房产预测任务期末展示

妙含笑 2022200230

在期中任务基础上的特征工程改进

异常值处理改进：对目标变量进行IQR、Z-score、孤立森林检测，从67306个样本中识别出2927个异常值

特征工程总体思路

原始特征 (30个)

- 数值特征处理 → 清洗规范化特征
- 文本特征处理 → TF-IDF/词频/主题特征
- 地理特征处理 → 聚类/距离特征
- 时间特征处理 → 周期性/间隔特征
- 房屋特征处理 → 效率/密度/朝向特征
- 交互特征生成 → 组合/乘积特征
- 统计特征构建 → 群体/聚合特征
- 高级特征扩展 → 多项式/降维/聚类特征

↓

工程化特征 (1028个)

特征工程具体改进

- **多模态语义挖掘：** TF-IDF语义向量化、词袋模型的增强、LDA主题建模，构建语义特征
- **地理特征处理：** 根据经纬度使用K-means聚类进行空间分区；使用balltree分区以构建空间密度特征
- **多项式特征构建：** 选择最重要的特征生成交互项
- **降维特征：** PCA & ICA
- **聚类特征：** 基于无监督学习，对特征进行多尺度K-means聚类
- **目标编码：** 基于目标变量构建编码特征，贝叶斯平滑避免过拟合

特征选择与特征重要性

- 文本特征的维度控制：去除过于稀少的词汇，控制特征数量避免过拟合
- 方差阈值选择：去除几乎不变的特征
- 互信息特征选择：捕获非线性关系，自适应调节阈值
- 基于模型的特征选择 ： 随机森林作为选择器

特征重要性表格

排名	特征	重要性
1	建筑面积	166.002786
2	lat	134.002247
3	lon	101.002069
4	到聚类中心距离	89.002732
5	总层数	43.500470
6	室	28.501761
7	卫	22.002151
8	区域_62.0	21.004612
9	城市_0.0	19.244831
10	城市_3.0	19.067405

模型构建与训练

- 基础模型池：
  - 线性模型：Linear Regression, Ridge, Lasso, ElasticNet
  - 树模型：Random Forest, Extra Trees, Gradient Boosting
  - 提升模型：XGBoost, LightGBM
  - 其他：SVR, MLP神经网络
- 并行训练
- 优化算法：
  - RandomizedSearchCV：快速参数搜索
  - 贝叶斯优化：高效参数空间探索
  - 网格搜索：精细化调参
- 单一模型性能比较

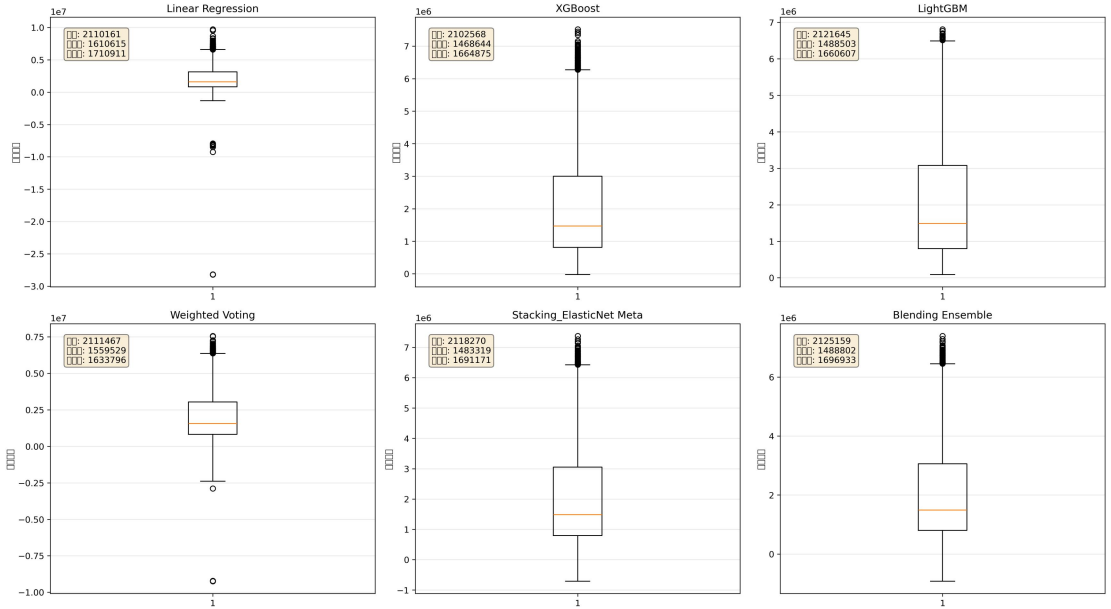
模型名称	训练集RMSE	验证集RMSE	交叉验证RMSE	训练集R2	验证集R2
LightGBM	278374.86	1670138.29	300417.98	0.9542	0.5964
XGBoost	242896.5	1672442.65	282820.93	0.9651	0.5953
Linear Regression	416418.23	1687842.74	492072.83	0.8975	0.5878

