# 《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 12 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

| 模型 | 预测得分 | 备注 |
|---|---|---|
| OLS | - | - |
| Lasso | 64 | $\alpha = 527$ |
| Ridge | 76 | $\alpha = 7870$ |
| ElasticNet | 24 | - |

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

| 模型 | 预测得分 | 样本内 MSE | 备注 |
|---|---|---|---|
| 决策树 | 76.798 | 0.0332 | |
| 随机森林 | 78.123 | 0.0260 | |
| XGBoost | 79.19 | 0.0004 | 最优 |
| 神经网络 | 75.713 | 0.0131 | 过拟合 |

表 2: 第二步模型的预测性能比较

# 未来的优化方向

- 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