《人工智能与 Python 程序设计》

期末建模报告

郭立为

中国人民大学经济学院

2025 年 6 月 15 日

中國人民大學
RENMIN UNIVERSITY OF CHINA

建模“两步法”

- 房屋价格的决定因素有很多，本次提供的数据文件中也包含异常丰富的特征信息
- 然而，这些信息并不是同等重要的（OLS 算法的局限性）
- 课程中学习了大量的建模方法，它们可以协助我判断哪些信息对于房屋价格有重要的影响
- 然而，机器是没有感情的，人类的经验和直觉往往是有效的

一个思考

地理位置在房屋价格的决定中有举足轻重的地位，其他任何因素都只是在此基础上对价格进行一定的增减。

建模" 两步法"

建模的基本思路

- 分两步拟合模型
- 第一步:计算房屋所在版块房价的平均值（实际是一个线性模型）
- 第二步:利用其他特征，对房屋在平均值上下浮动的百分比进行估计（更进一步的模型）

信息提取

利用 TF-IDF 处理文本信息

- 部分特征是描述性的而非定量的，相应特征也将会为购买者所直接阅读，因此适合使用文本处理方法
- 具体而言，采用 TF-IDF 工具，我们得以发现最重要的文本信息，从而将冗长的文本划分为特征变量

部分变量的特殊处理

- 户型：利用 KNN 算法填充
- 交易权属等：利用常识判断

第一步模型

- 预测变量：地理位置信息（城市、区域、板块）
- 模型类别：OLS、Lasso、Ridge、ElasticNet
- 利用 Optuna 进行参数优化

模型表现

由于本阶段去除了 0.2% 的极端值，因此测试集样本数为 16792.

模型	预测得分	样本内 MSE	样本外 MSE	交叉验证 MSE	备注
OLS	-	-	-	-	-
Lasso	64	68883147	70171902	562909480	$\alpha = 527$
Ridge	76	32813515	32530427	426538534	$\alpha = 7870$
ElasticNet	24	402502266	402703995	510278899	-

表 1: 第一步模型的预测性能比较

第二步模型

- 预测变量：除地理位置信息外的其他信息
- 模型类别：决策树模型、随机森林模型、XGBoost 模型、神经网络模型

模型表现

本阶段测试集样本数为 16826.

模型	预测得分	样本内 MSE	样本外 MSE	交叉验证 MSE	备注
决策树	76.798	0.0332	0.0444	0.0563	
随机森林	78.123	0.0302	0.0385	0.0485	
XGBoost	79.19	0.0003	0.0313	0.0498	最优
神经网络	75.713	0.0137	0.0416	0.0639	过拟合

表 2: 第二步模型的预测性能比较

未来的优化方向

因子	重要性	因子	重要性	因子	重要性
车库	0.096	一梯四十九户	0.058	别墅	0.055
两梯二十七户	0.046	一梯三十户	0.042	地下室	0.026
七厅	0.017	一梯三十六户	0.015	九卫	0.015
八梯三十八户	0.015	一梯十六户	0.015	一梯十七户	0.014
一梯十一户	0.013	六梯三十五户	0.011	十一室	0.011

表 3: 最优模型的因子重要性（前 15 个）

- 特征工程：许多信息还没有用到
- 模型选择和优化：可以尝试更多的复杂模型