Operations Research and Fuzziology 运筹与模糊学, 2023, 13(2), 734-744
Published Online April 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.132075

Hans汉斯

XGBoost算法在二手房价格预测中的应用

王　美[1]，朱珊艳[2]，陈鹏蕾[1]，武业文[1*]

[1]南京信息工程大学数学与统计学院，江苏　南京
[2]浙江科技学院理学院，浙江　杭州

收稿日期：2023年2月21日；录用日期：2023年4月7日；发布日期：2023年4月17日

摘　要

随着近些年社会经济的快速发展，房地产的开发也如火如荼展开，二手房市场也得到迅猛发展，在如今存量房时代，二手房交易成为房地产市场的重要部分，二手房的价格是诸多购房者的关注点，因此对二手房价格预测是有必要的。本文围绕XGBoost算法学习，借助网络爬虫技术，从链家网站采集了杭州九堡1500条在售二手房信息，将数据清洗并提取特征后，在Shiny平台进行展示，并以2023年2月江干区二手房均价34,381元/平米为分界线，将该地二手房分为高、低价格共两类。本文就影响二手房房价的因素进行深入研究，进而对房价进行分类预测。通过对影响二手房价格的特征因素提取并排序，结果显示第一重要的特征是房屋面积，其次是二手房屋关注的人数及发布时间、建筑结构、房间数量、房屋装修情况这四个特征，而房本时间、客厅和餐厅数量是重要性最弱的特征。本文基于XGBoost算法对房价预测，结果分类效果较为理想，说明算法的应用性较好，同时为后续我国二手房价格预测或其他问题的预测扩充探索的道路。

关键词

XGBoost，网络爬虫，分类预测，二手房

The Application of XGBoost Algorithm in Second-Hand House Price Prediction

Mei Wang[1], Shanyan Zhu[2], Penglei Chen[1], Yewen Wu[1*]

[1]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu
[2]College of Science, Zhejiang University of Science and Technology, Hangzhou Zhejiang

Received: Feb. 21st, 2023; accepted: Apr. 7th, 2023; published: Apr. 17th, 2023

*通讯作者。

文章引用: 王美, 朱珊艳, 陈鹏蕾, 武业文. XGBoost 算法在二手房价格预测中的应用[J]. 运筹与模糊学, 2023, 13(2): 734-744. DOI: 10.12677/orf.2023.132075

Abstract

With the rapid socio-economic development in recent years, real estate development has been in full swing and the secondary housing market has also developed rapidly. Nowadays, in the era of inventory, second-hand house transactions have become an important part of the real estate market, and the price of second-hand houses is a concern for many home buyers, so it is necessary to predict the price of second-hand houses. In this paper, around XGBoost algorithm, with the help of web crawler technology, 1500 second-hand houses for sale in Jiubao, Hangzhou are collected from the website of Chain Home, the data are cleaned and features are extracted and displayed in Shiny platform, and the average price of second-hand houses in Jianggan District in February 2023 is 34,381 Yuan every square meter as the dividing line, and the second-hand houses in the area are divided into two categories of high and low prices in total. This paper conducts an in-depth study on the factors affecting the price of second-hand houses, and then categorizes and predicts the price of houses. By extracting and ranking the feature factors affecting the price of second-hand houses, the results show that the first important feature is the house area, followed by the four features of the number of people concerned about second-hand houses and the release time, the building structure, the number of rooms, and the house decoration, while the time of the house book and the number of living and dining rooms are the features with the weakest importance. This paper is based on the XGBoost algorithm for house price prediction, and the results of the classification effect is more satisfactory, which indicates that the algorithm has better applicability, and also expands the path of exploration for the subsequent prediction of second-hand house price or other problems in China.

Keywords

XGBoost, Web Crawler, Classification Prediction, Second-Hand Houses

1. 引言

二手房市场以房源流动性强、交易活跃性高的特点始终保持着良好稳定的发展，在部分一线城市的市场占比已然超过新房市场。每一个房产销售者都会将李嘉诚说的话作为房产的卖点，即“决定房地产价值的因素，第一是地段，第二是地段，第三还是地段”。地段是影响房产当下价格的主要因素，也是影响房产未来一个时期价格变化的重要因素。如何综合考虑地段、房屋结构等因素，选择可量化的指标，对二手房价格进行合理有效的评估，是促进二手房交易市场稳定发展的关键。龚洪亮等人基于 XGBoost 方法对武汉市二手房价格进行预测，并与 LASSO 方法进行对比，结果表明 XGBoost 方法预测精度提升明显[1]。刘锋等人对重庆市 2012 年的统计年鉴房价数据，采用变系数模型预测，结论是拟合效果比现行回归模型更好[2]。魏云云等人通过采用灰度系数分析，选出了对西安房屋价格影响较大的一些因素，然后使用 BP 神经网络预测了价格，通过这种过程建立的预测模型加快了神经网络的训练速度，同时也得到了很好的预测结果[3]。

