



AI & Python Final Pre

汇报人：梁艺高

2025-06-04

数据处理 (基于midterm的改进)

时间变量

“距上次交易”

“上次交易”异常值 (2121-11-26 00:00:00)

缺失值填充为0，最小值控制为0

添加“交易年份”、“交易月份”

配套设施

“是否存在” (bool) 改为“存在数量” (int)

分类变量

按经纬度填充

detail

3k+填充8w+

'物业类别', '建筑年代', '绿化率', '容积率',
'物业费', '燃气费', '供热费', '停车位', '停车费用'

rent

'车位', '用水', '用电', '燃气', '采暖'

log

方差大，右偏

'log_area', 'log_last_trans', 'log_rent'

ANN

模型

- wide
 - 'leaky_relu' — 11(0.001)
 - 尝试 — wide增加重要特征('使用面积', '距上次交易', '平均租价', '室')权重 — datahub分数↓
- deep
 - 三层“残差连接”(residual_block) — 'swish'('silu') — BatchNormalization
 - 12(0.01) — Dropout(0.3)
- loss function — Huber Loss ($\delta=95\%$ 分位数)

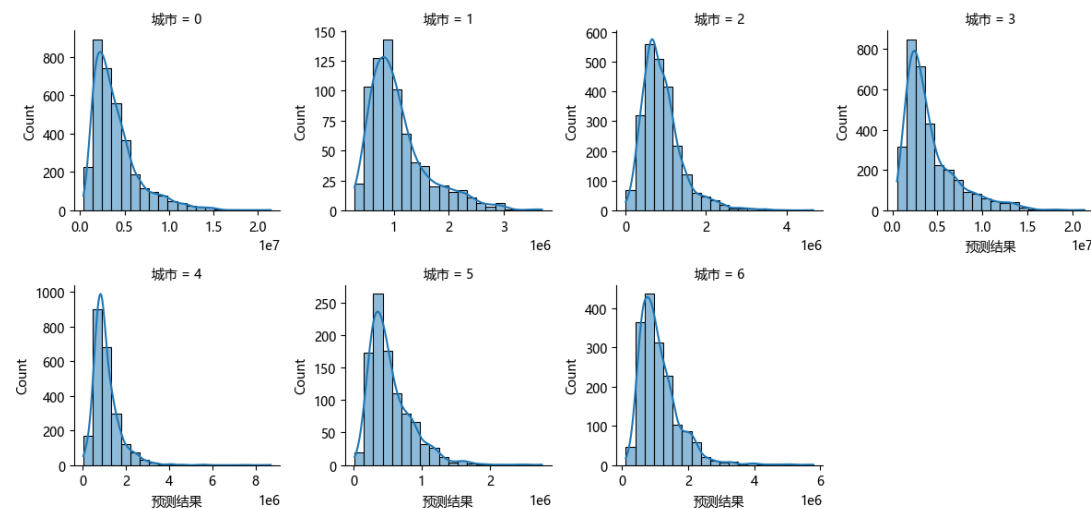
训练

- 早停策略
 - 监控MAE
 - patience=15
- 学习率动态调整
 - 初始学习率0.001
 - 学习率衰减比例0.5
 - patience=5 — 比早停策略更敏感
 - 最小学习率下限1e-5
- Test MAE — 15w左右

预测结果

- 未处理 — 78分+ (最高78.703分)
- 按“经纬度”(半径20km)缩尾 — datahub分数 < 未处理
- 按“城市”缩尾 — datahub分数 < 按“经纬度”缩尾

各城市预测结果直方图



XGBoost & stacking

Top 10重要特征:

	feature	importance
9	使用面积	52948.0
20	距上次交易	28151.0
1	交易月份	19855.0
8	总楼层	17330.0
35	log_area	15022.0
4	室	13726.0
2	lon	12764.0
34	平均租价	12161.0
18	户数	11124.0
19	elevators_per_household	11089.0

XGBoost

RMSE: 270,841.25
MAE: 151,242.94
R²: 0.9717

参数

- 核心参数 — 损失函数MSE — 最大树深8 — 学习率0.01
- 正则化参数
 - 节点分裂所需的最小损失减少值0.3
 - L1正则1 — L2正则1.5
 - 子节点所需的样本最小值3
- 随机性参数 — 每棵树训练时随机采样的数据/特征比例0.7（行/列采样）
- 其他
 - 监控指标rmse
 - 最大轮数8000 — 早停轮数50

Top10重要特征 — 74分+

stacking

- 元模型ridge(alpha=0.1)
- 76分+ — MAE: 115981.95

神经网络MAE: 125652.67401437927
XGBoost MAE: 122307.03279879095
元特征相关性:

	NN	XGB
NN	1.000000	0.994181
XGB	0.994181	1.000000

An aerial photograph of a city at sunset, with mountains in the background. The sky is a deep orange-red, and the city lights are visible. The text "感谢观看" and "Thanks for watching" is overlaid in the center.

感谢观看

Thanks for watching

汇报人：梁艺高