房价预测的机器学习方法

刘倡源

中国人民大学数学学院

2025 年 4 月 3 日

中國人民大學
RENMIN UNIVERSITY OF CHINA

整体处理思路

- 四种模型：OLS、Lasso、Ridge、ElasticNet
- 将 details 数据合并入训练集和测试集
- 对房价取 log
- 加入数值变量的对数项和二次项（平方项和交乘项）
- 分类变量缺失值单独分类处理，数值变量缺失值使用平均数填充
- 在 5% 和 95% 分位数上进行缩尾处理
- 在 Lasso、Ridge、ElasticNet 模型中，尝试使用不同的超参数，选择最优模型进行回归（使用 CV-R^2 进行比较）

预测变量概述

分类变量

- 已给出的分类变量，如城市、年份、房屋用途、交易权属等
- 根据“所在楼层”的高中低创建的分类变量
- 根据“房屋朝向”创建的方向分类变量

数值变量

- 从 float 或文本中提取数值，如建筑面积、燃气费等
- 对“房屋优势”等文本型信息，计算其逗号、顿号的个数，以衡量优点个数
- 对“交通出行”数据，计算文本中“线”、“路”二字出现的次数

预测结果展示

表 1: Metrics Table (MAE)

Metrics	In Sample	Out of Sample	Cross-Validation	Datahub Score
OLS	562317	567490	562449	59.911
Lasso	590178	581092	591451	58.779
Ridge	168409	183632	184816	83.898
ElasticNet	505804	498960	507701	63.743

表 2: Metrics Table (RMSE)

Metrics	In Sample	Out of Sample	Cross-Validation	Datahub Score
OLS	1361240	1532380	1365113	59.911
Lasso	1363015	1359713	1366506	58.779
Ridge	464373	538360	564927	83.898
ElasticNet	1258577	1229800	1263137	63.743