房价预测期末展示

汇报人：徐子禾

CatBoost-特征工程与建模架构的创新

1

全链路的特征工程体系设计：

结构性特征提取（如房屋户型解析出“房间数”“朝向” one-hot），复合派生特征（如“每房间面积”、“聚类估价”），时间维度提取（将“交易时间”拆分为年/月/周/天等时间特征）

2

空间建模：地理聚类与估价特征：

使用 KMeans 对经纬度聚类，为每条数据分配“聚类标签”以增强空间位置表达。

构建“聚类均价”与“聚类估价”特征，结合空间聚合均值辅助房价推测，弥补空间变量的局限。

3

丰富特征来源与高效编码机制：

引入外部详情文件（details.csv），通过“小区名称”左连接主数据，带入物业费、绿化率、楼栋总数等补充特征。

高基类变量 Target Encoding：对“区域”、“板块”、“小区名称”等使用 target 均值编码，缓解类别稀疏性、提升泛化能力。

4

模型结构为 Stacking 融合准备铺路：

输出 CatBoost 的 OOF 结构，同时架构中预留 ANN 模型通道，便于后续构建 stacking 或 blending 的二级元模型结构。

多模型集成预测具备良好扩展性与灵活性

CatBoost-工程实现与模型优化策略的亮点

1

Optuna 全自动超参搜索:

使用 Optuna 搜索 CatBoost 的多维参数（learning rate、depth、colsample、loss function 等）。
引入 MedianPruner 提升搜索效率，自动调参策略先进、灵活、高效。）

2

KFold 重训练 + OOF 验证预测框架:

全流程拆分为调参 ➡ OOF 训练 ➡ 平均预测 ➡ 提交预测，提高模型稳定性和可复现性。
每折验证 RMSE 单独评估，验证过程完整透明。

3

特征选择策略全面:

先用 LightGBM 获取重要性，筛除低贡献特征。
再进行相关性剔除（多重共线性处理），防止信息冗余。

4

工程化处理与预测安全性保障:

自定义 load_csv_robustly，自动尝试多种编码（如 UTF-8、GBK、GB18030），适配复杂中文数据集，避免数据读取失败。
对最终预测结果应用 1%~99% 分位数剪裁，控制异常极值，增强预测稳定性与实际部署安全性。

ANN-工程化能力与 AutoML 融合设计

1

多模型融合结构设计（Stacking-friendly 架构）：

保留 LightGBM 与 ANN（人工神经网络）双通道入口，具备 OOF 融合接口，为后续 stacking 融合（如 Meta-Learner）打通路径。

2

自动化特征筛选机制（Optuna + LightGBM）：

使用 LightGBM 获取特征重要性，并用 Optuna 自动确定阈值；
引入共线性检测机制，自动移除高相关冗余特征，提升模型稳定性。

3

自研式 Optuna 训练框架 + Pruning 策略：

全训练过程封装为 objective() 函数，具备完整 early-stop 判断；
支持自动 Prune Trial（自动裁剪欠佳模型），节省大量调参时间；
多线程+最优 trial 结构直接被用于最终训练与部署，贯穿模型生命周期。

4

高可控、可解释的 AutoML 策略调度器：

结构层次解耦：每层可单独设定单位数、激活函数、是否使用 BatchNorm；
支持选择多种损失函数（MSE / L1 / Huber）；
多种优化器（Adam / AdamW / RMSProp）与学习率调度器（CosineAnnealingLR / ReduceLROnPlateau）组合。

Stacking-CatBoost × ANN 双模融合策略解析

【原始数据】
├── train.csv
├── test.csv
└── details.csv
↓
【预处理 + 特征工程】
└── 基础清洗、派生特征、TE、聚类估价等
↓
【训练特征集构建】
└── train_features, test_features
↓
【第一层模型（Level-0）】
├── K-Fold ANN → OOF_ANN, Test_ANN_Preds
└── K-Fold CatBoost → OOF_CB, Test_CB_Preds
↓
【生成二级训练集】
├── X_meta_train = [OOF_ANN, OOF_CB]
└── X_meta_test = [avg(Test_ANN_Preds), avg(Test_CB_Preds)]
↓
【第二层模型（Level-1）】
└── LightGBM 元模型 → 拟合 X_meta_train, 预测 X_meta_test
↓
【最终输出】
└── Final Price = Pred_per_sqm × 面积

结果展示

模型	in-sample rmse	out-of-sample rmse	6-fold CV rmse	datahub score
CatBoost	458619	537865	605434	83.627
ANN	465031	556921	675432	84.544
blending both				84.612


中国人民大学
RENMIN UNIVERSITY OF CHINA
感谢指导!
中國人民大學
RENMIN UNIVERSITY OF CHINA