SUN YAT-SEN UNIVERSITY
中山大學
嶺南(大學)學院
LINGNAN (UNIVERSITY) COLLEGE

数据挖掘与机器学习报告

房源推荐算法——
基于 Airbnb 北京房价预测的机器学习模型

院系：中山大学岭南学院

课程编号：**LN3125**

课程名称：数据挖掘与机器学习

组别：第 **14** 组

何智钧　　**18333060**

刘懿瑾　　**18333126**

卢嘉婷　　**18332029**

彭礼敏　　**18333145**

宋红倩　　**18333160**

指导老师：王杉

完成日期：**2021** 年 **1** 月 **17** 日

房源推荐算法——
基于 Airbnb 北京房价预测的机器学习模型

摘要

我们对北京 Airbnb 房源价格进行预测，并试图定义以“推荐比”（Predict / Price，预测-实际价格比定义公平价格。首先，我们对房源进行凝聚式聚类。其次在各个组内我们使用线性回归、支持向量机、决策树、随机森林、极端随机树、梯度提升树、神经网络对价格进行预测并得到“推荐比”Predict / Price。我们发现影响价格的最主要因素是房屋大小（容纳人数、房间数等变量）。第三步，我们用评分相关变量对 Predict / Price 做拟合，发现性价比评分和 Predict / Price 有正相关关系，验证了 Predict / Price 对“性价比”的测度能力。最后，我们应用 LDA 主题模型挖掘评论中的文本信息，发现主题得分对于价格、Predict / Price、评分相关变量和聚类结果都有解释力度。

关键词 价格预测；凝聚式聚类；LDA 主题模型；线性回归；支持向量机；决策树；随机森林；极端随机树；梯度提升树；神经网络

目 录

1. 绪论

1.1. 爱彼迎公司概况

Airbnb 是 AirBed and Breakfast（"Air-b-n-b"）缩写。爱彼迎是一家连接旅游人士和家有空房出租的房主的服务型网站，可以为用户提供多样的住宿信息。Airbnb 成立于 2008 年 8 月，总部设在美国加州旧金山市。Airbnb 是一个旅行房屋租赁社区，用户可通过网络或手机应用程序发布、搜索度假房屋租赁信息并完成在线预定程序。据官网显示以及媒体报道，其社区平台在 191 个国家、65,000 个城市为旅行者们提供数以百万计的独特入住选择，不管是公寓、别墅、城堡还是树屋。Airbnb 被时代周刊称为“住房中的 EBay”。

图 1 爱彼迎理念

2015 年 2 月 28 日，美国短租网站 Airbnb 正在进行新一轮融资，而估值将达到 200 亿美元。2015 年 2 月，根据消息人士的说法，Airbnb 计划融资 10 亿美元，已完成了其中的一半，投资方包括富达、TPG、T. Rowe Price、Dragoneer、Founders Fund、红杉资本和俄罗斯 DST。2016 年 11 月 01 日，国外媒体报道美国短租服务公司 Airbnb 对中国用户表示，该公司将在本地存储他们的个人数据。

2017 年 1 月 27 日，总部位于旧金山的短期租赁网站 Airbnb 首次盈利，公司营业额增长超过 80%。爱彼迎平台表示，严厉禁止“刷单”等滥用平台评价机制的行为，一旦发现有入驻平台的商家、房东存在违规，将会采取房源排名降低、罚款甚至永久下线、封停账号等措施。

2019 年 11 月，Airbnb 和国际奥委会签订了一份价值约为五亿美元的协议，

涵盖了从 2020 年开始至 2028 年的五场夏季和冬季奥运会。Airbnb 将于东京奥运会开始为旅客以及奥运会工作人员提供住宿，可以看到其影响力越来越大。

1.2. 问题提出

1.2.1. 爱彼迎运营分析——SWOT 模型（基于中国市场）

1.2.1.1. 优势（strength）

1) 在世界范围内影响力广

目前，爱彼迎已覆盖全球 191 个国家和地区，6.5 万座城市，拥有 400 多万套房源，作为经营最早、范围最广的短租平台，在世界范围的影响力是目前其他短租平台无法短期内达到的。在国外，民宿已成为人们出行时住宿方式的主要选择之一。根据国家旅游数据中心发布的数据显示，2017 年全年外国游客入境旅游人数 2917 万人次，增长 3.6%。外国游客人数的增长，会对国内住宿有更多的需求。爱彼迎有着在世界范围的影响力，相比国内其他本土短租平台会吸引更多外国游客的入住。同样，中国游客到国外旅游看中爱彼迎在国外的影响力，认为其更加方便可靠而使用它，这也会对平台在中国的发展起到促进作用。

2) 服务项目多元化

提供民宿是爱彼迎的主要业务，另外，它不仅分享景点与美食攻略，还提供各种体验活动，如海上冲浪、野外露营等特色活动，让顾客深入地体验到与平时不一样的生活。住宿、活动、美食与景点一体化的服务平台，方便了用户的选择，让用户更加愿意在该平台进行消费。

1.2.1.2. 劣势（weakness）

1) 内部人员管理存在缺陷

中国区经营两年多，爱彼迎频繁更换负责人。负责人作为重要的企业管理人员，是战略的实施者，其能力、工作绩效对企业竞争力的提升有着密切关系。新上任的主要负责人需要花时间与精力去与了解企业的经营状况以及思考企业未来的发展。这有可能使企业战略实施方向与最初制定的不同，造成计划实施的延迟，可能要求员工调整或重做已完成的工作，数次之后员工丧失士气，不愿努力

工作甚至跳槽到竞争对手处，导致企业的人才流失。

2) 对房东的监管力度不够

爱彼迎通过收取服务费盈利，与其他平台不同的是它无须信用免押金，每个用户都可以免去预定时付押金的要求。只有当房屋出现毁损或其他意外情况，用户才会被房东要求索赔，平台上收取押金。但是，有房东会要求用户在使用房屋前在平台缴纳押金。如果房东私自收取的押金未及时归还，用户便成了弱势群体。另外，房东在房屋内安装监控侵犯他人隐私、无理由取消订单给用户造成损失等问题多次发生，如果这些问题不能及时有效地防范与解决，将会继续给爱彼迎带来严重的负面影响。

1.2.1.3. 机会（opportunities）

1) 科技的发展使民宿经营更加省心

随着科技的发展，已有民宿经营的一体化支持服务，将智能门锁直接与平台对接，构成方便快捷的云端物联网，用户下单支付后会自动收到密码可自行入住；退房后房东可在平台上一键呼叫保洁打扫房屋卫生为迎接新的客人，于是简化了房东经营的步骤，使得民宿经营更加省心，吸引更多房东入驻线上短租平台。

2) 乡村民宿市场的可开发性

越来越多城市居民希望在假期去体验乡村田园生活，呼吸新鲜空气，释放在快节奏生活中的压力。2018 年初发布的《中共中央国务院关于实施乡村振兴战略的意见》提出要实施休闲农业和乡村旅游精品工程，建设一批设施完备、功能多样的乡村民宿，发展乡村共享经济，助力乡村振兴。政策的出台将会推进乡村民宿的发展，若爱彼迎把握这次机遇，进行市场开发战略，将民宿情怀与文化传播到乡村，将会进一步扩大其在中国市场的综合影响力。

1.2.1.4. 威胁（Threats）

1) 国内信用体系不完整

爱彼迎在国外的顺利发展主要取决于国外比较完善的社会信用体系，而在我国，目前信用信息共享平台不够完善，也缺乏健全的信用法律体系保护个人信息

和解决纠纷问题。缺乏有效的信用监督管理制度，会造成房东和用户之间的信用危机，降低短租平台的使用率。

2) 国内短租平台的竞争激烈

这两年，国内新的短租平台如雨后春笋般发展起来，如途家、小猪短租等，虽然它们都是年轻的企业，但借助本土化发展的优势，它们能更好地掌握中国用户的消费特点。以竞争对手小猪短租公布的 2017 年业务数据情况为例，其成交总额增幅超过 350%，同时整体的订单增长幅度也超过 320%，每日新上线房源数超过 500 个，每日在线申请的用户房源达到 1500 个。面对一群国内快速发展的短租平台企业，爱彼迎应思考如何保持自己的竞争优势。

1.2.2. 爱彼迎运营分析——AARRR 海盗运营模型

AARRR（海盗模型）其实是用户生命周期的模型，但同时也有人拿来变成运营流程的模型：先拉新，其次促活，接着提高留存，然后获取收入，最后实现自转播。我们结合 AARRR 模型对爱彼迎进行分析。

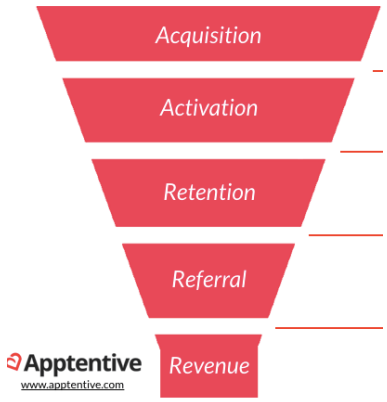

图 2 AARRR（海盗模型）示意图

1.2.2.1. 获取用户

获取用户指的是我们要了解目标用户群在哪，并且要最大程度地将他们转化成我们产品的用户。本阶段最主要的目的是将潜在的目标用户转化成我们产品的用户，并且开始使用产品。提高用户注册转化率的关键在于，调优产品的着陆页，要准确传达产品的核心价值。

1.2.2.2. 提高用户活跃度

活跃度的定义取决于产品，有的产品只要用户在指定时间内登录或启动一次就算用户活跃。对于移动应用产品，用户活跃度还有另外两个关键数据指标：每次启动平均使用时长和每个用户每日的平均启动次数。有时，产品除了登录和启动，还必须要求用户进行指定的操作才算用户活跃。

1.2.2.3. 提高用户留存率

用户留存率是非常重要的一个数据指标，留存率衡量着一个产品是否健康成长。留存率=登录用户数/新增用 户数*100%，其中，新增用户数是当前时间段内新注册并登录应用的用户数，登录用户数是当前时间段内至 少登录过一次的用户数。留存率反映的是一种转化率，由初期不稳定的用户转化为活跃用户、忠诚用户的过程。留存率一般有三个重要的指标：次日留存率=当天新增且在第 2 天还登录应用的用户数/当天新增的用户数。第 7 日留存率=当天新增且在第 7 天还登录应用的用户数/当天新增的用户数。第 30 日留存率=当天新增且在第 30 天还登录应用的用户数/当天新增的用户数。

1.2.2.4. 获取收入

获取收入就是要用户买单、消费，把留存用户转化为付费用户。本阶段的一个重要数据指标是 LTV，即用户给产品贡献的收入价值，是公司从用户所有的互动中所得到的全部经济收益的总和。

1.2.2.5. 自传播

自传播是指用户自发对产品进行口碑传播。自传播的数据指标是 K 因子（推荐系数），K=每个用户向朋友们发出的邀请数量*收到邀请的人转化为新用户的转化率。如果 K>1，用户群就会不断增加；如果 K<1，用户群就会逐步的停止增长。

表 1 自传播的关注指标以及问题关键

模型阶段	关注指标	关键
获取用户	新增用户、分渠道新增、分地域新增	获客成本低，转化留存高，ARPU 值高的
提高活跃度	AU(活跃用户)、活跃率、使用时长、启动次数	降低僵尸粉，做活动，更新信息
提高留存率	次日留存率，周、月留存率	一次用户少的渠道，定期激活
自传播	用户微信、微博的自传播	自传播转化
获取收入	ARPU（平均每用户收入）、消费用户比例、LTV（生命周期价值）	性价比，结合获客成本看

1.2.3. 问题确立

从 SWOT 分析与 AARRR 海盗模型分析中，我们可以看出爱彼迎等短租平台所面临的重要挑战是其性价比是否足够高，而随着短租平台竞争愈发激烈，爱彼迎的中国化路程似乎也不太顺利，我们在查阅资料之后，看到爱彼迎提供给房东的智能定价功能被广大房主吐槽，由于其定价机制非常不合理，且其价格并不能够最大化房东的利益。

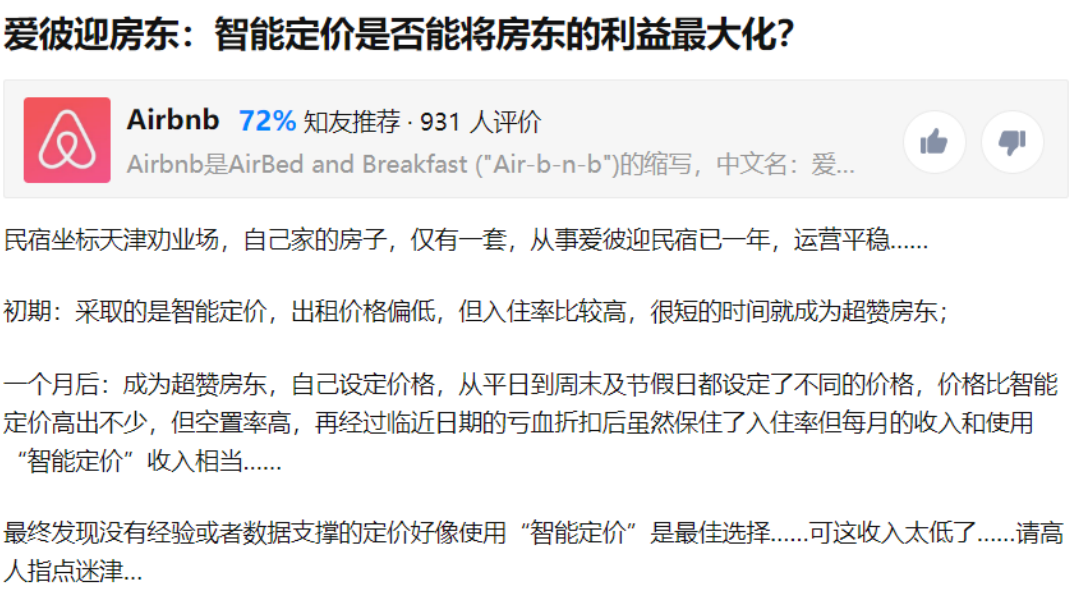

图 3 关于“智能定价”的看法

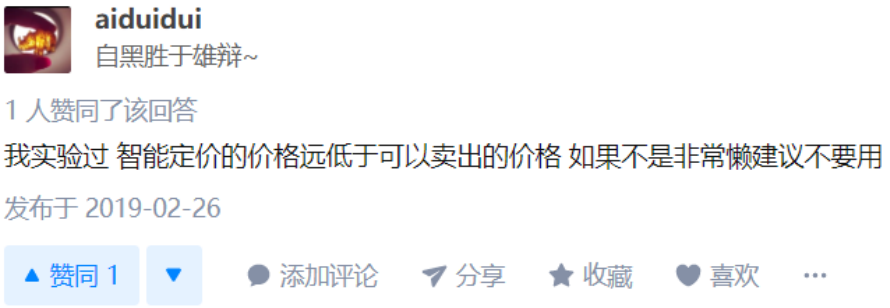

图 4 关于“智能定价”的看法

由此出发，印证了我们从运营模型中所发现的问题，爱彼迎在定价方面存在较大争议，且爱彼迎因为收费问题，在很多个国家受到过一定的处罚，其全球化平台的路径走的也并不容易。我们探寻爱彼迎的定价策略发现，爱彼迎定价主要分为四个部分：

1. **每晚价格**：由房东决定的每晚费用
2. **清洁费**：有些房东为支付其房源的清洁费用而收取的一次性费用
3. **额外房客费**：有些房东为支付与使用他们的房源有关的其他费用而收取的一次性费用
4. **其他费用**：

由于房东具有决定性的定价决策权，也就是说，爱彼迎平台上的房东收益情况往往需要他们主动对于定价进行监控和调整，而发布房源的房主通常通过两种方式进行定价，主要还是查看类似房源来对定价情况作以预估，而使用智能定价的房东反馈较为负面，智能定价在房东中真正的使用率非常低。

根据以上分析，我们确定研究爱彼迎的房源定价策略，对于短租房源的价格进行预测建模，通过这一举措，对于平台运营提出可借鉴建议。

1.3. 文献综述

P2P 等新技术的出现改变了城市的旅游发展模式，消费者可以通过基于共享经济理念的在线平台搜索和预订旅游住宿(Belk, 2014;Cusumano, 2015)。

房价解释特征的两组典型特征是：结构或物理特征(表面积、浴室数量等)和与房产位置或邻近地区相关的变量(Can, 1992)。然而，在房地产行业中，“位置、

位置、位置”是最重要的(Mueller & Loomis, 2008)。换句话说，一处房产的价格与邻近房产的价格密切相关(Gallin, 2008)。

对于酒店住宿，四种类型的因素被考虑:声誉属性(星级评级)、设施、消费者评级和位置属性(Blal et al., 2018; Castro & Ferreira, 2018)。设施，如餐厅、游泳池、电视、Wi-Fi 和健身房等，是文献中研究的另一个变量(Chen & Rothschild, 2010)。在线客户评论(消费者评级)的影响也与酒店客房价格有关(Andersson, 2010)。在这些模型中，最常用的可达性度量是距离城市中心的距离(Thrane, 2007)。旅游公寓定价的享乐模型也遵循类似的路线，包括结构性因素，如设施、消费者意见和地理因素(Yrigoy, 2018)。Kakar 等人(2018)将结构特征称为“出租列表特征”，将消费者意见称为“用户评论”，将区位因素称为“社区价值”。“在文献中，考虑与大小相关的结构特征也很常见，如卧室、浴室或床的数量(Kakar et al., 2018);Wi-Fi 或免费停车等便利设施(Lee et al., 2015);房客意见，如评论数量、评论分数等(Kakar et al., 2018);区位变量，如距离城市中心的距离(Gibbs, 2017)或最近的高速公路(Zhang, Chen, Han, & Yang, 2017)。最常用的方法是 OLS (Gibbs et al., 2017)，虽然也使用了其他方法，其中包括定量回归(Wang & Nicolau, 2017)和地理加权回归(Zhang et al., 2017)。

1.4. 分析思路

1.4.1. 逻辑思路

我们在这些理论基础的情况下，通过相应的不同类型变量，利用多种建模方式，进行对比与分析，最后陈述 Airbnb 房源价格预测带来的思考。以往的文献用回归较多，我们通过线性回归、SVM、Tree 模型、神经网络模型对 Airbnb 房源价格进行建模，并对不同的模型进行对比。

我们的前提假设为，条件相似的房源在定价上，**理论上不存在差别**，市场对于错误定价有一定的调节作用。我们所使用的数据集为北京市的爱彼迎房源，我们假定**大量的房源价格满足“公平定价”原则**，即不存在过高或过低的极端错误定价情况，进行建模，进而对于单个房源预测其价格。

1.4.2. 预期效果

1.4.2.1. 推荐性价比高的房源

提供推荐优先度排序模型，如在主页设立性价比较高的房源，提高该类优质房源的曝光度。我们以其预测价格与真实价格进行比较，若预测价格比真实价格低，表明其定价偏高，若预测价格比真实价格高，证明其定价偏低，可进入优先推荐的范围中。由此而来，让更多性价比高的房源被用户看到，提高用户**留存率**，同时对于平台本身的**盈利**起到积极的影响。

1.4.2.2. 优化智能定价功能

智能定价功能对于房主进行选择是一个参考作用，如何有效解决现有定价功能不够实用、无法实际代表房源价值的问题。应随着**房源价格、房源数量等**不断变化、迭代，对于房源价格的预测模型能够更加准确的提供定价推荐。

2. 数据挖掘相关理论

2.1. 无监督学习理论

2.1.1. 聚类

Agglomerative clustering 聚类是一种非监督式、自下而上的层次聚类算法(Hierarchical Clustering)。层次聚类通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。

Agglomerative clustering 聚类的流程如下：

a. 将每个样本都视为一个簇

b. 开始按一定规则，将相似度高的簇进行合并

c. 重复 b 过程，直至所有样本都形成一个簇或达到某一个条件

确定簇与簇之间相似度是该算法的要点，而相似度由簇间距离来确定。簇间距离小的相似度高，簇间距离大的相似度低。

簇间距离有四种表现形式，分别为 MIN、MAX、组平均与质心距离。

MIN 又称为“单链”，即不同簇之间的两个最近点之间的邻近度。从所有点作为单点簇开始，每次在两个簇的最近点之间增加一条链，先加最短的链，这些链将节点子集合并成为簇，单链技术适合处理非椭圆形状的簇，但是对于噪声和离群点十分敏感。簇 C_i 和 C_j 之间的组平均距离可以通过下式进行计算：

$$\text{proximity}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

MAX 又称为“全链”或“团”，即不同簇之间的两个最远点之间的邻近度。从所有点作为单点簇开始，每次在两个簇的最远点之间增加一条链，先加最短的链，这些链将节点子集连接成为一个簇（团）。全链技术生成的簇形状偏向于圆形，且不易受到噪声和离群点的影响，但是容易使大的簇破裂。簇 C_i 和 C_j 之间的组平均距离可以通过下式进行计算：

$$\text{proximity}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \text{proximity}(x, y)$$

组平均即不同簇之间的所有点对之间的邻近度的平均值，它是一种单链与全链之间折中的簇间距离计算方法。簇 C_i 和 C_j 之间的组平均距离可以通过下式进行计

算：

$$\text{proximity}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \text{proximity}(x, y)}{m, m_j}$$

依据对相似度（距离）的不同定义，可以将 Agglomerative Clustering 的聚类方法分为三种，分别为 Single-linkage,Complete-linkage 和 Group average。Single-linkage 是比较元素对之间的最小距离；Complete-linkage 比较元素对之间的最大距离；Group average 比较类之间的平均距离。

2.1.2. LDA

文本分析部分我们选择了 Latent Dirichlet Allocation (LDA)模型对评论的话题分布进行建模。代码的实现使用了 python 中的 gensim 包。

2.1.2.1. 模型设定

模型设定部分我们参考了 Bybee et al.(2019)和 Blei et al. (2003)的符号术语和数学表达。我们的语料库(corpus，D)由M篇文档(document, d)构成，每篇文档对应着一条房客的评论，可用集合D表示我们的文本库：

$$D = \{\mathrm{d}_1, \mathrm{d}_2, \dots, \mathrm{d}_M\}$$

模型输入：文档-词语矩阵**w**

w是$M \times V$矩阵，行对应M篇文档，列对应V个不重复的词。**w**中的元素 $w_{t,v}$ *代表*第v个单词在第t篇文档中出现的次数。

模型输出：文档-话题概率矩阵**Θ**和话题-词语概率矩阵 **Φ**

文档-主题概率矩阵$\mathbf{\Theta} = [\theta_1, \dots\dots, \theta_T]$是K $\times T$矩阵。对于第t个文档，θ_t是一K维向量，表示文档d_t在主题上的概率分布，即$\theta_{t,k} = P(Topic = k|document = d_t)$. 满足$\theta_{t,k} \geq 0$ 且 $\sum_k \theta_{t,k} = 1$.

主题-词语概率矩阵$\mathbf{\Phi} = [\phi_1, \dots\dots, \phi_K]'$ 是K $\times V$矩阵。对于第k个主题，ϕ_k是一个 V维向量，表示主题k在词语上的概率分布，即$\phi_{k,v} = P(word = v|Topic = k)$. 满足$\phi_{k,v} \geq 0$ 且$\sum_v \phi_{k,v} = 1$.

对第t篇文档d_t，对应词频向量w_t，w_t是V维向量，记录着每个词在文档d_t中

出现的频率。

模型假设w_t服从多项式分布：

$$w_t \sim Mult(\boldsymbol{\Phi}'\theta_t, N_t)$$

其中，N_t是文档d_t的总词数，Φ'是$\mathrm{V}\times K$矩阵，θ_t是K维列向量。

2.1.2.2. 模型原理

LDA 模型通过吉布斯抽样找到对θ_t，ϕ_k的估计，以产生和数据集最相似的信息。首先假设θ_t，ϕ_k的先验分布：

$$\theta_t \sim Dir(\alpha),\ \phi_k \sim Dir(\beta\quad)$$

比如第 t 篇文章共有N_t个词语，LDA 模型生成该文第一个词的过程如下所述：

①随机抽取 K 个主题中的一个，概率由θ_t（$(\theta_{t,1},\dots\dots,\theta_{t,K})'$）给出。把第一个词对应的主$z_{t,1}$视为随机变量，$z_{t,1} \sim mult(\theta_t, 1)$. 不妨设抽到了主题$k$ 即$z_{t,1}=k$，下面来抽取第一个词语。

②给定主题后，随机抽取 V 个词语中的一个，概率由ϕ_k（$(\phi_{k,1},\dots\dots,\phi_{k,V})'$）给出。把待抽取词语$x_{t,1}$视为随机变量，$x_{t,1} \sim mult(\phi_k, 1)$

更一般的有 $z_{t,i} \sim mult(\theta_t, 1)\ x_{t,i} \sim mult(\phi_{z_{t,i}}, 1)$ ，上述过程重复N_t 遍就生成了文档 t。

最后，$\theta_{t,k}$和$\phi_{k,v}$的估计可以由下式给出：

$$\hat{\theta}_{t,k}=\frac{\sum_{i=1}^{N_t}\mathbb{I}(\hat{z}_{t,i}=k)+\alpha}{\sum_{q=1}^{K}\sum_{i=1}^{N_t}\mathbb{I}(\hat{z}_{t,i}=q)+N_t\alpha},\quad \hat{\phi}_{k,v}=\frac{\sum_{t=1}^{T}\sum_{i=1}^{N_t}\mathbb{I}(\hat{x}_{t,i}=v)\mathbb{I}(\hat{z}_{t,i}=k)+\beta}{\sum_{q=1}^{K}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\mathbb{I}(\hat{x}_{t,i}=v)\mathbb{I}(\hat{z}_{t,i}=q)+K\beta}.$$

2.2. 有监督学习理论

2.2.1. 线性回归(Linear Regression)

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，根据自变量的多少可分为一元线性回归和多元线性回归。线性回归的模型为：

$$\mathrm{h(x)}=w_1x_1+w_2x_2+w_3x_3+\cdots+w_nx_n+\mathrm{b}$$

线性回归模型有很好的可解释性，从系数的大小能够直接看出每个变量特征对目标变量的影响方向和影响程度大小。当系数为正时，表示该变量对目标变量产生正向影响，反之则为负向影响。当系数的绝对值越大时，说明该变量对目标变量的影响程度越大，即变量在整个模型中的地位越重要。损失函数为：

$$\mathrm{f}=\frac{1}{n}\sum_{i=1}^{n}(y_i-h(y_i))^2$$

线性回归通常使用**最小二乘法**来减小损失函数值，从而使模型的拟合效果更好。线性回归模型的优点在于模型简单，建模迅速；具有很好的可解释性，有利于决策分析。但线性回归模型也存在缺点，对于非线性数据或数据特征之间具有多重共线性的数据难以进行有效的建模拟合分析。

线性回归是一个函数拟合的过程，为了防止过拟合现象的出现，在模型中会加入正则化项，不同的正则化项产生不同的回归方法。其中以 Ridge Regression（岭回归）和 Lasso 回归为经典，Ridge Regression（岭回归）加入 L2 正则化项，而 Lasso 加入 L1 正则化项，它们的目标函数如下：

$$\text{Lasso: }\beta^*=\underset{\beta}{\operatorname{argmin}}\frac{1}{n}\|y-X\beta\|_2^2+\lambda\|\beta\|_2$$

$$\text{Lasso: }\beta^*=\underset{\beta}{\operatorname{argmin}}\frac{1}{n}\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2$$

2.2.2. 支持向量机(SVM)

支持向量机是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。若样本本质为非线性可分，则需将其转化为线性可分样本再通过算法进行模型建立。其学习策略是间隔最大化，最终可转化为一个凸二次规划问题的求解。主要运用的算法有核函数和软间隔。

该模型的优点在于可用于线性或非线性分类，也可用于回归问题；泛化误差低；容易解释；计算的复杂度较低。

本文采用的支持向量机方法是 SVR。SVR 是从 SVM 中分支出来，一种用于回归的**二分类算法**。SVR 的原理是：在线性函数两侧制造一个“间隔带”，间距为 ϵ(即容忍偏差，是一个由人工设定的经验值)，对所有落入到间隔带内的样本不计算损失，即只有支持向量才会对其函数模型产生影响，最后通过最小化总损失和最大化间隔来得出优化后的模型。

SVR 的优化目标是：

$$\min_{w,b} \frac{1}{2}\|W\|^2$$

其中位于边界点内的点满足条件：

$$|y_i - (wx_i + b)| \leq \varepsilon$$

它的代价函数为：

$$\sum_{i=1}^{m} l_\varepsilon(f(x_i)\, y_i)$$

$$l_\varepsilon = \begin{cases} 0, & if|z| \leq \epsilon \\ |z| - \epsilon, otherwise \end{cases}$$

SVR 问题形式可化为：

$$\min_{w,b} \frac{1}{2}\|w\|^2 + loss$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_\varepsilon(f(x_i)\, y_i)$$

2.2.3. 决策树(Decision Tree)

决策树由决策结点、分支和叶结点组成，从根结点开始选择最合适的变量规则（即分类对象的属性）进行深度分支，沿决策树从上到下遍历最后会到达一个叶子结点，对数据进行分类，达到预测结果。

决策树的算法主要有 3 种，ID3、C4.5、CART 系数，通过算法来实现每一个分支上变量规则的选择，使决策树能达到更加准确的预测效果。ID3 算法通过计算每个属性的信息增益，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直到所有特征的信息增益均很小或没有特征可以选择为止，最后得到一个决策树。C4.5 算法与 ID3 算法相似，是对 ID3 算法的改进，使用信息增益

比来选择特征。CART 系数算法可以用于目标变量是多分类或连续型，对于连续型数据，使用均方误差作为划分标准，建立回归树进行预测分析；对于离散型数据使用基尼系数作为划分标准，建立分类树进行预测。

决策树的优点在于运算速度快，准确性高；适合高维数据。其缺点在于容易过拟合，泛化性能较差；忽略特征之间的相关性。

本文使用的是 Python 中的回归决策树 DecisionTreeRegressor，与普通二分类决策树(DecisionTreeClassifier)不同，DecisionTreeRegressor 使用均方误差(mse)或平均绝对误差(mae)作为标准(criterion)。计算公式如下：

$$\text{均方误差：MSE} = \frac{1}{M}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$\text{平均绝对误差：MAE} = \frac{1}{M}\sum_{i=1}^{m}|y_i - \hat{y}_i|^2$$

