**基于多元线性回归的北京市二手房价格预测**

张书宁1)

1)(北京市第四中学, 北京, 中国)

**摘 要** 房地产是我国经济发展的重要支柱产业，也是人民生活的重要保障。通过收集1206条北京市二手房价格样本，建立多元线性回归模型分析北京市二手房价格的影响因素，得出二手房的楼层、结构、房龄等因素与二手房价格的关系。

**关键词** 多元线性回归；房地产；机器学习

**0** 引言

房地产是我国经济发展的重要支柱产业，也是人民生活的重要保障。住房价格是观察国家经济发展的重要指标，对房地产市场的发展和交易有着重要的作用。在研究某一城市的经济发展情况时，房地产的发展和价格是重要的衡量指标之一。但是在房地产行业的迅速发展过程中也出现了诸多的问题，尤其是房价的高速增长，房地产企业泡沫和预售烂尾楼等等。房价的增长也对城市的居民带来了很大的影响，许多一线城市的房价对城市居民造成了巨大的负担。因此房地产的价格也成为了政府和居民的关注点之一，所以本文将在相关研究的基础上研究北京市房价的影响因素和其与房价的关系，并根据住房的特征对房价进行预测。

在《基于Hedonic模型的小区房价影响因素研究 ——以南京市鼓楼区为例》中，从小区房价的微观影响因素出发，分别从建筑特征、邻里特征、区位特征三个方面确定了10个特征价格变量，运用定量和定性相结合的方法，构建影响南京市鼓楼区小区房价的特征价格模型，并对回归结果 进行检验和分析。结果表明：建筑年代、容积率、交通干道等级是影响小区房价弹性的最大因素；主要街道 等级是影响小区房价的最主要微观因素。在《南京市二手房房价影响因素的多元线性回归分析》中，通过分析南京市 8个区 12479个二手房数据，建立了影响单位面积房价的多元回归模型。通过比较 8个变量 对单位面积房价的影响，得到区域和是否有电梯对单位面积房价影响较大，面积及卧室数对单位面积房价影响较小的结论。许多城市都有同类型论文及类似研究，但关于北京市的多元线性回归二手房价格预测相关研究较少，本文将尝试进行相关的研究。

**1** 数据选取

* 1. **数据来源**

本文研究中所使用的全部数据均来源于链家房地产网站(<https://bj.lianjia.com>）。 在网站中共选取北京市中6个较为有代表性的行政区（朝阳、东城、房山、海淀、通州、西城）的数据，并整合到一起作为北京市的数据。数据的选取方法为从链家网站中各区的二手房页面的首页依次向后爬取4页的数据。

* 1. **数据爬虫**

研究中采取编写爬虫的方式采集数据。爬虫程序使用python中的beautifulsoup4,requests,lxml,

json等库抓取和分析数据。爬虫还使用到了腾讯地图相关的开放平台WebService API接口(https://lbs.qq.com），如请求search和explore接口，用于查询住房所在小区的位置、附近相关设施和到某一地点的距离。爬虫代码见附录。

爬虫爬取的原始数据如下图：



1. 数据处理
   1. **数据清洗**

需要先清除掉爬虫获取的样本中的多余无效特征。其次数据中有一定的缺失某项特征的样本，需要进行数据清洗的步骤清除掉残缺和无效的数据。如部分二手房数据中缺少房龄，则需要删除掉这部分数据，保证数据的完整性。

* 1. **样本特征指标**

原始数据中有一些中文描述无法直接进行建模，需要进行量化处理，转换成量化指标，指标的具体取值范围及相应的描述见下表。

基于其他学者的研究及获得的数据的特征，采取了下表所示的指标：

|  |  |  |  |
| --- | --- | --- | --- |
| **指标名称** | **具体说明** | **取值范围** | **描述** |
| ROOM | 房间数 | 正整数 | 房间数量 |
| FLOR | 楼层高度 | 0，1，2 | 0-低楼层 1-中楼层 2-高楼层 |
| AREA | 面积 | 正数 | 面积（平方米） |
| TYPE | 建筑结构 | 0，1，2 | 0-塔楼 1-板塔结合 2-板楼 |
| EVAL | 是否有电梯 | 0，1 | 0-无电梯 1-有电梯 |
| DISC | 行政区 | 0，1，2，3，4，5 | 0-朝阳 1-东城 2-房山 3-海淀 4-通州 5-西城 |
| YEAR | 建造年代 | 正整数 | 建造时间 |
| SUBC | 周围1000m地铁站数量 | 正整数 | 周围1000m地铁站个数 |
| SUBD | 最近地铁站距离 | 正数 | 最近地铁站距离 |
| CEDT | 距市中心距离 | 正数 | 距市中心距离 |

