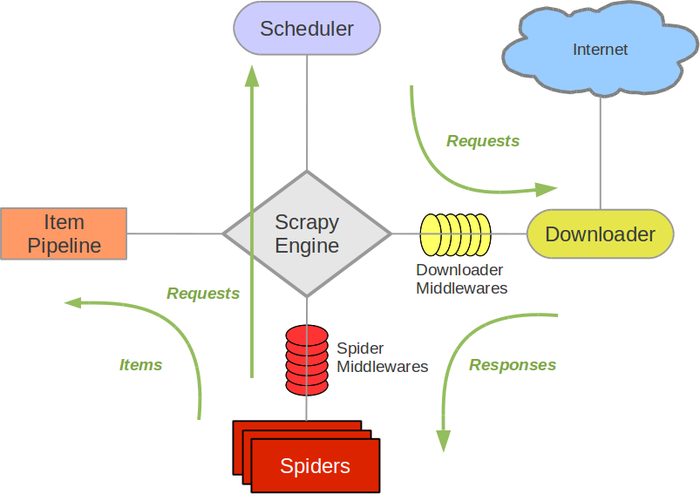
1. 数据采集

本次房价分析所选数据来自链家网站北京市8个行政区的二手房信息。数据采集方式选用python爬虫框架Scrapy实现，将采集的数据保存在MySQL数据库中。

* 1. Scrapy框架

Scrapy是用纯Python实现一个为了爬取网站数据、提取结构性数据而编写的应用框架，用途非常广泛。Scrapy 使用了 Twisted异步网络框架来处理网络通讯，可以加快我们的下载速度，不用单独去实现异步框架，并且包含了多种中间件接口，可以灵活的完成各种需求。



* 1. MySQL数据库

MySQL是一个轻量级关系型数据库管理系统，由瑞典MySQL AB公司开发，目前属于Oracle公司，因其体积小、速度快、总体拥有成本低，开放源码、免费等优点被广泛使用。

数据表设计： tbl\_house

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名称 | 字段含义 | 字段类型 | 备注 |
| Id | 房源编号 | Varchar（20） | 主键 |
| Title | 房源标题 | Varchar（255） |  |
| community | 小区名称 | Varchar（255） |  |
| model | 户型 | Varchar（255） |  |
| area | 面积 | Varchar（45） |  |
| Focus\_num | 关注人数 | Int |  |
| Watch\_num | 看房人数 | Int |  |
| Build\_time | 建造时间 | Varchar(255) |  |
| price | 总价 | Float |  |
| Average\_price | 每平均价 | Float |  |
| link | 房源链接 | Varchar(255) |  |
| Latitude | 房源纬度 | Float |  |
| Longitude | 房源经度 | Float |  |
| district | 所在行政区 | Varchar(45) |  |

1. 数据预处理

数据预处理过程主要包括缺失值处理、脏数据的清洗等等。如我们在数据中包含许多车位的信息，而不是房源信息，这部分数据约为29条，我们将其直接丢弃。

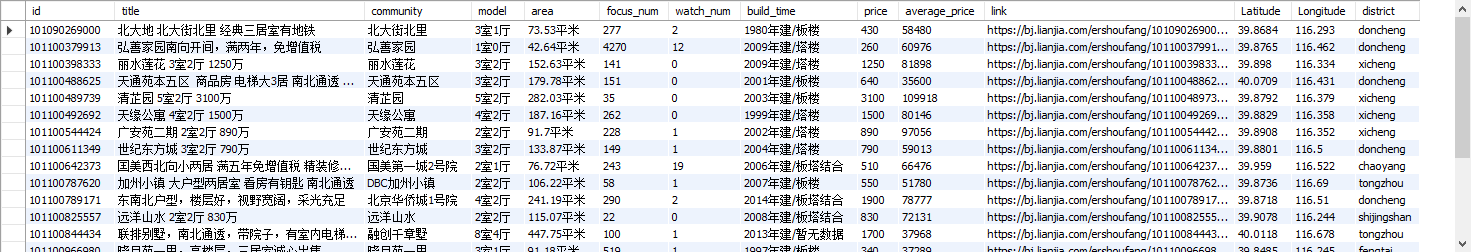
另外，在户型和房龄等计算时，有些字段内容为“暂无信息”，由于这些记录数量较少，我们也将这些记录直接删除。