2.2.4. 随机森林(Random Forests)

随机森林是以决策树为基学习器、使用 Bagging 抽样方法的一种集成学习方法。从样本中以有放回抽样的方式取样形成一个训练集，基于训练集在建立每棵决策树的每一个节点时随机选取有限个特征，利用 ID3、C4.5、CART 算法建立决策树。用未抽到的剩余样本作预测、评估训练出的决策树的预测误差。

随机森林的优点在于其不容易过拟合，训练可以高度并行，能够给出每个特征的重要性，对部分数据缺失不敏感。但其缺点在于当森林中的决策树数量多时，模型训练所需要的事件和空间会比较大；并且随机森林模型由许多黑箱难以解释，在噪音比较大的样本数据上容易陷入过拟合的状态。

随机森林是基于 bagging 框架下的决策树模型，随机森林包含的树的生成规则如下：

a. 如果训练集大小 N，对于每棵树而言，随机且有放回地从训练中抽取 N 个训练样本，作为该树的训练集，重复 K 次，生成 K 组训练样本集。

b. 如果每个特征的样本维度为 M，指定一个常数 m<<M，随机地从 M 个特征中选取 m 个特征。

c. 利用 m 个特征对每棵树尽最大程度的生长，并且没有剪枝过程。

d. 按照步骤 a~c 建立大量的决策树，就构成了随机森林。

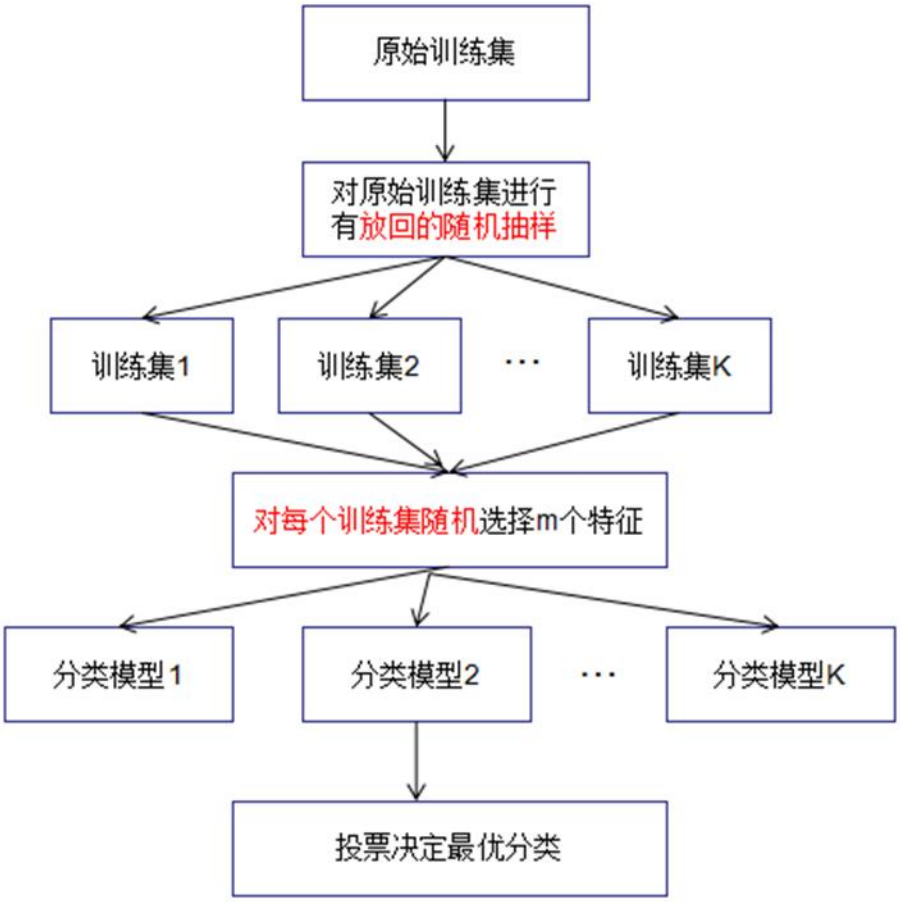

图 5 随机森林的分类算法流程

2.2.5. 极端随机树(Extremely Randomized Trees)

Extremely randomized trees 算法与随机森林算法十分相似，都是由许多决策树构成。ET 随机是指：特征随机 、参数随机 、模型随机（ID3 ,C4.5）、分裂随机。Extra tree 是随机森林（Random Forests）的一个变种，极限树与随机森林的主要区别：

a. 对于每个决策树的训练集，随机森林采用的是随机采样 bootstrap 来选择采样集作为每个决策树的训练集，而 Extra tree 一般不采用随机采样，即每个决策树采用原始训练集。Random Forest 应用的是 Bagging 模型，Extra Tree 使用的所有的样本，只是特征是随机选取的，因为分裂是随机的，所以在某种程度上比

随机森林得到的结果更加好。

b. 在选定了划分特征后，RF 的决策树会基于信息增益，gini（基尼系数），均方差等原则，选择一个最优的特征值划分点，这和传统的决策树相同。但是 Extremely randomized trees 比较的“激进”，会随机的选择一个特征值来划分决策树。

2.2.6. 梯度提升树(XGBoost)

XGBoost 是将分类回归树(CART 树)进行组合，它是一种树的集成学习方法。它的核心算法思想为：

a. 不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数 f(x)，去拟合上次预测的残差。

b.训练完成得到 k 棵树，根据样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，去预测一个样本的分数。

c.将每棵树对应的分数加起来得到该样本的预测值。

其预测模型为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

其中 K 为数的总个数，f_k表示第 k 棵树，$\hat{y}_i$表示样本x_i的预测结果。

损失函数为：

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

其中$l(y_i, \hat{y}_i$为样本x_i的训练误差，$\Omega(f_k)$表示第 k 棵树的正则项。

XGBoost 模型基于 GBDT 算法完成。GBDT 算法包括了 CART 回归树和梯度提升树(Gradient Boosting)，该算法的过程为：

a. 初始化弱学习器：

$$f_0(x) = arg\min_{c} \sum_{i=1}^{N} L(y_i, c)$$

b. 对 m=1,2,…M, 对每一个样本 i=1，2,…N，计算负梯度，即残差

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

c. 将残差作为样本的新的真实值，将数据（x_i,r_{im}）,i=1,2,…N 作为下棵树的训练数据，德奥一颗新的回归树$f_m(x)$，其对应的叶子节点区域为 R_{jm}，j=1,2,…J，其中 J 为回归树的叶子节点个数。

d. 对叶子区域 j=1,2,…J 计算最佳拟合值

$$\gamma_{im} = \arg min \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

e. 更新强学习器

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$$

f. 得到最终学习器

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$$

2.2.7. 神经网络(Neural Network)

神经网络是由众多的神经元可调的连接权值连接而成，具有大规模并行处理、分布式信息存储等特点。由输入层、隐藏层和输出层构成，每一层均分布有多个神经元。有单层神经网络与多层神经网络之分，单层神经网络只有一个隐藏层，多层神经网络有多个隐藏层，每一个隐藏层均被赋予一种激活函数。根据目标变量是否连续来决定输出层的神经元个数，分类型问题输出层有多个神经元，回归型问题输出层只有一个神经元。其中激活函数包括 Linear activation(即 f(net)=net)、Sigmoid activation(即 $f(net) = \frac{1}{1-e^{-net}}$)、Threshold activation（即 $sign(net) = \begin{cases} 1, if\ net \geq 0 \\ -1, if\ net < 0 \end{cases}$）、Hyperbolic tangent activation(即$f(net) = \tanh(net) = \frac{1-e^{-2net}}{1+e^{-2net}}$)。

神经元是神经网络的基本单元。每一个神经元先获得输入，对输入赋予一定权重并进行加和，再经过激活函数的计算产生一个输出，此输出将进入下一个神经元。通过 BP 算法（即 Backpropagation，反向传播算法）或梯度下降算法来调整每一个神经元的权重，循环算法，直至权重不再发生改变。

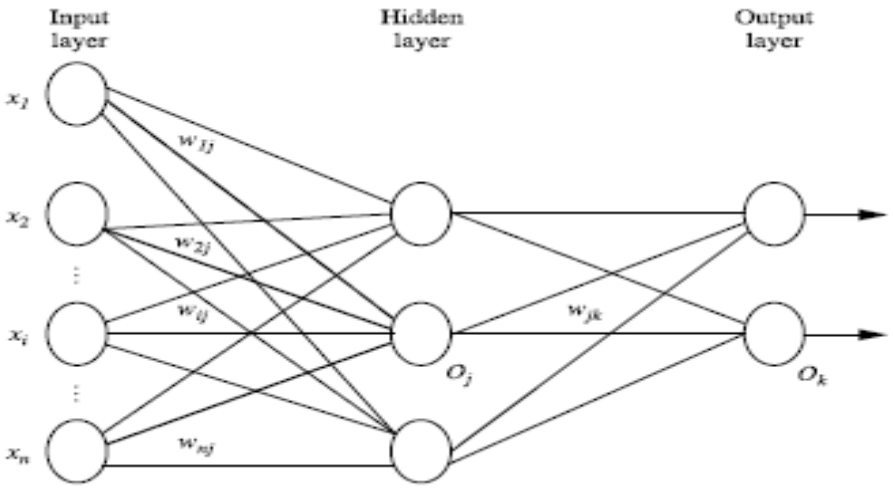

图 6 神经网络结构示意图

本文采用的是 Python 中 MLPRegressor（多层回归感知机）来建立神经网络模型。多层感知机基于 BP 算法，BP 算法利用可导的激活函数来描述输入与输出之间的关系，例如使用 sigmoid 函数作为激活函数。

输入：$\mathrm{net} = x_1w_1 + x_2w_2 + \cdots + x_nw_n$

输出：$\mathrm{y} = \mathrm{f(net)} = \frac{1}{1-e^{-net}}$

BP 算法由信号的正向传播和误差的反向传播两个过程组成。正向传播时，输入样本从输入层进入网络，经隐层逐层传递至输出层，如果输出层的实际输出与期望输出(导师信号)不同，则转至误差反向传播；如果输出层的实际输出与期望输出(导师信号)相同，结束学习算法。反向传播时，将输出误差(期望输出与实际输出之差)按原通路反传计算，通过隐层反向，直至输入层，在反传过程中将误差分摊给各层的各个单元，获得各层各单元的误差信号，并将其作为修正各单元权值的根据。这一计算过程使用梯度下降法完成，在不停地调整各层神经元的权值和阈值后，使误差信号减小到最低限度。

3. 房源推荐模式的实现

3.1. 数据概况

3.1.1. 数据集情况

表 2 数据集情况

数据集来源	数据类型	地区	变量数	样本量	是否存在缺失值
爱彼迎官网	字符、数值、日期	北京	74	27439	是

3.1.2. 变量格式与特征

表 3 变量格式与特征

序号	变量	数据类型	特征
1	*id*	字符	房源编码
2	*listing_url*	字符	房源 url 地址
3	*scrape_id*	字符	记录 id
4	*last_scraped*	日期	最后记录数据时间
5	*name*	字符	房源名称
6	*description*	字符	房源描述
7	*neighborhood_overview*	字符	描述周边
8	*picture_url*	字符	房源照片
9	*host_id*	字符	房东 id
10	*host_url*	字符	房东 url
11	*host_name*	字符	房东姓名
12	*host_since*	日期	房东入驻日期
13	*host_location*	字符	房东地址
14	*host_about*	字符	房东描述
15	*host_response_time*	字符	房东回复频率
16	*host_response_rate*	字符	房东回复率
17	*host_acceptance_rate*	字符	房东预定接受率

序号	变量	数据类型	特征
18	*host_is_superhost*	字符	房东是否是明星房东
19	*host_thumbnail_url*	字符	房东缩略图 url
20	*host_picture_url*	字符	房东图片 url
21	*host_neighbourhood*	字符	房东周边
22	*host_listings_count*	数字	房东发布数量
23	*host_total_listings_count*	数字	房东发布总数量
24	*host_verifications*	字符	是否有资质
25	*host_has_profile_pic*	字符	是否有外形图片
26	*host_identity_verified*	字符	是否通过认证
27	*neighbourhood*	字符	区域情况
28	*neighbourhood_cleansed*	字符	区
29	*neighbourhood_group_cleansed*	字符	区名称
30	*latitude*	数字	纬度
31	*longitude*	数字	经度
32	*property_type*	字符	房源类型
33	*room_type*	字符	房间类型
34	*accommodates*	数字	可容纳人数
35	*bathrooms*	数字	洗手间数量
36	*bathrooms_text*	字符	洗手间描述
37	*bedrooms*	数字	卧室
38	*beds*	数字	床的数量
39	*amenities*	字符	设施
40	*price*	数字	价格
41	*minimum_nights*	数字	最小预定
42	*maximum_nights*	数字	最大预订
43	*minimum_minimum_nights*	数字	最小预定的最小值
44	*maximum_minimum_nights*	数字	最小预定的最大值

序号	变量	数据类型	特征
45	*minimum_maximum_nights*	数字	最大预订的最小值
46	*maximum_maximum_nights*	数字	最大预订的最大值
47	*minimum_nights_avg_ntm*	数字	最短租时间均值
48	*maximum_nights_avg_ntm*	数值	最长租时间均值
49	*calendar_updated*	日期	最近一次日历更新时间
50	*has_availability*	字符	是否可租
51	*availability_30*	数字	30 天内出租天数
52	*availability_60*	数字	60 天内出租天数
53	*availability_90*	数字	90 天内出租天数
54	*availability_365*	数字	365 天出租天数
55	*calendar_last_scraped*	日期	日历相关信息最后抓取时间
56	*number_of_reviews*	数字	评论反馈数
57	*number_of_reviews_ltm*	数字	评论反馈数（12 个月）
58	*number_of_reviews_l30d*	数字	评论反馈数（30 天）
59	*first_review*	日期	首次反馈
60	*last_review*	日期	最新反馈
61	*review_scores_rating*	数字	评论等级
62	*review_scores_accuracy*	数字	评价：区位准确
63	*review_scores_cleanliness*	数字	评价：清洁程度
64	*review_scores_checkin*	数字	评价：登记入住
65	*review_scores_communication*	数字	评价：交流回复
66	*review_scores_location*	数字	评价：区位条件
67	*review_scores_value*	数字	评价：性价比
68	*license*	字符	是否具备房屋出租相关证件
69	*instant_bookable*	字符	是否可以闪定
70	*calculated_host_listings_count*	数字	房东出租的房屋数
71	*calculated_host_listings_count_entire_homes*	数字	房东出租的房屋数

序号	变量	数据类型	特征
72	*calculated_host_listings_count_private_rooms*	数字	房东出租的独立房间数
73	*calculated_host_listings_count_shared_rooms*	数字	房东出租的合住房间数
74	*reviews_per_month*	数字	每个月的反馈评价

3.2. 数据预处理

下载数据集包含三个数据文件，用到的是 listings.csv——包含了房屋、房主、房客评分的信息，和 reviews.csv——包含了历史上房客对于房屋的评论。

3.2.1. listings.csv

该数据集共包含 27439 个房源信息，包含多个不同类型的变量，如房屋类型及信息、房主信息、评论历史信息等。我们根据观察，对其进行以下初步处理：

（1）删除用不到的变量（照片、网址 url）

（2）缺失值处理：

把缺失值较多的 *reviews*评分变量按均值填补；

reviews_per_month 的缺失值填充为 0；

（3）数据类型转换：

时间变量(*first_review*, *last_review* 等)：转为到现在的时间(int)；

二元类别变量(*host_is_superhost* 等)：转为 0-1 变量(int)；

多元类别变量(*neighbourhood_cleansed* 等)：生成 0-1 虚拟变量(int)；

其它不规范的字符串变量(*price*, *bathrooms_text* 等)：去掉标点和文字，转为 float 或 datetime object；

字符串列表(*host_verifications*, *amenities*)：转化为列表长度(int)，如[“email”, “tel”]转为 2；

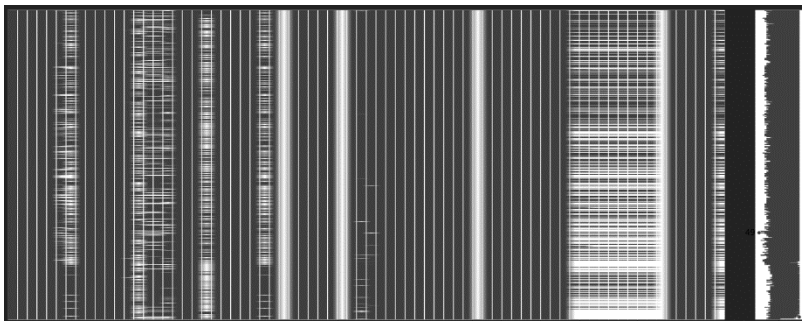

图 7 原始数据缺失情况

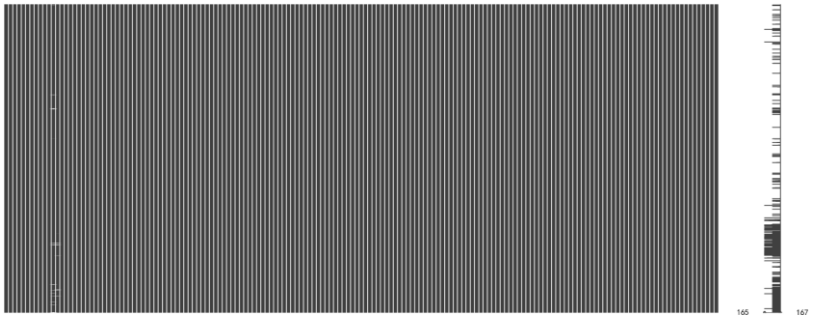

图 8 处理后数据缺失情况（后面的空白很多是虚拟变量）

3.2.2. reviews.csv

reviews.csv 一共有 171981 条，去缺失评论后有 171854 条，对应 14984 个 listing_id。

（1）文本翻译

```
zh      153190
en       16721
ko         402
else       375
fr         315
ja         258
es         232
de         157
ru         136
Name: lang, dtype: int64
```

图 9 评论语种

通过使用 langid 包我们得到了每个各条评论对应的语言。数据集评论中文为主(约 90%)，多个小语种。为了充分利用文本信息，我们改编了对谷歌翻译网站

的排重代码，对非中文评论进行批量翻译。尽管翻译的过程中有时确实会偏离中文原意，但大致意思是准确。

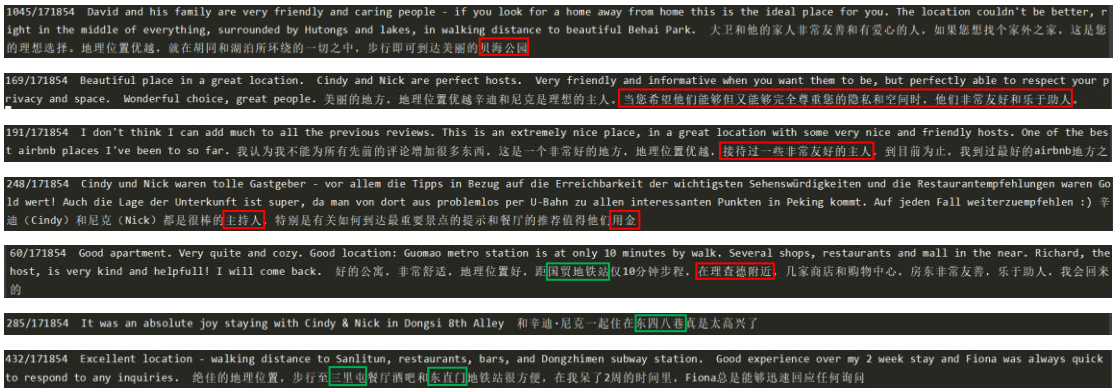

图 10 翻译过程截取（有偏离但大致准确）

我们使用 SnowNLP 包计算评论情感分数 sentiScore，发现和 7 个 reviews 评分都高度相关，最高有 0.59 的相关系数，这进一步说明了评论中很可能蕴含着有用的信息。

（2）jieba 分词

为了进一步对文本数据进行分析和 LDA 主题模型建模，我们使用 jieba 模块对评论数据进行了分词处理，主要涉及以下步骤:

词性筛选：保留了名词、人名、地名、机构团体名、其它专名、名词性惯用语等。

导入停用词词典[1]：去掉“了”、“的”、“都是”等常见停用词。

导入用户自定义词典：比如“古北水镇”就不能被 jieba 分词识别出，会分成“古北”和“水镇”。我们手动添加了一些可能出现的词，比如“冰糖葫芦”等。

[1] 词典来源: https://blog.csdn.net/u010533386/article/details/51458591. 我们还在其中加入了“房东”、“房间”等词，因为这些词在每个主题出现的可能性都比较大且对我们识别主题帮助不大。

3.2.3. calendar.csv

定义调整价格 *adjusted price*，且 *adjusted price* < *price*，通过下图，我们可以得到房源的日度、周度、月度平均价格以及淡旺季的调整（如过年期间价格下降等）等信息。这也体现了短租酒店行业的特点，与具体订房时间日期相关，择时订房对于房客来说更加重要。

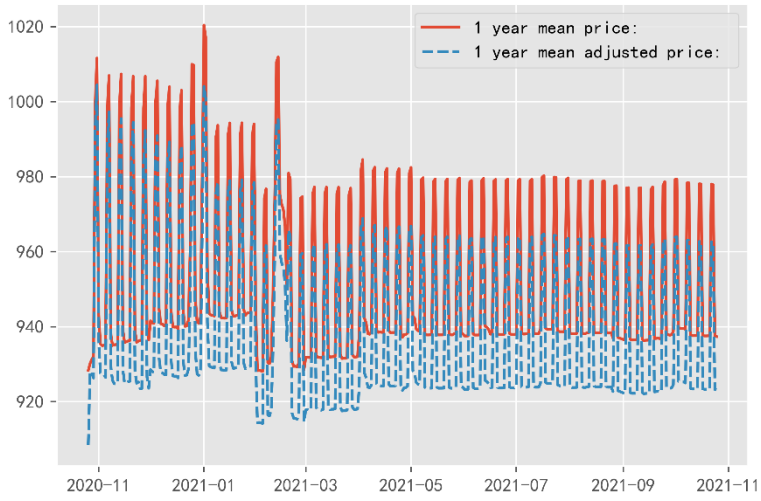

图 11 全部房源日度平均价格

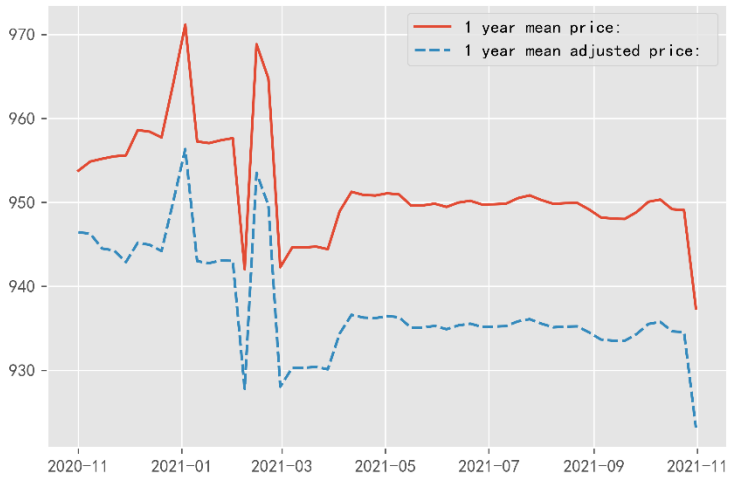

图 12 全部房源周度平均价格

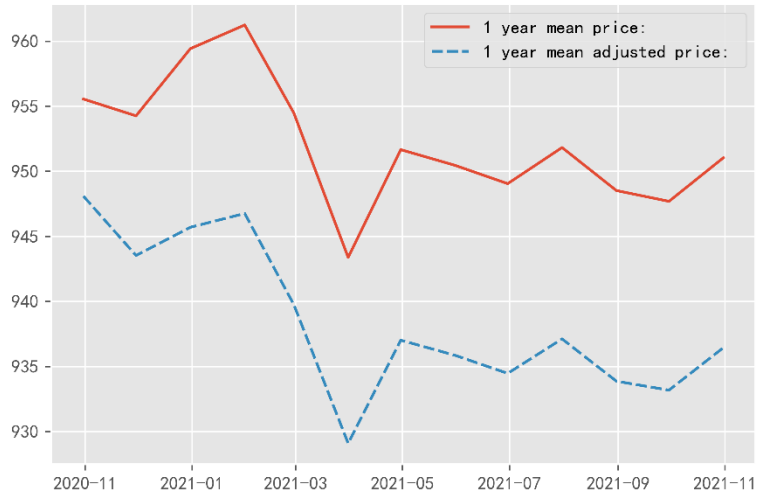

图 13 全部房源月度平均价格

3. 3. 探索性分析

3. 3. 1. 描述性统计

3.3.1.1. 区域相关

1) 房源区域分布

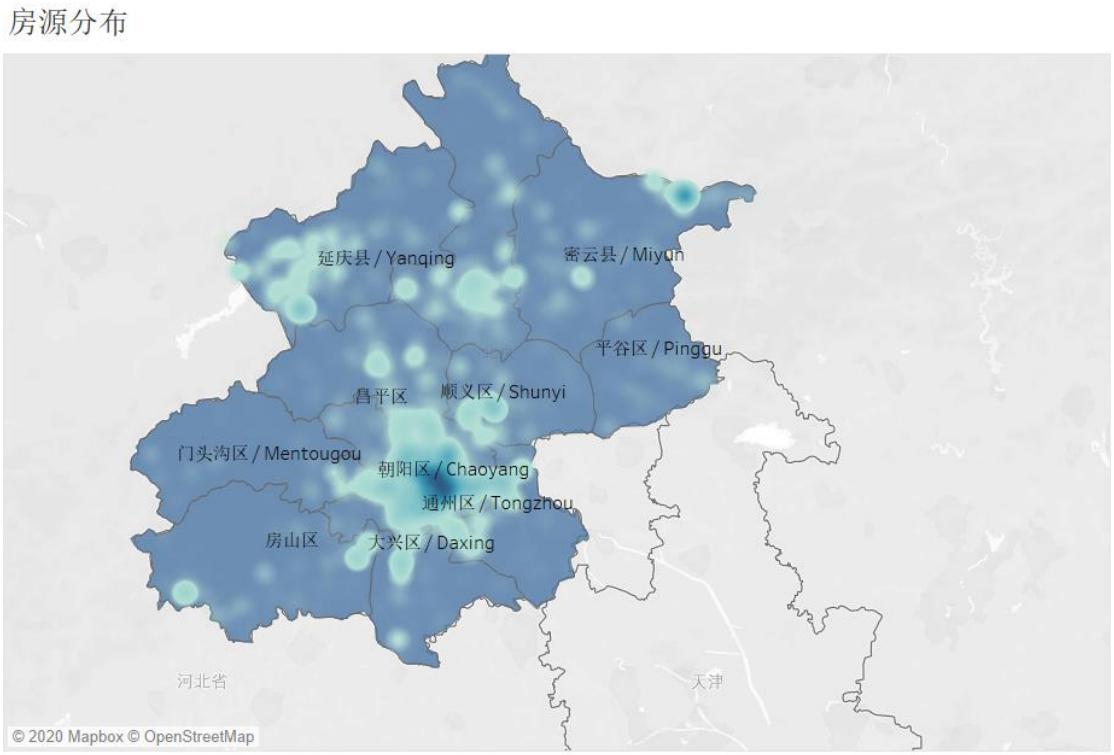

图 14 房源分布图

利用 tableau，我们对于房源的经纬度、区域进行匹配，得到房源的区域分布，东城区、西城区、朝阳区、海淀区房源较为丰富，更加集中；其他区域较为分散。在郊区部分中，延庆县相对而言房源更加丰富。

2) 不同区域均价差异

对于不同区域的房源来说，价格同样存在较大的差异。可以明显的看到，东西城、平谷区等地，房源均价较高。我们可以看到一些郊区的部分价格较高，可能是由于房源数量少，但房源面积较大，所以整体价格较高的情况。

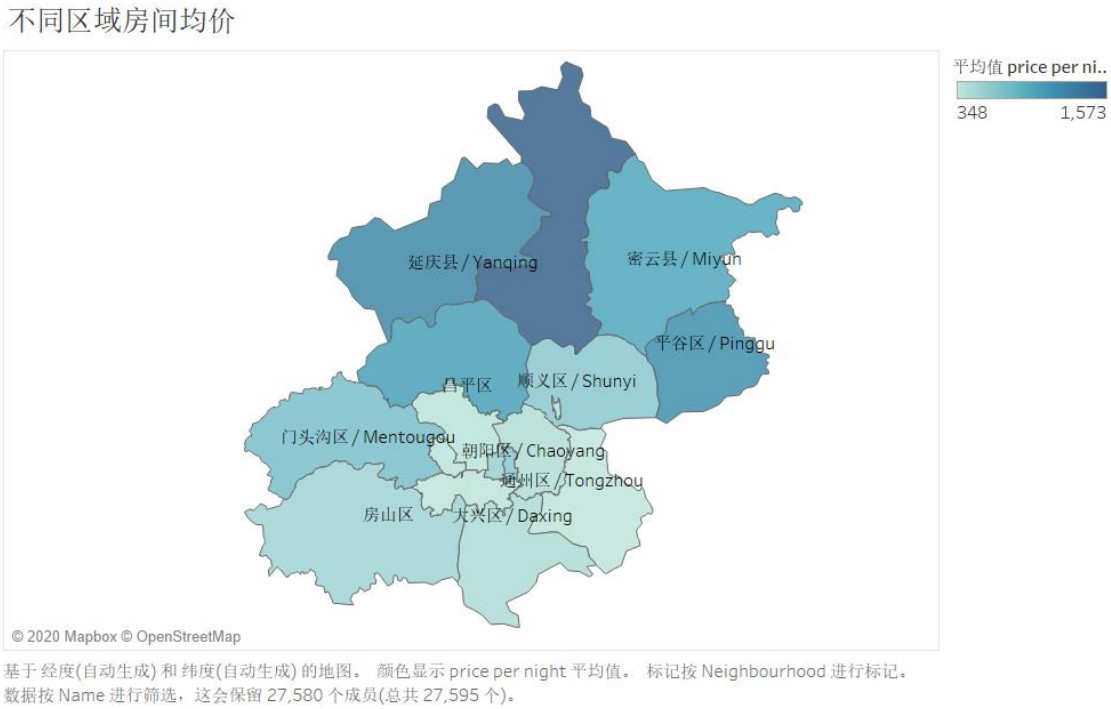

图 15 不同区域房间均价

3.3.1.2. Price 与房间规模

1) Price 分布

由下图可知，房源总体价格集中于 1000 元以下，符合常理。短租平台也常以其性价比为优势吸引用户，高于 1000 元的较少。有一些异常值，价格极高，可能是定价错误或房主短期内不进行租赁，以此来保持房源的留存。