英国统计学家 George E. P. Box 曾说“All models are wrong，but some are useful”。没有模型能够完全正确，但它们确实能够刻画或表达出我们想要的能够解释现象的规律，尽管，有些时候这类规律并不

如线性模型那般能够明确表示，而是像诸多机器学习算法一样，规律被隐藏于黑箱中。XGBoost (Extreme Gradient Boosting)是近几年在 Kaggle 竞赛中较为流行的机器学习算法，其本质仍是基于决策树的，以梯度提升为框架的算法。其应用范围广泛，可以帮助解决回归、分类、时间序列等预测问题。

本文是以 XGBoost 算法学习为出发点，将其应用在二手房价格预测中，学习其在分类预测中的算法流程。基于机器学习算法和网络爬虫技术，对二手房价格进行预测，本文的技术路线如图 1 所示。首先，借助 R 软件中的 rvest 包从链家网站爬取杭州九堡所有挂牌的二手房信息，共 1500 套房源信息。其次，对爬取的房源信息进行特征提取，借助 Shiny 搭建搜索平台，对房源信息进行展示；最后，以 2023 年 2 月江干区二手房均价 34,381 元/平米为分界线，将高于该价格的房源定位为高价格类，低于该价格的房源定位为低价格类，在划分为训练集和测试集后，进行 XGBoost 分类模型的训练和测试。

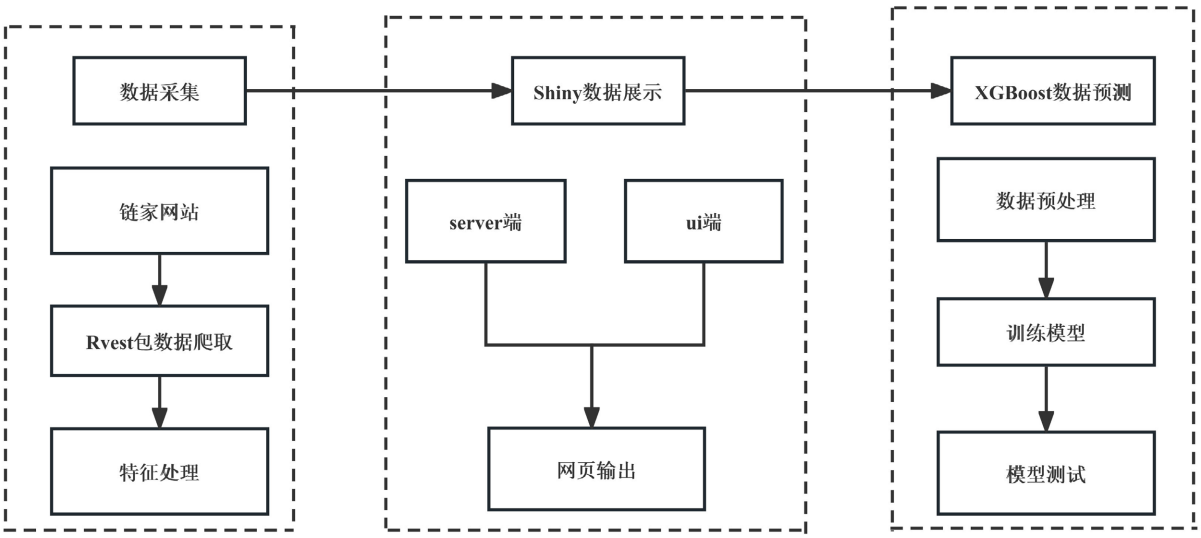

Figure 1. Technical roadmap
图 1. 技术路线图

2. XGBoost 理论介绍

2.1. XGBoost 预测原理介绍

XGBoost (Extreme Gradient Boosting)的基本框架是 boosting，其主要基础思想来源于决策树和集成学习，被广泛应用于分类和回归分析中。XGBoost 算法思想是：对所有特征的值进行排序，并存储为块结构，然后遍历所有分割点找到一个特征上最优的分割点，将数据分裂成左右子节点，其目标就是找到一组树，以这组树作为样本预测值，利用算法使得样本预测误差即目标函数达到最小的同时，还具有一定的鲁棒性和泛化能力。其公式推导流程如下：

