房价预测期末汇报

汇报人：赵一铭 2021200567

- 目标：基于多维数据构建高精度房价预测模型
- 数据来源：包含小区、房屋特征、文本描述等多种类型
- 主要挑战：异构数据融合、特征选择、模型性能提升
- 核心方法：融合嵌入式神经网络与预训练语言模型的深度回归策略

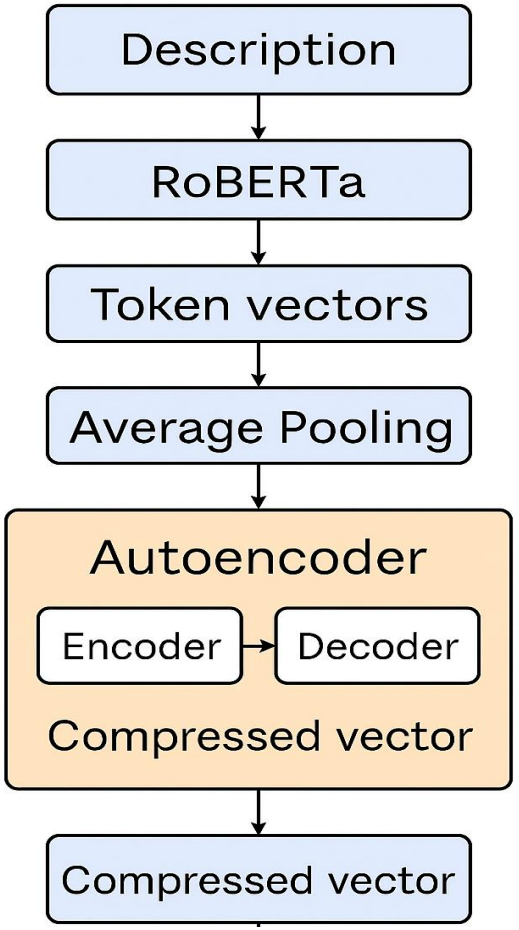
一、数据预处理

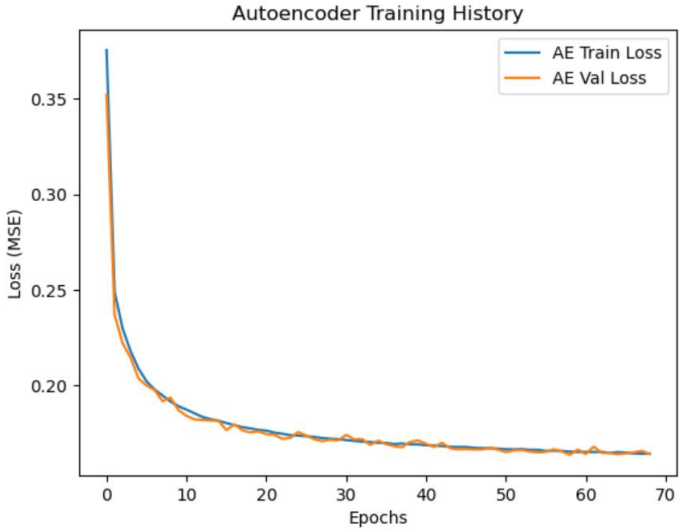
功能整合：集中完成数据清洗、字段转换、类别映射等任务

变量名称	变量来源/处理方式	缺失值处理方式	变量名称	变量来源/处理方式	缺失值处理方式
城市_1 ~ 城市_6	城市1-6的	无	城市1交叉 ~ 城市6交叉	城市i * 环线	无
医院	'周边配套'中包含关键词	无	总层数	提取所在楼层中的数值并转为int	无
公园	'周边配套'中包含关键词	无	室	房屋户型字段正则提取	众数
超市	'周边配套'中包含关键词	无	厅	同上	众数
商场	'周边配套'中包含关键词	无	厨	同上	众数
银行	'周边配套'中包含关键词	无	卫	同上	众数
学校	'周边配套'中包含关键词	无	建筑面积	正则提取数字	无
地铁	'交通出行'或'周边配套'中包含关键词	无	套内面积	正则提取数字	按拟合公式填补（0.8 * 建筑面积 + 0.38 （大于100）80.65（小于100））
公交	'交通出行'或'周边配套'中包含关键词	无	建筑面积正常	是否介于100~600（价格与建筑面积有较强关联）	无
高速	'交通出行'中包含关键词	无	建筑面积交叉	建筑面积正常 * 建筑面积	无
高铁	'交通出行'中包含关键词	无	lat2	lat的平方	无
机场	'交通出行'中包含关键词	无	lon2	lon的平方	无
上次年份	上次交易中的年份	交易年份减去平均交易时长	time2	(交易年份 - 2000)^2	无
交易年份	交易时间中的年份	无	配备电梯	映射为0/1	用平均数填充
交易时长	交易年份 - 上次年份	中位数填充	区域	未作处理	无
交易频率	数据中在交易年份这一年交易房产的数量	无	产权所属	映射为0或1	无
梯	梯户比例字段提取中文数字并转换	用25分位数填充	环线	映射为整数/浮点	按城市信息填入该城市75分位数
户	梯户比例字段提取中文数字并转换	用75分位数填充	房屋年限	映射为整数：'满五年'：0，'满两年'：2，'未满两年'：7（根据税率计算）	用平均数填充
梯户比例	正则化提取梯/户	用上述二值之比填充	朝南	'房屋朝向'字段是否包含'南'	默认不朝南
装修情况	映射为0-3：'精装'：3，'简装'：2，'毛坯'：1，'其他'：0	用中位数填充	别墅类型	映射为整数：None：0，'联排'：1，'叠拼'：2，'双拼'：3，'独栋'：4	0
建筑结构	映射为整数：'混合结构'：3，'钢混结构'：6，'砖混结构'：2，'钢结构'：5,'未知结构'：0，'砖木结构'：1，'框架结构'：4	用众数填充	楼层分类	映射为整数：'高楼层'：4，'中楼层'：3，'低楼层'：2，'顶层'：5，'底层'：1，'地下室'：0	无