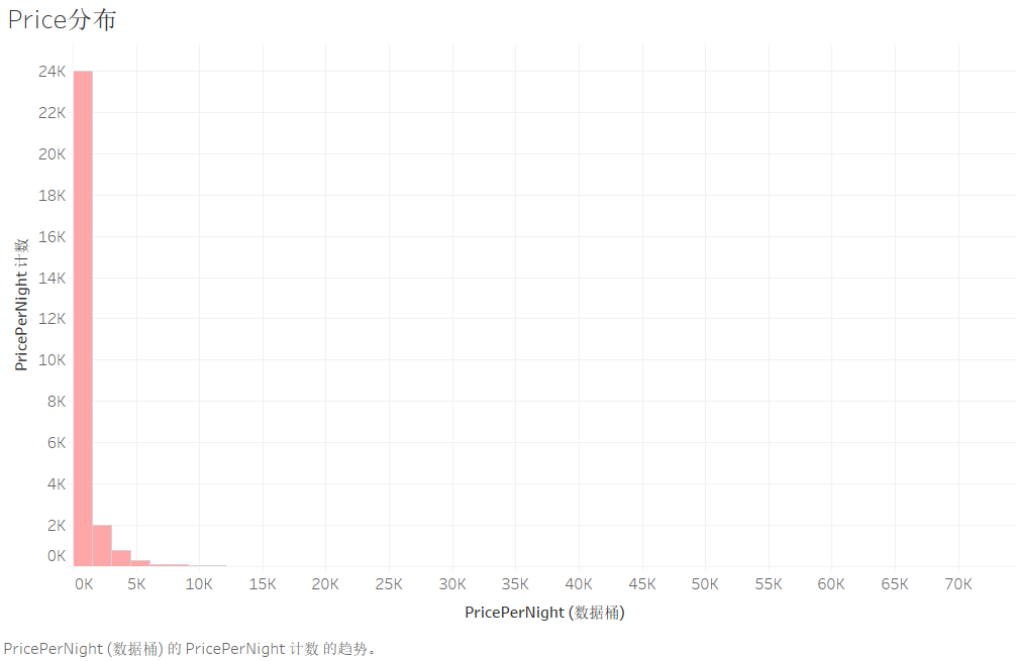

图 16 价格分布

2) Price 和 Accommodates

由下图可知，房间可容纳人数不同，房间均价不同，房间容纳数量越多，其均价呈现一个上升的趋势，如图所示，可容纳五人及以下的房源均价均位于 1000 元以下，其他可容纳人数较多的房间，也许是独栋别墅等，价格较高。

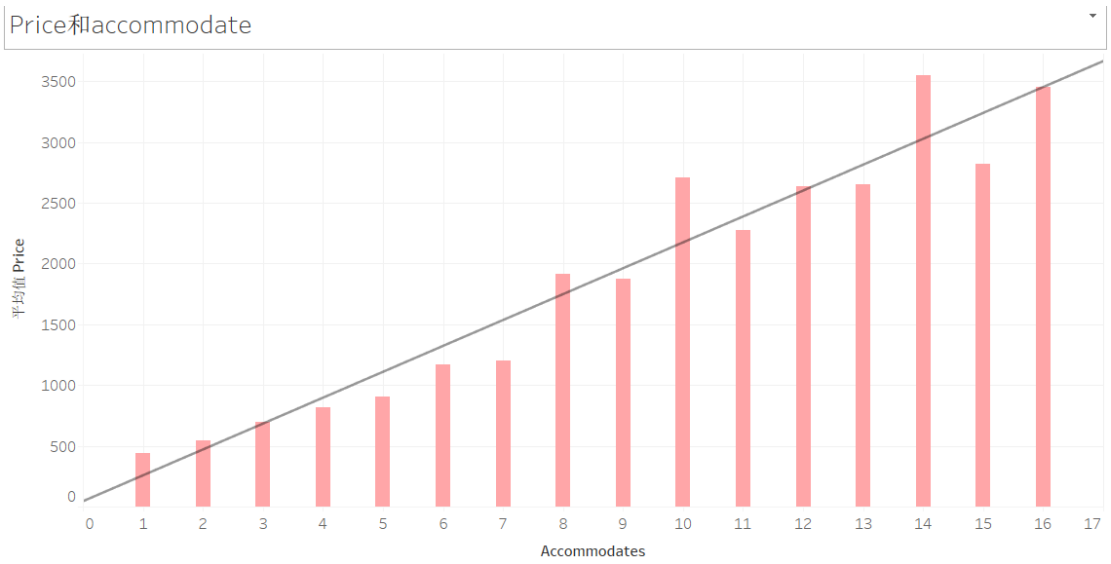

图 17 price 和 accommodates

3) Price 和 bathrooms

如图所示，不同的洗手间数量对应的房间价格均值不同，最大值出现在 6.5 个共享洗手间。同样是与房间规模有关的变量，整体与 Price 呈现正相关的趋势。

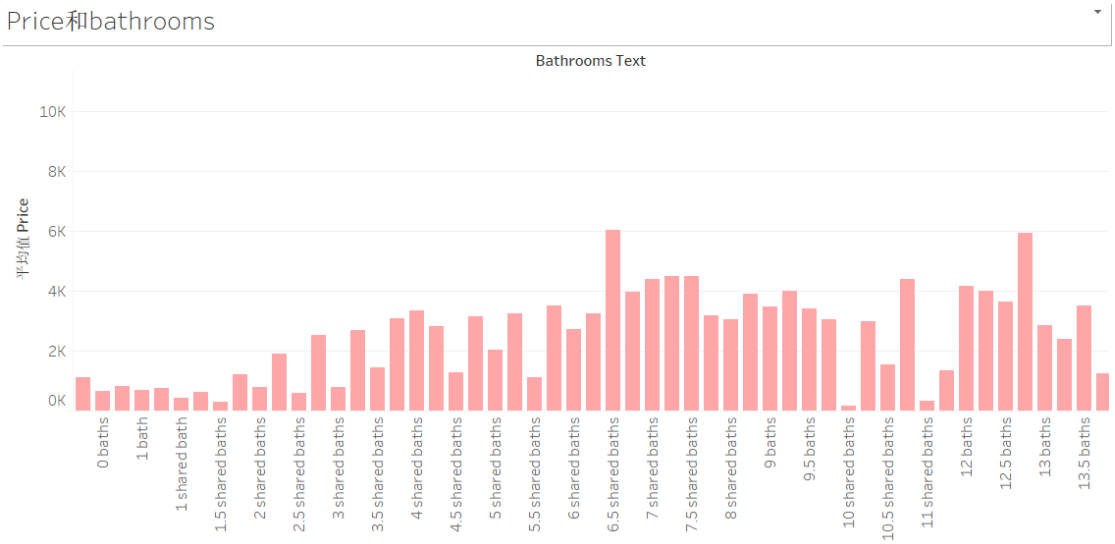

图 18 price 和 bathrooms

4) Price 和 bedrooms

与前文的 bathrooms 类似，bedrooms 更加直接的体现了房间的大小规模。由下图可知，总体来看，Price 和 bedrooms 大体上成正相关的关系。

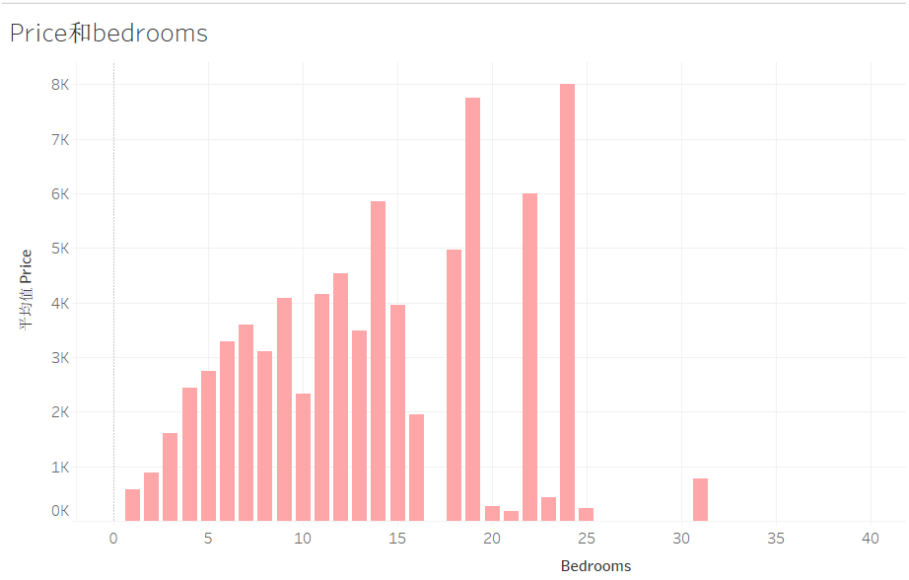

图 19 price 和 bedrooms

3.3.1.3. Price 与评价

1) Price 与评价：区位准确、评价：登记入住、评价：交流回复

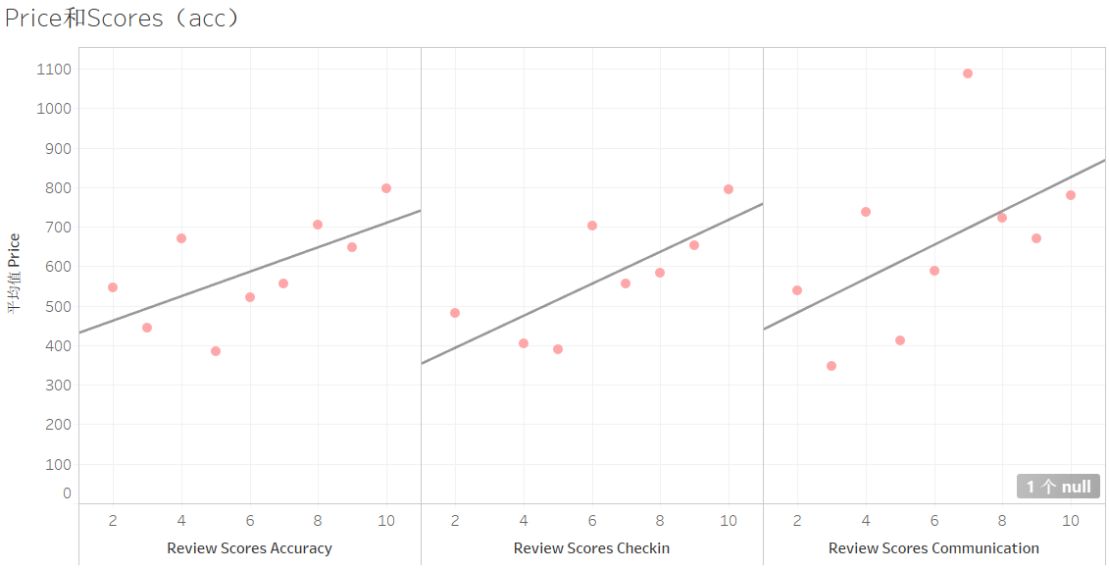

图 20 price 和 scores

如图所示，对于区位准确、登记入住、交流回复三者评论而言，对于不同的评论分数来看，总体上房间均价与评价均分成正相关关系。这三者主要是与房主相关的因素，是一个较为主观的评判，且不同的房客会有自己不同的标准。

2) Price 与评价：区位条件、评价：性价比、评价：清洁程度

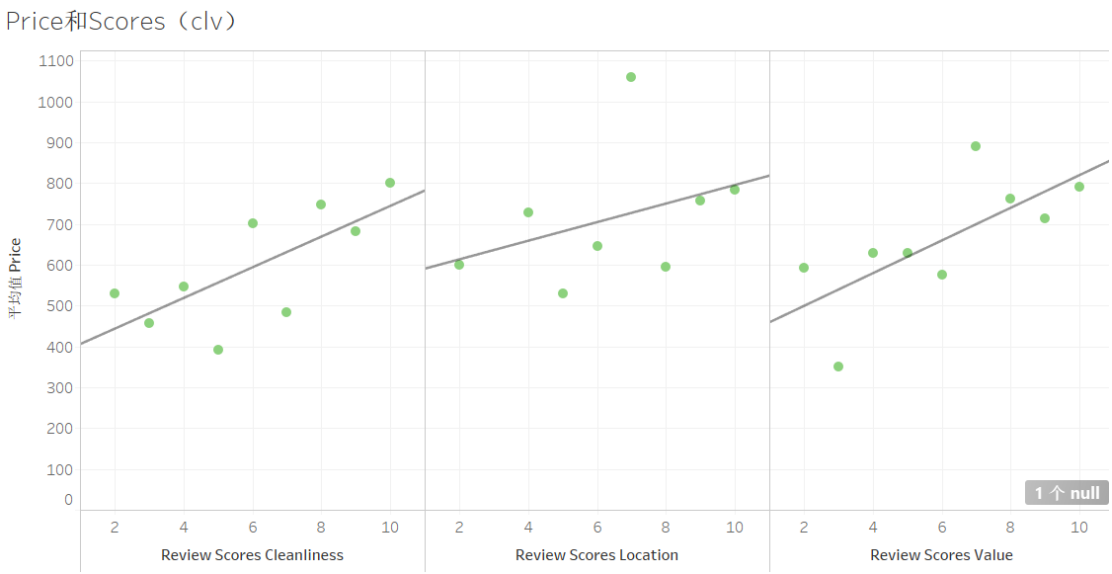

图 21 price 和 scores（clv）

相对而言，区位条件、性价比、清洁程度是针对房源的客观条件进行评价，我们从不同的分数来看，总体上房间均价与评价均分仍是正相关的关系。表明无论对与房主还是房源的特性，较好的水平都有较为高的价格均值。

3) Price 与评论数量

如图所示，Price 与评论数量几乎呈负相关关系，由逻辑推断，评论数较少的房源更多的是新房源，与老房源相比，新房源的价格会呈现偏高的情况。

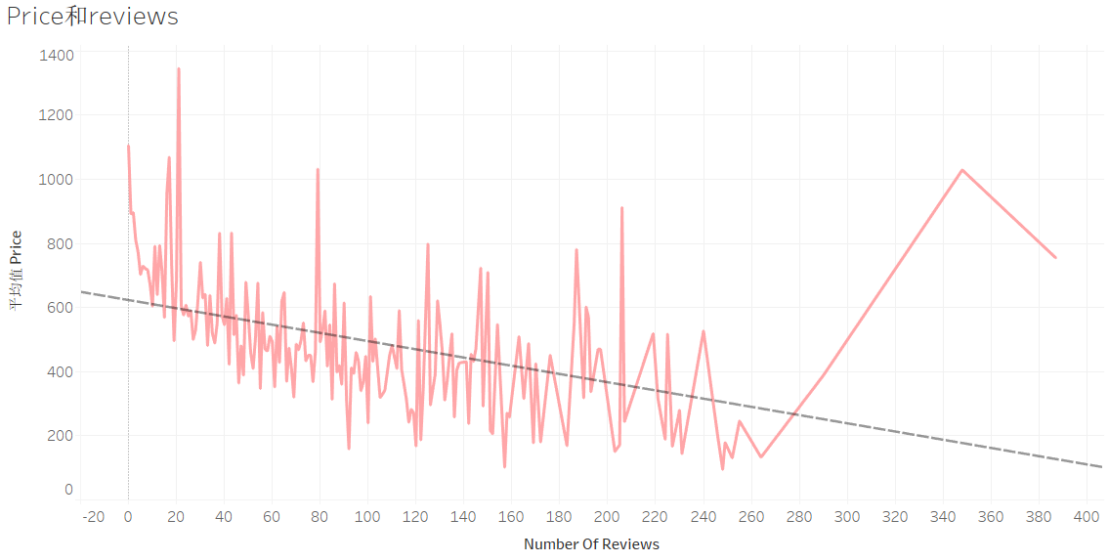

图 22 price 和 reviews

3.3.1.4. Price 与回复率

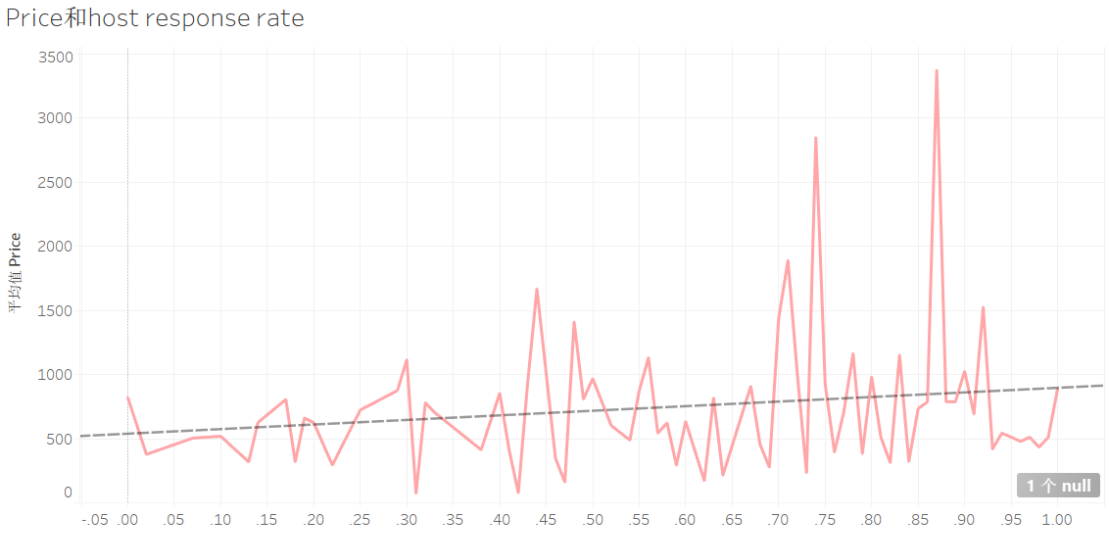

图 23 price 和 host response rate

从房东的回复率来看，回复率与房源价格有一个较弱的正相关，但其趋势不是非常明显。表明房源的价格与房东的回复率存在一定关系，但并不是决定其价格的一种因素，推理其对价格的影响应该为某种相关关系，而不是因果关系。

3.3.1.5. Price 与房东发布数量

如图所示，Price 与房东发布数量没有明显的相关关系，趋势线较为平缓，可以看到价格随着 hostlistingscount 的增大，其变化幅度较大，相关性较小。

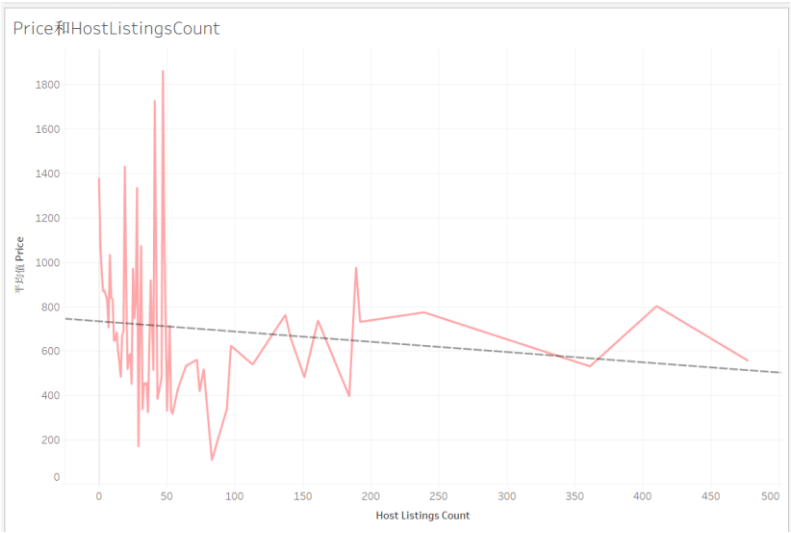

图 24 price 和 hostlistingcount

3.3.2. 相关性分析：以价格、评分为主

在初步的描述性统计探索之后，我们对于 Price 与各个变量的相关系数进行计算，可以看到价格与几种不同的评论指标系数有一定的相关性，但相关性并不是很强，推测是由于评分较为集中于 8-10 分的原因所致。

```
review_scores_rating            0.022953
review_scores_accuracy          0.027359
review_scores_cleanliness       0.025019
review_scores_checkin           0.026943
review_scores_communication     0.012642
review_scores_location          0.014851
review_scores_value             0.016444
Name: price, dtype: float64
```

图 25 Price 与评分变量的相关系数

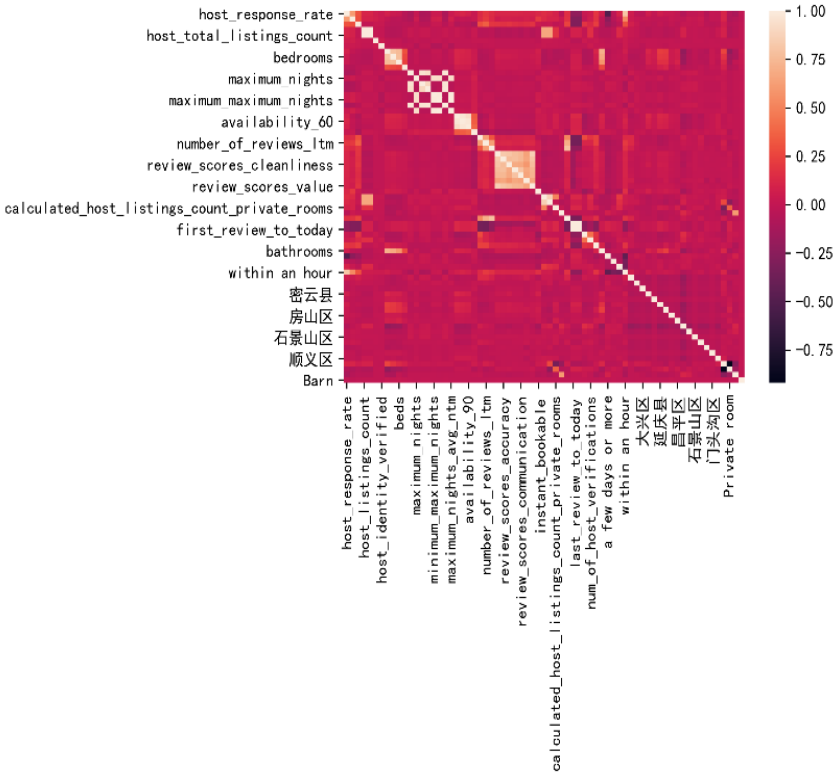

图 26 Price 与各个变量的相关系数图

进一步，我们可以从图看出，价格的主要影响因素是区位和大小——其与区位相关的变量和房间大小相关的变量均有较为明显的**相关关系**，这表明区位与房间规模大小是影响价格的主要因素。

以下的具体相关性分析，也同样得出相似的结论——价格的主要影响因素是关于区位与房间大小的因素，影响程度以相关系数度量。具体结果如下。

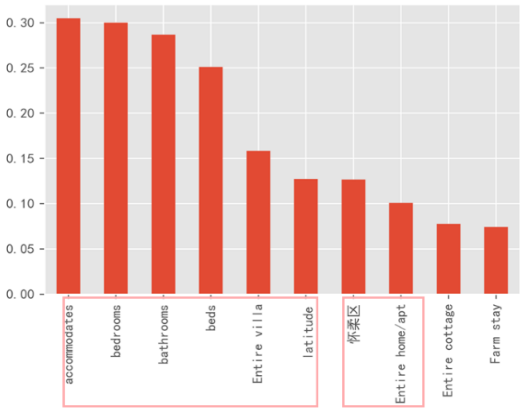

图 28 和价格（price）相关系数最高的 10 个变量

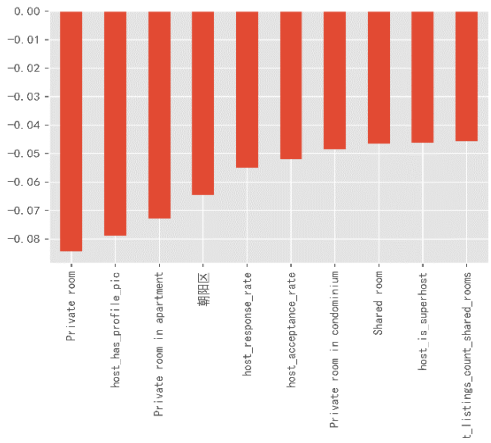

图 27 和价格（price）相关系数最低的 10 个变量

在对价格作为因变量分析之后，我们考虑，其他变量是否会与总评分有较大的相关关系？我们以相关系数为度量依据，得出与总评分相关系数最高的 10 个变量与相关系数最低的 10 个变量。

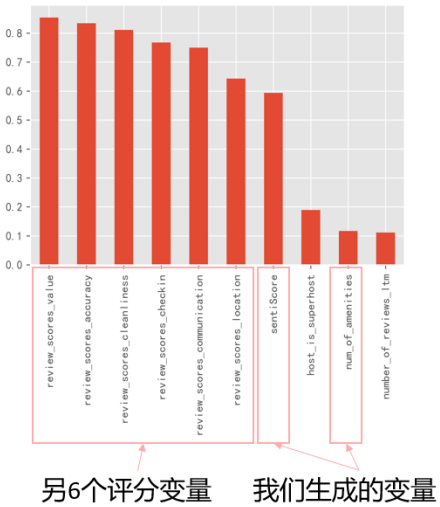

图 30 和总评分相关系数最高的 10 个变量

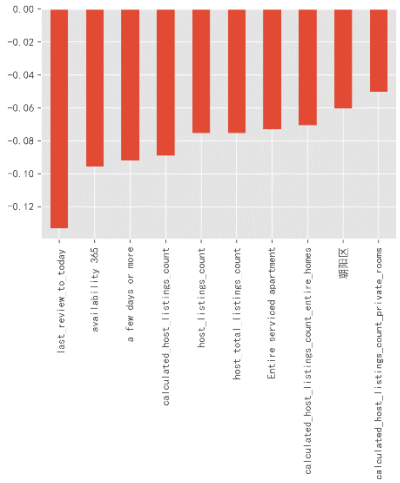

图 29 和总评分相关系数最低的 10 个变量

经过相关性分析，我们主要有以下发现：

（1） 评分和价格**没有显著的相关关系**；

（2） 评分最重要的影响因素是**房屋大小**——容纳人数（accommodates）、卧室厕所数量等，和**区位**——我们发现维度、所处地区对房屋价格有较强的影响；

（3） 几个评分指标之间相关性较大，使用 Snownlp 给出的情感分数（sentiScore）也和几个评论变量有较强的相关关系；此外，房东是否是“超级房东”（is_host_superhost）也对评分有正向的影响。

3.4. 价格预测模型建立

本文提出的房源推荐模式是基于对价格预测的机器学习模型，故本部分将对价格预测模型的建立、调参、检验等过程进行较为细致的阐析。

3.4.1. 无监督学习——聚类 Agglomerative Clustering

在对 Airbnb 房价建立回归模型之前，我们考虑到对于不同类型的房源来说，影响出租价格的因素可能不尽相同，故首先对所有样本进行了聚类。

在选择变量方面，在聚类时剔除了与**客户评分相关的变量**（如 *review_scores_rating*、*reviews_per_month* 等等），其主要原因有二：客户评分并不是房源固有的属性；同时，客户评分也将会用于**检验房源推荐模式**（并不会用在价格预测模型中）。除去客户评分相关的变量后，聚类时使用的变量包括两大类：一类是**房子本身的特征**，如 *bathrooms*、*latitude* 等，另一类是**房东的特征**，如 *host_response_rate*、*host_is_superhost* 等。

在对类别变量生成**虚拟变量**及对连续变量**归一化**处理后，本文采用了 Agglomerative clustering 的聚类方法：它是一种自下向上的层次聚类（a hierarchical clustering using a bottom up approach），其聚类结果较为稳定。

在对类别数目进行多次尝试和分析类别的经济意义后，我们最终把类别数目（n_clusters）设定为 3，并选取类别之间差异显著的部分变量及其均值列表如下。

表 4 聚类结果——部分变量均值列表

聚类	0	1	2	总体
样本个数	10155	7622	3320	21097
host_time	892.6595766	984.8272107	1013.499699	944.9746409
num_of_host_verifications	4.167306745	4.409866177	4.493373494	4.306252074
num_of_amenities	18.97744953	19.70952506	22.4939759	19.79532635
host_response_rate	0.871651403	0.87086329	0.975746988	0.887748021
host_acceptance_rate	0.902534712	0.909674626	0.986677711	0.91835569
host_listings_count	16.36405711	9.88913671	12.01957831	13.34109115
accommodates	4.746725751	2.919050118	4.50060241	4.047684505
bathrooms	1.532102413	1.248294411	1.437650602	1.414703512
bedrooms	2.033973412	1.3749672	1.861445783	1.768734891
beds	2.720827179	1.88743112	2.475301205	2.381096838
price	979.9955687	573.5011808	1012.716265	838.2850168
标签	**豪华**	**简易**	**豪华+热情**	-

从表格中我们可以看出，聚类 1 相对于其它两类来说价格较低，房间数、容纳人数（属于**房子本身的特征**）等也相对较少，算是“简易型”，而聚类 0 和聚类 2 则是“豪华型”；后二者的差异主要在于**房东的特征**，聚类 2 的房东从回复率等变量来看确实要比聚类 1 的房东“热情”不少。

3.4.2. 有监督学习

在聚类完成后，本部分将阐释建模预测价格的过程与结果。

我们首先使用了最简单的有监督学习算法——线性回归，对因变量（价格）与自变量的线性关系进行发掘；随后，我们使用了支持向量机、决策树（及基于树的集成算法）、神经网络等算法进一步探究变量之间的内在关联，并通过模型调参提高模型预测价格的准确率。

