爱彼迎短租房房价影响因素分析

——以北京地区为例

美国商业分析大赛队 – 梁之扬 陈乐偲 毛炫林

爱彼迎短租房房价影响因素分析

--以北京地区为例

摘要：本案例以爱彼迎平台在北京地区的短租房房源为研究对象，通过统计机器学习的方法，分析探究行政区划，地理位置，房源类型，便利设施等相关因素对短租房放假的影响做用，建立了对数线性模型来描述各因素与每晚短租房房价之间的联系，并使用决策树和随机森林模型尝试对每晚房价进行预测。结论表明，区位因素和短租房内部配置因素与短租房房价之间均有极强相关性，但并不是良好的决策变量。平台在实际决策房价时，还需考虑更多因素。

一、背景介绍与研究问题

2010 年，在线短租概念在中国兴起，中国第一家在线短租企业于 2011 年建立，随后更多中国企业加入在线租房行业。2012 年，在消费升级的大背景下，具有更好的居住体验和性价比的连锁品牌短租公寓一定程度上取代了传统的租房行业，整个行业陷入快速发展和品牌化时期。2015 年，美国企业爱彼迎入局中国短租行业及其市场份额的飞速发展吸引了众多企业加入在线短租市场。随着置业成本的增高和政府的支持政策，在线租房行业进入了另一个高速发展的新时期，开始认真起来经营中国市场的爱彼迎和美团榛果民宿等平台的入局推动行业发展走向新阶段。

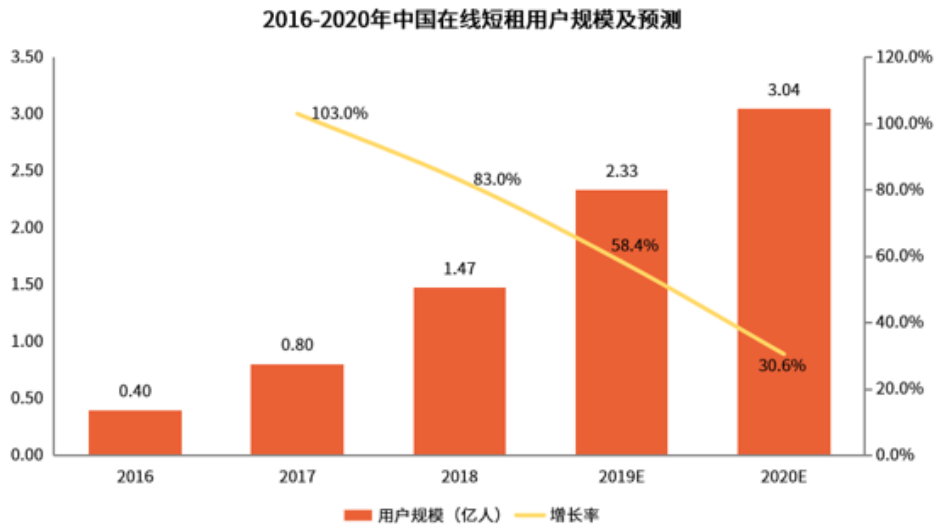

图 1-1 2016 年——2020 年中国在线短租用户规模及预测

在短租房平台的火爆和快速发展的同时，其背后存在的问题也在逐渐显露，如短租房带来了较为快速的人员流动、噪音、卫生健康、安全隐患、监管问题也随之而来。2020 年北京

市发布了《关于规范管理短租住房的通知（征求意见稿）》，首都功能核心区内禁止经营短租房，北京小区经营短租房或要同楼业主书面同意。同时，新冠疫情也对短租房行业造成冲击，据爱彼迎 2020 年 11 月 16 日发布的招股说明书显示，截至 2019 年 12 月 31 日和 2020 年 9 月 30 日，爱彼迎累计赤字分别为 14 亿美元和 21 亿美元，2020 年前三季度累计亏损 7 亿美元，前三季度营收 25.19 亿美元，相比 2019 年同期营收（36.98 亿）下滑 31.9%。在同质化竞争加剧、政策不断调整、监管逐步严格、疫情冲击仍未消退的情况下，爱彼迎需要及时做出调整来巩固业务、生存下去。

我们在对比爱彼迎不同店家房源中认为，商家常常会因为执着于某个或某些因素对价格的影响而错误定位自己的产品的价格区间，导致最终定价与实际可能市场均衡相去甚远。如何合理定价需要分析二手房的价格的影响因素。

