期末模型展示

欧阳语博 2022201462

中国人民大学

2025 年 6 月 11 日

特征工程

一、特征处理

舍去：套内面积、非结构文本数据、重复含义数据
挖掘：梯户比例：梯、户 | 房屋优势：地铁、房本 | 所在楼层：总层数、楼层位置 | 交易时间、上次交易：交易年份、交易季度、挂牌时间

二、高基数特征

对小区房屋、小区楼栋作分箱；物业公司、开放商样本数小于 10 的归类于“其它”

三、构建非线性特征（重要）

尝试构造衍生特征：总房间、楼栋密度、卫室比、舒适程度 ••••••

填补缺失值

首先依靠已知信息推断，最终再用中位数、众数填补

一、

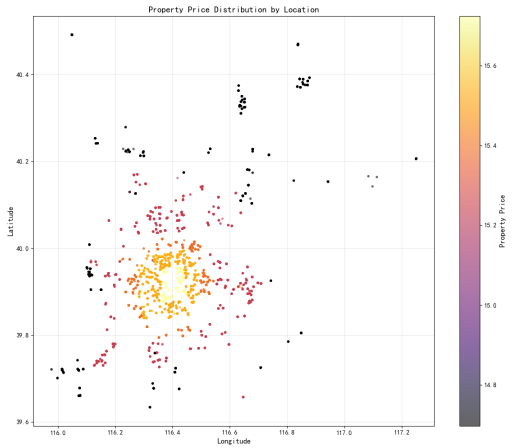
梯户比例、配备电梯和总层数 | 房屋年限 | 建筑结构和装修情况

二、

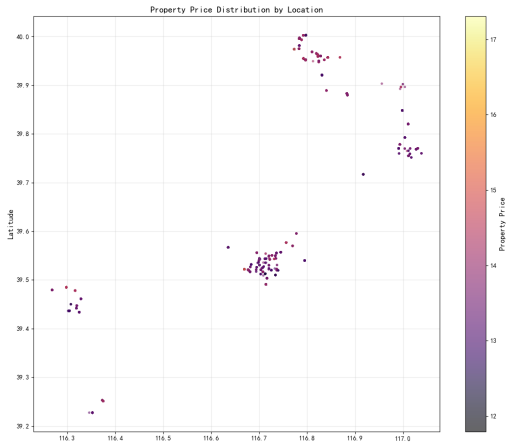
环线位置：

1、同一文本在不同城市有不同含义

2、先用同板块填补，再用同区域 KNN 填补



(a) 城市 0 价格热力图



(b) 城市 1 价格热力图

类别编码和自动编码器

类别编码非常重要！这决定了树模型的性能。在这里是标签（目标）编码
想想树的原理

一、排序

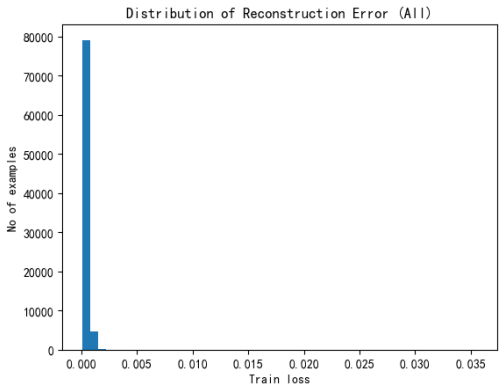
先验顺序：装修情况、建筑结构、交易权属、房屋用途、别墅类型；楼层位置
高基数特征：开发商、物业公司、板块、区域、房屋朝向