3.4.2.1. 线性回归 Linear Regressions

本文对 3 个类别的房源各进行了 4 种不同方式的线性回归，分别是最小二乘法（Ordinary Least Squares）、标准化后的最小二乘法、标准化后的 Ridge 回归、标准化后的 Lasso 回归，它们的拟合优度R^2如下表所示。

表 5 线性回归拟合优度表

拟合优度R^2	聚类 0	聚类 1	聚类 2
OLS	0.299	0.228	0.241
Standardlized OLS	0.299	0.228	0.241
Standardlized Ridge	0.294911997	0.217446651	0.189467894
Standardlized Lasso	0.296083787	0.221252865	0.224259515

其中，Ridge 和 Lasso 回归的惩罚项系数是由 5 折交叉验证（cross-validation）得出的最佳参数，具体如下。

表 6 Ridge 和 Lasso 回归的惩罚项系数表

Parameters	聚类 0	聚类 1	聚类 2
Ridge’s Alpha	2616	2765	5525
Lasso’s Alpha	0.006417745	0.009440305	0.020986613

拟合优度表明，单就**线性关系**而言，房源价格与现有的自变量之间关系较弱（R^2介于 0.2 和 0.3 之间）；3 个类别分开来看，则聚类 0 的因变量与自变量之间有相对较强的关系，聚类 2 次之，而聚类 1 的线性关系则最不理想，一定程度上说明简易型房源的价格影响机理相对“隐蔽”。

为了令变量之间的斜率系数更有可比性，下面着重关注**标准化**后的线性回归系数（使用标准化而不是归一化的目的则是让斜率系数保持**经济意义**，即该自变量变化 1 个标准差时以标准差衡量因变量的变化）。

对于聚类 0 而言，选取斜率系数绝对值较大（即经济意义显著）的变量如下。

表 7 聚类 0 经济意义显著的变量

Variables	标准化 OLS	标准化 RIDGE	标准化 LASSO
latitude	0.1646	0.042066	0.105331
accommodates	0.1662	0.125413	0.16614
bathrooms	0.1473	0.123799	0.145865
bedrooms	0.0732	0.093386	0.069654
beds	0.104	0.098185	0.101157
host_response_time_within an hour	-0.0392	-0.030519	-0.062032
property_type_Entire villa	0.0808	0.076161	0.083857
property_type_Private room in cottage	-0.0408	-0.029713	-0.067699
neighbourhood_cleansed_密云县 / Miyun	-0.1615	-0.037026	-0.089394

accommodates、*bathrooms*、*bedrooms*、*beds* 等 4 个变量与价格之间具有正向关系符合我们的常识；而令人费解的是，*latitude* 变量系数为正，说明对于聚类 0 来说，纬度越大（即越往北），Airbnb 北京的房源出租价格往往会越高。

用同样的方式整理聚类 1 和聚类 2 斜率系数绝对值较大（即经济意义显著）的变量得到下面的表格。

表 8 聚类 1 经济意义显著的变量

Variables	标准化 OLS	标准化 RIDGE	标准化 LASSO
accommodates	0.2292	0.148309	0.223378
bathrooms	0.1358	0.111065	0.125893
bedrooms	0.0496	0.063325	0.038197
property_type_Castle	0.1085	0.080332	0.098
property_type_Private room in farm stay	-0.0997	-0.066217	-0.07991
property_type_Room in boutique hotel	0.0926	0.068097	0.086588
property_type_Room in heritage hotel	-0.0406	-0.027642	-0.030689

Variables	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Shared room in boutique hotel	0.0695	0.051007	0.058699

表 9 聚类 2 经济意义显著的变量

Variables	标准化 OLS	标准化 RIDGE	标准化 LASSO
latitude	0.1238	0.035842	0.027057
host_listings_count	0.1102	0.022512	0.025578
accommodates	0.0628	0.043056	0.055895
bathrooms	0.0839	0.047305	0.086396
bedrooms	0.0709	0.040844	0.050265
maximum_nights	-0.0822	-0.032517	-0.065578
property_type_Entire villa	0.0309	0.024659	0.034395
property_type_Farm stay	0.2527	0.102124	0.246021
neighbourhood_cleansed_怀柔区 / Huairou	0.0757	0.071318	0.119359

从上面 3 个表格可以看出，无论对于何种类别的房源来说，*accommodates*、*bathrooms*、*bedrooms* 对出租价格都有明显的**正向线性关系**；对于豪华的房源（聚类 0 和聚类 2）来说，其纬度越高可能意味着出租价格越高；而某些特定的虚拟变量——房源类型（如 *Farm stay*、*Entire villa* 等）或是所在城区（如密云县、怀柔县等）也与价格有着较强的相关性。

完整的线性回归系数表详见附录。

3.4.2.2. 支持向量机 Support Vector Machines

支持向量机是一种传统的分类器算法，其算法简明且结果稳定；机器学习库——Scikit Learn 中将传统的分类器算法拓展成为了回归器算法，本文即采用此算法。

对变量标准化处理后，经过 5 折交叉验证调优的参数及拟合优度R^2如下。

表 10 支持向量机 Support Vector Machines 参数及拟合优度

Parameters	参数含义	聚类 0	聚类 1	聚类 2
C	Regularization parameter	4.6	3.8	7.9
gamma	Kernel coefficient	0.0056	0.0089	0.0115
kernel	Kernel type	rbf	rbf	rbf
5 折验证集平均得分	-	**0.377915**	**0.280287**	**0.28435**

交叉验证“不约而同”地为 3 个聚类选择了 rbf 作为核函数，但同时我们也可以发现，支持向量机在聚类 0 上的拟合优度要远远优于其在其它两类中的表现。

将支持向量按对偶问题系数（即权重）加权求和得到**分界面法向量**在各维度的系数（完整统计表详见附录），发现部分维度对于每一个类别来说都有明显异于 0 的系数，选取部分维度列举如下。

表 11 支持向量机 Support Vector Machines 部分维度系数

Variable	聚类 0	聚类 1	聚类 2
accommodates	228.2025771	286.4187229	91.97115144
bathrooms	145.7123066	183.3346458	75.28389271
bedrooms	189.1019799	138.1861846	120.7075062
beds	115.7060003	125.1752108	96.07803196
minimum_nights	-37.7847189	-36.37955819	-20.43770563
maximum_nights	-16.55736859	-12.61845776	-41.69245346
number_of_reviews	-50.03063399	-50.20028422	-37.42279689
number_of_reviews_ltm	-23.21345686	-67.4204772	-36.80294621

不难发现，*accommodates*、*bathrooms*、*bedrooms*、*beds* 等 4 个维度上的系数为正且数值较大，与上文线性回归中的结果相仿；而房源出租最小/大晚数和评论数维度上的系数为负，但未能找到合适的理论诠释这一现象。

3.4.2.3. 决策树及基于树的集成算法

本部分包括下述 4 种树及基于树的算法。

表 12 决策树及基于树的集成算法

非集成算法 Non-Ensemble Methods	集成算法 Ensemble Methods	
	Averaging Methods	**Boosting Methods**
决策树 Decision Tree	随机森林 Random Forests	梯度提升树 XGBoost
	极端随机树 Extremely Randomized Trees	

1) 决策树 Decision Tree

单棵决策树形式的回归器可以看作一个分段常数近似（a piecewise constant approximation）回归，虽然其稳定性和准确性都不如集成算法，但它们之间的调参思路相近，对单棵决策树的调优能给下文集成算法的调参带来灵感与启发。

经过 5 折交叉验证调优的参数及拟合优度R^2如下所示。

表 13 决策树 Decision Tree 参数及拟合优度

Parameters	参数含义	聚类 0	聚类 1	聚类 2
criterion	Function to measure the quality of a split	mae	mae	mse
max_depth	Maximum depth of the tree	18	105	72
max_features	Number of features to consider when looking for the best split	38	112	140
min_samples_leaf	Minimum number of samples required to be at a leaf node	16	12	14
min_samples_split	Minimum number of samples required to split an internal node	53	5	8
splitter	Strategy used to choose the split at each node	best	best	random
5 折验证集平均得分	-	**0.341784**	**0.277985**	**0.35209**

在评价标准 criterion 上，聚类 0 和聚类 1 选择了 mae 而聚类 2 选择了 mse（下文可以看到，极端随机树在交叉验证调优时也选择了同样的 criterion），mae

的运算速度虽然远远慢于 mse，但能减少异常值对结果的影响；所以，不妨合理猜测，聚类 0 和聚类 1 的样本分布不如聚类 2“均匀”。

在其它参数上，聚类 0 相比其它两类选择了更为“残忍”的防止过拟合参数——更少的深度和变量数、更大的叶子和分支所需样本数。

此外，决策树在聚类 0 和聚类 2（豪华房源）上表现明显优于聚类 1（简易房源）。

输出各变量的相对重要性数据后，我们选择各类重要性比较高的变量如下。

表 14 决策树 Decision Tree 相对重要性较高的变量列表

	Variables	**Importance**
聚类 0	*bedrooms*	26.62%
	latitude	22.70%
	accommodates	20.05%
聚类 1	*accommodates*	16.38%
	bathrooms	13.33%
聚类 2	*maximum_nights*	56.51%
	property_type_Farm stay	19.90%

在上表的结果中，最令我们感到意外的是决策树认为聚类 2 中能对预测价格起最关键作用的变量是 *maximum_nights* 和一个虚拟变量——*property_type_Farm stay*。虽然这不免让人开始思考结果的准确性，但下文极端随机树的结果却恰恰印证了这样匪夷所思的结果。

与此相对，在情理之中的是 *bedrooms*、*accommodates* 和 *bathrooms* 成为聚类 0 和聚类 1 相对重要的变量；而 *latitude* 作为聚类 0 第二重要的变量则恰恰说明 *latitude* 在线性回归的正系数或许确有其“道理”所在。

完整的相对重要性统计表详见附录。

2) 随机森林 Random Forests

随机森林是以决策树为基学习器、使用 Bagging 抽样方法的一种集成学习方法，它是决策树的集合，被用来解决决策树泛化能力弱的特点。随机森林的优点在于不容易过拟合，训练可以高度并行，能够给出每个特征的重要性，对部分数据缺失不敏感。

经过 5 折交叉验证调优的参数及拟合优度R^2如下所示。

表 15 随机森林 Random Forest 参数及拟合优度

Parameters	**参数含义**	**聚类 0**	**聚类 1**	**聚类 2**
criterion	Function to measure the quality of a split	mae	mse	mse
max_depth	Maximum depth of the tree	25	20	22
max_features	Number of features to consider when looking for the best split	18	25	16
min_samples_split	Minimum number of samples required to split an internal node	2	2	3
min_samples_leaf	Minimum number of samples required to be at a leaf node	2	2	2
n_estimators	Number of trees in the forest	600	700	100
5 折验证集平均得分	-	**0.404117**	**0.266969**	**0.319239**

在评价标准 criterion 上，聚类 1 和聚类 2 选择了 mse 而聚类 0 选择了 mae。而这三类在最大深度 max_depth 上均倾向于选择深度在 20 左右，与决策树的结果不同，聚类 1 和聚类 2 选择的深度比聚类 0 的小。

在最大特征数量 max_features 这一特征表现上，与决策树截然不同的是，三个聚类在随机森林中选择的最大特征数在 15-25 的范围内，显著小于决策树的结果。

在 n_estimators 的选择上，聚类 2 随机森林中树的个数为 100，明显小于聚类 0 和聚类 1 所选择的 n_estimators；同时，聚类 2 选择了 3 作为 min_samples_split 有关，当 min_samples_split 较大时（即切分节点所需的最小样本数较大），提前进行剪枝，防止构建的决策树过拟合。

经过 5 折交叉验证后验证集上的平均得分情况与单棵决策树的结果相似，随机森林在聚类 0 和聚类 2（豪华房源）上验证集平均得分表现均明显优于聚类 1

（简易房源）。

将最优模型中得到的各变量相对重要性数据输出，分别从三个聚类中挑选出相对重要性较高的前 3 个特征，结果如下表所示。

表 16 随机森林 Random Forest 相对重要性较高的变量列表

	Variables	Importance
聚类 0	*accommodates*	12.98%
	bedrooms	11.17%
	bathrooms	9.94%
聚类 1	*latitude*	19.25%
	accommodates	9.28%
	bedrooms	6.79%
聚类 2	*property_type_Farm stay*	8.43%
	latitude	8.10%
	longitude	7.30%

与单棵决策树相比，随机森林各变量的相对重要性相对均衡，没有出现个别占比极高（大于 50%）的变量——尤其在聚类 2 随机森林中，各变量相对重要性最为均衡，property_type_Farm stay 的重要性最高为 8.43%。

在聚类 0 和聚类 1 中，均出现了 *accommodates* 作为较重要的变量，并且 *accommodates* 的重要性往往高于 *bedrooms*，这说明在房子定价时比较关注房子所能容纳的人数，其次关注房间的数量。

在聚类 0（豪华房型）中，*bathrooms* 的重要性位居第三，通常浴室或卫生间会被当做房间的附加设施，当浴室或卫生间的数量较多时，房子更有可能作为豪华房型，房子的价格会更高，这一结果与实际情况相符，因此在豪华房型定价时会更多关注 *bathrooms* 的数量。

聚类 1 和聚类 2 均选择了 *latitude* 作为重要变量，其中聚类 1 中 *latitude* 的重要性显著高于 *accomodates*，说明简易房型（聚类 1）在定价时会首要考虑地理

位置，在北京市这种地价昂贵的地区，地理位置是决定房价的重要因素。

聚类 2（豪华房型+房东热情）选择 *property_type_Farm stay* 作为该类房型定价中最重要的变量特征——*property_type_Farm stay* 即一类类似于农场住宿的房型，这一结果或许与城市人们的偏好相关。

完整的相对重要性统计表详见附录。

3) 极端随机树 Extremely Randomized Trees

极端随机树与随机森林相近，都使用了候选特征的**随机子集**（a random subset of candidate features），但不是寻找最具鉴别能力的阈值（most discriminative thresholds），而是为每个候选特征**随机抽取阈值**，并从这些随机生成的阈值中选取最佳阈值（the best of these randomly-generated thresholds）作为分割规则，故其表现很可能较随机森林好。

经过 5 折交叉验证调优的参数及拟合优度R^2如下所示。

表 17 极端随机树 Extremely Randomized Trees 参数及拟合优度

Parameters	参数含义	聚类 0	聚类 1	聚类 2
criterion	Function to measure the quality of a split	mae	mae	mse
max_depth	Maximum depth of the tree	35	30	113
max_features	Number of features to consider when looking for the best split	25	20	88
min_samples_split	Minimum number of samples required to split an internal node	7	5	5
n_estimators	Number of trees in the forest	700	700	76
5 折验证集平均得分	-	**0.413164**	**0.295832**	**0.458936**

在评价标准 criterion 上，聚类 0 和聚类 1 选择了 mae 而聚类 2 选择了 mse，与决策树的交叉验证结果相同；而评价标准与其它参数似乎有某种内在“联系”——选择 mae 的极端随机树更倾向于选择更小的深度、更小的变量数以及更多的树，选择 mse 的极端随机树则更倾向于选择更大的深度、更多的变量数以及更少的树。

与单棵决策树的情况类似，极端随机树在聚类 0 和聚类 2（豪华房源）上表现明显优于聚类 1（简易房源）。

输出各变量的相对重要性数据后，我们选择各类重要性比较高的变量如下。

表 18 极端随机树 Extremely Randomized Trees 相对重要性较高的变量列表

	Variables	Importance
聚类 0	*accommodates*	15.27%
	bathrooms	10.17%
	bedrooms	6.73%
聚类 1	*accommodates*	8.59%
	bathrooms	4.63%
	latitude	4.29%
聚类 2	*minimum_nights*	23.66%
	maximum_nights	19.77%
	property_type_Farm stay	12.27%

与单棵决策树相比，极端随机树中各变量相对重要性结果相对均衡，没有出现个别占比极高（大于 50%）的变量，这表明集成算法在变量过滤方式上更为**稳健（robust）**。

在聚类 0 和聚类 1 中，极端随机树选择了 *accommodates*、*bathrooms* 和 *bedrooms* 作为最重要的变量，而 *latitude* 则在聚类 1 中位列第三，上述几个模型用数据说明了一个事实——纬度在 Airbnb 北京的房源价格中确实扮演着不可或缺的角色（虽然二者之间很可能不仅仅是线性关系）。

较为“扑朔迷离”的依然是聚类 2，*minimum_nights* 和 *maximum_nights* 对价格预测起着关键作用，而虚拟变量 *property_type_Farm stay* 的作用也不容小觑——但相信读者对此已不觉惊奇了，毕竟上文决策树的结果与此吻合。

完整的相对重要性统计表详见附录。

4) 梯度提升树 XGBoost

与普通的梯度下降算法相比，新兴的 XGBoost 是一种更灵活、功能更强大的算法，且由于其能够使用**并行计算**，大大提高了模型拟合的效率。

经过 5 折交叉验证调优的参数及拟合优度R^2如下所示。

表 19 梯度提升树 XGBoost 参数及拟合优度

Parameters	参数含义	聚类 0	聚类 1	聚类 2
eval_metric	Evaluation metrics for validation data	rmse	rmse	rmse
learning_rate	Step size shrinkage used in update	0.1	0.1	0.1
max_depth	Maximum depth of a tree	3	5	6
n_estimators	Number of rounds for boosting	80	145	650
reg_alpha	L1 regularization term on weights	6.4	0	0.1
reg_lambda	L2 regularization term on weights	2.2	1.5	1.2
subsample	Subsample ratio of the training instances	0.9	1	0.6
tree_method	Tree construction algorithm	exact	gpu_hist	gpu_hist
5 折验证集平均得分	-	**0.370026**	**0.345543**	**0.354294**

在 3 个类中，梯度提升树均选择了 rmse 作为评估标准，与决策树和极端随机树有明显不同。

在构建树的方法上，交叉验证为聚类 0 选择的方法为 exact（即贪婪算法），而为聚类 1 和聚类 2 选择了 gpu_hist（即使用 GPU 的近似算法）；而构建树的方法与其它参数也似乎有某种内在“联系”——使用 exact 的梯度提升树需要更小的深度、更小的迭代次数以及更大的惩罚项系数来防止过拟合，而使用 gpu_hist 的梯度提升树则不需要那么“残酷”的约束。

虽然梯度提升树在聚类 0 和聚类 2（豪华房源）上表现优于聚类 1（简易房源），但三类之间的差距相比上文决策树和极端随机树的结果大大缩小。

输出各变量的相对重要性数据后，我们选择各类重要性比较高的变量如下。

表 20 梯度提升树 XGBoost 相对重要性较高的变量列表

	Variables	Importance
聚类 0	*accommodates*	28.37%
	num_of_host_verifications	8.80%
	bedrooms	7.60%
聚类 1	*host_acceptance_rate*	20.77%
	neighbourhood_cleansed_西城区	9.46%
聚类 2	*minimum_nights*	36.23%
	property_type_Farm stay	23.33%
	neighbourhood_cleansed_怀柔区 / Huairou	16.54%
	maximum_nights	8.53%

梯度提升树对重要变量的选择结果与决策树和极端随机树的“大相径庭”，而之前最令我们意外的聚类 2 的重要变量却得到了一贯的肯定，如 *minimum_nights*、*property_type_Farm stay* 和 *maximum_nights*，此外还引入了另一个虚拟变量——*neighbourhood_cleansed_怀柔区 / Huairou*，其中的原因很可能是上述几种算法均为聚类 2 选择了 mse（等价于 rmse）作为评估标准。

在聚类 0 中，除了 *accommodates* 和 *bedrooms* 依旧作为重要变量外，还加入了之前“名不见经传”的 *num_of_host_verifications*，其含义是房东在各社交平台的注册（或认证）数目；而聚类 1 则发生了巨大的变化，*host_acceptance_rate* 和 *neighbourhood_cleansed_西城区*成为最重要的两个变量，说明梯度提升树在预测聚类 1 的出租价格时比较看重房东的接受率（表现房东“热情”程度的变量之一）而不是房子固有的特征。

完整的相对重要性统计表详见附录。

3.4.2.4. 神经网络

神经网络由众多的神经元通过可调的权重连接而成，由输入层、隐藏层和输出层构成，在每一层之间会附加一个激活函数，通过 BP 算法或梯度下降算法来

调整每一个神经元的权重，直至模型收敛，分类的准确性一般较高。但由于神经网络需要大量参数，难以搭配出适合的参数，有时可能会因为参数选择不恰当而导致神经网络模型达不到学习的效果，从而降低模型的拟合优度。

同样分别对三个聚类经过 5 折交叉验证调优参数，参数结果及拟合优度R^2如下表所示。

表 21 神经网络 Neural Network 参数及拟合优度

Parameters	参数含义	聚类 0	聚类 1	聚类 2
hidden_layer_sizes	The number of neurons in each hidden layer	(1500,)	(1500,)	(1500,1000)
activation	Activation function for the hidden layer	logistic	logistic	logistic
solver	Solver for weight optimization	adam	adam	adam
alpha	L2 penalty (regularization term) parameter	0.0001	0.0005	0.0001
learning_rate	Learning rate schedule for weight updates	constant	constant	constant
learning_rate_init	The initial learning rate used	0.0005	0.0005	0.0005
tol	Tolerance for the optimization	1e-06	1e-06	1e-06
5 折验证集平均得分	-	**0.349416**	**0.205707**	**0.045202**

由交叉验证结果可知，三个聚类在 activation、solver、learning_rate 这三个重要参数的选择上均表现一致——选择了 logistic（即 logistic-sigmoid 函数）作为激活函数，adam 算法（即随机梯度优化）作为神经元权重的优化的求解器，constant（即采用恒定的学习率）作为学习率对数据进行学习并拟合模型。

其中，聚类 0 的数据在模型拟合效果上表现最好；而聚类 2 在神经网络的表现最差，这一拟合优度的结果可能是由于人工难以选择到最适合的神经元数量和层数，导致神经网络无法达到学习的目的。

3.4.3. 模型比较

不妨将几个价格预测模型在验证集上的平均得分列表如下。

表 22 价格预测模型 5 折验证集平均得分表

模型	聚类 0	聚类 1	聚类 2
Support Vector Machines	0.377915	0.280287	0.28435
Decision Tree	0.341784	0.277985	0.35209
Random Forest	0.404117	0.266969	0.319239
Extremely Randomized Trees	0.413164	0.295832	0.458936
XGBoost	0.370026	0.345543	0.354294
Neural Network	0.349416	0.205707	0.045202

在对比多个模型的拟合优度之前，我们先不妨假设各模型的参数已经尽可能地调优。

从表格中看出，Support Vector Machines 虽然是比较传统的算法，但其在聚类 0 的拟合程度丝毫不逊色于 Decision Tree 甚至新兴的 XGBoost 算法；而两种基于树的集成算法（Extremely Randomized Trees 和 XGBoost）相比 Decision Tree 均有不同程度的提升，其中 Extremely Randomized Trees 在聚类 0 和聚类 2（豪华房源）提升较多，而 XGBoost 在聚类 1（简易房源）提升较多。

最终选择 Extremely Randomized Trees 作为聚类 0 和聚类 2 预测价格的算法，选择 XGBoost 作为聚类 1 预测价格的算法——价格预测模型在 3 个类的验证集上平均得分依次为：0.413164、0.345543 和 0.458936。

3.5. 推荐模式效果验证

定义房源出租的实际价格为 *Price*，机器学习模型预测的价格为 *Predict*，则推荐比为 *Predict / Price*。

根据房源 ID 匹配与客户评分有关的 7 个变量：*review_scores_rating*、*review_scores_accuracy* 、 *review_scores_cleanliness* 、 *review_scores_checkin* 、*review_scores_communication*、*review_scores_location*、*review_scores_value*，并作出 *Predict / Price* 与它们的相关系数矩阵（表格中的变量名作适当的删减）。

表 23 聚类 0 推荐比与客户评分相关系数矩阵

	Predict / Price	*rating*	*accuracy*	*cleanliness*	*checkin*	*communication*	*location*	*value*
Predict / Price	1.0000							
rating	-0.0189	1.0000						
accuracy	-0.0220	0.8462	1.0000					
cleanliness	-0.0130	0.8248	0.7695	1.0000				
checkin	-0.0220	0.7573	0.7676	0.6807	1.0000			
communication	-0.0087	0.7453	0.7287	0.6643	0.7959	1.0000		
location	-0.0423	0.6478	0.6596	0.5776	0.6684	0.6386	1.0000	
value	-0.0009	0.8616	0.8186	0.7832	0.7415	0.7164	0.6610	1.0000

在聚类 0 中，我们很失望地发现，每一个客户评分变量与 *Predict / Price* 的相关性都是负的；但在做了 *Predict / Price* 对 7 个客户评分变量的线性回归后却收到了符合预期的结果。

表 24 聚类 0 推荐比对客户评分线性回归系数

	Coefficients	**标准误差**	**t Stat**	**P-value**
Intercept	1.133199215	0.035682986	31.75740966	2.1709E-204
rating	-0.000868535	0.000641352	-1.354223693	0.17571715
accuracy	-0.00603557	0.006039884	-0.999285639	0.317697531
cleanliness	6.7226E-05	0.005045282	0.01332452	0.989369339
checkin	-0.005267389	0.006029133	-0.873656212	0.382341243
communication	0.009256917	0.00582623	1.588834899	0.112151527
location	-0.016044214	0.004581715	-3.50179217	0.000465579
value	0.018274592	0.00560746	3.258978531	0.001124509

在上表中，*review_scores_value*（代表客户对房源性价比的评分）的斜率显著

为正，即性价比评分越高，该房源的 *Predict / Price* 往往越高——客户的评分印证了房源推荐模式的有效性。

表 25 聚类 1 推荐比与客户评分相关系数矩阵

	Predict / Price	*rating*	*accuracy*	*cleanliness*	*checkin*	*communication*	*location*	*value*
Predict / Price	1.0000							
rating	-0.0022	1.0000						
accuracy	-0.0018	0.8052	1.0000					
cleanliness	-0.0088	0.7949	0.7517	1.0000				
checkin	-0.0051	0.7639	0.7779	0.6960	1.0000			
communication	0.0017	0.7440	0.7342	0.6610	0.7642	1.0000		
location	-0.0070	0.6389	0.6791	0.6092	0.6702	0.6747	1.0000	
value	0.0012	0.8367	0.7843	0.7474	0.7212	0.7150	0.6606	1.0000

表 26 聚类 2 推荐比与客户评分相关系数矩阵

	Predict / Price	*rating*	*accuracy*	*cleanliness*	*checkin*	*communication*	*location*	*value*
Predict / Price	1.0000							
rating	-0.0017	1.0000						
accuracy	-0.0021	0.7655	1.0000					
cleanliness	-0.0114	0.7051	0.7023	1.0000				
checkin	0.0018	0.6968	0.6721	0.5413	1.0000			
communication	0.0036	0.7270	0.7701	0.5928	0.7468	1.0000		
location	-0.0022	0.6147	0.5606	0.4803	0.5706	0.5722	1.0000	
value	0.0112	0.8068	0.6875	0.6276	0.6081	0.6509	0.5983	1.0000

在聚类 1 和聚类 2 中，*review_scores_value*（代表客户对房源性价比的评分）

与 *Predict / Price* 呈现正相关性（尽管比较“微弱”），能在一定程度上验证了房源推荐模式。

4. 总结与展望

4.1. 推荐模式价值与建议

在提出问题之际，我们提到推荐**性价比高**的房源在首页，让用户能够更加容易发现性价比高的房源，筛选优质房源不仅能够对于房东和房客而言是一件好事，对于平台来说也非常重要。识别优质的房源能够提升客户的满意度，增强客户黏性，对于长久经营来说是一个重中之重的方面。

推荐是不是仅仅推最便宜的房间呢？我们认为，在同样的房间质量（如设施、大小、区位等接近）的房源，其价格波动不应该太大，因此，这里更加注重对于**房源本身的性价比**进行识别。而同样，即使房东对于推优的规则完全了然于心，房东会因为成本及竞争的考虑，对于价格在合理范围内进行调整，不太可能出现过度的不理性行为。

其次，推优在短租平台真正的利用中，需要结合**更多的信息**进行建模，且大数据时代的用户定制化方案越来越普遍，在本文，我们仅以对于所能掌握的变量进行利用，在此基础上进行简单的模型演示与应用。在真正的企业运营中，考虑到的因素会更多，面临的也是更加复杂的环境，如何提高平台实用度，增强用户粘性，对于短租平台来说“性价比”是一个较为永恒的话题。

4.2. 模型应用

借鉴《投资学》中“**公平定价**”的概念，我们定义**推优比=公平定价与实际定价之比**，即上文算出的 *Predict / Price*。

公平定价原本是指在竞争和有效的市场中，市场汇聚大量投资者的信息，并且这些信息反映在证券的价格上，这一观念是投资者竞争的自然结果。

如果有信息表明，购买股票会得到正的净现值，那么获得这一信息的投资者会选择购买股票，他们的购买意图会驱动股价上升；按照同样的逻辑，投资者若得到信息表明卖出股票会得到正的净现值，他们将愿意卖出股票，从而引起股价下跌。

投资者之间的竞争会消除所有净现值为正的交易机会。这意味着给定投资者

知晓所有信息，基于证券的未来现金流，证券将被**公平定价**。

国内短租房市场竞争激烈。途家、蚂蚁短租等在线民宿预订平台是首批开拓者，2016 年 Airbnb 作为老牌劲旅强势崛起后行业经历了野蛮生长的蜜月期。如今，在资本、市场和技术的多方面因素助力下，中国的在线民宿行业竞争者甚多。途家以海内外 230 万套的房源排名第二，从 2017 年开始，先后并购蚂蚁短租、携程民宿和去哪儿民宿，去年又收购大鱼自助游，帮助其房源储备迅速达到现在的高度。排位第三的木鸟民宿，其海内外房源数量为 90 万套，成立 7 年的木鸟，深耕市场，房源数量也是逐步积累起来的。成立于 2017 年的榛果民宿，背靠美团的流量，从起初的 15 万套房源也发展到今天的 50 万套，房源扩张速度惊人。