本案例中，我们收集了爱彼迎平台于北京市内 24973 套短租房的相关数据，对其价格相关影响因素展开研究，报告如下。

二、数据来源与说明

本案例使用的是来自第三方网站 http://insideairbnb.com/的数据，据该网站称其数据均是来源于爱彼迎网站上所能公开爬取到的数据，不包含用户隐私信息。该数据集共 24973 条记录，数据采集时间为 2021 年 2 月 21 日。原始数据共包含 74 个变量，其中需要清洗的数据很多，例如 neighbourhood_group 列全为空，host_name，host_id 等为冗余信息，last_scraped 等列意义不明，license，review_detail 等列与我们需要研究的问题不符。

最终清洗过后，我们选取 10 个变量进行分析，其中连续型变量 1 个，离散型变量 9 个；因变量为短租房每晚房价，其他变量为自变量，又可以分为区位因素和内部配置因素，具体变量说明如表 2-1 所示。

表 2-1：数据变量说明表

变量类型	变量名	详细说明	取值范围	备注
因变量	每晚房价	单位：元/晚	36~70000	去除了异常值99999999和0
自变量	经度	单位：度	115.5~117.5	描述分析时和经度合并成了地理位置
	纬度	单位：度	39.5~41.0	描述分析时和纬度合并成了地理位置
区位因素	所属城区	定性变量：共16个水平	东城区，西城区，昌平区，大兴区，房山区 怀柔区，门头沟区，密云县，朝阳区，丰台区，海淀区 平谷区，延庆县，顺义区，通州区，石景山区	
配置因素	房源类型	定性变量：共3个水平	共享房间，私人房间，整套公寓	
	是否允许长租	定性变量：共2个水平	1代表允许长租，0代表不允许	96.6%允许长租
	是否有Wifi	定性变量：共2个水平	1代表有Wifi，0代表没有Wifi	96.0%有Wifi
	是否有空调	定性变量：共2个水平	1代表有空调，0代表没有空调	94.7%有空调
	是否有洗浴用品	定性变量：共2个水平	1代表有洗浴用品，0代表没有洗浴用品	90.2%有洗浴用品
	是否有暖气	定性变量：共2个水平	1代表有暖气，0代表没有暖气	90.2%有暖气

三、描述性分析

在对房价的影响因素进行模型探究之前，首先对各变量进行描述分析，初步判断房价的影响因素，为后续研究做铺垫。

（一）因变量：每晚房价

在本案例当中，每晚房价的最小值为 36 元，对应是朝阳区的女生合租四人间，所以对于个人的单价比较便宜；每晚房价最大值是 99999 元的四合院，临近天安门广场，这个可能是店家拿出来炫耀的，所以我们分析的时候还是把它去掉了。其余价格较高的数据都是比较高端的家庭及团队旅游住宿服务，这也符合我们的一般认知。