第一步，构建目标函数 Obj，由模型损失函数 + 正则化形式构成，其中通过添加正则项防止模型出现过拟合。

$$Obj=\sum_{i=1}^{n} l\left(y_i,\overline{y}_i\right)+\sum_{i=1}^{t}\Omega\left(f_i\right) \tag{1}$$

其中 Ω 为正则项，l 为损失函数，n 为样本量，y_i 为样本的实际值，$\overline{y}_i$ 为模型第 i 个样本的输出值，t 为决策树个数，f_i 为第 t 棵树用于到叶子点的映射。

第二步，对损失函数 l 进行泰勒展开，记为：

$$L\left(y_i,\bar{y}_i^{(t-1)}\right)=g_i f_t\left(x_i\right)+0.5h_i f_t^2\left(x_i\right) \tag{2}$$

其中 $f_t\left(x\right)=\omega_{q(x)},\omega\in R^{\mathrm{T}},q:R^d\to\{1,2,\cdots,T\}$，用于定义一棵树，$T$ 为叶子节点总数，$\Omega\left(F_k\right)$ 为正则项，用来定义一棵树的复杂度。

$$\Omega\left(F_k\right)=\gamma T+0.5\lambda\sum_{j=1}^{T}\omega_j^2 \tag{3}$$

第三步，对叶子节点进行分组的目标函数核函数为：

$$\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_j\right)\omega_j+0.5\left(\sum_{i\in I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T \tag{4}$$

定义 $G_j=\sum_{i\in I_j}g_j$，$H_j=\sum_{i\in I_j}h_i$，则其最优点为：

$$\omega_j^*=-\frac{G_j}{H_L+\lambda},Obj=\frac{-1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\lambda T \tag{5}$$

第四步，将数据集分成两组，数据集在分裂后各自的损失记为 L_1 和 L_2，计算分裂后的收益：

$$\mathrm{Gain}=L_1+L_2-L=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}+\frac{\left(G_R+G_L\right)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{6}$$

然后利用最优切分点划分算法遍历所有的决策树节点找到目标函数。

XGBoost 最优切分点划分算法常见的有两种：一种是贪婪算法，一种是近似算法[4] [5]。贪婪算法是从深度为 0 的树开始，对每个节点的所有样本按照每个特征的特征值进行升序排列后，找到最佳分裂点并记录分裂收益 Gain，选择 Gain 最大的特征为分裂特征，并以此作为分裂位置。并对该叶子节点上新分裂的两个叶节点关联其样本集。依次重复直至满足条件。由于贪婪算法当数据量太大时无法进入内存找到最优解。针对这一缺点，近似算法只考虑每个特征的分位点，利用分位数策略，大大减少计算的复杂度，从而找到最优解。

2.2. XGBoost 模型参数

XGBoost 算法的作者将模型参数分为三类，分别是通用参数、Booster 参数和学习目标参数。

通用参数也叫一般参数，作用是对宏观函数进行控制。silent 默认值是 0，静默模式为 1，0 为连续输出消息。Booster 默认值为 gbtree，上升模型为树模型，gblinear 是指上升模型为线性函数。num_pbuffer 和 num_feature 均由 XGBoost 自动设置。

Booster 参数用于控制每一步的 booster。eta 默认值为 0.3，可取范围是 0~1，用于更新补偿收缩，防止过度拟合。gamma 默认值为 0，指定最小损失减少应进一步划分树的叶节点。max_depth 默认为 6，是指一棵树的最大深度，参数范围 1 到 ∞。min_child_weight 默认为 1，是指在子树中指定最小的海塞权重的和，参数范围 0 到 ∞。max_delta_step 默认为 0，意味着没有约束。subsample 默认为 1，指定训练实例的子样品比，设置为 0.5 意味着 XGBoost 随机收集一半的数据实例来生成树来防止过度拟合，参数范围是 0~1。colsample_bytree 默认值为 1，指定列的子样品比，范围是 0~1。

