Final exam presentation

钱彦均 2023200773

- **异常值-减少数据损失**

  boxcox+删除部分异常值+标准化 - robustscaler
- **类别变量-非线性+减少维度**

  One-Hot Encoding - OrdinalEncoder
- **正则化表达式提取数据**

  建筑面积，房屋面积，套内面积

Bert - cuda+torch+transformer

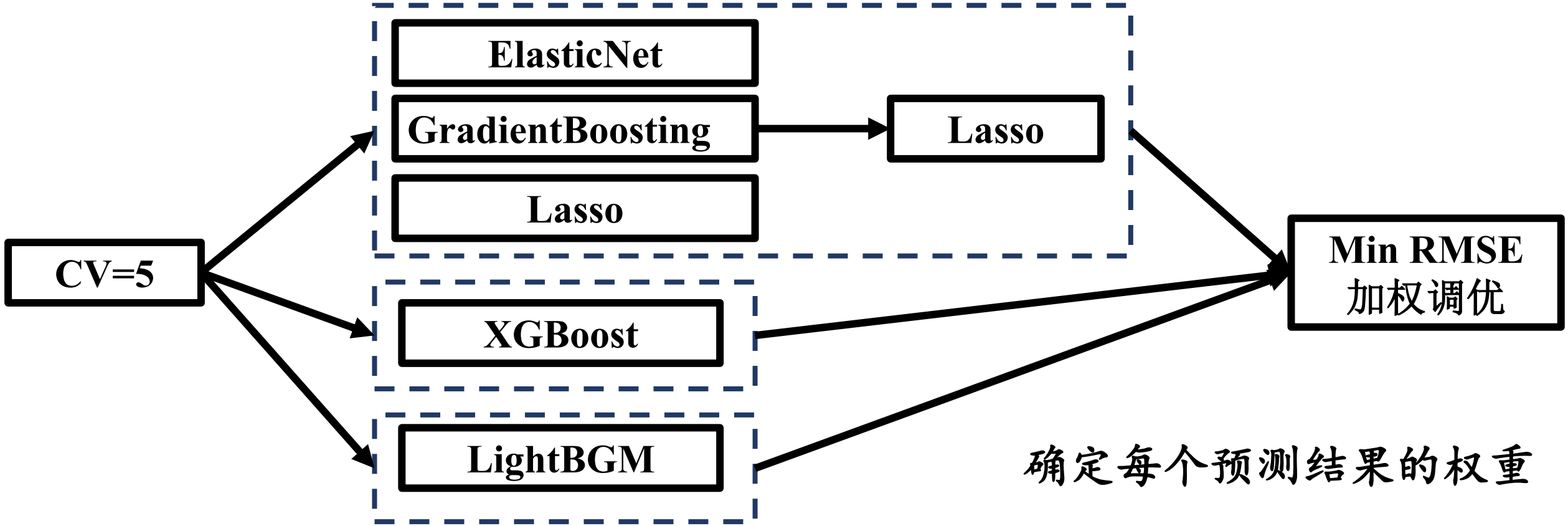
房屋优势 核心卖点 户型介绍 周边配套

提取每个token文本的CLS - PCA:768→200 - 200*4

OOM - batch_size=32 分批处理，防止爆内存

慢 - 加入文本分析后效果不明显

两层：基模型+元模型


ElasticNet
GradientBoosting
Lasso
Lasso
CV=5
XGBoost
LightBGM
Min RMSE
加权调优
确定每个预测结果的权重

Model(无文本)	RMSE
Stacking	0.129
XGBoost	0.155
LGBM	0.144
Ensemble (加权综合)	0.138

Model(加入文本分析)	RMSE
Stacking	0.161
XGBoost	0.163
LGBM	0.145
Ensemble (加权综合)	0.151