Midterm Presentation

Sno:2023200231 kaggleID-RhHannahZhao

innovation1: EDA: Dirty Data -> Clean Features -> Refined Models

- data cleansing:text(e.g. comment---->情感分析？ too complicated;risk of data leaking)
- 数据填充： numeric: median? knn? category: 众数&one hot code
- Handling Skewness: 对目标变量采用 `np.log1p`转换，使数据更符合线性模型的正态性假设

innovation2: Feature Engineering

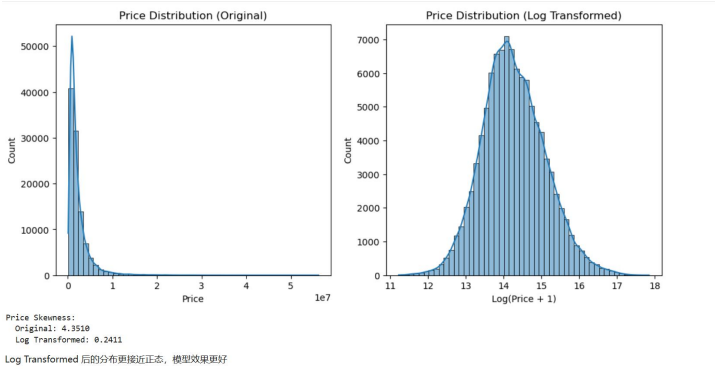
- `Regex`+`Pandas.apply`+`Sklearn Pipeline`
- Structured Parsing: `parse_layout_price` ("3室2厅1厨2卫") vs`parse_layout_rent` ("2室2厅1卫");`parse_floor_...` ("中楼层 (共23层)" / "4/6层")
- Text-Numeric Cleaning:`parse_area_price`("282.02㎡") ;`clean_numeric_text`("45%");`clean_parking`("免费")
- Derived Features:`HouseAge`=`TransactionYear`-`BuildingYear`
- use pipeline: simple imputer & standard scaler -- train on train set and process test set

innovation3: Cross-Task Feature Fusion

- `Pandas.groupby`+`merge`
- Price引入平均租金:`train_rent.csv`--`AvgRent_Community`,Rent引入平均售价:`train_price.csv`--`AvgPrice_Community`
- 为线性模型提供了强大的、基于租售比的金融信号

innovation4: Robust Validation

- `np.log1p`+`make_scorer`+`GridSearchCV`
- 自定义评估函数:`mae_on_original_scale`函数,用`np.expm1`将结果逆转换回原始价格后再计算 MAE
- Optimizing Real Metric:`mae_scorer`传递给`GridSearchCV(..., scoring=mae_scorer)`,确保超参数调优是直接优化最终的业务指标--原始价格MAE



模型名称	In sample	out of sample	CV	Kaggle Score
OLS (Price)	220,587	248,458	236,420	68.43
Lasso (Price)	411,051	413,690	412,289	62.26
Ridge (Price)	222,899	224,268	235,221	68.80
ElasticNet (Price)	338,494	340,599	340,950	64.78
LGBM (Price)	393,081	390,973	149,536	74.05

模型名称	In sample	out of sample	CV	Kaggle Score
OLS (Rent)	52,926	56,917	56,315	68.43
Lasso (Rent)	86,554	87,045	86,794	62.26
Ridge (Rent)	53,347	54,043	56,281	68.80
ElasticNet (Rent)	71,509	71,761	71,955	64.78
LGBM (Rent)	118,670	119,765	48,372	74.05

Midterm Presentation

学号:2023200193　kaggleID：xjfaaa-0193　姓名：熊俊沣

1.智能数据预处理与安全类型转换：

•**核心逻辑**：通过正则表达式匹配和基于列名的语义分析，自动判断该进行数值转换还是分类编码。对于疑似数值型的列（如包含“面积”、“数量”等关键词），使用正则表达式
•**安全转换**：safe_numeric_conversion函数是此环节的基石。它能智能处理带单位的字符串，例如将“123.7㎡”安全地转换为数值123.7，并对转换成功率进行监控，为后续处理提供依据。
•**容错机制**：整个流程被包裹在异常捕获（try-catch）中，单列处理的失败不会导致整个流程中断，极大增强了系统的鲁棒性

2.多策略混合编码应对高基数分类变量

低基数特征频率与目标编码：对于唯一值较少的分类变量（如“装修情况”），我们主要采用**频率编码**，用每个类别在数据集中出现的次数作为其数值。有时会辅以**目标编码**（在有目标变量且基数非常低时），用该类别对应目标变量的均值（经过平滑处理）来编码，更能体现其与房价的关系
高基数特征哈希与频率编码：对于唯一值很多的特征（如“周边配套”描述文本），独热编码会产生大量特征。我们采用**哈希编码**，将类别值映射到固定维度的空间（如1000维），有效控制维度。同时也会使用频率编码作为补充。

