中国人民大学
RENMIN UNIVERSITY OF CHINA

房价预测项目期末汇报

汇 报 人：曹馨元

学　　号：2021202449

专　　业：明德环境“经济学-科学”拔尖人才实验班

汇报日期：2025年6月5日

特征工程I：环线

中国人民大學 RENMIN UNIVERSITY OF CHINA

数据特点

- 城0、1、2、3、4、6环线变量在利用details表填充后存在空缺值
- 城5环线变量完全缺失

STEP 1 非城5环线空值处理

KNN

K-means

Ring-Minmax Boundary-Book

Ring-Median-Boundary-Book

经纬度

中心坐标 → 点位距离 → 环线类型

STEP 2 城5环线空值处理

信息填补

费马点(Weiszfeld算法)

城市中心坐标

点位距离

$$f(x)=\sum_{i}^{n} w_i \cdot \|x-x_i\|$$

$$x^{(k+1)}=\frac{\sum_{i=1}^{n}\frac{w_i x_i}{\|x^{(k)}-x_i\|}}{\sum_{i=1}^{n}\frac{w_i}{\|x^{(k)}-x_i\|}}$$

环数确定与环分类

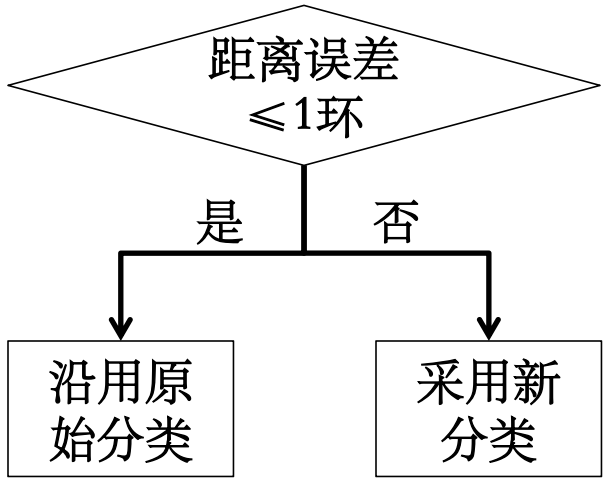
比较各城市点位距离分布

确定最相似城市（城6）

确定环数并填充环线类型

STEP 3 环线类型校验

校验标准



特征工程II：城市-区域-板块

中国人民大学 RENMIN UNIVERSITY OF CHINA

编码

hierarchical_entity_embedding

Binary_encoding

采用三位二进制 对7个城市进行编码			
城1	0	0	0
...	...	...	...

加入城市-区域-板块-小区的目标编码特征

作为新特征加入

STEP 1

计算各样本每平米对数房价Target。

STEP 2

对于Train/Test数据集每一条交易数据，根据交易日期索引前后半年该City/Area/District/Block的房屋交易记录。

STEP 3

计算所有索引记录target的平均值作为新特征补充。

Train / Test		
City	Time	City_target
0	2024-06-01	mean(Target)
...	...	...

Subset dataframe (Train)		
City	Time	Target
0	2024-01-01	k1
...	...	...
0	2025-01-01	kn
...	...	...

Test数据集填充

City_target → mean(City_target)

District_target → mean(District_target)

