# 基于 XGBoost 算法的房价预测模型

王冬雪，郭秀娟 *
（吉林建筑大学电气与计算机学院，吉林 长春 130118）

摘要：随着人工智能技术的快速发展，机器学习被广泛应用于各行各业，并在多个领域都取得了较好的成果。房价是一个影响因素复杂的热点问题，难以对其做出全面准确的预测。因此，本文尝试在一个相对稳定条件下将相关的机器学习算法应用于房价预测中。本文首先对数据进行缺失值处理、样本的因变量相关分析及标准化处理等一系列预处理，然后结合互联网数据和机器学习中的 XGBoost 算法对数据集进行建模和训练，最终实现对房价的精准预测。

关键词：房价预测；机器学习；XGBoost
中图分类号：F832.4 文献标识码：A 文章编号：2096-2118(2021)03-0079-04

## Housing Price Prediction Model Based on XGBoost

WANG Dongxue，GUO Xiujuan*
（*School of Electrical and Computer Engineering，Jilin Jianzhu University，Changchun Jilin* 130118，*China*）

**Abstract**：With the rapid development of artificial intelligence technology，machine learning has been widely used in all walks of life，and has achieved good results in many fields.Housing price is a hot issue with complicated influencing factors，so it is difficult to make a comprehensive and accurate forecast.Therefore，this paper attempts to apply the relevant machine learning algorithm to the housing price prediction under a relatively stable condition. In this paper，a series of preprocessing such as missing value processing，dependent variable correlation analysis of samples and standardization processing are carried out on the data at first.Then，the data set is modeled and trained by combining internet data and XGBoost algorithm in machine learning, and the accurate prediction of housing price is finally realized.

**Keywords**：housing price prediction；machine learning；XGBoost

## 0 引言

近年来，人们对房价的关注一直居高不下。由于城市化的加剧，对房屋租赁和购房的需求也持续增加，而房价问题不仅关系到人民生活水平，更是与国民经济发展息息相关。因此，对房价进行预测不仅对人们买卖房屋具有参考意义，而且对于政府进行房价调控也有积极作用[1]。因而确定一种可以精准反映房价走势的算法具有重要意义。

本文通过使用 XGBoost 算法来预测房价。通过对数据分析、预处理及基于 XGBoost 模型来构建房价预测模型。影响房价的因素多且复杂，如房屋面积、房屋地理位置、房屋户型等，本文仅选取对于房价影响较大的 79 个特征对房价预测模型进行评估，并选择 RMSLE 算法作为预测房价的评估算法。

## 1 数据预处理

在实际情况下，由于环境复杂等因素，我们获取的数据往往是存在缺失和异常的，因此，在建模前要对数据进行预处理。

### 1.1 数据集来源

收稿日期：2021-03-11
作者简介：王冬雪（1995-），女，吉林省长春市人，在读硕士研究生.研究方向：电气工程。
* 通讯作者：郭秀娟(1961~)，女，吉林省长春市人，教授，博士，研究方向：计算机图形图像处理，E-mail:779523836@qq.com

该数据采集来源于 Kaggle 2016 年竞赛项目，分为训练数据和测试数据。其中训练数据中有 1 460 个样本，每个样本中含有 80 个特征，测试数据中包含 1 459 个样本，每个样本中有 79 个特征。

### 1.2　缺失值处理

对于数据的缺失问题，要考虑 2 个重要因素：缺失数据的情况是否普遍，丢失的数据是规律的还是随机的。数据的缺失可能意味着样本量的减少，这可能对我们接下来的分析和建模产生阻碍。因此，对缺失值的处理是必需的。对于不同的数据缺失机制，处理的方法也是不同的，常用方法分为数据填充和删除样本两大类[2]。对缺失值的统计如图 1 所示。本文所使用的数据缺失值不能确定是否是随机的，例如缺失比例最高的游泳池质量，可能是由于很多房子里并没有游泳池。因此根据缺失值的统计结果，本文通过删除缺失比例较高的变量及带有缺失值的样本对缺失值进行处理。

