中国人民大学
RENMIN UNIVERSITY OF CHINA

期中建模情况汇报

姓名：郭立为
学号：2022202747

2025年4月3日

重点1-填补缺失值

- **以对“环线”数据的处理为例**

✓一开始：缺失值较多（42726/84133）

✓匹配“小区数据”中的“环线”变量（42841/84133）

✓利用相同小区/板块的“环线”变量填补（64098/84133）

✓利用经纬度和KNN算法填补（84133/84133）

- **其它的：例如使用常识填补**

重点2-改进原始数据以便建模

- **改进原始数据的不足-几个例子**

✓“城市”、“区域”、“板块”之间的包含关系

✓房屋价格后的单位需要去除

✓“户型”的信息需要进行拆分

✓虚拟变量：城市×环线、区域×板块

重点3-模型的运算

- **OLS、Ridge、LASSO回归**

✓运算能力的估计不足

✓模型表现有待优化

模型	样本内误差	样本外误差	六折交叉验证误差	Datahub分数
OLS	2729.72	2757.95	2773.23	75.631
LASSO	2763.92	2805.50	2824.19	58.658
Ridge	2760.41	2788.41	6895.53	59.789
最佳模型	2729.72	2757.95	2773.23	75.631