任务参数是指控制训练目标的表现，决定学习场景。base_score 默认值设置为 0.5，制定初始预测分数作为全局偏差。objective 默认值设置为 reg:liner，指定想要的学习类型，有线性回归、逻辑回归、泊松回归等。eval_metric 为指定验证数据的评估指标，例如回归中，默认 rmse，分类中，默认 error。seed 为

随机数种子，以便数据重现。

2.3. XGBoost 模型效果评价

在回归预测中，评价回归模型好坏的常见指标是 *MAE*、*MSE* 和 *RMSE*：

MAE (Mean Absolute Error)是平均绝对误差： $MAE = 1/m\sum_{i=1}^{m}\left|y_{test}^{(i)} - \hat{y}_{test}^{(i)}\right|$；

MSE (Mean Square Error)是均方误差： $MSE = 1/m\sum_{i=1}^{m}\left(y_{test}^{(i)} - \hat{y}_{test}^{(i)}\right)^2$；

RMSE (Root Mean Square Error)是均方均根误差： $RMSE = \sqrt{MSE_{test}}$。

在分类预测中，衡量学习器优劣的指标为 AUC (Area Under Curve)，通过绘制 ROC 曲线，计算曲线下方面积即为 AUC 的值。ROC 平面的横坐标是 false positive rate (FPR)，纵坐标是 true positive rate (TPR)。对某个分类器而言，我们可以根据其在测试样本上的表现得到一个 TPR 和 FPR 点对。将分类器映射成 ROC 平面上的一个点。调整分类器分类时使用的阈值，可以得到一个经过(0, 0)，(1, 1)的曲线，这就是此分类器的 ROC 曲线。一般情况下，这个曲线都应该处于(0, 0)和(1, 1)连线的上方。AUC 的取值范围是 0~1，越接近 1 说明分类效果越好，当小于 0.5 时，分类器意味着随机分类，没有意义。

3. 二手房价格分类预测

3.1. 网络爬虫

3.1.1. 基本原理及流程

本文进行数据采集的 R 语言包是 rvest，基本原理是通过编写自动化程序，在 html 树上找到相应的字符、数值等内容，存放回本地。

网络爬虫的流程如下：

1) 设置起始 URL 地址，放入待采集队列；
2) 自动化程序从待采集的队列中选一个 URL 地址，发送请求获取网页信息；
3) 信息采集后，将已采集的 URL 地址从队列中剔除；
4) 获取新的 URL 地址加入队列，直至遍历完所有待采集网页。

3.1.2. 功能实现

本文采集的是链家网站上杭州九堡二手房在售数据，2023 年 2 月 20 日关于该地区的在售二手房共 1500 套，共 50 页，每页 30 条信息。在此，批量采集二手房信息。

1) 初始 URL

杭州九堡二手房页面的初始网址为 http://hz.lianjia.com/ershoufang/jiubao/，这是首页，第二页的网络地址是 http://hz.lianjia.com/ershoufang/jiubao/pg2/，然后能够发现，当把/pg1/作为后缀加入首页地址后，其仍表示首页，因此，可以在 pg 后方设置页面循环参数 i，i 的取值范围为 1~50。

2) 批量爬取

本文较为关注的网页上二手房信息主要是六个 html_nodes，以下以房屋单价数据的批量爬取为例进行说明：

```
unitprice<-c()
for (i in 1:50){
  pathfile<-sub(pattern = "d", replacement = i, x = "http://XXXXX/pgd/")
  print(pathfile)
```

```
    position_name.temp <-read_html(pathfile) %>%
      html_nodes("div.positionInfo") %>%
      html_text()
    position_name<-c(position_name, position_name.temp)
    print(paste("... ... Page", i, "Done!"))
  }
```

首先，定义一个空的向量，用于存放爬取的批量数据；其次，编写循环函数，用于重复在多页上进行数据采集的行为，并设置网页上 unitprice 所在位置的 html_nodes；再次，将该 nodes 处的数据采集放入事先定义的空向量处；最后，完成以上步骤后，每爬取一页数据，输出一行提示：Page i Done!

