[文章编号] 1003－4684(2021)02-0104-06

改进的 XGBoost 模型在短租房价格预测中的应用

郑 列，穆新宇

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 对短租房价格原始数据集进行缺失值和异常值处理，针对短租房价格的影响因素构建包括 23 个特征的特征体系，使用 OLS 回归和分位数回归对这些因素的影响程度和影响方向进行分析，最后挑选具有较强显著性的 18 个特征构建 XGBoost 模型，用于预测房源价格。建模过程中采用网格搜索法调参。拟合优度这一指标在使用 XGBoost 模型进行价格预测时可以达到 0.60，而线性回归模型仅为 0.38。因此，使用 XGBoost 模型对短租房价格进行预测较优，将其与 OLS 回归和分位数回归相结合，既保留了传统统计模型的解释性，又提升了预测的精确度。

[关键词] 在线短租房；OLS 回归；分位数回归；XGBoost；网格搜索法

[中图分类号] F299.23 [文献标识码] A

在线短租房得到了空前的发展，并催生了途家、小猪、榛果民宿等知名短租平台[1]。从公开信息出发，建立在线短租房的价格预测模型，有助于在线短租产业的发展。XGBoost 模型自提出以来备受关注，不仅众多学者对其展开深入研究与改进[2-3]，而且在工业界取得了不错的成果[4-5]。相较于传统统计模型，XGBoost 无论在分类还是回归问题中均能取得较好效果，但是其解释性相对较差，不利于实际问题的具体分析。为解决这一弊端，本文在使用 XGBoost 预测短租房价格时，考虑先使用传统统计模型对问题做出较好的解释，再挑选表现较佳的变量构建预测模型。该模型能够为新加入短租平台的房源提供具有参考价值的定价范围，也能够帮助短租平台监管房源价格的异常情况，并及时做出调整，营造出公平合理的短租商业氛围。

1 数据来源及预处理

1.1 数据来源

本文的研究数据来源于阿里云天池大数据竞赛[6]，该数据是 Airbnb 公司于 2019 年 4 月 17 日公开的北京地区房源数据集。Airbnb 是全球知名的民宿短租平台，其房源范围覆盖 191 个国家和地区，以 Airbnb 平台的数据研究相关问题具有重要的参考价值。目前，在线短租业务主要分布在一线和省会城市，北京作为首都具有代表性，以其房源信息作为研究对象相对合理。

原始数据集包含 28452 个样本和 106 个属性，每一个样本对应一个房源，而每一个属性代表房源的一个特征，不过该数据集并未指定哪一个属性作为研究目标，因此为相关问题的研究提供了更多的可能性。本文拟研究在线短租房的价格影响因素及其预测模型，故将 price 属性作为目标变量，其原始值为数值型，表示房源的价格，而将其它属性作为房源的固有特征，其中包括数值型、分类型和文本型，主要涉及房源基本情况、房主情况和房客评价等方面。

1.2 数据预处理

在原始数据集中，存在缺失值和异常值，需要进行适当的数据清洗，对于缺失值一般采用样本填充法或属性删除法，如果某一属性的缺失值比例不大，那么会选择对有缺失值的样本进行填充，数值型数据采用均值填充，分类型数据采用众数填充，文本型数据暂不填充，而如果某一属性的缺失值比例较大，那么会选择删除该属性对应的所有数值。对于异常值的处理一般采用样本删除法，异常值会对后续分析产生很大干扰，可以根据 3σ 原则进行识别，进而删除存在异常值的样本。本文数据预处理的最终样本量为 23364，保留属性有 82 个。

2 特征体系构建

在线短租房是一种新兴产业，其发展模式介于传统酒店与传统租房之间，所以在研究短租房价格

[收稿日期] 2020－09－10

[基金项目] 教育部人文社会科学研究规划基金项目(17YJA790098)