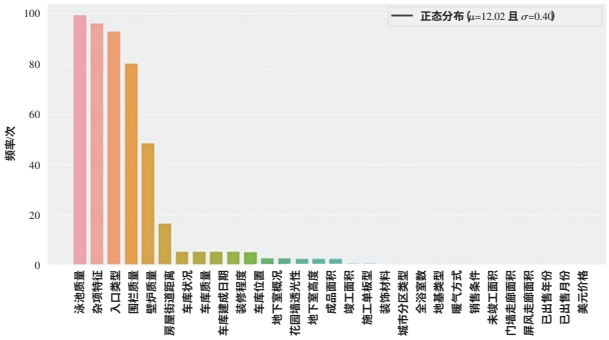



图 1　缺失数据统计图

### 1.3　样本因自变量相关分析

本文所使用的数据集虽然给出了 80 个自变量，但是有些因素对房价的影响非常小，如果不进行筛选可能会影响结果的准确性[3]。因此，进一步筛选后，得出对房价最具影响的特征依次为 :整体质量 (整体材料和装修) 、居住面积 (地面以上) 、车库容量、车库面积、地下室总面积、一楼面积、高档全浴室、客房总数 (不含浴室) 、建成年份。其相关矩阵热力图如图 2 所示。

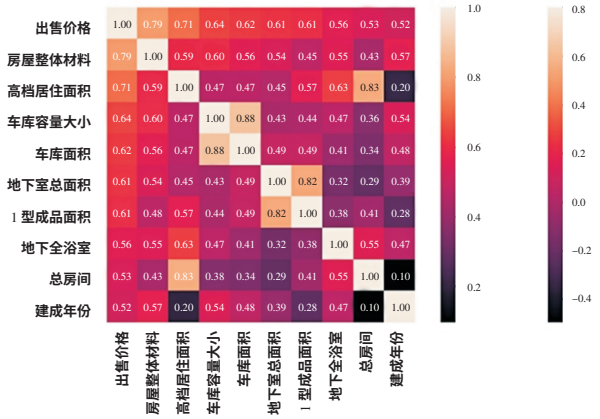



图 2　相关矩阵热力图

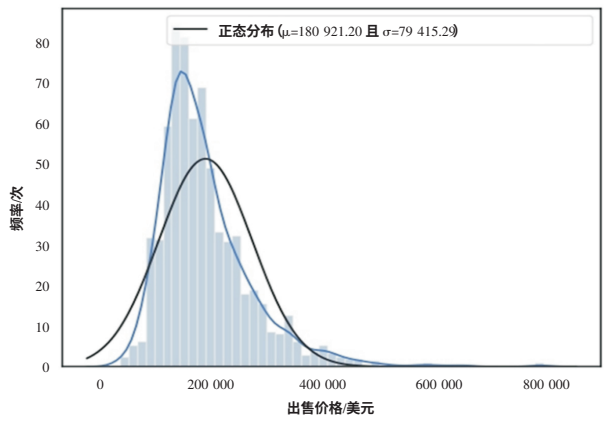
### 1.4　数据标准化处理

如图 3 所示，房屋的价格已经偏离了正态分布。因此，对于非正态分布，本文采用 Box–Cox 变换，使数据变得更“正常”。

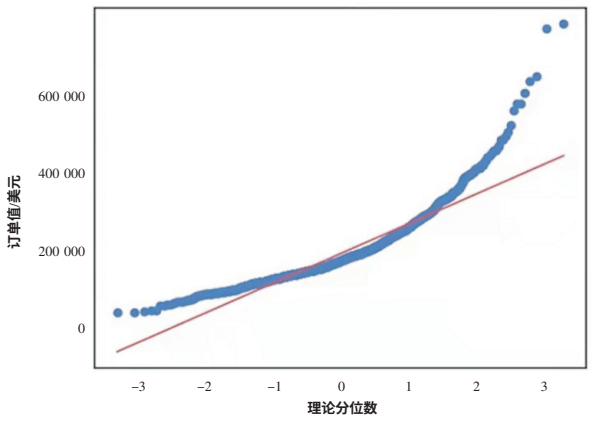
Box–Cox 变换的一般形式为：

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

式中 $y(\lambda)$ 为经 Box–Cox 变换后的响应变量，$y$ 为原始连续因变量，$\lambda$ 为变换参数。以上变换要求原始变量取值为正[4]。图 4 为经 Box–Cox 变换后分布。