其中样本目标为每平方米价格**(PRIC)**。

1. 多元线性回归
   1. **多元线性回归方程**

其中 为系数， 为特征， 为截距。

* 1. **建立模型**

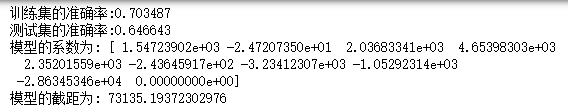
算法实现见附录[2]。

* 1. **目标预测**

建立模型后进行目标预测，需要先进行特征数据的标准化，代码见附录[3]。按照测试集占33%的比例划分测试集与训练集，设置随机数种子为66，代码见附录[4]。然后进行线性回归的训练及预测，代码见附录[5]。最后进行模型的准确性评估，代码见附录[6]，得出结论。

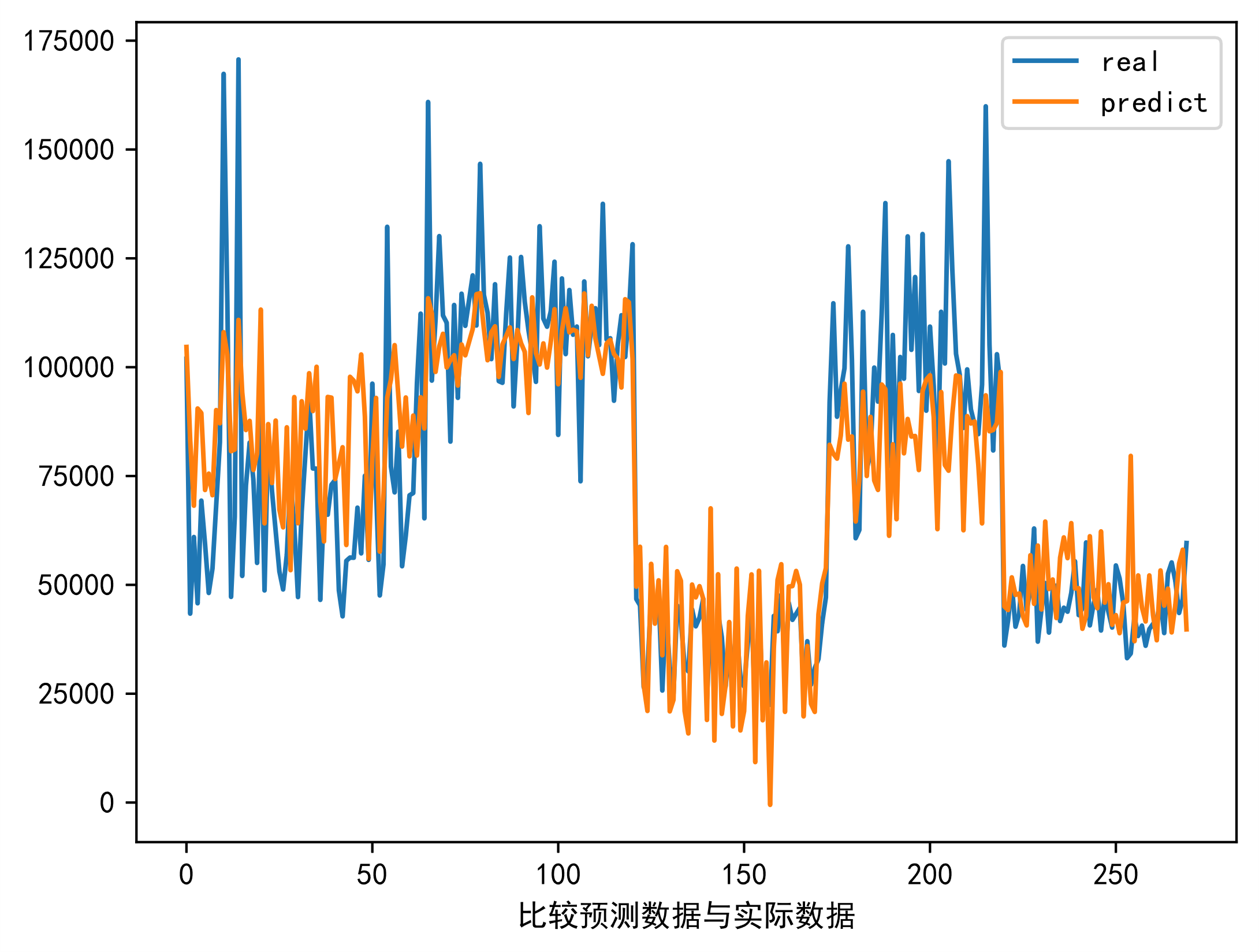
* 1. **结果分析**

1. 模型的系数与截距：

****

经评估，模型的训练集准确率为70.35%, 测试集准确率为64.66%，模型效果较为理想，与同类研究相比准确性较好。

1. 实际数值与预测数值对比：



可以看出实际数值与预测数值拟合度较好，准确性较高。但是在部分样本上存在较大的偏差，原因可能是部分本文研究中未考虑到的因素，如学区房、周边配套资源、商业价值等特征。

