网格化存量房评估模型简介

# 研究过程

## 数据采集

### 片区边界数据

#### 行政区划

行政区划数据来自电子地图。

#### 土地级别

土地级别数据来自国土局公布的官方数据。

#### 街道

街道数据使用专业的GIS软件QGIS进行绘制。

#### 学区

学区数据来自安徽各市教育局公布的官方数据。

### 影响因素数据

影响因素主要包括

## 数据分析

### 相关性分析

### 等级划分

对于有明确等级的因素，如土地级别、医院、景区等，直接以其对应的等级作为因素等级。而对于没有明确等级的因素，如学区、商场等，结合实地调查得到的规模、口碑等因素进行人为分级。

### 主成分分析

主成分分析是一种常见的数据降维手段，顾名思义，主成分分析就是找到数据中的最主要的方面来代替原始数据，能够起到降低数据噪声、去除冗余数据的作用。具体操作步骤如下：

#### 对样本数据进行中心化，即将样本中每个因素的均值置为0，；

#### 计算样本数据的协方差；

#### 求出协方差矩阵的特征值即对应的特征向量；

#### 将特征向量按照特征值大小进行排列成矩阵，取前k行组成矩阵P；

#### 即为将原始数据降到k维之后的结果。

## 数据处理

### 缺失值处理

在数据采集过程中难免会出现缺失值，如部分地区因为学校在建或者学区规划尚未公布而没有学区，这样就导致学区数据为NAN(not a number)。

### 偏态数据处理

线性回归理论中，模型的预测值和实际值之间会存在一个残差项： ，当这个残差项符合正态分布时才能获得一个无偏的模型。与之对应的训练数据X，预测目标Y都应该符合正态分布才能使模型更加准确。因此，在建立线性回归模型之前，对偏态分布的数据进行处理是很重要的一个步骤。

### 离群点处理

线性模型追求的是总体的误差最小，而原始数据中常常会出现离群点，比如某个市中的某个小区可能会出现一个房价远超该市均价的小区，为了使总体的误差最小，线性模型会偏向这种异常小区数据，所以这种小区会导致模型中对与该小区相似的小区价格被显著高估，导致模型失准，所以在训练模型时，要先去除这一类数据。

# 模型简介

## 岭回归模型

### 普通线性回归

线性回归模型的目的是将样本特征与样本标签之间的关系用线性方程的形式表示：

假设，则可以写成：

使用矩阵形式表示为:

而模型训练过程的优化目标是使预测数据与真实数据的差值的平方和最小，即最小二乘：

得到的优化结果是：

### 普通线性回归的问题

1. 一般的线性回归模型很容易出现过拟合的问题，所谓过拟合就是模型将当前训练集的特征作为了总体样本的特征，从而导致模型的推广能力变差，虽然模型在训练数据集中效果较好，但是对总体样本的其他部分评估能力较差。
2. 如果X本身存在线性相关关系（某些特征之间存在共线性）时，X的广义逆（）可能会变得非常不稳定。
3. 当变量比样本多的时候，回归系数会变得很大，无法求解。

### 岭回归原理

岭回归的模型方程与普通线性回归相同，但是在普通线性回归的平方误差项中添加了一个L2正则项，其损失函数为：

其优化结果为：

加入正则项主要有三个作用：

①岭回归系数解中的最小二乘法中X的广义逆（的基础上加入了一个小小的扰动，使得广义逆的求解变得稳定。

②岭回归损失函数中的保证了系数收缩过程的稳定性，W不会变得过大。

③岭回归是对参数W的有偏估计，它的结果使得残差平方和变大，但是能避免模型过多地学习数据集特征，减小过拟合的风险。

## 随机森林模型

### 回归决策树原理

#### 训练过程

输入：

数据集D；

房屋特征F={f\_1,f\_2,f\_3,…,f\_n };

价格P={p\_1,p\_2,p\_3,…,p\_n};

训练：

对将数据集D存入根节点

遍历 F集合中的每一个特征f\_m:

遍历 f\_m中每一个值a\_k：

按照a\_k的取值将数据集D分为f\_m≥a\_k和f\_m<a\_k两个数据集D1，D2；

计算DeltaError= Error(原数据集D)-Weighting\_Error(D\_1,D\_2);

从所有的DeltaError中筛选出最大的DeltaError，并根据对应的特征f和

取值a对数据集进行划分，得到数据集Set1，Set2；

将Set1存入左树，Set2存入右树；

分别对左树与右树进行划分；

直到数据分割完毕，每个叶节点中样本价格的平均数即为该叶节点的标记价格。

#### 预测过程

输入：

数据集D；

训练好的决策树T；

预测：

数据输入到决策树中；

根据每个节点的划分特征和划分值，将数据分割；

分割结束之后，每个样本最终进入的叶节点的标记价格即为该样本的预测价格。

### 随机森林原理

因为决策树采用的是一种贪心算法，单棵决策树泛化能力很差，需要通过集成学习的方式来提高泛化能力。集成学习方法要求基础分类器的性能不要太差，并且独立性高。

随机森林是一种通过建立多个独立性较强的决策树，再对所有的决策树预测结果进行整合的学习算法。

集成算法要求单棵决策树之间具有较高的独立性，但是所有的单棵决策树都是基于同一数据集的，为了提高独立性，随机森林算法采取了随机采样并随机选择特征集的解决方案。

#### 随机采样

随机采样使用的自助采样法（bootstrap sampling），从包含m个样本的数据集中进行分批次、有放回的随机采样，每个样本都有可能被多次选择或者不被选择，最终采集的样本数量与原数据集相同。

根据计算原数据集中将会有63.2%的数据出现在采样集中，未出现在采样集中的数据成为袋外数据，可以用于验证每棵决策树的准确率，对决策树进行剪枝或者初步筛选，避免太差的决策树拉低总体水平。

#### 随机选择特征

随机森林在传统的bagging算法中加入了随机选择特征的做法，进一步提高了单棵决策树之间的独立性，能够获得更好的集成效果，一般来说若总特征数为N个，每次建立单决策树都会随机选择log\_2⁡N个特征。

#### 结合策略

因为本模型是回归模型，所以采用平均法进行决策结果的整合，即将每个样本在不同决策树种的预测结果取平均值，即为最终的预测结果。

单棵决策树

单棵决策树

单棵决策树

结合模块

预测结果

数据集

另外可供采用的结合策略是学习法，即将每个样本在不同决策树种的预测结果作为新的特征，再根据新的特征和对应的房价训练出一个新的模型，将新模型的预测结果作为最终预测结果。这种方法计算较为复杂，在合肥的数据中表现与平均法相当。

# 模型效果

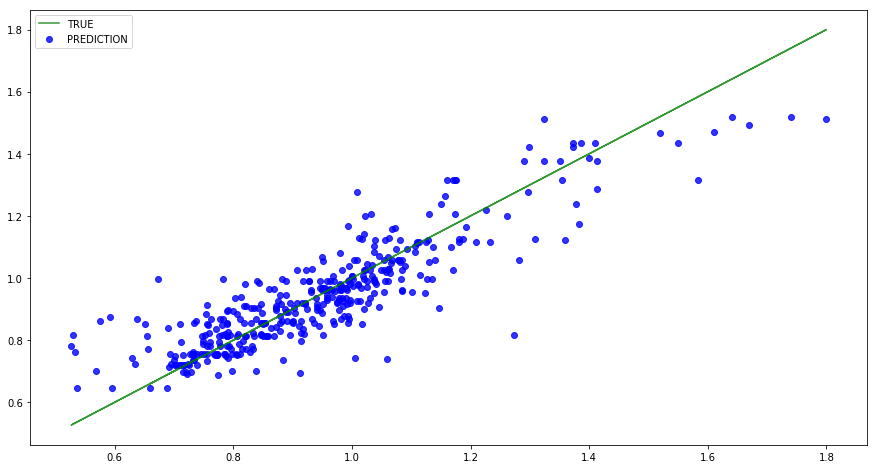
## 片区模型

片区因素

|  |  |
| --- | --- |
| 因素名称 | 因素重要性排行 |
| 行政区 | 3 |
| 街道 | 1 |
| 初中学区 | 4 |
| 小学学区 | 2 |
| 土地级别 | 5 |

