Advances in Applied Mathematics 应用数学进展, 2021, 10(8), 2862-2867
Published Online August 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.108298

Hans汉斯

基于随机森林模型的重庆市二手房价格预测研究

康嘉玲

成都信息工程大学应用数学学院，四川 成都
Email: 1913320474@qq.com

收稿日期：2021年7月23日；录用日期：2021年8月15日；发布日期：2021年8月26日

摘　要

二手房价格的精准预测对购房者和政府对房地产政策的调控具有重要意义，本文以重庆市九大主城区2015~2020成交的二手房价格为依据，将影响房价的微观因素与宏观因素有机结合，运用Python使用随机森林算法对二手房数据集进行训练建模，最后结合岭回归、Lasso回归对训练结果进行比较，实验结果显示随机森林模型的误差最小，应用效果比较好，值得推广和应用到房地产价格评估中。

关键词

随机森林，岭回归，Lasso回归，二手房

Research on Price Prediction of Second-Hand Housing in Chongqing Based on Random Forest Model

Jialing Kang

Department of Applied Mathematics, Chengdu University of Information Technology, Chengdu Sichuan
Email: 1913320474@qq.com

Received: Jul. 23rd, 2021; accepted: Aug. 15th, 2021; published: Aug. 26th, 2021

Abstract

The accurate prediction of the second-hand house price is of great significance to the buyers and the government's regulation of the real estate policy. This paper takes the prices of second-hand

文章引用: 康嘉玲. 基于随机森林模型的重庆市二手房价格预测研究[J]. 应用数学进展, 2021, 10(8): 2862-2867.
DOI: 10.12677/aam.2021.108298

houses that were traded between 2015 and 2020 in the nine major urban areas of Chongqing as reference, organically combines the micro and macro factors that affect the housing prices, and uses Python and random forest algorithm to conduct training modeling on the second-hand house data set. Finally, ridge regression and Lasso regression are combined to compare the training results. The experimental results show that the error of random forest model is the smallest, the application effect is better, and it is worth popularizing and applying to the real estate price evaluation.

Keywords

Random Forests, Ridge Regression, Lasso Regression, Second-Hand Houses

1. 引言

住房不管在任何时期都是人们最基本的生活需求，与人民的生活品质密切相关，然而近几年我国各地房价持续上涨成为了社会的热点关注话题。与此同时，由于城市内可以用来建筑新房的土地越来越少，二手房交易市场越来越活跃，逐步取代新房在房地产行业占据主导地位。因此对二手房价格进行精准的预测，不仅可以为老百姓买卖房屋提供指导意见，并且对政府合理调控房地产交易市场具有重要意义。

近年来，随着互联网、大数据等技术的发展，有学者开始尝试使用数据挖掘、机器学习的新技术去预测房价，与其他方法相比，机器学习在非线性和多元统计上表现出了良好的应用前景[1]。本文使用随机森林算法[2]对重庆市二手房价格进行预测，并结合岭回归、Lasso 回归和回归决策树模型对训练结果进行比较，使用 RMSE 和 R^2 对模型进行评估，最后结果显示随机森林模型的均方根误差 RMSE 最小和拟合优度 R^2 最大，对二手房价格的预测误差最小。

2. 数据分析及预处理

2.1. 数据集分析

由于考虑到城乡地区二手房价格有一定的差距，因此本文仅对重庆市九大主城区的二手房价格进行分析及预测。本文的数据集有 96,346 条样本，每个样本包含了 20 个特征，其中 14 个为数值型特征，6 个为类别型特征。被解释变量为单位面积房屋价格(元/平方米)，解释变量分为宏观因素与微观因素，宏观变量包括经济因素和人口因素相关的 7 个特征，如地区生产总值，全体居民人均可支配收入、年末常住人口等，微观因素包括 13 个房屋基本信息相关的特征，如成交年月、房屋面积、所属区域、房屋户型等。其中宏观变量来自于重庆市统计年鉴，微观变量是使用网络爬虫技术于重庆市链家网站上获取，之所以选择链家网站进行数据获取是因为链家是以二手房交易起步的房屋交易平台，其房源信息的真实性、完整性和丰富性要高于其他平台。