二、为什么不使用独热编码？

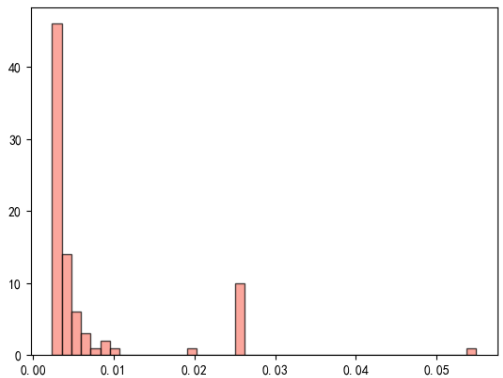
效率低、维度爆炸，还有呢？

自动编码器

识别错误处理的特征和错误样本



(a) 重建误差图



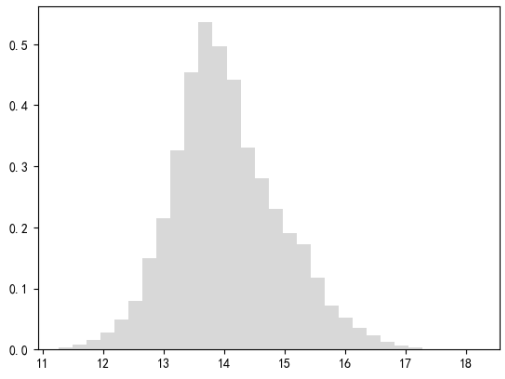
(b) 重建误差极端值图

尺度缩放

树模型有必要对目标变量尺度缩放吗？

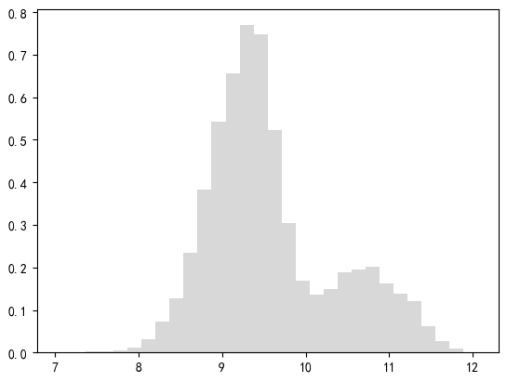
一、无

二、对数化价格



(a) 对数价格图

三、对数化单价



(b) 对数单价图

”index” 谜题和“理想”的树

Rank	Gain	Weight	Cover
1	环线	建筑面积	别墅类型
2	城市	房屋朝向	小区
3	房屋用途	总层数	lat
4	lat	楼层位置	交易权属
5	燃气费	梯户比例	lon
6	区域	板块	环线
7	lon	交易月份	房屋用途
8	开发商	装修情况	舒适程度
9	物业费	lon	停车位
10	小区	lat	燃气费
11	有无地铁	年份	物业公司
12	建筑结构	房屋总数	楼栋密度
13	供水	小区	绿化率
14	卫	物业费	房屋总数
15	板块	停车位	配备电梯
16	厨	楼栋密度	物业费
17	梯	开发商	开发商
18	房屋总数	交易年份	容积率
19	总层数	户	供水
20	物业公司	物业公司	楼栋总数
21	绿化率	卫室比	板块
22	楼栋总数	区域	有无地铁
23	楼栋密度	楼栋总数	供电
24	停车位	舒适程度	户
25	舒适程度	容积率	梯
26	容积率	室	建筑结构
27	供电	厅	区域
28	户	建筑结构	总层数
29	楼层位置	绿化率	厨
30	装修情况	燃气费	卫
31	别墅类型	房屋年限	梯户比例
32	梯户比例	房本	室
33	交易权属	产权所属	建筑面积
34	配备电梯	填充后	卫室比
35	室	卫	房屋朝向
36	建筑面积	城市	房屋年限
37	房屋年限	梯	交易年份
38	卫室比	交易权属	厅
39	厅	房屋用途	楼层位置
40	房本	有无地铁	房本
41	交易年份	厨	装修情况
42	年份	供水	城市
43	交易月份	供电	年份
44	房屋朝向	别墅类型	交易月份
45	产权所属	配备电梯	产权所属

index: 中高中 | 供热费：中低高 | 挂牌时间：低高高

模型表现和参数

一、模型表现

模型	Datahub	测 MAE	测 RMSE	训 MAE	训 RMSE	运行时间
XGBoost	85.124	143322	419884	69755	134204	30 s
随机森林	84.292	142941	442176	53375	167862	1 min 30 s

二、XGBoost 参数

max depth=6,eta=0.05,sunsample=0.8,colsample by tree=0.8,num rounds=7100

三、随机森林参数

max features=0.6,n estimator=200