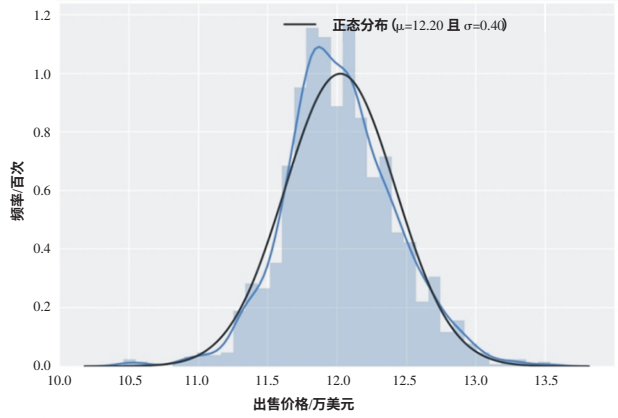



(a) 单价分布情况

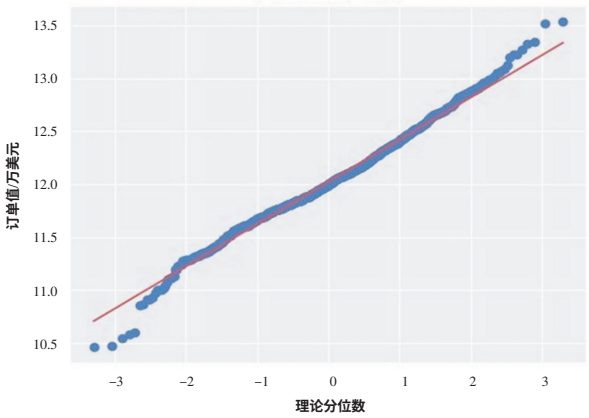



(b) 单价分布的 P–P 图

图 3　原始数据分布




(a) 变换后的单价分布情况




(b) 变换后的 P-P 图

图 4　变换后数据分布

## 2　XGBoost 模型

XGBoost 又称极端梯度上升，它是大规模并行 Boosted Tree，是 Gradient Boosting Machine 的扩展，在相同的环境和条件下，XGBoost 比同类算法快 10 倍以上[5]。XGBoost 还可以通过分布式运算，进一步提高训练速度[6]。

### 2.1　基本模型

XGBoost 是由 $k$ 个基模型组成的加法运算式：

$$\hat{y}_i=\sum_{t=1}^{k}f_t(x_i) \tag{2}$$

其中 $f_t$ 为 $k$ 个基模型，$\hat{y}_i$ 为第 $i$ 个样本的预测值。

损失函数可由预测值 $\hat{y}_i$ 与真实值 $y_i$ 进行表示：

$$loss=\sum_{i=1}^{n}l(y_i,\hat{y}_i) \tag{3}$$

其中 $n$ 为样本数量。

模型的偏差和方差共同决定了该模型的预测精度，模型的偏差具体表现为损失函数，模型越简单则其方差越小，所以目标函数由模型的损失函数 $loss$ 与抑制模型复杂度的正则项 $\Omega$ 组成，所以目标函数可表示为：

$$obj=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{t=1}^{k}\Omega(f_t) \tag{4}$$

其中 $\Omega$ 为模型的正则项。

以第 $t$ 步的模型为例，模型对第 $i$ 个样本 $x_i$ 的预测为：

$$\hat{y}_i^{t}=\hat{y}_i^{t-1}+f_t(x_i) \tag{5}$$

其中 $\hat{y}_i^{t-1}$ 是由第 $t-1$ 步的模型给出的预测值，$f_t(x_i)$ 是需要加入的新模型的预测值，则目标函数可写为：

$$obj^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{t})+\sum_{i=1}^{n}\Omega(f_t)=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{t-1}+f_t(x_i))+\sum_{i=1}^{t}\Omega(f_i) \tag{6}$$

而根据泰勒公式，可以把上述目标函数写为：

$$obj^{(t)}=\sum_{i=1}^{n}[l(y_i,\hat{y}_i^{t-1})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\sum_{i=1}^{t}\Omega(f_i) \tag{7}$$

其中 $g_i$ 为损失函数的一阶导，$h_i$ 为损失函数的二阶导。

将决策树定义为 $f_t(x)=w_{q(x)}$，$x$ 为某一样本，$q(x)$ 代表该样本所在的叶子结点，而 $w_q$ 则代表叶子结点取值 $w$，所以 $w_{q(x)}$ 代表每个样本的取值 $w$（即预测值），则目标函数的正则项可以定义为[7]：

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{8}$$

其中 $\lambda$ 和 $\gamma$ 是 XGBoost 定义的，其值可设定，值越大，表示越希望获得结构简单的树，$T$ 为叶子数。

再定义 $G_j=\sum_{i\in I_j}g_i$，$H_j=\sum_{i\in I_j}h_i$，则目标函数可记为：

$$obj^{(t)}=\sum_{j=1}^{T}[G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2]+\gamma T \tag{9}$$

