AI 与 Python 期中展示

——预测房价的线性机器学习模型

滕明阳

中国人民大学
经济学院

2025 年 4 月 3 日

数据处理

1. 改变名称显示层级结构（便于动态处理，为补全测试集中残缺值的上级法做准备）。

   [“城市”,“区域”,“板块”,“小区名称”]

   ⇒

   [“location1”,“location2”,“location3”,“location4”]

2. 正则表达式将有效信息分解而最大化利用。

   比如： fool=“低楼层 (共 28 层)”

   ⇒

   relative_height=“低楼层” total_floors=28

   再比如： directions=“东南北”

   ⇒

   east_south=True nouth=True south=False···

   再比如： area_gross=“52.3m^2” ⇒area_gross=“52.3”

   再比如： frame=“2 室 1 厅 1 厨 1 卫”

   ⇒

   “room”=2 “hall”=1 “kitchen”=1 “bathroom”=1

3. 缺失值补全。本数据集的缺失值是由多种原因造成的：1. 有时候缺失值可能代表了一种可能的类别，例如大多数的房屋的“别墅类型”是 NaN，标记为“非别墅”。2. 有时缺失值可以根据其他变量补全，例如可以先估计“建筑面积-套内面积”比，再根据“建筑面积”补全“套内面积”。更一般的变量，采用线性回归补全。

测试集中出现新类别怎么办？

建模中普遍遇到的问题是：如何估计测试集中不存在的区域、板块与小区！

1. 构建了 *get_location_relation*() 函数，得到城市、区域、板块与小区之间的树状从属关系。
2. 构建了 *get_neighbor_location_list*() 函数，对于测试集中出现的新地点，在训练集中找到它的邻居。
3. 寻找邻居的方法有两种：1. 上级法，共同属于同一个上级的便为邻居（简单方便，但无法控制邻居数，估计不精确）；2. 距离法，假定编号是连续排列的，相邻 n 个编号所代表的地点就是邻居（强烈依赖于连续排列的假定）。
4. 最终采用上级法。若有 m 个邻居，就在此 m 个邻居的虚拟变量上增加 $\frac{1}{m}$。

线性模型

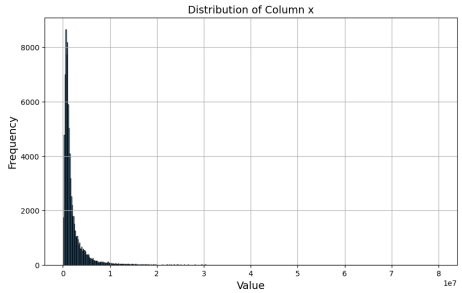

图: 房价的概率密度函数

1. 观察到房价具有非常明显的右偏分布，从而我们对其取对数处理。
2. 经比较发现，地理区分度越细致，预测效果越好，因此采用小区固定效应。
3. 环线在不同城市具有非常不同的含义，因此生成环线虚拟变量与城市虚拟变量的交互项。
4. 不同城市的房价走势可能非常不同，因此生成年份固定效应与城市效应的虚拟变量。

表 1: 模型性能指标

Model	In-sample MAE	Out-of-sample MAE	Cross-validation MAE	GRADE
OLS	178 332.0029	183 249.5126	0.099 102	82.994
Ridge	178 782.6243	194 134.9238	0.100 335	82.903
Lasso	179 426.3256	178 765.5761	0.109 196	78.711