总体来说，短租房行业还是以亲民的价格为主，但也不乏定价极高的高端产品。

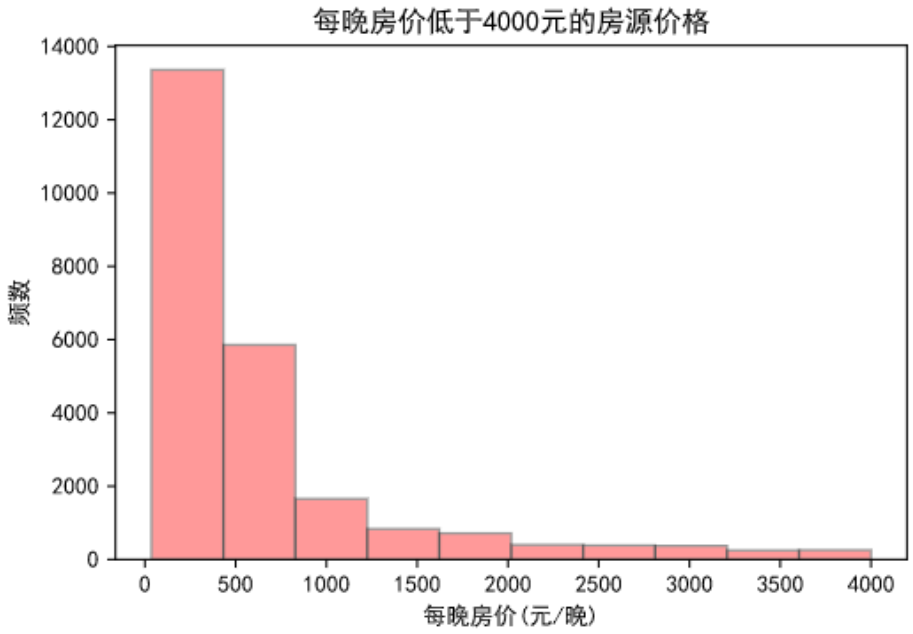

图 3-1 低于 4000 元的短租房房源价格分布

（二）自变量：区位因素

按照短租房每晚的平均价格，我们大致可以把价格分为三个梯队：怀柔区，平谷区，延庆县在 1600 元以上；门头沟区，昌平区，密云县在 1200 元以上；其余区在 500-900 元之间。以下从各梯队价格差异的示意图。有意思的是，第三梯队虽然平均数低，但是离群点所代表的价格却非常高，这说明不论在哪里高端短租房都依然有市场。