同时，我们将一些类别型特征进行了one-hot编码，这样处理可以有效使得模型在训练中能够捕捉到相应信息。在本次数据处理中主要将房屋建造类型（平房、板楼、塔楼、板塔结合）以及行政区（东城、西城、朝阳、海淀、石景山、丰台、昌平、通州）等进行了one-hot编码。

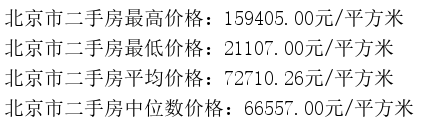
1. 探索性分析

3.1 房价整体分布

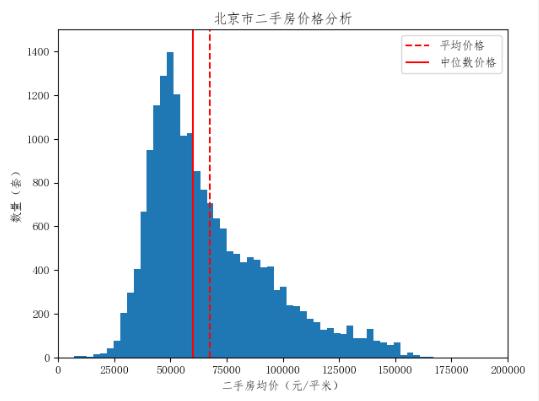
本次数据总计19359条，主要信息为小区、行政区、地址、面积、总价、建筑年份等等。示例数据如下所示：

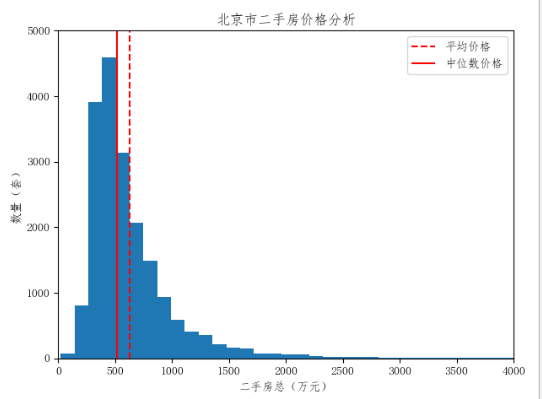


首先对北京市整体的房价水平做一个简单了解，包括全部房源的平均价格、中位数价格及价格分布。



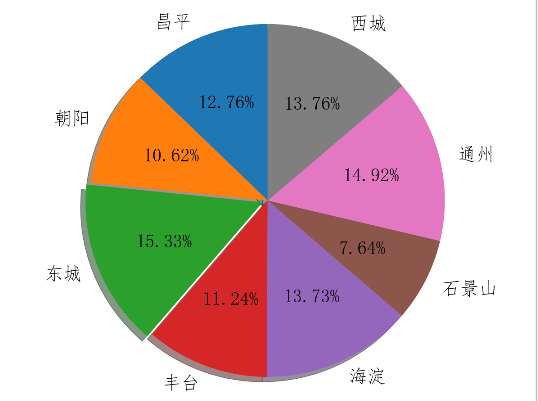
从上述统计结果来看，北京市二手房最高单价15.9万元/平米，最低价格2.1万/平米，房屋均价和中位数大致在6-7万元左右。下面是北京市二手房的整体价格分布：