[第一作者] 郑 列(1963－)，男，湖北英山人，湖北工业大学教授，研究方向为应用数学

[通信作者] 穆新宇(1996－)，女，江苏丰县人，湖北工业大学硕士研究生，研究方向为数据挖掘

的影响因素时,既要借鉴对传统酒店价格的研究,又要考虑短租房自身的特点。

本文查阅多篇有关短租房价格的文献[7-9],综合多个方面对短租房价格的影响因素构建了合理的特征体系。特征体系包括 5 个类别,分别为房源的基础设施、房源的基本属性、房主的基本情况、在线预定规则和房客的评论情况,共计 23 个变量,其详细的名称和含义见表 1。

表 1　短租房价格及其影响因素的描述

变量类别	变量名称	变量含义
目标变量	price	价格
房源的基础设施	bathrooms	浴室数
	bedrooms	卧室数
	beds	床数
	accommodates	可容纳人数
房源的基本属性	latitude	纬度
	longitude	经度
	is_location_exact	是否精确定位
	operation_days	经营天数
	property_type_is_apart	房产是否为公寓
	room_type_is_Entire	是否为整租
	neighbourhood_is_center	是否在市区
房主的基本情况	host_listings_count	拥有的房源总数
	host_has_profile_pic	是否提供肖像照片
	host_identity_verified	身份是否通过核实
	host_is_superhost	是否为超赞房主
在线预定规则	minimum_nights	最少入住天数
	instant_bookable	是否随时预定
	has_house_rules	是否有入住规定
	need_security_deposit	是否需要押金
	need_cleaning_fee	是否需要清洁费
	need_extra_people_fee	额外加人是否加费用
房客的评论情况	number_of_reviews_ltm	最近一年评论数
	review_days	评论天数

从表 1 可以看出,5 类特征涵盖了房源、房主和房客三个方面的信息,考虑的影响因素比较全面,而且 23 个变量中有 11 个数值型变量,12 个二分类型变量。二分类型变量主要通过短租平台的在线信息获取,体现了短租产业以互联网为重要媒介的特点。

3　影响因素分析

在线短租房的价格会受很多因素的影响,为了清楚地了解各影响因素对价格的影响程度和方向,需要建立合适的模型进行分析,本文主要采用传统统计模型中的 OLS 回归与分位数回归,其中 OLS 回归可以分析各因素对房源价格的综合影响情况,分位数回归可以分析各因素对不同价位房源的影响情况。

3.1　OLS 回归

OLS 回归,即最小二乘回归,它会将误差的平方和最小化,以此确定目标变量与影响因素之间的最佳线性关系,是各个学科研究中普遍使用的标准统计模型,其模型表达式为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

其中:y_i 被称作因变量;x_i 被称作自变量;β_0,β_1 是需要用最小二乘法确定的参数,也被称作回归系数;ε_i 被称作随机误差项。

使用 OLS 回归要求数据必须满足以下统计假设:①正态性,即对于固定的自变量值,因变量值成正态分布;②独立性,即个体之间相互独立;③线性相关,即因变量和自变量之间是线性相关的;④同方差性,即因变量的方差不随自变量的水平不同而变化,也就是说因变量的方差是恒定的。

OLS 回归因其思路简单、方便实现等特点,在各个学科广泛应用,不过它主要关注各影响因素与目标变量的条件均值之间的关系,没有充分考虑目标变量条件分布的整体性。

3.2　分位数回归

为了弥补 OLS 回归的局限性,18 世纪中期 Boscovich 首次提出中位数回归,在此基础上,Roger

Koenker 与 Gilbert Bassett 在 1978 年提出更具一般性的分位数回归，其模型表达式为

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)} \quad (2)$$

其中，$0 < p < 1$，表示数值小于第 p 分位数的比例。

y_i 在特定值 x_i 下的第 p 条件分位数为 $Q^{(p)}(y_i \mid x_i) = \beta_0^{(p)} + \beta_1^{(p)} x_i$，由此可知该模型要求误差项的第 p 分位数等于 0[10]。

分位数回归一般用来研究自变量与因变量的条件分位数之间的关系，得到的模型可以用前者来估计后者。它不是仅分析因变量的条件期望，而是比较全面地解释因变量的条件分布。

和 OLS 回归相比，分位数回归的使用条件更加宽泛，所获得的信息量更多，能够捕捉到条件分布形状对因变量的影响，可以全面地表现分布的情况，而且回归系数的估计更加稳健。