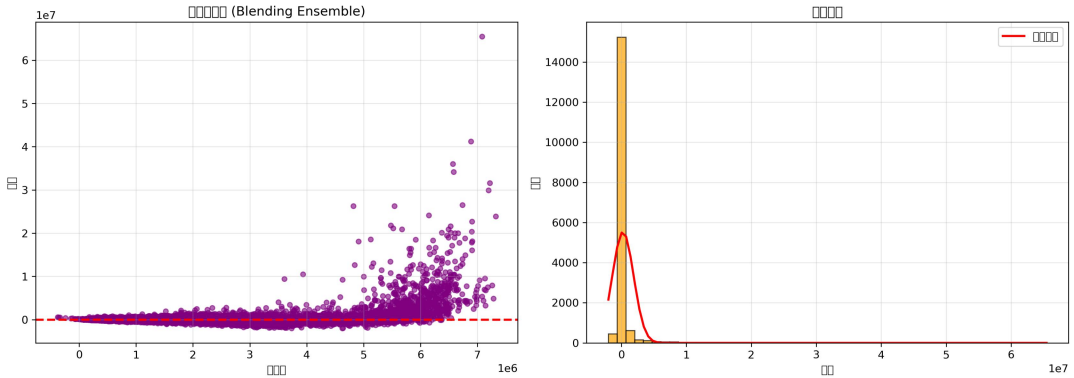
模型集成

- 三层集成框架：
  - Level 1: 基础模型训练（11个独立模型）
  - Level 2: 集成策略融合（3种集成方法）
  - Level 3: 最优模型选择（性能评估排序）
- 集成方法创新：
  - 加权投票集成：权重基于单一模型RMSE的倒数
  - 堆叠集成：Level-1: 基础预测器(5个最优模型)；Level-2: 元学习器(Ridge/RandomForest)
  - 混合集成：训练集分割：80%训练 + 20%混合；混合器：Ridge回归

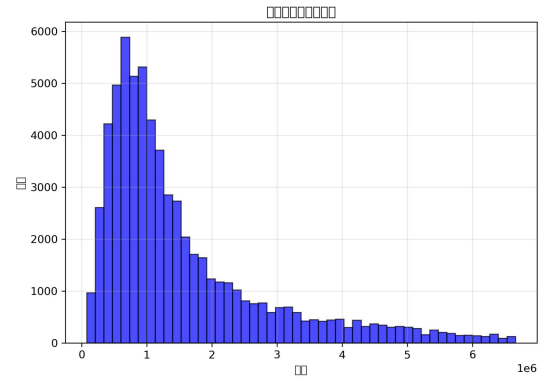
模型效果分析

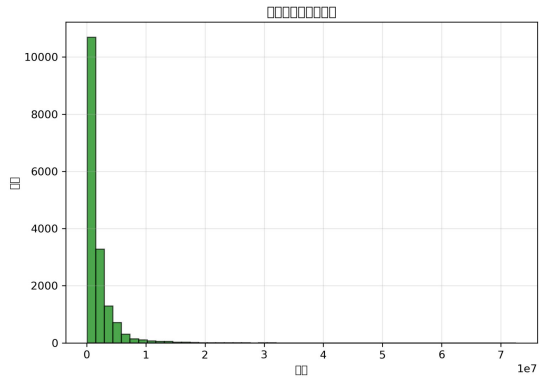
- 集成结果对比
  - 单模型最优：LightGBM，验证集RMSE = 1670138.29
  - 加权投票：验证集RMSE = 1662075.14
  - 堆叠集成：RMSE = 1656112.67
  - 混合集成：RMSE = 1652607.54
  - 综合最优模型：混合集成模型，hakson得分：64.528