从上图可以看到，无论是北京市二手房均价还是总价，其中位数和平均值十分接近，且数据分布类似于正态分布，呈现单峰特点，且中位数和平均值距离峰值很近。

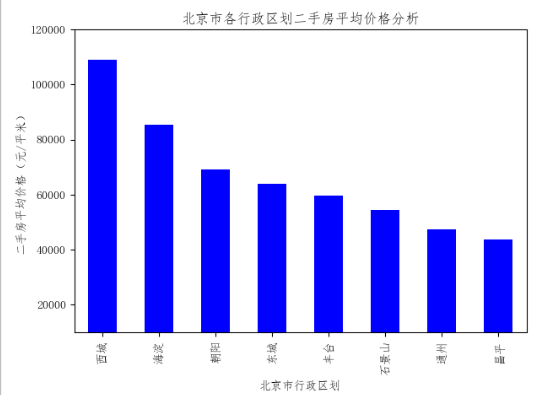
各行政区房源数量占比：



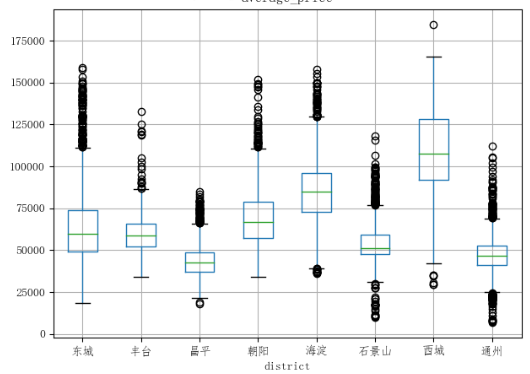
上图中，我们统计了所有采集的数据中二手房行政区的分布，总计包含8个行政区，基本呈现均匀分布，主要是东城、通州、西城、海淀，每个行政区平均占14.5%左右，其中石景山区二手房数量最少。

3.2 房价影响因素

* 行政区划影响

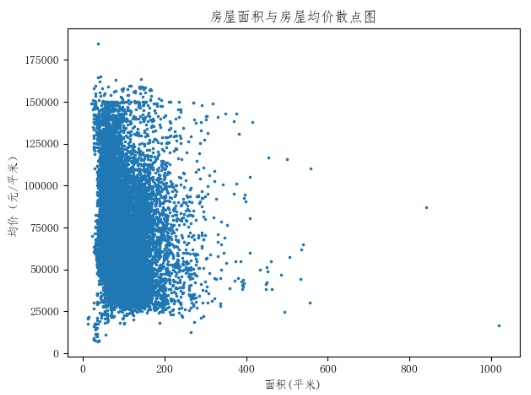


上图可以看出西城区的均价最贵，达到11万元/平米，其次是海淀区8.2万元/平米，然后朝阳区等等。由于房价分布的不均衡性，此处我们统计了各行政区内部的房价分布，绘成箱线图：



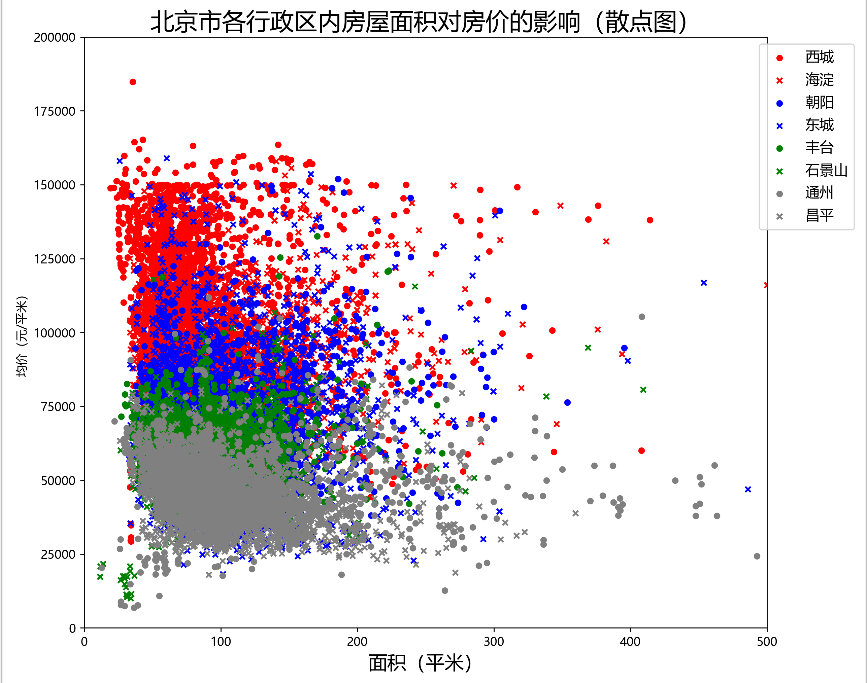
上图是北京市8个行政区二手房的箱形图，我们能看到东城区和通州区具有较多的上侧异常值，其中东城区最多，原因在于东城区作为传统中心城区，某些小区房价偏高。另外我们可以看到西城区的二手房房价均值最高，达到11万元/平米，这也与我们的常识相符。