3.3 回归结果与分析

Python 是一门简洁易懂的编程语言，其中有专门用于统计分析的封装库，对于统计分析非常方便，通过输入相关数据可以对在线短租房的价格及其影响因素做 OLS 回归与分位数回归，其中所选分位数依次为 0.1、0.25、0.5、0.75 和 0.9，两种结果的对比情况如表 2 所示。

表 2 OLS 回归与分位数回归结果对比

变量	OLS 回归	分位数回归				
		$q=0.1$	$q=0.25$	$q=0.5$	$q=0.75$	$q=0.9$
constant	2087.68	282.32	−897.79	−2465.39	−4795.14	−10320.0
bathrooms	282.80**	2.36	14.47**	83.58**	276.41**	502.14**
bedrooms	84.84**	29.30**	33.07**	40.05**	45.23**	95.37**
beds	−24.24**	−12.08**	−15.90**	−18.79**	−18.87**	−27.71**
accommodates	56.33**	22.03**	40.15**	55.92**	67.06**	73.72**
latitude	509.30**	96.30**	186.69**	356.76**	574.17**	892.93**
longitude	−193.03**	−34.06**	−54.87**	−100.07**	−154.91**	−217.34**
is_location_exact	−38.18**	−2.20	−3.26	−5.31*	−24.33**	−53.37**
operation_days	0.05**	0.01**	0.01**	0.02**	0.04**	0.06**
property_type_is_apart	−15.68**	−0.11	−3.89	−5.06*	−18.15**	−40.18**
room_type_is_Entire	124.79**	133.24**	137.18**	137.35**	129.77**	70.48**
neighbourhood_is_center	164.45**	39.61**	52.15**	85.72**	141.23**	249.88**
host_listings_count	1.48**	1.37**	1.53**	1.35**	1.81**	5.04**
host_has_profile_pic	−134.88*	−130.65**	−133.05**	−135.28**	−217.88**	−186.31
host_identity_verified	−28.11**	−5.32	−2.16	−12.90**	−21.31**	−32.67*
host_is_superhost	15.38*	13.81**	13.39**	9.16**	16.36**	1.08
minimum_nights	−3.55**	−2.01**	−2.65**	−2.64**	−2.56**	−4.71**
instant_bookable	16.26**	5.00*	4.27	7.97**	16.67**	−4.44
has_house_rules	−13.96**	4.76*	−1.05	−8.93**	−21.10**	−55.91**
need_security_deposit	45.57**	20.09**	24.00**	29.70**	31.71**	82.54**
need_cleaning_fee	−20.34**	19.92**	12.05**	0.99**	−9.35**	−35.95**
need_extra_people_fee	−48.29**	−22.53**	−27.63**	−26.24	−31.87	−36.08**
number_of_reviews_ltm	−3.59**	−0.45*	−0.51*	−1.18**	−2.18**	−3.35**
review_days	−0.10**	0.01	−0.01	−0.02**	−0.06**	−0.18**

* 表示在 0.05 水平下显著，** 表示在 0.01 水平下显著

由表 2 可见，所有因素在 OLS 回归中均显著，但在分位数回归中有个别因素不显著。通过这些结果不仅可以分析每个因素对于房源价格的影响情况，还可以针对不同价位的房源给出不同的解释。

从房源的基础设施来看，浴室数、卧室数、床数和可容纳人数无论在 OLS 回归中还是分位数回归中均较为显著，并且在分位数回归中，分位数越大各因素对价格的影响程度越大。浴室数、卧室数和可容纳人数对价格的回归系数是正值，所以这些因素的值越大，房源的价格越高，而床数对价格的回归系数是负值，说明房源的床数越多价格反而越低，这很可能是房主为吸引对价格敏感的房客而采用的营销策略，试图通过提供更多的入住机会来降低价格。

从房源的基本属性来看，是否精确定位和房产是否为公寓在低分位数回归时没有通过显著性检验，说明这两个特征对低价房源的价格并无显著影响，不过在高分位数回归中随着分位数的增大对价格的影响程度递增，且均为负向影响。其它因素在