而其中叶子结点 $j$ 对应的权值可表示为：

$$w_j^{*}=-\frac{G_j}{H_j+\lambda} \tag{10}$$

所以目标函数可简化为：

$$obj^{(t)}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{11}$$

记 $I_L$,$I_R$ 分别是数据集的左右结点,其中 $I=I_L\cup I_R$,则分裂后增益为:

$$L_{split}=Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}+\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\lambda \tag{12}$$

其中,$\frac{G_L^2}{H_L+\lambda}$,$\frac{G_R^2}{H_R+\lambda}$ 分别为左右树的分数,$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ 为分裂前分数,$\gamma$ 为惩罚项[8]。

XGBoost 在构建树的节点时,为每个节点添加了一个缺省方向,当样本缺失对应特征时,就会被归类到缺省方向上。如果样本存在特征缺失的情况,则只需分配到左右节点而无需遍历,故算法所需遍历的样本量大大减少。稀疏感知算法比 basic 算法速度快了超过 50 倍[9-10]。

### 2.2　模型评价

本文采用均方根对数误差(RMSLE)来作为模型评价的标准。其公式如下:

$$RMSLE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\log(\hat{y}_i+1)-\log(y_i+1)\right]^2} \tag{13}$$

其中预测值 $\hat{y}_i$ 为预测值,$y_i$ 为真实值,$n$ 为样本数量。

对训练集训练 100 次后的 RMSLE 为 0.041 646 875 398 8,如图 5 所示。XGBoost 能更好地适应不平衡的数据集,同时也更不容易过拟合,泛化能力较好,应用范围广泛,因此该模型基本可以实现对房价的精准预测。对于该预测模型,可应用到以下场景。

1)链家、安居客等二手房的交易。该模型更加有利于买卖双方看清房价接下来的走势,及时把握住期望成交价格。

2)房产投资的应用。近年来,从《新中产白皮书》中可以看到,新中产人群,除去自住房,投资性房地产占比是最多的。因此,该模型对于投资者有一定的指导性作用。

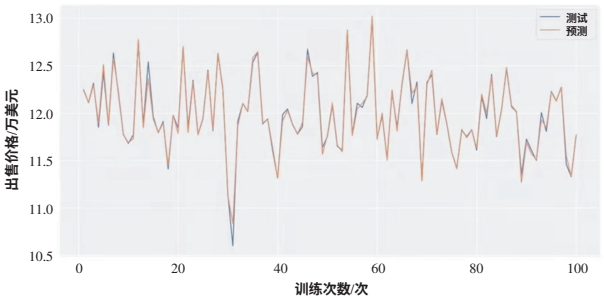



图 5　预测结果图

## 3　结论

房价预测问题本质上来说就是典型的回归问题。本文基于 XGBoost 算法进行房价预测,首先对数据进行缺失值处理、相关分析及标准化处理等一系列预处理,然后使用 XGBoost 算法对数据集进行建模和训练,最终实现对房价的精准预测。

## 参　考　文　献

[1]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):15-18.

[2]汪佳琦.基于机器学习方法的美国房屋价格影响因素分析[D].昆明:云南大学,2019.

[3]陈绵旺.基于 RS-SVM 的商品住宅价格预测研究[D].南昌:华东交通大学,2016.

[4]高建,周丽萍.基于 Box-Cox 变换的住宅特征价格理论研究[J].河北科技大学学报,2007(3):247-250.

[5]曾婷婷.基于机器学习的房价预测模型研究[D].绵阳:西南科技大学,2020.

[6]Byeonghwa Park,Jae Kwon Bae.Using machine learning algorithms for housing price prediction:The case of Fairfax County,Virginia housing data[J].2015,42(6):2928-2934.

[7]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.

[8]FLOREZ-LOPEZ R,RAMON-JERONIMO J M.Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment.A correlated-adjusted decision forest proposal [J].Expert Systems with Applications,2015,42(13):5737-5753.

[9]叶倩怡.基于 XGBoost 方法的实体零售业销售额预测研究[D].南昌:南昌大学,2016.

[10]Tianqi Chen,Carlos Guestrin.XGBoost:A Scalable Tree Boosting System[J].IEICE Transactions on Fundamentals of Electronics,Communications and Computer Sciences,2016,53(7):4903-4919.

编辑:杨　洋