* 房屋面积影响



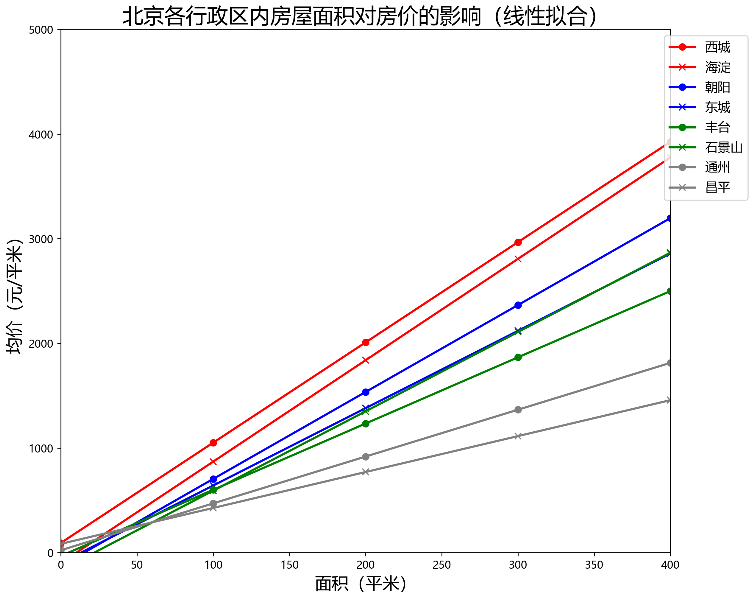
房屋面积也在一定程度上表达了房屋户型的相关信息，往往面积越大，户型卧室数量和客厅数量越多。从上图我们也能看出随着面积的变化，房屋均价变化不是很大，且面积小的房型均价高的更多，面积大的均价较低，这是由于小户型多集中在传统城区，老城区等，这些地区医疗、教育等资源较为充足，因此均价高。而大面积的别墅等多集中在郊区，这类户型每平米上均价较低。

各行政区内面积影响效果



从整个北京市的散点图看起来较为杂乱，我们结合行政区的信息，分别展示不同行政区所对应的散点图分布。同时为了体现各行政区房价分布的区别，图例按照平均房价的顺序排列，红色（西城、海淀）为第一梯队，蓝色（朝阳、东城）为第二梯队，绿色（丰台、石景山）为第三梯队，灰色（通州、昌平）为第四梯队。我们也能看到红色颜色的点多出现在图像左上方，另外三个梯度依次比上一个梯度整体向下偏移。

