

```
#房屋优势特征管道
def house_advantage_ext(X):
    """
    房屋优势特征提取函数
    将房屋优势字符串编码为四维向量 [装修, 房本满五年, 地铁, 房本满两年]
    每有一个特征就记为1, 例如"装修、房本满五年" -> [1,1,0,0]
    """
```

```
from sklearn.pipeline import make_pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import FunctionTransformer

FACILITY_FEATURES = ['冰箱', '天然', '天然气', '宽带', '床', '暖气', '洗衣机', '热水器', '电视', '空调', '衣柜']

def facility_ext(X):
    """
    从配套设施字符串中提取特征向量

    参数:
    X: 输入数据, 可以是Series、DataFrame或二维数组

    返回:
    numpy数组, 形状为(n_samples, n_features), 其中n_features=11
    """
```

```
#房屋户型处理管道
import re
class HouseTypeEncoder2(BaseEstimator, TransformerMixin):
    """房屋户型编码器 - 将'3室1厅1厨1卫'格式编码为四维数值特征"""
```

```
def direction_ext(X):
    """
    房屋朝向编码函数
    将房屋朝向字符串编码为四维向量 [东, 南, 西, 北]
    每有一个方向字符就记为1, 例如"东南" -> [1,1,0,0]
    """
```

Model	In_sample_MAE	Out_of_sample_MAE	CV_MAE	Kaggle
OLS	107418.6739	107258.4649	107593.1324	59.51
LASSO_Grid	110589.8347	110584.7792	110745.0865	56.15
RIDGE_Grid	107418.6908	107258.5021	107593.1213	57.9
OLS	539620.1426	546830.373	540624.8751	
LASSO_Grid	549402.8018	555203.9157	550135.6885	
RIDGE_Grid	539619.97	546830.424	540624.5991	
best model				59.51

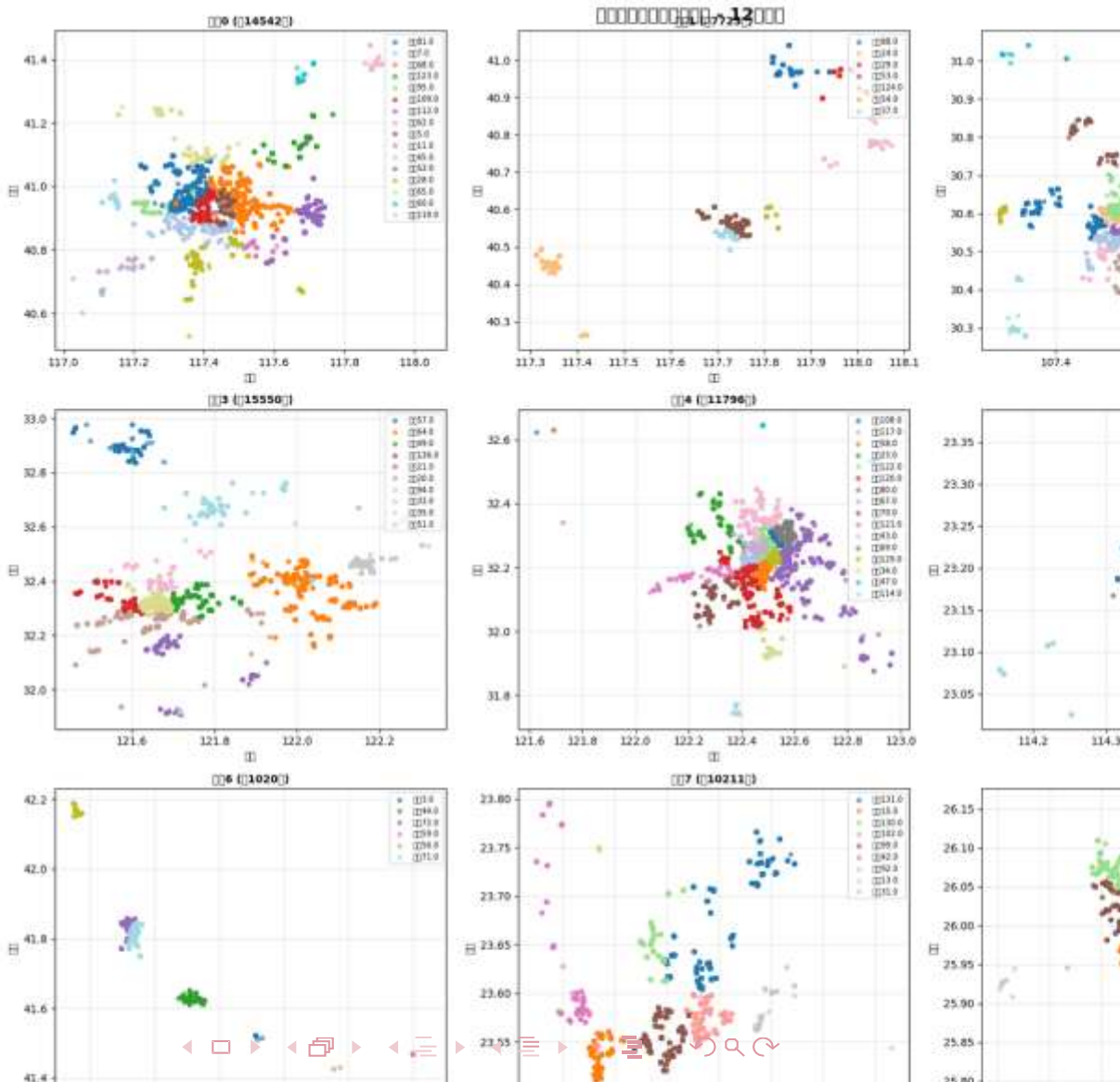
```
X_train, X_test, y_train, y_test = train_test_split(
    cleaned_rent.drop('Price', axis=1),
    np.log(cleaned_rent['Price']),
    test_size=0.2,
    random_state=111
)
```

```
from sklearn.compose import ColumnTransformer

rent_preprocessing=ColumnTransformer([
    ('poly',poly_pipeline2,['面积','绿化率_小数','物业费_提取','燃气费_提取','log_面积','log_绿化率','log_物业费','log_燃气费']),
    ('cat',cat_pipeline,['城市','付款方式','租赁方式','电梯','用水','用电','燃气','建筑结构','供水','供电']),
    ('fq',cat_frequency,['产权描述','租期','区县','板块','物业类别']),
    ('facility',facility_pipeline,['配套设施']),
    ('geo',cluster_simil,['lon','lat']),
    ('type',house_type_pipeline2,['户型']),
    ('floor',floor_pipeline2,['楼层']),
    ('direction',direction_pipeline,['朝向'])
])
```

