



# 房价预测期中展示



汇报人：徐子禾



中國人民大學  
RENMIN UNIVERSITY OF CHINA



# 模型构建

1

## 数据处理：

连续型变量如有效面积、聚类估价、楼层采用 1%-99% 分位截断，空缺值以中位数填充；无序分类变量如环线，缺失值映射为 4，如是否配备电梯，字段空缺设为 0。

2

## 变量选取：

通过计算各特征与房价的皮尔逊相关系数，选取绝对值大于 0.2 的变量作为自变量；对于与房价存在实际关系但皮尔逊系数偏小的变量，尝试与有效面积、总房间数等构造交互项，保留相关性提升后的特征。

3

## 共线性处理：

计算各特征的方差膨胀因子（VIF）评估多重共线性情况。对于 VIF 值显著大于 10 的变量，说明其与其他变量高度相关，可能导致模型系数不稳定、解释性下降，因此将这些变量剔除，保留信息量更独立的特征，以提升模型的稳健性和泛化能力。

4

## 模型训练与验证：

采用线性回归、LASSO、Ridge、Elastic Net 进行拟合，训练前对特征进行极差标准化处理，保证系数可比，防止例如“有效面积”这种大值变量掩盖其他变量影响，提升训练稳定性。



# 结果展示

Model	RMSE (In sample)	RMSE (Out of sample)	RMSE (CV)	MAE (In sample)	MAE (Out of sample)	MAE (CV)	Datahub Score
OLS	1049221.01	1012372.63	1041523.15	593318.91	574438.77	588529.3	62.008
LASSO	990561.53	994013.81	1007768.6	581422.03	588789.46	595056.53	62.032
Ridge	985841.21	953396.56	981508.69	583607.65	569048.27	580153.26	70.081
Elastic Net	1130114.97	1100069.61	1132316.65	702898.77	687817.55	699763.24	60.03



# 模型创新

1

针对非数值型变量，采用两种处理方式：一是映射为数值（如“房屋年限”转为 $1/2/3$ ）；二是提取结构化信息（如从“房屋户型”提取“总房间数”）。同时对高维特征简化处理，如将“房屋朝向”降维为是否朝南的0-1特征。

2

特征构建中优先使用套内面积（缺失则退回建筑面积）作为房屋面积指标，更贴近实际使用空间，减少公摊干扰，提升预测精度与用户居住体验的一致性。

3

基于房源经纬度引入地理聚类特征，使用KMeans将训练集中房源划分为10个空间簇，计算各簇单位面积房价均值后，生成反映区域单位房价水平的“聚类均单价”变量，其与真实价格相关性高达0.83，有效增强模型的空间感知能力。

4

引入多组交叉特征（如“房间数×聚类均单价”、“电梯×面积”、“精装×面积×均价”等），用于建模变量之间的交互关系。这些特征能够更细致地捕捉房屋结构、配套设施与地理估值之间的非线性影响，有效提升模型的拟合能力和价格预测准确性。



中国人民大学

RENMIN UNIVERSITY OF CHINA

感谢指导！



中國人民大學

RENMIN UNIVERSITY OF CHINA