工程与管理科学
Engineering and Management Science
第 8 卷◆第 2 期◆版本 1.0◆2026 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

基于堆叠集成学习的住宅单位面积能耗预测模型

陈子涵
同济大学 经济与管理学院 上海 200092

DOI：10.32629/ems.v8i2.18463

[摘 要] 为解决住宅建筑能耗预测中单一模型稳定性不足的问题，本文提出了一种基于堆叠集成学习的住宅单位面积能耗预测方法。以美国住宅能源消费调查(RECS)2020 年数据为研究对象，通过系统化的特征工程，采用 Spearman 相关性与 LightGBM-SHAP 的两阶段特征选择策略，构建了包含 LightGBM、XGBoost、CatBoost 和神经网络的异质基学习器组合，并以支持向量回归等模型作为元学习器进行二次融合。实验结果表明，以 GBDT 为元学习器的堆叠集成模型在源域测试中达到 R^2=0.6531、MAE=0.5450，较最优单一模型分别提升 3.3%和 1.9%，证明了堆叠集成策略在提升预测精度和稳健性方面的有效性。本研究为住宅能源管理和节能决策提供了具有工程适用性的技术路径。

[关键词] 住宅能耗预测；单位面积能耗；堆叠集成学习

引言

据国际能源署统计，建筑业消耗了全球超过三分之一的能源[1]，其相关二氧化碳排放量同样占全球总量的 30%以上。在美国和欧盟地区，这一比重甚至高达 40%左右[2, 3]。因此，提高建筑能效尤其是住宅建筑的能效已成为各国能源节约和减排政策的重点。

住宅建筑的能效评估与能耗预测被视为提升既有建筑能源效率、制定能源管理策略的重要手段。许多国家都出台了住宅改造计划，如爱尔兰计划在 2030 年前改造 50 万户住宅以提升能源效率[4]，进一步凸显了住宅能耗预测的现实意义。

随着机器学习方法被引入住宅能耗预测研究中，相关研究表明，树模型与神经网络模型等方法在捕捉非线性特征关系方面具有一定优势[5]。然而，单一模型往往对数据结构和参数设置较为敏感，在不同的特征组合和样本数据条件下预测稳定性不足，难以充分满足实际的应用要求。

为进一步提升住宅单位面积能耗预测的准确性和稳定性，集成学习方法逐渐受到关注。通过融合多个具有差异性的基学习器，集成模型能够综合不同模型的预测优势，在降低模型偏差和随机误差方面表现出较好的效果。其中，堆叠集成学习（stacking）通过引入元学习器对多个基模型的输出结果进行二次学习，为复杂能耗预测问题提供了一种有效的建模思路。

1 数据来源与特征工程

1.1 数据来源与预测目标

住宅能耗预测模型的准确性高度依赖于输入特征的质量和时效性，为保障模型在泛化预测中的稳健性，本研究选取了美国能源信息署（EIA）发布的住宅能源消费调查（RECS）数据集中的 2020 年样本。

预测目标变量为单位面积能耗（Energy Use Intensity，EUI），计算公式如下：

$$EUI = \frac{TOTALBTU}{SQFTEST}$$

其中，TOTALBTU 为建筑年度总能耗(英热单位)，SQFTEST 为建筑的总居住面积（平方英尺）。

1.2 数据预处理

数据预处理包括以下步骤：

（1）异常编码与缺失机制的统一处置

我们依据变量统计的方式和问卷跳题逻辑实施了专门的类型化处置：若该列本质为布尔型数据且仅取{0，1，-2}，则将“-2”视为“否”（置为 0），以保持二元一致性；若为

跳题型数值变量（取值离散、类别数较少），则新增缺失指示器（是否为-2）并将原列中的-2 置换为 0，以在不扭曲数值尺度的同时显示编码跳题信息;其余连续/计量变量则将“-2”统一替换成“NaN”，交由后续模型与变换统一处理。这一处理方式可以在保证可识别的缺失机制被显示编码的同时，避免以单一常数“硬填充”破坏分布形态，从而降低后续模型的偏差风险。

（2）气候变量的量纲归一与跨区可比

为削弱气候尺度差异对模型训练的影响，并提升不同气候区之间的可比性，我们对 HDD65、CDD65 及其 30 年平均值（HDD30YR_PUB、CDD30YR_PUB）实施 Min-Max 归一化，派生 HDD65_norm 、CDD65_norm 等标准化列。

（3）类别变量编码

类别变量统一进行标签编码（Label Encoding），并在基于树的模型中以类别型方式参与训练，以确保分割与统计聚合均在离散层面进行；对于神经网络模型，编码后的整数进一步通过嵌入/独热等方式处理。

1.3 特征选择策略

为在有限样本下兼顾可解释性与泛化能力,我们采用“先验相关性 + 模型驱动重要性”的两步筛选：

Spearman 相关性预筛

在处理完缺失机制与量纲归一后，首先对所有数值型备选特征与 EUI 计算 Spearman 等级相关与 p 值，出于对弱非线性与单调关系的敏感性考虑，我们采用宽松阈值（$|\rho| > 0.015$，$\rho < 0.1$）保留一批“方向正确、显著性初步达标”的候选变量，用以降低后续模型的噪声维度与计算负担。

LightGBM—SHAP 的两阶段压缩

在预筛集合上，以 LightGBM 为基学习器、以 L1 回归目标（MAE）为优化准则，采用 10 折交叉验证 与 黑箱超参搜索（Optuna）寻优后训练最终模型；基于“gain”型特征重要度按累计重要度 Top 90% 形成第一阶段精简集，随后在该集合上计算 SHAP 的平均绝对贡献度并按累计 90% 再做一次压缩，得到“模型一致性高、贡献稳定”的紧致特征子集。

