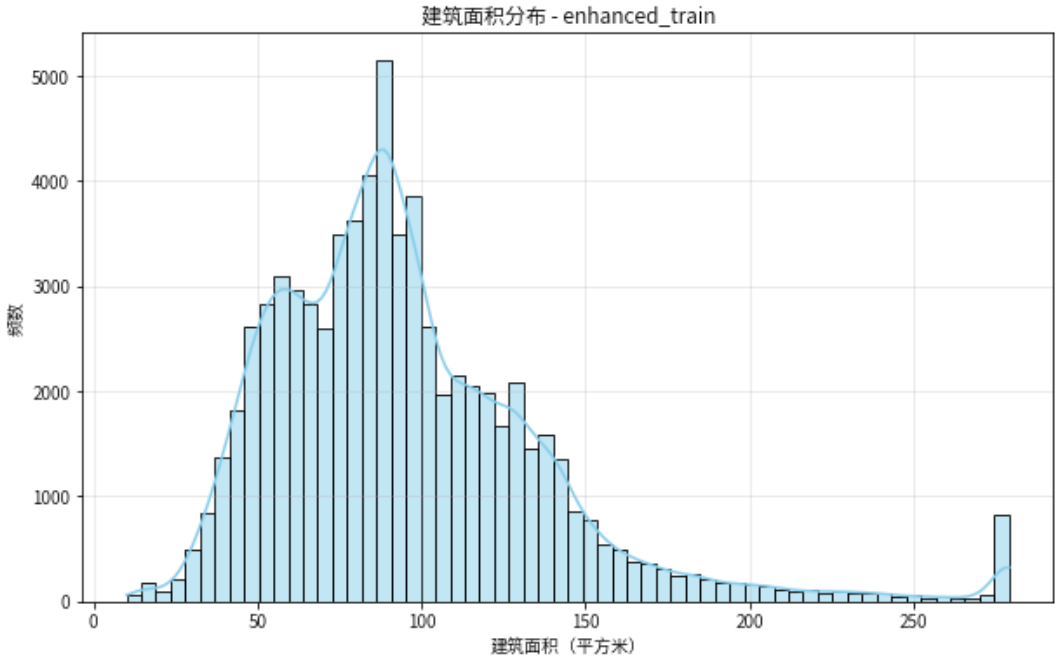
房地产价格预测期末汇报

XJY

2025.6

特征工程

- 新数据信息：房间数、房租、小区特征…
- 新变量构建：租金_面积、租金_卧室数、面积段…

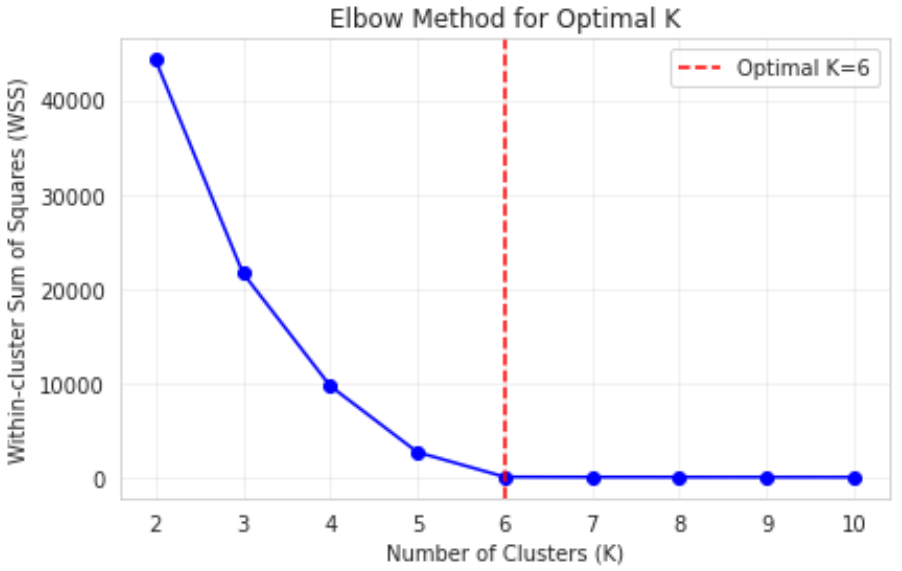


不同户型的单价可能存在差异，分段处理可以补充连续变量缺乏的信息，模型可以为每个区段学习一个独立的效应。

参考训练集数据分布，按70平米、110平米为界，划分了小户型、中户型和大户型3个面积段。

特征工程

- **地理位置特征工程：使用经纬度**
  - 采用KNN算法进行邻域分析；
  - 结合WSS肘部法则和轮廓系数确定最佳聚类数
  - 构建空间密度特征；地理中心距离