OLS 回归与分位数回归中均较为显著，且为正向影响，其中经营时间越长价格会越高，说明房主在经营经验的基础上可以打造出更有特色、更可靠的房源，另外，房源是否为整租对中等价位房源的价格影响程度最大，而是否在市区对高价房源的影响程度最大。

从房主的基本情况来看，房主的身份是否通过验证在低分位数回归时没有通过显著性检验，说明它对低价房源的价格并无显著影响，因为低价房源的安全性要稍微低一些，房客不会过于关注房主的身份情况。房主拥有的房源数越多价格越高，说明此类房源的房主可能是从事短租产业的专业房主，可以给房客提供更好的服务。另外，超赞房主拥有和专业房主同样的经营优势，房源价格自然会较高。

从在线预定规则来看，是否有入住规则和是否可以随时预定对中等价格房源的影响较为明显，而是否需要清洁费则对中等价位房源的价格无明显影响。另外，需要押金的房源价格更高，说明此类房源的设施和服务应该较好，价格自然攀升，而额外加人需要另收费的房源价格较低，说明此类房源更倾向于按人数收费。房源的最少入住天数越多价格越低，说明房主倾向于将房源租给长期房客，可以减少服务和沟通成本。

从房客的评论情况来看，评论天数对于低价房源的价格影响不显著，最近一年的评论数对于房源价格的整体影响均比较显著。不过两者对于房源价格的影响都是负向的，当评论数和评论天数增加时房源价格会下降，评论数在一定程度上能够反映房源的预定量，说明房主倾向于采用薄利多销的经营方式。

4　价格预测模型

房屋的价格预测有很多经典的预测模型，但是在线短租房与传统房屋在价格预测方面存在诸多不同，其中最主要的不同在于影响因素，模型的选择也会产生差异。随着机器学习技术的不断发展，目前倾向于选择新颖的模型来解决问题。

本文挑选出 OLS 回归和分位数回归中均有较强显著性的因素来构建价格预测模型，最终选取的是除 is_location_exact、property_type_is_apart、host_identity_verified、instant_bookable 和 review_days5 个之外的 18 个因素，主要采用 XGBoost 模型预测房源价格，并与线性回归模型的效果作比较，可以突出 XGBoost 在未调参与调参后的预测精度，最后通过 XGBoost 算法给出所有特征的重要性排序。

4.1　XGBoost 模型

XGBoost 算法是对 GBDT 算法的改进。原始的 GBDT 只利用了一阶的导数信息，而 XGBoost 则是对损失函数进行二阶泰勒展开，并在损失函数之外加入了正则项，可以针对整体计算最优解，用来衡量损失函数的下降以及模型的复杂度，避免过拟合，提高了模型的求解效率。XGBoost 算法的基本原理如下。

假定 (x_i,y_i)，$i=1,2,\cdots,n$ 是建模样本，$\hat{y}_i^{(t)}$ 是第 t 轮迭代后模型的预测结果，$f_t(x_i)$ 是第 t 个决策树的预测结果，则

$$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{3}$$

由于在第 t 轮迭代时 $t-1$ 轮的预测结果固定，模型目标函数的设定仅需考虑预测函数 $f_t(x_i)$，求解模型参数时最小化为如下目标函数：

$$S^{(t)}(\beta)=L(\beta)+D(f_t)+C \tag{4}$$

其中，

$$L(\beta)=\sum_{i=1}^{n}l(y_i,y_i^{(t-1)}+f_t(x_i)) \tag{5}$$

$$D(f_t)=\gamma T+0.5\lambda\sum_{j=1}^{T}\omega_j^2 \tag{6}$$

式(4)—(6)中：$L(\beta)$ 是测度模型拟合程度的损失函数，$D(f_t)$ 是测度模型复杂程度的正则化项，C 是常数项；$l(\cdot)$ 是测度样本预测准确性的损失函数；T 是决策树叶子节点数，ω_j 是叶子节点对应的预测结果，γ 和 λ 是对应的调整系数。将损失函数泰勒展开至二次项，利用贪婪算法或其它算法可以求解模型的参数[11]。

