

被审查小组: Group 2
审查小组: Group 1
审查日期: 2025 年 4 月 20 日

HW4 分工:

邓双贤: 提供预测代码基础版本+改进预测代码
曹越晰: 改进预测代码并形成最终版本+审查第二组代码并提出意见
刘倡源: 改进预测代码+审查第二组代码并提出意见
马瑜梓: 检查预测代码+审查第二组代码, 汇总形成报告

一、代码运行情况

1. 成功运行: 是
2. 报错或警告: 出现 Sklearn 的 ConvergenceWarning, 与 Lasso/ElasticNet 收敛有关。
3. 代码结构: 结构清晰, 流程完整, 注释基本齐备

综合评价:

代码结构良好, 逻辑清晰, 涵盖了数据加载、预处理、模型构建、评估到保存的全过程。运行成功, 但出现轻微的收敛警告, 建议进行模型调参或特征缩放检查, 以提高稳定性。并且未提供 requirements.txt, 建议附上运行所需的库和版本信息。此外, 项目原始数据是 csv 格式, 代码中将其转换为 parquet 格式, 但未说明如何进行该格式转换。建议保留转换过程的相关说明, 以确保实验的可复现性和严谨性。

二、数据处理逻辑

1. 缺失值处理: 对房屋户型、楼层、面积、朝向等特征进行了合理的缺失值填充
2. 正确划分了训练集和验证集: 使用 `train_test_split`, 比例合理
3. 数据泄露风险: 无明显泄露 (合理划分 + 使用 Pipeline)
4. 数据特征解释: 特征处理完整, 包括对房屋户型、楼层、面积、梯户比例等的解析与填充

综合评价:

数据预处理完整, 缺失值处理、特征转换等步骤合理。针对部分缺失值采用了动态填充策略 (如根据室内面积、楼层等特征填充), 并将所有特征转换为适合建模的格式。特征处理方法合理, 模型训练数据集和测试集的划分也较为清晰。但是, 关于环线的 KNN 填充部分, 代码中是在 3-14 循环范围内选择了最优值 3, 1 和 2 并未检验。建议扩大范围, 尝试 1 到 5 的 K 值, 以增加选择的合理性。

三、模型构建与选择

1. 模型类型: 线性回归 (OLS)、Lasso、Ridge、ElasticNet
2. 适用于房价预测任务: 合理
3. 模型丰富性: 使用了多个模型, 包括 OLS、Lasso、Ridge 和 ElasticNet
4. 过拟合: 考虑到过拟合, 使用了交叉验证选择最优模型和参数, 并计算了训练集、测试集的准确性

综合评价:

通过对四种不同的回归模型 (OLS、Lasso、Ridge、ElasticNet) 的比较, 成功利用交叉验证评估模型性能, 并计算了每个模型在训练集、测试集上的表现。每个模型的训练和测试集均表现较好。在回归过程中加入了所有的交乘项, 但部分交乘项的相关性较低, 建议进一步筛选这些项, 以提高模型精准度并减少计算时间。

四、验证方法与评估指标

1. 验证集划分方法：使用了`train_test_split`进行简单的 hold-out 划分
2. 评估指标：RMSE、 R^2 ，评估效果良好
3. 可视化对比：缺少可视化展示
4. 误差来源：无误差来源分析

综合评价：

通过`train_test_split`进行数据划分后，对每个模型进行了评估，使用 RMSE 和 R^2 作为指标，表现较为理想。然而，报告中缺少预测结果与实际值的可视化展示，无法直观反映模型的表现。建议增加误差分析，进一步评估模型的准确性，并分析具体误差来源，尤其是异常值对模型结果的影响。

五. 综合评价与建议

优点：

1. 数据预处理完整，缺失值处理、特征工程等做得细致
2. 多种回归模型的对比（OLS、Lasso、Ridge、ElasticNet）提供了多角度的评估，适合房价预测任务
3. 使用交叉验证选择最优 K 值，防止了过拟合问题

建议：

1. 尝试不同超参数的调优，特别是在 ElasticNet 和 Lasso 模型中，可能需要更精细的调整以提高收敛性。Lasso 和 ElasticNet 模型中出现了未收敛的警告，可能需要增加最大迭代次数或调整容差来使其收敛。
2. 增加预测与实际值的可视化对比图，并补充误差来源分析
3. 对于环线的 KNN 填充，建议检验 K 值的不同选择，尝试范围为 1 到 5，以增加 K 值选择的合理性。
4. 在进行 KNN 填充时未区分训练集和测试集，可能会污染测试集数据，建议优化为先分割数据再填充环线信息。
5. 样本分割-环线：主要涉及 CV 部分。该代码是先填充环线信息再 6-fold CV，所以实际上训练集的环线中已经利用了测试集的部分信息。建议优化代码，改为先分割再填充环线信息。
6. 在使用了小区数据 details 后，小区代表的元素已被使用，可以删除小区名称这一分类变量