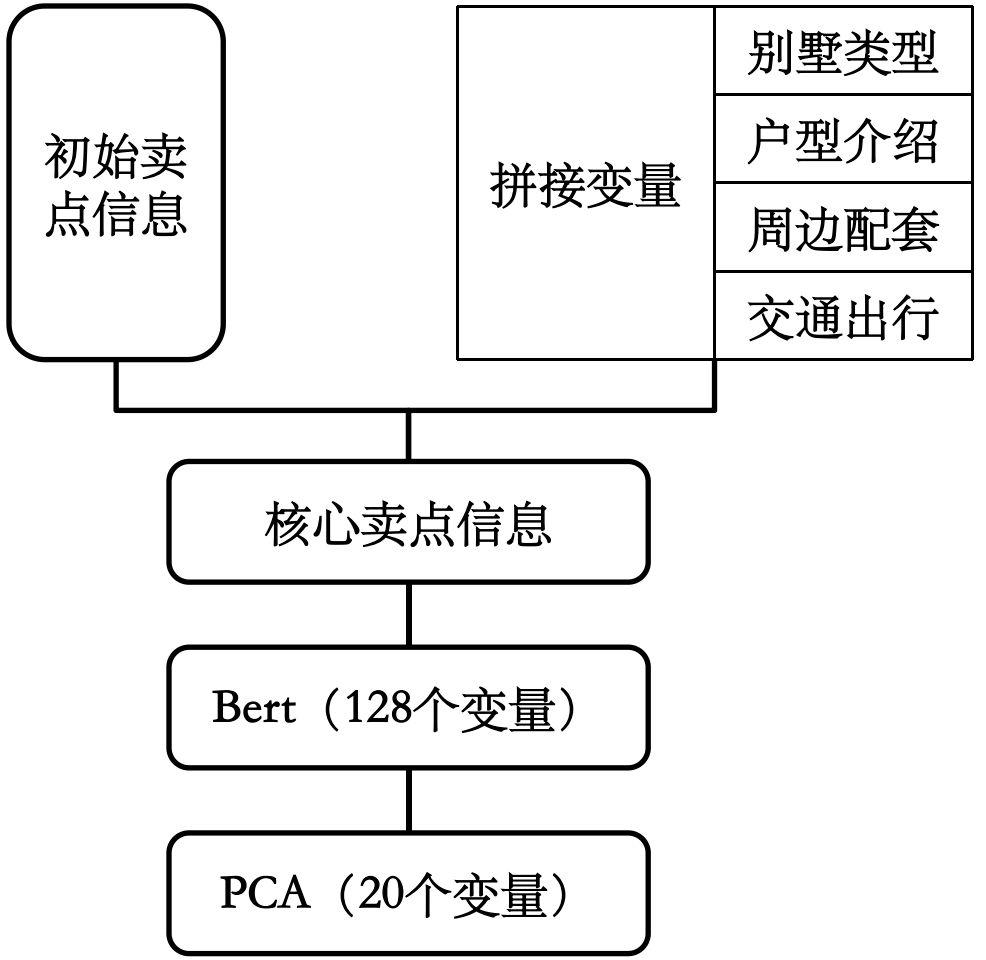
Block_target

特征工程III：租金以及核心卖点信息

目标：计算月租金信息rent

最近每平月租金
=实际租金÷折算系数÷期数÷面积

租金类型	月租金、季租金、 半年租金、年租金
折算系数	月租金（1.07） 季租金（1.0） 半年租金（0.965） 年租金（0.92）
期数	月租金（1） 季租金（3） 半年租金（6） 年租金（12）



