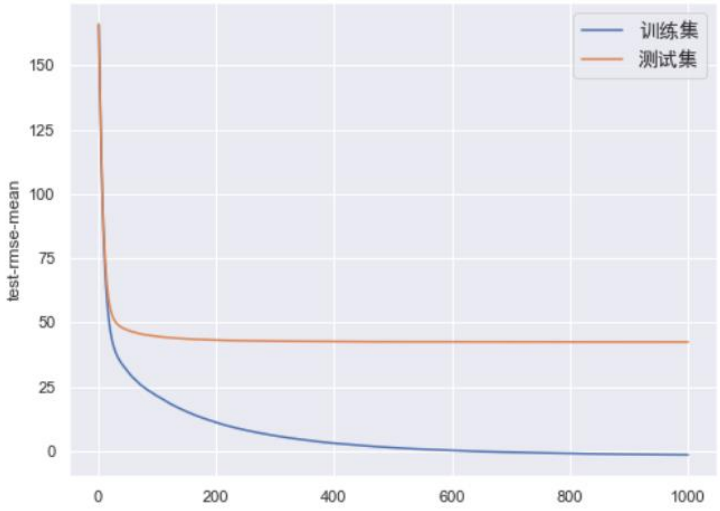


图 6.4 Lasso-XGBoost 模型迭代图

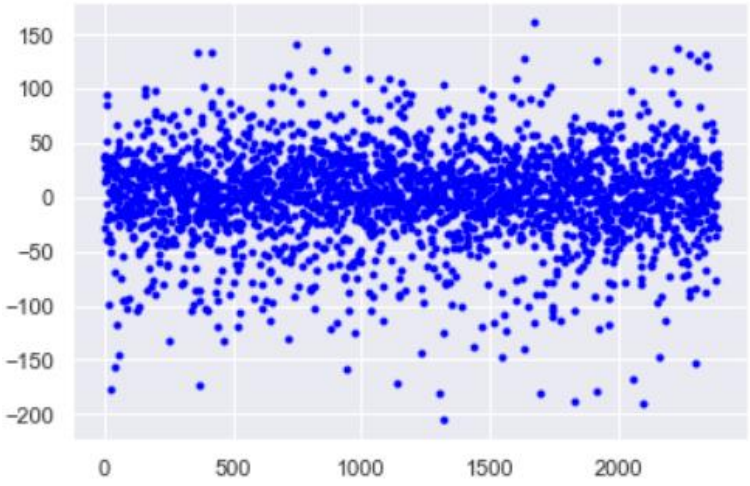


图 6.5 Lasso-XGBoost 模型测试集误差

XGBoost 模型作为预测模型算法较为复杂，只能得到最终的预测结果，而其中自变量与因变量之间的关系无从得知。因此，本文利用 Python 中 important 函数获取影响专业房东定价模型的前十个重要变量，从而得出各变量对预测结果的贡献大小。从图 6.6 细看按重要度从高到低分别为可容纳人数、是否是整租房源、卧室数、1km 内的景点数、是否是单人间、主题熵、1km 内的地铁站数、该区二手房均价、最近的景点距离、房东回复率。重要变量主要是来源于房源属性和区位优势。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