3.深度时空特征与地理网格创新

时间特征深度解析：从“交易时间”等日期字段中，不仅提取年份、月份、季度等基础信息，还创新性地生成**周期性编码**（通过正弦余弦转换月份等周期数据）和**市场季节性标志**（如标识3-5月、9-10月为租赁旺季），帮助模型捕捉市场的周期规律。
多粒度地理网格：利用经纬度信息，在不同精度级别（如0.001度, 0.01度）上创建**地理网格特征**，将连续的地理坐标转换为离散的网格ID，从而表征微观地理位置差异。
多中心距离计算：计算每个房源到多个城市重要节点（如市中心、商业中心、交通枢纽）的欧氏距离，生成一系列**区位特征**，量化了房源的相对区位价值

模型名称	In sample	out of sample	CV	Kaggle Score
OLS	0.4369	0.4333	0.4303	45.90
Lasso	0.5490	0.5452	0.5491	42.69
Ridge	0.4369	0.4332	0.4374	45.89
ElasticNet	0.4911	0.4872	0.4912	33.44

模型名称	In sample	out of sample	CV	Kaggle Score
OLS	0.4244	0.4256	0.4248	45.90
Lasso	0.5199	0.5164	0.5200	42.69
Ridge	0.4244	0.4256	0.4248	45.89
ElasticNet	0.4713	0.4694	0.4714	33.44

期中课堂展示

2023200190 徐国祥

创新1：数据处理

- 利用**‘客户反馈’**数据，借助**人工智能（豆包）**分析获取分主题关键词、情感词，按‘，’切割分句进行打分，并排除句子长短带来的影响

```
topic_keywords = {
    "设施硬件": ["光照","通风","网速","阳台","装修","照明灯","结构","厨房","储物","水"
    "空间布局": ["面积","布局","路况","地铁","楼间距","空间","风景","道路","房型","出行
    "环境体验": ["体验","素质","潮气","气味","交通","绿植","隔音","地段","绿化","通勤"
    "服务配套": ["服务","成熟度","安保","保安","维修","安防","管理","监控","智能","社区
}

positive_words = {"快","通透","稳定","时尚","预留","标准","顺畅","完善","新颖","舒适"

negative_words = {"老旧","松动","进风","漏水","渗水","裂","停用","坏","破损","异响",
```

创新2：特征工程

- 对**‘楼层位置’**分箱处理：所处楼层在总楼层数的位置分为低、中、高 楼层
- 对**‘朝向’**简化处理：将杂乱朝向分类提取为主要朝向
- 依据**‘城市’**分别处理：‘城市’编码不同，数据缺失特征也不同，按‘城市’分类填补缺失值

```
    df_chengshi_8 = max_min_filler(df_chengshi_8)
    print(get_high_missing_cols(df_chengshi_8))
[15]

…  共发现 6 列缺失值占比超过 50.0%:
     - 列 '装修': 缺失值 5973 个，占比 82.11%
     - 列 '车位': 缺失值 6115 个，占比 84.07%
     - 列 '采暖': 缺失值 6411 个，占比 88.14%
     - 列 '环线位置': 缺失值 7274 个，占比 100.00%
     - 列 '供热费下限': 缺失值 5834 个，占比 80.20%
     - 列 '供热费上限': 缺失值 5834 个，占比 80.20%
   ['装修', '车位', '采暖', '环线位置', '供热费下限', '供热费上限']
```

创新3：模型构建

- 依据**‘城市’**缺失特征构建**二值变量**，捕捉系统性缺失特征，并与原特征交乘

模型名称	In sample	out of sample	CV	Kaggle Score
OLS (Price)	310241.27	308818.99	309638.60	
Lasso (Price)				
Ridge (Price)				

模型名称	In sample	out of sample	CV	Kaggle Score
OLS (Rent)				
Lasso (Rent)				
Ridge (Rent)				
ElasticNet (Rent)				

期中课堂展示

Metrics	In sample	out of sample	Cross-validation	Kaggle Score
OLS	0.4686	0.4881	0.4686	30.96
Lasso	0.4686	0.4882	0.4687	30.96
Ridge	0.4686	0.4881	0.4686	30.96
Elastic Net	0.4686	0.4882	0.4686	30.96

Metrics	In sample	out of sample	Cross-validation	Kaggle Score
OLS	0.4430	0.4505	0.4431	30.96
Lasso	0.4430	0.4505	0.4431	30.96
Ridge	0.4430	0.4505	0.4431	30.96
Elastic Net	0.4430	0.4505	0.4431	30.96