数据分析和处理

数据清理

- 分月清除3σ的outliers
- 注意到存在所有属性相同，仅租金不同的数据，对这些数据进行合并，月租取mean

关于分类信息处理

- 地理相关分类信息分布较为均匀，档次房型等分类信息不均匀
- block存在同一个编号相距很远的情况，所以把block属性修改为street+block的属性
- street属性需要处理相同街道不同大小写的问题
- type属性不加'-'和加'-'分布差异显著，所以生成两个type属性，一个统一去除'-'，一个保留，防止忽略其中包含的信息
- 尝试对分类信息根据编号，均值，标准差，中位数进行聚集，仅保留标准差的效果最好，加上中位数和均值的效果次之（聚集后分布密度图）
- 建成日期当作分类属性处理
- block等属性存在数据样本过少容易使得模型过拟合的问题，先尝试对每个聚集后的block属性增加服从正态分布$(0, \frac{1}{\sqrt{n}})$，其中n为该类别个数的扰动误差，防止对某个别稀少类别过拟合，效果不够理想
- 将少于一定值(目前16)的类别全部设置为others，并将聚合方差设置为0，效果理想，同时对test中有train中没有的类别也视为others处理，获得泛化能力。

关于额外信息的处理

- 学校，超市等分布均匀，对结果影响已包含在地理信息内，额外增加的效果不好。
- 银行等股票波动和房价相关性不强。
- coe数据和房价增长曲线较为接近，对结果有提升

关于新增属性

- 注意到region对地理信息的划分过于粗糙，town，area等划分过于细小，新增根据DBSCAN进行聚类的标签属性作为分类属性，在$eps = 0.005, min_sample = 1$的情况下，对数据点划分出20个块。
- 利用$var = E(x^2) - E(x)^2$，用KNN分别计算月租和月租的平方，实现对周围Kl临近房价的标准差计算，分别计算对K=16，32，64，128作为新增属性，表示周围房价状况。

关于时间信息的处理

- 注意到均值标准差中位数都随着时间增加并且波动较大
- 尝试按月份拆分分别训练，由于每个月样本分布在2000左右太小，效果不理想
- 将月租分月标准化后作为categories的聚集属性，效果理想
- 月租按月标准化后会导致收敛的RMSE和原RMSE发生偏移，所以根据标准化后的月租处理完分类属性后，将其还原为原月租，后尝试减去每个月平均值（减去均值不影响RMSE计算结果），效果更好

模型采用

- 利用ms-flaml库，用ray库在soc-cluster的xcnf(AMD EPYC 7763-64core, 1TB ram)上进行分布式并行地分别对lgbm，xgboost，catboost，进行参数搜索
- 其中前两者移除原categories数据，只保留聚合后的标准差，catboost保留原category数据
- 根据三份的初步提交结果考虑权重，分别计算调和，几何，算术，平方平均数，其中调和，几何平均数会更倾向低值，平方平均数会更倾向高值，其中几何平均效果最好（但不多），说明高估情况

略多于低估。