3.2. 特征提取及数据展示

3.2.1. 特征提取

1) 本文采集了链家网页上与二手房相关的六个 html_nodes 的信息，分别命名为 position_name、totalprice、unitprice、houseinfo、tag 以及 attention。采集的少量结果如下表 1 所示：

Table 1. Display of second-hand house information collection results
表 1. 二手房信息采集结果展示

position_name	totalprice	unitprice	houseinfo	tag	attention
远洋心里－九堡	335	38,103 元/平	3 室 2 厅|87.92 平米|南|精装|11 层|板楼	近地铁 VR 看装修 房本满五年	8 人关注/ 12 天以前发布
泊林印象－九堡	350	39,375 元/平	3 室 2 厅|88.89 平米|南北|精装|18 层|板楼	VR 看装修 房本满五年	1 人关注/ 15 天以前发布
金雅苑社区－九堡	279	31,712 元/平	3 室 2 厅|87.98 平米|南西北|精装|11 层|板塔结合	近地铁 VR 看装修 房本满两年	35 人关注/ 23 天以前发布
保利梧桐语－九堡	535	38,833 元/平	4 室 2 厅|137.77 平米|南|精装|17 层|板楼	近地铁 VR 看装修 房本满五年	49 人关注/ 一年前发布

position_name 是指二手房屋所在的小区名称，如图 2 所示，每个小区名称之后都会附有小区所在的区域，本次数据采集的区域范围定在杭州九堡，因此该后缀统一，可用 gsub()函数以空格代替后缀进行处理，也能用 strsplit()函数以“-”对字符串进行划分，并保留前一组字符。前一种方法更简便，两种方法的具体代码如下：

- position_name<-gsub("-九堡","", position_name)
- position_name1<-strsplit(position_name, split = "-", fixed = T)

```
position_name2<-unlist(position_name1)
position_name3<-matrix(position_name2, ncol = 2, byrow = TRUE)
position_name4<-position_name3[,1]
position_name5<-gsub("      ", "", position_name4)
```

totalprice 是指二手房屋的总价，以万为单位，如图所示，该数据不需要进行清洗处理。

unitprice 是指二手房屋的单价，以元/平米为单位，如图所示，对该数据的清洗需要剔除“单价”和“元/平米”，同样也是使用 gsub()函数。

houseinfo 是指二手房屋结构信息，包括几室几厅、总面积、朝向、装修、楼层、建筑结构，共五个信息。信息之间以“|”进行分隔，使用 strsplit()函数对字符进行分割，并整理成每行六列的矩阵形式。

tag 是指二手房屋的标签，涉及到不动产权的时间，放初产权证明满两年和满五年所交的税额不同。部分二手房没有贴上产权证证明的时间，则被归为未满两年，在对分类数据进行赋值时需要用到。

attention 是指该套二手房屋关注的人数及发布时间。这两类信息间以“/”分割，处理方式与上述 houseinfo 一致，以 strsplit()函数进行分割。

首次数据清洗后的结果如下表 2 所示：

Table 2. Data display after the first cleaning
表 2. 首次清洗后数据展示

单价/元/平米	几室	几厅	面积/平米	朝向	装修	楼层高低	建筑结构	房本时间	关注人数	发布时间
38,103 元/平	3	2	87.92	南	精装	11 层	板楼	房本满五年	8 人关注	12 天以前发布
39,375 元/平	3	2	88.89	南|北	精装	18 层	板楼	房本满五年	1 人关注	15 天以前发布
31,712 元/平	3	2	87.98	南|西|北	精装	11 层	板塔结合	房本满两年	35 人关注	23 天以前发布
38,833 元/平	4	2	137.77	南	精装	17 层	板楼	房本满五年	49 人关注	一年前发布
24,648 元/平	2	1	58.83	南	简装	18 层	板楼	房本满两年	1 人关注	12 天以前发布

2) 数据清洗好后，需要将部分分类数据进行编码，便于后续预测分析。本文主要对装修、建筑结构以及房本时间三个特征进行赋值，具体如下表 3 所示：

Table 3. Description of classification feature assignment
表 3. 分类特征赋值说明

特征名称	含义	值
decoration	房屋装修情况	0：其他 1：毛坯 2：简装 3：精装
construction_type	建筑结构	1：板楼 2：塔楼 3：板塔结合
taxfree	房本时间	0：未满两年 1：满两年 2：满五年

特征赋值后的部分数据展示如下表 4：

Table 4. Data display for prediction and analysis after feature assignment
表 4. 特征赋值后用于预测分析的数据展示

totalprice	unitprice	room_num	living_num	room_square	decoration	construction_type	taxfree	attention
335	38,103 元/平	3	2	87.92	3	1	2	8
350	39,375 元/平	3	2	88.89	3	1	2	1
279	31,712 元/平	3	2	87.98	3	3	1	35
535	38,833 元/平	4	2	137.77	3	1	2	49
145	24,648 元/平	2	1	58.83	2	1	1	1

