**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** 本文摘要内容置于此处 (英文摘要中要有这些内容)，字体为**小 5 号宋体**。摘要主要用来概括文章研究的问题、使用的方法和重要结论等，不属于正文部分。摘要不要过长，控制字数在 **300 字**以内。

**关键词** 关键词；关键词；关键词；关键词

**Title**

Name1)  Name1)  Name1)

1)(xxx School, Beijing, China)

**Abstract** Abstract (200-300 英文单词，内容包含中文摘要的内容). 注意，请不要逐字逐句直译中文摘要内容，而是根**据中文摘要的整体意思**进行内容表述

**Key words** key word; key word; key word; key word

**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，如北京2022年的房价均价为60119元/㎡，上海2022年的房价均价为49184.09元/m²，深圳2022年的房价均价为57335元/㎡，普通居民很难负担的起这样的高房价和繁重的房贷。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。

**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个行政区的数据（朝阳、东城、房山、海淀、通州、西城），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。

* 1. **爬虫代码**

1. 数据处理
   1. **数据清洗**

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。同时也需要清除掉爬虫获取的样本中的多余无效特征。

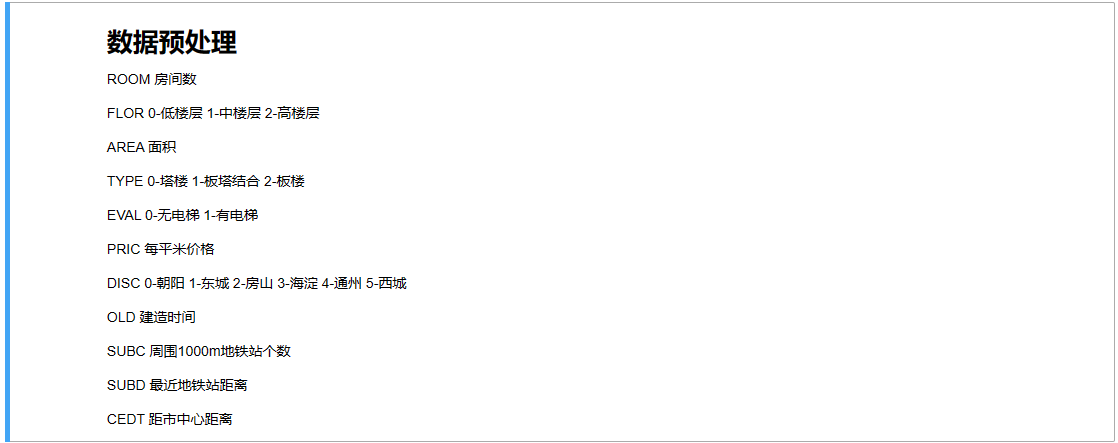
* 1. **加载读取数据**

具体代码如下：



* 1. **样本特征指标**

|  |  |  |
| --- | --- | --- |
| **指标名称** | **具体说明** | **取值范围** |
| **ROOM** | **房间数** | **正整数** |
| **FLOR** | **楼层高度** | **0，1，2** |
| **AREA** | **面积** | **正数** |
| **TYPE** | **建筑结构** | **0，1，2** |
| **EVAL** | **是否有电梯** | **0，1** |
| **DISC** | **行政区** | **0，1，2，3，4，5** |
| **YEAR** | **建造年代** | **四位正整数** |
| **SUBC** | **周围1000m地铁站数量** | **正整数** |
| **SUBD** | **最近地铁站距离** | **正数** |
| **CEDT** | **距市中心距离** | **正数** |