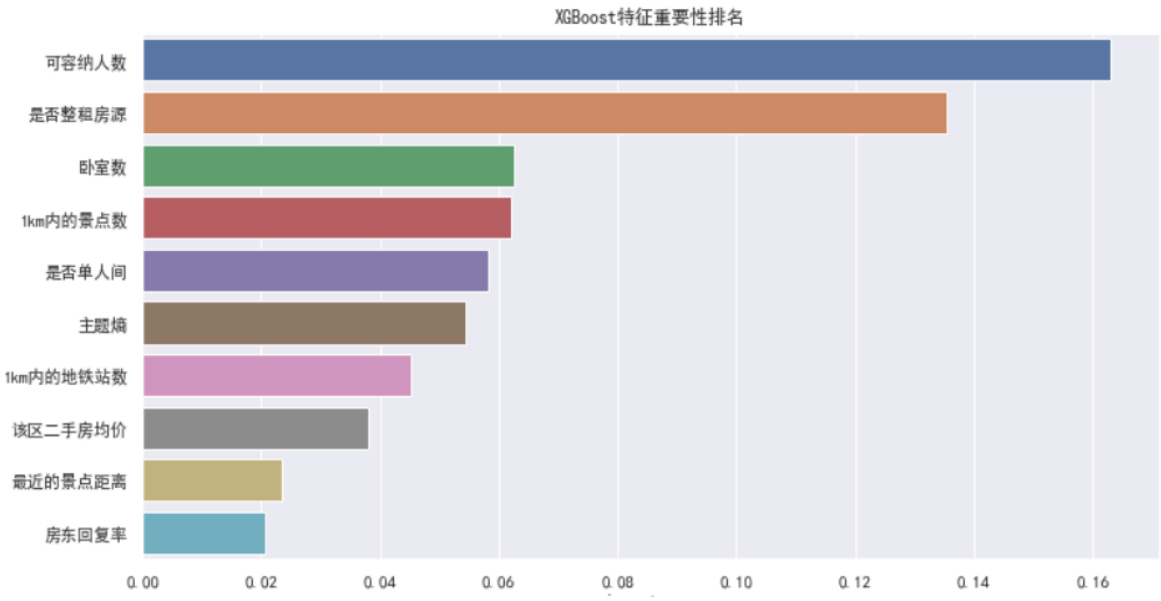


图 6.6 专业房东定价模型的前十个重要变量

由于机器学习预测模型都是黑盒模型，所以本文利用部分依赖图（PDP 图）来反映不同变量对于预测因变量的影响情况。以下四图分别是可容纳人数、最近景点距离、最近地铁站距离、1km 内的地铁站数的部分依赖图，可以发现其中可容纳人数、最近景点距离与房源人均价格呈现负向关系，最近地铁站距离、1km 内的地铁站数与房源人均价格呈现正向关系，这与上述研究中逐步回归的结果一致，并且符合经济意义。

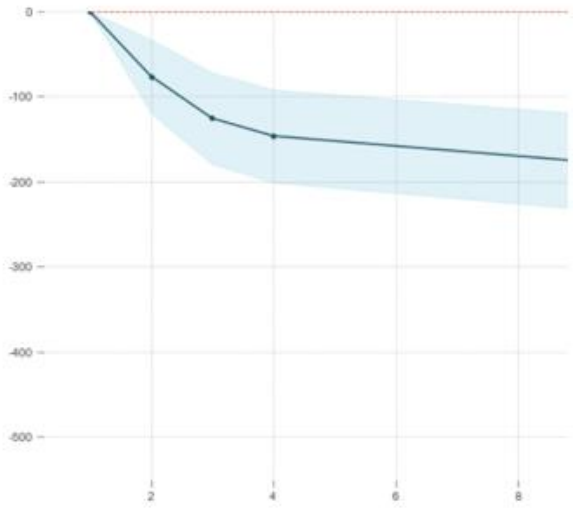


图 6.7 专业房东可容纳人数 PDP 图

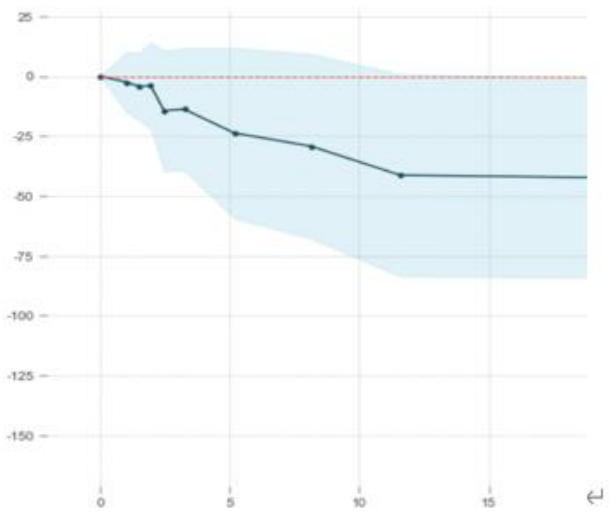


图 6.8 专业房东最近景点距离 PDP 图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

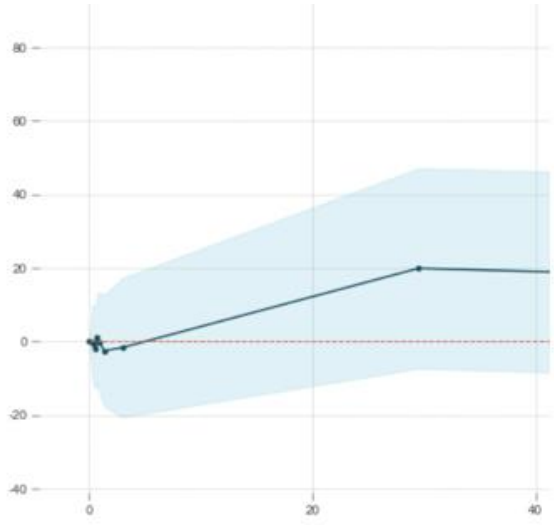


图 6.9 专业房东最近地铁站距离 PDP 图

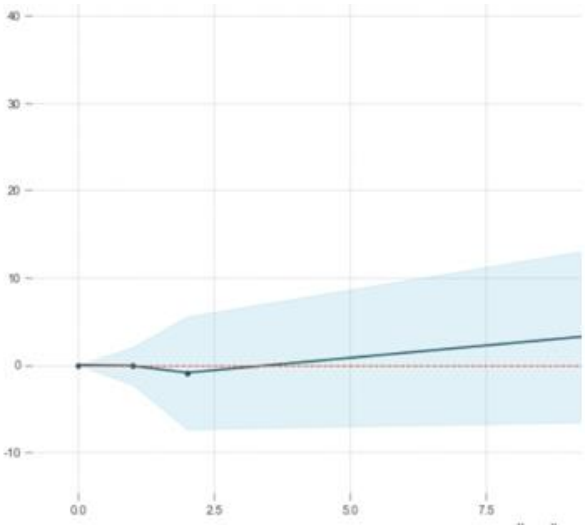


图 6.10 专业房东 1km 内的地铁站数 PDP 图

6.6 非专业房东定价模型评估

在对非专业房东的各种模型进行精细调参后得到表 6.5 中各定价模型的误差结果，通过对比决定结果可知 Lasso-CatBoost 模型更适合用来对非专业房东的房源进行定价预测。对 catboost 的重要参数进行网格搜索及交叉验证调参后得到最佳模型的参数取值如下表 6.7：