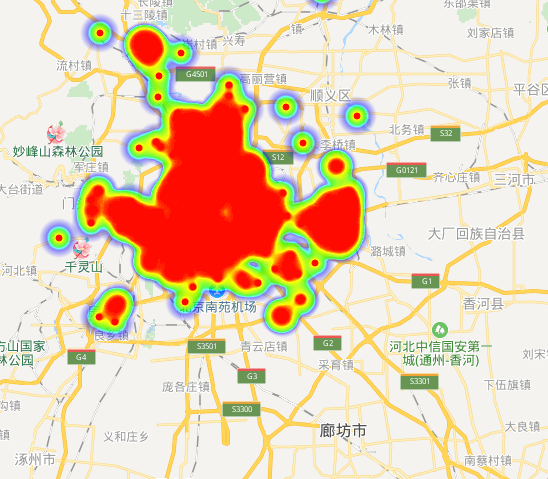
为了更明显的比较各行政区房屋面积对房价影响的规律，对各区的散点进行最小二乘线性拟合：

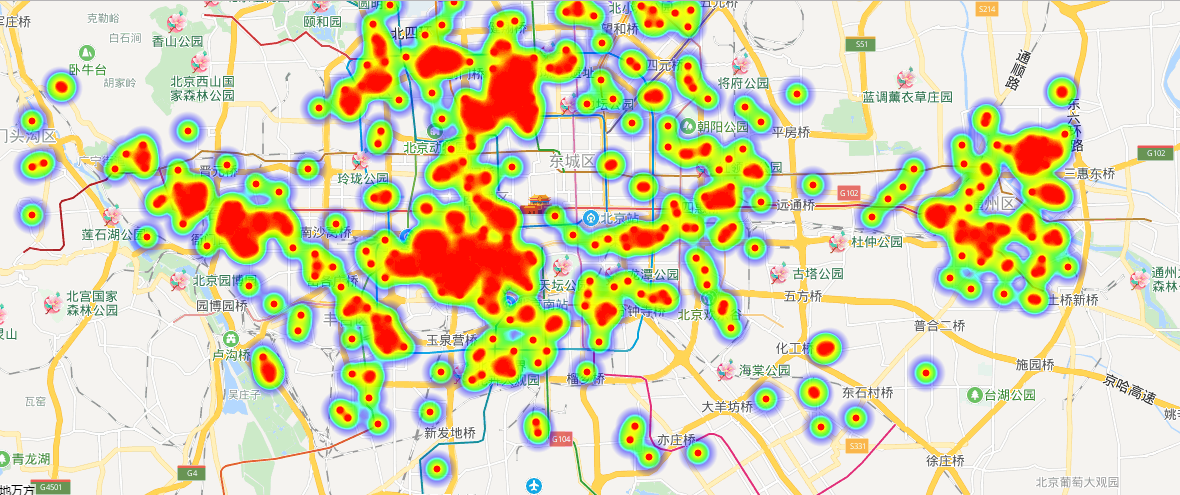


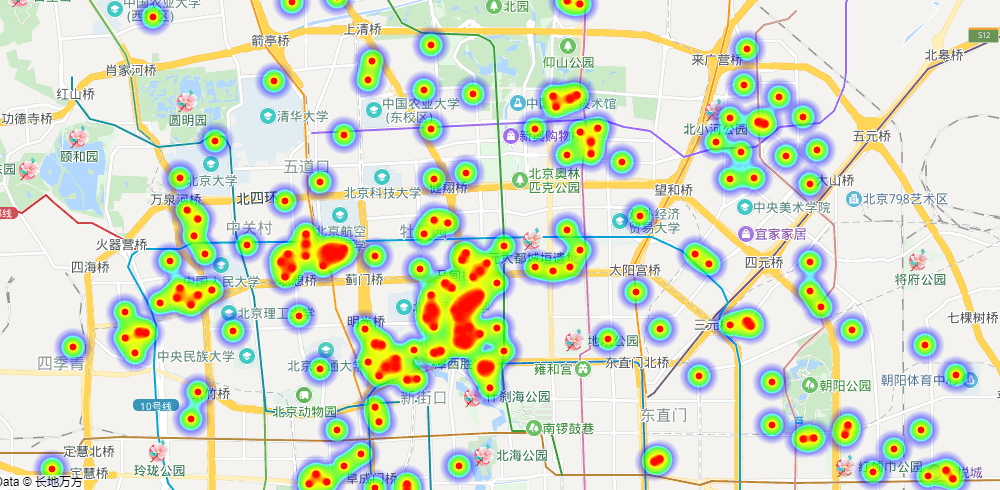
从线性拟合结果上来看，房价岁房屋面积变化规律与预期相符，即平均房价越高的行政区内，其房价随房屋面积变化的回归线越偏向左上方。