1. 结论

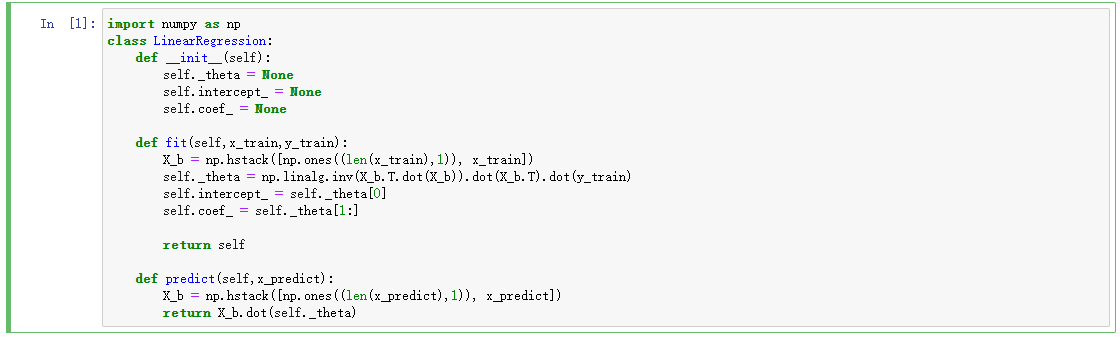
本文研究所建立的模型检验所得效果较为理想，可以获得一定的实际应用。但是研究中仍有很多不足之处，如有诸多的住房特征未在考虑范围之中，可能会对研究结果造成一定的影响。样本数据量不够大，可能使模型的预测不够精确。在此本文的基础上还会进行进一步的研究，改进研究的方法和样本量，优化模型。根据结果可知，在二手房房价的影响因素中，距市中心距离、楼房结构、房间数以及最近地铁站距离对房价影响较为明显，其中住房距市中心距离对房价的影响最为突出。

附录

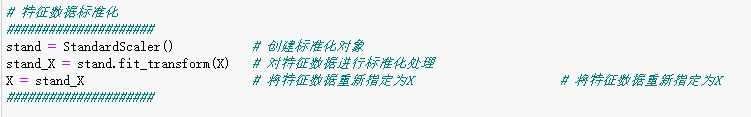
[1]数据处理代码



[2]算法实现

****

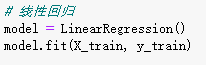
[3]数据标准化代码

****

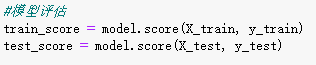
[4]划分测试集和训练集

****

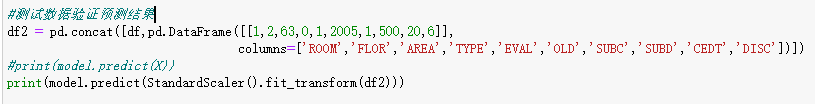
[5]线性回归代码

****

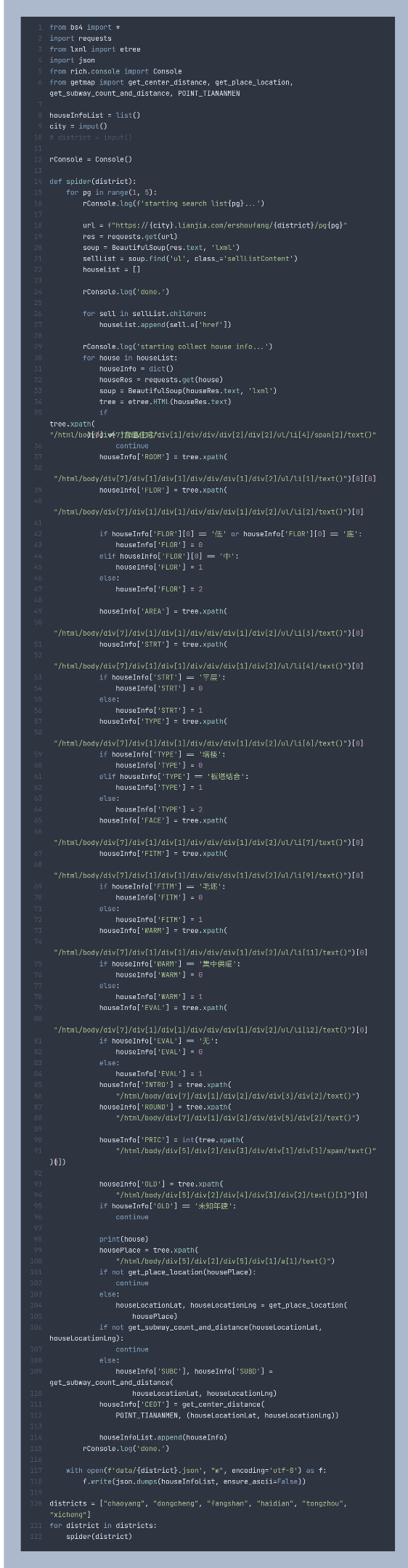
[6]模型评估代码

****

[7]比较验证预测结果

****

[8]爬虫代码



参考文献

1. 许基伟，马欣. 基于Hedonic模型的小区房价影响因素研究[A]
2. 陈多多，贾德铮. 上海徐汇区房价的影响因素分析[A]
3. 付东升. 基于机器学习的初中人工智能编程教学实践[A]