****

其中样本目标为每平方米价格（PRIC）。

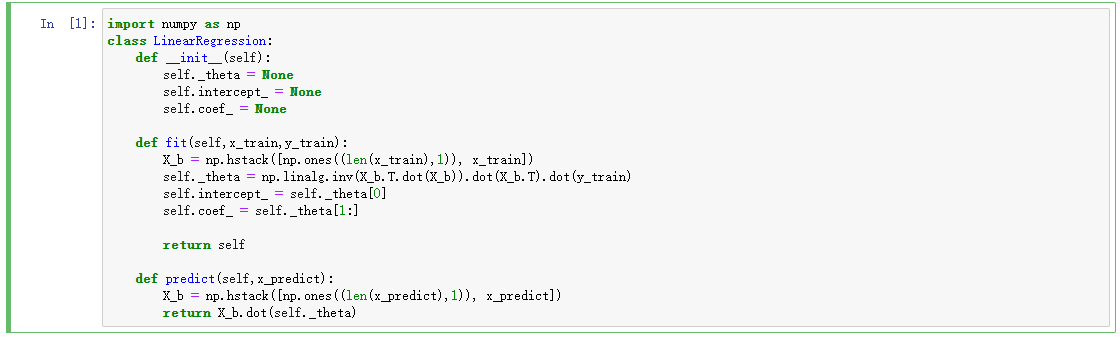
1. 多元线性回归
   1. **多元线性回归方程**

其中 为系数， 为特征， 为截距。

通过公式求出方程的系数与截距的最优解。

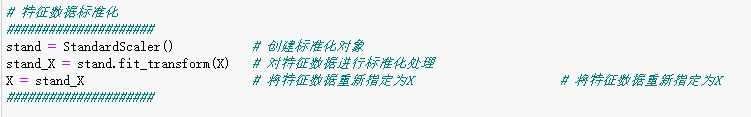
* 1. **建立模型**

以下为算法实现：

****

* 1. **目标预测**

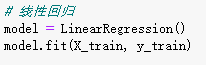
1. 特征数据标准化

****

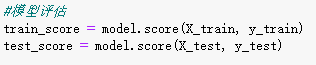
1. 划分训练集与测试集

****

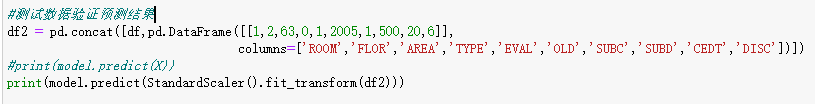
1. 线性回归

****

1. 模型评估

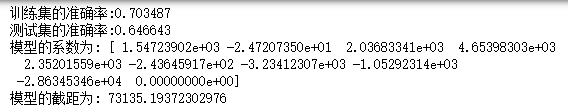
****

1. 验证预测结果

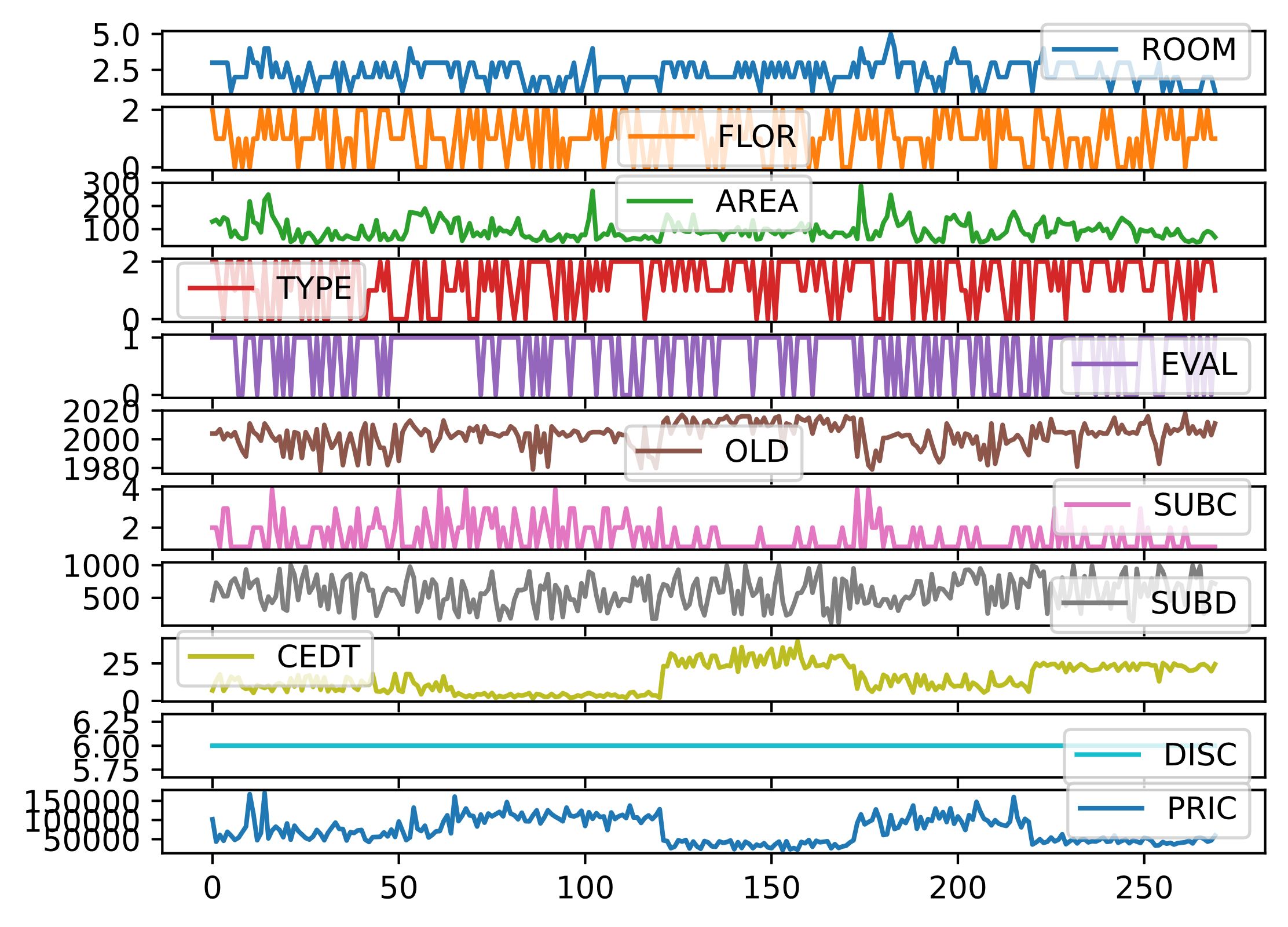
****

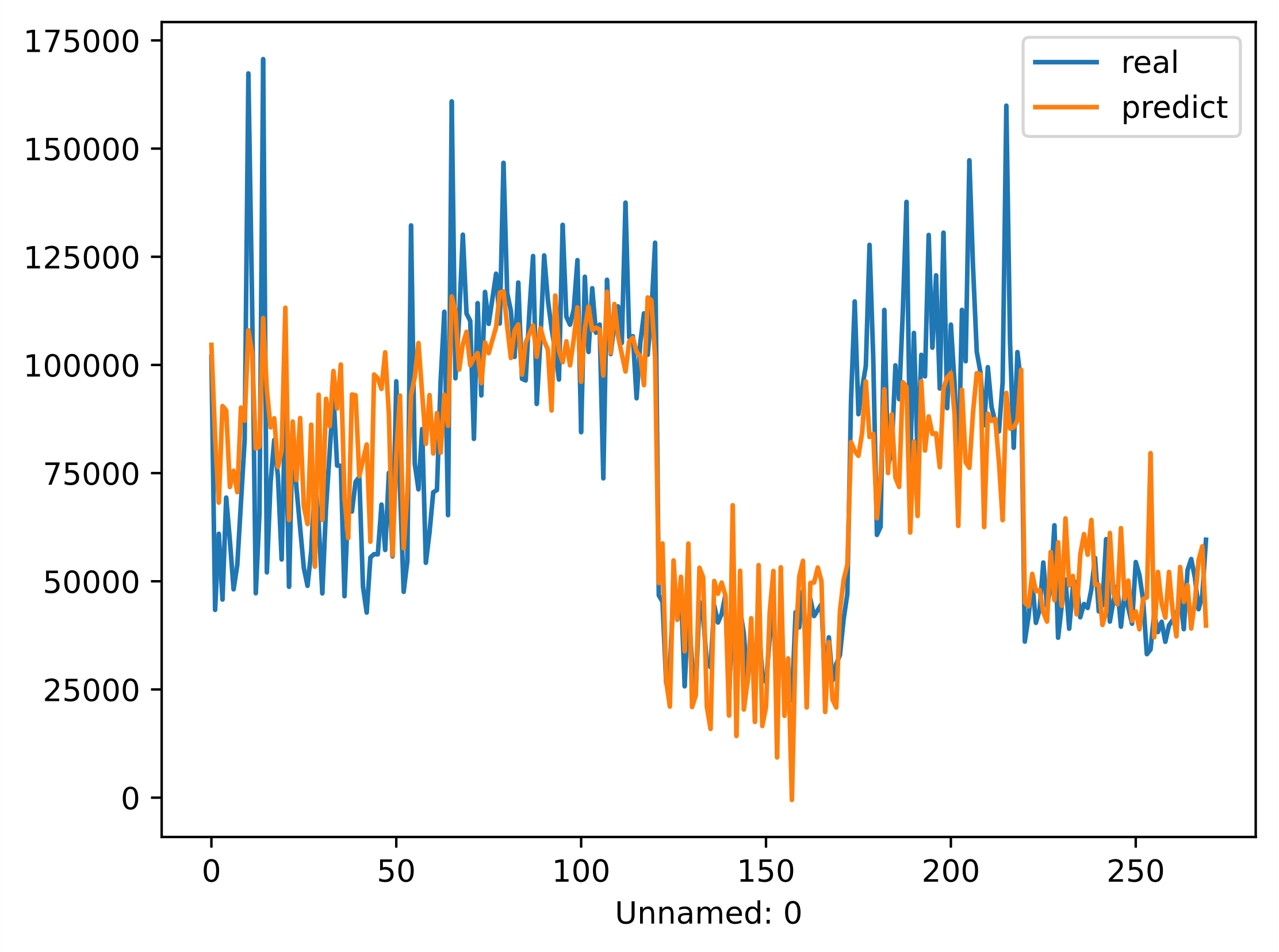
* 1. **结果分析**

1. 模型的系数与截距：

****

经评估，模型的训练集系数为0.703487, 测试集0.646643，模型效果显著，与同类研究相比准确性较好。

1. 特征与目标的折线图：
2. 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不错，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。

参考文献

1. 作者. 题目[J]. 刊名（全称）, 年, 期(卷):页码.
2. 作者. 文章题目[C]//Proceedings of the ... （会议名称）. 会议召开城市, 会议召开城市所在国家: [出版者不详], 年: 页码.
3. 作者. 文章题目[M]//编者. 文集标题. 出版地: 出版社, 出版年: 页码.
4. 作者. 书名[M]. 版次（初版不写）. 出版社地点: 出版社, 出版年.
5. 作者. 文章题目[D]. 单位地点: 单位名称, 年. [6] 作者. 文章题目[D]. 单位地点: 单位名称, 年.
6. 作者. 文章题目: 报告编号[R]. 单位地点: 单位, 年.
7. 专利拥有人. 专利名称, 专利授权国家[P]. 专利授权日期.