表 6.7 CatBoost 模型最终参数取值

参数	参数解释	最优取值
iterations	可以建立的最大树数。当使用其他限制迭代次数的参数时，树的最终数量可能少于此参数中指定的数量。	831
learning_rate	收缩率，防止过拟合	0.02
Depth	树的最大深度	12
l2_leaf_reg	代价函数的 L2 正则化项的系数。	3
Bagging temperature	定义贝叶斯 bootstrap 的设置，对象分配随机权重	0.2

在选定最优参数后进一步选择最优迭代次数，由图 6.11 可知，当 iterations（迭代次数）达到 831 时，迭代次数的增加并没有减小测试集上的误差，说明在考虑计算成本的情况下，最佳迭代次数 831。在选定重要参数下，拟合效果已经达到较好的水平，由图 6.12 得非专业房东的 Lasso-CatBoost 模型测试集的预测值和真实值比较接近，拟合绝对误差值集中的 0 附近，基本稳定在 50 以内。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

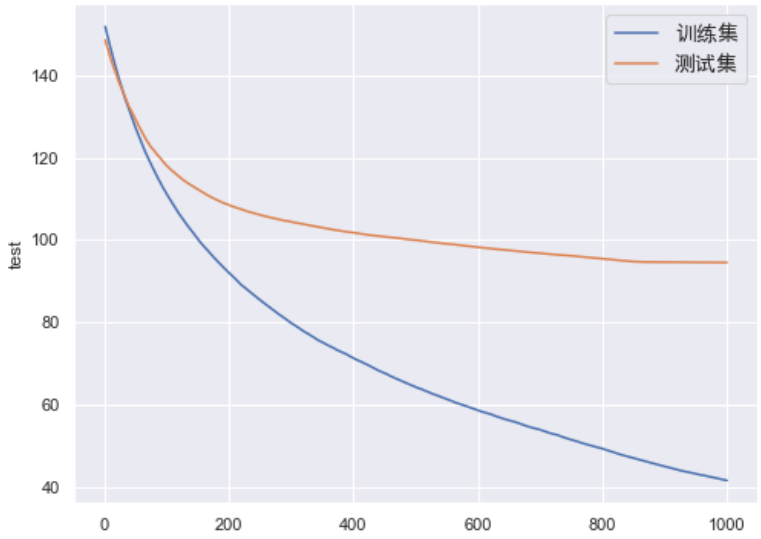


图 6.11 Lasso-CatBoost 模型迭代图

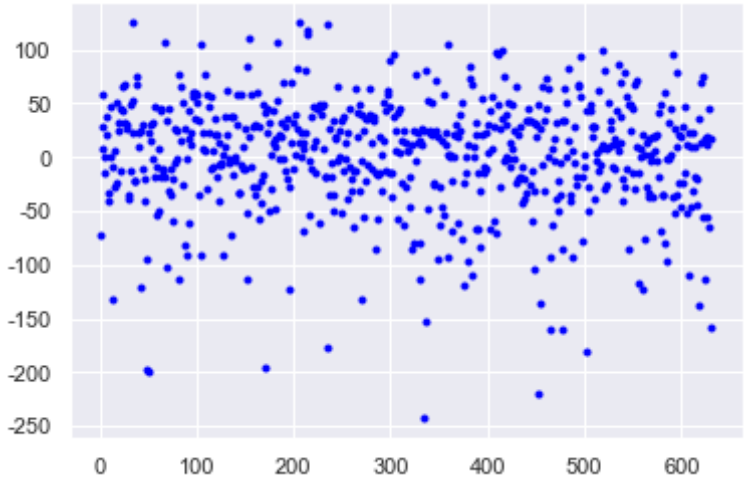


图 6.12 Lasso-CatBoost 模型测试集误差

利用 Python 中 important 函数获取非专业房东 Lasso-CatBoost 定价模型中重要度最高的前十个变量（见图 6.13），分别为可容纳人数、是否是整租房源、卧室数、最近的景点距离、最近的地铁站距离、实词数、房源设施数、房东注册天数、总评分、总评论数。重要性程度最大的前五个变量主要是房源属性变量（可容纳人数、房源类型、卧室数）以及区位优势特征。相较专业房东，非专业房东的定价模型重要变量中新增了对于在线评价以及房东自身属性的变量。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