3.2.2. Shiny 数据展示

Shiny 是 R 语言的一个基于 Web 框架的可视化应用，可以对接数据源，生成图表和配置仪表盘。Shiny

的结构包括两部分，一部分是 ui 端，用于构建整个应用的布局，可以添加控件，设置布局方式的排列等等。另一部分是 server 端，用于构建控件与图形的关系，在服务器端展示数据。执行结果以网页形式打开，其框架代码如下所示：

```
ui <- fluidPage(
   titlePanel("XXXXX") )
server <- function(input, output) {
      print(str(diamonds))}
shinyApp(ui = ui, server = server)
```

本文借助 Shiny 平台，将采集到的 1500 条二手房屋数据以列表形式进行展示，同时具备关键词搜索功能。如下图 2 所示：

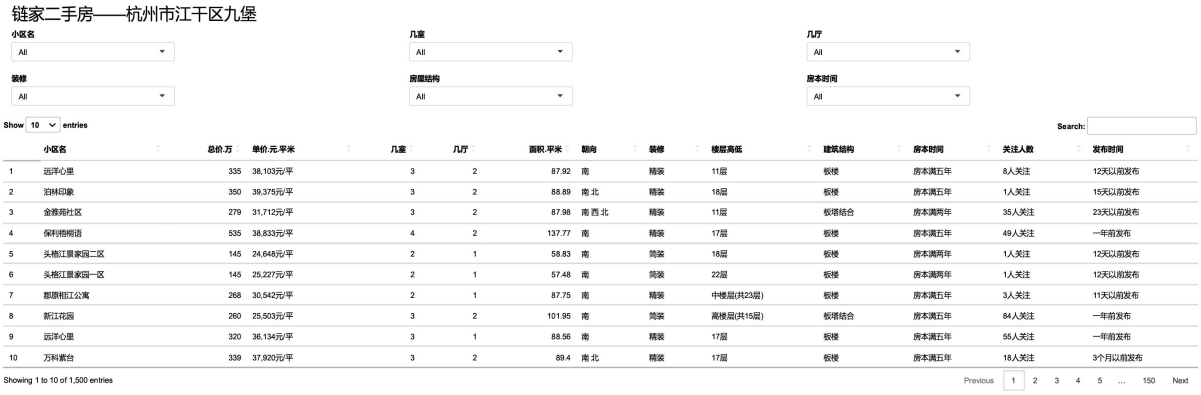

Figure 2. Shiny platform data display
图 2. Shiny 平台数据展示

可视化交互展示页面的标题是：链家二手房——杭州市江干区九堡。整体上分为上下两部分，上部分为六个控件，用于筛选，下部分为 table 输出，用于展示数据。另外，该页面还有 Search 控件，输入关键词能对信息进行选择性输出。

以下举个例子，设置条件：小区名保利梧桐语，4 室 2 厅，精装，板楼。如下图 3 所示，输出该小区符合条件的在售二手房为 3 套。

Figure 3. Display of condition filter results
图 3. 条件筛选结果展示

当在 Search 中输入“丽江公寓”进行搜索时，如下图 4 所示，共输出 61 条包含“丽江公寓”的二手房信息。

	小区名	总价.万	单价.元.平米	几室	几厅	面积.平米	朝向	装修	楼层高低	建筑结构	房本时间	关注人数	发布时间
1	远洋心里	335	38,103元/平	3	2	87.92	南	精装	11层	板楼	房本满五年	8人关注	12天以前发布
2	泊林印象	350	39,375元/平	3	2	88.89	南 北	精装	18层	板楼	房本满五年	1人关注	15天以前发布
3	金雅苑社区	279	31,712元/平	3	2	87.98	南 西 北	精装	11层	板塔结合	房本满两年	35人关注	23天以前发布
4	保利梧桐语	535	38,833元/平	4	2	137.77	南	精装	17层	板楼	房本满五年	49人关注	一年前发布
5	头格江景家园二区	145	24,648元/平	2	1	58.83	南	简装	18层	板楼	房本满两年	1人关注	12天以前发布
6	头格江景家园一区	145	25,227元/平	2	1	57.48	南	简装	22层	板楼	房本满两年	1人关注	12天以前发布
7	郡原相江公寓	268	30,542元/平	2	1	87.75	南	精装	中楼层(共23层)	板楼	房本满五年	3人关注	11天以前发布
8	新江花园	260	25,503元/平	3	2	101.95	南	简装	高楼层(共15层)	板塔结合	房本满五年	84人关注	一年前发布
9	远洋心里	320	36,134元/平	3	1	88.56	南	精装	17层	板楼	房本满五年	55人关注	一年前发布
10	万科紫台	339	37,920元/平	3	2	89.4	南 北	精装	17层	板楼	房本满五年	18人关注	3个月以前发布