2.2. 数据预处理

2.2.1. 数据清洗

在对数据进行分析处理之前，首先要对数据进行清洗工作，主要包括处理重复值、缺失值和异常值，

避免这些脏数据影响最终的预测结果。使用 Python 的工具包检查到本数据集不存在重复值与缺失值，然后使用箱线图分析法检测到三个异常点，对异常点进行删除，最终有效样本为 96,343 条。

2.2.2. 离散变量重编码

离散型变量是不能直接用于建模，需要对这些数据进行重编码，将字符型变量转化为数值型变量，比如本数据集中需要处理的字符型变量有户型、建筑结构、装修状况、所属区域、商圈等信息。

2.2.3. 特征归一化

对离散型变量进行数值化后，由于各个特征取值的大小不同，会造成特征空间中样本点的距离被个别特征值所主导，归一化[3]是为了将数据映射到 0~1 之间，去掉量纲的过程，使计算更加的合理，让不同维度之间的特征在数值上有一定的可比性，提高预测的准确性。公式如下：

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

2.2.4. 对非正态数据进行正态化

由于大多数机器学习不能很好的处理非正态数据，需要将偏态数据进行正态化，本文选择使用 Box-Cox 变换[4]，其一般变化形式为：

$$y(\lambda) = \begin{cases} \ln y, \lambda = 0 \\ \dfrac{y^{\lambda} - 1}{\lambda}, \lambda \neq 0 \end{cases}$$

其中 $y(\lambda)$ 表示变换后得到的新变量，y 表示原始连续变量，λ 表示变换参数，Box-Cox 变换要求原始变量 y 取值为正。经检查存在 3 个倾斜的数值型变量，使用 Box-Cox 变换将其进行转化，如图 1 为房屋面积转化前分布图和 P-P 图，图 2 为房屋面积转化后分布图和 P-P 图。

