RENMIN UNIVERSITY OF CHINA

AI & Python Midterm Pre

Enzo Liang 2021200657

DATA CLEAN

环线：根据经纬度，利用KNN模型填充缺失值
拆分训练集和测试集（7:3），在测试集上评估，测试集准确率: 99.94%

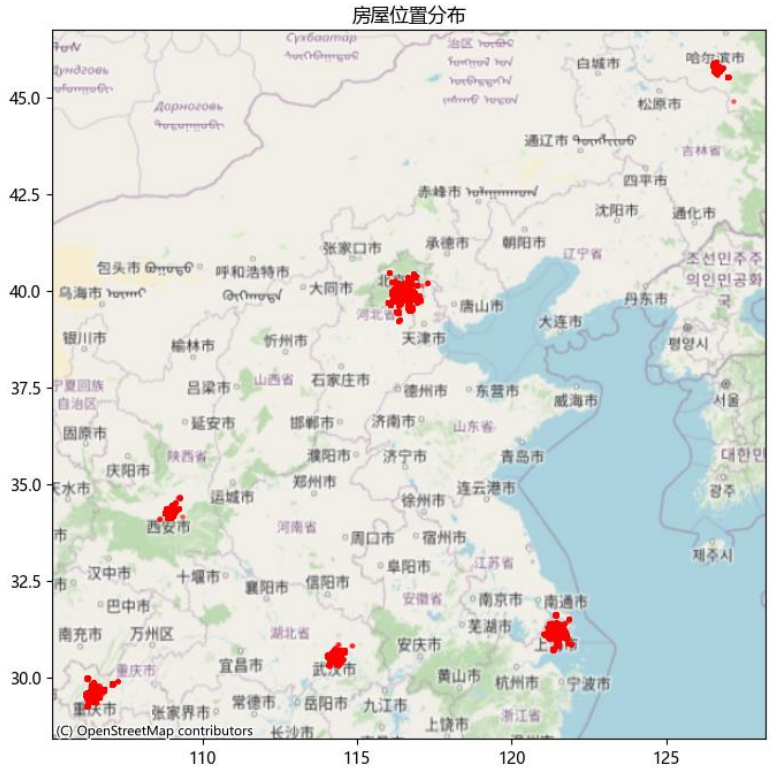
朝向、房屋优势：One-Hot Encoding

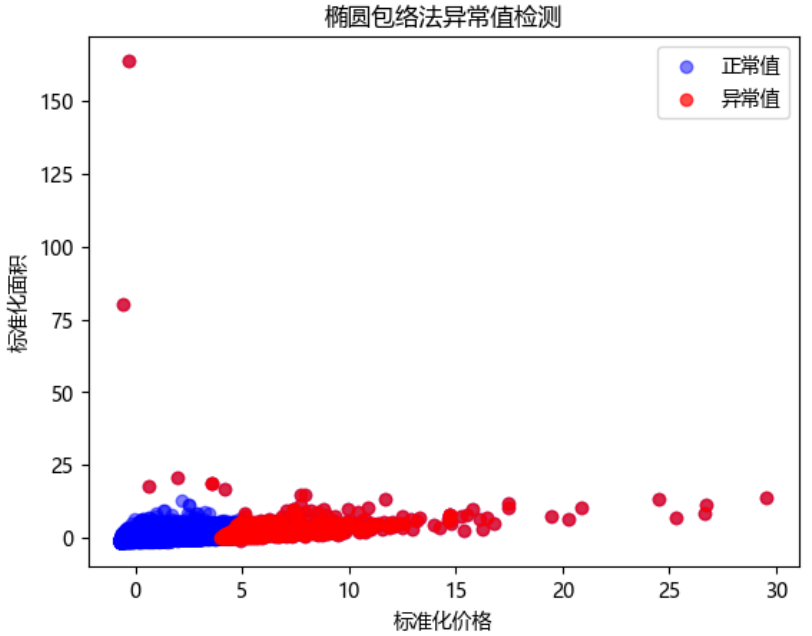
户均梯数：综合“配备电梯”、“梯户比例”
“配备电梯”notna，且“有”，计算“户均梯数”
“配备电梯”notna，且“无”，“户均梯数”计0
“配备电梯”na，“户均梯数”notna，计算“户均梯数”
“配备电梯”、“户均梯数”na，“户均梯数”计0
（已提前验证不存在“配备电梯”notna，“户均梯数”na的情况）

距上次交易：“交易时间”-“上次交易”，fillna(0)

周边配套、交通出行：
商业配套：int(bool(re.search(r'商场|购物|超市|市场|商圈|商业体|餐饮', text)))
教育配套、医疗配套、生态配套、金融配套
地铁线路数：len(re.findall(r'\d+号线', text)),
公交线路数：len(re.findall(r'\d+路', text))
主干道数、交通枢纽