XGBoost 能够获得青睐，取决于其优越性：1）支持二阶泰勒展开式，不仅能够增加精度，而且方便自定义损失函数；2）损失函数中添加正则项，能够控制模型的复杂度，防止发生过拟合现象，使训练出来的模型相对简洁；3）允许列抽样，既能够防止过拟合，又能够简化计算；4）支持并行计算，灵活性很强。

4.2　模型建立与调参

本文使用 Python 语言实现 XGBoost 算法。首先将预处理后的数据集 23364 个样本以 7∶3 的比例随机划分为训练集和验证集，接着构建 XGBoost 模型，然后使用网格搜索法对模型进行调参，最后根据评价指标选择较优的模型。

建模过程中的调参环节本质上是一个优化过程，可以使用随机搜索法、网格搜索法、遗传算法等。本文选择的网格搜索法需要给定参数的若干个值，然后将各参数的可能值进行排列组合，并将各组参数用来训练模型，同时采用交叉验证的方式评估各种组合的表现，选取效果最好的组合作为最优参数。

XGBoost 算法建模时共有三类参数：常规参

数、基础模型参数和学习任务参数[12]，本文研究过程中对常规参数全部选择默认。由于研究目标最终可以归结为一个回归问题，所以学习任务参数里的objective参数需要设置为“reg:linear”，其他选择默认即可，而基础模型参数是对模型效果影响较大的部分，也是调参的重点。本文根据各参数在模型中的重要性依次调节，结果见表3。

表3 XGBoost建模调参结果

参数	含义	最优值
learning_rate	学习率	0.1
n_estimators	回归器的数量	100
max_depth	树的最大深度	10
gamma	叶节点进一步划分所需的最小损失减少量	0.5
min_child_weight	生成一个子节点所需要的最少样本权重和	1
subsample	子样本占训练样本集的比例	0.7
colsample_bytree	建立树时对特征随机采样的比例	0.8

4.3 模型评价

模型的好坏需要根据评价指标来评判，对于任何问题而言都没有最优的模型，但是可以在已有的模型中选择较优的那一个。在回归预测问题中有一些常用的评价指标，比如：平均绝对误差(MAE)、均方误差(MSE)和拟合优度(R^2)，它们的计算公式分别如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|f_i-y_i| \tag{7}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(f_i-y_i)^2 \tag{8}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(f_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2} \tag{9}$$

其中，f_i 表示预测值，y_i 表示真实值，$\bar{y}$ 表示真实值的期望[13]。

本文为了清楚地表现XGBoost模型的预测精度，与线性回归模型进行了对比，验证集上各指标对比结果(表4)。

表4 各回归模型的预测精度对比

模型	线性回归	XGBoost回归	
		未调参	调参后
MAE 值	198.98	166.72	137.98
MSE 值	118801.75	95492.29	77343.10
R^2 值	0.38	0.50	0.60

由表4可知，利用线性回归模型对在线短租房的价格进行预测时精度较低，其 R^2 值仅为0.38，而XGBoost模型在未调参时 R^2 值可以达到0.50，通过网格搜索法调参可以达到0.60，而且 *MAE* 和 *MSE* 的值在XGBoost模型中也明显下降，可见XGBoost模型相较于线性回归模型来说，拟合效果得到了较大提升。图1展示了XGBoost模型中各特征的重要性排序。

通过图1可以看出，在众多影响因素中，浴室数和是否在市区对房源的价格影响最大，是否为整租、可容纳人数和卧室数对房源的价格影响也比较大，可见房源的基础设施和基本属性对房源价格起着决定性作用。房东的基本情况、在线预订规则和房客的评价信息虽然对房源价格有影响，但不会构成主导因素，不过在拥有同等房源的情况下，房主提高这些软实力必然会取得较好的收益。

