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

01 数据加载与合并（train/test/details/rent）

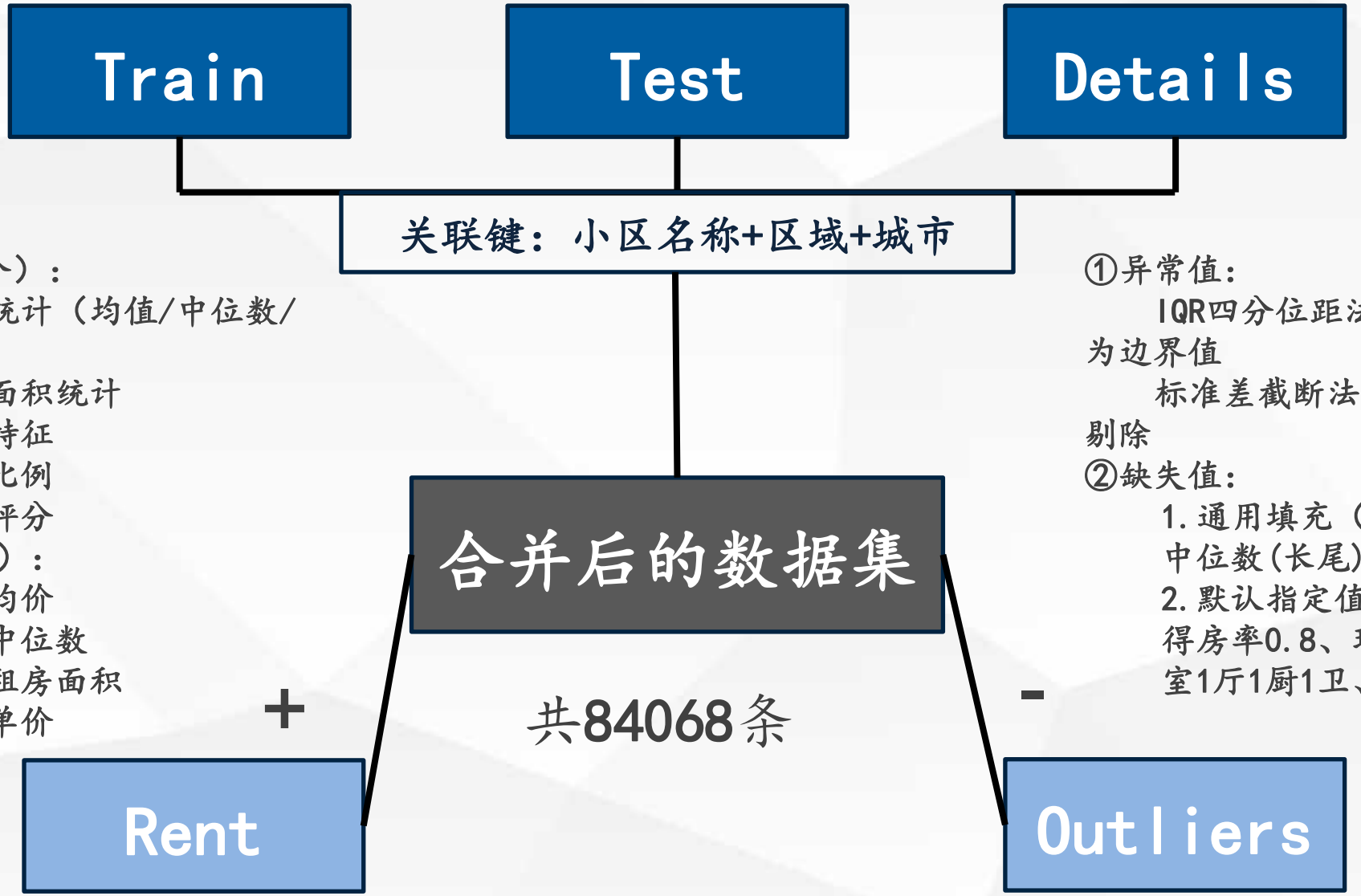
02 数据清洗与异常值剔除（IQR/Z-score）

03 数据特征提取+特征工程构建（基础/聚类/交互/统计）

04 特征重要性分析+特征选择（135）

05 模型训练（Lightgbm/extra_tree/Xgboost/OLS/Ridge）+ 堆叠集成

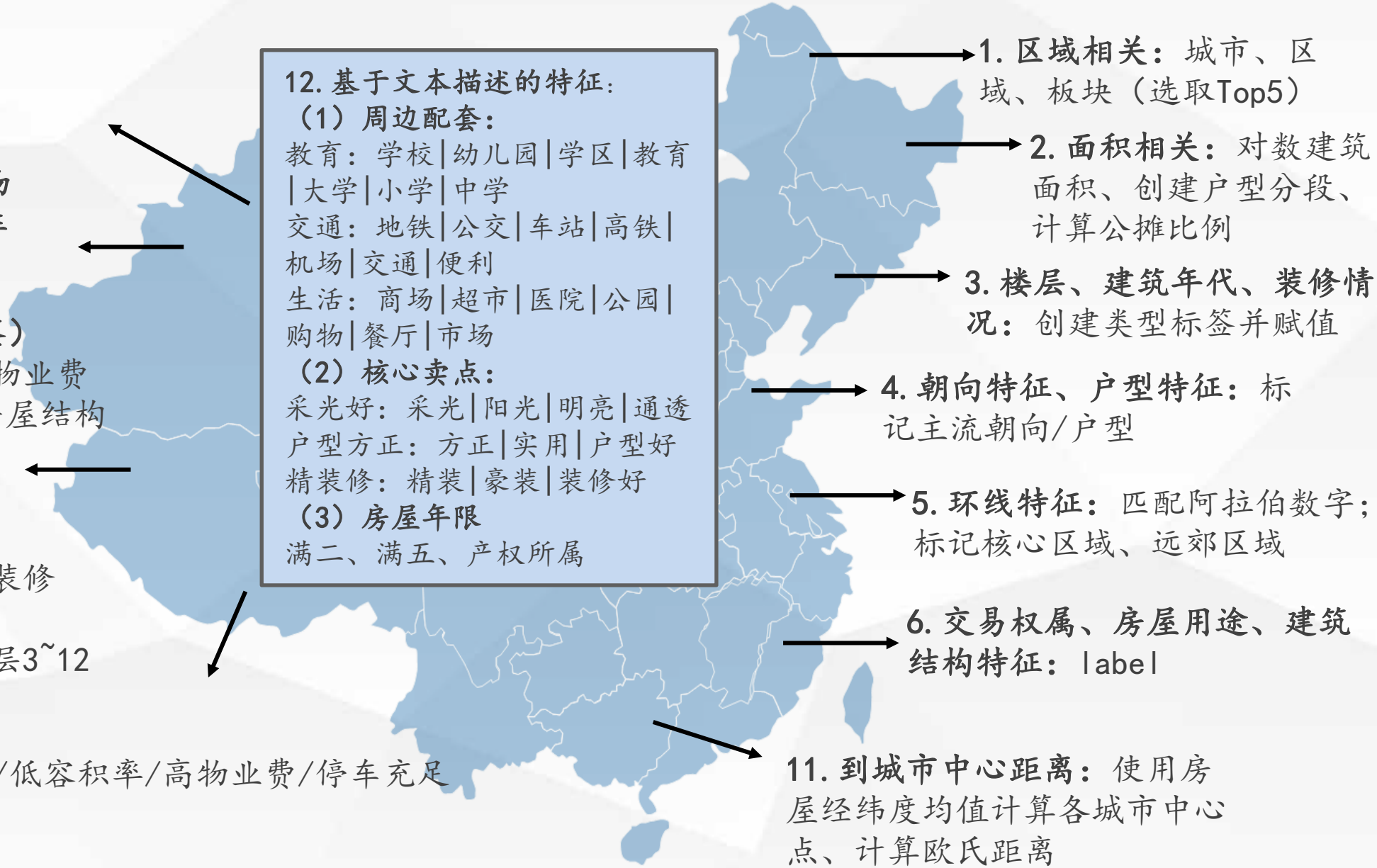
数据处理



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



1. 区域相关： 城市、区域、板块（选取Top5）

2. 面积相关： 对数建筑面积、创建户型分段、计算公摊比例

3. 楼层、建筑年代、装修情况： 创建类型标签并赋值

4. 朝向特征、户型特征： 标记主流朝向/户型

5. 环线特征： 匹配阿拉伯数字；标记核心区域、远郊区域

6. 交易权属、房屋用途、建筑结构特征： label

7. 电梯特征： *高层无电梯、计算梯户比

8. 容积率、绿化率、物业费、供暖方式、停车位、物业类别特征

9. 交互项：（基于标签）

- 户型*装修
- 容积率*物业费
- 楼层*电梯
- 房龄*房屋结构
- 房龄*装修
- 区位*面积
- 朝向*楼层
- 新房豪装=新房*高级装修
- 老房毛坯=老房*毛坯
- 电梯洋房=有电梯*楼层3~12

10. 特殊高端组合：

- **高端小区：** 高绿化率/低容积率/高物业费/停车充足
- **经济适用房**
- **学区房**

11. 到城市中心距离： 使用房屋经纬度均值计算各城市中心点、计算欧氏距离

12. 基于文本描述的特征：

（1）周边配套：

教育：学校|幼儿园|学区|教育|大学|小学|中学

交通：地铁|公交|车站|高铁|机场|交通|便利

生活：商场|超市|医院|公园|购物|餐厅|市场

（2）核心卖点：

采光好：采光|阳光|明亮|通透

户型方正：方正|实用|户型好

精装修：精装|豪装|装修好

（3）房屋年限

满二、满五、产权所属

特征重要性分析

Extra Tree Regressor

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
xgboost	458946	456791	455144	0.9641	8.77%	84068
Extra_trees	457036	456294	455073	0.9641	8.81%	84068
Lightgbm	437849	437950	435161	0.9672	8.49%	84068
OLS	857904	859063	859539	0.8724	18.63%	84068
Ridge	864417	863169	861200	0.8719	18.65%	84068
Lgbm集成	414628	414759	412384	0.9742	8.05%	84068

72.568

74.314

74.310

THANKS