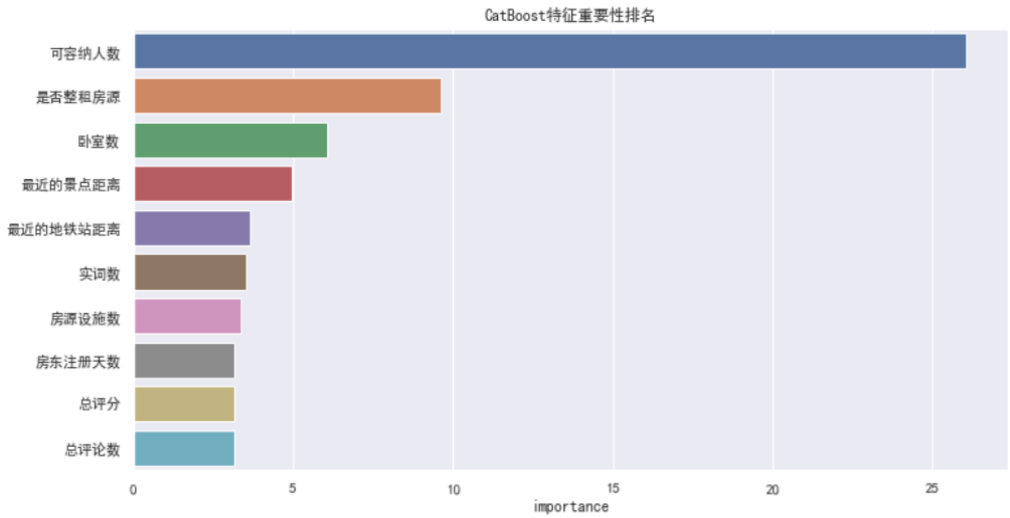


图 6.13 非专业房东定价模型的前十个重要变量

同时对于非专业房东的部分重要变量画出部分依赖图（PDP 图）来反映不同变量对于预测因变量的影响情况。以下四图分别是总评论数、房源设施数、实词数、房东注册天数的部分依赖图，可以发现这四个重要变量对于非专业房东房源的人均价格均呈现正向影响。与现实意义相符。

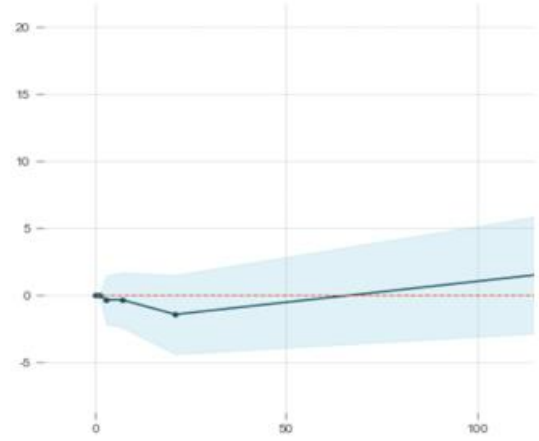


图 6.14 非专业房东总评论数 PDP 图

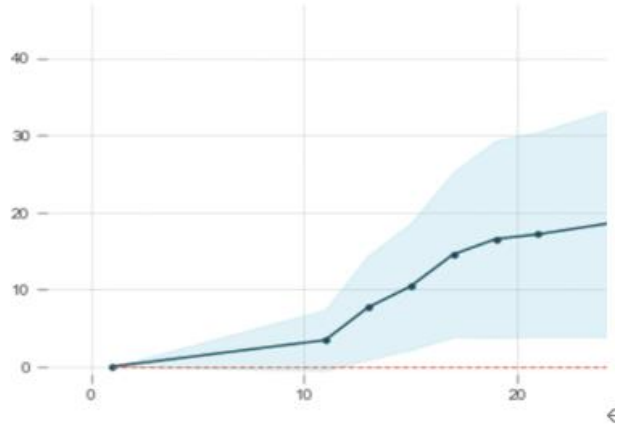


图 6.15 非专业房东房源设施数 PDP 图

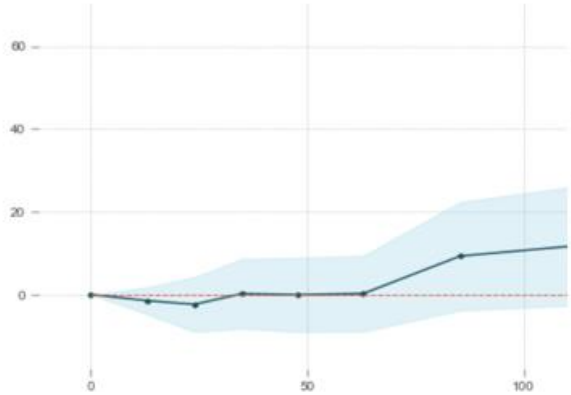


图 6.16 非专业房东实词数 PDP 图

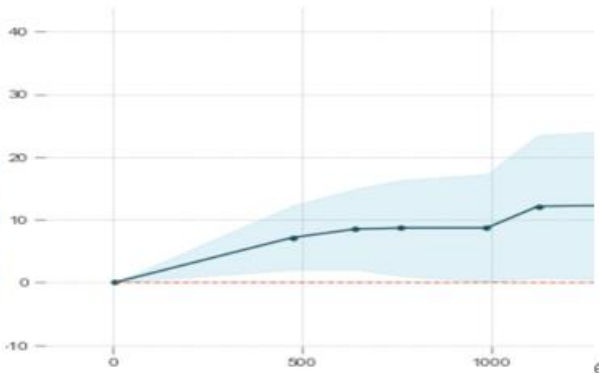


图 6.17 非专业房东注册天数 PDP 图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 7 章 结论及相关建议