坐标标准化 → 构建不同K值的KNN模型 → 查询每个房屋的K个最近邻 → 计算邻近房屋单价的均值/标准差特征

坐标标准化 → 寻找最优聚类K → 训练Kmeans模型 → 分配聚类标签 → 计算聚类统计特征

特征工程

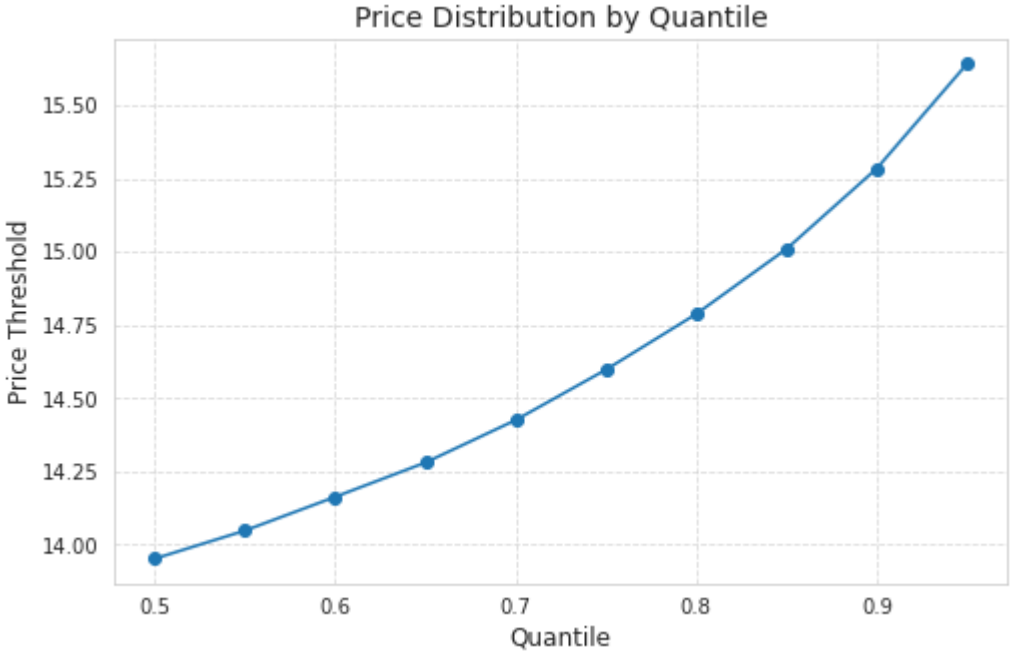
- **文本特征处理：**
  - 采用了jieba分词 + 正则表达式的双重策略
  - 基于高频词构建“周边配套”、“交通出行”、“房屋优势”的关键词字典
  - 核心卖点：采用TF-IDF将文本向量化

  - 引入了描述丰富度指标
    - 丰富度 = 唯一词汇数 / 总词汇数——越高表示内容越丰富
    - 将文本合并后分词，统计去重后的词汇数量得到唯一词汇

模型训练

- **线性模型**
- OLS、Ridge、Lasso：特征工程优化后，表现有所提升
- 尝试：
  - PCA降维：保留95%方差，效果明显变差
  - 更换损失函数：Huber Loss，效果一般
- **神经网络：** ANN
  - 3个隐藏层+1个输出层：
  - 输入维度设置基于特征数量，针对输入维度与dropout rate进行optuna调参
  - 损失函数结合了MSE与MAE
  - 模型表现：略差于树模型

模型训练



- **树模型：**
- Random Forest, XgBoost
  - 调参：Randomized Search, Optuna
- 尝试：
  - 自定义QuantitleGBDT（分位数加权损失下的梯度提升）
    - 对高价房样本给予更多关注，表现略优于基准模型（未加权的GBDT）

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot (y_i - \hat{y}_i)^2$$

$$w_i = \begin{cases} \text{high\_weight} & \text{if } y_i > \text{threshold} \\ 1.0 & \text{otherwise} \end{cases}$$

  - 选取重要特征后再训练：采用XGBoost的gain重要性指标选取前100个

模型集成

- 初步尝试：
  - 将选取重要特征后得到的Xgboost结果与使用所有变量得到的Xgboost结果加权（0.4/0.6），预测分数优于单独结果。
- Stacking：
  - 基模型为随机森林、Xgboost、Ridge、QuantitleGBDT
  - 元模型为线性模型（OLS、Ridge、Lasso）
    - 经尝试Lasso（α=0.1）效果最佳
  - 后基模型加入ANN：
    - 元模型为Lasso（α=1）效果相对“好”
    - 元模型为GBDT，效果好于用线性模型

基模型相关性较高，因此Stacking效果不明显，最终预测相对于Xgboost有微弱提升

预测值相关性矩阵：

	xgb	rf	quantitlegbdt	ridge	ols
xgb	1.000000	0.998012	0.998530	0.975494	0.975496
rf	0.998012	1.000000	0.997906	0.975674	0.975676
quantitlegbdt	0.998530	0.997906	1.000000	0.977070	0.977072
ridge	0.975494	0.975674	0.977070	1.000000	1.000000
ols	0.975496	0.975676	0.977072	1.000000	1.000000

误差相关性矩阵：

	xgb	rf	quantitlegbdt	ridge	ols
xgb	1.000000	0.892227	0.920195	0.504029	0.504032
rf	0.892227	1.000000	0.891214	0.514368	0.514365
quantitlegbdt	0.920195	0.891214	1.000000	0.551990	0.551988
ridge	0.504029	0.514368	0.551990	1.000000	0.999997
ols	0.504032	0.514365	0.551988	0.999997	1.000000

模型	交叉验证RMSE	Test MAE	Test RMSE	Datahub_score
Xgboost	521545	158838	575073	84.034
Random Forest	556659	163237	578929	83.462
QuantitleGBDT	538391	168363	580216	83.288
ANN	-	178676	603469	-
OLS	-	329055	931907	-

Xgboost与选择重要特征后的Xgboost加权模型，datahub分数为84.1
Stacking模型最好分数为84.08