从**竞争者数量**上来看，我们有理由假设短租房市场满足**有效市场**的概念，因此我们假定：

如果推优比大于 1：价格被低估，优先进行推荐；

推优比等于 1：价格满足假定的“公平定价”；

推优比小于 1：价格被高估，不优先推荐。

Airbnb 作为全球最大的民宿短租公寓预订平台，其平台基于背靠旅行、基于信任的共享经济，构建了从用户、房主抽取服务费的商业模式。

图 31 爱彼迎商业模式示意图

由于平台需要支付房东租金，也需要根据租金价格来收取手续费，价格在爱彼迎的短租平台中毋庸置疑起着至关重要的作用。

受疫情影响，公司 2020 年业绩波动，公司毛预订价值的增长率不断下跌，经营活动提供的现金流更是由正转负。公司如想在长远发展中提升市占率，需要平台创新驱动，持续**创新平台以优化房东和房客的体验**，使平台更加容易访问来

获取更多的房东与房客，同时提升**现有用户的忠诚度和社区参与度**。

而平台改进的核心要义就是平台应如何调整策略，**促进房源的“公平”定价**。

根据上述分析的得到的回归模型，我们可以改进**定价策略**，为平台、用户和房主提供良好的改进体验。

我们把之前选择的模型（Extremely Randomized Trees 作为聚类 0 和聚类 2 预测价格的算法，选择 XGBoost 作为聚类 1 预测价格的算法）和经过计算得出的 *Predict / Price* 设计成为一个能做到个性化推荐房屋的线上小产品——“租房指南”小程序。当大家在线上租房平台浏览房屋信息的时候，输入自己的所期望的租金水平、房屋的基本特征、房屋的地理信息和房屋的附加属性等信息，将根据需求利用已有房源做出各级推荐，这样既能帮助房客快速找到符合自己期望的房屋，同时还能了解自己所期望的房屋在市场上的交易情况。

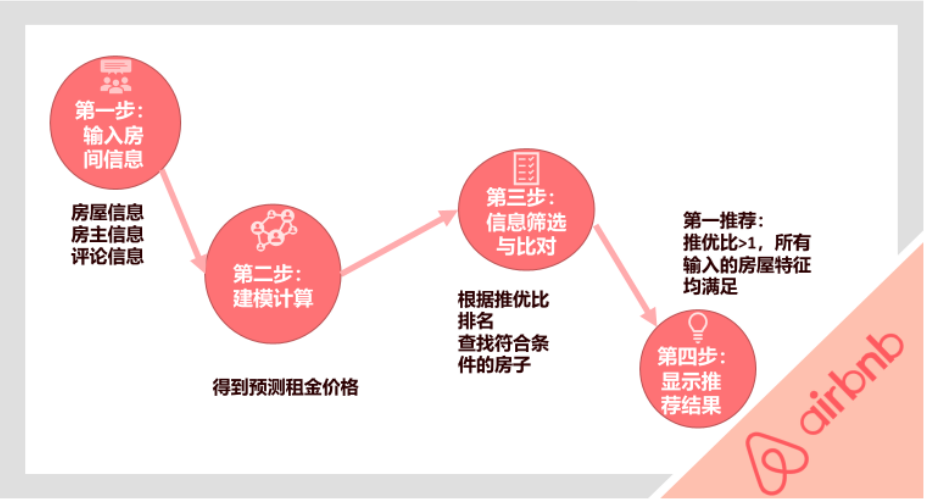

图 32 “租房指南”小程序运行过程

通过该应用，一方面用户可以拥有更加良好的用户体验，根据优先度排序选择心仪的“**性价比**之王”；另一方面，平台方可以实现对于房屋的**精准定价**，从而准确计算预估的支付房东租金，也可以根据预测的租金价格来计算手续费，达到对**全年营收及费用的预估**；房东也能够根据上传的信息在短租房中获得“公平定价”，实现房屋**出租利润最大化**。

4.3. 不足与展望

4.3.1. 数据选取

我们的数据选取了北京爱彼迎官网上的相关信息，包括 74 个变量，27439 个房源信息，变量可大致分为三种类型：**房间信息**——如房间设施、地理位置、租用天数限制等；**房主信息**——回复次数、房源数量等；**评论信息**——评分、评论数等。

但是由于租金的影响因素不胜枚举，还有很多因素可以影响房屋的定价水平，房源地段和退订规则等有值得深入挖掘的地方。比如**房源地段**中，我们考虑到了区位的影响因素，但是区位具体是如何影响定价策略的，小区位置（地处几环）、周边配套设施（如商圈、医院）、是否有**景点**等因素会对租金水平产生怎样的影响我们都没有进行深入细致地探究。

4.3.2. 模型构建

在模型构建部分，我们的研究包括了线性回归、决策树、支持向量机、神经网络、随机森林与 XGBoost 等模型。

但是现有的房租预测模型对于一些房屋的**预测结果不太准确**，未来可能还需要使用更多的房屋信息来拟合模型，或是利用房屋数据推出一个专门预测房屋价格的模型，将房屋预测的过程变得更加准确。

为了更深入地分析各因素对房屋租金价格的影响，建立租金价格关于基本属性、地理因素、附加属性的回归模型，使用定量化的方式更为精细地刻画各方面因素的影响作用大小，并且使用该模型来预测房屋租金。

4.3.3. 结论推广

爱彼迎的业务具有两大特色，这使得我们的结论可能并不具有普适性。

第一个是**独特的房东团体**。公司拥有超过 400 万房东提供超过 560 万各具特色的活跃房源，其中 23%的房东来自于过往租客。目前平台提供的房屋类型多样，包括木屋，农场，微型住宅、船、城堡、蒙古包、树屋、私人岛屿、灯塔和冰屋等，而北京地区的房源大多是民宿，若推行该模式需要注意到房源类型的差

异。

第二个是**全球网络**。公司在超过 220 个国家和地区提供服务，受新冠疫情影响公司仍能利用此点优势为全球用户提供国内旅行服务。国家与地区间的风俗文化使得公司的定价策略需要差异化定制。

我们的现有研究是基于北京地区数据得出的结论。若要将模型推广到其他城市，还要进一步考虑城市特有因素（如：在旅游城市是否为海景房等），考虑不同国家与文化间的风俗习惯的差异。

5. 对现有模型的研究补充

5.1. LDA

5.1.1. 调参和训练

LDA 模型的代码实现使用了 python 中的 gensim 包。我们的语料库 171854 条评论组成，每篇文档对应着一条房客的评论，进行翻译和分词后，使用 gensim 的 corpora 模块将处理后的文本生成文档-词语矩阵。

LDA 模型最核心的参数选择是话题数K的选择。我们使用 **10 折交叉验证**的方式，对每一个备选话题数进行 10 次估计，以 Blei et al. (2003)提出的困惑度(perplexity)作为衡量标准，即对 10%的训练样本外数据计算下式：

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{t=1}^{M} log\, p(\mathrm{d}_t)}{\sum_{t=1}^{M} N_t}\right\}$$

其中$p(d_t)$是文档t出现的概率，N_t是文档t的总词数。困惑度越小，就意味着样本外模型估计值生成的文档和原始文档越相近（即$p(\mathrm{d}_t)$越接近 1）。

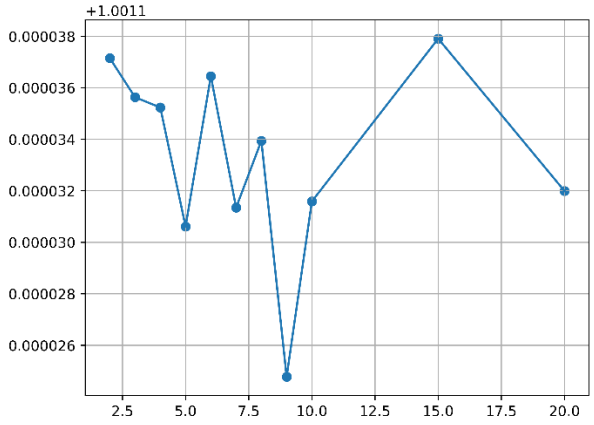

图 33 10 折交叉验证困惑度计算结果（横轴是主题数，纵轴是困惑度）

我们的结果显示最佳主题数为 9，但是为了权衡可解释性，我们选择 5 为主题数对全样本进行 LDA 建模。

5.1.2. 结果分析

5.1.2.1. LDA 主题与关键词

LDA 模型输出为文档-话题概率矩阵**Θ**和话题-词语概率矩阵 **Φ**。[2]

根据话题-词语概率矩阵 **Φ**的估计结果，我们得到 5 个主题的关键词排序。对于 5 个话题，下表给出了前 20 个概率最大的词语，从中我们可以初步看出：

Topic0 和房屋区位有关：如“位置”、“地铁站”等关键词；

Topic1 中有北京元素，可能和北京特色有关；

Topic2 和房屋设施有关：如“床”、“卫生间”等关键词；

Topic3 和正向情绪有关：如“热情”、“温馨”、“推荐”等关键词；

Topic4 和订房流程有关：如“入住”、“退房”、“接送”等关键词。

表 27 话题-词语概率矩阵

	Topic0	Prob0	Topic1	Prob1	Topic2	Prob2	Topic3	Prob3	Topic4	Prob4
0	干净	0.050546	院子	0.024922	舒服	0.022627	干净	0.044114	入住	0.026815
1	位置	0.044778	胡同	0.021449	小姐姐	0.021423	体验	0.032586	地方	0.021194
2	不错	0.034996	北京	0.01754	不错	0.018993	入住	0.028671	主人	0.013395
3	性价比	0.028738	玩	0.01705	干净	0.017115	热情	0.025405	时间	0.011784
4	高	0.023463	民宿	0.016963	床	0.014608	温馨	0.021451	很棒	0.010779
5	地铁站	0.021884	小院	0.015514	卫生间	0.013199	民宿	0.018467	提供	0.01076
6	交通	0.018122	孩子	0.014681	厨房	0.010995	真的	0.016945	房	0.009693
7	整洁	0.017973	管家	0.010818	很大	0.009867	推荐	0.016919	找	0.009643
8	设施	0.015447	朋友	0.008602	装修	0.00944	整洁	0.014828	退房	0.008633
9	地理位置	0.015115	适合	0.007061	感觉	0.009277	舒服	0.014584	接送	0.00848
10	便利	0.014886	四合院	0.007041	用品	0.008644	满意	0.014443	住宿	0.008044
11	小区	0.014098	吃	0.006957	整体	0.008446	北京	0.014256	北京	0.007965
12	齐全	0.013849	露台	0.006435	齐全	0.00832	感觉	0.013623	真的	0.007698
13	地铁	0.013179	停车	0.006433	做饭	0.007964	喜欢	0.012154	帮	0.007534
14	沟通	0.01256	喜欢	0.006419	卧室	0.007646	装修	0.011797	床单	0.007124

[2] 原始结果在 4LDA 模型代码和主要结果.zip/result_phi.csv 和 result_theta.csv 中

	Topic0	Prob0	Topic1	Prob1	Topic2	Prob2	Topic3	Prob3	Topic4	Prob4
15	环境	0.012467	感	0.006108	烧烤	0.007621	舒适	0.011785	建议	0.006311
16	舒适	0.012089	感受	0.00538	睡	0.007305	贴心	0.011662	一次性	0.006295
17	距离	0.011185	惬意	0.005358	空调	0.006564	姐姐	0.011022	预定	0.006045
18	回复	0.010341	故宫	0.004843	客厅	0.006564	老板	0.010663	找到	0.005922
19	推荐	0.009893	进	0.004708	空间	0.006371	位置	0.010435	问	0.005785

5.1.2.2. LDA 主题分数和价格、评分变量相关系数

表 28 LDA 主题分数和价格、评分变量相关系数

	price	*rating*	*value*	*accuracy*	*cleanliness*	*checkin*	*communication*	*location*
Topic0	-0.056	-0.01591	0.001573	-0.01241	-0.04094	-0.00482	-0.00033	-0.0085
Topic1	0.110752	0.042196	0.016346	0.038423	0.045498	0.032972	0.031329	0.038732
Topic2	0.02517	-0.21584	-0.20314	-0.1579	-0.17063	-0.12692	-0.11046	-0.11568
Topic3	-0.01295	0.284624	0.260715	0.239364	0.265676	0.209167	0.193764	0.188364
Topic4	-0.01396	-0.22024	-0.2035	-0.22013	-0.20025	-0.21752	-0.21939	-0.19524

仅仅通过关键词并不能让我们完整地识别出话题内涵，上表展示了 5 个话题的 LDA 分数和关键变量的关系。我们把 price 和 7 个评分变量作为重点关注对象。

从中发现:

1. Topic0 没有表现出和 price、评分的相关性；
2. Topic1 和价格正相关；
3. Topic2 和 Topic4 和评分负相关；
4. Topic3 和评分正相关，这与它的关键词识别吻合。

我们查看和每个主题相关性最高、最低的变量后识别出了**两个情绪主题**——Topic 4 和多个评分相关性负，Topic 3 和多个评分高度正相关。

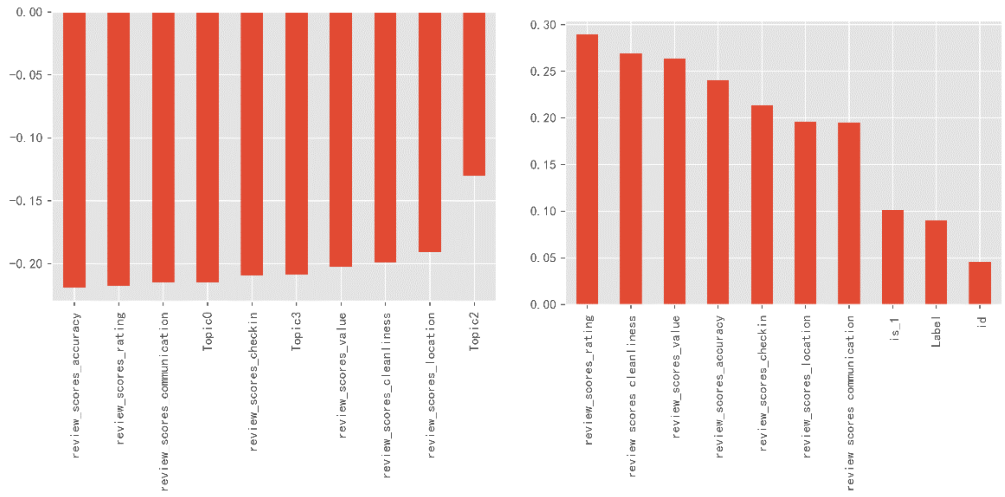

图 34 左：和 Topic 4 得分相关性最低的变量 右：和 Topic 3 得分相关性最高的变量

结合 3.1 和 3.2，我们可以将 5 个话题进行命名：

1. Topic0：区位；
2. Topic1：北京特色/情绪；
3. Topic2：设施；
4. Topic3：正向情绪；
5. Topic4：负向情绪。

5.1.2.3. 关于 LDA 主题分数和价格

接下来，我们通过 OLS 回归和 Logistic 回归来发现 LDA 得分和我们感兴趣的变量——价格、预测真实价格比、评分和聚类类别结果的关系。由于对每篇文章，主题分数的总和为 1，所以我们的自变量中去掉了在相关系数分析中表现“平庸”的 Topic0。又由于无截距项回归会使 R 方失去评估模型的能力，我们没有选择无截距项的回归。

下表展示了 LDA 主题分数和价格的 OLS 回归结果，我们发现，**主题 1-3 的得分均对价格有显著的正向结果**，其中**话题 1(北京特色/情绪)得分对于价格影响最大**。这与我们在 3.1 部分对相关系数的讨论一致。

OLS Regression Results

==

Dep. Variable: price R-squared: 0.040

Model: OLS Adj. R-squared: 0.040

Method: Least Squares F-statistic: 132.4

Date: Sat, 16 Jan 2021 Prob (F-statistic): 5.03e-111

==

	coef	std err	t	P>|t|	[0.025	0.975]
const	127.9403	48.291	2.649	0.008	33.283	222.598
Topic1	2583.6070	117.250	22.035	0.000	2353.779	2813.435
Topic2	858.5674	101.744	8.439	0.000	659.134	1058.000
Topic3	299.6183	79.052	3.790	0.000	144.665	454.572
Topic4	236.3295	126.630	1.866	0.062	-11.885	484.544

5.1.2.4. LDA 主题分数和预测-实际价格比

下表展示了 LDA 主题分数和预测-实际价格比的 OLS 回归结果。结果显示模型具有边缘整体显著性。其中 Topic4 (负向情绪) 的得分在 10%水平上显著为正——即 Topic4 (负向情绪) 得分越高，预测价格更有可能高于真实价格。

OLS Regression Results

==

Dep. Variable: Predictoprice R-squared: 0.001

Model: OLS Adj. R-squared: 0.000

Method: Least Squares F-statistic: 1.982

Date: Sat, 16 Jan 2021 Prob (F-statistic): 0.0943

==

	coef	std err	t	P>|t|	[0.025	0.975]

---

const	1.2085	0.105	11.509	0.000	1.003	1.414
Topic1	-0.2911	0.255	-1.142	0.254	-0.791	0.209
Topic2	-0.3125	0.221	-1.413	0.158	-0.746	0.121
Topic3	-0.1180	0.172	-0.686	0.493	-0.455	0.219
Topic4	0.4629	0.275	1.681	0.093	-0.077	1.003

==========

5.1.2.5. LDA 主题分数和评论

由于 7 个评分变量高度相关，它们分别作为因变量时的回归结果也十分相似，因此我们在这里只展示了 review_scores_rating 作为因变量时的回归结果。

在 7 个 OLS 回归中，话题得分自变量均具有 1%水平的显著性，且 Topic2 (设施) 和 Topic4 (负向情绪) 的斜率均为负，说明它们含着负向情绪。且验证了 Topic3 (正向情绪) 具有显著的正面情绪。

OLS Regression Results

==========

Dep. Variable:	review_scores_rating	R-squared:	0.143
Model:	OLS	Adj. R-squared:	0.142
Method:	Least Squares	F-statistic:	529.8
Date:	Sat, 16 Jan 2021	Prob (F-statistic):	0.00

==========

	coef	std err	t	P>|t|	[0.025	0.975]

---

const	0.9569	0.005	208.036	0.000	0.948	0.966
Topic1	0.0686	0.011	6.142	0.000	0.047	0.090
Topic2	-0.1971	0.010	-20.337	0.000	-0.216	-0.178
Topic3	0.1586	0.008	21.058	0.000	0.144	0.173

Topic4	-0.2740	0.012	-22.717	0.000	-0.298	-0.250

==========

5.1.2.6. LDA 主题分数和类别

最后，我们对 LDA 主题分数和聚类类别的虚拟变量进行 Logistic 回归。我们重点关注类别 1 和类别 2。

对于类别 1，Topic2 得分增加 1 单位会使得属于类别 1 的赔率降低 0.5568 单位。由于 Topic2 得分和房屋设施有关，**类别 1 可能存在设施不完善的情况**。这和之前聚类分析里类别 1 的“简易”特征相吻合。

Generalized Linear Model Regression Results

==========

Dep. Variable:	is_1	No. Observations:	12756
Model:	GLM	Df Residuals:	12751
Model Family:	Binomial	Df Model:	4
Link Function:	logit	Scale:	1.0000

==========

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6060	0.084	-19.052	0.000	-1.771	-1.441
Topic1	1.9719	0.190	10.385	0.000	1.600	2.344
Topic2	-0.5568	0.182	-3.067	0.002	-0.913	-0.201
Topic3	1.6575	0.132	12.534	0.000	1.398	1.917
Topic4	2.2387	0.206	10.866	0.000	1.835	2.642

==========

对于类别 2，Topic4 得分增加会显著降低房源属于类别 2 的赔率。由于 Topic4 含负面情绪，因此**类别 2 的房源评价可能大多具有正面情绪**。这和之前聚类分析

里的房东回复率高可能有一定关系。

Generalized Linear Model Regression Results

==========

Dep. Variable:	is_2	No. Observations:	12756
Model:	GLM	Df Residuals:	12751
Model Family:	Binomial	Df Model:	4
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6499.3

==========

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0519	0.090	-11.707	0.000	-1.228	-0.876
Topic1	-1.1569	0.235	-4.926	0.000	-1.617	-0.697
Topic2	0.0971	0.188	0.517	0.605	-0.271	0.465
Topic3	0.1641	0.145	1.129	0.259	-0.121	0.449
Topic4	-1.6856	0.267	-6.316	0.000	-2.209	-1.163

5.1.3. LDA 总结

①关于主题和价格、Predict/Price 的关系

1)Topic1 (负向情绪)得分对于价格的正面影响最大；Topic2 (设施)，3 (正向情绪) 对价格也有正面影响；

2)Topic4 (负向情绪) 得分越高，预测价格更有可能高于真实价格。

②关于主题和评分的关系

1)Topic2 (设施)和 Topic4 (负向情绪) 和评分显著负相关；

2)Topic3 (正向情绪) 和评分显著正相关。

③关于主题和房源类别的关系

1)类别 1 房源可能设施不完备，和聚类分析的“简易”特征相吻合；

2)类别 2 房源评论含正面情绪多，可能和聚类分析的“房东回复率高”有关。

5. 2. 链家成交数据匹配

为了更好地预测价格，我们还加 airbnb 和链家成交房源数据进行了匹配 。匹配的思路有两种：**（1）限定距离 r，找到圆周内内链家房源均值数据；（2）限定数量 N，找到距离最近的 N 个链家房源均值数据**。

对方法（1），我们对每一个 airbnb 房源匹配了附近链家成交房源的数量（nearby_num_m）、平均 DOM（DOM_m）、平均关注人数（followers_m）、平均总价（totalPrice_m）、平均单价（price_m）、平均面积（square_m）、平均地铁 0-1 变量取值（subway_m）、平均周边社区价格（communitiyAverage_m）、平均到现在成交时间（tradeTime_to_today）

对方法（2），我们对每一个 airbnb 房源匹配了类似的变量，只是将链家成交房源的数量（nearby_num_m）替换为了最近房源的距离（nearby_dist_min）和附近房源的平均距离（nearby_dist_mean）

但这两种方法得到的特征跟价格相关系数很低，我们猜测可能是因为链家成交房源的性质和 airbnb 房源性质有较大的区别，成家房源受供需影响，而 airbnb 房源缺少需求端的数据——需求端只有顾客平分还都非常集中。因而，从周边成交房源的数据中，很难挖掘出对 airbnb 房源价格有影响的指标。

由于方法（1）会产生很多缺失数据，我们最终选择了方法二，限定 N=20，但结果相近。下表展示了方法（2）拼配出来的附近链家成交房源与价格和评分等变量的相关关系。我们发现相关系数都比较低，成交总价和均价和 aiabnb 房源的相关系数是负的；而最近 20 个链家房源与 airbnb 房源距离和价格有正向关系，这些结果都和我们的直觉相左。

表 29 链家成交数据匹配结果

	price	*rating*	*accuracy*	*cleanliness*	*checkin*	*communication*	*location*	*value*
nearby_dist_mean	0.049622	0.03107	0.021117	0.044817	0.011212	0.010768	-0.01254	0.009378
nearby_dist_min	0.025725	0.020644	0.011499	0.034846	0.003568	0.003656	-0.01599	0.003161
DOM_m	-0.0563	-0.02575	-0.02161	-0.03603	-0.01457	-0.01377	0.048035	-0.01349
followers_m	-0.03594	-0.01702	-0.00547	-0.02749	-0.00348	-0.00093	0.030746	-0.00588
totalPrice_m	-0.04883	-0.02422	-0.01735	-0.03478	-0.00968	-0.00922	0.05068	-0.01193
price_m	-0.03525	-0.01842	-0.0126	-0.02408	-0.0058	-0.00222	0.052627	-0.00744
square_m	0.014417	0.007617	0.005542	0.005861	0.000728	-0.00511	-0.03951	0.001324
subway_m	-0.05531	-0.02697	-0.021	-0.03921	-0.01139	-0.01184	0.048355	-0.01611
communitiyAverage_m	-0.03834	-0.01961	-0.01331	-0.02613	-0.0064	-0.00316	0.053105	-0.00819
tradeTime_to_today	0.001176	-0.00246	-0.00145	-0.00421	-0.00376	-0.00856	-0.02666	-0.00265

6. 结语

本文首先探讨了 airbnb 的商业模式和定价策略，作为问题引入。其次，我们对房源进行凝聚式聚类，分为 3 类，从中识别出“低端类”和“房东热情类”。接下来，使用线性回归、支持向量机、树类模型、神经网络对房源价格进行预测，**并得到“推荐比”**Predict / Price。

作为验证，我们**用评分相关变量对 Predict / Price 做拟合**，发现性价比评分和 Predict / Price 有正相关关系，验证了 Predict / Price 对“性价比”的测度能力。此外，我们对评论文本直接进行建模，使用 LDA 主题模型识别出 5 个主题，发现房源在这 5 个主题上的得分与价格、Predict / Price、评分相关变量和聚类结果都有显著的相关关系，并进一步解释了识别主题和聚类的关键要素。

我们的贡献有以下几点：

1. **先聚类，然后在组内使用多种模型预测价格，获得了较好的拟合优度；**
2. **尝试定义公平定价指标 Predict / Price 并探讨其影响因素；**
3. **对评分、价格、Predict / Price 之间的相关关系进行了综合的讨论；**
4. **在数值数据外挖掘了文本信息，并分析了主题得分和评分、价格、Predict / Price 以及聚类结果之间的关系。**

我们的不足主要有以下两个方面：

1. **在数据选取上，airbnb 官网数据集的信息并不全面，没有体现供需关系的数据；**
2. **在方法上，有很多细节值得进一步探讨——比如文本预处理分词的方法、特征工程上的一些细节等，还有待提升。**

参考文献

[1] Bybee, L., et al., 2019. The Structure of Economic News. NBER Working Paper No. w26648.

[2] Blei, D.M., et al.,2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[3] Peng Ye,Julian Qian,Jieying Chen,Chen-hung Wu,Yitong Zhou,Spencer De Mars,Frank Yang,Li Zhang. Customized Regression Model for Airbnb Dynamic Pricing[P]. Knowledge Discovery & Data Mining,2018.

[4] Gibbs,Guttentag,Gretzel,Morton,Goodwill. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings[J]. Journal of Travel & Tourism Marketing,2018,35(1).

