中国人民大學
RENMIN UNIVERSITY OF CHINA

期中Python建模展示

邵远平 2022202672
中国人民大学经济学院

数据处理与模型构建

1

数据导入与特征创立

数据预处理

1. 数值信息：价格、建筑面积
2. 文字信息：环线、户型、朝向、楼层、装修情况

补充信息处理

1. 小区：单价、容积率、停车、位置
2. 租金：面积、租金、租金面积比

进一步处理

1. 位置：与市中心距离-平方、倒数
2. 面积：面积平方、面积平方根
3. 户型：单位房间面积、卧厅比、卧卫比、户型质量得分
4. 朝向：特征、朝向面积交互
5. 楼层：楼层比例、电梯楼层交互
6. 面积交互：环线、距离、装修
7. 价格：与位置的比值、波动率
8. 租金：租金回报率、波动率

2

构建线性模型

进一步处理数据：

无穷大、缺失值、极端值

特征选择：

1. 选择最重要的前40个特征
2. SelectKBest类- f_regression评分函数
3. 提取被选中的元素

最终选出的特征：

城市、年份、环线数值、建筑面积、房间数、厅数、卫数、总房间数、朝北、装修情况数值、小区平均租金、小区租金波动、小区租金面积比、小区均价、小区价格波动、小区房源数、小区每平米均价、到市中心距离、距离平方、距离倒数、中心区域、建筑面积平方、建筑面积平方根、平均房间面积、卧厅比、户型质量分、南北通透、南向大户型、朝向面积交互、高层无电梯惩罚、面积环线比、环内大户型、面积装修交互、建筑年龄、装修折旧、面积距离交互、小区价值指数、价格距离比、租金回报率

3

模型拟合与结果导出

模型拟合：

1. 划分训练集和测试集
2. 使用OLS、岭回归、LASSO、弹性网、随机森林五种方法拟合模型
3. 进行6折交叉验证

输出结果：

1. 使用五个模型分别预测房源价格
2. 计算训练集、测试集和交叉验证的平均绝对误差(MAE)和均方根误差(RMSE)反映误差情况

预测结果与模型评估

Metrics	In Sample	Out of Sample	Cross-validation	Datahub Score
OLS	3.968×10^5	3.918×10^5	0.1815	69.716
Ridge	3.965×10^5	3.917×10^5	0.1814	69.724
LASSO	3.953×10^5	3.954×10^5	0.1817	70.455
Elastic Net	3.956×10^5	3.957×10^5	0.1817	70.259
Random Forest	1.271×10^5	1.704×10^5	0.0881	79.512
Best Model	1.271×10^5	1.704×10^5	0.0881	79.512

注：此处使用平均绝对误差（MAE）来衡量模型的表现

模型创新点说明

对房源特征的深度刻画

1.将文字指标数值化：给房间类型、装修、朝向等进行赋值加权

2.深度处理数据，创建装修折旧、卧厅比、黄金朝向等指标，充分刻画数据特征

3.创建多种交互项，如面积装修交互、高层无电梯惩罚、环内大户型等

数据清洗与特征选择协同优化

1.针对数据中缺失值、异常值问题，多层次数据清洗，确保模型强健的稳定性

2.筛选40个重要性最高的特征，提高模型效率，改善模型过拟合状况

使用多种方法拟合模型

采用线性回归、Lasso、Ridge、弹性网与随机森林五种模型进行拟合和预测，可以解决处理高维数据、防止过拟合等问题，并比较模型表现