Showing 1 to 10 of 1,500 entries
Previous 1 2 3 4 5 ... 150 Next

Figure 4. Display of conditional search results
图 4. 条件搜索结果展示

3.3. XGBoost 模型分类预测

3.3.1. 数据预处理

本文用于分类预测的变量特征为七个，分别是 room_num、living_num、room_square、decoration、construction_type、taxfree 以及 attention。而其中分类的类指的是房屋单价的高低类别，2 月杭州江干区的房屋均价为 34,381 元/平米，将高于均价的房屋归为高价格类，记为 1，低于该均价的房屋归为低价格类记为 0。

运行下方代码，对数据集进行缺失值查找，输出结果显示，从链家采集的数据不含有缺失值。

```
missing_data = data.frame(lapply(infodata,function(x) sum(is.na(x))))
```

随机选取 75%的样本作为训练集，剩下 25%的样本作为测试集，因此，训练集中含有 1125 个样本信息，测试集中含有 375 个样本信息。

3.3.2. 分类预测

划分好训练集和测试集之后，进行建模。这里设置 xgboost()函数的部分参数如下：max_depth = 6，eta = 5，nround = 25，objective = “binary:logistic”，其余参数均为默认值。

测试集的分类结果如下表 5 所示：

Table 5. Classification results
表 5. 分类结果

True\Pre	0	1
0	143	54
1	46	132

从测试集的分类结果大致可以发现，实际上，此次分类较为理想。而判断分类结果好坏的指标是 AUC 的值以及 ROC 曲线。因此得到 AUC 取值为 0.734，效果理想。ROC 曲线如下图 5 所示：