7.1 结论

文本以房源为依据对 Airbnb 平台中的房东进行分类，对两类房东的房源分别建立逐步回归和分位数回归模型来研究不同类型房东的差异化定价机制。最后分别构建专业房东与非专业房东的最优定价模型，帮助短租平台实时更新企业数据库，为房东掌握平台的最新房价水平并提供动态定价服务。研究发现：

（1）对专业房东的短租房源进行逐步回归建模，最终保留了如下 32 个变量，其中包含房源显性和隐性变量共 11 个：是否闪订房源、是否是整租房、是否单人间、可容纳人数、床位数、卧室数、浴室数，最短预定天数、未来 30 天可定天数、实词数、主题熵；房东属性中保留的变量有 4 个：房东的注册天数、房东的认证方式数量、房东回复率、房东的回复时长；房源区位优势保留的变量有 8 个：该区二手房平均价格（万元/㎡）、1km 内的景点数、1km 内的地铁站数、1km 内的旅馆酒店数、最近公交站距离、最近旅馆距离、最近的景点距离、最近的地铁站距离；在线评价属性中保留的变量有 8 个：综合评分、干净整洁评分、沟通交流评分、位置便利评分、性价比评分、总评论数、每月评论数、最新评论时间。

（2）对非专业房东的短租房源进行逐步回归建模，最终保留了如下 23 个变量，其中包含房源属性变量有 9 个：床位数、浴室数、房源设施数、可容纳人数、是否是整租房、是否是单人间、最短预定天数、未来 30 天可定天数、实词数；房东属性中保留的变量有 4 个：是否超赞房东、房东注册天数、房东回复率、房东接受率；区域优势中保留的变量有 6 个：该区二手房平均价格（万元/㎡）、最近的地铁站距离、最近的景点距离、1km 内的旅馆酒店数、1km 内的地铁站数、1km 内的公交站数；在线评价中保留的变量有 4 个总评论数、每月评论数、最新评论时间、综合评分。

（3）**从专业房东和非专业房东逐步回归的差异性看：**对于专业房东来说，逐步回归后保留的变量明显多于非专业房东，说明专业房东在定价时会考虑到更多的因素，相比来说更加的成熟和全面。

针对于房源属性变量来说，专业房东模型中特有的显著变量为是否闪订房源、主题熵，并且均占有较大的权重。对于专业房东来说，他们对于自己的房源管理更加灵

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

活，其中房源的闪订政策，可以满足租户的任何时间段的租房需求，增加房源预定效率。另外专业房东更关注房源的主题丰富性，倾向于设计个性化房源，填补短租房源市场多样化风格，满足多元化租客的需求，来吸引更多的租户，实现自己的利润的最大化，且主题熵变量随着分位数点增加，对房源人均价格的影响从负向变为显著的正向，说明了专业房东的经营对房源产生的溢价性在高价区表现尤为突出。因此，平台可将这些个性化元素加入定价指标中并设定权重，制定一套易操作、可量化的定价指导系统，为个性化房源的溢价提供多维度信息支撑，帮助高价区房源合理定价，提倡房东削减部分超溢价，鼓励租客提升体验质感，主动承担合理溢价成本。

针对于房东属性变量，无论是专业房东和非专业房东，租户对于房东的信任度对最后是否预定房源起着非常重要的作用，进而影响到房东的定价。非专业房东模型中，房东的接受率也是一个显著性变量，可能是由于非专业房东由于房源数量较少，对于平台的关注量过少，即便是遇到租户预订的情形，也有可能因为错过回复时间而导致订单失败。

针对于区位优势变量来看，专业房东和非专业房东的显著性变量十分相似，对于最近的公交站距离、最近的景点距离、最近的地铁站距离，均随着房源人均价格分位数点的增加，负向影响呈现上升的趋势。相比较而言，专业房东考虑的会更加关注自身房源的地理位置优势，地铁站、公交站、旅馆或酒店、周边景点相关的区位优势变量均为显著变量。

针对于在线评价变量来说，专业房东和非专业房东相同的显著变量均随着房源人均价格分位数点的增加，正向影响越来也显著。专业房东逐步回归模型中独有的显著性变量是干净整洁评分、沟通交流评分、位置便利评分、性价比评分。对于专业房东来说，他们对于自己房源的运营更加的专业，他们会更加注重自己房源的整洁性、便利性，以及对于租户的礼貌性，通过租户给予的正向反馈，从而对自己的房源价格进行动态调整。

