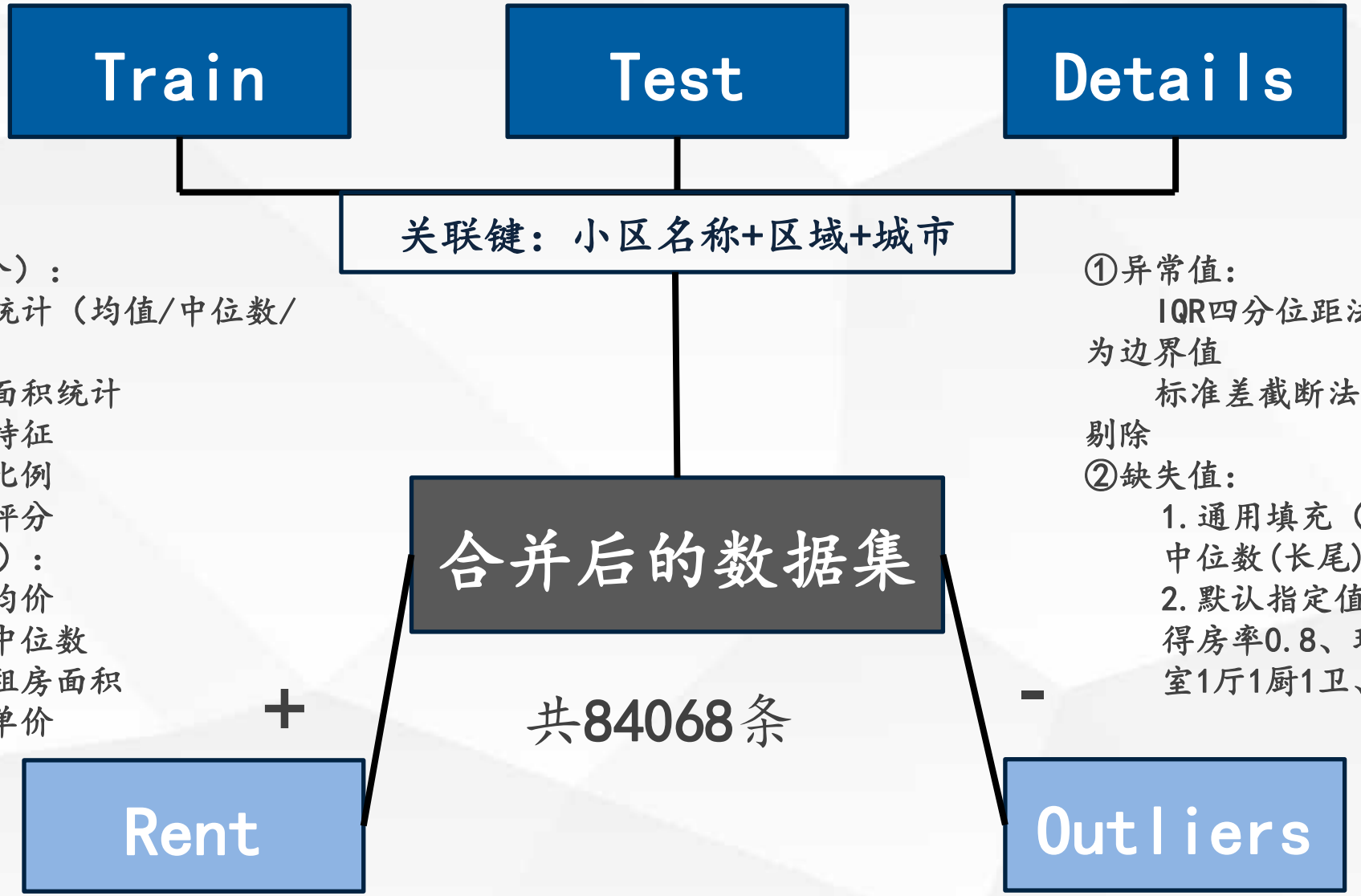
基于树模型和集成学习的房地产价格预测模型

——经济学院 程忆楠 2022202610

目录-模型结构

数据处理



两种粒度：
①小区级（23个）：
小区租金统计（均值/中位数/标准差等）
小区租房面积统计
小区户型特征
小区朝向比例
小区装修评分
②城市级（4个）：
城市租金均价
城市租金中位数
城市平均租房面积
城市租金单价

①异常值：
IQR四分位距法-超出1.5倍替换为边界值
标准差截断法-超出三倍标准差剔除
②缺失值：
1.通用填充（租金、面积等）：中位数(长尾)
2.默认指定值：
得房率0.8、环线中环、户型1室1厅1厨1卫、装修得分0.3等

特征变量选择

数值提取+正则化处理+标签评分

一、简单特征工程

1. **区域相关：**城市、区域、板块

2. **面积相关：**计算套内面积、对数建筑面积、平方根建筑面积、创建户型分段、标记异常值

3. **楼层、建筑年代、装修情况、朝向特征、建筑结构：**创建类型标签并赋值

4. **户型特征：**正则提取室厅厨卫；计算卧卫比；户型评分

5. **环线特征：**匹配阿拉伯数字；计算中心度

6. **小区品质特征：容积率、绿化率、物业费**

7. **文本特征：**统计标点符号数-视为特征数量

8. **数字特征：**房屋、楼栋、燃气费、供热费、停车位、停车费用

9. **交通出行：**正则化提取线/路数量

二、聚类特征 K-means

1. **地理位置聚类：**使用经纬度；颗粒度为15/30

2. **房屋特征聚类：**建筑面积、房龄、相对楼层、总楼层；颗粒度20

三、交互特征 乘积；比例

('建筑面积', '房龄'),
('相对楼层', '总楼层'),
('中心度', '装修分数'),
('朝向得分', '相对楼层'),
('核心区域', '建筑面积'),
('房龄', '装修分数'),
(租金', '建筑面积'),
(租金单价', '装修分数')

四、统计特征

1. **区域：**均价；标准差；中位数；样本数

2. **户型：**均价；标准价

特征重要性分析

树模型特征选择：计算每个分裂节点处特征对MSE的减少量并取均值

TOP 20

基本特征（基于LGBM模型）

租金均价×建筑面积 3508
建筑面积_对数 2846
建筑面积/房龄 2677
建筑面积×房龄 2621
租金均价/建筑面积 2540
相对楼层/总楼层 2205
总楼层 1991
相对楼层×总楼层 1924
房屋 1880
板块_单价均值 1869
板块_单价标准差 1779
租房房间数均值 1681
板块 1516
经度 1515
租房装修情况均值 1511
物业费 1448
租房数量 1417
容积率 1411
纬度 1410
租房面积均值 1369

TOP 10

租房特征

租金均价×建筑面积 3508
租金均价/建筑面积 2540
租房房间数均值 1681
租房装修情况均值 1511
租房数量 1417
租房面积均值 1369
租房南北朝向比例 1327
租房单价 1237
租金价格标准差 1226
租房南向比例 1223

预测结果

		In sample RMSE	Testing RMSE	CV RMSE	CV R^2	CV MAPE	Total N after dropping
84.11	xgboost	458946	456791	455144	0.9641	8.77%	84068
83.878	Extra_trees	457036	456294	455073	0.9641	8.81%	84068
84.332	Lightgbm	437849	437950	435161	0.9672	8.49%	84068
77.039	OLS	857904	859063	859539	0.8724	18.63%	84068
77.037	Ridge	864417	863169	861200	0.8719	18.65%	84068
84.36	Lgbm堆叠	414628	414759	412384	0.9742	8.05%	84068

THANKS