附　录

一、　主要代码摘录

1)　部分变量预处理

```
#host_since 转为到现在(2020/12/1)的时间
def to_today(date):
    present = datetime.datetime.strptime('2020-12-1',"%Y-%m-%d")
    if date<present:
        num=(present-date).days
        return num
    elif date>=present:
        num=(date-present).days
        return num
listing_data['host_since'] = pd.to_datetime(listing_data['host_since'])
listing_data['host_time'] = listing_data['host_since'].apply(lambda x:to_today(x))
listing_data['first_review'] = pd.to_datetime(listing_data['first_review'])
listing_data['first_review_to_today'] = listing_data['first_review'].apply(lambda x:to_today(x))
listing_data['last_review'] = pd.to_datetime(listing_data['last_review'])
listing_data['last_review_to_today'] = listing_data['last_review'].apply(lambda x:to_today(x))

#计算 Bathroom 的数量
def countBathroom(x):
    if type(x) is str:
        if (x=='Shared half-bath')|(x=='Private half-bath')|(x=='Half-bath'):
            num = 0.5
```

```
            else:
                num = float(re.findall(r'[\d+(\.\d+)?]',x)[0])
        else:
            num = 0
        return num
listing_data["bathrooms"] = listing_data.bathrooms_text.apply(lambda x: countBathroom(x))

#将百分数转化为数值
listing_data["host_response_rate"]=listing_data["host_response_rate"].str.strip("%").astype('floa
t')/100
listing_data["host_acceptance_rate"]=listing_data["host_acceptance_rate"].str.strip("%").astype('
float')/100

#将 reviews_per_month 的缺失值填充为 0
listing_data['reviews_per_month'].fillna(0, inplace = True)

# 字符串列表转化为长度
def getLen(x):
    li = x[2:-2].split(',')
    length = len(li)
    return length
listing_data["num_of_host_verifications"] = listing_data.host_verifications.apply(lambda
x:getLen(x))
listing_data["num_of_amenities"] = listing_data.amenities.apply(lambda x:getLen(x))

#price 去掉美元符号，转为 float
listing_data["price"]=listing_data["price"].str.replace(',', '').apply(lambda x : x[1:]).astype('float')
```

2) 聚类 Agglomerative Clustering

```
#聚类前归一化
cluster_scaler = MinMaxScaler(feature_range=(0, 1), copy=True)
cluster_data=cluster_scaler.fit_transform(cluster_data)
#聚类
cluster_model=AgglomerativeClustering(n_clusters=3, affinity='euclidean', memory=None,
connectivity=None, compute_full_tree='auto', linkage='ward', distance_threshold=None)
labels=cluster_model.fit_predict(cluster_data)
output_data=pd.concat([pd.DataFrame(labels,columns=["Label"]),new_listing_data],axis=1)
```

3) 线性回归 Linear Regressions（以聚类 0 为例）

```
#线性回归 statsmodels
modelLR2=sm.OLS(Y,X).fit()
print("Linear Regression")
print(modelLR2.summary())
print("\n\n")

#线性回归(Standardlized) statsmodels
modelLRStandardlized2=sm.OLS(YStandardlized,XStandardlized).fit()
print("Linear Regression (Standardlized)")
print(modelLRStandardlized2.summary())
print("\n\n")

#Ridge 回归(Standardlized)
modelRidgeStandardlized=RidgeCV(alphas=range(2600,2650),fit_intercept=False,
normalize=False, scoring=None, cv=5, gcv_mode=None, store_cv_values=False)
modelRidgeStandardlized.fit(XStandardlized,YStandardlized)
print("Ridge Regression (Standardlized)")
```

```
print("Alpha_\n{}".format(modelRidgeStandardlized.alpha_))
print("Score\n{}".format(modelRidgeStandardlized.score(XStandardlized,YStandardlized)))
print("Intercept\n{}".format(modelRidgeStandardlized.intercept_))
print("Coef")
for variable,coef in zip(X_names,modelRidgeStandardlized.coef_[0]):
    print("|{0:^50s}|{1:^30f}|".format(variable,coef))
print("\n\n")

#Lasso 回归(Standardlized)
modelLassoStandardlized=LassoCV(eps=1e-4, n_alphas=1000, alphas=None, fit_intercept=False,
normalize=False, precompute='auto', max_iter=100000000, tol=1e-8, copy_X=True, cv=5,
verbose=0, n_jobs=-1, positive=False, random_state=None, selection='cyclic')
modelLassoStandardlized.fit(XStandardlized,YStandardlized)
print("Lasso Regression (Standardlized)")
print("Alpha_\n{}".format(modelLassoStandardlized.alpha_))
print("Score\n{}".format(modelLassoStandardlized.score(XStandardlized,YStandardlized)))
print("Intercept\n{}".format(modelLassoStandardlized.intercept_))
print("Coef")
for variable,coef in zip(X_names,modelLassoStandardlized.coef_):
    print("|{0:^50s}|{1:^30f}|".format(variable,coef))
print("\n\n")
```

4)　支持向量机 Support Vector Machines（以聚类 0 为例）

```
#交叉验证选取最佳参数
modelSVR=GridSearchCV(SVR(coef0=0.0, tol=1e-6,epsilon=0.1,shrinking=True, cache_size=2000,
verbose=True, max_iter=-1,degree=3),
                      param_grid={"kernel":["rbf",],"gamma":[gamma/10000 for gamma in
range(50,60)],"C":[c/10 for c in range(41,51)]},
```

```
                                    scoring=None,cv=5,n_jobs=8,verbose=3)
modelSVR.fit(XStandardlized,YStandardlized)
print(modelSVR.best_params_)
print(modelSVR.best_score_)
best_model=modelSVR.best_estimator_
print("score\n{}".format(best_model.score(XStandardlized,YStandardlized)))
print("support_\n{}".format(best_model.support_))
print("support_vectors_\n{}".format(best_model.support_vectors_))
print("dual_coef_\n{}".format(best_model.dual_coef_[0]))
print("intercept_\n{}".format(best_model.intercept_))
vectors_weight=pd.concat([pd.DataFrame(best_model.support_vectors_,columns=X_names),pd.
DataFrame(best_model.dual_coef_[0],columns=["Dual Coef"])],axis=1)
vectors_weight.to_csv("SVR1219-0.csv",
                              sep=',', na_rep='',columns=None,
header=True,index=False,encoding="ANSI")
```

5) 决策树 Decision Tree（以聚类 0 为例）

```
#交叉验证选取最佳参数
modelDT=GridSearchCV(DecisionTreeRegressor(),
                     
param_grid={"criterion":["mae"],"splitter":["best"],'max_depth':range(18,19),"min_samples_leaf
":range(15,18),"max_features":range(38,41),"min_samples_split":range(45,66,2)},
                                    scoring=None,cv=5,n_jobs=8,verbose=3)
modelDT.fit(X,Y)
print(modelDT.best_params_)
print(modelDT.best_score_)
best_model=modelDT.best_estimator_
print("classes_\n{}".format(best_model.classes_))
```

```
print("max_features_\n{}".format(best_model.max_features_))
print("get_depth()\n{}".format(best_model.get_depth()))
print("get_n_leaves()\n{}".format(best_model.get_n_leaves()))
print("score\n{}".format(best_model.score(X,Y)))
print("feature_importances")
for variable,coef in zip(X_names,best_model.feature_importances_):
    print("|{0:^50s}|{1:^30f}|".format(variable,coef))
```

6)　随机森林 Random Forests（以聚类 0 为例）

```
#对 cluster0 的 随机森林调参建模
params_test2={
    'n_estimators':[300,400,500,600,700,800],
    'max_depth':[15,18,21,25,28,30],
    'max_features':[15,18,21,24,27,30],
    'criterion':['mse','mae'],
    }

regr_RF_tailor=RandomForestRegressor(min_weight_fraction_leaf=0.0,min_samples_split=2,
min_samples_leaf=2,min_impurity_decrease=0.0,min_impurity_split=None,bootstrap=True,
oob_score=True,n_jobs=6,verbose=2,warm_start=False)
gs2=GridSearchCV(regr_RF_tailor,param_grid=params_test2,scoring=None,n_jobs=6,cv=5,verbos
e=2)

gs2.fit(X,Y)
best_score=gs2.best_score_
best_params=gs2.best_params_
feature_importances_CV=gs2.best_estimator_.feature_importances_
feature_importances_CV_list=list(zip(cluster0.columns[2:],feature_importances_CV))
```

```
feature_importances_CV_list.sort(key=lambda x:x[1],reverse=True)
feature_importances_CV_sort=dict(feature_importances_CV_list)
print('Cluster 0')
print('RF_tailor Best Score:')
print(best_score)
print('RF_tailor Best Params:')
print(best_params)
print('RF_tailor Features Importance:')
print(feature_importances_CV_sort)
```

7) 极端随机树 Extremely Randomized Trees（以聚类 0 为例）

```
#交叉验证选取最佳参数
modelET=GridSearchCV(ExtraTreesRegressor(),
param_grid={"criterion":["mae"],'bootstrap':[False],"max_features":range(20,31,5),"n_estimators
":range(600,801,100),"max_depth":range(30,46,5),"min_samples_split":range(5,10,2),"verbose":
[3],"n_jobs":[8],"min_samples_leaf":[1]},
                                    scoring=None,cv=5,n_jobs=8,verbose=3)
modelET.fit(X,Y)
print(modelET.best_params_)
print(modelET.best_score_)
best_model=modelET.best_estimator_
print("score\n{}".format(modelET.score(X,Y)))
print("feature_importances")
for variable,coef in zip(X_names,best_model.feature_importances_):
    print("|{0:^50s}|{1:^30f}|".format(variable,coef))
```

8)　梯度提升树 XGBoost（以聚类 0 为例）

```
#交叉验证选取最佳参数
modelXGB=GridSearchCV(XGBRegressor(verbosity=1, min_split_loss=0,
objective='reg:squarederror',booster='gbtree',nthread=-1,gpu_id=0),

param_grid={"max_depth":[3],"n_estimators":range(80,81),"num_parallel_tree":[1],"learning_ra
te":[rate/10 for rate in range(1,2)],"subsample":[ratio/10 for ratio in range(9,10)],
"reg_alpha":[alpha/10 for alpha in range(56,66)], "reg_lambda":[beta/10 for beta in
range(22,23)],"eval_metric":["rmse",],"tree_method":["auto"],"sampling_method":["uniform"]},
                                scoring=None,cv=5,n_jobs=8,verbose=3)
modelXGB.fit(X,Y)
print(modelXGB.best_params_)
print(modelXGB.best_score_)
best_model=modelXGB.best_estimator_
print("score\n{}".format(modelXGB.score(X,Y)))
print("feature_importances")
for variable,coef in zip(X_names,best_model.feature_importances_):
    print("|{0:^50s}|{1:^30f}|".format(variable,coef))
```

9)　神经网络（以聚类 0 为例）

```
#对 cluster0 的 multilayer neural network (standardized)调参建模
params_test1={
    'activation':['logistic'],
    'learning_rate':['constant','adaptive'],
    'hidden_layer_sizes':[(1500,),(1600,),(1500,1000)],
    'alpha':[0.0001],
    'tol':[1e-6],
    'learning_rate_init':[0.0005],
```

```
    'solver':['adam'],
    }

regr_std_tailor=MLPRegressor(batch_size='auto',shuffle=True,
                             max_iter=5000,verbose=True,warm_start=False)
gs1=GridSearchCV(regr_std_tailor,param_grid=params_test1,scoring=None,n_jobs=8,cv=5,verbos
e=2)

gs1.fit(XStandardlized,YStandardlized)
print('Cluster 0')
print('NN_std_tailor Best Score:    {}'.format(str(gs1.best_score_)))
print('NN_std_tailor Best Params:    {}'.format(str(gs1.best_params_)))
print('\n\n')
```

二、 主要模型结果摘录

1) 线性回归 Linear Regressions

表 30 聚类 0 线性回归系数汇总

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
host_time	0.0269	0.0083	0.007897	0.004333
num_of_host_verifications	2.8733	0.0044	0.003975	0.002184
num_of_amenities	3.3679	0.0145	0.015834	0.010733
latitude	1079.599	0.1646	0.042066	0.105331
longitude	330.2351	0.0561	-0.00744	0
host_response_rate	-36.6581	-0.0053	-0.01216	0
host_acceptance_rate	108.6065	0.0138	0.007652	0.003346
host_listings_count	1.1315	0.0279	0.015716	0.007086
accommodates	78.3454	0.1662	0.125413	0.16614
bathrooms	200.271	0.1473	0.123799	0.145865

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
bedrooms	74.2153	0.0732	0.093386	0.069654
beds	59.6871	0.104	0.098185	0.101157
minimum_nights	-0.5399	-0.0051	-0.0047	0
maximum_nights	-0.1243	-0.0273	-0.02074	-0.02048
availability_30	5.0342	0.0349	0.00772	0.006392
availability_60	-5.4632	-0.0724	0.001685	0
availability_90	3.3041	0.0555	0.00494	0
availability_365	-0.0181	-0.0013	-0.00144	0
number_of_reviews	0.6556	0.0054	0.000315	0
number_of_reviews_ltm	-10.5032	-0.0227	-0.01639	-0.00825
calculated_host_listings_count	6.098	-0.0061	-0.00097	0
calculated_host_listings_count_entire_homes	-8.2412	-0.0188	-0.01084	0
calculated_host_listings_count_private_rooms	12.388	0.0521	0.041071	0.043836
calculated_host_listings_count_shared_rooms	1.9513	0.002	0.001738	0
host_response_time_a few days or more	-7518.71	0.0155	0.00866	0
host_response_time_within a day	-7455.41	0.0217	0.017996	0.000833
host_response_time_within a few hours	-7439.68	0.0258	0.022823	0.001401
host_response_time_within an hour	-7773.28	-0.0392	-0.03052	-0.06203
host_is_superhost_t	-15070	0.0052	0.002854	0
host_is_superhost_f	-15120	-0.0052	-0.00285	0
host_has_profile_pic_t	-14910	0.0032	0.002895	0.00027
host_has_profile_pic_f	-15280	-0.0032	-0.0029	0
host_identity_verified_t	-14930	0.0036	0.003062	0
host_identity_verified_f	-15250	-0.0036	-0.00306	0
property_type_Barn	-1151.29	-0.0067	-0.00381	0
property_type_Camper/RV	116.8244	0.0139	0.009998	0.008518
property_type_Campsite	572.3906	0.0154	0.011923	0.009675

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Casa particular	-843.665	-0.0042	-0.00344	0
property_type_Castle	-502.992	-0.0035	-0.00211	0
property_type_Cave	-2.7E-10	-1.3E-17	0	0
property_type_Dome house	-777.821	-0.0026	-0.00163	0
property_type_Earth house	735.1782	0.0172	0.015019	0.011871
property_type_Entire apartment	-394.038	-0.0186	-0.02161	0
property_type_Entire bed and breakfast	-482.814	-0.0015	-0.00134	0
property_type_Entire bungalow	-211.899	0.0107	0.013365	0.014145
property_type_Entire cabin	-453.307	-0.0035	-0.00176	0
property_type_Entire chalet	-483.197	-0.0013	-0.00013	0
property_type_Entire condominium	-440.737	-0.0193	-0.01878	-0.00337
property_type_Entire cottage	-47.6151	0.0247	0.024491	0.024282
property_type_Entire guest suite	-317.633	0.0004	-9.9E-05	0
property_type_Entire guesthouse	-271.223	0.0021	0.00082	0
property_type_Entire home/apt	-646.006	-0.0018	-0.00144	0
property_type_Entire house	-335.997	-0.0013	0.003289	0.005343
property_type_Entire loft	-443.85	-0.0214	-0.02119	-0.00871
property_type_Entire place	-582.383	-0.0038	-0.00361	0
property_type_Entire resort	-825.701	-0.004	-0.00367	0
property_type_Entire serviced apartment	-273.49	0.0089	0.002233	0.010899
property_type_Entire townhouse	-510.326	-0.013	-0.00668	-0.00177
property_type_Entire villa	271.174	0.0808	0.076161	0.083857
property_type_Farm stay	-513.067	-0.0199	-0.01044	-0.00709
property_type_Houseboat	-306.756	0.0001	-0.00036	0
property_type_Hut	-576.643	-0.0028	-0.00209	0
property_type_Igloo	-396.079	-0.0004	-0.00034	0
property_type_Kezhan	-324.969	0.0001	0.000329	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Minsu	40.7039	0.0052	0.006244	0
property_type_Pension	-8.9E-11	-3.8E-18	0	0
property_type_Private room	-3.8E-12	6.88E-18	0	0
property_type_Private room in apartment	1.63E-10	2.16E-18	0	0
property_type_Private room in barn	1.09E-11	-5.4E-18	0	0
property_type_Private room in bed and breakfast	1.4E-11	-6.3E-18	0	0
property_type_Private room in bungalow	-3.5E-11	-8.3E-19	0	0
property_type_Private room in cabin	6.15E-11	5.21E-18	0	0
property_type_Private room in camper/rv	-3.8E-11	9.59E-19	0	0
property_type_Private room in campsite	3.4E-12	-1.1E-18	0	0
property_type_Private room in casa particular	-1.2E-11	-7.7E-19	0	0
property_type_Private room in castle	2.45E-12	-4.3E-18	0	0
property_type_Private room in cave	-5.6E-13	-1.4E-18	0	0
property_type_Private room in chalet	-4.8E-13	3.77E-19	0	0
property_type_Private room in condominium	1.78E-14	-8.7E-19	0	0
property_type_Private room in cottage	-7019.08	-0.0408	-0.02971	-0.0677
property_type_Private room in dome house	-1E-14	-3.7E-18	0	0
property_type_Private room in earth house	2.96E-14	3.26E-18	0	0
property_type_Private room in farm stay	-1.9E-14	-6.2E-19	0	0
property_type_Private room in guest suite	1.45E-14	-1.4E-18	0	0
property_type_Private room in guesthouse	0	-9.8E-19	0	0
property_type_Private room in hostel	0	-8.8E-20	0	0
property_type_Private room in house	0	-6.9E-20	0	0
property_type_Private room in hut	0	-1.4E-19	0	0
property_type_Private room in kezhan	0	-4.2E-20	0	0
property_type_Private room in loft	0	-2.5E-19	0	0
property_type_Private room in minsu	0	4E-19	0	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Private room in nature lodge	0	0	0	0
property_type_Private room in resort	0	0	0	0
property_type_Private room in ryokan	0	0	0	0
property_type_Private room in serviced apartment	-1717.71	0.0149	0.009069	0
property_type_Private room in tent	0	0	0	0
property_type_Private room in tiny house	-1780.73	0.0145	0.008352	0
property_type_Private room in townhouse	0	0	0	0
property_type_Private room in treehouse	0	0	0	0
property_type_Private room in villa	0	0	0	0
property_type_Room in aparthotel	-149.335	0.0031	0.002215	0
property_type_Room in boutique hotel	-719.937	-0.005	-0.00277	0
property_type_Room in heritage hotel	0	0	0	0
property_type_Room in hotel	0	0	0	0
property_type_Shared room	0	0	0	0
property_type_Shared room in apartment	0	0	0	0
property_type_Shared room in bed and breakfast	-3897.35	0.0047	0.002588	0
property_type_Shared room in boutique hotel	0	0	0	0
property_type_Shared room in bungalow	0	0	0	0
property_type_Shared room in condominium	0	0	0	0
property_type_Shared room in cottage	-3884.87	0.0047	0.003828	0
property_type_Shared room in earth house	0	0	0	0
property_type_Shared room in farm stay	0	0	0	0
property_type_Shared room in guest suite	0	0	0	0
property_type_Shared room in guesthouse	0	0	0	0
property_type_Shared room in hostel	0	0	0	0
property_type_Shared room in house	0	0	0	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Shared room in hut	0	0	0	0
property_type_Shared room in kezhan	0	0	0	0
property_type_Shared room in loft	0	0	0	0
property_type_Shared room in nature lodge	0	0	0	0
property_type_Shared room in serviced apartment	0	0	0	0
property_type_Shared room in tent	0	0	0	0
property_type_Shared room in tiny house	0	0	0	0
property_type_Shared room in townhouse	0	0	0	0
property_type_Shared room in villa	0	0	0	0
property_type_Tent	-323.237	4.05E-05	-8.7E-05	0
property_type_Tiny house	-317.62	0.0003	-1.7E-05	0
property_type_Treehouse	0	0	0	0
property_type_Yurt	0	0	0	0
room_type_Entire home/apt	-11890	0.0208	0.016517	0
room_type_Private room	-10520	-0.0262	-0.0204	0
room_type_Shared room	-7782.22	0.0066	0.004537	0
instant_bookable_t	-15100	-0.0005	-0.0006	0
instant_bookable_f	-15090	0.0005	0.000601	0
neighbourhood_cleansed_昌平区	-1828.08	0.0045	0.014378	0
neighbourhood_cleansed_朝阳区 / Chaoyang	-1731.37	0.0355	0.006324	0.01106
neighbourhood_cleansed_大兴区 / Daxing	-1686.43	0.0214	-0.01273	0
neighbourhood_cleansed_东城区	-1633.77	0.0389	0.012345	0.012915
neighbourhood_cleansed_房山区	-1621.84	0.0257	-0.0163	-0.00175
neighbourhood_cleansed_丰台区 / Fengtai	-1825.89	0.006	-0.02045	-0.01085
neighbourhood_cleansed_海淀区	-1823.03	0.0058	-0.00892	-0.00418
neighbourhood_cleansed_怀柔区 / Huairou	-1880	-0.0015	0.04081	0.012808
neighbourhood_cleansed_门头沟区 / Mentougou	-1610.03	0.0158	-0.00081	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
neighbourhood_cleansed_密云县 / Miyun	-2908.43	-0.1615	-0.03703	-0.08939
neighbourhood_cleansed_平谷区 / Pinggu	-2502.37	-0.0192	-0.00175	-0.00183
neighbourhood_cleansed_石景山区	-1765.62	0.0061	-0.00606	0
neighbourhood_cleansed_顺义区 / Shunyi	-2124.92	-0.0351	-0.01767	-0.02462
neighbourhood_cleansed_通州区 / Tongzhou	-1848.92	0.0023	-0.00829	0
neighbourhood_cleansed_西城区	-1615.51	0.0316	0.008931	0.009879
neighbourhood_cleansed_延庆县 / Yanqing	-1780.89	0.013	0.037055	0.006366

表 31 聚类 1 线性回归系数汇总

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
host_time	0.0102	0.0048	0.000024	0
num_of_host_verifications	9.2023	0.0204	0.013456	0.006112
num_of_amenities	1.7328	0.0131	0.012567	0
latitude	-336.807	-0.0706	0.013165	0
longitude	125.8247	0.0294	-0.00234	0
host_response_rate	-426.577	-0.0907	-0.01811	-0.0129
host_acceptance_rate	-45.0594	-0.0077	-0.00872	-0.00307
host_listings_count	-2.2324	-0.0292	-0.01737	-0.01582
accommodates	105.6814	0.2292	0.148309	0.223378
bathrooms	174.4148	0.1358	0.111065	0.125893
bedrooms	41.9022	0.0496	0.063325	0.038197
beds	-13.4104	-0.0288	0.016161	-0.00108
minimum_nights	-0.7116	-0.0071	-0.00644	0
maximum_nights	0.0029	0.0009	-0.0017	0
availability_30	2.7049	0.026	0.009865	0.009096
availability_60	-1.183	-0.0218	0.005176	0
availability_90	0.5732	0.0135	0.004553	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
availability_365	-0.1607	-0.0166	-0.0105	-0.00251
number_of_reviews	-0.7059	-0.0121	-0.01186	-0.00366
number_of_reviews_ltm	-7.9468	-0.0266	-0.01948	-0.01926
calculated_host_listings_count	-1.9429	-0.0042	-0.00194	0
calculated_host_listings_count_entire_homes	9.1511	0.035	0.021169	0.015841
calculated_host_listings_count_private_rooms	-0.0584	-0.0169	-0.00841	0
calculated_host_listings_count_shared_rooms	-11.0356	-0.0228	-0.01992	-0.02343
host_response_time_a few days or more	-284.58	-0.0536	-0.00573	0
host_response_time_within a day	-74.0277	-0.0046	-0.00118	0
host_response_time_within a few hours	32.383	0.0192	0.007747	0
host_response_time_within an hour	21.0793	0.0257	-0.00053	0
host_is_superhost_t	-172.068	-0.0071	-0.00915	-0.00789
host_is_superhost_f	-133.078	0.0071	0.009146	0
host_has_profile_pic_t	-123.554	0.0004	0.000452	0
host_has_profile_pic_f	-181.592	-0.0004	-0.00045	0
host_identity_verified_t	-30.9724	0.0027	0.000842	0
host_identity_verified_f	-274.174	-0.0027	-0.00084	0
property_type_Barn	-3.4E-11	-2.7E-18	0	0
property_type_Camper/RV	-6.9E-13	-1.7E-17	0	0
property_type_Campsite	-3.8E-11	2.57E-17	0	0
property_type_Casa particular	-9.3E-12	5.27E-17	0	0
property_type_Castle	11490	0.1085	0.080332	0.098
property_type_Cave	6.59E-12	-6.5E-17	0	0
property_type_Dome house	-6.1E-12	-6.7E-17	0	0
property_type_Earth house	-3.4E-13	-2.4E-17	0	0
property_type_Entire apartment	-475.717	-0.0046	-0.00972	-0.00077
property_type_Entire bed and breakfast	-7.4E-12	1.17E-17	0	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Entire bungalow	41.6996	0.0054	0.00248	0
property_type_Entire cabin	8.17E-12	-1.2E-18	0	0
property_type_Entire chalet	1.07E-11	1.4E-17	0	0
property_type_Entire condominium	-554.798	-0.0047	-0.00687	0
property_type_Entire cottage	484.753	0.026	0.015204	0.012555
property_type_Entire guest suite	1.82E-11	3.83E-17	0	0
property_type_Entire guesthouse	1.42E-12	1.45E-17	0	0
property_type_Entire home/apt	-8.8E-12	-3.2E-17	0	0
property_type_Entire house	-162.765	0.0042	-0.00065	0
property_type_Entire loft	-625.186	-0.0085	-0.01172	-0.00228
property_type_Entire place	-8.6E-12	2.16E-17	0	0
property_type_Entire resort	-3.3E-12	-4.1E-17	0	0
property_type_Entire serviced apartment	-3.2E-12	-6.3E-19	0	0
property_type_Entire townhouse	-393.075	-0.0015	-0.00319	0
property_type_Entire villa	1632.468	0.0251	0.018563	0.015342
property_type_Farm stay	-118.851	0.0037	0.000287	0
property_type_Houseboat	1.8E-12	-1.2E-17	0	0
property_type_Hut	-4.5E-12	7.61E-19	0	0
property_type_Igloo	1.21E-12	3.73E-18	0	0
property_type_Kezhan	-3.3E-12	6.57E-19	0	0
property_type_Minsu	-7.5E-12	-1.5E-17	0	0
property_type_Pension	-1.5E-12	-1.4E-17	0	0
property_type_Private room	-578.597	-0.0117	-0.00982	-0.00237
property_type_Private room in apartment	-378.37	-0.0266	-0.02544	-0.01016
property_type_Private room in barn	-186.116	0.0018	0.000137	0
property_type_Private room in bed and breakfast	-280.508	0.0023	0.005426	0
property_type_Private room in bungalow	-250.688	0.0076	0.011058	0.007211

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Private room in cabin	-382.497	-0.002	0.000384	0
property_type_Private room in camper/rv	-513.29	-0.002	-0.0004	0
property_type_Private room in campsite	-305.386	-4.7E-05	-0.00026	0
property_type_Private room in casa particular	-637.85	-0.0054	-0.00433	0
property_type_Private room in castle	-287.274	0.001	-9.6E-05	0
property_type_Private room in cave	-184.573	0.0028	0.001415	0
property_type_Private room in chalet	-68.571	0.003	0.002553	0
property_type_Private room in condominium	-367.143	-0.0151	-0.01651	-0.00243
property_type_Private room in cottage	-325.931	-0.0031	0.003234	0
property_type_Private room in dome house	-519.891	-0.002	-0.00251	0
property_type_Private room in earth house	596.9646	0.0184	0.014868	0.01094
property_type_Private room in farm stay	-805.449	-0.0997	-0.06622	-0.07991
property_type_Private room in guest suite	-451.037	-0.0106	-0.00821	0
property_type_Private room in guesthouse	-346.774	-0.0045	-0.00521	0
property_type_Private room in hostel	-350.073	-0.0038	-0.00589	0
property_type_Private room in house	-365.409	-0.0164	-0.00928	0
property_type_Private room in hut	-92.5207	0.0038	0.004148	0
property_type_Private room in kezhan	-25.0918	0.0434	0.038501	0.041973
property_type_Private room in loft	-128.868	0.0232	0.018003	0.017908
property_type_Private room in minsu	-355.416	-0.0018	-0.00163	0
property_type_Private room in nature lodge	127.546	0.0335	0.029934	0.028769
property_type_Private room in resort	221.5187	0.045	0.032183	0.035001
property_type_Private room in ryokan	853.1823	0.015	0.012339	0.006208
property_type_Private room in serviced apartment	-248.232	0.009	0.005716	0.002988
property_type_Private room in tent	-521.887	-0.0035	-0.0023	0
property_type_Private room in tiny house	-573.01	-0.0109	-0.00743	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Private room in townhouse	-448.875	-0.0157	-0.00942	-0.00122
property_type_Private room in treehouse	-2.1E-12	1.02E-17	0	0
property_type_Private room in villa	-209.833	0.0152	0.014437	0.012108
property_type_Room in aparthotel	36.1362	0.0154	0.01324	0.007316
property_type_Room in boutique hotel	285.9279	0.0926	0.068097	0.086588
property_type_Room in heritage hotel	-3422.01	-0.0406	-0.02764	-0.03069
property_type_Room in hotel	-64.6391	0.0288	0.018036	0.018888
property_type_Shared room	7.89E-13	-9.6E-18	0	0
property_type_Shared room in apartment	-320.783	-0.0078	-0.01033	0
property_type_Shared room in bed and breakfast	-412.188	-0.0031	-0.00022	0
property_type_Shared room in boutique hotel	5072.387	0.0695	0.051007	0.058699
property_type_Shared room in bungalow	-254.343	0.0007	-0.00198	0
property_type_Shared room in condominium	-383.624	-0.0099	-0.01052	-0.00031
property_type_Shared room in cottage	-250.054	0.0006	-0.0003	0
property_type_Shared room in earth house	771.6663	0.0096	0.008708	0.001758
property_type_Shared room in farm stay	-477.633	-0.0038	-0.00111	0
property_type_Shared room in guest suite	-469.889	-0.0031	-0.00219	0
property_type_Shared room in guesthouse	-288.274	-0.0006	-0.00262	0
property_type_Shared room in hostel	-314.694	-0.0032	-0.0062	0
property_type_Shared room in house	-264.328	0.0007	-0.0013	0
property_type_Shared room in hut	0	0	0	0
property_type_Shared room in kezhan	-1346.14	-0.0171	-0.01013	-0.0075
property_type_Shared room in loft	-439.177	-0.0095	-0.0077	0
property_type_Shared room in nature lodge	0	0	0	0
property_type_Shared room in serviced apartment	334.2511	0.0137	0.00921	0.004887
property_type_Shared room in tent	-235.865	0.0003	-0.00077	0
property_type_Shared room in tiny house	0	0	0	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Shared room in townhouse	-361.245	-0.0018	0.000985	0
property_type_Shared room in villa	-432.389	-0.0041	-0.00052	0
property_type_Tent	0	0	0	0
property_type_Tiny house	0	0	0	0
property_type_Treehouse	0	0	0	0
property_type_Yurt	0	0	0	0
room_type_Entire home/apt	-292.317	-0.0205	0.00922	0
room_type_Private room	59.4935	0.0152	0.010234	0.00732
room_type_Shared room	-72.3222	-0.0095	-0.01346	-0.01715
instant_bookable_t	-153.939	-0.0005	0.000751	0
instant_bookable_f	-151.207	0.0005	-0.00075	0
neighbourhood_cleansed_昌平区	-144.131	-0.0212	-0.01878	-0.00126
neighbourhood_cleansed_朝阳区 / Chaoyang	-226.992	-0.0716	-0.03829	-0.01335
neighbourhood_cleansed_大兴区 / Daxing	408.0315	0.0491	0.047226	0.065766
neighbourhood_cleansed_东城区	28.1995	0.0141	0.021252	0.039334
neighbourhood_cleansed_房山区	-155.694	-0.0177	-0.00356	0
neighbourhood_cleansed_丰台区 / Fengtai	-273.568	-0.0351	-0.01693	-0.00441
neighbourhood_cleansed_海淀区	-160.26	-0.0306	-0.02104	-0.0019
neighbourhood_cleansed_怀柔区 / Huairou	323.8475	0.0806	0.03982	0.07026
neighbourhood_cleansed_门头沟区 / Mentougou	-93.7633	-0.0043	-0.00165	0
neighbourhood_cleansed_密云县 / Miyun	-16.8726	0.0025	-0.01233	0
neighbourhood_cleansed_平谷区 / Pinggu	333.0584	0.0373	0.030205	0.039615
neighbourhood_cleansed_石景山区	-126.783	-0.0075	-0.00414	0
neighbourhood_cleansed_顺义区 / Shunyi	-329.724	-0.038	-0.02199	-0.01128
neighbourhood_cleansed_通州区 / Tongzhou	-390.865	-0.0487	-0.0252	-0.01123
neighbourhood_cleansed_西城区	206.0488	0.0416	0.039416	0.057648
neighbourhood_cleansed_延庆县 / Yanqing	314.3198	0.0796	0.020997	0.046389