2 堆叠集成学习模型构建

2.1 堆叠集成框架设计

为提升住宅单位面积能耗（EUI）预测能力，本文采用两层堆叠集成学习（stacking ensemble learning）策略，将多种异质基学习器（base learners）的预测结果输入至元学习器（meta-learner）进行二次建模，从而充分利用不同模型在特征学习能力和泛化特性上的互补性。

堆叠集成框架由两层结构构成：

第一层（基学习器层）：选择四类表现互补的回归模型作为基学习器，分别为 LightGBM（LGBM）、XGBoost（XGB）、CatBoost（CatB）和神经网络模型（NN）。梯度提升树类模型在处理高维稀疏特征及建模非线性关系方面具有优势，而神经网络能够在非结构化特征交互建模中展现更强的特征表征能力。

第二层（元学习器层）：在第一层输出的基础上引入元学习器，以利用不同基学习器预测残差的互补信息。本文以支持向量回归（Support Vector Regression，SVR）作为主要元学习器，并将其与岭回归（Ridge Regression）、LightGBM 及 MLP 等其他候选模型进行性能对比，以确定最优组合。

2.2 折外预测与模型训练

为避免信息泄露并确保集成效果的可靠性，第一层基学习器采用 K 折交叉验证（K-fold cross-validation）生成折外预测（Out-of-Fold，OOF）作为元学习器输入。具体步骤如下：

（1）折外预测生成：在源域（2020 年数据）上对每个基学习器执行五折交叉验证,每一折使用 4/5 的数据进行训练，剩余 1/5 的数据用于生成折外预测。重复至全部样本均被预测一次，得到与训练集样本数一致的 OOF 特征矩阵。

（2）元学习器训练：将四个基学习器的 OOF 预测按列拼接，构成新的特征空间输入元学习器。通过比较不同元模型在验证集上的决定系数（R^2）和平均绝对误差（MAE），确

定最佳元学习器配置。

3 实验结果分析

我们在源域对四类基学习器——LightGBM、CatBoost、XGBoost 与前馈神经网络（NN）——进行 k 折交叉验证并结合 Optuna 进行超参数寻优（以验证集 MAE 最小化为目标），随后将各基模型的折外预测（OOF）拼接成二级特征，训练四种元学习器（Ridge、SVR、GBDT、MLP）构成堆叠集成。实验结果见表 1

表 1　基于堆叠集成学习的住宅单位面积能耗预测模型

Table 1　Comparison of prediction performance between base learners and stacking ensemble models

Model / Meta-learner	R^2	MAE
LightGBM	0.625	0.566
CatBoost	0.632	0.559
XGBoost	0.607	0.582
Neural Network	0.519	0.658
Stacking（Ridge）	0.6347	0.5575
Stacking（SVR）	0.6346	0.5573
Stacking（GBDT）	0.6531	0.5450
Stacking（MLP）	0.6242	0.5694

结果表明，三种梯度提升类基模型在 2020 年的表现相近，其中 CatBoost 在单模型中取得最高精度（R2=0.632，MAE =0.559），而神经网络由于对样本规模与特征非平稳性更为敏感，表现相对滞后（R2=0.519，MAE =0.658）。

在此基础上，堆叠集成相较任何单一基学习器均呈现一致的增益，尤其是以 GBDT 为元学习器的堆叠在源域达到全表最佳的 R2=0.6531、MAE =0.5450，说明异构基学习器的残差互补与元回归的非线性加权能够更充分地吸收交互项与高阶非线性结构，从而推高域内上限约 2%-3%。

4 结论

本文针对住宅单位面积能耗预测中单一模型稳定性不足的问题，提出了一种基于堆叠集成学习的预测方法，并在 RECS 2020 数据集上进行了系统验证。研究得到以下主要结论：

（1）堆叠集成的优越性。相较于单一机器学习模型，堆叠集成学习能够有效融合异质基学习器的互补优势。实验结果表明，以 GBDT 为元学习器的堆叠集成模型在源域测试中达到 R^2=0.6531、MAE=0.5450，较最优单一模型（CatBoost）分别提升 3.3%和 2.5%，验证了该方法在提升预测精度和稳健性方面的有效性。

（2）工程应用价值。本研究提出的方法具有良好的可扩展性和工程适用性，可为住宅能源管理部门、节能改造项目评估以及建筑能效政策制定提供技术支撑，对推动建筑领域节能减排具有实际意义。

[参考文献]

[1]YE Y，ZUO W，WANG G. A comprehensive review of energy-related data for U.S. commercial buildings [J]. Energy Build，2019，186：126-37.

[2]KUMAR MOHAPATRA S，MISHRA S，TRIPATHY H K，et al. A sustainable data-driven energy consumption assessment model for building infrastructures in resource constraint environment [J]. Sustainable Energy Technologies and Assessments，2022，53：102697.

[3]WANG Z，WANG Y，SRINIVASAN R S. A novel ensemble learning approach to support building energy use prediction [J]. Energy Build，2018，159：109-22.

[4]UIDHIR T M，ROGAN F，COLLINS M，et al. Improving energy savings from a residential retrofit policy：A new model to inform better retrofit decisions [J]. Energy Build，2020，209：109656.

[5]DAI Z，HUANG W. Improving energy management practices through accurate building energy consumption prediction：analyzing the performance of LightGBM，RF，and XGBoost models with advanced optimization strategies [J]. Electrical Engineering，2025.