（4）专业房东定价体系较成熟，稳站短租市场优势地位；非专业房东定价体系尚未成熟，不易实现精准定价。

在基于房东特征选择分类为专业房东和非专业房东后，本文分别构建了两类房东在逐步回归变量选择结果和 Lasso 变量选择下的线性回归模型、随机森林模型、XGBoost 模型、CatBoost 模型，共计八组定价模型。将数据集按 4:1 的比例随机抽样分

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

割为训练集和测试集，并在测试集中对模型进行评估。经比较，专业房东定价模型中，**Lasso-XGBoost 模型对于专业房东房源价格的预测精度最高**，RMSE 值最小且拟合优度 R 方达到了 84.95%；非专业房东定价模型中，**Lasso-CatBoost 模型对非专业房东房源价格的拟合性能最佳**，预测误差最小，模型的拟合优度 R 方达到了 **80.73%。**

但综合来看，专业房东的最优定价模型拟合效果比非专业房东更好，拟合误差和随机波动也更小。可能是由于短租平台上的专业房东群体拥有更丰富的短租选址、运营经验和更长的运营体验，他们能准确地把握短租市场的变动进而快速做出调价举措，确保房源的性价比长时间处于相对优势的地位，保持持续的高市场占有率和高投资回报率。而非专业房东自主定价体系尚未成熟，在管理房源时难免存在技能上的生疏，对于房源价格的影响因素的把控没有专业房东经验丰富，。

7.2 相关建议

（一）为短租平台建立了“房源+环境”的一体数据库

本文充分利用了短租平台及高德地图实现数据库中的平台内部数据和外部数据（例如房源1km内的地铁站数、公交站数、景点数、旅馆和酒店数）结合，为短租平台现有的数据创立了房源结合生活便利、交通便利的标准化数据库，使平台在推荐房源定价时能充分考虑房源的周边环境因素，智能定价。

（二）筛选不真实房源，加快核查速度

本文通过对短租平台的数据预处理发现大量不合常规、下架或不真实房源。这些房源在平台中将分散租客的注意力，延长选租时间，影响租客的消费体验。短租平台只需通过房源数据库与房东标价及对应的房源配置等条件进行比对，就够高效且方便地对价格异常的房源进行监督和核查。且使用者可以依据主要需求修改模型判定的阈值，调整对真实房源的评判标准，自动筛除无效房源，大大提高工作效率，节省了大量人力成本。

（三）为新房东提供房源定价推荐

本文基于逐步回归变量选择、Lasso变量选择与随机森林、XGBoost、CatBoost的组合模型，通过网格搜索和交叉验证的方法构建专业房东和非专业房东的最优定价模型。最后应用训练好的模型对剩余每个房源价格进行判断，计算出价格预测值。当有新房东加入短租平台时，只需提供真实有效的相关定价重要变量，便可得到一个较为

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

合理的预测价格作为参考定价，避免了房东搜寻参考大量其他房源信息的麻烦，或盲目定价造成损失。

（四）为专业/非专业房东提供差异化定价服务

短租平台可定时更新数据库中的房源信息，帮助房东快速掌握平台的最新短租行情和房价水平。同时，根据最新加入的房源数据进行模型的迭代，进行定价模型规则的迭代，为专业房东和非专业房东提供更加实时的差异化的定价建议。

7.3 研究不足与展望

本文基于 Airbnb 短租平台提供的北京市真实房源数据，结合爬虫以及文本挖掘的方法构建了包含房源属性、房东属性、区位优势、在线评价四个方面总共 45 个指标。通过逐步回归和分位数回归来探察专业房东和非专业房东所考虑的重要因素之间的差异，并从机器学习角度出发，构建了多个 Lasso+集成学习的组合模型，在与逐步回归变量选择的组合模型的对比后，为两类房东分别选择合理的定价模型。

但是本文仍然有一些不足之处。1、本文在选择研究对象的时候，只选取了北京市房源样本，在未来的研究中，可以考虑根据城市类别的分类，比如说从超一线城市、一线城市、二线城市出发，探究不同城市下房东定价模型之间是否有差异。2、本文的模型中房源人均价格，是以某段时期该房源的平均价格进行代替。所以定价模型的灵活性不足，无法对每一天或者是不同时段进行更加灵活的动态定价。3、在最后的集成学习定价模型中，本文只是对模型的部分重要参数进行了网格搜索调参，在未来的研究中，可以通过调整其他的参数，来进一步提高两类房东定价模型的预测精度。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] 国家信息中心,中国共享经济发展年度报告(2021)[R/OL]. (2021-02-19)[2021-08-30] http://www.sic.gov.cn/News /557/10779.htm.

[2] Wang D, Nicolau J L. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com[J]. International Journal of Hospitality Management, 2017(62): 120-131.

[3] 王春英,陈宏民.共享短租平台住宿价格及其影响因素研究——基于小猪短租网站相关数据的分析[J].价格理论与实践,2018(06):14-17.

[4] Zhang Z H, Chen R J C, Han L D, et al. Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach[J]. Sustainability, 2017, 9(9): 1635.