表 32 聚类 2 线性回归系数汇总

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
host_time	-0.059	-0.012	-0.00121	0
num_of_host_verifications	32.314	0.0336	0.010815	0.010398
num_of_amenities	1.3296	0.0045	0.002838	0
latitude	1239.84	0.1238	0.035842	0.027057
longitude	599.0976	0.0659	0.007059	0
host_response_rate	-1062.55	-0.0369	-0.00828	-0.00342
host_acceptance_rate	-425.287	-0.0073	0.002837	0
host_listings_count	16.1434	0.1102	0.022512	0.025578
accommodates	48.8615	0.0628	0.043056	0.055895
bathrooms	192.3104	0.0839	0.047305	0.086396
bedrooms	127.7646	0.0709	0.040844	0.050265
beds	21.4875	0.0215	0.038058	0.026553
minimum_nights	-2.1216	-0.0087	-0.00261	0
maximum_nights	-0.5553	-0.0822	-0.03252	-0.06558
availability_30	18.9452	0.0859	0.014871	0.008493
availability_60	0.5072	0.0043	0.006377	0
availability_90	-4.8955	-0.0547	0.002126	0
availability_365	-0.7048	-0.0346	-0.01476	-0.01356
number_of_reviews	-3.2274	-0.0271	-0.01454	-0.00662
number_of_reviews_ltm	-5.9582	-0.0183	-0.01583	-0.00832
calculated_host_listings_count	-12.6133	-0.0325	-0.0067	0
calculated_host_listings_count_entire_homes	1.0347	-0.0279	-0.0069	0
calculated_host_listings_count_private_rooms	-25.1179	-0.0349	-0.00158	0
calculated_host_listings_count_shared_rooms	11.4699	0.0013	-0.00121	0
host_response_time_a few days or more	-9504.67	-0.0212	-0.0018	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
host_response_time_within a day	-9157.78	-0.017	-0.00278	0
host_response_time_within a few hours	-8177.15	0.0538	0.014918	0.021757
host_response_time_within an hour	-9101.3	-0.0343	-0.01148	0
host_is_superhost_t	-17760	0.0062	0.000145	0
host_is_superhost_f	-18180	-0.0062	-0.00015	0
host_has_profile_pic_t	-35940	1.05E-16	0	0
host_has_profile_pic_f	9.12E-10	5.84E-17	0	0
host_identity_verified_t	-20250	-0.0151	-0.00681	-0.00799
host_identity_verified_f	-15690	0.0151	0.006809	0
property_type_Barn	1.44E-10	2.98E-17	0	0
property_type_Camper/RV	-1311.78	-0.0057	-0.00171	0
property_type_Campsite	-2708.31	-0.0266	-0.00902	-0.00035
property_type_Casa particular	2.18E-10	-5.1E-17	0	0
property_type_Castle	3.42E-10	5.72E-17	0	0
property_type_Cave	8.27E-10	-1.1E-17	0	0
property_type_Dome house	-4.9E-10	-1.4E-17	0	0
property_type_Earth house	-781.744	-0.0011	-0.00076	0
property_type_Entire apartment	-871.311	-0.0325	-0.01969	0
property_type_Entire bed and breakfast	-1457.14	-0.005	-0.00135	0
property_type_Entire bungalow	-240.616	0.0339	0.015863	0.031229
property_type_Entire cabin	-1057.1	-0.0086	-0.00249	0
property_type_Entire chalet	-3.7E-10	1.34E-17	0	0
property_type_Entire condominium	-972.616	-0.036	-0.01817	0
property_type_Entire cottage	-1095.5	-0.0218	-0.00151	0
property_type_Entire guest suite	-1027.1	-0.0062	-0.00296	0
property_type_Entire guesthouse	-830.117	-0.0015	0.001842	0
property_type_Entire home/apt	-498.133	0.0042	-0.00145	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Entire house	-833.272	-0.0144	0.000909	0
property_type_Entire loft	-964.286	-0.0356	-0.01873	-0.00389
property_type_Entire place	1.63E-10	6.02E-18	0	0
property_type_Entire resort	2.1E-10	1.36E-17	0	0
property_type_Entire serviced apartment	-954.677	-0.0238	-0.00875	0
property_type_Entire townhouse	-529.82	0.0092	0.004223	0
property_type_Entire villa	-292.291	0.0309	0.024659	0.034395
property_type_Farm stay	3405.586	0.2527	0.102124	0.246021
property_type_Houseboat	-1E-11	6.92E-18	0	0
property_type_Hut	3.96E-11	-4.2E-18	0	0
property_type_Igloo	-672.648	0.0002	-0.0007	0
property_type_Kezhan	-608.004	0.0012	0.002915	0
property_type_Minsu	8.74E-11	-1.1E-17	0	0
property_type_Pension	-276.412	0.0028	0.000386	0
property_type_Private room	-5.8E-11	1.1E-17	0	0
property_type_Private room in apartment	-2.1E-11	-6.1E-20	0	0
property_type_Private room in barn	4.09E-15	-1.5E-17	0	0
property_type_Private room in bed and breakfast	6.13E-13	-1E-17	0	0
property_type_Private room in bungalow	-3599.51	-0.0002	-0.0005	0
property_type_Private room in cabin	1.14E-12	-7.7E-18	0	0
property_type_Private room in camper/rv	-2.7E-13	4.59E-18	0	0
property_type_Private room in campsite	-3.7E-13	6.8E-18	0	0
property_type_Private room in casa particular	5.28E-13	-2.3E-19	0	0
property_type_Private room in castle	4.32E-14	2.34E-18	0	0
property_type_Private room in cave	0	0	0	0
property_type_Private room in chalet	0	1.15E-31	0	0
property_type_Private room in condominium	-3813.75	-0.0035	-0.00228	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Private room in cottage	0	-5.2E-32	0	0
property_type_Private room in dome house	0	3.9E-32	0	0
property_type_Private room in earth house	0	7.15E-58	0	0
property_type_Private room in farm stay	0	-8E-48	0	0
property_type_Private room in guest suite	0	6.42E-47	0	0
property_type_Private room in guesthouse	0	0	0	0
property_type_Private room in hostel	0	-1.3E-47	0	0
property_type_Private room in house	0	5.2E-47	0	0
property_type_Private room in hut	0	-4.1E-47	0	0
property_type_Private room in kezhan	0	9.22E-48	0	0
property_type_Private room in loft	0	0	0	0
property_type_Private room in minsu	0	1.61E-47	0	0
property_type_Private room in nature lodge	0	0	0	0
property_type_Private room in resort	0	0	0	0
property_type_Private room in ryokan	0	0	0	0
property_type_Private room in serviced apartment	0	0	0	0
property_type_Private room in tent	0	0	0	0
property_type_Private room in tiny house	0	0	0	0
property_type_Private room in townhouse	0	0	0	0
property_type_Private room in treehouse	-3633.73	-0.0004	-0.00045	0
property_type_Private room in villa	0	0	0	0
property_type_Room in aparthotel	-359.002	0.0059	0.001253	0
property_type_Room in boutique hotel	4615.596	0.0352	0.013676	0.016466
property_type_Room in heritage hotel	0	0	0	0
property_type_Room in hotel	0	0	0	0
property_type_Shared room	0	0	0	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
property_type_Shared room in apartment	0	0	0	0
property_type_Shared room in bed and breakfast	0	0	0	0
property_type_Shared room in boutique hotel	0	0	0	0
property_type_Shared room in bungalow	-3350.66	0.0005	-0.00122	0
property_type_Shared room in condominium	0	0	0	0
property_type_Shared room in cottage	0	0	0	0
property_type_Shared room in earth house	0	0	0	0
property_type_Shared room in farm stay	0	0	0	0
property_type_Shared room in guest suite	0	0	0	0
property_type_Shared room in guesthouse	0	0	0	0
property_type_Shared room in hostel	0	0	0	0
property_type_Shared room in house	-3637.19	-0.0017	-0.00118	0
property_type_Shared room in hut	0	0	0	0
property_type_Shared room in kezhan	0	0	0	0
property_type_Shared room in loft	0	0	0	0
property_type_Shared room in nature lodge	0	0	0	0
property_type_Shared room in serviced apartment	-4344.89	-0.009	-0.00161	0
property_type_Shared room in tent	0	0	0	0
property_type_Shared room in tiny house	0	0	0	0
property_type_Shared room in townhouse	0	0	0	0
property_type_Shared room in villa	0	0	0	0
property_type_Tent	-2278.19	-0.0105	-0.00391	0
property_type_Tiny house	-715.761	-0.0002	-0.00103	0
property_type_Treehouse	-246.519	0.0042	0.001419	0
property_type_Yurt	0	0	0	0
room_type_Entire home/apt	-13560	0.0054	0.003114	0
room_type_Private room	-11050	-0.003	-0.00222	0

Variables	OLS	标准化 OLS	标准化 RIDGE	标准化 LASSO
room_type_Shared room	-11330	-0.0047	-0.00218	0
instant_bookable_t	-17610	0.0274	0.004551	0
instant_bookable_f	-18330	-0.0274	-0.00455	0
neighbourhood_cleansed_昌平区	-2331.08	-0.0061	-0.00168	0
neighbourhood_cleansed_朝阳区 / Chaoyang	-2086.02	0.0269	-0.00633	0
neighbourhood_cleansed_大兴区 / Daxing	-2010.88	0.0171	-0.00768	0
neighbourhood_cleansed_东城区	-2021.74	0.0257	0.001881	0
neighbourhood_cleansed_房山区	-2109.88	0.0102	-0.01001	-0.00753
neighbourhood_cleansed_丰台区 / Fengtai	-2140.27	0.0115	-0.00911	0
neighbourhood_cleansed_海淀区	-2202.86	0.0042	-0.01071	0
neighbourhood_cleansed_怀柔区 / Huairou	-1331.95	0.0757	0.071318	0.119359
neighbourhood_cleansed_门头沟区 / Mentougou	-2196.84	0.0013	-0.00263	0
neighbourhood_cleansed_密云县 / Miyun	-3482.54	-0.1245	-0.00344	0
neighbourhood_cleansed_平谷区 / Pinggu	-2372.44	-0.0035	0.004218	0
neighbourhood_cleansed_石景山区	-2134.71	0.0049	-0.00553	0
neighbourhood_cleansed_顺义区 / Shunyi	-2597.16	-0.0295	-0.01044	0
neighbourhood_cleansed_通州区 / Tongzhou	-2336.43	-0.0062	-0.00915	0
neighbourhood_cleansed_西城区	-2025.17	0.0172	-0.00148	0
neighbourhood_cleansed_延庆县 / Yanqing	-2560.94	-0.0314	0.006078	0

2)　支持向量机 Support Vector Machines

表 33　支持向量机 Support Vector Machines 法向量系数汇总

Variable	聚类 0	聚类 1	聚类 2
host_time	4.947731809	-3.175386324	8.590122152
num_of_host_verifications	-16.19541052	13.84969538	15.06821204
num_of_amenities	66.58502295	68.07692538	34.07075019
latitude	17.93190111	37.56148278	12.25799422

Variable	聚类 0	聚类 1	聚类 2
longitude	-11.13784598	29.07725991	13.91670577
host_response_rate	35.78874364	-4.076138089	-16.45981276
host_acceptance_rate	22.3933836	6.657175992	18.30678113
host_listings_count	28.65886112	-18.36112783	0.736774326
accommodates	228.2025771	286.4187229	91.97115144
bathrooms	145.7123066	183.3346458	75.28389271
bedrooms	189.1019799	138.1861846	120.7075062
beds	115.7060003	125.1752108	96.07803196
minimum_nights	-37.7847189	-36.37955819	-20.43770563
maximum_nights	-16.55736859	-12.61845776	-41.69245346
availability_30	1.662896495	22.41434661	21.1165703
availability_60	14.81244788	18.20404942	16.11306381
availability_90	24.58513005	10.76972356	4.115276521
availability_365	24.54498851	-26.13152755	-20.51318015
number_of_reviews	-50.03063399	-50.20028422	-37.42279689
number_of_reviews_ltm	-23.21345686	-67.4204772	-36.80294621
calculated_host_listings_count	11.72930477	4.658374339	-0.73737462
calculated_host_listings_count_entire_homes	1.595104561	10.63017638	0.022174259
calculated_host_listings_count_private_rooms	44.45713376	4.935992007	-4.012010383
calculated_host_listings_count_shared_rooms	2.901426042	-23.85125574	1.047382701
host_response_time_a few days or more	-25.38165908	-6.264500893	-3.140495049
host_response_time_within a day	-11.23812436	28.88032779	4.870079337
host_response_time_within a few hours	32.66277937	17.27124203	-34.78121656
host_response_time_within an hour	3.706024244	-24.53507161	29.47708249
host_is_superhost_t	3.983829216	-39.47935969	-10.84122703
host_is_superhost_f	-3.983829216	39.47935969	10.84122703
host_has_profile_pic_t	22.95502172	36.77120066	0

Variable	聚类 0	聚类 1	聚类 2
host_has_profile_pic_f	-22.95502172	-36.77120066	0
host_identity_verified_t	21.82778398	5.403375984	-69.46312901
host_identity_verified_f	-21.82778398	-5.403375984	69.46312901
property_type_Barn	26.06185148	0	0
property_type_Camper/RV	-10.32819165	0	-12.51671306
property_type_Campsite	14.72606546	0	-8.336598672
property_type_Casa particular	-44.99429629	0	0
property_type_Castle	-10.47998583	331.7772173	0
property_type_Cave	0	0	0
property_type_Dome house	-70.48609356	0	0
property_type_Earth house	66.73236157	0	-3.244814088
property_type_Entire apartment	-10.12536137	-10.69326041	-8.991223166
property_type_Entire bed and breakfast	-27.65383781	0	-9.95471E-15
property_type_Entire bungalow	18.87231315	36.05594225	17.54830954
property_type_Entire cabin	-7.439951978	0	2.589174598
property_type_Entire chalet	-2.382666726	0	0
property_type_Entire condominium	-19.24471599	-12.23755949	-13.77787257
property_type_Entire cottage	-11.39548659	48.25893171	5.551604821
property_type_Entire guest suite	3.996108848	0	-9.907839195
property_type_Entire guesthouse	-4.664989226	0	25.70151603
property_type_Entire home/apt	-62.99644047	0	-6.771153514
property_type_Entire house	27.99387082	17.24735287	7.355131318
property_type_Entire loft	7.911527368	-13.83871678	-6.610020761
property_type_Entire place	-34.18309071	0	0
property_type_Entire resort	0.826053338	0	0
property_type_Entire serviced apartment	-32.75924181	0	-7.020628124
property_type_Entire townhouse	9.214350418	14.28734297	2.740925842

Variable	聚类 0	聚类 1	聚类 2
property_type_Entire villa	46.42472674	98.00924309	35.6314763
property_type_Farm stay	-8.318346452	25.6071514	12.89285252
property_type_Houseboat	-46.10117287	0	0
property_type_Hut	-21.04665857	0	0
property_type_Igloo	-51.32649613	0	-22.95058762
property_type_Kezhan	22.27907953	0	22.58665416
property_type_Minsu	63.42922765	0	0
property_type_Pension	0	0	-2.953781266
property_type_Private room	0	-15.3955214	0
property_type_Private room in apartment	0	-5.138893517	0
property_type_Private room in barn	0	-24.51614376	0
property_type_Private room in bed and breakfast	0	-11.6189373	0
property_type_Private room in bungalow	0	-3.9906639	-18.10308919
property_type_Private room in cabin	0	45.00744342	0
property_type_Private room in camper/rv	0	-17.516463	0
property_type_Private room in campsite	0	-26.77261961	0
property_type_Private room in casa particular	0	-24.50650305	0
property_type_Private room in castle	0	3.994321568	0
property_type_Private room in cave	0	-6.401700895	0
property_type_Private room in chalet	0	-7.428110074	0
property_type_Private room in condominium	0	-9.273848614	-22.05337926
property_type_Private room in cottage	-28.76173564	-8.143138111	0
property_type_Private room in dome house	0	-39.18428689	0
property_type_Private room in earth house	0	83.72055113	0
property_type_Private room in farm stay	0	-55.48500879	0
property_type_Private room in guest suite	0	-12.21095636	0
property_type_Private room in guesthouse	0	-8.959280698	0

Variable	聚类 0	聚类 1	聚类 2
property_type_Private room in hostel	0	-29.59793825	0
property_type_Private room in house	0	-7.760290725	0
property_type_Private room in hut	0	19.45156696	0
property_type_Private room in kezhan	0	21.48358662	0
property_type_Private room in loft	0	-9.191326301	0
property_type_Private room in minsu	0	-13.37040172	0
property_type_Private room in nature lodge	0	31.1953381	0
property_type_Private room in resort	0	5.53420556	0
property_type_Private room in ryokan	0	146.8361945	0
property_type_Private room in serviced apartment	-65.26077832	-9.013747343	0
property_type_Private room in tent	0	-2.529118352	0
property_type_Private room in tiny house	-50.16534424	-18.86790372	0
property_type_Private room in townhouse	0	11.71371906	0
property_type_Private room in treehouse	0	0	-24.92824579
property_type_Private room in villa	0	8.813736287	0
property_type_Room in aparthotel	5.852049209	-5.291285785	-12.44514234
property_type_Room in boutique hotel	5.811079085	54.4774379	130.9917229
property_type_Room in heritage hotel	0	-43.02255241	0
property_type_Room in hotel	0	14.92003975	0
property_type_Shared room	0	0	0
property_type_Shared room in apartment	0	-5.74841395	0
property_type_Shared room in bed and breakfast	-42.67565699	23.20127662	0
property_type_Shared room in boutique hotel	0	288.0475495	0
property_type_Shared room in bungalow	0	-15.89837122	-12.19499978
property_type_Shared room in condominium	0	-10.27154297	0
property_type_Shared room in cottage	10.49796724	-19.54820335	0

Variable	聚类 0	聚类 1	聚类 2
property_type_Shared room in earth house	0	96.22042093	0
property_type_Shared room in farm stay	0	0.187063626	0
property_type_Shared room in guest suite	0	-43.27668949	0
property_type_Shared room in guesthouse	0	-22.24202082	0
property_type_Shared room in hostel	0	-2.938503026	0
property_type_Shared room in house	0	2.666508313	-26.68620717
property_type_Shared room in hut	0	0	0
property_type_Shared room in kezhan	0	-36.16869361	0
property_type_Shared room in loft	0	-20.14027147	0
property_type_Shared room in nature lodge	0	0	0
property_type_Shared room in serviced apartment	0	16.99213718	-15.62488111
property_type_Shared room in tent	0	-40.23607245	0
property_type_Shared room in tiny house	0	0	0
property_type_Shared room in townhouse	0	11.64058547	0
property_type_Shared room in villa	0	0.336820183	0
property_type_Tent	-26.60936122	0	-25.41173306
property_type_Tiny house	-0.56698048	0	-18.63750761
property_type_Treehouse	0	0	-1.43592E-14
property_type_Yurt	0	0	0
room_type_Entire home/apt	67.47460575	94.57432193	44.06001642
room_type_Private room	-63.8585064	-24.65563962	-35.32513591
room_type_Shared room	-22.7541831	-3.482043978	-26.90983741
instant_bookable_t	8.977973424	9.51539558	13.48177898
instant_bookable_f	-8.977973424	-9.51539558	-13.48177898
neighbourhood_cleansed_昌平区	32.45163571	44.82252299	12.04186067
neighbourhood_cleansed_朝阳区 / Chaoyang	-0.267906421	-52.79194677	4.065911635
neighbourhood_cleansed_大兴区 / Daxing	-6.156114289	-11.50221596	-6.173320089

Variable	聚类 0	聚类 1	聚类 2
neighbourhood_cleansed_东城区	-1.620290934	16.34349082	19.05410659
neighbourhood_cleansed_房山区	2.462435963	-5.807434236	-11.93832158
neighbourhood_cleansed_丰台区 / Fengtai	-29.06723692	-20.93164233	-7.974269683
neighbourhood_cleansed_海淀区	-36.89031544	-14.21165621	-12.20146706
neighbourhood_cleansed_怀柔区 / Huairou	30.55565969	32.74220115	19.32595194
neighbourhood_cleansed_门头沟区 / Mentougou	-4.384664543	-22.0168479	-1.2014574
neighbourhood_cleansed_密云县 / Miyun	-23.12398126	-0.845045559	-12.81522428
neighbourhood_cleansed_平谷区 / Pinggu	16.02255073	95.93549973	9.498146466
neighbourhood_cleansed_石景山区	-11.48406286	-7.174944788	-7.670835146
neighbourhood_cleansed_顺义区 / Shunyi	-1.800407015	6.74501065	0.852389481
neighbourhood_cleansed_通州区 / Tongzhou	1.439596281	-34.56954524	-3.60481204
neighbourhood_cleansed_西城区	9.146614036	42.23048053	16.8383161
neighbourhood_cleansed_延庆县 / Yanqing	35.90757603	-13.71761964	-16.31506461

3)　决策树 Decision Tree

表 34　决策树 Decision Tree 重要性汇总

Variables	聚类 0	聚类 1	聚类 2
host_time	0.96%	3.14%	0.12%
num_of_host_verifications	0.53%	3.66%	0.02%
num_of_amenities	1.06%	1.85%	0.43%
latitude	22.70%	5.95%	0.27%
longitude	6.34%	5.83%	0.07%
host_response_rate	0.17%	1.59%	0.01%
host_acceptance_rate	0.74%	0.78%	0.01%
host_listings_count	0.92%	1.48%	0.02%
accommodates	20.05%	16.38%	7.81%
bathrooms	8.80%	13.33%	9.50%

Variables	聚类 0	聚类 1	聚类 2
bedrooms	26.62%	2.13%	0.07%
beds	0.37%	2.33%	1.94%
minimum_nights	0.14%	0.12%	0.00%
maximum_nights	0.00%	0.39%	56.51%
availability_30	0.12%	0.38%	0.10%
availability_60	0.34%	0.52%	0.01%
availability_90	0.88%	0.41%	0.01%
availability_365	0.70%	1.21%	0.04%
number_of_reviews	1.09%	0.59%	0.00%
number_of_reviews_ltm	0.08%	0.03%	0.39%
calculated_host_listings_count	1.31%	6.73%	0.01%
calculated_host_listings_count_entire_homes	0.90%	3.22%	0.00%
calculated_host_listings_count_private_rooms	0.70%	2.46%	0.00%
calculated_host_listings_count_shared_rooms	0.00%	4.71%	0.00%
host_response_time_a few days or more	0.01%	0.00%	0.00%
host_response_time_within a day	0.00%	0.34%	0.00%
host_response_time_within a few hours	0.01%	0.00%	1.60%
host_response_time_within an hour	0.19%	0.10%	0.00%
host_is_superhost_t	0.00%	0.23%	0.00%
host_is_superhost_f	0.00%	0.57%	0.00%
host_has_profile_pic_t	0.00%	0.00%	0.00%
host_has_profile_pic_f	0.00%	0.00%	0.00%
host_identity_verified_t	0.00%	0.00%	0.00%
host_identity_verified_f	0.00%	0.00%	0.00%
property_type_Barn	0.00%	0.00%	0.00%
property_type_Camper/RV	0.00%	0.00%	0.00%
property_type_Campsite	0.00%	0.00%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Casa particular	0.00%	0.00%	0.00%
property_type_Castle	0.00%	0.00%	0.00%
property_type_Cave	0.00%	0.00%	0.00%
property_type_Dome house	0.00%	0.00%	0.00%
property_type_Earth house	0.00%	0.00%	0.00%
property_type_Entire apartment	0.04%	0.00%	0.04%
property_type_Entire bed and breakfast	0.00%	0.00%	0.00%
property_type_Entire bungalow	0.00%	0.00%	0.01%
property_type_Entire cabin	0.00%	0.00%	0.00%
property_type_Entire chalet	0.00%	0.00%	0.00%
property_type_Entire condominium	0.00%	0.00%	0.00%
property_type_Entire cottage	0.10%	0.00%	0.00%
property_type_Entire guest suite	0.00%	0.00%	0.00%
property_type_Entire guesthouse	0.00%	0.00%	0.00%
property_type_Entire home/apt	0.00%	0.00%	0.00%
property_type_Entire house	0.09%	0.00%	0.00%
property_type_Entire loft	0.02%	0.00%	0.42%
property_type_Entire place	0.00%	0.00%	0.00%
property_type_Entire resort	0.00%	0.00%	0.00%
property_type_Entire serviced apartment	0.05%	0.00%	0.00%
property_type_Entire townhouse	0.00%	0.00%	0.03%
property_type_Entire villa	2.38%	0.00%	0.00%
property_type_Farm stay	0.00%	0.00%	19.90%
property_type_Houseboat	0.00%	0.00%	0.00%
property_type_Hut	0.00%	0.00%	0.00%
property_type_Igloo	0.00%	0.00%	0.00%
property_type_Kezhan	0.00%	0.00%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Minsu	0.00%	0.00%	0.00%
property_type_Pension	0.00%	0.00%	0.00%
property_type_Private room	0.00%	0.00%	0.00%
property_type_Private room in apartment	0.00%	0.79%	0.00%
property_type_Private room in barn	0.00%	0.00%	0.00%
property_type_Private room in bed and breakfast	0.00%	0.00%	0.00%
property_type_Private room in bungalow	0.00%	0.00%	0.00%
property_type_Private room in cabin	0.00%	0.00%	0.00%
property_type_Private room in camper/rv	0.00%	0.00%	0.00%
property_type_Private room in campsite	0.00%	0.00%	0.00%
property_type_Private room in casa particular	0.00%	0.00%	0.00%
property_type_Private room in castle	0.00%	0.00%	0.00%
property_type_Private room in cave	0.00%	0.00%	0.00%
property_type_Private room in chalet	0.00%	0.00%	0.00%
property_type_Private room in condominium	0.00%	0.00%	0.00%
property_type_Private room in cottage	0.00%	0.50%	0.00%
property_type_Private room in dome house	0.00%	0.00%	0.00%
property_type_Private room in earth house	0.00%	0.00%	0.00%
property_type_Private room in farm stay	0.00%	4.40%	0.00%
property_type_Private room in guest suite	0.00%	0.00%	0.00%
property_type_Private room in guesthouse	0.00%	0.00%	0.00%
property_type_Private room in hostel	0.00%	0.00%	0.00%
property_type_Private room in house	0.00%	0.09%	0.00%
property_type_Private room in hut	0.00%	0.00%	0.00%
property_type_Private room in kezhan	0.00%	0.00%	0.00%
property_type_Private room in loft	0.00%	0.00%	0.00%
property_type_Private room in minsu	0.00%	0.00%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Private room in nature lodge	0.00%	0.00%	0.00%
property_type_Private room in resort	0.00%	0.74%	0.00%
property_type_Private room in ryokan	0.00%	0.00%	0.00%
property_type_Private room in serviced apartment	0.00%	0.69%	0.00%
property_type_Private room in tent	0.00%	0.00%	0.00%
property_type_Private room in tiny house	0.00%	0.00%	0.00%
property_type_Private room in townhouse	0.00%	0.00%	0.00%
property_type_Private room in treehouse	0.00%	0.00%	0.00%
property_type_Private room in villa	0.00%	0.41%	0.00%
property_type_Room in aparthotel	0.00%	0.00%	0.00%
property_type_Room in boutique hotel	0.00%	1.93%	0.00%
property_type_Room in heritage hotel	0.00%	0.00%	0.00%
property_type_Room in hotel	0.00%	1.31%	0.00%
property_type_Shared room	0.00%	0.00%	0.00%
property_type_Shared room in apartment	0.00%	0.05%	0.00%
property_type_Shared room in bed and breakfast	0.00%	0.00%	0.00%
property_type_Shared room in boutique hotel	0.00%	0.00%	0.00%
property_type_Shared room in bungalow	0.00%	0.00%	0.00%
property_type_Shared room in condominium	0.00%	0.00%	0.00%
property_type_Shared room in cottage	0.00%	0.00%	0.00%
property_type_Shared room in earth house	0.00%	0.00%	0.00%
property_type_Shared room in farm stay	0.00%	0.00%	0.00%
property_type_Shared room in guest suite	0.00%	0.00%	0.00%
property_type_Shared room in guesthouse	0.00%	0.00%	0.00%
property_type_Shared room in hostel	0.00%	0.00%	0.00%
property_type_Shared room in house	0.00%	0.01%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Shared room in hut	0.00%	0.00%	0.00%
property_type_Shared room in kezhan	0.00%	0.00%	0.00%
property_type_Shared room in loft	0.00%	0.00%	0.00%
property_type_Shared room in nature lodge	0.00%	0.00%	0.00%
property_type_Shared room in serviced apartment	0.00%	0.00%	0.00%
property_type_Shared room in tent	0.00%	0.00%	0.00%
property_type_Shared room in tiny house	0.00%	0.00%	0.00%
property_type_Shared room in townhouse	0.00%	0.00%	0.00%
property_type_Shared room in villa	0.00%	0.00%	0.00%
property_type_Tent	0.00%	0.00%	0.00%
property_type_Tiny house	0.00%	0.00%	0.00%
property_type_Treehouse	0.00%	0.00%	0.00%
property_type_Yurt	0.00%	0.00%	0.00%
room_type_Entire home/apt	0.00%	0.00%	0.00%
room_type_Private room	0.00%	0.13%	0.00%
room_type_Shared room	0.00%	1.11%	0.00%
instant_bookable_t	0.04%	0.05%	0.02%
instant_bookable_f	0.08%	0.01%	0.00%
neighbourhood_cleansed_昌平区	0.06%	0.01%	0.01%
neighbourhood_cleansed_朝阳区 / Chaoyang	0.57%	0.67%	0.06%
neighbourhood_cleansed_大兴区 / Daxing	0.02%	0.00%	0.00%
neighbourhood_cleansed_东城区	0.00%	2.89%	0.04%
neighbourhood_cleansed_房山区	0.07%	0.00%	0.00%
neighbourhood_cleansed_丰台区 / Fengtai	0.08%	0.00%	0.00%
neighbourhood_cleansed_海淀区	0.00%	0.00%	0.00%
neighbourhood_cleansed_怀柔区 / Huairou	0.09%	3.14%	0.11%
neighbourhood_cleansed_门头沟区 / Mentougou	0.00%	0.00%	0.00%

Variables	聚类 0	聚类 1	聚类 2
neighbourhood_cleansed_密云县 / Miyun	0.54%	0.00%	0.38%
neighbourhood_cleansed_平谷区 / Pinggu	0.00%	0.00%	0.00%
neighbourhood_cleansed_石景山区	0.00%	0.00%	0.00%
neighbourhood_cleansed_顺义区 / Shunyi	0.00%	0.00%	0.00%
neighbourhood_cleansed_通州区 / Tongzhou	0.03%	0.00%	0.01%
neighbourhood_cleansed_西城区	0.01%	0.57%	0.03%
neighbourhood_cleansed_延庆县 / Yanqing	0.00%	0.00%	0.00%
总计	**100.00%**	**100.00%**	**100.00%**

