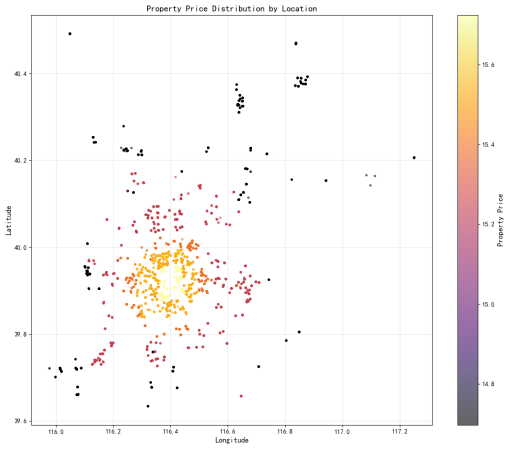
特征工程

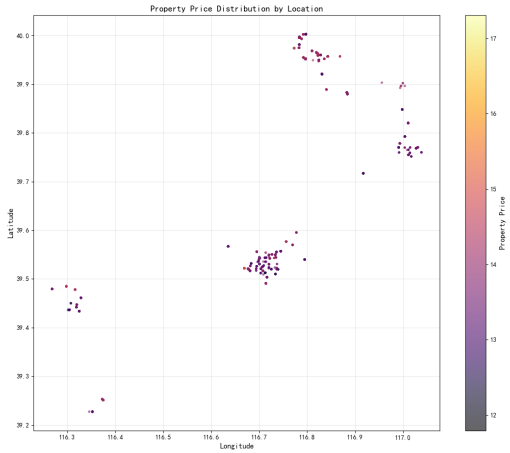
一、总体

舍弃缺失比例过高的数据；推断缺失值

二、环线位置



(a) 城市 0 价格热力图



(b) 城市 1 价格热力图

三、构建新的特征

比率：楼栋密度、卫室比；时间特征

目标编码

一、树的特性

树的每一次分裂是为了将不同的样本分开

二、分箱

处理离散数据

小区房屋、小区楼栋

三、处理高基数数据

设置一个阈值，防止过拟合：threshold=10

开发商、物业公司

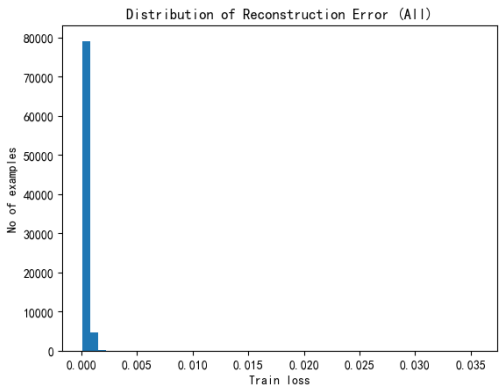
自动编码器

构造一个小型的神经网络，拟合特征之间的关系。以自身为训练目标

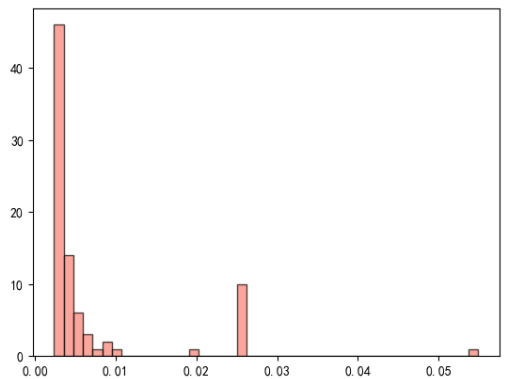
一、重构误差

二、回顾特征工程

三、检测异常值



(a) 重建误差分布图



(b) 重建误差极端值图

模型调整

一、XGboost

learning rate、num boost round

二、随机森林

max features、max depth、n estimators

三、集成

加权、Stacking