[5] Hong I S, Yoo C S. Analyzing Spatial Variance of Airbnb Pricing Determinants Using Multiscale GWR Approach[J]. Sustainability, 2020, 12(11): 4710.

[6] Gutt D, Herrmann P. Sharing Means caring? Hosts' price reaction to rating VisiBility[J]. ECIS 2015 Research-in-Progress Papers, 2015: 54.

[7] 吕姝.社交需求对共享经济平台成交价格影响研究[D].大连理工大学,2017.

[8] 王璐瑶.房东个人信息披露对共享住房价格的影响[D].大连理工大学,2019.

[9] 赵寒.共享短租房东面部表情对租赁意愿的影响[D].重庆邮电大学,2020.

[10] Tussyadiah I P. Strategic Self-presentation in the Sharing Economy: Implications for Host Branding[M]. Springer International Publishing, 2016: 695-708.

[11] Xie K, Mao Z X. The impacts of quality and quantity attributes of Airbnb hosts on listing performance[J]. International Journal of Contemporary Hospitality Management, 2017, 29(9): 2240-2260.

[12] Li Y, Pan Q, Yang T, et al, Reasonable Price Recommendation on Airbnb Using Multiscale Clustering[C]. 2016 35th Chinese Control Conference (CCC), 2016: 7038–7041.

[13] Eva M F, Pedro M J. Hotels that most rely on Booking.com–online travel agencies (OTAs) and hotel distribution channels[J]. Tourism Review, 2018, 73(4): 465-479.

[14] 牛阮霞,何砚.基于特征价格模型的共享住宿平台房源价格影响因素研究[J].企业经济,2020,39(07):27-36.

[15] 吴晓隽,裘佳璐.Airbnb 房源价格影响因素研究——基于中国 36 个城市的数据[J].旅游学刊,2019,34(04):13-28.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[16] 蒋钰洁.在线短租房价格预测[D].贵州财经大学,2019.

[17] 曹睿,廖彬,李敏,孙瑞娜.基于 XGBoost 的在线短租市场价格预测及特征分析模型[J].数据分析与知识发现,2021,5(06):51-65.

[18] 张利君.基于部分线性模型的房价预测[D].哈尔滨工业大学,2018.

[19] 李圆圆.基于 BP 神经网络的北京市二手房交易价格预测研究[D].首都经济贸易大学,2018.

[20] 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.

[21] 赵晨阳.基于集成学习的二手房交易价格预测研究[D].山西大学,2020.

[22] 慕钢,张宏烈,党佳俊,李广峰.基于 LightGBM 模型的二手房房价预测研究[J].高师理科学刊,2020,40(12):27-31.

[23] 张家棋,杜金.基于 XGBoost 与多种机器学习方法的房价预测模型[J].现代信息科技,2020,4(10):15-18.

[24] 张望舒,马立平.城市二手房价格评估方法研究——基于 Lasso-GM-RF 组合模型对北京市二手房价格的分析[J].价格理论与实践,2020(09):172-175+180.

[25] 王春英,陈宏民.共享短租平台房东定价行为——基于小猪短租平台的数据分析[J].系统管理学报,2021,30(02):363-372+383.

[26] Benitez-A B. Why are flexible booking policies priced negatively?[J]. Tourism Management, 2018, 67: 312-325.

[27] 张乐.共享短租中房客信任影响因素及信任传递研究[D].北京邮电大学,2019.

[28] 许业超.基于文本挖掘的管理科学热点识别与演化分析[D].哈尔滨工业大学,2019.

[29] 丁照银.基于机器学习的评论文本分析[D].安徽师范大学,2019.

[30] Balaguer J, Pernias J C. Relationship between spatial agglomeration and hotel prices. Evidence from business and tourism consumers[J]. Tourism Management, 2013(36): 391-400

[31] Ert E, Fleischer A, Magen N. Trust and reputation in the sharing economy: The role of personal photos in Airbnb[J]. Tourism Management, 2016, 55(8): 62-73.

[32] 陈子燕,邓丽.短租市场租赁平台定价机制研究——基于不同房东类型的分析[J].价格理论与实践,2019(05):149-152.

[33] 贺平,兰伟,丁月.我国股票市场可以预测吗?——基于组合 LASSO-logistic 方法的视角[J].统计研究,2021,38(05):82-96.

[34] 田蓉媛,张学锋.“互联网+”时代的在线共享短租经济模式的研究[J].福建电脑,2018,34(09):93-95.

[35] 张甜.基于特征价格法的共享住宿日租价格预测[D].北京交通大学,2020.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[36] Blei D M, Probabilistic topic models, Communications of the ACM, 2012, 55(4): 77-84.

[37] 轩源.基于网络数据挖掘的共享住宿空间格局特征及影响因素研究[D].南京师范大学,2020.

[38] 田亚男,郑瑞坤.新型租赁模式在线短租的决定性因素研究——以小猪短租平台为例[J].中国经贸导刊(中),2020(01):46-49.

[39] 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.

[40] 董倩,孙娜娜,李伟.基于网络搜索数据的房地产价格预测[J].统计研究,2014,31(10):81-88.