4)　随机森林 Random Forests

表 35 随机森林 Random Forests 重要性汇总

Variable	聚类 0	聚类 1	聚类 2
host_time	3.12%	4.32%	2.38%
num_of_host_verifications	1.65%	2.15%	1.99%
num_of_amenities	2.90%	3.07%	2.88%
latitude	7.19%	19.25%	8.12%
longitude	6.00%	3.68%	7.30%
host_response_rate	1.47%	1.70%	1.26%
host_acceptance_rate	1.38%	2.19%	0.48%
host_listings_count	2.71%	2.53%	6.60%
accommodates	12.98%	9.28%	6.55%
bathrooms	9.94%	6.57%	3.88%
bedrooms	11.17%	6.79%	4.81%
beds	8.17%	4.67%	4.81%
minimum_nights	0.47%	0.15%	1.89%
maximum_nights	0.90%	0.94%	5.78%
availability_30	1.18%	1.30%	2.78%

Variable	聚类 0	聚类 1	聚类 2
availability_60	1.25%	1.15%	2.96%
availability_90	1.35%	1.20%	3.13%
availability_365	1.82%	1.87%	3.33%
number_of_reviews	1.28%	1.67%	1.80%
number_of_reviews_ltm	0.91%	0.67%	0.96%
calculated_host_listings_count	2.58%	2.53%	2.28%
calculated_host_listings_count_entire_homes	1.94%	1.23%	2.28%
calculated_host_listings_count_private_rooms	2.24%	2.53%	0.89%
calculated_host_listings_count_shared_rooms	0.14%	1.23%	0.01%
host_response_time_a few days or more	0.27%	0.44%	0.00%
host_response_time_within a day	0.26%	0.13%	0.05%
host_response_time_within a few hours	0.36%	0.39%	0.56%
host_response_time_within an hour	0.45%	0.48%	0.46%
host_is_superhost_t	0.23%	0.15%	0.00%
host_is_superhost_f	0.22%	0.22%	0.00%
host_has_profile_pic_t	0.00%	0.00%	0.00%
host_has_profile_pic_f	0.00%	0.00%	0.00%
host_identity_verified_t	0.00%	0.02%	0.00%
host_identity_verified_f	0.00%	0.02%	0.00%
property_type_Barn	0.00%	0.00%	0.00%
property_type_Camper/RV	0.00%	0.00%	0.00%
property_type_Campsite	0.00%	0.00%	0.05%
property_type_Casa particular	0.00%	0.00%	0.00%
property_type_Castle	0.00%	0.43%	0.00%
property_type_Cave	0.00%	0.00%	0.00%
property_type_Dome house	0.00%	0.00%	0.00%
property_type_Earth house	0.00%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Entire apartment	0.59%	0.00%	0.42%
property_type_Entire bed and breakfast	0.00%	0.00%	0.00%
property_type_Entire bungalow	0.20%	0.00%	0.70%
property_type_Entire cabin	0.00%	0.00%	0.00%
property_type_Entire chalet	0.00%	0.00%	0.00%
property_type_Entire condominium	0.19%	0.00%	0.06%
property_type_Entire cottage	0.74%	0.02%	0.14%
property_type_Entire guest suite	0.20%	0.00%	0.00%
property_type_Entire guesthouse	0.02%	0.00%	0.00%
property_type_Entire home/apt	0.00%	0.00%	0.00%
property_type_Entire house	0.40%	0.02%	0.40%
property_type_Entire loft	0.36%	0.00%	0.09%
property_type_Entire place	0.00%	0.00%	0.00%
property_type_Entire resort	0.00%	0.00%	0.00%
property_type_Entire serviced apartment	0.20%	0.00%	0.02%
property_type_Entire townhouse	0.19%	0.00%	0.05%
property_type_Entire villa	2.53%	0.01%	1.00%
property_type_Farm stay	0.42%	0.00%	8.43%
property_type_Houseboat	0.00%	0.00%	0.00%
property_type_Hut	0.00%	0.00%	0.00%
property_type_Igloo	0.00%	0.00%	0.00%
property_type_Kezhan	0.01%	0.00%	0.00%
property_type_Minsu	0.03%	0.00%	0.00%
property_type_Pension	0.00%	0.00%	0.00%
property_type_Private room	0.07%	1.38%	0.00%
property_type_Private room in apartment	0.00%	0.53%	0.00%
property_type_Private room in barn	0.00%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Private room in bed and breakfast	0.00%	0.08%	0.00%
property_type_Private room in bungalow	0.00%	0.12%	0.00%
property_type_Private room in cabin	0.00%	0.00%	0.00%
property_type_Private room in camper/rv	0.00%	0.00%	0.00%
property_type_Private room in campsite	0.00%	0.00%	0.00%
property_type_Private room in casa particular	0.00%	0.00%	0.00%
property_type_Private room in castle	0.00%	0.00%	0.00%
property_type_Private room in cave	0.00%	0.00%	0.00%
property_type_Private room in chalet	0.00%	0.00%	0.00%
property_type_Private room in condominium	0.00%	0.15%	0.00%
property_type_Private room in cottage	0.07%	0.12%	0.00%
property_type_Private room in dome house	0.00%	0.00%	0.00%
property_type_Private room in earth house	0.00%	0.00%	0.00%
property_type_Private room in farm stay	0.00%	1.38%	0.00%
property_type_Private room in guest suite	0.00%	0.00%	0.00%
property_type_Private room in guesthouse	0.00%	0.04%	0.00%
property_type_Private room in hostel	0.00%	0.07%	0.00%
property_type_Private room in house	0.00%	0.15%	0.00%
property_type_Private room in hut	0.00%	0.00%	0.00%
property_type_Private room in kezhan	0.00%	0.46%	0.00%
property_type_Private room in loft	0.00%	0.39%	0.00%
property_type_Private room in minsu	0.00%	0.00%	0.00%
property_type_Private room in nature lodge	0.00%	0.28%	0.00%
property_type_Private room in resort	0.00%	0.23%	0.00%
property_type_Private room in ryokan	0.00%	0.00%	0.00%
property_type_Private room in serviced apartment	0.00%	0.14%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Private room in tent	0.00%	0.00%	0.00%
property_type_Private room in tiny house	0.00%	0.00%	0.00%
property_type_Private room in townhouse	0.00%	0.04%	0.00%
property_type_Private room in treehouse	0.00%	0.00%	0.00%
property_type_Private room in villa	0.00%	0.32%	0.00%
property_type_Room in aparthotel	0.04%	0.18%	0.00%
property_type_Room in boutique hotel	0.01%	1.94%	0.00%
property_type_Room in heritage hotel	0.00%	0.00%	0.00%
property_type_Room in hotel	0.00%	0.11%	0.00%
property_type_Shared room	0.00%	0.12%	0.00%
property_type_Shared room in apartment	0.00%	0.12%	0.00%
property_type_Shared room in bed and breakfast	0.00%	0.00%	0.00%
property_type_Shared room in boutique hotel	0.00%	0.12%	0.00%
property_type_Shared room in bungalow	0.00%	0.00%	0.00%
property_type_Shared room in condominium	0.00%	0.00%	0.00%
property_type_Shared room in cottage	0.00%	0.00%	0.00%
property_type_Shared room in earth house	0.00%	0.00%	0.00%
property_type_Shared room in farm stay	0.00%	0.00%	0.00%
property_type_Shared room in guest suite	0.00%	0.00%	0.00%
property_type_Shared room in guesthouse	0.00%	0.00%	0.00%
property_type_Shared room in hostel	0.00%	0.00%	0.00%
property_type_Shared room in house	0.00%	0.05%	0.00%
property_type_Shared room in hut	0.00%	0.00%	0.00%
property_type_Shared room in kezhan	0.00%	0.00%	0.00%
property_type_Shared room in loft	0.00%	0.00%	0.00%
property_type_Shared room in nature lodge	0.00%	0.00%	0.00%
property_type_Shared room in serviced apartment	0.00%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Shared room in tent	0.00%	0.00%	0.00%
property_type_Shared room in tiny house	0.00%	0.00%	0.00%
property_type_Shared room in townhouse	0.00%	0.00%	0.00%
property_type_Shared room in villa	0.00%	0.00%	0.00%
property_type_Tent	0.01%	0.00%	0.00%
property_type_Tiny house	0.00%	0.00%	0.00%
property_type_Treehouse	0.00%	0.00%	0.00%
property_type_Yurt	0.00%	0.00%	0.00%
room_type_Entire home/apt	0.08%	0.20%	0.00%
room_type_Private room	0.08%	0.21%	0.00%
room_type_Shared room	0.00%	0.29%	0.00%
instant_bookable_t	0.43%	0.45%	0.01%
instant_bookable_f	0.43%	0.50%	0.13%
neighbourhood_cleansed_昌平区	0.58%	0.10%	0.03%
neighbourhood_cleansed_朝阳区 / Chaoyang	0.28%	0.79%	0.11%
neighbourhood_cleansed_大兴区 / Daxing	0.18%	1.43%	0.00%
neighbourhood_cleansed_东城区	0.35%	0.50%	0.16%
neighbourhood_cleansed_房山区	0.29%	0.12%	0.02%
neighbourhood_cleansed_丰台区 / Fengtai	0.14%	0.02%	0.02%
neighbourhood_cleansed_海淀区	0.13%	0.09%	0.02%
neighbourhood_cleansed_怀柔区 / Huairou	1.36%	0.76%	7.16%
neighbourhood_cleansed_门头沟区 / Mentougou	0.08%	0.07%	0.00%
neighbourhood_cleansed_密云县 / Miyun	0.70%	0.17%	0.95%
neighbourhood_cleansed_平谷区 / Pinggu	0.03%	0.36%	0.00%
neighbourhood_cleansed_石景山区	0.03%	0.00%	0.00%
neighbourhood_cleansed_顺义区 / Shunyi	0.29%	0.01%	0.01%
neighbourhood_cleansed_通州区 / Tongzhou	0.27%	0.04%	0.02%

Variable	聚类 0	聚类 1	聚类 2
neighbourhood_cleansed_西城区	0.14%	2.26%	0.04%
neighbourhood_cleansed_延庆县 / Yanqing	1.36%	0.24%	0.24%
总计	**100.00%**	**100.00%**	**100.00%**

5)　极端随机树 Extremely Randomized Trees

表 35　极端随机树 Extremely Randomized Trees 重要性汇总

Variable	聚类 0	聚类 1	聚类 2
host_time	1.59%	2.21%	0.58%
num_of_host_verifications	1.45%	2.07%	1.59%
num_of_amenities	1.61%	2.10%	1.04%
latitude	3.68%	4.29%	1.57%
longitude	3.08%	2.64%	2.06%
host_response_rate	1.52%	2.00%	0.60%
host_acceptance_rate	1.57%	1.91%	0.38%
host_listings_count	2.05%	2.60%	4.05%
accommodates	15.27%	8.59%	6.02%
bathrooms	10.17%	4.63%	4.61%
bedrooms	6.73%	3.78%	2.11%
beds	4.94%	2.94%	1.45%
minimum_nights	0.94%	0.57%	23.66%
maximum_nights	1.25%	1.35%	19.77%
availability_30	0.99%	1.47%	0.81%
availability_60	0.96%	1.42%	0.68%
availability_90	0.96%	1.41%	1.00%
availability_365	1.13%	1.64%	0.44%
number_of_reviews	1.28%	1.17%	0.24%
number_of_reviews_ltm	0.89%	0.91%	0.16%

Variable	聚类 0	聚类 1	聚类 2
calculated_host_listings_count	2.12%	2.79%	1.03%
calculated_host_listings_count_entire_homes	1.94%	2.21%	2.33%
calculated_host_listings_count_private_rooms	2.24%	3.25%	0.69%
calculated_host_listings_count_shared_rooms	0.55%	0.99%	0.01%
host_response_time_a few days or more	0.68%	0.67%	0.00%
host_response_time_within a day	0.73%	0.68%	0.02%
host_response_time_within a few hours	0.90%	0.98%	1.09%
host_response_time_within an hour	0.90%	1.13%	0.41%
host_is_superhost_t	0.46%	0.88%	0.02%
host_is_superhost_f	0.45%	0.90%	0.01%
host_has_profile_pic_t	0.01%	0.00%	0.00%
host_has_profile_pic_f	0.01%	0.00%	0.00%
host_identity_verified_t	0.01%	0.02%	0.04%
host_identity_verified_f	0.01%	0.02%	0.04%
property_type_Barn	0.00%	0.00%	0.00%
property_type_Camper/RV	0.14%	0.00%	0.00%
property_type_Campsite	0.10%	0.00%	0.04%
property_type_Casa particular	0.00%	0.00%	0.00%
property_type_Castle	0.03%	0.43%	0.00%
property_type_Cave	0.00%	0.00%	0.00%
property_type_Dome house	0.00%	0.00%	0.00%
property_type_Earth house	0.13%	0.00%	0.00%
property_type_Entire apartment	0.99%	0.02%	0.09%
property_type_Entire bed and breakfast	0.01%	0.00%	0.00%
property_type_Entire bungalow	0.48%	0.02%	0.99%
property_type_Entire cabin	0.13%	0.00%	0.00%
property_type_Entire chalet	0.01%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Entire condominium	0.47%	0.00%	0.03%
property_type_Entire cottage	1.37%	0.14%	0.05%
property_type_Entire guest suite	0.13%	0.00%	0.00%
property_type_Entire guesthouse	0.12%	0.00%	0.01%
property_type_Entire home/apt	0.00%	0.00%	0.00%
property_type_Entire house	0.69%	0.05%	0.15%
property_type_Entire loft	0.77%	0.02%	0.09%
property_type_Entire place	0.01%	0.00%	0.00%
property_type_Entire resort	0.01%	0.00%	0.00%
property_type_Entire serviced apartment	0.43%	0.00%	0.01%
property_type_Entire townhouse	0.42%	0.01%	0.06%
property_type_Entire villa	5.01%	0.05%	0.96%
property_type_Farm stay	0.76%	0.03%	12.27%
property_type_Houseboat	0.00%	0.00%	0.00%
property_type_Hut	0.00%	0.00%	0.00%
property_type_Igloo	0.00%	0.00%	0.00%
property_type_Kezhan	0.15%	0.00%	0.00%
property_type_Minsu	0.16%	0.00%	0.00%
property_type_Pension	0.00%	0.00%	0.00%
property_type_Private room	0.00%	0.07%	0.00%
property_type_Private room in apartment	0.00%	1.48%	0.00%
property_type_Private room in barn	0.00%	0.01%	0.00%
property_type_Private room in bed and breakfast	0.00%	0.49%	0.00%
property_type_Private room in bungalow	0.00%	0.71%	0.00%
property_type_Private room in cabin	0.00%	0.04%	0.00%
property_type_Private room in camper/rv	0.00%	0.00%	0.00%
property_type_Private room in campsite	0.00%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Private room in casa particular	0.00%	0.00%	0.00%
property_type_Private room in castle	0.00%	0.11%	0.00%
property_type_Private room in cave	0.00%	0.01%	0.00%
property_type_Private room in chalet	0.00%	0.01%	0.00%
property_type_Private room in condominium	0.00%	0.72%	0.00%
property_type_Private room in cottage	0.12%	0.44%	0.00%
property_type_Private room in dome house	0.00%	0.00%	0.00%
property_type_Private room in earth house	0.00%	0.13%	0.00%
property_type_Private room in farm stay	0.00%	2.12%	0.00%
property_type_Private room in guest suite	0.00%	0.07%	0.00%
property_type_Private room in guesthouse	0.00%	0.14%	0.00%
property_type_Private room in hostel	0.00%	0.22%	0.00%
property_type_Private room in house	0.00%	0.64%	0.00%
property_type_Private room in hut	0.00%	0.04%	0.00%
property_type_Private room in kezhan	0.00%	1.75%	0.00%
property_type_Private room in loft	0.00%	0.36%	0.00%
property_type_Private room in minsu	0.00%	0.02%	0.00%
property_type_Private room in nature lodge	0.00%	0.50%	0.00%
property_type_Private room in resort	0.00%	0.94%	0.00%
property_type_Private room in ryokan	0.00%	0.05%	0.00%
property_type_Private room in serviced apartment	0.00%	1.10%	0.00%
property_type_Private room in tent	0.00%	0.02%	0.00%
property_type_Private room in tiny house	0.00%	0.03%	0.00%
property_type_Private room in townhouse	0.00%	0.31%	0.00%
property_type_Private room in treehouse	0.00%	0.00%	0.00%
property_type_Private room in villa	0.00%	0.66%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Room in aparthotel	0.04%	0.17%	0.00%
property_type_Room in boutique hotel	0.01%	2.36%	0.04%
property_type_Room in heritage hotel	0.00%	0.04%	0.00%
property_type_Room in hotel	0.00%	0.86%	0.00%
property_type_Shared room	0.00%	0.00%	0.00%
property_type_Shared room in apartment	0.00%	0.35%	0.00%
property_type_Shared room in bed and breakfast	0.00%	0.04%	0.00%
property_type_Shared room in boutique hotel	0.00%	0.38%	0.00%
property_type_Shared room in bungalow	0.00%	0.01%	0.00%
property_type_Shared room in condominium	0.00%	0.06%	0.00%
property_type_Shared room in cottage	0.00%	0.02%	0.00%
property_type_Shared room in earth house	0.00%	0.03%	0.00%
property_type_Shared room in farm stay	0.00%	0.01%	0.00%
property_type_Shared room in guest suite	0.00%	0.00%	0.00%
property_type_Shared room in guesthouse	0.00%	0.00%	0.00%
property_type_Shared room in hostel	0.00%	0.05%	0.00%
property_type_Shared room in house	0.00%	0.05%	0.00%
property_type_Shared room in hut	0.00%	0.00%	0.00%
property_type_Shared room in kezhan	0.00%	0.01%	0.00%
property_type_Shared room in loft	0.00%	0.03%	0.00%
property_type_Shared room in nature lodge	0.00%	0.00%	0.00%
property_type_Shared room in serviced apartment	0.00%	0.07%	0.00%
property_type_Shared room in tent	0.00%	0.00%	0.00%
property_type_Shared room in tiny house	0.00%	0.00%	0.00%
property_type_Shared room in townhouse	0.00%	0.03%	0.00%
property_type_Shared room in villa	0.00%	0.02%	0.00%
property_type_Tent	0.01%	0.00%	0.00%

Variable	聚类 0	聚类 1	聚类 2
property_type_Tiny house	0.06%	0.00%	0.00%
property_type_Treehouse	0.00%	0.00%	0.00%
property_type_Yurt	0.00%	0.00%	0.00%
room_type_Entire home/apt	0.11%	0.16%	0.00%
room_type_Private room	0.11%	1.16%	0.00%
room_type_Shared room	0.00%	1.66%	0.00%
instant_bookable_t	0.71%	0.89%	0.27%
instant_bookable_f	0.68%	0.85%	0.26%
neighbourhood_cleansed_昌平区	0.58%	0.50%	0.05%
neighbourhood_cleansed_朝阳区 / Chaoyang	0.66%	1.48%	0.47%
neighbourhood_cleansed_大兴区 / Daxing	0.39%	0.58%	0.00%
neighbourhood_cleansed_东城区	0.66%	1.82%	0.32%
neighbourhood_cleansed_房山区	0.64%	0.47%	0.06%
neighbourhood_cleansed_丰台区 / Fengtai	0.32%	0.18%	0.06%
neighbourhood_cleansed_海淀区	0.33%	0.43%	0.01%
neighbourhood_cleansed_怀柔区 / Huairou	3.05%	2.63%	4.25%
neighbourhood_cleansed_门头沟区 / Mentougou	0.25%	0.46%	0.00%
neighbourhood_cleansed_密云县 / Miyun	1.18%	0.58%	0.59%
neighbourhood_cleansed_平谷区 / Pinggu	0.13%	0.53%	0.01%
neighbourhood_cleansed_石景山区	0.09%	0.07%	0.00%
neighbourhood_cleansed_顺义区 / Shunyi	0.61%	0.30%	0.03%
neighbourhood_cleansed_通州区 / Tongzhou	0.61%	0.39%	0.01%
neighbourhood_cleansed_西城区	0.36%	1.37%	0.06%
neighbourhood_cleansed_延庆县 / Yanqing	2.67%	0.69%	0.23%
总计	**100.00%**	**100.00%**	**100.00%**

6) 梯度提升树 XGBoost

表 36 梯度提升树 XGBoost 重要性汇总

Variables	聚类 0	聚类 1	聚类 2
host_time	0.76%	1.13%	0.29%
num_of_host_verifications	8.80%	1.45%	0.30%
num_of_amenities	1.54%	0.74%	0.20%
latitude	1.93%	1.77%	0.69%
longitude	1.23%	0.53%	0.21%
host_response_rate	1.80%	2.71%	1.41%
host_acceptance_rate	1.58%	20.77%	0.07%
host_listings_count	0.47%	0.69%	0.18%
accommodates	28.37%	4.36%	1.43%
bathrooms	5.64%	1.41%	1.18%
bedrooms	7.60%	1.64%	1.06%
beds	4.93%	0.87%	0.60%
minimum_nights	5.73%	0.23%	36.23%
maximum_nights	1.32%	0.54%	8.53%
availability_30	0.33%	0.32%	0.16%
availability_60	0.00%	0.15%	0.19%
availability_90	0.00%	0.13%	0.05%
availability_365	0.40%	0.55%	0.20%
number_of_reviews	2.79%	0.92%	0.12%
number_of_reviews_ltm	0.58%	0.15%	0.14%
calculated_host_listings_count	2.42%	0.99%	0.16%
calculated_host_listings_count_entire_homes	0.99%	0.47%	0.09%
calculated_host_listings_count_private_rooms	1.51%	1.09%	0.14%
calculated_host_listings_count_shared_rooms	0.00%	4.32%	0.01%
host_response_time_a few days or more	0.00%	0.00%	0.00%
host_response_time_within a day	0.00%	0.52%	0.02%

Variables	聚类 0	聚类 1	聚类 2
host_response_time_within a few hours	0.00%	0.21%	0.57%
host_response_time_within an hour	1.41%	0.37%	0.04%
host_is_superhost_t	0.00%	0.83%	0.11%
host_is_superhost_f	0.00%	0.00%	0.01%
host_has_profile_pic_t	0.00%	0.00%	0.00%
host_has_profile_pic_f	0.00%	0.00%	0.00%
host_identity_verified_t	0.00%	0.00%	0.31%
host_identity_verified_f	0.00%	0.00%	0.00%
property_type_Barn	0.00%	0.00%	0.00%
property_type_Camper/RV	0.00%	0.00%	0.00%
property_type_Campsite	0.00%	0.00%	0.00%
property_type_Casa particular	0.00%	0.00%	0.00%
property_type_Castle	0.00%	0.00%	0.00%
property_type_Cave	0.00%	0.00%	0.00%
property_type_Dome house	0.00%	0.00%	0.00%
property_type_Earth house	0.00%	0.00%	0.02%
property_type_Entire apartment	0.00%	0.08%	0.07%
property_type_Entire bed and breakfast	0.00%	0.00%	0.00%
property_type_Entire bungalow	0.00%	0.00%	0.50%
property_type_Entire cabin	0.00%	0.00%	0.00%
property_type_Entire chalet	0.00%	0.00%	0.00%
property_type_Entire condominium	0.00%	0.00%	0.01%
property_type_Entire cottage	0.56%	0.20%	0.03%
property_type_Entire guest suite	0.00%	0.00%	0.02%
property_type_Entire guesthouse	0.00%	0.00%	0.07%
property_type_Entire home/apt	0.00%	0.00%	0.00%
property_type_Entire house	0.00%	0.00%	0.18%

Variables	聚类 0	聚类 1	聚类 2
property_type_Entire loft	0.63%	0.00%	0.04%
property_type_Entire place	0.00%	0.00%	0.00%
property_type_Entire resort	0.00%	0.00%	0.00%
property_type_Entire serviced apartment	0.00%	0.00%	0.02%
property_type_Entire townhouse	0.00%	0.00%	0.07%
property_type_Entire villa	3.24%	0.18%	0.18%
property_type_Farm stay	1.14%	0.00%	23.33%
property_type_Houseboat	0.00%	0.00%	0.00%
property_type_Hut	0.00%	0.00%	0.00%
property_type_Igloo	0.00%	0.00%	0.00%
property_type_Kezhan	0.00%	0.00%	0.02%
property_type_Minsu	0.00%	0.00%	0.00%
property_type_Pension	0.00%	0.00%	0.00%
property_type_Private room	0.00%	0.07%	0.00%
property_type_Private room in apartment	0.00%	0.15%	0.00%
property_type_Private room in barn	0.00%	0.00%	0.00%
property_type_Private room in bed and breakfast	0.00%	0.08%	0.00%
property_type_Private room in bungalow	0.00%	1.12%	0.00%
property_type_Private room in cabin	0.00%	0.04%	0.00%
property_type_Private room in camper/rv	0.00%	0.00%	0.00%
property_type_Private room in campsite	0.00%	0.00%	0.00%
property_type_Private room in casa particular	0.00%	0.00%	0.00%
property_type_Private room in castle	0.00%	0.00%	0.00%
property_type_Private room in cave	0.00%	0.00%	0.00%
property_type_Private room in chalet	0.00%	0.00%	0.00%
property_type_Private room in condominium	0.00%	0.63%	0.00%
property_type_Private room in cottage	0.83%	0.41%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Private room in dome house	0.00%	0.00%	0.00%
property_type_Private room in earth house	0.00%	0.12%	0.00%
property_type_Private room in farm stay	0.00%	5.38%	0.00%
property_type_Private room in guest suite	0.00%	0.05%	0.00%
property_type_Private room in guesthouse	0.00%	0.98%	0.00%
property_type_Private room in hostel	0.00%	0.52%	0.00%
property_type_Private room in house	0.00%	0.20%	0.00%
property_type_Private room in hut	0.00%	0.00%	0.00%
property_type_Private room in kezhan	0.00%	2.00%	0.00%
property_type_Private room in loft	0.00%	3.73%	0.00%
property_type_Private room in minsu	0.00%	0.00%	0.00%
property_type_Private room in nature lodge	0.00%	0.81%	0.00%
property_type_Private room in resort	0.00%	0.94%	0.00%
property_type_Private room in ryokan	0.00%	0.00%	0.00%
property_type_Private room in serviced apartment	0.00%	0.83%	0.00%
property_type_Private room in tent	0.00%	0.00%	0.00%
property_type_Private room in tiny house	0.00%	0.05%	0.00%
property_type_Private room in townhouse	0.00%	0.20%	0.00%
property_type_Private room in treehouse	0.00%	0.00%	0.00%
property_type_Private room in villa	0.00%	0.58%	0.00%
property_type_Room in aparthotel	0.00%	0.40%	0.01%
property_type_Room in boutique hotel	0.00%	4.07%	0.00%
property_type_Room in heritage hotel	0.00%	0.00%	0.00%
property_type_Room in hotel	0.00%	0.44%	0.00%
property_type_Shared room	0.00%	0.00%	0.00%
property_type_Shared room in apartment	0.00%	0.53%	0.00%

Variables	聚类 0	聚类 1	聚类 2
property_type_Shared room in bed and breakfast	0.00%	0.00%	0.00%
property_type_Shared room in boutique hotel	0.00%	0.71%	0.00%
property_type_Shared room in bungalow	0.00%	0.00%	0.00%
property_type_Shared room in condominium	0.00%	0.00%	0.00%
property_type_Shared room in cottage	0.00%	0.00%	0.00%
property_type_Shared room in earth house	0.00%	0.00%	0.00%
property_type_Shared room in farm stay	0.00%	0.00%	0.00%
property_type_Shared room in guest suite	0.00%	0.00%	0.00%
property_type_Shared room in guesthouse	0.00%	0.00%	0.00%
property_type_Shared room in hostel	0.00%	0.00%	0.00%
property_type_Shared room in house	0.00%	0.48%	0.00%
property_type_Shared room in hut	0.00%	0.00%	0.00%
property_type_Shared room in kezhan	0.00%	0.00%	0.00%
property_type_Shared room in loft	0.00%	0.00%	0.00%
property_type_Shared room in nature lodge	0.00%	0.00%	0.00%
property_type_Shared room in serviced apartment	0.00%	0.09%	0.00%
property_type_Shared room in tent	0.00%	0.00%	0.00%
property_type_Shared room in tiny house	0.00%	0.00%	0.00%
property_type_Shared room in townhouse	0.00%	0.00%	0.00%
property_type_Shared room in villa	0.00%	0.00%	0.00%
property_type_Tent	0.00%	0.00%	0.00%
property_type_Tiny house	0.00%	0.00%	0.01%
property_type_Treehouse	0.00%	0.00%	0.00%
property_type_Yurt	0.00%	0.00%	0.00%
room_type_Entire home/apt	0.00%	2.46%	0.01%
room_type_Private room	0.00%	0.14%	0.00%
room_type_Shared room	0.00%	0.09%	0.00%

Variables	聚类 0	聚类 1	聚类 2
instant_bookable_t	0.00%	0.23%	2.48%
instant_bookable_f	0.00%	0.00%	0.06%
neighbourhood_cleansed_昌平区	0.00%	0.13%	0.04%
neighbourhood_cleansed_朝阳区 / Chaoyang	0.00%	3.52%	0.78%
neighbourhood_cleansed_大兴区 / Daxing	0.00%	1.48%	0.00%
neighbourhood_cleansed_东城区	2.53%	1.60%	0.24%
neighbourhood_cleansed_房山区	0.00%	0.29%	0.01%
neighbourhood_cleansed_丰台区 / Fengtai	0.00%	0.00%	0.01%
neighbourhood_cleansed_海淀区	0.00%	0.07%	0.03%
neighbourhood_cleansed_怀柔区 / Huairou	0.97%	0.97%	16.54%
neighbourhood_cleansed_门头沟区 / Mentougou	0.00%	0.22%	0.01%
neighbourhood_cleansed_密云县 / Miyun	3.80%	0.32%	0.03%
neighbourhood_cleansed_平谷区 / Pinggu	0.00%	0.81%	0.02%
neighbourhood_cleansed_石景山区	0.00%	0.00%	0.00%
neighbourhood_cleansed_顺义区 / Shunyi	2.32%	0.62%	0.02%
neighbourhood_cleansed_通州区 / Tongzhou	0.00%	0.24%	0.01%
neighbourhood_cleansed_西城区	0.00%	9.46%	0.08%
neighbourhood_cleansed_延庆县 / Yanqing	1.84%	1.39%	0.30%
总计	**100.00%**	**100.00%**	**100.00%**