- 使用正则表达式分解有效信息
- 文本数据赋分
- 类别特征编码
- 分层级的缺失值补全方法
- 特殊城市的差异化处理

指标	样本内	样本外	交叉验证	Kaggle
OLS	320582.07	319596.14	328584.51	73.5
Ridge	601814.70	605654.64	621220.14	62.3
最佳线性模型	320582.07	319596.14	328584.51	73.5
指标	样本内	样本外	交叉验证	Kaggle
OLS	85179.74	88370.71	87848.58	73.5
Ridge	85174.49	88369.65	87640.89	62.3
最佳线性模型	85174.49	88369.65	87640.89	62.3



PRICE_Model	In sample MAE	Out of sample MAE	CV MAE	Kaggle Score
OLS	738418.52	726924.67	739022.16	54.31
LASSO	739308.43	727820.67	739856.26	54.33
Ridge	738418.52	726924.67	739022.16	54.31
ElasticNet	738621.47	727139.84	739215.67	54.31
SGD	740125.89	729320.10	745567.55	54.31
Huber	739523.46	729386.15	740130.92	54.04

RENT_Model	In sample MAE	Out of sample MAE	CV MAE
OLS	156664.86	157723.88	156845.09
LASSO	157715.75	158718.22	157893.30
Ridge	156664.88	157723.90	156845.11
ElasticNet	157609.73	158620.36	157788.61
SGD	159805.71	161424.36	158015.11
Huber	157479.68	158959.46	157748.79