房屋用途	映射为整数：'车库'：0，'商业'：1，'商业办公类'：1，'写字楼'：1,'底商'：2，'商住两用'：2,'老公寓'：3，'平房'：3,'酒店式公寓'：4，'住宅式公寓'：4，'公寓/住宅'：4，'公寓'：4，'公寓/公寓'：4，'公寓（住宅）'：4,'普通住宅'：5,'别墅'：6，'四合院'：6，'新式里弄'：6，'花园洋房'：6	用众数填充	交易权属	映射为整数：'使用权'：0，'集资房'：1，'拆迁还建房'：2,'动迁安置房'：2，'定向安置房'：2，'售后公房'：2,'安置房'：2,'经济适用房'：3，'限价商品房'：3，'自住型商品房'：3,'一类经济适用房'：3,'二类经济适用房'：3,'已购公房'：4，'房改房'：4，'央产房'：4,'自住型商品房'：5，'限价商品房'：5,'私产'：6，'商品房'：7	无

二、文本特征建模

- 使用 RoBERTa 提取文本深层语义
  - 输入字段：房源的“核心卖点”字段，通常为一句或几句话的自然语言描述；
  - 使用中文预训练模型 hfl/chinese-roberta-wwm-ext；
  - 采用平均池化方式将整个句子转为一个固定长度的句向量；
- 使用 Autoencoder 进行维度压缩
  - 原始RoBERTa输出为高维（如768维），对模型训练压力大；
  - 构建对称的Autoencoder神经网络结构进行降维压缩；
  - 编码器将原始嵌入压缩为低维表示（32维）；
  - 训练目标是重构原始RoBERTa向量；
- 提升非结构化文本利用效率
  - 将压缩后的文本向量与结构化数值特征/小区嵌入一同作为主模型输入；
  - 相比传统手工特征（如关键词提取、词频等），该方式更灵活、更具泛化能力；
  - 尤其对具有丰富描述的房源信息，提供了关键语义支持。



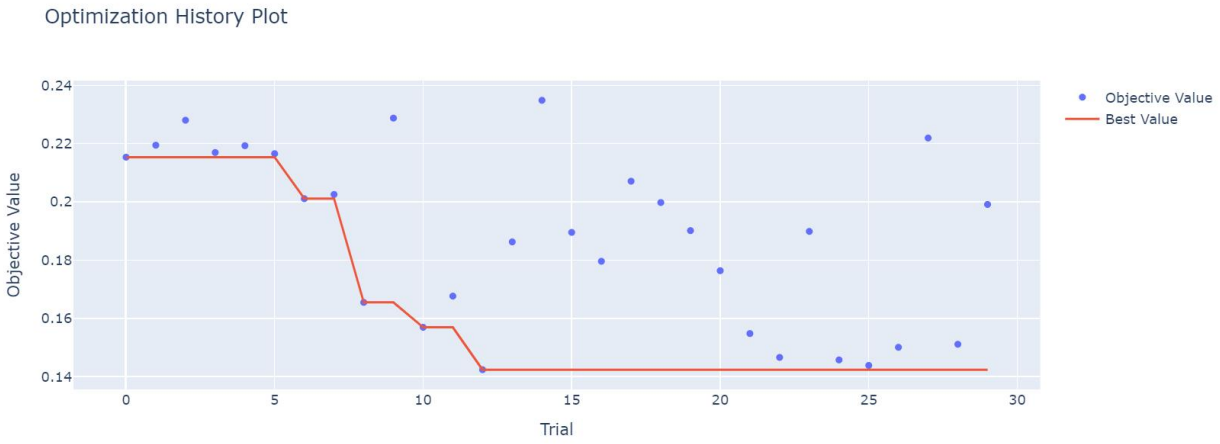


三、神经网络结构

- 双输入结构：数值特征 + 小区ID嵌入
  - 数值特征输入：包含上述的等连续变量；
  - 类别嵌入输入：小区名称通过Roberta转换为可学习的向量表示，捕捉小区语义位置和品质；
- 网络结构：Dense + BatchNorm + LeakyReLU + Dropout
  - 两层全连接网络：
    - Dense(dense1) → BatchNorm → LeakyReLU → Dropout(dropout_rate)
    - Dense(dense2) → BatchNorm → LeakyReLU → Dropout(dropout_rate)
- LeakyReLU 解决 ReLU 死神经问题；
- Dropout 防止过拟合，提升泛化能力；
- EarlyStopping 提高训练效率，避免过拟合
- ReduceLROnPlateau进行学习率调度，避免震荡或过早停止