1. 房价地图热点分析

由于我们在房屋数据采集过程中，实时获取了房源对应位置的百度地图地理编码（经纬度信息），因此我们可以将各房源位置及对应价格直观展示在地图上，以此来详细观察影响房屋价格的重要地理因素，为我们后续的价格分析与预测提供参考。







上面我们从不同的缩放级别可以看到热力图的不同分布情况，首先在第一个级别，我们可以明显的看到地图上行政区之间的地段价格较低，行政区中心区价格较高。而且，中心区要比近郊区高。后面放大一些之后，我们能看到高价格地区基本围绕在重点高校、优秀初高中学校、地铁站等附近，我们可以一次推断出交通、教育等资源对于房价的影响还是比较重要的。

1. 房价预测

经过上述分析，我们对可能影响房价的特征进行了单独分析，直观感受了每个特征对于房价的影响，现在我们将使用多种机器学习模型对所得特征进行建模，希望利用已有信息对房屋总价进行准确预测。

5.1 数据预处理

5.2 数据集划分

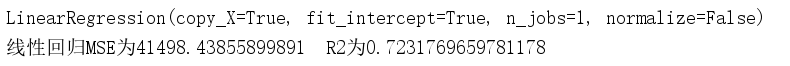
本数据集共计房源数据19359条，经清洗后有效数据18869条，按照0.7:0.3的比例将数据划分为训练集和测试集。划分后，训练集数量为13208条，测试集为5661条。实验中通过5折交叉验证的方式防止模型过拟合。

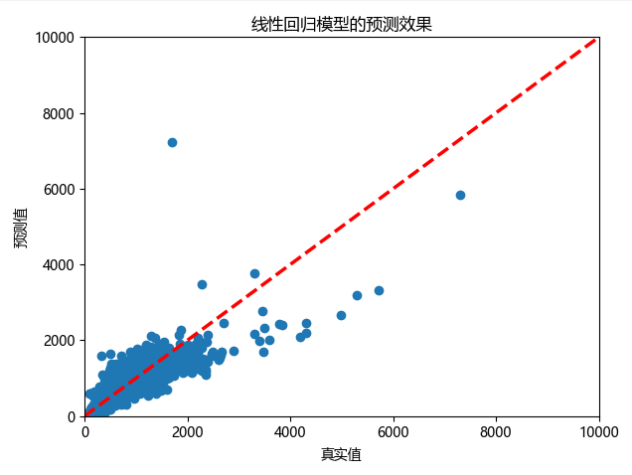
5.3 特征工程

本次房价预测模型中，我们将户型数据拆分为卧室数量和客厅数量，并根据房屋建造时间，计算房龄大小作为特征。此外房屋建造类型（平房、板楼、塔楼、板塔结合）作为类别型特征，使用one-hot进行编码，行政区（东城、西城、海淀、朝阳、石景山、丰台、昌平、通州）同样使用one-hot向量表达。

