中国人民大学
RENMIN UNIVERSITY OF CHINA

房价预测建模项目

汇报人：谢丽媛

2025年6月

模型亮点-数据预处理

“环线”预处理方案

- 统一表达：

将不同表达但含义相同的值统一（如"内环内"→"一环内"）

- 有序编码：

数值编码保留环线的距离关系（数值越大离市中心越远）

- 衍生特征：

创建是否核心区二元特征（编码值≤3的视为核心区）

“梯户比例"预处理方案

- 结构化提取：

从"X梯Y户"格式中提取电梯数量和每梯户数

示例："八梯六户" → 电梯数量=8，每梯户数=6

- 核心指标计算：

电梯服务密度 = 总户数 / 电梯数量

- 衍生特征创建：

拥挤度分级：非常宽松(≤4户/梯)、舒适、一般、拥挤、非常拥挤(>20户/梯)

- 异常值处理：

电梯数量超过20的设为20

每梯户数超过100的设为100

模型亮点-数据预处理

"建筑面积"和"套内面积"预处理方案

- 数据清洗：

移除"m²"符号并转换为float类型

处理极端小面积（<15m²设为15m²）

- 衍生特征：

得房率：套内面积/建筑面积（核心指标）

公摊面积：建筑面积-套内面积

面积等级：按市场标准分段（50/90/144/200m²为界）

是否豪宅二值变量：144m²为分界线（中国普通/非普通住宅标准）

- 缺失值处理：

使用中位数得房率估算缺失的套内面积

- 异常值处理：

确保套内面积≤建筑面积

模型亮点-数据预处理

训练集"房屋户型"预处理方案

- 数据清洗：

统一"房间"→"室"的表述

修复可能的文本错位（如"卫室"→"卫 室"）

- 异常值处理：

卧室数超过10的设为10（应对"16室"等异常）

卫生间数不超过卧室数+3（应对"14卫"等异常）

- 衍生特征：

总房间数：卧室+客厅

卧室卫生间比：反映居住舒适度

是否标准户型：标记功能齐全的户型

是否豪宅户型：卧室≥4的户型

训练集“房屋朝向"预处理方案

- 特征提取：

主朝向：按价值优先级选择（南>东南>东>...）

优质标志：南、东南、东朝向为优质

- 按现实逻辑编码：

中国购房者偏好：南 > 东南 > 东 > 西南 > 北 > 西 > 东北 > 西北

南北通透户型有显著溢价

模型评估

（一）线性模型

模型	RMSE	R^2
LinearRegression	6.44×10^{7} (异常)	-5.22×10^{15} (异常)
Ridge	**0.5186**	**0.6611**
Lasso	0.7881	0.2176

- 普通线性回归失效

R^2为负，比均值预测更差

- Ridge表现最优

说明共线性严重

→ 必须使用正则化

模型评估

（二）模型训练与选择

◘ 代码架构设计

```
pipeline = Pipeline([
    ('preprocessor', preprocessor),  # 特征工程管道
    ('model', model)                 # 预测模型
])
```

- **统一接口**：将特征处理与模型训练封装为端到端流程
- **可复用性**：支持快速切换不同模型

◘ 自动化调参流程

```
GridSearchCV(
    pipeline,                # 包含预处理+模型的完整流程
    param_grids[name],       # 对应模型的参数网格
    cv=5,                    # 5折交叉验证
    scoring='neg_rmse',      # 评估指标：负RMSE（越小越好）
    n_jobs=-1,               # 使用所有CPU核心并行计算
    verbose=1                # 输出调参过程日志
)
```

模型评估

（三）线性模型与树模型

模型名称	RMSE	R^2	最佳参数	训练耗时（拟合次数）
RandomForest	0.1917	0.9532	`{'model__max_depth': None}`	15 fits (3×5 folds)
LightGBM	0.2209	0.9379	`{'model__learning_rate': 0.1, 'model__num_leaves': 63}`	15 fits (3×5 folds)
Ridge	0.5146	0.6629	`{'model__alpha': 1}`	15 fits (3×5 folds)
Lasso	0.5452	0.6217	`{'model__alpha': 0.001}`	45 fits (9×5 folds)
ElasticNet	0.5321	0.6398	`{'model__alpha': 0.001, 'model__l1_ratio': 0.5}`	45 fits (9×5 folds)
XGBoost	0.2083	0.9426	`{'model__learning_rate': 0.1, 'model__max_depth': 6}`	30 fits (6×5 folds)

模型评估

（四）最优模型：随机森林

```
RandomForest - RMSE: 0.1917, R2: 0.9532
Best params: {'model__max_depth': None}
Fitting 5 folds for each of 6 candidates, totalling 30 fits
```

□ 模型性能评估：

预测较准确，拟合效果较好。

□ 最佳超参数：

Best params: {'model__max_depth': None}
•max_depth=None ： 决策树不限制最大深度。

□ 交叉验证：

•**5 folds**：使用了 5 折交叉验证（将数据分为 5 份，轮流用 4 份训练，1 份验证）。
•**3 candidates**：对 max_depth 参数尝试了 3 种可能的取值（[10, 20, None]）。
•**15 fits**：总共训练了 15 次模型（5 折 × 3 种参数组合）。

模型表现对比

（五）模型优化：随机森林

```
param_grids = {
    'RandomForest': {
        'n_estimators': [100, 150],          # ↑ 增加树的数量
        'max_depth': [15, 20, None],         # 动态深度控制
        'min_samples_split': [2, 5, 10],     # 新增分裂控制
        'min_samples_leaf': [1, 2, 4],       # 新增叶节点约束
        'max_features': ['sqrt', 0.8],       # 新增特征采样策略
        'bootstrap': [True, False]           # 新增样本采样开关
    }
}
# 采用RandomizedSearchCV加速搜索：
RandomizedSearchCV(n_iter=15, cv=3, n_jobs=4)
```

指标	初始版本	优化版本
RMSE	0.1917	0.1909
R^2	0.9532	0.9536
训练效率	15次拟合	45次拟合
最佳参数组合	仅max_depth	6参数联合优化

1. 增加树数量(n_estimators=150)未导致过拟合
2. 限制min_samples_split=5提升泛化能力
3. 深度控制 max_depth平衡模型复杂度与过拟合风险