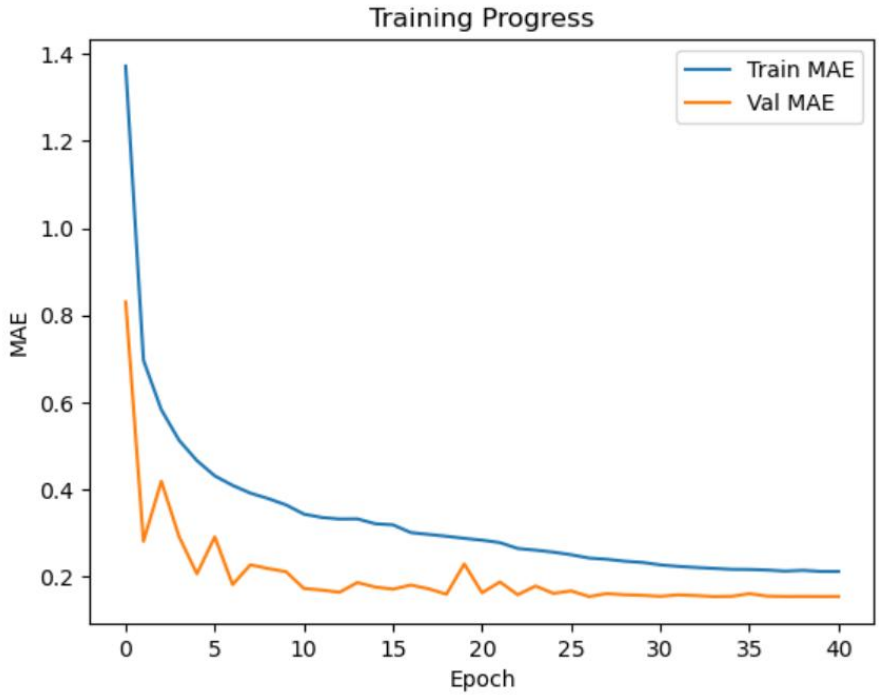
四、Optuna自动调参

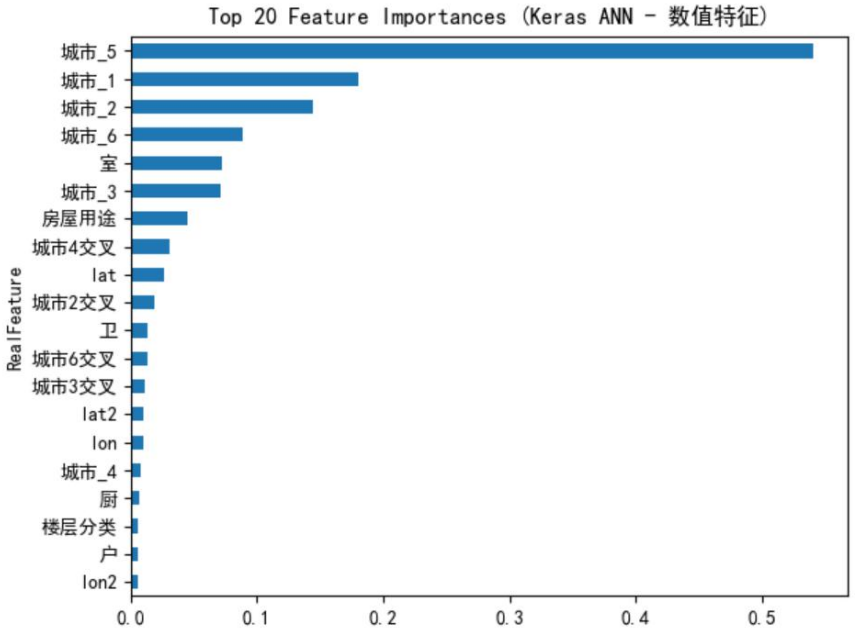
- 引入 Optuna 自动调参框架
  - Optuna 是一个现代化的超参数优化工具，支持分布式计算、早停策略和参数依赖；
  - 自动构建搜索空间并进行试验管理；
  - 与 Keras 完美集成，可快速应用于自定义模型。
- 调参维度覆盖关键模型结构和训练配置，包括：
  - dense1, dense2：隐藏层节点数，决定模型容量；
  - dropout_rate：随机失活比例，控制过拟合；
  - learning_rate：学习速率，影响收敛速度；
  - batch_size：每轮训练样本数，影响训练稳定性与效率；



结果与总结

- 评估指标：In-Sample MAE及Out-Sample MAE表现优异
- 模型优势：结构+文本多模态融合，结构灵活，语义感知强，自动调参提升效率与性能





Model	Hackthon Score	In-Sample MAE	Out-Sample MAE
ANN	74.538	241155.25	968022.70