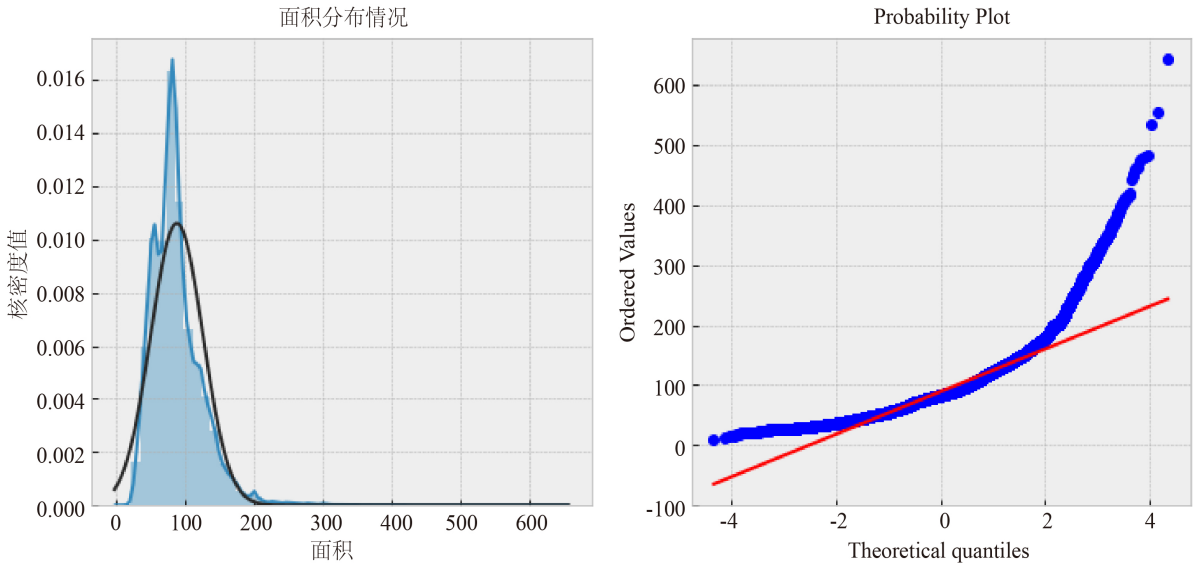

Figure 1. Curve: Distribution of raw data
图 1. 原始数据分布

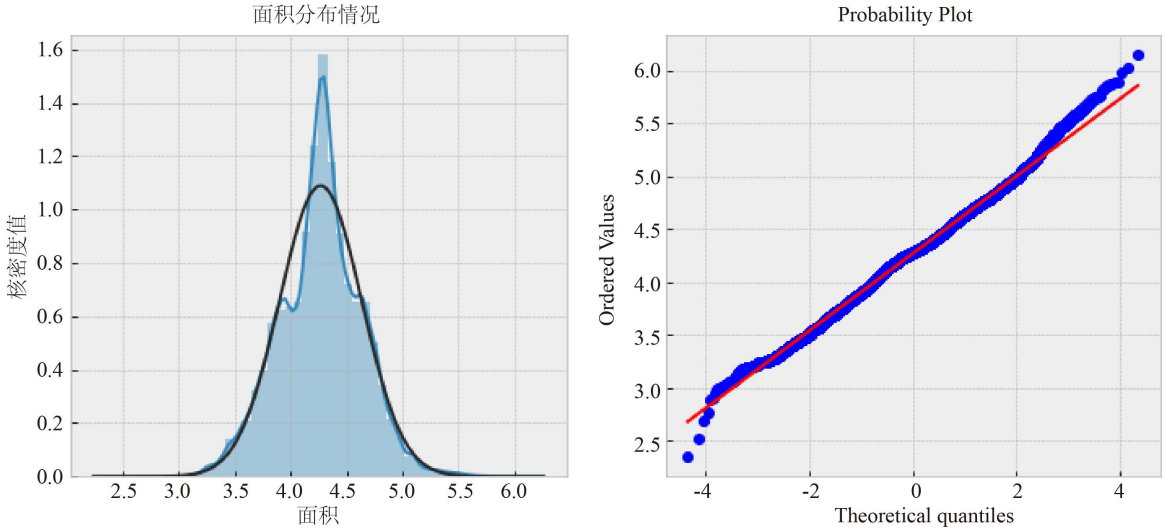

Figure 2. Curve: Data distribution after conversion
图 2. 转换后数据分布

3. 随机森林模型

随机森林是一种集成算法，既可以应用于分类问题也可以应用于回归问题，对于分类问题就是以少数服从多数的原则，将多颗决策树结果作为最终判断样本的类别；对于回归问题，样本最后的预测结果为多颗树的结果求平均值[5]。由于被解释变量单位面积房屋价格为连续型的数值，因此本文选择建立随机森林回归模型。由于计算量巨大，本文使用 Python 中的软件包 RandomForestRegressor 实现基于随机森林回归模型的二手房样本训练。

3.1. 相关参数的设置

使用 python 拟合随机森林回归模型时，主要调整的参数有随机森林所包含的决策树的个数(n_estimators)、决策树的最大深度(max_depth)、决策树根节点或中间节点能够继续分割的最小样本量(min_samples_split)、决策树叶节点的最小样本量(min_samples_leaf)。通过 10 重交叉验证的网格搜索，得到单棵回归决策树的最佳参数组合为 max_depth = 19，min_samples_leaf = 2，min_samples_split = 2，最后通过手动调参确定随机森林是决策树个数 n_estimators = 500，在测试集上得到的 RMSE 为 129.528，结果较为理想。

3.2. 模型评估

对模型的评估选择均方根误差 RMSE 和拟合优度 R^2 两种指标，均方根误差 RMSE 对模型的预测效果做定量的统计值，有关计算公式如下：

$$\mathrm{RMSE}=\sqrt{\mathrm{MSE}}=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n}}$$

其中 n 表示预测的样本量，y_i 表示被解释变量的真实值，$\hat{y}_i$ 表示被解释变量的预测值。MSE 或 RMSE 越小，说明模型对数据的拟合效果越好。拟合优度 R^2 是指对观测值的拟合程度，计算公式如下：

$$R^2=1-\frac{\sum_i\left(\hat{y}_i-y_i\right)^2}{\sum_i\left(\hat{y}_i-y_i^2\right)}=1-\frac{\sum_i\left(\hat{y}_i-y_i\right)^2/n}{\sum_i\left(\hat{y}_i-y_i^2\right)/n}=1-\frac{\mathrm{MSE}\left(\hat{y},y\right)}{\mathrm{var}\left(y\right)},$$