[41] Ma Y X, Zhang Z J, Ihler A, et al. Estimating Warehouse Rental Price using Machine Learning Techniques[J]. International Journal of Computers Communications & Control, 2018, 13(2): 235-250.

[42] Robert T. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[43] Chen T Q, Guestrin C. XGBoost: A Scalable Tree Boosting System[J]. Cornell Unive rsity, 2016(1): 785-794.

[44] Kalehbasti P R, Nikolenko L, Rezaei H. Airbnb Price Prediction Using Machine Learning and Sentiment Analysis[J]. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2019: 173-184.

[45] Barron K, Kung E, Proserpio D. The Sharing Economy and Housing Affordability: Evidence from Airbnb[J]. SSRN Electronic Journal, 2018(4): 126-135.

[46] 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.

[47] 袁博,刘石,姜连勋, et al.基于随机森林回归算法的住房租金预测模型[J].电脑编程技巧与维护,2020(01):23-25.

[48] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research Archive, 2003, 3, 993-1022.

[49] Breiman, L. Random forests. Machine Learning, 2001, 45(1): 5–32.

[50] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[51] Zervas G, Proserpio D, Byers J W. The rise of the sharing economy: Esti mating the impact of Airbnb on the hotel industry[J]. Journal of Marketing Research, 2017, 54(5): 687-705.

[52] 李阳.共享住宿平台 Airbnb 的定价策略研究[D].北京外国语大学,2019.

[53] Dogru T, Mody M, Suess C, et al. The Airbnb paradox: Positive employment effects in the hospitality

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

industry[J]. Tourism Management, 2020, 77.

[54] Aznar J P, Sayeras J M, Segarra G, et al. Airbnb landlords and price strategy: Have they learnt price discrimination from the hotel industry? Evidence from Barcelona[J]. International Journal of Tourism Sciences, 2018, 18(1): 16-28.

[55] Wu X J, Shen J B. A Study on Airbnb's Trust Mechanism and the Effects of Cultural Values—Based on a Survey of Chinese Consumers[J]. Sustainability, 2018, 10(9): 3041.

[56] 王悠.基于特征价格模型的杭州市共享住宿价格影响因素分析[D].天津师范大学,2020.

[57] Gutierrez J, Garcia-P J C, Romanillos G, et al. The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona[J]. Tourism management, 2017, 62(10):278-291.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致谢

弹指一挥间，两年半时光已流逝。我的青葱岁月，我最后的学生时代即将画上一个完美的句号。我很感谢能在这里度过我的研究生生涯，也很感恩一路走来，师长与同学对我的帮助。浙工商见证我的成长，我一路向前，也一路蜕变。在此，我对我的父母、老师、同学、室友表示由衷的感谢。

首先，特别要感谢我的导师徐静教授。感谢徐老师在入校时接纳那个基础知识薄弱的我。感谢徐老师牺牲自己的闲余时间精心安排讨论班、精心挑选国内外学术论文报告，让我们在交流中不断汲取新知识，在学习专家学者的论文时培养科研能力。徐老师不仅在学习上指引我们，在生活上和职业发展道路上也给我们讲授过来人的经验之谈，避免我们走弯路。当然，在我整个毕业论文的撰写过程中，徐老师一直耐心地指导和督促我，从选题、到整体框架、到论文的书写格式的每一个环节都细心指导和宏观把控，才使得我的毕业论文能够如期完成。再次感谢徐老师两年半的谆谆教导，我会永远铭记在心。

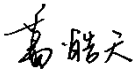
其次，我要感谢我的家人，不论是在生活中，还是学习中，你们都对我有最大的包容，无时不刻地给予我关怀和鼓励，让我在遇到困难时不再胆怯，在遇到失败时不再止步。你们在我选择继续攻读研究生后，从考研到研究生学习到研究生毕业，一直都在毫无保留地支持我，你们永远是我最坚强的后盾，因为有了你们的支持，我才能不惧困难、砥砺前行，安心完成我的学业并寻找自己喜欢的工作。

最后，也要感谢我的同门师兄弟、室友和其他同学们，茫茫人海能够相遇，本身就是一种缘分。感恩相遇，感恩一起陪伴的每一天、每一夜，我的研究生生涯正是因为有你们才变得如此精彩、如此丰富、如此有趣。相信多年后，当我们再次相遇时，也能欢快地谈论我们彼此之间共同的美好回忆。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

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，也不包含本人为获得浙江工商大学或其它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：  　　导师签名： 

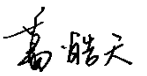
签字日期：2021 年 12 月 4 日　　签字日期：2021 年 12 月 4 日

关于论文使用授权的说明

本学位论文作者完全了解浙江工商大学有关保留、使用学位论文的规定：浙江工商大学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，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，并且本人电子文档的内容和纸质论文的内容相一致。

本论文提交□即日起/□半年/□一年以后，同意发布。

“内部”学位论文在解密后也遵守此规定。

学位论文作者签名：  　　导师签名： 

签字日期：2021 年 12 月 4 日　　签字日期：2021 年 12 月 4 日

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net