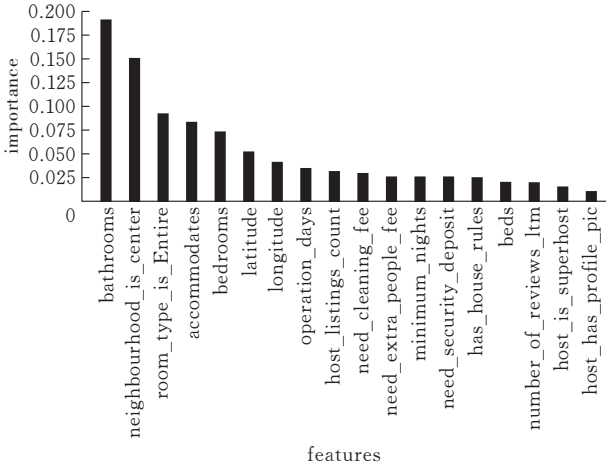

图1 XGBoost模型中各特征的重要性排名

5 结束语

本文通过OLS回归和分位数回归对短租房价格的影响因素展开研究，借鉴了传统的分析手段，解释性比较好，接着使用相比于线性回归模型精确度更高、更优越的XGBoost构建价格预测模型，当然还有其他机器学习算法值得探索。另外，在模型调参这个环节也有继续研究的必要。本文使用的网格搜索法比较费时，得到的是局部最优值，可以考虑使用其他优化算法进行调参，提高建模效率。

[参 考 文 献]

[1] 李鹏，陈雪均.国内共享住宿研究综述[J].商业经济，2020(6)：49-53.

[2] 宋玲玲，王时绘，杨超，等.改进的XGBoost在不平衡数据处理中的应用研究[J].计算机科学，2020(6)：98-103.

[3] Wang C，Deng C，Wang S. Imbalance-XGBoost：leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost[J]. Pattern Recognition Letters，2020(5)：35-37.

[4] 黄卿，谢合亮.机器学习方法在股指期货预测中的应用

研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析[J].数学的实践与认识，2018，48(8)：297-307.

[5] Parsa A，Movahedi A，Taghipour H，et al.Toward safer highways，application of XGBoost and SHAP for real-time accident detection and feature analysis[J]. Accident Analysis & Prevention，2020，136(4)：58-66.

[6] 佚名.短租数据集[EB/OL].(2020-09-01).https://tianchi. aliyun. com/competition/entrance/231715/information.

[7] Falk M，Larpin B，Scaglione M. The role of specific attributes in determining prices of Airbnb listings in rural and urban locations[J]. International Journal of Hospitality Management，2019，83(2)：132-140.

[8] 吴晓隽，裘佳璐.Airbnb 房源价格影响因素研究——基于中国 36 个城市的数据[J].旅游学刊，2019，34(4)：16-31.

[9] 薛洁，姚雨萌，吴霞.杭州共享住宿入住影响因素分析及预测——基于 Airbnb 爱彼迎平台数据[J].统计科学与实践，2018(12)：44-48.

[10] 郝令昕，丹尼尔奈曼.分位数回归模型[M].上海：人民出版社，2012.

[11] Chen T，Guestrin C. XGBoost：A scalable tree boosting system[J]. Knowledge Discovery and Data Mining，2016(8)：785-794.

[12] XGBoost.XGBoost parameters[EB/OL][2020-09-10]. https://xgboost. readthedocs. io/en/latest/parameter. html.

[13] 龚洪亮.基于 XGBoost 算法的武汉市二手房价格预测模型的实证研究[D].武汉：华中师范大学，2018.

The Application of improved XGBoost in the Prediction of Short-Term Rental Housing Price

ZHENG Lie，MU Xinyu

(*School of Sciences*，*Hubei Univ.of Tech.*，*Wuhan* 430068，*China*)

Abstract：This paper studies the prediction model of the housing price via the information of listings. Firstly，it processes the missing and abnormal values of the original data. Secondly，it constructs a reasonable feature system including 23 features for the influencing factors of short term rental housing price. Thirdly，it uses OLS regression and quantile regression to analyze the influence of these factors. Finally，it selects 18 outstanding features to construct the XGBoost model for predicting the price of listings. The model uses the grid search method to adjust parameters. The goodness of fit of XGBoost is 0.60，while that of linear regression is only 0.38. Therefore，the XGBoost combined with OLS regression and quantile regression，not only keeps the interpretation of the traditional model，but also improves the prediction accuracy.

Keywords：online short term rental；OLS regression；quantile regression；xgboost；grid search method

[责任编校：张　众]