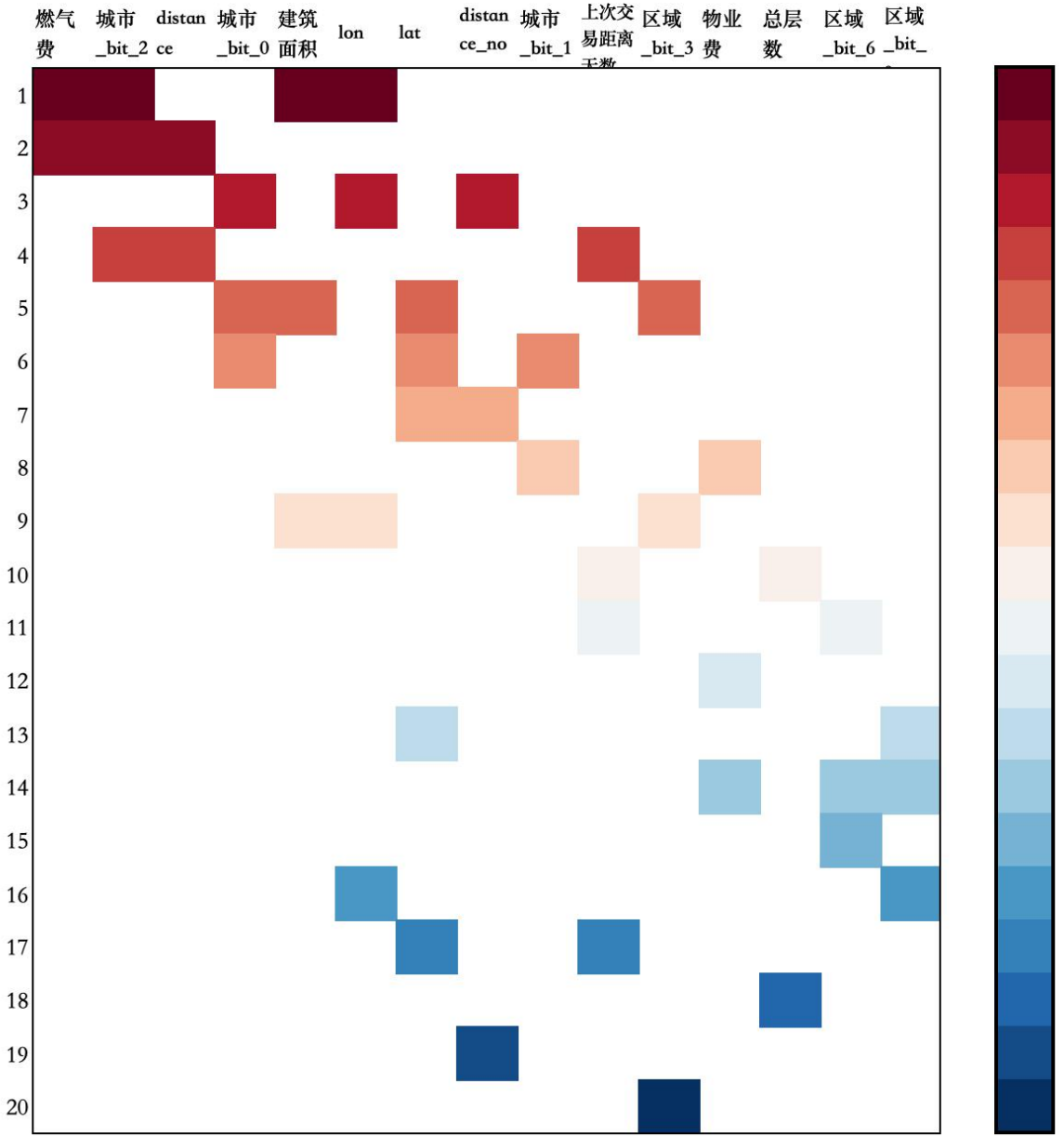
特征工程IV：辅助特征

External Data	
来源：Wind	
特征名称	
30大中城市商品房成交面积	贷款市场报价利率LPR_1年
非制造业PMI_建筑业	固定资产投资完成额
30大中城市商品房成交套数	M1同比
商品房销售面积	M2同比
十大城市成交面积	CPI当月同比
二手房出售挂牌量指数	全部工业品PPI当月同比
二手房出售挂牌价指数	GDP不变价
申万房地产指数	消费者信心指数
水泥价格指数	沪深300指数
钢材综合价格指数	城镇调查失业率
贷款市场报价利率LPR_5年	社会消费品零售总额

其他特征		
特征名称	处理方法	处理结果
房屋户型	正则提取数字	室、厅、厨、卫（4个特征）
所在楼层	文本解析	楼层位置（序数编码）、总层数
房屋朝向	独热编码	东、南、西、北（4个特征）
梯户比例	文本解析+计算	梯数、户数、每梯户数（3个特征）
交易时间	时间特征提取	交易年份、交易季度、交易月份
上次交易	计算时间差	上次交易距离天数
装修情况	序数编码	{毛坯:1, 简装:2, 精装:3, 其他:2}
楼层位置	序数编码	{地下室:0, 底层:1, 低楼层:2, 中楼层:3, 高楼层:4, 顶层:5}
配备电梯	二值编码	{有:1, 无:0}
房屋年限	序数编码	{未满两年:1, 满两年:2, 满五年:5}
产权所属	二值编码	{非共有:1, 共有:0}

模型构建

Models	Parameters	Range	Optuna
LASSO 30	alpha	0.1-1000	1.570
	max_iter	1000-20000	12284
	tol	1e-6 - 1e-3	2.028e-05
	selection	cyclic, random	cyclic
XGBoost 70	n_estimators	300-800	652
	max_depth	6-20	14
	learning_rate	0.02-0.15	0.022
	reg_alpha	0.0-5.0	0.666
	reg_lambda	3.0-10.0	3.866
	subsample	0.6-1.0	0.929
	colsample_bytree	0.6-1.0	0.771
	min_child_weight	1-10	10
Neural Network 20	n_layers	2-4	3
	dropout_rate	1e-4 - 1e-2	0.109
	learning_rate	10-100	0.0089
	batch_size	[128, 256, 512, 1024]	256
	weight_decay	1e-6 - 1e-3	0.000944



模型构建

中国人民大学 RENMIN UNIVERSITY OF CHINA

MAE	In sample	out of sample	Cross-validation	Datahub Score
LASSO	4.719269e+05	4.690274e+05	1.518839e+06	40.496
Random Forest	8.122901e+04	1.282786e+05	1.511168e+06	57.083
XGBoost	4.066294e+04	1.244846e+05	1.510086e+06	55.106
LightGBM	2.132834e+04	1.251043e+05	1.509259e+06	57.433
Neural Network	1.083373e+06	1.450076e+05	1.654973e+05	64.455

RMSE	In sample	out of sample	Cross-validation	Datahub Score
LASSO	1.060033e+06	7.757916e+05	2.873975e+06	40.496
Random Forest	4.214088e+05	2.804277e+05	2.169371e+06	57.083
XGBoost	1.282898e+05	2.728614e+05	2.164212e+06	55.106
LightGBM	5.175759e+04	2.708707e+05	2.160850e+06	57.433
Neural Network	1.888098e+06	3.084639e+05	1.984256e+06	64.455

感谢聆听

汇报人：曹馨元

汇报日期：2025年6月5日