基于机器学习算法的房地产价格预测模型

——经济学院 程忆楠 2022202610

模型结构与数据特征

01 数据加载与合并（train/test/details）

02 数据描述性统计->识别特征变量

03 特征工程（26类房产特征）

04 数据预处理 （数值变量/分类变量）

05 模型训练与参数调优 (OLS/LASSO/Ridge/ElasticNet)

06 最佳模型选择与最终预测

缺失值分析：

抵押信息、别墅类型、户型介绍、套内面积、环线缺失比例达 70%

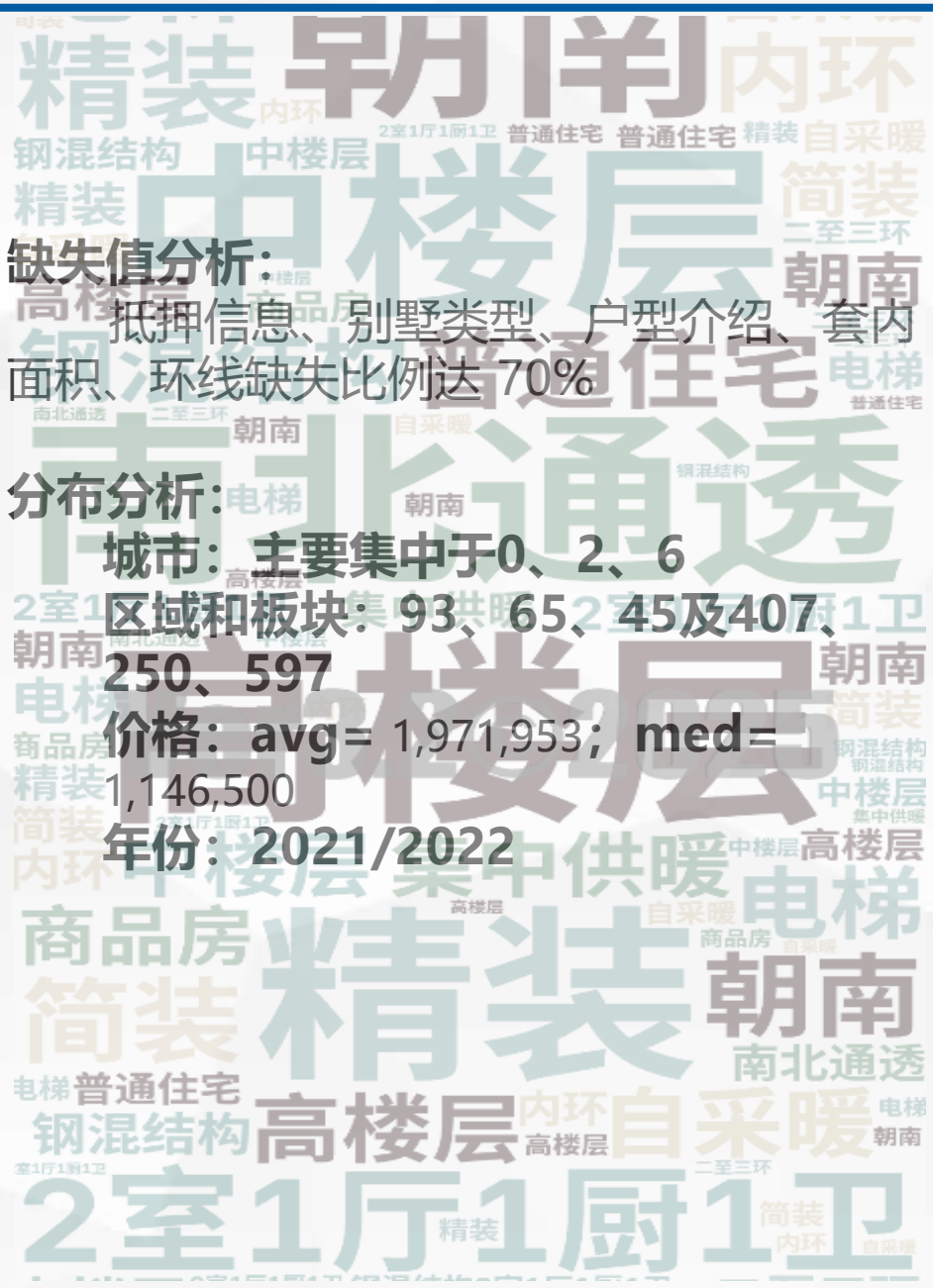
分布分析：

城市：主要集中于0、2、6

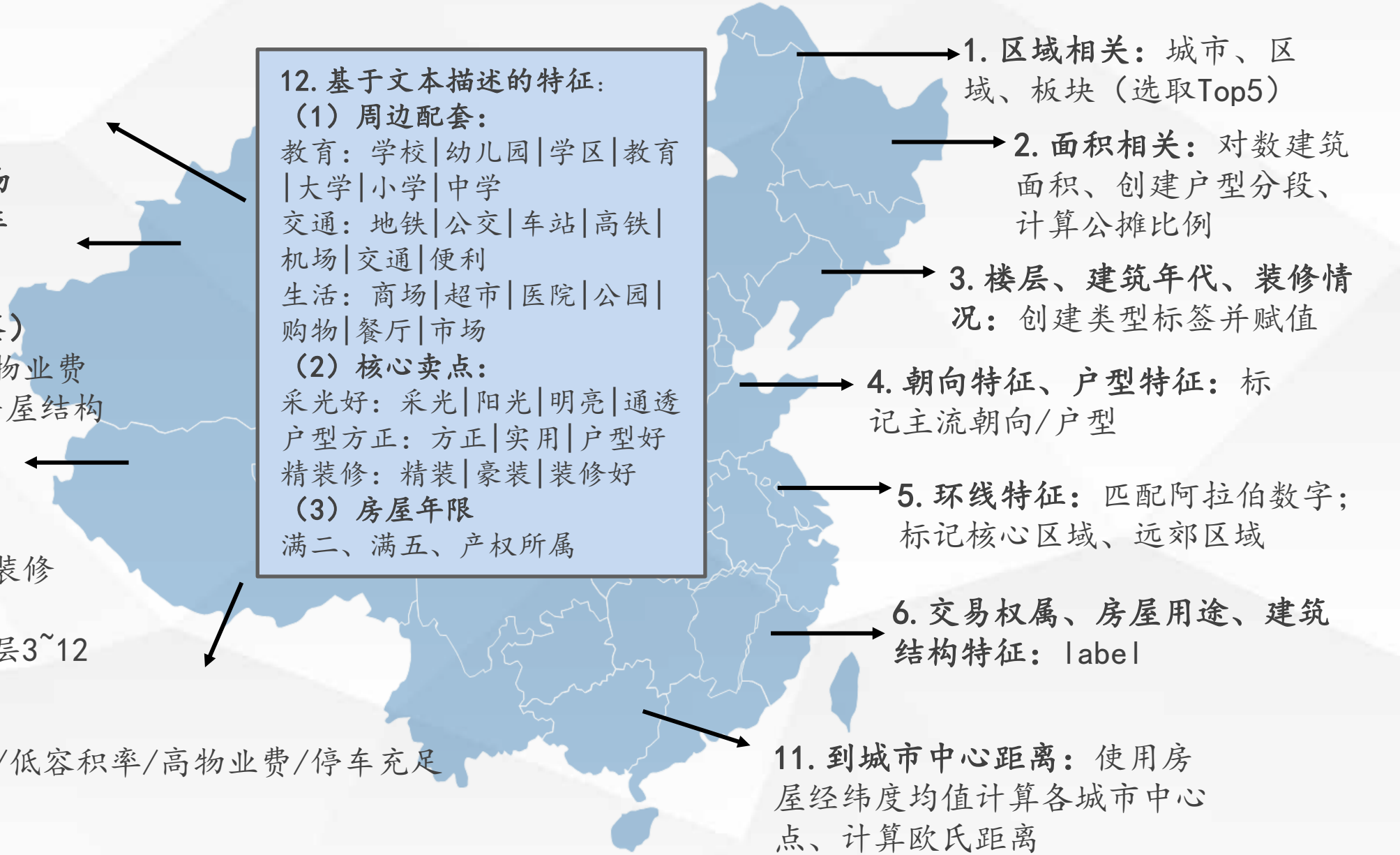
区域和板块：93、65、45及407、250、597

价格：avg= 1,971,953；**med=** 1,146,500

年份：2021/2022



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

预测结果

		训练MAE	训练RMSE	训练R^2	交叉验证 MAE	交叉验证 RMSE	交叉验证 R^2
74.314	OLS	305284.77	1007004.33	0.8545	310164.21	1021077.93	0.8495
72.568	Lasso	338482.66	1073296.87	0.8347	340422.81	1077763.69	0.8323
74.310	Ridge	305662.59	1007098.95	0.8545	310383.91	1021199.50	0.8494
71.783	ElasticNet	357025.51	1122555.17	0.8192	358516.84	1125494.45	0.8173

THANKS