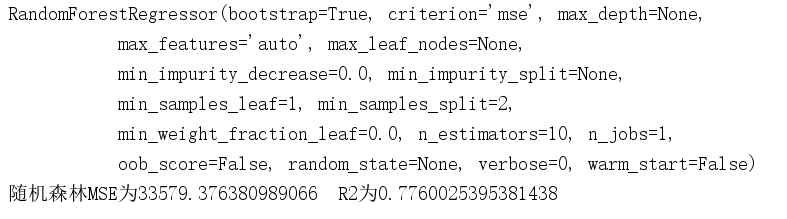
5.4 模型训练

* 线性模型预测如下：

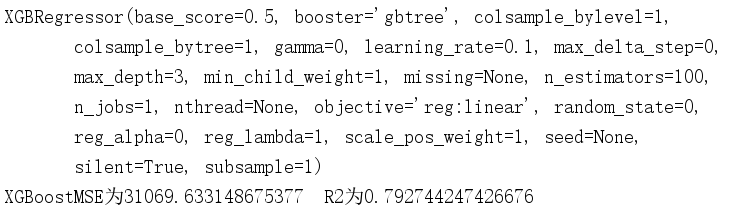


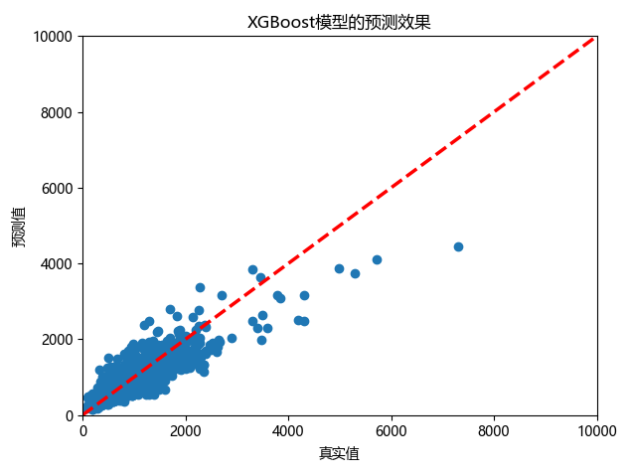


* 随机森林训练结果：

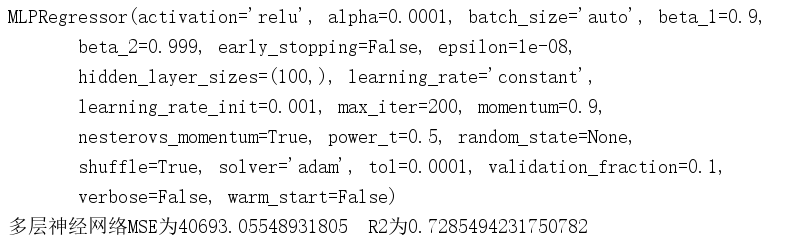


* XGBoost提升树模型





* 多层神经网络MLP训练结果：



5.5 效果与评价

* 评价指标

由于本问题是一个较为典型的回归问题，因此我们将使用均方误差（MSE）、来对模型进行评价。

假设为样本i的真实值，为模型预测值

均方误差计算公式：

计算公式：

越大越好，当我们的预测没有任何错误时，=1.它在统计学上用来表示模型的拟合程度。

* 各模型效果对比

|  |  |  |
| --- | --- | --- |
| 模型 | MSE | R2 |
| 线性回归 | 41498 | 0.7232 |
| 随机森林 | 33579 | 0.7760 |
| XGBoost | 31069 | 0.7927 |
| 神经网络 | 39583 | 0.7360 |

从上述模型的表现来看，集成模型（随机森林和XGBoost）表现要明显好于另外两种，原因在于集成模型使用多个弱模型来达到更好的效果。神经网络模型之所以没有表现出较好的效果，原因在于我们使用的数据量还比较小，而且特征数量偏少，并没有捕捉title中的文本信息，这可以作为后续的一个扩展方向。

1. 总结

本次数据分析中，我们使用了自己从链家网站爬取的北京市二手房近期的房源信息，并首先使用基础的统计方法对其进行了有益探索，同时结合经纬度信息对房价的分布有了直观的了解。之后我们使用机器学习的方法，通过提取特征，选择适合的模型，对其训练和预测评估，选择较好的结果。

当然，这其中还存在一些不足和改进之处。首先，我们本次并没有获取链家小区的有关信息，也就是小区地铁站数量、公交站数量、学校数量、商场数量等信息，这些信息实际上是有用的。同时本次爬取的数据量也稍微欠缺，仅有1.8w条左右，后续可以在特征和数据量方面持续改进。