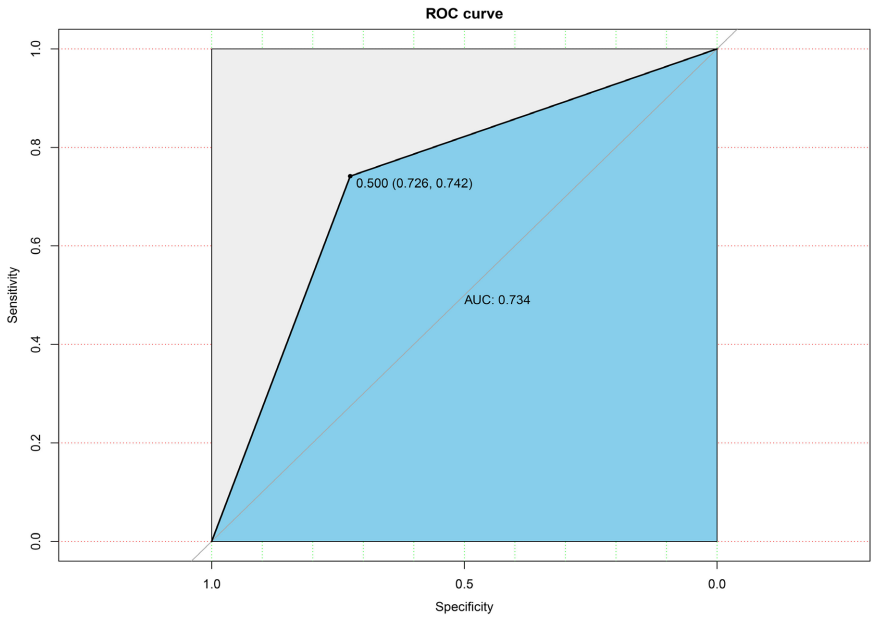

Figure 5. ROC curve
图 5. ROC 曲线

上图 5 中已显示 AUC 的取值为 0.734，高于 0.7，说明分类预测的有效果。

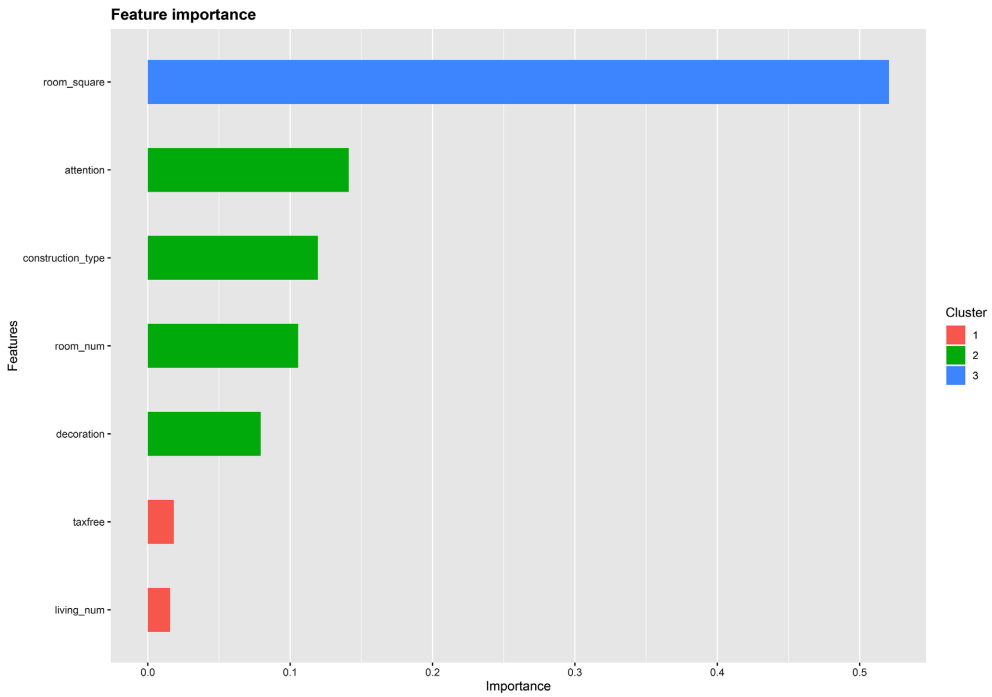

Figure 6. Characteristics importance ranking
图 6. 特征重要性排序

根据特征的重要性程度打分其中房屋面积(room_square)，可以分辨出对于模型来说哪些特征比较重要，可以基于这些特征完善特征选择的工作。图 6 所示的结果中可知：在此次预测中，房屋面积(room_square)是最靠前的特征，并且其重要性远远高于其他特征。与其对应的，重要性较低的特征为客厅和餐厅数量(living_room)和房本时间(taxfree)，即涉及到房产交易税费的特征。

4. 结论与不足

本文主要运用了机器学习中的网络爬虫技术和 XGBoost 算法，对从链家网站采集到的二手房数据进行清洗、建模并预测分类。在不进行参数调整等情况下，模型的分类效果较为理想。在本文选取的众多特征指标中，房屋面积是影响本文分类的重要因素。

结合本次实验的过程和结果，对本文的局限性提出以下几点思考：

第一，缺乏与其他分类预测技术的对比。理论上，XGBoost 算法能在 Kaggle 竞赛中风靡，主要是由于其预测精度高。而在本次实验中，缺乏这类对比，无法突出 XGBoost 算法的优越性。

第二，网络爬虫爬取的特征太少，最终参与分类模型的特征不足十个，并且众多影响房屋价格的因素并未列入，例如地理区位因素。后续实验在改进中，可考虑加入房屋的经纬度，将空间要素考虑在内，会更有参考价值。

第三，本次实验最终没有对模型参数进行调整，没有进一步调整得到关于该数据集的最优的分类结果。主要是时间有限，后续需要进一步完善。

参考文献

[1] 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]: [硕士学位论文]. 武汉: 华中师范大学, 2018.

[2] 刘锋, 张星, 张光锋. 重庆市房价变系数回归模型的建模与分析[J]. 重庆理工大学学报(自然科学), 2014, 28(4): 150-154.

[3] 魏云云, 张引娣, 陈晨. 基于灰色关联分析的 BP 神经网络对西安房价的预测分析[J]. 榆林学院学报, 2015, 25(4): 47-51. https://doi.org/10.16752/j.cnki.jylu.2015.04.036

[4] 山新们. XGBoost——从算法原理到近似计算[EB/OL]. https://zhuanlan.zhihu.com/p/94848125, 2020-09-20.

[5] 汤家正. 基于数据挖掘和 XGBoost 算法的量化多因子对冲模型研究[D]: [硕士学位论文]. 济南: 山东大学, 2020.