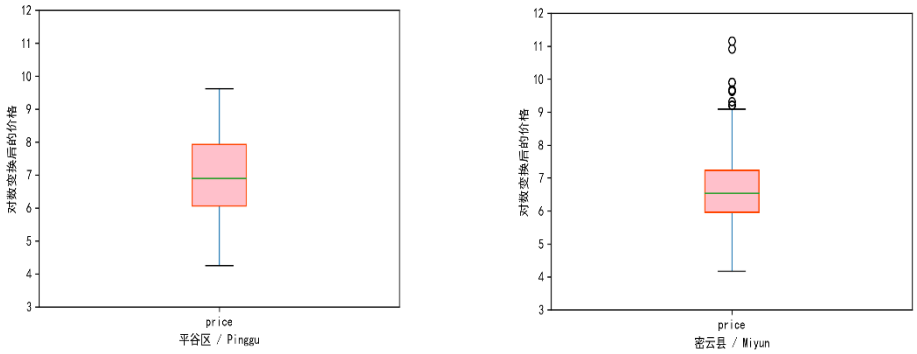

图 3-2 第一梯队和第二梯队的价格分布情况

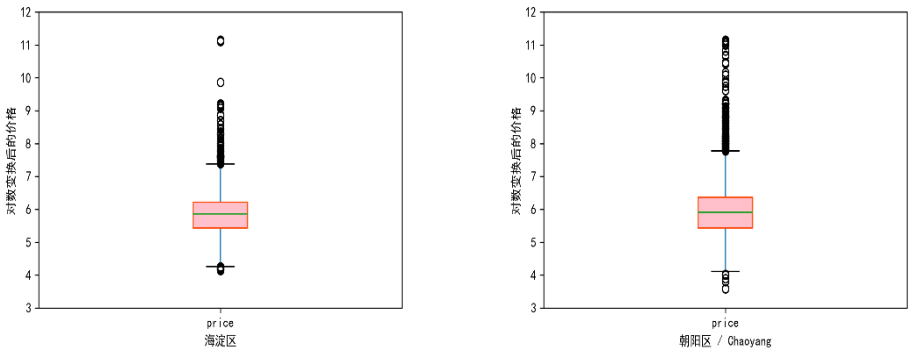

图 3-3 第三梯队的价格分布情况

我们一开始推测价格或许和房源数有关，密度高了，竞争激烈，商家想吸引客户的一个比较通常也比较简单的办法是，降低价格。

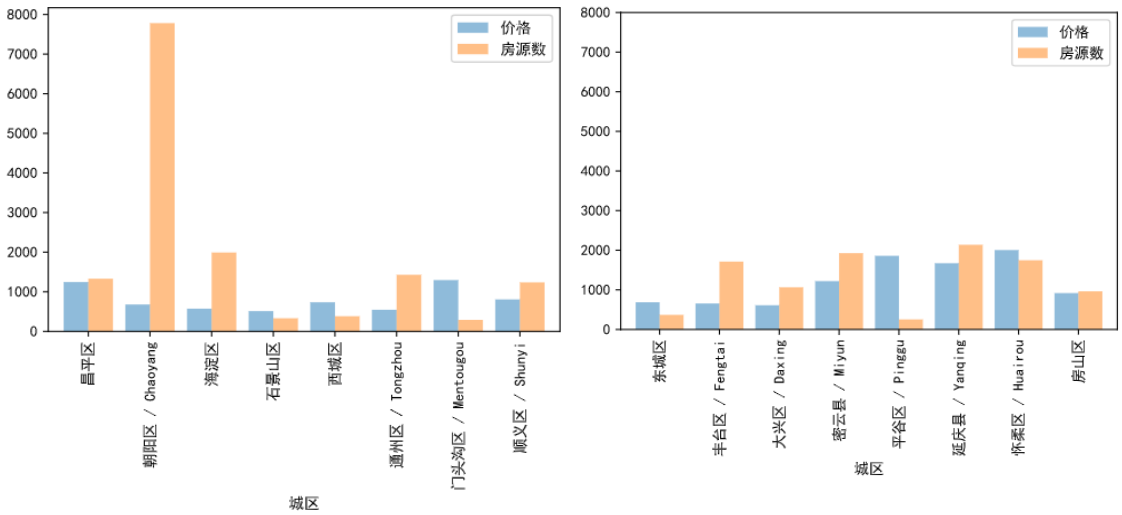

图 3-4 价格与房源数的关系

从图 3-4 可以看出，价格与房源数之间并没有很强的相关性，于是我们建模分析时，

应该考虑各行政区划中其他的因素，例如旅游景点，消费者人群分布等等因素，而不将房源数作为重要因素。

（三）自变量：内部配置因素

从图 3-5 可以看出，房源类型私人房间的短租房均价远高于共享房间，整套公寓的价格与私人房间相近，但离群点较私人房间更多。这一方面说明客户对舒适度和私密性要求高，且愿意为之付出更高价格；另一方面，也说明存在整套公寓利用低价吸引客户以及某些共享房间也提供高端产品，市场上短租房种类非常多，竞争十分激烈。

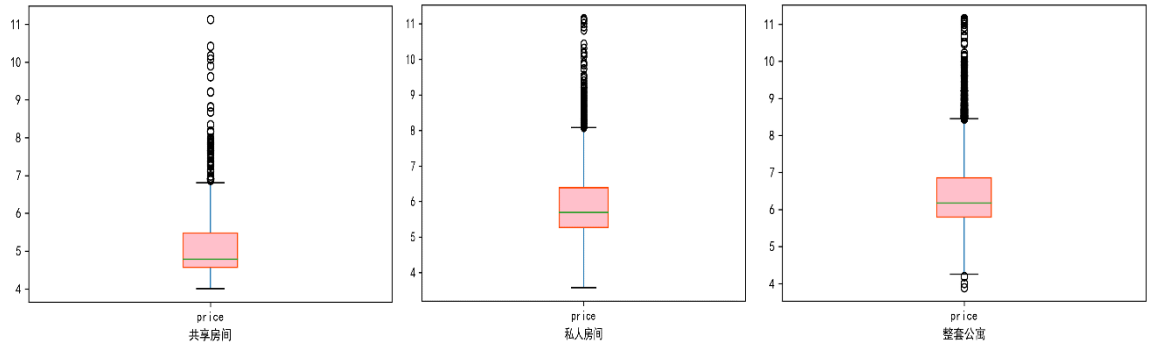

图 3-5 短租房价格和房源类型的关系

图 3-6 以有无暖气为例，描述了便利设施的有无与短租房房价之间的关系。虽然有暖气和无暖气在平均价格上并没有太大的区别，但是很明显有暖气的这一组离群点更多更密集，如果短租房想要定一个很高的价格，很有可能需要有充足完备的便利设施与之配套。值得一提的是，这些内部配置很多都是现代人几乎离不开的因素，例如 Wifi，但是在近两万五千房源当中，居然有百分之五没有 Wifi，也就是说近一千套房源没有配备，这或许就是他们不敢定高价的原因之一，进而导致其所对应的价格分布离群点少。

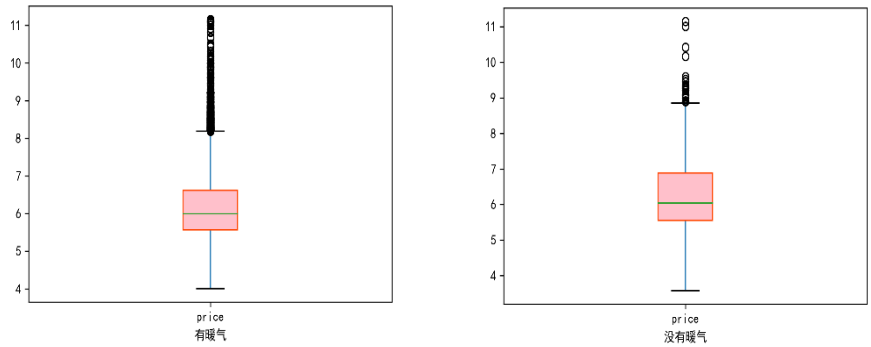

图 3-6 短租房价格和便利设施的关系（以有无暖气为例）

综上，通过对本案例数据的描述性分析，我们可以初步推断：对短租房房价可能会产生

影响的因素有：区位因素(行政区划、地理位置)和内部配置因素（房间类型、是否有 Wifi、暖气、洗浴用品等便利设施）；从影响作用来看，区位因素影响更为明显一些。

四、模型建立

为了更深入地分析各因素对短租房房价的影响，本案例将建立多个模型，尝试构建短租房房价和区位因素以及内部配置因素之间的联系，从而更为精细化地刻画这两方面因素的影响作用，并结合模型拟合和预测的结果，对模型进行说明与评价，以指导现实生活。