公式:

模型误差率：8.3298%



模型

## 小区模型

### 特征重要性排行

本模型的特征重要性计算原理如下：

对将特征A中的所有数值打乱，重新训练模型，计算新模型与原模型的平均绝对误差的差值，即为特征A的重要度。

表格 2 特征权重

|  |  |  |
| --- | --- | --- |
| 因素 | features | importance |
| 周边工厂（工业园区）数量 | FACTORY\_NUM | 0.051444 |
| 周边幼儿园数量 | KINDERGARTEN\_GRADE\_SUM | 0.025858 |
| 周边公交站数量 | BUS\_STOP\_NUM | 0.022881 |
| 周边学校数量 | SCHOOL\_NUM | 0.022187 |
| 建筑年代 | built\_year | 0.022097 |
| 开发商品牌 | BRAND | 0.021795 |
| 周边医院数量 | HOSPITAL\_NUM | 0.019988 |
| 公交站等级和 | BUS\_STOP\_GRADE\_SUM | 0.019678 |
| 周边银行数量 | BANK\_GRADE\_SUM | 0.017754 |
| 周边商店数量 | SHOP\_GRADE\_SUM | 0.017300 |
| 周边地标数量 | LANDMARK\_BUILDING\_NUM | 0.016696 |
| 周边餐馆等级 | RESTAURANT\_NUM | 0.016052 |
| 周边酒店等级和 | HOTEL\_GRADE\_SUM | 0.015847 |
| 周边景区等级和 | SCENIC\_AREA\_GRADE\_SUM | 0.015788 |
| 周边超市数量 | SUPER\_MARKET\_NUM | 0.015017 |
| 周边商场等级和 | SHOPPING\_MALL\_GRADE\_SUM | 0.014879 |
| 地铁一号线 | SUB\_WAY\_1 | 0.013541 |
| 周边政府等级 | GOVERNMENT\_GRADE\_SUM | 0.012806 |
| 周边长途汽车站数量 | BUS\_STATION\_NUM | 0.011727 |
| 是否职工宿舍 | DORMITORY | 0.011229 |
| 地铁二号线 | SUB\_WAY\_2 | 0.010521 |
| 周边公园等级和 | PARK\_GRADE\_SUM | 0.010330 |
| 街道面积 | AREA | 0.009395 |
| 周边医院等级和 | HOSPITAL\_GRADE\_SUM | 0.009228 |
| 地铁三号线 | SUB\_WAY\_3 | 0.007814 |
| 地铁四号线 | SUB\_WAY\_4 | 0.006033 |

### 重要因素类别

表格 3 重要因素类别

|  |  |  |
| --- | --- | --- |
| 排行 | 因素 | 备注 |
| 1 | 行政区 | 瑶海区在同等区位条件下，房价低于其他区，包河也有类似情况 |
| 2 | 学区 | 中学的影响大于小学 |
| 3 | 工厂 | 工业区房价普遍较低 |
| 4 | 幼儿园 | 周边有幼儿园的小区价格较高 |
| 5 | 土地级别 | 土地级别决定了地价 |
| 6 | 公交站点 | 交通方便 |
| 7 | 建筑年代 | 老小区即使地段学区好也难卖出高价 |
| 8 | 开发商 | 品牌开发商质量有保障，自然房价高 |
| 9 | 医疗 | 周边有大型医院能带动房价 |
| 10 | 人口密度 | 人口稠密地区房价较高 |
| 11 | 商业设施 | 银行、商店、餐馆、酒店等权重相近 |
| 12 | 景区公园 | 高端小区休闲设施较为完善 |
| 13 | 地铁 | 2016年合肥地铁交通尚不发达，对房价影响有限 |