价格-使用面积：椭圆包络法异常值检测
原始数据量：84133，过滤后：83291，异常值：842





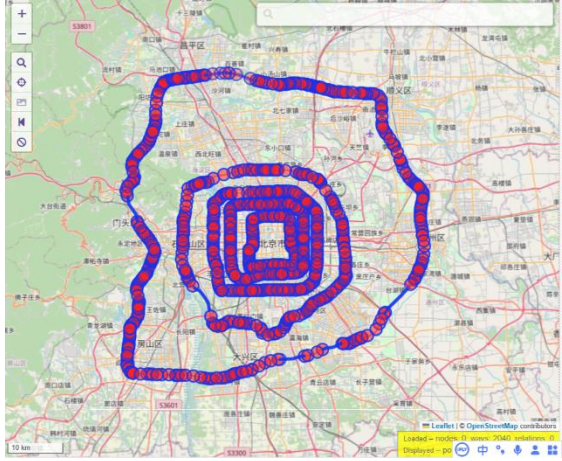
EXPERIENCE OF FAILURE

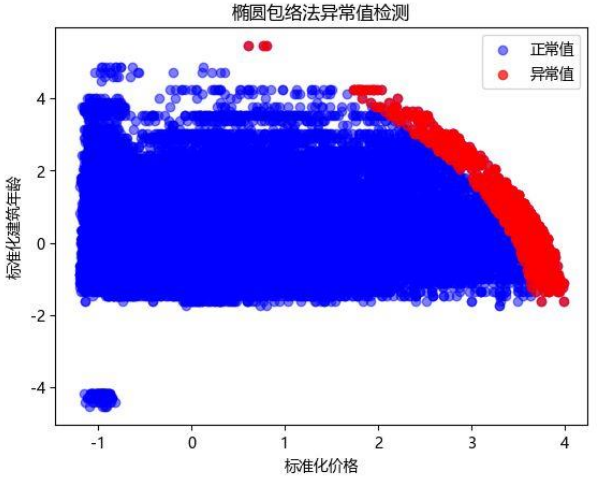
环线： 环线经纬数据（环线数据不闭合）
使用面积： “交易权属”平均得房率填充“套内面积”（缺失值过多）
电梯指数： 0.4*电梯有无+0.3*电梯需求+0.3*梯户密度（效果不好）
自然语言： TF-IDF（提取关键词质量差）
价格-建筑年龄： 椭圆包络法异常值检测（效果不好）
Interactions： selector = SelectKBest(mutual_info_regression, k=10)
（筛选慢，效果差）

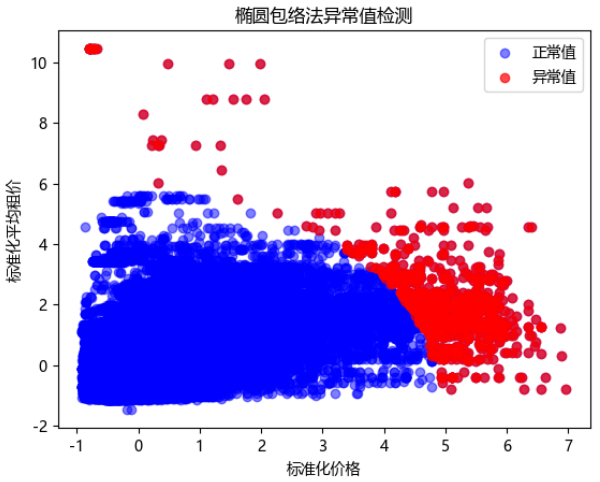
FINE-TUNE MODEL

Add features：
引入： 从Details、Rent根据经纬度引入对应的“建筑年份”、“价格”、“面积”
生成： 生成“建筑年龄”、“平均租价”
处理： 依次使用“小区”、“板块”平均填充“建筑年份”，then dropna
椭圆包络法剔除价格-平均租价outlier（最终：82363）

Add non-linearity and interactions：
('poly', PolynomialFeatures(degree=2, include_bias=False))
('poly', PolynomialFeatures(degree=2, interaction_only=True))







CHANGE HYPERPARAMETERS

网格搜索：

离线任务@2025/4/1 17:23:22

运行成功 · 开始时间：2025/04/01 17:25 · 运行时长： 10小时43分26秒

```
'Lasso':
        'regressor__alpha': [0.001, 0.01, 0.1, 1, 10],
        'regressor__max_iter': [5000, 10000],
        'regressor__tol': [1e-4, 1e-5]
'Ridge':
        'regressor__alpha': [0.1, 1, 10, 100, 1000]
'ElasticNet':
        'regressor__alpha': [0.001, 0.01, 0.1, 1],
        'regressor__l1_ratio': [0.2, 0.5, 0.8],
        'regressor__max_iter': [5000, 10000]

grid_search = GridSearchCV(models[name], param_grids[name],
                           cv=5, scoring='neg_mean_squared_error',
                           n_jobs=-1, verbose=1)
```

Lasso best params: 1, 10000, 0.0001
Ridge best params: 0.1
ElasticNet best params: 0.001, 0.8, 10000

性能指标：

主要性能指标报告：

模型	训练集R^2	测试集R^2	交叉验证R^2
OLS	0.920809	-8.37118e+12	-3.05623e+13
LASSO	0.920506	0.914823	0.913025
Ridge	0.920632	0.91488	0.913221
ElasticNet	0.914058	0.909978	0.907796

Negative
70.698（分数最高）
70.499
69.145

详细误差指标报告：

模型	训练集RMSE	测试集RMSE	交叉验证RMSE
OLS	460202	4.71745e+12	6.03521e+12
LASSO	461082	475856	482193
Ridge	460715	475697	481654
ElasticNet	479415	489202	496464

主要性能指标报告：

模型	训练集R^2	测试集R^2	交叉验证R^2
OLS	0.921934	0.916471	0.913856
LASSO	0.921555	0.916297	0.913617
Ridge	0.921744	0.916376	0.913888
ElasticNet	0.915091	0.911366	0.908351

70.127
70.592
70.512（R^2最高）
69.004

详细误差指标报告：

模型	训练集RMSE	测试集RMSE	交叉验证RMSE
OLS	456920	471229	479914
LASSO	458027	471720	480578
Ridge	457476	471498	479827
ElasticNet	476526	485415	495000