R^2 最大值为 1，R^2 值越接近 1，说明对数据的拟合效果越好。

将数据集按照 8:2 的比例划分训练集与测试集，经过调整参数之后在训练集上对模型进行训练，再将划分出来的 19,269 条测试集上的样本对二手房价格进行预测，预测结果如表 1 所示，模型在训练集和测试集上的评分如表 2 所示，由于测试集上的样本量非常多，因此表 1 仅选择 20 条进行展示，其中

$$\text{匹配度} = \frac{\hat{y}_i}{y_i}$$

反映了预测值($\hat{y}_i$)与真实值(y_i)之间的匹配情况。

$$\text{绝对误差} = \left|y_i - \hat{y}_i\right|$$

反映了预测值($\hat{y}_i$)与真实值之间的实际误差。从表 1 可以看到匹配度的取值在 0.9~1.01 之间，预测的效果较好。表 2 是分别在训练集与测试集上的均方根误差与拟合优度，从表 2 中可以看到测试集上的拟合优度为 0.99867，非常接近 1，说明建立的随机森林回归模型能较好挖掘训练集上各特征与被解释变量之间的关系并将其很好的拓展到了测试集样本上。

Table 1. Prediction results of random forest for test sets
表 1. 随机森林对测试集的预测结果

序号	实际成交价格(元/平方米)	预测价格(元/平方米)	匹配度	绝对误差
1	13,691.778276	13,693.030860	0.999909	1.252584
2	11,602.208059	11,584.158416	1.001558	18.049643
3	16,346.193633	16,390.281527	0.997310	44.087894
4	18,376.614015	18,348.623853	1.001525	27.990162
5	10,127.140774	10,135.135135	0.999211	7.994361
6	8589.388164	8603.104213	0.998406	13.716048
7	12,306.071281	12,291.666667	1.001172	14.404615
8	11,376.689881	11,372.549020	1.000364	4.140862
9	8008.016621	7936.507937	1.009010	71.508684
10	9774.693376	9805.903761	0.996817	31.210385
11	10,208.107475	10,203.090079	1.000492	5.017397
12	5132.411203	5173.865961	0.991988	41.454758
13	18,377.892153	18,419.556566	0.997738	41.664414
14	16,432.492217	16,437.355412	0.999704	4.863195
15	12,817.450620	12,813.852814	1.000281	3.597806
17	14,924.811060	14,917.127072	1.000515	7.683988
18	13,545.043557	13,542.688911	1.000174	2.354647
19	18,054.235728	18,040.435459	1.000765	13.800269
20	13,586.196783	13,584.905660	1.000095	1.291123

Table 2. Test set and training set score comparison
表 2. 训练集与测试集评分对比

	RMSE	R^2
训练集	57.91881	0.99973
测试集	129.52823	0.99869

4. 不同算法实例比较

对重庆市九大主城区二手房数据集同时使用岭回归模型、Lasso 回归模型[6]以及单个回归决策树模型，并与随机森林回归模型进行对比。各个模型在测试集上的表现如表 3 所示，从表 3 可以看到，在对测试集上的样本进行预测时，岭回归和 Lasso 回归 RMSE 略大于随机森林，回归决策树相对较好，但都不如随机森林算法。

Table 3. RMSE and R^2 of each model on the test set
表 3. 各模型在测试集上的 RMSE 和 R^2

模型	RMSE	R^2
岭回归	655.54513	0.966597
Lasso	658.21433	0.966325
回归决策树	251.25072	0.995093
随机森林	129.52823	0.998696

5. 结束语

本文以重庆市九大主城区二手房价格预测为基础，使用岭回归、Lasso 回归、回归决策树和随机森林，经实验证明，随机森林回归算法在测试集上取得的 RMSE 最小，R^2 最大，说明集成算法相较于传统的回归模型具有一定的优越性，基于二手房价格建立的随机森林回归模型可以推广应用到其他住宅价格预测中去，得到的房价预测具有一定的指导意义。

参考文献

[1] 曾婷婷. 基于机器学习的房价预测模型研究[D]: [硕士学位论文]. 绵阳: 西南科技大学, 2020.
[2] 李宇琪. 基于随机森林的房价预测模型[J]. 通讯世界, 2018(9): 306-308.
[3] 陈奕佳. 基于随机森林理论的北京市二手房估价模型研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2015.
[4] 张家棋, 杜金. 基于 XGBoost 与多种机器学习方法的房价预测模型[J]. 现代信息科技, 2020, 4(10): 15-18.
[5] 张倩. 基于随机森林回归模型的住房租金预测模型的研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2019.
[6] 时文静. 基于 Lasso 与数据挖掘方法的影响北京二手房价格的因素分析[D]: [硕士学位论文]. 北京: 北京工业大学, 2017.