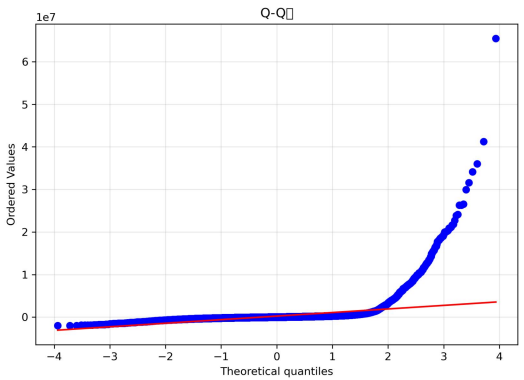

Linear Regression
평균: 2110161
중앙값: 1610615
표준편차: 1710911
XGBoost
평균: 2102568
중앙값: 1468644
표준편차: 1664875
LightGBM
평균: 2121645
중앙값: 1488503
표준편차: 1660607
Weighted Voting
평균: 2111467
중앙값: 1559529
표준편차: 1633796
Stacking_ElasticNet Meta
평균: 2118270
중앙값: 1483319
표준편차: 1691171
Blending Ensemble
평균: 2125159
중앙값: 1488802
표준편차: 1696933

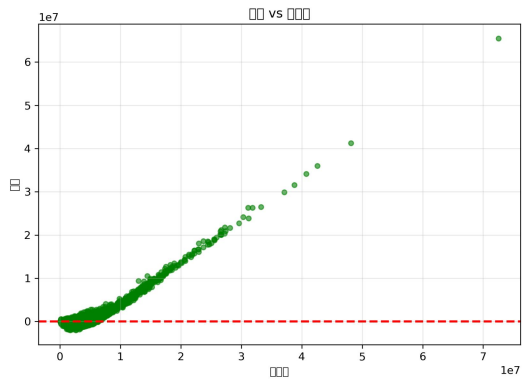

□□□□□ (Blending Ensemble)
□□□
□□□□

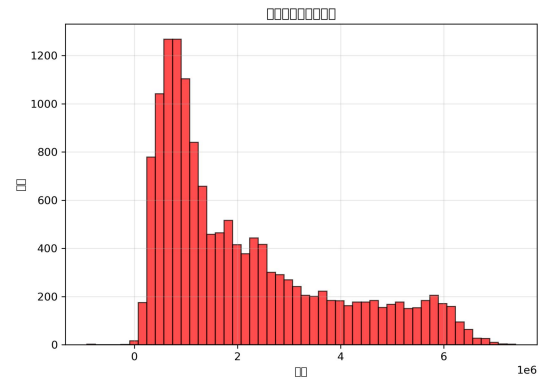
□□□□
□□□□
□□
□□

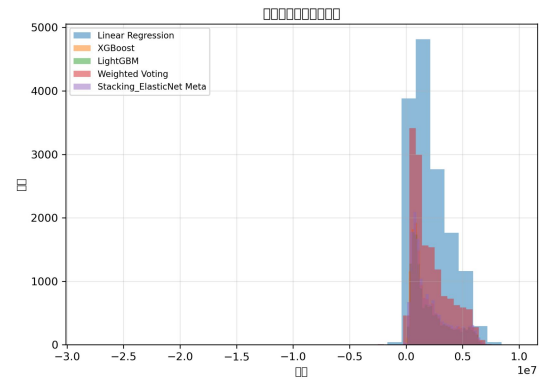

□□□□□□□□□□
□□
□□


□□□□□□□□□□
□□
□□


Q-Q□
Ordered Values
Theoretical quantiles


□□ vs □□□
□□□
□□


□□□□□□□□□□
□□
□□


□□□□□□□□□□□□
Linear Regression
XGBoost
LightGBM
Weighted Voting
Stacking_ElasticNet Meta
□□
□□