（一）线性回归与对数线性回归模型

利用卡方检验的手段，建立列联表，分别分析区位因素与内部配置因素与价格的相关性。以对区位因素与短租房价格的相关性分析为例，构建零假设为区位因素与价格无关，备选假设为区位因素与价格有关，计算得到卡方检验的 p 值远小于 0.001，因此在显著性水平为 0.001 的条件下，没有理由认为短租房所处的地理区位与价格无关。通过类似的方法可以分析内部配置与短租房价格的相关性，同样 p 值远小于 0.001，同样我们没有理由认为短租房的配置与价格无关。

进一步，我们尝试用回归分析的手段，分析短租房房价与自变量之间的关系。采用最简单的线性模型，采用线性回归的手段，得到的拟合优度小于 0.01，证明自变量与因变量之间绝非简单的线性关系。从现实的角度出发，作为连续性变量的住宿条件，住宿条件的好坏带给消费者的效用存在边际效用递减的规律，因此必然不是简单的线性关系。作为离散型变量的住宿条件，更不可能存在简单的线性关系。故我们最终舍弃这一模型，之后采用对数线性回归模型，拟合优度达到了 0.58，此次具有一定的参考价值。

（二）采用决策树和随机森林进行价格预测

尝试从区位因素和内部配置因素出发，预测短租房的价格，为商家提供现实参考。我们如表 4-1 所示，建立价格等级。先采用决策树模型，设置最大决策深度为 10，模型的拟合优度仅有 0.3， 说明考虑的决策变量不够全面，对于进行精准的决策来说远远不够。之后我们尝试使用随机森林的的手段进行预测，设置最大决策深度为 10，决策树数目为 10，最小分裂样本数为 2。拟合优度仅有 0.4，虽然比单纯的决策树提高了 0.1 的拟合优度，但该提升为集成学习带来的，且效果仍不佳，说明预测的瓶颈在于决策变量的选取。

表 4-1 价格等级

价格/（元/晚）	0-100	100-500	500-1000	1000-3000	3000-5000	5000-10000	>10000
价格等级	1	2	3	4	5	6	7

（三）模型结果说明与评价

所用模型以及模型的结果如表 4-2 所示。我们可以得出以下结论：

(1) 区位因素和内部配置因素都是短租房价格等级的相关性极强。

(2) 区位因素和内部配置因素都不是短租房价格的线性相关因素，决策时要综合更多信息。

(3) 区位因素和内部配置因素都是短租房价格的特征因素，且区位因素特征更为显著。

表 4-2 模型及结果

模型	任务	自变量	因变量	最高拟合优度	备注
线性回归	回归	内部配置	价格	<0.1	
对数线性回归	回归	内部配置	价格	0.58	
决策树	分类	所有因素	价格等级	0.3	
随机森林	分类	所有因素	价格等级	0.4	

五、商业应用与总结

本案例对爱彼迎平台，截至 2021 年 2 月 21 日的北京市内 24973 套短租房数据进行统计机器学习分析，认为影响短租房房价的主要因素有二：

(1)区位因素：行政区划，地理位置

(2)内部配置：房源类型，是否有配套的便利设施

但同时我们也注意到，对价格进行决策时，这两个因素均表现平平，说明背后存在其他原因有待研究确定。如作为线上短租平台，在相对公开可对比的市场上，消费者对于房东评价、房东性别、评论数与评论情况、房东营销手段等方式会同样较为敏感。这也就说明在定价问题上，房东不应只依据基于区位、硬件设施提供的个人偏好预期盲目定价，在互联网平台上也要着重注意客户关系与信任的建立、营销宣传手段的提高与市场同质化产品的对比综合定价。

另外，面对目前监管形式的不断增强、疫情期间消费者对安全状况要求标准的提高，如果想要摆脱同质化产品的困境，爱彼迎作为平台方需要考虑如何发挥平台监管、帮扶优势，一方面作为平台加强对房源的审核以最大可能控制风险、维护品牌形象；另一方面，在营销、定价、内部设施布置等重要且可调整缓环节对商家进行培训，提高品控。