## ①一次提取数值信息

→一次处理多列减少工作量

```
def extract_numeric_features(df):
    """提取数字特征"""
```

房屋总数	楼栋总数
1317户	19栋
2317户	40栋
1554户	20栋
66户	27栋

房屋总数  
1317

物业费
1.3-1.65元/月/m²
0.65元/月/m²
1.98-2.98元/月/m²

(1.3+1.65)/2

建筑年代
1955-2000年
2005年
2011-2012年
2009-2011年

建筑年代  
(1955+2000) / 2



建筑年代\_老旧: True

数据输入

数据预处理



特征工程



特征筛选



投入模型

## ②单列文字信息→更多的可用特征

房屋户型
2室1厅1厨1卫
3室1厅1厨1卫
3室2厅1厨2卫
6室3厅1厨3卫
1房间1卫

室厅厨卫  
2 1 1 1

房屋朝向
南北
南北
东南
东南西北

南 True 北 True

朝向个数: 2

南北通透: True





mialex2005

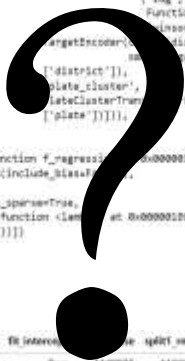
Alex Wong

黄之杰 2023200251

```
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=15; total time: 3.4s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=15; total time: 3.4s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=20; total time: 8.0s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=20; total time: 5.9s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=30; total time: 1.8s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=30; total time: 5.9s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=30; total time: 8.3s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=30; total time: 5.6s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 26.5s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 21.8s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 20.2s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 20.5s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 20.1s
[CV] END regressor__poly__degree=2, regressor__poly__interaction_only=True, regressor__selector_k=78; total time: 20.2s
```

```
## 训练 ##
最佳参数: ('regressor__poly__degree': 2, 'regressor__poly__interaction_only': False, 'regressor__selector_k': 30)
最佳 CV R^2: 0.886197787128167
```

```
## 训练 (已融合) ##
Pipeline(memory_location=C:\Users\121du\AppData\Local\Temp\sklearn_cache_8lqpdul\sklib\,
steps=[('preprocessor',
ColumnTransformer(transformers=[('geo_cluster',
LonLatClusterTransformer(n_clusters=40),
['lon', 'lat']),
('log_cols',
Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')),
('log',
FunctionTransformer(func=ufunc 'log'))],
['gas_feenum', 'park_area']),
('log1p_num',
P...
sparse_output=False))],
['city', 'way2rent', 'floor',
'wat_sup', 'ele_sup',
'build_struc']),
('district_target',
TargetEncoder(cols=['district'],
smoothing=15),
['district']),
('period2rent',
TargetEncoder(smoothing=15,
verbose='period2rent'),
['period2rent']),
('plate_cluster',
PlateClusterTransformer(),
['plate'])]]),
('selector',
SelectBest(k=30,
score_func=function F_regression,
score_func=0.0000001088456086)),
('poly', PolyomialFeatures(include_bias=True,
degree=2,
max_features=1000,
min_degree=1,
sparse_output=False,
transformer=FunctionTransformer(accept_sparse=True,
func=FunctionTransformer(accept_sparse=True,
func=function (X, y): 0.0000001088456086)),
('model', LinearRegression()))])
```



	geo_n_clusters	district	smoothing	plate_n_clusters	fit_intercept	split1_score	split2_score	split3_score	split4_score	split5_score	mean_score
8	58	54	11	True	1110873	1102881	1145309	1081228	1105181	1020895	1100871
7	31	48	14	False	1117018	1092608	1148632	1082603	1100300	1097088	1101283
9	47	50	12	True	1120708	1081057	1147455	1037470	1100880	1083888	1101356
6	36	51	10	True	1122580	1090218	1145743	1083719	1110388	1096044	1101430
10	51	57	6	False	1123103	1090309	1144040	1095455	1105165	1082312	1101722

Metrics (Price)	In sample	out of sample	Cross-validation	Kaggle Score
OLS	521784.319	517497.624	526609.124	64.85
LASSO L1:0.01	751114.486	734893.989	753207.844	57.65
Ridge L2:100	522947.226	518393.887	527682.534	64.63
ElasticNet Alpha:0.2 L1:0.01	755031.273	738051.991	716246.446	59.67

```
0 city 98899 non-null int64
1 Price 98899 non-null float64
2 area 98899 non-null float64
3 way2rent 98899 non-null object
4 elevator 98899 non-null float64
5 wat_sup 81159 non-null float64
6 ele_sup 81575 non-null float64
7 gas 94317 non-null float64
8 period2rent 15966 non-null object
9 lon 98899 non-null float64
10 lat 98899 non-null float64
11 year 98899 non-null float64
12 district 94222 non-null float64
13 plate 93755 non-null float64
14 built_year 72758 non-null float64
15 greening_rate 74497 non-null float64
16 plot_ratio 74819 non-null float64
17 property_fee 76748 non-null float64
18 build_struc 78158 non-null object
19 gas_feenum 73842 non-null float64
20 park_area 73420 non-null float64
21 floor 98835 non-null float64
22 room 93849 non-null float64
23 hall 93671 non-null float64
24 kitchen 0 non-null float64
25 bathroom 34765 non-null float64
26 east 98899 non-null int64
27 south 98899 non-null int64
28 west 98899 non-null int64
29 north 98899 non-null int64
30 transaction_year 98899 non-null int64
31 bed 98899 non-null float64
32 wardrobe 98899 non-null float64
33 air_condi 98899 non-null float64
34 wash_mach 98899 non-null float64
35 water_heat 98899 non-null float64
36 dwelling 98899 non-null float64
37 ground_comm 98899 non-null float64
38 commerce 98899 non-null float64
39 carport 98899 non-null float64
40 apart 98899 non-null float64
41 villa 98899 non-null float64
42 density_ratio 94268 non-null float64
43 reflect_sentiment 98899 non-null float64
```



```
Pipeline(steps=[('preprocessor',
ColumnTransformer(transformers=[('geo_cluster',
LonLatClusterTransformer(n_clusters=25),
['lon', 'lat']),
('log_cols',
Pipeline(steps=[('imputer',
SimpleImputer(strategy='median')),
('log',
FunctionTransformer(func=ufunc 'log'))],
['winson',
Winsorizer()),
('scaler',
StandardScaler()))],
['gas_feenum', 'park_area']),
('log1p_num',
P...
sparse_output=False))],
['city', 'way2rent', 'floor',
'wat_sup', 'ele_sup',
'build_struc']),
('district_target',
TargetEncoder(cols=['district'],
smoothing=15),
['district']),
('period2rent',
TargetEncoder(smoothing=15,
verbose='period2rent'),
['period2rent']),
('plate_cluster',
PlateClusterTransformer(),
['plate'])]]),
('scaler', StandardScaler()),
('regressor', LinearRegression()))])
```