其余因素影响并不显著

### 模型效果

在合肥数据集上多次运行，得到的平均绝对误差为**8.51%**。

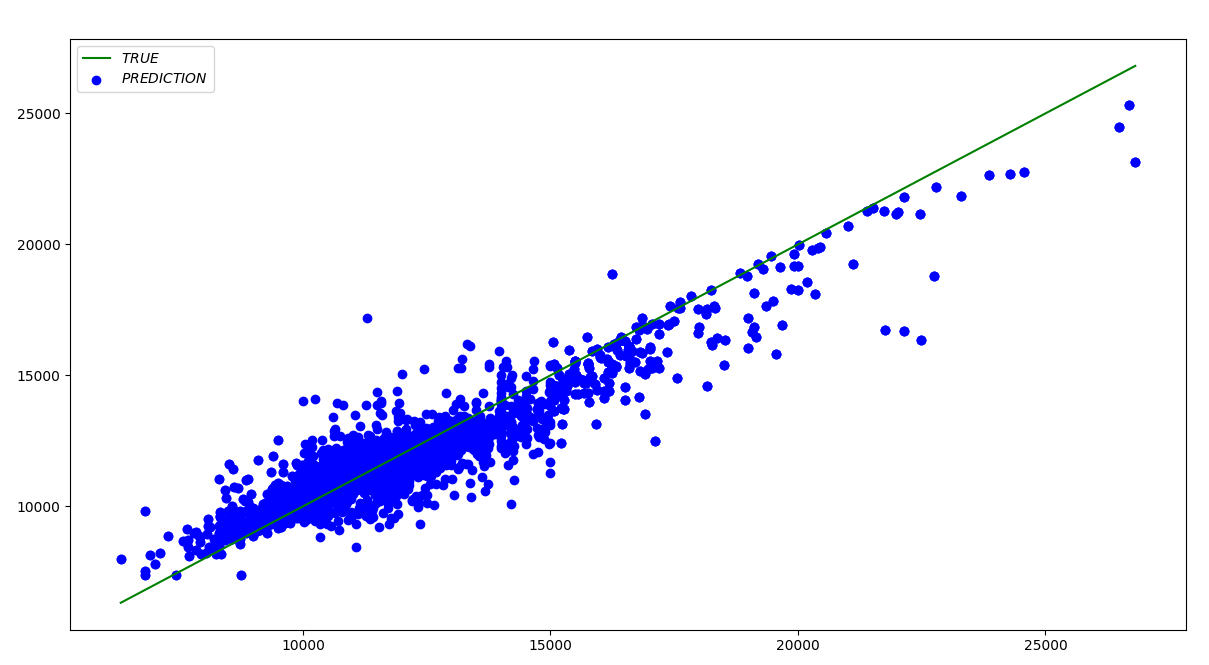
表格 4 模型参数

|  |  |
| --- | --- |
| 模型参数 | 参数值 |
| 决策树总数 | 500 |
| 叶节点最小样本数 | 2 |
| 选取特征数量 | （N为总特征数） |
| 划分依据 | 平均绝对误差 |
| 最小划分样本数 | 4 |
| 最大深度 | 20 |

表格 5 模型误差

|  |  |
| --- | --- |
| 绝对误差% | 占比% |
| ≤0.28 | 5 |
| ≤0.88 | 10 |
| ≤1.48 | 20 |
| ≤2.65 | 30 |
| ≤5.57 | 50 |
| ≤7.98 | 65 |
| ≤8.83 | 70 |
| ≤10.42 | 75 |
| ≤11.85 | 80 |
| ≤16.13 | 90 |
| ≤22.32 | 95 |
| ≤33.80 | 99 |

图表 2 模型误差的密度分布



图表 3 不同价位的房价平均值与误差值的关系